Bayesian Analysis (2017) **TBA**, Number TBA, pp. 1–26

Bayesian Analysis of RNA-Seq Data Using a Family of Negative Binomial Models

Lili Zhao[*], Weisheng Wu[†], Dai Feng[‡], Hui Jiang[§], and XuanLong Nguyen[¶]

Abstract. The analysis of RNA-Seq data has been focused on three main categories, including gene expression, relative exon usage and transcript expression. Methods have been proposed independently for each category using a negative binomial (NB) model. However, counts following a NB distribution on one feature (e.g., exon) do not guarantee a NB distribution for the other two features (e.g., gene/transcript). In this paper we propose a family of Negative Binomial models, which integrates the gene, exon and transcript analysis under a coherent NB model. The proposed model easily incorporates the uncertainty of assigning reads to transcripts and simplifies substantially the estimation for the relative usage. We developed simple Gibbs sampling algorithms for the posterior inference by exploiting fully tractable closed-forms of computation via suitable conjugate priors. The proposed models were investigated under extensive simulations. Finally, we applied our model to a real data set.

Keywords: Bayesian RNA-Seq, Chinese restaurant table distribution, differential test, exon usage, transcript analysis.

1 Introduction

High throughput sequencing technology has rapidly become the standard method for measuring RNA expression levels (Mortazavi et al., 2008). RNA-Seq uses next-generation sequencing (NGS) methods to sequence cDNA that has been derived from an RNA sample, and hence produces millions of short reads. These reads are mapped to a reference genome using a data alignment tool, such as TopHat (Kim et al., 2013). The number of reads mapping within a genomic feature of interest (such as a gene or an exon) is used as a measure of the abundance of that feature in the analyzed sample (Anders et al., 2012).

For the statistical analysis, we rely on a simple count matrix produced from the RNA-Seq experiments. In the count matrix the row represents a feature and the column a sample in a specific experimental condition. Then the data are normalized to account for different library sizes (the total number of mapped reads) across different samples. A detailed review of the normalization methods can be found in Dillies et al. (2013) and Rapaport et al. (2013). It is widely recognized that a negative binomial (NB) distribution

[*]Department of Biostatistics, University of Michigan, Ann Arbor, Michigan, U.S.A., zhaolili@umich.edu
[†]Department of Computational Medicine & Bioinformatics, University of Michigan
[‡]Biometrics Research Department, Merck Research Laboratories, U.S.A.
[§]Department of Biostatistics, University of Michigan, Ann Arbor, Michigan, U.S.A.
[¶]Department of Statistics, University of Michigan, Ann Arbor

 DOI: 10.1214/17-BA1055

provides a good fit to read counts data produced by NGS. Many statistical algorithms have been published using a NB model, including Anders and Huber (2010); Robinson et al. (2010); Di et al. (2011); Hardcastle and Kelly (2010); Leng et al. (2013); van de Wiel et al. (2012, 2014); Love et al. (2014); Wu et al. (2013); Trapnell et al. (2013); Niu et al. (2014). In all these works, the NB distribution was parameterized by a mean parameter μ and an overdispersion parameter α. Hypothesis tests were constructed to compare the μ's between conditions. However, all statistical inference rely on some form of approximation. Many treat the estimated dispersions as if they were known parameters, without allowing for uncertainty of the estimation. Other methods that account for the uncertainty in the dispersion rely on other approximations due to the lack of conjugacy in the parameter formulation (see Hardcastle and Kelly (2010); van de Wiel et al. (2012)). In this paper, we parameterize the NB distribution using a probability parameter p and a dispersion parameter r and treat both as unknown parameters. Based on the newly developed sampling algorithm in the field of topic modelling (Zhou and Carin, 2012, 2015), we estimate p and r using fully tractable closed-forms via suitable conjugate priors. More importantly, this parameterization allows us to define a family of NB models for each gene to unify the analysis of genes, exons and transcripts under a coherent statistical model.

The analysis of genes, exons and transcripts have been developed with separate NB models. For example, a NB distribution has been used for 1) modelling the total gene counts, such as edgeR (Robinson et al., 2010), DESeq (Anders and Huber, 2010), DESeq2 (Love et al., 2014), baySeq (Hardcastle and Kelly, 2010), NBPSeq (Di et al., 2011), and ShrinkBayes (van de Wiel et al., 2014), 2) modelling exon counts, such as DEXSeq (Anders et al., 2012), and 3) modelling transcript counts, such as Cuffdiff (Trapnell et al., 2013) and IUTA (Niu et al., 2014). However, under their model formulations, a NB model for one feature prohibits a NB distribution for the other two features. In particular, if different NB distributions are assumed for different exon counts within a particular gene, the total count in that gene by summing over the exon counts does not have a NB distribution.

In this paper, we propose a family of NB models (FNB for short), which unifies the exon, transcript and gene analysis under a coherent negative binomial modelling framework. Moreover, the FNB approach greatly simplifies the imputation of latent transcript counts and allows us to estimate the relative exon/transcript usage from a simple multinomial distribution.

2 Methods

2.1 Parameter estimation in a negative binomial distribution

Most negative binomial models developed for the analysis of RNA-Seq count data are parameterized by a mean parameter μ and an overdispersion parameter α. In this work, we characterize the NB distribution by parameters p and r. That is, $y|r,p \sim \text{NB}(r,p)$ has the probability mass function $f(y|r,p) = \frac{\Gamma(r+y)}{y!\Gamma(r)}(1-p)^r p^y$, where Γ denotes the gamma function, $p \in (0,1)$ and $r > 0$. The mean and overdispersion are obtained

by $\mu = rp/(1-p)$ and $r = 1/\alpha$. It is a classical fact that the NB distribution is equivalent to a Gamma-Poisson mixture distribution: we can obtain $y|r,p \sim \text{NB}(r,p)$ by first drawing $\lambda \sim \text{Gamma}(r,(1-p)/p)$ and then generating $y|\lambda \sim \text{Pois}(\lambda)$. A NB distribution can also be augmented under a compound Poisson representation, so that by endowing both parameters r and p with suitable conjugate priors, it is possible to draw samples for their posterior distribution in a tractable way, see Zhou and Carin (2012, 2015). Specifically, given the NB model $y_j|r,p \overset{iid}{\sim} \text{NB}(r,p)$, for $j = 1,\cdots,n$, the prior specifications $p \sim \text{Beta}(a_0,b_0)$, and $r \sim \text{Gamma}(e_0,f_0)$, where e_0 and f_0 are the shape and the rate parameters. Then the posterior distributions of p and r are obtained in the limit by iteratively applying the following Gibbs sampling steps:

$$(p|-) \sim \text{Beta}(a_0 + \sum\nolimits_j y_j, b_0 + nr), \tag{1}$$

$$(l_j|-) \sim \text{CRT}(y_j, r), \tag{2}$$

$$(r|-) \sim \text{Gamma}(e_0 + \sum\nolimits_j l_j, f_0 - n\log(1-p)). \tag{3}$$

In the above display "$|-$" denotes the conditional distributions given the data and all remaining parameters. The first step for sampling p follows from the Beta-Binomial conjugacy. The last two steps complete the sampling of parameter r, which also involves auxiliary variable l_j's. These variables represent (latent) counts distributed according to a Chinese restaurant table (CRT) distribution, which are defined as follows. We write $(l_j|-) \sim \text{CRT}(y_j,r)$, if

$$l_j = \sum_{m=1}^{y_j} b_m, \quad \text{and} \quad b_m \sim \text{Bernoulli}(r/(m-1+r)).$$

Now, given all l_j's, r can be sampled by the third equation, which is obtained by exploiting a Gamma-Poisson conjugacy.

2.2 Notation

We first introduce the notation for a particular gene. In general, let y_j be gene count in sample j ($j = 1,\cdots,n$), and r and p are gene-level parameters in the NB distribution. We use superscript "e" for exon-level data and parameters and superscript "t" for transcript-level data and parameters. Specifically, on the exon level, let $y_j^{e_i}$ be the read count in exon i ($i = 1,\cdots,E$) in sample j, r^{e_i} be the corresponding dispersion parameter. On the transcript level, let $y_j^{t_{i'}}$ be the read count in transcript i' ($i' = 1,\cdots,T$) in sample j, $r^{t_{i'}}$ be the corresponding dispersion parameter. We know that $y_j = \sum_{i=1}^{E} y_j^{e_i} = \sum_{i'}^{T} y_j^{t_{i'}}$ (i.e., the sum of exon counts and the sum of transcript counts are equal to the gene counts). In our proposed model (defined in the next section), we have $r = \sum_{i=1}^{E} r^{e_i} = \sum_{i'=1}^{T} r^{t_{i'}}$ (i.e., the sum of exon-level dispersion parameters and the sum of transcript-level dispersion parameters are equal to the gene-level dispersion parameter). There is a single parameter p for the gene, exons and transcripts within that gene. The mean expression for exons are $\mu^{e_i}(i = 1,\cdots,E)$ and for transcripts are $\mu_{i'}^{t}(i' = 1,\cdots,T)$.

	Transcript 1	Transcript 2	Exon counts
Exon 1	?	?	100
Exon 2	0	200	200
Exon 3	?	?	400
Transcript counts	?	?	700

$$M = \begin{bmatrix} 1 & 1 \\ 0 & 1 \\ 1 & 1 \end{bmatrix}$$

Table 1: The left table shows the hypothetical data for a particular gene in a given sample ("?" refers to unobserved data). The M matrix for this data is on the right.

We also need notation on the exon-transcript level to show that our model integrates the gene, exon and transcript analysis under a coherent NB model. Let $y^{e_i t_{i'}}$ be the read count in exon i and transcript i' ($i = 1, \cdots, E; i' = 1, \cdots, T$). Since not all exons are included in a particular transcript, we set $M_{e_i t_{i'}} = 1$ if exon i belongs to transcript i' and zero otherwise. Let S_{e_i} be the effective length of exon i, it follows that the effective length of transcript i' is $L_{t_{i'}} = \sum_i M_{e_i t_{i'}} S_{e_i}$.

