Title
Modeling and analysis of RNA-seq data: a review from a statistical perspective.

Journal
Quantitative Biology, 6(3)

ISSN
2095-4689

Authors
Li, Jingyi
Li, Wei Vivian

Publication Date
2018-09-01

DOI
10.1007/s40484-018-0144-7

Peer reviewed

Published in final edited form as:

Quant Biol. 2018 September ; 6(3): 195–209. doi:10.1007/s40484-018-0144-7.

Modeling and analysis of RNA-seq data: a review from a statistical perspective

Wei Vivian Li[1], **Jingyi Jessica Li**[1,2,*]

[1]Department of Statistics, University of California, Los Angeles, Los Angeles, CA 90095-1554, USA

[2]Department of Human Genetics, University of California, Los Angeles, Los Angeles, CA 90095-088, USA

Abstract

Background: Since the invention of next-generation RNA sequencing (RNA-seq) technologies, they have become a powerful tool to study the presence and quantity of RNA molecules in biological samples and have revolutionized transcriptomic studies. The analysis of RNA-seq data at four different levels (samples, genes, transcripts, and exons) involve multiple statistical and computational questions, some of which remain challenging up to date.

Results: We review RNA-seq analysis tools at the sample, gene, transcript, and exon levels from a statistical perspective. We also highlight the biological and statistical questions of most practical considerations.

Conclusions: The development of statistical and computational methods for analyzing RNA-seq data has made significant advances in the past decade. However, methods developed to answer the same biological question often rely on diverse statistical models and exhibit different performance under different scenarios. This review discusses and compares multiple commonly used statistical models regarding their assumptions, in the hope of helping users select appropriate methods as needed, as well as assisting developers for future method development.

Author summary:

In this review article, we provide an overview of the modeling and analysis of next-generation RNA sequencing (RNA-seq) data from a statistical perspective. We summarize state-of-the-art computational methods for RNA-seq data analysis at four different levels: sample, gene, transcript, and exon levels, and we focus on introducing and explaining their common statistical assumptions, models, and techniques. We also provide references to books and original papers for interested readers who would like to explore further technical details. Recommended readers include computational researchers focusing on methodology development and applied bioinformaticians interested in understanding the commonly used methods.

*Correspondence: jli@stat.ucla.edu.

COMPLIANCE WITH ETHICS GUIDELINES

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

The authors Wei Vivian Li and Jingyi Jessica Li declare that they have no conflict of interests.

Keywords

RNA-seq; statistical modeling; differentially expressed genes; alternatively spliced exons; isoform reconstruction and quantification

INTRODUCTION

RNA sequencing (RNA-seq) uses the next generation sequencing (NGS) technologies to reveal the presence and quantity of RNA molecules in biological samples. Since its invention, RNA-seq has revolutionized transcriptome analysis in biological research. RNA-seq does not require any prior knowledge on RNA sequences, and its high-throughput manner allows for genome-wide profiling of transcriptome landscapes [1,2]. Researchers have been using RNA-seq to catalog all transcript species, such as messenger RNAs (mRNAs) and long non-coding RNAs (lncRNAs), to determine the transcriptional structure of genes, and to quantify the dynamic expression patterns of every transcript under different biological conditions [1].

Due to the popularity of RNA-seq technologies and the increasing needs to analyze large-scale RNA-seq datasets, more than two thousand computational tools have been developed in the past ten years to assist the visualization, processing, analysis, and interpretation of RNA-seq data. The two most computationally intensive steps are data processing and analysis. In data processing, for organisms with reference genomes available, short RNA-seq reads are aligned (or mapped) to the reference genome and converted into genomic positions; for organisms without reference genomes, *de novo* transcriptome assembly is needed. Regarding the reference-based alignment, the RNA-seq Genome Annotation Assessment Project (RGASP) Consortium has conducted a systematic evaluation of mainstream spliced alignment programs for RNA-seq data [3]. We refer interested readers to this paper and do not discuss these alignment algorithms here, as statistical models are not heavily involved in the alignment step. In this paper, we focus on the statistical questions engaged in RNA-seq data analyses, assuming reads are already aligned to the reference genome. Depending on the biological questions to be answered from RNA-seq data, we categorize RNA-seq analyses at four different levels, which require three different ways of RNA-seq data summary. Sample-level analyses (*e.g.*, sample clustering) and gene-level analyses (*e.g.*, identifying differentially expressed genes [4] and constructing gene co-expression networks [5]) mostly require gene read counts, *i.e.*, number of RNA-seq reads mapped to each gene. Note that we refer to a transcribed genomic region as a “gene” throughout this review, and a multi-gene family (multiple transcribed regions that encode proteins with similar sequences) are referred to as multiple “genes”. Transcript-level analyses, such as RNA transcript assembly and quantification [6], often need read counts of genomic regions within a gene, *i.e.*, number of RNA-seq reads mapped to each region or each region-region junction, or even the exact position of each read. Exon-level analyses, such as identifying differential exon usage [7], usually require read counts of exons and exon-exon junctions. As these four levels of analyses use different statistical and computational methods, we will review the key statistical models and methods widely used at each level of RNA-seq analysis (Figure 1), with an emphasis on the identification of

differential expression and alternative splicing patterns, two of the most common goals of RNA-seq experiments.

This review does not aim to exhaustively enumerate all the existing computational tools designed for RNA-seq data, but to discuss the strategies of statistical modeling and application scopes of typical methods for RNA-seq analysis. We refer readers to Refs. [1,8] for an introduction to the development of RNA-seq technologies, and Ref. [9] for a comprehensive assessment of RNA-seq with a comparison to microarray technologies and other sequence-based platforms by the Sequencing Quality Control (SEQC) project. For considerations in experimental designs and more recent advances in computational tools, we refer readers to Ref. [10].

SAMPLE-LEVEL ANALYSIS: TRANSCRIPTOME SIMILARITY

The availability of numerous public RNA-seq datasets has created an unprecedented opportunity for researchers to compare multi-species transcriptomes under various biological conditions. Comparing transcriptomes of the same or different species can reveal molecular mechanisms behind important biological processes, and help one understand the conservation and differentiation of these molecular mechanisms in evolution. Researchers need similarity measures to directly evaluate the similarities of different samples (*i.e.*, transcriptomes) based on their genome-wide gene expression data summarized from RNA-seq experiments. Such similarity measures are useful for outlier sample detection, sample classification, and sample clustering analysis. When samples represent individual cells, similarity measures may be used to identify rare or novel cell types. In addition to gene expression, it is also possible to evaluate transcriptome similarity based on alternative splicing events [11]. Correlation analysis is a classical approach to measure transcriptome similarity of biological samples [12,13]. The most commonly used measures are Pearson and Spearman correlation coefficients. The analysis starts with calculating pairwise correlation coefficients of normalized gene expression between any two biological samples, resulting in a correlation matrix. Users can visualize the correlation matrix (usually as a heatmap) to interpret the pairwise transcriptome similarity of biological samples, or they may use the correlation matrix in downstream analysis such as sample clustering.

However, a caveat of using correlation analysis to infer transcriptome similarity is that the existence of housekeeping genes would inflate correlation coefficients. Moreover, correlation measures rely heavily on the accuracy of gene expression measurements and are not robust when the signal-to-noise ratios are relatively low. Therefore, we have developed an alternative transcriptome overlap measure, TROM [14], to find sparse correspondence of transcriptomes in the same or different species. The TROM method compares biological samples based on their “associated genes” instead of the whole gene population, thus leading to a more robust and sparse transcriptome similarity result than that of the correlation analysis. TROM defines the associated genes of a sample as the genes that have *z*-scores (normalized expression levels across samples per gene) greater than or equal to a systematically selected threshold. Pairwise TROM scores are then calculated by an overlap test to measure the similarity of associated genes for every pair of samples. The resulting TROM score matrix has the same dimensions as the correlation matrix, with rows and

columns corresponding to the samples used in the comparison, and the TROM score matrix can be easily visualized or incorporated into downstream analyses.

Aside from the correlation coefficients and the TROM scores, there are other statistical measures useful for measuring transcriptome similarity in various scenarios. First, partial correlation can be used to measure sample similarity after eliminating the part of the sample correlation attributable to another variable such as batch effects or experimental conditions [15]. Second, with evidence of a non-linear association between RNA-seq samples, it is suggested to use measures that can capture non-linear dependences, such as the mutual information (MI). Similarly, one may consider using the conditional mutual information (CMI) [16] or partial mutual information (PMI) [17] to remove the effects of other confounding variables. In addition to the direct calculation of the sample similarity matrix by applying a similarity measure to the high-dimensional gene expression data, sometimes it is helpful to visualize the gene expression data and investigate the sample similarities after dimension reduction. Popular dimension reduction methods include principal component analysis (PCA), t-stochastic neighbor embedding (t-SNE) [18], and multidimensional scaling (MDS) [19].

GENE-LEVEL ANALYSIS: GENE EXPRESSION DYNAMICS

RNA-seq technologies have enabled the measurement and comparison of genome-wide gene expression patterns across different samples without the restriction of known genes, which are required by microarray experiments. Profiling gene expression patterns is the key to investigating new biological processes in various tissues and cells of different organisms. A common and important question in a large cohort of biological studies is how to compare gene expression levels across different experimental conditions, time points, tissue and cell types, or even species. When a biological study concerns two different biological conditions, differential gene expression (DGE) analysis is useful for comparing RNA-seq samples of the two conditions. When the number of biological conditions far exceeds two, though DGE analysis can still be used to compare samples in a pairwise manner, a more useful way is to simultaneously measure the transcriptome similarity of multiple samples, as we have described in the previous Section: Sample-Level Analysis: Transcriptome Similarity.

