2018

Zero-Inflated Models for RNA-Seq Count Data

Morshed Alam
University of Nebraska Medical Center, morshed.alam@unmc.edu

Naim Al Mahi
University of Cincinnati

Munni Begum
Ball State University

Journal of Biomedical Analytics
Vol. 1, No. 2 (2018), pp. 55–70
doi:10.30577/jba.2018.v1n2.23

Research Article

Zero-Inflated Models for RNA-Seq Count Data

Morshed Alam[1], Naim Al Mahi[2], and Munni Begum[3,*]

[1]Department of Biostatistics, University of Nebraska Medical Center (UNMC), USA
[2]Department of Environmental Health, University of Cincinnati, USA
[3]Department of Mathematical Sciences, Ball State University, USA
[*]Corrresponding author: mbegum@bsu.edu

Received: May 8, 2018; revised: August 24, 2018; accepted: September 15, 2018.

Abstract: One of the main objectives of many biological studies is to explore differential gene expression profiles between samples. Genes are referred to as differentially expressed (DE) if the read counts change across treatments or conditions systematically. Poisson and negative binomial (NB) regressions are widely used methods for non-over-dispersed (NOD) and over-dispersed (OD) count data respectively. However, in the presence of excessive number of zeros, these methods need adjustments. In this paper, we consider a zero-inflated Poisson mixed effects model (ZIPMM) and zero-inflated negative binomial mixed effects model (ZINBMM) to address excessive zero counts in the NOD and OD RNA-seq data respectively in the presence of random effects. We apply these methods to both simulated and real RNA-seq datasets. The ZIPMM and ZINBMM perform better on both simulated and real datasets.

Keywords: RNA-seq, differential expression, zero inflated Poisson mixed effect model, zero inflated negative binomial mixed effect model, over-dispersed count data

1 Introduction

Deoxyribonucleic Acid (DNA) and Ribonucleic Acid (RNA) play a fundamental role in carrying genetic information in all living organisms on earth. RNA directly codes for amino acids and acts as a messenger between DNA and ribosomes to make proteins while DNA stores genetic information. RNA sequence unveils biological insights and characteristics of biological subjects. RNA-sequencing, also termed as "whole transcriptome shotgun sequencing", is a high-throughput sequencing technology for extracting information on RNA content from the sample by generating millions of short sequence reads. These reads are mapped or aligned

to a reference genome and the number of mapped reads within a gene is used as an approximate measure of gene expression (Oshlack *et al.*, 2010). One of the key purposes of RNA-seq experiments is to identify DE genes by comparing the expression measurements between two or more treatment conditions. DE genes have great biological importance since variation in gene expression levels may indicate a major source of evolutionary novelty. Specific types of gene expression may be responsible for certain types of diseases for both animals and plants and discriminate among diseased and healthy tissues in those species. Thus knowing DE genes in diverse conditions can help scientists to infer about the association of genes with certain diseases or conditions. Moreover, DE genes would facilitate developing the so called personalized medicine guided by individuals genetic mark-up.

A simple and commonly used RNA-seq study includes two treatment conditions in a completely randomized design, for example, treated versus untreated cells. Identification of DE genes unveils the complex functions of genes when cells respond to antithetical treatment conditions. To the latest, a number of methods have been suggested in the literature to detect DE genes from RNA-seq count data. Among these, binomial, Poisson, and negative binomial (NB) are the three extensively used discrete probability distributions to model the RNA-seq count data. When the number of read counts mapped to a given feature of interest is relatively small to the total number of reads, Poisson approximation to binomial distribution can be used (Kvam *et al.*, 2012). In the past RNA-seq studies with a single biological sample, Poisson models are considered as a good fit to the data (Marioni *et al.*, 2008; Bullard *et al.*, 2010). A number of methods namely, Fisher's exact test (Bloom *et al.*, 2009), Chi-square goodness-of-fit test (Marioni *et al.*, 2008) and likelihood ratio test (LRT) (Bullard *et al.*, 2010) were proposed based on Poisson distribution. A Poisson mixture model is used to identify DE genes if counts come from several distinct subpopulations (clusters) (Balzergue *et al.*). A Poisson mixed effect model is a useful tool for identification of DE genes (Blekhman *et al.*, 2010) in the presence of fixed and random effects.

In the presence of biological replications, RNA-seq data typically violate the equality of mean variance property for a Poisson model. If the observed variance exhibits lower value than the mean counts for some genes, then those genes are referred to as under-dispersed. Whereas if the observed variance is higher than the mean counts for some genes, then those genes are referred to as over-dispersed (Anders and Huber, 2010). In such cases, Poisson models are likely to produce high false positives. Some of the commonly used approaches such as, Bayesian methods (Baggerly *et al.*, 2004; Vêncio *et al.*, 2004), generalized linear models (Baggerly *et al.*, 2003; Srivastava and Chen, 2010) including NB models (Anders and Huber, 2010; Robinson and Smyth, 2007, 2008) assume that all gene counts are derived from an over-dispersed distribution, and fail to address the fact that some genes might have constant levels of transcription within treatment groups (Auer and Doerge, 2011; Oshlack *et al.*, 2010). Accordingly, the assumptions on the variations of each gene's expression can mislead the detection of a truly DE gene. Besides, excessive number of zero counts in the data is another situation when usual methods may not work well. With excessive number of zeros, standard Poisson or NB model is not a good fit to the count data. Zero inflated models (Zhou *et al.*, 2018; Zhang *et al.*, 2017; Chen and Li, 2016; Choi *et al.*, 2017) work as effective tools for identifying DE genes from the pool of genes with excessive number of zero counts.

In this study, we focus on the application of zero inflated Poisson mixed effects model (ZIPMM) and zero inflated negative binomial mixed effects model (ZINBMM) to identify DE genes from non-over-dispersed (NOD) and over-dispersed (OD) genes with excessive

number of zero counts respectively. The study also evaluates the performance of the models on real and simulated datasets. We use a number of *R* packages to analyze these data including *Poisson-Seq* (Li *et al.*, 2011), *GPseq* (Srivastava and Chen, 2010) , *edgeR* (Robinson and Smyth, 2007, 2008), *DESeq* (Anders and Huber, 2010), *sSeq* (Yu *et al.*, 2013), *NBPSeq* (Di *et al.*, 2011), *lme4* (Bates *et al.*, 2014), and *HTSDiff* (Balzergue *et al.*). Most of these packages are available in the Bioconductor web site: http://bioconductor.org (Gentleman *et al.*, 2004).

