Statistical Methods for Overdispersion in mRNA-Seq Count Data

Hui Zhang*, Stanley B. Pounds and Li Tang

Department of Biostatistics, St. Jude Children's Research Hospital, Memphis, TN 38105, USA

Abstract: Recent developments in Next-Generation Sequencing (NGS) technologies have opened doors for ultra high throughput sequencing mRNA (mRNA-seq) of the whole transcriptome. mRNA-seq has enabled researchers to comprehensively search for underlying biological determinants of diseases and ultimately discover novel preventive and therapeutic solutions. Unfortunately, given the complexity of mRNA-seq data, data generation has outgrown current analytical capacity, hindering the pace of research in this area. Thus, there is an urgent need to develop novel statistical methodology that addresses problems related to mRNA-seq data. This review addresses the common challenge of the presence of overdispersion in mRNA count data. We review current methods for modeling overdispersion, such as negative binomial, quasi-likelihood Poisson method, and the two-stage adaptive method; introduce related statistical theories; and discuss their applications to mRNA-seq count data.

Keywords: Count response, mRNA-seq, negative binomial theory, over-dispersion, Poisson, quasi-likelihood.

1. INTRODUCTION

DNA sequencing is a powerful technique that allows scientists to obtain key genetic information from biological system of interest. However, traditional sequencing methods such as capillary electrophoresis based Sanger sequencing have inherent limitations with regard to throughput, scalability, speed, resolution and cost. Next Generation Sequencing (NGS) has been developed to address such limitations, and this technology has triggered numerous ground-breaking discoveries and evoked a revolution in biomedical research [1]. NGS technologies enable sequencing the entire genome and transcriptome in a rapid and cost-effective manner, and can be applied to a wide range of biomedical applications such as variant discovery, profiling of histone modifications, identification of transcription factor binding sites, resequencing, and chararcterization of the transcriptome. The application of NGS to the transcriptome, also known as mRNA-seq, allows direct sequencing of a population of transcripts and these read counts linearly approximate target transcript abundance [2]. Therefore, mRNA-seq data can provide rich information on alterative splicing, allele-specific expression, unannotated exons, and differential expression.

In mRNA-seq data analysis, sequenced fragments are aligned with a reference sequence and the number of fragments (typically called *counts*) mapped to regions of interest is recorded, usually as count data [2]. However, unlike other types of discrete responses, count responses cannot be expressed in the form of several proportions. For count responses, the upper limit is infinite and the range is theoretically unbounded. Thus, methods for binomial responses do not apply. Log-linear models built upon Poisson distributions are most popular for analyzing the number of counts. An important mathematical feature of the Poisson distribution is that the mean equals its variance. However, in mRNA-seq data the variance of the count is often much larger than its mean, and this property is called overdispersion. When overdispersion occurs, the log-linear model does not hold, which results in biased and misleading conclusions. Therefore, alternative analytical strategies are needed to adequately address over dispersion of mRNA-seq data.

The regular log-linear regression and related goodness-of-fit tests will first be introduced in Section 2, followed by overdispersion and its detection in Section 3. In Section 4, we will review extended Poisson models and the newly developed two-stage adaptive strategy for overdispersion in mRNA-seq count reads. Finally, future directions will be discussed.

2. LOG-LINEAR REGRESSION MODEL FOR COUNT DATA

To assess differential continuous expression measures, such as microarray data, among various groups or treatments and other demographic factors, an ordinary linear regression can be written as follows:

$$Y_{ijg} = \mathbf{x}_i \boldsymbol{\beta}_{\mathrm{jg}} = \beta_{jg0} + \beta_{jg1} x_{i1} + \ldots + \beta_{jgp} x_{ip} + \varepsilon_{ijg}, \quad (1)$$

$$1 \le i \le n_j \text{ and } \varepsilon_{ijg} \sim N\left(0, \sigma_{jg}^2\right),$$

where Y_{ijg} denotes the expression level of gene g of the i^{th} replicate in the j^{th} treatment group and $\boldsymbol{\beta}_{jg}$ represents corresponding regression coefficients, which quantify group effects on gene expression levels. An example of $\mathbf{x}_i$ is

*Address correspondence to this author at the Department of Biostatistics, St. Jude Children's Research Hospital, 262 Danny Thomas PI, MS 768, Memphis, TN 38105, USA; Tel: 901-595-6736; Fax: 901--595-8843;
E-mail: hui.zhang@stjude.org

1874-0362/13

treatment assignment or gender. If x_{i1} represents a group indicators, β_{jg1} stands for difference in expression level between group j and the reference group. Similarly, if $\boldsymbol{x_{ip}}$ stands for gender, β_{jgp} represents the difference in the expression level of gene $\boldsymbol{g}$ between males and females. A t-test can be applied to evaluate whether a certain group effect, for example, β_{jg1} is equal to a constant C. In the case that $C = 0$, the test is the familiar one to assess the significance of the result,

$$\hat{\beta}_{jg0} \mid \beta_{jg1}, \sigma^2_{jg} \sim Normal\left(\beta_{jg1}, V_g \hat{\sigma}^2_{jg}\right) \Rightarrow T = \frac{\hat{\beta}_{jg1} - C}{\sqrt{V_g \hat{\sigma}^2_{jg}}} \overset{H_0:\beta_{jg1}=C}{\sim} t_{d_g},$$

where $V_g \hat{\sigma}^2_{jg}$ is the variance of $\hat{\beta}_{jg1}$ and t_{d_g} represents the T-distribution with degree of freedom d_g.

