Bioinformatics, 34(19), 2018, 3349–3356
doi: 10.1093/bioinformatics/bty330
Advance Access Publication Date: 24 April 2018
Original Paper

OXFORD

Gene expression

Bayesian negative binomial regression for differential expression with confounding factors

Siamak Zamani Dadaneh[1,*], Mingyuan Zhou[2,*] and Xiaoning Qian[1,*]

[1]Department of Electrical and Computer Engineering, TEES-AgriLife Center for Bioinformatics and Genomic Systems Engineering, Texas A&M University, College Station, TX 77843, USA and [2]Department of Information, Risk, and Operations Management, The University of Texas at Austin, Austin, TX 78712, USA

*To whom correspondence should be addressed.
Associate Editor: Bonnie Berger

Received on October 20, 2017; revised on March 27, 2018; editorial decision on April 19, 2018; accepted on April 20, 2018

Abstract

Motivation: Rapid adoption of high-throughput sequencing technologies has enabled better understanding of genome-wide molecular profile changes associated with phenotypic differences in biomedical studies. Often, these changes are due to multiple interacting factors. Existing methods are mostly considering differential expression across two conditions studying one main factor without considering other confounding factors. In addition, they are often coupled with essential sophisticated *ad-hoc* pre-processing steps such as normalization, restricting their adaptability to general experimental setups. Complex multi-factor experimental design to accurately decipher genotype-phenotype relationships signifies the need for developing effective statistical tools for genome-scale sequencing data profiled under multi-factor conditions.

Results: We have developed a novel Bayesian negative binomial regression (BNB-R) method for the analysis of RNA sequencing (RNA-seq) count data. In particular, the natural model parameterization removes the needs for the normalization step, while the method is capable of tackling complex experimental design involving multi-variate dependence structures. Efficient Bayesian inference of model parameters is obtained by exploiting conditional conjugacy via novel data augmentation techniques. Comprehensive studies on both synthetic and real-world RNA-seq data demonstrate the superior performance of BNB-R in terms of the areas under both the receiver operating characteristic and precision-recall curves.

Availability and implementation: BNB-R is implemented in R language and is available at https://github.com/siamakz/BNBR.

Contact: siamak@tamu.edu or mingyuan.zhou@mccombs.utexas.edu or xqian@ece.tamu.edu

Supplementary information: Supplementary data are available at *Bioinformatics* online.

1 Introduction

High-throughput sequencing technologies have become the basic practice for genomic studies in life science research (Wang *et al.*, 2009). In particular, RNA sequencing (RNA-seq), which measures the expression of each gene or genomic feature of interest by counting the number of sequence reads mapped to them, has been widely adopted for genotype-phenotype association studies. To identify the genes that are differentially expressed between different groups of samples as candidate biomarkers across different phenotypes or treatment conditions, a large number of statistical methods and tools have been developed (Anders and Huber, 2010; Dadaneh *et al.*, 2017; Law *et al.*, 2014; Li and Tibshirani, 2013; Love *et al.*, 2014; Robinson *et al.*, 2010).

While the majority of differential expression (DE) analyses are conducted with respect to a main treatment factor, the presence of potential confounding factors in real-world experiments makes it desirable to take them into account in the developed tools to derive unbiased genotype-phenotype association results. There exists a rich

set of methods on addressing this problem in microarray data analysis, such as the ones developed based on linear models (Smyth, 2004, 2005).

Unlike microarray data that are based on continuous intensity measurements, RNA-seq data have unique properties. The sequencing read counts are often skewed and highly over-dispersed (Datta and Nettleton, 2014), and most of the existing data are with a small number of samples due to high data collection cost. Analyzing RNA-seq data are challenging, especially if taking into account the potential confounding effects. A number of methods extend the statistical tools developed for microarray data to analyze RNA-seq read counts. For instance, `voom` (Law *et al.*, 2014) estimates the mean-variance relationship of the logarithmic transformation of counts, and after generating precision weights for each observation, exploits the empirical Bayes pipeline of `limma` (Smyth, 2005) for downstream analyses. Other statistical methods are specifically designed for RNA-seq count data. One of the most popular solutions in this category to account for over-dispersion due to biological variations is using the negative binomial (NB) distribution. Several DE analysis methods have employed generalized linear models (GLMs) to adapt the NB distribution to experiments with complex design. For example, two widely used methods, `edgeR` (Robinson *et al.*, 2010) and `DESeq2` (Love *et al.*, 2014), both use GLMs to model the mean of the NB distribution as a log-linear function of the covariates. The gene-wise dispersion parameters are then estimated using adjusted profile likelihood and GLM coefficients are estimated using Fisher scoring iterations.

For all these existing RNA-seq analysis methods, a common pre-processing step is to normalize the sequencing counts to compensate the variations of the sequencing depths across samples (Soneson and Delorenzi, 2013). For instance, `edgeR` (Robinson *et al.*, 2010) either calculates a trimmed mean of M-values between each pair of samples or uses an upper quantile of samples for normalization, while `DESeq2` (Love *et al.*, 2014) takes the median of the ratios of observed sample counts to the geometric mean across samples as a scaling factor for that specific sample. Normalizing the sequencing counts, however, makes the performance depend on whether the introduced normalization is appropriate for the structure of the RNA-seq data under study (Zyprych-Walczak *et al.*, 2015). In addition to normalization, another common pre-processing practice is to perform the surrogate variable analysis (SVA) to identify potential unknown factors that may help model batch effects, and then incorporate these surrogate variables (SVs) as additional covariates to adjust for the consequent DE analysis (Leek *et al.*, 2012; Leek, 2014).

Obviating the need to pre-process the data, `BNP-Seq` (Dadaneh *et al.*, 2017) uses a stochastic process based approach to model the observed sample-gene random count matrix of each group in a Bayesian non-parametric framework. More specifically, `BNP-Seq` algorithms model the gene counts using the gamma-negative binomial process (`GNBP`), which mixes the NB shape parameter for each gene with the distribution of the weight of an atom of a gamma process, or beta-negative binomial process (`BNBP`), which mixes the NB probability parameter of each gene with the distribution of the weight of an atom of a beta process (Zhou *et al.*, 2016). A limitation of the `BNP-Seq` methods is that they are designed for two-group comparative analysis, which only considers the main treatment factor, and cannot be applied to experiments with more complex design.

