Genomics 113 (2021) 1308–1324

Contents lists available at ScienceDirect

Genomics

journal homepage: www.elsevier.com/locate/ygeno

Original Article

SwarnSeq: An improved statistical approach for differential expression analysis of single-cell RNA-seq data

Samarendra Das [a,b,c], Shesh N. Rai [b,c,d,e,f,g,*]

[a] Division of Statistical Genetics, ICAR-Indian Agricultural Statistics Research Institute, PUSA, New Delhi 110012, India
[b] Biostatistics and Bioinformatics Facility, JG Brown Cancer Center, University of Louisville, Louisville, KY 40202, USA
[c] School of Interdisciplinary and Graduate Studies, University of Louisville, Louisville, KY 40292, USA
[d] Hepatobiology and Toxicology Center, University of Louisville, Louisville, KY 40202, USA
[e] Department of Bioinformatics and Biostatistics, University of Louisville, Louisville, KY 40202, USA
[f] Biostatistics and Informatics Facility, Center for Integrative Environmental Research Sciences, University of Louisville, Louisville, KY 40202, USA
[g] Christina Lee Brown Envirome Institute, University of Louisville, Louisville, KY 40202, USA

ARTICLE INFO

Keywords:
SwarnSeq
scRNA-seq
Zero inflated negative binomial
Dispersion
Differential expression
Capture rates

ABSTRACT

Single-cell RNA sequencing (scRNA-seq) is a powerful technology that is capable of generating gene expression data at the resolution of individual cell. The scRNA-seq data is characterized by the presence of dropout events, which severely bias the results if they remain unaddressed. There are limited Differential Expression (DE) approaches which consider the biological processes, which lead to dropout events, in the modeling process. So, we develop, SwarnSeq, an improved method for DE, and other downstream analysis that considers the molecular capture process in scRNA-seq data modeling. The performance of the proposed method is benchmarked with 11 existing methods on 10 different real scRNA-seq datasets under three comparison settings. We demonstrate that SwarnSeq method has improved performance over the 11 existing methods. This improvement is consistently observed across several public scRNA-seq datasets generated using different scRNA-seq protocols. The external spike-ins data can be used in the SwarnSeq method to enhance its performance.
Availability and implementation: The method is implemented as a publicly available R package available at https://github.com/sam-uofl/SwarnSeq.

1. Background

The advent of single-cell RNA sequencing (scRNA-seq) technology revolutionized transcriptomics through generating gene expression data at the single cell resolution level [1,2]. It has numerous advantages over bulk RNA-seq technology, which only characterize the global expression dynamics of genes in a tissue sample, while ignoring the inherent cell-cell heterogeneity [3,4]. Thus, it is pertinent to assess the variability that exists among the cells in a tissue sample as this is crucial to understand the complexity and dynamics of biological processes such as embryogenesis [1,5], cancer [6], *etc.* Through scRNA-seq technology, expression is quantified by mapping reads to a reference genome followed by counting the number of reads mapped to each gene [1]. Here, individual transcript molecules are attached with a Unique Molecular Identifier (UMI) tag; subsequently, counting the UMIs usually yields the number of transcripts for each gene in a cell [7]. Further, huge amounts of UMI count data are generated for several thousand(s) of genes across thousand(s) of cells and subsequently deposited in public domain databases by researchers across the globe. Hence, it is necessary to develop new, and innovative statistical approaches and tools for such data analysis to harness the potential of this new technology.

Small amounts of the mRNA molecules and imperfect procedures for capturing them in individual cells lead to dropout events, *i.e.,* genes show zero or very low expression, even though they are expressed in cells [8,9]. Further, it is well established that the capture rates vary between cells for a given scRNA-seq protocol, and this is a major source of unwanted technical variation that adds to the dropout events [10,11]. Addition of UMIs during the library preparation step reduces the amplification bias but has no effects on dropout events [12]. Further, the dropout events add more zeros to the output data, and can be categorized as either true/biological zeros (gene is not expressed in the cell), or false/technical zeros (gene is expressed but not detected) [13]. The

* Corresponding author at: Director at Biostatistics and Bioinformatics Facility, JG Brown Cancer Center, University of Louisville, Louisville, KY 40202, USA.
E-mail addresses: samarendra.das@louisville.edu (S. Das), shesh.rai@louisville.edu (S.N. Rai).

https://doi.org/10.1016/j.ygeno.2021.02.014
Received 14 September 2020; Received in revised form 22 January 2021; Accepted 22 February 2021
Available online 1 March 2021

presence of higher proportions of zeros and technical noise in scRNA-seq data can severely affect the performance of downstream Differential Expression (DE) analysis.

Bulk RNA-seq DE methods such as edgeR [14], and DESeq2 [15,16] have been used extensively for DE analysis of scRNA-seq data. These methods use the Negative Binomial (NB) model to capture the distributional nature of read counts under a Generalized Linear Model (GLM) framework. Further, Limma-Voom considers linear models for log-transformed counts data and observation-level weights to account for the dispersion of the transformed data [17,18], while DEGseq assumes the Poisson distribution of the read counts [19]. The use of such approaches in scRNA-seq data analysis raises serious concerns about their validity due to high dropout events [13], transcriptional bursting [20], lower molecular capturing in cells [9,21], and higher dispersion, *etc.* Therefore, dedicated scRNA-seq DE methods have been developed which use different strategies to cope with the above concerns [8,9,13,21–24]. For instance, SCDE uses a mixture model (*i.e.*, Poisson for dropout part and NB for amplification part) to capture the observed abundance of a given transcript in each cell [25]. SCDE assumes that the observed zero-count belongs to the dropout events with certainty. Further, MAST uses a hurdle model, *i.e.*, logistic regression for the level of gene expression and a Gaussian linear model for rate of expression conditioned on expression levels [8]. However, SCDE and MAST do not differentiate between biological and technical zeros during the model building. The BPSC approach [26] was developed to perform the DE analysis of scRNA-seq data through integrating Beta-Poisson model in the GLM framework. It does not consider the count nature of UMI data, and is severely affected by the dropout events [27]. These methods specifically consider the bi-modal distributional nature of the scRNA-seq data. Hence, a class of methods including D3E [28] and scDD [29] was developed to address the multimodal distributions of transformed scRNA-seq data, but they failed to consider the UMI count nature of the data and excluded the dropout events. Further, methods such as Monocle [23], Monocle2 [30], and NBID [31] were designed to handle the unique features of UMI in scRNA-seq experiments. They fit NB models directly to the observed UMI count data without any explicit focus on dropout events. Next, another class of specialized methods, such as ZINB-WaVE [13,32], DEsingle [22], and DECENT [9], were developed to handle excess zero inflation in scRNA-seq data. These methods are based on fitting of Zero Inflated Negative Binomial (ZINB) models to the UMI counts data. To be more specific, ZINB-WaVE [13] estimates observational level weights through Expectation-Maximization (EM) algorithm for adjusting bulk DE methods, *i.e.*, edgeR [14], DESeq2 [16]. The DEsingle [22] approach assumes ZINB models for observed UMI count data to estimate the parameters through Maximum Likelihood Estimation (MLE) method for two cellular populations separately. However, DECENT [9] assumes ZINB model for observed UMI count data and considers a Beta-Binomial model for the molecule capturing process. These methods ignore multimodal distributions of the observed expression data, estimate the DE parameters under parametric model assumptions, and are mostly focused on two-groups comparisons. Further, there is another class of DE methods which explicitly considers technical variation and molecular capturing processes, based on external spike-ins data. This class includes methods such as TASC [33], BASiCs [34], DECENT [9], and DESCEND [21]. Moreover, several comprehensive reviews and comparative analysis of DE methods covering all the above classes can be found in the literature [27,35–39].

It is evident that cells in scRNA-seq data behave variably and tend to be in different cell clusters [40], due to cell-cell heterogeneity. Biologically, these cell clusters are often different cell-types (*e.g.,* neurons and glia in brain sample) and correspond to different active states of cell types. Hence, descriptive data mining strategies (*e.g.,* clustering) have been adapted for scRNA-seq data analytics. In this study, we argue that the underlying cell clusters may have a significant effect on the means of non-zero counts of genes, and subsequently may influence the power of detection of DE genes. Further, there are limited methods available to date which consider the molecular capturing process, cell cluster information, and other cell-level auxiliary information for DE analysis. The incorporation of these data into the DE methods, may enhance their performance. This process requires building specific statistical models in order to perform statistical tests reliably.

We, therefore, propose a novel statistical approach, *i.e.,* SwarnSeq, for the DE analysis of scRNA-seq UMI count data. Here, we integrate the parametric ZINB model with a binomial molecular capture model in the presence of cell-level data. This allows us to detect DE genes and Differential Zero-Inflated (DZI) genes under a GLM framework. SwarnSeq can also classify the influential genes from scRNA-seq study into various groups. SwarnSeq can use external RNA spike-in data to adjust the distribution of the observed UMI counts with capture rates; however, it also works without spike-ins. In this paper, we describe SwarnSeq approach and benchmark it against 11 other existing methods, *i.e.*, DEGseq [19], edgeR [14], DEseq2 [15,16], LIMMA [18], Monocle2 [24], MAST [8], BPSC [26], SCDD [29], DEsingle [22], DECENT [9], and NODES [41] using 10 real scRNA-seq datasets. Our analytical results indicate that the SwarnSeq approach outperformed the competing existing methods on multiple real datasets, when assessed under 3 comparative settings.

2. Material and methods

2.1. Motivational data example

In scRNA-seq DE analysis, the cells are clustered, and these cell clusters are further divided into two groups (for example: group 1 has cluster *M* and group 2 has remaining clusters), as shown Supplementary Fig. S2. In existing analyses, this cell cluster information is kept out of the model building, and this may have a significant influence on the mean of non-zero counts. To test this claim, we took a toy example scRNA-seq dataset having 200 genes and 150 cells (Group 1: 50 cells; Group 2: 100 cells), available in DEsingle R package [22,42]. Then, we modeled the mean of non-zero counts under a GLM framework by providing group and cell cluster information as auxiliaries. The details of data description and analysis can be found in Supplementary Document S6. The results are shown in Supplementary Table S4. Our preliminary analysis indicated that the cell clusters have significant effect on the mean counts of the gene (Table S4). Hence, this toy data example motivated us to develop an innovative and novel statistical approach for DE analysis of scRNA-seq data through incorporating the cell clusters, other cell-level auxiliaries, and cell capture rates into the model building process under a GLM framework.

2.2. Single cell RNA-seq datasets

Our comprehensive analysis includes benchmarking of the proposed SwarnSeq method against 11 competitive existing methods on multiple real scRNA-seq datasets. This process starts with collection of publicly available scRNA-seq datasets from the GEO NCBI database (https://www.ncbi.nlm.nih.gov/geo). In our comparative analysis, we included the 10 UMI count gene expression datasets derived from 8 independent scRNA-seq studies. Further, the selected datasets include scRNA-seq data from lung cancer cells, pluripotent stem cells, liver cells, adipose stem/stromal cells, HEK cells from human, and embryonic stem cells, blood cells, and cells from mice. There are limited studies, where transcript concentration and external spike-in data are publicly available. Hence, we used the molecular concentration and ERCC spike-in data from Tung et al.'s experiment [12], available in https://github.com/jdblischak/singleCellSeq, to estimate the cell capture rates, while for other data cases, cell capture rates are estimated from the data *per se*. We used the processed UMI count data for these considered scRNA-seq studies as these datasets have gone through careful quality control steps by the authors of the original publications. The brief and detailed descriptions of the selected datasets are given in Table 1 and

Table 1
scRNA-seq datasets used in this study.

SN.	Data	Description	Accession	Protocol	#Genes	#Cells	References
01	Lung cancer	10× chromium sample from lung cells from three cell lines.	GSE111108	NextSeq 500	33,456	4000	[6]
02	Pluripotent stem cell	Human induced Pluripotent stem cell lines.	GSE77288	HiSeq	18,938	576	[12]
03	Mouse blood cell	Single cell profiling of mouse blood cells	GSE109999	CEL-Seq	19,903	383	[6]
04	Liver cell	single cell RNA sequencing by 10× Genomics of human liver cell lines	GSE115469	HiSeq	20,007	8444	[43]
05	Mouse cell	single-cell (ES and MEF) transcriptional landscape by highly multiplex RNA-Seq	GSE29087	SmartSeq	22,928	92	[7]
06	Adipose stem/stromal cells	Differentiating adipose cells by scRNA-Sequencing (Day 1 *vs* Day 2)	GSE53638	HiSeq	23,895	1835	[52]
07	Adipose stem/stromal cells	Differentiating adipose cells by scRNA-Sequencing (Day 1 *vs* Day 3)	GSE53638	HiSeq	23,895	2268	[52]
08	Adipose stem/stromal cells	Differentiating adipose cells by scRNA-Sequencing (Day 2 *vs* Day 3)	GSE53638	HiSeq	23,895	1613	[52]
09	Mouse embryonic cells	Mouse embryonic stem cells	GSE65525	DropSeq	24,174	1481	[5]
10	HEK cell	Single-cell RNA sequencing experiments of HEK cells	GSE92495	NextSeq	24,176	1453	[53]

#genes: number of genes, #cells: number of cells.

Supplementary Document S8, respectively. Further, we do not filter any cells in these datasets excluding the Tian et al. [6] and MacParland et al. [43] datasets, where we removed the low-quality cells with lower library sizes (Supplementary Document S8, Table S6). Further, we also performed filtering of very low abundance genes, *i.e.,* those that do not have at least five non-zero counts over the cells (Supplementary Document S8).

2.3. Model formulations

Notations: Let, Z_{ijk}: random variable (*rv*) representing the true (unknown) read (UMI) counts of k^{th} ($k = 1, 2, \ldots, K$) gene of j^{th} ($j = 1, 2, \ldots, M_i$) cell in i^{th} ($I = 1, 2, \ldots, N$) cell cluster; K: total number of genes; M_i: number of cells in i^{th} cell cluster; $M\left(=\sum_{i=1}^{N} M_i\right)$: total number of cells; N: number of cell clusters; μ_{ijk}: mean of k^{th} gene of j^{th} cell in i^{th} cell cluster for NB distribution; θ_{ijk}: size (=1/dispersion) parameter of k^{th} gene of j^{th} cell in i^{th} cell cluster for NB distribution; π_{ijk}: mixture probability (*i.e.,* the probability for a count to be an excess zero in a cell) parameter for k^{th} gene of j^{th} cell in i^{th} cell cluster.

