Kang *et al. BMC Bioinformatics* (2016) 17:248
DOI 10.1186/s12859-016-1111-9

BMC Bioinformatics

METHODOLOGY ARTICLE Open Access

multiDE: a dimension reduced model based statistical method for differential expression analysis using RNA-sequencing data with multiple treatment conditions

Guangliang Kang[1†], Li Du[1†] and Hong Zhang[1,2*]

Abstract

Background: The growing complexity of biological experiment design based on high-throughput RNA sequencing (RNA-seq) is calling for more accommodative statistical tools. We focus on differential expression (DE) analysis using RNA-seq data in the presence of multiple treatment conditions.

Results: We propose a novel method, multiDE, for facilitating DE analysis using RNA-seq read count data with multiple treatment conditions. The read count is assumed to follow a log-linear model incorporating two factors (i.e., condition and gene), where an interaction term is used to quantify the association between gene and condition. The number of the degrees of freedom is reduced to one through the first order decomposition of the interaction, leading to a dramatically power improvement in testing DE genes when the number of conditions is greater than two. In our simulation situations, multiDE outperformed the benchmark methods (i.e. edgeR and DESeq2) even if the underlying model was severely misspecified, and the power gain was increasing in the number of conditions. In the application to two real datasets, multiDE identified more biologically meaningful DE genes than the benchmark methods. An R package implementing multiDE is available publicly at http://homepage.fudan.edu.cn/zhangh/softwares/multiDE.

Conclusions: When the number of conditions is two, multiDE performs comparably with the benchmark methods. When the number of conditions is greater than two, multiDE outperforms the benchmark methods.

Keywords: RNA-seq, Differential expression, Multiple conditions

Background

High-throughput RNA sequencing (RNA-seq) technologies are emerging rapidly in recent years, which are widely used in biological studies [1]. One of the most important biological problems is to identify differentially expressed genes between multiple experimental conditions. As a result, the key of analyzing these data lies in establishing an appropriate statistical model for RNA-seq count data and consequently preforming differential expression (DE) analysis. Many statistical methods have been developed to fulfill this commission [2, 3]. Methods such as DEGSeq assume that the read count follows a Poisson distribution [4]. Nevertheless, this assumption is violated in the presence of variance overdispersion due to technological and biological variations. An additional parameter can be used to account for the overdispersion. DESeq2 and edgeR, for example, both assume a negative binomial distribution [5, 6]. Much more efficient estimates of the dispersion parameters are obtained using an empirical Bayesian method, resulting in a considerable improvement of DE analysis power and a better control of false positive rate if the sample size is very small. These two methods have been implemented in the Bioconductor packages edgeR and DESeq2, respectively.

In many biological studies, biologists are interested in identifying those genes differently expressed between

*Correspondence: zhanghfd@fudan.edu.cn
†Equal contributors
[1] Institute of Biostatistics, School of Life Sciences, Fudan University, 2005 Songhu Road, 200438 Shanghai, People's Republic of China
[2] State Key Laboratory of Genetic Engineering, School of Life Sciences, Fudan University, 2005 Songhu Road, 200438 Shanghai, People's Republic of China

multiple conditions. For example, the transcriptome of several stages of embryo development or several subtypes of cancer simultaneously enable them to get insight into the sophisticated biological mechanism on a much more comprehensive point [7, 8]. Moreover, the read counts of various conditions can be correlated with each other if they are generated from the same subject. This further complicates the statistical analysis. When the number of conditions, D, is greater than two, an analysis of variance (ANOVA) model can be used to detect those genes differentially expressed between the conditions. The chi-squared test based the ANOVA model has $D-1$ degrees of freedom.

To improve power of DE analysis, we recently proposed a 1-df test based on a rank-reduced logistic ANOVA model for logarithm of the expected read count, where the read count was assumed to follow a Poisson-lognormal distribution [9]. This method was termed PLN-ANOVA. In this paper, we present a novel framework committing to facilitate the DE analysis of RNA-seq read count data generated by experiments with multiple conditions, which takes the correlation between samples (if any) into account . The negative binomial distribution is used to model the read count, where the effects of gene and condition on the read counts are incorporated through a two-factor log-linear model with both main effects and interaction effects, while the association between gene and condition is quantified through D interaction terms for each of G genes. We propose to reduce the dimension of independent interaction parameters from $(D-1)(G-1)$ to $D+G-2$. A rigorous statistical estimation/test procedure is developed in this paper, which could be much more efficient compared with the ones without dimension reduction when $D > 2$. The major difference between the current paper and [9] are two-fold: in the distribution assumptions on the read count are different, and the former can deal with both matched samples and unmatched samples but the later can only deal with matched samples.

Methods

In this section, we describe a rank-reduced model for the RNA-seq read counts of biological samples drawn from multiple conditions, then develop an estimation/test procedure for DE analysis.

Notation and model

Let X denote the read count (i.e., the number of reads mapped to a reference) for any interested gene of any sample. Let δ denote the library size factor for that sample, then we can normalize the read count by $Y = [X/\delta]$, where $[a]$ is the integer closest to a. The size factor can be estimated by any normalization method to be described in the next subsection. As in many statistical methods for analyzing RNA-seq read counts, we assume that Y follows a negative binomial distribution. That is, there is an underlying gene expression Z that follows a gamma distribution, and Y follows a Poisson distribution with expectation Z. This way, Y marginally follows a compound Poisson-gamma distribution (i.e., negative binomial distribution). In the following, we extend the negative binomial distribution to model the read counts of biological samples drawn from multiple conditions, where the samples can be either independent or correlated with each other.

Consider a study involving D conditions, and n_d samples are drawn from the dth condition ($d = 1, \dots, D$). Let X_{idg} denote the read count of the gth gene for the ith sample in the dth condition. Let $Y_{idg} = [X_{idg}/\delta_{id}]$ be a normalized read count, where δ_{id} is a size factor for the ith sample in the dth condition. We assume the following conditions hold:

(C1) $Y_{i1g}, \dots, Y_{iDg}$ are conditionally independent of each other given their underlying gene expressions $Z_{i1g}, \dots, Z_{iDg}$, and Z_{idg} follows a gamma distribution with expectation μ_{dg} and variance $\phi_g \mu_{dg}^2$. As a result, Y_{idg} marginally follows the negative binomial distribution with expectation μ_{dg} and dispersion parameter ϕ_g (i.e., the variance of Y_{idg} is $\mu_{dg} + \phi_g \mu_{dg}^2$).

(C2) The correlation coefficient of Z_{id_1g} and Z_{id_2g}, $\rho_{d_1d_2}$, is free of gene g, which is equal to 0 for $d_1 \neq d_2$ in the unmatched situation.

(C3) As in [10] and [9], the following rank-reduced logistic ANOVA model holds for the logarithm of expected gene expression:

$$\log \mu_{dg} = \mu + \alpha_d + \beta_g + \gamma_{dg} = \mu + \alpha_d + \beta_g + u_d v_g, \quad (1)$$

where μ is the grand mean, α_d is the main effect for condition d, β_g is the main effect for gene g, and $\gamma_{dg} := u_d v_g$ is the interaction effect between gene g and condition d.

(C4) There are a sufficiently large number of DE genes between G genes, i.e., $\gamma_{dg} \neq 0$ for a large number of genes. As remarked below, this implies that $u_d \neq 0$ for at least one d.

Equality restrictions are imposed for the sake of parameter identifiability:

$$\sum_{d=1}^{D} n_d \alpha_d = 0, \quad (2)$$

$$\sum_{g=1}^{G} \beta_g = 0, \quad (3)$$

and $\sum_{d=1}^{D} n_d \gamma_{dg} = 0$, $u_1 = 1$, or equivalently

$$\sum_{d=1}^{D} n_d u_d = 0 \text{ and } u_1 = 1, \quad (4)$$

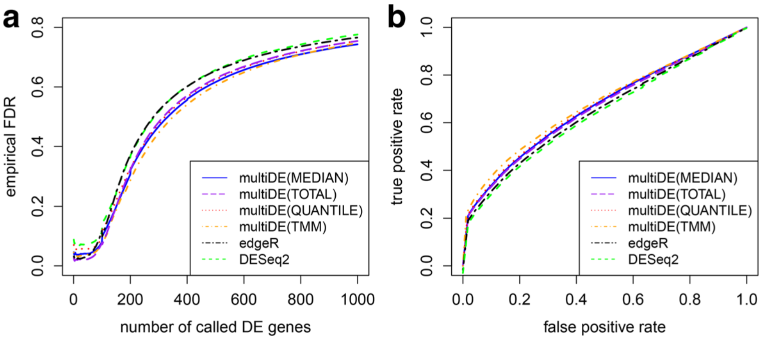

Fig. 1 Performance of various methods with simulation data based on the embryonic stem cells study. **a** False discovery rates; **b** ROC curves

$$\sum_{g=1}^{G} \nu_g = 0. \tag{5}$$

The restrictions (2) and (4) take the sample sizes into account, which will yield simple weighted least squares estimates described in the next section. Under the above restrictions, the main condition effect α_d should be equal to zero so that there is no systematic gene expression difference between the D conditions. The main genetic effect β_g quantifies the relative expression of gene g. Since $\log \mu_{d_1 g} - \log \mu_{d_2 g} = \gamma_{d_1 g} - \gamma_{d_2 g}$ provided $\alpha_d = 0$, γ_{dg} quantifies the relative association strength between condition d and the expression level for gene g. If there is at least one gene differentially expressed between the D conditions, then $\gamma_{dg} \neq 0$ for at least one g, hence $u_d \neq 0$ for at least one d under decomposition (1). Therefore, the relative association strength can be measured by ν_g, and the null hypothesis that the gth gene is differentially expressed between the D conditions can be formulated by

