Statistical Applications in Genetics and Molecular Biology

Volume 11, *Issue* 5 2012 *Article* 8

Detecting Differential Expression in RNA-sequence Data Using Quasi-likelihood with Shrunken Dispersion Estimates

Steven P. Lund, *Statistical Engineering Division, National Institute of Standards and Technology*
Dan Nettleton, *Department of Statistics, Iowa State University*
Davis J. McCarthy, *University of Oxford*
Gordon K. Smyth, *Walter and Eliza Hall Institute of Medical Research*

Recommended Citation:
Lund, Steven P.; Nettleton, Dan; McCarthy, Davis J.; and Smyth, Gordon K. (2012) "Detecting Differential Expression in RNA-sequence Data Using Quasi-likelihood with Shrunken Dispersion Estimates," *Statistical Applications in Genetics and Molecular Biology*: Vol. 11: Iss. 5, Article 8.

DOI: 10.1515/1544-6115.1826

Detecting Differential Expression in RNA-sequence Data Using Quasi-likelihood with Shrunken Dispersion Estimates

Steven P. Lund, Dan Nettleton, Davis J. McCarthy, and Gordon K. Smyth

Abstract

Next generation sequencing technology provides a powerful tool for measuring gene expression (mRNA) levels in the form of RNA-sequence data. Method development for identifying differentially expressed (DE) genes from RNA-seq data, which frequently includes many low-count integers and can exhibit severe overdispersion relative to Poisson or binomial distributions, is a popular area of ongoing research. Here we present quasi-likelihood methods with shrunken dispersion estimates based on an adaptation of Smyth's (2004) approach to estimating gene-specific error variances for microarray data. Our suggested methods are computationally simple, analogous to ANOVA and compare favorably versus competing methods in detecting DE genes and estimating false discovery rates across a variety of simulations based on real data.

KEYWORDS: differential expression, quasi-likelihood, RNA-seq

Author Notes: This work was supported by funding from the following sources: National Research Initiative of the USDA-CSREES Grant No. 2008-35600-18786 (to SPL and DN); National Science Foundation grant number 0820610 (to DN); General Sir John Monash Scholarship (to DJM); National Health and Medical Research Council Program Grant 490037 (to GKS). Disclaimer: The identification of any commercial products is given only for the sake of completely describing our experimental procedures. In no instance does such identification imply any recommendation by the National Institute of Standards and Technology; nor does it imply that the particular equipment identified is necessarily the best available for the described process. Steven P. Lund is also affiliated to Department of Statistics, Iowa State University. Davis J. McCarthy is also affiliated to Walter and Eliza Hall Institute of Medical Research. Gordon K. Smyth is also affiliated to Department of Mathematics and Statistics, University of Melbourne.

1 Introduction

Next-generation sequencing (NGS) technologies are powerful and increasingly popular tools used to identify differentially expressed genes, among other gene expression characteristics. RNA-Seq and SAGE technologies provide discrete count data serving as measures of messenger RNA (mRNA) expression levels through the following procedure. The mRNA is isolated from sample cells, fragmented, and copied to complementary DNA (cDNA). The cDNA fragments are then amplified and sequenced, and the resulting reads are aligned with a reference genome. The number of reads mapped within each reference gene provides the RNA-Seq count data. This paper considers integer counts, typically ranging from zero to many thousands, of single-end reads that uniquely map to a single gene. Because of the frequent presence of low integers, methods developed for analyzing microarray data, which can be modeled as a continuous response, are not generally appropriate for analyzing RNA-Seq data.

As NGS has grown in popularity among researchers exploring differential expression, many statistical methods have been proposed for handling the subsequent expression data. Poisson or binomial (with n fixed as the sample library size) generalized linear models (GLM) could certainly handle low integer counts present in RNA-Seq data. However, upon modeling data with biological replicates within experimental conditions, it is clear that the restrictive mean-variance relationships for the Poisson and binomial distributions do not adequately accommodate the variability present in RNA-Seq data. That is, the RNA-Seq data are overdispersed, exhibiting greater variability across biological replicates than Poisson or binomial models predict.

In the face of overdispersion, one option is to add random effects to the original GLM, creating a generalized linear mixed effects model, as demonstrated by Blekhman et al. (2010). Another option is to choose a more flexible distribution. Zhou et al. (2011) and Vêncio et al. (2004) use beta-binomial models to account for overdispersion. Several methods, including Lu et al. (2005); Robinson and Smyth (2007, 2008); McCarthy et al. (2012); Anders and Huber (2010); Di et al. (2011), are based on the negative-binomial distribution, which has two parameters (mean μ and dispersion ω) and a more flexible mean-variance relationship than the Poisson or binomial (with fixed n) distributions. Although the negative binomial distribution provides flexibility in modeling variances, existing popular methods based on this distribution fail to adequately account for uncertainty in parameter estimates. A simulation study described in Section 4 demonstrates that most of these methods produce an over-abundance of small p-values for tests with true null hypotheses, relative to a uniform distribution, even for data simulated from negative binomial distributions. Although it ignores uncertainty in its estimated dispersion parameters,

DESeq (Anders and Huber, 2010) produces too few small null p-values because its estimation procedure systematically overestimates negative binomial dispersion parameters. The resulting non-uniform distributions of null p-values obtained from these methods are shown to produce q-values (Storey and Tibshirani, 2003) that inaccurately estimate false discovery rates.

Tjur (1998) describes a general use quasi-likelihood (QL) approach to adjusting for overdispersion. To implement Tjur's method for RNA-Seq data, average counts for observations from the kth gene are modeled in the typical GLM fashion by specifying covariates and a link function. The variance of each observation from gene k is assumed to be a user-specified function of its modeled average, multiplied by a gene-specific quasi-dispersion parameter denoted by Φ_k. The QL approach then compares the ratio $LRT_k/(q\hat{\Phi}_k)$ to an appropriate F-distribution, where LRT_k is a quasi-likelihood ratio test statistic for the kth gene, q is the difference between the dimensions of the full and null-constrained parameter spaces, and $\hat{\Phi}_k$ is an estimate of the dispersion for the kth gene. Auer and Doerge (2011) suggest a two-stage Poisson model (TSPM), which first tests each gene for overdispersion (i.e. $\Phi_k > 1$) and then adjusts a Poisson model likelihood ratio test (LRT) for significantly overdispersed genes using a form of Tjur's QL method.

A drawback to using Tjur's QL approach with RNA-Seq data is that while many methods exist for estimating the quasi-dispersion for a single gene, there are often few degrees of freedom available for these estimates. In Section 2, we propose adapting Smyth's (2004) approach to estimating gene-specific error variances for microarray data in order to share information across genes when estimating gene-specific dispersion parameters for the QL approach. The resulting new methods are powerful, robust and fast, and accommodate all experimental designs that can be analyzed by an ordinary GLM. These suggested QL methods are analogous to ANOVA with shrunken variance estimates, where deviances are analogs to sums of squares.

In Section 3, we apply our new methods to real RNA-Seq data and compare results with several other popular methods. Section 4 describes simulation studies that demonstrate our recommended approach offers significantly improved differential expression significance rankings and better estimates of false discovery rates compared to competing methods when analyzing RNA-Seq data sets with small to moderate sample sizes common in practice. We provide brief commentary regarding the suggested methods and alternative methods based on exact tests in Section 5. Section 6 contains supplementary materials, including additional descriptive plots and example code for conducting two analyses of real data with the suggested methods of this article via the R (R Development Core Team, 2011) package QuasiSeq.

2 Method Description

2.1 Review of Related Methods

Auer and Doerge (2011) developed a quasi-likelihood approach for analyzing RNA-Seq data called TSPM. This approach first tests each gene for overdispersion, relative to a fitted Poisson model, and then adjusts the likelihood ratio test (LRT) for significantly overdispersed genes using a form of Tjur's QL method. Our simulation studies in Section 4 show that this approach will tend to correct for overdispersion only when it is severe and that this can lead to very liberal tests for differential expression. The proposed methods in this article use a more conservative approach to adjusting for overdispersion and provide the additional advantage of sharing information across genes when estimating dispersions.

The negative binomial distribution is popular among methods for analyzing RNA-Seq data. (See, for example, edgeR (Robinson and Smyth, 2007, 2008; Robinson et al., 2010; McCarthy et al., 2012), DESeq (Anders and Huber, 2010) and NBPSeq (Di et al., 2011), which all use negative binomial models to analyze RNA-Seq data.) For a detailed review of these methods, see McCarthy et al. (2012). While offering several ways to estimate negative binomial dispersion parameters, these methods all treat the resulting estimates as known constants when testing for differential expression and can be shown to produce liberal p-values, with the exception of DESeq, for which the distribution of p-values is often J-shaped. Among the popular methods based on the negative binomial distribution, the GLM version of edgeR is most closely related to the methods of this article in that it allows gene-specific dispersion estimates to vary around a central estimated trend and shares information across genes when estimating dispersions. The quasi-likelihood methods proposed in this article provide the additional advantages of incorporating uncertainty in estimated variances when testing for differential expression and providing a self-tuning approach to shrinking gene-specific dispersion estimates.

2.2 QL Method

We begin fitting a quasi-likelihood model by specifying a model for the mean and, up to a multiplicative constant, the variance for each observation as a function of its mean. Let Y_{ijk} represent the observed count for gene k in replicate j ($j = 1,\ldots,J$) of treatment group i ($i = 1,\ldots,I$), and let c_{ij} represent a normalization factor for the overall number of reads from replicate j in treatment group i (e.g., we set c_{ij} as the 0.75 quantile of reads from replicate j in treatment group i as recommended by Bullard et al. (2010)). Let $\mathrm{E}(Y_{ijk}|c_{ij}) = \mu_{ijk}$ where $\mu_{ijk} = \lambda_{ik}c_{ij}$ and λ_{ik} represents

the normalized expression level of gene k in treatment group i. In this framework, gene k is defined to be equivalently expressed (EE) across treatments i and i' if $\lambda_{ik} = \lambda_{i'k}$ and differentially expressed (DE) otherwise. More generally, we can model $\log(\mu_{ijk})$ as a known constant ($\log c_{ij}$) plus a linear function of covariates and treatment effects. Such extensions are straightforward and are not considered here to simplify the presentation.

Fitting a quasi-likelihood model requires specifying the variance of observed values, up to a proportionality constant, as a function of the modeled means. That is, one assumes $\text{Var}(Y_{ijk}) = \Phi_k V_k(\mu_{ijk})$, where $V_k(\mu_{ijk})$ is fully specified by the user and Φ_k is an unknown dispersion parameter that will be estimated from the data. Tables of commonly used variance functions, $V(\mu)$, and their corresponding quasi-likelihood functions can be found in McCullagh (1983) and McCullagh and Nelder (1983). For RNA-Seq data, it seems most reasonable to use $V_k(\mu_{ijk}) = \mu_{ijk} + \omega_k \mu_{ijk}^2$ (based on the negative binomial distribution, with some specified value of ω_k) or $V_k(\mu_{ijk}) = \mu_{ijk}$ (based on the Poisson distribution). However, our suggested methods can be used with any variance function for which there exists a corresponding quasi-likelihood function, $\ell(\mu_{ijk}|y_{ijk})$, that satisfies

$$\frac{\partial \ell(\mu_{ijk}|y_{ijk})}{\partial \mu_{ijk}} = \frac{y_{ijk} - \mu_{ijk}}{V_k(\mu_{ijk})}.$$

Note that both Φ_k and ω_k are dispersion parameters; ω_k (referred to as negative binomial dispersion) is a parameter of the negative binomial distribution, and Φ_k (referred to as quasi-likelihood dispersion) is a proportionality constant used in quasi-likelihood models. In a quasi-negative binomial model, both ω_k and Φ_k are used to model the variance of observations from gene k; i.e., $\text{Var}(Y_{ijk}) = \Phi_k \left(\mu_{ijk} + \omega_k \mu_{ijk}^2\right)$.

The use of a quasi-likelihood approach based on a negative binomial distribution may seem unnecessary, as the negative binomial distribution has two parameters and provides great flexibility in modeling mean-variance relationships. However, existing popular methods for detecting differential expression with RNA-Seq data based on the negative binomial distribution fail to adequately account for uncertainty in the modeled variance. The simulations in Section 4 demonstrate that ignoring this uncertainty produces an over-abundance of small p-values from EE genes, relative to a uniform distribution, even for data simulated from negative binomial distributions. These non-uniform distributions of p-values from EE genes are shown to produce q-values that substantially underestimate false discovery rates (FDR). A negative binomial implementation of the quasi-likelihood methods, using negative binomial dispersion parameter estimates from the GLM implementation of

edgeR (Robinson et al., 2010; McCarthy et al., 2012), however, was found to produce far more accurate q-values. The important benefit of using a quasi-likelihood approach based on a negative binomial distribution is not the additional flexibility in modeling variances but rather the incorporation of uncertainty in the modeled variances via the estimated quasi-likelihood dispersion parameter.

For each of K genes, parameters for the modeled means are estimated by maximizing

$$\ell_k(\hat{\boldsymbol{\mu}}_k|\boldsymbol{y}_k) = \sum_{i,j} \ell_k(\hat{\mu}_{ijk}|y_{ijk}), \tag{1}$$

where $\boldsymbol{y}_k = (y_{11k}, \ldots, y_{IJk})'$ is the vector of observations from the kth gene across samples, $\boldsymbol{\mu}_k = (\mu_{11k}, \ldots, \mu_{IJk})'$ is the vector of the corresponding means, and $\ell_k(\mu|y)$ is the quasi-likelihood function corresponding to the variance function chosen for gene k.

Conducting a hypothesis test for differential expression using the quasi-likelihood approach involves computing a quasi-likelihood ratio test statistic and estimating the dispersion parameter, Φ_k (the proportionality constant from the specified mean-variance relationship). The quasi-likelihood ratio test statistic is computed as

$$LRT_k = 2\left(\ell_k(\hat{\boldsymbol{\mu}}_k|\boldsymbol{y}_k) - \ell_k(\tilde{\boldsymbol{\mu}}_k|\boldsymbol{y}_k)\right), \tag{2}$$

where $\tilde{\mu}_{ijk}$ and $\hat{\mu}_{ijk}$ are the maximum quasi-likelihood estimates for μ_{ijk} under the null and alternative hypotheses, respectively. When the mean-variance function has been correctly specified, McCullagh (1983) shows that under the null hypothesis

$$LRT_k \sim \Phi_k \chi^2_q + O_p(n^{-1/2}), \tag{3}$$

where q is the difference between the dimensions of the full and null-constrained mean parameter spaces and n is the total number of samples.

The dispersion parameter, Φ_k, can be estimated as

$$\hat{\Phi}_k = \frac{2\left(\ell_k(\boldsymbol{y}_k|\boldsymbol{y}_k) - \ell_k(\hat{\boldsymbol{\mu}}_k|\boldsymbol{y}_k)\right)}{n-p}, \tag{4}$$

where p is the dimension of the full-model mean parameter space. This deviance based estimator of Φ_k is asymptotically independent of maximum likelihood estimates for the parameters used to model $\boldsymbol{\mu}_k$ (McCullagh, 1983). Although this estimator has a similar form to Equation 2, its asymptotic distribution does not follow from Equation 3 for as n tends to ∞, $n-p$ also tends to ∞, and the derivation of Equation 3 requires that q be finite. For distributions that are asymptotically normal, as $\mu \to \infty$, (including Poisson distributions, but not other negative binomial distributions) Tjur (1998) shows that as counts (rather than the number of samples,

n) tend to $\infty, \hat{\Phi}_k \sim \Phi_k \chi^2_{n-p}$ by approximating the quasi-likelihood models with non-linear regression models. Tjur suggests comparing the test statistic

$$F_{QL} = \frac{LRT_k/q}{\hat{\Phi}_k}$$

to an F-distribution with q and $n-p$ degrees of freedom. We refer to this approach as QL for quasi-likelihood.