Given the prevalence of model uncertainty in genomic studies, a Bayesian approach is often the only course possible (Boluki *et al.*, 2017a, b; Karbalayghareh *et al.*, 2017). In this paper, we propose a fully Bayesian negative binomial regression (BNB-R) method for DE analysis of RNA-seq data from experiments with complex multiple-factor design. Unlike all the existing DE methods based on the NB distribution, our method does not rely on *ad-hoc* approximations of various kinds, such as the fact that many statistical tests are only asymptotically valid (Law *et al.*, 2014). BNB-R quantifies the uncertainty of the estimations, and also allows for the incorporation of prior information. BNB-R directly models the influence from covariates of interest for DE analysis and therefore it does not need the SVA pre-processing step. Moreover, this new approach does not require the *ad-hoc* normalization step either, as the model accounts for the sequencing-depth heterogeneity of different samples automatically, similar to the mechanisms employed in the `BNP-Seq` algorithms.

By exploiting two novel data augmentation techniques (Zhou *et al.*, 2012), closed-form posterior inference of BNB-R model parameters is derived in a Gibbs sampling procedure. Specifically, the dispersion parameter of NB distribution is inferred using the augmentation technique of Zhou and Carin [2015], and regression coefficients are inferred in closed-forms by utilizing the Polya-Gamma (PG) distributed auxiliary variable technique of Polson and Scott [2011], removing the need for non-trivial Metropolis–Hastings correction steps (Chib and Greenberg, 1995). Comprehensive simulation results using synthetic and real-world RNA-seq datasets demonstrate the dominance of the proposed BNB-R over existing state-of-the-art tools.

The remainder of this paper is organized as follows. In Section 2, after presenting notations and a brief review of count regression methods, we introduce the model, inference, and DE procedure of BNB-R for genotype-phenotype association with multi-factor experiment design. In Section 3, we present experimental results on both synthetic and real-world benchmark RNA-seq data and show that the proposed NB regression algorithm outperforms the state of the art. We conclude the paper in Section 4.

2 Materials and methods

2.1 Notations and backgrounds

Throughout this paper, we denote scalars, vectors, and matrices by lower-case, bold lower-case and upper-case letters, respectively. We parameterize a NB random variable as $n \sim \mathrm{NB}(r,p)$, where r is the non-negative dispersion and p is the probability parameter. The probability mass function (pmf) of n is expressed as $f_N(n) = \frac{\Gamma(n+r)}{n!\Gamma(r)} p^n (1-p)^r$, where $\Gamma(\cdot)$ is the gamma function. The NB random variable $n \sim \mathrm{NB}(r,p)$ can be generated from a compound Poisson distribution as

$$n = \sum_{t=1}^{\ell} u_t,\ u_t \sim\ \mathrm{Log}(p),\ \ell \sim \mathrm{Pois}(-r\ln(1-p)),$$

where $u \sim \mathrm{Log}(p)$ corresponds to the logarithmic random variable (Johnson *et al.*, 2005), with the pmf $f_U(u) = -\frac{p^u}{u\ln(1-p)}$, $u = 1,2,\ldots$. As shown in Zhou and Carin [2015], given n and r, the distribution of ℓ is a Chinese Restaurant Table (CRT) distribution, $(\ell \,|\, n,r) \sim \mathrm{CRT}(n,r)$, whose random samples can be generated as $\ell = \sum_{t=1}^{n} u_t,\ u_t \sim \mathrm{Bernoulli}\left(\frac{r}{r+t-1}\right)$.

2.1.1 Count regression

A basic count regression model is Poisson regression (Winkelmann, 2013), which can be written as

$$n_j \sim \text{Pois}(\lambda_j), \quad \lambda_j = \exp\left(\boldsymbol{x}_j^T \boldsymbol{\beta}\right), \tag{1}$$

where $\boldsymbol{x}_j = [1, x_{j1}, \ldots, x_{jV}]^T$ is the covariate vector for sample j and $\boldsymbol{\beta} = [\beta_{k0}, \beta_{k1}, \ldots, \beta_{kV}]^T$ is the regression coefficient vector. Poisson regression makes an assumption of equal-dispersion, i.e. $\mathbb{E}[n_j|\boldsymbol{x}_j] = \text{Var}(n_j|\boldsymbol{x}_j)$, which limits its use in analyzing genomic count data that are often highly over-dispersed due to biological variations. To address this issue, a multiplicative random-effect term ϵ_j is commonly added to Poisson regression, expressed as $\lambda_j = \epsilon_j \exp\left(\boldsymbol{x}_j^T \boldsymbol{\beta}\right)$, to model over-dispersed counts. In particular, imposing a gamma prior on this random-effect term leads to a NB regression model (Hilbe, 2011; Winkelmann, 2013; Zhou *et al.*, 2012), which assumes a quadratic relationship between the variance and mean.

2.2 BNB-R: NB regression DE analysis

One important task of RNA-seq data analysis is to identify the genes that show significant changes in their expression levels under different phenotypes or treatment conditions. In this section, a BNB-R model is developed to discover differentially expressed genes, while taking into account biological variability, sequencing depth heterogeneity and experimental confounding factors simultaneously.

We denote the number of sequencing reads mapped to gene $k \in \{1, \ldots, K\}$ in sequencing sample $j \in \{1, \ldots, J\}$ by n_{kj}, and model this count as a NB random variable $n_{kj} \sim \text{NB}(r_j, p_{kj})$. The dispersion parameter r_j, which only depends on the sample index, can be considered as a parameter reflecting the heterogeneity of counts, due to the variation of the sequencing depths across different samples. This can be justified by the gene count expectation $\mathbb{E}[n_{kj}] = r_j \frac{p_{kj}}{1-p_{kj}}$, which is directly proportional to r_j. To establish the dependence between the gene expression and covariates (e.g. phenotypes, treatments and other potential confounding factors) in different experimental setups, we impose a linear relationship between the logit function of the probability and covariates as $\text{logit}(p_{kj}) = \boldsymbol{x}_j^T \boldsymbol{\beta}_k$, where $\boldsymbol{x}_j = [1, x_{j1}, \ldots, x_{jV}]^T$ is the covariate vector for sample j and $\boldsymbol{\beta}_k = [\beta_{k0}, \beta_{k1}, \ldots, \beta_{kV}]^T$ is the regression coefficient vector for gene k. In our proposed model, the covariate variables can be numerical or categorical. Consequently, the expected gene expression can be expressed as $\mathbb{E}[n_{kj}] = r_j \exp\left(\boldsymbol{x}_j^T \boldsymbol{\beta}_k\right)$, which resembles the familiar form of NB GLM (Gardner *et al.*, 1995). Thereby, the effects of different experimental factors on gene expression are captured through the regression coefficients $\boldsymbol{\beta}_k$. In particular, by utilizing the Bayesian framework, the posterior distributions of different combinations of the regression coefficients can be estimated via a Markov chain Monte Carlo (MCMC; Andrieu *et al.*, 2003) inference procedure to assess how the covariates impact the expression changes.