However, in mRNA-seq data, the expression level is often measured by count rather than a continuous value. Thus, using the model employed in (1) will likely result in a misleading conclusion. Although biomedical investigators previously transformed their count data and used tools appropriate for continuous data [3-5], it is now more suitable to employ more advanced statistical methods that better characterize the features of count data.

The Poisson distribution is a widely used model for the count response. It is determined by only one parameter λ, which is both the mean and variance of the distribution. However, this parameter often varies because of heterogenous underlying characteristics; for example, the reads for A/T-rich regions are usually low. Thus, this single-parameter model can no longer be applied if heterogeneity is high. Therefore, the log-linear regression model, an extension of the simple Poisson distribution, was developed to account for such heterogeneity [6].

Let Y_{ijg} denote the mRNA-seq count for gene g in the i^{th} replicate of the j^{th} treatment group and $\mathbf{x}_i = \left(x_{i1}, \ldots, x_{ip}\right)$ denote a vector of explanatory variables of the i th subject $\left(1 \le i \le n_j\right)$.

Given $\boldsymbol{x}_i$, the response variable Y_{ijg} follows a Poisson distribution with mean λ_{ijg},

$$Y_{ijg} \mid \mathbf{x}_i \sim \text{Poisson}\left(\lambda_{ijg}\right),\ 1 \le i \le n_j, \tag{2}$$

whereas the conditional mean λ_{ijg} is linked to a linear combination of predictors by the log function:

$$\log(\lambda_{ijg}) = \mathbf{x}_i \boldsymbol{\beta}_{jg} = \beta_{jg0} + \beta_{jg1} x_{i1} + \ldots + \beta_{jgp} x_{ip}, \tag{3}$$

where $\beta_{jg} = (\beta_{jg0}, \beta_{jg1}, \ldots, \beta_{jgp})^{\mathrm{T}}$ is the parameter vector, including interest and nuisance parameters. In the analysis of mRNA-seq data, a replicate-specific offset term is frequently used to account for replicate-to-replicate variation in technical factors such as errors and efficiency of laboratory operations such as RNA extraction and amplification [7]. Technical laboratory effects typically is manifest as differences in the total number of sequence reads generated for each replicate. Thus, the offset term is typically a simple function of the read count data of the replicate. The average, median, or upper quantile is often used [7]. Mathematically, the offset term may be absorbed into the intercept term of equation 3. Thus, we use the notation of equation 3 throughout with the understanding that the offset term is often required in practice.

Subsequently, with the log-linear model defined above, the variation of a count response can be usually partially explained by a vector of predictors.

Without loss of generality, in the rest of this review except Section 4.3, we will remove j and g by considering only 1 gene and 1 treatment group.

2.1. Inference About Model Parameters

In statistics, maximum-likelihood estimation (MLE) is used to estimate the parameters by finding optimal mathematical values that maximize the likelihood function (or the log-likelihood function). For the log-linear model, the log-likelihood function is

$$\begin{aligned} l(\beta) &= \sum_{i=1}^{n} \left\{ y_i \lambda_i - \exp\left(\lambda_i\right) - \log y_i! \right\} \\ &= \sum_{i=1}^{n} \left\{ y_i \mathbf{x}_i^T \beta - \exp\left(\mathrm{x}_i \beta\right) - \log y_i! \right\}. \end{aligned} \tag{4}$$

To locate the numbers maximizing the log-likelihood, we need to solve the score function, which sets the first derivative of the log-likelihood function with respect to β as 0:

$$\frac{\partial}{\partial \beta} l(\beta) = \sum_{i=1}^{n} \left\{ y_i \mathbf{x}_i^T - \exp\left(\mathbf{x}_i \beta\right) \mathbf{x}_i^{\mathrm{T}} \right\}. \tag{5}$$

As the second-order derivative is negative, the unique solution to the score equation (5) will be the MLE of β, denoted as $\hat{\beta}$. The MLE $\hat{\beta}$ follows an asymptotically normal distribution, that is $\hat{\beta} \sim AN\left(\beta, \frac{1}{n} \boldsymbol{I}^{-1}(\beta)\right)$, where $I(\beta)$ is the Fisher information matrix, defined as $I(\beta) = E\left[-\frac{\partial^2}{\partial \beta \partial \beta^{\mathrm{T}}} l\left(\beta \mid \mathbf{Y}\right) \right]$ [6]. For the log-linear model defined in (2) and (3),

$$E\left[I(\beta)\right] = E\left(\lambda_i \mathbf{x}_i \mathbf{x}_i^{\mathrm{T}}\right), \quad I(\beta) = \frac{1}{n} \sum_{i=1}^{n} \lambda_i \mathbf{x}_i \mathbf{x}_i^{\mathrm{T}}. \tag{6}$$

In practice, it is usually not feasible to evaluate $E\left[I(\beta)\right]$ in a closed form. Thus, the observed version of $I(\beta)$ with estimated $\hat{\beta}$ replacing β serves as a natural

estimator of $I(\beta)$ for inference. Then the MLEs can be obtained from iterative numerical optimization by taking advantage of modern computers.