Furthermore, we use subscript "k" to denote experimental conditions. For example, in condition k ($k = 1, 2$), the read count data are denoted by y_{jk}, $y_{jk}^{e_i}$ and $y_{jk}^{t_{i'}}$ on the gene, exon and transcript level, respectively, and the corresponding dispersion parameters are r_k, $r_k^{e_i}$ and $r_k^{t_{i'}}$, mean parameters are μ_k, $\mu_k^{e_i}$ and $\mu_k^{t_{i'}}$, and the probability parameter is p_k.

2.3 A family of negative binomial models for the analysis of gene, exon and transcript

In this section we focus on the modelling for a particular gene in a particular sample (we dropped indices of j and k for simplified notation). Table 1 shows a simple hypothetical data set with a total read count 700. This total count is decomposed into multiple counts on the exon level (i.e., $700 = 100 + 200 + 400$), or multiple counts on the transcript level. Here, "exon" is used loosely. More generally, "exon" can also be thought of as part of the exon (i.e., an exon counting bin in Anders et al. (2012), or as defined in Turro et al. (2011)). Regardless of these various definitions, counts in exons are observed, while counts in transcripts are unobserved. If we know how each exon count is allocated to different transcripts, we can impute the unobserved count for a given transcript by summing over exon counts belonging to that transcript. However, the allocation probabilities are unknown. Moreover, the allocation is under some restrictions because not all exons are included in a particular transcript. We use a M matrix to consider the restrictions. In the hypothetical example, the first row of M is $(1, 1)$ indicating that count 100 is allocated to two transcripts, and the second row is $(0, 1)$ indicating that count 200 is only allocated to transcript 2 since transcript 1 does not include exon 2.

In this paper, we exploit the infinitely divisible property of the negative binomial distribution and define a family of NB models for each gene. In a "gene" family, read counts in all family members (exons, transcripts and the corresponding gene) have NB

distributions. These NB distributions share a common probability parameter p, but each retain its own dispersion parameter (we call it a FNB model). The key assumption in our FNB model is that all family members share the same p. Biologically, the common p can characterize the effect of common transcription factors that control the rates of transcription in regulating the amounts of RNA products. Mathematically, this assumption greatly reduces the number of parameters in the model, which could gain efficiency for the parameter estimation, especially when the sample size is small as in RNA-Seq studies. Furthermore, this assumption allows us to unify the analyses of gene, exon and transcript under a coherent modeling framework. To investigate the robustness of the FNB model to this assumption, we conducted extensive simulation studies and performed goodness-of-fit tests in the real data analysis.

To illustrate that the gene, exon and transcript analyses are integrated under our proposed FNB model, we start from the exon-transcript level by assuming that the read count in exon i and transcript i' has a NB distribution. That is, $y^{e_i t_{i'}} \sim \text{NB}(M_{e_i t_{i'}} S_{e_i} \tilde{r}^{t_{i'}}, p)$, where $\tilde{r}^{t_{i'}}$ is the dispersion parameter per unit length of transcript i', and S_{e_i} is the effective length of exon i (Turro et al., 2011). Here, $y^{e_i t_{i'}} = 0$ if $M_{e_i t_{i'}} = 0$. The count in exon i has a NB distribution, which is given by

$$y^{e_i} = \sum_{i'}^{T} y^{e_i t_{i'}} \sim \text{NB}(S_{e_i} \sum_{i'}^{T} M_{e_i t_{i'}} \tilde{r}^{t_{i'}}, p), \quad i = 1, \cdots, E.$$

This is true because the sum of independent NB distributions with a common p is still a NB distribution with the same p and the dispersion parameter obtained by summing over the dispersion parameters in the independent NB distributions. Here, $r^{e_i} = S_{e_i} \sum_{i'}^{T} M_{e_i t_{i'}} \tilde{r}^{t_{i'}}$. It is worth noting that the M matrix and exon length are not necessary for the estimation of r^{e_i}, which can be directly estimated using observed exon counts.

Similarly, on the transcript level, we have

$$y^{t_{i'}} = \sum_{i}^{E} y^{e_i t_{i'}} \sim \text{NB}(L_{t_{i'}} \tilde{r}^{t_{i'}}, p),$$

where $L_{t_{i'}} = \sum_{i}^{E} M_{e_i t_{i'}} S_{e_i}$

Finally, the count in a gene is

$$y = \sum_{i} \sum_{i'} y^{e_i t_{i'}} \sim \text{NB}(\sum_{i'} L_{t_{i'}} \tilde{r}^{t_{i'}}, p).$$

Therefore, analyses of the gene, exon and transcript were integrated under a coherent NB modeling framework. The only unknown parameters in the FNB model are the probability and dispersion parameters, the data on the exon level, the exon and transcript length, and M matrix are known.

In the next two sections, we show that the above modelling framework 1) provides a simple way to infer the allocation probabilities in order to impute unobserved transcript

counts, and 2) reduces the complex problem of estimating the relative usage in a NB model to a simpler problem of estimating of the proportion of latent counts from a multinomial distribution.

2.4 Imputation of unobserved transcript counts

The read count in each transcript is not directly observed while similar transcripts can generate identical sequence reads (for example, the "?" in Table 1). Imputation of transcript counts is statistically very challenging especially under the NB model. Because of the difficulty, a two-step procedure is commonly used (Trapnell et al., 2013; Niu et al., 2014). That is, impute transcript counts in the first step, and then perform differential test in the second step based on the imputed counts obtained from the first step. Ideally, these two steps should be modelled simultaneously to avoid the potential bias and gain efficiency in the parameter estimation.

In our imputation, we treat $y^{e_i t_{i'}}$ $(i = 1, \cdots, E; i' = 1, \cdots, T)$ as unknown parameters, and estimate them together with other parameters in the NB distribution. As discussed in Section 2.3, $y^{e_i t_{i'}} \sim \text{NB}(M_{e_i t_{i'}} S_{e_i} \tilde{r}^{t_{i'}}, p)$, and it can also be reformulated using the Gamma-Poisson mixture formulation,

$$(y^{e_i t_{i'}} | \lambda^{e_i t_{i'}}) \sim \text{Poi}(\lambda^{e_i t_{i'}}), \tag{4}$$

$$(\lambda^{e_i t_{i'}} | -) \sim \text{Gamma}(M_{e_i t_{i'}} S_{e_i} \tilde{r}^{t_{i'}}, (1-p)/p). \tag{5}$$

At each Markov chain Monte Carlo (MCMC) iteration, we sample $\lambda^{e_i t_{i'}}$ $(i' = 1, \cdots, T)$ from Gamma distributions in (5). Based on the relationship between independent Poisson distributions and the multinomial distribution, we estimate transcript counts in exon i from

$$(y^{e_i t_1}, \cdots, y^{e_i t_{i'}}, \cdots, y^{e_i t_T}) \sim \text{Multinomial}(y^{e_i}, \mathbf{p}), \quad i = 1, \cdots, E,$$

where $\mathbf{p} = (\frac{\lambda^{e_i t_1}}{\sum_{i'} \lambda^{e_i t_{i'}}}, \cdots, \frac{\lambda^{e_i t_T}}{\sum_{i'} \lambda^{e_i t_{i'}}}, \cdots, \frac{\lambda^{e_i t_T}}{\sum_{i'} \lambda^{e_i t_{i'}}})$. This probability vector $\mathbf{p}$ contains the allocation probabilities, with which the count in exon i is distributed to different transcripts. The imputed count for transcript i' is simply $y^{t_{i'}} = \sum_i y^{e_i t_{i'}}$.

It is important to note that our imputation is naturally embedded in the whole estimation procedure. This procedure relies on the Gamma-Poisson mixture formulation of the NB distribution for imputation and the compound Poisson representation of the NB distribution for the parameter estimation.

2.5 Estimation of the relative usage

When the interest lies in estimating the fraction of the gene's reads that falls into the exon or transcript (i.e., the relative usage), read counts across all exons/transcripts in a particular gene need to be modelled simultaneously. The FNB model reduces the complex problem of estimating the relative usage in a NB model to a simple problem of estimating of the proportion of latent counts from a multinomial distribution. Specifically, the relative exon usage $\mathbf{Q^E} = (Q^{e_1}, \cdots, Q^{e_i}, \cdots, Q^{e_E})$ is sampled from a Dirichlet distribution,

$$\mathbf{Q^E} \sim \text{Dirichlet}(h_0 + \sum_j l_j^{e_1}, \cdots, h_0 + \sum_j l_j^{e_i}, \cdots, h_0 + \sum_j l_j^{e_E}), \tag{6}$$

where $l_j^{e_i} \sim \text{CRT}(y_j^{e_i}, r^{e_i})$, and h_0 is a fixed hyperparameter in the Dirichlet prior to quantify the prior belief of the relative usage. Conceptually, latent counts across the exons have a multinomial distribution. The relative usage of an exon is proportional to the latent counts in that exon (proof for (6) can be found in supplementary materials (Zhao et al., 2017)).

Similarly, the relative transcript usage is based on the latent counts inferred from the imputed transcript counts, that is, $\mathbf{Q^T} \sim \text{Dirichlet}(h_0 + \sum_j l_j^{t_1}, \cdots, h_0 + \sum_j l_j^{t_i'}, \cdots h_0 + \sum_j l_j^{t_T})$.

3 Parameter estimation and differential tests

In the previous sections, we defined our FNB model and emphasized its novelty through its simplicity to impute the unobserved transcript counts and infer the relative usage. In this section we discuss detailed steps for parameter estimations and differential tests on the gene, exon and transcript level, respectively.

