Yanming Di*, Sarah C. Emerson, Daniel W. Schafer, Jeffrey A. Kimbrel and Jeff H. Chang

Higher order asymptotics for negative binomial regression inferences from RNA-sequencing data

Abstract: RNA sequencing (RNA-Seq) is the current method of choice for characterizing transcriptomes and quantifying gene expression changes. This next generation sequencing-based method provides unprecedented depth and resolution. The negative binomial (NB) probability distribution has been shown to be a useful model for frequencies of mapped RNA-Seq reads and consequently provides a basis for statistical analysis of gene expression. Negative binomial exact tests are available for two-group comparisons but do not extend to negative binomial regression analysis, which is important for examining gene expression as a function of explanatory variables and for adjusted group comparisons accounting for other factors. We address the adequacy of available large-sample tests for the small sample sizes typically available from RNA-Seq studies and consider a higher-order asymptotic (HOA) adjustment to likelihood ratio tests. We demonstrate that 1) the HOA-adjusted likelihood ratio test is practically indistinguishable from the exact test in situations where the exact test is available, 2) the type I error of the HOA test matches the nominal specification in regression settings we examined via simulation, and 3) the power of the likelihood ratio test does not appear to be affected by the HOA adjustment. This work helps clarify the accuracy of the unadjusted likelihood ratio test and the degree of improvement available with the HOA adjustment. Furthermore, the HOA test may be preferable even when the exact test is available because it does not require ad hoc library size adjustments.

Keywords: RNA-Seq; higher-order asymptotics; negative binomial; regression; overdispersion; extra-Poisson variation.

***Corresponding author: Yanming Di,** Department of Statistics, Oregon State University, 44 Kidder Hall, Corvallis, OR 97330, USA, e-mail: diy@stat.oregonstate.edu
Sarah C. Emerson, Daniel W. Schafer, Jeffrey A. Kimbrel and Jeff H. Chang: Oregon State University

1 Introduction

RNA sequencing (RNA-Seq) is the current technology of choice for investigating gene expression (Mortazavi et al., 2008; Nagalakshmi et al., 2008; Wang et al., 2009); it provides unprecedented comprehensiveness, resolution and sensitivity. A typical RNA-Seq experiment involves isolating and randomly fragmenting mRNA (transcriptome), converting mRNA to complementary DNA (cDNA), preparing the cDNA for sequencing, and finally simultaneously sequencing cDNA fragments to produce hundreds of millions of short RNA-Seq reads. To infer gene expression, the RNA-Seq reads are aligned to sequence features in a reference database. The relative frequency of RNA-Seq reads that match sequence features of a gene serves as a measure of that gene's expression.

The frequencies of RNA-Seq reads cannot be adequately modeled by the most commonly used distributions such as normal, binomial or Poisson (Di et al., 2011). Although technical variability in read counts has been demonstrated to be Poisson (Marioni et al., 2008), practically useful models must also incorporate biological variability. The negative binomial (NB) distribution, which may be derived as a mixture of Poisson distributions, is a flexible and convenient choice. Consequently, it serves as the basis of several statistical packages for assessing differential expression from RNA-Seq data, including edgeR (Robinson et al., 2010), DESeq (Anders and Huber, 2010), NBPSeq (Di et al., 2011), and the recent version of Cuffdiff (http://cufflinks.cbcb.umd.edu/manual.html) in Cufflinks (Trapnell et al., 2010). For two-group comparisons, the NB distribution permits an exact test (Robinson and Smyth, 2007) that does not rely on large sample size asymptotic theory. The statistical packages exploit this exact test but differ in how they handle the efficient estimation of the NB dispersion parameter and how they deal with the problematic assumption of equal library sizes. (In this paper, library size refers to the total number of unambiguously aligned sequencing reads for each biological sample.)

The model for comparing two NB means extends easily to regression modeling of NB means. This extension is essential for exploring gene expression as a function of explanatory variables and for comparing groups after accounting for other factors. For example, the R package MASS (Venables and Ripley, 2002) and recent versions of DESeq and edgeR all include implementations of NB regression. The theory for the exact test for two-group comparisons does not extend to the regression setting, however. Asymptotic tests – most notably the Wald test and the likelihood ratio test – are available (Venables and Ripley, 2002; Hilbe, 2007; McCarthy et al., 2012), but are mathematically justified only for large sample sizes. Since RNA-Seq studies currently tend to be based on small sample sizes (for example, three biological replicates for each of two treatment groups, for a total sample size of six), there is an obvious need to examine the suitability of the tests derived for large samples.

In this paper we carry out this examination and, in particular, consider likelihood ratio tests with a higher-order asymptotic (HOA) adjustment (Skovgaard, 2001). Such tests have been shown to be very nearly exact in other situations, even for very small sample sizes. We demonstrate that 1) the HOA-adjusted likelihood ratio test *p*-values are practically indistinguishable from exact test *p*-values in situations where the exact test is available (i.e., two-group comparison), 2) via simulation, that the actual type I error of the test matches the nominal specification in regression settings, for which the exact test is unavailable, and 3) the power of the likelihood ratio test is not apparently affected very much by the HOA adjustment.

This work will help clarify the accuracy of the unadjusted likelihood ratio test and the degree of improvement available with the HOA adjustment. Because the HOA-adjusted test reduces to the unadjusted test when the sample sizes are large enough to ensure the desirable asymptotic test properties, there seems to be little harm in its automatic incorporation into RNA-Seq testing programs, relieving researchers of the need to decide whether or not a small sample adjustment is necessary. Furthermore, we believe the HOA test is a better choice than the exact test for two-group comparison because its theory does not require an assumption of equal library sizes and does not, therefore, require any additional adjustment to overcome the common violation of this assumption in practice.

This paper is organized as follows. Section 2 clarifies the NB regression model. Section 3 introduces higher-order asymptotic inferences. Section 4 shows numerical results. Section 5 includes discussion and conclusion. The Appendices include additional technical details.

2 Model

Let Y_{ij} represent the number of RNA-Seq reads from biological sample j attributed to gene i and let X_{jk} be the value of the k-th explanatory variable associated with biological sample j, for i=1, ..., m; j=1, ..., n; and k=1, ..., p. Let N_j be the total number of unambiguously aligned sequencing reads associated with biological sample j (i.e., $N_j = \sum_{i=1}^{m} Y_{ij}$), which we refer to as the *(observed) library size* of sample j. Our NB regression model for describing the mean expression of gene i as a function of explanatory variables includes the following three components:

1. A NB probability distribution for the frequency of reads:

$$Y_{ij} \sim \mathrm{NB}(\mu_{ij}, \phi_{ij}), \tag{1}$$

where μ_{ij} is the mean and ϕ_{ij} is the NB *dispersion parameter* such that $\mathrm{Var}(Y_{ij}) = \mu_{ij} + \phi_{ij}\mu_{ij}^2$. We assume that frequencies indexed by different i's and j's are independent of one another.

2. A log-linear regression model for the mean as a function of explanatory variables

$$\log(\mu_{ij}) = \log(N_j) + \log(R_j) + \sum_{k=1}^{p} \beta_{ik} X_{jk} \tag{2}$$

where β_{ik} are unknown regression coefficients associated with gene i and the R_j's are optional normalization factors, explained below.

3. A model for the dispersion parameter ϕ_{ij} as a function of the mean:

$$\log(\phi_{ij})=f(\mu_{ij}) \tag{3}$$

where $f(\cdot)$ is a specified model, such as $\alpha_0+\alpha_1\log(\mu_{ij})+\alpha_2(\log(\mu_{ij}))^2$, or the result of a nonparametric regression fitting algorithm, as clarified below.

Component 2 of the model involves two types of normalization, which are explained next. Component 3 is left in a fairly general form so that the extension of statistical tests to regression settings can be based on any of the dispersion parameter models used in the statistical packages (such as edgeR, DESeq and NBPSeq) mentioned in Section 1.

First we explain the normalization in component 2 of the model. The observed library sizes, N_j, differ due to chance variation in the preparation and sequencing of the samples. An appropriate parameter for comparing gene i's expression across different biological samples is the mean relative frequency, μ_{ij}/N_j. A log-linear model for this parameter,

$$\log(\mu_{ij}/N_j)=\sum_{k=1}^{p}\beta_{ik}X_{jk}$$

induces the model in (2) for the mean frequency (except for the presence of R_j, which will be explained next). It is important to note that this model directly accounts for variable library sizes, that inferences using this model do not require equal library sizes, and that the nuisance of library size adjustment [which is necessary for the exact NB test in Robinson and Smyth (2007)] is avoided with NB regression analysis.

The other part of normalization has to do with the *apparent* reduction or increase in expression of non-differentially expressing genes simply to accommodate the increased or decreased expression of truly differentially expressing genes. If, for example, only one gene expresses differently in two groups, its relative frequency of reads will be larger in one of the groups. The relative frequency of all other genes must necessarily be smaller, even though their biological behavior is the same in the two groups. The concern introduced in Robinson and Oshlack (2010) is that this reduction will give a false impression of biological relevance. Since the accommodation for relative frequencies summing to one is shared equally by a very large number of non-differentially expressing genes, we suspect that the effect is usually small, but examples where it is non-ignorable have been demonstrated (Robinson and Oshlack, 2010). Normalization factors, R_j, j=1,..., n, can be included as in model (2), if desired, to account for this possibility. Values of the R_j's can be estimated in a first stage of the analysis and then treated as known in the regression model (2). The "median of fold change" method in Anders and Huber (2010) and the "trimmed mean of M-values" method in Robinson and Oshlack (2010) are two possibilities. Since any choice for the R_j's can be used within the general model, a particular choice is not important for this paper.

Component 3 of the model has to do with important power gains available by pooling information about the unknown parameters associated with dispersion, such as the α's in the example following Eq. (3). These parameters are of little biological interest, but their estimation is a necessary step for producing correct statistical inference about the regression coefficients. For small sample sizes, the power of statistical tests for hypotheses about the β_{ik}'s could be substantially greater if these parameters were known than if they were estimated from the data.

If a parametric model $\log(\phi)=f(\mu;\alpha)$ is assumed, then the likelihood ratio test statistic for a hypothesis about β (the vector of regression coefficients for all genes) is $2l(\hat{\alpha},\hat{\beta})-2l(\tilde{\alpha},\tilde{\beta})$, where $l()$ is the log-likelihood function for counts from all genes combined (and treated as independent); $\hat{\alpha}$ and $\hat{\beta}$ are the unconstrained maximum likelihood estimates; and $\tilde{\alpha}$ and $\tilde{\beta}$ are the maximum likelihood estimates for the null-constrained model. Because the maximization based on all genes is unwieldy, it is more practical to approximate the MLEs by a one- or two-iteration two-step process in which the β's for each gene are estimated individually with the estimated α taken to be known and with α estimated from all genes based on estimated β's.

