Park et al. BMC Genomics (2019) 20:540
https://doi.org/10.1186/s12864-019-5851-6

BMC Genomics

METHODOLOGY ARTICLE Open Access

BALLI: Bartlett-adjusted likelihood-based linear model approach for identifying differentially expressed genes with RNA-seq data

Kyungtaek Park[1], Jaehoon An[2], Jungsoo Gim[3], Minseok Seo[4,5], Woojoo Lee[6], Taesung Park[1,7] and Sungho Won[1,2,8*]

Abstract

Background: 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 (DEGs) under two or more conditions with RNA-seq data. However, statistical analyses with RNA-seq data are often limited by 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-Adjusted 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-effects model, with and without adjusting small sample bias using Bartlkett'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 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.

Conclusions;: BALLI is statistically more efficient and valid than existing methods, and we conclude that it is useful for identifying DEGs in RNA-seq analysis.

Keywords: Differentially expressed genes, RNA sequencing, Linear mixed model, Bartlett's correction

Background

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 [1, 2]. 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 freedom from a priori chosen probes [3]. Furthermore, RNA-seq is robust against systematic errors and has therefore emerged as a successful alternative to microarray analysis [4].

RNA-seq quantifies the numbers of reads aligned to particular transcripts or genes, and various approaches have been proposed to manage RNA-seq data [5]. There are two different types of statistical methods: read-count-based approaches and transformation-based approaches. Read-count-based approaches assume that

* Correspondence: won1@snu.ac.kr
[1]Interdisciplinary Program of Bioinformatics, Seoul National University, Seoul 08826, South Korea
[2]Department of Public Health Science, Seoul National University, Seoul 08826, South Korea
Full list of author information is available at the end of the article

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 [6]. However, if biological replicates are available, means and variances of read counts differ 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 [7]. Thus, negative binomial distribution has often been utilized; edgeR [8] and DESeq2 [9] are such methods. Transformation-based approaches assume that the transformed read counts for each gene follow 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 normal distribution, assuming that the relative proportion of technical variances becomes smaller if the read count grows larger [10].

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 the 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 are assumed 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 [11, 12]. 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 [13]. Thus, their effects can distort the relationship between technical and biological variances. Multiple studies have shown that misspecified relationships can lead to substantial biases [14, 15]. For example, variance estimators for random effects, which are assumed to follow normal distribution, can be seriously biased unless they are normally distributed [14]. General approaches that are not sensitive to these problems are necessary.

In this article, we present new methods for identifying DEGs with RNA-seq data, namely, 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 [16, 17]. Thus, we considered log-cpm as response variables and used linear mixed-effects models to estimate technical and biological variance. Furthermore, Bartlett-adjusted likelihood ratio tests were applied to correct the small sample bias [18]. By allowing model comparisons among different models, our models enabled robust analyses for various scenarios. For our simulation studies, artificial RNA-seq data were 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 CRAN (http://cran.r-project.org) or http://healthstat.snu.ac.kr /software/balli/.

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; thus, we got $N = n_1 + \ldots + n_M$. We defined dummy variables for subject i in group m by

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

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

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 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. The effect of continuous variables can also be tested, and then some or all of the columns of the design matrix, **X**, become their realization. We assumed that expressed read counts were observed for G genes and were denoted by r_{gi} for gene g of subject i ($g = 1, ..., G$). Then, the library size for subject i, R_i, was equivalent to $R_i = \sum_g r_{gi}$. We transformed count data into the log-transformed read counts per million (log-cpm) using the *cpm* function in the edgeR R package. If we denoted the normalized R_i with the trimmed mean of the M-value [19] by R_i^*, the log-cpm of gene g for subject i were defined by:

$$\begin{aligned} y_{gi} &= \log_2\left(\frac{r_{gi}+d_i}{R_i^*+2*d_i}\times 10^6\right), \text{where } d_i \\ &= \frac{R_i^*}{\frac{1}{N}\sum_{i=1}^{N} R_i^*}\times 0.25... \end{aligned} \quad (1)$$

and their vector $\mathbf{Y_g}$ was defined as

$$\mathbf{Y_g}=\left(y_{g1}, y_{g2}, ..., y_{gN}\right)^t.$$

Technical and biological variances of y_{gi}

We assumed that r_{gi} followed a negative binomial distribution, and its mean and variance were μ_{gi} and $\mu_{gi}+\mu_{gi}^2\phi_g$, respectively. If we let the mean and variance of $\log_2 r_{gi}$ be λ_{gi} and s_{gi}^2, respectively, then var(y_{gi}) can be obtained by the first order approximation as follows [10]:

$$\begin{aligned} \mathrm{var}\left(y_{gi}\right) &\approx \mathrm{var}\left(\log_2 r_{gi}\right) \\ &= s_{gi}^2 \approx \mathrm{var}\left(\mu_{gi}+\frac{r_{gi}-\mu_{gi}}{\mu_{gi}}\right) = \frac{1}{{\mu_{gi}}^2}\, var\left(r_{gi}\right) \\ &= \frac{1}{\mu_{gi}}+\phi_g... \end{aligned} \quad (2)$$

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

$$\begin{aligned} \mu_{gi} &= E(r_{gi}) = \mathrm{E}\left(2^{\log_2 r_{gi}}\right) \\ &\approx E\left[2^{\lambda_{gi}} + \log_e 2\cdot 2^{\lambda_{gi}}\cdot\left(\log_2 r_{gi}-\lambda_{gi}\right)+\frac{1}{2}\left(\log_e 2\right)^2\cdot 2^{\lambda_{gi}}\cdot\left(\log_2 r_{gi}-\lambda_{gi}\right)^2\right] \\ &= 2^{\lambda_{gi}}\cdot\left\{1+\log_e 2\cdot E\left(\log_2 r_{gi}-\lambda_{gi}\right)+\frac{1}{2}\left(\log_e 2\right)^2\cdot E\left(\log_2 r_{gi}-\lambda_{gi}\right)^2\right\} \\ &= 2^{\lambda_{gi}}\times\left(1+\frac{1}{2}\left(\log_e 2\right)^2 s_{gi}^2\right). \end{aligned}$$

λ_{gi} and s_{gi}^2 are functionally related, and both were estimated with the method used for voom-trend [10] as follows:

i. For all genes, $g = 1, \ldots, G$, 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 $\hat{\lambda}_g = \overline{y}_g + \log_2\tilde{R} - \log_2 10^6$, where $\overline{y}_g$ is the average of y_{gi}; $\tilde{R}$ is the geometric mean of (R_i^*+1); and $g = 1, \ldots, G$.
iii. For $(\hat{\lambda}_g, \hat{s}_g^2), g = 1, ..., G$, 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:

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

Linear mixed-effects model

We denoted a design matrix for nuisance effects, including an intercept by **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:

$$\begin{aligned} \mathbf{Y_g} &= \mathbf{Z}\alpha_g + \mathbf{X}\beta_g + \mathbf{b_g} \\ &\quad + \mathbf{e_g}, \mathbf{b_g}\sim\mathrm{MVN}(\mathbf{0}, \psi_g\mathbf{\Sigma}_{g,\mathbf{b}}), \mathbf{e_g}\sim\mathrm{MVN}(\mathbf{0}, \mathbf{\Sigma}_{g,\mathbf{e}}), \end{aligned}$$

