*Briefings in Bioinformatics*, 00(00), 2020, 1–10

**doi: 10.1093/bib/bbaa314**
Method Review

# Goals and approaches for each processing step for single-cell RNA sequencing data

Zilong Zhang, Feifei Cui, Chunyu Wang, Lingling Zhao and Quan Zou

Corresponding authors: Quan Zou, Institute of Fundamental and Frontier Sciences, University of Electronic Science and Technology of China, Chengdu, China, and Hainan Key Laboratory for Computational Science and Application, Hainan Normal University, Haikou, China. Tel.: +86 170-9226-1008; E-mail: zouquan@nclab.net; Lingling Zhao, Computer Science and Technology, Harbin Institute of Technology, Harbin, Heilongjiang, China. E-mail: Zhaoll@hit.edu.cn

## Abstract

Single-cell RNA sequencing (scRNA-seq) has enabled researchers to study gene expression at the cellular level. However, due to the extremely low levels of transcripts in a single cell and technical losses during reverse transcription, gene expression at a single-cell resolution is usually noisy and highly dimensional; thus, statistical analyses of single-cell data are a challenge. Although many scRNA-seq data analysis tools are currently available, a gold standard pipeline is not available for all datasets. Therefore, a general understanding of bioinformatics and associated computational issues would facilitate the selection of appropriate tools for a given set of data. In this review, we provide an overview of the goals and most popular computational analysis tools for the quality control, normalization, imputation, feature selection and dimension reduction of scRNA-seq data.

**Key words:** single-cell RNA sequencing; quality control; normalization; imputation; feature selection; dimension reduction

## Introduction

RNA sequencing (RNA-seq) technology is a powerful tool for profiling gene expression patterns, which makes it possible for biologists to study different transcriptomes in pooled cells [1]. However, conventional bulk RNA-sequencing only quantifies the average expression signal for a large population of cells and does not reveal the heterogeneity of cells. Single-cell RNA sequencing (scRNA-seq) technology has become increasingly popular because it allows researchers to identify the transcriptomes profiles of individual cells [2–14]. To date, scRNA-seq studies have already shown great efficacy in the discovery of novel cell types, reconstruction of developmental trajectories and study of tumour heterogeneity [15–22]. The development of this technology is also providing new insights to support a better understanding of biological development and disease [23–25].

Although scRNA-seq provides convenience in biological studies, many drawbacks remain regarding this technology. Due to the extremely low amounts of transcripts in a single cell, low capture efficiency of mRNA and technical losses during reverse transcription (RT), numerous cycles of the cDNA-amplification process are required to meet the needs of sequencing [26, 27]. These factors lead to extremely noisy, highly dimensional and sparse gene expression matrices [28, 29]. Therefore, to fully exploit scRNA-seq technology, specifically designed

**Zilong Zhang** is currently working as a Postdoctoral Researcher in the University of Electronic Science and Technology of China. He received his PhD degree from the University of Tokyo, Japan, in 2020. His research interests include single-cell sequencing data analysis, bioinformatics and machine learning.
**Feifei Cui** received her PhD degree from the University of Tokyo, Japan. She is currently a Postdoctoral Researcher at the University of Electronic Science and Technology of China. Her research interests include bioinformatics, deep learning and biological data mining.
**Chunyu Wang** is an Associate Professor in the School of Computer Science and Technology, Harbin Institute of Technology. His research interests include bioinformatics and machine learning.
**Lingling Zhao** is a Lecturer in the School of Computer Science and Technology, Harbin Institute of Technology, Harbin, Heilongjiang. Her research interests include machine learning application and system biology.
**Quan Zou** is a Professor at the Institute of Fundamental and Frontier Sciences, University of Electronic Science and Technology of China. He received his PhD from the Harbin Institute of Technology, China, in 2009. He is a Senior Member of IEEE and ACM. His research is in the areas of bioinformatics, machine learning and parallel computing.

**Submitted:** 15 August 2020; **Received (in revised form):** 10 October 2020

computational tools for scRNA-seq data are needed. In recent years, the explosive increase in single-cell analysis tools has increased the difficulty of selecting appropriate tools for a given set of data [30, 31]. Although several user-friendly tools [32–34] have been developed to process and interpret scRNA-seq data, they are to some extent a 'black box' for users. We believe that an understanding of the computational methods for each step would help researchers to choose more suitable pipelines and tools for their data. Accordingly, bioinformatics reviews are needed to summarize the goals and computational methods that apply to these tools.

Many reviews have focused on downstream analyses (clustering, trajectory inference, visualization, etc.) [35–37] of scRNA-seq data, whereas few papers have summarized the specific processing steps (e.g. methods for data quality control—QC, normalization, imputation, feature selection and dimension reduction). However, these processing steps are extremely significant for downstream analysis, especially because of the high sparsity and technical noise of scRNA-seq data. In this review, we focus on summarizing the goals and popular computational approaches for each processing step of scRNA-seq data.

## Overview of scRNA-seq technology

Several scRNA-seq protocols have been developed over the past decade [38–46]. In general, the pipeline of scRNA-seq includes the following steps: (i) single cells are first isolated from a tissue; (ii) cell lysis is performed to obtain mRNA; (iii) mRNA molecule capture is performed; (iv) mRNA is converted to cDNA using RT; (v) cDNA is amplified by polymerase chain reaction (PCR); (vi) library preparation is performed and (vii) sequencing is performed [16, 45, 47–49] (Figure 1).

All protocols can be roughly divided into two categories based on their quantification method. The first category includes full-length protocols (e.g. SMART-seq2 [44]), in which the sequencing step attempts to produce uniform coverage for each transcript, and the second category includes tag-based protocols, which only capture and sequence 3′-end transcripts (e.g. inDrop [45] and Drop-seq [46]) or 5′-end transcripts (e.g. STRT-seq [43]). Normally, tag-based protocols are used in combination with unique molecular identifiers (UMIs) [14, 50, 51], which are used as barcodes for individual mRNA molecules to reduce the technical noise during the RT and amplification steps.

Based on the strategy of capturing cells, the protocols can also be divided into plate-based and droplet-based methods. Because of the high throughput, droplet-based technology (e.g. inDrop [45], Drop-seq [46] and 10X Genomics) has become the most popular strategy for isolating single-cell RNA. Droplet-based commercial chromium from 10X Genomics can achieve thousands of captured cells per run [49]. However, the high-throughput nature of the droplet-based protocol occurs at the cost of low sequencing depth (i.e. reduced total transcripts captured from each cell) [52]. Therefore, droplet-based protocols contain more technical noise than plate-based protocols, while plate-based protocols are preferable when dealing with rare cell types.