A further simplifying approximation involves estimating β's for each gene individually by treating the estimated ϕ's as known, via the estimated model $\log(\hat{\phi})=f(\hat{\mu};\hat{\alpha})$. Although there is additional uncertainty in this latter approximation due to the small-sample estimation of $\hat{\mu}$ for each gene, this is the approximation

that we will investigate in this paper. It permits the incorporation of the proposed HOA-adjusted likelihood ratio tests into the edgeR package (Robinson et al., 2010), which uses empirical Bayes estimates of ϕ's, optionally with an underlying trend of ϕ as a function of μ; the DESeq package (Anders and Huber, 2010), which estimates the ϕ's as a function of estimated μ's using nonparametric regression; and the NBPSeq approach (Di et al., 2011), which conducts likelihood analysis for the NBP parameterization of the NB model, in which the log of ϕ is a straight line function of the log of the mean. Our simulations suggest that the HOA adjustment is useful even with this approximation.

We discuss the logical next steps in the evolution of negative binomial regression for RNA-Seq data analysis in Section 5, with particular attention to the "known ϕ" issue. Lund et al. (2012) proposed quasi-likelihood methods to incorporate uncertainty in dispersion parameter estimates.

3 The likelihood ratio test and higher order asymptotic adjustment

The model in Section 2 implies a NB log-linear regression model for each gene, i. We suppress the index i here:

$$Y_j \sim NB(\mu_j, \phi_j)$$
$$\log(\mu_j) = \log(N_j) + \log(R_j) + \sum_{k=1}^{p} \beta_k X_{jk}$$

with the ϕ_j's taken to be known. We wish to test hypotheses about components of β, the vector of regression coefficients. Without loss of generality, suppose that $\beta=(\psi,\nu)$, where $\psi=(\beta_1,\ldots,\beta_q)$ and $\nu=(\beta_{q+1},\ldots,\beta_p)$, and the null hypothesis is $\psi=\psi_0$. In regard to this hypothesis, the q-dimensional parameter ψ is the parameter of interest and ν is a nuisance parameter. We let $\hat{\beta}(y)=\hat{\beta}=(\hat{\psi},\hat{\nu})$ be the maximum likelihood estimator of the full parameter vector and $\tilde{\beta}(y)=\tilde{\beta}=(\tilde{\nu},\psi_0)$ be the maximum likelihood estimator under the null hypothesis.

Under the usual regularity conditions, the likelihood ratio statistic,

$$\lambda = 2\left(l(\hat{\beta}) - l(\tilde{\beta})\right),$$

converges in distribution to a chi-square distribution with degrees of freedom q under the null hypothesis (Wilks, 1938). When ψ is one-dimensional (q=1), the signed square root of the likelihood ratio statistic λ, also called the directed deviance,

$$r = \text{sign}(\hat{\psi} - \psi_0)\sqrt{\lambda}, \tag{4}$$

converges to a standard normal distribution. The latest versions of DESeq (Anders and Huber, 2010) and edgeR (Robinson et al., 2010) both implemented the unadjusted likelihood ratio test for coefficients in NB regression models, but the two packages differ in how they pool information across genes to estimate dispersion parameters. Another R package BBSeq (Zhou et al., 2011) also used unadjusted likelihood ratio test for regression coefficients, but BBSeq models the count variation using a beta-binomial model instead of a negative binomial model.

For testing a one-dimensional parameter of interest (q=1), Barndorff-Nielsen (1986, 1991) showed that a *modified directed deviance*

$$r^* = r - \frac{1}{r}\log(z) \tag{5}$$

is, in wide generality, asymptotically standard normally distributed to a higher order of accuracy than the directed deviance r itself, where z is an adjustment term to be discussed below. Tests based on high-order asymptotic adjustment to the likelihood ratio statistic, such as r^* or its approximation (explained below), are

referred to as higher-order asymptotic (HOA) tests. They generally have better accuracy than corresponding unadjusted likelihood ratio tests, especially in situations where the sample size is small and/or when the number of nuisance parameters ($p-q$) is large.

However, the definition and computation of the adjustment term z in Barndorff-Nielsen's original formulation are difficult in practice. They rely on the specification of an (approximately) ancillary statistic and involve the computation of the *sample space derivatives* – differentiating the likelihood with respect to the $\hat{\beta}$ while holding fixed the value of the ancillary statistic. Except for full-rank exponential families and regression-scale models, the sample space derivatives have to be approximated. Although there is more than one way to approximate the sample space derivatives, we prefer Skovgaard's (1996) approach, which does not require the specification of an ancillary statistic and involves only calculations similar to those involved in computing the expected information. With Skovgaard's approximations, the HOA test becomes practical for general use.

For testing a one-dimensional component of the natural parameter of a full-rank exponential family, Pierce and Peters (1992) provided a relatively simple representation of the HOA adjustment [the term $\frac{1}{r}\log(z)$ in Eq. (5)] and pointed out that there are two aspects in the HOA adjustment: one reducing the effects of nuisance parameter estimation and the other improving the normal approximation to r when the information for the parameter of interest is small. While in general the NB regression model in Section 2 does not belong to a full-rank exponential family, the general point that the HOA adjustment consists of two aspects – a nuisance parameter adjustment and an information adjustment – is still valid and illuminating. (In the special case of two group comparison, when all library sizes are equal and all dispersion parameters are the same, it can be shown that the model in Section 2 belongs to a full-rank exponential family. See Appendix 6.2 for details.)

Skovgaard (2001) gave a comprehensive review of the development of the theory and practice of higher order asymptotics. That paper also presented a generalization of Barndorff-Nielsen's r^* statistic to test for multi-dimensional parameters ($q>1$). The test statistic

$$\lambda^{**}=\lambda\left(1-\frac{1}{\lambda}\log\gamma\right)^2 \tag{6}$$

is constructed by adding an HOA adjustment to the likelihood ratio statistic. In Appendix 6.1, we provide implementation details of the HOA test with Skovgaard's approximations for one-dimensional and multi-dimensional parameters in the context of the NB regression model.

4 Examples and numerical results

We present examples and numerical results to clarify the regression part of the model (2) and to compare the performance of three asymptotic tests: the HOA test based on the r^* statistic (5), the unadjusted likelihood ratio (LR) test based on the r statistic (4), and one other commonly used large sample test based on the Wald statistic Wald (1941, 1943),

$$\left(\hat{\psi}-\psi_0\right)^T\left(\left[\hat{i}\left(\beta\right)^{-1}\right]_{\psi\psi}\right)^{-1}\left(\hat{\psi}-\psi_0\right),$$

where $i(\beta)$ is the Fisher (expected) information. The p-values reported by the glm.nb program in the R package MASS are Wald test p-values. Under the null hypothesis, all three of these tests (HOA, LR, and Wald) have the same asymptotic chi-square distribution with q degrees of freedom.

In Section 4.1, we compare p-values from the three asymptotic tests to p-values from the exact NB test for assessing two-group differential gene expression using an Arabidopsis data set and computer generated data sets. In Section 4.2, we evaluate the performance of the three asymptotic tests by type I error simulations. The exact NB test does not extend to general regression settings. In Section 4.3, we present power simulation

results for two-group comparisons. In Section 4.4, we investigate the consequences of treating the estimated dispersion parameter as known. Finally, in Section 4.5, we apply the NB regression analysis to a bacterial data set to identify genes where the treatment and the genotype show strong interaction. We provide some details on the practical application of the NB regression model.

4.1 Comparison to the exact NB test

4.1.1 Overview

For two-group comparison, we can evaluate the accuracy of the asymptotic tests by comparing their p-values to the p-values from the exact NB test (Robinson and Smyth, 2007).

The regression model (2) includes the two-group comparison model as a special case. To test differential gene expression in two groups, two covariates X_1, X_2 are needed (p=2). One can define the intercept term X_{j1}=1 for all samples j=1,..., n, and define

$$X_{j2}=\begin{cases}1 & \text{if sample } j \text{ is from the treatment group}\\ 0 & \text{if it is sample } j \text{ is from the control group,}\end{cases}$$

then β_1 will represent the baseline log expression level of the gene as measured by mean relative frequency and β_2 will measure the effect of the treatment on the log expression level. For each gene, testing differential gene expression amounts to testing β_2=0.

For two-group comparison examples in this section, a continuity correction was made to the asymptotic tests by adding ±0.5 to the total counts in the two groups [see Pierce and Peters (1992) for details]. As pointed out by Lancaster (1961) among many others, the asymptotic p-values are meaningful even without continuity correction – they approximate the *mid-p-values* (i.e., when computing the p-value, only half the probability of the observed point is counted in the tail area calculation). Pierce and Peters (1999) also linked uncorrected HOA p-values to approximate conditional inference.

Even though an exact NB test is available for testing differential expression in two groups, there is good reason to prefer using a regression model with asymptotic inference instead of the exact test: the exact test is theoretically justified only for equal library sizes and requires awkward library size adjustments when library sizes are unequal (Robinson and Smyth, 2008; Anders and Huber, 2010; Di et al., 2011). The regression model (2) overcomes this limitation and is justified for both equal and unequal library sizes. We will demonstrate that the HOA test produces p-values essentially identical to those from the exact test.

4.1.2 Arabidopsis data

In Di et al. (2011), we examined a pilot Arabidopsis study consisting of three biological replicate samples in each of two treatment groups. In that paper, the exact NB test was used to identify genes that are differentially expressed in the two groups. In order to apply the exact NB test, we thinned the counts so that all library sizes were approximately the same after thinning. We estimated the dispersion parameters by modeling the log dispersion as a parametric function of the log mean. The estimated dispersion model is ϕ=1.5 $\mu^{-0.5}$ and the estimated dispersion ranges from about ϕ=0.05 at mean level μ=1000 to ϕ=0.5 at mean level μ=10. Details on thinning and dispersion estimation were presented in Di et al. (2011).