In bulk RNA-seq, the counts are usually modeled through a NB distribution. The Probability Mass Function (PMF) of the NB distribution is expressed as:

$$f_{NB}(z) = P\left[Z_{ijk} = z|\mu_{ijk}, \theta_{ijk}\right] = \frac{G(z+\theta_{ijk})}{G(z+1)G(\theta_{ijk})}\left(\frac{\theta_{ijk}}{\theta_{ijk}+\mu_{ijk}}\right)^{\theta_{ijk}}\left(\frac{\mu_{ijk}}{\theta_{ijk}+\mu_{ijk}}\right)^{z} \forall z = 0, 1, 2, \ldots \quad (1)$$

where, $\mu_{ijk} \geq 0$; $\theta_{ijk} > 0$ are the parameters of NB distribution, $G(.)$: Gamma function. The NB distribution becomes Poisson, when $\theta_{ijk} \to \infty$.

For any $\pi_{ijk} \in [0,1]$, the true read counts in scRNA-seq study is assumed to follow a ZINB distribution [9,13,22]. The PMF of the ZINB distribution can be expressed as:

$$f_{ZINB}(z) = P\left[Z_{ijk} = z\right] = \pi_{ijk}\delta_0(z) + \left(1-\pi_{ijk}\right)f_{NB}(z) \forall z = 0, 1, 2, \ldots \quad (2)$$

where, $f_{NB}(.)$: PMF of NB distribution; $\delta_0(.)$: Dirac's delta function. Here, $\delta_0(.)$ is used to model the excess zeros in the scRNA-seq data, and its PMF is equal to zero for every non-zero counts except zero-counts and can be expressed as:

$$\delta_0\left(Z_{ijk} = z\right) = \begin{cases} 1; z = 0 \\ 0; z \neq 0 \end{cases} \quad (3)$$

The PMF of the ZINB distribution, used to model the read counts from scRNA-seq data, can be given as:

$$P\left[Z_{ijk} = z|\pi_{ijk}, \mu_{ijk}, \theta_{ijk}\right] = \begin{cases} \pi_{ijk} + \left(1-\pi_{ijk}\right)\left(\frac{\theta_{ijk}}{\theta_{ijk}+\mu_{ijk}}\right)^{\theta_{ijk}} \; when\; z = 0 \\ \left(1-\pi_{ijk}\right)\frac{G(z+\theta_{ijk})}{G(z+1)G(\theta_{ijk})}\left(\frac{\theta_{ijk}}{\theta_{ijk}+\mu_{ijk}}\right)^{\theta_{ijk}}\left(\frac{\mu_{ijk}}{\theta_{ijk}+\mu_{ijk}}\right)^{z}; z > 0 \end{cases} \quad (4)$$

Now, $Z_{ijk} \sim ZINB(\pi_{ijk}, \mu_{ijk}, \theta_{ijk})$, then the expected value and variance of Z_{ijk} can be obtained as [Kindly see Supplementary Document S1 for proof]:

$$E\left(Z_{ijk}\right) = \left(1-\pi_{ijk}\right)\mu_{ijk} \text{ and } V\left(Z_{ijk}\right) = \left(1-\pi_{ijk}\right)\mu_{ijk}\left(1+\pi_{ijk}\mu_{ijk}+\frac{\mu_{ijk}}{\theta_{ijk}}\right) \quad (5)$$

If $\pi_{ijk} = 0$; $ZINB(\pi_{ijk}, \mu_{ijk}, \theta_{ijk}) \to NB(\mu_{ijk}, \theta_{ijk})$
If $\theta_{ijk} \to \infty$ (*No dispersion*); $ZINB(\pi_{ijk}, \mu_{ijk}, \theta_{ijk}) \to ZIP(\pi_{ijk}, \mu_{ijk})$

2.4. Proposed SwarnSeq method

2.4.1. Model adjustment for cell capture rates

Theorem. Let, Y_{ijk} be the *rv* for observed (known) read (UMI) counts of k^{th} gene of j^{th} cell in i^{th} cell cluster and ρ_{ijk} be the transcriptional capture rate *rv* for k^{th} gene of j^{th} cell in i^{th} cell cluster. If Z_{ijk} follows a *ZINB* $(\pi_{ijk}, \mu_{ijk}, \theta_{ijk})$ distribution, and ρ_{ijk} follows a binomial model with parameter p_{ijk} $(0 \leq p_{ijk} \leq 1)$, then Y_{ijk} will also follow *ZINB* distribution with parameters $(\pi_{ijk}, \mu_{ijk}p_{ijk}, \theta_{ijk})$.

Proof. Given that, $Z_{ijk} \sim ZINB(\pi_{ijk}, \mu_{ijk}, \theta_{ijk})$ and $\rho_{ijk} = (Y_{ijk}|\, Z_{ijk} = z) \sim B(z, p_{ijk})$.

Now, the PMF of Z_{ijk} is given in Eq. (4) and the PMF of ρ_{ijk} can be expressed in Eq. (6).

$$P\left[Y_{ijk} = y|Z_{ijk} = z\right] = \binom{z}{y} p_{ijk}{}^{y}\left(1-p_{ijk}\right)^{z-y} \quad (6)$$

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

$$P\left[Y_{ijk} = y, Z_{ijk} = z|\pi_{ijk}, \mu_{ijk}, \theta_{ijk}, p_{ijk}\right] = P\left[Y_{ijk} = y|Z_{ijk} = z, p_{ijk}\right]P\left[\, Z_{ijk} = z|\pi_{ijk}, \mu_{ijk}, \theta_{ijk}\right] \quad (7)$$

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

$$P\left[Y_{ijk} = y|\pi_{ijk}, \mu_{ijk}, \theta_{ijk}, p_{ijk}\right] = \sum_{z} P\left[Y_{ijk} = y|Z_{ijk} = z, p_{ijk}\right]P\left[\, Z_{ijk} = z|\pi_{ijk}, \mu_{ijk}, \theta_{ijk}\right] \quad (8)$$

Case-1. For zero count ($Y_{ijk} = 0$)

$$\begin{aligned} P[Y_{ijk} = 0|\pi_{ijk}, \mu_{ijk}, \theta_{ijk}, p_{ijk}] &= P[Y_{ijk} = 0|Z_{ijk} = z, p_{ijk}]P[Z_{ijk} \\ &= 0|\pi_{ijk}, \mu_{ijk}, \theta_{ijk}] + \sum_{z=1}^{\infty} P[Y_{ijk} = 0|Z_{ijk} = z, p_{ijk}]P[Z_{ijk} \\ &= z|\pi_{ijk}, \mu_{ijk}, \theta_{ijk}] = \pi_{ijk} + (1-\pi_{ijk})\left(\frac{\theta_{ijk}}{\theta_{ijk}+\mu_{ijk}p_{ijk}}\right)^{\theta_{ijk}} \\ &= \pi_{ijk} + (1-\pi_{ijk})\left(\frac{\theta_{ijk}}{\theta_{ijk}+\mu'_{ijk}}\right)^{\theta_{ijk}} \left(\mu_{ijk}p_{ijk} = \mu'_{ijk}\,(say)\right) \end{aligned} \tag{9}$$

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

$$\begin{aligned} P[Y_{ijk} = t|\pi_{ijk}, \mu_{ijk}, \theta_{ijk}, p_{ijk}] &= \sum_{z \geq t} P[Y_{ijk} = t|Z_{ijk} = z, p_{ijk}]P[Z_{ijk} = z|\pi_{ijk}, \mu_{ijk}, \theta_{ijk}] \\ &= (1 \\ &\quad -\pi_{ijk})\left(\frac{\theta_{ijk}}{\theta_{ijk}+\mu_{ijk}}\right)^{\theta_{ijk}} \sum_{z \geq t}\binom{z}{t} p_{ijk}{}^{t}(1-p_{ijk})^{z-t} \frac{G(z+\theta_{ijk})}{G(z+1)G(\theta_{ijk})}\left(\frac{\mu_{ijk}}{\theta_{ijk}+\mu_{ijk}}\right)^{z} \\ &= (1-\pi_{ijk})\frac{G(t+\theta_{ijk})}{G(t+1)G(\theta_{ijk})}\left(\frac{\theta_{ijk}}{\theta_{ijk}+\mu_{ijk}p_{ijk}}\right)^{\theta_{ijk}}\left(\frac{\mu_{ijk}p_{ijk}}{\theta_{ijk}+\mu_{ijk}p_{ijk}}\right)^{t} \\ &= (1-\pi_{ijk})\frac{G(t+\theta_{ijk})}{G(t+1)G(\theta_{ijk})}\left(\frac{\theta_{ijk}}{\theta_{ijk}+\mu'_{ijk}}\right)^{\theta_{ijk}}\left(\frac{\mu'_{ijk}}{\theta_{ijk}+\mu'_{ijk}}\right)^{t} \end{aligned} \tag{10}$$

Now, Eqs. (9) and (10) are in the form of Eq. (4), which indicates the distribution of Y_{ijk} is also $ZINB(\pi_{ijk}, \mu_{ijk}', \theta_{ijk})$. Kindly see Supplementary Document S2 for proof of this theorem. When $p_{ijk} = 1$ (under full capture rates), then $ZINB(\pi_{ijk}, \mu_{ijk}', \theta_{ijk}) \rightarrow ZINB(\pi_{ijk}, \mu_{ijk}, \theta_{ijk})$.

$$\boldsymbol{\varphi}_k = log\boldsymbol{\theta}_k \tag{13}$$

where, $logit(\pi) = log\left(\frac{\pi}{1-\pi}\right)$; $\boldsymbol{\alpha_k}$, $\boldsymbol{\tau_k}$ and $\boldsymbol{\varphi_k}$: $M \times 1$ vector of parameters for k^{th} gene; $\boldsymbol{X}$: $M \times G$ design matrix providing group information (first column consists of 1's to include intercept term); G: number of cellular groups (cell clusters are divided into G groups, if group is unknown); $\boldsymbol{R}$: $M \times N$ design matrix providing cell cluster information; $\boldsymbol{C}$: $M \times C$ design matrix providing cell level auxiliary information; $\boldsymbol{\gamma_k}$ and $\boldsymbol{\beta_k}$: $G \times 1$ vectors of cellular groups effects for k^{th} gene; $\boldsymbol{w_k}$ and $\boldsymbol{u_k}$: $N \times 1$ vectors of cell cluster effects for k^{th} gene; $\boldsymbol{s_k}$ and $\boldsymbol{v_k}$: $C \times 1$ vectors of effects for cell level co-variates like cell cycle, cell phase, *etc.* for the k^{th} gene; C: Levels of cell level auxiliaries. $\boldsymbol{O_\mu}$,$\boldsymbol{O_\pi}$: offsets for $\boldsymbol{\mu_k}'$ and $\boldsymbol{\pi_k}$respectively.

2.4.3. Estimation of model parameters with EM algorithm

The parameters in Eqs. (11)–(13) for k^{th} gene, *i.e.,* $\boldsymbol{\Omega_k} = \{\boldsymbol{\alpha_k}, \boldsymbol{\tau_k}, \boldsymbol{\varphi_k}\}$ can be estimated by using the MLE Method. However, no closed form solutions exist for the resulting log-likelihood equation in Eq. (14). Hence, we developed an EM algorithm to estimate the parameters for the given observed scRNA-seq count data, *i.e.,* $Y_{ijk} = y_{ijk}$. Now, the incomplete data likelihood function for k^{th} gene can be expressed as:

$$L(\boldsymbol{\Omega_k}; Y_{ijk} = y_{ijk}) = \prod_{i=1}^{N}\prod_{j=1}^{M_i}\{\pi_{ijk}\delta_0(y_{ijk}) + (1-\pi_{ijk})f_{NB}(y_{ijk})\} \tag{14}$$

Further, the EM algorithm recasts the ZINB model into a missing data problem by introducing a latent rv, V_{ijk}. The V_{ijk} can be defined as:

$$V_{ijk} = \begin{cases} 1 \text{ if the observed count data comes from the zero componet} \\ 0 \text{ if the observed count data comes from the count component} \end{cases}$$

Now, the joint likelihood function for complete data, *i.e.,* (Y_{ijk}, V_{ijk}) can be expressed in Eq. (15), as:

$$L(\boldsymbol{\Omega_k}; Y_{ijk}, V_{ijk}) = \prod_{i=1}^{N}\prod_{j=1}^{M_i}\left[\left\{\pi_{ijk} + (1-\pi_{ijk})\left(\frac{\theta_{ijk}}{\theta_{ijk}+\mu_{ijk}}\right)^{\theta_{ijk}}\right\}^{V_{ijk}}\left\{(1-\pi_{ijk})\frac{G(z+\theta_{ijk})}{G(z+1)G(\theta_{ijk})}\left(\frac{\theta_{ijk}}{\theta_{ijk}+\mu_{ijk}}\right)^{\theta_{ijk}}\left(\frac{\mu_{ijk}}{\theta_{ijk}+\mu_{ijk}}\right)^{y_{ijk}}\right\}^{1-V_{ijk}}\right] \tag{15}$$

Then, the log-likelihood function in Eq. (15) becomes:

$$\begin{aligned} l(\boldsymbol{\Omega_k}; Y_{ijk}, V_{ijk}) &= \sum_{i=1}^{N}\sum_{j=1}^{M_i} V_{ijk} log\left\{\pi_{ijk} + (1-\pi_{ijk})\left(\frac{\theta_{ijk}}{\theta_{ijk}+\mu_{ijk}}\right)^{\theta_{ijk}}\right\} + \sum_{i=1}^{N}\sum_{j=1}^{M_i}(1-V_{ijk}) log\left\{(1-\pi_{ijk})\frac{G(z+\theta_{ijk})}{G(z+1)G(\theta_{ijk})}\left(\frac{\theta_{ijk}}{\theta_{ijk}+\mu_{ijk}}\right)^{\theta_{ijk}}\left(\frac{\mu_{ijk}}{\theta_{ijk}+\mu_{ijk}}\right)^{y_{ijk}}\right\} \\ &= l_1(\boldsymbol{\Omega_k}; V_{ijk}) + l_2(\boldsymbol{\Omega_k}; Y_{ijk}, V_{ijk}) \end{aligned} \tag{16}$$

2.4.2. Generalized Linear Model framework in presence of cell capture rates

We estimated the parameters of the ZINB model, given in Eqs. (9) and (10), from the observed UMI count data under a GLM framework. We have shown that the observed UMI counts for k^{th} gene, Y_{ijk}, as a ZINB rv with parameters $\boldsymbol{\mu_k}' = (\mu_{11k}', \ldots, \mu_{1M1k}', \ldots \mu_{N1k}', \ldots, \mu_{NMNk}')$; $\boldsymbol{\pi_k} = (\pi_{11k}, \ldots, \pi_{1M1k}, \ldots, \pi_{N1k}, \ldots, \mu_{NMNk})$; $\boldsymbol{\theta_k} = (\theta_{11k}, \ldots, \theta_{1M1k}, \ldots, \theta_{N1k}, \ldots, \theta_{NMNk})$ and further the following GLMs (Eqs. (11)–(13)) are considered to model these parameters.

