 DOI: 10.1111/biom.12458

Modeling Overdispersion Heterogeneity in Differential Expression Analysis Using Mixtures

Elisabetta Bonafede,[1,*] Franck Picard,[2,] Stéphane Robin,[3,4,***] and Cinzia Viroli[1,****]**

[1]Department of Statistical Sciences, University of Bologna, 40126 Italy
[2]Laboratoire de Biométrie et Biologie Évolutive, UMR CNRS 5558 Univ. Lyon 1, F-69622 Villeurbanne, France
[3]AgroParisTech, UMR 518 MIA, Paris, France
[4]INRA, UMR 518 MIA, Paris, France
[*]*email:* elisabetta.bonafede@unibo.it
[**]*email:* franck.picard@univ-lyon1.fr
[***]*email:* robin@agroparistech.fr
[***]*email:* cinzia.viroli@unibo.it

Summary. Next-generation sequencing technologies now constitute a method of choice to measure gene expression. Data to analyze are read counts, commonly modeled using negative binomial distributions. A relevant issue associated with this probabilistic framework is the reliable estimation of the overdispersion parameter, reinforced by the limited number of replicates generally observable for each gene. Many strategies have been proposed to estimate this parameter, but when differential analysis is the purpose, they often result in procedures based on plug-in estimates, and we show here that this discrepancy between the estimation framework and the testing framework can lead to uncontrolled type-I errors. Instead, we propose a mixture model that allows each gene to share information with other genes that exhibit similar variability. Three consistent statistical tests are developed for differential expression analysis. We show through a wide simulation study that the proposed method improves the sensitivity of detecting differentially expressed genes with respect to the common procedures, since it reaches the nominal value for the type-I error, while keeping elevate discriminative power between differentially and not differentially expressed genes. The method is finally illustrated on prostate cancer RNA-Seq data.

Key words: Differential expression analysis; Mixture models; RNA-Seq data; ROC/AUC; Type-I error.

1. Introduction

Massive parallel sequencing has deeply changed our understanding of gene expression thanks to a higher resolution (Wang et al., 2009; Soon et al., 2013). From the analysis point of view, NGS experiments provide discrete *read counts* assigned to target genome regions measuring the expression level or the abundance of the target transcript. When the purpose of the assay is to perform *differential analysis*, that is comparing the counts of a given regions between conditions, the statistical task is then to provide an appropriate model to account for biological and technical variations, as well as a testing framework to test the hypothesis of no difference. Here we deal with the case where regions of interest are given *a priori*, contrarily to analysis where the regions themselves have to be discovered (Frazee et al., 2014). Models based on count distributions now constitute a consensus framework for the analysis, with the original Poisson distribution (Marioni et al., 2008; Wang et al., 2010) being replaced by the negative binomial model (Robinson and Smyth, 2008; Robinson et al., 2010). Indeed, the simplest choice of the Poisson distribution was rapidly identified as the cause of uncontrolled type-I errors, due to a poor adjustment to the larger observed variability compared with the equal mean-variance specification of the Poisson model (see, for a discussion, Anders and Huber (2010)). Since then, the correct modeling and estimation of this observed overdispersion has been a key issue in differential analysis.

Taking perspective from our past experience in micro-array analysis, the proper modeling of the dispersion parameter has long been a subject of debate in differential analysis, with a difficult trade-off between a common variance for every genes and gene-specific variances. Given the limited number of replicates, the first strategy provides robust estimates, but the testing procedure lacks of power and the model is not realistic, whereas the second is more sensitive at the price of increased type-I errors. Actually, the debate is still ongoing with the negative binomial framework, but the problem is much more difficult to solve due to this complex (and unknown) mean-variance relationship.

Several contributions have been proposed to find a trade-off between the common overdispersion and the gene-specific overdispersion frameworks. Robinson and Smyth (2008) and McCarthy et al. (2012) addressed the problem in a multi-step procedure called *edgeR*. They proposed to estimate the dispersions for each gene as a weighted combination of a common dispersion or by a "moderate-to-trend" approach that takes into account the dispersion-mean relationship. In a similar perspective, Law et al. (2014) estimate the mean–variance relationship on the base of the log-counts and generate weights for each set of observations; Di et al. (2011) solve

© 2015, The International Biometric Society

the problem of estimating the dispersions via nonparametric regression (the method is implemented in the R/Bioconductor packages *NBPSeq*). Anders and Huber (2010) proposed to use a mean-dependent local regression to smooth the gene-specific dispersion estimates, related to the idea that genes sharing a similar mean expression level have also a similar variance and therefore they can contribute to the estimation of the respective parameters. The method is implemented in the *DESeq2* R/Bioconductor package (Love et al., 2014). Wu et al. (2013) introduced an empirical Bayes shrinkage approach choosing a log-normal prior distribution on the dispersion parameters, and therefore imposing a negative binomial likelihood. Then the estimations are plugged-in the Wald statistic to perform the statistical test. The method is implemented in the *DSS* R library. In the Hardcastle and Kelly (2010) proposal called *baySeq*, the dispersion is iteratively estimated using the quasi-likelihood approach. Yu et al. (2013) and Soneson and Delorenzi (2013) evaluated the comparative performance of these approaches, together with other relevant proposals for differential expression analysis of RNA-Seq data, such as *NOISeq* (Tarazona et al., 2011), *PoissonSeq* (Li et al., 2012), *SAMSeq* (Li and Tibshirani, 2013), *QuasiSeq* (Lund et al., 2012), and *TSPM* (Auer and Doerge, 2011). More recently, Klambauer et al. (2013) addresses the issue of differential expression analysis with unknown conditions by mixtures of negative binomials. This method called *DEXUS* could be also adapted to the supervised situation with known conditions.

Despite the rapidly increasing diffusion of these statistical procedures, also thanks to the availability of the well documented R/Bioconductor packages, the estimation of the dispersion in NGS data remains a crucial and tricky issue because of the limited number of available observations for each gene. Moreover, less attention has been focused on the consistency between the *estimation* and the *testing* frameworks. Indeed, an important drawback of using plug-in estimators is that the expected variations of the test statistics are no longer controlled under the null hypothesis, which may result in an un-controlled level of the test. We will illustrate this point in a simulation study, by showing that most proposed methods do not always reach the nominal level of the test, whereas it is precisely what is expected to be controlled when performing standard hypothesis testing.

Our contribution is to explore and discuss a mixture model approach (McLachlan and Peel, 2000; Fraley and Raftery, 2002) based on the idea of sharing information among genes that exhibit similar dispersion. More specifically, mixtures of negative binomial distributions are investigated as a way to get more accurate estimation for the dispersion parameter of each gene, exploiting also the information provided by the others. Differently from *DEXUS* that assumes a mixture model for each transcript/gene with the aim of capturing the heterogeneity among conditions, the proposed method estimates a unified mixture model for all the genes with the aim of describing their "between" heterogeneity. Such an approach has already been considered in the same context for the differential analysis of microarray data (Delmar et al., 2005). A consistent statistical testing procedure is then developed within the unified model-based clustering framework. The proposed method improves the sensitivity of detecting differentially expressed genes (DE genes) with small replicates, and we show that our method controls the nominal level of the test by simulation. Compared to the other methods, it also shows a better overall discriminative performance between the truly differential and not differential expressed genes.

