

**BALLI: Bartlett-Adjusted Likelihood-based Linear Model Approach for Identifying  
Differentially Expressed Gene with RNA-seq Data**

**Kyungtaek Park<sup>1</sup>, Jaehoon An<sup>2</sup>, Jungsoo Gim<sup>3</sup>, Sungho Won<sup>1, 2, 3, \*</sup>**

<sup>1</sup>Interdisciplinary Program of Bioinformatics, Seoul National University, Seoul, 08826,  
Korea

<sup>2</sup>Department of Public Health Science, Seoul national university, Seoul, 08826, Korea

<sup>3</sup>Institute of Health and Environment, Seoul National University, Seoul, 08826, Korea.

\* To whom correspondence should be addressed.

## ABSTRACT

**Motivation:** Transcriptomic profiles can improve our understanding of the phenotypic molecular basis of biological research, and many statistical methods have been proposed to identify differentially expressed genes under two or more conditions with RNA-seq data. However, statistical analyses with RNA-seq data often suffer from small sample sizes, and global variance estimates of RNA expression levels have been utilized as prior distributions for gene-specific variance estimates, making it difficult to generalize the methods to more complicated settings. We herein proposed a Bartlett-Aadjusted Likelihood based Linear mixed model approach (BALLI) to analyze more complicated RNA-seq data. The proposed method estimates the technical and biological variances with a linear mixed effect model, with and without adjusting small sample bias using Bartlett's corrections.

**Results:** We conducted extensive simulations to compare the performance of BALLI with those of existing approaches (edgeR, DESeq2, and voom). Results from the simulation studies showed that BALLI correctly controlled the type-1 error rates at the various nominal significance levels, and produced better statistical power and precision estimates than those of other competing methods in various scenarios. Furthermore, BALLI was robust to variation of library size. It was also successfully applied to Holstein milk yield data, illustrating its practical value.

**Availability and Implementation:** BALLI is implemented as R package and freely available at <http://healthstat.snu.ac.kr/software/balli/>.

**Contact:** won1@snu.ac.kr

**Supplementary Information:** Supplementary data are available at *Bioinformatics* online

## INTRODUCTION

Transcriptomic profiles can improve our understanding of the phenotypic molecular basis of biological research, and many attempts have been made to identify differentially expressed genes (DEGs) by microarray analysis. However, microarray analysis often suffers from many systematic errors, such as hybridization and dye-based detection bias, hampering the detection of true DEGs (Dobbin, et al., 2005; Okoniewski and Miller, 2006). Recently, high-throughput sequencing technology has markedly improved. RNA sequencing (RNA-seq), also called whole-transcriptome shotgun sequencing, uses next-generation sequencing to quantify the abundance of transcripts with several desirable features, such as increased dynamic range and the freedom from *a priori* chosen probes (Zhao, et al., 2014). Furthermore, RNA-seq is robust against systematic errors and has therefore emerged as a successful alternative to microarray analysis (Mortazavi, et al., 2008).

RNA-seq quantifies the numbers of reads aligned to particular transcripts or genes, and various approaches have been proposed to manage the RNA-seq data (Dillies, et al., 2013). There are two different types of statistical methods: read-count-based approaches and transformation-based approaches. Read-count-based approaches assume that observed read counts follow negative binomial distribution, and generalized linear regression with a logarithm as a link function is utilized. These approaches typically assume that variances include biological and technical variances; the latter indicates variance observed among measurements of the same biological unit, and the former indicates variance between different biological units, such as different subjects or different tissues of the same subject. If technical replicates are analyzed, observed read counts from technical replicates have the same means under the same conditions. Marioni et al (2008) demonstrated that the data follow a Poisson distribution, and variances in technical replicates are expected to be the same as their means for each gene (Marioni, et al., 2008). However, if biological replicates

are available, means and variances of read counts are different among different biological units. Bullard et al (2010) carefully examined such variability and concluded that the biological variances were usually larger than technical variances, supporting the presence of overdispersion (Bullard, et al., 2010). Thus, negative binomial distribution has often been utilized; edgeR (Robinson, et al., 2010) and DESeq2 (Love, et al., 2014) are such methods. Transformation-based approaches assume that the transformed read counts for each gene follow the normal distribution. For example, voom calculated proportions of read counts for each gene per subject, and the log-transformed values were then assumed to follow the normal distribution, assuming that the relative proportion of technical variances becomes smaller if the read count grows larger (Law, et al., 2014).

Negative-binomial distributions for read counts and normal distributions for log-transformation of counts per million (CPM) successfully describe distributions of RNA-seq data. However, RNA-seq is relatively expensive compared with microarray, and thus, further adjustment has been made to handle the problem of small sample size. If sample size is small, the estimated variance can have large standard errors, and thus, multiple methods that incorporate prior knowledge have been proposed. For example, variances of read counts assume to be positively related to their means, and their relationships can be estimated by comparing the means and variances of read counts for all genes. This can often be utilized to shrink variance parameters (Robinson and Smyth, 2007; Tusher, et al., 2001). edgeR and DESeq2 estimate the overall dispersion parameter for all genes and are then combined with gene-wise dispersion parameters for each gene using empirical Bayesian rules. voom assumes that the variances of log-transformed CPM (log-cpm) are functionally related to their means. Locally weighted scatterplot smoothing (LOWESS) curves between the mean and residual variances of genes are then utilized to weight variance estimates of each gene.

Existing methods shrink the variance estimate of each gene toward global variance

estimates or use variance estimates based on the relationships between means and variances. Such assumptions are often very useful if the sample sizes are small. However, there are multiple factors that can distort these relationships, and if they are violated, the performance of existing approaches can be affected. For example, the quality of data is highly dependent on the preparation steps, and unexpected noise, such as noise from different storage periods or sequencing organization of samples, can occur during preparation steps. Moreover, read counts of cancer tissues are more heterogeneous than those of normal tissues, and biological variances can be affected by disease status (McCarthy, et al., 2012). Thus, their effects can distort the relationship between technical and biological variances. Multiple studies have shown that misspecified relationships can lead to substantial biases (Chavance and Escolano, 2016; Litière, et al., 2008). For example, variance estimators for random effects, which are assumed to follow a normal distribution, can be seriously biased unless they are normally distributed (Litière, et al., 2008). General approaches that are not sensitive to those problems are necessary.

In this article, we present new methods for identifying DEGs with RNA-seq data, BALLI and LLI. Statistical analyses with log-transformed read counts are often more powerful than other existing methods and are relatively insensitive to various errors (Seyednasrollah, et al., 2015; Sonesson and Delorenzi, 2013). Thus, we consider the log-cpm as response variables and used linear mixed effect models to estimate technical and biological variance. Furthermore, Bartlett-adjusted likelihood ratio tests were applied to correct the small sample bias (Bartlett, 1937). By allowing model comparisons among different models, our models enable robust analyses for various scenarios. For our simulations studies, artificial RNA-seq data are generated based on real data and negative binomial distributions. Our studies showed that the proposed method performed better than existing methods. The proposed methods were applied to Holstein milk yield data at the false discovery rate (FDR)-adjusted 0.1

significance level and uniquely produced significant results. The proposed methods were implemented as an R package and are freely downloadable at <http://healthstat.snu.ac.kr/software/balli/>.

## MATERIAL AND METHODS

### Notations

We assumed that there were  $M$  different groups, and the averages of the expressed read counts for each gene were compared among these groups. For case-control studies,  $M = 2$ . We assumed that there were  $n_m$  subjects in group  $m$  and denoted the total sample size by  $N$ . Then, we have  $N = n_1 + \dots + n_M$ . We defined dummy variables for subject  $i$  in group  $m$  by

$$x_{mi} = \begin{cases} 1 & \text{if subject } i \text{ is in group } m, \\ 0 & \text{o. w.} \end{cases},$$

$$m = 1, \dots, M - 1, \quad i = 1, 2, \dots, N.$$

A design matrix for group variables is defined by

$$\mathbf{X} = \begin{pmatrix} x_{11} & \cdots & x_{(M-1)1} \\ \vdots & \ddots & \vdots \\ x_{1N} & \cdots & x_{(M-1)N} \end{pmatrix}.$$

We assumed that indexes of all subjects were sorted in an ascending order of groups. Thus, the first  $n_1$  subjects were in group 1, the second  $n_2$  subjects were in group 2, and so on. We assumed that expressed read counts were observed for  $G$  genes and were denoted by  $r_{gi}$  for gene  $g$  of subject  $i$  ( $g = 1, \dots, G$ ). Then, the library size for subject  $i$ ,  $R_i$ , was equivalent to  $R_i = \sum_g r_{gi}$ . If we denoted the normalized  $R_i$  with the trimmed mean of the M-value (Robinson and Oshlack, 2010) by  $R_i^*$ , the log-transformed read counts per million of gene  $g$

for subject  $i$  were defined by:

$$y_{gi} = \log_2 \left( \frac{r_{gi+d}}{R_i^* + 2*d} \times 10^6 \right), \text{ where } d = \frac{R_i^*}{\frac{1}{N} \sum_{i=1}^N R_i^*} \times 0.25 \dots (1)$$

and their vector  $\mathbf{Y}_g$  was defined as

$$\mathbf{Y}_g = (y_{g1}, y_{g2}, \dots, y_{gN})^t.$$

### Technical and biological variances of $y_{gi}$

We assumed that  $r_{gi}$  followed a negative binomial distribution, and its mean and variance were  $\mu_{gi}$  and  $\mu_{gi} + \mu_{gi}^2 \phi_g$ , respectively. If we let the mean and variance of  $\log_2 r_{gi}$  be  $\lambda_{gi}$  and  $s_{gi}^2$ , respectively, then  $\text{var}(y_{gi})$  can be obtained by the first order approximation as follows:

$$\text{var}(y_{gi}) \approx \text{var}(\log_2 r_{gi}) = s_{gi}^2 \approx \frac{1}{\mu_{gi}} + \phi_g \dots (2).$$

