Gene Expression

Excess False Positive Rates in Methods for Differential Gene Expression Analysis using RNA-Seq Data

David M. Rocke¹,²,³,*, Luyao Ruan¹,²,³, Yilun Zhang¹, J. Jared Gossett²,³, Blythe Durbin-Johnson¹, and Sharon Aviran²,³

¹Division of Biostatistics, Department of Public Health Sciences, University of California, Davis
²Department of Biomedical Engineering, University of California, Davis
³Genome Center, University of California, Davis

Received on XXXXX; revised on XXXXX; accepted on XXXXX

Associate Editor: XXXXXXX

ABSTRACT

Motivation: An important property of a valid method for testing for differential expression is that the false positive rate should at least roughly correspond to the p-value cutoff, so that if 10,000 genes are tested at a p-value cutoff of 10⁻⁴, and if all the null hypotheses are true, then there should be only about 1 gene declared to be significantly differentially expressed. We tested this by resampling from existing RNA-Seq data sets and also by matched negative binomial simulations.

Results: Methods we examined, which rely strongly on a negative binomial model, such as edgeR, DESeq, and DESeq2, show large numbers of false positives in both the resampled real-data case and in the simulated negative binomial case. This also occurs with a negative binomial generalized linear model function in R. Methods that use only the variance function, such as limma-voom, do not show excessive false positives, as is also the case with a variance stabilizing transformation followed by linear model analysis with limma. The excess false positives are likely caused by apparently small biases in estimation of negative binomial dispersion and, perhaps surprisingly, occur mostly when the mean and/or the dispersion is high, rather than for low-count genes.

Contact: dmrocke@ucdavis.edu, lruan@ucdavis.edu, yilzhang@ucdavis.edu, gt4636b@gatech.edu, bpdurbin@ucdavis.edu, saviran@ucdavis.edu.

Supplementary Information: The computational tools developed for this study are freely available via our website http://dmrocke.ucdavis.edu/software.html. They can be downloaded as R code or run directly through an interactive web-based shiny application to reproduce the analysis presented here per a user’s choice of dataset and the methods to be evaluated.

¹To whom correspondence should be addressed.

1 INTRODUCTION AND BACKGROUND

1.1 Gene Expression

Transcriptome profiling is a powerful means of characterizing a cell’s state and as such, is key to elucidating cellular processes, dissecting genotype-phenotype relationships, and understanding health and disease. As genes are expressed at levels that can vary greatly both within a cellular sample and between samples, it is imperative not only to quantify these expression levels robustly but also to detect and assess between-sample differences with high fidelity.

For nearly two decades, cDNA microarrays have been the platform of choice for profiling complex mixtures of RNAs sampled from the transcriptome (Bertone et al. (2004), Subramanian et al. (2005)). Microarray technology transformed and accelerated gene expression studies by enabling large-scale and genome-wide measurements via parallel, automated, and cost-efficient hybridization-based detection of nucleic acids. While it accurately quantifies relative prevalence of transcripts that derive from a set of pre-identified genes, this technology also suffers major limitations. Primarily, it cannot support exploratory studies or facilitate de novo discoveries, as it requires a priori design and fabrication of the array with the genes that are to be probed. Such considerations essentially preclude unannotated genes or unknown splice variants from being detected as well as limit the number of variants that a single experiment can assay. Other issues include limited dynamic range and imperfections in signal acquisition, which warrant dedicated and non-trivial analysis methodology.

A fundamentally new approach to large-scale and comprehensive RNA analysis has emerged with the recent advent of next-generation DNA sequencers (Wold and Myers (2008)). While this technological advance was initially directed at massive genome sequencing, it was soon harnessed to provide sequencing-based
RNA transcript quantification, dubbed RNA-Seq (Mortazavi, Williams, McCue, Schaeffer, and Wold (2008), Nagalakshmi et al. (2008)). RNA-Seq was rapidly adopted and standardized due to the assay’s simplicity, affordability, enhanced information content, and scalability; it has become standard practice in molecular biology and biomedical research (Van Keuren-Jensen, Keats, and Craig (2014) S. Li et al. (2014)). This paradigm shift clearly contributed to the scaling trajectory of gene expression studies, but most importantly, it expanded their scope by allowing detection of transcripts de novo and at very low abundances. Yet, such unprecendented level of detail also presents multiple new informatics challenges, in particular with respect to transcript reconstruction, quantification, and comparative or differential analysis. In this paper, we show that certain methods used to test for differential expression in RNA-Seq have very poor type I error properties.

