frontiers in Genetics

METHODS
published: 05 February 2021
doi: 10.3389/fgene.2021.616686

Identifying Differentially Expressed Genes of Zero Inflated Single Cell RNA Sequencing Data Using Mixed Model Score Tests

Zhiqiang He[1], Yueyun Pan[2], Fang Shao[1] and Hui Wang[3]**

[1] Department of Biostatistics, School of Public Health, Nanjing Medical University, Nanjing, China, [2] First Clinical Medical College, Nanjing Medical University, Nanjing, China, [3] Department of Maternal and Child Health, School of Public Health, Peking University Health Science Center, Beijing, China

Single cell RNA sequencing (scRNA-seq) allows quantitative measurement and comparison of gene expression at the resolution of single cells. Ignoring the batch effects and zero inflation of scRNA-seq data, many proposed differentially expressed (DE) methods might generate bias. We propose a method, single cell mixed model score tests (scMMSTs), to efficiently identify DE genes of scRNA-seq data with batch effects using the generalized linear mixed model (GLMM). scMMSTs treat the batch effect as a random effect. For zero inflation, scMMSTs use a weighting strategy to calculate observational weights for counts independently under zero-inflated and zero-truncated distributions. Counts data with calculated weights were subsequently analyzed using weighted GLMMs. The theoretical null distributions of the score statistics were constructed by mixed Chi-square distributions. Intensive simulations and two real datasets were used to compare edgeR-zinbwave, DESeq2-zinbwave, and scMMSTs. Our study demonstrates that scMMSTs, as supplement to standard methods, are advantageous to define DE genes of zero-inflated scRNA-seq data with batch effects.

Keywords: score test, generalized linear mixed model, zero inflation, observational weights, differential expression analyses, single cell RNA sequencing

OPEN ACCESS

Edited by:
Alfredo Pulvirenti, University of Catania, Italy

Reviewed by:
Tiejun Tong, Hong Kong Baptist University, Hong Kong
Shiquan Sun, Xi'an Jiaotong University, China

****Correspondence:***
Fang Shao
shaofang@njmu.edu.cn
Hui Wang
huiwang@bjmu.edu.cn

Specialty section:
This article was submitted to Computational Genomics, a section of the journal Frontiers in Genetics

Received: *13 October 2020*
Accepted: *14 January 2021*
Published: *05 February 2021*

Citation:
He Z, Pan Y, Shao F and Wang H (2021) Identifying Differentially Expressed Genes of Zero Inflated Single Cell RNA Sequencing Data Using Mixed Model Score Tests. Front. Genet. 12:616686. doi: 10.3389/fgene.2021.616686

INTRODUCTION

In modern biology, transcriptomics has been widely used to elucidate the molecular basis of biological processes and diseases (Van den Berge et al., 2018). Previous transcriptome sequencing techniques (bulk RNA-seq) (Wang et al., 2009) might obscure the cell type heterogeneity in different samples. Because of the resolution, bulk RNA-seq hardly defines the rare cells, such as stem cells and tumor cells. Single cell RNA sequencing (scRNA-seq) enables researchers to study characteristics of gene expression in the resolution of individual cells (Kolodziejczyk et al., 2015). scRNA-seq has been treated as an effective method to study cellular heterogeneity in complex biological systems, and is being applied by more researchers in various biological processes, such as stem cell development and differentiation, embryonic organ development, tumors, immunology, and neurology (Tang et al., 2009; McEvoy et al., 2011; Zeisel et al., 2015; Chu et al., 2016; Papalexi and Satija, 2018; Sun et al., 2019). Identifying differentially expressed (DE) genes is one of the most common analysis of

both bulk RNA-seq and of scRNA-seq analysis (Robinson et al., 2010; Van den Berge et al., 2017, 2018; Sun et al., 2018).

For bulk RNA-seq and scRNA-seq data, batch effects conventionally were treated as the non-biological differences that occurs when samples or cells are measured in distinct batches. The measure of transcriptome can be influenced by different environments for cells (Luecken and Theis, 2019). Various methods to correct batch effects and preserve biological variability have been presented. Some methods directly remove or correct batch effects using linear models (Johnson et al., 2007; Tung et al., 2017; Somekh et al., 2019). ComBat (Johnson et al., 2007) is an empirical Bayes method which takes batch effects into a linear regression model of gene expression. ComBat was recommended for batch correction when groups or cell types and state compositions between batches are consistent (Luecken and Theis, 2019). Mutual nearest neighbors (MNNs) (Haghverdi et al., 2018) and canonical correlation analysis (CCA) (Butler et al., 2018) remove batch effects using nonlinear models. A method comparison study showed ComBat was the best one for both bulk RNA-seq and scRNA-seq data (Büttner et al., 2019). For DE analysis, it was recommended that DE testing should be conducted on measure data with covariates including the batch information in the model design, not on batch corrected data (Luecken and Theis, 2019).

Some studies directly used traditional bulk RNA-seq DE methods (Krieg et al., 2018; Roerink et al., 2018; Li et al., 2019; Mehtonen et al., 2020). Limma-voom (Ritchie et al., 2015) applies weighted linear regression models for log-transformed count data. edgeR (Robinson et al., 2010; McCarthy et al., 2012) and DESeq2 (Love et al., 2014) model the gene expression count data based on generalized linear models (GLMs) under negative binomial (NB) distributions. It was demonstrated that NB models overestimated the dispersion parameter with excess zero counts, which influenced the power to DE analysis (Van den Berge et al., 2018). Different to bulk RNA-seq data, dropout events cause excess zeros for scRNA-seq read count data (Finak et al., 2015; Hashimshony et al., 2016). Therefore, zero inflation or an excess of zeros is a particular feature of scRNA-seq data, and it is not considered for these methods. SCDE (Kharchenko and Fan, 2019) and MAST (Finak et al., 2015; McDavid et al., 2019) model the redundant zeros of scRNA-seq data by zero inflation and hurdle models, respectively. Both zinbwave (Risso et al., 2018; Van den Berge et al., 2018) and zingeR (Van den Berge et al., 2017) estimates observational weights based on a zero-inflated negative binomial (ZiNB) model and downweight excess zeros followed by classical bulk RNA-seq DE tools (e.g., edgeR and DESeq2). The performance of two combinations, edgeR-zinbwave and DESeq2-zinbwave, outperform other DE methods (Van den Berge et al., 2018).

Here, based on isoVCT (Yang et al., 2017) and SMMATs (Chen et al., 2019), we implement a series of efficient methods, the single cell mixed model score tests (scMMSTs), to identify DE genes for defined cell types in scRNA-seq data considering batch effects and zero inflation. isoVCT, a DE method for bulk RNA-seq, uses a random effect to consider the heterogeneous isoform effects. In large-scale whole-genome sequencing (WGS) studies, SMMATs are powerful and computationally efficient variant set tests for continuous and binary traits, which integrates the burden test and SKAT (Wu et al., 2011) under the framework of generalized linear mixed models (GLMMs).

METHODS

Generalized Linear Mixed Models

For a single gene, we consider the following:

$$g(\mu_i) = \alpha + g_i\mathbf{B}_i\boldsymbol{\beta} + \mathbf{B}_i\mathbf{b},$$

where $g(\cdot)$ is a monotonic differentiable link function for GLMs, $\mu_i = E\left(y_i|g_i, \mathbf{B}_i, \mathbf{b}\right)$ denotes the mean of phenotype or count y_i for subject or cell i for a given gene with sample size n to the intercept α, g_i is the group, cluster or cell type covariate dummy variable binary value for subject i, $\mathbf{B}_i$ is the row vector of dummy variables values of the batch or individual covariate for subject i, $\boldsymbol{\beta}$ is the group effects associated with bathes and $\mathbf{b}$ is the batch effects. In the above equation, the group effects $\boldsymbol{\beta}$ are assumed to follow the normal distribution $N(\beta_0\mathbf{1}_p, \sigma^2_\beta\mathbf{I}_p)$, where $\mathbf{1}_p$ is the $p \times 1$ dimensional vector whose elements are all 1, $\mathbf{I}_p$ is the $p \times p$ dimensional identity matrix, β_0 and σ^2_β are mean and variance of the normal distribution and p is the number of batches. If $\sigma^2_\beta > 0$, group effects are associated with the batches. We assume the batch random effects $\mathbf{b} \sim N(\mathbf{0_p}, \sigma^2_b\mathbf{I}_p)$, where $\mathbf{0}_p$ is the $p \times 1$ dimensional vector whose elements are all 0 and σ^2_b is the variance. We consider the binomial, quasi-binomial, Poisson, quasi-Poisson, and NB distributions to model y_i. Binary phenotypes are commonly modeled by binomial and quasi-binomial distributions and counts are commonly modeled by Poisson, quasi-Poisson, and NB distributions.

For single cell RNA-seq data of a given gene, y_i is the count for cell i. We identify DE genes for each defined cell type in the form of one-against-others, so g_i, the cell type covariate for cell i, is binary. GLMMs under Poisson, quasi-Poisson and NB distributions are appropriate in this scenario.

Single Cell Mixed Model Score Tests

Testing $H_0 : \boldsymbol{\beta} = \mathbf{0}$ is equivalent to testing $H_0 : \beta_0 = 0$ and $\sigma^2_\beta = 0$. Under the null hypothesis, the reduced GLMM is as follows.

$$g(\mu_{0i}) = \alpha + \mathbf{B}_i\mathbf{b},$$

where $\mu_{0i} = E\left(y_i|\mu_0, b_i\right)$.

We construct a variance component score test statistic T derived by testing $H_0' : \sigma^2_\beta = 0$ under the assumption $\beta_0 = 0$. SMMAT-O was also derived in the same manner. Under H_0' with the assumption $\beta_0 = 0$, we have the same reduced null model as that under $H_0 : \boldsymbol{\beta} = \mathbf{0}$. Therefore, our derived test statistic T is applicable for testing H_0. The test statistic T is shown as follows.

$$T = \frac{\left(\mathbf{y} - \widehat{\boldsymbol{\mu}}_0\right)^T\widehat{\boldsymbol{\Phi}}\mathbf{G_B}\mathbf{G_B^T}\widehat{\boldsymbol{\Phi}}\left(\mathbf{y} - \widehat{\boldsymbol{\mu}}_0\right)}{\widehat{\tau}},$$

where $\mathbf{y} = \left(y_1\, y_2 \cdots y_n\right)^T$ is an $n \times 1$ vector of counts or phenotypes, $\widehat{\boldsymbol{\mu}}_0 = g^{-1}\left(\widehat{\alpha} + \mathbf{B}_i\widehat{\mathbf{b}}\right)$ is the estimated mean vector of

the reduced null model under H_0, $\widehat{\alpha}$ and $\widehat{\mathbf{b}}$ are estimates of the α and $\mathbf{b}$, $\mathbf{\Phi} = diag\left\{1/\left(1+\left(\widehat{\mu}_{0i}/\widehat{\theta}\right)\right)\right\}$ for the NB distribution with the estimated dispersion parameter $\widehat{\theta}$ and $\widehat{\mathbf{\Phi}} = \mathbf{I}_n$ for other distributions mentioned, $\mathbf{B} = \left(\mathbf{B}_1^T \mathbf{B}_2^T \cdots \mathbf{B}_n^T\right)^T$ is an $n \times p$ design matrix of group covariate dummy variables values, $\mathbf{G_B} = \left(g_1\mathbf{B}_1^T g_2\mathbf{B}_2^T \cdots g_n\mathbf{B}_n^T\right)^T$ is an $n \times p$ design matrix of interactions of group and batch covariates with the multiplication of corresponding dummy variables values and $\widehat{\tau}$ is the estimate of dispersion parameter τ for quasi distributions, which is 1 for the binomial, Poisson and NB distributions and is estimated by the residual deviance divided by the degree of freedom of the reduced null model for quasi-binomial and quasi-Poisson distributions.

