Genomics & Informatics

REVIEW Open Access

Navigating single-cell RNA-sequencing: protocols, tools, databases, and applications

Ankish Arya[1†], Prabhat Tripathi[1†], Nidhi Dubey[1], Imlimaong Aier[1] and Pritish Kumar Varadwaj[1*]

Abstract

Single-cell RNA-sequencing (scRNA-seq) technology brought about a revolutionary change in the transcriptomic world, paving the way for comprehensive analysis of cellular heterogeneity in complex biological systems. It enabled researchers to see how different cells behaved at single-cell levels, providing new insights into the process. However, despite all these advancements, scRNA-seq also experiences challenges related to the complexity of data analysis, interpretation, and multi-omics data integration. In this review, these complications were discussed in detail, directly pointing at the optimization of scRNA-seq approaches and understanding the world of single-cell and its dynamics. Different protocols and currently functional single-cell databases were also covered. This review highlights different tools for the analysis of scRNA-seq and their methodologies, emphasizing innovative techniques that enhance resolution and accuracy at a single-cell level. Various applications were explored across domains including drug discovery, tumor microenvironment (TME), biomarker discovery, and microbial profiling, and case studies were discussed to explain the importance of scRNA-seq by uncovering novel and rare cell types and their identification. This review underlines a crucial aspect of scRNA-seq in the advancement of personalized medicine and highlights its potential to understand the complexity of biological systems.

Keywords Single-cell RNA sequencing, Databases, Protocols and tools, Cellular heterogeneity, Drug discovery

1 Introduction

Two centuries after Robert Hooke and Antonie van Leeuwenhoek, cells were redefined as the fundamental functional unit of life [1]. Since then, researchers have conducted numerous experiments and developed various techniques to study cells within complex multicellular systems for a more comprehensive understanding [2, 3]. Over the past decade, bulk RNA-sequencing technologies have been widely employed to investigate gene expression patterns on a population scale, which allowed researchers to analyze the transcriptome of a group of cells or tissues, providing insights into gene activity levels within the sample. The emergence of single-cell RNA-sequencing opened up remarkable prospects for investigating gene expression profiles at the individual cell level, when scRNA-seq was first reported in the 4-cell blastomere stage in the year of 2009. Consequently, in 2014, the first multiplexed scRNA-seq method was developed [4, 5]. In 2017, scRNASeqDB, a database dedicated to gene expression profiles for human single cells, was created [6]. In 2021, Asc-Seurat, a user-friendly web application for comprehensive scRNA-seq data analysis, was developed, which can perform complete analysis [7].

At present, scRNA-seq is increasingly being preferred when addressing crucial biological inquiries related to cell heterogeneity and early embryo development, particularly in cases involving a limited number of cells. In recent times, scientists have used scRNA-seq on a wide range of species, particularly in various human tissues, both healthy and cancerous. These studies have revealed that gene expression can differ significantly

†Ankish Arya and Prabhat Tripathi contributed equally to this work.

*Correspondence:
Pritish Kumar Varadwaj
pritish@iiita.ac.in
[1] Department of Applied Sciences, Indian Institute of Information Technology Allahabad, Jhalwa, Prayagraj 211015, Uttar Pradesh, India

from one cell to another. This insight helps understand how genes are active or inactive in individual cells, shedding light on the complexity of our biology [8, 9]. ScRNA-seq is a powerful technique for tackling issues related to the unpredictable behavior of genes. Currently, studies using scRNA-seq hold significant potential for uncovering previously unknown cell types, mapping out developmental pathways, and investigating the complexity of tumor diversity [10]. The key contrast between bulk RNA-seq and scRNA-seq is whether each library reflects an individual cell or a cell group, driven by challenges like scarce transcripts in single cells, inefficient mRNA capture, losses in reverse transcription, and bias in cDNA amplification due to the minute amounts involved [11, 12]. During quality control, when using scRNA-seq, it is important to identify and remove low-quality individual cells and any data that might represent multiple cells. In some methods like drop-based sequencing, background noise can also be removed. However, one should be careful when applying data normalization techniques designed for bulk RNA-sequencing because they can introduce errors into scRNA-seq data [13]. When it comes to aligning sequencing data, the tools commonly used for bulk RNA-sequencing can also be used for scRNA-seq data. However, dedicated alignment methods designed for scRNA-seq often offer advantages in terms of efficient use of computing resources and faster processing speed [11, 14]. There are often missing values in scRNA-seq data. To combat this issue, multiple imputation algorithms which rely on various models can be employed. When dealing with batch effects in scRNA-seq, it is crucial to account for both technical deviations and biological differences. The goal is to preserve the biological variation of interest by reducing unwanted variation. While scRNA-seq provides valuable advantages for biological research, it has notable limitations. Gene expression data obtained through this process is often noisy, high-dimensional, and sparsely populated. Consequently, to fully harness the potential of scRNA-seq technology, specialized computational tools tailored to scRNA-seq data are essential [11]. In recent years, the explosion of single-cell analysis tools has increased the difficulty of selecting the right tool for a given dataset [15, 16]. Many tools are designed to simplify the processing and comprehension of scRNA-sequencing data through user-friendly interfaces [17, 18]. Nevertheless, to the uninitiated, these tools and algorithms can resemble elusive "dark elixirs." Empowering researchers in their quest for the most fitting methods, algorithms, and tools demands a comprehensive review that unveils the inner workings of these computational marvels.

2 Different protocols for scRNA-seq: an overview

Many scRNA-seq approaches have been suggested for single-cell transcriptomic research (Table 1). After the initial scRNA-seq technique was published, several alternative scRNA-seq strategies emerged. These scRNA-seq technologies differ in at least one of the following areas: availability of Unique Molecular Identifiers (UMIs), cell isolation, cell lysis, reverse transcription, amplification, transcript coverage, and transcription. One obvious distinction between these scRNA-seq approaches is that some of these techniques can generate full-length (or nearly full-length) transcript sequencing data (e.g., Sn-drop, Smart-Seq2, Quartz-Seq2, MATQ-Seq, and Fluidigm C1), whereas others can only capture and sequence the transcripts 3′ or 5′ ends (e.g., REAP-Seq, Drop-Seq, inDrop, Seq-Well, DroNC-Seq, and SPLiT-Seq). Different scRNA-seq techniques each have unique benefits and restrictions. Numerous evaluations that have been published analyze thorough comparisons. According to one research, Smart-Seq2 performs better than other scRNA-seq technologies, including CEL-Seq2, MARS-Seq, Smart-Seq, and Drop-Seq procedures, in identifying more expressed genes. Furthermore, MATQ-Seq is superior to Smart-Seq2 in detecting low-abundance genes. Full- length scRNA-Seq methods offer unique advantages over 3′ end or 5′ end counting protocols. They excel in tasks like isoform usage analysis, allelic expression detection, and identifying RNA editing due to their comprehensive coverage of transcripts. Furthermore, in the detection of specific lowly expressed genes or transcripts, full-length scRNA-seq approaches may outperform 3′ end sequencing methods [19]. Droplet-based techniques like Drop-Seq, InDrop, and Chromium often enable a higher throughput of cells and a lower sequencing cost per cell as compared to whole-transcript scRNA-seq [20–22]. The ability to handle large numbers of cells makes droplet-based techniques particularly helpful for detecting various cell subpopulations inside complex tissues or tumor samples. While inDrop and CEL-Seq2 rely on in vitro transcription (IVT) for amplification, the remaining protocols utilize polymerase chain reaction (PCR) as their amplification method as described in Table 1.

3 Currently employed methodologies

The primary steps involved in scRNA-seq encompass single-cell isolation and capture, cell lysis, reverse transcription, cDNA amplification, and library preparation (Fig. 1).

Table 1 Comparison of protocols based on isolation strategy, transcript coverage, UMI usage, and amplification method

Protocols	Isolation strategy	Transcript coverage	UMI	Amplification method	Unique features
SnDrop [23]	Droplet-based	Full-length	Yes	PCR	Combines nuclei isolation with droplet microfluidics; reduces dissociation artifacts
REAP-Seq [24]	Droplet-based	3′-only	Yes	PCR	Allows simultaneous protein and RNA detection
Smart-Seq2 [25]	FACS	Full-length	No	PCR	Enhanced sensitivity for detecting low-abundance transcripts; generates full-length cDNA
Drop-Seq [26]	Droplet-based	3′-end	Yes	PCR	High-throughput and low cost per cell; scalable to thousands of cells simultaneously
inDrop [27]	Droplet-based	3′-end	Yes	IVT	Uses hydrogel beads; low cost per cell; efficient barcode capture
STRT-Seq [28]	FACS	5′-only	Yes	PCR	High-resolution mapping of transcription start sites
CEL-Seq2 [29]	FACS	3′-only	Yes	IVT	Linear amplification reduces bias compared to PCR
Seq-well [30]	Droplet-based	3′-only	Yes	PCR	Portable, low-cost, easily implemented without complex equipment
Quartz-Seq2 [31]	FACS	Full-length	No	PCR	Optimized reaction conditions for improved sensitivity
DroNC-Seq [32]	Droplet-based	3′-only	Yes	PCR	Specialized for single-nucleus sequencing, minimal dissociation bias
sci-RNA-Seq [33]	FACS	3′-only	Yes	PCR	Combinatorial indexing for ultra-high throughput without single-cell isolation equipment
SPLiT-Seq [9, 34]	Not required	3′-only	Yes	PCR	Combinatorial indexing without physical separation; highly scalable and low cost
MATQ-Seq [35]	Droplet-based	Full-length	Yes	PCR	Increased accuracy in quantifying transcripts; efficient detection of transcript variants
Fluidigm-C1 [36]	Droplet-based	Full-length	No	PCR	Microfluidics-based single-cell capture; precise cell handling

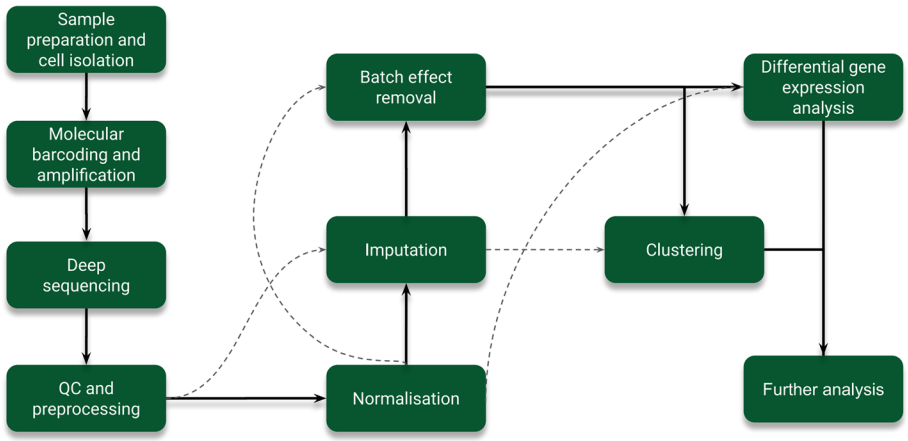

Fig. 1 A graphical overview of the steps involved in scRNA sequencing. There are certain tools that skip certain steps: DeepImpute and MAGIC skip normalization to perform batch effect correction, while EdgeR, MAST, and Monocle2 skip imputation in order to perform differential gene expression analysis. DESeq2 directly bypasses normalization and imputation to perform differential gene expression. Furthermore, clustering is done immediately by RaceID, SC3, and Monocle2 without the need for batch effect correction

3.1 Sample preparation and cell isolation

The initial stage of performing scRNA-seq involves the extraction of viable and individual cells from the specific tissue under investigation. Novel methodologies, such as the isolation of individual nuclei for RNA-seq (snRNA-seq), are used in conditions where tissue dissociation is challenging, or when samples are frozen or cells are fragile. Other methodologies include the use of "split-pooling" scRNA-seq techniques, which apply combinatorial indexing (cell barcodes) to single cells, offering distinct advantages over the isolation of intact single cells. These advantages include the ability to handle large sample sizes (up to millions of cells) and greater efficiency in parallel processing of multiple samples while

eliminating the need for expensive microfluidic devices. Subsequently, the individual cells are subjected to lysis in order to facilitate the capture of RNA molecules. Poly[T]-primers are frequently employed to selectively analyze polyadenylated mRNA molecules while minimizing the capture of ribosomal RNAs [37]. Table 2 gives an overview of different methods for cell-cell preparation and isolation strategy.

3.2 Molecular barcoding and amplification

Following the conversion of RNA into complementary DNA (cDNA), the resulting cDNA molecules undergo amplification by either the polymerase chain reaction (PCR) or in vitro transcription (IVT) methods. PCR, a non-linear amplification process, is utilized in several methodologies such as Smart-Seq, Smart-Seq2, Fluidigm C1, Drop-Seq, 10x Genomics, MATQ-Seq, Seq-Well, and DNBelab C4. At present, there are two distinct techniques for PCR amplification.

