MILITARY MEDICAL RESEARCH

REVIEW Open Access

# Data analysis guidelines for single-cell RNA-seq in biomedical studies and clinical applications

Min Su[1†], Tao Pan[2†], Qiu-Zhen Chen[1†], Wei-Wei Zhou[3†], Yi Gong[1,4], Gang Xu[2], Huan-Yu Yan[1], Si Li[2], Qiao-Zhen Shi[1], Ya Zhang[2], Xiao He[5], Chun-Jie Jiang[6], Shi-Cai Fan[7], Xia Li[3*], Murray J. Cairns[8,9*], Xi Wang[1*] and Yong-Sheng Li[2*]

**Abstract**

The application of single-cell RNA sequencing (scRNA-seq) in biomedical research has advanced our understanding of the pathogenesis of disease and provided valuable insights into new diagnostic and therapeutic strategies. With the expansion of capacity for high-throughput scRNA-seq, including clinical samples, the analysis of these huge volumes of data has become a daunting prospect for researchers entering this field. Here, we review the workflow for typical scRNA-seq data analysis, covering raw data processing and quality control, basic data analysis applicable for almost all scRNA-seq data sets, and advanced data analysis that should be tailored to specific scientific questions. While summarizing the current methods for each analysis step, we also provide an online repository of software and wrapped-up scripts to support the implementation. Recommendations and caveats are pointed out for some specific analysis tasks and approaches. We hope this resource will be helpful to researchers engaging with scRNA-seq, in particular for emerging clinical applications.

**Keywords:** Single-cell RNA-sequencing (scRNA-seq), Data analysis, Biomedical research, Clinical applications

## Background

Complex tissues consist of a variety of cell types that occur in a huge variety of mixtures and states. The functional genomic information contained within each cell is often quite different from the neighboring cell populations and even cells of the same type. This means that the molecular analyses of cell populations in bulk tissues are inherently unreliable and insensitive. The incredible sensitivity and specificity that can be achieved by quantifying molecular alterations at single-cell resolution have led to unprecedented opportunities for uncovering the molecular mechanisms underlying the pathogenesis and progression of the disease [1]. Since its inception, single-cell RNA-sequencing (scRNA-seq) has been shown to be a powerful tool for profiling gene expression in individual cells [2–4], in both physiogenesis [5, 6] and pathogenesis [7–9]. For example, by utilizing scRNA-seq

[†]Min Su, Tao Pan, Qiu-Zhen Chen and Wei-Wei Zhou contributed equally to this work

*Correspondence: lixia@hrbmu.edu.cn; murray.cairns@newcastle.edu.au; xiwang@njmu.edu.cn; liyongsheng@hainmc.edu.cn

[1] State Key Laboratory of Reproductive Medicine, Nanjing Medical University, Nanjing 211166, China
[2] College of Biomedical Information and Engineering, the First Affiliated Hospital of Hainan Medical University, Hainan Medical University, Haikou 571199, Hainan, China
[3] College of Bioinformatics Science and Technology, Harbin Medical University, Harbin 150081, Heilongjiang, China
[8] School of Biomedical Sciences and Pharmacy, Faculty of Health and Medicine, the University of Newcastle, University Drive, Callaghan, NSW 2308, Australia
Full list of author information is available at the end of the article

in cancer biology [10, 11], researchers have been able to determine the origin of cancer cells in various tumor types [12, 13]. Moreover, from the treatment and prognosis respect, subpopulations of malignant cells with clinically significant features, such as the poor prognosis in nasopharyngeal carcinoma with dual epithelial–immune characteristics have been discovered [14]. Similarly, strong epithelial-to-mesenchymal transition (EMT) and stemness signatures were observed in metastatic breast cancer cells [15, 16]. With the assistance of scRNA-seq, the quality and validity of organoid systems can also be accurately assessed and systematically evaluated [17–19]. Patient-derived organoid models are currently being applied to the dissection of disease pathology [20] and facilitating drug screening for personalized treatment [21, 22]. Furthermore, distinct cellular states along tumor progress were discovered and drug-resistant cell subsets were identified by joint application of patient-derived organoid and scRNA-seq [23, 24]. In the current coronavirus disease 2019 (COVID-19) pandemic, scRNA-seq accelerates the research for characterizing the molecular basis and, therefore, understanding the pathology of severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2). A variety of scRNA-seq-based studies have revealed the cell subtypes targeted by SARS-CoV-2 [25], profiled gene expression changes in immune cells upon infection [26, 27], quantified the alteration of cell-to-cell interaction between different cell types [26, 28], and provided important resources for the development of potential treatment of COVID-19 [26, 28].

Since the emergence of commercial single-cell platforms, including those offered by 10 × genomics [29, 30] and Singleron [31, 32], scRNA-seq services provided by core facilities of research institutes or third-party companies, are making the technology more accessible, affordable and in some cases a routine technique for biomedical researchers and clinicians [33]. While these service providers typically perform data quality-control and execute basic pipelines for data processing, the high-level data analysis needed for specific research objectives and scientific questions, is not usually available. Thus, most biomedical researchers need to come to grip with the full scope of scRNA-seq data analysis by identifying the most suitable computational tools to dissect their data.

To overcome the barriers in scRNA-seq data analysis, in particular for biomedical studies, this review aims to: 1) summarize the recent advances in algorithm development and benchmarking results for every analysis task in analyzing biomedical scRNA-seq data, and 2) introduce a workflow comprised of recommended software tools that are more appropriate for biomedical applications. The workflow covers basic scRNA-seq data processing, quality control (QC), feature selection, dimensionality reduction, cell clustering and annotation, trajectory inference, cell–cell communications (CCC), transcription factor (TF) active prediction and metabolic analysis. Along with the recommended workflow, we also provide example computational scripts together with the software environment setting, which may facilitate researchers to conduct the data analysis locally. The computational code is available at https://github.com/WXlab-NJMU/scrna-recom. To accommodate upcoming advanced approaches and more application scenarios, we will keep the computational scripts updated.

## General tasks of single-cell RNA-seq data analysis

Typical data analysis steps of scRNA-seq can be generally divided into three stages: raw data processing and QC, basic data analysis applicable for almost all scRNA-seq data sets, and advanced data analysis that should be tailored to specific research scenarios. While basic data analysis steps include data normalization and integration, feature selection, dimensionality reduction, cell clustering, cell type annotation and marker gene identification. The advanced data analysis tasks consist of trajectory inference, CCC analysis, regulon inference and TF activity prediction, and metabolic flux estimation.

### Experimental design

ScRNA-seq experiments need to be carefully designed to optimize the capability in addressing scientific questions [34]. Before starting the data analysis, the following information related to the experiment design needs to be gathered. (1) Species. For biomedical studies and clinical applications, human samples derived from patients are usually collected for sequencing [35–37]. In some cases, to study the underlying molecular mechanisms, mouse and other model organisms are also used [38]. Since the gene names and related data resources are different between humans and other species, it is important to specify the species for data analysis. For simplicity, we will focus on the data derived from human samples. (2) Sample origin. According to the scientific questions and sample accessibility, the sample types can be varied in different studies. For instance, to study solid tumors like hepatocellular carcinoma, tumor biopsies and peritumor samples are collected from patients for a case–control design [39]. Whereas the above design is feasible to some extent, peripheral blood mononuclear cells (PBMCs) are more easily accessible and widely used for scRNA-seq [40, 41]. In addition, cells from patient-derived organoids are often used to study the impact of personal genetic variants on the development of specific organs, which can also be the origin of particular diseases [42, 43]. Knowing the sample origin facilitates particular analysis, such as cell clustering and cell type annotation. (3)

Experiment design. To study disease pathogenesis and the effectiveness of particular treatments, a case–control design is mostly adopted, like the tumor-versus-peritumor design [39]. For diseases such as COVID-19, no normal samples can be obtained from the same patients, thus healthy people with matched age and gender serve as a control group [40]. To control possible covariates between the patients and the control groups, the number of individuals in each group needs to be carefully considered [44]. In (prospective) cohort studies, the sample size is usually considerably larger, so that scRNA-seq cannot be applied to every sample from individual donors; in this case, nested case–control studies [45] and sample multiplexing [46] are often applied. In general, data analysis strategies need to be adjusted according to the types of the experiment design.

### Raw data processing

Raw data processing steps include: sequencing read QC, read mapping [47], cell demultiplexing and cell-wise unique molecular identifier (UMI)-count table generation [48]. Whilst standardized data processing pipelines are provided with the release of scRNA-seq platforms, such as Cell Ranger for 10 × Genomics Chromium [49] and CeleScope (https://github.com/singleron-RD/CeleScope) for Singleron's systems, alternative tools including UMI-tools [48], scPipe [50], zUMIs [51], celseq2 [52], kallisto bustools [53], and scruff [54] can also be used for this procedure. The choice between these pipelines seems less important than the downstream steps according to a recent study benchmarking scRNA-seq analysis [55]. In any case, we would not recommend raw data processing on personal computers, as these pipelines need massive computational resources and are optimized for high-performance computing architectures [56]. Third-party companies usually provide processed data, including UMI count matrices and QC metrics, which enable the researchers to focus on downstream data analysis for addressing scientific questions.

### QC and doublet removal

The purpose of cell QC is to make sure all the 'cells' being analyzed are single and intact cells. Damaged cells, dying cells, stressed cells and doublets need to be discarded [57, 58]. In ultrahigh-throughput scRNA-seq, quantitative metrics used for bulk RNA-seq QC, including read mappability, fraction of reads mapped to exonic regions are computed at only the sample/library level, thus cannot be used for cell QC. Instead, the three mostly used metrics for cell QC are: the total UMI count (i.e., count depth), the number of detected genes, and the fraction of mitochondria-derived counts per cell barcode [56, 59]. Cell Ranger [49] and CeleScope (https://github.com/singleron-RD/CeleScope) usually perform a first-round cell QC, which distinguishes potentially authentic cells from background cell barcodes by examining the distribution of count depth in a scRNA-seq library. One caveat is that, when the damaged cells or cell debris take a considerable proportion in the library, the threshold of a minimum count depth for valid cells is hard to be determined. Possible solutions include the consideration of multiple QC metrics at the same time [56], and the application of more sophisticated approaches to rule out background and low-quality cells [60]. Typically, low numbers of detected genes and low count depth indicate damaged cells, whereas a high proportion of mitochondria-derived counts is indicative of dying cells. By contrast, too many detected genes and high count depth can be indicative of doublets [57, 58]. While R packages like Seurat [61–63] and Scater [64] implement functions to facilitate cell QC, the thresholds of the QC metrics are largely dependent on the tissue studied, cell dissociation protocol, library preparation protocol, etc.. Referring to publications with similar experiment designs would help to determine the thresholds, and advanced researchers may also inspect the joint distribution of the QC metrics. Notably, accumulated expression of genes encoding ribosomal proteins is not a typical QC metric, as the variation of ribosomal protein expression can be biologically meaningful [65].

In addition, various sources of contamination need to be considered and controlled during the QC step. For example, libraries derived from PBMCs and solid tissues can be contaminated by red blood cells, and thus cells expressing a high level of hemoglobin genes (e.g., *HBB*) are usually discarded [66, 67]. Another source of contamination is cell-free or ambient RNA, as evidenced by reads mapped back to specific genes in cell-free droplets or wells in high-throughput scRNA-seq [68, 69]. Methods and tools for estimating and removing such contamination have been recently developed, including SoupX [68], DecontX [69], fast correction for ambient RNA (FastCAR) [70] and CellBender [71]. Removal of the background signal caused by ambient RNA in single-cell gene expression improves downstream analyses and biological interpretation [69, 71].

In high-throughput scRNA-seq experiments, it is not uncommon to observe a high rate of doublets, which may reach up to 40% of cell barcodes [72, 73]. For this reason, a filtering step that only considers count depth and the number of detected genes is not adequate, particularly when the cell type composition is complex such that the count depth distribution of singlets is not distinct from that of doublets. Doublets composed of distinct cell types are likely to confound downstream analysis, particularly in cell clustering, differential expression analysis, and trajectory inference [56, 74]. Fortunately, a number of sophisticated approaches have been developed to disentangle these confounding signals [72]. These methods consider the gene expression profiles of individual cell barcodes and report doublet scores as an indicator. The doublet scores are calculated based on either artificial doublets [such as single-cell remover of doublets (Scrublet) [74], doubletCells [75], binary classification based doublet scoring (bcds) [76], DoubletDetection [77], DoubletFinder [78], Solo [73], DoubletDecon [79]] or gene co-expression [such as co-expression based doublet scoring (cxds) [76]]. In a recent study, benchmarking the available computational doublet-detection methods with a comprehensive set of synthetic and real data [72], the tool Doubletfinder [78] was recommended because it achieved both the highest detection accuracy and the best performance in downstream analysis.

### Expression normalization

The variability of total UMI counts per cell depends on a range of both technical and biological parameters [56]. The technical factors relate to the efficiency of RNA capture, reverse transcription, cDNA amplification and sequencing depth, whereas the biological factors mostly relate to cell size and cell cycle phase. Because of this variation, it is almost impossible to obtain the absolute number of RNA molecules unless external spike-in RNA control is added to the sequencing libraries [80, 81]. Like bulk RNA-seq, relative RNA abundance is commonly adopted for comparing gene expression profiles between individual cells; therefore, scRNA-seq data are typically normalized by global-scaling methods with scaling factors developed for bulk RNA-seq [82–84], which suppress partially the technical effects [56]. Popular global-scaling methods for bulk RNA-seq include transcript per million (TPM) [85], upper quartile (UQ) normalization [86], trimmed mean of M values (TMM) normalization [87], and the DESeq normalization method [88], however, are not appropriate for scRNA-seq due the tendency for distortion through zero inflation [81]. Normalization methods tailored for scRNA-seq, including single-cell differential expression (SCDE) [84] and model-based analysis of single-cell transcriptomics (MAST) [82], can specifically model dropout events in differential expression analysis of scRNA-seq data. Another approach, Scran [75], overcomes the issues of scaling factor estimation (affected by too many zero counts) by pooling cells of similar gene expression profiles [89]. Moreover, Census estimates the total number of RNA molecules per cell without spike-in controls and uses these estimates as the scaling factors [90]. While simulation studies carried out by Vallejos et al. [81] suggested Scran's pooling strategy outperforms compared tools in scaling factor estimation, the TPM-/count depth-scaling method is widely used in practice [91].

Following scaling factor-based normalization, the resulting values are typically added to one pseudo-count and log-transformed [56, 62]. This step is practically useful and statistically sound, as it mitigates the mean–variance relationship in scRNA-seq count data and also reduces the skewness in expression data [56, 64]. Toward better variance stabilization, SCTransform was recently developed by the Seurat team, which applies regularized negative binomial regression for scRNA-seq data normalization and variance stabilization [92].

Some known biological effects, such as cell cycle and cell stress (featured by overexpression of mitochondrial genes), may hinder the characterization of the particular biological signal of interest [56]. Hence, normalizing or correcting expression profiles against known biological may help interpret the data. For instance, correcting the effects of the cell cycle can improve developmental trajectory reconstruction [93, 94]. The procedure accounting for biological effects can be achieved by scoring related biological features (e.g., cell cycle scores [95]), followed by a simple linear regression against the calculated scores as implemented in Seurat [61, 62]. In addition, dedicated tools such as single-cell latent variable model (scLVM)/factorial single-cell latent variable model (f-scLVM) [93, 96] and cell growth correction (cgCorrect) [97] can also be used for this purpose. Of note, correcting biological effects for one particular analysis (e.g., cell differentiation) may unintentionally hinder the signals for another (e.g., cell proliferation) [56]; care should be taken when choosing data normalization strategies for particular analysis tasks.

### Data integration

As mentioned in the 'Experiment design' section, biomedical studies usually make case versus control comparisons [39]. Usually, batches of samples obtained from different medical centers or hospitals should be integrated before downstream analysis. For studies using patient-derived organoids, data integration also applies to cells harvested at different time points to depict organoid development [98]. In these cases, one other unwanted technical factor, batch effects, cannot be avoided because cells and library preparation were handled by different persons, at different time points, or with a different batch of reagents [91, 99]. In scRNA-seq, batch effects can be nonlinear, which may not be easily disentangled by state-of-the-art batch correction tools, such as ComBat [100]. Therefore, numerous methods have been recently developed for batch effect correction in scRNA-seq data integration, trying to relieve or remove the effects caused by batch-specific biases while preserving biological variations [56, 99]. The batch effect correction methods can be classified into a few categories: 1) tools developed for bulk expression analysis, including ComBat [100] and limma [101]; 2) approaches based on mutual nearest neighbors (MNN) in high-dimensional gene expression space or its subspace, such as mnnCorrect [102], fastMNN [102], Scanorama [103] and batch balanced k nearest neighbours (BBKNN) [104]; 3) methods that try to align cells with correlated/shared features in dimensionality-reduced spaces, including canonical correlation analysis (CCA) [61, 62], Harmony [105], and linked inference of genomic experimental relationships (LIGER) [106]; and 4) methods based on deep generative models, such as scGen [107]. Besides, depending on the choice of integration anchors, the algorithms can also be sorted into different types, such as genomic features as the anchor and cells as the anchor [108].

