

Differential Expression Analysis of Single-Cell RNA-Seq Data: Current Statistical Approaches and Outstanding Challenges

Samarendra Das ^{1,2,*} , Anil Rai ³ and Shesh N. Rai ^{4,5,6,7,8,9,*} 

- ¹ ICAR-Directorate of Foot and Mouth Disease, Arugul, Bhubaneswar 752050, India
 - ² International Centre for Foot and Mouth Disease, Arugul, Bhubaneswar 752050, India
 - ³ ICAR-Indian Agricultural Statistics Research Institute, PUSA, New Delhi 110012, India; anil.ra@icar.gov.in
 - ⁴ School of Interdisciplinary and Graduate Studies, University of Louisville, Louisville, KY 40292, USA
 - ⁵ Biostatistics and Bioinformatics Facility, Brown Cancer Center, University of Louisville, Louisville, KY 40202, USA
 - ⁶ Biostatistics and Informatics Facility, Center for Integrative Environmental Health Sciences, University of Louisville, Louisville, KY 40202, USA
 - ⁷ Data Analysis and Sample Management Facility, The University of Louisville Super Fund Center, University of Louisville, Louisville, KY 40202, USA
 - ⁸ Hepatobiology and Toxicology Center, University of Louisville, Louisville, KY 40202, USA
 - ⁹ Christina Lee Brown Envirome Institute, University of Louisville, Louisville, KY 40202, USA
- * Correspondence: samarendra.das@icar.gov.in or samarendra.das@louisville.edu (S.D.); shesh.ra@louisville.edu (S.N.R.)

Abstract: With the advent of single-cell RNA-sequencing (scRNA-seq), it is possible to measure the expression dynamics of genes at the single-cell level. Through scRNA-seq, a huge amount of expression data for several thousand(s) of genes over million(s) of cells are generated in a single experiment. Differential expression analysis is the primary downstream analysis of such data to identify gene markers for cell type detection and also provide inputs to other secondary analyses. Many statistical approaches for differential expression analysis have been reported in the literature. Therefore, we critically discuss the underlying statistical principles of the approaches and distinctly divide them into six major classes, i.e., generalized linear, generalized additive, Hurdle, mixture models, two-class parametric, and non-parametric approaches. We also succinctly discuss the limitations that are specific to each class of approaches, and how they are addressed by other subsequent classes of approach. A number of challenges are identified in this study that must be addressed to develop the next class of innovative approaches. Furthermore, we also emphasize the methodological challenges involved in differential expression analysis of scRNA-seq data that researchers must address to draw maximum benefit from this recent single-cell technology. This study will serve as a guide to genome researchers and experimental biologists to objectively select options for their analysis.

Keywords: scRNA-seq; differential expression analysis; classification; statistical approaches; challenges



Citation: Das, S.; Rai, A.; Rai, S.N. Differential Expression Analysis of Single-Cell RNA-Seq Data: Current Statistical Approaches and Outstanding Challenges. *Entropy* **2022**, *24*, 995. <https://doi.org/10.3390/e24070995>

Academic Editors: Alessandro Giuliani and William B. Sherwin

Received: 20 May 2022

Accepted: 9 July 2022

Published: 18 July 2022

Publisher's Note: MDPI stays neutral with regard to jurisdictional claims in published maps and institutional affiliations.



Copyright: © 2022 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (<https://creativecommons.org/licenses/by/4.0/>).

1. Background

High-throughput single-cell RNA-sequencing (scRNA-seq) has emerged as a promising technology to explore the dynamics of gene expression at the single-cell level. It has become extremely popular for answering the key questions of developmental biology, including cellular heterogeneity study [1], the discovery of novel cell types [2], and cell trajectory analysis [3], etc. To date, many single-cell sequencing protocols have been developed, of which two are very popular: (i) unique molecular identifier (UMI) tag-based protocols such as Drop-seq [4] and 10x Genomics Chromium [5]; and, (ii) full length, non-UMI-based protocols, e.g., Smart-seq2 and Fluidigm C1 [6,7]. UMI-based protocols sequence only the 5-prime or 3-prime end of the mRNA molecule compared with non-UMI protocols [8]. The former has lesser amplification bias (i.e., transcript isoforms within the

same gene) compared with the latter [9]. Irrespective of the sequencing protocols, scRNA-seq data have some peculiar features including high-level noises, excess overdispersion, low library sizes, sparsity, and a higher proportion of zeros (i.e., due to the lower capture of transcriptomic material and other sources of variation), etc., [10]. Through these single-cell protocols, a huge amount of gene expression data (over thousand(s) to millions of cells) are generated in each experiment and deposited in public domain databases. Such an unprecedented event requires novel and advanced statistical approaches and bioinformatics tools to extract relevant biological knowledge.

Differential expression analysis (DEA) is the primary downstream analysis performed on scRNA-seq data [11–13]. The DEA is useful for the detection of biomarkers for novel cell types or gene signatures for cellular heterogeneity, and also provides inputs for other secondary analyses including gene set or pathway, and network analysis. The initial practice of DEA in scRNA-seq involved borrowing methods from bulk RNA-seq, which usually did not consider the special features of the scRNA-seq data [10,14]. Hence, specialized approaches have been reported in the literature for DEA of scRNA-seq data. Software(s)/R packages were developed based on these statistical approaches. Each approach has its own benefits and drawbacks, i.e., DEA approaches have distinct features and disparate performances. Several computational experiments have been conducted to establish the same, as reported in the literature [10,11,14–18]. Excellent review(s) of the computational comparative studies can be found in [10,16]. However, the major chunk of the assessed approaches was imported from the bulk RNA-seq. For instance, Sonesson and Robinson (2019) and Das et al. (2021) considered ~50% approaches from the bulk RNA-seq to assess the DEA approaches' performance on scRNA-seq datasets [10,16]. There are limited studies available in the literature which mainly focuses on critically reviewing DEA approaches exclusively designed for single-cell studies.

Therefore, in this review, we aim to present: (i) state-of-the-art methods and tools available for DEA of scRNA-seq data along with their classification based on input data, fitted statistical models, and test statistic(s); (ii) discuss the unique features and limitations of each class of approaches; and (iii) describe the key challenges yet to be addressed in the DEA of the scRNA-seq data. This study will serve as a catalog and provide guidelines to genome researchers and experimental biologists for objectively choosing proper DEA approaches, based on several factors.

2. Current Statistical Approaches

The term “Differential Expression” has been extensively used in gene expression studies including microarrays [19], RNA-seq [20], and scRNA-seq [16]. The basic difference of DEA for scRNA-seq, compared with other studies, is that it is used to detect bio-markers across the cell types, while in other studies it is used to find differential genes across the case vs. control conditions [21,22]. The operational framework of the DEA in scRNA-seq study is shown in Figure 1. The operational procedure is mostly common to single-cell studies (Figure 1). It is beyond the scope of this article to discuss the large number of existing analytical approaches covered by the term ‘Differential Expression Analysis’ in gene expression studies. Therefore, this review only focuses on statistical approaches that were exclusively developed for single-cell expression studies, rather than on methods that perform DEA on any sequencing data.

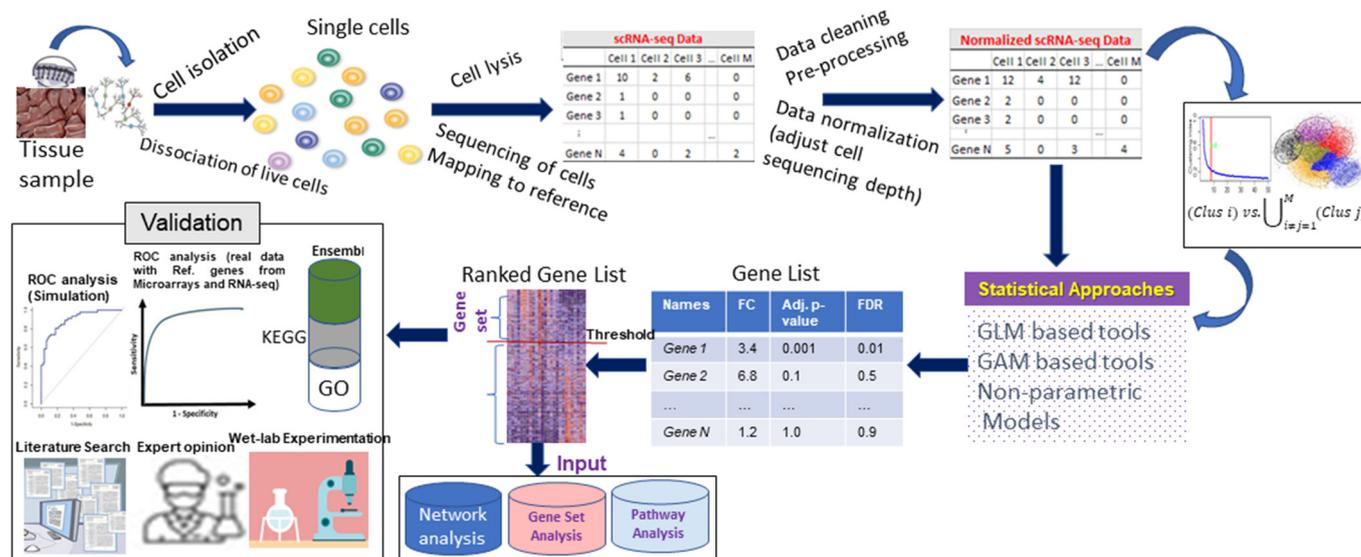


Figure 1. Operational framework of differential expression analysis of scRNA-seq data. Various steps in single-cell studies are shown. Pre-processing and various steps of DE analysis are also shown. Potential use and interpretation of obtained results are presented.

The existing approaches and tools for the DEA of scRNA-seq data and their availability are listed in Table 1 and Table S3. Table 1 also presents a comparative overview of the approaches in terms of different factors including statistical models, input, test statistic(s), and runtime, etc. We also classified these existing approaches and tools based on several factors including data requirement and background models, etc., as shown in Figure 2. For instance, the available approaches can be grouped based on the input data requirement, i.e., (i) group of approaches, which requires expression data as well as external spike-in data; (ii) other group of approaches, which only requires expression data (Figure 2). Further, the former category of tools is computationally expensive and more accurate compared with the latter [23], due to the implementation of efficient statistical models and spike-ins data requirement. The use of RNA spike-ins in the approaches tackles the issue of technical variability due to inefficient transcriptomes capture in single-cells [24–28]. In other words, the RNA spike-ins are RNA transcripts (with known sequences and quantity) that are applied to calibrate the measurements of RNA hybridization assays, such that scRNA-Seq and UMIs can theoretically enable the estimation of absolute molecular counts [29]. It is worthy to note that, if RNA spike-ins data are available, it is profitable to use them in DEA, using a suitable approach. Moreover, the classification of the approaches and tools based on other factors can be found in Figure 2.

Instead of individually reviewing the large number of DEA approaches (Table 1), our goal here is to classify the approaches based on the common statistical principles/models and discuss their relative merits. However, for the researcher desiring specific information about the individual tools, Supplementary Document S1 briefly presents reviews of the individual approaches. We also present class-wise critical reviews of the existing DEA approaches in the following section.

Table 1. Comparative overview of the DEA approaches for scRNA-seq data analysis.

SN.	Methods	Year	Model	Input	DE Test Stat.	Runtime	Platform	Ref.
1	NBID	2018	NB (GLM)	Counts	LRT	Medium	R code	[30]
2	ZINB-WaVE	2018	ZINB (GLM)	Counts	LRT	High	Bioconductor, GitHub	[31]
3	zingeR	2018	ZINB (GLM)	Counts	LRT	High	GitHub	[32,33]
4	DECENT	2019	ZINB (GLM)	Counts	LRT	High	GitHub	[24]
5	SwarnSeq	2021	ZINB (GLM)	Counts	LRT	High	GitHub	[13]
6	Tweedieverse	2021	ZITweedie (GLM)	Counts	Wald	High	GitHub	[34]
7	scMMST	2021	GLMM	Counts	Norm. score	High	NA	[35]
8	TPMM	2022	GLMM	Norm.	Wald/LRT	High	GitHub	[36]
9	Monocle2	2017	GAM	Norm.	LRT	Medium	Bioconductor	[37,38]
10	tradeSeq	2020	GAM	Counts	Wald	Medium	GitHub	[39]
11	MAST	2015	Hurdle	Norm.	LRT/Wald	Medium	Bioconductor	[40]
12	Random-Hurdle	2019	Hurdle	Counts	Chi-square test statistic	High	NA	[41]
13	SCDE	2014	Poisson-NB (MM)	Counts	Bayesian stat.	High	Bioconductor	[42]
14	BASiCS	2015	Poisson-Gamma (MM)	Norm.	Posterior prob.	High	Bioconductor	[25]
15	D3E	2016	Poisson-Beta (MM)	Counts	CM/KS test	High	GitHub	[43]
16	BPSC	2016	Beta-Poisson (MM)	Counts	LRT	Medium	GitHub	[12]
17	TASC	2017	Logistic, Poisson Models (MM)	UMI	LRT	High	GitHub	[26]
18	DESCEND	2018	Poisson-Alpha (MM)	Counts	Normalized Gini Score	High	GitHub	[28]
19	SC2P	2018	ZIP, Poisson-Lognormal (MM)	Counts	Posterior prob.	High	GitHub	[44]
20	ZIAQ	2020	Logistic and quantile Regression (MM)	Norm.	Fisher's test	Medium	GitHub	[45]
21	SimCD	2021	Gamma-NB (MM)	Counts	Bayesian	High	GitHub	[46]
22	ZIQRank	2022	Zero-inflated model, quantile regression (MM)	Cont.	Rank-score test	High	NA	[47]
23	Seurat	2015	NB (TCP)	Counts	LRT	Low	CRAN	[48,49]
24	scDD	2016	Multi-modal Bayesian (TCP)	Norm.	Bayesian stat.	High	Bioconductor	[50]
25	DEsingle	2018	ZINB (TCP)	Counts	LRT	High	Bioconductor, GitHub	[51]
26	NYMP	2019	Logistic regression (TCP)	Cont.		Medium	GitHub	[52]
27	<i>t</i> -test		logCPM (TCP)	Norm.	T stat	Low	CRAN	[10]

Table 1. Cont.

