Biostatistics (2013), **14**, 2, *pp.* 232–243
doi:10.1093/biostatistics/kxs033
Advance Access publication on September 22, 2012

A new shrinkage estimator for dispersion improves differential expression detection in RNA-seq data

HAO WU†

Department of Biostatistics and Bioinformatics, Emory University, Atlanta, GA 30322, USA

CHI WANG†

Department of Biostatistics and Markey Cancer Center, University of Kentucky, Lexington, KY 40536, USA

ZHIJIN WU*

Department of Biostatistics, Brown University, Providence, RI 02912, USA

zhijin_wu@brown.edu

Summary

Recent developments in RNA-sequencing (RNA-seq) technology have led to a rapid increase in gene expression data in the form of counts. RNA-seq can be used for a variety of applications, however, identifying differential expression (DE) remains a key task in functional genomics. There have been a number of statistical methods for DE detection for RNA-seq data. One common feature of several leading methods is the use of the negative binomial (Gamma–Poisson mixture) model. That is, the unobserved gene expression is modeled by a gamma random variable and, given the expression, the sequencing read counts are modeled as Poisson. The distinct feature in various methods is how the variance, or dispersion, in the Gamma distribution is modeled and estimated. We evaluate several large public RNA-seq datasets and find that the estimated dispersion in existing methods does not adequately capture the heterogeneity of biological variance among samples. We present a new empirical Bayes shrinkage estimate of the dispersion parameters and demonstrate improved DE detection.

Keywords: Differential expression; Empirical Bayes; RNA sequencing; Shrinkage estimator.

1. Introduction

The recent developments in RNA-sequencing (RNA-seq) technology have led to a rapid increase in gene expression data in the form of counts. RNA-seq enables some applications not achievable by microarrays, including discovering novel transcripts and alternative splicing. One fundamental question in functional genomics, however, remains the regulation of gene expression under the different conditions (treatments, genotypes, environments, or developmental stages). Therefore, identifying differential expression (DE) is still a key task in transcriptomics research.

*To whom correspondence should be addressed.
† Authors with equal contribution.

The measurement of expression in RNA-seq is the number of sequenced fragments that can be mapped to a genomic unit of interest, which can be a gene, an exon, or any region of interest. We use *gene* thereafter for convenience. One major difference in the DE detection methods for RNA-seq and microarray data is that data from microarrays are often modeled as a Gaussian distribution after proper preprocessing. These methods do not take into account the discrete nature of RNA-seq data and thus may not be directly applicable. There are nonetheless some common features between the two technologies. For example, the advancement in technology does not eliminate biological variability, which early applications of RNA-seq often neglected (Hansen *and others*, 2011). In fact, the much higher dynamic range of RNA-Seq data makes the heteroskedasticity, i.e. the gene-specific dispersion, even more important than it was with microarrays.

In both types of gene expression data, there are two sources of variation: technical variation, which represents the measurement error inherent to the technology and sample preparation, and biological variation, which represents the heterogeneity among samples in the same treatment group or population. If all subjects within the same group have an identical expression level for a given gene, i.e. lacking biological variation, the counts for that gene can be modeled as a Poisson random variable, as done by Marioni *and others* (2008) and Wang *and others* (2010). The earliest tests for DE essentially tested for whether the difference of average counts in two groups exceeds what is expected by Poisson sampling error. However, since gene expression is a stochastic process, biological replicates in the same treatment group do not share identical expression levels. The presence of biological variation leads to the "over-dispersion" problem, e.g. the read counts show variation greater than expected from Poisson random variables.

A common approach to address the over-dispersion problem is to use the negative binomial (NB) model. The NB model is a Gamma–Poisson mixture and can be interpreted as the following: the Gamma distribution models the unobserved true expression levels in each biological sample, and conditioning on the expression level, the measurement from the sequencing machine follows a Poisson distribution (Robinson and Smyth, 2007; Anders and Huber, 2010; Hardcastle and Kelly, 2010). The variance from an NB distribution depends on the mean μ in the relationship:

$$\text{var} = \mu + \mu^2\phi, \tag{1.1}$$

where the first term represents variance due to Poisson sampling error and the second term represents variance due to variation between biological replicates. The parameter ϕ is referred to as the dispersion parameter. Note that ϕ is the reciprocal of the shape parameter in the Gamma distribution, and thus is the squared coefficient of variation (CV). Therefore, ϕ represents the variation of a gene's expression relative to its mean.

Most statisticians agree that the over-dispersion problem needs to be addressed. The difference is how the dispersion is modeled, estimated, and used in inference. Robinson and Smyth (2008) assume a common dispersion for all genes and use information from all genes to estimate a global ϕ. This stabilizes the estimation for ϕ but a common dispersion means an identical CV for all genes, while it is known that some genes are more tightly controlled (e.g. housekeeping genes) and other genes vary much more relative to their means (e.g. immune-modulated and stress-induced genes (Pritchard *and others*, 2001)). The gene-specific biological variation is reproducible across technologies (Hansen *and others*, 2011) and is not reflected by a common dispersion. As a result, genes that are naturally more variable are more likely to be reported as DE due to an underestimation of their dispersion. Anders and Huber (2010) introduce DESeq, another method based on the NB model. Instead of assuming the second term in (1.1) to be proportional to μ^2 with a dispersion ϕ, they let the variance be a smooth function of the mean (with proper offset accounting for the sequencing depth). However, the model in Anders and Huber (2010) still assumes that conditioning on the mean expression, the variance is constant. That is, two genes with the same mean expression level also has the same variance. There have been many updates of the DESeq package and the current version uses the greater value between the empirical gene-specific

dispersion and the mean-dependent fitted value. Robinson and Smyth (2007) modify the common dispersion model and introduce an empirical Bayes shrinkage estimate using weighted conditional log-likelihood. This new estimate shrinks the sample dispersion for each gene toward a common prior instead of shrinking them completely to the common dispersion. The method is implemented in the widely used R/Bioconductor package edgeR (Robinson *and others*, 2010), with recent extensions to multifactor experiments (McCarthy *and others*, 2012). Hardcastle and Kelly (2010) use NB distribution for the counts in a Bayesian hierarchical model setting. Non-parametric prior distributions for the NB parameters are empirically determined. This is a much more computationally intensive method and, possibly for this reason, not as widely applied as edgeR and DESeq.