Recently, Tran et al. [99] compared 14 batch-effect correction methods available at that time on 10 datasets under 5 different integration scenarios. Among them, Harmony [105], LIGER [106], and CCA implemented in Seurat 3 [62] were recommended according to their overall performance [99]. Together with our experience, it is suggested to perform data integration with Harmony, Seurat3/4-CCA, and LIGER in order. This is because there is no clear winner among the three strategies when dealing with distinct datasets [99]. Harmony runs faster than the other tools, suitable for initial exploration; Seurat3/4-CCA is moderate in mixing cells from different batches, whereas LIGER makes the best efforts in batch mixing, sometimes at the cost of cell type purity. Of note, if one wants to evaluate the effectiveness of batch-effect correction or assess the extent of the batch effects in the data, it can be achieved by comparing clustering or visualization results based on batch-effect corrected analysis and that from directly merging cells derived from multiple samples (e.g., merge function in Seurat), and by computing test metrics such as k-nearest-neighbor batch-effect test (kBET) [91].

### Feature selection

While cell QC removes background cells and problematic cells, the feature section is concerning genes. In the human genome, more than 20,000 genes are annotated, and mapped reads are counted for individual gene loci to yield the UMI count matrix. However, not all the > 20,000 genes are informative in characterizing cell-to-cell heterogeneity or distinguishing cell types/states [56]. Therefore, the term 'feature selection' was borrowed from the fields of statistics and machine learning to describe the process of selecting biologically informative genes for downstream analysis. This process is typically unsupervised, meaning that no information related to cell types or other biological processes of interest is needed.

Considering the relatively high noise level in scRNA-seq data, feature selection usually identifies genes with stronger biological variability than technical noise [58]. Since the technical noise largely depends on the mean expression of genes [109], highly variable genes (HVGs) were originally identified by examining the relationship between the coefficient of variation and expression means [58]. Due to its usefulness in reducing technical noise and relieving the computational demand in downstream analysis, such as cell clustering and dimensionality reduction for visualization [110], many other tools for HVG identification were developed and comparatively evaluated [111–113]. Instead of identifying HVGs, alternative feature selection methods consider dropouts and prioritize genes with a higher-than-expected number of observed zeros [114].

The number of genes selected for downstream analysis is theoretically dependent on the complexity of cellular composition in the samples studied. While approaches for HVG identification can determine the number of HVGs at a given significance level, identifying a fixed number of HVGs is becoming popular, and typically the HVG number is between 1000 and 5000 [56]. Studies have shown that downstream analysis is not sensitive to the exact number of HVGs [110, 115]. Notably, some unfavorable covariates such as batch effect may distort HVG identification [82]. Therefore, HVG selection should be performed after correction for the covariates. In the presence of batch effects, feature selection may also be conducted in individual samples before data integration [56].

### Dimensionality reduction and visualization

With 1000–5000 HVGs selected, the dimensionality of the expression data is still high, thus obstructing manual inspection of the dataset, such as visualization, clustering and cell type annotations [116]. To this end, the dimensions of the expression matrixes can be further reduced by dimensionality reduction techniques, which project the cells from a high-dimensional space into a low-dimensional embedding space, and preserve the biological information on cell-to-cell variability [56, 59]. The widely used methods for dimensionality reduction include principal component analysis (PCA) [117], non-negative matrix factorization (NMF) [118], multidimensional scaling (MDS) [119], t-distributed stochastic neighbor embedding (t-SNE) [120] and uniform manifold approximation and projection (UMAP) [121].

PCA is a general technique for dimensionality reduction and denoising, and has been widely used in scRNA-seq data analysis [122, 123]. With the linear projection of the original expression matrix to its subspace, PCA gives the principal components (PCs) in order of significance. While the first two or three PCs can be used for visualization, a few more PCs are typically retained for downstream analysis, such as cell clustering and trajectory inference. The number of PCs for retention largely depends on the complexity of the dataset [59], and can be determined by the "elbow" method [56] or the jackstraw permutation-test-based method [95, 124]. Nevertheless, PCA cannot take into account the dropout events in the analysis, which leads to the development of several new methods. Zero-inflated factor analysis (ZIFA) is one of such methods based on factor analysis, which explicitly models the dropout characteristics and outperforms the comparative methods [125]. Similar to PCA, NMF is a linear projection method for dimensionality reduction, and showed robust performance in cell clustering based on scRNA-seq [118].

For visualization, nonlinear dimensionality reduction methods are more suitable, which allow a global nonlinear embedding in a two-/three-dimensional space [126]. MDS is one of the nonlinear dimensionality reduction methods and preserves the distance among the cells in the original space [119]. However, MDS can be not scalable to large-scale scRNA-seq data because calculating the pairwise distances becomes computationally demanding when the number of cells is huge [127]. Emerging evidence suggests t-SNE and UMAP are more suitable for scRNA-seq data, which have been widely used in single-cell analysis for data visualization and cell population identification. However, t-SNE usually suffers from limitations such as slow computation time for large-scale scRNA-seq datasets [128] and global data structure was not preserved [121]. With advantages in the above two respects, UMAP currently becomes the most popular choice for dimensionality reduction. UMAP not only helps visualize the cell clusters but also facilitates annotating the cell clusters. It is worth noting, however, that while UMAP strikes a balance between preserving global data structure and capturing local similarity, the cell-to-cell distance in the resulted space is not preserved. Hence, downstream analysis like clustering and pseudotime inference is typically executed based on the PCA results with several to dozens of PCs.

### Identification of cell subpopulations

One of the key applications in single-cell transcriptomics is to determine cell subpopulations based on cell clustering or classification [129, 130]. Due to the high level of noise in the scRNA-seq data, applying dimensionality reduction approaches to scRNA-seq matrix data may facilitate cell clustering. Whilst PCA is commonly used for bulk RNA-seq, the true biological variability of gene expression among cell subpopulations may not be readily distinguished by a small number of PCs. To better account for this variation, NMF was adapted to disentangle subpopulations in single-cell transcriptome data [118, 131], and has been shown to outperform PCA with greater accuracy and robustness (Fig. 1). Likewise, SinNLRR was developed to provide robust clustering of gene expression subspace by non-negative and low-rank representation [132].

State-of-the-art clustering methods, such as the k-means algorithm, have also been applied to scRNA-seq datasets, and based on this application, the single-cell consensus clustering (SC3) approach was developed [133] (Fig. 1). Another category of popularly used methods for cell clustering in scRNA-seq is community detection methods based on a nearest-neighbor network for the cells [134], and was adopted and implemented in the Seurat R package [61] (Fig. 1). Besides, the community has developed a diversity of approaches for cell clustering. For instance, BackSPIN takes advantage of the biclustering technique to avoid unfavorable pairwise comparisons in hierarchical clustering [135], single-cell interpretation via multikernel learning (SIMLR) is based on multi-kernel learning [136], clustering through imputation and dimensionality reduction (CIDR) [137] utilizes imputation to mitigate the impact of dropouts in scRNA-seq, and Single-cell Aggregated Clustering via Mixture Model Ensemble clustering (SAME-clustering) [138] ensembles clustering results from multiple methods. Nevertheless, two independent benchmarking studies have shown that SC3 and the clustering method in Seurat perform similarly to each other and outperform all other comparative methods [139, 140].

Similarity or distance metrics are crucial for clustering cells in scRNA-seq, which can be specific to experiment platforms or particular samples. It has been shown that, compared to unsupervised clustering methods, supervised methods for cell type identification suffered less from batch effects, number of cell types, and imbalance in cell population composition [141]. Mechanistically, the supervised methods rely on a comprehensive reference database with known cell types annotated, based on which a classification model is trained for predicting the cell types in an unannotated dataset [142, 143]. CellAssign [144], scmap [145], single cell recognition (SingleR) [146], characterization of cell types aided by hierarchical classification (CHETAH) [147], and SingleCellNet [148] are methods of this category. Albeit the clear strength of

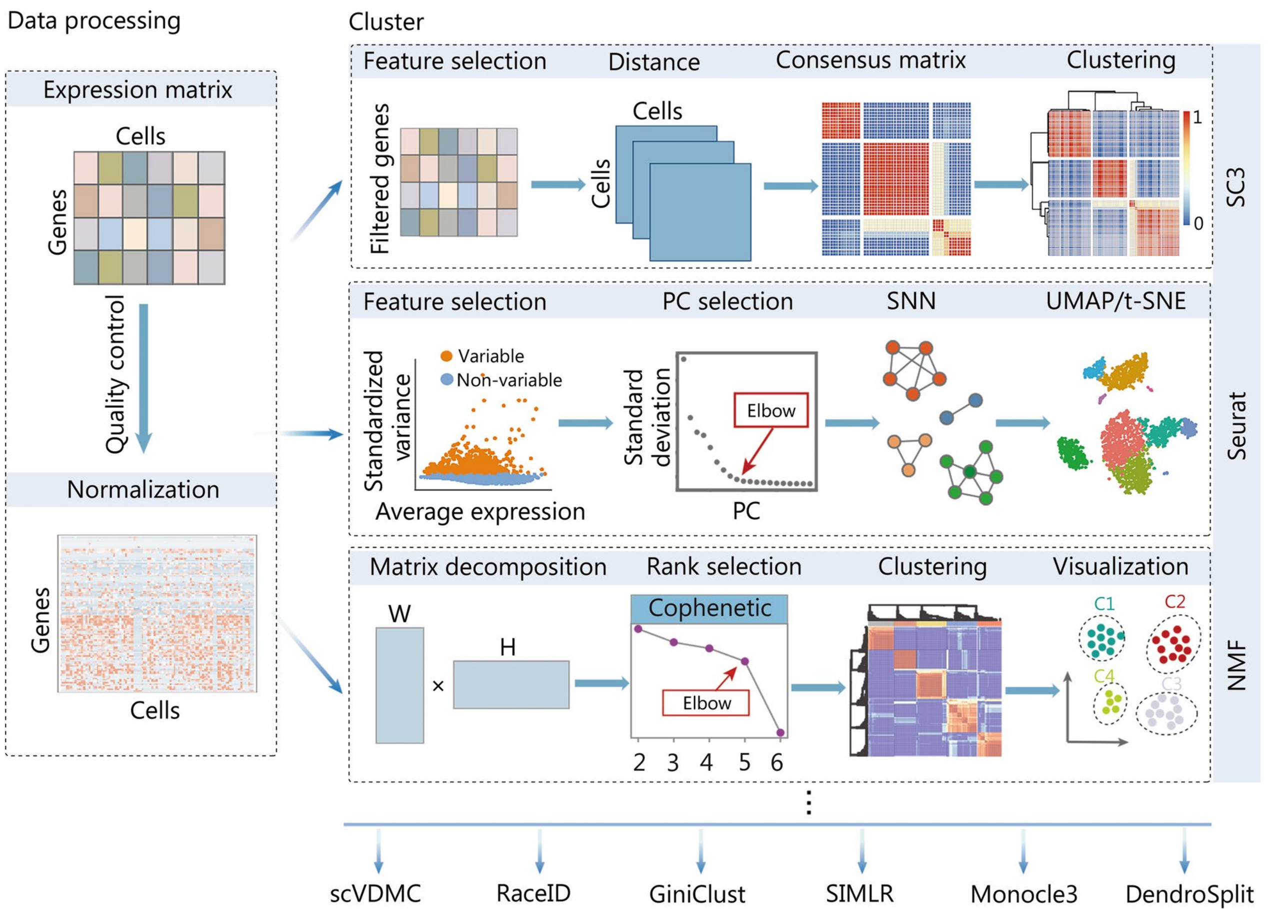

**Fig. 1** Typical computational strategies and methods for clustering cells using scRNA-seq data. With the processed scRNA-seq data, the SC3 approach, the Seurat clustering implementation based on the community detection method, and the NMF method are popular choices. scRNA-seq single-cell RNA sequencing, SC3 single-cell consensus clustering, NMF non-negative matrix factorization, PC principal component, SNN shared nearest neighbor, scVDMC variance-driven multitask clustering of scRNA-seq data, SIMLR single-cell interpretation via multikernel learning, UMAP uniform manifold approximation and projection, t-SNE t-distributed stochastic neighbor embedding

the supervised methods, unsupervised methods are generally better at identifying unknown cell types and have higher computational efficiency [141]. Therefore, the clustering methods implemented in Seurat have the best overall performance, and are suggested as the first choice of cell type identification [141].

Another important issue for single-cell clustering analysis is the detection of rare cell types, which play an important role in complex diseases but have a low abundance. RaceID [129], GiniClust [149], SINCERA [150] and DendroSplit [151] are clustering algorithms specifically designed to identify rare cell types in scRNA-seq data analysis.

### Cell type annotation

Assigning cell identities to cell subpopulations, a process known as cell type annotation, is a critical step in scRNA-seq data analysis [152]. Manual annotation of cell types is time-consuming and potentially subjective. Thus, emerging computational tools have been developed for automatic cell type annotation [143, 152]. These computation methods usually can be classified into three main groups (Fig. 2).

The first type is marker gene-based, which relies on the availability of cell type-specific markers in public databases or literature. CellMarker [153] and PanglaoDB [154] are commonly used online resources storing the markers for a large variety of cell types in the tissues of humans and mouse. CellMarker deposits over 13,000 cell markers of about 500 cell types of humans by manually curating over 100,000 published papers [153], and PanglaoDB is a community-curated cell marker compendium, containing 6000 markers for different cell types from over 1000 scRNA-seq experiments [154]. Moreover, the TF-Marker database was developed for

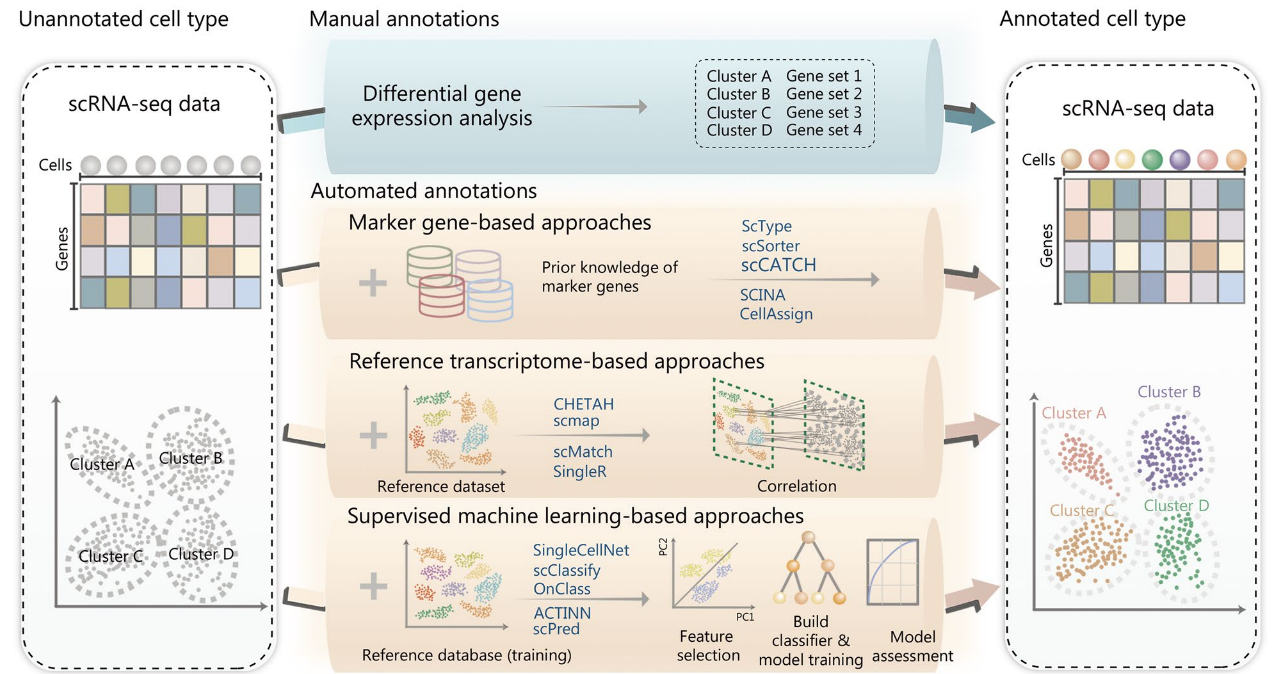

**Fig. 2** Typical strategies and representative methods for annotating cell subpopulations identified by scRNA-seq. In addition to manual annotation, which is potentially time-consuming and subjective, automated cell type annotation can be mainly sorted into three categories: marker gene-based, reference transcriptome-based, and supervised machine learning-based approaches. The example approach names are listed in the plot. scRNA-seq single-cell RNA sequencing, scCATCH single-cell cluster-based automatic annotation toolkit for cellular heterogeneity, SCINA semi-supervised category identification and assignment, CHETAH characterization of cell types aided by hierarchical classification, SingleR single cell recognition, OnClass ontology-based single cell classification, ACTINN automated cell type identification using neural networks

providing cell or tissue-specific TFs and related markers for humans [155]. These databases are valuable resources for cell type annotations. Meanwhile, a number of tools have been developed to use the marker genes for cell type annotations, such as ScType [156], scSorter [157], semi-supervised category identification and assignment (SCINA) [158], single-cell cluster-based automatic annotation toolkit for cellular heterogeneity (scCATCH) [159] and CellAssign [144]. Some of these methods apply sophisticated statistical models to make use of the prior knowledge of marker genes. For example, SCINA builds a semi-supervised model to exploit previously identified marker genes with the expectation–maximization (EM) algorithm [158], and CellAssign leverages a probabilistic graphical model to annotate cells into predefined or novel cell types based on prior knowledge of cell-type marker genes, while accounting for batch and sample effects [144].