## 1.2 Statistical Tests for Differential Expression

Perhaps the most fundamental choice in a pipeline for differential expression analysis is the basic statistical test used for each gene. Many packages and methods, including DESeq ((Anders & Huber, 2010); Anders and Huber (2012)), DESeq2 (Love, Huber, and Anders (2014)), edgeR (Robinson, McCarthy, and Smyth (2010), McCarthy, Chen, and Smyth (2012), Robinson and Smyth (2007), Robinson and Smyth (2008)) and edgeR-robust (X. Zhou, Lindsay, McCarthy, Chen, and Smyth (2012), Robinson and Smyth (2007), Robinson and Smyth (2010); Anders (2014) ), edgeR ( Robinson, McCarthy, and Smyth (2010) , McCarthy, Chen, and Smyth (2012), Robinson and Smyth (2007), Robinson and Smyth (2008) ) , are based on a negative binomial model for the counts in the expression matrix and explicitly use the negative binomial likelihood. The mathematics of this is detailed in most of the papers supporting these methods, but the rationale seems to be as follows:

1. RNA-Seq data are in the form of counts.
2. The simplest model for count data is the Poisson, in which the mean and the variance are the same.
3. But in most cases, the data are over-dispersed, meaning that the variance is greater than the mean. This may occur when individual observations for a given gene under given conditions are each Poisson, but with a mean that varies from observation to observation.
4. Perhaps the simplest distribution for over-dispersed count data is the negative binomial, which is obtained when each observation is Poisson but the mean varies from sample to sample following a gamma distribution.

One consequence of this assumption is a relationship between the mean and the variance. If a random variable X has a negative binomial distribution with mean $\mu$ and variance $\sigma^2$, then $\sigma^2 = \mu + \theta \mu^2$, where $\theta$ can be called the dispersion parameter.

The limma package (Smyth (2005)), originally developed for gene expression microarray analysis, is based on fitting a standard linear model for each gene. By default, in a standard linear model, the error variance is the same for all observations from the same gene, for treatment or control. After a log-like transformation (B. P. Durbin, Hardin, Hawkins, and Rocke (2002); B. Durbin and Rocke (2003); P. Durbin and Rocke (2004); Rocke and Durbin (2001); Rocke and Durbin (2003)), this might be a plausible assumption for microarray data, though care in the choice of the transformation is important. For RNA-Seq data also, the variance tends to rise with the mean in a very marked fashion, so that the assumption of equal variance cannot hold in general without a transformation.

Taking logarithms can help with this problem, but it fails for low counts, especially zero counts. For the basic treatment vs. control comparison, limma conducts what is actually a pooled two-sample $t$-test, which assumes equal variance. In order to handle the problem of variances trending with the mean, the voom function (Law, Chen, Shi, and Smyth (2014)) was written as a front end for limma, yielding a method called limma-voom. The voom function calculates variance-based weights that, together with a started log transformation, are meant to stabilize the variance.

In addition, we consider a straightforward negative-binomial-based statistical test arising from the generalized linear model regression formulation with the negative binomial distribution as implemented in the R function glm.nb() in the MASS package (Ripley (2011)). This is the likelihood ratio test (Neyman and Pearson (1928)) which compares the model to a simpler model, which is similar to the ANOVA $F$-test in ordinary linear regression (Snedecor (1934)). In the case of a two-condition experiment, we compare a model with the condition as a predictive factor to one that has only the mean as a predictor. We could also use the Wald test on the coefficient, though the performance of this test is even worse than that of the likelihood ratio test.

We will also consider the application of a transformation to the counts before applying a test. As described in Bartlett’s 1947 paper (Bartlett (1947)), once a variance function (relationship between the variance and the mean) has been defined, one can derive a transformation that approximately stabilizes the variance. After some derivation it can be shown that if $\sigma^2 = \mu + \theta \mu^2$, then the transformation

$$f(x) = \ln \left[ x + \sqrt{x^2 + x \theta + \frac{1}{2\theta}} \right]$$

(1.1)

approximately stabilizes the variance, meaning that the variance no longer depends systematically on the mean. Thus, after transformation, we can use methods that don’t explicitly model the variance as a function of the mean, such as linear regression, the analysis of variance, or in the simple two-condition case, the pooled $t$-test. One way to estimate the best value of $\theta$ is to regress the gene-and-condition-specific variance on the gene-and-condition-specific mean, and find a value of $\theta$ for which the variance neither increases systematically with the mean nor decreases systematically. We can operationalize this by finding a value of $\theta$ for which the slope of the regression line is zero. This procedure can be used instead of the voom function and then the data can be submitted to limma. We refer to this procedure as limma-trans.

