Variance component testing for identifying differentially expressed genes in RNA-seq data

Sheng Yang*, Fang Shao*, Weiwei Duan, Yang Zhao and Feng Chen

Department of Biostatistics, School of Public Health, Nanjing Medical University, China

* These authors contributed equally to this work.

ABSTRACT

RNA sequencing (RNA-Seq) enables the measurement and comparison of gene expression with isoform-level quantification. Differences in the effect of each isoform may make traditional methods, which aggregate isoforms, ineffective. Here, we introduce a variance component-based test that can jointly test multiple isoforms of one gene to identify differentially expressed (DE) genes, especially those with isoforms that have differential effects. We model isoform-level expression data from RNA-Seq using a negative binomial distribution and consider the baseline abundance of isoforms and their effects as two random terms. Our approach tests the global null hypothesis of no difference in any of the isoforms. The null distribution of the derived score statistic is investigated using empirical and theoretical methods. The results of simulations suggest that the performance of the proposed set test is superior to that of traditional algorithms and almost reaches optimal power when the variance of covariates is large. This method is also applied to analyze real data. Our algorithm, as a supplement to traditional algorithms, is superior at selecting DE genes with sparse or opposite effects for isoforms.

Subjects Bioinformatics, Statistics

Keywords RNA-seq, Differentially expressed (DE), Generalized mixed linear model (GLMM), Variance component test (VCT)

Submitted 13 April 2017
Accepted 21 August 2017
Published 8 September 2017

Corresponding author
Feng Chen, fengchen@njmu.edu.cn

Academic editor
Jun Chen

Additional Information and Declarations can be found on page 12

DOI **10.7717/peerj.3797**

INTRODUCTION

The availability of massively parallel transcriptome sequencing technology has improved our understanding of the central dogma processes of transcription and translation and the molecular mechanisms of complex diseases (*Koboldt et al., 2013*; *Modelska, Quattrone & Re, 2015*; *Wang, Gerstein & Snyder, 2009*). RNA sequencing (RNA-Seq) has shown that coding genes are stochastically spliced into different transcripts (called isoforms), including partial exons or selected exons (*Kalsotra & Cooper, 2011*; *Kanitz et al., 2015*; *Oshlack, Robinson & Young, 2010*; *Pan et al., 2008*). This process is referred to as alternative splicing (AS). The different sequence characteristics of isoforms may exhibit different expression patterns, and these differences further influence the translation of proteins and affect cellular phenotypes (*Li et al., 2014*).

The elementary problem of transcriptomic data analysis is accurate identification of differentially expressed (DE) genes (*Garber et al., 2011*). Emerging software packages

and pipelines for such analyses have been designed and developed, including methods that rely on gene-level measurement, such as DESeq, edgeR and the two-stage passion model (TSPM), and others based on isoform-level measurements, such as Cuffdiff2, IsoDE and EBSeq (*Al Seesi et al., 2014*; *Anders & Huber, 2010*; *Lehmann et al., 2011*; *Li & Tibshirani, 2013*; *Robinson, McCarthy & Smyth, 2010*; *Trapnell et al., 2013*). The gene-level quantification methods assume that the sum of isoform expression levels constitutes the expression of one gene, which may ignore the variability of different transcripts (*Trapnell et al., 2013*; *Trapnell et al., 2010*). Thus, the development of a method for considering isoform-level measurements is of increasing importance. For example, considering both the splicing structure and expression levels of isoforms, Cuffdiff2 uses the beta negative binomial distribution to define DE isoforms (*Trapnell et al., 2013*). Unfortunately, testing for individual isoforms separately neglects the relationships between each isoform (*Dialsingh, Austin & Altman, 2015*).

Currently, the idea of a set test, which supposes that some related variables form one set and tests the set, likely overcomes the drawbacks of gene-level measurement methods and the disadvantage of testing isoforms individually. The theoretical basis is that the score statistic of the variance component in a generalized linear mixed model (GLMM) follows the mixture chi-squared distribution with the null hypothesis (*Lin, 1997*). The sequencing kernel association test (SKAT), a popular algorithm in genome-wide association studies (GWASs), assumes that the effects of each locus are in the same functional region as random effects and uses the score statistics to test the set (*Wu et al., 2011*). Since the estimation process is under the null hypothesis, this method reduces computation time. The test for the effect of gene set (TEGS), which regresses gene expression against phenotypes, is widely used to select DE genes in a microarray platform. The random effect and the residual error of the model are due to the disease effect and the correlation between genes, respectively (*Huang & Lin, 2013*). The set test is effectively applied to analyze other types of high-dimensional genomic data (*Ionita-Laza et al., 2013*; *Wu et al., 2016*).

Considering the disadvantages of traditional algorithms in selecting heterogeneous genes and the accessibility of the set test, we apply the idea of the set test to identify DE genes in isoform-level measurements. This study involved three parts. First, the relationship between the expression level of one gene and the phenotype is formulated as a GLMM with the assumption of a Poisson and negative binomial (NB) distribution. Second, we study the empirical and theoretical distributions of score statistics and measure their statistical performance for different parameter settings. Third, we analyze level 3 mRNA-Seq data on lung squamous cell carcinoma (LUSC) from The Cancer Genome Atlas (TCGA). Comparisons with traditional algorithms are described in the 'Simulations' and 'Real data analysis'.

METHODS

Model

Assume that there are N subjects in the studied sample, and suppose that for all individuals there is one sequenced gene with p isoforms. For the ith individual, $Y_{i1}, Y_{i2}, \ldots, Y_{ip}$ denotes

the RNA-seq data of one gene. The vector $\mathbf{Y}$ is assumed as a Poisson or NB distribution. With the assumption of independent samples, x_i is the covariate of individual i. For example, its value is 0 or 1 in a case-control study. We assume α_j to be a random term that follows a normal distribution. Its variance represents the heterogeneity of the baseline abundance of the isoforms. This distinct assumption is the difference between gene-level measurement methods and isoform-level measurement methods. The disease effect is also defined as a random effect. Based on the above assumptions, a GLMM of two random terms is constructed as follows:

$$g\left(E\left(Y_{ij}\right)\right)=\mu+\alpha_j+\beta_j x_i,\ i=1,\ldots,N \text{ and } j=1,\ldots,p$$

where $g(.)$ is a monotonic differentiable link function, such as $g(x)=\log(x)$ for the Poisson regression. If x_i is equal to zero in a case-control study, $\mu+\alpha_j$ indicates the expression of the jth isoform in the control group, and β_j indicates the disease effect.