The novel method will be described in Section 2. We will derive three statistical tests and describe the procedure for performing the differential analysis in Section 3. We will show through a large simulation study in Section 4 that the proposed statistical test procedure outperforms the existing methods in the literature, because it is very good in reaching the nominal value for the type-I error, while keeping elevate discriminative power, thus indicating its inferential reliability. The method will be applied on prostate cancer data in Section 5. A final discussion is presented in Section 6.

2. The Proposed Method

2.1. *The Data Problem*

Suppose we observe the read counts of p genes in d biological conditions. For each condition, assume that n_j replicates are available, with $j = 1, \ldots, d$. Without loss of generality, here we consider $d = 2$. We denote Y_{ijr} the random variable that expresses the read counts, say y_{ijr}, mapped to gene i (i=1, ..., p), in condition j, in sample r (r= 1, ..., n_j). Let $\mathbf{Y}_i$ be the random vector of length $n = \sum_{j=1}^{d} n_j$ denoting the expression profile of a gene. Let $\mathbf{y}_i$ be the observed value. Analogously, $\mathbf{y}_{ij}$, is the vector of length n_j containing the observed gene profile under condition j. Let us consider the negative binomial distribution (NB), $NegBin(\lambda, \alpha)$ with parameters λ, $\alpha > 0$, expectation λ and variance $\lambda\left(1 + \frac{1}{\alpha}\lambda\right)$, so that λ is the location parameter and α is the dispersion parameter. For a random variable $Y \sim NegBin(\lambda, \alpha)$ we have

$$f(y|\lambda, \alpha) = \binom{y + \alpha - 1}{\alpha - 1} \left(\frac{\lambda}{\lambda + \alpha}\right)^{y} \left(\frac{\alpha}{\lambda + \alpha}\right)^{\alpha}. \qquad (1)$$

By assuming that the read counts can be described by the distribution in (1), ideally we would fit the model

$$Y_{ijr} \sim NegBin(\lambda_{ij}, \alpha_i)$$

for each gene and then we would test the null hypothesis $H_0 : \lambda_{i1} = \lambda_{i2}$. In most experiments, replicates are too few to accurately estimate both parameters for each gene (usually less than 20). To overcome the limitedness of the available observations, we propose to aggregate information of genes sharing similar variability thought a mixture model that jointly considers all the genes. The proposal constitutes a trade-off between the gene specific variance model and the one with a common dispersion parameter.

2.2. *The Model*

An interesting characteristic of the NB parametrization in (1) is that it can be derived from a Poisson–Gamma mixed distribution. More precisely, if the random variable U_{ijr} follows the Gamma distribution with unit mean

$$U_{ijr} \sim Gamma(\alpha_i, \alpha_i),$$

and the random variable Y_{ijr} conditional on $U_{ijr} = u_{ijr}$ follows the Poisson distribution

$$Y_{ijr}|U_{ijr} = u_{ijr} \sim Poisson(\lambda_{ij}u_{ijr}),$$

then unconditionally Y_{ijr} is distributed according to the NB, $Y_{ijr} \sim NegBin(\lambda_{ij}, \alpha_i)$. This means that the parameter λ_{ij} of the Poisson component rules the expectation of the negative binomial distribution, and the parameter α_i of the gamma component controls the heterogeneity, thus allowing overdispersion.

Instead of fitting p different NB models, we consider a unified mixture model approach for all the genes. In order to develop it, notice that the data can be viewed in a multilevel structure, where the first-level units are the replicates $(r = 1, ..., n_j)$, at the second level we have the conditions $(j = 1, ..., d)$ and at the third level there are the genes $(i = 1, ..., p)$. By considering the replicates independent draws within and between conditions we have

$$f(\mathbf{y}_i) = \prod_{j=1}^{d} f(\mathbf{y}_{ij}) = \prod_{j=1}^{d}\prod_{r=1}^{n_j} f(y_{ijr}). \tag{2}$$

We consider a mixture model for the units at the third level of this hierarchical structure. More precisely, we assume that the heterogeneity component $\mathbf{u}_i$ contains n independent replicates with distribution given by a mixture of K gamma components:

$$f(\mathbf{u}_i) = \sum_{k=1}^{K} w_k f_k(\mathbf{u}_i) = \sum_{k=1}^{K} w_k \prod_{j=1}^{d}\prod_{r=1}^{n_j} Gamma(u_{ijr}; \alpha_k, \alpha_k), \tag{3}$$

It is possible to show that if $Y_{ijr}|U_{ijr} = u_{ijr} \sim Poisson(\lambda_{ij}u_{ijr})$ then $\mathbf{Y}_i$ follows a mixture of NB distributions:

$$f(\mathbf{y}_i) = \sum_{k=1}^{K} w_k f_k(\mathbf{y}_i) = \sum_{k=1}^{K} w_k \prod_{j=1}^{d}\prod_{r=1}^{n_j} NegBin(y_{ijr}; \lambda_{ij}, \alpha_k). \tag{4}$$

Proof is given in the Web Appendix A.

In the mixture model (4) the p genes are clustered according to their variability only. Conditionally on each group k, the density of the generic read count y_{ijr} is a negative binomial with gene-wise and condition-wise mean equal to λ_{ij} and variance equal to $\lambda_{ij}\left(1 + \frac{1}{\alpha_k}\lambda_{ij}\right)$.

2.3. *Model Estimation*

Let $\boldsymbol{\Theta} = \{\lambda_{ij}, w_k, \alpha_k\}_{i=1,...p;j=1,...,d;k=1,...,K}$ be the whole set of model parameters. The log-likelihood of the model is given by

$$\ln L(\theta) = \ln \prod_{i=1}^{p} \sum_{k=1}^{K} w_k \prod_{j=1}^{d}\prod_{r=1}^{n_j} NegBin(y_{ijr}; \lambda_{ij}, \alpha_k). \tag{5}$$

A direct maximization of (5) is not analytically possible, but the maximum likelihood estimates can be derived by the EM algorithm (Dempster et al., 1977). The algorithm alternates between the expectation and the maximization steps until convergence, with the aim of maximizing the conditional expectation of the so-called complete log-likelihood given the observable data. The complete log-likelihood is the joint log-density of the observable data and of the missing data of the model. In the proposed mixture model, the hidden data consists of two sets of latent variables, which are (1) the latent variable u deriving by considering the NB distribution as a Gamma–Poisson mixed process and (2) the latent allocation K-vector, $\mathbf{z}_i$, that contains the 1 if the gene belongs to the group k and zero otherwise. The complete log-likelihood can be defined as

$$\begin{aligned}\ln L_c(\theta) = \ln f(\mathbf{y}, \mathbf{u}, \mathbf{z}) &= \ln \prod_{i=1}^{p} f(\mathbf{y}_i|\mathbf{u}_i) f(\mathbf{u}_i|\mathbf{z}_i) f(\mathbf{z}_i) \\ &= \ln \prod_{i}\prod_{j}\prod_{r} f(y_{ijr}|u_{ijr}) f(u_{ijr}|\mathbf{z}_i) f(\mathbf{z}_i),\end{aligned} \tag{6}$$

where $f(y_{ijr}|u_{ijr}) = Poisson(y_{ijr}; \lambda_{ij}u_{ijr})$, $f(u_{ijr}|z_{ik} = 1) = Gamma(u_{ijr};\alpha_k, \alpha_k)$ and $f(\mathbf{z}_i)$ is the multinomial distribution

$$f(\mathbf{z}_i) = \prod_{k=1}^{K} w_k^{z_{ik}}.$$

In the EM algorithm we maximize the conditional expectation of the complete density given the observable data, using a fixed set of parameters $\boldsymbol{\Theta}'$:

$$\arg\max_{\boldsymbol{\Theta}} E_{\mathbf{z},\mathbf{u}|\mathbf{y};\boldsymbol{\Theta}'} \left[\log f(\mathbf{y}, \mathbf{u}, \mathbf{z}|\boldsymbol{\Theta})\right], \tag{7}$$

which leads to iterating the E and M steps until convergence. The details of the algorithm are described in the Web Appendix A.

