Annual Review of Biomedical Data Science

RNA Sequencing Data: Hitchhiker's Guide to Expression Analysis

Koen Van den Berge,[1,*] Katharina M. Hembach,[2,*] Charlotte Soneson,[2,3,*] Simone Tiberi,[2,*] Lieven Clement,[1,†] Michael I. Love,[4,†] Rob Patro,[5,†] and Mark D. Robinson[2,†]

[1]Bioinformatics Institute Ghent and Department of Applied Mathematics, Computer Science and Statistics, Ghent University, 9000 Ghent, Belgium

[2]Institute of Molecular Life Sciences and SIB Swiss Institute of Bioinformatics, University of Zurich, 8057 Zurich, Switzerland; email: mark.robinson@imls.uzh.ch

[3]Current Affiliation: Friedrich Miescher Institute for Biomedical Research and SIB Swiss Institute of Bioinformatics, 4058 Basel, Switzerland

[4]Department of Biostatistics and Department of Genetics, University of North Carolina, Chapel Hill, North Carolina 27514, USA

[5]Department of Computer Science, Stony Brook University, Stony Brook, New York 11794, USA

Annu. Rev. Biomed. Data Sci. 2019. 2:139–73

First published as a Review in Advance on April 30, 2019

The *Annual Review of Biomedical Data Science* is online at biodatasci.annualreviews.org

https://doi.org/10.1146/annurev-biodatasci-072018-021255

*These authors contributed equally to this article

†These authors contributed equally to this article

Keywords

RNA sequencing, gene expression, high-dimensional data, differential expression analysis, expression quantification

Abstract

Gene expression is the fundamental level at which the results of various genetic and regulatory programs are observable. The measurement of transcriptome-wide gene expression has convincingly switched from microarrays to sequencing in a matter of years. RNA sequencing (RNA-seq) provides a quantitative and open system for profiling transcriptional outcomes on a large scale and therefore facilitates a large diversity of applications, including basic science studies, but also agricultural or clinical situations. In the past 10 years or so, much has been learned about the characteristics of the RNA-seq data sets, as well as the performance of the myriad of methods developed. In this review, we give an overview of the developments in RNA-seq data analysis, including experimental design, with an explicit focus on the quantification of gene expression and statistical approaches

for differential expression. We also highlight emerging data types, such as single-cell RNA-seq and gene expression profiling using long-read technologies.

INTRODUCTION: OVERVIEW OF THE RNA SEQUENCING ASSAY

> After that it gets a bit complicated, and there's all sort of stuff going on in dimensions thirteen to twenty-two that you really wouldn't want to know about. All you really need to know for the moment is that the universe is a lot more complicated than you might think, even if you start from a position of thinking it's pretty damn complicated in the first place.
>
> —*Mostly Harmless* by Douglas Adams

Molecular biologists use gene expression studies to get a snapshot of the RNA molecules present in a biological system, which dictates what cells are doing or are capable of. The original RNA sequencing (RNA-seq) protocols, published over 10 years ago (1–5), described the sequencing of complementary DNA (cDNA) fragments on a large scale from a population of cells. Since then, the system has been optimized for different types and qualities of starting material, as well as different research questions, and many mature protocols are available.

A basic overview of the main steps in a standard RNA-seq experiment is given in **Figure 1**. The first step is the extraction and purification of RNA from a sample, followed by an enrichment of target RNAs. Most commonly used is poly(A) capture, to select for polyadenylated RNAs, or ribosomal depletion, to deplete ribosomal and transfer RNAs that are highly abundant in a cell (approximately 95% of total RNA) (6) and are usually not of primary interest (7). The selected

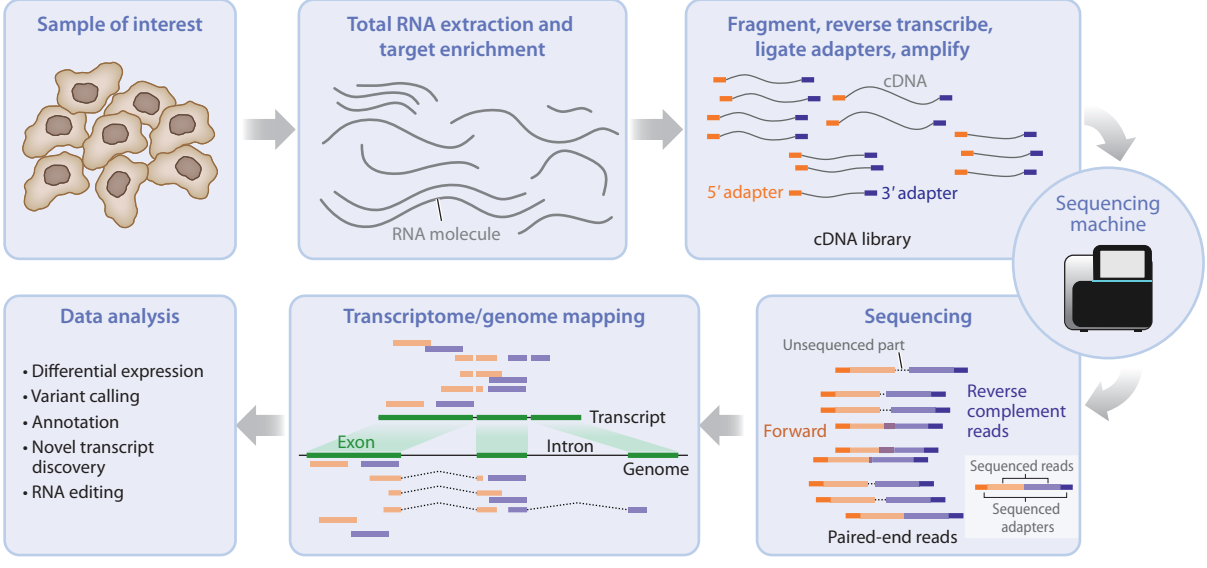

Figure 1

Overview of the experimental steps in an RNA sequencing (RNA-seq) protocol. The complementary DNA (cDNA) library is generated from isolated RNA targets and then sequenced, and the reads are mapped against a reference genome or transcriptome. Downstream data analysis depends on the goal of the experiment and can include, among other things, assessing differential expression, variant calling, or genome annotation.

RNAs are then chemically or enzymatically fragmented to molecules of appropriate size (e.g., 300–500 bp for Illumina's TruSeq). Current dominant systems (e.g., Illumina) only sequence DNA; single-stranded target RNAs are thus reverse-transcribed to cDNA (first strand), the RNA is then degraded, and the first-strand cDNA is complemented to a double strand. Adapter sequences are either ligated to the 3′ and 5′ end of the double-stranded cDNA or used as primers in the reverse transcription reaction. The final cDNA library consists of cDNA inserts flanked by an adapter sequence on each end. In the last step, the cDNA library is amplified by polymerase chain reaction (PCR) using parts of the adapter sequences as primers.

For Illumina sequencing, the library is loaded onto a flow cell where the cDNAs bind to short oligonucleotides complementary to the adapter sequence. Bridge amplification creates dense clonal clusters of each cDNA loaded (8). The sequence of each cluster is determined by a process called sequencing by synthesis (9): Single-stranded templates are read as the complementary strand is generated. A single fluorescently labeled deoxynucleoside triphosphate (dNTP) is added in each step. The label acts as a terminator and prevents the incorporation of more than one dNTP at the same time. After the fluorescent label has been imaged, it is enzymatically cleaved and the next dNTP can bind to the chain. Base calls are inferred directly from the measured fluorescent signal intensity.

cDNA libraries can be sequenced in one of two modes: single-end or paired-end. In single-end mode, only one end of the cDNA insert is sequenced, whereas in paired-end mode, both ends are sequenced, yielding two reads in opposite orientation, one from each end.

There are protocols for unstranded and stranded RNA-seq (10, 11), where the latter preserves information about the coding strand of each fragment, which is useful in compact genomes or with expressed RNAs that originate from opposite strands of the same genomic locus. One possibility to construct a stranded library is to use deoxyuridine triphosphates (dUTPs) in the generation of the second strand cDNA and to degrade the dUTP-labeled cDNA before PCR amplification (12). Other protocols use alternative adapters to distinguish between 5′ and 3′ ends of the RNA (13).

RNA-seq has greatly evolved over time, with early experiments having 35-bp reads and modern (Illumina-based) experiments typically employing 50-bp (single-end) or 100-bp (paired-end) reads (**Figure 2*a***). Most RNA-seq experiments comprise between 10 and 100 million reads, with a trend toward deeper sequencing over time (**Figure 2*b***). The number of samples per project has remained constant over the years, with a median of around eight samples (**Figure 2*c***). Rapid enhancements in sequencing technology have enabled not only longer read lengths (e.g., 250–300 bp for Illumina's MiSeq) and much higher throughput for the same cost, but also much lower amounts of required starting material. Meanwhile, third-generation technologies, such as Pacific Biosciences (PacBio) and Oxford Nanopore Technologies (ONT), allow the sequencing of single molecules and have now been used for sequencing full-length transcripts on a transcriptome-wide scale (14). Further developments are summarized below in the section titled Long-Read Transcriptome Sequencing. In addition, single-cell RNA-seq (scRNA-seq) is a rapidly emerging technique that can be used to sequence the sparse transcriptome of individual cells. Some of the early developments in this area are captured in the section titled Single-Cell Transcriptome Sequencing.

Design Aspects of RNA-seq

The basics of scientific experimental design apply equally for RNA-seq experiments (e.g., see Reference 16). For example, whether the desired experiment is a simple two-group design or a full factorial design, one should consider randomizing experimental units to treatments to avoid confounding factors (e.g., via blocking over batches). If the experiment is run in multiple batches

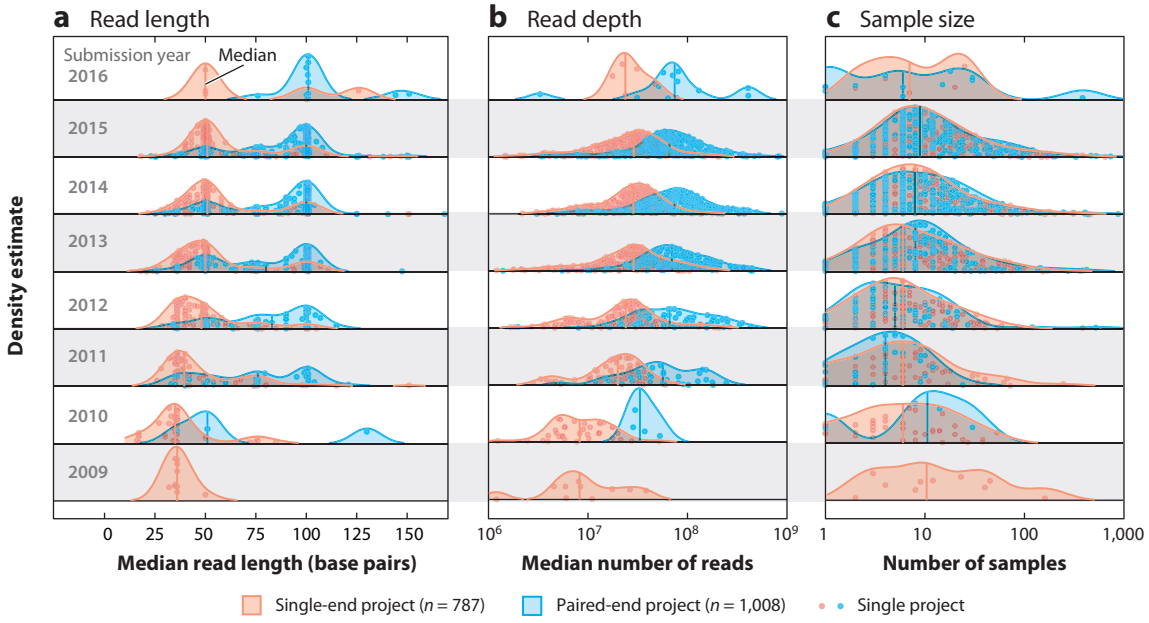

Figure 2

Ridge plots showing the progression of read length, depth, and sample size in Sequence Read Archive (SRA) projects using the `recount` package (15). The projects are separated by the submission year of the biosample. (*a*) Median read length of all samples per project and year. (*b*) Median number of reads across all samples per project and year. (*c*) Number of samples in each project. Each point represents one project.

(e.g., a limited number of samples per run), it is critical to represent every experimental condition in each batch, so that, when comparing conditions, differences within a batch can be averaged over in the statistical modeling.

Specific aspects to be considered while designing an RNA-seq experiment include the number of replicates and the depth of sequencing. Ultimately, in modern genomic experiments where resources (e.g., material from subjects) are scarce and the RNA-seq experiment is itself a hypothesis-generating tool, the first driver of sample size is budget. Many RNA-seq studies use as few as three replicates per condition (**Figure 2*c***), near the minimum required to do any statistical analysis.

Sample size calculators can compute the required number of samples to achieve a user-defined power for detecting differential expression (DE) (17–20). However, the user must define many parameters, such as the expected alignment rate, the desired power, the significance level, and the log-fold change (LFC) of DE genes. A recent study concluded that the recommended sample sizes vary across tools, even when estimates from pilot data are available (21). Another issue with sample size calculators is how to precisely define the outcome: Do we want to find as many DE genes as possible? Do we want a certain power for the lowly expressed genes or the highly expressed ones? In many cases, RNA-seq experiments are exploratory and thus a means to further experimentation.

Nonetheless, there is a trade-off between the number of samples and the sequencing depth in terms of discovery performance. Increasing the number of reads might seem always beneficial, but a large proportion of the reads originate from a small pool of highly expressed genes, and there is effectively no signal saturation. **Figure 3** highlights that more than 80% of reads are attributed to the 10% most expressed genes, acknowledging that transcript length also plays a role (22). An increased number of reads only marginally increases the coverage of lowly expressed genes,

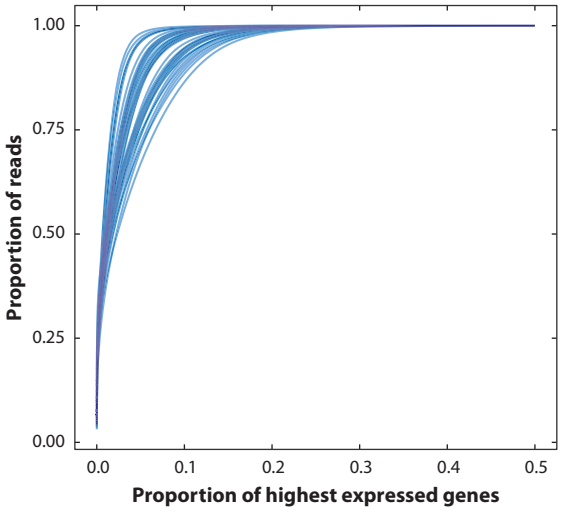

Figure 3

Cumulative proportion of reads among the top expressed genes. The *x*-axis orders genes according to the total number of reads they attract, and the *y*-axis displays the cumulative fraction of total reads. Each line represents a single sample. Counts were downloaded from `recount` (15), and 50 samples were randomly selected from accession number SRP060416.

and therefore, the statistical power to detect DE does not improve considerably, especially if the experiment already comprises ~10 million reads per sample (23). In most cases, the budget is better spent on replicates. For example, Schurch et al. (24) showed that a higher number of replicates is required to identify DE genes with low-fold change, and that ideally at least six replicates per condition should be used.

There are options for additional capture of genes with low expression, but these require additional labor and cost. In targeted RNA-seq (RNA CaptureSeq), specific regions are first captured by probes that are complementary to the region of interest and these selected regions are prepared and sequenced (25, 26). After capture, the quantitative nature of the assay is maintained (25); such capture is especially useful in degraded samples (e.g., patient material stored in paraffin blocks) where the poly(A) tails may not be present.

RNA-seq Applications

Clearly, the popularity of RNA-seq is driven by its large number of applications. One obvious application area is genome annotation. Even the well-studied transcriptomes of humans or model organisms such as mice, zebrafish, or fruit flies are not complete. Thus, transcriptomics is used to annotate novel transcriptional events, such as exon skipping, alternative 3′ acceptor or 5′ donor sites, or intron retention, and to understand their usage in normal, developmental, or pathological conditions. Transcriptomic studies identified previously unknown phenomena, such as microexons (27), cryptic exons (28), so-called skiptic exons (29), circular RNAs (30), enhancer RNAs (31), fusion genes (32), and so-called epitranscriptomics involving RNA base modifications (33).

