

A COMPARISON OF STATISTICAL METHODS FOR DETECTING DIFFERENTIALLY EXPRESSED GENES FROM RNA-SEQ DATA¹

VANESSA M. KVAM, PENG LIU², AND YAQING SI

Department of Statistics, Iowa State University, Snedecor Hall, Ames, Iowa 50011-1210 USA

RNA-Seq technologies are quickly revolutionizing genomic studies, and statistical methods for RNA-seq data are under continuous development. Timely review and comparison of the most recently proposed statistical methods will provide a useful guide for choosing among them for data analysis. Particular interest surrounds the ability to detect differential expression (DE) in genes. Here we compare four recently proposed statistical methods, edgeR, DESeq, baySeq, and a method with a two-stage Poisson model (TSPM), through a variety of simulations that were based on different distribution models or real data. We compared the ability of these methods to detect DE genes in terms of the significance ranking of genes and false discovery rate control. All methods compared are implemented in freely available software. We also discuss the availability and functions of the currently available versions of these software.

Key words: baySeq; DESeq; edgeR; false discovery rate; RNA-seq; statistical test; two-stage Poisson model.

The advent of next-generation sequencing (NGS) technologies has revolutionized the way genomic study is progressing. One important application of NGS technologies is to study the transcriptome, and the resulting technology is called RNA-seq. RNA-seq technologies have been employed for studying different plants including, for example, maize (Li et al., 2010), soybean (Severin et al., 2010), rice (Zhang et al., 2010; Lu et al., 2010), and *Arabidopsis* (Filichkin et al., 2010). Compared with the hybridization-based microarray technologies that have been the dominate approach to study gene expression in the past decade, RNA-seq technologies offer several advantages, including a wider range of expression levels, more information to detect allele-specific expression, novel promoters, and isoforms, less noise, and higher throughput (Wang et al., 2010; Oshlack et al., 2010). For these reasons, RNA-seq is poised to replace microarray technology and become the major platform to study gene expression in the coming years.

In a typical RNA-seq experiment, a sample of RNA is converted to a library of cDNA fragments and then sequenced on a high-throughput commercially available platform, such as Illumina's Genome Analyzer, Helicos BioSciences' HeliScope, Applied Biosystems' SOLiD, Pacific Biosciences' SMRT or Roche's 454 Life Sciences sequencing systems (Wang et al., 2010). The raw data comprise large amounts of sequences of DNA fragments that are termed reads, and such data undergo a series of steps of analysis. Oshlack et al. (2010) provides an excellent review of the analysis pipeline that includes mapping the reads, summarizing read counts for each gene, normalization, and detecting differentially expressed genes. Also, Table 1 in Oshlack et al. (2010) provides a list of software for each step of analysis. Typically, the reads generated by an RNA-seq study are assigned to a gene (or other class) based on their mapping to

a common region of the target genome or de novo assembled transcriptome. There are several methods, still under research, for quantifying gene expression levels from RNA-seq data (Oshlack et al., 2010; Bullard et al., 2010). Complexities with alternatively spliced transcripts and isoform expression make this an active area of research. While isoform detection is not the focus of this paper, interested readers may refer to Hiller et al. (2009) and Salzman et al. (2011) and their references for estimating isoform abundance for RNA-seq data. Gene is a general term that we adopt throughout the remainder of the paper, which can refer to an exon, a subset of exons or all exons for a gene model. Gene expression is measured by the number of reads mapped to a gene. Thus, RNA-seq results in a discrete measurement for gene expression, which is different from the fluorescence intensity measurement from microarray technologies that has been treated as a continuous variable. Consequently, the statistical methods used to analyze microarray data are not directly applicable, and new statistical approaches that are appropriate to handle the RNA-seq data are urgently needed due to the huge amount of data being generated.

Detecting differentially expressed (DE) genes across treatments/conditions is an essential step and sometimes the major goal in the statistical analysis of RNA-seq data. The identification of DE genes helps to elucidate the function of genes when cells respond to different treatments/conditions. In addition, detecting DE genes can be a prestep for clustering gene expression profiles or testing gene set enrichments. Due to the short history of RNA-seq and its continuous development, there are no standard methods available yet to detect DE genes based on such data. Many statisticians have been working on this subject. Several papers have been published, and more are probably on their way. In this article, we first review the currently available methods to detect DE genes including edgeR (Robinson and Smyth, 2007; Robinson and Smyth, 2008), DESeq (Anders and Huber, 2010), baySeq (Hardcastle and Kelly, 2010), and a method based on a two-stage Poisson model (TSPM, Auer and Doerge, 2011). We provide information on how to download the corresponding packages or codes to apply these methods in freely available software R. We then compare their performance on significance ranking of genes through simulation studies

¹Manuscript received 21 July 2011; revision accepted 1 December 2011.

The authors thank the referees for comments that improved this manuscript. This material is based upon work supported in part by the National Science Foundation (grant no. 0701736).

²Author for correspondence (e-mail: pliu@iastate.edu)

TABLE 1. Mean true positive rate (TPR) and mean true proportion of false positives among discoveries (V/R) with respective replicate size calculated for 100 simulations following simulation setting 1 at controlled 5% false discovery rate (FDR) via the Benjamini and Hochberg (BH) method.

| Replicates | $n = 2$ | | $n = 4$ | |
|------------|---------------|---------------|---------------|---------------|
| | TPR (SE) | V/R (SE) | TPR (SE) | V/R (SE) |
| edgeR | 0.237 (0.11%) | 0.217 (0.16%) | 0.343 (0.13%) | 0.178 (0.13%) |
| DESeq | 0.254 (0.11%) | 0.219 (0.16%) | 0.346 (0.13%) | 0.168 (0.13%) |
| TSPM | 0.322 (0.11%) | 0.281 (0.16%) | 0.366 (0.12%) | 0.136 (0.12%) |

under various settings mimicking real data. We also examine the false discovery rate control of different procedures, which is a necessary step in such a high-dimensional testing problem in genomic data analysis.

Our results have indicated baySeq has the highest true positive rates at low rates of false positives. We also have found that TSPM does not perform nearly as well as the other methods under comparison when sample size is as small as 2. Among those methods that control for the false discovery rate (FDR), we have found situations where the true FDR is actually much larger than expected. Our study allows for transparency in the strengths and weaknesses of these methods, which can be useful for scientists analyzing future data obtained from RNA-seq studies.

