Bioinformatics, 35(24), 2019, 5155–5162
doi: 10.1093/bioinformatics/btz453
Advance Access Publication Date: 14 June 2019
Original Paper

OXFORD

Gene expression

DECENT: differential expression with capture efficiency adjustmeNT for single-cell RNA-seq data

Chengzhong Ye[1,2,3], Terence P. Speed[1,4] and Agus Salim[1,5,6,*]

[1]Bioinformatics Division, Walter and Eliza Hall Institute of Medical Research, Parkville, VIC 3052, Australia, [2]Department of Medical Biology, The University of Melbourne, Parkville, VIC 3010, Australia, [3]School of Medicine, Tsinghua University, Haidian District, Beijing 100084, China, [4]Department of Mathematics and Statistics, The University of Melbourne, Parkville, VIC 3010, Australia, [5]Department of Mathematics and Statistics, La Trobe University, Bundoora, VIC 3086, Australia and [6]Baker Heart and Diabetes Institute, Melbourne, VIC 3004, Australia

*To whom correspondence should be addressed.
Associate Editor: Bonnie Berger

Received on October 23, 2018; revised on May 8, 2019; editorial decision on May 24, 2019; accepted on June 6, 2019

Abstract

Motivation: Dropout is a common phenomenon in single-cell RNA-seq (scRNA-seq) data, and when left unaddressed it affects the validity of the statistical analyses. Despite this, few current methods for differential expression (DE) analysis of scRNA-seq data explicitly model the process that gives rise to the dropout events. We develop DECENT, a method for DE analysis of scRNA-seq data that explicitly and accurately models the molecule capture process in scRNA-seq experiments.

Results: We show that DECENT demonstrates improved DE performance over existing DE methods that do not explicitly model dropout. This improvement is consistently observed across several public scRNA-seq datasets generated using different technological platforms. The gain in improvement is especially large when the capture process is overdispersed. DECENT maintains type I error well while achieving better sensitivity. Its performance without spike-ins is almost as good as when spike-ins are used to calibrate the capture model.

Availability and implementation: The method is implemented as a publicly available R package available from https://github.com/cz-ye/DECENT.

Contact: a.salim@latrobe.edu.au

Supplementary information: Supplementary data are available at *Bioinformatics* online.

1 Introduction

Recent developments in sequencing technology have enabled high-throughput whole-transcriptome profiling at single-cell resolution. Single-cell RNA-seq (scRNA-seq) allows the quantification of gene expression of thousands of individual cells in a single experiment. It has already led to profound new discoveries that could not be have been made using data from bulk transcriptome sequencing, ranging from the identification of novel cell types to the study of global patterns of stochastic gene expression (Kolodziejczyk *et al.*, 2015; Wagner *et al.*, 2016). However, there are still many statistical challenges in drawing inferences from scRNA-seq data. Due to the small amount of starting material and the imperfect capturing of RNA molecules in current scRNA-seq experiments, failure to detect expressed transcripts in single cells is still common. This gives rise to the characteristic dropout phenomenon in scRNA-seq data, in which a gene shows zero or very low abundance in a fraction of cells in spite of moderate to high expression in others (Finak *et al.*, 2015; Hashimshony *et al.*, 2012; Ramskold *et al.*, 2012). Also, the capture rates can vary between cells and across genes (Brennecke *et al.*, 2013), showing as a major source of unwanted variation in

scRNA-seq data, with the first principal component of raw counts typically exhibiting high correlation with the proportions of zero counts (Risso *et al.*, 2018). This unique feature of scRNA-seq will hinder downstream analyses if not properly modeled. Lots of effort has been made in order to alleviate this issue, including specialized normalization methods (Bacher *et al.*, 2017; Lun *et al.*, 2016), clustering algorithms (Kiselev *et al.*, 2017; Wang *et al.*, 2018; Zeisel *et al.*, 2015) and methods for differential expression analysis (Finak *et al.*, 2015; Jia *et al.*, 2017; Kharchenko *et al.*, 2014).

One way to resolve this is through explicit modeling of the molecule capturing process and hence separating the biological variation of interest from unwanted variation in the experimental procedures. For instance, several methods (Huang *et al.*, 2018; van Dijk *et al.*, 2018; Wang *et al.*, 2018) are designed to recover the pre-dropout expression matrix by modeling the process from RNA molecule to read count. However, a difficulty in modeling the molecule capturing and dropout events is that this process is usually mixed up with other sources of technical variation, such as amplification and sequencing biases (Wagner *et al.*, 2016). The unique molecular identifier (UMI) barcoding approach has become increasingly popular in scRNA-seq experiments as an effective way to address this issue (Islam *et al.*, 2014; Svensson *et al.*, 2017). Random barcodes are attached to cDNA molecules during reverse transcription. Each individual molecule from a particular gene in each cell is expected to have a distinct UMI (Islam *et al.*, 2014). Therefore, after sequencing, by counting UMI barcodes instead of reads per se, the resulting UMI counts will be a more faithful representation of the original cDNA counts, with amplification and sequencing bias largely avoided. But the UMI count will still show as zero if an RNA molecule failed to convert to cDNA, or was completely lost in amplification and sequencing.

As a consequence, the main source of technical variation left in UMI counts is the loss of molecules during the experimental procedure, namely, dropouts. Hence, UMI count data provides us with an opportunity to model the molecule capturing process in depth. Also, given the distinct features of UMI-based data, it is necessary to build specific models in order to perform statistical tests reliably.