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

Here, $\psi_g\mathbf{\Sigma}_{g,\,\mathbf{b}}$ and $\boldsymbol{\sigma}_{\mathbf{g}}^{\mathbf{2}}$ indicate technical and biological variances, respectively, according to Eq. (2), and the random effect, $\mathbf{b}_g$, and measurement error, $\mathbf{e}_g$, were used to model the technical and biological variances, respectively. The proposed linear mixed effects model may be conceptually useful for understanding the variance structure of the RNA-seq data. Notably, elements of $\mathbf{\Sigma}_{g,\,\mathbf{b}}$ are

obtained from Eq. (3), and ψ_g and $\boldsymbol{\sigma}_{\boldsymbol{g}}^{\mathbf{2}} = (\sigma_{g1}^2, \ldots, \sigma_{gM}^2)$ are estimated. Equation (2) shows that ψ_g becomes 1, and we assumed that $\sigma_{g1}^2 = \ldots = \sigma_{gM}^2$. Then, our final model becomes

$$\mathbf{Y}_{\boldsymbol{g}} = \mathbf{Z}\alpha_{\boldsymbol{g}} + \mathbf{X}\beta_{\boldsymbol{g}} + \mathbf{b}_{\boldsymbol{g}} + \mathbf{e}_{\boldsymbol{g}}, \mathbf{b}_{\boldsymbol{g}} \sim \mathrm{MVN}(\mathbf{0}, \boldsymbol{\Sigma}_{g,\mathbf{b}}), \mathbf{e}_{\boldsymbol{g}} \sim \mathrm{MVN}(\mathbf{0}, \sigma_g^2\mathbf{I}),$$

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

which is equivalent to

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

Bartlett-adjusted profile likelihood ratio tests

The likelihood ratio test (LRT) is very flexible and can be utilized for various purposes such as testing two or more variables at once. However, statistical analyses with RNA-seq data often involve few samples, and the LRT statistic has a bias with order $O_p(N^{-1})$ to its null distribution. In such cases, an adjustment can be applied to reduce the bias, such as the Bartlett adjustment or Cox-Reid adjustment [18, 20]. As Zucker et al. (2000) showed that the latter estimated p values more conservative than the former did in linear mixed models [21], we selected the Bartlett-adjusted LRT for identifying DEGs, which reduces the order of bias to $O_p(N^{-2})$ and controls type-1 error rates well when the sample size is small [18]. If we let $\boldsymbol{\beta}_{\mathbf{g}} = (\beta_{g,1}, \ldots, \beta_{g,M-1})^t$, $\mathbf{V}_{\mathbf{g}} = \boldsymbol{\Sigma}_{g,\mathbf{b}} + \sigma_g^2\mathbf{I}$ and $\boldsymbol{\theta}_{\boldsymbol{g}} = (\boldsymbol{\alpha}_{\boldsymbol{g}}, \sigma_g^2)$, the likelihood for the proposed linear mixed model is

$$L(\theta_{\boldsymbol{g}}, \beta_{\boldsymbol{g}}) \propto |\mathbf{V}_{\mathbf{g}}|^{-\frac{1}{2}} \exp\left(-\frac{1}{2}(\mathbf{Y}_{\boldsymbol{g}} - \mathbf{Z}\alpha_{\boldsymbol{g}} - \mathbf{X}\beta_{\boldsymbol{g}})^t \mathbf{V}_{\mathbf{g}}^{-1}(\mathbf{Y}_{\boldsymbol{g}} - \mathbf{Z}\alpha_{\boldsymbol{g}} - \mathbf{X}\beta_{\boldsymbol{g}})\right).$$

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

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

The Bartlett-adjusted LRT (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) [22], as follows:

$$C_g = \mathrm{D_g}^{-1}\left(-\frac{1}{2}\mathrm{M_g} + \frac{1}{4}\mathrm{P_g} - \frac{1}{2}\nu_{\mathrm{g}}\tau_{\mathrm{g}}\right).$$

Here, $\mathrm{D_g}$, $\mathrm{M_g}$, $\mathrm{P_g}$, ν_{g}, and τ_{g} are scalars, and if we let $\mathbf{X}^{'} = [\mathbf{I} - \mathbf{Z}(\mathbf{Z}^{\mathbf{T}}\mathbf{V_g}^{-1}\mathbf{Z})^{-1}\mathbf{Z}^t\mathbf{V_g}^{-1}]\mathbf{X}$ 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

$$\mathrm{D_g} = -\frac{1}{2}\mathrm{tr}(\mathrm{V_g}^{-2}),$$

$$\mathrm{M_g} = 2\mathrm{tr}\left(\left(\mathrm{X}'^t\mathrm{V_g}^{-1}\mathrm{X}'\right)^{-1}\left(\mathrm{X}'^t\mathrm{V_g}^{-3}\mathrm{X}' - \dot{\mathrm{X}}'^t\mathrm{V_g}^{-2}\mathrm{X}'\right)\right),$$

$$\mathrm{P_g} = \mathrm{tr}\left(\left(\mathrm{X}'^t\mathrm{V_g}^{-2}\mathrm{X}'\left(\mathrm{X}'^t\mathrm{V_g}^{-1}\mathrm{X}'\right)^{-1}\right)^2\right),$$

$$\nu_{\mathrm{g}} = -\mathrm{tr}\left((\mathrm{Z}^t\mathrm{V_g}^{-1}\mathrm{Z})^{-1}\mathrm{Z}^t\mathrm{V_g}^{-2}\mathrm{Z}\right)$$

and

$$\tau_{\mathrm{g}} = -\mathrm{tr}\left(\left(\mathrm{X}'^t\mathrm{V_g}^{-1}\mathrm{X}'\right)^{-1}\mathrm{X}'^t\mathrm{V_g}^{-2}\mathrm{X}'\right).$$

The forms of $\mathrm{D_g}$, $\mathrm{M_g}$, $\mathrm{P_g}$, ν_{g}, and τ_{g} depend on the structure of $\mathbf{V_g}$, and counterparts to general $\mathbf{V_g}$ s are shown in Additional file 1.

Parameter estimation

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

$$\log L(\theta_{\boldsymbol{g}}, \beta_{\boldsymbol{g}}) = \mathrm{C} - \frac{1}{2}\log|\mathbf{V_g}| - \frac{1}{2}(\mathbf{Y_g} - \mathbf{Z}\alpha_{\boldsymbol{g}} - \mathbf{X}\beta_{\boldsymbol{g}})^t\mathbf{V_g}^{-1}(\mathbf{Y_g} - \mathbf{Z}\alpha_{\boldsymbol{g}} - \mathbf{X}\beta_{\boldsymbol{g}}), \mathrm{C}: \text{someconstants.}$$

$\hat{\boldsymbol{\theta}}_{\boldsymbol{g}}$ and $\hat{\boldsymbol{\beta}}_{\boldsymbol{g}}$ 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 σ_g^2 becomes

$$l_P(\sigma_g^2) = \mathrm{C} - \frac{1}{2}\log|\mathbf{V_g}| - \frac{1}{2}\mathbf{Y_g}^t\mathbf{P}\mathbf{Y_g}.$$

Here, $\mathbf{V_g}$ is a function of σ_g^2, and $\hat{\sigma}_g^2$ can be obtained by maximizing $l_P(\sigma_g^2)$ with Fisher's scoring method. $\hat{\sigma}_g^{2(m)}$ at the m step was updated by

$$\hat{\sigma}_g^{2(m)} = \hat{\sigma}_g^{2(m-1)} + \left[I_{o(m-1)}\right]^{-1} U\left(\hat{\sigma}_g^{2(m-1)}\right)$$

, where $I_{o(m-1)} = E(-\frac{\partial^2 l_P}{\partial \hat{\sigma}_g^{2(m-1)} \partial \hat{\sigma}_g^{2(m-1)t}})$ and $U(\hat{\sigma}_g^{2(m-1)}) = \frac{\partial l_P}{\partial \hat{\sigma}_g^{2(m-1)}}$. We found that Fisher's scoring method was sometimes unsuccessful for estimating $\hat{\sigma}_g^{2(m)}$, and in such cases, we used Brent's derivative-free method with the *optimize* function in R [23]. It always converged and successfully estimated parameters, at least in our simulation. We assumed that $\hat{\sigma}_g^2$ was non-negative. $\hat{\mathbf{V}}_\mathbf{g}=\mathbf{\Sigma}_{g,\mathbf{b}} + \hat{\sigma}_g^2\mathbf{I}$, and if $\hat{\sigma}_g^2$ is equal to zero, $\hat{\mathbf{V}}_\mathbf{g}$ becomes $\mathbf{\Sigma}_{g,\,\mathbf{b}}$. Then, $\hat{\beta}_g$ and $\hat{\alpha}_g$ can be obtained by

$$\begin{pmatrix} \hat{\alpha}_g \\ \hat{\beta}_g \end{pmatrix} = ((\mathbf{Z},\mathbf{X})^t \hat{\mathbf{V}}_\mathbf{g}^{-1} \begin{pmatrix} \mathbf{Z},\mathbf{X} \end{pmatrix})(\mathbf{Z},\mathbf{X})^t \hat{\mathbf{V}}_\mathbf{g}^{-1}\mathbf{Y}.$$