To complete the hierarchal model, we place a gamma prior on each sequencing scaling parameter r_j and independent zero-mean normal priors on the regression coefficients $\boldsymbol{\beta}_k$. The full model is expressed as:

$$\begin{aligned} &n_{kj} \sim \text{NB}(r_j, p_{kj}), \quad \psi_{kj} := \text{logit}(p_{kj}) = \boldsymbol{x}_j^T \boldsymbol{\beta}_k \\ &\boldsymbol{\beta}_k \sim \prod_{v=0}^{V} \text{N}(0, \alpha_v^{-1}), \quad \alpha_v \sim \text{Gamma}(c_0, 1/d_0) \\ &r_j \sim \text{Gamma}(a_0, 1/h), \quad h \sim \text{Gamma}(b_0, 1/g_0). \end{aligned} \tag{2}$$

In addition to controlling the effects of multiple experimental factors via the regression coefficients $\boldsymbol{\beta}_k$, in BNB-R, the precision parameters of the normal distributions over these coefficients are shared between all genes to borrow signal strengths, a desirable property of the model that makes it robust especially in RNA-seq data analysis with a small sample size. In the following, we present our efficient MCMC inference of model parameters, which takes advantage of two novel data augmentation techniques, leading to closed-form parameter updates.

2.2.1 Parameter inference

We start by the inference of the dispersion parameter r_j, by using the data augmentation technique introduced in Zhou and Carin [2015]. In the first step of MCMC inference, we draw latent counts corresponding to gene expression as

$$(\ell_{kj}|-) \sim \text{CRT}(n_{kj}, r_j). \tag{3}$$

It can be shown that the ℓ_{kj} can be considered as the Poisson random count, expressed as $\ell_{kj} \sim \text{Pois}(-r_j \ln(1-p_{kj}))$, used in the compound Poisson representation of the NB distribution $n_{jk} \sim \text{NB}(r_j, p_{kj})$. Hence, by taking advantage of the gamma-Poisson conjugacy, in each Gibbs sampling iteration, the parameter r_j can be updated as

$$(r_j|-) \sim \text{Gamma}\left(\sum_k \ell_{kj} + a_0, \frac{1}{h - \sum_k \ln(1-p_{kj})}\right). \tag{4}$$

The second challenge is the inference of the regression coefficients, for which the lack of conditional conjugacy precludes immediate closed-form inference. Resorting to the methods such as Metropolis–Hastings (Chib and Greenberg, 1995), however, requires a careful choice of the proposal distributions to avoid suffering from high rejection rates and subsequently slow convergence. To address these issues, we adopt an augmentation technique to infer the regression coefficients $\boldsymbol{\beta}_k$, relying on the PG data augmentation of Polson and Scott [2011]. Denote ω_{kj} as a random variable drawn from the PG distribution as $\omega_{kj} \sim \text{PG}(n_{kj} + r_j, 0)$. We have $\mathbb{E}_{\omega_{kj}}\left[\exp\left(-\omega_{kj}\psi_{kj}^2/2\right)\right] = \cosh^{(n_{kj}+r_j)}\left(\psi_{kj}^2/2\right)$. Thus the likelihood of ψ_{kj} in Equation (2) can be expressed as

$$\begin{aligned} \mathcal{L}\left(\psi_{kj}\right) &\propto \frac{\left(e^{\psi_{kj}}\right)^{n_{kj}}}{\left(1+e^{\psi_{kj}}\right)^{n_{kj}+r_j}} \\ &\propto \exp\left(\frac{n_{kj}-r_j}{2}\psi_{kj}\right)\mathbb{E}_{\omega_{kj}}\left[\exp\left(-\omega_{kj}\psi_{kj}^2/2\right)\right]. \end{aligned} \tag{5}$$

Exploiting the exponential tilting of the PG distribution in Polson and Scott [2011], we draw ω_{kj} as

$$(\omega_{kj}|-) \sim \text{PG}\left(n_{kj} + r_j, \psi_{kj}\right). \tag{6}$$

Given the values of the auxiliary variables ω_{kj} for $j = 1, \ldots, J$ and the prior in Equation (2), the conditional posterior of $\boldsymbol{\beta}_k$ can be expressed as

$$p(\boldsymbol{\beta}_k|-) \propto \text{N}(0, A^{-1}) \prod_{j=1}^{J} e^{-\frac{\omega_{kj}}{2}\left(\psi_{kj} - \frac{n_{kj}-r_j}{2\omega_{kj}}\right)^2}, \tag{7}$$

where $A = \text{diag}(\alpha_1, \ldots, \alpha_P)$. Thus in each Gibbs sampling iteration, we update the gene-wise regression coefficients $\boldsymbol{\beta}_k$ as

$$(\boldsymbol{\beta}_k|-) \sim \text{N}(\boldsymbol{\mu}_k, \Sigma_k), \tag{8}$$

where the covariance and mean of this multi-variate normal distribution are defined as $\Sigma_k = \left(\sum_{j=1}^{J} \omega_{kj}\boldsymbol{x}_j\boldsymbol{x}_j^T + A\right)^{-1}$ and $\boldsymbol{\mu}_k = \Sigma_k\left(\sum_{j=1}^{J}\left(\frac{n_{kj}-r_j}{2}\right)\boldsymbol{x}_j\right)$, respectively.

Using the gamma-gamma conjugacy with respect to the gamma scale parameter, we have

$$\begin{aligned}(\alpha_v|-) &\sim \text{Gamma}\left(K/2+c_0, \frac{1}{d_0+\sum_k \beta_{kv}^2/2}\right),\ v=0,\ldots,V.\\ (h|-) &\sim \text{Gamma}\left(b_0+Ja_0, \frac{1}{g_0+\sum_j r_j}\right).\end{aligned} \tag{9}$$

The Gibbs sampling steps in Equations (3)–(9) are summarized in Supplementary Algorithm 1.

2.2.2 DE analysis

To detect differentially expressed genes using the inferred NB regression model, we notice in the prior that

$$\mathbb{E}[n_{kj}] = r_j \exp\left(\boldsymbol{x}_j^T \boldsymbol{\beta}_k\right) \tag{10}$$

and in the conditional posterior shown in Equation (4)

$$\mathbb{E}[r_j|-] = \frac{\sum_k \ell_{kj} + a_0}{h + \sum_k \ln\left(1+\exp\left(\boldsymbol{x}_j^T \boldsymbol{\beta}_k\right)\right)}. \tag{11}$$

