Chapter 2
Using RNA-seq Data to Detect Differentially Expressed Genes

Douglas J. Lorenz, Ryan S. Gill, Ritendranath Mitra, and Susmita Datta

Abstract RNA-sequencing (RNA-seq) technology has become a major choice in detecting differentially expressed genes across different biological conditions. Although microarray technology is used for the same purpose, statistical methods available for identifying differential expression for microarray data are generally not readily applicable to the analysis of RNA-seq data, as RNA-seq data comprise discrete counts of reads mapped to particular genes. In this chapter, we review statistical methods uniquely developed for detecting differential expression among different populations of RNA-seq data as well as techniques designed originally for the analysis of microarray data that have been modified for the analysis of RNA-seq data. We include a very brief description of the normalization of RNA-seq data and then elaborate on parametric and nonparametric testing procedures, as well as empirical and fully Bayesian methods. We include a brief review of software available for the analysis of differential expression and summarize the results of a recent comprehensive simulation study comparing existing methods.

2.1 Introduction: RNA-seq Data

RNA-seq is a next generation sequencing (NGS) procedure of the entire transcriptome by which one can measure the expression of several features such as gene expression, allelic expression, and intragenic expression. The number of reads mapped to a given gene or transcript is considered to be the estimate

D.J. Lorenz • R. Mitra • S. Datta (✉)
Department of Bioinformatics and Biostatistics, School of Public Health and Information Science, University of Louisville, 485 E. Gray St., Louisville, KY 40205, USA
e-mail: susmita.datta@louisville.edu

R.S. Gill
Department of Mathematics, University of Louisville, Louisville, KY 40292, USA

S. Datta and D. Nettleton (eds.), *Statistical Analysis of Next Generation Sequencing Data*, Frontiers in Probability and the Statistical Sciences, DOI 10.1007/978-3-319-07212-8_2,

Table 2.1 Table of read counts from a hypothetical RNA-seq experiment

	Population 1			Population 2	
Gene	Sample 1	Sample 2	Sample 3	Sample 1	Sample 2
1	22	26	15	66	44
2	4	1	20	1	4
3	75	113	281	116	97
⋮	⋮	⋮	⋮	⋮	⋮
10,000	0	9	0	1	2
Total	824,015	782,345	1,345,387	693,428	923,450

In this example, there are $K = 2$ populations, $J_1 = 3$ samples in the first population, $J_2 = 2$ samples in the second population, and $G = 10,000$ genes. The final row lists the cumulative read counts for each sample, frequently referred to as the library size for a sample

of the expression level of that feature using this technology [24]. Microarray technology has been the method of choice to measure gene expression since the nineteen-nineties. However, RNA-seq is generally acknowledged to be a better platform for transcription profiling for several reasons [8, 22, 25, 26, 28, 43, 50]. RNA-seq is believed to have a wider range of signal detection. The resolution of microarray expression measures cannot go beyond the probe level. In contrast, the majority of the reads from NGS technology map to the reference genome with single base resolution and consequently RNA-seq can be evaluated at single-base resolution. Moreover, in microarray technology one needs to have knowledge of the target sequences to construct the probe sets. Hence, RNA-seq is more suitable for the discovery of novel transcripts.

The end-product of a RNA-seq experiment is a sequence of read counts, typically represented as a matrix with rows representing genes and columns representing samples from one or more populations, as in Table 2.1. When RNA-seq data are generated from two or more populations, interest often is in the detection of differentially expressed genes among the populations, i.e., genes for which read count distributions differ among populations. Methods for detecting differential expression in microarray data are well-established but generally not applicable to RNA-seq data, as the data from a RNA-seq experiment are discrete counts rather than continuous measures of expression levels.

A challenge in the detection of differential expression for RNA-seq data results from the way in which reads are mapped to features such as genes, transcripts or exons. One of the issues is that the expression quantification from short reads using RNA-seq data depends on the length of the features; longer features usually produce more reads. Normalization by dividing by the length of the transcript [25] alleviates this problems somewhat but not completely [54]. The expression value used by Mortazavi et al. [25] is referred to as Reads per Kilobase per Million reads (RPKM). Differential RNA-seq analysis using an empirical Bayes procedure by the *limma* method [38] uses log-counts per million (log-cpm), analogous to the log-intensity values in microarray studies.

Differential expression analysis may also be affected by the sequence depth of the NGS data generation. Sequence depth can be calculated as $N \times L/G$, where N is the number of reads, L is the average read length and G is the length of the original genome. For example, $N = 8$ reads for a genome with $G = 2,000$ base pairs at average length $L = 500$ will have a sequence depth of 2 or 2× redundancy. This also is equivalent to the percentage of genome covered by reads and the average number of times a base is read. Higher coverage can improve the power to identify differential expression using RNA-seq data. However, read counts are subject to technical variation in which the overall read count for a sample, referred to as the library size, can substantially vary among repeated NGS experiments on the same sample. In order to accommodate this source of variability, log-cpm values need to be adjusted by accounting for mean-variance trends typically observed in RNA-seq data, particularly among genes with lower counts. Zero counts are augmented by a small positive value to avoid taking the logarithm of zero, ensuring non-missing log-cpm and reducing the variability at lower count values.

An additional challenge is that some of genes may exhibit very large read counts while the rest of the reads are distributed among the remaining genes. Hence, even if library sizes are identical between samples, some genes may mask the expression of others which may be moderately equivalently expressed. Thus, the expression signals of genes or transcripts in RNA-seq data not only depend on sequence depth, but also are dependent on the expression levels of other transcripts. Because of this and the technical variation of NGS experiments noted above, raw read counts from different populations are not necessarily directly comparable in an analysis of differential expression without adjustment for technical variation. In other words, simply viewing the count for a given gene and sample as proportional to the sample's total read count is problematic because a few genes may have extremely large counts that artificially inflate a sample's total read count. Alternative complex normalization schemes for RNA-seq data have been proposed by Bullard et al. [6], Anders and Huber [1], and Robinson and Oshlack [32]. In these methods, there are additional sample specific normalizations combined with library sizes. There are other methods of normalization as well. A thorough evaluation of many normalization methods for RNA-seq data is provided in Dillies et al. [11]. Trimmed mean of M-values normalization (TMM) [32] and the normalization scheme provided by Anders and Huber [1] are among the easiest to use and provide a decent solution to the normalization problem of RNA-seq data. However, even these methods assume that very few genes are differentially expressed between different populations and those are equivalently spread between the up- and down-regulated genes. Other types of normalization strategies deal with the GC content of the reads. Normalization for this specific reason transforms RNA-seq data in such a way that it no longer remains count data and should be dealt with differently in terms of further analysis for finding differentially expressed genes. *Cufflinks/Cuffdiff* [48] provides a normalization scheme in their integrated differential analysis algorithm. For a more thorough discussion of normalization methods, we encourage the reader to consult the chapter on normalization in this volume.

The focus of this chapter is to provide a comprehensive review of the methods related to the analysis of differential expression for RNA-seq data. In recent years, a number of reviews of RNA-seq data analysis methods have been published, and they all are effective in communicating the current status of the analysis of RNA-seq data [2,40,53]. Much further work will be devoted to developing statistical methods for the detection of differentially expressed genes for RNA-seq data. In this chapter, we review statistical methods for detecting differential expression in RNA-seq data, including the application of techniques for analyzing microarray data to RNA-seq data, parametric and nonparametric tests, and empirical and fully Bayesian methods. We summarize the results of several simulation studies, including a recently published thorough examination of several of these methods. We briefly describe some existing open source R and Bioconductor software for testing differential expression for RNA-seq data. We conclude the chapter with a discussion section.

2.2 Statistical Methods for Testing Differential Expression

For consistency of notation in what follows, we have established a single unifying notation for the RNA-seq read counts. As a result, the notation we use here is frequently different from the source works. We consider read counts for G genes measured in K populations. Let Y_{ijg} denote the number of RNA-seq reads mapped to gene g in replicate j of population i, where $1 \leq i \leq K$, $1 \leq j \leq J_i$, and $1 \leq g \leq G$. We will generally refer to “genes” as that which are being tested for differential expression, with the understanding that other features (transcripts, exome expression, etc.) may be tested as well. While the developments below will focus on detection of differential expression between two populations, several of the methods have natural extensions permitting the comparison of more than two populations.

