DECENT: Differential Expression with Capture Efficiency AdjustmeNT for Single-Cell RNA-seq Data

Chengzhong Ye[2,3], Terence P Speed[2,4], Agus Salim[1,2]

[1] *Mathematics and Statistics, La Trobe University, Melbourne, Australia*
[2] *Bioinformatics Division, Walter and Eliza Hall Institute of Medical Research*
[3] *School of Medicine, Tsinghua University, Beijing, People's Republic of China*
[4] *School of Mathematics and Statistics, University of Melbourne*

Abstract

Due to the small amount of starting material and the low capture efficiency of the current technologies, scRNA-seq data contains a large number of zero counts due to the dropout phenomenon. The unique features of scRNA-seq data have led to the development of novel methods for differential expression (DE) analysis. However, few of the existing DE methods for scRNA-seq data estimate the number of molecules pre-dropout and therefore do not explicitly distinguish technical and biological zeroes. We develop DECENT, a DE method for scRNA-seq data that adjusts for the imperfect capture efficiency by estimating the number of molecules pre-dropout. Using simulated and real datasets, we saw the performance of our method is better compared to previously published methods. DECENT uses raw UMI-count data as input and does not require spike-ins, but when spike-ins are available, they can be used to improve its performance.

Background

Recent developments in sequencing technology have enabled high-throughput sequencing of transcriptomes at the single-cell level. Single-cell technologies have already led to profound new discoveries that could not be have been made using data from bulk transcriptome sequencing. Examples of these discoveries range from the identification of novel cell types to the study of global patterns of stochastic gene expression (Stegle, Teichmann and Marioni, 2015). However, there are still considerable statistical challenges in drawing inferences from single-cell RNA-seq (scRNA-seq) data. Due to the small amount of starting material and the low capture efficiency of the current technologies used to generate scRNA-seq data, the phenomenon of dropout, in which only a small fraction of the mRNA molecules in the cell is captured and amplified, is still quite common. This leads to scRNA-seq data containing a large number of zero counts and exhibiting bimodal distributions. Also, dropout rates can vary between cells and across genes, with less abundant genes and cells with smaller amounts of starting material exhibiting higher proportions of zero counts (Brennecke *et al.*, 2014).

Dropout is a also 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.*, 2017). The large numbers of zero counts has led to the need to develop novel normalization methods for scRNA-seq data (Lun *et al.*, 2015 ; Bacher *et al.*, 2017).

After performing quality control and normalization steps, the first task when analysing scRNA-seq data is usually identifying cell-types, using one or more of a variety of clustering methods. Once clusters are identified, differential expression (DE) analysis is often performed, to identify marker genes for each cluster found. The unique features of scRNA-seq data have led to the development of novel methods for differential expression analysis. These methods adjust for technical variation due to dropout in different ways. Kharchenko *et al.* (2015) use a mixture model to distinguish counts affected by dropout from the rest of the data, and follow that by the use of a cell-specific zero-inflated negative binomial (ZINB) model on genes with large enough expression levels to be less affected by dropout. Finak *et al.* (2015) proposed a two-part generalized linear model in which the dropout rates are adjusted by the inclusion of the proportion of genes with non-zero counts as a term in their regression model. BPSC (Vu *et al.*, 2016) proposed a model based on a Beta-Poisson mixture, to capture the bimodality of scRNA-seq data. One consequence of imperfect capture efficiency is the existence of both *technical* and *biological* zeroes in scRNA-seq data. *Biological zeroes* occurs when the zero counts comes from genes that are truly silent. On the other hand, *technical zeroes* occur when the zero counts come from genes that are expressed but due to the low capture efficiency, none of the molecules from these genes are sequenced. Statistically, these two kinds of zeroes need to be treated differently. For example, if we are to use imputation methods (e.g., Huang *et al.*, 2017) to estimate the true number of molecules pre-dropout, ideally only the technical zeroes need to be imputed.

Few of the existing DE methods for scRNA-seq data distinguish technical and biological zeroes. For example, SCDE (Kharchenko et al, 2015) almost always assigns a probability of 1 to zero counts belonging to the dropout component, thus assuming all zeroes are technical zeroes. MAST (Finak *et al.*, 2015) does not differentiate the two kinds of zeroes, instead it simply categorizes all zero counts as 'unexpressed'. Even for genes with non-zero count, a low capture efficiency has effect of reducing the number of molecules captured. This effect 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. More recently, Jia *et al.* (2017) proposed a method that relies on spike-ins to estimate the number of molecules pre-dropout and performs DE analysis based on the true number of molecules. Using simulated data, they showed the superior performance of their method compared to methods that perform DE analysis directly on the observed data. In this paper, we propose a method for performing DE analysis on scRNA-seq data. As in Jia *et al.* (2017), our method estimates the number of molecules pre-dropout but unlike theirs, our method works even when spike-ins are not available. Another difference is that our method assumes a zero-inflated distribution for the pre-dropout counts, to cater for the high proportion of biological zeroes in otherwise highly-expressed genes.

Methods

DECENT Models

DECENT assumes that unique molecular identifiers (UMI, Islam *et al.*, 2013) have been used in the scRNA-seq experiment and that the data is available in UMI-count format. To allow differentiation between *biological zeroes* and *technical zeroes*, we first assume that in the idealized setting where all molecules are captured (no dropout), the observed count y_{ij} for gene i in cell j can be modelled as a zero-inflated negative binomial (ZINB) random variable with parameters $\theta_{ij} = (\pi_{0i}, \mu_{ij}, s_j, \psi_i)$, where π_{0i} is a gene-specific zero-inflation parameter, ψ_i is gene-specific dispersion parameter, and the mean parameter is $s_j\mu_{ij}$, where s_j is the size factor for cell j that measures differences in the amount of starting material between cells.

We use parameterization in Ridout, Hinde and Demétrio (2001) for defining the ZINB distribution. Under this paramerization, the probability of of observing k molecules in the idealized setting with no dropout is given by,

$$p(y_{ij} = k; \theta_{ij}) \quad = \quad (1 - \pi_{0i}) \frac{\Gamma(\psi_i + k)}{k! \Gamma \psi_i} \left(\frac{\psi_i}{\psi_i + s_j \mu_{ij}} \right)^{\psi_i} \left(\frac{s_j \mu_{ij}}{\psi_i + s_j \mu_{ij}} \right)^{k}, \qquad k>0$$

while the probability of a biological zero is given by,

$$p(y_{ij} = 0; \theta_{ij}) \quad = \quad \pi_{0i} + (1 - \pi_{0i}) \left(\frac{\psi_i}{\psi_i + s_j \mu_{ij}} \right)^{\psi_i}$$

For lowly-expressed genes with small mean parameter μ_{ij}, the contribution from the second component above can be quite significant, but for highly-expressed genes, the probability of a biological zero largely depends on π_{0i}, with larger values of this parameter being associated with higher probabilities of a biological zero.

The mean parameter μ_{ij} ('standardized' for unit cell size) is assumed to depend on the cell-type through a log-linear model

$$\log \mu_{ij} = \beta_{0i} + \beta_{1i} x_j$$

where in the simple two cell type comparison, x_j is simply the binary indicator of cell-type and β_{1i} has interpretation as the log-fold change (log FC) parameter for gene i.