For the illustrative purpose, we assume two experimental conditions and use index k to denote condition, but the algorithms can be easily extended to deal with multiple conditions. Recall that y_{jk} denotes the normalized total count for a particular gene in sample j in condition k ($j = 1, \cdots, n_k$; $k = 1, 2$). The read counts on the gene, exon and transcript level all have NB distributions and they share a common p_k, but they have their own dispersion parameters (see Section 2.2 for the relevant notation). In the sampling algorithms below, the gene-level dispersion parameters are assumed to be the same between conditions (i.e., $r_1 = r_2 = r$, which is a commonly used assumption in practice), but the algorithms can be easily modified to allow different r's.

3.1 Gene analysis

We proposed two algorithms to estimate the expression for a particular gene. Algorithm 1A is a simple extension of (1)–(3) for studies with two experimental conditions. After iterating the four steps in Algorithm 1A, we obtain posterior distributions of the mean gene expression, μ_1 and μ_2, for the two conditions (we removed the "|–" in the algorithms below to simplify the notation). Based on these posterior distributions, we obtain the posterior distribution of the log fold change, $\beta = \log(\mu_2) - \log(\mu_1)$, and compute a posterior probability $P_\beta = 2 \times \min\{P(\beta \leq 0), P(\beta > 0)\}$, with a smaller value of P_β indicating stronger evidence of differential expression (DE) for that gene.

For the multiplicity control, we can convert $P(\beta \leq 0)$'s, or $P(\beta > 0)$'s, over all genes to obtain the posterior expected false discovery rates (FDR) using methods in Lewin et al. (2007); León-Novelo et al. (2013) (see supplementary materials (Zhao et al., 2017)). These FDRs are derived under an one-sided test: $H_1 : \mu_2 \leq \mu_1$ or $H_1 : \mu_2 > \mu_1$. If r is assumed to be the same between conditions, we can derive the FDRs under a two-sided test (i.e., $H_0 : \mu_2 = \mu_1$ vs. $H_1 : \mu_2 \neq \mu_1$), see Algorithm 1B. If r is common between

Algorithm 1A Gene Analysis

Sample $p_k \sim$ Beta$(a_0 + \sum_j y_{jk}, b_0 + n_k r)$, $k = 1, 2$
Sample $l_{jk} \sim$ CRT(y_{jk}, r), $j = 1, \cdots, n_k$
Sample $r \sim$ Gamma$(e_0 + \sum_k \sum_j l_{jk}, f_0 - \sum_k n_k \log(1 - p_k))$
Calculate $\mu_k = rp_k/(1 - p_k)$

Algorithm 1B Gene Analysis

Sample $p_k \sim$ Beta$(a_0 + \sum_j y_{jk}, b_0 + n_k r)$, $j = 1, \cdots, n; k = 1, 2$
Sample $p_0 \sim$ Beta$(a_0 + \sum_k \sum_j y_{jk}, b_0 + r \sum_k n_k)$
Calculate $m_0 =$ B$(a_0 + \sum_k \sum_j y_{jk}, b_0 + r \sum_k n_k)$, and $m_1 = \prod_k$ B$(a_0 + \sum_j y_{jk}, b_0 + n_k r)$
Calculate $\hat{p} = (1 + \frac{\pi_0 m_0}{\pi_1 m_1})^{-1}$
Sample $z \sim$ Bernouli$(\hat{p})$; if $z = 1$, then $p_k = p_k$, else $p_k = p_0$
The rest (sampling l_{jk} and r) are the same as in **Algorithm 1A**

conditions, the hypothesis test for DE is reduced to $H_0 : p_2 = p_1$ vs. $H_1 : p_2 \neq p_1$. Let $z \in \{0, 1\}$ indicates choosing the null or alternative model, the posterior probability of selecting H_1 is given by

$$Pr(z = 1|-) = \left(1 + \frac{\pi_0}{\pi_1} \times \frac{m_0(y)}{m_1(y)}\right)^{-1},$$

where π_0 and π_1 are the prior probabilities for null and alternative model. The marginal likelihood functions, $m_0(y)$ and $m_1(y)$ under the null and alternative model, are obtained by integrating out the p_k's with respect to their prior distributions. Under our parametrization of the NB distribution, the marginal functions have simple forms due to the fact that a NB distribution with a beta prior for the probability parameter is a beta negative binomial distribution, which is given by

$$Pr(z = 1|-) = \left(1 + \frac{\pi_0}{\pi_1} \frac{B(a_0, b_0) B(a_0 + r \sum_k n_k, b_0 + \sum_j \sum_k y_{jk})}{\prod_k B(a_0 + n_k r, b_0 + \sum_j y_{jk})}\right)^{-1}, \quad (7)$$

where $B(,)$ denotes a beta function; derivation of the formula (7) is in supplementary materials (Zhao et al., 2017). Posterior probabilities $Pr(z = 1|-)$'s over all genes, obtained from Algorithm 1B, are then converted to the posterior expected FDRs.

When the number of replicates per condition is small, it is beneficial to allow each gene to have its own dispersion estimate while borrowing information from the other genes. In the Bayesian framework, this is achieved by adding a hierarchical structure for gene-level $r's$. Here, we add index g to represent gene g. That is, $r_g \sim$ Gamma(R, c_0) $g = 1, \cdots, G$. Then the sampling of r_g is

$$(r_g|-) \sim \text{Gamma}(R + \sum_k \sum_j l_{gjk}, c_0 - \sum_k n_k \log(1 - p_{gk})), \quad (8)$$

$$(l'_g|-) \sim \text{CRT}(\sum_k \sum_j l_{gjk}, R), \quad (9)$$

$$(R|-) \quad \sim \quad \text{Gamma}(e_0 + \sum\nolimits_g l'_g, f_0 - \sum\nolimits_g \log(1-p'_g)), \tag{10}$$

where $p'_g = \frac{-\sum_k n_k \log(1-p_{gk})}{c_0 - \sum_k n_k \log(1-p_{gk})}$. Parameters e_0 and f_0 in a gamma distribution quantify the prior belief for the population parameter R, and c_0 controls the shrinkage of the gene-level r_g's to the population average.

3.2 Exon analysis

Algorithm 2 contains steps to sample parameters in the exon analysis.

Algorithm 2 Exon Analysis

Sample $p_k \sim \text{Beta}(a_0 + \sum_j y_{jk}, b_0 + n_k r),\ k = 1, 2$
Sample $l_{jk}^{e_i} \sim \text{CRT}(y_{jk}^{e_i}, r_k^{e_i}),\ i = 1, \cdots E; j = 1, \cdots, n_k$
Sample $\mathbf{Q}_k^E \sim \text{Dirichlet}(h_0 + \sum_j l_{jk}^{e_1}, \cdots h_0 + \sum_j l_{jk}^{e_E})$
Sample $r \sim \text{Gamma}(e_0 + \sum_k \sum_j \sum_i l_{jk}^{e_i}, f_0 - \sum_k n_k \log(1-p_k))$
Calculate $r_k^{e_i} = rQ_k^{e_i}$, and $\mu_k^{e_i} = r_k^{e_i} p_k/(1-p_k)$

MCMC sampling allows us to carry out essentially all posterior inference of interest.

1. To test differential expression for exon i, we calculate the posterior probability $P(\mu_2^{e_i} > \mu_1^{e_i})$ or $P(\mu_2^{e_i} < \mu_1^{e_i})$.
2. To test differential relative usage for exon i, we calculate the posterior probability $P(Q_2^{e_i} > Q_1^{e_i})$ or $P(Q_2^{e_i} < Q_1^{e_i})$.
3. To test the difference in the overall exon usage, we use a compositional metric (Aitchison et al., 2000) (denoted by Adist) to compare the two probability vectors $\mathbf{Q}_1^E$ vs. $\mathbf{Q}_2^E$. This metric is widely recognized as the most appropriate geometry for the compositional data. Specifically, Adist $= (\sum_{i=1}^E (\log \frac{Q_1^{e_i}}{g(Q_1)} - \log \frac{Q_2^{e_i}}{g(Q_2)})^2)^{1/2}$, where $g(Q) = (\prod_{i=1}^E Q^{e_i})^{1/E}$. Then we calculate the posterior probability $P(\text{Adist} > \text{threshold})$ to quantify the difference (for example, threshold $= 1$).

Finally, we can convert all posterior probabilities to FDRs for the multiplicity control under an one-sided hypothesis. To borrow data information in the whole genome, we can add a hierarchical prior for the gene-level r's as specified in (8)–(10). Similarly, we can add a gamma hierarchical prior for r's in the two conditions such that the r's are not the same, but could be similar between conditions.

3.3 Transcript analysis

Algorithm 3 contains steps to sample parameters in the transcript analysis, which are similar to Algorithm 2 for the exon analysis except that we added two more steps to impute the unobserved transcript counts.

Algorithm 3 Transcript Analysis

Sample $p_k \sim \text{Beta}(a_0 + \sum_j y_{jk}, b_0 + n_k r),\ k = 1, 2$

Sample $l_{jk}^{t_{i'}} \sim \text{CRT}(y_{jk}^{t_{i'}}, r_k^{t_{i'}}),\ i' = 1, \cdots T; j = 1, \cdots, n_k$

Sample $\mathbf{Q}_k^T \sim \text{Dirichlet}(h_0 + \sum_j l_{jk}^{t_1}, \cdots h_0 + \sum_j l_{jk}^{t_T})$

Sample $r \sim \text{Gamma}(e_0 + \sum_k \sum_j \sum_{i'} l_{jk}^{t_{i'}}, f_0 - \sum_k n_k \log(1 - p_k))$

Calculate $r_k^{t_{i'}} = rQ_k^{t_{i'}}$, and $\tilde{r}_k^{t_{i'}} = r_k^{t_{i'}} / L_{t_{i'}}$

Sample $\lambda_k^{e_i t_{i'}} \sim \text{Gamma}(M_{e_i t_{i'}} S_{e_i} \tilde{r}_k^{t_{i'}}, (1 - p_k)/p_k),\ i = 1, \cdots E$

Sample $(y_{jk}^{e_i 1}, \cdots, y_{jk}^{e_i t_T}) \sim \text{Multinomial}(y_{jk}^{e_i}, (\frac{\lambda_k^{e_i t_1}}{\sum_{i'} \lambda_k^{e_i t_{i'}}}, \cdots, \frac{\lambda_k^{e_i t_T}}{\sum_{i'} \lambda_k^{e_i t_{i'}}}))$, and calculate $y_{jk}^{t_{i'}} = \sum_i y_{jk}^{e_i t_{i'}}$

The differential tests are the same as in the exon analysis with the following modifications based on suggestions in Cuffdiff (Trapnell et al., 2013).

