RNA sequencing data: hitchhiker's guide to expression analysis

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

1 Bioinformatics Institute 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
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, UNC-Chapel Hill, Chapel Hill, NC, USA
5 Department of Computer Science, Stony Brook University, NY, USA

* these authors contributed equally
† these authors contributed equally
‡ correspondence to: mark.robinson@imls.uzh.ch

Abstract

Gene expression is the fundamental level at which the result 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 datasets as well as the performance of the myriad of methods developed. In this review, we give an overall view of the developments in RNA-seq data analysis, including experimental design, with an explicit focus on 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.

Abbreviations:
DE - differential expression / differentially expressed
DGE - differential gene expression
DTE - differential transcript expression
DTU - differential transcript usage
GLM - generalized linear model
DS - differential splicing
LRT - likelihood ratio test
TPM - transcripts per million

SJ - splice junction
bp - base pairs
NB - negative binomial
ML - maximum likelihood
MM - method of moments
FDR - false discovery rate
VB - variational Bayes
LRTS - long-read transcript sequencing
APL - approximate profile likelihood
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." - from Mostly Harmless by Douglas Adams

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

![Diagram of RNA sequencing workflow]

Figure 1: Overview of the experimental steps in a RNA-seq protocol. The cDNA library is generated from isolated RNA targets, 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.

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 of interest 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 RNAs are then chemically or enzymatically fragmented to molecules of appropriate length.
size (e.g., Illumina TruSeq: 300-500 bp). Current dominant systems (e.g., Illumina) only sequence DNA; single-stranded target RNAs are thus reverse transcribed to cDNA (first-strand), the RNA is 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.

**Figure 2:** Ridge plot showing the progression of read length, depth and sample size in Sequence Read Archive (SRA) projects from the recount package [14]. The projects are separated by the submission year of the biosample. (A) Median read length of all samples per project and year. The color of the ridge indicates the library type of the project and the black vertical line marks the median. There are 787 single-end and 1008 paired-end projects. Each point represents one project. (B) Median number of reads across all samples per project and year. (C) Number of samples in each project.

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 unstripped 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 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 reads around 35bp long and modern (Illumina-based) experiments typically employing 50bp (single-end) or 100bp (paired-end) reads (Figure 2A). Most RNA-seq experiments comprise between 10 and 100 million reads, but the projects highlight a trend towards deeper sequencing over time (Figure 2B). However, the number of samples per project remained constant over the years with a median of around 8 samples (Figure 2C). Rapid enhancements in sequencing technology have enabled not only longer read lengths (e.g., Illumina MiSeq at 250-300bp) 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 [15]. Further developments are summarized below in the section ‘Long Read Transcriptome Sequencing’. In addition, single cell RNA-seq is a rapidly emerging technique that can be used to sequence the sparse transcriptome of individual cells instead of that from a potentially heterogeneous mixture of cells. Some of the early developments in this area are captured in the section ‘Single Cell Transcriptome Sequencing’.

Figure 3: Cumulative proportion of reads amongst 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 RNA-seq sample. Counts were downloaded from recount2 [14] and 50 samples were randomly selected from Accession SRP060416.

Design aspects of RNA-seq

The basics of experimental design apply equally for RNA-seq as they do for other scientific experiments (e.g., see [16]). For example, whether the desired experiment is a simple two-group design or a full factorial design, considerations towards randomizing experimental units to treatments and to avoid confounding factors (e.g., via blocking over batches) should be respected. In an experiment that must be run in multiple batches (e.g., 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 in itself is a hypothesis-generating tool, the first driver of sample size is budget. For better or worse, many RNA-seq studies use as few as three replicates per condition (Figure 2C), near to 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 gene expression [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 of differentially expressed genes. A recent study came to the conclusion that the recommended sample sizes vary from tool to tool, even when estimates from pilot data are available [21]. Another issue with sample size calculators is that it might not be obvious 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 tradeoff between the number of samples and the sequencing depth in terms of discovery performance. Increasing the number of reads might seem always beneficial, but it is only part of the equation. In an RNA-seq experiment, a large proportion of the reads originate from a small pool of highly expressed genes and unlike earlier gene expression technologies, there is 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 and therefore the statistical power to detect differential expression (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. show that a higher number of replicates is required to identify DE of genes with low fold change and ideally at least 6 replicates per condition should be used [24].

There are options for additional capture of genes with low expression, but this implies 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], “skiptic” exons [29], circular RNAs [30], enhancer RNAs [31], fusion genes [32] and so-called epi-transcriptomics involving RNA base modifications [33].

One of the main application areas is that of 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 set of genes that change in expression, hopefully leading to some understanding of the molecular pathways that are used or altered or the regulatory components that are utilized.