We could also model the zero-inflation parameter as a function of cell type through a logistic regression model,

$$\log \frac{\pi_{0ij}}{1 - \pi_{0ij}} = \alpha_{0i} + \alpha_{1i} x_j$$

However, in the current version of DECENT we do not implement this option, but assume that $\alpha_{1i} = 0$ for all genes, resulting in equal π_{0ij} across cells.

In reality, except for the spike-in transcripts, we do not know y_{ij}, instead we have the observed counts z_{ij} that are what remains of the y_{ij} after dropout. DECENT uses a cell-specific beta-binomial distribution to model the dropout process and empirical evidence shows that the model fits spike-ins data from experiments rather well (see Figure 3(a)). We suppose,

given that y_{ij} molecules are available before RT, that the observed-count z_{ij} follow a beta-binomial distribution

$$P(z_{ij} = l \mid y_{ij} = k) = \binom{k}{l} \frac{B(l + a_j, k - l + b_j)}{B(a_j, b_j)},$$

where $B(.,.)$ is the Beta function and a_j and b_j are cell-specific parameters that satisfy $\frac{a_j}{a_j+b_j} = \eta_j$ and $a_j + b_j + 1 = \rho_j^{-1}$. The cell-specific capture efficiency parameter η_j can be estimated using spike-ins, see the **DECENT without spike-ins** section below for how to use DECENT with datasets without spike-ins. The ρ_j parameter controls the amount of over-dispersion. We found that setting this parameter in the range of 0.15-0.25 equally across all cells gives a good performance. When $\rho_j = 0$, the Beta-binomial model will reduce to a Binomial model, with the implicit assumption that molecules are captured independently rather than in clusters.

Since the pre-dropout counts are not observed, we use an EM algorithm (see Appendix) to estimate the gene-specific parameters $\theta_i = \{\alpha_{01}, \beta_{0i}, \beta_{1i}, \psi_i\}$ and cell-specific parameters s_j. Not surprisingly, the E-step involves evaluating the conditional probability of an observed zero count being a biological zero, $P(E_{ij} = 0 \mid z_{ij} = 0; \theta_i^0, s_j^0) = 1 - P(E_{ij} = 1 \mid z_{ij} = 0; \theta_i^0, s_j^0)$. Supplementary Figure 1 shows that under Beta-Binomial dropout model, the overall probability of observing a zero count is higher for genes with low expression (Supp. Figure 1(a)) and cells with smaller size factors (less transcriptional activity). However, the conditional probability of a zero count being a biological zero is higher for more highly expressed genes (Supp. Figure 1(b)) and for cells with larger size factors (Supp. Figure 1(d)). This reflects the fact that while the proportion of zero counts is lower more for highly-expressed genes and for cells with larger size, if we observe a zero count there, it is more than likely that the zero count is a biological rather than a technical zero.

Estimating Capture Efficiency Parameter Using Spike-ins

Some scRNA-seq experiments have the ability to add spike-ins with known concentrations to the single-cell mRNA mix. When spike-ins are available, we can use them to estimate the capture efficiency. Suppose that prior to RT step, we added T spike-in transcripts at the known concentrations $c_1, c_2, \ldots c_T$ into cell j and subsequently observe $u_{1j}, u_{2j}, \ldots u_{Tj}$ molecules respectively, for these transcripts. Then, assuming that these molecules are captured independently and that the capture efficiency for different transcripts is constant, the capture efficiency for any cell can be estimated as the proportion of molecules observed after sequencing relative to the total number of molecules initially added,

$$\hat{\eta}_j = \frac{\sum_{i=1}^{T} u_{ij}}{\sum_{i=1}^{T} c_i}$$

A better estimate for capture efficiency can also be obtained by applying the beta-binomial dropout model to the spike-in transcripts. Assuming that $U_{ij} \sim \texttt{Beta-Binomial}(c_i, a_j, b_j)$, we can estimate the parameters of the Beta-binomial distribution numerically using maximum-likelihood method, and then estimate the capture efficiency by

$$\hat{\eta}_j = \frac{\hat{a}_j}{\hat{a}_j + \hat{b}_j}$$

where $\hat{a}_j$ and $\hat{b}_j$ are the MLEs of the parameters a_j and b_j.

If there is evidence that capture-efficiency depends on factors such as GC-content or the length of the poly-A tail of the transcripts, we can allow capture-efficiency to vary between transcripts by using a regression model on gene and cell-specific Beta-binomial parameters a_{ij} and b_{ij}. For example, we can use the log-linear models,

$$\begin{aligned} \log a_{ij} &= \nu_{0j} + \nu_{1j}\texttt{GC}_i \\ \log b_{ij} &= \kappa_{0j} + \kappa_{1j}\texttt{GC}_i \end{aligned}$$

Given estimates of $(\nu_{0j}, \nu_{1j}, \kappa_{0j}, \kappa_{1j})$, transcript-specific capture efficiencies can be calculated for all transcripts, including the endogenous transcripts. Currently, this regression option is not implemented in DECENT.

DECENT without Spike-ins

There are some cell-sorting technologies such as the droplet-based ones where the addition of spike-ins is not currently feasible. For single-cell RNA-seq data generated without spike-ins, DECENT can still be used to carry out DE analyses. We are fortunate that due to a property of the ZINB distribution that we will explain here, our approximation turns out to be quite robust.

Let Y be the pre-dropout count where $Y \sim \text{ZINB}(\pi_0, s\mu, \psi)$. Given Y, the observed data Z follows Beta-Binomial distribution, $Z \mid Y = y \sim \text{Beta-Binomial}(y, a, b)$, so that the capture efficiency is $\eta = \frac{a}{a+b}$. It turns out that the marginal distribution of Z can be approximated by the marginal distribution of Z', so that

$$F_Z(z; \pi_0, s, \mu, \psi, \eta) \approx F_{Z'}(z; \pi_0, s', \mu, \psi, \eta')$$

where $Z' \mid Y' = y' \sim \text{Beta-Binomial}(y', a', b')$ and $Y' \sim \text{ZINB}(\pi_0, s'\mu, \psi)$, with $a' = \frac{\frac{s'}{s}\eta}{\frac{s'}{s}-\eta}$ and $b' = \frac{s'}{s}b$, so that the new capture efficiency is $\eta' = \frac{s'}{s}\eta$. Figure 1 shows that this approximation is actually exact when molecules are captured independently, leading to Binomial dropout model but even when molecules are captured as clusters and there is considerable over-dispersion, the approximation is still quite good.

The above result means that if we misspecify the capture efficiency by using η' rather than η, the misspecification can be corrected by scaling the size-factor estimates accordingly. Of course getting the capture-efficiency estimates as close as possible to the true value is preferable. Motivated by the above results and our experience with real datasets showing that capture-efficiency is the biggest factor contributing to library size variations, we devised a method to estimate capture-efficiency when spike-ins are not available:

- Compute library size for each cell and denote these as $L_1, L_2, \ldots L_N$. To minimize the impact of a few genes having very large counts, we can also use trimmed sums instead of full sums here.
- Order the library size from the smallest to the largest and denote these by $L_{(1)}, L_{(2)}, \ldots L_{(N)}$.
- Generate N random capture efficiency parameters from a uniform $U(min, max)$ distribution and order these random numbers as $u_{(1)}, u_{(2)}, \ldots u_{(N)}$

- Assign $u_{(1)}$ as the capture efficiency parameter for the cell with the smallest library size, $u_{(2)}$ as capture efficiency parameter for cell with the second smallest library size, and so on.