Differential gene expression analysis

The main approach to comparing two biological conditions is to find “differentially expressed” (DE) genes. A gene is defined as DE if it is transcribed into different amounts of mRNA molecules per cell under the two conditions [20]. However, since we do not observe the true amounts of mRNA molecules, statistical tests are principled approaches that help biologists understand to what extent a gene is DE.

It is commonly acknowledged that normalization is a crucial step prior to DGE analysis due to the existence of batch effects, which could arise from different sequencing depths or various protocol-specific biases in different experiments [21]. The reads per kilobase per million mapped reads (RPKM) [22], the fragments per kilobase per million mapped reads (FPKM) [23], and the transcripts per million mapped reads (TPM) [24] are the three most frequently used units for gene expression measurements from RNA-seq data, and they

remove the effects of total sequencing depths and gene lengths. The main difference between RPKM and FPKM is that the former is a unit based on single-end reads, while the latter is based on paired-end reads and counts the two reads from the same RNA fragment as one instead of two. The difference between RPKM/FPKM and TPM is that the former calculates sample-scaling factors before dividing read counts by gene lengths, while the latter divides read counts by gene lengths first and calculates sample-scaling factors based on the length-normalized read counts. If researchers would like to interpret gene expression levels as the proportions of RNA molecules from different genes in a sample, TPM has been suggested as a better unit than RPKM/FPKM [25]. Even though in these units, gene expression data may still contain protocol-specific biases[26], and further normalization is often needed. There are two main categories of normalization methods: distribution-based and gene-based. Distribution-based normalization methods aim to make the distribution of all or most gene expression levels similar across different samples, and such methods include the quantile normalization [27], DESeq [28], and TMM [29]. Gene-based normalization methods aim to make non-DE genes or housekeeping genes have the same expression levels in different samples, and such methods include a method by Bullard *et al.* [21] and PoissonSeq [30]. For a comprehensive comparison of the assumptions and performance of these normalization methods, we refer readers to Refs. [20,21,31].

How to form a proper statistical hypothesis test is the core question in the development of a DGE method. Most existing methods use the Poisson distribution [32] or the Negative Binomial (NB) distribution [28,33,34] to model the read counts of an individual gene in different samples (Figure 2A and D). In our discussion here, we focus on the NB distribution because it is commonly used to account for the observed over-dispersion of RNA-seq read counts. Throughout this section, we consider two biological conditions $k = 1, 2$, each with J_k samples. $Y_{k,ij}$ denotes the read count of gene i in the j-th sample of condition k. The basic assumption is that

$$Y_{k,ij} \sim \mathrm{NB}\left(\text{mean} = S_{kj}\theta_{ki}, \text{dispersion} = \phi_i\right), \quad (1)$$

where s_{kj} is the size factor of the j-th sample of condition k, θ_{ki} is the true expression level of gene i under condition k, and ϕ_i is the dispersion of gene i. It is necessary to consider the size factor s_{kj} because it accounts for varying numbers of sequenced reads in various samples. The dispersion parameter ϕ_i controls the variability of the expression levels of gene i across biological samples. The estimation of the parameters s_{kj}, θ_{ki}, and ϕ_i is the key step to investigating the differential expression of gene i between the two conditions. Bayesian modeling is often used, and prior distributions and relationships of s_{kj}, θ_{ki}, and ϕ_i are often assumed. Note that assuming s_{kj} being independent of gene i simplifies the problem, but it can be advantageous to calculate gene-specific factors $s_{k,ij}$ to account for technical biases dependent on gene-specific GC contents or gene lengths [35]. The DGE analysis is carried out by testing

$$H_0: \theta_{1i} = \theta_{2i} \; vs. \; H_1: \theta_{1i} \neq \theta_{2i} \quad (2)$$

for each gene i.

Starting from the model (Equation (1)), most methods include six steps. First, they estimate θ_{ki} and ϕ_i for each gene. Under the NB distribution, the dispersion parameter characterizes the mean-variance relationship, consistent with the observation that genes with similar true expression levels exhibit similar variances [33,35]. When the sample sizes are small (Figure 2C and F), one may consider using shrinkage estimation of ϕ_i's to borrow information across genes or to incorporate prior knowledge, for the purpose of obtaining more robust results [36]. Second, they construct a test statistic based on the estimators to reflect the mean difference between the two conditions. Third, they derive the null distribution of the test statistic under H_0. Fourth, they calculate the observed value of the test statistic for each gene. Fifth, they convert the observed values of the test statistic into p-values based on the null distribution. Sixth, they perform multiple-testing correction on the p-values to determine a reasonable threshold, and the genes with p-values under that threshold would be called as DE.

For example, edgeR [33] first estimates the dispersion parameter using a conditional maximum likelihood, and it then develops a test analogous to the Fisher's exact test. DESeq2 [35] adds a layer to the model by estimating $(\theta_{2i} - \theta_{1i})$ using a generalized linear model with a logarithmic link function, $Y_{k,ij}$ as the response variable, and the condition as a binary predictor (*i.e.*, whether the condition $k = 2$). This generalized linear model setup can easily incorporate the information of experimental design as additional predictors. In the testing step, DESeq2 transforms the problem into testing if the condition predictor has significant effects on the logarithmic fold change of gene expression, which is equivalent to testing whether $\theta_{2i} - \theta_{1i} = 0$. EBSeq [37] and ShrinkSeq [38] are also based on the model (Equation (1)), but under a Bayesian framework they use hyper-parameters to borrow information across genes, and they directly calculate the posterior probability of a gene being differentially expressed, *i.e.*, $\mathbb{P}(\theta_{1i} \neq \theta_{2i} \mid Y_{1,i1}, \ldots, Y_{1,in1}, Y_{2,i1}, \ldots, Y_{2,in2})$.

There are other DGE methods that do not assume the NB distribution as in the model (Equation (1)) but take a different approach by assuming that $\log(Y_{k,ij})$ follows a Normal distribution, which has much more tractable mathematical theory than count distributions (such as the NB distribution) have. For example, the voom method[39] estimates the mean-variance relationship of $\log(Y_{k,ij})$ and generates a precision weight for each observation. Then voom inputs $\log(Y_{k,ij})$ and precision weights into the limma empirical Bayes analysis pipeline [40], which is designed for microarray data and has multiple modeling advantages: using linear modeling to analyze complex experiments with multiple treatment factors, using quantitative weights to account for variations in the precision of different observations, and using empirical Bayes methods to borrow information across genes. Another method sleuth [41] is applicable to finding both differentially expressed genes and transcripts between two conditions. Here we describe sleuth in the context of DGE analysis. Sleuth uses a linear model with $\log(Y_{k,ij})$ as the response variable, and it decomposes the variance of $\log(Y_{k,ij})$ into three components: the variance explained by the condition predictor (whose coefficient is the parameter of interest and indicates differential expression if non-zero), the variance of "biological noise" (which accounts for the variance of true gene expression across samples of the same condition), and the variance of "inferential noise" (which accounts for the

additional variance of observed gene expression due to the uncertainty in gene expression estimation). Sleuth assumes both the “biological noise” and the “inferential noise” follow independent zero-mean Gaussian distributions. For every gene, sleuth estimates the variance of the “inferential noise” by bootstrapping RNA-seq reads to estimate the variance of the expression estimates of that gene. Accurate estimation of the variance of the “inferential noise” allows better estimation of the null distribution of the test statistic, *i.e.*, the estimator of the coefficient of the condition predictor, in the third step, thus leading to more accurate estimates of the *p*-values and false discovery rates.

Remark 1. A common scenario is that a study only includes a small number of RNA-seq replicates [42]. Even though most methods introduced in this section are technically applicable to data with as few as two replicates per condition, there is no guarantee of good performance for these methods with a small number of replicates. In fact, it was observed that many methods did not have a good control on false discovery rates (FDRs) under this scenario [42] (Figure 2). We suggest users carefully check themselves or consult a statistician if the assumptions of a method are reasonable for their study before using the method, as a way to reduce the chance of misusing statistics.

Remark 2. Comparisons of DGE methods show that none of the methods is optimal in all circumstances, and methods can produce very different results (regarding both the ranking and number of DE genes) on the same dataset [4,31]. In some applications, users are more concerned about the ranking of DE genes than the resulting *p*-values of genes, especially when setting a reasonable threshold on the *p*-values is difficult. In other applications where thresholding on *p*-values is required to control the probability that a gene is falsely discovered as DE, users need to address the multiple-testing issue, as testing for tens of thousands of genes simultaneously could lead to a large number of false discoveries even at a small *p*-value threshold. Common approaches to address the multiple-testing issue include the Bonferroni correction [43], the Holm-Bonferroni method [44], and the Benjamini-Hochberg FDR correction [45], with a decreasing level of conservatism. The first two methods aim to control the family-wise error rate (the probability of making one or more false discoveries), while the third method aims to control the expected proportion of false discoveries among the discoveries.

Remark 3. In studies where researchers are interested in temporal dynamics of transcriptomes, RNA-seq data are produced at multiple time points of the same tissue or cell type. To identify the genes whose expression levels along the time course change significantly between two conditions, a previous approach maSigPro [46] is based on a linear model where gene expression level is modeled as the response variable, while time points and conditions are considered as predictors. The identification of DE genes is then formulated as the problem of testing whether the condition variable has a non-zero coefficient for each gene. Another previous work based on microarray data provided a two-sample multivariate empirical Bayes statistic (MB statistic) for replicated microarray time course data [47]. The MB statistic can be used to test the null hypothesis that the expected temporal expression profiles of one gene under two conditions are the same, and it is thus a criterion to rank genes in the order of evidence of non-zero mean difference between two

conditions, incorporating the correlation structure of time points, moderation, and replication.