One of the main application areas is gene regulation. RNA-seq enables the comparison of gene/transcript/exon expression between different tissues, cell types, genotypes, stimulation conditions, time points, disease states, growth conditions, and so on. Ultimately, the goal of such comparisons

is to identify the genes that change in expression to understand the molecular pathways that are used or altered or the regulatory components that are utilized.

Gene expression has been used for the molecular subclassification of cancer since the early days of microarrays (34). RNA-seq offers this same capacity but at higher resolution and can include, for example, categorization by splicing (35). There is considerable interest in using RNA-seq in clinical applications to augment or corroborate the information given by genome sequencing (36, 37). Other applications include spatial transcriptomics, where cellular positional information is maintained in the preparation of cDNA fragments (38); host–pathogen interactions via dual RNA-seq, where the transcriptomes of both host and pathogen are simultaneously assayed (39); the analysis of genetic variation among expressed genes (40); RNA editing events (41); the characterization of long noncoding RNAs (42); and metatranscriptomics (43).

Despite the many use cases for bulk RNA-seq, there are applications where single-cell resolution is desired, especially when studying heterogeneous tissues that consist of more than one cell type. While bulk RNA-seq can be computationally deconvoluted to estimate the composition of cells present (44), it is not possible to discover new cell types or perform cell-type-specific analyses with bulk RNA-seq, and thus scRNA-seq opens the door to new applications.

Outline

This review focuses on data analysis aspects, the computational steps involved (focusing on DE), various statistical and computational challenges, and the approaches that have been proposed to address them. We focus on Illumina-based RNA-seq data on model organisms, as that is the dominant application area. There are already excellent reviews for major application or computational areas, such as de novo (or reference-based) transcriptome assembly (45), allele-specific expression analyses (46), expression quantitative trait loci mapping (47), splicing (48), analysis of gene regulatory networks (49), and pathway analyses (50, 51). In most applications, the overarching goal is to identify DE, at either the gene, transcript, or exon level. The set of DE entities provides a snapshot into the molecular underpinnings of a stimulus, a disease condition, a genetic mutation, or any other perturbation being interrogated. In most cases, DE is only an intermediate (although critical) step to understanding the biological system under study.

The review is organized as follows. First, we discuss alignment and quantification, where RNA-seq reads are placed in the context of the genome or annotation catalogs and the relative expression level of each target is assessed. Following quantification, we discuss the basics of DE, to lay the foundation for the current frameworks, and variants of DE, to highlight the diverse conceptual tools available to run the discovery process. Finally, we discuss two rapidly evolving research areas that have experienced considerable activity in recent years, single-cell transcriptome sequencing and long-read transcriptome sequencing (LRTS).

ALIGNMENT AND QUANTIFICATION

After an experiment has been conducted, the analyst is presented with files containing up to billions of short cDNA fragments. Following sufficient quality control of the sequencing reactions, alignment to a reference genome or (de novo assembled) transcriptome is one of the critical steps in translating the raw data into something quantitative.

Because the sequenced fragments are derived from cDNA corresponding to fully (or partially) spliced transcripts, reads often span the boundaries of splice junctions (SJs), resulting in so-called junction-spanning reads (**Figure 4**). This results in contiguous read sequences whose constituent subsequences may be separated by tens of thousands of nucleotides. This poses a considerable

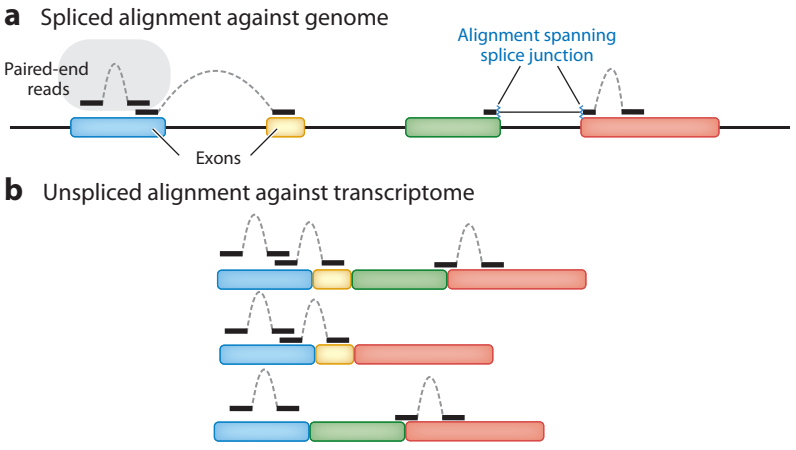

Figure 4

An illustration of spliced alignment of RNA sequencing (RNA-seq) fragments to a genome (*a*) and direct alignment to a transcriptome (*b*). Reads are designated by thick solid lines, while dashed arcs represent the pairing relationship between paired-end reads. This illustration depicts alignment to a single four-exon gene consisting of three distinct transcripts. In the spliced alignment (*a*), the left read of the rightmost pair is a junction-spanning alignment to the red–green exon boundary. In the direct alignment to the transcriptome (*b*), one observes how the same alignment (e.g., the alignment to the blue exon) is repeated for each transcript.

computational challenge, as the position of SJs in spanning reads needs to be accurately identified for a read to be properly aligned. There are two main approaches for handling spliced reads, each with its own challenges and benefits: a spliced alignment against a reference genome or an unspliced alignment against a reference transcriptome (a database of all isoforms). A main challenge in spliced alignment against a reference genome is the proper alignment of reads that span an SJ, especially when these junctions are not annotated a priori. Meanwhile, the main challenge in unspliced alignment to a transcriptome is the redundant sequence among related isoforms, which often leads to a high multimapping rate.

Spliced Alignment to a Reference Genome

A popular solution for handling RNA-seq alignments is to use a splice-aware aligner. Early RNA-seq aligners, e.g., TopHat (52), made use of DNA-seq aligners, such as Bowtie (53), by first building a catalog of putative SJs to which the reads can be directly aligned.

More recent splice-aware alignment tools (54–64) account for read splicing directly. They also can utilize the locations of known SJs and discover previously unannotated SJs. When a read partially aligns, the annotated SJ database is consulted to check if the alignment ends prematurely as the result of the read spanning a known splice site. In this case, compatible downstream splice sites can be considered as candidate loci to align the remaining portion of the read. Even if no annotated splice site exists at the point where the alignment ends, the tool can interrogate the terminal nucleotides in the partial alignment to see if they are compatible with known canonical (or user-provided) donor or acceptor sites, providing evidence that the partial alignment stops as the result of a splicing event.

One of the primary difficulties in aligning reads across SJs is that only a small portion of the read spans into one of the exons. Splice-aware aligners including STAR (55), HISAT(2) (56),

Subread (57), and GMAP (58) attempt to deal with such cases by using evidence from reads that confidently align across SJs. In such strategies, new SJs are added to the index when they display high-confidence evidence, i.e., when multiple reads with a sufficient anchoring sequence span the SJ. Trusted SJs are then used to help align reads that start or end near junction boundaries.

Unspliced Alignment to a Reference Transcriptome

In organisms where transcriptomes are well characterized, an alternative to splice-aware genome alignment is direct transcriptome alignment, which consists of aligning against a set of known transcripts. Since the transcript sequences are already spliced, reads should align contiguously, and many of the computationally expensive steps and heuristics can be avoided. Moreover, when no reasonable quality reference genome is available for reference-based transcript assembly (e.g., when a transcriptome has been assembled de novo), alignment directly to the assembled transcripts is the only available option. However, transcriptome alignment induces a high degree of multimapping, and dealing with this becomes a primary computational challenge. For example, if a gene has three distinct isoforms, a constitutive exon of this gene will appear three times in the transcriptome reference (e.g., **Figure 4*b***). Additionally, mapping only to annotated transcripts does not allow one to find novel splicing or expression patterns (e.g., novel exons), and it becomes difficult to assess retained introns or partial splicing; of course, it is possible to augment the transcriptome with unspliced variants. The choice of genome versus transcriptome alignment is largely driven by the desired target application and the constraints of downstream analyses.

Gene- and Transcript-Level Quantification From RNA-seq Data

One of the main uses of RNA-seq is to assess gene- and transcript-level abundances. Accurate abundance estimation is crucial to common downstream applications, including assessing all the notions of DE. Most commonly, abundances are estimated at the level of genes, but recently transcript-level abundances have become more widely used, and there are trade-offs in choosing between the two levels of resolution.

Gene-level quantification consists of assigning fragments (reads or read pairs) to genes, where the gene is often taken to represent the amalgamation of all transcripts produced from a specific strand at a specific locus (65), which typically share some exons or parts of exons. The total expression of a gene is the sum of the expression of its isoforms. Any fragment arising from any isoform of a gene is assigned to the underlying gene. There are typically two paths that can be taken to obtain gene-level quantifications: direct fragment overlap counting of gene features, and transcript-level quantification followed by aggregation to the gene level.

Direct fragment counting of gene features is done by first mapping RNA-seq reads to the genome with a splice-aware aligner, and then using a tool like featureCounts (66), HTSeq (67), or the built-in capability of STAR (55) to assess how many fragments overlap each gene; the same approach can be used to quantify other disjoint genomic features, such as nonoverlapping exonic segments. Even in this basic pipeline, there are many ways certain conditions can be handled. For example, should a fragment reside completely within a feature to be counted? If a fragment maps to multiple features, should it be discarded, counted toward each feature, or somehow partially allocated? Of course, direct fragment counting approaches exhibit desirable features: They are conceptually simple and typically quite fast. Conversely, they suffer from various disadvantages: They have no principled way of handling multimapping reads (e.g., arising from paralogous genes), and they are oblivious to potentially important compositional changes not reflected directly in gene-level read counts (e.g., isoform switching). Additionally, since such methods assess the frequency

of reads overlapping a gene, they must grapple with the definition of a gene. For example, should a gene be the union or the intersection of exons of all transcripts of the gene? Should intronic reads be included? Although the concept of a gene is a useful abstraction, transcripts are assayed in RNA-seq and so present a conceptually cleaner target for quantification.

Transcript-level quantification consists of the assignment of fragments to specific transcripts, which is more challenging but has a number of advantages. It admits a clear interpretation, since transcripts are what the cell expresses; it allows for improved biological resolution and decoding of potentially important biological changes, such as isoform switching; it is the most appropriate level to model and correct for technical biases (68–71); and it provides a proper model for handling reads that multimap, as failing to do so can lead to systematically poor quantification for genes in gene families (72). Solving the transcript-level abundance estimation problem requires a principled solution to aggregating to gene-level estimates (73–75). Conversely, transcript-level quantification is not without disadvantages. Alternative splicing implies that many fragments are ambiguous in their origin, and they must be assigned probabilistically, necessitating the adoption of a model, which may fail to adequately capture reality; this read ambiguity translates to additional uncertainty in the estimated transcript abundances.

Transcript Quantification

Methods for transcript quantification are based primarily on defining a generative model of RNA-seq reads and then trying to perform inference on this model to obtain the relevant quantities (i.e., transcript abundances); see **Figure 5**. There has been a tremendous amount of research on quantifying transcript-level abundance from high-throughput sequencing data; here we describe a few major highlights.

Initial probabilistic frameworks for transcript identification and abundance estimation using EST (expressed sequence tags) data were already being developed before Illumina-based sequencing (76), but Jiang & Wong (77) were among the first to attempt isoform-level abundance estimation using RNA-seq data. They defined counts over exons and exon junctions as arising according to a Poisson model and viewed transcripts as vectors of inclusion and exclusion of these exons and junctions. By expressing the likelihood of the model parameters given the observed data, they posed a statistical model that admits efficient inference, for which they obtained the point estimate by gradient ascent and provided estimates of the posterior distributions of the parameters via importance sampling. This work represents one of the first proper statistical formulations of the problem. However, the approach does not account for fragments that map to multiple genes, and it requires annotations of transcripts in terms of the gene–transcript relationship, as well as the exon and junction read inclusion matrix.

Li et al. (73, 78) proposed one of the most widely adopted generative models for transcript quantification, RSEM. They defined a fragment-level model of RNA-seq experiments in terms of sampling molecules from an underlying population, proportional to the product of their abundance and length, and then generated fragments from the sampled molecules. Primary quantities of interest are estimated, including the nucleotide fractions (the fraction of all sequenced nucleotides deriving from each transcript) and the transcript fractions (the fraction of all transcripts in the initial population comprising each transcript species); these quantities can be directly converted into popular abundance units, such as transcripts per million (TPM) or estimated counts. Notably, they proposed computing the maximum likelihood (ML) estimates using an expectation-maximization (EM) algorithm (see **Figure 5**) and introduced a modified Gibbs sampling procedure to allow estimating credible intervals for the abundance estimates (73). The model is quite general: It works at the fragment level, and it can account for numerous protocol-related aspects,

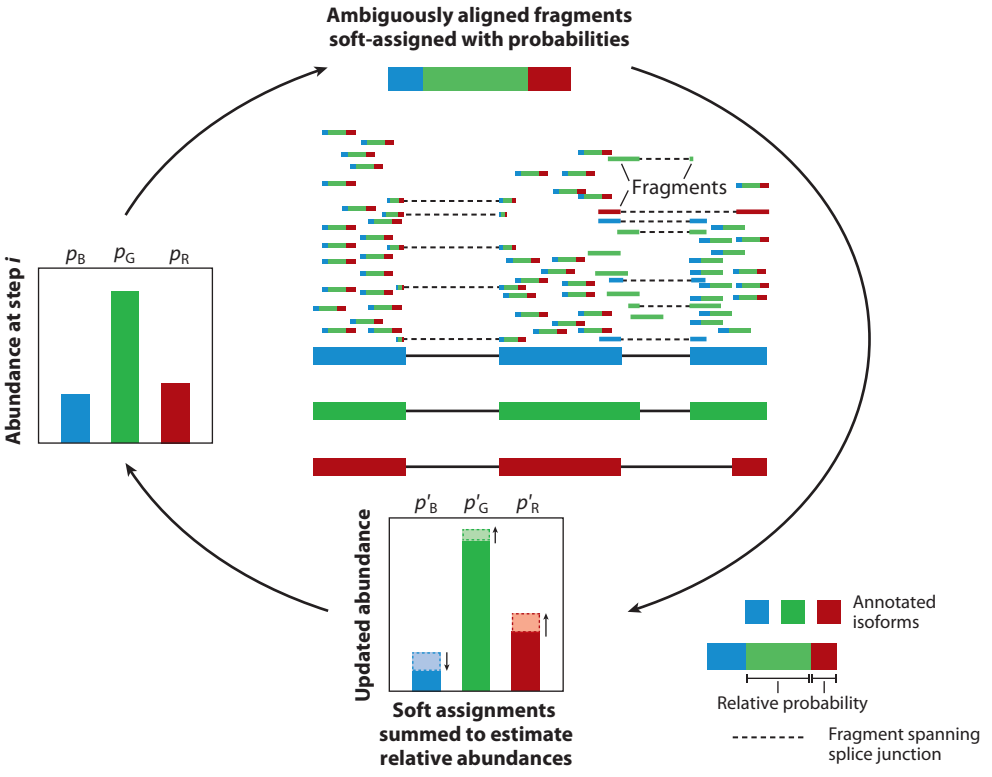

Figure 5

An illustration of the alignment of various reads to a gene with three isoforms: blue (B), green (G), and red (R). In this example, we wish to estimate the abundances of these isoforms, but most reads have ambiguous origins and need to be probabilistically assigned to the transcripts (relative probabilities for each read are shown by the magnitudes of the three colors). Some reads are consistent only with the B and G transcripts, and a few reads uniquely align to a single transcript (single color). In the expectation-maximization (or related) algorithm, given the current abundance estimates, fragments are probabilistically assigned to transcripts, and then estimated abundances are updated by summarizing the (proportional) allocations over all fragments; transcript abundance estimates are determined by iterating the procedure until convergence.

including single-end and paired-end sequencing, directional versus unstranded protocols, various coverage biases, etc. Further, the model relies only on the transcript sequences and not on the relationships to genes or annotations of exons and SJs. Thus, it can be easily applied to both well-characterized and newly assembled transcriptomes. One drawback of a fragment-level model, however, is that each EM iteration scales with the total number of alignments, which is indeed large in most RNA-seq experiments.