The utilization of SMART technology involves the exploitation of the transferase and strand-switch activity of Moloney murine leukemia virus reverse transcriptase. This enzyme is employed to integrate template-switching oligos as adaptors for subsequent PCR amplification. The aforementioned approach was widely employed for the amplification of cDNA. The alternative approach involves the ligation of common adaptors to the 5′ end of cDNA, using either poly(A) or poly(C), in order to facilitate the subsequent PCR reaction. The IVT method is a technique employed in many sequencing procedures like as CEL-Seq, MARS-Seq, and inDrop-Seq. It serves as an amplification strategy and facilitates linear amplification of genetic material. A second iteration of reverse transcription of the amplified RNA is necessary, leading to the emergence of further 3′ coverage biases. Both methodologies have the potential to result in amplification biases. In order to mitigate biases associated with amplification, a technique called UMIs was implemented. UMIs are used to label each individual mRNA molecule within a cell during the reverse transcription process. This approach enhances the quantitative aspect of scRNA-seq and improves the accuracy of data interpretation by effectively eliminating biases introduced by PCR amplification. The CEL-Seq, MARS-Seq, Drop-Seq, inDrop-Seq, 10x Genomics, MATQ-Seq, Seq-Well, and DNBelab C4 methods have incorporated UMIs [3].

3.3 Deep sequencing

After the generation of single cell-barcoded cDNAs from individual cells or nuclei, the subsequent sequencing of the cDNA can be performed using several advanced sequencing technologies. Regarding high-throughput sequencing using DNA nanoballs (DNBseq), the DNA fragments chosen were subjected to repair processes to achieve a blunt end and were subsequently changed at the three ends to generate a dATP overhang. Following this, the dTTP-tailed adapter sequence was utilized to ligate each end of the DNA fragment. The ligation result was subsequently subjected to a few cycles of amplification, followed by a single-strand cycle. A specific segment of the PCR product was subjected to reverse complementation using a specialized molecule, followed by ligation with a single-stranded molecule using DNA ligase. This process ultimately resulted in the generation of a circular DNA library consisting of single-stranded molecules [3]. Some other sequencing platforms which are relevant to scRNA-seq are Illumina Sequencing (short-read sequencing), DNBseq (DNA Nanoball sequencing, short-read sequencing), Oxford Nanopore Technologies (long-read sequencing), and Pacific Biosciences (PacBio, single molecule real-time sequencing) [42]. Table 3 gives a comparative overview of sequencing platforms.

3.4 Quality check and pre-processing

Performing quality control is important to remove consistent technical variations that might have been introduced in generating the data, to focus on the biological variations of cells. The random sampling procedure and the limited RNA content in data increase noise, compared to bulk RNA-sequencing [45–47]. Dropout

Table 2 Summary of different methods of cell preparation and isolation strategy

Method	Principle	Advantages	Limitations	Applications
FACS [38]	Fluorescence-based sorting using specific cell markers	Highly selective, precise isolation	Expensive, cellular stress	Targeted cell populations
Microfluidics (droplet-based) [39]	Encapsulation of cells in droplets with barcoded beads	High-throughput, efficient, automated	High cost, transcript loss	Large-scale profiling, general use
Split-pooling [40]	Combinatorial barcoding without physical isolation	Cost-effective, highly scalable	Complex data handling, barcode collisions	Large-scale studies, multiplexed samples
snRNA-seq [41]	Isolation of nuclei instead of intact cells	Minimal dissociation stress, suitable for frozen tissues	Lower RNA yield, excludes cytoplasmic transcripts	Difficult to dissociate tissues, archival samples

Table 3 Comparison of sequencing platforms based on key parameters

Sequencing platform	Principle	Throughput	Read length	Accuracy	Cost	Error type	Application
Illumina [43]	SBS	Very high	Short (50–300 bp)	Very high (~ 99.9%)	Cost-effective	Substitutions	High-throughput gene quantification
DNBSeq [3]	DNA Nanoball sequencing	Very high	Short (50–200 bp)	Very high (~ 99.9%)	Cost-effective	Substitutions	High-throughput gene quantification
Oxford Nanopore [44]	Nanopore sequencing	Moderate–high	Long (up to > 100 kb)	Moderate (90–98%)	Moderate, declining	Indels/substitutions	Full-length isoform detection
Pacific Biosciences [44]	Single molecule real-time (SMRT)	Moderate–low	Long (5–30 kb or more)	Moderate to very high (90–99.9%)	Higher, declining	Indels/substitutions	Isoform characterization and full-length transcript profiling

events (genes not detected due to low RNA content) create excessive zero counts [48] and can lead to highly sparse datasets, making it difficult to detect genuine biological variations within individual cells. Filtering out lower quality cells is crucial for expression analysis; criteria vary based on cell and tissue type [3, 20, 49].

Two primary quality metrics in scRNA-seq are expressed counts and total library size. Low counts suggest poor RNA capture, while high counts may indicate multiple cells captured erroneously, depicted visually via violin plots [17, 19, 50]. Filtering out potential doublets or multiplets in scRNA-seq involves setting thresholds based on expressed feature counts, affected by both biological and technical factors. Sequencing depth influences read and feature counts, and using robust statistics like median absolute deviations aids in identifying outlier cells [51–53]. There are certain established methods for detecting doublets [54], which are given in Table 4. Mitochondrial gene presence in the data can also affect the quality of data; its proportion can vary from tissue to tissue like heart cells typically exhibit around 30%, contrasting with lymphocytes [55, 56].

In droplet-based scRNA-seq protocols, another source of unwanted signals comes from ambient RNAs. These are RNA molecules that are freely floating in the cell lysate due to the breakdown of dead or dying cells before the droplets are separated. Since these ambient mRNAs are found everywhere, they add extra background noise and can greatly muddle the quality of the data and the true biological signals we are trying to capture [68]. Methods like SoupX, DecontX, and CellBender effectively remove ambient RNA influences in single-cell RNA-seq. SoupX uses known negative markers (genes not expressed in specific cell types), while DecontX employs Bayesian inference, and CellBender uses deep generative models [63, 68].

Low-abundance genes should be excluded as they do not provide enough information for reliable analysis. Additionally thresholds should be set based on the number of cells expressing a gene or the genes average expression level [18, 45, 65]. Depending on the analysis, non-coding genes may be excluded to simplify the data. In scRNA-seq data, mitochondrial genes are discarded after quality control to avoid biases, as mitochondrial transcripts are not usually expressed in the nucleus [114, 115].

Among quality-checking and preprocessing tools, Solo was reported to have an accuracy of 83.59%, with a runtime of 13–17 min and memory usage of 7–12 GB across different standard datasets [61]. Scrublet demonstrated an accuracy of 99% and supported scalability but lacked data on runtime and memory consumption [52]. DropletQC can process 100 million reads in under 133 s on 8 CPUs with 16 GB RAM, highlighting its efficiency for nuclear fraction analysis [60]. ScPipe required 10 h to process a dataset with 112 million reads while consuming 540 GB of RAM, making computation highly resource-intensive [66]. CellBender exhibited adaptability, with runtime varying from 20 min to 1 h depending on dataset size and GPU availability, though precise memory usage data was unavailable [64]. DoubletFinder was notable for its limitation in detecting homotypic doublets and the need for parameter optimization, but it performed better when integrated with sample multiplexing [67].

3.5 Normalization

Normalization is an important process in scRNA-seq data analysis, as it helps focus on meaningful information by fixing issues like differences in how well genes are captured, the depth of sequencing, and other technical variations that can affect the data. There are two main categories of normalization: within sample normalization and between sample normalization [12, 116] (Table 4).

Table 4 List of tools used in scRNA-seq analysis

Tool name	Input format	Algorithm framework	Output format	Specific use case	Language	Availability
Pre-processing and quality control						
FastQC [57]	FASTQ	Identifies overrepresented short sequences (k-mers)	HTML, ZIP	NA	JAVA	https://www.bioinformatics.babraham.ac.uk/projects/fastqc/
PopsicleR [58]	Expression matrix (e.g., CSV, TSV)	Uses a mixture model of mitochondrial content and gene expression levels to classify cells as low or high-quality	Filtered expression matrix, plots (e.g., PDF, PNG)	Suitable for detecting rare cell populations	R	https://github.com/bicciatolab/popsicleR
miQC [59]	Expression matrix, metadata (e.g., CSV)	Uses probabilistic approach to identify empty droplets or droplets with low RNA content by modeling the distribution of RNA molecules per droplet	Filtered expression matrix, plots (e.g., PNG)	Mitochondrial contamination assessment	R	https://www.bioconductor.org/packages/release/bioc/html/miQC.html
DropletQC [60]	Expression matrix (e.g., CSV, TSV)	Uses supervised learning algorithms to detect and classify doublets by comparing observed gene expression with simulated doublets	Quality-controlled data (e.g., CSV, HDF5)	Identifies empty droplets and low-quality cells in droplet-based scRNA-seq data	R	https://github.com/powellgenomicslab/DropletQC
Solo [61]	Expression matrix, BAM	Uses supervised learning algorithms to detect and classify doublets by comparing observed gene expression with simulated doublets	Filtered expression matrix, summary stats	Efficiently remove doublets	Python	https://docs.scvi-tools.org/en/stable/user_guide/models/solo.html
Trim Galore [62]	FASTQ	Uses Cutadapt for adapter trimming and FastQC for quality control	Trimmed FASTQ	NA	Python	https://github.com/FelixKrueger/TrimGalore
DecontX [63]	Expression matrix	Bayesian hierarchical model	Decontaminated count matrix	Removes ambient RNA contamination	R	https://bioconductor.org/packages/release/bioc/html/decontX.html
CellBender [64]	Count matrices (HDF5)	Variational inference, deep learning	Cleaned count matrix	Droplet-based scRNA-seq ambient RNA removal	Python	https://github.com/broadinstitute/CellBender
Scater [65]	Single-cell expression matrix (e.g., CSV, TSV)	Uses threshold-based filtering and visualization to assess quality metrics	Quality-controlled expression matrix, plots (e.g., PNG, PDF)	NA	R	https://bioconductor.org/packages/release/bioc/html/scater.html
Seurat [49]	Expression matrix (e.g., CSV, TSV)	Uses heuristic filtering based on gene count, mitochondrial content, and feature expression to remove low-quality cells	Filtered and normalized data, plots, R objects	NA	R	https://satijalab.org/seurat/

Table 4 (continued)

Tool name	Input format	Algorithm framework	Output format	Specific use case	Language	Availability
Scanpy [50]	Expression matrix (e.g., CSV, TSV, H5 AD)	Uses various metrics like gene count, mitochondrial gene percentage, and UMI count for filtering	Processed AnnData (H5 AD), plots	NA	Python	https://scanpy.readthedocs.io/en/stable/
Scrublet [52]	Expression matrix (e.g., CSV, TSV)	Uses a k-NN classifier to detect doublets by comparing observed gene expression profiles with simulated doublets	Doublet scores, data Filtered	Detects doublets	Python	https://github.com/swolock/scrublet
ScPipe [66]	FASTQ, BAM, GTF	Uses a series of filtering steps based on read alignment quality, gene count, and expression levels to remove low-quality cells	Processed expression matrix (e.g., CSV, HDF5)	NA	R	https://github.com/LuyiTian/scPipe
Doublet Finder [67]	Expression matrix (e.g., CSV, TSV)	Machine learning approach to identify doublets by creating artificial doublets and comparing them with observed data	Doublet scores, data filtered	Doublet detection	R	https://github.com/chris-mcginnis-ucsf/DoubletFinder
SoupX [68]	Expression matrix (e.g., CSV, TSV)	Bayesian model to estimate and subtract ambient RNA contamination	Corrected expression matrix (e.g., CSV, HDF5)	Corrects ambient RNA contamination in droplet-based dataset	R	https://github.com/constantAmateur/SoupX
Normalization tools						
Dino [69]	Expression matrix (e.g., CSV, TSV)	Uses a Bayesian latent variable model to estimate and remove technical noise from gene expression data	Normalized expression matrix	Designed to improve signal recovery and account for technical variations	Python	https://www.bioconductor.org/packages/release/bioc/html/Dino.html
Scran [70]	Expression matrix (e.g., CSV, TSV)	Uses a deconvolution approach to compute size factors for normalization, adjusting for cell-specific biases	Normalized expression matrix, size factors	Effective for addressing cell-specific biases	R	https://www.bioconductor.org/packages/release/bioc/html/scran.html
Seurat [49]	Expression matrix (e.g., CSV, TSV)	Uses log-normalization, where counts are divided by total counts per cell and multiplied by a scaling factor, followed by log transformation	Normalized expression matrix, R objects	Provides multiple normalization methods	R	https://satijalab.org/seurat/

Table 4 (continued)

