Genomics 113 (2021) 606–619

Contents lists available at ScienceDirect

Genomics

journal homepage: www.elsevier.com/locate/ygeno

Review

A hitchhiker's guide to single-cell transcriptomics and data analysis pipelines

Richa Nayak, Yasha Hasija *

Department of Biotechnology, Delhi Technological University, Delhi 110042, India

ARTICLE INFO

Keywords:
Single-cell transcriptomics
Single-cell RNA sequencing
Single-cell data analysis
Computational approach
Machine learning

ABSTRACT

Single-cell transcriptomics (SCT) is a tour de force in the era of big omics data that has led to the accumulation of massive cellular transcription data at an astounding resolution of single cells. It provides valuable insights into cells previously unachieved by bulk cell analysis and is proving crucial in uncovering cellular heterogeneity, identifying rare cell populations, distinct cell-lineage trajectories, and mechanisms involved in complex cellular processes. SCT data is highly complex and necessitates advanced statistical and computational methods for analysis. This review provides a comprehensive overview of the steps in a typical SCT workflow, starting from experimental protocol to data analysis, deliberating various pipelines used. We discuss recent trends, challenges, machine learning methods for data analysis, and future prospects. We conclude by listing the multitude of scRNA-seq data applications and how it shall revolutionize our understanding of cellular biology and diseases.

1. Introduction

As an elementary school textbook would exclaim, cells are the fundamental, structural, and functional unit of all living organisms. Understanding the biology of cells has been at the center of our pursuit to unravel the complexities that make an organism. Cell biology research has undergone a remarkable transformation in recent years with the advent of single-cell multi-omics technology. The genome structure of every cell is essentially the same for any given individual organism; however, the genome's expression pattern determines the physiological fate of the cell. The observed diversity of phenotypes is due to the genotype and the varying expression pattern, abnormalities in which form the basis of various diseases. Mapping of this unique genotype-phenotype relationship requires transcriptome profiling, and recent progress made in high throughput sequencing technologies has enabled the measurement of transcriptomic information at an unprecedented resolution of single cells [1].

Transcriptome profiling has revealed that, for any given cell, the transcriptome information reveals the activity of merely a subset of genes [2], and each cell type has a unique transcriptomic fingerprint. Earlier, transcriptome profiling was based on the assumption that all cells from any given tissue are homogenous, and bulk population sequencing followed by average expression analysis would provide us with sufficient information to understand gene expression in both standard and abnormal cell states. Increasing evidence suggests that even in similar cells, the gene expression pattern can be heterogeneous [3,4]. Although bulk expression analysis could simultaneously assess gene expression levels and differentiate between abundant known cell types, it could obscure the identification of rare cell types, subtypes and fail to distinguish cell to cell variability [5]. Thus, the understanding of stochastic cellular processes necessitated a more precise transcriptome analysis technique to overcome the averaging phenomenon inherent to bulk analysis. Unabated technological advancements in NGS, molecular biology, cell biology, and bioinformatics has fostered a new wave of profiling single cells at genomics, transcriptomics, proteomics, and epigenomics level [6,7].

Single-cell transcriptomics (SCT) involves profiling the complete set of RNA transcripts of each individual cell for a given population of cells [8]. Transcriptome analysis of single-cell was pioneered two decades ago, in two separate historical experiments, one by Norman N. Iscove [9], and James Eberwine and group [10,11], that laid the groundwork for single-cell transcriptome analysis based on high throughput sequencing technologies [12]. scRNA-sequencing is a fast-growing and promising technology for SCT [13] and has rendered microarrays and qPCR obsolete.

The volume and complexity of scRNA-seq data make it a paradigm of big data [1], it has opened doors to a multitude of possibilities in biomedical research, but we have only tapped a fraction of the potential

* Corresponding author.
E-mail address: yashahasija@dtu.ac.in (Y. Hasija).

https://doi.org/10.1016/j.ygeno.2021.01.007
Received 9 August 2020; Received in revised form 30 December 2020; Accepted 18 January 2021
Available online 22 January 2021

of such a large and versatile dataset. scRNA-seq transcriptome profiles have paved the way for identification of rare cell types in complex tissues [14], cell lineage relationships in early development [15], antigen specificity of immune cells [16], inferring cellular trajectory [17], determination of cell fate [18], distinguishing between normal and abnormal cell [19], understanding tumor heterogeneity [20], identifying regulatory signatures in cancer [21], deciphering immune repertoire for infectious diseases [22], elucidating the mechanism for drug resistance, and relapse in cancer treatment [23]. With better analysis methods, we are uncovering more applications.

scRNA-seq data, although highly potent, poses many challenges on various fronts owing to its big-data characteristics such as sophisticated data acquisition techniques, data storage, management, and analysis [24]. A single scRNA-seq experiment generates a larger volume of high-dimensional raw data than bulk sequencing methods as it retains the information of the stochastic expression of genes for individual cells. In addition, scRNA-seq experimental protocols have more steps compared to bulk sequencing, which gives rise to more technical biases and artifacts.

Experimental techniques for scRNA-seq have mushroomed and improved over time, which has led to the generation of a massive amount of data and an increasing demand for computational techniques for data analysis. That has led to a spike in developing new experimental protocols, algorithms, and tools to analyze the raw data. Several research groups and commercial companies have designed software tools and packages for data preprocessing and downstream analysis. Machine Learning (ML) approaches, preeminent in big data analysis, have been a noteworthy addition to the list of approaches used for the underlying computational challenges of dimensionality reduction, clustering, and differential expression (DE) analysis [25,26]. Furthermore, there are choices between various programming languages like R, Java, MATLAB, C++, and Python. The development of analysis tools is still in its infancy, and current tools have many shortcomings. The challenge to improve the reciprocation between speed and accuracy in analysis remains. Despite the abundance of techniques, it is hard to establish a standard that can be used across disciplines, and it is crucial to make an informed decision while proceeding for analysis as it can have a tremendous impact on the findings. The scRNA-seq analysis tool choice can influence detecting a biological signal comparable to quadrupling the sample size [27].