In most mRNA-seq analyses, we are primarily interested in testing whether a predictor of interest is associated with the gene expression level, such as cancerous versus normal tissue. Group comparisons can be made using Wald, score, or likelihood ratio tests.

2.2. Goodness of Fit

Departures from the Poisson assumption that mean equals variance are frequently seen even in well-designed and controlled practical studies, especially in mRNA-seq count data. Therefore, it is important to evaluate whether the log-linear model fits the data appropriately. In this review, we discuss 2 goodness-of-fit tests for log-linear regression: the Pearson's χ^2 test and the deviance test.

2.2.1. Pearson's χ^2 Statistic

Pearson's χ^2 statistic is the sum of normalized squared residuals. A residual is defined as the difference between the observed and model-fitted values of the response variable, and the sum of its squares asymptotically follows a chi-square distribution under regular conditions. For instance, let y_i be the count response and $\hat{\lambda}_i = \exp\left(\mathbf{x}_i \hat{\beta}\right)$ be the fitted value under the log-linear model $(1 \le i \le n)$. Since the mean and variance are the same for the Poisson distribution, we may also estimate the variance as $\hat{\lambda}_i$. Subsequently, the normalized residual for the i th subject is $\frac{y_i - \hat{\lambda}_i}{\sqrt{\hat{\lambda}_i}}$, and the Pearson statistic is simply $P = \sum_{i=1}^{n} \frac{\left(y_i - \hat{\lambda}_i\right)^2}{\hat{\lambda}_i}$. It can be proven that the Poisson distribution converges to a normal distribution when the mean λ grows unbounded [8]. If we ignore the sampling variability in $\hat{\beta}$, we have

$$\frac{y_i - \hat{\lambda}_i}{\sqrt{\hat{\lambda}_i}} \sim AN(0,1), \quad \text{if } \hat{\lambda}_i \to \infty, \quad 1 \le i \le n.$$

It follows that for fixed n,

$$P = \sum_{i=1}^{n} \frac{\left(y_i - \hat{\lambda}_i\right)^2}{\hat{\lambda}_i} \sim asymptotic\ \chi^2_{n-p}, \quad \text{if } \hat{\lambda}_i \to \infty \text{forall} 1 \le i \le n. \quad (7)$$

where p is the number of parameters, that is the dimension of β.

Given both $y_i \sim$ Poisson(λ_i) and λ_is are all large, Pearson's statistic approximately follows a chi-square distribution with degrees of freedom of $n - p$. Note that the asymptotic distribution of Pearson's statistic holds when $\lambda_i \to \infty$ while n is fixed. This is somewhat in comparison with the common large sample theory, because most asymptotic results typically hold with a large sample size n. In mRNA-seq data, λ_is may not be large; thus, inference based on the asymptotic chi-square distribution may be invalid.

2.2.2. Deviance Statistic

The deviance statistic is defined as twice the difference between the maximum achievable log-likelihood and the value of the log-likelihood evaluated at the MLEs of the model, that is,

$$D(\mathbf{y},\theta) = 2\left[l(\mathbf{y},\mathbf{y}) - l(\mathbf{y},\theta)\right],$$

where $\mathbf{y} = (y_1,\ldots,y_n)^{\mathrm{T}}$. $l(\mathbf{y},\mathbf{y})$ is the log-likelihood assuming the model is a perfect fit to the data and $l(\boldsymbol{y},\theta)$ is the log-likelihood of the model of consideration. For log-linear regression, the deviance statistic is defined as

$$D(\mathbf{y},\theta) = 2\sum_{i=1}^{n}\left[y_i \log\left(\frac{y_i}{\hat{\lambda}_i}\right) - \left(y_i - \hat{\lambda}_i\right)\right], \quad \hat{\lambda}_i = \exp\left(\mathbf{x}_i \hat{\theta}\right).$$

If the Poisson model holds, $D(\mathbf{y},\theta)$ is approximated by a chi-squared random variable with degrees of freedom $n - p$. When the deviance divided by the degrees of freedom is significantly larger than 1, which is a rare case under the null hypothesis of good fit with Poisson, it is indicative of lack of fit.

3. OVERDISPERSION AND ITS DETECTION

Although the Poisson model is most widely employed for count data analysis, a common violation of it is overdispersion, which is seen in mRNA-seq count data. As discussed at the beginning of Section 2, the mean and variance of a Poisson distribution should be the same. This is actually a very stringent restriction, and in many applications, the variance $\sigma^2 = Var(y)$ often exceeds the mean $\lambda = E(y)$, causing overdispersion and making it inappropriate to model such data using the Poisson model. When overdispersion occurs, the standard errors of parameter estimates of the Poisson log-linear model are artificially deflated, leading to exaggerated biased estimates and false positive findings. In mRNA-seq data, overdispersion may occur because of correlated gene counts, clustering of subjects and within group heterogeneity. If the underlying reason for overdispersion is uncertain, a common approach is to use a parametric model with 1 additional parameter to account for variance inflation or use a nonparametric robust estimate for the asymptotic distribution of model estimate and make an inference based on the corrected asymptotic distribution.