$$\boldsymbol{\alpha_k} = log\boldsymbol{\mu_k}' = \boldsymbol{X\gamma_k} + \boldsymbol{Rw_k} + \boldsymbol{Cs_k} + \boldsymbol{O_\mu} \tag{11}$$

$$\boldsymbol{\tau_k} = logit\boldsymbol{\pi_k} = \boldsymbol{X\beta_k} + \boldsymbol{Ru_k} + \boldsymbol{Cv_k} + \boldsymbol{O_\pi} \tag{12}$$

where, $l_1(.)$: log-likelihood due to the zero-component of the model and $l_2(.)$: log-likelihood due to the count-component of the model. Hence, the expected value of the log-likelihood function in Eq. (16) can be expressed as:

$$\begin{aligned} Q &= E[l(\boldsymbol{\Omega_k}; Y_{ijk}, V_{ijk})] \\ &= \sum_{i=1}^{N}\sum_{j=1}^{M_i} E\left(V_{ijk}|Y_{ijk}\boldsymbol{\Omega_k}\right) log\left\{\pi_{ijk} + (1-\pi_{ijk})\left(\frac{\theta_{ijk}}{\theta_{ijk}+\mu_{ijk}}\right)^{\theta_{ijk}}\right\} \\ &\quad + \sum_{i=1}^{N}\sum_{j=1}^{M_i}(1 - E(V_{ijk}|Y_{ijk}\boldsymbol{\Omega_k}))log\left\{(1 \right. \\ &\quad \left. -\pi_{ijk})\frac{G(z+\theta_{ijk})}{G(z+1)G(\theta_{ijk})}\left(\frac{\theta_{ijk}}{\theta_{ijk}+\mu_{ijk}}\right)^{\theta_{ijk}}\left(\frac{\mu_{ijk}}{\theta_{ijk}+\mu_{ijk}}\right)^{y_{ijk}}\right\} \end{aligned} \tag{17}$$

Further, the posterior probabilities in Eq. (17) for the observations originate from the count component of the model and can be given as:

$$w_{ijk} = P\left[V_{ijk} = 0 | Y_{ijk}, \boldsymbol{\Omega}_k\right] = \frac{(1-\pi_{ijk}) f_{NB}\left(y_{ijk}; \mu_{ijk}, \theta_{ijk}\right)}{\pi_{ijk}\delta_0\left(y_{ijk}\right) + (1-\pi_{ijk}) f_{NB}\left(y_{ijk}; \mu_{ijk}, \theta_{ijk}\right)} \tag{18}$$

where, $f_{NB}(.)$ is the PMF of NB distribution given in Eq. (1).

A. ***E-step***: The E-step in the EM algorithm involves in evaluating the expected value of the log-likelihood of the complete data (Eq. (17)), given the observed data with the current estimates of the parameters. In our proposed approach, for each gene, given the observed data and a current estimate of the ZINB parameters, the expected value of the log-likelihood is calculated. Let, $\widehat{\boldsymbol{\Omega}}_k^{\,c} = \left\{\widehat{\boldsymbol{\alpha}}_k^{\,c}, \widehat{\boldsymbol{\tau}}_k^{\,c}, \widehat{\boldsymbol{\varphi}}_k^{\,c}\right\}$ be the given current estimate of the parameters, then the expected value of log likelihood in Eq. (17) at step (c + 1), *i.e.,* Q^{c+1} is calculated. The conditional expectation, *i.e.,* $E(V_{ijk}\left|Y_{ijk}\widehat{\boldsymbol{\Omega}}_k\right.)$ in Eq. (17) can be given as:

$$E\left(V_{ijk}\middle|Y_{ijk}, \widehat{\boldsymbol{\Omega}}_k\right) = \frac{\widehat{\pi}_{ijk} + \left(1-\widehat{\pi}_{ijk}\right)\left(\frac{\widehat{\theta}_{ijk}}{\widehat{\theta}_{ijk}+\widehat{\mu}_{ijk}}\right)^{\widehat{\theta}_{ijk}}}{\widehat{\pi}_{ijk}\delta_0\left(y_{ijk}\right) + \left(1-\widehat{\pi}_{ijk}\right) f_{NB}\left(y_{ijk}\middle|\widehat{\mu}_{ijk}, \widehat{\theta}_{ijk}\right)} \tag{19}$$

B. ***M*-step:** Maximize Q^{c+1} to update the parameter estimates.

i. The parameters from the count component of the model, $\left\{\widehat{\boldsymbol{\mu}}_k^{'}, \widehat{\boldsymbol{\theta}}_k\right\}$ are updated within the GLM framework, and can be expressed as:

$$log\boldsymbol{\mu}_k^{'} = \boldsymbol{X}\boldsymbol{\gamma}_k + \boldsymbol{R}\boldsymbol{w}_k + \boldsymbol{C}\boldsymbol{s}_k + \boldsymbol{O}_{\boldsymbol{\mu}} \tag{20}$$

The updated value of the estimates of parameters at step (c + 1) is obtained by providing the observation wise weights, $\widehat{w}_{ijk}^{(c)}{}_{ijk}$, given in Eq. (18) and parameters estimates at c-step. For this purpose, the *glm.nb* function in MASS R package was executed.

ii. The zero-inflation probability, $\widehat{\pi}_{ijk}$, is updated with the logistic regression, can be expressed as:

$$logit(\boldsymbol{\pi}_k) = \boldsymbol{X}\boldsymbol{\beta}_k + \boldsymbol{R}\boldsymbol{u}_k + \boldsymbol{C}\boldsymbol{v}_k + \boldsymbol{O}_{\boldsymbol{\pi}} \tag{21}$$

The updated value of $\widehat{\pi}_{ijk}$ at step (c + 1) is obtained by incorporating the observation level weights, $\widehat{w}_{ijk}^{(c)}{}_{ijk}$ given in Eq. (18) and the parameters estimate at c-step. For this, *glm*(…, family = 'binomial') function in stat R package was executed.

C. **Starting values for EM algorithm**

The success of an iterative algorithm, *e.g.,* EM, depends on the provision of initial values for the parameters. In our SwarnSeq method, we provide the initial values for the estimators for each gene by estimating through Generalized Linear (GL) Poisson and GL Binomial models for non-zero and zero counts, respectively. For this purpose, the *glm* function implemented in stats package was executed.

D. **Assessing convergence**

The EM algorithm iterates over an Expectation (E) step and Maximization (M) step for each gene until convergence achieved [13,44,45]. Let, $\widehat{\boldsymbol{\Omega}}_k^{\,c} = \left\{\widehat{\boldsymbol{\alpha}}_k^{\,c}, \widehat{\boldsymbol{\tau}}_k^{\,c}, \widehat{\boldsymbol{\varphi}}_k^{\,c}\right\}$ be the vector parameter estimates for k^{th} gene. The criteria for convergence can be expressed as:

$$\left| Q\left(\widehat{\boldsymbol{\Omega}}_k^{\,c+1}; Y_{ijk}, V_{ijk}\right) - Q\left(\widehat{\boldsymbol{\Omega}}_k^{\,c}; Y_{ijk}, V_{ijk}\right) \right| < \epsilon \tag{22}$$

where, ϵ is the threshold for convergence (*e.g.,* in SwarnSeq R package, the default for $\epsilon = 10^{-10}$ and maximum iteration is 10^3). It is important to note that for some genes, the EM algorithm may fail to converge or may be not successful; therefore, we used Nelder's optimization algorithm [46] implemented in *optim* function of stats R package to estimate the MLE of parameters.

2.4.4. Differential expression analysis

The gene-wise mean parameter depends on the cellular groups through the model given in Eq. (11). Further, the factors such as cell clusters and cell co-variates are included in the model to remove the unwanted effects. For DE analysis, two group comparisons were made and the model in Eq. (11) can be expanded as:

$$\mathrm{Log}\left(\mu_{ijk}\right) = \gamma_{0k} + \gamma_{1k}x_{ijk} + w_{1k}r_{1jk} + \ldots + w_{Nk}r_{Njk} + s_{1k}c_{1jk} + \ldots + s_{mk}c_{mjk} + O_{\mu_k} \tag{23}$$

where, x_{ijk}: binary indicator for cellular group membership, γ_{0k}: (intercept term) logarithm of mean parameter for k^{th} gene in the reference cellular group, γ_{1k}: regression parameter for cellular group effect of k^{th} gene, w_{ik}: regression co-efficient for i^{th} cell cluster for k^{th} gene, r_{ijk}: indicator variable for cell cluster membership of j^{th} cell in i^{th} cluster for k^{th} gene, s_{mk}: regression co-efficient for m^{th} cell co-variates of k^{th} gene, c_{mjk}: indicator variable for m^{th} co-variates of j^{th} cell for k^{th} gene and O_{μ_k}: offset term.

To decide whether, the k^{th} gene is DE or not, the following hypotheses were tested.

$$H_0 : \gamma_{1k} = 0 \; vs. H_1 : \gamma_{1k} \neq 0$$

The above test can be performed by using likelihood ratio test statistic, and can be expressed as:

$$DS_k = -2\left\{l\left(\boldsymbol{\Omega}_k = \widehat{\boldsymbol{\Omega}}_{k0}\right) - l\left(\boldsymbol{\Omega}_k = \widehat{\boldsymbol{\Omega}}_k\right)\right\} \tag{24}$$

where, $\widehat{\boldsymbol{\Omega}}_{k0}$: 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. The test statistic, DS_k, follows a Chi-square distribution with 1 degree of freedom (for 2 groups) under H_0. Further, based on the distribution of DS_k, the *p*-value, adjusted *p*-value and FDR for k^{th} gene can be computed after adjustment for multiple hypothesis testing.

2.4.5. Testing for differential zero inflation

Through generalized likelihood ratio statistical test, we have shown that genes in scRNA-seq data are highly zero inflated, described in Supplementary Document S3. Therefore, to facilitate DZI analysis in the SwarnSeq method, the gene-wise zero inflation parameter depends on the cellular groups through the model given in Eq. (12). Further, factors such as cell clusters and other cell-level auxiliaries are included in the model to remove the unwanted effects. For DZI analysis, two group comparisons were made and the model in Eq. (12) can be written as:

Table 2
Classification of influential genes using SwarnSeq method.

	Differentially expressed		
Differentially Zero Inflated		Yes	No
	Yes	DEZI	DZI
	No	DE	None

DEZI: Differentially Expressed as well as Differentially Zero Inflated; DZI: Differentially Zero Inflated; DE: Differentially Expressed.

$$logit(\pi_{ijk}) = \beta_{0k} + \beta_{1k}x_{ijk} + u_{1k}r_{1jk} + \ldots + u_{Nk}r_{Njk} + v_{1k}c_{1jk} + \ldots + v_{mk}c_{mjk} + O_{\pi_k} \quad (25)$$

where, x_{ijk}: binary indicator for cellular group membership, β_{0k}: intercept term, β_{1k}:regression co-efficient of cellular group effect for k^{th} gene, u_{ik}: regression co-efficient for i^{th} cell cluster for k^{th} gene, r_{ijk}: indicator variable for cell cluster membership of j^{th} cell in i^{th} cluster for k^{th} gene, v_{mk}: regression co-efficient for m^{th} cell co-variates of k^{th} gene, c_{mjk}: indicator variable for m^{th} co-variates of j^{th} cell and O_{π_k}: offset term.

To decide whether the k^{th} gene is DZI or not, the following hypotheses were tested.

$$H_{10} : \beta_{1k} = 0 \; vs. H_1 : \beta_{1k} \neq 0$$

A similar test statistic to that given in Eq. (28) can also be developed for testing of DZI of genes.

2.4.6. Classification of influential genes

The SwarnSeq method can divide the detected influential genes into different classes, as shown in Table 2. For instance, the H_0: $\gamma_{1k} = 0$ detects all the genes that are DE across two cellular groups, while $H_{10} : \beta_{1k} = 0$ detects the DZI genes. Further, the SwarnSeq detects a class of influential genes with both H_0 and H_{10} rejected, which indicates there is a significant difference in the number of cells with zero values for genes across the cellular groups, but the (non-zero counts) expressions in the remaining cells also show significant differences. We call such a group of influential genes as 'DEZI' genes. The second type of genes are those for which H_0 is rejected, but H_{10} is not. This means that there is no significant difference in the number of cells whose expressions are zeros across cellular conditions for genes, but they are expressed differentially. We call this group of genes as 'DE'. Further, the third type (*i.e.,* DZI) of genes is that for which H_{10} is rejected, but H_0 is not. It includes the genes for which, there is a significant difference in the number of cells with real zero values across the two cellular conditions, but the expression in the remaining cells shows no significant difference.

2.5. Estimation of capture rates parameter

The distribution of the observed scRNA-seq UMI counts depends on the cell specific transcriptional efficiency parameter, p_{ijk}. For computational simplicity, we assume $p_{ij1} = p_{ij2} = \ldots = p_{ijK} = p_{ij}$, *i.e.,* the cell specific efficiency parameters remain same across all the genes. The proposed procedure for estimation of cell capture rate parameters is described as follows.

Case 1. When RNA spike-ins are available

Suppose n RNA spike-ins are added to each cell's lysate and spike-in transcripts are processed in parallel. This will result a set of UMI counts for spike-in transcripts. Let, C_1, C_2, …, C_n be the respective mRNA concentrations of n spike-in transcripts added to j^{th} ($j = 1, 2, \ldots, M_i$) cell of i^{th} ($i = 1, 2, \ldots, N$) cell cluster and let $R_{ij1}, R_{ij2}, \ldots, R_{ijn}$ be the observed UMI counts of the n spike-in transcripts for j^{th} cell. Now, the transcriptional capture rate for j^{th} cell in i^{th} cell cluster can be estimated through a linear regression equation, given in Eq. (26).

$$R_{ijk} = p_{ij0} + p_{ij1}C_k + \epsilon_k \quad (26)$$

Here, $\widehat{p}_{ij1}$, regression co-efficient, is the estimate of the capture rate for j^{th} cell in i^{th} cell cluster.

