Goodness-of-Fit Tests and Model Diagnostics for Negative Binomial Regression of RNA Sequencing Data

Goodness-of-Fit Tests and Model Diagnostics for Negative Binomial Regression of RNA Sequencing Data

Gu Mi[1]*, Yanming Di[1,2], Daniel W. Schafer[1]

1 Department of Statistics, Oregon State University, Corvallis, Oregon, United States of America, 2 Molecular and Cellular Biology Program, Oregon State University, Corvallis, Oregon, United States of America

* neo.migu@gmail.com

Abstract

This work is about assessing model adequacy for negative binomial (NB) regression, particularly (1) assessing the adequacy of the NB assumption, and (2) assessing the appropriateness of models for NB dispersion parameters. Tools for the first are appropriate for NB regression generally; those for the second are primarily intended for RNA sequencing (RNA-Seq) data analysis. The typically small number of biological samples and large number of genes in RNA-Seq analysis motivate us to address the trade-offs between robustness and statistical power using NB regression models. One widely-used power-saving strategy, for example, is to assume some commonalities of NB dispersion parameters across genes via simple models relating them to mean expression rates, and many such models have been proposed. As RNA-Seq analysis is becoming ever more popular, it is appropriate to make more thorough investigations into power and robustness of the resulting methods, and into practical tools for model assessment. In this article, we propose simulation-based statistical tests and diagnostic graphics to address model adequacy. We provide simulated and real data examples to illustrate that our proposed methods are effective for detecting the misspecification of the NB mean-variance relationship as well as judging the adequacy of fit of several NB dispersion models.

Introduction

The negative binomial (NB) model has been widely adopted for regression of count responses because of its convenient implementation and flexible accommodation of extra-Poisson variability. Let Y represent a univariate count response variable and X a p-dimensional vector of known explanatory variables. Then an NB log-linear regression model specifies that the probability distribution of Y is NB with mean μ and dispersion parameter ϕ, with $\log(\mu) = X' \beta$ where β is a p-dimensional vector of unknown regression coefficients.

The NB distribution can be derived as a Poisson-gamma mixture model. For the conventional parameterization (which we refer to as NB2), suppose v is a gamma-distributed random variable with E $(v) = \mu$ and Var $(v) = \phi\mu^2$, and that $Y|v \sim$ Poisson (v), then the marginal

distribution of Y is NB with mean μ and variance $\mu + \phi\mu^2$ (see, for example, [1]). The NB2 probability mass function (p.m.f.) has the form:

$$f(y|\mu,\phi) = \frac{\Gamma(y+\theta)}{\Gamma(\theta)\Gamma(y+1)} \left(\frac{\theta}{\mu+\theta}\right)^\theta \left(\frac{\mu}{\mu+\theta}\right)^y,$$

where $\theta = 1/\phi$.

Other NB parameterizations follow from different parameterizations for the gamma mixing distribution. A general form, called NBP [2, 3], follows from the assumption that the gamma variance is $\phi\mu^\alpha$, and has the same form of p.m.f. $f(y|\mu,\phi)$, but with θ replaced by $\phi^{-1}\mu^{2-\alpha}$. In this parameterization, $E(Y) = \mu$ and $Var(Y) = \mu + \phi\mu^\alpha$. We note that (1) for identically distributed count variables the NBP distribution is over-parameterized, but in a regression setting it offers additional flexibility in mean-variance modeling, which is useful in the RNA sequencing (RNA-Seq) analysis that follows; and (2) NBP includes the well-known NB1 ($\alpha = 1$) and NB2 ($\alpha = 2$) parameterizations as well as others. Greene [2] specified the symbol "P" for our α, which is why this parameterization is called "NBP".

RNA-Seq analysis [4] may be performed on biological units from any of the traditional forms of life science study, such as randomized experiments with multiple treatments and co-variates, or observational studies with multiple observed explanatory variables. The response variable for each unit is a vector of relative frequencies, each of which is a measure of a gene's (or an isoform's) expression level. Although much of the statistical attention to RNA-Seq analysis has so far been focused on the two-group problem—and, therefore, on identification of differentially expressed genes—there is a clear need for regression analysis for identifying differential expression after accounting for other variables, and for identifying patterns of expression and differential expression as a function of explanatory variables.

Future statistical techniques might be derived for the multivariate regression on all genes simultaneously, but the problem is currently tackled by the simpler univariate regression on each gene individually, with appropriate attention to false discovery rate. The response for a single gene is the number of RNA-Seq reads corresponding to that gene (Y) out of a total number of reads for the particular biological unit (s). Although there is evidence that the "technical variability" in Y—meaning the variability in the RNA-Seq technical procedure repeated on a single biological unit—can be described by a Poisson distribution [5], the observed variability from multiple biological units in the same observational or experimental group is greater than Poisson (see, for example, [6, 7]). The gamma mixture of Poissons, as described above, is a conceptually appealing alternative because the gamma mixing represents "biological variability". Practically, the NB model is both flexible and convenient.

The primary statistical challenge involves simultaneous regression fitting for tens of thousands of genes from fairly small numbers of biological samples (e.g., less than twenty). An important power and efficiency issue in this case involves the modeling of the NB2 dispersion parameter ϕ. Five possibilities, for example, are (1) ϕ is constant for all genes; (2) ϕ is allowed to differ between genes but is constant within gene under all conditions; (3) ϕ is allowed to differ for all gene/condition combinations; (4) ϕ is taken to be a function of μ; and (5) ϕ is taken to have a trend as a function of μ, but with some additional between-gene variability. More flexible models are much more likely to fit the data, of course, but at the expense of tens of thousands of nuisance parameters. If a more specific model fits, based on vastly fewer nuisance parameters, it could offer substantial power and efficiency gains (for improved "true discovery" rates of differential expression, for example). Because of the very large number of hypothesis tests performed in a single RNA-Seq study and the very large number of RNA-Seq studies

being performed world-wide, even a small improvement in power can have an important impact on the overall rate of scientific learning from the RNA-Seq technology.

Traditional tools for model diagnostics in generalized linear models (GLM), such as deviance and Pearson residuals and goodness-of-fit (GOF) tests, are suitable for binomial and Poisson regression if the means are large, i.e., the adequacy of the normal and χ^2 null distributions for residuals and GOF test statistics, respectively, are justified under central-limit-theorem-like asymptotics rather than large sample size asymptotics [8]. Such GOF tests are not appropriate for small means (which are typical for the majority of genes in RNA-Seq analysis), and the theory for the null sampling distribution of the residuals and GOF test statistics does not extend to NB regression.

In this article, we propose a goodness-of-fit test statistic for NB regression based on Pearson residuals, and the calculation of a p-value using Monte Carlo-estimated null sampling distributions. The same simulations are used to estimate expected ordered residuals for an empirical probability plot. For RNA-Seq diagnostics, the GOF p-values from all genes are examined in a uniform QQ plot and combined via the Fisher's combined probability test (Fisher's method [9]).

Background

Dispersion Modeling

Let Y_{ij} denote an RNA-Seq read count for the ith gene ($i = 1, \cdots, m$) of the jth experimental or observational unit ($j = 1, \cdots, n$), and $\mathbf{X}_j$ the associated p-dimensional explanatory variable. Suppose $Y_{ij} \sim \mathrm{NB}(\mu_{ij}, \phi_{ij})$ where μ_{ij} is the mean and ϕ_{ij} is the dispersion parameter in the NB2 parameterization. Suppose also that

$$\log(\mu_{ij}) = \log(s_j) + \log(R_j) + \log(\pi_{ij}),$$

with $\pi_{ij} = \exp(\mathbf{X}_j'\beta_i)$, where s_j is the library size (the number of RNA-Seq reads in the biological sample from unit j), and R_j is an optional normalization factor estimated beforehand [6, 10, 11] and treated as known. In this formulation, π_{ij} is the mean relative frequency of occurrence of RNA-Seq reads associated with gene i, which is taken to be the expression level of gene i associated with observational or experimental unit j.