REVIEW OF METHODS TO DETECT DIFFERENTIALLY EXPRESSED GENES BASED ON RNA-SEQ DATA

To detect which genes are differentially expressed, statistical hypothesis tests are in order. Many statistical methods have been developed for normalized gene expression measurements from microarray based on normal distributions. For example, the popularly applied moderated- t test implemented in the R package limma (Smyth, 2005) is developed based on normal assumptions. As previously stated, the RNA-seq technology results in discrete measurement of gene expression, and hence, statistical methods developed based on normal distributions are not directly applicable. Logarithm transformation may bring the highly skewed, discrete RNA-seq data closer to normal distribution. However, an arbitrary, small number has to be added to those genes with zero counts in some samples to do the log transformation. Even so, the transformed data may still not fit well with normal distributions. Instead of focusing on finding transformations so that existing methods with microarray analysis can be applied, researchers have developed several methods based on discrete distributions that can be used to model the gene counts directly.

Three discrete probability distributions have been proposed to model the count data from RNA-seq studies: binomial, Poisson, and negative binomial (NB). It can be shown mathematically that if the number of reads is big, which is true for RNA-seq data, and the probability of a read mapped to a given gene is small, then the binomial distribution can be well approximated by the Poisson distribution. In early RNA-seq studies using a single source of RNA, the distribution of counts across technical replicates for the majority of genes was reported to fit well to a Poisson distribution (Marioni et al., 2008; Bullard et al., 2010). However, one property of the Poisson distribution is that

the variance is equal to the mean. When there are biological replicates, RNA-seq data may exhibit more variability than expected by Poisson distribution, i.e., the variance is likely to exceed the mean for considerably many genes (Anders and Huber, 2010). This phenomenon is called overdispersion. For overdispersed data, Poisson-based analyses will be prone to high false positive rates resulting from underestimation of sampling error. Quasi-likelihood approach can be applied because it introduces a scaling factor for the variance so that it is allowed to differ from the mean. Assuming a NB model instead of Poisson is another way to deal with overdispersed data because the NB distribution specifies that the variance is greater than the mean. Because biological replicates are essential in reaching biologically meaningful results, we hope all experiments will be designed to include biological replicates. Because of this, we only review the currently available methods that have the ability to handle overdispersion.

We should mention that methods that were developed for serial analysis of gene expression (SAGE) data could be applied in RNA-seq data analysis. For example, the edgeR method was originally developed for SAGE data analysis and is now applied to RNA-seq data analysis (Robinson and Smyth, 2007). A comprehensive analysis of methods used in analyzing SAGE data are beyond the scope of this paper, but interested readers can refer to Lu et al. (2005), Robinson and Smyth (2008), and Baggerly et al. (2004).

Method based on Poisson distribution—Recently, Auer and Doerge (2011) proposed a method based on a two-stage Poisson model (TSPM). Their rationale is that some genes may demonstrate overdispersion while others may not. Therefore, the first stage of their method is to test for overdispersion for each gene. If such a test suggests overdispersion, a quasi-Poisson likelihood approach is applied to test for DE in the second stage. Otherwise, a test based on Poisson model is used in the second stage. They control the false discovery rate separately on two lists of genes because two different methods to assess the significance of DE are applied to different genes.

There is an interesting paper by Srivastava and Chen (2010) that employs a generalized Poisson (GP) distribution to model the position-level read counts. Their method, implemented in the GPseq package, takes into consideration the potential positional bias in conducting DE analysis, which is different from all other methods reviewed here. The real data sets that are accessible to us do not offer the position-level counts, while the gene-wise counts cannot be handled by the GPseq package. Therefore, the GPseq method is not included in our analysis.

Methods based on NB distribution—Three methods that are implemented in R packages, edgeR, DESeq, and baySeq, are based on the NB model. The method used in edgeR (Robinson and Smyth, 2007, 2008) is the first proposed one, and it was originally developed for SAGE data, which can be considered as a smaller scale of RNA-seq data. The negative binomial distribution has been used as a natural extension of the Poisson distribution, requiring an additional dispersion parameter and allows extra variability than the Poisson distribution does. Because costly replicates induce very small sample sizes for RNA-seq studies, the estimation for the dispersion parameter is a challenging issue. Robinson and Smyth (2008) proposed using a common dispersion for all genes to achieve a better estimate of the dispersion parameter. If the assumption that the dispersion parameter, which measures the extra variance

compared with the mean, is the same across all genes, the common dispersion parameter can be estimated very accurately because a lot of data are used for this estimation. However, a common dispersion for all genes might rarely be an appropriate assumption in practice. Perhaps a better strategy is to allow different genes to have different dispersion parameters, while the estimation of these dispersion parameters can be improved by borrowing information across genes using some appropriate statistical methods. Such strategies were used in microarray data analysis, and many tests were developed that borrow information across genes to better estimate the variance (for example, Smyth, 2004; Cui et al., 2005) or both the mean expression and the variance (Hwang and Liu, 2010). Those tests were shown to perform favorably when compared with tests that do not borrow information. Following similar strategies, a moderated test was proposed for RNA-seq data and can be implemented in the package edgeR (Robinson and Smyth, 2007).

Anders and Huber (2010) also tried to borrow information across genes to better estimate the dispersion parameter. They assumed a locally linear relationship between variance and the mean expression levels. This assumption allows the variance (or equivalently the overdispersion parameter) to be estimated using pooled data with similar expression levels. This method is executed in DESeq as an R/Bioconductor package. Both edgeR and DESeq provide p -values for the tests based on exact test or approximation of exact test.

The method proposed by Hardcastle and Kelly (2010) also assumes the data follow a NB distribution but differs from the other two methods in the estimate of significance. They follow an empirical Bayes approach that ranks genes by the estimates of the posterior probabilities for a model that defines DE for each gene. The prior distribution on the NB parameters is found using the data. Samples behaving similarly to each other should possess the same prior distribution on the underlying parameters of the gene, while samples behaving differently should possess different prior distributions. This method is applied in the baySeq R/Bioconductor package.