This review outlines the general workflow involved in single-cell RNA-seq protocols and discusses the popular and promising new computational tools for analysis. It provides a comprehensive account of each step of the analysis, starting from data preprocessing, imputation of dropouts to tools used for pseudotime ordering, and rare cell type identification. It also discusses ML approaches in the analysis steps, wherever applicable. The review concludes with a discussion of applications across fields in biological sciences, remaining challenges, and prospects.

2. Single-cell RNA sequencing technology

A wide range of scRNA-seq protocols has been developed to accommodate the high demand for improved techniques with high throughput. New methods are being developed to counter batch effects and technical noise since it is important to regulate the initial steps to alleviate the computational burden during data analysis.

scRNA-seq technologies currently in use can be divided into four broad classes based on transcript coverage approach [28]: (i) full-length transcript sequencing [example- MATQ-seq [29], SMART-seq2 [30], ICELL8 [31] SUPeR-seq [32]], (ii) 5′-end transcript sequencing [example- STRT-seq [33,34]], (iii)3′-end transcript sequencing [example- Chromium [35] 10X Genomics, Fluidigm C1 [36], Drop-seq [37], inDrop [38]]. With full-length transcript sequencing approach, there is an issue of resolution, speed, and sequencing cost. On the other hand, a major drawback of cDNA sequencing prioritizing either 5′ or 3′-end transcripts of the DNA is incapable of examining allele-specific expression or alternative splice forms. Some methods rely on FACS based sorting, such as MARS-seq, that make them reliant on a larger initial volume [39] and is disadvantageous when the initial volume is low, as in fine-needle aspirates. Another drawback of using FACS is the requirement of antibodies that target specific proteins for sorting; this poses problems while sorting rare cell subtypes [40]. Thus, each protocol has its sets of advantages and disadvantages that determine the "depth" (reads/cell) of a given dataset, and it could ultimately affect the statistical and biological insight [41]. scRNA-seq is not a "one-size-fits-all" technique like the bulk sequencing approach since the depth can vary with the protocol being used, cell types being examined, capture method, sequencing technique, and alignment stringency during library construction [42]. We discuss a typical scRNA-seq experimental workflow and complementary technologies in use.

2.1. scRNA-seq workflow

2.1.1. Single-cell isolation

The generation of scRNA-seq data from a tissue sample involves multiple steps. First, the tissue is digested to ensure dissociation, which gives rise to the single-cell suspension from which single-cells are isolated so that each cell's mRNA can be profiled separately. Single-cell isolation technique predominantly used for scRNA-seq are plate-based techniques and microfluidic-based techniques. Plate-based techniques involve capturing or sorting cells on multi-well plates or microfuge tubes, followed by FACS based sorting. Some full-length scRNA-seq techniques like SUPeR-seq, SMART-seq2, MATQ-seq, Cell-seq rely on plate-based techniques. However, there are many limitations, one being fewer cells per assay than droplet-based technologies. Microfluidic technology involves capturing cells in its microfluidic droplet. Microfluidics-based techniques have swiftly gained popularity amongst single-cell isolation techniques as it requires less initial volume, is cost-effective, and aids in massively parallel quantification of single-cell gene expression profiles [43,44]. Microfluidics techniques can be of two types, i.e. (i) continuous-flow microfluidics, like the Fluidigm's C1 Single-Cell Auto Preparation System, (ii) droplet-based microfluidics like InDrop, Drop-seq, and 10X Genomics. Comparison between single-cell isolation techniques is well depicted by Chen X. et al. 2018 [45] and Ziegenhain C. et al. 2017 [46]. Droplet-based platforms are readily automated and easily optimizable to suit individual experimental needs depending upon the number of cells to be captured and sequenced. Recent studies have an added advantage of not requiring zero inflation over plate-based techniques that need zero inflation for accurate simulation [27].

Another technique for single-cell isolation from solid tissue is Laser Capture Microdissection, LCM-seq, a laser system-aided isolation of cells directly from solid tissue, coupled with in-situ RNA-sequencing techniques [47,48], which conserves spatial information of mRNA expression within the morphology of a tissue. This method enables isolation of rare cell types even in highly heterogeneous clinical samples, with increased accuracy and understanding of dynamic cellular systems.

In the lookout to overcome these techniques' inefficiencies, nanowell based single-cell isolation like Seq-Well promises cost-effectiveness, throughput, and portability and requires only nanoliter sized initial volume [49]. Some newer techniques eliminate the single-cell isolation step, like in SPLiT-seq (split-pool ligation-based transcriptome sequencing) and sci-RNA-seq (Single-cell Combinatorial Indexing RNA sequencing [50]). SPLiT-seq allows for simplified and low-cost transcriptome profiling compatible with fixed cells or nuclei and offers high-resolution [51].