3. Differential Analysis

In this section, the procedure to identify the genes $(i = 1, ..., p)$ that differentially express under two $(j = 1, 2)$ different biological conditions is described. This aim can be accomplished in different ways: one could be interested in evaluating the equality between the two population means, or in checking wether their ratio is equal to 1, or wether the log-ratio is zero. The common null hypothesis can be rephrased in the following way for the three scenarios:

1) "Difference": $H_0 : \lambda_{i1} - \lambda_{i2} = 0$
2) "Ratio": $H_0 : \frac{\lambda_{i1}}{\lambda_{i2}} = 1$
3) "Log Ratio": $H_0 : \ln \frac{\lambda_{i1}}{\lambda_{i2}} = \ln(\lambda_{i1}) - \ln(\lambda_{i2}) = 0$

For each case, we can evaluate a test-statistic based on the EM estimates. Since the EM estimators are maximum likelihood estimators, the test-statistics are asymptotically distributed according to the standard Gaussian under the null

hypothesis:

1) For the "Difference" test statistic: $\frac{\widehat{\lambda}_{i1}-\widehat{\lambda}_{i2}}{\sqrt{Var(\widehat{\lambda}_{i1}-\widehat{\lambda}_{i2})}}|H_0 \rightsquigarrow N(0,1)$,
2) for the "Ratio" test statistic: $\frac{\frac{\widehat{\lambda}_{i1}}{\widehat{\lambda}_{i2}}-1}{\sqrt{Var\left(\frac{\widehat{\lambda}_{i1}}{\widehat{\lambda}_{i2}}\right)}}|H_0 \rightsquigarrow N(0,1)$,
3) for the "Log Ratio" test statistic: $\frac{\ln\widehat{\lambda}_{i1}-\ln\widehat{\lambda}_{i2}}{\sqrt{Var(\ln\widehat{\lambda}_{i1}-\ln\widehat{\lambda}_{i2})}}|H_0 \rightsquigarrow N(0,1)$,

where $\widehat{\lambda}_{i1}$ and $\widehat{\lambda}_{i2}$ are the EM-estimators.

The "Difference" statistical test For computing the test statistic we need to estimate the variance at the denominator of the test statistics. First of all, we notice that $Var(\widehat{\lambda}_{i1}-\widehat{\lambda}_{i2}) = Var(\widehat{\lambda}_{i1}) + Var(\widehat{\lambda}_{i2}) - 2Cov(\widehat{\lambda}_{i1},\widehat{\lambda}_{i2}) = Var(\widehat{\lambda}_{i1}) + Var(\widehat{\lambda}_{i2})$ since the covariance is zero by the model assumptions. The specific variances $Var(\widehat{\lambda}_{ij})$ are a function of y_{ijr}. In particular, $\widehat{\lambda}_{ij} = \frac{\sum_{r=1}^{n_j} y_{ijr}}{n_j}$ and denoting $\sum_r y_{ijr}$ as y_{ij+}

$$Var(\widehat{\lambda}_{ij}) = \frac{1}{n_j^2} Var(y_{ij+}). \tag{8}$$

The variance $Var(y_{ij+})$ can be computed by observing that the replicates y_{ijr}, with $r = 1,\ldots,n_j$ are independently distributed according to a mixture of NB, so that $Var(y_{ij+}) = n_j Var(y_{ijr})$ and for mixture models the general formula for the variance holds

$$Var(y_{ijr}) = E[Var(y_{ijr}|z_{ik})] + Var[E(y_{ijr}|z_{ik})],$$

where $Var[E(y_{ijr}|z_{ik})] = 0$ because the expectation is not component varying; as regards $E[Var(y_{ijr}|z_{ik})]$ we considered the conditional expectation given the observed data because of the multilevel structure of the data, and therefore

$$\begin{aligned} Var(y_{ijr}) &= E_{\mathbf{z}_i|\mathbf{y}_i}[Var(y_{ijr}|z_{ik}=1)] \\ &= \widehat{\lambda}_{ij}\left(1+\widehat{\lambda}_{ij}\sum_k \frac{f(z_{ik}|\mathbf{y}_i)}{\alpha_k}\right). \end{aligned} \tag{9}$$

This formula enlightens the effect of the mixture model we propose: the over-dispersion term is a weighted average of the (estimated) over-dispersion terms λ_{ij}/α_k one would get in each component of the mixture. These terms are weighted according to the posterior probability for observation i to belong to each component k: $f(z_{ik}|\mathbf{y}_i)$.

The "Ratio" statistical test As for the computation of $Var\left(\frac{\widehat{\lambda}_{i1}}{\widehat{\lambda}_{i2}}\right)$ we can use the Delta method (van der Vaart, 2000), and by simple computations we get:

$$Var\left(\frac{\widehat{\lambda}_{i1}}{\widehat{\lambda}_{i2}}\right) \approx \frac{Var(\widehat{\lambda}_{i1})}{E(\widehat{\lambda}_{i2})^2} + \frac{E(\widehat{\lambda}_{i1})^2}{E(\widehat{\lambda}_{i2})^4} Var(\widehat{\lambda}_{i2})$$

All the needed quantities can be computed easily. With regards to $E(\widehat{\lambda}_{ij})$ we can use the EM estimates for λ_{ij} because they are correct. For the variances we make use of the procedure described above.

The "Log Ratio" statistical test The variance at the denominator of the statistical test can be decomposed as $Var(\ln\widehat{\lambda}_{i1} - \ln\widehat{\lambda}_{i2}) = Var(\ln\widehat{\lambda}_{i1}) + Var(\ln\widehat{\lambda}_{i2})$. Now we observe that $Var(\ln\lambda_{ij}) = Var\left(\ln\left(\frac{y_{ij+}}{n_j}\right)\right) = Var(\ln(y_{ij+}))$. According to the Delta method $Var(g(y_{ij+})) \approx Var(y_{ij+})\left(\frac{\partial}{\partial y_{ij+}} g(y_{ij+})\right)^2$ and in our case the Delta approximation for the variance is $Var(\ln y_{ij+}) = \frac{1}{y_{ij+}^2} Var(y_{ij+})$.

4. Simulation Study

The performance of the proposed strategy is evaluated by a large simulation study comprising several data generating processes with the double aim of: (a) assessing the capability of the proposed mixture model to estimate the variances with a specified number of components (simulation A) and (b) evaluating the discriminative performance between the truly differential and not differential expressed genes and the accuracy of the three statistical test procedures in terms of type-I error (simulation B). In particular, in a set of multiple p tests in the absence of correction, a reliable statistical test should reach the nominal significance level as n increases.