While other dispersion estimators have better understood asymptotic distributions, we originally chose Equation 4 due to its symmetry with Equation 2. The numerator of F_{QL} is twice the difference between quasi-likelihoods of the full and reduced models, divided by the difference between the dimensions of the unconstrained and null-constrained parameter spaces. That is, the numerator is an estimate of the average change in deviance per constrained parameter. The denominator of F_{QL} is the estimated dispersion and, when the suggested deviance estimator is used, is equal to twice the difference between quasi-likelihoods of the saturated and full models, divided by the residual degrees of freedom. That is, the denominator is an estimate of the average change in deviance per residual degree of freedom. F_{QL} thus provides the average number of residual degrees of freedom each parameter constrained by the null hypothesis is worth in terms of change in deviance. This is an exact parallel to the F-tests produced in standard ANOVA tables and, as the simulation studies described in Section 4 demonstrate, makes the QL method robust to model misspecification.

Among alternative dispersion estimators, the most popular may be Pearson's estimator,

$$\hat{\Phi}_k^{Pearson} = \frac{1}{n-p} \sum_{i=1}^{I} \sum_{j=1}^{J} \left(Y_{ijk} - \hat{\mathrm{E}}(Y_{ijk})\right)^2 / \widehat{\mathrm{Var}}(Y_{ijk}).$$

We examined the performance of our suggested methods using Pearson's dispersion estimator in place of the deviance estimator. In general, Pearson dispersion estimates tended to be smaller than the corresponding deviance based dispersion estimates, and using the Pearson estimates led to liberal results (i.e. over-abundance of small p-values from EE genes and q-values that underestimated empirical FDRs), particularly for the quasi-negative binomial methods. We therefore recommend the deviance dispersion estimator when using methods described in this paper.

2.3 QLShrink Method

It is common for $n-p$ to be small in RNA-Seq experiments, so the QL approach often can be substantially improved by sharing information across genes when estimating dispersion parameters. We suggest adapting the method described in Smyth

(2004) for estimating gene-specific error variances for multiple linear models. Our approach places a scaled-inverse χ^2 prior distribution with d_0 degrees of freedom and scaling factor Φ_0 on each gene's dispersion, such that

$$d_0\Phi_0/\Phi_k \sim \chi^2_{d_0}. \tag{5}$$

We further assume that

$$(n-p)\hat{\Phi}_k/\Phi_k|\Phi_k \sim \chi^2_{n-p}, \tag{6}$$

based on, but not theoretically justified by, Equations 3 and 4. These assumptions produce an inverse-gamma posterior distribution such that

$$1/\Phi_k|\hat{\Phi}_k \sim \text{gamma}\left[.5(d_0+n-p), .5\left(d_0\Phi_0+(n-p)\hat{\Phi}_k\right)\right].$$

The hyperparameters d_0 and Φ_0 can be estimated from the distribution of $\hat{\Phi}_k$ using a method of moments approach described by Smyth (2004). A natural estimator of Φ_k can be formed by using the estimated posterior expectation as follows:

$$\hat{\Phi}^s_k = \hat{\text{E}}^{-1}(\Phi_k^{-1}|\hat{\Phi}_k) = \frac{\hat{d_0}\hat{\Phi}_0+(n-p)\hat{\Phi}_k}{\hat{d_0}+(n-p)}. \tag{7}$$

We compare the test statistic $LRT_k/(q\hat{\Phi}^s_k)$ to an F-distribution with q and $\hat{d_0}+n-p$ degrees of freedom. Given that Marioni et al. (2008) showed that variability among technical replicates is consistent with a Poisson model, we do not expect RNA-Seq data from biological replicates to be underdispersed. Thus, when using a quasi-Poisson model, we suggest using $\tilde{\Phi}^s_k = \max(1, \hat{\Phi}^s_k)$ as an estimator of Φ_k and comparing the test statistic $LRT_k/(q\tilde{\Phi}^s_k)$ to an F-distribution with q and $\hat{d_0}+n-p$ degrees of freedom. We refer to this approach as QLShrink.

2.4 QLSpline Method

A clear relationship is often present between estimated dispersions and average counts. (See Figure 1, for example.) In this scenario, it is beneficial to define a prior scaling factor, Φ_{0k}, for each gene as a function of the gene's average count. We recommend fitting a cubic spline to $\log(\hat{\Phi}_k)$ versus $\log(\bar{y}_{\cdot\cdot k})$, using cross-validation to determine the appropriate degrees of freedom to allow when fitting the spline. Let $S_0(\cdot)$ be the resulting continuous function, and let $\hat{\Phi}_{0k} = \exp[S_0(\log \bar{y}_{\cdot\cdot k})]$. Under the assumption that the distribution of $\Phi_k|\hat{\Phi}_{0k}$ is defined by

$$d_0'\hat{\Phi}_{0k}/\Phi_k|\hat{\Phi}_{0k} \sim \chi^2_{d_0'}$$

and that Equation 6 holds, the ratio $\hat{\Phi}_k/\hat{\Phi}_{0k}|\hat{\Phi}_{0k}$ follows an F-distribution with $n-p$ and d_0' degrees of freedom for all k. When the cubic spline is fit on the log scale, we

recommend allowing added flexibility by assuming $\hat{\Phi}_k/\hat{\Phi}_{0k}|\hat{\Phi}_{0k}$ follows a scaled F-distribution, with scaling factor γ. We then apply Smyth's method of moments approach to the set $\{\hat{\Phi}_k/\hat{\Phi}_{0k}\}_{k=1}^K$ to obtain estimates $\hat{d}_0'$ and $\hat{\gamma}$. Our suggested estimator for the kth gene's dispersion is

$$\hat{\Phi}_k^{(spline)} = \frac{\hat{d}_0'\hat{\Phi}_{0k}\hat{\gamma} + (n-p)\hat{\Phi}_k}{\hat{d}_0' + (n-p)}. \tag{8}$$

Fitting the cubic spline to $\log\left(\hat{\Phi}_k\right)$, as opposed to $\hat{\Phi}_k$, reduces the influence of extreme estimates on the spline fit but also produces estimates $\hat{\Phi}_{0k}$ that are too small. The additional scaling factor γ serves as a correction for using the log-scale and is strongly recommended by the authors. Fixing $\hat{\gamma} = 1$ in Equation 8 causes methods using the estimator to produce liberal results, particularly for small sample sizes. (e.g. For simulations with total sample sizes less than six, $\hat{\gamma}$ was often around 1.5.)

This estimation procedure shrinks $\hat{\Phi}_k$ toward $\hat{\Phi}_{0k}\hat{\gamma}$, which is a scale-adjusted, spline-based estimate of Φ_k. The extent of shrinkage depends on $\hat{d}_0'$ relative to $n-p$. As the degree of scatter around the spline fit (like that in Figure 1) decreases, $\hat{d}_0'$ increases and $\hat{\Phi}_{0k}\hat{\gamma}$ is more heavily weighted in $\hat{\Phi}_k^{(spline)}$. Conversely, as the scatter around the spline fit increases or as $n-p$ increases, the dispersion estimate based on the data for the kth gene, $\hat{\Phi}_k$, is more heavily weighted in $\hat{\Phi}_k^{(spline)}$. We then compare $LRT_k/(q\hat{\Phi}_k^{(spline)})$ to an F-distribution with q and $\hat{d}_0' + n - p$ degrees of freedom. As before, when using a quasi-Poisson model, we recommend letting $\tilde{\Phi}_k^{(spline)} = \max(1, \hat{\Phi}_k^{(spline)})$ and comparing $LRT_k/(q\tilde{\Phi}_k^{(spline)})$ to an F-distribution with q and $\hat{d}_0' + n - p$ degrees of freedom. We refer to this approach as QLSpline.

For this article, we consider Poisson and negative binomial implementations of the QL, QLShrink and QLSpline methods and use prefixes "Pois" and "NegBin" to denote which distribution was used when discussing results. Using a quasi-negative binomial model requires providing the negative binomial dispersion parameter ω_k in the equation $\text{Var}(Y_{ijk}) \propto \mu_{ijk} + \omega_k\mu_{ijk}^2$. The provided estimate $\hat{\omega}_k$ is treated as a known constant when estimating mean parameters (by maximizing Equation 1) and the quasi-dispersion parameter (according to Equation 4). For this paper, we provide estimates obtained from edgeR (Robinson et al., 2010) using the 'estimateGLMTrendedDisp' (McCarthy et al., 2012) function. Unless otherwise specified, we use the default settings of this function.

3 Data Analysis

3.1 Fly Embryo data set

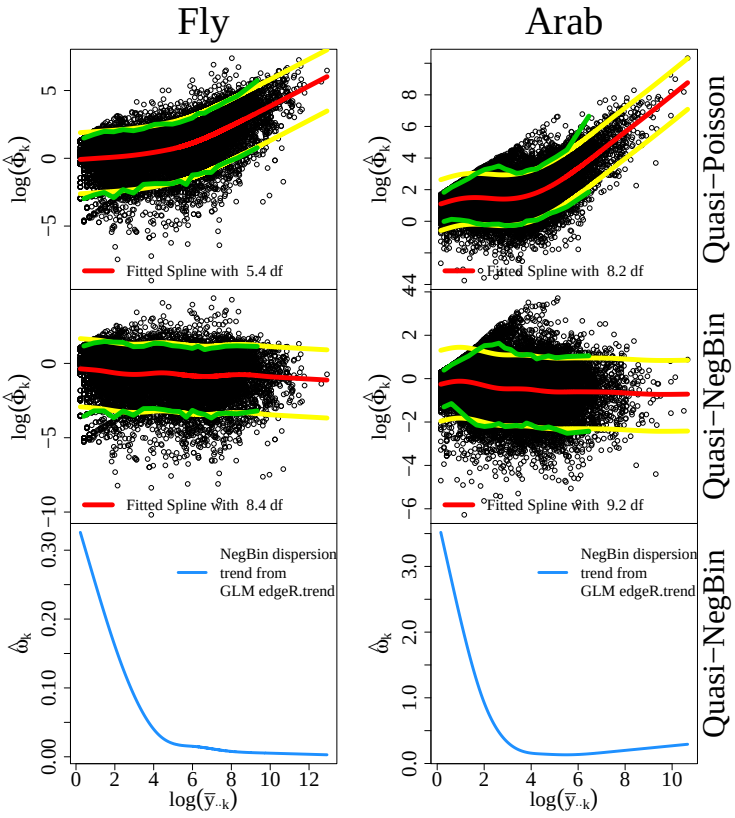

Figure 1: Estimated quasi-dispersions $\left(\hat{\Phi}_k\right)$ from quasi-Poisson (top) and quasi-negative binomial (middle) models versus average count with fitted splines for fly embryo (left) and Arabidopsis (right) data. Green and yellow lines provide 0.05/0.95 quantiles from empirical and fitted scaled-F distributions, respectively. Estimated dispersions for negative binomial distribution $(\hat{\omega}_k)$ from GLM edgeR.trend are shown in bottom row.

We first examine the fly embryo data set provided in Anders and Huber (2010) from RNA-Seq experiments on fly embryos conducted by B. Wilczynski, Y.-H. Liu, N. Delhomme, and E. Furlong. The data set includes count data for two biological replicates in each of two treatment groups labeled A and B, respectively. The left side of Figure 1 contains a scatterplot of the estimated quasi-Poisson and

quasi-NegBin dispersions versus the average count for each gene for these data, along with the corresponding fitted cubic-splines used in the QLSpline methods. There is little relation between quasi-dispersion estimates, $\hat{\Phi}_k$, and the average count for the quasi-NegBin model. This is not surprising because the negative binomial dispersion parameter estimates, $\hat{\omega}_k$, used in the quasi-NegBin model come from the edgeR trend, also shown in Figure 1, which already captures the relationship between dispersion and average count.

Figure 1 also provides a comparison between the 0.05 and 0.95 quantiles of the empirical and estimated scaled-F distributions of $\hat{\Phi}_k$. The quantiles for the estimated scaled F-distribution are shown as yellow curves and are given by $\hat{\Phi}_{0k}\hat{\gamma}F_{\alpha,n-p,\hat{d}_0'}$, for $\alpha = 0.05, 0.95$. Quantiles from the empirical distribution of $\hat{\Phi}_k$ appear as green curves and are computed by sorting all included genes into 20 bins according to their total count and taking the 0.05 and 0.95 quantile for $\hat{\Phi}_k$ within each bin. These curves indicate there is good agreement between the empirical and modeled distributions of $\hat{\Phi}_k$.

The data set contains 13230 genes with average counts greater than one and for which there were at least two samples with non-zero counts. For the purpose of comparing different methods of analysis, we tested these genes for differential expression between groups A and B with the following methods: DESeq (Anders and Huber, 2010), TSPM (Auer and Doerge, 2011), NBPSeq (Di et al., 2011), six implementations of edgeR (Robinson et al., 2010) formed by factorial combinations of testing procedure (exact (Robinson and Smyth, 2007, 2008) or GLM (McCarthy et al., 2012)) and dispersion estimation method (common dispersion [com], non-trended tagwise [tgw], or trended tagwise [trend]), and the QL, QLShrink and QLSpline methods applied to quasi-Poisson and quasi-negative binomial models. For each method, its recommended approach was used to account for differences in library sizes. The QL method group and TSPM used the 0.75 quantile of the read count distribution from each sample, as recommended by Bullard et al. (2010). Throughout this manuscript, library size offsets were computed after filtering out genes with average counts less than or equal to 1 or fewer than 2 samples with non-zero counts.

The analyses in this report used the following R packages to implement their corresponding methods: DESeq (version 1.8.3), edgeR (version 2.6.7) and NBPSeq (version 0.1.6). Code for implementing the TSPM method was taken from the website provided by Auer and Doerge (2011). Except where otherwise stated, the default settings for these packages were used during analysis.