Gene expression has been employed for the molecular sub-classification 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]. Not surprisingly, there is considerable interest in using RNA-seq in clinical applications, to augment or corroborate the information that genome sequencing gives [36,37].

Among others, further 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], characterization of long non-coding RNAs [42] and meta-transcriptomics [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], the uptake of direct single cell RNA-seq is growing rapidly.

Outline

The focus of this review is on data analysis aspects, with a view towards equipping the reader with an overall view of the computational steps involved (focus on DE), but also shining a light on various statistical and computational challenges and the range of approaches that have been proposed to address them. Not surprisingly, the review is largely geared towards 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] or pathway analyses [50,51]. In a large majority of applications, the overarching goal is to identify DE, whether that be at 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 (though 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 and/or annotation catalogs and the relative expression level of each target is assessed. Following quantification, we split the rather broad topic of DE into ‘Basics of differential expression’, which lays the foundation for the current frameworks, and ‘Variants of differential expression’, to highlight the diverse conceptual tools available to run the discovery process. Following this, we discuss two rapidly evolving research areas, namely ‘Single Cell Transcriptome Sequencing’ and ‘Long Read Transcriptome Sequencing’, which have both experienced considerable activity in the last few years.

Alignment and Quantification

After an experiment has been designed and executed, the analyst is presented with files containing potentially millions 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.

Figure 4: Illustration of spliced alignment of RNA-seq fragments to a genome (top) and direct alignment to a transcriptome (bottom). 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 4 exon gene consisting of 3 distinct transcripts. In the spliced alignment (top), the left read of the rightmost pair is a spliced alignment to the red-green exon boundary. In the direct alignment to the transcriptome (bottom), one observes how the same alignment (e.g., the alignment to the blue exon) is repeated for each transcript.

Because the sequenced fragments are derived from cDNA corresponding to fully (or partially) spliced transcripts, reads will often span the boundaries of splice junctions (SJ), resulting in so-called “junction-spanning” reads (e.g., the read spanning the green and red exons in Figure 4). This results in contiguous read sequences whose constituent sub-sequences may be separated by tens of thousands of nucleotides on the genome. This poses a considerable computational challenge, as the position of splice junctions in spanning reads needs to be accurately identified for a read to be properly aligned. There are two main approaches for
handling spliced reads, with each approach having its own challenges and benefits: a spliced alignment against a reference genome or 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 a 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 multi-mapping rate. We describe these approaches and challenges in further detail below.

Spliced alignment to a reference genome

Perhaps the most popular solution for handling alignments of RNA-seq data is to use a splice-aware aligner. For example, early RNA-seq aligners, e.g., TopHat [52], make 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 in the algorithms used for alignment. These tools can be provided with a genome annotation that marks the locations of known SJs. Additionally, they have the ability to discover previously unannotated SJs. When a read partially aligns, the annotated SJ database is consulted to determine if the contiguous 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 there does not exist an annotated splice site at the point where a read stops aligning to the reference, the alignment 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 may span into one of the exons. When both exons contain a sufficient length of the read, there is strong evidence for the SJ, since two relatively long read pieces are unlikely to align to other locations by chance. However, when the SJ occurs very close to the beginning or end of a read, the situation becomes more difficult. For example, if in a 100bp read, 96bp align confidently to the upstream exon, ending in a donor site, the alignment tool is left with the task of determining where the remaining 4 nucleotides should align. Even if the search is restricted (e.g., within 100kb), there will often be many places the corresponding 4-mer appears following an acceptor site. Splice-aware aligners like STAR [55], HISAT(2) [56], Subread [57] and GMAP [58], among others, 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 sufficient anchoring sequence span the SJ. Subsequently, the trusted SJs are used to help align reads that start or end near the junction boundaries. The main distinction is whether the set of newly-discovered SJs is mapped “online” (called 1.5-pass) or done in a separate phase (denoted 2-pass).
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 indexing and aligning against the 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 multi-mapping and dealing with this becomes a primary computational challenge. For example, if a gene has 3 distinct isoforms, a constitutive exon of this gene will appear 3 times in the transcriptome reference (e.g., the blue exon in Fig 4). Additionally, mapping only to annotated transcripts gives no capability 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 also become more widely used, and there are inherent tradeoffs in choosing between the two levels of resolution.

Gene-level quantification consists of assigning fragments (reads or read pairs) to “genes”, where 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 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 non-overlapping exonic segments. Even in this basic pipeline, there are many variations of how certain conditions should 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 are typically quite fast.
Conversely, they suffer from various disadvantages: they have no principled way of handling multi-mapping reads (e.g., arising from paralogous genes), and they are oblivious to potentially important compositional changes that are 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 concept of gene definition. For example, should a gene be considered to be the union of exons of all transcripts of the gene, or the intersection? Should intronic reads be included? Though the notion of the gene is a useful abstraction, transcripts are the biological entities that are assayed in RNA-seq, and so present a conceptually cleaner target for quantification.