$$H_g : \nu_g = 0. \tag{6}$$

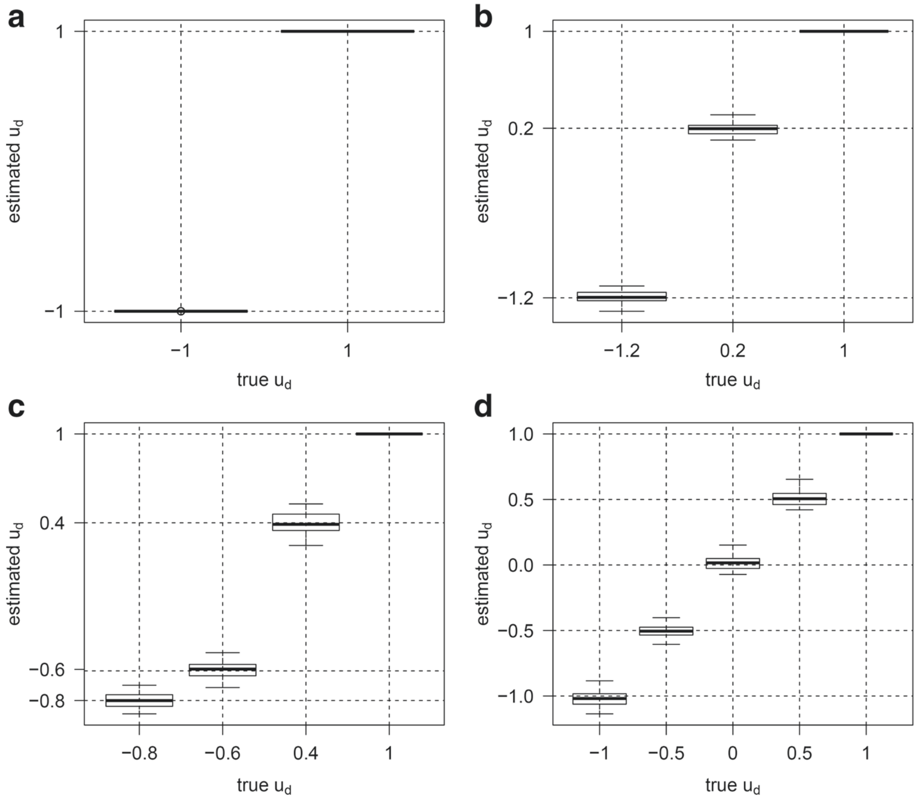

Fig. 2 Estimates of u_d in matched sample situation. **a** D=2; **b** D=3; **c** D=4; **d** D=5

The key idea of model (1) is to approximate the interaction effect γ_{dg} using the product of two terms that depend on d and g separately. This reduces the dimension of test problem from $D - 1$ (the corresponding null hypothesis is $H_g : \gamma_{1g} = \ldots = \gamma_{Dg}$) to one (the corresponding null hypothesis is $H_g : v_g = 0$) for gene g. Obviously, the resultant test could be much more powerful than the chi-square test based on the ANOVA model without such decomposition if $D > 2$.

Read count normalization

In this subsection, we consider the calculation of the size factor δ_{id}. It is widely known that raw counts are not directly comparable between genes due to differential gene lengths and sequencing depths, and reads per kilobase per million reads (RPKM) can be used to correct the resultant technical bias [11]. In DE analysis between multiple conditions, the gene length does not affect the analysis result since such DE analysis focuses on the same gene. However, the condition comparison could greatly suffer from sample specific effects such as sequencing depth and sample specific GC-content effect. The sample specific GC-content effect could arise if two or more samples are sequenced in the same lane. Several within-lane normalization methods (i.e., regression normalization, global-scaling normalization, and full-quantile normalization) can be used to correct the resultant technical bias [12]. On the other hand, such effect can be absorbed into sample specific sequencing depth if only a single sample is sequenced in each lane, and the following four between-lane normalization methods are designed for correcting the technical bias due to sequencing depth.

The first one is the median normalization, denoted by MEDIAN, which takes the form [5]

$$\delta_{id} = \frac{m_{id}}{\left(\prod_{d'=1}^{D}\prod_{i'=1}^{n_{d'}} m_{i'd'}\right)^{\left(\sum_{d'=1}^{D} n_{d'}\right)^{-1}}}, \qquad (7)$$

where m_{id} is the sample median of $\{X_{id1}, \ldots, X_{idG}\}$.

The second one is the total count normalization, denoted by TOTAL, which takes the form [13]

$$\delta_{id} = \frac{s_{id}}{\left(\sum_{d'=1}^{D} n_{d'}\right)^{-1}\sum_{d'=1}^{D}\sum_{i'=1}^{n_{d'}} s_{i'd'}}, \qquad (8)$$

where s_{id} is the sum of $\{X_{id1}, \ldots, X_{idG}\}$.

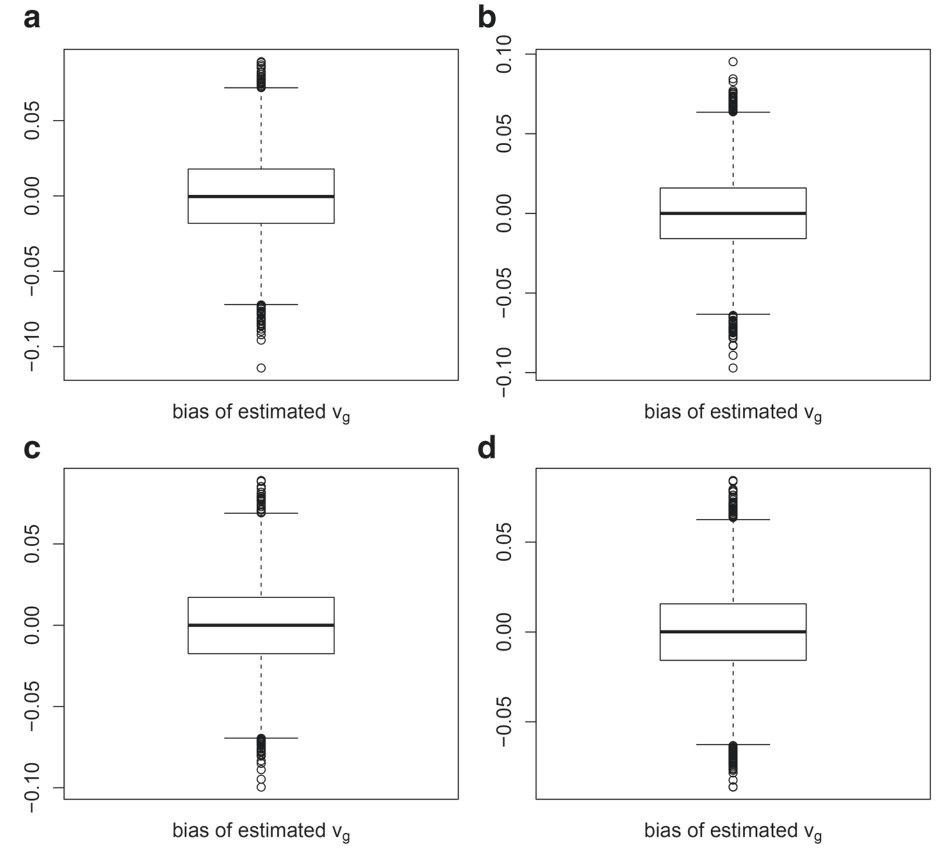

Fig. 3 Estimation biases of v_g in matched sample situation. **a** D=2; **b** D=3; **c** D=4; **d** D=5

The third one is the quantile normalization, denoted by QUANTILE, which takes the form [2]

$$\delta_{id} = \frac{q_{id}}{(\sum_{d'=1}^{D} n_{d'})^{-1} \sum_{d'=1}^{D} \sum_{i'=1}^{n_{d'}} q_{i'd'}}, \tag{9}$$

where q_{id} is the 75th percentile of $\{X_{id1}, \ldots, X_{idG}\}$.

The fourth is a weighted trimmed mean of M-values, denoted by TMM, which has been implemented in the Bioconductor package edgeR [14]. For sample i in condition d and sample i' in condition d', the log-fold changes (M-values) are defined as

$$M_{idg}^{i'd'} = \frac{\log_2(X_{idg}/N_{id})}{\log_2(X_{i'd'g}/N_{i'd'})}, \quad g = 1, \ldots, G,$$

where $N_{id} = \sum_{g=1}^{G} X_{idg}$. If one uses sample i_0 in condition d_0 as a reference sample, then the size factor for sample i in condition d is defined as

$$\delta_{id} = \frac{\sum_{g \in G^*} w_{idg}^{i_0 d_0} M_{idg}^{i_0 d_0}}{\sum_{g \in G^*} w_{idg}^{i_0 d_0}}, \tag{10}$$

where G^* is the set of those genes with the upper and lower 30 % of the M-values removed. In edgeR that implements TMM, the sample with upper quartile closest to the mean upper quartile across all samples is used as the default reference. To minimize the variance of δ_{id}, here the weight $w_{idg}^{i_0 d_0}$ is defined as inverse of the estimated variance of M-values:

$$w_{idg}^{i_0 d_0} = \frac{N_{id} - X_{idg}}{N_{id} X_{idg}} + \frac{N_{i_0 d_0} - X_{i_0 d_0 g}}{N_{i_0 d_0} X_{i_0 d_0 g}}.$$

Using any of the above four normalization methods, the read count X_{idg} can be normalized by

$$Y_{idg} = \left[\frac{X_{idg}}{\delta_{id}} \right].$$

Parameter estimation

In this subsection, we derive estimators of μ, α_d, β_g, u_d, ν_g, ϕ_g, and $\rho_{d_1 d_2}$. We first describe two existing methods for estimating ϕ_g. Then we develop a simple estimation equation method for estimating μ, α_d, β_g, u_d, and ν_g. Finally, we propose a method for estimating $\rho_{d_1 d_2}$ in the matched sample situation.

