Next-Generation Sequencing
Standard Operating Procedures and Applications

Edited by
Prashanth N. Suravajhala and Jeffrey W. Bizzaro

First edition published 2025

ISBN: 978-1-032-39262-2 (hbk)
ISBN: 978-1-032-40635-0 (pbk)
ISBN: 978-1-003-35406-2 (ebk)

Chapter 7

Best Practices in Single-Cell RNA-seq Data Analysis

DOI: 10.1201/9781003354062-7

The funder is Luxembourg National Research Fund (FNR) and EUROSTARS (EUREKA Network)

CRC Press
Taylor & Francis Group
Boca Raton London New York

CRC Press is an imprint of the
Taylor & Francis Group, an **informa** business

7 Best Practices in Single-Cell RNA-seq Data Analysis

Mathias Galati, Xinhui Wang, Muhammad Shoaib, Rajesh Rawal, Irina Balaur, Shaman Narayanasamy, and Venkata Satagopam

INTRODUCTION

Single-cell (SC) research is rapidly ascending to the forefront of biomedicine and biotechnology, due to its ability to elucidate cell-level variation and heterogeneity, particularly in the context of human diseases. Consequently, considerable advances have been made in SC research, spanning from experimental methodologies, technological breakthroughs, data analytics, and outcomes [1]. Single-cell RNA sequencing (scRNA-seq) provides unprecedented depth in exploring cellular heterogeneity [2], unveiling intricate variability within seemingly homogeneous cell populations. Moreover, it facilitates understanding cell trajectories through pseudo-time analysis, which allows the reconstruction of cellular development or differentiation pathways [3]. Furthermore, scRNA-seq can uncover intricate disease mechanisms. Researchers can dissect signalling networks between different cell types, offering novel insights into disease progression and potential therapeutic targets [4]. In oncology, it offers a higher-resolution picture of the intricate and diverse cell populations within tumours, aiding in the identification of unique transcriptional signatures that could be associated with therapy response or disease prognosis [5], a critical asset in the development of personalised medicine strategies.

Historically, the progress of SC studies was limited by the cost, labour, and time-intensive nature of low-throughput SC isolation methods [6]. Today, high-throughput techniques have mitigated those challenges [7]. The success of SC technologies can be largely attributed to advancement in two main techniques: (i) cellular separation and (ii) barcoding [8,9]. Figure 7.1 illustrates an SC protocol that utilises those strategies. Briefly, high-throughput cell isolation methods such as fluorescence-activated cell sorting (FACS) in SC suspensions, plate-based segregation, and robotic droplet microfluidic devices [10] work in tandem with state-of-the-art barcoding to label cells derived from specific samples, subjects, experimental conditions, tissue, or cell types. It is

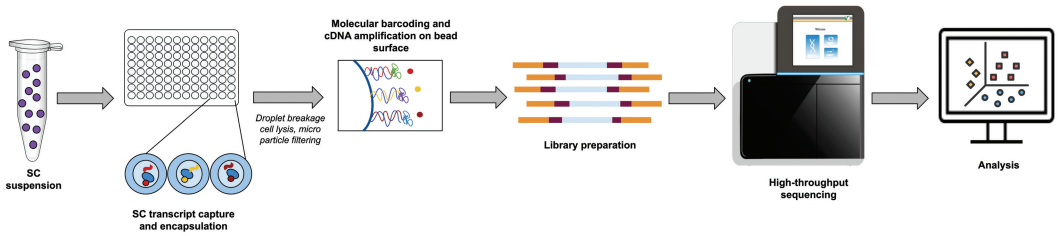

FIGURE 7.1 SC transcriptomics workflow. Tissue samples are preprocessed and SC suspension is collected in vials. Each cell is wrapped with a bead inside a nanoscale droplet (each bead contains a unique molecular identifier [UMI] barcode attached with a transcript). Cell lysis, droplet breakage, and microparticle filtering occurs within each droplet. Transcripts together with the relevant barcodes and adapter bind to the beads followed by amplification of complementary DNA (cDNA). The cDNA then undergoes library preparation procedure for sequencing, which in turn is followed by data analysis. (Graphical elements were obtained from Bioicons https://bioicons.com/ [Accessed 16 October 2024].)

DOI: 10.1201/9781003354062-7

important to note that barcoding can be performed on various biomolecules (e.g., DNA, RNA, proteins, etc.) and staggered or nested to reflect multiple aspects, e.g., separating subjects and cell types [11]. Upon barcoding, the cells can then be lysed to extract the biomolecules of interest (e.g., DNA, RNA, proteins, etc.). Barcoded biomolecules can then be pooled into a single library for an associated high-throughput run, e.g., next-generation sequencing (NGS), thus significantly reducing the overall cost [12]. The process of cellular separation to pooled library preparation can be collectively referred to as multiplexing. To that end, there are various multiplexing strategies [13–16] and platforms [17–19] currently available.

The cell transcriptome, i.e., the collection of all RNA molecules within a cell, exhibits considerable diversity and dynamics relative to the genome and epigenome while offering superior coverage relative to readouts from spectrometry-based proteomics. These hallmark characteristics highlight the versatility of the transcriptome as an effective tool for SC studies. As such, SC transcriptomics enables us to investigate cell-wide gene expression by quantifying mRNA expression at the SC level [20], thereby producing detailed gene expression profiles. These reflect a snapshot of a cell's transcriptomic activity at a particular time point. Such profiles can be applied in various contexts, including but not limited to exploring intracellular dynamics [21], alternative splicing [22], and cell typing [23].

In this chapter, we focus on SC transcriptomics analyses generated using scRNA-seq technologies and the various strategies and tools for the analyses, spanning from large-scale preprocessing of NGS reads (primary analysis) to downstream secondary analysis, such as differential expression. We aim to educate and guide readers who are looking to explore high-throughput SC data for analysis (Figure 7.2).

PRIMARY ANALYSIS

The primary analysis involves the large-scale processing of NGS reads derived from scRNA-seq runs. It covers demultiplexing, quality control (QC), and the alignment and quantification of the data.

Demultiplexing

Demultiplexing [24,25] is a process applied to the raw scRNA-seq data to separate individual cells and distinct samples within a multiplexed library using barcode information (see Figure 7.1 and the 'Introduction'). Briefly, demultiplexing can be carried out based on barcodes (i.e., artificially introduced nucleotides that tag specific samples) [14,15,26–30] or single-nucleotide polymorphisms (SNPs, which are naturally occurring) [31–34]. Popular barcode-based demultiplexing tools were found to be of comparable performance [25], with HTODemux [14] edging out the others for barcode-based demultiplexing. It is important to note that this process can be skipped if a given scRNA-seq dataset is already in a demultiplexed form.