Datasets

- Simulated Dataset. We simulated five independent datasets where the pre-dropout molecule counts were generated from ZINB distributions. The gene-specific zero-inflated parameters were generated from a Beta(3,27) distribution so that the average proportion of biological zeroes across genes is approximately 10%. The gene-specific mean transcript counts were generated from a Gamma(1,20) distribution so the average transcript count across genes is 20, and the size parameters were generated from a U(3,8) distribution. We generated counts for 1000 genes and 200 cells belonging to two cell-types, 100 cells from each type. We also assumed that the second cell-type had on average smaller size factors than the first cell-type. Approximately 25% of the genes are differentially-expressed with the fold-change parameters generated from a Gamma(3,1/3) distribution. 92 ERCC spike-in transcripts with known numbers of molecules are also added to each cell. Both the endogenous and the spike-ins molecules were then subjected to imperfect capture, with cell-specific capture rates randomly generated between 2% and 25%). The beta-binomial model with correlation parameter 0.001 and cell-specific capture rate was used to generate the observed numbers of molecules.

- Tung dataset: This dataset is part of the Tung *et al.* (2017) benchmarking scRNA-seq data and is publicly-available from the GEO website (ID: GSE77288) and their GitHub repository (https://github.com/jdblischak/singleCellSeq). The full dataset consists of single-cell and bulk RNA-seq data from three Yoruba (YRI) induced pluripotent stem cell (iPSC) lines. The single cells were captured using a Fluidigm C1 machine with 96 wells and for each cell line, three independent C1 collections were made. ERCC spike-ins (Jiang *et al.*, 2011) were added to each single cell and unique molecular identifiers (UMI) were used. Samples were sequenced using an Illumina Hiseq 2000 machine. Each replicate was accompanied by a matched bulk sample using the same reagents. For the purpose of our illustration, we used only data for the NA19101 and NA19239 cell lines and we compared the single-cell gene expression profiles between the two cell lines. Reference DE genes were derived by analyzing the bulk data using `limma-voom` and ranking the DE genes by their p-values.

- Soumillon dataset: This dataset is publicly-available from the GEO website (ID: GSE53638). Cells were collected at different stages and different timepoints of directed differentiation of human adipose-derived stem/stromal cells. Cells were FACS-sorted into 384-well plates with three plates per condition. UMI were used but no spike-ins were added. After sorting, the samples were sequenced using the SCRB-seq protocol. For our purposes, we compare the stage-3 differentiated cells at day 0 (baseline) versus day 7. We chose this as the only subset of the single cell data for which matched bulk data was available. We wanted to use the bulk data to derive reference DE genes.

Results

Simulated Dataset

Compared with MAST and SCDE, DECENT seems better at identifying DE genes when data are simulated under our model, see Figure 2. Since spike-ins were available for the simulated dataset, we could use them to estimate the capture efficiency (Figure 2, black solid line). When we ignore the spike-in data and instead use library size information (see DECENT without spike-ins) to estimate capture efficiency, we see only a slight decrease in the performance of DECENT (Figure 2, various dotted lines) and DECENT still performs better than MAST and SCDE when spike-ins are not available.

Supplementary Figure 2(a) shows that for one of the simulated datasets, DECENT imputes the observed counts in an unbiased manner, and on average recovers the pre-dropout counts well. The size factors, log mean expression and the log fold-changes are also estimated in an unbiased manner (Supp. Figure 2(b)-(d)). In particular, the size factor estimates reflect that on average the second cell-type has lower transcriptional content than the first. The results for other simulated datasets are similar and so are not shown.

Tung Dataset

We downloaded the filtered molecule count matrix from the GitHub repository and performed comparisons between cells from individual NA19101 and NA19239. There are 201 cells from the NA19101 and 221 cells from the NA19239 cell line that passed QC filtering. We compared DECENT with and without spike-ins to the following methods for scRNA-seq data: MAST (Finak *et al.*, 2015), SCDE (Kharchenko *et al.*, 2015), Monocle (Trapnell *et al.*, 2014), zingeR (Van Den Berge *et al.*, 2017) and edgeR (Robinson, McCarthy and Smyth, 2010) for bulk RNA-seq data. Figure 3(a) shows that our estimated dropout model fits the dropout patterns in the spike-in data really well. The volcano plot (Figure 3(b)) is also well-behaved, with very few genes having small log FC being statistically-significant. Overall DECENT has the highest power to detect differentially-expressed genes with SCDE having the second highest and Monocle the lowest power (Figure 3(c)). When we examine the low and high abundance genes separately, where high abundance here is defined as having median cpm ≥ 1, it becomes quite clear that DECENT has higher power for detecting low abundance DE genes (Figure 3(d)). This suggests that while there is an overall benefit from imputing, that benefit is more pronounced for low abundance genes.

The violin plots of observed and imputed count for two DE genes that were found among the top 25 genes with the smallest p-value by DECENT, but not so by the other methods, clearly demonstrates why imputation increases the chance of discovering these genes (Supp. Figure 3).

Soumillon Dataset

We downloaded the raw count data from the NCBI website and removed cells with log library size smaller than 3 MAD away from the median log library size, and number of expressed genes smaller than 3 MAD away from the median of number of expressed genes. Genes with average count less than 0.05 were also removed, resulting in 1949 cells and 8586 genes included for

analysis. To estimate the size factors for this dataset, we used both the maximum-likelihood estimate (see Appendix) and the TMM normalization method (Robinson and Oshlack, 2010). The TMM method was developed to normalize bulk RNA-seq data that have very different library compositions across experimental conditions. We compare the effectiveness of the two estimates of size factors by examining the distribution of log fold-change for single-cell housekeeping genes (scHK, list obtained from Pengyi Yang, University of Sydney). These single-cell housekeeping genes have been identified as genes that are expressed relatively stably across various cell-types and experimental conditions, and we thus expect the better size factor estimates will result in a distribution of log fold-changes for these genes centred around zero. Figure 4(d) shows that size factors estimated using the TMM method results in the log fold-change distribution for the housekeeping genes that was centred more around zero.

Using the TMM normalization method to estimate size factors, DECENT outperforms other methods for identifying the reference DE genes (Figure 4(a)). The reference DE genes here were defined as the 1000 genes with the largest magnitude of log fold-change in the matching bulk RNA-seq experiment.

Discussion

We have developed a differential expression (DE) method that adjusts for the imperfect capture efficiency by estimating the number of molecules pre-dropout in scRNA-seq datasets. Using simulated and real datasets, we saw the performance of our method is better compared to previously published methods. DECENT does not require spike-ins, but when available, they can be used to improve its performance.

As with all other data, scRNA-seq data can be affected by batch effects (Hicks *et al.*, 2015 ; Tung *et al.*, 2017). Since DECENT estimates and performs statistical testing based on the inferred number of molecules pre-dropout, another advantage is that it minimizes differences between cells of the same type analyzed in different batches *if* the differences between batches are primarily driven by differences in capture rates. Since DECENT uses the generalized linear model (GLM) framework, it is possible to extend the DECENT model to adjust for unwanted factors in the data such as cell-cycle variation (Scialdone *et al.*, 2015 ; Buettner *et al.*, 2015). One way to adjust for cell-cycle differences would be including latent variables estimated by leading eigenvectors of a between-cell covariance matrix that has been calculated using only data from cell-cycle annotated genes. We have not thoroughly tested this idea but plan to do so in the near future.