First, we estimate the dispersion parameter ϕ_g using an empirical Bayes method or a parametric method via a robust gamma-family generalized linear model. The two dispersion estimation methods have been implemented in the Bioconductor packages edgeR and DESeq2, respectively.

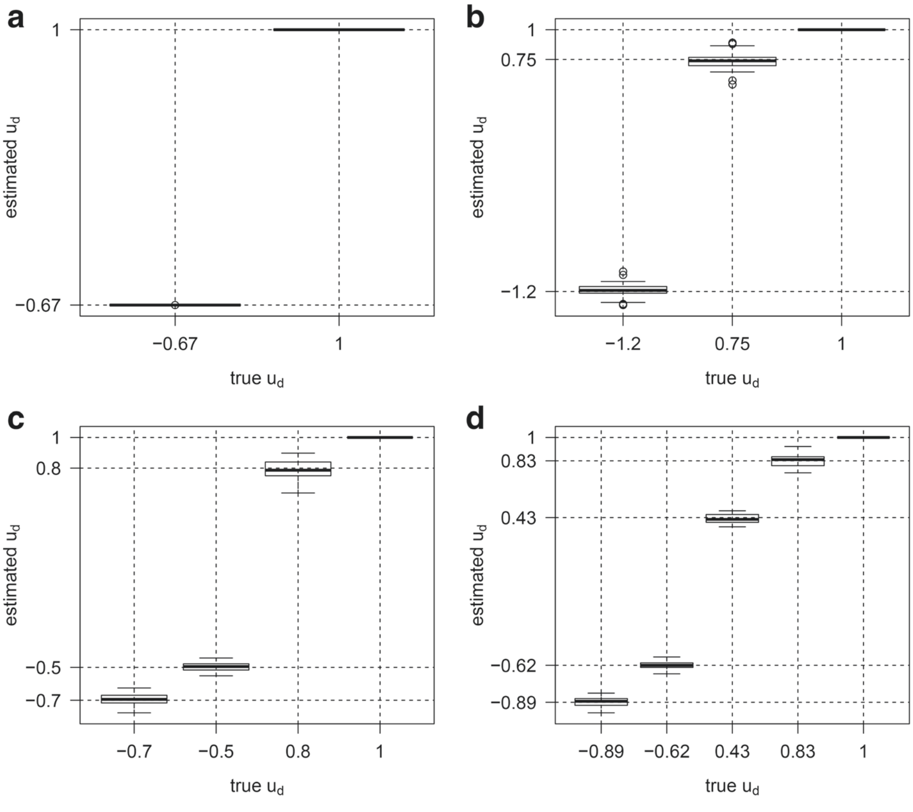

Fig. 4 Estimates of u_d in unmatched sample situation. **a** D=2; **b** D=3; **c** D=4; **d** D=5

Next, we describe a novel method for estimating μ, α_d, β_g, u_d, and v_g. Denote $\eta_{dg} = \log\mu_{dg}$, which can be estimated by

$$\hat{\eta}_{dg} = \log\hat{\mu}_{dg} = \log\left(\frac{1}{n_d}\sum_{i=1}^{n_d} Y_{idg}\right). \quad (11)$$

Under restrictions (2)-(5), we have the moment estimators of μ, α_d, β_g, and γ_{dg}:

$$\hat{\mu} = \frac{\sum_{d=1}^{D}\sum_{g=1}^{G} n_d\hat{\eta}_{dg}}{G\sum_{d=1}^{D} n_d}, \quad (12)$$

$$\hat{\alpha}_d = \frac{1}{G}\sum_{g=1}^{G}\hat{\eta}_{dg} - \hat{\mu}, \quad (13)$$

$$\hat{\beta}_g = \frac{\sum_{d=1}^{D} n_d\hat{\eta}_{dg}}{\sum_{d=1}^{D} n_d} - \hat{\mu}, \quad (14)$$

$$\hat{\gamma}_{dg} = \hat{\eta}_{dg} - \hat{\mu} - \hat{\alpha}_d - \hat{\beta}_g.$$

Here we adopt a weighted least squares approach to estimating u_d and v_g [15]. Noting that $\hat{\gamma}_{dg}$ has an asymptotic variance proportional to n_d^{-1}, we can minimize the weighted sum of squares

$$l(u_d, v_g) = \sum_{d=1}^{D}\sum_{g=1}^{G} n_d(\hat{\gamma}_{dg} - u_d v_g)^2$$

subject to restrictions (2)-(5). The resultant solution $\{\hat{u}_d : d = 1,\ldots,D; \hat{v}_g : g = 1,\ldots,G\}$ satisfies the following equations:

$$\hat{u}_d = \frac{\tilde{u}_d}{\tilde{u}_1}\text{with } \tilde{u}_d = \frac{\sum_{g=1}^{G}\hat{\gamma}_{dg}\hat{v}_g}{\sum_{g=1}^{G}\hat{v_g^2}}, \quad (15)$$

$$\hat{v}_g = \frac{\sum_{d=1}^{D}\hat{\gamma}_{dg} n_d\hat{u}_d}{\sum_{d=1}^{D} n_d\hat{u_d^2}}. \quad (16)$$

In real situations, most genes are equally expressed with $v_g = 0$ and the corresponding information is pure noise for estimating u_d. To eliminate the impact of such noise, we can modify (15) as

$$\hat{u}_d = \frac{\tilde{u}_d}{\tilde{u}_1}\text{with } \tilde{u}_d = \frac{\sum_{g\in S}\hat{\gamma}_{dg}\hat{v}_g}{\sum_{g\in S}\hat{v_g^2}}, \quad (17)$$

where S is a DE gene set determined by any existing method such as edgeR. Equations (16) and (17) can be easily solved in an iterative manner, and the algorithm for all genes can be greatly sped up via vectorization.

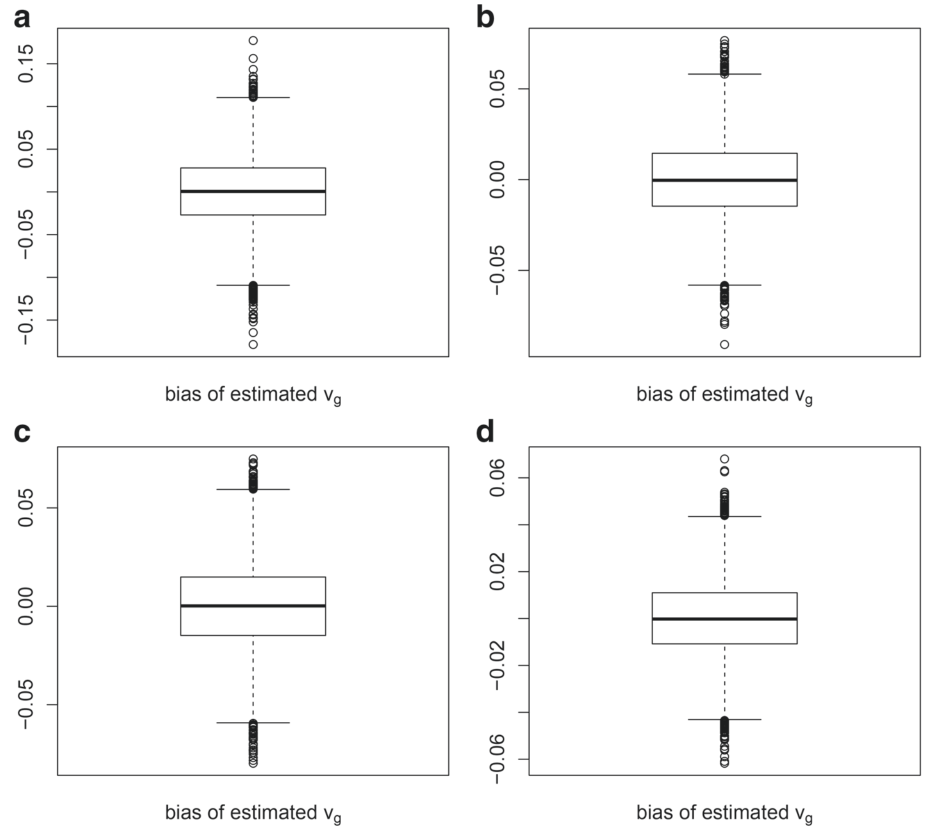

Fig. 5 Estimation biases of v_g in unmatched sample situation. **a** D=2; **b** D=3; **c** D=4; **d** D=5

Finally, we derive an estimator of the correlation coefficient $\rho_{d_1d_2}$ in the matched sample situation. By conditions (C1) and (C2), we have that

$$\begin{aligned}\text{corr}(Y_{id_1g}, Y_{id_2g}) &= \frac{\text{cov}(Z_{id_1g}, Z_{id_2g})}{\{\text{var}(Y_{id_1g})\text{var}(Y_{id_2g})\}^{1/2}}\\ &= \frac{\rho_{d_1d_2}\mu_{d_1g}\mu_{d_2g}\phi_g}{\{\mu_{d_1g}+\mu^2_{d_1g}\phi_g)(\mu_{d_2g}+\mu^2_{d_2g}\phi_g)\}^{1/2}}.\end{aligned}$$

Since $Y_{1dg}, \ldots, Y_{ndg}$ are identically distributed, we can estimate $\text{corr}(Y_{id_1g}, Y_{id_2g})$ with the sample correlation coefficient of $(Y_{1d_1g}, Y_{1d_2g}), \ldots, (Y_{nd_1g}, Y_{nd_2g})$ times a continuity correct term proposed by [16], which is denoted by $r_{d_1d_2g}$. Here the continuity correct term is used to reduce the estimation bias due to a very small sample size. An estimator of $\rho_{d_1d_2}$ takes the form

$$\frac{1}{G}\sum_{g=1}^{G}\frac{r_{d_1d_2g}\{(\hat\mu_{d_1g}+\hat\mu^2_{d_1g}\hat\phi_g)(\hat\mu_{d_2g}+\hat\mu^2_{d_2g}\hat\phi_g)\}^{1/2}}{\hat\mu_{d_1g}\hat\mu_{d_2g}\hat\phi_g}.$$

To construct a Wald test statistic for testing $H_g : \nu_g = 0$, we need to estimate the variance of $\hat\nu_g$, as detailed in the next subsection.