Currently scRNA-seq experiments mainly focus on cell-wise analyses such as clustering and trajectory inference for studying heterogeneity within cellular populations (Qiu *et al.*, 2017; Trapnell *et al.*, 2014; Zeisel *et al.*, 2015). Nevertheless, differential gene expression (DE), as one of the most common gene-wise analyses, still plays an essential role in complementing these analyses. For example, it is used to identify cluster-specific markers for identifying the cell types. It is also used to derive disease-associated gene signatures (Savas *et al.*, 2018; Sun *et al.*, 2018; Zhao *et al.*, 2017). However, DE methods originally designed for bulk RNA-seq tend to produce unreliable results due to failing to account for the extra variation in single-cell data (Jia *et al.*, 2017; Van den Berge *et al.*, 2018). Driven by this, a few DE methods have been designed specifically for scRNA-seq data. All of them use some strategy to deal with the large variation and number of zero observations. However, most of them do not distinguish biological from technical factors that are causing the phenomenon. For example, SCDE (Kharchenko *et al.*, 2014) uses a mixture model to distinguish counts affected by dropout from the rest of the data. This model almost always assigns a probability of one that a zero count belongs to the dropout component, in essence assuming all observed zeroes to be technical. MAST (Finak *et al.*, 2015) uses a two-part generalized linear model (GLM) in which the dropout rates are adjusted by the inclusion of the observed fraction of non-zero counts as a term in their regression model. This still does not differentiate the dropouts from real biological zeros. Additionally, the effect of dropout events is likely to be non-linear, especially for genes with low to moderate expression (Bacher *et al.*, 2017), and so the inclusion of simple linear term that represents capture rates in the regression model is unlikely to be optimal. Zero-inflated negative binomial (ZINB)-WaVE (Van den Berge *et al.*, 2018) uses a zero-inflated model directly fitted to the observed data to derive observation weights for adjusting bulk DE methods. Only Jia *et al.* (2017) proposed a DE method, TASC, that relies on external RNA spike-in data (Jiang *et al.*, 2011) to fit a technical variation model in order to explicitly cater for dropouts, thus enabling separation of the biological variation for DE analysis. They showed improved performance of their method compared with methods that perform DE analysis directly using the observed data. Note that the methods mentioned so far are not specifically designed for UMI-count data. There are two existing methods that considers the unique features of UMI-based experiments: Monocle2 (Qiu *et al.*, 2017; Trapnell *et al.*, 2014) and NBID (Chen *et al.*, 2018). They both fit negative binomial (NB) models directly to the observed UMI count without any explicit modeling of dropouts.

Here we propose a novel model for the DE analysis of UMI-based scRNA-seq data. Leveraging the features UMI-count data, we are able to model the molecule capturing process precisely. We build a capture model to account for the gene- and cell-specific properties of molecule capturing. This allows us to perform DE analysis on the inferred pre-dropout distributions of RNA molecules. We named our method *D*ifferential *E*xpression with Capture *E*fficiency adjustme*NT* (DECENT). DECENT can use the external RNA spike-in data to calibrate the capture model, but also works without spike-ins. In this paper, we describe the DECENT model and benchmark it against existing methods using both simulated data and four published UMI-based scRNA-seq datasets. The results showed im proved performance of DECENT in various settings when compared with existing methods.

2 Materials and methods

2.1 Model formulation

DECENT assumes that UMIs (Islam *et al.*, 2014) have been used in the scRNA-seq experiment. Our statistical model is hierarchical, involving modelling the observed count Z_{ij} of mRNA molecules 'captured' from gene i in cell j, and the unobserved total mRNA count Y_{ij} of all mRNA molecules from gene i in cell j that could have been captured had there been no molecule dropout. We will subsequently use the term 'pre-dropout count' when referring to Y_{ij}. Cells will in general be of more than one type, and our inferences will concern the mean parameters μ_{ij} of the distribution of the unobserved pre-dropout count across cell types. Other parameters in our model for the unobserved pre-dropout count are zero inflation parameters π_{0i} and (over) dispersion parameters ψ_i specific to gene i, and a size-factor parameter s_j specific to cell j, to account for differences in total mRNA molecules across cells. To summarize, we use ZINB distribution with the following probability density function to model the unobserved pre-dropout count:

$$P(Y_{ij}=k;\pi_{0i},s_j,\mu_{ij},\psi_i)=\begin{cases}\pi_{0i}+(1-\pi_{0i})\left(\frac{1}{1+\psi_i s_j\mu_{ij}}\right)^{\psi_i^{-1}}, & k=0.\\ (1-\pi_{0i})\frac{\Gamma\left(\psi_i^{-1}+k\right)}{k!\Gamma\left(\psi_i^{-1}\right)}\left(\frac{1}{1+\psi_i s_j\mu_{ij}}\right)^{\psi_i^{-1}}\left(\frac{\psi_i s_j\mu_{ij}}{1+\psi_i s_j\mu_{ij}}\right)^{k}, & k>0.\end{cases} \quad (1)$$

Note that, as $\psi_i \to 0$, our ZINB model reduced to a zero-inflated Poisson model (ZIP). When $\pi_{0i}=0$, our model reduces to the NB

model and finally when $\pi_{0i} = 0$ and $\psi_i \to 0$, our model reduces to the Poisson model.

The second part of the DECENT model involves the specification of the capture model for modelling the distribution of the observed data Z_{ij} given the unobserved pre-dropout count Y_{ij}. We use a binomial model for the molecule capture, assuming $Z_{ij}|Y_{ij} = k; \eta_{ij} \sim \text{Binomial}(\eta_{ij}, k)$ where η_{ij} is the capture rates for gene i within cell j. The probability density function for the capture model is given by,

$$P(Z_{ij} = l|Y_{ij} = k; \eta_{ij}) \propto \eta_{ij}^{l}(1-\eta_{ij})^{k-l} \tag{2}$$

Finally, to account for variability in the capture rate, we assume a beta prior for the capture rate parameter η_{ij} with the prior mean equal to the cell-specific capture rates η_j and the prior variance characterized by a dispersion parameter $\rho_{ij}, 0 \leq \rho_{ij} \leq 1$. When $\rho_{ij} = 0$, our capture model reduces to the standard Binomial model with all genes within a cell having the same capture rate, $\eta_{ij} = \eta_j, \forall i$. In our DECENT model, backed up by empirical evidence from published scRNA-seq datasets, we assume that the dispersion parameter is related to gene abundance through a logistic linear model:

$$\log \frac{\rho_{ij}}{1-\rho_{ij}} = \tau_{0j} + \tau_{1j} \log \{\mu_{ij}(1-\pi_{0i})\} \tag{3}$$

where τ_{0j} and τ_{1j} are cell-specific parameters estimated from the data. In our experience with real datasets, we usually found that assuming global parameters $\tau_{0j} = \tau_0$ and $\tau_{1j} = \tau_1, \forall j$ is adequate. Furthermore, from real datasets we also found that $\tau_1 < 0$, which suggests that within a cell, capture rates for highly abundant genes vary around the cell-specific capture rates much less than the capture rates for low or moderately abundant genes. As an alternative to estimating a global $\boldsymbol{\tau} = (\tau_0, \tau_1)^{\mathrm{T}}$ parameter, DECENT also has an option for estimating cell-specific $\boldsymbol{\tau}_j = (\tau_{0j}, \tau_{1j})^{T}$ if there is evidence that the parameters vary significantly between cells.

