Biostatistics (2012), **13**, 3, *pp.* 523–538
doi:10.1093/biostatistics/kxr031
Advance Access publication on October 14, 2011

Normalization, testing, and false discovery rate estimation for RNA-sequencing data

JUN LI*

Department of Statistics, Stanford University, Stanford, CA 94305, USA
junli07@stanford.edu

DANIELA M. WITTEN

Department of Biostatistics, University of Washington, Seattle, WA 98195, USA
dwitten@u.washington.edu

IAIN M. JOHNSTONE

Department of Statistics, Stanford University, Stanford, CA 94305, USA
imj@stanford.edu

ROBERT TIBSHIRANI

Department of Health Research & Policy, and Statistics, Stanford University, Stanford, CA 94305, USA
tibs@stanford.edu

SUMMARY

We discuss the identification of genes that are associated with an outcome in RNA sequencing and other sequence-based comparative genomic experiments. RNA-sequencing data take the form of counts, so models based on the Gaussian distribution are unsuitable. Moreover, normalization is challenging because different sequencing experiments may generate quite different total numbers of reads. To overcome these difficulties, we use a log-linear model with a new approach to normalization. We derive a novel procedure to estimate the false discovery rate (FDR). Our method can be applied to data with quantitative, two-class, or multiple-class outcomes, and the computation is fast even for large data sets. We study the accuracy of our approaches for significance calculation and FDR estimation, and we demonstrate that our method has potential advantages over existing methods that are based on a Poisson or negative binomial model. In summary, this work provides a pipeline for the significance analysis of sequencing data.

Keywords: Differential expression; FDR; Overdispersion; Poisson log-linear model; RNA-Seq; Score statistic.

1. INTRODUCTION

In the past 15 years, it has become well understood that disease states and biological conditions are characterized by distinct patterns of gene expression (see for instance DeRisi *and others*, 1997; Spellman *and others*, 1998; Eisen and Brown, 1999; Brown and Botstein, 1999). During most of that time, the

*To whom correspondence should be addressed.

microarray has been the primary tool for assessing gene expression. In recent years, a new approach called RNA sequencing (RNA-Seq) has been developed (see e.g. Mortazavi *and others*, 2008; Nagalakshmi *and others*, 2008; Wang *and others*, 2009; Wilhelm and Landry, 2009).

Figure 1 shows the protocol of a typical RNA-Seq experiment. Messenger RNAs (mRNAs) from a biological sample are randomly fragmented into small pieces, which are then reverse transcribed into complementary DNA (cDNA) using random primers. This cDNA library is then amplified using PCR and sequenced by a sequencing machine, resulting in millions of short sequence read-outs called "reads." These reads can then be mapped to the genome or transcriptome or used for *de novo* assembly (not illustrated in Figure 1). The number of reads mapped to a particular region of interest provides a measure of expression of that region. In this paper, we will for simplicity refer to all such regions of interest as "genes," though in practice other regions, such as exons, may be of interest. RNA-Seq overcomes some major limitations of the microarray by discovering *de novo* transcripts efficiently (since no reference genome is used to generate reads) and by avoiding problems associated with cross hybridization (since no hybridization is used). For these reasons, it is expected that the technique's popularity will continue to increase (Shendure, 2008).

RNA-Seq, like microarrays, is often used for comparative experiments, where expression measurements are collected for multiple samples, each of which is associated with an "outcome." This outcome often takes 1 of 3 forms: (i) two class, such as tumor versus normal, (ii) multiple class, such as tumor type A versus tumor type B versus tumor type C, or (iii) quantitative, such as blood pressure. In a comparative experiment, one is typically interested in finding genes that are "associated" with the outcome—genes that are overexpressed in one or several classes or genes that exhibit increased expression as the quantitative outcome increases. We will refer to such genes as "differentially expressed." Suppose comparative experiments are done on n experiments and each experiment measures expression levels of m genes, the data can be written as an $n \times m$ matrix $\mathbf{N}$ for which N_{ij} is the measure of expression for Gene j in Experiment i. For RNA-Seq data, N_{ij} is a nonnegative integer, that is, the number of reads mapped to this gene; for microarray data, N_{ij} is a real number. We also denote the outcome measurement associated with Experiment i by $y_i, i = 1, \ldots, n$.

Many proposals have been made for identifying differentially expressed genes using microarray data (see e.g. Dudoit *and others*, 2002; Kerr *and others*, 2000; Newton *and others*, 2001; Tusher *and others*,

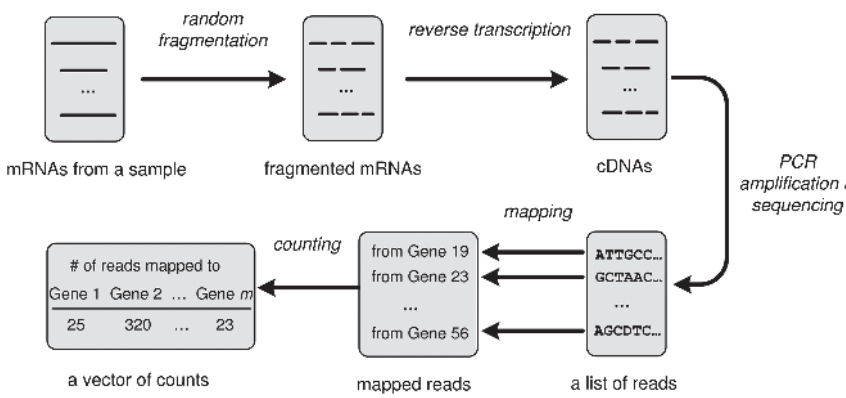

Fig. 1. Pipeline for a typical RNA-Seq experiment. Firstly, mRNA is randomly fragmented into small pieces. These small pieces are then reverse transcribed into a cDNA library by random priming. Then this cDNA library is amplified via PCR and sequenced by a sequencing machine, producing a list of reads. These reads are mapped to a known transcriptome which consists of m genes, and the number of reads mapping to each gene is used as a measure of gene expression. Thus, the results of an RNA-Seq experiment are summarized by a vector of m counts.

2001). However, these approaches are not directly applicable to RNA-Seq data due to intrinsic differences between the technologies. The RNA-Seq data matrix is composed of nonnegative integers instead of real numbers and so proposals based on a Gaussian assumption do not directly apply. Secondly, each RNA-Seq experiment generates a different total number of reads, so count N_{ij} depends not only on the expression of Gene j but also on the total number of reads generated by Experiment i ($\sum_{j=1}^{m} N_{ij}$). For example, if no genes are differentially expressed between Experiments 1 and 2, but $\sum_{j=1}^{m} N_{1j} = 1 \times 10^6$ and $\sum_{j=1}^{m} N_{2j} = 2 \times 10^6$, then it is likely that $N_{2j} \simeq 2N_{1j}$, for any $j = 1, \ldots, m$. In this case, we say that the "sequencing depth" of Experiment 2 is twice that of Experiment 1. The sequencing depth is a measure of the *relative* number of reads generated by an experiment. Counts from each experiment should be "normalized" by the sequencing depth of that experiment before any comparison is made between experiments. However, accurate estimation of the sequencing depth is often not trivial (reviewed in Section 2.1). Given sequencing depth estimates, the normalization is often done implicitly (i.e. including the sequencing depth as a term in the model) rather than explicitly (i.e. scaling each count by the corresponding sequencing depth and using the scaled data for detecting differential expression).