Variance estimation and Wald test

Since $\hat\mu$, $\hat\alpha_d$, and $\hat u_d$ use the information across a sufficiently large number of genes, it is reasonable to assume that their variances are ignorable compared with those of the other estimators. Therefore, we have

$$\text{var}(\hat\nu_g) \approx \frac{\text{var}(\sum_d \hat\eta_{dg} n_d \hat u_d)}{(\sum_d n_d \hat u_d^2)^2}.$$

Using the delta method, we have that

$$\text{var}(\hat\eta_{dg}) \approx \mu_{dg}^{-2}\text{var}(\hat\mu_{dg}) = n_d^{-1}(\hat\mu_{dg}^{-1}+\hat\phi_g).$$

Therefore, we can estimate $\text{var}(\hat\nu_g)$ by

$$\widehat{\text{var}}(\hat\nu_g) = \frac{\sum_{d=1}^{D} n_d \hat u_d^2(\hat\mu_{dg}^{-1}+\hat\phi_g)}{(\sum_{d=1}^{D} n_d \hat u_d^2)^2} \tag{18}$$

in the unmatched sample situation. In the matched sample situation with $n_d = n$ for $d = 1, \ldots, D$, we need to further estimate $\text{cov}(\hat\eta_{d_1g}, \hat\eta_{d_2g})$ for $d_1 \neq d_2$. The resultant estimate of $\text{var}(\hat\nu_g)$ in the matched sample situation, $\widehat{\text{var}}(\hat\nu_g)$, takes the form

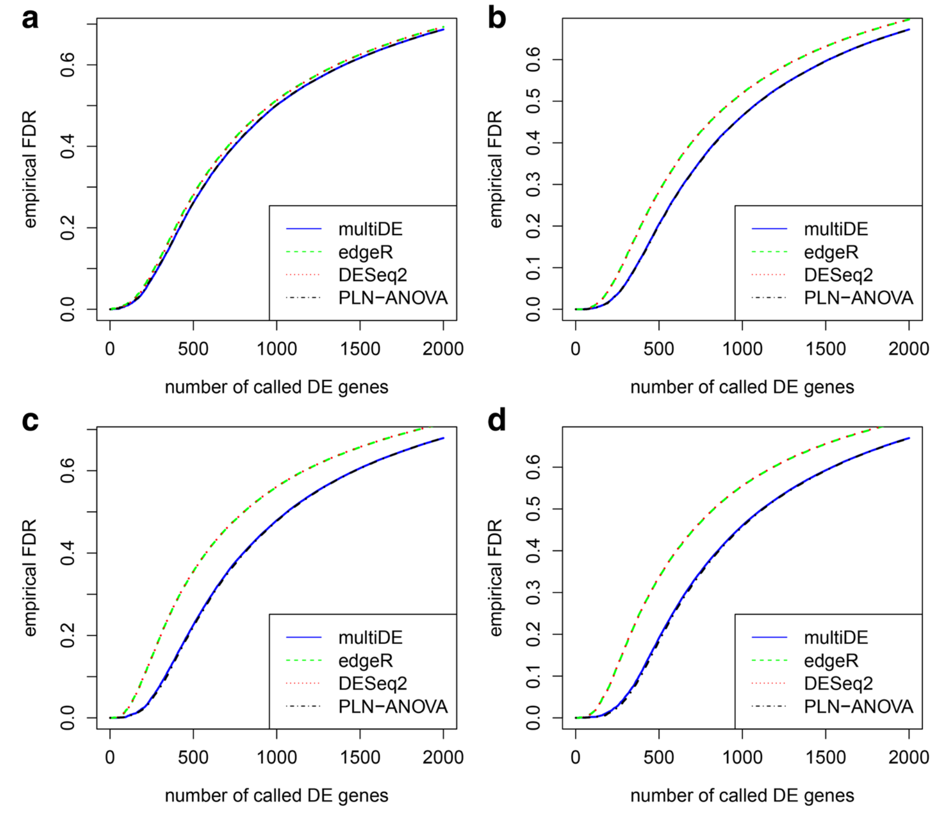

Fig. 6 False discovery rates in matched sample situation. **a** $D = 2$; **b** $D = 3$; **c** $D = 4$; **d** $D = 5$

$$\left\{n\left(\sum_{d=1}^{D}\hat{u}_d^2\right)^2\right\}^{-1}\left\{\sum_{d=1}^{D}\hat{u}_d^2(\hat{\mu}_{dg}^{-1}+\hat{\phi}_g)\right.$$
$$\left.+2\sum_{1\le d_1<d_2\le D}\hat{u}_{d_1}\hat{u}_{d_2}\hat{\phi}_g\hat{\rho}_{d_1d_2}\right\}. \qquad (19)$$

With the estimator (18) or (19), we can construct a Wald test statistic

$$T_g=\frac{\hat{v}_g^2}{\widehat{\text{var}}(\hat{v}_g)}, \qquad (20)$$

whose null limiting distribution is the chi-squared distribution with one degree of freedom.

Results

To evaluate the performance of the proposed method, we conducted both simulation studies and real data analyses. For comparison purpose, we considered two benchmark methods edgeR (version 3.12.1) and DESeq2 (version 1.10.1). The functions *estimateGLMTagwiseDisp* and *estimateDispersions* were used to estimate tagwise negative binomial dispersions in edgeR and DESeq2, respectively. In these two methods, the likelihood ratio tests were used for DE analysis and the default parameters were adopted. We also considered our recently developed method PLN-ANOVA, which was designed for matched samples. In both simulation studies and real data applications for multiDE, the dispersion estimates given by edgeR and DESeq2 produced DE analysis results that were very close to each other. Therefore, in the following studies, we only present the results of multiDE with the dispersion estimates given by edgeR.

Simulation studies

We considered the situation where only a single sample was sequenced in each lane, as commonly done in real applications, so that the effect of sample specific GC-content can be absorbed into the library size. First we evaluated the four normalization methods using simulation data generated from a real dataset. Then we conducted a comparison study between two benchmark methods (i.e., edgeR and DESeq2) and multiDE equipped with MEDIAN using simulation data generated under model (1). Finally, we conducted a sensitivity analysis by generating data without the assumption of model (1).

Simulation based on a real dataset

In multiDE, any of the four methods (namely, MEDIAN , TOTAL, QUANTILE , and TMM) can be used to normalize the read counts.

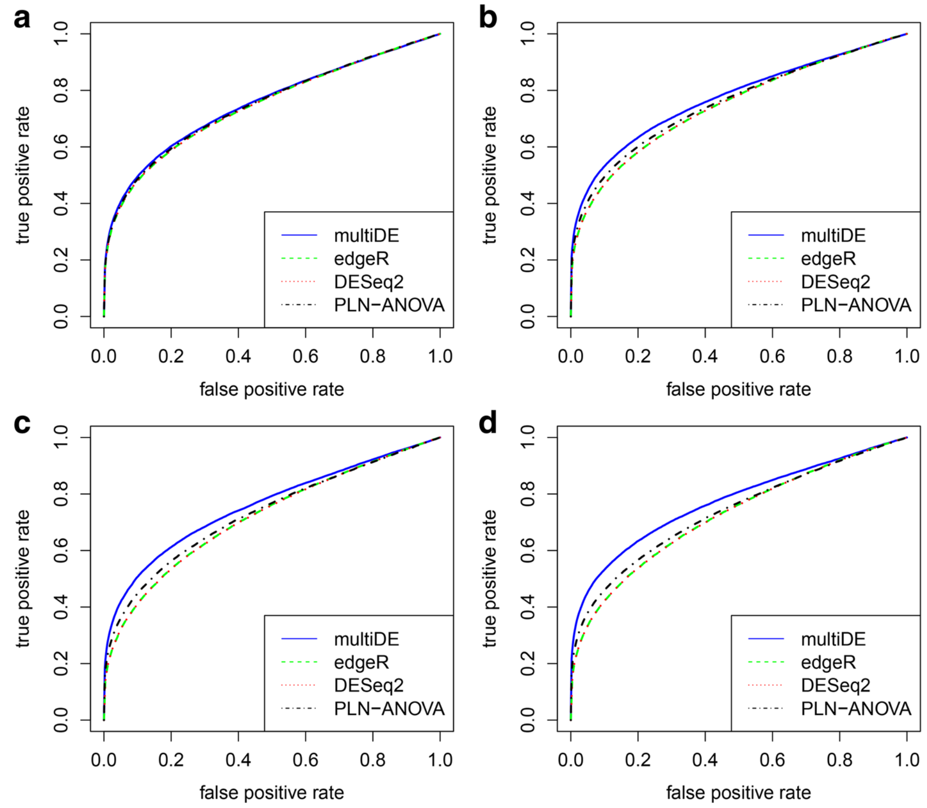

Fig. 7 ROC curves in matched sample situation. **a** $D = 2$; **b** $D = 3$; **c** $D = 4$; **d** $D = 5$

We generated simulation data using a real data set from an embryonic stem cells study described in the next subsection. In this study, RNA-seq read counts at 6,526 genes were available from nine unrelated individuals. Three different conditions were considered, and each condition had three individuals. Refer to the next subsection for detailed description of this study. We randomly permutated the condition information for the nine samples, then randomly selected 600 DE genes. In each of these DE genes, the read counts of the second condition were the original ones multiplied by a factor of 1.1, and the read counts of the third condition were the original ones multiplied by a factor of 0.9. The rest 5,926 genes were regarded as equally expressed genes. Such experiment was repeated for 400 times, and the four normalization methods were applied to these datasets.