2.2.1 Simple Approaches

An early treatment [6] of differential expression for RNA-seq data examined the performance of Fisher’s exact test and test statistics derived from generalized linear models used to derive and normalize expression counts. We temporarily extend our notation and let Y_{ijgk} denote the read count for gene g along lane k in sample j of population i. A Poisson generalized linear model for Y_{ijgk} is

$$\log\left(E[Y_{ijgk} \mid d_{ijk}]\right) = \log\left(d_{ijk}\right) + \lambda_{ijg} + \theta_{ijgk}, \tag{2.1}$$

relating the logarithm of the expected read count for gene g in lane k as a linear function of the gene g rate in sample j of population i (λ_{ijg}), an offset term adjusting for variation in lane depths (d_{ijk}), and other unspecified technical effects that vary by gene, lane, and sample (θ_{ijgk}). Tests of differential expression are derived from this model through a likelihood ratio test (LRT) or t-tests of the maximum likelihood estimates (MLE) of the expression parameters λ_{ijg}. The performance of these tests as well as Fisher's exact test in detecting differentially expressed genes was evaluated on a gold standard data set [7]. Two variants of the GLM-derived t-tests—one using the variance of the MLE of λ_{ijg} and one using variance calculated via the delta method—exhibited reduced detection rates. Fisher's exact test and the LRT performed equivalently and exhibited uniformly greater true positive rates (TPR) than the t-tests. The authors noted that screening genes based on read counts improved the performance of both the t-test and LRT. When genes with read counts lower than 20 were filtered out, detection rates for the LRT and t-test greatly improved and were roughly equivalent. The filtering threshold, however, was arbitrarily selected and tested only on the single gold standard data set.

A recently developed R software package, *DEGseq* [51], also employs Fisher's exact test as well as the two versions of the likelihood ratio test noted by Bullard et al. [6]. Additionally, *DEGseq* introduces two tests based on the thresholding of plots of log fold change as a function of mean log expression level (MA plots) commonly used in microarray data, one for analyses based on single samples in each population and one for analyses based on technical replicates. These MA plot-based tests are based upon binomial assumptions for the read counts and a normal approximation of the conditional distribution of the log count ratio between populations (M) and average of log counts (A) between populations.

Another simple two-sample test can be constructed by assuming a Poisson distribution for the read counts. To this end, suppose that the $Y_{ijg} \sim POI(c_{ij}\lambda_{ig})$, where λ_{ig} represents the relative rate parameter for gene g in population i and c_{ij} is a replicate-specific constant. The constant c_{ij} is included to account for variation in read intensity among biological replicates, which can artificially inflate overall library sizes for replicates with high intensity. The within-population and overall read counts are defined as $Y_{i\cdot g} = \sum_j Y_{ijg}$ and $Y_{\cdot\cdot g} = \sum_{i,j} Y_{ijg}$, which follow $POI(\sum_j \lambda_{ig} c_{ij})$ and $POI(\sum_{i,j} \lambda_{ig} c_{ij})$ distributions, respectively, under the Poisson assumption for the individual read counts. The null hypothesis for testing differential expression for gene g is that of equal relative rates of expression, which takes the form $H_{0,g} : \lambda_{1g} = \lambda_{2g}$. Under the null, the conditional distribution of the read count for gene g in population i ($Y_{i\cdot g}$) given the total read count for gene g ($Y_{\cdot\cdot g}$) is binomial with size $Y_{\cdot\cdot g}$ and success probability $\pi_0 = \sum_j c_{ij} / \sum_{i,j} c_{ij}$, which is common to all G genes. The test of $H_{0,g}$ is then any binomial test (e.g. asymptotic, exact, Clopper-Pearson) of $Y_{i\cdot g}$ successes in $Y_{\cdot\cdot g}$ trials against null probability π_0. Adjustment of p-values from the G tests to control the false discovery rate (FDR) can be achieved via the Benjamini–Hochberg [4] correction, or any other suitable method.

Fisher's exact test, GLM-based tests, the MA plot tests of *DEGSeq*, and the conditional binomial test have received little attention, in large part due to the practical infeasibility of assumptions about the marginal or conditional distributions of the read counts. In particular, the Poisson assumption for read count distributions and the binomial assumption for conditional read count distributions have proven infeasible for real data. The variation in replicate samples is typically far greater than that modeled by the Poisson distribution even after adjustment for read intensity. Other tests for differential expression have focused on extensions of the Poisson model for read counts or alternative discrete probability distributions.

2.2.2 Tests Based on Extensions of the Poisson Distribution

Srivastava and Chen [41] proposed a test of differential expression based upon the generalized Poisson distribution. In terms of RNA-seq data, the generalized Poisson model is

$$P(Y_{ijg} = y) = \lambda_{ig}\,(\lambda_{ig} + \theta_{ig} y)^{y-1}\, e^{-\lambda_{ig} - \theta_{ig} y}/y!, \tag{2.2}$$

where λ_{ig} is the read intensity parameter for gene g in population i and θ_{ig} is a parameter referred to by the authors as the average bias caused by the sample preparation and sequencing process. The authors note that the bias parameter θ_{ig} serves as a shrinkage factor relative to the Poisson distribution, as $E[Y_{ijg}] = \lambda_{ig}(1 - \theta_{ig})^{-1}$ and $Var[Y_{ijg}] = \lambda_{ig}(1 - \theta_{ig})^{-3}$. To construct a likelihood ratio test based on the generalized Poisson (GP) model, the intensity and sequencing-bias parameters $(\lambda_{ig}, \theta_{ig})$ are first estimated freely. The intensity parameters are then estimated under the restriction $\lambda_{2g} = w\lambda_{1g}$, where w represents a normalization constant accounting for different sequencing depths between populations. In practice, this normalization constant w is chosen as the ratio of the total amount of sequenced RNA in the two populations. This in turn is estimated in each population as a weighted sum over all genes of the unrestricted MLE of the λ_{ig}, with weights defined by gene lengths. The LRT statistic calculated from the restricted ($\lambda_{i2} = w\lambda_{i1}$) and unrestricted likelihoods approximately follows the χ^2_1 distribution. Using a standard data set [37], the GP test was shown to be more sensitive than the Poisson LRT as well as LRT derived from generalized linear models under Poisson, negative binomial, and quasi-Poisson distributions. The generalized Poisson distribution does permit negative intensities λ_{ig} which are not interpretable in a practical sense. The authors note that the GP test fails when data produce a negative estimate of λ_{ig} as likelihoods become zero and maximum likelihood estimation fails, a notable drawback to the applicability of the GP test.

Auer and Doerge [3] introduced the two-stage Poisson model (TSPM), in which gene counts are first screened for overdispersion and different test statistics are calculated for genes determined to be overdispersed/not overdispersed. In the first stage of TSPM, genes are filtered so that those with small cumulative counts over replicates and populations are not given further consideration. The authors arbitrarily select 10 as the cutoff, but do note that the cutoff can be varied based on the number of replicates and overall read intensity. After filtering, a random effects Poisson model is fitted to the gene counts assuming no overdispersion, and an adjusted score statistic is calculated to test the null hypothesis of no overdispersion per gene, $H_{0g} : \phi_g = 1$, where ϕ_g is the overdispersion parameter for gene g, $1 \leq g \leq G$. The quantiles of the adjusted score statistic are compared to theoretical quantiles from the χ^2_1 distribution. Genes for which the adjusted score statistic is greater than the upper bound of the Working-Hotelling simultaneous confidence band for the theoretical χ^2_1 quantiles are classified as overdispersed. All other genes are classified as not overdispersed. In the second stage of the TSPM, genes classified as overdispersed are tested using a likelihood ratio test derived from fitting overdispersed quasi-likelihood models under the null and alternative hypotheses of no differential expression and differential expression, respectively. Genes classified as not overdispersed in stage 1 are tested using a standard likelihood ratio test from a Poisson model. The authors recommend that corrections for FDR control be applied separately within the sets of genes found to be overdispersed and not overdispersed as a power-saving strategy, diverging from common implementation of methods for FDR control. In a simulation study, the authors show that the TSPM exhibited improved power over a negative binomial model and a quasi-likelihood approach in settings where some genes were overdispersed and others not.

Pounds et al. [30] proposed two procedures for identifying differentially expressed genes using both a likelihood ratio test with a Poisson distribution and a quasi-likelihood model which adjusts for overdispersion. Both procedures are based on the adaptive histogram estimator of empirical Bayesian probabilities of no differential expression and of no overdispersion. The Assumption Adequacy Averaging (AAA) procedure uses the law of total probability to estimate the empirical Bayesian probabilities of no differential expression for each gene. These estimates are based on a weighted average of the empirical Bayesian probabilities of no differential expression for the gene using the Poisson and quasi-likelihood models, with weights based on the empirical Bayesian probability of no overdispersion. The Empirical Best Test (EBT) procedure alternatively selects the best test based on the empirical Bayesian probabilities of no overdispersion for each gene. The EBT procedure then applies the adaptive histogram estimator to obtain the empirical Bayesian probabilities based on the set of p-values for the tests for differential expression, using the best test for each individual gene. The authors present simulation studies which evaluate the performance of these two procedures based on various performance metrics and scenarios, and also compare them to the Poisson model, the quasi-likelihood model, TSPM, and negative binomial and Bayesian tests discussed below. The authors also discuss some nice theoretical properties of the two proposed procedures.