SN.	Methods	Year	Model	Input	DE Test Stat.	Runtime	Platform	Ref.
28	IDEAS	2022	NB/ZINB/Kernel Density estimation/ Cumulative distribution function (TCP)	Counts/Cont.	Jensen–Shannon Divergence/ Wasserstein distance	High	GitHub	[53]
29	SAMstrt	2013	NP	Counts		Medium	GitHub	[54]
30	Wilcox		NP	Counts/Norm.	Sum ranks	Low	CRAN	[10]
31	SINCERA	2015	NP	Norm.	Welch (LS)/ Wilcox (SS)	High	GitHub	[55]
32	NODES	2016	NP	Norm.	Wilcox	Medium	Dropbox	[56]
33	EMDomics	2016	NP	Norm.	Euclidean distance	High	Bioconductor	[57]
34	sigEMD	2018	NP	Norm.	Distance measure	High	GitHub	[58]
35	DTWscore	2017	NP	FPKM	Distance	Medium	GitHub	[59]
36	ROSeq	2021	NP	Counts/Norm.	Wald	High	Bioconductor, GitHub	[60]
37	scDEA ¹	2021	12 Models (Hybrid)	Counts	Lancaster’s test (Chi)	High	GitHub	[61]

CM: Cramér–von Mises test; Counts: read/UMI counts; Cont.: continuous values, e.g., FPKM, log(CPM), RPKM; NA: source codes are not freely available; Norm.: normalized; GLM: generalized linear model; NB: negative binomial; GLMM: generalized linear mixed model; NP: non-parametric; GAM: generalized additive model; MM: mixture model; TCP: two-class parametric. ¹: Integrated approach.

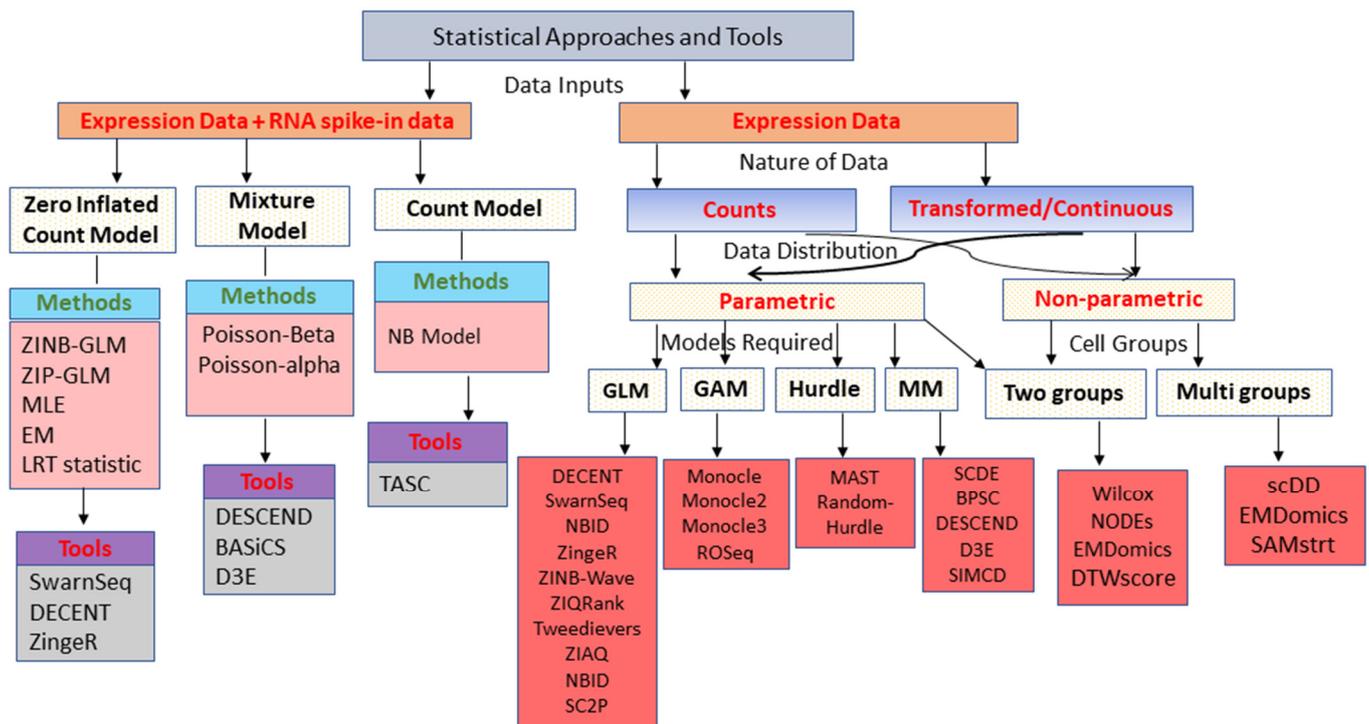


Figure 2. Classification of available statistical approaches and tools used for DEA in single-cell studies. Classification of the approaches is conducted based on the requirement of input data, data distribution, and statistical models, etc. DE analytic tools belonging to each category are presented in pink colored boxes.

3. Classes of Statistical Approaches for DEA

Notation: Y_{ij} : random variable (rv) represents the observed expression (i.e., read, UMI) counts of i th ($i = 1, 2, \dots, N$) gene in j th ($j = 1, 2, \dots, M$) cell; N : total number of genes; M : total number of cells; μ_i : mean of i th gene for NB distribution (count part of the model); φ_i and $\theta_i (= \varphi_i^{-1})$: dispersion and size parameters, respectively, for i th gene; $\pi_i (\in [0, 1])$: mixture probability (zero inflation probability) of i th gene; s_j : library size of j th cell; Z_{ij} : rv represents the true (unknown) expression counts for i th gene in j th cell; \mathbf{X} : design matrix for cell group information, whose j th row: $X_j = [X_{j1}, X_{j2}, \dots, X_{jN}]$; W_{ij} : indicator rv representing the rate of expression for i th gene in j th cell, i.e., $W_{ij} = 0 : Y_{ij} = 0$; $W_{ij} = 1 : Y_{ij} > 0$. $\Omega_i = \{\mu_i, \theta_i, \pi_i\}$: parametric space for i th gene.

3.1. Generalized Linear Model-Based Approaches

Generalized linear model (GLM)-based approaches assume that: (i) expression counts follow certain exponential family distribution [62]; and, (ii) a non-linear function (known as *link function*) relates the expected expression counts of genes to the linear component of the model. In other words, every GLM has three components: (i) expression count distribution of the gene (sometimes called the error structure); (ii) linear predictor that involves the explanatory cell variables or covariates including cell group information; and, (iii) a link function ($g(\cdot)$) that connects the linear predictor to the natural mean of expression counts of genes. The GLM class of DEA approaches includes NBID [30], DECENT [24], ZINB-WaVE [31], ZingeR [32,33], Tweedieverse [34], and SwarnSeq [13], to name a few. The operational layout of this class of approaches is shown in Figure 3.

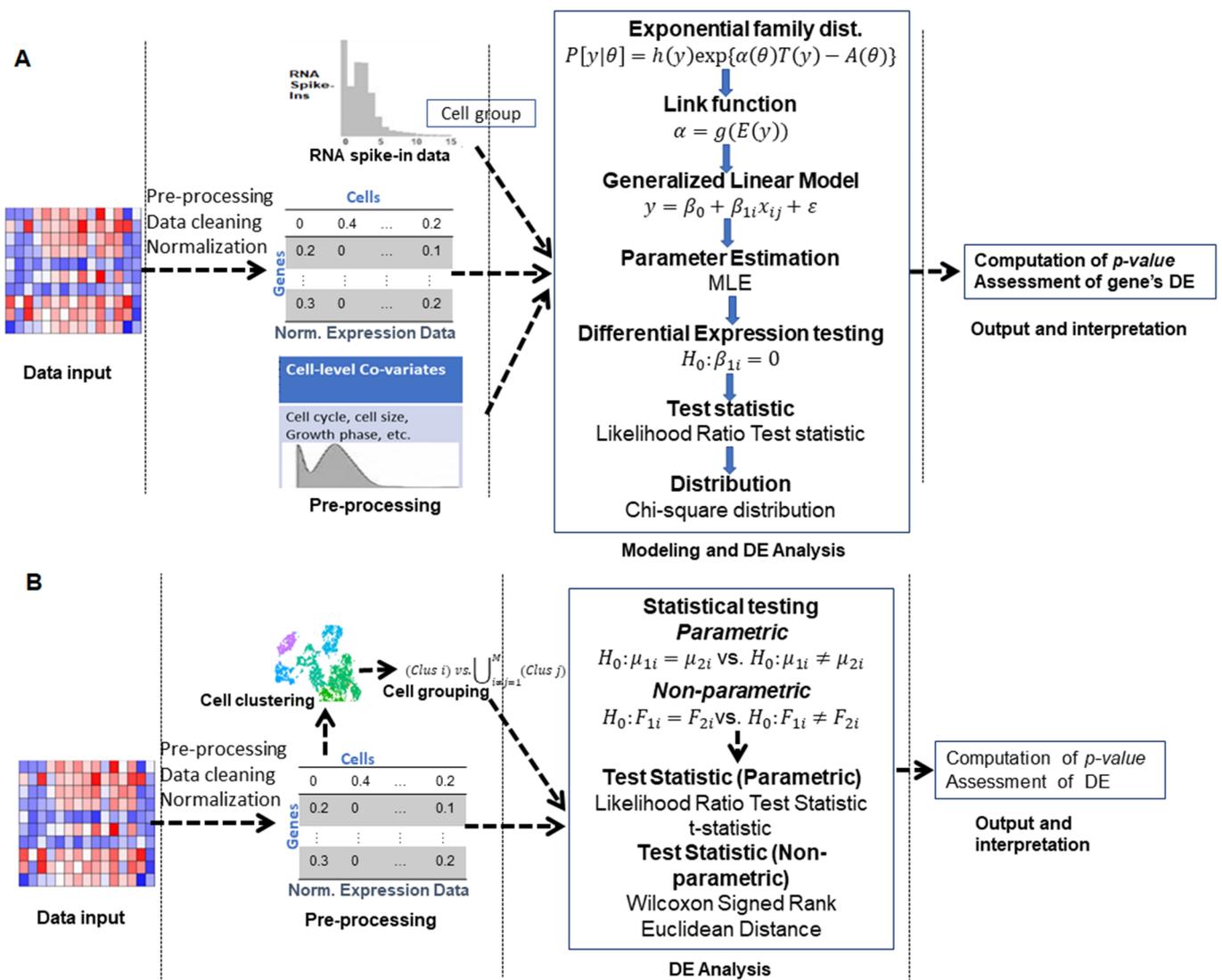


Figure 3. Operational outlines of DE analytic GLM and two-class comparison approaches in scRNA-seq studies. (A) Workflow of steps for GLM-based DE approaches. (B) Workflow of steps for two-class comparison approaches. In both classes, the framework can be divided into four major parts, namely: (i) input (data provided as input to tools); (ii) pre-processing of data, this step involves data cleaning, outlier removal, normalization, etc.; (iii) model fitting and computation of DE test statistic, various distributional/model (e.g., GLM, simple statistical distribution or distribution-free) assumptions are made about the expression data, parameters of the models are estimated, and DE test statistic(s) for genes and their corresponding p-values are computed; and, (iv) assessment and interpretation of DE results.

The GLM-based DEA approaches can be divided into two categories: first, expression counts follow certain exponential family distributions, e.g., negative binomial (NB) and Poisson, etc. Second, expression counts follow zero-inflated models, e.g., zero inflated negative binomial (ZINB), and zero inflated Poisson (ZIP) (Supplementary Document S2). For instance, NBID [30] and IDEAS [53] approaches use the NB model, and probability mass function (PMF) in Equation (1), to fit the single-cell expression counts data. The expected value and variance of the observed expression counts is given in Equation (2).

$$P[Y_{ij} = y] = \frac{G(y + \theta_{ij})}{G(y + 1)G(\theta_{ij})} \left(\frac{\theta_{ij}}{\theta_{ij} + \mu_{ij}} \right)^{\theta_{ij}} \left(\frac{\mu_{ij}}{\theta_{ij} + \mu_{ij}} \right)^y \quad \forall y = 0, 1, 2, \dots \quad (1)$$

where, $\mu_{ij} \geq 0$; $\theta_{ij} > 0$ are the mean and size parameters of NB distribution, $G(\cdot)$: Gamma function. Then, the expected value and variance of Y_{ij} is shown as:

$$E(Y_{ij}) = \mu_{ij} \text{ and } V(Y_{ij}) = \mu_{ij} + \frac{\mu_{ij}^2}{\theta_{ij}} = \mu_{ij} + \varphi_{ij} \quad (2)$$

The NBID uses a non-linear link function to model the expected value of expression counts with the explanatory variables, such as cell group labels and other potential covariates, given in Equation (3).

$$g(\mu_{ij}) = \log \mu_{ij} = \mathbf{X}\boldsymbol{\beta} \quad (3)$$

where, $g(\cdot)$: link function, and $\boldsymbol{\beta}$: parameters of the model.