The Bartlett-adjusted LRT requires $\hat{\boldsymbol{\theta}}_{g0}$, 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 σ_g^2 becomes

$$l_{P0}(\sigma_g^2) = \mathrm{C} - \frac{1}{2}\log|\mathbf{V_g}| - \frac{1}{2}\mathbf{Y}_g{}^t\mathbf{P_0}\mathbf{Y}_g.$$

$\hat{\sigma}_{g0}^2$ was estimated with Fisher's scoring method, and if we incorporated $\hat{\sigma}_{g0}^2$ to $\mathbf{V_{g0}}$ and denoted it as $\hat{\mathbf{V}}_\mathbf{g0}$, $\hat{\boldsymbol{\alpha}}_\mathbf{g}$ could be obtained by

$$\hat{\alpha}_g = (\mathbf{Z}^t\hat{\mathbf{V}}_\mathbf{g0}{}^{-1}\mathbf{Z})\mathbf{Z}^t\hat{\mathbf{V}}_\mathbf{g0}{}^{-1}\mathbf{Y}.$$

Datasets

We considered two real datasets consisting of samples from unrelated Nigerian people and Holstein cattle, respectively. Nigerian subjects participated in the International HapMap Project and comprised 29 male and 40 female participants [24]. The read counts were downloaded from the ReCount website [25]. Holstein data were obtained to identify genes associated with milk yield and consisted of high- and low-milk-yielding groups, with nine and 12 samples per group, respectively [26]. Furthermore, parity and lactation periods were available and were considered as covariates. We obtained the read counts from Gene Expression Omnibus (GSE60575). Based on count data, we generated simulation data; the steps are described in Additional file 2.

Results

Simulation studies with RNA-seq data from Nigerian individuals

We applied the proposed linear mixed models to the simulated data based on Nigerian individuals' RNA-seq data and calculated empirical type-1 error rates and statistical powers with these models. The data were then compared with DESeq2 (v1.20.0), edgeR (v3.22.3), and voom (v3.36.5). Table 1, Additional file 3, and Fig. 1 show results from simulation studies based on Nigerian individuals' RNA-seq data. Nigerian individuals' 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 50 replicates. Table 1 and Additional file 3 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 Additional file 3, BALLI and voom always controlled the nominal type-1 error rates correctly if N was greater than or equal to 12. LLI also successfully controlled the nominal type-1 error rates if N was greater than or equal to 20. However, if N was less than 20, 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.

Figure 1 shows estimated powers and precisions at the FDR-adjusted 0.1 significance level when $\delta = 0.5\sigma$ or 1σ, and $N = 12$, 16, 20, 24, 28, 40, 64, or 68. Figure 1a and c show the statistical power estimates, and Fig. 1b and d show the precision. Precision indicates the proportions of DEGs among genes for which FDR-adjusted p values are less than 0.1. According to Fig. 1a and c, LLI outperformed other methods in terms of power (Fig. 1a and c). However, it should be noted that this method showed worse precision than BALLI and voom if N was less than 40 (Fig. 1b and d), suggesting that LLI had larger false-positive rates than those of BALLI and voom when N was less than 40. 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 0.597, followed by voom (0.548) and DESeq2 (0.399). The estimated precision of BALLI was 0.940, and those of voom and DESeq2 were 0.930 and 0.907, respectively. If $N = 40$ and $\delta = 0.5\sigma$, the estimated power and precision of BALLI were 0.342 and 0.943, which were higher than those of DESeq2 (0.205, 0.897) and voom (0.298, 0.932).

Our method was also applied to simulation data based on RNA-seq data from Holstein cows. The results were similar to those of the simulation data based on Nigerian people's data. Additional file 4 shows that BALLI and

Table 1 Estimated type-1 error rates with simulation data based on Nigerian people's RNA-seq 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**. [a]The type-1 error rates are marked by bold font if their 95% confidence levels include or lower than the nominal significant level ***α***

	BALLI	DESeq2	edgeR	LLI	voom	BALLI	DESeq2	edgeR	LLI	voom
α	*N* = 12					*N* = 16				
0.1	**0.09217[a] (0.08173,0.10262)**	**0.08221 (0.07144,0.09298)**	**0.10536 (0.09595,0.11477)**	0.12695 (0.11443,0.13947)	**0.09381 (0.08306,0.10456)**	**0.08976 (0.08033,0.09919)**	**0.08794 (0.07827,0.09761)**	0.11070 (0.10223,0.11916)	0.11613 (0.10522,0.12704)	**0.09618 (0.08683,0.10552)**
0.05	**0.04485 (0.03814,0.05155)**	**0.04374 (0.03633,0.05115)**	**0.05495 (0.04877,0.06114)**	0.06636 (0.05767,0.07506)	**0.04473 (0.03798,0.05148)**	**0.04329 (0.03764,0.04895)**	**0.04672 (0.04041,0.05302)**	0.05762 (0.05218,0.06305)	0.06072 (0.05355,0.06790)	**0.04645 (0.04078,0.05212)**
0.01	**0.00899 (0.00686,0.01112)**	**0.01208 (0.00913,0.01503)**	0.01441 (0.01186,0.01697)	0.01662 (0.01328,0.01996)	**0.00860 (0.00653,0.01066)**	**0.00796 (0.00643,0.00948)**	**0.01202 (0.00980,0.01424)**	0.01393 (0.01212,0.01575)	0.01349 (0.01118,0.01580)	**0.00850 (0.00706,0.00993)**
0.005	**0.00456 (0.00328,0.00584)**	0.00741 (0.00531,0.00952)	0.00891 (0.00709,0.01072)	0.00942 (0.00720,0.01163)	**0.00424 (0.00301,0.00548)**	**0.00381 (0.00305,0.00458)**	0.00695 (0.00555,0.00835)	0.00809 (0.00692,0.00925)	0.00714 (0.00575,0.00853)	**0.00397 (0.00326,0.00468)**
α	*N* = 20					*N* = 24				
0.1	**0.08698 (0.07689,0.09707)**	**0.08700 (0.07653,0.09747)**	0.10962 (0.10028,0.11896)	**0.10564 (0.09442,0.11686)**	**0.09289 (0.08246,0.10332)**	**0.08480 (0.07651,0.09309)**	**0.08744 (0.07802,0.09686)**	0.10874 (0.10024,0.11724)	**0.10067 (0.09133,0.11000)**	**0.09124 (0.08196,0.10051)**
0.05	**0.04113 (0.03455,0.04771)**	**0.04607 (0.03870,0.05344)**	0.05774 (0.05147,0.06401)	**0.05408 (0.04633,0.06183)**	**0.04517 (0.03844,0.05191)**	**0.03919 (0.03450,0.04387)**	**0.04518 (0.03922,0.05114)**	0.05881 (0.05343,0.06419)	**0.04960 (0.04402,0.05519)**	**0.04390 (0.03849,0.04931)**
0.01	**0.00752 (0.00548,0.00955)**	**0.01175 (0.00890,0.01460)**	0.01381 (0.01152,0.01609)	**0.01157 (0.00879,0.01436)**	**0.00848 (0.00643,0.01053)**	**0.00647 (0.00550,0.00744)**	**0.01050 (0.00873,0.01228)**	0.01317 (0.01163,0.01472)	**0.00955 (0.00817,0.01093)**	**0.00778 (0.00648,0.00907)**
0.005	**0.00368 (0.00248,0.00489)**	**0.00669 (0.00475,0.00863)**	0.00789 (0.00643,0.00934)	**0.00605 (0.00428,0.00783)**	**0.00420 (0.00305,0.00535)**	**0.00293 (0.00248,0.00338)**	**0.00575 (0.00477,0.00672)**	0.00728 (0.00636,0.00819)	**0.00474 (0.00400,0.00547)**	**0.00363 (0.00295,0.00431)**