## 2 METHODS—MEASURING PERFORMANCE WHEN THERE ARE NO TRUE GROUP DIFFERENCES

### 2.1 Packages and Estimators Used

This paper is not meant to be a comprehensive review of packages and methods for differential expression analysis of RNA-Seq data. It is instead meant to make a specific point, which is that methods based on estimates of the parameters of the negative binomial distribution can result in highly inflated false positive levels. We chose one commonly used package based on the negative binomial...
significant at p = 10. With, say, 10,000 genes, we should have on the order of 1 gene
there are no systematic differences, we should find that the fraction
many procedures on other packages and methods such as DESeq2 and
edgeR-robust with essentially similar results. Of course, many

We use three publicly available data sets, the Bottomly data
(2010), Montgomery et al. (2010)), as obtained from  the recount
site (http://bowtie-bio.sourceforge.net/recount/). The Bottomly
and the Pickrell/Montgomery (“MontPick”) data (Pickrell et al.
(2010)), as obtained from the recount site, we filtered the MontPick data to eliminate any row with a

2.2 Resampling from Existing Data Sets
We use three publicly available data sets, the Bottomly data
(Bottomly et al. (2011)), the Cheung data (Cheung et al. (2010)),
and the Pickrell/Montgomery (“MontPick”) data (Pickrell et al.
(2010), Montgomery et al. (2010)), as obtained from the recount
site (http://bowtie-bio.sourceforge.net/recount/). The Bottomly
data consist of RNA-Seq results from 21 samples of gene expres-
sion from the CEPH (Utah residents with ancestry from northern and western Europe) (CEU)
subset, and 17 of the samples were female and 24 were male.
The MontPick data consist of RNA-Seq results from 129 human
samples of B-cell gene expression from the HapMap project. There were 60 from the above-referenced CEU subset and 69 from the
Yoruba people in Ibadan, Nigeria (YRI).

Our resampling method is to choose a sample size \( n = n_1 = n_2 \) such that \( n_1 + n_2 \) is no larger than the smaller of the two group sizes. We
then select \( n_1 \) samples from group 1 at random and compare them with \( n_2 \) samples randomly selected from the same group. Since
there are no systematic differences, we should find that the fraction
significant is within chance range of the nominal significance level.
With, say, 10,000 genes, we should have on the order of 1 gene
significant at \( p = 10^{-4} \). Of course we could be unlucky and the 3 vs.
3 could have some factor that is more similar within these random
groups than across them, but if we repeat the random selection a
sufficient number of times, that should average out. Thus, if these
methods are performing as advertised, when we conduct a “null”
experiment like this, we should not have substantially more significant
genes than is implied by the significance level. In this case, we did this 100 times in group 1, 100 times in group 2, and
aggregated the results. For each of the three data sets, we conducted this
experiment with \( n = 3, 4, 5 \) in group 1 and then also for group 2.
For the Cheung data set, we also add an 8 vs. 8 comparison and for
the MontPick data set a 10 vs. 10 comparison, since the larger
sample sizes allow this. In this case, we sampled each trial inde-
pendently so there can be duplicate samples in some cases (that is
the comparison of \( n_1 \) vs. \( n_2 \) might be duplicated. We used this
method because it is simple, easy to understand and interpret, and
easy to reproduce. If desired, the duplicates could be eliminated with hashing or using the R package partitions (Hankin (2006)).

We then compared the number of nominally significant genes with
the expected number of significant genes under the null hypothesis.
A method with well-calibrated \( p \)-values should have a number of
significant genes close to the nominal number. With 10,000 genes
and 100 replications in each of the two groups, we have 2 million
\( p \)-values. The expected number less than \( 10^{-4} \) (for example) is then 200.

We chose the \( 10^{-4} \) and \( 10^{-6} \) levels to make allowance for multiple
comparisons. For comparison, with 10,000 genes, the cutoff with
the Bonferroni method for a 5% test is \( 5 \times 10^{-4} \). We did not use the
FDR-adjusted \( p \)-values because each of those is a complex func-
tion of the entire vector of \( p \)-values, and this can make interpreta-
tion more difficult, but the raw \( p \)-values for a typical method in this
data set for FDR-adjusted \( p \)-values of 5% are in the range of
\( 10^{-4} \) and lower. Note that if the null hypothesis is always true, then
all genes identified as significant are false positives. If we ask for a
set of genes with FDR control at 5%, that means that the number of
false positives should be no more than 5% of those identified as
significant, and since all genes identified as significant are false
positives, that means that an FDR-controlling method should never
identify any genes as positive, at least on the average.