The NB-GLM approaches may not suitable to fit the scRNA-seq counts data due to the presence of excess zeros (Supplementary Documents S3-S5) [10,13,32], thus may compromise the statistical power to detect true differentially expressed genes [13,31]. Hence, ZIM was introduced in DEA bioinformatics tools to fit the observed scRNA-seq count data [1-33]. The ZIM-GLM-based approaches include tools such as DECENT [24], ZINB-WaVE [31], ZingeR [32,33], and SwarnSeq [13], which assume the UMI counts of genes follow a ZINB distribution. PMF is given in Equation (4).

$$P[Y_{ij} = y] = \begin{cases} \pi_{ij} + (1 - \pi_{ij}) \left(\frac{\theta_{ij}}{\theta_{ij} + \mu_{ij}} \right)^{\theta_{ij}} & \text{when } y = 0 \\ (1 - \pi_{ij}) \frac{G(y + \theta_{ij})}{G(y + 1)G(\theta_{ij})} \left(\frac{\theta_{ij}}{\theta_{ij} + \mu_{ij}} \right)^{\theta_{ij}} \left(\frac{\mu_{ij}}{\theta_{ij} + \mu_{ij}} \right)^y & ; y > 0 \end{cases} \quad (4)$$

The expected value and variance of Y_{ij} is expressed in Equations (5) and (6).

$$E(Y_{ij}) = (1 - \pi_{ij})\mu_{ij} \quad (5)$$

$$V(Y_{ij}) = (1 - \pi_{ij})\mu_{ij} \left(1 + \pi_{ij}\mu_{ij} + \frac{\mu_{ij}}{\theta_{ij}} \right) \quad (6)$$

If $\pi_{ij} = 0 \Rightarrow \text{ZINB}(\pi_{ij}, \mu_{ij}, \theta_{ij}) \rightarrow \text{NB}(\mu_{ij}, \theta_{ij})$

If $\varphi_{ij} \rightarrow 0 \Rightarrow \text{ZINB}(\pi_{ij}, \mu_{ij}, \theta_{ij}) \rightarrow \text{ZIP}(\pi_{ij}, \mu_{ij})$

The linear predictors including cellular group [24], cell type [13], and other cell-level auxiliaries [31] are included in the GLM. Various link functions including log and logit functions are used to model the gene specific mean and zero-inflation parameters, respectively. For instance, DECENT, SwarnSeq, ZINB-Wave, and ZingeR consider log-link function to connect the mean with cell group, cell-level auxiliary predictors and *logit*-link function, to model the zero-inflation parameter, as given in Equations (7) and (8).

$$\log \mu_i = \mathbf{X}\boldsymbol{\gamma}_i + \mathbf{R}\boldsymbol{w}_i + \mathbf{C}\boldsymbol{s}_i + \mathbf{O}_\mu \quad (7)$$

$$\text{logit} \pi_i = \mathbf{X}\boldsymbol{\beta}_i + \mathbf{R}\boldsymbol{u}_i + \mathbf{C}\boldsymbol{v}_i + \mathbf{O}_\pi \quad (8)$$

where, $\text{logit}(\pi) = \log\left(\frac{\pi}{1-\pi}\right)$; $\boldsymbol{\alpha}_i$, $\boldsymbol{\tau}_i$ and $\boldsymbol{\varphi}_i$: $M \times 1$ vector of parameters for i th gene; \mathbf{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); \mathbf{R} : $M \times N$ design matrix providing cell cluster information; \mathbf{C} : $M \times C$ design matrix providing cell level auxiliary information; $\boldsymbol{\gamma}_i$ and $\boldsymbol{\beta}_i$: $G \times 1$ vectors of cellular groups effects for i th gene; \boldsymbol{w}_i and \boldsymbol{u}_i : $N \times 1$ vectors of cell cluster effects for i th gene; \boldsymbol{s}_i and \boldsymbol{v}_i : $C \times 1$ vectors of effects for cell level co-variates, such as cell cycle, cell phase, etc., for the i th gene; \mathbf{C} : levels of cell level auxiliaries; and, \mathbf{O}_μ , \mathbf{O}_π : offsets for μ_i and π_k , respectively.

The DECENT approach only considers the cell group information as predictor in the model (Equation (9)), while the SwarnSeq, ZingeR and ZINB-WaVE approaches consider group as well as other cell level data as linear predictors in the model (Equations (7) and (8)). All GLM-based approaches use the maximum likelihood estimation (MLE) method

to estimate the model parameters through minimizing the goodness of fit criterion [63]. However, it is very difficult to obtain the exact solution (i.e., closed form) of the MLE's objective function [13,33]. Therefore, iterative expected maximization (EM) techniques are implemented in these approaches to estimate the parameters [13,32]. For instance, DECENT, ZingeR, ZINB-WaVE, and SwarnSeq use EM algorithms to estimate the gene specific parameters including mean, dispersion, and zero-inflation parameters [13,24,32,33]. For DEA of genes (two cell groups' simple comparison), the models in Equations (2) and (7) can be written as:

$$\log\mu_{ij} = \beta_{i0} + \beta_{i1}x_{ij} + O_{\mu} \quad (9)$$

where, β_{i0} : intercept term; β_{i1} : regression co-efficient for cell group; and, O_{μ} : offset term. The model in Equation (9) can also be expanded to accommodate other cell-level co-variates including cell type, cell cycle, cell growth phase, etc., [13]. To test whether the i th gene is differentially expressed or not across the cell groups, GLM-based approaches test the following null hypothesis:

$$H_0 : \beta_{i1} = 0 \text{ vs. } H_1 : \beta_{i1} \neq 0$$

Importantly, all GLM-based approaches use the likelihood ratio test (LRT) statistic. Mathematically, the LRT statistic (for i th gene) is $-2\log L$ (L : likelihood function) (i.e., deviance divided by the scale parameter ϕ called scaled deviance) which can be expressed as:

$$-2\log L(\Omega_i|y_{ij}) = \frac{1}{\phi} D(y_{ij}|\Omega_i) = 2l(y_{ij}|y_{ij}) - 2l(\Omega_i|y_{ij}) \sim \chi^2_{(M-p)} \quad (10)$$

where, $l(\Omega_i|y_{ij})$: log-likelihood function; and $l(y_{ij}|y_{ij})$: log-likelihood function for the saturated model (i.e., maximum likelihood achievable when the fitted values are exactly equal to the observed data for exponential family distribution). The test statistic in Equation (10) follows a Chi-square distribution with certain degrees of freedom.

Additionally, modifications in the GLM have been performed by adding random components for different cell-level factors to build generalized linear mixed models. Recently, such models have been implemented in two-part mixed model (TPMM) and scMMST approaches.

Limitations: There are three major limitations of this class of approaches. Firstly, *strict model assumptions*: the GLM class of approaches requires several distributional assumptions about the expression counts, which may not be satisfied by the real single-cell data. For instance, GLM requires the counts to be generated by exponential family distributions; the link function must be invertible, continuous, and differentiable; and it linearly depends on cell co-variates. These strict assumptions restrict the utility of GLM-based DEA approaches for real data analysis. In most cases, the users simply apply these techniques without testing or violating these assumptions, which causes the results to be misleading.

For the second limitation, *multi-modality*, several previous studies [37,38,45,47,64–67] report multi-modal distributions of scRNA-seq gene expressions, which may be due to a gene's expression deriving from multiple cell states or from a series of biological processes [65]. For instance, a cancer-suppressor gene is over-expressed (i.e., higher counts) in some cells and its expression is suppressed (by its regulator genes) causing low expression in other cells. This negative feedback causes oscillations in gene expression across the cells [64], leading to multiple modes in scRNA-seq data [67]. The NB-GLM approaches fail to handle the multi-modal distribution of scRNA-seq data, while the ZIM-GLM-based approaches are able to tackle the bio-modality (i.e., modes due to biological zeros and non-zero counts) of underlying data. In other words, GLM-based tools cannot handle the multi-modality (>2) of expression counts distribution, which is inherent to scRNA-seq studies.

For the third limitation, *computational complexity*, this class of tools is computationally intensive due to implementation of iterative techniques for parameters estimation. For instance, DECENT, SwarnSeq and ZingeR, etc., approaches usually take more than 10–12 h

to analyze the scRNA-seq data (with a few thousand genes and hundreds of cells) [10]. In addition, the EM algorithm employed in these tools fails to converge in most of the genes, causing the computational process to be slow and cumbersome. Furthermore, complex statistical models are fitted for each gene individually in a large dataset, making the implemented tools' runtime inefficient. Further features and limitations of this class of tools are listed in Table 2.

3.2. Generalized Additive Model-Based Approaches

Generalized additive models (GAM) are natural extensions of the GLM, where the link function is additive but each term/predictor non-linearly depends on the mean and zero-inflation parameters of the gene. GAMs are similar to GLMs but allow testing of variables in response to a numerically estimated trend in the predictors, alleviating the burden of specifying their distribution. While this necessitates some approximations in downstream testing, it has proven to be highly effective in many settings, particularly when one wishes to model the response variable as a function of both categorical (e.g., cellular group) and continuous (e.g., cell growth phase, cycle, etc.) predictors. The operational framework of this class of approaches is shown in Figure 4.

The GAM uses smooth (non-parametric (NP)) spline functions to capture the relationships between individual cell co-variables and the expression of gene, which can be linear or nonlinear. In other words, these smooth relationships can be simultaneously estimated, and then, the expected expression values predicted, by simply adding them. The GAM class of DEA approaches includes Monocle [37], Monocle2 [38], and tradeSeq [39], etc.

The impact of the predictive variables is captured through smooth functions, depending on the underlying patterns in the data, which can be nonlinear:

$$g(\mu_{ij}) = \beta_0 + s_1(x_1) + s_2(x_2) + \dots + s_p(x_p) \quad (11)$$

where, x_l : cell level co-variables/predictors; and $S_{il}(\cdot)$: smooth function.

The GAM-based approaches use a *log* link function that depends on the pseudo-time of the cell, as shown in Equation (12).

$$\log(\mu_{ij}) = \sum_{l=1}^L s_{il}(t_{lj})Z_{lij} + O_{\mu_{ij}} \quad (12)$$

where, $s_{il}(\cdot)$: smooth function for i th gene at l th cell lineage, which are functions of pseudo-time t_{lj} , $\forall l \in \{1, 2, \dots, L\}$; and, Z_{lij} : binary variable for gene expression, i.e., $Z_{lij} = 1$; $y_{lij} > \tau$ or 0 else (τ : hard threshold). The smoothing spline function in Equation (12) can be represented as a linear combination of K cubic basis functions ($b_k(\cdot)$).

$$s_{il}(t_{lj}) = \sum_{k=1}^K b_k(t) \beta_{ilk} \quad (13)$$

For testing i th gene at l th cell lineage, the null hypothesis, $H_0 : \beta_{ilk} = 0$ is tested for its possible rejection through the Wald test statistic (e.g., tradeSeq) or LRT (e.g., Monocle). Further, Monocle (also Monocle 2, 3) only considers one cell lineage (i.e., $L = 1$) while tradeSeq considers multiple cell lineages (i.e., $L \geq 2$). Additionally, the latter provides provision for the DEA of genes within and across cell lineages. The special features of the GAM-based approaches are listed in Table 2.

Table 2. Classes of statistical approaches and tools extensively used in DEA of scRNA-seq data.

SN.	Class	Features	Limitations	Tools
1	GLM	<ul style="list-style-type: none"> Gene expression can have any form of exponential distribution type. Suitable for bi-modality of data. Able to deal with categorical predictors, e.g., cell type, cell cycle, etc. Easy to interpret and allows a clear understanding of how each of the predictors are influencing the gene parameters. Can be generalized to multi-cell group comparisons. Less susceptible to model over-fitting. 	<ul style="list-style-type: none"> Strict exponential family distributional assumptions about the data. Needs relatively large datasets (with more predictor and large number of cells). Sensitive to outliers. Sensitive to dropout events. Not suitable for low expressed genes. Cannot handle multi-modality of the data. ZIM-GLM approaches are not able to handle zero-deflation at any level of a factor and will result in parameter estimates of infinity for the logistic component. Higher computational cost especially for large datasets. 	NBID, ZingeR ZINB-WaVE, DECENT, SwarnSeq, scMMST, TPM,MM, Tweedieverse
2	GAM	<ul style="list-style-type: none"> Predictor functions are automatically derived during model estimation. Marginal impact of a single variable does not depend on the values of the other variables in the model. Flexibility in choosing the type of functions, which will help in finding patterns missed in a parametric model. Allows controlling smoothness of the predictor functions to prevent model over-fitting. By controlling the wiggleness of the predictor functions, we can directly tackle the bias/variance tradeoff. Highly effective in many settings, particularly when one wishes to model the response variable as a function of both categorical (e.g., cell groups) and continuous predictors (e.g., cell-level auxiliary variables). Considers both linear and non-linear functions of cell-level predictors to model gene parameters. Each lineage is represented by a separate cubic smoothing spline, and its flexibility allows adjustment for other covariates or confounders as fixed effects in the model. 	<ul style="list-style-type: none"> Approaches such as Monocle can only handle a single lineage of cells. Lack of interpretability, to infer differences in expression between lineages of cells. Assumes the dropout events to be linear; however, the effect of dropout events is likely to be non-linear, especially for genes with low to moderate expression. Computationally complex. 	Monocle, Monocle2, Monocle3, tradeSeq

Table 2. Cont.