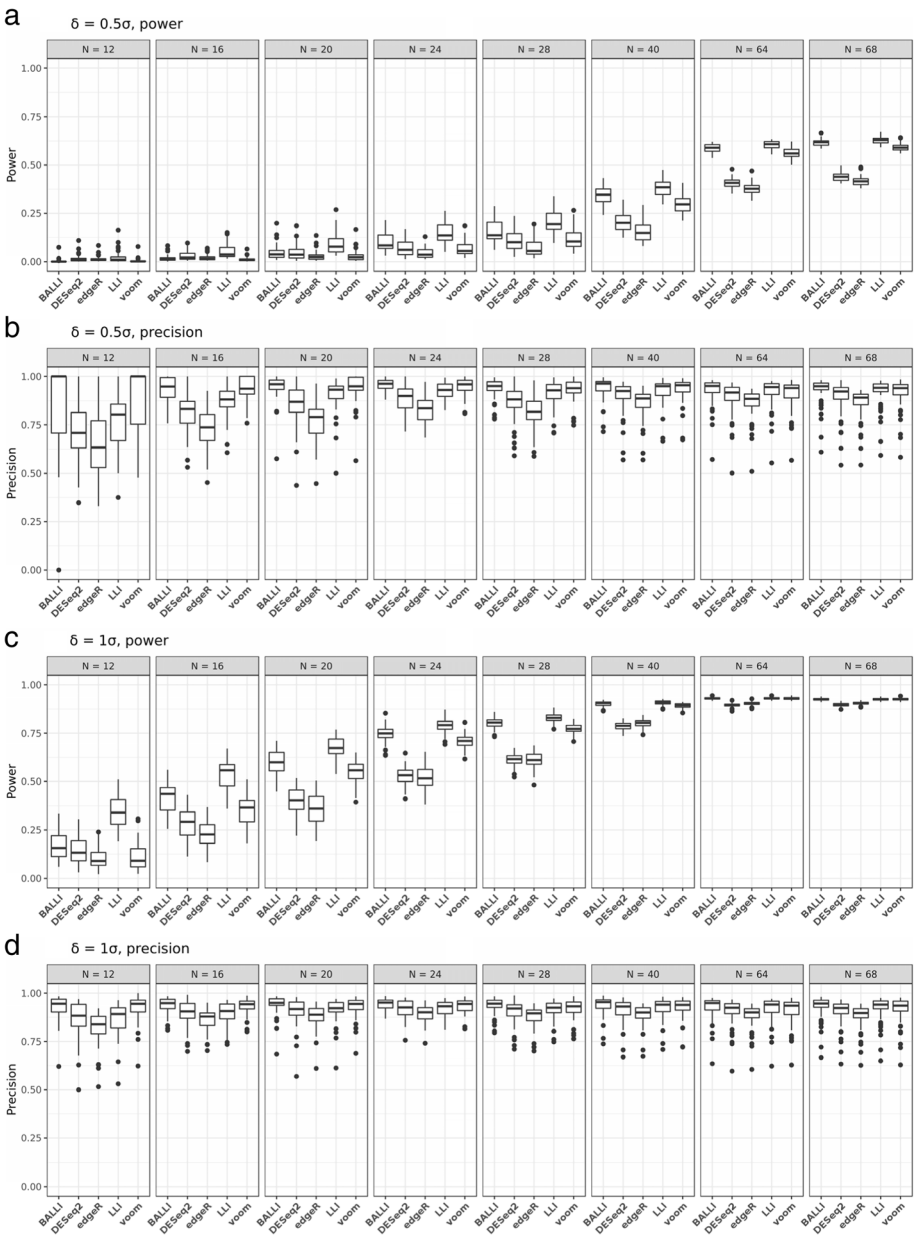

Fig. 1 Estimated powers and precisions with simulation data based on Nigerian people's RNA-seq data. Statistical powers of BALLI, DESeq2, edgeR, LLI, and voom were estimated at FDR-adjusted 0.1 significance level when $\delta = 0.5\sigma$ or 1σ and N = 12, 16, 20, 24, 28, 40, 64, and 68. **a** Estimated power when δ=0.5σ. **b** Estimated precision when δ=0.5σ. **c** Estimated power when δ=1σ. **d** Estimated precision when δ=1σ

LLI controlled the nominal type-1 error rates if $N \geq 8$ and if $N \geq 20$, respectively. Power estimates were highest for LLI, but it always had a smaller precision value than those of BALLI and voom (Additional file 5).

Simulation studies with randomly generated RNA-seq data

RNA-seq data are generally known to follow 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 real RNA-seq data. Estimated type-1 error rates by voom and BALLI usually maintained the nominal significance levels (Table 2 and Additional file 6). *P* values obtained from LLI and edgeR tended to be inflated. DESeq2 generally showed deflation of type-1 error rates at a 0.1 nominal significance level. Second, we

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**. [a]The type-1 error rates are marked by bold font if their 95% confidence levels include or lower than the nominal significant level ***α***

	BALLI	DESeq2	edgeR	LLI	voom	BALLI	DESeq2	edgeR	LLI	voom
α	*N* = 12					*N* = 16				
0.1	**0.09550[a] (0.09394, 0.09706)**	**0.08417 (0.08278, 0.08556)**	0.12172 (0.11984, 0.12360)	0.13198 (0.13030, 0.13367)	**0.09912 (0.09767, 0.10057)**	**0.09211 (0.09097, 0.09325)**	**0.08428 (0.08307, 0.08549)**	0.12028 (0.11911, 0.12145)	0.11853 (0.11729, 0.11977)	**0.09977 (0.09868, 0.10086)**
0.05	**0.04863 (0.04757, 0.04969)**	**0.04334 (0.04243, 0.04425)**	0.05604 (0.05492, 0.05716)	0.07063 (0.06926, 0.07200)	**0.05013 (0.04899, 0.05126)**	**0.04508 (0.04429, 0.04587)**	**0.04333 (0.04261, 0.04404)**	0.05616 (0.05528, 0.05703)	0.06345 (0.06239, 0.06452)	**0.05013 (0.04937, 0.05089)**
0.01	**0.00989 (0.00939, 0.01039)**	**0.00997 (0.00954, 0.01040)**	0.01267 (0.01209, 0.01326)	0.01858 (0.01798, 0.01919)	**0.00997 (0.00957, 0.01037)**	**0.00893 (0.00847, 0.00939)**	**0.01008 (0.00961, 0.01056)**	0.01276 (0.01225, 0.01328)	0.01469 (0.01419, 0.01519)	**0.00996 (0.00969, 0.01024)**
0.005	**0.00524 (0.00492, 0.00557)**	0.00549 (0.00517, 0.00581)	0.00689 (0.00650, 0.00729)	0.01043 (0.00992, 0.01094)	**0.00488 (0.00456, 0.00519)**	**0.00445 (0.00414, 0.00475)**	0.00543 (0.00515, 0.00572)	0.00682 (0.00653, 0.00711)	0.00803 (0.00759, 0.00846)	**0.00506 (0.00483, 0.00528)**
α	*N* = 20					*N* = 24				
0.1	**0.09198 (0.09051, 0.09346)**	**0.08419 (0.08282, 0.08557)**	0.11844 (0.11718, 0.11971)	0.11073 (0.10925, 0.11221)	**0.10023 (0.09908, 0.10139)**	**0.09276 (0.09156, 0.09395)**	**0.08521 (0.08386, 0.08657)**	0.11846 (0.11686, 0.12005)	0.10867 (0.10747, 0.10987)	**0.09963 (0.09833, 0.10092)**
0.05	**0.04463 (0.04376, 0.04549)**	**0.04296 (0.04205, 0.04387)**	0.05748 (0.05625, 0.05871)	0.05780 (0.05662, 0.05898)	**0.05048 (0.04936, 0.05159)**	**0.04468 (0.04394, 0.04541)**	**0.04268 (0.04188, 0.04348)**	0.06050 (0.05953, 0.06147)	0.05518 (0.05427, 0.05608)	**0.04907 (0.04815, 0.05000)**
0.01	**0.00861 (0.00812, 0.00911)**	**0.00986 (0.00936, 0.01036)**	0.01271 (0.01204, 0.01338)	0.01328 (0.01267, 0.01388)	**0.01016 (0.00954, 0.01077)**	**0.00844 (0.00810, 0.00878)**	**0.00940 (0.00899, 0.00982)**	0.01242 (0.01203, 0.01281)	0.01212 (0.01167, 0.01257)	**0.00957 (0.00907, 0.01007)**
0.005	**0.00457 (0.00425, 0.00489)**	0.00536 (0.00501, 0.00572)	0.00688 (0.00647, 0.00730)	0.00707 (0.00661, 0.00753)	**0.00521 (0.00484, 0.00559)**	**0.00405 (0.00386, 0.00425)**	**0.00497 (0.00469, 0.00525)**	0.00662 (0.00639, 0.00685)	0.00625 (0.00595, 0.00655)	**0.00474 (0.00440, 0.00507)**