One current limitation of DECENT is it requires the raw data to be in UMI-count format. Islam *et al.* (2013) have shown that using UMI in scRNA-seq experiments reduces technical noise as it minimizes the bias due to the PCR amplification step. Many current scRNA-seq technologies, especially the increasingly popular droplet-based technologies, allow the use of UMI tags. However, we do wish to cater for technologies that do not use UMI, and we are currently exploring the possibility of extending DECENT to read-count data.

Availability

DECENT is implemented as an R package and is freely available from GitHub. The package can be installed using the following command:

```
devtools::install_github("cz-ye/DECENT")
```

Appendix: EM algorithm

Let E_{ij} be the binary indicator that gene i is truly expressed in cell j, i.e. that $y_{ij}>0$. Our complete data is (y_{ij}, E_{ij}). Note that we can only be certain that $E_{ij} = 1$ when $z_{ij}>0$ because due to imperfect capture $y_{ij} \geq z_{ij}$. However, we cannot be sure of the value of E_{ij} when $z_{ij} = 0$. In neither case do we directly observe the true number y_{ij} of molecules. Our interest is in estimating the parameters of the ZINB distribution for the true number of molecules, but because we have incomplete data we will estimate the ZINB parameters using the EM algorithm (McLachlan and Krishnan, 1997).

E-step

For each gene, we start with an initial estimate of the ZINB parameters $\theta_i^0 = (\alpha_{0i}^0, \beta_{0i}^0, \beta_{1i}^0, \psi_i^0)$ and size-factor parameters s_j^0 and evaluate the expected value of the complete data log-likelihood given this initial estimate. Note that under the ZINB model, the complete data likelihood is given by,

$$L(\theta_i, s_j) = \prod_j \left[\pi_{0i} + (1-\pi_{0i})\left(\frac{\psi_i}{\psi_i + s_j\mu_{ij}}\right)^{\psi_i}\right]^{1-E_{ij}} \left[(1-\pi_{0i})\frac{\Gamma(\psi_i + y_{ij})}{y_{ij}!\Gamma\psi_i}\left(\frac{\psi_i}{\psi_i + s_j\mu_{ij}}\right)^{\psi_i}\left(\frac{s_j\mu_{ij}}{\psi_i + s_j\mu_{ij}}\right)^{y_{ij}}\right]^{E_{ij}}$$

which means that the complete data log-likelihood can be written as

$$\begin{aligned}
\ell(\theta_i, s_j) &= \sum_j (1-E_{ij}) \log\left[\pi_{0i} + (1-\pi_{0i})\left(\frac{\psi_i}{\psi_i + s_j\mu_{ij}}\right)^{\psi_i}\right] + E_{ij}\log\left[(1-\pi_{0i})\left(\frac{\psi_i}{\psi_i + s_j\mu_{ij}}\right)^{\psi_i}\right] \\
&+ E_{ij}\log\left[\frac{\Gamma(\psi_i + y_{ij})}{y_{ij}!\Gamma\psi_i}\right] + y_{ij}E_{ij}\log\left[\left(\frac{s_j\mu_{ij}}{\psi_i + s_j\mu_{ij}}\right)\right] \\
&= \sum_j (1-E_{ij}) \log\left[\pi_{0i} + (1-\pi_{0i})\left(\frac{\psi_i}{\psi_i + s_j\mu_{ij}}\right)^{\psi_i}\right] + E_{ij}\log\left[(1-\pi_{0i})\left(\frac{\psi_i}{\psi_i + s_j\mu_{ij}}\right)^{\psi_i}\right] \\
&+ E_{ij}\log\left[\frac{1}{y_{ij}}\frac{\Gamma(\psi_i + y_{ij})}{\Gamma(y_{ij})\Gamma\psi_i}\right] + y_{ij}E_{ij}\log\left[\left(\frac{s_j\mu_{ij}}{\psi_i + s_j\mu_{ij}}\right)\right] \qquad (1)
\end{aligned}$$

From the last equation, it can be seen that to evaluate the expected value of the complete data log-likelihood given the observed data at the current estimate, $E(\ell(\theta_i, s_j) \mid z_{ij}; \theta_i^0, s_j^0)$, we need to evaluate $E(E_{ij} \mid z_{ij} = 0; \theta_i^0, s_j^0) = P(E_{ij} = 1 \mid z_{ij} = 0; \theta_i^0, s_j^0)$ and $E(y_{ij}E_{ij} \mid z_{ij}; \theta_i^0, s_j^0)$

It can be shown that these expected values are given by,

$$
\begin{aligned}
P(E_{ij} = 1 \mid z_{ij} = 0; \theta_i^0, s_j^0) &= \frac{P(E_{ij} = 1, z_{ij} = 0; \theta_i^0, s_j^0)}{P(z_{ij} = 0; \theta_i^0, s_j^0)} \\
&= \frac{\sum_{k;k\geq 1} P(z_{ij} = 0 \mid y_{ij} = k; \theta_i^0, s_j^0) P(y_{ij} = k; \theta_i^0, s_j^0)}{P(z_{ij} = 0; \theta_i^0, s_j^0)} \\
&= \frac{\sum_{k;k\geq 1} P(z_{ij} = 0 \mid y_{ij} = k; \theta_i^0, s_j^0) P(y_{ij} = k; \theta_i^0, s_j^0)}{P(z_{ij} = 0, E_{ij} = 1; \theta_i^0) + P(z_{ij} = 0 \mid y_{ij} = 0; \theta_i^0, s_j^0) P(y_{ij} = 0; \theta_i^0, s_j^0)} \\
&= \frac{\sum_{k;k\geq 1} P(z_{ij} = 0 \mid y_{ij} = k; \theta_i^0, s_j^0) P(y_{ij} = k; \theta_i^0, s_j^0)}{P(z_{ij} = 0, E_{ij} = 1; \theta_i^0) + P(y_{ij} = 0; \theta_i^0, s_j^0)}
\end{aligned}
$$

where

$$
P(z_{ij} = 0, E_{ij} = 1; \theta_i^0, s_j^0) = \sum_{k;k\geq 1} P(z_{ij} = 0 \mid y_{ij} = k; \theta_i^0, s_j^0) P(y_{ij} = k; \theta_i^0, s_j^0)
$$

and

$$
E(y_{ij} \mid z_{ij} = l; \theta_i^0, s_j^0) = \sum_{k;k\geq l} k P(y_{ij} = k; \theta_i^0, s_j^0) \frac{P(z_{ij} = l \mid y_{ij} = k; \theta_i^0, s_j^0)}{\sum_{m;m\geq l} P(z_{ij} = l \mid y_{ij} = m; \theta_i^0, s_j^0) P(y_{ij} = m; \theta_i^0, s_j^0)}
$$

where the summation needs to be taken only for $k \geq l$ to reflect the fact that the number of captured molecules cannot be more than the total number of molecules available.