The method described in this article is implemented in the software package `MixtNB`, available from the website CRAN/R.

4.1. *Simulation A*

In the first simulation study, we evaluated the capability of the proposed mixture model to estimate the variances of the genes as the number of components, K, increases. Indeed, our purpose is to account for heterogeneity among the overdispersion parameters and we do not necessarily believe that groups of differently overdispersed genes do exist. Eventually, the number of overdispersion parameters could be equal to the number of genes, and we want to test if our strategy will adapt to such a situation. The details of the simulation design are reported in the Web Appendix B. Results indicate that 4 or 5 components are generally enough to accurately estimate the variances. Moreover in a large comparison with other methods that explicitly estimate the dispersions, the proposed mixture model and the method *DSS* outperform the others approaches in terms of accuracy in the variance estimation. See the detailed results in the Web Appendix B.

4.2. *Simulation B*

The aim of this simulation study is to evaluate the ability to control the type-I error rate at an imposed level, the ability to rank the DE and non-DE genes and the Area Under the ROC Curve (AUC). By "controlling the type-I error," we mean having an empirical type-I error close to the nominal level of the test, not having a small type-I error. This criterion aims at checking if the null distribution of the test is correct. Also, we remind that the power of testing procedures can only be compared if type-I error rate are similar. So checking the control of the type-I error is also a way to ensure that the power of different procedures can be compared.

We compared the proposed MixtNB method with nine methods that perform DE analysis on count data directly and provide p-values, namely: *edgeR*, *DESeq2*, *DSS*, *DEXUS*, *SAMSeq*, *PoissonSeq*, *NBPSeq*, *QuasiSeq*, and *TSPM*. We evaluated the methods in different scenarii with the aim of assessing the effect of factors such as sample size, dimensionality and outliers. We generated synthetic data by the `compcodeR` package (Soneson, 2014). The package has the desirable property to generate synthetic data that reflect the main characteristics of real data and it allows to consider a variety of settings. We considered $p = 1,000, 5,000, 10,000$ (sequencing depth of 1e+05, 5e+05, and 1e+06, respectively) and different number of replicates $n_j = 3, 5, 10$ in order to cover a wide range of experimental settings. We set the number of differentially expressed genes at 10%, with the same proportion of upregulated and downregulated genes for evaluating the empirical type-I errors. In the ROC/AUC simulation study we also considered the situations with 20 and 40% of DE genes. We evaluated the robustness of the methods in alternative settings, more precisely:

1) The baseline simulation study with data generated by `compcodeR` with parameter choices previously described.
2) A simulation study with mild overdispersion obtained by halving the dispersion parameters of the baseline scenario.
3) A simulation study with strong overdispersion obtained by doubling the dispersion parameters of the baseline scenario.
4) A simulation study with outliers (a fraction of 10% of outliers has been obtained by multiplying the counts by a random number between 5 and 10).

The outliers have been generated by using the option "single" implemented in the `compcodeR` package. In order to evaluate the capability of controlling the type-I error rate, we generated 1,000 datasets in each setting previously described. The mixture model with $K = 5$ components has been estimated on the data and the three proposed test statistics have been computed. The other methods have been implemented with the parameter choice described in the Supplementary materials (Web Appendix D). The computational time in seconds for fitting each method with $p = 1,000, 5,000, 10,000$ is reported in the Web Appendix D Section.

Control of type-I error rate. The adequateness of the statistical procedures can be evaluated by observing the approximation towards the nominal significance level under the null hypothesis as the number of replicates increases. More specifically, with reference to synthetic data with $p = 1{,}000$, we have computed the empirical type-I errors across the 1,000 datasets for each of the 900 non-DE genes.

Figure 1 contains the box-plots of the empirical type-I errors obtained in the baseline scenario by the "Difference" test statistic and by the other considered approaches for the nominal level of 0.05, as the number of replicates increases.

Not surprisingly, the empirical type-I errors of most methods approximate the nominal level as the number of replicates increase, with the exception of *DEXUS*, which is too conservative, and *TSPM* that, on the contrary, has a too high probability of rejecting the null hypothesis when it is true independently on n_j. With very few replicates ($n_j = 3$ or $n_j = 5$) the methods that consistently reach the nominal level are the *MixtNB*, the nonparametric *SAMSeq* and the method *QuasiSeq*. The boxplots obtained in the other three scenarii are reported in the Supplementary materials.

The means, medians, and standard errors of the type-I errors at different confidence levels (0.05, 0.01, and 0.001) for the baseline scenario have been reported in Table 1. It is important to underline that the distribution of the type-I errors is usually skewed for most methods, for this reason we computed both the means and the medians. The corresponding tables for the other tested settings are shown in the Supplementary materials.

Area under the ROC. In order to evaluate the capability of the different methods to discriminate between truly DE and non-DE genes, we considered as a score of evidence for differential expression between the two conditions the nominal p-values provided by each method in each experimental setting, for $n_j = 3, 5, 10$, $p = 1,000, 5,000, 10,000$ and a varying proportion of true DE expression genes 10, 20, and 40%. For all possible score thresholds the false positive rate and the true positive rate have been computed and represented in a ROC curve. In Figure 2 the ROC curves of the different methods computed under the baseline scenario with 5 replicates and $p = 10,000$ genes (10% DE) are shown.

The area under the curve (AUC) can be used as a measure of the overall discriminative capability. Table 2 contains the AUC computed in the baseline setting for all the evaluated methods. The AUC obtained in the other scenarii are reported in the Web Appendix B.

These results indicate that the best performance in terms of discriminative ranking is obtained by the proposed mixture model with the "Difference" test statistic and by the *PoissonSeq* method. Similar results are obtained in the other settings (as reported in the Supplementary materials). *MixtNB* seems to outperform all the methods when the number of replicates is very low and overdispersion due to outliers is high. In the other situations, the *Poissonseq* test is slightly better than *MixtNB*, maybe because when the overdispersion is low, *MixtNB* seeks to estimate over-dispersion parameters that does actually not exist.

Overall, by comparing all the results, the *Quasi-Seq* method is the best strategy in terms of control of type-I error rate, but it is not good as other methods in terms of ability to ranking the truly DE genes ahead the non-DE genes. On the other hand, the *Poissonseq*, which is the best with *MixtNB* in terms of AUC, shows the tendency to be too conservative in terms of type-I empirical errors. The proposed *MixtNB* with the "Difference" test gives very good performance in all the evaluated settings provided that K is sufficiently large.

5. Application to Prostate Cancer Data

We have analyzed data on RNA-Seq data on prostate cancer cells collected in two different conditions: a group of patients has been treated with androgens, and the second one with an inactive compound. The data have been sequenced and analyzed by Li et al. (2008). It is well known that androgen hormones stimulate some genes, and they also have a positive effect in curing prostate cancer cells.