1.1 Data Description

In order to compare gene expression patterns and exon usage across sexes within and between species, Blekhman *et al.* (2010) used RNA-seq technology to extract transcript levels in humans, chimpanzees, and rhesus macaques, using liver RNA samples from three males and three females from each species. RNA from each sample were prepared for high-throughput Illumina sequencing technology. Each sample was sequenced using two lanes of the Genome-Analyzer II (GA2), producing a total of 36 lanes. After the preprocessing steps of alignment, summarizing, and quality control, digital gene expression (DGE) count data contain the number of reads in 36 columns extracted from 18 liver tissue samples each with 2 replications. The authors considered 6 samples from each of the three species namely human (HS), chimpanzee (PT), and rhesus macaque (RM). From each species, they collected information on 3 males and 3 females. Details on samples, data collection and associated protocols are available in method section as well as supplemental Tables S1 and S2 and Figure S1 of Blekhman *et al.* (2010).

An illustration of the experimental design as given in the supplementary materials of Blekhman *et al.* (2010) is reproduced below in Figure.1.

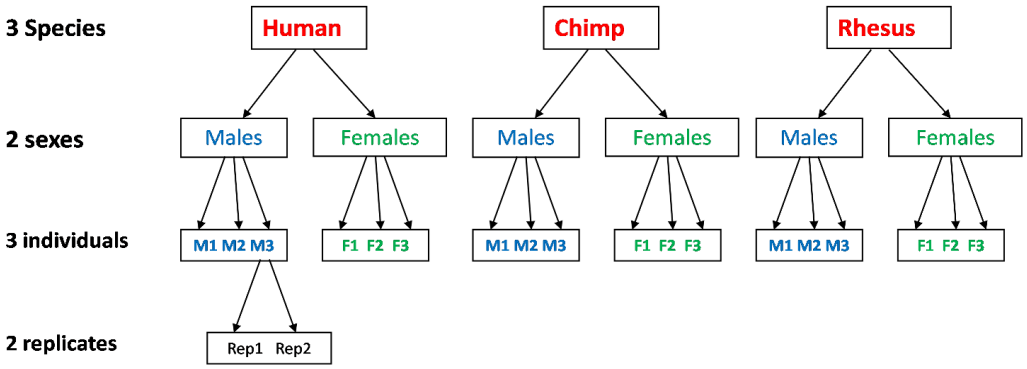

Figure 1: Graphical representation of the experimental design (Blekhman *et al.*, 2010).

Rest of the paper is organized as follows: section 2 presents the methods for applying appropriate statistical models to identify differentially expressed genes across sexes within and between species in the presence of limited and excessive number of zeros. Section 3 presents results and discussion and section 4 concludes the paper.

2 Methods

The count matrix from (Blekhman *et al.*, 2010) is arranged as follows: columns 1-12 contain read counts on 3 males and 3 females with 2 replications from HS. Next 13-24 and 25-36 columns contain read counts from PT and RM respectively. Out of a total of 20,689 genes, we exclude 2,803 genes having zero read counts in every cell. Based on over-dispersion test (section 2.2), data are divided into two sets. We refer to these datasets (i) Poisson with NOD genes and NB with OD genes.

We check the distribution of zero counts for each gene in the overall data set and in OD and NOD data sets. We found that for a notable number of genes, the number of zero counts was substantially high. Initially, we fit a Poisson mixed effects model (PMM) on NOD data set and a negative binomial mixed effects model (NBMM) on OD data set by using *glmer* function of *R* package *lme4*, but encountered frequent interruptions in fitting these models. Existing literature suggests (Auer and Doerge, 2011; Mahi and Begum, 2016) that standard Poisson and NB models are not viable methods for data with excessive number of zeros. We adopted an empirical approach to check the candidacy for zero-inflated models by examining the error rates in the presence of 1/3, 1/2, and 3/4 of zeros in the data. Both models exhibit high frequency of errors when the proportions of zeros were 1/2 and 3/4. When the proportion of zeros was 1/3 or less, the error rate was almost negligible for both models. This leads us to split the data by setting a threshold of proportion of zeros as 1/3. Genes containing zero counts more than 33.33% was considered as zero inflated data set for the rest of the study. Finally, the data set was split up into four parts and we refer to these as: Poisson data, NB data, zero inflated Poisson (ZIP) data, and zero inflated negative binomial (ZINB) data.

2.1 Statistical Models for Identifying DE Genes

For each gene, data contain read counts from two replicates of males and females from three species. Thus the individuals are nested within species. To identify DE genes with respect to species, sex, and their interactions, we apply Poisson mixed effects model (PMM) for NOD genes with limited number of zero counts, and zero-inflated Poisson mixed effects model (ZIPMM) for genes with excessive number of zero counts. For OD data set, we apply negative binomial mixed effects mdoel (NBMM) and zero-inflated negative binomial mixed effects model (ZINBMM) based on limited and excessive number of zero counts respectively. In all the models, we consider species and sex as fixed effects, and the variability due to individuals as random effect. The zero inflated Poisson (ZIP) model can be expressed as:

$$Y_i = \begin{cases} 0, & \text{with probability } \pi_i + (1-\pi_i)e^{-\mu_i} \\ k, & \text{with probability } (1-\pi_i)e^{-\mu_i} \end{cases},$$

where, $k = 1, 2, 3, ...$ and π represents the probability of inflated zeros and μ is the mean parameter. Similarly the zero inflated negative binomial (ZINB) model can be expressed as:

$$P(Y_i = y_i) = \begin{cases} \pi_i + (1-\pi_i)(1+\tau\mu_i)^{-\frac{1}{\tau}}; & y = 0 \\ (1-\pi_i)\dfrac{\Gamma(y_i+\frac{1}{\tau})(\tau\mu_i)^{y_i}}{\Gamma(y_i+1)\Gamma(\frac{1}{\tau})(1+\tau\mu_i)^{y_i+\frac{1}{\tau}}}; & y = 1, 2, 3, ... \end{cases}$$

where π is the probability of inflated zeros, μ is the mean, and τ is the dispersion parameter.