The last expected value, $E(y_{ij}E_{ij} \mid z_{ij}; \theta_i^0, s_j^0)$, can be evaluated using the rule of conditional expectation $E(X) = E[E(X|Y)]$,

$$
\begin{aligned}
E(y_{ij}E_{ij} \mid z_{ij}; \theta_i^0, s_j^0) &= E(y_{ij}E_{ij} \mid z_{ij}, ; E_{ij} = 1; \theta_i^0, s_j^0) P(E_{ij} = 1 \mid z_{ij}; \theta_i^0, s_j^0) \\
&+ E(y_{ij}E_{ij} \mid z_{ij}, E_{ij} = 0; \theta_i^0, s_j^0) P(E_{ij} = 0 \mid z_{ij}; \theta_i^0, s_j^0) \\
&= E(y_{ij} \mid z_{ij}; \theta_i^0, s_j^0) P(E_{ij} = 1 \mid z_{ij}; \theta_i^0, s_j^0)
\end{aligned}
$$

where, by definition of E_{ij} and the fact that if we observe at least one molecule then the gene must have been expressed, $P(E_{ij} = 1 \mid z_{ij}; \theta_i^0, s_j^0) = 1$ if $z_{ij} \geq 1$ and $P(E_{ij} = 1 \mid z_{ij} = 0; \theta_i^0, s_j^0)$ is given above.

M-step

At the M-step, we maximize the expected value of the complete data log-likelihood given observed data, $E(\ell(\theta_i) \mid z_{ij})$ to update $\theta_i^0 \to \theta_i^1$ and $s_j^0 \to s_j^1$. In practice, this is done by first replacing E_{ij} and $y_{ij}E_{ij}$ in equation (1) with their respective values from the E-step and then maximization step is carried out in the following way:

- Updating $\theta_i^0 \to \theta_i^1$: For each gene, fixing the cell-specific size factors $s_j = s_j^0, j = 1, 2, \ldots N$, we maximize $E(\ell(\theta_i, s_j) \mid z_{ij}; s_j = s_j^0)$ w.r.t $\gamma_{0i}, \beta_{0i}, \beta_{1i}, \psi_i$ using a Quasi-Newton algorithm

- Updating $s_j^0 \to s_j^1$: Differentiating $E(\ell(\theta_i, s_j) \mid z_{ij}; \theta_i = \theta_i^1)$ w.r.t s_j and setting it to zero, the updated estimate of s_j is given by

$$
s_j^1 \propto \frac{\sum_i w_{ij} E(y_{ij}E_{ij} \mid z_{ij}; \theta_i^0, s_j^0)}{\sum_i w_{ij}}
$$

where the weight w_{ij} is given by

$$w_{ij} = \frac{P(y_{ij} = 0; \theta_i^1, s_j^0) - \pi_{0i}^0 P(E_{ij} = 1 \mid z_{ij} = 0; \theta_i^1, s_j^0)}{P(y_{ij} = 0; \theta_i^1, s_j^0)}(1 - \pi_{0i}^1)$$

In the case where $\pi_{0i}^1 = 0, \forall i$, the weights are all 1 and the estimated size-factor for cell j is simply proportional to the library size of the imputed data $E(y_{ij}E_{ij} \mid z_{ij}; \theta_i^1, s_j^0)$ for that cell.

Assessing Convergence

The EM algorithm iterates between the E-step and M-step until convergence. Let $\Theta = \{\theta_1, \theta_2, \ldots \theta_G\}$ be the vector of parameters for G genes, we use the log-likelihood of the incomplete data $\ell_I(\Theta)$ to assess convergence.

$$\ell_I(\Theta) \quad = \quad \sum_{i=1}^{G} \ell_I(\theta_i) = \sum_{i=1}^{G} \sum_{j=1}^{N} \log P(z_{ij}; \theta_i, s_j)$$

where

$$P(z_{ij}; \theta_i, s_j) = \sum_k P(z_{ij} \mid y_{ij} = k; \theta_i, s_j) P(y_{ij} = k; \theta_i, s_j)$$

DE analysis

Differential expression across cell-types for the ith gene is assessed by testing the hypotheses:

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

using the likelihood ratio test (LRT) statistic,

$$-2\{\ell_I(\theta_i = \hat{\theta}_i^{H_0}) - \ell_I(\theta_i = \hat{\theta}_i)\} \sim \chi_1^2$$

where $\theta_i^{H_0}$ is the MLE of θ_i under the restriction that $\beta_{1i} = 0$, and $\hat{\theta}_i$ is the MLE under the unrestricted model.

References

Bacher R, Chu LF, Leng N et al. SCnorm: robust normalization of single-cell RNA-seq data. *Nat Methods.* 2017 Jun ;**14(6)**:584-586. doi: 10.1038/nmeth.4263.

Brennecke P, Anders S, Kim JK et al. Accounting for technical noise in single-cell RNA-seq experiments. *Nat Methods.* 2013 Nov; **10(11)**:1093-5. doi: 10.1038/nmeth.2645.

Buettner F, Natarajan KN, Casale FP et al. Computational analysis of cell-to-cell heterogeneity in single-cell RNA-sequencing data reveals hidden subpopulations of cells. *Nat Biotechnol.* 2015 Feb; **33(2)**:155-60. doi: 10.1038/nbt.3102.

Finak G, McDavid A, Yajima M. MAST: a flexible statistical framework for assessing transcriptional changes and characterizing heterogeneity in single-cell RNA sequencing data. *Genome Biol.* 2015 Dec 10;**16**:278. doi: 10.1186/s13059-015-0844-5.

Hicks, S. C., Teng, M. & Irizarry, R. A. et al. On the widespread and critical impact of systematic bias and batch effects in single-cell RNA-seq data. bioRxiv 2015.doi: 10.1101/025528

Huang M, Wang J, Torre E et al. Gene Expression Recovery For Single Cell RNA Sequencing. bioRxiv 2017. doi: https://doi.org/10.1101/138677

Jia C, Hu Y, Kelly D et al. Accounting for technical noise in differential expression analysis of single-cell RNA sequencing data. *Nucleic Acids Res.* 2017 Nov;**45(19)**:10978-10988. doi: 10.1093/nar/gkx754.

Jiang L, Schlesinger F, Davis CA, Zhang Y, Li R, Salit M, Gingeras TR, Oliver B. Synthetic spike-in standards for RNA-seq experiments. *Genome Res* 2011; **21(9)**:1543-51. doi: 10.1101/gr.121095.111.

Kharchenko PV, Silberstein L, Scadden DT. Bayesian approach to single-cell differential expression analysis. *Nat Methods.* 2014 Jul; **11(7)**:740-2. doi: 10.1038/nmeth.2967.

Lun AT, Bach K, Marioni JC. Pooling across cells to normalize single-cell RNA sequencing data with many zero counts. *Genome Biol.* 2016 Apr ;**17**:75. doi: 10.1186/s13059-016-0947-7.

McLachlan, G.J. and Krishnan, T. (1997). *The EM Algorithm and Extensions.* New York: Wiley. xvii + 274 pp.

Ridout M, Hinde J, Demétrio CG. A score test for testing a zero-inflated Poisson regression model against zero-inflated negative binomial alternatives. *Biometrics.* 2001 Mar;**57(1)**:219-23.

Risso D, Perraudeau F, Gribkova S et al. ZINB-WaVE: A general and flexible method for signal extraction from single-cell RNA-seq data. bioRxiv 2017. doi: https://doi.org/10.1101/125112

Robinson MD, McCarthy DJ, Smyth GK. edgeR: a Bioconductor package for differential expression analysis of digital gene expression data *Bioinformatics.* 2010 Jan; **26(1)**: 139140. doi: 10.1093/bioinformatics/btp616

Robinson MD, Oshlack A. A scaling normalization method for differential expression analysis of RNA-seq data. *Genome Biol.* 2010; **11(3)**:25.

