Briefings in Bioinformatics, 22(5), 2021, 1–11

https://doi.org/10.1093/bib/bbab105
Method Review

Critical downstream analysis steps for single-cell RNA sequencing data

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

Corresponding authors.Chunyu Wang, Institute of Computer Science and Technology, Harbin Institute of Technology, 92 Xidazhi Street. Harbin City, 150001, China. E-mail: chunyu@hit.edu.cn; Quan Zou, Institute of Fundamental and Frontier Sciences, University of Electronic Science and Technology of China, 4 North Jianshe Road, Chengdu City, 610054, China, E-mail: zouquan@nclab.net.

Abstract

Single-cell RNA sequencing (scRNA-seq) has enabled us to study biological questions at the single-cell level. Currently, many analysis tools are available to better utilize these relatively noisy data. In this review, we summarize the most widely used methods for critical downstream analysis steps (i.e. clustering, trajectory inference, cell-type annotation and integrating datasets). The advantages and limitations are comprehensively discussed, and we provide suggestions for choosing proper methods in different situations. We hope this paper will be useful for scRNA-seq data analysts and bioinformatics tool developers.

Key words: single-cell RNA sequencing; clustering; trajectory inference; cell type annotation; integrating datasets

INTRODUCTION

As a fast-growing technology, single-cell RNA sequencing (scRNA-seq) offers the detection of gene expression at the single-cell level, which allows researchers to perform research at the single-cell level instead of at the organism level [1–3]. Over the last decade, scRNA-seq technology has shown great power in uncovering unexpected biological discoveries, revealing new cell types and investigating disease development [4–10]. However, due to the low amount of mRNA in a single cell and the technical noise introduced during different protocols, scRNA-seq data are generally noisy [11, 12]. Therefore, scRNA-seq data usually have the characteristic of 'high-dimensional' and 'large numbers of zeros' [13]. Staring from a sparse and noisy counts or unique molecular identifiers (UMIs) matrix, distinguishing the technical noise from true biological difference is crucial for revealing interesting information [14, 15]. To better utilize this powerful technology, several user-friendly analysis pipelines (e.g. Seurat [16, 17] and SCANPY [18]) have been developed for analysis and achieve relatively good performance; however, due to the complexity of scRNA-seq data, alternative tools may help us obtain interesting findings. Moreover, the explosion in data analysis tools developed for scRNA-seq data has made it hard for users to choose the proper workflow for their study [19–21].

In our previous work, we reviewed the goals and approaches for processing scRNA-seq data from raw data [22]. In this review, we summarize popular methods for critical steps of downstream analyses (i.e. clustering, trajectory inference, cell-type annotation and integrating datasets) for scRNA-seq data. Specifically, we focus on the methods that are either most widely used or

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, machine learning and data mining.
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.
Chen Lin is a Professor in Xiamen University. She received her PhD degree from Fudan University in China. Her research interests include social network analysis, bioinformatics and data mining.
Lingling Zhao is an Assistant Professor in the Harbin Institute of Technology. She received her PhD degree from the Harbin Institute of Technology in China. Her research interests include bioinformatics and machine learning.
Chunyu Wang is an Associate Professor in the Harbin Institute of Technology. He received his PhD degree from the Harbin Institute of Technology in China. His research interest is bioinformatics.
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, P.R. 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: 5 January 2021; **Received (in revised form):** 20 February 2021

show great performance, as demonstrated in related benchmark papers. In each following section, we first discuss the goals and precautions for each step overall and then elaborate on the advantages and limitations of different popular methods. Finally, we provide suggestions for how to choose proper methods in different situations.

PIPELINE FOR ANALYSING scRNA-seq TECHNOLOGY

The whole pipeline of scRNA-seq technology analysis is shown in Figure 1. Owing to the noisy character of scRNA-seq data, data processing steps are generally important; however, elaboration for this part of the analysis is beyond the scope of this review, and we recommend that readers refer to these articles for further information [22–25]. In this review, we focus on downstream analysis, including clustering, trajectory inference, cell-type annotation and dataset integration. As shown in Figure 1, we divide popular methods into several categories for better discussion. In each section, we provide our perspectives for choosing proper tools in different situations.

QUALITY CONTROL AND DIFFERENTIAL GENE EXPRESSION ANALYSIS

Single-cell RNA sequencing data are well-known for their characteristic of sparse and noisy. Generally, this may lead to two problems for downstream analysis. Firstly, as most of the low-quality cells have extremely low expression values across all genes, they have a high chance to form a cluster during downstream analysis due to their similar expression pattern. Therefore, this can give us an erroneous indication that they belong to a new cell type or subtype. Secondly, cells with low expression values across all genes are very sensitive to noise, which can lead to substantial impacts on dimension reduction [22]. Consequently, performing quality control before downstream analysis is essential for analysing single-cell data, and we recommend the readers to refer to the cited references at the end of this sentence for more details [23, 25, 26].

Differential gene expression analysis is a common task for scRNA-seq data. Although approaches developed for bulk-cell, such as DESeq2 [27], edgeR [28] and limma [29], are still in current use, scRNA-seq data have many different characteristics (e.g. higher level of noise and larger amount of zero counts) requiring new developed methods. The most popular tool Seurat utilizes a popular nonparametric test named Wilcoxon rank sum test as default method to perform differential expression tests by 'FindMarkers' function [16, 30]. As the first differential expression analysis designed for single-cell data, single-cell differential expression uses a mixture probabilistic model for expression gene values, specifically, a negative binomial distribution for normal genes and a passion distribution for dropout genes [31]. Taking dropouts and binodal expression patterns into account, model-based analysis of single-cell transcriptomic proposed a two-part generalized linear model while modelling changes of gene expression upon technical noise [32]. Another popular tool named beta-Poisson model for the single-cell gene expression data utilizes beta-Poisson mixture model for capturing the bimodality of single-cell data [33]. Although many approaches exit for differential expression analysis, a comprehensive benchmark conducted by Soneson *et al.* [34] found that most of these popular methods showed similar results, and even the bulk RNA-seq analysis approaches generally perform well.

CLUSTERING

Generally, downstream analysis of scRNA-seq data starts with a clustering step. The main goal of clustering is to find discrete cell types using similar expression patterns across different cells (of note, the clusters are different from the cell types, and we further discuss this question in the following chapter) [19]. This is carried out computationally by unsupervised learning without using any prior knowledge; however, the trickiest question in this step is to decide the best number of clusters. Strictly speaking, the answer would be 'there is no best cluster number', and we can even say that 'there is no best cell type number' for specific scRNA-seq data. The reason is that researchers may just want to get to know the major cell types in some circumstances, while they may be interested in the subtypes or even *de novo* subtypes in other circumstances.