The asymptotic distribution of the statistic T under H_0 is derived as follows. Following the theoretical results of mixed models (Harville, 1977; Breslow and Clayton, 1993; Santos Nobre and da Motta Singer, 2007; Chen et al., 2016), we have $\widehat{\mathbf{e}} = (\mathbf{y} - \widehat{\boldsymbol{\mu}}_0)/\sqrt{\widehat{\tau}}$ asymptotically following a n-dimensional multivariate normal distribution $\mathbf{M}VN_{\mathbf{n}}(\mathbf{0}, \widehat{\mathbf{D}}^{-1}\widehat{\mathbf{V}}\widehat{\mathbf{P}}\widehat{\mathbf{\Sigma}}\widehat{\mathbf{P}}\widehat{\mathbf{V}}\widehat{\mathbf{D}}^{-1})$ under H_0, where $\widehat{\mathbf{D}} = diag\left\{g'(\widehat{\mu}_{0i})\right\}$, whose diagonal elements are the first order derivative of the link function $g(\cdot)$ evaluated at $\widehat{\mu}_{0i}$, $\widehat{\mathbf{P}}$ is the $n \times n$ projection matrix of the reduced null model $\widehat{\mathbf{P}} = \widehat{\mathbf{\Sigma}}^{-1} - \widehat{\mathbf{\Sigma}}^{-1}\mathbf{1}_n\left(\mathbf{1}_n^{\mathrm{T}}\widehat{\mathbf{\Sigma}}^{-1}\mathbf{1}_n\right)^{-1}\mathbf{1}_n^{\mathrm{T}}\widehat{\mathbf{\Sigma}}^{-1}$ with $\widehat{\mathbf{\Sigma}} = \widehat{\mathbf{V}} + \widehat{\sigma}_b^2\mathbf{B}\mathbf{B}^{\mathbf{T}}$, $\widehat{\mathbf{V}} = diag\left\{\left(g'(\widehat{\mu}_{0i})\right)^2\widehat{Var}(y_i)\right\}$, the first order derivative function of the link function $g'(\cdot)$ and the estimated variance of y_i, $\widehat{Var}(y_i)$. For binomial and quasi-binomial distributions, $\left(g'\left(\widehat{\mu}_{0i}\right)\right)^2\widehat{Var}(y_i) = 1/[\widehat{\mu}_{0i}\left(1-\widehat{\mu}_{0i}\right)]$. For Poisson and quasi-Poisson distributions, $\left(g'\left(\widehat{\mu}_{0i}\right)\right)^2\widehat{Var}(y_i) = 1/\widehat{\mu}_{0i}$. For NB distributions, $\left(g'\left(\widehat{\mu}_{0i}\right)\right)^2\widehat{Var}\left(y_i\right) = \left(1/\widehat{\mu}_{0i}\right) + \left(1/\widehat{\theta}\right)$. Since $\widehat{\mathbf{P}}\widehat{\mathbf{\Sigma}}\widehat{\mathbf{P}} = \widehat{\mathbf{P}}$ and $\widehat{\mathbf{\Phi}} = \widehat{\mathbf{V}}^{-1}\widehat{\mathbf{D}}$, the asymptotic distribution can be simplified as $\mathbf{M}VN_{\mathbf{n}}\left(\mathbf{0}, \widehat{\mathbf{\Phi}}^{-1}\widehat{\mathbf{P}}\widehat{\mathbf{\Phi}}^{-1}\right)$. Therefore, under H_0, T, a quadratic form of $\widehat{\mathbf{e}}$, asymptotically follows a mixture Chi-square distribution $\sum_{i=1}^{p}\xi_i\chi_{1,i}^2$, where $\chi_{1,i}^2$ are independent Chi-square distributions with 1 degree of freedom, and ξ_i are the eigenvalues of $\mathrm{E} = \mathbf{G}_{\mathbf{B}}^{\mathbf{T}}\widehat{\mathbf{P}}\mathbf{G}_{\mathbf{B}}$. Notably, $\widehat{\mathbf{\Sigma}}$ in $\widehat{\mathbf{P}}$ has a simple structure which makes $\widehat{\mathbf{\Sigma}}^{-1}$ to be solved explicitly and E to be calculated efficiently. The p-value of the test can be calculated soon after the estimation of the reduced null model. More details of the computational efficiency of scMMSTs are discussed in section "Performance Evaluation". The estimation procedure of $\widehat{\mu}_{0i}$ is the same for binomial and quasi-binomial distribution pair and the Poisson and quasi-Poisson distribution pair. Thus, we implement quasi distributions to allow flexibility. In the followings, unless specified otherwise, "binomial" stands for both binomial and quasi-binomial and "Poisson" stands for both Poisson and quasi-Poisson.

There is zero inflation in scRNA-seq count data. Therefore, following the idea of ZINB-WaVE, a weighting strategy is implemented. Firstly, observational weights are calculated for all counts independently with details shown in sections "Zero-Inflated and Zero-Truncated Distributions for Counts" and "Calculations of Observational Weights for scMMSTs." Afterward, counts data with calculated weights are analyzed under the weighted GLMMs. Accordingly, a weighted version test statistic T_w for scMMSTs is proposed as follows with above notations.

$$T_w = \frac{\left(\mathbf{y} - \widehat{\boldsymbol{\mu}}_0\right)^T\widehat{\mathbf{\Phi}}\mathbf{W}\mathbf{G}_{\mathbf{B}}\mathbf{G}_{\mathbf{B}}^{\mathbf{T}}\mathbf{W}\widehat{\mathbf{\Phi}}\left(\mathbf{y} - \widehat{\boldsymbol{\mu}}_0\right)}{\widehat{\tau}},$$

where $\mathbf{W} = diag\{w_i\}$ and w_i is the given weights for count y_i. The estimation is based on the weighted GLLMs for the reduced null model. We denote $\mathbf{1}_{\mathbf{w},n} = \mathbf{W}^{\frac{1}{2}}\mathbf{1}_n$, $\mathbf{B}_{\mathbf{w}} = \mathbf{W}^{\frac{1}{2}}\mathbf{B}$, $\widehat{\mathbf{V}}_{\mathbf{w}} = \mathbf{W}^{-\frac{1}{2}}\widehat{\mathbf{V}}\mathbf{W}^{-\frac{1}{2}}$, $\widehat{\mathbf{\Sigma}}_{\mathbf{w}} = \widehat{\mathbf{V}} + \widehat{\sigma}_b^2\mathbf{B}_{\mathbf{w}}\mathbf{B}_{\mathbf{w}}^{\mathbf{T}}$, $\widehat{\widetilde{\mathbf{\Sigma}}}_{\mathbf{w}} = \mathbf{W}^{-\frac{1}{2}}\widehat{\mathbf{\Sigma}}_{\mathbf{w}}\mathbf{W}^{-\frac{1}{2}}$ and $\widehat{\mathbf{P}}_{\mathbf{w}} = \mathbf{W}^{\frac{1}{2}}\widehat{\mathbf{\Sigma}}_{\mathbf{w}}^{-1}\mathbf{W}^{\frac{1}{2}} - \mathbf{W}^{\frac{1}{2}}\widehat{\mathbf{\Sigma}}_{\mathbf{w}}^{-1}\mathbf{1}_{\mathbf{w},n}\left(\mathbf{1}_{\mathbf{w},n}^{\mathrm{T}}\widehat{\mathbf{\Sigma}}_{\mathbf{w}}^{-1}\mathbf{1}_{\mathbf{w},n}\right)^{-1}\mathbf{1}_{\mathbf{w},n}^{\mathrm{T}}\widehat{\mathbf{\Sigma}}_{\mathbf{w}}^{-1}\mathbf{W}^{\frac{1}{2}} = \widehat{\widetilde{\mathbf{\Sigma}}}_{\mathbf{w}}^{-1} - \widehat{\widetilde{\mathbf{\Sigma}}}_{\mathbf{w}}^{-1}\mathbf{1}_n\left(\mathbf{1}_n^{\mathrm{T}}\widehat{\widetilde{\mathbf{\Sigma}}}_{\mathbf{w}}^{-1}\mathbf{1}_n\right)^{-1}\mathbf{1}_n^{\mathrm{T}}\widehat{\widetilde{\mathbf{\Sigma}}}_{\mathbf{w}}^{-1}$. Based on the theoretical results of weighted GLMMs (Harville, 1977; Breslow and Clayton, 1993; Santos Nobre and da Motta Singer, 2007; Chen et al., 2016), if H_0 and $\mathbf{W}$ are true, we have $\widehat{\mathbf{e}}$ asymptotically normally distributed as $\mathbf{M}VN_{\mathbf{n}}(\mathbf{0}, \widehat{\mathbf{D}}^{-1}\widehat{\mathbf{V}}_{\mathbf{w}}\widehat{\mathbf{P}}_{\mathbf{w}}\widehat{\widetilde{\mathbf{\Sigma}}}_{\mathbf{w}}\widehat{\mathbf{P}}_{\mathbf{w}}\widehat{\mathbf{V}}_{\mathbf{w}}\widehat{\mathbf{D}}^{-1})$. Since $\widehat{\mathbf{P}}_{\mathbf{w}}\widehat{\widetilde{\mathbf{\Sigma}}}_{\mathbf{w}}\widehat{\mathbf{P}}_{\mathbf{w}} = \widehat{\mathbf{P}}_{\mathbf{w}}$, $\widehat{\mathbf{\Phi}} = \widehat{\mathbf{V}}^{-1}\widehat{\mathbf{D}}$ and $\widehat{\mathbf{D}}^{-1}\widehat{\mathbf{V}}_{\mathbf{w}} = \widehat{\mathbf{D}}^{-1}\mathbf{W}^{-\frac{1}{2}}\widehat{\mathbf{V}}\mathbf{W}^{-\frac{1}{2}} = \widehat{\mathbf{\Phi}}^{-1}\mathbf{W}^{-1}$, where $\mathbf{W}^{-\frac{1}{2}}$, $\widehat{\mathbf{D}}^{-1}$, $\widehat{\mathbf{V}}$ are diagonal matrices, the asymptotic distribution can be simplified as $\mathbf{M}VN_{\mathbf{n}}(\mathbf{0}, \widehat{\mathbf{\Phi}}^{-1}\mathbf{W}^{-1}\ \widehat{\mathbf{P}}_{\mathbf{w}}\mathbf{W}^{-1}\widehat{\mathbf{\Phi}}^{-1})$. If H_0 and $\mathbf{W}$ are true, T_w, a quadratic form of $\widehat{\mathbf{e}}$, asymptotically follows a mixture Chi-square distribution $\sum_{i=1}^{p}\xi_i\chi_{1,i}^2$, where $\chi_{1,i}^2$ are independent Chi-square distributions with 1 degree of freedom, and ξ_i are the eigenvalues of $\mathrm{E}_{\mathbf{w}} = \mathbf{G}_{\mathbf{B}}^{\mathbf{T}}\widehat{\mathbf{P}}_{\mathbf{w}}\mathbf{G}_{\mathbf{B}}$. Note that $\widehat{\mathbf{\Sigma}}_{\mathbf{w}}$ in $\widehat{\mathbf{P}}_{\mathbf{w}}$ does not have the simple structure of $\widehat{\mathbf{\Sigma}}$, which makes it hard to analytically and explicitly solve $\widehat{\mathbf{\Sigma}}_{\mathbf{w}}^{-1}$. Therefore, we propose $\mathrm{E}'_{\mathbf{w}} = \mathbf{G}_{\mathbf{B}}^{\mathbf{T}}\mathbf{W}\widehat{\mathbf{P}}\mathbf{W}\mathbf{G}_{\mathbf{B}}$ to approximate $\mathrm{E}_{\mathbf{w}}$ for simplicity and efficiency, where we treat $\widehat{\mathbf{e}}$ as it is estimated by GLMMs without weights. Calculated weights are 1 for nonzero counts and between 0 and 1 for zero counts. Thus, this approximation performs worse when there are more redundant zeros, which might influence the performance of scMMSTs.