The goal here is to assess the accuracy of the three large-sample tests. We performed two-sided HOA, LR, and Wald tests on the same thinned counts as used for the exact NB test, using the same estimates of the dispersion parameters. Figure 1 shows the relative differences between p-values from the three large-sample tests and p-values from the exact NB test. The relative difference between a large-sample p-value p_1 and an exact test p-value p_0 is defined as

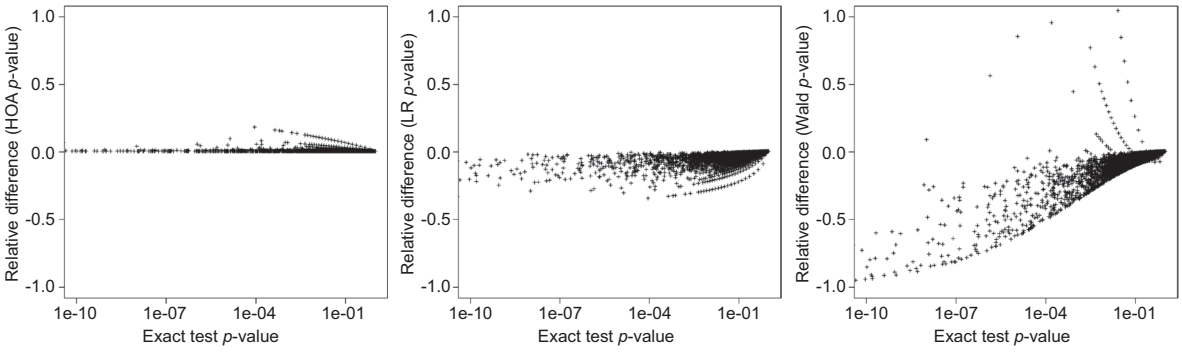

Figure 1 Relative differences between the *p*-values from the (a) HOA (left), (b) LR (middle) and (c) Wald (right) tests and the *p*-values from the exact NB test for differential gene expression on roughly 25,000 genes from a two-group data set with sample size 3 in each group.

$$\frac{p_1 - p_0}{p_0}.$$

Note that there is a *p*-value for each of the approximately 25,000 genes to indicate evidence of differential expression. Overall, the *p*-values from the HOA test and the exact test were very similar. The relative differences were smaller than 20% for all genes. The relative differences were smaller than 10% for 99.9% of the genes and smaller than 2% for 92% of the genes. All but one of the 30 cases for which the relative error was >10% were characterized by frequencies of 0 in one of the groups. The LR test and Wald test provided poorer approximations to the exact test *p*-value. The inaccuracy was present for *p*-values of all levels. The Wald test was the least accurate. For comparison purpose, the limits on the *y*-axes were set to ±1, but in 66 cases, the relative differences between the Wald test *p*-values and exact test *p*-values were out of this range.

4.1.3 Computer generated data sets

To further clarify factors that affect the accuracy of the asymptotic tests, we compared the performance of exact NB, HOA, LR and Wald tests on computer generated data sets. We generated a series of two-group data sets representing different mean levels and differences in means. Each data set – mimicking read counts from a single gene – contained 6 counts in two groups of sizes 2 and 4. For each data set, we used one-sided exact NB, HOA, LR and Wald tests to test whether the two group means are the same. We assumed that the dispersion parameter ϕ is 1 for all read counts. We also assumed that the counts are from samples of equal library sizes [i.e., N_j's in model (2) are the same for all read counts] so that the exact NB test can be used without the need for library size adjustment.

In Tables 1 and 2, we compare the *p*-values from the three asymptotic tests and from the exact NB tests for two possible one-sided alternative hypotheses. Figures 2 and 3 give graphical presentations of the same results. When the group sizes are not balanced, it is typical that the asymptotic test *p*-values are more accurate in one tail of the test statistic distribution than in the other. Using one-sided tests and unbalanced group sizes enabled us to investigate the different behaviors of the asymptotic *p*-values in the two tails. Tables 1 and 2 list only the total counts in the two groups, which we call S_1 and S_2, since they are the sufficient statistics for comparing two group means when the library sizes are the same. In fact, the exact NB test is based on the conditional distribution of $S_1|S_1+S_2$.

In this set of examples, the HOA test *p*-values show consistent accuracy over a wide range of mean levels and mean differences, even though the total sample size is only 6. Only in extreme cases (meaning one of the group totals is close to 0), are the relative differences between the HOA *p*-values and the exact *p*-values over 5%. An appealing feature of the HOA test theory is that it is the *relative errors* of the *p*-values that are bounded.

Table 1 Comparisons between the *p*-values from HOA, LR, and Wald tests and the *p*-values from the exact NB test for comparing means in simulated two-group data sets. Each data set contains six counts, divided into two groups: one of size 2 and one of size 4. For each data set, the first two columns list the total counts in the two groups, the third column list the *p*-value of the exact NB test, and the last three columns list the relative differences between the *p*-values of HOA, LR, Wald tests and the *p*-value of the exact NB test. The one-sided alternative hypothesis is that the group with size 2 (group 1) has a smaller mean.

			Relative differences		
S_1	S_2	Exact *p*-value	HOA	LR	Wald
1	9	2.42e-01	0.37%	−17.93%	−17.18%
2	8	4.07e-01	−0.15%	−11.30%	−11.41%
3	7	5.66e-01	−0.28%	−6.94%	−6.95%
1	99	5.39e-03	2.46%	−44.10%	−67.82%
10	90	9.89e-02	0.17%	−24.69%	−40.75%
20	80	2.81e-01	−0.09%	−14.73%	−17.64%
30	70	4.84e-01	−0.14%	−8.57%	−8.65%
1	999	5.93e-05	3.81%	−58.40%	−99.10%
10	990	1.28e-03	1.59%	−49.04%	−98.90%
100	900	8.32e-02	0.18%	−26.07%	−48.42%
200	800	2.65e-01	−0.08%	−15.34%	−19.21%
300	700	4.73e-01	−0.14%	−8.83%	−8.95%

Table 2 Comparisons between the *p*-values from HOA, LR, and Wald tests and the *p*-values from the exact NB test for comparing means in simulated two-group data sets. Each data set contains six NB counts, divided into two groups: one of size 2 and one of size 4. For each data set, the first two columns list the total counts in the two groups, the third column lists the *p*-value of the exact NB test, and the last three columns list the relative differences between the *p*-values of HOA, LR, Wald tests and the *p*-value of the exact NB test. The one-sided alternative hypothesis is that the group with size 2 (group 1) has a greater mean.

			Relative differences		
S_1	S_2	Exact *p*-value	HOA	LR	Wald
7	3	1.00e-01	1.88%	7.73%	14.20%
8	2	4.70e-02	3.06%	4.87%	21.11%
9	1	1.70e-02	5.60%	−0.17%	47.21%
70	30	3.70e-02	0.89%	5.22%	20.82%
80	20	9.35e-03	1.26%	0.13%	22.70%
90	10	9.64e-04	1.97%	−8.71%	13.06%
99	1	5.19e-06	7.86%	−26.93%	327.48%
700	300	3.14e-02	0.89%	4.68%	22.12%
800	200	6.97e-03	1.24%	−0.97%	21.73%
900	100	5.01e-04	1.79%	−10.98%	−5.76%
990	10	1.18e-07	3.39%	−34.07%	−94.75%
999	1	5.91e-10	8.88%	−43.91%	28.75%

The LR test and Wald test are less accurate. Furthermore, the Wald test *p*-values sometimes display non-monotonicity as the test statistic values tend towards the extremes (see Figure 4). It can be confusing in practice when more extreme differences in count means yield less extreme *p*-values. Similar behavior of the Wald test has been observed and characterized in other families of statistical distributions (Væth, 1985; Fears et al., 1996).

4.2 Type I error simulations

As additional evidence of the exceptional performance of the HOA adjustments, we present results from type I error simulations under a range of NB regression models.

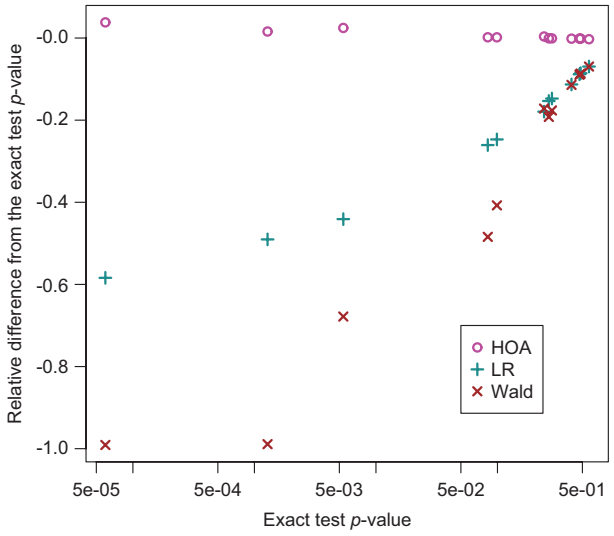

Figure 2 A graphical presentation of the results in Table 1: relative differences between the *p*-values of the asymptotic tests and the *p*-values from the exact NB test. See the caption of Table 1 for further description of the data.

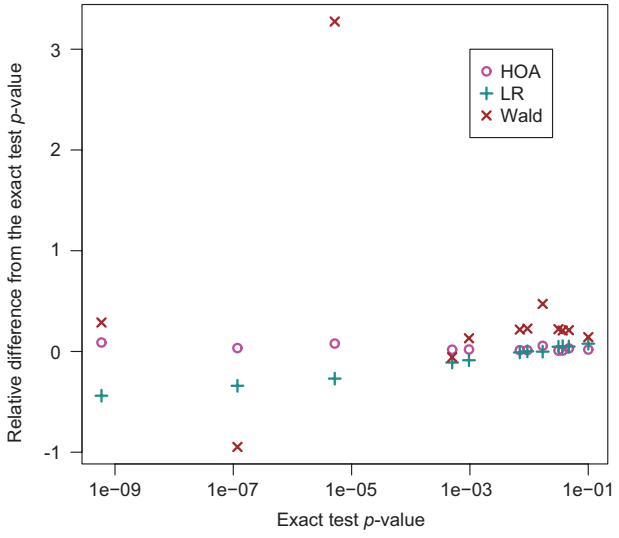

Figure 3 A graphical presentation of the results in Table 2: relative differences between the *p*-values of the asymptotic tests and the *p*-values from the exact NB test. See the caption of Table 2 for further description of the data.

4.2.1 One-dimensional tests

Table 3 shows Monte Carlo type I error rates for the tests from 10,000 simulated NB two-group comparisons, indicating remarkable agreement of the HOA test with the exact test and the superiority of these two – in producing accurate type I error rates – over the Wald and the unadjusted LR tests.