Gene co-expression network analysis

A gene co-expression network (GCN) is an undirected graph, where nodes correspond to genes, and edges connecting the nodes denote the co-expression relationships between genes. GCNs can help people learn the functional relationships between genes and infer and annotate the functions of unknown genes. To the best of our knowledge, the first GCN analysis on a genome-wide scale across multiple organisms was completed in 2003, enabled by the availability of high-throughput microarray data [48]. One of the most commonly used GCN analysis methods, WGCNA, was initially developed for micro-array data but can also be used on normalized RNA-seq data [49]. It is widely applied to gene expression datasets to detect gene clusters and modules and to investigate gene connectivity by analyzing correlation networks. Here we introduce the GCN methods based on the framework proposed in [50]. We denote the gene expression matrix as $\boldsymbol{X}_{N\times J}$, where the N rows represent genes, and the J columns represent samples. The N genes are considered as N nodes in the co-expression network. The first step is to construct a symmetric adjacency matrix $\boldsymbol{A}_{N\times N}$, where A_{ij} is a similarity score in the range from 0 to 1 between genes i and j. A_{ij} measures the level of concordance between gene expression vectors $\boldsymbol{X}_{i\cdot}$ and $\boldsymbol{X}_{j\cdot}$, the i-th and j-th rows of $\boldsymbol{X}$: As discussed in the Section of Sample-level Analysis: Transcriptome Similarity, the similarity measure can be calculated based on the correlation coefficients, the TROM measure, or the mutual information measures, depending on the type of gene co-expression relationships of interest in the analysis. The elements in the adjacency matrix only consider each pair of genes when evaluating their similarity in expression profiles. However, it is important to consider the relative connectedness of gene pairs with respect to the entire network in order to detect co-expression gene modules. Therefore, one needs to calculate the topological overlap matrix $\boldsymbol{T}_{N\times N}$, where T_{ij} is the topological overlap between nodes i and j. One such example used in previous studies is [51]:

$$T_{ij}=\frac{\sum_{k=1}^{N}A_{ik}A_{kj}+A_{ij}}{\min\left\{\sum_{k=1}^{N}A_{ik},\sum_{k=1}^{N}A_{jk}\right\}+1-A_{ij}}.$$

The final distance between nodes i and j is defined as $d_{ij}=1-T_{ij}$. Clustering methods can then be applied to search for gene modules based on the resulting distance matrix. The identified gene modules are of great biological interest in many applications. For example, the modules can serve as a prioritizer to evaluate functional relationships between known disease genes and candidate genes [52]. Gene modules can also be used to detect regulatory genes and study the regulatory mechanisms in various organisms [53].

TRANSCRIPT-LEVEL ANALYSIS: TRANSCRIPT RECONSTRUCTION AND QUANTIFICATION

An important use of RNA-seq data is to recover full-length mRNA transcript structures and expression levels based on short RNA-seq reads. This application involves two major tasks.

The first task, identification of novel transcripts in RNA-seq samples, is commonly referred to as transcript/isoform reconstruction, discovery, assembly, or identification. This is one of the most challenging problems in this area due to the large search space of candidate isoforms (especially for complex genes) and inadequate information contained in short reads (Figure 3A). The second task, estimation of the expression of known or newly discovered transcripts, is usually referred to as transcript/isoform quantification or abundance estimation. In recent years, it is a common practice to combine the two tasks into one step, and many popular computational tools simultaneously perform transcript reconstruction and quantification [54]. This is usually achieved by estimating the expression levels of all the candidate isoforms with penalty or regularity constraints, and the resulting isoforms with non-zero estimated expression are treated as reconstructed isoforms. Therefore, we introduce these two tasks together in this review, as they can be tackled by the same statistical framework in many existing tools. We focus on the basic models that are commonly used by multiple methods, while selectively introducing characteristics of individual methods. These models are generally annotation-based and assume that a reference genome is available for the organism of interest.

The transcript reconstruction and quantification are performed separately for individual genes, so the following discussion applies to one gene. Throughout this section, we index the isoforms of a gene as $\{1, 2, \ldots, J\}$. In the reconstruction setting, J is the total number of candidate isoforms to be considered; in the quantification setting, J is the number of annotated (or newly discovered) isoforms to be quantified. We index the exons of the gene as $\{1, 2, \ldots, I\}$. Suppose that a total of n (single-end or paired-end) reads are mapped to the gene, and they are denoted as $\boldsymbol{R} = \{r_1, r_2, \ldots, r_n\}$. The goal of most methods is to estimate $\boldsymbol{\Theta}=(\theta_1, \theta_2, \ldots, \theta_J)^T$, where

$$\begin{aligned} \theta_j &= \text{fraction of isoform } j \\ &= \mathbb{P}(\text{a random read is from isoform } j). \end{aligned} \tag{3}$$

Likelihood-based methods

The first type of transcript discovery and quantification methods estimates transcript abundance by maximizing the likelihood or the posterior based on a statistical model. These methods are flexible and can be easily modified to incorporate prior biological information into the posterior to improve quantification accuracy. The statistical models are further divided into three categories: region-based, read-based, and fragment-based models.

Region-based models summarize the read counts based on the genomic regions of interest, such as exons and exon-exon junctions. Suppose that S is the index set that denotes all the regions of interest. Read counts can be summarized as $\boldsymbol{X}=\{X_s|s\in S\}$, where X_s is the total number of reads mapped to region s. The basic model assumes that X_s follows a Poisson distribution with parameter λ_s. Given the structures of isoforms and their compatibility with the regions, it is reasonable to assume λ_s as a linear function of the θ_j's: $\lambda_s = \sum_{j=1}^{J} a_{sj}\theta_j$. The likelihood function can then be derived, and the task of estimating $\boldsymbol{\Theta}$ reduces to a maximum likelihood estimation (MLE) problem:

$$L(\boldsymbol{\Theta}|\boldsymbol{X}) = \prod_{s \in S} \frac{e^{-\lambda_s} \lambda_s^{X_s}}{X_s!} = \prod_{s \in S} \frac{\exp\left\{-\sum_{j=1}^{J} a_{sj}\theta_j\right\}\left(\sum_{j=1}^{J} a_{sj}\theta_j\right)^{X_s}}{X_s!}, \quad (4)$$

$$\widehat{\boldsymbol{\Theta}} = \left(\hat{\theta}_1, \hat{\theta}_2, \ldots, \hat{\theta}_J\right)^T = \underset{\Theta}{\operatorname{argmax}} \log L(\boldsymbol{\Theta}|\boldsymbol{X}).$$

The first isoform quantification method [55] uses a region-based model.

In contrast to region-based models, read-based methods directly use the likelihood as a product of the probability densities of individual reads instead of first summarizing reads into region counts.

$$\begin{aligned} L(\boldsymbol{\Theta}|R) &= \prod_{i=1}^{n} p(r_i|\boldsymbol{\Theta}) \\ &= \prod_{i=1}^{n} \sum_{j=1}^{J} p(r_i|\text{isoform } j)\theta_j \\ &= \prod_{i=1}^{n} \sum_{j=1}^{J} p(r_i|\text{isoform } j)p(\ell_{ij}|\text{isoform } j)\theta_j, \end{aligned} \quad (5)$$

where s_i is the starting position and ℓ_{ij} is the read length (for single-end reads) or fragment length (for paired-end reads) of read r_i if it belongs to isoform j (Figure 3B). While many methods do not explicitly state it, they assume that the s_i and ℓ_{ij} are independent in the above model. If the two ends of read r_i are mapped to the same exon or two neighboring exons, its corresponding fragment length can be determined and remains the same for all its compatible isoforms. Otherwise, the corresponding fragment length ℓ_{ij} of read r_i could be different for a different compatible isoform j (Figure 3B). Even though each read has the same weight in the likelihood model, the reads that are mapped to two non-neighboring exons play a critical role in the detection of splicing junctions and the reconstruction of full-length transcripts. Cufflinks [56], eXpress [57], RSEM [24], and Kallisto [58] all adapted or extended the above model in their quantification step, and they mainly differ in how they model $p(s_i|\text{isoform } j)$ and $p(\ell_{ij}|\text{isoform } j)$ to incorporate sequencing bias adjustment. One feature distinguishing Kallisto from the other methods is that Kallisto speeds up the processing by pseudoaligning the reads and circumventing the computation costs of exact alignment of individual bases. To estimate $\boldsymbol{\Theta}$ by maximizing the likelihood in Equation (5), the expectation-maximization (EM) algorithm [59] is the standard optimization algorithm.

Some other methods, including WemIQ [60], Salmon[61], iReckon [62], and MSIQ [63], introduce hidden variables to denote the isoform origins of reads and use these variables to simplify the form of the likelihood function. Suppose that the isoform origins of reads $\boldsymbol{R}=\{r_1, \ldots, r_n\}$ are denoted as $\boldsymbol{Z}=(Z_1, Z_2, \ldots, Z_n)^T$, where $Z_i = j$ if read r_i comes from isoform j. Then the joint probability density of $\boldsymbol{R}$ and $\boldsymbol{Z}$ can be written as

$$p(\boldsymbol{R}, \boldsymbol{Z}|\boldsymbol{\Theta}) = \prod_{i=1}^{n} p(r_i, Z_i|\boldsymbol{\Theta}) = \prod_{i=1}^{n} \sum_{j=1}^{J} \left[p(r_i|\text{isoform } j)\theta_j\right]^{\mathbb{I}\{Z_i = j\}}. \quad (6)$$

This model formulation is especially useful when one would like to estimate $\boldsymbol{\Theta}$ under the Bayesian framework (Figure 3C), as what has been done in MISO [64], Salmon[61], and MSIQ [63]. Prior knowledge on $\boldsymbol{\Theta}$ can be incorporated via modeling the prior distribution of $\boldsymbol{\Theta}$, and $\boldsymbol{\Theta}$ would be estimated as the maximum-a-posteriori (MAP) estimator. As shown in Figure 3C, another advantage of the Bayesian framework is that the model can be easily extended to incorporate multiple RNA-seq samples and borrow isoform abundance information across samples [63].