Quality Control

Standard NGS QC of scRNA-seq reads is typically performed using tools such as FastQC [35] to ensure sufficient quality for downstream processes. Additional scRNA-seq-specific QC steps involve, for instance, setting a threshold for UMIs per cell to exclude low-quality cells. UMIs are 'barcodes' that tag individual molecules during library preparation [36]. Tools such as EmptyDrop identify droplets with ambient or 'free-floating' RNA, thus reducing the false discovery rates [26]. Genes from cells that are either dying, dead, or below the detection threshold can be removed to further improve reliability [37]. Last, but not least, the doublets or multiplets, i.e., artefacts of two or more cells encapsulated in a single droplet and sequenced as if they were a single cell (see Figure 7.1) can be removed using tools such as Scrublet [38] and DoubletFinder [39]. Finally, Hong

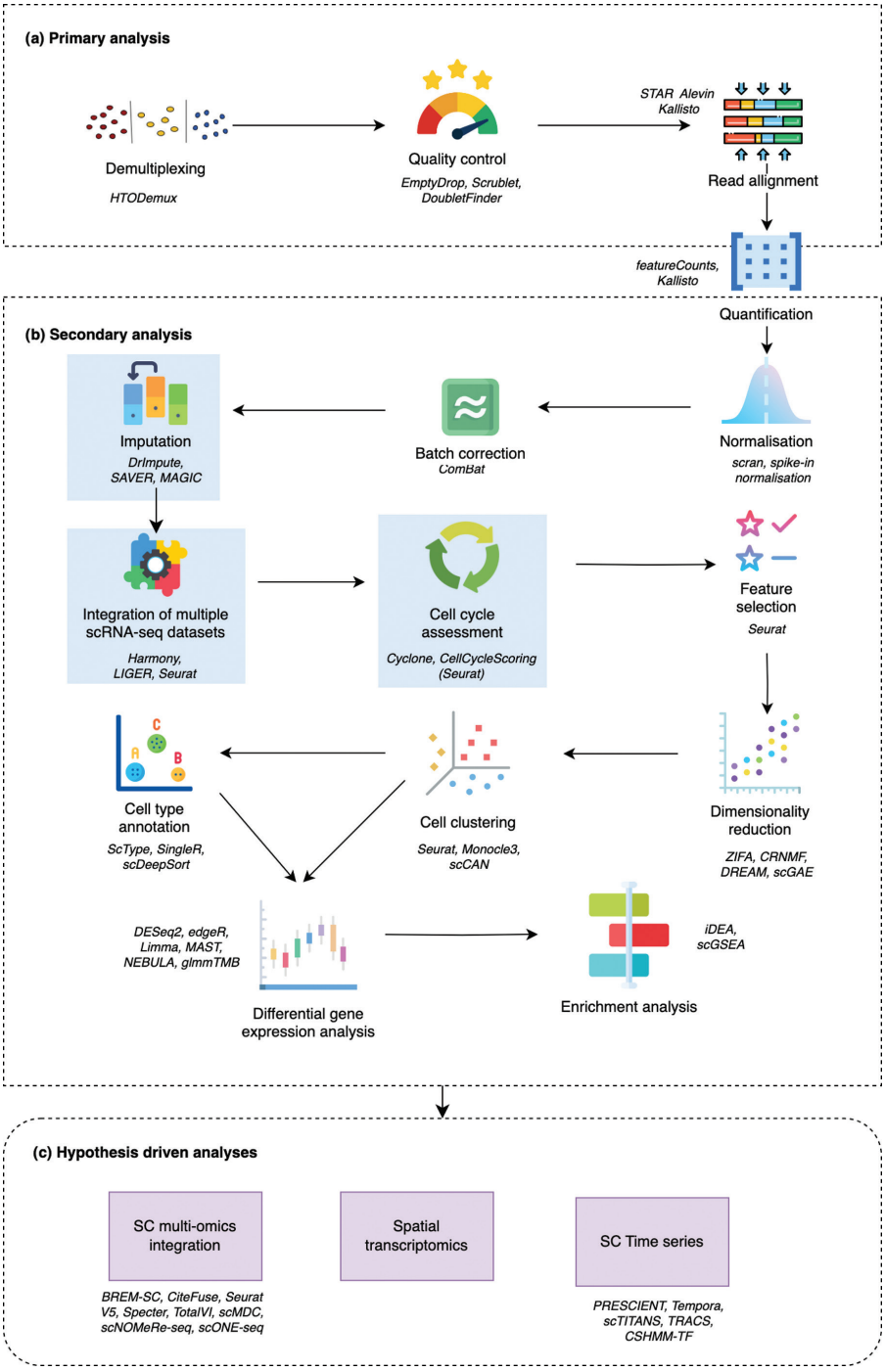

FIGURE 7.2 Overview of the scRNA-seq data processing and analysis workflow. (a) Primary analysis steps include processing of the scRNA-seq reads such as demultiplexing, quality control, and read alignment. (b) Secondary analysis steps that are used to explore and interpret scRNA-seq data, spanning from data normalisation up to functional interpretation. Blue boxes highlight optional steps. (c) Downstream hypothesis-driven analysis could include more advanced multi-omics integration, temporal or spatially resolved analyses. Examples of tools for each step are provided as italic text. (Illustrative icons were obtained from Flaticons https://www.flaticon.com/ and Bioicons https://bioicons.com/ [Accessed 16 October 2024].)

et al. provide an updated and extensive review of various QC tools for scRNA-seq to guide researchers working on such data [40].

Read Alignment and Quantification

NGS reads that pass the QC process then proceed to the read alignment or mapping step. This involves aligning the reads to a reference genome, e.g., a human genome reference. Briefly, alignment programmes require the NGS reads and an indexed reference genome. This process is typically resource-intensive and highly dependent on the size of the reference genome and the depth of sequencing. STAR [41] is a common tool used for transcriptomic analysis that relies on a classical alignment algorithm [42]. However, the resource intensive (i.e., CPU, RAM, and storage) nature of classical read alignment (e.g., in tools like STAR) has led to the prominence of pseudo-alignment tools such as Alevin [43] and Kallisto [44] in scRNA-seq data.

The quantification of reads from scRNA-seq data is a critical step in downstream analysis, and it can be approached in two main ways. First, reads aligned with tools such as STAR result in a sequence/binary alignment map in SAM/BAM file format, which contains extensive information such as read mapping position, quality of read, and quality of the alignment. The SAM/BAM formats are, in turn, used with annotation information (e.g., in GTF format) within tools such as featureCounts to quantify gene expression, which can be further normalised prior to secondary analysis and processing (e.g., differential expression). Alternatively, pseudo-alignment tools such as Kallisto can directly 'align' reads to a set of transcripts to output normalised estimated count tables (e.g., transcripts per million [TPM]) that can be used in secondary analysis (e.g., differential expression) with minimal additional processing. Pseudo-alignment tools may be preferred for scRNA-seq data processing as they are less resource intensive, require less intermediate processing, and are more in line with the goals of scRNA-seq primary analysis, which is gene quantification.

SECONDARY ANALYSIS

Normalisation, Batch Effect Correction, and Imputation

ScRNA-seq requires normalisation procedures to account for technical variabilities, specifically the fact that different cells or sequencing experiments may yield different numbers of reads. After initial quantification, which produces raw count matrices, tools such as scran [45] estimate size factors that are subsequently used to normalise gene expression values. Spike-in normalisation [46] is another strategy that can be used when spike-in controls are added to the samples, providing a reference for correcting technical noise. Moreover, different normalisation strategies are used for varying scales of data and different downstream analyses. For instance, SCnorm is applied to low-throughput (low number of samples), high-depth sequencing data, as it is robust in dealing with high variability [47]. Conversely, for high-throughput (large number of samples), low-depth data, where the emphasis is on capturing a wide range of cellular diversity, the sctransform method is often utilised [48]. Finally, log-normalisation and z-score transformation are typically applied across all datasets to standardise gene expression values, facilitating comparative analysis across cells and conditions.

Diverse sources of technical variability may introduce so-called batch effects in scRNA-seq data, potentially obscuring true biological signals of interest. Batch effects may stem from many sources, for instance, different or distinct extractions, experiments, sequencing runs, and/or reagent lots. To tackle these challenges, numerous batch effect correction tools have been developed [49]. Among them, ComBat [50] is often favoured due to its robustness and simplicity. However, correction methodologies such as this were originally developed for bulk transcriptomics, and thus assume equal cell composition across batches, which is often inaccurate in the context of scRNA-seq data. Hence, batch correction relying on mutual nearest neighbours (MNNs) may lead to more accurate and precise results [51]. Alternatively, the task of integrating multiple scRNA-seq

datasets is addressing similar challenges, thus the methodologies and computational strategies often overlap. For more details, please refer to the subsequent section on 'Integration of Multiple scRNA-seq Datasets'.