Figure 5: Illustration of alignment of various reads to a gene with 3 isoforms (B - blue; G - green; R - red). In this example, we wish to estimate the abundances of these isoforms, but the majority of reads have ambiguous origins and need to be probabilistically assigned to the transcripts (relative probabilities for each read is shown by the magnitude of the three colors). Some reads are consistent only with the B and G transcripts (colored blue and green, respectively) and a small number of reads uniquely align to a single transcript (single color). In the EM (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 after iterating the procedure until convergence.

Transcript-level quantification consists of the assignment of fragments to specific transcripts, which is fundamentally 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 allows 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]; it provides a proper model for handling reads that multi-map; solving the transcript-level abundance estimation problem implies a principled solution to aggregating to gene-level estimates [72–74]. 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 schematic in Figure 5.

Jiang et al. [75] were among the first to attempt isoform-level abundance estimation from RNA-seq data. They define counts over exons and exon junctions as arising according to a Poisson model, and view 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 pose a statistical model that admits efficient inference, for which they both obtain the point estimate by gradient ascent, and provide 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 requires annotations of transcripts in terms of the gene-transcript relationship as well as the exon and junction read inclusion matrix.

Li et al. [72,76] proposed one of the most widely-adopted generative models for transcript quantification, RSEM. They define 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 generate fragments from the sampled molecules. Primary quantities of interest, 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 that consists of each transcript species), are estimated, and can be directly converted into popular abundance units, such as transcripts per million (TPM) or estimated counts. Notably, they propose computing the maximum likelihood (ML) estimates using an expectation-maximization (EM) algorithm (see Figure 5), and introduce a modified Gibbs sampling procedure to allow estimating credible intervals for the abundance estimates [72]. The model is quite general: it works at the fragment level; it can account for numerous protocol-related aspects, including single-end and paired-end sequencing, directional vs. unstranded protocols, various coverage biases, etc. Further, the model relies only on knowing
the transcript sequences and not the relationships to genes or annotations of exons and SJs. Thus, it can be easily applied to both well-characterized or newly-assembled transcriptomes. One drawback of adopting a fragment-level model, however, is that each EM iteration scales in the total number of alignments, which is indeed large in most RNA-seq experiments.

Instead of trying to model each fragment individually, MMSeq focuses on modeling sufficient statistics [77,78]. 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 giving 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, they also introduce a Gibbs sampling procedure that allows estimating transcript abundances using summary statistics from samples of the estimated posterior, which also allows the assessment of uncertainty in the transcript-level abundance estimation and for assessing groups of transcripts with correlated posterior estimates. It is worth noting that 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 [79], or sacrificed equality in an attempt to balance efficiency and fidelity [80,81]. 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 [82] [83].

Cufflinks is perhaps most widely known as a reference-guided transcript assembly tool [84], but it also performs quantification; either restricted to a reference annotation, or allowing 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, it is worth mentioning that, given the close relationship between transcript identification (assembly) and quantification, numerous approaches attempt to solve both problems, either stagewise or jointly [80,85–92].

A model similar to RSEM that jointly performs quantification and DE, together with fully Bayesian inference, was introduced in BitSeq [93]. BitSeq focused on sampling from the posterior distributions of transcript abundances, given the fragment alignments, giving accurate estimates [94] and useful information about posterior uncertainty and posterior correlation, which is used in the DE step [93]. To combat the heavy computational requirements, Hensman et al. introduced a variational Bayesian (VB) approximation [95] that can be efficiently optimized. A VB approach to the transcript abundance estimation problem was introduced in TiGAR [96], where the VB EM algorithm was shown to outperform the standard EM algorithm. However, Hensman et al. [95] introduced a novel optimization procedure called VBNG (Variational Bayesian Natural Gradient), which is a gradient ascent algorithm that takes into account the information geometry [97] of the underlying problem. They also suggest that EM-based methods
tend to find solutions near the boundary of the parameter space, and their quantifications are less robust than either fully Bayesian or variational Bayesian estimates [95].