The matrix form is written as follows:

$$g(E(\mathbf{Y}))=\boldsymbol{\mu}+\mathbf{K}\boldsymbol{\alpha}+\mathbf{X}\boldsymbol{\beta}$$

$\mathbf{Y}=\left(\mathbf{Y}_1^T,\mathbf{Y}_2^T,\ldots,\mathbf{Y}_N^T\right)^T$ is an $N\times p$ vector consisting of N $\mathbf{Y}_i$ vectors ($\mathbf{Y}_i=(Y_{i1},Y_{i2},\ldots,Y_{ip})^T$). $\boldsymbol{\mu}$ is also an $N\times p$ vector, in which all elements are $\mu\cdot\mathbf{K}=\left(\mathbf{I}_p,\ldots,\mathbf{I}_p\right)^T$ is an $Np\times p$ matrix in which $\mathbf{I}_p$ is the p-dimensional identity matrix, $\boldsymbol{\alpha}=\left(\alpha_1,\alpha_2,\ldots,\alpha_p\right)^T$, $\mathbf{X}=\left(x_1\mathbf{I}_p,x_2\mathbf{I}_p,\ldots,x_N\mathbf{I}_p\right)^T$, and $\boldsymbol{\beta}=\left(\beta_1,\beta_2,\ldots,\beta_p\right)^T$.

Theoretical null distributions of statistics

In fact, the test of the disease effect is performed to test the variance component of the second random effect. As the basic assumption of mixed models, the third term in the above equation follows a normal distribution ($\boldsymbol{\beta}\sim N\left(0,\ \tau^2\mathbf{I}_p\right)$) (*Gad & El Kholy, 2012*). Therefore, the test of $\boldsymbol{\beta}$ equals the test of its variance component τ. The final null hypothesis (H_0) of the model is written as $\tau=0$. The idea of a score test is used to test τ. Based on the theory of score statistics in a GLMM and the special assumption of this model, the null distribution and parameter estimation of the statistic U is as follows:

First, the matrix form of the first order derivate of the log-likelihood is as follows:

$$\frac{\partial l(\boldsymbol{\alpha},\tau)}{\partial\tau}=\frac{1}{2}\left((\mathbf{Y}-\boldsymbol{\mu}-\mathbf{K}\boldsymbol{\alpha})^T\Delta^{-1}\mathbf{W}\mathbf{X}\mathbf{X}^T\Delta^{-1}\mathbf{W}(\mathbf{Y}-\boldsymbol{\mu}-\mathbf{K}\boldsymbol{\alpha})-tr\left(\mathbf{W}_0\mathbf{X}\mathbf{X}^T\right)\right)$$

where $\Delta^{-1},\mathbf{W}$ and $\mathbf{W}_0$ are block diagonal matrixes, and $\mathbf{W}=E(\mathbf{W}_0)$. Each block of the three matrices is a diagonal matrix whose diagonal elements are similar. Assuming $\gamma_i=E\left(y_i\right)$, their diagonal elements are $\delta_i=1/g'(\gamma_i),w_i=V(\gamma_i)^{-1}\delta_i^2$ and $w_{0i}=w_i+e_i\left(y_i-\gamma_i\right)$, where $e_i=\left(V'(\gamma_i)g'(\gamma_i)+V(\gamma_i)g''(\gamma_i)\right)/\left(V^2(\gamma_i)\left(g'(\gamma_i)\right)^3\right)$. With different variances and the same link function, Δ^{-1} of Poisson and NB distribution is the same, but $\mathbf{W}$ and $\mathbf{W}_0$ are different. For the Poisson assumption, the product of Δ^{-1} and $\mathbf{W}$ is an identity matrix, and $\mathbf{W}=\mathbf{W}_0$. For the NB assumption, we take the product of Δ^{-1} and $\mathbf{W}$ as Φ, which has diagonal elements of $1/(1+\phi\gamma_i)$. ϕ is the scale parameter of an NB. The diagonal elements of the block of $\mathbf{W}_0$ are $w_{0i}=\gamma_i/(1+\phi\gamma_i)+\left(\phi\gamma_i\left(y_i-\gamma_i\right)\right)/(1+\phi\gamma_i)^2$.

Second, the statistic U depends on the first derivate of the log-likelihood function. The formula for the U statistic depends on an assumption. Under the Poisson assumption, U is equal to $U_{Pois} = \frac{1}{2}\left(\left(\mathbf{Y} - g^{-1}(\hat{\mu} + \mathbf{K}\hat{\alpha})\right)^T \mathbf{X}\mathbf{X}^T \left(\mathbf{Y} - g^{-1}(\hat{\mu} + \mathbf{K}\hat{\alpha})\right) - tr\left(\hat{\mathbf{W}}\mathbf{X}\mathbf{X}^T\right)\right)$, while under the NB assumption, U is formulated as $U_{NB} = \frac{1}{2}\left(\left(\mathbf{Y} - g^{-1}(\hat{\mu} + \mathbf{K}\hat{\alpha})\right)^T \hat{\Phi}\mathbf{X}\mathbf{X}^T \hat{\Phi}\right.$ $\left.\left(\mathbf{Y} - g^{-1}(\hat{\mu} + \mathbf{K}\hat{\alpha})\right) - tr\left(\hat{\mathbf{W}}_{\mathbf{0}}\mathbf{X}\mathbf{X}^T\right)\right)$. We estimate $\hat{\mu}$ and $\hat{\alpha}$ of the model under H_0.

Third, the chi-square statistic is shown as $\chi^2 = U\widetilde{I}(\hat{\alpha})^{-1}U$. The formula of the information matrix is as follows:

$$\widetilde{I} = I_{\tau\tau} - I_{\widetilde{\alpha}\tau} I_{\widetilde{\alpha}\widetilde{\alpha}}^{-1} I_{\widetilde{\alpha}\tau}$$

where $I_{\tau\tau} = E\left(\frac{\partial l}{\partial \tau}\frac{\partial l}{\partial \tau}\right), I_{\widetilde{\alpha}\tau} = E\left(\frac{\partial l}{\partial \widetilde{\alpha}}\frac{\partial l}{\partial \tau}\right), I_{\widetilde{\alpha}\widetilde{\alpha}} = E\left(\frac{\partial l}{\partial \widetilde{\alpha}}\frac{\partial l}{\partial \widetilde{\alpha}}\right)$, and $\widetilde{\alpha} = (\mu, \alpha)$. Then, the formula is rewritten as follows:

$$g(E(\mathbf{Y})) = \widetilde{\mathbf{K}}\widetilde{\alpha} + \mathbf{X}\beta$$

where $\widetilde{\mathbf{K}} = \left(\widetilde{\mathbf{I}}_p, \ldots, \widetilde{\mathbf{I}}_p\right)^T, \widetilde{\mathbf{I}}_p = \left(1, \mathbf{I}_p\right)$.