Analysis results from the fly embryo data are summarized in the left side of Figure 2. For each method, we assigned p-values to bins of width 0.05 and used the number of p-values assigned to each bin to construct histogram curves. We applied

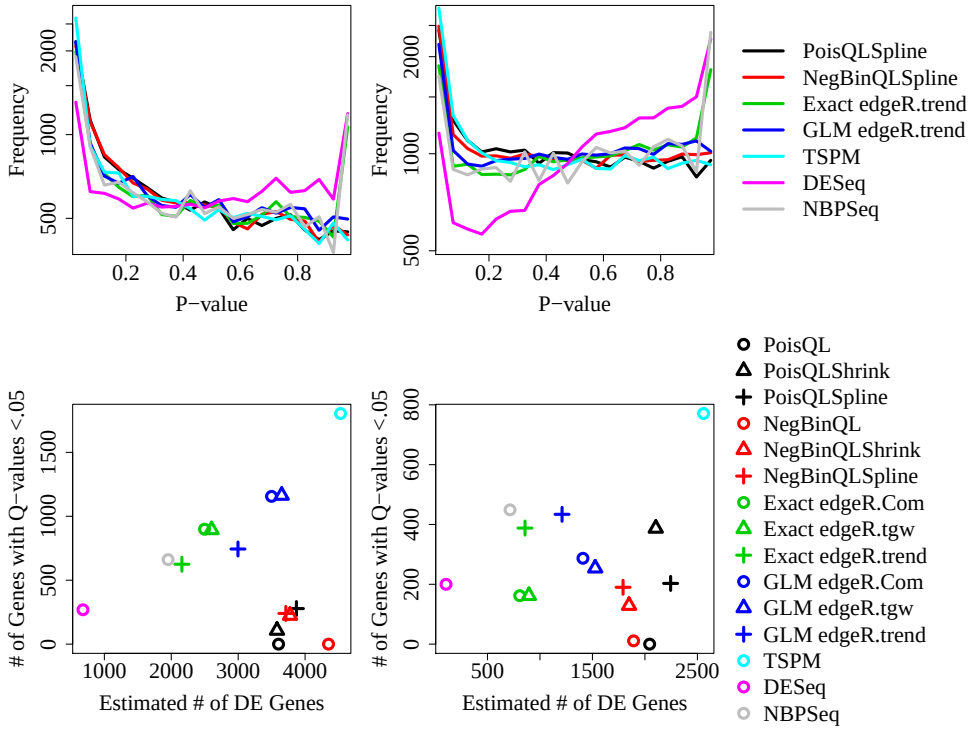

Figure 2: Histograms of p-values (top) and number of genes with q-values less than 0.05 versus estimated number of DE genes (bottom) for fly embryo (left) and Arabidopsis (right) data.

the method of Nettleton et al. (2006) to the distribution of p-values resulting from the application of each method in order to obtain q-values and estimates of the total number of DE genes. The methods produced drastically different estimates of the total number of DE genes (from 681 to 4530) and the number of genes with q-values less than .05 (from 0 to 1804). The p-value histograms for methods that used the exact test of Robinson and Smyth (2007) (i.e. exact edgeR, DESeq and NBPSeq) exhibited a spike for large p-values, which led to conservative estimates of the total number of DE genes.

The scatterplots in Figure 2 are primarily intended to display the large differences between the results from the considered methods. By themselves, these results do not provide sufficient information to evaluate each method. Generally speaking, the method with greatest power to detect differential expression is preferred, so long as the method allows researchers to accurately estimate or control false discovery rates. It is not possible to assess the performance of error rate control

Table 1: Overlap in methods' lists of top 200 genes for fly embryo (top) and Arabidopsis (bottom) data.

Method	1	2	3	4	5	6
PoisQLSpline(1)	200					
NegBinQLSpline(2)	189	200				
Exact edgeR.trend(3)	173	171	200			
GLM edgeR.trend(4)	177	177	183	200		
TSPM(5)	77	77	71	66	200	
DESeq(6)	168	161	149	150	81	200
NBPSeq(7)	153	151	158	165	50	133
Method	1	2	3	4	5	6
PoisQLSpline(1)	200					
NegBinQLSpline(2)	177	200				
Exact edgeR.trend(3)	158	160	200			
GLM edgeR.trend(4)	160	160	187	200		
TSPM(5)	25	26	14	15	200	
DESeq(6)	164	157	159	154	12	200
NBPSeq(7)	100	105	123	113	0	113

or estimation when the true status (EE or DE) of each analyzed gene is unknown, which is why we evaluate method performance through simulation studies.

In most cases when the goal of analyzing RNA-Seq data is to identify DE genes, resource constraints limit the number of genes that researchers will follow up with further study. Thus, a list of a fixed number of the most significant genes is a potentially important summary of the results of an analysis method. For the purpose of assessing similarities and differences among methods, the top half of Table 1 provides the size of pairwise intersections of lists containing the 200 most significant genes from each of seven methods.

3.2 Arabidopsis data set

We also examined the Arabidopsis data set provided as "arab" in the R package NBPSeq (Di et al., 2011). The data set includes count data for three biological replicates in each of two treatments in which leaves were inoculated with either a *Pseudomonas syringae* DC3000 mutant bacteria strain or a mock inoculant. The right side of Figure 1 contains a scatterplot of the estimated quasi-Poisson and

quasi-NegBin dispersions versus the average count for each gene for these data, along with the corresponding fitted cubic-splines used in the QLSpline methods. The data set contains 21185 genes with average counts greater than one and for which there were at least two samples with non-zero counts. We tested these genes for differential expression between two treatment conditions with the same methods used to analyze the fly embryo data set. Code and corresponding output for implementing the PoisQL, PoisQLShrink and PoisQLSpline methods for these data via the R (R Development Core Team, 2011) package QuasiSeq is shown in Section 6.1.1.

The right side of Figure 2 summarizes analysis results from the Arabidopsis data set when assuming a completely randomized experimental design (i.e. no replicate effects), as was done in Di et al. (2011). The methods produced drastically different estimates of the total number of DE genes (from 105 to 2559) and the number of genes with q-values less than 0.05 (from 0 to 771). The p-value histogram for DESeq was severely J-shaped, and NBPSeq and exact edgeR again exhibited a spike for large p-values, which led to conservative estimates of the total number of DE genes. The bottom half of Table 1 provides the size of pairwise intersections of lists containing the 200 most significant genes from each of seven methods.

Describing the experiment behind the Arabidopsis data set, Cumbie et al. (2011) writes, "Each treatment was done as biological triplicates with each pair of replicates done at separate times..." This description suggests that block effects should be included when analyzing these data, unless there is evidence that block effects are insignificant. The exact test of Robinson and Smyth (2007) examines differences between two levels of a common factor and does not accommodate nuisance factors, so the exact edgeR, NBPSeq and DESeq methods are unable to incorporate (or test for) block effects. The TSPM, GLM edgeR and QL methods are all built from GLMs and can accommodate nuisance factors by using an appropriate design matrix when estimating parameters.

When block effects are included in the model, estimating the variance for a gene in a reasonable manner requires having at least three total samples that have non-zero counts, with at least one of those samples coming from each treatment group. (Otherwise, the full model provides the same fitted values as the saturated model, so the variance is estimated to be essentially zero.) We analyzed the 20224 genes contained in the Arabidopsis data that met this criteria and that had an average count greater than one. Figure 3 provides estimates of the total number of non-null genes and the numbers of genes with q-values less than 0.05 resulting from tests for block and treatment effects, respectively, in the Arabidopsis data. These results provide strong evidence that block effects are present and that incorporating block effects significantly improves power to detect differential expression between treatments for these data.

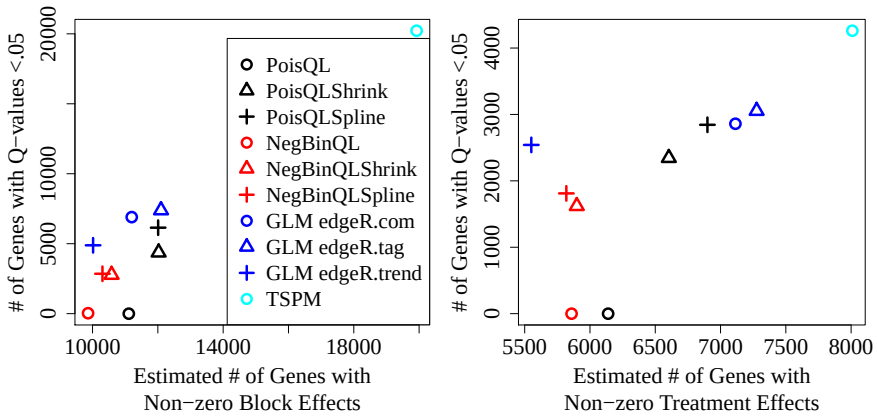

Figure 3: Number of genes with q-values less than 0.05 versus estimated number of non-null genes based on p-values testing for presence of block effects (left) and treatment effects (right) for Arabidopsis data.

4 Simulation Study

4.1 Simulation Descriptions

To examine the effectiveness of our suggested approach, we conducted a series of simulations for sample sizes of 4, 6 and 10, split evenly between two treatment groups. Simulated genes with average counts less than 1 were replaced with new simulated data before analyzing, as the count data for these genes contain little or no information about differential expression that can be detected with any method. Each simulation scenario was repeated 200 times, and each data set contained simulated counts for 1000 DE and 4000 EE genes. When analyzing simulated data, we set min.n=100 in 'estimateGLMTrendedDisp' in order to provide more points for edgeR to use when identifying a trend between the negative binomial dispersion estimates and average simulated counts.

4.1.1 Negative Binomial Simulations

We simulated negative binomial data using parameters guided by sample averages and dispersion estimates from the fly embryo and Arabidopsis data sets. For the fly embryo and Arabidopsis data sets, let $\bar{y}_{\cdot\cdot k}$ denote the sample average of the four and six observations, respectively, from gene k. For simulations based on the fly embryo data set, let $\hat{\omega}_k$ denote the estimated dispersion parameter for the negative

binomial variance function $\left(\text{Var}(Y_{ijk}) = \mu_{ijk} + \omega_k \mu_{ijk}^2\right)$ obtained for gene k from the edgeR exact test tagwise dispersion estimation procedure with the trend option and a prior.n specification of 1. For simulations based on the Arabidopsis data set, let $\hat{\omega}_k$ denote the estimated negative binomial dispersion parameter for gene k obtained by fitting a model with both treatment and block effects via the edgeR GLM trended tagwise dispersion estimation procedure with a prior.n specification of 1. Figure 4 displays plots of $\hat{\omega}_k$ versus $\bar{y}_{\cdot\cdot k}$ that were used in these simulations.

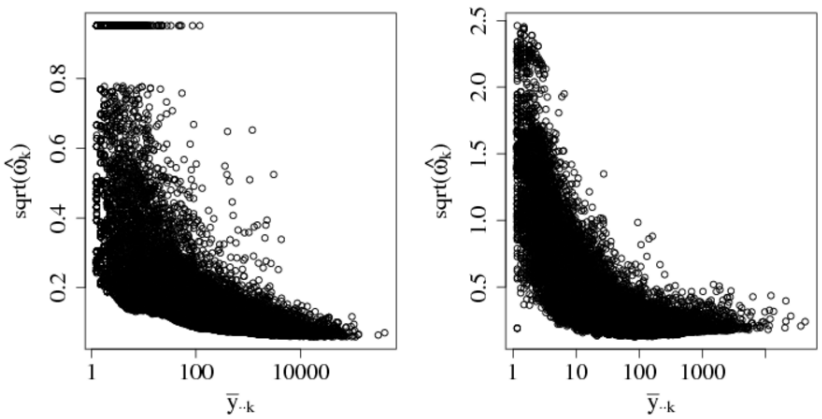

Figure 4: Sample averages and negative binomial dispersion parameter estimates used for simulations based on fly embryo (left) and Arabidopsis (right) data.

Data were simulated from negative binomial distributions for the kth gene in the following manner. Let k' index a gene randomly selected from the real data set. If the kth simulated gene was to be EE, we let $\lambda_{ik} = \bar{y}_{\cdot\cdot k'}$ for $i = 1, 2$. If the kth simulated gene was to be DE, for simulations based on the fly embryo data, we sampled a fold change factor, B_k, in the following manner. We set $B_k = B_{k1} + B_{k2}$, where B_{k1} was sampled from an inverse-gamma distribution with rate parameter 1 and shape parameter $S\bar{y}_{\cdot\cdot k'}^{1/8}$ and B_{k2} was sampled from a uniform distribution with endpoints L and U. (Values for L and U are provided in Table 2. We adjusted the severity of simulated fold changes to maintain moderate separation of EE and DE genes by using $S = 1.25, 1.5, 2$ for $n = 4, 6, 10$, respectively.) For simulations based on the Arabidopsis data, B_{k1} was sampled from an inverse-gamma distribution with rate parameter 1 and shape parameter $S\log\left(\bar{y}_{\cdot\cdot k'}\right)^{1/8}$. Small and large expression levels of $\bar{y}_{\cdot\cdot k'}/\sqrt{B_k}$ and $\bar{y}_{\cdot\cdot k'}\sqrt{B_k} + 5$, respectively, were randomly assigned between λ_{1k} and λ_{2k}. Library size factors were simulated according to $\log_2 c_{ij} \sim \text{Normal}(0, 0.125^2)$, where c_{ij} is the simulated library size factor for replicate j in treatment i. Final

counts were simulated from a negative binomial distribution with mean $\mu_{ijk} = \lambda_{ik} c_{ij}$ and variance $\mu_{ijk} + \hat{\omega}_{k'} \mu_{ijk}^2$.

The techniques for simulating fold changes were chosen to reproduce the relationship between estimated fold change and average count seen in the fly embryo and Arabidopsis data for the $n = 4$ and $n = 6$ simulations, respectively. (See Figure 5.)

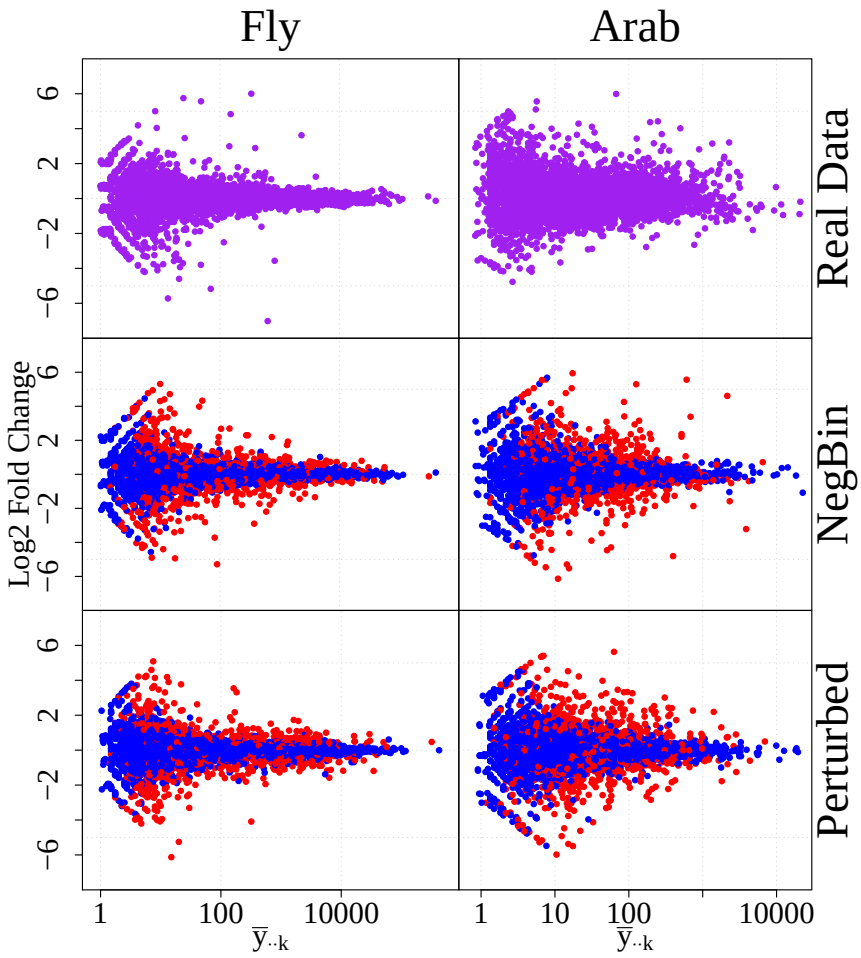

Figure 5: Estimated log fold change versus log average count for actual (top), negative binomial simulated (middle) and perturbed simulated (bottom) data from fly embryo (left, $n = 4$) and Arabidopsis (right, $n = 6$) data sets. For simulated data sets, DE and EE genes are marked with red and blue dots, respectively, and library size factors were simulated according to $\log_2 c_{ij} \sim \text{Normal}(0, 0.125^2)$.