2.2.3 Negative Binomial and Quasi-Likelihood Tests

Rather than extend the Poisson distribution [41] or work around overdispersion via screening [3], several authors have proposed differential expression methods based on the negative binomial distribution. The use of the negative binomial distribution was motivated by observation that real RNA-seq data sets typically exhibited greater variability than could be modeled via the Poisson distribution. Robinson and Smyth [33] assume a negative binomial distribution for the read counts for all genes with a common dispersion parameter, (i.e.) $Y_{ijg} \sim NB(\mu_{ijg}, \phi)$, where $\mu_{ijg} = m_{ij}\lambda_{ig}$, m_{ij} the library size for sample j in population i, and λ_{ig} a relative abundance parameter for gene g in population i, which is assumed to be common to the replicate samples within a population. The dispersion parameter ϕ is estimated by maximizing the conditional likelihood given the sum of the counts in each population. This conditional maximization is straightforward when library sizes are assumed to be equal within each population. When this is not the case, a quantile adjustment is applied to the library sizes, adjusting observed counts to the geometric mean of the replicates. These adjusted library sizes are then used in the maximization of the conditional likelihood for the dispersion parameter, a process referred to as quantile adjusted conditional maximum likelihood (qCML) estimation. The null hypothesis for the test of differential expression is the equality of the relative abundance parameters, $H_{0g} : \lambda_{1g} = \lambda_{2g}, g = 1, \ldots, G$. The authors suggest an exact negative binomial test based on the same quantile adjustment used in estimating the dispersion parameter, in which the “pseudosum” of adjusted counts for a given population is conditioned on the pseudosum of counts across populations, and a p-value calculated as the probability of observing counts greater than those observed.

The assumption of a dispersion parameter ϕ common to all genes is frequently biologically implausible. As such, Robinson and Smyth [34] extended their original negative binomial approach and suggested the use of gene specific dispersion parameters ϕ_g, so that the distributional assumption on read counts becomes $Y_{ijg} \sim NB(\mu_{ijg}, \phi_g)$. The authors suggested estimation of the ϕ_g via a weighted likelihood approach, approximating an empirical Bayes procedure. The weighted likelihood for ϕ_g is defined as the weighted sum of the likelihood with gene-specific overdispersion (ϕ_g) and the common likelihood function with common overdispersion (ϕ). The weight parameter α determines the weight assigned to the common likelihood relative to the gene-specific likelihood. In practice, the parameter α is selected based on a Bayesian normal hierarchical model for the gene-specific dispersion parameters ϕ_g. The authors demonstrate that when dispersions do not differ among genes, this approach results in greater values of α, which gives greater weight to the common likelihood in the weighted likelihood equations and thus shrinks the gene-specific ϕ_g to a common value. A simulation study demonstrated that the ability of the exact test [33] to detect differentially expressed genes improved when the empirical Bayes estimation of the gene-specific dispersion parameters was implemented, and was equivalent to the performance of a Wald test from an overdispersed log-linear model when genes were commonly overdispersed. Further, the exact test with empirical Bayes adjustment was better able to control false discovery rates when gene-specific overdispersion was introduced.

Anders and Huber [1] noted that in practice, dispersion often varies with expected read count, and suggested an extended negative binomial model in which the variances of the read count are defined as a nonparametric function of their expectation. Formally, $Y_{ijg} \sim NB(\mu_{ijg}, \phi_{\mu})$, where, as in the Robinson–Smyth approach, $\mu_{ijg} = m_{ij}\lambda_{ig}$, and m_{ij} is a library size parameter accounting for the sampling depth in replicate j in population i. The notation ϕ_{μ} is understood to imply that dispersion varies in an unspecified fashion with the expectation. Under this approach, $Var(Y_{ijg}) = \mu_{ijg}(1 + \phi_{\mu}\mu_{ijg})$, which departs from the Robinson–Smyth [33] negative binomial approach for which $Var(Y_{ijg}) = \mu_{ijg}(1 + \phi\mu_{ijg})$. As noted above, Robinson and Smyth [34] extended the standard negative binomial approach by estimating gene-specific dispersion parameters via empirical Bayes weighted likelihood estimation, in which gene-specific dispersion parameter estimates were shrunk toward a common dispersion. Anders and Huber [1] employ a gamma-family generalized linear local regression to model the mean-dispersion relationship. The null hypothesis in the test of differential expression, $H_{0g} : \lambda_{1g} = \lambda_{2g}$, is tested via an exact test constructed similarly to the Robinson and Smyth test. The Robinson and Smyth approach adjusts counts by qCML to achieve equal pseudocounts per replication. The equality of the pseudocounts is then used in the construction of exact negative binomial test statistics. In contrast, Anders and Huber approximated the distribution of the sum of negative binomial random variables assuming unequal library sizes. The authors demonstrated their method on four standard data sets, noting that both approaches were effective at controlling false discovery rates, while a Poisson-based χ^2 test failed. The authors note that the overall sensitivities of their test and the common-dispersion version of the Robinson and Smyth test were roughly equivalent. However, the Robinson and Smyth test was less conservative for weakly expressed genes and more conservative for strongly expressed genes, an apparent product of the flexibility of the nonparametric variance estimator in the Anders and Huber test.

Di et al. [9] applied a generalized negative binomial distribution, known as the negative binomial power (NBP) distribution, to test for differential expression. The NBP distribution is a gamma mixture of Poisson distributions; if $Y|Z \sim POI(Z)$ and $Z \sim \Gamma$ with mean μ and variance $\phi\mu^{\alpha}$, then marginal distribution of Y is NBP. The authors note that by assuming NBP-distributed read counts, $Var(Y_{ijg}) = \mu_{ig}(1 + \phi(\mu_{ig})^{\alpha-1})$. While the dispersion parameter is common to all genes, the mean-variance relationship is given flexibility via the power parameter α. This is in contrast to the Robinson–Smyth and Anders–Huber approaches, in which the dispersion parameters themselves are varied. The NBP tests is constructed as an exact test based on the NBP assumption. The null hypothesis is $\lambda_{1g} = \lambda_{2g}$, where, as in the other negative binomial tests, $\mu_{ijg} = m_{ij}\lambda_{ig}$, and m_{ij} represents the library size for replicate j in population i. Under the assumption of equal library sizes, the authors estimate the relative frequency parameters λ_{ig} as simple averages over replicates weighted by the common library size. The dispersion parameters ϕ and α are estimated via maximum likelihood conditional on the sum of read counts within each population and the estimated λ_{ig}. The exact test is constructed in the fashion of Robinson and Smyth [33], based on the conditional distribution of the read count sums in one population given the read count sum over both populations. To permit varying library sizes, the authors randomly sample read counts to force equal active library sizes, a process they term "thinning". A simulation study was conducted in which read counts were simulated from the Poisson and several variants of the negative binomial distribution, under different assumptions on the functional form of the negative binomial variance. The authors noted that each of the negative binomial tests, including their own, appeared adequate at controlling the false discovery rate under their simulation settings, while the NBP test appeared to be most powerful, particularly under a simulation model in which the log-dispersion parameter was defined as a quadratic function of the log-mean.

Lund et al. [23] noted that while methods based on extensions of the Poisson distribution or the negative binomial distribution provide added flexibility in modeling read count overdispersion, these methods fail to properly account for uncertainty arising from estimating this overdispersion. In general, this results in overly liberal tests of differential expression and skewed p-value distributions when genes are not differentially expressed. The authors suggest modeling read counts via quasi-likelihood (QL) by defining the read count variance to be proportional to a user-defined function—$Var(Y_{ijg}) = \Phi_g V_g(\mu_{ijg})$, where Φ_g is a quasi-dispersion parameter to be estimated from the data, and the variance function $V_g()$ must possess a corresponding quasi-likelihood function satisfying $\partial l(\mu_{ijg}|y_{ijg})/\partial\mu_{ijg} = (y_{ijg} - \mu_{ijg})/V_g(\mu_{ijg})$. Differential expression is tested through a quasi-likelihood ratio test, for which three methods for estimating the QL dispersion parameter Φ_g are discussed. The first is a standard deviance-based estimator. The second is an empirical Bayes estimator, adapted from an approach introduced by Smyth [38], which borrows information across genes in estimating gene specific dispersions by placing a scaled inverse χ^2 prior distribution on the QL dispersion parameter. The third approach accounts for mean-variance relationships in the read counts by fitting a cubic spline of the logarithm of the deviance-based QL dispersion estimator against the log-average counts. A preliminary estimator of the QL dispersion is derived from the spline function, and the aforementioned empirical Bayes approach of Smyth is employed to arrive at the spline-based estimator of the QL dispersion. The authors note that the latter two methods, termed QLShrink and QLSpline, can be characterized as shrinkage estimators—weighted averages of the deviance-based and Bayesian or spline estimators. Lund et al. [23] conducted a simulation study demonstrating the liberal nature of existing Poisson and negative binomial tests, and noted that of the three proposed QL methods, the spline-based method (QLSpline) appeared to perform best.