Finally, suppose $\mathbf{A} = \mathbf{K}^T\mathbf{K}$ with a diagonal element of a_{ii}. a_{ii} consists of a vector, denoted as $\widetilde{\mathbf{a}}$. Let κ_{2i}, κ_{3i} and κ_{4i} be the cumulants of $\mathbf{Y}_i$. With the assumption of an exponential family of distributions, the relationships between the three cumulants are as follows:

$$\begin{aligned} \kappa_{2i} &= V(\gamma_i) \\ \kappa_{3i} &= V'(\gamma_i)V(\gamma_i) \\ \kappa_{3i} &= \left(V''(\gamma_i)V(\gamma_i) + \left(V'(\gamma_i)\right)^2\right)V(\gamma_i). \end{aligned}$$

Let $\mathbf{R} \in \mathbb{R}^{Np \times Np}$ be a block diagonal matrix. In each block, the diagonal elements and non-diagonal elements are calculated as $r_{ii}^j = w_i^4\delta_i^{-4}\kappa_{4i} + 2w_i^2 + e_i^2\kappa_{2i} - 2w_i^2\delta_i^{-2}e_i\kappa_{3i}$ and $r_{ii'}^j = 2w_iw_{i'}, i \neq i'$, respectively. $\mathbf{C} \in \mathbb{R}^{Np \times Np}$ is composed of a p diagonal matrix, C^j, whose diagonal elements are $c_{ii}^j = w_i^4\delta_i^{-4}\kappa_{3i} + 2w_i^2 + e_i^2\kappa_{2i} - 2w_i^2\delta_i^{-2}e_i\kappa_{3i}$. Then, the elements of the information matrix are as follows:

$$I_{\tau\tau} = \frac{1}{4}\mathbf{J}^T(\mathbf{ARA})\mathbf{J}, \quad I_{\widetilde{\alpha}\tau} = \frac{1}{2}\widetilde{\mathbf{K}}\mathbf{C}\widetilde{\mathbf{a}}, \quad I_{\widetilde{\alpha}\widetilde{\alpha}} = \widetilde{\mathbf{K}}^T\mathbf{W}\widetilde{\mathbf{K}}$$

where $\mathbf{J}$ is an Np vector, and its elements are 1. The difference in variances leads to different expressions for $\mathbf{R}$ and $\mathbf{C}$. For the Poisson assumption, the diagonal and off-diagonal elements of $\mathbf{R}$ are $r_{ii}^j = \gamma_i + 2\gamma_i^2$ and $r_{i\widetilde{i}} = 2\gamma_i\gamma_{\widetilde{i}}, i \neq \widetilde{i}$, respectively. The diagonal elements of $\mathbf{C}$ are $c_{ii}^j = \gamma_i$. For the NB assumption, the diagonal and off-diagonal elements of $\mathbf{R}$ are $r_{ii}^j = \frac{2\gamma_i^3\phi^2 + 3\gamma_i^3 + 4\gamma_i\phi + 2\gamma_i^2 + \gamma_i}{(1+\gamma_i\phi)^3}$ and $r_{i\widetilde{i}} = 2\frac{\gamma_i}{1+\gamma_i\phi}\frac{\gamma_{\widetilde{i}}}{1+\gamma_{\widetilde{i}}\phi}, i \neq \widetilde{i}$, respectively. The diagonal elements of $\mathbf{C}$ are $c_{ii}^j = \frac{\gamma_i + \gamma_i^2\phi}{(1+\gamma_i^2\phi)^2}$.

Empirical distribution of statistics

The permutation algorithm generates the empirical distribution. The two different assumptions cause different score statistics. For the Poisson assumption, the statistic is $U_{Pois} = \frac{1}{2}\left(\left(\mathbf{Y} - g^{-1}(\hat{\mu} + \mathbf{K}\hat{\alpha})\right)^T \mathbf{X}\mathbf{X}^T \left(\mathbf{Y} - g^{-1}(\hat{\mu} + \mathbf{K}\hat{\alpha})\right) - tr\left(\hat{\mathbf{W}}\mathbf{X}\mathbf{X}^T\right)\right)$. For the NB assumption, the statistic is equal to $U_{NB} = \frac{1}{2}\left(\left(\mathbf{Y} - g^{-1}(\hat{\mu} + \mathbf{K}\hat{\alpha})\right)^T \hat{\Phi}\mathbf{X}\mathbf{X}^T \hat{\Phi}(\mathbf{Y} - g^{-1}\right.$

$(\hat{\mu}+\mathbf{K}\hat{\alpha}))-tr(\hat{\mathbf{W}}_{\mathbf{0}}\mathbf{X}\mathbf{X}^{T})\Big)$. The hypothesis testing of U follows two steps: (a) construction of the empirical distribution by shuffling the label and (b) calculation of the P value of the original statistic U. The estimation of $\boldsymbol{\mu}$ and $\boldsymbol{\alpha}$ is similar to the above. To improve robustness, the default setting of the number of permutations is 5,000.

Simulations

We perform simulations to examine the type I error and power of the proposed score statistics, U, for identifying differential expression under a range of scenarios. The parameter settings are based on the analysis of real data and references (*Lehmann et al., 2011*). We assume that the RNA-Seq data follow NB and Poisson distributions. Five parameters are involved: the variance of $\boldsymbol{\alpha}$ (s), the variance of the disease effect (l), the number of isoforms (p), the logarithm of expression levels ($\exp(M)$) and the dispersion parameter (d). s shows the heterogeneity of the baseline abundance of each isoform. l refers to the heterogeneity of the disease effect. The other three parameters describe the characteristics of the expression levels. The variance of $\boldsymbol{\alpha}$ varies from 0.25 to 1.75 in steps of 0.75. The variance of $\boldsymbol{\beta}$ varies from 0.20 to 1.00 in steps of 0.4. The number of isoforms can be 2, 4 or 8. d is set at 0.5 or 2. We let $\exp(M)$ be 2.5 or 5. The sample size is fixed at 40. In the type I error simulations, the effect size is set to zero ($l=0$).

We compare our approach with three traditional algorithms: DESeq, edgeR and TSPM (*Anders & Huber, 2010*; *Lehmann et al., 2011*; *Robinson, McCarthy & Smyth, 2010*). The sum of isoforms is assumed to represent the expression level of one gene in these methods. Therefore, the loop number is the gene number in simulations. When l is equal to 0, the proportion of significant genes is the false positive rate.