We denote the count response for the gth gene of ith species, jth sex, kth individual in lth replicate by $Y_g^{ijkl}(i = 1, 2, 3; j = 1, 2; k = 1, 2, 3; l = 1, 2)$. Y_g^{ijkl} follows four separate distributions under four distinct scenarios as given below:

$$Y_g^{ijkl} \sim \begin{cases} Poisson(\mu_g^{ijk}), & \text{for NOD genes with limited zeros} \\ NB(\mu_g^{ijk}, \tau_g), & \text{for OD genes with limited zeros} \end{cases}$$

$$Y_g^{ijkl} \sim \begin{cases} ZIP(\mu_g^{ijk}, \pi_g^{ijkl}), & \text{for NOD genes with excessive zeros} \\ ZINB(\mu_g^{ijk}, \pi_g^{ijkl}, \tau_g), & \text{for OD genes with excessive zeros} \end{cases}$$

Under each scenario a mixed model $log(\mu)$ can be expressed as:

$$log(\mu_g^{ijk}) = \mu_g + \theta_g^i + \delta_g^j + \gamma_g^{k(i)} + (\theta\delta)_g^{ij} \tag{1}$$

Where, the superscript i indicates species with 1=HS, 2=PT, and 3=RM. Superscript j indicates sex with 1= Male, and 2= Female; and superscript $k(i)$ represents k^{th} individual nested within i^{th} species. In the model for gth gene, μ_g is the overall gene expression across all individuals, θ_g^i is the species specific fixed effect, δ_g^j is the sex specific fixed effect, $\gamma_g^{k(i)}$ is the per individual random effect, and $(\theta\delta)_g^{ij}$ is the sex-by-species interaction effect. The following sets of hypothesis are tested:

$$\begin{cases} H_0 : \theta_g^i = 0, \delta_g^j \neq 0, (\theta\delta)_g^{ij} = 0 \\ H_1 : \theta_g^i \neq 0, \delta_g^j \neq 0, (\theta\delta)_g^{ij} = 0 \end{cases}$$

$$\begin{cases} H_0 : \theta_g^i \neq 0, \delta_g^j = 0, (\theta\delta)_g^{ij} = 0 \\ H_1 : \theta_g^i \neq 0, \delta_g^j \neq 0, (\theta\delta)_g^{ij} = 0 \end{cases}$$

$$\begin{cases} H_0 : \theta_g^i \neq 0, \delta_g^j \neq 0, (\theta\delta)_g^{ij} = 0 \\ H_1 : \theta_g^i \neq 0, \delta_g^j \neq 0, (\theta\delta)_g^{ij} \neq 0 \end{cases}$$

By the first set of hypotheses, we test whether a particular gene is DE across species (HS versus PT, and HS versus RM). The second set of hypotheses allows us to detect DE genes with respect to sex such as aggregate male versus aggregate female for all three species. Finally, by the third set, we test DE with respect to sex-species interactions. Due to design constraint, we test only the interactions: HS-Male versus PT-Female and HS-Male versus RM-Female. We apply the same model and test the same hypotheses for all four types of our data where we fit PMM, ZIPMM, NBMM, and ZINBMM models. For all the tests, the level of significance is considered as 5%. Because we are testing thousands of genes, we take into account the false discovery rate (FDR) which is the expected proportion of false positives among the rejected null hypotheses. To control FDR, we compute Benjamini-Hochberg (BH) (Benjamini and Hochberg, 1995) adjusted *P*-values. Normalization is another important pre-processing step in RNA-seq data analysis while investigating DE. In order to produce a comparable data across different treatment groups, we use trimmed mean of M-values (TMM) (Robinson *et al.*, 2010) as a normalization method in this study.

All computations are carried out under **R** environment. The package *DCluster* is used for over dispersion test, *lme4* for PMM and NBMM, *glmmADMB* for ZIPMM and ZINBMM models, *VennDiagram* for drawing Venn diagrams, *gplots* and *RcolorBrewer* for heatmaps, *gamlss.dist*, *gamlss*, and *ROCR* are considered for drawing receiver operating characteristic (ROC) curves.

We consider zero inflated models for DE genes identification with multiple treatment conditions (3 species here) at a time. In order to evaluate the performance of these models, we construct heatmaps for a subset of DE genes identified by the zero inflated models. Venn diagrams are created to identify the common DE genes across species and sex. Finally, performance of the zero inflated models are evaluated with a simulation study. Overall computational work-flow for the data manipulation is presented pictorially in Figure 2.