The second group of methods is reference transcriptome-based, which uses cell type-labeled scRNA-seq datasets as input for cell type annotation, via the search for the best correlation between the queried data and the reference data. Popular tools of this group include CHETAH [147], scmap [145], scMatch [160] and SingleR [146]. The CHETAH algorithm is based on a hierarchical tree built by reference profiles of known cell types, and searches for a cell's best annotation by stepwise traversing the tree from the root node to a leaf node [147]. By calculating the correlation coefficients between the input cell and two tree branches under consideration based on the 200 most discriminating genes for the two branches, a profile score and confidence score are calculated for selecting tree branches to continue tree traversing. The SingleR approach correlates each unannotated single-cell transcriptome with the reference transcriptomes of known cell types based on HVGs among cell types in the reference data [146]. SingleR assigns cell identity in an iterative manner, and in each iteration the reference set is reduced to refine the assignment. Notably, the comprehensiveness of the reference transcriptomics data is critical for this group of methods. The reference data from Blueprint [161], Encode [162] and the Human Primary Cell Atlas [163] are commonly used.

Lastly, the third group leverages supervised machine learning-based approaches, where classifiers trained by a labeled reference are then applied to predict cell types of unannotated cells. For instance, SingleCellNet uses multi-class random forest classifiers [148], automated cell type identification using neural networks (ACTINN) uses artificial neural networks [164], scPred uses support vector machine (SVM) [165], and scClassify uses ensemble learning [166] for cell type annotation. Furthermore, ontology-based single cell classification (OnClass) may also accurately annotate cell types absent in the training dataset, through identifying the nearest cell type in low-dimensional embeddings resulting from the Cell Ontology and the unannotated cells [167].

Automated methods for cell type annotation have been applied in a broad range of biomedical studies, including cancer research. However, a recent benchmarking study has demonstrated that every computational method possesses specific advantages over the others under different scenarios [142], making it however difficult for clinical users to select the appropriate tools. Integrating the annotation results from multiple tools may be a solution to the above issue, and probably achieve more accurate cell types annotation. Therefore, ImmCluster has been developed recently for immune cell clustering and annotation, integrating seven reference-based and four marker gene-based computational methods, supported by manually curated marker gene sets [168]. Comparative studies have shown that ImmCluster provides more accurate and stable cell type annotation than individual methods [168].

### Marker gene identification

Marker genes of a particular cell cluster or cell type are an important resource for characterizing its function. In reverse, as shown above, marker genes can also be used for cell type annotation. The typical methods to identify cell cluster/type-specific genes are those to identify differentially expressed genes (DEGs) among the clusters based on statistical tests. For example, the scRNA-seq analysis pipelines Seurat [169] and SINCERA [150] use the nonparametric Wilcoxon's rank-sum test to identify highly expressed genes of specific cell types. It has been shown that Wilcoxon's rank-sum test is of low false positive rates than dedicated methods for sequencing-based DEG analysis [e.g., DESeq2 [170] and empirical analysis of digital gene expression (DGE) in R (edgeR) [171] when the sample size is large [172]]. In addition, the nonparametric Kruskal–Wallis test was adopted in SC3 [133] for comparisons of more than two groups of cells. Considering dropouts in scRNA-seq and differences in gene expression distribution between cell types or status, many other methods have been developed for marker genes identification, such as MAST [82], SCDE [84], and DEsingle [173].

There is one more category of methods, which identify cell-specific genes simultaneously with the process of cell clustering rather than a step thereafter. As introduced in the earlier section, BackSPIN is based on a biclustering approach [135], which clusters highly expressed genes together when clustering cells. Similarly, iterative clustering and guide-gene selection (ICGS) first identifies guide genes by pairwise correlation of expressed genes, and then performs iterative clustering with the guide genes [174]. Moreover, DendroSplit considers marker genes' significance level in identifying sub-clusters [151]. Finally, statistically modeling the distribution of gene expression across individual cells, methods like variance-driven multitask clustering of scRNA-seq data (scVDMC) [175], BPSC [176] and bias-corrected sequencing analysis (BCseq) [177] have been developed to improve both cell subtype identification and differential expression analysis.

Regarding the best choice of DEG tools in scRNA-seq, a recent study compared 36 approaches and found fundamental differences between the methods compared [178]. It has been pointed out that prefiltering of lowly expressed genes may help DEG analysis, and the methods used for bulk RNA-seq analysis in general have comparable performance to those specifically developed for scRNA-seq. Overall, the nonparametric Wilcoxon's rank-sum test ranks high in most application scenarios, except for complex experimental designs.

### Functional enrichment analysis

To facilitate the interpretation and organization of marker genes identified in each cell type, functional enrichment analysis is commonly performed. Computational methods developed for bulk transcriptomics can be easily applied to this analysis, such as Database for Annotation, Visualization, and Integrated Discovery (DAVID) [179]. This kind of analysis requires a hard cutoff on statistical significance to define the marker genes; in contrast, the widely-used gene set enrichment analysis (GSEA) is a cutoff-free approach [180, 181]. GSEA begins with ordering genes based on differential expression statistics between cell populations of interest, followed by statistically assessing if a functionally meaningful gene set or pathway is significantly overrepresented toward the top or bottom of the ranked list. To facilitate GSEA analysis, Molecular Signatures Database (MSigDB) provides a series of annotated gene sets, including pathways and hallmark gene signatures [182].

Besides the above scenarios where the functional annotation is performed based on marker genes or differential expression between two groups of cells, this analysis can also be carried out at the single-cell level. Single sample GSEA (ssGSEA) and gene set variation analysis (GSVA) [183], which are analogues to GSEA and designed for enrichment analysis of single bulk samples, have now been widely used in scRNA-seq to compute signature scores [184, 185]. Besides, accounting for its characteristics in scRNA-seq, more specific tools including Vision [186], Pagoda2 [187], AUCell [188], single-cell signature explorer (SCSE) [189] and jointly assessing signature mean and inferring enrichment (JASMINE) [190] have been proposed, and in general more suitable for signature scoring in scRNA-seq [190]. In addition, these signature-scoring methods can also be used for pathway activity inference [185].

### Trajectory inference and RNA velocity

In addition to the cell-to-cell heterogeneity that can be captured by scRNA-seq, the dynamics of transcriptomes may also reflect the developmental trajectory or cell state transitions. Trajectory inference [191], pseudo-time estimation [192], and RNA velocity modeling [193] are all helpful to reveal molecular characteristics and regulatory mechanisms during cell differentiation or activation.

Trajectory inference is a popular research field in the past years, with approximately a hundred computational tools developed [191], facilitating studies in developmental biology, as well as cancer development and immune response status alterations. Furthermore, applying this category of methods may also facilitate the objective identification of new cell types [194], and the inference of regulatory networks during the development or status transition [188]. According to the types of trajectories, the trajectory inference methods can also be classified into different categories, including linear methods [e.g., SCORPIUS [195], tools for single cell analysis (TSCAN) [196], Wanderlust [197]], bifurcating methods [e.g., diffusion pseudotime (DPT) [198], Wishbone [199]], multifurcation methods [e.g., FateID [200], STEMNET [201], mixtures of factor analysers (MFA) [202]], tree methods (e.g., Slingshot [203], scTite [204], Monocle [205]), and graph methods [e.g., partition-based graph abstraction (PAGA) [206], rare cell type identification (RaceID) [129], selective locally linear inference of cellular expression relationships (SLICER) [207]]. Currently, the trajectory inference methods are maturing, particularly for the linear and bifurcating methods [191]. Based on a recent benchmarking study, guidelines for practical applications are given so that biomedical researchers can choose the appropriate methods according to prior knowledge on the expected topology in the data [191]; otherwise, PAGA, Monocle, RaceID, and Slingshot are recommended for an initial investigation.

Per existing biological knowledge on the starting point of inferred developmental or transition trajectory, cells along the trajectory can be ordered in a pseudo-temporal order. If there are bifurcation, multifurcation, or tree structures in the trajectory, multiple routes should be applied to go through tree branches separately. In this manner, it is easy to investigate gene expression dynamics along the pseudo time. Methods have been developed to conduct the trajectory-/pseudotime-based differential expression analysis [208, 209], which may reveal the dynamic regulation of lineage/status specification.

An alternative way to capture transcriptome dynamics is to use RNA velocity, which is based on the relationship between matured and unmatured transcripts (i.e., with unspliced introns) in the same cell. If there are relatively more unspliced transcripts in a cell, the gene is under upregulation, and vice versa. Jointly quantifying the ratio between matured and unmatured transcripts, and the gene expression changes during status changes, the direction of cell transition can be thus determined [192]. This rationale has been realized in the first RNA velocity method Velocyto [210], and improved in the follow-up method scVelo, where a likelihood-based dynamical model was adopted [211]. Furthermore, recently developed methods [212, 213] have combined RNA velocity with trajectory inference, resulting in directed trajectory inference independent of prior knowledge. For instance, CellRank takes advantage of both the robustness of trajectory inference and the directional information from RNA velocity, enabling the detection of previously unknown trajectories and cell states [212]. CellPath is another method integrating single-cell gene expression dynamics and RNA velocity information for trajectory inference [213].

### Cell–cell communications

CCC events play important roles in organism development and homeostasis, as well as disease generation and progression. For example, tumor microenvironments are complex ecosystems composed of tumor cells, stromal cells and a variety of immune cells, such that abnormal or disrupted communication among these cells may promote tumor growth. To this end, various computational tools have been developed to infer CCC using scRNA-seq data [214]. The communication between cells commonly depends on ligand-receptor (LR) interactions, which are usually quantified by LR co-expression.

To facilitate the above investigation, known ligand-receptor interactions (LRIs) have been manually curated and deposited in databases (Fig. 3a). To date, there are quite a few LRI databases, including CellPhoneDB [215], ICELLNET [216], CellTalkDB [217], SingleCellSignalR [218] and Omnipath [219]. The last updated CellPhoneDB (version 4) includes nearly 2000 high-confidence interactions between ligand and receptor proteins, as well as heteromeric protein complexes [215, 220]. CellTalkDB is another comprehensive LRI database in humans and mouse, including 3398 human LR pairs and 2033 mouse LR pairs [217]. Meanwhile, scRNA-seq data are processed using methods mentioned previously for cell clustering and annotation (Fig. 3b). Integrating the annotated scRNA-seq data with known LRIs, sample-specific LR scores are typically calculated, quantifying the interaction potential. Based on LR co-expression, there are a few categories of LR scoring functions [221], including expression thresholding, expression correlation, expression product, and a combination of differential expression [222]. For example, Camp et al. [223] only considered LR pairings if the expression values of both the ligand and receptor were above a certain threshold [$\log_2(\text{FPKM}) \geq 5$]. By contrast, the method SingleCellSignalR is based on the product of LR gene expression levels [218].

Recently, computational methods for predicting CCC based on scRNA-seq data have been continuously developed [221]. The CCC inference tools can be categorized into three main classes according to their special features (Fig. 3c), that is network-based, machine learning-based and spatial information-based approaches [221]. Network-based approaches, including NicheNet [224], cell–cell communication explorer (CCCExplorer) [225], scConnect [226] and network analysis toolkit for

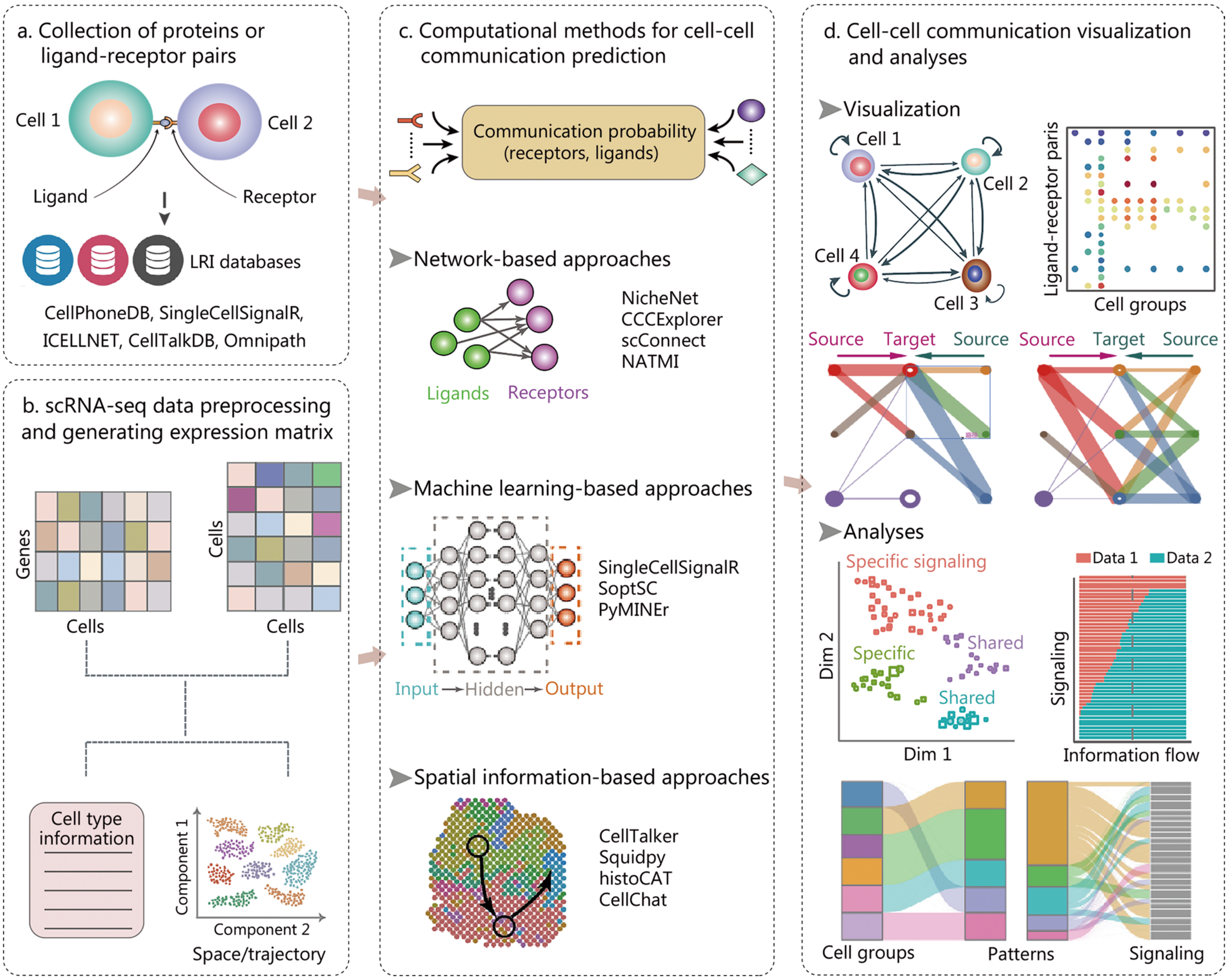

**Fig. 3** The data resources, computational pipelines, and visualization methods used for cell–cell communication (CCC) inference with scRNA-seq data. Typical analysis steps include the collection of ligand-receptor pairs (**a**), cell clustering and annotation in scRNA-seq (**b**), computational prediction of CCC (**c**), followed by results visualization and downstream analysis (**d**). The CCC inference tools can be categorized into three main classes: network-based, machine learning-based and spatial information-based approaches. LRI ligand-receptor interaction, scRNA-seq single-cell RNA sequencing, CCCExplorer cell–cell communication explorer, NATMI network analysis toolkit for multicellular interactions, histoCAT histology topography cytometry analysis toolbox, SoptSC similarity matrix-based optimization for single-cell data analysis, PyMINEr Python maximal information network exploration resource, Squidpy spatial quantification of molecular data in Python

multicellular interactions (NATMI) [227], leverage the connection network between genes to predict CCC. For instance, NicheNet integrates single-cell expression data with prior knowledge of signaling pathways and gene regulatory networks [224], featured by the application of personalized PageRank algorithm, which was used to calculate ligand–target regulatory potential scores [228]. Various types of machine learning algorithms are adopted in the machine learning-based approaches, such as SingleCellSignalR [218], similarity matrix-based optimization for single-cell data analysis (SoptSC) [229] and Python maximal information network exploration resource (PyMINEr) [230]. Besides, reference component analysis (RCA)-CCA [231], linear regression [232] and decision tree classifiers [233] were also used for CCC prediction. Cell localization in space or spatial proximity between cells is the prerequisite of CCC; hence, accounting for spatial information would improve the accuracy of CCC inference. With the rapid development of spatial transcriptomics, many CCC inference approaches integrate scRNA-seq data with spatial transcriptomic and/or image data for identifying CCC. CellTalker scored communication among cell types by counting the number of LRIs, which was then assessed by spatial proximity between cells using image data [234]. In addition, spatial quantification of molecular data in Python (Squidpy) [235] and histology topography cytometry analysis toolbox (histoCAT) [236] provide analysis frameworks for spatial omics data, where intercellular communication can be investigated through cellular proximity or neighborhood analysis. Moreover, the authors of CellChat take the spatial information as the gold standard to evaluate different CCC inference approaches, and showed that CellChat performs better at predicting stronger interactions [237]. Finally, the inference results are usually visualized by heatmap, circus plot, Sankey plot and bubble plot (Fig. 3d).

The emerging computational methods for identifying CCC have improved our understanding of the microenvironment for disease development. However, all the methods depend on prior knowledge of LRIs and statistical or machine learning models to predict potential CCC events. Alternatively choosing LRI resources and prediction approaches may result in different results, yet the impact of the choice on the results is largely unknown. To address this issue, one recent study systematically compared 16 resources and 7 methods for CCC inference, as well as the consensus of the compared methods [214]. The comparison demonstrated that different LRI resources covered a varying fraction of the collective prior knowledge, and the predicted CCC were largely inconsistent with each other, suggesting the need for continued efforts to improve CCC-inference resources and tools.