Since these publications, many RNA-seq studies have emerged and the accumulated RNA-seq datasets with a reasonable number of replicates encourage us to re-evaluate the estimation of dispersion in RNA-seq. We find that existing methods often capture the overall level of dispersion but do not adequately reflect the heterogeneity in dispersion among genes. In this paper, we propose a new empirical Bayes method to shrink the dispersion parameter. We demonstrate using real data-based simulation that the proposed method provides a better estimate for the gene-specific dispersion and improves DE detection. The rest of the paper is organized as follows. In Section 2, we describe the datasets that provide both the motivation for this research and the basis for the evaluation of DE detection. In Section 3, we demonstrate the limitation of current estimation of the dispersion parameter and motivate our model for the prior distribution of the dispersion parameter. In Section 4, we present the hierarchical model, estimation, and testing procedure. In Section 5, we compare the DE detection using the proposed method with the leading alternatives. Finally, in Section 6, we discuss the interpretation of the dispersion, the connection with microarray data, and future directions.

2. DATA DESCRIPTION

Two real RNA-seq datasets are used in the paper for estimating simulations parameters and showing motivational plots. The first is from Cheung *and others* (2010), which quantifies the expressions of lymphoblastoid cell lines from 41 CEU individuals in International HapMap Project (Gibbs *and others*, 2003). The second is from Blekhman *and others* (2010), which includes six liver samples from both males and females. These datasets are referred to as "Cheung data" and "Gilad data", respectively, thereafter. For both datasets, the sequence reads are summarized into counts within genes from Ensembl 61 annotation. The gene counts are obtained from the ReCount webpage presented by Frazee *and others* (2011). Samples in both datasets include biological replicates, so the variance in observed data is a result of both biological variation and technical variation.

We also use data from the MicroArray Quality Control Project (MAQC) (Shi *and others*, 2006) phase III, also known as Sequencing Quality Control. Two biological samples, human brain and universal human reference sample, are assayed using seven lanes each. The MAQC data use samples from the same library preparation protocols, and thus represent data with technical replicates only. This dataset provides an extreme case in which the biological variation reaches the lower bound.

3. MOTIVATION

The NB distribution is widely used in modeling sequence count data because the Gamma–Poisson mixture provides a natural hierarchy for the generation of sequencing counts. The Gamma distribution is a flexible continuous distribution that characterizes the expression rate for a gene. Given the expression rate in a biological sample, the sequencing technology provides a count from a Poisson process with the mean proportional to the expression rate and an offset related to sequencing depth and possibly other factors affecting the detection efficiency (Hansen *and others*, 2012). When there are a large number of replicates,

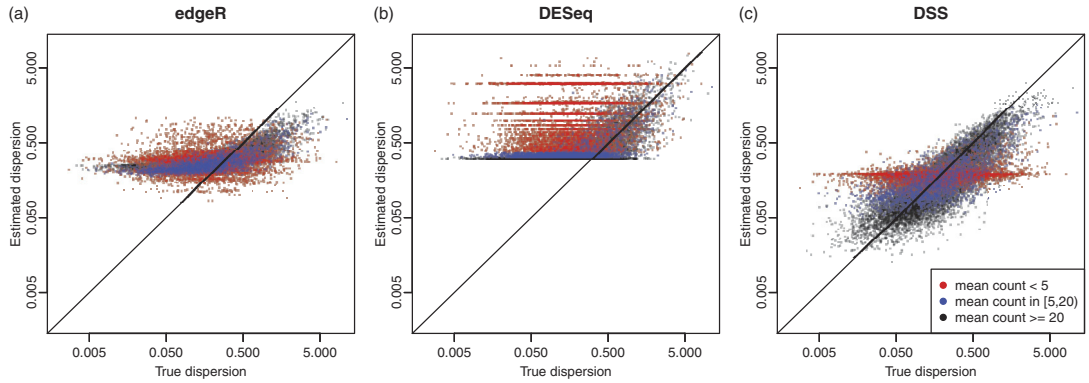

Fig. 1. Estimated versus true dispersion from (a) edgeR, (b) DESeq and (c) DSS, based on simulation. Simulated data were generated for 20 000 genes and two treatment groups with four replicates in each group. The dispersions were randomly simulated from log-normal distribution with mean -1.72 and standard deviation 1.07, which match those parameters observed from the Cheung data. Different colors represent genes with different levels of average read counts. A figure showing the genes in different color strata in separate panels is included as supplementary material available at *Biostatistics* online, Figure S2.

the dispersion parameter can be estimated from the data and over-dispersion is a well-addressed problem in the generalized linear model framework. The challenge in RNA-seq data is that there are often very limited replicates, such that the sample estimate of gene-specific dispersion is unstable. In this case, it is often useful to combine information from other genes to improve the estimation. DESeq and edgeR are the most widely used methods that provide shrinkage estimates of the dispersions by borrowing information across genes. To assess how well the dispersions are estimated from these methods, we conduct a simulation based on real data. First the gene-specific means and dispersions are estimated from the Cheung data. Pseudo-datasets of 20 000 genes, in two treatment groups with four replicates each, are constructed under the NB model with parameters taken from the real data. The mean expressions are directly sampled from estimated gene-specific means non-parametrically. The dispersions are generated from log-normal$(-1.72, 1.07^2)$, which well approximates the empirical distribution of the dispersion observed in the Cheung data (supplementary material available at *Biostatistics* online, Figure S1).