Zero-Inflated and Zero-Truncated Distributions for Counts

Zero-Inflated Distributions for Counts

A zero-inflated distribution for counts is a mixture distribution with two components, which are a point mass at zero and a conventional random variable distribution for counts, e.g., Poisson and NB distributions. The probability mass function (pmf) of a zero-inflated distribution for counts is as follows.

$$f_{ZI}\left(y; \boldsymbol{\theta}, \pi\right) = \pi\delta_0\left(y\right) + (1-\pi)f\left(y; \boldsymbol{\theta}\right),\ \forall y \in \mathbb{N},$$

where $\pi \in [0, 1]$ indicates the probability of zero inflation, $\delta_0(\cdot)$ the Dirac function, $f(\cdot; \theta)$ the pmf of a conventional distribution with parameter vector $\boldsymbol{\theta}$. The observational weights of the counts can be calculated under a zero-inflated distribution model as the conditional probability that a given count y belongs to the conventional distribution with parameter estimates $\widehat{\boldsymbol{\theta}}$, $\widehat{\pi}$:

$$w = \frac{(1-\widehat{\pi})f\left(y; \widehat{\boldsymbol{\theta}}\right)}{f_{ZI}\left(y; \widehat{\boldsymbol{\theta}}, \widehat{\pi}\right)}.$$

Note that w is 1 for nonzero counts and $\in (0, 1)$ for zeros counts. All the weights for counts under the conventional distribution

are 1. Under a zero-inflated distribution, we take the weights of nonzero counts remain 1 and downweight zero counts from 1 to the conditional probability that a given count y belongs to the conventional distribution. Counts with observational weights are subsequently analyzed under the weighted version of models for the conventional distribution. In ZINB-WaVE, this weighting strategy is applied and the above formula is applied to calculate observational weights under the ZiNB distribution (Van den Berge et al., 2018).

Zero-Truncated Distributions for Counts

A zero-truncated distribution for counts is a distribution for counts with random variable values truncated at zero, i.e., only counts larger than zero can be observed. In the followings, we refer to zero-truncated distributions as truncated distributions for short. The pmf of a truncated distribution for counts is as follows.

$$f_{Tr}\left(y;\boldsymbol{\theta}\right)=\frac{f\left(y;\boldsymbol{\theta}\right)}{P_f(t>0;\boldsymbol{\theta})}=\frac{f\left(y;\boldsymbol{\theta}\right)}{\sum_{t=1}^{+\infty}f\left(t;\boldsymbol{\theta}\right)},\ \forall y\in\mathbb{N}_+,$$

where $f\left(\cdot;\boldsymbol{\theta}\right)$ denotes the pmf of a conventional distribution for counts with parameter vector$\boldsymbol{\theta}$. The observational weights of nonzero counts are 1 and weights of zero counts can be calculated under a truncated distribution model as following:

$$w=\frac{n_1 f\left(y=0;\widehat{\boldsymbol{\theta}}\right)}{n_0\sum_{t=1}^{+\infty}f\left(t;\widehat{\boldsymbol{\theta}}\right)},$$

where n_1 is the number of nonzero counts, n_0 is the number of the zero counts in the whole sample and $\widehat{\boldsymbol{\theta}}$ is the parameter vector estimate.

The derivation of the above formula is as follows. Nonzero counts follow the truncated distribution with parameter $\boldsymbol{\theta}$ which is the also the parameter for the corresponding conventional distribution. Therefore, the probability of zero counts is estimated as $f\left(y=0;\widehat{\boldsymbol{\theta}}\right)$. All the weights for counts under the conventional distribution are 1. However, since excess zeros are presented, the observational weights of nonzero counts remain 1 and zero counts are reweighted from 1 to w, so that $\frac{w\cdot n_0}{w\cdot n_0+1\cdot n_1}=f\left(y=0;\widehat{\boldsymbol{\theta}}\right)$. The resulting formula for observational weights w is derived by solving the equation. Counts are then analyzed with observational weights calculated under the weighted version of models for the conventional distribution.

Calculations of Observational Weights for scMMSTs

In ZINB-WaVE, the weighting strategy shown in the previous section is applied and observational weights are estimated by the ZiNB regression (Van den Berge et al., 2018). For our methods, the truncated Poisson (TrPois), zero-inflated Poisson (ZiPois), truncated negative binomial (TrNB), and ZiNB distributions are considered. Following the weighting strategy mentioned and $H_0:\boldsymbol{\beta}=\mathbf{0}$, we estimate parameters for counts in each batch and calculate the weights accordingly using the formulas in section "Zero-Inflated and Zero-Truncated Distributions for Counts" for simplicity with the assumption of no group effects.

For zero-inflated distributions, weights are the conditional probabilities that a count y belongs to the corresponding conventional distribution. We directly use ZINB-WaVE for the ZiNB distribution, and implement the algorithm in Appendix A of the paper (Böhning et al., 1999) for the ZiPois distribution. In ZINB-WaVE, no mixed models are involved. Thus, we treat batch effects as fixed effects in the ZiNB regression without group effects to calculate weights using all counts data, when using ZINB-WaVE. For TrPois distribution, since the pmf $f_{TrPois}\left(y\right)=\frac{f_{Pois}\left(y\right)}{1-e^{-\lambda}}=\frac{\lambda^y e^{-\lambda}}{y!1-e^{-\lambda}}$, we can derive the method of moment estimate and maximum likelihood estimate $\widehat{\lambda}$ and they are identical by numerically solve the equation $\frac{\widehat{\lambda}}{1-e^{-\widehat{\lambda}}}=\bar{y}$, where $\bar{y}$ is the sample mean for the truncated sample. For each batch, the weights are $w_i=\frac{n_1 e^{-\widehat{\lambda}}}{n_0\left(1-e^{-\widehat{\lambda}}\right)}$ for a zero count and $w_i=1$ for nonzero y_i, where n_1 is truncated sample size for the batch and n_0 is the number of the zero counts in the batch. TrPois and ZiPois perform very close to each other. For TrNB distribution, we implement the formulas in section "Results" of the paper (Rider, 1955) to estimate the mean parameter μ and the dispersion parameter θ for each batch. The common dispersion parameter θ is estimated by the harmonic mean of the estimated $\widehat{\theta}$ for each batch. However, this algorithm is not robust for small θ ($\theta<2$, based on simulations). The weights are $w_i=\frac{n_1\left(\widehat{\theta}/\left(\widehat{\theta}+\widehat{\mu}\right)\right)^{\widehat{\theta}}}{n_0\left(1-\left(\widehat{\theta}/\left(\widehat{\theta}+\widehat{\mu}\right)\right)^{\widehat{\theta}}\right)}$ for zero counts in each batch, where $\widehat{\theta}$ and $\widehat{\mu}$ are respectively the estimated dispersion and mean parameters for the NB distribution using counts in the batch, and $w_i=1$ for nonzero y_i for each corresponding batch.

After weights are calculated, counts data with weights are analyzed under weighted GLMMs shown in section "Single Cell Mixed Model Score Tests." Note that weights are calculated independently of GLMMs. Theoretically, the weights are 1 under conventional distributions. The calculated observational weights for nonzero counts remain 1. If there are calculated weights of zero counts far from 1 and closer to 0, it indicates that there are excess zeros. If calculated weights of zero counts are close to 1, the results for conventional distributions are similar to those considering zero inflation. In ZiNB-Wave, weights are calculated through the ZiNB regressions on all counts. However, the weights for TrPois, ZiPois, and TrNB are calculated using counts for each batch with smaller sample sizes. Therefore, although the calculation of weights for TrPois, ZiPois and TrNB is easier to implement and time saving, it is less accurate and less reliable than that for ZiNB-Wave and the performances of scMMSTs are affected.