Tool name	Input format	Algorithm framework	Output format	Specific use case	Language	Availability
Scanpy [50]	Expression matrix (e.g., CSV, TSV, H5 AD)	Supports multiple methods including log-normalization, total count scaling, and variance-stabilizing transformations	Normalized AnnData (H5 AD)	Provides multiple normalization methods	Python	https://scanpy.readthedocs.io/en/stable/
DESeq2 [71]	Count matrix (e.g., CSV, TSV)	Uses a median of ratios method to normalize counts, adjusting for differences in sequencing depth and RNA composition between cells	Normalized count matrix, DE results (CSV, TSV)	NA	R	https://bioconductor.org/packages/release/bioc/html/DESeq2.html
SCnorm [72]	Expression matrix (e.g., CSV, TSV)	Uses a quantile regression approach to normalize gene expression, correcting for gene-specific biases across cells	Normalized Expression Matrix	Address gene-specific biases in expression levels	R	https://bioconductor.org/packages/release/bioc/html/SCnorm.html
LIGER [73]	Expression matrix (e.g., CSV, TSV, HDF5)	Uses iNMF to normalize and integrate data across different datasets, enabling joint analysis of multiple datasets	Integrated and normalized matrix	Harmonize multiple dataset	R	https://github.com/welch-lab/liger
Imputation tools						
IGSimpute [74]	Expression matrix (e.g., CSV, TSV)	Uses a graph-based method, where gene similarities are calculated, and missing values are imputed based on similar genes' expression patterns	Imputed expression matrix	Preserving gene-gene relationships	R	https://github.com/ericcombiolab/IGSimpute
SciGAN [75]	Expression matrix (e.g., CSV, TSV)	Uses GANs to model the distribution of observed data and impute missing values by generating realistic data points	Imputed expression matrix	Maintains biological variability	Python	https://github.com/xuyungang/scIGANs
CMF-Impute [76]	Expression matrix (e.g., CSV, TSV)	Uses coupled matrix factorization, where data from multiple related matrices is factorized simultaneously to estimate missing values	Imputed expression matrix	NA	R	https://github.com/xujunlin123/CMFImpute
scIMC [77]	Expression matrix (e.g., CSV, TSV)	Uses IMC learning to impute missing values by leveraging shared information across multiple cell clusters	Imputed expression matrix	NA	Python	https://server.wei-group.net/scIMC/

Table 4 (continued)

Tool name	Input format	Algorithm framework	Output format	Specific use case	Language	Availability
DrImpute [78]	Expression matrix (e.g., CSV, TSV)	Uses clustering-based imputation where similar cells are grouped, and missing values are estimated based on the expression of similar cells	Imputed expression matrix	NA	R	https://github.com/gongx030/DrImpute
Deep Impute [76]	Expression matrix (e.g., CSV, TSV)	Uses deep neural networks to learn the underlying data structure and impute missing values	Imputed expression matrix	NA	Python	https://github.com/lanagarmire/deepimpute
ScLRTC [79]	Expression matrix (e.g., CSV, TSV)	Uses LRTC, where the data is modeled as a tensor, and missing values are imputed by completing the low-rank tensor	Imputed expression matrix	NA	R	https://github.com/jianghuaijie/scLRTC
ScHinter [80]	Expression matrix (e.g., CSV, TSV)	Uses hierarchical Bayesian models to impute missing data, incorporating information from both gene expression and cell clustering	Imputed expression matrix	NA	R	https://github.com/BMILAB/scHinter
MAGIC [81]	Expression matrix (e.g., CSV, TSV)	Uses Markov affinity-based graph imputation of cells, where data is imputed by DGE values across a graph representing cell similarities	Imputed expression matrix	Preserve local gene-gene interactions	Python	https://github.com/KrishnaswamyLab/MAGIC
Batch-effect correction						
Batchelor [82]	Expression matrix (e.g., CSV, TSV)	Uses methods like MNN to correct batch effects by aligning similar cells across batches	Batch-corrected expression matrix	Corrects batch effects while preserving biological variance	R	https://bioconductor.org/packages/release/bioc/html/batchelor.html
Beer [83]	Expression matrix (e.g., CSV, TSV)	Employs a two-stage model where data is modeled first and then batch effects are removed	Batch-corrected expression matrix	Integrates multiple datasets while reducing batch-related noise	R	https://github.com/jumphone/BEER/releases
Seurat [49]	Expression matrix (e.g., CSV, TSV)	Uses CCA and MNN to integrate datasets and correct for batch effects	Batch-corrected expression matrix, R objects	To integrate datasets from different conditions	R	https://satijalab.org/seurat/

Table 4 (continued)

Tool name	Input format	Algorithm framework	Output format	Specific use case	Language	Availability
scVI [84]	Expression matrix (e.g., CSV, TSV, HDF5)	Uses a VAE framework to model and correct batch effects by learning a latent space that separates biological signal from technical noise	Batch-corrected latent space, imputed matrix	Handling complex batch effects	Python	https://github.com/scverse/scvi-tools
MMD-ResNet [85]	Expression matrix (e.g., CSV, TSV)	Uses MMD combined with residual networks to align data across batches and remove batch effects	Batch-corrected expression matrix	NA	R	https://github.com/ushaham/BatchEffectRemoval
scScope [86]	Expression matrix (e.g., CSV, TSV)	Utilizes deep learning to learn a shared latent space that corrects batch effects while preserving biological variation	Batch-corrected expression matrix	NA	Python	https://github.com/AltschulerWu-Lab/scScope
Harmony [87]	Expression matrix (e.g., CSV, TSV)	Uses an iterative algorithm to align cell embeddings across batches while maintaining cell type structure	Batch-corrected embeddings	Preserve biological heterogeneity	R	https://github.com/immunogenomics/harmony
LIGER [73]	Expression matrix (e.g., CSV, TSV, HDF5)	Uses iNMF to jointly factorize data from multiple batches, effectively correcting for batch effects	Batch-corrected and integrated expression	Multi-omics data harmonization	R	https://github.com/welch-lab/liger
scBatch [88]	Expression matrix (e.g., CSV, TSV)	Uses a combination of clustering and empirical Bayesian methods to remove batch effects from single-cell RNA-seq data	Batch-corrected expression matrix	Correct batch effects while retaining meaningful biological signals	R	https://github.com/tengfei-emory/scBatch
Clustering tools						
SC3 [89]	Expression matrix (e.g., CSV, TSV)	Uses consensus clustering by combining results from multiple clustering methods to achieve robust cell groupings	Clustering labels, plots	NA	R	https://github.com/hemberg-lab/SC3
SCENIC [90]	Expression matrix, Gene Networks Regulatory (GRN)	Combines GRN inference with clustering, identifying co-expressed gene sets that drive cell type clustering	Cell clusters, regulatory network activity scores (e.g., CSV, RData)	Identifies regulatory networks and clusters cells based on gene regulatory interactions rather than direct gene expression	R	https://github.com/aertslab/SCENIC

Table 4 (continued)

Tool name	Input format	Algorithm framework	Output format	Specific use case	Language	Availability
BackSPIN [91]	Expression matrix (e.g., CSV, TSV)	Uses a biclustering approach that alternates between clustering genes and samples to achieve optimal partitions	Cell clusters, plots	Useful for complex cellular hierarchies	R	https://github.com/linnarsson-lab/BackSPIN
SIMLR [92]	Expression matrix (e.g., CSV, TSV)	Uses similarity learning to infer cell similarities and clusters by learning a low-dimensional space from the high-dimensional data	Clustering labels, latent embedding, plots	NA	R	https://github.com/BatzoglouLabSU/SIMLR
SAIC [93]	Expression matrix (e.g., CSV, TSV)	Iterative clustering method	Adjusted cell counts, cell clusters, plots	Ambient RNA correction	R	https://github.com/xiweiwu/SAIC.
GiniClust [94]	Expression matrix (e.g., CSV, TSV)	Gini index-based clustering	Cell clusters	Detects rare cell types	R and Python	https://github.com/lanjiangboston/GiniClust
RaceID [95]	Expression matrix (e.g., CSV, TSV)	k-medoids clustering, outlier detection	Cell clusters, CSV, RDS	Rare cell type identification	R	https://github.com/dgrun/RaceID
CIDR [96]	Expression matrix (e.g., CSV, TSV)	Clustering through imputation and dimensionality reduction	Clustering results, CSV, RDS	NA	R and C++	https://github.com/VCCRI/CIDR
GRACE [97]	Expression matrix (e.g., CSV, TSV)	Graph convolutional network (GCN) based clustering	Cell-type assignments CSV, HDF5	NA	R	https://github.com/th00516/GRACE
CosTaL [98]	Expression matrix (e.g., CSV, TSV)	KNN-graph based clustering	Corrected count matrix, CSV, HDF5	Removes technical variations	Python	https://github.com/li000678/CosTaL
DESC [99]	Expression matrix (e.g., CSV, TSV)	Deep learning based clustering	Cell clusters, CSV, RDS	NA	Python	https://eleozzr.github.io/desc/
scziDesk [100]	Expression matrix (e.g., CSV, TSV)	Deep learning (autoencoder-based)	Cluster assignments CSV, HDF5	NA	Python and R	https://github.com/xuebaliang/scziDesk
scVAE [101]	Count matrices	Variational autoencoder (VAE)	Latent representations CSV, HDF5	NA	Python	https://github.com/scvae/scvae
scDeep Cluster [102]	Count matrices	Deep learning-based clustering	Cluster assignments CSV, HDF5	NA	Python	https://github.com/keras-team/keras
SINCERA [103]	Expression matrix (e.g., CSV, TSV)	Utilizes hierarchical clustering and other standard clustering techniques to identify cell types and subtypes in single-cell RNA-seq data	Clustering labels, plots	Identify cell types and subtypes	R	https://research.cchmc.org/pbge/sincera.html

Table 4 (continued)

Tool name	Input format	Algorithm framework	Output format	Specific use case	Language	Availability
SEURAT [49]	Expression matrix (e.g., CSV, TSV)	Uses graph-based clustering, where cells are connected based on shared nearest neighbors, and clusters are identified as connected components in the graph	Clustering labels, UMAP/t-SNE embeddings, plots	NA	R	https://satijalab.org/seurat/
Monocle [104]	Expression matrix (e.g., CSV, TSV)	Uses a density peak clustering approach combined with pseudotime ordering to cluster cells and infer trajectories	Pseudotime trajectories, clusters, plots	NA	R	https://github.com/cole-trapnell-lab/monocle-release
SCRL [105]	Expression matrix (e.g., CSV, TSV)	Uses self-supervised learning to perform clustering in a reduced dimensional space, improving clustering accuracy	Clustering labels, plots	NA	C++	https://github.com/SuntreeLi/SCRL/
MultiK [106]	Expression matrix (e.g., CSV, TSV)	Uses k-means clustering on multiple data representations, combining results to identify robust clusters	Clustering labels, plots	Automates the selection of the optimal number of clusters for better accuracy	R	https://github.com/perou-lab/MultiK
Secuer [107]	Expression matrix (e.g., CSV, TSV)	Uses ensemble clustering methods to stabilize clustering results by combining multiple clustering algorithms	Clustering labels, plots	NA	Python	https://github.com/nanaweei11/Secuer
Scanpy [50]	Expression matrix (e.g., CSV, TSV, H5 AD)	Uses Louvain or Leiden algorithms for clustering, which are graph-based methods that identify clusters based on cell-to-cell similarity networks	Clustering labels, UMAP/t-SNE embeddings, plots	NA	Python	https://scanpy.readthedocs.io/en/stable/
Differential gene expression						
Seurat [49]	Count matrix (e.g., CSV, TSV)	Uses a non-parametric Wilcoxon rank-sum test by default for DE analysis, with options for other statistical tests	Differential expression results (e.g., CSV)	NA	R	https://satijalab.org/seurat/
Monocle [104]	Count matrix, experimental design	Uses a GLM framework to identify genes that vary across cell states or pseudotime, often based on negative binomial distribution	Differential expression results (e.g., CSV), pseudotime analysis	Helps in studying dynamic gene expression changes during cell differentiation	R	http://cole-trapnell-lab.github.io/monocle-release/

Table 4 (continued)