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 $\left(0.2 + a_{gi}^{-1/2}\right)^2 \times \delta_g$, where $40/\delta_g \sim \chi^2_{40}$. 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 Additional file 7, and the general pattern was similar to that in Fig. 2, except that DESeq2 was conservative at a significance level of 0.05 and liberal at 0.005 (Additional file 7). Therefore, we could conclude that the performances of BALLI and LLI were robust, and we recommend using BALLI if $10 \le N \le 40$, and LLI if $N > 40$.

Figure 3 and Additional file 8 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 real RNA-seq data. For example, when $u = 0.2$, $N = 28$, and $\delta = 0.5\sigma$, the estimated power by BALLI was 0.438, whereas those for DESeq2 and voom were 0.317 and 0.352, respectively (Fig. 3a). Results when $u = 0.2$ and $\delta = 1\sigma$ in Fig. 3c are very similar as those for Fig. 3a. Figure 3b and d also show that BALLI achieves the best estimated precisions. Similar patterns were observed when $u = 0.4$, 0.6, 0.8, and 1 (Additional file 8). In summary, we can conclude that BALLI shows better performance than that of other methods.

DEGs of Holstein milk data

Holstein milk data of 21 Holstein cows were generated to detect genes related to the daily productivity of milk. High and low milk yields were considered the primary exposure variables, and parity and lactation period were included as covariates, as in the original study [26]. In this study, 12 tentative DEGs were chosen and

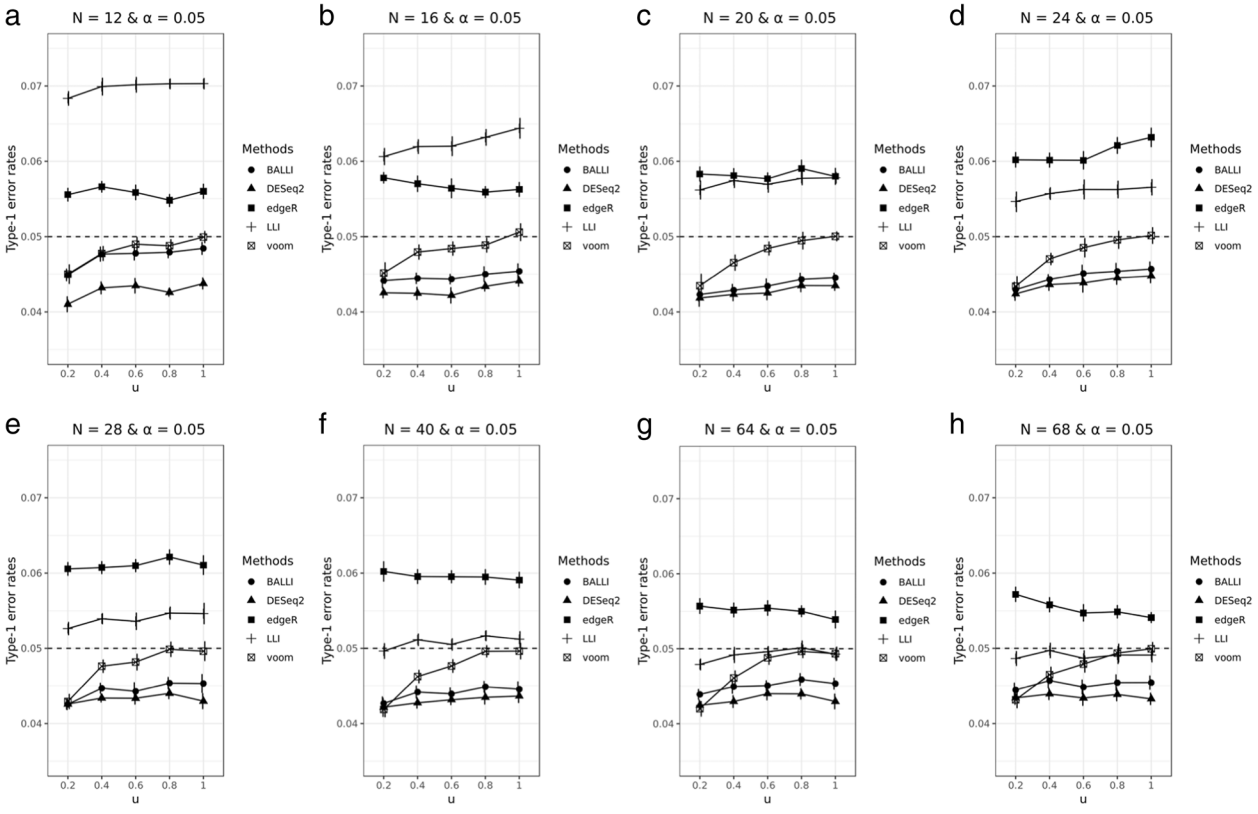

Fig. 2 Effect of varying library sizes on the type-1 error rates. Type-1 error rates were estimated by BALLI, DESeq2, edgeR, LLI, and voom when u = 0.2, 0.4, 0.6, 0.8, and 1 and sample size (N) is 12 (**a**), 16 (**b**), 20 (**c**), 24 (**d**), 28 (**e**), 40 (**f**), 64 (**g**), and 68 (**h**) at the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