SN.	Class	Features	Limitations	Tools
3	Hurdle Model	<ul style="list-style-type: none"> • Considers the excess zeros while model building. • Can handle zero-inflation as well as zero-deflation present in data. • Models the bimodality of gene expression distribution. 	<ul style="list-style-type: none"> • Does not differentiate the generating process for excessive zeros versus sampling zeros. • Fails to consider the multi-modality of gene expression distribution. • Requires higher runtime. 	MAST, Random Hurdle
4	Mixture-Model	<ul style="list-style-type: none"> • Considers bi-modal or multi-modal nature of single-cell data. • Can differentiate between major sources of variation in single-cell data. 	<ul style="list-style-type: none"> • Certain approaches including BPSC, SC2P cannot consider the zero-inflation in single-cell data. • Mostly uses linear models for DEA, which is cumbersome. • Higher runtime and computationally intensive. 	SCDE, D3E, BPSC, BASiCS, DESCEND, SC2P, ZIAQ, ZIQRank, SimCD
5	Non-parametric (two-class)	<ul style="list-style-type: none"> • Distribution-free approaches. • Considers the multi-modality of the data. • Computationally not cumbersome (less runtime). • Estimates the parameters without fitting any distribution for genes. • Performs DEA with distance-like metrics across two cell types. • Performs well when there are lesser proportions of zeros in the data. 	<ul style="list-style-type: none"> • Mostly focuses on two cellular groups' comparison. • Computationally complex for multi-groups. • Performance severely affected due to high dropouts (some methods exclude dropouts). • Cannot separate between true/biological and false/dropout zeros. • Sensitive to sparsity. • Methods such as D3E, scDD fail to consider UMI count nature of the data. • Cannot separate confounding factors from each other. 	Wilcox, NODES, ROTS, EMDomics, ROSeq, SINCERA, sigEMD, DTWscore, SAMstrt
6	Parametric (two-class)	<ul style="list-style-type: none"> • Easy to understand and execute. • Lesser runtime. • Particularly suitable for larger datasets. 	<ul style="list-style-type: none"> • Makes strict distributional assumption about the data. • Cannot generalize to multi-group comparisons. • Ignores the multi-modal distributions of the scRNA-seq data. • Sensitive to sparsity or dropout events. • Cannot differentiate between the major sources of variability in the data. 	scDD, DEsingle, <i>t</i> -test, NYMP, IDEAS

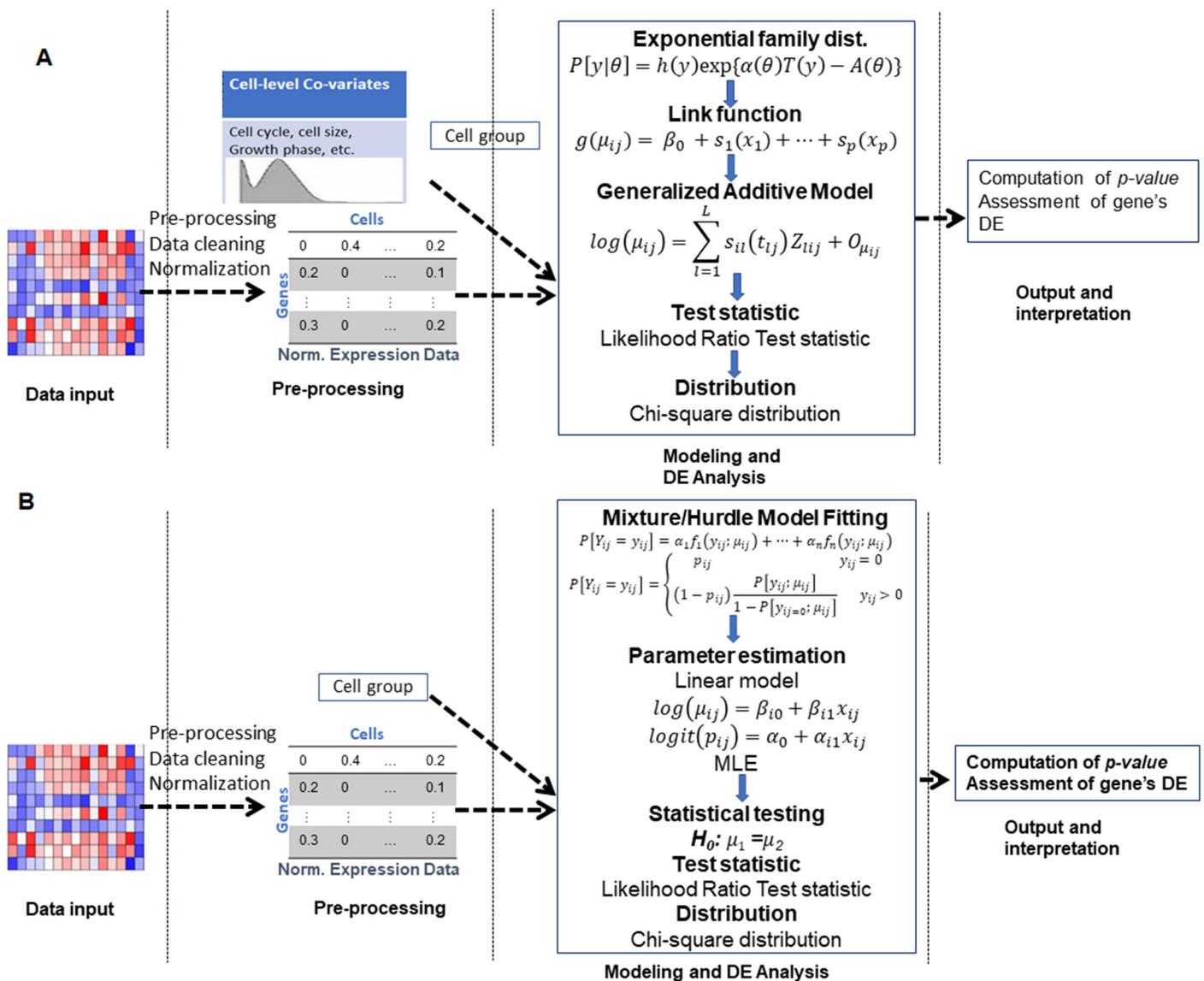


Figure 4. Operational outlines of DE analytic GAM, Hurdle and mixed model class of approaches in scRNA-seq studies. **(A)** Workflow of steps for GAM-based DEA approaches. **(B)** Workflow of steps for Hurdle and mixed-model-based approaches. In both classes, the framework can be divided into four major parts, namely: (i) input (data provided as input to tools); (ii) pre-processing of data, this step involves data cleaning, outlier removal, normalization, etc.; (iii) model fitting and computation of DEA test statistic, various distributional/model (e.g., GAM, Hurdle or mixture model) assumptions are made about the expression data, parameters of the models are estimated, DEA test statistic(s) for genes and their corresponding *p*-values are computed; and (iv) assessment and interpretation of DEA results.

Limitations: (i) *Pseudo-time dependent:* Approaches including Monocle heavily depend on the accuracy of the pseudo-time-ordering of cells. In other words, in single-cell studies, expression of genes in each cell is a function of time, therefore, cells can be ordered by the time. Single-cell analytical tools use existing algorithms including Wanderlust [68] to order the single-cells along discrete paths. These paths do not represent real time but rather a pseudo-time variable (due to short life cycles of cells), which usually represents the intrinsic cellular process. Further, computational experiments indicate that differences in the temporal ordering of the single-cells from different approaches affect the results, and thus interpretations [69]. The use of pseudo-temporal ordering along with expression

data has been useful in some studies, but it has also faced criticism. For instance, Moris et al. 2016 [70] questioned the underlying assumptions of smooth and continuous cell state transitions, which are required by pseudo-time-ordering algorithms. Moreover, such data may not be readily available for the users, thus making it difficult to apply in general cases; (ii) Similar to the GLM classes of approaches, this class is also unable to consider the multi-modal nature of single-cell data; and, (iii) The GAM class of approaches is computationally intensive, due to implementation of complex statistical models fitted individually for each gene.

3.3. Mixture Model-Based Approaches

The observed gene expressions in a scRNA-seq experiment are the noisy reflections of true expression levels due to various biological and technical sources. Hence, the mixture model (MM) framework (shown in Figure 4) assumes that the distributions of the observed expression counts are decomposed into multiple parts or a mixture of probability distributions, as shown in Equation (14):

$$P[Y_{ij} = y_{ij}] = \alpha_1 f_1(y_{ij}; \Omega_{ij}) + \alpha_2 f_2(y_{ij}; \Omega_{ij}) + \dots + \alpha_n f_n(y_{ij}; \Omega_{ij}) \quad (14)$$

where, $f_1(\cdot)$, $f_2(\cdot)$, \dots , $f_n(\cdot)$ are the probability distributions associated with various components of single-cell studies, e.g., dropout events, amplification, etc.; and $\alpha_1, \alpha_2, \dots, \alpha_n$ are the corresponding weights of the distribution functions such that $\alpha_i > 0$ and $\sum_{i=1}^n \alpha_i = 1$. In other words, the PMF of the expression counts is expressed in terms of the linear combination of the distribution functions of various components of single-cell studies (a convex combination).

The MM class includes popular approaches and tools such as SCDE [42], D3E [43], BPSC [12], BASiCS [25], DESCEND [28], and SC2P [44], etc. For instance, SCDE models the expression (in terms of reads per million) of i th gene in j th cell using a mixture of Poisson and NB distributions, and the PMF for the SCDE can be written as:

$$P[Y_{ij} = y_{ij}] = \alpha f_{NB}(y_{ij}; \mu_{ij}, \theta_{ij}) + (1 - \alpha) f_{PD}(y_{ij}; \gamma_{ij}) \quad (15)$$

where, f_{NB} and f_{PD} : PMF of NB and Poisson distributions, respectively; γ_{ij} : parameter of Poisson distribution; and $\alpha = \log(e)$ (e : expected expression magnitude).

Further, approaches such as BPSC and D3E use the Beta-Poisson model to fit the expression counts of genes in scRNA-seq data to capture the cell burst size and burst frequency of the gene level expression data. However, the BPSC uses a linear model framework (to model μ_{ij} through log-link function) to perform DEA, while D3E employs three statistical tests, i.e., Cramer-von Mises test, KS test or the LRT for differential expression testing of genes. It is interesting to note that BPSC can be generalized to multi-cellular group comparison, while D3E is limited to only two group comparison. Moreover, the DESCEND, BASiCS, SC2P tools use the Poisson-Alpha, Poisson-Gamma, and Poisson-Lognormal MMs, respectively, to fit the observed scRNA-seq data.

Additionally, approaches including ZIAQ [45] and ZIQRank [47] use logistic regression and quantile regression to model the dropout events/zero-inflation and non-zero expression counts (shifted quantiles), respectively [47]. It is worthy to note that this class of approaches uses different setups to perform DEA of genes across the cell groups. For instance, ZIQRank and ZIAQ use the Cauchy and Fisher test, respectively, to compute the p -values for genes [45,47]. The major pros and cons of this class of approaches are listed in Table 2.

3.4. Hurdle Model-Based Approaches

In Hurdle model-based approaches, the expression counts of genes are modeled in two parts, namely, (i) zero counts ($[W_{ij} = 0] \sim f_1(\cdot; \theta_i)$) and (ii) non-zero counts ($[Y_{ij} | W_{ij} = 1 \sim f_2(\cdot; \theta_i)$). In other words, the first part of the model fits the zero counts, while the second part models the probability of the non-zero expression values (through a truncated (at zero)

probability distribution function). Mathematically, the PMF of the Hurdle model is shown in Equation (16):

$$P[Y_{ij} = y_{ij}] = \begin{cases} p_{ij} & y_{ij} = 0 \\ (1 - p_{ij}) \frac{P[y_{ij}; \mu_{ij}]}{1 - P[y_{ij}=0; \mu_{ij}]} & y_{ij} > 0 \end{cases} \quad (16)$$

where, p_{ij} : probability of expression counts of i th gene in j th cell belongs to the zero component.

Statistically, the ZIM and Hurdle models differ based on their conceptualization of the zeros in scRNA-seq data and interpretation of model parameters. In other words, the ZIM model always assumes that zero counts are derived from a mixture of two distributions: (i) the first part produces zero counts which are often called “structural zeros” or “excessive zeros” (e.g., absence of mRNA of a gene) (modeled using Dirac’s delta function); and, (ii) the second part produces zero counts termed as “sampling zeros” (e.g., absence of expression counts due to inefficient amplification/limited sequencing depth) (modeled using the count data model).

The Hurdle model-based approaches model the gene parameters using linear models, as shown in Equations (17) and (18).

$$\log(\mu_{ij}) = \beta_{i0} + \beta_{i1}x_{ij} \quad (17)$$

$$\text{logit}(p_{ij}) = \alpha_0 + \alpha_{i1}x_{ij} \quad (18)$$

For the DEA of i th gene, the null hypothesis is tested for its possible rejection using the test statistic given in Equation (10). This class includes popular approaches such as MAST [40] and Random Hurdle [41]. Further, the MAST approach uses a logistic regression model to fit the indicator variable (for zeros), W_{ij} and the Gaussian linear model for the continuous variable (non-zero expression values) ($Y_{ij}|W_{ij} = 1$), independently. The unique features of this class of approaches are listed in Table 2.