Tool name	Input format	Algorithm framework	Output format	Specific use case	Language	Availability
MAST [108]	FASTQ, expression matrix (e.g., CSV, TSV)	Uses a hurdle model that combines a logistic regression model for zero inflation and a Gaussian linear model for positive counts	Differential expression results (e.g., CSV)	Handles dropout events and sparse data	R	https://github.com/RGLab/MAST
scDE [48]	Expression matrix (e.g., CSV, TSV)	Uses a Bayesian framework that models the dropout events and expression levels, distinguishing between technical noise and true differential expression	Differential expression results (e.g., CSV)	Reduce technical noise	R	https://www.bioconductor.org/packages/release/bioc/html/scde.html
EdgeR [109]	Expression matrix (e.g., CSV, TSV)	Statistical methods based on the negative binomial distribution, including empirical Bayes estimation, exact tests, generalized linear models, and quasi-likelihood tests	Differential expression results (e.g., CSV)	NA	R	https://bioconductor.org/packages/release/bioc/html/edgeR.html
DESeq2 [71]	Count matrix (e.g., CSV, TSV)	Uses a negative binomial GLM for count-based data, estimating dispersion and normalizing counts to perform DE analysis	Differential expression results (e.g., CSV)	NA	R	https://bioconductor.org/packages/release/bioc/html/DESeq2.html
Scotty [110]	Count matrix, experimental design	Allowing users to plan experiments by estimating the power to detect differential expression given certain parameters	Power analysis results, experimental design recommendations (e.g., CSV, TXT)	NA	R	https://github.com/mbusby/Scotty
Myrna [111]	FASTQ, expression matrix (e.g., CSV, TSV)	Uses a TMM normalization method and tests for differential expression using a modified version of the edgeR pipeline	Differential expression results (e.g., CSV)	Cloud-based	R	https://github.com/BenLangmead/myrna
GREIN [112]	Expression matrix (e.g., CSV, TSV)	Provides an online platform using a variety of methods (e.g., DESeq2, edgeR) for DE analysis	Differential expression results (e.g., CSV)	Web-based	R	http://ilincs.org/apps/grein/
D3E [113]	Expression matrix (e.g., CSV, TSV)	Statistical modeling of transcriptional dynamics to detect differential gene expression	Differential expression results (e.g., CSV)	NA	Python	https://hemberg-lab.github.io/D3E/

Within-cell normalization methods like calculating TPM (Transcripts Per Kilobase Million) or RPKM (Reads Per Kilobase of transcript per Million mapped reads)/ FPKM (Fragments Per Kilobase of transcript per Million mapped reads) are commonly used to address sequencing depth within individual cells. But, these methods may not be suitable for certain downstream analyses as they do not account for changes in RNA content and can be misleading when analyzing differentially expressed genes. However, studies in bulk RNA-seq have emphasized the vital role of between-sample normalization. The non-linear normalization method, without UMIs, effectively explores cellular heterogeneity and accurately analyzes scRNA-seq data with high library sizes. This method computes individual normalization factors for each cell and gene by using information from multiple genes and cells, reducing technical biases in scRNA-seq. It is more flexible than traditional size factor normalization, as it estimates using many genes with minimal constraints [117]. For example, DINO works on the principle of non-linear method [69].

In scRNA-seq normalization, two main methods are employed: cell-based and gene-based. The cell-based approach calculates a specific size factor for each cell, used to normalize its gene expression. "Scran" uses this by pooling cells for robust size factor estimation and reducing the impact of excessive zeros. On the other hand, gene-based methods like SCnorm and SCTransform in Seurat adjust genes based on their sequencing depths or abundance levels. This distinction enables precise normalization in scRNA-seq analysis [45, 72, 118, 119].

Among normalization tools for scRNA-seq, Scanpy and Seurat normalization stands in terms of scalability [49, 50]. DESeq2, which is a differential expression tool, has an inbuilt method for normalization and presents concern with false positive rates, which can affect downstream analysis [120]. LIGER is a very useful tool but has limitations in integrating diverse features like gene expression and intergenic methylation, potentially reducing its effectiveness for multiomics studies [73].

3.6 Imputation method

ScRNA-seq data often exhibit missing values, notably zero expression counts for numerous genes. Some of these zeros have biological significance, indicating either gene inactivity or mRNA degradation post-expression. Additionally, technical and sampling factors in scRNA-seq contribute to non-biological zeros, arising from issues like reverse transcription failures, low mRNA quantities, inefficient amplification, or restricted sequencing depth. These non-biological zeros introduce intercellular variability, impede the detection of gene relationships, and exert a notable influence on downstream analyses. In contrast, employing imputation, i.e., replacing missing values with estimated alternatives, proves an effective strategy for addressing these missing data.

Effectively classifying and comparing popular imputation methods is paramount in providing users with informed guidance for various datasets and unique needs (Table 4). These methods can be broadly categorized into four distinct groups:

- Model-based methods rely on statistical models encompassing technical and biological variability, estimating parameters to perform imputation.
- Low-rank matrix-based approaches utilize a low-rank matrix to uncover spatial representations of cells, capturing linear relationships and reconstructing a less sparse expression matrix.
- Data smoothing methods, on the other hand, leverage gene expression values from similar cells to adjust all values, including zeros and non-zeros, employing a smoothing technique.
- Deep learning methods use advanced techniques to identify potential spatial representations of cells and reconstruct the observed expression matrix based on these estimated representations. This classification provides users with a structured framework for selecting the most suitable imputation method tailored to their specific dataset and analytical requirements [121].

Ruochen Jiang stresses the importance of tailoring imputation methods to the specific attributes of single-cell data [122]. For UMIs-based sequencing, which lacks zero-inflation, using imputation methods designed for zero-inflated models is inappropriate. While tasks like cell dimensionality reduction or clustering can often be performed without imputation at the cell level, selecting the correct imputation method is critical for optimal performance in differential expression (DE) analysis. Conversely, non-UMI data benefits from imputation utilizing a non-zero inflation model, particularly for tasks like cell down scaling or clustering. In DE analysis, any imputation method outperforms no imputation or binarization. Ultimately, in cases where the cell library is sufficiently extensive, indicating ample sequencing depth, imputation may be unnecessary. This underscores the importance of aligning imputation strategies precisely with the distinctive characteristics and analytical objectives of single-cell datasets [12].

IGSimpute offers GPU-accelerated imputation but is unsuitable for rare cell types, with training times ranging from 4 min (100,000 cells) to 64 min (1,000,000 cells) using a batch size of 1000 [74]. SciGAN requires

large datasets with at least a few thousand training samples but lacks detailed runtime or scalability data [75]. CMF-Impute may introduce bias when dropout events are abundant, with runtimes ranging from 6 to 12.6 min (0.21 hours), depending on dataset size [123]. DrImpute, achieving an accuracy of 96%, focuses on cell-level correlations but ignores gene-level correlations, limiting its precision. It requires approximately 750 s for 10,000 cells [78]. DeepImpute shows a low mean squared error (MSE = 0.0259) and high correlation (0.984) with ground truth, making it a robust option, running in about 12 min on a dataset with 50,000 cells using 10 GB of RAM on an 8-core machine [76]. ScLRTC provides clustering performance indicators (ARI = 0.7, NMI = 0.8) and takes approximately 8000 s to process 12,500 cells [79]. ScHinter offers a high ARI of 0.9 and a highly variable runtime, from 0.3 to 56.22 s, depending on dataset complexity [80].

The choice of an imputation tool depends on dataset size, computational constraints, and the level of accuracy needed, with DeepImpute, DrImpute, and ScHinter appearing as strong contenders for balancing accuracy and efficiency.

3.7 Batch effect

Variations in single-cell RNA-sequencing data are known to be influenced by technical factors. In some cases, the measurement of biological variations among the samples is affected by these technical factors, making it difficult to address the research problems. Confounding factors in single-cell RNA-sequencing data encompass experimental biases and batch effects. Systematic technical biases, like unequal PCR amplification, cell lysis discrepancies, variable reverse transcriptase enzyme efficiency, and stochastic molecular sampling during sequencing, are unavoidable sources of variation [124] (Table 4). Therefore, batch effect is the major challenge that needs to be resolved before downstream analysis.

Batchelor demonstrates a linear increase in CPU time with dataset size, requiring 2 min for 7000 cells and up to 20 min for 70,000 cells, indicating scalability but potential inefficiencies for extremely large datasets [82]. Beer is a lightweight tool with a runtime of only 1–5 min but lacks further performance details [83]. scVI completes batch correction on 100,000 cells in 25 min, leveraging an NVIDIA Tesla K80 GPU with 24 GB RAM, making it suitable for large-scale analyses [84]. scScope takes under 100 min for 8000 cells across five iterations on a high-performance Xeon E5 CPU with 64 GB RAM and an Nvidia Titan X GPU, suggesting it is computationally demanding [86]. Harmony processes 500,000 cells in 68 min while consuming 7.2GB RAM, offering a balance between efficiency and resource usage [87]. LIGER provides batch correction but has a limitation in multiomics integration, which may affect studies requiring diverse data integration [73]. scBatch is not recommended for highly imbalanced study designs and exhibits scalability issues, handling a few hundred cells in minutes but taking hours for datasets exceeding 1000 cells [88].

The choice of a batch correction tool depends on dataset size, computational resources, and integration requirements, with Harmony and scVI standing out for large-scale studies, while Beer offers a quick but less detailed solution.

3.8 Feature selection and dimensionality reduction

In managing high-dimensional data, dimensionality reduction stands out as a crucial strategy, alongside feature selection (Table 4). When dealing with single-cell RNA-sequencing data, a dual-step approach is often required. Initially, principal component analysis (PCA) is employed to simplify the data. Subsequently, techniques like t-distributed stochastic neighbor embedding (t-SNE) or Uniform Manifold Approximation and Projection (UMAP) are utilized to create visual representations for enhanced comprehension [3]. PCA, a potent mathematical tool, adeptly handles large datasets, preserving both local and long-range patterns. Each principal component acts as a unique axis, orthogonal to the others, enabling the reconstruction of the overall genetic makeup. Determining the number of principal components involves identifying the top ones explaining 80 to 90% of the total variances, or discerning an "elbow point" in the analysis [125].

To address missing or dropout data readings, adaptations of PCA have emerged, incorporating the zero-inflated negative binomial distribution (ZINB) [126]. t-SNE, a non-linear technique, excels in preserving local relationships among data points, effectively segregating clusters. Nevertheless, it may not accurately represent long-range relationships or structures in the data [125, 127]. Diffusion map (DM), another widely used non-linear technique, condenses both nearby and far-reaching patterns into a lower dimension, specially designed to track subtle shifts and transformations in a transcriptome [128]. UMAP, a computationally efficient method, surpasses DM and t-SNE. It captures both local and long-range patterns, recovering global structures in single-cell RNA-sequencing data [129]. Non-linear projection techniques like DM, t-SNE, and UMAP can compress data into 2 or 3 dimensions, but they may introduce distortions and non-biological artifacts, making them primarily recommended for visualization [11].

3.9 Cell clustering

ScRNA-seq clustering helps elucidate cell-to-cell heterogeneity and uncover cell sub-groups and cell dynamics at the group level. Different methods have been created to find different types of cells in single-cell RNA-seq data (Table 4). There are five methods of clustering: K-means clustering, hierarchical clustering, graph based clustering, density-based clustering, and deep learning-based clustering [130]. K-means clustering is a widely used method for grouping data. It works by repeatedly finding a set number of cluster centers (called centroids) in a way that minimizes the total squared distance between each data point and its closest centroid. This method is efficient even with large datasets, as it scales well with the number of data points [131]. SAIC and RaceID both the clustering tools are based on k-means clustering [93, 95].

Hierarchical clustering is widely used in single-cell RNA-seq analysis. It comes in two types: agglomerative, where cells merge based on similarity, and divisive, where clusters are recursively split. These strategies form a hierarchical structure, aiding in identifying rare cell types. Unlike some methods, hierarchical clustering does not require any pre-determined number of clusters, or make assumptions about data distributions. Thus, many single-cell RNA-seq clustering methods include hierarchical clustering [4]. CIDR, BackSPIN, and SINCERA are the clustering tools which are based on hierarchical clustering [91, 96, 103].

DBSCAN is a popular density-based clustering algorithm capable of identifying clusters with arbitrary shapes and outliers. Unlike many clustering methods, DBSCAN does not require the pre-specification of the number of clusters. However, it demands users to set two parameters: ϵ (eps) and the minimum number of points (minPts) to define dense regions that influence DBSCAN clustering [132]. GiniClust and Monocle2 are two tools which are based on DBSCAN clustering [130].

Graph-based clustering, also known as community-based clustering, plays a crucial role in disciplines like sociology, biology, and systems analysis. It is particularly applicable to scenarios represented as interconnected nodes and edges. In single-cell RNA-seq data, nodes represent cells, and connections are determined by pairwise cell-cell distances. The approach involves isolating the branch with the highest weights (cell-to-cell distances) in a dense graph, reflecting cellular relationships. The three primary methods for community detection-based clustering are the clique algorithm, spectral clustering, and the Louvain algorithm [133]. GRACE and CosTaL are well known graph based clustering tools [98, 134].