Many of the mentioned approaches, among others, simplify the model or improve the efficiency of the inferential procedure, but they all rely on full alignments of each read, which can be a computationally intensive and time consuming process. Recently, a number of 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 [98]. Since the $k$-mers are completely known in advance, the relevant equivalence classes can be pre-computed, which reduces the inferential problem to one of simply counting $k$-mers and performing inference via an EM algorithm, e.g., the SQUAREM algorithm [99]. 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, known as “sigmers” [100]. 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 pseudo-alignments to directly compute the sufficient statistics of the equivalence-class-based model of transcript abundance estimation [101]. This approach uses $k$-mers to identify the transcripts with which fragments are compatible, but does not treat the $k$-mers independently. The pseudo-alignments can be computed in such a way that equivalence class counts are generated without the need to consider or compute 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, making the approach very efficient and allowing 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 [102]; it uses a two-phase algorithm for transcript abundance estimation: an online phase using a stochastic collapsed VB inference algorithm [103], where abundances and auxiliary parameters (e.g., GC bias parameters, sequence-specific bias parameters, fragment length distribution) are estimated; updates are then made 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 mis-identification 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 [81] in a way that allows one to tradeoff 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 [96] or a traditional EM algorithm. Combining the ideas of efficiently determining fragment-transcript
compatibility with the sigmer concept of RNA-skim, Fleximer uses a new matching algorithm that makes use of sets of sigmers to determine the likely loci of origin of reads, instead of treating each sigmer independently [104]. 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 pre-computed automaton. This process produces a set of transcript equivalence classes, along with a corresponding count for each, which is sometimes referred to as transcript compatibility counts (TCC); these are used in conjunction 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 [105–108] suggest that, in addition to being fast, such methods are capable of producing accurate abundance estimates — at least to the extent that simulation-based studies, sometimes adopting the assumed generative models of the quantification approaches, can be relied upon to assess such accuracy. However, work still remains to be done with respect to fast, accurate, and robust transcript-level quantification. For example, it is likely the case that the underlying models can be further improved to account for complexities in the fragmentation patterns of molecules prior to sequencing [109], to better balance robustness and sample-specific accuracy [110], or to address as-yet-uncharacterized biases. Also, 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, though it is possible to computationally flag transcripts whose estimates are unreliable [111].

Basics of differential expression

Following alignment and quantification, assessing differential expression (DE) from the estimated feature abundances is often the next challenge. We will 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. As such, we elaborate on further variations of the overall workflow in the ‘Variants of differential expression’ section.

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.
Then, we discuss various notable deviations, including alternative estimation and inference frameworks and additional strategies to ensure robustness.

![Schematic overview of a DE analysis for RNA-seq data.](https://doi.org/10.7287/peerj.preprints.27283v1)

Figure 6: Schematic overview of a DE analysis for RNA-seq data. The red boxes correspond to pipelines for count-based models (e.g., edgeR, DESeq2), while the blue boxes correspond to a linear model based pipeline (e.g., limma-voom).

It is worth first considering the magnitude of the inferential problem. Typically, only a limited number of replicates are available (e.g., 3-5 replicates per condition). One can ponder on the achievable statistical power from such small sample sizes, even for a single feature, with the real interest lying in thousands of features simultaneously. This parallel inference challenge is common to various genome-scale experiments and the statistical community has contributed strategies to at least improve the overall performance; from this, a few themes have emerged. For example, in estimating parameters for a given feature, it is often beneficial to consider the information coming from the other features in the dataset [112]. In general, genomics data is ripe for the use of empirical Bayes methods to moderate estimates, where priors for a feature are derived from a suitable set of other features measured in the dataset. In addition, it has become clear that 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 this occurs in an ad hoc strategy [113] or in formal hierarchical models [114]. 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, e.g., 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. [115] showed that filtering is valid as long as it is independent of the DE test statistic; thus, filtering on residual variance is invalid, while filtering on general 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 across-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 [3,116]. However, variation in library preparation or RNA composition between samples also contribute to between-sample variability, and must be accounted for [117]. 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 [117,118] for each sample. This can be considered to be a robust global fold change between the current sample and a (pseudo-)reference sample derived from all samples. DESeq’s median-of-ratio and edgeR’s trimmed mean of M-values (TMM, where M-values denote empirical fold changes between two samples) method are the most popular normalization approaches [119]. Both procedures assume that the majority of genes are indeed not DE, and adopt robust summarization methods to calculate the size factors in order to reduce the impact of DE genes (TMM: a trimmed weighted mean; DESeq: 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 normalisation, e.g., CQN: [120]; EDASEQ: [121]. Moreover, methods based on external spike-in features have been introduced to address normalisation for applications where many features are DE or where the basic assumptions of conventional normalization methods are violated [122,123]. 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 [124].

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, while they are actually random variables that have been estimated from the data [125], and it is unclear how ignoring their associated uncertainty affects the downstream DE analysis.
Figure 7: A) A mean-dispersion plot of the RNA-seq experiment from [126], as processed in [127]. The dispersion smoothly decreases for genes with higher expression and eventually reaches an asymptote, which can be considered as the biological variability that is present in the dataset. B) MA-plot of the RNA-seq experiment from [126], as processed in [127]. The y-axis shows the log fold change (M) and the x-axis shows the mean of normalized counts (A). The variability on the fold changes is higher for lowly expressed genes, which is intrinsic to count data.