Our hierarchical DECENT model specification can be summarized as follows:

$$Y_{ij}; \pi_{0i}, s_j, \mu_{ij}, \psi_i \sim \text{ZINB}(\pi_{0i}, s_j\mu_{ij}, \psi_i) \tag{4}$$

$$Z_{ij}|Y_{ij} = k; \eta_{ij} \sim \text{Binomial}(\eta_{ij}, k) \tag{5}$$

$$\eta_{ij} \sim \text{Beta}(a_{ij}, b_{ij}) \tag{6}$$

where the parameters of the beta prior for η_{ij} satisfy

$$\eta_j = \frac{a_{ij}}{a_{ij}+b_{ij}}$$
$$\log \frac{\rho_{ij}}{1-\rho_{ij}} = \tau_{0j} + \tau_{1j} \log \{\mu_{ij}(1-\pi_{0i})\} = -\log(a_{ij}+b_{ij})$$

Next, we describe empirical evidence that motivated us to choose this specification for DECENT.

2.1.1 Empirical evidence for beta-binomial capture model

We investigated the appropriateness of our capture model using six ERCC spike-in datasets, consisting of three plate-based (Grun *et al.*, 2014; Tung *et al.*, 2017; Zeisel *et al.*, 2015) and three droplet-based experiments (Klein *et al.*, 2015; Macosko *et al.*, 2015; Zheng *et al.*, 2017) (Supplementary Table S1). We use spike-ins because their average (nominal) pre-dropout count are known, with no biological variation between cells expected. To model the distribution of their unobserved pre-dropout counts, we thus use a Poisson distribution with mean equal to the nominal count for each spike-in. The full results from all six datasets are shown in Supplementary Figure S2.

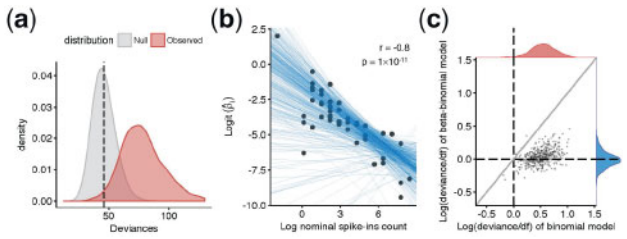

Fig. 1. Modeling extra-binomial variation in the molecule capturing process. We evaluate the binomial and beta-binomial capture models using the ERCC spike-in data from the Tung *et al.* experiment. (**a**) The observed distribution (red) of deviances with cell-wise binomial capture model shows notable deviation from the expected χ^2 distribution under the null hypothesis. This indicates inadequacy of the binomial capture model. (**b**) Modeling the relationship between the spike-in nominal count c_i and the dispersion parameter ρ in the beta-binomial capture model. If the parameter is estimated in a spike-in specific manner, a high correlation between the ρ_i estimates and the true pre-dropout mean abundance, namely the nominal count c_i, can be observed, which are shown as black points. We build a cell-wise linear model to characterize this relationship. Each blue line represents a fitted cell-wise model, which is shown to adequately describe this relationship. (**c**) A scatter plot comparing the cell-wise deviances under the binomial and beta-binomial capture models to assess goodness-of-fit. Deviances were standardized by dividing by the degrees of freedom to enable comparison, and logged. The blue and red marginal densities represent the observed distributions of deviances under the two models, respectively. It can be seen that the beta-binomial capture model fits better than the binomial model in the majority of the cells

But as an example, we will use results from the Tung *et al.* data as shown in Figure 1. Figure 1a shows that the simple Binomial capture model does not provide an adequate fit to the data because the deviance statistic (see Supplementary Materials, pp. 13–14) for this model is much larger than the expected distribution when the model is adequate. Further analyses suggests that variation in capture rates between spikes contribute to this inadequate fit (Supplementary Fig. S1). We therefore model this extra variation by allowing capture rates to have a beta prior with cell-specific mean and dispersion parameter that differs from one spike to another. We find that the dispersion parameter can be well-approximated by a linear logistic model as a function of the spikes' abundance (Fig. 1b). We then compare the fit of the beta-binomial capture model with the simple binomial model and find that the beta-binomial model provides much better fit to the data (Fig. 1c). The same analyses were carried out using the spike-in data from the other five experiments and similar results were obtained (Supplementary Fig. S2).

2.1.2 Empirical evidence for pre-dropout count distribution

The ZINB distribution has two extra parameters when compared with the simple Poisson distribution for count data, namely the overdispersion and the zero-inflation parameters. To examine whether the ZINB distribution is needed for modelling the unobserved pre-dropout count, we used two scRNA-seq datasets (Tung *et al.*, 2017; Zeisel *et al.*, 2015). To investigate the need for overdispersion parameter, we first fit the DECENT model assuming a Poisson pre-dropout distribution to the data without considering zero-inflation. We found that the expected variances of most genes were noticeably lower than the observed values for the Zeisel *et al.* dataset. This extra variation was well-modeled by assuming NB as pre-dropout count distribution, hence indicating the need for the dispersion parameter (Supplementary Fig. S3a). For the Tung *et al.* data, the expected variances under the Poisson pre-dropout count assumption were already close to the observed values for most genes, showing

little need for the extra parameter (Supplementary Fig. S3b). This suggests that overdispersion in pre-dropout counts is dataset-specific and depends on the amount of biological variability in the sample. The Tung *et al.* data used here are from one iPSC cell line where cells were highly homogeneous and hence lack biological variation. On the other hand, the Zeisel *et al.* data are from mouse brain tissue, which has a complex cellular composition. To investigate the need for zero-inflation parameter, we further fitted the DECENT model assuming a ZINB pre-dropout count distribution and compared this with the model that assumes NB pre-dropout count distribution. We performed chi-square goodness-of-fits test on the DECENT models with ZINB and NB pre-dropout count distribution to assess their adequacy. Consistent with previous findings (Chen *et al.*, 2018; Vieth *et al.*, 2017), the majority of genes do not appear to require zero-inflated model. However we still found a small number of genes in both datasets in which models with ZINB provide a more adequate fit than NB (Supplementary Fig. S4).

We also investigated the appropriateness of the ZINB distribution using a single-molecule fluorescence *in situ* hybridization (smFISH) dataset. The smFISH technology allows precise quantification of RNA molecules from a list of targeted genes. This technology can achieve nearly 100% sensitivity detection of the RNA molecules (Raj *et al.*, 2008). Hence, the smFISH count data is a good approximation to the pre-dropout molecule counts that would normally be unobserved. We used the data from an experiment that profiled 33 marker genes in mouse somatosensory cortex (Codeluppi *et al.*, 2018). We examined three of the clusters identified by the authors, Oligodendrocyte Mature, Pyramidal L4 and Inhibitory Vip, finding most of the gene count distributions to be significantly overdispersed relative to the Poisson (Supplementary Fig. S5a). Yet we did not find zero-inflated genes in these clusters. This is quite possibly because the targeted genes are all canonical markers, which are expected to mostly exhibit constitutive expression and hence unlikely to have inflated zeros caused by transcriptional bursting. However, heterogeneity within a population can also result in zero-inflation, which is common in actual DE analysis. When we looked at 3 major cell types with higher within-group heterogeneity (Oligodendrocytes, Pyramidal neurons and Inhibitory neurons), we identified 2, 1 and 2 out of the 33 genes to have significant zero-inflation (Supplementary Fig. S5b).