This deep learning model utilizes a denoising autoencoder to reconstruct uncorrupted data from intentionally corrupted inputs, enabling robust handling of noisy observations. By introducing random Gaussian noise, it simulates minor data variations. The encoder and decoder functions, implemented with rectifier-activated neural networks, process the corrupted input [102]. DESC, scziDesk, scVAE, and scDeepCluster are well known methods that come under deep learning clustering [102, 130]. The evaluation of clustering performance commonly relies on metrics like adjusted R and index for correctness, normalized mutual information (NMI) and Jaccard index for similarity, and Silhouette coefficient and Dunn index for compactness and separateness of clusters [12].

scRNA-seq analysis tools offer diverse capabilities, each with strengths and limitations in terms of computational efficiency, scalability, and sensitivity. SCENIC demonstrates high specificity (0.99) and sensitivity (0.88) for cell-type identification but demands significant memory (128 GB) [90]. SIMLR and Monocle provide efficient analysis times, with Monocle processing 8365 cells in 9 min, making them suitable for rapid computations [92]. DESC achieves high clustering accuracy (adjusted Rand index of 0.919–0.970) and efficiently processes large datasets (30,000 cells) using a NVIDIA TITAN Xp GPU [100]. However, RaceID and CIDR have limitations, with RaceID showing reduced sensitivity to low-expressed genes and CIDR's accuracy decreasing with higher dropout rates [96, 97]. MultiK excels in identifying rare cell populations, even as small as 0.5%, but is computationally expensive [106]. Secuer outperforms traditional clustering methods, being five to twelve times faster than k-means and Louvain/Leiden, making it highly suitable for ultra-large datasets [107]. scDeepCluster scales well for up to 100,000 cells but requires substantial computational resources [102].

Ultimately, tool selection depends on dataset size, computational constraints, and the need for rare cell-type identification, making comparative evaluation crucial for optimizing scRNA-seq analysis.

3.10 Differential expression (DE) analysis

DE analysis is crucial for identifying genes that have significant differences in expression levels between distinct subpopulations, groups of cells, or under specific disease conditions in single-cell RNA-sequencing experiments [135, 136]. Differentially expressed genes (DEGs) play a vital role in understanding the biological differences between compared conditions [120]. Cell states within a population lead to unique gene expression patterns, and data processing methods have considerable impact on the analysis of differential expression, enabling the evaluation of their performance using the results of the analysis [12].

DESeq2 and EdgeR, initially developed for bulk RNA-seq experiments, are also widely utilized in single-cell RNA-seq studies [71, 109]. DESeq2 employs a generalized linear model (GLM) for each gene, incorporating shrinkage estimation for stabilizing variances and fold changes. It applies statistical tests like the Wald test or likelihood ratio (LR) test to assess significance [71, 137]. In contrast, EdgeR fits a GLM with negative binomial (NB) noise for each gene, estimates dispersions using conditional maximum likelihood, and employs a tailored exact test suitable for over dispersed data to identify DEGs [109, 137].

Various methods have emerged to address challenges posed by dropouts and the presence of multiple expression modes in single-cell RNA-sequencing data analysis (Table 4). For example, MAST utilizes a GLM and accounts for dropouts by fitting them to a bimodal distribution, while Monocle incorporates a Tobit model to address dropout events and employs a generalized additive model (GAM) for effective data fitting [108]. SCDE models gene expression as a combination of Zero-Inflated Negative Binomial (ZINB) distributions and uses Bayesian techniques to estimate posterior probabilities for differentially expressed (DE) genes [48]. D3E introduces a novel perspective by modeling the distribution of gene expression through the bursting model of transcriptional regulation. scDD employs a multi-modal Bayesian modeling framework to capture the diverse distributions found in single cells, providing a versatile solution in this complex field of study [45]. Recently, Soneson and Robinson conducted a comprehensive assessment of [36] DE methods, including those designed for both single-cell RNA-seq and bulk RNA-seq data. Their evaluation highlighted substantial differences among these approaches, particularly in terms of the characteristics and quantity of identified DEGs [138]. As the field continues to advance, an increasing number of tools dedicated to the analysis of differential expression in single-cell RNA-sequencing data will be developed. In order to accurately identify DEGs, it is important to select tools specifically designed for scRNA-seq.

Monocle demonstrates the greatest sensitivity (0.765) but also generates a high number of false positives, making it less reliable for certain datasets [104]. MAST, on the other hand, offers high precision but lower sensitivity (0.198) and struggles with highly multi-modal data [108]. EdgeR and DESeq2, originally designed for bulk RNA-seq, achieve intermediate sensitivity (0.58 and 0.695, respectively), but they may not optimally handle zero counts or multi-modality in scRNA-seq [71, 109]. However, DESeq2 performs better than EdgeR, with a higher true positive rate (TPR). D3E shows high sensitivity (0.722) but also introduces false positives [113]. Scotty focuses on optimizing experimental design but may introduce bias against genes with low read counts [110]. GREIN lacks functionalities for downstream analyses, while Myrna lacks available data [111, 112]. Overall, the choice of tool depends on the balance between precision, sensitivity, and computational trade-offs in the analysis of scRNA-seq data.

3.11 Further analysis step

Following differential expression analysis and clustering, several downstream analyses can provide deeper insights into cellular mechanisms. Trajectory inference is employed to map cell differentiation processes, elucidating the progression of cellular states. Cell-cell communication analysis, based on ligand-receptor interactions, helps in understanding intercellular signaling networks. Gene regulatory network construction enables the identification of key transcriptional regulators governing gene expression. Additionally, pathway and functional enrichment analysis facilitates the identification of critical biological pathways associated with cellular functions and disease mechanisms. Furthermore, metabolic and functional state analysis provides a comprehensive view of cellular metabolism and functional alterations, offering insights into disease progression and potential therapeutic targets.

4 Single cell databases

There are multiple databases which offer invaluable resources for researchers delving into various facets of single-cell transcriptomics (Table 5). scRNASeqDB is a repository housing 38 human single-cell transcriptome datasets, which provides researchers access to gene expression profiles across 200 distinct cell types, totaling 13,440 samples [6]. TMExplorer, on the other hand, specializes in TME scRNA-seq datasets, providing access to 48 datasets representing 28 different cancer forms [139]. scREAD is a pivotal resource for Alzheimer's disease research, offering access to 73 datasets across 10 brain regions, providing researchers with information on cell-type predictions and DGEs analyses [140]. SC2 disease is a curated database for exploring cellular heterogeneity across diverse cell types in various diseases, containing 9,46,481 entries categorized into 341 specific cell types, 29 distinct tissues, and 25 different diseases [141]. PlantscRNAdb uniquely focuses on plant species, featuring 26,326 marker genes spanning 128 distinct cell types [142]. EndoDB specializes in endothelial cells, providing curated data from 360 datasets, comprising 4741 bulk and 5847 single-cell endothelial transcriptome [143]. Lastly, SCAD-Brain integrates data from 17 projects related to Alzheimer's disease, encompassing 21 datasets with 359 samples, enabling analyses such as cell marker

Table 5 Databases for scRNA-seq studies and their summary

Database name	Host institute	Nature of database	Link
scRNAseqDB [6]	School of Biomedical Informatics and School of Public Health, University of Texas Health Science Center, USA	Human single cell gene expression datasets	https://bioinfo.uth.edu/scrnaseqdb/
scREAD [140]	Bioinformatics and Mathematical Biosciences Lab, The Ohio University, Ohio	Alzheimer's disease dataset	https://bmbls.bmi.osumc.edu/scread/
SC2 disease [141]	School of Computer Science, Northwestern Polytechnical University, China	Comprehensive datasets	http://easybioai.com/sc2disease/
DRscDB [144]	DRSC-Harvard Medical School, USA	Comprehensive datasets	https://www.flyrnai.org/tools/single_cell/
PlantscRNAdb [142]	Institute of Crop Sciences/Institute of Bioinformatics, Zhejiang University, China	Plant dataset	http://ibi.zju.edu.cn/plantscrnadb/
EndoDB [143]	Carmeliet Lab, VIB - KU Leuven Center for Cancer Biology, Belgium	Endothelial cell transcriptomics data	https://vibcancer.be/software-tools/endodb
SCAD-Brain [145]	Hu Lab, School of Medicine, WUST, China	Datasets of human and mouse brains with Alzheimer's disease	https://www.bioinform.cn/SCAD/
PanglaoDB [146]	Integrated Cardio Metabolic Centre Karolinska Institutet, Blickagången 6, 141 57 Huddinge, Sweden	Comprehensive datasets	https://panglaodb.se/index.html
Single Cell Expression Atlas	EMBL-EBI	Comprehensive datasets	https://www.ebi.ac.uk/gxa/sc/home
Single Cell Portal [147]	Broad Institute of MIT and Harvard	Comprehensive datasets	https://singlecell.broadinstitute.org/single_cell
CELLxGENE [148]	Chan Zuckerberg Initiative, 1180 Main Street, Redwood City, CA 94063, USA	Comprehensive datasets	https://cellxgene.cziscience.com/
Allen Brain Cell Atlas	Allen Institute for Brain Science	Mammalian brain	https://portal.brain-map.org/atlases-and-data/bkp/abc-atlas
CellMarker 2.0 [149]	College of Bioinformatics Science and Technology, Harbin Medical University	Comprehensive datasets	http://bio-bigdata.hrbmu.edu.cn/CellMarker/

analysis, gene expression analysis, and pathway enrichment [143]. These databases collectively provide crucial information on cellular heterogeneity, gene expression profiles, and disease-specific transcriptomic patterns.

5 Current research and gaps

With the help of scRNA-seq, data analysis at single-cell resolution is possible to some extent and is expected to advance in the future, proving to be a vital technique for data analysis. These techniques rely on identifying cellular differences, understanding their communication, and recognizing unique or rare cellular states. We will discuss these approaches to scRNA-seq, how they can be implemented in various research domains, and provide some examples of their applications.

5.1 Cellular heterogeneity

ScRNA-seq is a technique used to explore single-cell expression among a population of cells, characterizing cellular heterogeneity. It identifies unique gene expression profiles, highlighting specific cells that can serve as biomarker for disease diagnosis [150]. Additionally, it can reveal significant transcriptomic changes in disease individuals compared to a healthy ones, suggesting their role in disease progression [151]. For example, different transcriptional profiles in tumors identify immune-evading clones, drug-resistant subpopulations, and cancer stem-like cells, all of which aid in the advancement of the disease and resistance to treatment. Functional heterogeneity among T cells, B cells, and myeloid cells has been revealed by scRNA-seq in the immune system, outlining how distinct immune cells react to infections, inflammatory cues, and antigenic challenges. Therefore, scRNA-seq thoroughly examines important gene expression patterns and uncovers significant biomarkers, receptors, ligands, and transcription factors, which lay the foundation for the functional analysis of cells [75].

5.2 Cell-cell communication

The utilization of single-cell RNA-sequencing data for the purpose of examining cell-to-cell communication is a powerful methodology for elucidating inter-cellular communication pathways. Nevertheless, prevailing approaches commonly do this analysis by focusing on cell categories or clusters, disregarding the intricate details at the level of individual single cells. A novel

approach introduced for analyzing interactions at the single-cell level was Scriabin [152].

Using this method, scientists can map cellular interactions in real time and pinpoint biologically important pathways across a range of tissues. Scriabin has been utilized in immuno-oncology to analyze immune-tumor crosstalk and identify ligand-receptor interactions that promote immune evasion. Precision immunotherapies have been made possible by the discovery of novel checkpoint inhibitors and tumor-supporting stromal signals by researchers using scRNA-seq data analysis. It employs a combination of curated ligand-receptor interaction databases [153, 154], models of downstream intracellular signaling [155], anchor-based dataset integration [49], and gene network analysis [156] to examine intricate communication pathways at the resolution of individual cells. This approach allows for the identification of biologically significant connections between cells at a single-cell level.

5.3 Cell type identification

ScRNA-seq provides a valuable avenue for the complete sequencing and annotation of cell types within various tissues of a given species [156–158]. This technique facilitates the identification of both known and novel cell types, hence enabling an understanding of their associated biological processes and molecular activities. For example, in a sample size of around 25,000 bipolar cells in mice, researchers found two distinct types of unique mouse retinal bipolar cells. Notably, one of these cell types had a shape that deviated from the conventional structure often observed in bipolar cells [159]. Furthermore, significant cellular heterogeneity within the retinal bipolar cell population was revealed by scRNA-seq. As a result, using computational clustering and differential gene expression analyses, researchers were able to identify hidden molecular signatures associated with these morphologically distinct bipolar cells. This method identified a type of atypical bipolar cells that might have important effects on visual processing by playing particular roles in retinal signaling pathways. One excellent example of how scRNA-seq can be used to enhance cellular classifications and find new targets for additional research into retinal development and related visual disorders is this kind of characterization. Moreover, computational approaches for cell type detection do not need manual annotation. Alternatively, these tools may be utilized to make direct predictions of cell types based on publicly available resources of scRNA-seq data [160].

6 Applications

As we discussed about current research in scRNA-seq, it offers plenty of implementation in various domains. It includes drug discovery, microbial profiling, tumor study, stem cell research, and so on. Here, we will describe the importance and how it can be implemented. The following are the applications.