2.2.4 Other Methods

Parametric approaches to modeling RNA-seq data based on discrete distributions for counts can be adversely affected by model misspecifications and the presence of outliers. A nonparametric approach to the identification of differentially expressed genes in RNA-seq data was proposed by Li and Tibshirani [20]. A modified two-sample Wilcoxon statistic

$$T_g^* = \frac{1}{S}\sum_{s=1}^{S}\left\{\sum_j R_{1jg}(Y'^s) - \frac{J_1(J+1)}{2}\right\} \tag{2.3}$$

based on a multiple Poisson sampling procedure over S iterations is used to examine the differential expression of the gth feature (gene) in two-class data. As in the previous section, Y_{ijg} denotes the RNA-seq count for the gth gene in the jth experimental observation in population i, J_i is the number of observations in the ith population for $i = 1,2$, and we define $J = J_1 + J_2$. The rank statistic $R_{ijg}(Y)$ gives the rank of Y_{ijg} in the set $Y = \left\{Y_{11g},\ldots,Y_{1J_1g},Y_{21g},\ldots,Y_{2J_2g}\right\}$. The use of equation (2.3) requires equal sequencing depths, so the authors suggest Poisson sampling of the read counts, replacing original counts Y_{ijg} with random variables Y'_{ijg} resampled from a Poisson distribution with mean $\bar{d}Y_{ijg}/d_{ij}$ for $i = 1,2$ and $j = 1,\ldots,J_i$ where the d_{ij} represent the original sequencing depths for replicate j in population i, and $\bar{d} = \left(\prod_{i,j} d_{ij}\right)^{1/n}$ is the geometric mean of all sequencing depths. This Poisson sampling procedure is repeated S times and the resulting average test statistic is computed to alleviate limitations resulting from the additional randomness introduced by resampling and by tie-breaking procedures for the rank statistic. Since the distribution of the average of the Wilcoxon statistics is complicated, the false discovery rate (FDR) is estimated based on a permutation plug-in estimate. The FDR estimates for this test are more conservative than for parametric alternatives, and were shown to be accurate in simulated data with outliers for which some parametric models greatly underestimated the FDR. In overdispersed data sets with outliers, parametric methods often identified features with a small number of very large count values as differentially expressed, whereas the Li and Tibshirani test tended to identify features where the counts in one class were consistently larger than the counts in the other class.

Tarazona et al. [44] introduced a nonparametric approach designed to be robust against sequencing depth effects. The empirical distributions of fold-change differences $M^g = \log_2(\tilde{Y}_{1\cdot g}/\tilde{Y}_{2\cdot g})$ and absolute expression differences $D^g = |\tilde{Y}_{1\cdot g} - \tilde{Y}_{2\cdot g}|$ are used to estimate the probability that the gth gene is differentially expressed, where the $\tilde{Y}_{i\cdot g}$ represent cumulative read counts normalized to correct for different sequencing depths and adjusted to avoid zero counts. Genes are declared to be differentially expressed if the estimated probability exceeds a specified threshold; 0.8 is used by the authors. The empirical probabilities are computed using technical replicates when available, or through technical replicates simulated from the multinomial distribution when not available. Tarazona et al. [44] examined the effect of sequencing depth on the identification of expressed genes via their nonparametric test, sequencing noise, transcript length, and genes declared to be differentially expressed. A thorough comparison to other novel methods [1, 14, 35] as well as Fisher's Exact Test was made. The authors found that the number of differentially expressed genes as well as the length, fold-change, and expression level of the discovered genes strongly depended on the sequencing depth for the parametric methods, while their nonparametric method was relatively consistent. Further, the authors noted an increase in the number of false positives as the sequencing depth increased for the parametric methods, which was also found by Li and Tibshirani [20], while their nonparametric method was able to control the rate of false discovery.

Recently, a Markov random field approach was proposed by Yang et al. [52]. Consider the set $X = \{x_1,\dots,x_G\}$ of binary random variables defining indicators x_i which equal 1 if a gene is differentially expressed and equal 0 otherwise. A vector $Y = \{y_1,\dots,y_G\}$ of observed discretized FDRs are computed for the individual genes using the Anders and Huber [1] test, and the joint probability of X given Y is modeled as proportional to the product $\prod_{(i,j)\in E} \psi_{(i,j)}(x_i,x_j) \prod_{i=1}^{G} \phi_i(x_i)$ where E is the set of vertices with coexpressed gene database (COXPRESdb) correlations $c_{i,j}$ larger than a specified value [27] and $\psi_{(i,j)}(x_i,x_j) = e^{c_{i,j}}$ if $x_i = x_j$ and 1 otherwise. The unary function $\phi_i(x_i)$ are defined to be $P(x_i = 1|y_i)/P(x_i = 0|y_i)$ if $P(x_i = 1|y_i) > P(x_i = 0|y_i)$ and $x_i = 1$, $P(x_i = 0|y_i)/P(x_i = 1|y_i)$ if $P(x_i = 0|y_i) > P(x_i = 1|y_i)$ and $x_i = 0$, and 1 otherwise. It is shown that these clique potential functions of this pairwise Markov random field model are selected so that maximum a posteriori estimation of the differentially expressed genes is reduced to a maximum flow problem discussed in Kolmogorov and Zabih [15]. By including information about the dependence of gene expressions, Yang et al. [52] show through simulation studies and real data examples that this method exhibited improved sensitivity without a loss of precision. Through the inclusion of additional coexpression information, this method additionally helped remove bias against detection of genes with low read counts.

Zhou et al. [55] proposed a beta-binomial model where the probabilities that a single read in each sample is mapped to gene g is a vector $\theta_{g\cdot}$ of beta random variables for which the logits of the expected values are modeled linearly by XB_g. The design matrix X is flexible and can include columns indicating group assignments for experimental conditions as well as any other desired covariates. The vector of regression coefficients B_g corresponds to the effects of the variables in the columns of X for the gth gene. Two approaches are considered—(1) a free model where the likelihood function is directly maximized, and (2) a shrinkage approach with a constrained model where the overdispersion ϕ_g of the beta distribution is modeled as a polynomial function of the mean. The authors additionally suggest an automatic correction for outliers. While other penalized approaches and the constrained model offer some advantages for very small sample sizes, simulation studies and a real data example support direct parametric modeling with the free model.

2.2.5 Bayesian and Empirical Bayes Approaches

A number of fully Bayesian and empirical Bayes methods have been developed for analyzing differential expression. Typically inferences on differential expression span across multiple genes and conditions, each characterized by its own set of parameters. It is frequently natural to express these parameters as a mixture over two latent states. The states may imply the presence or absence of differential effects and hence define the primary objects of inference. In Bayesian approaches, such gene-specific parameters are assigned prior distributions, which are in turn indexed by a common hyper-parameter. The model is then completed by assuming specific sampling models for normalized count data conditional on these parameters. As in frequentist settings, these distributions are chosen to allow for overdispersion, which poses a critical challenge in analyzing RNA-seq data. All Bayesian models typically follow this common hierarchy.

However, empirical Bayes and fully Bayesian methods differ sharply in their approaches to inference and shrinkage. The former estimates the relevant hyper-parameters directly from the data and through this combined estimate, pools information among genes. In contrast, fully Bayesian methods borrow strength by fixing the hyper-parameter at the highest level of the Bayesian hierarchy and sharing the parameters themselves across different levels. For example, one could achieve some shrinkage by simply assuming a common probability for the presence of indicators. More generally, the extent and nature of shrinkage vary with the desired level. Shrinkage is highly relevant in differential expression settings, where we have multiple genes but very few replicates per gene. In the following discussion, we shall review some commonly used empirical Bayes approaches introduced by van de Wiel et al. [49], Leng et al. [19], and Hardcastle and Kelly [14], and conclude by describing a fully-Bayesian method [17].

The sampling model considered by van de Wiel et al. [49] is a zero-inflated negative binomial regression: $Y_{ijg} \sim ZI - NB(\mu_{ijg}, \phi_g, w_{0g})$ and $\mu_{ijg} = h^{-1}(\beta_{g0} + \sum_k \beta_{gk} x_{ijk})$, where g indexes the genes, h is a link function, ϕ_g the negative binomial overdispersion parameter, and w_{0g} a zero-inflation parameter. The zero-inflation parameter is defined to be a probability mixing the negative binomial distribution $NB(\mu_{ijg}, \phi_g)$ with probability $1 - w_{0g}$ and a point mass at zero with probability w_{0g}. The regression coefficients are permitted to have their own normal random effects. The covariates typically correspond to different conditions or populations corresponding to possible differential expression. In assigning priors, van de Wiel et al. examined several different choices. Both flat and mixture priors were considered for β_{gl}, while the prior for $\log(\phi_g)$ was assumed to be a mixture. Each parameter family had its own associated set of hyper-parameters. A conventional method of estimating hyper-parameters in an empirical Bayes framework is by maximizing the marginal likelihood. As an alternative, van de Wiel et al. [49] utilize the fact that the likelihood estimator α approximately satisfies $\pi_\alpha(\cdot) = (1/G)\sum_{g=1}^{G} \pi_\alpha(\cdot|\mathbf{Y}_g)$, where G is the number of genes and $\mathbf{Y}_g$ the vector of read counts for gene g. This approximation can be seen by setting the derivative of the log-marginal likelihood to 0. Since the model includes multiple parameter families (e.g overdispersion, regression coefficients), this generic procedure was extended to an iterative algorithm which conditioned on a given set of parameters at each step. Shrinkage of overdispersion is treated separately. Since the overdispersion and mean are intertwined in NB models, a univariate shrinkage of the former may not work. The authors suggest shrinking the individual ϕ_g through a prior that regresses them against the gene counts. Specifically, they assume $\phi_g = h(c_g) + \varepsilon_g$ where c_g is the log of the gene count and the function h is left unspecified and estimated via LOESS. Initial values required by this iterative algorithm are fixed at the posterior mean estimates of the ϕ_g obtained under a flat prior. Having obtained these estimates, the shrinkage prior was assigned as $\phi_g - \hat{\phi}_g \sim N(0, \sigma^2)$ where σ^2 was also estimated from the iterative procedure. The authors also suggest the importance of the zero-inflation component in this context, describing it as a potential reason for overdispersion. Indeed, including factors accounting for zero inflation was shown to effectively account for the residual trends of ϕ_g in simulation settings. Finally, posterior estimates of the specific contrasts involving the regression coefficients are computed, and then Bayesian and local false discovery rates are applied to these estimates to infer differential expression.