Instead of modeling each fragment individually, MMSeq models sufficient statistics (79, 80). Reads are categorized into equivalence classes, where two reads are equivalent if they align to the same set of transcripts. The approach works both within and across genes and does not require the shared regions that give rise to the equivalence classes to correspond to any known annotation (e.g., exon or SJ). MMSeq uses an EM method that works directly over these equivalence classes, allowing efficient inference of transcript-level abundance in this model. In addition to this ML approach, a Gibbs sampling procedure was introduced that can estimate transcript abundances using summary statistics from samples of the estimated posterior, which also allows one to assess uncertainty in the transcript-level abundance estimation and groups of transcripts with correlated

posterior estimates. The underlying likelihood function of the equivalence class–based model is not equivalent to that of the fragment-level model in RSEM, although subsequent work explored other factorizations of the full fragment-level likelihood that either preserved equality with the RSEM model while speeding up inference (81) or sacrificed equality to balance efficiency and fidelity (82, 83). eXpress demonstrated how fragment-level inference could be made much more efficient by modifying the inferential algorithm itself (i.e., online EM), rather than the factorization of the underlying likelihood function (84, 85).

Cufflinks is widely known as both a reference-guided transcript assembly algorithm and a quantification tool (86). Quantification either is restricted to a reference annotation or allows new transcripts to be identified via alignments; transcript abundances are estimated via an EM algorithm to determine the ML estimates given the observed data. While we do not focus on assembly methods here, given the close relationship between transcript identification (assembly) and quantification, numerous approaches attempt to solve both problems together, either stagewise or jointly (82, 87–94).

BitSeq introduced a model similar to RSEM that jointly performs quantification and DE, together with fully Bayesian inference (95). BitSeq focused on sampling from the posterior distributions of transcript abundances, given the fragment alignments, giving accurate estimates (96) and useful information about posterior uncertainty and posterior correlation, which is used in the DE step (95). To combat the heavy computational requirements, Hensman et al. (97) introduced a variational Bayesian (VB) approximation that can be efficiently optimized. TIGAR introduced a VB approach to the transcript abundance estimation problem (98), and the VB EM algorithm was shown to outperform the standard EM algorithm. However, Hensman et al. (97) introduced a novel optimization procedure called VBNG (VB natural gradient), which is a gradient ascent algorithm that considers the information geometry (99) of the underlying problem. Hensman et al. also suggested that EM-based methods tend to find solutions near the boundary of the parameter space, and that their quantifications are less robust than either fully Bayesian or VB estimates (97).

Many of these approaches, among others, simplify the model or improve the efficiency of the inferential procedure, but all rely on full alignments of each read, which can be computationally intensive and time consuming. Recently, several new methods bypass the alignment step and instead adopt lightweight models for quantification. Sailfish defines the transcript abundance likelihood in terms of the constituent *k*-mers of the underlying transcriptome and their abundance in the read data (100). Since the *k*-mers are completely known in advance, the relevant equivalence classes can be precomputed, which reduces the inferential problem to one of simply counting *k*-mers and performing inference via an EM algorithm such as the SQUAREM algorithm (101). This approach increases the speed of abundance estimation by over an order of magnitude compared to full alignment approaches. Building on the idea of *k*-mer-based abundance estimation, RNA-Skim takes the approach of Sailfish even further, identifying sets of distinctive *k*-mers, termed sigmers (102). Transcripts are clustered into groups, and sigmers are identified as *k*-mers that are unique to (and indicative of) each cluster. Quantification is then performed by counting the sigmers in the read data, instead of all *k*-mers, and the EM algorithm is used to estimate transcript abundances from sigmer equivalence class counts. While very fast, these *k*-mer-based approaches do not retain the coherence of the *k*-mers along a read, which can reduce specificity, and they cannot easily estimate certain aspects of the generative model, like the fragment length distribution. Addressing these shortcomings, kallisto relies on the use of pseudoalignments to directly compute the sufficient statistics of the equivalence class–based model of transcript abundance estimation (103). This approach uses *k*-mers to identify the transcripts with which fragments are compatible, but does not treat the *k*-mers independently. The pseudoalignments can be computed in such a way that

equivalence class counts are generated without considering or computing individual fragment-to-transcript alignments, and this can often be achieved by querying only a small number of the *k*-mers present in a fragment, allowing for efficient and accurate estimation in the equivalence class–based model using an EM algorithm. Salmon is another lightweight quantification approach that avoids full alignments, although they can still be used as input (104). It uses a two-phase algorithm for transcript abundance estimation: an online phase using a stochastic collapsed VB inference algorithm (105), where abundances and auxiliary parameters are estimated (e.g., GC bias parameters, sequence-specific bias parameters, fragment length distribution), and an update using mini batches of mappings. Salmon uses a lightweight mapping algorithm to compute the likely transcripts, positions, and orientations of origin of each fragment and adopts a fragment-level GC bias modeling approach (71), which reduces misidentification of expressed isoforms when read coverage is not uniform along the transcripts due to GC content. In the offline phase, a factorized likelihood function is optimized until parameter convergence. The granularity of the likelihood factorization used by Salmon can be adjusted (83) in a way that allows one to trade off between the fragment model of RSEM and the count-based model of MMSeq. In the offline phase, the factorized likelihood is optimized using a VB EM algorithm (98) or a traditional EM algorithm. Combining the efficient determination of fragment–transcript compatibility with RNA-Skim's sigmer concept, Fleximer uses a new matching algorithm that uses sets of sigmers to determine the likely loci of origin of reads, instead of treating each sigmer independently (106). A generalized suffix tree is used to organize the reference sequences, and a segment graph that demonstrates how segments of sequence are shared among reference transcripts is used to select an informative and robust set of sigmers for quantification. Reads are mapped against the reference by matching them to sigmers using a precomputed automaton. This process produces a set of transcript equivalence classes, along with a corresponding count for each sometimes termed the transcript compatibility count; this is used with an EM algorithm to estimate transcript abundances.

Due to their vastly improved speed, ease of use, and reduced computational requirements, alignment-free approaches have become popular for assessing transcript- and gene-level abundance using RNA-seq data. Recent benchmarks (107–110) suggest that, in addition to being fast, such methods can produce accurate abundance estimates—at least to the extent that simulation-based studies, which sometimes adopt the assumed generative models of the quantification approaches, can be relied upon to assess such accuracy. However, there remain opportunities for improving transcript-level quantification methods. For example, the underlying models can likely be further enhanced to account for complexities in the fragmentation patterns of molecules prior to sequencing (111), to better balance robustness and sample-specific accuracy (112), and to address uncharacterized biases. Additionally, most of these approaches (lightweight and otherwise) assume that the annotation of transcripts to be quantified is complete. The accuracy of quantification can suffer when this is not the case, although it is possible to computationally flag transcripts whose estimates are unreliable (113).

BASICS OF DIFFERENTIAL EXPRESSION

Following alignment and quantification, the next challenge often is assessing DE from the estimated feature abundances. We first present a general context and describe the statistical frameworks and overall workflow. The starting point is a count table with rows representing features (e.g., genes) and columns representing samples (i.e., experimental units). The goal of DE is to formulate and test a statistical hypothesis for each feature. Depending on the experimental design, the context, and the research question, more complex analyses are often required. We elaborate on further variations of the overall workflow in the section titled Variants of Differential Expression.

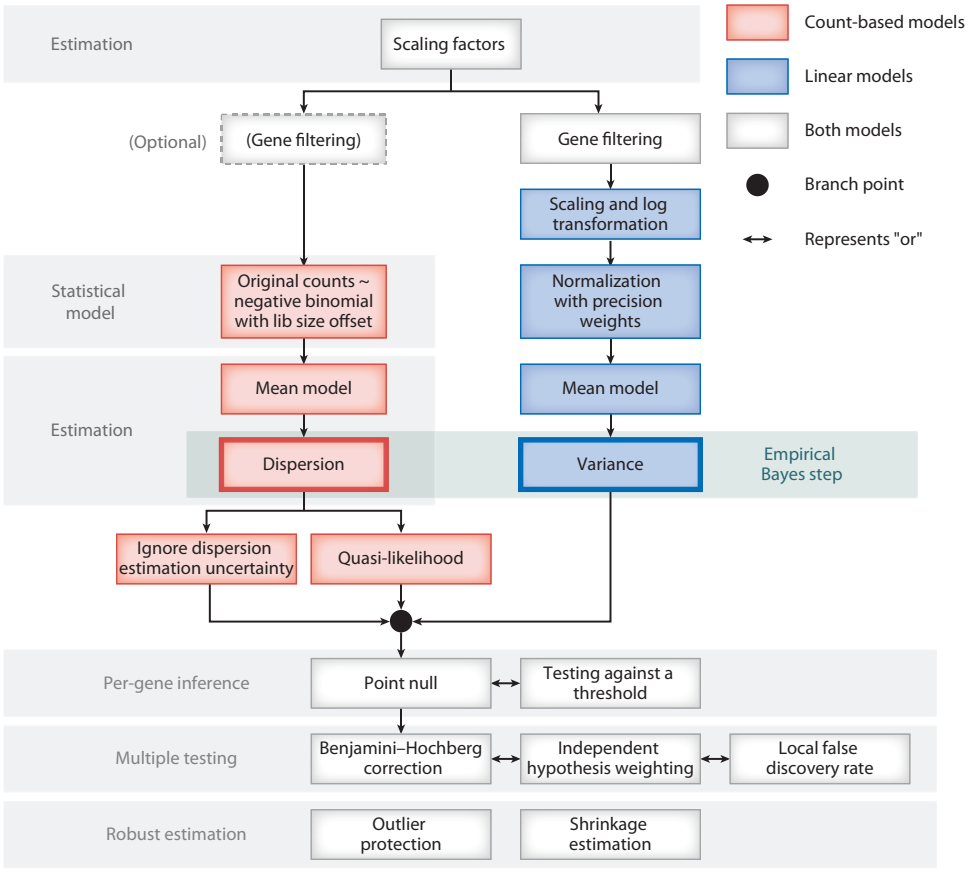

Figure 6

Schematic overview of a DE analysis for RNA sequencing data. Red boxes correspond to pipelines for count-based models (e.g., edgeR, DESeq2), while blue boxes correspond to a linear-model-based pipeline (e.g., limma-voom).

The general workflow involves the following steps (see **Figure 6**): filtering and normalization (preprocessing), specification of the statistical model and estimation of model parameters, statistical inference on the relevant parameters, and adjustment for multiple testing. We introduce this general workflow from the perspective of classical models for count regression. We then discuss various notable deviations, including alternative estimation and inference frameworks and additional strategies to ensure robustness.

Typically, only a limited number of replicates are available (e.g., three to five replicates per condition). The achievable statistical power from such small sample sizes can be low, even for a single feature, with the real interest lying in inference on thousands of features simultaneously. This parallel inference challenge is common to various genome-scale experiments, and the statistical community has contributed strategies to improve the overall performance, from which a few themes have emerged. For example, in estimating parameters for a given feature, one should consider the information coming from the other features in the data set (114). In general, genomics data are ripe for using empirical Bayes methods to moderate estimates, where priors for a feature are derived from a suitable set of other features measured in the data set. In addition,

moderating variance parameters is critical, and indeed, much of the success of earlier parallel inference frameworks (e.g., for microarrays) can be attributed to variance moderation, whether in an ad hoc strategy (115) or in hierarchical models (116). Other tricks such as regularization of regression parameters or considerations for robustness provide additional performance benefits. Taken together, the challenges associated with vast parallel inference can be greatly eased by adopting one or more of these strategies.

Preprocessing: Filtering and Normalization

The vast number of features in a typical RNA-seq experiment leads to a large multiple testing burden. However, many features are largely uninformative; for example, features with low expression provide little evidence for DE. Therefore, filtering strategies are employed that predominantly remove uninformative features and reduce the multiple testing burden. Bourgon et al. (117) showed that filtering is valid if it is independent of the DE test statistic; thus, filtering on residual variance is invalid, while filtering on expression strength, as is commonly done, is valid.

The observed counts of the features cannot be directly compared across samples since there are differences in sequencing depth across libraries. Several methods have been developed to normalize counts to facilitate cross-sample comparisons, although in most count-based models, the counts themselves are not modified and instead scaling factors accompany the analysis. Initial attempts focused on a simple correction for sequencing depth, using the total sum of counts for each sample (i.e., the library size) as a scaling factor (3, 118). However, variation in library preparation or RNA composition between samples also contributes to cross-sample variability and should be accounted for (119). In addition, a few highly expressed genes can largely drive the sampling of fragments, thus leading to inaccurate scaling of the counts. A popular approach is to calculate a size factor (119, 120) for each sample. This can be considered a robust global fold change between the current sample and a (pseudo) reference sample derived from all samples. DESeq's median of ratios method and edgeR's trimmed mean of *M*-values (TMM) method (where *M*-values denote empirical fold changes between two samples) are the most popular scaling approaches (121). Both procedures assume that most genes are not DE and adopt robust summarization methods to calculate the size factors (effective library sizes) to reduce the impact of DE genes (TMM uses a trimmed weighted mean; DESeq uses the median of the log-expression ratios). More advanced normalization methods have since emerged to address other technical artifacts such as GC content and transcript length effects and to accommodate within- and between-lane normalization, e.g., CQN (122) and EDASEQ (123). Moreover, methods based on external spike-in features have been introduced to address normalization for applications where many features are DE or where the basic assumptions of conventional normalization methods are violated (124–126). Recently, a normalization technique has been proposed for RNA-seq data with large differences between conditions that assumes similar distributions in biological replicates, while accommodating for differences between conditions (127).

The normalization size factors are built into the DE analysis workflow as offsets in the statistical models (see below). Notably, size factors are treated as fixed and known, although they are actually random variables estimated from the data (128), and it is unclear how ignoring their associated uncertainty affects the downstream DE analysis.

Modeling and Estimation

Because of the typically small sample size, DE tools mainly implement parametric methods (120, 129–132). Initially, count data were log-transformed and linear models were used for DE analysis

(4). However, log-transformed counts suffer from heteroscedasticity (a systematic mean-variance trend) intrinsic to count data, rendering the standard linear model, which assumes homoscedasticity, suboptimal. In addition, fitting continuous models to (transformed) count data introduces a further approximation. Therefore, discrete count distributions have gained more traction in the initial frameworks.