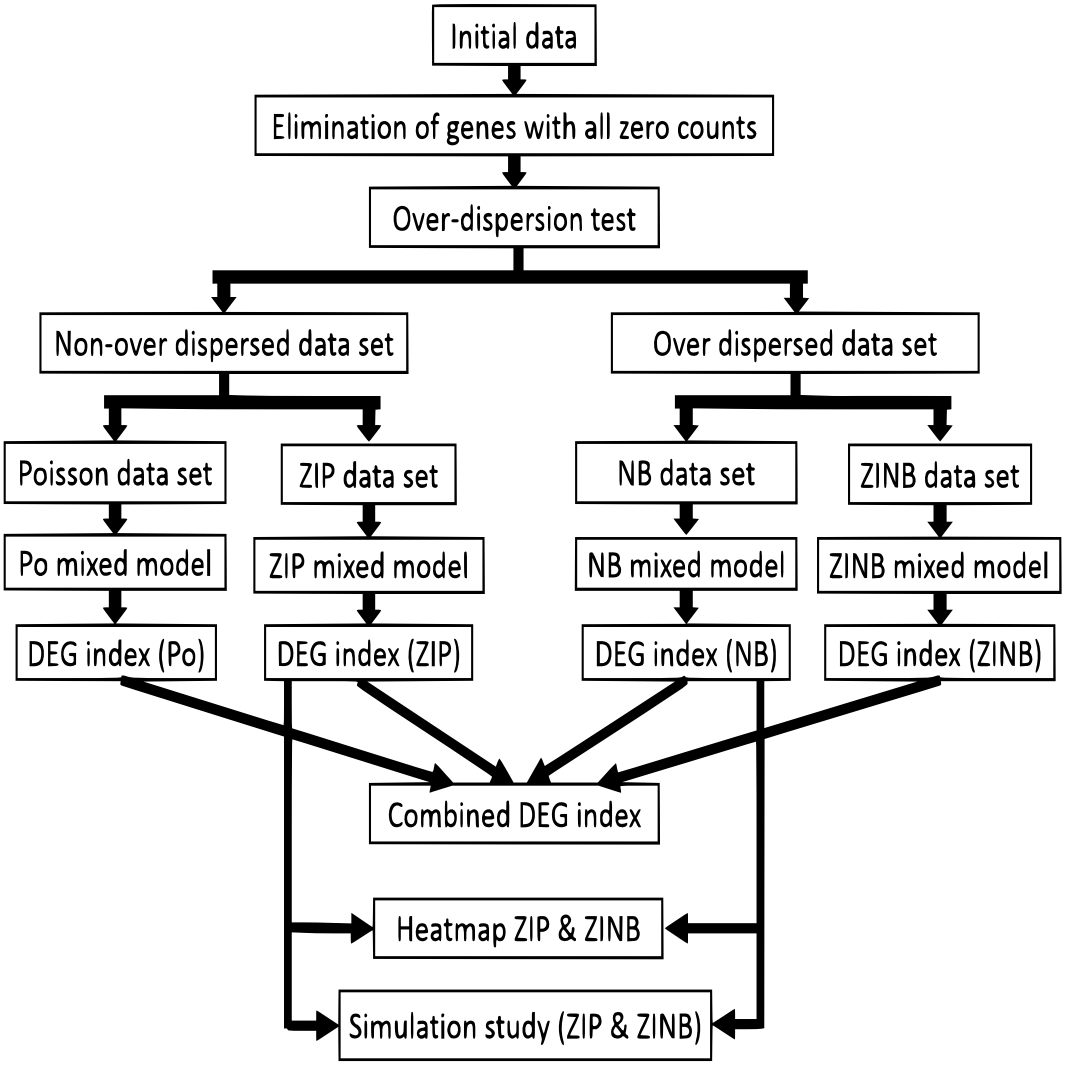

Figure 2: Computational work-flow of the study.

2.2 Test for Over-dispersion

Poisson model is a natural choice for count response. However, read counts generated from multiple biological replications tend to be over-dispersed. Over-dispersion arises when variance of count data exceeds the mean. In the presence of over-dispersion, Poisson model does

not perform well for DE analysis (Di *et al.*, 2011). In our test for over-dispersion, Poisson is the null model against any other alternative model with over-dispersion. Following the notations of Deans and Lawless (Dean and Lawless, 1989), we let Y_i be the response from the ith subject with covariates X_i. Then Y_i is distributed as Poisson with mean $\mu_i = \mu_i(X_i, \beta)$, where β is a *p*-dimensional vector of unknown coefficients. We denote the possible extra-Poisson variation by v_i, in the presence of which the standard Poisson model becomes a random or mixed effects Poisson model. Thus, for given X_i and $v_i, Y_i \sim Poisson(\mu_i v_i)$, where v_i's are continuous positive valued random variables that are independent and identically distributed with some finite mean $E(v_i)$ and variance $Var(v_i) = \tau$. If we let $E(v_i)$ to be 1 and $Var(Y_i|X_i) = \mu_i + \tau\mu_i^2$, then the null hypothesis for testing over-dispersion becomes, $H_0 : \tau = 0$. Failure to reject the null hypothesis leads to the Poisson model. In this study we perform, Dean's P_B test (Collings and Margolin, 1985) for over-dispersion using the *R* package *DCluster* for each gene. Rejection of H_0 for a specific gene justifies the application of a negative binomial model for detecting DE genes.

3 Results and Discussion

This section includes summary results of our analysis on the observed data. In the exploratory data analysis section, we observe the distribution of zeros on overall data as well as for Poisson and negative binomial data, based on which further classification of data is made. The next section presents summary statistics on DE genes obtained by using these models. Venn diagrams illustrate overlapping genes across the species and sex-by-species interactions. Since the main focus of this study is to adopt zero inflated models as an appropriate approach for DE genes identification, the final section reveals the performance of the zero inflated models. In particular, heatmaps for two sets of DE genes obtained by using ZIPMM and ZINBMM are presented.

3.1 Exploratory Analysis of the Data

We exclude the genes with zero counts in all the 36 cells and obtain a set of 17,886 genes for which number of zero counts ranges from 0 to 35. Figure 3(a) presents the distribution of genes with respect to number of zero counts they contain. The histogram depicts that most of the genes (more than 12,000) contain 0-5 zero counts. The number of genes for the range of 8-30 zero counts is substantially lower while a remarkable number of genes contain more than 30 zero counts. Dean's P_B test (Collings and Margolin, 1985) for overdispersion generates NOD and OD datasets. The NOD data set consists of 5,444 genes and is referred to as Poisson data set. The OD data set consists of 12,442 genes is referred to as negative binomial dataset.

In Figure 3(b), the distribution of zero counts in Poisson data reveals that most of the genes contain higher number of zero counts ranging from 15-35. In particular, more than 2,000 genes contain 30-35 zero counts. In Figure 3(c) it is evident that approximately 12,000 genes have zero counts ranging from 0 to 5. Therefore, there are fewer number of genes that contain excessive number of zeros on negative binomial data. However, the excessive number of zero counts leads to the classification of the data again and to the application of appropriate models. Finally, four data sets namely Poisson, ZIP, NB, and ZINB are created.

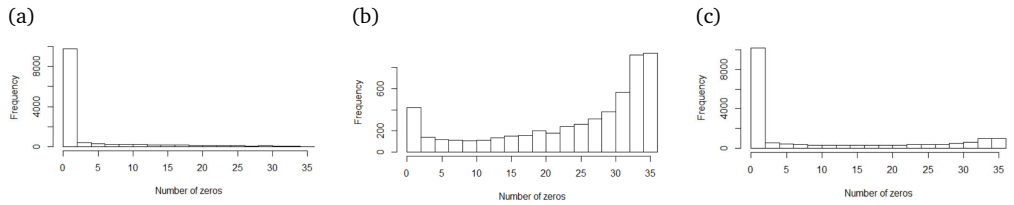

Figure 3: (a) Distribution of zero counts on aggregated data. (b) Distribution of zero counts on Poisson data. (c) Distribution of zero counts on negative binomial data

For each gene in Poisson data set, we fit a PMM. On the basis of BH adjusted *P*-values (< 0.05), we identify whether the particular gene is DE or not. Finally, index sets of DE genes for HS vs PT, HS vs RM, Male vs Female, HS male vs PT female, and HS male vs RM female are obtained. We obtain similar sets of DE genes for each of the four data sets. Table 1 illustrates the summary of DE genes in our study produced by all models. Out of 17,886 genes 27.47% gens are DE between HS and PT, 34% genes are DE between HS and RM, and 10.49% genes are DE between aggregate Male and Female. Out of 4,441 zero inflated NOD genes, ZIPMM model identifies 106 HS genes as DE with respect to PT. From a set 1,273 zero inflated OD genes, ZINBMM identifies 300 genes as DE between HS and PT. While Compared with HS vs RM, the ZIPMM detects 196 genes as DE from a set of 4,441 and ZINBMM finds 262 genes as DE from a total of 1,273 in ZINB set. The table also depicts that out of all genes 14.68% are DE between HS male vs PT female and 13.09% are DE between HS male vs RM female.