In this paper, we develop a method for detecting differentially expressed genes on the basis of RNA-Seq data. Our proposal is based on a Poisson log-linear model. Unlike existing methods, our method can be used not only in the case of a two-class outcome but also for a multiple-class or quantitative outcome. A new type of normalization is performed in our method. We also develop a new permutation plug-in approach for estimating the false discovery rate (FDR), and we demonstrate that the resulting FDR estimates are more accurate than past methods from the literature.

The rest of this paper is organized as follows. In Section 2, we review existing work. In Section 3, we present a log-linear model for sequencing data and an associated score statistic for quantifying differential expression. In Section 4, we suggest a strategy for FDR estimation. We study the performance of our proposed method in a simulation study in Sections 5 (Poisson data) and 6 (negative binomial data). Section 7 contains an application to two real data sets, and Section 8 contains the Discussion.

2. A REVIEW OF EXISTING WORK

2.1 *Normalization for sequence data*

As mentioned previously, RNA-Seq data must be normalized by the sequencing depth before any comparison of the counts between experiments can be made. We let d_i denote the sequencing depth for Experiment i; this is a measure of the number of reads generated by Experiment i. It seems natural to estimate d_i by $\sum_{j=1}^{m} N_{ij}$, the total number of reads generated by Experiment i. We will refer to this approach as "total-count normalization." However, total-count normalization can lead to very poor results, as one can see from a toy example with two experiments. Suppose that each experiment measures the expression of 101 genes. Suppose also that $N_{1j} = 100$ and $N_{2j} = 80$ for $j = 1, \ldots, 100$ and that $N_{1j} = 0$ and $N_{2j} = 2000$ for $j = 101$. In this case, $\sum_{j=1}^{m} N_{1j} = \sum_{j=1}^{m} N_{2j} = 10\,000$, so total-count normalization gives $\hat{d}_1 = \hat{d}_2$. This suggests that counts from the two experiments are directly comparable, and all genes are differentially expressed. However, it seems much safer to believe that only Gene 101 is differentially expressed and that a better estimate of sequencing depth is $\hat{d}_1 = 1.25\hat{d}_2$.

Several past proposals have sought to overcome the problems associated with total-count normalization. Trimmed mean of M values (TMM) normalization (Robinson and Oshlack, 2010) estimates sequencing depth after excluding genes for which the ratio of counts between a pair of experiments is too extreme or for which the average expression is too extreme. Quantile normalization (Bullard *and others*, 2010) estimates the sequencing depth of an experiment by the upper quantile of its counts. In DESeq (Anders and Huber, 2010), the sequencing depth is estimated by the count of the gene with the median count ratio across all genes. These three methods give $\hat{d}_1 = 1.25\hat{d}_2$ in the above example.

2.2 *Identification of differentially expressed genes from RNA-Seq data*

A number of proposals have been made for identifying differentially expressed genes from RNA-Seq data in the case of a two-class outcome. With the exception of baySeq (Hardcastle and Kelly, 2010), these methods yield a p-value for each gene to indicate the extent to which it is differentially expressed. Bloom *and others* (2009) apply a *t*-test to the total-count normalized data. 't Hoen *and others* (2008) take a square-root transformation for the total-count normalized data to stabilize the variance and then apply a *t*-test. A number of methods based on a Poisson distribution have also been developed. Marioni *and others* (2008) propose a Poisson log-linear model and use the classical likelihood ratio test (LRT) to calculate the p-value. Here, the total-count normalization is used implicitly. Bullard *and others* (2010) take a similar approach but use quantile normalization rather than total-count normalization. In DEGseq proposed by Wang *and others* (2010), it is assumed that log ratios of the counts have a normal distribution and a z-score is calculated. Log-linear models based on a negative binomial distribution have also been developed in order to deal with overdispersion in some RNA-Seq data sets. Robinson *and others* (2010) adapt methods previously developed for SAGE data (Robinson and Smyth, 2007, 2008; Baggerly *and others*, 2004; Lu *and others*, 2005) to RNA-Seq data. Their method, edgeR, can use either total-count normalization or TMM normalization to estimate the sequencing depth, and it estimates the dispersion of the negative binomial distribution from replicates in each class. The dispersion parameters can be estimated for each gene or can be common to all genes, making this method quite flexible. Another method, DESeq by Anders and Huber (2010), also uses a negative binomial distribution and uses local regression to estimate the relationship between the variance and the mean.

2.3 *Corrections for multiple comparisons*

Due to the large number of genes in a typical RNA-Seq data set, correction for multiple comparisons is very important. The FDR (Benjamini and Hochberg, 1995) provides an attractive measure of control for multiple testing in genomic settings. The true FDR is unknown on real data sets; one can estimate it using the procedure of Benjamini and Hochberg (1995), which requires the availability of accurate p-values. Past papers on RNA-Seq data have used the asymptotic distributions of the test statistics in order to obtain p-values; we will refer to these as "theoretical p-values." Unfortunately, if the sample size is small or if the model assumed for the data does not hold, these theoretical p-values can be quite wrong, leading to inaccurate estimates of FDR.

A solution is a permutation plug-in procedure (Tusher *and others*, 2001; Storey, 2002; Storey and Tibshirani, 2003, 2002; Storey, 2003), which is often used in the analysis of microarray data. It uses permutations to generate the null distribution of the test statistics and then derives p-values from this data-specific null distribution. We will show that it is not straightforward to adapt this simple approach to sequencing data and that doing so in a naive way can lead to incorrect results.

3. POISSON LOG-LINEAR MODEL, ESTIMATION, AND TESTING

3.1 *A log-linear model for sequencing data*

We again let $\mathbf{N}$ denote an $n \times m$ data matrix, where N_{ij} is the count for Gene j in Experiment i. Let $N_{i\cdot} = \sum_{j=1}^{m} N_{ij}$, $N_{\cdot j} = \sum_{i=1}^{n} N_{ij}$, $N_{\cdot\cdot} = \sum_{i=1}^{n}\sum_{j=1}^{m} N_{ij}$. We assume that $N_{ij} \sim \text{Poisson}(\mu_{ij})$, where the form of μ_{ij} is given by a log-linear model depending on the type of outcome. In the case of a quantitative outcome $y_i \in \mathbb{R}$ which we assume to be centered, ($\sum_{i=1}^{n} y_i = 0$), μ_{ij} takes the form

$$\log \mu_{ij} = \log d_i + \log \beta_j + \gamma_j y_i. \tag{3.1}$$

Here, d_i is the sequencing depth for Experiment i, and we assume that $\sum_{i=1}^{n} d_i = 1$, without loss of generality. β_j captures the expression level of Gene j, and γ_j is the slope for the association of Gene j with the outcome. $\gamma_j = 0$ means that Gene j is unrelated to the outcome. A gene is associated with the outcome if $\gamma_j \neq 0$.