The most popular clustering methods for scRNA-seq data are shown in Table 1. The clustering algorithm chosen by Seurat is graph-based clustering. The greatest advantage of graph-based clustering over other clustering methods is scalability, which is significantly important for scRNA-seq data analysis due to the growth of cell numbers in recent years. By treating each cell as a node, a graph can be built by finding the k-nearest neighbour for each node. The edges in the graph represent the similarity relationships between the cells. The main drawback of this clustering algorithm used in Seurat is that the result is relatively sensitive to the parameter (resolution), and the default algorithm (Louvain method [35]) may generate false clusters in some cases. Similarly, another popular R package, scanpy [18], also utilizes the Louvain algorithm for clustering. However, both Seurat and scanpy perform poorly when dealing with small datasets. As a terrific graph-based denoising method, MAGIC (Markov Affinity-based Graph Imputation of Cells) is also often utilized for data visualization and clustering [36, 37]. MAGIC learns the manifold of high dimensional data and uses graphs for smoothing. As the easiest and most popular clustering method, k-means [38] clustering is famous for its fast computational speed. However, determining the number of k remains a challenge to reveal biologically meaningful clusters. To overcome these drawbacks, SC3 [39] performs k-means clustering several times with different initial values and obtains the consensus as the final result. In addition, k-means clustering is extremely sensitive to outlier cells. Therefore, low-quality data and doublets should be thoroughly excluded before performing k-means clustering. Based on this problem, by detecting outliers first before performing k-means clustering, RaceID [9] shows great performance in rare cell-type identification. Hierarchical clustering tries to build a hierarchy of clusters using either a 'bottom-up' or a 'top-down' approach. One advantage of this method compared with graph-based and k-means clustering is that it is deterministic. Nevertheless, because of the high computational complexity, hierarchical clustering could only be used to analyse very small scRNA-seq datasets. CIDR [40] takes dropout events into consideration and performs hierarchical clustering after dimensionality reduction by principal coordinates analysis (PCoA). TSCNA (Time reconstruction in Single-Cell RNA-seq ANalysis) is a model-based clustering approach which utilizes a mixture of multivariate normal distributions of single-cell data, and the posterior probability is calculated for assigning cells to clusters [41]. Another type of clustering is based on regulation network, which is chosen by popular tool SCENIC [42]. SCENIC has shown its great performance for transcription factor analysis, which is especially useful for facilitating researchers to find key genes for diseases.

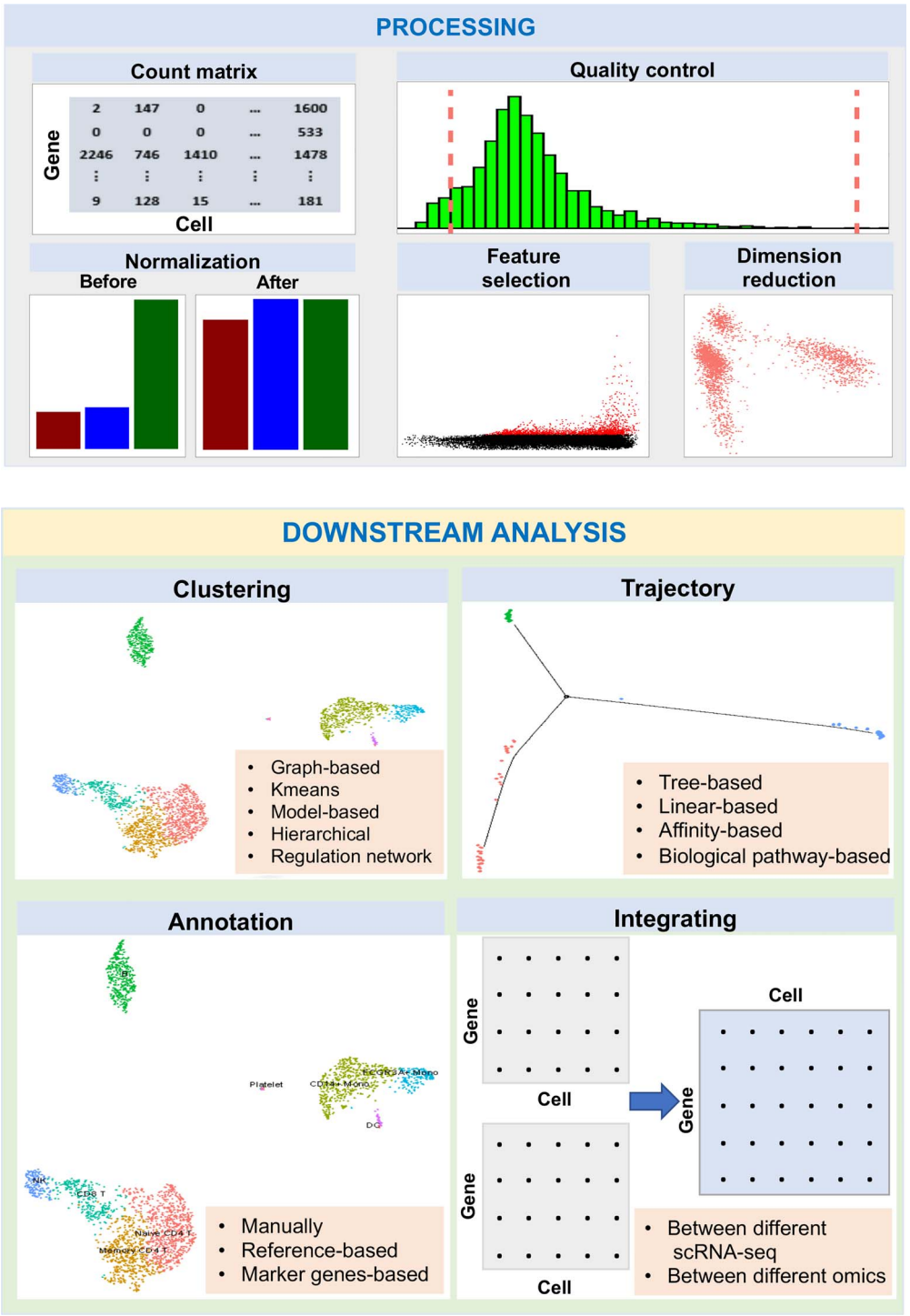

Figure 1. General pipeline of single-cell RNA sequencing analysis. The analysis generally started from count or UMIs matrix. Due to the noisy characteristics of single-cell data, the matrix firstly needs to be processed by quality control, normalization, feature selection and dimension reduction.

As clustering is extremely important for the further analysis, we conducted a benchmark work including the approaches mentioned above. We collected five high-quality single-cell datasets containing both human and mouse samples (i.e. Darmanis data [43], Kolodziejczyk data [44], Li data [45], Deng data [46] and Goolam data [47]). Cell-type labels provided by the authors were treated as ground truth, and then, we calculated the adjusted rand index (ARI) between the clustering result from each method and the ground truth [48]. ARI penalizes both false positive and false negative decisions, where a larger ARI value means a higher agreement between two clusters. The maximum ARI value is 1, and the minimal value is 0 in the case of random clusters.