2.2 Estimating capture rate parameters

Our capture model requires estimates of a cell-specific capture rate that will be used as part of the beta prior in the beta-binomial capture model. This capture rate needs to be estimated externally outside the main algorithm, for reasons that we will explain below. When the dataset contains spike-ins, their data are used to estimate the capture rates. Suppose we added n_s spike-ins at the known concentrations $c_1, c_2, \ldots c_{n_s}$ into cell j and subsequently observe $z_{1j}, z_{2j}, \ldots z_{n_s j}$ molecules respectively. The cell-specific capture efficiency for any cell j is estimated as the proportion of molecules observed after sequencing relative to the total number of molecules initially added:

$$\hat{\eta}_j = \frac{\sum_{i=1}^{n_s} z_{ij}}{\sum_{i=1}^{n_s} c_i},$$

This is the method of moments estimator of the capture rate η_j under the beta-binomial-Poisson model for spike-ins.

However, many scRNA-seq data do not have spike-ins. Interestingly, we found that if we specified a set of inexact capture rates, other components of the model will compensate for the inaccuracy and produce DE results almost as reliable as if we had the correct values. This is due to a property of the our model that is explained below:

Let Y be the unobserved pre-dropout count where $Y \sim \text{ZINB}(\pi_0, s\mu, \psi)$ and $Z|Y = k; a, b \sim \text{Beta} - \text{Binomial}(k, a, b)$, with the capture rate given by $\eta = \frac{a}{a+b}$. It turns out that the marginal distribution of Z in this case is almost indistinguishable from the marginal distribution of Z when the capture rate is η' and the size-factor parameter of the ZINB distribution is $s' = s(\eta/\eta')$ (see Supplementary Fig. S6). This feature means that if we incorrectly specify the capture rate as η' rather than η, the misspecification can be approximately corrected by scaling the size factor estimates accordingly. Nevertheless, it is still preferable to get capture efficiency estimates as close as possible to the true value. This identifiability issue involving a cell's size-factor and capture rate parameter also means that it is nearly impossible to simultaneously estimate these parameters inside the main algorithm. Our approach here is to estimate the capture rate parameters before invoking the expectation conditional maximization (ECM) algorithm, and given these, we estimate the size-factor parameters within the ECM algorithm.

Motivated by the above results and our experience with real datasets showing that capture efficiency is the biggest factor contributing to the variation in the observed library sizes, we devised a method for generating functional capture rates when spike-ins are not available. We refer to this as the ranked random capture efficiency. To use this method, users will need to specify an interval of plausible capture rate parameters in their experiment. DECENT uses $(0.02, 0.10)$ as its default and we have found this to work well in all the datasets without spike-ins that we have analyzed. As an alternative, users can refer to Ziegenhain *et al.* (2017) for guidance on plausible range of capture rate parameters across various different sequencing protocols.

Let the lower and upper bounds of capture rates be $\eta_{\min}$ and $\eta_{\max}$, respectively. The cell-specific capture rates are specified as follows:

- Compute library size for each cell and denote the $\log_{10}$ of these by $L_1, L_2, \ldots L_n$, where n is the number of cells. To minimize the impact of a few genes having very large counts, we can also use trimmed sums instead of full sums here. Denote the minimum and maximum $\log_{10}$ library size as $L_{\min}$ and $L_{\max}$.
- Calculate weight for cell j as $w_j = L_j - L_{\min} / L_{\max} - L_{\min}$.
- Estimate the capture efficiency for cell j as $(1 - w_j)\eta_{\min} + w_j\eta_{\max}$. This ensures that cells with larger library size will have larger capture efficiency and the capture efficiency estimates are bounded within the $(\min_\eta, \max_\eta)$ interval.

2.3 Parameter estimation and DE analyses

DECENT's main aim is to estimate parameters of the unobserved pre-dropout count and perform statistical inference on these parameters for the purpose of identifying differentially expressed genes (DEGs). Because the pre-dropout count is unobserved, we use an ECM algorithm to estimate these parameters (see Supplementary Materials for details). The algorithm works as follows:

1. First, capture rate parameters $\boldsymbol{\eta} = (\eta_1, \eta_2, \ldots, \eta_n)^T$ are estimated prior to invoking the ECM algorithm.
2. Given the capture rates, initial estimates of cell-specific size factor parameters s_j and gene-specific parameters $\boldsymbol{\theta_i} = (\pi_{0i}, \mu_{ij}, \psi_i)^T$,

as well as $\boldsymbol{\tau}_j = (\tau_{0j}, \tau_{1j})^T$, the parameters that control the uncertainties in the beta prior for capture rates are calculated using method-of-moment approaches, assuming no DE genes and a simple Binomial capture model.

3. Perform an E-step to estimate the following conditional expectations given the observed count: $P(Y_{ij} < 0 | Z_{ij} = 0; s_j, \boldsymbol{\theta}_i, \boldsymbol{\tau}_j, \eta_j)$ and $E(Y_{ij} | Z_{ij}; s_j, \boldsymbol{\theta}_i, \boldsymbol{\tau}_j, \eta_j)$.
4. For each gene, perform an M-step to update the gene-specific parameters $\boldsymbol{\theta}_i$ (see Supplementary Materials for details).
5. For each cell, perform an M-step to update the cell-specific size-factor parameters s_j (see Supplementary Materials for details).
6. Perform an M-step to update the (possibly) cell-specific parameters of the beta prior for capture rates, $\boldsymbol{\tau}_j$ (see Supplementary Materials for details).
7. Iterate between step [2] to [5] until convergence is achieved.

To facilitate DE analysis, the gene-wise mean parameter $\boldsymbol{\mu} = (\mu_{ij})$ is assumed to depend on the cell type or group through a log-linear model:

$$\log \boldsymbol{\mu} = \mathbf{X}\boldsymbol{\beta} + \mathbf{W}\boldsymbol{\gamma}$$

where $\mathbf{X}$ is the design matrix providing group information and $\boldsymbol{\beta}$ are the coefficients. We also allow the mean parameter to depend on cell-wise covariates $\mathbf{W}$ to remove unwanted variation (e.g. batch effects, cell-cycle phases). In the most common two group comparisons, we have

$$\log \mu_{ij} = \beta_{0i} + \beta_{1i} x_j + \boldsymbol{\gamma}_i^{\mathrm{T}} \mathrm{w}_j$$

where x_j is simply the binary indicator of cellular group, β_{0i} is logarithm of mean parameter for gene i in the reference cell type, β_{1i} is the log fold-change parameter for gene i and the $\boldsymbol{\gamma}_i$ is the gene-specific regression coefficients that adjust the DE analysis for the unwanted, cell-specific factors w_j.