For this set of simulations, we used mid-*p*-values from the exact NB test and did not apply continuity correction to the asymptotic tests. Due to discreteness, using ordinary *p*-values from the exact test will give slightly conservative results. As discussed in Section 4.1.1, the asymptotic *p*-values without continuity correction are more comparable with the mid- *p*-values.

Table 4 shows another set of Monte Carlo type I error rates for NB two-group comparisons, with higher mean levels. Comparing this set to Table 3, we see that, unlike the Poisson case, the test accuracy did not improve with increased mean value (for fixed dispersion values).

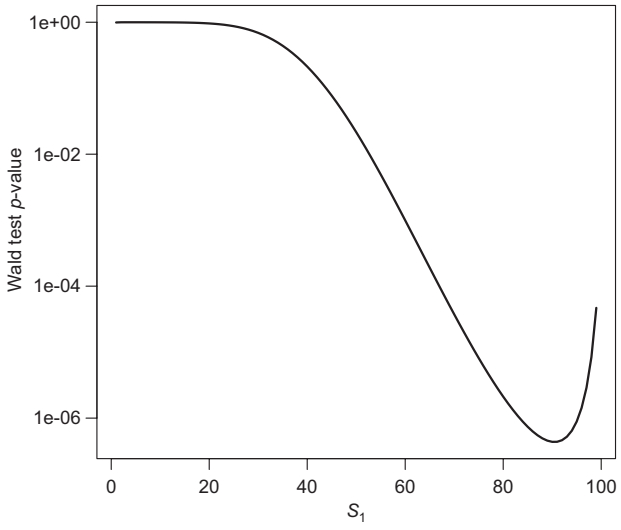

Figure 4 Erratic behavior of the Wald test. The plot shows the *p*-values of one-sided Wald test for comparing two group means when applied to a series of two-group data sets. In each data set, the group sizes are 2 and 4, the total count in both groups is 100, the total count in group 1 (with group size 2) ranges from 1 to 99 (shown on the *x*-axis). The alternative hypothesis being tested is that group 1 has a greater mean. We assume that the read counts are from samples of equal library sizes and the dispersion parameter is 0.1 for all counts. As S_1 approaches 100, the Wald test *p*-values display non-monotonicity.

Table 3 Monte Carlo type I error rates of one-sided tests for two-group comparisons at nominal levels (alpha) 1%, 5%, 10%, and 20%. Results are based on 10,000 simulated two-group data sets. Each data set contains six NB counts, divided into two groups: one of size 2 and one of size 4. The simulations were performed under the null hypothesis: both groups were simulated to have the same means (20 for all counts). The dispersion parameters were simulated to be 1.0, 0.3 or 0.1 for all counts. The alternative hypothesis is that the group with size 2 has a smaller [tables (a), (c) and (e)] or greater [tables (b), (d) and (f)] mean.

	Estimated type I error rates					Estimated type I error rates			
Alpha	**Exact**	**HOA**	**LR**	**Wald**	**Alpha**	**Exact**	**HOA**	**LR**	**Wald**
(a) ϕ=1.0					(b) ϕ=1.0				
0.01	0.010	0.010	0.016	0.027	**0.01**	0.010	0.010	0.010	0.009
0.05	0.050	0.050	0.071	0.088	**0.05**	0.049	0.049	0.047	0.041
0.10	0.099	0.099	0.126	0.144	**0.10**	0.102	0.101	0.094	0.086
0.20	0.196	0.196	0.233	0.245	**0.20**	0.208	0.207	0.190	0.182
(c) ϕ=0.3					(d) ϕ=0.3				
0.01	0.011	0.011	0.013	0.015	**0.01**	0.010	0.010	0.010	0.008
0.05	0.046	0.046	0.055	0.062	**0.05**	0.049	0.049	0.046	0.042
0.10	0.095	0.095	0.109	0.115	**0.10**	0.100	0.100	0.093	0.089
0.20	0.196	0.196	0.214	0.220	**0.20**	0.205	0.205	0.192	0.190
(e) ϕ=0.1					(f) ϕ=0.1				
0.01	0.010	0.011	0.011	0.012	**0.01**	0.010	0.010	0.009	0.008
0.05	0.048	0.049	0.052	0.053	**0.05**	0.046	0.046	0.045	0.043
0.10	0.096	0.096	0.103	0.104	**0.10**	0.096	0.096	0.091	0.089
0.20	0.190	0.190	0.202	0.204	**0.20**	0.199	0.199	0.194	0.192

Table 5 shows similar simulation results for a regression setting. Table 6 shows simulation results for testing the interaction term in an experiment with a 2×2 treatment structure and with three replicates per treatment group. For these settings, exact tests are not available. These results show similar conclusions – the HOA test produces type I error rates consistently and, in some cases, substantially closer to the nominal error rates than the Wald and unadjusted LR tests.

Table 4 Monte Carlo type I error rates of one-sided tests for two-group comparisons at nominal levels (alpha) 1%, 5%, 10%, and 20%. Results are based on 10,000 simulated two-group data sets. Each data set contains six NB counts, divided into two groups: one of size 2 and one of size 4. The simulations were performed under the null hypothesis: both groups were simulated to have the same means (1000 for all counts). The dispersion parameters were simulated to be 1.0, 0.3 or 0.1 for all counts. The alternative hypothesis is that the group with size 2 has a smaller [tables (a), (c) and (e)] or greater [tables (b), (d) and (f)] mean.

	Estimated type I error rates					Estimated type I error rates			
Alpha	**Exact**	**HOA**	**LR**	**Wald**	**Alpha**	**Exact**	**HOA**	**LR**	**Wald**
(a) ϕ=1.0					(b) ϕ=1.0				
0.01	0.011	0.011	0.018	0.035	**0.01**	0.011	0.011	0.011	0.009
0.05	0.051	0.051	0.070	0.095	**0.05**	0.050	0.050	0.047	0.041
0.10	0.103	0.103	0.132	0.151	**0.10**	0.099	0.099	0.093	0.086
0.20	0.200	0.200	0.240	0.252	**0.20**	0.193	0.192	0.177	0.171
(c) ϕ=0.3					(d) ϕ=0.3				
0.01	0.009	0.009	0.012	0.019	**0.01**	0.009	0.009	0.008	0.006
0.05	0.050	0.050	0.057	0.067	**0.05**	0.049	0.049	0.046	0.040
0.10	0.098	0.098	0.110	0.120	**0.10**	0.102	0.102	0.095	0.090
0.20	0.198	0.198	0.215	0.221	**0.20**	0.201	0.201	0.191	0.186
(e) ϕ=0.1					(f) ϕ=0.1				
0.01	0.012	0.012	0.013	0.016	**0.01**	0.008	0.008	0.008	0.007
0.05	0.053	0.053	0.057	0.061	**0.05**	0.052	0.052	0.049	0.046
0.10	0.105	0.105	0.111	0.116	**0.10**	0.104	0.104	0.099	0.094
0.20	0.205	0.205	0.215	0.218	**0.20**	0.205	0.205	0.199	0.196

Table 5 Monte Carlo type I error rates of one-sided tests for the coefficient of X in a NB log-linear regression model, $\log(\mu)=\log(N)+\beta_0+\beta_1 X$, at nominal levels (alpha) 1%, 5%, 10%, and 20%. Results are based on 10,000 simulated NB samples of size 6, with the predictor X=1, 2, 4, 8, 16, and 32. These simulations were performed under the true null hypothesis where β_1=0, β_0=−11.5 and N=10^6, so the mean frequencies were μ=10 for all counts. The dispersion parameter was simulated to be 1.0, 0.3 or 0.1. The alternative hypotheses are β_1<0 (left side of table) and β_1 > 0 (right side of table).

	Estimated type I error rates				Estimated type I error rates		
Alpha	**HOA**	**LR**	**Wald**	**Alpha**	**HOA**	**LR**	**Wald**
(a) ϕ=1.0				(b) ϕ=1.0			
0.01	0.010	0.016	0.010	**0.01**	0.008	0.006	0.007
0.05	0.049	0.070	0.091	**0.05**	0.049	0.036	0.038
0.10	0.100	0.137	0.164	**0.10**	0.098	0.075	0.076
0.20	0.201	0.252	0.272	**0.20**	0.202	0.163	0.162
(c) ϕ=0.3				(d) ϕ=0.3			
0.01	0.011	0.015	0.013	**0.01**	0.010	0.009	0.007
0.05	0.051	0.061	0.067	**0.05**	0.052	0.045	0.039
0.10	0.101	0.118	0.124	**0.10**	0.104	0.091	0.087
0.20	0.204	0.229	0.235	**0.20**	0.207	0.182	0.181
(e) ϕ=0.1				(f) ϕ=0.1			
0.01	0.009	0.011	0.009	**0.01**	0.009	0.008	0.007
0.05	0.048	0.053	0.049	**0.05**	0.048	0.044	0.042
0.10	0.101	0.112	0.108	**0.10**	0.100	0.091	0.091
0.20	0.205	0.220	0.220	**0.20**	0.199	0.186	0.187

In these simulations, we set the dispersion size to be 1.0, 0.3 or 0.1. In general, as the dispersion decreases, the performance of LR and Wald tests relative to HOA will improve, but the accuracy of these tests are also affected by other factors, such as the design matrix. The HOA test shows consistent accuracy across all simulations. In our Arabidopsis data, the dispersion size ranges from about 0.05 when μ=1000 to 0.5 when μ=10.

Table 6 Monte Carlo Type I Error Rates of one-sided tests for positive (left side of table) or negative (right side of table) interaction in a 2×2 design at nominal levels (alpha) 1%, 5%, 10%, and 20%. The corresponding regression model is $\log(\mu)=\log(N)+\beta_0+\beta_1X_1+\beta_2X_2+\beta_3X_1X_2$, where X_1 and X_2 are indicator variables for two two-level factors. Results are based on 10,000 simulated NB samples of size 12 (three replicates per treatment group). These simulations were performed under the true null hypothesis, so there was no interaction effect in the simulations ($\beta_3=0$). Other parameters were specified as $N=10^6$, $\beta_0=-11.5$, $\beta_1=0.41$, $\beta_2=0.69$, and the mean frequencies range from 10 to 120. The dispersion parameter was simulated to be 1.0, 0.3 or 0.1.