We label some of the ways to model the nuisance parameters ϕ_{ij} as follows:

1. Genewise: $\phi_{ij} = \phi_i$ (constant within each gene i across all conditions j), with m parameters for NB dispersion.

2. Common: $\phi_{ij} = \phi$ (constant for all gene/condition combinations), with one parameter for NB dispersion.

3. NBP: $\log(\phi_{ij}) = \alpha_0 + \alpha_1 \log(\pi_{ij})$, equivalent to assuming NBP response distribution discussed in Di $et\ al.$ [7], with two parameters for NB dispersion.

We also introduce here a new approach, in which the dispersion parameter trend is quadratic on the log scale:

4 NBQ: $\log(\phi_{ij}) = \alpha_0 + \alpha_1 \log(\pi_{ij}) + \alpha_2 [\log(\pi_{ij})]^2$, with three parameters for NB dispersion.

An important related method estimates the ϕ_{ij}'s via non-parametric regression:

5 Non-parameteric: ϕ_{ij} is estimated in a first step as a smooth function of $\log(\hat{\phi}_{ij})$ on $\log(\hat{\mu}_{ij})$, and then treated as known in the second step of regression coefficient inference.

In addition, there are variants that use an average of trend and individually-estimated dispersion parameters, based on empirical Bayes considerations [12]:

6   Tagwise-common: ϕ_{ij} is estimated as a weighted average of the common and genewise estimates, based on empirical Bayes calculations.

7   Tagwise-trend: ϕ_{ij} is estimated as a weighted average of the non-parametric and genewise estimates, based on empirical Bayes calculations.

Methods for inference from the genewise, common, non-parametric, tagwise-common, and tagwise-trend approaches are available in the `edgeR` Bioconductor package [13, 14]. The non-parametric method is also available in the DESeq and NBPSeq packages [6, 15]. The NBP and NBQ approaches are implemented in NBPSeq [15, 16].

The details of estimation for these methods are important but are not relevant to the proposed diagnostic tools and so are not discussed here. The adequacy of the models for RNA-Seq data is not yet well understood. We wish to use the model diagnostic tools proposed in this article to judge the degree of fit of the various models on real RNA-Seq data—particularly the fit of simple parametric models for the trend of $\log(\phi)$ as a function of π and the degree of noise, if any, about this trend, so that realistic robustness and power studies can follow.

To further clarify this point, Fig. 1 shows a log-log scatter plot of method-of-moments-like estimated NB2 dispersion parameters, $\hat{\phi}$, versus estimated mean relative frequencies, $\hat{\pi}$, for each of 19,623 genes from a single sample of size three of a pilot Arabidopsis RNA-Seq study examined in [7]. The curves on the plot are estimated dispersion trends based on the models described above. Polynomial gamma log-linear regression models of $\hat{\phi}$ on $\log(\hat{\pi})$ were used for quick-and-dirty testing and quantification of the trend, as follows. The linear model explains 24.1% of the variability in logged dispersion parameter estimates. A quadratic term (with p-value < 0.0001) explains an additional 7.2% of variability. A cubic term (p-value < 0.0001) in a full cubic model explains less than 0.1% additional variability. This plot and informal analysis suggest that the common ϕ model is inadequate; the trend is primarily, but not entirely, linear; and that a quadratic model captures essentially all of the trend in this particular dataset.

A simple model for trend in NB dispersion parameter ϕ as a function of mean relative frequency π is a good starting point for reducing the number of nuisance parameters, but the evidence of a trend does not imply that the ϕ's fall exactly on the trend; there may be additional variability in ϕ for genes with the same value of π. The main questions we wish to address with diagnostic tools are the following: (1) Does the NB assumption hold for a very rich model (for both regression and dispersion)? (2) What relatively simple models are adequate for describing ϕ as a function of π? (3) Is there evidence of additional biological variability in ϕ between genes having the same value of π?

Other Related Work on Model Diagnostics

Best *et al.* [17] extended Anscombe's tests of fit for the NB distribution by using fourth order smooth tests, but these tests don't extend in an obvious way to regression models for non-exponential family response distributions. The test we propose in this paper gives similar results to theirs for independent and identically distributed samples and can also be used for the procedures that involve non-parametric trend fitting and empirical Bayes averaging. Esnaola *et al.* [18] proposed a larger family of response distributions for RNA-Seq analysis, which permits the testing of NB as a special case; but we do not believe the approach (validated under extensively replicated experiments) is suitable for the small sample sizes we have in mind here.

Similar regression diagnostic approaches that use Monte Carlo or resampling to derive null sampling distributions of diagnostic quantities have been previously proposed for several
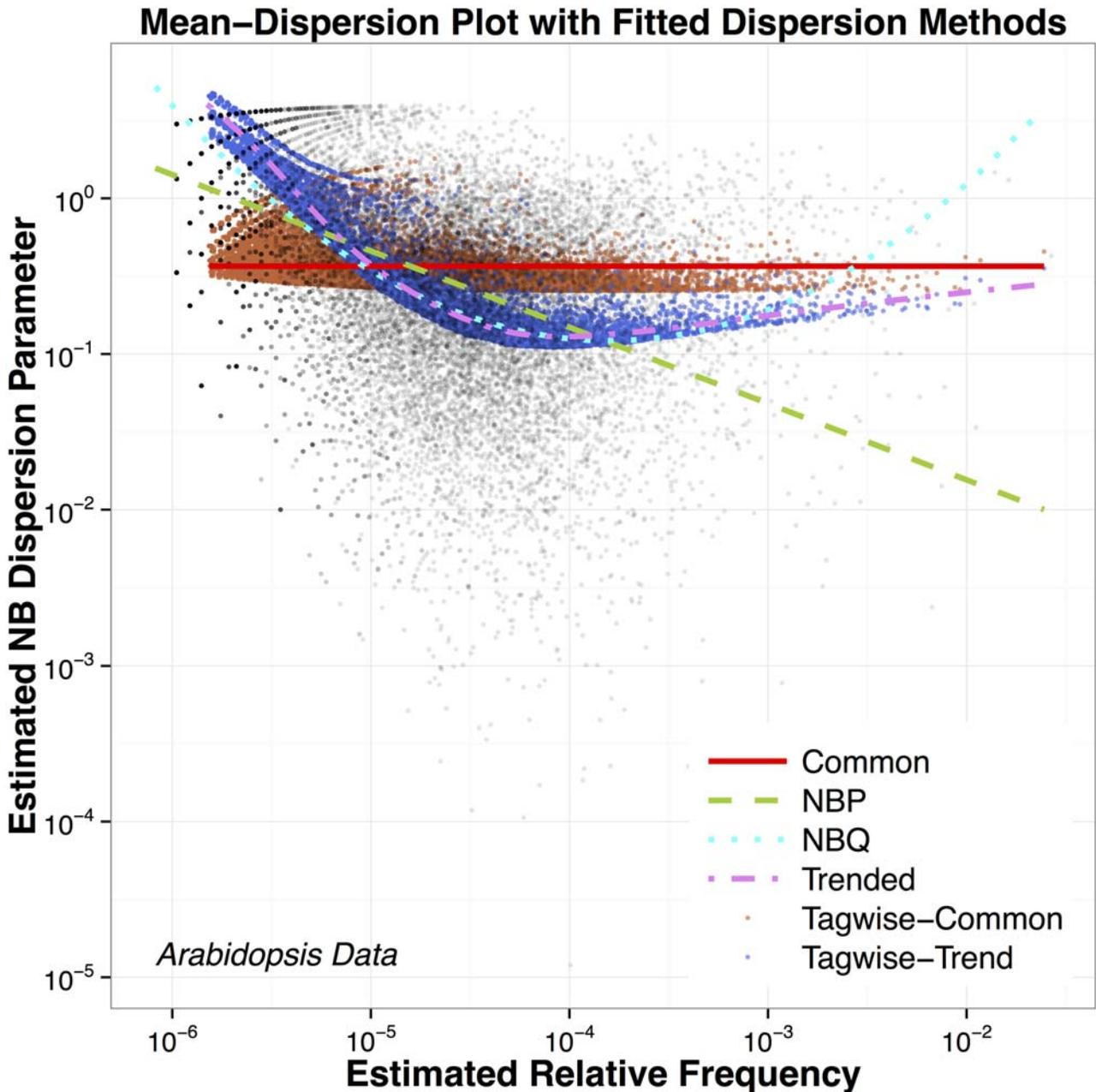
Fig 1. Mean-Dispersion Plot with Fitted Dispersion Models. The mean-dispersion plot with six fitted dispersion models (common, NBP, NBQ, trended, tagwise-common and tagwise-trend) for the Arabidopsis RNA-Seq dataset (19,623 genes from three biological samples in the mock treatment group). For better visualization, we use dots instead of curves for the tagwise procedures to indicate the variability of individual NB2 dispersion parameters about the trend.