### Regulon inference and TF activity prediction

TFs play essential roles in gene expression regulation, and are involved in various physiological and pathological processes of humans [238]. It has been realized in scRNA-seq to identify co-expression modules that were directly regulated by TFs of interest, and these modules were defined as regulons [188]. Therefore, it has been made possible to chart the cell type-specific regulons and to reconstruct regulation-based regulatory networks in individual cells (Fig. 4).

One important resource in recognizing regulons is the TF-target databases. The Animal Transcription Factor DataBase (AnimalTFDB) [239], JASPAR [240], transcriptional regulatory relationships unravelled by sentence-based text-mining (TRRUST) [241], KnockTF [242], and Cistrome Data Browser (Cistrome DB) [243] are widely applied TF annotation databases, covering most human and mouse TFs. Based on these databases, a simple way to build cell type-specific transcriptional regulatory networks is to identify up-regulated TFs and/or differentially expressed TF-target genes. For instance, a recent scRNA-seq study identified differentially expressed TFs based on AnimalTFDB TF annotation, and revealed that the reactivation of TFs expressed in fetal epithelium may be the cause of Crohn's disease [244].

Integrating single-cell gene expression and the comprehensive TF-target information, there have been many methods developed for inferring regulons and TF activity. Coexpression analysis, such as weighted gene co-expression network analysis (WGCNA) [245], has been widely used in bulk samples to detect gene modules that likely are regulated by the same TF(s). Recently, this approach has also been applied to scRNA-seq data, to discover, for example, the gene modules whose expression changed significantly over the course of HIV infection [246]. The single cell regulatory network information and clustering (SCENIC) method is the earliest method for regulon inference based on scRNA-seq data [188], and has now been used to study regulatory networks of many diseases such as cancer and COVID-19 [247, 248]. In SCENIC, co-expression modules between TFs and their target genes are first inferred with machine learning methods such as random forest regression, followed by regulon identification through TF's binding motif analysis, and only their direct targets in the co-expression modules are kept to form the regulons. Finally, binarized scores are calculated to indicate TF's activity in each cell. The other methods, including SCODE [249] and SINCERITIES [250], take advantage of the pseudo-temporal information

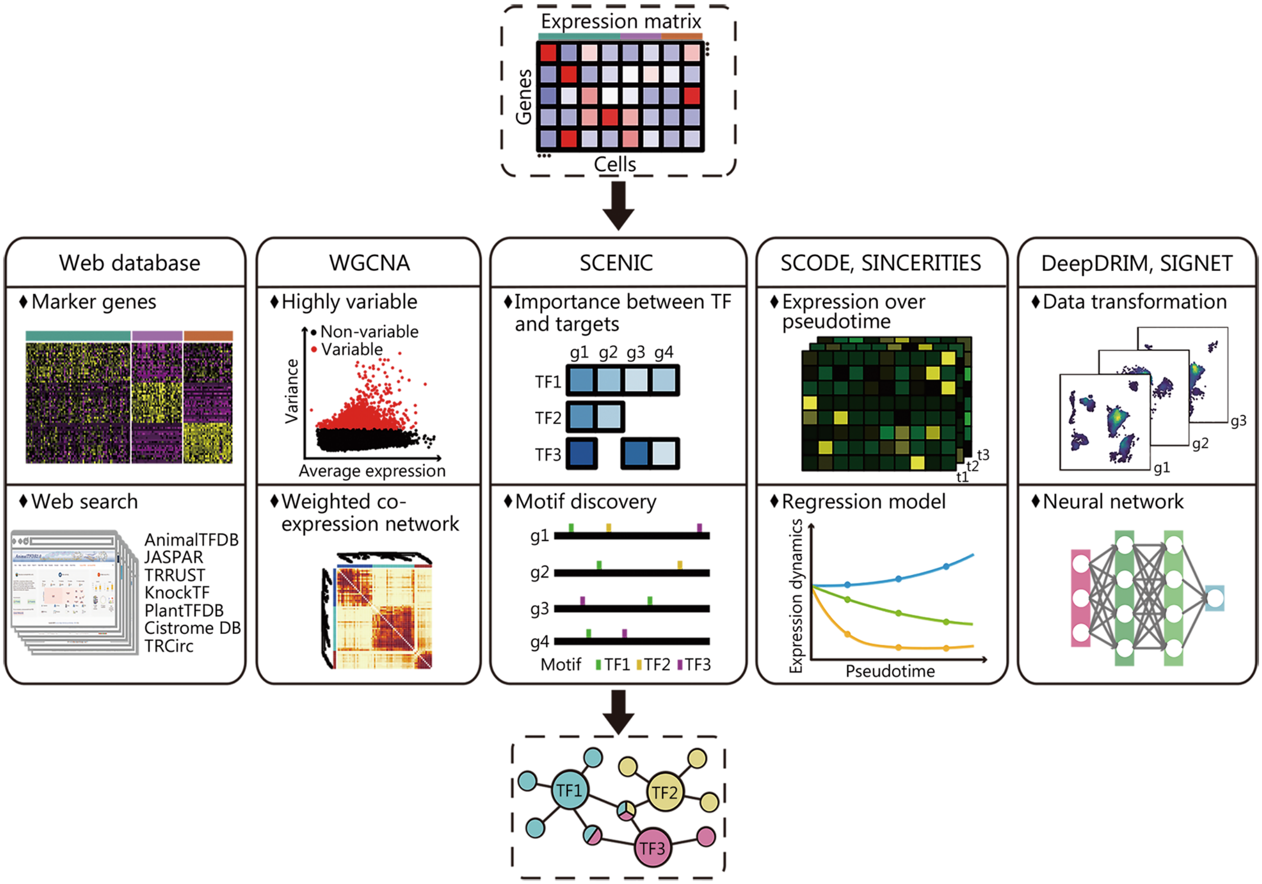

**Fig. 4** Different strategies and approaches developed for regulon inference and TF activity prediction with scRNA-seq. To achieve regulon and TF activity prediction, the TF databases and TF-target databases are important resources, and the computational strategies include co-expression gene module identification, dynamic and stochastic modeling of TF versus target expression changes, and application of machine learning approaches. TF transcription factor, scRNA-seq single-cell RNA sequencing, AnimalTFDB Animal Transcription Factor DataBase, Cistrome DB Cistrome Data Browser, WGCNA weighted gene co-expression network analysis, SCENIC single cell regulatory network information and clustering, TRRUST transcriptional regulatory relationships unravelled by sentence-based text-mining

reconstructed in scRNA-seq and infer TF-target regulatory networks based on ordinary differential equations or stochastic differential equation models. Moreover, machine learning techniques have also been applied for transcriptional regulation analysis. For example, while SIGNET [251] adopts multiple-layer perceptron bagging to identify regulons, DeepDRIM [252] utilizes supervised deep neural network to reconstruct gene regulatory networks. In particular, DeepDRIM is shown to be tolerant to dropout events in scRNA-seq and identify distinct regulatory networks of B cells in COVID-19 patients with mild and severe symptoms.

Despite many methods developed for gene regulation analysis based on scRNA-seq, a rigorous judgment on the inferred results needs to be made, due to the complexity of transcriptional regulation and the insufficient information provided by scRNA-seq data. Performing validation experiments may make the inferred results more solid [253, 254].

### Metabolic analysis

Metabolism is at the core of all biological processes, and metabolic dysregulation is a hallmark of many diseases including cancer, diabetes, and cardiovascular disease [255]. Although single-cell metabolomics technologies are under rapid development, they are now too premature for large-scale applications [256]. Instead, metabolic analysis based on single-cell transcriptomics is a promising alternative approach. For example, researchers may use scRNA-seq to monitor the gene expression changes of key metabolic genes under different treatments [257] or during important physiological/pathological processes [258].

The computational tools for scRNA-seq-based metabolic analysis can be classified into two major categories: pathway-based analysis and flux balance analysis (FBA)-based methods [256] (Fig. 5). For the first category, the standard functional enrichment analysis approaches are generally used (refer to the subsection entitled Functional enrichment analysis). In particular, the R package scMetabolism provides an integrated framework for quantitative analysis of metabolic pathway activity in scRNA-seq, with the ability to account for dropouts, and compatible with multiple tools designed for single-cell functional enrichment analysis [259], including ssGSEA [183, 184], Vision [186], and AUCell [188].

The other category is the FBA-based methods, where constraint-based mathematical models are utilized to systematically simulate metabolism in reconstructed metabolic networks [260]. The reconstruction of metabolic networks is usually based on curated databases, such as Kyoto Encyclopedia of Genes and Genomes (KEGG) [261] and Reactome [262]; thereafter, FBA computes static metabolic fluxes in the system with constraints on the input and output fluxes satisfied [263]. Expression levels of individual enzymes in single cells may not directly affect metabolic fluxes in the networks, because they are mostly dependent on the network topology and constraints [256]. To our knowledge, single-cell flux balance analysis (scFBA) was the first computational tool that combines scRNA-seq data and FBA to estimate single-cell fluxomes [264]. Later, Compass [265] and single-cell flux estimation analysis (scFEA) [255] were proposed. Compass is based on Recon2's reconstruction of human metabolism [266] and solves constraint-based optimization problems with linear programming, to score the potential activity of each metabolic reaction in individual cells [265]. By contrast, scFEA introduces a probabilistic model to consider the flux balance constraints, a multiplayer neural network to model the nonlinearity of flux changes and enzymatic gene expression changes, and a graph neural network to solve the optimization problem [255]. The analysis result by scFEA enables a variety of biologically meaningful downstream analysis, such as cell–cell metabolic communications.

## A collected resource for scRNA-seq data analysis with biomedical applications

With the above overview of the analysis steps and tools for scRNA-seq data, this review may help biomedical researchers to design the data processing and analysis frameworks. However, it would still be challenging for researchers without a bioinformatics background to implement the analysis tasks for their data. For instance, scRNA-seq data analysis requires the installation of specific software tools and running through the scripts written in programming languages such as R and Python. To this end, we collected a range of widely-used software tools in scRNA-seq, and provided practical guidance for installing and running through the analysis with simple commands. The software collection, practical examples, brief description of the analysis results are available at https://github.com/WXlab-NJMU/scrna-recom. Notably, due to time and space constraints, we are unable to incorporate all popular tools into the analysis pipelines on the GitHub site; however, we provide a list of currently available tools with accessible links for users' convenience (Additional file 1: Table S1). We are also open to suggestions from the community and will adjust the pipelines accordingly. Currently, there are still a few research domains in scRNA-seq data analysis that are under positive development, we will keep updating related software and adjusting the scripts to implement the favorable progress made in these research domains.

## Discussion

Focusing on single-cell transcriptomics, we have reviewed almost all respects of typical analysis of scRNA-seq data, ranging from QC, basic data processing, to high-level analysis including trajectory inference, CCC estimation and metabolic analysis. To facilitate researchers conducting the analysis on their data, we have constructed an online software/script repertoire for these analysis steps, and will keep it updated to cover more research scenarios. We also offer a step-by-step command line interface (CLI) for wrapping up the R and Python scripts for scRNA-seq analysis. The step-wise commands can be flexibly combined and tailored for specific applications due to the diversity on scientific questions and experimental design. Moreover, incorporating cutting-edge technologies, the analysis steps reviewed above may not cover every specifically required task. Indeed, additional analysis pipeline (https://github.com/WXlab-NJMU/scPolylox) was necessary to process the scRNA-seq data for identifying *Polylox* transcript variants in lineage tracing [267].

In this review, we did not mention the task for gene expression imputation aiming to alleviate the impact of the well-known dropout issue in scRNA-seq [268]. This is because all the analyses reviewed in this article can be carried out without data imputation, and moreover one comparative study reported that the imputation results did not improve downstream analysis compared to no imputation [269]. Nevertheless, expression data imputation may help when the expression diversity of important genes or gene pairs needs to be investigated [270]. Additionally, the data integration step for removing the effect of covariants can also be optional. For instance, in a complex experimental design where tumor tissues and

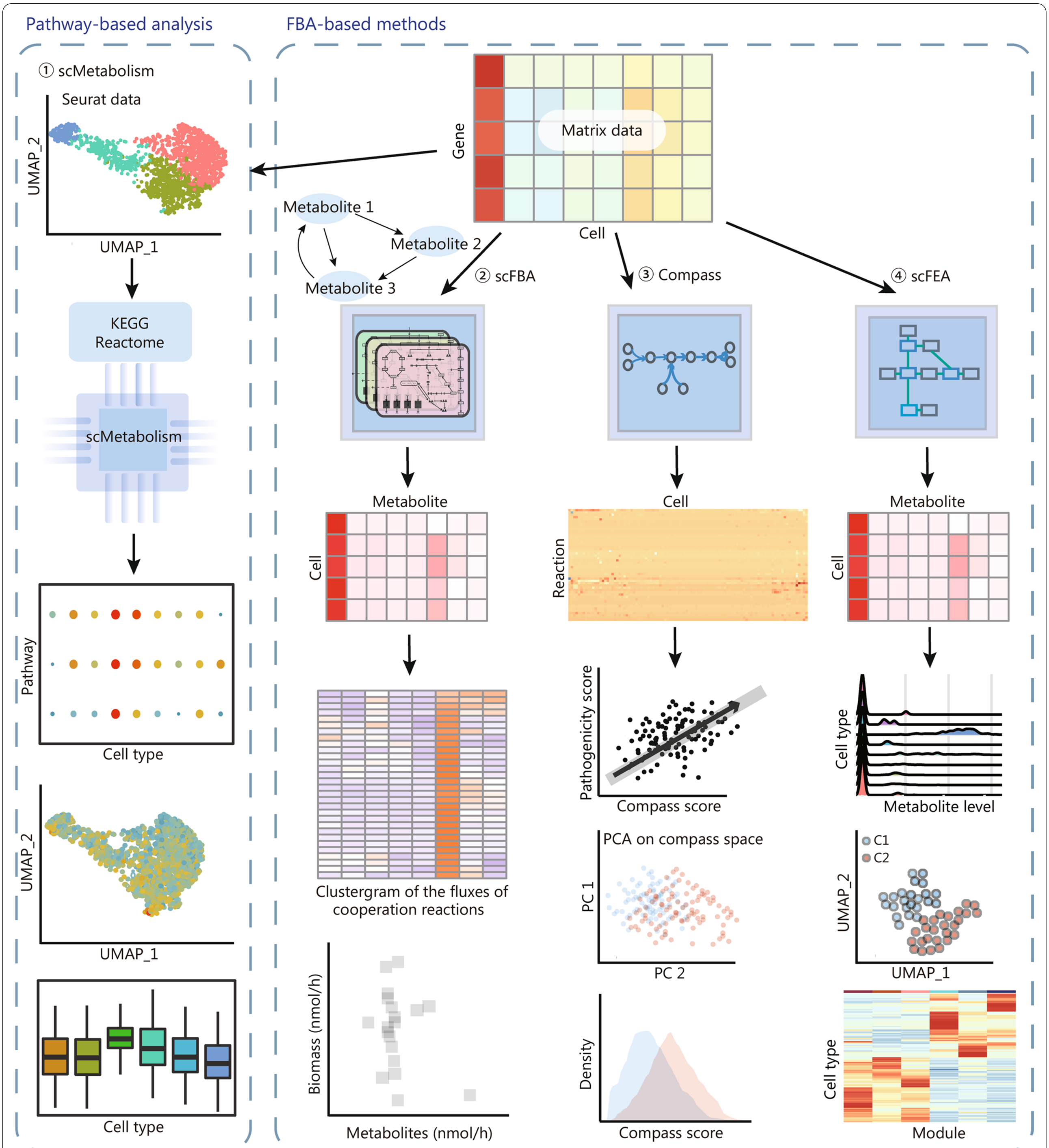

**Fig. 5** Two main types of metabolic analysis within scRNA-seq: pathway-based functional enrichment analysis and flux balance analysis of metabolic flow. While the former makes use of standard functional enrichment analysis, and the latter utilizes constraint-based mathematical models to systematically simulate metabolism in metabolic networks. Methods including scFBA, Compass, and scFEA employed different implementation strategies for flux balance analysis of metabolic flow. FBA flux balance analysis, KEGG Kyoto Encyclopedia of Genes and Genomes, UMAP uniform manifold approximation and projection, scRNA-seq single-cell RNA sequencing, scFBA single-cell flux balance analysis, scFEA single-cell flux estimation analysis, PCA principal component analysis

peritumor tissues are collected from liver cancer patients of different cancer subtypes, the strategies to integrate the datasets may be different depending on whether the common feature of the liver cancer or the subtype-specific feature is interesting.

Previous research has classified the downstream scRNA-seq data analysis methods into cell-level and gene-level analysis [56], which is intuitive and helpful for understanding. While cell-level analysis is typically concerned with the cell composition of given tissues or samples, gene-level analysis focuses on gene expression differences and heterogeneity. As a result, cell clustering for subpopulation identification, trajectory analysis, and CCC inference are examples of cell-level analysis, whereas differential expression, functional enrichment analysis, regulon inference, and metabolic flux analysis are primarily concerned with gene-level information. In contrast to bulk RNA-seq, single-cell RNA-seq allows for cell-level analysis with unprecedented accuracy and throughput, which in turn inspires a few types of gene-level analysis, such as marker gene identification and gene expression dynamics along inferred trajectories.