2.2. scRNA-seq library preparation

Much like any RNA library preparation, it roughly entails reverse transcription of captured mRNA into first-strand cDNA synthesis,

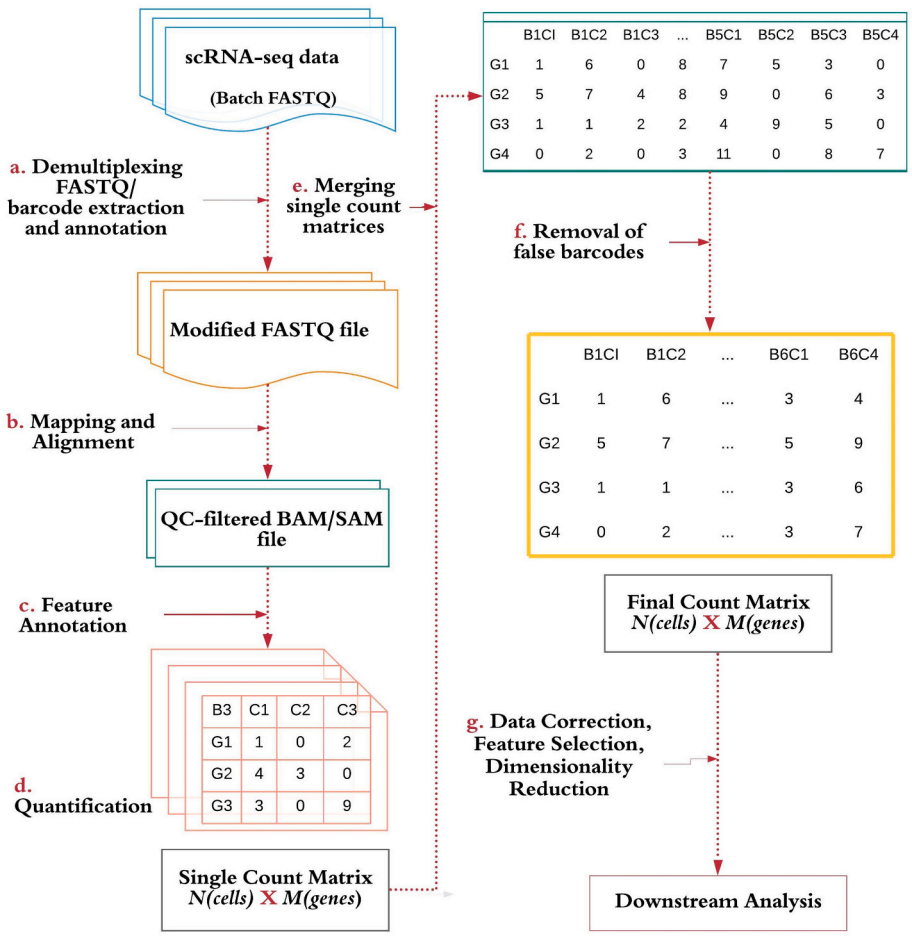

Fig. 1. Preprocessing (a) The first step after obtaining the scRNA-seq data is demultiplexing FASTQ batch data. (b) The demultiplexed data (modified FASTQ file) is mapped against respective genome using an alignment program. (c) Feature annotation is carried out using gene annotation file that contains all information on genes, exons, introns and regions of interest (RefSeq, GENCODE). All the reads are filtered, keeping reads that align to the forward and reverse strand with less than three mismatches and mapped only once to the reference genome. (d) Once the gene names for specific reads are obtained, how many reads correspond to which genes can be determined. Tabular count matrix depicting genes/ features as rows and cell labels as columns is generated. (e) Multiple single count matrices are joined together to form a combined matrix. (f) Not all batches use same barcodes thus barcodes are filtered for the combined matrix to ensure there is no cross-contamination and a final QC-filtered count matrix is generated. (g) Feature selection and dimensionality reduction is carried out on the matrix carried out on the count matrix followed by downstream analysis.

second-strand synthesis, and cDNA amplification followed by sequencing [52]. But it is more challenging as the amount of RNA per cell is low compared to bulk RNA-seq experiments. A thorough analysis of scRNA-seq required the profiling of a large number of represented individual cells, which is a task worthy of Sisyphus and significantly adds to the cost of carrying out sequencing. The use of Unique Molecular Identifiers (UMI) or cellular barcodes somewhat simplified the process. Cell barcodes are primarily designed to be able to distinguish between read transcripts originating from different cells. To fully determine the uniqueness of the reads, UMIs (short molecular tags composed of a unique random sequence) are added to the reverse transcription step (5'end in template switching or 3′ end in oligo-dT primer) [53]. They constitute the second portion of a barcode and primarily detect and quantify unique mRNA transcripts [46] such that amplicons of the same transcript are only counted once. This allows for multiplexing of scRNA-seq, even for low abundant transcripts that show poor reproducibility with previous quantification methods based on the number of sequencing reads [54]. Still, current UMI based approaches are poorly suited for the identification of allele-specific expression or alternative splice variants [2]. Post library construction, cDNA libraries labeled with cellular barcodes are pooled for sequencing, mapping, and alignment. Current RNA-seq technologies rely on pooled-sequencing in order to generate high throughput data. This allows for amplification and sequencing of multiple cells parallelly in the same pool that generates batch-specific output data containing sequences from multiple cells.

3. Preprocessing scRNA-seq data

Sequencing generates reads (raw data) that need to undergo quality control (QC) before downstream analysis. scRNA-seq data contains many technical artifacts that may arise due to cell bursting leading to RNA leakage, multiple cells sticking together, and lowly expressed RNA leading to dropouts, amplification bias, transcriptional bursting, RNA degradation, and batch effect. Before performing downstream analysis, it is crucial to ensure that all the cellular barcode data obtained from scRNA-seq correspond to viable cells [55]. Another challenge is to prevent false interpretation of technical artifacts (cells that show technical noise that appears as distinguishable gene expression pattern) as biological heterogeneity [56].

