Supplementary Materials

of

SwarnSeq: An Improved Statistical Approach for Differential Expression Analysis of Single-Cell RNA-Seq Data

Samarendra Das[1 – 3], Shesh N. Rai[2 – 7, *]

[1]Division of Statistical Genetics, ICAR-Indian Agricultural Statistics Research Institute, PUSA, New Delhi 110012, India

[2]Biostatistics and Bioinformatics Facility, JG Brown Cancer Center, University of Louisville, Louisville, KY 40202, USA

[3]School of Interdisciplinary and Graduate Studies, University of Louisville, Louisville, KY 40292, USA

[4]Hepatobiology and Toxicology Center, University of Louisville, Louisville, KY 40202, USA

[5]Department of Bioinformatics and Biostatistics, University of Louisville, Louisville, KY 40202, USA

[6]Biostatistics and Informatics Facility, Center for Integrative Environmental Research Sciences, University of Louisville, Louisville, KY, 40202, USA

[7]Christina Lee Brown Envirome Institute, University of Louisville, Louisville, KY 40202, USA

Authors' email addresses:

SD: samarendra.das@louisville.edu

SNR: shesh.rai@louisville.edu

***To whom correspondence should be addressed-** email: shesh.rai@louisville.edu; Telephone: +1 502-426-001

Contents

1. Supplementary Document S1: Count data models

1.1 *Negative Binomial Distribution*

Most of the popular Differential Expression (DE) analysis tools, *e.g.* DESeq [1], DESeq2 [2], edgeR [3], *etc*., for bulk RNA-sequencing (RNA-seq) study assume the RNA-seq read counts to follow a Negative Binomial (NB) distribution, and subsequently, DE analysis is performed under a Generalized Linear Model (GLM) framework.

Let, Z_{ij}: random variable (*rv*) representing the RNA-seq read counts of i^{th} (i = 1, 2, …, N) gene of j^{th} (j = 1, 2, …, M) cell; μ: mean of i^{th} gene of j^{th} cell; θ: size (=1/dispersion) parameter of i^{th} gene of j^{th} cell.

Further, the Probability Mass Function (PMF) of the NB distribution is expressed as:

$$f_{NB}(z) = P[Z_{ij} = z] = \frac{G(z+\theta)}{G(z+1)G(\theta)}\left(\frac{\theta}{\theta+\mu}\right)^{\theta}\left(\frac{\mu}{\theta+\mu}\right)^{z} \quad \forall\, z = 0, 1, 2, \ldots \tag{1}$$

where, $\mu \geq 0$; $\theta > 0$ are the parameters of NB distribution, G(.): Gamma function. The NB distribution boils down to Poisson distribution, when $1/\theta$ =0. The NB distribution can be parameterized as $Z_{ij} \sim NB(\mu, \theta)$, and its mean and variance can be expressed in Eq. 2.

$$E(Y_{ij}) = \mu \text{ and } Var(Y_{ij}) = \mu + \frac{\mu^2}{\theta} \tag{2}$$

1.2 **Zero Inflated Negative Binomial Distribution**

The proportions of zeros in single cell RNA-sequencing (scRNA-seq) data are higher as compared to bulk RNA-seq data due to low efficiency of mRNA capture, lower abundance of transcriptomics in single cell, amplification bias, *etc*. Therefore, the application of NB based bulk RNA-seq DE tools leads to several technical problems including statistical power to detect DE genes in scRNA-seq studies [4,5]. So, specialized scRNA-seq DE tools, *e.g.* ZINB-Wave [6], DEsingle [7], DECENT [8], *etc*. are developed based on assuming the observed scRNA-seq read counts follow Zero Inflated Negative Binomial (ZINB) Distribution.

Let, Z_{ij}: random variable (*rv*) representing the read (UMI) counts in scRNA-seq data of i^{th} (i = 1, 2, …, N) gene of j^{th} (j = 1, 2, …, M) cell; μ: mean of i^{th} gene of j^{th} cell; θ: size (=1/dispersion) parameter of i^{th} gene of j^{th} cell; π: zero inflation (*i.e.* the probability for a count to be an excess zero in a cell) parameter for i^{th} gene of j^{th} cell.

The PMF of the ZINB distribution can be expressed as:

$$f_{ZINB}(z) = P\left[Z_{ijk} = z\right] = \pi I(z = 0) + (1 - \pi) f_{NB}(z) \;\forall\; z = 0, 1, 2, \ldots \tag{3}$$

$$= \begin{cases} \pi + (1 - \pi)\left(\dfrac{\theta}{\theta + \mu}\right)^{\theta} & when\ z = 0 \\ (1 - \pi)\dfrac{G(z + \theta)}{G(z + 1)G(\theta)}\left(\dfrac{\theta}{\theta + \mu}\right)^{\theta}\left(\dfrac{\mu}{\theta + \mu}\right)^{z}; & z > 0 \end{cases}$$

where, π is the proportion of constant zeros for i^{th} gene in the group of cells, $G(.)$ is the Gamma function, $I(z = 0)$ is an indicator function which takes value 1 for $z = 0$ and 0 for $z \neq 0$ and $f_{NB}(.)$ is the PMF of NB distribution.

The mean and variance of ZINB distribution can be derived as:

$$E(Z) = \sum_{z=0}^{\infty} z\, f_{zinb}(z)$$

$$= 0 f_{zinb}(0) + \sum_{z=1}^{\infty} z\, f_{zinb}(z)$$

$$= (1 - \pi) \textstyle\sum_{z=0}^{\infty} z\, f_{NB}(z) = (1 - \pi)\mu \tag{4}$$

$$Var(Z) = E(Z^2) - \{E(Z)\}^2$$

$$E(Z^2) = \sum_{z=0}^{\infty} z^2 f_{zinb}(z) = (1 - \pi)(\mu + \frac{\mu^2}{\theta} + \mu^2)$$

Now, $Var(Z) = (1 - \pi)\left(\mu + \frac{\mu^2}{\theta} + \mu^2\right) - (1 - \pi)^2\mu^2$

$$= (1 - \pi)\mu\left(1 + \pi\mu + \frac{\mu}{\theta}\right) \tag{5}$$

1.3 **Poisson Distribution**

Poisson distributions are also extensively used for analysis of count data obtained from bulk RNA-seq or scRNA-seq experiments. The PMF of Poisson distribution can be expressed as:

$$f_{PD}(z) = P\left[Z_{ij} = z\right] = \frac{e^{-\mu}\mu^{z}}{G(z+1)} \quad \forall\, z = 0, 1, 2, \ldots \tag{6}$$

1.4 **Zero Inflated Poisson Distribution**

Poisson model has very strict assumptions. One that is often violated is that the mean equals the variance. When the variance is too large because there are many 0s as well as a few very high values [9]. In this case, a better solution is often the Zero-Inflated Poisson (ZIP) model.

The PMF of ZIP distribution can be expressed as:

$$f_{ZIPD}(z) = P\left[Z_{ij} = z\right] = \pi I(z = 0) + (1 - \pi) f_{PD}(z) \qquad \forall\, z = 0, 1, 2, \ldots \tag{7}$$

$$= \begin{cases} \pi + (1 - \pi)e^{\mu} & when\ z = 0 \\ (1 - \pi)\frac{e^{-\mu}\mu^{z}}{G(z+1)}; & z > 0 \end{cases} \tag{8}$$

The mean and variance of ZIP distribution can be obtained as:

$$E(Z) = (1 - \pi)\mu \tag{9}$$

$$Var(Z) = (1 - \pi)\mu(1 + \pi\mu) \tag{10}$$

1.5 **Hermite Distribution**

Hermite Distribution (HD) can be used to model the counts data [10]. Further, the PMF of HD is given as:

$$f_{HD}\left(Z_{ij} = z \middle| \alpha, \beta\right) = e^{-(\alpha+\beta)} \sum_{k=0}^{\left[\frac{z}{2}\right]} \frac{\alpha^{\,y-2k}\beta^{k}}{G(z-2k+1)G(k+1)} \qquad \forall\, z = 0, 1, 2, \ldots \tag{11}$$

Further, the mean (μ), variance (σ^2) and dispersion index (φ) (*i.e.*, ratio between mean and variance) of rv Z_{ij}~HD can be given as:

$$E(Z_{ij}) = \mu = f(\alpha, \beta) = (\alpha_{ig} + 2\beta_{ig})$$

(12)

$$Var(Z_{ij}) = (\alpha_{ig} + 4\beta_{ig})$$

(13)

$$\varphi = g(\alpha, \beta) = 1 + 2\beta_{ig}/(\alpha_{ig} + 2\beta_{ig})$$

(14)

2. Supplementary Document S2: Models for scRNA-seq read counts

2.1 Binomial Transcriptional Capturing Model

Dropout events in single cell RNA-sequencing (scRNA-seq) data are mostly due to the low capture efficiency of mRNA molecules presently in single cells [11–14]. It may go to the extent that the mRNA products inherently present in single cells sometimes goes totally uncaptured by the single cell protocols due to their low capture efficiency [15–17]. Here, we refer capture efficiency procedure to both the mRNA molecules capturing in reverse transcription step and the capturing of cDNA molecules in the amplification step. We tried to integrate the transcriptional capturing procedure through a simple Binomial model with the ZINB model for observed scRNA-seq read counts for estimation of different parameters for each gene. The Binomial Transcriptional Capturing Model can be described as:

Let, Z_{ij}: random variable (*rv*) representing the true (unknown) read (UMI) counts of i^{th} (i = 1, 2, …, N) gene of j^{th} (j = 1, 2, …, M) cell; Y_{ij}: *rv* for observed (known) read (UMI) counts of i^{th} gene of j^{th} cell; ; ρ_{ijk}: *rv* representing transcriptional capture of i^{th} gene of j^{th} cell with efficiency parameter p_{ij} ($0 \leq p_{ij} \leq 1$). The transcriptional capturing model can be expressed as:

$$P\left[Y_{ij} = y \middle| Z_{ij} = z, p_{ij}\right] = \binom{z}{y} {p_{ij}}^{y} (1 - p_{ij})^{z-y}$$

(15)

2.2 Distribution for the observed scRNA-seq read (UMI) counts

We assumed that the true UMI counts from the scRNA-seq study derived from a ZINB population and the transcriptional capture process is a Binomial process. Then, the distribution of observed read (UMI) counts in the scRNA-seq study can be obtained as follows:

From the above, we have $Z_{ijk} \sim ZINB(\pi_{ijk}, \mu_{ijk}, \theta_{ijk})$ and $\rho_{ijk} = (Y_{ijk} | Z_{ijk} = z) \sim B(z, p_{ijk})$

Now,

$$P\left[Z_{ijk} = z\right] = \begin{cases} \pi_{ijk} + (1 - \pi_{ijk}) \left(\frac{\theta_{ijk}}{\theta_{ijk} + \mu_{ijk}}\right)^{\theta_{ijk}} & when\ z = 0 \\ (1 - \pi_{ijk}) \frac{G(z + \theta_{ijk})}{G(z+1)G(\theta_{ijk})} \left(\frac{\theta_{ijk}}{\theta_{ijk} + \mu_{ijk}}\right)^{\theta_{ijk}} \left(\frac{\mu_{ijk}}{\theta_{ijk} + \mu_{ijk}}\right)^{z}; & z > 0 \end{cases}$$

(16)

$$P\left[Y_{ijk} = y \middle| Z_{ijk} = z\right] = \binom{z}{y} {p_{ijk}}^{y} (1 - p_{ijk})^{z-y}$$

(17)

The joint probability distribution of Y_{ijk} and Z_{ijk} can be written as:

$$P\left[Y_{ijk} = y, Z_{ijk} = z \middle| \pi_{ijk}, \mu_{ijk}, \theta_{ijk}, p_{ijk}\right] = P\left[Y_{ijk} = y \middle| Z_{ijk} = z, p_{ijk}\right] P\left[Z_{ijk} = z \middle| \pi_{ijk}, \mu_{ijk}, \theta_{ijk}\right]$$

(18)

Now, the marginal probability distribution of Y_{ijk} can be given as:

$$P\left[Y_{ijk} = y \middle| \pi_{ijk}, \mu_{ijk}, \theta_{ijk}, p_{ijk}\right] = \sum_z P\left[Y_{ijk} = y \middle| Z_{ijk} = z, p_{ijk}\right] P\left[Z_{ijk} = z \middle| \pi_{ijk}, \mu_{ijk}, \theta_{ijk}\right]$$

(19)

<u>Case-1</u>: For zero count $(Y_{ijk} = 0)$ **case**

$$\left[Y_{ijk}=0\middle|\pi_{ijk},\mu_{ijk},\theta_{ijk},p_{ijk}\right]=\sum_{z}P\left[Y_{ijk}=0\middle|Z_{ijk}=z,p_{ijk}\right]P\left[\,Z_{ijk}=z\middle|\pi_{ijk},\mu_{ijk},\theta_{ijk}\right]$$

$$=P\left[Y_{ijk}=0\middle|Z_{ijk}\right.$$

$$\left.=z,p_{ijk}\right]P\left[\,Z_{ijk}=0\middle|\pi_{ijk},\mu_{ijk},\theta_{ijk}\right]$$

$$+\sum_{z=1}^{\infty}P\left[Y_{ijk}=0\middle|Z_{ijk}=z,p_{ijk}\right]P\left[\,Z_{ijk}=z\middle|\pi_{ijk},\mu_{ijk},\theta_{ijk}\right]$$

$$=\pi_{ijk}+\left(1-\pi_{ijk}\right)\left(\frac{\theta_{ijk}}{\theta_{ijk}+\mu_{ijk}}\right)^{\theta_{ijk}}$$

$$+\sum_{z=1}^{\infty}\left\{(1-p_{ijk})^{z}\left(1-\pi_{ijk}\right)\frac{G\left(z+\ \theta_{ijk}\right)}{G(z+1)G\left(\theta_{ijk}\right)}\left(\frac{\theta_{ijk}}{\theta_{ijk}+\mu_{ijk}}\right)^{\theta_{ijk}}\left(\frac{\mu_{ijk}}{\theta_{ijk}+\mu_{ijk}}\right)^{z}\right\}$$

$$=\pi_{ijk}+\left(1-\pi_{ijk}\right)\left(\frac{\theta_{ijk}}{\theta_{ijk}+\mu_{ijk}}\right)^{\theta_{ijk}}\left[\sum_{z=0}^{\infty}\left\{(1-p_{ijk})^{z}\frac{G\left(z+\ \theta_{ijk}\right)}{G(z+1)G\left(\theta_{ijk}\right)}\left(\frac{\mu_{ijk}}{\theta_{ijk}+\mu_{ijk}}\right)^{z}\right\}\right]$$

$$=\pi_{ijk}+\left(1-\pi_{ijk}\right)\left(\frac{\theta_{ijk}}{\theta_{ijk}+\mu_{ijk}}\right)^{\theta_{ijk}}\left[\sum_{z=0}^{\infty}\frac{G\left(z+\ \theta_{ijk}\right)}{G(z+1)G\left(\theta_{ijk}\right)}\left(\frac{\mu_{ijk}(1-p_{ijk})}{\theta_{ijk}+\mu_{ijk}}\right)^{z}\left(1\right.\right.$$

$$\left.\left.-\frac{\mu_{ijk}(1-p_{ijk})}{\theta_{ijk}+\mu_{ijk}}\right)^{\theta_{ijk}}\right]\left(1-\frac{\mu_{ijk}(1-p_{ijk})}{\theta_{ijk}+\mu_{ijk}}\right)^{-\theta_{ijk}}$$

$$=\pi_{ijk}+\left(1-\pi_{ijk}\right)\left(\frac{\theta_{ijk}}{\theta_{ijk}+\mu_{ijk}}\right)^{\theta_{ijk}}\left(1-\frac{\mu_{ijk}(1-p_{ijk})}{\theta_{ijk}+\mu_{ijk}}\right)^{-\theta_{ijk}}$$

$$=\pi_{ijk}+\left(1-\pi_{ijk}\right)\left(\frac{\theta_{ijk}}{\theta_{ijk}+\mu_{ijk}p_{ijk}}\right)^{\theta_{ijk}}$$

(20)

<u>Case-2</u>: For non-zero counts, *i.e.*, $Y_{ijk}(>0)=t=1,2,3,\ldots$

$$P\left[Y_{ijk}=t\middle|\pi_{ijk},\mu_{ijk},\theta_{ijk},p_{ijk}\right]=\sum_{z\geq t}P\left[Y_{ijk}=t\middle|Z_{ijk}=z,p_{ijk}\right]P\left[\,Z_{ijk}=z\middle|\pi_{ijk},\mu_{ijk},\theta_{ijk}\right]$$

$$= \sum_{z \geq t} \binom{z}{t} {p_{ijk}}^{t} (1 - p_{ijk})^{z-t} (1 - \pi_{ijk}) \frac{G(z + \theta_{ijk})}{G(z+1)G(\theta_{ijk})} \left(\frac{\theta_{ijk}}{\theta_{ijk} + \mu_{ijk}}\right)^{\theta_{ijk}} \left(\frac{\mu_{ijk}}{\theta_{ijk} + \mu_{ijk}}\right)^{z}$$

$$= (1 - \pi_{ijk}) \left(\frac{\theta_{ijk}}{\theta_{ijk} + \mu_{ijk}}\right)^{\theta_{ijk}} \sum_{z \geq t} \binom{z}{t} {p_{ijk}}^{t} (1 - p_{ijk})^{z-t} \frac{G(z + \theta_{ijk})}{G(z+1)G(\theta_{ijk})} \left(\frac{\mu_{ijk}}{\theta_{ijk} + \mu_{ijk}}\right)^{z}$$

$$= \frac{(1 - \pi_{ijk})}{G(t+1)G(\theta_{ijk})} \left(\frac{\theta_{ijk}}{\theta_{ijk} + \mu_{ijk}}\right)^{\theta_{ijk}} \left(\frac{p_{ijk}}{1 - p_{ijk}}\right)^{t} \sum_{z \geq t} \frac{G(z + \theta_{ijk})}{G(z - t + 1)} \left(\frac{\mu_{ijk}}{\theta_{ijk} + \mu_{ijk}}\right)^{z} (1 - p_{ijk})^{z}$$

Let, $z' = z - t$

$$= \frac{(1 - \pi_{ijk})}{G(t+1)G(\theta_{ijk})} \left(\frac{\theta_{ijk}}{\theta_{ijk} + \mu_{ijk}}\right)^{\theta_{ijk}} \left(\frac{p_{ijk}}{1 - p_{ijk}}\right)^{t} \sum_{z'=0} \frac{G(z' + t + \theta_{ijk})}{G(z'+1)} \left(\frac{\mu_{ijk(1-p_{ijk})}}{\theta_{ijk} + \mu_{ijk}}\right)^{z'+t}$$

$= (1$

$$- \pi_{ijk}) \frac{G(t + \theta_{ijk})}{G(t+1)G(\theta_{ijk})} \left(\frac{\theta_{ijk}}{\theta_{ijk} + \mu_{ijk}}\right)^{\theta_{ijk}} \left(\frac{\mu_{ijk} p_{ijk}}{\theta_{ijk} + \mu_{ijk}}\right)^{t} \sum_{z'=0} \frac{G(z' + t + \theta_{ijk})}{G(z'+1)G(t + \theta_{ijk})} \left(\frac{\mu_{ijk}(1 - p_{ijk})}{\theta_{ijk} + \mu_{ijk}}\right)^{z'}$$

$$= (1 - \pi_{ijk}) \frac{G(t + \theta_{ijk})}{G(t+1)G(\theta_{ijk})} \left(\frac{\theta_{ijk}}{\theta_{ijk} + \mu_{ijk}}\right)^{\theta_{ijk}} \left(\frac{\mu_{ijk} p_{ijk}}{\theta_{ijk} + \mu_{ijk}}\right)^{t} \left(1 - \frac{\mu_{ijk}(1 - p_{ijk})}{\theta_{ijk} + \mu_{ijk}}\right)^{-(t + \theta_{ijk})}$$

$$= (1 - \pi_{ijk}) \frac{G(t + \theta_{ijk})}{G(t+1)G(\theta_{ijk})} \left(\frac{\theta_{ijk}}{\theta_{ijk} + \mu_{ijk} p_{ijk}}\right)^{\theta_{ijk}} \left(\frac{\mu_{ijk} p_{ijk}}{\theta_{ijk} + \mu_{ijk} p_{ijk}}\right)^{t}$$

(21)

Here, the Eq. 20 and 21 are in the form of Eq. 3, which indicates the distribution of Y_{ijk} is also $ZINB\left(\pi_{ijk}, \mu_{ijk} p_{ijk}, \theta_{ijk}\right)$. Now, the PMF of Y_{ijk} can be expressed in Eq. 22.

Let, $\mu_{ijk} p_{ijk} = {\mu'}_{ijk}$

$$P[Y_{ijk} = y | \pi_{ijk}, \mu_{ijk}, \theta_{ijk}, p_{ijk}] = \begin{cases} \pi_{ijk} + (1 - \pi_{ijk}) \left(\frac{\theta_{ijk}}{\theta_{ijk} + {\mu'}_{ijk}}\right)^{\theta_{ijk}} \\ (1 - \pi_{ijk}) \frac{G(t + \theta_{ijk})}{G(t+1)G(\theta_{ijk})} \left(\frac{\theta_{ijk}}{\theta_{ijk} + {\mu'}_{ijk}}\right)^{\theta_{ijk}} \left(\frac{{\mu'}_{ijk}}{\theta_{ijk} + {\mu'}_{ijk}}\right)^{y} \end{cases}$$

(22)

When $p_{ijk} = 1$ ($capture\ rate\ is$ 100 %), *i.e.,* when all the mRNA molecules in the cell are captured (a perfect deep sequencing of cells), the distribution of observed read counts is same as the distribution of

true read counts. Further, the expected value and variance of Y_{ijk} can be derived as shown in Eq. 4 and 5 and are expressed in Eq. 23 and 24.

$$E(Y_{ijk}) = (1 - \pi_{ijk})\mu_{ijk}p_{ijk} \tag{23}$$

$$Var(Y_{ijk}) = (1 - \pi_{ijk})\mu_{ijk}p_{ijk}(1 + \pi_{ijk}\mu_{ijk}p_{ijk} + \frac{\mu_{ijk}p_{ijk}}{\theta_{ijk}}) \tag{24}$$

2.3 *Extraction of signals from observed scRNA-seq UMI data*

The expected values of gene-wise sample mean, and sample variance of the observed scRNA-seq count data can be obtained as follows. Here, we assume that the observed count data are arise from the NB distribution as given in Eq. 2 and the transcriptional capture efficiencies across the gene remain same, *i.e.*, $p_{ij1} = p_{ij2} = \cdots = p_{ijK} = p_{ij}$. (25)

Now, the model parameters, μ_{ijk} and θ_{ijk} for the genes remain same across the cells. Let, $\bar{Y}_{..k}$: be the sample mean expression values of *k-th* gene and its expected values can be given as:

$$E(\bar{Y}_{..k}) = \frac{1}{N}\sum_{i=1}^{N}\frac{1}{M_i}\sum_{j=1}^{M_i}E(Y_{ijk}) = \frac{1}{N}\sum_{i=1}^{N}\frac{1}{M_i}\sum_{j=1}^{M_i}E\{E(Y_{ijk}|Z_{ijk})\}$$

$$= \frac{1}{N}\sum_{i=1}^{N}\frac{1}{M_i}\sum_{j=1}^{M_i}E(Z_{ijk}p_{ijk}) = \frac{1}{N}\sum_{i=1}^{N}\frac{1}{M_i}\sum_{j=1}^{M_i}(\mu_{ijk}p_{ijk})$$

(26)

Under the assumption given in Eq. 25, the expression of expected value of sample mean can be written as:

$$E(\bar{Y}_{..k}) = \frac{1}{N}\sum_{i=1}^{N}\frac{1}{M_i}\sum_{j=1}^{M_i}(\mu_k p_{ij}) = \frac{1}{N}\mu_k\sum_{i=1}^{N}\frac{1}{M_i}\sum_{j=1}^{M_i}p_{ij} = \mu_k\bar{p}_{..}$$

(27)

Further, the variance of the observed scRNA-seq data can be obtained as:

$$V(Y_{ijk}) = E\{V(Y_{ijk}|Z_{ijk})\} + V\{E(Y_{ijk}|Z_{ijk})\}$$

$$= E\left(Z_{ijk}p_{ijk}\right)\left(1-p_{ijk}\right) + V\left(Z_{ijk}p_{ijk}\right)$$

$$= p_{ijk}\left(1-p_{ijk}\right)\mu_{ijk} + p_{ijk}^2\left(\mu_{ijk} + {\mu_{ijk}^2}/{\theta_{ijk}}\right)$$

$$= \mu_{ijk}p_{ijk}\left(1 + {\mu_{ijk}p_{ijk}}/{\theta_{ijk}}\right) \tag{28}$$

Under the assumption in Eq. 25, the $V\left(Y_{ijk}\right)$ becomes:

$$V\left(Y_{ijk}\right) = \mu_k p_{ij}\left(1 + {\mu_k p_{ij}}/{\theta_k}\right) \tag{29}$$

Let, S_k^2 be the sample variance of k^{th} gene. Then its expected value can be derived as follows.

$$E(S_k^2) = \frac{1}{N}\sum_{i=1}^{N}\frac{1}{(M_i-1)}\sum_{j=1}^{M_i}\left\{V\left(Y_{ijk}\right) + E\left(Y_{ijk}\right)^2\right\}$$

$$-\frac{1}{N(N-1)}\sum_{i\neq i'=1}^{N}\frac{1}{M_i(M_i-1)}\sum_{j\neq j'=1}^{M_i}E\left(Y_{ijk}\right)E\left(Y_{i'j'k}\right)$$

$$= \mu_k\bar{p}_{..} + \frac{{\mu_k}^2}{\theta_k}\overline{p_{ij}^2} + {\mu_k}^2 var(p_{ij}) \tag{30}$$

3. <u>Supplementary Document S3</u>: Testing for zero inflation parameters of genes

In order to test the statistical significance of the zero inflation parameters of k^{th} gene π_{ijk} of the ZINB model, we adopt the following Generalized Likelihood Ratio Test (GLRT) procedure. Here, for the testing purpose, we define the following null hypothesis.

$$H_0: \pi_{ijk} = 0 \; vs. \; H_1: \pi_{ijk} \neq 0$$

where, H_0: null hypothesis; H_1: alternate hypothesis. In other words, null hypothesis tells us that k^{th} gene is not zero inflated, and subsequently, the scRNA-seq data structure is same as RNA-seq data. Further, if we fail to reject H_0, then the RNA-seq DE tools can be used for DE analysis of scRNA-seq data with the expectation of satisfactory results.

The above-mentioned test, $H_0\ vs. H_1$, can be tested through GLRT and the test statistic is given in Eq. 31.

$$-2ln\alpha = -2\{l(\boldsymbol{\Omega}_k = \widehat{\boldsymbol{\Omega}}_{k0};\ Y_{ijk}) - l(\boldsymbol{\Omega}_k = \widehat{\boldsymbol{\Omega}}_k;\ Y_{ijk})\} \quad (31)$$

where, $\widehat{\boldsymbol{\Omega}}_{k0}$: Maximum Likelihood Estimator (MLE) of $\boldsymbol{\Omega}_k$ for k^{th} gene under the constraint of H_0 and $\widehat{\boldsymbol{\Omega}}_k$: unconstrained MLE of $\boldsymbol{\Omega}_k$ for k^{th} gene, $\boldsymbol{\Omega}_k$: parametric space for k^{th} gene, *i.e.*, $\boldsymbol{\Omega}_k = \{\mu_k, \theta_k, \pi_k\}$. The test statistic in Eq. 31 is asymptotically distributed as Chi-square distribution with 1 degree of freedom under H_0.

We applied the above procedure to Tung *et al.*'s scRNA-seq data (Table 1) to test the statistical significance of the zero inflation parameters of genes. The results are shown in Figure S1. It can be observed that most of the genes in scRNA-seq data is found to be zero inflated as their corresponding *p-values* are less than the (1%) level of significance value (Figure S1). This finding motivates us to develop a statistical approach for testing of differential zero inflation of genes in single-cell studies.

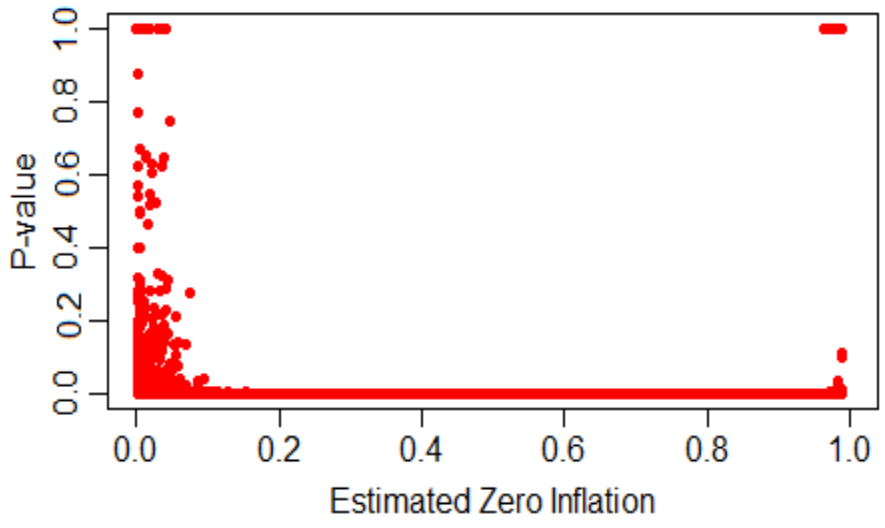


Figure S1. Plotting of estimated value of zero inflation parameter and their corresponding *p-values*. X-axis represents estimated values of zero inflation, (higher value of zero inflation parameter means more number of zeros found in the expression vector of that gene) and Y-axis represents the computed statistical significance value for the zero inflation parameter, lesser the value represents the gene is more zero inflated.