DESeq function `estimateDispersions with the default settings (sharingMode="maximum" and fitType="parametric")` and edgeR functions `estimateCommonDisp` followed by `estimateTagwiseDisp` with the default settings (`trend="movingave", prop.used=0.3, method="grid"`) are applied to the simulated counts to obtain estimated dispersions. Figure 1(a) and (b) shows the estimated versus true dispersions from both methods. Clearly, the estimates do not correlate well with the truth. Although both edgeR and DESeq estimate the central magnitude of the dispersions well, they fail to capture the variations in dispersions among genes. Changing the settings in the edgeR function `estimateTagwiseDisp` leads to different estimates (supplementary material available at *Biostatistics* online, Figure S4), but does not appear to substantially increase the accuracy of the estimated dispersion. The near horizontal clusters of points show that the dispersion of many genes are estimated to be approximately the same, which indicates over-shrinkage of the dispersions.

Since the dispersion parameter represents biological variation, having a good estimate of ϕ is crucial to finding the DE that is beyond natural biological variation. In order to account for gene-specific dispersion, we take advantage of real RNA-seq datasets with replicates and note that the distribution of the logarithm of sample dispersion, $\log(\phi)$, is approximately Gaussian as shown in Figure 2. In the next section, we describe a novel shrinkage estimator for ϕ using a log-normal prior and an NB likelihood. In Section 5,

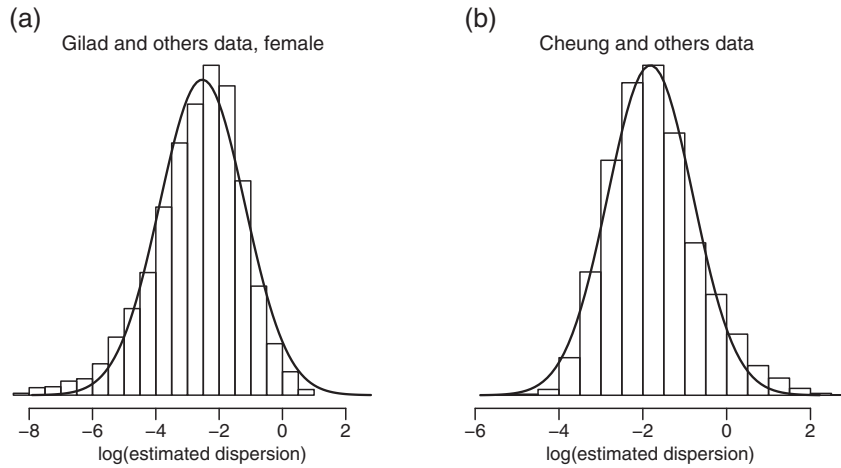

Fig. 2. Histogram for the logarithm of estimated gene-specific dispersions (ϕ_g) from (a) Gilad and (b) Cheung data. Only female samples are used for the Gilad data. The solid lines are density curves for normal distribution with parameters estimated from $\log(\phi_g)$. It can be seen that ϕ_g can be approximately modeled as a log-normal distribution.

we show that the proposed method greatly improves the estimation of ϕ and leads to better detection of the DE.

4. METHODS

Denote the observed counts from gene g in sample i by Y_{gi}, and the unobserved expression rate by θ_{gi}. We assume the following hierarchical model:

$$\begin{aligned} Y_{gi}|\theta_{gi} &\sim \text{Poisson}(\theta_{gi}s_i), \\ \theta_{gi}|\phi_g &\sim \text{Gamma}(\mu_{g,k(i)}, \phi_g), \\ \phi_g &\sim \text{log-normal}(m_0, \tau^2). \end{aligned}$$

Here the Gamma distribution is parameterized with mean and dispersion (ϕ_g is the reciprocal of the shape parameter in the common parameterization); $k(i)$ denotes the treatment group of sample i; and s_i represents the normalizing factor, such as the relative library size, for sample i. This normalization factor could also be a gene-specific factor, taking into account the varying sequencing efficiency due to the guanine-cytosine content and/or length (Hansen *and others*, 2012; Risso *and others*, 2012), but would nonetheless be treated as a known constant. Based on the hierarchical model, the marginal distribution of Y_{gi} given $\mu_{g,k(i)}$ and ϕ_g is NB with mean $\nu_{gi} = \mu_{g,k(i)}s_i$ and dispersion ϕ_g. In the simplest setting where there are only two groups, the goal of DE detection is to test, for each gene, whether the mean expressions are identical in both groups, e.g. $\mu_{g,1} = \mu_{g,2}$, as noted by other researchers (Robinson and Smyth, 2007, 2008; Anders and Huber, 2010; Robinson *and others*, 2010).

Estimating ϕ_g is a crucial step in DE detection and shrinkage estimators have been shown to be useful in typical RNA-seq experiments with a small number of replicates (Anders and Huber, 2010; Hardcastle and Kelly, 2010; Robinson *and others*, 2010). As discussed in previous works (Robinson and Smyth, 2007), there is no conjugate prior for ϕ_g to ease the computation of a posterior distribution. Thus, we decide to use a prior that approximates the ϕ_g in real RNA-seq data. We choose the log-normal distribution, motivated by Figure 2.