After sequencing, the first step of scRNA-seq data analysis is to generate a gene expression matrix wherein each row represents a gene and each column represents a cell. Briefly, this process is accomplished by mapping each sequencing read to a reference genome and counting the number of mapped reads [53]. The initial output of FASTQ files obtained from the sequencing step are first pre-processed by QC and read alignment. The most popular tools for QC include FASTQC [54] and Cutadapt [55]. FASTQC takes sequencing FASTQ files obtained from the sequencing machine as input and returns a reads quality report, then Cutadapt could be used for trimming the reads to improve the reads quality. For pipelines using UMIs, UMI-tools [30] can also be performed to trim the barcode. As errors in the UMI sequence are common, UMI-tools introduce network-based methods to account for these errors when identifying PCR duplicates. The end-to-end pipeline Alevin [56] extends UMI-tools and is used as an alternative method to Cell Ranger pipeline. Compared with Cell Ranger, Alevin achieves a higher accuracy and is considerably faster. Due to the widely existence of deletion and mismatch errors in cell barcodes, identifying cell barcodes from all sequencing data is a very challenging computational task. As for the scRNA-seq data generated by 10X genomics, a known 'whitelist' of sequences (i.e. the list of all known barcode sequences that have been included in the assay kit) could be used as prior knowledge to simplify error-correction and read assignment [57]. With regard to the barcodes generated through split-pool synthesis (e.g. Drop-seq), multiple sequence alignment may be the best choice for detecting and correcting the errors [58]. For aligners, STAR [59], Tophat2 [60] and HISAT [61] are the most widely used. Generally, STAR shows both higher mapping accuracy and speed than TopHat2 and HISAT, while HISAT uses lower memory usage. STAR is used by the popular software Cell Ranger from 10X genomics for mapping and quantifying. Recently, the authors of STAR have presented a new addition named STARsolo [62] to STAR alignment program. Compared with Cell Ranger, STARsolo improves the computing speed and saves significant computing resources in barcode demultiplexing and UMI counting. In addition to the gene counts, STARsolo can also calculate counts for pre-mRNA counts, which is useful for single-nucleus RNA-seq. To reveal the true biological signals, processing steps must be performed on the gene expression matrix before further downstream analysis. In the section below, we summarize the goals and state-of-the-art methods for scRNA-seq processing steps: QC, normalization, imputation, feature selection and dimension reduction.

## Data cleaning by QC

Even after performing several QC steps during the preprocessing step, the raw count matrix usually still contains many low-quality cells, which can be caused by a variety of reasons, such as dead cells during cell isolation, inefficient RT or PCR amplification. These low-quality cells may cause several problems for downstream analysis: (i) the inclusion of low-quality cells, which have low expression values across all genes could lead to a similar pattern of gene expression and the formation of a cluster and thus may erroneously indicate a new cell type; and (ii) low expression values across cells make the analyses extremely sensitive to noise. For instance, contaminating with a low-quality exogenous transcript could lead to substantial impacts on variance estimation or dimension reduction.

Consequently, applying a filter for removing low-quality cells is extremely important before further analysis. Currently, widely used QC metrics include library size, expressed gene detection, proportion of reads or UMIs mapped to the mitochondrial genome and proportion of reads mapped to External RNA Control Consortium (ERCC) spike-in transcripts. Library size consists of the overall endogenous reads count or UMI count for each cell. Filtering lower library sizes for QC is easy to understand since low-quality cells generally have lower expression values. Moreover, cells with an extremely high library size also need

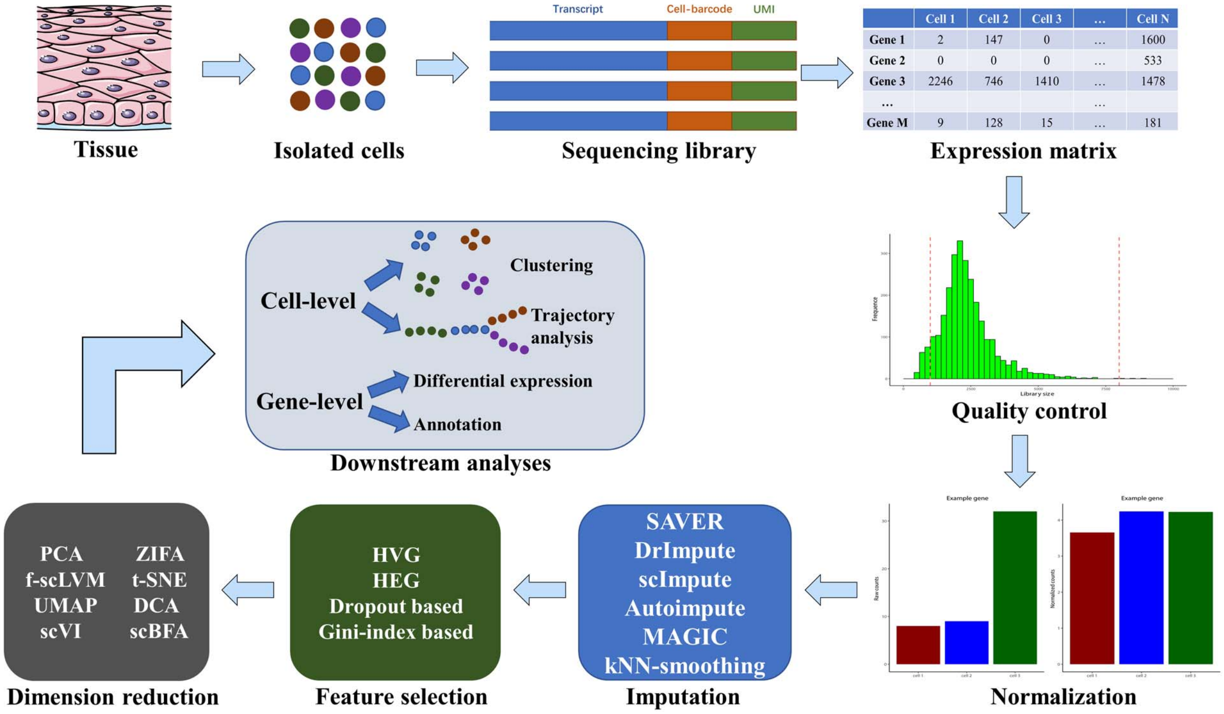

**Figure 1**. Workflow of scRNA-seq technology. Single cells are first isolated from biological tissue, and mRNA is captured and amplified to obtain the sequencing library. After sequencing, the gene expression matrix is processed by QC, normalization, imputation, feature selection and dimension reduction in a step-by-step process. Finally, the processed expression matrix can be used for cell-level and gene-level downstream analyses.

to be removed because they may be caused by a doublet (i.e. two or more cells captured in one droplet with the same cell barcode). Another QC metric is expressed gene detection, which means the number of endogenously expressed genes that were detected in each cell, and it is applied similarly to library size. A high proportion of mitochondrial RNA also indicates low-quality cells because in a damaged cell, mRNA transcripts efflux from the cell membrane, while the mitochondria are too large to escape, which leads to a high proportion of reads or UMIs mapped to the mitochondrial genome. Alternatively, synthetic spike-in RNA molecules from (ERCC) can also be used to reveal the true biological differences among cells [15, 63]. Similarly, when the same amount of ERCC spike-in is add to each cell, if the proportion of reads mapped to ERCC is too high in one cell, then it is identified as a low-quality cell. However, ERCC spike-ins are an alternative for scRNA-seq experiments, indicating that this metric does not apply to every dataset [64].

To remove low-quality cells, fixed thresholds for each metric are widely used (Figure 2). These thresholds are generally determined by experimental experience with the protocols and biological systems. Alternatively, adaptive thresholds could also be used based on the median absolute deviation (MAD) for each metric. Specifically, a cell is regarded as an outlier of high-quality cells and should be removed if it is more than three MADs from the median in any metric. Generally, the strategy of adaptive thresholds does not require a lot of experience to determine appropriate thresholds; hence, it is more friendly to a non-expert.