Case 2. When RNA spike-ins are not available

Transcriptional capture efficiency parameters of cells are the key factors for variation in the observed cell specific library sizes [47]. Hence, the observed cell library sizes can be used to empirically compute the cell specific capture rates, which is given as:

Let, (ρ_1, ρ_2) be the range of capture rates and S_{ij} be the library size of j^{th} cell in i^{th} cell cluster and, $L_{ij} = log_{10}(S_{ij}) \; \forall \; i, j$

$$\text{Now, } L_{min} = \min_{j} L_{ij} \text{ and } \quad L_{max} = \max_{j} L_{ij}$$
$$\widehat{p}_{ij} = \rho_1 + (\rho_2 - \rho_1)\frac{L_{ij} - L_{min}}{L_{max} - L_{min}} \quad (27)$$

2.6. Performance evaluation metrics

The performance of SwarnSeq method for identifying genuine DE genes was evaluated with respect to 11 existing competitive methods (Table S7) using the Area Under Receiver Operating Characteristic (AUROC) curve (*i.e.,* true positive rate (TPR) *vs.* false positive rate (FPR)), and other performance metrics on 10 real scRNA-seq datasets (Table 1). The layout of this comparative study is shown in Supplementary Fig. S7. Further, the performance metrics (Eqs. (28)–(34)) were computed by comparing the DE genes obtained through each method with the reference genes (*i.e.,* true DE genes) for each dataset. For instance, we defined True Positive (TP) in Eq. (28) as the significant genes those were found to be true DE genes and False Positive (FP) in Eq. (29) as the genes that were found significant but were not true DE genes. Similarly, True Negative (TN) in Eq. (29) were defined as genes that were not true DE and were not found significant, and False Negative (FN) in Eq. (28) were defined as genes that were true DE but were not found significant. The evaluation metrics are defined in Eqs. (28)–(34), as:

$$TPR(\text{ or } Sensitivity) = \frac{TP}{TP + FN} \quad (28)$$

$$FPR(\text{ or } 1 - Specificity) = \frac{FP}{FP + TN} \quad (29)$$

$$PPR = \frac{TP}{TP + FP} \quad (30)$$

$$FDR = \frac{FP}{FP + TP} \quad (31)$$

$$NPV = \frac{TN}{TN + FN} \quad (32)$$

$$ACC = \frac{TP + TN}{TP + TN + FP + FN} \quad (33)$$

$$F1 = \frac{2TP}{2TP + FP + FN} \quad (34)$$

where, TPR: True Positive Rate; FPR: False Positive Rate; PPR: Positive Prediction Rate; FDR: False Discovery Rate; NPV: Negative Prediction Value; ACC: Accuracy; F1: F-score

3. Results

3.1. Preliminary analytical results

We considered two publicly available zero inflated and overdispersed datasets to show the suitability and goodness of fit of different count data models, *viz.* NB, ZINB, Poisson Distribution (PD), Hermite Distribution (HD) and Zero Inflated Poisson Distribution (ZIPD) [48,49]. The descriptions of these models, datasets, and the results from fitting the above count models are given in Supplementary Documents S1, S4. Our preliminary analytical results indicated that the expected frequencies computed from the ZINB model are much closer to their observed counter parts as compared to other models (Supplementary Tables S1, S2). Further, the ZINB provides the best fit to the given datasets as compared to NB, PD, ZIPD, and HD when assessed through different model fitting criteria (Supplementary Tables S1, S2). At this preliminary stage, we inferred that the ZINB model best suits to the zero inflated and overdispersed data (*e.g.,* scRNA-seq data) as compared to

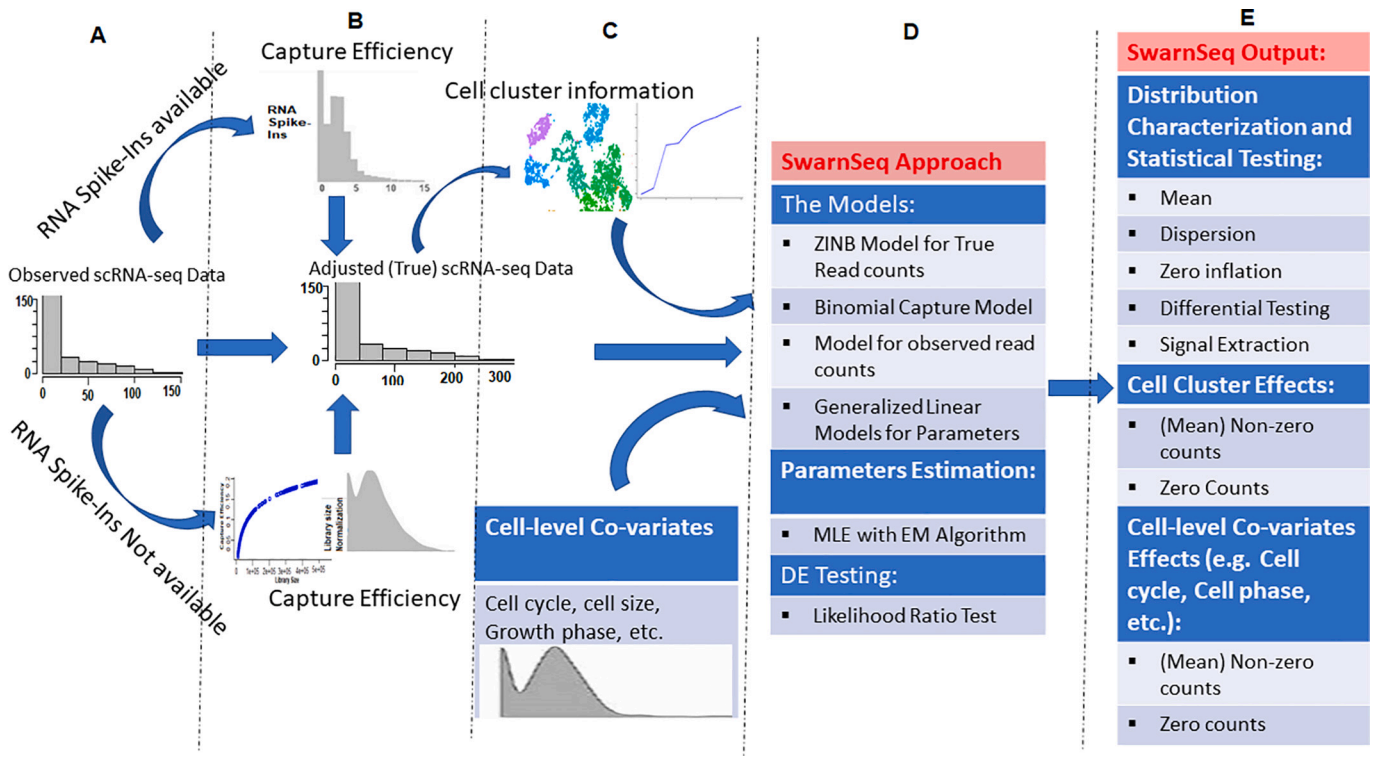

Fig. 1. Illustration of the operational framework of the SwarnSeq method. (A) cross-cell distribution of observed scRNA-seq counts; (B) cross-cell distribution of true/adjusted scRNA-seq counts with capture efficiency with respect to spike-ins information; (C) Auxiliary information such as cell cluster and cell level co-variates as inputs to the SwarnSeq; (D) Details of SwarnSeq method fitted for each gene; (E) For each gene, the output of SwarnSeq includes the distribution characterization (*i.e.* mean, dispersion and zero inflation) over cell populations, differential expression testing between two cell populations, differential zero inflation testing between two cell populations, effects of cell clusters on zero-inflation parameter and mean of non-zero counts, effects of cell level auxiliary information on zero-inflation parameter and mean of non-zero counts.

the NB model extensively used in RNA-seq data analysis.

We also tested the ability of NB and ZINB models to estimate the mean and dispersion parameters for scRNA-seq count data through simulation (Supplementary Document S5). For this purpose, parameter estimates for the BTG4 gene from human preimplantation of embryonic scRNA-seq data, available in DEsingle R package [22,42] was used to simulate count expression data through ZINB model and the results are shown in Supplementary Table S3. Our preliminary analysis indicates that the NB model underestimates the mean and overestimates the dispersion for scRNA-seq data. Further, the ZINB model provides better estimates of mean and dispersion with lower bias as compared to NB. This indicates better suitability of the ZINB model for modeling the zero inflated and overdispersed UMI counts data (Tables S1-S3). This is due to the fact that the NB model accommodated excess zeros in scRNA-seq data by underestimating the mean and overestimating the dispersion, which further jeopardizes the statistical power to detect DE genes. The detailed analysis and results are given in Supplementary Documents S4, S5.

3.2. Proposed model overview

Fig. 1 and Supplementary Fig. S6 give an overview of the computational framework and major analytical steps of the proposed SwarnSeq method. The observed UMI counts are the noisy reflection of the true expression of genes due to low transcriptional capturing. We modeled the observed UMI counts, Y_{ijk} of k^{th} gene in j^{th} cell in i^{th} cluster, as the joint distribution of k^{th} gene's true expression Z_{ijk} and transcriptional capture rate (p_{ijk}) of j^{th} cell in i^{th} cell cluster. In other words, after incorporating the transcriptional capturing procedure in the modeling process, the mean of non-zero counts in the ZINB model depends on cell capturing rate parameter. The relation between the capture efficiency with the distribution of the observed read counts is shown in Fig. 2. The relation among the means of count part in the ZINB model before and after incorporation of the capturing procedure is found to be $\mu_{ijk} > \mu_{ijk}'$. In other words, the distribution of observed scRNA-seq read counts shift more towards zeros after incorporation of the transcriptional capturing process (Fig. 2). This means that more zeros are found in observed data and will be from the count part of the model. Further, the expected value and variance of the observed UMI counts of genes depends (*i.e.,* directly proportional) on the cell capture rates (See Supplementary Document S2 for proof) and is expressed in Eqs. (35) and (36). Thus, when p_{ijk} becomes smaller both mean and variance of Y_{ijk} also becomes smaller.

$$E(Y_{ijk}) = (1-\pi_{ijk})\mu_{ijk}p_{ijk} \tag{35}$$

$$V(Y_{ijk}) = (1-\pi_{ijk})\mu_{ijk}p_{ijk}\left(1+\pi_{ijk}\mu_{ijk}p_{ijk}+\frac{\mu_{ijk}p_{ijk}}{\theta_{ijk}}\right) \tag{36}$$

The mixture probability and dispersion parameters for the observed UMI counts remain unchanged after the incorporation of the molecular capture procedure in the modeling process (Fig. 2, Document S2). For instance, when $p_{ijk} = 1$ (100% capture), the genes in a cell will have zero counts which are not truly expressed (*i.e.,* biological zeros); this is expected under a perfect deep sequencing scenario. In other words, observed UMI counts are the true expected counts of genes in a cell under a perfect deep sequencing. When $p_{ijk} < 1$ (*real case*), the zeros in the observed UMI counts are the mixture of dropouts and true zeros. It may be noted that π_{ijk} remains unaffected by the capture rate parameter, hence, the $\hat{\pi}_{ijk}$ from observed data can be used to measure the proportions of true zeros of genes in the data (Fig. 2, Document S2). The relation among the various parameters estimated through the SwarnSeq method is given in Supplementary Document S13.

SwarnSeq allows the modeling of the effects of cellular groups, cell

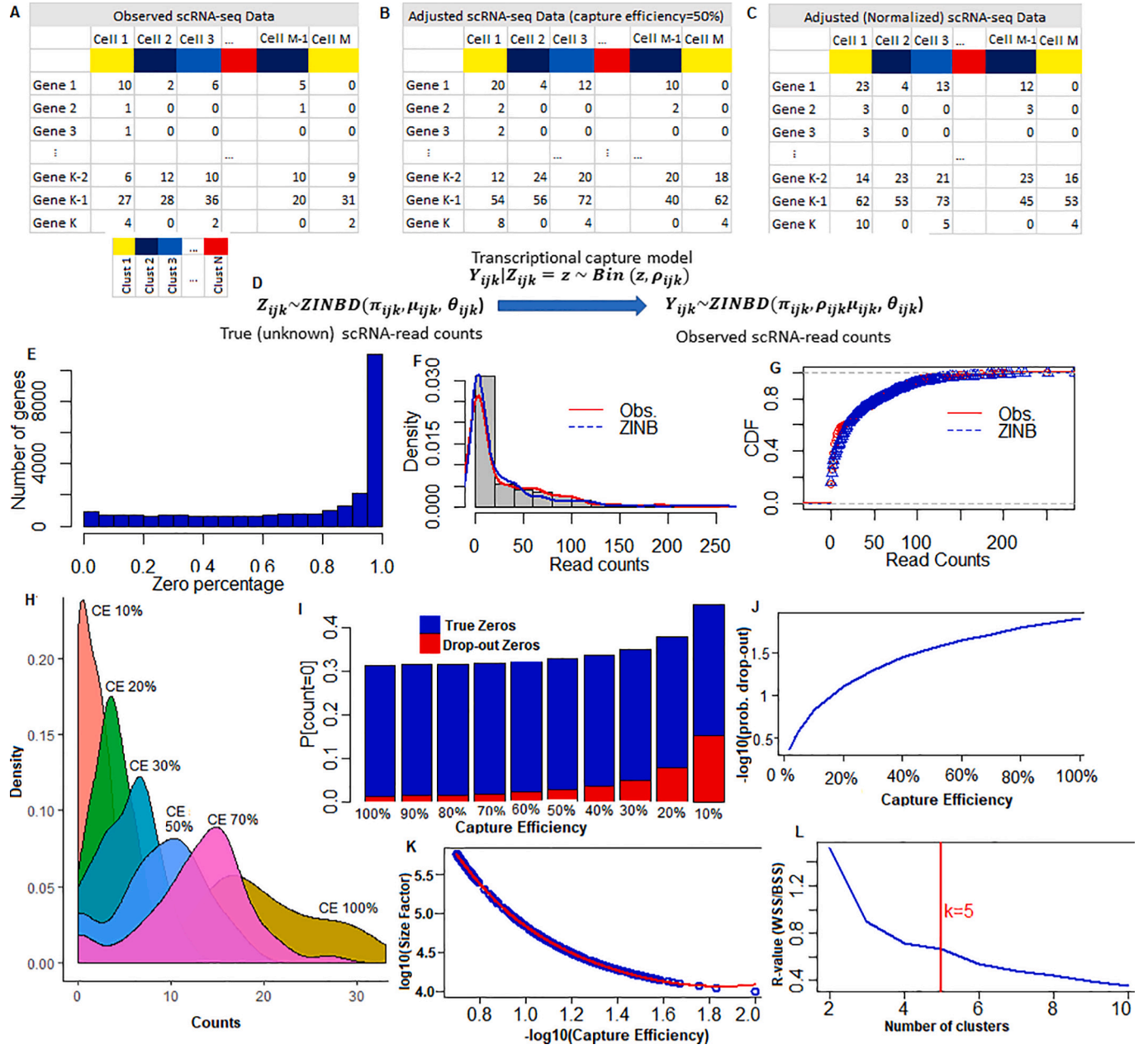

Fig. 2. Data structures, Models, and Distributions used in the SwarnSeq method. (A) Structure of the observed scRNA-seq data; (B) Input structure of the true scRNA-seq data adjusted with capture efficiency; (C) Input structure of the normalized true scRNA-seq data; (D) ZINB and transcriptional capture models used in SwarnSeq approach; (E) Histogram of zero percentages of all expressed genes in a real scRNA-seq dataset; (F) An example of ZINB model fitting for scRNA-seq data. The fitting of observed and theoretical ZINB models are shown for real scRNA-seq data for a gene; (G) Cumulative distribution function fitting for observed and theoretical ZINB models; (H) Theoretical ZINB distribution of observed scRNA-seq counts of a gene with different random capture efficiency. The distributions are shown for capture efficiencies 100%, 70%, 50%, 30%, 20% and 10%. Here, the 100% capture efficiency represents the distribution of true scRNA-seq counts; (I) The histograms of zero probabilities for different capture efficiencies are shown. The red colour bars represent the probability density of real true zero expressions. The blue bars represent the probability density of the NB part of the ZINB model. (J) The plot shows the relation between the probability of drop-out events and capture efficiencies of cells. (K) The relation between the library sizes and the capture efficiencies of the cells is shown. (L) Deciding the number of optimum cell clusters for a real scRNA-seq data. CE: Capture Efficiency; Clust.: Cell Cluster. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

clusters and other cell-level covariates on both the zero-inflation and mean parameters. So, for fitting the SwarnSeq model, we performed cluster analysis on all the 10 datasets and determined the optimum number of cell clusters through the proposed method described in Supplementary Document S9. The results are shown in Supplementary Fig. S3. When the cell level auxiliary information is specified, the SwarnSeq uses a log-linear model for the covariate effect on mean and a logit model for the covariate effect on zero-inflation in a GLM framework. Further, SwarnSeq performs DE analysis of genes for a given two groups situation and can be generalized to multiple groups comparison. Moreover, DZI analysis of genes of scRNA-seq data is allowed in the SwarnSeq method, which leads to the identification of severely zero-inflated genes over the cellular populations. Additionally, the genes in scRNA-seq data are classified into different gene types based on DE and DZI analysis (Table 2).