Here,  $\mu_{gi}^{-1}$  and  $\phi_g$  indicate the variances attributable to the technical and biological replicates, respectively. By the second order approximation, the technical variance,  $1/\mu_{gi}$ , can be expressed in terms of  $\lambda_{gi}$  and  $s_{gi}^2$  as follows:

$$\mu_{gi} = E(2^{\log_2 r_{gi}}) \approx 2^{\lambda_{gi}} \times \left( 1 + \frac{1}{2} (\log_e 2)^2 s_{gi}^2 \right).$$

$\lambda_{gi}$  and  $s_{gi}^2$  are functionally related, and both were estimated with the method used for voom-trend (Law, et al., 2014) as follows:

- i. For all genes,  $g = 1, \dots, G$ , fit linear regressions,  $y_{gi} = \alpha_g + x_i^t \beta_g + \epsilon_{gi}$ , and calculate  $\hat{y}_{gi}$ . Residual variances are used as  $\hat{s}_g^2$ . If environmental effects affect  $y_{gi}$ ,

then they should be included as covariates.

- ii. Calculate  $\hat{\lambda}_g = \bar{y}_g + \log_2 \tilde{R} - \log_2 10^6$  where  $\bar{y}_g$  is an average of  $y_{gi}$ ,  $\tilde{R}$  is a geometric mean of  $(R_i^* + 1)$  and  $g = 1, \dots, G$ .
- iii. For  $(\hat{\lambda}_g, \hat{s}_g^2)$ ,  $g = 1, \dots, G$ , obtained from (i) and (ii), fit LOWESS curve  $\hat{s}_g^{1/2}$  on  $\hat{\lambda}_g$ .
- iv. Calculate  $\hat{\lambda}_{gi} = \log_2 \hat{r}_{gi} = \hat{y}_{gi} + \log_2(R_i^* + 1) - \log_2 10^6$  and apply LOWESS curve from (iii) to obtain  $\hat{s}_{gi}^{1/2}$ .
- v. Calculate  $\hat{\mu}_{gi}$  by incorporating  $\hat{\lambda}_{gi}$  and  $\hat{s}_{gi}$  to the following equation:

$$\hat{\mu}_{gi} \approx 2^{\hat{\lambda}_{gi}} \times \left(1 + \frac{1}{2} (\log_e 2)^2 \hat{s}_{gi}^2\right) \dots \quad (3).$$

### Linear mixed effect model

We denoted a design matrix for nuisance effects including an intercept by  $\mathbf{Z}$ . We let  $\mathbf{b}_g$  and  $\mathbf{e}_g$  be vectors for random effects and measurement errors, respectively. Denoting a  $w \times w$  dimensional identity matrix by  $\mathbf{I}_w$ , we considered the following linear mixed effects model:

$$\mathbf{Y}_g = \mathbf{Z}\boldsymbol{\alpha}_g + \mathbf{X}\boldsymbol{\beta}_g + \mathbf{b}_g + \mathbf{e}_g, \mathbf{b}_g \sim \text{MVN}(\mathbf{0}, \psi_g \boldsymbol{\Sigma}_{g,\mathbf{b}}), \mathbf{e}_g \sim \text{MVN}(\mathbf{0}, \boldsymbol{\Sigma}_{g,\mathbf{e}}),$$

$$\boldsymbol{\Sigma}_{g,\mathbf{b}} = \begin{pmatrix} \mu_{g1}^{-1} & 0 & \dots & 0 \\ 0 & \mu_{g2}^{-1} & \dots & 0 \\ & & \ddots & \\ 0 & 0 & \dots & \mu_{gN}^{-1} \end{pmatrix}, \boldsymbol{\Sigma}_{g,\mathbf{e}} = \begin{pmatrix} \sigma_{g1}^2 \mathbf{I}_{n_1} & \mathbf{0} & \dots & \mathbf{0} \\ \mathbf{0} & \sigma_{g2}^2 \mathbf{I}_{n_2} & \dots & \mathbf{0} \\ & & \ddots & \\ \mathbf{0} & \mathbf{0} & \dots & \sigma_{gM}^2 \mathbf{I}_{n_M} \end{pmatrix}.$$

Here,  $\psi_g \boldsymbol{\Sigma}_{g,\mathbf{b}}$  and  $\boldsymbol{\sigma}_g^2$  indicate technical and biological variations, respectively, according to equation (2). Notably, elements of  $\boldsymbol{\Sigma}_{g,\mathbf{b}}$  are obtained from equation (3), and  $\psi_g$  and  $\boldsymbol{\sigma}_g^2 = (\sigma_{g1}^2, \dots, \sigma_{gM}^2)$  are should be estimated. Equation (2) shows that  $\psi_g$  becomes 1, and we



assumed that  $\sigma_{g1}^2 = \dots = \sigma_{gM}^2$ . Then, our final model becomes

$$\mathbf{Y}_g = \mathbf{Z}\boldsymbol{\alpha}_g + \mathbf{X}\boldsymbol{\beta}_g + \mathbf{b}_g + \mathbf{e}_g, \mathbf{b}_g \sim \text{MVN}(\mathbf{0}, \boldsymbol{\Sigma}_{g,\mathbf{b}}), \mathbf{e}_g \sim \text{MVN}(\mathbf{0}, \sigma_g^2 \mathbf{I}),$$

$$\boldsymbol{\Sigma}_{g,\mathbf{b}} = \begin{pmatrix} \hat{\mu}_{g1}^{-1} & 0 & \dots & 0 \\ 0 & \hat{\mu}_{g2}^{-1} & \dots & 0 \\ & & \ddots & \\ 0 & 0 & \dots & \hat{\mu}_{gN}^{-1} \end{pmatrix},$$

which is equivalent to

$$\mathbf{Y}_g = \mathbf{Z}\boldsymbol{\alpha}_g + \mathbf{X}\boldsymbol{\beta}_g + \mathbf{e}'_g, \mathbf{e}'_g \sim \text{MVN}(\mathbf{0}, \boldsymbol{\Sigma}_{g,\mathbf{b}} + \sigma_g^2 \mathbf{I}), \quad \sigma_g^2 \geq 0.$$

### Bartlett-adjusted profile likelihood ratio tests

Statistical analyses with RNA-seq data often use small samples, and we selected the Bartlett-adjusted likelihood ratio test for identifying DEGs. Bartlett's adjustments make the likelihood ratio statistic close to its null distribution with reducing the order of approximation error from  $O(N^{-1})$  to  $O(N^{-2})$  and control the type-1 error rates well when the sample size is small (Bartlett, 1937). If we let  $\boldsymbol{\beta}_g = (\beta_{g,1}, \dots, \beta_{g,M-1})^t$ ,  $\mathbf{V}_g = \boldsymbol{\Sigma}_{g,\mathbf{b}} + \sigma_g^2 \mathbf{I}$  and  $\boldsymbol{\theta}_g = (\boldsymbol{\alpha}_g, \sigma_g^2)$ , the likelihood for the proposed linear mixed model is

$$L(\boldsymbol{\theta}_g, \boldsymbol{\beta}_g) \propto |\mathbf{V}_g|^{-\frac{1}{2}} \exp\left(-\frac{1}{2} (\mathbf{Y}_g - \mathbf{Z}\boldsymbol{\alpha}_g - \mathbf{X}\boldsymbol{\beta}_g)^t \mathbf{V}_g^{-1} (\mathbf{Y}_g - \mathbf{Z}\boldsymbol{\alpha}_g - \mathbf{X}\boldsymbol{\beta}_g)\right).$$

If we let  $\hat{\boldsymbol{\theta}}_{g0}$  be a maximum likelihood estimate (mle) under the parameter space for the null hypothesis  $H_0: \boldsymbol{\beta}_g = \mathbf{0}$ , and  $(\hat{\boldsymbol{\theta}}_g, \hat{\boldsymbol{\beta}}_g)$  be mles of  $(\boldsymbol{\theta}_g, \boldsymbol{\beta}_g)$  under the parameter space for null or alternative hypothesis, the likelihood ratio test for the null hypothesis  $H_0: \boldsymbol{\beta}_g = \mathbf{0}$  can be obtained by

$$LR_g = -2\{\log L(\hat{\boldsymbol{\theta}}_{g0}, 0) - \log L(\hat{\boldsymbol{\theta}}_g, \hat{\boldsymbol{\beta}}_g)\} \sim \chi^2(df = M - 1) \text{ under } H_0.$$

The Bartlett-adjusted likelihood ratio test ( $LR_g^*$ ) for gene  $g$  can be expressed by

$$LR_g^* = \frac{LR_g}{1+C_g/(M-1)} \sim \chi^2(df = M - 1) \text{ under } H_0.$$

$C_g$  can be obtained based on the results of Melo et al. (2009) (Melo, et al., 2009), as follows:

$$C_g = D_g^{-1} \left( -\frac{1}{2}M_g + \frac{1}{4}P_g - \frac{1}{2}v_g\tau_g \right).$$