	Estimated type I error rates				Estimated type I error rates		
Alpha	**HOA**	**LR**	**Wald**	**Alpha**	**HOA**	**LR**	**Wald**
(a) ϕ=1.0				(b) ϕ=1.0			
0.01	0.008	0.013	0.014	**0.01**	0.009	0.014	0.016
0.05	0.045	0.058	0.060	**0.05**	0.049	0.061	0.064
0.10	0.092	0.110	0.112	**0.10**	0.101	0.114	0.117
0.20	0.196	0.215	0.215	**0.20**	0.204	0.219	0.221
(c) ϕ=0.3				(d) ϕ=0.3			
0.01	0.010	0.012	0.012	**0.01**	0.012	0.014	0.015
0.05	0.048	0.052	0.052	**0.05**	0.052	0.054	0.055
0.10	0.098	0.105	0.105	**0.10**	0.106	0.109	0.110
0.20	0.195	0.202	0.202	**0.20**	0.209	0.210	0.211
(e) ϕ=0.1				(f) ϕ=0.1			
0.01	0.011	0.012	0.011	**0.01**	0.010	0.010	0.011
0.05	0.051	0.054	0.053	**0.05**	0.049	0.048	0.049
0.10	0.100	0.105	0.104	**0.10**	0.096	0.095	0.096
0.20	0.199	0.205	0.204	**0.20**	0.194	0.191	0.192

From our experience, this range is typical of real RNA-Seq data sets. We have kept simulations with dispersion parameter of 1.0, even though it may not be practically realistic, to emphasize the relationship of test accuracy and dispersion parameter size. It is also common that the dispersion tends to decrease with the mean level (Anders and Huber, 2010; Di et al., 2011), so in practice, the accuracy of LR and Wald test will improve with increased mean levels.

4.2.2 Multi-dimensional tests

The main result of this paper is about the potentially useful improvements from HOA for tests about a single parameter. We have also implemented the multi-parameter HOA test. Although more simulations in a variety of settings are needed to better understand the behavior for multi-parameter tests, the following results provide an initial investigation.

For testing multi-dimensional parameters ($q > 1$), the forms of the unadjusted LR and Wald test remain the same. For the HOA test, Skovgaard's approximation (6) can be used. Skovgaard (2001) cautioned that producing accurate test results for multi-dimensional parameters can be challenging. In the examples investigated in Skovgaard (2001), asymptotic tests – with or without HOA adjustments – are less accurate in multi-dimensional examples than in the one-dimensional examples.

In the context of the NB regression model (2), factors influencing the accuracy of the asymptotic tests include the amount of dispersion, sample size, and the structure of the design matrix. Table 7 shows simulation results for testing the interaction terms in an experiment with a 3×4 treatment structure (i.e., two factors with 3 and 4 levels for each factor) and with two replicates per treatment group. In this example, there are n=24 observations, p=12 parameters and the dimension of the test is q=6. The dispersion parameter is 0.5, 0.3 or 0.1. The HOA test provides noticeable improvement in this quite challenging example.

Table 8 shows type I error simulation results for testing the interaction terms in an experiment with a 4×5 design and with two replicates per treatment. In this example, there are n=40 observations, p=20 parameters

Table 7 Monte Carlo type I error rates of one-sided tests for the 6-dimensional parameter for interaction in a 3×4 design at nominal levels (alpha) 1%, 5%, 10%, and 20%. The corresponding regression model is $\log(\mu)=\log(N)+\beta_0+\beta_1X_1+\ldots+\beta_5X_5+\beta_6X_1X_3+\ldots+\beta_{11}X_2X_5$, where indicator variables (X_1, X_2) and (X_3, X_4, X_5) encode levels of two factors. Results are based on 10,000 simulated NB samples of size 24 (two replicates from each of the 12 treatment groups). These simulations were performed under the true null hypothesis, so there was no interaction effect in the simulations ($\beta_6=\cdots=\beta_{11}=0$). Other parameters were specified as $N=10^6$, $\beta_0=-11.5$, (β_1, β_2)=(−1.0, 0.5), (β_3, β_4, β_5)=(0.1, 0.55, 1.0), and the mean frequencies range from 3 to 44. The dispersion parameter was simulated to be 0.5, 0.3 or 0.1 for all counts.

	Estimated type I error rates		
Alpha	**HOA**	**LR**	**Wald**
(a) ϕ=0.5			
0.01	0.009	0.019	0.013
0.05	0.044	0.070	0.058
0.10	0.091	0.133	0.114
0.20	0.189	0.247	0.224
(b) ϕ=0.3			
0.01	0.007	0.012	0.008
0.05	0.046	0.062	0.049
0.10	0.095	0.119	0.103
0.20	0.188	0.225	0.202
(c) ϕ=0.1			
0.01	0.008	0.009	0.006
0.05	0.048	0.055	0.042
0.10	0.101	0.113	0.093
0.20	0.198	0.215	0.192

Table 8 Monte Carlo type I error rates of one-sided tests for the 12-dimensional parameter for interaction in a 4×5 design at nominal levels 1%, 5%, 10%, and 20%. The corresponding regression model is $\log(\mu)=\log(N)+\beta_0+\beta_1X_1+\ldots+\beta_7X_7+\beta_8X_1X_4+\ldots+\beta_{19}X_3X_7$, where indicator variables (X_1, …, X_3) and (X_4, …, X_7) encode levels of two factors. Results are based on 10,000 simulated NB samples of size 40 (two replicates from each of the 20 treatment groups). These simulations were performed under the true null hypothesis, so there was no interaction effect in the simulations ($\beta_8=\cdots=\beta_{19}=0$). Other parameters were specified as $N=10^6$, $\beta_0=-11.5$, (β_1, β_2, β_3)=(−1.0, −0.25, 0.5), (β_4, β_5, β_6, β_7)=(0.1, 0.4, 0.7, 1.0), and the mean frequencies range from 3 to 44. The dispersion parameter was simulated to be 0.5, 0.3 or 0.1 for all counts.

	Estimated type I error rates		
Alpha	**HOA**	**LR**	**Wald**
(a) ϕ=0.5			
0.01	0.006	0.015	0.009
0.05	0.035	0.073	0.055
0.10	0.077	0.133	0.115
0.20	0.163	0.247	0.223
(b) ϕ=0.3			
0.01	0.007	0.014	0.009
0.05	0.042	0.064	0.049
0.10	0.086	0.119	0.102
0.20	0.181	0.232	0.209
(c) ϕ=0.1			
0.01	0.009	0.013	0.007
0.05	0.049	0.058	0.041
0.10	0.097	0.111	0.087
0.20	0.196	0.217	0.186

and the dimension of the test is q=12. The dispersion parameter is 0.5, 0.3 or 0.1. The large size of q relative to n makes this an extreme challenge for the asymptotic tests. The accuracy will improve as the dispersion decreases or as the number of replicates increases.

Table 9 Monte Carlo power (calibrated) for two-group comparisons for tests with level alpha (0.01, 0.05, 0.10, 0.20). Each data set contains six NB counts, divided into two groups, one of size 2 and one of size 4. The simulations were performed under an alternative hypothesis. Group 1, with 4 observations, was simulated to have mean 10, and group 2, with 2 observations, was simulated to have mean 20. The dispersion parameters were simulated to be 1.0 or 0.1 for all counts and were treated as known in performing the tests. One-sided (left side of table) and two-sided (right side of table) tests were performed, and the proportion of 100,000 simulated data sets for which the p-value was less than a critical value $c(\alpha)$ is reported. $c(\alpha)$ was estimated from a separate Monte Carlo simulation under the null hypothesis.

	Estimated power (one-sided)				Estimated power (two-sided)		
Alpha	**HOA**	**LR**	**Wald**	**Alpha**	**HOA**	**LR**	**Wald**
(a) ϕ=1.0				(b) ϕ=1.0			
0.01	0.059	0.059	0.059	**0.01**	0.037	0.194	0.194
0.05	0.194	0.194	0.194	**0.05**	0.127	0.109	0.079
0.10	0.310	0.310	0.311	**0.10**	0.208	0.185	0.154
0.20	0.468	0.468	0.469	**0.20**	0.340	0.312	0.289
(c) ϕ=0.1				(d) ϕ=0.1			
0.01	0.364	0.364	0.366	**0.01**	0.286	0.266	0.262
0.05	0.630	0.630	0.631	**0.05**	0.509	0.496	0.490
0.10	0.746	0.746	0.748	**0.10**	0.629	0.618	0.613
0.20	0.860	0.860	0.860	**0.20**	0.748	0.738	0.738

Table 10 Monte Carlo power (calibrated) for two-group comparisons for tests with level alpha (0.01, 0.05, 0.10, 0.20). Each data set contains six NB counts, divided into two groups, one of size 2 and one of size 4. The simulations were performed under an alternative hypothesis. Group 1, with 2 observations, was simulated to have mean 10, and group 2, with four observations, was simulated to have mean 20. The dispersion parameters were simulated to be 1.0 or 0.1 for all counts and were treated as known in performing the tests. One-sided (left side of table) and two-sided (right side of table) tests were performed, and the proportion of 100,000 simulated data sets for which the p-value was less than a critical value $c(\alpha)$ is reported. $c(\alpha)$ was estimated from a separate Monte Carlo simulation under the null hypothesis.

	Estimated power (one-sided)				Estimated power (two-sided)		
Alpha	**HOA**	**LR**	**Wald**	**Alpha**	**HOA**	**LR**	**Wald**
(a) ϕ=1.0				(b) ϕ=1.0			
0.01	0.035	0.035	0.033	**0.01**	0.018	0.022	0.027
0.05	0.146	0.146	0.144	**0.05**	0.084	0.100	0.116
0.10	0.259	0.259	0.258	**0.10**	0.156	0.177	0.198
0.20	0.439	0.439	0.438	**0.20**	0.280	0.310	0.328
(c) ϕ=0.1				(d) ϕ=0.1			
0.01	0.307	0.307	0.299	**0.01**	0.217	0.226	0.230
0.05	0.592	0.594	0.589	**0.05**	0.452	0.465	0.471
0.10	0.732	0.732	0.731	**0.10**	0.590	0.603	0.606
0.20	0.861	0.861	0.861	**0.20**	0.731	0.742	0.744

4.3 Power simulations

The optimal power properties of the likelihood ratio test are not preserved with the HOA adjustment. The following tables compare the power of the HOA adjusted likelihood ratio test to that of the ordinary LR and Wald tests in a few situations to get a glimpse into power issues. Power comparisons were done using Monte Carlo simulation, generating a large number of data sets under various alternatives and comparing the proportion of data sets for which each test would reject at a given level α. Two different alternative scenarios are presented, as the relative performance of the tests depends somewhat upon whether the larger or smaller group has the larger mean. This asymmetric behavior was also observed in the type I error simulations. Since some of these tests are not perfectly calibrated at the sample sizes considered (see Subsection 4.2), we report

calibrated power. For calibrated power, the critical value $c(\alpha)$ is chosen so that when the null hypothesis is true (the means in the two groups are equal) exactly α of the resulting p-values are less than $c(\alpha)$; that is, $c(\alpha)$ is the α quantile of the null distribution of p-values, where the null distribution is also obtained through Monte Carlo simulation. This allows a more fair comparison between tests, as tests that have very small p-values even under the null hypothesis will typically appear to have better power due to the tendency to produce small p-values, but this apparent power is not necessarily truly distinguishing between the null and alternative. In these simulations, the dispersion ϕ was simulated to be 1.0 or 0.1 and treated as known when performing the tests.