For gene g, the conditional posterior distribution of ϕ_g given all observed counts and means, $p(\phi_g|Y_{gi}, \nu_{gi}, i = 1, \ldots, n)$ satisfies (detailed derivation can be found in supplementary material available at *Biostatistics* online):

$$\begin{aligned}\log[p(\phi_g|Y_{gi}, \nu_{gi}, i = 1, \ldots, n)] \propto & \sum_i \psi(\phi_g^{-1} + Y_{gi}) - n\psi(\phi_g^{-1}) - \phi_g^{-1} \sum_i \log(1 + \nu_{gi}\phi_g) \\ & + \sum_i Y_{gi}[\log(\nu_{gi}\phi_g) - \log(1 + \nu_{gi}\phi_g)] \\ & - \frac{[\log(\phi_g) - m_0]^2}{2\tau^2} - \log(\phi_g) - \log(\tau), \end{aligned} \tag{4.1}$$

where $\psi(.)$ is the log gamma function and n is the number of samples. To obtain a point estimate of ϕ_g, we could compute the posterior mean by using numerical methods such as importance sampling. It is, however, too computationally intensive to be practical for RNA-seq data. Instead, we compute the posterior mode by maximizing an approximate of (4.1) and denote the estimate by $\tilde{\phi}_g$.

Specifically, we substitute ν_{gi} by $\hat{\nu}_{gi} \equiv \hat{\mu}_{g,k(i)} s_i$ where $\hat{\mu}_{g,k(i)} = (\sum_{j:k(j)=k(i)} Y_{gj}/s_j)/n_{k(i)}$, and $n_{k(i)}$ is the number of samples in the same treatment group as i. The estimate $\hat{\mu}_{g,k(i)}$ is the same as the sample estimate of true concentration in Anders and Huber (2010). We plug in the hyper-parameters m_0, τ^2 estimated from the data, and maximize the approximated Equation (4.1) using the Newton–Raphson method. In practice, we use "optimize" function in *R* , which provides excellent computational performance.

The hyper-parameters m_0, τ^2 are estimated from the data by pooling information from all genes (details in the next subsection). The estimate $\tilde{\phi}_g$ is therefore an empirical Bayes estimate shrunken toward the common prior. It can also be viewed as an estimate that maximizes a penalized pseudo-likelihood, as we are maximizing an approximate of the penalized likelihood (4.1), with penalty $-[\log(\phi_g) - m_0]^2/(2\tau^2) - \log(\phi_g)$, by replacing the nuisance parameter ν_{gi} by its estimate $\hat{\nu}_{gi}$. The first term in the penalty, $[\log(\phi_g) - m_0]^2/(2\tau^2)$, penalizes values that deviate far from the common prior m_0 and the second term adds an additional penalty for large values of ϕ_g. As mentioned in Section 1, ϕ_g has the interpretation of the squared CV. Based on RNA-seq data from human populations (Cheung *and others*, 2010; Pickrell *and others*, 2010), we observe that around 70% of the genes have a CV less than 0.5. The CV in inbred animal models or established cell lines are likely even lower. Thus, penalizing larger ϕ_g appears a desirable feature.

Estimating hyper-parameters. We start with a simple dispersion estimate for each gene motivated by the method of moments. Define a new random variable $z_{gi} \equiv (Y_{gi}^2 - Y_{gi})/s_i^2$; we have $\mathrm{E}[z_{gi}|\phi_g] = \mu_{g,k(i)}^2(\phi_g + 1)$, which is identical for every sample in a certain treatment group. Using the sample mean to estimate the first moment of z_{gi} and pooling data from different treatment groups together, one can obtain an estimator $\hat{\phi}_g = \sum_i z_{gi}/\sum_i \hat{\mu}_{g,k(i)}^2 - 1$ (detailed derivation in supplementary material available at *Biostatistics* online).

Next we estimate m_0 and τ using the empirical distribution of $\hat{\phi}_g$. We use the median of $\log(\hat{\phi}_g)$'s, $\hat{m}_0$, to estimate m_0. The estimation of τ^2 is less straightforward as the sample variance of $\log(\hat{\phi}_g)$ overestimates τ^2. The sample variance of $\log(\hat{\phi}_g)$ comprises two parts: τ^2, the variance representing the heterogeneity of $\log(\phi_g)$'s among genes, and $\mathrm{var}\{\log(\hat{\phi}_g)|\log(\phi_g)\}$, the variation due to estimating $\log(\phi_g)$. Even when $\tau^2 = 0$ (all genes share common dispersion, $\phi_g = \phi_0\ \forall g$), we still observe non-zero $\mathrm{var}\{\log(\hat{\phi}_g)\}$. Therefore, it is important to take into account this inflation in sample variance of $\log(\hat{\phi}_g)$. We use an *ad hoc* adjustment that behaves well in practice evaluated by extensive simulation. Specifically, we create a pseudo-dataset with $\tau^2 = 0$ by simulating Y'_{gi} from $\mathrm{NB}(\hat{\nu}_{gi}, \hat{\phi}_0)$, where $\hat{\phi}_0 = \exp\{\hat{m}_0\}$ is used as common dispersion for all genes. We then compute the sample dispersion $\hat{\phi}'_g$ for each gene, and estimate $\mathrm{var}\{\log(\hat{\phi}_g)|\log(\phi_0)\}$ as

$S^2 = [\text{IQR}\{\log(\hat{\phi}'_g)\}/1.349]^2$. From a number of pseudo-datasets, we obtain the mean $\bar{S}^2$ as the baseline. Lastly, we estimate τ^2 as $\hat{\tau}^2 = \max([\text{IQR}\{\log(\hat{\phi}_g)\}/1.349]^2 - \bar{S}^2, c_0)$. Here c_0 is the lower bound for $\hat{\tau}^2$. In practice, we use $c_0 = 0.01$.