Here,  $D_g$ ,  $M_g$ ,  $P_g$ ,  $v_g$ , and  $\tau_g$  are scalars, and if we let  $\mathbf{X}' = [\mathbf{I} - \mathbf{Z}(\mathbf{Z}^T\mathbf{V}_g^{-1}\mathbf{Z})^{-1}\mathbf{Z}^t\mathbf{V}_g^{-1}]\mathbf{X}$

and  $\dot{\mathbf{X}}' = \mathbf{Z}(\mathbf{Z}^t\mathbf{V}_g^{-1}\mathbf{Z})^{-1}\mathbf{Z}^t\mathbf{V}_g^{-2}\mathbf{X}'$ , they are

$$D_g = -\frac{1}{2}\text{tr}(\mathbf{V}_g^{-2}),$$

$$M_g = 2\text{tr}\left(\left(\mathbf{X}'^t\mathbf{V}_g^{-1}\mathbf{X}'\right)^{-1}\left(\mathbf{X}'^t\mathbf{V}_g^{-3}\mathbf{X}' - \dot{\mathbf{X}}'^t\mathbf{V}_g^{-2}\mathbf{X}'\right)\right),$$

$$P_g = \text{tr}\left(\left(\mathbf{X}'^t\mathbf{V}_g^{-2}\mathbf{X}'\left(\mathbf{X}'^t\mathbf{V}_g^{-1}\mathbf{X}'\right)^{-1}\right)^2\right),$$

$$v_g = -\text{tr}\left(\left(\mathbf{Z}^t\mathbf{V}_g^{-1}\mathbf{Z}\right)^{-1}\mathbf{Z}^t\mathbf{V}_g^{-2}\mathbf{Z}\right) \text{ and}$$

$$\tau_g = -\text{tr}\left(\left(\mathbf{X}'^t\mathbf{V}_g^{-1}\mathbf{X}'\right)^{-1}\mathbf{X}'^t\mathbf{V}_g^{-2}\mathbf{X}'\right).$$

The forms of  $D_g$ ,  $M_g$ ,  $P_g$ ,  $v_g$ , and  $\tau_g$  depend on the structure of  $\mathbf{V}_g$  and counterparts to general  $\mathbf{V}_g$ s are shown in the Supplementary Text A.

### Parameter estimation

The log-likelihood function for our final model is given by

$\log L(\boldsymbol{\theta}_g, \boldsymbol{\beta}_g) = C - \frac{1}{2} \log |\mathbf{V}_g| - \frac{1}{2} (\mathbf{Y}_g - \mathbf{Z}\boldsymbol{\alpha}_g - \mathbf{X}\boldsymbol{\beta}_g)^t \mathbf{V}_g^{-1} (\mathbf{Y}_g - \mathbf{Z}\boldsymbol{\alpha}_g - \mathbf{X}\boldsymbol{\beta}_g)$ , C: some constants.

$\widehat{\boldsymbol{\theta}}_g$  and  $\widehat{\boldsymbol{\beta}}_g$  can be estimated by maximizing the log-likelihood function. Then, if we let  $\mathbf{P} = \mathbf{V}_g^{-1} - \mathbf{V}_g^{-1}(\mathbf{Z}, \mathbf{X})((\mathbf{Z}, \mathbf{X})^t \mathbf{V}_g^{-1}(\mathbf{Z}, \mathbf{X}))(\mathbf{Z}, \mathbf{X})^t \mathbf{V}_g^{-1}$ , the profile log-likelihood of  $\sigma_g^2$  becomes

$$l_p(\sigma_g^2) = C - \frac{1}{2} \log |\mathbf{V}_g| - \frac{1}{2} \mathbf{Y}_g^t \mathbf{P} \mathbf{Y}_g.$$

Here,  $\mathbf{V}_g$  is a function of  $\sigma_g^2$ , and  $\widehat{\sigma}_g^2$  can be obtained by maximizing  $l_p(\sigma_g^2)$  with Fisher's Scoring method. If we let  $\hat{c}_{gi}^{(m-1)}$  be  $i$ th component of  $\mathbf{Z}\widehat{\boldsymbol{\alpha}}_g^{(m-1)} - \mathbf{X}\widehat{\boldsymbol{\beta}}_g^{(m-1)}$ ,  $\widehat{\sigma}_g^{2(m)}$  at the  $m$  step was updated by

$$\widehat{\sigma}_g^{2(m)} = \frac{\sum_{i=1}^N (\hat{\mu}_{gi}^{-1} + \widehat{\sigma}_g^{2(m-1)})^{-2} \left\{ (y_{gi} - \hat{c}_{gi}^{(m-1)})^2 - \hat{\mu}_{gi}^{-1} \right\}}{\sum_{i=1}^N (\hat{\mu}_{gi}^{-1} + \widehat{\sigma}_g^{2(m-1)})^{-2}}$$

(Viechtbauer, 2007). We found that Fisher's Scoring method was sometimes unsuccessful for estimating  $\widehat{\sigma}_g^{2(m)}$ , and in such cases, we used Brent's derivative free method (Brent, 1973) with the *optimize* function in R. We assumed that  $\widehat{\sigma}_g^2$  was non-negative.  $\widehat{\mathbf{V}}_g = \boldsymbol{\Sigma}_{g,b} + \widehat{\sigma}_g^2 \mathbf{I}$ , and if  $\widehat{\sigma}_g^2$  is equal to zero,  $\widehat{\mathbf{V}}_g$  becomes  $\boldsymbol{\Sigma}_{g,b}$ . Then,  $\widehat{\boldsymbol{\beta}}_g$  and  $\widehat{\boldsymbol{\alpha}}_g$  can be obtained by

$$\begin{pmatrix} \widehat{\boldsymbol{\alpha}}_g \\ \widehat{\boldsymbol{\beta}}_g \end{pmatrix} = \left( (\mathbf{Z}, \mathbf{X})^t \widehat{\mathbf{V}}_g^{-1} (\mathbf{Z}, \mathbf{X}) \right)^{-1} (\mathbf{Z}, \mathbf{X})^t \widehat{\mathbf{V}}_g^{-1} \mathbf{Y}.$$

The Bartlett-adjusted likelihood ratio test requires  $\widehat{\boldsymbol{\theta}}_{g0}$ , which maximizes the likelihood under the null hypothesis. Under the null hypothesis, if we let  $\mathbf{P}_0 = \mathbf{V}_g^{-1} - \mathbf{V}_g^{-1} \mathbf{Z}(\mathbf{Z}^t \mathbf{V}_g^{-1} \mathbf{Z}) \mathbf{Z}^t \mathbf{V}_g^{-1}$ , the profile log-likelihood of  $\sigma_g^2$  becomes

$$l_{P_0}(\sigma_g^2) = C - \frac{1}{2} \log |\mathbf{V}_g| - \frac{1}{2} \mathbf{Y}_g^t \mathbf{P}_0 \mathbf{Y}_g.$$

$\hat{\sigma}_{g_0}^2$  was estimated with the Fisher's scoring method, and if we incorporated  $\hat{\sigma}_{g_0}^2$  to  $\mathbf{V}_{g_0}$  and denoted it as  $\hat{\mathbf{V}}_{g_0}$ ,  $\hat{\boldsymbol{\alpha}}_g$  could be obtained by

$$\hat{\boldsymbol{\alpha}}_g = \left( \mathbf{Z}^t \hat{\mathbf{V}}_{g_0}^{-1} \mathbf{Z} \right)^{-1} \mathbf{Z}^t \hat{\mathbf{V}}_{g_0}^{-1} \mathbf{Y}.$$

## Datasets

We considered two real datasets consisting of unrelated Nigerian and Holstein samples, respectively. Nigerian subjects were participated in the International HapMap Project and were composed of 29 males and 40 females (Pickrell, et al., 2010). The read counts were downloaded from the ReCount website (Frazee, et al., 2011). Holstein data were obtained to identify genes associated with milk yield and consisted of high and low milk yielding groups whose number of subjects is 9 and 12, respectively (Seo, et al., 2016). Furthermore, parity and lactation periods were available and were considered as covariates. Steps for transformation from the raw sequence data to read counts are shown in Supplementary Text B. Based on count data, we generated simulation data and the steps are described in the Supplementary Text C.

## RESULTS

### Simulation studies with Nigerian RNA-seq data

We applied the proposed linear mixed models to the simulated data based on Nigerian RNA-seq data and calculated empirical type-1 error rates and statistical powers with these models.

The data were then compared with DESeq2 (v1.14.1), edgeR (v3.16.5) and voom (v3.30.13). Table 1, Supplementary Table 1 and Figure 1 show results from simulation studies based on Nigerian RNA-seq data. Nigerian RNA-seq data consisted of 52,580 genes, and after filtering genes whose total read counts across samples were smaller than one tenth of the sample size, each replicate had around 10,000–10,500 genes. Empirical type-1 error rates and powers were estimated with 20 replicates. Table 1 and Supplementary Table 1 assumed  $\delta = 0$ , and thus, their estimates indicated the empirical type-1 error rates. For the proposed methods, we assumed that  $\psi_g = 1$  and  $\sigma_{g1}^2 = \sigma_{g2}^2$ , and the proposed methods with and without Bartlett’s corrections are denoted as BALLI and LLI, respectively, for the remainder of this article. According to Table 1 and Supplementary Table 1, BALLI and voom always controlled the nominal type-1 error rates correctly. LLI also successfully controlled the nominal type-1 error rates if  $N$  was larger than or equal to 20. However, if  $N = 12$  or 16,  $p$  values by LLI were inflated. edgeR showed the least performance, and the estimated type-1 error rates were always inflated at 0.05, 0.01, and 0.005 nominal significance levels. Interestingly, DESeq2 tended to be conservative at 0.1 and 0.05, but liberal at 0.01 and 0.005 nominal significance levels. Thus, we could conclude that the proposed linear mixed model with Bartlett’s correction reasonably controlled the type-1 error, and Bartlett’s correction was required if the sample size was less than 20.