Dropout events, wherein a gene appears to be unexpressed due to technical limitations, pose significant challenges in scRNA-seq data interpretation. As a result, several imputation methods have been developed to handle this issue, typically replacing dropout events (usually represented as zeros) with estimated values to reduce noise and facilitate downstream data analysis. DrImpute [52] and SAVER [53], for example, employ statistical models to predict missing values based on observed gene expression patterns. Alternatively, methods such as MAGIC [54] use diffusion-based algorithms to smooth gene expression data across similar cells, helping to mitigate noise. While these methods can reduce data noise and improve interpretability, it is important to note potential limitations, namely, the over-reliance on imputation could inadvertently introduce artificial structures or patterns, leading to overfitting. Therefore, these techniques should be employed judiciously, always considering the context of the biological question.

Integration of Multiple scRNA-Seq Datasets

The integration of datasets (also referred as P-integration) in scRNA-seq is a critical task in the analysis of SC genomics data, as it allows for the comparison and consolidation of data from multiple batches, experimental conditions, or technologies (such as scATAC-seq). Various computational tools have been developed to facilitate such integration, including Harmony [55], LIGER [56], and Seurat [57]. Harmony, an algorithm and R package, creates a harmonised embedding of cells where they cluster by cell type rather than by the batch or dataset of origin, and is known for its speed and efficiency. Linked Inference of Genomic Experimental Relationships (LIGER), on the other hand, identifies shared and dataset-specific (or condition-specific) cell types using integrative non-negative matrix factorisation, thereby uncovering shared and unique features across datasets. Finally, Seurat, an R package, offers various methods for data integration, including canonical correlation analysis (CCA)-based integration and an integration method based on MNNs. All three methods have proven effective for batch correction in scRNA-seq data and have been recommended based on benchmarking studies, such as the one by Tran et al. [49], each with their unique strengths and use cases.

Cell Cycle Assignment

Cell cycle stages considerably influence gene expression profiles. Specifically, different phases of the cell cycle will result in distinct transcriptional signatures, thus generating an inherent variability in the expression of genes. This variability, if unaccounted for, can introduce confounding factors into downstream analysis, especially when the objective is to compare cell types or states, which may also be represented by disparate proportions of cells in each phase of the cell cycle [47].

To that end, scRNA-seq data analysis should incorporate the assignment of cell cycle stage information. For instance, during an SC differential gene expression analysis, a population of cells predominantly in the S or G2/M phases (phases marked by DNA replication and thus high expression of replication-related genes), compared to a population largely in other stages, may lead to misleading interpretations. In this event, replication-related genes may seem differentially expressed due to the disproportion in cell cycle stages rather than any true inherent biological differences between the cell populations.

By implementing cell cycle assignment, such as through tools like Cyclone [58] and CellCycleScoring [59] from the Seurat package, cells can be categorised into their respective cell cycle phases. This step provides an opportunity to account for this intrinsic variability. Post-assignment, general linear models (GLM) can be employed to 'regress out' cell cycle associated differences, thereby neutralising their impact on downstream analysis [60]. Some studies may require

explicit modelling of cell cycle effects or conduct separate intracell cycle analyses. This process not only minimises the potential confounding influence of these effects but also permits focus on biologically relevant variations over those from different cell cycle stages, enabling accurate and biologically meaningful interpretations.

Feature Selection

The scRNA-seq dataset often contains a vast number of genes, but not all of these genes or cell types may be adequately represented during a given experiment. This underlines the importance of feature selection, a step in the analysis process that focuses on retaining only strong biological signals compared to technical noise. Genes (also referred to as 'features' in this context) that are either too noisy or present in a small number of samples are often eliminated based on parameters like the mean and variance of gene expression or a high frequency of zero values [48,57].

When dealing with multiple datasets, the feature selection process is first performed on an individual dataset. Following this, features are prioritised across multiple experiments based on the number of datasets where they were independently identified as highly variable. This integrative approach allows for a more robust identification of relevant features, ultimately enhancing the potential for meaningful biological insights.

It is important to note that the feature selection process in scRNA-seq is crucial for reducing data complexity and improving computational efficiency. However, it also requires careful consideration, as it can influence downstream analyses and interpretation of the results. Methods available range from univariate filters (such as t-statistics), requiring predefined cell type labels, to less restrictive approaches that filter highly variable genes (HVGs), implemented in frameworks like Seurat. Several other approaches for analysing scRNA-seq data are emerging. These include multivariate methods such as gene pair filtering and COMET (a deviance statistic-based gene filtering technique). Additionally, researchers are developing wrapper and embedded methods that build upon classic analytical techniques [61].

Dimensionality Reduction

The processed scRNA-seq data is typically organised into an expression matrix, where each column represents a single cell, and each row represents a specific gene or transcript. Each element in this matrix shows the expression level of a particular gene in a specific cell. The size of this matrix can be substantial, with the number of cells ranging from thousands to millions, while the number of genes depends on the reference transcriptome (e.g., about 30,000 for humans). This results in a large, high-dimensional dataset, which presents significant challenges for effective visualisation, interpretation, and further analysis. To address these challenges, researchers use dimensionality reduction (DR) techniques. These DR methods allow the data to be represented and visualised in two dimensions, typically as a graph or plot, making it easier to identify patterns and relationships within the complex scRNA-seq data.

DR is an integral step in scRNA-seq analysis, as it facilitates the identification of clusters in the data that are presumed to represent similar cell types. Traditional DR methods, such as principal component analysis (PCA) [62], t-Distributed Stochastic Neighbour Embedding (t-SNE) [63], and especially uniform manifold approximation and projection (UMAP) [64], have been commonly employed in scRNA-seq data analysis, as they generally work well. However, in some cases, those conventional methods do not account for characteristics specific to scRNA-seq data, such as the prevalence of zero values (dropouts), which are false-zero gene expression values in some cells due to technical limitations.

To that end, methods such as Zero Inflated Factor Analysis (ZIFA) [65], Constrained Robust Non-negative Matrix Factorisation (CRNMF) [66], and an improved Deep Variational Autoencoder model (DREAM) [67] are being developed to mitigate SC-specific DR issues. These methods claim

to be capable of preserving long-distance relationships in a latent space such as single-cell Graph Autoencoder (scGAE) [68].

To conclude, the choice of suitable methods depends on the specifics of the data and the research question at hand, while future rapid advancements in SC technologies will continue introducing novel techniques for DR.

Cell Clustering

An essential part of scRNA-seq analysis is cell clustering, a process that groups cells based on their features, facilitating the discovery of new cell (sub)types and/or marker genes. However, it can be challenging due to the noisy and high-dimensional nature of the data. Seurat clustering functions [57] and Monocle [38] are the most widely used methods among several developed approaches to address this problem, each offering distinct advantages, drawbacks, and areas of application. A novel approach is scCAN [69], which utilises non-negative kernel autoencoders, stacked variational autoencoders, and graph-based techniques. Authors have demonstrated that scCAN outperforms existing methods in accuracy and scalability.

Nevertheless, while multiple methods exist for clustering scRNA-seq data, choosing the right approach is dependent on the nature and specific needs of the analysis. The advancement of computational methods continues to enhance the accuracy and efficiency of scRNA-seq clustering. As a result, it is advisable to consult recent benchmark studies, such as the article by Yu et al. [70], to make an informed decision.

Cell Type Annotation

Cell type annotation involves associating clusters (generated from the cell clustering step described in the previous section) with different gene expression profiles to distinct cell types. Given the time-consuming and subjective nature of manual annotation, the development of standardised and automated tools for cell type identification has been a significant advancement in the field [71].