Several methods have been developed to detect overdispersion, 4 of which are discussed here. First, overdispersion detection could be taken as the goodness-of-fit test to evaluate how well a Poisson model fits the data. In fact, the 2 goodness-of-fit tests, the deviance and Pearson's chi-square statistics, discussed in Section 2, can be used to verify whether overdispersion occurs. However, the deviance and Pearson's statistics can be quite different if the mean is not large enough. Second, Vuong's statistic [9], which tests the Poisson model against its nested model such as the negative binomial model, can be an efficient way to test overdispersion. Third, a recently proposed generalized ANOVA [10] detects overdispersion by extending U-statistics with asymptotic theory. Finally, for mRNA-seq data analysis, Auer and Doerge [11] adopted a score test method proposed by Dean and Lawless [12, 13]. They defined $T = \frac{1}{2}\sum_{i=1}^{n}\left(\left(y_i - \widehat{\lambda}_i\right)^2 - y_i\right)$ under the assumption of a correct specification of the mean response λ_i. This statistic is motivated by assuming a form of extra Poisson variation, $Var(y_i) = \lambda_i + \tau \lambda_i^2$, and then testing the null hypothesis of the Poisson model $H_0 : \tau = 0$. When the sample size $n \to \infty$, the following statistic

$$T_1 = \frac{\sum_{i=1}^{n}\left[\left(y_i - \widehat{\lambda}_i\right)^2 - y_i\right]}{\sqrt{2\sum_{i}^{n}\left(\widehat{\lambda}_i\right)^2}}, \quad (8)$$

approximately normally distributed under the null hypothesis that y_i follows a Poisson distribution. If the sample size n is fixed, but $\lambda_i \to \infty$ for all $1 \le i \le n$, T is asymptotically equivalent to

$$T_2 = \frac{\sum_{i=1}^{n}\left(y_i - \widehat{\lambda}_i\right)^2}{\frac{1}{n}\sum_{i=1}^{n}\widehat{\lambda}_i}.$$

Dean and Lawless [12] showed that the limiting distribution of T_2 is a linear combination of chi-squares as $\lambda_i \to \infty \ (1 \le i \le n)$.

4. METHODS FOR OVERDISPERSED COUNT DATA

When overdispersion is detected and the appropriateness of the Poisson model is in serious doubt, the model-based asymptotic variance no longer indicates the variability of the MLE and inference based on the likelihood approach may be invalid. A different and more appropriate model can be used to fit the data if the underlying cause for overdispersion is known. In previously published sequence count data analyses, multiple approaches have been employed to address overdispersion [14-18]. These overdispersion modeling methods can be divided into 3 categories: the maximum likelihood based parametric method, the maximum quasi-likelihood based nonparametric method and the two-stage analysis strategy.

4.1. Parametric Method

As mentioned in Section 2, mRNA-seq reads frequently show overdispersion because of reasons such as correlated gene counts, clustering of subjects, and within-group heterogeneity. Therefore, it is natural to add a random effect r_i to standard Poisson models to capture additive heterogeneity. Equation (2) can be rewritten as

$$Y_i \mid \mathbf{x}_i, r_i \sim \text{Poisson}\left(r_i \mu_i\right), \ 1 \le i \le n.$$

As stated above, if the random effect r_i is known, the count response y_i follows standard Poisson distribution. However, when r_i is unknown or not observed, we assume it as a random that follows a parametric distribution. Statisticians have shown that it is convenient and reasonable to assume that r_i has a gamma distribution [8]. Therefore, it is easy to derive the marginal distribution of response Y_i,

$$P(Y_i = y_i \mid \mathbf{x}_i) = \frac{\Gamma\left(\alpha + y_i\right)\theta^{\alpha}\ \mu_i^{y_i}}{y_i!\Gamma(\alpha)\ (\mu_i + \theta)^{\alpha + y_i}}, \quad (9)$$

where α and θ are the parameters of gamma distribution. Equation (9) is the probability mass function of a negative binomial distribution with mean $\frac{\alpha\theta}{\mu_i}$ and variance $\frac{\alpha\theta}{\mu_i}\left(1 + \frac{\theta}{\mu_i}\right)$, which can be rewritten as mean λ_i and variance $\lambda_i + \tau \lambda_i^2$ to match the test in (8). Unless $\tau = 0$, this variance is always larger than the mean λ. Therefore, the negative binomial model adds an extra term $\tau \lambda^2$ to the variance of Poisson model to account for overdispersion. For this reason, τ is the parameter to control dispersion or the shape of this distribution. The negative binomial model has been widely used for sequence count data analyses, such as egeR [14-16], DESeq [17] and baySeq [18].