Table 2: Parameters used to simulate fold changes.

Model	Data Set	L	U
NegBin	Fly	0.5	1
NegBin	Arab	0.25	0.75
Perturbed	Fly	0.25	0.75
Perturbed	Arab	0.5	1

Variability in library size among samples can strongly impact method performance. In particular, as variability in library size among samples increases, models based on the wrong mean-variance relationship will suffer. To understand this point, consider a single gene that is EE (because p-values are based on the modeled behavior of counts when the null hypothesis is true). If all library sizes were the same, and if there were no confounding variables, then the modeled means would all be the same across samples. Regardless of what function of the mean is chosen, the modeled variances would also be constant across samples. In this case it does not matter whether the mean-variance relationship is modeled as linear (overdispersed Poisson) or quadratic (negative binomial) because a line and a parabola both provide adequate flexibility to intersect a single point on the mean-variance plane. As library size variability increases, the range of modeled means for a single gene becomes wider, and the functional form of the mean-variance relationship becomes more important. For this reason, we repeat the negative binomial simulations simulating library size factors according to $\log_2 c_{ij} \sim \text{Normal}(0,1)$. These simulations are referred to as "extreme NegBin" and provide an assessment of performance under a scenario with somewhat extreme variation in library size.

To examine method sensitivity to the data-generating model, we also simulated data from slight perturbations of negative binomial distributions using parameters guided by sample averages and dispersion estimates from the fly embryo and Arabidopsis data sets. These simulations began by sampling a mean and dispersion pair from the real data set $(\bar{y}_{\cdot\cdot k'}, \hat{\omega}_k)$, using library size factors simulated according to $\log_2 c_{ij} \sim \text{Normal}(0,1)$ and, for DE genes, generating a fold change factor, B_k, in exactly the same way as was done in the negative binomial simulations, using the parameter values given in Table 2. Let $\lambda'_{ijk} = \bar{y}_{\cdot\cdot k'} c_{ij}$ if gene k was simulated as EE and let $\lambda'_{ijk} = \bar{y}_{\cdot\cdot k'} c_{ij} / \sqrt{B_k}$ (or $\bar{y}_{\cdot\cdot k'} c_{ij} \sqrt{B_k} + 5$) if gene k was simulated as DE.

To modify the data-generating model, we generated a perturbation effect, $\varsigma_k \sim \text{Normal}(0,\ 0.1)$, and simulated means λ_{ijk} from a gamma distribution with shape parameter ${\lambda'_{ijk}}^{\varsigma_k} / \hat{\omega}_k$ and rate parameter ${\lambda'_{ijk}}^{\varsigma_k - 1} / \hat{\omega}_k$. Final counts were sim-

ulated as $Y_{ijk}^{sim}|\lambda_{ijk} \sim \text{Poisson}(\lambda_{ijk})$. The final counts have conditional mean and variance $\text{E}(Y_{ijk}^{sim}|\lambda'_{ijk}, \varsigma_k) = \lambda'_{ijk}$ and $\text{Var}(Y_{ijk}^{sim}|\lambda'_{ijk}, \varsigma_k) = \lambda'_{ijk} + \hat{\omega}_k {\lambda'_{ijk}}^{2-\varsigma_k}$, which is a slight variation from the mean-variance relationship of the negative binomial distribution. We refer to these simulations as "extreme perturbed."

4.2 Simulation Results

We evaluated each method's performance according to two criteria: separation of DE and EE genes in significance rankings as seen in discovery versus false discovery curves and uniformity of the empirical distribution of p-values coming from EE genes. We also observed the effect that non-uniform null p-value distributions can have on estimated false discovery rates by comparing empirical FDRs (eFDR) to q-values. We report simulation results through a combination of plots and tables. The plotted curves describe average behavior over 200 iterations for each simulation scenario. For each curve, solid thin lines located $\pm$ two standard errors around the mean are also included, providing approximate 95% pointwise confidence intervals, although most of the standard error lines have merged with their corresponding mean line.

We began our simulation study with every method whose results are reported for the fly embryo and Arabidopsis data sets. To control the number of results to report and to increase the speed of conducting simulations, we kept only the best performing methods from each of the following four classes: Poisson QL, negative binomial QL, GLM edgeR, and exact edgeR. Across most scenarios, the QLSpline method exhibited the best performance of the quasi-Poisson methods. Under a quasi-negative binomial model, the QLShrink and QLSpline methods performed similarly well. This was not surprising because only a slight relationship was present between quasi-likelihood dispersion estimates, $\hat{\Phi}_k$, and average counts for the quasi-NegBin model. We chose to include the QLSpline approach. The trend implementations of the exact test and GLM versions of edgeR generally outperformed their constant dispersion and non-trend tagwise dispersion counterparts. We also included results from TSPM, DESeq and NBPSeq.

The solid curves in Figure 6 display curves relating number of false discoveries to total number of discoveries for the $n = 6$ simulations. Plots for $n = 4$ and $n = 10$ show similar qualitative traits and are provided as Figures 16 and 17, respectively, in Section 6. It is difficult to assess relative performance from these plots, although it is clear that the top five methods are GLM edgeR.trend, exact edgeR.trend, DESeq, PoisQLSpline, and NegBinQLSpline. To better examine the relative performance among the top five methods for each simulation scenario, we subtract the average number of discoveries (across 200 simulation iterations) for

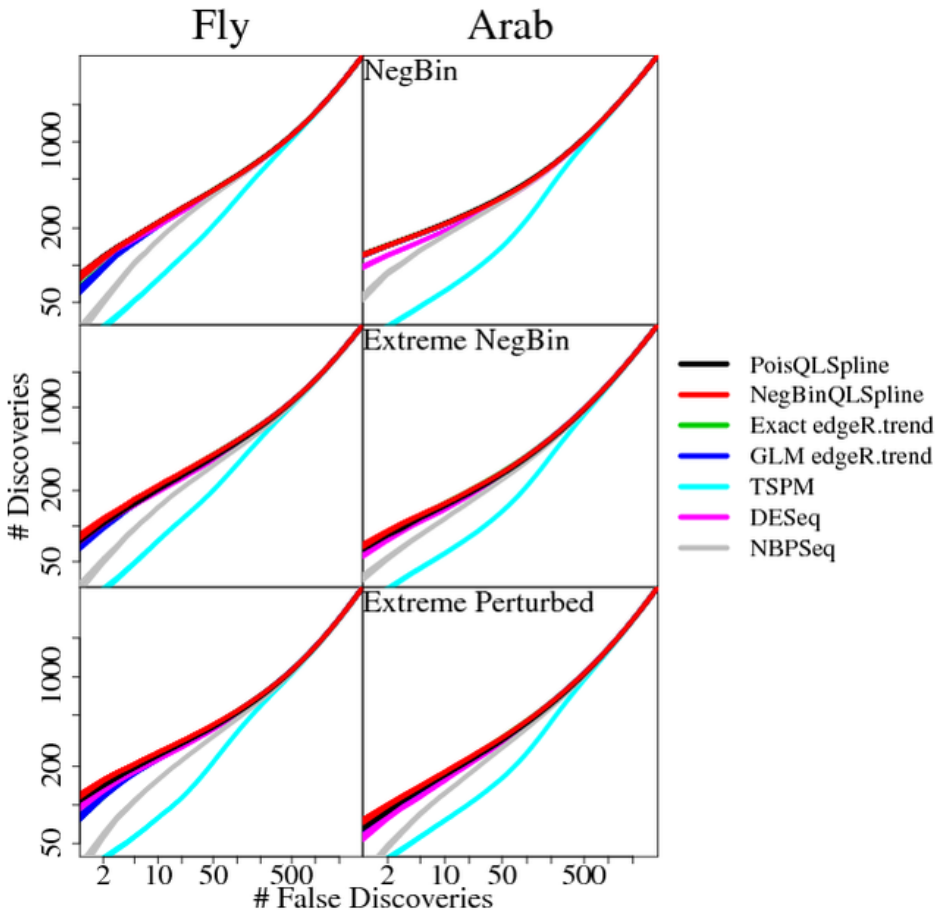

Figure 6: Curves relating average number of total discoveries to average number of false discoveries for negative binomial (top) and perturbed NegBin (bottom) simulations based on fly embryo (left) and Arabidopsis (right) data sets with $n = 6$.

the PoisQLSpline method from the curve for each of the top five methods and plot the differences in Figures 7 through 9. Figure 7 shows that PoisQLSpline and NegBinQLSpline provided the best significance rankings among the most significant genes in the simulations with moderate differences between library size factors. For simulation scenarios using extreme differences between library sizes, NegBinQLSpline either closely followed the exact and GLM edgeR.trend methods or outperformed them for small (< 20) numbers of false discoveries. As an example, in the $n = 10$ extreme perturbed simulations based on the fly embryo data, NegBinQLSpline identified between 25 and 50 more true positives than the non-QL methods

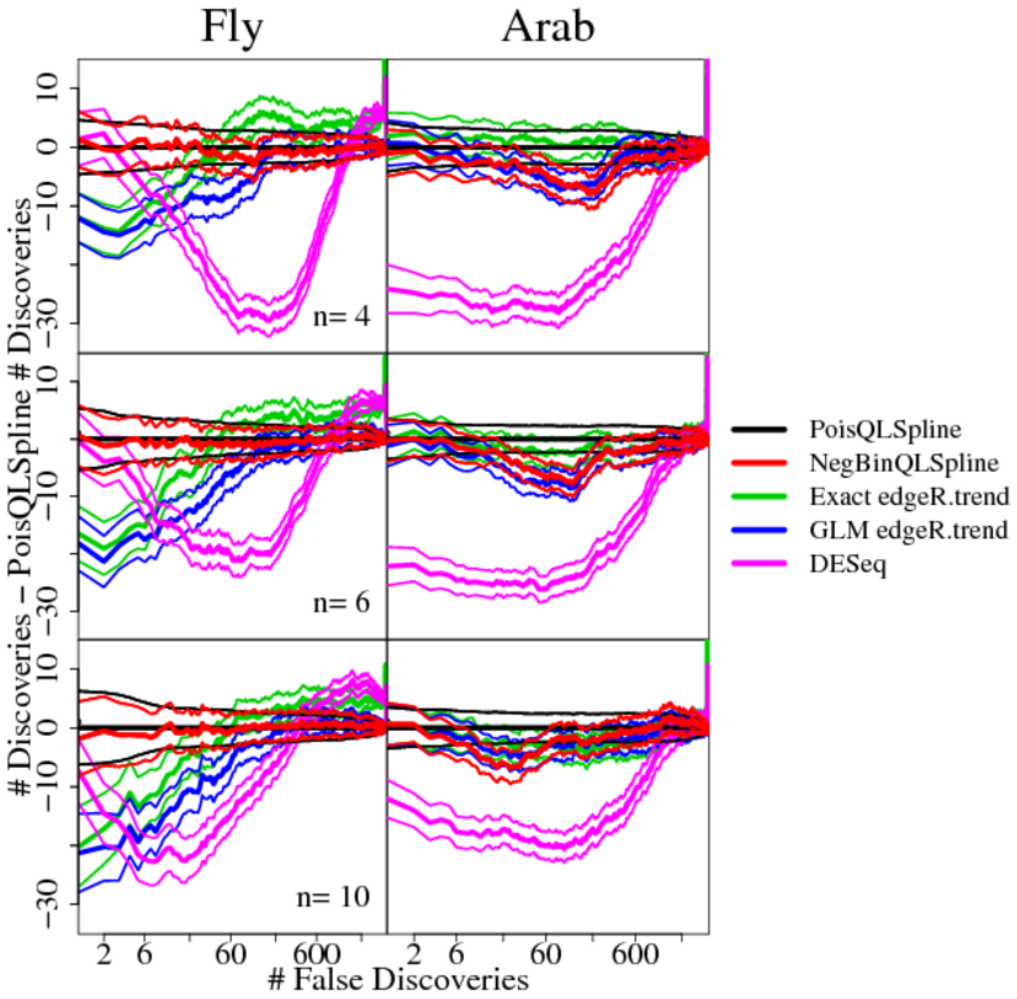

Figure 7: Curves relating difference in average number of total discoveries to average number of false discoveries for negative binomial simulations based on fly embryo (left) and Arabidopsis (right) data sets with $n = 4$ (top), $n = 6$ (middle) and $n = 10$ (bottom).

over a range of 0 to 10 false discoveries. Curves for PoisQLSpline and DESeq were generally lower than the other three methods in simulations using extreme library size differences.

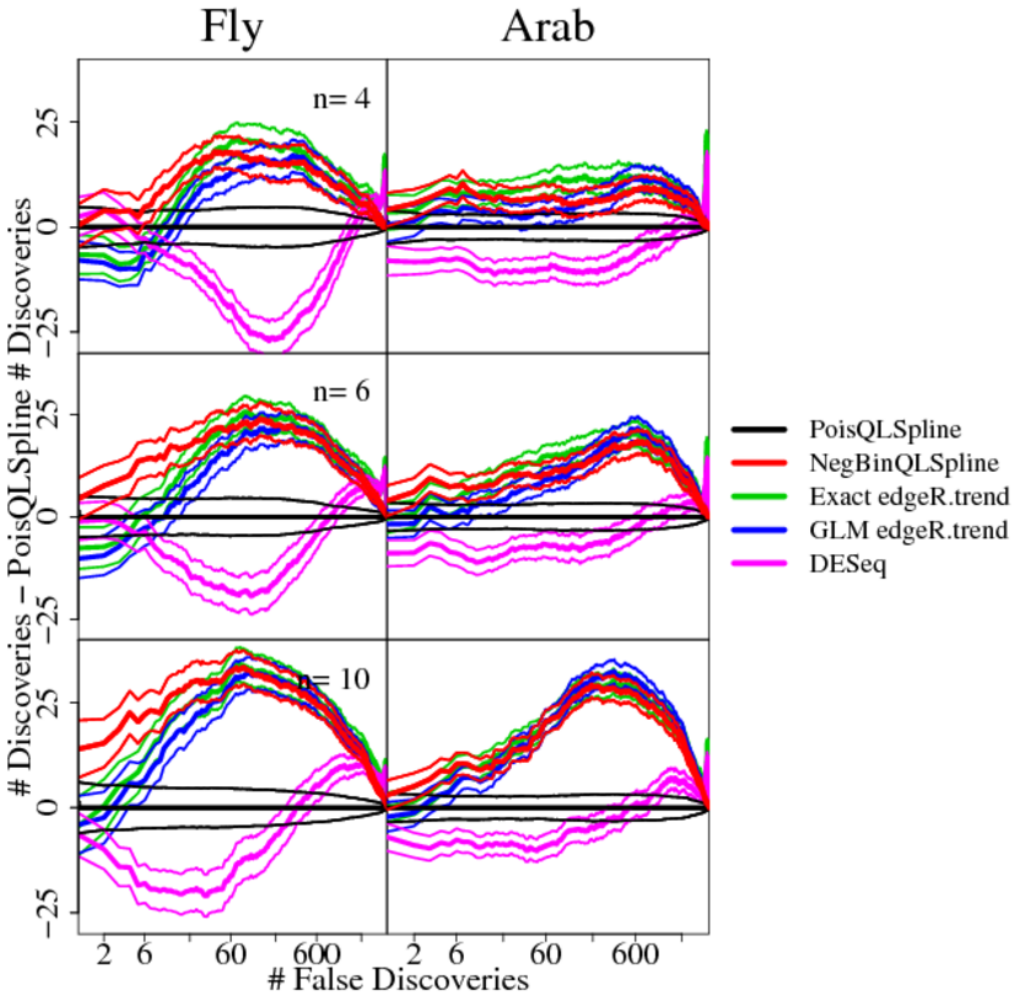

Figure 8: Curves relating difference in average number of total discoveries to average number of false discoveries for extreme NegBin simulations based on fly embryo (left) and Arabidopsis (right) data sets with $n = 4$ (top), $n = 6$ (middle) and $n = 10$ (bottom).