**Table 1.** Estimated type-1 error rates with simulation data based on Nigerian data. Estimated type-1 error rates by BALLI, DESeq2, edgeR, LLI and voom and their 95% confidence levels were estimated for  $N = 12, 16, 20$  and 24. The type-1 error rates are marked by bold font if their 95% confidence levels include or lower than the nominal significant level  $\alpha$ .

$\alpha$	$N = 12$					$N = 16$				
	BALLI	DESeq2	edgeR	LLI	voom	BALLI	DESeq2	edgeR	LLI	voom
0.1	<b>0.10722</b> ( <b>0.08803</b> , <b>0.12640</b> )	<b>0.09581</b> ( <b>0.07622</b> , <b>0.11540</b> )	0.12502( 0.10744, 0.14260)	0.14288( 0.12020, 0.16556)	<b>0.10497</b> ( <b>0.08608</b> , <b>0.12386</b> )	<b>0.09810</b> ( <b>0.07994</b> , <b>0.11625</b> )	<b>0.09133</b> ( <b>0.07252</b> , <b>0.11015</b> )	0.12072( 0.10319, 0.13826)	0.12429( 0.10349, 0.14509)	<b>0.10640</b> ( <b>0.08854</b> , <b>0.12425</b> )
0.05	<b>0.05313</b> ( <b>0.04158</b> , <b>0.06469</b> )	<b>0.05203</b> ( <b>0.03888</b> , <b>0.06518</b> )	0.06816( 0.05662, 0.07970)	0.07703( 0.06158, 0.09247)	<b>0.05147</b> ( <b>0.03992</b> , <b>0.06303</b> )	<b>0.04811</b> ( <b>0.03602</b> , <b>0.06021</b> )	<b>0.04928</b> ( <b>0.03602</b> , <b>0.06293</b> )	0.06605( 0.05395, 0.07814)	0.06655( 0.05194, 0.08117)	<b>0.05291</b> ( <b>0.04107</b> , <b>0.06474</b> )

0.01	<b>0.01000(0.00748, 0.01252)</b>	<b>0.01393(0.00942, 0.01843)</b>	0.01785(0.01386, 0.02184)	0.01917(0.01441, 0.02392)	<b>0.00957(0.00672, 0.01241)</b>	<b>0.00977(0.00603, 0.01351)</b>	<b>0.01356(0.00790, 0.01922)</b>	0.01688(0.01240, 0.02136)	0.01573(0.01035, 0.02112)	<b>0.01065(0.00701, 0.01429)</b>
0.005	<b>0.00504(0.00371, 0.00638)</b>	0.00834(0.00553, 0.01116)	0.01063(0.00811, 0.01314)	0.01016(0.00748, 0.01284)	<b>0.00449(0.00303, 0.00594)</b>	<b>0.00481(0.00273, 0.00689)</b>	<b>0.00787(0.00410, 0.01164)</b>	0.00974(0.00689, 0.01260)	0.00862(0.00527, 0.01198)	<b>0.00531(0.00323, 0.00740)</b>
$\alpha$	<i>N</i> = 20					<i>N</i> = 24				
	BALLI	DESeq2	edgeR	LLI	voom	BALLI	DESeq2	edgeR	LLI	voom
0.1	<b>0.09292(0.07721, 0.10863)</b>	<b>0.09016(0.07371, 0.10662)</b>	0.11795(0.10291, 0.13299)	<b>0.11330(0.09560, 0.13100)</b>	<b>0.10581(0.08793, 0.12369)</b>	<b>0.09300(0.07477, 0.11123)</b>	<b>0.08930(0.06998, 0.10862)</b>	<b>0.11535(0.09727, 0.13344)</b>	<b>0.10886(0.08888, 0.12883)</b>	<b>0.09738(0.07824, 0.11651)</b>
0.05	<b>0.04441(0.03436, 0.05446)</b>	<b>0.04736(0.03597, 0.05875)</b>	0.06353(0.05333, 0.07372)	<b>0.05834(0.04603, 0.07065)</b>	<b>0.05343(0.04226, 0.06459)</b>	<b>0.04528(0.03330, 0.05726)</b>	<b>0.04822(0.03470, 0.06174)</b>	0.06513(0.05302, 0.07725)	<b>0.05621(0.04247, 0.06995)</b>	<b>0.04771(0.03541, 0.06002)</b>
0.01	<b>0.00797(0.00537, 0.01058)</b>	<b>0.01156(0.00750, 0.01563)</b>	0.01541(0.01212, 0.01870)	<b>0.01221(0.00858, 0.01583)</b>	<b>0.00949(0.00699, 0.01200)</b>	<b>0.00846(0.00512, 0.01181)</b>	<b>0.01269(0.00739, 0.01800)</b>	0.01681(0.01244, 0.02117)	<b>0.01224(0.00762, 0.01685)</b>	<b>0.00921(0.00570, 0.01272)</b>
0.005	<b>0.00367(0.00231, 0.00504)</b>	<b>0.00632(0.00378, 0.00885)</b>	0.00873(0.00677, 0.01068)	<b>0.00626(0.00414, 0.00838)</b>	<b>0.00443(0.00311, 0.00575)</b>	<b>0.00419(0.00236, 0.00602)</b>	<b>0.00717(0.00385, 0.01050)</b>	0.00962(0.00690, 0.01235)	<b>0.00625(0.00368, 0.00883)</b>	<b>0.00432(0.00248, 0.00617)</b>

**Figure 1.** Estimated Powers and Precisions with simulation data based on Nigerian RNA-seq data. Statistical powers of BALLI, DESeq2, edgeR, LLI and voom were estimated at FDR-adjusted 0.1 significance level when  $\delta = 0.8\sigma$ ,  $1\sigma$  and

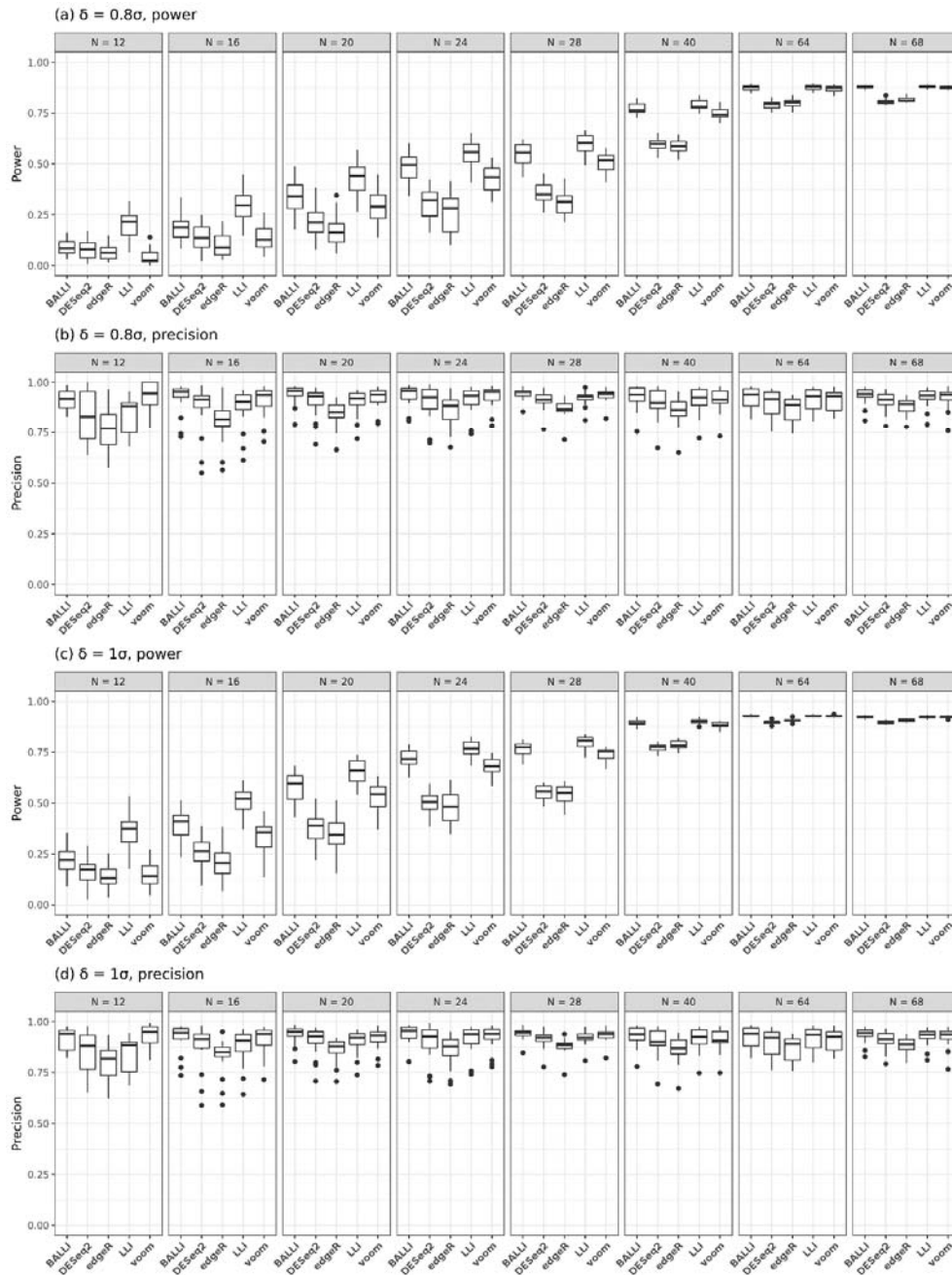


Figure 1 shows estimated powers and precisions at the FDR-adjusted 0.1 significance level when  $\delta = 0.8\sigma$  or  $1\sigma$  and  $N = 12, 16, 20, 24, 28, 40, 64$ , or  $68$ . Figure 1a and 1c show the statistical power estimates, and Figure 1b and 1d show the precision. Precision indicates the