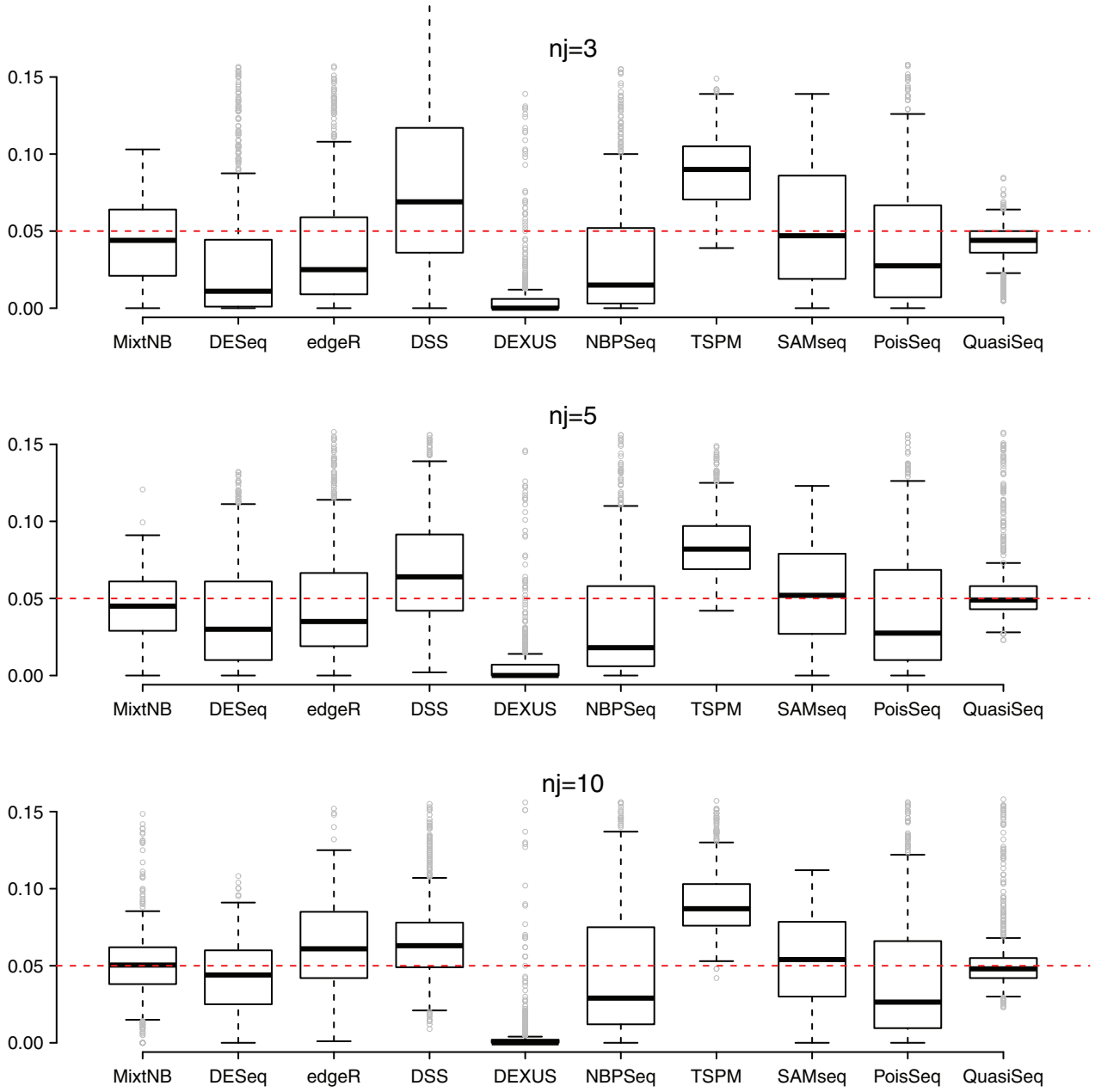

Figure 1. Box-plots of the distribution of the type-I errors computed on the null genes in the first scenario. Comparison between the performances of the proposed "Difference" test statistic and the other methods as n_j increases. The dashed line indicates the nominal value of 0.05 that has been considered.

Therefore, the connection between these stimulated genes and the survival of these cells is a largely studied issue. Seven biological replicates of prostate cancer cells (three for the androgen-treated condition and four for the control-group) for $37,435$ genes have been sequenced using the Illumina 1G Genome Analyzer. Then they have been mapped to the NCBI36 build of the human genome using Bowtie (allowing up to two mismatches) and then the number of reads that corresponded to each Ensembl gene (version 53) was counted. The resulting read count table is available from `https://sites.google.com/site/davismcc/useful-documents`. For the analysis we have considered the $p = 16,424$ genes with mean count greater than 1, because they provide sufficient statistical information on the differential analysis. In order to account for the bias introduced by the different lanes of the experiment and the eventual effect the gene length, we preliminarily normalized the data using quantile-based normalization scheme as described in the Web Appendix C.

5.1. *Analysis and Results*

The proposed NB mixture model has been fitted on the data with a number of components K ranging from 1 to 6. The BIC and AIC criteria suggested $K = 3$ components.

Table 1

Simulation B: means, medians, and standard errors of type-I errors at different confidence levels for scenario 1 across 1,000 simulated datasets

Statistic	$n_j = 3$			$n_j = 5$			$n_j = 10$		
	Confidence level = 0.05								
	Mean	Median	s.e.	Mean	Median	S.E.	Mean	Median	S.E.
MixtNB Diff	0.043	0.044	0.026	0.045	0.045	0.020	0.052	0.051	0.026
MixtNB Ratio	0.084	0.076	0.048	0.075	0.068	0.042	0.063	0.057	0.036
MixtNB LogRatio	0.054	0.076	0.043	0.048	0.068	0.038	0.056	0.057	0.035
DESeq	0.030	0.011	0.043	0.038	0.030	0.033	0.043	0.044	0.023
edgeR	0.043	0.025	0.053	0.048	0.035	0.042	0.063	0.061	0.030
DSS	0.084	0.069	0.064	0.072	0.064	0.044	0.070	0.063	0.038
DEXUS	0.025	0.000	0.100	0.025	0.000	0.101	0.019	0.000	0.090
NBPSeq	0.045	0.015	0.077	0.050	0.018	0.080	0.060	0.029	0.082
TSPM	0.089	0.090	0.022	0.084	0.082	0.020	0.093	0.087	0.025
SAMseq	0.051	0.047	0.037	0.052	0.052	0.032	0.053	0.054	0.030
PoisSeq	0.057	0.027	0.084	0.058	0.027	0.083	0.057	0.026	0.083
QuasiSeq	0.042	0.044	0.012	0.065	0.049	0.044	0.063	0.048	0.055
	Confidence level = 0.01								
	Mean	Median	S.E.	Mean	Median	S.E.	Mean	Median	S.E.
MixtNB Diff	0.011	0.010	0.010	0.009	0.007	0.008	0.011	0.011	0.008
MixtNB Ratio	0.054	0.042	0.045	0.044	0.032	0.039	0.030	0.021	0.031
MixtNB LogRatio	0.018	0.042	0.021	0.014	0.032	0.018	0.014	0.021	0.020
DESeq	0.009	0.001	0.019	0.011	0.005	0.013	0.010	0.009	0.008
edgeR	0.013	0.003	0.028	0.013	0.006	0.020	0.017	0.015	0.013
DSS	0.024	0.014	0.027	0.018	0.014	0.016	0.017	0.014	0.015
DEXUS	0.000	0.000	0.001	0.000	0.000	0.000	0.000	0.000	0.000
NBPSeq	0.018	0.001	0.050	0.019	0.002	0.052	0.023	0.004	0.054
TSPM	0.033	0.033	0.018	0.024	0.023	0.010	0.026	0.024	0.011
SAMseq	0.010	0.002	0.015	0.010	0.007	0.009	0.010	0.008	0.008
PoisSeq	0.014	0.001	0.044	0.014	0.000	0.043	0.013	0.001	0.041
QuasiSeq	0.012	0.010	0.007	0.013	0.010	0.009	0.020	0.010	0.033
	Confidence level = 0.001								
	Mean	Median	S.E.	Mean	Median	S.E.	Mean	Median	S.E.
MixtNB Diff	0.003	0.001	0.003	0.001	0.000	0.002	0.001	0.001	0.002
MixtNB Ratio	0.033	0.021	0.039	0.024	0.013	0.031	0.016	0.007	0.025
MixtNB LogRatio	0.006	0.021	0.010	0.003	0.013	0.008	0.003	0.007	0.012
DESeq	0.003	0.000	0.008	0.002	0.000	0.004	0.002	0.001	0.002
edgeR	0.004	0.000	0.013	0.003	0.000	0.008	0.003	0.002	0.004
DSS	0.004	0.001	0.007	0.003	0.001	0.003	0.002	0.002	0.003
DEXUS	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000
NBPSeq	0.007	0.000	0.031	0.007	0.000	0.033	0.009	0.000	0.035
TSPM	0.014	0.011	0.011	0.006	0.005	0.005	0.005	0.004	0.003
SAMseq	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000
PoisSeq	0.002	0.000	0.016	0.002	0.000	0.018	0.002	0.000	0.018
QuasiSeq	0.003	0.001	0.004	0.001	0.001	0.002	0.002	0.001	0.003