Modeling and Estimation

Because of the typically small sample size, DE tools mainly implement parametric methods [118,128–131]. Initially, count data were log-transformed and linear models were used for DE analysis [4]. However, log-transformed counts suffer from 

heteroskedasticity (a systematic mean-variance trend) intrinsic to count data, rendering analysis with 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 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 [116], for which the variance is equal to the mean by definition. Biological replication introduces additional between-sample variability and analysis frameworks therefore resorted to one of the natural extensions, the gamma-Poisson or negative binomial (NB) distribution, which has an additional dispersion parameter and a quadratic mean-variance relationship:

\[ Y_{fi} \sim NB(\mu_{fi}, \phi_f) \]

\[ Var(Y_{fi}) = \mu_{fi} + \phi_f \mu_{fi}^2, \]
where \( Y_{ji} \) denotes the read count of feature \( f \) in sample \( i \), \( \phi_j \) is the dispersion for feature \( f \) and \( \mu_{ji} = s_j\theta_{ji} \) represents the average expression, which is driven by the true (relative) mRNA concentration in the sample, \( \theta_{ji} \), multiplied by a normalization scaling factor, \( s_j \); this also implies a characteristic dispersion-mean trend in RNA-seq datasets (Figure 7A). Initial implementations focused on two-group comparisons [118,132] and were later extended to the generalized linear model (GLM) framework, an extension of classical linear models to non-Gaussian responses [133]. GLMs allow for the inclusion of multiple treatments or covariates, thus broadening the applicability. The NB GLM model can be formulated as:

\[
Y_{ji} \sim NB(\mu_{ji}, \phi_j)
\]

\[
\log(\mu_{ji}) = \eta_{ji} \\
\eta_{ji} = X_i\beta_j + \log(s_j)
\]

where \( \eta_{ji} \) is the linear predictor, \( X_i \) denotes the design matrix, \( \beta_j \) represents the regression parameters and \( \log(s_j) \) are offsets to correct for differences in sequencing depth. Regardless of the model, the parameters \( \theta_{ji} \) or, equivalently, (a linear contrast of) \( \beta_j \), would represent the parameter(s) of interest for inference.

Reliable estimation of the dispersion parameter \( \phi_j \) is non-trivial due to limited sample sizes. Traditional ML estimators for the dispersion are negatively biased [134], 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 [134], with the rationale to obtain a stable estimate by borrowing strength over all genes. However, the common dispersion assumption was shown to be unrealistic and relaxed estimation schemes were proposed, such as moderation toward a common dispersion [132], or estimation in strata of similar expression strength [118]. For example, DESSeq adopts a method of moments (MM) estimator and assumes the dispersion to be a smooth function of the mean. The approach provides more stable dispersion estimates by fitting across genes with similar expression. In order to avoid too liberal inference, the dispersion is then set 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 [135,136]. Later approaches resorted to an approximate conditional inference scheme, the Cox-Reid adjusted profile likelihood (APL) [137], to correct for the bias in the ML estimator [133]. Again, stable estimation is provided by leveraging information across genes (Figure 8). In particular, edgeR uses a maximized weighted APL to tradeoff between gene-specific and shared dispersion estimators upon estimating the dispersion-mean trend across all genes (similar to DESSeq). The weighted likelihood:

\[
APL\phi_j + G_0APL\phi_j
\]

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 [132]. The weight given to the prior likelihood (\( G_0 \)) can also be estimated from the data [138]. Analogously, Dispersion Shrinkage for Sequencing (DSS) and DESSeq2 model the \( \log(\phi_j) \) as a Gaussian random variable, and Bayes formula is applied to
generate a posterior mode for each gene [128,139]. Hyperparameters for the (Gaussian) prior are inferred from the data, using either the MM or Cox-Reid estimator across all genes.

Figure 8: Empirical Bayes. In an RNA-seq experiment, the observed differences in gene expression across groups of samples with respect to within-group variance is assessed. The unobserved population distribution for the true within-group variance of each gene is depicted with a purple curve. Variances are estimated from limited sample size experiments, and so there is sampling variance in our estimate of the variance (orange curves). A ML estimate or a bias-corrected estimator for expression variance can be used (black arrow). Thousands of genes are typically observed and estimates are made for each, providing an empirical distribution of ML estimates across all genes (red curve). This empirical distribution of ML estimates 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. The maximum a posteriori (MAP), or posterior mode, represents a "shrunken estimate" (blue arrow), where the amount of shrinkage is determined by the shape of the likelihood and the width of the prior distribution.