Differential expression across two cellular groups for the gene i is assessed by testing the hypotheses:

$$H_0 : \beta_{1i} = 0 \quad \text{versus} \quad H_1 : \beta_{1i} \neq 0$$

using the likelihood ratio test statistic,

$$-2\{\ell_I(\theta_i = \theta_i^{H_0}) - \ell_I(\theta_i = \hat{\theta}_i)\}$$

where $\theta_i^{H_0}$ is the maximum likelihood estimator (MLE) of θ_i under the restriction that $\beta_{1i} = 0$, $\hat{\theta}_i$ is the MLE under the unrestricted model and ℓ_I is the log-likelihood based on the observed data Z_{ij} [see Supplementary Materials; Equation (9)]. For simple two cell type comparisons, the statistic is approximately distributed as χ_1^2 under H_0. More generally, when performing DE across p different cell types or conditions, the statistic is approximately distributed as χ_{p-1}^2 under H_0.

3 Results

3.1 Benchmarking using simulated data

We simulated 20 datasets, each consisting of 500 cells belonging to 2 cell types (224 cells from cell type 1 versus 276 cells from cell type 2) with 3000 endogenous genes and 50 spike-ins. The observed count were generated under the DECENT model using parameters estimated from Tung's dataset (see Supplementary Materials for details). In each dataset, we set ~10% of the genes to be DEGs. Figure 2 shows that DECENT estimates gene-specific pre-dropout proportion of zeroes and variance, as well as the actual pre-dropout counts unbiasedly. Figure 3 shows that DECENT's performance in

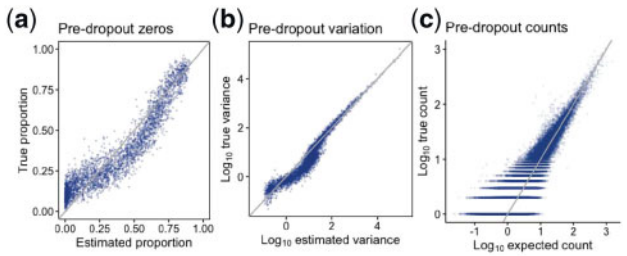

Fig. 2. Inferring pre-dropout molecule counts in simulation. (**a**) Scatter plot comparing for each gene the estimated proportion of zeros of the fitted pre-dropout distribution with the true proportion of zeros in the pre-dropout counts. (**b**) Scatter plot comparing the expected variance of the fitted pre-dropout distribution with the true gene-wise variance in the pre-dropout counts. (**c**) Scatter plot comparing the expected value of pre-dropout counts (see Supplementary Methods for details) under the fitted model with the true pre-dropout counts. We shows a random subsample of 5% of all the non-zero counts. The estimated pre-dropout counts used to calculate (**a**) and (**b**) were based on single imputation, i.e. drawing a single value from the conditional pre-dropout distribution for each gene and each cell given the observed data. The estimated pre-dropout counts shown in (**c**) were calculated as the expected value of the conditional pre-dropout distribution (see Supplementary Methods)

detecting DEGs also appear to be competitive when compared with existing methods, namely SCDE (Kharchenko *et al.*, 2014), MAST (Finak *et al.*, 2015), Monocle2 (Qiu *et al.*, 2017; Trapnell *et al.*, 2014), ZINB-WaVE adjusted edgeR (Van den Berge *et al.*, 2018) and edgeR (McCarthy *et al.*, 2012). Over the 20 datasets, the mean(SD) of the partial area under the receiver operating characteristic (pAUROC) for DECENT is 0.708(0.001), followed by MAST with 0.687(0.001) (see Supplementary Table S3). DECENT's performance also appears to be relatively robust to misspecification of capture rates parameters (Supplementary Fig. S7).

3.2 Benchmarking using real data

We further benchmarked our method against existing methods using real datasets. The difficulty in benchmarking using real datasets is that the genuine DEGs are usually unknown. In order to obtain a credible list of genuine DEGs, we searched for scRNA-seq datasets that have matching bulk RNA-seq experiments, which means a bulk RNA-seq was also performed using cells from exactly the same tissues or cell lines. We found four such experiments in total that also used UMI. Then a DEG list derived from these bulk data can be used as the reference set for benchmarking. These includes two plate-based experiments and two droplet-based experiments, with different scales, sources of tissues or cell lines and observed proportion of zero counts (Supplementary Table S2) (Chen *et al.*, 2018; Savas *et al.*, 2018; Soumillon *et al.*, 2014; Tung *et al.*, 2017). We use these datasets to benchmark DECENT against existing methods (see Supplementary Materials: Benchmarking for more details about these datasets).

The same existing methods were benchmarked using all four datasets, except that we also applied TASC to the Tung *et al.* data where spike-ins are available. All of these datasets have gone through careful quality control steps by the authors of the original publications. Therefore, we do not further filter any cells. The only further filtering we perform is filtering very low abundance genes that correspond to the second peak in the histogram of average gene count distribution (Supplementary Fig. S13). We fitted the DECENT model to all four datasets and found that a global $\boldsymbol{\tau}$ parameter was adequate for all datasets, except Tung's where

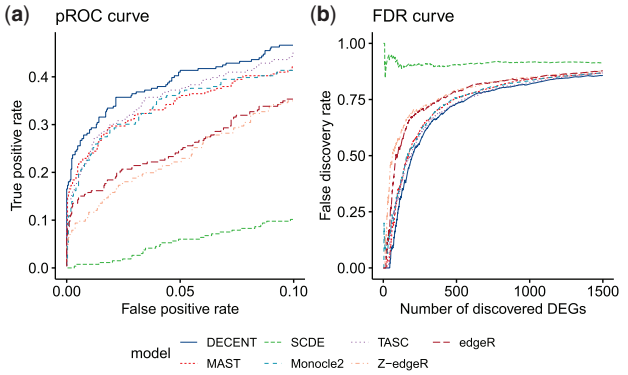

Fig. 3. Differential expression analysis of simulated data. (**a**) Partial receiver operating characteristic curve for differential expression methods on the simulated data. (**b**) False discovery rate curves for differential expression methods on the simulated data. Both curves only focus on the low *P*-value region, since other regions were of little interest in actual DE analysis. Z-edgeR stands for ZINB-WaVE-adjusted edgeR