The major limitations of the Hurdle model-based approaches are: (i) The Hurdle model considers the sources of zeros in single-cell studies without making any distinctions. In other words, the Hurdle model can have restrictive assumptions that fail to distinguish between structural zeros and sampling zeros [8], which can be quite detrimental when the assumptions are violated. By design, the Hurdle model will always predict the same number of zeros as observed in the scRNA-seq data without telling their sources; (ii) Several of these approaches require a data transformation including the use of a pseudo-count and log-transformation, but this has recently been shown to introduce false variation in downstream analyses [71,72]; and, (iii) These approaches are computationally intensive especially for large single-cell datasets, as they fit models individually for each gene, such as GLM-based approaches. Further limitations are listed in Table 2.

3.5. Two-Class Comparison (Parametric) Approaches

The above four classes of approaches consider the inflated zero counts inherent to the scRNA-seq data through various provisions in the underlying statistical models. However, they use complex linear models to perform DEA of genes, which require more computational time to individually fit the models for each gene. Therefore, another class of parametric approaches is reported in the literature, which is straight forward. In other words, this class of approaches is quite simple to execute, as they compare the mean expressions/estimated parameters of genes across two cell groups/populations. Further, their modes of execution require two simple steps: (i) estimation of gene-level parameters including mean, dispersion, etc., using a parametric model; and, (ii) comparison of the estimated mean parameters of genes between the two cell groups through a parametric statistical test. The operational procedure for this class of approaches is shown in Figure 3.

This class of approaches includes, scDD [50], DEsingle [51], NYMP [52], IDEAS [53], and, t -test [10], etc. For instance, DEsingle assumes the UMI counts must follow ZINB distribution [51], as given in Equation (4), but does not use GLM framework to model the

mean parameter. Further, DEsingle uses the EM algorithm to estimate the parameters of the ZINB model across the two cell types, while scDD utilizes a conjugate Dirichlet normal distribution to fit the expression data, thus handles the hidden cellular heterogeneity. The underlying statistical models and unique features of the other approaches are listed in Table 1.

The underlying null hypothesis for this class of approaches can be expressed as:

$$H_0 : \mu_{i1} = \mu_{i2} \text{ vs. } H_1 : \mu_{i1} \neq \mu_{i2} \quad (19)$$

where, μ_{il} is the mean expression of i th ($i = 1, 2, \dots, N$) gene in l th ($l = 1, 2$) cell populations/groups.

This class of approaches statistically tests the estimated value of the mean parameter of genes across the two cell populations using various test statistic(s) [10,16,51]. For instance, DEsingle uses the LRT (following a Chi-square distribution), while the scDD uses the Bayesian posterior probability. The test statistic(s) for other approaches is given in Table 1. Though this class of approaches is simple and quick in terms of execution, it suffers from serious limitations.

Limitations: (i) *Only two groups:* This class of approaches cannot be generalized to accommodate multiple cellular groups, though it is clear that scRNA-seq data are characterized by the presence of multiple cell types/groups, which these methods are unable to consider. This is due to the fact that other classes of methods including GLM, GAM, Hurdle, and MM, consider the GLM to model the mean parameter of genes, which can accommodate the multi-group comparison; (ii) *Cell-level auxiliary data:* The incorporation of cell-level confounding covariates including cell type, cell cycle, cell growth phase, etc., in the DEA improves the statistical power to detect true differentially expressed genes in single-cell studies. Therefore, this class of approaches cannot accommodate such auxiliary data in the analysis, and thus has poor performance compared with other class of approaches [10,16]; and, (iii) Many aforementioned approaches consider the inflated zero counts through parametric models (e.g., ZINB) which might not be sufficient to capture the heterogeneity in the scRNA-seq data. Further limitations and unique features of this class of approaches are listed in Table 2.

3.6. Non-Parametric Approaches

The approaches described in the previous five classes assume that the expression counts follow a well-defined parametric distribution. These approaches are typically slower due to the implementation of complex statistical models and iterative algorithms of parameter estimation. Thus, parametric models may not be ideally suited to data from many hundreds or thousands of single-cells [47,56,60]. Furthermore, parametric tests including LRT, Wald test, etc., have been used to compute the statistical significance of differentially expressed genes across the cell population. However, in statistics, NP methods have statistical power at par or greater than parametric methods, if the data violate the underlying assumptions of the parametric methods and a large number of samples/cells exist [73]. Under these circumstances, NP approaches can be better alternatives to their parametric counterparts, and are ideally suited to large single-cell datasets [74]. Further, the NP class of approaches can capture the multi-modal nature of the single-cell data. The major pros and cons of this class of approaches are described in Table 2.

The NP class of approaches includes NODES [56], Wilcoxon signed rank test (Wilcox) [10,16], ROTS [75], EMDomics [57], ROSeq [60], and SINCERA [55], etc. This class of approaches estimates the parameters that can quantify the distribution of expression profiles and makes comparisons between two cell groups. The null hypothesis of this class of approaches can be expressed as:

$$H_0 : F_i = G_i \text{ vs. } H_1 : F_i \neq G_i$$

where, F_i and G_i are the distributions of i th gene in the first and second cell groups, respectively. The above null hypothesis indicates that the two cell populations have the same distributions or derive from a same cell population. For instance, NODES and Wilcox usually test significant differences of genes' mean expressions across the two cell groups [10,16]. The former requires pseudo-counted quantile normalized gene expression values for DEA, while the latter can be used for counts or normalized data. Further, these approaches use test statistic(s) based on ranks (e.g., Wilcox, ROSeq), quantiles or percentiles (e.g., NODES), or distance measures (e.g., EMDomics), etc., for DEA of genes. For instance, Wilcox compares the ranks of the expression values that derive from the two cell groups. This rank-based test mostly ignores the magnitude of the expression deviations of genes between the two cell groups. Moreover, this class of approaches mostly uses permutation or bootstrap procedure to compute the p -values for genes (e.g., NODES [56], ROTS [75]). These approaches are relatively simple to understand and easy to execute for large scRNA-seq datasets with relatively lesser runtime required, compared with the other five classes.

Limitations: (i) *Lesser statistical power:* If all of the assumptions of the parametric approaches are apparently met by the single-cell data, and the DEA hypothesis can be tested with a parametric approach, then NP approaches may not be suitable. The degree of unsuitableness can be expressed in terms of lesser statistical power. Previous studies indicate that ZIM (Supplementary Documents S3–S5) fits well to the single-cell data [13,34]. Subsequently, DEA approaches based on ZIM usually have better performance over NP approaches [10,16]; (ii) NP approaches are not systematic, whereas parametric approaches have been systematized, and different tests are simply variations on a central theme; (iii) Another objection to NP approaches is related with convenience. Tables necessary to implement NP tests are scattered widely and appear in different formats; and, (iv) The results may or may not provide an accurate answer because they are distribution free. Further limitations and special features of this class of approaches are listed in Table 2.

4. Outstanding Challenges

The challenges in DEA of scRNA-seq data can be divided into two broad categories: (i) biological challenges, and (ii) methodological challenges.

4.1. Biological Challenges

We believe that development of the DEA approaches will require improvement of the existing annotations of genes on real single-cell data applications. Therefore, it is necessary to create accurate, high resolution knowledge bases with detailed annotations of genes. These knowledge bases will help investigators assess their DEA approach's performance based on the biological ground truth, and will also help in understanding the biological process from a systems biology point of view.

4.1.1. Proper Biological Benchmarking

Simulation techniques are usually used to validate the performance of the DEA approaches in scRNA-seq studies [11,14,16–18,24,51]. In simulation studies, the ground truth (reference genes) is (artificial) known and biological data are mimicked through statistical models. Then, these artificial single-cell data are used to assess the performance of the statistical approaches, given the artificial reference genes. In other words, DEA approach's performance is assessed through comparing the obtained differentially expressed genes with the given reference genes using sensitivity–specificity-based performance metrics [10–18,24,33,34]. However, Glazko et al. (2009) showed that statistical method's performance on simulated and real biological data are significantly different [76], which raises several questions about the performance assessment of methods using simulation as a benchmark. The attributable reason may be that the biology is more complicated than artificial scenarios and is influenced by factors such as the absence of an exclusive division into classes, presence of outliers, biological or technical hidden factors, environ-

mental influence(s), and random errors, etc. This aspect of performance evaluation is highly questionable among stakeholders, and needs further exploration.

To tackle this issue, researchers started using reference genes from microarrays [13,17,42] and bulk RNA-seq [10,13,24] (for the same bulk cell lines) to validate the performance of the DEA approaches on real single-cell data. However, this technique of performance validation has faced many criticisms from researchers, as scRNA-seq is the latest technology and obtaining reference genes from the old techniques incurs a technical lag. Therefore, to assess the performance of DEA approaches, proper biological benchmarking platform is required in single-cell studies.

4.1.2. Annotation

Biology-based techniques have been utilized to assess the performance of DEA methods in microarrays data analysis [21,22]. In other words, the genes from microarrays are validated using bio-knowledge bases (KEGG, STRING, GO terms, pathways, QTLs, etc.) in order to find biological processes and pathways that are relevant to the underlying condition [77]. For instance, the QTLs, GO, etc., tools were previously used to validate the performance of DEA methods in microarrays under a biological framework [77,78]. Further, these bio-knowledge bases were well established for microarrays and RNA-seq studies [77,78]. The scRNA-seq technique has shifted the paradigm of gene expression dynamics to the single-cell resolution-level. Therefore, the current annotation databases need to be updated with respect to these high-resolution techniques. It is essential that they also begin specifying gene and cell level annotation information. Such information will provide a better platform for assessing scRNA-seq DEA approaches from biological perspectives.

In addition to annotation, other information including literature support and expert interpretation can be considered while assessing DEA approaches. Further, statisticians and bioinformaticians must work closely with experimental biologists to validate their in-silico findings in wet-lab conditions.

4.2. Methodological Challenges

In addition to the above biological challenges, we also highlight the methodological challenges involved in DEA of scRNA-seq data.

4.2.1. Gold Standard scRNA-seq Data

The huge availability of statistical approaches for DEA of scRNA-seq data has prompted the search for methods which produce biologically accurate results. To address this, computational biologists have turned to simulations to mimic the biological ground truth through which DEA approaches could be benchmarked. Further, simulations require the specification of a model through which scRNA-seq data are generated. Differences in model parameters specifications have led researchers to generate irreproducible results [79]. These lacunae motivate for requirement a sound epistemological framework for DEA of scRNA-seq data [79,80]. To address this, Squair et al. (2021) suggested the quantification of performance of the DEA methods across multiple datasets in which the experimental ground truth was known, and also identified the principles/factors responsible for their performance differences [79]. This framework of gold standard data (i.e., data with known biological ground-truth) may provide a suitable platform for assessing the performance of the DEA approaches from a biological perspective.

4.2.2. Excess Heterogeneity

The scRNA-seq data tend to have abundance of zero counts, complicated underlying distributions, and huge heterogeneity. Subsequently, the heterogeneity between and within cell populations poses greater challenges to the DEA of scRNA-seq data [53]. Further, this cellular heterogeneity in data will increase multifold if the single-cell data are collected over individuals/patients. Previously, researchers have usually considered bulk RNA-seq methods for DEA of scRNA-seq data, which may not be sufficient to handle the huge

heterogeneity in the data. The implemented models in the existing single-cell DEA approaches (Supplementary Document S1) could best solve cellular heterogeneity. However, this may not be sufficient to tackle the heterogeneity in the data for single-cell studies over individuals/patients. Therefore, novel statistical approaches and tools are required for DEA of highly heterogeneous scRNA-seq data.

4.2.3. Dropouts or Excess Zeros of Single-Cell Data

Existing approaches of bulk RNA-seq DEA have been optimized for bulk tissue samples, and either perform poorly on single-cell data or do not accommodate the special features brought out by the revolutionary single-cell technology [81]. The zeros in scRNA-seq data are mainly due to biological and non-biological sources, a well-known challenge in scRNA-seq data analysis, and how to best tackle it remains a controversial topic. It is very difficult to distinguish between biological and non-biological zeros in scRNA-seq data without pre-defined knowledge or spike-in control [8]. Therefore, researchers started using ZIM or Hurdle models to tackle the issue of zero-inflation or excess zeros [8,13]. For instance, the Hurdle model failed to consider the sources of zeros, and assumed them to be from a single source. A significant portion of these zeros is due to dropout events, which need to be addressed in the modelling process for better DEA. Another strategy of handling the dropout/false zeros (attributed to inefficient sample preparation and sequencing protocol) is through suitable data imputation tools (e.g., scImpute and DrImpute, etc.) [82]. Furthermore, lower transcriptional captures in single-cells also contribute to dropout events in the data. For instance, efficient protocols of single-cell sequencing can capture 1–10% of the transcriptomes present in the cell [13,24,83]. Hence, different capture rate models including Binomial and Hypergeometric, etc., [13,24] can be used to adjust the cellular capture rate while modelling the observed UMI counts. These efforts will help mitigate the issues associated with single-cell data that limit the utilization of existing DEA approaches.

4.2.4. Pre-Processing of scRNA-seq Data

The DEA seems to be a single-step process, but actually, unavoidably, is a multi-step process, and its success highly depends on the pre-processing of scRNA-seq data. For instance, scRNA-seq data usually have poor quality or outlier cells, which may bias the analytical findings if included in the analysis. Therefore, researchers remove the cells whose library size lies below a certain threshold [24], which is an empirical approach and does not consider the statistical distributions of the cell library sizes. Hence, bioinformatics tool developers must consider the pre-processing steps applied to input data, and the DEA may be performed on the processed data. For example, the popular Seurat package uses many data pre-processing steps before DEA of genes [48]. These pre-processing steps include filtering low-quality genes and cells, data normalization, pre-feature selection, dimensionality reduction, and cell clustering [49]. Hence, if DEA tool developers are not aware of these pre-processing steps, their bioinformatics tools may not identify true differentially expressed genes. In other words, accuracy and reproducibility of the DEA tool will depend on pre-processing of scRNA-seq data. Therefore, tool developers must consider data pre-processing as an integral part of the DEA for real data applications.