Performance Evaluation

Performances of DE methods considered are assessed in terms of the per-comparison error rate (PCER), which refers to type I error rate (i.e., the proportion of false positives), line plots of the true positive rate (TPR) vs. the false discovery proportion (FDP) and the areas under the receiver operating characteristic (ROC)

curves [i.e., the TPR vs. the false positive rate (FPR) curves] (AUCs) with definitions as follows.

$$TPR = \frac{TP}{P},\ FPR = \frac{FP}{N}, FDP = \frac{FP}{\max(1, FP + TP)'}$$

where we use the following abbreviations for empirical quantities: FP (the number of false positives), TP (the number of true positives), N (the number of negative samples), P (the number of positive samples). FDP-TPR curves for adjusted *p*-values are plotted by *iCOBRA* Bioconductor R package (version 1.12.1) (Soneson and Robinson, 2016) and AUCs for adjusted *p*-values are calculated by *pROC* R package (version 1.16.2) (Robin et al., 2011). Unless otherwise stated, the adjusted *p*-values for all DE methods considered are calculated by the Benjamini and Hochberg method (Benjamini and Hochberg, 1995) for FDR control.

Comparison Methods

The 12 methods considered for comparisons are Poisson, TrPois, ZiPois, NB, TrNB, NB-zinb, DESeq2, DESeq2-zinb, edgeR, edgeR-zinb, limma-voom, and MAST. The first six methods are our implemented methods of scMMSTs s under GLMMs assumptions and the last six methods are the state-of-the-art DE methods, where Tr, Zi, Pois, NB, and zinb are abbreviations of truncated, zero-inflated, Poisson, ZINB-WaVE, respectively. We follow the implementations of the last six DE methods above in the *zinbwave* paper (Van den Berge et al., 2018) and the R packages used are *edgeR* (version 3.28.1), *DESeq2* (version 1.26.0), *limma* (version 3.42.2), *MAST* (version 1.12.0), and *zinbwave* (version 1.8.0), which was developed to deal with zero inflation for scRNA-seq data by a weighting strategy and was used in edgeR-zinb, DESeq2-zinb, and NB-zinb. The binomial distribution scMMST is implemented, however, not covered in the simulations and real data analysis since only methods for count data are considered in thisarticle.

The implementations of scMMSTs are available in **Supplementary Data S1**. Codes for simulations and real data analysis are partially based on the GitHub repositories[1,2] of papers (Yang et al., 2017; Van den Berge et al., 2018) and the *GMMAT* R package (version 1.3.0) (Chen et al., 2016, 2019). R packages *doParallel* (version 1.0.15) (Corporation and Weston, 2019) and *BiocParallel* (version 1.20.1) (Morgan et al., 2019) are used for parallel computation. The reduced null model is estimated by *lme4* R package (version 1.1.23) and *p*-values are calculated by *CompQuadForm* R package (version 1.4.3). Simulated single cell datasets are generated by *splatter* R package (version 1.10.1) (Zappia et al., 2017). Additionally, the code to reproduce all analyses, figures and tables reported in this manuscript is attached in **Supplementary Data S1**.

Simulations

We perform simulations to evaluate performances of scMMSTs, which are our methods of association tests under the proposed GLMMs, comparing with state-of-art DE methods under a range of scenarios. We simulate the scRNA-seq data based on GLMMs directly and by the R package *splatter*. *Splatter* can directly estimate model parameters for real scRNA-seq data and generate quality controlled simulated mock datasets with DE genes easily and can add batch effects, which are not associated with group effects, to the simulated data. The simulated number of genes for one dataset by *splatter* and GLMMs is 10,000 and the number of cells is 250 with balanced two groups and five batches. In the DE genes simulations, the proportion of the DE genes is set to be 0.1.

Additional parameters of *splatter* simulations, batch.facLoc–batch factor location, batch.facScale–batch factor scale, and out.prob–the expression outlier probability, are set to be 0.5. For DE gene simulations, de.facLoc, DE factor location, is set to 2 and de.facScale, DE factor scale, is set to be 0.5.

The procedure to simulate datasets based on the proposed GLMMs is as follows. We assume that the scRNA-seq count data follow Poisson and NB distributions and generate y_i based on the GLMM shown with the parameters setting and generate a Bernoulli random variable z_i with parameter $\pi_i = logit^{-1}(\mu_\pi + \mathbf{B}_i\mathbf{b})$. Larger values of parameter μ_π causes smaller baseline proportions of zeros. If $z_i = 0$, then $y_i = 0$, and y_i remains the same otherwise. The parameter settings for simulations are based on the real data analysis and references (Yang et al., 2017). Seven parameters are considered: the variance of the batch or individual effects $\mathbf{b}(\sigma_b^2)$, the variance of the group or cell type effects $\boldsymbol{\beta}(\sigma_\beta^2)$, the baseline group effect (β_0), the number of batches (p), the dispersion parameter ($\theta = 1/\phi$) for NB distributions and the intercepts (μ_0) and (μ_π) for the GLMM and logstic regression for excess zeros, respectively. σ_b^2 shows the heterogeneity of batch effects in different batches. σ_β^2 shows the heterogeneity of group effects in different batches. β_0 shows the baseline group effect. The larger the $|\beta_0|$, the larger the baseline group effect is. Other parameters describe the features of the gene expression and zero inflation. σ_b^2 is set to be 0.25 and σ_β^2 varies in 0, 0.01, 0.25, and 1. β_0 varies in 0, 0.01, 0.1, 0.3, and 0.5. θ varies in 0.5, 1, and 2. μ_π varies in -1, 0, and 2. $p = 5$ and $\mu_0 = 5$.

Real Data Sets

Usoskin Dataset

This scRNA-seq dataset contains mouse neuronal cells in the dorsal root ganglion (Usoskin et al., 2015). The processed expression values were downloaded from the Github respiratory[3] of the *zinbwave* paper. Following the process procedures given in the *zinbwave* paper, the authors considered 622 cells with a classification of 11 neuronal cell-types, which were denoted as NF1 to NF5, NP1 to NP3, PEP1, PEP2 and TH. Genes with less than 20 counts were removed and a total of 12,132 genes are considered for the following analyses with 68% zero counts. The authors showed the existence of a batch effect related to the picking session for the cells. Thus, the picking session covariate (with values Cold, RT-1, and RT-2) in this dataset was considered as a batch covariate for real data analysis. The batch effect was associated with expression measures and the relationship between zero inflation and sequencing depth, which was shown

[1]https://github.com/biostat0903/RNAseq-Data-Analysis

[2]https://github.com/statOmics/zinbwaveZinger

[3]https://github.com/statOmics/zinbwaveZinger/tree/master/datasets

in **Figure 5** of the *zinbwave* paper (Hicks et al., 2015; Van den Berge et al., 2018). We repeated the results of Figures 5A,B of the *zinbwave* paper in **Supplementary Figures S1A,B**. There is a large variation in the depth of sequencing among batches, which weaken the overall association with zero inflation when pooling cells across batches (**Supplementary Figure S1A**). Zero inflation was also identified for the Usoskin dataset. Histograms of observational weights for nonzero counts, which were calculated by the ZINB-WaVE model including the cell type as a covariate with and without the batch effect as fixed effects, are shown in **Supplementary Figure S1B**. Calculated weights of nonzero counts with and without the batch effect both have high modes near zero. This suggests zero inflation in the Usoskin dataset. The real data analysis of the processed Usoskin dataset was done to identify DE genes for defined 11 cell types vs. the rest. Simulated datasets based on this dataset were generated by *spaltter* with estimated corresponding parameters. For a null dataset without DE genes, we created 10,000 genes, 250 cells, five balanced batches and two balanced groups for cells. Twelve methods were implemented to identify DE genes between the two groups for each of the 30 simulated null data sets. A gene was declared to be DE if its unadjusted p-value was less than or equal to 0.05. Declared DE genes were false positives for these simulated null datasets. The empirical PCER of each method was calculated as the proportion of declared DE genes and was compared to the 0.05 nominal PCER.

Tung Dataset

This scRNA-seq dataset is for induced pluripotent stem cells from three individuals from HapMap (Tung et al., 2017). Following the *splatter* paper (Zappia et al., 2017), the matrix of molecules (UMIs) was treated as counts and was used directly. This dataset is available from GEO (accession GSE77288)[4] and the Github respiratory[5] of the *splatter* paper. No batch information is available for this dataset. Genes with less than 20 counts were removed and a total of 14,893 genes with 864 cells containing 44% zero counts were considered. Zero inflation was identified for the Tung dataset. Histograms of observational weights of nonzero counts of two filtered datasets (18,726 genes with more than 0 count and 14,893 genes with more than 19 counts, respectively), which were calculated by the ZINB-WaVE model, are shown in **Supplementary Figures S1C,D**. There are moderate proportion s of calculated weights of nonzero counts close to zero. This suggests zero inflation in the Tung dataset. Comparing to the Usoskin dataset, the Tung dataset is less zero inflated. We generated 30 simulated null datasets and identified DE genes using the same procedures for the Usoskin dataset with *spaltter*.

RESULTS

Method Overview

Single cell mixed model score tests are computationally efficient DE analysis tools for scRNA-seq data considering batch effects and zero inflation. Bath effects are estimated as random effects under the reduced null models of GLMMs. A weighting strategy is implemented to characterize excess zeros. The score statistics are derived on theoretical asymptotic distributions. First, we estimated normalization factors of count matrix by the function *calcNormFactors* in *edgeR* after counts per million (CPM) normalization. Second, the estimation of the observational weights is efficient. We use *zinbwave* to fit NB-zinb which might be the most time-consumed assumption. Third, we use *lme4* for the estimation, the most efficient method to fit GLMM, to estimation the parameters in the null hypothesis (Eddelbuettel and François, 2011; Eddelbuettel, 2013; Eddelbuettel and Balamuta, 2017). Considering the real data, the estimation procedure of mixed model is not related to the number of groups or cell types. Compared to the traditional estimation procedure, scMMSTs use three strategies to decrease memory usage and computation time. First, scMMSTs do not need to store $n \times n$ matrices $\widehat{\mathbf{P}}$ and $\widehat{\mathbf{\Sigma}}$ explicitly. The p-value is efficiently calculated by *CompQuadForm* with eigenvalues of E or $\mathrm{E}'_{\mathbf{w}}$, which is only a $p \times p$ matrix. Second, scMMSTs use an analytical form to calculate the inverse of $\widehat{\mathbf{\Sigma}}$ which might be the most time consumption procedure in the estimation of T or T_w. Third, scMMSTs is implemented for parallel computing. Therefore, although more complicated models GLMMs are considered, scMMSTs are computationally affordable compared to other DE methods.

Simulations by Real Datasets and *Splatter*

Simulated datasets generated by the *splatter* used parameters estimated from two publicly available real scRNA-seq datasets, the Usoskin (Usoskin et al., 2015) and Tung (Tung et al., 2017) datasets.