Table 1: Distribution of model-wise total number of genes and DE genes.

Model	Total number of genes	Number of DE genes				
		HS vs PT	HS vs RM	M vs F	HSM vs PTF	HSM vs RMF
PMM	1,003	203	369	100	82	87
ZIPMM	4,441	106	196	59	37	30
NBMM	11,169	4,305	5,254	1,580	2,370	2,138
ZINBMM	1,273	300	262	139	137	88
Total	17,886	4,914	6,081	1,877	2,626	2,343
Percentage	100%	27.47%	34%	10.49%	14.68%	13.09%

In order to identify the common DE genes across species and their combinations from the overall data, as well as from the model specific data, we present Venn diagrams. Venn diagrams allow us to visualize the number of overlapping DE genes with respect to two or more classification traits and the specific number of DE genes in each trait. Figure 4(a) presents the number of DE genes between HS and PT as well as between HS and RM for aggregate data. The Venn diagram shows that 2,768 genes are commonly differentially expressed for both PT and RM with respect to HS. On the other hand, 3,313 genes are only DE between HS and RM and 2,146 genes are only DE between HS and PT. Figure 4(b)

presents the number of DE genes between HS male, and PT female as well as between HS male and RM female for aggregate data. The diagram shows that 819 genes are commonly differentially expressed for both the groups HS male vs PT female, and HS male vs RM female. On the other hand 1,807 genes are only DE between HS male and PT female, and 1,524 genes are only DE between HS male and RM female. In particular, as we consider the application of ZIPMM and ZINBMM to model read counts data with excessive number of zeros, we focus only on these two models for our further analysis. For simplicity, we consider DE genes across species only.

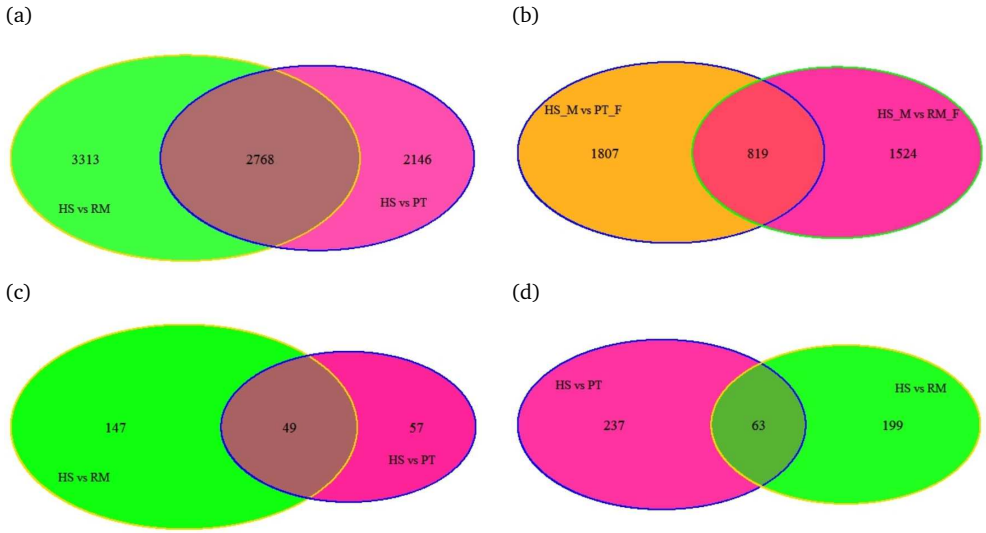

Figure 4: Venn diagrams of DE genes with respect to (a) Human vs Chimpanzee and Human vs Rhesus Macaque. (b) Sex-species interactions. (c) Human vs Chimpanzee and Human vs Rhesus Macaque obtained from ZIPMM. (d) Human vs Chimpanzee and Human vs Rhesus Macaque obtained from ZINBMM.

Figure 4(c) shows the number of DE genes between HS and PT as well as between HS and RM for the zero inflated Poisson data set. The Venn diagram reveals that only 49 genes are commonly DE for both PT and RM when compared against HS. On the other hand, 147 genes between HS and RM, and 57 genes between HS and PT are DE. Figure 4(d) represents the number of DE genes between HS and PT as well as between HS and RM from the zero inflated negative binomial data set. Only 63 genes are commonly DE for both PT and RM when compared against HS. On the other hand, 199 genes are only DE between HS and RM and 237 genes are only DE between HS and PT.

A heatmap is a pictorial representation of the data where the individual values contained in a data matrix are expressed as colors. To construct heatmaps, we consider only 49 and 63 DE genes common in HS versus PT and HS versus RM obtained by ZIPMM and ZINBMM respectively. We consider these portions of DE genes obtained by the aforesaid models as an illustration. For both the heatmaps below, we consider log(mean counts of RM/mean counts of HS) and log(mean counts of PT/mean counts of HS) for each of the genes.

The heatmaps in Figure 5(a) and Figure 5(b) are constructed with the DE genes obtained by ZIPMM and ZINBMM respectively. In both of the heatmaps, left side indicates comparison