Real data analysis

We download the LUSC Batch 101 mRNA sequencing data from TCGA (*Network, 2012*). The sample size is 12 for the case-control traits. The number of genes is 20,533. Each sample has four files: *gene expression*, *normalized gene expression*, *isoform expression* and *normalized isoform expression*. We only used the *gene expression* and *isoform expression* datasets. These data are attached as File S1. To utilize the advantages of isoVCT, the candidate genes are the genes showing a higher expression level than the total sample, with isoforms that exhibit heterogeneous effect sizes. Finally, we select 6,134 genes.

Software and algorithms

This algorithm is completed using the Microsoft R Open (v3.3.0; Microsoft Corp., Seattle, WA, USA). The functions *glmer* and *glmer.nb* in the *lme4* package fit the mixed models in our algorithm. The function *ginv* in the MASS package is employed to calculate the generalized inverse of the singular matrix. The packages *DESeq* and *edgeR* are both from Bioconductor. In consideration of the potential computing time of the simulations, the *doParallel* package is used for parallel computation.

This algorithm provides self-adaptation for real data analysis. If the fitness of the NB assumption fails or the dispersion parameter is close to one, the Poisson assumption is used. Due to the application of the idea of the variance component test, our method is referred to isoVCT. The distribution of the score test statistic is fitted via theoretical and

Table 1 The type I error rate of five algorithms in NB assumptions.

M	s	p	$\phi = 2$					$\phi = 0.5$				
			The	Emp	DESeq	edgeR	TSPM	The	Emp	DESeq	edgeR	TSPM
2.5	0.20	2	0.034	0.056	0.033	0.053	0.051	0.029	0.049	0.044	0.068	0.062
		4	0.024	0.049	0.041	0.056	0.053	0.031	0.051	0.033	0.051	0.050
		8	0.020	0.049	0.034	0.047	0.045	0.020	0.063	0.037	0.060	0.052
	0.60	2	0.022	0.053	0.041	0.059	0.057	0.013	0.039	0.047	0.058	0.052
		4	0.014	0.048	0.035	0.055	0.051	0.024	0.051	0.041	0.064	0.049
		8	0.010	0.056	0.037	0.057	0.050	0.013	0.047	0.036	0.045	0.043
	1.00	2	0.024	0.049	0.033	0.050	0.047	0.032	0.050	0.034	0.050	0.059
		4	0.019	0.048	0.041	0.060	0.052	0.019	0.058	0.038	0.057	0.054
		8	0.007	0.042	0.036	0.061	0.055	0.000	0.045	0.029	0.051	0.045
5.0	0.20	2	0.029	0.046	0.043	0.056	0.051	0.019	0.071	0.050	0.070	0.059
		4	0.010	0.052	0.042	0.062	0.057	0.000	0.044	0.036	0.062	0.054
		8	0.011	0.052	0.051	0.065	0.059	0.000	0.056	0.047	0.055	0.055
	0.60	2	0.024	0.056	0.044	0.057	0.054	0.002	0.033	0.028	0.045	0.041
		4	0.006	0.056	0.048	0.070	0.059	0.000	0.050	0.046	0.060	0.058
		8	0.014	0.050	0.038	0.061	0.054	0.000	0.059	0.032	0.047	0.039
	1.00	2	0.021	0.046	0.032	0.052	0.044	0.000	0.049	0.038	0.053	0.045
		4	0.010	0.053	0.038	0.061	0.047	0.000	0.045	0.046	0.062	0.053
		8	0.019	0.046	0.038	0.060	0.052	0.000	0.057	0.043	0.067	0.059

empirical methods, which we term isoVCT-The and isoVCT-Emp, respectively. The R program of isoVCT and the simulation code are attached as File S2.

RESULTS

Simulations with the NB assumption

The results for the empirical type I error for the NB assumption are shown in Table 1 and Fig. 1. s, l and p may be unrelated to type I error rates, but ϕ may show an inverse correlation to the type I error rate. isoVCT-The is exactly conservative in specific scenarios. The type I error rate of isoVCT-Emp is close to 0.05, which might mean that the permutation can fit the distribution of U in a small sample size setting. DESeq can control type I errors; however, the type I error rates of edgeR and TSPM are both dispersed, especially when $\phi = 0.5$. Generally, isoVCT can control type I error rates around the nominal level for the NB assumption.

The results regarding empirical power for the NB assumption are shown in Tables 2, 3 and Fig. 2. Three findings of this analysis are as follows: (a) M exhibits a positive correlation with power; (b) p and s exhibit a negative correlation with power; and (c) l also exhibits a positive correlation with power, but its increase is much sharper than those of p and M. Our results show that isoVCT-Emp is more advantageous than traditional methods, especially when the effect size is small and the effect heterogeneity is large.