4. Supplementary Document S4: Application of count data models to zero-inflated and overdispersed real data

4.1 *European red mite data*

Earlier experimental studies have shown that the scRNA-seq (UMI) read count data is zero inflated and over dispersed [11–17]. Hence, we consider a published zero-inflated and over-dispersed data on counts of European red mites on apple leaves [18] to study the suitability of different discrete data models, *i.e.*, NBD, ZINB, PD, HD, and ZIPD, for fitting the data. This example data characterizes the distribution of number of adult female European red mite on 150 apple leaves [18]. Here, 25 leaves were selected at random from each of the six McIntosh trees in a single orchard receiving the same spray treatment and the number of adult females counted on each leaf [18,19]. The expected frequencies are computed for each of the models and are shown in Table S1. The goodness of fit to the experimental data is assessed through Akaike Information Criterion (AIC) and Bayesian Information Criterion (BIC).

Table S1. Fitting of well-known discrete models to over-dispersed and zero-inflated European red mite data.

Read	Obs. Freq.	NBD	ZINB	PD	ZIPD	HD
0	70	68.49	69.1	47.65	69	64.65
1	38	37.6	35.01	54.64	28.67	34.71
2	17	20.1	20.65	31.33	25.68	29.02
3	10	12.7	11.21	11.97	15.34	12.25
4	9	5.69	5.91	3.43	6.87	6.07
5	3	3.02	3.06	0.79	2.46	2.14
6	2	1.6	1.57	0.15	0.74	0.81
7	1	0.85	3.79	0.02	1.19	0.25
8	0	0.6	0.1	0.1	0.67	0.02
Total	150	150.65	150.4	150.08	150.62	149.92
Parameter Estimates (MLE)		μ=1.147 θ=1.025	$\mu = 1.283$ $\theta = 1.39$ $\pi = 0.107$	$\mu = 1.146$	$\mu = 1.791$ $\pi = 0.367$	$\mu = 1.147$ $\varphi = 1.757$
#Parameter		2	3	1	2	2
Likelihood		-224.71	-223.43	-242.86	-226.44	-225.98
AIC		453.40	452.80	487.72	456.80	455.95
BIC		453.75	453.33	487.90	457.15	456.31

μ: Mean; θ: size; π: zero-inflation probability; φ: dispersion index (ratio of variance to mean); AIC: Akaike Information Criterion; BIC: Bayesian Information Criterion; Obs. Freq: Observed Frequencies

It is observed that the expected frequencies computed from ZINB model are closer to their observed values as compared to other existing models, *i.e*., NBD, PD, ZIPD and HD. Further, the AIC and BIC values for the ZINB model is lowest followed by NBD for the given zero inflated and over dispersed European red mite data as compared to PD, ZIPD and HD (Table S1). This indicates, for fitting over-dispersed and zero inflated datasets like scRNA-seq data, the ZINB model provides a better fit as compared to other count models. Moreover, we validate the above claim by using another overdispersed and zero-inflated dataset.

3.2 Application of ZINB model to embryonic mouse cysts data

We consider another publicly available zero-inflated and over-dispersed data on counts of cysts in embryonic mouse [20] to study the suitability of fitting different discrete models to the underlying data. Here, we considered data on counts of cysts in embryonic mouse kidneys which

had been subjected to steroids, taken from McElduff *et al.* [20]. This data reveals the details on the effect of a low protein diet in mice on kidney development in their offspring. Data on counts of cysts in embryonic mouse kidneys which had been subjected to steroid were featured in this study. Then, the count data models, such as NBD, ZINB, PD, ZIPD and HD are fitted on this data. Further, the parameters of these models are estimated through MLE. The results are shown in Table S2. The goodness of fit to the experimental data is assessed through AIC and BIC.

Table S2. Fitting of discrete models to over-dispersed and zero-inflated cyst count data.

Read	**Obs. Freq.**	**NBD**	**ZINB**	**PD**	**ZIPD**	**HD**
0	65	63.29	64.99	25.1	65.03	45.36
1	14	17.56	14.01	37.32	5.1	13.75
2	10	8.98	9.11	27.74	8.87	28.92
3	6	5.72	6.27	13.74	10.28	8.35
4	4	3.91	4.44	5.11	8.93	9.19
5	2	2.79	3.2	1.52	6.21	2.53
6	2	2.04	2.33	0.38	3.6	1.94
7	2	1.52	1.71	0.08	1.79	0.51
8	1	1.15	1.26	0.01	0.78	0.31
9	1	0.88	0.93	0	0.3	0.08
10	1	0.68	0.69	0	0.1	0.04
11	2	0.52	0.52	0	0.03	0.01
12	1	0.41	0.38	0	0.01	0
Total	111	110.95	110.84	111	111.03	110.99
Parameters (MLE)		μ=1.49 θ=0.31	$\mu = 2.285$ $\theta = 0.698$ $\pi = 0.349$	$\mu = 1.486$	$\mu = 3.476$ $\pi = 0.572$	$\mu = 1.487$ $\varphi = 1.796$
#Parameters		2	3	1	2	2
Likelihood		-175.22	-172.8	-263.25	-191.9	-202.84
AIC		354.44	351.60	528.50	387.80	409.68
BIC		354.53	351.74	528.55	387.89	409.77

#Parameters: number of parameters; μ: Mean; θ: size; π: zero-inflation probability; φ: dispersion index (ratio of variance to mean); AIC: Akaike Information Criterion; BIC: Bayesian Information Criterion; Obs. Freq: Observed Frequency

It is observed that the expected frequencies computed from the ZINB model are closer to their observed values as compared to other models. Further, the AIC and BIC values for the ZINB model are lowest followed by NBD for the given zero inflated and over dispersed cyst count data as compared to PD, ZIPD and HD (Table 2). This indicates, for fitting over-dispersed

and zero inflated datasets like scRNA-seq data, ZINB model provides a better fit as compared to other count models, *i.e.,* NBD, PD, ZIPD and HD (Tables S1, S2).

Usually, the NBD is extensively used for modeling and fitting of RNA-seq count data. But it performed poor when the data is simultaneously zero inflated and overdispersed. It can be observed that, for fitting overdispersed and zero inflated datasets like scRNA-seq data, ZINB model provides a better fit as compared to other count data models. Further, we also test the ability of NBD, and ZINB models to estimate the mean and dispersion parameters for scRNA-seq count data through simulation as described in Document S5.

5. Supplementary Document S5: Bulk RNA-seq *vs.* scRNA-seq data

NBD model is implemented in popular RNA-seq DE tools, such as DESeq2, edgeR, baySeq, cuffdiff, [2,3,21,22] *etc*., for analysis of RNA-seq count data. Such datasets are overdispersed. It is pertinent to test its performance on scRNA-seq datasets, which is both zero inflated and overdispersed. Hence, we applied NBD and ZINB models to test their ability estimate the mean and dispersion parameters for scRNA-seq count data under an artificial setup. For this purpose, parameter estimates for BTG4 gene from a public scRNA-seq dataset of human preimplantation of embryonic cells, available in DEsingle R package [7] is used to simulate count expression data through ZINB model. We assume the mean and dispersion values for the same gene as true values of the parameters to simulate count data over 500 cells. Then both the models, *i.e*., NBD and ZINB are applied on the simulated data to estimate the parameters through MLE. For this purpose, we executed *zinb* and *glm.nb* functions implemented in pscl [23]and MASS statistical R packages respectively. The simulation is repeated for 100 time and the results are shown in Table S3.

Table S3. Comparative analysis of NBD and ZINB for estimation of parameters from scRNA-seq data.

		NBD				ZINB			
Parameters	**True values**	**Estimate**	**Bias**	**MSE**	**95% CI**	**Estimate**	**Bias**	**MSE**	**95% CI**
Mean (μ)	**2.28**	1.483	-0.797	0.649	(1.325, 1.641)	2.328	0.046	0.132	(2.254, 2.410)
Dispersion (θ^{-1})	**1.45**	3.315	1.865	3.597	(2.943, 3.687)	1.433	0.048	0.263	(1.333, 1.534)
Zero inflation prob. (π)	**0.35**	-	-	-	-	0.353	0.003	0.011	(0.331, 0.371)

Number of cells: 500; number of simulations: 100; MSE: Mean Standard Error; CI: Confidence Interval

Our analysis indicates that NBD model underestimate the mean and over-estimate the dispersion parameters for scRNA-seq data (Table S3). Further, the ZINB model provides better estimates of mean and dispersion, which are close to their true values for scRNA-seq data. Further, ZINB model has lower bias and MSE as compared to NBD model (Table S3). It is interesting to note that 95 % confidence interval of parameters for NBD does not contain the true values of the parameters. While this observation is quite satisfactory for ZINB model. This indicates the better suitability of ZINB model for modeling the zero-inflated and over-dispersed scRNA-seq count data and provides better estimates of the parameters. The reason may be attributed as NBD model thus accommodates excess zeros by underestimating the mean and overestimating the dispersion parameters. This phenomenon jeopardizes the statistical power of the NBD model-based RNA-seq DE tools to discover the DE genes in the presence of zero-inflation, when applied to scRNA-seq UMI dataset. Hence, it is necessary to develop the ZINB model based DE methods and tools for downstream analysis of the scRNA-seq data.

6. <u>Supplementary Document S6</u>: Illustration of the effects of cell clusters on mean parameter

In the DE analysis of the scRNA-seq data, the cells are clustered into cell clusters and further these cell clusters are divided into two cellular groups (taking one cluster as one group and the remaining clusters as another group), as shown in Figure S2. Further, this cell cluster information is usually kept out during the model development process of the DE analysis. These cell clusters may have significant influence on the mean of non-zero scRNA-seq counts. To test this claim, we took a toy example scRNA-seq dataset having 200 genes and 150 cells (Group 1: 50; Group 2: 100), available in DEsingle R package [7]. This example data is a subset of publicly available human preimplantation embryonic cells scRNA-seq dataset, available in ArrayExpress with accession number E-MTAB-3929 [24]. Using this toy example data, we determined the optimum of number of cell clusters by following the proposed technique described in Document S9 of this Supplementary Materials. In other words, the 150 cells present in the toy data are clustered using *K*-means clustering through executing *kmeans* function implemented in stats R package for different values of *K*. Here, we have set the values of K as h=2, 3, 4, …, 20 and then computed the values of clustering index, r_h, given in Document S9, for each value of h. Through the Graphical method, the value of optimum number of cell cluster is chosen as 9 (*i.e.*, the point, where the curve gets flatten), as shown in Figure S2A.

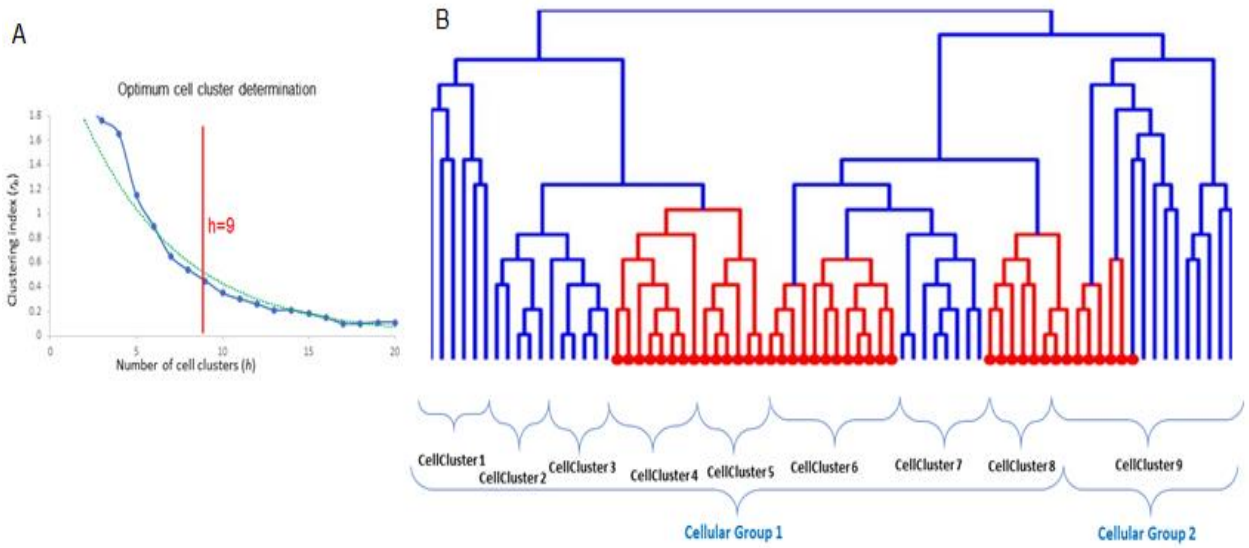


Figure S2 Schematic representations of determination of number of cell clusters and division of cell clusters into cellular groups for DE analysis. (A) Determination of number of cell clusters; (B) Detection of cell groups. The number of cell clusters are determined by using the method described in Document S9. First, the cells are clustered into different cell clusters by using the count expression data. Then, randomly picking one cell cluster as one group and remaining cell clusters as another group, for detection of DE genes among these two cellular groups.

Through *kmeans* clustering function implemented in stats R package, we divide the 150 cells into 9 cell clusters (Figure S2A). For DE analysis, we took cluster 9 as cellular group 2 and remaining cell clusters as cellular group 1 (Figure S2B). Then, we model the mean of non-zero counts using ZINB model under a GLM framework by providing group and cell cluster information as auxiliaries. For this purpose, we executed *pscl* function implemented in pscl R package [23]. The effects cellular group and cell clusters on the mean of non-zero counts of a single gene, *i.e.*, BTG4 gene are shown in Table S4.

Table S4. Effect of cell clusters and groups on mean of non-zero scRNA-seq counts of BTG4 gene.

	No. cells	**Max**	**Min**	**#Zeros**	**Avg. Exp.**	**Co-efficient**	**Z-value**	**Sig.[a]**
Intercept	-	-			-	7.35	11.81	***
Group 1	50					Ref.	Ref.	Ref.
Group 2	100	-			-	-2.8016	1.703	*
Cell Cluster 1	16	179	0	8	27.9375	-2.1	-1.875	*
Cell Cluster 2	47	437	0	30	23.70213	-3.55	-4.277	***
Cell Cluster 3	3	45	0	2	15	-2.8527	-1.626	NS
Cell Cluster 4	8	145	0	6	23	-2.97	-2.78	**
Cell Cluster 5	6	3496	466	0	1557.167	-2.18	-1.356	NS
Cell Cluster 6	14	13	0	11	1.857	-3.005	-3.84	***
Cell Cluster 7	32	497	0	23	24.375	-4.67	-3.7	***
Cell Cluster 8	16	308	0	6	49.937	-3.041	-3.89	***
Cell Cluster 9	8	20	0	5	3.125	-3.477	-2.735	**
log(theta)	-					-0.841	-2.706	**

Max: Maximum read count; Min: Minimum read count; Avg. Exp.: Average expression values; a: comparing group 1 vs. remaining Cell clusters (e.g., Cell cluster 1 vs. Cell clusters 2-9) *, **, *** represents values significant at 5%, 1% and 0.1 % level of significance; (.,.): number of cells in [a]: The above table describes the analysis for BT4 gene, if there are *K* genes *K* such tables can be formulated. Also, if there are *N* cell clusters, there will be N(N-1)/2 grouping combinations and there will be *KN(N-1)/2* such tables.

Our illustration indicated that cellular group 2 has significant effects on the mean counts of BTG4 gene, which means that BTG4 is expressed differentially with respect to group 1. Further, most of the cell clusters have significant effects on the mean count of BTG4 gene. Specifically, all the cell clusters except cell clusters 3 and 5 have significant effects on mean count of BTG4. Here, we can hypothesize, the cell clusters may influence the DE analysis of genes in scRNA-seq data. This toy data example motivates us to develop statistical approach and tool for DE analysis of scRNA-seq data by integrating cell cluster information in the modeling process. Further, we found that for 63 genes (= 32 % of total 200 genes present in this toy data), all the 9 cell clusters have significant effect on the mean transcript abundance (Figure S2.1A). In other words, all the cell clusters have significant effects on the mean of non-zero read counts for 63 genes at 5 % level of statistical significance as shown in Figure S2.1A.

Further, to get a realistic idea about the effects of cell clusters on the mean of non-zero read counts, we used a large real scRNA-seq UMI counts data from Tung et al. (with 15955 genes and 576 cells) (described in Tables 1, S6 and Supplementary Document S8). The results for this data are shown in Figure S2.1B. The results indicated that for 3651 genes out of 15955 genes (~ 23 % of total genes), all the cell clusters have significant effects on the mean of non-zero read counts of the genes at 1% level of significance (Figure 2.1 B). For both the data cases, it was observed that all the cell clusters have significant effects on the mean counts of quite larger proportion of genes. Hence, this cell cluster information must be incorporated in the statistical model building process for the differential expression analysis of genes in single-cell datasets.

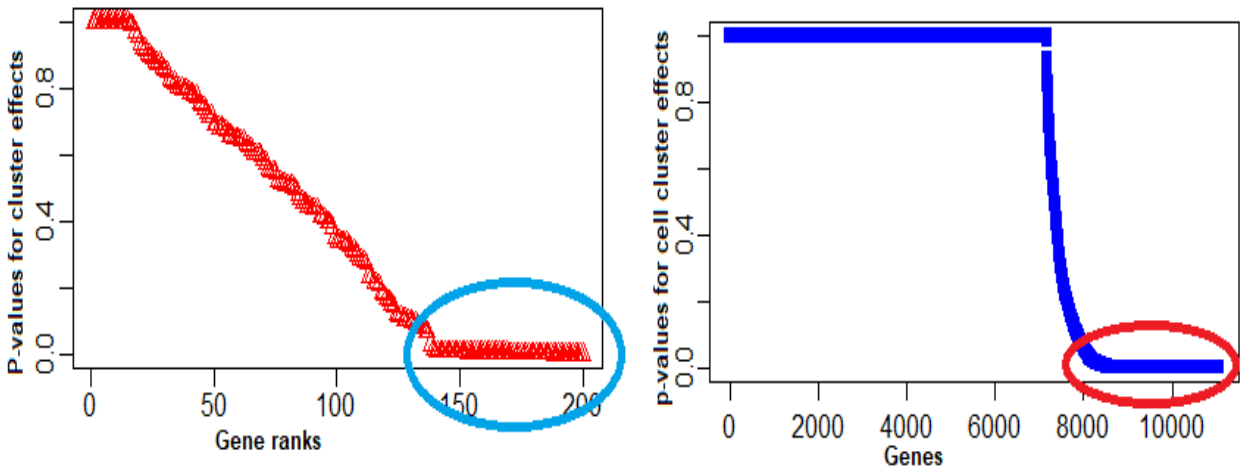


Figure S2.1. Effects of cell clusters on the mean of non-zero read counts of genes in the toy example data.

7. Supplementary Document S7: Normalization of scRNA-seq read (UMI) count data

7.1 *Trimmed mean of M-values (TMM) method*

SwarnSeq method integrates TMM normalization method originated from edgeR R package [25] to normalize the read (UMI) counts data. The details about the TMM can be found in Robinson and Oshlack, 2010's work [26]. It deals with calculation of effective libraries sizes, which are then used as part of the per-sample normalization.

TMM normalization adjusts library sizes based on the assumption that most genes are not differentially expressed. Therefore, it is important not to make subsets of the count data before doing statistical analysis or visualization, as this can lead to differences being normalized away.

Algorithm

- 'log CPM' (Counts per Million) values are calculated for each gene. The CPM calculation uses the effective library sizes as calculated by the TMM normalization.
- After this first normalization, a second one is performed across samples for each gene: the counts for each gene are mean centered and scaled to unit variance.
- Genes or transcripts with zero expression across all samples or invalid values (NaN or +/- Infinity) are removed.

7.2 *DESeq.norm*

SwarnSeq method integrates a modified median normalization method originated from DESeq to normalize the read (UMI) counts data. Let Z_{jk} denote the read counts of gene *k* in cell j, then the size factor of cell j is estimated by using Eq. 32.

$$\widehat{S_j} = \text{median}_k \frac{Z_{jk}}{(\prod_{v=1}^{K} Z_{kv})^{1/K}} \tag{32}$$

where, $Z_{jk} \neq 0$ namely only the non-zero read counts of each gene are used. Finally, the normalized read counts of gene *k* in cell j is calculated by $\frac{Z_{jk}}{\widehat{S_j}}$.

8. Supplementary Document S8: Collection and Pre-processing of scRNA-seq counts datasets

A. ***Lung cancer data*** (www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE111108)

This dataset is publicly available from Gene Expression Omnibus (GEO) repository GSE111108 [27]. The ScRNA-seq data are generated for an equal mixture of cells from the 3 Human Lung Adenocarcinoma cell lines through 10X Genomics protocol and sequenced with Illumina NextSeq 500. At the preliminary stage, we removed the cells whose library size is less than 1800 and also further removed the genes which have non-zero expressions in ≤ 5 cells. Through this process, we selected expression counts of 17326 genes over 2126 cells for further analysis. Hence, we used the o*ptimcluster* function implemented in SwarnSeq R package to decide the number of optimum cell clusters [see Document S9 for details]. For this purpose, we set the seed value at 1712 and, found that the 2126 cells are clustered into 8 optimum cell clusters (Figure S3). Further, for Differential Expression (DE) analysis, we took cell cluster 3 as group 1 and remaining cell clusters as group 2. A brief description about this data is given in Table S6.

B. ***Pluripotent stem cell data*** (www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE77288)

This dataset is publicly available in NCBI GEO database with accession id GSE77288 [28] and used for benchmarking scRNA-seq experiment. We downloaded the filtered UMI count matrix from their GitHub repository (https://github.com/jdblischak/singleCellSeq). The full dataset contains three Yoruba (YRI) induced pluripotent stem cell (iPSC) lines, with three 96-well plates per individual. Here, we used the ERCC spike-ins, UMI and molecular concentration data were used. We only used data of two individual cell lines NA19101 (288 cells) and NA19239 (288 cells) for further analyses. Here, we have not removed any cells from analysis. also further removed the genes which have non-zero expressions in ≤ 5 cells. Through this process, we selected expression counts of 15955 genes over 576 cells for further analysis. The SwarnSeq modeling requires the cluster information. Hence, we used the *optimcluster* function implemented in SwarnSeq R package to decide the number of optimum cell clusters [see Document S9 for details]. For this purpose, we set the seed value at 108 and, found that the 576 cells are clustered into 10 optimum cell clusters (Figure). Further, for the DE analysis, we took two cell lines (*i.e.*, NA19101 and NA19239) as two cellular groups. A brief description about this data is given in Table S6.

C. *Mouse blood cell data* (https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE109999)

This dataset is publicly available in NCBI GEO database with accession GSE109999 [27]. For this experiment, we downloaded the count expression data as it has undergone rigorous preprocessing by the authors of the original publication. Here, the blood cells are derived from B lymphocytes (B220+ FSC-Alow), erythroblasts (Ter119+ CD44+, FSC-Amid/high), granulocytes (Mac1+ Gr1+) and high-end progenitor/stem (Lin- Kit+ Sca1+) from the bone marrow of a C57BL/6 10-13-week-old female. T cells (CD3+ FSC-Alow) were isolated from the thymus of the same mouse. Gene expression in mouse blood cells from sorted via FACS analysis

into a 384-well plate were profiled using a modified CEL-seq2 protocol. Here, we have not removed any cells from analysis. also further removed the genes which have non-zero expressions in ≤ 5 cells. Through this process, we selected expression counts of 13055 genes over 383 cells for further analysis. The SwarnSeq modeling requires the cluster information. Hence, we used the *optimcluster* function implemented in SwarnSeq R package to decide the number of optimum cell clusters [see Document S9 for details]. For this purpose, we set the seed value at 110 and, found that the 383 cells are clustered into 9 optimum cell clusters (Figure S3). Further, for the DE analysis, we took cell cluster 2 (180 cells) as one group and remaining cell clusters (203 cells) as other group. A brief description about this data is given in Table S6.

D. ***Liver cell data*** (https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE115469)

This dataset is publicly available, and we downloaded counts expression dataset from the NCBI GEO database with accession GSE115469 [29]. We directly took the count data, as it has undergone rigorous pre-procession, mapping, and other data analysis procedures by the authors of the original publication. In this study, the fractionated fragile, fresh hepatic tissue from human livers is obtained to get viable parenchymal and non-parenchymal cells. Then, expression profiling of cells is done by high throughput sequencing. The data consists the counts expression data of 8444 cells. To reduce the size of the data, we removed the cells whose sizes are less than 1500 and genes which have non-zero counts in 5 cells. Through this, the expressions data of 17316 genes over 5466 cells for further analysis. The SwarnSeq modeling requires the cluster information. Hence, we used the *optimcluster* function implemented in SwarnSeq R package to decide the number of optimum cell clusters [see Document S9 for details]. For this purpose, we set the seed value at 222 and, found that the 5466 cells are clustered into 16 optimum cell clusters (Figure S3). Further, for the DE analysis, we took cell cluster 3 (1852 cells) as one group

and remaining cell clusters (3614 cells) as other group. A brief description about this data is given in Table S6.

E. ***Mouse cell data*** (https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE29087)

The dataset is publicly available in NCBI GEO database with accession GSE29087 [11] and widely used for benchmarking of scRNA-seq DE methods and tools. Single-cell RNA-Seq expression profiles were generated for 92 cells, and clustered to form a two-dimensional cell map onto which expression data was projected. Further, overall design of 92 single cells (48 mouse ES cells, 44 mouse embryonic fibroblasts and 4 negative controls) were analyzed by single cell tagged reverse transcription (STRT). Negative control cells are removed from the further analysis. For DE analysis, ES and MEF cell lines are considered as two cellular groups. Here, we have not removed any cells from analysis, and further removed the genes which have non-zero expressions in ≤ 5 cells. Through this process, we selected expression counts of 11436 genes over 92 cells. The SwarnSeq modeling requires the cluster information. Hence, we used the *optimcluster* function implemented in SwarnSeq R package to decide the number of optimum cell clusters [see Document S9 for details]. The results are shown in Table S6, Figure S3.

F. ***Adipose stem/stromal cell data*** (www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE53638)

This dataset is publicly available from NCBI Gene Expression Omnibus (GEO) database with Accession GSE53638 [30]. Initially, cells are collected during directed differentiation of human adipose-derived stem/stromal cells and further, 11, 116 cells are profiled by the authors of the original publication. Here, cells were collected at different stages and different time points (day 0, day 3 and day 7) of differentiation of human adipose-derived stem/stromal cells. FACS sorted cells were sequenced using the SCRB-seq protocol with UMI. To study the performance of scRNA-seq DE tools, we used two group comparison settings based on different time points, i)

Data 1 (Day 0 (1245 cells, baseline) vs. Day 3 (590 cells), (ii) Data 2 (Day 0 (1245 cells, baseline) vs. Day 7 (1023 cells), and (iii) Data 3 (Day 7 (1023 cells, baseline) vs. Day 3 (590 cells). Here, we have not removed any cells from analysis, and further removed the genes which have non-zero expressions in ≤ 5 cells. Through this process, we selected expression counts of 14863, 15637, and 15015 genes over 1835, 2268, 1613 cells for Data 1, Data 2, and Data 3, respectively. The SwarnSeq modeling requires the cluster information. Hence, we used the *optimcluster* function implemented in SwarnSeq R package to decide the number of optimum cell clusters [see Document S9 for details]. The results are shown in Table S6, Figure S3.

G. *Mouse embryonic cell data* (www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE65525)

Mouse embryonic cell data is a publicly available scRNA-seq count data in NCBI GEO database with accession id GSE65525 [31]. Here, the mouse embryonic stem cells expressions were profiled through high throughput sequencing using a droplet-microfluidic approach. The experimental design includes A total of 8 single cell data sets are submitted: 3 for mouse embryonic stem (ES) cells (1 biological replicate, 2 technical replicates); 3 samples following LIF withdrawal (days 2,4, 7); one pure RNA data set (from human lymphoblast K562 cells); and one sample of single K562 cells. We took the counts data, which are already undergone through several quality control steps by the authors of the original publication. Further, at the preliminary stage, we removed the cells whose library size is less than 1500 and also removed the genes which have non-zero expressions in ≤ 5 cells. Through this process, we selected expression counts of 23971 genes over 1481 cells for further analysis. The SwarnSeq modeling requires the cluster information. Hence, we used the *optimcluster* function implemented in SwarnSeq R package to decide the number of optimum cell clusters [see Document S9 for details]. For this purpose, we set the seed value at 108 and, found that the 1481 cells are clustered into 11

optimum cell clusters (Figure S3). Further, for the DE analysis, we took day 4 (cells 683) as group 1 and day 7 (798 cells) as group 2. A brief description about this data is given in Table S6.