The comparison results were shown in Figure 2 and Supplementary Table 1, see Supplementary Data available online at http://bib.oxfordjournals.org/. Figure 2 showed that overall SC3 showed highest score of ARI, followed by the widely

Table 1. Clustering methods for single-cell RNA sequencing data

Method	Category	Description	Availability	Refs
Seurat	Graph-based	As the most popular R package for single-cell sequencing data analysis pipeline, Seurat utilized Louvain method for clustering.	https://satijalab.org/seurat/	[16, 17]
Scanpy	Graph-based	A python package also widely used for single-cell data analysis, which utilizes Louvain algorithm for clustering.	https://scanpy.readthedocs.io/en/stable/	[18]
SC3	k-means	Performs k-means clustering several times with different initial values and obtains the consensus as the final result. In addition, k-means clustering is extremely sensitive to outlier cells.	http://bioconductor.org/packages/release/bioc/html/SC3.html	[39]
RaceID	k-means	Shows great performance in rare cell-type identification. Hierarchical clustering tries to build a hierarchy of clusters using either a 'bottom-up' or a 'top-down' approach.	https://github.com/dgrun/RaceID	[9]
TSCNA	Model-based	Utilizes a mixture of multivariate normal distributions of single-cell data, and the posterior probability is calculated for assigning cells to clusters	https://github.com/zji90/TSCAN	[41]
CIDR	Hierarchical	Takes dropout events into consideration and performs hierarchical clustering after dimensionality reduction by PCoA.	https://github.com/VCCRI/CIDR	[40]
SCENIC	Regulation network	Has shown its great performance for transcription factor analysis, which is especially useful for facilitating researchers to find key genes for diseases.	https://github.com/aertslab/SCENIC	[42]

used package Seurat. We found that our results are consistent with studies conducted by Duò *et al.* [37], who also got the conclusion that SC3 and Seurat showed the most favorable results.

In summary, as the first step for downstream analysis of scRNA-seq data, clustering is significantly important for uncovering biologically meaningful information from gene expression data. Generally, no clustering method can perform well in all circumstances [49]. We recommend that biological users employ SC3 and the Louvain method integrated in Seurat or scanpy in most situations while choosing hierarchical clustering methods (e.g. CIDR and BackSPIN) for small datasets.

TRAJECTORY INFERENCE

Unlike the underlying discrete status assumed in clustering, in some situations, the cellular state can be treated as a continuum of dynamic changes [25]. Trajectory inference can provide tremendous benefits for understanding the cell cycle, cell-type differentiation and cell activation [50, 51]. The 'pseudotime' generally used for trajectory analysis is just a number that describes the position of a cell thorough the trajectory, while branched trajectories consist of multiple pseudotimes thorough different trajectories. Ideally, the trajectory can be interpreted as cellular states one after another. Marker-based methods utilized in conventional bulk RNA sequencing are insufficient for cell-level resolution [52]; consequently, approaches tailored for single-cell data are needed.

To date, more than 70 tools are available for trajectory analysis of scRNA-seq data [53]; hence, choosing the proper method is challenging (Table 2). Many trajectory methods are designed based on dimension reduction. Monocle [54] tries to build a minimum spanning tree (MST) based on the reduced dimensions (independent component analysis) of gene expression data and then finds the longest path as 'pseudotime' (i.e. the relative position of a cell in the trajectory). Partition-based graph abstraction (PAGA) [55] is a tree-based algorithm that estimates the connectivity between clusters by generating a graph-like map where each node represents one cluster and each edge represents a neighbourhood relation. PAGA showed great performance on dealing with the so-called 'short circuiting' with a relatively fast running speed [53]. Slingshot [56] is also a tree-based method such as Monocle; besides utilizing a cluster-based MST like other tree-based methods, Slingshot then fits smooth branching curves to global lineages to obtain cell-level pseudotime. Similar to Slingshot, pseudotime reconstruction in single-cell RNA-seq ANalysis (TSCAN) [41] employs a cluster-based MST approach for ordering, while the MST is built on cluster centroids instead of individual cells. Monocle 3 [57] uses the dimension reduction method UMAP [58] to initialize trajectory inference and then utilizes graph theory to learn a principal in a dynamic process to refine the trajectory inference result. SCOEPIUS [59] iteratively refines the shortest path through cluster centroids for the low-dimensionality data (the dimensions are reduced by multidimensional scaling) and then identifies the key genes using the prediction score of ordering by the random forest algorithm [60]. Tempora [61] is a newly presented method for novel cell trajectory inference that shows great performance, while prior information for time-series data is available. PHATE (Potential of Heat-diffusion for Affinity-based Trajectory Embedding) uses an affinity-preserving embedding method for visualizing high-dimensional biological data, which is famous for both keeping the global and local information of the input data [62]. Another interesting method using RNA abundance as an indicator needs to be mentioned is RNA velocity [63]. By calculating the proportion of unspliced mRNA and spliced mRNA in each single cell, the state of mRNA change (i.e. the time derivative of the gene expression state) for each cell could be estimated. The advantage

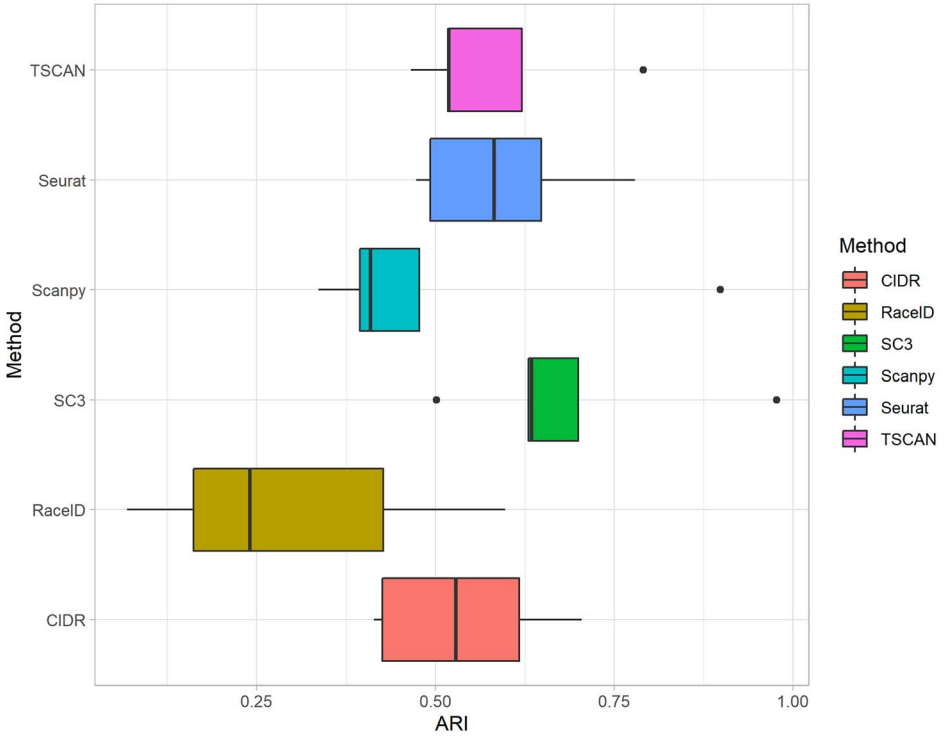

Figure 2. Boxplot of ARI score for different clustering methods.

of RNA velocity is that it treats the state of each single cell differentiation as a vector, which is theoretically more accurate than just estimating the starting point and ending point of cell differentiation [64].