proportions of DEGs among genes for which FDR-adjusted  $p$  values are less than 0.1. According to Figure 1a and 1c, LLI outperformed other methods in terms of power (Figure 1a and 1c). However, it should be noted that this method showed worse precision than BALLI and voom if  $N$  was less than 20 (Figure 1b and 1d), suggesting that LLI had larger false-positive rates than BALLI and voom when  $N$  was less than 20. The precision of LLI was increased if  $N$  was sufficiently large. In terms of both power and precision, the best performance was always obtained by BALLI. For example, when  $N = 20$  and  $\delta = \sigma$ , the estimated power of BALLI was 0.577, followed by voom (0.526) and DESeq2 (0.376). The estimated precision of BALLI was 0.936, and those of voom and DESeq2 were 0.919 and 0.906, respectively. If  $N = 16$  and  $\delta = 0.8\sigma$ , the estimated power and precision of BALLI were 0.188 and 0.926, which were higher than those of DESeq2 (0.137, 0.870) and voom (0.138, 0.910).

### **Simulation studies with randomly generated RNA-seq data**

RNA-seq data are generally known to follow the negative binomial distribution, and we conducted simulation studies with RNA-seq data generated from negative binomial distributions. First, we assumed that library sizes were the same among subjects. The overall trend of the estimated type-1 error rate was similar to that of simulation studies based on Nigerian RNA-seq data. Estimated type-1 error rates by voom and BALLI usually maintained the nominal significance levels (Table 2 and Supplementary Table 2).  $P$  values obtained from LLI and edgeR tended to be inflated, but the amount of inflation by LLI was small compared with that of edgeR. DESeq2 generally showed deflation of type-1 error rates at 0.1 and 0.05 nominal significance levels. Second, we considered the effects of library size variance on statistical analyses. Data with unequal library sizes among subjects were generated by



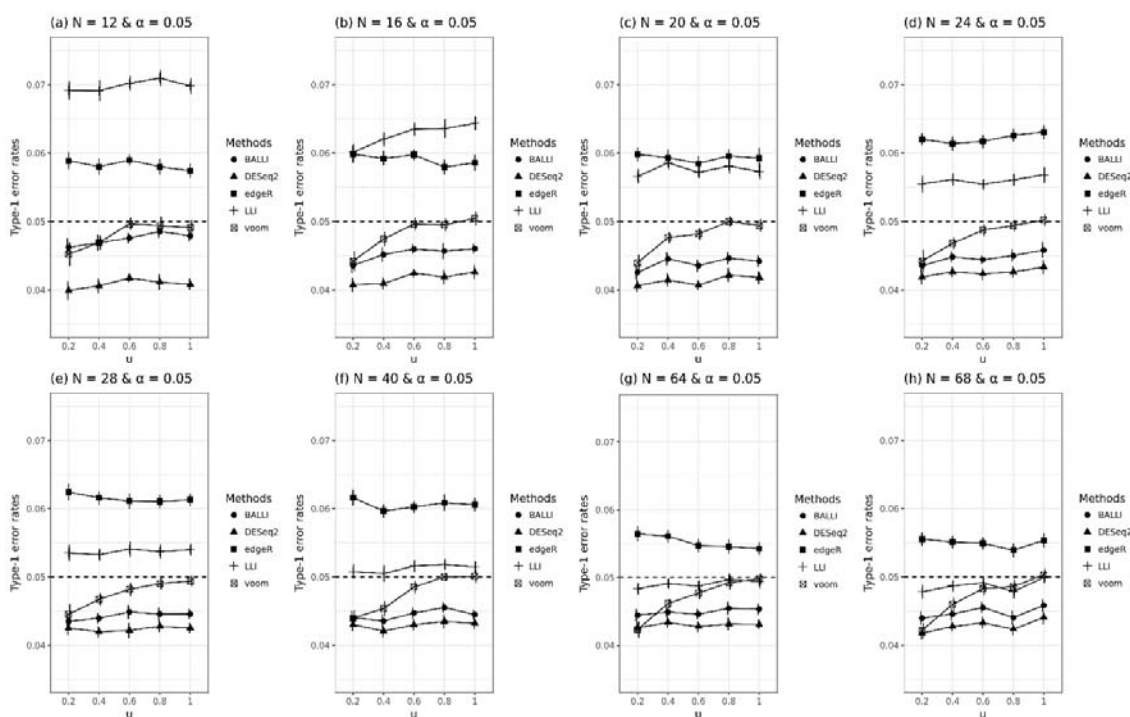
negative binomial distribution whose mean parameters ( $a_{gi}$ ) were the product of mean estimates, under the equal library size assumption, and random numbers from  $U(u, 2 - u)$ , where  $u = 0.2, 0.4, 0.6, 0.8$  or  $1$ , and dispersion parameters were estimated from  $(0.2 + a_{gi}^{-1/2})^2 \times \delta_g$ , where  $40/\delta_g \sim \chi_{40}^2$ . If  $u$  became smaller, the library size had larger variances. Figure 2 shows the estimated type-1 error rates at the 0.05 significance level according to different choices of  $u$ . Figure 2 shows that voom was sensitive to the amount of library size variation and became conservative in the context of large library size variation. Compared with voom, BALLI and LLI were robust with regard to  $u$ . The estimated type-1 error rates of LLI were affected by sample size. If  $N$  was larger than or equal to 40, LLI controlled the type-1 error rates the most correctly and was not affected by the library size variation. BALLI was slightly conservative, but the amount remained constant. Results at the 0.005 significance level are provided in Supplementary Figure 1, and the general pattern was similar to that in Figure 2, except that DESeq2 was relatively closer to the base line (Supplementary Figure 1). Therefore, we could conclude that the performances of BALLI and LLI were robust, and we recommend using BALLI if  $10 \leq N \leq 40$  and LLI if  $N > 40$ .

**Table 2.** Estimated type-1 error rates with simulation data based on simulated RNA-seq data from negative binomial distribution. Estimated type-1 error rates by BALLI, DESeq2, edgeR, LLI and voom and their 95% confidence levels were estimated for  $N = 12, 16, 20$  and  $24$ . The type-1 error rates are marked by bold font if their 95% confidence levels include or lower than the nominal significant level  $\alpha$ .

$\alpha$	$N = 12$					$N = 16$				
	BALLI	DESeq2	edgeR	LLI	voom	BALLI	DESeq2	edgeR	LLI	Voom
0.1	<b>0.09728</b> ( <b>0.09619</b> , <b>0.09838</b> )	<b>0.08127</b> ( <b>0.08013</b> , <b>0.08241</b> )	0.12351 (0.12234, 0.12468)	0.13228 (0.13118, 0.13339)	<b>0.09869</b> ( <b>0.09750</b> , <b>0.09989</b> )	<b>0.09214</b> ( <b>0.09054</b> , <b>0.09374</b> )	<b>0.08275</b> ( <b>0.08145</b> , <b>0.08405</b> )	0.12219 (0.12094, 0.12345)	0.11895 (0.11711, 0.12079)	<b>0.10026</b> ( <b>0.09872</b> , <b>0.10180</b> )
0.05	<b>0.04792</b> ( <b>0.04708</b> , <b>0.04876</b> )	<b>0.04082</b> ( <b>0.03997</b> , <b>0.04168</b> )	0.05743 (0.05644, 0.05842)	0.06983 (0.06877, 0.07089)	<b>0.04915</b> ( <b>0.04806</b> , <b>0.05023</b> )	<b>0.04607</b> ( <b>0.04536</b> , <b>0.04677</b> )	<b>0.04261</b> ( <b>0.04163</b> , <b>0.04359</b> )	0.05859 (0.05744, 0.05974)	0.06435 (0.06333, 0.06536)	<b>0.05046</b> ( <b>0.04950</b> , <b>0.05142</b> )
0.01	<b>0.00990</b> ( <b>0.00943</b> , <b>0.01037</b> )	<b>0.00877</b> ( <b>0.00839</b> , <b>0.00915</b> )	0.01272 (0.01227, 0.01317)	0.01798 (0.01743, 0.01853)	<b>0.00997</b> ( <b>0.00948</b> , <b>0.01046</b> )	<b>0.00955</b> ( <b>0.00911</b> , <b>0.00999</b> )	<b>0.00980</b> ( <b>0.00934</b> , <b>0.01027</b> )	0.01379 (0.01332, 0.01427)	0.01575 (0.01525, 0.01625)	<b>0.01032</b> ( <b>0.00992</b> , <b>0.01073</b> )
0.005	<b>0.00524</b> ( <b>0.00488</b> , <b>0.00559</b> )	<b>0.00497</b> ( <b>0.00463</b> , <b>0.00530</b> )	0.00703 (0.00668, 0.00737)	0.01034 (0.00986, 0.01082)	<b>0.00522</b> ( <b>0.00489</b> , <b>0.00555</b> )	<b>0.00480</b> ( <b>0.00453</b> , <b>0.00506</b> )	0.00550 (0.00523, 0.00576)	0.00757 (0.00726, 0.00788)	0.00857 (0.00816, 0.00899)	<b>0.00517</b> ( <b>0.00490</b> , <b>0.00544</b> )
$\alpha$	$N = 20$					$N = 24$				
	BALLI	DESeq2	edgeR	LLI	voom	BALLI	DESeq2	edgeR	LLI	voom
0.1	<b>0.09188</b>	<b>0.08277</b>	0.11842	0.11062	<b>0.09941</b>	<b>0.09491</b>	<b>0.08516</b>	0.11899	0.11056	<b>0.10126</b>