Thus one may consider that the NB sample-specific dispersion parameter r_j, which depends on all the gene counts of sample j through latent counts ℓ_{kj}, accounts for the sequencing depth of sample j, and the quantity $\exp\left(\boldsymbol{x}_j^T \boldsymbol{\beta}_k\right)$ represents the expression of gene k in sample j after removing the sequencing-depth effect. To assess whether a certain experimental factor v causes significant expression differences across samples for gene k, we collect posterior MCMC samples for regression coefficients $\boldsymbol{\beta}_k$ and use these MCMC samples to measure the distance between the posterior distributions of $\exp(\beta_{k0})$ and $\exp(\beta_{k0}+\beta_{kv})$. More precisely, we use the symmetric Kullback–Leibler (KL) divergence defined between two discrete distributions P and Q as

$$\text{KL}(P,Q) = \sum\nolimits_x [p(x)-q(x)]\log[p(x)/q(x)].$$

To calculate this distance, we follow the same steps as in Dadaneh *et al.* [2017], and construct a discrete probability vector for each group of collected MCMC samples, referred to as $\boldsymbol{\pi}^{(1)}$ and $\boldsymbol{\pi}^{(2)}$ for the first and second groups under comparison, respectively. Finally, with a small constant set as $\epsilon = 10^{-10}$, we calculate the symmetric KL-divergence as

$$\text{KL}\left(\boldsymbol{\pi}^{(1)}, \boldsymbol{\pi}^{(2)}\right) = \sum_{i=1}^{N} \left(\pi_i^{(1)} - \pi_i^{(2)}\right) \log\left(\frac{\pi_i^{(1)}+\epsilon}{\pi_i^{(2)}+\epsilon}\right). \tag{12}$$

3 Results

To evaluate our BNB-R DE analysis algorithm, referred to as BNB-R, we compare its performance on both synthetic and real-world benchmark data with those of `edgeR` (Robinson *et al.*, 2010), `DESeq2` (Love *et al.*, 2014) and `voom` included in the package `limma` (Law *et al.*, 2014), three widely used methods capable of handling biomedical studies with complex experimental design. As it is common in practice, before applying these methods to real-world RNA-seq data, we first perform a SVA to introduce SVs as additional covariates to model potential unwanted batch effects (Leek, 2014) and then use them to adjust for these artifacts for unbiased DE analysis. We first consider synthetic RNA-seq data in simulated experiments with multiple factors, and we demonstrate that the proposed BNB-R consistently outperforms the other approaches. We then consider the real-world benchmark RNA-seq data extracted from the SEquencing Quality Control (SEQC) project (SEQC/MAQC-III Consortium, 2014). While this dataset does not possess explicit confounding factors, the results support the outstanding performance of BNB-R for DE analysis method in general. On both synthetic and real-world RNA-seq count data, different methods are compared in terms of both the receiver operating characteristic (ROC) and precision-recall (PR) curves and the area under these curves (AUC). Finally, we test BNB-R on a RNA-seq dataset of Th17 cell differentiation to study how incorporating the temporal information can lead to more meaningful biological discoveries.

3.1 Synthetic data

3.1.1 Incorporating covariates improves DE detection

We generate synthetic RNA-seq data with the NB regression generative model. To make the synthetic data closely resemble real-world RNA-seq data, the parameters of the NB regression model are first inferred from the SEQC dataset and then synthetic sequencing counts are generated using these inferred model parameters. Throughout the simulations, we consider three experimental factors as *condition*, *gender* and *dosage*, where condition and gender are categorical covariates with labels {*treated*, *untreated*} and {*male*, *female*}, respectively, and dosage is a numeric covariate in the interval $[0,1]$, generated uniformly at random for each sample.

In the first simulation setting, the expression of gene k in sample j is simulated from $\text{NB}\left(r_j, \frac{1}{1+\exp\left(-\boldsymbol{x}_j^T \boldsymbol{\beta}_k\right)}\right)$, where for sample $j \in \{1,2,\ldots,J\}$, the covariate vector is $\boldsymbol{x}_j = [x_{j0}, x_{j1}, x_{j2}, x_{j3}]$. The variable x_{jv} represents the value of covariate v for sample j. In the first simulation setup, $v=0$ corresponds to the intercept term, and $v=1,2,3$ correspond to *condition*, *gender* and *dosage* covariates respectively. We use a binary scheme for coding the categorical covariates x_{j1} and x_{j2}. More precisely, $x_{j1}=0$ if no treatment has been applied to sample j, and $x_{j1}=1$ if this sample is under treatment. Also, $x_{j2}=0$ if sample j belongs to a female individual and $x_{j2}=1$ if it belongs to a male.

The effect of covariate v on the expression level of gene k is adjusted through the regression coefficient β_{kv}. We simulate this coefficient according to a zero-mean normal distribution with precision parameter α_v. For the *condition* covariate, we draw the precision parameter as $\alpha_1 \sim \text{Gamma}(1.7\text{e}5, 1/1\text{e}4)$. Under this setting, the absolute value of β_{k1} is larger than 0.4 with probability 10%. Thus on average, 10% of the genes exhibit an expression fold-change of at least $\exp(\beta_{k1}) = 1.5$ between two different conditions. In subsequent ROC and PR analyses, we consider gene k as true differentially expressed if $|\beta_{k1}| \geq 0.4$ and not differentially expressed otherwise. The other three precision parameters are simulated as follows:

$$\begin{aligned}\alpha_0 &\sim \text{Gamma}(2.7\text{e}6, 1/1\text{e}4)\\ \alpha_2 &\sim \text{Gamma}(3\text{e}3, 1/1\text{e}4)\\ \alpha_3 &\sim \text{Gamma}(3\text{e}5, 1/1\text{e}4),\end{aligned} \tag{13}$$

where α_0 determines the baseline gene expression independent of experimental factors, and α_2 and α_3 adjust the heterogeneity of gene expressions due to the *gender* and *dosage* factors, respectively. Finally, to simulate the effect of different sequencing depths for different samples, the dispersion parameters r_j are independently drawn from $\text{Gamma}(50, 1/5)$, which is close to the posterior distribution of r_j inferred from the Beijing Genomics Institute (BGI) dataset of the SEQC benchmark.