Scialdone A, Natarajan KN, Saraiva LR et al. Computational assignment of cell-cycle stage from single-cell transcriptome data. *Methods.* 2015 Sep 1;**85**:54-61. doi: 10.1016/j.ymeth.2015.06.021.

Stegle O, Teichmann SA, Marioni JC. Computational and analytical challenges in single-cell transcriptomics. *Nat Rev Genet.* 2015 Mar;**16(3)**:133-45. doi: 10.1038/nrg3833.

Trapnell C, Cacchiarelli D, Grimsby J et al. The dynamics and regulators of cell fate decisions are revealed by pseudotemporal ordering of single cells. *Nat Biotechnol.* 2014 Apr; **32(4)**:381-386. doi: 10.1038/nbt.2859.

Tung P-Y, Blischak JD, Hsiao JC. Batch effects and the effective design of single-cell gene expression studies. *Scientific Reports* 2017 **7**: 39921. doi:10.1038/srep39921

Van den Berge K, Soneson C, Love MI et al. zingeR: unlocking RNA-seq tools for zero-inflation and single cell applications. bioRxiv 2017. doi: https://doi.org/10.1101/157982

Vu TN, Wills QF, Kalari KR et al. Beta-Poisson model for single-cell RNA-seq data analyses. *Bioinformatics.* 2016 Jul: **32(14)**:2128-35. doi: 10.1093/bioinformatics/btw202.

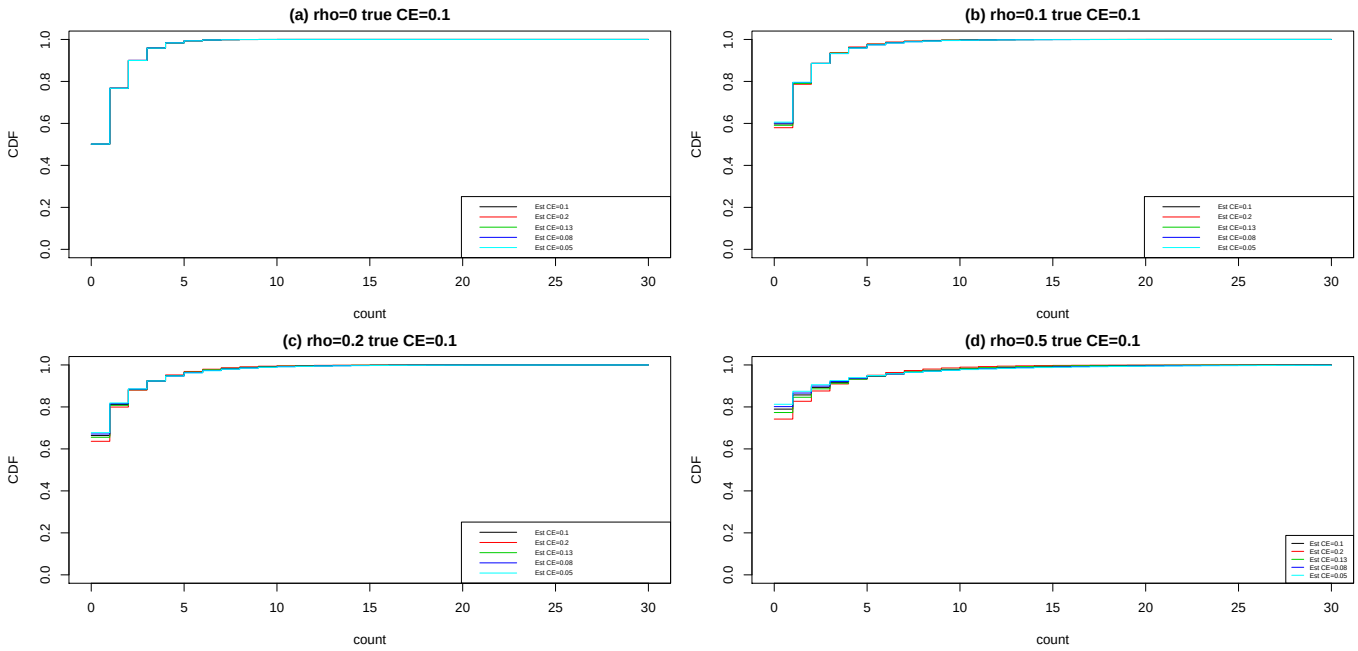


Figure 1: Cumulative distribution function of observed counts under ZINB ($\pi_0 = 0.1, \mu = 10, \psi = 2$) model with true (black line) and various levels of misspecified capture efficiency.

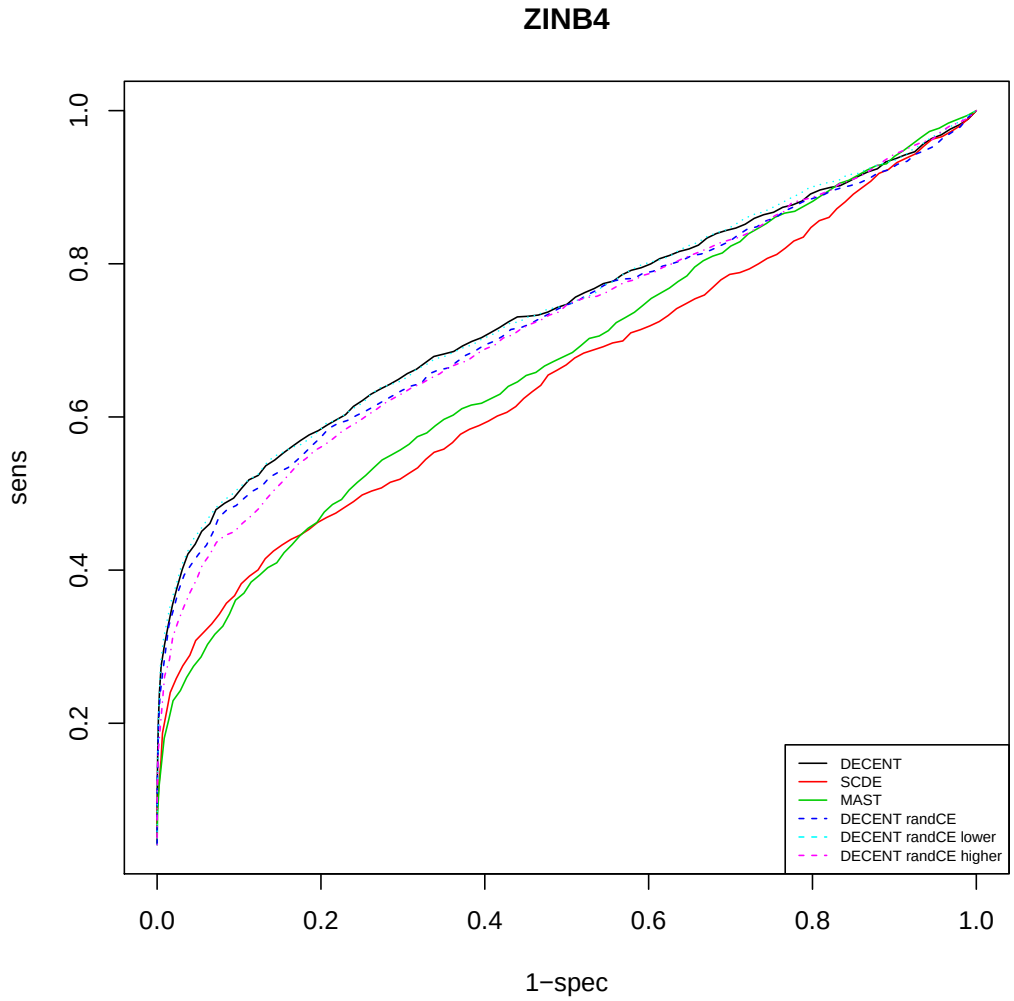


Figure 2: Receiver Operating Characteristics (ROC) curves of DECENT, MAST and SCDE for our simulated datasets.