	<b>(0.09069,</b> <b>0.09307)</b>	<b>(0.08179,</b> <b>0.08375)</b>	(0.11718, 0.11967)	(0.10960, 0.11164)	<b>(0.09834,</b> <b>0.10048)</b>	<b>(0.09349,</b> <b>0.09634)</b>	<b>(0.08383,</b> <b>0.08650)</b>	(0.11755, 0.12043)	(0.10909, 0.11203)	<b>(0.09972,</b> <b>0.10280)</b>
0.05	<b>0.04421</b> <b>(0.04335,</b> <b>0.04506)</b>	<b>0.04184</b> <b>(0.04088,</b> <b>0.04280)</b>	0.05927 (0.05780, 0.06075)	0.05732 (0.05625, 0.05840)	<b>0.04942</b> <b>(0.04853,</b> <b>0.05031)</b>	<b>0.04585</b> <b>(0.04474,</b> <b>0.04696)</b>	<b>0.04335</b> <b>(0.04233,</b> <b>0.04437)</b>	0.06308 (0.06207, 0.06409)	0.05684 (0.05568, 0.05799)	<b>0.05021</b> <b>(0.04936,</b> <b>0.05105)</b>
0.01	<b>0.00887</b> <b>(0.00852,</b> <b>0.00922)</b>	<b>0.00967</b> <b>(0.00913,</b> <b>0.01021)</b>	0.01344 (0.01299, 0.01389)	0.01326 (0.01295, 0.01356)	<b>0.01005</b> <b>(0.00970,</b> <b>0.01040)</b>	<b>0.00897</b> <b>(0.00849,</b> <b>0.00944)</b>	<b>0.00993</b> <b>(0.00955,</b> <b>0.01032)</b>	0.01377 (0.01328, 0.01426)	0.01265 (0.01217, 0.01313)	<b>0.01007</b> <b>(0.00963,</b> <b>0.01051)</b>
0.005	<b>0.00455</b> <b>(0.00432,</b> <b>0.00479)</b>	0.00536 (0.00504, 0.00568)	0.00738 (0.00698, 0.00778)	0.00718 (0.00687, 0.00748)	<b>0.00503</b> <b>(0.00477,</b> <b>0.00529)</b>	<b>0.00440</b> <b>(0.00419,</b> <b>0.00462)</b>	<b>0.00515</b> <b>(0.00484,</b> <b>0.00547)</b>	0.00733 (0.00696, 0.00770)	0.00678 (0.00640, 0.00716)	<b>0.00498</b> <b>(0.00466,</b> <b>0.00530)</b>

**Figure 2.** Effect of varying library sizes on the type-1 error rates. Type-1 error rates were estimated by BALLI, DESeq2, edgeR, LLI and voom when  $u = 0.2, 0.4, 0.6, 0.8$  and  $1$  and sample size ( $N$ ) is  $12, 16, 20, 24, 28, 40, 64$  or  $68$  at the  $0.05$  nominal significance level.



**Figure 3.** Effect of varying library sizes on the statistical power and precision. Statistical powers and precision for BALLI, DESeq2, edgeR, LLI and voom were empirically estimated at FDR-adjusted 0.1 significance level when  $u = 0.2$ ,  $\delta = 0.8\sigma$  or  $1\sigma$  and sample size ( $N$ ) is 12, 16, 20, 24, 28, 40, 64 or 68.

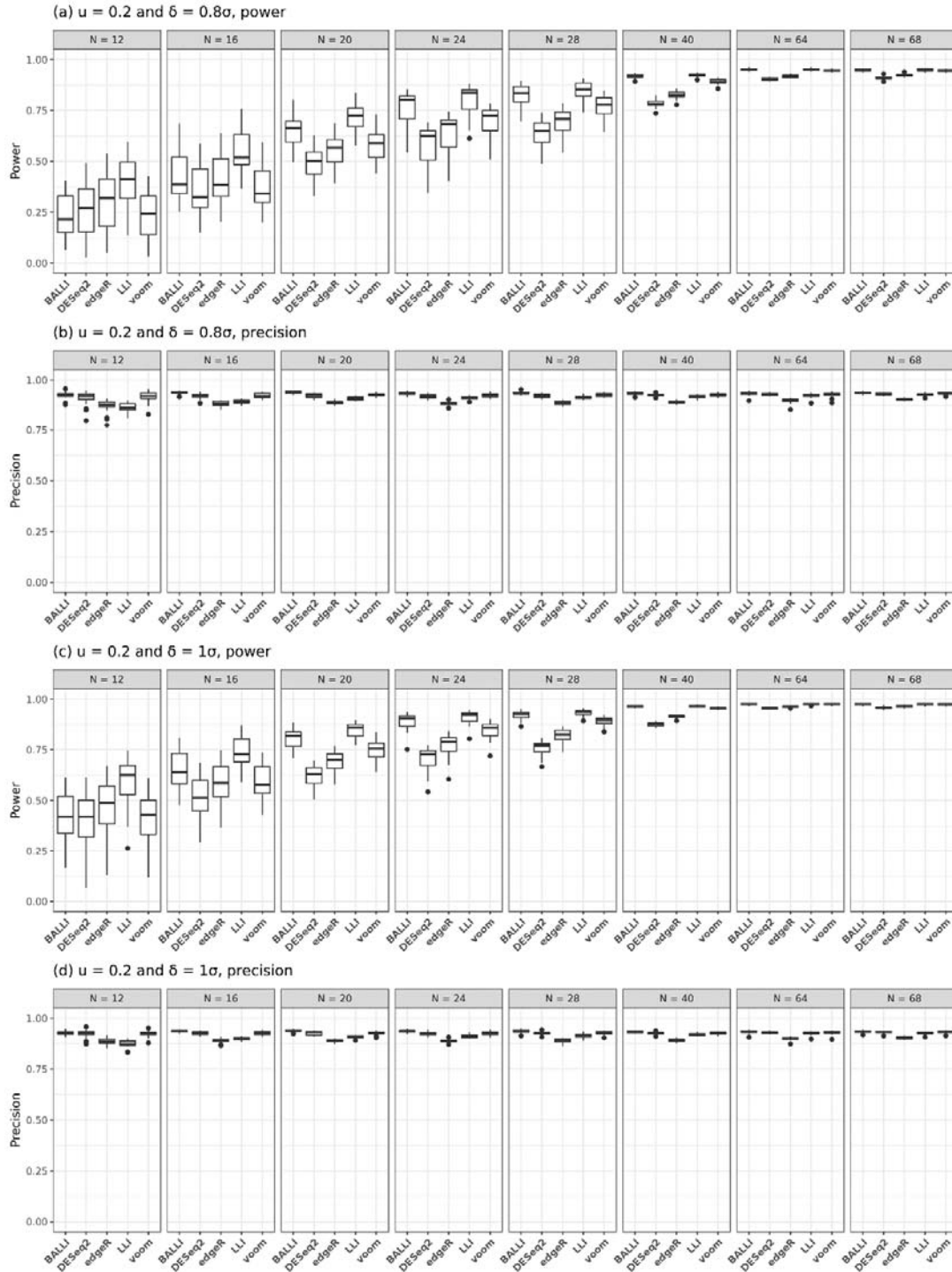


Figure 3 and Supplementary Figure 2 show the estimated statistical powers and precision according to different choices of  $u$ . BALLI usually had the best estimated power and precision, as was observed in simulation studies based on Nigerian RNA-seq data. For example, when  $u = 0.2$ ,  $N = 20$ , and  $\delta = 0.8\sigma$ , the estimated power by BALLI was 0.641, whereas those for DESeq2 and voom were 0.480 and 0.571, respectively (Figure 3a). Results when  $u = 0.2$  and  $\delta = 1\sigma$  in Figure 3c are very similar as those for Figure 3a. Figure 3b and Figure 3d also shows that BALLI achieve the best estimated precisions. Similar patterns were observed when  $u = 0.4, 0.6, 0.8, \text{ and } 1$  (Supplementary Figure 2). In summary, we can conclude that BALLI shows better performance than other methods.