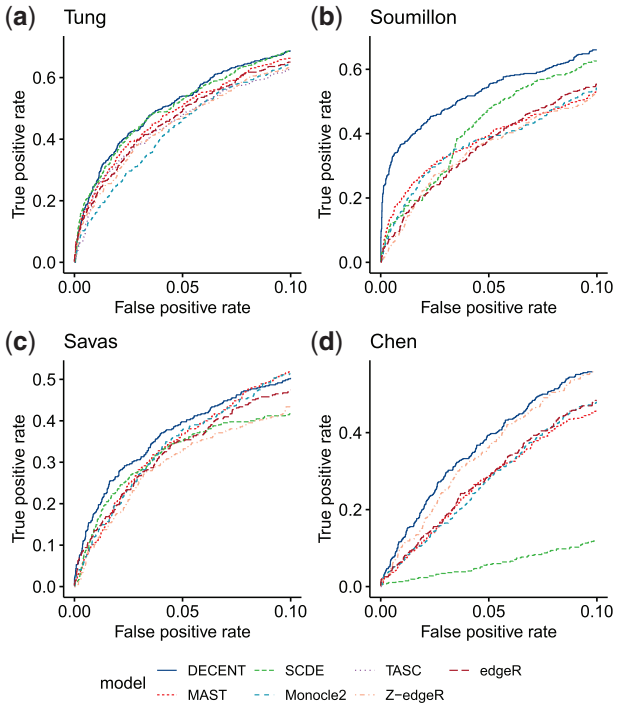

Fig. 4. Partial receiver operating characteristic curves for differential expression methods on real datasets. Evaluating the performance of different methods by partial receiver operating characteristic curves using the (**a**) Tung *et al.,* (**b**) Soumillon *et al.,* (**c**) Savas *et al.* and (**d**) Chen *et al.* datasets. DEGs from matching bulk RNA-seq data were used as gold-standard for benchmarking. DECENT achieves competitive accuracy of identifying genuine DEGs in all four datasets. We used pROC to focus on the low *P*-value region with high specificity. DE methods are denoted by different colors. Z-edgeR stands for ZINB-WaVE-adjusted edgeR. TASC requires spike-ins and was only evaluated using the Tung *et al.* data

substantial variation in the τ_0 and τ_1 parameters was observed (Supplementary Fig. S14). As shown in Figures 4 and 5, DECENT showed superior performance on all four datasets. In particular, DECENT performs better than other methods for the Soumillon and Chen datasets where the dispersion parameters for the beta prior are larger and thus the capture models are more overdispersed relative to the Binomial model (Supplementary Fig. S15).

Among the other methods, MAST showed stable and generally good performance across datasets, while the performance of SCDE appeared to be dataset-specific, showing inadequacy for droplet-based experiments. The Monocle NB-based model based on observed UMI counts did not show satisfactory performance. The ZINB-WaVE adjustment of edgeR did not show noticeable improvements over standard edgeR for three out of four datasets. But it remarkably outperformed edgeR on the Chen *et al.* data, where both molecule counts and the cell numbers were high. To demonstrate the merit of performing DE analysis using a inferred pre-dropout rather than the observed expression, we selected a few genuine DEGs in the Tung *et al.* data that are detected by our method and not the others and compared their expression levels between the two cellular groups using either the observed counts or inferred pre-dropout counts. We discovered that the differential expression between two groups became more prominent in the pre-dropout counts (Supplementary Fig. S8).

ERCC spike-ins were available in Tung *et al.* data. We thus used capture rates estimated from spike-ins for the result shown. This dataset also enabled us to examine how specifying the ranked random capture rates impacts DE performance on real data. We performed DECENT DE analysis again using the ranked random capture rates specifying the range as half, the same and 1.5 times the range of the spike-in estimates. The results turned out to be in concordance with the simulation studies. Although optimal performance was achieved when capture rates estimated from spike-ins were used, there were only small decreases in performance when using the ranked random capture rates (Supplementary Fig. S9). This convincingly demonstrated the viability of using the spike-in capture rates for endogenous RNA and that DECENT's DE performance is also robust to misspecified capture rates.

For the Soumillon *et al.* data, the median of the log fold-change estimates deviates from zero when the standard MLEs were used to estimate the cell size factors s_j. This default size factor estimator effectively performs library size normalization on the inferred pre-dropout counts. The bias is greatly reduced when using the trimmed mean of *M* values method (Robinson and Oshlack, 2010) to estimate the size factors instead and the overall performance of DECENT was slightly improved (Supplementary Fig. S10). This suggests that different datasets may require different normalization strategies, and highlights the flexibility of our method with regards to normalization.

The benchmarking so far was based on two group comparisons. DECENT performs statistical tests under the under the well-established GLM framework and can readily accommodate more complex experimental designs. The Soumillon *et al.*'s data are a time course experiment, with three time points involved in adipose stem cell differentiation. This allowed us to have a glance at how different DE methods perform on more complex UMI-based scRNA-seq experiments beyond two-group comparisons. We tested the hypothesis that expression of a gene is constant across the three time points. Except for SCDE, which is designed only for two group comparison, and TASC, which requires spike-ins, other methods were compared in this setting. The reference genuine DEGs across the three time points were also derived from the matching bulk experiments. DECENT again outperformed all other methods with an even more pronounced advantage (Supplementary Fig. S11).

In terms of controlling type I error, our in-silico investigation (see Supplementary Materials and Fig. S12) demonstrates that DECENT controls the type I error quite well.