A typical scRNA-seq dataset constitutes of three files, genes quantified (gene IDs), cells quantified (cellular barcode data), and a count matrix, irrespective of technology or pipeline used. These files are crucial for building quality matrices for QC assessment. The barcodes are extracted and annotated, called demultiplexing or barcode extraction, followed by mapping and alignment of the read data using read processing pipelines. Post alignment, feature annotation, and quantification are carried out on the data to generate gene expression matrices (*N(cells) x M(genes)*) indicative of the level of gene expression in each cell, based on the molecular counts or read counts Fig. 1. These correspond to high mapping quality exonic loci [55]. A list of preprocessing and downstream analysis tools is shown in Table 1.

3.1. Quality control

QC filtering can be performed using a combination of strategies. Some common strategies are based on assigned barcodes- number of counts per barcode and number of genes per barcode [57]. Cells that show unique gene counts and genes expressed in very few cells are not always indicative of biological heterogeneity, low or high count-depth are indicative of quiescent/damaged cells or doublets/multiplets,

Table 1
List of some popular preprocessing pipelines and downstream analysis tools used for scRNA-seq data.

Analysis category	Pipeline	Environment	Description
Overall analysis	Seurat [78]	R	A comprehensive tool to perform QC analysis on diverse types of single-cell data. Spatial interference using in situ RNA patterns as a reference. Compatible with multimodal data. http://satijalab.org/seurat/
	Scanpy [79]	Python	A scalable analytical framework for scRNA-seq analysis starting from preprocessing, visualization, DE, clustering, and TI works in tandem with AnnData-annotated data matrices. Fast and efficient for large data sets. https://github.com/theislab/scanpy
	PyMINEr [80]	Python	An automated tool for cell-type identification, filtering enriched genes, network pathway analysis, and visualization of analysis. Non-cell type determining gene expression may influence cellular graphs. https://www.sciencescott.com/pyminer
Pre-processing	dropEst [81]	R	Accurate estimation, quality control, correcting composition bias, and sequencing error for droplet-based scRNA-seq data. Provides configuration options for accommodating different scRNA-seq protocols. More efficient with smaller datasets than larger datasets. https://github.com/hms-dbmi/dropEst
Visualization	Cerebro [82]	R	An interactive environment, compatible with Seurat objects. https://github.com/romanhaa/Cerebro
	iSEE [83]		Interactive visualization, reproducible, and compatible with existing R/Bioconductor packages. https://github.com/csoneson/iSEE
Imputation	bayNorm [73]	R	An integrated platform for normalization, imputation, and batch effect correction. Improves accuracy and sensitivity of DE analysis. https://bioconductor.org/packages/release/bioc/html/bayNorm.html
	DeepImpute [84]		Employs a deep neural network-based algorithm that allows for improved speed, accuracy, and scalability. It is also well suited for large ever-increasing datasets. https://github.com/lanagarmire/DeepImpute
Batch effect, merging	BERMUDA [85]	R	Deep Autoencoders. Data is obtained from scRNA-seq data, and one common cell type is required at the least from batches for the process https://github.com/txWang/BERMUDA
	MNN [86]	R	Mutual Nearest Neighbor for batch correction of single cells. Based on the established assumption that a batch is orthogonal to biology and that MNN exists between batches. https://github.com/MarioniLab/MNN2017/
Normalization	SCnorm [63]	R	Makes use of quantile regression to approximate the dependency of expression of a transcript on the depth of sequencing per gene. Similar dependency genes are clustered together, and then a second quantile regression is utilized to approximate scale parameters in every group. For depth sequencing, then in-group correction is achieved by using the approximate scale parameters to deliver normalized estimations of expression. https://www.biostat.wisc.edu/~kendzior/SCNORM/
Dimensionality Reduction	DR-A [87]	Python	Dimensionality Reduction with Adversarial variational autoencoder
	SAUCIE [88]	Python	Deep Multitasking Neural Networks for DR
Cell clustering	Refer to Table 3		
Cluster Annotation	scMAP [89]	R	Automated cluster annotation technique follows a projection-based approach where scRNA-seq data is projected onto a previously annotated cell type or dataset.
Pseudotime trajectory inference/ reconstruction	TinGa [90]	R	TI based on Growing Neural Graphs. Scalable, time-efficient, and accurate on complex trajectories. Does not require prior specification of the topology by the user. https://github.com/Helena-todd/TinGa
	ReCAT [91]	R	Hidden Markov model-based method for reconstructing cell cycle pseudotime in time-series data https://github.com/tinglab/reCAT
Differential Expression	MAST	R	Uses a linear hurdle model to account for confounders, and DE is determined using the likelihood ratio test.

(*continued on next page*)

Table 1 (*continued*)

Analysis category	Pipeline	Environment	Description
	SCDE	R	https://github.com/RGLab/MAST\ Uses a Bayesian approach that incorporates an evidence-based approach to evaluate the likelihood of the average level of gene expression for individual cells and measure the fold changes. Highly sensitive. https://hms-dbmi.github.io/scde/index.html

*Note:*https://www.scrna-tools.org/ is a catalog of tools for analyzing single-cell RNA sequencing data.

respectively. Another covariate used for QC is the fraction of mitochondrial genes per barcode. Elevated levels of mitochondrial genes (above 5–10%) in a cell is an indication that the cell may have broken and the cytoplasmic mRNA content has leaked. Furthermore, RNA spike-ins (synthetically generated short RNA polymers of known quantity) are used for calibration purposes, where a low mapping ratio between endogenous RNA and spike-ins is indicative of a low-quality library [58].