4.2.5. Lack of Biological Relevant Criteria

The performances of the scRNA-seq DEA approaches are usually assessed on simulated data through sensitivity–specificity-based criteria. Though these criteria are statistically strong, they fail to state the biological relevance of the stated approaches. To address this, biologically relevant criteria (based on GO, QTL, and pathways, etc.) under a sound statistical framework have been developed for microarrays studies [21,77,78]. However, such comparative indices are missing in single-cell data analytics. In other words, such an assessment will answer the question of whether the differences between DEA approaches can impact the functional interpretation of transcriptomics experiments. Hence, Squair et al. (2021) used the GO term enrichment analyses in bulk vs. scRNA-seq DEA to assess the

biological relevance of approaches [79]. However, strong statistical criteria are required for this purpose, based on biologically relevant information including GO, QTL or pathways information for single-cell datasets.

4.2.6. Statistical Methods for DEA across Individuals

The standard practice in DEA of scRNA-seq studies is to collect many cells from one or a few individuals, and finding differentially expressed genes through the comparison of gene expression between two cell groups/clusters. Several methods have been developed for this purpose [10,11,16,24,51]. Currently, the scRNA-seq technique is slowly becoming a standard practice, and many investigators have started generating scRNA-seq data from multiple individuals. Thus, the DEA of genes across groups of individuals (i.e., comparison between case and controls) has opened new avenues, which requires novel and innovative statistical approaches and tools. The existing DEA approaches are inappropriate for individual level differential expression testing, as the sampling units for these approaches are cells, not individuals. For this purpose, IDEAS [53] is the only recently developed technique which performs DEA of genes across individuals, by capturing their cell type-specific gene expressions. However, computational biologists and bioinformaticians may focus on developing novel approaches and tools using multi-level hierarchical linear models.

4.2.7. False Discoveries in DEA

If a method fails to account for cell–cell variations in DEA, then it could produce false discoveries in the presence of a real biological perturbation. The false discoveries may also arise in the absence of any biological difference. For instance, recent computational studies confirmed that single-cell methods produced a systematic excess of false positives compared with the bulk of RNA-seq DEA methods [79]. In addition, they found that the genes falsely identified as differentially expressed corresponded to those with the highest variability between replicates [79]. This exposes a fundamental pitfall for DEA in single-cell transcriptomics. In other words, the single-cell studies, especially in human, would exhibit greater variability between biological replicates, and consequently would be more vulnerable to false discoveries in DEA. These false discoveries are poised to mislead investigators. Therefore, novel and innovative statistical approaches and tools are of paramount importance to address this issue in DEA of scRNA-seq data.

4.2.8. Improved Methods for Dispersion Estimation

In most of the DEA tools including DECENT, DEsingle, and SwarnSeq, etc., the MLE method has mostly been used to estimate the dispersion parameter through iterative algorithms (e.g., EM and ECM) [13,24,34,51]. The dispersion parameter represents the cellular variability, thus obtaining its good estimate is crucial to finding the true differentially expressed genes. For this purpose, Empirical Bayes (EB) shrinkage estimation using weighted conditional log-likelihood method was used in bulk RNA-seq DEA methods [84]. This type of estimate shrinks the dispersion estimates toward a common prior, instead of shrinking them completely to the common dispersion. Therefore, any forms of the EB method in the estimation of dispersion parameter may be incorporated in approaches including DECENT, SwarnSeq, ZINB–WaVE, and ZingeR, to have better performance. Given that scRNA-seq data are very sparse, it may be expected that there is potential benefit in using the EB to improve the existing approaches performance. Such an attempt will stabilize estimates of the gene-specific dispersion parameter. For instance, MAST [40] used the EB to shrink the gene-specific variance parameter. Therefore, it is imperative to implement the EB or equivalent methods within scRNA-seq DEA approaches to stabilize the dispersion, for better analysis.

4.2.9. Random/Mixed Effect Models

The statistical models implemented in scRNA-seq DEA tools assume various factors (e.g., cell group and cell-level auxiliary variables) affecting the gene parameters, including

that mean and zero-inflation have fixed effects [13,24]. For instance, DECENT and SwarnSeq assume the cellular groups, cell types and cell-level auxiliaries, etc., have fixed effects on gene mean and zero-inflation. The trend is the same for all of the developed approaches. Sometimes, these assumptions are necessary for methodological derivations, but are highly unrealistic in biology, as some factors may have random effects. This is due to the fact that cell biology is a highly dynamic system, and factors affecting genes' expression have random/mixed effects. Therefore, researchers may think of implementing random or mixed effect models in DEA approaches, for better results.

4.2.10. Optimal Combination of Algorithms

Previous studies have shown that no statistical approach performed better for all the single-cell datasets [10,11,14–18]. It has even been found that some bulk RNA-seq DEA methods performed at par, or even better, compared with their single-cell counterparts [10,16]. In other words, statistical tools have their limitations and distributional assumptions about the data, which make them sensitive to real data applications. Thus, the differentially expressed genes identified by them are quite different from each other [10], making them data dependent, which leads to unstable and sometimes inaccurate results [61]. To tackle this issue, a more reliable strategy is to apply all methods at hand, and form a community prediction for better analytical findings. Alternatively, combining an assortment of different DEA approaches can be a better choice for finding true genes from single-cell data. For instance, Li et al. (2022) developed an ensemble learning-based computational framework to produce more stable and accurate results through combining results from 12 individual approaches [61]. This finding has opened the quest to obtain the optimal combination of state-of-the-art individual approaches, for better results. This aspect of obtaining better results through combining algorithm(s) is at infant stage, and more computational studies are needed in single-cell studies. We have listed the strengths and weaknesses of each class of methods in Table 2. A natural extension may be that a suitable combination of approaches can be a good strategy for the finding of true differentially expressed genes, as the methods may mask each other's weaknesses.

4.2.11. Integration of Multi-Omics Data

Single-cell multi-omics profiling technologies are rapidly evolving, bringing newer techniques to improve our understanding of the unique function of the basic atom of life. Recently, progress has been made in single-cell analytics to more accurately detect cell types, performing downstream analyses, correcting technical sources of error, and delineating cell lineages and cell-state transitions, etc., [17,18,69,85–91]. For instance, other high-throughput genomic studies, including genome-wide association study, have emerged as a powerful approach to identify risk variants; hence, such data can be integrated with scRNA-seq data for better identification of true marker genes.

In this direction, a computational framework has been developed to integrate association data with scRNA-seq data for the identification of novel cell types and marker genes in COVID-19 infected patients [92]. Such integrated data require advanced tools for DEA. Additionally, other single-cell cross platform datasets are available due to the advancement of genomic technologies, which can be integrated with scRNA-seq data for identifying true biological differentially expressed genes. For instance, fluorescence in situ hybridization (FISH) methods [93] provide data on the spatial distribution of single-cells, which can be used as priors in the modelling of gene expressions, preferably through a Bayesian approach.

Another advantage of integrating FISH data with sequencing data is that the former is extremely accurate and free from dropout events [93], which will compensate the high dropout events in the scRNA-seq. Further, in the absence of, or minimal, dropout events, it will be possible to accurately model the observed expression counts of genes. It is also possible to integrate phenotypic data of cellular activation with scRNA-seq data for effective modelling of gene expression. Broadly, integration of single-cell sequencing (e.g.,

scRNA-seq) approaches with high throughput single-molecule imaging (e.g., FISH) or GWAS has a better chance to identify true DE genes at single-cell level, which requires innovative statistical approaches.

4.2.12. Slow Computational Processing

For DEA, one must consider the computational processing speed for large-scale scRNA-seq datasets, as one may not wait for several hours to obtain the results [10,13,24,51]. For instance, the Drop-seq-based system generates expression data of thousand(s) of genes over a large number of cells (e.g., from 10,000 to million cells), and the existing approaches fit complex statistical models individually to each gene, which costs a lot of computational time. Further, these approaches and tools employ complex iterative algorithms (e.g., EM, ECM) to estimate the gene specific parameters (most cases do not converge). Through computational experiments, it was found that methods including ZingeR, DECENT, DEsingle, and SwarnSeq, etc., require several hours for even a small dataset [10]. This situation will be more worrisome for real large experimental single-cell datasets. Additionally, researchers use the artificial setup with a small number of genes over hundreds of cells to assess the performance of their methods, which is far from the experimental reality. Thus, computationally efficient tools are required for the DEA of scRNA-seq data. Sometimes, to speed up the analysis, researchers consider pre-selected genes or dimensionality reduction technique, which restricts the analytical inference to prior knowledge or ignoring other important genes. Hence, future DEA approaches and tools must consider the scalability issue in single-cell studies.

5. Conclusions

DEA has become the primary downstream analysis of scRNA-seq data for extracting valuable biological insights into high-throughput gene expression measurements. This analysis also provides input to other secondary bioinformatics analyses including gene set analysis, gene network analysis, and pathways analysis, etc. To date, several statistical approaches and tools have been developed in the literature based on various statistical principles. This paper discusses the critical reviews of state-of-the-art methods available for DEA of scRNA-seq data, and distinctly classifies them into six major classes based on the underlying statistical models. Although the first three classes of methods, namely, GLM, GAM, MM, and Hurdle model approaches are extremely popular due to their ability to accommodate cell-level auxiliaries, they are computationally complex and runtime intensive.

Despite these developments, certain challenges exist in the DEA of scRNA-seq data, which need to be addressed in the future to develop improved classes of methods. We grouped the existing challenges of DEA into biological and methodological challenges. Under the biological challenges, a lack of proper biological benchmarking and incomplete annotations of genes in single-cell studies restrict the ability to assess the performance of DEA approaches for speaking the biological ground truth. We also reported several methodological challenges in DEA. The bioinformatics community must address these challenges to develop novel and innovative classes of DEA approaches and tools. These new approaches will utilize the features of the relatively new high-throughput single-cell technologies in order to better understand large biological systems.

Supplementary Materials: The following supporting information can be downloaded at: <https://www.mdpi.com/article/10.3390/e24070995/s1>, Supplementary Document S1: scRNA-seq DE Analysis Approaches; Supplementary Document S2: Count data models; Supplementary Document S3: Testing for zero inflation parameters for genes in scRNA-seq data; Supplementary Document S4: Statistical testing for overdispersion parameters in scRNA-seq data; Supplementary Document S5: Application of Count Data Models to Zero-Inflated and Overdispersed Real Datasets. References [94–105] are cited in the supplementary materials.

Author Contributions: Conceived and designed the study, contributed materials, carried out the study, drafted the manuscript, S.D.; corrected the manuscript, S.D., A.R. and S.N.R.; funding acquisition, S.D., A.R. and S.N.R. All authors provided critical feedback and helped shape the research, analysis, and manuscript. All authors have read and agreed to the published version of the manuscript.

Funding: This study was supported in part by the Science Education Research Board (SERB), New Delhi, India (grant CRG/2021/004960); the ICAR-Indian Agricultural Statistics Research Institute (ICAR-IASRI), New Delhi, India (grant AGEDIASRISIL202101800189); and the Wendell Cherry Chair of the Clinical Trial Research Fund (SNR).

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: Not applicable.

Acknowledgments: The authors duly acknowledge the help and support obtained from the Director and Head (Statistical Genetics) of ICAR-IASRI, New Delhi, India, and the Director of ICAR-ICFMD-DFMD, Bhubaneswar, India.

Conflicts of Interest: The authors declare no conflict of interest.