Once dispersion estimates are available, the parameters of the mean model, $\beta_j$, can be estimated using standard algorithms for GLMs.

**Statistical inference**

After fitting a GLM to each feature, the statistical inference involves testing the null hypothesis $H_0$ that there is no DE between conditions, e.g., that the log fold change is zero ($LFC = 0$), against the alternative $H_1$ that the LFC differs from zero ($LFC \neq 0$). In the GLM framework, the null hypothesis can be represented either as a single regression parameter or a linear combination of parameters (contrast), which is defined by a vector $L$ so that $H_0: LFC=L\beta_j=0$.

There are multiple hypothesis tests available for GLMs with known (asymptotic) distribution under the null. 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 chi-squared 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 $W = \frac{LFC}{\text{se}(LFC)}$ asymptotically follows a standard
normal distribution under $H_0$, where $\text{se}(\cdot)$ indicates the standard error and $\widehat{LF}C$ 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 [140]; 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 family wise error rate (FWER) 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 in order 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., $\text{FDR} = \mathbb{E}[V/\max(R,1)]$, where $V$ is the number of false positive rejections and $R$ the total number of detections. The FDR was introduced in the seminal paper of Benjamini & Hochberg [141] 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 the genes [142] 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 inference 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 the many 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 that mean and variance are specified to be able to make inference on the mean model parameters [143]. The QL method adopts the same mean model structure as the NB, but introduces an additional overdispersion parameter, such that $\text{var}(Y) = \psi_Y (\mu_Y + \phi_Y \mu_Y^2)$. $\psi_Y$ 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 negative binomial power distribution, which adds an additional parameter [144] 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) [145,146] and where fitting was accomplished using Integrated nested Laplace approximations (INLA) [147], 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-transformed normalized counts using a linear model while adjusting for heteroskedasticity via weighted regression, where the observation weights are computed from the observed variance-mean relationship [148]. In this case, moderated t- and F-statistics are used for inference. Finally, non-parametric methods have been developed, which are more robust to outliers and do not require distributional assumptions, such as SAMSeq [149], which adopts the Wilcoxon test to assess DE between groups, and uses resampling procedures to adjust for differences in sequencing depth.

**Robust LFC 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: i) ratios of smaller counts result in more variable LFCs (Figure 7B); ii) estimation of LFC can be sensitive to outliers. This makes a ranking of genes according to LFC difficult, since lowly expressed or outlier-affected genes are likely to dominate the top list. In order to derive more robust LFC estimates, several approaches have been adopted. First, the use of “prior counts” 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 datasets. Second, edgeR-robust [136], 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 [128]. Lastly, priors can be imposed on the LFC parameters. Bayesian methods produce moderated estimates, in that their magnitude is lower than standard ML estimates, but generally offer reduced variability at the cost of some bias. DESeq2 includes a zero-centered Gaussian prior in the NB GLM and provides the posterior mode of LFC as output [128]. The width of the prior is set conservatively, using a weighted upper quantile of the observed log fold changes. New alternative shrinkage estimators in DESeq2 incorporate priors with heavier tails that introduce less bias, using either a mixture of Normals [150] or a Cauchy distribution [151].

**Accounting for unobserved effects.** As mentioned, (G)LMs can adjust for known confounders. However, genomic data can also be affected by unknown, 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 [152,153] and RUV [123], for instance, estimate *surrogate variables* through singular value decomposition on control features or on a matrix of model residuals so as to avoid that the phenotypic effect of interest is 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: $H_0: \text{LFC}=0$. 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 [154]. To address these practical and theoretical concerns, several tools have adopted tests relative to a LFC threshold, a procedure initially proposed for microarray data [155]. This results in a composite null hypothesis, such as $H_{\alpha}: |\text{LFC}|<\alpha$. Implementations differ; for example, DESeq2 replaces the composite null with a simple null hypothesis at the boundary of the parameter space [128]; edgeR uses a modified likelihood ratio test or a quasi-likelihood F-test against a threshold [156].

**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 can lead to inflated FDR. Initial implementations provided exact tests [134], but these can only be applied in simple designs. Another strategy is “small sample asymptotics”, essentially making use of higher-order approximations that are still compatible with the GLM framework [157].

**Multiple testing.** While the FDR achieves a more reasonable sensitivity-specificity tradeoff than family wise 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 [158], while others adopt a data-driven weighting of the p-values in the FDR correction [159]. Although the FDR is deeply rooted in statistical theory, it is not clear that DE 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 co-regulation of genes induces intergene correlations, which can alter the null distribution of the statistical test [160]; local FDR approaches were introduced that empirically estimate the null distribution. 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 prioritisation, since genes typically have a different ranking for each hypothesis. Stage-wise testing procedures can be interpreted as a generalisation of analysis of variance with post-hoc tests towards a high throughput context [161,162], which allows a natural ordering of the genes according to an “omnibus” (all effects of interest) test, 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 dataset) and hypothesis tests of interest are conducted, with an adjustment for multiple testing. In this section, we unravel a set of additional approaches to interrogate RNA-seq data in terms of DE.