One more important point in scRNA-seq data analysis is data presentation and interpretation. Although there are no standard protocols for presenting and interpreting the analysis results, these procedures directly link the data with scientific conclusions. In particular, choosing the most appropriate plots would make the message conveyed more straightforwardly. For instance, if one wants to compare the expression levels of a particular gene between tumor and peritumor samples, violin plots showing the two distributions of the expression levels would be more appropriate than t-SNE or UMAP visualizing individual cells with color scales indicating the expression levels. Moreover, using t-SNE or UMAP visualization to compare the composition of cell origins (e.g., from tumor samples or peritumor samples) in a cell subtype of interest might be misleading, although it is more intuitive. This is because massive cells are usually profiled in a scRNA-seq experiment, and consequently cell points can be buried by some others in the two-dimensional visualization. Other types of plots that directly and more quantitatively demonstrate the composition would be more suitable.

Many other aspects of scRNA-seq data analysis are advancing rapidly. ScAPAtrap [271], Sierra [272], dynamic analysis of alternative polyadenylation (APA) from single-cell RNA-seq (scDaPars) [273], SCAPTURE [274], and single cell alternative polyadenylation using expectation–maximization (SCAPE) [275], for example, take advantage of the fact that sequencing reads in 3' tag-based scRNA-seq are distributed near the polyadentation sites of individual transcripts to analyze alternative polyadentation and differential usage of 3'UTR isoforms between cells or cell types. Alternative UTR isoform usage is an important post-transcriptional regulatory mechanism in many physiological and pathological processes, affecting the rate of RNA degradation and the status of translation [276, 277]. Currently, many research groups have been combining scRNA-seq with long-read sequencing technologies to enable high-confidence isoform profiling at the single-cell level [278–280]. Such studies have paved the way for the examination of alternative splicing and transcript fusions between cells and/or cell types, as well as during the progression of diseases [278].

In addition to gene expression regulation by TFs, trans-factors like RNA binding proteins (RBPs) and microRNAs typically bind to the 3'UTR of genes to modulate RNA stability, which also contributes to cellular RNA concentration. Based on collections of RBP and microRNA target genes [281, 282], RBP and microRNA regulons can be investigated similarly to the TF regulons [283] in scRNA-seq. In fact, this kind of co-expression module-based analysis can be extended to the examination of cellular signaling pathway activities. Furthermore, in conjunction with CCC inference [214] and ligand–target regulatory potential scores [224], the activation of certain signaling pathways may also be inferred using scRNA-seq data.

Very recently, Live-seq has been developed to convert scRNA-seq from an end-point type assay to a temporal analysis workflow, by keeping cells alive while extracting RNA from individual cells [284]. It is anticipated that Live-seq will address a number of additional biological questions beyond scRNA-seq. In addition, other sequencing-based single-cell profiling technologies are under rapid development. Aiming at better understanding the dysregulation of altered gene expression in diseases conditions, single-cell assay for transposase-accessible chromatin using sequencing (ATAC-seq) [285], single-cell DNA methylation profiling [286], and single-cell Hi-C [287] are all useful to dissect the underlying regulatory mechanisms from different angles at the single-cell resolution. Algorithms have also been developed to integrate these multimodal single-cell data [63], capable of better resolving cell states and defining novel cell subtypes. Moreover, single-cell multi-omics approaches enable simultaneously profiling a couple of omics in identical cells [288], providing information on both regulatory elements and consequential gene expression levels for individual cells. The datasets generated by these technologies may help biomedical researchers to discover disease-specific regulatory programs, possibly in the subset of certain cell types [289]. Furthermore, although still in the developmental stage, spatial transcriptomics is a promising technique for considering the cellular context in characterizing molecular features of a particular cell [290]. With ever-increasing resolution in spatial transcriptomics, we anticipate gaining more in-depth knowledge in analyzing cell microenvironment and cell–cell interactions in health and disease. Collectively, with technologies continuously advancing, especially those that resolve molecular properties and interactions at the single-cell resolution, we will be able to better understand the pathogenesis of a variety of diseases and enable personalized therapies in the near future.

### Abbreviations

AnimalTFDB: Animal Transcription Factor DataBase; ACTINN: Automated cell type identification using neural networks; APA: Alternative polyadenylation; ATAC-seq: Assay for transposase-accessible chromatin using sequencing; BBKNN: Batch balanced k nearest neighbours; bcds: Binary classification based doublet scoring; Bcseq: Bias-corrected sequencing analysis; CCA: Canonical correlation analysis; CgCorrect: Cell growth correction; CCC: Cell–cell communications; CHETAH: Characterization of cell types aided by hierarchical classification; Cistrome DB: Cistrome Data Browser; CIDR: Clustering through imputation and dimensionality reduction; CLI: Command line interface; COVID-19: Coronavirus disease 2019; cxds: Co-expression based doublet scoring; DAVID: Database for Annotation, Visualization, and Integrated Discovery; DEGs: Differentially expressed genes; DGE: Digital gene expression; DPT: Diffusion pseudotime; EMT: Epithelial-to-mesenchymal transition; EM: Expectation-maximization; f-scLVM: Factorial single-cell latent variable model; FastCAR : Fast correction for ambient RNA; FBA: Flux balance analysis; GSEA: Gene set enrichment analysis; GSVA: Gene set variation analysis; HVGs: Highly variable genes; HistoCAT: Histology topography cytometry analysis toolbox; ICGS: Iterative clustering and guide-gene selection; JASMINE: Jointly assessing signature mean and inferring enrichment; kBET: k-nearest-neighbor batch-effect test; KEGG: Kyoto Encyclopedia of Genes and Genomes; LR: Ligand-receptor; LRI: Ligand-receptor interaction; LIGER: Linked inference of genomic experimental relationships; MAST: Model-based analysis of single-cell transcriptomics; MSigDB: Molecular Signatures Database; MDS: Multi-dimensional scaling; MFA: Mixtures of factor analysers; MNN: Mutual nearest neighbors; NATMI: Network analysis toolkit for multicellular interactions; NMF: Non-negative matrix factorization; OnClass: Ontology-based single cell classification; PAGA : Partition-based graph abstraction; PBMCs: Peripheral blood mononuclear cells; PCA: Principal component analysis; PCs: Principal components; PyMINEr: Python maximal information network exploration resource; QC: Quality control; RaceID: Rare cell type identification; RCA: Reference component analysis; RBPs: RNA binding proteins; SCINA: Semi-supervised category identification and assignment; SARS-CoV-2: Severe acute respiratory syndrome coronavirus 2; SCAPE: Single cell alternative polyadenylation using expectation–maximization; SingleR: Single cell recognition; SCENIC: Single cell regulatory network information and clustering; SCSE: Single-cell signature explorer; ssGSEA: Single sample GSEA; SAME-clustering: Single-cell Aggregated Clustering via Mixture Model Ensemble clustering; scCATCH: Single-cell cluster-based automatic annotation toolkit for cellular heterogeneity; SC3: Single-cell consensus clustering; scDaPars: Dynamic analysis of APA from single-cell RNA-seq; SCDE: Single-cell differential expression; scFBA: Single-cell flux balance analysis; scFEA: Single-cell flux estimation analysis; scLVM: Single-cell latent variable model; Scrublet: Single-cell remover of doublets; scRNA-seq: Single-cell RNA sequencing; scVDMC: Variance-driven multitask clustering of scRNA-seq data; SIMLR: SAingle-cell interpretation via multikernel learning; SLICER: Selective locally linear inference of cellular expression relationships; SNN: Shared nearest neighbor; SoptSC: Similarity matrix-based optimization for single-cell data analysis; Squidpy: Spatial quantification of molecular data in Python; SVM: Support vector machine; TSCAN: Tools for single cell analysis; t-SNE: t-distributed stochastic neighbor embedding; TF: Transcription factor; TPM: Transcript per million; TMM: Trimmed mean of M values; TRRUST: Transcriptional regulatory relationships unravelled by sentence-based text-mining; UMAP: Uniform manifold approximation and projection; UMI: Unique molecular identifier; UQ: Upper quartile; WGCNA: Weighted gene co-expression network analysis; ZIFA: Zero-inflated factor analysis.

## Supplementary Information

The online version contains supplementary material available at https://doi.org/10.1186/s40779-022-00434-8.

**Additional file 1. Table S1**: Tools for analyzing single-cell RNA-seq data, with references and links.

### Acknowledgements

Not applicable.

### Authors' contributions

XW and YSL conceived the project. MS constructed the online repository of software and scripts. TP, QZC, WWZ, YG, GX, QZS, SL, HYY, YZ, and XW collected the literature and drafted the manuscript. MJC, XH, CJJ, SCF, and XL commented on the manuscript. MS, MJC, YSL, and XW revised the manuscript. All authors read and approved the final manuscript.

### Funding

This work was supported by the National Key Research and Development Program of China (2022YFC2702502), the National Natural Science Foundation of China (32170742, 31970646, and 32060152), the Start Fund for Specially Appointed Professor of Jiangsu Province, Hainan Province Science and Technology Special Fund (ZDYF2021SHFZ051), Hainan Provincial Natural Science Foundation of China (820MS053), the Start Fund for High-level Talents of Nanjing Medical University (NMUR2020009), Marshal Initiative Funding of Hainan Medical University (JBGS202103), Hainan Province Clinical Medical Center (QWYH202175), Bioinformatics for Major Diseases Science Innovation Group of Hainan Medical University, and Shenzhen Science and Technology Program, China (JCYJ20210324140407021).

### Availability of data and materials

The online repository of software and wrapped-up command line interface (CLI) is available at https://github.com/WXlab-NJMU/scrna-recom.

## Declarations

### Ethics approval and consent to participate

Not applicable.

### Consent for publication

Not applicable.

### Competing interests

The authors declare that they have no competing interests.

### Author details

[1]State Key Laboratory of Reproductive Medicine, Nanjing Medical University, Nanjing 211166, China. [2]College of Biomedical Information and Engineering, the First Affiliated Hospital of Hainan Medical University, Hainan Medical University, Haikou 571199, Hainan, China. [3]College of Bioinformatics Science and Technology, Harbin Medical University, Harbin 150081, Heilongjiang, China. [4]Department of Immunology, Nanjing Medical University, Nanjing 211166, China. [5]Department of Laboratory Medicine, Women and Children's Hospital of Chongqing Medical University, Chongqing 401174, China. [6]Baylor College of Medicine, Houston, TX 77030, USA. [7]Shenzhen Institute for Advanced Study, University of Electronic Science and Technology of China, Shenzhen 518110, Guangdong, China. [8]School of Biomedical Sciences and Pharmacy, Faculty of Health and Medicine, the University of Newcastle, University Drive, Callaghan, NSW 2308, Australia. [9]Precision Medicine Research Program, Hunter Medical Research Institute, New Lambton Heights, NSW 2305, Australia.

Received: 27 September 2022 Accepted: 18 November 2022
Published online: 02 December 2022

## References

1. Sklavenitis-Pistofidis R, Getz G, Ghobrial I. Single-cell RNA sequencing: one step closer to the clinic. Nat Med. 2021;27(3):375–6.
2. Shapiro E, Biezuner T, Linnarsson S. Single-cell sequencing-based technologies will revolutionize whole-organism science. Nat Rev Genet. 2013;14(9):618–30.
3. Kolodziejczyk AA, Kim JK, Svensson V, Marioni JC, Teichmann SA. The technology and biology of single-cell RNA sequencing. Mol Cell. 2015;58(4):610–20.
4. Nawy T. Single-cell sequencing. Nat Methods. 2014;11(1):18.
5. Griffiths JA, Scialdone A, Marioni JC. Using single-cell genomics to understand developmental processes and cell fate decisions. Mol Syst Biol. 2018;14(4):e8046.
6. Briggs JA, Weinreb C, Wagner DE, Megason S, Peshkin L, Kirschner MW, et al. The dynamics of gene expression in vertebrate embryogenesis at single-cell resolution. Science. 2018;360(6392):eaar5780.
7. Jerby-Arnon L, Shah P, Cuoco MS, Rodman C, Su MJ, Melms JC, et al. A cancer cell program promotes T cell exclusion and resistance to checkpoint blockade. Cell. 2018;175(4):984-97.e24.
8. Kuppe C, Ibrahim MM, Kranz J, Zhang X, Ziegler S, Perales-Paton J, et al. Decoding myofibroblast origins in human kidney fibrosis. Nature. 2021;589(7841):281–6.
9. Bossel Ben-Moshe N, Hen-Avivi S, Levitin N, Yehezkel D, Oosting M, Joosten LaB, et al. Predicting bacterial infection outcomes using single cell RNA-sequencing analysis of human immune cells. Nat Commun. 2019;10(1):3266.
10. Li Y, Jin J, Bai F. Cancer biology deciphered by single-cell transcriptomic sequencing. Protein Cell. 2022;13(3):167–79.
11. Jia Q, Chu H, Jin Z, Long H, Zhu B. High-throughput single-cell sequencing in cancer research. Signal Transduct Target Ther. 2022;7(1):145.
12. Vladoiu MC, El-Hamamy I, Donovan LK, Farooq H, Holgado BL, Sundaravadanam Y, et al. Childhood cerebellar tumours mirror conserved fetal transcriptional programs. Nature. 2019;572(7767):67–73.
13. Blanpain C. Tracing the cellular origin of cancer. Nat Cell Biol. 2013;15(2):126–34.
14. Jin S, Li R, Chen MY, Yu C, Tang LQ, Liu YM, et al. Single-cell transcriptomic analysis defines the interplay between tumor cells, viral infection, and the microenvironment in nasopharyngeal carcinoma. Cell Res. 2020;30(11):950–65.
15. Pastushenko I, Brisebarre A, Sifrim A, Fioramonti M, Revenco T, Boumahdi S, et al. Identification of the tumour transition states occurring during EMT. Nature. 2018;556(7702):463–8.
16. Chung W, Eum HH, Lee HO, Lee KM, Lee HB, Kim KT, et al. Single-cell RNA-seq enables comprehensive tumour and immune cell profiling in primary breast cancer. Nat Commun. 2017;8:15081.
17. Kim J, Koo BK, Knoblich JA. Human organoids: model systems for human biology and medicine. Nat Rev Mol Cell Biol. 2020;21(10):571–84.
18. Wang R, Mao Y, Wang W, Zhou X, Wang W, Gao S, et al. Systematic evaluation of colorectal cancer organoid system by single-cell RNA-Seq analysis. Genome Biol. 2022;23(1):106.
19. Wu H, Uchimura K, Donnelly EL, Kirita Y, Morris SA, Humphreys BD. Comparative analysis and refinement of human PSC-derived kidney organoid differentiation with single-cell transcriptomics. Cell Stem Cell. 2018;23(6):869-81.e8.
20. Neal JT, Li X, Zhu J, Giangarra V, Grzeskowiak CL, Ju J, et al. Organoid modeling of the tumor immune microenvironment. Cell. 2018;175(7):1972-88.e16.
21. Vlachogiannis G, Hedayat S, Vatsiou A, Jamin Y, Fernandez-Mateos J, Khan K, et al. Patient-derived organoids model treatment response of metastatic gastrointestinal cancers. Science. 2018;359(6378):920–6.
22. Broutier L, Mastrogiovanni G, Verstegen MM, Francies HE, Gavarro LM, Bradshaw CR, et al. Human primary liver cancer-derived organoid cultures for disease modeling and drug screening. Nat Med. 2017;23(12):1424–35.
23. Krieger TG, Le Blanc S, Jabs J, Ten FW, Ishaque N, Jechow K, et al. Single-cell analysis of patient-derived PDAC organoids reveals cell state heterogeneity and a conserved developmental hierarchy. Nat Commun. 2021;12(1):5826.
24. Guillen KP, Fujita M, Butterfield AJ, Scherer SD, Bailey MH, Chu Z, et al. A human breast cancer-derived xenograft and organoid platform for drug discovery and precision oncology. Nat Cancer. 2022;3(2):232–50.
25. Ziegler CGK, Allon SJ, Nyquist SK, Mbano IM, Miao VN, Tzouanas CN, et al. SARS-CoV-2 receptor ACE2 is an interferon-stimulated gene in human airway epithelial cells and is detected in specific cell subsets across tissues. Cell. 2020;181(5):1016-35.e19.
26. Stephenson E, Reynolds G, Botting RA, Calero-Nieto FJ, Morgan MD, Tuong ZK, et al. Single-cell multi-omics analysis of the immune response in COVID-19. Nat Med. 2021;27(5):904–16.
27. Tian Y, Carpp LN, Miller HER, Zager M, Newell EW, Gottardo R. Single-cell immunology of SARS-CoV-2 infection. Nat Biotechnol. 2022;40(1):30–41.
28. Melms JC, Biermann J, Huang H, Wang Y, Nair A, Tagore S, et al. A molecular single-cell lung atlas of lethal COVID-19. Nature. 2021;595(7865):114–9.
29. Zhang X, Li T, Liu F, Chen Y, Yao J, Li Z, et al. Comparative analysis of droplet-based ultra-high-throughput single-cell RNA-seq systems. Mol Cell. 2019;73(1):130-42.e5.
30. Wang X, He Y, Zhang Q, Ren X, Zhang Z. Direct comparative analyses of 10x genomics chromium and Smart-seq2. Genomics Proteom Bioinform. 2021;19(2):253–66.
31. Wu F, Fan J, He Y, Xiong A, Yu J, Li Y, et al. Single-cell profiling of tumor heterogeneity and the microenvironment in advanced non-small cell lung cancer. Nat Commun. 2021;12(1):2540.
32. Xu K, Wang R, Xie H, Hu L, Wang C, Xu J, et al. Single-cell RNA sequencing reveals cell heterogeneity and transcriptome profile of breast cancer lymph node metastasis. Oncogenesis. 2021;10(10):66.
33. Haque A, Engel J, Teichmann SA, Lonnberg T. A practical guide to single-cell RNA-sequencing for biomedical research and clinical applications. Genome Med. 2017;9(1):75.
34. Lafzi A, Moutinho C, Picelli S, Heyn H. Tutorial: guidelines for the experimental design of single-cell RNA sequencing studies. Nat Protoc. 2018;13(12):2742–57.
35. Kinker GS, Greenwald AC, Tal R, Orlova Z, Cuoco MS, Mcfarland JM, et al. Pan-cancer single-cell RNA-seq identifies recurring programs of cellular heterogeneity. Nat Genet. 2020;52(11):1208–18.
36. Suva ML, Tirosh I. Single-cell RNA sequencing in cancer: lessons learned and emerging challenges. Mol Cell. 2019;75(1):7–12.
37. Ramachandran P, Matchett KP, Dobie R, Wilson-Kanamori JR, Henderson NC. Single-cell technologies in hepatology: new insights into liver biology and disease pathogenesis. Nat Rev Gastroenterol Hepatol. 2020;17(8):457–72.
38. Ni J, Wang X, Stojanovic A, Zhang Q, Wincher M, Buhler L, et al. Single-cell RNA sequencing of tumor-infiltrating NK cells reveals that inhibition of transcription factor HIF-1α unleashes NK cell activity. Immunity. 2020;52(6):1075-87.e8.
39. Zheng C, Zheng L, Yoo JK, Guo H, Zhang Y, Guo X, et al. Landscape of infiltrating T cells in liver cancer revealed by single-cell sequencing. Cell. 2017;169(7):1342-56.e16.
40. Wilk AJ, Rustagi A, Zhao NQ, Roque J, Martinez-Colon GJ, Mckechnie JL, et al. A single-cell atlas of the peripheral immune response in patients with severe COVID-19. Nat Med. 2020;26(7):1070–6.
41. Wang Z, Xie L, Ding G, Song S, Chen L, Li G, et al. Single-cell RNA sequencing of peripheral blood mononuclear cells from acute Kawasaki disease patients. Nat Commun. 2021;12(1):5444.
42. Clevers H. Modeling development and disease with organoids. Cell. 2016;165(7):1586–97.
43. Salahudeen AA, Choi SS, Rustagi A, Zhu J, van Unen V, de la OS, et al. Progenitor identification and SARS-CoV-2 infection in human distal lung organoids. Nature. 2020;588(7839):670–5.
44. Perez RK, Gordon MG, Subramaniam M, Kim MC, Hartoularos GC, Targ S, et al. Single-cell RNA-seq reveals cell type-specific molecular and genetic associations to lupus. Science. 2022;376(6589):eabf1970.
45. Ernster VL. Nested case-control studies. Prev Med. 1994;23(5):587–90.