Gene expression variability across technical replicates (i.e., resequencing the same sample), so-called shot noise, has been shown to approximately follow a Poisson distribution (118), for which the variance is equal to the mean. Biological replication introduces additional cross-sample variability, and analysis frameworks therefore have resorted to one of the natural extensions, the gamma-Poisson or the negative binomial (NB) distribution, which has an additional dispersion parameter and a quadratic mean-variance relationship,

$$Y_{fi} \sim \mathrm{NB}(\mu_{fi},\ \varphi_f),$$
$$\mathrm{Var}(Y_{fi}) = \mu_{fi} + \varphi_f \mu_{fi}^2,$$

where Y_{fi} denotes the read count of feature f in sample i, φ_f is the dispersion for feature f, and $\mu_{fi} = s_i \theta_{fi}$ represents the average expression, which is driven by the true (relative) mRNA concentration in the sample, θ_{fi}, multiplied by a normalization scaling factor, s_i; there also exists a characteristic dispersion-mean trend in RNA-seq data sets (**Figure 7*a***). Initial implementations focused on two-group comparisons (120, 133) and were later extended to the generalized linear model (GLM) framework, an extension of classical linear models to non-Gaussian responses (134). GLMs allow for the inclusion of multiple treatments or covariates, thus broadening the

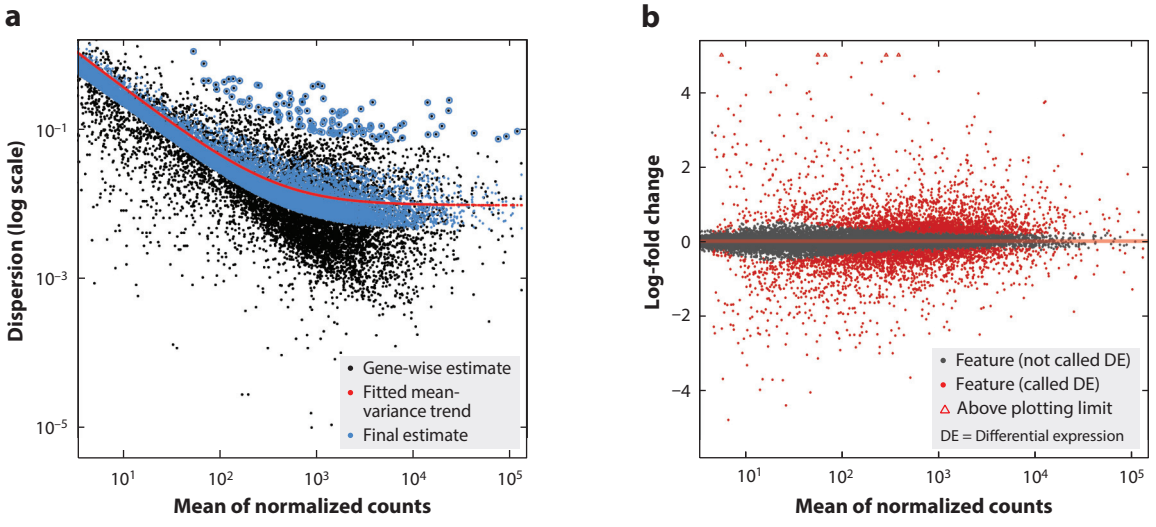

Figure 7

(*a*) A dispersion-mean plot of the RNA sequencing (RNA-seq) experiment from Reference 135, as processed in Reference 136. The dispersion trend smoothly decreases for genes with higher expression and eventually reaches an asymptote, which can be considered the biological variability present in the data set for a typical gene. (*b*) MA (log ratio over mean) plot of the same RNA-seq experiment. The *y*-axis shows the moderated log-fold change and the *x*-axis shows the mean of normalized counts. Red points denote differential expression detection according to a nominal false discovery rate threshold of 0.05.

applicability. The NB GLM model can be formulated as

$$Y_{fi} \sim \text{NB}(\mu_{fi},\ \varphi_f),$$

$$\log \mu_{fi} = \eta_{fi},$$

$$\eta_{fi} = X_i \beta_f + \log s_i,$$

where η_{fi} is the linear predictor, X_i denotes the design matrix, β_f represents the regression parameters, and $\log s_i$ are scaling (normalization) offsets. Regardless of the model, the parameters θ_{fi} or, equivalently, (a linear contrast of) β_f would represent the parameter(s) of interest for inference.

Reliable estimation of the dispersion parameter φ_f is nontrivial due to limited sample sizes. Traditional ML estimators for the dispersion are negatively biased (137) since they do not account for the fact that the mean is also estimated from the data. Early implementations estimated a single common dispersion parameter for all features (137), with the rationale to obtain a stable estimate by borrowing strength over all genes. However, the common dispersion assumption is unrealistic and relaxed estimation schemes were proposed, such as moderation toward a common dispersion (133) or estimation in strata of similar expression strength (120). For example, DESeq adopts a method of moments (MM) estimator and assumes the dispersion to be a smooth function of the mean. To avoid too liberal inference, one then sets the dispersion as the maximum between the smooth fit and the gene-wise MM estimate; however, while robust to outliers, this method tends to overestimate the variance and is therefore conservative (138, 139). Later approaches resorted to an approximate conditional inference scheme, the Cox–Reid adjusted profile likelihood (APL) (140), to correct for the bias in the ML estimator (134). Again, stable estimation is provided by leveraging information across genes (**Figure 8**). In particular, edgeR uses a maximized weighted APL to trade off between gene-specific and shared dispersion estimators upon estimating the dispersion-mean trend across all genes (similar to DESeq). The weighted likelihood,

$$\text{APL}_f(\varphi_f) + G_0 \text{APL}_{sf}(\varphi_f),$$

consists of the APL for a specific feature f (first component) and a shared likelihood (second component), which can be interpreted as a prior from a Bayesian perspective, thus representing an approximate empirical Bayes solution (133). The weight given to the prior likelihood, G_0, can also be estimated from the data (141). Analogously, DSS (Dispersion Shrinkage for Sequencing) and DESeq2 model the log φ_f as a Gaussian random variable, and Bayes' formula is applied to generate a posterior mode for each gene (129, 142). Hyperparameters for the (Gaussian) prior are inferred from the data using either the MM or the Cox–Reid estimator across all genes. Once dispersion estimates are available, the parameters of the mean model, β_f, can be estimated using standard algorithms for GLMs.

Statistical Inference

After fitting a GLM to each feature, the statistical inference typically involves testing the null hypothesis H_0 that there is no DE between conditions, i.e., that the LFC is zero, against the alternative H_1 that the LFC differs from zero. In the GLM framework, the null hypothesis can be represented as either a single regression parameter or a linear combination of parameters (contrasts), which is defined by a vector or matrix L such that H_0 is the hypothesis that $L\beta_f = 0$. Indeed,

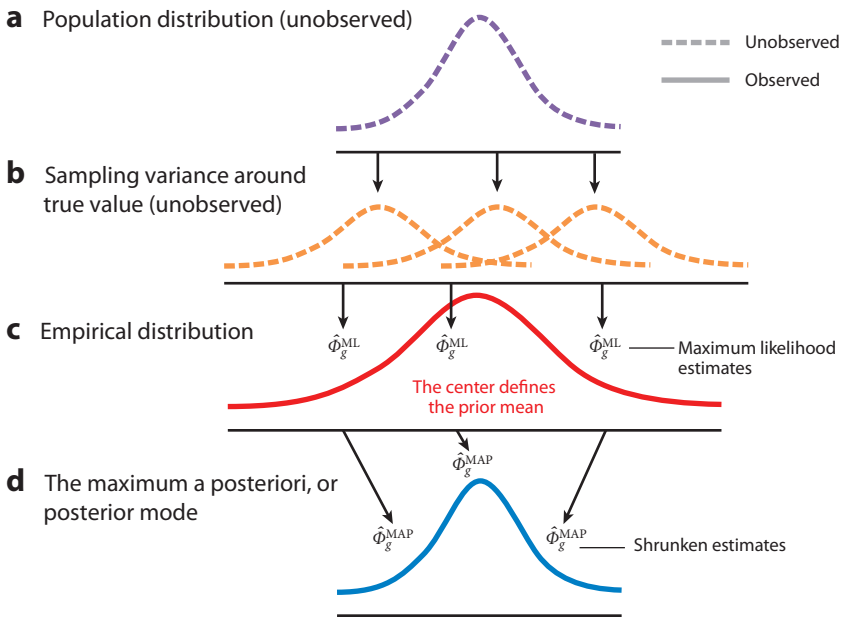

Figure 8

Steps in an empirical Bayes model. In an RNA sequencing experiment, one assesses the observed differences in gene expression across groups of samples with respect to within-group variance. (*a*) The unobserved population distribution for the true within-group variance of each gene. (*b*) Variances are estimated from limited sample size experiments, and so there is sampling variance in our estimate of the variance. A maximum likelihood estimate (MLE) or a bias-corrected estimator for expression variance can be used. (*c*) Thousands of genes are typically observed and estimates are made for each, providing an empirical distribution of MLEs across all genes. This empirical distribution of MLEs can be used to determine a prior distribution for empirical Bayes analysis; the posterior distribution for the variance of each gene is calculated using Bayes' formula. (*d*) Distribution of the maximum a posteriori (MAP), or posterior mode, estimates of variance over all genes. The posterior modes represent shrunken estimates, where the amount of shrinkage is determined by the shape of the likelihood and the width of the prior distribution.

a regression parameter in an NB GLM with a canonical link function can be interpreted as an LFC between groups and thus provides a measure of effect size.

There are multiple hypothesis tests available for GLMs with known (asymptotic) distribution under the null hypothesis. Likelihood ratio tests (LRTs) compare the likelihood of a full model, upon estimating all parameters without constraints, with the likelihood of a reduced model, where one or some of the parameters are constrained according to H_0. LRT statistics are asymptotically χ^2-distributed under H_0, and this type of test is implemented in both edgeR and DESeq2. By default, however, DESeq2 adopts a Wald test. Wald tests are attractive from a computational point of view since they only require fitting the full model and calculating the variance–covariance matrix of the regression coefficients. The Wald test statistic for a single model parameter or a single contrast, $W = \widehat{LFC}\ /\mathrm{se}(LFC)$, asymptotically follows a standard normal distribution under H_0, where se(*LFC*) is the standard error of *LFC* and $\widehat{LFC}$ is the ML estimate of *LFC*. From ML theory, it is known that LRTs have better properties (e.g., invariance to transformation) than Wald tests in GLMs (143); however, RNA-seq tools moderate dispersion estimates and do not re-estimate them under H_0, so it is unclear whether these benefits carry over to RNA-seq data analysis in practice.

Multiple Testing

The *p*-values obtained from the statistical inference must be corrected for multiple testing to avoid excess false positives. While it is possible to control the probability of returning at least one false positive in the list of detections by adopting familywise error rate corrections, this stringent form of correction is overly conservative. Indeed, when screening many thousands of features, one is typically willing to tolerate a certain proportion of false positives to obtain a larger number of true positives. The false discovery rate (FDR), which gained significant popularity, controls the expected fraction of false positives in the detected set of features, i.e., $FDR = E[V/\max(R, 1)]$, where V is the number of false positive rejections and R is the total number of detections. The FDR was introduced by Benjamini & Hochberg (144) and has become common practice in high-dimensional data analyses because of its simplicity and solid theoretical justification. Indeed, it can be shown that the FDR is justified under a range of dependency structures between genes (145) and can be approached from both frequentist and Bayesian perspectives.

Variations to the General Workflow

There is a large and growing number of alternatives to the basic framework mentioned above: different inferences based on the same models, alternative models, more robust approaches, different testing regimes, variations on multiple testing corrections, and so on. In this section, we summarize some of these developments.

Alternative models (inference frameworks). NB count models, which underpin many DE tools, assume a quadratic mean-variance relationship. Inference, however, may benefit from a more flexible variance structure, and for this, other models have been proposed. One strategy uses quasi-likelihood (QL), which requires only that mean and variance are specified to be able to make inference on the mean model parameters (146). The QL method adopts the same mean model structure as the NB but introduces an additional overdispersion parameter such that $\mathrm{Var}(Y_{fi}) = \psi_f(\mu_{fi} + \varphi_f \mu_{fi}^2)$, where ψ_f is estimated using a moderated MM estimator. QL naturally allows (asymptotic) hypothesis tests based on *t*- and *F*-statistics, thus accommodating the uncertainty in the estimation of the additional QL dispersion parameter. Another variation is the use of a more flexible distribution, such as the NB power distribution, which adds an additional parameter (147) to the NB. Within the NB framework itself, Bayesian methods have also been developed. A fully Bayesian approach has the benefit that various aspects of the posterior can be reported (e.g., credible intervals), and the degree of parameter shrinkage naturally depends on the amount of information available for that gene (a trade-off between expression magnitude, dispersion, and residual degrees of freedom). One of the early methods was ShrinkBayes, a fully Bayesian approach that included multiple mixture priors (e.g., Gaussian) (148, 149) and where fitting was accomplished using integrated nested Laplace approximations (150), which avoids the Markov chain Monte Carlo sampling. Another alternative is to remain within computationally and inferentially efficient Gaussian linear models, after suitably transforming the (normalized) count data. For example, limma-voom models log-transform normalized counts using a linear model while adjusting for heteroskedasticity via weighted regression, where the observation weights are computed from the observed mean-variance relationship (151). In this case, moderated *t*- and *F*-statistics are used for inference. Finally, nonparametric methods have been developed, which are more robust to outliers and do not require distributional assumptions. For example, SAMSeq (152) adopts the Wilcoxon test to assess DE between groups and uses resampling procedures to adjust for differences in sequencing depth.

Robust log-fold change estimation. The standard NB workflow typically makes use of APL NB likelihood for parameter estimation, combined with empirical Bayes procedures to borrow strength across features when estimating the dispersion parameter. There are two related challenges: (*a*) Ratios of smaller counts result in more variable LFCs (**Figure 7*b***), and (*b*) the estimation of LFC can be sensitive to outliers. This makes it difficult to rank genes according to LFC since lowly expressed or outlier-affected genes are likely to dominate the top list. To derive more robust LFC estimates, researchers have adopted several approaches. First, prior counts have been used in the numerator and denominator of the LFC; effective shrinkage is accomplished by augmenting each count with a carefully chosen value, although the optimal value may vary across data sets. Second, edgeR-robust (139), for instance, adopts an M-estimation approach by iteratively downweighting outlying observations within the GLM fitting procedure, dampening the effect of outliers on both mean and variance estimates. Alternatively, outliers can be identified and removed and/or imputed by taking advantage of the remaining data for a feature (129). Lastly, priors can be imposed on the LFC parameters. For example, DESeq2 includes a zero-centered Gaussian prior in the NB GLM and provides the posterior mode of LFC as output (129). The width of the prior is set conservatively, using a weighted upper quantile of the observed LFCs. New alternative shrinkage estimators in DESeq2 incorporate priors with heavier tails that introduce less bias, using either a mixture of normal distributions (153) or a Cauchy distribution (154).

Accounting for unobserved effects. As mentioned above, (G)LMs can adjust for known confounders. However, genomic data can also be affected by unknown, and hence unobserved, confounders. This problem is widespread in publicly available data, which typically do not contain sufficient metadata on potential batch effects caused by lab, protocol, date, etc. Batch correction methods can leverage the parallel structure of high-throughput transcriptomic data to identify unknown and unobserved systematic effects. SVA (surrogate variable analysis) (155, 156) and RUV (remove unwanted variation) (125) methods, for instance, estimate surrogate variables through singular value decomposition on control features or on a matrix of model residuals so that the phenotypic effect of interest is not captured by the surrogates. RUV also has the option to exploit information in replicate samples. The estimated surrogate variables can subsequently be included as predictors in the statistical model to adjust for the batch effects.

Statistical inference by testing against a threshold. The standard approach for detecting DE in RNA-seq involves a simple null hypothesis H_0 that the LFC is zero. However, statistical significance does not guarantee that the fold changes are large enough to be biologically relevant. Analysts often produce candidate gene lists by applying a threshold on the magnitude of the LFC, but the statistical properties of this approach are unclear. The FDR is a set property and has no interpretation when the set, post-FDR calculation is altered (157). To address these practical and theoretical concerns, researchers have adopted several tests relative to an LFC threshold, a procedure initially proposed for microarray data (158). This results in a composite null hypothesis H_0, such as $|LFC| < a$. Implementations differ: DESeq2 replaces the composite null with a simple null hypothesis at the boundary of the parameter space (129), whereas edgeR uses a modified likelihood ratio test or a quasi-likelihood F-test against a threshold (159).

Small-sample inference. The null distributions for Wald or LRT statistics for count models are only valid asymptotically, and the number of replicates is often too low for these approximations to be fully effective, which may lead to an inflated FDR. Initial implementations provided exact tests (137), but these can only be applied in simple designs. Another strategy is small-sample

asymptotics, which makes use of higher-order approximations that are still compatible with the GLM framework (160).

Multiple testing. While the FDR achieves a more reasonable sensitivity–specificity trade-off than familywise error rate correction approaches, other developments beyond simple filtering aim to further reduce the multiple testing burden. Storey's *q*-value, for instance, estimates the proportion of true null hypotheses from the data to increase power (161), while others adopt a data-driven weighting of the *p*-values in the FDR correction (162). Although the FDR is deeply rooted in statistical theory, it is not guaranteed that methods will control error rates at the nominal level in real applications. NB methods, for instance, rely on the asymptotic theory, which might not hold for applications with low sample sizes. A study has suggested that coregulation of genes induces intergene correlations, which can alter the null distribution of the statistical test (163); local FDR approaches were introduced that empirically estimate the null distribution (164). Other developments address issues in testing many hypotheses for every gene (e.g., multifactorial designs). The conventional approach is to control the FDR on each hypothesis, but this does not allow for straightforward prioritization since genes typically have a different ranking for each hypothesis. Stagewise testing procedures can be interpreted as generalizations of analyses of variance with post hoc tests for high-throughput contexts (165, 166), thus allowing a natural ordering of the genes according to an omnibus test (all effects of interest) while providing FDR control at the gene level.