Automated cell type annotation methodologies can be classified into three primary categories: (i) marker gene database-based annotation such as ScType [23], (ii) correlation-based methods such as SingleR [72], and (iii) supervised classification such as scDeepSort [73] or scClassify [74]. The first approach utilises specific marker genes to identify cell types, while correlation-based methods compare gene expression profiles of cells with reference datasets to infer cell types. Supervised classification employs machine learning algorithms trained on labelled datasets to classify cells into known cell types [75].

Despite the availability of these automation methods, manual annotation remains necessary in cases of conflicting or absent cell labels, or for novel and critically discussed cell type subclasses. This underscores the importance of refining cell type annotation methods. Useful guidelines are discussed in a review by Clarke et al. [76].

Differential Gene Expression Analysis

Differentially expressed genes (DEGs) refer to genes that exhibit significant variations in their expression levels across two or more biological states or conditions. The identification of these DEGs is fundamental to the analysis of gene expression data, as they are often implicated in the biological processes or the diseases being studied.

In the context of scRNA-seq, the analysis of DEGs is even more important, as it enables researchers to identify differences in gene expression at the level of individual cells. Such granular insights can help in understanding the diversity within cell populations and the specific roles of different cell types.

Classical methods for differential gene expression in scRNA-seq include DESeq2 [77] and edgeR [78], which are based on negative binomial models. In contrast, limma [79] utilises a linear model

for data processing. Originally designed for bulk RNA-seq data, these techniques have also found applicability in scRNA-seq analyses despite not being specifically tailored for the task.

The Model-based Analysis of Single-cell Transcriptomics (MAST) approach is interesting because it offers a two-part generalised linear model that simultaneously models the rate of expression over the background of various transcripts and the positive expression mean. By leveraging a hurdle model and by introducing the concept of cellular detection rate, they provide a mechanism for advanced, multiparametric modelling of SC transcriptomics data that can account for both technical and biological factors affecting gene expression and thus potentially enabling more accurate interpretation of experimental results [80].

According to a recent study from Gagnon et al. [81], the NEBULA [82] and glmmTMB [83] methods were found to outperform the commonly used methods in the literature. Nevertheless, selecting an appropriate method depends on the dataset characteristics and the research questions available. Researchers are advised to consult original papers and software documentation to understand the optimal use of these methodologies.

Enrichment Analysis (EA)

EA in scRNA-seq is a computational method used to identify groups or classes of significantly overrepresented or underrepresented features in the data. The goal is to determine if the genes from the previous DEG analysis, whether comparing conditions (e.g., healthy vs. diseased) or cell types, are enriched in certain biological pathways, molecular functions, cellular components, or other defined gene sets (such as MSigDB [84], Gene Ontology [85], etc.), thus providing a deeper understanding of the underlying biological processes being activated or repressed in the given dataset.

To perform an EA, the list of genes is compared to predefined gene sets or databases to see if there is an overlap greater than what is expected by chance, using statistical methods such as Fisher's exact test. There are common tools and software available for EA, such as g:Profiler [86], DAVID [87], and GSEA [88,89]. Some of the most recent and tailored tools for scRNA-seq data are iDEA [90] and scGSEA [91]. iDEA is a statistical method that uses summary statistics from gene expression data, including measures of gene expression changes (such as fold change or effect size estimates) and their associated standard errors. What distinguishes iDEA is its ability to combine differential expression analysis with gene set enrichment analysis within a single, unified statistical framework. scGSEA combines latent data representations and gene set enrichment scores to detect coordinated gene activity at single-cell resolution.

Overall, EA is a form of hypothesis-generating procedure, and while it provides useful biological insights, it does not prove causal relationships, and so the results need to be interpreted with caution and ideally validated with further experiments.

SC Analytic Frameworks

scRNA-seq data analysis is multifaceted, with various factors influencing the success of a given study. We highly recommend using established analytical methodologies (such as those highlighted in this chapter) as a sturdy foundation for analysis and later moving to more tailored approaches that may be better suited to the given experimental design and/or data. To that end, there are several specialised frameworks to facilitate scRNA-seq data analyses that comprise a suite of preselected tools conveniently packaged together and maintained by experts in the field [92]. These frameworks offer a smoother onboarding experience for those who want to delve into scRNA-seq data analysis. They can also be divided by focus: whether it is on primary or secondary analysis.

Primary analysis frameworks typically consist of large-scale processing tools (see the section 'Primary Analysis') that are strung together using workflow automation tools such as Snakemake [93] or Nextflow [94,95]. These workflow languages enhance reproducibility, automation, and scalability, potentially saving considerable time, while being kept up to date by a community of experts.

Secondary analysis frameworks include various secondary analysis steps (see the section 'Secondary Analysis') and can be further divided by platform, such as whether they are R or Python based. Seurat, a package built on R [96], is arguably the most widely used implementation for scRNA-seq secondary analysis due to its flexibility, high-quality implementation, and frequent updates. And for those familiar with R, platforms such as Bioconductor, Seurat, and tidy transcriptomics are highly recommended for their excellent documentation and very active user communities that can provide support and answer questions. Similarly, tidyseurat [97] (within the tidytranscriptomics suite [98]) offers a collection of powerful and cohesive packages for data manipulation, exploration, and visualisation under the R tidyverse framework [98]. This unified framework enables consistent syntax and data structures for efficient, intuitive, and reproducible scRNA-seq analysis. Furthermore, packages such as clustree [99] should be considered to facilitate visualisations and decision-making, while SCpubr [100] could aid in generating publication-quality figures. Similarly, SCEDAR [101] and scanpy [102] offer equivalent secondary analysis capabilities in Python, with the latter further offering an R implementation.

Researchers should consult available resources when starting their scRNA-seq analysis journey; high-quality guides are available for Python [103,104] and R Bioconductor [105]. More importantly, the landscape of scRNA-seq data analysis is highly dynamic, therefore researchers should always refer to the most recent analysis procedures, guidelines, and benchmarks.

BEST PRACTICES FOR SC DATA ANALYSIS

It is necessary for researchers to approach scRNA-seq data analysis with a dynamic mindset, considering the rapid advances in the field. Depending on the time of reading, one may encounter certain tools highlighted within this chapter that have become either outdated or significantly improved. Thus, we aim to address crucial aspects of SC transcriptomics data analysis from two angles: (i) the selection of tools and (ii) the reproducibility of analysis pipelines.

When selecting analytical tools, two factors should be considered: the goals of the study and the quality and implementation of the tools. Always consider the objectives of your analysis and the characteristics of the generated SC data when selecting tools. Given the vast range of available tools, it is generally prudent to select those with a proven track record of generating high-quality output, demonstrating stable implementation, providing regular support and updates, and ensuring interoperability with other tools through standard output formats. Another factor of paramount importance is a collaborative multidisciplinary mindset when designing scRNA-seq experiments, so that you include input from biologists, statisticians, and bioinformaticians to ensure a biologically, statistically, and methodologically robust downstream analysis. Moreover, remember that the specifics of your project – such as the biological question, sample type, and sequencing platform – will often determine the most suitable tools and analysis strategies.

The reproducibility of analysis pipelines applied within an scRNA-seq study should adhere to general open-science best practices [106,107]. Researchers should have a long-term view on being able to replicate a given study even years after publication. Therefore, code for a given analysis protocol should be deposited into repositories such as GitHub, while raw NGS data (e.g., scRNA-seq data) should be uploaded into public archives like the National Center for Biotechnology Information's (NCBI) Sequence Read Archive (SRA) or the European Bioinformatics Institute's (EBI) European Nucleotide Archive (ENA) with detailed accompanying metadata, while gene expression data should be provided on NCBI's Gene Expression Omnibus (GEO) platform. If applicable, non-standard data formats could be uploaded on Zenodo [108]. Finally, containerisation (e.g., using Docker or Singularity) of specific software versions could further facilitate reproducibility.