The FPR control was assessed by the PCER. Results are shown in **Figure 1**. For the Usoskin dataset, the estimated common dispersion parameter value of biological coefficient of variation (BCV) was $\hat{\phi} = 1/\widehat{\theta} = 1.89$. TrNB and Poisson failed to control the FPR. The PCERs of NB-zinb, DESeq2, edgeR-zinb, and edgeR were a little inflated. DESeq2-zinb and MAST controlled the FPRs with large variability, especially for MAST. Other methods were a little conservative with PCERs smaller than the nominal level 0.05. For the Tung dataset, the estimated common dispersion parameter value of BCV was $\hat{\phi} = 1/\widehat{\theta} = 0.11$. Poisson failed to control the FPR. The PCERs of TrNB and edgeR were a little inflated. Other methods conservatively controlled FPRs, especially for NB, DESeq2-zinb, and NB-zinb. We treated "NA" p-values of DE methods as 1, thus, there are peak bars at 1 for some methods in the unadjusted p-value histograms shown in **Figures 1B,D**. In summary, standard DE methods can control the FPRs and scMMSTs except Poisson and TrNB can conservatively control the FPRs. FPRs of scMMSTs increase as the dispersion parameter θ decreases.

False discovery proportion-true positive rate curves for adjusted p-values are shown in **Figure 2**. For the Usoskin dataset, bulk RNA-seq DE methods are shown to

[4]https://github.com/jdblischak/singleCellSeq

[5]https://github.com/Oshlack/splatter-paper

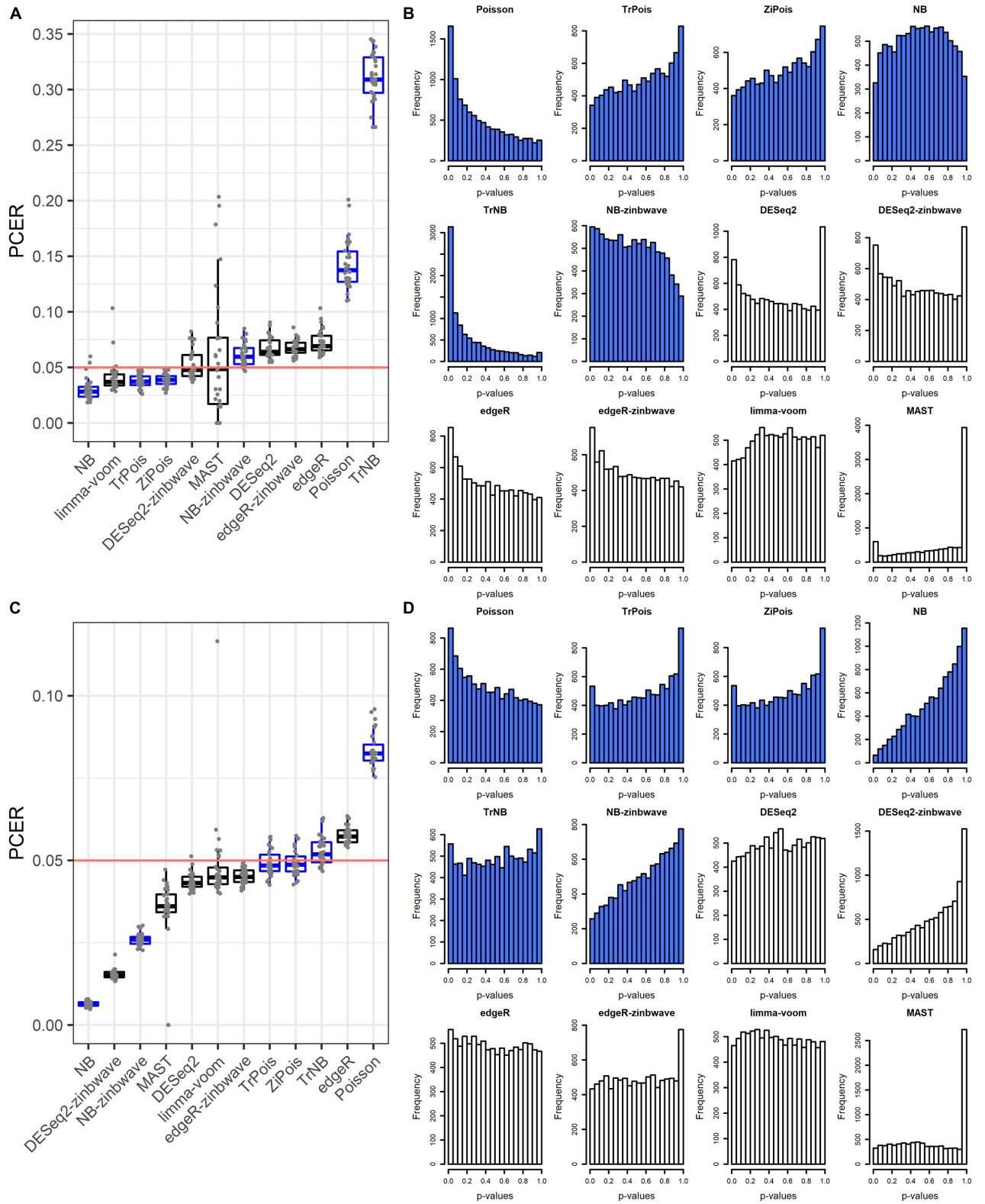

FIGURE 1 | False positive rate control on simulated null Usoskin datasets and Tung datasets. **(A)** Boxplot of PCER for 30 simulated null Usoskin datasets generated by *splatter* for each of 12 DE methods. scMMSTs are marked in blue. **(B)** Histogram of uncorrected *p*-values for one dataset in panel **A**. **(C)** Boxplot of PCER for 30 simulated null Tung datasets generated by *splatter* for each of 12 DE methods. scMMSTs are marked in blue. **(D)** Histogram of uncorrected *p*-values for one dataset in panel **C**. PCER, per-comparison error rate; DE, differential expression; scMMST, single cell mixed model score test.

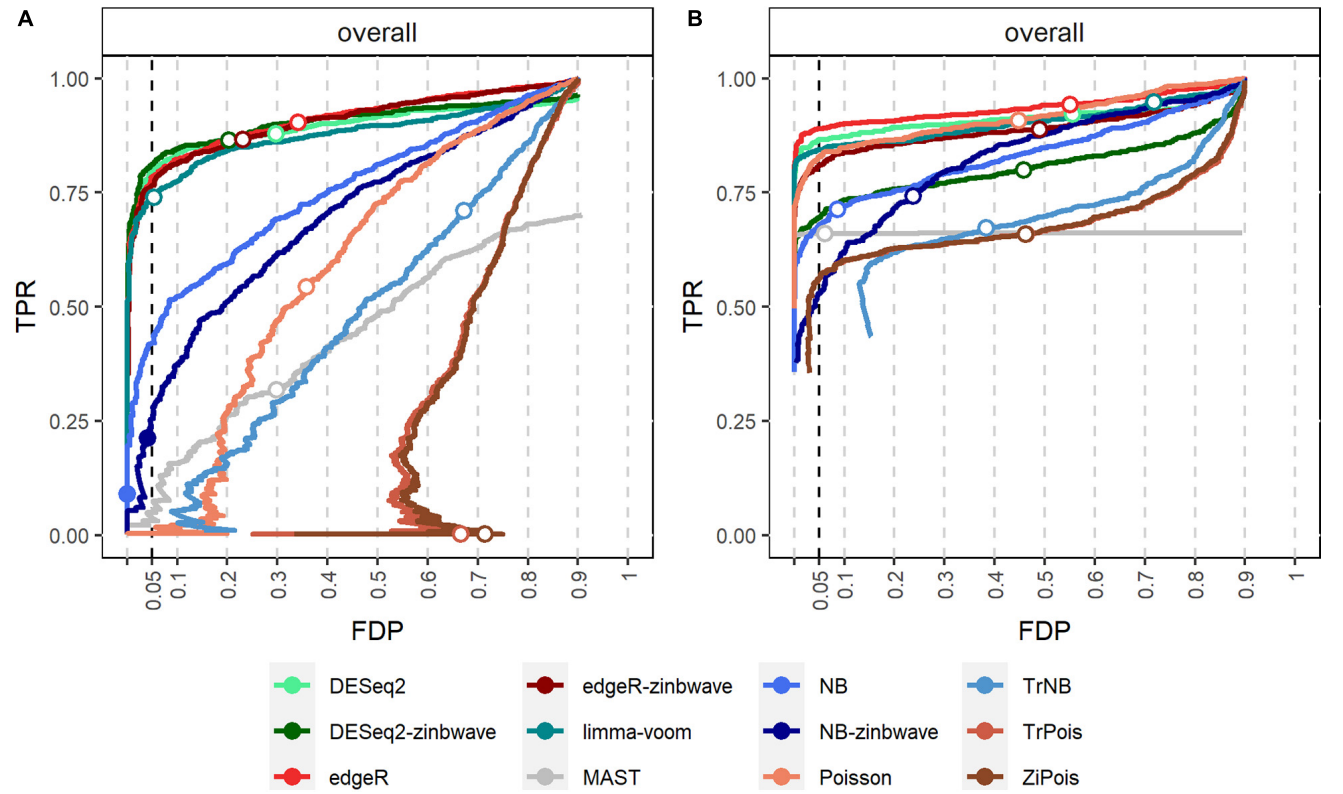

FIGURE 2 | FDP-TPR curves of DE methods on simulated Usoskin datasets and Tung datasets. **(A)** Line plot of the FDP-TPR curves for simulated Usoskin datasets generated by *splatter* for each of 12 DE methods. **(B)** Line plot of the FDP-TPR curves for simulated Tang datasets generated by *splatter* for each of 12 DE methods. Circles represent values at a 0.05 nominal FDR threshold and are filled in if the FDP (i.e., empirical FDR) is less than 0.05. DE, differential expression; TPR, true positive rate; FDP, false discovery proportion; FDR, false discovery rate.

perform well, possibly due to the high proportion of zeros and low counts (Van den Berge et al., 2018). In general, standard DE methods except MAST perform better than scMMSTs when the batch effects is not associated with group effects.

Simulations by GLMMs

Results of PCERs are shown in **Supplementary Figures S2, S3** and **Supplementary Table S1**. Methods performances of the FPR control were similar to those in simulations by *splatter*. Based on FDP-TPR curves for adjusted *p*-values shown in **Figure 3**, scMMSTs performed better than standard DE methods when batch effects were associated with weak group effects. NB-zinb was the best among all methods considered for comparisons. EdgeR-zinb and DESeq2-zinb were the best two methods among the six standard DE methods considered. TrPois and ZiPois perform very close to each other. **Figure 4** demonstrates bar plots of AUCs for adjusted *p*-values. $|\beta_0|$, σ^2_β , θ and μ_π exhibited positive correlations with AUCs. Our scMMSTs performed better when the group effect size and its heterogeneity are larger and the counts dispersion BCV and proportion of zeros are smaller. Similar results are obtained to those of FDP-TPR curves. Therefore, our results demonstrate that scMMSTs performs better than standard DE methods when the group effect size is small with large group effect heterogeneity.