As a parametric model belonging to the exponential family, negative binomial inference can be performed with maximum likelihood. Its log-likelihood is derived as

$$\begin{aligned} l(\beta, \alpha) &= \sum_{i=1}^{n} \log f_{NB}\left(y_i \mid \mathbf{x}_i, \beta, \tau\right) \\ &= \sum_{i=1}^{n}\left\{ y_i!\left[\log g_1^{-1}\left(\mathbf{x}_{1i}\beta\right) - \log\left(\frac{1}{\tau} + g_1^{-1}\left(\mathbf{x}_{1i}\beta\right)\right)\right]\right\} \\ &+ \sum_{i=1}^{n}\left[\tau \log\left(1 + \tau g_1^{-1}\left(\mathbf{x}_{1i}\beta\right)\right) + \log\Gamma\left(y_i + \frac{1}{\tau}\right)\right] \\ &- \sum_{i=1}^{n}\left(\log y_i! - \log\Gamma\left(\frac{1}{\tau}\right)\right). \end{aligned} \quad (10)$$

By maximizing this log-likelihood, we readily obtain the MLEs $\widehat{\beta}$, as well as their asymptotic distribution for inference.

By testing the null $H_0:\tau=0$, we can evaluate whether there is overdispersion in the data. However, since its range is nonnegative, 0 is a boundary point. Thus, the asymptotic theory of the MLE cannot be applied directly for inference about τ, as 0 is not an interior point in the parameter space. Statisticians refer to this problem as *inference under nonstandard condition*. Under certain conditions that are satisfied by most applications, including the overdispersion test being discussed here, inference about the boundary point can be based on a modified asymptotic distribution. For example, to test the null hypothesis, $H_0:\tau=0$, the revised asymptotic distribution is an equal mixture of a point mass at 0 and the positive half of the asymptotic normal distribution of $\widehat{\tau}$ under the null hypothesis. Intuitively, the lower half of the asymptotic distribution of $\widehat{\tau}$ is "folded" into a point mass at 0, since negative values of τ are not allowed under the null hypothesis.

4.2. Non-Parametric Method

When overdispersion exists and the assumptions of negative binomial are satisfied, the parametric method and related maximum likelihood are very efficient to model mRNA-seq counts. However, negative binomial may not always hold for all mRNA-seq reads and therefore a robust inference is required. Robert Wedderburn introduced a quasi-likelihood function in 1974 [19]. Despite its properties being similar to that of the log-likelihood function, it only requires the mean and variance of the response to follow a certain pattern, without having to specify any parametric distribution. Quasi-likelihood models can be fitted using a straightforward extension of the algorithms for generalized linear models.

To perform inference, the challenge is to estimate the asymptotic variance of estimated parameters. The most popular method is the sandwich variance estimate, which is derived based on the estimating equations. Scaled variance is another approach.

4.2.1. Sandwich Estimate for Asymptotic Variance

Let

$$E\left(y_i \mid \mathbf{x}_i\right)=\lambda_i\left(\mathbf{x}_i\boldsymbol{\beta}\right),\quad D_i=\frac{\partial\lambda_i}{\partial\beta},\quad B=\frac{1}{n}\sum_{i=1}^{n}D_iV_i^{-1}D_i^{\mathrm{T}},\qquad(11)$$

where $\lambda_i\left(\mathbf{x}_i\beta\right)$ indicates that the mean λ_i is an arbitrary function of $\mathbf{x}_i\beta$. Let $\widetilde{\beta}$ be the estimating equation estimate of β, i.e., $\widetilde{\beta}$ is the solution of

$$\sum_{i=1}^{n}D_iV_i^{-1}\left(y_i-\lambda_i\right)=\mathbf{0},\qquad(12)$$

where λ_i and D_i are defined in (11), and V_i is also a function of λ_i. It has been proven that $\widetilde{\beta}$ is consistent and asymptotically normal regardless of the distribution of y_i and choice of V_i, as long as the mean function $E\left(y_i \mid \mathbf{x}_i\right)=\lambda_i\left(\mathbf{x}_i\beta\right)$ is correct [16]. Further, if y_i given $\mathbf{x}_i$ is modeled parametrically using a member of the exponential family, the estimating equations in (12) are essentially the same as the score equations of the log-likelihood with an appropriate selection of V_i, which β is the same as $\widetilde{\beta}$.

With the sandwich estimate, the asymptotic variance of $\widetilde{\beta}$ is given by

$$\boldsymbol{\Phi}_{\beta}=\frac{1}{n}B^{-1}\left(\frac{1}{n}\sum_{i=1}^{n}D_iV_i^{-2}Var\left(y_i \mid \mathbf{x}_i\right)D_i^{\mathrm{T}}\right)B^{-1}.\qquad(13)$$

If the conditional distribution of y_i given x_i follows a Poisson with mean $\lambda_i=\exp\left(\mathbf{x}_i\beta\right)$, then

$$D_i=\frac{\partial\lambda_i}{\partial\beta}=\lambda_i\mathbf{x}_i,\quad Var\left(y_i \mid \mathbf{x}_i\right)=\lambda_i,\quad B=E\left(\lambda_i\mathbf{x}_i\mathbf{x}_i^{\mathrm{T}}\right).\qquad(14)$$