Although it may be obvious that DE is of interest, this can manifest or be defined in multiple ways (see Schematic Figure 9). It is important to remember that while one may want to cast inferences to the gene level, 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, either by comparing accumulated transcript-per-million (TPM) estimates or by comparing raw counts while including an adjustment for average transcript length via offsets [73]. 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 referred to as differential transcript usage (DTU) or, more generally, differential splicing (DS). A surrogate for DTU, differential exon usage (DEU), 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 [163].

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 an aggregated output (i.e., DGE). Or, transcript-level expression can be represented as a gene-wise multivariate outcome and isoform switches considered collectively, i.e., by assessing DTU, which is not affected, in either direction, by DGE. DTU implies DTE while the opposite is not necessary true. Generally speaking, we favor the strategy of two clear but orthogonal analyses (DGE and DTU), over a catch-all DTE analysis [73], 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 [73]. Thus, transcript quantifications inferred by popular tools such as RSEM, Salmon or kallisto, carry a higher degree of uncertainty, which should be accounted for in the downstream DE analysis.

Figure 9: Schematic illustration of some examples of DGE, DTE and DTU for a gene with three isoforms (Isoform 1, 2 and 3) in a two group comparison (A vs B). Bars marked with an asterisk indicate DTE in group B relative to group A.

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 also DGE) analyses; the method quantifies transcript abundances via a likelihood model and EM algorithm and tests of DE are performed by applying the delta method on the abundance parameters [74,84]. Bayesian approaches for identifying DTE based on estimated counts, e.g., ranking via Bayes factors, include EBSeq [131], which uses an empirical Bayesian hierarchical model, and MMSeq [77,164], 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 [93,165]. 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 [166].

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 [84]. 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 [167]. Such departure-from-parallelism modeling was introduced in earlier analyses of probe-level microarray data for DTU [168]; on RNA-seq data, this approach was further formalized with DEXSeq [169], where a NB model on exon-level counts is formulated. 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 [170]; in addition, DEXSeq also works well with transcript quantifications as input [171].

In a similar vein, DRIMSeq [172] and LeafCutter [173] 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 [174]. Several tools neglect the uncertainty in estimated transcript-level counts and this is perhaps the reason for inflated FDRs [171]. To address this, RATs 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 [175]. Bayesian methods, such as cjBitSeq [165], instead of considering estimated counts and their uncertainties, 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 (PSI) values. 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 [176], which uses a LRT, and SUPPA2 [177], 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 splice junctions 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 [169,177,178].

Multi stage 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. Stage-wise testing procedures [161,162], 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 allowing them to interpret results at the transcript level [171]. The same procedure can be applied replacing transcript-level tests with exon- or event-specific tests.

Single Cell Transcriptome Sequencing
One of the emerging data types in transcriptomics is single cell RNA-seq (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 [179]. Experimentally, capture and reverse transcription efficiency become important, given that the number of mRNA transcripts per mammalian cell is estimated to vary between 50k and 300k [180].

Two main experimental approaches are used: plate-based, where cells are sorted into individual wells for lysis and library preparation; or, droplet-based, where each cell is absorbed (together with reagents) and processed within an oil droplet [181]. 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: i) generally low depth of sampling for each cell (due to cost, but also due to lower diversity of cDNA fragments); ii) so-called “dropout” where a cell expresses a transcript but it is unobserved; and, iii) higher levels of biological (since no averaging) and technical (e.g., more amplification) variation.

Nonetheless, researchers are able to distinguish cell “identities”, where identity represents the combined effects of cell “type” (permanent feature) and cell “state” (transient feature) [182]. The Human Cell Atlas, amongst other projects, opens the door for exploring spatial context [183], developmental patterns [184], immune responses [185], response to therapy [186], and an increasing range of basic science and clinical investigations [187–189].

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.

To date, several methods have been developed to decipher DE between cell types, many of which have been comparatively assessed in recent benchmarks [190,191]. Many of these single-cell-specific methods are extensions or variations of existing bulk approaches. For example, SCDE formulated the RPM (read-per-million) data for a given gene across cells as a mixture of Poisson and negative binomial components; using a Bayesian approach, probabilities of observing a given fold change are converted into empirical p-values [192]. MAST uses a hurdle model on log(TPM+1) data, where a logistic regression is used to model whether a gene is expressed or not and conditional on expression, a Gaussian linear model is used. Inferences for the two sets of regression parameters are done in a Bayesian framework that also provides regularization [193]. Again, extending existing approaches, Van den Berge et al. [194] proposed a zero-inflated NB (ZINB) model; model fitting is done within the ZINB-WaVE framework [195], 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. But since cell subsets are being compared, we may not expect to have simple shifts in the mean. Instead, it may be informative to detect and understand changes in expression variability across conditions [196]. Alternatively, full distributions (instead of means or variances) can be compared, as was proposed in a Bayesian framework in scDD [197], 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 non-parametric tests in comparing single cell subsets [190].