46. Mandric I, Schwarz T, Majumdar A, Hou K, Briscoe L, Perez R, et al. Optimized design of single-cell RNA sequencing experiments for cell-type-specific eQTL analysis. Nat Commun. 2020;11(1):5504.
47. Dobin A, Davis CA, Schlesinger F, Drenkow J, Zaleski C, Jha S, et al. STAR: ultrafast universal RNA-seq aligner. Bioinformatics. 2013;29(1):15–21.
48. 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.
49. Zheng GX, Terry JM, Belgrader P, Ryvkin P, Bent ZW, Wilson R, et al. Massively parallel digital transcriptional profiling of single cells. Nat Commun. 2017;8:14049.
50. Tian L, Su S, Dong X, Amann-Zalcenstein D, Biben C, Seidi A, et al. scPipe: a flexible R/Bioconductor preprocessing pipeline for single-cell RNA-sequencing data. PLoS Comput Biol. 2018;14(8):e1006361.
51. Parekh S, Ziegenhain C, Vieth B, Enard W, Hellmann I. zUMIs—a fast and flexible pipeline to process RNA sequencing data with UMIs. Gigascience. 2018;7(6):giy059.
52. Hashimshony T, Senderovich N, Avital G, Klochendler A, de Leeuw Y, Anavy L, et al. CEL-Seq2: sensitive highly-multiplexed single-cell RNA-Seq. Genome Biol. 2016;17:77.
53. Melsted P, Booeshaghi AS, Liu L, Gao F, Lu L, Min KHJ, et al. Modular, efficient and constant-memory single-cell RNA-seq preprocessing. Nat Biotechnol. 2021;39(7):813–8.
54. Wang Z, Hu J, Johnson WE, Campbell JD. scruff: an R/Bioconductor package for preprocessing single-cell RNA-sequencing data. BMC Bioinform. 2019;20(1):222.
55. You Y, Tian L, Su S, Dong X, Jabbari JS, Hickey PF, et al. Benchmarking UMI-based single-cell RNA-seq preprocessing workflows. Genome Biol. 2021;22(1):339.
56. Luecken MD, Theis FJ. Current best practices in single-cell RNA-seq analysis: a tutorial. Mol Syst Biol. 2019;15(6):e8746.
57. Stegle O, Teichmann SA, Marioni JC. Computational and analytical challenges in single-cell transcriptomics. Nat Rev Genet. 2015;16(3):133–45.
58. Brennecke P, Anders S, Kim JK, Kolodziejczyk AA, Zhang X, Proserpio V, et al. Accounting for technical noise in single-cell RNA-seq experiments. Nat Methods. 2013;10(11):1093–5.
59. Andrews TS, Kiselev VY, Mccarthy D, Hemberg M. Tutorial: guidelines for the computational analysis of single-cell RNA sequencing data. Nat Protoc. 2021;16(1):1–9.
60. Ilicic T, Kim JK, Kolodziejczyk AA, Bagger FO, Mccarthy DJ, Marioni JC, et al. Classification of low quality cells from single-cell RNA-seq data. Genome Biol. 2016;17:29.
61. Butler A, Hoffman P, Smibert P, Papalexi E, Satija R. Integrating single-cell transcriptomic data across different conditions, technologies, and species. Nat Biotechnol. 2018;36(5):411–20.
62. Stuart T, Butler A, Hoffman P, Hafemeister C, Papalexi E, Mauck WM 3rd, et al. Comprehensive integration of single-cell data. Cell. 2019;177(7):1888-902.e21.
63. Hao Y, Hao S, Andersen-Nissen E, Mauck WM 3rd, Zheng S, Butler A, et al. Integrated analysis of multimodal single-cell data. Cell. 2021;184(13):3573-87.e29.
64. Mccarthy DJ, Campbell KR, Lun AT, Wills QF. Scater: pre-processing, quality control, normalization and visualization of single-cell RNA-seq data in R. Bioinformatics. 2017;33(8):1179–86.
65. Guimaraes JC, Zavolan M. Patterns of ribosomal protein expression specify normal and malignant human cells. Genome Biol. 2016;17(1):236.
66. Oelen R, de Vries DH, Brugge H, Gordon MG, Vochteloo M, Ye CJ, et al. Single-cell RNA-sequencing of peripheral blood mononuclear cells reveals widespread, context-specific gene expression regulation upon pathogenic exposure. Nat Commun. 2022;13(1):3267.
67. Zhong S, Ding W, Sun L, Lu Y, Dong H, Fan X, et al. Decoding the development of the human hippocampus. Nature. 2020;577(7791):531–6.
68. Young MD, Behjati S. SoupX removes ambient RNA contamination from droplet-based single-cell RNA sequencing data. Gigascience. 2020;9(12):giaa151.
69. Yang S, Corbett SE, Koga Y, Wang Z, Johnson WE, Yajima M, et al. Decontamination of ambient RNA in single-cell RNA-seq with DecontX. Genome Biol. 2020;21(1):57.
70. Berg M, Petoukhov I, Van Den Ende I, Meyer KB, Guryev V, Vonk JM, et al. FastCAR: fast correction for Ambient RNA to facilitate differential gene expression analysis in single-cell RNA-sequencing datasets. bioRxiv. 2022. https://doi.org/10.1101/2022.07.19.500594
71. Fleming SJ, Chaffin MD, Arduini A, Akkad AD, Banks E, Marioni JC, et al. Unsupervised removal of systematic background noise from droplet-based single-cell experiments using CellBender. bioRxiv. 2022. https://doi.org/10.1101/791699.
72. Xi NM, Li JJ. Benchmarking computational doublet-detection methods for single-cell RNA sequencing data. Cell Syst. 2021;12(2):176-94.e6.
73. Bernstein NJ, Fong NL, Lam I, Roy MA, Hendrickson DG, Kelley DR. Solo: doublet identification in single-cell RNA-Seq via semi-supervised deep learning. Cell Syst. 2020;11(1):95-101.e5.
74. Wolock SL, Lopez R, Klein AM. Scrublet: computational identification of cell doublets in single-cell transcriptomic data. Cell Syst. 2019;8(4):281-91.e9.
75. 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.
76. Bais AS, Kostka D. scds: computational annotation of doublets in single-cell RNA sequencing data. Bioinformatics. 2020;36(4):1150–8.
77. Park J, Choi W, Tiesmeyer S, Long B, Borm LE, Garren E, et al. Cell segmentation-free inference of cell types from in situ transcriptomics data. Nat Commun. 2021;12(1):3545.
78. McGinnis CS, Murrow LM, Gartner ZJ. DoubletFinder: doublet detection in single-cell RNA sequencing data using artificial nearest neighbors. Cell Syst. 2019;8(4):329-37.e4.
79. DePasquale EAK, Schnell DJ, Van Camp PJ, Valiente-Alandi I, Blaxall BC, Grimes HL, et al. DoubletDecon: deconvoluting doublets from single-cell RNA-sequencing data. Cell Rep. 2019;29(6):1718-27.e8.
80. Deeke JM, Gagnon-Bartsch JA. Stably expressed genes in single-cell RNA sequencing. J Bioinform Comput Biol. 2020;18(1):2040004.
81. Vallejos CA, Risso D, Scialdone A, Dudoit S, Marioni JC. Normalizing single-cell RNA sequencing data: challenges and opportunities. Nat Methods. 2017;14(6):565–71.
82. Finak G, Mcdavid A, Yajima M, Deng J, Gersuk V, Shalek AK, et al. MAST: a flexible statistical framework for assessing transcriptional changes and characterizing heterogeneity in single-cell RNA sequencing data. Genome Biol. 2015;16:278.
83. Grun D, van Oudenaarden A. Design and analysis of single-cell sequencing experiments. Cell. 2015;163(4):799–810.
84. Kharchenko PV, Silberstein L, Scadden DT. Bayesian approach to single-cell differential expression analysis. Nat Methods. 2014;11(7):740–2.
85. Li B, Ruotti V, Stewart RM, Thomson JA, Dewey CN. RNA-Seq gene expression estimation with read mapping uncertainty. Bioinformatics. 2010;26(4):493–500.
86. Bullard JH, Purdom E, Hansen KD, Dudoit S. Evaluation of statistical methods for normalization and differential expression in mRNA-Seq experiments. BMC Bioinform. 2010;11:94.
87. Robinson MD, Oshlack A. A scaling normalization method for differential expression analysis of RNA-seq data. Genome Biol. 2010;11(3):R25.
88. Anders S, Huber W. Differential expression analysis for sequence count data. Genome Biol. 2010;11(10):R106.
89. 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.
90. Qiu X, Hill A, Packer J, Lin D, Ma YA, Trapnell C. Single-cell mRNA quantification and differential analysis with Census. Nat Methods. 2017;14(3):309–15.
91. Buttner M, Miao Z, Wolf FA, Teichmann SA, Theis FJ. A test metric for assessing single-cell RNA-seq batch correction. Nat Methods. 2019;16(1):43–9.
92. Hafemeister C, Satija R. Normalization and variance stabilization of single-cell RNA-seq data using regularized negative binomial regression. Genome Biol. 2019;20(1):296.
93. Buettner F, Natarajan KN, Casale FP, Proserpio V, Scialdone A, Theis FJ, et al. Computational analysis of cell-to-cell heterogeneity in single-cell RNA-sequencing data reveals hidden subpopulations of cells. Nat Biotechnol. 2015;33(2):155–60.

94. Vento-Tormo R, Efremova M, Botting RA, Turco MY, Vento-Tormo M, Meyer KB, et al. Single-cell reconstruction of the early maternal-fetal interface in humans. Nature. 2018;563(7731):347–53.
95. Macosko EZ, Basu A, Satija R, Nemesh J, Shekhar K, Goldman M, et al. Highly parallel genome-wide expression profiling of individual cells using nanoliter droplets. Cell. 2015;161(5):1202–14.
96. Buettner F, Pratanwanich N, McCarthy DJ, Marioni JC, Stegle O. f-scLVM: scalable and versatile factor analysis for single-cell RNA-seq. Genome Biol. 2017;18(1):212.
97. Blasi T, Buettner F, Strasser MK, Marr C, Theis FJ. cgCorrect: a method to correct for confounding cell–cell variation due to cell growth in single-cell transcriptomics. Phys Biol. 2017;14(3): 036001.
98. Kanton S, Boyle MJ, He Z, Santel M, Weigert A, Sanchis-Calleja F, et al. Organoid single-cell genomic atlas uncovers human-specific features of brain development. Nature. 2019;574(7778):418–22.
99. Tran HTN, Ang KS, Chevrier M, Zhang X, Lee NYS, Goh M, et al. A benchmark of batch-effect correction methods for single-cell RNA sequencing data. Genome Biol. 2020;21(1):12.
100. Johnson WE, Li C, Rabinovic A. Adjusting batch effects in microarray expression data using empirical Bayes methods. Biostatistics. 2007;8(1):118–27.
101. Smyth GK, Speed T. Normalization of cDNA microarray data. Methods. 2003;31(4):265–73.
102. Haghverdi L, Lun ATL, Morgan MD, Marioni JC. Batch effects in single-cell RNA-sequencing data are corrected by matching mutual nearest neighbors. Nat Biotechnol. 2018;36(5):421–7.
103. Hie B, Bryson B, Berger B. Efficient integration of heterogeneous single-cell transcriptomes using Scanorama. Nat Biotechnol. 2019;37(6):685–91.
104. Polański K, Young MD, Miao Z, Meyer KB, Teichmann SA, Park JE. BBKNN: fast batch alignment of single cell transcriptomes. Bioinformatics. 2020;36(3):964–5.
105. Korsunsky I, Millard N, Fan J, Slowikowski K, Zhang F, Wei K, et al. Fast, sensitive and accurate integration of single-cell data with Harmony. Nat Methods. 2019;16(12):1289–96.
106. Welch JD, Kozareva V, Ferreira A, Vanderburg C, Martin C, Macosko EZ. Single-cell multi-omic integration compares and contrasts features of brain cell identity. Cell. 2019;177(7):1873-87.e17.
107. Lotfollahi M, Wolf FA, Theis FJ. scGen predicts single-cell perturbation responses. Nat Methods. 2019;16(8):715–21.
108. Argelaguet R, Cuomo ASE, Stegle O, Marioni JC. Computational principles and challenges in single-cell data integration. Nat Biotechnol. 2021;39(10):1202–15.
109. Grun D, Kester L, Van Oudenaarden A. Validation of noise models for single-cell transcriptomics. Nat Methods. 2014;11(6):637–40.
110. Su K, Yu T, Wu H. Accurate feature selection improves single-cell RNA-seq cell clustering. Brief Bioinform. 2021;22(5):bbab034.
111. Townes FW, Hicks SC, Aryee MJ, Irizarry RA. Feature selection and dimension reduction for single-cell RNA-Seq based on a multinomial model. Genome Biol. 2019;20(1):295.
112. Yang P, Huang H, Liu C. Feature selection revisited in the single-cell era. Genome Biol. 2021;22(1):321.
113. Yip SH, Sham PC, Wang J. Evaluation of tools for highly variable gene discovery from single-cell RNA-seq data. Brief Bioinform. 2019;20(4):1583–9.
114. Andrews TS, Hemberg M. M3Drop: dropout-based feature selection for scRNASeq. Bioinformatics. 2019;35(16):2865–7.
115. Klein AM, Mazutis L, Akartuna I, Tallapragada N, Veres A, Li V, et al. Droplet barcoding for single-cell transcriptomics applied to embryonic stem cells. Cell. 2015;161(5):1187–201.
116. Sun S, Zhu J, Ma Y, Zhou X. Accuracy, robustness and scalability of dimensionality reduction methods for single-cell RNA-seq analysis. Genome Biol. 2019;20(1):269.
117. Ringner M. What is principal component analysis? Nat Biotechnol. 2008;26(3):303–4.
118. Shao C, Hofer T. Robust classification of single-cell transcriptome data by nonnegative matrix factorization. Bioinformatics. 2017;33(2):235–42.
119. Tzeng J, Lu HH, Li WH. Multidimensional scaling for large genomic data sets. BMC Bioinform. 2008;9:179.
120. Kobak D, Berens P. The art of using t-SNE for single-cell transcriptomics. Nat Commun. 2019;10(1):5416.
121. Becht E, Mcinnes L, Healy J, Dutertre CA, Kwok IWH, Ng LG, et al. Dimensionality reduction for visualizing single-cell data using UMAP. Nat Biotechnol. 2019;37(1):38–44.
122. Gogolewski K, Sykulski M, Chung NC, Gambin A. Truncated robust principal component analysis and noise reduction for single cell RNA sequencing data. J Comput Biol. 2019;26(8):782–93.
123. Tsuyuzaki K, Sato H, Sato K, Nikaido I. Benchmarking principal component analysis for large-scale single-cell RNA-sequencing. Genome Biol. 2020;21(1):9.
124. Chung NC, Storey JD. Statistical significance of variables driving systematic variation in high-dimensional data. Bioinformatics. 2015;31(4):545–54.
125. Pierson E, Yau C. ZIFA: dimensionality reduction for zero-inflated single-cell gene expression analysis. Genome Biol. 2015;16:241.
126. Shi J, Luo Z. Nonlinear dimensionality reduction of gene expression data for visualization and clustering analysis of cancer tissue samples. Comput Biol Med. 2010;40(8):723–32.
127. Petegrosso R, Li Z, Kuang R. Machine learning and statistical methods for clustering single-cell RNA-sequencing data. Brief Bioinform. 2020;21(4):1209–23.
128. van Unen V, Li N, Molendijk I, Temurhan M, Hollt T, van der Meulen-de Jong AE, et al. Mass cytometry of the human mucosal immune system identifies tissue- and disease-associated immune subsets. Immunity. 2016;44(5):1227–39.
129. Grun D, Lyubimova A, Kester L, Wiebrands K, Basak O, Sasaki N, et al. Single-cell messenger RNA sequencing reveals rare intestinal cell types. Nature. 2015;525(7568):251–5.
130. Jaitin DA, Kenigsberg E, Keren-Shaul H, Elefant N, Paul F, Zaretsky I, et al. Massively parallel single-cell RNA-seq for marker-free decomposition of tissues into cell types. Science. 2014;343(6172):776–9.
131. Zhang W, Xue X, Zheng X, Fan Z. NMFLRR: clustering scRNA-Seq Data by integrating nonnegative matrix factorization with low rank representation. IEEE J Biomed Health Inform. 2022;26(3):1394–405.
132. Zheng R, Li M, Liang Z, Wu FX, Pan Y, Wang J. SinNLRR: a robust subspace clustering method for cell type detection by non-negative and low-rank representation. Bioinformatics. 2019;35(19):3642–50.
133. Kiselev VY, Kirschner K, Schaub MT, Andrews T, Yiu A, Chandra T, et al. SC3: consensus clustering of single-cell RNA-seq data. Nat Methods. 2017;14(5):483–6.
134. Levine JH, Simonds EF, Bendall SC, Davis KL, El Amir AD, Tadmor MD, et al. Data-driven phenotypic dissection of AML reveals progenitor-like cells that correlate with prognosis. Cell. 2015;162(1):184–97.
135. Zeisel A, Munoz-Manchado AB, Codeluppi S, Lonnerberg P, La Manno G, Jureus A, et al. Brain structure. Cell types in the mouse cortex and hippocampus revealed by single-cell RNA-seq. Science. 2015;347(6226):1138–42.
136. Wang B, Zhu J, Pierson E, Ramazzotti D, Batzoglou S. Visualization and analysis of single-cell RNA-seq data by kernel-based similarity learning. Nat Methods. 2017;14(4):414–6.
137. 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.
138. Huh R, Yang Y, Jiang Y, Shen Y, Li Y. SAME-clustering: single-cell aggregated clustering via mixture model ensemble. Nucleic Acids Res. 2020;48(1):86–95.
139. Duo A, Robinson MD, Soneson C. A systematic performance evaluation of clustering methods for single-cell RNA-seq data. F1000Res. 2018;7:1141.
140. Freytag S, Tian L, Lonnstedt I, Ng M, Bahlo M. Comparison of clustering tools in R for medium-sized 10x Genomics single-cell RNA-sequencing data. F1000Res. 2018;7:1297.
141. Sun X, Lin X, Li Z, Wu H. A comprehensive comparison of supervised and unsupervised methods for cell type identification in single-cell RNA-seq. Brief Bioinform. 2022;23(2):bbab567.
142. Abdelaal T, Michielsen L, Cats D, Hoogduin D, Mei H, Reinders MJT, et al. A comparison of automatic cell identification methods for single-cell RNA sequencing data. Genome Biol. 2019;20(1):194.