Convergence has been obtained with 50 iterations of the EM-algorithm at the log-likelihood of −384,886 (BIC = 1,088,660, AIC = 835,479). Differential expression analysis has been conducted by computing the three proposed test statistics. For comparative purposes we have performed differential analysis using the *DESeq*, *edgeR* and *DSS* methods with the versions and parameters described in the Web Appendix D.

All the obtained p-values have been adjusted following the procedure of Benjamini and Hochberg (1995) in order keep under control the total type-I error in multiple comparison testing. In Table 3 the number of genes declared DE by each method at the confidence levels of 0.05, 0.01, and 0.001 is shown. The different methods detect a proportion of DE genes ranging from about 10 to 25%. In order to investigate the degree of accordance between two methods, we measured the proportion between the number of genes declared DE jointly by both methods and the average number of the genes declared DE marginally at a certain confidence level.

The first panel of Figure 3 shows the pairwise comparison between the proposed "Difference" test statistic and the

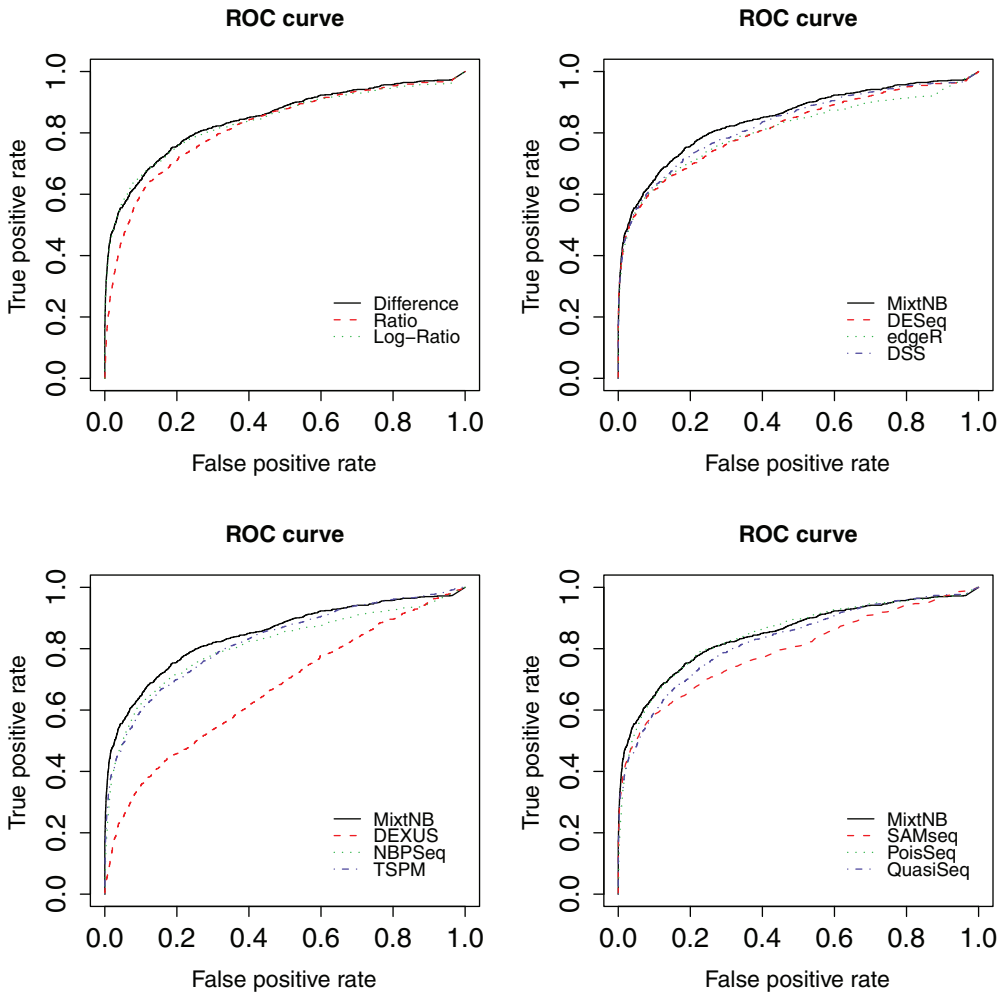

Figure 2. ROC curve of the different methods in the baseline scenario with $p = 10,000$ and $n_j = 5$.

DESeq, *edgeR*, and *DSS* methods. The other two pictures of Figure 3 show the same results for the "Ratio" and "Log Ratio" test statistics, respectively.

It is clear from these graphs that the proposed test statistics provide results that are strongly consistent with the ones obtained by *edgeR* and *DSS*, with a degree of accordance of about 90% when the "Difference" or the "Log Ratio" is used. The set of DE genes detected by *DESeq* seems to be slightly different by the ones selected by all the other methods, even if the accordance level is between 60 and 80%.

Web Figures 5, 6, and 7 in the Supplementary material show the comparison between the rankings of the adjusted p-values between each proposed test statistics and the other considered methods. These graphs confirm the results that has been described above.

6. Discussion and Concluding Remarks

We proposed a novel framework for the differential analysis of count data in the negative binomial setting, especially designed for the analysis of RNA-Seq data. Like several others already proposed, our approach accounts for the heterogeneity of the overdispersion parameter across genes, but the use of a mixture model to the aim is novel. Our approach is fully consistent in terms of parameter estimation and hypothesis testing. As a result, the type-I error of the proposed test is controlled and the discriminative power between differentially and not differentially expressed genes is elevate and outperforms most of alternative methods. The comparative study we performed shows that some popular testing procedures like *DEseq* or *edgeR* actually do not control the type-I error. This lack of control is likely to be due to the post-processing of the overdispersion parameter, which is not accounted for by the null-distribution of the test statistics. The control achieved by *DSS*, which also combines consistent parameter estimation and testing methods, is similar to ours.