In two-sided comparisons, if the group with the larger mean has the larger sample size, the LR test has better calibrated power than the HOA test, and the Wald test has slightly better calibrated power than the LR test. If the group with the larger mean has the smaller sample size, this trend is reversed: the HOA test has better power than LR, and the LR test has better power than the Wald test. There is virtually no difference in the calibrated power of any of the tests – HOA, LR, and Wald – when testing one-sided alternatives. There is also virtually no difference in the calibrated power if the two groups have the same sample sizes (results not shown). These results indicate that when the test statistic distribution is skewed the extreme values in either tail are ranked similarly by these large sample tests, but the extreme values from the two tails, when combined, are not ranked the same way by these tests. No test is dominant in all situations.

4.4 Consequences of treating dispersion estimates as known

In the type I error and power simulations in Sections 4.2 and 4.3, we used the actual (data generating) value of the dispersion ϕ in place of the estimate for evaluating the tests. In practice, we would estimate a model for ϕ from all genes combined, then estimate ϕ for each individual gene from this fit and from the estimated μ for the particular gene, and would treat the estimated ϕ's as known in tests for regression coefficients. A more practically realistic simulation strategy would have been to simulate counts from a complete set of genes (such as 25,000 genes) in each Monte Carlo step, but that strategy would have been very time consuming.

To understand the consequences of treating estimated ϕ's as known, we simulated one complete set of genes under a two-group comparison setting and estimated the dispersion model from all genes combined (details given below). In this, we found that the type I error rates of the three large sample tests did not depend very much on whether true or estimated ϕ's were used. The improvement by HOA adjustment was still evident.

Table 11 shows Monte Carlo type I error rates of tests for two-group comparisons at nominal levels (alpha) 1%, 5%, 10%, and 20%. Results are based on a simulated two-group data set containing NB counts for 25,000 genes. The two groups are of sizes 2 and 4. The data were simulated under the null hypothesis: for each gene i, read counts from all six samples were simulated to have the same mean, which was drawn from a log normal distribution ($\log_{10}\mu_i \sim N(2,0.5)$), and the same dispersion, which was determined from the equation $\phi_i = 1.5\mu_i^{-0.5}$. This dispersion model mimicked the parametric dispersion model estimated from the Arabidopsis data (see Section 4.1.2).

When performing the test, we first estimate the dispersion model from all 25,000 genes combined using method described in Di et al. (2011). The estimated dispersion model is $\phi_i = 1.279\mu_i^{-0.468}$. For each gene i, we then plugged in $\hat{\mu}_i$ (simply the row means in this setting) into this estimated mean-dispersion function to get the fitted value $\hat{\phi}_i$ of ϕ_i. Treating this estimated value $\hat{\phi}_i$ as known, we performed HOA, LR and Wald tests to test three possible alternative hypotheses: (a) the group with size 2 has a smaller mean; (b) the group with size 2 has a greater mean; and (c) the two groups have different means. Table 11 summarizes the type I error rates of the tests over all 25,000 simulated genes. The type I error rates show the same trends as in the simulations where we treated ϕ as known. In particular, the HOA test gives more accurate type I error rates.

Table 11 Monte Carlo type I error rates of tests for two-group comparisons at nominal levels (alpha) 1%, 5%, 10%, and 20%. Results are based on one simulated two-group data set containing 25,000 genes. The data set consisted of six samples, divided into two groups: one of size 2 and one of size 4. The data were simulated under the null hypothesis: both groups were simulated to have the same means, the mean values were drawn from a log normal distribution ($\log_{10}\mu_i \sim N(2,0.5)$, the true dispersion is determined by $\phi_i = 1.5\,\mu_i^{-0.5}$, and all counts are independent. We performed HOA, LR and Wald tests to test three possible alternative hypotheses as described in table captions. In the tests, dispersion is estimated from a fitted model $\hat{\phi}_i = 1.279\,\hat{\mu}_i^{-0.468}$ and treated as known.

	Estimated type-I error rates		
Alpha	**HOA**	**LR**	**Wald**
(a) The alternative hypothesis is that the group with size 2 has a smaller mean			
0.01	0.010	0.012	0.015
0.05	0.050	0.056	0.062
0.10	0.101	0.109	0.114
0.20	0.199	0.213	0.217
(b) The alternative hypothesis is that the group with size 2 has a greater mean			
0.01	0.010	0.009	0.008
0.05	0.050	0.048	0.044
0.10	0.099	0.095	0.091
0.20	0.197	0.189	0.187
(c) The alternative hypothesis is that the two groups have different mean			
0.01	0.009	0.010	0.011
0.05	0.052	0.054	0.055
0.10	0.100	0.104	0.106
0.20	0.200	0.204	0.205

4.5 Bacterial data

We used RNA-Seq to study the transcriptome changes of the plant pathogenic bacterium *Pseudomonas syringae* pv *tomato* DC3000. The experiment had a 2×2 structure: two genotypes of the pathogen, wild-type and a Δ*hrpL* mutant, were individually grown for 7 hours in either King's B (KB) rich medium or *hrp*-inducing medium (Huynh et al., 1989). Three biological replicates were used in each of the four groups, for a total sample size of 12. HrpL is an alternative sigma factor important for the virulence of this pathogen (Lindeberg et al., 2006). KB represses, whereas *hrp*-inducing medium induces, *hrpL* expression. The goal was to identify genes that were differentially regulated in an *hrp*-inducing medium-dependent and *hrpL*-independent manner.

Total RNA was extracted from *in vitro* grown bacteria using Trizol (Invitrogen, Carlsbad, CA, USA). Samples were depleted of rRNA using MicrobExpress (Ambion, Austin, TX, USA) and Ribominus (Invitrogen, Carlsbad, CA, USA). The enriched mRNA were prepared according to instructions and sequenced on an Illumina HiSeq (Illumina, San Diego, CA, USA). Reads were mapped to the reference sequence (NC_004578.1, Buell et al. 2003) using CASHX (Cumbie et al., 2011).

We use this example to outline the steps needed for practical application of the NB regression model (see Section 2). We focus on one biological question of interest: identify genes for which the effect of media treatment differs in the two genotypes. The statistical task is to test the interaction term in a 2×2 design, which requires the use of the NB regression model. For the design matrix, we will let the intercept term $X_{j1}=1$ for all samples $j=1,\ldots,J$, $X_{j2}=1$ for one of the media types, $X_{j3}=1$ for one of the genotypes, and $X_{j4}=X_{j2}\times X_{j3}$. We test the media-genotype interaction by testing the regression coefficient corresponding to X_{j4}. There is no exact NB test available for testing the interaction term.

The regression model can account for differences in observed library sizes, so there is no need to adjust read counts to make the library sizes equal. For estimating the normalization factors R_j, we used the method described in Anders and Huber (2010), which assumes the median log fold change in expression levels

between any two samples is 0. For this data set, this method gives similar results as the "trimmed mean of M-values" method of Robinson and Oshlack (2010).

For estimating the dispersion parameter, we assume a parametric dispersion model

$$\log(\phi_{ij})=\alpha_0+\alpha_1\log(\pi_{ij}), \tag{7}$$

where the $\pi_{ij}=\frac{\mu_{ij}}{N_j R_j}$ is the mean relative frequency after normalization. To estimate the parameters $\alpha=(\alpha_0, \alpha_1)$ in the dispersion model, we maximize the *adjusted profile likelihood (APL)*:

$$l_p(\alpha)=l\left(\alpha,\hat{\beta}_\alpha\right)-\frac{1}{2}\log\det\left\{j_{\beta\beta}\left(\hat{\beta}_\alpha;\alpha\right)\right\}, \tag{8}$$

where $l(\alpha, \beta)$ is the full model likelihood, $\hat{\beta}_\alpha$ maximizes the constrained likelihood $l(\alpha, \beta)$ for fixed α and $j_{\beta\beta}$ the observed information matrix for estimating $\hat{\beta}_\alpha$. The expression of the likelihood function and other details are provided in Appendix 6.3. Adjusted profile likelihood was introduced by Cox and Reid (1987). It can be viewed as an approximation to the conditional likelihood of α given the MLE of β. In general, maximizing adjusted profile likelihood will give less biased estimates than maximizing unadjusted profile likelihood [i.e., $l(\alpha,\hat{\beta}_\alpha)$]. Robinson and Smyth (2008) discussed its use in estimating the dispersion parameter in a constant dispersion model. For the bacteria data set, (α_0, α_1) are estimated to be (–1.857, 0.037).

The estimated normalization factors R_j and dispersion parameters ϕ_{ij} are treated as known in the regression model 2, and the regression parameters are then estimated separately for each gene. We applied the LR and HOA tests to all genes and used false discovery rate (FDR, Storey and Tibshirani, 2003) to control for multiple testing. With a FDR cutoff of 0.05, the HOA test identified 153 genes where the effect of the media has significant interaction with the genotype. For the 10 genes displaying the greatest interaction effect, Table 12 shows the estimated log fold changes in their expression levels in the two media types (KB rich versus minimal), separately for the two genotypes (wild type and hrpL mutant). The data provide strong evidence that for these genes the effect of media treatment differs for the two genotypes. For this data set, the LR and HOA tests give similar rankings of the genes, but the LR test *p*-values tended to be smaller than corresponding HOA *p*-values. LR test identified 161 genes as showing interaction at the specified FDR of 5%, including all 153 identified by the HOA test. This is consistent with the simulation results, which showed that type I error rates for the LR test tended to be slightly larger than the nominal values (so the LR test would produce more than 5% false discoveries when FDR is set at 5%).