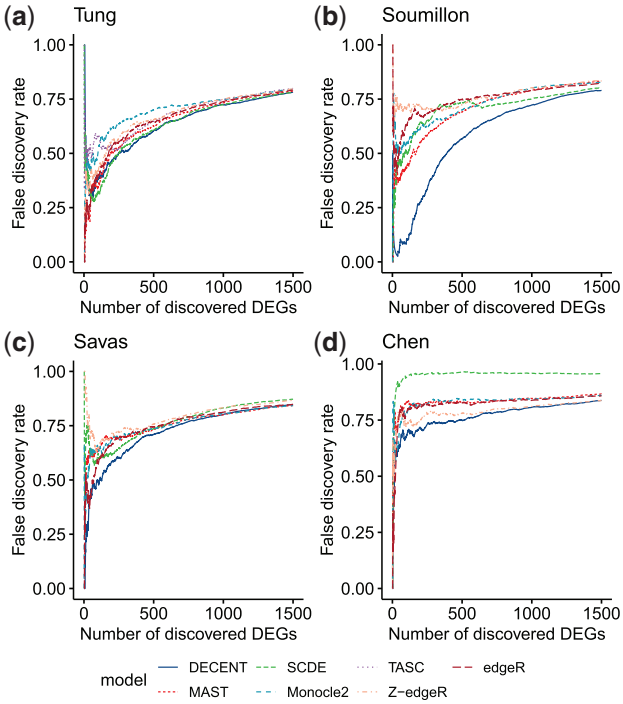

Fig. 5. False discovery rate (FDR) curves for differential expression methods on real datasets. Evaluating the performance different method by FDR curves using the (**a**) Tung et *al.,* (**b**) Soumillon *et al.,* (**c**) Savas *et al.* and (**d**) Chen *et al.* datasets. Bulk DEGs were considered as conditional positives. DECENT consistently showed low number of false discoveries at the same number of declared DEGs across all four datasets. Again, only the top one thousand DEGs were considered to focus on the region of interested. DE methods are denoted by different colors. Z-edgeR denotes ZINB-WaVE-adjusted edgeR. TASC requires spike-ins and was only evaluated using the Tung *et al.* data

4 Discussion

We presented DECENT, a novel statistical method for performing DE analysis on UMI-based scRNA-seq data. UMI count data have provided us with an excellent opportunity to model the molecule capturing process. We found that the technical variation arising in this process can be characterized by a gene and cell-specific beta-binomial capture model. We were able to perform DE analysis on the inferred pre-dropout counts, and achieve good performance. Our model is usable either with or without spike-ins and is compatible with different normalization strategies. Also, we can draw on established statistical theory and use the model for analyzing data from experiments and studies more complex than two group comparisons. Used in a three group setting, the model gave promising results. Adding more cell-level covariates is straightforward (see Section 2) and is catered for in our software. External RNA spike-ins, such as the ERCC spike-ins (Jiang *et al.*, 2011) can give valuable insights into the technical variation in scRNA-seq data. They have been used in this way in some scRNA-seq methods (Jia *et al.*, 2017; Lun *et al.*, 2017). However, spiked-in molecules differ from endogenous transcripts in properties such as overall length, and length of the poly(A) tract, and in their technical variation such as capture rates (Svensson *et al.*, 2017). This raises the question of whether and how to use spike-ins in analyses like ours. When they are available we use them to estimate the capture rates in our model, that is, to center the beta-binomial distribution in each cell. During the development of our beta-binomial capture model, we found more variation in the data than would be found in a cell-specific binomial capture model. This extra variation is more likely to be due to spike-specific biases in capture rates rather than due to random noise (Supplementary Fig. S2). However, unlike cell-specific capture rates, the estimated spike-specific biases cannot be generalized to endogenous genes. Indeed we are unable to estimate the such gene-specific biases using only gene abundance, because it is not separable from a gene's mean expression. Such a separation would only be achievable if extra information was available. For example, it is plausible that capture rates would depend on features of the gene sequence such as GC-content and the length of the poly(A) tract. A more refined capture model might then be built by modeling the relationship between these gene-specific features and the gene-specific biases of capture rates. Fortunately, our method is flexible enough to permit the amount of over-dispersion in the capture process to differ between the spike-ins and endogenous genes to reflect any differences in the capture process of the two types of molecules. In this way we deal with the issue just mentioned.

Although multilevel models fitted with an ECM algorithm are intrinsically computationally intensive, DECENT has achieved acceptable speed with a series of acceleration approaches such as a gaussian quadrature approximation for large integration and parallelization of all the main steps. For instance, our simulated data with 500 cells and 3000 genes took 18 min, while the largest dataset, Chen *et al.* with 6875 cells and 12 929 genes took ∼8 h to finish on a 28-core XENON Radon Duo R1885 server node with Intel(R) Xeon(R) E5-2690 v4 CPUS @ 2.60 GHz. Some existing models for scRNA-seq allow tests beyond the comparison of means, such as comparisons of zero fractions, of biological variation or even the overall distributions (Korthauer *et al.*, 2016; Wang *et al.*, 2018; Wu *et al.*, 2018). But there remains a difficulty in assessing the type I error control and power of these tests due to lack of ground-truth. Single molecule FISH technology is under rapid development and is able to produce accurate measurements of distributions, biological variation and the zero fractions. As the amount of such data and number of genes profiled in this way increases, we should soon have the opportunity to assess these tests objectively.

While DECENT focuses on performing reliable statistical tests concerning gene mean abundance, it can be easily be extended to carry out other types of tests. For example, we can permit the zero-inflation parameter in the pre-dropout distribution to be a function of cell type. Then a linear logistic model can be used to test for biological differences in zero inflation. However, some alteration of the parameter estimation strategy may be needed to achieve valid testing results.

We have not incorporated any forms of Empirical Bayes (EB) in the estimation of DECENT model parameters. Several bulk RNA-seq methods for differential expression such as edgeR (McCarthy *et al.*, 2012) and DESeq2 (Love *et al.*, 2014) use EB shrinkage to stabilize estimates of gene-specific dispersion parameter. Among methods for scRNA-seq data, MAST (Finak *et al.*, 2015) uses EB to shrink the gene-specific variance parameter. Given that scRNA-seq data are very sparse, we certainly think there is potential benefit in using EB to improve DECENT's performance. One major challenge is there is no natural conjugate prior for the dispersion parameter of the ZINB or NB distribution that we use to model the count data. This is in contrast to MAST that uses Gaussian distribution to model the log(TPM) + 1 data and therefore able to take advantage of the inverse Gamma conjugate prior for the variance parameter. DECENT also estimates the parameters associated with the unobserved rather than the observed data, which makes it less straightforward to introduce Bayes or EB variants. Our current thinking is that to implement EB within DECENT, we need to either use a form of

variational Bayes (Beal, 2003) or something similar to the weighted likelihood method (Robinson and Smyth, 2007) that has been successfully used for bulk RNA-seq data.

Acknowledgements

We would like to thank three anonymous reviewers for their constructive comments that have improved the quality of this article and Alexander Rhys Hayes for his assistance in preparing some of the Supplementary Figures.

Funding

T.P.S. was awarded by Australian National Health and Medical Research Council Program Grant 105461.

Conflict of Interest: none declared.

References

Bacher,R. *et al.* (2017) Scnorm: robust normalization of single-cell RNA-seq data. *Nat. Methods*, **14**, 584.

Beal,M.J. (2003) Variational algorithms forapproximate bayesian inference. PhD Thesis, The Gatsby Computational Neuroscience Unit, University College London, London.