Recently, the sharing of raw and/or processed NGS data from human subjects has raised regulatory and ethical concerns [109], hence researchers must adhere to policies such as the General Data Protection Regulation (GDPR) in the European Union [110] or the Health Insurance Portability and Accountability Act (HIPAA) in the United States, among other equivalent regional or international

data protection policies. Overall, striking a balance between protecting the subjects of data collection and maintaining research transparency is critical. The Findability, Accessibility, Interoperability, and Reuse (FAIR) principles and resources serve as an excellent guide for data discovery, use, and sharing [111–115].

In general, the availability of code, data, and software alongside detailed methodological descriptions or documentation (e.g., in an article) should provide sufficient material for reproducing a given study, which in turn could lead to downstream integration into other independent studies (e.g., meta-analysis; see the section 'Integration of Multiple scRNA-seq Datasets'), scaling up existing studies, and knowledge transfer.

ADVANCED SCRNA-SEQ ANALYSES

Following the scRNA-seq data analysis approach outlined in this chapter, it is imperative to underscore the numerous avenues for further analyses, which include the following:

Cell-to-cell communication analysis can unravel the intricate interactions that occur within and between cellular communities [116]. LIANA is a computational tool worth mentioning here, as it integrates multiple ligand-receptor databases and offers various statistical methods to infer and analyse cell-cell interactions from scRNA-seq data [4].

Gene regulatory network inference is a process that models the complex regulation of gene expression in individual cells. Tools such as SCENIC perform this inference, offering insights into the fundamental mechanisms underpinning cellular diversity and the dynamics of biological processes [117].

Genotyping analysis can supplement scRNA-seq data by relating genetic variations to cellular-level expression, providing insights into their phenotypic consequences. As such, various tools can be utilised to perform copy number variation (CNV) [118], single nucleotide variation (SNV) [119], and expression quantitative trait locus (eQTL) [120] analysis.

RNA velocity is an approach that employs scRNA-seq to approximate the derivative of gene expression over time, thereby elucidating the dynamism of gene expression within individual cells. Through the quantification of RNA expression 'velocity', researchers can predict cellular fate and comprehend developmental trajectories with tools such as scVelo [121].

INTEGRATED APPROACHES IN SC ANALYSIS: TIME-SERIES, SPATIAL, AND MULTI-OMICS

Unveiling biological complexity requires the intersection of the various dimensions of SC analysis, namely time-series, spatial, and multi-omics approaches.

The time-series aspect of SC analysis unravels dynamic changes in cell states and gene expression, elucidating the evolution of cellular processes [122]. Current capabilities extend to inferring cell trajectory [123], predicting cell fates over varying time points and populations [124], assigning transcription factors to activation points [125], and mitigating cell asynchrony [126].

Spatial transcriptomics can be used to complement high-resolution scRNA-seq information to mitigate the inability of SC methods to retain spatial information [127,128]. This union reveals cell-specific changes within their native tissue context, aiding in the exploration of disease biomarkers [127]. Challenges such as data availability and computational complexity are hurdles on the path towards full integration [128,129].

Lastly, SC multi-omics technologies offer the opportunity to analyse multiple types of biomolecules from individual cells, providing an extensive understanding of cellular processes and heterogeneity [130–132]. As such, scRNA-seq derived gene expression can be integrated with genomic, DNA methylation, chromatin accessibility, and/or protein abundance information using a multitude of cutting-edge analytical methods and tools [130,133–137], despite existing technical and computational challenges [138].

Together, these integrated approaches in SC analysis offer powerful tools for comprehensive insights into cellular dynamics, spatial organisation, and multilayered molecular signatures.

THE EXPANDING RESOURCES AND APPLICATIONS OF SC TECHNOLOGIES

scRNA-seq has been applied in various research areas and clinical applications as it enables the high-resolution analysis of individual cells, providing deep details about the cellular heterogeneity and gene expression of complex biological systems and disease mechanisms [139,140].

Precision medicine is one of the most promising applications of scRNA-seq. By elucidating the intricacies of cellular dynamics, it helps identify therapeutic targets and develop innovative disease diagnoses and treatments [140]. In addition, scRNA-seq has enabled the culmination of major resources, including the Human Cell Atlas (HCA) [141], which aims to develop comprehensive reference maps for all human cells, detailing their features, location, and biological functions [129,142]; and the Brain Research through Advancing Innovative Neurotechnologies (BRAIN) Initiative's Cell Census Network (BICCN) [143], which aims for the equivalent with a focus on brain cells. These initiatives have catalysed the development of robust SC analytical tools [144], further enriching our understanding of cellular complexity and heterogeneity across different species and different scales.

From a clinical perspective, scRNA-seq can improve diagnostics, therapeutic choices, and disease monitoring in various fields of medicine. The potential impact of these technologies on precision medicine could see them becoming a key component of the diagnostic process in the future [145].

The application of scRNA-seq in cancer research has been particularly noteworthy with its ability to study individual cells in a tumour, providing insights into cancer biology, with the potential to enhance diagnosis, prognosis, and therapeutics specific to the tumour heterogeneity [146,147].

Moreover, scRNA-seq holds significant promise for drug discovery and development by enabling detailed descriptions of cellular heterogeneity and the related disease mechanisms, and it can help with the identification of new therapeutic targets and understanding cellular responses to treatments [139].

Finally, scRNA-seq facilitates the study of rare cell (sub)types that might have a significant impact on health or disease. These cells are often missed in bulk sequencing techniques but can be effectively studied with scRNA-seq [148]. Also, the integration of scRNA-seq with spatial transcriptomics provides a holistic understanding of the cellular environment, maintaining crucial information about the spatial organisation of the tissue.

In summary, scRNA-seq has a wide range of applications and holds immense promise for advancing our understanding of complex biological systems and disease mechanisms, potentially transforming various fields of research and clinical practice.

CONFLICT OF INTEREST

Authors declare no conflict of interest.

AUTHOR CONTRIBUTIONS

VS conceived and supervised the effort. MG, XW, MS, RR, and IB reviewed the literature. MG and SN compiled the first draft. MS, RR, and IB prepared the figures. All authors wrote and reviewed the final manuscript.

ACKNOWLEDGEMENTS

The authors acknowledge the funding agencies and grants: Luxembourg National Research Fund (FNR) PRIDE grant i2TRON to MG; Eurostars E! 113726 grant to SN; FNR/ERA PerMed (INTER/ERAPerMed/19/13589768) grant to IB; Innovative Medicines Initiative 2 Joint Undertaking, ImmUniverse IMI2-RI-853995 to XW; FNR NCER-PD grant FNR/NCER13/BM/11264123 to RR, and EU Horizon 2020 grant no. 101016072 to MS.