Real Data Analysis

Table 1 and **Supplementary Figure S4** show the numbers of DE genes detected by the 12 methods considered in simulations for 11 cell types in the Usoskin dataset. This dataset was also analyzed in the *zinbwave* paper. MAST failed for some cell-types, so no DE gene was detected. NB-zinb defined smallest number of DE genes in general. The results of Venn diagrams and Upset plots by R packages *VennDiagram* (version 1.6.20) (Chen, 2018) and *upsetR* (version 1.4.0) (Gehlenborg, 2019) are shown in **Supplementary Figures S5–S15**. Since NB-zinb is conservative for FDR, the DE genes only detected by NB-zinb highly likely have weak group effects with their heterogeneity across batches. In general, scMMSTs, as supplement to standard methods, are superior at selecting DE genes with weak group effects and their heterogeneity in different batches for scRNA-seq data.

Computational Time

To demonstrate the computation time scale of DE methods considered, we benchmarked two different simulated null datasets by *splatter* with parameters estimated by the Usoskin

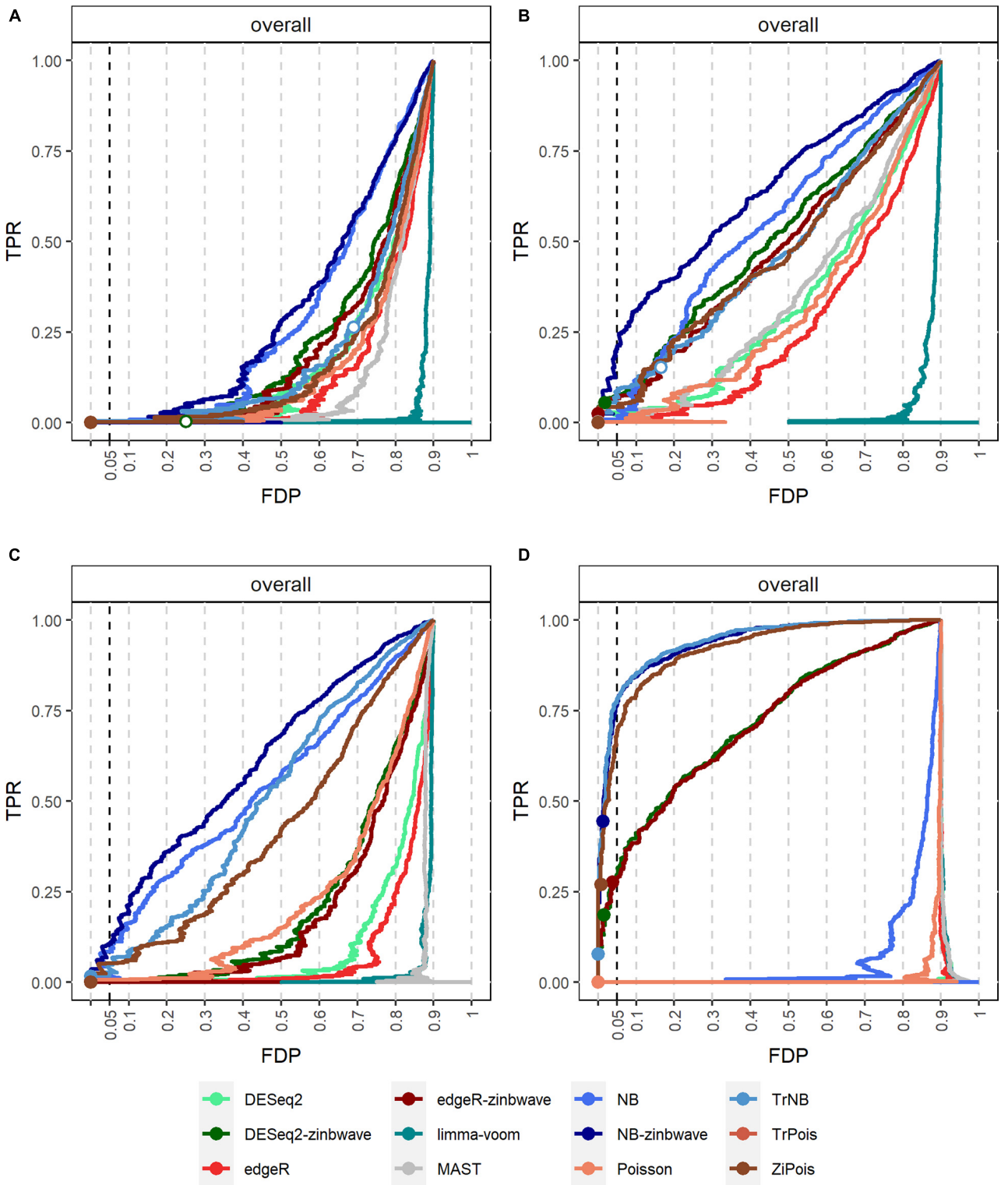

FIGURE 3 | FDP-TPR curves of DE methods on simulated datasets generated by GLMMs with $\mu_\pi = 0$. **(A)** Line plot of the FDP-TPR curves for simulated datasets based on NB GLMMs for each of 12 DE methods with the dispersion parameter $\theta = 0.5$. **(B)** Line plot of the FDP-TPR curves for simulated datasets based on negative binomial (NB) GLMMs for each of 12 DE methods with $\theta = 1$. **(C)** Line plot of the FDP-TPR curves for simulated datasets based on NB GLMMs for each of 12 DE methods with $\theta = 2$. **(D)** Line plot of the FDP-TPR curves for simulated datasets based on Poisson GLMMs for each of 12 DE methods with $\beta_0 = \sigma^2_\beta = 0.01$. Circles represent values at a 0.05 nominal FDR threshold and are filled in if the FDP (i.e., empirical FDR) is less than 0.05. DE, differential expression; GLMM, generalized linear mixed model; NB, negative binomial; TPR, true positive rate; FDP, false discovery proportion; FDR, false discovery rate.

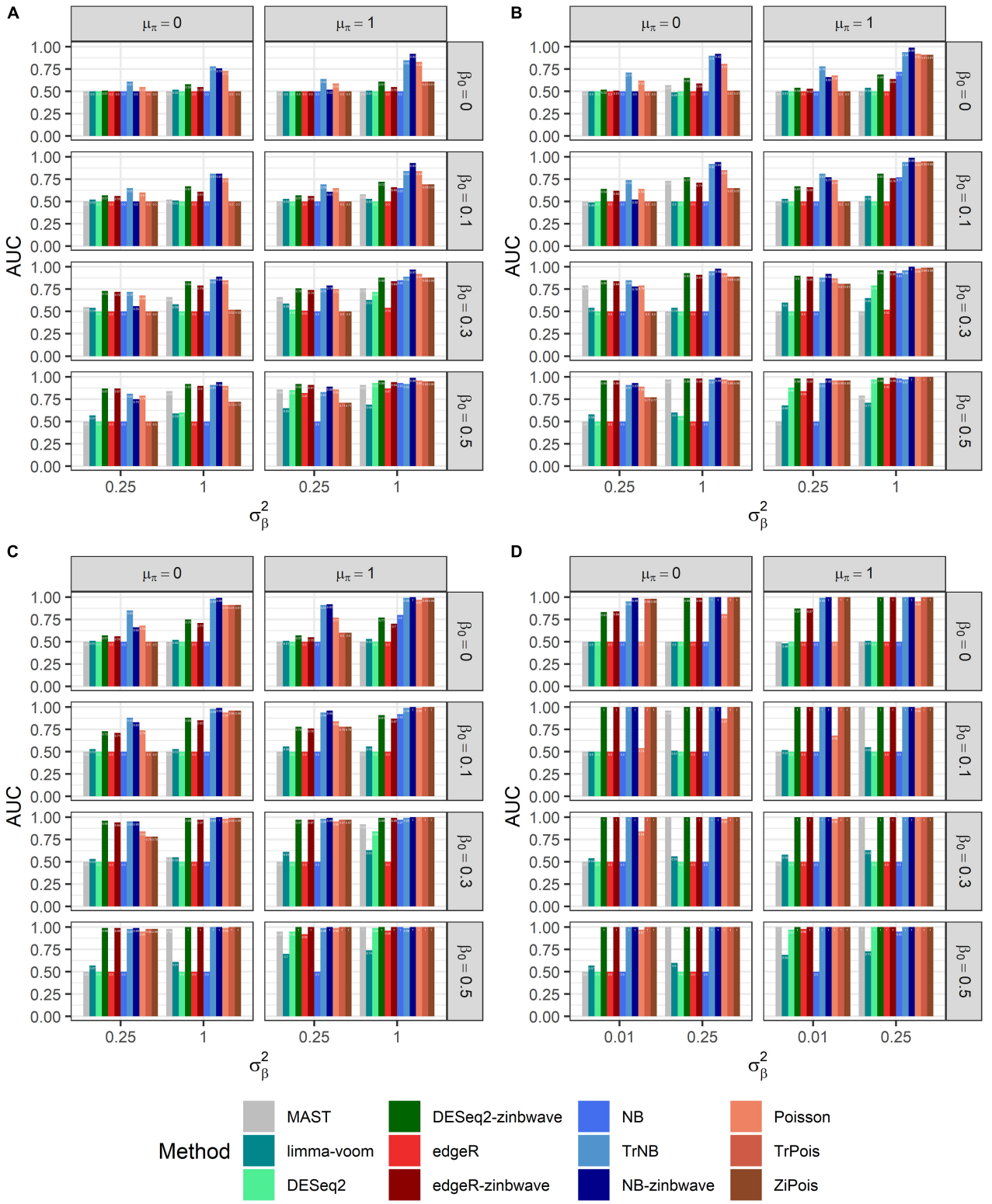

FIGURE 4 | AUCs of DE methods for simulated datasets generated by GLMMs with $\mu_\pi = 0$. Adjusted *p*-values are used as predictors. **(A)** Bar plot of AUCs for simulated datasets generated by NB GLMMs for each of 12 DE methods with the dispersion parameter $\theta = 0.5$. **(B)** Bar plot of AUCs for simulated datasets generated by NB GLMMs for each of 12 DE methods with $\theta = 1$. **(C)** Bar plot of AUCs for simulated datasets generated by NB GLMMs for each of 12 DE methods with $\theta = 2$. **(D)** Bar plot of AUCs for simulated datasets generated by Poisson GLMMs for each of 12 DE methods. AUC, area under curve; DE, differential expression; GLMM, generalized linear mixed model; NB, negative binomial.

TABLE 1 | Numbers of declared differentially expressed genes by 12 methods for 11 defined cell types vs. the rest in the Usoskin dataset ($n = 622$ cells).