situations. For ordinary linear regression, Atkinson [19] proposed half normal plots of jack-knife residuals. For logistic regression, Landwehr *et al.* [20] proposed an "empirical probability plot" in which ordered residuals from the observed data are plotted against their expected values (or median values), as computed by Monte Carlo simulations. Their simulation procedure, which resembles parametric bootstrapping, is based on the estimated parameters from the fitted model. Similar graphical displays were adopted as informal checks of various count models.

For example, Svetliza *et al.* [21] considered normal probability plots for log-linear Poisson, log-linear NB and non-linear NB models. Garay *et al.* [22] evaluated GOF between zero-inflated Poisson (ZIP) and zero-inflated NB (ZINB) models. Both of these used simulated envelopes in their plots, but with standard normal quantiles (instead of quantiles from simulations) on the *x*-axis. None of the aforementioned papers provided statistical tests for evaluating model lack-of-fit.

Materials and Methods

## GOF Tests for Univariate NB Regression and the Empirical Probability Plot

We first consider univariate NB regression in this section and then return to the RNA-Seq problem of NB regression for each of many genes later. For regression data with counted response, we wish to determine whether any NBP model fits and, because of the convenience of NB2 estimation programs, whether the NB2 model fits in particular. We propose two GOF tests and an associated residual plot. The null hypothesis is that the count data follow the assumed NB regression model. In particular, the means follow the log-linear regression model, and the dispersion parameters follow the specified dispersion model (i.e., NB2, NBP, etc.). The test *p*-values provide an overall assessment of fit and the plot shows whether a small GOF *p*-value might be due to a small portion of the data. We use the same notation as in the "Background/Dispersion Modeling" subsection, but without the subscript *i*. In the RNA-Seq context, the methods of this section apply to a single gene. We start with Pearson residuals: $r_j = (y_j - \hat{\mu}_j)/\hat{s}_j$, where $\hat{\mu}_j$ is the estimated NB mean and $\hat{s}_j$ is the estimated NB standard deviation of y_j from the particular model being tested, for $j = 1, \cdots, n$.

We first propose an empirical probability plot of the ordered Pearson residuals $r_{(j)}$ versus the sampling distribution medians for each ordered Pearson residual, $\text{Med}[r_{(j)}]$, assuming the proposed NB regression model is correct. To approximate the medians, we simulate a large number of NB regression datasets of the same size and form as the observed one, using the data estimates as parameters for simulation; fit the same NB regression model to each simulated dataset; extract the ordered Pearson residuals; and retain the sample medians for each ordered residual. This is exactly the Landwehr *et al.*'s approach [20] applied to NB regression. A 95% pointwise prediction envelope (in dashed blue lines) can be formed from the similarly estimated 2.5[th] and 97.5[th] percentiles of the ordered residuals. Fig. 2 from the Results section shows an empirical probability plot for an earthquake event dataset.

We then wish to provide a global GOF test to accompany the empirical probability plot of Pearson residuals. A natural starting point is a test based on the Pearson statistic, i.e., the sum of squared Pearson residuals. The classical use of the χ^2 reference distribution is not appropriate here, but the null sampling distribution may be approximated by the Monte Carlo estimate. A *p*-value can be obtained as the proportion of simulated samples that produce a Pearson statistic as extreme or more extreme than the observed one. This is similar to the approach of Best *et al.* [17] in its application of parametric bootstrap to obtain a GOF *p*-value. We have found that our procedure and theirs give very similar results for samples with a common mean, but the approach discussed in Best *et al.* [17] requires estimation of higher order moments, which is difficult for the regression models we have in mind for RNA-Seq analysis.

Since the simulations provide estimated sampling distributions for each ordered residual, a finer test statistic is available as the sum of squared differences of the ordered residuals from their sampling distribution medians. We believed this was a worthwhile test statistic to consider given that we had already obtained approximate sampling distributions for each ordered residual to obtain the type of diagnostic plot discussed in Landwehr *et al.* [20]. The test is also
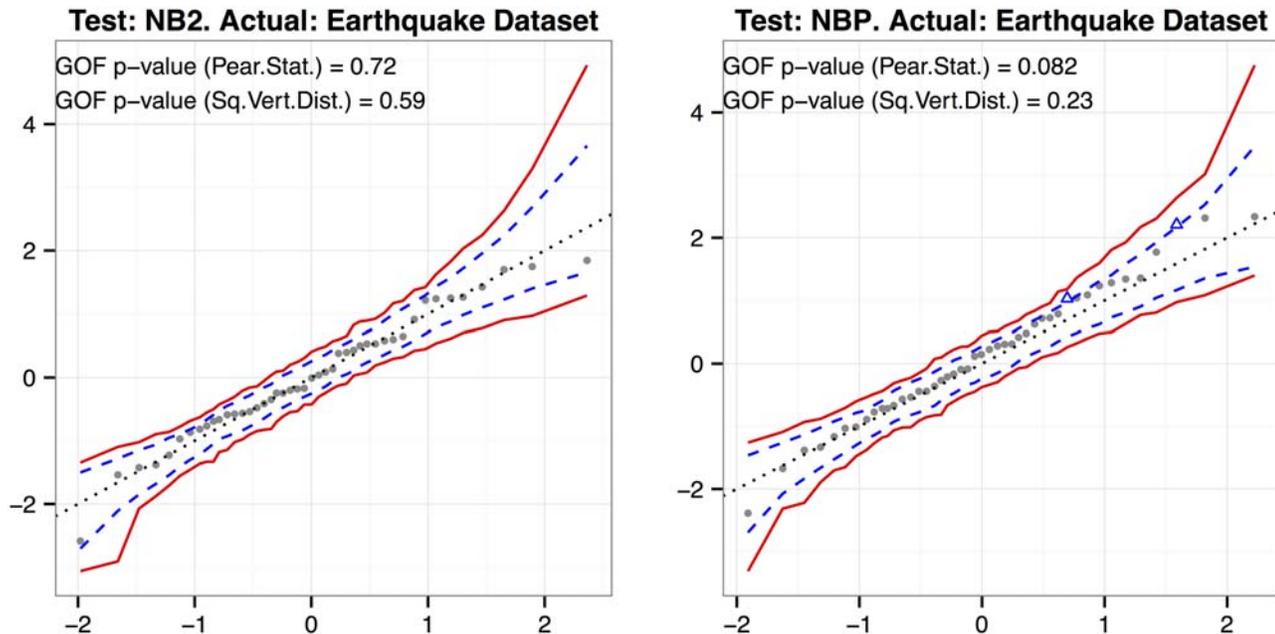
Fig 2. Empirical Probability Plots for Evaluating NB2 and NBP Model Fits. Empirical probability plots with GOF test *p*-values for evaluating NB2 and NBP model fits on the earthquake dataset (sample size: 45), with 95% prediction envelope in dashed blue lines and 95% simultaneous prediction band in solid red lines (based on 999 Monte Carlo simulations). Points outside the prediction envelope are flagged as blue triangles.

doi:10.1371/journal.pone.0119254.g002

related to the SAM graphical procedure of Tusher *et al.* [23] for identifying differential gene expression from microarray.