REFERENCES

1. Stuart T & Satija R. Integrative single-cell analysis. *Nat Rev Genet*. 2019;**20**:257–72.
2. Choi YH & Kim JK. Dissecting cellular heterogeneity using single-cell RNA sequencing. *Mol Cells*. 2019;**42**:189–99.
3. Wang L et al. Current progress and potential opportunities to infer single-cell developmental trajectory and cell fate. *Curr Opin Syst Biol*. 2021;**26**:1–11.
4. Dimitrov D et al. Comparison of methods and resources for cell-cell communication inference from single-cell RNA-seq data. *Nat Commun*. 2022;**13**:3224.
5. Noureen N, Ye Z, Chen Y, Wang X & Zheng S. Signature-scoring methods developed for bulk samples are not adequate for cancer single-cell RNA sequencing data. *eLife*. 2022;**11**:e71994.
6. Hu P, Zhang W, Xin H & Deng G. Single cell isolation and analysis. *Front Cell Dev Biol*. 2016;**4**:4–116. doi: 10.3389/fcell.2016.00116
7. Lähnemann D et al. Eleven grand challenges in single-cell data science. *Genome Biol*. 2020;**21**:31.
8. Cao J et al. The single-cell transcriptional landscape of mammalian organogenesis. *Nature* 2019;**566**: 496–502.
9. Zilionis R et al. Single-cell barcoding and sequencing using droplet microfluidics. *Nat Protoc*. 2017; **12**:44–73.
10. Prakadan SM, Shalek AK & Weitz DA. Scaling by shrinking: Empowering single-cell 'omics' with microfluidic devices. *Nat Rev Genet*. 2017;**18**:345–61.
11. Rosenberg AB et al. Single-cell profiling of the developing mouse brain and spinal cord with split-pool barcoding. *Science*. 2018;**360**:176–82.
12. Svensson V, Vento-Tormo R & Teichmann SA. Exponential scaling of single-cell RNA-seq in the past decade. *Nat Protoc*. 2018;**13**:599–604.
13. Gehring J, Hwee Park J, Chen S, Thomson M & Pachter L. Highly multiplexed single-cell RNA-seq by DNA oligonucleotide tagging of cellular proteins. *Nat Biotechnol*. 2020;**38**:35–38.
14. Stoeckius M et al. Cell hashing with barcoded antibodies enables multiplexing and doublet detection for single cell genomics. *Genome Biol*. 2018;**19**:224.
15. McGinnis CS et al. MULTI-seq: Sample multiplexing for single-cell RNA sequencing using lipid-tagged indices. *Nat Methods*. 2019;**16**:619–26.
16. Single Cell Gene Expression Flex. 10x Genomics https://www.10xgenomics.com/products/single-cell-gene-expression-flex.
17. Jariani A et al. A new protocol for single-cell RNA-seq reveals stochastic gene expression during lag phase in budding yeast. *eLife*. 2020;**9**:e55320.
18. 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 (ed. Christian, S. L.) 261–278 (Springer US, 2022). doi:10.1007/978-1-0716-2376-3_19
19. Wen L et al. Single-cell technologies: From research to application. *The Innovation*. 2022;**3**:100342.
20. 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**:75.
21. Schwabe D, Formichetti S, Junker JP, Falcke M & Rajewsky N. The transcriptome dynamics of single cells during the cell cycle. *Mol Syst Biol*. 2020;**16**:e9946.
22. Arzalluz-Luque Á & Conesa A. Single-cell RNAseq for the study of isoforms—how is that possible? *Genome Biol*. 2018;**19**:110.
23. Ianevski A, Giri AK & Aittokallio T. Fully-automated and ultra-fast cell-type identification using specific marker combinations from single-cell transcriptomic data. *Nat Commun*. 2022;**13**:1246.
24. Zhang Y et al. Sample-multiplexing approaches for single-cell sequencing. *Cell Mol Life Sci*. 2022; **79**:466.
25. Howitt G et al. Benchmarking single-cell hashtag oligo demultiplexing methods. 2023. Preprint at doi: 10.1101/2022.12.20.521313
26. Lun ATL et al. EmptyDrops: Distinguishing cells from empty droplets in droplet-based single-cell RNA sequencing data. *Genome Biol*. 2019;**20**:63.
27. Xin H et al. GMM-Demux: Sample demultiplexing, multiplet detection, experiment planning, and novel cell-type verification in single cell sequencing. *Genome Biol*. 2020;**21**:188.
28. Tuddenham JF et al. A cross-disease human microglial framework identifies disease-enriched subsets and tool compounds for microglial polarization. 2022. Preprint at doi: 10.1101/2022.06.04.494709
29. Boggy GJ et al. BFF and cellhashR: Analysis tools for accurate demultiplexing of cell hashing data. *Bioinformatics* 2022;**38**:2791–801.

30. 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 Systems*. 2020;**11**(1): 95–101.e5. doi: 10.1016/j.cels.2020.05.010
31. Kang HM et al. Multiplexed droplet single-cell RNA-sequencing using natural genetic variation. *Nat Biotechnol*. 2018;**36**:89–94.
32. Neavin D et al. Demuxafy: Improvement in droplet assignment by integrating multiple single-cell demultiplexing and doublet detection methods. *Genome Biology*. doi: 10.1186/s13059-024-03224-8
33. Heaton H et al. Souporcell: Robust clustering of single-cell RNA-seq data by genotype without reference genotypes. *Nat Methods*. 2020;**17**:615–20.
34. Xu J et al. Genotype-free demultiplexing of pooled single-cell RNA-seq. *Genome Biol*. 2019;**20**:290.
35. Babraham bioinformatics - FastQC A quality control tool for high throughput sequence data. https://www.bioinformatics.babraham.ac.uk/projects/fastqc/
36. Chen W et al. UMI-count modeling and differential expression analysis for single-cell RNA sequencing. *Genome Biol*. 2018;**19**:70.
37. Ilicic T et al. Classification of low quality cells from single-cell RNA-seq data. *Genome Biol*. 2016;**17**:29.
38. Wolock SL, Lopez R & Klein AM. Scrublet: Computational identification of cell doublets in single-cell transcriptomic data. *Cell Syst*. 2019;**8**:281–91.
39. McGinnis CS, Murrow LM & Gartner ZJ. DoubletFinder: Doublet detection in single-cell RNA sequencing data using artificial nearest neighbors. *Cell Syst*. 2019;**8**:329–37.
40. Hong R et al. Comprehensive generation, visualization, and reporting of quality control metrics for single-cell RNA sequencing data. *Nat Commun*. 2022;**13**:1688.
41. Dobin A & Gingeras TR. Mapping RNA-seq reads with STAR. *Curr Protoc Bioinforma*. 2015;**51**: 11.14.1–11.14.19.
42. Malhotra A, Das S & Rai SN. Analysis of single-cell RNA-sequencing data: A step-by-step guide. *BioMedInformatics*. 2022;**2**:43–61.
43. Srivastava A, Malik L, Smith T, Sudbery I & Patro R. Alevin efficiently estimates accurate gene abundances from dscRNA-seq data. *Genome Biol*. 2019;**20**:65.
44. Melsted P et al. Modular, efficient and constant-memory single-cell RNA-seq preprocessing. *Nat Biotechnol*. 2021;**39**:813–18.
45. Lun ATL, McCarthy DJ & Marioni JC. A step-by-step workflow for low-level analysis of single-cell RNA-seq data with Bioconductor. 2016. Preprint at doi: 10.12688/f1000research.9501.2
46. DeBerardine M. BRGenomics for analyzing high-resolution genomics data in R. *Bioinformatics*. 2023; **39**(6):btad331. doi: 10.1093/bioinformatics/btad331
47. Andrews TS, Kiselev VY, McCarthy D & Hemberg M. Tutorial: Guidelines for the computational analysis of single-cell RNA sequencing data. *Nat Protoc*. 2021;**16**:1–9.
48. Hafemeister C & Satija R. Normalization and variance stabilization of single-cell RNA-seq data using regularized negative binomial regression. *Genome Biol*. 2019;**20**:296.
49. Tran HTN et al. A benchmark of batch-effect correction methods for single-cell RNA sequencing data. *Genome Biol*. 2020;**21**:12.
50. Zhang Y, Parmigiani G & Johnson WE. ComBat-seq: Batch effect adjustment for RNA-seq count data. *NAR Genom Bioinform*. 2020;**2**(3):lqaa078. doi: 10.1093/nargab/lqaa078
51. Haghverdi L, Lun ATL, Morgan MD & Marioni JC. Batch effects in single-cell RNA-sequencing data are corrected by matching mutual nearest neighbors. *Nat Biotechnol*. 2018;**36**:421–27.
52. 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**:220.
53. Huang M et al. SAVER: Gene expression recovery for single-cell RNA sequencing. *Nat Methods*. 2018;**15**:539–42.
54. van Dijk D et al. Recovering gene interactions from single-cell data using data diffusion. *Cell*. 2018; **174**:716–29.e27.
55. Korsunsky I et al. Fast, sensitive and accurate integration of single-cell data with harmony. *Nat Methods*. 2019;**16**:1289–96.
56. Welch JD et al. Single-cell multi-omic integration compares and contrasts features of brain cell identity. *Cell*. 2019;**177**:1873–87.e17.
57. Stuart T et al. Comprehensive integration of single-cell data. *Cell*. 2019;**177**:1888–902.e21.
58. Scialdone A et al. Computational assignment of cell-cycle stage from single-cell transcriptome data. *Methods*. 2015;**85**:54–61.
59. 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**:411–20.