### **DEGs of Holstein milk data**

Holstein milk data, consisting of 21 Holstein cows, were generated to detect genes related to the productivity of daily milk. High and low milk yields were considered the primary exposure variables, and parity and lactation period were included as covariates (Seo, et al., 2016). In this study, twelve tentative DEGs were chosen, and technically validated using quantitative real time polymerase chain reaction (qRT-PCR). qRT-PCR was conducted with QuantiTect SYBR Green RT-PCR Master Mix (Qiagen, Valencia, CA, USA), and a 7500 Fast Sequence Detection System (Applied Biosystems, Foster City, CA, USA) was used to confirm whether the twelve tentative genes were true DEGs. Among the twelve genes, nine (*TOX4*, *HNRNPL*, *SPTSSB*, *NOS3*, *C25H16orf88*, *KALRN*, *SLC4A1*, *NLN*, and *PMCH*) were significantly validated. According to Seo et al (2016), however, no DEGs including the nine genes were found at FDR 0.1 significance level by DESeq2 and voom as well as their methods due to the lack of statistical power (Seo, et al., 2016). Our proposed methods and existing methods (DESeq2, edgeR, and voom) were applied to the data analysis. LLI only

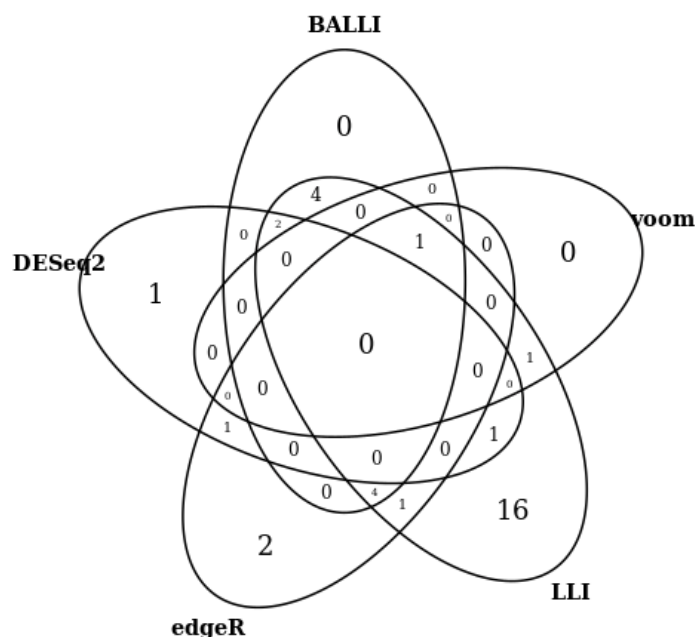
detected significant differences for the *TOX4* gene between the high and low milk yield groups at the FDR-adjusted 0.1 significance level, but other methods did not detect any significant genes. The FDR-adjusted  $p$  value of *TOX4* by BALLI was 0.1272, which was much smaller than those of DESeq2, edgeR, and voom. Table 3 shows  $p$  values for the nine genes, including *TOX4*.  $P$  values for the nine genes obtained by LLI and BALLI were small compared with those obtained from other methods. We also analyzed all genes with proposed methods; Figure 4 shows the number of genes that were significant at the 0.001 nominal significance level. There were no DEGs that were commonly significant only for all existing methods (DESeq2, edgeR and voom). Four genes, including *HNRNPL*, were detected as DEGs by only BALLI and LLI (Figure 4). Table 4 shows eight genes that were commonly significant by BALLI, DESeq2, edgeR, LLI, and voom at the 0.005 nominal significance level. Of the eight genes, all genes had the lowest  $p$  values in LLI, and three genes had lower  $p$  values in BALLI than in DESeq2, edgeR, and voom. Simulation studies revealed that LLI tended to be liberal, and the results may be inflated. However, BALLI controlled the nominal significance level, and  $p$  values by BALLI were expected to be statistically valid. Therefore, we concluded that the proposed method, BALLI, worked well for real data analysis.

**Table 3.** True DEG analysis results of Holstein milk data. Holstein milk data was analyzed by BALLI, DESeq2, edgeR, LLI and voom and their  $p$  values (FDRs) are provided

	BALLI	DESeq2	edgeR	LLI	voom
TOX4	$1.063 \times 10^{-5}$ ( $1.272 \times 10^{-1}$ )	$3.032 \times 10^{-4}$ ( $9.997 \times 10^{-1}$ )	$2.797 \times 10^{-2}$ (1)	$7.476 \times 10^{-7}$ ( $8.947 \times 10^{-3}$ )	$4.980 \times 10^{-3}$ ( $9.999 \times 10^{-1}$ )
HNRNPL	$3.914 \times 10^{-4}$ ( $8.400 \times 10^{-1}$ )	$1.602 \times 10^{-3}$ ( $9.997 \times 10^{-1}$ )	$1.684 \times 10^{-1}$ (1)	$6.796 \times 10^{-5}$ ( $1.787 \times 10^{-1}$ )	$4.677 \times 10^{-2}$ ( $9.999 \times 10^{-1}$ )
SPTSSB	$4.040 \times 10^{-4}$ ( $8.400 \times 10^{-1}$ )	$8.792 \times 10^{-3}$ ( $9.997 \times 10^{-1}$ )	$2.158 \times 10^{-4}$ (1)	$8.686 \times 10^{-5}$ ( $1.787 \times 10^{-1}$ )	$1.037 \times 10^{-3}$ ( $9.999 \times 10^{-1}$ )
NOS3	$4.460 \times 10^{-4}$ ( $8.400 \times 10^{-1}$ )	$1.083 \times 10^{-3}$ ( $9.997 \times 10^{-1}$ )	$2.549 \times 10^{-4}$ (1)	$8.957 \times 10^{-5}$ ( $1.787 \times 10^{-1}$ )	$8.693 \times 10^{-4}$ ( $9.999 \times 10^{-1}$ )
SLC4A1	$2.142 \times 10^{-2}$ (1)	$1.221 \times 10^{-2}$ ( $9.997 \times 10^{-1}$ )	$1.025 \times 10^{-1}$ (1)	$9.856 \times 10^{-3}$ ( $7.101 \times 10^{-1}$ )	$7.579 \times 10^{-2}$ ( $9.999 \times 10^{-1}$ )
NLN	$9.512 \times 10^{-2}$ (1)	$2.511 \times 10^{-1}$ ( $9.997 \times 10^{-1}$ )	$3.789 \times 10^{-1}$ (1)	$6.084 \times 10^{-2}$ ( $9.722 \times 10^{-1}$ )	$1.214 \times 10^{-1}$ ( $9.999 \times 10^{-1}$ )
KALRN	$9.790 \times 10^{-2}$ (1)	$9.661 \times 10^{-2}$ ( $9.997 \times 10^{-1}$ )	$8.815 \times 10^{-2}$ (1)	$6.307 \times 10^{-2}$ ( $9.740 \times 10^{-1}$ )	$1.054 \times 10^{-1}$ ( $9.999 \times 10^{-1}$ )
PMCH	$1.635 \times 10^{-1}$ (1)	$2.019 \times 10^{-1}$ ( $9.997 \times 10^{-1}$ )	$2.353 \times 10^{-1}$ (1)	$1.176 \times 10^{-1}$ (1)	$1.758 \times 10^{-1}$ ( $9.999 \times 10^{-1}$ )

C25H16orf88	$1.765 \times 10^{-1}$ (1)	$1.527 \times 10^{-1}$ ( $9.997 \times 10^{-1}$ )	$2.627 \times 10^{-1}$ (1)	$1.289 \times 10^{-1}$ (1)	$1.516 \times 10^{-1}$ ( $9.999 \times 10^{-1}$ )
-------------	-------------------------------	--	-------------------------------	-------------------------------	--

**Figure 4.** Significant genes of Holstein milk data. Venn diagram was provided with significant genes at the 0.001 nominal significance level by BALLI, DESeq2, edgeR, LLI and voom



**Table 4.** Significant genes in all methods of Holstein milk data. Gene lists of Holstein milk data significant in nominal 0.005 significant level for all methods (BALLI, DESeq2, edgeR, LLI and voom) and their p values (FDRs) are provided.

	BALLI	DESeq2	edgeR	LLI	voom
NOS3	$4.460 \times 10^{-4}$ ( $8.400 \times 10^{-1}$ )	$1.083 \times 10^{-3}$ ( $9.997 \times 10^{-1}$ )	$2.549 \times 10^{-4}$ (1)	$8.957 \times 10^{-5}$ ( $1.787 \times 10^{-1}$ )	$8.693 \times 10^{-4}$ ( $9.999 \times 10^{-1}$ )
SPESP1	$1.065 \times 10^{-3}$ ( $8.570 \times 10^{-1}$ )	$4.901 \times 10^{-3}$ ( $9.997 \times 10^{-1}$ )	$1.669 \times 10^{-3}$ (1)	$2.579 \times 10^{-4}$ ( $2.385 \times 10^{-1}$ )	$4.001 \times 10^{-3}$ ( $9.999 \times 10^{-1}$ )
CHST1	$1.330 \times 10^{-3}$ ( $8.570 \times 10^{-1}$ )	$3.074 \times 10^{-3}$ ( $9.997 \times 10^{-1}$ )	$1.872 \times 10^{-3}$ (1)	$3.166 \times 10^{-4}$ ( $2.385 \times 10^{-1}$ )	$6.108 \times 10^{-4}$ ( $9.999 \times 10^{-1}$ )
LEPREL1	$1.369 \times 10^{-3}$ ( $8.570 \times 10^{-1}$ )	$3.775 \times 10^{-3}$ ( $9.997 \times 10^{-1}$ )	$3.297 \times 10^{-3}$ (1)	$3.334 \times 10^{-4}$ ( $2.385 \times 10^{-1}$ )	$1.487 \times 10^{-3}$ ( $9.999 \times 10^{-1}$ )
JUB	$1.447 \times 10^{-3}$ ( $8.570 \times 10^{-1}$ )	$2.559 \times 10^{-3}$ ( $9.997 \times 10^{-1}$ )	$2.445 \times 10^{-3}$ (1)	$3.674 \times 10^{-4}$ ( $2.385 \times 10^{-1}$ )	$2.804 \times 10^{-3}$ ( $9.999 \times 10^{-1}$ )
MIA	$1.504 \times 10^{-3}$ ( $8.570 \times 10^{-1}$ )	$3.792 \times 10^{-4}$ ( $9.997 \times 10^{-1}$ )	$2.247 \times 10^{-3}$ (1)	$3.666 \times 10^{-4}$ ( $2.385 \times 10^{-1}$ )	$2.432 \times 10^{-3}$ ( $9.999 \times 10^{-1}$ )
CLDN6	$2.646 \times 10^{-3}$ (1)	$2.096 \times 10^{-3}$ ( $9.997 \times 10^{-1}$ )	$2.414 \times 10^{-3}$ (1)	$7.377 \times 10^{-4}$ ( $3.270 \times 10^{-1}$ )	$1.542 \times 10^{-3}$ ( $9.999 \times 10^{-1}$ )