In the case of a two-class or multiple-class outcome, let K denote the number of classes, $y_i \in \{1, \ldots, K\}$ the class label for Experiment i, and C_k the indices of the experiments in class k, that is, $C_k = \{i : y_i = k\}$. Then, μ_{ij} takes the form

$$\log \mu_{ij} = \log d_i + \log \beta_j + \sum_{k=1}^{K} \gamma_{jk} I_{(i \in C_k)}. \tag{3.2}$$

Then, $\gamma_{j1} = \ldots = \gamma_{jK} = 0$ indicates that the expression of Gene j is not associated with the class labels. When $K = 2$, this is equivalent to past proposals (Marioni *and others*, 2008; Bullard *and others*, 2010; Witten *and others*, 2010; Wang *and others*, 2010).

3.2 *A new method for data normalization*

We fit the model via the following two-step procedure:

Step 1. We fit the model under the null hypothesis that no gene is associated with the outcome:

$$\log \mu_{ij} = \log d_i + \log \beta_j. \tag{3.3}$$

Let the fit from this model be $\hat{\mu}_{ij} = N_{ij}^{(0)}$.

Step 2. We fit an additional term to the model in order to accommodate differential expression:

$$\log \mu_{ij} = \log N_{ij}^{(0)} + \gamma_j y_i \tag{3.4}$$

in the case of a quantitative outcome and

$$\log \mu_{ij} = \log N_{ij}^{(0)} + \sum_{k=1}^{K} \gamma_{jk} I_{(i \in C_k)} \tag{3.5}$$

in the case of a two-class or multiple-class outcome.

We now consider the problem of fitting the model in Step 1. Step 2 will be considered in Section 3.3. We fit β_j by maximum likelihood: $\hat{\beta}_j = N_{\cdot j}$. We could also estimate d_i by maximum likelihood, resulting in $\hat{d}_i = \frac{N_{i\cdot}}{N_{\cdot\cdot}}$, which is exactly the total-count normalization approach previously shown to be problematic. Instead, we seek a set S of genes that are not differentially expressed and use the estimate

$$\hat{d}_i = \frac{\sum_{j \in S} N_{ij}}{\sum_{j \in S} N_{\cdot j}}. \tag{3.6}$$

Note that total-count normalization is a special case of the above estimate when S is taken to be the full set of genes. We employ a Poisson goodness-of-fit statistic to estimate which genes belong to S. Given the estimate $\hat{d}_i$ as well as the maximum likelihood estimate $\hat{\beta}_j = N_{\cdot j}$, the expected value of N_{ij} is $\hat{d}_i N_{\cdot j}$. So the goodness-of-fit statistic is

$$\mathrm{GOF}_j = \sum_{i=1}^{n} \frac{(N_{ij} - \hat{d}_i N_{\cdot j})^2}{\hat{d}_i N_{\cdot j}}. \tag{3.7}$$

We set S to be the genes whose GOF_j values are in the $(\epsilon, 1-\epsilon)$ quantile of all GOF_j values, where $\epsilon \in (0, \frac{1}{2})$ is a fixed constant. Using this S in (3.6) results in an updated $\hat{d}_i$, which in turn can be used in (3.7). We use the maximum likelihood estimate for d_i as an initial estimate and then iterate to get the final estimate. The estimate of d_i converges quickly: In our simulation study, five iterations suffice. The choice of ϵ will affect the performance of our method, since a larger ϵ excludes more genes, leading to an estimate with less bias, but more variance. To obtain the results reported in this paper, we used $\epsilon = 0.25$. That is, we used half of the genes to estimate the sequencing depth. In Section 5, we show that this yields accurate results relative to existing methods for normalizing the experiments. Given $\hat{d}_i$ and $\hat{\beta}_j$, $N_{ij}^{(0)}$ is simply their product.

3.3 *Score statistics for hypothesis testing*

In Step 2 of the model-fitting procedure described in Section 3.2, we fit an additional term to the log-linear model (3.3) that measures the extent to which each gene is associated with the outcome (3.4, 3.5). While we are somewhat interested in the resulting parameter estimate, our main interest lies in determining whether the estimate is nonzero. A number of tools are available for hypothesis testing in log-linear models (see e.g. Agresti, 2002); these include the likelihood ratio statistic, the Wald statistic, and the score statistic. In this paper, we propose the use of a score statistic, as it does not require estimation of the parameter being tested. In the interest of brevity, we will refer to genes that are not truly differentially expressed as "null" genes and will refer to differentially expressed genes as "non-null."

We begin with the case of a quantitative outcome. In (3.4), given $N_{ij}^{(0)}$, the log likelihood is given by $L = \sum_{i=1}^{n} \sum_{j=1}^{m} \left[N_{ij} (\log N_{ij}^{(0)} + \gamma_j y_i) - N_{ij}^{(0)} \exp(\gamma_j y_i) \right]$, and so the score statistic for Gene j is

$$S_j = \frac{\left(\frac{\partial L}{\partial \gamma_j} |_{\gamma_j = 0} \right)^2}{-\mathbf{E} \frac{\partial^2 L}{\partial \gamma_j^2} |_{\gamma_j = 0}} = \frac{\left[\sum_{i=1}^{n} y_i (N_{ij} - N_{ij}^{(0)}) \right]^2}{\sum_{i=1}^{n} y_i^2 N_{ij}^{(0)}}. \tag{3.8}$$

For a two-class or multiple-class outcome, given $N_{ij}^{(0)}$ in (3.5), the log likelihood is $L = \sum_{i=1}^{n} \sum_{j=1}^{m} \left[N_{ij} (\log N_{ij}^{(0)} + \sum_{k=1}^{K} \gamma_{jk} I_{(i \in C_k)}) - N_{ij}^{(0)} \exp(\sum_{k=1}^{K} \gamma_{jk} I_{(i \in C_k)}) \right]$ and so the score statistic for Gene j is given as

$$S_j = \sum_{k=1}^{K} \frac{\left(\frac{\partial L}{\partial \gamma_{jk}} |_{\gamma_{jk} = 0} \right)^2}{-\mathbf{E} \frac{\partial^2 L}{\partial \gamma_{jk}^2} |_{\gamma_{jk} = 0}} = \sum_{k=1}^{K} \frac{\left[\sum_{i \in C_k} (N_{ij} - N_{ij}^{(0)}) \right]^2}{\sum_{i \in C_k} N_{ij}^{(0)}}. \tag{3.9}$$

These score statistics are unsigned. In the case of a quantitative or two-class outcome, a signed score statistic can be obtained (see Supplementary Material in Section 1, available at *Biostatistics* online).

We have illustrated by simulation (Supplementary Material in Section 8, available at *Biostatistics* online) that when the Poisson log-linear model holds exactly, the empirical sampling distribution of the above score statistic for a null gene seems to closely follow the chi-squared law (with appropriate degrees of freedom).

4. ESTIMATION OF THE FDR

We now consider the standard permutation plug-in estimate for FDR (see e.g. Tusher *and others*, 2001; Storey, 2002; Storey and Tibshirani, 2003), and we show that this approach must be modified in order to yield accurate results for the Poisson model.