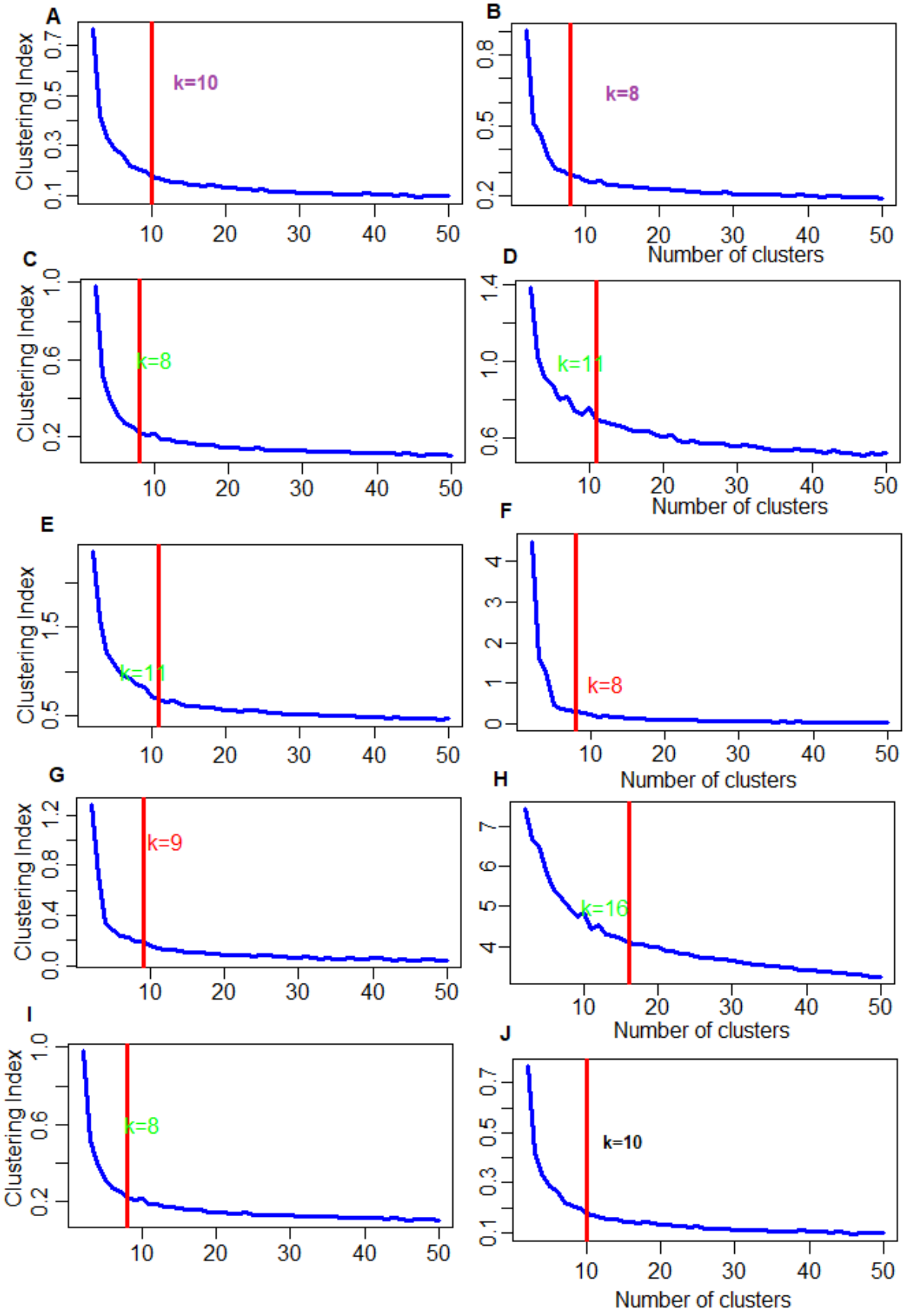
H. *Human Embryonic Kidney cell data*

(https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE92495)

This dataset is publicly available in NCBI GEO database with accession id GSE92495 [32]. This experiment contains scRNA-seq dataset of Mouse and Human Embryonic Kidney (HEK)/3T3 mixing, PBMCs, and TB-exposed cell Macrophages generated through Seq-Well, a portable, low-cost platform for single-cells. For benchmarking scRNA-seq DE tools, we considered count dataset for HEK cells, as it has expressions of thousands of genes over a large number of cells. Further, at the preliminary stage, we removed the cells whose library size is less than 1500 and also removed the genes which have non-zero expressions in ≤ 5 cells. Through this process, we selected expression counts of 15524 genes over 1453 cells for further analysis. The SwarnSeq modeling requires the cluster information. Hence, we used the *optimcluster* function implemented in SwarnSeq R package to decide the number of optimum cell clusters [see Document S9 for details]. For this purpose, we set the seed value at 208 and, found that the 1453 cells are clustered into 8 optimum cell clusters (Figure S3). Further, for the DE analysis, we took cell cluster 8 (cells 537) as group 1 and remaining cell clusters (916 cells) as group 2. A brief description about this data is given in Table S6.

Figure S3. Cluster analysis and determination of optium number of cell clusters for real scRNA-seq datasets.

Clustering analysis is performed on the scRNA-seq datasets through k-means algorithm. The optimum number of cell clusters is determined for each scRNA-seq datasets through OptimCluster function implemented in SwarnSeq R package. X-axis represents number of cell clusters. Y-axis represents clustering index. Here, the number of cell clusters is kept in the range of [2, 50]. Vertical lines represent the optimum number of cell clusters with k value. The graphs are shown for for (A) GSE53638 (Data 1); (B) GSE77728; (C) GSE53638 (Data 3); (D) GSE53638 (Data 2); (E) GSE29087; (F) GSE65525 (G) GSE111108; (H) GSE92495; (I) GSE115469; (J) GSE109999.


A
Clustering Index
k=10
B
k=8
Number of clusters
C
k=8
D
k=11
Number of clusters
E
k=11
F
k=8
Number of clusters
G
k=9
H
k=16
Number of clusters
I
k=8
J
k=10
Number of clusters

9. Supplementary Document S9: Determination of number of cell clusters and cellular groups for DE analysis

For performing DE analysis through the tested methods, the cells present in the data needs to be grouped into two cellular populations/types. When the cell types information is not known *a-priori*, then we used K-means clustering technique to determine the cell types. Further, the SwarnSeq method also required the cell cluster information to model the observed UMI counts. Therefore, we developed an algorithm to determine the optimum number of cell clusters that the cells need to be grouped based on the observed UMI count data, which is given as follows.

Let, $Y_{ij.}$: mean expression value of j^{th} cell in i^{th} cell cluster; $Y_{i..}$: mean expression value of i^{th} cell cluster, and $Y_{...}$ be the over-all mean; N: Number of cell clusters and M_i: Number of cells in i^{th} cell cluster.

Then, Total Sum of Squares (TSS) can be expressed as:

$$TSS = \sum_{i=1}^{N} \sum_{j=1}^{M_i} \left(Y_{ij.} - Y_{...}\right)^2$$

$$= \sum_{i=1}^{N} \sum_{j=1}^{M_i} \left(Y_{ij.} - Y_{i..}\right)^2 + \sum_{i=1}^{N} M_i \left(Y_{i..} - Y_{...}\right)^2$$

$$= WSS + BSS \tag{33}$$

where, WSS: Within cluster Sum of Squares, BSS: Between cluster Sum of Squares

Now, the proposed index can be given as:

$$r_h = \frac{WSS}{BSS} \tag{34}$$

For different values of number of clusters (h) in the scRNA-seq data, the r-measure is computed. The h value which provides the maximum value of r, can be chosen as the estimator for optimum cell clusters for the observed scRNA-seq data. The optimum value of "h" can be obtained through graphically by plotting h vs. r_h and choosing the point in x-axis where the curve flattens.

In this study, we determined cell types through the following steps:

Step 1 Determine the optimum number of cell cluster: first, it determines the optimum number of cell clusters needed to group the cells present in the data through the developed algorithm.

Step 2 Clustering of the cells: once the optimum number of cell cluster is determined then, the cells are grouped to that many clusters through K-means clustering using the observed UMI counts. Both the steps are implemented in *optimcluster* function in SwarnSeq R package. The flowchart for this process is given in Figure S4.

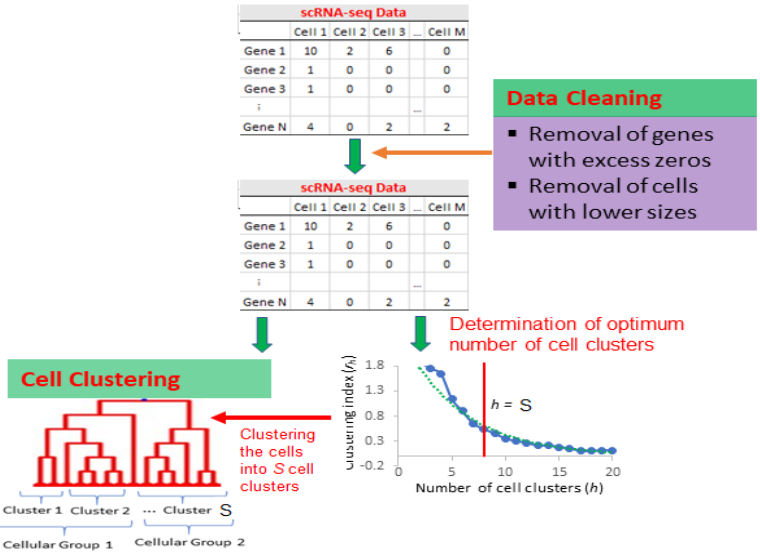


Figure S4. Flowchart for determining optimum number of cell clusters and detection of cell types for DE analysis.

For instance, we applied the above discussed technique to Lung cancer UMI data (GSE111108) with 17326 genes over 2126 cells. Through executing the *optimcluster* function implemented in SwarnSeq R package, the number of optimum cell clusters was determined based on the proposed method described in Supplementary Document S9. The result is shown Figure S5.

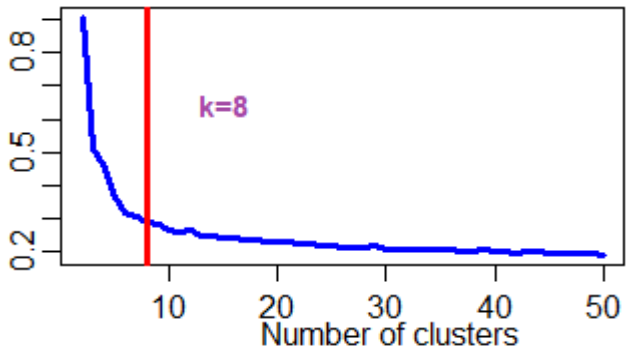


Figure S5. Determination of optimum number of cell clusters.

Here, we set the seed value at 110 and, found that the 2126 cells are clustered into 8 optimum cell clusters. Further, for Differential Expression (DE) analysis, we took cell cluster 3 (cells: 987) as group 2 and remaining cell clusters (cells: 1139) as group 1, as shown in Table S5.

Table S5. Cellular group and cellular clusters information.

Cell clusters	1	2	4	5	6	7	8	3
Number of cells	158	59	373	50	156	119	224	987
	Group 1 (1139)							Group 2 (987)

The algorithm for determining the optimum number of cell clusters was applied to all the 10 real scRNA-seq UMI counts datasets and the results are shown in Figure S3. If the cellular group/type information is not known *a-priori*, then the cellular groups 1 and 2 are determined manually by taking one cluster (mostly with highest number of cells) as group 1 (*i.e.*, cell type 1) and the remaining cell clusters as group 2. The results from the cell type detection in the 10 real datasets is summarized in Table S6. Then this cell group information is used in all the tested methods for DE analysis of single-cell data.

Table S6. Pre-processing of scRNA-seq data and number of optimum cell clusters.

Sl. No.	Accession	#Cells*	#Genes*	Seed value	#Optimum cluster	#Cells in G1	#Cells in G2
01	GSE111108	2126	17326	1712	8	987	1139
02	GSE77288	576	15955	108	10	288	288
03	GSE109999	383	13054	110	9	203	180
04	GSE115469	5466	17316	222	16	1852	3614
05	GSE29087	92	11436	110	8	48	44
06	GSE53638	1835	14863	222	11	1245	590
07	GSE53638	2268	15637	555	10	1245	1023
08	GSE53638	1613	15015	2222	8	590	1023
09	GSE65525	1481	23971	108	11	683	798
10	GSE92495	1453	15524	208	8	537	916

#Cells*, #Genes* are the number of cells and genes in the scRNA-seq data after data pre-processing.

10. Supplementary Document S10: Selection of reference genes

To capture the real distributional of the single-cell data, we assessed the performance of the tested methods including the proposed SwarnSeq method on 10 different publicly available real scRNA-seq UMI counts datasets having cells ranging from few hundreds to several thousands (Table S6). These datasets are briefly described in Document S8. Therefore, such an approach of performance evaluation was capable of taking the biological ground truth of real DE genes for the benchmarking of performance the tested methods including the proposed one. However, it is more complicated to obtain the available list of real true DE genes for each of the considered dataset. To tackle this problem, we selected few known biological relevant datasets, such as Microarrays, bulk RNA-seq, to get the list of reference DE genes for which single-cell RNA-seq datasets are available for the same cell-lines. The detail procedure for obtaining the reference genes for each of the 10 considered real datasets to validate the performance of the tested methods is described as below.

***Mouse cell data* (GSE29087):** For this real scRNA-seq dataset, we obtained the reference DE genes (the same cell lines for which single-cell dataset was generated) from the biologically relevant Microarray dataset. Further, we collected the Microarray datasets available for the mouse Embryonic Stem (ES) cell types (class label: 1), and Embryonic Fibroblasts (EF) cell types (class label: -1). This dataset is publicly available at http://carlosibanezlab.se//Data/Moliner_CELfiles.zip [33]. We used the Support Vector Machine-Recursive Feature Elimination (SVM-RFE) [34] to select the genes, which are differentially expressed between the ES and EF cell lines from the Microarray expression data and assumed these selected differentially expressed genes as reference genes for benchmarking the tested methods. The SVM-RFE is a machine learning algorithm, widely used for informative gene selection in Microarray data analysis and has superior performance compared to other gene selection methods, as established in the previous study [35–40]. Through the SVM-RFE, the top-ranked 3000 genes were selected and considered as the reference genes for performance analysis of the scRNA-seq DE methods.

Pluripotent stem cell data (GSE77288): For this single-cell dataset, we obtained the reference DE genes (for the NA19101 and NA19239 cell lines for which single-cell dataset is available) from the relevant bulk RNA-seq data. This bulk RNA-seq data, generated for the NA19101 and NA19239 cell lines, is publicly available at https://github.com/jdblischak/singleCellSeq. Here, we used the EBSeq method [41] to select the differentially expressed between the NA19101 and NA19239 cell lines from the bulk RNA-seq and assumed these selected differentially expressed genes as reference genes for benchmarking the tested methods. The rationale behind using EBSeq technique is that it is based efficient Bayesian algorithm and it has superior performance for detecting DE genes from bulk RNA-seq data, as established in our previous study [48]. Based

on the computed posterior probability values (through EBSeq), we selected the top ranked 3000 genes as reference genes for benchmarking the performance of the single-cell methods including the SwarnSeq.

Other scRNA-seq datasets: For the remaining eight single-cell datasets, we selected the reference genes from these data itself. Hence, to obtain the list of reference genes, we used Fold Change (FC) criterion (*i.e.,* ratios of mean expressions of genes over the two cellular groups). The detailed descriptions about the identification of cellular groups for each of the remaining eight datasets including Lung cancer data, Mouse blood cell data, Liver cell data, Adipose stem/stromal cell data, Mouse embryonic cell data, Human Embryonic Kidney (HEK) cell data is described in Supplementary Document S8. For each of the 10 datasets, we selected the top 3000 genes based on the FC criterion as reference gene lists. In is important to note that the selection of reference genes from these datasets is completely independent of the detection of DE genes through the tested DE methods. The selection of reference genes through FC criterion provides an equal platform for the comparative analysis of the tested method on each of the datasets [48].

11. <u>Document S11</u>: Analytical steps of the proposed SwarnSeq method and Layout of the comparative study with the existing ones on real single-cell datasets.

The proposed statistical approach, *i.e.*, SwarnSeq, is described in the section "2.4 *Proposed SwarnSeq Method"* of the main text. Further, the major analytical steps of the proposed SwarnSeq method are given in Figure S6.

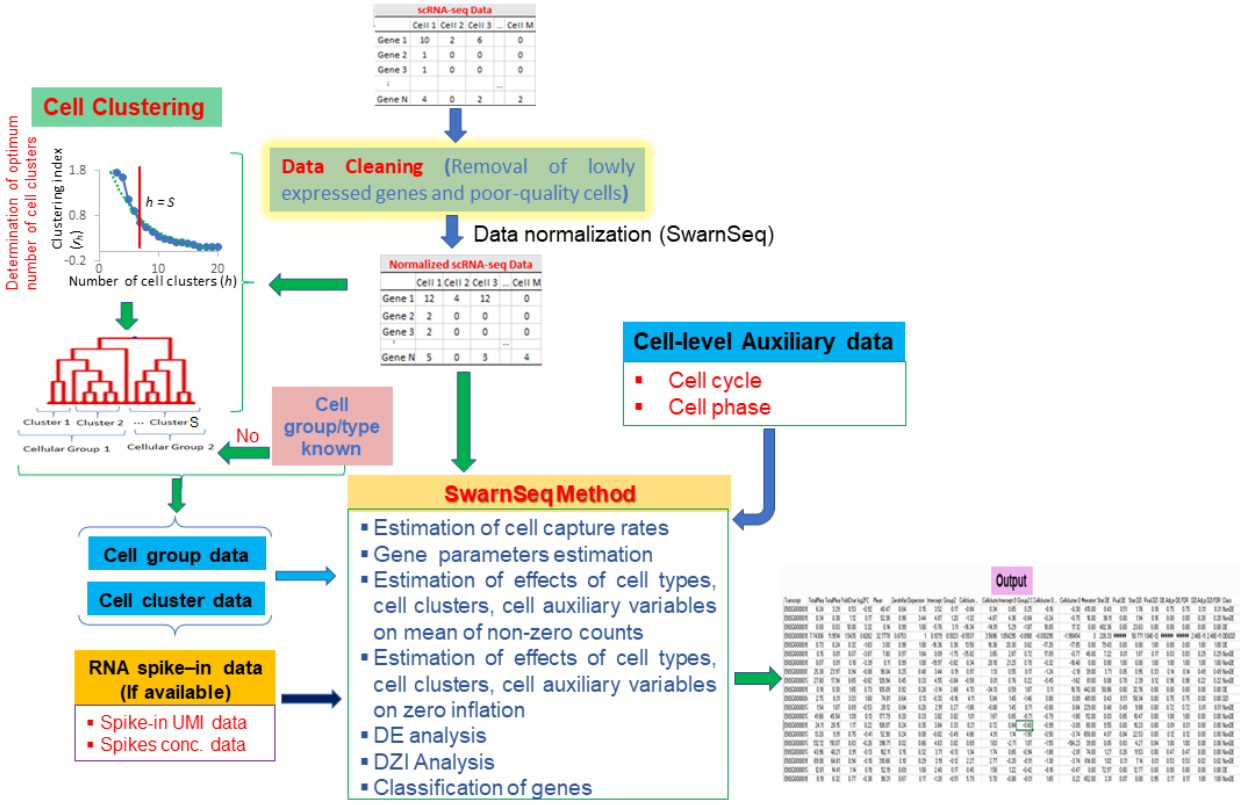


Figure S6. Analytical steps of the proposed SwarnSeq method for analysis of single-cell data.

Layout of the comparative study:

The single-cell datasets and their respective comparison designs were used to detect the DE genes through each of the 12 tested methods including the proposed SwarnSeq approach (Table S5). For instance, the Islam data [11], the experimental design involves DE analysis of genes between 48 mouse embryonic stem cells and 44 mouse embryonic fibroblast cells. In other words, we selected the DE gene sets of sizes 500, 1000, …, 3000 through the tested methods from the GSE29078 data (Table S6). The availability of the proposed and existing methods is briefly described in Table S7. Further, the parameter settings used in each of the tested methods including the proposed SwarnSeq approach is given in Table S8. Then, the performance metrics such as TP, FP, PPR, TPR, FPR, ACC, and F1, were computed for each of the methods by comparing the detected DE genes with the reference genes for each dataset. The expressions of the performance metrics are given in Eq. 28-34 of the main text. The layout of the undertaken comparative study is briefly described in Figure S7.

Table S7. Availability of the methods used in the comparative study.

SN.	Methods	Year	R package	Availability	Ref.
01	DESEQ	2010	DESeq	https://bioconductor.org/packages/release/bioc/html/DESeq.html	[42]
02	DEGSEQ	2009	DEGseq	https://bioconductor.org/packages/release/bioc/html/DEGseq.html	[43]
03	LIMMA	2002	limma	https://bioconductor.org/packages/release/bioc/html/limma.html	[44]
04	EDGER-LRT	2010	edgeR	https://bioconductor.org/packages/release/bioc/html/edgeR.html	[3]
05	MONOCLE	2017	monocle	www.bioconductor.org/packages/release/bioc/html/monocle.html	[45]
06	MAST	2015	MAST	https://github.com/RGLab/MAST	[46]
07	SCDD	2016	scDD	https://www.bioconductor.org/packages/release/bioc/html/scDD.html	[47]
08	DESINGLE	2018	DEsingle	https://www.bioconductor.org/packages/release/bioc/html/DEsingle.html	[7]

09	DECENT	2019	DECENT	https://github.com/cz-ye/DECENT	[8]
10	BPSC	2016	BPSC	https://github.com/nghiavtr/BPSC	[48]
11	NODES	2016	NODES	https://goo.gl/Ndx07M	[49]
12	SwarnSeq	2021	SwarnSeq	https://github.com/sam-uofl/SwarnSeq	#

method reported in this study

Table S8. DE tools/R packages along with their parameter's settings used in this study.

Sl. No.	R package	Version	Parameters	Utility	DE test stat	Sel. criteria	Ref.
01	DEGSeq	1.42.0	method="LRT"	Bulk cell	LRT	Adj. p-value	[43])
02	edgeR	3.30.3	Default settings	Bulk cell	LRT	Adj. p-value	[25]
03	DESeq2	1.39.0	test="LRT", reduced= ~ 1	Bulk cell	LRT	FDR	[2]
04	limma	3.44.1	Default settings	Bulk cell	Wald t-test	p-value	[50]
05	Monocle	2.16.0	UMI = T, n.cores = 2	Single cell	LRT	Adj. p-value	[51,52]
06	MAST	1.14.0	Default settings	Single cell	LRT	Adj. p-value	[46]
07	BPSC	0.99.2	Default settings	Single cell	LRT	Adj. p-value	[48]
08	ScDD	2.16.0	UMI = T, n.cores = 2	Single cell	Bayesian Stat.	Adj. p-value	[47]
09	NODES	0.0.0.9010	Default settings	Single cell	Wilcoxon test stat.	Adj. pvalue	[49]
10	DEsingle	1.8.2	Default settings	Single cell	LRT	FDR	[7]
11	DECENT	1.1.0	CE.range = c(0.02, 0.1), n.cores =7	Single cell	LRT	Adj. pvalue	[8]
12	*SwarnSeq	0.1.0	CE.range=c(0.1, 0.4), norm.method = "TMM", CellAuxil=NULL, maxit=1000, eps=1e-10, muoffset=NULL, phioffset=NULL, weights=NULL, p.adjust.method="BH"	Single cell	LRT	FDR	Proposed

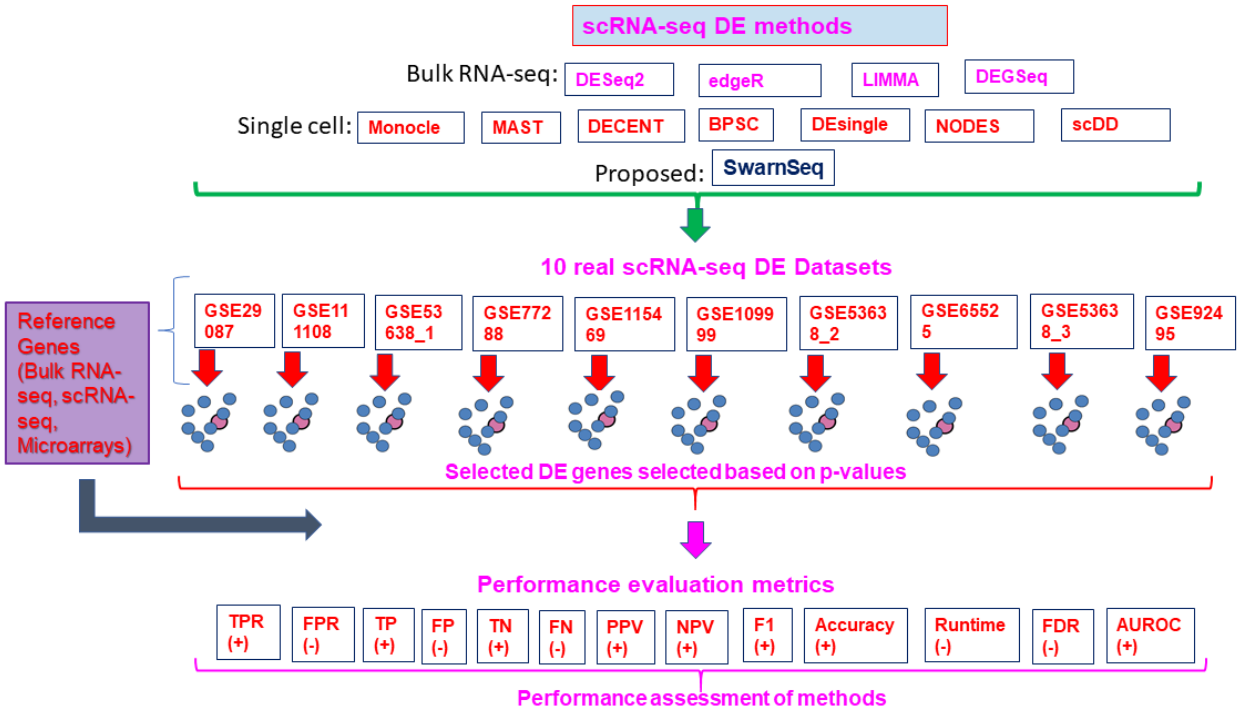


Figure S7. Layout of the comparative study used for benchmarking of the proposed SwarnSeq method.

12. Supplementary Document S12: Performance analysis of proposed SwarnSeq method based on runtime criterion

In order to develop a novel method for current scRNA-seq data, it is important to consider the computational processing speed for the analysis of large-scale scRNA-seq data. From single-cell data analysis point of view, computational processing time is important. It is because that the current single-cell data especially generated by the droplet-based system (*e.g.*, 10x genomics) have several thousand(s) of cells compared to datasets generated from bulk RNA-seq studies. Therefore, we considered the Liver cell experimental dataset, described in Document S8, (GSE115469 data having thousand(s) of genes over thousand(s) of cells) with 17316 of genes over 5466 cells to evaluate the performance of the tested methods, including the proposed SwarnSeq method, based on runtime criterion. Here, the runtime refers to the amount of

computational time required to get the informative Differentially Expressed Genes (DEG) of particular size through providing the gene expression UMI data as input to the R functions of the respective methods. To measure this, we ran the code written in R (*v* 4.0.2) for each of the tested methods by following the instructions and recommendations of their respective R software packages. The required average CPU time (over 10 runs for each program) was observed for each of the methods for analyzing the large experimental the Liver cell scRNA-seq (GSE115469) dataset. The detail descriptions about the dataset can be found in Supplementary Document S8. All these analyses were performed on a 10-core 32 GB DELL PC with Windows 10 OS and Intel(R) Core (TM) i3-6100U CPU clock rate as 2.93 GHz.

The runtime-based scores of the tested methods can be computed through the following procedure. Let, p_r be the rank of r^{th} DE method based on the computational time required to get the DE gene set of size 1000 from the considered liver cell data. The rank score (based on computational processing time) for r^{th} DE method can be calculated by using:

$$score_r = f(p_r) = \frac{1+S-p_r}{S} \tag{35}$$

where, $score_r$: rank score for r^{th} DE method, S: total number of methods used in the analysis. The result from the runtime-based analysis of scRNA-seq DE methods is shown in Table S9.

Table S9. Runtime based analysis of the tested scRNA-seq DE methods.