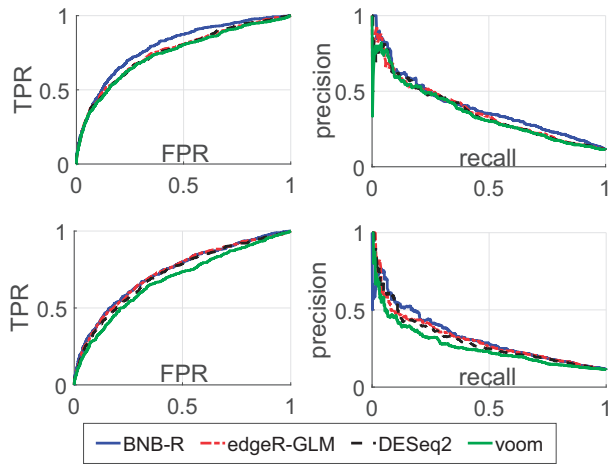

Fig. 1. Left panel: ROC curve, Right panel: PR curve. Performance comparison of different methods in detecting differentially expressed genes generated under a NB regression model with covariates: *condition*, *gender* and *dosage*. Panels in the top row correspond to the case that full covariate information is used in DE analysis. Panels in the bottom row correspond to the case that only condition covariate is used in DE analysis

Table 1. AUC of ROC and PR curves presented in the panels, in the top row of Figure 1

Method	AUC-ROC	AUC-PR
BNB-R	0.7952	0.3922
edgeR-GLM	0.7563	0.3622
DESeq2	0.7533	0.3587
voom	0.7450	0.3499

In the first simulation setting, the gene-expression counts for a total of $K = 5000$ genes and $J = 12$ samples, with three males and three females in each of the two conditions, are generated. We evaluate the performance of BNB-R based on this synthetic data, and compare it to `edgeR`, `DESeq2` and `voom`. For BNB-R, model parameters are inferred via Gibbs sampling, where in each run of the algorithm, we collect 1000 MCMC samples after 1000 burn-in iterations and then rank the genes using the symmetric KL-divergence measure developed in Section 2.2.2. For `edgeR`, `DESeq2` and `voom`, we follow the standard analysis pipelines and rank the genes using the computed *P*-values.

Panels in the top row of Figure 1 illustrate the ROC and PR curves of BNB-R, `edgeR`, `DESeq2` and `voom` under the first simulation setting, when all covariates are employed. The AUCs of these curves are presented in Table 1. The panels in the bottom row of Figure 1 represent the performance of BNB-R, `edgeR`, `DESeq2` and `voom` on the synthetic data when using the *condition* covariate as the single experimental factor, while neglecting all the other covariates. Table 2 provides the AUCs of the curves in the latter scenario. Methods that exploit covariates' information clearly outperform the ones that only rely on the *condition* factor to identify differentially expressed genes, in terms of both the ROC and PR curves. This observation demonstrates the benefit of incorporating available experimental design information to better capture the heterogeneity of gene expression counts. In particular, BNB-R with covariates has the best performance with a significant margin over all the other algorithms. This may be explained by the hierarchical structure of BNB-R, where borrowing information from all genes to estimate precision parameters makes it robust in modeling overdispersed

Table 2. AUC of ROC and PR curves presented in the panels, in the bottom row of Figure 1

Method	AUC-ROC	AUC-PR
BNB-R	0.7343	0.3188
edgeR-GLM	0.7302	0.3087
DESeq2	0.7193	0.2999
voom	0.6832	0.2617

count data. In addition, we have also applied the `BNBP` and `GNBP` methods (Dadaneh *et al.*, 2017), which use only the *condition* factor to determine DE, to the synthetic data in this simulation (results not included in Fig. 1 to not overwhelm it but it can be found in the Supplementary Material). These two methods also perform closely to the algorithms exploiting only the *condition* factor, confirming the observation that integrating additional covariates into a DE model can achieve more accurate and robust DE analysis for genotype-phenotype association.

3.1.2 Sensitivity to experimental design

To assess the sensitivity of BNB-R to the experimental design assumption employed in the DE analysis model, we consider a simulation setting with a more complex combination of experimental factors, including an interaction term between the *gender* and *condition* covariates. Similar to the previous simulation, the expression of gene k in sample j is drawn from $\mathrm{NB}\left(r_j, \frac{1}{1+\exp\left(-\boldsymbol{x}_j^T\boldsymbol{\beta}_k\right)}\right)$, where for sample $j = 1, 2, \ldots, J$, the covariate vector is $\boldsymbol{x}_j = \left[x_{j0}, x_{j1}, \ldots, x_{j4}\right]^T$. In this simulation setup, the elements x_{jv} in the covariate vector for $v = 0, 1, .., 4$ correspond to intercept, *gender*, *condition*, *dosage* and the interaction between *gender* and *condition*, respectively. We employ the same binary coding scheme for the categorical covariates as those used in the previous simulation setting. Thus, for example, $x_{j4} = 1$ if sample j has been under treatment and belongs to a male individual, and $x_{j4} = 0$ otherwise. We also generate the *dosage* covariates x_{j3} from a uniform distribution in interval $[0, 1]$.

The presence of the interaction term in the regression model leads to the dependence of gene DE on both the *condition* and *gender* covariates. More precisely, in this simulation setting, the expected expression fold-change of gene k across two treatment conditions, for a female is $\exp(\beta_{k2})$ and for a male is $\exp(\beta_{k2} + \beta_{k4})$. Hence in ROC and PR analyses, gene k with $|\beta_{k2}| > 0.4$ is considered as truly differentially expressed across conditions for females and when $|\beta_{k2} + \beta_{k4}| > 0.4$, it is considered as truly differentially expressed across conditions for males. We simulate the regression coefficient β_{kv} according to a zero-mean normal distribution with the precision parameter α_v, and we place the following Gamma distributions on the precision parameters:

$$\begin{aligned}
\alpha_0 &\sim \mathrm{Gamma}(2.7e6, 1/1e4)\\
\alpha_1 &\sim \mathrm{Gamma}(1e6, 1/1e4)\\
\alpha_2 &\sim \mathrm{Gamma}(1.8e5, 1/1e4)\\
\alpha_3 &\sim \mathrm{Gamma}(3e5, 1/1e4)\\
\alpha_4 &\sim \mathrm{Gamma}(1.2e6, 1/1e4).
\end{aligned} \tag{14}$$

RNA-seq counts for a total of $K = 5000$ genes and $J = 12$ samples, with three males and three females in each treatment condition, are generated. In this synthetic dataset, 516 genes are differentially expressed across treatment conditions for females and 653 genes are differentially expressed for males. First, we evaluate the

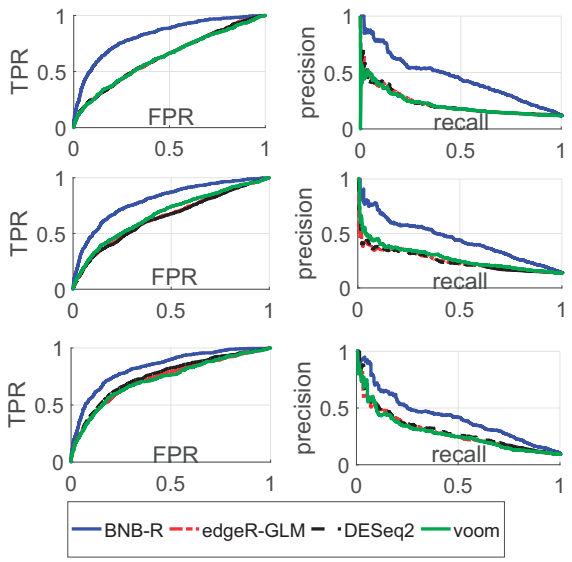

Fig. 2. Left panels: ROC curve, Right panels: PR curve. Performance comparison of different methods in detecting differentially expressed genes generated under the NB regression model with covariates: *condition, gender, dosage* and interaction of *condition* and *gender*. The panels in the top and middle rows correspond to differentially expressed genes across conditions for males and females, respectively. The panels in the bottom row correspond to differentially expressed genes for the case that full covariate information is not employed, with the interaction term excluded from DE analyses by all the methods

performance of BNB-R, `edgeR`, `DESeq2` and `voom` on this synthetic data, assuming that the true design matrix used for data generation is provided for all algorithms. Differentially expressed genes are identified using the same protocol as described in the previous subsection. The top and middle panels of Figure 2 illustrate the ROC and PR curves for the detection of differentially expressed genes across conditions in males and females, respectively. BNB-R clearly outperforms the other methods in terms of both ROC and PR, for gender-specific DE analyses.