Trajectory inference has shown great power in taking full advantage of single-cell RNA-seq technology to uncover gradual changes in cell differential processes and could also be used to reveal the mechanisms of the cell cycle and the immune activation system [12, 65, 66]. A comprehensive benchmark presented by Saelens *et al.* [53] has been treated as a guideline for choosing trajectory inference tools for scRNA-seq data. According to their work, although Monocle is the most classical and most widely used, the overall best performance (considering accuracy, scalability, stability and usability) is provided by PAGA, Slingshot and SCOEPIUS.

CELL-TYPE ANNOTATION

Clustering can group cells into different clusters based on computational algorithms; yet, the clusters have no biological insights. Identifying cell types for each cluster is one of the most challenging and time-consuming steps in the pipeline of scRNA-seq data analysis. As we mentioned earlier, even the concept of 'cell type' is controversial since many cell types can be further divided into more subtypes. Hence, the correct answer to cell-type annotation also changes, as researchers want to interpret their data. The current most widely used method for cell-type annotation is annotating manually by finding the marker genes in the dataset and then matching them to canonical cell-type markers. Obviously, this method is time-consuming and relies heavily on prior biological knowledge.

To overcome these inconveniences, various automatic annotation tools have been developed (Table 3). Generally, these tools can be grouped into two categories: reference-based and marker-based. SingleR [67] is one of the most widely used reference-based tools for annotation. Essentially, SingleR conducts the labelling process in three steps. First, SingleR downloads a reference scRNA-seq dataset with cell-type labels and identifies marker genes based on the fold change. Then, it calculates Spearman's correlation coefficient between cells to be identified and cells with labels. Last, SingleR eliminates the cell type with the least correlation score and iteratively repeats the previous steps until only one cell type remains. scMAP [68] first utilizes the feature selection M3Drop [69] method to find highly variable genes (HVGs) and then calculates the similarities between new cells and reference cells. scPred [70] leverages orthogonalization and dimensionality reduction of expression values to generate a support-vector machine (SVM) model for prediction [71]. Another reference-based method, scMatch [72], instead of annotating after clustering, directly annotates single cells by identifying the best match in the reference datasets. SingleCellNet is another reference-based method, which utilizes multi-class random forest algorithm for cell-type annotation [73]. In addition, the classical classification method SVM also showed great performance for cell identification according to the terrific benchmark study by Abdelaal *et al.* [74, 75]. The other category of automatic tools is proposed based on prior knowledge of marker genes. Cellassign [76] utilizes prior cell-type marker genes to build a probabilistic model for annotation. Similarly, SCSA [77] combines the prior knowledge of marker genes and user-defined information to annotate cell types based on an annotation model.

Due to the significant roles of canonical marker genes play in cell annotation, it is worth summarizing useful marker genes resources. CellMarker database was developed by Zhang *et al.* in 2019, which has become the most popular marker gene database for cell annotation [78]. Another famous single-cell

Table 2. Trajectory inference methods for single-cell RNA sequencing data

Method	Category	Description	Availability	Refs
Monocle	Tree-based	Tries to build a MST based on the reduced dimensions (independent component analysis) of gene expression data and then finds the longest path as 'pseudotime'.	http://cole-trapnell-lab.github.io/monocle-release/	[54, 57]
PAGA	Tree-based	A graph-based algorithm that estimates the connectivity between clusters by generating a graph-like map where each node represents one cluster and each edge represents a neighbourhood relation.	https://github.com/theislab/paga	[55]
Slingshot	Tree-based	Besides utilizing a cluster-based MST like other tree-based methods, Slingshot then fits smooth branching curves to global lineages to obtain cell-level pseudotime.	https://github.com/kstreet13/slingshot	[56]
TSCAN	Linear-based	Employs a cluster-based MST approach for ordering, while the MST is built on cluster centroids instead of individual cells.	https://github.com/zji90/TSCAN	[41]
SCOEPIUS	Linear-based	Iteratively refines the shortest path through cluster centroids for the low-dimensionality data and then identifies the key genes using the prediction score of ordering by the random forest algorithm	https://github.com/rcannood/SCORPIUS	[59]
Tempora	Biological pathway-based	A newly presented method for novel cell trajectory inference that shows great performance, while prior information for time-series data is available.	https://github.com/BaderLab/Tempora	[61]
PHATE	Affinity-based	Uses an affinity-preserving embedding method for visualizing high-dimensional biological data, which is famous for both keeping the global and local information of the input data.	https://github.com/KrishnaswamyLab/PHATE	[62]

Table 3. Cell-type annotation methods for single-cell RNA sequencing data

Method	Category	Description	Availability	Refs
SinlgeR	Reference-based	Downloads a reference scRNA-seq dataset with cell-type labels and identifies marker genes based on the fold change. Then, it calculates Spearman's correlation coefficient between cells to be identified and cells with labels. Last, SingleR eliminates the cell type with the least correlation score and iteratively repeats the previous steps until only one cell type remains.	https://github.com/dviraran/SingleR	[67]
scMAP	Reference-based	Utilizes the feature selection M3Drop method to find HVGs and then calculates the similarities between new cells and reference cells.	https://github.com/hemberg-lab/scmap	[68]
scPred	Reference-based	Leverages orthogonalization and dimensionality reduction of expression values to generate an SVM model for prediction.	https://github.com/powellgenomicslab/scPred	[70]
scMatch	Reference-based	Directly annotates single cells by identifying the best match in the reference datasets.	https://github.com/asrhou/scMatch	[72]
SingleCellNet	Reference-based	Utilizes multi-class random forest algorithm for cell-type annotation.	https://github.com/pcahan1/singleCellNet	[73]
Cellassign	Marker-based	Utilizes prior cell-type marker genes to build a probabilistic model for annotation.	https://github.com/Irrationone/cellassign	[76]
SCSA	Marker-based	Combines the prior knowledge of marker genes and user-defined information to annotate cell types based on an annotation model.	https://github.com/bioinfo-ibms-pumc/SCSA	[77]

sequencing database, PanglaoDB, also provides a large amount of maker genes information in addition to single-cell dataset [79]. Han *et al.* [80] constructed a comprehensive single-cell atlas named human cell landscape (HCL) provided useful resource including marker genes and data reference for human biology. The well-known biological research institution, Broad Institute, developed Single Cell Portal (SCP) which could be used for cell annotation based on searching genes. Moreover, the European Bioinformatics Institute (EMBL-EBI) also provided similar single cell expression atlas like SCP for cell annotation. All the data resources listed above and their official websites were shown in Table 4.