A more recent isoform quantification method alpine[65] belongs to the third fragment-based category. Alpine is specifically designed to adjust for multiple sources of sequencing biases in isoform quantification. It considers all potential fragments with lengths within the middle of the fragment length distribution, at all possible positions within every isoform. For each fragment, alpine counts the number of reads compatible with it. Then alpine models the fragment counts using a Poisson generalized linear model, whose predictors are bias features including the length, the relative position, the read start sequence bias, the GC content and the presence of long GC stretches within every fragment. Alpine estimates the read start sequence biases using the variable length Markov model (VLMM) proposed by Roberts *et al.* [66] and implemented in Cufflinks [56]. After estimating bias parameters, alpine outputs bias-corrected isoform abundance estimates. The Poisson parameter λ_s for a potential fragment s is assumed to be $\lambda_s = \sum_{j=1}^{J} a_{sj}\theta_j$, similar to what is assumed in region-based models. Hence, θ_j's are estimated based on the bias-corrected estimates $\hat{\lambda}_s$'s.

The above approaches, however, would not lead to accurate isoform reconstruction results when directly used to discover new isoforms, because the number of candidate isoforms can be huge when the number of exons is large. A common practice is to add penalty terms before maximizing the objective function, *i.e.*, the likelihood or the posterior. The regularization aims to enforce sparsity on the estimated $\widehat{\boldsymbol{\Theta}}$, whose non-zero entries indicate the discovered isoforms. Two such reconstruction methods are iReckon [62] and NSMAP[67].

Regression-based methods

The second type of statistical methods for isoform discovery and quantification is regression-based. These methods formulate the isoform quantification problem as a linear or generalized linear model and treat the region-based read count (or proportion) as the response variable, candidate isoforms as predictor variables, and isoform abundances as coefficients (parameters) to be estimated. Regression-based methods include rQuant [68], SLIDE[69], IsoLasso [70], and CIDANE [54].

The basic model is a linear model with region-based read count proportions as the responses. As for the design matrix, IsoLasso uses a binary matrix to denote the compatibility between the isoforms and genomic regions (*i.e.*, a value of 1 indicating that an isoform and a region are compatible, and 0 otherwise), while the other three methods consider a conditional probability matrix, for which the read proportions are modeled as:

$$\begin{aligned}\frac{X_s}{n} &= \sum_{j=1}^{J} \mathbb{P}(\text{a random read falls into region } s \,|\, \text{isoform } j)\, \mathbb{P}(\text{isoform } j) + \varepsilon_s \\ &= \sum_{j=1}^{J} F_{sj}\theta_j + \varepsilon_s, \quad s \in S,\end{aligned} \tag{7}$$

where ε_s represents independent random noise with mean 0. As in the likelihood-based methods, the probability F_{sj} depends on the structure of region s and the length of isoform j. Especially when region s spans alternative splicing junctions (*e.g.*, region s skips the middle exon but includes the two end exons), the estimation accuracy of F_{sj} is critical in the modeling. Then the estimation task reduces to a penalized least-squares problem

$$\widehat{\boldsymbol{\Theta}} = \underset{\boldsymbol{\Theta} \geqslant 0}{\operatorname{argmin}} \sum_{s=1}^{S} \left(\frac{X_s}{n} - \sum_{j=1}^{J} F_{sj}\theta_j \right)^2 + \text{penalty}, \tag{8}$$

where the penalty term is only needed for isoform discovery and often excluded for isoform quantification. For example, IsoLasso sets the penalty term as $\lambda \sum_{j=1}^{J} \frac{n\theta_j}{L_j}$, where L_j is the length of isoform j, while SLIDE uses $\lambda \sum_{j=1}^{J} \frac{\theta_j}{m_j}$, where m_j is the number of exons in isoform j. For both methods, λ is a tuning parameter to control the level of regularization. IsoLasso selects l based on the resulting number of isoforms with non-zero estimated expression, while SLIDE uses a stability criterion [71].

Remark 4. There are isoform discovery methods that reconstruct mRNA transcripts based on deterministic graph methods. Examples include a *de novo* approach Trinity [72], and reference genome-based approaches Scripture [73], Cufflinks [23], and Stringtie [74], which all construct splice graphs based on aligned reads and then use various criteria to parse the constructed graph into transcripts in a deterministic way, without resorting to statistical models.

Remark 5. Despite many methods developed for isoform quantification, not all of them discuss the estimation uncertainty of isoform abundance levels. Even though the point estimates of expression levels have led to new scientific discoveries in many biological studies, it is important to consider estimation uncertainty, especially when the differential expression analysis is of interest, or when some candidate isoforms are highly similar in structures (related to the collinearity issue in linear model estimation). One way to evaluate the uncertainty in Bayesian methods is to construct posterior or credible intervals of isoform

abundance levels [63,75]. In regression-based methods, it is possible to calculate the standard errors of the abundance estimates (the coefficients in regression models). However, we have to note that assumptions, which are not always practical, are needed for uncertainty estimation. This explains why hypothesis tests about the same population abundance levels can give different p-values when they use different assumptions.

Remark 6. There have been many efforts to quantify transcripts for better accuracy based on multiple RNA-seq samples (especially biological replicates), thanks to reduced sequencing costs and the rapid accumulation of publicly available RNA-seq samples. Model-based methods include CLIIQ [76], MITIE [77], FlipFlop [78], and MSIQ [63]. These methods generalize the models designed for isoform quantification based on a single sample, and their results show that aggregating the information from multiple samples can achieve better accuracy in isoform abundance estimation. It has been noted in MSIQ [63] that it is important to consider the possible heterogeneity in the quality of different samples to obtain robust and accurate estimation results.

Remark 7. Current statistical methods differ in their perspectives to formulate the isoform quantification problem, the trade-offs between the complexity and flexibility of models, and the approaches to adjust for various sources of sequencing biases and errors. Because of the complexity of transcript-level analysis and the noise and biases in RNA-seq samples, it is impossible to identify a method that has superior performance on all real datasets. We suggest that users consider their preferences on the precision and recall rates in isoform discovery problems, and evaluate the assumptions of different methods for RNA-seq read generation and bias correction, before selecting the appropriate computatinonal method. For a computational comparison of mutiple methods mentioned above, please refer to Refs. [6,79].

EXON-LEVEL ANALYSIS: EXON INCLUSION RATES IN ALTERNATIVE SPLICING

Since transcript-level analysis of complex genes in eukaryotic organisms remains a great challenge [79], there are approaches focusing on exon-level signals, seeking to study alternative splicing based on exons and exon-exon junctions instead of full-length transcripts. When transcriptomic studies focus on the exon-level, a primary step is usually to estimate the *percentage spliced in* (PSI or Ψ, [64]) of an exon of interest. Our discussion below applies to an individual exon. Considering two isoforms, one includes the exon and the other skips the exon, the goal of model-based methods is to estimate

$$\Psi = \text{exon's inclusion rate} = \frac{\text{fraction of the inclusion isoform}}{\text{fraction of the inclusion isoform+fraction of the exclusion isoform}} = \frac{\mathbb{P}(\text{a random read is from the inclusion isoform})}{\mathbb{P}(\text{a random read is from the inclusion isoform}) + \mathbb{P}(\text{a random read is from the exclusion isoform})}. \quad (9)$$

A direct estimator of PSI is

$$\widehat{\Psi} = \frac{\frac{C_I}{L_I}}{\frac{C_I}{L_I} + \frac{C_E}{L_E}},$$

where C_I denotes the number of reads supporting the inclusion isoform (*i.e.*, reads spanning the upstream splicing junction, the exon of interest, and the downstream splicing junction), and C_E denotes the number of reads supporting the exclusion isoform (*i.e.*, reads spanning parts of the upstream and downstream exons but skipping the exon of interest). L_I and L_E denote the lengths or the adjusted lengths (after accounting for constraints on read and isoform lengths, *i.e.*, isoform lengths–read length) of the inclusion and exclusion isoforms, respeetively.

To evaluate the estimation uncertainty, methods including MISO [64], SpliceTrap [80], and rMATS [81] use different statistical models. Both MISO and Splice-Trap construct models similar to the model (Equation (6)) under the Bayesian framework, with Ψ as the parameter of interest. Bayesian confidence intervals of Ψ can then be obtained based on its posterior distribution. rMATS integrates the information from multiple replicates through the following hierarchical model

$$\begin{aligned} C_{Ik} \mid \Psi_k &\sim \text{Binomial}\left(n = C_{Ik} + C_{Ek},\ p = f(\Psi_k)\right), \\ \text{logit}(\Psi_k) &\sim \text{Normal}\left(\mu = \text{logit}(\Psi),\ \sigma^2\right), \end{aligned} \quad (10)$$

where C_{Ik} (C_{Ek}) is the number of reads supporting inclusion (exclusion) isoform in replicate k ($k = 1, 2, \ldots, K$); Ψ_k is the PSI of the exon of interest in replicate k; Ψ and σ^2 are the mean and variance of PSI in the biological condition of interest; f is a function to normalize Ψ_k based on the effective length of the exon. Since both MISO and rMATS can estimate Ψ_k and the uncertainty of $\widehat{\Psi}_k$, it follows that they can detect differential exon usage between two biological conditions through statistical testing.

Remark 8. The above discussion mainly focuses on the scenario where only two alternative isoforms are involved, and does not extend easily to more complex alternative splicing

patterns with more than two alternative splice forms. A proposed remedy is DiffSplice[82], which identifies alternative splicing modules (ASMs) from the splice graph to study splicing patterns that may involve multiple exons. However, one limitation of DiffSplice is that it does not address the estimation uncertainty of the expression levels of ASMs. DEXSeq[83] is another method that studies differential exon usage, but it focuses more on exon-level expression and less on splice junctions.