It is readily checked that the estimating equation in (12) is identical to the score equations of the log-likelihood of the Poisson log-linear regression in (5), if $V_i=Var\left(y_i \mid \mathbf{x}_i\right)$ and the asymptotic variance of the estimating equation estimate in (13) simplifies to

$$\begin{aligned}\boldsymbol{\Phi}_{\beta}&=\frac{1}{n}B^{-1}\left(\frac{1}{n}\sum_{i=1}^{n}D_iV_i^{-2}Var\left(y_i \mid \mathbf{x}_i\right)D_i^{\mathrm{T}}\right)B^{-1}\\&=\frac{1}{n}B^{-1}\left(\frac{1}{n}\sum_{i=1}^{n}D_iV_i^{-2}V_iD_i^{\mathrm{T}}\right)B^{-1}=\frac{1}{n}I^{-1}\left(\beta\right).\end{aligned}\qquad(15)$$

Then, the estimating equation yields the same inference as the MLE. However, as the estimating equation estimate $\widehat{\beta}$ and its associated asymptotic variance $\boldsymbol{\Phi}_{\beta}$ in (13) are derived independent of such distributional models, it still provides a valid inference even when the conditional distribution of y_i given $\mathbf{x}_i$ is not Poisson. For example, in the presence of overdispersion, $Var\left(y_i \mid \mathbf{x}_i\right)$ is larger than λ_i, biasing the asymptotic variance in (15) based on the MLE. In contrast, the estimating equation based asymptotic variance $\boldsymbol{\Phi}_{\beta}$ in (13) still provides valid inference about $Var\left(\beta\right)$.

After estimating various parameters in (13) using corresponding moments that

$$\widehat{Var}\left(y_i \mid \mathbf{x}_i\right)=\left(y_i-\widehat{\lambda}_i\right)^2,\quad \widehat{B}=\frac{1}{n}\sum_{i=1}^{n}\widehat{\lambda}_i\,\mathbf{x}_i\mathbf{x}_i^{\mathrm{T}},$$

$$\widehat{D}_i=\widehat{\lambda}_i\,\mathbf{x}_i,\quad \widehat{V}_i=\widehat{\lambda}_i,$$

we get the sandwich variance estimate

$$\widehat{\Phi}_\beta = \frac{1}{n}\widehat{B}^{-1}\left(\frac{1}{n}\sum_{i=1}^{n}\widehat{D}_i\widehat{V}_i\widehat{Var}\left(y_i \mid \mathbf{x}_i\right)\widehat{D}_i\right)\widehat{B}^{-1}$$

$$= \frac{1}{n}\left(\frac{1}{n}\sum_{i=1}^{n}\widehat{\lambda}_i\,\mathbf{x}_i\mathbf{x}_i^{\mathrm{T}}\right)^{-1}\left(\frac{1}{n}\sum_{i=1}^{n}\widehat{\lambda}_i^{2}\,\widehat{Var}\left(y_i \mid \mathbf{x}_i\right)\mathbf{x}_i\mathbf{x}_i^{\mathrm{T}}\right)\left(\frac{1}{n}\sum_{i=1}^{n}\widehat{\lambda}_i\,\mathbf{x}_i\mathbf{x}_i^{\mathrm{T}}\right)^{-1}. \quad (16)$$

Thus, if overdispersion is detected, we can still use the MLE to estimate β, but need to switch to the sandwich variance estimate in (16) to ensure valid inference.

4.2.2. Scaled Variance

For the log-linear model, a widely used alternative to address overdispersion is to add an additional scale parameter to inflate the variance of y_i. Specifically, this scaled variance approach assumes the same conditional mean but a scaled conditional variance of y_i given $\mathbf{x}_i$ as following:

$$\lambda_i = \exp\left(\mathbf{x}_i\beta\right), \quad Var\left(y_i \mid \mathbf{x}_i\right) = \lambda^2\lambda_i.$$

If $\lambda^2 = 1$, $Var\left(y_i \mid \mathbf{x}_i\right) = \lambda_i$ and the modified variance approach reduces to the Poisson model. In the presence of overdispersion, $\lambda^2 > 1$ and $Var\left(y_i \mid \mathbf{x}_i\right) > \lambda_i$, accounting for overdispersion.

Under this scaled-variance approach, we first estimate β using either the MLE or the estimating equation approach. Then, we estimate the scale parameter λ^2 by

$$\widehat{\lambda}^{2} = \frac{1}{n}\sum_{i=1}^{n} r_i^2, \quad r_i = \sqrt{\widehat{\lambda}_i}\left(y_i - \widehat{\lambda}_i\right).$$

where r_is are known as the *Pearson* residuals. By substituting $\widehat{Var}\left(y_i \mid \mathbf{x}_i\right)$ with $\widehat{\lambda}^{2}\widehat{\lambda}_i$ in the sandwich variance estimate $\widehat{\Phi}_\beta$ in (16), we obtain a consistent estimate of the asymptotic variance of the estimating equation (or maximum likelihood) estimate of β:

$$\widetilde{\Phi}_\beta = \left(\sum_{i=1}^{n}\widehat{\lambda}_i\,\mathbf{x}_i\mathbf{x}_i^{\mathrm{T}}\right)^{-1}\left(\sum_{i=1}^{n}\widehat{\lambda}^{2}\,\widehat{\lambda}_i\,\mathbf{x}_i\mathbf{x}_i^{\mathrm{T}}\right)\left(\sum_{i=1}^{n}\widehat{\lambda}_i\,\mathbf{x}_i\mathbf{x}_i^{\mathrm{T}}\right)^{-1}$$

$$= \widehat{\lambda}^{2}\left(\sum_{i=1}^{n}\widehat{\lambda}_i\,\mathbf{x}_i\mathbf{x}_i^{\mathrm{T}}\right)^{-1}. \quad (17)$$

Alternatively, we can apply the deviance statistic to estimate λ^2 in (17) to obtain a slightly different yet consistent estimate of the asymptotic variance of the estimating equation/maximum likelihood estimate of β.