The approach of Hardcastle and Kelly [14] deals directly with the latent indicators of differential expression. In the most general version of this approach, a broad space of models is encompassed, each corresponding to a hypothesis to be tested. For simplicity of exposition, we consider here just two exclusive models: (1) no differential expression and (2) differential expression. Each gene in the data set then has an associated latent indicator identifying whether it is differentially expressed. A key difference with the method of van de Wiel et al. [49] is that Hardcastle and Kelly [14] do not explicitly estimate a hyper-parameter. Instead, their method estimates the entire prior distribution through resampling and quasi-likelihood. The pooling of prior probabilities for the different indicators is done through iterative estimation. The sampling model in this approach is negative binomial, with the probabilities weighted by library sizes. Posterior probabilities are obtained as the final step.

Leng et al. [19] introduced an empirical Bayes method that not only models differential expression among genes but also among isoforms of the same gene. In this setup, let Y_{ijgl} denote the read counts in isoform l of gene g in sample j of population i. This count is assumed to follow a negative binomial distribution, where the parameters of the negative binomial can vary across genes, isoforms, and biological conditions. The prior distribution of the negative binomial mean-variance ratio is assumed to be $Beta(\alpha, \beta^{I_g})$, where I_g denotes a grouping of genes. The hyper-parameter α is shared across all isoforms and genes, while β varies by gene group (I_g). These gene groups can be defined freely to provide flexibility to the approach; for example, genes can be grouped by the number of their isoforms. As in other differential expression approaches, the full model was expressed as mixture over two latent states. In the EB step, the four global hyper-parameters (each pair corresponding to a state) are estimated via the EM algorithm. Conditioned on these estimates, the state-specific posterior probabilities are calculated.

Lee et al. [17] proposed a fully Bayesian hierarchical model that diverged from existing approaches in that the cumulative read count of gene g at each genomic position l in population i is explicitly modeled as $Y_{i \cdot gl} \sim Bin(Y_{\cdot \cdot gl}; p_{gl})$, independently across the positions. The binomial probability p_{gl} is modeled by adding another layer in the hierarchy and assuming $p_{gl} \sim (1 - w_{gl})Beta(\alpha_g; \beta_g) + w_{gl}Beta(.5, .5)$. In this formulation, w_{gl} expresses the outlier effect, while the α_g and β_g are gene specific and centered around a mixture prior. This mixture is over three possible indicators encoding for high, low, and non-differential expression. The parameters corresponding to each indicator are assigned their own Gaussian priors. These priors allow for the usual inter-gene pooling as in previous hierarchical setups. However, this method implements full posterior inference using MCMC methods. Final results are obtained by direct posterior sampling of the latent indicators. This approach offers a number of advantages. First, prior normalization of the mapped read counts is not required. Rather, normalization and differential calling are done simultaneously via the model. Second, this approach effectively downweights outliers at the position level through the w_{gl}. The authors showed that this step played a significant role in increasing the specificity and sensitivity of differential expression calls. Third, the pooling across positions increased the effective sample size per gene per sample. Importantly, this model uses each position in the gene as a data point, thus we have multiple observations per gene in the absence of replicates. This can be relevant for many cost-prohibitive RNA-seq studies where replicates are difficult to obtain.

2.3 Software for Differential Expression in RNA-seq Data

Several of the novel methods for detecting differential expression in RNA-seq data have associated software packages, most of which have been released via the open source R [31] and Bioconductor [12] software environments. Below we provide a brief summary of R and Bioconductor implementations of the different techniques for detecting differential expression in RNA-seq data. We do not discuss other methods such as Fisher's exact test, two sample t-tests, GLM-derived tests, and methods for microarray data analysis applied to RNA-seq data. The package names we use in the discussion below can be used to load the R and Bioconductor libraries for the associated methods, via the commands `library(`*pkgname*`)` for R packages (after local installation) and `biocLite(`*pkgname*`)` for Bioconductor packages.

The general convention for formatting RNA-seq data for use in frequentist analyses is as a $G \times J$ matrix for G genes measured in J samples, with the columns typically arranged so that the first few columns are read counts of replicates from population 1 and the remaining columns read counts of replicates from population 2. Most functions for detecting differential expression accept two arguments at a minimum—the matrix of read counts and a vector defining a population identifier for the columns (e.g. 1 or 2).

The R package *GPseq* [42] implements the generalized Poisson test via the function `estimate_differential_expression`. The interface for this function differs somewhat from other implementations, in that the function accepts an annotated read count matrix as well as exon and gene annotation matrices for unraveling the annotated read count matrix. The *GPseq* package also includes functions to calculate chi-square goodness-of-fit tests for the generalized Poisson distribution and workhorse functions for the generalized Poisson likelihood and likelihood ratios, and functions for permutation tests of the generalized Poisson test statistic. The other Poisson test, based on the two-stage Poisson model [3], is not available through an R library, rather as an R function downloadable from the authors' website (http://www.stat.purdue.edu/~doerge/software/TSPM.R). The function reads in a matrix of read counts and indicators defining populations for the columns of the matrix, and returns adjusted and unadjusted p-values as well as vectors of indicators defining genes found to be overdispersed. The R functions used to implement the procedures introduced by Pounds et al. [30] utilize some of the code for TSPM and are available on the personal website (http://www.stjuderesearch.org/site/depts/biostats/software/ebshtpasced).

DEGseq is a Bioconductor package implementing Fisher's exact test, two likelihood ratio tests, and tests based on MA plots [51], all through the function `DEGseq`. This function also does not follow the convention of accepting matrices of read counts. Rather, `DEGseq` accepts mapping files for samples from two populations as well as arguments specifying characteristics of the RNA-seq data files. Additional arguments specify the differential expression test to be conducted and customize the characteristics of said tests, such as p- and q-value thresholds and thresholds for tests derived from MA plots.

Libraries for the negative binomial tests [1, 9, 33, 34] are available in R and Bioconductor. The Robinson–Smyth test can be found in the Bionconductor package *edgeR* [35]. To obtain the Robinson–Smyth test, users of *edgeR* format a matrix of counts into a package-specific object that is then fed to the function `estimateCommonDisp`, which estimates the common dispersion parameters and outputs a matrix of pseudocounts and pseudo-library sizes. The object created by this function is fed to the function `exactTest` which calculates p-values from the exact negative binomial tests based on the quantile-adjusted counts. Additional functions in the *edgeR* library provide tests based on the assumption of gene-specific dispersion parameters, utility functions for RNA-seq data, workhorse functions for estimation and testing, and additional functions for the analysis of RNA-seq data. We refer the reader to this book's chapter on the *edgeR* package for further details. The test of Anders and Huber [1] is available via the Bioconductor package *DESeq*. To test differential expression using *DESeq*, users must create a package-specific object containing the read count matrix via the function `newCountDataSet`, normalize the counts using the function `estimateSizeFactors`, estimate overdispersion using `estimateDispersions`, and then conduct the negative binomial test using `nbinomTest`. *DESeq* includes additional functions for conducting the negative binomial test directly on count matrices, as well as functions for graphics (e.g. MA plots), variance stabilizing transformations, and negative binomial GLM tests per gene. The R package *NBPSeq* [10] implements the NBP test [9]. The function `nbp.test` accepts a matrix of read counts and vector of indicators for group membership. Normalization of read counts by the random resampling process ("thinning" as termed by the authors) is accomplished internally within `nbp.test`. Additional functions in *NBPSeq* estimate the negative binomial dispersion parameters (`estimate.disp`) and normalization factors (`estimate.norm.factors`), perform exact negative binomial tests (`exact.nb.test`) and GLM-based tests (`nb.glm.test`), and perform utility functions on package-specific objects. Most of these functions are workhorses for `nbp.test`. The quasi-likelihood approach of Lund et al. [23] is implemented in the R package *QuasiSeq*, which requires the *edgeR* library. The function `QL.fit` accepts a matrix of read counts and a list containing design matrices for full and reduced models. Additional options permit customization of the QL model and estimation of dispersion parameters. The list object returned by `QL.fit` can be fed to the function `QL.results`, which produces lists of p-values and q-values.