Improved significance rankings lead to fewer false positives (and more true positives) appearing on a list containing a fixed number of genes. This is important as resource constraints limit the number of genes that researchers can follow up on in future studies. To facilitate a direct comparison among the methods, the average number of DE genes in the 200 most significant genes for each method are

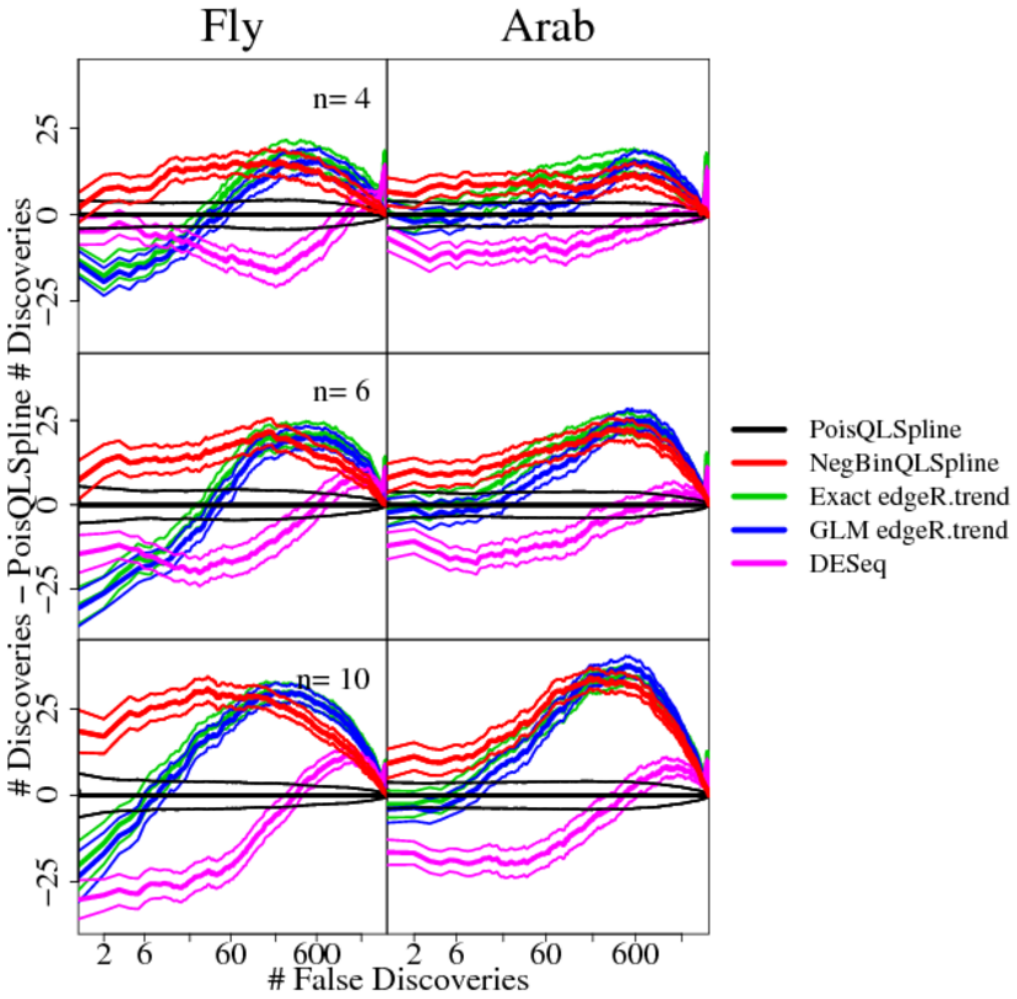

Figure 9: Curves relating difference in average number of total discoveries to average number of false discoveries for extreme perturbed simulations based on fly embryo (left) and Arabidopsis (right) data sets with $n = 4$ (top), $n = 6$ (middle) and $n = 10$ (bottom).

provided in Tables 3-8. These numbers are useful for putting the power and sensitivity of the methods into a practical perspective. In the $n = 4$ extreme perturbed simulations based on the fly embryo data set, for example, PoisQLSpline and NegBinQLSpline averaged 185.4 and 187.8 DE genes, respectively, while the closest competitor, exact edgeR.trend, averaged 184.6 DE genes in their respective lists of

200 most significant genes. For simulations with moderate library size differences, NegBinQLSpline and PoisQLSpline produced the fewest false positives. For simulations with extreme library size differences, NegBinQLSpline produced the fewest false positives. In general, the significance rankings produced by NegBinQLSpline were as good as or better than those produced by each of the non-QL methods in each simulation scenario.

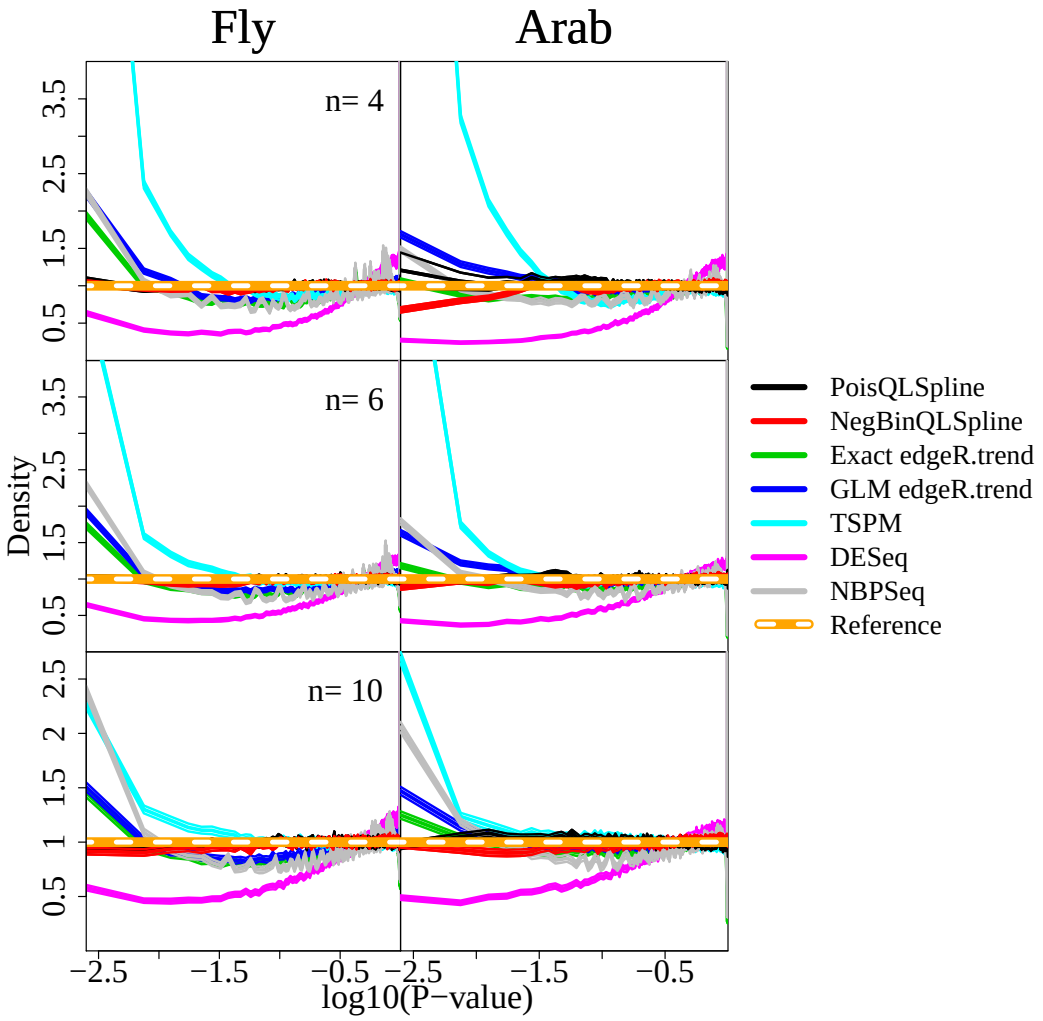

Figure 10: Histograms of p-values for EE genes in negative binomial simulations based on fly embryo (left) and Arabidopsis (right) data sets with $n = 4$ (top), $n = 6$ (middle) and $n = 10$ (bottom).

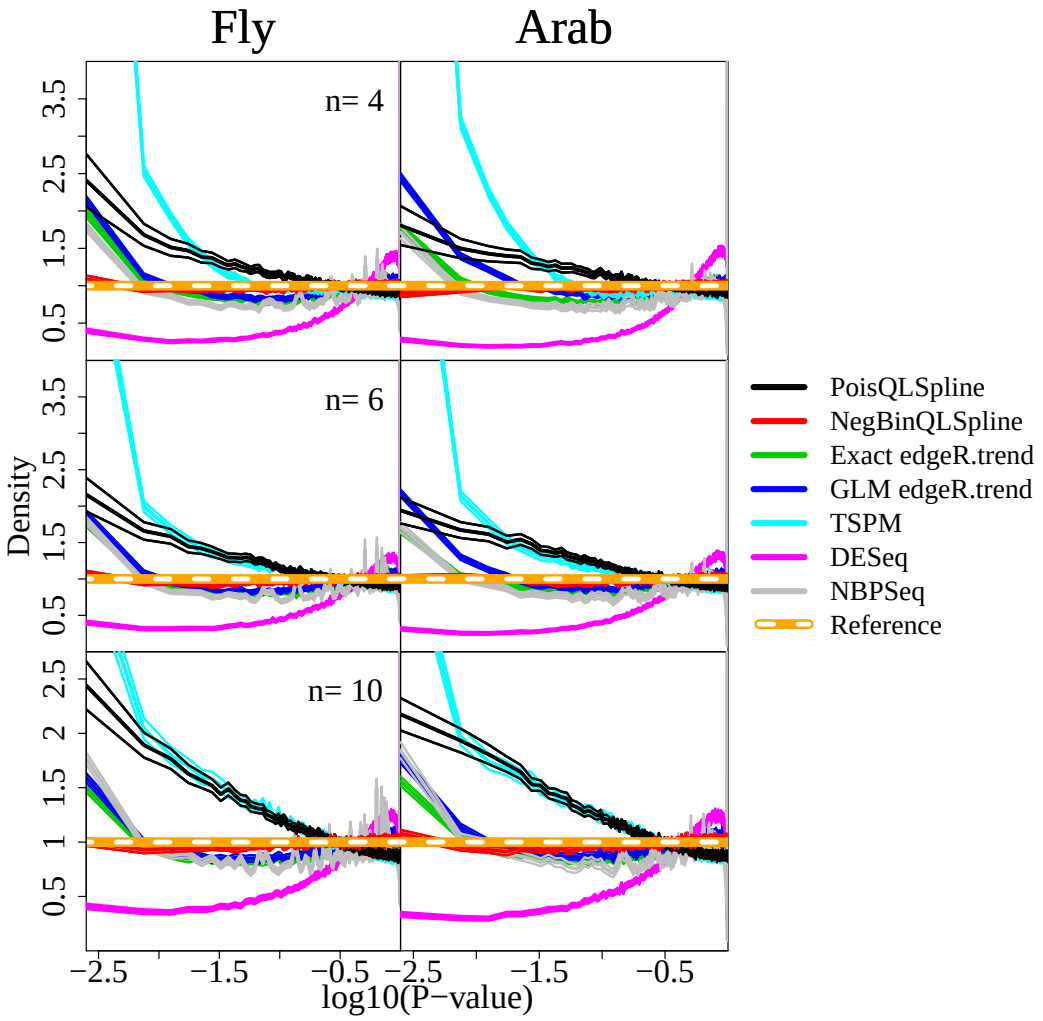

Figure 11: Histograms of p-values for EE genes in extreme NegBin simulations based on fly embryo (left) and Arabidopsis (right) data sets with $n = 4$ (top), $n = 6$ (middle) and $n = 10$ (bottom).

We next examine the distribution of p-values for simulated EE genes. For each method in each simulation, p-values from the 4000 EE genes were assigned to bins of width 0.005, and the number of p-values assigned to each bin was recorded. Figures 10 through 12 display histogram curves, providing the average density of p-values assigned to each bin. The dashed orange line provides a reference for comparison with the uniform distribution. For the purposes of estimating false

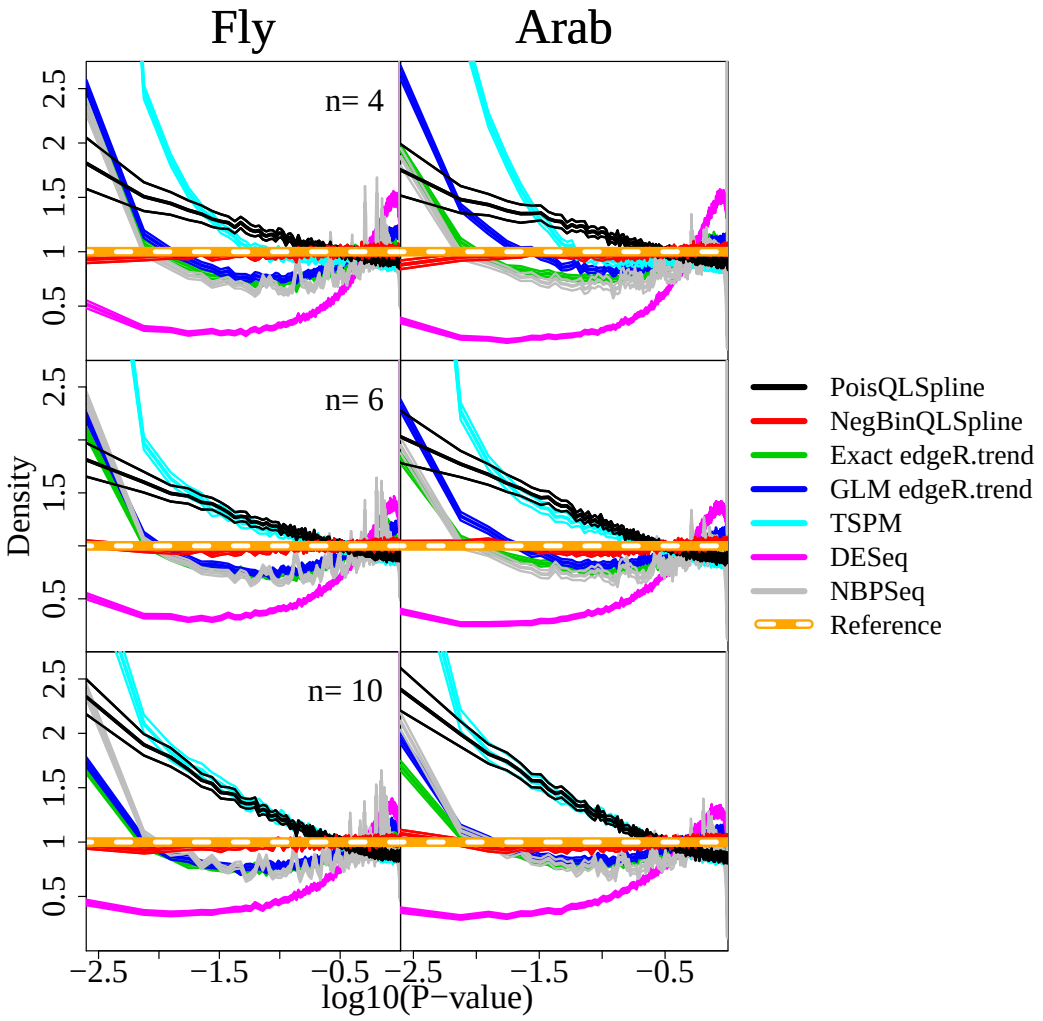

Figure 12: Histograms of p-values for EE genes in extreme perturbed simulations based on fly embryo (left) and Arabidopsis (right) data sets with $n = 4$ (top), $n = 6$ (middle) and $n = 10$ (bottom).

discovery rates, the most influential deviation from uniformity occurs when there are too many small p-values. These plots display the p-value axis on a log-scale in order to focus on the distribution of null p-values between 0 and 0.1.