The quality metrices are visualized to determine the outlier cells. Low-quality cells are filtered out by setting appropriate thresholds. While filtering out outlier cells, multiple independent variables must be considered together rather than individual ones as it can lead to misinterpretation of biological heterogeneity. Thresholds can be fixed or adaptive. Setting fixed thresholds requires experience as suitable thresholds may vary for each experimental protocol or biological systems [59]. An alternative is adaptive thresholds that are decided based on the outlier peaks for the QC covariates. It is essential to reevaluate the QC metrices after filtering before proceeding for further analysis.

3.2. Normalization

UMI-based protocols inherently reduce amplification biases, and the addition of spike-ins enables assessment of sensitivity. However, cell-level (counts comparable between cells) and gene-level (counts comparable between genes) normalization is carried out to cater to sampling effects or technical biases that remain in the data due to variability in the protocols. Once the count matrix is obtained, normalization sought to address the gene expression variability between cells in count data to prevent the highly expressed genes from influencing the analysis. Popular normalization methods have been derived from bulk RNA-seq analysis methods and have been successfully applied to scRNA-seq data such as DESeq and Trimmed Mean of M-values [60]. Popularly, raw scRNA-seq read library normalization is carried out using read count normalization/ CPM (Counts Per Million) methods like RPKM (Read Per Kilobase Million), TPM (Transcripts Per Million), and FPKM (Fragments Per Kilobase Million). The scaling factors in these methods are based on the assumption that the majority of the genes are not differentially expressed, so they might fail when fold-change of DE genes is high across cell populations under study [60]. These library-based or global-scaling normalization methods derived from bulk RNA analysis, other than being computationally intensive, have their shortcomings when used for scRNA-seq data. Given the added complexity of scRNA-seq data due to data sparsity and high heterogeneity, it requires advanced normalization strategies to address specific biases.

SINCERA is a commonly used normalization pipeline, in which gene normalization is performed by z-score, while cell normalization is performed by trimmed mean. For example, during clustering, BISCUIT uses iterative normalization by learning features representing technical modifications. RaceID (Rare cell type identification) normalizes the total transcript count within each cell to the median transcript number across cells [61]. Clustering based on Transcript Compatibility Counts (TCC) uses equivalence classes in place of genes as parameters and normalizes each parameter by distributing the total count across all the cells. Sctransform pipeline interfaces with Seurat and Pearson residuals from negative binomial regression, which has been regularized. In regression, sequencing depth is used as a covariate to eliminate technical artifacts [62]. SCnorm uses quantile regression to approximate the dependency of expression of transcript or depth of sequencing per gene [63]. Similar dependency genes are clustered together, and then a second quantile regression is utilized to approximate scale parameters in every group. In-group correction is achieved by using the approximate scale parameters to deliver normalized estimations of expression.

The diversity of scRNA-seq protocols makes it difficult to standardize any one normalization method. It has been observed that different normalization methods perform optimally for different datasets, and the same goes for cell-level and gene-level normalization. Post normalization, log(x + 1) transformed count matrices are obtained that give a simplified account of expression levels in terms of log-fold changes and bring down the skewness of the data [55]. Downstream analyses that are based on the assumption that scRNA-seq data is normally distributed and perform analysis on log-transformed data may sometimes result in counterfeiting DE effects. Thus, there is a pressing need to develop more precise and robust normalization methods designed explicitly for scRNA-seq data.

3.3. Data correction

Normalized data successfully removes amplification and count depth biases; however, a few challenging technical and biological biases remain. Data correction deals with batch effects, dropouts, and biological effects. scRNA-seq data is prone to zero-inflated values, otherwise known as dropouts [64], resulting from low sensitivity of scRNA-seq protocols, inefficient capture of mRNA, low amounts of mRNA in cells, or transient gene expression [2]. The dependency of downstream analysis on the accuracy of gene expression profiles makes the imputation of dropouts a crucial step. Many analysis pipelines account for dropouts during analysis; however, recent findings can change how we look at dropouts. Imputation is carried out either by direct expression analysis or model-based. Newer and more robust ML-based algorithms have taken over popular imputation techniques like Markov Transition Matrix-based MAGIC [65], Clustering-based DrImpute [66], LASSO regression-based ScImpute [64]. SAVER(Single-cell Analysis Via Expression Recovery) uses gene-gene relationships to recover true expression levels of each cell [67], RESCUE (REcovery of Single-Cell Under-detected Expression) enhances cell-type identification based on an ensemble-based method to minimize feature selection bias and count error and perform imputation by comparing gene expression levels of other cells with similar patterns [68]. SCRABBLE, a matrix regularization framework, uses bulk RNA seq data as a constraint that improves the accuracy and estimation of gene expression distribution across cells compared to scRNA-seq analysis in isolation [69]. For datasets that suffer from imbalance and limited sample sizes, scHinter, with a hierarchical framework for random interpolation by leveraging minority oversampling technique [70], proves to be a robust technique than its predecessors. A recent study explores the opposite view of dropouts, where instead of imputation, it can be used as a signal. It was observed that binary dropout patterns prove almost equally informative as quantitative expression patterns of highly variable genes in cell type identification [71].