The nonparametric approach of Li and Tibshirani [20] is implemented by the R package *samr* [46]. The function `SAMseq` is specifically designed for the analysis of count data, whereas `samr` and other functions in the package are designed for microarray data analysis. `SAMseq` permits flexibility in the type of analysis to be conducted via the `resp.type` argument, which can be used to request paired and unpaired two-class comparisons, comparison of three or more classes, analysis of association with a quantitative predictor, and analysis of a survival outcome. Other functions in *samr* can be used to estimate sequencing depths and normalize read counts. Registered academic users can also download a supplementary Addin for Microsoft Excel from the developers web page (http://www-stat.stanford.edu/~tibs/SAM/). The nonparametric test of Tarazona et al. [44] is implemented in the Bioconductor package *NOISeq* [45] using the functions `noiseq` and `noiseqbio`. These functions, which operate on package-specific objects containing the read counts, include options for handling data with technical and biological replicates, as well as data with no replicates. The function outputs a list of differentially expressed genes based on the desired threshold probability. This package also provides several exploratory plots for biotype detection, sequencing depth and expression quantification, and sequencing bias that are useful for detecting potential problems that need to be corrected by normalization procedures and several plots which summarize the differentially expressed genes identified by the algorithm.

The Markov random field approach of Yang et al. [52], termed *MRFSeq*, is implemented as C++ code and distributed from the author's website (http://www.cs.ucr.edu/~yyang027/mrfseq.htm). *MRFSeq* depends upon the coexpressed gene database COXPRESdb [27], available at http://coxpresdb.jp, and *DESeq*, the Bioconductor package for the negative binomial test of Anders and Huber [1]. The beta-binomial test of Zhou et al. [55] is available in the R package *BBSeq*, available only from the author's webpage (http://www.bios.unc.edu/research/genomic_software/BBSeq/). The separate functions `free.estimate` and `constrained.estimate` compute parameter estimates and estimate p-values based on the corresponding likelihood and shrinkage approaches discussed in the previous section. An additional utility function (`outlier.flag`) is included to identify potential outliers among the read counts.

Among the Bayesian methods, the multiple shrinkage priors approach of van de Wiel et al. [49] is implemented in the R package *ShrinkBayes*, available from the primary author's webpage (http://www.few.vu.nl/~mavdwiel/ShrinkBayes.html). The function `ShrinkSeq` is used to fit the multiple shrinkage priors model based on specification of a model formula, the model parameters to be shrunk, whether or not a mixture prior for overdispersion is to be implemented, and the family of distributions used to fit the data (zero-inflated negative binomial being the default). Since *ShrinkSeq* is computationally intensive, parallel computing is implemented, and the user is permitted to specify the number of processors to be used in parallel. Formal documentation of the functions comprising *ShrinkSeq* are unavailable, but thorough examples of code usage are provided in the package documentation. Use of the *ShrinkBayes* package requires the installation of *inla* [36], an R package for Bayesian modeling via integrated nested Laplace approximation.

The Bioconductor package *baySeq* [13] implements the empirical Bayes method of Hardcastle and Kelly [14]. The functions `getPriors` and `getLikelihood` are the two most important functions in this package. The first constructs the empirical priors by bootstrapping, while the second yields posterior probabilities. *baySeq* offers a fair amount of choice in analysis, e.g., in the number of bootstrap samples and in techniques for re-estimating priors. *baySeq* can be run in parallel mode, via the independent R package *snow* [47] for networking workstations. *EBSeq* [18] is the Bioconductor package implementing the method of Leng et al. [19]. The `EBtest` function in this package uses the EM algorithm to obtain posterior probabilities for the detection of two-condition differential expression. The function `EBMultitest` extends this utility for multiple conditions. The underlying model in *EBSeq* is assumed to be negative binomial. *EBSeq* offers the users a range of simulated datasets upon which to test the algorithm. The R package *BMDE* implements the fully Bayesian method of Lee et al. [17], and is available for download at http://health.bsd.uchicago.edu/yji/soft.html. Since *BMDE* uses full posterior inference, it is able to provide the entire set of posterior samples, allowing the users to choose their own posterior summaries. Unlike the empirical Bayes algorithms mentioned above, *BMDE* relies on certain hyper-parameter settings. The users are provided the flexibility to choose them and examine the sensitivity of results based on selections for the hyper-parameters.

2.4 Comparison of Methods for Detecting Differential Expression

In most of the source works for the methods detailed in Sect. 2.2, simulation studies and/or analyses of live RNA-seq data sets were conducted to evaluate the detection capabilities of the proposed methods and to make comparisons to existing methods.

These simulation studies were largely designed to highlight special features of the proposed methods and demonstrate the superiority of these methods under specific conditions. More general comparative simulation studies have been conducted to compare these methods; we discuss three such studies below. We note that these comparative studies generally implemented default settings that were defined in the software packages corresponding to each method, that these default settings can change over time with new package version releases, and that the conclusions reached by each the comparative studies may be version specific.

Bullard et al. [6] compared the performance of Fisher's exact test and three tests derived from the generalized linear model in (2.1)—the likelihood ratio test (LRT), and t-tests based on the GLM-derived variance and the delta method variance. The authors compared RNA-seq data from two biological samples from the MicroArray Quality Control (MAQC) Project [37]. The detection capability of these four tests were compared using the results of analysis of 375 genes by qRT-PCR gold standard for differential expression. The authors found that the LRT and Fisher's exact test performed comparably in detecting differential expression, while the two t-tests were also comparable but exhibited substantially reduced detection rates relative to the LRT and Fisher tests. A notable contribution of this paper was the impact of filtering genes with low read counts on the detection of differential expression. After removing 186 genes with read counts less than 20 and repeating the analysis of the MAQC data, the authors noted that the detection rate of both the LRT and the t-test with GLM-based variance improved greatly and, in particular, the detection rate of the t-test was roughly equivalent to that of the LRT.

Kvam et al. [16] conducted a comparative study of the two-stage Poisson model [3] and three tests based on the negative binomial distribution—*edgeR* [35], *DESeq* [1], and *baySeq* [14]. The authors simulated data under four models—Poisson read counts with half of the genes simulated from an overdispersed Poisson model, following a simulation conducted by Auer and Doerge [3] to evaluate the TSPM, counts generated from the Poisson or negative binomial distribution with mean and dispersion parameters estimated from a known plant data set [21], and counts generated from a data set of human lymphoblastoid cell lines [29] with randomly-induced differential expression. The authors noted that the three negative binomial tests *edgeR*, *DESeq*, and *baySeq* performed similarly under each simulation setting. The performance of the TSPM test was notably affected by the number of replicates simulated, as detection capability was severely reduced for two replicates per population. Further, the TSPM notably underperformed relative to the negative binomial tests when all counts were simulated from the negative binomial distribution. An analysis of a plant data set [21] showed that *edgeR* and *DESeq* largely identified the same genes as differentially expressed, while most of the genes identified by the TSPM were not declared differentially expressed by *edgeR* or *DESeq*.

A recently published study by Soneson and Delorenzi [40] comprehensively examined via simulation the performance of nine tests—*DESeq*, *edgeR*, *NBPSeq*, *TSPM*, *baySeq*, *EBSeq*, *NOISeq*, *SAMSeq*, *ShrinkSeq*—and two tests based on the empirical Bayes linear model *limma* [38, 39] after variance-stabilizing or logarithmic transformation. Using the negative binomial distribution with common dispersion between the two populations as a foundation for simulating RNA-seq data, the authors compared the performance of the 11 tests and noted the impact on performance of mixing in Poisson-simulated counts, adding high-count outlier genes, varying the number of differentially expressed genes, the direction of differential expression (up- or down-regulated), sample size, and altering the dispersion parameter in one of the populations.

Under these multiple simulation scenarios, the authors compared the methods in terms of true positive rates (TPR), ranking of differential expression, type I error control, and false discovery rate control. We paraphrase the general characteristics of each test here, and refer the reader to the source work [40] for more detailed explanations. Among the negative binomial tests, *DESeq* was generally conservative, exhibiting low detection capability but strong FDR control, even in the presence of outliers except for when sample sizes were small (two per population). Both *edgeR* and *NBPSeq* were liberal, particularly when outliers were present. *edgeR* exhibited greater sensitivity than *NBPSeq* in most settings, and became less liberal under large sample sizes while *NBPSeq* was liberal for all sample sizes. Both were poor at controlling the FDR, and *NBPSeq* often ranked truly non-differentially expressed genes as the most differentially expressed. The hallmark characteristic of the *TSPM*, which relies on asymptotic theory for its test of differential expression, was its sample-size dependence. For small samples the *TSPM* was poor at controlling FDR and ranking differentially expressed genes, although performance improved greatly with minimal increases in sample size and outliers were generally non-problematic. The *TSPM* performed poorly in terms of differential expression rankings when all genes were overdispersed, but this was improved when non-overdispersed genes were mixed in.