Sl. No.	Methods	Run Time	Ranks (p_r)	Rank score ($score_r$)
1	BPSC	35 Min.	5	0.67
2	DECENT	12 hr. 15 Min.	12	0.08
3	DEGSeq	20 Min.	2	0.92
4	DESeq2	30 Min.	3	0.83
5	DEsingle	9 hr. 20 Min	11	0.17
6	edgeR	33 Min.	4	0.75
7	Limma	18 Min.	1	1.00
8	MAST	1 hr. 10 Min.	8	0.42
9	Monocle	1 hr. 5 Min.	7	0.50
10	scDD	45 Min.	6	0.58
11	NODES	1 hr. 15 Min.	9	0.33

12	SwarnSeq	6 hr. 20 Min.	10	0.25

hr. Hours requires to analyze the data; Min.: Minutes required

The findings indicated that zero inflated model based single-cell methods, such as DECENT and DEsingle required more computational processing time as compared to other methods and thus considered as computationally inefficient for analyzing large scale single-cell data (Table S9). But the developed SwarnSeq method required lesser computational runtime, *i.e.*, computationally efficient, compared to other zero inflation model-based methods including DEsingle and DECENT. However, the SwarnSeq method required more runtime compared to bulk RNA-seq methods, such as DESeq2, edgeR, Limma and other single-cell methods, such as BPSC, scDD, MAST, Monocle, NODES (Table S9). The reason may be attributed as it is based on an iterative Expected Maximization algorithm to estimate the gene specific parameters (*i.e.*, mean, dispersion, zero inflation) including the effects of the cell types, cell clusters, cell level auxiliary variables on the mean and zero inflation parameters of the genes. Though the developed SwarnSeq method required more runtime to analyze the large single-cell data compared to bulk RNA-seq methods, but it has much better performance in terms of detecting genuine DE genes compared to other tested methods. Besides, it is capable of handling/incorporating the cell level auxiliary data such as cell cycle, cell phase, etc. in the gene specific model development process. Apart from the DE analysis, it does other type analysis such as estimation of cell capture rates, determining the optimum number of cell clusters, differential zero inflation analysis, and classification of detected influential genes.

13. Supplementary Document S13: Relation among the genes and cells parameters estimated through the SwarnSeq model.

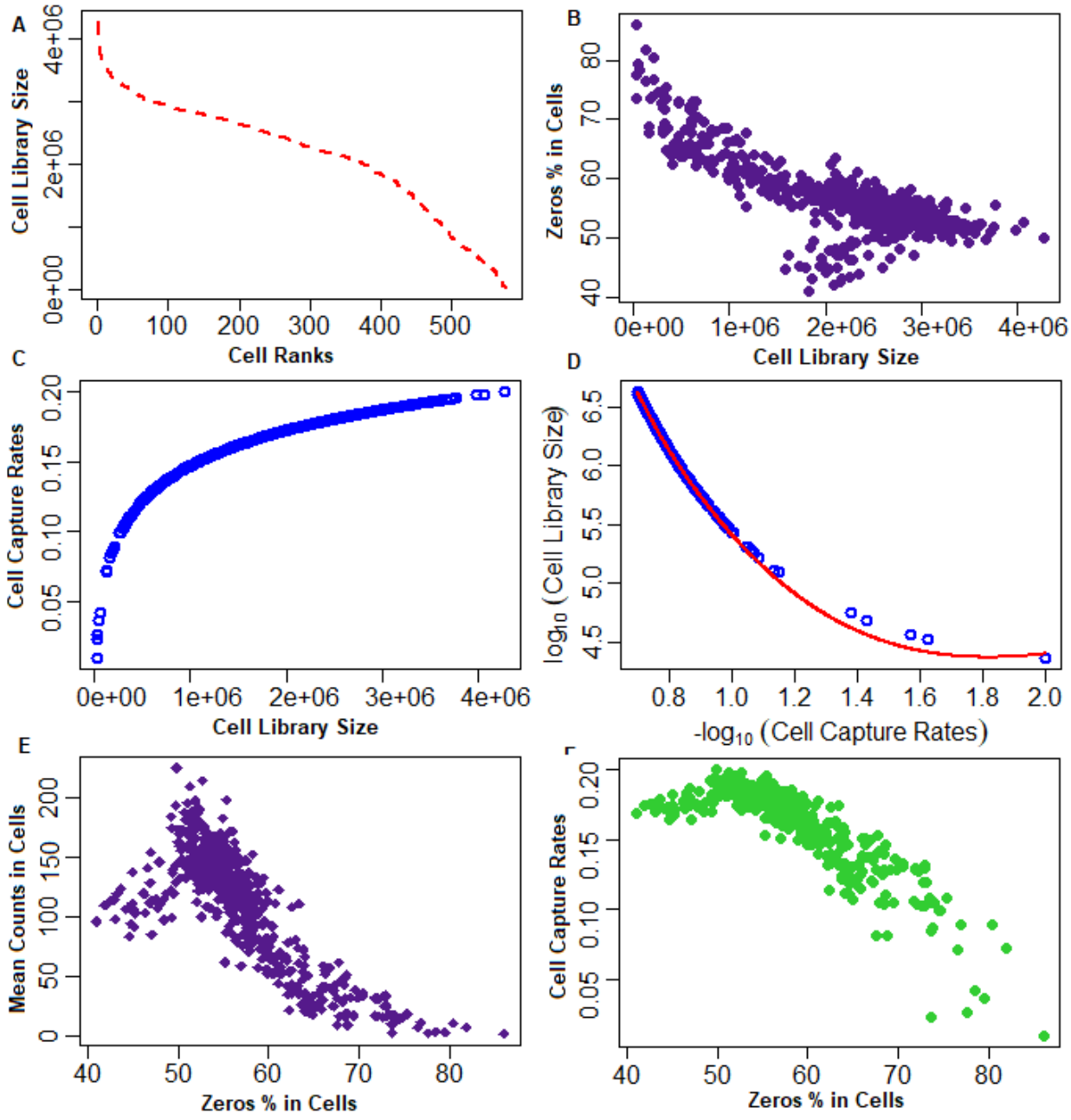


Figure S8. Characteristics of single cells in GSE77288 scRNA-seq data derived from Human induced pluripotent stem cell lines. We have taken the scRNA-seq data over 576 cells

belonging to two individual cell types, such as NA19101 (288 cells), and NA19239 (288 cells). **(A) Cell library size distribution over cells.** In X-axis, the ranks of the cells are shown, and Y-axis shows the library sizes of the cells. Here, the underlying distribution is S-shaped. **(B) Zero inflation vs. Cell library sizes plot.** Relationship between the cell library sizes and the zero counts in cell are shown. Here, X-axis represents the cell library sizes and Y-axis represents the zero counts percentages in the cells. It can be shown that every cell has higher zero counts (> 40%) as expression of genes, due to the availability of lower concentration of mRNA molecules. Further, the cell library sizes are inversely proportional to the zero percentage in cells. In other words, the cells with higher library sizes contain lesser percentages of zeros as expression of genes and *vice-versa*. **(C) Cell library size vs. Cell capture rates plot.** The relation between the library sizes and the capture efficiencies of the cells is shown. The library sizes and mRNA molecules capture rates of the cells are represented in X-axis and Y-axis, respectively. The cell capture rates are estimated through SwarnSeq model. Here, the library sizes are directly proportional to the capture rate of cells. In other words, the cells with higher capture rates have higher library sizes and *vice-versa.* **(D) Cell library size vs Cell capture efficiency plot.** The log-transformation of cell library sizes is plotted with log transformation of cell capture rates and a curve is fitted shown in red color. **(E) Mean non-zero counts of cells *vs.* zeros percentage in cells plot.** The relation between the mean of non-zero counts and zero percentages in cells is shown. X-axis shows the zero percentages in cells and Y-axis shows the mean of non-zero counts expression of genes in cell. Here, the relation between the zero percentages and mean of non-zero counts in cells is reciprocal. In other words, the cells with higher mean non-zero counts contain lesser percentages of zeros as expression of genes and *vice-versa.* **(F) Zeros percentage in cell vs. Cell capture rates plot.** The relation between the zeros' present in the cell with the cell capture rates is shown. X-axis represents the Zeros percentage in cell and Y-axis represents the cell capture rates. Here, the relation is inversely proportional, means cells with higher capture rates have lesser zeros as counts expression the cell. In other words, cells with higher capture rates can ably capture mRNA molecules leading to fewer dropout events.

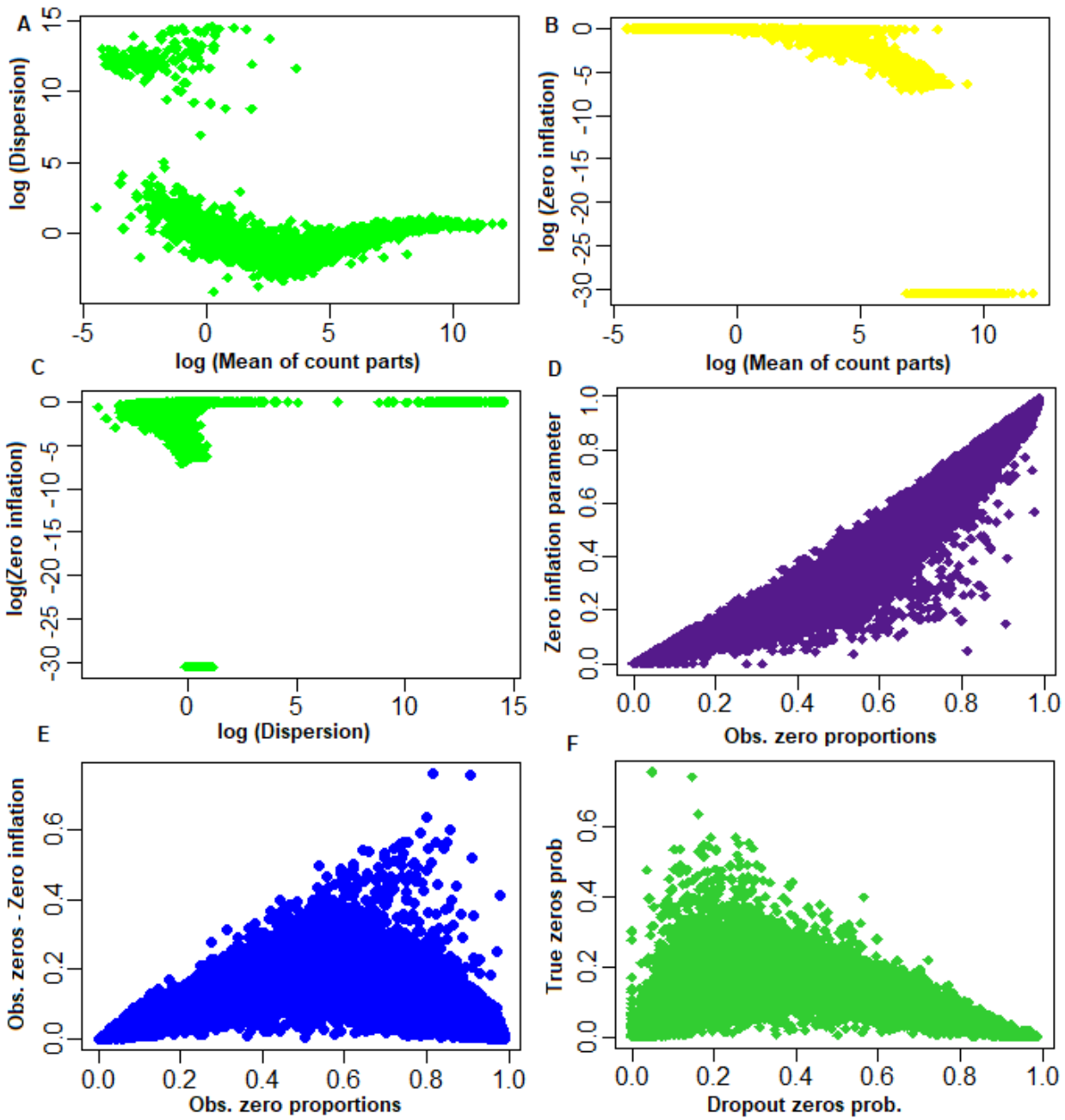


Figure S9. Estimated values of SwarnSeq model's parameters for GSE77288 data from Human pluripotent stem cell lines. We have taken the scRNA-seq data over 576 cells belonging to two individual cell types, such as NA19101 (288 cells), and NA19239 (288 cells). Here, the parameters of the SwarnSeq model (shown in Eq. 9-13) are estimated through MLE (EM) method. (A)

Mean of count parts vs. Dispersion plot. The relation between the mean count parts (NB part) with dispersion parameters of the SwarnSeq model is shown. X-axis represents the log transformed value of mean count parts and Y-axis represents the log transformed value of the estimated dispersion parameter. (B) **Mean of count parts vs. Zero inflation probability plot.** The relation between the mean count parts (NB part) with the zero inflation probability parameters of the SwarnSeq model is shown. X-axis represents the log transformed value of mean count parts and Y-axis represents the log transformed value of the estimated zero inflation probability parameter. (C) **Dispersion vs. Zero inflation parameter plot**. The relation between the dispersion and zero inflation probability parameters of the SwarnSeq model is shown. X-axis represents the log transformed value of dispersion and Y-axis represents the log transformed value of the estimated zero inflation probability parameter. (D) **Observed zero proportions vs. Estimated zero inflation probability plot**. The relation between the Zero inflation probability parameter of the SwarnSeq model with the observed zero proportions present in the scRNA-seq data is shown. X-axis represents the observed zero proportions for genes in the Human pluripotent stem cell data and Y-axis represents the zero-inflation probability parameter. (E) **Observed zero proportions vs. (Observed zero - zero inflation) probability plot**. The relation between the observed zero proportions presents in the scRNA-seq data and the (Observed zero - zero inflation) probability is shown. X-axis represents the observed zero proportions for genes in the Human pluripotent stem cell data and Y-axis represents the difference between the observed zero proportion and zero-inflation probability parameter. **(F) Dropout zeros vs. True Zeros plot.** The relation between the dropout zeros, *i.e.*, excess zeros present in scRNA-seq data modeled through Dirac's delta function in Eq. 3, and the True zeros from the NB model (Eq. 1) is shown. X-axis represents the estimated values of dropout zeros probability and Y-axis represents the estimated values of true zeros probability parameter.

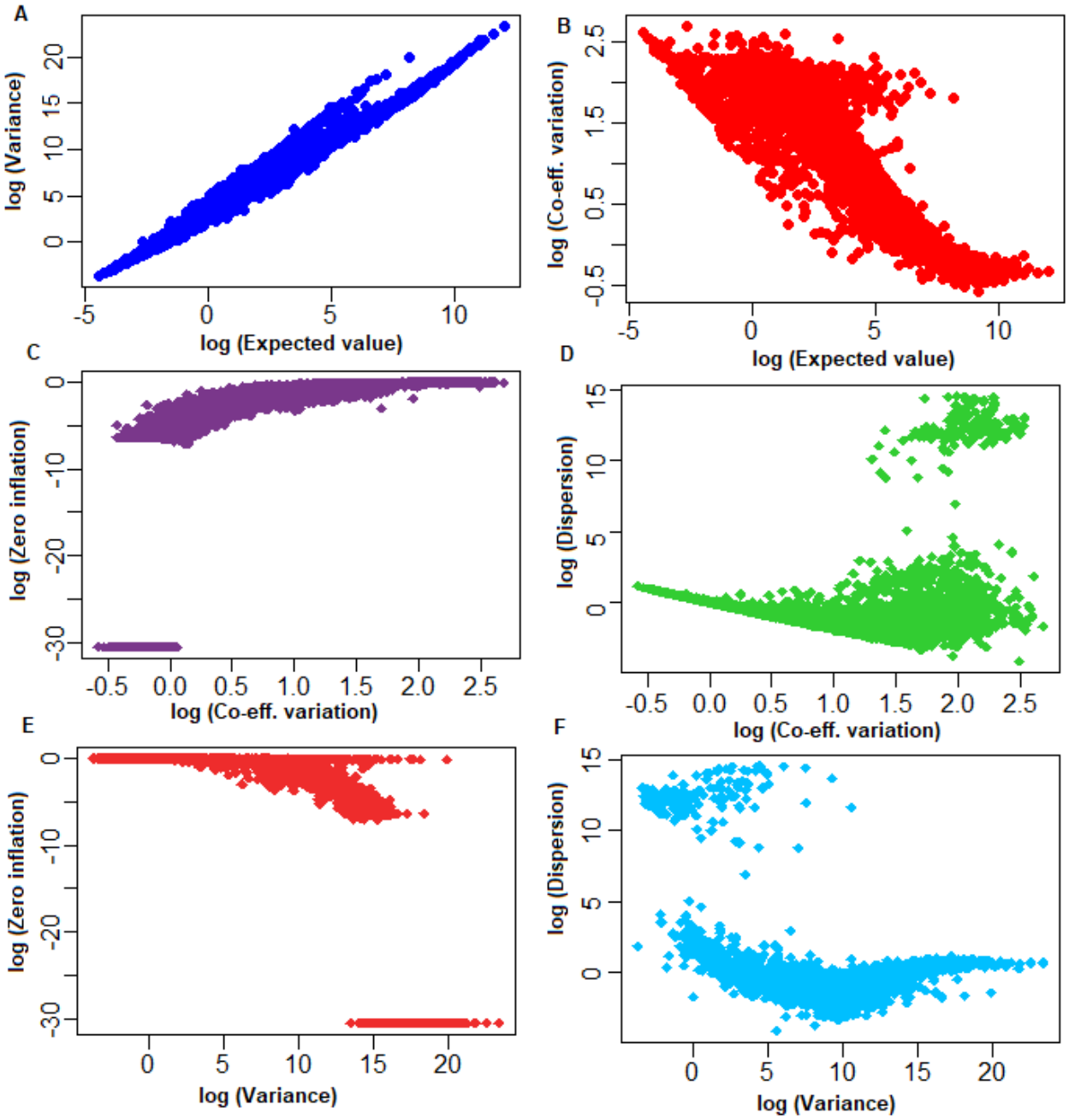


Figure S10. Relation among the expected values, variances, and co-efficient variations of the genes estimated through the SwarnSeq approach. We have taken the GSE77288 data over 576 cells belonging to two individual cell types, such as NA19101 (288 cells), and

NA19239 (288 cells) from Human pluripotent stem cell lines. **(A) Expected values vs. Variances of the UMI counts of genes.** The relation between the expected values and variances of the observed UMI counts of genes estimated through SwarnSeq model given in Eq. 35, 36 is shown. The expressions for the expected values and variances are shown in Eq. 35 and 36. X-axis represents the log transformed expected values and Y-axis represents the log transformed variances of observed UMI counts of genes. **(B) Estimated expected values vs. co-efficient of variations from the SwarnSeq model.** The relation between the estimated expected values and co-efficient of variations of the SwarnSeq model is shown. The expressions for the expected value and co-efficient of variation variance are shown in Eq. 35 and 36, respectively. X-axis represents the log transformed expected values and Y-axis represents the log transformed values of co-efficient of variation. **(C) Co-efficient of Variation vs. Zero inflation plot.** The relation between the estimated co-efficient of variations and zero inflation probability parameters of the SwarnSeq model is shown. X-axis represents the log transformed Co-efficient of Variation and Y-axis represents the log transformed value of Zero inflation probability. **(D) Co-efficient of Variation vs. Dispersion plot.** The relation between the Co-efficient of Variation and the Dispersion parameter of the SwarnSeq model is shown. X-axis represents the log transformed co-efficient of variation and Y-axis represents the log transformed estimated values of dispersion parameter. **(E) Variance vs. Zero inflation plot.** The relation between the Variance and the estimated value Zero inflation probability parameter, of the SwarnSeq model is shown. X-axis represents the log transformed Variance values and Y-axis represents the log transformed estimated values of Zero inflation probability. **(F) Variance vs. Dispersion plot.** The relation between the Variance and the dispersion parameter, of the SwarnSeq model is shown. X-axis represents the log transformed Variance and Y-axis represents the log transformed value of the dispersion.

14. Supplementary Document S14: A brief tutorial of SwarnSeq R Package

SwarnSeq is a package for the differential expression analysis of Unique Molecular Identifier (UMI) count gene expression data, that is, obtained from single cell RNA sequencing (scRNA-seq) technologies. It is especially designed for differential expression analyses of scRNA-Seq or bulk RNA-seq data. It also performs differential zero-inflation analysis of genes of scRNA-seq data. Further, it classifies the influential genes into various categories based on their differential expression and zero inflation across different cell types. SwarnSeq implements novel statistical methods based on the Zero Inflated Negative Binomial (ZINB) distribution as a model for count variability, including likelihood ratio tests, and generalized linear models. This method is capable of incorporating the molecular capturing process through a binomial model while modeling the observed UMI counts. Here, the gene specific ZINB parameters are estimated through the expected-maximization algorithm the generalized linear model framework. SwarnSeq method is capable of performing various analyses, such as differential expression and zero inflation, through incorporating the cell cluster and cell level auxiliary information in the parametric modeling of gene specific parameters. The graphical layout of the operational procedure for the SwarnSeq R package is given in Figure S11.

Availability: https://github.com/sam-uofl/SwarnSeq

Dependent packages: stats, MASS, Matrix, edgeR, foreach

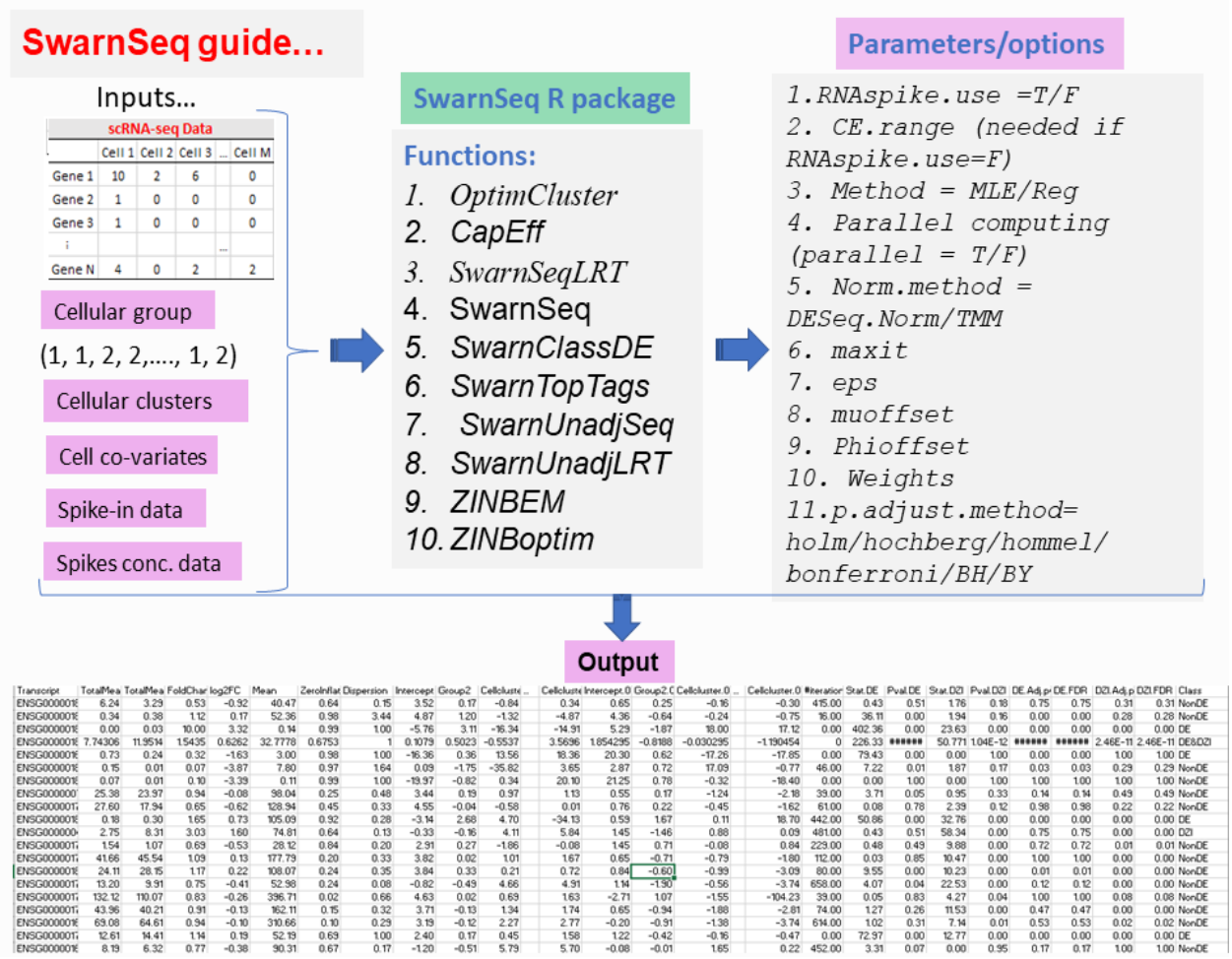


Figure S11. Graphical layout for the SwarnSeq R software package.

A. Estimation of cell capture rates

Function

> CapEff

1.1* Description**: The *CapEff* function estimates the capture efficiencies of the cells from the single-cell RNA-seq studies. It takes input the ERCC spike-in transcript and molecular concentration data, if available or count expression data, if spike-ins are not available. The detail method for estimation of cell capture rates is described in section "2.5 ***Estimation of capture rates parameter" of the main text.

Usage: `CapEff(CountData, CE.range, RNAspike.use, spikes, spike.conc, method)`

Arguments

Inputs	Descriptions
`CountData`	Observed count data matrix for genes, rows represent genes, columns represent cells.
`CE.range`	Two-element vector representing the lower and upper limits for the estimated range of capture efficiencies (ONLY needed if RNAspike.use =FALSE, default [0.1, 0.40]).
`RNAspike.use`	Logical value indicating TRUE/FALSE, if TRUE, spikes and spike.conc information must be provided.
`spikes`	Observed count matrix for spike-in transcripts, rows represent spike-in transcripts, columns represent cells. Only needed if RNAspike.use =TRUE).
`spike.conc`	Vector of theoretical count for each spike-in transcript in one cell (ONLY needed if RNAspike.use = TRUE).
`method`	Character representing the methods to be used for computation of capture efficiencies for cells.

Output: Returns a vector of estimated capture efficiencies for cells given in the scRNA-seq data.

B. Determination of optimum number of cell clusters

Function

> optimcluster

Description: The *optimcluster* function decides the number of optimum cell clusters for the given experimental scRNA-seq data. The detail methods for deciding the optimum number of cell clusters is given in Supplementary Document S9.

Usage: `optimcluster(CountData, n, seed, Threshold, plot = TRUE)`

Arguments

Inputs	Description
`CountData`	Observed count data matrix for genes, rows represent genes, columns represent cells.
`n`	Maximum value for number of cell clusters.
`seed`	value for random cluster generation.
`Threshold`	Threshold value for deciding the optimum number of cell clusters.
`plot`	Logical variable taking value either TRUE or FALSE, default is FALSE.

Value: A list containing the clustering index, delta and the optimum cluster number.

Examples

```
##Load the test count data given for SwarnSeq.
counts <- matrix(rnbinom(2000, size=0.2, mu=3.4), 50)

results <- optimcluster(CountData=counts, n = 10, seed = 108, Threshold =
0.3, plot = FALSE)
```

C. Estimation of gene specific parameters, effects of cell clusters and cell level auxiliary variables on mean of non-zero counts and zero inflation parameters

Function:

> SwarnSeq

Description: The *SwarnSeq* function is used to estimate the parameters for genes between two specified groups of cells in a raw read (UMI) count matrix of single-cell RNA-seq (scRNA-seq) data. It takes a non-negative integer matrix of scRNA-seq raw read counts as input. So, the users should map the reads (obtained from sequencing libraries of the samples) to the corresponding genome and count the reads mapped to each gene according to the gene annotation to get the raw read counts matrix in advance.

Usage

```
SwarnSeq(
  CountData,
  RNAspike.use,
  spikes,
  spike.conc,
  CE.range,
  parallel,
  norm.method,
  group,
  CellCluster,
  CellAuxil,
  maxit,
  eps,
  muoffset,
  phioffset,
  weights
)
```

Arguments

Inputs	Descriptions
CountData	Observed count data matrix for genes, rows represent genes, columns represent cells.
RNAspike.use	Logical value indicating TRUE/FALSE, if TRUE, spikes and spike.conc information must be provided.
spikes	Observed count matrix for spike-in transcripts, rows represent spike-in transcripts, columns represent cells. Only needed if RNAspike.use =TRUE).
spike.conc	Vector of theoretical count for each spike-in transcript in one cell (ONLY needed if

	RNAspike.use = TRUE).
CE.range	Two-element vector representing the lower and upper limits for the estimated range of capture efficiencies (ONLY needed if RNAspike.use =FALSE, default [0.1, 0.40]).
parallel	If FALSE (default), no parallel computation is used; if TRUE, parallel computation is performed.
norm.method	Method for normalizing the scRNA-seq count expression data, either 'DESeq2' (maximum likelihood, Ye et al., 2017) or 'TMM' (Robinson et al., 2010).
group	Vector which specifies the membership of the cells, *i.e.*, two groups to be compared, corresponding to the columns in the count data matrix.
CellCluster	Vector which specifies the cluster memberships of the cells, *i.e.*, each entry represents memberships of the columns of the count data matrix.
CellAuxil	Vector of cell level auxiliary information, corresponding to the columns in the counts matrix, default is NULL.
maxit	Maximum number of iterations for Expected-Maximization (EM) algorithm.
eps	Convergennce criteria for EM algorithm.
muoffset	Offset parameter for mean (mu) parameter, default is NULL.
phioffset	Offset parameter for zero inflation (phi) parameter, default is NULL.
weights	Observation wise weights for the cells, default is unity vector.