Next, instead of assuming knowing the true underlying data generation mechanism, we exclude the interaction term used for data generation for DE analysis with different methods and use the covariate vector $\boldsymbol{x}_j = [x_{j0}, x_{j1}, x_{j2}, x_{j3}]$ for sample j, where the elements x_{jv} for $v = 0, 1, 2, 3$ represent the same covariates as in the data generation procedure. As a consequence of using this design, detected differentially expressed genes are not specific to a gender. Hence to evaluate the performance of BNB-R, `edgeR`, `DESeq2` and `voom` when using this design matrix, we need to compare the detected genes to those that are truly differentially expressed across conditions independent of gender. In this simulation, there are 400 genes that are differentially expressed across the treatment conditions for both male and female groups. We consider these genes as truly differentially expressed independent of gender, and the rest of the genes as not differentially expressed. The ROC and PR curves plotted based on this setting are shown in the bottom row of Figure 2. In this case, BNB-R again exhibits the best performance in terms of the ROC and PR curves, confirming its superior performance even if the true mechanism of data generation is not fully known.

3.2 SEQC benchmark

In this section, we evaluate the performance of the proposed BNB-R method using SEQC benchmark (SEQC/MAQC-III Consortium, 2014). Specifically, we use the RNA-seq data from BGI provided in the R package SEQC on Bioconductor (Gentleman *et al.*, 2004), containing the counts for about 26 000 genes. In our experiments, we employ sample groups A and B, which are derived from the Agilent's Universal Human Reference RNA and Life Technologies' Human Brain Reference RNA cell lines, respectively. We collect the counts from the first flow cells of the sequencing machines on five replicates for each group.

To evaluate the DE analysis methods, we note that in the SEQC project, the same RNA samples for a comprehensive group of control genes are analyzed based on quantitative Reverse Transcription Polymerase Chain Reaction (qRT-PCR) using TaqMan assays (Joyce, 2002), which is referred as the TaqMan benchmark data (Maqc Consortium and Others, 2006; SEQC/MAQC-III Consortium, 2014). More precisely, for sample groups A and B, the expression intensity values of 955 selected control genes have been derived in the TaqMan qPT-PCR analysis for sequencing benchmarking. In the absence of the knowledge on the genes that are truly differentially expressed across different conditions, we follow the approach in Rapaport *et al.* [2013] to threshold the qRT-PCR expression ratios across different conditions at a certain value to define the ground-truth set of differentially expressed genes. Based on these 955 genes in the TaqMan data, we evaluate the performance of different DE analysis pipelines.

Before applying `edgeR`, `DESeq2` and `voom` to this dataset, we first perform a SVA to adjust for un-modeled artifacts such as batch effects (Leek, 2014). More precisely, we use svaseq function of R package sva (Leek *et al.*, 2012) with two introduced SVs. In the downstream DE analysis, we use these two SVs as extra confounding factors for `edgeR`, `DESeq2` and `voom`. Our experiment shows that incorporation of the SVs slightly improves the performance of these methods (Supplementary Fig. S5). Note that although for BNB-R no explicit experimental factor other than a sample's group is used in this experiment, our results suggest the performance of the proposed BNB-R DE analysis method is superior to those of stochastic processes inspired models in `BNP-Seq`, all of which achieve better ROC and PR curves than `edgeR`, `DESeq2` and `voom` in conjunction with SVA, as described in detail below.

While truly differentially expressed genes are unknown for the SEQC RNA-seq data, we rely on the qRT-PCR expression intensity of the 955 genes in the TaqMan data and set different cut-offs for the binary logarithm (log2) of the qRT-PCR expression ratio to define 'truly' differentially expressed genes. We increase this log2 cut-off value gradually from 0.5 to 2, and calculate both AUC-ROC and AUC-PR. For the analysis of the dataset BGI on a single cluster node with Intel Xeon 2.5 GHz E5-2670 v2 processor, it took around 2 h for BNB-R method with 2000 MCMC iterations. The posterior distributions of the regression coefficients are used to assess DE. In addition to the methods used for synthetic data, we also include `BNBP` and `GNBP` (Dadaneh *et al.*, 2017), both of which are generative models designed specifically for a single factor setting. As shown in the bottom panels of Figure 3, the BNB-R method outperforms all the other methods in both ROC and PR analyses, followed very closely by `BNBP` and `GNBP`. Note that the performance gains of the three generative models over the other methods become more significant as one increase the log2 cut-off for the qRT-PCR expression ratio, which reduces the number of genes that are considered as truly differentially expressed.

To further investigate the experimental results, we fix the log2 cut-off value at 2 for the qRT-PCR expression intensity of the 955 genes in the TaqMan data, and illustrate the ROC and PR curves for the BGI dataset in the top panels of Figure 3. It is clear the BNB-R method along with `GNBP` and `BNBP` not only have higher AUC-ROC

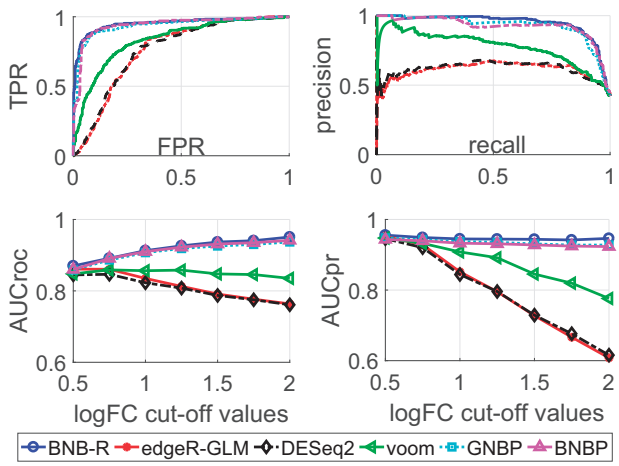

Fig. 3. Top row: ROC and PR curves for a fixed cut-off, Bottom row: AUC of ROC and PR curves for different cut-off values. Performance comparison of different methods in detecting differentially expressed genes on real-world benchmark RNA-seq data from the SEQC project. edgeR, DESeq2 and voom are applied in conjunction with SVA with two SVs

and AUC-PR, but also outperform `edgeR`, `DESeq2`, and `voom` used together with SVA in almost all regions of the ROC and PR curves.