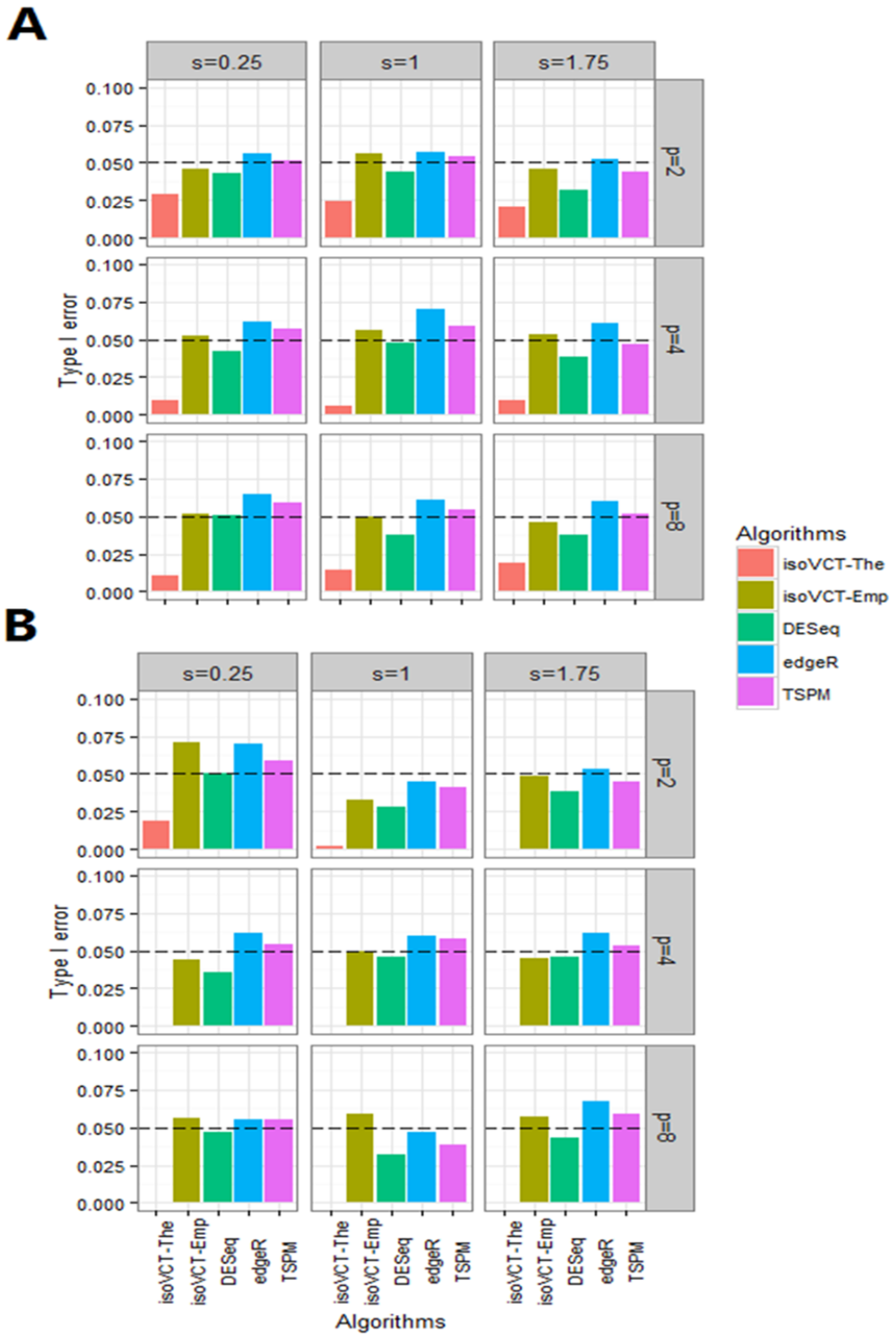

Figure 1 **Plots of type I error of five algorithms in NB assumption.** (A) The parameter setting is mu = 5 and phi = 2. (B) The parameter setting is mu = 5 and phi = 0.5.

Table 2 The power of five algorithms in NB assumptions ($\exp(M) = 5$).

l	s	p	$\phi = 2$					$\phi = 0.5$				
			The	Emp	DESeq	edgeR	TSPM	The	Emp	DESeq	edgeR	TSPM
0.2	0.25	2	0.057	0.182	0.124	0.127	0.141	0.018	0.089	0.065	0.080	0.083
		4	0.029	0.277	0.139	0.096	0.151	0.000	0.109	0.071	0.087	0.079
		8	0.020	0.391	0.130	0.103	0.144	0.000	0.129	0.058	0.074	0.065
	1.00	2	0.031	0.186	0.130	0.104	0.148	0.005	0.065	0.063	0.091	0.073
		4	0.029	0.243	0.123	0.090	0.142	0.000	0.095	0.059	0.092	0.077
		8	0.010	0.349	0.119	0.105	0.140	0.000	0.105	0.065	0.108	0.080
	1.75	2	0.039	0.167	0.116	0.116	0.141	0.002	0.069	0.069	0.095	0.085
		4	0.021	0.228	0.119	0.109	0.133	0.000	0.094	0.067	0.107	0.093
		8	0.010	0.323	0.120	0.100	0.141	0.000	0.106	0.070	0.119	0.083
0.6	0.25	2	0.518	0.691	0.461	0.490	0.490	0.144	0.394	0.258	0.290	0.260
		4	0.390	0.890	0.465	0.482	0.475	0.011	0.530	0.219	0.256	0.242
		8	0.107	0.980	0.459	0.482	0.468	0.000	0.734	0.244	0.275	0.247
	1.00	2	0.420	0.640	0.447	0.488	0.481	0.059	0.285	0.233	0.278	0.247
		4	0.286	0.876	0.470	0.506	0.525	0.002	0.410	0.212	0.265	0.241
		8	0.045	0.981	0.483	0.520	0.503	0.000	0.631	0.209	0.273	0.225
	1.75	2	0.400	0.652	0.472	0.503	0.503	0.043	0.269	0.227	0.279	0.248
		4	0.241	0.847	0.443	0.487	0.475	0.001	0.417	0.224	0.289	0.252
		8	0.030	0.978	0.432	0.480	0.502	0.000	0.591	0.202	0.268	0.233
1.0	0.25	2	0.792	0.882	0.616	0.508	0.617	0.383	0.645	0.367	0.396	0.371
		4	0.809	0.979	0.636	0.555	0.648	0.083	0.814	0.393	0.425	0.418
		8	0.730	0.999	0.656	0.558	0.648	0.000	0.958	0.397	0.422	0.372
	1.00	2	0.709	0.839	0.628	0.555	0.691	0.224	0.529	0.431	0.481	0.447
		4	0.729	0.975	0.641	0.546	0.666	0.027	0.727	0.378	0.439	0.439
		8	0.579	1.000	0.622	0.544	0.665	0.000	0.920	0.378	0.446	0.405
	1.75	2	0.683	0.849	0.656	0.558	0.673	0.203	0.539	0.427	0.482	0.459
		4	0.710	0.968	0.633	0.536	0.694	0.015	0.744	0.364	0.433	0.415
		8	0.505	1.000	0.627	0.557	0.673	0.000	0.928	0.374	0.437	0.386

Simulations with the Poisson assumption

The results of the type I error and empirical power under the Poisson assumption are shown in File S3.

Real data analysis

The results of real data analysis are shown in Fig. 3. Only 4,422 genes are fit to the GLMM. The three traditional algorithms, DESeq, edgeR and TSPM, define 364, 259 and six DE genes, respectively. The intersection of DESeq and edgeR is 221 genes, representing a high proportion of the DE genes that these algorithms define. TSPM is almost ineffective for these genes.

The results of isoVCT are as follows. First, isoVCT-Emp defines 263 DE genes; of these genes, isoVCT-Emp specifically selected 242 DE genes, and only a small portion of these genes is also found by the other algorithms. Second, the DE genes identified by isoVCT-Emp included those that are identified by isoVCT-The. In general, isoVCT is superior at selecting heterogeneous genes.