References

1. Liu, S.; Trapnell, C. Single-cell transcriptome sequencing: Recent advances and remaining challenges. *F1000Research* **2016**, *5*, 182. [[CrossRef](#)] [[PubMed](#)]
2. Kiselev, V.Y.; Andrews, T.S.; Hemberg, M. Challenges in unsupervised clustering of single-cell RNA-seq data. *Nat. Rev. Genet.* **2019**, *20*, 273–282. [[CrossRef](#)] [[PubMed](#)]
3. Saliba, A.-E.; Westermann, A.J.; Gorski, S.A.; Vogel, J. Single-cell RNA-seq: Advances and future challenges. *Nucleic Acids Res.* **2014**, *42*, 8845–8860. [[CrossRef](#)] [[PubMed](#)]
4. Macosko, E.Z.; Basu, A.; Satija, R.; Nemes, J.; Shekhar, K.; Goldman, M.; Tirosh, I.; Bialas, A.R.; Kamitaki, N.; Martersteck, E.M.; et al. Highly Parallel Genome-wide Expression Profiling of Individual Cells Using Nanoliter Droplets. *Cell* **2015**, *161*, 1202–1214. [[CrossRef](#)] [[PubMed](#)]
5. Zheng, G.X.Y.; Terry, J.M.; Belgrader, P.; Ryvkin, P.; Bent, Z.W.; Wilson, R.; Ziraldo, S.B.; Wheeler, T.D.; McDermott, G.P.; Zhu, J.; et al. Massively parallel digital transcriptional profiling of single cells. *Nat. Commun.* **2017**, *8*, 14049. [[CrossRef](#)]
6. Picelli, S.; Faridani, O.R.; Björklund, Å.K.; Winberg, G.; Sagasser, S.; Sandberg, R. Full-length RNA-seq from single cells using Smart-seq2. *Nat. Protoc.* **2014**, *9*, 171–181. [[CrossRef](#)]
7. Pollen, A.A.; Nowakowski, T.J.; Shuga, J.; Wang, X.; Leyrat, A.A.; Lui, J.H.; Li, N.; Szpankowski, L.; Fowler, B.; Chen, P.; et al. Low-coverage single-cell mRNA sequencing reveals cellular heterogeneity and activated signaling pathways in developing cerebral cortex. *Nat. Biotechnol.* **2014**, *32*, 1053–1058. [[CrossRef](#)]
8. Jiang, R.; Sun, T.; Song, D.; Li, J.J. Statistics or biology: The zero-inflation controversy about scRNA-seq data. *Genome Biol.* **2022**, *23*, 31. [[CrossRef](#)]
9. Svensson, V. Reply to: UMI or not UMI, that is the question for scRNA-seq zero-inflation. *Nat. Biotechnol.* **2021**, *39*, 160. [[CrossRef](#)]
10. Das, S.; Rai, A.; Merchant, M.L.; Cave, M.C.; Rai, S.N. A Comprehensive Survey of Statistical Approaches for Differential Expression Analysis in Single-Cell RNA Sequencing Studies. *Genes* **2021**, *12*, 1947. [[CrossRef](#)]
11. Mou, T.; Deng, W.; Gu, F.; Pawitan, Y.; Vu, T.N. Reproducibility of Methods to Detect Differentially Expressed Genes from Single-Cell RNA Sequencing. *Front. Genet.* **2020**, *10*, 1331. [[CrossRef](#)] [[PubMed](#)]
12. Vu, T.N.; Wills, Q.F.; Kalari, K.R.; Niu, N.; Wang, L.; Rantalainen, M.; Pawitan, Y. Beta-Poisson model for single-cell RNA-seq data analyses. *Bioinformatics* **2016**, *32*, 2128–2135. [[CrossRef](#)] [[PubMed](#)]
13. Das, S.; Rai, S.N. SwarnSeq: An improved statistical approach for differential expression analysis of single-cell RNA-seq data. *Genomics* **2021**, *113*, 1308–1324. [[CrossRef](#)] [[PubMed](#)]
14. Dal Molin, A.; Baruzzo, G.; Di Camillo, B. Single-cell RNA-sequencing: Assessment of differential expression analysis methods. *Front. Genet.* **2017**, *8*, 62. [[CrossRef](#)]
15. Wang, T.; Li, B.; Nelson, C.E.; Nabavi, S. Comparative analysis of differential gene expression analysis tools for single-cell RNA sequencing data. *BMC Bioinform.* **2019**, *20*, 40. [[CrossRef](#)]
16. Sonesson, C.; Robinson, M.D. Bias, robustness and scalability in single-cell differential expression analysis. *Nat. Methods* **2018**, *15*, 255–261. [[CrossRef](#)]
17. Jaakkola, M.K.; Seyednasrollah, F.; Mehmood, A.; Elo, L.L. Comparison of methods to detect differentially expressed genes between single-cell populations. *Brief. Bioinform.* **2016**, *18*, 735–743. [[CrossRef](#)]
18. Miao, Z.; Zhang, X. Differential expression analyses for single-cell RNA-Seq: Old questions on new data. *Quant. Biol.* **2016**, *4*, 243–260. [[CrossRef](#)]

19. Cui, X.; Churchill, G.A. Statistical tests for differential expression in cDNA microarray experiments. *Genome Biol.* **2003**, *4*, 210. [[CrossRef](#)]
20. Costa-Silva, J.; Domingues, D.; Lopes, F.M. RNA-Seq differential expression analysis: An extended review and a software tool. *PLoS ONE* **2017**, *12*, e0190152. [[CrossRef](#)]
21. Das, S.; Rai, A.; Mishra, D.C.; Rai, S.N. Statistical approach for selection of biologically informative genes. *Gene* **2018**, *655*, 71–83. [[CrossRef](#)]
22. Das, S.; Rai, S.N. Statistical approach for biologically relevant gene selection from high-throughput gene expression data. *Entropy* **2020**, *22*, 1205. [[CrossRef](#)] [[PubMed](#)]
23. Pratapa, A.; Jalihal, A.P.; Law, J.N.; Bharadwaj, A.; Murali, T.M. Benchmarking algorithms for gene regulatory network inference from single-cell transcriptomic data. *Nat. Methods* **2020**, *17*, 147–154. [[CrossRef](#)] [[PubMed](#)]
24. Ye, C.; Speed, T.P.; Salim, A. DECENT: Differential expression with capture efficiency adjustmeNT for single-cell RNA-seq data. *Bioinformatics* **2019**, *35*, 5155–5162. [[CrossRef](#)] [[PubMed](#)]
25. Vallejos, C.A.; Marioni, J.C.; Richardson, S. BASiCS: Bayesian Analysis of Single-Cell Sequencing Data. *PLoS Comput. Biol.* **2015**, *11*, e1004333. [[CrossRef](#)] [[PubMed](#)]
26. Jia, C.; Hu, Y.; Kelly, D.; Kim, J.; Li, M.; Zhang, N.R. Accounting for technical noise in differential expression analysis of single-cell RNA sequencing data. *Nucleic Acids Res.* **2017**, *45*, 10978–10988. [[CrossRef](#)]
27. Das, S.; Rai, S.N. Statistical methods for analysis of single-cell RNA-sequencing data. *MethodsX* **2021**, *8*, 101580. [[CrossRef](#)]
28. Wang, J.; Huang, M.; Torre, E.; Dueck, H.; Shaffer, S.; Murray, J.; Raj, A.; Li, M.; Zhang, N.R. Gene expression distribution deconvolution in single-cell RNA sequencing. *Proc. Natl. Acad. Sci. USA* **2018**, *115*, E6437–E6446. [[CrossRef](#)]
29. The External RNA Controls Consortium: A progress report. *Nat. Methods* **2005**, *2*, 731–734. [[CrossRef](#)]
30. Chen, W.; Li, Y.; Easton, J.; Finkelstein, D.; Wu, G.; Chen, X. UMI-count modeling and differential expression analysis for single-cell RNA sequencing. *Genome Biol.* **2018**, *19*, 70. [[CrossRef](#)]
31. Risso, D.; Perraudeau, F.; Gribkova, S.; Dudoit, S.; Vert, J.-P. A general and flexible method for signal extraction from single-cell RNA-seq data. *Nat. Commun.* **2018**, *9*, 284. [[CrossRef](#)] [[PubMed](#)]
32. Van den Berge, K.; Sonesson, C.; Love, M.I.; Robinson, M.D.; Clement, L. zingeR: Unlocking RNA-seq tools for zero-inflation and single cell applications. *bioRxiv* **2017**. [[CrossRef](#)]
33. Van den Berge, K.; Perraudeau, F.; Sonesson, C.; Love, M.I.; Risso, D.; Vert, J.-P.; Robinson, M.D.; Dudoit, S.; Clement, L. Observation weights unlock bulk RNA-seq tools for zero inflation and single-cell applications. *Genome Biol.* **2018**, *19*, 24. [[CrossRef](#)]
34. Mallick, H.; Chatterjee, S.; Chowdhury, S.; Chatterjee, S.; Rahnavard, A.; Hicks, S.C. Differential expression of single-cell RNA-seq data using Tweedie models. *Stat. Med.* **2022**, *41*, 3492–3510. [[CrossRef](#)]
35. He, Z.; Pan, Y.; Shao, F.; Wang, H. Identifying Differentially Expressed Genes of Zero Inflated Single Cell RNA Sequencing Data Using Mixed Model Score Tests. *Front. Genet.* **2021**, *12*, 616686. [[CrossRef](#)]
36. Shi, Y.; Lee, J.-H.; Kang, H.; Jiang, H. A Two-Part Mixed Model for Differential Expression Analysis in Single-Cell High-Throughput Gene Expression Data. *Genes* **2022**, *13*, 377. [[CrossRef](#)] [[PubMed](#)]
37. Trapnell, C.; Cacchiarelli, D.; Grimsby, J.; Pokharel, P.; Li, S.; Morse, M.; Lennon, N.J.; Livak, K.J.; Mikkelsen, T.S.; Rinn, J.L. The dynamics and regulators of cell fate decisions are revealed by pseudotemporal ordering of single cells. *Nat. Biotechnol.* **2014**, *32*, 381–386. [[CrossRef](#)] [[PubMed](#)]
38. Qiu, X.; Mao, Q.; Tang, Y.; Wang, L.; Chawla, R.; Pliner, H.A.; Trapnell, C. Reversed graph embedding resolves complex single-cell trajectories. *Nat. Methods* **2017**, *14*, 979–982. [[CrossRef](#)]
39. Van den Berge, K.; Roux de Bézieux, H.; Street, K.; Saelens, W.; Cannoodt, R.; Saeys, Y.; Dudoit, S.; Clement, L. Trajectory-based differential expression analysis for single-cell sequencing data. *Nat. Commun.* **2020**, *11*, 1201. [[CrossRef](#)]
40. Finak, G.; McDavid, A.; Yajima, M.; Deng, J.; Gersuk, V.; Shalek, A.K.; Slichter, C.K.; Miller, H.W.; McElrath, M.J.; Prlic, M.; et al. MAST: A flexible statistical framework for assessing transcriptional changes and characterizing heterogeneity in single-cell RNA sequencing data. *Genome Biol.* **2015**, *16*, 278. [[CrossRef](#)]
41. Sekula, M.; Gaskins, J.; Datta, S. Detection of differentially expressed genes in discrete single-cell RNA sequencing data using a hurdle model with correlated random effects. *Biometrics* **2019**, *75*, 1051–1062. [[CrossRef](#)] [[PubMed](#)]
42. Kharchenko, P.V.; Silberstein, L.; Scadden, D.T. Bayesian approach to single-cell differential expression analysis. *Nat. Methods* **2014**, *11*, 740–742. [[CrossRef](#)] [[PubMed](#)]
43. Delmans, M.; Hemberg, M. Discrete distributional differential expression (D3E)-A tool for gene expression analysis of single-cell RNA-seq data. *BMC Bioinform.* **2016**, *17*, 110. [[CrossRef](#)] [[PubMed](#)]
44. Wu, Z.; Zhang, Y.; Stitzel, M.L.; Wu, H. Two-phase differential expression analysis for single cell RNA-seq. *Bioinformatics* **2018**, *34*, 3340–3348. [[CrossRef](#)] [[PubMed](#)]
45. Zhang, W.; Wei, Y.; Zhang, D.; Xu, E.Y. ZIAQ: A quantile regression method for differential expression analysis of single-cell RNA-seq data. *Bioinformatics* **2020**, *36*, 3124–3130. [[CrossRef](#)]
46. Niyakan, S.; Hajiramezanali, E.; Boluki, S.; Zamani Dadaneh, S. SimCD: Simultaneous Clustering and Differential expression analysis for single-cell transcriptomic data. *arXiv* **2021**, arXiv:2104.01512.
47. Ling, W.; Zhang, W.; Cheng, B.; Wei, Y. Zero-inflated quantile rank-score based test (ZIQRank) with application to scRNA-seq differential gene expression analysis. *Ann. Appl. Stat.* **2021**, *15*, 1673–1696. [[CrossRef](#)]