Output

A data frame containing the parameters from the EM algorithm for each gene, rows are genes and columns contain the following items:

- 1 totalMean_1, totalMean_2, Adj_Mean_1, Adj_Mean_2, AdjNormMean_1, AdjNormMean_2 are the total mean, adjusted mean, normalized mean for cellular groups 1 and 2 respectively.
- 2 Mean, ZeroInflation, Dispersion are the MLE of the parameters of whole cellular population characetrized by the ZINB model.
- 3 Intercept, Group2 are the co-efficients of the intercept and group effect on mean of non-zero counts.
- 4 Cellcluster 2, Cellcluster 3, .., Cellcluster M are the effects of cell clusters on mean of non-zero counts.
- 5 CellAuxil 2, CellAuxil 3, ..., CellAuxil N are the effects of cell-level auxiliary information on mean of non-zero counts (if included in the model).
- 6 Intercept.0 Group2.0 are the co-efficients of the intercept and group effect on zero inflation probability.
- 7 Cellcluster 2.0, Cellcluster 3.0, .., Cellcluster M.0 are the effects of cell clusters on zero-inflation probability.
- 8 CellAuxil 2.0, CellAuxil 3.0, ..., CellAuxil N.0 are the effects of cell-level auxiliary information on zero-inflation probability (if included in the model).
- 9 #iteration number of iterations required for convergence for each gene.

Examples

```
#Load the test count data, spike-in counts and spike-in concentration data
for SwarnSeq.
data(TestData)
counts <- TestData$CountData
Spikes <- TestData$SpikeCounts
SpikeConc <- TestData$SpikeConc

#specifying the group information, the group 1 and 2 have two hundred cells
each.
group <- c(rep(1, 200), rep(2, 200))
#Specifying the cluster memberships of the cells in columns of countData.
cellcluster <- c(rep(1, 60), rep(2, 40), rep(3, 50),
                    rep(4, 50), rep(5, 30),
                    rep(6, 90),
                    rep(7, 80))

#parameters from EM algorithm for each gene (when Spike-in information
available).
#results <- SwarnSeq(CountData=counts, RNAspike.use=TRUE, spikes=Spikes,
spike.conc=SpikeConc,
                        #parallel=FALSE, norm.method="TMM", group=group,
CellCluster=cellcluster,
                            #CellAuxil=NULL, maxit=500, eps=1e-10,
                            #muoffset=NULL, phioffset=NULL, weights=NULL)

#When Spike-in information not available.
#results <- SwarnSeq(CountData=counts, RNAspike.use=FALSE, CE.range=c(0.1,
0.4)
                        #parallel=FALSE, norm.method="TMM", group=group,
CellCluster=cellcluster,
                            #CellAuxil=NULL, maxit=500, eps=1e-10,
                            #muoffset=NULL, phioffset=NULL, weights=NULL)
```

D. Differential Expression analysis when observed UMI counts are adjusted with cell molecular capture process

Function:

> SwarnSeqLRT

Description: The *SwarnSeqLRT* function is used to detect differentially expressed genes between two specified groups of cells in a raw UMI count matrix of scRNA-seq data. It takes a non-negative integer matrix of scRNA-seq raw read counts object as input. So, the users should map the reads (obtained from sequencing libraries of the samples) to the corresponding genome and count the reads mapped to each gene according to the gene annotation to get the raw read counts matrix in advance.

Usage:

```
SwarnSeqLRT(
  CountData,
  RNAspike.use,
  CE.range,
  spikes,
  spike.conc,
```

```
  parallel,
  norm.method,
  group,
  CellCluster,
  CellAuxil,
  maxit,
  eps,
  muoffset,
  phioffset,
  weights,
  p.adjust.method
)
```

Arguments

Inputs	Descriptions
CountData	Observed count data matrix for genes, rows represent genes, columns represent cells.
RNAspike.use	Logical value indicating TRUE/FALSE, if TRUE, spikes and spike.conc information must be provided.
CE.range	Two-element vector representing the lower and upper limits for the estimated range of capture efficiencies (ONLY needed if RNAspike.use = FALSE, default [0.1, 0.40]).
spikes	Observed count matrix for spike-in transcripts, rows represent spike-in transcripts, columns represent cells. Only needed if RNAspike.use =TRUE).
spike.conc	Vector of theoretical count for each spike-in transcript in one cell (ONLY needed if RNAspike.use = TRUE).
parallel	If FALSE (default), no parallel computation is used; if TRUE, parallel computation is performed.
norm.method	Method for normalizing the scRNA-seq count expression data, either 'DESeq2' (maximum likelihood, Ye et al., 2017) or 'TMM' (Robinson et al., 2010).
group	Vector which specifies the membership of the cells, *i.e.*, two groups to be compared, corresponding to the columns in the count data matrix.
CellCluster	Vector which specifies the cluster memberships of the cells, *i.e.*, each entry represents memberships of the columns of the count data matrix.
CellAuxil	Vector of cell level auxiliary information, corresponding to the columns in the counts matrix, default is NULL.
maxit	Maximum number of iterations for Expected-Maximization (EM) algorithm.
eps	Convergence criteria for EM algorithm.
muoffset	Offset parameter for mean (mu) parameter, default is NULL.
phioffset	Offset parameter for zero inflation (phi) parameter, default is NULL.
weights	Observation wise weights for the cells, default is unity vector.
p.adjust.method	Logical variable represents the method used for multiple hypothesis correction. It can be any value from ("holm", "hochberg", "hommel", "bonferroni", "BH", "BY").

Output

A data frame containing the results from differential expression analysis, rows are genes and columns contain the following items:

- 1 totalMean_1, totalMean_2, Adj_Mean_1, Adj_Mean_2, AdjNormMean_1, AdjNormMean_2 are the total mean, adjusted (capture efficiency) mean, normalized mean for cellular groups 1 and 2 respectively.
- 2 FoldChange, log2FC, AdjNormFC, log2AdjNormFC are the fold change, log fold change, and log normalized fold change for the genes respectively.
- 3 Stat.DE, Pval.DE, DE.Adj.pval, and DE.FDR are values of DE statistic, p-value, adjusted p-value, false discovery rate, obtained from DE analysis, for the genes.
- 4 Stat.DZI, Pval.DZI, DZI.Adj.pval, DZI.FDR are Differential Zero Inflation (DZI) statistic, DZI p-value, DZI adjusted p-value, DZI false discovery rate results obtained for each gene from DZI analysis.

Examples

```
#Load the test count data, spike-in counts and spike-in concentration data
for SwarnSeq.
data(TestData)
counts <- TestData$CountData
Spikes <- TestData$SpikeCounts
SpikeConc <- TestData$SpikeConc

#specifying the group information, the group 1 and 2 have two hundred cells
each.
group <- c(rep(1, 200), rep(2, 200))
#Specifying the cluster memberships of the cells in columns of countData.
cellcluster <- c(rep(1, 60), rep(2, 40), rep(3, 50),
                    rep(4, 50), rep(5, 30),
                    rep(6, 90),
                    rep(7, 80))

#Do not run
#parameters from EM algorithm for each gene.
#results <- SwarnSeqLRT(CountData=counts, RNAspike.use=TRUE, spikes=Spikes,
spike.conc=SpikeConc,
                       #parallel=FALSE, norm.method="TMM", group=group,
CellCluster=cellcluster,
                           #CellAuxil=NULL, maxit=500, eps=1e-10,
                           #muoffset=NULL, phioffset=NULL, weights=NULL,
p.adjust.method="hochberg")

#When Spike-in information not available.
#results <- SwarnSeqLRT(CountData=counts, RNAspike.use=FALSE, CE.range=c(0.1,
0.4)
                       #parallel=FALSE, norm.method="TMM", group=group,
CellCluster=cellcluster,
                           #CellAuxil=NULL, maxit=500, eps=1e-10,
                           #muoffset=NULL, phioffset=NULL, weights=NULL,
p.adjust.method ="hochberg")
```

E. Estimation of gene specific parameters, effects of cell clusters and cell level auxiliary variables on mean of non-zero counts and zero inflation parameters when observed UMI counts are not adjusted with cell molecular capture process

Function:

> *SwarnUnadjSeq*

Description: The *SwarnUnadjSeq* function is used to compute the estimate the parameters of genes between two specified groups of cells in a raw UMI count matrix of scRNA-seq study. It takes a non-negative integer matrix of scRNA-seq raw read counts object as input. So, the users should map the reads (obtained from sequencing libraries of the samples) to the corresponding genome and count the reads mapped to each gene according to the gene annotation to get the raw read counts matrix in advance.

Usage

```
SwarnUnadjSeq(
  CountData,
  parallel,
  norm.method,
  group,
  CellCluster,
  CellAuxil,
  maxit,
  eps,
  muoffset,
  phioffset,
  weights
)
```

Arguments

Inputs	Descriptions
CountData	Observed count data matrix for genes, rows represent genes, columns represent cells.
parallel	If FALSE (default), no parallel computation is used; if TRUE, parallel computation is performed.
norm.method	Method for normalizing the scRNA-seq count expression data, either 'DESeq2' (maximum likelihood, Ye et al., 2017) or 'TMM' (Robinson et al., 2010).
group	Vector which specifies the membership of the cells, *i.e.*, two groups to be compared, corresponding to the columns in the count data matrix.
CellCluster	Vector which specifies the cluster memberships of the cells, *i.e.*, each entry represents memberships of the columns of the count data matrix.
CellAuxil	Vector of cell level auxiliary information, corresponding to the columns in the counts matrix, default is NULL.
maxit	Maximum number of iterations for Expected-Maximization (EM) algorithm.
eps	Convergence criteria for EM algorithm.
muoffset	Offset parameter for mean (mu) parameter, default is NULL.

phioffset	Offset parameter for zero inflation (phi) parameter, default is NULL.
weights	Observation wise weights for the cells, default is unity vector.

Output:

A data frame containing the parameters from the EM algorithm for each gene, rows are genes and columns contain the following items:

- totalMean_1, totalMean_2 are the total mean, normalized mean for cellular groups 1 and 2 respectively.
- Mean, ZeroInflation, Dispersion are the MLE of the parameters of whole cellular population characetrized by the ZINB model.
- Intercept, Group2 are the co-efficients of the intercept and group effect on mean of non-zero counts.
- Cellcluster 2, Cellcluster 3, .., Cellcluster M are the effects of cell clusters on mean of non-zero counts.
- CellAuxil 2, CellAuxil 3, ..., CellAuxil N are the effects of cell-level auxiliary information on mean of non-zero counts (if included in the model).
- Intercept.0 Group2.0 are the co-efficients of the intercept and group effect on zero inflation probability.
- Cellcluster 2.0, Cellcluster 3.0, .., Cellcluster M.0 are the effects of cell clusters on zero-inflation probability.
- CellAuxil 2.0, CellAuxil 3.0, ..., CellAuxil N.0 are the effects of cell-level auxiliary information on zero-inflation probability (if included in the model).
- #iteration number of iterations required for convergence for each gene.

Examples

```
#Load the test count data, spike-in counts and spike-in concentration data
for SwarnSeq.
data(TestData)
counts <- TestData$CountData

#specifying the group information, the group 1 and 2 have two hundred cells
each.
group <- c(rep(1, 200), rep(2, 200))
#Specifying the cluster memberships of the cells in columns of countData.
cellcluster <- c(rep(1, 60), rep(2, 40), rep(3, 50),
                    rep(4, 50), rep(5, 30),
                    rep(6, 90),
                    rep(7, 80))

#Do not run
#parameters from EM algorithm for each gene.
#results <- SwarnUnadjSeq(CountData=counts, parallel=FALSE,
norm.method="TMM", group=group,
                       # CellCluster=cellcluster, CellAuxil=NULL, maxit=500,
eps=1e-10,
                           #muoffset=NULL, phioffset=NULL, weights=NULL)
```

F. Differential expression analysis of genes when observed UMI counts are not adjusted with cell molecular capture process

Function

> SwarnUnadjLRT

Description: The *SwarnUnadjLRT* function is used to detect differentially expressed genes between two specified groups of cells in a raw UMI count matrix from scRNA-seq study without adjustment for capture efficiency. It takes a non-negative integer matrix of scRNA-seq raw read counts object as input. So, the users should map the reads (obtained from sequencing libraries of the samples) to the corresponding genome and count the reads mapped to each gene according to the gene annotation to get the raw read counts matrix in advance.

Usage

```
SwarnUnadjLRT(
  CountData,
  parallel,
  norm.method,
  group,
  CellCluster,
  CellAuxil,
  maxit,
  eps,
  muoffset,
  phioffset,
  weights,
  p.adjust.method
)
```

Arguments

Inputs	Descriptions
CountData	Observed count data matrix for genes, rows represent genes, columns represent cells.
parallel	If FALSE (default), no parallel computation is used; if TRUE, parallel computation is performed.
norm.method	Method for normalizing the scRNA-seq count expression data, either 'DESeq2' (maximum likelihood, Ye et al., 2017) or 'TMM' (Robinson et al., 2010).
group	Vector which specifies the membership of the cells, i.e., two groups to be compared, corresponding to the columns in the count data matrix.
CellCluster	Vector which specifies the cluster memberships of the cells, i.e., each entry represents memberships of the columns of the count data matrix.
CellAuxil	Vector of cell level auxiliary information, corresponding to the columns in the counts matrix, default is NULL.
maxit	Maximum number of iterations for Expected-Maximization (EM) algorithm.

eps	Convergence criteria for EM algorithm.
muoffset	Offset parameter for mean (mu) parameter, default is NULL.
phioffset	Offset parameter for zero inflation (phi) parameter, default is NULL.
weights	Observation wise weights for the cells, default is unity vector.
p.adjust.method	Character variable represents the method used for multiple hypothesis correction. It can be any value from ("holm", "hochberg", "hommel", "bonferroni", "BH", "BY").

Output:

A data frame containing the results from differential expression analysis, rows are genes and columns contain the following items:

- 1 totalMean_1, totalMean_2, NormMean_1, NormMean_2 are the total mean, normalized mean for cellular groups 1 and 2 respectively.
- 2 FoldChange, log2FC, NormFC, log2NormFC are the fold change, log fold change, and log normalized fold change for the genes respectively.
- 3 Stat.DE, Pval.DE, DE.Adj.pval, and DE.FDR are values of DE statistic, p-value, adjusted p-value, false discovery rate, obtained from DE analysis, for the genes.
- 4 Stat.DZI, Pval.DZI, DZI.Adj.pval, DZI.FDR are Differential Zero Inflation (DZI) statistic, DZI p-value, DZI adjusted p-value, DZI false discovery rate results obtained for each gene from DZI analysis.

Examples

```
#Load the test count data, spike-in counts and spike-in concentration data
for SwarnSeq.

data(TestData)
counts <- TestData$CountData

#specifying the group information, the group 1 and 2 have two hundred cells
each.
group <- c(rep(1, 200), rep(2, 200))
#Specifying the cluster memberships of the cells in columns of countData.
cellcluster <- c(rep(1, 60), rep(2, 40), rep(3, 50),
                    rep(4, 50), rep(5, 30),
                    rep(6, 90),
                    rep(7, 80))

#Do not run
#parameters from EM algorithm for each gene.
#results <- SwarnUnadjLRT(CountData=counts, parallel=FALSE,
norm.method="TMM", group=group,
                       # CellCluster=cellcluster, CellAuxil=NULL, maxit=500,
eps=1e-10,
                           #muoffset=NULL, phioffset=NULL, weights=NULL,
p.adjust.method="hochberg")
```

G. Selection of top marker genes from DE analysis

Function:

> SwarnTopTags

Description: The *SwarnTopTags* function selects the top marker genes from differential expression analysis of scRNA-seq UMI count data.

Usage: SwarnTopTags(results, m)

Arguments

Inputs	Descriptions
results	A output data frame from SwarnSeqLRT or SwarnUnadjLRT which contains the unclassified differential expression analysis results.
m	A scalar representing the number of top performing genes to be selected from the scRNA-seq data.

Output: A list of the top marker genes along with their test statistic(s).

Examples

```
 #Load the test count data, spike-in counts and spike-in concentration data
for SwarnSeq.
data(TestData)
counts <- TestData$CountData
Spikes <- TestData$SpikeCounts
SpikeConc <- TestData$SpikeConc

#specifying the group information, the group 1 and 2 have two hundred cells
each.
group <- c(rep(1, 200), rep(2, 200))
#Specifying the cluster memberships of the cells in columns of countData.
cellcluster <- c(rep(1, 60), rep(2, 40), rep(3, 50),
                    rep(4, 50), rep(5, 30),
                    rep(6, 90),
                    rep(7, 80))

#results <- SwarnSeqLRT(CountData=counts, RNAspike.use=TRUE, spikes=Spikes,
spike.conc=SpikeConc,
                       #parallel=FALSE, norm.method="TMM", group=group,
CellCluster=cellcluster,
                           #CellAuxil=NULL, maxit=500, eps=1e-10,
                           #muoffset=NULL, phioffset=NULL, weights=NULL,
p.adjust.method="hochberg")

#DEGtypes <- SwarnTopTags(results, m = 100)
```

H. Classification of influential genes detected from scRNA-seq study

Function:

> *SwarnClassDE*

Description: The *SwarnClassDE* function is used to classify the influential genes of scRNA-seq data obtained from SwarnSeqLRT or SwarnUnadjLRT.

Usage

```
SwarnClassDE(results, alpha)
```

Arguments

Inputs	Descriptions
results	A output data frame from SwarnSeqLRT or SwarnUnadjLRT which contains the unclassified differential expression analysis results.
alpha	A number in (0, 0.05) to specify the threshold of adjusted p-values.

Outputs: A list containing the results along with the classes of influential genes in scRNA-seq data.

Examples

```
 #Load the test count data, spike-in counts and spike-in concentration data
for SwarnSeq.
data(TestData)
counts <- TestData$CountData
Spikes <- TestData$SpikeCounts
SpikeConc <- TestData$SpikeConc

#specifying the group information, the group 1 and 2 have two hundred cells
each.
group <- c(rep(1, 200), rep(2, 200))
#Specifying the cluster memberships of the cells in columns of countData.
cellcluster <- c(rep(1, 60), rep(2, 40), rep(3, 50),
                    rep(4, 50), rep(5, 30),
                    rep(6, 90),
                    rep(7, 80))

#results <- SwarnSeqLRT(CountData=counts, RNAspike.use=TRUE, spikes=Spikes,
spike.conc=SpikeConc,
                       #parallel=FALSE, norm.method="TMM", group=group,
CellCluster=cellcluster,
                           #CellAuxil=NULL, maxit=500, eps=1e-10,
                           #muoffset=NULL, phioffset=NULL, weights=NULL,
p.adjust.method="hochberg")

#DEGtypes <- SwarnClassDE(results, alpha = 0.0005)
```

I. Estimation of parameters for a single gene

Function:

> *ZINBEM*

Description: The *ZINBEM* function estimates the MLE of the parameters for a single gene from a vector of count expression values through Expected-Maximization (EM) algorithm.

Usage

```
ZINBEM(
  Count,
  group,
  CellCluster,
  CellAuxil,
  maxit,
  eps,
  muoffset,
  phioffset,
  weights
)
```

Arguments

Inputs	Descriptions
`Count`	Vector of count expression data of a gene over the cells.
`group`	Vector which specifies the membership of the cells, *i.e.*, two groups to be compared, corresponding to the cells in the count.
`CellCluster`	Vector which specifies the cluster memberships of the cells, *i.e*., each entry represents memberships of the entries of the count.
`CellAuxil`	Vector of cell level auxiliary information, corresponding to the entries in the counts, default is NULL.
`maxit`	Maximum number of iterations for Expected-Maximization (EM) algorithm.
`eps`	Convergence criteria for EM algorithm.
`muoffset`	Offset parameter for mean (mu) parameter, default is NULL.
`phioffset`	Offset parameter for zero inflation (phi) parameter, default is NULL.
`weights`	Observation wise weights for the cells, default is unity vector.

Output: A list containing the estimates of the parameters of a gene.

J. Optimization the likelihood function for a gene, supplies starting values for EM algorithm (parameters estimation when EM algorithm does not converge)

Function:

> *ZINBoptim*

Description: This function optimizes the likelihood function for a gene, supplies starting values for EM algorithm.

Usage:

```
ZINBoptim(Count, group, CellCluster, CellAuxil, muoffset, phioffset, weights)
```

Arguments

Inputs	Descriptions
`Count`	Vector of count expression data of a gene over the cells.
`group`	Vector which specifies the membership of the cells, *i.e.*, two groups to be compared, corresponding to the cells in the count.
`CellCluster`	Vector which specifies the cluster memberships of the cells, *i.e.*, each entry represents memberships of the entries of the count.
`CellAuxil`	Vector of cell level auxiliary information, corresponding to the entries in the counts, default is NULL.
`muoffset`	Offset parameter for mean (mu) parameter, default is NULL.
`phioffset`	Offset parameter for zero inflation (phi) parameter, default is NULL.
`weights`	Observation wise weights for the cells, default is unity vector.

Output: A list containing the estimates of the parameters of a gene.

K. Probability function of zero inflated negative binomial distribution

Function:

> *dzinb*

Description: The *dzinb* function computes the distribution of the zero inflated negative binomial distribution.

Usage

```
dzinb(x, size, mu, rho, log)
```

Arguments

Inputs	Descriptions
`x`	Vector of (non-negative integer) quantiles.
`size`	size parameter.
`mu`	mean parameter

`rho`	Zero inflation parameter.
`log`	Logical variable either TRUE/FALSE; if TRUE, probabilities p's are given as log(p).

Output:

Returns the distribution function for the zero inflated negative binomial distribution with mean, size, zero inflation parameters.

Example:

```
dzinb(x =100, size = 3.5, mu = 2.45, rho = 0.34, log = FALSE)
```

Table S6. List of differentially zero inflated genes identified through SwarnSeq method. The results are provided in separate excel sheet.

15. Supplementary Tables

Table S10. Performance evaluation metrics for GSE92495 scRNA-seq data.

							NDEG = 500				
Methods	**TP**	**FP**	**TN**	**FN**	**TPR**	**FPR**	**FDR**	**PPR**	**NPV**	**ACC**	**F1**
SwarnSeq	292	208	12316	2708	0.097	0.017	0.416	0.584	0.820	0.812	0.167
DEGSeq	193	307	12221	2807	0.064	0.025	0.614	0.386	0.813	0.799	0.110
DESeq2	45	455	12069	2955	0.015	0.036	0.910	0.090	0.803	0.780	0.026
DESingle	150	350	12174	2850	0.050	0.028	0.700	0.300	0.810	0.794	0.086
EdgeR	1	499	12025	2999	0.000	0.040	0.998	0.002	0.800	0.775	0.001
Limma	226	274	12438	2774	0.075	0.022	0.548	0.452	0.818	0.806	0.129
DECENT	310	190	12351	2690	0.103	0.015	0.380	0.620	0.821	0.815	0.177
MAST	9	491	12033	2991	0.003	0.039	0.982	0.018	0.801	0.776	0.005
Monocle	273	227	12302	2727	0.091	0.018	0.454	0.546	0.819	0.810	0.156
NODES	18	482	12042	2982	0.006	0.038	0.964	0.036	0.802	0.777	0.010
ScDD	29	471	12053	2971	0.010	0.038	0.942	0.058	0.802	0.778	0.017
BPSC	1	499	12025	2999	0.000	0.040	0.998	0.002	0.800	0.775	0.001
							NDEG = 1000				
SwarnSeq	506	494	12030	2494	0.169	0.039	0.494	0.506	0.828	0.808	0.253
DEGSeq	423	577	11951	2577	0.141	0.046	0.577	0.423	0.823	0.797	0.212
DESeq2	55	945	11579	2945	0.018	0.075	0.945	0.055	0.797	0.749	0.028
DESingle	150	850	11674	2850	0.050	0.068	0.850	0.150	0.804	0.762	0.075
EdgeR	1	999	11525	2999	0.000	0.080	0.999	0.001	0.794	0.742	0.001
Limma	435	565	12282	2565	0.145	0.044	0.565	0.435	0.827	0.802	0.218
DECENT	595	405	12137	2405	0.198	0.032	0.405	0.595	0.835	0.819	0.298
MAST	12	988	11536	2988	0.004	0.079	0.988	0.012	0.794	0.744	0.006
Monocle	550	450	12086	2450	0.183	0.036	0.450	0.550	0.831	0.813	0.275
NODES	37	963	11561	2963	0.012	0.077	0.963	0.037	0.796	0.747	0.019
ScDD	52	948	11576	2948	0.017	0.076	0.948	0.052	0.797	0.749	0.026
BPSC	1	999	11525	2999	0.000	0.080	0.999	0.001	0.794	0.742	0.001
							NDEG = 1500				
SwarnSeq	813	687	11837	2187	0.271	0.055	0.458	0.542	0.844	0.815	0.361
DEGSeq	631	869	11673	2369	0.210	0.069	0.579	0.421	0.831	0.792	0.280
DESeq2	61	1439	11085	2939	0.020	0.115	0.959	0.041	0.790	0.718	0.027
DESingle	150	1350	11174	2850	0.050	0.108	0.900	0.100	0.797	0.729	0.067
EdgeR	1	1499	11025	2999	0.000	0.120	0.999	0.001	0.786	0.710	0.000
Limma	678	822	12112	2322	0.226	0.064	0.548	0.452	0.839	0.803	0.301
DECENT	862	638	11905	2138	0.287	0.051	0.425	0.575	0.848	0.821	0.383
MAST	21	1479	11045	2979	0.007	0.118	0.986	0.014	0.788	0.713	0.009
Monocle	775	725	11821	2225	0.258	0.058	0.483	0.517	0.842	0.810	0.344
NODES	68	1432	11092	2932	0.023	0.114	0.955	0.045	0.791	0.719	0.030
ScDD	74	1426	11098	2926	0.025	0.114	0.951	0.049	0.791	0.720	0.033
BPSC	1	1499	11025	2999	0.000	0.120	0.999	0.001	0.786	0.710	0.000

						NDEG = 2000					
SwarnSeq	1113	887	11637	1887	0.371	0.071	0.444	0.557	0.860	0.821	0.445
DEGSeq	846	1154	11397	2154	0.282	0.092	0.577	0.423	0.841	0.787	0.338
DESeq2	189	1811	10750	2811	0.063	0.144	0.906	0.095	0.793	0.703	0.076
DESingle	150	1850	10674	2850	0.050	0.148	0.925	0.075	0.789	0.697	0.060
EdgeR	1	1999	10525	2999	0.000	0.160	1.000	0.001	0.778	0.678	0.000
Limma	936	1064	11909	2064	0.312	0.082	0.532	0.468	0.852	0.804	0.374
DECENT	1085	915	11682	1915	0.362	0.073	0.458	0.543	0.859	0.819	0.434
MAST	30	1970	10554	2970	0.010	0.157	0.985	0.015	0.780	0.682	0.012
Monocle	988	1012	11548	2012	0.329	0.081	0.506	0.494	0.852	0.806	0.395
NODES	102	1898	10626	2898	0.034	0.152	0.949	0.051	0.786	0.691	0.041
ScDD	99	1901	10623	2901	0.033	0.152	0.951	0.050	0.785	0.691	0.040
BPSC	3	1997	10527	2997	0.001	0.159	0.999	0.002	0.778	0.678	0.001
						NDEG = 2500					
SwarnSeq	1550	950	11574	1450	0.517	0.076	0.380	0.620	0.889	0.845	0.564
DEGSeq	1066	1434	11117	1934	0.355	0.114	0.574	0.426	0.852	0.783	0.388
DESeq2	275	2225	10351	2725	0.092	0.177	0.890	0.110	0.792	0.682	0.100
DESingle	150	2350	10174	2850	0.050	0.188	0.940	0.060	0.781	0.665	0.055
EdgeR	1	2499	10025	2999	0.000	0.200	1.000	0.000	0.770	0.646	0.000
Limma	1174	1326	11667	1826	0.391	0.102	0.530	0.470	0.865	0.803	0.427
DECENT	1367	1133	11468	1633	0.456	0.090	0.453	0.547	0.875	0.823	0.497
MAST	43	2457	10067	2957	0.014	0.196	0.983	0.017	0.773	0.651	0.016
Monocle	1171	1329	11251	1829	0.390	0.106	0.532	0.468	0.860	0.797	0.426
NODES	133	2367	10157	2867	0.044	0.189	0.947	0.053	0.780	0.663	0.048
ScDD	129	2371	10153	2871	0.043	0.189	0.948	0.052	0.780	0.662	0.047
BPSC	5	2495	10029	2995	0.002	0.199	0.998	0.002	0.770	0.646	0.002
						NDEG = 3000					
SwarnSeq	2050	950	11574	950	0.683	0.076	0.317	0.683	0.924	0.878	0.683
DEGSeq	1277	1723	10833	1723	0.426	0.137	0.574	0.426	0.863	0.778	0.426
DESeq2	371	2629	9954	2629	0.124	0.209	0.876	0.124	0.791	0.663	0.124
DESingle	150	2850	9674	2850	0.050	0.228	0.950	0.050	0.772	0.633	0.050
EdgeR	2	2998	9526	2998	0.001	0.239	0.999	0.001	0.761	0.614	0.001
Limma	1395	1605	11391	1605	0.465	0.123	0.535	0.465	0.877	0.799	0.465
DECENT	1641	1359	11246	1359	0.547	0.108	0.453	0.547	0.892	0.826	0.547
MAST	64	2936	9588	2936	0.021	0.234	0.979	0.021	0.766	0.622	0.021
Monocle	1328	1672	10933	1672	0.443	0.133	0.557	0.443	0.867	0.786	0.443
NODES	178	2822	9702	2822	0.059	0.225	0.941	0.059	0.775	0.636	0.059
ScDD	162	2838	9686	2838	0.054	0.227	0.946	0.054	0.773	0.634	0.054
BPSC	5	2995	9529	2995	0.002	0.239	0.998	0.002	0.761	0.614	0.002

NDEG: Number of differentially expressed genes; TP: True Positive; FP: False Positive; TN: True Negative; FN: False Negative; TPR: True Positive Rate; FPR: False Positive Rate; FDR: False Discovery Rate; PPR: Positive Prediction Rate; NPV: Negative Prediction Value; ACC: Accuracy; F1: F-score

Table S11. Performance evaluation metrics for GSE53638 (Data 1) scRNA-seq data.