2.3 Negative Binomial Simulations
In this case, we took one relatively homogeneous data set (the 33
females from the CEU samples in the MontPick data set) and made
a crude estimate of the mean \( \mu \) and dispersion \( \theta \) of each gene. We
estimated \( \mu \) by the sample mean. If the sample variance was smaller
than the mean, we used the Poisson distribution for that gene.
Otherwise we estimated \( \theta \) by

\[
\hat{\theta} = \frac{s^2 - \bar{x}}{\bar{x}^2}
\]

Although this may not be the optimal method for estimating negative
binomial parameters, it does not in this case matter, since the
only real purpose was to obtain a distribution of means and disper-
sions that somewhat resembled real data. Before generating the
subset, we filtered the MontPick data to eliminate any row with a
mean count of less than 1 or where the variance of the CEU female
counts were 0, leaving 8122 genes. Clearly, the data set used to
obtain the means and variances of the counts is not highly i-

3 RESULTS

3.1 Results for resampling from existing data sets.
Figure 1 shows the results for the resampling experiment with the MontPick data set and \( n = 5 \). The dramatic departures from the null can be seen for edgeR and the likelihood ratio test from the R function glm.nb. As mentioned previously, the Wald test from glm.nb is even worse.

Table 1. False positives at \( p = 10^{-4} \) for four estimators in null simulations obtained by resampling three data sets.

<table>
<thead>
<tr>
<th>Data Set</th>
<th>n</th>
<th>Expected</th>
<th>edgeR</th>
<th>Limma voom</th>
<th>Limma trans</th>
<th>lr</th>
</tr>
</thead>
<tbody>
<tr>
<td>Bottomly 3</td>
<td>237</td>
<td>4560</td>
<td>1110</td>
<td>508</td>
<td>2790</td>
<td></td>
</tr>
<tr>
<td>Bottomly 4</td>
<td>237</td>
<td>1103</td>
<td>124</td>
<td>200</td>
<td>951</td>
<td></td>
</tr>
<tr>
<td>Bottomly 5</td>
<td>237</td>
<td>2155</td>
<td>571</td>
<td>426</td>
<td>1328</td>
<td></td>
</tr>
<tr>
<td>Cheung 3</td>
<td>181</td>
<td>3798</td>
<td>198</td>
<td>326</td>
<td>14113</td>
<td></td>
</tr>
<tr>
<td>Cheung 4</td>
<td>181</td>
<td>2634</td>
<td>123</td>
<td>171</td>
<td>9006</td>
<td></td>
</tr>
<tr>
<td>Cheung 5</td>
<td>181</td>
<td>2845</td>
<td>200</td>
<td>227</td>
<td>6359</td>
<td></td>
</tr>
<tr>
<td>Cheung 8</td>
<td>181</td>
<td>1824</td>
<td>117</td>
<td>146</td>
<td>2707</td>
<td></td>
</tr>
<tr>
<td>MontPick 3</td>
<td>162</td>
<td>2099</td>
<td>144</td>
<td>194</td>
<td>11537</td>
<td></td>
</tr>
<tr>
<td>MontPick 4</td>
<td>162</td>
<td>1755</td>
<td>168</td>
<td>164</td>
<td>8953</td>
<td></td>
</tr>
<tr>
<td>MontPick 5</td>
<td>162</td>
<td>1450</td>
<td>129</td>
<td>135</td>
<td>6047</td>
<td></td>
</tr>
<tr>
<td>MontPick 10</td>
<td>162</td>
<td>1146</td>
<td>139</td>
<td>116</td>
<td>1215</td>
<td></td>
</tr>
</tbody>
</table>

Table 2. False positives at \( p = 10^{-6} \) for four estimators in null simulations obtained by resampling three data sets.