However, unlike the sandwich estimate $\widehat{\Phi}_\beta$ in (16), the estimate $\widetilde{\Phi}_\beta$ is derived based on a particular variance model for overdispersion. If this variance model is incorrectly specified, inference based on this asymptotic variance estimate $\widetilde{\Phi}_\beta$ may not be reliable.

4.3. Two-Stage Analysis

Typically, an mRNA-seq dataset can have thousands to millions of genetic features (exons, genes, or other genetic loci to which sequences are mapped). Some genes may have substantial variation in transcription levels, whereas others, such as housekeeping genes, may have very stable transcription levels, with no overdispersion. Therefore, a Two-Stage Poisson Model (TSPM) was proposed to first examine overdispersion in each individual gene before making inferences in order to preserve high statistical power, especially for small sample sizes [12].

After estimating the offset terms, a nonspecific filtering is performed to remove genes with total counts less than 10 across all treatment groups (a pre-specified arbitrary cutoff) to satisfy the conditions required by asymptotic theory in Section 3. Then each gene that passes the filtering process, such as gene g, is evaluated first for overdispersion by using the Dean and Lawless' method [12], introduced in Section 3, with the hypothesis set as

$$H_{0g}: \tau_g = 0 \qquad \text{v.s.} \qquad H_{1g}: \tau_g > 0,$$

and with the statistic

$$T_g = \frac{1}{2N}\left[\sum_{i,j}\frac{\left[\left(y_{ijg} - \widehat{\lambda}_{ijg}\right)^2 - y_{ijg} + \widehat{h}_{ijg}\,\widehat{\lambda}_{ijg}\right]}{\widehat{\lambda}_{ijg}}\right]^2 \overset{H_0}{\sim} \chi_1^2, \quad (18)$$

where $\widehat{h}_{ijg}$ is the corresponding diagonal element of the hat matrix (i.e. $H = X\left(X^TX\right)^{-1}X^T$ and X is the covariate vector). At the second stage, if the H_{0g} is rejected, suggesting that gene g has overdispersion, a quasi-likelihood method will be employed. Otherwise, a standard log-linear model will be fitted which has a greater statistical power than the quasi-likelihood method, since the MLE is most efficient when the parametric model holds.

Recently, Pounds *et al*. [20] further extended the TSPM by using concepts of the Assumption Adequacy Averaging method [21]. Briefly, Dean's test for overdispersion, a test based on a standard Poisson general linear model, and a quasi-likelihood test are applied to the count data of each genetic feature. A set of empirical Bayesian probabilities [22] is computed for each test and these empirical Bayesian probabilities are either combined in a weighted average or used to select the empirically best test for each genetic feature. These intermediate results are used to obtain a final empirical Bayesian probability that is a type of local false discovery rate metric for each genetic feature. The approach by Pounds *et al*. performs better than TSPM [20], but it

cannot be applied to a gene with small counts due to its dependence on asymptotic theories for detecting overdispersion [12].

5. FUTURE STUDIES

Uniformly applying the negative binomial or the quasi-likelihood Poisson model to mRNA-seq count data addresses the overdispersion and fixes the bias resulting from standard Poisson models. However, it also reduces the analysis power for genes that do not exhibit overdispersion. The two-stage analysis strategy seems to address this limitation, but it can be used only when genes have large counts. Therefore, improvements in statistical methods are warranted, especially for genes with lower counts.

Another challenge is that many mRNA-seq data carry excessive zeroes, while statistical models described in this review generally no longer hold. Also, a widely recognized practical problem is to distinguish between a *random zero,* occurring in *a gene with low expression*, from a *structural zero,* occurring in an *unexpressed gene*. The zero-inflated models may be useful when there is an excess of zeroes in mRNA-seq data [23]. These models introduce an additional parameter to model the probability that a gene is unexpressed, which serves as a mathematical representation for structural zeroes whereas random zeroes reflect genes with low expressions.

mRNA-seq data have posed numerous statistical challenges, and translating the rich information derived from mRNA-seq data into clinical knowledge will be a long-standing direction for biomedical and statistical research. This review aims to aid researchers to analyze regular and overdispersed mRNA-seq count data more appropriately and direct attention to practical issues in such data.

CONFLICT OF INTEREST

The author confirms that this article content has no conflicts of interest.

ACKNOWLEDGEMENTS

We gratefully acknowledge funding from the American Lebanese Syrian Associated Charities (ALSAC).

REFERENCES

[1] S.C. Schuster, "Next-generation sequencing transforms today's biology", *Nat. Methods,* vol. 5, pp.16-18, 2008.