1. The differential tests are based on transcript-length adjusted estimates $\mu^{*t_{i'}}$ and $Q^{*t_{i'}}$, specifically, $\mu^{*t_{i'}} = \frac{\mu_{i'}^t / L_{t_{i'}}}{\sum_{i'} \mu^{t_{i'}} / L_{t_{i'}}}$ and $Q^{*t_{i'}} = \frac{Q^{t_{i'}} / L_{t_{i'}}}{\sum_{i'} Q^{t_{i'}} / L_{t_{i'}}}$, where $L_{t_{i'}}$ is the effective length of transcript i'.

2. The expression for a particular gene is estimated using estimates of the transcript expression, that is, $\mu_k^* = \sum_{i'} \mu_k^{*t_{i'}}$ $(k = 1, 2)$. This estimate has a larger variance compared to the estimate directly using the total gene count (see Section 3.1), since it considers the uncertainty of assigning reads to transcripts.

4 Simulation studies

In this section, we ran extensive simulations to evaluate our model. First, we investigated our Gibbs sampling algorithms for the gene expression analysis. Secondly, we tested the robustness of our FNB model when the key assumption of the model is violated. Finally, we evaluated our FNB model for the exon analysis and compared simulation results to DEXSeq.

4.1 Gene expression analysis

There are many existing algorithms to perform the differential test for gene expression data. We set up simulation studies as described in Soneson and Delorenzi (2013), and compared our method to three popular methods, DESeq, edgeR and baySeq. The differential test for the gene expression using the total counts in a gene does not require the key assumption in the FNB model. Hence, we directly simulated gene counts from NB distributions with mean and dispersion parameters estimated from two real RNA-Seq data sets. One data set contains RNA-Seq data from 69 unrelated Nigerian individuals (Pickrell et al., 2010). The other data set contains RNA-Seq data from 41 unrelated Caucasian individuals of European descent (Cheung et al., 2010). For each data set, the

maximum likelihood estimates of the mean and the dispersion were obtained for each gene using all samples. Then we merged all pairs of mean and dispersion estimates to form a population of true parameters to be sampled in the simulation studies.

We studied two scenarios with 2 or 5 samples per condition. For each scenario, we simulated 10 data sets, with each data set consisting of 10,000 genes with mean and dispersion parameters randomly sampled from the population pool. The number of differentially expressed genes was set to 20%, and the $\log 2$ fold changes for the differentially expressed genes were generated from a standard normal distribution. In these simulations, the dispersions in both conditions were assumed to be identical. Both algorithms 1A and 1B were used for the estimation, with a hierarchical structure on the gene-level dispersion parameters, as specified in (8)–(10). We set $a_0 = b_0 = 1$ in the beta prior and $e_0 = f_0 = c_0 = 0.1$ in the gamma prior (vague priors). With a burn-in of 4000 iterations, an additional 6000 iterations were used for inference. We observed that the chain mixes well. Figure 1 shows that our method is comparable to DESeq, edgeR and baySeq. It is not surprising that all four methods give similar performances because all methods used a NB model for the differential test. When n is larger ($n = 5$), the FNB model seems to have a slightly higher area under the receiver operating characteristic (ROC) curve than DESeq2 and edgeR, and slightly smaller false discoveries than the baySeq. Our results show that our Gibbs sampling algorithms provide competing alternatives for the differential test in gene expression analysis.

4.2 Robustness of the FNB model

The key assumption of the FNB model is that the counts in exons/transcripts share the same parameter p. This assumption greatly simplifies the estimation of relative usage and the imputation of transcript counts. In this section, we conducted simulation studies to assess the robustness of the FNB model when this assumption is violated.

In all simulations, we generated datasets consisting of 24,000 exons with 2 conditions and 3 samples per condition. Each exon count was generated from a NB distribution with a mean and a probability parameter, p. We randomly sampled 24,000 means from the mean estimates derived from the pasilla dataset (Brooks et al., 2010). The proportion of differentially expressed exons and the distribution of fold change were the same as the gene expression simulations in Section 4.1. Parameter p's were simulated under four scenarios. In scenario I, p's were fixed to be 0.8 for all exons in both conditions. In the remaining three scenarios, p's were generated from a beta distribution in both conditions. Specifically, we used beta$(8, 2)$, beta$(4, 1)$, and beta$(1, 1)$ for scenario II, III, and IV, respectively. The mean of the distribution for p is 0.8 in scenarios II and III, which was determined based on the real dataset. In scenario IV, p's are uniformly distributed between 0 and 1. When p's have a larger variance, the assumption (i.e., the same p's for exons in the same gene) is more likely to be violated. Hence, scenarios II, III and IV studied the presence of mild, moderate and severe departure of the assumption, respectively.

In each scenario, there is a read count matrix of 24,000 rows, with each row representing an exon and a column representing a sample in a condition. Next, we used

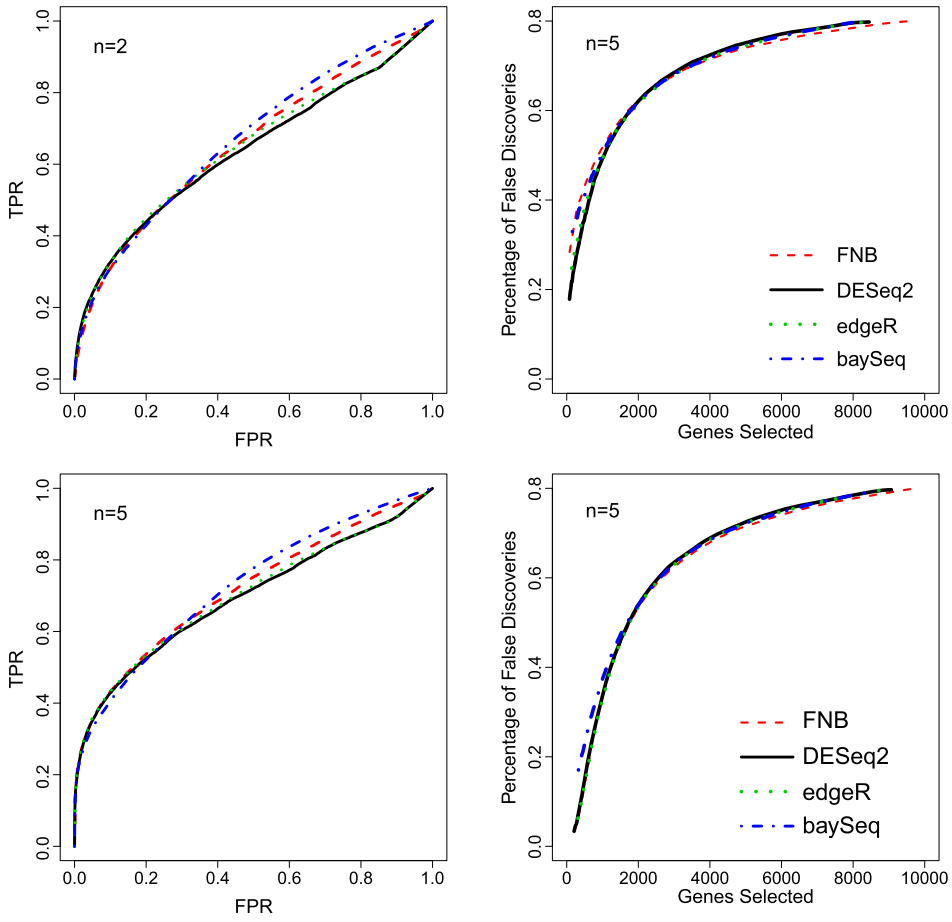

Figure 1: The results of our FNB model are based on P_β's obtained from Algorithm 1A. The left two plots show the averaged ROC curves, where the y axis is the true positive rate (TPR) and the x axis is the false positive rate (FPR). The right two plots show the averaged percentage false discoveries given the number of selected genes. Results based on Algorithm 1B are very similar to the results presented here.

six different ways to assign gene IDs, resulting in six different datasets: 1) every 30 exons share an unique gene ID, 2) every 10 exons share an unique gene ID, 3) every 6 exons share an unique gene ID, 4) every 3 exons share an unique gene ID, 5) every 2 exons share an unique gene ID and 6) each exon has an unique gene ID. It is important to note that the six dastasets under each scenario have exactly the same read count matrices but different gene IDs. Since the FNB model assumes that all exons in the same gene share a common p, the first five datasets allow us to investigate whether the FNB model is robust to the violation of the assumption when every 30, 10, 6, 3 or 2 exons share a common p if p's are slightly, moderately and dramatically different (scenarios II–IV). Among these studied cases, the first dataset in scenario IV represents

the strongest deviation from the assumption, because the model assumes that 30 exons have the same p's when the actual p's are likely to be very different. In scenario I, all six datasets satisfy the assumption, and the last dataset in scenarios II to IV also satisfies the assumption since each exon has its own p.

Finally, we applied our FNB model to 24 datasets (6 datasets $\times$ 4 scenarios), and results for the six datasets were compared in each scenario. In the FNB model, we used Algorithm 2 with a gamma hierarchical prior for r's in the two conditions. We set $a_0 = b_0 = 1$, $e_0 = f_0 = h_0 = 0.1$, and $c_0 = 0.01$ (weak priors). With a burn-in of 4000 iterations, an additional 6000 iterations were used for inference. We also applied DESeq2 to the last dataset in each scenario and compared to the FNB model.