Normalization—For using RNA-seq data to compare expression between samples, normalization is performed to adjust for varying sequencing depths and potentially other technical effects across replicates and is needed in all four methods reviewed above. One example normalizes the read counts by the total number of reads for each sample and the length of the gene. This method quantifies transcript levels in reads per kilobase of exon model per million mapped reads (RPKM) (Mortazavi et al., 2008). However, when conducting DE analysis of the same genes between samples and not to compare genes to genes, the normalization with respect to gene length is not important (i.e., the biases will affect the same gene in the same way in different samples). If we do not consider the gene length, currently available normalization methods can be done with some scaling factor to the mean expression level. The simplest and most commonly used normalization factor is the total number of reads in the library, accounting for the fact that more reads will be assigned to each gene if a sample is sequenced to a greater depth (Oshlack et al., 2010). However, typically the total number of reads is mostly contributed by a small group of abundantly expressed genes (Bullard et al., 2010). If this group of genes is differentially expressed, then using the total number of reads would dramatically affect the results of detecting DE genes. Bullard et al. (2010) compared several normalization methods and found that using the 75th percentile of nonzero

count distribution within each lane as a normalization factor is a more robust choice over the standard total-count normalization, and the overall performance is best among those methods studied. The R package DESeq estimates the normalization factor by the median of scaled counts, which is a similar idea to the 75-percentile normalization. The R/Bioconductor package, edgeR, uses a weighted trimmed mean of the log expression ratios (trimmed mean of M values [TMM]) (Robinson and Oshlack, 2010), which is another robust normalization method. Based on our experience, the 75th percentile and the TMM method perform similarly.

SIMULATION RESULTS

In this section, we present simulation studies to compare the four statistical methods reviewed: TSPM, edgeR, DESeq, and baySeq. Simulation study is one way to examine properties of certain statistical methods. An ideal simulation for our purpose would generate data similar to those produced by real RNA-seq studies. We designed our simulation settings with the goal to mimic, as closely as possible, the data that can occur in an actual RNA-seq study. Given the simulated data, we applied different tests as if the data arise from a real RNA-seq study. The benefit of simulated data is that we know the true underlying mechanism to generate the data; therefore, we can assess the results, for example, whether a declared positive is indeed correct (true positive).

We applied the following packages or codes to implement different tests: Robinson and Smyth's edgeR (version 2.2.5), Anders and Huber's DESeq (version 1.0.6), Hardcastle and Kelly's baySeq (version 1.6.0), and the code for the recently developed TSPM method by Auer and Doerge (available at <http://www.stat.purdue.edu/~doerge/software/TSPM.R>). The packages edgeR, DESeq, and baySeq are available as Bioconductor packages from the Bioconductor web site (<http://bioconductor.org>) (Gentleman et al., 2004). All analyses were implemented in the statistical environment R (version 2.13.1, available at <http://www.R-project.org>). We estimate normalization factors for baySeq and TSPM as the third quartile of counts following Bullard et al. (2010). DESeq and edgeR have their own method of estimating normalization factors built in their respective packages.

We compare and evaluate the analysis results of different statistical methods based on two criteria. First, we look at the significance ranking of genes. Each test reviewed above produces an ordered list of genes based on p -values or posterior probabilities. An ideal method would rank all the truly differentially expressed genes on the top of the list and then genes without change in expression levels. However, if we select, for example, genes with p -values smaller than 0.001 and declare them to be differentially expressed, our findings might include both truly differentially expressed genes (true positives) and nondifferentially expressed genes (false positives). Given a list of declared DE genes and the information about which genes were generated to be true DE and which genes are not, we can calculate the true positive rate (TPR) and false positive rate (FPR). TPR is defined as the proportion of true DE genes that are declared to be DE, while FPR is the proportion non-DE genes that are also declared to be DE. As the number of declared findings change, or equivalently, the threshold for significance changes, different (TPR, FPR) pairs are generated. Plotting TPR vs. FPR results in a receiver operating characteristic (ROC) curve. Methods that rank the genes better would give an ROC curve with higher TPR than others for the same value of FPR. So examining ROC

curves tells us about the ability of the approaches to rank the genes in order of differential expression. Second, we compare the results from different methods after controlling the false discovery rate (FDR). FDR is a multiple testing error, and its control is important for high-dimensional genomic data and has been popularly applied in microarray and RNA-seq data analysis. We compared the three methods that produce p -values: edgeR, DESeq and TSPM. We applied the Benjamini-Hochberg (BH) procedure to control FDR at a 5% level and computed the TPR and the true proportion of false positives out of the discoveries. Because baySeq does not provide p -values and Hardcastle and Kelly (2010) did not discuss how to control FDR with their method, we did not include baySeq under comparison for performance on FDR control.

For each simulation setting described next, 100 independent data sets were generated, and each data set included at least 10000 genes. Reported results were averaged over these 100 data sets.

Simulation 1—We adapted the simulation setting two studied by Auer and Doerge (2011) where approximately half of the genes were derived from Poisson and the other half from overdispersed Poisson with different degrees of overdispersion. This is the one setting closest to real data among the three included in Auer and Doerge (2011). Auer and Doerge (2011) simulated data with four different numbers of replicates. Here we only simulated data with either two or four replicates in each of the two treatment groups because most plant RNA-seq data published to date only have a couple or several replicates.

Figure 1 plots the ROC curves. The results from baySeq generated the best ranking of genes. TSPM is the next best performing method if we only look at the range of small FPRs, but it becomes the poorest if we go beyond 4% of FPR or so. One possible reason for its poor performance is the small number of replicates. The authors also made this observation in their results. All four methods gain power from $n_i = 2$ to $n_i = 4$, with baySeq increasing the most and also performing most powerfully. Applying the Benjamini-Hochberg (BH) procedure to the

list of p -values generated by edgeR, DESeq, and TSPM, respectively, to control FDR at a 5% level, we see the true proportion of false positives out of the discoveries is unfavorably high in all three methods (Table 1). So the FDR is not in fact controlled to the desired level. Note that the standard errors are all very small; hence, the differences between the four methods are statistically significant. This comment also applies to the results from all other simulation settings.

Simulation 2—The second and third settings were generated based on a real data set from a maize study. Li et al. (2010) carefully selected a developing leaf from corn plant that captures multiple stages of photosynthetic differentiation at a single time point. They exploited Illumina sequencing technologies to profile gene expression from four representative sections of the leaf blade and also from the laser-capture microdissected (LCM) bundle sheath and mesophyll cells from the mature zone (tip of the leaf). There were two biological replicates in their studies. We based our simulation settings 2 and 3 on the fitted statistics obtained from LCM samples. For simulation 2, we follow the three steps described next to generate an RNA-seq data set.

Step 1: Estimation—We first estimated the overall mean expression levels across both cell types from the LCM data and fitted a gamma distribution using maximum likelihood approach. The estimated parameters were $\alpha = 0.28$ and $\beta = 666$ (gamma mean = $\alpha\beta = 186.67$). We also estimated the half log of the fold-changes between the bundle sheath and mesophyll cells and then fitted a three-component normal mixture distribution using an expectation-maximization (EM) algorithm. One component is set to be zero, so it corresponds to the non-DE genes. The other components have means 0.96 or -0.96 and the same standard deviation 0.725, so we have both upregulated and downregulated genes.