The following algorithm defines the diagnostic empirical probability plot of residuals and the Monte Carlo GOF test *p*-value based on the second test statistic.

#1 Fit an NB regression model from the data $\mathbf{Y}^{(0)} = (Y_1, \cdots, Y_n)^T$; estimate all unknown dispersion parameters, e.g., $\hat{\alpha}_0, \hat{\alpha}_1, \cdots$, and regression coefficients $\hat{\beta}^{(0)}$; calculate Pearson residuals $\mathbf{r}^{(0)} = (r_1^{(0)}, \cdots, r_n^{(0)})$ and the mean vector $\hat{\mu}^{(0)}$.

#2 For $h = 1, \cdots, R$:

   a. Simulate a random vector $\mathbf{Y}^{(h)}$ from $\mathrm{NB}(\hat{\mu}^{(0)}, \hat{\alpha}_0, \hat{\alpha}_1, \cdots)$.

   b. Compute and retain Pearson residuals $\mathbf{r}^{(h)}$.

#3 Find the median, 2.5^{th} and 97.5^{th} percentiles of the Monte Carlo sampling distribution for each ordered residual, denoted by $\tilde{r}_{(j)}^{50}, \tilde{r}_{(j)}^{2.5}$ and $\tilde{r}_{(j)}^{97.5}$, respectively; plot the ordered residuals from the observed data against the Monte Carlo medians; draw a 95% pointwise prediction envelope on the plot by connecting the $\tilde{r}_{(j)}^{2.5}$'s (for lower bound) and the $\tilde{r}_{(j)}^{97.5}$'s (for upper bound).

#4 Compute the sum of squared deviations of ordered residuals from the medians of their sampling distributions $d^{(h)} = \sum_{j=1}^{n} (r_{(j)}^{(h)} - \tilde{r}_{(j)}^{50})^2$ for the observed data ($h = 0$) and for

the simulated samples ($h = 1, \cdots, R$); compute a Monte Carlo GOF test p-value by:

$$P^{MC}_{1\,\text{sided}} = \frac{\sum_{h=1}^{R} \mathbb{1}\left(d^{(h)} \geq d^{(0)}\right) + 1}{R + 1}, \tag{1}$$

where $\mathbb{1}(A)$ is the indicator function equal to 1 if the event A is true and 0 otherwise [24].

The Pearson GOF p-value is computed in the same way, but using the sum of squared residuals as the test statistic. The two statistics are visualized on the empirical probability plots (see, e.g., Fig. 2) as the sum of squared deviations about the $y = 0$ line and the sum of squared deviations about the $y = x$ dotted line. For this reason, we call the latter statistic the sum of squared vertical distances.

In using Monte Carlo simulation in lieu of theory to obtain the sampling distributions of the ordered residuals and test statistics, it would be ideal, but impossible, to use the *true* parameter values rather than their *estimates* from the data. Using estimated parameters in the simulation may lead to conservative test results, especially in small sample situations. As the sample sizes increase, we expect the sampling distribution of the *residuals* to be about the same.

We use $R = 999$ Monte Carlo samples so that the (binomial) standard error in p-value estimation is 0.016 for p-values near 0.5, 0.007 for p-values near 0.05, and 0.003 for p-values near 0.01. As pointed out in North *et al.* [25], adding 1 to both the numerator and the denominator in Equation (1) produces a slightly biased estimate of the true p-value but with the correct Type-I error rate, in contrast to the unbiased but anti-conservative p-value obtained without adding the 1.

Diagnostic Tools for RNA-Seq Modeling

A major step for improving RNA-Seq analysis is the comparative evaluation of the models and methods for incorporating commonalities of NB dispersion parameters within and across genes, as described in the Introduction section and displayed in Fig. 1. For studying these models on a given RNA-Seq dataset, we propose fitting them, calculating the squared vertical distance NB GOF p-value for each of a randomly selected sample of genes (i.e., testing the univariate NB regression model fit for each gene individually, using the NB2 dispersion parameter estimated according to the particular dispersion model), drawing a uniform QQ plot of the p-values, and calculating a single p-value using the Fisher's method.

Let p_i be the GOF p-value for gene i, based on the parameter estimates from the global model being tested. Fisher's method produces a single GOF p-value by testing the conformity of the p_i's from m genes to a standard uniform distribution. If the m single-gene p-values are independent and all follow standard uniform distribution, the test statistic $X^2 = -2 \sum_{i=1}^{m} \log(p_i)$ follows a $\chi^2_{(2m)}$ distribution. In this multiple genes case, the null hypothesis is that the NB counts for each gene follow the assumed NB regression model (the same as in the univariate case) and the NB dispersion parameters across all genes follow the specified dispersion model.

Although it is possible to base this on all genes, we elect to reduce the computational burden by selecting a random sample m^* genes, and use $m^* = 1,000$ as a computationally tolerable value. We do not have a direct way to study the suitability of this sample size for testing whether the p-values follow a uniform(0,1) distribution with Fisher's method; but we do have an indirect approach that helps. Let P be the proportion of genes with p-values less than 0.05. A binomial 95% confidence interval for P from a sample of 1,000 genes has half-width 0.0135, so we would be likely to detect lack-of-fit to the uniform(0,1) if the actual proportion is 0.0635 or greater. Although we use Fisher's method rather than this (arbitrary) binomial test, the binomial calculation provides some clarification of the type of departure from the uniform(0,1) that

we are likely to detect with a sample of 1,000 genes. The m p-values are not exactly independent, as required for the theory of Fisher's combined test, but are approximately so because the global parameter estimates are based on such a large number of genes.

As we noted earlier, the Pearson statistic can give misleading results if there are combinations of under- and over-dispersion relative to the response distributional model being tested. We have found this problem to be exacerbated in the RNA-Seq setting and so focus only on the squared vertical distance estimator, which does not suffer from the same problem.

An extremely small (or large) X^2 value can be due to either a small number of extreme single-gene p-values or a large number of moderately small (or large) p-values. The Fisher's method itself cannot distinguish the different possibilities. Along with the Fisher's combination of p-values, we suggest a uniform QQ plot of individual p-values to help reveal the nature of any lack-of-fit, indicated by a higher than expected proportion of small p-values. The proportion of genes with small p-values may have some effect on the thinking about appropriate models.