<table>
<thead>
<tr>
<th>Data Set</th>
<th>n</th>
<th>Expected</th>
<th>edgeR</th>
<th>Limma voom</th>
<th>Limma trans</th>
<th>lr</th>
</tr>
</thead>
<tbody>
<tr>
<td>Bottomly 3</td>
<td>2.37</td>
<td>1071</td>
<td>41</td>
<td>11</td>
<td>457</td>
<td></td>
</tr>
<tr>
<td>Bottomly 4</td>
<td>2.37</td>
<td>160</td>
<td>1</td>
<td>1</td>
<td>391</td>
<td></td>
</tr>
<tr>
<td>Bottomly 5</td>
<td>2.37</td>
<td>350</td>
<td>18</td>
<td>4</td>
<td>322</td>
<td></td>
</tr>
<tr>
<td>Cheung 3</td>
<td>1.81</td>
<td>674</td>
<td>2</td>
<td>12</td>
<td>9398</td>
<td></td>
</tr>
<tr>
<td>Cheung 4</td>
<td>1.81</td>
<td>351</td>
<td>0</td>
<td>1</td>
<td>6350</td>
<td></td>
</tr>
<tr>
<td>Cheung 5</td>
<td>1.81</td>
<td>353</td>
<td>1</td>
<td>5</td>
<td>4509</td>
<td></td>
</tr>
<tr>
<td>Cheung 8</td>
<td>1.81</td>
<td>136</td>
<td>0</td>
<td>0</td>
<td>1896</td>
<td></td>
</tr>
<tr>
<td>MontPick 3</td>
<td>1.62</td>
<td>410</td>
<td>5</td>
<td>9</td>
<td>7045</td>
<td></td>
</tr>
<tr>
<td>MontPick 4</td>
<td>1.62</td>
<td>307</td>
<td>5</td>
<td>7</td>
<td>6001</td>
<td></td>
</tr>
<tr>
<td>MontPick 5</td>
<td>1.62</td>
<td>166</td>
<td>5</td>
<td>6</td>
<td>4319</td>
<td></td>
</tr>
<tr>
<td>MontPick 10</td>
<td>1.62</td>
<td>75</td>
<td>1</td>
<td>0</td>
<td>1215</td>
<td></td>
</tr>
</tbody>
</table>

Table 3 shows the number of false positives at \( p = 10^{-4} \) in 1000 simulated null negative binomial data sets, each with 8122 genes for four sample sizes and for edgeR and limma-voom. It is apparent that the number of false positives for edgeR is large even under the negative binomial distribution, though the number does decline as \( n \) gets larger. Limma-voom (and limma-trans, not shown) do not have large numbers of false positives.

Table 3. False positives at \( p = 10^{-4} \) for edgeR and limma-voom in negative binomial simulated data.

<table>
<thead>
<tr>
<th>n</th>
<th>Expected</th>
<th>edgeR</th>
<th>limma-voom</th>
</tr>
</thead>
<tbody>
<tr>
<td>3</td>
<td>812</td>
<td>7138</td>
<td>541</td>
</tr>
<tr>
<td>4</td>
<td>812</td>
<td>6311</td>
<td>949</td>
</tr>
<tr>
<td>5</td>
<td>812</td>
<td>5847</td>
<td>1062</td>
</tr>
<tr>
<td>10</td>
<td>812</td>
<td>3575</td>
<td>915</td>
</tr>
</tbody>
</table>

An interesting characteristic of p-values from edgeR is that the false-positive p-values are not just small, many of them are extremely small. A histogram of the p-values (say for \( n = 5 \)) does not show any apparent excess of false positives (not shown). Only 8% of the p-values are less than 0.10 instead of an expected 10%. But if we examine the p-values that are less than 0.10, then 15% of them are less than 0.01 (with again an expected fraction of 10%). This continues all the way down. Of the p-values less than \( 10^{-3} \), 54% are less than \( 10^{-4} \). Thus, there are a large number of very small p-values.

3.3 Which Genes are Generating the False Positives?
Under the negative binomial, the only parameters determining the distribution of counts are the mean $\mu$ and dispersion $\theta$. It is commonly assumed that it is the low-count genes that cause trouble in the analysis, and it is true that there are often convergence problems in genes with low mean counts, but the false positives are arising from the other end of the parameter distribution.

In Table 4, we have partitioned the means in the CEU female data set into quintiles, and also partitioned the dispersions in to quintiles. The row headings show the geometric mean of the sample means in the quintile and the column headings show the square root of the mean of the estimated dispersions in the quintile. The number in each cell is the number of false positives at $p = 10^{-4}$ for genes in that cell divided by the number of genes falling into that cell for $n = 5$ in 1000 trials of edgeR with negative binomial simulated data. Under the null, the numbers in each cell should be approximately 0.1—the shaded cells are the ones more than twice the nominal.