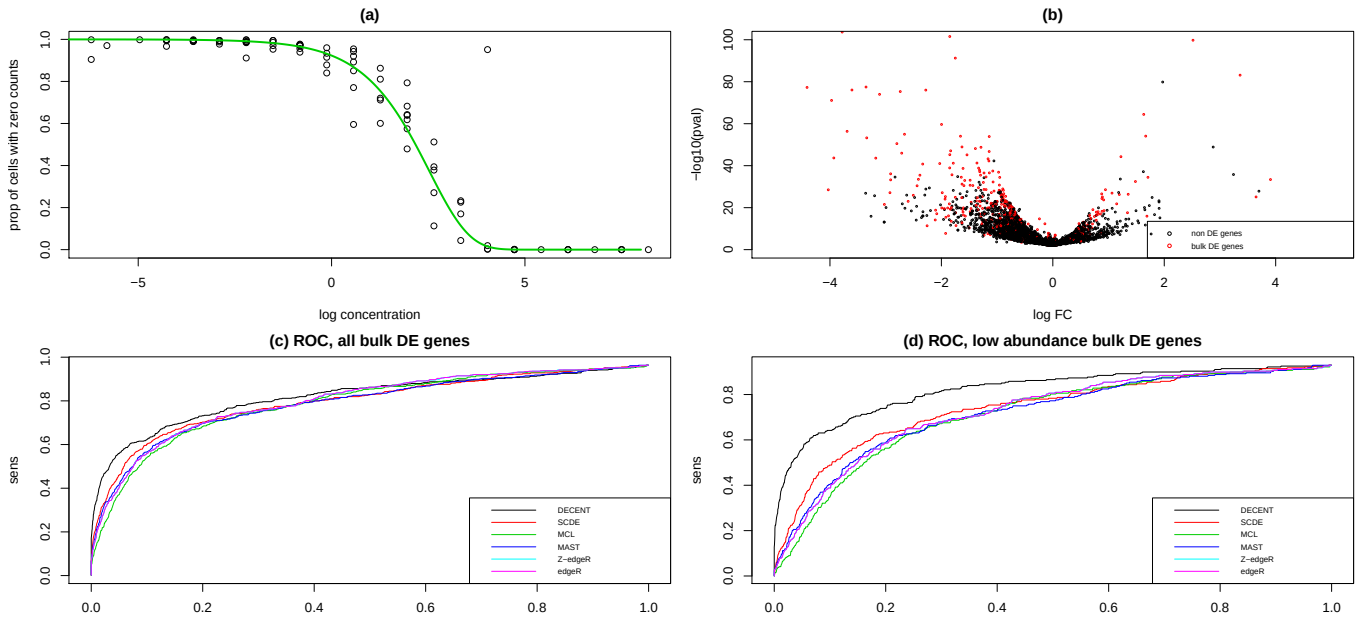


Figure 3: Results for Tung's dataset: (a) Observed vs predicted dropout rates for the ERCC spike-ins transcripts, (b) DECENT's Volcano Plot, (c) ROC for all bulk DE genes, (d) ROC for low abundance bulk DE genes