The usual permutation plug-in estimate for FDR is as follows:

1. Compute (unsigned) statistics $S_1, S_2, \ldots, S_m$ based on the data.
2. Permute the n outcome values B times. In the bth permutation, compute statistics $S_1^b, S_2^b, \ldots, S_m^b$ based on the permuted data.
3. For a range of values of a cut-point C for the statistic, let $\hat{R} = \sum_{j=1}^{m} I_{(S_j > C)}$, $\hat{V} = \frac{\hat{\pi}_0}{B} \sum_{j=1}^{m} \sum_{b=1}^{B} I_{(S_j^b > C)}$.
4. Estimate the FDR at a cut-point C by $\hat{\text{FDR}}_C = \hat{V}/\hat{R}$.

In Step 3 above, $\hat{\pi}_0$ is an estimate of π_0, the true proportion of null genes in the population. The estimation is typically made by comparing the numbers of observed and permutation statistics that fall in the smaller (nonsignificant) range of values. In particular, the usual estimate is $\hat{\pi}_0 = \sum_{j=1}^{m} I_{(S_j \leqslant q_{2\zeta})}/(m(1-2\zeta))$. Here $q_{2\zeta}$ is the 2ζ quantile of the distribution of permuted values S_j^b; typically $\zeta = 0.25$ is used. Notice that this is a "pooled" estimate of the permutation distribution, which uses permutation values from all genes to estimate the null distribution for all genes.

The use of the pooled permutation distribution to estimate FDR works well when S_j is a two-sample t-statistic computed from Gaussian data (i.e. each N_{ij} is normally distributed). However, a problem arises in the Poisson setting considered in this paper since the null and non-null genes have very different permutation distributions. To illustrate this point, we generated two-class data from a Poisson model with $n = 12$ and $m = 20\,000$ and 2000 of the genes differentially expressed (the exact details of the simulation are given in the next section). Figure 2 shows histograms of the actual null distribution of the score statistics (left), the permutation distribution of the score statistics using the non-null genes (middle) and null genes (right). We see that while the permutation distribution from null genes is very similar to the actual null distribution, the permutation distribution from non-null genes has much heavier tails. This is likely due to the dependence between the mean and variance in the Poisson distribution. As a result, the estimate of FDR based on the permutation distribution of all genes greatly overestimates the true FDR. (see Sections 5 and 6 for a detailed simulation study and Supplementary Material Sections 2 and 3, available at *Biostatistics* online, for more details and a theoretical analysis.) Thus, we would like to use the

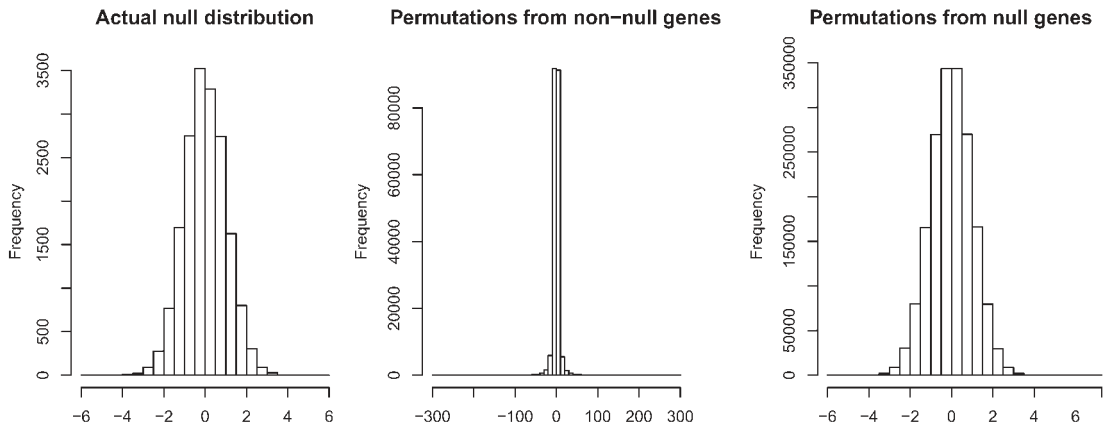

Fig. 2. Histograms of score statistics for simulated data with a two-class outcome. Here, we use the signed version of the score statistic for a more clear display. The permutation distribution of the non-null genes (middle) is much wider than the permutation distribution of the null genes (right), which is very similar to the true null distribution (left).

permutation distribution from the null genes only; however, this requires knowledge of which genes are truly null. So, instead we use the permutations from genes whose observed score is small.

The details are as follows. As before, we permute the outcome B times, producing statistics S_j^b for genes $j = 1, 2, \ldots, m$ and permutations $b = 1, 2, \ldots, B$. Then to estimate π_0, we let M be the indices of the values of S_j falling in the $(\eta, 1-\eta)$ quantiles of its distribution. Let $q'_{2\zeta}$ be the 2ζ quantile of the collection of permuted values $S_j^b, b = 1, 2, \ldots, B, j \in M$. Then, $\hat{\pi}_0' = \sum_{j=1}^m I_{(S_j \leqslant q'_{2\zeta})}/(m(1-2\zeta))$. We use $\zeta = \eta = 0.25$ in this paper.

Finally, we let D denote the indices of the values of S_j that are less than the $\hat{\pi}_0'$ quantile of its distribution; set D contains genes that are likely to be null. To estimate the FDR for a cut-point C for the score statistic, we compute

$$\hat{R} = \sum_{j=1}^{m} I_{(S_j > C)}, \quad \hat{V}' = \frac{1}{B} \sum_{j \in D} \sum_{b=1}^{B} I_{(S_j^b > C)}, \quad \hat{\text{FDR}}_C = \hat{V}'/\hat{R}. \tag{4.1}$$

We study the accuracy of this estimate in the next section. We refer to the procedure outlined in Sections 3 and 4 as "PoissonSeq."

5. A SIMULATION STUDY FOR POISSON-DISTRIBUTED DATA

5.1 *Simulation design*

We now present the results of a simulation study to evaluate PoissonSeq under the Poisson log-linear model of Section 3. Details of the simulation setup are in Supplementary Materials (available at *Biostatistics* online) Section 6, but we provide a brief overview here. There are $m = 20\,000$ genes and $n = 12$ experiments. We generate d_i and β_j, so that the total number of counts per gene and the total number of counts per experiment roughly match the numbers observed in real RNA-Seq data sets. Ten percent of the genes are taken to be non-null; of these non-null genes, 80% are upregulated and the other 20% are downregulated.

5.2 *A comparison of FDR estimates*

In the top three panels of Figure 3, true and estimated FDR curves obtained using PoissonSeq are shown (averaged over 100 simulations) for different outcome types. We also show the FDR estimates obtained using the usual permutation plug-in strategy and the theoretical p-value approach (i.e. we convert the score statistics to p-values using their asymptotic distributions and then use the R package "qvalue" (Storey and Tibshirani, 2003) to convert p-values to FDRs). The true FDR curves for these three methods are the same since all use the score statistics defined by (3.8) and (3.9). Both our FDR estimation method and the theoretical p-value strategy accurately estimate FDR, whereas the usual permutation plug-in method substantially overestimates FDR for reasons discussed in the previous section. Note that in this simulation, the theoretical p-value approach estimates FDR accurately since the p-values upon which it relies are accurate. However, in Section 6, we will show that reliance on theoretical p-values can lead to poor estimation of FDR if the data violates the Poisson model.