To evaluate the model performance, we calculated the area under the ROC curve (AUC). Additionally, we computed model selection criteria to assess the goodness-of-fit of the six FNB models within each scenario. We chose log-pseudo marginal likelihood (LPML) (Ibrahim et al., 2001; Chen et al., 2000) and Watanabe–Akaike information criterion (WAIC) (Watanabe, 2010). LPML is a cross-validated leave-one-out measure of a model's ability to predict the data. It is valid for small and large samples and does not suffer from a heuristic justification based on large sample normality. Based on Gelfand and Dey, 1994, the LPML implicitly includes a similar dimensional penalty as AIC asymptotically. WAIC was proposed recently and can also be viewed as an improvement over the standard DIC and it estimates pointwise out-of-sample prediction accuracy from a fitted Bayesian model (see supplementary materials (Zhao et al., 2017)). The best model should have the smallest WAIC and |LMPL|.

As shown in Figure 2, the FNB model were very robust in all studied scenarios. The differences in AUC are less than 0.005 even in scenario IV where p's are vastly different. Moreover, when p's are the same as in scenario I, the FNB model gained power as more exons were used to estimate the p parameter. To our surprise, a similar gain in power was also observed in scenario II where p's are different (p ranges from 0.186 to 0.999). Compared to the AUC in DESeq2, the FNB model is better in 3 out of 4 studied scenarios.

The model fit statistics in Figure 3 appear to tell us the same story as in the AUC results. The differences in LPML and WAIC statistics are very small in all studied scenarios.

4.3 Exon-level data analysis

We conducted extensive simulation studies to evaluate our FNB model for the exon analysis. In the simulations, we compared the FNB model to DEXSeq. DEXSeq is a commonly used package for testing the exon relative usage (i.e., the fraction of the gene's reads that falls into the exon) between conditions. It estimates the relative usage for each exon separately. Specifically, each exon is a factor with two levels, this (i.e., the number of reads to the exon in question) and others (i.e.,the read counts from all other exons of the same gene), and an interaction between exon and condition is tested in a generalized linear model. In contrast, the FNB model estimates the relative usage for all exons simultaneously. Furthermore, it tests a global hypothesis for the overall relative

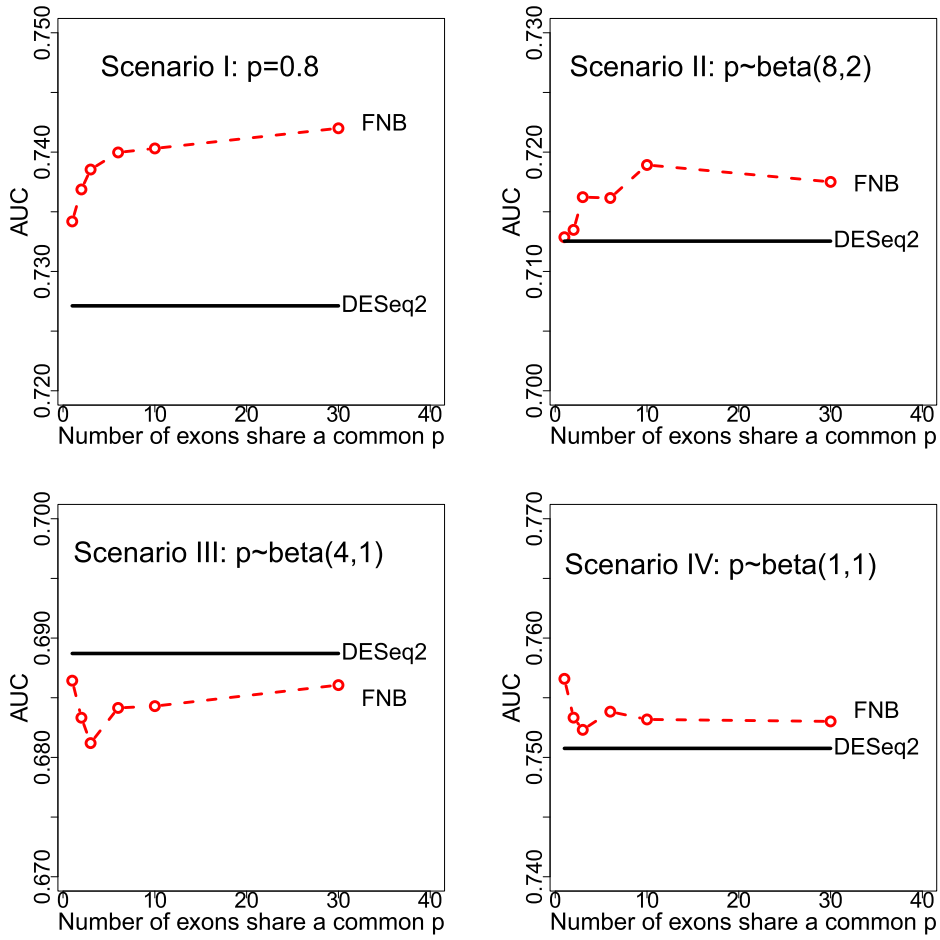

Figure 2: Simulation results under four scenarios. In DESeq2, each exon was treated as a single gene. This value was plotted over different number of exons to visually compare with the FNB models under different assumptions.

usage. In addition to the exon usage, the FNB model also tests for the differential expression of exons.

In all simulation studies, we generated datasets with 2 conditions with 3 samples per condition. Each simulated dataset consists of 10,000 genes with mean expressions randomly sampled from two real RNA-Seq data sets as described in 4.1. The mean expression for each exon was calculated based on its relative usage, which was simulated from a Dirichlet distribution. Finally, read counts in each exon were generated from a NB distribution.

We assumed that 3 or 6 exons per gene in each dataset, which produces a dataset with 30,000 or 60,000 exons (call it a 3-exon or 6-exon dataset). To test for the exon

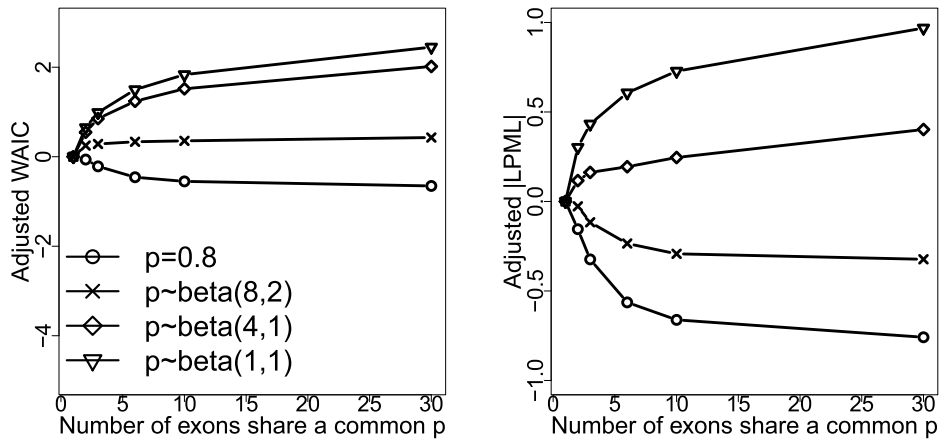

Figure 3: Each curve indicates a different scenario. When each exon was assumed to have a different p (i.e., number of exons sharing a common p is one), |LMPL| statistics are 24.4, 26.4, 24.8 and 22.2 for scenario I to IV, respectively, and WAIC statistics are 47.0, 47.6, 47.8,and 43.4 for scenario I to IV, respectively (call these statistics as the baseline values). The WAIC and |LMPL| statistics on the y-axis were adjusted by the baseline values such that the first value in each scenario is zero.

expression and the relative usage at the same test, we generated 70% genes with the same expression and same exon relative usage between conditions, 20% genes with the same expression and differential exon relative usage between conditions, and 10% genes with differential expression and same exon relative usage between conditions. The exon expression is calculated by multiplying the gene expression and the relative usage. Thus, each dataset includes exons with the same and different expressions/usages between conditions, and genes with the same and different overall exon relative usages.

To generate gene expressions, we used a similar approach as described in Section 4.1. For DE genes, $\log 2$ fold changes were simulated from a normal distribution with mean 0 and standard deviation 2. The relative usages were generated from a Dirichlet distribution in both conditions. Specifically, a Dirichlet(1,4,8) for the 3-exon dataset and a Dirichlet(1,4,8,1,2,2) for the 6-exon dataset. The averaged usage is (0.08 0.31 0.61) and (0.06,0.22,0.44,0.06,0.11,0.11), respectively. Table 2 shows concentration parameters in the Dirichlet distributions used to generate the relative usage in condition 2. For the 3-exon dataset, we studied two scenarios. In scenario I, usages of the first two exons are switched in condition 2, and the differences in usage between conditions are about 23% (i.e., 0.31-0.08=23%) on average. For the 6-exon dataset, we studied three scenarios. In the last scenario, all six exons have different usages between conditions.

Based on the mean gene expression and the relative usage described as above, we calculated the mean expression for each exon. To generate exon read counts from a NB distribution, we also need to determine a probability parameter p. We generated p's from three distributions. The first distribution assumes that p's for exons in the same

Exon per gene	Scenario	Condition 1	Condition 2	Num. exon with different usage
3	I	(1,4,8)	(4,1,8)	2
3	II	(1,4,8)	(8,4,1)	3
6	I	(1,4,8,1,2,2)	(4,1,8,1,2,2)	2
6	II	(1,4,8,1,2,2)	(8,4,1,1,2,2)	3
6	III	(1,4,8,1,2,2)	(2,2,1,8,4,1)	6

Table 2: Concentration parameters in the Dirichlet distribution for generating the exon relative usage in two conditions.

gene are the same. The last two distributions allow p's to be different (a more realistic assumption). To determine the p's in the first distribution, we first sampled gene-level dispersion parameters, r's, from the two real RNA-Seq data sets as described in Section 4.1, and then obtained the p by $p = \mu/(\mu + r)$ (here, μ is the mean gene expression). In the other two distributions, p's were simulated from a beta distribution, i.e., beta(8,2) or a beta(1,1) as described in Section 4.2.