VARIANTS OF DIFFERENTIAL EXPRESSION

The previous section introduced count-based DE in general terms: Each row of a count matrix is submitted to a statistical model (often by first estimating moderated variance parameters over the whole data set) and hypothesis tests of interest are conducted, with an adjustment for multiple testing. In this section, we discuss additional approaches to interrogate RNA-seq data in terms of DE.

Although DE is of obvious interest, this can manifest or be defined in multiple ways (see **Figure 9**). One may want to cast inferences to the gene level, but measurements are made at the fragment level. We use the term "differential gene expression" (DGE) to refer to hypothesis testing related to the total outcome of an annotated gene, by comparing either accumulated TPM estimates or raw counts while including an adjustment for average transcript length via offsets (74). If the expression of transcripts is the feature of interest (independent of other transcripts), differential transcript expression (DTE) analyses can be conducted. Alternatively, one could be interested in whether at least one transcript from a gene is DE. This requires statistical testing at the transcript level and then aggregation to the gene level. Yet another strategy is to consider whether the relative abundance (i.e., proportions) of transcripts for a specific genomic locus changes between conditions, which is commonly termed differential transcript usage (DTU) or, more generally, differential splicing (DS). A surrogate for DTU, differential exon usage, is conducted on exon-level quantifications; in this case, the goal is to identify exons that deviate from proportional expression to separate differential usage from DE. Yet another alternative is to quantify and test differences at the event level, where reads supporting (or not supporting) an event (e.g., inclusion of a cassette exon) are summarized and compared (167).

There is certainly a question of which analysis path to choose. Conceptually, pure DTE points to all kinds of DE, and while casting a wide net of potentially interesting genes might seem appealing, there are some considerations to be made. For example, if a given transcript is DE, often the question becomes, What happens to the expression of the other transcripts for this gene? Are all transcripts changing in the same direction? If so, it may be better in terms of sensitivity to detect

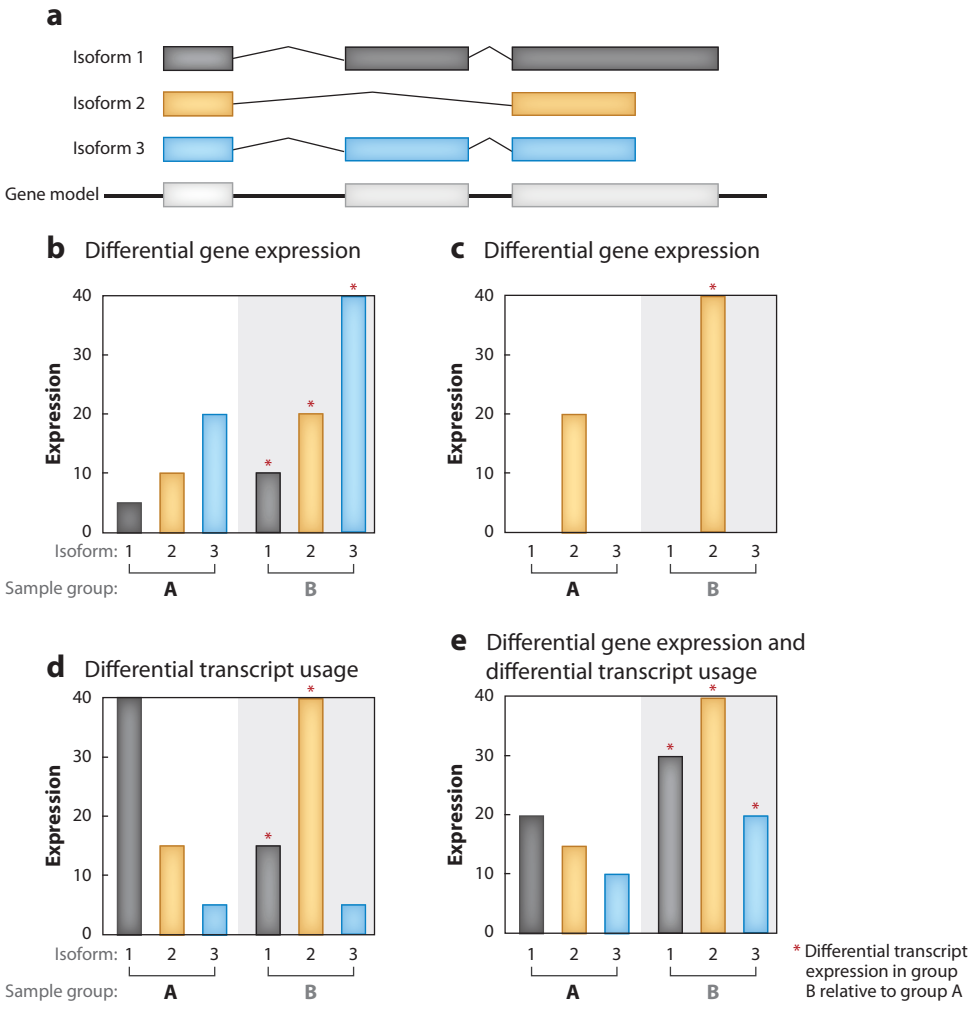

Figure 9

Schematic illustration of some examples of differential gene expression, differential transcript expression, and differential transcript usage for a gene with three isoforms (1, 2, and 3) in a two-group comparison (A versus B).

an aggregated output (i.e., DGE). Or, transcript-level expression can be represented as a genewise multivariate outcome and isoform switches can be considered collectively, i.e., by assessing DTU, which is not affected in either direction by DGE. DTU implies DTE while the opposite is not necessarily true. We generally favor two clear but orthogonal analyses (DGE and DTU) over a catchall DTE analysis (74), but this will ultimately be application dependent, and scientists should clearly define their question of interest in advance.

Differential Transcript Expression

Modeling transcript-level count data for DE presents some additional challenges due to increased variability and resolution compared to gene-level analyses. For example, transcript-level

abundance estimates are considerably more variable than gene-level counts due to ambiguous assignment of fragments to isoforms (74). Thus, transcript quantifications inferred by popular tools such as RSEM, Salmon, or kallisto carry more uncertainty, which should be accounted for in the downstream DE analysis.

Transcript quantifications still have many of the properties of count data (e.g., mean-variance relationships) and thus could be used as inputs to the frameworks mentioned above. However, quantifications are estimates that may obscure inference when plugging them into count-based RNA-seq tools. Cufflinks was one of the first methods to use estimated abundances and their corresponding standard errors to perform DTE (and DGE) analyses; the method quantifies transcript abundances via a likelihood model and an EM algorithm, and tests of DE are performed by applying the delta method on the abundance parameters (75, 86). Bayesian approaches for identifying DTE based on estimated counts, e.g., ranking via Bayes factors, include EBSeq (132), which uses an empirical Bayesian hierarchical model, and MMSeq (79, 168), which fits a linear mixed model to data via Markov chain Monte Carlo techniques. Similarly, BitSeq (and later cjBitSeq) introduced a generative model that couples both quantification and DE using fully Bayesian inference (95, 169). Most recently, with the advent of ultrafast transcript quantification algorithms, sleuth uses bootstrap samples of each sample of reads to determine the so-called inferential variance and integrates this into the DE calculation through a variance components model on the log-transformed scale (170).

Differential Transcript Usage (Differential Splicing)

One of the first statistical models for DTU, cuffdiff, calculates the square root of the Jensen–Shannon divergence on estimated transcript proportions and uses the delta method to estimate the variance of this metric under the null hypothesis of no change in proportions (86). Another conceptually distinct approach formulates a Poisson mixed effects model on exon- and junction-level quantifications and searches for exon–condition interactions that represent differential usage (171). Such departure-from-parallelism modeling was introduced in earlier analyses of probe-level microarray data for DTU (172); on RNA-seq data, this approach was further formalized with DEXSeq (173), which uses an NB model on exon-level counts. Exon by exon, DEXSeq tests whether an improvement in fit is achieved by adding a single exon–condition interaction, which represents the differential usage of that exon across conditions. A comparison study showed that DEXSeq has a good performance in well-annotated transcriptomes and that filtering of lowly expressed transcripts improves error control (174); in addition, DEXSeq also works well with transcript quantifications as input (175).

In a similar vein, DRIMSeq (176) and LeafCutter (177) employ the Dirichlet-multinomial (DM) distribution to perform the same inference task but treat the output of a gene's expression as a multivariate response; Bayesian inference for the DM model has also been considered in BayesDRIMSeq (178). Several tools neglect the uncertainty in estimated transcript-level counts, and this is perhaps the reason for inflated FDRs (175). To address this, RATs (relative abundance of transcripts) uses bootstrapped (transcript-level) quantifications to infer DTU via a *G*-test of independence, based on the multinomial distribution, on the two groups' isoform counts (179). Instead of considering estimated counts and their uncertainties, Bayesian methods such as cjBitSeq (169) focus on the group of transcripts that each read is compatible with (i.e., equivalence classes). In this way, quantification is not required because the DS tools treat the transcript allocation of reads as an unknown latent variable.

Event-Level Analyses Based on Percent Spliced-In

Some methods perform differential analyses based on percent spliced-in values (PSIs). PSIs can be computed either for specific events (retained intron, cassette exon, etc.) or at the transcript level and indicate the fraction of RNA-seq reads supporting the event, obtained as the ratio between the number of reads including the event and the total number of reads including and excluding the event. The difference of the PSIs between conditions is then used to assess DS, performed separately for each event (or transcript). Some of the main DS tools based on PSIs include rMATS (180), which uses an LRT, and SUPPA2 (181), whose test is based on comparing the observed difference in PSIs across conditions to the empirical cumulative density function of the within-replicates differences of PSIs of SJs from similarly expressed transcripts.

Event-level analysis, similar to DEXSeq's exon-level approach, separately focuses on each splicing event, and results could be aggregated to the gene level by considering the most significant event- or transcript-level test, appropriately adjusted for multiple testing (173, 181, 182).

Multistage Testing

As mentioned, DS analyses can be approached at the gene-, transcript- or event-specific level. While gene-level tests often have higher sensitivity, testing each individual transcript provides increased resolution. However, neither gene- nor transcript-level tests guarantee FDR control on the full set. Stagewise testing procedures (165, 166) instead first screen for significant genes and only consider significant transcripts from those genes. This procedure gives gene-level FDR control and allows researchers to leverage the power from gene-level tests while interpreting results at the transcript level (175). The same procedure can be applied by replacing transcript-level tests with exon- or event-specific tests.

SINGLE-CELL TRANSCRIPTOME SEQUENCING

One of the emerging data types in transcriptomics is scRNA-seq, whereby the expressed content of individual cells is prepared and sequenced. In this case, experimental design is again of critical importance to avoid confounding the data (183). Experimentally, capture and reverse transcription efficiency become important, given that the number of mRNA transcripts per mammalian cell is estimated to be between 50,000 and 300,000 (184).

Two main experimental approaches are used: plate-based, where cells are sorted into individual wells for lysis and library preparation, and droplet-based, where each cell is absorbed (together with reagents) and processed within an oil droplet (185). Several variations of these protocols are now available, increasing the number of cells assayed, but ultimately only a small fraction of the expressed RNAs (cDNAs), often the most highly expressed transcripts, are captured. The features that distinguish scRNA-seq from bulk RNA-seq data include (*a*) generally low depth of sampling for each cell (due to cost, but also due to lower diversity of cDNA fragments); (*b*) so-called dropout, where a cell expresses a transcript but it is unobserved; and (*c*) higher levels of biological (since no averaging) and technical (e.g., more amplification) variation.

Nonetheless, researchers can distinguish cell identities, where identity represents the combined effects of cell type (permanent features) and cell state (transient features) (186). The Human Cell Atlas, among other projects, opens the door for exploring spatial context (187), developmental patterns (188), immune responses (189), response to therapy (190), and an increasing range of basic science and clinical investigations (191–193).

Although many computational aspects of scRNA-seq data are beyond the scope of this review (e.g., dimensionality reduction techniques, ordering cells into lineages), one connected application

area that has already received considerable attention is DE analysis. In the simplest setting, cells are first partitioned into different classes (e.g., assumed to correspond to different cell types) via clustering, with the subsequent aim of finding markers for each cluster (e.g., to annotate cell types). To perform this task, a statistical model uses cells as experimental units, as opposed to samples in bulk analyses; thus, it is worth considering the population to which the conclusions extrapolate.

To date, several methods have been developed to decipher DE between cell types, many of which have been comparatively assessed in recent benchmarks (194, 195). Many of these single cell–specific methods are extensions or variations of existing bulk approaches. For example, SCDE formulated the RPM (reads per million) data for a given gene across cells as a mixture of Poisson and NB components; using a Bayesian approach, probabilities of observing a given fold change are converted into empirical *p*-values (196). MAST uses a hurdle model on $\log(TPM + 1)$ data, where a logistic regression is used to model whether a gene is expressed, and a Gaussian linear model is used conditional on expression. Inferences for the two sets of regression parameters are done in a Bayesian framework that also provides regularization (197). Again, extending existing approaches, Van den Berge et al. (198) proposed a zero-inflated NB (ZINB) model; model fitting is done within the ZINB-WaVE (wanted variation extraction) framework (199), estimating cell- and gene-specific posterior probabilities for counts to belong to the NB count component of the ZINB mixture model. These probabilities are used as observation weights in the downstream estimation of regression parameters in the classical NB framework.

Nonetheless, many DE methods focus on assessing changes in the mean parameters. However, since cell subsets are being compared, we may not have simple shifts in the mean. Instead, it may be informative to detect and understand changes in expression variability across conditions (200). Alternatively, full distributions (instead of means or variances) can be compared, as was proposed in a Bayesian framework in scDD (201), highlighting not only DE but also differential proportions (change in the relative usage of low and high expression), differential modality (change in the number and place of the mode of expression), or some combination thereof.

In many applications of single-cell DE analysis, the sample sizes (numbers of cells) are generally larger than those commonly used within the optimized frameworks built for bulk RNA-seq data, and thus it seems that the distributional assumptions play less of a role for effective inference. Indeed, a recent comparison highlighted decent performance of *t*-tests and Wilcoxon rank sum nonparametric tests in comparing single-cell subsets (194).

Beyond comparing cell types, which may involve multiple experimental units (e.g., patients), it will be of increasing interest to compare expression levels of genes across biological replicates and conditions. For example, it may be of interest to understand cell-type-specific immune responses following a stimulus. A recent study compared multiple patients across stimulated and unstimulated conditions by first computationally separating immune cell types (189); to do this, the researchers aggregated cells from a given cell type into a pseudobulk RNA-seq data set and performed DE using standard tools.

LONG-READ TRANSCRIPTOME SEQUENCING

The short read length of Illumina-based RNA-seq complicates the unambiguous placement of reads to the genome, especially in repeat regions (202), and adds difficulties to the assembly, identification, and quantification of expressed isoforms (203–205). In contrast, so-called third-generation, or long-read, sequencing technologies, led by PacBio (206) and ONT, can generate much longer reads. By sequencing single molecules, they can also forego PCR amplification, hence reducing coverage biases (207, 208). Currently, long-read technologies incur a higher average cost

and a higher error rate than short-read sequencing (203, 209). However, this is a rapidly developing field, and improvements in error rates and throughput are to be expected.

The strategies employed by PacBio and ONT to generate long sequencing reads of single molecules differ in many ways. PacBio, with its RSII and Sequel instruments, uses SMRT (single molecule real time) sequencing (210), where the reactions take place inside so-called zero-mode waveguides (ZMWs) (211). At the bottom of each ZMW, there is a single DNA polymerase molecule. As the polymerase processes a DNA fragment, the incorporation of each nucleotide leads to a fluorescent signal, which is detected by the ZMW and converted to a base call. A specific characteristic of the PacBio library preparation system is the creation of SMRTbell templates (212), which are obtained by ligating SMRTbell hairpin adapters. The result is a circular construct, where the two strands of the template are separated by adapters with known sequences. As the construct is processed by the polymerase in the ZMW, the original template can be passed multiple times. Since the sequencing errors are largely random (213), the base-level error rate can be considerably reduced by forming a consensus over these passes.