Note that $\hat{\phi}_g$ is only used in estimating the hyper-parameters and not directly used to obtain the shrinkage estimate $\tilde{\phi}_g$. Although each $\hat{\phi}_g$ is a rather crude estimate of the dispersion, especially for a small number of samples, the large number of genes in RNA-seq data allows us to obtain stable estimates of the hyper-parameters. Once the prior is established in the empirical Bayesian method, our shrinkage estimate $\tilde{\phi}_g$ directly comes from maximizing (4.1).

Test statistic and false discovery rate. With all parameters estimated, a hypothesis testing for the comparison of any two groups can be carried using a Wald test or exact test. Bullard *and others* (2010) show that the likelihood ratio test and the exact test perform similarly. Our simulation show that the exact test and the Wald test provide similar performance (supplementary material available at *Biostatistics* online, Figure S9). We choose to use the Wald test for its simplicity.

The Wald test statistic for two-group comparison is constructed as

$$t_g = \frac{\hat{\mu}_{g,1} - \hat{\mu}_{g,2}}{\sqrt{\hat{\sigma}^2_{g,1} + \hat{\sigma}^2_{g,2}}},$$

where $\hat{\sigma}^2_{g,1} \equiv (1/n_1^2)[\hat{\mu}_{g,1}(\sum_{j:k(j)=1}(1/s_j)) + n_1\hat{\mu}^2_{g,1}\tilde{\phi}_g]$, is the estimated variance for $\hat{\mu}_{g,1}$ (detailed derivation in supplementary material available at *Biostatistics* online), and $\hat{\sigma}^2_{g,2}$ similarly defined. When there is no need for normalization ($s_j = 1\ \forall j$), $\hat{\sigma}^2_{g,1}$ reduces to $(\hat{\mu}_{g,1} + \hat{\mu}^2_{g,1}\tilde{\phi}_g)/n_1$, an estimate of the variance in (1.1), standardized by the sample size in group 1.

It is not trivial to derive the null distribution for the test statistic and the asymptotic distribution is irrelevant for most RNA-seq data. However, the empirical distribution of t_g is Gaussian-like (Figure 5(a)) and we can estimate a local false discovery rate (FDR) based on the empirical null (Efron, 2004) estimated from the data.

We implemented the proposed method in an R package titled "DSS", standing for Dispersion Shrinkage for Sequencing. The package is available from Bioconductor.

5. RESULTS

We compare results from the proposed method with two most widely used software packages, edgeR (version 2.6.2) and DESeq (version 1.8.2) using their latest release versions. For DESeq, we call the functions `estimateDispersions` and `nbinomTest` with default settings. For edgeR, we call functions `estimateCommonDisp` followed by `estimateTagwiseDisp` to estimate the dispersion parameter, and `exactTest` to test for DE.

First, we show an improvement in estimating the gene-specific dispersion parameter ϕ_g. All three methods use the common NB model, and estimating ϕ_g is the key challenge and what determines the ability to rank the truly DE genes above the null genes. In Figure 1(c), we show the estimates $\tilde{\phi}_g$ plotted against true ϕ_g. It is clear that the shrinkage estimates from DSS track the real dispersion much better than those from edgeR or DESeq. Note that the dispersion parameter estimates are shrunk more for genes with lower counts, as expected, since for those genes there is little information about ϕ in the data.

To quantify the improvement,we further use a similar approach as done in Robinson and Smyth (2007) to compare ϕ_g estimation. The mean squared error (MSE) of the estimates, in the scale of $\phi_g/(1+\phi_g)$, were calculated from 50 simulations. Instead of using fixed mean expressions as in Robinson and Smyth (2007), we use a more realistic approach and randomly sample from mean expressions computed from the Cheung data. DSS has the lowest MSE in both settings (Figure 3), and in all strata of mean expression

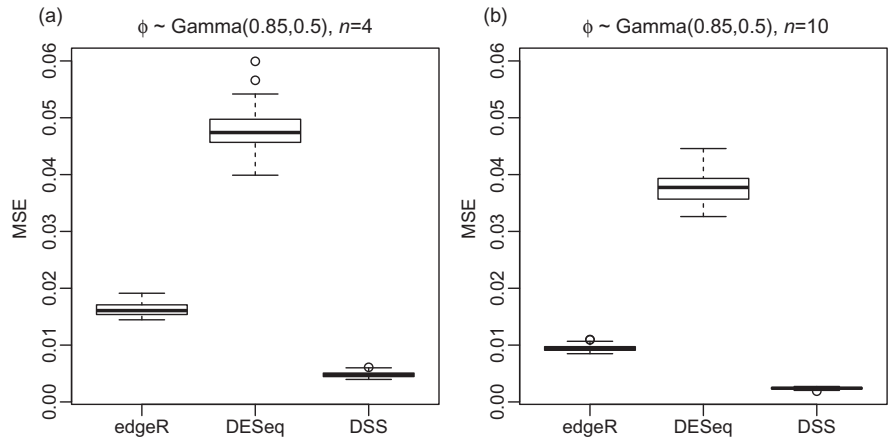

Fig. 3. Boxplots comparing the distribution of MSE for dispersion estimations from edgeR, DESeq and DSS over 50 simulations. In all simulations, dispersions ϕ_g are randomly generated from Gamma distribution with shape 0.85 and scale 0.5. The mean expressions are randomly sampled from the means calculated from Cheung data. Simulations are performed under different sample sizes: (a) 4 replicates, and (b) 10 replicates. All simulations are for 2000 genes.