## Normalization

As previously mentioned, due to the different cDNA capture efficiency and PCR amplification, a great deal of technical bias occurs across cells. Normalization is a significant step to remove these biases across cells for downstream analyses. Historically, library size normalization is usually sufficient for bulk sequencing data [64]. The simplest library size method is to transform each read count or UMI count to counts per million (CPM). The CPM values are simply calculated by dividing the total count values for each sample and then multiplying by 1 million. The other confounder that may bias the results is gene length because longer genes will naturally have more reads mapped to them than shorter genes. To remove this bias, transcripts per million (TPM) [65], reads per kilobase million (RPKM) and fragments per kilobase million (FPKM) [66], which are similar to CPM, have been proposed. RPKM and FPKM are two very closely related terms and the only difference between them is that RPKM is for single end sequencing while FPKM is for paired end sequencing. Unlike RPKM and FPKM normalize for sequencing depth before normalizing for gene length, TPM first normalize for gene length and then normalize for sequencing depth, thus all transcripts in a sample shall add up to 1 million. However, these methods may all hide the biological signals we are interested in if the highly expressed genes (HEGs) are also the differentially expressed (DE) genes. Many other cell-specific scaling factor methods have been developed for bulk RNA-seq data, such as DESeq [67] and DESeq-2 [68], which calculate the size factor (SF) for each cell using the geometric mean of each gene.

However, compared with bulk RNA-seq data, scRNA-seq data are much sparser and consist of a high proportion of low and zero values caused by both biological differences and technical noise; therefore, the high proportion of zero counts reduces the accuracy of calculated geometric means. When the relationship between transcript expression and library size is not similar across genes, the 'SF' method tends to overcorrect the low and moderately expressed genes.

**Table 1.** Normalization methods for scRNA-seq data

| Method | Description | Availability | Refs |
|---|---|---|---|
| BASiCS | Uses spike-in genes to remove technical noise and normalizes the data using an estimated cell-specific constant | https://github.com/catavallejos/BASiCS | [69] |
| GRM | uses ERCC spike-in genes and FPKM values of reads to fits an GRM and then estimates the molecular concentration. | http://wanglab.ucsd.edu/star/GRM | [70] |
| scran | Pools cells by similar library size and then estimates an SF | https://bioconductor.org/packages/release/bioc/html/scran.html | [33] |
| Linnorm | utilizes a set of homogenously expressed gens as reference, and then calculates normalisation parameters by ignoring zero values. | http://www.jjwanglab.org/Linnorm/ | [71] |
| SCnorm | Uses quantile regression to group similar dependence genes, then performs within-group adjustment for library size to estimate scale factors | https://github.com/rhondabacher/SCnorm | [72] |
| Census | Converts conventional measures of relative gene expression levels (e.g. TPM) into relative transcript counts without the need for spike-in standards or UMIs | https://github.com/cole-trapnell-lab/monocle2-rge-paper | [73] |

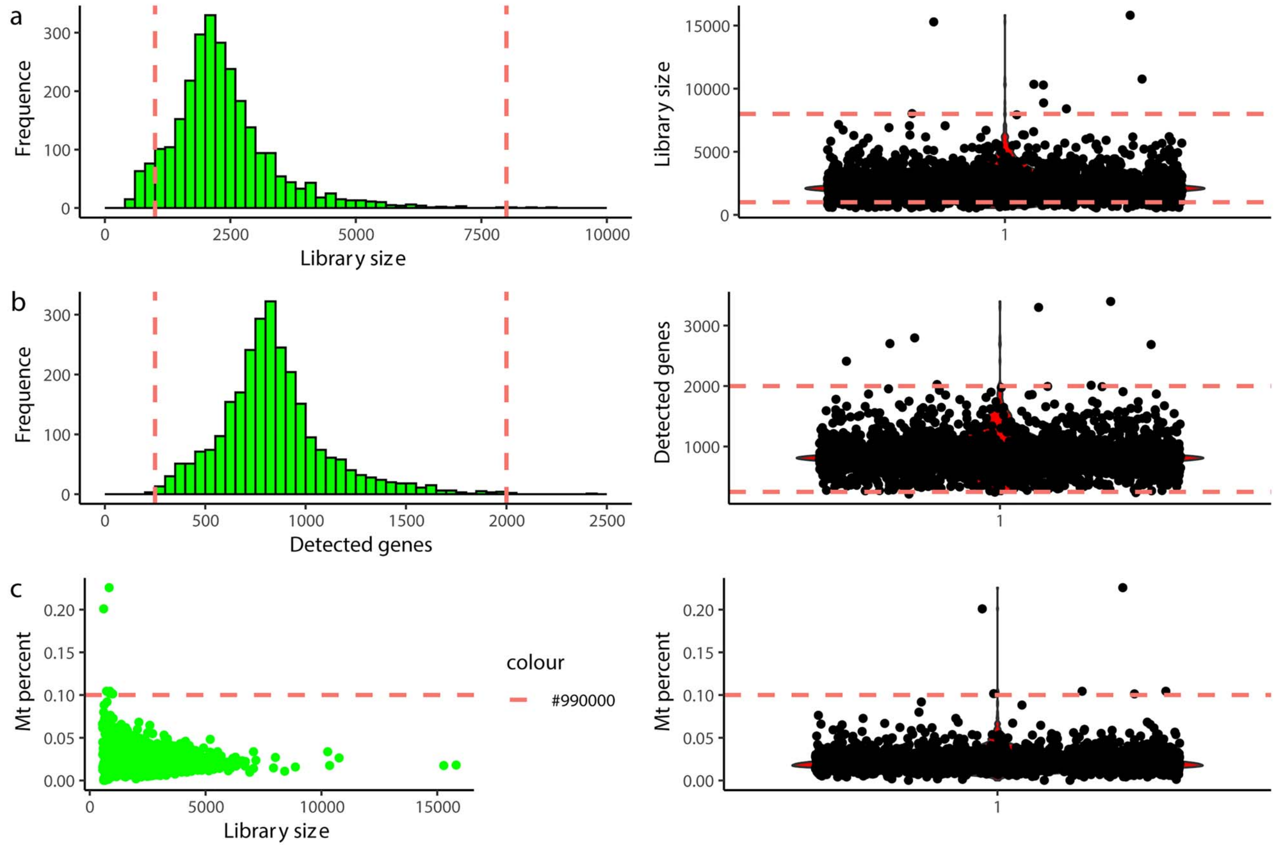

**Figure 2**. Example of QC for scRNA-seq data using fixed thresholds. (a) Low-quality cells are filtered out using the library size. As mentioned in the main text, cells with an extremely low library size and extremely high library size should be removed. In this example, library sizes larger than 1000 UMIs and smaller than 8000 are retained (regions between the two dashed lines). (b) Low-quality cells are filtered out using detected genes. Similar to part (a) but detected genes between 250 UMIs and 2000 UMIs are retained (regions between the two dashed lines). (c) Low-quality cells are filtered out using the proportion of reads or UMIs mapped to the mitochondrial genome (MT percent). Here, the cells with mitochondrial percentages higher than 10% were removed.