Methods	NF1	NF2	NF3	NF4	NF5	NP1	NP2	NP3	PEP1	PEP2	TH
edgeR	826	1206	348	646	1070	1877	880	362	1833	328	2424
DESeq2	906	963	218	402	782	1988	748	407	2649	102	2387
limma-voom	5427	3762	3777	721	2572	2505	4857	203	7892	173	4800
MAST	0	0	0	2	0	85	5	2	10	0	112
edgeR-zinb	509	778	244	550	985	1871	987	486	2475	185	3225
DESeq2-zinb	555	1003	319	453	1235	1985	786	392	2249	153	3166
NB	295	517	186	365	555	462	329	218	592	145	533
TrNB	910	703	596	1763	885	1127	2139	2254	3752	537	1986
NB-zinb	192	308	77	295	364	976	467	270	2004	100	878
Pois	745	1214	410	881	1195	1401	745	583	2104	339	1942
TrPois	242	298	82	345	321	602	756	444	3353	54	708
ZiPois	337	311	81	487	376	607	1019	446	3350	137	704

and Tung datasets. Other settings remained the same as those in the simulations for PCERs. Results are shown in **Figure 5**. For both datasets, the fastest method was limma-voom. DESeq2 was slower than edgR, thus, DESeq2-zinb was also slower than edgeR-zinb. Our scMMSTs performed in the same scale of DESeq2-zinb and DESeq2-zinb. The computation times of simulated null Tung datasets were shorter than those of simulated null Usoskin datasets with the same number of cores. More cores used in the parallel computation made our scMMSTs faster. With eight cores, the computation times of Poisson related methods were close to MAST, edgeR, and DESeq2. In summary, our scMMSTs are computationally affordable compared to other DE methods especially when parallel computing is allowed. All computations were done on a cluster with 24 Intel Xeon Processor (Skylake, IBRS) at 2.60 GHz (2593 MHz) and 128 GB RAM.

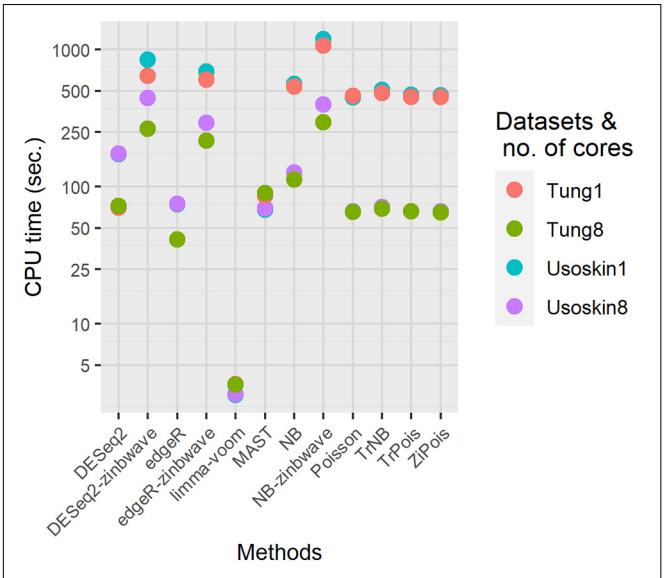

FIGURE 5 | Computational times for differential expression methods on the simulated null Usoskin and Tung datasets, which were generated by *splatter*. The number of cores were set to be 1 and 8 on a cluster with 24 Intel Xeon Processor (Skylake, IBRS) at 2.60 GHz (2593 MHz) and 128 GB RAM.

DISCUSSION

We proposed scMMSTs to identify DE genes, considering batch effect and zero inflation of scRNA-seq data. Both simulations and real data indicated that these methods have advantages in selecting DE genes with weak group effects and their heterogeneity in different batches. In simulations, scMMSTs conservatively controlled FPRs or type I error rates in each setting under assumptions of NB and Poisson distributions, except TrNB and Poisson assumption. However, TrNB controlled FPRs when θ is large. Second, following the model assumption, scMMST was the best one when $|\beta_0|$ was small and σ^2_β was large, especially when θ was large. In real data analysis, the Venn diagrams and Upset plots of DE genes (**Supplementary Figures S5–S15**) directly indicated the relationships among the DE methods. scMMATs defined smaller numbers of DE genes and NB-zinb defined the smallest. Since scMMATs are conservative, the DE genes only defined by NB-zinb are likely to have the small group effect size with its heterogeneity across batches.

Furthermore, scMMSTs exhibited three innovations. First, scMMSTs derived the association test score statistics and their theoretical null distributions in the framework of GLMMs under the binomial, Poisson and NB assumptions. Second, the group effect $\boldsymbol{\beta}$ was modeled as random effects associated with batches in the framework of GLMMs. Third, scMMSTs verified their effectiveness to detect DE genes with the weak group effect and its heterogeneity in different batches. However, scMMSTs have some limitations. scMMSTs performed worse than other standard DE methods to detect DE genes without group effect heterogeneity across batches. scMMSTs performed worse when the dispersion parameter θ was small, especially for the TrNB method, this may due to the non-robust estimation of θ. scMMSTs, in fact, are derived to test H_0' under the assumption $\beta_0 = 0$, not to jointly test $\beta_0 = 0$ and $\sigma^2_\beta = 0$. This decreases the power of testing H_0 for scMMSTs. For association tests, the Mixed effects Score Test (MiST), which jointly tests H_0, is more powerful. Therefore, scMMSTs may be extended using the framework of GLMM-MiST (Sun et al., 2013) in future work to overcome these drawbacks. E'_w is used to approximate E_w for the statistic $\mathbf{T_w}$ of scMMSTs. This approximation performs

worse when there are more excess zeros. Better approximations of E_w or methods to efficiently calculate E_w may improve the performance of scMMSTs. The weighting strategy implemented may be explained in a Bayesian framework and scMMSTs may be extended accordingly. In addition, following the idea of PEA (Shao et al., 2019), scMMSTs may be extended to efficiently identify gene-pathway interactions without permutations of test statistics. In conclusion, scMMSTs, supplements to standard single cell DE methods, are advantageous at selecting genes with the weak group effect and its heterogeneity across batches for scRNA-seq data analysis.

DATA AVAILABILITY STATEMENT

Publicly available datasets were analyzed in this study. These data can be found here: the dataset (Usoskin) analyzed for this study can be found in the [Github respiratory of the *zinbwave* paper (Van den Berge et al., 2018)] (https://github.com/statOmics/zinbwaveZinger/blob/master/datasets/esetUsoskin.RData); the dataset (Tung) can be found in the [Github respiratory of the splatter paper (Zappia et al., 2017)](https://github.com/Oshlack/splatter-paper/blob/master/data.tar.gz).

AUTHOR CONTRIBUTIONS

FS and HW conceived and supervised the study. ZH and FS implemented the software, conducted the simulations, analyzed the data, and wrote the manuscript. ZH and YP prepared figures and tables. ZH, YP, HW, and FS modified and reviewed the manuscript. All authors contributed to the article and approved the submitted version.

FUNDING

This work was supported by the National Natural Science Foundation of China (Nos. 81703321 and 81502888), the Jiangsu Shuangchuang Plan 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.

ACKNOWLEDGMENTS

The authors would like to thank reviewers for their valuable feedback and comments which significantly improved the article's quality. The authors are grateful to YiDuCloud Tech. Ltd., for the support on high performance computation resources.

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fgene.2021.616686/full#supplementary-material

REFERENCES

Benjamini, Y., and Hochberg, Y. (1995). Controlling the false discovery rate: a practical and powerful approach to multiple testing. *J. R. Stat. Soc. Ser. B-Methodol.* 57, 289–300. doi: 10.1111/j.2517-6161.1995.tb02031.x

Böhning, D., Dietz, E., Schlattmann, P., Mendonça, L., and Kirchner, U. (1999). The zero-inflated Poisson model and the decayed, missing and filled teeth index in dental epidemiology. *J. R. Stat. Soc. Ser. A* 162, 195–209. doi: 10.1111/1467-985X.00130

Breslow, N. E., and Clayton, D. G. (1993). Approximate inference in generalized linear mixed models. *J. Am. Stat. Assoc.* 88, 9–25. doi: 10.2307/2290687

Butler, A., Hoffman, P., Smibert, P., Papalexi, E., and Satija, R. (2018). Integrating single-cell transcriptomic data across different conditions, technologies, and species. *Nat. Biotechnol.* 36, 411–420. doi: 10.1038/nbt.4096

Büttner, M., Miao, Z., Wolf, F. A., Teichmann, S. A., and Theis, F. J. (2019). A test metric for assessing single-cell RNA-seq batch correction. *Nat. Methods* 16, 43–49. doi: 10.1038/s41592-018-0254-1

Chen, H. (2018). *VennDiagram: Generate High-Resolution Venn and Euler Plots*. Available online at: https://CRAN.R-project.org/package=VennDiagram (accessed June 8, 2020).

Chen, H., Huffman, J. E., Brody, J. A., Wang, C., Lee, S., Li, Z., et al. (2019). Efficient variant set mixed model association tests for continuous and binary traits in large-scale whole-genome sequencing studies. *Am. J. Hum. Genet.* 104, 260–274. doi: 10.1016/j.ajhg.2018.12.012

Chen, H., Wang, C., Conomos, M. P., Stilp, A. M., Li, Z., Sofer, T., et al. (2016). Control for population structure and relatedness for binary traits in genetic association studies via logistic mixed models. *Am. J. Hum. Genet.* 98, 653–666.

Chu, L.-F., Leng, N., Zhang, J., Hou, Z., Mamott, D., Vereide, D. T., et al. (2016). Single-cell RNA-seq reveals novel regulators of human embryonic stem cell differentiation to definitive endoderm. *Genome Biol.* 17:173. doi: 10.1186/s13059-016-1033-x

Corporation, M., and Weston, S. (2019). *doParallel: Foreach Parallel Adaptor for the "Parallel" Package*. Available online at: https://CRAN.R-project.org/package=doParallel (accessed June 8, 2020).

Eddelbuettel, D. (2013). *Seamless R and C++ Integration with Rcpp*. New York, NY: Springer, doi: 10.1007/978-1-4614-6868-4

Eddelbuettel, D., and Balamuta, J. J. (2017). Extending extitR with extitC++: A Brief Introduction to extitRcpp. *PeerJ. Prepr.* 5:e3188v1. doi: 10.7287/peerj.preprints.3188v1

Eddelbuettel, D., and François, R. (2011). Rcpp: Seamless R and C++ Integration. *J. Stat. Softw.* 40, 1–18. doi: 10.18637/jss.v040.i08

Finak, G., Mcdavid, A., Yajima, M., Deng, J., Gersuk, V. H., Shalek, A. K., et al. (2015). MAST: a flexible statistical framework for assessing transcriptional changes and characterizing heterogeneity in single-cell RNA sequencing data. *Genome Biol.* 16, 278–278. doi: 10.1186/s13059-015-0844-5

Gehlenborg, N. (2019). *UpSetR: A More Scalable Alternative to Venn and Euler Diagrams for Visualizing Intersecting Sets*. Available online at: https://CRAN.R-project.org/package=UpSetR (accessed June 8, 2020).