The Earthquake Event Dataset

We consider an earthquake event dataset when illustrating the empirical probability plot and the GOF tests for univariate NB regression. The dataset (provided in the NBGOF package) contains the frequencies of all earthquakes of a given magnitude (reported to one decimal place) for magnitudes from 4.5 to 9.1, that occurred between January 1, 1964 to December 31, 2012 (Source: Composite Earthquake Catalog, Advanced National Seismic System, Northern California Earthquake Data Center (NCEDC), http://quake.geo.berkeley.edu/cnss/). The empirical probability plots with GOF test results, based on a log-linear regression of mean number of earthquakes on magnitude, are shown in Fig. 2. Neither the NB2 nor NBP model shows lack-of-fit.

The Arabidopsis Study

We use an Arabidopsis dataset when demonstrating the diagnostic tools for RNA-Seq modeling. *Arabidopsis thaliana* has been intensively studied as a model organism in plant biology. The Arabidopsis data discussed in Di *et al.* [7] contain RNA-Seq reads that aligned to more than 25,000 genes from two groups of Arabidopsis samples of size three each. The two groups of size three each were derived from plants inoculated with $\Delta hrcC$ of *Pseudomonas syringae* pv tomato DC3000 or 10 mM MgCl$_2$ (mock). The dataset used in this article comes from Di *et al.* [7], which is a subset of the data described in Cumbie *et al.* [26].

Simulation Parameter Specifications

We specify the parameters in the simulation studies in the subsection "Results/Error Rates of GOF Tests in Simulations" (Tables 1 and 2) as follows: the mean is determined by $\mu = \exp(X' \beta)$ with the coefficient $\beta = (15, -1.5)$. The design matrix X takes an intercept and a covariate equally spaced from 4 to 8 of length $n = 5, 10, 50$ and 100. The resulting mean levels approximately range from 20 to 8,100.

For the NB2 model fit: the NB2 responses are simulated under $\alpha_0 = \log(0.1)$, $\alpha_1 = 0$ and $\phi = 0.1$. The NB1 responses are obtained by simulating NB2 with $\alpha_0 = \log(0.5)$, $\alpha_1 = -1$, and $\phi = 0.5/\mu$. The "NB2 plus outliers" responses are simulated with the same dispersions as in NB2, except we randomly double 20% of responses (as outliers). The "NB2 plus noise" responses are simulated with the same dispersions as in NB2, except ϕ is specified as $0.1 \cdot \exp(G)$, where $G \sim \mathcal{N}(0,1)$. The α in the variance $\mu + \phi\mu^{\alpha}$ is determined by $\alpha = \alpha_1 + 2$.

For the NBP model fit: the specifications are almost the same as in the NB2 model fit above, except for the simulated NB2 data, we use $\alpha_0 = \log(0.05)$ so that $\phi = 0.05$.

Results

GOF Tests for Univariate NB Regression and the Empirical Probability Plot

Application to the Earthquake Dataset. This paper is mainly about RNA-Seq analysis, but in order to convey the univariate version of the problem, we wish to use a non-biological example, the earthquake dataset, that better illustrates the main issues (because of its larger sample size). Fig. 2 shows the empirical probability plot, along with a 95% pointwise prediction envelope (in dashed blue lines), for the NB2 and NBP regression of 48-year earthquake frequencies on magnitude (i.e., the Gutenberg-Richter Law, for 45 magnitudes from 4.5 to 9.1). Note that the band is formed from prediction intervals for the corresponding sample quantiles (not confidence intervals for their expected values). If the model fits, we would expect about 95% of the ordered residuals to fall within the band. We also superimpose a 95% simultaneous prediction band in solid red using the simulation method discussed in Buja *et al.* [27], so that for 95% of samples *all* ordered residuals should be contained in the red band. See the Materials and Methods section for details of the earthquake dataset.

The *p*-values for testing the NB2 and NBP models on the earthquake dataset are shown in Fig. 2. The suggestive Pearson statistic *p*-value for goodness-of-fit of the NBP model is due to two outliers, corresponding to the frequencies of earthquakes of magnitudes 7.1 and 7.8. Although NB2 (with variance function $\mu + \phi\mu^2$) and NBP (in which the variance function is estimated to be $\mu + \phi\mu^{2.5}$) produce nearly identical fits, the standard errors (in the denominators of the Pearson residuals) from the NBP fit tend to be smaller for the smaller counts, which is why the NBP, but not the NB2 diagnostic, is detecting some potential lack-of-fit of the simple log-linear model in the region of magnitudes between 7 and 8 (corresponding to relatively small frequencies).

Illustration on Simulated Datasets with Known Response Distributions. In Fig. 3 we demonstrate the NBP and NB2 empirical probability plots and Monte Carlo GOF test *p*-values on four simulated regression datasets with known response distributions. The regression structure is taken to be the estimated log-linear model from the earthquake dataset of Fig. 2, with sample size 45. For the first two scenarios, we generate NBP responses (with variance function $\mu + \phi\mu^{\alpha}$) as follows: (1) NB1, with variance 2μ and (2) NB2, with variance $\mu + 0.1\mu^2$. For the third scenario, we simulate NB2 responses as (2) above, but introduce "outliers": (3) "NB2 + Outliers", by randomly doubling three of the 45 counts. For the last one, we generate a mixture of NB2 distributions with different dispersion parameters: (4) "NB2 + Noise", with conditional variance $\mu + [0.1\exp(G)] \cdot \mu^2$, where $G \sim \mathcal{N}(0, 2^2)$. In this case the response counts are still gamma mixtures of Poissons, but the gamma variances are not constant.

In general, the two tests correctly indicate or fail to indicate lack-of-fit. An exception is that the Pearson test doesn't do as well at detecting the lack-of-fit in the last two columns. As evident in the last column of the empirical probability plots in Fig. 3, the ordered Pearson residuals are larger in magnitude than expected in some regions and smaller in others. While the ordered residuals do not seem to behave as a sample from the tested distribution, the sum of their squares is moderated by the combination of small and large magnitudes. We are particularly concerned about this cancellation aspect of the Pearson test in the extension to RNA-Seq data, in which both under- and over-dispersion (relative to the tested model) may be present in subsets of genes. This issue is not relevant to the squared vertical distance test.

Error Rates of GOF Tests in Simulations. Table 1 shows the Monte Carlo Type-I error rates for 0.05-level tests using the NB2 and NBP Monte Carlo GOF tests on 1,000 simulated samples of NB1 and NB2 response distributions. The parameter specifications are detailed in the Materials and Methods section. The standard error of simulation is approximately 0.007.

Table 1. Type-I Error Rate Evaluations.

GOF Test For	Simulated Data	n:	Sq.Vert.D.				Pear.Stat.			
			5	10	50	100	5	10	50	100
NB2	NB2		0.045	0.055	0.049	0.054	0.041	0.032	0.051	0.052
NBP	NB1		0.031	0.042	0.057	0.041	0.025	0.034	0.049	0.042
	NB2		0.040	0.034	0.056	0.060	0.044	0.030	0.047	0.044

Type-I error rates for 0.05-level NB2 and NBP GOF tests based on squared vertical distance ("Sq.Vert.D.") or Pearson statistics ("Pear.Stat."), from 1,000 simulated samples from each of several conditions. The standard error of simulation is approximately 0.007 for the Type-I error evaluations. The simulation conditions are detailed in the Materials and Methods section.

doi:10.1371/journal.pone.0119254.t001

Table 2. Rejection Rate (Power) Evaluations.