With the development of scRNA-seq technology, several specifically tailored single-cell normalization methods have been proposed (Table 1). BASiCS (Bayesian Analysis of Single-Cell Sequencing data) [69] uses ERCC spike-in genes to remove technical noise and normalize the data using an estimated cell-specific constant. Gamma Regression Model (GRM) [70] uses ERCC spike-in genes and FPKM values of reads to fits an GRM and then estimates the molecular concentration. Scran [33] pools cells by similar library size and then estimates an SF, thereby overcoming the problem of scRNA data being dominated by low and zero counts. Linnorm [71] utilizes a set of homogenously expressed gens as reference and then calculates normalization

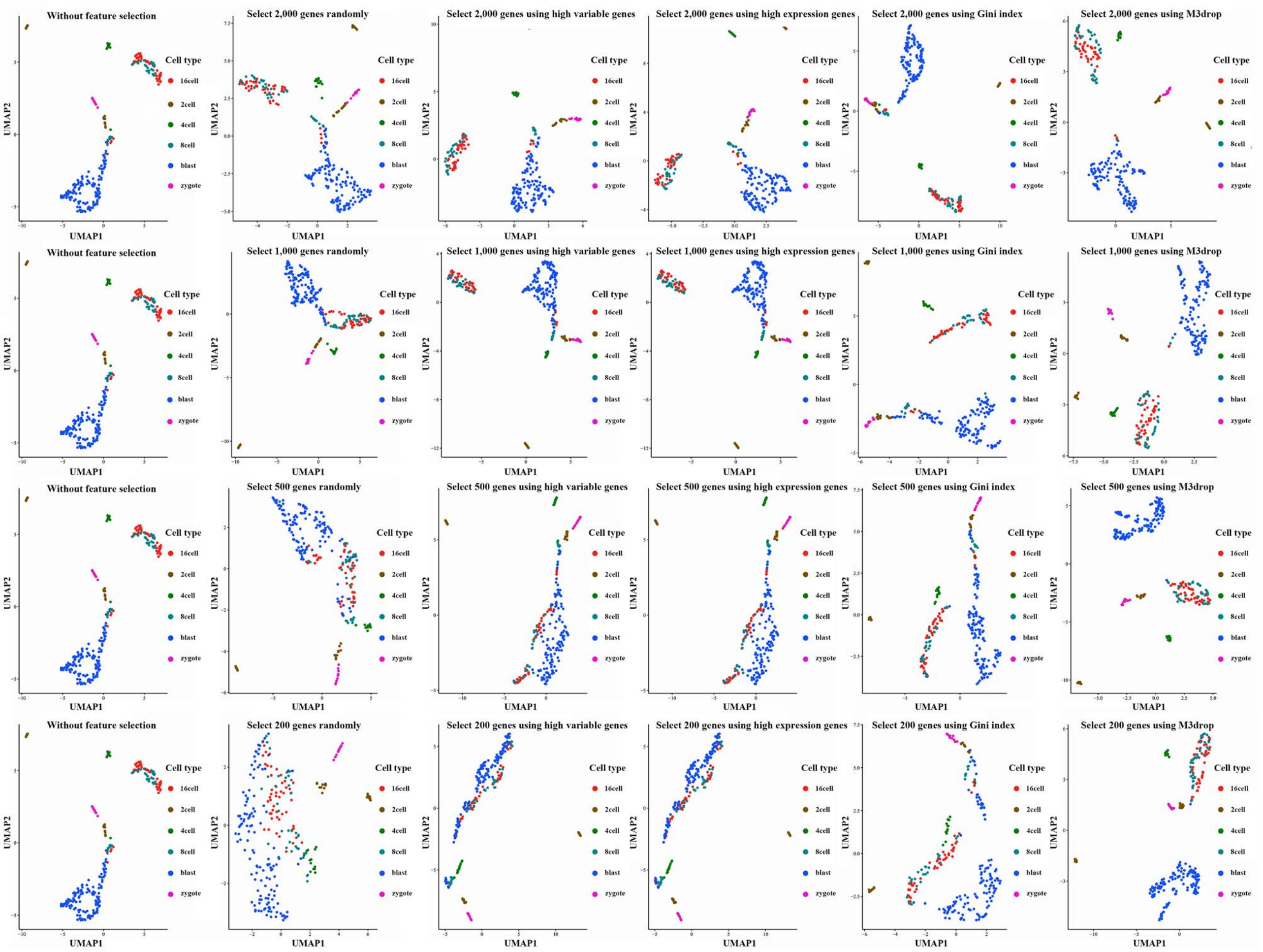

**Figure 3**. Comparison of visualization schemes using different feature selection methods and different numbers of genes in mammalian cells [91]. Each column indicates different feature selection methods, with the first using all genes without feature selection and the second using a certain number of random genes. The following feature selection methods are mentioned in the main text: HVG-based, HEG-based, Gini-index-based and Dropout-based methods. Each row uses different gene numbers: 2000, 1000, 500 and 200. After gene selection by each method, we applied the visualization method UMAP [92] for final visualization.

parameters by ignoring zero values. SCnorm [72] first uses quantile regression to group similar dependence genes and then performs within-group adjustment for library size to estimate scale factors. Census [73] converts conventional measures of relative gene expression levels (e.g. TPM) into relative transcript counts without the need for spike-in standards or UMIs.

In summary, due to the sparse characteristic of scRNA-seq data, traditional bulk RNA-seq normalization methods are not applicable. Specialized normalization methods tailored for scRNA-seq data, such as scran and SCnorm, would help the researchers to obscure true biological heterogeneity rather than technical biases.

## Imputation

Dropout events are widely existing phenomena in scRNA-seq data. A dropout event occurs when a gene is observed in one cell but shows zero or near zero expression values in another cell. Dropout events not only increase the cell-to-cell variability but also obscure gene–gene relationships [74, 75]. To alleviate the influence of widely existing dropout events, several imputation and smoothing methods have been developed (Table 2). The primary difference between imputation methods and smoothing methods is that imputation methods typically only focus on the zero value for correction while smoothing methods try to correct all values in the dataset. Some researchers argue that since technical noise affects the whole transcriptome instead of just zero values, smoothing methods are a more reasonable choice [76].

SAVER (Single-cell Analysis Via Expression Recovery) [77] was developed for UMI-based scRNA-seq data, and it recovers expression values by borrowing information across genes and cells using a Bayesian model. DrImpute [78] first identifies similar cells using clustering methods and then performs imputation by averaging the expression values from similar cells. scImpute [79] attempts to reduce imputation noise by identifying the dropout events first and only performing imputation on these values using a gamma-normal mixture model. Inspired by the success of autoencoder in collaborative filtering, AutoImpute [80] imputes scRNA-seq data using an over-complete autoencoder model. MAGIC (Markov Affinity-based Graph Imputation of Cells) [81] imputes scRNA-seq data by borrowing information from neighbouring cells using a Markov affinity-based graph. kNN-smoothing (k-nearest neighbour smoothing) [82] first identifies the k nearest neighbours in a step-wise fashion and then performs smoothing by aggregating gene-specific UMI counts.

Of note, imputation is not a necessary part of the analysis pipeline and is usually not recommended before DE genes analysis, as the artificial changes of the expression values may introduce new noise. Some researchers recently argued that droplet-based scRNA-seq data are not zero-inflated, which eliminates the need for an imputation step [83]. Moreover, some researchers argued that certain terminology (e.g. 'dropout', 'missing data', etc.) should not be used because the high proportion of zero values is mostly caused by biological variance rather than technical noise [84].