The TSPM, NBPSeq, GLM edgeR.trend, and exact edgeR.trend methods display an over-abundance of small p-values relative to a uniform distribution in all simulation scenarios. This can be explained by the fact that the edgeR and NBPSeq

methods do not account for uncertainty in their negative binomial dispersion parameter estimates, and the TSPM method uses a Poisson-based approach that only adjusts for overdispersion for genes in which overdispersion is found to be statistically significant. Although DESeq also fails to account for uncertainty in its negative binomial dispersion parameter estimates, it generally produced strongly conservative results (i.e. small p-values are under-represented in the distributions of null p-values from DESeq). DESeq computes gene-specific negative binomial dispersion estimates in addition to fitting a trend to the relationship between dispersion and average count. As a final dispersion estimate for each gene, DESeq uses the maximum between the original estimated dispersion and the corresponding point on the trend. This approach is designed to avoid underestimating negative binomial dispersions and explains why DESeq consistently produced strongly conservative results even when failing to account for uncertainty in the estimated dispersion. Null p-values from PoisQLSpline were roughly uniformly distributed for simulations with moderate library size differences, but displayed a severe overabundance of small p-values in simulations with extreme library size differences. The distribution of null p-values from NegBinQLSpline closely matched the uniform distribution for all simulations.

A surplus of very small (<0.005) p-values can drastically affect false discovery rate estimates. As a demonstration, we compare empirical false discovery rates (eFDR) to q-values. The eFDR of gene k reports the proportion of genes that were EE from the set of genes that have p-values as small as or smaller than the p-value of gene k. Q-values are obtained by applying the method of Nettleton et al. (2006) to the distribution of p-values resulting from the application of each method. In this section we refer to methods as being liberal or conservative when their distributions of null p-values lead to q-values that underestimate or overestimate FDRs, respectively. It should be noted that R packages for many competing methods include an approach, such as the Benjamini and Hochberg procedure, to control, rather than estimate, FDRs. We are not investigating the performance of FDR control approaches from each package, but examining the impact of non-uniform null p-values on q-values. In this sense, if a method is neither conservative nor liberal, then the q-value for any given gene should closely match its eFDR. For example, if the gene with the Mth smallest p-value has a corresponding q-value of .05, then roughly 5% of the M genes with p-values as small or smaller should be EE.

To examine if this characteristic held for each method, we plotted average eFDRs versus q-values for each scenario. The solid curves in Figures 13 through 15 display curves from the negative binomial, extreme NegBin and extreme perturbed simulations, respectively. To construct these plots, we rounded each q-value to the nearest 0.001 before plotting. When multiple genes produced identical rounded q-

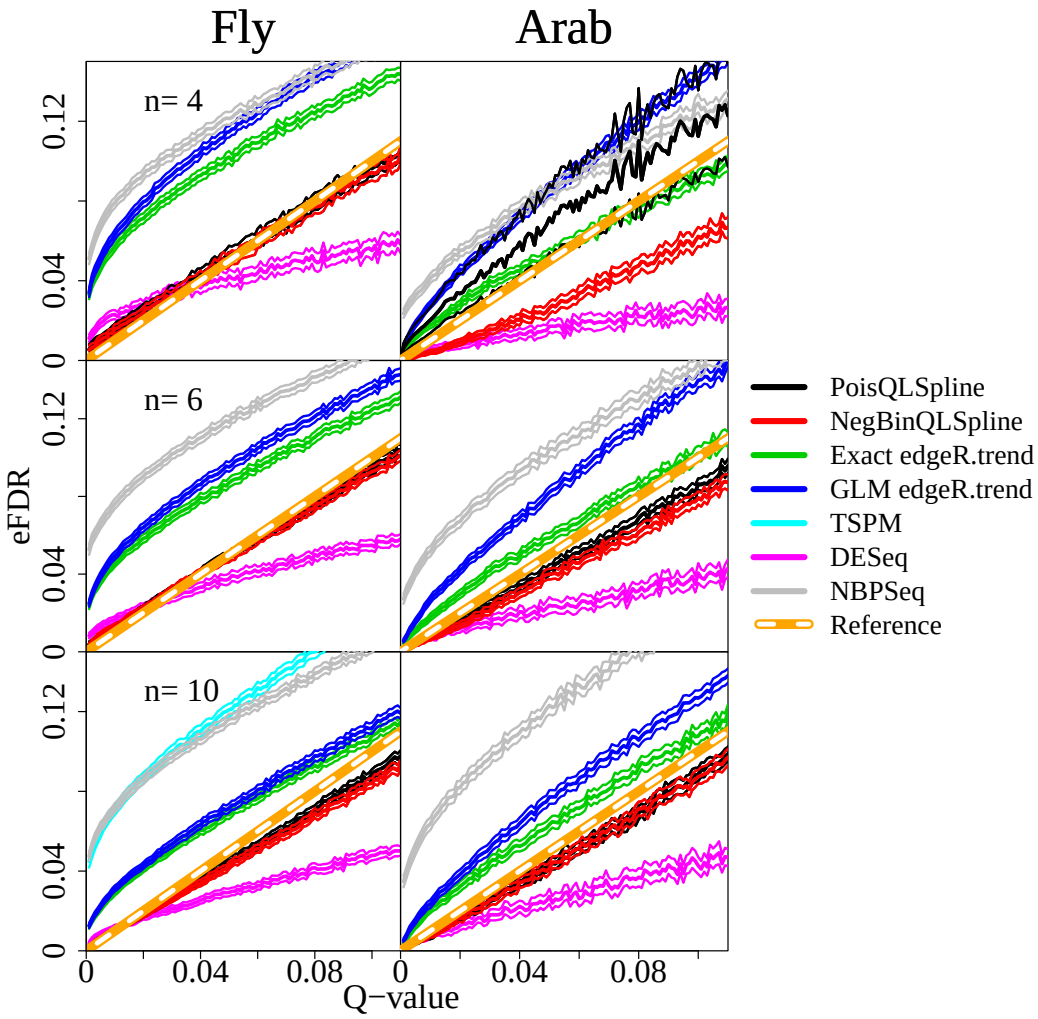

Figure 13: Curves relating average eFDR to q-values for negative binomial simulations based on fly embryo (left) and Arabidopsis (right) data sets with $n = 4$ (top), $n = 6$ (middle) and $n = 10$ (bottom).

values for a given method, the eFDR of the gene with the largest original p-value was used to represent the set. (This technique facilitated averaging eFDRs across simulations and computing standard errors at each rounded q-value.) If a method was neither conservative nor liberal, its line should closely follow the dashed orange $y = x$ diagonal. Lines appearing substantially above or below the diagonal indicate the corresponding method was liberal or conservative, respectively.

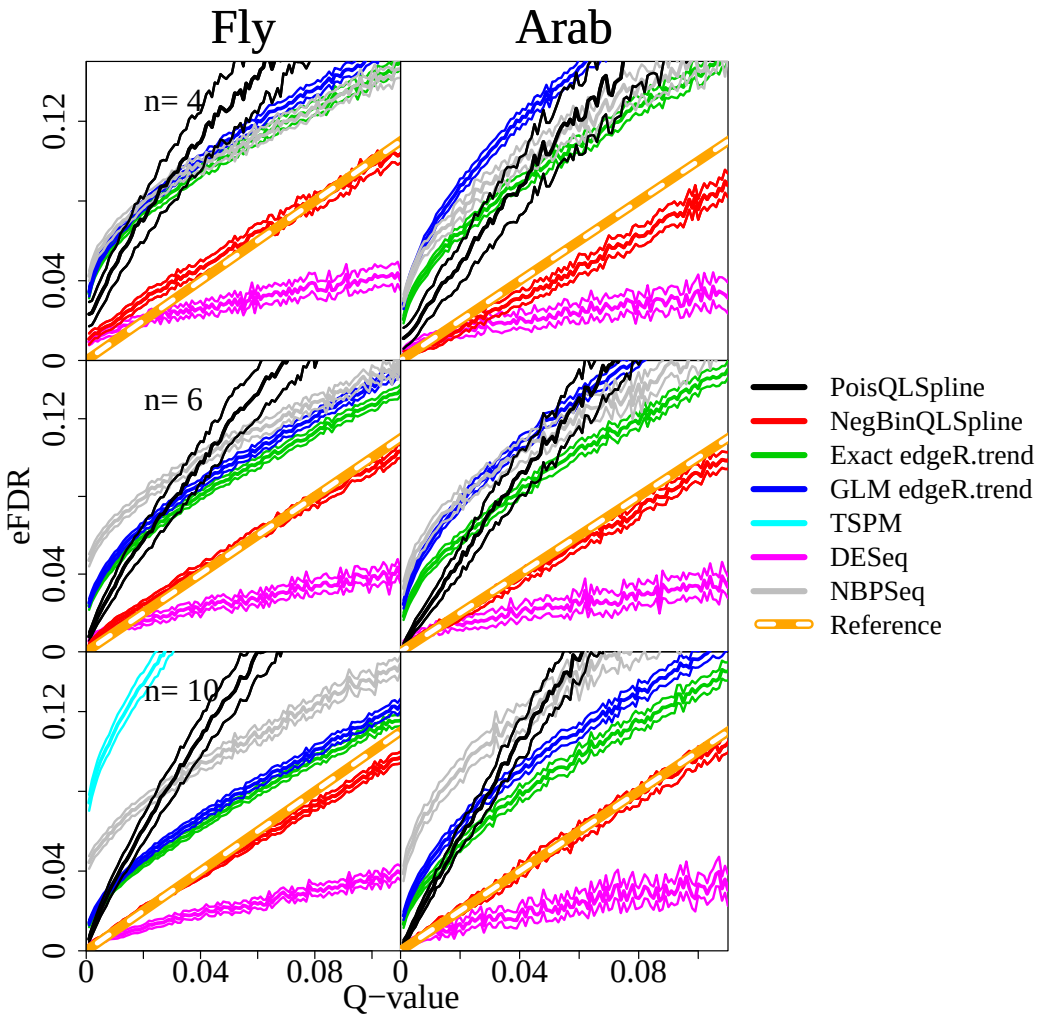

Figure 14: Curves relating average eFDR to q-values for extreme NegBin simulations based on fly embryo (left) and Arabidopsis (right) data sets with $n = 4$ (top), $n = 6$ (middle) and $n = 10$ (bottom).

The average eFDR curves for the TSPM, NBPSeq, exact edgeR.trend, and GLM edgeR.trend methods are substantially above the dotted orange $y = x$ diagonal in every simulation scenario, indicating these methods produced liberal results for these data. DESeq was strongly conservative in these simulations. In simulations with moderate library size differences, PoisQLSpline produced accurate q-values. In simulations with extreme library size differences, PoisQLSpline

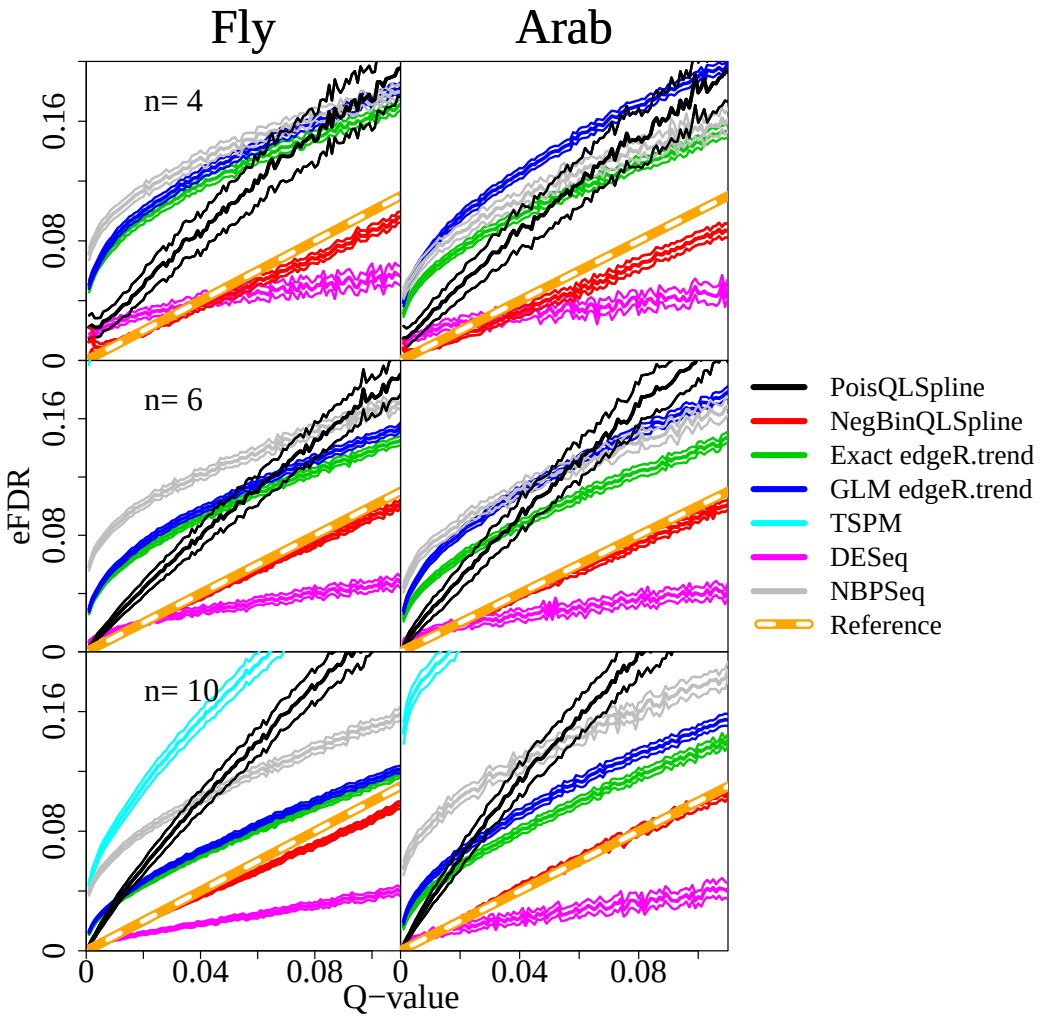

Figure 15: Curves relating average eFDR to q-values for extreme perturbed simulations based on fly embryo (left) and Arabidopsis (right) data sets with $n = 4$ (top), $n = 6$ (middle) and $n = 10$ (bottom).

produced severely liberal q-values. Q-values for NegBinQLSpline generally were accurate or moderately conservative across the simuations.