(supplementary material available at *Biostatistics* online, Figure S3). We next compare the ability to detect the DE under a variety of situations. To evaluate DE detection under known truth, and to conduct the comparison under the most realistic scenario encountered in RNA-seq experiments, we simulated data semi-parametrically based on real RNA-seq datasets. In all simulations, the mean expressions were randomly sampled from the means calculated from the source dataset, Cheung or Gilad. The dispersion parameters were generated from parametric distributions, in which the parameters were computed from the source dataset (details in supplementary material available at *Biostatistics* online). To demonstrate robustness to distributional assumption, we simulated ϕ_g from both log-normal and Gamma distribution. In all simulations, 5% of all genes were true DE genes as this proportion is expected to be low in most gene expression experiments (Smyth, 2004). The logarithm of fold changes for the DE genes were randomly sampled from $N(0, 1)$.

First, we evaluate the ability to rank true DE genes above non-DE genes. Since DE detection is often used as a hypothesis generating tool, and the goal is to have as many true positives as possible in the top-ranked genes, we compare the percentage of true DEs (i.e. 1 − false discovery proportion) in the top-ranked genes up to top 1000. DSS shows a much higher proportion of true positives among the top-ranked genes when the dispersion is based on biological replicates as seen in the Gilad and Cheung data (Figure 4). All three methods provide identical performances when the biological variations are near zero transcriptome wide, as expected in technical replicates such as MAQC data in supplementary material available at *Biostatistics* online, Figure S11.

To further evaluate the performance of DSS in ranking genes under different dispersion patterns, we performed additional simulations with dispersion from different distributions. We found that in general when the variation in dispersions is large, DSS provides better performance. The average dispersion level does not cause much differences among the results from the three methods. This is consistent with the findings that DESeq and edgeR provide reasonable estimates for the average dispersion, but underestimate the variations in dispersion due to over-shrinkage. The results for these simulations are presented in supplementary material available at *Biostatistics* online (Section 10, Figures S6 and S7).

In addition to better sensitivity/specificity, another important task in DE detection is to provide a measure of statistical significance to guide the choice of a cut-off. In genomics, this is typically done using the FDR or some variant of the FDR. The Wald statistic in DSS is Gaussian-like in the center (Figure 5(a))

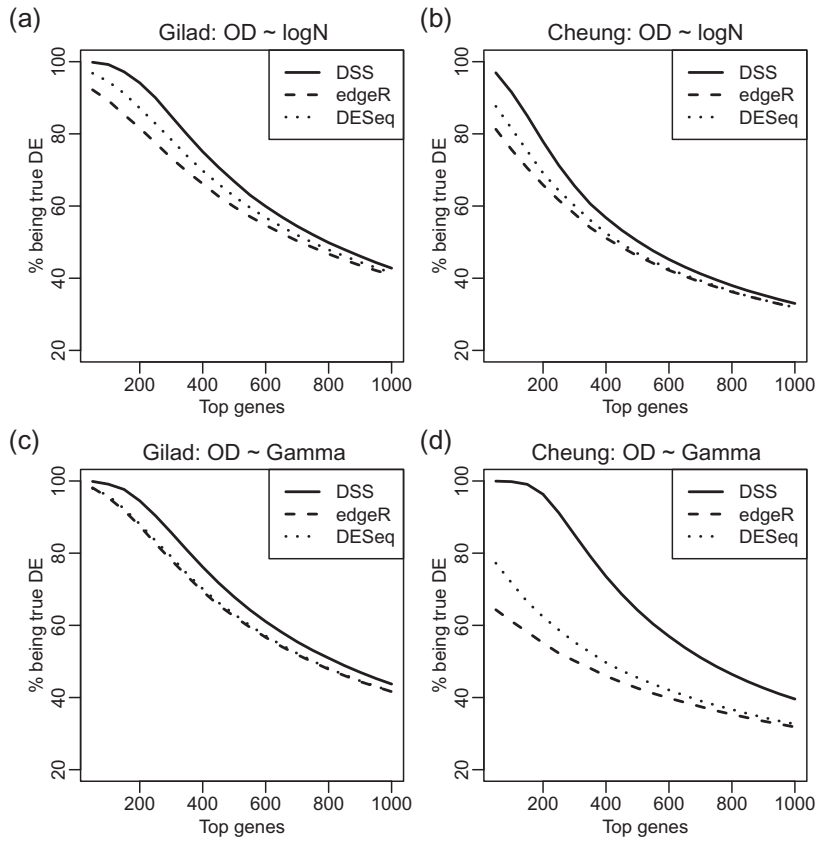

Fig. 4. DE detection accuracies from DSS, edgeR and DESeq for top 1000 ranked genes. Genes are ranked by the absolute Wald-statistic, from high to low, for DSS; ranked by p-values, from low to high, for edgeR and DESeq. The proportion of true discovery among top-ranked genes is plotted against the number of top-ranked genes. (a) The dispersion parameter simulated from log-normal distribution, with parameters estimated from the Gilad data. (b) The dispersion parameter simulated from log-normal distribution, with parameters estimated from the Cheung data. (c) The dispersion parameter simulated from Gamma distribution, with parameters estimated from the Gilad data. (d) The dispersion parameter simulated from Gamma distribution, with parameters estimated from the Cheung data.

and satisfies the assumption in local FDR calculation (Efron, 2004). Based on the connection between the local FDR and the FDR (Efron, 2004), we convert the local FDR to the scale of the classical FDR for comparison. For edgeR and DESeq, we use the FDR reported from the packages. Figure 5(b) plots the reported (dashed lines) and true FDRs (solid lines) for the top-ranked genes from the simulation. The simulation settings are the same as those used in Figure 4(b). The curves are averaged from 50 simulations. For DSS, the reported FDR curve is close to the true FDR. In contrast, both edgeR and DESeq give misleading FDR estimations. The underestimation of the FDR in the top-ranked genes gives an over-optimistic certainty in the reported DE. Supplementary material available at *Biostatistics* online, Figure S8, shows the comparison of the FDR under more simulation settings.