## Feature selection

The dimensionality of scRNA-seq data refers to the gene number in the count matrix. Although the gene expression matrix of scRNA-seq data normally has more than 20 000 genes, not all genes are equally important. Some genes (e.g. house-keeping genes) show similar expression values in all cells; hence, most of them are not useful for investigating cellular heterogeneity [85]. Feature selection refers to excluding these uninformative or noise genes and only focusing on the biological variance genes. This process not only reduces the noise that obscures the biological structure but also reduces the size of the count matrix, thereby improving the computational efficiency for downstream analyses. In general, the methods for feature selection can be divided into four categories, which are discussed below [86] (Table 3).

### Highly variable gene-based methods

This type of method attempts to identify the most variable genes for further analysis [87]. In scRNA-seq data analysis, the Fano factor is widely used to measure variability, and it is defined as the ratio between the variance and the mean. Seurat is one of the most popular scRNA-seq methods [32], and it uses highly variable genes (HVGs) to perform feature selection. Of note, although the Fano factor performs well in most circumstances, it is not suitable for identifying rare cell types [32].

### HEG-based methods

This method calculates the highest average expression levels across cells and is sometimes also used for informatics-based gene selection [35].

### Gini-index-based methods

GiniClust [88] takes the characteristics of scRNA-seq data (zero values or low expression levels across most of the samples) into consideration to normalize the Gini index. The novel gene selection methods developed by GiniClust show good performance in analysing rare cell types.

### Dropout-based methods

Previous studies demonstrated that dropout rates are strongly correlated with expression levels [83]. M3Drop [89] fits a Michaelis–Menten function to the relationship between mean expression and dropout rate.

We compared all of the feature selection methods mentioned above with different gene numbers (Figure 3). Concentrating on the first row of Figure 3, which selects 2000 genes using each method, we found that all methods still reveal biological difference in different cell types. Surprisingly, even a random selection of 2000 genes preserved most of the biological variance across cells. This phenomenon also enhanced the feasibility and necessity of feature selection. When the number of selected genes was reduced, most feature selection methods performed worse. However, the dropout-based method M3drop exhibited great performance with only 200 genes, and the visualization result was even better than that for raw data using all the genes. The reason for this result may be that genes with high dropout rates introduce more technical noise than actual biological differences. Therefore, we highly recommend that researchers choose the M3drop method when they only want to select low numbers of genes for downstream analysis.

To further validate this conclusion, we also assessed these feature selection methods using another two scRNA-seq datasets. The results shown in Supplementary Figure 1 were generated by Li *et al.* [90] containing 561 cell samples in eight cell types. The results showed a consistent conclusion with Figure 3 that M3drop performed best in all of four circumstances, especially with 200 genes. We then used a dataset generated by Klein *et al.* [45] containing mouse embryo stem cells to check the influence of feature selection methods on trajectory analysis. The results in Supplementary Figure 2 illustrated that visualization results using M3drop as feature selection method showed clearly dynamic changes (i.e. from day 0 to day 7) in all of four different circumstances, which beats the visualization results using other feature selection methods and the raw data without feature selection step. To sum up, we think that M3drop method is the best feature selection method for both clustering analysis and trajectory analysis.

## Dimension reduction

Dimension reduction refers to methods that are designed to capture the underlying structure of the expression matrix. A low-dimension underlying structure is also known as embedding high-dimension data, such as scRNA-seq data [93, 94]. In other words, dimension reduction attempts to identify the low-dimension biological manifold of scRNA-seq data.

To overcome the 'curse of dimension', dimension reduction is a crucial step for further analysis of scRNA-seq data. scRNA-seq data often contain many highly correlated genes, even after the feature selection step [23]. Because redundant genetic information is not helpful for downstream analysis, dimension reduction methods aim to represent high-dimension data via low-dimension embedding, which is both more computationally efficient and reliable. We divided the dimension reduction methods into the four categories discussed below (Table 4).

### Linear model

Principal component analysis (PCA) [95] is historically the most traditional dimension reduction method for high-dimensional data, and it tries to reduce the dimensionality of the data by identifying the largest amount of variance. However, PCA can only model linear patterns and cannot easily analyse complex scRNA-seq data. Taking the noisy characteristics of scRNA-seq data into account, zero-inflated factor analysis (ZIFA) [96] is also widely used for scRNA-seq data dimension reduction. ZIFA uses a zero-inflated model to account for the high frequency of zero values in scRNA-seq data. The factorial single- cell latent variable model (f-scLVM) [97] creates the embedding by explicitly modelling annotated gene sets; therefore, the reduced dimensions are more interpretable.

**Table 2.** Imputation and smoothing methods for scRNA-seq data

| Method | Description | Availability | Refs |
|---|---|---|---|
| SAVER | Recovers expression values by borrowing information across genes and cells using a Bayesian model | https://github.com/mohuangx/SAVER | [77] |
| DrImpute | Finds similar cells using clustering methods, then performs imputation by averaging the expression values from similar cells | https://github.com/gongx030/DrImpute | [78] |
| scImpute | Reduces the imputation noise by identifying dropout events first and only performing imputation on these values using a gamma-normal mixture model | https://github.com/Vivianstats/scImpute | [79] |
| AutoImpute | Imputes scRNA-seq data with an over-complete autoencoder model | https://github.com/kearnz/autoimpute | [80] |
| MAGIC | Imputes by borrowing information from neighbour cells using a Markov affinity-based graph | https://github.com/KrishnaswamyLab/MAGIC | [81] |
| kNN-smooth | Identifies k nearest neighbours in a step-wise fashion and then performs smoothing by aggregating gene-specific UMI counts. | https://github.com/yanailab/knn-smoothing | [82] |

**Table 3.** Feature selection methods for scRNA-seq data

| Method | Category | Availability | Refs |
|---|---|---|---|
| Seurat | HVG | https://satijalab.org/seurat/ | [32] |
| GiniClust | Gini-index based | https://github.com/lanjiangboston/GiniClust | [88] |
| M3Drop | Dropout-based | https://github.com/tallulandrews/M3Drop | [89] |

**Table 4.** Dimension reduction methods for scRNA-seq data

| Method | Category | Availability | Refs |
|---|---|---|---|
| PCA | Linear | https://github.com/dhamvi01/Principal-Component-Analysis-PCA---Python/blob/master/PCA.ipynb | [95] |
| ZIFA | Linear | https://github.com/epierson9/ZIFA | [96] |
| f-scLVM | Linear | https://github.com/scfurl/f-scLVM | [97] |
| t-SNE | Nonlinear | https://github.com/shivanichander/tSNE/blob/master/Code/tSNE%20Code.ipynb | [99] |
| UMAP | Nonlinear | https://github.com/lmcinnes/umap | [92] |
| DCA | Deep learning | https://github.com/theislab/dca | [103] |
| scVI | Deep learning | https://github.com/shahcompbio/scvis | [98] |
| scBFA | Others | https://github.com/quon-titative-biology/scBFA | [104] |

## Nonlinear model

scRNA-seq data have a nonlinear structure, and nonlinear models have the potential to show better performance [98]. The famous t-SNE (t-distributed stochastic neighbourhood embedding) [99, 100] method has become one of the most popular nonlinear dimension reduction techniques, and it also widely used for scRNA-seq data. However, t-SNE can only deal with local structures, such as trajectory analysis, thus limiting its performance for further study. In contrast, the newly developed method uniform manifold approximation and projection (UMAP) [92] preserves both global and local structures by preserving the high-dimensional topology in low-dimensional space.