When differential expression occurred in a uniform direction (e.g. all genes up-regulated in one population), *baySeq* exhibited highly variable performance for each metric (TPR, FDR control, type I error control). This effect was mitigated when differential expression was mixed. *baySeq* was largely conservative with good FDR control, except when sample sizes were low. *EBSeq* provided a liberal test with good sensitivity and poor FDR control, and was particularly resistant to the effect of outliers. Control of the FDR for *NOISeq* was unevaluated due to lack of clarity in how thresholds could be set, but it was noted that *NOISeq* was particularly adept at ranking genes when populations were differentially overdispersed. *SAMSeq* was non-sensitive at low sample sizes, but power rapidly increased with sample size, and *SAMSeq* was particularly resistant to the presence of outliers. *ShrinkSeq* exhibited high sensitivity and poor FDR control at default settings, but featured a user-controlled fold-change thresholding procedure that could conceivably offer stronger FDR control.

limma with transformation exhibited strong control of type I error that was resistant to outliers. Control of FDR was also strong and resistant to outliers, except under settings in which a large proportion of genes were uniformly upregulated in one population and when populations were differentially overdispersed. The *limma* method was relatively conservative, particularly under low sample sizes, where no genes were declared differentially expressed when only two samples per population were available.

In addition to the simulation study, Soneson and Delorenzi [40] analyzed RNA-seq data from two mouse strains [5] to compare methods. *ShrinkSeq* and *SAMSeq* called the most genes as differentially expressed, while *baySeq*, *DESeq*, and *EBSeq* were particularly conservative. Among the negative binomial methods and *TSPM*, all genes called as differentially-expressed by *DESeq* were also called by one or more of the other methods. *NBPSeq*, *TSPM*, and *edgeR* called a substantial number of the same genes, but also called a non-trivial number of distinct genes not called by the other methods. Genes called by *baySeq* were a subset of those called by the log-transformed *limma* method, and the genes called by the variance-stabilized *limma* method contained most genes called by log-transformed *limma*. Genes called by *EBSeq* were effectively a subset of the variance-stabilized *limma* method, although *EBSeq* called a substantial number of unique genes. A resampled analysis of one of the mouse-strains, under which no genes would be expected to be differentially expressed, showed the tendency of *TSPM* to be too liberal, as the average number of genes called differentially expressed by *TSPM* was far greater than the other methods.

2.5 Discussion

The challenge in analyzing RNA-seq data, particularly in the detection of differential expression, has three primary sources. The first is the inherent problem with the technology; the second is the laboratory or experimental errors causing technical variation across samples. However, these sources of error are usually present in any relatively new technology. The third and the most important challenge is that current costs of producing RNA-seq data are prohibitive to the generation of many biological replicates, which poses a problem for statistical data analysis. Very small sample sizes for a typical RNA-seq study prevent the appropriate use of asymptotic statistical inference commonly employed for count data analysis. Frequently, due to these reasons, estimated false discovery rates (FDR) are not less than the selected FDR cut-off. Thus, asymptotic tests are adversely affected by small sample size in the analysis of RNA-seq data.

Small sample sizes (two samples per condition) imposed problems also for the methods that were indeed able to find differentially expressed genes, thereby leading to false discovery rates sometimes widely exceeding the desired threshold implied by the FDR cut-off. For the parametric methods, this may also be due to inaccuracies in the estimation of mean and dispersion parameters. In the previous section, we noted that TSPM stood out as the method being most affected by sample size, potentially due to the use of asymptotic statistics. Currently, RNA-seq experiments are often too expensive to allow extensive replication in scientific experiments. Hence, we strongly suggest that the differentially expressed genes found between small sample studies be interpreted with caution and that the true FDR may be several times higher than the selected FDR threshold. The negative binomial methods [1, 9, 33, 34] tests are based on similar principles and work relatively well. However, due to differences in the estimation of the overdispersion parameters, lists of differentially expressed genes produced by these methods at the same FDR level were different.

In Sect. 2.4, we summarized the results of a detailed comparison of many existing methods and the resulting guidelines to users about the suitability of one method over others for a given data type. We advocate that those testing differential expression in RNA-seq data be cognizant of the characteristics of their data, particularly with regard to the simulation settings evaluated by Soneson and Delorenzi [40]—sample size, direction of regulation, presence of outliers, degree and variability of overdispersion. Awareness of these characteristics will permit a more informed choice of test for differential expression. We also advocate that analysts not rely on a single test of differential expression nor on a single setting for a given test, and rather perform several tests or several settings of a given test based on their suitability for the data set at hand and compare lists of differentially expressed genes. We have also provided brief descriptions of existing software and their respective functionality in analysis of RNA-seq data. We hope that this review will provide a comprehensive description of the current status of the analysis of RNA-seq data.

References

[1] Anders, S., Huber, W.: Differential expression analysis for sequence count data. Genome Biol. **11**, R106 (2010)

[2] Anders, S., McCarthy, D.J., Chen, Y., Okoniewski, M., Smyth, G.K., Huber, W., Robinson, M.D.: Count-based differential expression analysis of RNA sequencing data using R and bioconductor. Nat. Protocol. **8**, 1765–1786 (2013)

[3] Auer, P.L., Doerge, R.W.: A two-stage poisson model for testing RNA-seq data. Stat. Appl. Genet. Mol. Biol. **10(1)**, 26 (2011)

[4] Benjamini, Y., Hochberg, Y.: Controlling the false discovery rate: a practical and powerful approach to multiple testing. J. Roy. Stat. Soc. Ser. B **57**, 289–300 (1995)

[5] Bottomly, D., Walter, N.A., Hunter, J.E., Darakjian, P., Kawane, S., Buck, K.J., Searles, R.P., Mooney, M., McWeeney, S.K., Hitzermann, R.: Evaluating gene expression in C57BL/6J and DBA/2J mouse striatum using RNA-seq and microarrays. PLoS One **6**(3), e17820 (2011)

[6] Bullard, J.H., Purdom, E., Hansen, K.D., Dudoit, S.: Evaluation of statistical methods for normalization and differential expression in mRNA-seq experiments. BMC Bioinform. **11**, 94 (2010)

[7] Canales, R.D., Luo, Y., Willey, J.C., Austermiller, B., Barbacioru, C.C., Boysen, C., Hunkapiller, K., Jensen, R.V., Knight, C.R., Lee, K.Y., et al.: Evaluation of DNA microarray results with quantitative gene expression platforms. Nat. Biotech. **24**(9), 1115–1122 (2006)

[8] Cloonan, N., Forrest, A.R.R., Kolle, G., Gardiner, B.B.A., Faulkner, G.J., Brown, M.K., Taylor, D.F., Steptoe, A.L., Wani, S., Bethel, G., et al.: Stem cell transcriptome profiling via massive-scale mRNA sequencing. Nat. Meth. **5**, 613–619 (2008)

[9] Di, Y., Schafer, D.W., Cumbie, J.S., Chang, J.H.: The NBP negative binomial model for assessing differential gene expression from RNA-seq. Stat. Appl. Genet. Mol. Biol. **10**(1), 24 (2011)

[10] Di, Y., Schafer, D.W, Cumbie, J.S., Chang, J.H. NBPSeq: negative binomial models for RNA-sequencing data. R Package Version 0.1.8. (2012). http://CRAN.R-project.org/package=NBPSeq

[11] Dillies, M.A., Rau, A., Aubert, J., Hennequet-Antier, C., Jeanmougin, M., Servant, N., Keime, C., Marot, G., Castel, D., Estelle, J., et al.: A comprehensive evaluation of normalization methods for Illumina high-throughput RNA sequencing data analysis. Brief Bioinform. (2012). doi:10.1093/bib/bbs046

[12] Gentleman R., Carey V.J., Bates D.M., Bolstad B., Dettling M., Dudoit S., Ellis B., Gautier L., Ge Y., Others: Bioconductor: open software development for computational biology and bioinformatics. Genome Biol. **5**, R80 (2004)

[13] Hardcastle, T.J.: baySeq: empirical Bayesian analysis of patterns of differential expression in count data. R Package Version 1.16.0. (2012)

[14] Hardcastle, T.J., Kelly, K.A.: baySeq: empirical Bayesian methods for identifying differential expression in sequence count data. BMC Bioinform. **11**, 422 (2010)

[15] Kolmogorov, V., Zabih, R.: What energy functions can be minimized via graph cuts? IEEE Trans. Pattern Anal. Mach. Intell. **26**, 147–159 (2004)

[16] Kvam, V.M., Liu, P., Si, Y.: A comparison of statistical methods for detecting differentially expressed genes from RNA-seq data. Am. J. Botany **99**(2), 248–256 (2012)

[17] Lee, J., Ji, Y., Liang, S., Cai, G., Muller, P.: On differential gene expression using RNA-seq data. Cancer Inform. **10**, 205–215 (2011)

[18] Leng, N.: EBSeq: an R package for gene and isoform differential expression analysis of RNA-seq data. R Package Version 1.2.0 (2013)