3.3. SwarnSeq as differential expression tool

We benchmarked the proposed SwarnSeq approach for DE analysis on 10 different real scRNA-seq datasets (Table 1). The problem in benchmarking of scRNA-seq DE methods on real scRNA-seq datasets is the unavailability of reference genes. Hence, to obtain the list of reference genes, we used the corresponding Microarrays, bulk RNA-seq and scRNA-seq datasets for each of the 10 considered datasets. For instance, GSE29087 data consists of 22,928 genes over 92 cells (48: mouse embryonic stem cells; 44 mouse embryonic fibroblasts cells), and then we filtered out the low expressed genes, *i.e.,* genes which do not have non-zero expressions in at least 5 cells. Further, the reference genes for the same cell lines were collected from the Microarray study available at http://carlosibanezlab.se//Data/Moliner_CELfiles.zip [50]. The selection of reference gene lists is described in detailed for each of the datasets in Supplementary Document S10. The 12 tested DE methods, including proposed SwarnSeq (Supplementary Tables S7, S8), were benchmarked using all the 10 datasets (Table 1, Supplementary Document S8), and the SwarnSeq method was also applied to GSE77288 data from Tung et al. [12], where spike-ins are available. The layout of this comparative study is briefly described in Supplementary Fig. S7.

3.3.1. Benchmarking based on receiver operating characteristic

This comparison setting used the experimental designs and the 10 count datasets from 8 independent scRNA-seq studies (Table 1) for performance analysis of scRNA-seq DE methods (Tables S7, S8). For instance, mouse cell data (GSE29087) [7] were used to detect DE genes between 48 mouse embryonic stem cells and 44 mouse embryonic fibroblast cells (Table 1). Then, the 12 competitive methods, including SwarnSeq, were compared in terms of their AUROC using the identified reference gene lists for each of 10 datasets. Basically, through each of the methods, DE gene sets of size 3000 were selected for each of the datasets. Then, the AUROC values were computed by executing *proc* function implemented in pROC R package [51] using the output (*i.e., p*-values or adjusted *p*-values) of each method as predictor, and a binary vector, which indicates whether a gene belongs to the reference gene list, as the response.

The ROC curves of different methods are shown in Figs. 3, and S12 along with the AUROC values. In this comparison setting for GSE53638 (data 1), the SwarnSeq (0.76) produced the highest AUROC values followed by DECENT (0.66), MAST (0.61), DEsingle (0.61), Monocle (0.54), and BPSC (0.52) among single cell specific tools (Fig. 3A). The scDD performed the worst among the scRNA-seq DE tools for this dataset. Further, edgeR (0.54) had a higher AUROC values followed by Limma (0.52), DEGseq (0.51) and DESeq2 (0.48) in the bulk RNA-seq tool category (Fig. 3A). It was found that the SwarnSeq performed better than other methods of both bulk and scRNA-seq DE tools. For GSE53638 (data 3), the SwarnSeq (0.73) had the highest AUROC value followed by DECENT (0.70) and performed best among bulk and single-cell tools (Fig. 3C). Moreover, among the bulk RNA-seq DE tools, edgeR (0.54) had higher AUROC followed by Limma (0.52), DEGseq (0.51) for the GSE53638 (3) data (Fig. 3C). Similarly, for data from Tung et al. (GSE77288), the AUROC for SwarnSeq method was highest (0.83) among other competitive bulk and single cell DE tools (Fig. 3B). Among the bulk RNA-seq DE tools, Limma had higher AUROC (0.62), when applied to Tung et al. scRNA-seq data. Similar interpretations can be made for other datasets, as shown in Figs. 3 and S12. Our analysis indicated that under AUROC settings, our SwarnSeq method performed better in 8 datasets (with rank 1) and competitive with other methods in remaining 2 datasets (rank 2 and 3) (Figs. 3, S12). In other words, the performance of SwarnSeq method is consistently better than other competitive methods on real scRNA-seq datasets.

3.3.2. Benchmarking based on FDR

The second comparison setting included assessment of the 12 tested methods (Table S7) through computation of FDRs for different DE gene sets on the 10 different real scRNA-seq datasets (Table 1). For this purpose, different DE gene sets of sizes 100, 200, 300, …, 3000 were selected based on the *p*-values/adjusted *p*-values computed through each of the 12 methods including SwarnSeq. Then, the selected DE gene sets were compared with respect to the reference gene list to compute FDRs for each of the 10 datasets. The results are shown in Figs. 4 and S13. In this comparison setting, it was found that the FDR computed for the SwarnSeq method was found to be lower as compared to other competitive methods for GSE53638 (data 1) (Fig. 4A). Similar findings were observed across all the selected DE gene sets for the same data (Fig. 4). This indicates that the proposed SwarnSeq performed better to detect DE genes as compared to other competitive methods. Similar interpretations can be made for other remaining datasets (Figs. 4, S13). Under this FDR based comparison setting on multiple real scRNA-seq datasets, we demonstrated our SwarnSeq method was consistently better and more robust to detect the DE genes of various sizes with respect to bulk and scRNA-seq DE tools.

3.3.3. Benchmarking based on other performance metrics

This comparison setting included the performance evaluation of the 12 scRNA-seq DE tools (Table S7) based on performance metrics, *viz.* TP, TN, FN, FP, FPR, NPV, F1, and ACC on the 10 scRNA-seq datasets (Table 1). For this purpose, the DE methods were applied to each dataset following the instructions and recommendations of their respective software packages. Genes were declared as DE based on their computed *p*-values/adjusted *p*-values and subsequently DE gene sets of sizes 500, 1000, 1500, …, 3000 were selected for each of the datasets. Then, the performance metrics were computed for the DE gene sets from different datasets and the results are given in Tables 3, S10-S18.

In this comparison setting, for a DE gene set of size 500, the SwarnSeq method identified more TP genes, followed by DECENT as compared to other competitive methods in GSE29087 data (Table 3). Further, the value of FP, FN, and FPR for the SwarnSeq was observed to be lower than other competitive methods. Moreover, the values of TPR, NPV, ACC, and F1 for SwarnSeq method were found to be higher than from other methods (Table 3). This finding indicated the better performance of our proposed method in terms of various computed performance metrics for the GSE29087 dataset. We demonstrated the consistently similar findings for our method over other DE gene sets of sizes 1000, 1500, 2000, 2500, and 3000 (Table 3). Similar interpretations can be made for other datasets, as shown in Supplementary Tables S10-S18. The comparative analysis under this setting gave us confidence that our SwarnSeq method can detect the genes, which are truly DE in wide range of real datasets. Furthermore, its performance was consistently better over the considered competitive scRNA-seq DE methods, when assessed through various performance metrics.

3.3.4. Benchmarking based on runtime metric

The computational speed for processing the large-scale counts data is an important factor in single-cell data analytics. Therefore, we evaluated the proposed SwarnSeq method with respect to the existing techniques (Tables S7, S8) based on runtime metric, where the runtime refers to the computational time required to analyze the data. Through this, the method which requires less runtime was considered better and *vice-versa*. To measure this, we ran the code written in R (*v* 4.0.2) for each tested method 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 a large experimental single-cell dataset, *i.e.*, GSE115469 data with 5466 cells and 17,316 genes. 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 detailed runtime-based performance analysis of the methods including the proposed SwarnSeq method is shown in Supplementary Document S12. This performance analysis indicated that the DECENT was the slowest and more computationally intensive method followed by DEsingle

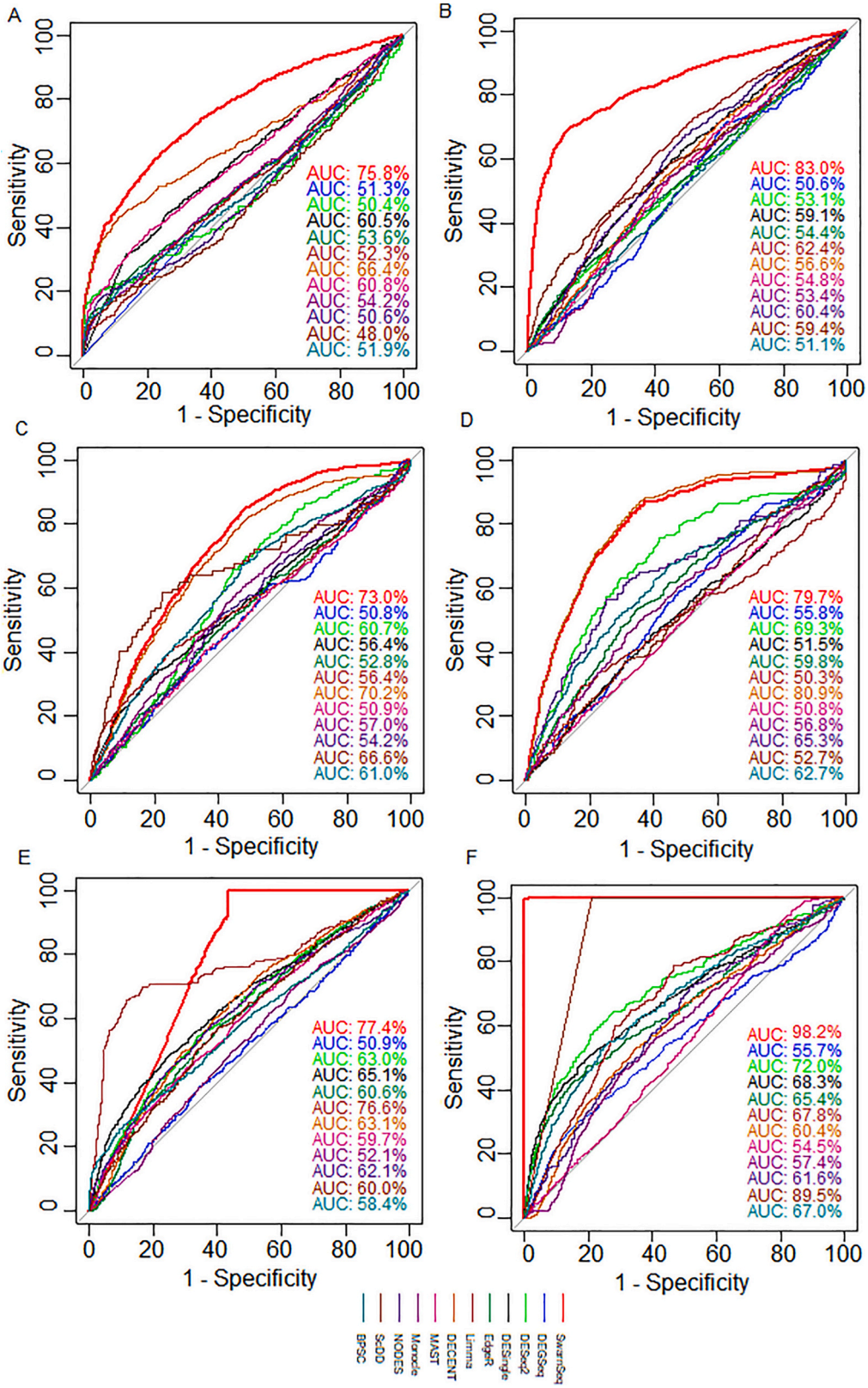

Fig. 3. Differential expression analysis of real scRNA-seq data. Receiver Operating Characteristic curves for differential expression methods on different real scRNA-seq data. Evaluation of the performance of different methods based on Area Under Receiver Operating Characteristic Curves (AUROC) is shown for (A) GSE53638 (Data 1); (B) GSE77728; (C) GSE53638 (Data 3); (D) GSE53638 (Data 2); (E) GSE29087; (F) GSE65525. Different reference gene lists, prepared based on the procedure given in Supplementary Document S10, used for benchmarking various differential expression analysis methods on different real scRNA-seq datasets. SwarnSeq achieves competitive and better accuarcy for identifying genuine differential gene lists in all six different real datasets. DE methods are denoted by different colors.