## Deep learning-based methods

Recently, deep learning methods, which can also capture the nonlinear features, have become increasingly popular for the analysis of scRNA-seq data [101, 102]. Deep count autoencoder (DCA) [103] uses the denoising autoencoder to denoise the scRNA-seq dataset with a zero-inflated negative binominal model. Single-cell variational inference (scVI) [98] creates a probabilistic model using a neural network to quantify the uncertainty of each gene expression estimate, which preserves both the local and global structures of the data.

## Other methods

Single-cell binary factor analysis (scBFA) [104] aims at reducing the dimensions of large scRNA-seq data by ignoring the quantified counts value. The author assumes that when the sample size is ultra-large, the gene quantification value (i.e. counts) will be too low due to technical noise. Because the low signal to noise ratio decreases the accuracy of the gene quantification, scRFA only employs gene detection to perform dimension reduction.

In summary, linear models could capture the linear patterns and preserve the global structure of the data. However, due to the nonlinear structure of scRNA-seq data, nonlinear model and deep learning-based methods are more suitable for dimension reduction, especially new developed method UMAP, which could capture both global and local structures.

## Conclusions and outlook

As a highly promising technology, scRNA-seq allows researchers to study the heterogeneity of gene expression in individual cells in large cell populations, thus enabling the identification of the dynamics underlying tissue and organism development. Moreover, due to the large scale of sparse and noisy data produced by scRNA-seq, high-efficiency computational tools are essential. In this review, we concentrated on the data processing steps for analysing these noisy sequencing data. Specifically, we reviewed the goals and popular tools for QC, normalization, imputation, feature selection and dimension reduction to apply for analyses of scRNA-seq data. We hope that this study will help researchers choose suitable processing methods for analysing scRNA-seq data.

Unfortunately, because single-cell sequencing technologies are relatively new, standardization is lacking for analyses in this field. More comparisons of processing work need to be performed to evaluate existing methods for addressing the complexities of scRNA-seq data. Moreover, due to the emergence of increasing scRNA-seq datasets, data integration and analysis approaches would be more and more important. Although there have already been many computational tools and workflows for analysing scRNA-seq data, comprehensive comparisons of different tools and best practices workflows are still needed for better utilizing this technology.

### Key Points

- This paper reviewed the goals and various computational analysis tools (quality control, normalization, imputation, feature selection and dimension reduction) for processing single-cell RNA sequencing (scRNA-seq) data.
- We briefly discussed the advantages and disadvantages of the methods introduced in the article.
- Processing steps are extremely significant for downstream analysis of scNRA-seq data, and we provided a comprehensive description and the URLs of processing tools for scRNA-seq data.
- This paper also provided a guide for non-specialists who aim to utilize scRNA-seq technology.

## Supplementary Data

Supplementary data are available online at *Briefings in Bioinformatics*.

## Code availability

Code used to perform the benchmarking for feature selection is available from https://github.com/ZilongZhang44/Review_benchmark.

## Funding

National Key R&D Program of China (2018YFC0910405); National Natural Science Foundation of China (61922020, 61771331, 91935302).

## Conflict of interest

The authors declare that they have no conflicts of interest.

## References

1. Wang Z, Gerstein M, Snyder M. RNA-Seq: a revolutionary tool for transcriptomics. *Nat Rev Genet* 2009;**10**(1):57–63.
2. Ramsköld D, Luo S, Wang Y-C, *et al*. Full-length mRNA-Seq from single-cell levels of RNA and individual circulating tumor cells. *Nat Biotechnol* 2012;**30**(8):777–82.
3. Chen KH, Boettiger AN, Moffitt JR, *et al*. RNA imaging. Spatially resolved, highly multiplexed RNA profiling in single cells. *Science* 2015;**348**(6233):aaa6090.
4. Habib N, Avraham-Davidi I, Basu A, *et al*. Massively parallel single-nucleus RNA-seq with DroNc-seq. *Nat Methods* 2017;**14**(10):955–8.
5. Villani AC, Satija R, Reynolds G, *et al*. Single-cell RNA-seq reveals new types of human blood dendritic cells, monocytes, and progenitors. *Science* 2017;**356**(6335):eaah4573.
6. Kowalczyk MS, Tirosh I, Heckl D, *et al*. Single-cell RNA-seq reveals changes in cell cycle and differentiation programs upon aging of hematopoietic stem cells. *Genome Res* 2015;**25**(12):1860–72.
7. Jaitin DA, Weiner A, Yofe I, *et al*. Dissecting immune circuits by linking CRISPR-pooled screens with single-cell RNA-Seq. *Cell* 2016;**167**(7):1883–1896.e15.
8. Joost S, Zeisel A, Jacob T, *et al*. Single-cell transcriptomics reveals that differentiation and spatial signatures shape epidermal and hair follicle heterogeneity. *Cell Syst* 2016;**3**(3):221–237.e9.
9. Petropoulos S, Edsgärd D, Reinius B, *et al*. Single-cell RNA-Seq reveals lineage and X chromosome dynamics in human preimplantation embryos. *Cell* 2016;**167**(1):285.
10. Faridani OR, Abdullayev I, Hagemann-Jensen M, *et al*. Single-cell sequencing of the small-RNA transcriptome. *Nat Biotechnol* 2016;**34**(12):1264–6.
11. Tirosh I, Venteicher AS, Hebert C, *et al*. Single-cell RNA-seq supports a developmental hierarchy in human oligodendroglioma. *Nature* 2016;**539**(7628):309–13.
12. Avraham R, Haseley N, Brown D, *et al*. Pathogen cell-to-cell variability drives heterogeneity in host immune responses. *Cell* 2015;**162**(6):1309–21.
13. Muraro MJ, Dharmadhikari G, Grün D, *et al*. A single-cell transcriptome atlas of the human pancreas. *Cell Syst* 2016;**3**(4):385–394.e3.
14. Grün D, Kester L, van Oudenaarden A. Validation of noise models for single-cell transcriptomics. *Nat Methods* 2014;**11**(6):637–40.
15. Liu S, Trapnell C. Single-cell transcriptome sequencing: recent advances and remaining challenges. *F1000Res* 2016;**5**:182.
16. Zheng GX, Terry JM, Belgrader P, *et al*. Massively parallel digital transcriptional profiling of single cells. *Nat Commun* 2017;**8**:14049.
17. Stegle O, Teichmann SA, Marioni JC. Computational and analytical challenges in single-cell transcriptomics. *Nat Rev Genet* 2015;**16**(3):133–45.
18. Choi YH, Kim JK. Dissecting cellular heterogeneity using single-cell RNA sequencing. *Mol Cells* 2019;**42**(3):189–99.
19. Seweryn MT, Pietrzak M, Ma Q. Application of information theoretical approaches to assess diversity and similarity in single-cell transcriptomics. *Comput Struct Biotechnol J* 2020;**18**:1830–7.