6.1 Drug discovery and development

Since the introduction of whole-transcriptome profiling of a single cell in 2009 [4], this technology has evolved enormously by giving results at the single-cell level. One such example is drug discovery and development, in which scRNA-seq investigations were carried out by Van de Sande et al. [161] on brain tissues obtained from both healthy mice and mouse models of Alzheimer's disease, which revealed the presence of disease-associated microglia. There are differential expression patterns seen in such microglia clusters that point to novel molecular targets that might be taken advantage of to control negative neuroinflammation reactions. The study revealed evidence that particular microglial subsets are preferentially activated under disease conditions, implying that focused treatments could specifically target these pathogenic populations. This high resolution data made possible by scRNA-seq which highlights the need of spotting rare cell types possibly crucial for the course of neuro-degeneration. These findings indicate that a therapeutic approach targeting specific cell states may hold potential benefits for those afflicted with Alzheimer's disease. In the end, scRNA-seq enables the early identification and characterization of therapeutic targets linked to diseases. Early detection of potential issues can ultimately decrease the occurrence of clinical failures, hence, enhancing the efficiency of the drug development process.

6.2 Tumor microenvironment (TME)

ScRNA-seq is a very powerful technology that enables the study of heterogeneous single-cell populations in a TME [162, 163]. Moreover, it gives clarity to marked differences among cells in a cell population. It allows for the comprehensive examination of gene expression patterns in individual cells, which may not be readily apparent in bulk analysis. Furthermore, throughout a study, it enables researchers to identify and analyze the diverse cellular composition within a TME. In the study of glioma, Li et al. identified 14 glioma cellular sub-populations and 7 primary cell types [164], demonstrating the intricate and varied nature of the TME. Since the different cell types exhibit different gene expression patterns, metabolic adaptations, and immune evasion strategies, this heterogeneity is absolutely important in determining

tumor evolution and response to treatment. The study also highlighted how interactions between glioma cells and immune components such as regulatory T cells and tumor-associated macrophages (TAMs) create an immunosuppressive environment that promotes tumor growth. Glioma cells also show metabolic plasticity, meaning they can change between glycolysis and oxidative phosphorylation depending on microenvironmental signals, enabling their adaptation to changes in therapeutic pressure and nutrient availability. Another important discovery was the identification of specific molecular markers and signaling pathways linked to several glioma subtypes, thus providing possible targets for precision treatments. Another study conducted by Ding et al. from Harvard Medical School utilized single-cell profiling techniques to investigate the presence of sub-clonal heterogeneity and identify aggressive disease states in triple-negative breast cancer (TNBC) [165]. Initially, on untreated TNBC tumors, the investigators confirmed the presence of cellular heterogeneity within primary TNBCs with the help of scRNA-seq. Furthermore, employing clustering methods, they have successfully identified five discrete clusters of cells. As this method provides more diverse cells, it is suitable for evaluating cellular heterogeneity in TME. Thus, scRNA-seq could be an appropriate platform for research on TME [166].

6.3 Biomarker discovery

Biomarker discovery is a crucial step for disease diagnosis, and it can be effectively carried out using scRNA-seq analysis. This technology allows for the identification gene biomarkers, their regulatory factors, and their signaling pathways involved in disease mechanisms [167]. In a study of clear cell renal cell carcinoma (ccRCC) [168], Narayanan et al. identified SLC6 A3 as a potent diagnostic and prognostic biomarker using scRNA-seq, paving a way for improved diagnosis and treatment strategies. Its specificity for malignant tissues is demonstrated by the fact that SLC6 A3 expression is markedly increased in ccRCC tumor cells but absent in immune cells and benign kidney tubules. This finding was confirmed by a number of datasets, such as scRNA-seq profiles, microarray datasets (GSE40435, GSE53757), and TCGA pancancer expression data, thereby confirming its potential as a diagnostic marker. SLC6 A3 is a promising candidate for early disease detection because of its high sensitivity and specificity in differentiating ccRCC from normal kidney tissue, as further demonstrated by receiver operating characteristics (ROC) analysis. Therefore, it is important to utilize this technique in biomarker discovery, as it plays a vital role in uncovering the molecular mechanism of diseases and determines strategies for tackling them [169].

6.4 Microbes profiling

Different subpopulations of bacteria within a community can exhibit diverse gene expression patterns and dynamically adjust to challenging conditions. This expression heterogeneity, which is prevalent in natural micro-biota, is difficult to capture using bulk sequencing approaches. However, microbes present challenges, such as cell size and cell wall variation, as well as low of mRNA content per cell, and the absence of poly(A) tails in mRNA. To address these issues, Pu et al. developed a novel method called Ribosomal RNA-derived cDNA Depletion (RiboD), integrated into the PETRI-Seq technique, to capture single-cell transcriptomes of Gram-positive and Gram-negative bacteria with high purity and low bias [170, 171]. This method, known as RiboD-PETRI, offers a high-throughput, cost-effective solution for bacterial scRNA-seq. It allows precise exploration of bacterial population heterogeneity in biofilms and microbiomes revealing subpopulations with distinct gene expression profiles that influence the dynamics and behavior of biofilms communities. Using RiboD-PETRI, scientists can now see microbial communities more clearly and thoroughly, especially in settings like microbiomes and biofilms where cellular heterogeneity is essential for adaptation and survival. By identifying transcriptional alterations that reflect metabolic changes, the emergence of antibiotic resistance, and the regulation of quorum sensing, the method makes it easier to identify functionally distinct bacterial subpopulations within biofilms. A deeper understanding of how subpopulations contribute to community structure and function is made possible by scRNA-seq with RiboD-PETRI, which reveals rare phenotypic variants in contrast to bulk sequencing, which averages gene expression across a population. These findings highlight the importance of understanding microbial heterogeneity in developing therapeutics [170].

6.5 Stem cell research

The use of single-cell sequencing technology offers distinct advantages in comprehending the occurrence and progression of stem cells. In the course of stem cell growth, there exists temporal variation in gene expression, posing challenges that have been arduous to address using conventional methodologies. The utilization of single-cell sequencing enables researchers to direct their attention toward a solitary cell, whether it is seen as an independent entity within a larger cell population or as a representative of a certain subpopulation across several developmental phases. The integration of single-cell sequencing with other sophisticated methodologies holds great potential for further enhancing scientific inquiry. For instance, the combination of scRNA-seq with patch-clamp technology presents

an intriguing avenue for investigating the underlying mechanisms of neuropsychiatric disorders, therefore unraveling its fundamental essence [172].

Human primordial germ cells (hPGCs) serve as the progenitors for fully developed germ cells. The transcriptomes of hPGCs at the single-cell level exhibit a notable degree of homogeneity during both the migration and gonad phases [173]. Li et al. endeavored to elucidate the trajectory of development and variability of fetal female germ cells [174]. A comprehensive analysis was conducted on over 2000 germ cells and their corresponding gonadal niche cells using scRNA-seq across many developmental stages. The primary findings of this investigation encompass the identification of distinct transcriptome attributes shown by transcription factor networks across several stages of development. It allows researchers to identify gene regulatory networks and stage specific transcription factors that govern the transition from pleuripotency to lineage commitment. The study shed light on the sequential activation of important pathways involved in germ cell maturation by identifying different transcriptomic signatures in early migrating hPGCs compared to those that had reached the gonadal ridge. Additionally, the use of scRNA-seq has revealed previously unexplored subpopulations in the germline during development, emphasizing the existence of transcriptionally unique cells that might have specialized functions in gametogenesis.

7 Conclusion

In conclusion, the evolution of scRNA-seq has revolutionized our understanding of cellular complexities and heterogeneity, paving the way for advanced research across various biological landscapes. With the establishment of novel methodologies, tools, and databases, researchers can now delve deeper into the mechanisms governing gene expression at an individual cell level, addressing pivotal challenges in drug discovery, TMEs, and cellular communication. As we continue to explore the vast potential of scRNA-seq, it becomes increasingly essential to adopt tailored computational tools that enhance data accuracy, mitigate biases, and refine analysis techniques. By harnessing the latest advancements and remaining cognizant of existing gaps, the scientific community can leverage scRNA-seq to uncover critical biological insights, ultimately driving forward our understanding of health and disease in unprecedented ways.

Acknowledgements
The authors acknowledge the Systems Biology Lab, Indian Institute of Information Technology, Allahabad, Uttar Pradesh, India, for providing computing facility.

Authors' contributions
A.A. and P.T. contributed equally as joint first authors. A.A. and P.T. conceptualized the study, conducted literature survey, and drafted the manuscript. N.D. and I.A. performed literature survey and helped in organizing the manuscript. P.K.V. provided guidance throughout and finalized the manuscript as the corresponding author. All authors have read and approved the final manuscript.

Funding
Not applicable.

Data availability
No datasets were generated or analyzed during the current study.

Declarations

Ethics approval and consent to participate
Not applicable

Consent for publication
The authors are giving their consent for this publication.

Competing interests
The authors declare no competing interests.

Received: 22 January 2025 Accepted: 7 April 2025
Published online: 17 May 2025

References

1. Ribatti D. An historical note on the cell theory. Exp Cell Res. 2018;364(1):1–4.
2. Coons AH, Creech HJ, Jones RN. Immunological properties of an antibody containing a fluorescent group. Proc Soc Exp Biol Med. 1941;47(2):200–2.
3. Jovic D, Liang X, Zeng H, Lin L, Xu F, Luo Y. Single-cell RNA sequencing technologies and applications: a brief overview. Clin Transl Med. 2022;12(3):e694.
4. Wu X, Yang B, Udo-Inyang I, Ji S, Ozog D, Zhou L, et al. Research techniques made simple: single-cell RNA sequencing and its applications in dermatology. J Investig Dermatol. 2018;138(5):1004–9.
5. Tang F, Barbacioru C, Wang Y, Nordman E, Lee C, Xu N, et al. mRNA-Seq whole-transcriptome analysis of a single cell. Nat Methods. 2009;6(5):377–82.
6. Cao Y, Zhu J, Jia P, Zhao Z. scRNASeqDB: a database for RNA-Seq based gene expression profiles in human single cells. Genes. 2017;8(12):368.
7. Pereira WJ, Almeida FM, Conde D, Balmant K, Triozzi P, Schmidt H, et al. Asc-Seurat: analytical single-cell Seurat-based web application. BMC Bioinformatics. 2021;22:1–14.
8. Chen G, Ning B, Shi T. Single-cell RNA-seq technologies and related computational data analysis. Front Genet. 2019;10:317.
9. Rosenberg AB, Roco CM, Muscat RA, Kuchina A, Sample P, Yao Z, et al. Single-cell profiling of the developing mouse brain and spinal cord with split-pool barcoding. Science. 2018;360(6385):176–82.
10. Liu S, Trapnell C. Single-cell transcriptome sequencing: recent advances and remaining challenges. F1000Research. 2016;5(F1000 Faculty Rev):182.
11. Kiselev VY, Andrews TS, Hemberg M. Challenges in unsupervised clustering of single-cell RNA-seq data. Nat Rev Genet. 2019;20(5):273–82.
12. Lu J, Sheng Y, Qian W, Pan M, Zhao X, Ge Q. scRNA-seq data analysis method to improve analysis performance. IET Nanobiotechnology. 2023;17(3):246–56.
13. Cole MB, Risso D, Wagner A, DeTomaso D, Ngai J, Purdom E, et al. Performance assessment and selection of normalization procedures for single-cell RNA-seq. Cell Syst. 2019;8(4):315–28.
14. Kharchenko PV. The triumphs and limitations of computational methods for scRNA-seq. Nat Methods. 2021;18(7):723–32.