Table 3 The power of five algorithms in NB assumptions ($\exp(M) = 2.5$).

l	s	p	$\phi = 2$					$\phi = 0.5$				
			The	Emp	DESeq	edgeR	TSPM	The	Emp	DESeq	edgeR	TSPM
0.2	0.25	2	0.057	0.181	0.110	0.086	0.130	0.014	0.081	0.052	0.071	0.078
		4	0.022	0.242	0.113	0.110	0.134	0.000	0.104	0.060	0.084	0.080
		8	0.019	0.365	0.111	0.096	0.124	0.000	0.129	0.063	0.084	0.077
	1.00	2	0.031	0.143	0.121	0.094	0.142	0.006	0.079	0.072	0.094	0.092
		4	0.027	0.208	0.115	0.103	0.141	0.000	0.099	0.069	0.100	0.089
		8	0.010	0.309	0.105	0.109	0.127	0.000	0.101	0.058	0.098	0.074
	1.75	2	0.029	0.144	0.111	0.090	0.127	0.007	0.073	0.068	0.100	0.092
		4	0.020	0.203	0.107	0.091	0.137	0.000	0.086	0.072	0.107	0.085
		8	0.010	0.285	0.121	0.103	0.146	0.000	0.087	0.066	0.103	0.073
0.6	0.25	2	0.527	0.705	0.429	0.460	0.452	0.152	0.392	0.210	0.261	0.236
		4	0.325	0.864	0.445	0.470	0.452	0.011	0.522	0.210	0.239	0.220
		8	0.069	0.978	0.459	0.484	0.465	0.000	0.735	0.218	0.255	0.235
	1.00	2	0.357	0.592	0.448	0.473	0.492	0.060	0.300	0.231	0.281	0.272
		4	0.208	0.831	0.443	0.478	0.484	0.002	0.400	0.217	0.276	0.245
		8	0.020	0.970	0.455	0.494	0.473	0.000	0.614	0.194	0.247	0.217
	1.75	2	0.330	0.584	0.449	0.488	0.471	0.034	0.254	0.200	0.245	0.240
		4	0.158	0.786	0.445	0.487	0.479	0.001	0.383	0.201	0.259	0.219
		8	0.022	0.957	0.441	0.482	0.504	0.000	0.572	0.200	0.283	0.237
1.0	0.25	2	0.760	0.857	0.592	0.512	0.620	0.319	0.625	0.352	0.404	0.386
		4	0.773	0.974	0.630	0.525	0.643	0.067	0.822	0.368	0.396	0.375
		8	0.589	0.999	0.617	0.532	0.610	0.000	0.957	0.385	0.410	0.360
	1.00	2	0.663	0.820	0.650	0.524	0.660	0.180	0.505	0.382	0.432	0.417
		4	0.623	0.958	0.610	0.536	0.623	0.024	0.717	0.407	0.463	0.423
		8	0.415	0.997	0.634	0.537	0.666	0.000	0.933	0.368	0.442	0.391
	1.75	2	0.593	0.791	0.636	0.539	0.658	0.145	0.499	0.389	0.447	0.443
		4	0.595	0.9520	0.628	0.518	0.663	0.013	0.716	0.396	0.446	0.426
		8	0.345	0.9955	0.623	0.569	0.668	0.000	0.914	0.393	0.454	0.405

DISCUSSION

Here, we propose isoVCT, a variance component score test for testing the coefficients of covariates to select DE genes. Simulations and real data analysis both suggest that this method is advantageous in selecting weak effects or heterogeneous genes. The results of the simulations are discussed from the following two aspects. First, isoVCT controls type I errors in each setting under any of the distribution assumptions. The type I error rate of the empirical distribution is almost controlled. However, the type I error rate of isoVCT-The is strictly controlled, which is likely caused by an inaccurate estimation of the U statistic in the small sample size setting. The type I error rates of edgeR and TSPM increase in some settings, especially with the NB assumption (*Yang et al., 2015*). Second, the power of isoVCT is higher than that of other methods; furthermore, the strength is evident for the small l and NB assumption. isoVCT-Emp is superior to other methods in each setting,

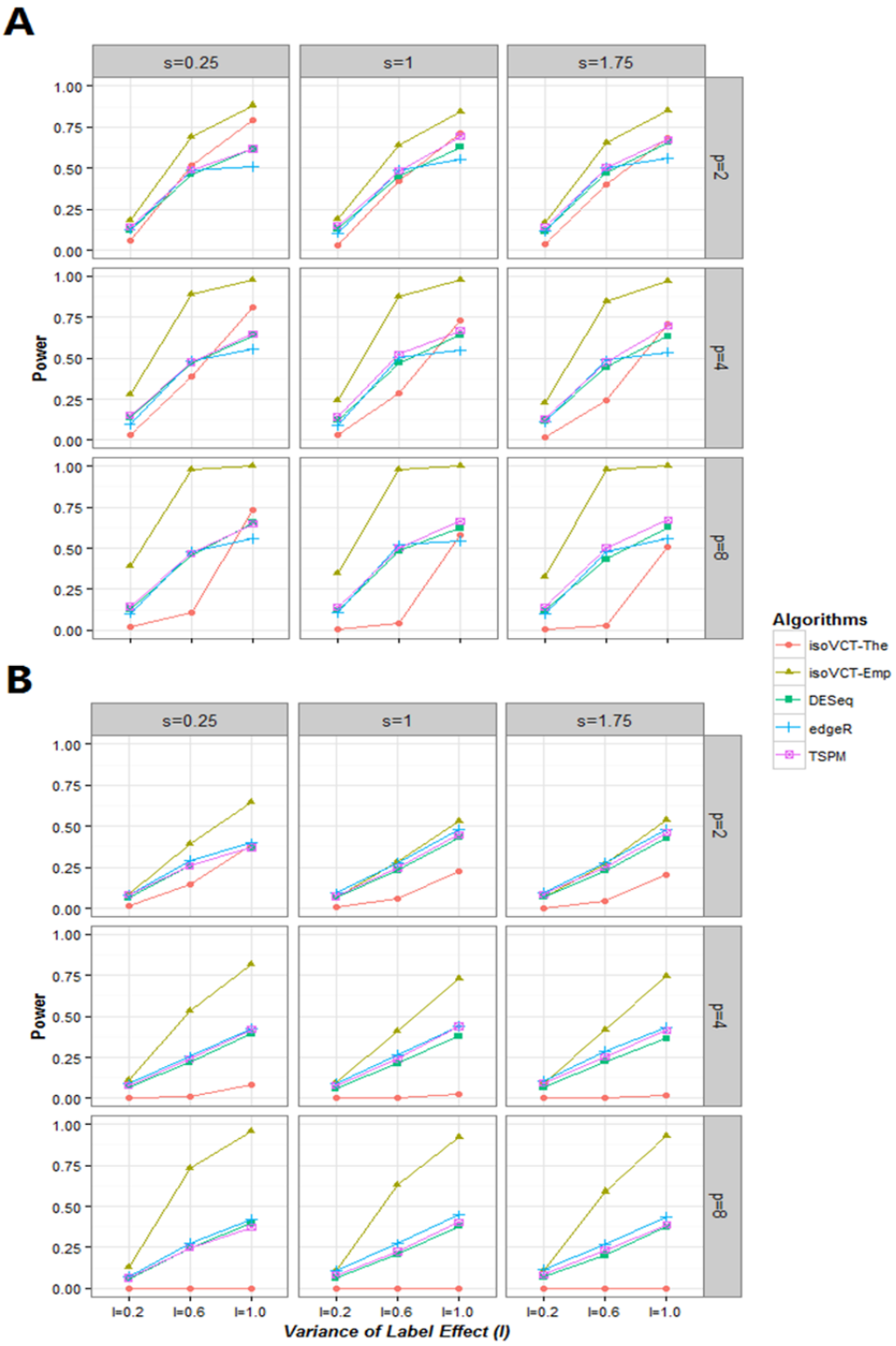

Figure 2 Plots of power of five algorithms in NB assumption. (A) The parameter setting is mu = 5 and phi = 2. (B) The parameter setting is mu = 5 and phi = 0.5.