PALMD	$3.733 \times 10^{-3}$ (1)	$1.676 \times 10^{-3}$ ( $9.997 \times 10^{-1}$ )	$2.648 \times 10^{-3}$ (1)	$1.127 \times 10^{-3}$ ( $3.700 \times 10^{-1}$ )	$3.032 \times 10^{-3}$ ( $9.999 \times 10^{-1}$ )
-------	-------------------------------	--	-------------------------------	--	--

## DISCUSSION

In this article, we suggested new methods, designated BALLI and LLI, for identifying DEGs with RNA-seq data. We assumed that log-cpm values of read counts asymptotically followed normal distributions, and the linear mixed effects model with Bartlett's correction was proposed. The proposed methods were compared with existing methods, such as DESeq2, edgeR, and voom, with extensive simulation studies. According to our results, negative-binomial-based approaches often failed to preserve the nominal type-1 error rates. For example,  $p$  values from edgeR were inflated. DESeq2 tended to be conservative and suffered from large false-negative rates. However, the proposed method with Bartlett's correction, BALLI, preserved the nominal type-1 error rates and was the most powerful method other than LLI. Unless sample sizes were small, LLI controlled the type-1 error rates as well and was the most powerful method. Therefore, we recommend using LLI if the sample size is sufficiently large (e.g., larger than 40); otherwise, it is better to use BALLI.

Furthermore, we evaluated the effects of library size variations on statistical analyses. We found that library size variance could affect the estimated type-1 error rates, and the effect was the largest for voom. Library sizes are affected by multiple factors, such as the amount of mRNA and the sequencing instrument, which can generate substantial variation among library sizes for subjects. Our simulation studies showed that BALLI was robust with regard to library size variation in samples of various sizes and was a reasonable choice if large library size variance was observed.

The proposed methods assumed that log-cpm values of read counts asymptotically followed a normal distribution and that their variances were approximately equal to  $1/\mu + \phi$  with the first order approximation. In addition, voom considered log-cpm value as a response and assumed that they were normally distributed. However, our simulation studies revealed

the superiority of the proposed methods compared with voom, which was found to be attributable to their different variance structures. For the proposed methods,  $1/\mu + \phi$  was derived from the first-order approximation of the negative binomial distribution and thus may be a natural assumption for RNA-seq data. Furthermore, for  $1/\mu + \phi$ ,  $\phi$  obviously indicates the overdispersion parameter, and biological and technical variances can be estimated with BALLI. However, voom assumes  $\phi/\mu$ , and the amount attributable to biological or technical variances cannot be clearly defined.

We also suggested the most flexible and general linear mixed model for log-cpm. The proposed model assumed that the variance of log-cpm was  $\varphi/\mu + \phi$  and had the most generalized variance parameter space. Incorporation of  $\varphi = 1$  yielded BALLI and LLI, and  $\phi = 0$  yielded voom. We found that BALLI was the most efficient in the considered scenarios; however, in real data analyses, various factors affected variance structure. For example, subjects with different ethnicities can cause  $\varphi$  to be larger than 1, and thus, a better model may differ according to RNA-seq data.  $\varphi$  and  $\phi$  can be estimated with the proposed linear mixed model by implementing only a simple modification, and thus, we can choose the best model using AIC or likelihood ratio tests. The selected models can then be utilized to identify DEGs. This model was implemented as an R package and can be downloaded from <http://healthstat.snu.ac.kr/software/balli/>. Furthermore, the proposed methods can be easily extended to various scenarios via a simple modification. For example, repeatedly observed data or multivariate phenotypes can be analyzed by adding some random effects. Maximizing the likelihood for negative binomial distributions with random effects is computationally intensive, but the proposed methods can easily obtain variance parameter estimates using existing R packages, such as lme4 and nlme.

With simulation studies for various scenarios, we showed that the proposed methods were usually the most efficient. However, results from simulation studies obviously depended on



various factors. Our results were obtained from simulation data based on Nigerian RNA-seq data and random samples from negative binomial distributions, but any systematic differences in RNA-seq data could generate different results, depending on sequencing errors or differences in preparation steps. Multiple studies have revealed some possible differences in these relationships, and our conclusions based on simulation studies could be limited to the considered scenarios. However, despite such limitations, we believe that our results illustrate the practical value of the proposed methods. Further studies are needed to confirm our findings and expand on the work presented herein.

## **FUNDING**

This research was supported by a grant of the Korea Health Technology R&D Project through the Korea Health Industry Development Institute (KHIDI), funded by the Ministry of Health & Welfare, Republic of Korea (H15C2165).

## REFERENCE

- Bartlett, M.S. (1937) Properties of sufficiency and statistical tests, *Proceedings of the Royal Society of London. Series A, Mathematical and Physical Sciences*, 268-282.
- Brent, R.P. (1973) Algorithms for minimization without derivatives.
- Bullard, J.H., *et al.* (2010) Evaluation of statistical methods for normalization and differential expression in mRNA-Seq experiments, *BMC bioinformatics*, **11**, 94.
- Chavance, M. and Escolano, S. (2016) Misspecification of the covariance structure in generalized linear mixed models, *Statistical methods in medical research*, **25**, 630-643.
- Dillies, M.A., *et al.* (2013) A comprehensive evaluation of normalization methods for Illumina high-throughput RNA sequencing data analysis, *Brief Bioinform*, **14**, 671-683.
- Dobbin, K.K., *et al.* (2005) Characterizing dye bias in microarray experiments, *Bioinformatics*, **21**, 2430-2437.
- Frazee, A.C., Langmead, B. and Leek, J.T. (2011) ReCount: a multi-experiment resource of analysis-ready RNA-seq gene count datasets, *BMC bioinformatics*, **12**, 449.
- Law, C.W., *et al.* (2014) voom: Precision weights unlock linear model analysis tools for RNA-seq read counts, *Genome biology*, **15**, R29.
- Litière, S., Alonso, A. and Molenberghs, G. (2008) The impact of a misspecified random-effects distribution on the estimation and the performance of inferential procedures in generalized linear mixed models, *Statistics in medicine*, **27**, 3125-3144.
- 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.
- Marioni, J.C., *et al.* (2008) RNA-seq: an assessment of technical reproducibility and comparison with gene expression arrays, *Genome research*, **18**, 1509-1517.
- McCarthy, D.J., Chen, Y. and Smyth, G.K. (2012) Differential expression analysis of multifactor RNA-Seq experiments with respect to biological variation, *Nucleic acids research*, **40**, 4288-4297.
- Melo, T.F., Ferrari, S.L. and Cribari-Neto, F. (2009) Improved testing inference in mixed linear models, *Computational Statistics & Data Analysis*, **53**, 2573-2582.
- Mortazavi, A., *et al.* (2008) Mapping and quantifying mammalian transcriptomes by RNA-Seq, *Nature methods*, **5**, 621-628.
- Okoniewski, M.J. and Miller, C.J. (2006) Hybridization interactions between probesets in short oligo microarrays lead to spurious correlations, *BMC bioinformatics*, **7**, 276.
- Pickrell, J.K., *et al.* (2010) Understanding mechanisms underlying human gene expression variation with RNA sequencing, *Nature*, **464**, 768-772.
- 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.
- Robinson, M.D. and Oshlack, A. (2010) A scaling normalization method for differential expression analysis of RNA-seq data, *Genome biology*, **11**, R25.
- Robinson, M.D. and Smyth, G.K. (2007) Moderated statistical tests for assessing differences in tag abundance, *Bioinformatics*, **23**, 2881-2887.
- Seo, M., *et al.* (2016) RNA-seq analysis for detecting quantitative trait-associated genes, *Scientific reports*, **6**.
- Seyednasrollah, F., Laiho, A. and Elo, L.L. (2015) Comparison of software packages for detecting differential expression in RNA-seq studies, *Briefings in bioinformatics*, **16**, 59-70.
- Soneson, C. and Delorenzi, M. (2013) A comparison of methods for differential expression analysis of RNA-seq data, *BMC bioinformatics*, **14**, 91.
- Tusher, V.G., Tibshirani, R. and Chu, G. (2001) Significance analysis of microarrays applied to the ionizing radiation response, *Proceedings of the National Academy of Sciences of the United States of America*, **98**, 5116-5121.
- Viechtbauer, W. (2007) Hypothesis tests for population heterogeneity in meta-analysis, *British Journal of Mathematical and Statistical Psychology*, **60**, 29-60.
- Zhao, S., *et al.* (2014) Comparison of RNA-Seq and microarray in transcriptome profiling of activated T cells, *PLoS one*, **9**, e78644.