Furthermore, studies have been conducted to establish that the excess of dropouts is consistent with stochastic sampling of molecular counts, and any additional zero values may result from biological variation [72]. Such studies suggested that a negative binomial distribution model for UMI based scRNA-seq count data would suffice, and zero-inflation may not always be necessary [46,72,73]. It can also be

inferred that the number of dropouts can be decreased by increasing the depth of sequencing or increasing global count with more efficient capturing methods.

Apart from dropouts, several other technical covariates like batch effects and biological covariates need to be considered. Removal of such biases must be carried out simultaneously as there might be a dependency between multiple covariates under consideration. Batch effects arise from data handling in different experiment batches or time points and are highly nonlinear variations. Batch effects can have a significant impact on DE analysis. Some methods include aggregation-based methods [74] that pool cells from batches to form a pseudo-bulk sample and use bulk analysis approaches or nested fixed [75] and mixed effect models [76] that treat batch effects as fixed effects nested within each group or random-effects shared between cells from each batch, respectively. A comparison between different batch correction methods is given in Chen et al., 2020 [77], and some popular batch correction methods are listed in table 1. Upon removal of batch effects, data is merged for further unbiased analysis. Biological covariates may arise due to important biological processes like cell cycle effects that affect cell-size and mRNA counts. Correcting for such variation helps reveal important biological signals and processes. Linear regression against a cell cycle score, and correction for cell size during normalization are some ways to remove the effects of the cell cycle [55]. However, data correction for biological effects may not always be in the best interest, and correction for one effect may mask another. Thus, it is advised first to evaluate the study's objective and context before deciding on data correction measures.

Despite the numerous QC measures, it is hard to determine each step's stringency before assessing its effect on downstream analysis. Thus, a feedback system should be followed to regulate QC stringency alongside downstream analysis.

3.4. Dimensionality Reduction (DR)

scRNA-seq data is computationally intensive, noisy, and suffers from the curse of dimensionality. scRNA-seq expression metrics are of high dimension, but not all genes are required for meaningful classification of cellular expression profiles, and it can practically be explained in fewer dimensions, focusing only on relevant biological signals. DR enables better data visualization and resolves the statistical issue of data sparsity. An effective low-dimensional representation should summarize the data in a few optimal dimensions that must retain the underlying structure in the data to describe the variability of the dataset. Some of the popular techniques for DR are Principal Component Analysis (PCA), t-distributed Stochastic Neighbor Embedding(t-SNE), Uniform Manifold Approximation, and Projection (UMAP), Self Organizing Maps (SOM), and Model embedded dimension reduction [92]. Some clustering pipelines come integrated for both dropout imputation and DR like CIDR [93].DR has two components: feature selection, where one selects a smaller subset from the original set of variables, and the other part is feature extraction, where the high dimension data is projected to a lower dimension. Feature selection is carried out based on the expression variability of genes according to the assumption that genes showing high variability correspond to biological variation. Per-gene variation can be quantified by calculating the variance of log-normalized value. A subset of highly variable genes (HVG), depending on the type of dataset, is selected for further analysis. Selecting a larger subset of HVGs may increase the noise but reduces the risk of discarding biologically relevant signals. Seurat performs feature selection by modeling the mean-variance relationship [94].

After feature selection, DR is carried out using linear or nonlinear techniques. PCA is a linear projection method that linearly transforms the original dataset into PCs ranked in decreasing order of variance, and the data's variance is maximized in the lower dimensional space. It is computationally efficient and removes redundant features, but since scRNA-seq has a highly nonlinear structure, PCA alone is not best suited

Table 2
Classes of clustering algorithms used in scRNA-seq following ways [92].

Class of clustering algorithm	Principle	Limitations	Pipeline
Distance-matrix	Unsupervised learning algorithms like k- means falls in this category. This algorithm first identifies k centroids or means iteratively, and data points are assigned to the cluster around the nearest centroids. During cluster allocation of datapoints, the in-cluster sum of squares is reduced, and the position of centroids is iteratively optimized. Scalable and time-efficient.	Sensitive to outliers, biased towards data shape/cluster shape, and the number of clusters must be specified beforehand.	SCUBA [102], PCAKmeans, pcaReduce, SAIC [103], scVCMD [104]
Hierarchical clustering	Generates clusters into a hierarchical structure and is popular in gene expression analysis. It overcomes the limitation of k-means of specifying the number of clusters a priori and handling different shapes of clusters. No assumptions are made about the distribution of data points, and each cluster links to another by branches and is nested like a hierarchical tree in the form of a dendrogram. This representation facilitates meaningful data interpretation.	Time intensive	BackSPIN, cellTree [105], DendroSplit [106] CIDR [93]
Graph-based	Supervised-learning algorithms. Projects a graph representation of data in which the nodes correspond to datapoints/cells, and edges correspond to pairwise similarity between the datapoints, for example, K-Nearest Neighbor. Clusters are based on neighboring cell pairs. Graph-based clustering techniques have various subtypes like Louvain	Reliance on heuristic solutions sometimes leads to spurious results, and iteration sometimes masks small communities.	Seurat [78] scanpy [79] SNN-Cliq [108]

(continued on next page)

Table 2 (*continued*)