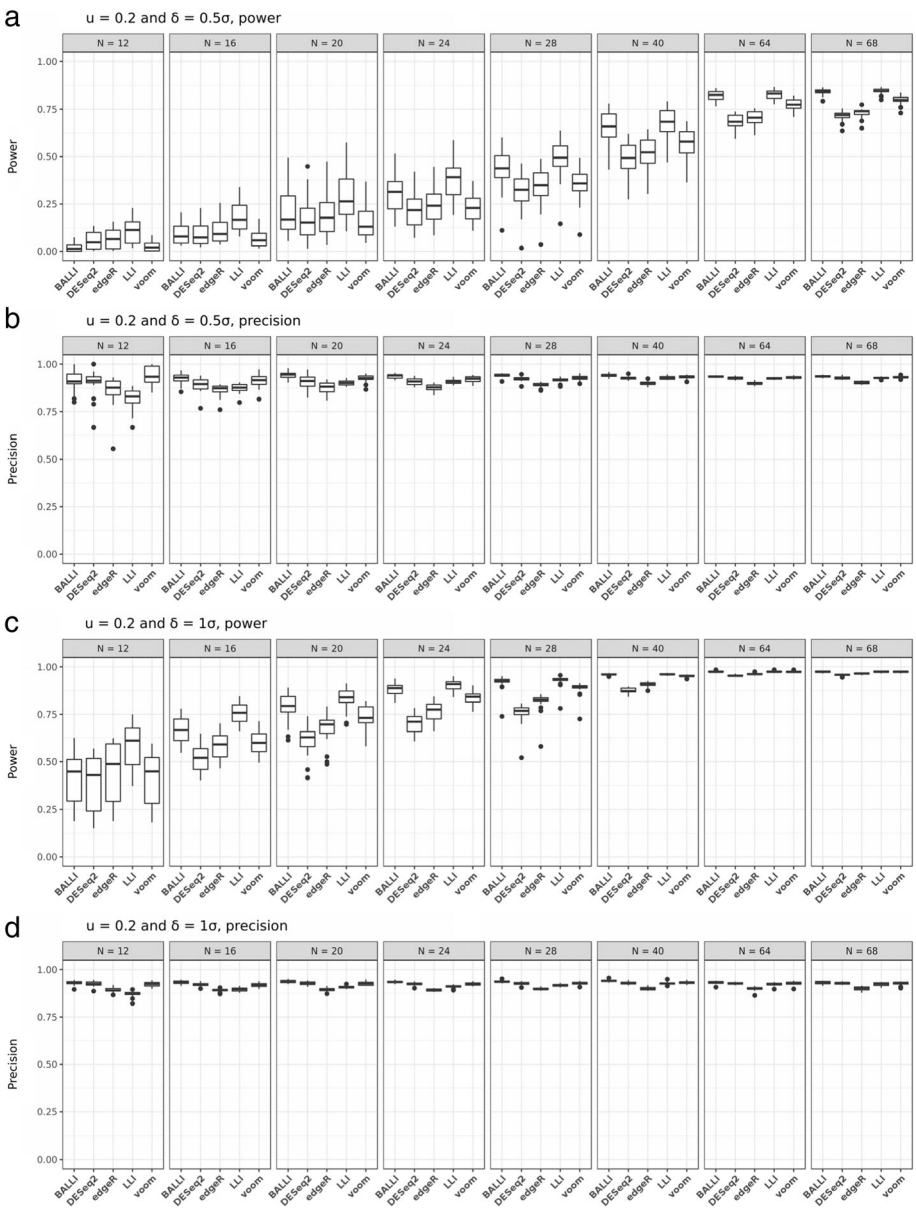

Fig. 3 Effect of varying library sizes on the statistical power and precision. Statistical powers and precisions for BALLI, DESeq2, edgeR, LLI, and voom were empirically estimated at FDR-adjusted 0.1 significance level when u = 0.2, δ = 0.5σ or 1σ and sample size (N) is 12, 16, 20, 24, 28, 40, 64, and 68. **a** Estimated power when u=0.2 and δ=0.5σ. **b** Estimated precision when u=0.2 and δ=0.5σ. **c** Estimated power when u=0.2 and δ=1σ. **d** Estimated precision when u=0.2 and δ=1σ

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 12 tentative genes were true DEGs. Among the 12 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 an FDR 0.1 significance level by DESeq2, voom, and their methods owing to the lack of statistical power [26]. 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 0.1267, which was much smaller than those of DESeq2, edgeR, and voom. Table 3 shows *p* values for the nine genes, including *TOX4*. *P*

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	1.058×10^{-5} (1.267×10^{-1})	2.949×10^{-4} (6.298×10^{-1})	2.797×10^{-2} (1)	7.476×10^{-7} (8.947×10^{-3})	4.980×10^{-3} (9.999×10^{-1})
HNRNPL	3.913×10^{-4} (9.222×10^{-1})	1.551×10^{-3} (9.997×10^{-1})	1.684×10^{-1} (1)	6.796×10^{-5} (1.787×10^{-1})	4.677×10^{-2} (9.999×10^{-1})
SPTSSB	4.457×10^{-4} (9.222×10^{-1})	2.660×10^{-3} (9.997×10^{-1})	2.160×10^{-4} (1)	8.686×10^{-5} (1.787×10^{-1})	1.037×10^{-3} (9.999×10^{-1})
NOS3	4.676×10^{-4} (9.222×10^{-1})	2.396×10^{-4} (6.928×10^{-1})	2.549×10^{-4} (1)	8.957×10^{-5} (1.787×10^{-1})	8.693×10^{-4} (9.999×10^{-1})
SLC4A1	2.145×10^{-2} (9.999×10^{-1})	8.753×10^{-2} (9.997×10^{-1})	1.025×10^{-1} (1)	9.856×10^{-3} (7.179×10^{-1})	7.579×10^{-2} (9.999×10^{-1})
NLN	9.513×10^{-2} (9.999×10^{-1})	3.028×10^{-1} (9.997×10^{-1})	3.789×10^{-1} (1)	6.084×10^{-2} (9.774×10^{-1})	1.214×10^{-1} (9.999×10^{-1})
KALRN	9.792×10^{-2} (9.999×10^{-1})	8.943×10^{-2} (9.997×10^{-1})	8.815×10^{-2} (1)	6.307×10^{-2} (9.790×10^{-1})	1.054×10^{-1} (9.999×10^{-1})
PMCH	1.635×10^{-1} (9.999×10^{-1})	2.225×10^{-1} (9.997×10^{-1})	2.353×10^{-1} (1)	1.176×10^{-1} (9.999×10^{-1})	1.758×10^{-1} (9.999×10^{-1})
C25H16orf88	1.765×10^{-1} (9.999×10^{-1})	1.494×10^{-1} (9.997×10^{-1})	2.627×10^{-1} (1)	1.289×10^{-1} (9.999×10^{-1})	1.516×10^{-1} (9.999×10^{-1})

values for most of the nine genes obtained by LLI and BALLI were smaller than those obtained by other methods. The nine genes were not significantly affected by parity or lactation period (Additional file 9). We also analyzed all genes by the proposed methods; Fig. 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). Three genes, including *HNRNPL*, were detected as DEGs by only BALLI and LLI (Fig. 4). Table 4 shows 12 genes that were commonly significant by BALLI, DESeq2, edgeR, LLI, and voom at the 0.005 nominal significance level. Of the 12 genes, all genes except *C4BPA* had the lowest *p* values in LLI, and three genes had lower *p* values in BALLI than in DESeq2, edgeR, and voom. This analysis with BALLI took 153.35 s using an Intel Xeon E7–4820 2.00 GHz processor, which was quite a bit longer than other methods, voom (2.81 s), DESeq2 (27.21 s) and edgeR (33.25 s) (Additional file 10). However, BALLI can conduct the multi-threaded analyses with a simple option unlike other methods and analyses of whole genes can be completed within a reasonable time. For example, the analyses with 20 cores took only 11.99 s under the same conditions (Additional file 10). Simulation studies revealed that LLI tended to be liberal, with the results