60. Piper M, Mistry M, Liu J, Gammerdinger W & Khetani R. hbctraining/scRNA-seq_online: scRNA-seq lessons from HCBC (first release). Zenodo. 2022. doi:10.5281/zenodo.5826256
61. Yang P, Huang H & Liu C. Feature selection revisited in the single-cell era. *Genome Biol*. 2021;**22**:321.
62. Pearson KL III. On lines and planes of closest fit to systems of points in space Lond. *Edinb Dublin Philos Mag J Sci*. 1901;**2**:559–72.
63. vanMaaten L & Hinton G. Visualizing data using t-SNE. *J Mach Learn Res*. 2008;**9**:2579–605.
64. McInnes L, Healy J & Melville J. UMAP: Uniform manifold approximation and projection for dimension reduction. 2020. Preprint at doi: 10.48550/arXiv.1802.03426
65. Pierson E & Yau C. ZIFA: Dimensionality reduction for zero-inflated single-cell gene expression analysis. *Genome Biol*. 2015;**16**:241.
66. Zhang S, Yang L, Yang J, Lin Z & Ng MK. Dimensionality reduction for single cell RNA sequencing data using constrained robust non-negative matrix factorization. NAR Genom Bioinform. 2020;**2**(3):lqaa064. doi: 10.1093/nargab/lqaa064
67. Jiang J et al. Dimensionality reduction and visualization of single-cell RNA-seq data with an improved deep variational autoencoder. *Brief Bioinform*. 2023;**24**:bbad152.
68. Luo Z, Xu C, Zhang Z & Jin W. A topology-preserving dimensionality reduction method for single-cell RNA-seq data using graph autoencoder. *Sci Rep*. 2021;**11**:20028.
69. Tran B, Tran D, Nguyen H, Ro S & Nguyen T. scCAN: Single-cell clustering using autoencoder and network fusion. *Sci Rep*. 2022;**12**:10267.
70. Yu L, Cao Y, Yang JYH & Yang P. Benchmarking clustering algorithms on estimating the number of cell types from single-cell RNA-sequencing data. *Genome Biol*. 2022;**23**:49.
71. Sreenivasan VKA, Henck J & Spielmann M. Single-cell sequencing: Promises and challenges for human genetics. *Med Genet*. 2022;**34**:261–73.
72. Aran D et al. Reference-based analysis of lung single-cell sequencing reveals a transitional profibrotic macrophage. *Nat Immunol*. 2019;**20**:163–72.
73. Shao X et al. scDeepSort: A pre-trained cell-type annotation method for single-cell transcriptomics using deep learning with a weighted graph neural network. *Nucleic Acids Res*. 2021;**49**:e122.
74. Lin Y et al. scClassify: Sample size estimation and multiscale classification of cells using single and multiple reference. *Mol Syst Biol*. 2020;**16**(6). doi: 10.15252/msb.20199389
75. Pasquini G, Rojo Arias JE, Schäfer P & Busskamp V. Automated methods for cell type annotation on scRNA-seq data. *Comput Struct Biotechnol J*. 2021;**19**:961–9.
76. Clarke ZA et al. Tutorial: Guidelines for annotating single-cell transcriptomic maps using automated and manual methods. *Nat Protoc*. 2021;**16**:2749–64.
77. Love MI, Huber W & Anders S. Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. *Genome Biol*. 2014;**15**:550.
78. 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. doi: 10.1093/bioinformatics/btp616
79. Ritchie ME et al. Limma powers differential expression analyses for RNA-sequencing and microarray studies. *Nucleic Acids Res*. 2015;**43**:e47.
80. Finak G et al. MAST: A flexible statistical framework for assessing transcriptional changes and characterizing heterogeneity in single-cell RNA sequencing data. *Genome Biol*. 2015;**16**:278.
81. Gagnon J et al. Recommendations of scRNA-seq differential Gene expression analysis based on comprehensive benchmarking. *Life*. 2022;**12**:850.
82. He L et al. NEBULA is a fast negative binomial mixed model for differential or co-expression analysis of large-scale multi-subject single-cell data. *Commun Biol*. 2021;**4**:1–17.
83. Brooks ME et al. glmmTMB balances speed and flexibility among packages for zero-inflated generalized linear mixed modeling. *R J*. 2017;**9**:378–400.
84. Liberzon A et al. The molecular signatures database (MSigDB) hallmark gene set collection. *Cell Syst*. 2015;**1**:417–25.
85. The Gene Ontology Consortium. The gene ontology knowledgebase in 2023. *Genetics*. 2023;**224**:iyad031.
86. Raudvere U et al. g:Profiler: A web server for functional enrichment analysis and conversions of gene lists (2019 update). *Nucleic Acids Res*. 2019;**47**:W191–8.
87. Sherman BT et al. DAVID: A web server for functional enrichment analysis and functional annotation of gene lists (2021 update). *Nucleic Acids Res*. 2022;**50**:W216–21.
88. Subramanian A et al. Gene set enrichment analysis: A knowledge-based approach for interpreting genome-wide expression profiles. *PNAS*. 2005;**102**(43):15545–50. doi: 10.1073/pnas.0506580102
89. Mootha VK et al. PGC-1α-responsive genes involved in oxidative phosphorylation are coordinately downregulated in human diabetes. *Nat Genet*. 2003;**34**:267–73.