Finally, we applied the FNB model to six 3-exon datasets (2 scenarios for relative usage $\times$ 3 distributions for p), and nine 6-exon datasets (3 scenarios for relative usage $\times$ 3 distributions for p). As shown in Table 3, the FNB model tests for the exon expression, relative usage and the overall relative usage simultaneously. In contrast, DEXSeq only tests for the exon relative usage. With regarding to the relative usage, the FNB model exhibited a dramatically better performance than DEXSeq when the model assumption is satisfied. It performed similarly to DEXSeq when the assumption is violated, and it seems to be slightly better if the violation is not so strong (all ROC and FDR curves can be found in supplementary materials (Zhao et al., 2017)). Additionally, tests for the overall relative usage are always more powerful than the tests for a single exon usage.

In simulation studies we did not compare FNB to Cuffdiff, because we directly generated read count data from R package and it's technically challenging to compare to Cuffdiff which has its own pipeline that takes raw sequencing reads. However, we illustrated the advantage of FNB over Cuffdiff in the real data analysis.

5 Real data analysis

In this section we applied our FNB model on a real dataset (Brooks et al., 2010). The experiment investigated the effect of siRNA knock-down of pasilla, a gene that is known to bind to mRNA in the spliceosome, which is thought to be involved in the regulation of splicing. The dataset is available in DEXSeq package, which contains 3 biological replicates of the knockdown as well as 4 biological replicates for the untreated control. BDGP5.gtf was used as the reference transcriptome annotation.

5.1 Gene expression analysis

To test the differential gene expression, we used Algorithm 1B with the same priors as in the simulation studies. For each gene, we estimated prior π_1/π_0 to be 0.15 using the

Simulation set-up			DEXSeq	FNB		
exon per gene	probability parameters	scenario	exon usage	exon usage	overall usage	exon expression
3	Same	I	75.4	76.3	79.8	70.4
3	Same	II	79.4	81	88.7	72.9
3	beta(8,2)	I	80.6	81.8	86.1	75.3
3	beta(8,2)	II	87.5	87.5	93.6	79.1
3	beta(1,1)	I	85.1	85.2	89.5	78
3	beta(1,1)	II	88.6	88.3	94.9	81.1
6	Same	I	72.4	74.6	76.6	65
6	Same	II	73.6	76.7	82.7	66.9
6	Same	III	72	74.4	87.1	68.7
6	beta(8,2)	I	78.8	79.1	85.3	69.3
6	beta(8,2)	II	82.8	82.6	89.3	72.8
6	beta(8,2)	III	79.4	79	91.7	74.7
6	beta(1,1)	I	84.6	84.8	89.3	74.3
6	beta(1,1)	II	85.5	84.9	93.3	76.2
6	beta(1,1)	III	83.9	83.6	95.5	80.4

Table 3: AUC for the analysis of exon relative usage by the FNB model and DEXSeq. The FNB model also provides the analysis for the exon expression and overall relative usage.

qvalue package (Storey and Tibshirani, 2003). When the FDR was controlled at 5%, we compared the number of DE genes selected from the FNB model to that identified in DESeq, edgeR and baySeq. Figure 4 shows a good overlap among the set of DE genes selected by the four methods.

5.2 Exon analysis

In the data set, 70% genes have less than 6 exons and 90% genes have less than 11 exons. We applied the FNB model to the pasilla data for the exon analysis. We normalized the count data using the default method in DEXSeq and deleted exons with the total sample counts ≤ 5. We used Algorithm 2 for the statistical inference with $a_0 = b_0 = 1$ and $e_0 = f_0 = h_0 = 0.1$ (r's on the gene-level were assumed to be the same between conditions based on the goodness-of-fit statistics). With a burn-in of 4000 iterations, an additional 6000 iterations were used for inference.

The model simultaneously provided test results for exon expression, relative usage for each exon, and the overall relative usage within a gene. We compared our results to DEXSeq (Anders et al., 2012) for the relative usage of each exon. In DEXSeq, the relative usage for a particular exon is tested through testing the interaction between that exon (defined as "this" and "others"; see Section 2.5) and condition in a generalized linear model. This test is the same as comparing $Q_2^{e_i}$ and $Q_1^{e_i}$ $(i = 1, \cdots, E)$ in the FNB

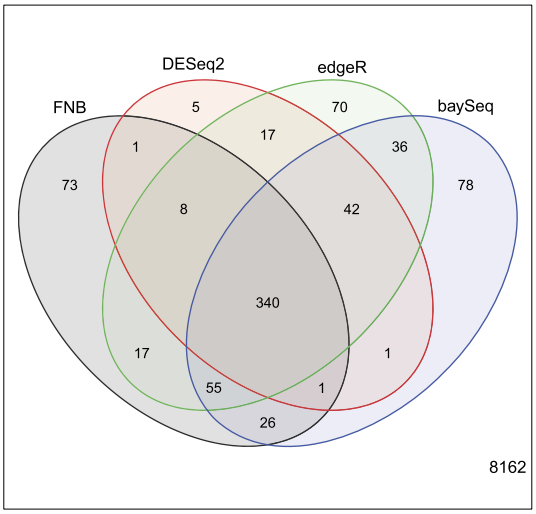

Figure 4: Overlap among the set of DE genes found by four methods. The number at the bottom right is the number of genes not selected by any of the four methods.

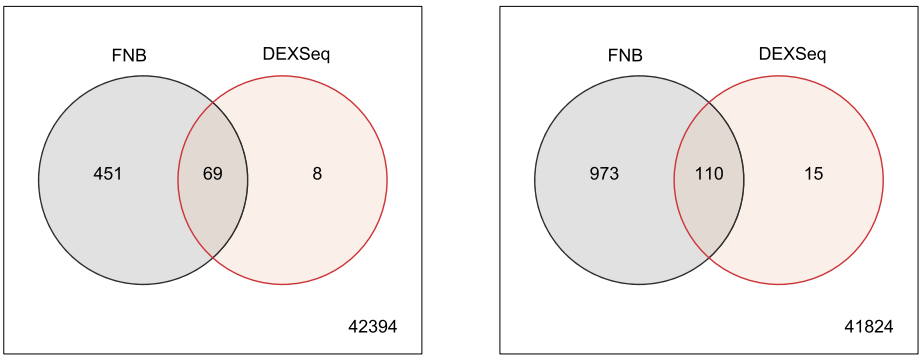

Figure 5: Overlap among the set of DE exons found by the two methods for under-used (left) and over-used exons (right) in the pasilla data. The number at the bottom right is the number of exons not selected by FNB and DEXSeq.

model as both methods compare whether the fraction of the gene's reads that fall into the exon differs significantly between conditions.

Figure 5 shows a very good overlap between the two methods. The FNB model identified 520 ($\approx 1\%$) significantly under-used exons (i.e., $Q_2^{e_i} < Q_1^{e_i}$; $k = 1$ for the controls) while DEXSeq identified 77. When the FDR was controlled at 2.5%. About 90% of those exons identified by DEXSeq were also selected by FNB. A similar conclusion was obtained for the over-used exons (i.e., $Q_2^{e_i} > Q_1^{e_i}$).

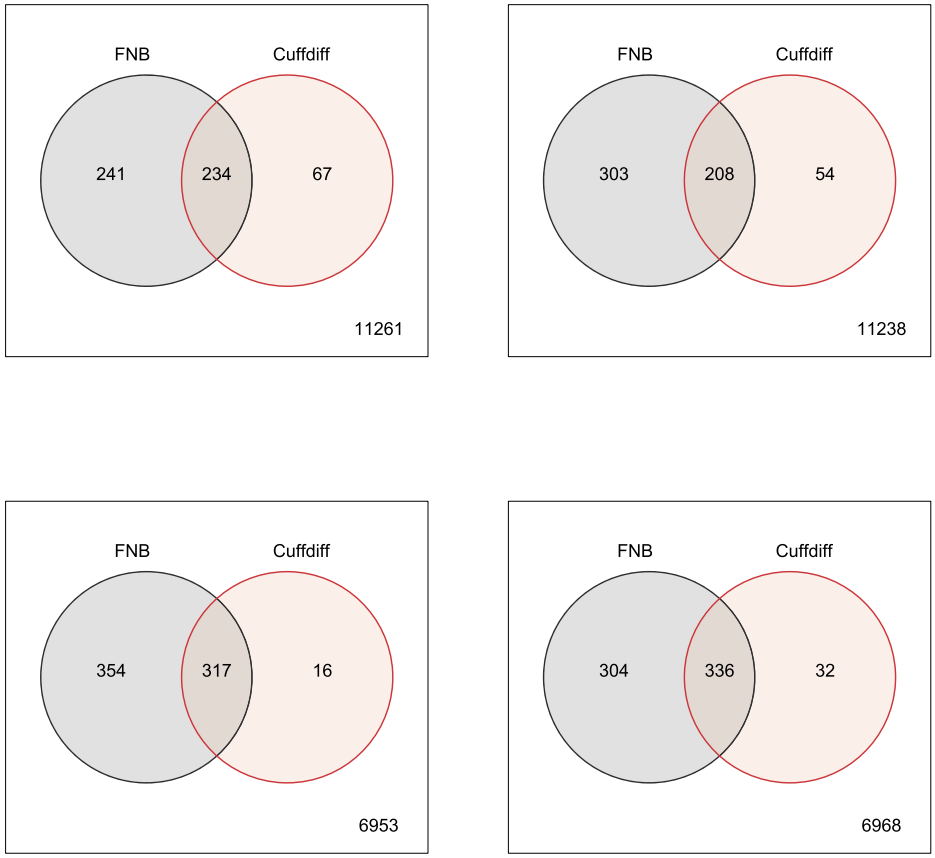

Figure 6: Overlap among the set of DE transcripts (top) and DE genes (bottom) found by the two methods for under-expressed (left) and over-expressed genes/transcripts (right).