6. DISCUSSION

We present a new shrinkage estimator for the dispersion parameter in the NB model for RNA-seq data. We show that this new estimator better captures the variation in gene-specific dispersion and, as a result,

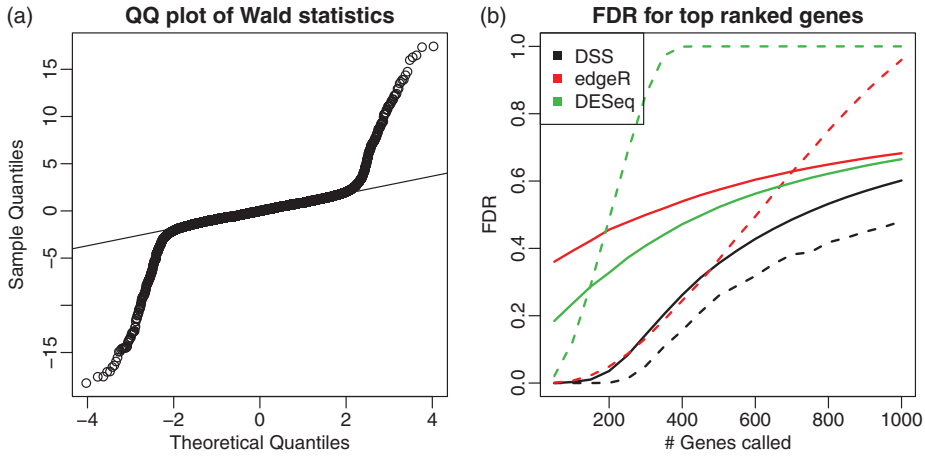

Fig. 5. Statistical inferences from simulation. Dispersions used in simulations were generated from Gamma distributions with parameters estimated from Cheung data. (a) Quantile–Quantile normal plot for Wald statistics from DSS. (b) Reported and true FDR curves for top-ranked genes. Solid lines represent true FDRs. Dashed lines represent reported FDRs.

leads to a better detection in the DE compared with two most popular DE detection softwares DESeq and edgeR. In addition, the nominal FDR based on the local FDR method is much closer to the actual FDR. We do not include another method, baySeq, in the comparisons because the computational time becomes formidable for the large number of simulations we perform.

We find that a good estimate of ϕ_g is critical in DE detection when biological replicates are involved. This is not surprising as we have seen in the past decade that improving the estimation of the gene-specific variance in microarray data has greatly improved DE detection by using moderated Wald tests, including the most widely used methods LIMMA (Smyth, 2004) and SAM (Tusher *and others*, 2001). In fact, we find that ϕ_g has a similar interpretation as the gene-specific variance in microarray data. Let Y represent an untransformed expression level in microarray data. It is often assumed that $\log(Y)|\mu \sim N(\mu, \sigma^2)$, where μ and σ^2 are the mean and variance for the log scale expression measure, respectively. In the original scale, the squared CV $\mathrm{CV}(Y)^2 = e^{\sigma^2-1} \approx \sigma^2$ for $\sigma^2 \ll 1$, which is satisfied in most genes (McCall *and others*, 2011). Thus, σ^2 is approximately the squared CV in the untransformed scale. In comparison, RNA-seq data are often modeled by the NB distribution, i.e. a Gamma–Poisson mixture, in which the Gamma distribution models the biological variation and the Poisson distribution models the variation due to the counting process in sequencing. The dispersion parameter ϕ_g is the squared CV of the Gamma distribution, i.e. ϕ_g represents the gene-specific biological variation, and is directly linked to σ_g^2 in log-transformed microarray expression data.

This interpretation explains several findings in our analysis of RNA-seq data with biological replicates. It has been well established that genes have different biological variance, and thus we should not expect a constant ϕ_g for all genes. In experiments with a small number of replicates, an empirical Bayes method that combines the information from the observed sample and shrinks toward a common prior of ϕ_g helps stabilize the estimate of ϕ_g. However, the extent of shrinkage should not be uniform as the amount of information about ϕ_g varies between genes. The variation of counts from genes with low expression is dominated by the Poisson counting error, and provides less information of ϕ_g (supplementary material available at *Biostatistics* online, Figure S10); thus the estimates of ϕ_g for these genes should be shrunken more heavily . DSS does exactly what we desire, shrinking ϕ_g less for high-expression genes and more for low-expression genes (Figure 1(c)).

It has been reported that the biological variation, thus the dispersion parameter, may depend on the expression level. DESeq and EdgeR allow trended estimation of the dispersion depending on the mean level. To account for this dependence, we also provide an option in DSS by letting the hyper-parameter m_0 be a smooth function of expression. The details of the trended estimation is provided in supplementary material available at *Biostatistics* online.