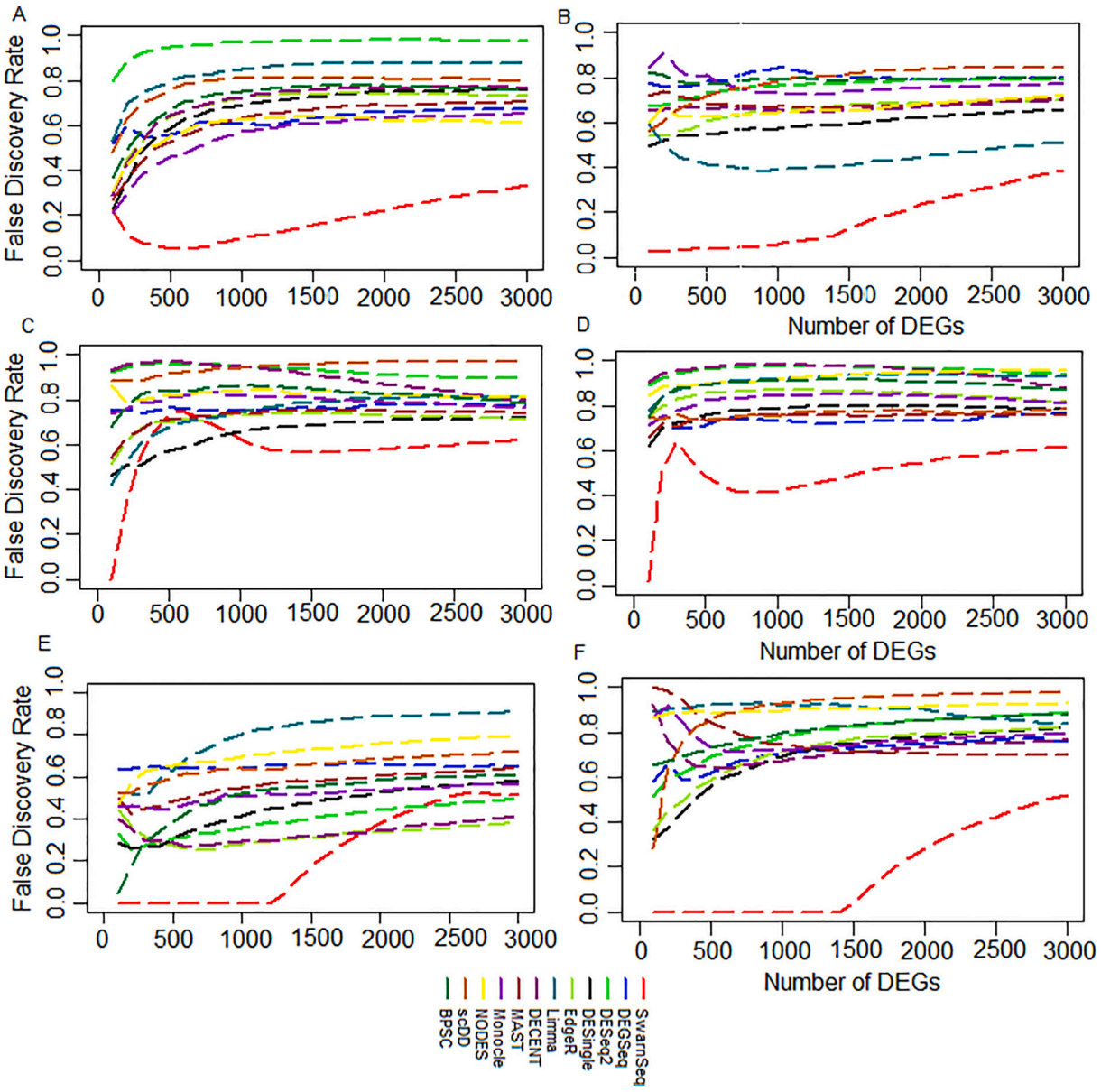

Fig. 4. FDR based Performance analysis of DE methods on real scRNA-seq data. FDR curves for differential expression methods on different real scRNA-seq data are shown. Evaluation of the performance of different methods based on false discovery rate is shown for (A) GSE53638 (Data 1); (B) GSE77728; (C) GSE53638 (Data 3); (D) GSE53638 (Data 2); (E) GSE29087; (F) GSE65525. Different reference gene lists, prepared based on the procedure given in Supplementary Document S10, used for benchmarking various differential expression analysis methods on different real scRNA-seq datasets. SwarnSeq achieves robust performance for identifying genuine differential gene lists in all four different real datasets. DE methods are denoted by different colors.

(Table S9) for a large single-cell dataset such as GSE115469 data. Further, it was found that the proposed SwarnSeq method required relatively lesser computational processing time compared to other zero inflated model-based methods, such as DECENT and DEsingle (Table S9).

3.3.5. *Effect of spike-in on performance*

We evaluated the performance of the SwarnSeq method on (GSE77288) data from Tung et al., for which spike-in and molecular concentration data is publicly available. For this purpose, we considered the following comparison settings: (a) spike-in data available; (b) spike-in not available (capture rates estimated from the data); (c) data unadjusted with cell capture rates. In other words, this comparison setting allowed us to examine the impact of external spike-ins and further capture rates on the performance of SwarnSeq method. The results are shown in Fig. 5 and Table 4. It was observed that the SwarnSeq performed better when capture rates were estimated from external spike-ins as assessed in terms of AUROC (Fig. 5). However, there was a decrease in AUROC value when the capture rates of cells were estimated from the count data (Fig. 5). Further, the SwarnSeq had the least AUROC when the observed counts were not adjusted with cell capture rates.

Under the FDR based comparison setting, the SwarnSeq had the smallest FDR values, when the capture rates of cells were estimated from the spike-in data (Fig. 5). Further, SwarnSeq performed poorly when the observed counts were not adjusted with capture rates of the cells, as compared to the adjusted scRNA-seq counts. The results from the third comparison setting, *i.e.,* comparative analysis based on performance metrics, are shown in Table 4. It was found that when the capture rates were estimated from the spike-ins and incorporated in SwarnSeq, its performance was better as compared to other two situations (Fig. 5, Table 4). Thus, we have convincingly demonstrated the viability of using the external spike-in capture rates for endogenous RNA in SwarnSeq

Table 3
Performance evaluation metrics for GSE 29087 scRNA-seq data.

Methods	TP	FP	TN	FN	TPR	FPR	PPR	NPV	ACC	F1
NDEG = 500										
SwarnSeq	500	0	8436	2500	0.167	0.000	1.000	0.771	0.781	0.286
DEGSeq	181	319	8140	2819	0.060	0.038	0.362	0.743	0.726	0.103
DESeq2	346	154	8282	2654	0.115	0.018	0.692	0.757	0.754	0.198
DEsingle	344	156	8280	2656	0.115	0.018	0.688	0.757	0.754	0.197
EdgeR	364	136	8302	2636	0.121	0.016	0.728	0.759	0.758	0.208
Limma	182	318	8188	2818	0.061	0.037	0.364	0.744	0.727	0.104
DECENT	355	145	8291	2645	0.118	0.017	0.710	0.758	0.756	0.203
MAST	258	242	8195	2742	0.086	0.029	0.516	0.749	0.739	0.147
Monocle	275	225	8212	2725	0.092	0.027	0.550	0.751	0.742	0.157
NODES	174	326	8110	2826	0.058	0.039	0.348	0.742	0.724	0.099
scDD	202	298	8138	2798	0.067	0.035	0.404	0.744	0.729	0.115
BPSC	308	192	8244	2692	0.103	0.023	0.616	0.754	0.748	0.176
NDEG = 1000										
SwarnSeq	1000	0	8436	2000	0.333	0.000	1.000	0.808	0.825	0.500
DEGSeq	357	643	7846	2643	0.119	0.076	0.357	0.748	0.714	0.179
DESeq2	641	359	8077	2359	0.214	0.043	0.641	0.774	0.762	0.321
DEsingle	585	415	8021	2415	0.195	0.049	0.585	0.769	0.753	0.293
EdgeR	718	282	8164	2282	0.239	0.033	0.718	0.782	0.776	0.359
Limma	198	802	8126	2802	0.066	0.090	0.198	0.744	0.698	0.099
DECENT	706	294	8142	2294	0.235	0.035	0.706	0.780	0.774	0.353
MAST	449	551	7894	2551	0.150	0.065	0.449	0.756	0.729	0.225
Monocle	495	505	7934	2505	0.165	0.060	0.495	0.760	0.737	0.248
NODES	301	699	7737	2699	0.100	0.083	0.301	0.741	0.703	0.151
scDD	362	638	7798	2638	0.121	0.076	0.362	0.747	0.714	0.181
BPSC	481	519	7917	2519	0.160	0.062	0.481	0.759	0.734	0.241
NDEG = 1500										
SwarnSeq	1242	258	8178	1758	0.414	0.031	0.828	0.823	0.824	0.552
DEGSeq	510	990	7539	2490	0.170	0.116	0.340	0.752	0.698	0.227
DESeq2	886	614	7822	2114	0.295	0.073	0.591	0.787	0.761	0.394
DEsingle	782	718	7718	2218	0.261	0.085	0.521	0.777	0.743	0.348
EdgeR	1037	463	7997	1963	0.346	0.055	0.691	0.803	0.788	0.461
Limma	212	1288	8052	2788	0.071	0.138	0.141	0.743	0.670	0.094
DECENT	1025	475	7961	1975	0.342	0.056	0.683	0.801	0.786	0.456
MAST	630	870	7589	2370	0.210	0.103	0.420	0.762	0.717	0.280
Monocle	720	780	7663	2280	0.240	0.092	0.480	0.771	0.733	0.320
NODES	403	1097	7339	2597	0.134	0.130	0.269	0.739	0.677	0.179
scDD	513	987	7449	2487	0.171	0.117	0.342	0.750	0.696	0.228
BPSC	671	829	7607	2329	0.224	0.098	0.447	0.766	0.724	0.298
NDEG = 2000										
SwarnSeq	1320	680	7803	1680	0.440	0.080	0.660	0.823	0.794	0.528
DEGSeq	682	1318	7238	2318	0.227	0.154	0.341	0.757	0.685	0.273
DESeq2	1117	883	7553	1883	0.372	0.105	0.559	0.800	0.758	0.447
DEsingle	946	1054	7382	2054	0.315	0.125	0.473	0.782	0.728	0.378
EdgeR	1242	758	7678	1758	0.414	0.090	0.621	0.814	0.780	0.497
Limma	228	1772	7978	2772	0.076	0.182	0.114	0.742	0.644	0.091
DECENT	1314	686	7757	1686	0.438	0.081	0.657	0.821	0.793	0.526
MAST	795	1205	7289	2205	0.265	0.142	0.398	0.768	0.703	0.318
Monocle	925	1075	7376	2075	0.308	0.127	0.463	0.780	0.725	0.370
NODES	485	1515	6925	2515	0.162	0.180	0.243	0.734	0.648	0.194
scDD	633	1367	7069	2367	0.211	0.162	0.317	0.749	0.673	0.253
BPSC	832	1168	7268	2168	0.277	0.138	0.416	0.770	0.708	0.333
NDEG = 2500										
SwarnSeq	1601	899	7612	1399	0.534	0.106	0.640	0.845	0.800	0.582
DEGSeq	874	1626	6966	2126	0.291	0.189	0.350	0.766	0.676	0.318
DESeq2	1327	1173	7263	1673	0.442	0.139	0.531	0.813	0.751	0.483
DEsingle	1103	1397	7039	1897	0.368	0.166	0.441	0.788	0.712	0.401
EdgeR	1242	1258	7178	1758	0.414	0.149	0.497	0.803	0.736	0.452
Limma	255	2245	7915	2745	0.085	0.221	0.102	0.742	0.621	0.093
DECENT	1548	952	7499	1452	0.516	0.113	0.619	0.838	0.790	0.563
MAST	945	1555	6973	2055	0.315	0.182	0.378	0.772	0.687	0.344
Monocle	1114	1386	7071	1886	0.371	0.164	0.446	0.789	0.714	0.405
NODES	556	1944	6517	2444	0.185	0.230	0.222	0.727	0.617	0.202
scDD	744	1756	6680	2256	0.248	0.208	0.298	0.748	0.649	0.271
BPSC	999	1501	6935	2001	0.333	0.178	0.400	0.776	0.694	0.363
NDEG = 3000										
SwarnSeq	1837	1163	7386	1163	0.612	0.136	0.612	0.864	0.799	0.612
DEGSeq	1055	1945	6681	1945	0.352	0.225	0.352	0.775	0.665	0.352
DESeq2	1502	1498	6938	1498	0.501	0.178	0.501	0.822	0.738	0.501
DEsingle	1249	1751	6685	1751	0.416	0.208	0.416	0.792	0.694	0.416
EdgeR	1450	1550	6886	1550	0.483	0.184	0.483	0.816	0.729	0.483
Limma	279	2721	7813	2721	0.093	0.258	0.093	0.742	0.598	0.093

(*continued on next page*)

Table 3 (*continued*)

Methods	TP	FP	TN	FN	TPR	FPR	PPR	NPV	ACC	F1
DECENT	1754	1246	7217	1246	0.585	0.147	0.585	0.853	0.783	0.585
MAST	1074	1926	6651	1926	0.358	0.225	0.358	0.775	0.667	0.358
Monocle	1303	1697	6769	1697	0.434	0.200	0.434	0.800	0.704	0.434
NODES	633	2367	6106	2367	0.211	0.279	0.211	0.721	0.587	0.211
scDD	845	2155	6281	2155	0.282	0.255	0.282	0.745	0.623	0.282
BPSC	1181	1819	6617	1819	0.394	0.216	0.394	0.784	0.682	0.394

TP: True Positive; FP: False Positive; TN: True Negative; FN: False Negative; TPR: True Positive Rate;
FPR: False Positive Rate; PPR: Positive Prediction Rate; NPV: Negative Prediction Value;
ACC: Accuracy; F1: F-score.

modeling, and subsequently found its performance is both robust and better. The relation among the gene and cell parameters estimated through the SwarnSeq method for GSE77288 data, when external spike-ins data are available, is also shown in Fig. 5.

3.4. SwarnSeq as differential zero inflation tool

The SwarnSeq method provides an excellent platform for performing DZI analysis of genes. DZI genes are detected using the ZINB model under the GLM framework. For identification of DZI genes on real dataset, we set 1E-10 as threshold for adjusted *p*-values computed through the SwarnSeq. The results are shown in Supplementary Table S19. For instance, at this threshold value we identified 2936 genes as DZI for GSE29087 data (Table S19). This means, 2936 genes have a significant difference in the number of cells whose expressions are zeros across two cellular groups. Similar interpretations can be made for other datasets (Table S19).

Our SwarnSeq model provides an opportunity to classify the influential genes into gene types with respect to their differential zero inflation and expression. Through this, the identified genes can be grouped into various gene types, and the results are shown in Table 5. For instance, the GSE29087 data, the SwarnSeq identified 4930 genes as DE, 2789 genes as DEZI, and 149 as only DZI (Table 5). This means that out of 15,234 genes, the mean expression of non-zero counts of 4930 genes are expressed differentially across the two cellular groups. While, for 2789 genes, there is a significant proportion of cells whose expressions are zero across two cellular populations (however the mean of non-zeros counts of these genes in the remaining cells are significantly different) and only 149 genes had a significant number of cells as zero expressions across the cellular populations (Table 5). Similar type of interpretations can be made for other datasets, as shown in Table 5.