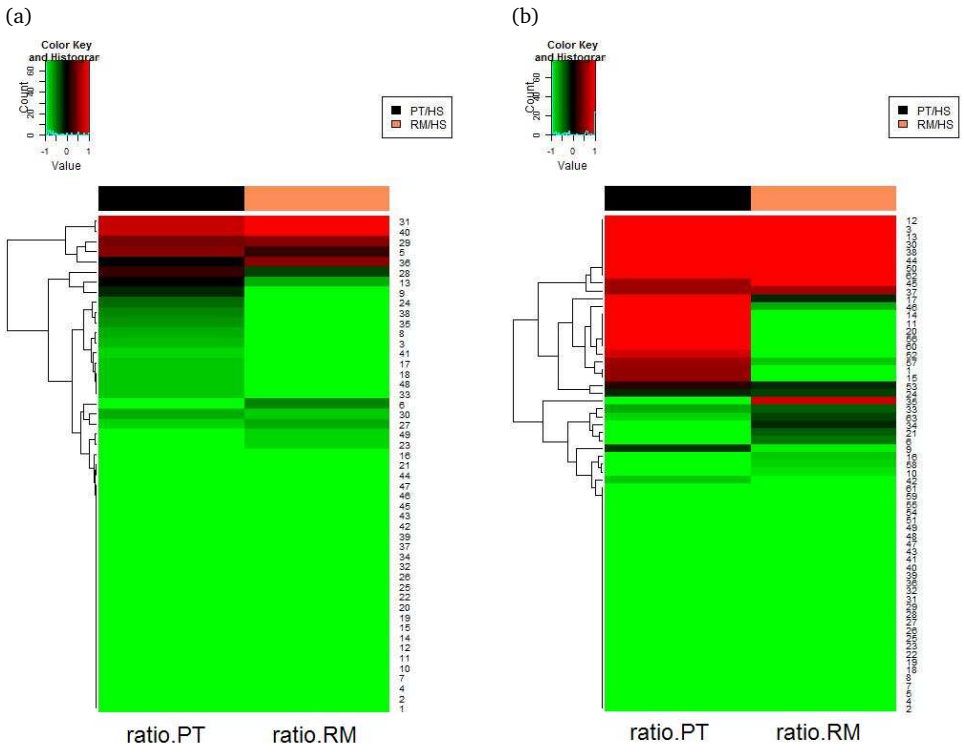

Figure 5: Heatmap of the DE genes with respect to Human vs Chimpanzee and Human vs Rhesus Macaque obtained from (a) ZIPMM and (b) ZINBMM. The numbers specified in the right side of the heatmaps are gene IDs.

between PT and HS and right side for the comparison of RM and HS. Inside the heatmaps, red color in the left represents PT is more prominent to be DE and green color represents HS is more prominent to be DE, while black color indicates those genes are not DE with respect to HS and PT. Similar comparison are made in the right side for RM and HS. Red color in the right side express dominance of RM and green color represents dominance of HS genes to be differentiable. Substantially smaller portions of black color in the heatmaps indicate misclassification by the ZIPMM and ZINBMM are relatively small. Overall, the heatmaps for DE genes obtained by using ZIPMM and ZINBMM illustrate good performance of the models for identifying DE genes across species when the data contain excessive number of zeros.

3.2 Simulation Study on the Performance of ZIPMM and ZINBMM

An efficient practice of simulating RNA-seq read counts is to empirically estimate the simulation parameters from a real RNA-seq data and generate read counts using these values. In our simulation study, we consider 49 genes that were found commonly DE in HS versus PT, and HS versus RM comparisons by ZIPMM as the true DE genes. These genes are a part of the non-over-dispersed zero inflated data. For each gene we estimate the probability of zero inflation, π and the mean parameter μ by using the method of moments (MM) technique.

Thus, $\hat{\mu} = \bar{y} + \frac{S^2}{\bar{y}} - 1$ and $\hat{\pi} = \frac{S^2 - \bar{y}}{\bar{y}^2 + S^2 - \bar{y}}$, where, $\bar{y}$ and S^2 are the species-specific mean and variance respectively. By keeping counts for PT and RM as usual, we take four fold changes of HS data to create DE data set with respect to species. For each of the DE genes, we simulate 10 genes and thus we have a set of 490 DE genes. We create another set of 490 genes where all of them are non-DE. Finally, we have a set of 980 genes out of which first 490 are DE and the remaining are non-DE.

Similarly as a base ZINB sample, we consider 63 common DE genes found in HS versus PT, and HS versus RM comparisons by ZINBMM. These DE genes are extracted from a set of over-dispersed zero inflated gene set. For each gene we estimate the probability of zero inflation π, the over-dispersion and mean parameters τ and λ respectively. Also, we use species specific average of $\hat{\pi}$s obtained from all 63 genes as an estimate of the probability of zero inflation. The other two parameters are estimated by using the following MM estimators: $\hat{\mu} = \frac{\bar{y}}{1 - \hat{\pi}}$ and $\hat{\tau} = \frac{\bar{y} - (S^2/\bar{y}) - 1}{\hat{\lambda}} - 1$, where $\bar{y}$ and S^2 are the species-specific mean and variance respectively. Similar to the ZIPMM simulation, we keep counts for PT and RM as it is and take four fold changes for HS data to create DE data set with respect to species. According to the same principle as before , we generate 1,260 genes where the first 630 are DE and the rest are non-DE.

Hundred data sets are generated for each of these two simulation schemes. ZIPMM is applied to the first scheme and ZINBMM to the second. In each case, we calculate Area under the curve (AUC), the average area under a receiver operating characteristic curve (ROC) to see if our proposed zero inflated models can identify the true DE genes. We also calculate true positive rate (TPR) and false positive rate (FPR).

3.3 Simulation Result

We evaluate the performance of ZIPMM and ZINBMM based on the efficiency of ranking true DE genes. To compare DE genes under the simulation set up and three scenarios: HS versus PT, HS versus RM and overall Male versus female, we compute TPR, FPR, and the AUC under each scenario. Table 2 presents the average AUC values, average TPR, and average FPR along with their corresponding standard errors obtained from the two models. Average AUC and TPR values indicate better performance of ZIPMM for identifying DE genes with respect to HS and PT, while ZINBMM outperforms ZIPMM with respect to HS vs RM. FPR is reasonably low (less than 6%) in identifying DE genes across species and sex based on ZIPMM. However, FPR is substantially high (more than 16%) in all the cases for the ZINBMM. Standard errors for AUC, TPR, and FPR are low for both models. However, none of the models perform well for detecting DE genes across males versus female overall.

The ROC curves present TPR against FPR. Figure 6(a) and 6(b) illustrate that at 0.0% FPR, the ZIPMM identifies more than 75% DE genes across HS versus PT, and about 60% DE genes across HS versus RM correctly. In addition we see that while FPR increases up to 10%, the TPR remains at a constant level of approximately 80% and 60% respectively for HS versus PT and for HS versus RM. The area under the curve for HS versus PT is bigger than that of HS versus RM which indicates that the model works better for HS versus PT for detection of true DE genes. Figure 6(c) and 6(d) depict that ZINBMM performs quite similarly in detecting true DE gens for both HS versus PT, and HS versus RM. The ROC curves for HS versus PT also show that as the FPR increases up to 10%, the TPR also increases for

Table 2: Model-wise average AUC, average TPR, and average FPR (standard error within parenthesis).