15. 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.
16. Wu Y, Zhang K. Tools for the analysis of high-dimensional single-cell RNA sequencing data. Nat Rev Nephrol. 2020;16(7):408–21.
17. 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.
18. Lun AT, McCarthy DJ, Marioni JC. A step-by-step workflow for low-level analysis of single-cell rna-seq data with bioconductor. F1000Research. 2016;5:2122.
19. Ziegenhain C, Vieth B, Parekh S, Reinius B, Guillaumet-Adkins A, Smets M, et al. Comparative analysis of single-cell RNA sequencing methods. Mol Cell. 2017;65(4):631–43.
20. 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.
21. 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.
22. 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(1):14049.
23. Lake BB, Chen S, Sos BC, Fan J, Kaeser GE, Yung YC, et al. Integrative single-cell analysis of transcriptional and epigenetic states in the human adult brain. Nat Biotechnol. 2018;36(1):70–80.
24. Peterson VM, Zhang KX, Kumar N, Wong J, Li L, Wilson DC, et al. Multiplexed quantification of proteins and transcripts in single cells. Nat Biotechnol. 2017;35(10):936–9.
25. Picelli S. Full-length single-cell RNA sequencing with smart-seq2. Single Cell Methods Sequencing Proteomics. 2019;1979:25–44.
26. Bageritz J, Raddi G. Single-cell RNA sequencing with drop-seq. Single Cell Methods: Sequencing and Proteomics. 2019;1979:73–85.
27. Zilionis R, Nainys J, Veres A, Savova V, Zemmour D, Klein AM, et al. Single-cell barcoding and sequencing using droplet microfluidics. Nat Protocol. 2017;12(1):44–73.
28. Natarajan KN. Single-cell tagged reverse transcription (STRT-Seq). Single Cell Methods Sequencing Proteomics. 2019;1979:133–53.
29. 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;17:1–7.
30. Gierahn TM, Wadsworth MH, Hughes TK, Bryson BD, Butler A, Satija R, et al. Seq-Well: portable, low-cost RNA sequencing of single cells at high throughput. Nat Methods. 2017;14(4):395–8.
31. Sasagawa Y, Danno H, Takada H, Ebisawa M, Tanaka K, Hayashi T, et al. Quartz-Seq2: a high-throughput single-cell RNA-sequencing method that effectively uses limited sequence reads. Genome Biol. 2018;19:1–24.
32. Habib N, Avraham-Davidi I, Basu A, Burks T, Shekhar K, Hofree M, et al. Massively parallel single-nucleus RNA-seq with DroNc-seq. Nat Methods. 2017;14(10):955–8.
33. Cao J, Packer JS, Ramani V, Cusanovich DA, Huynh C, Daza R, et al. Comprehensive single-cell transcriptional profiling of a multicellular organism. Science. 2017;357(6352):661–7.
34. Kuchina A, Brettner LM, Paleologu L, Roco CM, Rosenberg AB, Carignano A, et al. Microbial single-cell RNA sequencing by split-pool barcoding. Science. 2021;371(6531):eaba5257.
35. Sheng K, Cao W, Niu Y, Deng Q, Zong C. Effective detection of variation in single-cell transcriptomes using MATQ-seq. Nat Methods. 2017;14(3):267–70.
36. Kim J, Marignani PA. Single-cell RNA sequencing analysis using Fluidigm C1 platform for characterization of heterogeneous transcriptomes. In: Cancer cell biology: methods and protocols. New York: Springer, Humana; 2022. pp. 261–78.
37. Haque A, Engel J, Teichmann SA, Lönnberg T. A practical guide to single-cell RNA-sequencing for biomedical research and clinical applications. Genome Med. 2017;9:1–12.
38. Segeren HA, Andree KC, Oomens L, Westendorp B. Collection of cells for single-cell RNA sequencing using high-resolution fluorescence microscopy. STAR Protoc. 2021;2(3):100718.
39. Zhou WM, Yan YY, Guo QR, Ji H, Wang H, Xu TT, et al. Microfluidics applications for high-throughput single cell sequencing. J Nanobiotechnology. 2021;19:1–21.
40. Kuijpers L, Hornung B, van den Hout-van MCGN, van IJcken WFJ, Grosveld F, Mulugeta E, et al. Split Pool Ligation-based Single-cell Transcriptome sequencing (SPLiT-seq) data processing pipeline comparison. BMC Genomics. 2024;25(1):361.
41. Slyper M, Porter CB, Ashenberg O, Waldman J, Drokhlyansky E, Wakiro I, et al. A single-cell and single-nucleus RNA-Seq toolbox for fresh and frozen human tumors. Nat Med. 2020;26(5):792–802.
42. Mandlik JS, Patil AS, Singh S. Next-generation sequencing (NGS): platforms and applications. J Pharm Bioallied Sci. 2024;16(Suppl 1):S41–5.
43. Slatko BE, Gardner AF, Ausubel FM. Overview of next-generation sequencing technologies. Curr Protocol Mol Biol. 2018;122(1):e59.
44. Udaondo Z, Sittikankaew K, Uengwetwanit T, Wongsurawat T, Sonthirod C, Jenjaroenpun P, et al. Comparative analysis of PacBio and Oxford nanopore sequencing technologies for transcriptomic landscape identification of Penaeus monodon. Life. 2021;11(8):862.
45. Wang M, Song WM, Ming C, Wang Q, Zhou X, Xu P, et al. Guidelines for bioinformatics of single-cell sequencing data analysis in Alzheimer's disease: review, recommendation, implementation and application. Mol Neurodegener. 2022;17(1):17.
46. Marinov GK, Williams BA, McCue K, Schroth GP, Gertz J, Myers RM, et al. From single-cell to cell-pool transcriptomes: stochasticity in gene expression and RNA splicing. Genome Res. 2014;24(3):496–510.
47. Brennecke P, Anders S, Kim JK, Kołodziejczyk AA, Zhang X, Proserpio V, et al. Accounting for technical noise in single-cell RNA-seq experiments. Nat Methods. 2013;10(11):1093–5.
48. Kharchenko PV, Silberstein L, Scadden DT. Bayesian approach to single-cell differential expression analysis. Nat Methods. 2014;11(7):740–2.
49. Stuart T, Butler A, Hoffman P, Hafemeister C, Papalexi E, Mauck WM, et al. Comprehensive integration of single-cell data. Cell. 2019;177(7):1888–902.
50. Wolf FA, Angerer P, Theis FJ. SCANPY: large-scale single-cell gene expression data analysis. Genome Biol. 2018;19:1–5.
51. Krentz NA, Lee MY, Xu EE, Sproul SL, Maslova A, Sasaki S, et al. Single-cell transcriptome profiling of mouse and hESC-derived pancreatic progenitors. Stem Cell Rep. 2018;11(6):1551–64.
52. Wolock SL, Lopez R, Klein AM. Scrublet: computational identification of cell doublets in single-cell transcriptomic data. Cell Syst. 2019;8(4):281–91.
53. Bais AS, Kostka D. scds: computational annotation of doublets in single-cell RNA sequencing data. Bioinformatics. 2020;36(4):1150–8.
54. Xi NM, Li JJ. Benchmarking computational doublet-detection methods for single-cell RNA sequencing data. Cell Syst. 2021;12(2):176–94.
55. Galow AM, Kussauer S, Wolfien M, Brunner RM, Goldammer T, David R, et al. Quality control in scRNA-Seq can discriminate pacemaker cells: the mtRNA bias. Cell Mol Life Sci. 2021;78(19–20):6585–92.
56. Osorio D, Cai JJ. Systematic determination of the mitochondrial proportion in human and mice tissues for single-cell RNA-sequencing data quality control. Bioinformatics. 2021;37(7):963–7.
57. Andrews S, et al. FastQC: a quality control tool for high throughput sequence data. Cambridge, UK; 2010. https://github.com/sandrews/FastQC.
58. Grandi F, Caroli J, Romano O, Marchionni M, Forcato M, Bicciato S. popsicleR: AR package for pre-processing and quality control analysis of single cell RNA-seq data. J Mol Biol. 2022;434(11):167560.
59. Hippen AA, Falco MM, Weber LM, Erkan EP, Zhang K, Doherty JA, et al. miQC: an adaptive probabilistic framework for quality control of single-cell RNA-sequencing data. PLoS Comput Biol. 2021;17(8):e1009290.
60. Muskovic W, Powell JE. DropletQC: improved identification of empty droplets and damaged cells in single-cell RNA-seq data. Genome Biol. 2021;22:1–9.
61. 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.
62. Krueger F. Trim Galore!: a wrapper around Cutadapt and FastQC to consistently apply adapter and quality trimming to FastQ files, with extra functionality for RRBS data. Babraham Institute; 2015. https://github.com/FelixKrueger/TrimGalore.

63. 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;21:1–15.
64. 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 Cell Bender. Nat Methods. 2023;20(9):1323–35.
65. 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.
66. 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.
67. 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.
68. Young MD, Behjati S. SoupX removes ambient RNA contamination from droplet-based single-cell RNA sequencing data. Gigascience. 2020;9(12):giaa151.
69. Brown J, Ni Z, Mohanty C, Bacher R, Kendziorski C. Normalization by distributional resampling of high throughput single-cell RNA-sequencing data. Bioinformatics. 2021;37(22):4123–8.
70. Lytal N, Ran D, An L. Normalization methods on single-cell RNA-seq data: an empirical survey. Front Genet. 2020;11:41.
71. Love MI, Huber W, Anders S. Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. Genome Biol. 2014;15:1–21.
72. Bacher R, Chu LF, Leng N, Gasch AP, Thomson JA, Stewart RM, et al. SCnorm: robust normalization of single-cell RNA-seq data. Nat Methods. 2017;14(6):584–6.
73. 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.
74. Xu K, Cheong C, Veldsman WP, Lyu A, Cheung WK, Zhang L. Accurate and interpretable gene expression imputation on scRNA-seq data using IGSimpute. Brief Bioinforma. 2023;24(3):bbad124.
75. Xu X, Hua X, Mo H, Hu S, Song J. Single-cell RNA sequencing to identify cellular heterogeneity and targets in cardiovascular diseases: from bench to bedside. Basic Res Cardiol. 2023;118(1):7.
76. Arisdakessian C, Poirion O, Yunits B, Zhu X, Garmire LX. DeepImpute: an accurate, fast, and scalable deep neural network method to impute single-cell RNA-seq data. Genome Biol. 2019;20:1–14.
77. Dai C, Jiang Y, Yin C, Su R, Zeng X, Zou Q, et al. scIMC: a platform for benchmarking comparison and visualization analysis of scRNA-seq data imputation methods. Nucleic Acids Res. 2022;50(9):4877–99.
78. Gong W, Kwak IY, Pota P, Koyano-Nakagawa N, Garry DJ. DrImpute: imputing dropout events in single cell RNA sequencing data. BMC Bioinformatics. 2018;19:1–10.
79. Pan X, Li Z, Qin S, Yu M, Hu H. ScLRTC: imputation for single-cell RNA-seq data via low-rank tensor completion. BMC Genomics. 2021;22:1–19.
80. Ye P, Ye W, Ye C, Li S, Ye L, Ji G, et al. scHinter: imputing dropout events for single-cell RNA-seq data with limited sample size. Bioinformatics. 2020;36(3):789–97.
81. 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.
82. Haghverdi L, Lun AT, 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.
83. Zhang F, Wu Y, Tian W. A novel approach to remove the batch effect of single-cell data. Cell Discov. 2019;5(1):46.
84. Gayoso A, Lopez R, Xing G, Boyeau P, Valiollah Pour Amiri V, Hong J, et al. A Python library for probabilistic analysis of single-cell omics data. Nat Biotechnol. 2022;40(2):163–6.
85. Shaham U, Stanton KP, Zhao J, Li H, Raddassi K, Montgomery R, et al. Removal of batch effects using distribution-matching residual networks. Bioinformatics. 2017;33(16):2539–46.
86. Deng Y, Bao F, Dai Q, Wu LF, Altschuler SJ. Scalable analysis of cell-type composition from single-cell transcriptomics using deep recurrent learning. Nat Methods. 2019;16(4):311–4.
87. 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.
88. Fei T, Yu T. scBatch: batch-effect correction of RNA-seq data through sample distance matrix adjustment. Bioinformatics. 2020;36(10):3115–23.
89. 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.
90. 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.
91. Zeisel A, Muñoz-Manchado AB, Codeluppi S, Lönnerberg P, La Manno G, Juréus A, et al. Cell types in the mouse cortex and hippocampus revealed by single-cell RNA-seq. Science. 2015;347(6226):1138–42.
92. 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.
93. Yang L, Liu J, Lu Q, Riggs AD, Wu X. SAIC: an iterative clustering approach for analysis of single cell RNA-seq data. BMC Genomics. 2017;18:9–17.
94. 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;17:1–13.
95. Grün 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.
96. Lin P, Troup M, Ho JW. CIDR: ultrafast and accurate clustering through imputation for single-cell RNA-seq data. Genome Biol. 2017;18:1–11.
97. Yu H, Wang Y, Zhang X, Wang Z. GRACE: a comprehensive web-based platform for integrative single-cell transcriptome analysis. NAR Genomics Bioinforma. 2023;5(2):lqad050.
98. Li Y, Nguyen J, Anastasiu DC, Arriaga EA. CosTaL: an accurate and scalable graph-based clustering algorithm for high-dimensional single-cell data analysis. Brief Bioinformatics. 2023;24(3):bbad157.
99. Li X, Wang K, Lyu Y, Pan H, Zhang J, Stambolian D, et al. Deep learning enables accurate clustering with batch effect removal in single-cell RNA-seq analysis. Nat Commun. 2020;11(1):2338.
100. Chen L, Wang W, Zhai Y, Deng M. Deep soft K-means clustering with self-training for single-cell RNA sequence data. NAR Genomics Bioinforma. 2020;2(2):lqaa039.
101. Grønbech CH, Vording MF, Timshel PN, Sønderby CK, Pers TH, Winther O. scVAE: variational auto-encoders for single-cell gene expression data. Bioinformatics. 2020;36(16):4415–22.
102. Tian T, Wan J, Song Q, Wei Z. Clustering single-cell RNA-seq data with a model-based deep learning approach. Nat Mach Intell. 2019;1(4):191–8.
103. 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.
104. 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.
105. Li X, Chen W, Chen Y, Zhang X, Gu J, Zhang MQ. Network embedding-based representation learning for single cell RNA-seq data. Nucleic Acids Res. 2017;45(19):e166.
106. Liu S, Thennavan A, Garay JP, Marron J, Perou CM. MultiK: an automated tool to determine optimal cluster numbers in single-cell RNA sequencing data. Genome Biol. 2021;22:1–21.
107. Wei N, Nie Y, Liu L, Zheng X, Wu HJ. Secuer: ultrafast, scalable and accurate clustering of single-cell RNA-seq data. PLOS Comput Biol. 2022;18(12):e1010753.
108. 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;16:1–13.
109. 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.
110. Busby MA, Stewart C, Miller CA, Grzeda KR, Marth GT. Scotty: a web tool for designing RNA-Seq experiments to measure differential gene expression. Bioinformatics. 2013;29(5):656–7.