4. Discussion

We presented the SwarnSeq, an improved statistical method, for performing analysis on UMI counts data obtained from scRNA-seq study. Our method is capable of performing reliable statistical tests on gene mean abundance, zero inflation, and classification of influential genes in scRNA-seq expression data. It uses the ZINB model to model the observed UMI counts. Further, the UMI provides an excellent opportunity to model the transcriptional capturing process through the SwarnSeq. In other words, the observed counts data are adjusted with cell capture rates through a binomial model in the proposed approach. Moreover, RNA spike-ins data including the ERCC spike-ins [33], can give valuable insights into the technical variation in scRNA-seq study. This raises a key question of whether and how to use spike-ins in data analyses. For instance, when they are available, they can be used to estimate the capture rates for cells. This property is well integrated in our SwarnSeq approach. Thus, the SwarnSeq is capable of modeling capture rates using spike-ins data, if they are available and can estimate the capture rates from the observed data, if spike-ins are not available. We established a statistical theory for adjusting the UMI counts data with the molecular capturing process derived from real scRNA-seq experiments. Moreover, the SwarnSeq operates through various analytical steps including, pre-processing, normalization, estimation of gene parameters, detection of DE genes, and DZI genes, selection of top genes, and classification of genes into sub-types. The SwarnSeq method employs different normalization methods such as modified median normalization [16] and trimmed mean of M values [14] to remove the amplification bias from the scRNA-seq count data. Thus, SwarnSeq is compatible with different normalization strategies.

Here, we established the statistical basis for the distributional nature of the observed scRNA-seq count in presence of cell capture rates. Further, we empirically showed the suitability of the ZINB model for fitting zero inflated, and overdispersed count data over other models, such as NB, PD, HD, and ZIPD. Moreover, the study of ZINB *vs.* NB model for estimation of parameters indicated that the latter overestimated the dispersion to accommodate excess overdispersion and underestimated the mean to accommodate the zero inflation present in scRNA-seq data. In scRNA-seq data, factors such as technical noise, dropout events, and low molecular capturing have substantial overdispersion and zero-inflation, and a NB model is not appropriate. Hence, we implemented the ZINB model in our SwarnSeq method to fit the observed scRNA-seq count data and to obtain better estimates of the gene-wise parameters.

The SwarnSeq method models the unwanted variation in mean transcript abundance of genes attributed to different sources, such as cellular groups, cell clusters, and other cell co-variates. This means, that provides reliable MLEs of the effects of the cellular groups, cell clusters, and cell co-variates using the EM algorithm. Further, it detects the influential genes which are DE under a GLM framework. Here, these genes were identified based on the statistical significance values adjusted over multiple hypothesis testing. This provides statistically meaningful and biologically interpretable values in [0,1] for genes in scRNA-seq data. The benchmarking of methods indicated the better performance of our SwarnSeq method over other existing competitive methods. This comparative analysis was carried out on three different comparison settings, *i.e.,* AUROC, FDR, and other performance metrics on multiple real scRNA-seq datasets.

The SwarnSeq method can also be extended to carry out other types of tests, including the differential testing of zero proportions of genes across the cell populations. Here, we considered the zero-inflation parameter of genes as a function of the effects of cellular populations, cell clusters, and other cell co-variates. Then, a linear logit model was used to test for biological differences in zero inflation. To statistically measure this, a statistical significance value adjusted over multiple hypothesis testing, was assigned to each gene. This measure provided biologically interpretable values to genes and showed there were significant differences in the proportion of zeros across the cellular populations. The available scRNA-seq tools are mostly focused on performing the DE analysis of genes and ignores the zero-inflation analysis which is an integral part of the scRNA-seq experiments. Our SwarnSeq method can perform DZI analysis including DE analysis of genes using the observed UMI counts data adjusted over molecular capturing process. Additionally, it provides options for classifying the detected influential genes into various gene types according to their differential expression and zero inflation.

Multilevel statistical models fitted with an EM algorithm are computationally intensive and time consuming. Further, the ZINB

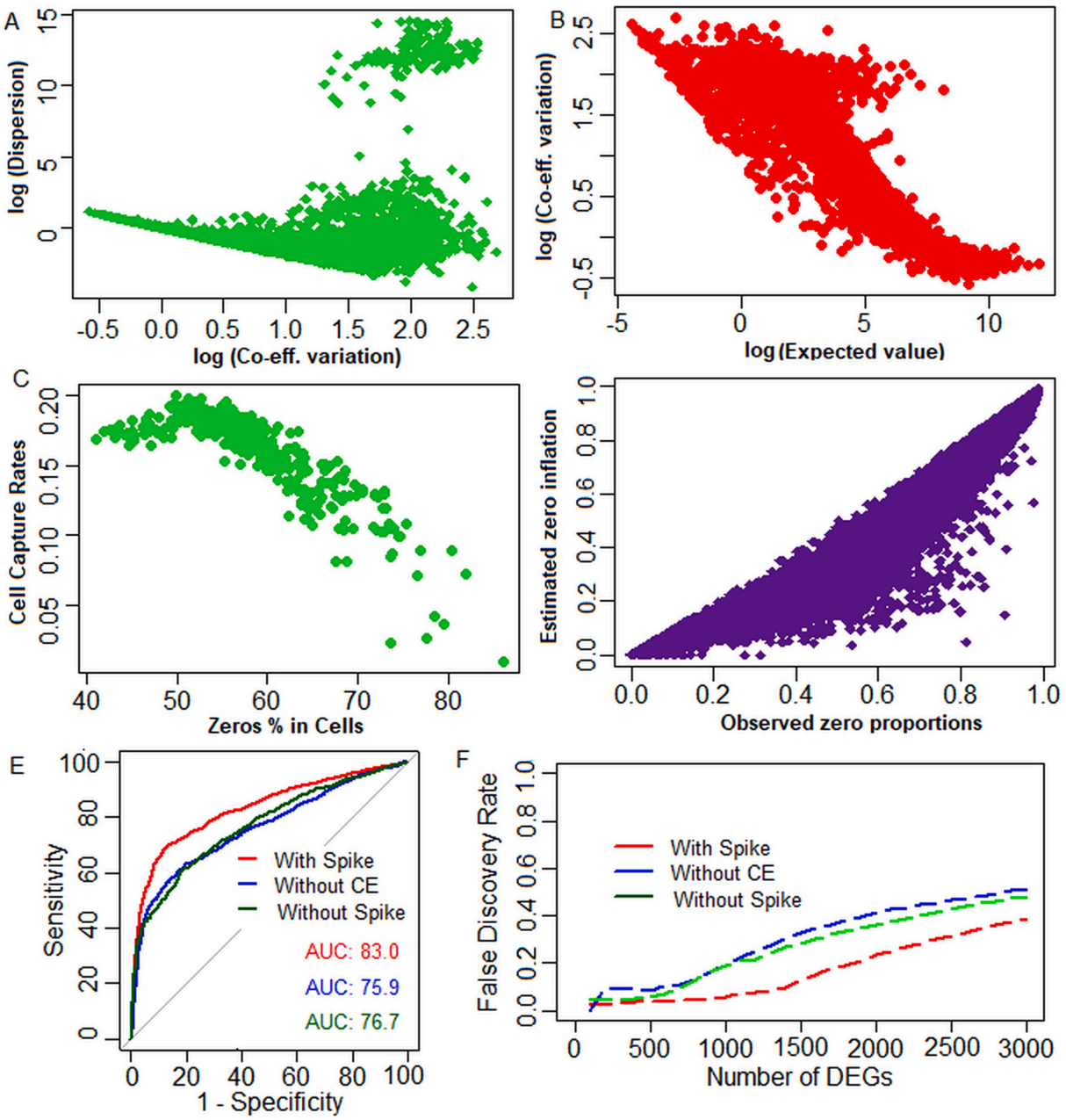

Fig. 5. Performance analysis of SwarnSeq method on real scRNA-seq data in presence of spike-ins. (A) Scatter plot showing the relationship of mean and dispersion parameters of the genes. (B) Scatter plot comparing the observed value of zero proportions and estimated zero inflation parameters of genes (C) The ROC curves are shown for SwarnSeq method (i) when spike-in information is considered (red); (ii) when spike-in data are not considered and capture efficiencies are estimated from the data (green); and (iii) Unadjusted for capture efficiency. (D) The false discovery rate curves of SwarnSeq method on real scRNA-seq data under different conditions: (i) when spike-in information is considered (red); (ii) when spike-in data are not considered and capture efficiencies are estimated from the data (green); (iii) Unadjusted for capture efficiency. (E) Various performance measures are listed for SwarnSeq method under different conditions of (i), (ii) and (iii). Here, real data from GSE77288 are considered, as ERCC spike-in data is available for this experiment. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

Table 4
Effects of external RNA spike-ins data on performance of the SwarnSeq method.

	TP	FP	TN	FN	TPR	FPR	PPR	NPV	ACC	F1
NDEG = 500										
With Spike	482	18	12,937	2518	0.161	0.001	0.964	0.837	0.841	0.275
Unadjusted	457	43	12,912	2543	0.152	0.003	0.914	0.835	0.838	0.261
Without spike	468	27	12,928	2532	0.156	0.002	0.945	0.836	0.840	0.268
NDEG = 1000										
With Spike	946	54	12,901	2054	0.315	0.004	0.946	0.863	0.868	0.473
Unadjusted	809	191	12,764	2191	0.270	0.015	0.809	0.853	0.851	0.405
Without spike	795	187	12,768	2205	0.265	0.014	0.810	0.853	0.850	0.399
NDEG = 1500										
With Spike	1311	189	12,766	1689	0.437	0.015	0.874	0.883	0.882	0.583
Unadjusted	1008	492	12,463	1992	0.336	0.038	0.672	0.862	0.844	0.448
Without spike	1059	416	12,539	1941	0.353	0.032	0.718	0.866	0.852	0.473
NDEG = 2000										
With Spike	1532	468	12,487	1468	0.511	0.036	0.234	0.766	0.895	0.879
Unadjusted	1175	825	12,130	1825	0.392	0.064	0.413	0.588	0.869	0.834
Without spike	1257	713	12,242	1743	0.419	0.055	0.362	0.638	0.875	0.846
NDEG = 2500										
With Spike	1713	787	12,168	1287	0.571	0.061	0.685	0.904	0.870	0.623
Unadjusted	1342	1158	11,797	1658	0.447	0.089	0.537	0.877	0.824	0.488
Without spike	1414	1053	11,902	1586	0.471	0.081	0.573	0.882	0.835	0.517
NDEG = 3000										
With Spike	1846	1154	11,801	1154	0.615	0.089	0.615	0.911	0.855	0.615
Unadjusted	1466	1534	11,421	1534	0.489	0.118	0.489	0.882	0.808	0.489
Without spike	1542	1424	11,531	1458	0.514	0.110	0.520	0.888	0.819	0.517

With Spike: Observed read counts are adjusted with capture efficiency when spike-in data is available; Without spike: Observed read counts are adjusted with capture efficiency estimated from the data itself; Unadjusted: Observed read counts are not adjusted with capture efficiency; TP: True Positive; FP: False Positive; TN: True Negative; FN: False Negative; TPR: True Positive Rate; FPR: False Positive Rate; PPR: Positive Prediction Rate; NPV: Negative Prediction Value; ACC: Accuracy; F1: F-score.

Table 5
Classification of DE and DZI genes.

Datasets	DE	DEZI	DZI	Non-DE
GSE29087	4930	2789	149	3567
GSE53638dt1	2406	278	408	11,771
GSE53638dt2	1831	2789	5013	6004
GSE53638dt3	1733	3101	4673	5507
GSE65525	2033	15,194	5799	929
GSE75790	3993	9874	2865	3852
GSE92495	757	324	5	14,438
GSE109999	5694	6386	71	903
GSE111108	27	7187	87	10,021
GSE115469	24	7745	6296	3231
GSE77288	1426	119	619	13,791

DE: Differentially Expressed; DZI: Differentially Zero Inflated;
DEZI: Both DE and DZI.

models are implemented in several tools like DEsingle [22], DECENT [9], which are time consuming for a PC. However, the SwarnSeq method required less computational time than DECENT and DEsingle with better performance in terms of detecting DE genes along with additional features. Moreover, it can even be used on a PC or workstation computer for analyzing large scRNA-seq datasets with better and robust performance.

5. Software implementation

Our novel SwarnSeq approach is implemented in an R software package, which is available at https://github.com/sam-uofl/SwarnSeq. This R package provides *optimcluster* function for getting the optimum number of cell clusters from scRNA-seq count data. Additionally, it also provides option for estimation of capture rates of cells using different methods, *e.g.,* MLE, regression, *etc.*, whether RNA spike-in data is available or not. The function SwarnSeq implemented in SwarnSeq R package can be executed for estimating the parameters for each gene, *i.e.,* mean, dispersion, zero inflation, effects of groups, cell clusters and cell level auxiliary information on zero-inflation as well as means of non-zero counts. SwarnSeqLRT function provides option for results from DE analysis and DZI analysis, when the observed UMI counts are adjusted for molecular capture rates. Moreover, functions like SwarnUnadjSeq and SwarnUnadjLRT are implemented for parameter estimation and DE, and DZI analyses respectively, when the users do not need to adjust the count data for capture efficiency. The top influential genes detected through SwarnSeq approach can be selected and classified through the implemented SwarnClass and SwarnTopTags functions, respectively. Different options are provided in the SwarnSeq R package for adjusting the capturing process and correcting amplification bias through different normalization methods. The detail manual for the usage of the SwarnSeq tool with suitable examples is given in Supplementary Document S14.

6. Conclusion

In this study, we proposed an improved and novel statistical approach for analysis of scRNA-seq data. This approach can perform analysis including DE, DZI, classification of genes, estimation of cell capture rate, and determination of optimum number of cell clusters with strong statistical basis. Here, we provided all the background statistical theory, data example, preliminary data and real experimental data analysis results for our SwarnSeq approach. The benchmarking of the SwarnSeq method on multiple real datasets over a wide range of statistical criteria indicated its better performance over the existing methods. Further, the SwarnSeq method will surely help the experimental biologist and genome researchers to identify true DE genes for their experiments. Our comparison framework may be adopted for further comparative study of scRNA-seq DE tools. In future, parameter estimation procedure, like Empirical Bayes shrinkage method can be implemented in the SwarnSeq tool to estimate the gene specific dispersion, and that will enhance its performance. The SwarnSeq assumes the factors, such as cellular groups, cell clusters and other covariates, have fixed effects on means and zero inflations. This

assumption may be unrealistic from a biological standpoint (some may have random effects). Therefore, researchers may think of random or mixed effect models in SwarnSeq in the future to improve its performance.

Funding

Samarendra Das: Indian Council of Agricultural Research (ICAR), New Delhi, India (Netaji Subhas-ICAR International Fellowship, OM No. 18(02)/2016-EQR/Edn) Shesh N. Rai: Clinical Trial Research Fund (Wendell Cherry Chair), JG Brown Cancer Center, USA; multiple National Institutes of Health (NIH), USA grants (5P20GM113226, PI: McClain; 1P42ES023716, PI: Srivastava; 5P30GM127607-02, PI: Jones; 1P20GM125504-01, PI: Lamont; 2U54HL120163, PI: Bhatnagar/Robertson; 1P20GM135004, PI: Yan; 1R35ES0238373-01, PI: Cave; 1R01ES029846, PI: Bhatnagar; 1P30ES030283, PI: States); Kentucky Council on Postsecondary Education grant, USA (PON2 415 1900002934, PI: Chesney)

Availability of data and materials

All the datasets used in this study are publicly available in NCBI GEO database. For Tung et al. dataset, the UMI count, ERCC spike-ins and molecular concentration datasets were taken from the GitHub repository (https://github.com/jdblischak/singleCellSeq). An R software package for our novel SwarnSeq method is available at https://github.com/sam-uofl/SwarnSeq for this manuscript.