90. Ma Y et al. Integrative differential expression and gene set enrichment analysis using summary statistics for scRNA-seq studies. *Nat Commun*. 2020;**11**:1585.
91. Franchini M, Pellecchia S, Viscido G & Gambardella G. Single-cell gene set enrichment analysis and transfer learning for functional annotation of scRNA-seq data. *NAR Genomics Bioinforma*. 2023;**5**:lqad024.
92. Zappia L & Theis FJ. Over 1000 tools reveal trends in the single-cell RNA-seq analysis landscape. *Genome Biol*. 2021;**22**:301.
93. Snakemake workflow: single-cell-rna-seq. 2023.
94. Peltzer A et al. nf-core/scrnaseq: nf-core/scrnaseq v2.3.2 Sepia Samarium Salmon. 2023. doi:10.5281/ZENODO.3568187
95. Khozoie C et al. scFlow: A scalable and reproducible analysis pipeline for single-cell RNA sequencing data. *Authorea*. August 19, 2021. doi: 10.22541/au.162912533.38489960/v1
96. https://www.r-project.org/
97. Mangiola S, Doyle MA & Papenfuss AT. Interfacing Seurat with the R tidy universe. *Bioinformatics*. 2021;**37**:4100–7.
98. Mangiola S, Molania R, Dong R, Doyle MA & Papenfuss AT. Tidybulk: An R tidy framework for modular transcriptomic data analysis. *Genome Biol*. 2021;**22**:42.
99. Zappia L & Oshlack A. Clustering trees: A visualization for evaluating clusterings at multiple resolutions. *GigaScience*. 2018;**7**:giy083.
100. https://www.biorxiv.org/content/10.1101/2022.02.28.482303v1
101. Zhang Y, Kim MS, Reichenberger ER, Stear B & Taylor DM. Scedar: A scalable python package for single-cell RNA-seq exploratory data analysis. *PLOS Comput Biol*. 2020;**16**:e1007794.
102. Wolf FA, Angerer P & Theis FJ. SCANPY: Large-scale single-cell gene expression data analysis. *Genome Biol*. 2018;**19**:15.
103. Luecken MD & Theis FJ. Current best practices in single-cell RNA-seq analysis: A tutorial. *Mol Syst Biol*. 2019;**15**:e8746.
104. https://www.sc-best-practices.org/introduction/prior_art.html.
105. https://satijalab.org/seurat/articles/pbmc3k_tutorial.html.
106. Sandve GK, Nekrutenko A, Taylor J & Hovig E. Ten simple rules for reproducible computational research. *PLOS Comput Biol*. 2013;**9**:e1003285.
107. Niarakis A et al. Addressing barriers in comprehensiveness, accessibility, reusability, interoperability and reproducibility of computational models in systems biology. *Brief Bioinform*. 2022;**23**:bbac212.
108. Martínez Arbas S et al. Challenges, strategies, and perspectives for reference-independent longitudinal multi-omic microbiome studies. *Front Genet*. 2021;**12**:666244.
109. https://www.frontiersin.org/articles/10.3389/fgene.2020.00303/full.
110. Dove ES. The EU general data protection regulation: Implications for international scientific research in the digital era. *J Law Med Ethics*. 2018;**46**:1013–30.
111. Wilkinson MD et al. The FAIR guiding principles for scientific data management and stewardship. *Sci Data*. 2016;**3**:160018.
112. https://www.nature.com/articles/s41597-023-02166-3.
113. Welter D et al. FAIR in action — A flexible framework to guide FAIRification. *Sci Data*. 2023;**10**:291.
114. https://fairplus.github.io/the-fair-cookbook/content/home.html.
115. https://eosc-portal.eu/.
116. Liu Z, Sun D & Wang C. Evaluation of cell-cell interaction methods by integrating single-cell RNA sequencing data with spatial information. *Genome Biol*. 2022;**23**:218.
117. Aibar S et al. SCENIC: Single-cell regulatory network inference and clustering. *Nat Methods*. 2017;**14**:1083–86.
118. https://github.com/broadinstitute/infercnv/blob/master/inst/CITATION, accessed on 16/10/2024, accessed on 16 October 2024,
119. Bahonar S & Montazeri H. Somatic Single-Nucleotide Variant Calling from Single-Cell DNA Sequencing Data Using SCAN-SNV in Variant Calling: Methods and Protocols (eds. Ng C & Piscuoglio S) 267–277 (Springer US, 2022). doi:10.1007/978-1-0716-2293-3_17
120. Hu Y, Xi X, Yang Q & Zhang X. SCeQTL: An R package for identifying eQTL from single-cell parallel sequencing data. *BMC Bioinformatics*. 2020;**21**:184.
121. Bergen V, Lange M, Peidli S, Wolf FA & Theis FJ. Generalizing RNA velocity to transient cell states through dynamical modeling. *Nat Biotechnol*. 2020;**38**:1408–14.
122. Ding J, Sharon N & Bar-Joseph Z. Temporal modelling using single-cell transcriptomics. *Nat Rev Genet*. 2022;**23**:355–68.

123. Tran TN & Bader GD. Tempora: Cell trajectory inference using time-series single-cell RNA sequencing data. *PLOS Comput. Biol*. 2020;**16**:e1008205.
124. Yeo GHT, Saksena SD & Gifford DK. Generative modeling of single-cell time series with PRESCIENT enables prediction of cell trajectories with interventions. *Nat Commun*. 2021;**12**:3222.
125. Lin C, Ding J & Bar-Joseph Z. Inferring TF activation order in time series scRNA-seq studies. *PLOS Comput Biol*. 2020;**16**:e1007644.
126. Shao L et al. Identify differential genes and cell subclusters from time-series scRNA-seq data using scTITANS. *Comput Struct Biotechnol J*. 2021;**19**:4132–41.
127. Longo SK, Guo MG, Ji AL & Khavari PA. Integrating single-cell and spatial transcriptomics to elucidate intercellular tissue dynamics. *Nat Rev Genet*. 2021;**22**:627–44.
128. Moffitt JR, Lundberg E & Heyn H. The emerging landscape of spatial profiling technologies. *Nat Rev Genet*. 2022;**23**:741–59.
129. Elmentaite R, Domínguez Conde C, Yang L & Teichmann SA. Single-cell atlases: Shared and tissue-specific cell types across human organs. *Nat Rev Genet*. 2022;**23**:395–410.
130. Lee J, Hyeon DY & Hwang D. Single-cell multiomics: Technologies and data analysis methods. *Exp Mol Med*. 2020;**52**:1428–42.
131. Baysoy A, Bai Z, Satija R & Fan R. The technological landscape and applications of single-cell multi-omics. *Nat Rev Mol Cell Biol*. 2023;1–19 doi:10.1038/s41580-023-00615-w
132. Vandereyken K, Sifrim A, Thienpont B & Voet T. Methods and applications for single-cell and spatial multi-omics. *Nat Rev Genet*. 2023;1–22 doi:10.1038/s41576-023-00580-2
133. Wang X et al. BREM-SC: A Bayesian random effects mixture model for joint clustering single cell multi-omics data. *Nucleic Acids Res*. 2020;**48**:5814–24.
134. Kim HJ, Lin Y, Geddes TA, Yang JYH & Yang P. CiteFuse enables multi-modal analysis of CITE-seq data. *Bioinformatics*. 2020;**36**:4137–43.
135. Do VH, Ringeling FR & Canzar S. Linear-time cluster ensembles of large-scale single-cell RNA-seq and multimodal data. *Genome Res*. 2021;**31**:677–88.
136. Gayoso A et al. Joint probabilistic modeling of single-cell multi-omic data with totalVI. *Nat Methods*. 2021;**18**:272–82.
137. Lin X, Tian T, Wei Z & Hakonarson H. Clustering of single-cell multi-omics data with a multimodal deep learning method. *Nat Commun*. 2022;**13**:7705.
138. Marx V. How single-cell multi-omics builds relationships. *Nat Methods*. 2022;**19**:142–6.
139. Van de Sande B et al. Applications of single-cell RNA sequencing in drug discovery and development. *Nat Rev Drug Discov*. 2023;**22**:496–520.
140. Jovic D et al. Single-cell RNA sequencing technologies and applications: A brief overview. *Clin Transl Med*. 2022;**12**:e694.
141. Regev A et al. The human cell atlas. *eLife*. 2017;**6**:e27041.
142. Rood JE, Maartens A, Hupalowska A, Teichmann SA & Regev A. Impact of the human cell atlas on medicine. *Nat Med*. 2022;**28**:2486–96.
143. BRAIN Initiative Cell Census Network (BICCN). A multimodal cell census and atlas of the mammalian primary motor cortex. *Nature*. 2021;**598**:86–102.
144. Papatheodorou I et al. Expression atlas update: From tissues to single cells. *Nucleic Acids Res*. 2020;**48**:D77–D83.
145. Lim J et al. Transitioning single-cell genomics into the clinic. *Nat Rev Genet*. 2023;1–12. doi:10.1038/s41576-023-00613-w
146. Sun G et al. Single-cell RNA sequencing in cancer: Applications, advances, and emerging challenges. *Mol Ther Oncolytics*. 2021;**21**:183–206.
147. Li L et al. What are the applications of single-cell RNA sequencing in cancer research: A systematic review. *J Exp Clin Cancer Res*. 2021;**40**:163.
148. Ke M, Elshenawy B, Sheldon H, Arora A & Buffa FM. Single cell RNA-sequencing: A powerful yet still challenging technology to study cellular heterogeneity. *BioEssays*. 2022;**44**:2200084.