[2] A. Mortazavi, B.A. Williams, K. McCue, L. Schaeffer, and B. Wold, "Mapping and quantifying mammalian transcriptomes by RNA-Seq", *Nat. Methods.,* vol. 7, pp. 621-628, 2008.

[3] N. Cloonan, A.R. Forrest, G. Kolle, B.B. Gardiner, G.J. Faulkner, M.K. Brown, D.F. Taylor, A.L. Steptoe, S. Wani, G. Bethel, A.J. Robertson, A.C. Perkins, S.J. Bruce, C.C. Lee, S.S. Ranade, H.E. Peckham, J.M. Manning, K.J. McKernan, and S.M. Grimmond, "Stem cell transcriptome profiling *via* massive-scale mRNA sequencing", *Nat. Methods,* vol. 5, pp.613-619, 2008.

[4] B. Langmead, K.D. Hansen and J.T. Leek, "Cloud-scale RNA-sequencing differential expression analysis with Myrna", *Genome. Biol.*, vol. 11, pp. 83, 2010.

[5] M. Ashburner, C.A. Ball, J.A. Blake, D. Botstein, H. Butler, J.M. Cherry, A.P. Davis, K. Dolinski, S.S. Dwight, J.T. Eppig, M.A. Harris, D.P. Hill, L. Issel-Tarver, A. Kasarskis, S. Lewis, J.C. Matese, J.E. Richardson, M. Ringwald, G.M. Rubin and G. Sherlock, "Gene ontology: tool for the unification of biology. The Gene Ontology Consortium", *Nat. Genet.,* vol. 25, pp.25-29, 2000.

[6] P. McCullagh and J.A. Nelder, *Genaralized Linear Models*. Chapman & Hall/CRC, 1999.

[7] J.H. Bullard, E. Purdom, K.D. Hansen and S. Dudoit, "Evaluation of statistical methods for normalization and differential expression in mRNA-Seq experiments", *BMC Bioinformatics.,* vol. 11, pp. 94-107, 2010.

[8] G. Casella and R.L. Berger, *Statistical Inference*, 2nd ed. Thomson Learning: Pacific Grove, CA, 2002.

[9] Q.H. Vuong, "Likelihood Ratio Tests for Model Selection and non-nested Hypotheses", *Econometrica,* vol. 57, pp.307-333, 1989.

[10] H. Zhang and X.M. Tu, "The Generalized ANOVA -- A classic song sung with modern lyrics", In press by *Institute of Mathematical Statistics Collection Series,* 2013.

[11] P.L. Auer and R. Doerge, "A two-stage Poisson model for testing RNA-seq data", *Statis. Appl. Genet. Mol. Biol.,* vol. 10, pp.1-26, 2011.

[12] C.B. Dean and J.F. Lawless, "Tests for detecting overdispersion in poisson regression models", *J. Am. Stat. Assoc.,* vol. 84, pp.467-472, 1989.

[13] C.B. Dean, "Testing for overdispersion in poisson and binomial regression models", *J. Am. Stat. Assoc.,* vol. 87, pp.451-457, 1992.

[14] M.D. Robinson and A. Oshlack, "A scaling normalization method for differential expression analysis of RNA-seq data", *Genome. Biol.,* vol. 11, pp.1-9, 2010.

[15] M.D. Robinson and G.K. Smyth, "Moderated statistical tests for assessing differences in tag abundance", *Bioinformatics,* vol. 23, pp. 2881-2887, 2007.

[16] M.D. Robinson and G.K. Smyth, "Small-sample estimation of negative binomial dispersion, with applications to SAGE data", *Biostatistics,* vol. 9, pp.321-332, 2008.

[17] S. Anders and W. Huber, "Differential expression analysis for sequence count data", *Genome Biol.,* vol. 11, article 106, 2010.

[18] T.J. Hardcastle and K.A. Kelly, "baySeq: empirical Bayesian methods for identifying differential expression in sequence count data", *BMC. Bioinformatics,* vol. 11, pp. 422, 2010.

[19] R.W.M. Wedderburn RWM, "Quasi-likelihood functions, generalized linear models, and the Gauss---Newton method", *Biometrika,* vol. 61, pp.439-447, 1974.

[20] S.B. Pounds, C.L. Gao and H. Zhang, "Empirical bayesian selection of hypothesis testing procedures for analysis of sequence count expression data", *Stat. Appl. Genet. Mol. Biol.,* vol. 11, article 5, 2012.

[21] S.B. Pounds and S.N. Rai, "Assumption adequacy averaging as a concept to develop more robust methods for differential gene expression analysis", *Comput.l Stat. Data Anal.,* vol. 53, pp.1604-1612, 2009.

[22] S.B. Pounds and S. Morris, "Estimating the occurrence of false positives and false negatives in Microarray studies by approximating and partitioning the empirical distribution of p-values", *Bioinformatics,* vol. 19, pp. 1236-1242, 2003.

[23] D. Lambert, "Zero-Inflated poisson regression, with an application to defects in manufacturing", *Technometrics,* vol. 34, pp. 1-14, 1992.

Received: August 6, 2013 Revised: September 6, 2013 Accepted: September 15, 2013