					NDEG = 500						
Methods	**TP**	**FP**	**TN**	**FN**	**TPR**	**FPR**	**FDR**	**PPR**	**NPV**	**ACC**	**F1**
SwarnSeq	475	25	11732	3525	0.119	0.002	0.050	0.950	0.769	0.775	0.211
DEGSeq	220	280	11460	3780	0.055	0.024	0.560	0.440	0.752	0.742	0.098
DESeq2	26	474	11266	3974	0.007	0.040	0.948	0.052	0.739	0.717	0.012
DESingle	207	293	11447	3793	0.052	0.025	0.586	0.414	0.751	0.740	0.092
EdgeR	181	319	11421	3819	0.045	0.027	0.638	0.362	0.749	0.737	0.080
Limma	102	398	11342	3898	0.026	0.034	0.796	0.204	0.744	0.727	0.045
DECENT	177	323	11417	3823	0.044	0.028	0.646	0.354	0.749	0.737	0.079
MAST	236	264	11476	3764	0.059	0.022	0.528	0.472	0.753	0.744	0.105
Monocle	271	229	11511	3729	0.068	0.020	0.458	0.542	0.755	0.749	0.120
NODES	225	275	11465	3775	0.056	0.023	0.550	0.450	0.752	0.743	0.100
scDD	122	378	11362	3878	0.031	0.032	0.756	0.244	0.746	0.730	0.054
BPSC	164	336	11404	3836	0.041	0.029	0.672	0.328	0.748	0.735	0.073
					NDEG = 1000						
SwarnSeq	907	93	11664	3093	0.227	0.008	0.093	0.907	0.790	0.798	0.363
DEGSeq	396	604	11136	3604	0.099	0.051	0.604	0.396	0.755	0.733	0.158
DESeq2	33	967	10773	3967	0.008	0.082	0.967	0.033	0.731	0.687	0.013
DESingle	307	693	11047	3693	0.077	0.059	0.693	0.307	0.749	0.721	0.123
EdgeR	275	725	11015	3725	0.069	0.062	0.725	0.275	0.747	0.717	0.110
Limma	150	850	10890	3850	0.038	0.072	0.850	0.150	0.739	0.701	0.060
DECENT	274	726	11014	3726	0.069	0.062	0.726	0.274	0.747	0.717	0.110
MAST	366	634	11106	3634	0.092	0.054	0.634	0.366	0.753	0.729	0.146
Monocle	429	571	11170	3571	0.107	0.049	0.571	0.429	0.758	0.737	0.172
NODES	373	627	11113	3627	0.093	0.053	0.627	0.373	0.754	0.730	0.149
scDD	189	811	10929	3811	0.047	0.069	0.811	0.189	0.741	0.706	0.076
BPSC	249	751	10989	3751	0.062	0.064	0.751	0.249	0.746	0.714	0.100
					NDEG = 1500						
SwarnSeq	1264	236	11521	2736	0.316	0.020	0.157	0.843	0.808	0.811	0.460
DEGSeq	556	944	10797	3444	0.139	0.080	0.629	0.371	0.758	0.721	0.202
DESeq2	37	1463	10277	3963	0.009	0.125	0.975	0.025	0.722	0.655	0.013
DESingle	410	1090	10650	3590	0.103	0.093	0.727	0.273	0.748	0.703	0.149
EdgeR	390	1110	10631	3610	0.098	0.095	0.740	0.260	0.747	0.700	0.142
Limma	195	1305	10435	3805	0.049	0.111	0.870	0.130	0.733	0.675	0.071
DECENT	364	1136	10605	3636	0.091	0.097	0.757	0.243	0.745	0.697	0.132
MAST	493	1007	10734	3507	0.123	0.086	0.671	0.329	0.754	0.713	0.179
Monocle	593	907	10834	3407	0.148	0.077	0.605	0.395	0.761	0.726	0.216
NODES	540	960	10780	3460	0.135	0.082	0.640	0.360	0.757	0.719	0.196
scDD	287	1213	10527	3713	0.072	0.103	0.809	0.191	0.739	0.687	0.104
BPSC	341	1159	10581	3659	0.085	0.099	0.773	0.227	0.743	0.694	0.124
					NDEG = 2000						

SwarnSeq	1554	446	11311	2446	0.389	0.038	0.223	0.777	0.822	0.816	0.518
DEGSeq	687	1313	10428	3313	0.172	0.112	0.657	0.344	0.759	0.706	0.229
DESeq2	41	1959	9781	3959	0.010	0.167	0.980	0.021	0.712	0.624	0.014
DESingle	497	1503	10237	3503	0.124	0.128	0.752	0.249	0.745	0.682	0.166
EdgeR	522	1478	10263	3478	0.131	0.126	0.739	0.261	0.747	0.685	0.174
Limma	243	1757	9983	3757	0.061	0.150	0.879	0.122	0.727	0.650	0.081
DECENT	465	1535	10206	3535	0.116	0.131	0.768	0.233	0.743	0.678	0.155
MAST	622	1378	10366	3378	0.156	0.117	0.689	0.311	0.754	0.698	0.207
Monocle	737	1263	10480	3263	0.184	0.108	0.632	0.369	0.763	0.713	0.246
NODES	757	1243	10498	3243	0.189	0.106	0.622	0.379	0.764	0.715	0.252
scDD	385	1615	10125	3615	0.096	0.138	0.808	0.193	0.737	0.668	0.128
BPSC	452	1548	10192	3548	0.113	0.132	0.774	0.226	0.742	0.676	0.151
NDEG = 2500											
SwarnSeq	1792	708	11049	2208	0.448	0.060	0.283	0.717	0.833	0.815	0.551
DEGSeq	826	1674	10070	3174	0.207	0.143	0.670	0.330	0.760	0.692	0.254
DESeq2	54	2446	9294	3946	0.014	0.208	0.978	0.022	0.702	0.594	0.017
DESingle	618	1882	9859	3382	0.155	0.160	0.753	0.247	0.745	0.666	0.190
EdgeR	655	1845	9903	3345	0.164	0.157	0.738	0.262	0.748	0.670	0.202
Limma	308	2192	9548	3692	0.077	0.187	0.877	0.123	0.721	0.626	0.095
DECENT	573	1927	9819	3427	0.143	0.164	0.771	0.229	0.741	0.660	0.176
MAST	761	1739	10011	3239	0.190	0.148	0.696	0.304	0.756	0.684	0.234
Monocle	899	1601	10143	3101	0.225	0.136	0.640	0.360	0.766	0.701	0.277
NODES	948	1552	10189	3052	0.237	0.132	0.621	0.379	0.770	0.708	0.292
scDD	476	2024	9716	3524	0.119	0.172	0.810	0.190	0.734	0.648	0.146
BPSC	589	1911	9829	3411	0.147	0.163	0.764	0.236	0.742	0.662	0.181
NDEG = 3000											
SwarnSeq	2003	997	10762	1997	0.501	0.085	0.332	0.668	0.843	0.810	0.572
DEGSeq	988	2012	9734	3012	0.247	0.171	0.671	0.329	0.764	0.681	0.282
DESeq2	79	2921	8819	3921	0.020	0.249	0.974	0.026	0.692	0.565	0.023
DESingle	736	2264	9477	3264	0.184	0.193	0.755	0.245	0.744	0.649	0.210
EdgeR	813	2187	9568	3187	0.203	0.186	0.729	0.271	0.750	0.659	0.232
Limma	372	2628	9112	3628	0.093	0.224	0.876	0.124	0.715	0.603	0.106
DECENT	701	2299	9448	3299	0.175	0.196	0.766	0.234	0.741	0.645	0.200
MAST	890	2110	9646	3110	0.223	0.179	0.703	0.297	0.756	0.669	0.254
Monocle	1039	1961	9785	2961	0.260	0.167	0.654	0.346	0.768	0.687	0.297
NODES	1164	1836	9907	2836	0.291	0.156	0.612	0.388	0.777	0.703	0.333
scDD	608	2392	9348	3392	0.152	0.204	0.797	0.203	0.734	0.633	0.174
BPSC	731	2269	9471	3269	0.183	0.193	0.756	0.244	0.743	0.648	0.209

NDEG: Number of differentially expressed genes; TP: True Positive; FP: False Positive; TN: True Negative; FN: False Negative; TPR: True Positive Rate; FPR: False Positive Rate; FDR: False Discovery Rate; PPR: Positive Prediction Rate; NPV: Negative Prediction Value; ACC: Accuracy; F1: F-score

Table S12. Performance evaluation metrics for GSE53638 (Data 2) scRNA-seq data.

					NDEG = 500						
Methods	**TP**	**FP**	**TN**	**FN**	**TPR**	**FPR**	**FDR**	**PPR**	**NPV**	**ACC**	**F1**
SwarnSeq	258	242	12393	2742	0.086	0.019	0.484	0.516	0.819	0.809	0.147
DEGSeq	145	355	12280	2855	0.048	0.028	0.710	0.290	0.811	0.795	0.083
DESeq2	20	480	12155	2980	0.007	0.038	0.960	0.040	0.803	0.779	0.011
DESingle	122	378	12257	2878	0.041	0.030	0.756	0.244	0.810	0.792	0.070
EdgeR	72	428	12207	2928	0.024	0.034	0.856	0.144	0.807	0.785	0.041
Limma	54	446	12189	2946	0.018	0.035	0.892	0.108	0.805	0.783	0.031
DECENT	14	486	12149	2986	0.005	0.038	0.972	0.028	0.803	0.778	0.008
MAST	130	370	12265	2870	0.043	0.029	0.740	0.260	0.810	0.793	0.074
Monocle	89	411	12224	2911	0.030	0.033	0.822	0.178	0.808	0.788	0.051
NODES	51	449	12186	2949	0.017	0.036	0.898	0.102	0.805	0.783	0.029
ScDD	130	370	12265	2870	0.043	0.029	0.740	0.260	0.810	0.793	0.074
BPSC	55	445	12190	2945	0.018	0.035	0.890	0.110	0.805	0.783	0.031
					NDEG = 1000						
SwarnSeq	585	415	12220	2415	0.195	0.033	0.415	0.585	0.835	0.819	0.293
DEGSeq	272	728	11907	2728	0.091	0.058	0.728	0.272	0.814	0.779	0.136
DESeq2	26	974	11661	2974	0.009	0.077	0.974	0.026	0.797	0.747	0.013
DESingle	214	786	11849	2786	0.071	0.062	0.786	0.214	0.810	0.772	0.107
EdgeR	129	871	11764	2871	0.043	0.069	0.871	0.129	0.804	0.761	0.065
Limma	81	919	11716	2919	0.027	0.073	0.919	0.081	0.801	0.755	0.041
DECENT	17	982	11653	2983	0.006	0.078	0.983	0.017	0.796	0.746	0.009
MAST	242	758	11877	2758	0.081	0.060	0.758	0.242	0.812	0.775	0.121
Monocle	155	845	11790	2845	0.052	0.067	0.845	0.155	0.806	0.764	0.078
NODES	84	916	11719	2916	0.028	0.072	0.916	0.084	0.801	0.755	0.042
ScDD	248	752	11883	2752	0.083	0.060	0.752	0.248	0.812	0.776	0.124
BPSC	89	911	11724	2911	0.030	0.072	0.911	0.089	0.801	0.756	0.045
					NDEG = 1500						
SwarnSeq	767	733	11902	2233	0.256	0.058	0.489	0.511	0.842	0.810	0.341
DEGSeq	417	1083	11552	2583	0.139	0.086	0.722	0.278	0.817	0.766	0.185
DESeq2	40	1460	11175	2960	0.013	0.116	0.973	0.027	0.791	0.717	0.018
DESingle	303	1197	11438	2697	0.101	0.095	0.798	0.202	0.809	0.751	0.135
EdgeR	208	1292	11343	2792	0.069	0.102	0.861	0.139	0.802	0.739	0.092
Limma	100	1400	11235	2900	0.033	0.111	0.933	0.067	0.795	0.725	0.044
DECENT	35	1464	11171	2965	0.012	0.116	0.977	0.023	0.790	0.717	0.016
MAST	363	1137	11498	2637	0.121	0.090	0.758	0.242	0.813	0.759	0.161
Monocle	224	1276	11359	2776	0.075	0.101	0.851	0.149	0.804	0.741	0.100
NODES	94	1406	11229	2906	0.031	0.111	0.937	0.063	0.794	0.724	0.042
ScDD	342	1158	11477	2658	0.114	0.092	0.772	0.228	0.812	0.756	0.152
BPSC	128	1372	11263	2872	0.043	0.109	0.915	0.085	0.797	0.729	0.057
					NDEG = 2000						

SwarnSeq	904	1096	11539	2096	0.301	0.087	0.548	0.452	0.846	0.796	0.362
DEGSeq	533	1467	11168	2467	0.178	0.116	0.734	0.267	0.819	0.748	0.213
DESeq2	71	1929	10706	2929	0.024	0.153	0.965	0.036	0.785	0.689	0.028
DESingle	403	1597	11038	2597	0.134	0.126	0.799	0.202	0.810	0.732	0.161
EdgeR	287	1713	10922	2713	0.096	0.136	0.857	0.144	0.801	0.717	0.115
Limma	128	1872	10763	2872	0.043	0.148	0.936	0.064	0.789	0.697	0.051
DECENT	86	1913	10722	2914	0.029	0.151	0.957	0.043	0.786	0.691	0.034
MAST	482	1518	11117	2518	0.161	0.120	0.759	0.241	0.815	0.742	0.193
Monocle	316	1684	10951	2684	0.105	0.133	0.842	0.158	0.803	0.721	0.126
NODES	108	1892	10743	2892	0.036	0.150	0.946	0.054	0.788	0.694	0.043
ScDD	449	1551	11084	2551	0.150	0.123	0.776	0.225	0.813	0.738	0.180
BPSC	190	1810	10825	2810	0.063	0.143	0.905	0.095	0.794	0.705	0.076
NDEG = 2500											
SwarnSeq	1031	1469	11166	1969	0.344	0.116	0.588	0.412	0.850	0.780	0.375
DEGSeq	628	1872	10763	2372	0.209	0.148	0.749	0.251	0.819	0.729	0.228
DESeq2	115	2385	10250	2885	0.038	0.189	0.954	0.046	0.780	0.663	0.042
DESingle	520	1980	10655	2480	0.173	0.157	0.792	0.208	0.811	0.715	0.189
EdgeR	411	2089	10546	2589	0.137	0.165	0.836	0.164	0.803	0.701	0.149
Limma	159	2341	10294	2841	0.053	0.185	0.936	0.064	0.784	0.669	0.058
DECENT	185	2314	10321	2815	0.062	0.183	0.926	0.074	0.786	0.672	0.067
MAST	609	1891	10744	2391	0.203	0.150	0.756	0.244	0.818	0.726	0.221
Monocle	426	2074	10561	2574	0.142	0.164	0.830	0.170	0.804	0.703	0.155
NODES	117	2383	10252	2883	0.039	0.189	0.953	0.047	0.781	0.663	0.043
ScDD	568	1932	10703	2432	0.189	0.153	0.773	0.227	0.815	0.721	0.207
BPSC	271	2229	10406	2729	0.090	0.176	0.892	0.108	0.792	0.683	0.099
NDEG = 3000											
SwarnSeq	1145	1855	10780	1855	0.382	0.147	0.618	0.382	0.853	0.763	0.382
DEGSeq	717	2283	10352	2283	0.239	0.181	0.761	0.239	0.819	0.708	0.239
DESeq2	192	2808	9827	2808	0.064	0.222	0.936	0.064	0.778	0.641	0.064
DESingle	648	2352	10283	2352	0.216	0.186	0.784	0.216	0.814	0.699	0.216
EdgeR	552	2448	10187	2448	0.184	0.194	0.816	0.184	0.806	0.687	0.184
Limma	213	2787	9848	2787	0.071	0.221	0.929	0.071	0.779	0.643	0.071
DECENT	368	2628	10007	2632	0.123	0.208	0.877	0.123	0.792	0.664	0.123
MAST	724	2276	10359	2276	0.241	0.180	0.759	0.241	0.820	0.709	0.241
Monocle	564	2436	10199	2436	0.188	0.193	0.812	0.188	0.807	0.688	0.188
NODES	136	2864	9771	2864	0.045	0.227	0.955	0.045	0.773	0.634	0.045
ScDD	659	2341	10294	2341	0.220	0.185	0.780	0.220	0.815	0.701	0.220
BPSC	398	2602	10033	2602	0.133	0.206	0.867	0.133	0.794	0.667	0.133

NDEG: Number of differentially expressed genes; TP: True Positive; FP: False Positive; TN: True Negative; FN: False Negative; TPR: True Positive Rate; FPR: False Positive Rate; FDR: False Discovery Rate; PPR: Positive Prediction Rate; NPV: Negative Prediction Value; ACC: Accuracy; F1: F-score

Table S13. Performance evaluation metrics for GSE53638 (Data 3) for scRNA-seq data.

					NDEG = 500						
Methods	**TP**	**FP**	**TN**	**FN**	**TPR**	**FPR**	**FDR**	**PPR**	**NPV**	**ACC**	**F1**
SwarnSeq	136	364	11648	2864	0.045	0.030	0.728	0.272	0.803	0.785	0.078
DEGSeq	114	386	11626	2886	0.038	0.032	0.772	0.228	0.801	0.782	0.065
DESeq2	21	479	11533	2979	0.007	0.040	0.958	0.042	0.795	0.770	0.012
DESingle	213	287	11725	2787	0.071	0.024	0.574	0.426	0.808	0.795	0.122
EdgeR	151	349	11663	2849	0.050	0.029	0.698	0.302	0.804	0.787	0.086
Limma	161	339	11673	2839	0.054	0.028	0.678	0.322	0.804	0.788	0.092
DECENT	14	486	11526	2986	0.005	0.040	0.972	0.028	0.794	0.769	0.008
MAST	145	355	11657	2855	0.048	0.030	0.710	0.290	0.803	0.786	0.083
Monocle	94	406	11606	2906	0.031	0.034	0.812	0.188	0.800	0.779	0.054
NODES	88	412	11600	2912	0.029	0.034	0.824	0.176	0.799	0.779	0.050
scDD	42	458	11554	2958	0.014	0.038	0.916	0.084	0.796	0.772	0.024
BPSC	81	419	11593	2919	0.027	0.035	0.838	0.162	0.799	0.778	0.046
					NDEG = 1000						
SwarnSeq	374	626	11386	2626	0.125	0.052	0.626	0.374	0.813	0.783	0.187
DEGSeq	249	751	11261	2751	0.083	0.063	0.751	0.249	0.804	0.767	0.125
DESeq2	49	951	11061	2951	0.016	0.079	0.951	0.049	0.789	0.740	0.025
DESingle	342	658	11354	2658	0.114	0.055	0.658	0.342	0.810	0.779	0.171
EdgeR	271	729	11283	2729	0.090	0.061	0.729	0.271	0.805	0.770	0.136
Limma	253	747	11265	2747	0.084	0.062	0.747	0.253	0.804	0.767	0.127
DECENT	52	948	11064	2948	0.017	0.079	0.948	0.052	0.790	0.740	0.026
MAST	257	743	11269	2743	0.086	0.062	0.743	0.257	0.804	0.768	0.129
Monocle	184	816	11196	2816	0.061	0.068	0.816	0.184	0.799	0.758	0.092
NODES	163	837	11175	2837	0.054	0.070	0.837	0.163	0.798	0.755	0.082
scDD	57	943	11070	2943	0.019	0.078	0.943	0.057	0.790	0.741	0.029
BPSC	140	860	11152	2860	0.047	0.072	0.860	0.140	0.796	0.752	0.070
					NDEG = 1500						
SwarnSeq	656	844	11168	2344	0.219	0.070	0.563	0.437	0.827	0.788	0.292
DEGSeq	358	1142	10870	2642	0.119	0.095	0.761	0.239	0.804	0.748	0.159
DESeq2	100	1400	10612	2900	0.033	0.117	0.933	0.067	0.785	0.714	0.044
DESingle	469	1031	10982	2531	0.156	0.086	0.687	0.313	0.813	0.763	0.208
EdgeR	397	1103	10909	2603	0.132	0.092	0.735	0.265	0.807	0.753	0.176
Limma	316	1184	10828	2684	0.105	0.099	0.789	0.211	0.801	0.742	0.140
DECENT	136	1364	10648	2864	0.045	0.114	0.909	0.091	0.788	0.718	0.060
MAST	373	1127	10885	2627	0.124	0.094	0.751	0.249	0.806	0.750	0.166
Monocle	286	1214	10799	2714	0.095	0.101	0.809	0.191	0.799	0.738	0.127
NODES	250	1250	10762	2750	0.083	0.104	0.833	0.167	0.796	0.734	0.111
scDD	62	1438	10575	2938	0.021	0.120	0.959	0.041	0.783	0.709	0.028
BPSC	229	1271	10741	2771	0.076	0.106	0.847	0.153	0.795	0.731	0.102
					NDEG =2000						

SwarnSeq	847	1153	10859	2153	0.282	0.096	0.577	0.424	0.835	0.780	0.339
DEGSeq	433	1567	10445	2567	0.144	0.130	0.784	0.217	0.803	0.725	0.173
DESeq2	178	1822	10190	2822	0.059	0.152	0.911	0.089	0.783	0.691	0.071
DESingle	593	1407	10606	2407	0.198	0.117	0.704	0.297	0.815	0.746	0.237
EdgeR	541	1459	10553	2459	0.180	0.121	0.730	0.271	0.811	0.739	0.216
Limma	390	1610	10402	2610	0.130	0.134	0.805	0.195	0.799	0.719	0.156
DECENT	260	1740	10272	2740	0.087	0.145	0.870	0.130	0.789	0.702	0.104
MAST	505	1495	10517	2495	0.168	0.124	0.748	0.253	0.808	0.734	0.202
Monocle	426	1574	10439	2574	0.142	0.131	0.787	0.213	0.802	0.724	0.170
NODES	357	1643	10369	2643	0.119	0.137	0.822	0.179	0.797	0.714	0.143
scDD	69	1931	10082	2931	0.023	0.161	0.966	0.035	0.775	0.676	0.028
BPSC	351	1649	10363	2649	0.117	0.137	0.825	0.176	0.796	0.714	0.140
NDEG = 2000											
SwarnSeq	1005	1495	10517	1995	0.335	0.124	0.598	0.402	0.841	0.768	0.365
DEGSeq	562	1938	10075	2438	0.187	0.161	0.775	0.225	0.805	0.709	0.204
DESeq2	260	2240	9772	2740	0.087	0.186	0.896	0.104	0.781	0.668	0.095
DESingle	717	1783	10230	2283	0.239	0.148	0.713	0.287	0.818	0.729	0.261
EdgeR	689	1811	10201	2311	0.230	0.151	0.724	0.276	0.815	0.725	0.251
Limma	482	2018	9994	2518	0.161	0.168	0.807	0.193	0.799	0.698	0.175
DECENT	416	2084	9928	2584	0.139	0.173	0.834	0.166	0.793	0.689	0.151
MAST	635	1865	10147	2365	0.212	0.155	0.746	0.254	0.811	0.718	0.231
Monocle	558	1942	10071	2442	0.186	0.162	0.777	0.223	0.805	0.708	0.203
NODES	469	2031	9981	2531	0.156	0.169	0.812	0.188	0.798	0.696	0.171
scDD	75	2425	9588	2925	0.025	0.202	0.970	0.030	0.766	0.644	0.027
BPSC	478	2022	9990	2522	0.159	0.168	0.809	0.191	0.798	0.697	0.174
NDEG = 3000											
SwarnSeq	1146	1854	10158	1854	0.382	0.154	0.618	0.382	0.846	0.753	0.382
DEGSeq	685	2315	9698	2315	0.228	0.193	0.772	0.228	0.807	0.692	0.228
DESeq2	317	2683	9329	2683	0.106	0.223	0.894	0.106	0.777	0.643	0.106
DESingle	844	2156	9857	2156	0.281	0.179	0.719	0.281	0.821	0.713	0.281
EdgeR	862	2138	9874	2138	0.287	0.178	0.713	0.287	0.822	0.715	0.287
Limma	577	2423	9589	2423	0.192	0.202	0.808	0.192	0.798	0.677	0.192
DECENT	618	2382	9630	2382	0.206	0.198	0.794	0.206	0.802	0.683	0.206
MAST	785	2215	9797	2215	0.262	0.184	0.738	0.262	0.816	0.705	0.262
Monocle	710	2290	9723	2290	0.237	0.191	0.763	0.237	0.809	0.695	0.237
NODES	578	2422	9590	2422	0.193	0.202	0.807	0.193	0.798	0.677	0.193
scDD	89	2911	9102	2911	0.030	0.242	0.970	0.030	0.758	0.612	0.030
BPSC	645	2355	9657	2355	0.215	0.196	0.785	0.215	0.804	0.686	0.215

NDEG: Number of differentially expressed genes; TP: True Positive; FP: False Positive; TN: True Negative; FN: False Negative; TPR: True Positive Rate; FPR: False Positive Rate; FDR: False Discovery Rate; PPR: Positive Prediction Rate; NPV: Negative Prediction Value; ACC: Accuracy; F1: F-score

Table S14. Performance evaluation metrics for GSE65525 scRNA-seq data.