Beyond comparing cell types, which may or may not 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 [185]; to do this, cells from a given cell type were aggregated into a “pseudo-bulk” RNA-seq dataset and DE was performed using standard tools.
Long Read Transcriptome Sequencing

The short read length of Illumina-based RNA-seq complicates unambiguous placement of reads to the genome, especially in repeat regions [198], and adds difficulties to the assembly, identification and quantification of expressed isoforms [199–201]. In contrast, so-called third-generation, or long-read, sequencing technologies, led by Pacific Biosciences (PacBio) [202] and Oxford Nanopore Technologies (ONT), are able to generate reads that are much longer. By sequencing single molecules, they can also avoid the need for PCR amplification, hence reducing coverage biases [203,204]. Currently, long-read technologies incur a higher average cost and a higher error rate than short read sequencing [199,205]. 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 [206], where the reactions take place inside so-called zero-mode waveguides (ZMWs) [207]. 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 [208], 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 [209], 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 [210] 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 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 [211]. In contrast to PacBio, ONT also offers direct sequencing of RNA [212]. 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 [213]. 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 [15,199,214–218], as well as performance evaluations based on synthetic transcript catalogs (ERCC with 92 sequences, or SIRV, with 68). Long-read transcript sequencing (LRTS) offers the potential that every read represents a full-length transcript. If this was indeed true, de novo (reference-free) identification of the full set of observed isoforms would be straightforward, and only require grouping together reads expected to differ only by sequencing errors (which, depending on the error rate and isoform similarity, may of course still not be trivial). This is, however, not currently the case, both due to fragmentation and degradation of template molecules during library preparation and because of early termination of the sequencing, which leads to ambiguities in transcript identification [219]. In particular, 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 of the consensus sequence within each cluster [15,220–223]. 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 [224]. A recent study comparing PacBio, ONT and Illumina data [199], showed that the long-read technologies were indeed much better at correctly identifying expressed SIRV transcripts than de novo assembly of short reads.

The rapid technological developments in LRTS also mean that the read generation process, e.g., biases affecting the ability of observing 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. In particular for the PacBio RSII instrument, the selection of transcript molecules is biased towards short sequences [219]. 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 [199].

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 is 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 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.

References


85. Li W, Feng J, Jiang T. IsoLasso: a LASSO regression approach to RNA-Seq based
approach to identify alternatively spliced isoforms from RNA-seq data. Bioinformatics.
2018;34: 56–63.
89. Pertea M, Pertea GM, Antonescu CM, Chang T-C, Mendell JT, Salzberg SL. StringTie
90. Toomescu A, Kuosmanen A, Rizzi R, Mäkinen V. A novel min-cost flow method for
91. Bernard E, Jacob L, Maïral J, Vert J-P. Efficient RNA isoform identification and quantification
92. Shao M, Kingsford C. Accurate assembly of transcripts through phase-preserving graph
93. Glaus P, Honkela A, Rattray M. Identifying differentially expressed transcripts from RNA-seq
94. SEQC/MAQC-III Consortium. A comprehensive assessment of RNA-seq accuracy,
reproducibility and information content by the Sequencing Quality Control Consortium. Nat
approximate inference of transcript expression from RNA-seq data. Bioinformatics.
2015;31: 3881–3889.
method with gapped alignment of RNA-Seq data by variational Bayesian inference.
97. Amari S-I, Nagaoka H. Methods of information geometry, volume 191 of Translations of
98. Patro R, Mount SM, Kingsford C. SaliFish enables alignment-free isoform quantification from
99. Varadhan R, Roland C. Squared Extrapolation Methods (SQUAREM): A New Class of
Simple and Efficient Numerical Schemes for Accelerating the Convergence of the EM
Algorithm. bepress; 2004; Available: https://biostats.bepress.com/jhubiostat/paper63/
102. Patro R, Duggal G, Love MI, Irizarry RA, Kingsford C. Salmon provides fast and
Bayesian Inference for Latent Dirichlet Allocation. Proceedings of the 19th ACM SIGKDD


153. Leek JT. svaSeq: removing batch effects and other unwanted noise from sequencing data. Nucleic Acids Res. 2014;42. doi:10.1093/nar/gku864