Table 4. Popular data resources for cell-type annotation

Database	Websites
CellMarker	http://biocc.hrbmu.edu.cn/CellMarker/
PanglaoDB	https://panglaodb.se/
HCL	http://bis.zju.edu.cn/HCL/
Single Cell Portal	https://singlecell.broadinstitute.org/single_cell
EMBL-EBI	https://www.ebi.ac.uk/gxa/sc/experiments

Taken together, although many automatic annotation tools have already been developed, there is still an urgent need for new tools to be developed with better accuracy. Manual cell-type annotation has become the most time-consuming step in single-cell data analysis. Although various automatic annotation tools have been developed, the results obtained from both reference- and marker-based methods are still not quite accurate [74]. Therefore, to guarantee the accuracy of cell-type annotation, manual assignment is still necessary. More importantly, no matter reference- and marker-based automatic annotation tools require priori biological knowledge, which would be impossible for identifying new cell types. Consequently, for now, we would recommend that single-cell data analysts utilize automatic tools, such as SingleR and Cellassign, as auxiliary tools to save some time while checking the results by manual confirmation.

INTEGRATING DATASETS

With the rapid development of scRNA-seq technology, a large amount of scRNA-seq data has accumulated [81]. However, how to integrate these data across different batches or different experiments is a complex question due to batch effects [82, 83]. Batch effects generally refer to the gene expression values in one batch differing systematically from gene expression values in another batch, which may be caused by several factors, such as different cell dissociation protocols, reagent quality, operators or technology platforms [84–86]. Linear computational tools developed for bulk RNA sequencing and microarray data such as sva [87] and limma [29] have been utilized for batch correction of scRNA data; however, these tools are based on the assumption that different batches have identical cell compositions, which is incorrect in most scRNA-seq data due to 'dropout events' or amplification bias [26]. In addition to defining the transcriptome, single-cell sequencing technologies can also define the epigenome [88–91], proteome [92–95] and genome [96–98]. Integrative analysis of multimodal data (i.e. single-cell transcriptomics, epigenomics, proteomics and genomics sequencing data) enables a comprehensive understanding of cellular and regulatory processes [81]. For example, CITE-seq [99] can simultaneously obtain transcriptional modules and surface protein markers within a single cell. In this case, surface protein markers could be complementary information for distinguishing complex cell groups (e.g. memory and regulatory subsets of T cell groups).

As such, many methods tailed for scRNA-seq have been developed for integration (Table 5). Mutual nearest neighbours (MNNs) correction [100] identifies MNN pairs between datasets and uses this information to compute the batch effect. The MNN algorithm assumes that the batch effect is orthogonal to the biological difference, which makes the MNN model more reasonable than the linear model, as in high-dimension space, most random vectors are orthogonal. However, finding the neighbours in such high-dimension space requires huge computational resources. Moreover, another limitation of this method is that it is hard to identify MNNs with strong batch effects in the scRNA-seq data. Consequently, FastMNN [101] is often used in practice due to its shorter running time, which is achieved by performing dimension reduction before finding neighbours. Another widely known method, MultiCCA [16], which employs canonical correlation analysis (CCA) [102], is implemented in the popular R toolkit Seurat V2. Harmony [103] projects high-dimensional data into a subspace by PCA and iteratively corrects the batch effect by maximizing the diversity of batches of similar cells within each cluster. Accompanied by the rapid development of multimodal profiling technologies, there is an urgent need for computational approaches to integrate these data. In Seurat V3 [17], CCA is utilized to generate a subspace, and MNNs are then calculated as 'anchors' to correct the batch effect. In the latest version of Seurat (V4) [104], weighted nearest neighbour (WNN) was proposed by utilizing unsupervised strategy to define cellular state with the weighted combination of each modalities. LIGER [105] employs integrative nonnegative matrix factorization to obtain shared and dataset-specific factors, which are further utilized for clustering and identifying cells by matching to a reference dataset.

In summary, with the rapid development of protocols and lower cost, single-cell technology has been preferred in many laboratories in multiomics studies. Consequently, integrating these datasets seems promising. By gathering the results of several benchmark works for single-cell data integration [81, 84, 106, 107], we conclude that, in addition to the popular R package Seurat, the newly developed tools Harmony and LIGER are also alterative choices that show great results during data integration.

OUTLOOK

In addition to the maturity of scRNA-seq technology, sequencing protocols for the epigenome, proteome and genome at single-cell resolution are developing rapidly. Since obtaining multi-omics information could provide us with a much more comprehensive view of the cell, further technological progress in multimodal profiling simultaneously within a single cell is likely to be made in the next few years. Correspondingly, computation methods intended for analysing multiomics data are also likely to appear.

Recently, single-nuclei RNA sequencing (snRNA-seq) has become more and more popular. The advantage of snRNA-seq is that it extracts nuclei instead of integrated cells, meaning that it can be utilized to samples that are hard-to-dissociate or frozen [108]. The main difference between processing snRNA-seq data and scRNA-seq data may be the importance of using mitochondrial proportion for quality control since nuclei should not contain any mitochondrial [109]. As the computational analysis of snRNA-seq data is very similar to scRNA-seq for downstream analysis, tools mentioned above could be used for snRNA-seq analysis directly [25].

Single-cell RNA sequencing has shown great effects in revealing cell heterogeneity; however, positional information is generally missing. To overcome this issue, combining spatial transcriptomics with scRNA-seq seems extremely promising. Although spatial methods have achieved several results, there are still many limitations, such as throughput and spatial resolution [110–112]. We believe that in the near feature, spatial

Table 5. Integrating datasets methods for single-cell RNA sequencing data

Method	Category	Description	Availability	Refs
MNN correction	Between different scRNA-seq datasets	Identifies MNN pairs between datasets and uses this information to compute the batch effect. The MNN algorithm assumes that the batch effect is orthogonal to the biological difference, which makes the MNN model more reasonable than the linear model, as in high-dimension space, most random vectors are orthogonal.	https://github.com/MarioniLab/MNN2017/	[100]
FastMNN	Between different scRNA-seq datasets	Is often used in practice due to its shorter running time, which is achieved by performing dimension reduction before finding neighbours.	https://marionilab.github.io/FurtherMNN2018/theory/description.html	[101]
MutiCCA	Between different scRNA-seq datasets	Employs CCA, which is implemented in the popular R toolkit Seurat V2.	https://satijalab.org/seurat/	[16]
Harmony	Between different scRNA-seq datasets	Projects the high dimensional data into a subspace by PCA and iteratively correct batch effect by maximizing the diversity of batches of similar cells within each cluster.	https://github.com/immunogenomics/harmony	[103]
Seurat V3 and Seurat V4	Between multimodal datasets	Seurat V3 utilizes prior cell-type marker genes to build a probabilistic model for annotation, while Seurat V4 proposes WNN algorithm.	https://satijalab.org/seurat/	[17, 104]
LIGER	Between multimodal datasets	Employs integrative nonnegative matrix factorization to obtain shared and dataset-specific factors, which are further utilized for clustering and identifying cells by matching to a reference dataset.	https://github.com/welch-lab/liger	[105]

transcriptomics would play a more important role in better utilizing single-cell sequencing technology.