GOF Test For	Simulated Data	n:	Sq.Vert.D.				Pear.Stat.			
			5	10	50	100	5	10	50	100
NB2	NB1		0.17	0.17	0.37	0.55	0.17	0.19	0.48	0.71
	NB2 + Outliers		0.05	0.13	0.45	0.70	0.03	0.15	0.56	0.84
	NB2 + Noise		0.06	0.12	0.33	0.54	0.05	0.08	0.13	0.17
NBP	NB2 + Outliers		0.07	0.18	0.74	0.95	0.08	0.15	0.84	0.99
	NB2 + Noise		0.05	0.06	0.36	0.61	0.05	0.04	0.16	0.22

Rejection rates for 0.05-level NB2 and NBP GOF tests based on squared vertical distance ("Sq.Vert.D.") or Pearson statistics ("Pear.Stat."), from 1,000 simulated samples from each of several conditions. The maximum standard error of simulation for the power evaluations is approximately 0.016. The simulation conditions are detailed in the Materials and Methods section.

doi:10.1371/journal.pone.0119254.t002

The Type-I error rates are smaller than the nominal values for both tests at the small sample sizes. As the sample size increases, the Monte Carlo evidence is consistent with actual Type-I error rates matching the nominal values. The severity of the small-sample conservatism is slightly greater for the Pearson test than for the squared vertical distance test.

Table 2 shows the estimated statistical power of the NB2 and NBP Monte Carlo GOF tests under several alternative distributions. In the "NB2 plus noise" alternative, we add random $\mathcal{N}(0,1)$ noise to $\log(\phi)$ as described in the previous subsection, which means the data are a mixture of negative binomials with different dispersion parameters ϕ. In the "NB2 plus outliers" alternative, we randomly double 20% of the counts. The details of the generated distributions are also provided in the Materials and Methods section. The results do not indicate major power differences between the two tests, but the squared vertical distance test is more powerful in detecting the "NB2 plus noise" alternative.

Diagnostic Tools for RNA-Seq Modeling

In this section, we apply our proposed GOF test and diagnostic graphics to the analysis of an Arabidopsis RNA-Seq dataset.

The following are the GOF p-values (in parentheses) from fitting the seven dispersion models (described in the "Background/Dispersion Modeling" subsection) to a random sample of 1,000 genes: common (< 0.0001); NBP (0.04); NBQ (0.94); trended (0.21);
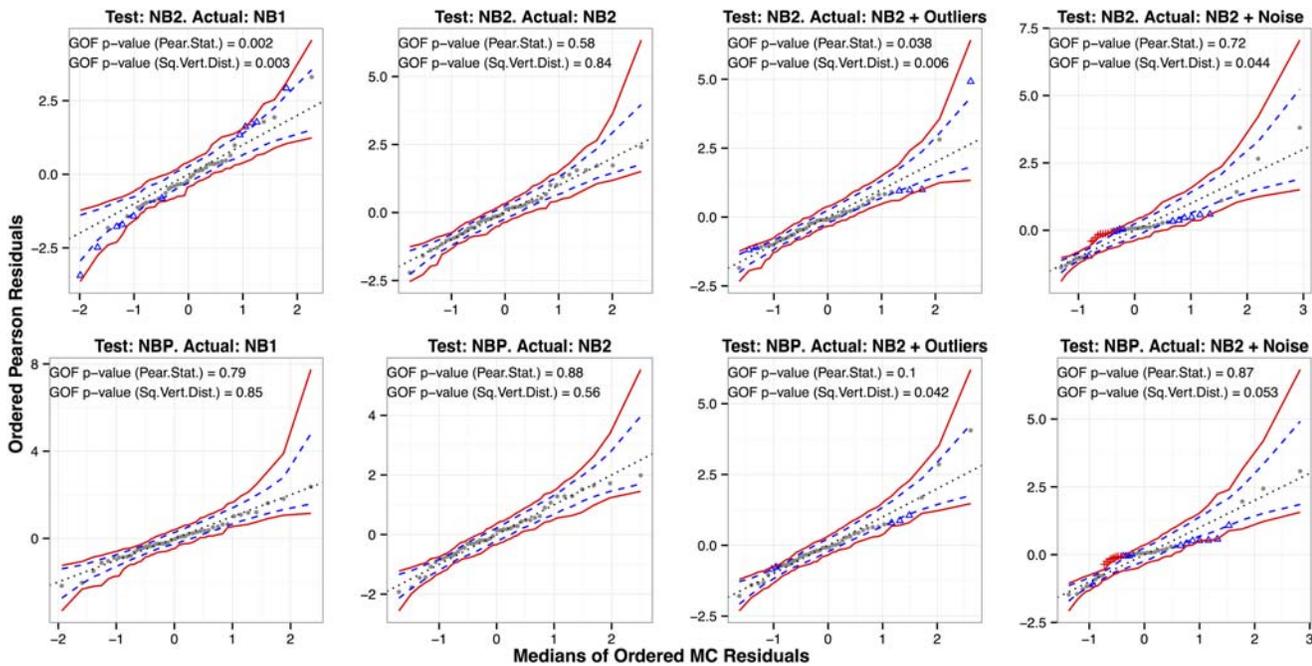
Fig 3. Empirical probability plots and GOF *p*-values for testing NB2 (top row) and NBP (bottom) on four simulated datasets with sample size = 45. The simulated response distributions are (left to right): NB1, NB2, NB2 with outliers and NB2 with random $\mathcal{N}(0, 2^2)$ noise added to $\log(\phi)$. We superimpose 95% prediction envelopes in dashed blue lines and 95% simultaneous prediction bands in solid red lines (based on 999 Monte Carlo simulations). Points outside the prediction envelope but inside the simultaneous confidence bands are flagged as blue triangles, and points outside the simultaneous confidence bands are flagged as red crosses.

doi:10.1371/journal.pone.0119254.g003

genewise (> 0.9999); tagwise-common (> 0.9999) and tagwise-trend (> 0.9999). The corresponding uniform QQ plots of *p*-values are shown in Fig. 4.

The *p*-values are greater than 0.9999 for the genewise and tagwise models. These unusually large *p*-values indicate that our GOF test may be conservative for these models. We note that for these models, the number of dispersion parameters to be estimated increases with the number of genes. In this sense, it is inherently more challenging to judge goodness-of-fit for these models. As can be seen in the uniform QQ plots of single-gene *p*-values, there are fewer small *p*-values than expected from a uniform(0,1) distribution. The extra large Fisher's *p*-value is most likely due to conservativeness in the tests from small sample sizes, as evident in Table 1, when individual (genewise and tagwise) dispersion parameters are estimated from the small sample. Even a slight degree of conservativeness in the individual NB GOF tests can produce a very small Fisher combination statistic when there are so many *p*-values being combined. The evidence from Table 1, and our experience with simulations and other RNA-Seq datasets, suggest that this conservativeness diminishes with increasing sample size.

The consistency of the data with the NBQ trend model and the non-parametric "trended" approach suggests that models without noise about the trend may be adequate. We need to exercise caution with this conclusion, though—since the lack of evidence for lack-of-fit does not prove "fit" and that the test may be conservative at this sample size. Nevertheless, the apparent fit of NBQ in conjunction with the evidence of lack-of-fit for NBP is intriguing.