CRediT authorship contribution statement

Samarendra Das: Conceptualization, Investigation, Data curation, Formal analysis, Methodology, Software, Validation, Visualization, Writing - original draft, writing-review & editing. **Shesh N. Rai:** Project administration, Supervision, Funding acquisition, Writing - review & editing.

Declaration of Competing Interest

Authors declare that they have no competing interests.

Acknowledgement

Authors duly acknowledge the help and support obtained from Education Division, ICAR, New Delhi, India and ICAR-Indian Agricultural Statistics Research Institute, New Delhi, India. Also, the authors acknowledge Mrs. Marion McClain for the language checking and editing related helps.

Appendix A. Supplementary data

Supplementary data to this article can be found online at https://doi.org/10.1016/j.ygeno.2021.02.014.

References

[1] C. Trapnell, Defining cell types and states with single-cell genomics, Genome Res. (2015), https://doi.org/10.1101/gr.190595.115.
[2] B. Tasic, V. Menon, T.N. Nguyen, T.K. Kim, T. Jarsky, Z. Yao, et al., Adult mouse cortical cell taxonomy revealed by single cell transcriptomics, Nat. Neurosci. 19 (2016) 335–346, https://doi.org/10.1038/nn.4216.
[3] G. Chen, B. Ning, T. Shi, Single-cell RNA-seq technologies and related computational data analysis, Front. Genet. 10 (2019), https://doi.org/10.3389/fgene.2019.00317.
[4] A. Zeisel, A.B. Muñoz-Manchado, S. Codeluppi, P. Lönnerberg, Manno G. La, A. Juréus, et al., Cell types in the mouse cortex and hippocampus revealed by single-cell RNA-seq, Science (80-) (2015), https://doi.org/10.1126/science.aaa1934.
[5] A.M. Klein, L. Mazutis, I. Akartuna, N. Tallapragada, A. Veres, V. Li, et al., Droplet barcoding for single-cell transcriptomics applied to embryonic stem cells, Cell (2015), https://doi.org/10.1016/j.cell.2015.04.044.
[6] L. Tian, S. Su, X. Dong, D. Amann-Zalcenstein, C. Biben, A. Seidi, et al., scPipe: a flexible R/Bioconductor preprocessing pipeline for single-cell RNA-sequencing data, PLoS Comput. Biol. (2018), https://doi.org/10.1371/journal.pcbi.1006361.
[7] S. Islam, U. Kjällquist, A. Moliner, P. Zajac, J.B. Fan, P. Lönnerberg, et al., Characterization of the single-cell transcriptional landscape by highly multiplex RNA-seq, Genome Res. (2011), https://doi.org/10.1101/gr.110882.110.
[8] G. Finak, A. McDavid, M. Yajima, J. Deng, V. Gersuk, A.K. Shalek, et al., MAST: a flexible statistical framework for assessing transcriptional changes and characterizing heterogeneity in single-cell RNA sequencing data, Genome Biol. 16 (2015) 278, https://doi.org/10.1186/s13059-015-0844-5.
[9] C. Ye, T.P. Speed, A. Salim, DECENT: differential expression with capture efficiency adjustmeNT for single-cell RNA-seq data. Berger B, editor, Bioinformatics. 35 (2019) 5155–5162, https://doi.org/10.1093/bioinformatics/btz453.
[10] D. Ramsköld, S. Luo, Y.C. Wang, R. Li, Q. Deng, O.R. Faridani, et al., Full-length mRNA-Seq from single-cell levels of RNA and individual circulating tumor cells, Nat. Biotechnol. (2012), https://doi.org/10.1038/nbt.2282.
[11] T. Hashimshony, F. Wagner, N. Sher, I. Yanai, CEL-Seq: single-cell RNA-Seq by multiplexed linear amplification, Cell Rep. (2012), https://doi.org/10.1016/j.celrep.2012.08.003.
[12] P.-Y. Tung, J.D. Blischak, C.J. Hsiao, D.A. Knowles, J.E. Burnett, J.K. Pritchard, et al., Batch effects and the effective design of single-cell gene expression studies, Sci. Rep. 7 (2017) 39921, https://doi.org/10.1038/srep39921.
[13] K. Van den Berge, F. Perraudeau, C. Soneson, M.I. Love, D. Risso, J.-P. Vert, et al., Observation weights unlock bulk RNA-seq tools for zero inflation and single-cell applications, Genome Biol. 19 (2018) 24, https://doi.org/10.1186/s13059-018-1406-4.
[14] M.D. Robinson, D.J. McCarthy, G.K. Smyth, edgeR: a bioconductor package for differential expression analysis of digital gene expression data, Bioinformatics 26 (2010) 139–140, https://doi.org/10.1093/bioinformatics/btp616.
[15] S. Anders, W. Huber, Differential expression analysis for sequence count data, Genome Biol. 11 (2010) R106, https://doi.org/10.1186/gb-2010-11-10-r106.
[16] M.I. Love, S. Anders, W. Huber, Differential analysis of count data - the DESeq2 package, Genome Biol. (2014), https://doi.org/10.1186/s13059-014-0550-8.
[17] C.W. Law, Y. Chen, W. Shi, G.K. Smyth, voom: precision weights unlock linear model analysis tools for RNA-seq read counts, Genome Biol. 15 (2014) R29, https://doi.org/10.1186/gb-2014-15-2-r29.
[18] M.E. Ritchie, B. Phipson, D. Wu, Y. Hu, C.W. Law, W. Shi, et al., Limma powers differential expression analyses for RNA-sequencing and microarray studies, Nucleic Acids Res. 43 (2015), https://doi.org/10.1093/nar/gkv007 e47.
[19] L. Wang, Z. Feng, X. Wang, X. Wang, X. Zhang, DEGseq: an R package for identifying differentially expressed genes from RNA-seq data, Bioinformatics (2009), https://doi.org/10.1093/bioinformatics/btp612.
[20] K. Fujita, M. Iwaki, T. Yanagida, Transcriptional bursting is intrinsically caused by interplay between RNA polymerases on DNA, Nat. Commun. (2016), https://doi.org/10.1038/ncomms13788.
[21] J. Wang, M. Huang, E. Torre, H. Dueck, S. Shaffer, J. Murray, et al., Gene expression distribution deconvolution in single-cell RNA sequencing, Proc. Natl. Acad. Sci. U. S. A. (2018), https://doi.org/10.1073/pnas.1721085115.
[22] Z. Miao, K. Deng, X. Wang, X. Zhang, DEsingle for detecting three types of differential expression in single-cell RNA-seq data. Berger B, editor, Bioinformatics 34 (2018) 3223–3224, https://doi.org/10.1093/bioinformatics/bty332.
[23] C. Trapnell, D. Cacchiarelli, J. Grimsby, P. Pokharel, S. Li, M. Morse, et al., The dynamics and regulators of cell fate decisions are revealed by pseudotemporal ordering of single cells, Nat. Biotechnol. (2014), https://doi.org/10.1038/nbt.2859.
[24] X. Qiu, A. Hill, J. Packer, D. Lin, Y.-A. Ma, C. Trapnell, Single-cell mRNA quantification and differential analysis with census, Nat. Methods 14 (2017) 309–315, https://doi.org/10.1038/nmeth.4150.
[25] P.V. Kharchenko, L. Silberstein, D.T. Scadden, Bayesian approach to single-cell differential expression analysis, Nat. Methods 11 (2014) 740–742, https://doi.org/10.1038/nmeth.2967.
[26] T.N. Vu, Q.F. Wills, K.R. Kalari, N. Niu, L. Wang, M. Rantalainen, et al., Beta-Poisson model for single-cell RNA-seq data analyses, Bioinformatics (2016), https://doi.org/10.1093/bioinformatics/btw202.
[27] T. Mou, W. Deng, F. Gu, Y. Pawitan, T.N. Vu, Reproducibility of methods to detect differentially expressed genes from single-cell RNA sequencing, Front. Genet. (2020), https://doi.org/10.3389/fgene.2019.01331.
[28] M. Delmans, M. Hemberg, Discrete distributional differential expression (D3E) – a tool for gene expression analysis of single-cell RNA-seq data, BMC Bioinformatics (2016), https://doi.org/10.1186/s12859-016-0944-6.
[29] K.D. Korthauer, L.F. Chu, M.A. Newton, Y. Li, J. Thomson, R. Stewart, et al., A statistical approach for identifying differential distributions in single-cell RNA-seq experiments, Genome Biol. (2016), https://doi.org/10.1186/s13059-016-1077-y.
[30] X. Qiu, Q. Mao, Y. Tang, L. Wang, R. Chawla, H.A. Pliner, et al., Reversed graph embedding resolves complex single-cell trajectories, Nat. Methods (2017), https://doi.org/10.1038/nmeth.4402.
[31] W. Chen, Y. Li, J. Easton, D. Finkelstein, G. Wu, X. Chen, UMI-count modeling and differential expression analysis for single-cell RNA sequencing, Genome Biol. 19 (2018) 70, https://doi.org/10.1186/s13059-018-1438-9.
[32] D. Risso, F. Perraudeau, S. Gribkova, S. Dudoit, J.-P. Vert, A general and flexible method for signal extraction from single-cell RNA-seq data, Nat. Commun. 9 (2018) 284, https://doi.org/10.1038/s41467-017-02554-5.

[33] L. Jiang, F. Schlesinger, C.A. Davis, Y. Zhang, R. Li, M. Salit, et al., Synthetic spike-in standards for RNA-seq experiments, Genome Res. (2011), https://doi.org/10.1101/gr.121095.111.
[34] C.A. Vallejos, J.C. Marioni, S. Richardson, BASiCS: Bayesian Analysis of Single-Cell Sequencing Data. Morris Q, editor, PLoS Comput. Biol. 11 (2015), https://doi.org/10.1371/journal.pcbi.1004333 e1004333.
[35] C. Soneson, M.D. Robinson, Bias, robustness and scalability in single-cell differential expression analysis, Nat. Methods (2018), https://doi.org/10.1038/nmeth.4612.
[36] A. Dal Molin, G. Baruzzo, B. Di Camillo, Single-cell RNA-sequencing: assessment of differential expression analysis methods, Front. Genet. (2017), https://doi.org/10.3389/fgene.2017.00062.
[37] Z. Miao, X. Zhang, Differential expression analyses for single-cell RNA-Seq: old questions on new data, Quant. Biol. (2016), https://doi.org/10.1007/s40484-016-0089-7.
[38] M.K. Jaakkola, F. Seyednasrollah, A. Mehmood, L.L. Elo, Comparison of methods to detect differentially expressed genes between single-cell populations, Brief. Bioinform. (2016), https://doi.org/10.1093/bib/bbw057 bbw057.
[39] T. Wang, B. Li, C.E. Nelson, S. Nabavi, Comparative analysis of differential gene expression analysis tools for single-cell RNA sequencing data, BMC Bioinformatics (2019), https://doi.org/10.1186/s12859-019-2599-6.
[40] A. Duò, M.D. Robinson, C. Soneson, A systematic performance evaluation of clustering methods for single-cell RNA-seq data, F1000Research 7 (2018), https://doi.org/10.12688/f1000research.15666.1, 1141.
[41] D. Sengupta, N.A. Rayan, M. Lim, B. Lim, S. Prabhakar, Fast, scalable and accurate differential expression analysis for single cells, bioRxiv (2016), https://doi.org/10.1101/049734.
[42] S. Petropoulos, D. Edsgärd, B. Reinius, Q. Deng, S.P. Panula, S. Codeluppi, et al., Single-cell RNA-seq reveals lineage and X chromosome dynamics in human preimplantation embryos, Cell (2016), https://doi.org/10.1016/j.cell.2016.03.023.
[43] S.A. MacParland, J.C. Liu, X.Z. Ma, B.T. Innes, A.M. Bartczak, B.K. Gage, et al., Single cell RNA sequencing of human liver reveals distinct intrahepatic macrophage populations, Nat. Commun. (2018), https://doi.org/10.1038/s41467-018-06318-7.
[44] A.P. Dempster, N.M. Laird, D.B. Rubin, Maximum likelihood from incomplete data via the EM algorithm, J. R. Stat. Soc. Ser. B 39 (1977) 1–22, https://doi.org/10.1111/j.2517-6161.1977.tb01600.x.
[45] K. Van den Berge, C. Soneson, M.I. Love, M.D. Robinson, L. Clement, zingeR: unlocking RNA-seq tools for zero-inflation and single cell applications, BioRxiv (2017), https://doi.org/10.1101/157982.
[46] K.I.M. McKinnon, Convergence of the Nelder-Mead simplex method to a nonstationary point, SIAM J. Optim. (1998), https://doi.org/10.1137/S1052623496303482.
[47] C. Ziegenhain, B. Vieth, S. Parekh, B. Reinius, A. Guillaumet-Adkins, M. Smets, et al., Comparative analysis of single-cell RNA sequencing methods, Mol. Cell (2017), https://doi.org/10.1016/j.molcel.2017.01.023.
[48] D. Moriña, M. Higueras, P. Puig, M. Oliveira, Generalized Hermite Distribution Modelling with the R Package Hermite, R J, 2015, https://doi.org/10.32614/rj-2015-035.
[49] J.S. Long, J. Freese, Regression models for categorical dependent variables using STATA, Sociol. J. Br. Sociol. Ass. (2001), https://doi.org/10.1186/2051-3933-2-4.
[50] A. Moliner, P. Enfors, C.F. Ibáñez, M. Andäng, Mouse embryonic stem cell-derived spheres with distinct neurogenic potentials, Stem Cells Dev. (2008), https://doi.org/10.1089/scd.2007.0211.
[51] X. Robin, N. Turck, A. Hainard, N. Tiberti, F. Lisacek, J.C. Sanchez, et al., pROC: an open-source package for R and S+ to analyze and compare ROC curves, BMC Bioinformatics (2011), https://doi.org/10.1186/1471-2105-12-77.
[52] M. Soumillon, D. Cacchiarelli, S. Semrau, A. van Oudenaarden, T.S. Mikkelsen, Characterization of directed differentiation by high-throughput single-cell RNA-Seq, bioRxiv (2014), https://doi.org/10.1101/003236.
[53] T.M. Gierahn, M.H. Wadsworth, T.K. Hughes, B.D. Bryson, A. Butler, R. Satija, et al., Seq-well: portable, low-cost rna sequencing of single cells at high throughput, Nat. Methods (2017), https://doi.org/10.1038/nmeth.4179.