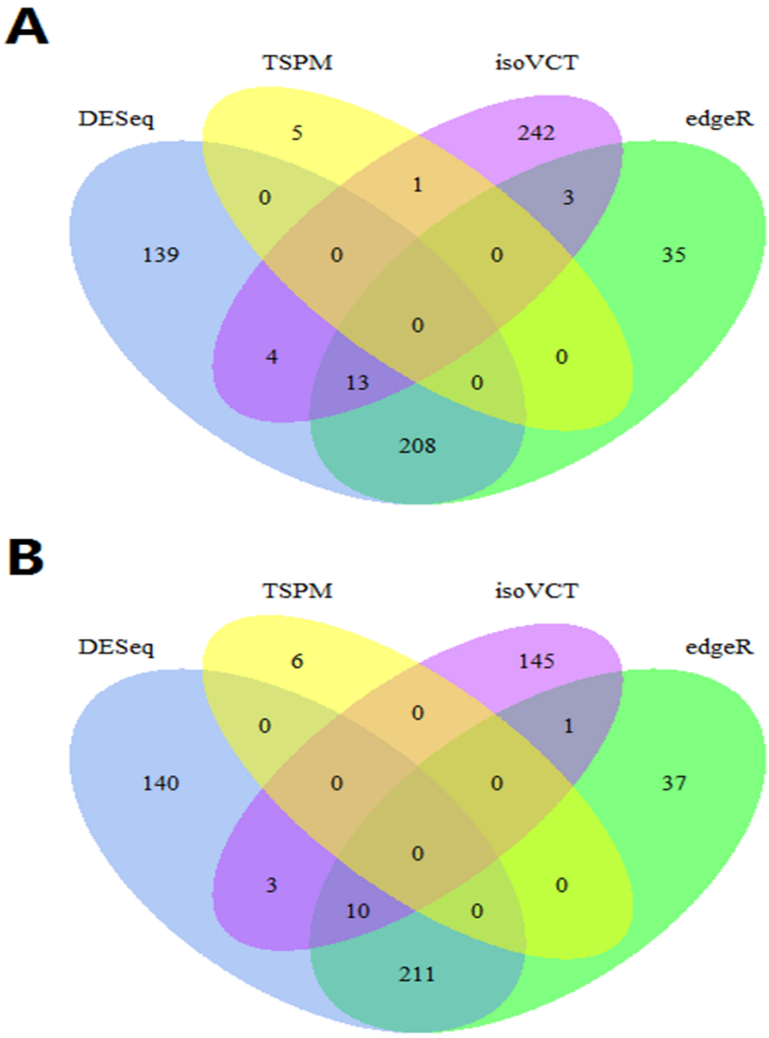

Figure 3 Venn diagrams of different methods in real data analysis. (A) Venn diagram of DESeq, edgeR, TSPM and isoVCT-Emp of non-normalized data; (B) Venn diagram of DESeq, edgeR, TSPM and isoVCT-The of non-normalized data.

especially for $l = 0.6$. The small sample size setting may cause the inverse proportion between p and the power of isoVCT. Nevertheless, the powers of five algorithms are high for the Poisson assumption.

In real data analysis, the Venn diagram of DE genes directly indicates the relationships among the four algorithms. The fact that DESeq and edgeR use the same distribution assumption and the same idea for testing likely results in the number of intersections representing a majority of the DE genes identified by these algorithms. TSPM defines the smallest number of DE genes. However, the use of different ideas leads to the different result of isoVCT. isoVCT-Emp specifically defines 242 heterogeneous or small effect size genes. For example, *CASP7* and *STAT6* are related to LUSC (*Dubey & Saini, 2015*; *Lee et al., 2009*).

Furthermore, isoVCT exhibits three innovations. First, isoVCT derives the empirical and theoretical distribution of the variance component score statistics in the framework of a GLMM and evaluates the performance of score statistics. Second, the random effect α represents the correlation of isoforms in the framework of the GLMM. Third, isoVCT further verifies the effectiveness of the set test in RNA-Seq data and supplies a new view of biological functions with isoform information.

However, isoVCT has some limitations. In the derivation of the score statistic, the random term α is regarded as a fixed parameter, which may cause isoVCT-The to be quite conservative. For the NB assumption, the power of isoVCT is very low in some settings, because the estimation of the dispersion parameter may be biased. The methods of weighted likelihood and quasi-likelihood likely overcome this drawback in estimating the NB mixed model (*Lund et al., 2012*).

In conclusion, isoVCT, a supplement to DESeq or edgeR, is powerful and robust at selecting small effect and heterogeneous genes.

ADDITIONAL INFORMATION AND DECLARATIONS

Funding

This work was supported by the National Natural Science Foundation of China (No. 81502888, 81473070 and 81373102), the Jiangsu Shuangchuang Plan, the Science and Technology Development Fund Key Project of Nanjing Medical University (2014NJMUZD003 and 2016NJMUZD014), and the Priority Academic Program Development of Jiangsu Higher Education Institutions (PAPD). The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Grant Disclosures

The following grant information was disclosed by the authors:
National Natural Science Foundation of China: 81502888, 81473070, 81373102.
Science and Technology Development Fund Key Project of Nanjing Medical University: 2014NJMUZD003, 2016NJMUZD014.

Competing Interests

The authors declare there are no competing interests.

Author Contributions

- Sheng Yang conceived and designed the experiments, performed the experiments, wrote the paper, prepared figures and/or tables.
- Fang Shao analyzed the data, contributed reagents/materials/analysis tools.
- Weiwei Duan and Yang Zhao reviewed drafts of the paper.
- Feng Chen conceived and designed the experiments.

Data Availability

The following information was supplied regarding data availability:

The raw data has been supplied as Supplemental Files.

Supplemental Information

Supplemental information for this article can be found online at http://dx.doi.org/10.7717/peerj.3797#supplemental-information.

REFERENCES

Al Seesi S, Tiagueu YT, Zelikovsky A, Mandoiu II. 2014. Bootstrap-based differential gene expression analysis for RNA-Seq data with and without replicates. *BMC Genomics* **15(Suppl 8)**:S2 DOI 10.1186/1471-2164-15-S8-S2.