NDEG = 500											
Methods	**TP**	**FP**	**TN**	**FN**	**TPR**	**FPR**	**FDR**	**PPR**	**NPV**	**ACC**	**F1**
SwarnSeq	500	0	20955	2500	0.167	0.000	0.000	1.000	0.893	0.896	0.286
DEGSeq	191	309	20646	2809	0.064	0.015	0.618	0.382	0.880	0.870	0.109
DESeq2	155	345	20610	2845	0.052	0.016	0.690	0.310	0.879	0.867	0.089
DESingle	221	279	20676	2779	0.074	0.013	0.558	0.442	0.882	0.872	0.126
EdgeR	206	294	20661	2794	0.069	0.014	0.588	0.412	0.881	0.871	0.118
Limma	38	462	20495	2962	0.013	0.022	0.924	0.076	0.874	0.857	0.022
DECENT	178	322	20633	2822	0.059	0.015	0.644	0.356	0.880	0.869	0.102
MAST	77	423	20532	2923	0.026	0.020	0.846	0.154	0.875	0.860	0.044
Monocle	134	366	20589	2866	0.045	0.017	0.732	0.268	0.878	0.865	0.077
NODES	59	441	20514	2941	0.020	0.021	0.882	0.118	0.875	0.859	0.034
scDD	72	428	20527	2928	0.024	0.020	0.856	0.144	0.875	0.860	0.041
BPSC	135	365	20590	2865	0.045	0.017	0.730	0.270	0.878	0.865	0.077
NDEG=1000											
SwarnSeq	1000	0	20955	2000	0.333	0.000	0.000	1.000	0.913	0.917	0.500
DEGSeq	294	706	20249	2706	0.098	0.034	0.706	0.294	0.882	0.858	0.147
DESeq2	226	774	20183	2774	0.075	0.037	0.774	0.226	0.879	0.852	0.113
DESingle	306	694	20261	2694	0.102	0.033	0.694	0.306	0.883	0.859	0.153
EdgeR	288	712	20243	2712	0.096	0.034	0.712	0.288	0.882	0.857	0.144
Limma	85	915	20043	2915	0.028	0.044	0.915	0.085	0.873	0.840	0.043
DECENT	326	674	20281	2674	0.109	0.032	0.674	0.326	0.884	0.860	0.163
MAST	255	745	20210	2745	0.085	0.036	0.745	0.255	0.880	0.854	0.128
Monocle	283	717	20238	2717	0.094	0.034	0.717	0.283	0.882	0.857	0.142
NODES	105	895	20060	2895	0.035	0.043	0.895	0.105	0.874	0.842	0.053
scDD	72	928	20027	2928	0.024	0.044	0.928	0.072	0.872	0.839	0.036
BPSC	204	796	20161	2796	0.068	0.038	0.796	0.204	0.878	0.850	0.102
NDEG = 1500											
SwarnSeq	1442	58	20897	1558	0.481	0.003	0.039	0.961	0.931	0.933	0.641
DEGSeq	402	1098	19857	2598	0.134	0.052	0.732	0.268	0.884	0.846	0.179
DESeq2	263	1237	19722	2737	0.088	0.059	0.825	0.175	0.878	0.834	0.117
DESingle	371	1129	19826	2629	0.124	0.054	0.753	0.247	0.883	0.843	0.165
EdgeR	348	1152	19803	2652	0.116	0.055	0.768	0.232	0.882	0.841	0.155
Limma	123	1377	19583	2877	0.041	0.066	0.918	0.082	0.872	0.822	0.055
DECENT	438	1062	19893	2562	0.146	0.051	0.708	0.292	0.886	0.849	0.195
MAST	431	1069	19886	2569	0.144	0.051	0.713	0.287	0.886	0.848	0.192
Monocle	381	1119	19836	2619	0.127	0.053	0.746	0.254	0.883	0.844	0.169
NODES	146	1354	19601	2854	0.049	0.065	0.903	0.097	0.873	0.824	0.065
scDD	72	1428	19527	2928	0.024	0.068	0.952	0.048	0.870	0.818	0.032
BPSC	259	1241	19718	2741	0.086	0.059	0.827	0.173	0.878	0.834	0.115
NDEG = 2000											

SwarnSeq	1442	558	20397	1558	0.481	0.027	0.279	0.721	0.929	0.912	0.577
DEGSeq	503	1497	19458	2497	0.168	0.071	0.749	0.252	0.886	0.833	0.201
DESeq2	294	1706	19253	2706	0.098	0.081	0.853	0.147	0.877	0.816	0.118
DESingle	440	1560	19395	2560	0.147	0.074	0.780	0.220	0.883	0.828	0.176
EdgeR	423	1577	19378	2577	0.141	0.075	0.789	0.212	0.883	0.827	0.169
Limma	200	1800	19160	2800	0.067	0.086	0.900	0.100	0.872	0.808	0.080
DECENT	532	1468	19487	2468	0.177	0.070	0.734	0.266	0.888	0.836	0.213
MAST	591	1409	19546	2409	0.197	0.067	0.705	0.296	0.890	0.841	0.236
Monocle	473	1527	19428	2527	0.158	0.073	0.764	0.237	0.885	0.831	0.189
NODES	173	1827	19128	2827	0.058	0.087	0.914	0.087	0.871	0.806	0.069
scDD	72	1928	19027	2928	0.024	0.092	0.964	0.036	0.867	0.797	0.029
BPSC	298	1702	19257	2702	0.099	0.081	0.851	0.149	0.877	0.816	0.119
NDEG = 2500											
SwarnSeq	1442	1058	19897	1558	0.481	0.050	0.423	0.577	0.927	0.891	0.524
DEGSeq	589	1911	19044	2411	0.196	0.091	0.764	0.236	0.888	0.820	0.214
DESeq2	323	2177	18782	2677	0.108	0.104	0.871	0.129	0.875	0.797	0.117
DESingle	496	2004	18951	2504	0.165	0.096	0.802	0.198	0.883	0.812	0.180
EdgeR	488	2012	18943	2512	0.163	0.096	0.805	0.195	0.883	0.811	0.177
Limma	351	2149	18811	2649	0.117	0.103	0.860	0.140	0.877	0.800	0.128
DECENT	617	1883	19072	2383	0.206	0.090	0.753	0.247	0.889	0.822	0.224
MAST	750	1750	19205	2250	0.250	0.084	0.700	0.300	0.895	0.833	0.273
Monocle	553	1947	19008	2447	0.184	0.093	0.779	0.221	0.886	0.817	0.201
NODES	199	2301	18654	2801	0.066	0.110	0.920	0.080	0.869	0.787	0.072
scDD	72	2428	18527	2928	0.024	0.116	0.971	0.029	0.864	0.776	0.026
BPSC	326	2174	18785	2674	0.109	0.104	0.870	0.130	0.875	0.798	0.119
NDEG = 3000											
SwarnSeq	1442	1558	19397	1558	0.481	0.074	0.519	0.481	0.926	0.870	0.481
DEGSeq	719	2281	18674	2281	0.240	0.109	0.760	0.240	0.891	0.810	0.240
DESeq2	350	2650	18309	2650	0.117	0.126	0.883	0.117	0.874	0.779	0.117
DESingle	552	2448	18507	2448	0.184	0.117	0.816	0.184	0.883	0.796	0.184
EdgeR	545	2455	18500	2455	0.182	0.117	0.818	0.182	0.883	0.795	0.182
Limma	496	2504	18456	2504	0.165	0.119	0.835	0.165	0.881	0.791	0.165
DECENT	704	2296	18659	2296	0.235	0.110	0.765	0.235	0.890	0.808	0.235
MAST	906	2094	18861	2094	0.302	0.100	0.698	0.302	0.900	0.825	0.302
Monocle	632	2368	18587	2368	0.211	0.113	0.789	0.211	0.887	0.802	0.211
NODES	217	2783	18172	2783	0.072	0.133	0.928	0.072	0.867	0.768	0.072
scDD	72	2928	18027	2928	0.024	0.140	0.976	0.024	0.860	0.756	0.024
BPSC	364	2636	18323	2636	0.121	0.126	0.879	0.121	0.874	0.780	0.121

NDEG: Number of differentially expressed genes; TP: True Positive; FP: False Positive; TN: True Negative; FN: False Negative; TPR: True Positive Rate; FPR: False Positive Rate; FDR: False Discovery Rate; PPR: Positive Prediction Rate; NPV: Negative Prediction Value; ACC: Accuracy; F1: F-score

Table S15. Performance evaluation metrics for GSE109999 scRNA-seq data.

Methods	TP	FP	TN	FN	TPR	FPR	FDR	PPR	NPV	ACC	F1
NDEG = 500											
SwarnSeq	329	166	9888	2671	0.110	0.017	0.335	0.665	0.787	0.783	0.188
DEGSeq	175	325	9729	2825	0.058	0.032	0.650	0.350	0.775	0.759	0.100
DESingle	382	118	9936	2618	0.127	0.012	0.236	0.764	0.791	0.790	0.218
EdgeR	164	336	9735	2836	0.055	0.033	0.672	0.328	0.774	0.757	0.094
Limma	0	500	9591	3000	0.000	0.050	1.000	0.000	0.762	0.733	0.000
DECENT	207	293	9764	2793	0.069	0.029	0.586	0.414	0.778	0.764	0.118
MAST	212	288	9766	2788	0.071	0.029	0.576	0.424	0.778	0.764	0.121
Monocle	224	276	9778	2776	0.075	0.027	0.552	0.448	0.779	0.766	0.128
NODES	295	205	9849	2705	0.098	0.020	0.410	0.590	0.785	0.777	0.169
scDD	7	493	9561	2993	0.002	0.049	0.986	0.014	0.762	0.733	0.004
BPSC	369	131	9923	2631	0.123	0.013	0.262	0.738	0.790	0.788	0.211
NDEG = 1000											
SwarnSeq	556	433	9621	2444	0.185	0.043	0.438	0.562	0.797	0.780	0.279
DEGSeq	326	674	9383	2674	0.109	0.067	0.674	0.326	0.778	0.744	0.163
DESingle	627	373	9681	2373	0.209	0.037	0.373	0.627	0.803	0.790	0.314
EdgeR	331	669	9414	2669	0.110	0.066	0.669	0.331	0.779	0.745	0.166
Limma	0	1000	9146	3000	0.000	0.099	1.000	0.000	0.753	0.696	0.000
DECENT	336	664	9394	2664	0.112	0.066	0.664	0.336	0.779	0.745	0.168
MAST	366	634	9420	2634	0.122	0.063	0.634	0.366	0.781	0.750	0.183
Monocle	416	584	9472	2584	0.139	0.058	0.584	0.416	0.786	0.757	0.208
NODES	425	575	9479	2575	0.142	0.057	0.575	0.425	0.786	0.759	0.213
scDD	13	987	9067	2987	0.004	0.098	0.987	0.013	0.752	0.696	0.007
BPSC	588	412	9642	2412	0.196	0.041	0.412	0.588	0.800	0.784	0.294
NDEG = 1500											
SwarnSeq	716	768	9286	2284	0.239	0.076	0.518	0.482	0.803	0.766	0.319
DEGSeq	450	1050	9008	2550	0.150	0.104	0.700	0.300	0.779	0.724	0.200
DESingle	853	647	9407	2147	0.284	0.064	0.431	0.569	0.814	0.786	0.379
EdgeR	481	1019	9088	2519	0.160	0.101	0.679	0.321	0.783	0.730	0.214
Limma	0	1500	8786	3000	0.000	0.146	1.000	0.000	0.745	0.661	0.000
DECENT	452	1048	9023	2548	0.151	0.104	0.699	0.301	0.780	0.725	0.201
MAST	471	1029	9026	2529	0.157	0.102	0.686	0.314	0.781	0.727	0.209
Monocle	583	917	9151	2417	0.194	0.091	0.611	0.389	0.791	0.745	0.259
NODES	505	995	9059	2495	0.168	0.099	0.663	0.337	0.784	0.733	0.224
scDD	43	1457	8597	2957	0.014	0.145	0.971	0.029	0.744	0.662	0.019
BPSC	804	696	9358	2196	0.268	0.069	0.464	0.536	0.810	0.778	0.357
NDEG = 2000											
SwarnSeq	847	1136	8918	2153	0.282	0.113	0.573	0.427	0.806	0.748	0.340
DEGSeq	560	1440	8634	2440	0.187	0.143	0.720	0.280	0.780	0.703	0.224
DESingle	1053	947	9107	1947	0.351	0.094	0.474	0.527	0.824	0.778	0.421

EdgeR	618	1382	8773	2382	0.206	0.136	0.691	0.309	0.786	0.714	0.247
Limma	0	2000	8324	3000	0.000	0.194	1.000	0.000	0.735	0.625	0.000
DECENT	550	1450	8623	2450	0.183	0.144	0.725	0.275	0.779	0.702	0.220
MAST	559	1441	8616	2441	0.186	0.143	0.721	0.280	0.779	0.703	0.224
Monocle	760	1240	8839	2240	0.253	0.123	0.620	0.380	0.798	0.734	0.304
NODES	569	1431	8623	2431	0.190	0.142	0.716	0.285	0.780	0.704	0.228
scDD	65	1935	8119	2935	0.022	0.192	0.968	0.033	0.734	0.627	0.026
BPSC	986	1014	9040	2014	0.329	0.101	0.507	0.493	0.818	0.768	0.394
NDEG = 2500											
SwarnSeq	982	1500	8554	2018	0.327	0.149	0.604	0.396	0.809	0.731	0.358
DEGSeq	674	1826	8259	2326	0.225	0.181	0.730	0.270	0.780	0.683	0.245
DESingle	1220	1280	8777	1780	0.407	0.127	0.512	0.488	0.831	0.766	0.444
EdgeR	780	1720	8487	2220	0.260	0.169	0.688	0.312	0.793	0.702	0.284
Limma	10	2490	7948	2990	0.003	0.239	0.996	0.004	0.727	0.592	0.004
DECENT	663	1837	8243	2337	0.221	0.182	0.735	0.265	0.779	0.681	0.241
MAST	659	1841	8217	2341	0.220	0.183	0.736	0.264	0.778	0.680	0.240
Monocle	892	1608	8483	2108	0.297	0.159	0.643	0.357	0.801	0.716	0.324
NODES	652	1848	8207	2348	0.217	0.184	0.739	0.261	0.778	0.679	0.237
scDD	90	2410	7644	2910	0.030	0.240	0.964	0.036	0.724	0.592	0.033
BPSC	1167	1333	8721	1833	0.389	0.133	0.533	0.467	0.826	0.757	0.424
NDEG = 3000											
SwarnSeq	1125	1855	8199	1875	0.375	0.185	0.622	0.378	0.814	0.714	0.376
DEGSeq	806	2194	7893	2194	0.269	0.218	0.731	0.269	0.782	0.665	0.269
DESingle	1401	1599	8458	1599	0.467	0.159	0.533	0.467	0.841	0.755	0.467
EdgeR	950	2050	8194	2050	0.317	0.200	0.683	0.317	0.800	0.690	0.317
Limma	100	2900	7597	2900	0.033	0.276	0.967	0.033	0.724	0.570	0.033
DECENT	770	2230	7862	2230	0.257	0.221	0.743	0.257	0.779	0.659	0.257
MAST	753	2247	7812	2247	0.251	0.223	0.749	0.251	0.777	0.656	0.251
Monocle	1035	1965	8143	1965	0.345	0.194	0.655	0.345	0.806	0.700	0.345
NODES	701	2299	7765	2299	0.234	0.228	0.766	0.234	0.772	0.648	0.234
scDD	126	2874	7180	2874	0.042	0.286	0.958	0.042	0.714	0.560	0.042
BPSC	1319	1681	8373	1681	0.440	0.167	0.560	0.440	0.833	0.742	0.440

NDEG: Number of differentially expressed genes; TP: True Positive; FP: False Positive; TN: True Negative; FN: False Negative; TPR: True Positive Rate; FPR: False Positive Rate; FDR: False Discovery Rate; PPR: Positive Prediction Rate; NPV: Negative Prediction Value; ACC: Accuracy; F1: F-score

Table S16. Performance evaluation metrics for GSE111108 scRNA-seq data.

					NDEG = 500						
Methods	**TP**	**FP**	**TN**	**FN**	**TPR**	**FPR**	**FDR**	**PPR**	**NPV**	**ACC**	**F1**
SwarnSeq	495	5	14321	2506	0.165	0.000	0.010	0.990	0.851	0.855	0.283
DEGSeq	147	353	13973	2854	0.049	0.025	0.706	0.294	0.830	0.815	0.084
DESingle	12	488	13838	2989	0.004	0.034	0.976	0.024	0.822	0.799	0.007
EdgeR	55	445	13881	2946	0.018	0.031	0.890	0.110	0.825	0.804	0.031
Limma	120	380	13946	2881	0.040	0.027	0.760	0.240	0.829	0.812	0.069
DECENT	17	483	13843	2984	0.006	0.034	0.966	0.034	0.823	0.800	0.010
MAST	12	488	13838	2989	0.004	0.034	0.976	0.024	0.822	0.799	0.007
Monocle	118	382	13944	2883	0.039	0.027	0.764	0.236	0.829	0.812	0.067
NODES	43	457	13869	2958	0.014	0.032	0.914	0.086	0.824	0.803	0.025
scDD	22	478	13848	2979	0.007	0.033	0.956	0.044	0.823	0.800	0.013
BPSC	23	477	13849	2978	0.008	0.033	0.954	0.046	0.823	0.801	0.013
					NDEG = 1000						
SwarnSeq	995	5	14321	2006	0.332	0.000	0.005	0.995	0.877	0.884	0.497
DEGSeq	288	712	13614	2713	0.096	0.050	0.712	0.288	0.834	0.802	0.144
DESingle	12	988	13338	2989	0.004	0.069	0.988	0.012	0.817	0.770	0.006
EdgeR	107	893	13433	2894	0.036	0.062	0.893	0.107	0.823	0.781	0.053
Limma	227	773	13553	2774	0.076	0.054	0.773	0.227	0.830	0.795	0.113
DECENT	63	937	13389	2938	0.021	0.065	0.937	0.063	0.820	0.776	0.031
MAST	25	975	13351	2976	0.008	0.068	0.975	0.025	0.818	0.772	0.012
Monocle	234	766	13560	2767	0.078	0.053	0.766	0.234	0.831	0.796	0.117
NODES	64	936	13390	2937	0.021	0.065	0.936	0.064	0.820	0.776	0.032
scDD	43	957	13369	2958	0.014	0.067	0.957	0.043	0.819	0.774	0.021
BPSC	49	951	13375	2952	0.016	0.066	0.951	0.049	0.819	0.775	0.024
					NDEG = 1500						
SwarnSeq	1495	5	14321	1506	0.498	0.000	0.003	0.997	0.905	0.913	0.664
DEGSeq	414	1086	13240	2587	0.138	0.076	0.724	0.276	0.837	0.788	0.184
DESingle	12	1488	12838	2989	0.004	0.104	0.992	0.008	0.811	0.742	0.005
EdgeR	170	1330	12996	2831	0.057	0.093	0.887	0.113	0.821	0.760	0.076
Limma	322	1178	13148	2679	0.107	0.082	0.785	0.215	0.831	0.777	0.143
DECENT	122	1378	12948	2879	0.041	0.096	0.919	0.081	0.818	0.754	0.054
MAST	38	1462	12864	2963	0.013	0.102	0.975	0.025	0.813	0.745	0.017
Monocle	322	1178	13148	2679	0.107	0.082	0.785	0.215	0.831	0.777	0.143
NODES	74	1426	12900	2927	0.025	0.100	0.951	0.049	0.815	0.749	0.033
scDD	70	1430	12896	2931	0.023	0.100	0.953	0.047	0.815	0.748	0.031
BPSC	78	1422	12904	2923	0.026	0.099	0.948	0.052	0.815	0.749	0.035
					NDEG = 2000						
SwarnSeq	1995	5	14321	1006	0.665	0.000	0.003	0.998	0.934	0.942	0.798
DEGSeq	535	1465	12861	2466	0.178	0.102	0.733	0.268	0.839	0.773	0.214

DESingle	12	1988	12338	2989	0.004	0.139	0.994	0.006	0.805	0.713	0.005
EdgeR	222	1778	12548	2779	0.074	0.124	0.889	0.111	0.819	0.737	0.089
Limma	406	1594	12732	2595	0.135	0.111	0.797	0.203	0.831	0.758	0.162
DECENT	180	1820	12506	2821	0.060	0.127	0.910	0.090	0.816	0.732	0.072
MAST	60	1940	12386	2941	0.020	0.135	0.970	0.030	0.808	0.718	0.024
Monocle	398	1602	12724	2603	0.133	0.112	0.801	0.199	0.830	0.757	0.159
NODES	82	1918	12408	2919	0.027	0.134	0.959	0.041	0.810	0.721	0.033
scDD	90	1910	12416	2911	0.030	0.133	0.955	0.045	0.810	0.722	0.036
BPSC	108	1892	12434	2893	0.036	0.132	0.946	0.054	0.811	0.724	0.043
NDEG = 2500											
SwarnSeq	2495	5	14321	506	0.831	0.000	0.002	0.998	0.966	0.971	0.907
DEGSeq	674	1826	12500	2327	0.225	0.127	0.730	0.270	0.843	0.760	0.245
DESingle	12	2488	11838	2989	0.004	0.174	0.995	0.005	0.798	0.684	0.004
EdgeR	295	2205	12121	2706	0.098	0.154	0.882	0.118	0.817	0.717	0.107
Limma	481	2019	12307	2520	0.160	0.141	0.808	0.192	0.830	0.738	0.175
DECENT	250	2250	12076	2751	0.083	0.157	0.900	0.100	0.814	0.711	0.091
MAST	88	2412	11914	2913	0.029	0.168	0.965	0.035	0.804	0.693	0.032
Monocle	469	2031	12295	2532	0.156	0.142	0.812	0.188	0.829	0.737	0.171
NODES	96	2404	11922	2905	0.032	0.168	0.962	0.038	0.804	0.694	0.035
scDD	113	2387	11939	2888	0.038	0.167	0.955	0.045	0.805	0.696	0.041
BPSC	141	2359	11967	2860	0.047	0.165	0.944	0.056	0.807	0.699	0.051
NDEG = 3000											
SwarnSeq	2949	51	14275	52	0.983	0.004	0.017	0.983	0.996	0.994	0.983
DEGSeq	803	2197	12129	2198	0.268	0.153	0.732	0.268	0.847	0.746	0.268
DESingle	12	2988	11338	2989	0.004	0.209	0.996	0.004	0.791	0.655	0.004
EdgeR	360	2640	11686	2641	0.120	0.184	0.880	0.120	0.816	0.695	0.120
Limma	571	2429	11897	2430	0.190	0.170	0.810	0.190	0.830	0.720	0.190
DECENT	324	2676	11650	2677	0.108	0.187	0.892	0.108	0.813	0.691	0.108
MAST	125	2875	11451	2876	0.042	0.201	0.958	0.042	0.799	0.668	0.042
Monocle	543	2457	11869	2458	0.181	0.172	0.819	0.181	0.828	0.716	0.181
NODES	125	2875	11451	2876	0.042	0.201	0.958	0.042	0.799	0.668	0.042
scDD	134	2866	11460	2867	0.045	0.200	0.955	0.045	0.800	0.669	0.045
BPSC	173	2827	11499	2828	0.058	0.197	0.942	0.058	0.803	0.674	0.058

NDEG: Number of differentially expressed genes; TP: True Positive; FP: False Positive; TN: True Negative; FN: False Negative; TPR: True Positive Rate; FPR: False Positive Rate; FDR: False Discovery Rate; PPR: Positive Prediction Rate; NPV: Negative Prediction Value; ACC: Accuracy; F1: F-score

Table S17. Performance evaluation metrics for GSE115469 scRNA-seq data.

Methods	TP	FP	TN	FN	TPR	FPR	FDR	PPR	NPV	ACC	F1
NDEG = 500											
SwarnSeq	401	96	14218	2599	0.134	0.007	0.193	0.807	0.845	0.844	0.229
DEGSeq	182	318	13996	2818	0.061	0.022	0.636	0.364	0.832	0.819	0.104
DESingle	35	465	13849	2965	0.012	0.032	0.930	0.070	0.824	0.802	0.020
EdgeR	59	441	13873	2941	0.020	0.031	0.882	0.118	0.825	0.805	0.034
Limma	144	356	13958	2856	0.048	0.025	0.712	0.288	0.830	0.814	0.082
DECENT	480	20	14294	2520	0.160	0.001	0.040	0.960	0.850	0.853	0.274
MAST	93	407	13907	2907	0.031	0.028	0.814	0.186	0.827	0.809	0.053
Monocle	217	283	14031	2783	0.072	0.020	0.566	0.434	0.834	0.823	0.124
NODES	155	345	13969	2845	0.052	0.024	0.690	0.310	0.831	0.816	0.089
scDD	12	488	13826	2988	0.004	0.034	0.976	0.024	0.822	0.799	0.007
NDEG = 1000											
SwarnSeq	786	206	14108	2214	0.262	0.014	0.208	0.792	0.864	0.860	0.394
DEGSeq	345	655	13659	2655	0.115	0.046	0.655	0.345	0.837	0.809	0.173
DESingle	125	875	13439	2875	0.042	0.061	0.875	0.125	0.824	0.783	0.063
EdgeR	196	804	13510	2804	0.065	0.056	0.804	0.196	0.828	0.792	0.098
Limma	252	748	13566	2748	0.084	0.052	0.748	0.252	0.832	0.798	0.126
DECENT	834	166	14148	2166	0.278	0.012	0.166	0.834	0.867	0.865	0.417
MAST	252	748	13566	2748	0.084	0.052	0.748	0.252	0.832	0.798	0.126
Monocle	440	560	13754	2560	0.147	0.039	0.560	0.440	0.843	0.820	0.220
NODES	301	699	13615	2699	0.100	0.049	0.699	0.301	0.835	0.804	0.151
scDD	17	983	13331	2983	0.006	0.069	0.983	0.017	0.817	0.771	0.009
NDEG = 1500											
SwarnSeq	1027	462	13852	1973	0.342	0.032	0.310	0.690	0.875	0.859	0.458
DEGSeq	523	977	13337	2477	0.174	0.068	0.651	0.349	0.843	0.801	0.232
DESingle	251	1249	13065	2749	0.084	0.087	0.833	0.167	0.826	0.769	0.112
EdgeR	345	1155	13159	2655	0.115	0.081	0.770	0.230	0.832	0.780	0.153
Limma	367	1133	13181	2633	0.122	0.079	0.755	0.245	0.834	0.782	0.163
DECENT	1063	437	13877	1937	0.354	0.031	0.291	0.709	0.878	0.863	0.472
MAST	396	1104	13210	2604	0.132	0.077	0.736	0.264	0.835	0.786	0.176
Monocle	615	885	13429	2385	0.205	0.062	0.590	0.410	0.849	0.811	0.273
NODES	445	1055	13259	2555	0.148	0.074	0.703	0.297	0.838	0.791	0.198
scDD	32	1468	12846	2968	0.011	0.103	0.979	0.021	0.812	0.744	0.014
NDEG = 2000											
SwarnSeq	1194	791	13523	1806	0.398	0.055	0.398	0.602	0.882	0.850	0.479
DEGSeq	715	1285	13029	2285	0.238	0.090	0.643	0.358	0.851	0.794	0.286
DESingle	407	1593	12721	2593	0.136	0.111	0.797	0.204	0.831	0.758	0.163
EdgeR	530	1470	12844	2470	0.177	0.103	0.735	0.265	0.839	0.772	0.212
Limma	496	1504	12810	2504	0.165	0.105	0.752	0.248	0.836	0.769	0.198

DECENT	1229	771	13543	1771	0.410	0.054	0.386	0.615	0.884	0.853	0.492
MAST	535	1465	12849	2465	0.178	0.102	0.733	0.268	0.839	0.773	0.214
Monocle	806	1194	13120	2194	0.269	0.083	0.597	0.403	0.857	0.804	0.322
NODES	606	1394	12920	2394	0.202	0.097	0.697	0.303	0.844	0.781	0.242
scDD	40	1960	12354	2960	0.013	0.137	0.980	0.020	0.807	0.716	0.016
NDEG = 2500											
SwarnSeq	1335	1147	13167	1665	0.445	0.080	0.462	0.538	0.888	0.838	0.487
DEGSeq	870	1630	12684	2130	0.290	0.114	0.652	0.348	0.856	0.783	0.316
DESingle	579	1921	12393	2421	0.193	0.134	0.768	0.232	0.837	0.749	0.211
EdgeR	689	1811	12503	2311	0.230	0.127	0.724	0.276	0.844	0.762	0.251
Limma	620	1880	12434	2380	0.207	0.131	0.752	0.248	0.839	0.754	0.225
DECENT	1358	1142	13172	1642	0.453	0.080	0.457	0.543	0.889	0.839	0.494
MAST	662	1838	12476	2338	0.221	0.128	0.735	0.265	0.842	0.759	0.241
Monocle	977	1523	12791	2023	0.326	0.106	0.609	0.391	0.863	0.795	0.355
NODES	747	1753	12561	2253	0.249	0.122	0.701	0.299	0.848	0.769	0.272
scDD	56	2444	11870	2944	0.019	0.171	0.978	0.022	0.801	0.689	0.020
NDEG = 3000											
SwarnSeq	1460	1520	12794	1540	0.487	0.106	0.510	0.490	0.893	0.823	0.488
DEGSeq	1057	1943	12371	1943	0.352	0.136	0.648	0.352	0.864	0.776	0.352
DESingle	731	2269	12045	2269	0.244	0.159	0.756	0.244	0.841	0.738	0.244
EdgeR	856	2144	12170	2144	0.285	0.150	0.715	0.285	0.850	0.752	0.285
Limma	732	2268	12046	2268	0.244	0.158	0.756	0.244	0.842	0.738	0.244
DECENT	1477	1523	12791	1523	0.492	0.106	0.508	0.492	0.894	0.824	0.492
MAST	758	2242	12072	2242	0.253	0.157	0.747	0.253	0.843	0.741	0.253
Monocle	1113	1887	12427	1887	0.371	0.132	0.629	0.371	0.868	0.782	0.371
NODES	935	2065	12249	2065	0.312	0.144	0.688	0.312	0.856	0.761	0.312
scDD	66	2934	11380	2934	0.022	0.205	0.978	0.022	0.795	0.661	0.022

NDEG: Number of differentially expressed genes; TP: True Positive; FP: False Positive; TN: True Negative; FN: False Negative; TPR: True Positive Rate; FPR: False Positive Rate; FDR: False Discovery Rate; PPR: Positive Prediction Rate; NPV: Negative Prediction Value; ACC: Accuracy; F1: F-score

Table S18. Performance evaluation metrics for Tung's (GSE77288) data.