The average eFDR with a corresponding q-value of 0.05 for each method are provided in Tables 3 through 8. Average eFDRs for DESeq and NegBinQLSpline were most often contained in (0.02, 0.03) and (0.04, 0.05), respectively. Average eFDRs for other methods were often substantially greater than 0.05. In the $n =$

Table 3: Summary of simulation results for negative binomial fly embryo simulations. Legend $\sim$ # DE Top 200: Number of truly DE genes contained in list of 200 most significant genes; $\text{eFDR}_{Q<.05}$: empirical FDR for list of all genes with q-values less than .05; $N_{Q<.05}$: Number of genes with q-values less than .05; $\hat{N}_{DE}$: Estimated number of DE genes; Max SE: Maximum standard error of averages.

Method	# DE Top 200	$\text{eFDR}_{Q<.05}$	$N_{Q<.05}$	$\hat{N}_{DE}$
	$n = 4$			
PoisQLSpline	189.6	0.0523	198.3	689
NegBinQLSpline	189.7	0.0495	194.6	652
Exact edgeR.trend	188.6$^{\star\circ}$	0.101	306.8	455
GLM edgeR.trend	188$^{\star\circ}$	0.114	323.1	506
TSPM	141.9$^{\star\circ}$	0.453	490.8	789
DESeq	188.1$^{\star\circ}$	0.0402	159.9	187
NBPSeq	184.3$^{\star\circ}$	0.122	307.4	445
Max SE	0.43	0.00176	2.4	7.71
	$n = 6$			
PoisQLSpline	192.9	0.0481	241.3	683
NegBinQLSpline	192.7	0.0479	238.1	647
Exact edgeR.trend	191.7$^{\star\circ}$	0.0854	328.4	483
GLM edgeR.trend	191.1$^{\star\circ}$	0.0956	335.4	520
TSPM	153.8$^{\star\circ}$	0.278	316.8	774
DESeq	191.3$^{\star\circ}$	0.0369	186.3	259
NBPSeq	185.7$^{\star\circ}$	0.119	331.9	478
Max SE	0.4	0.00183	1.6	6.77
	$n = 10$			
PoisQLSpline	197	0.044	317.5	698
NegBinQLSpline	197	0.041	311.1	660
Exact edgeR.trend	196.3$^{\star\circ}$	0.0684	373.8	518
GLM edgeR.trend	196$^{\star\circ}$	0.0707	373.1	539
TSPM	184.4$^{\star\circ}$	0.12	315.1	769
DESeq	195.8$^{\star\circ}$	0.0291	234.6	331
NBPSeq	188.4$^{\star\circ}$	0.115	380	516
Max SE	0.26	0.00158	1.6	5.21

$^{\circ}$ paired t-test comparing reported average to that of PoisQLSpline yielded two-sided p-value$<$0.01

* paired t-test comparing reported average to that of NegBinQLSpline yielded two-sided p-value$<$0.01

4 extreme perturbed simulations based on the fly embryo data set, for example, TSPM, NBPSeq, and both edgeR methods all had average eFDRs greater than 0.12.

Table 4: Summary of simulation results for negative binomial Arabidopsis simulations. See Table 3 for legend.

Method	# DE Top 200	eFDR$_{Q<.05}$	$N_{Q<.05}$	$\hat{N}_{DE}$
	$n=4$			
PoisQLSpline	185.9	0.0375	146.8	608
NegBinQLSpline	185.6	0.033	138.9	509
Exact edgeR.trend	185.8	0.0955	235.4	347
GLM edgeR.trend	184.9$^{\star\circ}$	0.129	272.7	455
TSPM	141.3$^{\star\circ}$	0.536	503.3	776
DESeq	182.7$^{\star\circ}$	0.0275	115.8	81.2
NBPSeq	177.9$^{\star\circ}$	0.132	233.7	353
Max SE	0.44	0.00182	1.9	7.96
	$n=6$			
PoisQLSpline	186.2*	0.0419	157.9	616
NegBinQLSpline	185.6°	0.042	157.1	491
Exact edgeR.trend	185.1$^{\star\circ}$	0.0909	226.7	363
GLM edgeR.trend	184.4$^{\star\circ}$	0.116	252.9	433
TSPM	124.9$^{\star\circ}$	0.412	269.6	786
DESeq	183.4$^{\star\circ}$	0.0308	123.3	132
NBPSeq	175.3$^{\star\circ}$	0.149	239	371
Max SE	0.5	0.00258	1.5	7.55
	$n=10$			
PoisQLSpline	184.6*	0.0421	147.7	634
NegBinQLSpline	183.5°	0.0455	148.4	491
Exact edgeR.trend	183.8°	0.0843	205.2	372
GLM edgeR.trend	182.9$^{\star\circ}$	0.0987	220	419
TSPM	153.3$^{\star\circ}$	0.217	164	727
DESeq	182$^{\star\circ}$	0.0294	110.9	158
NBPSeq	170.5$^{\star\circ}$	0.17	232.4	380
Max SE	0.39	0.00241	1.3	6.24

$^{\circ}$ paired t-test comparing reported average to that of PoisQLSpline yielded two-sided p-value$<$0.01

* paired t-test comparing reported average to that of NegBinQLSpline yielded two-sided p-value$<$0.01

To produce the most accurate q-values, we recommend using p-values obtained from NegBinQLSpline.

Interestingly, although many of the negative binomial modeling methods had liberal eFDRs compared to their q-values, they all underestimated the number of DE genes (1000) in every simulation scenario. DESeq was most conservative in

Table 5: Summary of simulation results for extreme NegBin simulations based on fly embryo data. See Table 3 for legend.

Method	# DE Top 200	eFDR$_{Q<.05}$	$N_{Q<.05}$	$\hat{N}_{DE}$
	$n=4$			
PoisQLSpline	187.4*	0.125	305.4	1040
NegBinQLSpline	189.3°	0.0544	201.9	617
Exact edgeR.trend	188.3$^{\star\circ}$	0.102	308.8	470
GLM edgeR.trend	187.6*	0.11	314.1	504
TSPM	141.2$^{\star\circ}$	0.484	578.6	1140
DESeq	186.4$^{\star\circ}$	0.0285	119.6	111
NBPSeq	182.1$^{\star\circ}$	0.105	238.5	374
Max SE	0.42	0.00817	13.1	29.7
	$n=6$			
PoisQLSpline	190.4*	0.116	329.9	1080
NegBinQLSpline	192.2°	0.0499	238.5	619
Exact edgeR.trend	191.2$^{\star\circ}$	0.0884	323.7	484
GLM edgeR.trend	190.6*	0.0954	329.3	514
TSPM	150.1$^{\star\circ}$	0.32	395.8	1160
DESeq	189.2$^{\star\circ}$	0.0252	131.7	168
NBPSeq	181$^{\star\circ}$	0.108	237.5	391
Max SE	0.44	0.00598	9.8	26
	$n=10$			
PoisQLSpline	194.7*	0.126	426	1190
NegBinQLSpline	196.2°	0.0456	307.3	648
Exact edgeR.trend	195.7$^{\star\circ}$	0.0705	364	516
GLM edgeR.trend	195.4$^{\star\circ}$	0.0758	369.7	536
TSPM	179.9$^{\star\circ}$	0.19	415.4	1260
DESeq	192.8$^{\star\circ}$	0.0218	159.3	238
NBPSeq	183.5$^{\star\circ}$	0.1	253	418
Max SE	0.39	0.00542	8.6	24.3

$^{\circ}$ paired t-test comparing reported average to that of PoisQLSpline yielded two-sided p-value$<$0.01

* paired t-test comparing reported average to that of NegBinQLSpline yielded two-sided p-value$<$0.01

this regard, with estimates ranging between 32 and 331. For DESeq, the number of genes with q-values less than 0.05 frequently exceeded the estimated total number of DE genes, which can be explained by the J-shape seen in its distribution of p-values from null simulated genes.

Table 6: Summary of simulation results for extreme NegBin simulations based on Arabidopsis data. See Table 3 for legend.

Method	# DE Top 200	eFDR$_{Q<.05}$	$N_{Q<.05}$	$\hat{N}_{DE}$
		$n=4$		
PoisQLSpline	183.1*	0.103	205.1	918
NegBinQLSpline	184.8°	0.0378	143.7	489
Exact edgeR.trend	185°	0.0979	233.5	357
GLM edgeR.trend	184.1$^{\star\circ}$	0.135	267.6	463
TSPM	139.8$^{\star\circ}$	0.562	589.8	1090
DESeq	180.4$^{\star\circ}$	0.0204	88.9	32.4
NBPSeq	176.3$^{\star\circ}$	0.108	186.7	289
Max SE	0.61	0.00797	10.8	29.8
		$n=6$		
PoisQLSpline	182.2*	0.107	212.9	1020
NegBinQLSpline	184.2°	0.0461	155.2	483
Exact edgeR.trend	184.5°	0.0933	221.2	364
GLM edgeR.trend	183.2$^{\star\circ}$	0.119	246.4	436
TSPM	119.9$^{\star\circ}$	0.458	351	1160
DESeq	179.7$^{\star\circ}$	0.0244	88.5	59.6
NBPSeq	173.1$^{\star\circ}$	0.108	174.9	297
Max SE	0.68	0.00677	6.8	25.2
		$n=10$		
PoisQLSpline	179*	0.126	216.4	1090
NegBinQLSpline	181.7°	0.0487	145.2	482
Exact edgeR.trend	182.2°	0.0865	200.1	370
GLM edgeR.trend	181.6°	0.101	214.7	417
TSPM	147.6$^{\star\circ}$	0.28	232.7	1170
DESeq	176.4$^{\star\circ}$	0.0205	73.6	76.8
NBPSeq	166$^{\star\circ}$	0.129	151.5	297
Max SE	0.54	0.00546	5.3	21.8

$^\circ$ paired t-test comparing reported average to that of PoisQLSpline yielded two-sided p-value$<$0.01

* paired t-test comparing reported average to that of NegBinQLSpline yielded two-sided p-value$<$0.01

The impact of the suggested quasi-likelihood approaches can be illustrated by comparing results from NegBinQLSpline and GLM edgeR.trend, which are closely related. The methods use similar estimates for the negative binomial dispersion of each gene; $\hat{\omega}_k$ values for GLM edgeR.trend are shrunken toward a fitted trend, while NegBinQLSpline uses $\hat{\omega}_k$ lying directly on the same fitted trend (see

Table 7: Summary of simulation results for extreme perturbed simulations based on fly embryo data. See Table 3 for legend.

Method	# DE Top 200	eFDR$_{Q<.05}$	$N_{Q<.05}$	$\hat{N}_{DE}$
		$n = 4$		
PoisQLSpline	185.4*	0.0993	224.8	976
NegBinQLSpline	187.8°	0.0443	159.7	612
Exact edgeR.trend	184.6$^{\star\circ}$	0.126	302.4	423
GLM edgeR.trend	184$^{\star\circ}$	0.133	309.9	451
TSPM	139.8$^{\star\circ}$	0.48	532.4	1080
DESeq	184$^{\star\circ}$	0.0431	126.1	99.2
NBPSeq	176.6$^{\star\circ}$	0.142	250.2	357
Max SE	0.44	0.007	10	28.4
		$n = 6$		
PoisQLSpline	194.7*	0.0941	347.3	1050
NegBinQLSpline	195.8°	0.0468	280.4	627
Exact edgeR.trend	193.4$^{\star\circ}$	0.0994	372.1	475
GLM edgeR.trend	193.2$^{\star\circ}$	0.106	377.5	491
TSPM	154.5$^{\star\circ}$	0.29	435.4	1140
DESeq	193.2$^{\star\circ}$	0.0305	193.3	195
NBPSeq	184.1$^{\star\circ}$	0.13	313.4	417
Max SE	0.48	0.00472	6.5	22.3
		$n = 10$		
PoisQLSpline	199.7*	0.119	544.6	1220
NegBinQLSpline	199.9°	0.0444	444	682
Exact edgeR.trend	199.5$^{\star\circ}$	0.0712	491.9	558
GLM edgeR.trend	199.4$^{\star\circ}$	0.0745	496.4	567
TSPM	192.4$^{\star\circ}$	0.173	565.1	1310
DESeq	199.4$^{\star\circ}$	0.0209	306.7	315
NBPSeq	194.4$^{\star\circ}$	0.11	415.8	505
Max SE	0.22	0.00419	5.9	20.7

$^{\circ}$ paired t-test comparing reported average to that of PoisQLSpline yielded two-sided p-value$<$0.01

* paired t-test comparing reported average to that of NegBinQLSpline yielded two-sided p-value$<$0.01

bottom of Figure 1). Also, both methods use asymptotic tests for differential expression. Although both methods generally performed well, NegBinQLSpline has clear advantages. In most simulation scenarios, the average number of truly DE genes contained in the list of 200 most significant genes was greater for NegBinQLSpline than for GLM edgeR.trend. While q-values for GLM edgeR.trend underestimated

Table 8: Summary of simulation results for extreme perturbed simulations based on Arabidopsis data. See Table 3 for legend.

Method	# DE Top 200	eFDR$_{Q<.05}$	$N_{Q<.05}$	$\hat{N}_{DE}$
	$n=4$			
PoisQLSpline	184.9*	0.0995	219.2	1020
NegBinQLSpline	186.7°	0.0406	145.6	584
Exact edgeR.trend	185.9$^{\star\circ}$	0.105	268.1	413
GLM edgeR.trend	185.2*	0.137	308.6	517
TSPM	153.6$^{\star\circ}$	0.529	627.3	1200
DESeq	183$^{\star\circ}$	0.0325	95.6	57.9
NBPSeq	177.8$^{\star\circ}$	0.114	211.2	346
Max SE	0.6	0.00698	10.1	28.4
	$n=6$			
PoisQLSpline	187*	0.109	269.7	1140
NegBinQLSpline	189°	0.0468	186.3	580
Exact edgeR.trend	188.2$^{\star\circ}$	0.0956	274.1	438
GLM edgeR.trend	187.3*	0.12	304.7	506
TSPM	135.7$^{\star\circ}$	0.417	412.8	1280
DESeq	184.3$^{\star\circ}$	0.0258	103.1	102
NBPSeq	177$^{\star\circ}$	0.121	213.4	361
Max SE	0.67	0.006	9.9	25.2
	$n=10$			
PoisQLSpline	186*	0.132	296.4	1210
NegBinQLSpline	188.4°	0.0498	189	591
Exact edgeR.trend	187.6$^{\star\circ}$	0.0891	257.2	445
GLM edgeR.trend	187.1$^{\star\circ}$	0.101	274.7	485
TSPM	158.9$^{\star\circ}$	0.25	313.2	1290
DESeq	182$^{\star\circ}$	0.022	86.7	123
NBPSeq	172.1$^{\star\circ}$	0.134	196.3	375
Max SE	0.53	0.00591	7.4	22.7

$^\circ$ paired t-test comparing reported average to that of PoisQLSpline yielded two-sided p-value$<$0.01

* paired t-test comparing reported average to that of NegBinQLSpline yielded two-sided p-value$<$0.01

eFDRs in every simulation scenario, q-values for NegBinQLSpline were most often accurate or slightly conservative. The advantages of NegBinQLSpline are most clearly evident in the ”extreme perturbed” simulations, which demonstrates the robustness of the QL methods to model misspecification.