Table 12 Estimated log (base 2) fold changes in expression levels in the two media types (KB rich versus minimal), separately for the two genotypes (DC3000 and hrpL), for 10 genes displaying the strongest interaction between the media effect and the genotype as identified by the HOA test.

	Wild type	hrpL mutant
PSPTO_1411	−4.38	3.37
PSPTO_t55	−6.04	1.33
PSPTO_2670	−3.18	1.77
PSPTO_0513	−3.82	1.21
PSPTO_2667	−4.32	0.64
PSPTO_2665	−5.27	0.49
PSPTO_4462	−5.06	0.22
PSPTO_4128	−5.55	−0.82
PSPTO_2664	−4.38	1.00
PSPTO_1412	−4.85	0.07

5 Discussion and conclusion

We envision NB testing for differential expression as an exploratory tool in an initial phase of research and also as a more formal inferential tool in a follow-up RNA-Seq study to focus more seriously on those genes identified in the exploratory phase. Since the goal of the exploratory phase is simply to identify a manageable set of genes for further study, one might question whether a type I error that departs from the nominal value by only about 10% (as the unadjusted likelihood ratio test does for small sample sizes) is in need of improvement for that purpose. We believe that it is because of the sheer magnitude of tests being performed. If, hypothetically, each of 2000 labs conducts five RNA-Seq studies per year, each testing for differential expression in 30,000 genes, that amounts to 300,000,000 statistical significance tests performed per year. If all tests are controlled to expect 5% false positives but the actual false positive rate is 5.5%, then 1.5 million unexpected false positives would be realized per year. While it may not matter much if an individual study produces 105 rather than 100 candidate genes for further investigation, the elimination of the huge number of unexpected false positives can have an impact on the long-term efficiency in overall learning from RNA-Seq analysis. The benefit of accurate testing in the follow-up studies is justified by more traditional arguments for accurate and optimal statistical inference (which are similarly best conveyed in long-term scientific learning rather than on the effect on an individual study).

On the other hand, the desirable properties of the HOA-adjusted LR test depend on the NB probability model being correct, on the dispersion parameter being functionally dependent on the mean, and on the treatment of estimated dispersion parameters as free of error. We will discuss each of these conditions individually.

First, we believe there are good reasons to believe the NB model is adequate. Although the RNA-Seq technical variability from repeated measurement of a gene from a single biological sample is Poisson (Marioni et al., 2008), the variability among biological samples has been unambiguously observed to be greater than Poisson. The Poisson means from the different biological samples within the same treatment or observational group must vary around the overall group mean, and the marginal distribution is necessarily a mixture of Poissons. Of the choices for modeling the mixing distribution, the gamma is advantageous because it leads to the convenient NB marginal distribution and because it is rich enough to encompass a wide range of possibilities for the distribution of Poisson means. Even if the gamma is not exactly the right mixing distribution for biological variability, it should, at least, be a good approximation and the NB assumption should be pretty safe.

A more controversial assumption, we believe, is that the NB dispersion parameters are functionally dependent on the means. We believe the empirical evidence for this model (Anders and Huber, 2010; Di et al., 2011) is strong enough to proceed with this type of approach, especially because the power benefits (from estimating nuisance parameters from all genes combined) are potentially substantial. To further resolve this issue, though, we are currently preparing diagnostic tools and a goodness of fit test for judging the adequacy of the NB assumption and of the model for dispersion. We will also be conducting further simulations to better articulate the power benefits.

Finally, we comment on the treatment of estimated ϕ's as being free of error. We believe this is an unnecessary simplification because likelihood inferences can be made with the more realistic and weaker assumption that the nuisance parameters in the model for the dispersion parameter are known, as discussed in Section 2. The implementation of current approaches that assume a relationship between the dispersion and the mean, though, involves fitting some kind of regression model for the mean of estimated dispersion parameters as a function of estimated NB means, from all genes combined. There is nothing wrong with using estimated means as predictor variables, but the predicted ϕ's will necessarily contain some resulting prediction error, even if the regression model is estimated perfectly. From a few preliminary simulation studies we found there to be little difference in the differential expression test properties depending on whether true or estimated ϕ's were used, but we believe this issue is in need of closer attention as NB regression methods for RNA-Seq data continue to evolve.

The programs for performing HOA test in NB regression model have been released in the latest version of the R package NBPSeq (version 0.1.6) and available through CRAN R Development Core Team (2012). The inferential techniques we develop will also be integrated into a comprehensive computational pipeline GENE-counter (Cumbie et al., 2011).

6 Appendix

6.1 Implementation details of the HOA test

Here we give the expression of the adjustment term z in Barndorff-Nielsen's r^* statistic with Skovgaard's approximations.

Let $\beta=(\psi, \nu)$ where $\psi=(\beta_1, ..., \beta_q)$ is the parameter of interest and $\nu=(\beta_{q+1}, ..., \beta_p)$ is a nuisance parameter and we wish to test the null hypothesis $\psi=\psi_0$. We let $\hat{\beta}=(\hat{\psi},\hat{\nu})$ denote the maximum likelihood estimate of the full parameter vector β and $\tilde{\beta}=(\psi_0,\tilde{\nu})$ denote the maximum likelihood estimate of ν under the null hypothesis. Let $l(\beta)=l(\beta; y)$ denote the log-likelihood, $D_1(\beta; y)$ denote the score vector

$$D_1(\beta;y)=\frac{\partial}{\partial\beta}l(\beta;y),$$

and $j(\beta)$and $i(\beta)$denote denote the observed and the Fisher information matrices, respectively:

$$j(\beta)=j(\beta;y)=-\frac{\partial^2}{\partial\beta^2}l(\beta;y).$$
$$i(\beta)=\mathrm{Var}_\beta D_1(\beta;y)=-E_\beta(j(\beta;y)).$$

With Skovgaard's approximations plugged in, the general expression for the adjustment term z in Barndorff-Nielsen's r^* statistic $r^*=r-\frac{1}{r}\log(z)$ is

$$z\approx|j(\hat{\beta})|^{-1/2}|i(\hat{\beta})||\hat{S}|^{-1}|j(\tilde{\beta})_{\nu\nu}|^{1/2}\frac{r}{[\hat{S}^{-1}\hat{q}]_\psi}, \tag{9}$$

where $j(\tilde{\beta})_{\nu\nu}$ refers to the submatrix corresponding to ν and the $[\hat{S}^{-1}\hat{q}]_\psi$ refers to the component corresponding to ψ. The two unfamiliar quantities in (9),

$$\hat{S}=\mathrm{Cov}_{\hat{\beta}}(D_1(\hat{\beta};y),D_1(\tilde{\beta};y))$$

and

$$\hat{q}=\mathrm{Cov}_{\hat{\beta}}(D_1(\hat{\beta};y),l(\hat{\beta};y)-l(\tilde{\beta};y)),$$

are approximations to the so-called sample space derivatives. Note that the quantities involved in computing z are similar to those involved in computing the observed and Fisher information matrices.

Under the NB log-linear regression model (2) introduced in Section 2,

$$Y_j\sim\mathrm{NB}(\mu_j,\phi_j),$$
$$\log(\mu_j)=\log(N_jR_j)+X_j^T\beta,$$

where $X_j=(X_{j1},..., X_{jp})^T$, $\beta=(\beta_1,..., \beta_p)^T$, and ϕ_j's are taken to be known. The likelihood from a single observation y_j, up to a constant that does not depend on β, is

$$l_j(\beta; y_j)=y_j\log(\mu_j(\beta))-(y_j+\kappa_j)\log(\mu_j(\beta)+\kappa_j) \tag{10}$$

where $\kappa_j=1/\phi_j$ and $\mu_j(\beta)=N_jR_j\exp(x_j^T\beta)$. For a set of independent RNA-Seq counts $y=(y_1,..., y_n)$,

$$l(\beta;y)=\sum_{j=1}^{n}l_j(\beta;y_j). \tag{11}$$

The score vector is

$$D_1(\beta;y)=\frac{\partial l}{\partial \beta}=\sum_j \frac{\partial l_j}{\partial \mu_j}\frac{\partial \mu_j}{\partial \beta}=\sum_j \frac{y_j-\mu_j}{\sigma_j^2}\mu_j x_j,$$

the observed information is

$$j(\beta)=-\frac{\partial^2 l}{\partial \beta^2}=\sum_{j=1}^{n}\frac{\partial^2 l}{\partial \mu_j^2}\frac{\partial \mu_j}{\partial \beta}\frac{\partial \mu_j^T}{\partial \beta}+\frac{\partial l}{\partial \mu_j}\frac{\partial^2 \mu_j}{\partial \beta \partial \beta^T} \tag{12}$$

$$=\sum_{j=1}^{n}\left[\left(\frac{y_j}{\mu_j^2}-\frac{y_j+\kappa_j}{(\mu_j+\kappa_j)^2}\right)\mu_j^2-\frac{(y_j-\mu_j)\mu_j}{\sigma_j^2}\right]x_j x_j^T, \tag{13}$$

and the Fisher information

$$i(\beta)=\sum_{j=1}^{n}\frac{\mu_j^2}{\sigma_j^2}x_j x_j^T,$$

where $\sigma_j^2=\mu_j+\phi_j\mu_j^2$. The covariance of the score vectors evaluated at $\hat{\beta}$ and at $\tilde{\beta}$ is

$$\hat{S}=\mathrm{Cov}_{\hat{\beta}}(D_1(\hat{\beta};Y),D_1(\tilde{\beta};Y))=\sum_j \hat{\mu}_j\tilde{\mu}_j\tilde{\sigma}_j^2 x_j^T x_j=\sum_j(1-\hat{p}_j)(1-\tilde{p}_j)\hat{\sigma}_j^2 x_j^T x_j.$$

and

$$\hat{q}=\mathrm{Cov}_{\hat{\beta}}(D_1(\hat{\beta}),l(\hat{\beta})-l(\tilde{\beta}))=\sum_j \hat{\mu}_j\log(\hat{p}_j/\tilde{p}_j)x_j.$$

where $p_j=p_j(\beta)=\log(\mu_j/(\mu_j+\kappa))$ and $\hat{p}_j=p_j(\hat{\beta})$ and $\tilde{p}_j=p_j(\tilde{\beta})$.