Next, we compare the true and estimated FDR curves for PoissonSeq with those obtained using four popular methods. Since most existing methods can be applied only to a two-class outcome, we limit the comparison to that setting. The methods to which we compare PoissonSeq are (The version numbers of R packages we used are listed in Supplementary Material in Section 14, available at *Biostatistics* online): (1) SAM applied to the square root of the total-count normalized data. SAM refers to the modified t-statistic approach of Tusher *and others* (2001). This is similar to the normal distribution–based model of 't Hoen

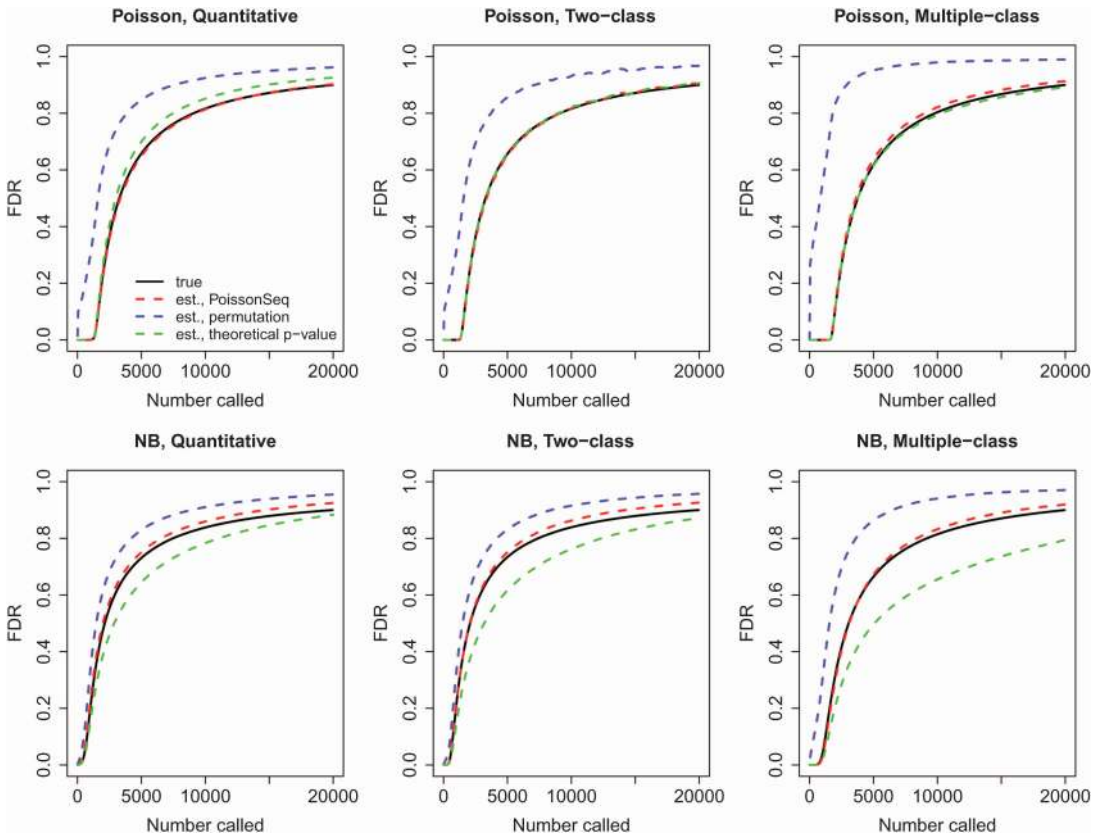

Fig. 3. FDR curves for simulated (top three panels) Poisson distributed data and (bottom three panels) negative binomial distributed data. The left, middle, and right panels show results from data with different types of outcome (averaged over 100 simulations). The solid curves show the true FDRs; the broken curves are estimates. All curves are based on the score statistic (3.8 and 3.9) but use different methods to estimate FDRs: PoissonSeq, the usual permutation plug-in method (permutation), and the theoretical p-value method (theoretical p-value). The true FDR curve, which is the same for all three procedures, is also shown. In the Poisson case, both PoissonSeq and the theoretical p-value method give much more accurate FDR estimates than the usual permutation plug-in method. In the negative binomial case, the PoissonSeq estimate of FDR is much more accurate than the other two estimates.

and others (2008). (2) The method of Marioni *and others* (2008), implemented in the R package DEGSeq (Wang *and others*, 2010). This method uses the same Poisson log-linear model as in PoissonSeq, but the experiments are normalized using total counts, and theoretical p-values combined with the method of Benjamini and Hochberg (1995) are used to estimate FDRs. (3) The edgeR method with total count normalization. This method assumes a negative binomial distribution for the counts. (4) The edgeR method with TMM normalization. Again, a negative binomial distribution is assumed.

The resulting true and estimated FDR curves are shown in the left panel of Figure 4. Our method and edgeR with TMM normalization yield almost identical true FDRs, which are lower than those obtained using the other methods. Our method accurately estimates FDR, whereas edgeR with TMM normalization overestimates it. SAM on square-root transformed data substantially overestimates the FDR, since it uses the usual permutation-based plug-in estimate, which is conservative. Marioni *and others* (2008) and edgeR with total-count normalization severely underestimate FDR.

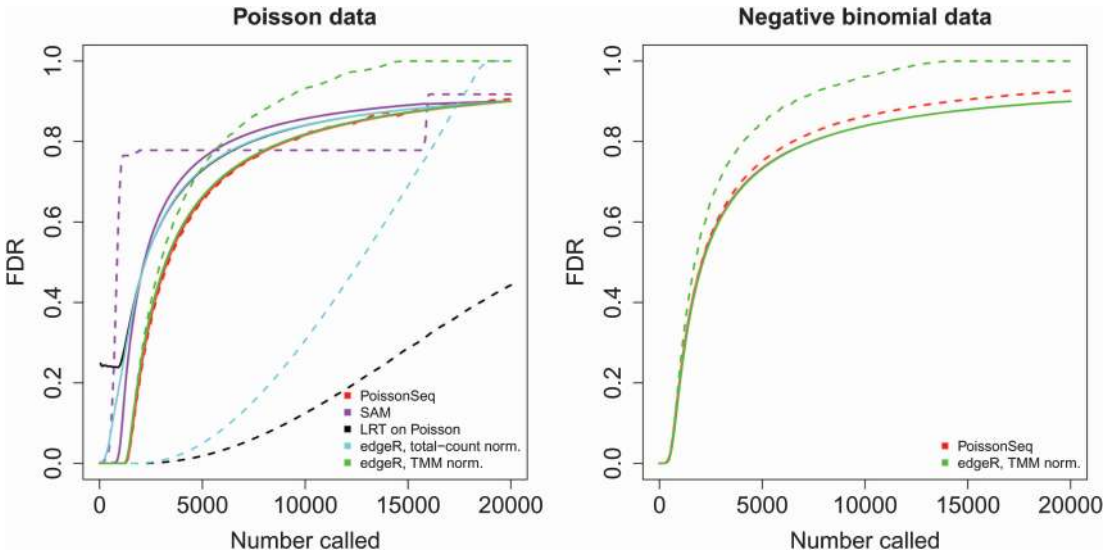

Fig. 4. FDR curves for simulated (left) Poisson-distributed data and (right) negative binomial distributed data. The solid curves show the true FDRs; the broken curves are estimates. These are results (averaged over 100 simulations) on data with two-class outcome using different methods: our method (PoissonSeq), SAM applied to the square root of total-count normalized data (SAM), the method proposed by Marioni *and others* (2008) (LRT on Poisson), edgeR with the default total-count normalization (edgeR, total-count norm.), and edgeR with TMM normalization (edgeR, TMM norm). In the Poisson case, we see that only edgeR with TMM normalization and our PoissonSeq method yield accurate FDR estimates. PoissonSeq and edgeR with TMM normalization also yield much lower true FDRs than the other methods. In the negative binomial case, only the results using PoissonSeq and edgeR with TMM normalization are shown. We see that the true FDR curves of the two methods are almost the same, while our estimate is more accurate.