Fig. 1 shows the mean-dispersion plot (log-log scale) with six fitted dispersion models (common, NBP, NBQ, trended, tagwise-common and tagwise-trend) based on the mock treatment group alone. The genewise estimates are not included since there is no implied trend associated with that method.
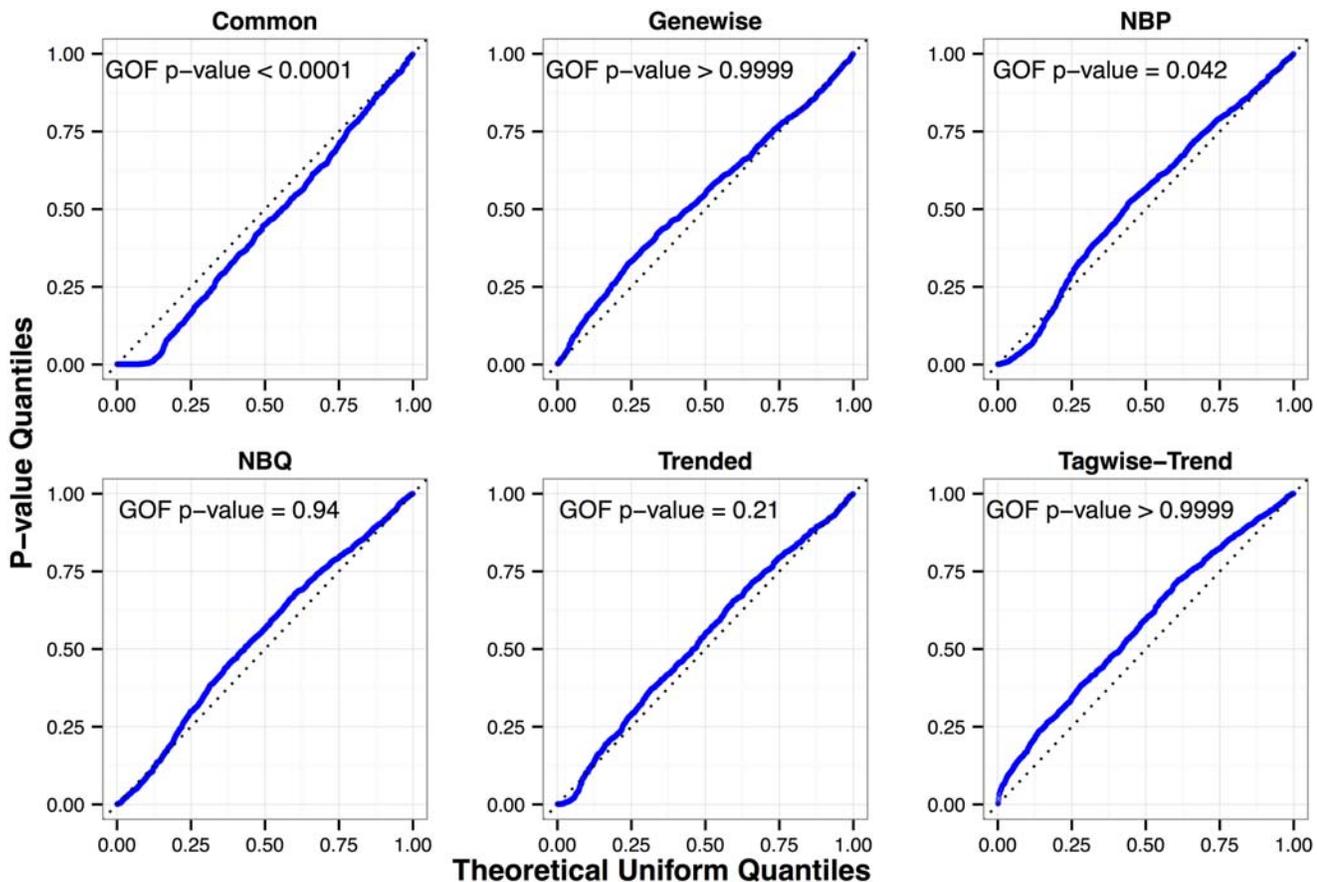
Fig 4. Uniform QQ plots of individual GOF test *p*-values for the Arabidopsis dataset. The results are based on a random sample of 1,000 genes from six experimental units in two experimental groups. The tagwise-common model (not shown) has a very similar pattern to the tagwise-trend model.

doi:10.1371/journal.pone.0119254.g004

In addition to the real RNA-Seq data analyses, we also performed simulation studies where the datasets are generated according to the "NB2 + noise" model. We then tested for GOF of the simple models and the tagwise approaches. Details on the simulation specifications and the results are provided in Supporting Information S1 File. The simulation results are as expected from the simulation setting.

Conclusion and Discussion

In this article, we proposed a simulation-based GOF test and associated graphical displays for assessing NB model adequacy for NB regression, and we showed a way to combine those tests from multiple genes or gene isoforms in RNA-Seq datasets. We believe the results may be useful for ordinary regression with count responses, but our concentration is on the RNA-Seq setting.

We are interested in the potential power and efficiency gains in inferences from NB regression fits of individual genes when we adopt a global model that reduces the number of NB nuisance parameters. In this article we proposed methodology for judging such models. It is important to understand that there are two kinds of trend models relating the NB2 dispersion parameter to the mean. In one, represented by the NBP approach, the NB2 dispersion parameter is taken to be a simple function of the mean, so that the NB2 dispersion parameter will differ

on the same gene for observations in different treatment groups if there are different expression levels in the different groups. For the trended approach and the related non-parametric approach in the `DESeq` and NBPSeq packages [6, 15], the NB2 dispersion parameter is taken to be constant for all observations on a single gene and that constant dispersion parameter is thought to be a smooth function of the average of means for that gene. It is not theoretically obvious whether the NB2 dispersion parameter should or should not be constant for a gene or, for that matter, whether the observed trend in dispersion parameter as a function of the mean is exact. We intend that the diagnostic analysis, performed on a variety of RNA-Seq datasets, will help provide an empirical clarification. The resolution of model adequacy is not, of course, the final piece of the puzzle. As in data analysis more generally, we do not expect models to fit exactly; we just need them to fit well enough for accurate and efficient inference. The diagnostic tools should help clarify models so that more comprehensive robustness and power studies can be used to compare the usefulness of the various inferential procedures upon which they are based.

The NBP model—in which the log of the NB2 dispersion parameter is a straight line function of the log of the mean—does not fully capture the trend in the RNA-Seq data we have examined. For that reason, we introduced the NBQ to allow for the next simplest model. We see evident improvement in model fit to the Arabidopsis data when the quadratic term is included. Note that the NBQ model also avoids the need for user-specified tuning parameters. Although the results for NBQ on the Arabidopsis data are intriguing, no strong generalizations about model adequacy emerge from the analysis of this single dataset or from the simulations based on the conditions of the dataset, which includes a very small sample size.

We are currently applying the diagnostic tools to a variety of RNA-Seq studies on different organisms. In that regard, we believe a useful picture emerges from the following set of diagnostic tools: (a) A plot of estimated NB2 dispersion parameter estimates with various model fits (as in Fig. 1). (b) The informal gamma log-linear regression analysis associated with that plot, including successive testing of polynomial terms and estimates of the proportion of variation in dispersion parameter estimates explained by polynomial models (as discussed in the "Background/Dispersion Modeling" subsection). (c) The NB GOF p-value from the fits to various models, such as the seven models reported in the real data example. (d) The estimate of σ in the noise model, in which the log of the NB2 dispersion parameter is the sum of a trend component (from NBQ trend or from trend estimated non-parametrically) and an individual component from a $\mathcal{N}(0, \sigma^2)$ distribution (as a measure of "noise" about the trend). In Mi and Di [28], we proposed a method for estimating σ^2 and studied the connection between magnitude of σ^2 and the performance of different DE test methods. Similar noise models had also been discussed in Wu *et al.* [29].

Finally, the goodness-of-fit of a dispersion model is only one of the factors that will affect the DE analysis. If a simpler parametric or nonparametric dispersion model shows good fit, it is a good indication that potential power can be saved in DE tests by using methods that "borrow information" between genes. When performing the DE tests, one still needs to carefully account for the uncertainty in the estimated dispersion model or dispersion parameters. The discussion of such DE test methods is beyond the scope of this paper. We refer readers to, for example, Lund *et al.* [30] and Love *et al.* [31], and corresponding software packages `Quasi-Seq` and DESeq2.