Step 2: Draw model parameters—The overall mean expression level for each gene j , λ_j , was drawn independently from

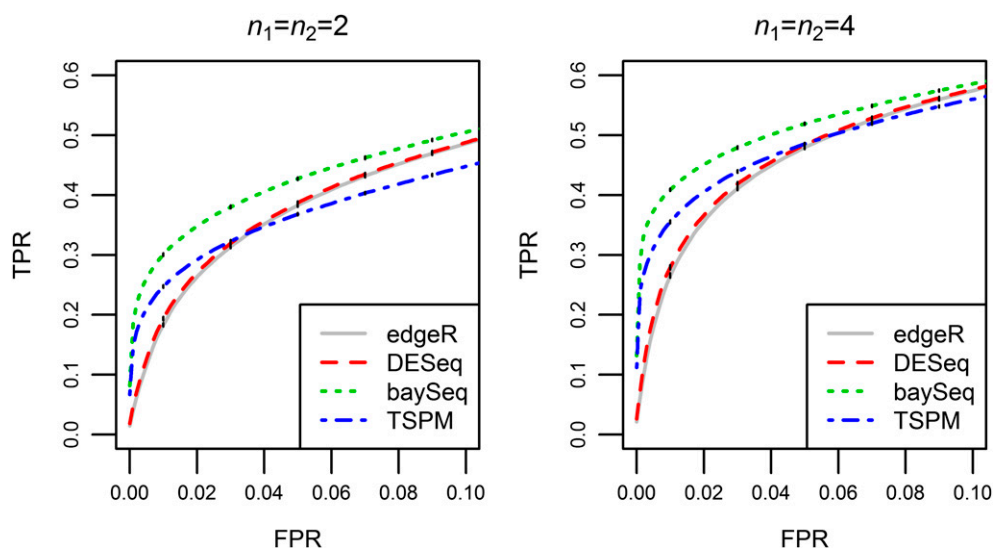


Fig. 1. Mean receiver operating characteristic (ROC) curves comparing the performance of edgeR, DESeq, baySeq, and TSPM to detect differential expression based on 100 simulations from simulation setting 1 where half of the genes follow Poisson distributions and the other half follow overdispersed Poisson distributions. Left panel is from two replicates; right panel is from four replicates. Vertical bars at odd levels of false positive rate (FPR) are ± 2 times the standard error to the value of the estimated corresponding true positive rate (TPR).

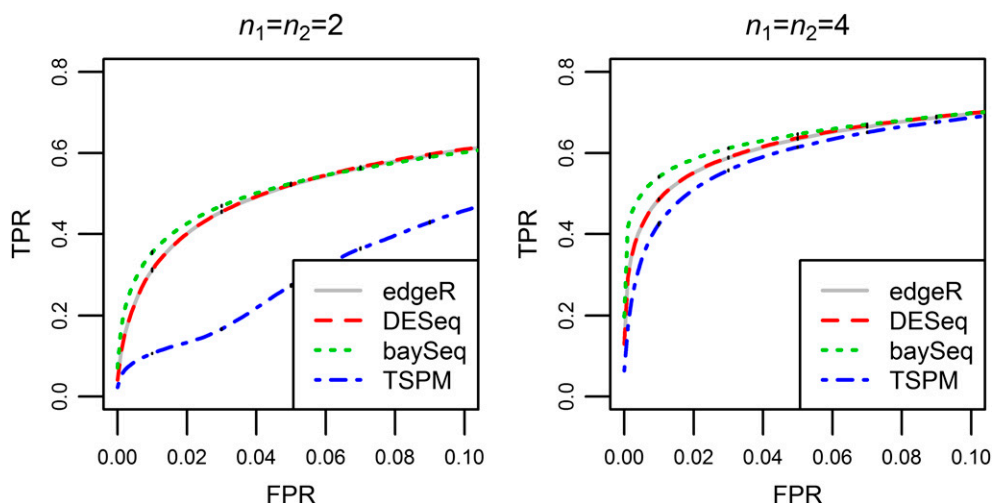


Fig. 2. Mean receiver operating characteristic (ROC) curves, based on 100 simulations from simulation setting 2, comparing the performance of edgeR, DESeq, baySeq, and TSPM to detect differential expression. Simulation based on counts simulated from NB distribution with parameters from empirically estimated distributions from the maize data. Left panel is from two replicates, and right panel is from four replicates. Vertical bars at odd levels of false positive rate (FPR) are ± 2 times the standard error to the value of the estimated corresponding true positive rate (TPR).

the gamma distribution estimated from step 1. Then, the half log fold-change between expression means for each gene j , denoted by δ_j , was drawn independently from the fitted 3-component normal distribution where each component had an equal chance to be chosen. Then, of the simulated δ_j , $j = 1, \dots, 10000$, we randomly selected a third, and set these δ_j to zero whether it was originally zero or not. As a consequence, the amount of non-DE genes varies between 1/3 and 2/3 of the total number of genes. Gene-specific dispersions ϕ_j were simulated as independent and identically distributed draws from a gamma distribution with parameters $\alpha = 0.85$ and $\beta = 0.5$ (gamma mean = $\alpha\beta = 0.425$) following simulations by Hardcastle and Kelly (2010).

Step 3: Simulate count data—Let Y_{ijk} denote the count measure for treatment i , gene j , and experimental unit k within treatment i ($i = 1, 2; j = 1, \dots, 10000; k = 1, \dots, n_i$) where $n_i = 2$ or 4. Then conditional on δ_j , λ_j , and ϕ_j obtained from step 2, the counts, Y_{ijk} , were simulated from NB distribution with parameters mean = $\lambda_j \exp\{(-1)^i \delta_j\}$ and dispersion = ϕ_j , so that variance = $\lambda_{ijk} + \phi_j \lambda_{ijk}^2$.