Remark 9. There is a trade-off in alternative splicing studies concerning whether to use transcript-level or exon-level information. Full-length transcripts provide global information on splicing patterns that directly lead to knowledge on protein isoforms, but accurate quantification of transcripts suffer from the limited information in short RNA-seq reads. On the other hand, exon-level analysis results in the more accurate quantification of individual splicing events, but limits the scope of studies to local genomic regions. As mentioned in the Section of Transcript-level Analysis: Transcript Reconstruction And Quantification, the accumulation of multiple RNA-seq samples and the increasingly large databases of annotated transcripts [84] might provide a solution to this dilemma: combining information from multiple samples with prior knowledge on transcripts to assist the reconstruction and quantification of full-length isoforms from short RNA-seq reads.

OUTLOOK

RNA-seq has become the standard experimental method for transcriptome profiling, and its application to numerous biological studies have led to new scientific discoveries in various biomedical fields. We have summarized the key statistical considerations and methods involved in sample-level, gene-level, transcript-level, and exon-level RNA-seq analyses. Despite the fact that continuous efforts on the development of new tools have improved the accuracy of analyses at all levels, challenges posted by relatively short RNA-seq reads remain in studying full-length transcripts, making it difficult to fully understand the dynamics of mRNA isoforms and their protein products. In complex transcriptomes, probabilistic models have limited power in distinguishing different but highly similar transcripts. It has been noted that identification of all constituent exons of a gene is not always successful, and in cases where these exons are correctly reported, it is challenging to assemble them into complete transcripts with high accuracy [79]. Given the current read lengths in NGS, we emphasize the importance of jointly using multiple samples (i.e., technical or biological replicates) to aggregating information on alternative splicing and sequencing noise. Naïve pooling or averaging methods have been shown inadequate in the multiple-sample analysis [63], and statistical discussion on this topic is still insufficient. On the other hand, new sequencing technologies such as PacBio [85] and Nanopore [86,87] sequencing technologies can produce longer reads with average lengths of 2–3 kb[88]. A primary barrier of the current long-read sequencing technologies is their relatively high error rates and sequencing costs [89]. One current approach to take advantage of these new technologies is to combine the information in next-generation short reads and third-generation long reads in isoform analysis [88].

To demonstrate the efficiency of statistical methods developed for RNA-seq data, method developers must show the reproducibility and interpretability of these methods. As we have

discussed in *Remarks* 2 and 7, there is hardly a method that is superior in every application. However, a useful method should at least demonstrate its advantages under specific assumptions or on a particular type of datasets. Meanwhile, no matter how complicated a statistical model is, its general framework and logical reasoning should be interpretable to users (*e.g.*, biologists). Also, comparison of different methods on benchmark data can be beneficial for the development of new methods. Experimentally validated benchmark data for RNA-seq experiments are still limited on the genome-wide scale.

Aside from the analysis tasks introduced and discussed in this review article, RNA-seq is also widely applied to other research problems like RNA-editing analysis [90,91], non-coding RNA discovery and characterization [92,93], expression quantitative trait loci (eQTL) mapping [94], and prediction of disease progression [95], with interesting statistical questions involved. Transcriptomic data can also be integrated with genomic and epigenomic data to advance our understanding of gene regulation and other biological processes [96]. In recent years, the emerging single-cell RNA sequencing (scRNA-seq) technologies enable the investigation of transcriptomic landscapes at the single-cell resolution, bringing RNA-seq analyses to a new stage [97]. In contrast to scRNA-seq data, the RNA-seq data we have reviewed in this article are now referred to as bulk RNA-seq data, where the data are generated from RNA molecules in multiple cells in a batch. The analysis of scRNA-seq data is complicated by excess zero counts, the so-called dropouts due to the low amounts of mRNA sequenced within individual cells. Therefore, current usage of scRNA-seq data focuses on gene-level analysis, and frequently discussed statistical topics include clustering [98], dimension reduction [99], and imputation [100]. Since the signal-to-noise ratio in scRNA-seq data is much lower than that in bulk RNA-seq, many models developed for bulk RNA-seq data cannot be directly applied to scRNA-seq data, calling for the development of new computational and statistical tools. With the ongoing efforts to build the Human Cell Atlas [101], new scRNA-seq and other single-cell level data (*e.g.*, imaging data) will help researchers more thoroughly understand human cell types and their molecular mechanisms. People can also refer to The Human Cell Atlas White Paper for the detailed discussion of statistical challenges in analyzing these data [102].

ACKNOWLEDGEMENTS

This work was supported by the following grants: National Science Foundation DMS-1613338, NIH/NIGMS R01GM120507, PhRMA Foundation Research Starter Grant in Informatics, Johnson & Johnson WiSTEM2D Award, and Sloan Research Fellowship (to J.J.L) and the UCLA Dissertation Year Fellowship (to W.V.L). The authors would like to thank the insightful feedbacks from Dr. Lior Pachter at California Institute of Technology and Dr. Michael I. Love at University of North Carolina at Chapel Hill.

REFERENCES

1. Wang Z, Gerstein M and Snyder M (2009) RNA-seq: a revolutionary tool for transcriptomics. Nat. Rev. Genet, 10, 57–63 [PubMed: 19015660]
2. Zhao S, Fung-Leung W-P, Bittner A, Ngo K and Liu X (2014) Comparison of RNA-seq and microarray in transcriptome profiling of activated t cells. PLoS One, 9, e78644 [PubMed: 24454679]
3. Engström PG, Steijger T, Sipos B, Grant GR, Kahles A, The RGASP Consortium, Rätsch G, Goldman N, Hubbard TJ, Harrow J, et al. (2013) Systematic evaluation of spliced alignment programs for RNA-seq data. Nat. Methods, 10, 1185–1191 [PubMed: 24185836]

4. Soneson C and Delorenzi M (2013) A comparison of methods for differential expression analysis of RNA-seq data. BMC Bioinformatics, 14, 91 [PubMed: 23497356]
5. Giorgi FM, Del Fabbro C and Licausi F (2013) Comparative study of RNA-seq- and microarray-derived coexpression networks in *Arabidopsis thaliana*. Bioinformatics, 29, 717–724 [PubMed: 23376351]
6. Kanitz A, Gypas F, Gruber AJ, Gruber AR, Martin G and Zavolan M (2015) Comparative assessment of methods for the computational inference of transcript isoform abundance from RNA-seq data. Genome Biol, 16, 1–26 [PubMed: 25583448]
7. Tourasse NJ, Millet JRM, and Dupuy D (2017) Quantitative RNA-seq meta-analysis of alternative exon usage in *C. elegans*. Genome Res, 27, 2120–2128 [PubMed: 29089372]
8. Li JJ, Huang H, Qian M and Zhang X (2015) Advanced Medical Statistics, 2nd ed., chapter 24, pp. 915–936. World Scientific
9. Seqc/Maqc-Iii Consortium (2014) A comprehensive assessment of RNA-seq accuracy, reproducibility and information content by the Sequencing Quality Control Consortium. Nat. Biotechnol, 32, 903–914 [PubMed: 25150838]
10. Conesa A, Madrigal P, Tarazona S, Gomez-Cabrero D, Cervera A, McPherson A, Szcześniak MW, Gaffney DJ, Elo LL, Zhang X et al. (2016) A survey of best practices for RNA-seq data analysis. Genome Biol, 17, 1 [PubMed: 26753840]
11. Gao R and Li JJ (2017) Correspondence of *D. melanogaster* and *C. elegans* developmental stages revealed by alternative splicing characteristics of conserved exons. BMC Genomics, 18, 234 [PubMed: 28302059]
12. Arbeitman MN, Furlong EEM, Imam F, Johnson E, Null BH, Baker BS, Krasnow MA, Scott MP, Davis RW and White KP (2002) Gene expression during the life cycle of *Drosophila melanogaster*. Science, 297, 2270–2275 [PubMed: 12351791]
13. Necsulea A, Soumillon M, Warnefors M, Liechti A, Daish T, Zeller U, Baker JC, Grützner F and Kaessmann H (2014) The evolution of lncRNA repertoires and expression patterns in tetrapods. Nature, 505, 635–640 [PubMed: 24463510]
14. Li WV, Chen Y and Li JJ (2017) Trom: a testing-based method for finding transcriptomic similarity of biological samples. Stat. Biosci, 9, 105–136 [PubMed: 28781712]
15. de la Fuente A, Bing N, Hoeschele I and Mendes P (2004) Discovery of meaningful associations in genomic data using partial correlation coefficients. Bioinformatics, 20, 3565–3574 [PubMed: 15284096]
16. Wyner AD (1978) A definition of conditional mutual information for arbitrary ensembles. Inf. Control, 38, 51–59
17. Zhao J, Zhou Y, Zhang X and Chen L (2016) Part mutual information for quantifying direct associations in networks. Proc. Natl. Acad. Sci. USA, 113, 5130–5135 [PubMed: 27092000]
18. van der Maaten L and Hinton G (2008) Visualizing data using t-SNE. J. Mach. Learn. Res, 9, 2579–2605
19. Kruskal JB and Wish M (1978) Multidimensional Scaling, volume 11 Sage
20. Evans C, Hardin J and Stoebel DM (2017) Selecting between-sample RNA-seq normalization methods from the perspective of their assumptions. Brief. Bioinform, bbx008
21. Bullard JH, Purdom E, Hansen KD and Dudoit S (2010) Evaluation of statistical methods for normalization and differential expression in mRNA-seq experiments. BMC Bioinformatics, 11, 94 [PubMed: 20167110]
22. Mortazavi A, Williams BA, McCue K, Schaeffer L and Wold B (2008) Mapping and quantifying mammalian transcriptomes by RNA-seq. Nat. Methods, 5, 621–628 [PubMed: 18516045]
23. Trapnell C, Pachter L and Salzberg SL (2009) Tophat: discovering splice junctions with RNA-seq. Bioinformatics, 25, 1105–1111 [PubMed: 19289445]
24. Li B and Dewey CN (2011) RSEM: accurate transcript quantification from RNA-seq data with or without a reference genome. BMC Bioinformatics, 12, 323 [PubMed: 21816040]
25. Wagner GP, Kin K and Lynch VJ (2012) Measurement of mRNA abundance using RNA-seq data: RPKM measure is inconsistent among samples. Theory Biosci, 131, 281–285 [PubMed: 22872506]