ONT, in contrast, uses a different sequencing strategy based on protein nanopores placed in a polymer membrane (214) for its MinION and PromethION sequencers. A current is passed through the nanopores, and as the template molecule is passed through the pore by a motor protein, each combination of bases induces a change in the current. Analyzing the exact nature of this change allows for the identification of the template sequence. By adding a hairpin sequence to the end of the double-stranded cDNA fragment before denaturing it into a single-stranded molecule and passing it through the nanopore, both the template sequence and its complement are included in a single read and can be combined at the base calling step to generate a higher-quality, so-called 2D, read (215). In contrast to PacBio, ONT also offers direct sequencing of RNA (216). Advantages of this include that the reverse transcription step is avoided, which may reduce biases, and that RNA modifications can be directly observed, since they also change the current passing through the nanopore in characteristic ways (217). However, at present, the required amount of starting material is considerably higher than for cDNA protocols.

Applications to cDNA (RNA) include both transcriptome-wide sequencing and characterization of specific genes via targeted sequencing (14, 203, 218–222), as well as performance evaluations based on synthetic transcript catalogs [ERCC (External RNA Controls Consortium) with 92 sequences or SIRV (spike-in RNA variant) with 68]. With LRTS, every read potentially represents a full-length transcript. If this were indeed true, de novo (reference-free) identification of the full set of observed isoforms would be straightforward and would only require grouping reads expected to differ only by sequencing errors (which, depending on the error rate and isoform similarity, may not be trivial). However, this is not currently the case, due to fragmentation and degradation of template molecules during library preparation and early termination of the sequencing, which leads to ambiguities in transcript identification (223). This means that it is not easy to determine whether truncated variants are present.

Transcript identification from LRTS can be either reference-based or reference-free. The latter typically involves clustering reads based on similarity, followed by polishing the consensus sequence within each cluster (14, 224–227). Since LRTS is still a young field, methods and tools for reference-based alignment are still emerging but so far include a mix of established tools, such as GMAP (58), and new innovations, such as minimap2 (228). A recent study comparing PacBio, ONT, and Illumina data (203) showed that the long-read technologies were indeed much better at correctly identifying expressed SIRV transcripts than de novo assemblies of short reads.

The rapid technological developments in LRTS also mean that the read generation process (e.g., biases affecting the observation of a given read) is still largely unknown. In addition, read lengths are extremely variable, error rates are relatively high, and throughput is still relatively low.

For the PacBio RSII instrument, the selection of transcript molecules is biased toward short sequences (223). Thus, samples are typically size-fractionated before sequencing, which distorts the abundance estimates. Taken together, these and other aspects make accurate transcript quantification from LRTS more difficult, and new models and tools will be needed. Encouragingly, a recent study showed that by combining LRTS and Illumina data, more accurate quantifications for the artificial SIRV transcripts could be achieved (203).

Since abundance estimation for long reads returns values in the form of read (or transcript) counts, it is plausible that the DE machinery developed for short-read data can be applied in a similar way. The quality of the DE calls will be directly dependent on the accuracy of the abundance estimates. However, the current low depth of sequencing compared to short-read data sets will ultimately lead to low power to detect DE features.

SUMMARY

> I'm a scientist and I know what constitutes proof. But the reason I call myself by my childhood name is to remind myself that a scientist must also be absolutely like a child. If [they] see a thing, [they] must say that [they] see it, whether it was what [they] thought [they] were going to see or not. See first, think later, then test. But always see first. Otherwise you will only see what you were expecting. Most scientists forget that.
>
> –adapted from *The Ultimate Hitchhiker's Guide to the Galaxy* by Douglas Adams

In this review, we gave an overview of the data science of gene expression analysis, with a focus on methods to estimate transcript-level abundance and statistical tools for assessing DE. Notably, RNA-seq data are often an intermediate discovery step where the detected molecular changes represent candidates for further follow-up. Nonetheless, the analysis of RNA-seq data for gene expression is already very mature, due to a deep understanding of the biases present, to the implementation of efficient data structures and algorithms for processing the data into (estimated) count tables, and to a refined understanding of how well tools perform via the many benchmarks available.

Ultimately, the success of RNA-seq lies in its wide range of applications, and it is likely that Illumina-based short-fragment RNA-seq will continue to be the workhorse for the field for many years. With the increasing fidelity of single-cell protocols, many tools are emerging to deal with the additional complexities of single-cell measurements, and these will be further refined in the coming years. Similarly, with the decreasing costs and lower error rates of long-read technologies, it may be possible to characterize alternative transcription quantitatively with full-length transcript sequencing, thus considerably reducing read-to-transcript ambiguity; however, much still needs to be learned about the biases present.

DISCLOSURE STATEMENT

R.P. is a cofounder of Ocean Genomics.

LITERATURE CITED

1. Lister R, O'Malley RC, Tonti-Filippini J, Gregory BD, Berry CC, et al. 2008. Highly integrated single-base resolution maps of the epigenome in *Arabidopsis*. *Cell* 133:523–36
2. Nagalakshmi U, Wang Z, Waern K, Shou C, Raha D, et al. 2008. The transcriptional landscape of the yeast genome defined by RNA sequencing. *Science* 320:1344–49
3. Mortazavi A, Williams BA, McCue K, Schaeffer L, Wold B. 2008. Mapping and quantifying mammalian transcriptomes by RNA-Seq. *Nat. Methods* 5:621–28

4. Cloonan N, Forrest ARR, Kolle G, Gardiner BBA, Faulkner GJ, et al. 2008. Stem cell transcriptome profiling via massive-scale mRNA sequencing. *Nat. Methods* 5:613–19
5. Wilhelm BT, Marguerat S, Watt S, Schubert F, Wood V, et al. 2008. Dynamic repertoire of a eukaryotic transcriptome surveyed at single-nucleotide resolution. *Nature* 453:1239–43
6. Palazzo AF, Lee ES. 2015. Non-coding RNA: What is functional and what is junk? *Front. Genet.* 6:2
7. Zhao W, He X, Hoadley KA, Parker JS, Hayes DN, Perou CM. 2014. Comparison of RNA-Seq by poly (A) capture, ribosomal RNA depletion, and DNA microarray for expression profiling. *BMC Genom.* 15:419
8. Bentley DR, Balasubramanian S, Swerdlow HP, Smith GP, Milton J, et al. 2008. Accurate whole human genome sequencing using reversible terminator chemistry. *Nature* 456:53–59
9. Ju J, Kim DH, Bi L, Meng Q, Bai X, et al. 2006. Four-color DNA sequencing by synthesis using cleavable fluorescent nucleotide reversible terminators. *PNAS* 103:19635–40
10. Levin JZ, Yassour M, Adiconis X, Nusbaum C, Thompson DA, et al. 2010. Comprehensive comparative analysis of strand-specific RNA sequencing methods. *Nat. Methods* 7:709–15
11. Zhao S, Zhang Y, Gordon W, Quan J, Xi H, et al. 2015. Comparison of stranded and non-stranded RNA-seq transcriptome profiling and investigation of gene overlap. *BMC Genom.* 16:675
12. Parkhomchuk D, Borodina T, Amstislavskiy V, Banaru M, Hallen L, et al. 2009. Transcriptome analysis by strand-specific sequencing of complementary DNA. *Nucleic Acids Res.* 37:e123
13. Mamanova L, Turner DJ. 2011. Low-bias, strand-specific transcriptome Illumina sequencing by on-flowcell reverse transcription (FRT-seq). *Nat. Protoc.* 6:1736–47
14. Wang B, Tseng E, Regulski M, Clark TA, Hon T, et al. 2016. Unveiling the complexity of the maize transcriptome by single-molecule long-read sequencing. *Nat. Commun.* 7:11708
15. Collado-Torres L, Nellore A, Kammers K, Ellis SE, Taub MA, et al. 2017. Reproducible RNA-seq analysis using *recount2*. *Nat. Biotechnol.* 35:319–21
16. Lazic SE. 2017. *Experimental Design for Laboratory Biologists: Maximising Information and Improving Reproducibility*. Cambridge, UK: Cambridge Univ. Press. 1st ed.
17. Hart SN, Therneau TM, Zhang Y, Poland GA, Kocher J-P. 2013. Calculating sample size estimates for RNA sequencing data. *J. Comput. Biol.* 20:970–78
18. Guo Y, Zhao S, Li CI, Sheng Q, Shyr Y. 2014. RNAseqPS: a web tool for estimating sample size and power for RNAseq experiment. *Cancer Inform.* 13:1–5
19. Zhao S, Li C-I, Guo Y, Sheng Q, Shyr Y. 2018. RnaSeqSampleSize: real data based sample size estimation for RNA sequencing. *BMC Bioinform.* 19:191
20. Busby MA, Stewart C, Miller CA, Grzeda KR, Marth GT. 2013. Scotty: a web tool for designing RNA-Seq experiments to measure differential gene expression. *Bioinformatics* 29:656–57
21. Poplawski A, Binder H. 2018. Feasibility of sample size calculation for RNA-seq studies. *Brief. Bioinform.* 19:713–20
22. Oshlack A, Wakefield MJ. 2009. Transcript length bias in RNA-seq data confounds systems biology. *Biol. Direct.* 4:14
23. Liu Y, Zhou J, White KP. 2014. RNA-seq differential expression studies: More sequence or more replication? *Bioinformatics* 30:301–4
24. Schurch NJ, Schofield P, Gierliński M, Cole C, Sherstnev A, et al. 2016. How many biological replicates are needed in an RNA-seq experiment and which differential expression tool should you use? *RNA* 22:839–51
25. Mercer TR, Clark MB, Crawford J, Brunck ME, Gerhardt DJ, et al. 2014. Targeted sequencing for gene discovery and quantification using RNA CaptureSeq. *Nat. Protoc.* 9:989–1009
26. Cabanski CR, Magrini V, Griffith M, Griffith OL, McGrath S, et al. 2014. cDNA hybrid capture improves transcriptome analysis on low-input and archived samples. *J. Mol. Diagn.* 16:440–51
27. Irimia M, Weatheritt RJ, Ellis JD, Parikshak NN, Gonatopoulos-Pournatzis T, et al. 2014. A highly conserved program of neuronal microexons is misregulated in autistic brains. *Cell* 159:1511–23
28. Eom T, Zhang C, Wang H, Lay K, Fak J, et al. 2013. NOVA-dependent regulation of cryptic NMD exons controls synaptic protein levels after seizure. *eLife* 2:e00178

29. Fratta P, Sivakumar P, Humphrey J, Lo K, Ricketts T, et al. 2018. Mice with endogenous TDP-43 mutations exhibit gain of splicing function and characteristics of amyotrophic lateral sclerosis. *EMBO J.* 37:e98684
30. Salzman J, Gawad C, Wang PL, Lacayo N, Brown PO. 2012. Circular RNAs are the predominant transcript isoform from hundreds of human genes in diverse cell types. *PLOS ONE* 7:e30733
31. Kim T-K, Hemberg M, Gray JM. 2015. Enhancer RNAs: a class of long noncoding RNAs synthesized at enhancers. *Cold Spring Harb. Perspect. Biol.* 7:a018622
32. Parker BC, Zhang W. 2013. Fusion genes in solid tumors: an emerging target for cancer diagnosis and treatment. *Chin. J. Cancer* 32:594–603
33. Frye M, Jaffrey SR, Pan T, Rechavi G, Suzuki T. 2016. RNA modifications: What have we learned and where are we headed? *Nat. Rev. Genet.* 17:365–72
34. Perou CM, Sørlie T, Eisen MB, van de Rijn M, Jeffrey SS, et al. 2000. Molecular portraits of human breast tumours. *Nature* 406:747–52
35. Climente-González H, Porta-Pardo E, Godzik A, Eyras E. 2017. The functional impact of alternative splicing in cancer. *Cell Rep.* 20:2215–26
36. Cieślik M, Chinnaiyan AM. 2017. Cancer transcriptome profiling at the juncture of clinical translation. *Nat. Rev. Genet.* 19:93–109
37. Pedersen G, Kanigan T. 2016. Clinical RNA sequencing in oncology: Where are we? *Per Med.* 13:209–13
38. Ståhl PL, Salmén F, Vickovic S, Lundmark A, Navarro JF, et al. 2016. Visualization and analysis of gene expression in tissue sections by spatial transcriptomics. *Science* 353:78–82
39. Westermann AJ, Gorski SA, Vogel J. 2012. Dual RNA-seq of pathogen and host. *Nat. Rev. Microbiol.* 10:618–30
40. Piskol R, Ramaswami G, Li JB. 2013. Reliable identification of genomic variants from RNA-seq data. *Am. J. Hum. Genet.* 93:641–51
41. Park E, Williams B, Wold BJ, Mortazavi A. 2012. RNA editing in the human ENCODE RNA-seq data. *Genome Res.* 22:1626–33
42. Uszczynska-Ratajczak B, Lagarde J, Frankish A, Guigó R, Johnson R. 2018. Towards a complete map of the human long non-coding RNA transcriptome. *Nat. Rev. Genet.* 19:535–48
43. Bashiardes S, Zilberman-Schapira G, Elinav E. 2016. Use of metatranscriptomics in microbiome research. *Bioinform. Biol. Insights* 10:19–25
44. Racle J, de Jonge K, Baumgaertner P, Speiser DE, Gfeller D. 2017. Simultaneous enumeration of cancer and immune cell types from bulk tumor gene expression data. *eLife* 6:e26476
45. Martin JA, Wang Z. 2011. Next-generation transcriptome assembly. *Nat. Rev. Genet.* 12:671–82
46. Castel SE, Levy-Moonshine A, Mohammadi P, Banks E, Lappalainen T. 2015. Tools and best practices for data processing in allelic expression analysis. *Genome Biol.* 16:195
47. Sun W, Hu Y. 2013. eQTL mapping using RNA-seq data. *Stat. Biosci.* 5:198–219
48. Alamancos GP, Agirre E, Eyras E. 2014. Methods to study splicing from high-throughput RNA sequencing data. *Methods Mol. Biol.* 1126:357–97
49. van Dam S, Võsa U, van der Graaf A, Franke L, de Magalhães JP. 2018. Gene co-expression analysis for functional classification and gene-disease predictions. *Brief. Bioinform.* 19:575–92
50. Khatri P, Sirota M, Butte AJ. 2012. Ten years of pathway analysis: current approaches and outstanding challenges. *PLOS Comput. Biol.* 8:e1002375
51. de Leeuw CA, Neale BM, Heskes T, Posthuma D. 2016. The statistical properties of gene-set analysis. *Nat. Rev. Genet.* 17:353–64
52. Trapnell C, Pachter L, Salzberg SL. 2009. TopHat: discovering splice junctions with RNA-Seq. *Bioinformatics* 25:1105–11
53. Langmead B, Trapnell C, Pop M, Salzberg SL. 2009. Ultrafast and memory-efficient alignment of short DNA sequences to the human genome. *Genome Biol.* 10:R25
54. Engström PG, Steijger T, Sipos B, Grant GR, Kahles A, et al. 2013. Systematic evaluation of spliced alignment programs for RNA-seq data. *Nat. Methods* 10:1185–91
55. Dobin A, Davis CA, Schlesinger F, Drenkow J, Zaleski C, et al. 2013. STAR: ultrafast universal RNA-seq aligner. *Bioinformatics* 29:15–21