It is important to note that the comparison to DEXSeq is just to show that the FNB model provided reasonable results for the real data analysis. However, the important feature of the FNB model is its ability to test multiple hypotheses. For example, the FNB model shows that 2.4% gene have large differences in the overall usage between two conditions (median Adist larger than 20). Another example of the FNB model used to evaluate composite hypotheses was presented in the transcript analysis.

5.3 Transcript analysis

Algorithm 3 was used for the inference with the same priors as in the exon analysis. With a burn-in of 4000 iterations, an additional 6000 iterations were used for inference. Posterior distributions of $\mu_k^{*t_{i'}}$ and μ_k^* $(k = 1, 2)$ were used to test differential transcript and gene expression, respectively, as discussed in Section 3.3. Then we compared DE

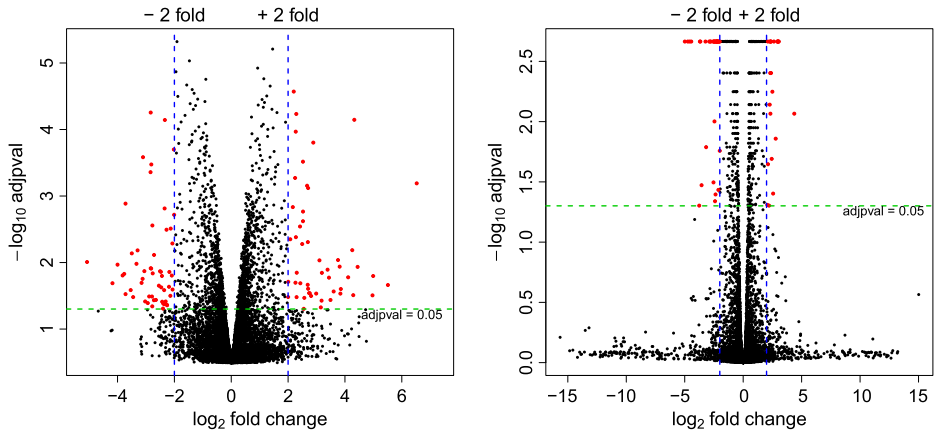

Figure 7: $\log_2$ fold changes are on the x-axis fold-change and $\log_{10}$-transformed adjusted p-values are on the y-axis.The red points indicate DE transcripts that display both large-magnitude fold-changes as well as high statistical significance. The dashed green-line shows the adjusted p-value cutoff. The vertical dashed blue lines shows 2-fold changes.

transcripts and genes identified by the FNB model to that selected by Cuffdiff (Trapnell et al., 2013) (see details of the parameter setting in Cuffdiff in supplementary materials (Zhao et al., 2017)). The comparisons were based on all transcripts/genes that were testable by Cuffdiff (the status in the output file showed "OK"). As shown in Figure 6, there is a good overlap among the set of DE transcripts/genes selected by the FNB model and Cuffdiff, and the FNB model identified more DE transcripts and genes.

It is important to note that Cuffdiff relies on a two-step method for inference (i.e., impute transcript counts + estimate expression), while the FNB model performs the two steps simultaneously. Figure 7 plots the statistical significance versus estimated transcript fold-changes for both methods. It seems that Cuffdiff missed out majority of transcripts with a large magnitude of fold change.

One appealing feature of the FNB model is its flexibility to test various hypotheses. It not only allows us to test the relative usage for a particular transcript, but also allows a test for the overall usage for a particular gene (see Section 3.3). Combined with the test for gene expression, we can categorize each gene into one of the four groups: (1) Non-DE gene and transcript usage, (2) DE gene and Non-DE transcript usage, (3) Non-DE gene and DE transcript usage, or (4) DE gene and transcript usage (Shi and Jiang (2013) considered (3) and (4) as a single group). Figure 8 depicts a representative gene for each of the four groups.

5.4 Goodness of fit of the FNB model

In the exon analysis, we compared the LPML and WAIC statistics for our FNB model, a Poisson model, and a NB model, in which each exon has its own dispersion and

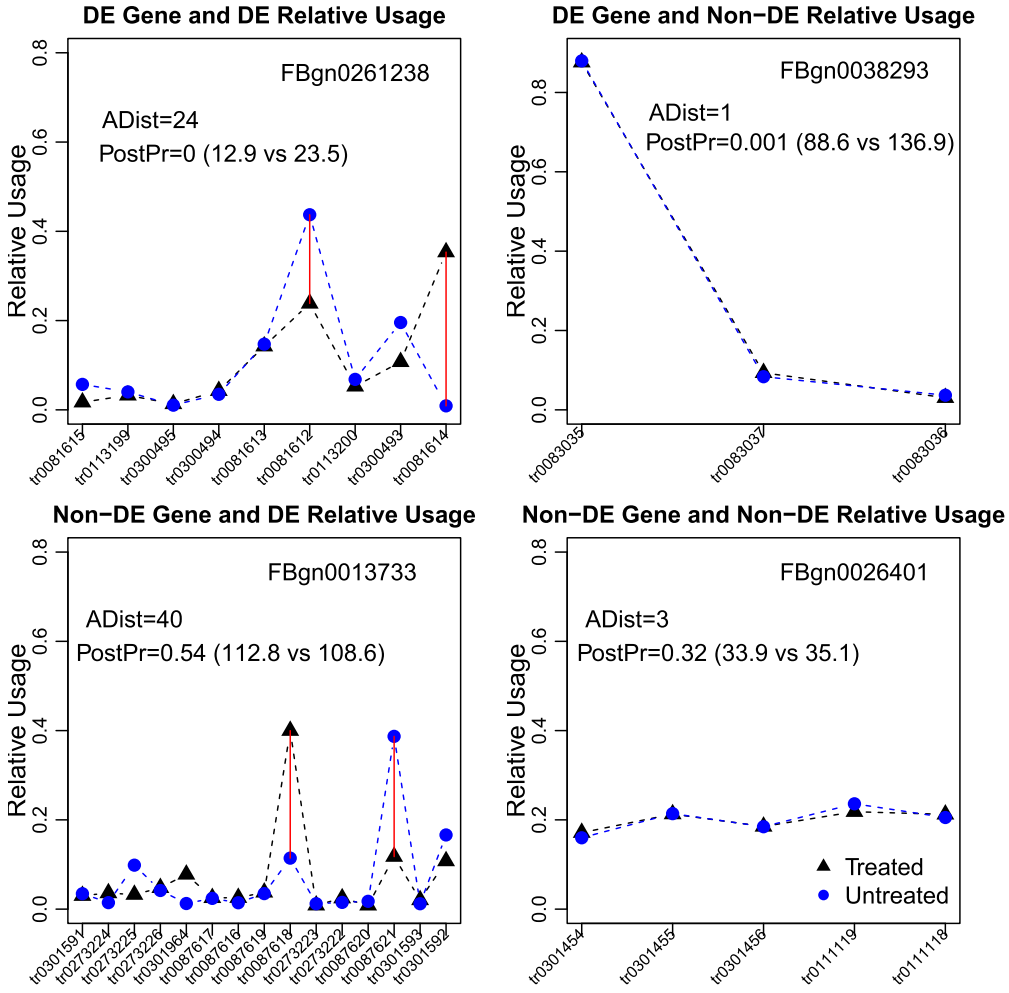

Figure 8: Plot of four possible scenarios for the analysis of gene expression and relative transcript usage. The transcripts with differential relative usage are indicated by the red lines. The difference in the overall relative usage is indicated by Adist, with a large value suggesting a large difference; PostPr= $P(\mu_2^* > \mu_1^*)(\mu_2^*$ vs $\mu_1^*)$, and a value of PostPr close to 0 or 1 suggests an under- or over-expressed gene.

probability parameter (call it an IndNB model). In the Poisson model, the Gamma-Poisson conjugacy was used to update the rate parameter of the poisson distribution for each exon. In the IndNB model, Algorithm 1A was used to estimate r^e and p_k^e by treating each exon as a single gene. In the transcript analysis, we compared the FNB model to a Poisson model based on the fit on the exon-level observed data. In the Poisson model, a multinomial distribution was used to impute the latent transcript counts as described in Turro et al. (2011). In this comparison, we did not evaluate an IndNB model

		|LMPL|		**WAIC**	
Analysis	Model	Non-est	Mean	Non-est	Mean
Exon	FNB	0	28	0	51
	IndNB	0	28	0	56
	Poisson	91	50	82	108
Transcript	FNB	1097	275	1054	469
	Poisson	4518	626	4380	754

Table 4: Goodness-of-fit statistics for the exon analysis in pasilla data. Non-est denotes the number of non-estimable exons having the worst fit (the fit statistic is Infinity). Mean is the averaged |LMPL| or WAIC among estimable exons. The best model should have the smallest |LMPL| and WAIC. The non-estimable exons (1097/1054) in the FNB model are a subset of the non-estimable exons (4518/4380) in the Poisson model.

(i.e., transcripts have different NB distributions), because the transcript imputation and parameter estimation can not be performed simultaneously under such model.

Table 4 shows that the FNB model dramatically improved the fit of the data compared to the Poisson model, and the fit statistics are similar between FNB and IndNB model. Compared to the IndNB model, the FNB model is much more appealing, in that 1) greatly simplifies the estimation for (overall) relative usage, and 2) allows straightforward imputation of the unobserved transcript counts.

6 Discussions

We have developed a family of NB models that addressed the downstream analysis tasks for count-based RNA-Seq data. To our best knowledge, this is the only method which integrates the gene, exon and transcript analysis under an unified NB modelling framework. The computational algorithms are very simple because of the available conjugate forms in our FNB model. It easily incorporates the uncertainty of assigning reads to transcripts using the Gamma-Poisson mixture formulation of the NB distribution. Furthermore, it greatly simplifies the estimation of the relative exon/transcript usage, by reducing a complex problem of estimating the relative usage in a NB model to a simple problem of estimating of the proportion of latent counts from a multinomial distribution.