26. Dillies M-A, Rau A, Aubert J, Hennequet-Antier C, Jean-mougin M, Servant N, Keime C, Marot G, Castel D, Estelle J, et al. (2013) A comprehensive evaluation of normalization methods for illumina high-throughput RNA sequencing data analysis. Brief. Bioinform, 14, 671–683 [PubMed: 22988256]
27. Bolstad BM, Irizarry RA, Astrand M and Speed TP (2003) A comparison of normalization methods for high density oligonucleotide array data based on variance and bias. Bioinformatics, 19, 185–193 [PubMed: 12538238]
28. Anders S and Huber W (2010) Differential expression analysis for sequence count data. Genome Biol, 11, R106 [PubMed: 20979621]
29. Robinson MD and Oshlack A (2010) A scaling normalization method for differential expression analysis of RNA-seq data. Genome Biol, 11, R25 [PubMed: 20196867]
30. Li J, Witten DM, Johnstone IM and Tibshirani R (2012) Normalization, testing, and false discovery rate estimation for RNA-sequencing data. Biostatistics, 13, 523–538 [PubMed: 22003245]
31. Rapaport F, Khanin R, Liang Y, Pirun M, Krek A, Zumbo P, Mason CE, Socci ND and Betel D (2013) Comprehensive evaluation of differential gene expression analysis methods for RNA-seq data. Genome Biol, 14, 3158
32. Bloom JS, Khan Z, Kruglyak L, Singh M and Caudy AA (2009) Measuring differential gene expression by short read sequencing: quantitative comparison to 2-channel gene expression microarrays. BMC Genomics, 10, 221 [PubMed: 19435513]
33. Robinson MD, McCarthy DJ and Smyth GK (2010) edger: a bioconductor package for differential expression analysis of digital gene expression data. Bioinformatics, 26, 139–140 [PubMed: 19910308]
34. Hardcastle TJ and Kelly KA (2010) baySeq: Empirical Bayesian methods for identifying differential expression in sequence count data. BMC Bioinformatics, 11, 422 [PubMed: 20698981]
35. Love MI, Huber W and Anders S (2014) Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. Genome Biol, 15, 550 [PubMed: 25516281]
36. Yu D, Huber W and Vitek O (2013) Shrinkage estimation of dispersion in negative binomial models for RNA-seq experiments with small sample size. Bioinformatics, 29, 1275–1282 [PubMed: 23589650]
37. Leng N, Dawson JA, Thomson JA, Ruotti V, Rissman AI, Smits BMG, Haag JD, Gould MN, Stewart RM and Kendziorski C (2013) Ebseq: an empirical Bayes hierarchical model for inference in RNA-seq experiments. Bioinformatics, 29, 1035–1043 [PubMed: 23428641]
38. Van De Wiel MA, Leday GGR, Pardo L, Rue H, Van Der Vaart AW and Van Wieringen WN (2013) Bayesian analysis of RNA sequencing data by estimating multiple shrinkage priors. Biostatistics, 14, 113–128 [PubMed: 22988280]
39. Law CW, Chen Y, Shi W and Smyth GK (2014) voom: precision weights unlock linear model analysis tools for RNA-seq read counts. Genome Biol, 15, R29 [PubMed: 24485249]
40. Smyth GK (2005) Limma: linear models for microarray data In Bioinformatics and Computational Biology Solutions Using R and Bioconductor, pp. 397–420. Springer
41. Pimentel H, Bray NL, Puente S, Melsted P and Pachter L (2017) Differential analysis of RNA-seq incorporating quantification uncertainty. Nat. Methods, 14, 687–690 [PubMed: 28581496]
42. Schurch NJ, Schofield P, Gierliński M, Cole C, Sherstnev A, Singh V, Wrobel N, Gharbi K, Simpson GG, Owen-Hughes T, et al. (2016) How many biological replicates are needed in an RNA-seq experiment and which differential expression tool should you use? RNA, 22, 839–851 [PubMed: 27022035]
43. Neyman J and Pearson ES (1928) On the use and interpretation of certain test criteria for purposes of statistical inference: Part I. Biometrika, 20, 175–240
44. Holm S (1979) A simple sequentially rejective multiple test procedure. Scand. J. Stat, 6, 65–70
45. Benjamini Y and Hochberg Y (1995) Controlling the false discovery rate: a practical and powerful approach to multiple testing. J. R. Stat. Soc. B, 57, 289–300
46. Nueda MJ, Martorell-Marugan J, Martí C, Tarazona S and Conesa A (2018) Identification and visualization of differential isoform expression in RNA-seq time series. Bioinformatics, 34, 524–526 [PubMed: 28968682]

47. Tai YC and Speed TP (2006) A multivariate empirical Bayes statistic for replicated microarray time course data. Ann. Stat, 34, 2387–2412
48. Stuart JM, Segal E, Koller Dand Kim SK (2003) A gene-coexpression network for global discovery of conserved genetic modules. Science, 302, 249–255 [PubMed: 12934013]
49. Langfelder P and Horvath S (2008) WGCNA: an R package for weighted correlation network analysis. BMC Bioinformatics, 9, 559 [PubMed: 19114008]
50. Zhang B and Horvath S (2005) A general framework for weighted gene co-expression network analysis. Stat. Appl. Genet. Mol. Biol, 4, Article 17
51. Ravasz E, Somera AL, Mongru DA, Oltvai ZN and Barabási A-L (2002) Hierarchical organization of modularity in metabolic networks. Science, 297, 1551–1555 [PubMed: 12202830]
52. Oti M, van Reeuwijk J, Huynen MA and Brunner HG (2008) Conserved co-expression for candidate disease gene prioritization. BMC Bioinformatics, 9, 208 [PubMed: 18433471]
53. Segal E, Shapira M, Regev A, Pe'er D, Botstein D, Koller D and Friedman N (2003) Module networks: identifying regulatory modules and their condition-specific regulators from gene expression data. Nat. Genet, 34, 166–176 [PubMed: 12740579]
54. Canzar S, Andreotti S, Weese D, Reinert K and Klau GW (2016) CIDANE: comprehensive isoform discovery and abundance estimation. Genome Biol, 17, 16 [PubMed: 26831908]
55. Jiang H and Wong WH (2009) Statistical inferences for isoform expression in RNA-seq. Bioinformatics, 25, 1026–1032 [PubMed: 19244387]
56. Trapnell C, Williams BA, Pertea G, Mortazavi A, Kwan G, van Baren MJ, Salzberg SL, Wold BJ and Pachter L (2010) Transcript assembly and quantification by RNA-seq reveals unannotated transcripts and isoform switching during cell differentiation. Nat. Biotechnol, 28, 511–515 [PubMed: 20436464]
57. Roberts A and Pachter L (2013) Streaming fragment assignment for real-time analysis of sequencing experiments. Nat. Methods, 10, 71–73 [PubMed: 23160280]
58. Bray NL, Pimentel H, Melsted P and Pachter L (2016) Near-optimal probabilistic RNA-seq quantification. Nat. Biotechnol, 34, 525–527 [PubMed: 27043002]
59. Dempster AP, Laird NM and Rubin DB (1977) Maximum likelihood from incomplete data via the EM algorithm. J. R. Stat. Soc. B, 39, 1–38
60. Zhang J, Jay Kuo C-C and Chen L (2014) WEMIQ: an accurate and robust isoform quantification method for RNA-seq data. Bioinformatics, 31, 878–885 [PubMed: 25406327]
61. Patro R, Duggal G, Love MI, Irizarry RA and Kingsford C (2017) Salmon provides fast and bias-aware quantification of transcript expression. Nat. Methods, 14, 417–419 [PubMed: 28263959]
62. Mezlini AM, Smith EJM, Fiume M, Buske O, Savich GL, Shah S, Aparicio S, Chiang DY, Goldenberg A and Brudno M (2013) iReckon: simultaneous isoform discovery and abundance estimation from RNA-seq data. Genome Res, 23, 519–529 [PubMed: 23204306]
63. Li WV, Zhao A, Zhang S and Li JJ (2017) Msiq: joint modeling of multiple RNA-seq samples for accurate isoform quantification. Ann. Appl. Stat, 12, 510–539
64. Katz Y and Eric T (2010) Analysis and design of RNA sequencing experiments for identifying isoform regulation. Nat. Methods, 7, 1009–1015 [PubMed: 21057496]
65. Love MI, Hogenesch JB and Irizarry RA (2016) Modeling of RNA-seq fragment sequence bias reduces systematic errors in transcript abundance estimation. Nat. Biotechnol, 34, 1287–1291 [PubMed: 27669167]
66. Roberts A, Trapnell C, Donaghey J, Rinn JL and Pachter L (2011) Improving RNA-seq expression estimates by correcting for fragment bias. Genome Biol, 12, R22 [PubMed: 21410973]
67. Xia Z, Wen J, Chang C-C and Zhou X (2011) Nsmap: a method for spliced isoforms identification and quantification from RNA-seq. BMC Bioinformatics, 12, 162 [PubMed: 21575225]
68. Bohnert R and Rätsch G (2010) rQuant. web: a tool for RNA-seq-based transcript quantitation. Nucleic Acids Res, 38, W348–W351 [PubMed: 20551130]
69. Li JJ, Jiang C-R, Brown JB, Huang H and Bickel PJ (2011) Sparse linear modeling of next-generation mRNA sequencing (RNA-seq) data for isoform discovery and abundance estimation. Proc. Natl. Acad. Sci. USA, 108, 19867–19872 [PubMed: 22135461]