Class of clustering algorithm	Principle	Limitations	Pipeline
	clustering that is used by popular scRNA-seq analysis pipeline. Louvain algorithm is largely getting taken over by the Leiden algorithm for cluster detection, which uses a smart local move approach faster and shows more proficiency in detecting well-connected communities [107].		
Mixture Models	Clustering is based on the probability distribution of the datapoints. It is well suited for the identification of subpopulations and integrates prior knowledge as assumptions of probability distributions.	Computationally intensive and relies on the accuracy of assumptions of probability distribution.	BISCUIT [109], Seurat, TSCAN [110]
Density-based	Density-based algorithms assign clusters based on high-density regions of datapoints. It is a highly efficient clustering technique but is sensitive to parameters.	Sensitive to parameters, time-intensive	Monocle2 [111] (for the identification of outlier cells)
Neural Network	Supervised-learning methods, inspired by the neural network of the human nervous system. Highly efficient in performing clustering and classification tasks. Kohonen networks are bilayer networks that use competitive learning for clustering. Deep learning and autoencoders are also used. These are efficient, scalable, and information on relationships amongst clusters can be incorporated.	Sensitive to parameters	SAUCIE [88] scDeepCluster [112]
Ensemble clustering	Ensemble clustering algorithms are a solution to the lack of any one optimal algorithm as it makes use of multiple clustering algorithms on the same dataset, and a consensus result is obtained that is more precise than independent algorithms.	Disadvantages of individual algorithms add up	SAFE-clustering uses SC3, CIDR, Seurat, and t-SNE + k-means for clustering and then combines the result to obtain one consensus result by using a hypergraph partitioning algorithm [113]

for data visualization. However, t-SNE, a graph-based nonlinear technique, is much better suited for the task. It is often performed in tandem with PCA. t-SNE is based on a probabilistic distance model. It creates probability distributions to establish a relationship between two data points in high dimensional space and reconstructs it in a lower-dimensional space by optimizing using gradient descent. Although t-SNE is favorable for data visualization, good with nonlinear datasets, and preserves local structures in lower dimensions, it often ignores the global structure, which may lead to misinterpretation of differences between cell populations. Owing to these drawbacks, a superior manifold learning-based DR technique has emerged called Uniform Manifold Approximation and Projection (UMAP) [95]. It is principally similar to t-SNE but also preserves the global structure. It is also computationally more efficient but may sometimes result in spurious signals in smaller datasets. For complex dataset visualization, partition-based graph abstraction (PAGA) is used with UMAP. ZIFA [96] and ZINB-WaVE [97] are model-embedded dimension reduction algorithm carried out on zero-inflated data. ML methods, like Deep Learning and Autoencoders, have shown efficiency in DR problems. Latent Dirichlet Allocation (LDA), a Natural Language Processing-based algorithm, and SAUCIE [88], a neural network-based algorithm, are promising new approaches. DR-A, an autoencoder based framework, provides precise low dimensional representation, enhances downstream clustering performance, and could potentially be used for lineage estimation [87].

4. Downstream analysis

The unique features of single-cell data make downstream analysis elaborate and diversified. There are different stages of preprocessed data such as log-transformed data, batch corrected data, feature selected data, dimensionality reduced data, etc. Depending on the requirement specifications of downstream analysis and availability, certain preprocessed data is chosen. Some open-source scRNA-seq repositories or reference databases are Human Cell Atlas [98], Broad Institute's Single Cell Portal [99], EMBL-EBI's Single Cell Expression Atlas [100], PanglaoDB [101], and OmicX Jingle Bells. 10× Genomics offers datasets at various preprocessed levels for downstream analysis.

Downstream analysis of scRNA-seq data can be of cellular level or gene level. This section will discuss various analysis tools and their applications that have led to prominent discoveries.

4.1. Cell- level analysis

The cell-level analysis is about understanding cell subtypes, cell differentiation patterns, cell lineage, cellular trajectories, identify novel cellular markers, and many unique cellular features. It helps characterize known properties of cells and uncover previously unidentified characteristics. Cell-level analysis is not exclusive to gene-level analysis. We briefly discuss important cell-level analysis techniques that have found major implications in advancing cell biology research.

4.1.1. Cluster analysis

Clustering entails categorizing cells into clusters to enable the identification of cell types and subtypes. It is ideally performed on a

dimensionally reduced dataset, and clusters are made based on cell-specific molecular profiles. Results of a clustering analysis can itself be of much significance or can serve as a covariate in other downstream analyses. There has been an attempt to develop robust cluster analysis algorithms (as shown in Table 2) that can address the vast heterogeneity of cells, but there is still a lack of an optimal algorithm that fares well across datasets.

After clustering, the clusters need to be annotated in order to give them biological relevance. *Cluster annotation* can be achieved either by thorough examination of literature, reference cell databases, or by identifying gene signatures or differentially expressed marker genes. CellAssign [114] and scMAP [89] rely on the former technique for cluster annotation, but since most marker genes have been identified by bulk analysis in the reference datasets, it is limited to giving a classical view of the cell types. Also, it is not necessary that cell types in reference databases will correspond to all the cell types present in the dataset under investigation. Analysis pipelines like Seurat, SC3, scVDMC, make use of differential gene expression approach. Full gene expression profiles are used in DE analysis for marker gene identification and cluster annotation, that is performed using simple statistical tests. The quantitative levels in gene expression are measured amongst clusters and all the cells in the dataset. Based on statistical tests like Wilcoxon rank sum test (used in Seurat), Welch's *t*-test, Kruskal-Wallis test, etc. marker gene sets are identified i.e., the top-ranked genes from these tests. DE analysis could either be carried out in succession to clustering like in Seurat, simultaneously like in scVDMC and DendroSplit, or by using DE software like MAST [115], SCDE [116], and ZingeR. Gene set enrichment analysis is carried out against reference gene sets (set of genes grouped as they share common chromosome location, biological function, or regulation) using statistical parameters like Jaccard index, and clusters are annotated accordingly. The important thing to note here is the *p*-value that is based on the assumption that the marker gene identified represents the biological phenomenon, but the *p*-value is often inflated and leads to an overestimation of marker genes. Most existing GSEA methods have been developed for bulk RNA seq analysis and perform poorly in case of scRNA-seq, thus Ma, Y., Sun, S., Shang, X. et al. came up with an integrative DE-GSE analysis technique called iDEA [117] that makes use of DE summary statistic and thus easy to use with current DE methods and efficiently produces well-assessed *p*-values for enriched gene set detection. Nowadays, automated cluster annotation techniques are becoming increasingly available like scMAP [89], which follows a projection-based approach where scRNA-seq data is projected onto a previously annotated cell type or dataset. Garnett uses a supervised classification approach for rapid annotation [118], scCATCH [119] makes use of CellMatch reference database for annotation followed by evidence-based scoring for increased performance. Another important goal is to identify rare cell types that may appear as outliers in clustering results that only consider global differences in gene expression. RaceID [61] and GiniClust [120] are clustering algorithms sensitive to identifying rare cell types. RaceID is based on the assumption that a given cell-type must express some genes that are specific to the cell type and appear as outliers but if the focus is shifted from global to such cells and the technical and the biological noise are accounted for by setting appropriate thresholds, it will enable the identification of rare cell types.