Software Information

The proposed approach is implemented as an R package named `NBGOF` (version 0.1.6, available in the Unix-like platforms) released at the first author's github page: https://github.com/

gu-mi/NBGOF, under GPL-2 License. The package also includes all datasets analyzed in this article. The R codes for reproducing all results in this article are available at the first author's github page.

Supporting Information

S1 File. Supplementary Information for Simulation Results. Description of "NB2 + Noise" Simulated Datasets and S1 Fig. (uniform QQ plots of individual GOF test *p*-values for the simulated "NB2+noise" dataset) are provided in the Supporting Information S1 File.
(PDF)

Acknowledgments

We thank Jeff H. Chang for preparing the Arabidopsis dataset, and Sarah C. Emerson for helpful discussions. This article is part of a doctoral dissertation written by the first author, under the supervision of the other two.

Author Contributions

Conceived and designed the experiments: GM YD DWS. Performed the experiments: GM YD DWS. Analyzed the data: GM. Contributed reagents/materials/analysis tools: GM YD. Wrote the paper: GM YD DWS.

References

1. Lawless JF (1987) Regression Methods for Poisson Process Data. Journal of the American Statistical Association 82: 808–815. doi: 10.1080/01621459.1987.10478502

2. Greene W (2008) Functional forms for the negative binomial model for count data. Economics Letters 99: 585–590. doi: 10.1016/j.econlet.2007.10.015

3. Hilbe JM (2011) Negative Binomial Regression. Cambridge University Press.

4. Wang Z, Gerstein M, Snyder M (2009) RNA-Seq: a revolutionary tool for transcriptomics. Nature Reviews Genetics 10: 57–63. doi: 10.1038/nrg2484 PMID: 19015660

5. Marioni JC, Mason CE, Mane SM, Stephens M, Gilad Y (2008) RNA-seq: an assessment of technical reproducibility and comparison with gene expression arrays. Genome Research 18: 1509–1517. doi: 10.1101/gr.079558.108 PMID: 18550803

6. Anders S, Huber W (2010) Differential expression analysis for sequence count data. Genome Biology 11: R106. doi: 10.1186/gb-2010-11-10-r106 PMID: 20979621

7. Di Y, Schafer DW, Cumbie JS, Chang JH (2011) The NBP Negative Binomial Model for Assessing Differential Gene Expression from RNA-Seq. Statistical Applications in Genetics and Molecular Biology 10: 1–28. doi: 10.2202/1544-6115.1637

8. Pierce DA, Schafer DW (1986) Residuals in Generalized Linear Models. Journal of the American Statistical Association 81: 977–986. doi: 10.1080/01621459.1986.10478361

9. Fisher RA (1970) Statistical Methods for Research Workers. Oliver and Boyd Edinburgh, 14th ed.

10. Robinson MD, Oshlack A (2010) A scaling normalization method for differential expression analysis of RNA-seq data. Genome Biology 11: R25. doi: 10.1186/gb-2010-11-3-r25 PMID: 20196867

11. Bullard JH, Purdom E, Hansen KD, Dudoit S (2010) Evaluation of statistical methods for normalization and differential expression in mRNA-Seq experiments. BMC Bioinformatics 11: 94. doi: 10.1186/1471-2105-11-94 PMID: 20167110

12. McCarthy DJ, Chen Y, Smyth GK (2012) Differential expression analysis of multifactor RNA-Seq experiments with respect to biological variation. Nucleic Acids Research 40: 4288–4297. doi: 10.1093/nar/gks042 PMID: 22287627

13. Robinson MD, McCarthy DJ, Smyth GK (2010) edgeR: a Bioconductor package for differential expression analysis of digital gene expression data. Bioinformatics 26: 139–140. doi: 10.1093/bioinformatics/btp616 PMID: 19910308

14. Gentleman RC, Carey VJ, Bates DM, Bolstad B, Dettling M, et al. (2004) Bioconductor: open software development for computational biology and bioinformatics. Genome Biology 5: R80. doi: 10.1186/gb-2004-5-10-r80 PMID: 15461798

15. Di Y, Schafer DW, with contributions from Jason S Cumbie, Chang JH (2014) NBPSeq: Negative Binomial Models for RNA-Sequencing Data. R package version 0.3.0. http://CRAN.R-project.org/package = NBPSeq.

16. R Core Team (2014) R: A Language and Environment for Statistical Computing. R Foundation for Statistical Computing, Vienna, Austria. URL http://www.R-project.org/.

17. Best D, Rayner J, Thas O (2009) Anscombe's Tests of Fit for the Negative Binomial Distribution. Journal of Statistical Theory and Practice 3: 555–565. doi: 10.1080/15598608.2009.10411946

18. Esnaola M, Puig P, Gonzalez D, Castelo R, Gonzalez JR (2013) A flexible count data model to fit the wide diversity of expression profiles arising from extensively replicated RNA-seq experiments. BMC Bioinformatics 14: 254. doi: 10.1186/1471-2105-14-254 PMID: 23965047

19. Atkinson A (1981) Two graphical displays for outlying and influential observations in regression. Biometrika 68: 13–20. doi: 10.1093/biomet/68.1.13

20. Landwehr JM, Pregibon D, Shoemaker AC (1984) Graphical Methods for Assessing Logistic Regression Models. Journal of the American Statistical Association 79: 61–71. doi: 10.2307/2288339

21. Svetliza CF, Paula GA (2003) Diagnostics in Nonlinear Negative Binomial Models. Communications in Statistics-Theory and Methods 32: 1227–1250. doi: 10.1081/STA-120021329

22. Garay AM, Hashimoto EM, Ortega EM, Lachos VH (2011) On estimation and influence diagnostics for zero-inflated negative binomial regression models. Computational Statistics & Data Analysis 55: 1304–1318. doi: 10.1016/j.csda.2010.09.019

23. Tusher VG, Tibshirani R, Chu G (2001) Significance analysis of microarrays applied to the ionizing radiation response. Proceedings of the National Academy of Sciences 98: 5116–5121. doi: 10.1073/pnas.091062498

24. Davison AC (1997) Bootstrap Methods and their Application, volume 1. Cambridge University Press.

25. North BV, Curtis D, Sham PC (2002) A Note on the Calculation of Empirical P Values from Monte Carlo Procedures. American Journal of Human Genetics 71: 439. doi: 10.1086/341527 PMID: 12111669

26. Cumbie JS, Kimbrel JA, Di Y, Schafer DW, Wilhelm LJ, et al. (2011) GENE-Counter: A Computational Pipeline for the Analysis of RNA-Seq Data for Gene Expression Differences. PLOS ONE 6: e25279. doi: 10.1371/journal.pone.0025279 PMID: 21998647

27. Buja A, Rolke W (2003) Calibration for Simultaneity: (Re)Sampling Methods for Simultaneous Inference with Applications to Function Estimation and Functional Data.

28. Mi G, Di Y (2015) The Level of Residual Dispersion Variation and the Power of Differential Expression Tests for RNA-Seq Data. PLOS ONE: submitted.

29. Wu H, Wang C, Wu Z (2013) A new shrinkage estimator for dispersion improves differential expression detection in RNA-seq data. Biostatistics 14: 232–243. doi: 10.1093/biostatistics/kxs033 PMID: 23001152

30. Lund S, Nettleton D, McCarthy D, Smyth G, et al. (2012) Detecting differential expression in RNA-sequence data using quasi-likelihood with shrunken dispersion estimates. Statistical Applications in Genetics and Molecular Biology 11: 8. doi: 10.1515/1544-6115.1826

31. Love MI, Huber W, Anders S (2014) Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. Genome Biology 15: 550. doi: 10.1186/s13059-014-0550-8 PMID: 25516281