143. Huang Q, Liu Y, Du Y, Garmire LX. Evaluation of cell type annotation R packages on single-cell RNA-seq data. Genom Proteom Bioinform. 2021;19(2):267–81.
144. Zhang AW, O'flanagan C, Chavez EA, Lim JLP, Ceglia N, Mcpherson A, et al. Probabilistic cell-type assignment of single-cell RNA-seq for tumor microenvironment profiling. Nat Methods. 2019;16(10):1007–15.
145. Kiselev VY, Yiu A, Hemberg M. scmap: projection of single-cell RNA-seq data across data sets. Nat Methods. 2018;15(5):359–62.
146. Aran D, Looney AP, Liu L, Wu E, Fong V, Hsu A, et al. Reference-based analysis of lung single-cell sequencing reveals a transitional profibrotic macrophage. Nat Immunol. 2019;20(2):163–72.
147. de Kanter JK, Lijnzaad P, Candelli T, Margaritis T, Holstege FCP. CHETAH: a selective, hierarchical cell type identification method for single-cell RNA sequencing. Nucleic Acids Res. 2019;47(16):e95.
148. 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):207-13.e2.
149. Jiang L, Chen H, Pinello L, Yuan GC. GiniClust: detecting rare cell types from single-cell gene expression data with Gini index. Genome Biol. 2016;17(1):144.
150. Guo M, Wang H, Potter SS, Whitsett JA, Xu Y. SINCERA: a pipeline for single-cell RNA-Seq profiling analysis. PLoS Comput Biol. 2015;11(11):e1004575.
151. Zhang JM, Fan J, Fan HC, Rosenfeld D, Tse DN. An interpretable framework for clustering single-cell RNA-seq datasets. BMC Bioinform. 2018;19(1):93.
152. Pasquini G, Rojo Arias JE, Schäfer P, Busskamp V. Automated methods for cell type annotation on scRNA-seq data. Comput Struct Biotechnol J. 2021;19:961–9.
153. Zhang X, Lan Y, Xu J, Quan F, Zhao E, Deng C, et al. Cell Marker: a manually curated resource of cell markers in human and mouse. Nucleic Acids Res. 2019;47(D1):D721–8.
154. Franzén O, Gan LM, Björkegren JLM. Panglao DB: a web server for exploration of mouse and human single-cell RNA sequencing data. Database. 2019;2019:baz046.
155. Xu M, Bai X, Ai B, Zhang G, Song C, Zhao J, et al. TF-Marker: a comprehensive manually curated database for transcription factors and related markers in specific cell and tissue types in human. Nucleic Acids Res. 2022;50(D1):D402–12.
156. Ianevski A, Giri AK, Aittokallio T. Fully-automated and ultra-fast cell-type identification using specific marker combinations from single-cell transcriptomic data. Nat Commun. 2022;13(1):1246.
157. Guo H, Li J. scSorter: assigning cells to known cell types according to marker genes. Genome Biol. 2021;22(1):69.
158. Zhang Z, Luo D, Zhong X, Choi JH, Ma Y, Wang S, et al. SCINA: a semi-supervised subtyping algorithm of single cells and bulk samples. Genes. 2019;10(7):531.
159. Shao X, Liao J, Lu X, Xue R, Ai N, Fan X. scCATCH: automatic annotation on cell types of clusters from single-cell RNA sequencing data. iScience. 2020;23(3):100882.
160. Hou R, Denisenko E, Forrest ARR. scMatch: a single-cell gene expression profile annotation tool using reference datasets. Bioinformatics. 2019;35(22):4688–95.
161. Stunnenberg HG, International Human Epigenome C, Hirst M. The international human epigenome consortium: a blueprint for scientific collaboration and discovery. Cell. 2016;167(5):1145–9.
162. Consortium EP. An integrated encyclopedia of DNA elements in the human genome. Nature. 2012;489(7414):57–74.
163. Mabbott NA, Baillie JK, Brown H, Freeman TC, Hume DA. An expression atlas of human primary cells: inference of gene function from coexpression networks. BMC Genomics. 2013;14:632.
164. Ma F, Pellegrini M. ACTINN: automated identification of cell types in single cell RNA sequencing. Bioinformatics. 2020;36(2):533–8.
165. Alquicira-Hernandez J, Sathe A, Ji HP, Nguyen Q, Powell JE. scPred: accurate supervised method for cell-type classification from single-cell RNA-seq data. Genome Biol. 2019;20(1):264.
166. Lin Y, Cao Y, Kim HJ, Salim A, Speed TP, Lin DM, et al. scClassify: sample size estimation and multiscale classification of cells using single and multiple reference. Mol Syst Biol. 2020;16(6): e9389.
167. Wang S, Pisco AO, Mcgeever A, Brbic M, Zitnik M, Darmanis S, et al. Leveraging the cell ontology to classify unseen cell types. Nat Commun. 2021;12(1):5556.
168. Jiang T, Zhou W, Sheng Q, Yu J, Xie Y, Ding N, et al. ImmCluster: an ensemble resource for immunology cell type clustering and annotations in normal and cancerous tissues. Nucleic Acids Res. 2022;22:gkac922.
169. Satija R, Farrell JA, Gennert D, Schier AF, Regev A. Spatial reconstruction of single-cell gene expression data. Nat Biotechnol. 2015;33(5):495–502.
170. 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.
171. 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.
172. Li Y, Ge X, Peng F, Li W, Li JJ. Exaggerated false positives by popular differential expression methods when analyzing human population samples. Genome Biol. 2022;23(1):79.
173. Miao Z, Deng K, Wang X, Zhang X. DEsingle for detecting three types of differential expression in single-cell RNA-seq data. Bioinformatics. 2018;34(18):3223–4.
174. Olsson A, Venkatasubramanian M, Chaudhri VK, Aronow BJ, Salomonis N, Singh H, et al. Single-cell analysis of mixed-lineage states leading to a binary cell fate choice. Nature. 2016;537(7622):698–702.
175. Zhang H, Lee CaA, Li Z, Garbe JR, Eide CR, Petegrosso R, et al. A multitask clustering approach for single-cell RNA-seq analysis in Recessive Dystrophic Epidermolysis Bullosa. PLoS Comput Biol. 2018;14(4):e1006053.
176. Vu TN, Wills QF, Kalari KR, Niu N, Wang L, Rantalainen M, et al. Beta-Poisson model for single-cell RNA-seq data analyses. Bioinformatics. 2016;32(14):2128–35.
177. Chen L, Zheng S. BCseq: accurate single cell RNA-seq quantification with bias correction. Nucleic Acids Res. 2018;46(14):e82.
178. Soneson C, Robinson MD. Bias, robustness and scalability in single-cell differential expression analysis. Nat Methods. 2018;15(4):255–61.
179. Dennis G Jr, Sherman BT, Hosack DA, Yang J, Gao W, Lane HC, et al. DAVID: Database for Annotation, Visualization, and Integrated Discovery. Genome Biol. 2003;4(5):P3.
180. Subramanian A, Tamayo P, Mootha VK, Mukherjee S, Ebert BL, Gillette MA, et al. Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles. Proc Natl Acad Sci USA. 2005;102(43):15545–50.
181. Wang X, Cairns MJ. SeqGSEA: a Bioconductor package for gene set enrichment analysis of RNA-Seq data integrating differential expression and splicing. Bioinformatics. 2014;30(12):1777–9.
182. Liberzon A, Birger C, Thorvaldsdottir H, Ghandi M, Mesirov JP, Tamayo P. The Molecular Signatures Database (MSigDB) hallmark gene set collection. Cell Syst. 2015;1(6):417–25.
183. Hanzelmann S, Castelo R, Guinney J. GSVA: gene set variation analysis for microarray and RNA-seq data. BMC Bioinformatics. 2013;14:7.
184. Jin Y, Wang Z, He D, Zhu Y, Chen X, Cao K. Identification of novel subtypes based on ssGSEA in immune-related prognostic signature for tongue squamous cell carcinoma. Cancer Med. 2021;10(23):8693–707.
185. Zhang Y, Ma Y, Huang Y, Zhang Y, Jiang Q, Zhou M, et al. Benchmarking algorithms for pathway activity transformation of single-cell RNA-seq data. Comput Struct Biotechnol J. 2020;18:2953–61.
186. Detomaso D, Jones MG, Subramaniam M, Ashuach T, Ye CJ, Yosef N. Functional interpretation of single cell similarity maps. Nat Commun. 2019;10(1):4376.
187. Fan J, Salathia N, Liu R, Kaeser GE, Yung YC, Herman JL, et al. Characterizing transcriptional heterogeneity through pathway and gene set overdispersion analysis. Nat Methods. 2016;13(3):241–4.
188. Aibar S, González-Blas CB, Moerman T, Huynh-Thu VA, Imrichova H, Hulselmans G, et al. SCENIC: single-cell regulatory network inference and clustering. Nat Methods. 2017;14(11):1083–6.
189. Pont F, Tosolini M, Fournie JJ. Single-Cell Signature Explorer for comprehensive visualization of single cell signatures across scRNA-seq datasets. Nucleic Acids Res. 2019;47(21):e133.