Anders S, Huber W. 2010. Differential expression analysis for sequence count data. *Genome Biology* **11**: Article R106 DOI 10.1186/gb-2010-11-10-r106.

Dialsingh I, Austin SR, Altman NS. 2015. Estimating the proportion of true null hypotheses when the statistics are discrete. *Bioinformatics* **31**:2303–2309 DOI 10.1093/bioinformatics/btv104.

Dubey R, Saini N. 2015. STAT6 silencing up-regulates cholesterol synthesis via miR-197/FOXJ2 axis and induces ER stress-mediated apoptosis in lung cancer cells. *Biochimica Et Biophysica Acta* **1849**:32–43 DOI 10.1016/j.bbagrm.2014.10.002.

Gad AM, El Kholy RB. 2012. Generalized linear mixed models for longitudinal data. *International Journal of Probability and Statistics* **1**:41–47 DOI 10.5923/j.ijps.20120103.03.

Garber M, Grabherr MG, Guttman M, Trapnell C. 2011. Computational methods for transcriptome annotation and quantification using RNA-seq. *Nature Methods* **8**:469–477 DOI 10.1038/nmeth.1613.

Huang YT, Lin X. 2013. Gene set analysis using variance component tests. *BMC Bioinformatics* **14**:210 DOI 10.1186/1471-2105-14-210.

Ionita-Laza I, Lee S, Makarov V, Buxbaum JD, Lin X. 2013. Sequence kernel association tests for the combined effect of rare and common variants. *American Journal of Human Genetics* **92**:841–853 DOI 10.1016/j.ajhg.2013.04.015.

Kalsotra A, Cooper TA. 2011. Functional consequences of developmentally regulated alternative splicing. *Nature Reviews Genetics* **12**:715–729 DOI 10.1038/nrg3052.

Kanitz A, Gypas F, Gruber AJ, Gruber AR, Martin G, Zavolan M. 2015. Comparative assessment of methods for the computational inference of transcript isoform abundance from RNA-seq data. *Genome Biology* **16**: Article 150 DOI 10.1186/s13059-015-0702-5.

Koboldt DC, Steinberg KM, Larson DE, Wilson RK, Mardis ER. 2013. The next-generation sequencing revolution and its impact on genomics. *Cell* **155**:27–38 DOI 10.1016/j.cell.2013.09.006.

Lee WK, Kim JS, Kang H, Cha SI, Kim DS, Hyun DS, Kam S, Kim CH, Jung TH, Park JY. 2009. Polymorphisms in the Caspase7 gene and the risk of lung cancer. *Lung Cancer* **65**:19–24 DOI 10.1016/j.lungcan.2008.10.022.

Lehmann BD, Bauer JA, Chen X, Sanders ME, Chakravarthy AB, Shyr Y, Pietenpol JA. 2011. Identification of human triple-negative breast cancer subtypes and preclinical models for selection of targeted therapies. *Journal of Clinical Investigation* **121**:2750–2767 DOI 10.1172/JCI45014.

Li HD, Menon R, Omenn GS, Guan Y. 2014. The emerging era of genomic data integration for analyzing splice isoform function. *Trends in Genetics* **30**:340–347 DOI 10.1016/j.tig.2014.05.005.

Li J, Tibshirani R. 2013. Finding consistent patterns: a nonparametric approach for identifying differential expression in RNA-Seq data. *Statistical Methods in Medical Research* **22**:519–536 DOI 10.1177/0962280211428386.

Lin X. 1997. Variance component testing in generalised linear models with random effects. *Biometrika* **84**:309–326 DOI 10.1093/biomet/84.2.309.

Lund SP, Nettleton D, McCarthy DJ, Smyth GK. 2012. Detecting differential expression in RNA-sequence data using quasi-likelihood with shrunken dispersion estimates. *Statistical Applications in Genetics and Molecular Biology* **11**:307–314 DOI 10.1515/1544-6115.1826.

Modelska A, Quattrone A, Re A. 2015. Molecular portraits: the evolution of the concept of transcriptome-based cancer signatures. *Briefings in Bioinformatics* **16**:1000–1007 DOI 10.1093/bib/bbv013.

Network CGAR. 2012. Comprehensive genomic characterization of squamous cell lung cancers. *Nature* **489**:519–525 DOI 10.1038/nature11404.

Oshlack A, Robinson MD, Young MD. 2010. From RNA-seq reads to differential expression results. *Genome Biology* **11**: Article 220 DOI 10.1186/gb-2010-11-12-220.

Pan Q, Shai O, Lee LJ, Frey BJ, Blencowe BJ. 2008. Deep surveying of alternative splicing complexity in the human transcriptome by high-throughput sequencing. *Nature Genetics* **40**:1413–1415 DOI 10.1038/ng.259.

Robinson MD, McCarthy DJ, Smyth GK. 2010. edgeR: a bioconductor package for differential expression analysis of digital gene expression data. *Bioinformatics* **26**:139–140 DOI 10.1093/bioinformatics/btp616.

Trapnell C, Hendrickson DG, Sauvageau M, Goff L, Rinn JL, Pachter L. 2013. Differential analysis of gene regulation at transcript resolution with RNA-seq. *Nature Biotechnology* **31**:46–53 DOI 10.1038/nbt.2450.

Trapnell C, Williams BA, Pertea G, Mortazavi A, Kwan G, Van Baren MJ, Salzberg SL, Wold BJ, Pachter L. 2010. Transcript assembly and quantification by RNA-Seq reveals unannotated transcripts and isoform switching during cell differentiation. *Nature Biotechnology* **28**:511–515 DOI 10.1038/nbt.1621.

Wang Z, Gerstein M, Snyder M. 2009. RNA-Seq: a revolutionary tool for transcriptomics. *Nature Reviews Genetics* **10**:57–63 DOI 10.1038/nrg2484.

Wu C, Chen J, Kim J, Pan W. 2016. An adaptive association test for microbiome data. *Genome Medicine* **8**: Article 56 DOI 10.1186/s13073-016-0302-3.

Wu MC, Lee S, Cai T, Li Y, Boehnke M, Lin X. 2011. Rare-variant association testing for sequencing data with the sequence kernel association test. *American Journal of Human Genetics* **89**:82–93 DOI 10.1016/j.ajhg.2011.05.029.

Yang S, Guo L, Shao F, Zhao Y, Chen F. 2015. A systematic evaluation of feature selection and classification algorithms using simulated and real miRNA sequencing data. *Computational and Mathematical Methods in Medicine* **2015**: Article 178572 DOI 10.1155/2015/178572.