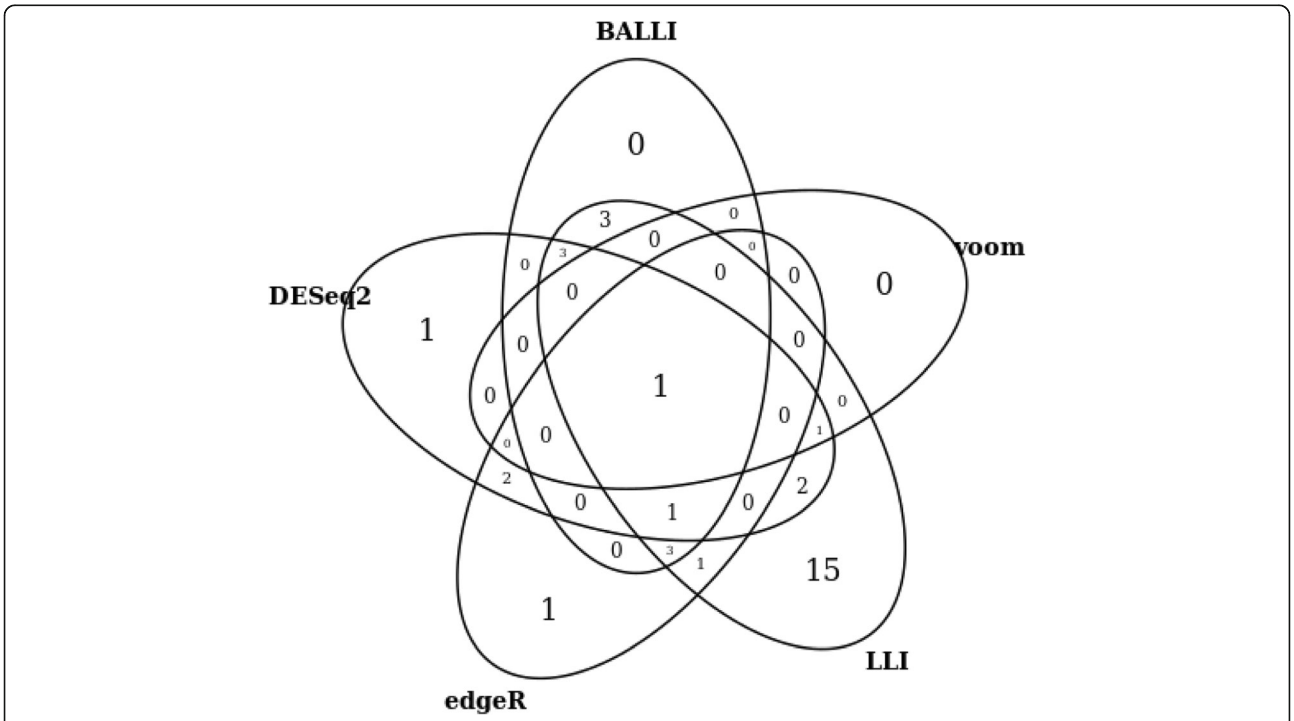

Fig. 4 Significant genes of Holstein milk data. Venn diagram was provided with significant genes at the 0.001 nominal significance level by BALLI, DESeq2, edgeR, LLI, and voom

Table 4 Significant genes in all methods of Holstein milk data. Gene lists of Holstein milk data siginificant 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
SPTSSB	4.457×10^{-4} (9.222×10^{-1})	2.660×10^{-3} (9.997×10^{-1})	2.160×10^{-4} (1)	8.686×10^{-5} (1.787×10^{-1})	1.037×10^{-3} (9.999×10^{-1})
NOS3	4.676×10^{-4} (9.222×10^{-1})	2.396×10^{-4} (6.298×10^{-1})	2.549×10^{-4} (1)	8.957×10^{-5} (1.787×10^{-1})	8.693×10^{-4} (9.999×10^{-1})
FXYD3	7.198×10^{-4} (9.507×10^{-1})	2.813×10^{-4} (6.298×10^{-1})	5.182×10^{-4} (1)	1.513×10^{-4} (2.012×10^{-1})	2.097×10^{-3} (9.999×10^{-1})
SPESP1	1.103×10^{-3} (9.507×10^{-1})	4.834×10^{-3} (9.997×10^{-1})	1.669×10^{-3} (1)	2.579×10^{-4} (2.385×10^{-1})	4.001×10^{-3} (9.999×10^{-1})
CHST1	1.339×10^{-3} (9.507×10^{-1})	8.624×10^{-4} (9.383×10^{-1})	1.872×10^{-3} (1)	3.166×10^{-4} (2.385×10^{-1})	6.108×10^{-4} (9.999×10^{-1})
LEPREL1	1.387×10^{-3} (9.507×10^{-1})	2.554×10^{-3} (9.997×10^{-1})	3.297×10^{-3} (1)	3.334×10^{-4} (2.385×10^{-1})	1.487×10^{-3} (9.999×10^{-1})
JUB	1.486×10^{-3} (9.507×10^{-1})	1.755×10^{-3} (9.997×10^{-1})	2.445×10^{-3} (1)	3.674×10^{-4} (2.385×10^{-1})	2.804×10^{-3} (9.999×10^{-1})
MIA	1.509×10^{-3} (9.507×10^{-1})	4.210×10^{-4} (6.298×10^{-1})	2.247×10^{-3} (1)	3.666×10^{-4} (2.385×10^{-1})	2.432×10^{-3} (9.999×10^{-1})
C4BPA	2.241×10^{-3} (9.999×10^{-1})	3.190×10^{-4} (6.298×10^{-1})	1.307×10^{-3} (1)	6.000×10^{-4} (2.951×10^{-1})	2.866×10^{-3} (9.999×10^{-1})
CLDN6	2.653×10^{-3} (9.999×10^{-1})	1.282×10^{-3} (9.997×10^{-1})	2.414×10^{-3} (1)	7.377×10^{-4} (3.270×10^{-1})	1.542×10^{-3} (9.999×10^{-1})
PALMD	3.737×10^{-3} (9.999×10^{-1})	1.151×10^{-3} (9.997×10^{-1})	2.648×10^{-3} (1)	1.127×10^{-3} (3.776×10^{-1})	3.032×10^{-3} (9.999×10^{-1})
KLK12	3.840×10^{-3} (9.999×10^{-1})	2.766×10^{-3} (9.997×10^{-1})	9.826×10^{-4} (1)	1.225×10^{-3} (3.776×10^{-1})	3.162×10^{-3} (9.999×10^{-1})

likely to 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.

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 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 $1/\mu + \phi$, ϕ obviously indicates the overdispersion parameter, and biological and technical variances can be estimated with BALLI. However, voom assumes ϕ/μ, 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 $\phi/\mu + \phi$ and had the most generalized variance parameter space. Incorporation of $\phi = 1$ 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 ϕ to be larger than 1, and thus, a better model may differ according to RNA-seq data. ϕ and ϕ 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 LRTs. The selected models can then be utilized to identify DEGs. This model was implemented as an R package and can be downloaded from CRAN (http://cran.r-project.org) or http://healthstat.snu.ac.kr/software/balli/. Furthermore, in contrast to methods based on a generalized linear mixed model such as MACAU [27], the proposed methods can be easily extended to various scenarios with 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 or Poisson 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, Bartlett's correction seems inappropriate when $N \leq 6$ (Additional files 3, 4, and 6) and the corrected LR statistic was smaller than expected in some cases, especially in simulation studies based on a negative binomial distribution (Table 2). Further studies are necessary to adjust this. Additionally, results from simulation studies obviously depended on various factors. Our results were obtained from simulation data based on RNA-seq data from Nigerian individuals and Holstein cows 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.

Conclusions

In this article, we proposed likelihood-based linear mixed model approaches with and without Bartlett's correction to analyze more complicated RNA-seq data. The proposed methods consider log-cpm values of genes as response variables, and technical and biological variances are estimated with a linear mixed model. According to our simulation studies and real data analysis, our methods are statistically more efficient than existing methods and correctly control the type-1 error rates. We found that the statistical performance of our method, BALLI and LLI, depends on the sample size; we recommend using LLI if the sample size is larger than 40 and otherwise using BALLI.