Steps 2 and 3 were repeated to obtain 100 independently simulated data sets. The average ROC curves plotted in Fig. 2 show that baySeq generates the best ranking of genes for low levels of FPR. The ROC curves corresponding to edgeR and DESeq are indistinguishable for this simulation, and both performed closely to baySeq. The method of TSPM performs noticeably poorer than the other three, especially in the smaller replication (Fig. 2). Because the three methods, except TSPM, are all based on NB distribution, this simulation shows that indeed they perform better when their model assumption is correct. Although TSPM incorporated tests to handle overdispersion, it does not seem to do well for NB-distributed data with small number of replicates. The performances of edgeR and DESeq in terms of FDR control are similar, and the FDR is not controlled but close to the desired level (Table 2). The FDR control for TSPM is not reliable for small number of replicates ($n_i = 2$), but the FDR for TSPM is controlled to the desirable level when the number of replicates equals 4.

Simulation 3—Different from simulation 2, we did not simulate λ_j and δ_j from fitted gamma or normal distributions for simulation 3. Instead, we used the empirically estimated values from the same data (Li et al., 2010).

Step 1: Estimate model parameters—The parameters λ_j and δ_j were empirically estimated from the maize data set. Estimation of λ_j was set to be overall mean across treatment groups for each gene j . To estimate δ_j , we first implemented edgeR to determine DE genes. Given a p -value greater than 0.1, genes were declared not DE, thus, the half log-fold change, δ_j , was set to zero. Those genes with p -values less than 0.01 were declared DE and δ_j was estimated from the empirical half log fold-change. We also applied edgeR to estimate separate dispersions for individual genes using the option “tagwise” available from the package. These tagwise dispersion estimates were used as the dispersion parameter, ϕ_j , applied to the NB model. Given that edgeR was used to estimate parameters for this simulation, there may be a potential that the results could be biased toward edgeR.

Step 2: Simulate count data—Still, let Y_{ijk} denote the count measure for treatment i , gene j , and experimental unit k within treatment i ($i = 1, 2; j = 1, \dots, g; k = 1, \dots, n_i$). Note that we only looked at the case with two replicates ($n_i = 2$), the same as the maize data, for each treatment group. Two sets of counts, Y_{ijk} ,

TABLE 2. Mean true positive rate (TPR) and mean true proportion of false positives among discoveries (V/R) with respective replicate size calculated for 100 simulations following simulation setting 2 at controlled 5% false discovery rate (FDR) via the Benjamini and Hochberg (BH) method.

| Replicates | $n = 2$ | | $n = 4$ | |
|------------|---------------|----------------|---------------|----------------|
| | TPR (SE) | V/R (SE) | TPR (SE) | V/R (SE) |
| edgeR | 0.398 (0.10%) | 0.0596 (0.05%) | 0.565 (0.08%) | 0.0503 (0.03%) |
| DESeq | 0.459 (0.10%) | 0.0793 (0.06%) | 0.591 (0.08%) | 0.0616 (0.04%) |
| TSPM | 0.167 (0.05%) | 0.2472 (0.13%) | 0.491 (0.09%) | 0.0386 (0.04%) |

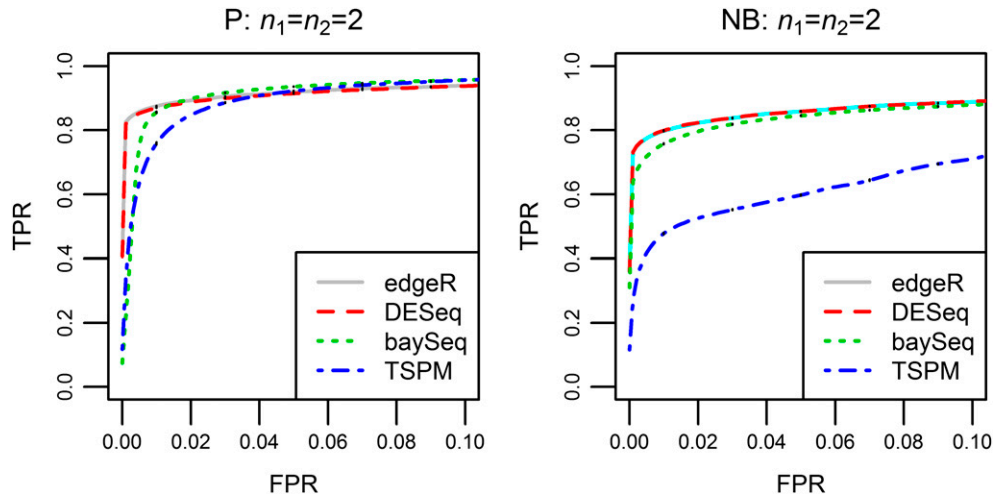


Fig. 3. Mean receiver operating characteristic (ROC) curves, based on 100 simulations from simulation setting 3, comparing the performance of edgeR, DESeq, baySeq, and TSPM to detect differential expression. Simulation based on Poisson (left panel) or NB (right panel) distribution. Parameters for Poisson or NB distributions were empirically estimated from the maize data. Vertical bars at odd levels of false positive rate (FPR) are ± 2 times the standard error to the value of the estimated corresponding true positive rate (TPR).

were simulated according to the data models: $Poisson(\lambda_j, \exp\{(-1)^i \delta_j\})$ and $NB(\lambda_j, \exp\{(-1)^i \delta_j\}, \phi_j)$.

All methods perform comparably well in detecting DE from Poisson-distributed data. However, all suffer a decrease in power when data were simulated from a NB distribution. The difference between TSPM results from Poisson and that from NB distributed data are most noticeable. At 5% FPR, TPR drops 32.5% for TSPM (Fig. 3). At a controlled FDR level of 5%, both edgeR and DESeq give a conservative result with FDR at about half of the desired level for both Poisson and NB cases (Table 3). The FDR control for TSPM, however, still shows a liberal behavior with higher true FDR than the desired level especially for the NB case for which the proportion of false positives among discoveries is 23% instead of the controlled level of 5% (Table 3).

Simulation 4—The final simulation setting utilizes a real RNA-seq data set obtained from 69 lymphoblastoid cell lines (LCL) derived from unrelated Nigerian individuals (Pickrell et al., 2010). Each sample was sequenced at two separate laboratories (Argonne and Yale) on Illumina Genome Analyzer II instruments, but the two laboratories generated reads with different lengths. For our simulation, we only selected one lane for each individual from those sequenced at Yale. After accounting for zero counts across all 69 samples, we reduced the data from 41 466 to 33 442 genes.