3.3 Case study: Th17 cell differentiation

To further illustrate its potential biological significance when integrating other covariates in BNB-R for biomarker identification applications, we provide a case study with our BNB-R method on a RNA-seq dataset of early human T helper 17 (Th17) cell differentiations and T-cell activation (Th0). Th17 cells play an essential role in the pathogenesis of auto-immune and inflammatory diseases, and have been the focus of many recent research efforts (Tuomela *et al.*, 2012). In particular, the knowledge of the early phase of Th17 differentiation helps to gain insight into the process of signal propagation through various pathways and gene regulatory networks (Äijö *et al.*, 2014). We use the RNA-seq dataset of Tuomela *et al.* [2016] and Chan *et al.* [2016], which contains gene expression profiling of Th0 and Th17 cells at the following five time points: 0 h, 12 h, 24 h, 48 h and 72 h after cell activation and stimulation, with three biological replicates at each time point. The data is obtained from *Gene Expression Omnibus*, with accession *GSE52260*.

The design matrix of the analysis is formed from an additive model formula as in our simulation studies, accounting for condition and time point factors. More precisely, for sample $j = 1, 2, \ldots, 15$ the covariate vector is $\boldsymbol{x}_j = [x_{j0}, x_{j1}, x_{j2}]^T$, where x_{j0} is the intercept, x_{j1} is the cell category (i.e. Th0 versus Th17) and x_{j2} is the sample time point. We apply BNB-R to identify differentially expressed genes, where after 1000 burn-in iterations, 1000 posterior samples are collected to calculate the symmetric KL-divergence between the posterior distributions of $\exp(\beta_{k0})$ and $\exp(\beta_{k0} + \beta_{k1})$ to rank the genes. The run-time of BNB-R with 2000 MCMC sampling iterations for the Th17 dataset on the cluster node with configuration provided in Section 3.2 is around 6 h.

We consider the top 100 genes ranked by the symmetric KL-divergence and perform *Gene Ontology* (GO) analysis using LAGO (Boyle *et al.*, 2004) software (available at http://go.princeton.edu/cgi-bin/LAGO), focusing on the ontology of biological processes. The top five significantly enriched GO terms discovered by LAGO, with their corresponding adjusted *P*-values shown in Table 3, illustrating the association between the differentially expressed genes and immune system activation and response to stimulus.

Table 3. Top five enriched GO terms associated with top 100 differentially expressed genes in TH17 dataset detected by BNB-R

GO-ID	Term	*P*-value
GO: 0002376	Immune system process	4.74695e–13
GO: 0046649	lymphocyte activation	3.33415e–11
GO: 0006955	Immune response	3.90728e–11
GO: 0045321	Leukocyte activation	1.6007e–10
GO: 0050896	Response to stimulus	1.89798e–10

In a closer look at the results, the top differentially expressed gene identified by BNB-R is gene COL6A3, an important organizer of the extracellular matrix proteins, contributing to adipose tissue inflammation (Pasarica *et al.*, 2009). Also, the up-regulation of COL6A3 gene in Th17-polarizing cells is confirmed by microarray and RT-PCR assays in Tuomela *et al.* [2012]. The third ranked gene, Leukemia Inhibitory Factor (LIF), belongs to the IL-6 family of cytokines and resides within the core regulatory circuitry of T cells (Metcalfe, 2011). The fourth gene, RORC, is a Th17 lineage-specific transcription factor (Diveu *et al.*, 2009), whose DE is also verified in the microarray study in Tuomela *et al.* [2012]. In addition, Western blotting results in Tuomela *et al.* [2012] show that genes BATF, CTSL1, VDR, KDSR, ATP1B1 and BASP1 were highly expressed in Th17 cells compared with their expression in Th0 cells at various time points during the first 3 days of polarization. The rankings of these genes obtained by our BNB-R are 11, 13, 15, 20, 24 and 41, respectively, which confirms the significance of their expression changes. Moreover, microarray studies of Tuomela *et al.* [2012] found out the up-regulation of CXCR5 and LMNA in CD4+ T cells cultured under Th17-polarizing conditions compared with Th0 cells, and flow cytometric detection of CD52 at 48 h and 72 h showed down-regulation of this protein in CD4+ T cells cultured under Th17-polarizing conditions. These genes are ranked 14, 44 and 60, respectively, in our DE analysis, supporting their potential roles in Th17 cells differentiation process.

Next, to examine how incorporating the time course information changes the DE analysis results, we apply BNB-R on the Th17 dataset, considering only the condition factor but ignoring the temporal information of different samples. Although out of the top 100 differentially expressed genes, there are 84 genes common between these two differential analysis results, the GO analysis, when the time factor is neglected, results in a total of 36 significantly enriched terms with known annotations, which is less than 40 annotated enriched terms when including the time factor. Some of the GO terms missed include *cytokine-mediated signaling pathway*, *positive regulation of JAK-STAT*, *STAT cascades* and *T cell activation involved in immune response*, which are all related to the immune system and can potentially lead to new hypotheses. In addition, BNB-R considering the time factor leads to smaller *P*-values overall in comparison to the analysis without time information, and hence more significantly enriched GO terms. For instance, the adjusted *P*-value obtained for *T cell differentiation* by the former analysis is 1.910e − 4, while the latter returns 2.554e − 3.

4 Conclusions

We propose a BNB-R method for DE analysis of sequencing count data. On one hand, BNB-R is capable of handling complex experiments involving multiple factors. On the other hand, it does not require an *ad-hoc* normalization pre-processing step. By taking advantage of novel data augmentation techniques, BNB-R possesses

efficient closed-form Gibbs sampling update equations and ranks differentially expressed genes based on a symmetric KL-divergence measure, exploiting the full posterior distributions of the model parameters. Experimental results on both synthetic and real-world RNA-seq data demonstrate the state-of-the-art performance of BNB-R in DE analysis of RNA-seq data.

Acknowledgements

This project was partially supported by the USDA-SCRI competitive grant 2017-51181-26834. XQ was partially supported by Award CCF-1553281 from the National Science Foundation and the USDA NIFA Award 06-505570-01006. The authors thank Texas A&M High Performance Research Computing and Texas Advanced Computing Center for providing computational resources to perform experiments in this paper.