<table>
<thead>
<tr>
<th>Dispersion</th>
<th>Mean</th>
<th>0.4</th>
<th>0.5</th>
<th>0.6</th>
<th>0.9</th>
<th>1.7</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0.011</td>
<td>0.198</td>
<td></td>
</tr>
<tr>
<td>8</td>
<td>0</td>
<td>0</td>
<td>0.024</td>
<td>0.154</td>
<td>3.24</td>
<td></td>
</tr>
<tr>
<td>32</td>
<td>0</td>
<td>0.015</td>
<td>0.098</td>
<td>0.863</td>
<td>10.676</td>
<td></td>
</tr>
<tr>
<td>108</td>
<td>0.014</td>
<td>0.070</td>
<td>0.358</td>
<td>2.370</td>
<td>11.848</td>
<td></td>
</tr>
<tr>
<td>489</td>
<td>0.032</td>
<td>0.267</td>
<td>0.926</td>
<td>3.583</td>
<td>6.909</td>
<td></td>
</tr>
</tbody>
</table>

What we can see from this table is that the false positives are entirely generated from genes with high means, high dispersions, or both. An implication of this is that the $p$-value distribution under then null is quite different for genes with different means and dispersions, so $p$-values obtained by edgeR and other negative-binomial-based estimators cannot be used to rank genes. Consider the 147 genes that fall into the second quintile in mean and the second quintile in dispersion. Out of 147,000 statistical tests, none is significant at the $p = 10^{-4}$ level, instead of an expected 15. On the other hand, the 146 genes that fall into the fourth quintile in the means and the fourth quintile in the dispersions had 346 tests significant at the $p = 10^{-4}$ level. This means that genes in the higher quintiles of means and dispersions are far more likely to be top genes than those in the lower quintiles, all other things being equal, and therefore that the ranking of genes by $p$-value is unreliable, not just the level of the test.

4 DISCUSSION

There have been numerous reviews and evaluations of methods for RNA-Seq data, many quite comprehensive, and a number of them have compared the power and type I error rates, with varying conclusions. Soneson and Delorenzi (2013), Kvam, Liu, and Si (2012), and van de Wiel, Neerincx, Buffart, Sie, and Verheul (2014) compared the null performance of several methods and found poor control of false discovery rate (FDR) in general. Robles et al. (2012) likewise compared 3 methods and found inflated type I error rates for some methods and conservative performance from others. Reeb and Steibel (2013) compared 3 methods using “plas-modes” (resampled data) and found inflated type I error rates for small significance levels. Guo, Li, Ye, and Shyr (2013) compare 6 methods and conclude that all “suffer from over-sensitivity”. However, Rapaport et al. (2013) compared 7 different methods and concluded that most methods provide good FDR control at the 5% level. Finally, Nookaew et al. (2012) compared 5 different methods with each other and with microarray data and concluded that agreement was good between methods and platforms, and Guo et al. (2013) concluded that agreement between the methods compared was “reasonable”.

Mostly, the examination of Type I error control has been limited to $p$-values such as 0.05 which are important in analysis of a single outcome. But when testing 10,000 genes, Type I error for small $p$-values such as $10^{-8}$ or $10^{-6}$ are more important. Control of the false discovery rate after multiplicity adjustment has also been frequently examined, but this is harder to interpret since each individual FDR-adjusted $p$-value (q-value) depends on the entire vector of $p$-values. In addition, if all the null hypotheses are true, then the actual false discovery rate of any non-empty set of “significant” genes is 100%.

Most likely, the phenomenon of excess false positives in methods based strongly on the negative binomial distribution is a consequence of the known downward bias in the MLE for the negative binomial dispersion (Piegorsch, 1990). While this bias is generally small when viewed from the point of view of parameter estimation, it apparently has a large effect on the tail areas of the test statistic. In any case, it appears that the degree of violation of Type I error control in many existing methods has been greatly underestimated. For the time being, it seems safer to use methods like limma-voom that do not depend explicitly on the negative binomial likelihood.

AUTHORS’ CONTRIBUTIONS

DMR directed the study, did some of the computations, and was primary author of the manuscript. LR wrote much of the code and performed most of the simulation experiments. JG and YZ each wrote some of the code, and did bibliographic research. BDJ and SA wrote parts of the manuscript and edited it and did background research. All participants contributed to the research plan and the
outline of the manuscript in extensive and frequent meetings and discussions.

**FUNDING**

This work was supported by National Institutes of Health (NIH) grants R33AI080604 and UL1 RR024146 (David Rocke) and RO0 HG006860 (Sharon Aviran)

**REFERENCES**


Excess False Positive Rates in Methods for Differential Gene Expression Analysis using RNA-Seq Data