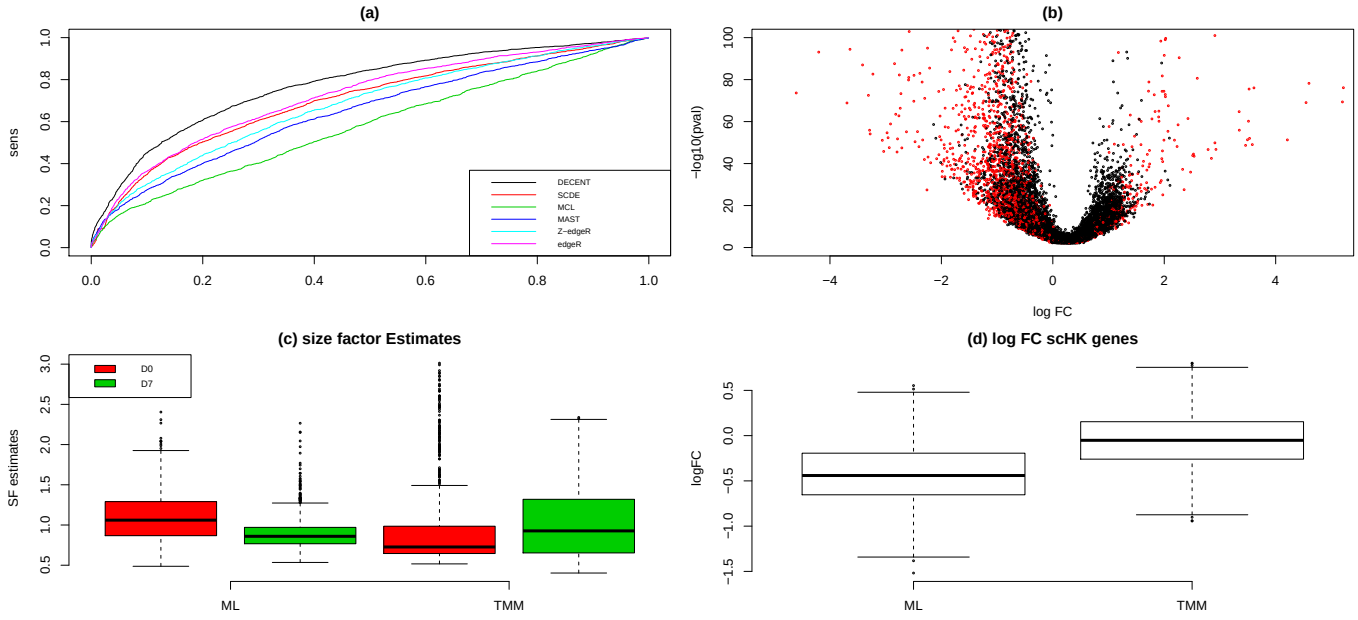
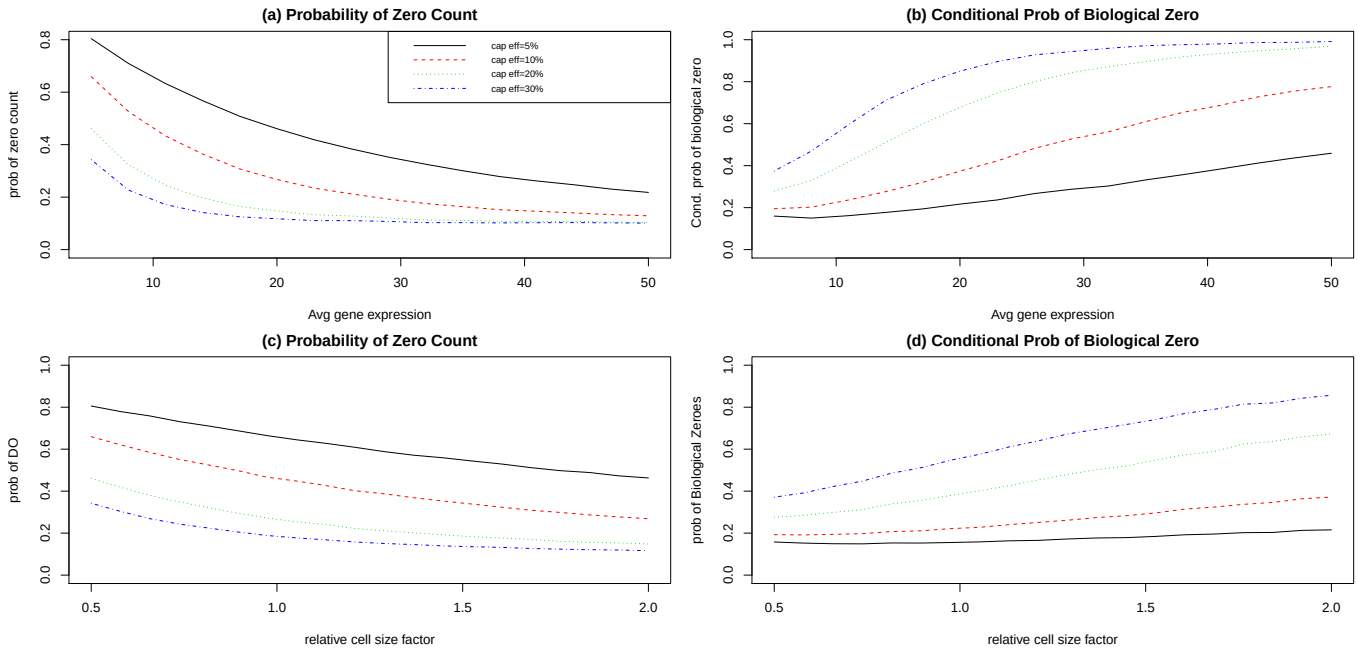
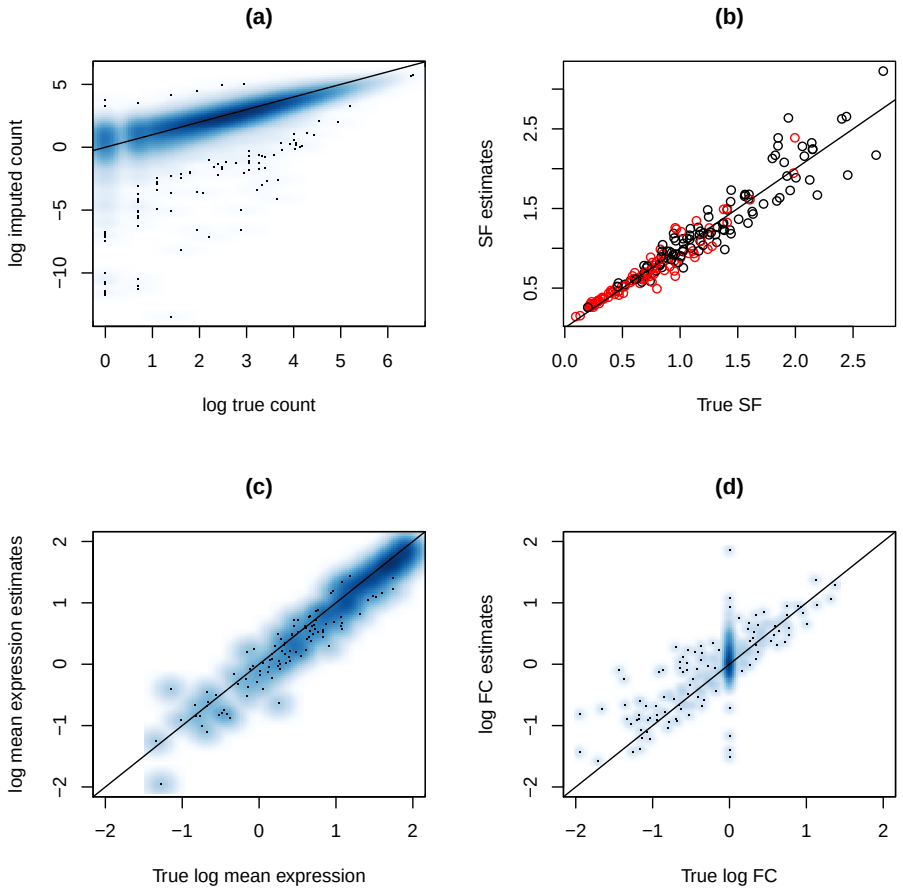


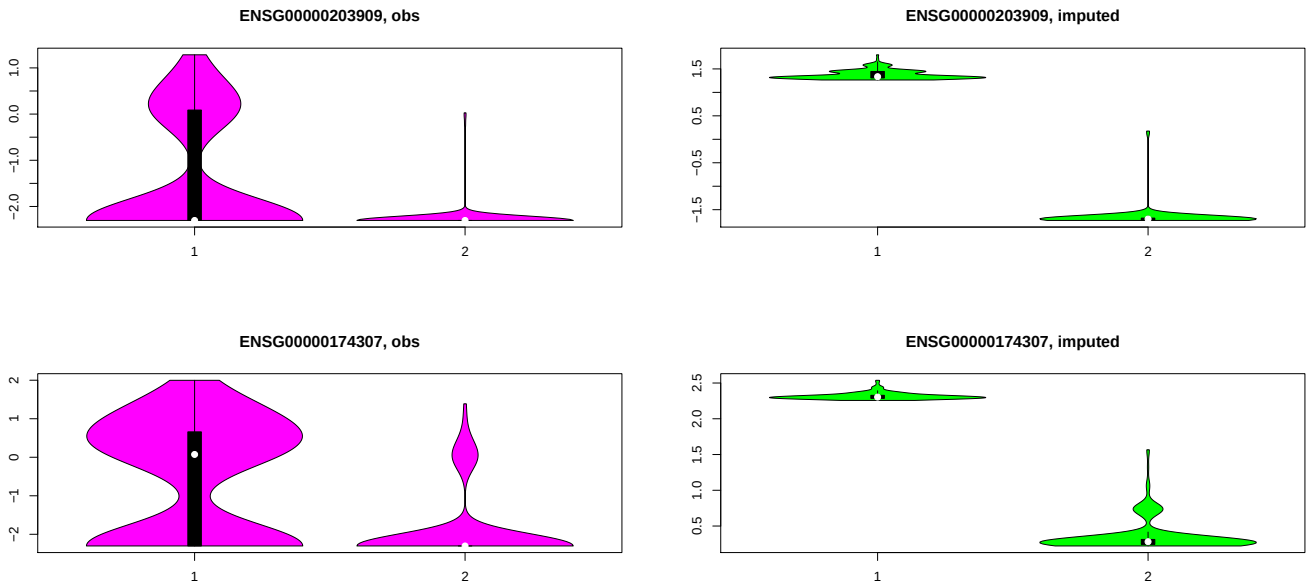
Figure 4: Results for Soumillon's dataset: (a) ROC, (b) DECENT's volcano plot, (c) Comparison of size factor estimates using ML and TMM methods (d) Distributions of log FC estimates for single-cell housekeeping (scHK) genes



Supp. Figure 1



Supp. Figure 2



Supp. Figure 3