Methods	TP	FP	TN	FN	TPR	FPR	FDR	PPR	NPV	ACC	F1
NDEG = 500											
SwarnSeq	482	18	12937	2518	0.161	0.001	0.036	0.964	0.837	0.841	0.275
DEGSeq	107	393	12570	2893	0.036	0.030	0.786	0.214	0.813	0.794	0.061
DESeq2	134	366	12589	2866	0.045	0.028	0.732	0.268	0.815	0.797	0.077
DESingle	227	273	12682	2773	0.076	0.021	0.546	0.454	0.821	0.809	0.130
EdgeR	199	301	12660	2801	0.066	0.023	0.602	0.398	0.819	0.806	0.114
Limma	292	208	12747	2708	0.097	0.016	0.416	0.584	0.825	0.817	0.167
DECENT	172	328	12627	2828	0.057	0.025	0.656	0.344	0.817	0.802	0.098
MAST	159	341	12614	2841	0.053	0.026	0.682	0.318	0.816	0.801	0.091
Monocle	96	404	12551	2904	0.032	0.031	0.808	0.192	0.812	0.793	0.055
NODES	187	313	12642	2813	0.062	0.024	0.626	0.374	0.818	0.804	0.107
scDD	141	359	12596	2859	0.047	0.028	0.718	0.282	0.815	0.798	0.081
BPSC	114	386	12569	2886	0.038	0.030	0.772	0.228	0.813	0.795	0.065
NDEG = 1000											
SwarnSeq	946	54	12901	2054	0.315	0.004	0.054	0.946	0.863	0.868	0.473
DEGSeq	159	841	12136	2841	0.053	0.065	0.841	0.159	0.810	0.770	0.080
DESeq2	234	766	12189	2766	0.078	0.059	0.766	0.234	0.815	0.779	0.117
DESingle	426	574	12381	2574	0.142	0.044	0.574	0.426	0.828	0.803	0.213
EdgeR	347	653	12398	2653	0.116	0.050	0.653	0.347	0.824	0.794	0.174
Limma	607	393	12562	2393	0.202	0.030	0.393	0.607	0.840	0.825	0.304
DECENT	351	649	12306	2649	0.117	0.050	0.649	0.351	0.823	0.793	0.176
MAST	330	670	12285	2670	0.110	0.052	0.670	0.330	0.821	0.791	0.165
Monocle	263	737	12218	2737	0.088	0.057	0.737	0.263	0.817	0.782	0.132
NODES	355	645	12310	2645	0.118	0.050	0.645	0.355	0.823	0.794	0.178
scDD	218	782	12173	2782	0.073	0.060	0.782	0.218	0.814	0.777	0.109
BPSC	213	787	12168	2787	0.071	0.061	0.787	0.213	0.814	0.776	0.107
NDEG = 1500											
SwarnSeq	1311	189	12766	1689	0.437	0.015	0.126	0.874	0.883	0.882	0.583
DEGSeq	293	1207	11784	2707	0.098	0.093	0.805	0.195	0.813	0.755	0.130
DESeq2	338	1162	11793	2662	0.113	0.090	0.775	0.225	0.816	0.760	0.150
DESingle	612	888	12067	2388	0.204	0.069	0.592	0.408	0.835	0.795	0.272
EdgeR	486	1014	12140	2514	0.162	0.077	0.676	0.324	0.828	0.782	0.216
Limma	887	613	12342	2113	0.296	0.047	0.409	0.591	0.854	0.829	0.394
DECENT	518	982	11973	2482	0.173	0.076	0.655	0.345	0.828	0.783	0.230
MAST	497	1003	11952	2503	0.166	0.077	0.669	0.331	0.827	0.780	0.221
Monocle	398	1102	11853	2602	0.133	0.085	0.735	0.265	0.820	0.768	0.177
NODES	523	977	11978	2477	0.174	0.075	0.651	0.349	0.829	0.784	0.232
scDD	279	1221	11734	2721	0.093	0.094	0.814	0.186	0.812	0.753	0.124
BPSC	321	1179	11776	2679	0.107	0.091	0.786	0.214	0.815	0.758	0.143
NDEG = 2000											

SwarnSeq	1532	468	12487	1468	0.511	0.036	0.234	0.766	0.895	0.879	0.613
DEGSeq	420	1580	11426	2580	0.140	0.121	0.790	0.210	0.816	0.740	0.168
DESeq2	426	1574	11381	2574	0.142	0.121	0.787	0.213	0.816	0.740	0.170
DESingle	758	1242	11714	2242	0.253	0.096	0.621	0.379	0.839	0.782	0.303
EdgeR	625	1375	11872	2375	0.208	0.104	0.688	0.313	0.833	0.769	0.250
Limma	1104	896	12059	1896	0.368	0.069	0.448	0.552	0.864	0.825	0.442
DECENT	664	1336	11632	2336	0.221	0.103	0.668	0.332	0.833	0.770	0.266
MAST	642	1358	11597	2358	0.214	0.105	0.679	0.321	0.831	0.767	0.257
Monocle	491	1509	11446	2509	0.164	0.116	0.755	0.246	0.820	0.748	0.196
NODES	636	1364	11591	2364	0.212	0.105	0.682	0.318	0.831	0.766	0.254
scDD	327	1673	11282	2673	0.109	0.129	0.837	0.164	0.808	0.728	0.131
BPSC	416	1584	11371	2584	0.139	0.122	0.792	0.208	0.815	0.739	0.166
NDEG = 2500											
SwarnSeq	1713	787	12168	1287	0.571	0.061	0.315	0.685	0.904	0.870	0.623
DEGSeq	514	1986	11043	2486	0.171	0.152	0.794	0.206	0.816	0.721	0.187
DESeq2	525	1975	10980	2475	0.175	0.152	0.790	0.210	0.816	0.721	0.191
DESingle	901	1599	11359	2099	0.300	0.123	0.640	0.360	0.844	0.768	0.328
EdgeR	769	1731	11582	2231	0.256	0.130	0.692	0.308	0.838	0.757	0.280
Limma	1293	1207	11748	1707	0.431	0.093	0.483	0.517	0.873	0.817	0.470
DECENT	791	1709	11305	2209	0.264	0.131	0.684	0.316	0.837	0.755	0.288
MAST	772	1728	11227	2228	0.257	0.133	0.691	0.309	0.834	0.752	0.281
Monocle	598	1902	11053	2402	0.199	0.147	0.761	0.239	0.821	0.730	0.217
NODES	752	1748	11207	2248	0.251	0.135	0.699	0.301	0.833	0.750	0.273
scDD	394	2106	10849	2606	0.131	0.163	0.842	0.158	0.806	0.705	0.143
BPSC	514	1986	10969	2486	0.171	0.153	0.794	0.206	0.815	0.720	0.187
NDEG = 3000											
SwarnSeq	1846	1154	11801	1154	0.615	0.089	0.385	0.615	0.911	0.855	0.615
DEGSeq	611	2389	10657	2389	0.204	0.183	0.796	0.204	0.817	0.702	0.204
DESeq2	629	2371	10584	2371	0.210	0.183	0.790	0.210	0.817	0.703	0.210
DESingle	1024	1976	10986	1976	0.341	0.152	0.659	0.341	0.848	0.752	0.341
EdgeR	909	2091	11290	2091	0.303	0.156	0.697	0.303	0.844	0.745	0.303
Limma	1456	1544	11412	1544	0.485	0.119	0.515	0.485	0.881	0.806	0.485
DECENT	901	2099	10997	2099	0.300	0.160	0.700	0.300	0.840	0.739	0.300
MAST	886	2114	10841	2114	0.295	0.163	0.705	0.295	0.837	0.735	0.295
Monocle	698	2302	10655	2302	0.233	0.178	0.767	0.233	0.822	0.711	0.233
NODES	836	2164	10791	2164	0.279	0.167	0.721	0.279	0.833	0.729	0.279
scDD	461	2539	10416	2539	0.154	0.196	0.846	0.154	0.804	0.682	0.154
BPSC	614	2386	10569	2386	0.205	0.184	0.795	0.205	0.816	0.701	0.205

NDEG: Number of differentially expressed genes; TP: True Positive; FP: False Positive; TN: True Negative; FN: False Negative; TPR: True Positive Rate; FPR: False Positive Rate; FDR: False Discovery Rate; PPR: Positive Prediction Rate; NPV: Negative Prediction Value; ACC: Accuracy; F1: F-score

Table S19. List of differentially zero inflated genes for all the ten scRNA-seq datasets. (given in separate .xlsx file)

16. Supplementary Figures

Figure S12. Differential expression (DE) analysis of real scRNA-seq data. Receiver operating characteristic curves for differential expression methods on different real scRNA-seq data. Evaluation of the performance of different methods based on Area Under Receiver Operating Characteristic Curves (AUC) is shown for (A) GSE111108; (B) GSE92495; (C) GSE115469; (D) GSE109999. Different goldstandard gene lists are prepared based on the fold change values for benchmarking different differential expression analysis methods on different real scRNA-seq datasets. Swarnseq achieves competitive and better accuarcy for identifying genuine differential gene lists in all four different real datasets. DE methods are denoted by different colors.

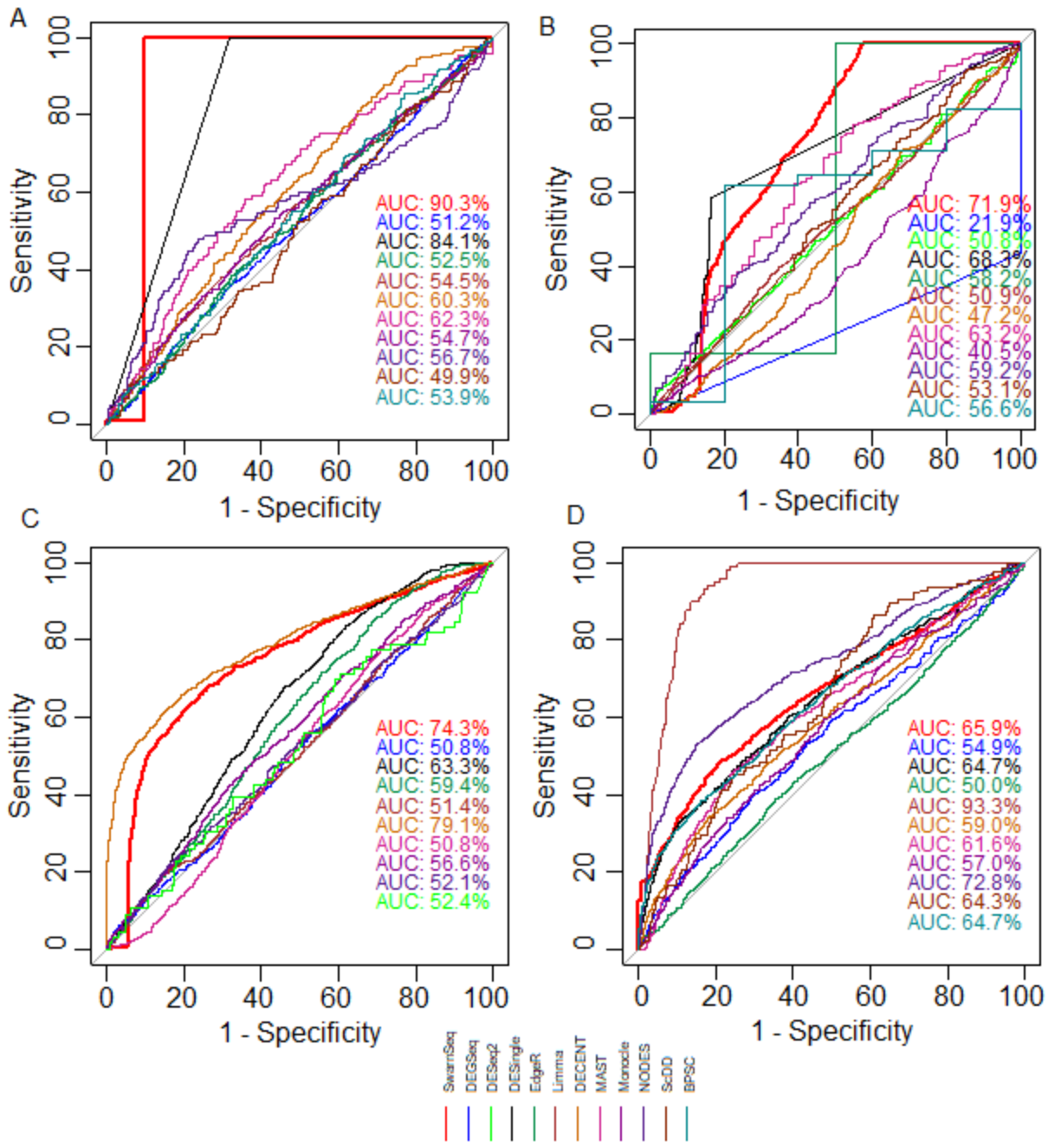
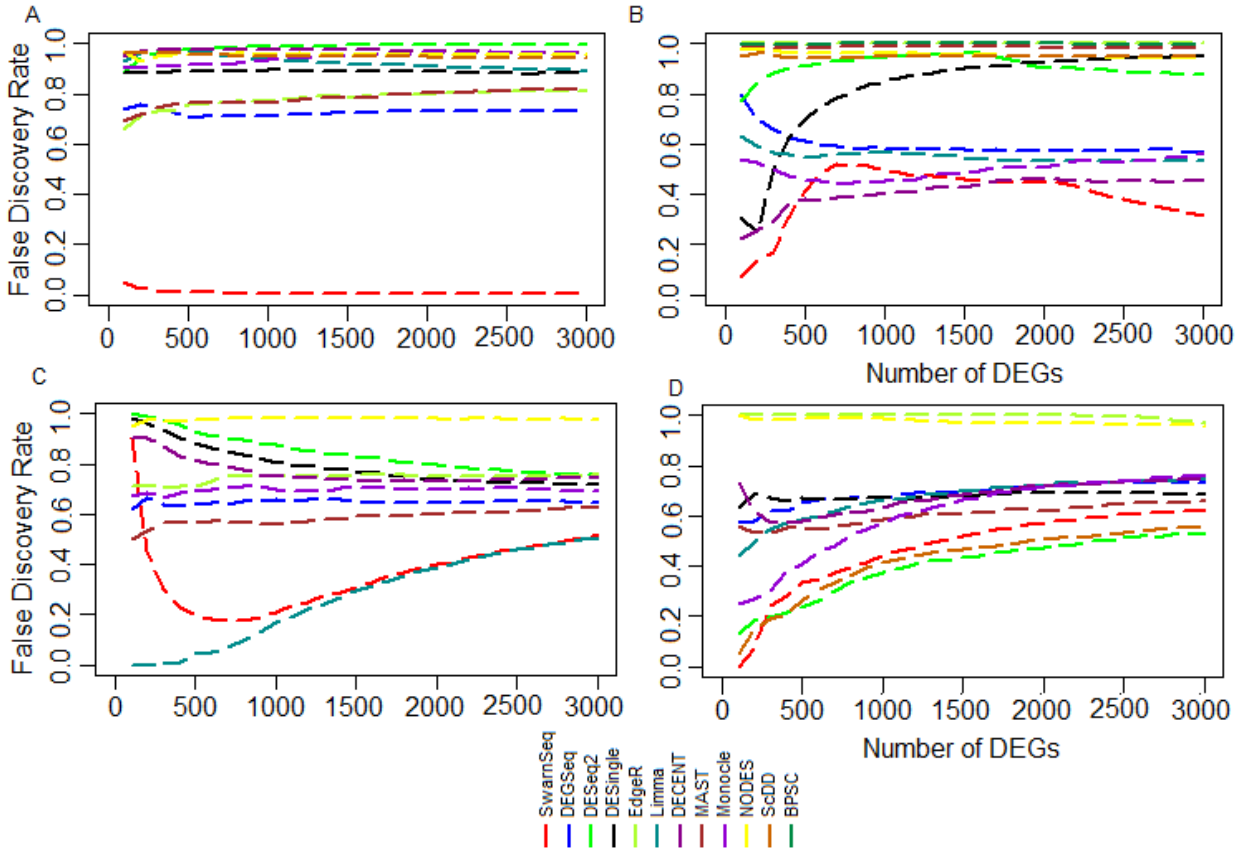

A
Sensitivity
1 - Specificity
0
20
40
60
80
100
AUC: 90.3%
AUC: 51.2%
AUC: 84.1%
AUC: 52.5%
AUC: 54.5%
AUC: 60.3%
AUC: 62.3%
AUC: 54.7%
AUC: 56.7%
AUC: 49.9%
AUC: 53.9%
B
Sensitivity
1 - Specificity
AUC: 71.9%
AUC: 21.9%
AUC: 50.8%
AUC: 68.3%
AUC: 58.2%
AUC: 50.9%
AUC: 47.2%
AUC: 63.2%
AUC: 40.5%
AUC: 59.2%
AUC: 53.1%
AUC: 56.6%
C
Sensitivity
1 - Specificity
AUC: 74.3%
AUC: 50.8%
AUC: 63.3%
AUC: 59.4%
AUC: 51.4%
AUC: 79.1%
AUC: 50.8%
AUC: 56.6%
AUC: 52.1%
AUC: 52.4%
D
Sensitivity
1 - Specificity
AUC: 65.9%
AUC: 54.9%
AUC: 64.7%
AUC: 50.0%
AUC: 93.3%
AUC: 59.0%
AUC: 61.6%
AUC: 57.0%
AUC: 72.8%
AUC: 64.3%
AUC: 64.7%
SwarmSeq
DEGSeq
DESeq2
DESingle
EdgeR
Limma
DECENT
MAST
Monocle
NODES
ScDD
BPSC

Figure S13. FDR based Performance analysis of DE methods on real scRNA-seq data. FDR curves for differential expression methods on different real scRNA-seq data are shown. Evaluation of the performance of different methods based on false discovery rate is shown for (A) GSE111108; (B) GSE92495; (C) GSE115469; (D) GSE109999.



17. References

1. Love MI, Huber W, Anders S. Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. Genome Biol. 2014. doi:10.1186/s13059-014-0550-8
2. Love MI, Anders S, Huber W. Differential analysis of count data - the DESeq2 package. Genome Biology. 2014. doi:110.1186/s13059-014-0550-8
3. Robinson MD, McCarthy DJ, Smyth GK. edgeR: a Bioconductor package for differential expression analysis of digital gene expression data. Bioinformatics. 2010;26: 139–140. doi:10.1093/bioinformatics/btp616
4. Van den Berge K, Perraudeau F, Soneson C, Love MI, Risso D, Vert J-P, et al. Observation weights unlock bulk RNA-seq tools for zero inflation and single-cell applications. Genome Biol. 2018;19: 24. doi:10.1186/s13059-018-1406-4
5. Risso D, Perraudeau F, Gribkova S, Dudoit S, Vert J-P. A general and flexible method for signal extraction from single-cell RNA-seq data. Nat Commun. 2018;9: 284. doi:10.1038/s41467-017-02554-5
6. Van den Berge K, Soneson C, Love MI, Robinson MD, Clement L. zingeR: unlocking RNA-seq tools for zero-inflation and single cell applications. doi.org. 2017. doi:10.1101/157982
7. Miao Z, Deng K, Wang X, Zhang X. DEsingle for detecting three types of differential expression in single-cell RNA-seq data. Berger B, editor. Bioinformatics. 2018;34: 3223–3224. doi:10.1093/bioinformatics/bty332
8. Ye C, Speed TP, Salim A. DECENT: differential expression with capture efficiency adjustmeNT for single-cell RNA-seq data. Berger B, editor. Bioinformatics. 2019;35: 5155–5162. doi:10.1093/bioinformatics/btz453

9. Zeileis A, Kleiber C, Jackman S. Regression models for count data in R. J Stat Softw. 2008. doi:10.18637/jss.v027.i08

10. KEMP CD, KEMP AW. Some properties of the “Hermite” distribution. Biometrika. 1965;52: 381–394. doi:10.1093/biomet/52.3-4.381

11. Islam S, Kjällquist U, Moliner A, Zajac P, Fan JB, Lönnerberg P, et al. Characterization of the single-cell transcriptional landscape by highly multiplex RNA-seq. Genome Res. 2011. doi:10.1101/gr.110882.110

12. Boon WC, Petkovic-Duran K, Zhu Y, Manasseh R, Horne MK, Aumann TD. Increasing cDNA yields from single-cell quantities of mRNA in standard laboratory reverse transcriptase reactions using acoustic microstreaming. J Vis Exp. 2011. doi:10.3791/3144

13. Macaulay IC, Voet T. Single Cell Genomics: Advances and Future Perspectives. PLoS Genetics. 2014. doi:10.1371/journal.pgen.1004126

14. Marinov GK, Williams BA, McCue K, Schroth GP, Gertz J, Myers RM, et al. From single-cell to cell-pool transcriptomes: Stochasticity in gene expression and RNA splicing. Genome Res. 2014. doi:10.1101/gr.161034.113

15. Kharchenko P V., Silberstein L, Scadden DT. Bayesian approach to single-cell differential expression analysis. Nat Methods. 2014;11: 740–742. doi:10.1038/nmeth.2967

16. Pierson E, Yau C. ZIFA: Dimensionality reduction for zero-inflated single-cell gene expression analysis. Genome Biol. 2015. doi:10.1186/s13059-015-0805-z

17. Wang Y, Navin NE. Advances and Applications of Single-Cell Sequencing Technologies. Molecular Cell. 2015. doi:10.1016/j.molcel.2015.05.005

18. Chakraborty S. On some distributional properties of the family of weighted generalized poisson distribution. Commun Stat - Theory Methods. 2010.

doi:10.1080/03610920903129141

19. Satheesh Kumar C, Ramachandran R. On some aspects of a zero-inflated overdispersed model and its applications. J Appl Stat. 2020. doi:10.1080/02664763.2019.1645098

20. McElduff F, Cortina-Borja M, Chan SK, Wade A. When t-tests or Wilcoxon-Mann-Whitney tests won't do. Am J Physiol - Adv Physiol Educ. 2010. doi:10.1152/advan.00017.2010

21. Hardcastle TJ, Kelly KA. BaySeq: Empirical Bayesian methods for identifying differential expression in sequence count data. BMC Bioinformatics. 2010. doi:10.1186/1471-2105-11-422

22. Trapnell C, Hendrickson DG, Sauvageau M, Goff L, Rinn JL, Pachter L. Differential analysis of gene regulation at transcript resolution with RNA-seq. Nat Biotechnol. 2013. doi:10.1038/nbt.2450

23. Jackman S, Tahk A, Zeileis A, Maimone C, Fearon J, Maintainer ZM. Package "pscl": Political Science Computational Laboratory. R Packag Vignettes Collect. 2017.

24. Petropoulos S, Edsgärd D, Reinius B, Deng Q, Panula SP, Codeluppi S, et al. Single-Cell RNA-Seq Reveals Lineage and X Chromosome Dynamics in Human Preimplantation Embryos. Cell. 2016. doi:10.1016/j.cell.2016.03.023

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

26. Robinson MD, Oshlack A. A scaling normalization method for differential expression analysis of RNA-seq data. Genome Biol. 2010. doi:10.1186/gb-2010-11-3-r25

27. Tian L, Su S, Dong X, Amann-Zalcenstein D, Biben C, Seidi A, et al. scPipe: A flexible

R/Bioconductor preprocessing pipeline for single-cell RNA-sequencing data. PLoS Comput Biol. 2018. doi:10.1371/journal.pcbi.1006361

28. Tung P-Y, Blischak JD, Hsiao CJ, Knowles DA, Burnett JE, Pritchard JK, et al. Batch effects and the effective design of single-cell gene expression studies. Sci Rep. 2017;7: 39921. doi:10.1038/srep39921

29. MacParland SA, Liu JC, Ma XZ, Innes BT, Bartczak AM, Gage BK, et al. Single cell RNA sequencing of human liver reveals distinct intrahepatic macrophage populations. Nat Commun. 2018. doi:10.1038/s41467-018-06318-7

30. Soumillon M, Cacchiarelli D, Semrau S, van Oudenaarden A, Mikkelsen TS. Characterization of directed differentiation by high-throughput single-cell RNA-Seq. bioRxiv. 2014. doi:10.1101/003236

31. Klein AM, Mazutis L, Akartuna I, Tallapragada N, Veres A, Li V, et al. Droplet barcoding for single-cell transcriptomics applied to embryonic stem cells. Cell. 2015. doi:10.1016/j.cell.2015.04.044

32. Gierahn TM, Wadsworth MH, Hughes TK, Bryson BD, Butler A, Satija R, et al. Seq-Well: Portable, low-cost rna sequencing of single cells at high throughput. Nat Methods. 2017. doi:10.1038/nmeth.4179

33. Moliner A, Enfors P, Ibáñez CF, Andäng M. Mouse embryonic stem cell-derived spheres with distinct neurogenic potentials. Stem Cells Dev. 2008. doi:10.1089/scd.2007.0211

34. Liang Y, Zhang F, Wang J, Joshi T, Wang Y, Xu D. Prediction of Drought-Resistant Genes in Arabidopsis thaliana Using SVM-RFE. Zhu D, editor. PLoS One. 2011;6: e21750. doi:10.1371/journal.pone.0021750

35. Das S, Meher PK, Rai A, Bhar LM, Mandal BN. Statistical Approaches for Gene

Selection, Hub Gene Identification and Module Interaction in Gene Co-Expression Network Analysis: An Application to Aluminum Stress in Soybean (Glycine max L.). Tian Z, editor. PLoS One. 2017;12: e0169605. doi:10.1371/journal.pone.0169605

36. Das S, Pandey P, Rai A, Mohapatra C. A computational system biology approach to construct gene regulatory networks for salinity response in rice (Oryza sativa). Indian J Agric Sci. 2015;85: 1546–1552.

37. Das S, McClain CJ, Rai SN. Fifteen Years of Gene Set Analysis for High-Throughput Genomic Data: A Review of Statistical Approaches and Future Challenges. Entropy. 2020;22: 427. doi:10.3390/e22040427

38. Das S, Rai SN. Statistical approach for biologically relevant gene selection from high-throughput gene expression data. Entropy. 2020;22. doi:10.3390/e22111205

39. Das S, Rai A, Mishra DC, Rai SN. Statistical Approach for Gene Set Analysis with Trait Specific Quantitative Trait Loci. Sci Rep. 2018;8: 2391. doi:10.1038/s41598-018-19736-w

40. Das S, Rai A, Mishra DC, Rai SN. Statistical approach for selection of biologically informative genes. Gene. 2018;655. doi:10.1016/j.gene.2018.02.044

41. Leng N, Dawson JA, Thomson JA, Ruotti V, Rissman AI, Smits BMG, et al. EBSeq: an empirical Bayes hierarchical model for inference in RNA-seq experiments. Bioinformatics. 2013;29: 1035–1043. doi:10.1093/bioinformatics/btt087

42. Anders S, Huber W. Differential expression analysis for sequence count data. Genome Biol. 2010;11: R106. doi:10.1186/gb-2010-11-10-r106

43. Wang L, Feng Z, Wang X, Wang X, Zhang X. DEGseq: An R package for identifying differentially expressed genes from RNA-seq data. Bioinformatics. 2009.

doi:10.1093/bioinformatics/btp612

44. Smyth GK. limma: Linear Models for Microarray Data. Bioinformatics and Computational Biology Solutions Using R and Bioconductor. doi:10.1007/0-387-29362-0_23

45. Qiu X, Hill A, Packer J, Lin D, Ma Y-A, Trapnell C. Single-cell mRNA quantification and differential analysis with Census. Nat Methods. 2017;14: 309–315. doi:10.1038/nmeth.4150

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

47. Korthauer KD, Chu LF, Newton MA, Li Y, Thomson J, Stewart R, et al. A statistical approach for identifying differential distributions in single-cell RNA-seq experiments. Genome Biol. 2016. doi:10.1186/s13059-016-1077-y

48. Vu TN, Wills QF, Kalari KR, Niu N, Wang L, Rantalainen M, et al. Beta-Poisson model for single-cell RNA-seq data analyses. Bioinformatics. 2016. doi:10.1093/bioinformatics/btw202

49. Sengupta D, Rayan NA, Lim M, Lim B, Prabhakar S. Fast, scalable and accurate differential expression analysis for single cells. bioRxiv. 2016. doi:10.1101/049734

50. Ritchie ME, Phipson B, Wu D, Hu Y, Law CW, Shi W, et al. limma powers differential expression analyses for RNA-sequencing and microarray studies. Nucleic Acids Res. 2015;43: e47–e47. doi:10.1093/nar/gkv007

51. Trapnell C, Cacchiarelli D, Grimsby J, Pokharel P, Li S, Morse M, et al. The dynamics

and regulators of cell fate decisions are revealed by pseudotemporal ordering of single cells. Nat Biotechnol. 2014. doi:10.1038/nbt.2859

52. Qiu X, Mao Q, Tang Y, Wang L, Chawla R, Pliner HA, et al. Reversed graph embedding resolves complex single-cell trajectories. Nat Methods. 2017. doi:10.1038/nmeth.4402

48. Das, Samarendra (2020). "Statistical approaches of gene set analysis with quantitative trait loci for highthroughput genomic studies." Electronic Theses and Dissertations. Paper 3537. doi.org/10.18297/etd/3537