Haghverdi, L., Lun, A. T., Morgan, M. D., and Marioni, J. C. (2018). Batch effects in single-cell RNA-sequencing data are corrected by matching mutual nearest neighbors. *Nat. Biotechnol.* 36, 421–427. doi: 10.1038/nbt.4091

Harville, D. A. (1977). Maximum likelihood approaches to variance component estimation and to related problems. *J. Am. Stat. Assoc.* 72, 320–338. doi: 10.1080/01621459.1977.10480998

Hashimshony, T., Senderovich, N., Avital, G., Klochendler, A., De Leeuw, Y., Anavy, L., et al. (2016). CEL-Seq2: sensitive highly-multiplexed single-cell RNA-Seq. *Genome Biol.* 17, 77–77. doi: 10.1186/s13059-016-0938-8

Hicks, S. C., Teng, M., and Irizarry, R. A. (2015). On the widespread and critical impact of systematic bias and batch effects in single-cell RNA-Seq data. *BioRxiv*[Preprint] 025528. doi: 10.1101/025528

Johnson, W. E., Li, C., and Rabinovic, A. (2007). Adjusting batch effects in microarray expression data using empirical Bayes methods. *Biostatistics* 8, 118–127. doi: 10.1093/biostatistics/kxj037

Kharchenko, P., and Fan, J. (2019). *scde: Single Cell Differential Expression*. Available online at: http://pklab.med.harvard.edu/scde (accessed June 8, 2020).

Kolodziejczyk, A. A., Kim, J. K., Svensson, V., Marioni, J. C., and Teichmann, S. A. (2015). The technology and biology of single-cell RNA sequencing. *Mol. Cell* 58, 610–620. doi: 10.1016/j.molcel.2015.04.005

Krieg, C., Nowicka, M., Guglietta, S., Schindler, S., Hartmann, F. J., Weber, L. M., et al. (2018). High-dimensional single-cell analysis predicts response to anti-PD-1 immunotherapy. *Nat. Med.* 24:144. doi: 10.1038/nm.4466

Li, Q., Cheng, Z., Zhou, L., Darmanis, S., Neff, N. F., Okamoto, J., et al. (2019). Developmental heterogeneity of microglia and brain myeloid cells revealed by deep single-cell RNA sequencing. *Neuron* 101, 207–223.e10. doi: 10.1016/j.neuron.2018.12.006

Love, M. I., Huber, W., and Anders, S. (2014). Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. *Genome Biol.* 15, 550–550. doi: 10.1186/s13059-014-0550-8

Luecken, M. D., and Theis, F. J. (2019). Current best practices in single-cell RNA-seq analysis: a tutorial. *Mol. Syst. Biol.* 15:e8746. doi: 10.15252/msb.20188746

McCarthy, D. J., Chen, Y., and Smyth, G. K. (2012). Differential expression analysis of multifactor RNA-Seq experiments with respect to biological variation. *Nucleic Acids Res.* 40, 4288–4297. doi: 10.1093/nar/gks042

McDavid, A., Finak, G., and Yajima, M. (2019). *MAST: Model-based Analysis of Single Cell Transcriptomics*. Available online at: https://github.com/RGLab/MAST/ (accessed June 8, 2020).

McEvoy, J., Flores-Otero, J., Zhang, J., Nemeth, K., Brennan, R., Bradley, C., et al. (2011). Coexpression of normally incompatible developmental pathways in retinoblastoma genesis. *Cancer Cell* 20, 260–275. doi: 10.1016/j.ccr.2011.07.005

Mehtonen, J., Teppo, S., Lahnalampi, M., Kokko, A., Kaukonen, R., Oksa, L., et al. (2020). Single cell characterization of B-lymphoid differentiation and leukemic cell states during chemotherapy in ETV6-RUNX1 positive pediatric leukemia identifies drug-targetable transcription factor activities. *bioRxiv*[Preprint] doi: 10.1186/s13073-020-00799-2

Morgan, M., Obenchain, V., Lang, M., Thompson, R., and Turaga, N. (2019). *BiocParallel: Bioconductor Facilities for Parallel Evaluation*. Available online at: https://github.com/Bioconductor/BiocParallel (accessed June 8, 2020).

Papalexi, E., and Satija, R. (2018). Single-cell RNA sequencing to explore immune cell heterogeneity. *Nat. Rev. Immunol.* 18, 35–45. doi: 10.1038/nri.2017.76

Rider, P. R. (1955). Truncated binomial and negative binomial distributions. *J. Am. Stat. Assoc.* 50, 877–883. doi: 10.1080/01621459.1955.10501973

Risso, D., Perraudeau, F., Gribkova, S., Dudoit, S., and Vert, J.-P. (2018). A general and flexible method for signal extraction from single-cell RNA-seq data. *Nat. Commun.* 9:284. doi: 10.1038/s41467-017-02554-5

Ritchie, M. E., Phipson, B., Wu, D., Hu, Y., Law, C. W., Shi, W., et al. (2015). limma powers differential expression analyses for RNA-sequencing and microarray studies. *Nucleic Acids Res.* 43:e47. doi: 10.1093/nar/gkv007

Robin, X., Turck, N., Hainard, A., Tiberti, N., Lisacek, F., Sanchez, J.-C., et al. (2011). pROC: an open-source package for R and S+ to analyze and compare ROC curves. *BMC Bioinformatics* 12:77. doi: 10.1186/1471-2105-12-77

Robinson, M. D., McCarthy, D. J., and Smyth, G. K. (2010). edgeR: a Bioconductor package for differential expression analysis of digital gene expression data. *Bioinformatics* 26, 139–140. doi: 10.1093/bioinformatics/btp616

Roerink, S. F., Sasaki, N., Lee-Six, H., Young, M. D., Alexandrov, L. B., Behjati, S., et al. (2018). Intra-tumour diversification in colorectal cancer at the single-cell level. *Nature* 556, 457–462. doi: 10.1038/s41586-018-0024-3

Santos Nobre, J., and da Motta Singer, J. (2007). Residual analysis for linear mixed models. *Biom. J. J. Math. Methods Biosci.* 49, 863–875. doi: 10.1002/bimj.200610341

Shao, F., Wang, Y., Zhao, Y., and Yang, S. (2019). Identifying and exploiting gene-pathway interactions from RNA-seq data for binary phenotype. *BMC Genet.* 20:36. doi: 10.1186/s12863-019-0739-7

Somekh, J., Shenorr, S. S., and Kohane, I. S. (2019). Batch correction evaluation framework using a-priori gene-gene associations: applied to the GTEx dataset. *BMC Bioinformatics* 20:268. doi: 10.1186/s12859-019-2855-9

Soneson, C., and Robinson, M. D. (2016). iCOBRA: open, reproducible, standardized and live method benchmarking. *Nat. Methods* 13:283. doi: 10.1038/nmeth.3805

Sun, J., Zheng, Y., and Hsu, L. (2013). A unified mixed-effects model for rare-variant association in sequencing studies. *Genet. Epidemiol.* 37, 334–344. doi: 10.1002/gepi.21717

Sun, S., Zhu, J., Mozaffari, S., Ober, C., Chen, M., and Zhou, X. (2018). Heritability estimation and differential analysis of count data with generalized linear mixed models in genomic sequencing studies. *Bioinformatics* 35, 487–496. doi: 10.1093/bioinformatics/bty644

Sun, X., Sun, S., and Yang, S. (2019). An efficient and flexible method for deconvoluting bulk RNA-Seq data with single-cell RNA-seq data. *Cells* 8:1161. doi: 10.3390/cells8101161

Tang, F., Barbacioru, C., Wang, Y., Nordman, E., Lee, C. C., Xu, N., et al. (2009). mRNA-Seq whole-transcriptome analysis of a single cell. *Nat. Methods* 6, 377–382. doi: 10.1038/nmeth.1315

Tung, P., Blischak, J. D., Hsiao, C. J., Knowles, D., Burnett, J. E., Pritchard, J. K., et al. (2017). Batch effects and the effective design of single-cell gene expression studies. *Sci. Rep.* 7, 39921–39921. doi: 10.1038/srep39921

Usoskin, D., Furlan, A., Islam, S., Abdo, H., Lonnerberg, P., Lou, D., et al. (2015). Unbiased classification of sensory neuron types by large-scale single-cell RNA sequencing. *Nat. Neurosci.* 18, 145–153. doi: 10.1038/nn.3881

Van den Berge, K., Perraudeau, F., Soneson, C., Love, M. I., Risso, D., Vert, J.-P., et al. (2018). Observation weights unlock bulk RNA-seq tools for zero inflation and single-cell applications. *Genome Biol.* 19:24. doi: 10.1186/s13059-018-1406-4

Van den Berge, K., Soneson, C., Love, M. I., Robinson, M. D., and Clement, L. (2017). zingeR: unlocking RNA-seq tools for zero-inflation and single cell applications. *bioRxiv*[Preprint] doi: 10.1101/157982

Wang, Z., Gerstein, M., and Snyder, M. (2009). RNA-Seq: a revolutionary tool for transcriptomics. *Nat. Rev. Genet.* 10, 57–63. doi: 10.1038/nrg2484

Wu, M. C., Lee, S., Cai, T., Li, Y., Boehnke, M., and Lin, X. (2011). Rare-variant association testing for sequencing data with the sequence kernel association test. *Am. J. Hum. Genet.* 89, 82–93. doi: 10.1016/j.ajhg.2011.05.029

Yang, S., Shao, F., Duan, W., Zhao, Y., and Chen, F. (2017). Variance component testing for identifying differentially expressed genes in RNA-seq data. *PeerJ* 5:e3797. doi: 10.7717/peerj.3797

Zappia, L., Phipson, B., and Oshlack, A. (2017). Splatter: simulation of single-cell RNA sequencing data. *Genome Biol.* 18:174. doi: 10.1186/s13059-017-1305-0

Zeisel, A., Muñoz-Manchado, A. B., Codeluppi, S., Lönnerberg, P., La Manno, G., Juréus, A., et al. (2015). Cell types in the mouse cortex and hippocampus revealed by single-cell RNA-seq. *Science* 347, 1138–1142. doi: 10.1126/science.aaa1934

Conflict of Interest: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.