TABLE 3. Mean true positive rate (TPR) and mean true proportion of false positives among discoveries (V/R) with respective replicate size calculated for 100 simulations following simulation setting 3 at controlled 5% false discovery rate (FDR) via the Benjamini and Hochberg (BH) method.

| Model | Poisson | | NB | |
|-------|---------------|----------------|---------------|----------------|
| | TPR (SE) | V/R (SE) | TPR (SE) | V/R (SE) |
| edgeR | 0.859 (0.04%) | 0.0192 (0.03%) | 0.778 (0.05%) | 0.0233 (0.04%) |
| DESeq | 0.856 (0.04%) | 0.0205 (0.03%) | 0.792 (0.05%) | 0.0318 (0.04%) |
| TSPM | 0.947 (0.04%) | 0.2102 (0.11%) | 0.575 (0.09%) | 0.2339 (0.10%) |

There may be dissimilarities between human and plant gene expression profiles. However, to evaluate different statistical methods, the number of individuals (biological replicates) of the LCL data set allows us to apply a data-based simulation strategy. This strategy generates data without making any distributional assumptions and thus, likely mimics real data better than data that were simulated based on probability models.

Step 1: Select samples for hypothetical treatment group—For each simulation, we randomly selected n_i samples and assigned them to each of two hypothetical treatment groups. Next, after further excluding those genes with zero counts across the selected hypothetical treatment groups, we randomly selected 10 000 genes for each simulation. We expect no differential expression for these genes because the samples were randomly chosen from the same population of 69 individuals.

Step 2: Simulate DE—We generated DE genes to a random sample of 20% of the total genes by scaling counts by $\exp\{(-1)^i \delta_g\}$ where treatment group is indexed by $i = 1, 2$. We drew δ_i for each intended DE gene independently from a two-component normal distribution with parameters $\mu = (-0.5, 0.5)$ and $\sigma = (0.7, 0.7)$. We employed this simulation design for studies with $n_i = 2, 4,$ and 5 replicates. This is a similar strategy to that applied in Nettleton et al. (2008) where they looked at gene set analysis for microarray data.

The three methods based on NB models, edgeR, DESeq, and baySeq, perform comparably, while baySeq is obviously better for small FPR values. The range of small FPR values are of the most practical importance because, after controlling multiple testing error such as FDR, we rarely go down the significance list to a position with more than 5% FPR. For the smallest replication of $n_i = 2$, TSPM performs much poorer than the other three, followed by a comparable performance in terms of power as replication increases (Fig. 4). When we look at the FDR control, all methods do not control FDR well in this simulation setting (Table 4). TSPM performs closest to a desired level of control at four and five replicates among the four methods.

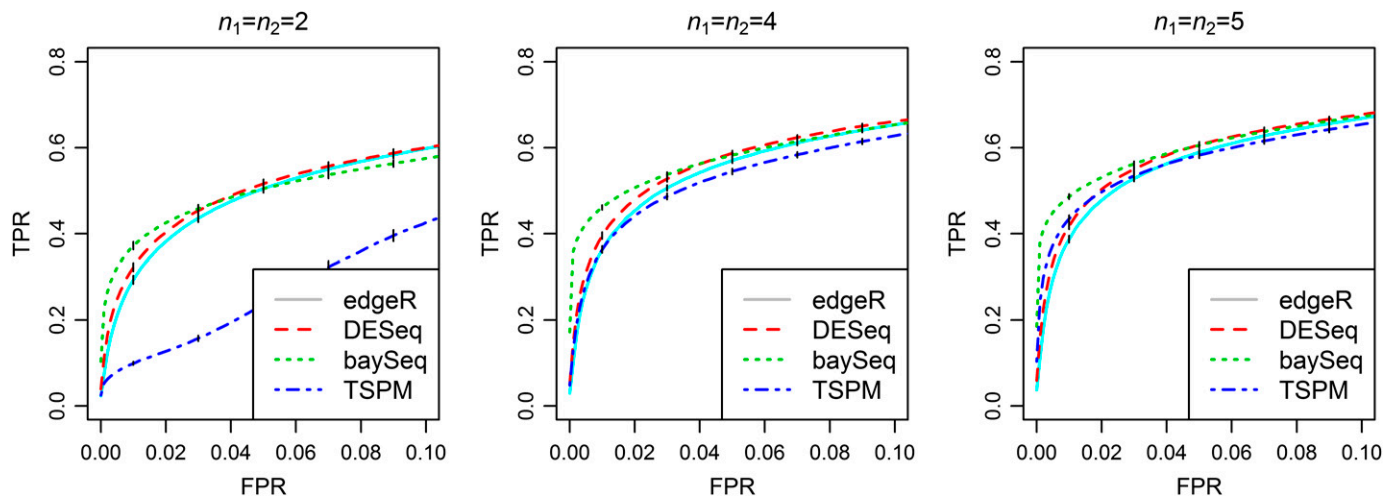


Fig. 4. Mean receiver operating characteristic (ROC) curves, based on 100 simulations from simulation setting 4, comparing the performance of edgeR, DESeq, baySeq, and TSPM to detect differential expression. Simulation based on real data experiment sequencing lymphoblastoid cell lines. Left panel is from two replicates, middle panel is from four replicates, and right panel is from five replicates. Vertical bars at odd levels of false positive rate (FPR) are ± 2 times standard error to the value of the estimated corresponding true positive rate (TPR).

DATA ANALYSIS

We analyzed the real RNA-seq data set that compares bundle sheath and mesophyll cells obtained from the laser-capture microdissected (LCM) samples from the tip of the maize leaf (Li et al., 2010). There were two biological replicates in their studies. The same data set was used in simulations 2 and 3. We performed tests to this data set in the same way as described in the Review of Methods section, and the FDR levels were estimated by the Benjamini-Hochberg procedure. The numbers of detected DE genes at different FDR levels (0.01 and 0.05) are shown in Fig. 5A and 5B, respectively. Among the three compared testing methods, TSPM detected the most DE genes at estimated FDR level of 1%, and DESeq detected most at 5%. DESeq and edgeR displayed more agreement than TSPM did with either of them. For example, as shown in Fig. 5A, when the FDR is controlled at 1%, 86% of the genes detected as DE by DESeq were also declared DE by edgeR. And 91% of the genes detected by edgeR were also declared DE by DESeq. In the same figure, 70% of the genes detected by TSPM were undetected by edgeR nor DESeq. This result agrees with the simulation results that the behavior of edgeR and DESeq are comparable, while TSPM differs the most from other tests.

DISCUSSION

Due to the differences in generating microarray and RNA-seq data, methods appropriately considering the discrete distribution

of RNA-seq data are important in obtaining meaningful results. In addition, the RNA-seq technologies and statistical methods for RNA-seq data are under continuous development. Timely review and comparison of the most recently proposed procedures are important in helping practitioners decide which procedure to apply. In this manuscript, we compared four recently proposed statistical methods to detect differentially expressed genes based on RNA-seq data. We performed a variety of simulations that were based on different distribution models—Poisson, overdispersed Poisson, or NB—and real data. We believe that results from simulated data that mimic real data would provide a useful guide in choosing methods for analysis of data from RNA-seq studies.