20. Ma A, Wang C, Chang Y, *et al.* IRIS3: integrated cell-type-specific regulon inference server from single-cell RNA-Seq. *Nucleic Acids Res* 2020;**48**(W1):W275–86.
21. Zhang Y, Wan C, Wang P, *et al.* M3S: a comprehensive model selection for multi-modal single-cell RNA sequencing data. *BMC Bioinformatics* 2019;**20**(Suppl 24):672.
22. Wan C, Chang W, Zhang Y, *et al.* LTMG: a novel statistical modeling of transcriptional expression states in single-cell RNA-Seq data. *Nucleic Acids Res* 2019;**47**(18): e111.
23. Iacono G, Mereu E, Guillaumet-Adkins A, *et al.* bigSCale: an analytical framework for big-scale single-cell data. *Genome Res* 2018;**28**(6):878–90.
24. Cao J, Packer JS, Ramani V, *et al.* Comprehensive single-cell transcriptional profiling of a multicellular organism. *Science* 2017;**357**(6352):661–7.
25. Gawad C, Koh W, Quake SR. Single-cell genome sequencing: current state of the science. *Nat Rev Genet* 2016;**17**(3):175–88.
26. Kiselev VY, Andrews TS, Hemberg M. Challenges in unsupervised clustering of single-cell RNA-seq data. *Nat Rev Genet* 2019;**20**(5):273–82.
27. Hicks SC, Townes FW, Teng M, *et al.* Missing data and technical variability in single-cell RNA-sequencing experiments. *Biostatistics* 2018;**19**(4):562–78.
28. Angerer P, Haghverdi L, Büttner M, *et al.* Destiny: diffusion maps for large-scale single-cell data in R. *Bioinformatics* 2016;**32**(8):1241–3.
29. DeTomaso D, Yosef N. FastProject: a tool for low-dimensional analysis of single-cell RNA-Seq data. *BMC Bioinformatics* 2016;**17**(1):315.
30. Smith T, Heger A, Sudbery I. UMI-tools: modeling sequencing errors in unique molecular identifiers to improve quantification accuracy. *Genome Res* 2017;**27**(3):491–9.
31. Wu Y, Zhang K. Tools for the analysis of high-dimensional single-cell RNA sequencing data. *Nat Rev Nephrol* 2020;**16**(7):408–21.
32. Butler A, Hoffman P, Smibert P, *et al.* Integrating single-cell transcriptomic data across different conditions, technologies, and species. *Nat Biotechnol* 2018;**36**(5):411–20.
33. Lun AT, McCarthy DJ, Marioni JC. A step-by-step workflow for low-level analysis of single-cell RNA-seq data with Bioconductor. *F1000Res* 2016;**5**:2122.
34. Trapnell C, Cacchiarelli D, Grimsby J, *et al.* The dynamics and regulators of cell fate decisions are revealed by pseudotemporal ordering of single cells. *Nat Biotechnol* 2014;**32**(4):381–6.
35. Duo A, Robinson MD, Soneson C. A systematic performance evaluation of clustering methods for single-cell RNA-seq data. *F1000Res* 2018;**7**:1141.
36. Qi R, Ma A, Ma Q, *et al.* Clustering and classification methods for single-cell RNA-sequencing data. *Brief Bioinform* 2020;**21**(4):1196–208.
37. Wang Z, Ding H, Zou Q. Identifying cell types to interpret scRNA-seq data: how, why and more possibilities. *Brief Funct Genomics* 2020;**19**(4):286–291.
38. Tang F, Barbacioru C, Wang Y, *et al.* mRNA-Seq whole-transcriptome analysis of a single cell. *Nat Methods* 2009;**6**(5):377–82.
39. Jaitin DA, Kenigsberg E, Keren-Shaul H, *et al.* Massively parallel single-cell RNA-seq for marker-free decomposition of tissues into cell types. *Science* 2014;**343**(6172):776–9.
40. Sasagawa Y, Nikaido I, Hayashi T, *et al.* Quartz-Seq: a highly reproducible and sensitive single-cell RNA sequencing method, reveals non-genetic gene-expression heterogeneity. *Genome Biol* 2013;**14**(4):R31.
41. Hashimshony T, Wagner F, Sher N, *et al.* CEL-Seq: single-cell RNA-Seq by multiplexed linear amplification. *Cell Rep* 2012;**2**(3):666–73.
42. Hashimshony T, Senderovich N, Avital G, *et al.* CEL-Seq2: sensitive highly-multiplexed single-cell RNA-Seq. *Genome Biol* 2016;**17**:77.
43. Islam S, Kjällquist U, Moliner A, *et al.* Characterization of the single-cell transcriptional landscape by highly multiplex RNA-seq. *Genome Res* 2011;**21**(7):1160–7.
44. Picelli S, Bjorklund AK, Faridani OR, *et al.* Smart-seq2 for sensitive full-length transcriptome profiling in single cells. *Nat Methods* 2013;**10**(11):1096–8.
45. Klein AM, Mazutis L, Akartuna I, *et al.* Droplet barcoding for single-cell transcriptomics applied to embryonic stem cells. *Cell* 2015;**161**(5):1187–201.
46. Macosko EZ, Basu A, Satija R, *et al.* Highly parallel genome-wide expression profiling of individual cells using Nanoliter droplets. *Cell* 2015;**161**(5):1202–14.
47. Haque A, Engel J, Teichmann SA, *et al.* A practical guide to single-cell RNA-sequencing for biomedical research and clinical applications. *Genome Med* 2017;**9**(1):75.
48. Hwang B, Lee JH, Bang D. Single-cell RNA sequencing technologies and bioinformatics pipelines. *Exp Mol Med* 2018;**50**(8):96.
49. Dal Molin A, Di Camillo B. How to design a single-cell RNA-sequencing experiment: pitfalls, challenges and perspectives. *Brief Bioinform* 2019;**20**(4):1384–94.
50. Islam S, Zeisel A, Joost S, *et al.* Quantitative single-cell RNA-seq with unique molecular identifiers. *Nat Methods* 2014;**11**(2):163–6.
51. Fan HC, Fu GK, Fodor SP. Expression profiling. Combinatorial labeling of single cells for gene expression cytometry. *Science* 2015;**347**(6222):1258367.
52. Dillies MA, Rau A, Aubert J, *et al.* A comprehensive evaluation of normalization methods for Illumina high-throughput RNA sequencing data analysis. *Brief Bioinform* 2013;**14**(6):671–83.
53. Lun AT, Bach K, Marioni JC. Pooling across cells to normalize single-cell RNA sequencing data with many zero counts. *Genome Biol* 2016;**17**:75.
54. Andrews S. FASTQC. A quality control tool for high throughput sequence data. 2010 [Online] https://www.bioinformatics.babraham.ac.uk/projects/fastqc/.
55. Martin M. Cutadapt removes adapter sequences from high-throughput sequencing reads. *EMBnetjournal* 2011;**17**(1):3.
56. Srivastava A, Malik L, Smith T, *et al.* Alevin efficiently estimates accurate gene abundances from dscRNA-seq data. *Genome Biol* 2019;**20**(1):65.
57. Tambe A, Pachter L. Barcode identification for single cell genomics. *BMC Bioinformatics* 2019;**20**(1):32.
58. Zorita E, Cuscó P, Filion GJ. Starcode: sequence clustering based on all-pairs search. *Bioinformatics* 2015;**31**(12):1913–9.
59. Dobin A, Davis CA, Schlesinger F, *et al.* STAR: ultrafast universal RNA-seq aligner. *Bioinformatics* 2013;**29**(1):15–21.
60. Kim D, Pertea G, Trapnell C, *et al.* TopHat2: accurate alignment of transcriptomes in the presence of insertions, deletions and gene fusions. *Genome Biol* 2013;**14**(4):R36.