The simulation results of multiDE equipped with the four normalization methods were summarized through receiver operating characteristic (ROC) curves and false discovery rates (FDRs) (Fig. 1). In terms of both FDRs and ROC curves, TMM slightly outperformed the other three normalization methods. As stated in [14], TMM is robust against outlying read counts and a significant proportion of DE genes, while the other three normalization methods assume that most genes are not differentially expressed. Actually, these three methods performed relatively poorer than TMM in the application to the above datasets that have about 10 % DE genes. This agrees with the simulation results of [14].

We also applied the two benchmark methods edgeR and DESeq2 to the above simulation datasets. Evidently, multiDE outperformed edgeR and DESeq2 even using the MEDIAN (Fig. 1), we therefore focus on this normalization method in the next two subsections.

Simulation based on model (1)

We considered two situations for generating the read count X_{idg}, one was for unmatched samples and the other was for matched samples. The parameters shared the two situations were set to be the same. We let D = 2, 3, 4, or 5, and fixed the total number of genes to be $G = 10^4$. Then we set $n_d = D + d - 1$ for $d = 1, \ldots, D$ in the unmatched sample situation, and $n = 4$ in the matched sample situation.

To generate RNA-seq read count data, we set $\mu = 4$, $\alpha_d = 0$ for $d = 1, \ldots, D$,

$$(u_1, \ldots, u_D) = \begin{cases} (1, -\frac{2}{3}), & D = 2 \\ (1, \frac{3}{4}, -\frac{6}{5}), & D = 3 \\ (1, \frac{4}{5}, -\frac{3}{6}, -\frac{5}{7}), & D = 4 \\ (1, \frac{5}{6}, \frac{3}{7}, -\frac{5}{8}, -\frac{8}{9}), & D = 5 \end{cases}$$

in the unmatched sample situation, and

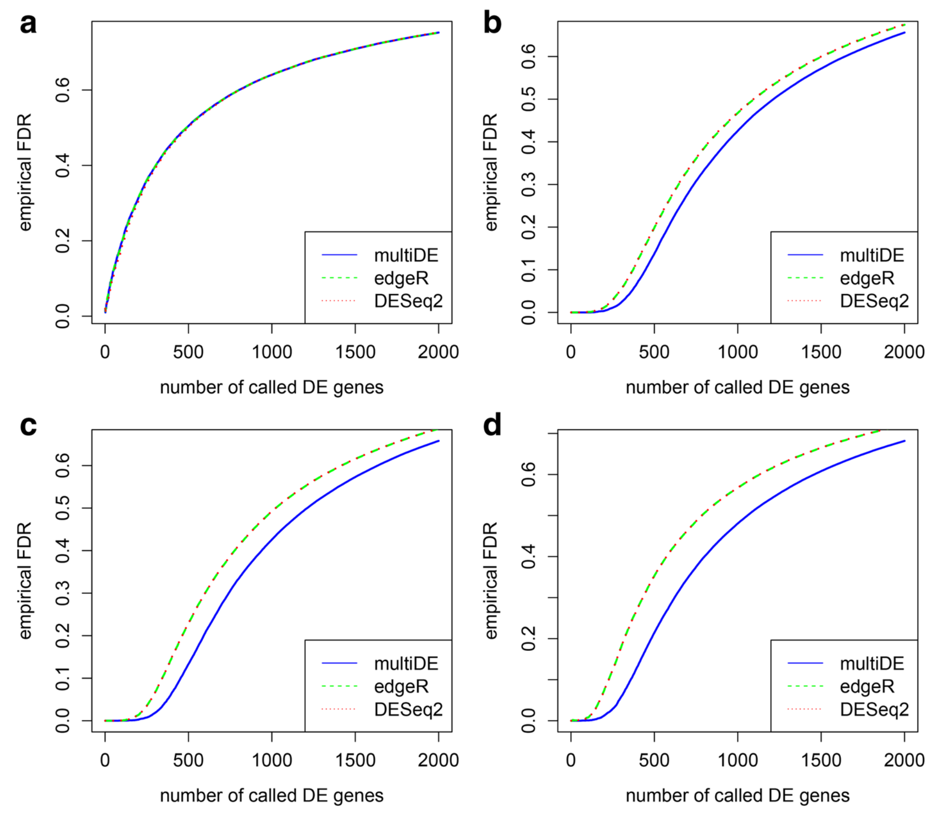

Fig. 8 False discovery rates in unmatched sample situation. **a** $D = 2$; **b** $D = 3$; **c** $D = 4$; **d** $D = 5$

$$(u_1, \ldots, u_D) = \begin{cases} (1, -1), & D = 2 \\ (1, 0.2, -1.2), & D = 3 \\ (1, 0.4, -0.6, -0.8), & D = 4 \\ (1, 0.5, 0, -0.5, -1), & D = 5 \end{cases}$$

in the matched sample situation. We let $v_g = 0$ for equally expressed genes $g = 1, \ldots, 9000$, $v_g \sim -|N(0, 0.32)|$ for up-regulated genes $g = 9001, \ldots, 9500$, and $v_g \sim |N(0, 0.32)|$ for down-regulated genes $g = 9501, \ldots, 10000$. In addition, we randomly generated size factors δ_{id} from the log-normal distribution with mean and varariance parameters 0 and 0.25^2, main genetic effects β_g from the normal distribution with mean 0 and variance 0.25^2, dispersion parameters ϕ_g from the gamma distribution with shape parameter 5 and rate parameter 20, where the shape and rate parameters were close to those for the embryonic stem cells data to be analyzed in the next subsection.

With the above parameters, we then generated underlying gene expressions $(Z_{i1g}, \ldots, Z_{iDg})$ from multivariate gamma distribution with correlation parameters $\{\rho_{d_1 d_2} : 1 \leq d_1 \leq d_2 \leq D\}$ using the Bioconductor package copula [17], and generated the read counts X_{idg} from the Poisson distribution with expectation Z_{idg}. We let $\rho_{d_1 d_2} = 0$ in the unmatched situation and $\rho_{d_1 d_2} \sim U(0.2, 0.4)$ in the matched situation.

For each parameter combination, we generated 50 datasets. For each dataset, u_d and v_g were estimated using multiDE. To evaluate the estimation accuracy of multiDE, we report in Figs. 2, 3, 4 and 5 the u_d estimates and the biases of v_g estimates in both matched and unmatched sample situations. Overall, the u_d estimates were virtually close to the true ones and the estimation biases of v_g were minor.

We obtained the p-values for testing $H_0 : v_g = 0$ using the four considered methods (PLN-ANOVA was applied only in the matched sample situation), and evaluated the performance of these methods through ROC curves and FDRs (Figs. 6, 7, 8 and 9). For $N = 1, \ldots, 2000$, we evaluated the FDRs of top N identified DE genes. The FDR-adjusted p-values were calculated using the R function *p.adjust*, and the empirical FDRs were consequently obtained. In terms of FDRs, the four methods performed quite comparably when $D = 2$. On the other hand, multiDE and PLN-ANOVA had smaller FDRs compared with edgeR and DESeq2 when $D > 2$. According to the ROC curves, when $D = 2$, multiDE was comparable with the other two methods in the unmatched sample situation,

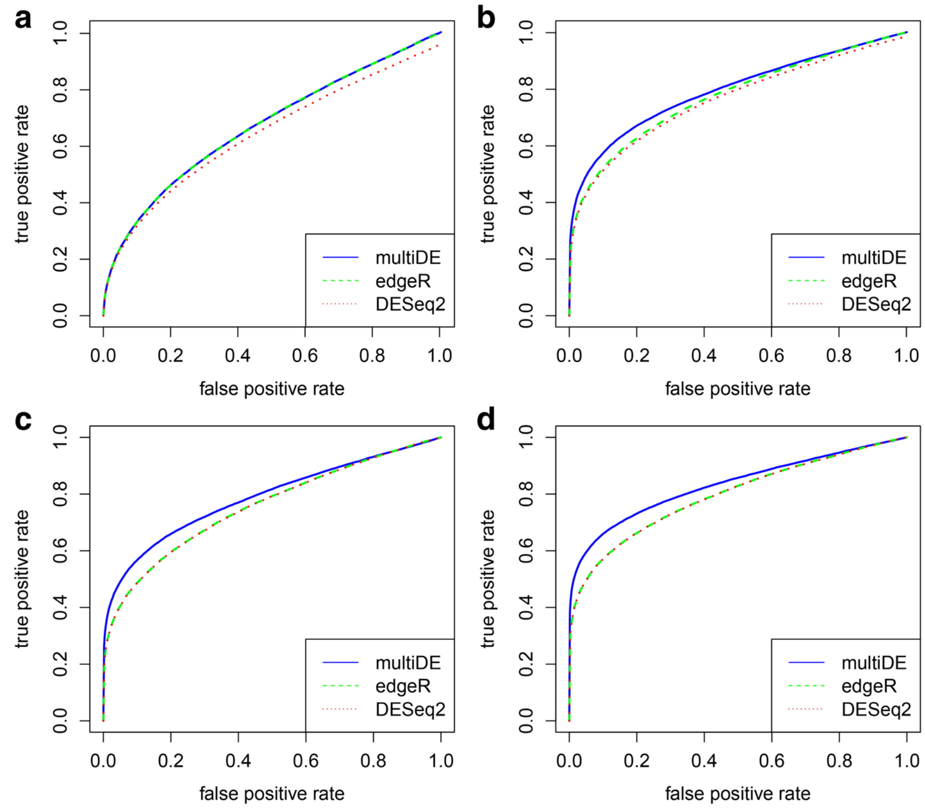

Fig. 9 ROC curves in unmatched sample situation. **a** $D = 2$; **b** $D = 3$; **c** $D = 4$; **d** $D = 5$

and the former was even slightly more powerful in the matched sample situation. When $D > 2$, multiDE was evidently more powerful than the other methods including PLN-ANOVA, and the power gain steadily went up as the number of conditions increased.