Interestingly, edgeR with two different types of normalization gives quite different results, in terms of both true FDRs and estimated FDRs. This indicates that improper normalization (specifically, total-count normalization) cannot only give more (true) false positives, as previously reported (Robinson and Oshlack, 2010; Bullard *and others*, 2010), but can also seriously underestimate FDRs.

Since the other three methods we considered do not accurately estimate FDR, hereafter, we will limit our comparison of PoissonSeq to edgeR with TMM normalization.

5.3 *A comparison of normalization approaches*

In the previous section, we noted that the choice of normalization can have a major effect. We now evaluate the accuracy of different methods for estimating sequencing depth. Suppose that the true and estimated sequencing depths are $d_1, \ldots, d_n$ and $\hat{d}_1, \ldots, \hat{d}_n$, respectively. Without loss of generality, we assume $\sum_{i=1}^n d_i = \sum_{i=1}^n \hat{d}_i = 1$. We use the chi-square distance $d = \sum_{i=1}^n \frac{(\hat{d}_i - d_i)^2}{d_i}$, which seems to be a reasonable choice for measuring differences in percentages. The results are shown in Table 1. We see that the error of our normalization procedure is about 1% of the errors of TMM normalization or normalization used by DESeq, although TMM normalization and normalization used by DESeq are more accurate than total-count normalization and quantile normalization.

We also simulated data with the same mean function but following a negative binomial distribution (simulation details are in the next section). The results are also shown in Table 1. Interestingly, our nor-

Table 1. *The performance of different methods for estimating the sequencing depth*

	Poisson data		Negative binomial data	
	Mean of error	SE of error	Mean of error	SE of error
Total count	4.0×10^{-3}	4.0×10^{-3}	5.3×10^{-3}	5.2×10^{-3}
Quantile	7.9×10^{-4}	1.3×10^{-4}	8.0×10^{-4}	1.7×10^{-4}
TMM	8.1×10^{-5}	1.3×10^{-5}	8.9×10^{-4}	1.3×10^{-4}
DESeq	5.6×10^{-5}	4.8×10^{-6}	5.2×10^{-4}	9.1×10^{-5}
Ours	6.2×10^{-7}	2.5×10^{-7}	1.5×10^{-4}	6.9×10^{-5}

malization procedure, which assumes a Poisson model for the data, still gives the smallest error among the five methods.

Our method, the TMM method, and the DESeq method for estimating sequencing depth implicitly assume that a minority of the genes are differentially expressed. To explore whether our method still outperforms other methods when a substantial fraction of genes are non-null, we simulate data with 50% of genes differentially expressed, in both the Poisson setting and the negative binomial ($\phi = 0.25$) setting. As shown in Supplementary Table 1 (available at *Biostatistics* online), our method still substantially outperforms all the other methods.

6. Overdispersed data

The log-linear model (3.1, 3.2) assumes that counts are sampled from a Poisson distribution, leading to a simple form of the score statistic. However, counts in real data sets may be overdispersed, especially for data from biological replicates. To apply our method to overdispersed data, we use a simple transformation to make the data follow a Poisson distribution more closely and then apply our method to the transformed data.

We define the overall overdispersion of the data as $\mathcal{O} = \sum_{j \in S} \mathrm{GOF}_j - (1 - 2\epsilon)(n - 1)(m - 1)$, where GOF_j is defined by (3.7) and S is the set of genes whose GOF is in the $(\epsilon, 1 - \epsilon)$ quantile of all GOF_j's. If the data is Poisson distributed and the genes are independent of each other, then $\sum_{j \in S} \mathrm{GOF}_j$ should approximately follow a χ^2 distribution with $(1 - 2\epsilon)(n - 1)(m - 1)$ degrees of freedom. Thus, the expectation of $\mathcal{O}$ is approximately 0. Therefore, we seek a power transformation $N_{ij} \leftarrow N_{ij}^{\theta}$, where θ is a constant, such that $\mathcal{O} \simeq 0$ so that the data approximately fits the model (3.1, 3.2). The transformed data does not take on integer values, but the formulas from the previous section can still be applied.

Simulations show that our method gives a very accurate estimate of θ if the data are truly a power of the Poisson distribution (Supplementary Material Section 9, available at *Biostatistics* online). A more realistic model for overdispersed RNA-Seq data is a negative binomial distribution, in which the mean μ and variance σ^2 are linked by $\sigma^2 = \mu + \phi\mu^2$, where ϕ is the "dispersion" parameter. The Poisson distribution is a special case of the negative binomial distribution with $\phi = 0$. We generated negative binomial data with a particular mean and dispersion and used our method to estimate θ. The resulting transformed data have approximately a Poisson distribution (Supplementary Material Section 10, available at *Biostatistics* online).

In the description above, we assume that θ is the same constant for all genes. However, this assumption may be too restrictive. Simulations suggest (see Supplementary Material Section 11, available at *Biostatistics* online) that when ϕ is a constant, the relationship between $\log \mu$ and the proper value of θ^{-1} is roughly linear, and larger μ requires larger θ^{-1} or smaller θ. Therefore, instead of using the same θ for all genes, we divide the genes into 10 groups according to the value of $N_{\cdot j}$, and we estimate a value of θ

for each group of genes. Then, we fit a natural cubic spline (or a straight line if we assume the dispersion of each gene is the same) for these 10 pairs of θ^{-1} and $\log N_{\cdot j}$ and use it to predict θ for each gene.

To evaluate the performance of this approach, we generated counts from the negative binomial distribution with $\phi = 0.25$. The means for the cells are the same as for the Poisson data in the previous section. The bottom three panels of Figure 3 display the FDR curves for different outcome types and different methods for estimating FDR: our PoissonSeq procedure, the usual permutation plug-in method, and the theoretical p-value method. We see that our PoissonSeq method is the only one that gives accurate FDR estimates. The right panel of Figure 4 compares our PoissonSeq method and edgeR with TMM normalization (with the dispersion set as constant). We find that the true FDR curves are almost identical. While our PoissonSeq method gives uniformly accurate estimates of FDR curve, edgeR overestimates the higher part of the curve.

7. PERFORMANCE ON REAL DATA SETS

We applied our method to the RNA-Seq data set in Marioni *and others* (2008). This data set contains 5 human kidney samples and 5 human liver samples. The reads were mapped to all human genes. After filtering genes with no more than 5 reads total, we are left with 18 228 genes. We applied our PoissonSeq method and edgeR with TMM normalization to the data. The estimated FDR curves are shown in the left panel of Figure 5. Both methods suggest that the FDR is almost 0 for the $\sim$11 000 most significant genes. However, the estimated FDRs become quite different if more than 11 000 genes are called significant. When all genes are called significant, PoissonSeq's FDR estimate is about 0.34 and that of edgeR is 1. The latter estimate is inconsistent with the estimate that the first 11 000 genes are essentially all differentially expressed, since if we assume that this is true, then the FDR for all 18 228 genes cannot exceed $1 - 11\,000/18\,228 \simeq 0.40$. (In fact, this is consistent with the estimate given by PoissonSeq.)