[19] Leng, N., Dawson, J., Thomson, J., Ruotti, V., Rissman, A., Smits, B., Haag, J., Gould, M., Stewart, R., Kendziorski, C.: EBSeq: an empirical bayes hierarchical model for inference in RNA-seq experiments. Technical Report 226. Department of Biostatistics and Medical Informatics, University of Wisconsin, Madison (2012). http://www.biostat.wisc.edu/Tech-Reports/pdf/tr_226.pdf

[20] Li, J., Tibshirani, R.: Finding consistent patterns: a nonparametric approach for identifying differential expression in RNA-seq data. Stat. Meth. Med. Res. **22**(5), 519–536 (2011)

[21] Li, P., Ponnala, L., Gandotra, N., Wang, L., Si, Y. Tausta, S.L., Kebrom, T.H., et al. The developmental dynamics of the maize leaf transcriptome. Nat. Genet. **42**, 1060–1067 (2010)

[22] Lister, R., O'Malley, R.C., Tonti-Filippini, J., Gregory, B.D., Berry, C.C., Millar, A.H., Ecker, J.R.: Highly integrated single-base resolution maps of the epigenome in Arabidopsis. Cell **133**, 523–536 (2008)

[23] Lund, S.P., Nettleton, D., McCarthy, D.J., Smyth, G.K.: Detecting differential expression in RNA-sequence data using quasi-likelihood with shrunken dispersion estimates. Stat. Appl. Genet. Mol. Biol. **11**(5), Article 8 (2012)

[24] Marioni, J.C., Mason, C.E., Mane, S.M., Stephens, M., Gilad, Y.: RNA-seq: an assessment of technical reproducibility and comparison with gene expression arrays. Genome Res. **18**, 1509–1517 (2008)

[25] Mortazavi, A., Williams, B.A., McCue, K., Schaeffer, L., Wold, B.: Mapping and quantifying mammalian transcriptomes by RNA-seq. Nat. Meth. **5**, 621–628 (2008)

[26] Nagalakshmi, U., Wang, Z., Waern, K., Shou, C., Raha, D., Gerstein, M., Snyder, M.: The transcriptional language of the yeast genome defined by RNA sequencing. Science **320**(5881), 1344–1349 (2008)

[27] Obayashi, T., Kinoshuta, K.: Coxpresdb: a database to compare gene coexpression in seven model animals. Nucleic Acids Res. **39**, D1016–D1022 (2011)

[28] Pan, Q., Shai, O., Lee, L.J., Frey, B.J., Blencowe, B.J.: Deep surveying of alternative splicing complexity in the human transcriptome by high-throughput sequencing. Nat. Genet. **40**, 1413–1415 (2008)

[29] Pickrell, J.K. , Marioni, J.C., Pai, A.A., Degner, J.F., Engelhardt B.E., Nkadori, E., Veyrieras, J.B., et al.: Understanding mechanisms underlying human gene expression variation with RNA sequencing. Nature **464**, 768–772 (2010)

[30] Pounds, S.B., Gao, C.L., Zhang, H.: Empirical Bayesian selection of hypothesis testing procedures for analysis of sequence count expression data. Stat. Appl. Genet. Mol. Biol. **11**(5), Article 7 (2012)

[31] R Core Team: R: A Language and Environment for Statistical Computing. R Foundation for Statistical Computing, Vienna, Austria (2013). http://www.R-project.org/

[32] Robinson, M.D., Oshlack, A.: A scaling normalization method for differential expression analysis of RNA-seq data. Genome Biol. **11**, R25 (2010)

[33] Robinson, M.D., Smyth, G.K.: Moderated statistical tests for assessing differences in tag abundance. Bioinformatics **23**, 2881–2887 (2007)

[34] Robinson, M.D., Smyth, G.K.: Small-sample estimation of negative binomial dispersion, with applications to SAGE data. Biostatistics **9**, 321–332 (2008)

[35] Robinson, M.D., McCarthy, D.J., Smyth, G.K.: edgeR: a bioconductor package for differential expression analysis of digital gene expression data. Bioinformatics **26**, 139–140 (2010)

[36] Rue, H., Martino, S., Chopin, N.: Approximate Bayesian inference for latent Gaussian models using integrated nested Laplace approximations (with discussion). JRSSB **71**(2), 319–392 (2009)

[37] Shi, L., Reid, L.H., Jones, W.D., Shippy, R., Warrington, J.A., Baker, S.C., Collins, P.J., de Longueville, F., Kawasaki, E.S., Lee, K.Y., et al.: The microarray quality control (MAQC) project shows inter- and intraplatform reproducibility of gene expression measurements. Nat. Biotech. **24**, 1151–1161 (2006)

[38] Smyth, G.K.: Linear models and empirical bayes methods for assessing differential expression in microarray experiments. Stat. Appl. Genet. Mol. Biol. **3**, Article 3 (2004)

[39] Smyth, G.K.: Limma: linear models for microarray data. In: Gentleman, R., Carey, V., Dudoit, S., Irizarry, R., Huber, W. (eds.) Bioinformatics and Computational Biology Solutions Using R and Bioconductor, pp. 397–420. Springer, New York (2005)

[40] Soneson, C., Delorenzi, M.: A comparison of methods for differential expression analysis of RNA-seq data. BMC Bioinform. **14**, 91 (2013)

[41] Srivastava, S., Chen, L.: A two-parameter generalized Poisson model to improve the analysis of RNA-seq data. Nucleic Acids Res. **38**(17), e170 (2010)

[42] Srivastava, S., Chen, L.: GPseq: using the generalized Poisson distribution to model sequence read counts from high throughput sequencing experiments. R Package Version 0.5. (2011). http://CRAN.R-project.org/package=GPseq

[43] Sultan, M., Schulz, M.H., Richard, H., Magen, A., Klingenhoff, A., Scherf, M., Seifert, M., Borodina, T., Soldatov, A., Parkhomchuk, D., et al.: A global view of gene activity and alternative splicing by deep sequencing of the human transcriptome. Science **321**, 956–960 (2008)

[44] Tarazona, S., García-Alcalde, F., Dopazo, J., Ferrer, A., Conesa, A.: Differential expression in RNA-seq: a matter of depth. Genome Res. **21**, 2213–2223 (2011)

[45] Tarazona, S., Furio-Tari, P., Ferrer, A., Conesa, A.: NOISeq: Exploratory analysis and differential expression for RNA-seq data. R Package Version 2.2.1 (2012)

[46] Tibshirani, R., Chu, G., Narasimhan, B., Li, J.: samr: SAM: significance analysis of microarrays. R Package Version 2.0. (2011). http://CRAN.R-project.org/package=samr

[47] Tierney, L., Rossini, A.J., Li, N., Sevcikova, H.: snow: simple Network of Workstations. R Package Version 0.3–13 (2013). http://CRAN.R-project.org/package=snow

[48] Trapnell, C., Williams, B.A., Pertea, G., Mortazavi, A., Kwan, G., van Baren, M.J., Salzberg, S.L., Wold, B.J., Pachter, L.: Transcript assembly and quantification by RNA-seq reveals unannotated transcripts and isoform switching during cell differentiation. Nat. Biotech. **28**, 511–515 (2010)

[49] van de Wiel, M.A., Leday, G.G.R., Pardo, L., Rue, H., van der Vaart, A.W., Van Wieringen, W.N.: Bayesian analysis of RNA sequencing data by estimating multiple shrinkage priors. Biostatistics **14**, 113–128 (2012)

[50] Wang, Z., Gerstein, M., Snyder, M.: RNA-seq: a revolutionary tool for transcriptomics. Nat. Rev. Genet. **10**, 57–63 (2009)

[51] Wang, L., Feng, Z., Wang, X., Wang, X., Zhang, X.: DEGseq: an R package for identifying differentially expressed genes from RNA-seq data. Bioinformatics **26**, 136–138 (2010)

[52] Yang, E., Girke, T., Jiang, T.: Differential gene expression analysis using coexpression and RNA-seq data. Bioinformatics **29**(17), 2153–2161 (2013). doi:10.1093/bioinformatics/btt363

[53] Yendrek, Y.R., Ainsworth, A.A., Thimmaruram, J.: The bench scientist's guide to statistical analysis of RNA-seq data. BMC Res. Notes **5**, 506 (2012)

[54] Young, M.D., Wakefield, M.J., Smyth, G.K., Oshlack, A.: Gene ontology analysis for RNA-seq: accounting for selection bias. Genome Biol. **11**, R14 (2010). doi:10.1186/gb-2010-11-2-r14

[55] Zhou, Y., Xia, K., Wright, F.A.: A powerful and flexible approach to the analysis of RNA sequence count data. Bioinformatics **27**(19), 2672–2678 (2011)

Springer

http://www.springer.com/978-3-319-07211-1

Statistical Analysis of Next Generation Sequencing Data
Datta, S.; Nettleton, D. (Eds.)
2014, XIV, 432 p. 87 illus., 68 illus. in color., Hardcover
ISBN: 978-3-319-07211-1