	ZIPMM			ZINBMM		
DE type	AUC	TPR	FPR	AUC	TPR	FPR
HS vs PT	0.793 (0.014)	0.758 (0.015)	0.048 (0.010)	0.754 (0.018)	0.663 (0.018)	0.163 (0.018)
HS vs RM	0.643 (0.017)	0.613 (0.018)	0.047 (0.009)	0.715 (0.017)	0.652 (0.017)	0.162 (0.017)
M vs F	0.551 (0.020)	0.112 (0.014)	0.053 (0.011)	0.469 (0.017)	0.124 (0.013)	0.166 (0.015)

both cases. Finally, it is clearly evident from the ROC curves that ZIPMM performs better than ZINBMM for identifying truly DE genes.

4 Conclusion

Poisson and negative binomial regression models are widely used methods for non-over-dispersed and over-dispersed count data respectively. However, in the presence of excessive number of zeros in the count data, these methods tend to perform poor. We applied zero-inflated Poisson and negative binomial models in the presence excessive zero counts in the RNA-seq data for detecting DE genes across sexes and within and between species. In order to include variations due individuals within each species these models are adapted as mixed-effects models for both non-over-dispersed and over-dispersed count data. We apply the zero inflated models to both simulated and real RNA-seq data set and examine their performances. Since a single model is not adequate to capture all the underlying features of the data, we split the entire data based on their underlying characteristics.

The current study implements a four-step approach to detect DE genes based on RNA-Seq count data. The first step deals with the test of over-dispersion. This test generates two sets of genes: OD and NOD genes. In the second step, we examine number of zero counts in each part. Based on the number of zeros, we divide both OD and NOD data sets into two groups with limited number of zeros and with excessive number of zeros. We consider genes with more than 33.33% zero counts as zero inflated on empirical basis of our data. Then all the genes are tested for differential expression. Poisson type models are used for testing NOD genes. For OD genes, we apply negative binomial models. In particular to address data with excessive zero counts we apply zero inflated Poisson mixed model (ZIPMM) for non-over-dispersed genes and zero inflated negative binomial mixed model (ZINBMM) for over-dispersed genes.

In order to assess performance of the zero inflated models, we check heatmaps of the DE genes for the real data. Relatively small portions of black color represent that the misidentification rates of DE genes are relatively small for both ZIPMM and ZINBMM models. ROC curves and AUC derived from the simulated data indicate that both ZIPMM and ZINBMM perform satisfactorily to identify DE genes across species. However, overall performance of ZIPMM is better than ZINBMM. In case of ZIPMM, FPR is less than 5% whereas for ZINBMM it is fairly large (greater than 16%). These results are consistent with Blekhman *et al.* (2010).

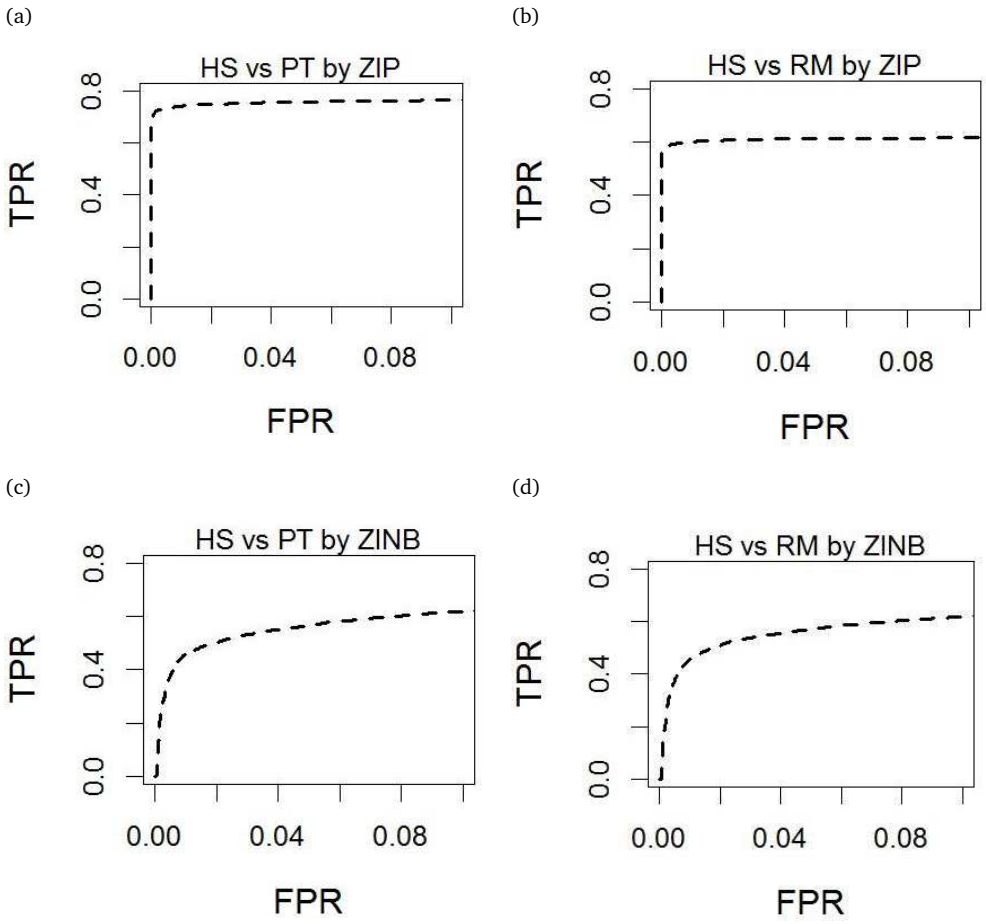

Figure 6: ROC curves with respect to (a) Human vs Chimpanzee obtained from ZIPMM. (b) Human vs Rhesus Macaque obtained from ZIPMM. (c) Human vs Chimpanzee obtained from ZINBMM. (d) Human vs Rhesus Macaque obtained from ZINBMM

However, they only considered two species as the two treatment conditions in their model. In our study, we considered three species as three treatment conditions which is relatively a complex situation for differential expression analysis.

Genes containing more than 33.33% zero read counts are considered as zero inflated which is merely empirical for our data. Theoretical basis for this cut point is not justified. ZIPMM and ZINBMM count data simulation and model execution become quite time intensive due to the consideration of species, sex and individuals as covariate effects. To the current date in RNA-seq experiments, none of the methods work distinctively better under all experimental conditions. Apart from the limitations listed above, zero-inflated models perform better under a mixture of NOD and OD genes with excessive number of zeros. Thus zero-inflated models serve as valuable tools for analyzing RNA-seq data for complex study design and in the presense of excessive number of zero counts.

Acknowledgments

The authors would like to thank the anonymous referee for critical reading and helpful comments which improve the paper significantly. All authors contributed equally toward producing the manuscript.