70. Li W, Feng J and Jiang T (2011) IsoLasso: a LASSO regression approach to RNA-seq based transcriptome assembly. J. Comput. Biol, 18, 1693–1707 [PubMed: 21951053]
71. Meinshausen N and Bühlmann P (2010) Stability selection. J.R. Stat. Soc. Series B Stat. Methodol, 72, 417–473
72. Grabherr MG, Haas BJ, Yassour M, Levin JZ, Thompson DA, Amit I, Adiconis X, Fan L, Raychowdhury R, Zeng Q, et al. (2011) Full-length transcriptome assembly from RNA-seq data without a reference genome. Nat. Biotechnol, 29, 644–652 [PubMed: 21572440]
73. Guttman M, Garber M, Levin JZ, Donaghey J, Robinson J, Adiconis X, Fan L, Koziol MJ, Gnirke A, Nusbaum C, et al. (2010) Ab initio reconstruction of cell type-specific transcriptomes in mouse reveals the conserved multi-exonic structure of lincrnas. Nat. Biotechnol, 28, 503–510 [PubMed: 20436462]
74. Pertea M, Pertea GM, Antonescu CM, Chang T-C, Mendell JT and Salzberg SL (2015) Stringtie enables improved reconstruction of a transcrip-tome from RNA-seq reads. Nat. Biotechnol, 33, 290–295 [PubMed: 25690850]
75. Wang X, Wu Z and Zhang X (2010) Isoform abundance inference provides a more accurate estimation of gene expression levels in RNA-seq. J. Bioinform. Comput. Biol, 8 (Supp. 1), 177–192 [PubMed: 21155027]
76. Lin Y-Y, Dao P, Hach F, Bakhshi M, Mo F, Lapuk A, Collins C and Cenk Sahinalp S (2012) Cliiq: accurate comparative detection and quantification of expressed isoforms in a population In Algorithms in Bioinformatics, pp. 178–189. Springer
77. Behr J, Kahles A, Zhong Y, Sreedharan VT, Drewe P and Rätsch G (2013) MITIE: Simultaneous RNA-seq-based transcript identification and quantification in multiple samples. Bioinformatics, 29, 2529–2538 [PubMed: 23980025]
78. Bernard E, Jacob L, Mairal J and Vert J-P (2014) Efficient RNA isoform identification and quantification from RNA-seq data with network flows. Bioinformatics, 30, 2447–2455 [PubMed: 24813214]
79. Steijger T, Abril JF, Engström PG, Kokocinski F, Abril JF, Akerman M, Alioto T, Ambrosini G, Antonarakis SE, Behr J, et al. (2013) Assessment of transcript reconstruction methods for RNA-seq. Nat. Methods, 10, 1177–1184 [PubMed: 24185837]
80. Wu J, Akerman M, Sun S, McCombie WR, Krainer AR and Zhang MQ (2011) Splicetrap: a method to quantify alternative splicing under single cellular conditions. Bioinformatics, 27, 3010–3016 [PubMed: 21896509]
81. Shen S, Park JW, Lu Z, Lin L, Henry MD, Wu YN, Zhou Q and Xing Y (2014) rMATS: robust and flexible detection of differential alternative splicing from replicate RNA-seq data. Proc. Natl. Acad. Sci. USA, 111, E5593–E5601 [PubMed: 25480548]
82. Hu Y, Huang Y, Du Y, Orellana CF, Singh D, Johnson AR, Monroy A, Kuan P-F, Hammond SM, Makowski L, et al. (2013) Diffsplice: the genome-wide detection of differential splicing events with RNA-seq. Nucleic Acids Res, 41, e39–e39 [PubMed: 23155066]
83. Anders S, Reyes A and Huber W (2012) Detecting differential usage of exons from RNA-seq data. Genome Res, 22, 2008–2017 [PubMed: 22722343]
84. Harrow J, Frankish A, Gonzalez JM, Tapanari E, Diekhans M, Kokocinski F, Aken BL, Barrell D, Zadissa A, Searle S, et al. (2012) GENCODE: the reference human genome annotation for the ENCODE project. Genome Res, 22, 1760–1774 [PubMed: 22955987]
85. Rhoads A and Au KF (2015) Pacbio sequencing and its applications. Genom. Proteom. Bioinf, 13, 278–289
86. Branton D, Deamer DW, Marziali A, Bayley H, Benner SA, Butler T, Di Ventra M, Garaj S, Hibbs A, Huang X, et al. (2008) The potential and challenges of nanopore sequencing. Nat. Biotechnol, 26, 1146–1153 [PubMed: 18846088]
87. Byrne A, Beaudin AE, Olsen HE, Jain M, Cole C, Palmer T, DuBois RM, Forsberg EC, Akeson M and Vollmers C (2017) Nanopore long-read RNA-seq reveals widespread transcriptional variation among the surface receptors of individual B cells. Nat. Commun, 8, 16027 [PubMed: 28722025]
88. Au KF, Sebastiano V, Afshar PT, Durruthy JD and Lee L Williams BA, van Bakel H, Schadt EE, Reijo-Pera RA, Underwood JG, et al. (2013) Characterization of the human ESC transcriptome by hybrid sequencing. Proc. Natl. Acad. Sci. USA, 110, E4821–E4830 [PubMed: 24282307]

89. Bleidorn C (2016) Third generation sequencing: technology and its potential impact on evolutionary biodiversity research. Syst. Biodivers, 14, 1–8

90. Ramaswami G, Lin W, Piskol R, Tan MH, Davis C and Li JB (2012) Accurate identification of human *Alu* and non-*Alu* RNA editing sites. Nat. Methods, 9, 579–581 [PubMed: 22484847]

91. Bahn JH, Lee J-H, Li G, Greer C, Peng G and Xiao X (2012) Accurate identification of A-to-I RNA editing in human by transcriptome sequencing. Genome Res, 22, 142–150 [PubMed: 21960545]

92. Iyer MK, Niknafs YS, Malik R, Singhal U, Sahu A, Hosono Y, Barrette TR, Prensner JR, Evans JR, Zhao S, et al. (2015) The landscape of long noncoding RNAs in the human transcriptome. Nat. Genet, 47, 199–208 [PubMed: 25599403]

93. Hezroni H, Koppstein D, Schwartz MG, Avrutin A, Bartel DP and Ulitsky I (2015) Principles of long noncoding RNA evolution derived from direct comparison of transcriptomes in 17 species. Cell Reports, 11, 1110–1122 [PubMed: 25959816]

94. Pickrell JK, Marioni JC, Pai AA, Degner JF, Engelhardt BE, Nkadori E, Veyrieras J-B, Stephens M, Gilad Y and Pritchard JK (2010) Understanding mechanisms underlying human gene expression variation with RNA sequencing. Nature, 464, 768–772. [PubMed: 20220758]

95. Zak DE, Penn-Nicholson A, Scriba TJ, Thompson E, Suliman S, Amon LM, Mahomed H, Erasmus M, Whatney W, Hussey GD, et al. (2016) A blood RNA signature for tuberculosis disease risk: a prospective cohort study. Lancet, 387, 2312–2322 [PubMed: 27017310]

96. Hawkins RD, Hon GC and Ren B (2010) Next-generation genomics: an integrative approach. Nat. Rev. Genet, 11, 476–486 [PubMed: 20531367]

97. Kolodziejczyk AA, Kim JK, Svensson V, Marioni JC and Teichmann SA (2015) The technology and biology of single-cell RNA sequencing. Mol. Cell, 58, 610–620 [PubMed: 26000846]

98. Xu C and Su Z (2015) Identification of cell types from single-cell transcriptomes using a novel clustering method. Bioinformatics, 31, 1974–1980 [PubMed: 25805722]

99. Pierson E and Yau C (2015) Zifa: dimensionality reduction for zero-inflated single-cell gene expression analysis. Genome Biol, 16, 241 [PubMed: 26527291]

100. Li WV and Li JJ (2018) An accurate and robust imputation method scimpute for single-cell RNA-seq data. Nat. Commun, 9, 997 [PubMed: 29520097]

101. Regev A, Teichmann SA, Lander ES, Amit I, Benoist C, Birney E, Bodenmiller B, Campbell P, Carninci P, Clatworthy M, et al. (2017) The human cell atlas. eLife, 6, e27041 [PubMed: 29206104]

102. The Human Cell Atlas Consortium. (2017) The human cell atlas white paper

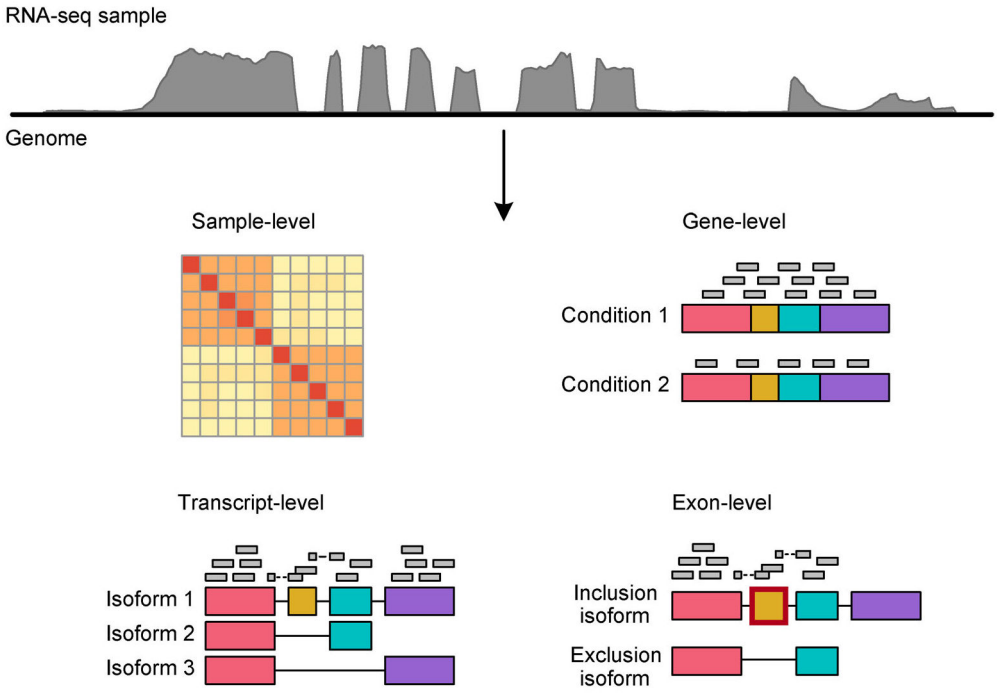

Figure 1. RNA-seq analyses at four different levels: sample-level, gene-level, transcript-level, and exon-level.