190. Noureen N, Ye Z, Chen Y, Wang X, Zheng S. Signature-scoring methods developed for bulk samples are not adequate for cancer single-cell RNA sequencing data. Elife. 2022;11:e71994.
191. Saelens W, Cannoodt R, Todorov H, Saeys Y. A comparison of single-cell trajectory inference methods. Nat Biotechnol. 2019;37(5):547–54.
192. Ding J, Sharon N, Bar-Joseph Z. Temporal modelling using single-cell transcriptomics. Nat Rev Genet. 2022;23(6):355–68.
193. Bergen V, Soldatov RA, Kharchenko PV, Theis FJ. RNA velocity-current challenges and future perspectives. Mol Syst Biol. 2021;17(8):e10282.
194. Schlitzer A, Sivakamasundari V, Chen J, Sumatoh HR, Schreuder J, Lum J, et al. Identification of cDC1- and cDC2-committed DC progenitors reveals early lineage priming at the common DC progenitor stage in the bone marrow. Nat Immunol. 2015;16(7):718–28.
195. Cannoodt R, Saelens W, Sichien D, Tavernier S, Janssens S, Guilliams M, et al. SCORPIUS improves trajectory inference and identifies novel modules in dendritic cell development. bioRxiv. 2016. https://doi.org/10.1101/079509
196. Ji Z, Ji H. TSCAN: pseudo-time reconstruction and evaluation in single-cell RNA-seq analysis. Nucleic Acids Res. 2016;44(13):e117.
197. Bendall SC, Davis KL, El Amir AD, Tadmor MD, Simonds EF, Chen TJ, et al. Single-cell trajectory detection uncovers progression and regulatory coordination in human B cell development. Cell. 2014;157(3):714–25.
198. Haghverdi L, Buttner M, Wolf FA, Buettner F, Theis FJ. Diffusion pseudotime robustly reconstructs lineage branching. Nat Methods. 2016;13(10):845–8.
199. Setty M, Tadmor MD, Reich-Zeliger S, Angel O, Salame TM, Kathail P, et al. Wishbone identifies bifurcating developmental trajectories from single-cell data. Nat Biotechnol. 2016;34(6):637–45.
200. Herman JS, Sagar D, Grun D. FateID infers cell fate bias in multipotent progenitors from single-cell RNA-seq data. Nat Methods. 2018;15(5):379–86.
201. Velten L, Haas SF, Raffel S, Blaszkiewicz S, Islam S, Hennig BP, et al. Human haematopoietic stem cell lineage commitment is a continuous process. Nat Cell Biol. 2017;19(4):271–81.
202. Campbell KR, Yau C. Probabilistic modeling of bifurcations in single-cell gene expression data using a Bayesian mixture of factor analyzers. Wellcome Open Res. 2017;2:19.
203. Street K, Risso D, Fletcher RB, Das D, Ngai J, Yosef N, et al. Slingshot: cell lineage and pseudotime inference for single-cell transcriptomics. BMC Genomics. 2018;19(1):477.
204. Gan Y, Guo C, Guo W, Xu G, Zou G. Entropy-based inference of transition states and cellular trajectory for single-cell transcriptomics. Brief Bioinform. 2022;23(4):bbac225.
205. Cao J, Spielmann M, Qiu X, Huang X, Ibrahim DM, Hill AJ, et al. The single-cell transcriptional landscape of mammalian organogenesis. Nature. 2019;566(7745):496–502.
206. Wolf FA, Hamey FK, Plass M, Solana J, Dahlin JS, Gottgens B, et al. PAGA: graph abstraction reconciles clustering with trajectory inference through a topology preserving map of single cells. Genome Biol. 2019;20(1):59.
207. Welch JD, Hartemink AJ, Prins JF. SLICER: inferring branched, nonlinear cellular trajectories from single cell RNA-seq data. Genome Biol. 2016;17(1):106.
208. Van den Berge K, Roux De Bézieux H, Street K, Saelens W, Cannoodt R, Saeys Y, et al. Trajectory-based differential expression analysis for single-cell sequencing data. Nat Commun. 2020;11(1):1201.
209. Song D, Li JJ. PseudotimeDE: inference of differential gene expression along cell pseudotime with well-calibrated *P*-values from single-cell RNA sequencing data. Genome Biol. 2021;22(1):124.
210. La Manno G, Soldatov R, Zeisel A, Braun E, Hochgerner H, Petukhov V, et al. RNA velocity of single cells. Nature. 2018;560(7719):494–8.
211. Bergen V, Lange M, Peidli S, Wolf FA, Theis FJ. Generalizing RNA velocity to transient cell states through dynamical modeling. Nat Biotechnol. 2020;38(12):1408–14.
212. Lange M, Bergen V, Klein M, Setty M, Reuter B, Bakhti M, et al. Cell Rank for directed single-cell fate mapping. Nat Methods. 2022;19(2):159–70.
213. Zhang Z, Zhang X. Inference of high-resolution trajectories in single-cell RNA-seq data by using RNA velocity. Cell Rep Methods. 2021;1(6):100095.
214. Dimitrov D, Türei D, Garrido-Rodriguez M, Burmedi PL, Nagai JS, Boys C, et al. Comparison of methods and resources for cell–cell communication inference from single-cell RNA-seq data. Nat Commun. 2022;13(1):3224.
215. Efremova M, Vento-Tormo M, Teichmann SA, Vento-Tormo R. Cell PhoneDB: inferring cell–cell communication from combined expression of multi-subunit ligand-receptor complexes. Nat Protoc. 2020;15(4):1484–506.
216. Noël F, Massenet-Regad L, Carmi-Levy I, Cappuccio A, Grandclaudon M, Trichot C, et al. Dissection of intercellular communication using the transcriptome-based framework ICELLNET. Nat Commun. 2021;12(1):1089.
217. Shao X, Liao J, Li C, Lu X, Cheng J, Fan X. CellTalkDB: a manually curated database of ligand-receptor interactions in humans and mice. Brief Bioinform. 2021;22(4):bbaa269.
218. Cabello-Aguilar S, Alame M, Kon-Sun-Tack F, Fau C, Lacroix M, Colinge J. SingleCellSignalR: inference of intercellular networks from single-cell transcriptomics. Nucleic Acids Res. 2020;48(10):e55.
219. Turei D, Valdeolivas A, Gul L, Palacio-Escat N, Klein M, Ivanova O, et al. Integrated intra- and intercellular signaling knowledge for multicellular omics analysis. Mol Syst Biol. 2021;17(3):e9923.
220. Garcia-Alonso L, Lorenzi V, Mazzeo CI, Alves-Lopes JP, Roberts K, Sancho-Serra C, et al. Single-cell roadmap of human gonadal development. Nature. 2022;607(7919):540–7.
221. Peng L, Wang F, Wang Z, Tan J, Huang L, Tian X, et al. Cell–cell communication inference and analysis in the tumour microenvironments from single-cell transcriptomics: data resources and computational strategies. Brief Bioinform. 2022;23(4):bbac234.
222. Armingol E, Officer A, Harismendy O, Lewis NE. Deciphering cell–cell interactions and communication from gene expression. Nat Rev Genet. 2021;22(2):71–88.
223. Camp JG, Sekine K, Gerber T, Loeffler-Wirth H, Binder H, Gac M, et al. Multilineage communication regulates human liver bud development from pluripotency. Nature. 2017;546(7659):533–8.
224. Browaeys R, Saelens W, Saeys Y. NicheNet: modeling intercellular communication by linking ligands to target genes. Nat Methods. 2020;17(2):159–62.
225. Choi H, Sheng J, Gao D, Li F, Durrans A, Ryu S, et al. Transcriptome analysis of individual stromal cell populations identifies stroma-tumor crosstalk in mouse lung cancer model. Cell Rep. 2015;10(7):1187–201.
226. Jakobsson JET, Spjuth O, Lagerström MC. scConnect: a method for exploratory analysis of cell–cell communication based on single cell RNA sequencing data. Bioinformatics. 2021;37(20):3501–8.
227. Hou R, Denisenko E, Ong HT, Ramilowski JA, Forrest ARR. Predicting cell-to-cell communication networks using NATMI. Nat Commun. 2020;11(1):5011.
228. Lamurias A, Ruas P, Couto FM. PPR-SSM: personalized PageRank and semantic similarity measures for entity linking. BMC Bioinform. 2019;20(1):534.
229. Wang S, Karikomi M, Maclean AL, Nie Q. Cell lineage and communication network inference via optimization for single-cell transcriptomics. Nucleic Acids Res. 2019;47(11):e66.
230. Tyler SR, Rotti PG, Sun X, Yi Y, Xie W, Winter MC, et al. PyMINEr finds gene and autocrine-paracrine networks from human islet scRNA-seq. Cell Rep. 2019;26(7):1951-64.e8.
231. Lee HO, Hong Y, Etlioglu HE, Cho YB, Pomella V, Van den Bosch B, et al. Lineage-dependent gene expression programs influence the immune landscape of colorectal cancer. Nat Genet. 2020;52(6):594–603.
232. Zhou JX, Taramelli R, Pedrini E, Knijnenburg T, Huang S. Extracting intercellular signaling network of cancer tissues using ligand-receptor expression patterns from whole-tumor and single-cell transcriptomes. Sci Rep. 2017;7(1):8815.
233. Kumar MP, Du J, Lagoudas G, Jiao Y, Sawyer A, Drummond DC, et al. Analysis of single-cell RNA-seq identifies cell–cell communication associated with tumor characteristics. Cell Rep. 2018;25(6):1458–68e4.
234. Cillo AR, Kürten CHL, Tabib T, Qi Z, Onkar S, Wang T, et al. Immune landscape of viral- and carcinogen-driven head and neck cancer. Immunity. 2020;52(1):183-99.e9.

235. Palla G, Spitzer H, Klein M, Fischer D, Schaar AC, Kuemmerle LB, et al. Squidpy: a scalable framework for spatial omics analysis. Nat Methods. 2022;19(2):171–8.
236. Schapiro D, Jackson HW, Raghuraman S, Fischer JR, Zanotelli VRT, Schulz D, et al. histoCAT: analysis of cell phenotypes and interactions in multiplex image cytometry data. Nat Methods. 2017;14(9):873–6.
237. Jin S, Guerrero-Juarez CF, Zhang L, Chang I, Ramos R, Kuan CH, et al. Inference and analysis of cell–cell communication using Cell Chat. Nat Commun. 2021;12(1):1088.
238. Lambert SA, Jolma A, Campitelli LF, Das PK, Yin Y, Albu M, et al. The human transcription factors. Cell. 2018;172(4):650–65.
239. Hu H, Miao YR, Jia LH, Yu QY, Zhang Q, Guo AY. AnimalTFDB 3.0: a comprehensive resource for annotation and prediction of animal transcription factors. Nucleic Acids Res. 2019;47(D1):D33-8.
240. Fornes O, Castro-Mondragon JA, Khan A, van der Lee R, Zhang X, Richmond PA, et al. JASPAR 2020: update of the open-access database of transcription factor binding profiles. Nucleic Acids Res. 2020;48(D1):D87–92.
241. Han H, Cho JW, Lee S, Yun A, Kim H, Bae D, et al. TRRUST v2: an expanded reference database of human and mouse transcriptional regulatory interactions. Nucleic Acids Res. 2018;46(D1):D380–6.
242. Feng C, Song C, Liu Y, Qian F, Gao Y, Ning Z, et al. KnockTF: a comprehensive human gene expression profile database with knockdown/knockout of transcription factors. Nucleic Acids Res. 2020;48(D1):D93–100.
243. Mei S, Qin Q, Wu Q, Sun H, Zheng R, Zang C, et al. Cistrome Data Browser: a data portal for ChIP-Seq and chromatin accessibility data in human and mouse. Nucleic Acids Res. 2017;45(D1):D658–62.
244. Elmentaite R, Ross ADB, Roberts K, James KR, Ortmann D, Gomes T, et al. Single-cell sequencing of developing human gut reveals transcriptional links to childhood Crohn's disease. Dev Cell. 2020;55(6):771-83.e5.
245. Langfelder P, Horvath S. WGCNA: an R package for weighted correlation network analysis. BMC Bioinform. 2008;9:559.
246. Kazer SW, Aicher TP, Muema DM, Carroll SL, Ordovas-Montanes J, Miao VN, et al. Integrated single-cell analysis of multicellular immune dynamics during hyperacute HIV-1 infection. Nat Med. 2020;26(4):511–8.
247. Liao M, Liu Y, Yuan J, Wen Y, Xu G, Zhao J, et al. Single-cell landscape of bronchoalveolar immune cells in patients with COVID-19. Nat Med. 2020;26(6):842–4.
248. Cheng S, Li Z, Gao R, Xing B, Gao Y, Yang Y, et al. A pan-cancer single-cell transcriptional atlas of tumor infiltrating myeloid cells. Cell. 2021;184(3):792-809.e23.
249. Matsumoto H, Kiryu H, Furusawa C, Ko MSH, Ko SBH, Gouda N, et al. SCODE: an efficient regulatory network inference algorithm from single-cell RNA-Seq during differentiation. Bioinformatics. 2017;33(15):2314–21.
250. Papili Gao N, Ud-Dean SMM, Gandrillon O, Gunawan R. SINCERITIES: inferring gene regulatory networks from time-stamped single cell transcriptional expression profiles. Bioinformatics. 2018;34(2):258–66.
251. Luo Q, Yu Y, Lan X. SIGNET: single-cell RNA-seq-based gene regulatory network prediction using multiple-layer perceptron bagging. Brief Bioinform. 2022;23(1):bbab547.
252. Chen J, Cheong C, Lan L, Zhou X, Liu J, Lyu A, et al. DeepDRIM: a deep neural network to reconstruct cell-type-specific gene regulatory network using single-cell RNA-seq data. Brief Bioinform. 2021;22(6):bbab325.
253. Pratapa A, Jalihal AP, Law JN, Bharadwaj A, Murali TM. Benchmarking algorithms for gene regulatory network inference from single-cell transcriptomic data. Nat Methods. 2020;17(2):147–54.
254. Chen S, Mar JC. Evaluating methods of inferring gene regulatory networks highlights their lack of performance for single cell gene expression data. BMC Bioinform. 2018;19(1):232.
255. Alghamdi N, Chang W, Dang P, Lu X, Wan C, Gampala S, et al. A graph neural network model to estimate cell-wise metabolic flux using single-cell RNA-seq data. Genome Res. 2021;31(10):1867–84.
256. Artyomov MN, Van den Bossche J. Immunometabolism in the single-cell era. Cell Metab. 2020;32(5):710–25.
257. Gubin MM, Esaulova E, Ward JP, Malkova ON, Runci D, Wong P, et al. High-dimensional analysis delineates myeloid and lymphoid compartment remodeling during successful immune-checkpoint cancer therapy. Cell. 2018;175(4):1014-30.e19.
258. Ariss MM, Islam ABMMK, Critcher M, Zappia MP, Frolov MV. Single cell RNA-sequencing identifies a metabolic aspect of apoptosis in Rbf mutant. Nat Commun. 2018;9(1):5024.
259. Wu Y, Yang S, Ma J, Chen Z, Song G, Rao D, et al. Spatiotemporal immune landscape of colorectal cancer liver metastasis at single-cell level. Cancer Discov. 2022;12(1):134–53.
260. Raman K, Chandra N. Flux balance analysis of biological systems: applications and challenges. Brief Bioinform. 2009;10(4):435–49.
261. Kanehisa M, Furumichi M, Sato Y, Ishiguro-Watanabe M, Tanabe M. KEGG: integrating viruses and cellular organisms. Nucleic Acids Res. 2021;49(D1):D545–51.
262. Jassal B, Matthews L, Viteri G, Gong C, Lorente P, Fabregat A, et al. The reactome pathway knowledgebase. Nucleic Acids Res. 2020;48(D1):D498–503.
263. Orth JD, Thiele I, Palsson BØ. What is flux balance analysis? Nat Biotechnol. 2010;28(3):245–8.
264. Damiani C, Maspero D, Di Filippo M, Colombo R, Pescini D, Graudenzi A, et al. Integration of single-cell RNA-seq data into population models to characterize cancer metabolism. PLoS Comput Biol. 2019;15(2):e1006733.
265. Wagner A, Wang C, Fessler J, Detomaso D, Avila-Pacheco J, Kaminski J, et al. Metabolic modeling of single Th17 cells reveals regulators of autoimmunity. Cell. 2021;184(16):4168-85.e21.
266. Thiele I, Swainston N, Fleming RMT, Hoppe A, Sahoo S, Aurich MK, et al. A community-driven global reconstruction of human metabolism. Nat Biotechnol. 2013;31(5):419–25.
267. Pei W, Shang F, Wang X, Fanti AK, Greco A, Busch K, et al. Resolving fates and single-cell transcriptomes of hematopoietic stem cell clones by polyloxexpress barcoding. Cell Stem Cell. 2020;27(3):383-95.e388.
268. Basharat Z, Majeed S, Saleem H, Khan IA, Yasmin A. An overview of algorithms and associated applications for single cell RNA-Seq data imputation. Curr Genomics. 2021;22(5):319–27.
269. Hou W, Ji Z, Ji H, Hicks SC. A systematic evaluation of single-cell RNA-sequencing imputation methods. Genome Biol. 2020;21(1):218.
270. Van Dijk D, Sharma R, Nainys J, Yim K, Kathail P, Carr AJ, et al. Recovering gene interactions from single-cell data using data diffusion. Cell. 2018;174(3):716–29.e27.
271. Wu X, Liu T, Ye C, Ye W, Ji G. scAPAtrap: identification and quantification of alternative polyadenylation sites from single-cell RNA-seq data. Brief Bioinform. 2021;22(4):bbaa273.
272. Patrick R, Humphreys DT, Janbandhu V, Oshlack A, Ho JWK, Harvey RP, et al. Sierra: discovery of differential transcript usage from polyA-captured single-cell RNA-seq data. Genome Biol. 2020;21(1):167.
273. Gao Y, Li L, Amos CI, Li W. Analysis of alternative polyadenylation from single-cell RNA-seq using scDaPars reveals cell subpopulations invisible to gene expression. Genome Res. 2021;31(10):1856–66.
274. Li GW, Nan F, Yuan GH, Liu CX, Liu X, Chen LL, et al. SCAPTURE: a deep learning-embedded pipeline that captures polyadenylation information from 3'tag-based RNA-seq of single cells. Genome Biol. 2021;22(1):221.
275. Zhou R, Xiao X, He P, Zhao Y, Xu M, Zheng X, et al. SCAPE: a mixture model revealing single-cell polyadenylation diversity and cellular dynamics during cell differentiation and reprogramming. Nucleic Acids Res. 2022;50(11):e66.
276. Wang X, Hou J, Quedenau C, Chen W. Pervasive isoform-specific translational regulation via alternative transcription start sites in mammals. Mol Syst Biol. 2016;12(7):875.
277. He Y, Chen Q, Zhang J, Yu J, Xia M, Wang X. Pervasive 3'-UTR isoform switches during mouse oocyte maturation. Front Mol Biosci. 2021;8:727614.
278. Philpott M, Watson J, Thakurta A, Brown T Jr, Brown T Sr, Oppermann U, et al. Nanopore sequencing of single-cell transcriptomes with scCOLOR-seq. Nat Biotechnol. 2021;39(12):1517–20.
279. Tian L, Jabbari JS, Thijssen R, Gouil Q, Amarasinghe SL, Voogd O, et al. Comprehensive characterization of single-cell full-length isoforms in human and mouse with long-read sequencing. Genome Biol. 2021;22(1):310.
280. Rebboah E, Reese F, Williams K, Balderrama-Gutierrez G, McGill C, Trout D, et al. Mapping and modeling the genomic basis of differential RNA isoform expression at single-cell resolution with LR-Split-seq. Genome Biol. 2021;22(1):286.
281. Li J, Pan T, Chen L, Wang Q, Chang Z, Zhou W, et al. Alternative splicing perturbation landscape identifies RNA binding proteins as potential therapeutic targets in cancer. Mol Ther Nucleic Acids. 2021;24:792–806.

282. Huang HY, Lin YC, Li J, Huang KY, Shrestha S, Hong HC, et al. miRTarBase 2020: updates to the experimentally validated microRNA-target interaction database. Nucleic Acids Res. 2020;48(D1):D148–54.
283. Jiang T, Zhou W, Chang Z, Zou H, Bai J, Sun Q, et al. ImmReg: the regulon atlas of immune-related pathways across cancer types. Nucleic Acids Res. 2021;49(21):12106–18.
284. Chen W, Guillaume-Gentil O, Rainer PY, Gabelein CG, Saelens W, Gardeux V, et al. Live-seq enables temporal transcriptomic recording of single cells. Nature. 2022;608(7924):733–40.
285. Zhang K, Hocker JD, Miller M, Hou X, Chiou J, Poirion OB, et al. A single-cell atlas of chromatin accessibility in the human genome. Cell. 2021;184(24):5985-6001.e19.
286. Karemaker ID, Vermeulen M. Single-cell DNA methylation profiling: technologies and biological applications. Trends Biotechnol. 2018;36(9):952–65.
287. Zhang R, Zhou T, Ma J. Multiscale and integrative single-cell Hi-C analysis with Higashi. Nat Biotechnol. 2022;40(2):254–61.
288. Lee J, Hyeon DY, Hwang D. Single-cell multiomics: technologies and data analysis methods. Exp Mol Med. 2020;52(9):1428–42.
289. Long Z, Sun C, Tang M, Wang Y, Ma J, Yu J, et al. Single-cell multiomics analysis reveals regulatory programs in clear cell renal cell carcinoma. Cell Discov. 2022;8(1):68.
290. Marx V. Method of the year: spatially resolved transcriptomics. Nat Methods. 2021;18(1):9–14.