Declarations

Funding: None.
Conflict of interest: None.
Ethical approval: Not applicable.

References

Anders S, Huber W (2010). "Differential expression analysis for sequence count data." *Genome Biol*, **11**(10), R106. doi:10.1186/gb-2010-11-10-r106.

Auer PL, Doerge RW (2011). "A two-stage Poisson model for testing RNA-seq data." *Statistical Applications in Genetics and Molecular Biology*, **10**(1), 1–26. doi:10.2202/1544-6115.1627.

Baggerly KA, Deng L, Morris JS, Aldaz CM (2003). "Differential expression in SAGE: accounting for normal between-library variation." *Bioinformatics*, **19**(12), 1477–1483. doi:10.1093/bioinformatics/btg173.

Baggerly KA, Deng L, Morris JS, Aldaz CM (2004). "Overdispersed logistic regression for SAGE: modelling multiple groups and covariates." *BMC bioinformatics*, **5**(1), 144. doi:10.1186/1471-2105-5-144.

Balzergue S, Rigaill G, Brunaud V, Blondet E, Rau A, Rogier O, Caius J, Maugis-Rabusseau C, Soubigou-Taconnat L, Aubourg S, *et al.* (????). "Differential expression analysis of RNA-seq data with the HTSDiff package."

Bates D, Mächler M, Bolker B, Walker S (2014). "Fitting linear mixed-effects models using lme4." URL https://arxiv.org/pdf/1406.5823.pdf.

Benjamini Y, Hochberg Y (1995). "Controlling the false discovery rate: a practical and powerful approach to multiple testing." *Journal of the Royal Statistical Society. Series B (Methodological)*, pp. 289–300. URL https://www.jstor.org/stable/2346101.

Blekhman R, Marioni JC, Zumbo P, Stephens M, Gilad Y (2010). "Sex-specific and lineage-specific alternative splicing in primates." *Genome Research*, **20**(2), 180–189. doi:doi/10.1101/gr.099226.109.

Bloom JS, Khan Z, Kruglyak L, Singh M, Caudy AA (2009). "Measuring differential gene expression by short read sequencing: quantitative comparison to 2-channel gene expression microarrays." *BMC genomics*, **10**(1), 221. doi:10.1186/1471-2164-10-221.

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**(1), 94. doi:10.1186/1471-2105-11-94.

Chen EZ, Li H (2016). "A two-part mixed-effects model for analyzing longitudinal microbiome compositional data." *Bioinformatics*, **32**(17), 2611–2617. doi:10.1093/bioinformatics/btw308.

Choi H, Gim J, Won S, Kim YJ, Kwon S, Park C (2017). "Network analysis for count data with excess zeros." *BMC genetics*, **18**(1), 93. doi:10.1186/s12863-017-0561-z.

Collings BJ, Margolin BH (1985). "Testing goodness of fit for the Poisson assumption when observations are not identically distributed." *Journal of the American Statistical Association*, **80**(390), 411–418. doi:10.2307/2287906.

Dean C, Lawless JF (1989). "Tests for detecting overdispersion in Poisson regression models." *Journal of the American Statistical Association*, **84**(406), 467–472. doi:10.2307/2289931.

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), 1–28. doi:10.2202/1544-6115.1637.

Gentleman RC, Carey VJ, Bates DM, Bolstad B, Dettling M, Dudoit S, Ellis B, Gautier L, Ge Y, Gentry J, *et al.* (2004). "Bioconductor: open software development for computational biology and bioinformatics." *Genome Biology*, **5**(10), R80. doi:10.1007/0-387-29362-0_21.

Kvam VM, Liu P, Si Y (2012). "A comparison of statistical methods for detecting differentially expressed genes from RNA-seq data." *American Journal of Botany*, **99**(2), 248–256. doi:10.3732/ajb.1100340.

Li J, Witten DM, Johnstone IM, Tibshirani R (2011). "Normalization, testing, and false discovery rate estimation for RNA-sequencing data." *Biostatistics*, p. kxr031. doi:10.1093/biostatistics/kxr031.

Mahi NA, Begum M (2016). "A two-step integrated approach to detect differentially expressed genes in RNA-Seq data." *Journal of Bioinformatics and Computational Biology*, **14**(6), 18. doi:10.1142/s0219720016500347.

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**(9), 1509–1517. doi:10.1101/gr.079558.108.

Oshlack A, Robinson MD, Young MD, *et al.* (2010). "From RNA-seq reads to differential expression results." *Genome Biol*, **11**(12), 220. doi:10.7717/peerj.3566/fig-2.

Robinson MD, Oshlack A, *et al.* (2010). "A scaling normalization method for differential expression analysis of RNA-seq data." *Genome Biol*, **11**(3), R25. doi:10.1186/gb-2010-11-3-r25.

Robinson MD, Smyth GK (2007). "Moderated statistical tests for assessing differences in tag abundance." *Bioinformatics*, **23**(21), 2881–2887. doi:10.1093/bioinformatics/btm453.

Robinson MD, Smyth GK (2008). "Small-sample estimation of negative binomial dispersion, with applications to SAGE data." *Biostatistics*, **9**(2), 321–332. doi:10.1093/biostatistics/kxm030.

Srivastava S, Chen L (2010). "A two-parameter generalized Poisson model to improve the analysis of RNA-seq data." *Nucleic Acids Research*, **38**(17), e170–e170. doi:10.1093/nar/gkq670.

Vêncio RZ, Brentani H, Patrão DF, Pereira CA (2004). "Bayesian model accounting for within-class biological variability in Serial Analysis of Gene Expression (SAGE)." *BMC bioinformatics*, **5**(1), 119. doi:10.1186/1471-2105-5-119.

Yu D, Huber W, Vitek O (2013). "Shrinkage estimation of dispersion in Negative Binomial models for RNA-seq experiments with small sample size." *Bioinformatics*, **29**(10), 1275–1282. doi:10.1093/bioinformatics/btt143.

Zhang X, Mallick H, Tang Z, Zhang L, Cui X, Benson AK, Yi N (2017). "Negative binomial mixed models for analyzing microbiome count data." *BMC bioinformatics*, **18**(1), 4. doi:10.3389/fmicb.2018.01683.

Zhou Y, Wan X, Zhang B, Tong T (2018). "Classifying next-generation sequencing data using a zero-inflated Poisson model." *Bioinformatics*, **34**(8), 1329–1335. doi:10.1093/bioinformatics/btx768.