Key Points

- This paper summarized popular approaches of critical steps (clustering, trajectory inference, cell-type annotation and integrating datasets) for downstream analysis of single-cell RNA sequencing data.
- All methods were grouped into several categories for better discussion, and advantages and limitations were further discussed.
- We provided our suggestions for how to choose proper tools in different situations, and URLs for all the tools mentioned in this paper are also given.
- This paper could also be beneficial for bioinformatic tool developers who aim to develop new tools for single-cell sequencing data.

Supplementary Data

Supplementary data are available online at https://academic.oup.com/bib.

Code availability

All tools mentioned in this review are listed in a github repository which is available from (https://github.com/ZilongZhang44/single-cell-downstream-analysis).

Funding

National Key R&D Program of China (2018YFC0910405); National Natural Science Foundation of China (No. 61922020, No. 61771331, No. 91935302).

Conflict of interest

The authors declare that they have no conflict of interest.

References

1. Tang F, Barbacioru C, Wang Y, *et al*. mRNA-Seq whole-transcriptome analysis of a single cell. *Nat Methods* 2009;**6**(5):377–82.
2. Potter SS. Single-cell RNA sequencing for the study of development, physiology and disease. *Nat Rev Nephrol* 2018;**14**(8):479–92.
3. Pradeep C, Nandan D, Das AA, *et al*. Comparative transcriptome profiling of disruptive technology, single-molecule direct RNA sequencing. *Curr Bioinforma* 2020;**15**(2):165–72.
4. Chen KH, Boettiger AN, Moffitt JR, *et al*. RNA imaging. Spatially resolved, highly multiplexed RNA profiling in single cells. *Science* 2015;**348**(6233):aaa6090.
5. 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.
6. Rosenberg AB, Roco CM, Muscat RA, *et al*. Single-cell profiling of the developing mouse brain and spinal cord with split-pool barcoding. *Science* 2018;**360**(6385):176–82.
7. 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.
8. Wu YE, Pan L, Zuo Y, *et al*. Detecting activated cell populations using single-cell RNA-Seq. *Neuron* 2017;**96**(2):313–329.e6.

9. Grün D, Lyubimova A, Kester L, *et al*. Single-cell messenger RNA sequencing reveals rare intestinal cell types. *Nature* 2015;**525**(7568):251–5.
10. Li F, *et al*. Single cell RNA and immune repertoire profiling of COVID-19 patients reveal novel neutralizing antibody. *Protein Cell* 2020;1–5. 10.1007/s13238-020-00807-6.
11. Grün D, van Oudenaarden A. Design and analysis of single-cell sequencing experiments. *Cell* 2015;**163**(4):799–810.
12. Papalexi E, Satija R. Single-cell RNA sequencing to explore immune cell heterogeneity. *Nat Rev Immunol* 2018;**18**(1):35–45.
13. 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.
14. Grün D, Kester L, van Oudenaarden A. Validation of noise models for single-cell transcriptomics. *Nat Methods* 2014;**11**(6):637–40.
15. Islam S, Zeisel A, Joost S, *et al*. Quantitative single-cell RNA-seq with unique molecular identifiers. *Nat Methods* 2014;**11**(2):163–6.
16. 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.
17. Stuart T, Butler A, Hoffman P, *et al*. Comprehensive integration of single-cell data. *Cell* 2019;**177**(7):1888–1902.e21.
18. Wolf FA, Angerer P, Theis FJ. SCANPY: large-scale single-cell gene expression data analysis. *Genome Biol* 2018;**19**(1):15.
19. Kiselev VY, Andrews TS, Hemberg M. Challenges in unsupervised clustering of single-cell RNA-seq data. *Nat Rev Genet* 2019;**20**(5):273–82.
20. Svensson V, Vento-Tormo R, Teichmann SA. Exponential scaling of single-cell RNA-seq in the past decade. *Nat Protoc* 2018;**13**(4):599–604.
21. Vieth B, Parekh S, Ziegenhain C, *et al*. A systematic evaluation of single cell RNA-seq analysis pipelines. *Nat Commun* 2019;**10**(1):4667.
22. Zhang Z, Cui F, Wang C, *et al*. Goals and approaches for each processing step for single-cell RNA sequencing data. *Brief Bioinform* 2020. 10.1093/bib/bbaa314.
23. Luecken MD, Theis FJ. Current best practices in single-cell RNA-seq analysis: a tutorial. *Mol Syst Biol* 2019;**15**(6): e8746.
24. Andrews TS, *et al*. Tutorial: guidelines for the computational analysis of single-cell RNA sequencing data. *Nat Protoc* 2020;**16**:1–9.
25. Amezquita RA, Lun ATL, Becht E, *et al*. Orchestrating single-cell analysis with Bioconductor. *Nat Methods* 2020;**17**(2):137–45.
26. Wu Y, Zhang K. Tools for the analysis of high-dimensional single-cell RNA sequencing data. *Nat Rev Nephrol* 2020;**16**(7):408–21.
27. 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.
28. Robinson MD, McCarthy DJ, Smyth GK. edgeR: a Bioconductor package for differential expression analysis of digital gene expression data. *Bioinformatics* 2010;**26**(1): 139–40.
29. Ritchie ME, Phipson B, Wu D, *et al*. Limma powers differential expression analyses for RNA-sequencing and microarray studies. *Nucleic Acids Res* 2015;**43**(7):e47.
30. Haynes W. In: Dubitzky W *et al*. (eds). *Wilcoxon Rank Sum Test, in Encyclopedia of Systems Biology*. New York, NY: Springer, 2013, 2354–5.
31. Kharchenko PV, Silberstein L, Scadden DT. Bayesian approach to single-cell differential expression analysis. *Nat Methods* 2014;**11**(7):740–2.
32. Finak G, McDavid A, Yajima 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.
33. Vu TN, Wills QF, Kalari KR, *et al*. Beta-Poisson model for single-cell RNA-seq data analyses. *Bioinformatics* 2016;**32**(14):2128–35.
34. Soneson C, Robinson MD. Bias, robustness and scalability in single-cell differential expression analysis. *Nat Methods* 2018;**15**(4):255–61.
35. Blondel VD, Guillaume JL, Lambiotte R, *et al*. Fast unfolding of communities in large networks. *J Stat Mech* 2008;**2008**(10):P10008.
36. 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.
37. Duò A, Robinson MD, Soneson C. A systematic performance evaluation of clustering methods for single-cell RNA-seq data. *F1000Res* 2018;**7**:1141.
38. Lloyd S. Least squares quantization in PCM. *IEEE Trans Inf Theory* 1982;**28**(2):129–37.
39. Kiselev VY, Kirschner K, Schaub MT, *et al*. SC3: consensus clustering of single-cell RNA-seq data. *Nat Methods* 2017;**14**(5):483–6.
40. Lin P, Troup M, Ho JW. CIDR: ultrafast and accurate clustering through imputation for single-cell RNA-seq data. *Genome Biol* 2017;**18**(1):59.
41. Ji Z, Ji H. TSCAN: pseudo-time reconstruction and evaluation in single-cell RNA-seq analysis. *Nucleic Acids Res* 2016;**44**(13):e117.
42. Aibar S, González-Blas CB, Moerman T, *et al*. SCENIC: single-cell regulatory network inference and clustering. *Nat Methods* 2017;**14**(11):1083–6.
43. Darmanis S, Sloan SA, Zhang Y, *et al*. A survey of human brain transcriptome diversity at the single cell level. *Proc Natl Acad Sci U S A* 2015;**112**(23):7285–90.
44. Kolodziejczyk AA, Kim JK, Tsang JCH, *et al*. Single cell RNA-sequencing of pluripotent states unlocks modular transcriptional variation. *Cell Stem Cell* 2015;**17**(4): 471–85.
45. 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.
46. Deng Q, Ramskold D, Reinius B, *et al*. Single-cell RNA-seq reveals dynamic, random monoallelic gene expression in mammalian cells. *Science* 2014;**343**(6167):193–6.
47. Goolam M, Scialdone A, Graham SJL, *et al*. Heterogeneity in Oct4 and Sox2 targets biases cell fate in 4-cell mouse embryos. *Cell* 2016;**165**(1):61–74.
48. Hubert L, Arabie P. Comparing partitions. *J Classif* 1985;**2**(1):193–218.
49. Kleinberg J. An impossibility theorem for clustering. In: *Proceedings of the 15th International Conference on Neural Information Processing Systems*. Cambridge, MA, USA: MIT Press, 2002, 463–70.
50. Tanay A, Regev A. Scaling single-cell genomics from phenomenology to mechanism. *Nature* 2017;**541**(7637):331–8.
51. Etzrodt M, Endele M, Schroeder T. Quantitative single-cell approaches to stem cell research. *Cell Stem Cell* 2014;**15**(5):546–58.