We also applied our method to the Tag-Seq data set in 't Hoen *and others* (2008). The technology of Tag-Seq experiments is slightly different from RNA-Seq, but they both involve identifying differentially expressed genes on the basis of sequencing data. We choose this data set since it is known to be overdispersed (see the manual of the edgeR package). This data set contains 4 samples in each group, and the

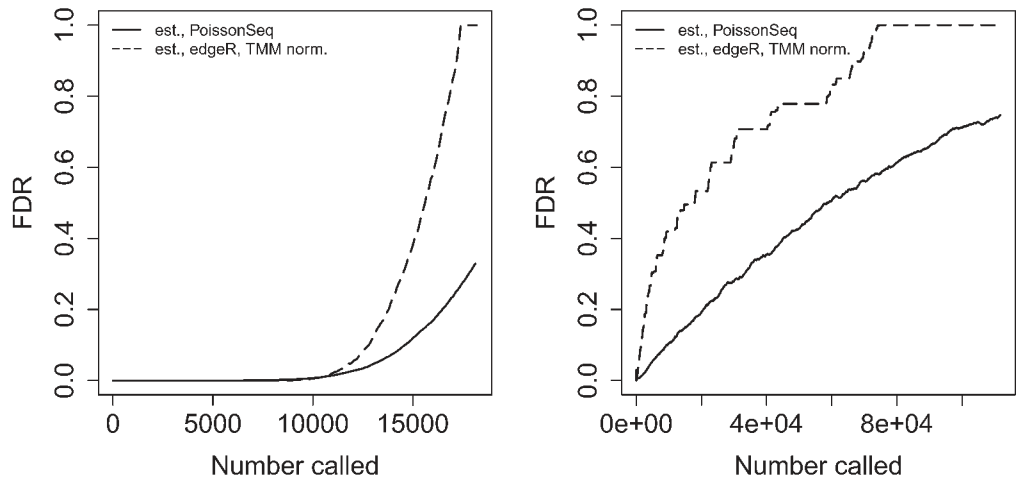

Fig. 5. FDR curves estimated by PoissonSeq and edgeR with TMM normalization for two data sets: (left) the data set from Marioni *and others* (2008) and (right) the data set from 't Hoen *and others* (2008).

counts are mapped to 844 316 tags. 111 809 tags are left for analysis after tags with less than 5 total reads are filtered. The right panel of Figure 5 shows the FDR estimates for PoissonSeq and edgeR with TMM normalization. The FDR estimates from PoissonSeq are lower than those from edgeR. Since this is real data, we do not know which estimate is more accurate.

8. Discussion

We have proposed a Poisson log-linear model for sequencing data in the case of a general outcome type. In the case of a two-class outcome, our model is equivalent to past proposals in the literature; however, our model also accommodates other outcome types. We propose a new and more accurate method for estimating sequencing depth, which is known to play a crucial rule in the performance of any model for sequencing data (Robinson and Oshlack, 2010). Furthermore, we proposed a new method for estimating FDR. Our proposed method does not rely on theoretical p-values, which can be grossly inaccurate for sequencing data, and it also overcomes the poor behavior of the traditional permutation-based plug-in estimates for FDR in the Poisson setting. We have shown that our method outperforms competitors in terms of both true and estimated FDR.

Our method is fast: on a two-class problem, FDR estimation for a data set with 20 000 genes and 12 experiments on the basis of 100 permutations takes ∼15 s on a Windows 7 laptop with a 2.40 GHz processor and 2 GB of memory. As a comparison, edgeR takes ∼2.5 min. Our method is available in an extension package called "PoissonSeq" for the R statistical environment (R Development Core Team, 2011).

Our method is based on the Poisson distribution, and possible overdispersion in the data is handled by taking a power transformation. Since the likelihood of the Poisson distribution is very simple, the Poisson log-linear model is easily extended to different outcome types and potentially to different experimental designs. On the other hand, the negative binomial model rarely has explicit solutions, and estimation of the dispersion parameter by current algorithms for RNA-Seq differential expression analysis like edgeR and DESeq require duplicates (multiple experiments with the same outcome), which are not generally available for quantitative outcomes. To more accurately model RNA-Seq data, other aspects of the data could be incorporated in the Poisson log-linear model, such as transcript length bias (Oshlack and Wakefield, 2009), bias in sequencing rates due to nucleotide content (see e.g. Li *and others*, 2010; Hansen *and others*, 2010; Srivastava and Chen, 2010), and more. We leave this to future work.

We have proposed a new procedure for estimating FDR. This approach attempts to exclude non-null genes from the pooled permutation distribution used in FDR estimation. This procedure could potentially also improve FDR estimation for other problems that are characterized by a mean–variance dependence in the underlying distribution.

Our method for estimating FDRs has some limitations. Firstly, when dealing with overdispersed data, we assume that the transformation power θ depends only on the gene expression. Secondly, we assume that the libraries are totally exchangeable so that permutation gives equivalent data sets under the null hypothesis (see Supplementary Material Section 13, available at *Biostatistics* online). Third, our approach, like other methods from the literature, assumes that all genes are independent from each other (see Supplementary Materials Section 12). Fourth, there are a number of tunable parameters in the procedure; the simple default choices used in this paper may not perform well in all cases. In recent years, DNA sequencing, chromatin immunoprecipitation sequencing, and other approaches related to RNA sequencing have risen in popularity. The methods that we have proposed should be applicable to many of these related technologies.

Supplementary material

Supplementary material is available at http://biostatistics.oxfordjournals.org.

ACKNOWLEDGMENTS

We thank the editor and two referees for helpful comments. An associate editor's suggestions greatly improved the quality of the manuscript. We also thank Art B. Owen for fruitful discussions. *Conflict of Interest:* None declared.

FUNDING

National Science Foundation (DMS-9971405 to R.T.); National Institutes of Health (N01-HV-28183 to R.T., BIB R01EB1988 to I.M.J.).

REFERENCES

AGRESTI, A. (2002). *Categorical Data Analysis*, 2nd edition. New York: Wiley.

ANDERS, S. AND HUBER, W. (2010). Differential expression analysis for sequence count data. *Genome Biology* **11**, R106.

BAGGERLY, K. A., DENG, L., MORRIS, J. S. AND ALDAZ, C. M. (2004). Overdispersed logistic regression for sage: modelling multiple groups and covariates. *BMC Bioinformatics* **5**, 144.

BENJAMINI, Y. AND HOCHBERG, Y. (1995). Controlling the false discovery rate: a practical and powerful approach to multiple testing. *Journal of the Royal Statistical Society Series B* **85**, 289–300.

BLOOM, J. S., KHAN, Z., KRUGLYAK, L., SINGH, M. AND CAUDY, A. A. (2009). Measuring differential gene expression by short read sequencing: quantitative comparison to 2-channel gene expression microarrays. *BMC Genomics* **10**, 221.