In the sample-level analysis, the results are usually summarized into a similarity matrix, as introduced in the Section of Sample-level Analysis: Transcriptome Similarity. Taking a 4-exon gene as an example, the gene-level analysis summarizes the counts of RNA-seq reads mapped to genes in samples of different conditions, and it subsequently compares genes' expression levels calculated based on read counts; the transcript-level analysis focuses on reads mapped to different isoforms; the exon-level analysis mostly considers the reads mapped to or skipping the exon of interest (the yellow exon marked by a red box in this example).

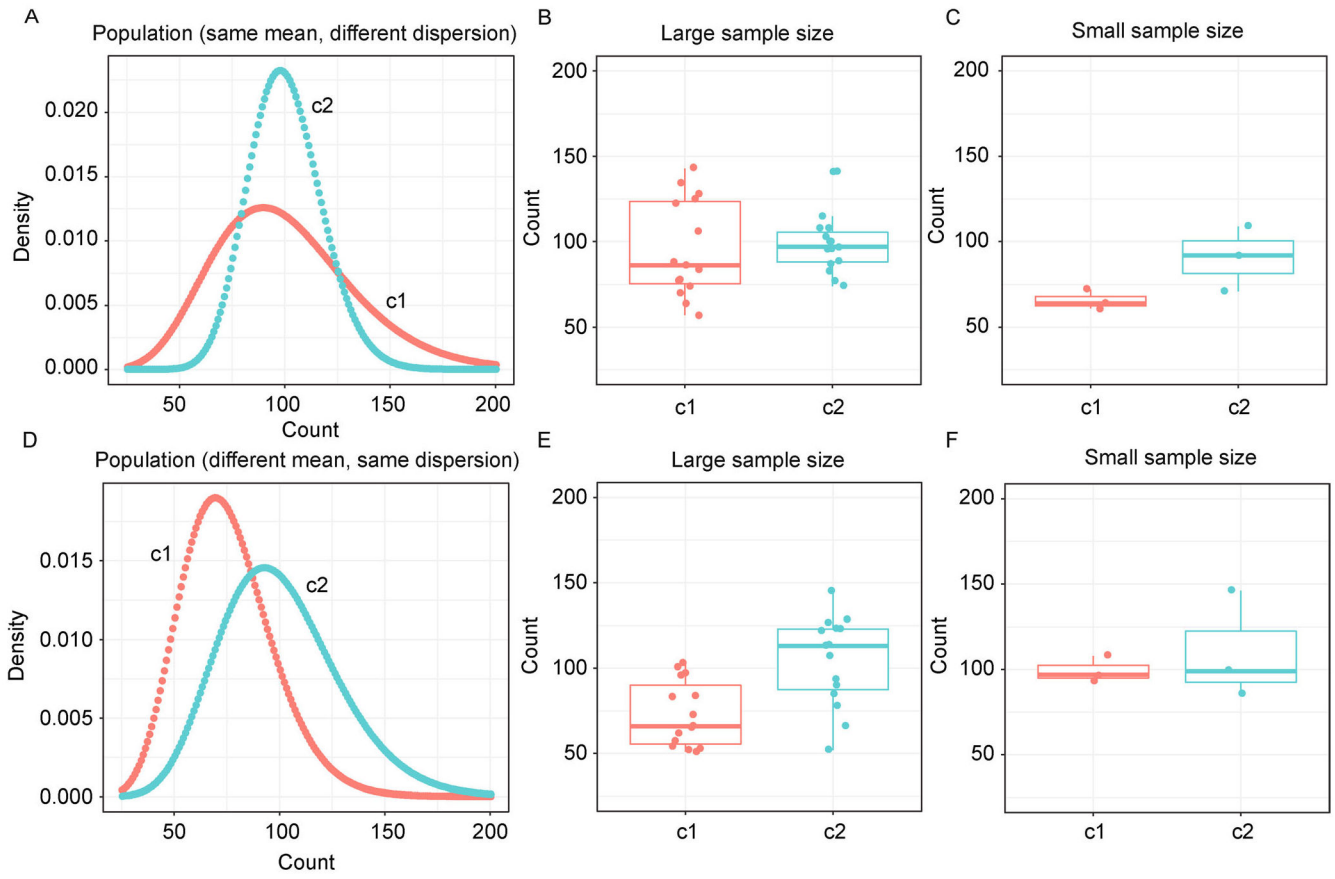

Figure 2. Illustration of read counts as samples drawn from unobservable populations.

(A) The population read count distribution of a hypothetical gene 1 under conditions c1 and c2, based on the NB model. The two population distributions have the same mean parameter but different dispersion parameters. (B and C) The observed read counts of gene 1 are independent samples (B: with large sample sizes; C: with small sample sizes) drawn from the two unobservable distributions. When the sample size is small, a statistical test about whether the two samples have the same population mean will possibly lead to a false positive result (gene 1 found as a DE gene). (D) The population read count distribution of a hypothetical gene 2 under conditions c1 and c2, based on the NB model. The two population distributions have different mean parameters but the same dispersion parameter. (E and F) The observed read counts of gene 2 are independent samples (E: with large sample sizes; F: with small sample sizes) drawn from the two unobservable distributions. When the sample size is small, a statistical test about whether the two samples have the same population mean will possibly lead to a false negative result (gene 2 found as a non-DE gene).

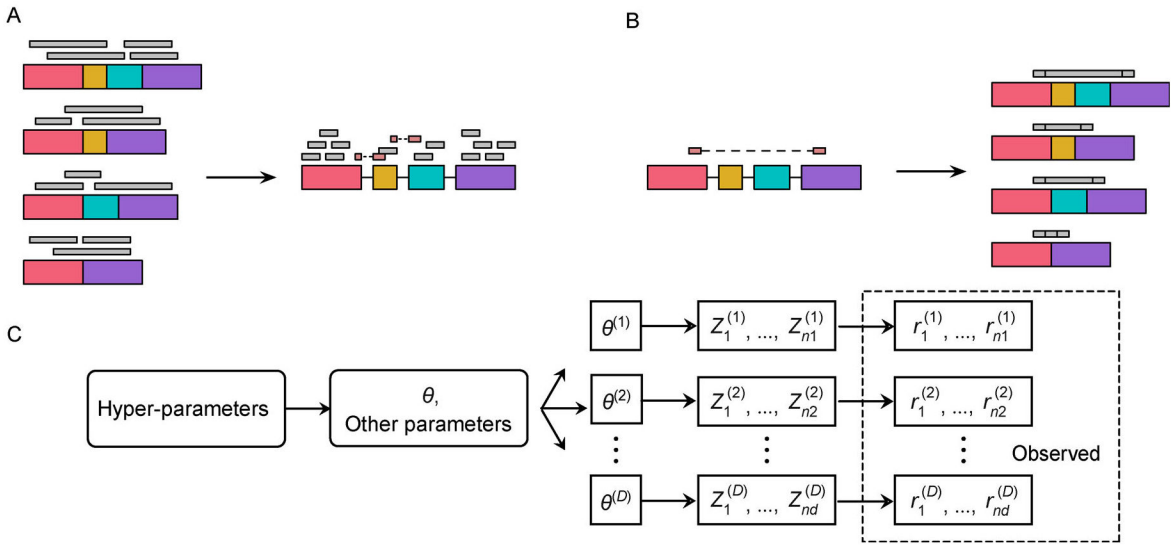

Figure 3. Illustration of data and modeling issues in transcript-level RNA-seq analysis.

(A) Taken this 4-exon gene as an example, the observed RNA-seq reads are sequenced from fragments of the true but unobservable isoforms. The read length is fixed in each experiment, but the fragment lengths can vary. Since only the two ends of each fragment are sequenced as paired-end reads, this leads to information loss in RNA-seq experiments. (B) Given the paired-end reads mapped to the 4-exon gene (one end mapped to the first exon and the other end mapped to the fourth exon), the inferred fragment length could be different when assuming different isoform origins of the read. (C) An example Bayesian framework to estimate the population isoform proportions θ of a gene given D samples. $\theta^{(1)}$, …, $\theta^{(D)}$ are considered as the realization of θ in D samples. $Z_i^{(d)}$ denotes the isoform origin of read $r_i^{(d)}$ in sample d. Only the reads $r_i^{(d)}$'s are observed information, other random variables are hidden, and parameters need estimation.