Additional files

Additional file 1: The forms of C_g's components depending on the structure of V_g. (DOCX 17 kb)

Additional file 2: Steps of generating simulation data. (DOCX 18 kb)

Additional file 3: Estimated type-1 error rates with simulation data for *N* = 4, 6, 8, 28, 40, 64, and 68 based on Nigerian people's data. (DOCX 21 kb)

Additional file 4: Estimated type-1 error rates with simulation data for *N* = 4, 6, 8, 12, 16 and 20 based on Holstein cow's data. (DOCX 20 kb)

Additional file 5: Estimated powers and precisions with simulation data when $\delta = 0.5\sigma$ or 1σ and *N* = 12, 16 and 20 based on Holstein cow's data. (DOCX 197 kb)

Additional file 6: Estimated type-1 error rates with simulation data for *N* = 4, 6, 8, 28, 40, 64, and 68 based on simulated data from negative binomial distribution. (DOCX 21 kb)

Additional file 7: Effect of varying library sizes on the type-1 error rates when u = 0.2, 0.4, 0.6, 0.8, and 1 and *N* = 12, 16, 20, 24, 28, 40, 64, or 68 at the 0.005 nominal significance level. (DOCX 371 kb)

Additional file 8: Effect of varying library sizes on the statistical power and precision when u = 1, 0.8, 0.6, or 0.4, $\delta = 1\sigma$ and *N* = 12, 16, 20, 24, 28, 40, 64, or 68. (DOCX 593 kb)

Additional file 9: Analysis of covariables, parity and lactation period, for true nine DEGs of Holstein cow's data. (DOCX 18 kb)

Additional file 10: Computation time for each method when analyzing Holstein cow's data. (DOCX 14 kb)

Abbreviations

BALLI: Bartlett-Adjusted Likelihood based LInear mixed model approach; CPM: Counts Per Million; DEGs: Differentially Expressed Genes; FDR: False Discovery Rate; Log-CPM: Log-transformed CPM; LOWESS: Locally WEighted Scatterplot Smoothing; LRT: Likelihood Ratio Test; MLE: Maximum Likelihood Estimate; qRT-PCR: quantitative Real Time Polymerase Chain Reaction; RNA-seq: RNA sequencing

Acknowledgements

Sungho Won and Taesung Park contributed equally to this work as corresponding authors.

Authors' contributions

KP, TP, and SW designed and led the study, and wrote the manuscript. KP developed BALLI R package. JA, JG, MS, and WL advised on statistical methods and data analysis. All authors have read and approved the final version of the manuscript.

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 (HI16C2037) and the National Research Foundation of Korea (2017M3A9F3046543). The funding body was not involved in the study design, data collection, analysis and interpretation, and writing the manuscript.

Availability of data and materials

We downloaded Nigerian samples from ReCount website (http://bowtie-bio.sourceforge.net/recount/countTables/montpick_count_table.txt) and Holstein samples from Gene Expression Omnibus (GSE60575).

Ethics approval and consent to participate

Data used in this article were publicly available.

Consent for publication

Not applicable.

Competing interests
The authors declare that they have no competing interests.

Author details
[1]Interdisciplinary Program of Bioinformatics, Seoul National University, Seoul 08826, South Korea. [2]Department of Public Health Science, Seoul National University, Seoul 08826, South Korea. [3]Department of Biomedical Science, Chosun University, Gwangju 61452, South Korea. [4]Channing Division of Network Medicine, Brigham and Women's Hospital, Boston, MA, USA. [5]Department of Medicine, Harvard Medical School, Boston, MA, USA. [6]Department of Statistics, Inha University, Incheon 22212, South Korea. [7]Department of Statistics, Seoul National University, Seoul 08826, South Korea. [8]Institute of Health and Environment, Seoul National University, Seoul 08826, South Korea.

Received: 23 October 2018 Accepted: 28 May 2019
Published online: 02 July 2019

References

1. Okoniewski MJ, Miller CJ. Hybridization interactions between probesets in short oligo microarrays lead to spurious correlations. BMC Bioinf. 2006;7:276.
2. Dobbin KK, Kawasaki ES, Petersen DW, Simon RM. Characterizing dye bias in microarray experiments. Bioinformatics. 2005;21(10):2430–7.
3. Zhao S, Fung-Leung W-P, Bittner A, Ngo K, Liu X. Comparison of RNA-Seq and microarray in transcriptome profiling of activated T cells. PLoS One. 2014;9(1):e78644.
4. Mortazavi A, Williams BA, McCue K, Schaeffer L, Wold B. Mapping and quantifying mammalian transcriptomes by RNA-Seq. Nat Methods. 2008;5(7):621–8.
5. Dillies MA, Rau A, Aubert J, Hennequet-Antier C, Jeanmougin M, Servant N, Keime C, Marot G, Castel D, Estelle J, et al. A comprehensive evaluation of normalization methods for Illumina high-throughput RNA sequencing data analysis. Brief Bioinform. 2013;14(6):671–83.
6. Marioni JC, Mason CE, Mane SM, Stephens M, Gilad Y. RNA-seq: an assessment of technical reproducibility and comparison with gene expression arrays. Genome Res. 2008;18(9):1509–17.
7. Bullard JH, Purdom E, Hansen KD, Dudoit S. Evaluation of statistical methods for normalization and differential expression in mRNA-Seq experiments. BMC Bioinf. 2010;11(1):94.
8. Robinson MD, McCarthy DJ, Smyth GK. edgeR: a Bioconductor package for differential expression analysis of digital gene expression data. Bioinformatics. 2010;26(1):139–40.
9. Love MI, Huber W, Anders S. Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. Genome Biol. 2014;15(12):550.
10. Law CW, Chen Y, Shi W, Smyth GK. Voom: precision weights unlock linear model analysis tools for RNA-seq read counts. Genome Biol. 2014;15(2):R29.
11. Robinson MD, Smyth GK. Moderated statistical tests for assessing differences in tag abundance. Bioinformatics. 2007;23(21):2881–7.
12. Tusher VG, Tibshirani R, Chu G. Significance analysis of microarrays applied to the ionizing radiation response. Proc Natl Acad Sci U S A. 2001;98(9):5116–21.
13. McCarthy DJ, Chen Y, Smyth GK. Differential expression analysis of multifactor RNA-Seq experiments with respect to biological variation. Nucleic Acids Res. 2012;40(10):4288–97.
14. Litière S, Alonso A, Molenberghs G. The impact of a misspecified random-effects distribution on the estimation and the performance of inferential procedures in generalized linear mixed models. Stat Med. 2008;27(16):3125–44.
15. Chavance M, Escolano S. Misspecification of the covariance structure in generalized linear mixed models. Stat Methods Med Res. 2016;25(2):630–43.
16. Seyednasrollah F, Laiho A, Elo LL. Comparison of software packages for detecting differential expression in RNA-seq studies. Brief Bioinform. 2015;16(1):59–70.
17. Soneson C, Delorenzi M. A comparison of methods for differential expression analysis of RNA-seq data. BMC Bioinf. 2013;14:91.
18. Bartlett MS. Properties of sufficiency and statistical tests. Proc R Soc Lond A Math Phys Sci. 1937;160(901):268–82.
19. Robinson MD, Oshlack A. A scaling normalization method for differential expression analysis of RNA-seq data. Genome Biol. 2010;11(3):R25.
20. Cox DR, Reid N. Parameter orthogonality and approximate conditional inference. J R Stat Soc Ser B Methodol. 1987;49(1):1–18.
21. Zucker DM, Lieberman O, Manor O. Improved small sample inference in the mixed linear model: Bartlett correction and adjusted likelihood. J R Stat Soc Series B Stat Methodol. 2000;62(4):827–38.
22. Melo TF, Ferrari SL, Cribari-Neto F. Improved testing inference in mixed linear models. Comput Stat Data Anal. 2009;53(7):2573–82.
23. Brent RP. Algorithms for minimization without derivatives; 1973.
24. Pickrell JK, Marioni JC, Pai AA, Degner JF, Engelhardt BE, Nkadori E, Veyrieras JB, Stephens M, Gilad Y, Pritchard JK. Understanding mechanisms underlying human gene expression variation with RNA sequencing. Nature. 2010;464(7289):768–72.
25. Frazee AC, Langmead B, Leek JT. ReCount: a multi-experiment resource of analysis-ready RNA-seq gene count datasets. BMC Bioinf. 2011;12:449.
26. Seo M, Kim K, Yoon J, Jeong JY, Lee H-J, Cho S, Kim H. RNA-seq analysis for detecting quantitative trait-associated genes. Sci Rep. 2016;6:24375.
27. Sun S, Hood M, Scott L, Peng Q, Mukherjee S, Tung J, Zhou X. Differential expression analysis for RNAseq using Poisson mixed models. Nucleic Acids Res. 2017;45(11):e106.

Publisher's Note