52. Chen J, Rénia L, Ginhoux F. Constructing cell lineages from single-cell transcriptomes. *Mol Asp Med* 2018;**59**:95–113.
53. Saelens W, Cannoodt R, Todorov H, *et al*. A comparison of single-cell trajectory inference methods. *Nat Biotechnol* 2019;**37**(5):547–54.
54. 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.
55. Wolf FA, Hamey FK, Plass M, *et al*. PAGA: graph abstraction reconciles clustering with trajectory inference through a topology preserving map of single cells. *Genome Biol* 2019;**20**(1):59.
56. Street K, Risso D, Fletcher RB, *et al*. Slingshot: cell lineage and pseudotime inference for single-cell transcriptomics. *BMC Genomics* 2018;**19**(1):477.
57. Cao J, Spielmann M, Qiu X, *et al*. The single-cell transcriptional landscape of mammalian organogenesis. *Nature* 2019;**566**(7745):496–502.
58. Melville LMJHJ. UMAP: uniform manifold approximation and projection for dimension reduction. *arXiv e-prints* 2018;1802.03426. https://arxiv.org/abs/1802.03426.
59. Cannoodt R, Saelens W, Sichien D, *et al*. SCORPIUS improves trajectory inference and identifies novel modules in dendritic cell development. *bioRxiv* 2016 2016. doi: 10.1101/079509 preprint: not peer reviewed.
60. Breiman L. Random forests. *Mach Learn* 2001;**45**(1):5–32.
61. Tran TN, Bader GD. Tempora: cell trajectory inference using time-series single-cell RNA sequencing data. *PLoS Comput Biol* 2020;**16**(9):e1008205.
62. Moon KR, van Dijk D, Wang Z, *et al*. *Visualizing structure and transitions in high-dimensional biological data*. *Nat Biotechnol*, 2019;**37**:1482–1492.
63. la Manno G, Soldatov R, Zeisel A, *et al*. RNA velocity of single cells. *Nature* 2018;**560**(7719):494–8.
64. Bergen V, Lange M, Peidli S, *et al*. Generalizing RNA velocity to transient cell states through dynamical modeling. *Nat Biotechnol* 2020;**38**(12):1408–14.
65. Holguera I, Desplan C. Neuronal specification in space and time. *Science* 2018;**362**(6411):176–80.
66. Kester L, van Oudenaarden A. Single-cell transcriptomics meets lineage tracing. *Cell Stem Cell* 2018;**23**(2):166–79.
67. Aran D, Looney AP, Liu L, *et al*. Reference-based analysis of lung single-cell sequencing reveals a transitional profibrotic macrophage. *Nat Immunol* 2019;**20**(2):163–72.
68. Kiselev VY, Yiu A, Hemberg M. Scmap: projection of single-cell RNA-seq data across data sets. *Nat Methods* 2018;**15**(5):359–62.
69. Andrews TS, Hemberg M. M3Drop: dropout-based feature selection for scRNASeq. *Bioinformatics* 2019;**35**(16):2865–7.
70. Alquicira-Hernandez J, Sathe A, Ji HP, *et al*. scPred: accurate supervised method for cell-type classification from single-cell RNA-seq data. *Genome Biol* 2019;**20**(1):264.
71. Jiang Q, Wang G, Jin S, *et al*. Predicting human microRNA-disease associations based on support vector machine. *Int J Data Min Bioinform* 2013;**8**(3):282–93.
72. Hou R, Denisenko E, Forrest ARR. scMatch: a single-cell gene expression profile annotation tool using reference datasets. *Bioinformatics* 2019;**35**(22):4688–95.
73. Tan Y, Cahan P. SingleCellNet: a computational tool to classify single cell RNA-Seq data across platforms and across species. *Cell Syst* 2019;**9**(2):207–213.e2.
74. Abdelaal T, Michielsen L, Cats D, *et al*. A comparison of automatic cell identification methods for single-cell RNA sequencing data. *Genome Biol* 2019;**20**(1):194.
75. Hearst MA. Support Vector Machines 1998;**13**(4%J IEEE Intelligent Systems):18–28.
76. Zhang AW, O'Flanagan C, Chavez EA, *et al*. Probabilistic cell-type assignment of single-cell RNA-seq for tumor microenvironment profiling. *Nat Methods* 2019;**16**(10):1007–15.
77. Cao Y, Wang X, Peng G. SCSA: a cell type annotation tool for single-cell RNA-seq data. *Front Genet* 2020;**11**:490.
78. Zhang X, *et al*. CellMarker: a manually curated resource of cell markers in human and mouse. *Nucleic Acids Res* 2018;**47**(D1):D721–8.
79. Franzén O, Gan LM, Björkegren JLM. PanglaoDB: a web server for exploration of mouse and human single-cell RNA sequencing data. *Database (Oxford)* 2019;**2019**:baz046.
80. Han X, Zhou Z, Fei L, *et al*. Construction of a human cell landscape at single-cell level. *Nature* 2020;**581**(7808):303–9.
81. Forcato M, Romano O, Bicciato S. Computational methods for the integrative analysis of single-cell data. *Brief Bioinform* 2020. 10.1093/bib/bbaa042.
82. 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.
83. Risso D, Perraudeau F, Gribkova S, *et al*. A general and flexible method for signal extraction from single-cell RNA-seq data. *Nat Commun* 2018;**9**(1):284.
84. Tran HTN, Ang KS, Chevrier M, *et al*. A benchmark of batch-effect correction methods for single-cell RNA sequencing data. *Genome Biol* 2020;**21**(1):12.
85. Zhang ZM, Tan JX, Wang F, *et al*. Early diagnosis of hepatocellular carcinoma using machine learning method. *Front Bioeng Biotechnol* 2020;**8**:254.
86. Zhang ZM, Wang JS, Zulfiqar H, *et al*. Early diagnosis of pancreatic ductal adenocarcinoma by combining relative expression orderings with machine-learning method. *Front Cell Dev Biol* 2020;**8**:582864.
87. Leek JT, Johnson WE, Parker HS, *et al*. The sva package for removing batch effects and other unwanted variation in high-throughput experiments. *Bioinformatics* 2012;**28**(6):882–3.
88. Schwartzman O, Tanay A. Single-cell epigenomics: techniques and emerging applications. *Nat Rev Genet* 2015;**16**(12):716–26.
89. Kelsey G, Stegle O, Reik W. Single-cell epigenomics: recording the past and predicting the future. *Science* 2017;**358**(6359):69–75.
90. Wei L, Chen H, Su R. M6APred-EL: a sequence-based predictor for identifying N6-methyladenosine sites using ensemble learning. *Mol Ther Nucleic Acids* 2018;**12**:635–44.
91. Liu K, Cao L, du P, *et al*. im6A-TS-CNN: identifying the N(6)-Methyladenine site in multiple tissues by using the convolutional neural network. *Mol Ther Nucleic Acids* 2020;**21**:1044–9.
92. Wu M, Singh AK. Single-cell protein analysis. *Curr Opin Biotechnol* 2012;**23**(1):83–8.
93. Shen Y, Tang J, Guo F. Identification of protein subcellular localization via integrating evolutionary and physicochemical information into Chou's general PseAAC. *J Theor Biol* 2019;**462**:230–9.
94. Wei L, Wan S, Guo J, *et al*. A novel hierarchical selective ensemble classifier with bioinformatics application. *Artif Intell Med* 2017;**83**:82–90.