56. Kim D, Langmead B, Salzberg SL. 2015. HISAT: a fast spliced aligner with low memory requirements. *Nat. Methods* 12:357–60
57. Liao Y, Smyth GK, Shi W. 2013. The Subread aligner: fast, accurate and scalable read mapping by seed-and-vote. *Nucleic Acids Res.* 41:e108
58. Wu TD, Watanabe CK. 2005. GMAP: a genomic mapping and alignment program for mRNA and EST sequences. *Bioinformatics* 21:1859–75
59. Lin H-N, Hsu W-L. 2017. DART: a fast and accurate RNA-seq mapper with a partitioning strategy. *Bioinformatics* 34:190–97
60. Sedlazeck FJ, Rescheneder P, von Haeseler A. 2013. NextGenMap: fast and accurate read mapping in highly polymorphic genomes. *Bioinformatics* 29:2790–91
61. Medina I, Tárraga J, Martínez H, Barrachina S, Castillo MI, et al. 2016. Highly sensitive and ultrafast read mapping for RNA-seq analysis. *DNA Res.* 23:93–100
62. Baruzzo G, Hayer KE, Kim EJ, Di Camillo B, FitzGerald GA, Grant GR. 2017. Simulation-based comprehensive benchmarking of RNA-seq aligners. *Nat. Methods* 14:135–39
63. Bonfert T, Kirner E, Csaba G, Zimmer R, Friedel CC. 2015. ContextMap 2: fast and accurate context-based RNA-seq mapping. *BMC Bioinform.* 16:122
64. Wu TD, Nacu S. 2010. Fast and SNP-tolerant detection of complex variants and splicing in short reads. *Bioinformatics* 26:873–81
65. Aken BL, Ayling S, Barrell D, Clarke L, Curwen V, et al. 2016. The Ensembl gene annotation system. *Database* 2016:baw093
66. Liao Y, Smyth GK, Shi W. 2014. featureCounts: an efficient general purpose program for assigning sequence reads to genomic features. *Bioinformatics* 30:923–30
67. Anders S, Pyl PT, Huber W. 2015. HTSeq—a Python framework to work with high-throughput sequencing data. *Bioinformatics* 31:166–69
68. Roberts A, Trapnell C, Donaghey J, Rinn JL, Pachter L. 2011. Improving RNA-Seq expression estimates by correcting for fragment bias. *Genome Biol.* 12:R22
69. Hansen KD, Brenner SE, Dudoit S. 2010. Biases in Illumina transcriptome sequencing caused by random hexamer priming. *Nucleic Acids Res.* 38:e131
70. Liu X, Shi X, Chen C, Zhang L. 2015. Improving RNA-Seq expression estimation by modeling isoform- and exon-specific read sequencing rate. *BMC Bioinform.* 16:332
71. Love MI, Hogenesch JB, Irizarry RA. 2016. Modeling of RNA-seq fragment sequence bias reduces systematic errors in transcript abundance estimation. *Nat. Biotechnol.* 34:1287–91
72. Robert C, Watson M. 2015. Errors in RNA-Seq quantification affect genes of relevance to human disease. *Genome Biol.* 16:177
73. Li B, Dewey CN. 2011. RSEM: accurate transcript quantification from RNA-Seq data with or without a reference genome. *BMC Bioinform.* 12:323
74. Soneson C, Love MI, Robinson MD. 2015. Differential analyses for RNA-seq: transcript-level estimates improve gene-level inferences. *F1000Research* 4:1521
75. Trapnell C, Roberts A, Goff L, Pertea G, Kim D, et al. 2012. Differential gene and transcript expression analysis of RNA-seq experiments with TopHat and Cufflinks. *Nat. Protoc.* 7:562–78
76. Xing Y, Yu T, Wu YN, Roy M, Kim J, Lee C. 2006. An expectation-maximization algorithm for probabilistic reconstructions of full-length isoforms from splice graphs. *Nucleic Acids Res.* 34:3150–60
77. Jiang H, Wong WH. 2009. Statistical inferences for isoform expression in RNA-Seq. *Bioinformatics* 25:1026–32
78. Li B, Ruotti V, Stewart RM, Thomson JA, Dewey CN. 2010. RNA-Seq gene expression estimation with read mapping uncertainty. *Bioinformatics* 26:493–500
79. Turro E, Su S-Y, Gonçalves Â, Coin LJM, Richardson S, Lewin A. 2011. Haplotype and isoform specific expression estimation using multi-mapping RNA-seq reads. *Genome Biol.* 12:R13
80. Richard H, Schulz MH, Sultan M, Nürnberger A, Schrinner S, et al. 2010. Prediction of alternative isoforms from exon expression levels in RNA-Seq experiments. *Nucleic Acids Res.* 38:e112
81. Nicolae M, Mangul S, Măndoiu II, Zelikovsky A. 2011. Estimation of alternative splicing isoform frequencies from RNA-Seq data. *Algorithms Mol. Biol.* 6:9

82. Mezlini AM, Smith EJM, Fiume M, Buske O, Savich GL, et al. 2013. iReckon: simultaneous isoform discovery and abundance estimation from RNA-seq data. *Genome Res.* 23:519–29
83. Zakeri M, Srivastava A, Almodaresi F, Patro R. 2017. Improved data-driven likelihood factorizations for transcript abundance estimation. *Bioinformatics* 33:i142–51
84. Roberts A, Pachter L. 2013. Streaming fragment assignment for real-time analysis of sequencing experiments. *Nat. Methods* 10:71–73
85. Cappé O, Moulines E. 2009. On-line expectation–maximization algorithm for latent data models. *J. R. Stat. Soc. B* 71:593–613
86. Trapnell C, Williams BA, Pertea G, Mortazavi A, Kwan G, et al. 2010. Transcript assembly and quantification by RNA-Seq reveals unannotated transcripts and isoform switching during cell differentiation. *Nat. Biotechnol.* 28:511–15
87. Li W, Feng J, Jiang T. 2011. IsoLasso: a LASSO regression approach to RNA-Seq based transcriptome assembly. *J. Comput. Biol.* 18:1693–707
88. Canzar S, Andreotti S, Weese D, Reinert K, Klau GW. 2016. CIDANE: comprehensive isoform discovery and abundance estimation. *Genome Biol.* 17:16
89. Maretty L, Sibbesen JA, Krogh A. 2014. Bayesian transcriptome assembly. *Genome Biol.* 15:501
90. Shi X, Wang X, Wang T-L, Hilakivi-Clarke L, Clarke R, Xuan J. 2018. SparseIso: a novel Bayesian approach to identify alternatively spliced isoforms from RNA-seq data. *Bioinformatics* 34:56–63
91. Pertea M, Pertea GM, Antonescu CM, Chang T-C, Mendell JT, Salzberg SL. 2015. StringTie enables improved reconstruction of a transcriptome from RNA-seq reads. *Nat. Biotechnol.* 33:290–95
92. Tomescu AI, Kuosmanen A, Rizzi R, Mäkinen V. 2013. A novel min-cost flow method for estimating transcript expression with RNA-Seq. *BMC Bioinform.* 14(Suppl. 5):S15
93. Bernard E, Jacob L, Mairal J, Vert J-P. 2014. Efficient RNA isoform identification and quantification from RNA-Seq data with network flows. *Bioinformatics* 30:2447–55
94. Shao M, Kingsford C. 2017. Accurate assembly of transcripts through phase-preserving graph decomposition. *Nat. Biotechnol.* 35:1167–69
95. Glaus P, Honkela A, Rattray M. 2012. Identifying differentially expressed transcripts from RNA-seq data with biological variation. *Bioinformatics* 28:1721–28
96. SEQC/MAQC-III Consort. 2014. A comprehensive assessment of RNA-seq accuracy, reproducibility and information content by the Sequencing Quality Control Consortium. *Nat. Biotechnol.* 32:903–14
97. Hensman J, Papastamoulis P, Glaus P, Honkela A, Rattray M. 2015. Fast and accurate approximate inference of transcript expression from RNA-seq data. *Bioinformatics* 31:3881–89
98. Nariai N, Hirose O, Kojima K, Nagasaki M. 2013. TIGAR: transcript isoform abundance estimation method with gapped alignment of RNA-Seq data by variational Bayesian inference. *Bioinformatics* 29:2292–99
99. Amari S-I, Nagaoka H. 2000. *Methods of Information Geometry*, transl. D Harada. Transl. Math. Monogr. 191. Oxford: Am. Math. Soc.
100. Patro R, Mount SM, Kingsford C. 2014. Sailfish enables alignment-free isoform quantification from RNA-seq reads using lightweight algorithms. *Nat. Biotechnol.* 32:462–64
101. Varadhan R, Roland C. 2004. *Squared extrapolation methods (SQUAREM): a new class of simple and efficient numerical schemes for accelerating the convergence of the EM algorithm*. Work. Pap. 63, Johns Hopkins Univ. Dep. Biostat., Baltimore, MD. **https://biostats.bepress.com/jhubiostat/paper63/**
102. Zhang Z, Wang W. 2014. RNA-Skim: a rapid method for RNA-Seq quantification at transcript level. *Bioinformatics* 30:i283–92
103. Bray NL, Pimentel H, Melsted P, Pachter L. 2016. Near-optimal probabilistic RNA-seq quantification. *Nat. Biotechnol.* 34:525–27
104. Patro R, Duggal G, Love MI, Irizarry RA, Kingsford C. 2017. Salmon provides fast and bias-aware quantification of transcript expression. *Nat. Methods* 14:417–19
105. Foulds J, Boyles L, DuBois C, Smyth P, Welling M. 2013. Stochastic collapsed variational Bayesian inference for latent Dirichlet allocation. In *Proceedings of the 19th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, ed. R Ghani, TE Senator, P Bradley, R Parekh, J He, pp. 446–54. New York: Assoc. Comput. Mach.

106. Ju CJ-T, Li R, Wu Z, Jiang J-Y, Yang Z, Wang W. 2017. Fleximer: accurate quantification of RNA-Seq via variable-length k-mers. In *Proceedings of the 8th ACM International Conference on Bioinformatics, Computational Biology, and Health Informatics*, pp. 263–72. New York: Assoc. Comput. Mach.
107. Kanitz A, Gypas F, Gruber AJ, Gruber AR, Martin G, Zavolan M. 2015. Comparative assessment of methods for the computational inference of transcript isoform abundance from RNA-seq data. *Genome Biol*. 16:150
108. Germain P-L, Vitriolo A, Adamo A, Laise P, Das V, Testa G. 2016. RNAontheBENCH: computational and empirical resources for benchmarking RNAseq quantification and differential expression methods. *Nucleic Acids Res*. 44:5054–67
109. Teng M, Love MI, Davis CA, Djebali S, Dobin A, et al. 2016. A benchmark for RNA-seq quantification pipelines. *Genome Biol*. 17:74
110. Zhang C, Zhang B, Lin L-L, Zhao S. 2017. Evaluation and comparison of computational tools for RNA-seq isoform quantification. *BMC Genom*. 18:583
111. Prakash C, Haeseler AV. 2017. An enumerative combinatorics model for fragmentation patterns in RNA sequencing provides insights into nonuniformity of the expected fragment starting-point and coverage profile. *J. Comput. Biol.* 24:200–12
112. Jones DC, Kuppusamy KT, Palpant NJ, Peng X, Murry CE, et al. 2016. Isolator: accurate and stable analysis of isoform-level expression in RNA-Seq experiments. bioRxiv 088765. **https://doi.org/10.1101/088765**
113. Soneson C, Love MI, Patro R, Hussain S, Malhotra D, Robinson MD. 2018. A junction coverage compatibility score to quantify the reliability of transcript abundance estimates and annotation catalogs. *Life Sci. Alliance* 2:e201800175
114. Efron B, Hastie T. 2016. *Computer Age Statistical Inference: Algorithms, Evidence, and Data Science*. New York: Cambridge Univ. Press. 1st ed.
115. Tusher VG, Tibshirani R, Chu G. 2001. Significance analysis of microarrays applied to the ionizing radiation response. *PNAS* 98:5116–21
116. Smyth GK. 2004. Linear models and empirical Bayes methods for assessing differential expression in microarray experiments. *Stat. Appl. Genet. Mol. Biol.* 3:3
117. Bourgon R, Gentleman R, Huber W. 2010. Independent filtering increases detection power for high-throughput experiments. *PNAS* 107:9546–51
118. Marioni JC, Mason CE, Mane SM, Stephens M, Gilad Y. 2008. RNA-seq: an assessment of technical reproducibility and comparison with gene expression arrays. *Genome Res*. 18:1509–17
119. Robinson MD, Oshlack A. 2010. A scaling normalization method for differential expression analysis of RNA-seq data. *Genome Biol*. 11:R25
120. Anders S, Huber W. 2010. Differential expression analysis for sequence count data. *Genome Biol*. 11:R106
121. Dillies M-A, Rau A, Aubert J, Hennequet-Antier C, Jeanmougin M, et al. 2013. A comprehensive evaluation of normalization methods for Illumina high-throughput RNA sequencing data analysis. *Brief. Bioinform.* 14:671–83
122. Hansen KD, Irizarry RA, Wu Z. 2012. Removing technical variability in RNA-seq data using conditional quantile normalization. *Biostatistics* 13:204–16
123. Risso D, Schwartz K, Sherlock G, Dudoit S. 2011. GC-content normalization for RNA-Seq data. *BMC Bioinform*. 12:480
124. Lovén J, Orlando DA, Sigova AA, Lin CY, Rahl PB, et al. 2012. Revisiting global gene expression analysis. *Cell* 151:476–82
125. Risso D, Ngai J, Speed TP, Dudoit S. 2014. Normalization of RNA-seq data using factor analysis of control genes or samples. *Nat. Biotechnol.* 32:896–902
126. Taruttis F, Feist M, Schwarzfischer P, Gronwald W, Kube D, et al. 2017. External calibration with *Drosophila* whole-cell spike-ins delivers absolute mRNA fold changes from human RNA-Seq and qPCR data. *Biotechniques* 62:53–61
127. Hicks SC, Okrah K, Paulson JN, Quackenbush J, Irizarry RA, Bravo HC. 2018. Smooth quantile normalization. *Biostatistics* 19:185–98

128. Vallejos CA, Risso D, Scialdone A, Dudoit S, Marioni JC. 2017. Normalizing single-cell RNA sequencing data: challenges and opportunities. *Nat. Methods* 14:565–71
129. Love MI, Huber W, Anders S. 2014. Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. *Genome Biol.* 15:550
130. Robinson MD, McCarthy DJ, Smyth GK. 2010. edgeR: a Bioconductor package for differential expression analysis of digital gene expression data. *Bioinformatics* 26:139–40
131. Hardcastle TJ, Kelly KA. 2010. baySeq: empirical Bayesian methods for identifying differential expression in sequence count data. *BMC Bioinform.* 11:422
132. Leng N, Dawson JA, Thomson JA, Ruotti V, Rissman AI, et al. 2013. EBSeq: an empirical Bayes hierarchical model for inference in RNA-seq experiments. *Bioinformatics* 29:1035–43
133. Robinson MD, Smyth GK. 2007. Moderated statistical tests for assessing differences in tag abundance. *Bioinformatics* 23:2881–87
134. McCarthy DJ, Chen Y, Smyth GK. 2012. Differential expression analysis of multifactor RNA-Seq experiments with respect to biological variation. *Nucleic Acids Res.* 40:4288–97
135. Himes BE, Jiang X, Wagner P, Hu R, Wang Q, et al. 2014. RNA-Seq transcriptome profiling identifies CRISPLD2 as a glucocorticoid responsive gene that modulates cytokine function in airway smooth muscle cells. *PLOS ONE* 9:e99625
136. Love MI, Anders S, Kim V, Huber W. 2016. RNA-Seq workflow: gene-level exploratory analysis and differential expression. *F1000Research* 4:1070
137. Robinson MD, Smyth GK. 2008. Small-sample estimation of negative binomial dispersion, with applications to SAGE data. *Biostatistics* 9:321–32
138. Soneson C, Delorenzi M. 2013. A comparison of methods for differential expression analysis of RNA-seq data. *BMC Bioinform.* 14:91
139. Zhou X, Lindsay H, Robinson MD. 2014. Robustly detecting differential expression in RNA sequencing data using observation weights. *Nucleic Acids Res.* 42:e91
140. Cox DR, Reid N. 1987. Parameter orthogonality and approximate conditional inference. *J. R. Stat. Soc. B* 49:1–39
141. Chen Y, Lun ATL, Smyth GK. 2014. Differential expression analysis of complex RNA-seq experiments using edgeR. In *Statistical Analysis of Next Generation Sequencing Data*, ed. S Datta, D Nettleton, pp. 51–74. Cham, Switz.: Springer Int.
142. Wu H, Wang C, Wu Z. 2013. A new shrinkage estimator for dispersion improves differential expression detection in RNA-seq data. *Biostatistics* 14:232–43
143. Nelder JA, Wedderburn RWM. 1972. Generalized linear models. *J. R. Stat. Soc. A* 135:370–84
144. Benjamini Y, Hochberg Y. 1995. Controlling the false discovery rate: a practical and powerful approach to multiple testing. *J. R. Stat. Soc. B* 57:289–300
145. Benjamini Y, Yekutieli D. 2001. The control of the false discovery rate in multiple testing under dependency. *Ann. Stat.* 29:1165–88
146. Lund SP, Nettleton D, McCarthy DJ, Smyth GK. 2012. Detecting differential expression in RNA-sequence data using quasi-likelihood with shrunken dispersion estimates. *Stat. Appl. Genet. Mol. Biol.* 11:5
147. Di Y, Schafer DW, Cumbie JS, Chang JH. 2011. The NBP negative binomial model for assessing differential gene expression from RNA-Seq. *Stat. Appl. Genet. Mol. Biol.* 10:24
148. van de Wiel MA, Leday GGR, Pardo L, Rue H, van der Vaart AW, van Wieringen WN. 2013. Bayesian analysis of RNA sequencing data by estimating multiple shrinkage priors. *Biostatistics* 14:113–28
149. van de Wiel MA, Neerincx M, Buffart TE, Sie D, Verheul HMW. 2014. ShrinkBayes: a versatile R-package for analysis of count-based sequencing data in complex study designs. *BMC Bioinform.* 15:116
150. Rue H, Martino S, Chopin N. 2009. Approximate Bayesian inference for latent Gaussian models by using integrated nested Laplace approximations. *J. R. Stat. Soc. B* 71:319–92
151. Law CW, Chen Y, Shi W, Smyth GK. 2014. voom: precision weights unlock linear model analysis tools for RNA-seq read counts. *Genome Biol.* 15:R29
152. Li J, Tibshirani R. 2013. Finding consistent patterns: a nonparametric approach for identifying differential expression in RNA-Seq data. *Stat. Methods Med. Res.* 22:519–36