Among the three proposed statistical test procedures the "Difference" test seems to be preferable. The simulation results indicate that it is the more consistent in approximating the nominal p-value in all the analyzed settings, while the "Ratio" and the "Log Ratio" tests have generally higher probability of rejecting the null hypothesis when it is true. At the same time, it is also the best in terms of discriminative power between DE and not DE genes as indicated by the ROC/AUC. A possible interpretation of this is related to the theoretical fact that, despite the "Ratio" and the "Log

Table 2
Simulation B: Area Under the ROC Curve (AUC) in the baseline scenario

	$p = 1000$								
	$n_j = 3$			$n_j = 5$			$n_j = 10$		
% of DE:	10%	20%	40%	10%	20%	40%	10%	20%	40%
MixtNB Diff	0.737	0.779	0.814	0.824	0.799	0.856	0.906	0.900	0.902
MixtNB Ratio	0.712	0.752	0.790	0.807	0.777	0.820	0.879	0.889	0.895
MixtNB LogRatio	0.733	0.781	0.808	0.815	0.796	0.844	0.909	0.903	0.905
DESeq	0.711	0.753	0.790	0.829	0.784	0.834	0.860	0.871	0.891
edgeR	0.682	0.717	0.760	0.753	0.666	0.729	0.796	0.791	0.795
DSS	0.698	0.735	0.776	0.810	0.765	0.821	0.852	0.855	0.880
DEXUS	0.704	0.724	0.755	0.761	0.666	0.728	0.813	0.795	0.778
NBPSeq	0.618	0.602	0.624	0.667	0.621	0.658	0.691	0.689	0.699
TSPM	0.693	0.718	0.762	0.758	0.668	0.723	0.807	0.794	0.788
SAMseq	0.668	0.724	0.754	0.781	0.750	0.808	0.837	0.837	0.861
PoisSeq	0.741	0.787	0.815	0.837	0.809	0.862	0.915	0.908	0.914
QuasiSeq	0.708	0.739	0.785	0.819	0.778	0.831	0.853	0.861	0.884
	$p = 5000$								
	$n_j = 3$			$n_j = 5$			$n_j = 10$		
% of DE:	10%	20%	40%	10%	20%	40%	10%	20%	40%
MixtNB Diff	0.797	0.799	0.795	0.847	0.852	0.834	0.889	0.899	0.902
MixtNB Ratio	0.753	0.771	0.772	0.818	0.821	0.816	0.868	0.882	0.893
MixtNB LogRatio	0.781	0.787	0.789	0.842	0.843	0.826	0.887	0.897	0.901
DESeq	0.771	0.785	0.778	0.828	0.826	0.817	0.875	0.882	0.880
edgeR	0.754	0.762	0.730	0.808	0.794	0.739	0.861	0.763	0.666
DSS	0.756	0.772	0.764	0.816	0.812	0.808	0.863	0.869	0.867
DEXUS	0.753	0.773	0.744	0.807	0.797	0.747	0.856	0.766	0.675
NBPSeq	0.639	0.635	0.634	0.676	0.662	0.659	0.692	0.698	0.697
TSPM	0.746	0.763	0.727	0.819	0.798	0.741	0.871	0.753	0.657
SAMseq	0.744	0.752	0.743	0.793	0.794	0.785	0.846	0.847	0.842
PoisSeq	0.806	0.807	0.806	0.850	0.855	0.844	0.895	0.903	0.908
QuasiSeq	0.757	0.776	0.772	0.821	0.815	0.811	0.870	0.873	0.875
	$p = 10{,}000$								
	$n_j = 3$			$n_j = 5$			$n_j = 10$		
% of DE:	10%	20%	40%	10%	20%	40%	10%	20%	40%
MixtNB Diff	0.803	0.791	0.784	0.848	0.842	0.842	0.905	0.907	0.894
MixtNB Ratio	0.771	0.762	0.761	0.821	0.812	0.818	0.888	0.891	0.883
MixtNB LogRatio	0.797	0.781	0.774	0.841	0.834	0.837	0.900	0.906	0.892
DESeq	0.780	0.772	0.763	0.831	0.824	0.822	0.885	0.886	0.874
edgeR	0.756	0.737	0.724	0.808	0.783	0.771	0.865	0.848	0.808
DSS	0.766	0.758	0.751	0.817	0.809	0.807	0.872	0.872	0.861
DEXUS	0.766	0.748	0.731	0.811	0.788	0.776	0.862	0.849	0.814
NBPSeq	0.640	0.628	0.621	0.667	0.664	0.653	0.696	0.701	0.683
TSPM	0.758	0.732	0.722	0.824	0.788	0.776	0.870	0.856	0.810
SAMseq	0.744	0.734	0.733	0.793	0.790	0.788	0.850	0.844	0.840
PoisSeq	0.801	0.800	0.799	0.846	0.844	0.850	0.904	0.909	0.906
QuasiSeq	0.767	0.756	0.753	0.826	0.815	0.813	0.877	0.881	0.866

Ratio," the "Difference" statistical test does not require the use of the Delta approximation method, and for this reason, it works better if the number of available observations is very low. A deep investigation on the empirical behavior of the three statistics on simulated data has shown that the rate of convergence of the three statistics is strongly affected by the level of dispersion in the data. Usually, the "Difference" statistics is approximately Gaussian even with very limited number of replicates (3 or 4), the "LogRatio" needs at least a double number of samples and the "Ratio" has typically a worse behavior, thus indicating that its use is not advisable in case of few replicates.

The proposed method has been applied to prostate cancer cells data previously analyzed in Li et al. (2008) where 2 groups of patients (with $n_1 = 3$ and $n_2 = 4$ replicates) have been observed. When the number of replicates is very low, as

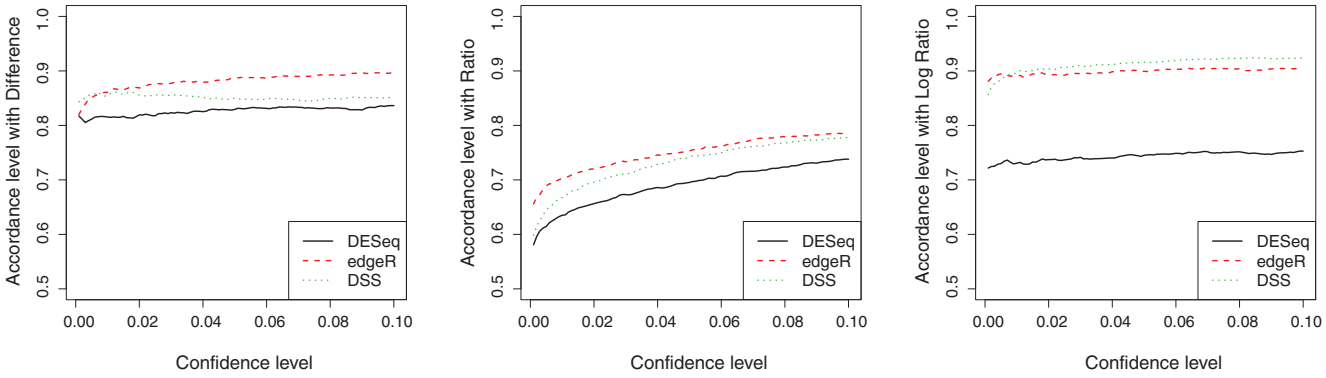

Figure 3. Proportion of genes declared DE as the confidence level increases for the different methods where the proposed "Difference" (panel a), "Ratio" (panel b) and "Log Ratio" test statistic (panel c) are taken as baseline.