For testing a hypothesis involving several parameters, Skovgaard (2001) gave two generalizations:

$$\lambda^*=\lambda-2\log\gamma$$

and

$$\lambda^{**}=\lambda\left(1-\frac{1}{\lambda}\log\gamma\right)^2$$

which are both approximately distributed as $\chi^2(q)$, where γ can be approximated by

$$|\tilde{i}|^{1/2}|\hat{i}|^{1/2}|\hat{S}|^{-1}|\tilde{j}_{vv}|^{1/2}|[\tilde{i}\hat{S}^{-1}\hat{j}\hat{i}\hat{S}]_{vv}|^{-1/2}\frac{\{\tilde{D}_1^T\hat{S}^{-1}\tilde{i}\tilde{j}^{-1}\hat{S}\tilde{i}^{-1}\tilde{D}_1\}^{q/2}}{w^{q/2-1}\tilde{D}_1^T\hat{S}^{-1}\hat{q}}.$$

The second statistic λ^{**} reduces to the square of r^* when q=1.

6.2 NB regression models and full-rank exponential families

For comparing mean relative frequencies between two groups of NB counts, $y_1,\ldots,y_{n_1}$ and $y_{n_1+1},\ldots,y_{n_1+n_2}$, if all effective library sizes $N_j R_j$ are the same and all dispersion parameters κ_j are the same (say, equal to κ), the log likelihood of the two group means μ_1 and μ_2 is [cf. Eqs. (10) and (11)]

$$\sum_{j=1}^{n_1} y_j \log\frac{\mu_1}{\mu_1+\kappa} - \sum_{j=n_1+1}^{n_2} y_j \log\frac{\mu_2}{\mu_2+\kappa} - n_1\kappa\log(\mu_1+\kappa) - n_2\kappa\log(\mu_2+\kappa),$$

which belongs to a two-dimensional full-rank exponential family. [In this special case, a simpler formula from Pierce and Peters (1992) can be used to compute the adjustment term z, yielding identical results as using Eq. (9) in Appendix 6.1.]

For more general NB regression models, each library will have a different mean frequency μ_j and a different dispersion parameter κ_j, so the joint distribution of Y_j's does not belong to a full-rank exponential family but rather to a curved exponential family.

6.3 Implementation details of the adjusted profile likelihood

In the adjusted profile likelihood (8), the expression for $j_{\beta\beta}$ is given in (13). The probability mass function of a single NB random variable Y with mean μ and shape parameter κ (the reciprocal of the dispersion) is

$$\Pr(Y=y;\mu,\kappa)=\frac{\Gamma(\kappa+y)}{\Gamma(\kappa)\Gamma(1+y)}\left(\frac{\mu}{\mu+\kappa}\right)^y\left(\frac{\kappa}{\mu+\kappa}\right)^\kappa.$$

The likelihood from a single observation under the (μ, κ) parameterization is

$$l(\mu, \kappa; y)=\log(\Gamma(\kappa+y))-\log(\Gamma(\kappa))+y\log(\mu)+\kappa\log(\kappa)-(y+\kappa)\log(\mu+\kappa).$$

The full model likelihood of (α, β) is given by summing this over all observations after expressing μ and κ in terms of α and β according to the dispersion model (7) and the NB regression model (2). In the dispersion model (7), preliminary estimates of μ_{ij} are needed. One can estimate these quantities from the regression model (2) by assuming a constant dispersion (e.g., ϕ=0.1). Precise mean estimates are unnecessary for the dispersion model since under that model, the dispersion changes gradually with the mean.

Acknowledgements: Research reported in this publication was partially supported by the National Institute of General Medical Sciences of the National Institutes of Health under Award Number R01GM104977 (to YD, SCE, and JHC). The content is solely the responsibility of the authors and does not necessarily represent the official views of the National Institutes of Health. Work in the Chang lab is supported by the National Research Initiative Competitive Grants Program Grant no. 2008-35600-04691 and Agriculture and Food Research Initiative Competitive Grants Program Grant no. 2011-67019-30192 from the USDA National Institute of Food and Agriculture, National Science Foundation (Grant no. IOS-1021463), and the Agricultural Research Foundation. We would like to thank Prof. Don Pierce for many helpful discussions. We would also like to thank Mark Dasenko, Chris Sullivan, and Matthew Peterson of the CGRB core facility for their assistance with RNA-Seq preparation and data processing. We thank the reviewers for their insightful and constructive comments.

References

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

Barndorff-Nielsen, O. (1986): "Infereni on full or partial parameters based on the standardized signed log likelihood ratio," Biometrika, 73, 307–322.

Barndorff-Nielsen, O. (1991): "Modified signed log likelihood ratio," Biometrika, 78, 557–563.

Buell, C., V. Joardar, M. Lindeberg, J. Selengut, I. Paulsen, M. Gwinn, R. Dod-son, R. Deboy, A. Durkin, J. Kolonay, et al. (2003): "The complete genome sequence of the Arabidopsis and tomato pathogen Pseudomonas syringae pv. tomato DC 3000," Proc. Natl. Acad. Sci. USA, 100, 10181.

Cox, D. R. and N. Reid (1987): "Parameter orthogonality and approximate conditional inference," J. R. Stat. Soc. Series B Stat. Methodol., 49, 1–39.
Cumbie, J., J. Kimbrel, Y. Di, D. Schafer, L. Wilhelm, S. Fox, C. Sullivan, A. Curzon, J. Carrington, T. Mockler, et al. (2011): "GENE-Counter: A computational pipeline for the analysis of RNA-Seq data for gene expression differences," PLoS ONE, 6, e25279.
Di, Y., D. Schafer, J. Cumbie, and J. Chang (2011): "The NBP negative binomial model for assessing differential gene expression from RNA-Seq," Stat. Appl. Genet. Mol. Biol., 10, 24.
Fears, T., J. Benichou, and M. Gail (1996): "A reminder of the fallibility of the wald statistic," American Statistician, 50, 226–227.
Hilbe, J. M. (2007): Negative Binomial Regression. Cambridge, UK: Cambridge University Press.
Huynh, T., D. Dahlbeck, and B. Staskawicz (1989): "Bacterial blight of soybean: regulation of a pathogen gene determining host cultivar specificity," Science, 245, 1374.
Lancaster, H. (1961): "Significance tests in discrete distributions," J. Am. Stat. Assoc., 56, 223–234.
Lindeberg, M., S. Cartinhour, C. Myers, L. Schechter, D. Schneider, and A. Collmer (2006): "Closing the circle on the discovery of genes encoding Hrp regulon members and type III secretion system effectors in the genomes of three model pseudomonas syringae strains," Mol. Plant Microbe Interact., 19, 1151–1158.
Lund, S., D. Nettleton, D. McCarthy, and G. Smyth (2012): "Detecting differential expression in RNA-sequence data using quasi-likelihood with shrunken dispersion estimates," Stat. Appl. Genet. Mol. Biol., 11, in press. ISSN online (1544–6115).
Marioni, J. C., C. E. Mason, S. M. Mane, M. Stephens, and Y. Gilad (2008): "RNA-seq: an assessment of technical reproducibility and comparison with gene expression arrays," Genome Res., 18, 1509–1517.
McCarthy, D., Y. Chen, and G. Smyth (2012): "Differential expression analysis of multifactor RNA-Seq experiments with respect to biological variation," Nucleic Acids Res., 40, 4288–4297.
Mortazavi, A., B. A. Williams, K. McCue, L. Schaeffer, and B. Wold (2008): "Mapping and quantifying mammalian transcriptomes by RNA-Seq," Nat. Methods, 5, 621–628.
Nagalakshmi, U., Z. Wang, K. Waern, C. Shou, D. Raha, M. Gerstein, and M. Snyder (2008): "The transcriptional landscape of the yeast genome defined by RNA sequencing," Science, 320, 1344–1349.
Pierce, D. and D. Peters (1992): "Practical use of higher order asymptotics for multiparameter exponential families," J. R. Stat. Soc. Series B Stat. Methodol., 54, 701–737.
Pierce, D. and D. Peters (1999): "Improving on exact tests by approximate conditioning," Biometrika, 86, 265–277.
Robinson, M. D. and G. K. Smyth (2007): "Moderated statistical tests for assessing differences in tag abundance," Bioinformatics, 23, 2881–2887.
Robinson, M. D. and G. K. Smyth (2008): "Small-sample estimation of negative binomial dispersion, with applications to SAGE data," Biostat., 9, 321–332.
Robinson, M. D. and A. Oshlack (2010): "A scaling normalization method for differential expression analysis of RNA-seq data," Genome Biol., 11, R25.
Robinson, M. D., D. J. McCarthy, and G. K. Smyth (2010): "edgeR: a Bioconductor package for differential expression analysis of digital gene expression data," Bioinformatics, 26, 139–140.
R Development Core Team (2012): R: A Language and Environment for Statistical Computing, R Foundation for Statistical Computing, Vienna, Austria, URL http://www.R-project.org/ , ISBN 3-900051-07-0.
Skovgaard, I. (1996): "An explicit large-deviation approximation to one-parameter tests," Bernoulli, 2, 145–165.
Skovgaard, I. (2001): "Likelihood asymptotics," Scandinavian Journal of Statistics, 28, 3–32.
Storey, J. D. and R. Tibshirani (2003): "Statistical significance for genomewide studies," Proc. Natl. Acad. Sci. USA, 100, 9440–9445.
Trapnell, C., B. Williams, G. Pertea, A. Mortazavi, G. Kwan, M. Van Baren, S. Salzberg, B. Wold, and L. Pachter (2010): "Transcript assembly and quantification by RNA-Seq reveals unannotated transcripts and isoform switching during cell differentiation," Nat. Biotechnol., 28, 511–515.
Væth, M. (1985): "On the use of Wald's test in exponential families," Int. Stat. Rev., 53, 199–214.
Venables, W. and B. Ripley (2002): Modern applied statistics with S. New York, USA: Springer verlag.
Wald, A. (1941): "Asymptotically most powerful tests of statistical hypotheses," Ann. Math. Statist., 12, 1–19.
Wald, A. (1943): "Tests of statistical hypotheses concerning several parameters when the number of observations is large," Trans. Amer. Math. Soc., 54, 426–482.
Wang, Z., M. Gerstein, and M. Snyder (2009): "RNA-Seq: a revolutionary tool for transcriptomics," Nat. Rev. Genet., 10, 57–63.
Wilks, S. (1938): "The large-sample distribution of the likelihood ratio for testing composite hypotheses," Ann. Math. Statist., 9, 60–62.
Zhou, Y., K. Xia, and F. Wright (2011): "A powerful and flexible approach to the analysis of rna sequence count data," Bioinformatics, 27, 2672–2678.