Brennecke,P. *et al.* (2013) Accounting for technical noise in single-cell RNA-seq experiments. *Nat. Methods*, **10**, 1093–1095.

Chen,W. *et al.* (2018) Umi-count modeling and differential expression analysis for single-cell RNA sequencing. *Genome Biol.*, **19**, 70.

Codeluppi,S. *et al.* (2018) Spatial organization of the somatosensory cortex revealed by osmFISH. *Nat. Methods*, **15**, 932–935.

Finak,G. *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.

Grun,D. *et al.* (2014) Validation of noise models for single-cell transcriptomics. *Nat. Methods*, **11**, 637–640.

Hashimshony,T. *et al.* (2012) Cel-seq: single-cell RNA-seq by multiplexed linear amplification. *Cell Rep.*, **2**, 666–673.

Huang,M. *et al.* (2018) Saver: gene expression recovery for single-cell RNA sequencing. *Nat. Methods*, **15**, 539–542.

Islam,S. *et al.* (2014) Quantitative single-cell RNA-seq with unique molecular identifiers. *Nat. Methods*, **11**, 163–166.

Jia,C. *et al.* (2017) Accounting for technical noise in differential expression analysis of single-cell RNA sequencing data. *Nucleic Acids Res.*, **45**, 10978–10988.

Jiang,L. *et al.* (2011) Synthetic spike-in standards for RNA-seq experiments. *Genome Res.*, **21**, 1543–1551.

Kharchenko,P.V. *et al.* (2014) Bayesian approach to single-cell differential expression analysis. *Nat. Methods*, **11**, 740–742.

Kiselev,V.Y. *et al.* (2017) Sc3: consensus clustering of single-cell RNA-seq data. *Nat. Methods*, **14**, 483.

Klein,A.M. *et al.* (2015) Droplet barcoding for single-cell transcriptomics applied to embryonic stem cells. *Cell*, **161**, 1187–1201.

Kolodziejczyk,A.A. *et al.* (2015) The technology and biology of single-cell RNA sequencing. *Mol. Cell*, **58**, 610–620.

Korthauer,K.D. *et al.* (2016) A statistical approach for identifying differential distributions in single-cell RNA-seq experiments. *Genome Biol.*, **17**, 222.

Love,M.I. *et al.* (2014) Moderated estimation of fold change and dispersion for RNA-seq data with deseq2. *Genome Biol.*, **15**, 550.

Lun,A.T. *et al.* (2016) Pooling across cells to normalize single-cell RNA sequencing data with many zero counts. *Genome Biol.*, **17**, 75.

Lun,A.T. *et al.* (2017) Assessing the reliability of spike-in normalization for analyses of single-cell RNA sequencing data. *Genome Res.*, **27**, 1795–1806.

Macosko,E.Z. *et al.* (2015) Highly parallel genome-wide expression profiling of individual cells using nanoliter droplets. *Cell*, **161**, 1202–1214.

McCarthy,D.J. *et al.* (2012) Differential expression analysis of multifactor RNA-seq experiments with respect to biological variation. *Nucleic Acids Res.*, **40**, 4288–4297.

Qiu,X. *et al.* (2017) Reversed graph embedding resolves complex single-cell trajectories. *Nat. Methods*, **14**, 979.

Raj,A. *et al.* (2008) Imaging individual mRNA molecules using multiple singly labeled probes. *Nat. Methods*, **5**, 877.

Ramskold,D. *et al.* (2012) Full-length mRNA-seq from single-cell levels of RNA and individual circulating tumor cells. *Nat. Biotechnol.*, **30**, 777–782.

Risso,D. *et al.* (2018) A general and flexible method for signal extraction from single-cell RNA-seq data. *Nat. Commun.*, **9**, 284.

Robinson,M.D., and Oshlack,A. (2010) A scaling normalization method for differential expression analysis of RNA-seq data. *Genome Biol.*, **11**, R25.

Robinson,M.D., and Smyth,G.K. (2007) Moderated statistical tests for assessing differences in tag abundance. *Bioinformatics*, **23**, 2881–2887.

Savas,P. *et al.* (2018) Single-cell profiling of breast cancer t cells reveals a tissue-resident memory subset associated with improved prognosis. *Nat. Med.*, **24**, 986–993.

Soumillon,M. *et al.* (2014) Characterization of directed differentiation by high-throughput single-cell RNA-seq. doi: 10.1101/003236.

Sun,Z. *et al.* (2018) Single-cell RNA sequencing reveals gene expression signatures of breast cancer-associated endothelial cells. *Oncotarget*, **9**, 10945–10961.

Svensson,V. *et al.* (2017) Power analysis of single-cell RNA-sequencing experiments. *Nat. Methods*, **14**, 381–387.

Trapnell,C. *et al.* (2014) The dynamics and regulators of cell fate decisions are revealed by pseudotemporal ordering of single cells. *Nat. Biotechnol.*, **32**, 381.

Tung,P.-Y. *et al.* (2017) Batch effects and the effective design of single-cell gene expression studies. *Sci. Rep.*, **7**, 39921.

Van den Berge,K. *et al.* (2018) Observation weights unlock bulk RNA-seq tools for zero inflation and single-cell applications. *Genome Biol.*, **19**, 24.

van Dijk,D. *et al.* (2018) Recovering gene interactions from single-cell data using data diffusion. *Cell*, **173**, 716–729.e27.

Vieth,B. *et al.* (2017) powsimr: power analysis for bulk and single cell RNA-seq experiments. *Bioinformatics*, **33**, 3486–3488.

Wagner,A. *et al.* (2016) Revealing the vectors of cellular identity with single-cell genomics. *Nat. Biotechnol.*, **34**, 1145–1160.

Wang,J. *et al.* (2018) Gene expression distribution deconvolution in single-cell RNA sequencing. *Proc. Natl. Acad. Sci.*, **115**, E6437–E6446.

Wu,Z. *et al.* (2018) Two-phase differential expression analysis for single cell RNA-seq. *Bioinformatics*, **34**, 3340–3348.

Zeisel,A. *et al.* (2015) Cell types in the mouse cortex and hippocampus revealed by single-cell RNA-seq. *Science*, **347**, 1138–1142.

Zhao,X. *et al.* (2017) Single-cell RNA-seq reveals a distinct transcriptome signature of aneuploid hematopoietic cells. *Blood*, **130**, 2762–2773.

Zheng,G.X.Y. *et al.* (2017) Massively parallel digital transcriptional profiling of single cells. *Nat. Commun.*, **8**, 14049.

Ziegenhain,C. *et al.* (2017) Comparative analysis of single-cell RNA sequencing methods. *Mol. Cell*, **65**, 631–643.e4.