95. Wei L, Xing P, Shi G, *et al*. Fast prediction of protein methylation sites using a sequence-based feature selection technique. *IEEE/ACM Trans Comput Biol Bioinform* 2019;**16**(4):1264–73.
96. Gawad C, Koh W, Quake SR. Single-cell genome sequencing: current state of the science. *Nat Rev Genet* 2016;**17**(3):175–88.
97. Guo F, Wang D, Wang L. Progressive approach for SNP calling and haplotype assembly using single molecular sequencing data. *Bioinformatics* 2018;**34**(12):2012–8.
98. Wei L, Liao M, Gao Y, *et al*. Improved and promising identification of human MicroRNAs by incorporating a high-quality negative set. *IEEE/ACM Trans Comput Biol Bioinform* 2014;**11**(1):192–201.
99. Stoeckius M, Hafemeister C, Stephenson W, *et al*. Simultaneous epitope and transcriptome measurement in single cells. *Nat Methods* 2017;**14**(9):865–8.
100. Haghverdi L, Lun ATL, Morgan MD, *et al*. Batch effects in single-cell RNA-sequencing data are corrected by matching mutual nearest neighbors. *Nat Biotechnol* 2018;**36**(5):421–7.
101. Lun A. *A description of the theory behind the fastMNN algorithm*. 2019. Available from: https://marionilab.github.io/FurtherMNN2018/theory/description.html.
102. Hardoon DR, Szedmak S, Shawe-Taylor J. Canonical correlation analysis: an overview with application to learning methods. *Neural Comput* 2004;**16**(12):2639–64.
103. Korsunsky I, Millard N, Fan J, *et al*. Fast, sensitive and accurate integration of single-cell data with harmony. *Nat Methods* 2019;**16**(12):1289–96.
104. Hao Y, *et al*. *Integrated analysis of multimodal single-cell data*, 2020, 2020.10.12.335331.
105. Welch JD, Kozareva V, Ferreira A, *et al*. Single-cell multiomic integration compares and contrasts features of brain cell identity. *Cell* 2019;**177**(7):1873–1887.e17.
106. Luecken M, *et al*. *Benchmarking atlas-level data integration in single-cell genomics*, 2020, 2020.05.22.111161.
107. Jiang L, Wang C, Tang J, *et al*. LightCpG: a multi-view CpG sites detection on single-cell whole genome sequence data. *BMC Genomics* 2019;**20**(1):306.
108. Slyper M, Porter CBM, Ashenberg O, *et al*. A single-cell and single-nucleus RNA-Seq toolbox for fresh and frozen human tumors. *Nat Med* 2020;**26**(5):792–802.
109. Wu H, Kirita Y, Donnelly EL, *et al*. Advantages of single-nucleus over single-cell RNA sequencing of adult kidney: rare cell types and novel cell states revealed in fibrosis. *J Am Soc Nephrol* 2019;**30**(1):23–32.
110. Edsgärd D, Johnsson P, Sandberg R. Identification of spatial expression trends in single-cell gene expression data. *Nat Methods* 2018;**15**(5):339–42.
111. Moffitt JR, Bambah-Mukku D, Eichhorn SW, *et al*. Molecular, spatial, and functional single-cell profiling of the hypothalamic preoptic region. *Science* 2018;**362**(6416) eaau5324.
112. Lein E, Borm LE, Linnarsson S. The promise of spatial transcriptomics for neuroscience in the era of molecular cell typing. *Science* 2017;**358**(6359):64–9.