5 Discussion

The QL methods are only supported by asymptotic theory in special cases, as discussed in Section 2. However, this did not adversely affect their performance in our simulation study. Indeed, the NegBinQLSpline method provided significance rankings as good as or better than each competing method, and its q-values were more accurate than those for every alternative method. Other methods, like edgeR, DESeq and NBPSeq, can test for differential expression between two treatments in a one-factor design using the exact test of Robinson and Smyth (2007). However, these methods also treat parameter estimates as true parameter values for their corresponding negative binomial distributions, which is also inaccurate and can produce an over-abundance of small p-values coming from EE genes. EdgeR, DESeq and NBPSeq methods use different dispersion estimates for each gene (for details, see McCarthy et al. (2012)), and regardless of estimation procedures, these estimates will have non-negligible uncertainties or biases for data sets with small values of $n - p$. While edgeR provides an option to assume a constant dispersion parameter common among all genes, this assumption has not been met in data sets we have examined.

When a relationship between estimated quasi-likelihood dispersions (as opposed to the dispersion in the variance function of the negative binomial distribution) and sample averages is present, the QLSpline method is generally preferable to the QLShrink method. The number of additional denominator degrees of freedom used in the QLShrink approach, $\hat{d}_0$, is estimated from the scatter of $\hat{\Phi}_k$ around a single constant for all k. The number of additional denominator degrees of freedom used in the QLSpline approach, $\hat{d}_0'$, is estimated from the scatter of $\hat{\Phi}_k$ around a spline fit to the (log-scale) relationship between $\hat{\Phi}_k$ and $\bar{y}_{\cdot\cdot k}$ for all k. When a relationship exists between sample means and estimated dispersions, the QLSpline method associates less random scatter with each $\hat{\Phi}_k$ than does the QLShrink method, which causes $\hat{d}_0'$ to be greater than $\hat{d}_0$. In the fly embryo data set, for which $n - p = 2$, the PoisQLSpline and PoisQLShrink approaches produced estimates $\hat{d}_0' = 7.1$ and $\hat{d}_0 = 2.4$, respectively. Having more denominator degrees of freedom helps to increase the power of the QLSpline method over that of the QLShrink method. Separately, failing to account for the relationship with sample means when shrinking estimated dispersions can induce bias. For example, if there is an increasing relationship between average counts and dispersion, then shrinking each estimated dispersion toward a single central value will systematically underestimate (overestimate) dispersions for genes with large (small) average counts.

When implementing the QLSpline methods, we suggest restricting the set of analyzed genes to include only those for which the average count across all samples is at least one and for which at least two samples have positive counts. This general

guideline has been appropriate for both real and simulated data originating from single factor experimental designs with a moderate number of levels. Experimental designs with more than one factor, like the analysis of the Arabidopsis data set that included block effects, may require more selective criteria when estimating dispersions (see, for example, Section 3.2).

The best significance rankings in most simulation scenarios came from the QLSpline method applied to either a quasi-Poisson or quasi-negative binomial model, and p-values from one of the QLSpline methods also produced q-values that most closely followed empirical FDRs. For moderate differences among library sizes, the QLSpline methods both produced similar results. For data sets with large differences between library sizes, NegBinQLSpline clearly outperformed PoisQLSpline. We therefore recommend NegBinQLSpline among the methods included in QuasiSeq. Intuitively pleasing, the QL (QLShrink and QLSpline) methods quantify the effect of parameter constraints in terms of residual degrees of freedom in an approach analogous to ANOVA (with shrunken variance estimates) and are robust to model misspecification. The implementation of the suggested methods via the QuasiSeq package is fast, simple and flexible enough to handle all models that can be analyzed by an ordinary GLM.

6 Additional Materials

6.1 QuasiSeq Package Demonstration on Arabidopsis data set

The authors have developed an R (R Development Core Team, 2011) package called QuasiSeq, available from the CRAN website, used to implement the suggested methods of this article. Code used to analyze the Arabidopsis data set described in Section 3 with the quasi-Poisson model and some selected results are shown below.

6.1.1 Analysis of Arabidopsis data without block effects

```
##Load QuasiSeq
library(QuasiSeq);
##Load data
library(NBPSeq); data(arab); counts=arab;

## Only use genes with an average count greater than 1
## and with at least 2 samples with positive counts
counts< −as.matrix(counts[rowSums(counts>0)>1&rowMeans(counts)>1,])
```

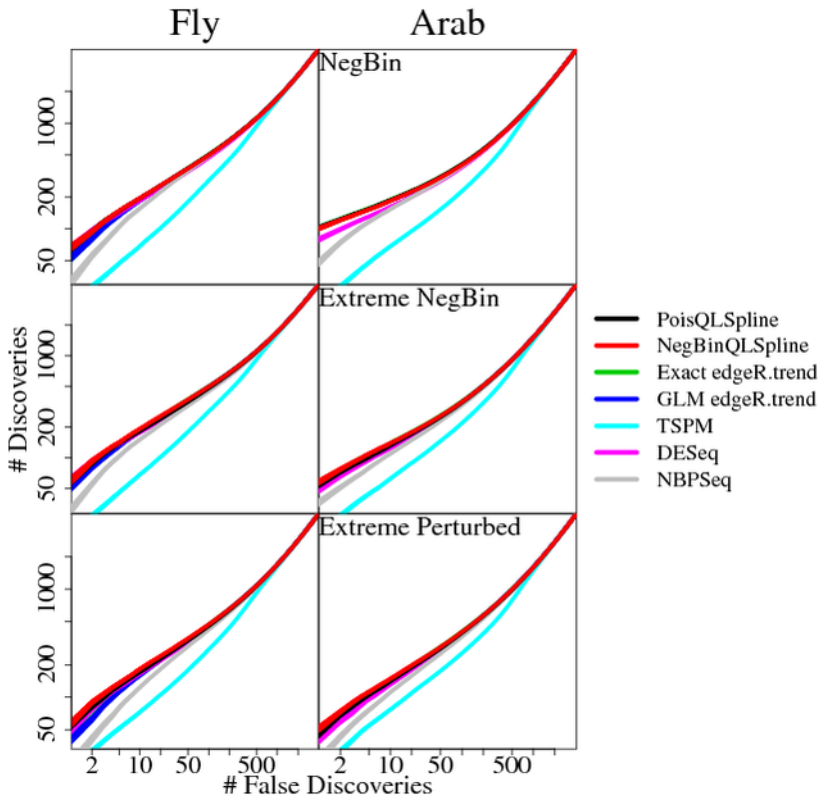

Figure 16: Curves relating average number of total discoveries to average number of false discoveries for negative binomial (top) and perturbed NegBin (bottom) simulations based on fly embryo (left) and Arabidopsis (right) data sets with $n = 4$.

View first 6 rows of data
head(counts)

	mock1	mock2	mock3	hrcc1	hrcc2	hrcc3
AT1G01010	35	77	40	46	64	60
AT1G01020	43	45	32	43	39	49
AT1G01030	16	24	26	27	35	20
AT1G01040	72	43	64	66	25	90
AT1G01050	49	78	90	67	45	60
AT1G01060	0	15	2	0	21	8

Define models under alternative and null hypotheses
design.list< −vector(”list”, 2)
Model under alternative hypothesis (DE gene)

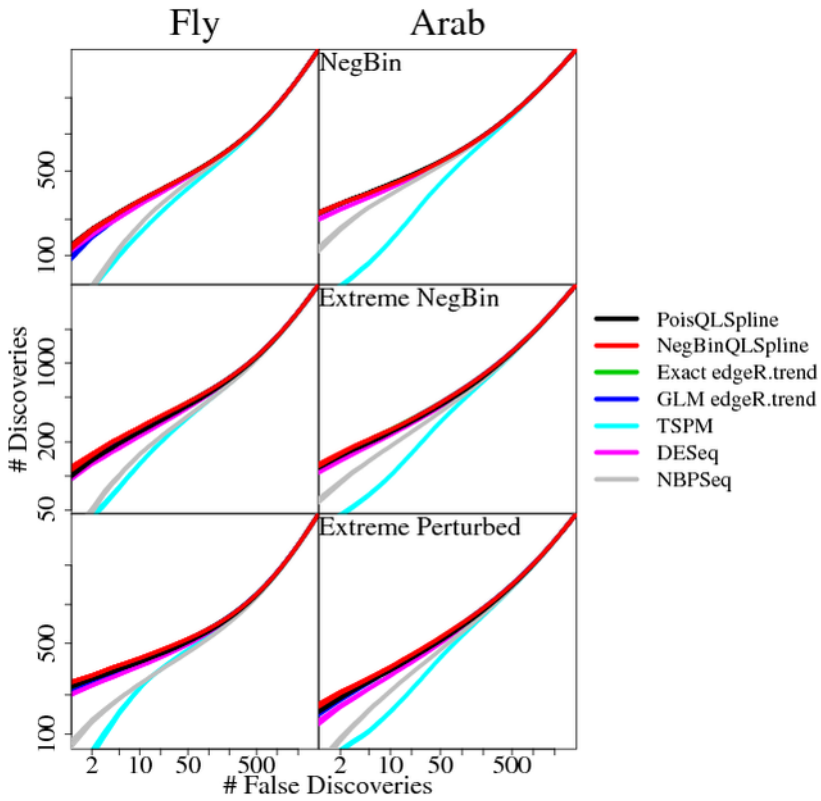

Figure 17: Curves relating average number of total discoveries to average number of false discoveries for negative binomial (top) and perturbed NegBin (bottom) simulations based on fly embryo (left) and Arabidopsis (right) data sets with $n = 10$.

design.list[[1]]< −rep(1:2, each=3)
Model under null hypothesis (EE gene)
design.list[[2]]< −rep(1, ncol(counts))

Estimate library size factors
size< −apply(counts, 2, quantile, .75)

Fit data
fit< −QL.fit(counts, design.list, log.offset=log(size), Model=”Poisson”)
results< −QL.results(fit)
[1] ”Spline scaling factor: 1.13041121749461”

How many additional degrees of freedom are obtained by QLShrink and QLSpline?

```
results$d0
  QLShrink    QLSpline
  3.305914    7.082764

## How many genes have q-values less than 0.05?
sapply(results$Q.values,FUN=function(qval) sum(qval<.05))
  QL    QLShrink    QLSpline
   0         386         203

## What is the estimated total number of DE genes?
t(round(nrow(counts)-results$m0))
              QL    QLShrink    QLSpline
 LRT12      2045        2103        2242
```

6.1.2 Analysis of Arabidopsis data with block effects

```
## Only use genes with at least 3 total samples (at least 1 sample from both treatments)
## with positive counts and an average count greater than 1
counts< −as.matrix(counts[rowSums(counts>0)>2&rowSums(counts[,1:3]>0)>0
&rowSums(counts[,4:6]>0)>0&rowMeans(counts)>1,])

## Define block and treatment levels
block< −rep(1:3,2); trt< −rep(1:2,each=3)

## Define model designs
design.list< −vector("list",3)

## Full model includes both block and treatment effects
design.list[[1]]< −model.matrix( as.factor(block)+as.factor(trt))

## Test for block effects using design with only treatment effects
design.list[[2]]< −trt

## Test for treatment effects using design with only block effects
design.list[[3]]< −block

test.mat< −rbind(1:2,c(1,3));
row.names(test.mat)< −c("Block","Trt")

fit< −QL.fit(counts,design.list,test.mat,log.offset=log(size), Model="NegBin")
results2< −QL.results(fit)
[1] "Spline scaling factor: 1.60291608692618"
```

```
## How many additional degrees of freedom are obtained by QLShrink and QLSpline?
results$d0
  QLShrink   QLSpline
  3.305914   7.082764

## How many genes have q-values less than 0.05 for the tests of block and trt effects?
sapply(results2$Q.values,FUN=function(qval) colSums(qval<.05))
          QL   QLShrink   QLSpline
Block     23       2665       2870
Trt        0       1594       1780

## What is the estimated number of total genes with block effects and trt effects?
round(nrow(counts)-results2$m0)
           Block   Trt
QL          9664  5772
QLShrink   10494  5939
QLSpline   10444  5804
```

References

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

Auer, P. L. and R. W. Doerge (2011): "A two-stage poisson model for testing RNAseq data," *Statistical Applications in Genetics and Molecular Biology*, 10.

Blekhman, R., J. C. Marioni, P. Zumbo, M. Stephens, and Y. Gilad (2010): "Sex-specific and lineage-specific alternative splicing in primates," *Genome Research*, 20, 180–189.

Bullard, J. H., E. Purdom, K. D. Hansen, and S. Dudoit (2010): "Evaluation of statistical methods for normalization and differential expression in mRNA-seq experiments," *BMC Bioinformatics*, 11.

Cumbie, J. S., J. A. Kimbrel, Y. Di, D. W. Schafer, L. J. Wilhelm, S. E. Fox, C. M. Sullivan, A. D. Curzon, J. C. Carrington, T. C. Mockler, and J. H. Chang (2011): "GENE-counter: A computational pipeline for the analysis of RNA-seq data for gene expression differences," *PLoS ONE*, 6.

Di, Y., D. W. Schafer, J. S. Cumbie, and J. H. Chang (2011): "The NBP negative binomial model for assessing differential gene expression from RNA-seq," *Statistical Applications in Genetics and Molecular Biology*, 10.

Lu, J., J. K. Tomfohr, and T. B. Kepler (2005): "Identifying differential expression in multiple SAGE libraries: an overdispersed log-linear model approach," *Bioinformatics*, 6.

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

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

McCullagh, P. (1983): "Quasi-likelihood functions," *Annals of Statistics*, 11, 59–67.

McCullagh, P. and J. A. Nelder (1983): *Generalized Linear Models*, New York: Chapman and Hall, first edition.

Nettleton, D., J. T. G. Hwang, R. A. Caldo, and R. P. Wise (2006): "Estimating the number of true null hypotheses from a histogram of p-values," *Journal of Agricultural, Biological, and Environmental Statistics*, 11, 337–356.

R Development Core Team (2011): *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.

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.

Robinson, M. D. and G. K. Smyth (2007): "Small-sample estimation of negative binomial dispersion, with applications to SAGE data," *Biostatistics*, 9, 321–332.

Robinson, M. D. and G. K. Smyth (2008): "Moderated statistical tests for assessing differences in tag abundance," *Bioinformatics*, 23, 2881–2887.

Smyth, G. K. (2004): "Linear models and empirical bayes methods for assessing differential expression in microarray experiments," *Statistical Applications in Genetics and Molecular Biology*, 3.

Storey, J. D. and R. Tibshirani (2003): "Statistical significance for genome wide studies," *Proceedings of the National Academy of Sciences*, 100, 9440–9445.

Tjur, T. (1998): "Nonlinear regression, quasi likelihood, and overdispersion in generalized linear models," *American Statistician*, 52, 222–227.

Vêncio, R. Z., H. Brentani, D. F. Patrão, and C. A. Pereira (2004): "Bayesian model accounting for within-class biological variability in serial analysis of gene expression (SAGE)," *BMC Bioinformatics*, 5, 119–131.

Zhou, Y.-H., K. Xia, and F. A. Wright (2011): "A powerful and flexible approach to the analysis of RNA sequence count data," *Bioinformatics*.