153. Stephens M. 2016. False discovery rates: a new deal. *Biostatistics* 18:275–94
154. Zhu A, Ibrahim JG, Love MI. 2018. Heavy-tailed prior distributions for sequence count data: removing the noise and preserving large differences. *Bioinformatics*. In press. **https://doi.org/10.1093/bioinformatics/bty895**
155. Leek JT, Storey JD. 2007. Capturing heterogeneity in gene expression studies by surrogate variable analysis. *PLOS Genet*. 3:12
156. Leek JT. 2014. svaseq: removing batch effects and other unwanted noise from sequencing data. *Nucleic Acids Res*. 42:e161
157. Finner H, Roters M. 2001. On the false discovery rate and expected type I errors. *Biomet. J.* 43:985–1005
158. McCarthy DJ, Smyth GK. 2009. Testing significance relative to a fold-change threshold is a TREAT. *Bioinformatics* 25:765–71
159. Chen Y, Lun ATL, Smyth GK. 2016. From reads to genes to pathways: differential expression analysis of RNA-Seq experiments using Rsubread and the edgeR quasi-likelihood pipeline. *F1000Research* 5:1438
160. Di Y, Emerson SC, Schafer DW, Kimbrel JA, Chang JH. 2013. Higher order asymptotics for negative binomial regression inferences from RNA-sequencing data. *Stat. Appl. Genet. Mol. Biol.* 12:49–70
161. Storey JD. 2003. The positive false discovery rate: a Bayesian interpretation and the *q*-value. *Ann. Stat.* 31:2013–35
162. Ignatiadis N, Klaus B, Zaugg JB, Huber W. 2016. Data-driven hypothesis weighting increases detection power in genome-scale multiple testing. *Nat. Methods* 13:577–80
163. Efron B. 2004. Large-scale simultaneous hypothesis testing. *J. Am. Stat. Assoc.* 99:96–104
164. Efron B. 2007. Size, power and false discovery rates. *Ann. Stat*. 35:1351–77
165. Van den Berge K, Soneson C, Robinson MD, Clement L. 2017. stageR: a general stage-wise method for controlling the gene-level false discovery rate in differential expression and differential transcript usage. *Genome Biol*. 18:151
166. Heller R, Manduchi E, Grant GR, Ewens WJ. 2009. A flexible two-stage procedure for identifying gene sets that are differentially expressed. *Bioinformatics* 25:1019–25
167. Kakaradov B, Xiong HY, Lee LJ, Jojic N, Frey BJ. 2012. Challenges in estimating percent inclusion of alternatively spliced junctions from RNA-seq data. *BMC Bioinform*. 13(Suppl. 6):S11
168. Turro E, Astle WJ, Tavaré S. 2014. Flexible analysis of RNA-seq data using mixed effects models. *Bioinformatics* 30:180–88
169. Papastamoulis P, Rattray M. 2018. A Bayesian model selection approach for identifying differentially expressed transcripts from RNA sequencing data. *J. R. Stat. Soc. C* 67:3–23
170. Pimentel H, Bray NL, Puente S, Melsted P, Pachter L. 2017. Differential analysis of RNA-seq incorporating quantification uncertainty. *Nat. Methods* 14:687–89
171. Blekhman R, Marioni JC, Zumbo P, Stephens M, Gilad Y. 2010. Sex-specific and lineage-specific alternative splicing in primates. *Genome Res*. 20:180–89
172. Purdom E, Simpson KM, Robinson MD, Conboy JG, Lapuk AV, Speed TP. 2008. FIRMA: a method for detection of alternative splicing from exon array data. *Bioinformatics* 24:1707–14
173. Anders S, Reyes A, Huber W. 2012. Detecting differential usage of exons from RNA-seq data. *Genome Res*. 22:2008–17
174. Soneson C, Matthes KL, Nowicka M, Law CW, Robinson MD. 2016. Isoform prefiltering improves performance of count-based methods for analysis of differential transcript usage. *Genome Biol*. 17:12
175. Love MI, Soneson C, Patro R. 2018. Swimming downstream: statistical analysis of differential transcript usage following Salmon quantification. *F1000Research* 7:952
176. Nowicka M, Robinson MD. 2016. DRIMSeq: a Dirichlet-multinomial framework for multivariate count outcomes in genomics. *F1000Research* 5:1356
177. Li YI, Knowles DA, Humphrey J, Barbeira AN, Dickinson SP, et al. 2017. Annotation-free quantification of RNA splicing using LeafCutter. *Nat. Genet.* 50:151–58
178. Papastamoulis P, Rattray M. 2017. Bayesian estimation of differential transcript usage from RNA-seq data. *Stat. Appl. Genet. Mol. Biol.* 16:367–86

179. Froussios K, Mourão K, Simpson GG, Barton GJ, Schurch NJ. 2017. Identifying differential isoform abundance with RATs: a universal tool and a warning. bioRxiv 132761. **https://doi.org/10.1101/132761**
180. Shen S, Park JW, Lu Z-X, Lin L, Henry MD, et al. 2014. rMATS: robust and flexible detection of differential alternative splicing from replicate RNA-Seq data. *PNAS* 111:E5593–601
181. Trincado JL, Entizne JC, Hysenaj G, Singh B, Skalic M, et al. 2018. SUPPA2: fast, accurate, and uncertainty-aware differential splicing analysis across multiple conditions. *Genome Biol*. 19:40
182. Yi L, Pimentel H, Bray NL, Pachter L. 2018. Gene-level differential analysis at transcript-level resolution. *Genome Biol*. 19:53
183. Hicks SC, Townes FW, Teng M, Irizarry RA. 2017. Missing data and technical variability in single-cell RNA-sequencing experiments. *Biostatistics* 19:562–78
184. Marinov GK, Williams BA, McCue K, Schroth GP, Gertz J, et al. 2014. From single-cell to cell-pool transcriptomes: stochasticity in gene expression and RNA splicing. *Genome Res*. 24:496–510
185. Svensson V, Vento-Tormo R, Teichmann SA. 2018. Exponential scaling of single-cell RNA-seq in the past decade. *Nat. Protoc.* 13:599–604
186. Wagner A, Regev A, Yosef N. 2016. Revealing the vectors of cellular identity with single-cell genomics. *Nat. Biotechnol.* 34:1145–60
187. Moor AE, Itzkovitz S. 2017. Spatial transcriptomics: paving the way for tissue-level systems biology. *Curr. Opin. Biotechnol.* 46:126–33
188. Kumar P, Tan Y, Cahan P. 2017. Understanding development and stem cells using single cell-based analyses of gene expression. *Development* 144:17–32
189. Kang HM, Subramaniam M, Targ S, Nguyen M, Maliskova L, et al. 2018. Multiplexed droplet single-cell RNA-sequencing using natural genetic variation. *Nat. Biotechnol.* 36:89–94
190. Paulson KG, Voillet V, McAfee MS, Hunter DS, Wagener FD, et al. 2018. Acquired cancer resistance to combination immunotherapy from transcriptional loss of class I HLA. *Nat. Commun.* 9:3868
191. Giladi A, Amit I. 2018. Single-cell genomics: a stepping stone for future immunology discoveries. *Cell* 172:14–21
192. Trapnell C. 2015. Defining cell types and states with single-cell genomics. *Genome Res*. 25:1491–98
193. Sandberg R. 2014. Entering the era of single-cell transcriptomics in biology and medicine. *Nat. Methods* 11:22–24
194. Soneson C, Robinson MD. 2018. Bias, robustness and scalability in single-cell differential expression analysis. *Nat. Methods* 15:255–61
195. Jaakkola MK, Seyednasrollah F, Mehmood A, Elo LL. 2017. Comparison of methods to detect differentially expressed genes between single-cell populations. *Brief. Bioinform.* 18:735–43
196. Kharchenko PV, Silberstein L, Scadden DT. 2014. Bayesian approach to single-cell differential expression analysis. *Nat. Methods* 11:740–42
197. Finak G, McDavid A, Yajima M, Deng J, Gersuk V, et al. 2015. MAST: a flexible statistical framework for assessing transcriptional changes and characterizing heterogeneity in single-cell RNA-seq data. *Genome Biol*. 16:278
198. Van den Berge K, Perraudeau F, Soneson C, Love MI, Risso D, et al. 2018. Observation weights unlock bulk RNA-seq tools for zero inflation and single-cell applications. *Genome Biol.* 19:24
199. Risso D, Perraudeau F, Gribkova S, Dudoit S, Vert J-P. 2018. A general and flexible method for signal extraction from single-cell RNA-seq data. *Nat. Commun.* 9:284
200. Eling N, Richard AC, Richardson S, Marioni JC, Vallejos CA. 2018. Correcting the mean-variance dependency for differential variability testing using single-cell RNA sequencing data. *Cell Syst*. 7:284–94.e12
201. Korthauer KD, Chu L-F, Newton MA, Li Y, Thomson J, et al. 2016. A statistical approach for identifying differential distributions in single-cell RNA-seq experiments. *Genome Biol.* 17:222
202. Treangen TJ, Salzberg SL. 2012. Repetitive DNA and next-generation sequencing: computational challenges and solutions. *Nat. Rev. Genet.* 13:36–46
203. Weirather JL, de Cesare M, Wang Y, Piazza P, Sebastiano V, et al. 2017. Comprehensive comparison of Pacific Biosciences and Oxford Nanopore Technologies and their applications to transcriptome analysis. *F1000Research* 6:100

204. Steijger T, Abril JF, Engström PG, Kokocinski F, RGASP Consort., et al. 2013. Assessment of transcript reconstruction methods for RNA-seq. *Nat. Methods* 10:1177–84
205. Tilgner H, Grubert F, Sharon D, Snyder MP. 2014. Defining a personal, allele-specific, and single-molecule long-read transcriptome. *PNAS* 111:9869–74
206. Gonzalez-Garay ML. 2016. Introduction to isoform sequencing using Pacific Biosciences Technology (Iso-Seq). In *Transcriptomics and Gene Regulation*, ed. J Wu, pp. 141–60. Dordrecht, Neth.: Springer Neth.
207. Laver T, Harrison J, O'Neill PA, Moore K, Farbos A, et al. 2015. Assessing the performance of the Oxford Nanopore Technologies MinION. *Biomol. Detect. Quantif.* 3:1–8
208. Ross MG, Russ C, Costello M, Hollinger A, Lennon NJ, et al. 2013. Characterizing and measuring bias in sequence data. *Genome Biol.* 14:R51
209. Teng JLL, Yeung ML, Chan E, Jia L, Lin CH, et al. 2017. PacBio but not Illumina technology can achieve fast, accurate and complete closure of the high GC, complex *Burkholderia pseudomallei* two-chromosome genome. *Front. Microbiol.* 8:1448
210. Eid J, Fehr A, Gray J, Luong K, Lyle J, et al. 2009. Real-time DNA sequencing from single polymerase molecules. *Science* 323:133–38
211. Levene MJ, Korlach J, Turner SW, Foquet M, Craighead HG, Webb WW. 2003. Zero-mode waveguides for single-molecule analysis at high concentrations. *Science* 299:682–86
212. Travers KJ, Chin C-S, Rank DR, Eid JS, Turner SW. 2010. A flexible and efficient template format for circular consensus sequencing and SNP detection. *Nucleic Acids Res.* 38:e159
213. Carneiro MO, Russ C, Ross MG, Gabriel SB, Nusbaum C, DePristo MA. 2012. Pacific Biosciences sequencing technology for genotyping and variation discovery in human data. *BMC Genom.* 13:375
214. Wang Y, Yang Q, Wang Z. 2014. The evolution of nanopore sequencing. *Front. Genet.* 5:449
215. Quick J, Quinlan AR, Loman NJ. 2014. A reference bacterial genome dataset generated on the MinION^{TM} portable single-molecule nanopore sequencer. *Gigascience* 3:22
216. Garalde DR, Snell EA, Jachimowicz D, Sipos B, Lloyd JH, et al. 2018. Highly parallel direct RNA sequencing on an array of nanopores. *Nat. Methods* 15:201–6
217. Smith AM, Jain M, Mulroney L, Garalde DR, Akeson M. 2017. Reading canonical and modified nucleotides in 16S ribosomal RNA using nanopore direct RNA sequencing. bioRxiv 132274. **http://doi.org/10.1101/132274**
218. Sharon D, Tilgner H, Grubert F, Snyder M. 2013. A single-molecule long-read survey of the human transcriptome. *Nat. Biotechnol.* 31:1009–14
219. Oikonomopoulos S, Wang YC, Djambazian H, Badescu D, Ragoussis J. 2016. Benchmarking of the Oxford Nanopore MinION sequencing for quantitative and qualitative assessment of cDNA populations. *Sci. Rep.* 6:31602
220. Abdel-Ghany SE, Hamilton M, Jacobi JL, Ngam P, Devitt N, et al. 2016. A survey of the sorghum transcriptome using single-molecule long reads. *Nat. Commun.* 7:11706
221. Au KF, Sebastiano V, Afshar PT, Durruthy JD, Lee L, et al. 2013. Characterization of the human ESC transcriptome by hybrid sequencing. *PNAS* 110:E4821–30
222. Treutlein B, Gokce O, Quake SR, Südhof TC. 2014. Cartography of neurexin alternative splicing mapped by single-molecule long-read mRNA sequencing. *PNAS* 111:E1291–99
223. Rhoads A, Au KF. 2015. PacBio sequencing and its applications. *Genom. Proteom. Bioinform.* 13:278–89
224. Marchet C, Lecompte L, Da Silva C, Cruaud C, Aury JM, et al. 2017. *De novo* clustering of long reads by gene from transcriptomics data. *Nucleic Acids Res.* 47:e2
225. Workman RE, Myrka AM, Wong GW, Tseng E, Welch KC Jr., Timp W. 2018. Single-molecule, full-length transcript sequencing provides insight into the extreme metabolism of the ruby-throated hummingbird *Archilochus colubris*. *Gigascience* 7:1–12
226. An D, Cao HX, Li C, Humbeck K, Wang W. 2018. Isoform sequencing and state-of-art applications for unravelling complexity of plant transcriptomes. *Genes* 9:43
227. Gordon SP, Tseng E, Salamov A, Zhang J, Meng X, et al. 2015. Widespread polycistronic transcripts in fungi revealed by single-molecule mRNA sequencing. *PLOS ONE* 10:e0132628
228. Li H. 2018. Minimap2: pairwise alignment for nucleotide sequences. *Bioinformatics* 34:3094–100

Annual Review of Biomedical Data Science

Volume 2, 2019

Contents

Errata

An online log of corrections to *Annual Review of Biomedical Data Science* articles may be found at http://www.annualreviews.org/errata/biodatasci