Sensitivity analysis

The proposed method multiDE is based on model (1). In real situations, this model might not hold. Therefore, we conducted a sensitive analysis by generating data via the following model:

$$\log \mu_{dg} = \mu + \alpha_d + \beta_g + \gamma_{dg}, \tag{21}$$

where γ_{dg} was generated from the uniform distribution on the interval $(-0.75, 0.75)$ for genes 9,001 through 10,000. Other parameters setting and data generation process were the same as before. We only present the results for unmatched sample situation since the results for matched sample situation were similar. Again, multiDE outperformed edgeR and DESeq2 in terms of both FDRs and ROC curves (Figs. 10 and 11). Interestingly, model misspecification did not alter the performance advantage of multiDE over the other methods.

Real data analyses

In order to evaluate their relative performance on real data, multiDE, edgeR, and DESeq2 were applied to data from two biological experiments with multiple treatment conditions. Furthermore, PLN-ANOVA was applied to the first dataset with matched samples. In each of these two experiments, only a single sample was sequenced in each lane, so we only used between-lane normalization methods.

Psoriatic study

In this study, the major interest was to detect the influence of aryl hydrocarbon receptor (AhR) on RNA expression profiles of psoriatic lesion cells [18]. Each of eight patients were treated with culture treatment of DMSO (vehicle control), AhR agonist FICZ, and AhR antagonist CH-2233191. RNA-seq data were obtained using Illumina Genome Analyzer II platform for each of three treated lesion tissue samples. Therefore, this was a matched sample design. The RNA-seq read counts were derived from the GEO database (accession ID: GSE47944). We kept 13,416 genes with maximal read counts greater than 50 in each treatment condition.

With multiDE, the estimated u_d for the three conditions (vehicle control, AhR agonist, and AhR antagonist) were 1, −4.20, and 3.20, respectively, which coincided with

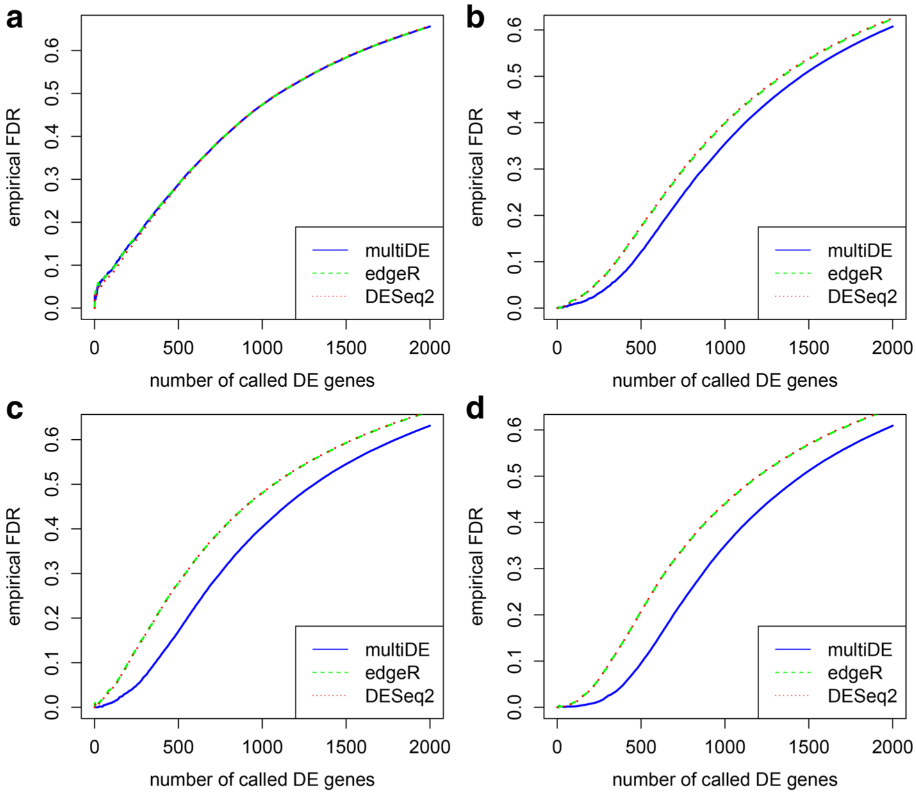

Fig. 10 False discovery rates for model sensitivity analysis. **a** $D = 2$; **b** $D = 3$; **c** $D = 4$; **d** $D = 5$

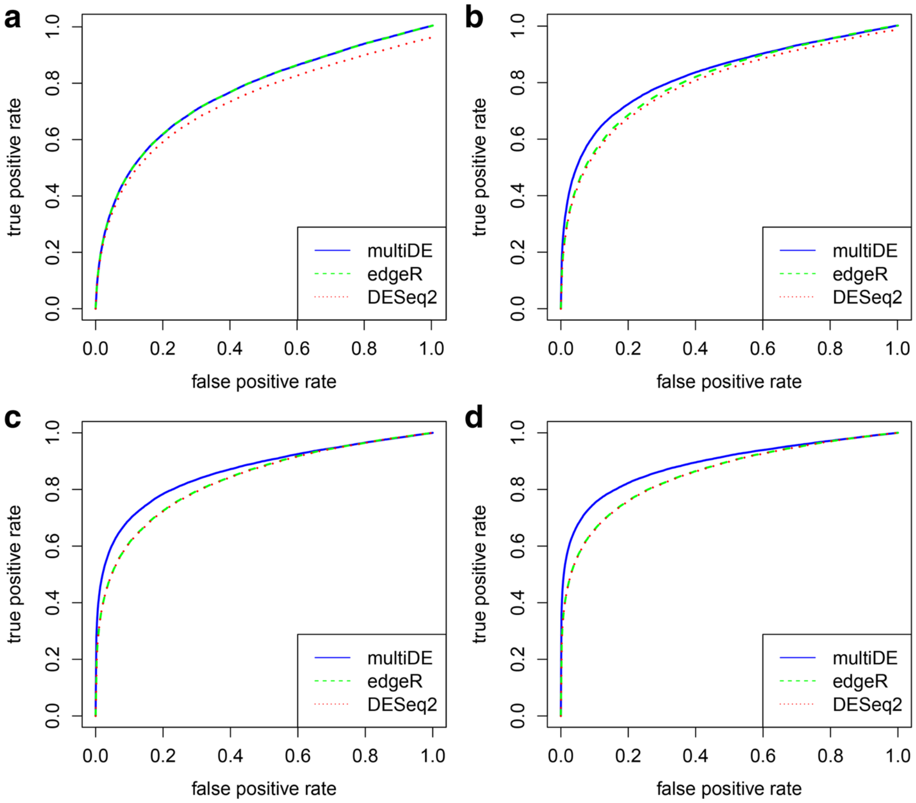

Fig. 11 ROC curves for model sensitivity analysis. **a** $D = 2$; **b** $D = 3$; **c** $D = 4$; **d** $D = 5$

the fact that AhR-activating ligands reduced inflammation in the lesion of psoriasis patients and AhR antagonists upregulated inflammation.

At significance level 0.05, multiDE, PLN-ANOVA, edgeR, and DESeq2 identified 919, 836, 688, and 872 DE genes, respectively. After Bonferroni adjustment, multiDE, PLN-ANOVA, edgeR, and DESeq2 identified 39, 32, 17, and 24 DE genes, respectively (Fig. 12). Evidently, multiDE detected most DE genes. Of the five genes (BATF2, HRNR, SIGLEC1, SLC4A11, CXCL10) uniquely identified by multiDE (with Bonferroni adjustment), four were found to be closely related to psoriasis. In detail, BATF2 could induce the development of CD8α+ dendritic cells, the most powerful antigen presenting cell during inflammation [19]. The up regulation of HRNR, which encoded granular layer keratin bundling proteins, was closely related to psoriatic lesions [20]. As a chemoattractant of a serial of immune cells, encoded protein of CXCL10 also played an important role in psoriatic immune abnormality [21].

It is well known that housekeeping genes maintain basic cellular functions, and they are expressed in all types of cells of an organism. Some typical housekeeping genes (eg., ACTB, GAPDH, NONO, PGK1, PPIH) have relatively constant expressions in most non-pathological situations [22], which can be used to evaluate the false positive performance of DE analysis methods. We found that the aforementioned five reference housekeeping genes were not identified to be DE genes by any of the four considered methods (Table 1).

Embryonic stem cells study

The second dataset on a study of Homo sapiens hormone embryonic stem cells was downloaded from the NCBI GEO database (accession ID: GSE36552). To find causal relationship between gene expression network and cellular phenotype, Yan et al. derived embryonic stem cells from donated human pre-implantation embryos, prepared cDNA and sequenced them by Illumina HiSeq 2000 [23].