The results from ROC curves suggest that baySeq performs best in terms of ranking genes according to their significance to be declared DE, especially for smaller values of FPR, which is of most practical importance. Both edgeR and DESeq perform similarly and close to baySeq. The results from TSPM are most variable and often the poorest when the number of replicates is small. Please note that our simulated data comprised two, four, and five replicates because those are the numbers we encounter in our collaborative works with plant biologists. The behavior of TSPM would be improved for data with more replicates as shown in Auer and Doerge (2011). TSPM always performs noticeably better as replicate size increases.

In terms of FDR control, the results give a warning that in many cases, the FDR may not be controlled well as we often obtained liberal rates of true FDR. New investigations into the reasons why FDR is not controlled well and development of

TABLE 4. Mean true positive rate (TPR) and mean true proportion of false positives among discoveries (V/R) with respective replicate size calculated for 100 simulations following simulation setting 4 at controlled 5% false discovery rate (FDR) via the Benjamini and Hochberg (BH) method.

| Replicates | $n = 2$ | | $n = 4$ | | $n = 5$ | |
|------------|---------------|---------------|---------------|---------------|---------------|---------------|
| | TPR (SE) | V/R (SE) | TPR (SE) | V/R (SE) | TPR (SE) | V/R (SE) |
| edgeR | 0.337 (0.60%) | 0.156 (0.50%) | 0.458 (0.30%) | 0.158 (0.35%) | 0.492 (0.24%) | 0.159 (0.39%) |
| DESeq | 0.361 (0.50%) | 0.149 (0.60%) | 0.439 (0.32%) | 0.122 (0.35%) | 0.465 (0.27%) | 0.120 (0.35%) |
| TSPM | 0.166 (0.28%) | 0.529 (0.74%) | 0.334 (0.41%) | 0.087 (0.40%) | 0.396 (0.32%) | 0.059 (0.35%) |

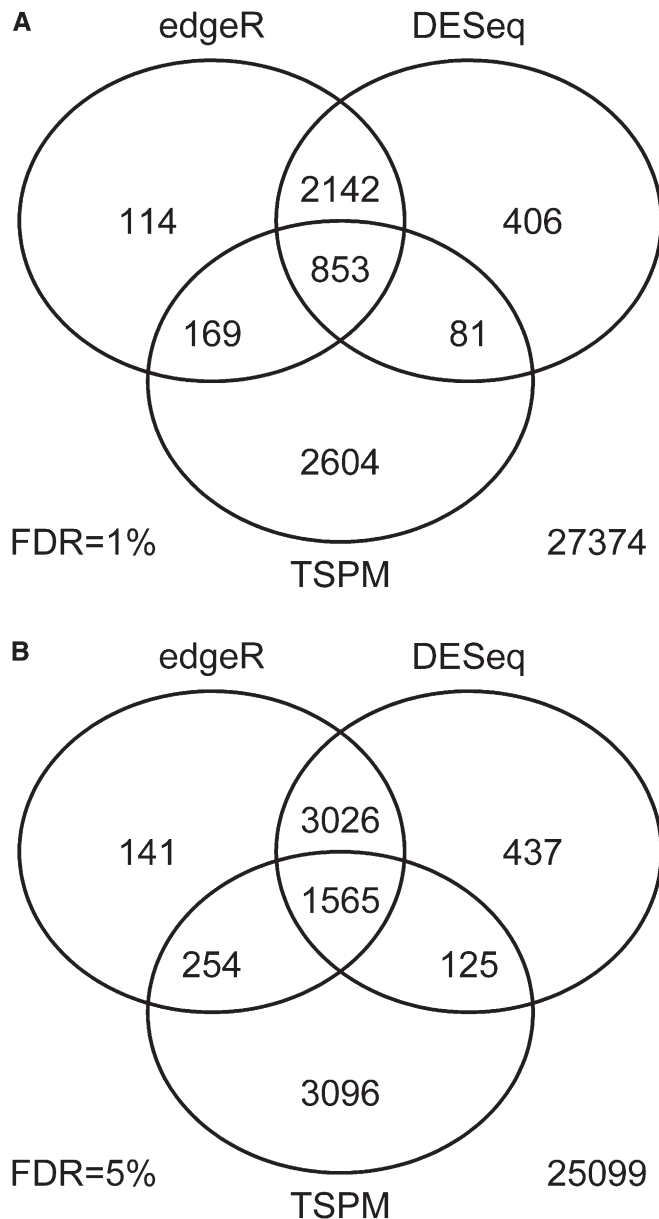


Fig. 5. Venn diagram of the number of detected differentially expressed genes. (A) For the laser-capture microdissected data from Li, et al. (2010) with false discovery rate (FDR) controlled at 1% using the Benjamin-Hochberg procedure. (B) For the laser-capture microdissected data from Li, et al. (2010) with false discovery rate (FDR) controlled at 5% using the Benjamin-Hochberg procedure.

better methods are needed. Meanwhile, we suggest practitioners use a more stringent FDR control to avoid too many false discoveries.

Among the R packages that are freely available and compared in our manuscript, the flexibility of handling different experimental designs vary. All can deal with the simple case that compares two treatment groups from a completely randomized design, which is the experimental design used in all of our simulation settings. Some methods offer more flexibility in allowing more complex experimental designs and others allow alternative modes of estimation. In the baySeq package, users can analyze

experimental designs involving multiple treatment groups. The package edgeR can be applied to two or more groups where at least one of the groups has replicated measurements. However, testing for differential expression is only allowed for pairwise comparisons (Robinson et al., 2010). There are two options to estimate the dispersion parameter in edgeR—common and tag-wise. In our simulation, we ran edgeR in common-dispersion mode because we think this might be the better case when the number of replicates is very small. However, we don't expect that results using tagwise estimation would differ far from those obtained in our simulation studies based on the observation of Anders and Huber (2010). All methods require replicates in at least one of the treatment groups except DESeq, which allows analysis of experiments with no biological replicates in either of the experimental conditions. However, a design without replication is not recommended. Given these options, the choice of method also depends on the experimental design.