48. Satija, R.; Farrell, J.A.; Gennert, D.; Schier, A.F.; Regev, A. Spatial reconstruction of single-cell gene expression data. *Nat. Biotechnol.* **2015**, *33*, 495–502. [[CrossRef](#)]
49. Hao, Y.; Hao, S.; Andersen-Nissen, E.; Mauck, W.M.; Zheng, S.; Butler, A.; Lee, M.J.; Wilk, A.J.; Darby, C.; Zager, M.; et al. Integrated analysis of multimodal single-cell data. *Cell* **2021**, *184*, 3573–3587.e29. [[CrossRef](#)]
50. Korthauer, K.D.; Chu, L.F.; Newton, M.A.; Li, Y.; Thomson, J.; Stewart, R.; Kendziorski, C. A statistical approach for identifying differential distributions in single-cell RNA-seq experiments. *Genome Biol.* **2016**, *17*, 222. [[CrossRef](#)]
51. Miao, Z.; Deng, K.; Wang, X.; Zhang, X. DEsingle for detecting three types of differential expression in single-cell RNA-seq data. *Bioinformatics* **2018**, *34*, 3223–3224. [[CrossRef](#)]
52. Ntranos, V.; Yi, L.; Melsted, P.; Pachter, L. A discriminative learning approach to differential expression analysis for single-cell RNA-seq. *Nat. Methods* **2019**, *16*, 163–166. [[CrossRef](#)] [[PubMed](#)]
53. Zhang, M.; Liu, S.; Miao, Z.; Han, F.; Gottardo, R.; Sun, W. IDEAS: Individual level differential expression analysis for single-cell RNA-seq data. *Genome Biol.* **2022**, *23*, 33. [[CrossRef](#)] [[PubMed](#)]
54. Katayama, S.; Töhönen, V.; Linnarsson, S.; Kere, J. SAMstr: Statistical test for differential expression in single-cell transcriptome with spike-in normalization. *Bioinformatics* **2013**, *29*, 2943–2945. [[CrossRef](#)] [[PubMed](#)]
55. Guo, M.; Wang, H.; Potter, S.S.; Whitsett, J.A.; Xu, Y. SINCERA: A Pipeline for Single-Cell RNA-Seq Profiling Analysis. *PLoS Comput. Biol.* **2015**, *11*, e1004575. [[CrossRef](#)]
56. Sengupta, D.; Rayan, N.A.; Lim, M.; Lim, B.; Prabhakar, S. Fast, scalable and accurate differential expression analysis for single cells. *bioRxiv* **2016**, 049734. [[CrossRef](#)]
57. Nabavi, S.; Schmolze, D.; Maitituoheti, M.; Malladi, S.; Beck, A.H. EMDomics: A robust and powerful method for the identification of genes differentially expressed between heterogeneous classes. *Bioinformatics* **2016**, *32*, 533–541. [[CrossRef](#)]
58. Wang, T.; Nabavi, S. SigEMD: A powerful method for differential gene expression analysis in single-cell RNA sequencing data. *Methods* **2018**, *145*, 25–32. [[CrossRef](#)]
59. Wang, Z.; Jin, S.; Liu, G.; Zhang, X.; Wang, N.; Wu, D.; Hu, Y.; Zhang, C.; Jiang, Q.; Xu, L.; et al. DTWscore: Differential expression and cell clustering analysis for time-series single-cell RNA-seq data. *BMC Bioinform.* **2017**, *18*, 270. [[CrossRef](#)]
60. Gupta, K.; Lalit, M.; Biswas, A.; Sanada, C.; Greene, C.; Hukari, K.; Maulik, U.; Bandyopadhyay, S.; Ramalingam, N.; Ahuja, G.; et al. Modeling expression ranks for noise-tolerant differential expression analysis of scRNA-seq data. *Genome Res.* **2021**, *31*, 689–697. [[CrossRef](#)]
61. Li, H.-S.; Ou-Yang, L.; Zhu, Y.; Yan, H.; Zhang, X.-F. scDEA: Differential expression analysis in single-cell RNA-sequencing data via ensemble learning. *Brief. Bioinform.* **2022**, *23*, bbab402. [[CrossRef](#)] [[PubMed](#)]
62. Müller, M. Generalized Linear Models. In *XploRe—Learning Guide*; Springer: Berlin/Heidelberg, Germany, 2000; pp. 205–228.
63. McCullagh, P.; Nelder, J.A. *Generalized Linear Models*; Springer: Boston, MA, USA, 1989; ISBN 978-0-412-31760-6.
64. Kaern, M.; Elston, T.C.; Blake, W.J.; Collins, J.J. Stochasticity in gene expression: From theories to phenotypes. *Nat. Rev. Genet.* **2005**, *6*, 451–464. [[CrossRef](#)] [[PubMed](#)]
65. Birtwistle, M.R.; Rauch, J.; Kiyatkin, A.; Aksamitiene, E.; Dobrzyński, M.; Hoek, J.B.; Kolch, W.; Ogunnaike, B.A.; Kholodenko, B.N. Emergence of bimodal cell population responses from the interplay between analog single-cell signaling and protein expression noise. *BMC Syst. Biol.* **2012**, *6*, 109. [[CrossRef](#)] [[PubMed](#)]
66. Singer, Z.S.; Yong, J.; Tischler, J.; Hackett, J.A.; Altinok, A.; Surani, M.A.; Cai, L.; Elowitz, M.B. Dynamic Heterogeneity and DNA Methylation in Embryonic Stem Cells. *Mol. Cell* **2014**, *55*, 319–331. [[CrossRef](#)] [[PubMed](#)]
67. Dobrzyński, M.; Nguyen, L.K.; Birtwistle, M.R.; von Kriegsheim, A.; Blanco Fernández, A.; Cheong, A.; Kolch, W.; Kholodenko, B.N. Nonlinear signalling networks and cell-to-cell variability transform external signals into broadly distributed or bimodal responses. *J. R. Soc. Interface* **2014**, *11*, 20140383. [[CrossRef](#)]
68. Bendall, S.C.; Davis, K.L.; Amir, E.D.; Tadmor, M.D.; Simonds, E.F.; Chen, T.J.; Shenfeld, D.K.; Nolan, G.P.; Pe’er, D. Single-Cell Trajectory Detection Uncovers Progression and Regulatory Coordination in Human B Cell Development. *Cell* **2014**, *157*, 714–725. [[CrossRef](#)]
69. Bacher, R.; Kendziorski, C. Design and computational analysis of single-cell RNA-sequencing experiments. *Genome Biol.* **2016**, *17*, 63. [[CrossRef](#)]
70. Moris, N.; Pina, C.; Arias, A.M. Transition states and cell fate decisions in epigenetic landscapes. *Nat. Rev. Genet.* **2016**, *17*, 693–703. [[CrossRef](#)]
71. Hafemeister, C.; Satija, R. Normalization and variance stabilization of single-cell RNA-seq data using regularized negative binomial regression. *Genome Biol.* **2019**, *20*, 296. [[CrossRef](#)]
72. Townes, F.W.; Hicks, S.C.; Aryee, M.J.; Irizarry, R.A. Feature selection and dimension reduction for single-cell RNA-Seq based on a multinomial model. *Genome Biol.* **2019**, *20*, 295. [[CrossRef](#)]
73. Anders, S.; Huber, W. Differential expression analysis for sequence count data. *Genome Biol.* **2010**, *11*, R106. [[CrossRef](#)] [[PubMed](#)]
74. Klein, A.M.; Mazutis, L.; Akartuna, I.; Tallapragada, N.; Veres, A.; Li, V.; Peshkin, L.; Weitz, D.A.; Kirschner, M.W. Droplet barcoding for single-cell transcriptomics applied to embryonic stem cells. *Cell* **2015**, *161*, 1187–1201. [[CrossRef](#)] [[PubMed](#)]
75. Seyednasrollah, F.; Rantanen, K.; Jaakkola, P.; Elo, L.L. ROTS: Reproducible RNA-seq biomarker detector-Prognostic markers for clear cell renal cell cancer. *Nucleic Acids Res.* **2016**, *44*, e1. [[CrossRef](#)]
76. Glazko, G.V.; Emmert-Streib, F. Unite and conquer: Univariate and multivariate approaches for finding differentially expressed gene sets. *Bioinformatics* **2009**, *25*, 2348–2354. [[CrossRef](#)] [[PubMed](#)]

77. Das, S.; McClain, C.J.; Rai, S.N. Fifteen Years of Gene Set Analysis for High-Throughput Genomic Data: A Review of Statistical Approaches and Future Challenges. *Entropy* **2020**, *22*, 427. [[CrossRef](#)]
78. Das, S.; Rai, A.; Mishra, D.C.; Rai, S.N. Statistical Approach for Gene Set Analysis with Trait Specific Quantitative Trait Loci. *Sci. Rep.* **2018**, *8*, 2391. [[CrossRef](#)]
79. Squair, J.W.; Gautier, M.; Kathe, C.; Anderson, M.A.; James, N.D.; Hutson, T.H.; Hudelle, R.; Qaiser, T.; Matson, K.J.E.; Barraud, Q.; et al. Confronting false discoveries in single-cell differential expression. *Nat. Commun.* **2021**, *12*, 5692. [[CrossRef](#)] [[PubMed](#)]
80. Mehta, T.; Tanik, M.; Allison, D.B. Towards sound epistemological foundations of statistical methods for high-dimensional biology. *Nat. Genet.* **2004**, *36*, 943–947. [[CrossRef](#)]
81. Chen, S.; Mar, J.C. Evaluating methods of inferring gene regulatory networks highlights their lack of performance for single cell gene expression data. *BMC Bioinform.* **2018**, *19*, 232. [[CrossRef](#)]
82. Hou, W.; Ji, Z.; Ji, H.; Hicks, S.C. A systematic evaluation of single-cell RNA-sequencing imputation methods. *Genome Biol.* **2020**, *21*, 218. [[CrossRef](#)]
83. Ziegenhain, C.; Vieth, B.; Parekh, S.; Reinius, B.; Guillaumet-Adkins, A.; Smets, M.; Leonhardt, H.; Heyn, H.; Hellmann, I.; Enard, W. Comparative Analysis of Single-Cell RNA Sequencing Methods. *Mol. Cell* **2017**, *65*, 631–643.e4. [[CrossRef](#)] [[PubMed](#)]
84. Robinson, M.D.; Smyth, G.K. Moderated statistical tests for assessing differences in tag abundance. *Bioinformatics* **2007**, *23*, 2881–2887. [[CrossRef](#)] [[PubMed](#)]
85. Sandberg, R. Entering the era of single-cell transcriptomics in biology and medicine. *Nat. Methods* **2014**, *11*, 22–24. [[CrossRef](#)] [[PubMed](#)]
86. Trapnell, C. Defining cell types and states with single-cell genomics. *Genome Res.* **2015**, *25*, 1491–1498. [[CrossRef](#)]
87. Islam, S.; Kjällquist, U.; Moliner, A.; Zajac, P.; Fan, J.B.; Lönnerberg, P.; Linnarsson, S. Characterization of the single-cell transcriptional landscape by highly multiplex RNA-seq. *Genome Res.* **2011**, *21*, 1160–1167. [[CrossRef](#)]
88. Luecken, M.D.; Theis, F.J. Current best practices in single-cell RNA-seq analysis: A tutorial. *Mol. Syst. Biol.* **2019**, *15*, e8746. [[CrossRef](#)]
89. Tung, P.-Y.; Blischak, J.D.; Hsiao, C.J.; Knowles, D.A.; Burnett, J.E.; Pritchard, J.K.; Gilad, Y. Batch effects and the effective design of single-cell gene expression studies. *Sci. Rep.* **2017**, *7*, 39921. [[CrossRef](#)]
90. Kolodziejczyk, A.A.; Kim, J.K.; Svensson, V.; Marioni, J.C.; Teichmann, S.A. The Technology and Biology of Single-Cell RNA Sequencing. *Mol. Cell* **2015**, *58*, 610–620. [[CrossRef](#)]
91. Stegle, O.; Teichmann, S.A.; Marioni, J.C. Computational and analytical challenges in single-cell transcriptomics. *Nat. Rev. Genet.* **2015**, *16*, 133–145. [[CrossRef](#)]
92. Ma, Y.; Qiu, F.; Deng, C.; Li, J.; Huang, Y.; Wu, Z.; Zhou, Y.; Zhang, Y.; Xiong, Y.; Yao, Y.; et al. Integrating single-cell sequencing data with GWAS summary statistics reveals CD16+monocytes and memory CD8+T cells involved in severe COVID-19. *Genome Med.* **2022**, *14*, 16. [[CrossRef](#)]
93. Cui, C.; Shu, W.; Li, P. Fluorescence In situ Hybridization: Cell-Based Genetic Diagnostic and Research Applications. *Front. Cell Dev. Biol.* **2016**, *4*, 89. [[CrossRef](#)] [[PubMed](#)]
94. Love, M.I.; Huber, W.; Anders, S. Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. *Genome Biol.* **2014**, *15*, 550. [[CrossRef](#)] [[PubMed](#)]
95. Malhotra, A.; Das, S.; Rai, S.N. Analysis of Single-Cell RNA-Sequencing Data: A Step-by-Step Guide. *BioMedInformatics* **2022**, *2*, 43–61. [[CrossRef](#)]
96. Robinson, M.D.; McCarthy, D.J.; Smyth, G.K. EdgeR: A Bioconductor package for differential expression analysis of digital gene expression data. *Bioinformatics* **2010**, *26*, 139–140. [[CrossRef](#)] [[PubMed](#)]
97. Zeileis, A.; Kleiber, C.; Jackman, S. Regression Models for Count Data in R. *J. Stat. Softw.* **2008**, *27*, 1–25. [[CrossRef](#)]
98. Kempc, D.; Kempa, W. Some properties of the “Hermite” distribution. *Biometrika* **1965**, *52*, 381–394.
99. Boon, W.C.; Petkovic-Duran, K.; Zhu, Y.; Manasseh, R.; Horne, M.K.; Aumann, T.D. Increasing cDNA Yields from Single-cell Quantities of mRNA in Standard Laboratory Reverse Transcriptase Reactions using Acoustic Microstreaming. *J. Vis. Exp.* **2011**, *53*, e3144. [[CrossRef](#)]
100. Macaulay, I.C.; Voet, T. Single Cell Genomics: Advances and Future Perspectives. *PLoS Genet.* **2014**, *10*, e1004126. [[CrossRef](#)]
101. Marinov, G.K.; Williams, B.A.; McCue, K.; Schroth, G.P.; Gertz, J.; Myers, R.M.; Wold, B.J. From single-cell to cell-pool transcriptomes: Stochasticity in gene expression and RNA splicing. *Genome Res.* **2013**, *24*, 496–510. [[CrossRef](#)]
102. Pierson, E.; Yau, C. ZIFA: Dimensionality reduction for zero-inflated single-cell gene expression analysis. *Genome Biol.* **2015**, *16*, 241. [[CrossRef](#)]
103. Wang, Y.; Navin, N.E. Advances and Applications of Single-Cell Sequencing Technologies. *Mol. Cell* **2015**, *58*, 598–609. [[CrossRef](#)] [[PubMed](#)]
104. McElduff, F.; Cortina-Borja, M.; Chan, S.-K.; Wade, A. When t-tests or Wilcoxon-Mann-Whitney tests won’t do. *Adv. Physiol. Educ.* **2010**, *34*, 128–133. [[CrossRef](#)] [[PubMed](#)]
105. Qiu, X.; Hill, A.; Packer, J.; Lin, D.; Ma, Y.-A.; Trapnell, C. Single-cell mRNA quantification and differential analysis with Census. *Nat. Methods* **2017**, *14*, 309–315. [[CrossRef](#)] [[PubMed](#)]