Some datasets are generated to investigate technical reproducibility and do not capture the biological variation in a population. Even in these cases with technical replicates only, such as the MAQC data, we find that analysis assuming a not-too-small dispersion provides improvement in DE detection (supplementary material available at *Biostatistics* online, Figure S12). When biological replicates are from established cell lines or inbred animal models, the variance among replicates may be lower than the variance in a random human population. However, the result on MAQC data with technical replicates suggests that even when there is no strong evidence for over-dispersion in the data, it is beneficial to analyze the data under a positive dispersion model.

We find that most of the improvement obtained in DSS is due to the different dispersion estimate, as passing the DSS estimates to edgeR/DESeq yields very similar results as in DSS (supplementary material available at *Biostatistics* online, Figure S9). Both the edgeR and the DESeq methods have been expanded to now accommodate multiclass comparisons. Our test is currently limited to two-class comparison and it is our immediate plan to extend the dispersion estimators to multifactor designs. With an estimate of the dispersion, one can use generalized linear models as done in McCarthy *and others* (2012).

In all comparisons presented here, we use semiparametric simulation based on real RNA-seq datasets for lack of real data example with known truth. Although the NB model is the current predominant choice for RNAseq data, most models only approximate reality. Real data with a reasonable number of known DE, for example, using spike-in genes, will enable better comparison.

SUPPLEMENTARY MATERIAL

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

ACKNOWLEDGMENTS

The authors thank the anonymous reviewers for their insightful and constructive comments. *Conflict of Interest*: None declared.

FUNDING

This research was supported by National Science Foundation (DBI-1054905).

REFERENCES

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

BLEKHMAN, R., MARIONI, J. C., ZUMBO, P., STEPHENS, M. AND GILAD, Y. (2010). Sex-specific and lineage-specific alternative splicing in primates. *Genome Research* **20**, 180–189.

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.

CHEUNG, V. G., NAYAK, R. R., WANG, I. X., ELWYN, S., COUSINS, S. M., MORLEY, M. AND SPIELMAN, R. S. (2010). Polymorphic Cis- and Trans-regulation of human gene expression. *PLoS Biology* **8**, e1000480.

Efron, B. (2004). Large-scale simultaneous hypothesis testing. *Journal of the American Statistical Association* **99**, 96–104.

Frazee, A., Langmead, B. and Leek, J. (2011). Recount: a multi-experiment resource of analysis-ready RNA-seq gene count datasets. *BMC Bioinformatics* **12**, 449.

Gibbs, R. A., Belmont, J. W., Hardenbol, P., Willis, T. D., Yu, F., Yang, H., Ch'ang, L. Y., Huang, W., Liu, B., Shen, Y. *and others*. (2003). The international hapmap project. *Nature* **426**, 789–796.

Hansen, K. D., Irizarry, R. A. and Wu, Z. (2012). Removing technical variability in RNA-seq data using conditional quantile normalization. *Biostatistics* **13**, 204–216.

Hansen, K. D., Wu, Z., Irizarry, R. A. and Leek, J. T. (2011). Sequencing technology does not eliminate biological variability. *Nature Biotechnology* **29**, 572–573.

Hardcastle, T. and Kelly, K. (2010). Bayseq: empirical bayesian methods for identifying differential expression in sequence count data. *BMC Bioinformatics* **11**, 422.

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 Research* **18**, 1509–1517.

McCall, M. N., Uppal, K., Jaffee, H. A., Zilliox, M. J. and Irizarry, R. A. (2011). The gene expression barcode: leveraging public data repositories to begin cataloging the human and murine transcriptomes. *Nucleic Acids Research* **39**(Suppl 1), D1011.

McCarthy, D. J., Chen, Y. and Smyth, G. K. (2012). Differential expression analysis of multifactor RNA-seq experiments with respect to biological variation. *Nucleic Acids Research* doi:10.1093/nar/gks042.

Pickrell, J. K., Marioni, J. C., Pai, A. A., Degner, J. F., Engelhardt, B. E., Nkadori, E., Veyrieras, J.-B., Stephens, M., Gilad, Y. and Pritchard, J. K. (2010). Understanding mechanisms underlying human gene expression variation with RNA sequencing. *Nature* **464**, 768–772.

Pritchard, C. C., Hsu, L., Delrow, J. and Nelson, P. S. (2001). Project normal: defining normal variance in mouse gene expression. *Proceedings of the National Academy of Sciences* **98**, 13266.

Risso, D., Schwartz, K., Sherlock, G. and Dudoit, S. (2012). GC-content normalization for RNA-seq data. *BMC Bioinformatics* **12**, 480–496.

Robinson, M. D., McCarthy, D. J. and Smyth, G. K. (2010). edgeR: a Bioconductor package for differential expression analysis of digital gene expression dataSMotn. *Bioinformatics* **26**, 139–140.

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.

Shi, L., Reid, L. H., Jones, W. D., Shippy, R., Warrington, J. A., Baker, S. C., Collins, P. J., De Longueville, F., Kawasaki, E. S., Lee, K. Y. *and others*. (2006). The microarray quality control (maqc) project shows inter-and intraplatform reproducibility of gene expression measurements. *Nature Biotechnology* **24**, 1151–1161.

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.

Tusher, V. G., Tibshirani, R. and Chu, G. (2001). Significance analysis of microarrays applied to the ionizing radiation response. *Proceedings of the National Academy of Sciences* **98**, 5116.

Wang, L., Feng, Z., Wang, X., Wang, X. and Zhang, X. (2010). Degseq: an r package for identifying differentially expressed genes from RNA-seq data. *Bioinformatics* **26**, 136–138.

[*Received February 25, 2012; revised August 17, 2012; accepted for publication August 19, 2012*]