in this illustrative real example, the results from simulation study suggest that *DEseq* and *edgeR* tend to be conservative in detecting DE genes, while *DSS* appears to reject the true null hypothesis too often (see Table 1 and the first plot of graph 1). This behavior seems to be confirmed by the empirical results in Table 3 that reports the number of genes declared DE for all the compared methods at different confidence levels, even if these numbers likely include both the true DE genes and the false positive. An important insight of the real application is that, despite this remarkable differences in the general inferential behavior of the tested methods that should be kept in mind, the degree of accordance between the methods in detecting the DE genes in generally high. As shown in the first plot of graph 3 the "Difference" statistical test has a degree of accordance between 80 and 90% with the other evaluated methods. Only 10–15% of genes receive a different characterization and only on them further investigation could be addressed.

In this article, we focus on two sample comparison, but the procedure can indeed be adapted to any contrast, in an obvious manner, especially when using the "Difference" statistics. In a similar way, because our approach can be cast in the general linear framework, normalization or correction for some exogenous effects could also be considered.

7. Supplementary Materials

Web Appendix A referred in Section 2 contains the methodological steps of the EM algorithm. This Appendix and Web Appendix B and D, referenced in Section 4, and Web Appendix C, referenced in Section 5, are available with this article at the *Biometrics* website on Wiley Online Library.

Table 3
Number of genes declared DE for all the compared methods at different confidence levels (adjusted p-values)

Statistic	$\alpha = 0.05$	$\alpha = 0.01$	$\alpha = 0.001$
Difference	3167	2146	1360
Ratio	3538	2591	1914
Log ratio	4254	2941	2024
DESeq	2695	1828	1271
edgeR	3918	2774	1886
DSS	4215	2737	1737

References

Anders, S. and Huber, W. (2010). Differential expression analysis for sequence count data. *Genome Biology* **11**, R106.

Auer, P. and Doerge, R. (2011). A two-stage poisson model for testing RNA-seq data. *Statistical Applications in Genetics and Molecular Biology* **10**, 1–26.

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 (Methodological)* **57**, 289–300.

Delmar, P., Robin, S., Roux, T., and Daudin, J. (2005). Mixture model on the variance for the differential analysis of gene expression data. *Journal of the Royal Statistical Society, Series C (Applied Statistics)* **54**, 31–50.

Dempster, A. P., Laird, N. M., and Rubin, D. B. (1977). Maximum likelihood from incomplete data via the EM algorithm. *Journal of the Royal Statistical Society, Series B (Methodological)* **39**, pages 1–38.

Di, Y., Schafer, D., Cumbie, J., and Chang, J. (2011). The NBP negative binomial model for assessing differential gene expression from RNA-Seq. *Statistical Applications in Genetics and Molecular Biology* **10**, 1–28.

Fraley, C. and Raftery, A. E. (2002). Model-based clustering, discriminant analysis and density estimation. *Journal of the American Statistical Association* **97**, 611–631.

Frazee, A. C., Sabunciyan, S., Hansen, K. D., Irizarry, R. A., and Leek, J. T. (2014). Differential expression analysis of RNA-seq data at single-base resolution. *Biostatistics* pages 413–426.

Hardcastle, T. and Kelly, K. (2010). BaySeq: Empirical Bayesian methods for identifying differential expression in sequence count data. *BMC Bioinformatics* **11**, 1–15.

Klambauer, G., Unterthiner, T., and Hochreiter, S. (2013). DEXUS: Identifying differential expression in RNA-Seq studies with unknown condtions. *Nucleics Acids Research* **42**, 1–11.

Law, C., Chen, Y., Shi, W., and Smyth, G. (2014). Voom: Precision weights unlock linear model analysis tools for RNA-seq read counts. *Genome Biology* **15**, R29.

Li, H., Lovci, M. T., Kwon, Y. S., Rosenfeld, M. G., Fu, X. D., and Yeo, G. W. (2008). Determination of tag density required for

digital transcriptome analysis: Application to an androgen-sensitive prostate cancer model. *Proceedings of the National Academy of Sciences* **105**, 20179–20184.

Li, J. and Tibshirani, R. (2013). Finding consistent patterns: A nonparametric approach for identifying differential expression in RNA-Seq data. *Statistical Methods in Medical Research* **22**, 519–523.

Li, J., Witten, D., Johnstone, I., and Tibshirani, R. (2012). Normalization, testing, and false discovery rate estimation for RNA-sequencing data. *Biostatistics* **13**, 523–538.

Love, M., Huber, W., and Anders, S. (2014). Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. *Genome Biology* **15**, 550.

Lund, S., Nettleton, D., McCarthy, D., and Smyth, G. (2012). Detecting differential expression in RNA-sequence data using quasi-likelihood with shrunken dispersion estimates. *Statistical Applications in Genetics and Molecular Biology* **11**, 8.

Marioni, J., Mason, C., Mane, S., Stephens, M., and Gilad, Y. (2008). RNA-seq: An assessment of techincal reproducibility and comparison with gene expression arrays. *Genome Research* **18**, 1509–1517.

McCarthy, D., Chen, Y., and Smyth, G. (2012). Differential expression analysis of multifactor RNA-seq experiments with respect to biological variation. *Nucleics Acids Research* **40**, 4288–4297.

McLachlan, G. and Peel, D. (2000). *Finite Mixture Models, Willey Series in Probability and Statistics*. New York: John Wiley & Sons.

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 Smyth, G. K. (2008). Small-sample estimation of negative binomial dispersion, with application to SAGE data. *Biostatistics* **9**, 321–332.

Soneson, C. (2014). compcodeR – an R package for benchmarking differential expression methods for RNA-seq data. *Bioinformatics* **30**, 2517–2518.

Soneson, C. and Delorenzi, M. (2013). A comparison of methods for differential expression analysis of RNA-seq data. *BMC Bioinformatics* **14**, 91.

Soon, W., Hariharan, M., and Snyde, M. (2013). High-throughput sequencing for biology and medicine. *Molecular Systems Biology* **9**, 1–14.

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

van der Vaart, A. (2000). *Asymptotic Statistics*. Cambridge Series in Statistical and Probabilistic Mathematics. Cambridge University Press.

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.

Wu, H., Wang, C., and Wu, Z. (2013). A new shrinkage estimator for dispersion improves differential expression detection in RNA-seq data. *Biostatistics* **14**, 232–243.

Yu, D., Huber, W., and Vitek, O. (2013). Shrinkage estimation of dispersion in negative binomial models for RNA-seq experiments with small sample size. *Bioinformatics* **29**, 1275–1282.

Received October 2014. *Revised September* 2015. *Accepted October* 2015.