Funding

This work was supported by Award CCF-1553281 from the National Science Foundation and the USDA NIFA Award 06-505570-01006.

Conflict of Interest: none declared.

References

Äijö,T. *et al*. (2014) Methods for time series analysis of RNA-seq data with application to human Th17 cell differentiation. *Bioinformatics*, **30**, i113–i120.

Anders,S. and Huber,W. (2010) Differential expression analysis for sequence count data. *Genome Biol*., **11**, R106.

Andrieu,C. *et al*. (2003) An introduction to mcmc for machine learning. *Mach. Learn*., **50**, 5–43.

Boluki,S. *et al*. (2017a) Constructing pathway-based priors within a gaussian mixture model for bayesian regression and classification. In: *IEEE/ACM Transactions on Computational Biology and Bioinformatics*, IEEE.

Boluki,S. *et al*. (2017b) Incorporating biological prior knowledge for Bayesian learning via maximal knowledge-driven information priors. *BMC Bioinformatics*, **18**, 552.

Boyle,E.I. *et al*. (2004) GO:: termFinder—open source software for accessing Gene Ontology information and finding significantly enriched Gene Ontology terms associated with a list of genes. *Bioinformatics*, **20**, 3710–3715.

Chan,Y.H. *et al*. (2016) A subpopulation model to analyze heterogeneous cell differentiation dynamics. *Bioinformatics*, **32**, 3306–3313.

Chib,S. and Greenberg,E. (1995) Understanding the Metropolis-Hastings algorithm. *Am. Stat*., **49**, 327–335.

Dadaneh,S.Z. *et al*. (2017) `BNP-Seq`: Bayesian nonparametric differential expression analysis of sequencing count data. *J. Am. Stat. Assoc*., doi: 10.1080/01621459.2017.1328358.

Datta,S. and Nettleton,D. (2014) *Statistical Analysis of Next Generation Sequencing Data*. Springer, New York, USA.

Diveu,C. *et al*. (2009) IL-27 blocks RORc expression to inhibit lineage commitment of Th17 cells. *J. Immunol*., **182**, 5748–5756.

Gardner,W. *et al*. (1995) Regression analyses of counts and rates: poisson, overdispersed Poisson, and negative binomial models. *Psychol. Bull*., **118**, 392.

Gentleman,R.C. *et al*. (2004) Bioconductor: open software development for computational biology and bioinformatics. *Genome Biol*., **5**, R80.

Hilbe,J.M. (2011) *Negative Binomial Regression*. Cambridge University Press, Cambridge, United Kingdom.

Johnson,N.L. *et al*. (2005) *Univariate Discrete Distributions, Volume 444*. Wiley, New Jersey, USA.

Joyce,C. (2002) Quantitative RT-PCR: a review of current methodologies, RT-PCR Protocols. *Methods Mol. Biol*., doi: 10.1385/1-59259-283-X:083.

Karbalayghareh,A. *et al*. (2017) Classification of Gaussian trajectories with missing data in Boolean gene regulatory networks. In: *2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, IEEE, New Orleans, LA, USA, pp. 1078–1082.

Law,C.W. *et al*. (2014) Voom: precision weights unlock linear model analysis tools for RNA-seq read counts. *Genome Biol*., **15**, R29.

Leek,J.T. (2014) Svaseq: removing batch effects and other unwanted noise from sequencing data. *Nucleic Acids Res*., **42**, e161–e161.

Leek,J.T. *et al*. (2012) The sva package for removing batch effects and other unwanted variation in high-throughput experiments. *Bioinformatics*, **28**, 882–883.

Li,J. and Tibshirani,R. (2013) Finding consistent patterns: a nonparametric approach for identifying differential expression in RNA-Seq data. *Stat. Methods Med. Res*., **22**, 519–536.

Love,M.I. *et al*. (2014) Moderated estimation of fold change and dispersion for RNA-seq data with `DESeq2`. *Genome Biol*., **15**, 550.

Maqc Consortium and Others. (2006) The MicroArray Quality Control (MAQC) project shows inter-and intraplatform reproducibility of gene expression measurements. *Nat. Biotechnol*., **24**, 1151.

Metcalfe,S. (2011) LIF in the regulation of T-cell fate and as a potential therapeutic. *Genes Immun*., **12**, 157.

Pasarica,M. *et al*. (2009) Adipose tissue collagen VI in obesity. *J. Clin. Endocrinol. Metab*., **94**, 5155–5162.

Polson,N.G. and Scott,J.G. (2011) Default Bayesian analysis for multi-way tables: a data-augmentation approach. *arXiv preprint arXiv: 1109.4180*.

Rapaport,F. *et al*. (2013) Comprehensive evaluation of differential gene expression analysis methods for RNA-seq data. *Genome Biol*., **14**, R95.

Robinson,M.D. *et al*. (2010) `edgeR`: a Bioconductor package for differential expression analysis of digital gene expression data. *Bioinformatics*, **26**, 139–140.

SEQC/MAQC-III Consortium. (2014) A comprehensive assessment of RNA-seq accuracy, reproducibility and information content by the Sequencing Quality Control Consortium. *Nat. Biotechnol*., **32**, 903–914.

Smyth,G.K. (2004) Linear models and empirical bayes methods for assessing differential expression in microarray experiments. *Stat. Appl. Genet. Mol. Biol*., **3**, 1–25.

Smyth,G.K. (2005) Limma: linear models for microarray data. In: *Bioinformatics and Computational Biology Solutions Using R and Bioconductor*. Springer, New York, USA, pp. 397–420.

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

Tuomela,S. *et al*. (2012) Identification of early gene expression changes during human Th17 cell differentiation. *Blood*, **119**, e151–e160.

Tuomela,S. *et al*. (2016) Comparative analysis of human and mouse transcriptomes of Th17 cell priming. *Oncotarget*, **7**, 13416.

Wang,Z. *et al*. (2009) RNA-Seq: a revolutionary tool for transcriptomics. *Nat. Rev. Genet*., **10**, 57–63.

Winkelmann,R. (2013) *Econometric Analysis of Count Data*. Springer Science and Business Media, Berlin, Germany.

Zhou,M. and Carin,L. (2015) Negative binomial process count and mixture modeling. *IEEE Trans. Pattern Anal. Mach. Intel*., **37**, 307–320.

Zhou,M. *et al*. (2012) Lognormal and gamma mixed negative binomial regression. In: *ICML 2012*, NIH Public Access, Edinburgh, Scotland.

Zhou,M. *et al*. (2016) Priors for random count matrices derived from a family of negative binomial processes. *J. Am. Stat. Assoc*., **111**, 1144–1156.

Zyprych-Walczak,J. *et al*. (2015) The impact of normalization methods on rna-seq data analysis. *BioMed Res. Int*., **2015**, 1.