LITERATURE CITED

- ANDERS, S., AND W. HUBER. 2010. Differential expression analysis for sequence count data. *Genome Biology* 11: R106.
- AUER, P. L., AND R. W. DOERGE. 2011. A two-stage Poisson model for testing RNA-Seq data. *Statistical Applications in Genetics and Molecular Biology* 10: 26.
- BAGGERLY, K. A., L. DENG, J. S. MORRIS, AND C. M. ALDAZ. 2004. Overdispersed logistic regression for SAGE: Modeling multiple groups and covariates. *BMC Bioinformatics* 5: 144.
- BULLARD, J. H., E. PURDOM, K. D. HANSEN, AND S. DUDOIT. 2010. Evaluation of statistical methods for normalization and differential expression in mRNA-Seq experiments. *BMC Bioinformatics* 11: 94.
- CUI, X., J. T. G. HWANG, J. QIU, N. J. BLADES, AND G. A. CHURCHILL. 2005. Improved statistical tests for differential gene expression by shrinking variance components estimates. *Biostatistics* 6: 59–75.
- FILICHKIN, S. A., H. D. PRIEST, S. A. GIVAN, R. SHEN, D. W. BRYANT, S. E. FOX, W. WONG, AND T. C. MOCKLER. 2010. Genome-wide mapping of alternative splicing in *Arabidopsis thaliana*. *Genome Research* 20: 45–58.
- GENTLEMAN, R. C., V. J. CAREY, D. M. BATES, B. BOLSTAD, M. DETTLING, S. DUDOIT, B. ELLIS, ET AL. 2004. Bioconductor: Open software development for computational biology and bioinformatics. *Genome Biology* 5: R80.
- HARDCASTLE, T. J., AND K. A. KELLY. 2010. baySeq: Empirical Bayesian methods for identifying differential expression in sequence count data. *BMC Bioinformatics* 11: 422.
- HILLER, D., H. JIANG, W. XU, AND W. WONG. 2009. Identifiability of isoform deconvolution from junction arrays and RNA-Seq. *Bioinformatics* 25: 3056–3059.
- HWANG, J. T. G., AND P. LIU. 2010. Optimal tests shrinking both means and variances applicable to microarray data analysis. *Statistical Applications in Genetics and Molecular Biology* 9: 36.
- LI, P., L. PONNALA, N. GANDOTRA, L. WANG, Y. SI, S. L. TAUSTA, T. H. KEBROM, ET AL. 2010. The developmental dynamics of the maize leaf transcriptome. *Nature Genetics* 42: 1060–1067.
- LU, J., J. K. TOMFOHR, AND T. B. KEPLER. 2005. Identify differential expression in multiple SAGE libraries: An overdispersed log-linear model approach. *BMC Bioinformatics* 6: 165.
- LU, T., G. LU, D. FAN, C. ZHU, W. LI, Q. ZHAO, Q. FENG, ET AL. 2010. Function annotation of the rice transcriptome at single-nucleotide resolution by RNA-seq. *Genome Research* 20: 1238–1249.
- MARIONI, J. C., C. E. MASON, S. M. MANE, M. STEPHENS, AND Y. GILAD. 2008. RNA-seq: An assessment of technical reproducibility and comparison with gene expression arrays. *Genome Research* 18: 1509–1517.
- MORTAZAVI, A., B. A. WILLIAMS, K. MCCUE, L. SCHAEFFER, AND B. WOLD. 2008. Mapping and quantifying mammalian transcriptomes by RNA-Seq. *Nature Methods* 5: 621–628.
- NETTLETON, D., J. RECKNOR, AND J. M. REECY. 2008. Identification of differentially expressed gene categories in microarray studies using non-parametric multivariate analysis. *Bioinformatics* 24: 192–201.

- OSHLACK, A., M. D. ROBINSON, AND M. D. YOUNG. 2010. From RNA-seq reads to differential expression results. *Genome Biology* 11: 220.
- PICKRELL, J. K., J. C. MARIONI, A. A. PAI, J. F. DEGNER, B. E. ENGELHARDT, E. NKADORI, J. B. VEYRIERAS, ET AL. 2010. Understanding mechanisms underlying human gene expression variation with RNA sequencing. *Nature* 464: 768–772.
- ROBINSON, M. D., D. J. MCCARTHY, AND G. K. SMYTH. 2010. edgeR: A Bioconductor package for differential expression analysis of digital gene expression data. *Bioinformatics (Oxford, England)* 26: 139–140.
- ROBINSON, M. D., AND A. OSHLACK. 2010. A scaling normalization method for differential expression analysis of RNA-seq data. *Genome Biology* 11: R25.
- ROBINSON, M. D., AND G. K. SMYTH. 2007. Moderated statistical tests for assessing difference in tag abundance. *Bioinformatics (Oxford, England)* 23: 2881–2887.
- ROBINSON, M. D., AND G. K. SMYTH. 2008. Small-sample estimation of negative binomial dispersion, with applications to SAGE data. *Biostatistics* 9: 321–332.
- SALZMAN, J., H. JIANG, AND W. WONG. 2011. Statistical modeling of RNA-Seq data. *Statistical Science* 226: 62–83.
- SEVERIN, A. J., J. L. WOODY, Y. BOLON, B. JOSEPH, B. W. DIERS, A. D. FARMER, G. J. MUEHLBAUER, ET AL. 2010. RNA-Seq Atlas of *Glycine max*: A guide to the soybean transcriptome. *BMC Plant Biology* 10: 160.
- SMYTH, G. K. 2004. Linear models and empirical Bayes methods for assessing differential expression in microarray experiments. *Statistical Applications in Genetics and Molecular Biology* 3: 3.
- SMYTH, G. K. 2005. Limma: Linear models for microarray data. In R. Gentleman, V. Carey, S. Dudoit, R. Irizarry, and W. Huber [eds.], *Bioinformatics and computational biology solutions using R and bioconductor*, 397–420. Springer, New York, New York, USA.
- SRIVASTAVA, S., AND L. CHEN. 2010. A two-parameter generalized Poisson model to improve the analysis of RNA-seq data. *Nucleic Acids Research* 38: e170.
- WANG, L., P. LI, AND T. P. BRUTNEL. 2010. Exploring plant transcriptomes using ultra high-throughput sequencing. *Briefings in Functional Genomics* 9: 118–128.
- ZHANG, G., G. GUO, X. HU, Y. ZHANG, Q. LI, R. LI, R. ZHUANG, ET AL. 2010. Deep RNA sequencing at single base-pair resolution reveals high complexity of the rice transcriptome. *Genome Research* 20: 646–654.