61. Kim D, Langmead B, Salzberg SL. HISAT: a fast spliced aligner with low memory requirements. *Nat Methods* 2015;**12**(4):357–60.
62. Ash Blibaum JW. Alexander Dobin, STARsolo: single-cell RNA-seq analyses beyond gene expression[version 1; not peer reviewed]. *F1000Research* 2019, 1896;**8**.
63. Jiang L, Schlesinger F, Davis CA, *et al.* Synthetic spike-in standards for RNA-seq experiments. *Genome Res* 2011;**21**(9):1543–51.
64. Amezquita RA, Lun ATL, Becht E, *et al.* Orchestrating single-cell analysis with Bioconductor. *Nat Methods* 2020;**17**(2):137–45.
65. Li B, Ruotti V, Stewart RM, *et al.* RNA-Seq gene expression estimation with read mapping uncertainty. *Bioinformatics* 2010;**26**(4):493–500.
66. Lee S, Seo CH, Lim B, *et al.* Accurate quantification of transcriptome from RNA-Seq data by effective length normalization. *Nucleic Acids Res* 2010;**39**(2):e9–9.
67. Anders S, Huber W. Differential expression analysis for sequence count data. *Genome Biol* 2010;**11**(10):R106.
68. Love MI, Huber W, Anders S. Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. *Genome Biol* 2014;**15**(12):550.
69. Vallejos CA, Marioni JC, Richardson S. BASiCS: Bayesian analysis of single-cell sequencing data. *PLoS Comput Biol* 2015;**11**(6):e1004333.
70. Ding B, Zheng L, Zhu Y, *et al.* Normalization and noise reduction for single cell RNA-seq experiments. *Bioinformatics* 2015;**31**(13):2225–7.
71. Yip SH, Wang P, Kocher JA, *et al.* Linnorm: improved statistical analysis for single cell RNA-seq expression data. *Nucleic Acids Res* 2017;**45**(22):e179.
72. Bacher R, Chu LF, Leng N, *et al.* SCnorm: robust normalization of single-cell RNA-seq data. *Nat Methods* 2017;**14**(6):584–6.
73. Qiu X, Hill A, Packer J, *et al.* Single-cell mRNA quantification and differential analysis with census. *Nat Methods* 2017;**14**(3):309–15.
74. Chen G, Ning B, Shi T. Single-cell RNA-Seq technologies and related computational data analysis. *Front Genet* 2019;**10**:317.
75. Bacher R, Kendziorski C. Design and computational analysis of single-cell RNA-sequencing experiments. *Genome Biol* 2016;**17**:63.
76. Ronen J, Akalin A. netSmooth: network-smoothing based imputation for single cell RNA-seq. *F1000Res* 2018;**7**:8.
77. Huang M, Wang J, Torre E, *et al.* SAVER: gene expression recovery for single-cell RNA sequencing. *Nat Methods* 2018;**15**(7):539–42.
78. Gong W, Kwak IY, Pota P, *et al.* DrImpute: imputing dropout events in single cell RNA sequencing data. *BMC Bioinformatics* 2018;**19**(1):220.
79. Li WV, Li JJ. An accurate and robust imputation method scImpute for single-cell RNA-seq data. *Nat Commun* 2018;**9**(1):997.
80. Talwar D, Mongia A, Sengupta D, *et al.* AutoImpute: autoencoder based imputation of single-cell RNA-seq data. *Sci Rep* 2018;**8**(1):16329.
81. van Dijk D, Sharma R, Nainys J, *et al.* Recovering gene interactions from single-cell data using data diffusion. *Cell* 2018;**174**(3):716–729 e27.
82. Wagner F, Yan Y, Yanai I. K-nearest neighbor smoothing for high-throughput single-cell RNA-Seq data. *bioRxiv* 2018;217737.
83. Svensson V. Droplet scRNA-seq is not zero-inflated. *Nat Biotechnol* 2020;**38**(2):147–50.
84. Sarkar A, Stephens M. Separating measurement and expression models clarifies confusion in single cell RNA-seq analysis. *bioRxiv* 2020; 2020.04.07.030007.
85. Ramskold D, Wang ET, Burge CB, *et al.* An abundance of ubiquitously expressed genes revealed by tissue transcriptome sequence data. *PLoS Comput Biol* 2009;**5**(12): e1000598.
86. Townes FW, Hicks SC, Aryee MJ, *et al.* Feature selection and dimension reduction for single-cell RNA-Seq based on a multinomial model. *Genome Biol* 2019;**20**(1):295.
87. Love MI, Anders S, Kim V, *et al.* RNA-Seq workflow: gene-level exploratory analysis and differential expression. *F1000Res* 2015;**4**:1070.
88. Jiang L, Chen H, Pinello L, *et al.* GiniClust: detecting rare cell types from single-cell gene expression data with Gini index. *Genome Biol* 2016;**17**(1):144.
89. Andrews TS, Hemberg M. M3Drop: dropout-based feature selection for scRNASeq. *Bioinformatics* 2019;**35**(16):2865–7.
90. Li H, Courtois ET, Sengupta D, *et al.* Reference component analysis of single-cell transcriptomes elucidates cellular heterogeneity in human colorectal tumors. *Nat Genet* 2017;**49**(5):708–18.
91. Deng Q, Ramsköld D, Reinius B, *et al.* Single-cell RNA-seq reveals dynamic, random monoallelic gene expression in mammalian cells. *Science* 2014;**343**(6167):193–6.
92. McInnes L, Healy J, Melville J. UMAP: uniform manifold approximation and projection for dimension reduction. *arXiv e-prints* 2018;1802.03426.
93. Luecken MD, Theis FJ. Current best practices in single-cell RNA-seq analysis: a tutorial. *Mol Syst Biol* 2019;**15**(6):e8746.
94. Heimberg G, Bhatnagar R, El-Samad H, *et al.* Low dimensionality in gene expression data enables the accurate extraction of transcriptional programs from shallow sequencing. *Cell Syst* 2016;**2**(4):239–50.
95. Jolliffe I. In: Lovric M (ed). *Principal Component Analysis, in International Encyclopedia of Statistical Science*. Berlin, Heidelberg: Springer Berlin Heidelberg, 2011, 1094–6.
96. Pierson E, Yau C. ZIFA: dimensionality reduction for zero-inflated single-cell gene expression analysis. *Genome Biol* 2015;**16**:241.
97. Buettner F, Pratanwanich N, McCarthy DJ, *et al.* F-scLVM: scalable and versatile factor analysis for single-cell RNA-seq. *Genome Biol* 2017;**18**(1):212.
98. Lopez R, Regier J, Cole MB, *et al.* Deep generative modeling for single-cell transcriptomics. *Nat Methods* 2018;**15**(12):1053–8.
99. van der Maaten LJP, Hinton GE. Visualizing high-dimensional data using t-SNE. *J. Mach. Learn. Res.* 2008:**9**:2579–2605.
100. Aliverti E, Tilson JL, Filer DL, *et al.* Projected t-SNE for batch correction. *Bioinformatics* 2020;**36**(11):3522–7.
101. Eraslan G, Avsec Z, Gagneur J, *et al.* Deep learning: new computational modelling techniques for genomics. *Nat Rev Genet* 2019;**20**(7):389–403.
102. Min S, Lee B, Yoon SJBIB. Deep learning in bioinformatics. *Brief. Bioinform.* 2017;**18**:851–69.
103. Eraslan G, Simon LM, Mircea M, *et al.* Single-cell RNA-seq denoising using a deep count autoencoder. *Nat Commun* 2019;**10**(1):390.
104. Li R, Quon G. scBFA: modeling detection patterns to mitigate technical noise in large-scale single-cell genomics data. *Genome Biol* 2019;**20**(1):193.