BROWN, P. O. AND BOTSTEIN, D. (1999). Exploring the new world of the genome with DNA microarrays. *Nature genetics* **21**, 33–37.

BULLARD, J. H., PURDOM, E., HANSEN, K. D. AND DUDOIT, S. (2010). Evaluation of statistical methods for normalization and differential expression in mRNA-seq experiments. *BMC Bioinformatics* **11**, 94.

DERISI, J. L., IYER, V. R. AND BROWN, P. O. (1997). Exploring the metabolic and genetic control of gene expression on a genomic scale. *Science* **278**, 680–686.

DUDOIT, S., YANG, Y. H., CALLOW, M. J. AND SPEED, T. P. (2002). Statistical methods for identifying differentially expressed genes in replicated cDNA microarray experiments. *Statistica Sinica* **12**, 111–139.

EISEN, MICHAEL AND BROWN, PATRICK. (1999). DNA arrays for analysis of gene expression. *Methods in Enzymology* **303**, 179–205.

HANSEN, K. D., BRENNER, S. E. AND DUDOIT, S. (2010). Biases in illumina transcriptome sequencing caused by random hexamer priming. *Nucleic Acids Research* **38**, e131.

HARDCASTLE, T. J. AND KELLY, K. A. (2010). bayseq: Empirical Bayesian methods for identifying differential expression in sequence count data. *BMC Bioinformatics* **11**, 422.

KERR, M. K., MARTIN, G. AND CHURCHILL, G. (2000). Analysis of variance for gene expression microarray data. *Journal of Computational Biology* **7**, 819–837.

LI, J., JIANG, H. AND WONG, W. H. (2010). Modeling non-uniformity in short-read rates in RNA-seq data. *Genome Biology* **11**, R50.

LU, J., TOMFOHR, J. K. AND KEPLER, T. B. (2005). Identifying differential expression in multiple sage libraries: an overdispersed log-linear model approach. *BMC Bioinformatics* **6**, 165.

MARIONI, J. C., MASON, C. E., MANE, S. M., STEPHENS, M. AND GILAD, Y. (2008). Rna-seq: an assessment of technical reproducibility and comparison with gene expression arrays. *Genome Res* **18**, 1509–1517.

MORTAZAVI, A., WILLIAMS, B. A., MCCUE, K., SCHAEFFER, L. AND WOLD, B. (2008). Mapping and quantifying mammalian transcriptomes by RNA-seq. *Nature Methods* **5**, 621–628.

NAGALAKSHMI, U., WONG, Z., WAERN, K., SHOU, C., RAHA, D., GERSTEIN, M. AND SNYDER, M. (2008). The transcriptional landscape of the yeast genome defined by RNA sequencing. *Science* **302**, 1344–1349.

NEWTON, M. A., KENDZIORSKI, C. M., RICHMOND, C. S., BLATTER, F. R. AND TSUI, K. W. (2001). On differential variability of expression ratios: improving statistical inference about gene expression changes from microarray data. *Journal of Computational Biology* **8**, 37–52.

OSHLACK, A. AND WAKEFIELD, M. J. (2009). Transcript length bias in RNA-seq data confounds systems biology. *Biology Direct* **4**, 14.

R DEVELOPMENT CORE TEAM. (2011). *R: A Language and Environment for Statistical Computing*. Vienna, Austria: R Foundation for Statistical Computing.

ROBINSON, M. D., MCCARTHY, D. J. AND SMYTH, G. K. (2010). edger: a bioconductor package for differential expression analysis of digital gene expression data. *Bioinformatics* **26**, 139–140.

ROBINSON, M. D. AND OSHLACK, A. (2010). A scaling normalization method for differential expression analysis of RNA-seq data. *Genome Biology* **11**, R25.

ROBINSON, M. D. AND SMYTH, G. K. (2007). Moderated statistical tests for assessing differences in tag abundance. *Bioinformatics* **23**, 2881–2887.

ROBINSON, M. D. AND SMYTH, G. K. (2008). Small-sample estimation of negative binomial dispersion, with applications to sage data. *Biostatistics* **9**, 321–332.

SHENDURE, J. (2008). The beginning of the end for microarrays? *Nature Methods* **5**, 585–587.

SPELLMAN, P. T., SHERLOCK, G., IYER, V. R., ZHANG, M., ANDERS, K., EISEN, M. B., BROWN, P. O. AND BOTSTEIN, B. D. AND FUTCHER, B. (1998). Comprehensive identification of cell cycle-regulated genes of the yeast saccharomyces by microarray hybridization. *Molecular Cell Biology* **9**, 3273–3975.

SRIVASTAVA, S. AND CHEN, L. (2010). A two-parameter generalized poisson model to improve the analysis of RNA-seq data. *Nucleic Acids Research* **38**, e170.

STOREY, J. (2002). A direct approach to false discovery rates. *Journal of the Royal Statistical Society B* **64**, 479–498.

STOREY, J. (2003). The positive false discovery rate: A Bayesian interpretation and the q-value. *Annals of Statistics* **31**, 2013–2025.

STOREY, J. AND TIBSHIRANI, R. (2002). SAM thresholding and false discovery rates for detecting differential gene expression in DNA microarrays. In: Parmigiani, G., Garrett, E. S., Irizarry, R. A. and Zeger, S. L. (editors), *The Analysis of Gene Expression Data: Methods and Software.* New York: Springer.

STOREY, J. AND TIBSHIRANI, R. (2003). Statistical significance for genomewide studies. *Proceedings of the National Academy of Sciences of the United States of America* **100**, 9440–9445.

'T HOEN, P. A. C., ARIYUREK, Y., THYGESEN, H. H., VREUGDENHIL, E., VOSSEN, R. H., DE MENEZES, R. X., BOER, J. M., VAN OMMEN, G. J. AND DEN DUNNEN, J. T. (2008). Deep sequencing-based expression analysis shows major advances in robustness, resolution and inter-lab portability over five microarray platforms. *Nucleic Acids Research* **36**, e141.

TUSHER, V., TIBSHIRANI, R. AND CHU, G. (2001). Significance analysis of microarrays applied to transcriptional responses to ionizing radiation. *Proceedings of the National Academy of Sciences of the United States of America* **98**, 5116–5121.

WANG, L., FENG, Z., WANG, X. AND ZHANG, X. (2010). Degseq: an r package for identifying differentially expressed genes from RNA-seq data. *Bioinformatics* **26**, 136–138.

WANG, Z., GERSTEIN, M. AND SNYDER, M. (2009). Rna-seq: a revolutionary tool for transcriptomics. *Nature Reviews Genetics* **10**, 57–63.

WILHELM, B. T. AND LANDRY, J. R. (2009). RNA-Seq—quantitative measurement of expression through massively parallel RNA-sequencing. *Methods* **48**, 249–257.

WITTEN, D., TIBSHIRANI, R., GU, S. G., FIRE, A. AND LUI, W. O. (2010). Ultra-high throughput sequencing-based small RNA discovery and discrete statistical biomarker analysis in a collection of cervical tumours and matched controls. *BMC Biology* **8**, 58.

[*Received November 2, 2010; revised August 26, 2011; accepted for publication August 27, 2011*]