RNA-seq samples were obtained from the embryonic stem cells of nine unrelated individuals, so this was an unmatched design and PLN-ANOVA was not applicable. The embryonic stem cells were obtained at the 2-cell stage, three at the 4-cell stage, and the other three at the 8-cell stage. We aligned the downloaded RNA-seq reads to human reference genome hg19 (UCSC release) using the bioinformatics tool *TopHat* [24], and counted the reads for each gene using the Python program *htseq-count* [25]. Altogether, 6,526 genes with the maximal

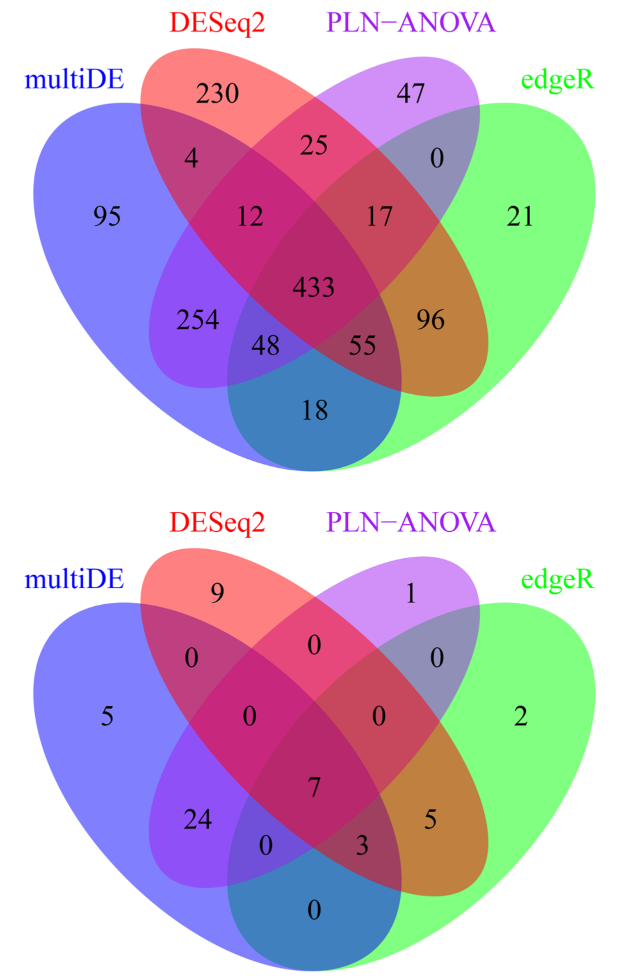

Fig. 12 Number of identified DE genes in the psoriatic study. *Upper panel*: without Bonferroni adjustment; *Down panel*: with Bonferroni adjustment

counts greater than 50 in each treatment condition were kept.

With multiDE, the estimated u_d for the three conditions (2-cell stage, 4-cell stage, and 8-cell stage) were 1, 0.94 and, −1.94, respectively, indicating that the gene expression difference between the 2-cell and 8-cell stages could be generally large, while the difference between the 2-cell and the 4-cell stages was generally minor.

Presented in Fig. 13 are the numbers of DE genes identified by multiDE, edgeR, and DESeq2 at significance level 0.05 with or without Bonferroni adjustment. Among all three methods, multiDE detected most DE genes without Bonferroni adjustment, with a number of 3,392, compared with 3,038 and 3,195 by edgeR and DESeq2, respectively. With Bonferroni adjustment, DESeq2 identified most DE genes, with a number of 1,092, compared with 1,058 and 667 by multiDE and edgeR, respectively.

Using the identified DE genes (with Bonferroni adjustment), we then conducted GO analysis with DAVID [26, 27]. The numbers of identified GO terms are presented in Fig. 14. Evidently, multiDE enriched most GO terms, and the GO terms uniquely enriched by multiDE included protein catabolic process, protein ligase activity, and so on (Table 2). These uniquely enriched GO terms were found to be closely related to the development of embryo. First, ligases always play multiple important roles

Table 1 *P*-values of DE analysis for five housekeeping genes in the psoriatic study

Method	ACTB	GAPDH	NONO	PGK1	PPIH
multiDE	0.1081	0.3010	0.8429	0.1925	0.5984
edgeR	0.1814	0.2139	0.9340	0.1907	0.8923
DESeq2	0.2247	0.1852	0.8984	0.1608	0.8084
PLN-ANOVA	0.1411	0.3862	0.8282	0.2563	0.6134

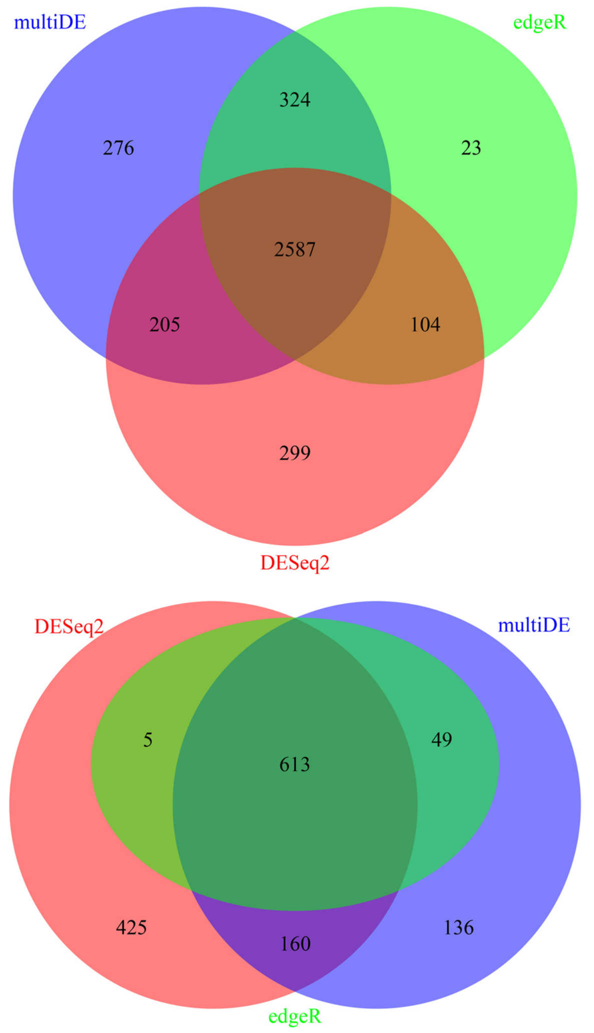

Fig. 13 Number of identified DE genes in the embryonic stem cells study. *Upper panel*: without Bonferroni adjustment; *Down panel*: with Bonferroni adjustment

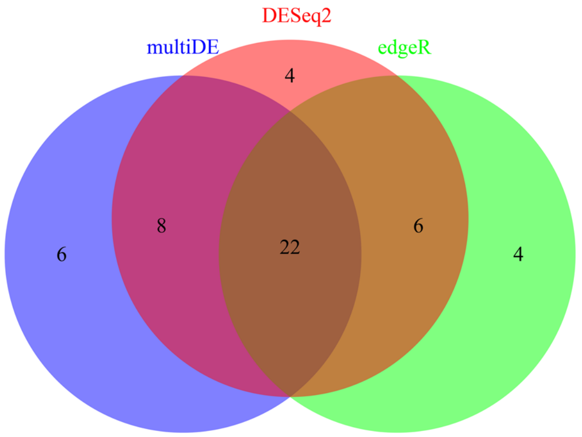

Fig. 14 Number of detected GO terms using identified DE genes in the embryonic stem cells study

in embryo development. For instance, the deficiency of DNA Ligase IV in mice might lead to defective neurogenesis and embryonic lethality [28]. Besides, hSmurf1, a ubiquitin ligase, was shown to have the ability of controlling both embryonic development and a wide variety of cellular responses [29]. Second, the balance of metabolic and protein catabolic was in subtle poise during the development of embryo. As an alternative emergency way supplying energy, catabolism was of significant importance when embryos were facing the threat of nutrient deficiency, especially in their early stage [30]. Third, it was evident that the majority of human embryos donated for research were suffering from various cellular defects, thus the chance of innate embryo malnutrition would be greatly enhanced [31].

Table 2 Exclusively enriched GO terms using identified DE genes in the embryonic stem cells study

	GO code[1]	GO term description	*P*-value[2]
multiDE	GO:0030163 [BP]	Protein catabolic process	1.72e-04
	GO:0044257 [BP]	Cellular protein catabolic process	5.60e-04
	GO:0005829 [CC]	Cytosol	4.00e-03
	GO:0030530 [CC]	Nuclear ribonucleoprotein complex	6.16e-03
	GO:0019787 [MF]	Small conjugating protein ligase activity	2.71e-02
	GO:0031967 [CC]	Organelle envelope	4.32e-02
edgeR	GO:0010468 [BP]	Regulation of gene expression	1.29e-03
	GO:0022618 [BP]	Ribonucleoprotein complex assembly	1.27e-02
	GO:0044452 [CC]	Nucleolar part	1.61e-02
	GO:0006413 [BP]	Translational initiation	4.86e-02
DESeq2	GO:0000279 [BP]	M phase	1.31e-04
	GO:0005819 [CC]	spindle	3.70e-04
	GO:0030880 [CC]	RNA polymerase complex	2.07e-03
	GO:0006259 [BP]	DNA metabolic process	2.83e-03

[1] [BP], biological process ontology; [CC], cellular component ontology; [MF], molecular function ontology.
[2] Bonferroni adjusted *p*-value

As in the psoriatic study, we used the five reference housekeeping genes to verify the performance of the above three DE methods. Since the gene NONO had maximal counts less than 50 in at least one treatment condition, we excluded this gene in the DE analyses. The other four genes (ACTB, GAPDH, PGK1, PPIH) were not identified to be DE genes using multiDE and edgeR. On the other hand, DESeq2 identified GAPDH as a DE gene (Table 3), suggesting that DESeq2 had more false positive findings.