111. Langmead B, Hansen KD, Leek JT. Cloud-scale RNA-sequencing differential expression analysis with Myrna. Genome Biol. 2010;11:1–11.
112. Mahi NA, Najafabadi MF, Pilarczyk M, Kouril M, Medvedovic M. GREIN: an interactive web platform for re-analyzing GEO RNA-seq data. Sci Rep. 2019;9(1):7580.
113. Delmans M, Hemberg M. Discrete distributional differential expression (D 3 E)-a tool for gene expression analysis of single-cell RNA-seq data. BMC Bioinformatics. 2016;17:1–13.
114. Mathys H, Davila-Velderrain J, Peng Z, Gao F, Mohammadi S, Young JZ, et al. Single-cell transcriptomic analysis of Alzheimer's disease. Nature. 2019;570(7761):332–7.
115. Schirmer L, Velmeshev D, Holmqvist S, Kaufmann M, Werneburg S, Jung D, et al. Neuronal vulnerability and multilineage diversity in multiple sclerosis. Nature. 2019;573(7772):75–82.
116. 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.
117. Wu Z, Su K, Wu H. Non-linear normalization for non-UMI single cell RNA-Seq. Front Genet. 2021;12:612670.
118. L Lun AT, Bach K, Marioni JC. Pooling across cells to normalize single-cell RNA sequencing data with many zero counts. Genome Biol. 2016;17:1–14.
119. 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.
120. McDavid A, Finak G, Chattopadyay PK, Dominguez M, Lamoreaux L, Ma SS, et al. Data exploration, quality control and testing in single-cell qPCR-based gene expression experiments. Bioinformatics. 2013;29(4):461–7.
121. Wang M, Gan J, Han C, Guo Y, Chen K, Shi YZ, et al. Imputation methods for scRNA sequencing data. Appl Sci. 2022;12(20):10684.
122. Jiang R, Sun T, Song D, Li JJ. Statistics or biology: the zero-inflation controversy about scRNA-seq data. Genome Biol. 2022;23(1):31.
123. Xu J, Cai L, Liao B, Zhu W, Yang J. CMF-Impute: an accurate imputation tool for single-cell RNA-seq data. Bioinformatics. 2020;36(10):3139–47.
124. 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;21:1–32.
125. Van der Maaten L, Hinton G. Visualizing data using t-SNE. J Mach Learn Res. 2008;9(86):2579–605.
126. Risso D, Perraudeau F, Gribkova S, Dudoit S, Vert JP. A general and flexible method for signal extraction from single-cell RNA-seq data. Nat Commun. 2018;9(1):284.
127. Andrews TS, Hemberg M. Identifying cell populations with scRNASeq. Mol Asp Med. 2018;59:114–22.
128. Haghverdi L, Büttner M, Wolf FA, Buettner F, Theis FJ. Diffusion pseudotime robustly reconstructs lineage branching. Nat Methods. 2016;13(10):845–8.
129. Becht E, McInnes L, Healy J, Dutertre CA, Kwok IW, Ng LG, et al. Dimensionality reduction for visualizing single-cell data using UMAP. Nat Biotechnol. 2019;37(1):38–44.
130. Zhang S, Li X, Lin J, Lin Q, Wong KC. Review of single-cell RNA-seq data clustering for cell-type identification and characterization. Rna. 2023;29(5):517–30.
131. Lloyd S. Least squares quantization in PCM. IEEE Trans Inf Theory. 1982;28(2):129–37.
132. Ester M, Kriegel HP, Sander J, Xu X, et al. A density-based algorithm for discovering clusters in large spatial databases with noise. In: KDD-96 Proceedings. Menlo Park: AAAI Press; 1996. vol. 96. pp. 226–231.
133. Blondel VD, Guillaume JL, Lambiotte R, Lefebvre E. Fast unfolding of communities in large networks. J Stat Mech Theory Exp. 2008;2008(10):P10008.
134. Guan J, Li RY, Wang J. Grace: a graph-based cluster ensemble approach for single-cell rna-seq data clustering. IEEE Access. 2020;8:166730–41.
135. 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.
136. Shalek AK, Satija R, Adiconis X, Gertner RS, Gaublomme JT, Raychowdhury R, et al. Single-cell transcriptomics reveals bimodality in expression and splicing in immune cells. Nature. 2013;498(7453):236–40.
137. Rostom R, Svensson V, Teichmann SA, Kar G. Computational approaches for interpreting sc RNA-seq data. FEBS Lett. 2017;591(15):2213–25.
138. Soneson C, Robinson MD. Bias, robustness and scalability in single-cell differential expression analysis. Nat Methods. 2018;15(4):255–61.
139. Christensen E, Naidas A, Chen D, Husic M, Shooshtari P. TMExplorer: a tumour microenvironment single-cell RNAseq database and search tool. PLoS ONE. 2022;17(9):e0272302.
140. Jiang J, Wang C, Qi R, Fu H, Ma Q. scREAD: a single-cell RNA-Seq database for Alzheimer's disease. Iscience. 2020;23(11).
141. Zhao T, Lyu S, Lu G, Juan L, Zeng X, Wei Z, et al. SC2disease: a manually curated database of single-cell transcriptome for human diseases. Nucleic Acids Res. 2021;49(D1):D1413–9.
142. Chen H, Yin X, Guo L, Yao J, Ding Y, Xu X, et al. PlantscRNAdb: a database for plant single-cell RNA analysis. Mol Plant. 2021;14(6):855–7.
143. Khan S, Taverna F, Rohlenova K, Treps L, Geldhof V, de Rooij L, et al. EndoDB: a database of endothelial cell transcriptomics data. Nucleic Acids Res. 2019;47(D1):D736–44.
144. Hu Y, Tattikota SG, Liu Y, Comjean A, Gao Y, Forman C, et al. DRscDB: a single-cell RNA-seq resource for data mining and data comparison across species. Comput Struct Biotechnol J. 2021;19:2018–26.
145. Li XW, Duan TT, Chu JY, Pan SY, Zeng Y, Hu FF. SCAD-Brain: a public database of single cell RNA-seq data in human and mouse brains with Alzheimer's disease. Front Aging Neurosci. 2023;15:1157792.
146. Franzén O, Gan LM, Björkegren JL. PanglaoDB: a web server for exploration of mouse and human single-cell RNA sequencing data. Database. 2019;2019:baz046.
147. Tarhan L, Bistline J, Chang J, Galloway B, Hanna E, Weitz E. Single Cell Portal: an interactive home for single-cell genomics data. BioRxiv. 2023.
148. Megill C, Martin B, Weaver C, Bell S, Prins L, Badajoz S, et al. Cellxgene: a performant, scalable exploration platform for high dimensional sparse matrices. BioRxiv. 2021;2021–04.
149. Hu C, Li T, Xu Y, Zhang X, Li F, Bai J, et al. CellMarker 2.0: an updated database of manually curated cell markers in human/mouse and web tools based on scRNA-seq data. Nucleic Acids Res. 2023;51(D1):D870–6.
150. Amancherla K, Schlendorf KH, Chow N, Sheng Q, Freedman JE, Rathmell JC. Single-cell RNA-sequencing identifies unique cell-specific gene expression profiles in high-grade cardiac allograft vasculopathy. BioRxiv. 2024.
151. Russell-Hallinan A, Cappa O, Kerrigan L, Tonry C, Edgar K, Glezeva N, et al. Single-cell RNA sequencing reveals cardiac fibroblast-specific transcriptomic changes in dilated cardiomyopathy. Cells. 2024;13(9):752.
152. Wilk AJ, Shalek AK, Holmes S, Blish CA. Comparative analysis of cell-cell communication at single-cell resolution. Nat Biotechnol. 2024;42(3):470–83.
153. Türei 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.
154. Türei D, Korcsmáros T, Saez-Rodriguez J. OmniPath: guidelines and gateway for literature-curated signaling pathway resources. Nat Methods. 2016;13(12):966–7.
155. Browaeys R, Saelens W, Saeys Y. NicheNet: modeling intercellular communication by linking ligands to target genes. Nat Methods. 2020;17(2):159–62.
156. The Tabula Muris Consortium, Overall coordination, Logistical coordination, Organ collection and processing, Library preparation and sequencing, Computational data analysis, et al. Single-cell transcriptomics of 20 mouse organs creates a *Tabula Muris*. Nature. 2018;562(7727):367–372.
157. Han X, Wang R, Zhou Y, Fei L, Sun H, Lai S, et al. Mapping the mouse cell atlas by microwell-seq. Cell. 2018;172(5):1091–107.
158. Han X, Zhou Z, Fei L, Sun H, Wang R, Chen Y, et al. Construction of a human cell landscape at single-cell level. Nature. 2020;581(7808):303–9.
159. Shekhar K, Lapan SW, Whitney IE, Tran NM, Macosko EZ, Kowalczyk M, et al. Comprehensive classification of retinal bipolar neurons by single-cell transcriptomics. Cell. 2016;166(5):1308–23.
160. Xie B, Jiang Q, Mora A, Li X. Automatic cell type identification methods for single-cell RNA sequencing. Comput Struct Biotechnol J. 2021;19:5874–87.

161. Van de Sande B, Lee JS, Mutasa-Gottgens E, Naughton B, Bacon W, Manning J, et al. Applications of single-cell RNA sequencing in drug discovery and development. Nat Rev Drug Discov. 2023;22(6):496–520.
162. Kashima Y, Sakamoto Y, Kaneko K, Seki M, Suzuki Y, Suzuki A. Single-cell sequencing techniques from individual to multiomics analyses. Exp Mol Med. 2020;52(9):1419–27.
163. Wang Y, Navin NE. Advances and applications of single-cell sequencing technologies. Mol Cell. 2015;58(4):598–609.
164. Li X, Chen S, Ding M, Ding H, Yang K. Decoding the glioma microenvironment: single-cell RNA sequencing reveals the impact of cell-to-cell communication on tumor progression and immunotherapy response. 2024.
165. Ding S, Chen X, Shen K. Single-cell RNA sequencing in breast cancer: understanding tumor heterogeneity and paving roads to individualized therapy. Cancer Commun. 2020;40(8):329–44.
166. Kashima Y, Togashi Y, Fukuoka S, Kamada T, Irie T, Suzuki A, et al. Potentiality of multiple modalities for single-cell analyses to evaluate the tumor microenvironment in clinical specimens. Sci Rep. 2021;11(1):341.
167. Sultana A, Alam MS, Liu X, Sharma R, Singla RK, Gundamaraju R, et al. Single-cell RNA-seq analysis to identify potential biomarkers for diagnosis, and prognosis of non-small cell lung cancer by using comprehensive bioinformatics approaches. Transl Oncol. 2023;27:101571.
168. Narayanan SP, Gopal R, Jenifer SA, Masoodi TA. Single-cell RNA sequencing identifies SLC6A3 as a biomarker and prognostic marker in clear cell renal cell carcinoma. bioRxiv. 2023;2023–08.
169. Cochain C, Vafadarnejad E, Arampatzi P, Pelisek J, Winkels H, Ley K, et al. Single-cell RNA-seq reveals the transcriptional landscape and heterogeneity of aortic macrophages in murine atherosclerosis. Circ Res. 2018;122(12):1661–74.
170. Pu Y, Yan X, Liao H, Wang C, Huang C, Zhang W, Guo C. An advanced bacterial single-cell rna-seq reveals biofilm heterogeneity. eLife. 2024;13:RP97543.
171. Blattman SB, Jiang W, Oikonomou P, Tavazoie S. Prokaryotic single-cell RNA sequencing by in situ combinatorial indexing. Nat Microbiol. 2020;5(10):1192–201.
172. Chen T, Li J, Jia Y, Wang J, Sang R, Zhang Y, et al. Single-cell sequencing in the field of stem cells. Curr Genomics. 2020;21(8):576–84.
173. Conrad S, Azizi H, Skutella T. Single-cell expression profiling and proteomics of primordial germ cells, spermatogonial stem cells, adult germ stem cells, and oocytes. Stem Cells Biol Eng. 2018;1083:77–87.
174. Li L, Dong J, Yan L, Yong J, Liu X, Hu Y, et al. Single-cell RNA-seq analysis maps development of human germline cells and gonadal niche interactions. Cell Stem Cell. 2017;20(6):858–73.

Publisher's Note