Our simulation studies showed that our FNB model is very robust to the violation of the model assumption. In real data analysis, we compared the FNB model to 1) DESeq, edgeR and baySeq for the gene expression analysis, 2) DEXSeq for the exon usage analysis, and 3) Cuffdiff for the differential transcript analysis. We have demonstrated that the FNB model provided a good fit to the real data. It also allows us to test various hypotheses of interest at the same time, including the exon/transcript expression, exon/transcript relative usage and overall relative usage.

Our method is relatively computationally intensive, but has been implemented to take advantage of parallel processing. It took approximately 3 hours for the transcript and exon analysis in the pasilla dataset, running on a quad-core Intel Xeon 2.10 GHz

8 GB RAM x64 computer. Frequentist approaches, such as DESeq, edgeR, baySeq, or DEXSeq, took just a couple of minutes. However, these methods only test a single hypothesis, while the FNB model is able to test multiple hypotheses simultaneously and provides a more comprehensive analysis than the previous mentioned methods.

In the transcript analysis, the FNB model relies on the M matrix to impute the count data in transcripts. In this work the M matrix was created using DEXSeq. Additionally, this M matrix can be obtained using other software, such as rSeq (Salzman et al., 2011; Shi and Jiang, 2013) and MMSEQ (Turro et al., 2011). In MMSEQ, the insert size from paired-end data can be considered to construct the M matrix.

7 Data and software

The R code and data are available from http://www-personal.umich.edu/~zhaolili/FNBSeq/.

Supplementary Material

Supplementary Materials for "Bayesian Analysis of RNA-Seq Data Using a Family of Negative Binomial Models" (DOI: 10.1214/17-BA1055SUPP; .pdf).

References

Aitchison, J., Barcelo-Vidal, C., Martin-Fernandez, J. A., and Pawlowsky-Glahn, V. (2000). "Logratio analysis and compositional distance." *Mathematical Geology*, 32: 271–275. 9

Anders, S. and Huber, W. (2010). "Differential expression analysis for sequence count-data." *Genome Biology*, 11: R106. 2

Anders, S., Reyes, A., and Huber, W. (2012). "Detecting differential usage of exons from RNA-seq data." *Genome Research*, 22: 2008–2017. 1, 2, 4, 17

Brooks, A. N., Yang, L., Duff, M. O., Hansen, K. D., Park, J. W., Dudoit, S., Brenner, S. E., and Graveley, B. R. (2010). "Conservation of an RNA regulatory map between Drosophila and mammals." *Genome Research*, 193–202. 11, 16

Chen, M.-H., Shao, Q., and Ibrahim, J. G. (2000). *Monte Carlo Methods in Bayesian Computation*. New York: Springer. MR1742311. doi: http://dx.doi.org/10.1007/978-1-4612-1276-8. 13

Cheung, V. G., Nayak, R. R., Wang, I. X., Elwyn, S., Cousins, S. M., Morley, M., and Spielman, R. S. (2010). "Polymorphic cis- and trans-regulation of human gene expression." *PLoS Biology*, 8: e1000480. 10

Di, Y., Schafer, D. W., and nd J. H. Chang, J. S. C. (2011). "The NBP negative binomial model for assessing differential gene expression from RNA-seq." *Statistical Applications in Genetics and Molecular Biology*, 10: Article 24. 2

Dillies, M.-A., Rau, A., Aubert, J., Hennequet-Antier, C., Jeanmougin, M., Servant, N., Keime, C., Marot, G., Castel, D., Estelle, J., Guernec, G., Jagla, B., Jouneau, L., Laloë, D., Gall, C. L., Schaëffer, B., Crom, S. L., Guedj, M., and Jaffrézic, F. (2013). "A comprehensive evaluation of normalization methods for Illumina high-throughput RNA sequencing data analysis." *Briefings in Bioinformatics*, 14: 671–683. 1

Gelfand, A. E. and Dey, D. K. (1994). "Bayesian model choice: asymptotics and exact calculations." *Journal of the Royal Statistical Society*, 56: 501–514. 13

Hardcastle, T. J. and Kelly, K. A. (2010). "baySeq: empirical Bayesian methods for identifying differential expression in sequence count data." *BMC Bioinforma*, 11: 422. 2

Ibrahim, J. G., Chen, M.-H., and Sinha, D. (2001). *Bayesian Survival Analysis*. New York: Springer. 13

Kim, D., Pertea, G., Trapnell, C., Pimentel, H., Kelley, R., and SL, S. L. S. (2013). "TopHat2: accurate alignment of transcriptomes in the presence of insertions, deletions and gene fusions." *Genome Biology*, 14: R36. 1

Leng, N., Dawson, J. A., Thomson, J. A., Ruotti, V., Rissman, A. I., Smits, B. M., Haag, J. D., Gould, M. N., Stewart, R. M., and Kendziorski, C. (2013). "EBSeq: An empirical Bayes hierarchical model for inference in RNA-seq experiments." *Bioinformatics*. 2

León-Novelo, L. G., Müller, P., Arap, W., Kolonin, M., Sun, J., Pasqualini, R., and Do, K.-A. (2013). "Semi-parametric Bayesian inference for phage display data." *Biometrics*, 69: 174–183. 7

Lewin, A., Bochkina, N., and Richardson, S. (2007). "Fully Bayesian mixture model for differential gene expression: simulations and model checks." *Statistical Applications in Genetics and Molecular Biologys*, 6: Article36. 7

Love, M. I., Huber, W., and Anders, S. (2014). "Moderated estimation of fold change and dispersion for RNA-Seq data with DESeq2." *Genome Biology*, 15: 550. 2

Mortazavi, A., Williams, B. A., McCue, K., Schaeffer, L., and Wold, B. (2008). "Mapping and quantifying mammalian transcriptomes by RNA-Seq." *Nature Methods*, 5: 621–628. 1

Niu, L., Huang, W., Umbach, D. M., and Li, L. (2014). "IUTA: a tool for effectively detecting differential isoform usage from RNA-Seq data." *BMC Genomics*, 15: 862. 2, 6

Pickrell, J. K., Marioni, J. C., Pai, A. A., Degner, J. F., Engelhardt, B. E., Nkadori, E., Veyrieras, J. B., Stephens, M., Gilad, Y., and Pritchard, J. K. (2010). "Understanding mechanisms underlying human gene expression variation with RNA sequencing." *Nature*, 464: 768–772. 10

Rapaport, F., Khanin, R., Liang, Y., Pirun, M., Krek, A., Zumbo, P., Mason, C. E., Socci, N. D., and Betel, D. (2013). "Comprehensive evaluation of differential gene expression analysis methods for RNA-seq data." *Genome Biology*, 14: R95. 1

Robinson, M. D., McCarthy, D. J., and Smyth, G. K. (2010). "edgeR: a Bioconductor package for differential expression analysis of digital gene expression data." *Bioinformatics*, 26: 139–140. 2

Salzman, J., Jiang, H., and Wong, W. H. (2011). "Statistical Modeling of RNA-Seq Data." *Statistical Science*, 26. 23

Shi, Y. and Jiang, H. (2013). "rSeqDiff: Detecting Differential Isoform Expression from RNA-Seq Data Using Hierarchical Likelihood Ratio Test." *PLoS ONE*, 8: e79448. 20, 23

Soneson, C. and Delorenzi, M. (2013). "A comparison of methods for differential expression analysis of RNA-seq data." *Bioinformatics*, 14: 91. 10

Storey, J. D. and Tibshirani, R. (2003). "Statistical significance for genomewide studies." *Proceedings of the National Academy of Sciences of the United States of America*, 100: 9440–9445. 17

Trapnell, C., Hendrickson, D. G., Sauvageau, M., Goff, L., Rinn, J. L., and Pachter, L. (2013). "Differential analysis of gene regulation at transcript resolution with RNA-seq." *Nature Biotechnology*, 31: 46–53. 2, 6, 10, 20

Turro, E., Su, S.-Y., Goncalves, A., Coin, L. J., Richardson, S., and Lewin, A. (2011). "Haplotype and isoform specific expression estimation using multi-mapping RNA-seq reads." *Genome Biology*, 12: R13. 4, 5, 21, 23

van de Wiel, M. A., Leday, G., Pardo, L., Rue, H., der Vaart, A. W. V., and Wieringen, W. N. V. (2012). "Bayesian analysis of RNA sequencing data by estimating multiple shrinkage priors." *Biostatistics*, 14: 113–128. 2

van de Wiel, M. A., Neerincx, M., Buffart, T. E., Sie, D., and Verheul, H. M. (2014). "ShrinkBayes: a versatile R-package for analysis of count-based sequencing data in complex study designs." *Bioinformatics*, 15: 116. 2

Watanabe, S. (2010). "Asymptotic equivalence of Bayes cross validation and widely applicable information criterion in singular learning theory." *Journal of Machine Learning Research*, 11: 3571–3594. MR2756194. 13

Wu, H., Wang, C., and Wu, Z. (2013). "A new shrinkage estimator for dispersion improves differential expression detection in RNA-seq data." *Biostatistics*, 14: 232–243. 2

Zhao, L., Wu, W., Feng, D., Jiang, H., and Nguyen, X.L. (2017). "Supplementary Materials for "Bayesian Analysis of RNA-Seq Data Using a Family of Negative Binomial Models"" *Bayesian Analysis*. doi: http://dx.doi.org/10.1214/17-BA1055SUPP. 7, 8, 13, 16, 20

Zhou, M. and Carin, L. (2012). "Augment-and-conquer negative binomial processes." *NIPS*. 2, 3

Zhou, M. and Carin, L. (2015). "Negative binomial process count and mixture modelling." *IEEE*, 37: 307–320. 2, 3

Acknowledgments

The authors gratefully acknowledge the constructive comments of three referees, an associate editor and an editor, and thank Mingyuan Zhou at The University of Texas at Austin for helpful discussions on the negative binomial process.