Discussion

Models for fitting the distribution of read count data are essential for detecting DE genes. In experiments involving multiple conditions, it would be of great interest to detect those genes that are differentially expressed between at least two conditions. The traditional statistical methods are generally based on an ANOVA like framework, and the number of the degrees of freedom for detecting DE genes is equal to $D - 1$ (D is the number of conditions). In this paper, we propose to reduce the number of the degrees of freedom from $D - 1$ to one based on a new dimension reduced model. The new method multiDE based on this model can handle both matched and unmatched samples. If $D > 2$, multiDE greatly outpermed the existing methods in our simulation studies, even if the model used to generate data was severely misspecified.

If only a single sample is sequenced in each lane, one needs only to correct the technical bias due to library size effect since the lane-specific efect can be absorbed into the library size factors. In multiDE, there are various options for estimating size factors. Four between-lane normalization methods (i.e., MEDIAN , TOTAL, QUANTILE, and TMM) can be used to estimate size factors. In our simulation study based on a real dataset, TMM slightly outperformed the other three normalization methods. Furthermore, in [14], TMM had been shown to be robust against outlying read counts and DE genes, and it outperformed other methods in simulation studies. Therefore,

Table 3 *P*-values of DE analysis for four housekeeping genes in embryonic stem cells study

Methods	ACTB	GAPDH	PGK1	PPIH
multiDE	0.5419	0.1064	0.2056	0.7460
edgeR	0.5351	0.1845	0.4914	0.9021
DESeq2	0.6540	0.0151	0.6484	0.9666

we recommend TMM in real data applications. If two or more samples are sequenced in the same lane, it would be advantageous to use any within-lane normalization method before between-lane normalization [12].

In multiDE, two methods implemented in edgeR and DESeq2 can be used to estimate dispersion parameters. The two dispersion estimation methods performed comparably in our simulation studies. When estimating u_d, the size of gene set S can be specified to be the number of significant DE genes (after FDR adjustment) by any existing method such as edgeR or DESeq2.

Conclusions

In this paper, the new method multiDE is developed based on a dimension-reduced model for the purpose of detecting DE genes between multiple conditions. Through both simulation studies and real data applications, multiDE was shown to outperform the existing benchmark methods. The proposed method multiDE has been implemented in an R package. This package requires that each condition has at least two biological replications, it takes RNA-seq read counts as input data and can be used to estimate fold changes and to conduct Wald tests for detecting DE genes between various conditions. Three functions are provided in multiDE, namely *normalization*, *dispersion*, and *multiDE*, which can be used to calculate size factors using four normalization methods (i.e., MEDIAN , TOTAL, QUANTILE, and TMM), to estimate dispersion parameters using two methods provided in edgeR and DESeq2, respectively, and to calculate DE p-values and fold changes and their standard errors. Using a desktop computer with a 3.20GHz CPU, it took multiDE only a few seconds to analyze two real datasets.

Abbreviations

AhR, aryl hydrocarbon receptor; DE, differentially expressed; FDR, false discovery rate; NB, negative binomial; ROC, receiver operating characteristic

Funding

HZ was partially supported by the State Key Development Program for Basic Research of China 2012CB316505 and the National Natural Science Foundation of China 11371101.

Availability of data and materials

Source code repository: http://homepage.fudan.edu.cn/zhangh/softwares/multiDE/. Psoriatic study dataset: Downloaded from the GEO database with accession ID GSE47944. Embryonic stem cells study dataset: Downloaded from the GEO database with accession ID GSE36552.

Authors' contributions

GK contributed to drafting the manuscript, developing the new statistical method, and analyzing both the simulation data and real data. LD contributed to analyzing the real data analysis and drafting the manuscript. HZ conceived the idea and designed the model. All authors read and approved the final manuscript.

Competing interests

The authors declare that they have no competing interests.

Consent for publication

Not applicable.

Ethics approval and consent to participate

Not applicable.

Received: 30 January 2016 Accepted: 2 June 2016
Published online: 22 June 2016

References

1. 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.
2. Bullard JH, Purdom E, Hansen KD, Dudoit S. Evaluation of statistical methods for normalization and differential expression in mRNA-Seq experiments. BMC Bioinformatics. 2010;11(1):94.
3. Kvam VM, Liu P, Si Y. A comparison of statistical methods for detecting differentially expressed genes from RNA-seq data. Am J Bot. 2012;99(2):248–56.
4. Wang L, Feng Z, Wang X, Wang X, Zhang X. DEGseq: an R package for identifying differentially expressed genes from RNA-seq data. Bioinformatics. 2010;26(1):136–8.
5. Anders S, Huber W. Differential expression analysis for sequence count data. Genome Biol. 2010;11(10):106.
6. 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.
7. Brandenberger R, Wei H, Zhang S, Lei S, Murage J, Fisk GJ, Li Y, Xu C, Fang R, Guegler K, et al. Transcriptome characterization elucidates signaling networks that control human ES cell growth and differentiation. Nat Biotechnol. 2004;22(6):707–16.
8. Goldhirsch A, Wood W, Coates A, Gelber R, Thürlimann B, Senn HJ, et al. Strategies for subtypes–dealing with the diversity of breast cancer: highlights of the St Gallen International Expert Consensus on the Primary Therapy of Early Breast Cancer 2011. Ann Oncol. 2011;22(8):1736–47.
9. Zhang H, Xu J, Jiang N, Hu X, Luo Z. PLNseq: a multivariate Poisson lognormal distribution for high-throughput matched RNA-sequencing read count data. Stat Med. 2015;34(9):1577–89.
10. Jung Y, Huang JZ, Hu J. Biomarker detection in association studies: modeling SNPs simultaneously via logistic ANOVA. J Am Stat Assoc. 2014;109(508):1355–67.
11. Trapnell C, Williams BA, Pertea G, Mortazavi A, Kwan G, Van Baren MJ, Salzberg SL, Wold BJ, Pachter L. Transcript assembly and quantification by RNA-Seq reveals unannotated transcripts and isoform switching during cell differentiation. Nat Biotechnol. 2010;28(5):511–5.
12. Risso D, Schwartz K, Sherlock G, Dudoit S. GC-content normalization for RNA-Seq data. BMC Bioinformatics. 2011;12(1):480.
13. Li J, Witten DM, Johnstone IM, Tibshirani R. Normalization, testing, and false discovery rate estimation for rna-sequencing data. Biostatistics. 2011;13(3):523–38.
14. Robinson MD, Oshlack A, et al. A scaling normalization method for differential expression analysis of RNA-seq data. Genome Biol. 2010;11(3):25.
15. Ruppert D, Wand MP. Multivariate locally weighted least squares regression. Ann Stat. 1994;22(3):1346–70.
16. Olkin I, Pratt JW. Unbiased estimation of certain correlation coefficients. Ann Math Stat. 1958;29(1):201–11.
17. Kojadinovic I, Yan J, et al. Modeling multivariate distributions with continuous margins using the copula R package. J Stat Softw. 2010;34(9):1–20.
18. Di Meglio P, Duarte JH, Ahlfors H, Owens ND, Li Y, Villanova F, Tosi I, Hirota K, Nestle FO, Mrowietz U, Gilchrist MJ, Stockinger B. Activation of the aryl hydrocarbon receptor dampens the severity of inflammatory skin conditions. Immunity. 2014;40(6):989–1001.
19. Pasparakis M, Haase I, Nestle FO. Mechanisms regulating skin immunity and inflammation. Nat Rev Immunol. 2014;14(5):289–301.
20. Roberson ED, Bowcock AM. Psoriasis genetics: breaking the barrier. Trends Genet. 2010;26(9):415–23.
21. Nomura I, Gao B, Boguniewicz M, Darst MA, Travers JB, Leung DY. Distinct patterns of gene expression in the skin lesions of atopic dermatitis and psoriasis: a gene microarray analysis. J Allergy Clin Immunol. 2003;112(6):1195–202.
22. Zhu J, He F, Hu S, Yu J. On the nature of human housekeeping genes. Trends Genet. 2008;24(10):481–4.
23. Yan L, Yang M, Guo H, Yang L, Wu J, Li R, Liu P, Lian Y, Zheng X, Yan J, et al. Single-cell RNA-Seq profiling of human preimplantation embryos and embryonic stem cells. Nat Struct Mol Biol. 2013;20(9):1131–9.

24. Trapnell C, Pachter L, Salzberg SL. TopHat: discovering splice junctions with RNA-Seq. Bioinformatics. 2009;25(9):1105–11.
25. Anders S, Pyl PT, Huber W. HTSeq–A Python framework to work with high-throughput sequencing data. Bioinformatics. 2004;31(2):166–169.
26. Huang D, Sherman BT, Lempicki RA. Systematic and integrative analysis of large gene lists using DAVID bioinformatics resources. Nat Protoc. 2008;4(1):44–57.
27. Huang DW, Sherman BT, Lempicki RA. Bioinformatics enrichment tools: paths toward the comprehensive functional analysis of large gene lists. Nucleic Acids Res. 2009;37(1):1–13.
28. Frank KM, Sharpless NE, Gao Y, Sekiguchi JM, Ferguson DO, Zhu C, Manis JP, Horner J, DePinho RA, Alt FW. DNA ligase IV deficiency in mice leads to defective neurogenesis and embryonic lethality via the p53 pathway. Mol Cell. 2000;5(6):993–1002.
29. Zhu H, Kavsak P, Abdollah S, Wrana JL, Thomsen GH. A SMAD ubiquitin ligase targets the BMP pathway and affects embryonic pattern formation. Nature. 1999;400(6745):687–93.
30. Jurisicova A, Acton BM. Deadly decisions: the role of genes regulating programmed cell death in human preimplantation embryo development. Reproduction. 2004;128(3):281–91.
31. Antczak M, Van Blerkom J. Temporal and spatial aspects of fragmentation in early human embryos: possible effects on developmental competence and association with the differential elimination of regulatory proteins from polarized domains. Hum Reprod. 1999;14(2):429–47.