To an extent, the determination of cell-types is dependent on user-defined criteria, since for different researchers, the level of cluster resolution may vary, and in some cases, sub clustering of clusters may also be required. The choice of the resolution also has a significant effect on the results. DendroSplit, a clustering framework, allows the user to cluster using feature selection that enables identifying multiple levels of biologically meaningful cell populations in a dataset, also suitable for detecting rare cells. [106]. Despite continual attempts on developing new algorithms, clustering and cluster annotations suffer from various challenges in both biological and computational front. It is advisable to use a cocktail of automated and manual cluster annotation measures to get precise results. But since clustering is an unsupervised learning approach, several parameters are needed for reliable evaluation. Statistical and experimental validation is often needed. Transient biological states make it difficult to identify cell states. However, the generation of more comprehensive and extensive cell atlases will facilitate better clustering, cluster annotation results, and better computational approaches to help overcome technical challenges.

4.1.2. Trajectory inference

Trajectory Inference (TI), also known as pseudo-temporal ordering, is a process of characterization of underlying dynamic cellular processes. Although clustering successfully builds discrete clusters of cell types and subtypes, it does not account for the variability due to dynamic cellular processes like transient cell states in cell differentiation, cell cycle, or environmental effect. TI deals with this blind spot by ordering cells along a continuous path that minimizes transcriptional changes between successive cell pairs, called pseudotime (one dimensional manifold), that represents the progression of the cell through its dynamic processes measured in terms of transcriptional changes that a cell undergoes during a biological process. Some datasets have an expected temporal component such as cells from developing embryos, immune cells during an immune response, tumor cells, progenitor stem cells, etc. Understanding of cell differentiation via bulk RNA analysis gave us the impression that cell differentiation occurs in discrete stages, but in reality, it is a continuous process that may appear chaotic but can be ordered along continuous trajectories. Following the cells along a pseudo-temporal trajectory and analyzing gene expression changes yields valuable insights into the cellular regulatory processes, dynamic states, and abnormal cell states. The progression of cells in a given cellular process is rarely synchronized. Capturing this asynchrony poses a unique challenge of deciphering the sequence of regulatory events. TI takes into account a snapshot view of these events and uses computational techniques to infer the order of the cells along their developmental trajectories. The derived trajectory topologies can be linear, bifurcating, multifurcating, complex tree structures, or graph structures [121]. TI is mostly used for cell-lineage construction. During cellular developmental stages, cells express unique cellular markers and various lineage marks that can serve helpful for tracing lineage along pseudotime trajectories, such as somatic mutations, single nucleotide polymorphisms, copy number variants, microsatellites, transposons, and retroviral sequences.

Two broad strategies used for TI are DR-based methods (Monocle [122], Wishbone [123]), which use the reduced latent space as the first phase in inference and assigns pseudotime to individual cells, or clustering-based methods (TSCAN [110], SCUBA [102], ÉCLAIR [124]) which builds a network connecting clusters and applies pseudo-temporal ordering of clusters [125]. Monocle is a pioneering pseudo-temporal ordering algorithm to demonstrate how pseudotime analysis can reveal important cellular regulatory interactions [122]. It uses an unsupervised learning approach that constructs minimum spanning trees (MST) for ordering cells along pseudotime. Several other algorithms like Wishbone, TSCAN, versions of Monocle (currently Monocle 3) have been developed that are more robust and accurate, and a detailed comparative account can be found in Saelens et al. 2018 [121]. While most TI algorithms are unsupervised learning-based models, Ouija is a supervised learning-based algorithm that was developed keeping in mind that several confounding factors that affect biological processes like cell-cycle and apoptosis sometimes need to be accounted for to get biologically plausible pseudotime trajectories [126]. It uses switch-like marker genes and can be used as a complementary method with existing methods owing to its consideration of gene-specific behaviors as opposed to unsupervised methods. Despite the availability of more than 70 TI methods, researchers find it challenging to determine which is best suited for their analysis. Selection depends on the task like types of biological processes being studied, whether it is a cell differentiation process (Wishbone), lineage trees (MerLOT [127]), cell cycle (Cyclone [128], reCAT [91]), or downstream analysis, like inferring

GRN (SCODE [129]), DE. Each trajectory method has its pros and cons, and there is a lack of standardization. Dynverse is a collection of R packages specifically designed to address this issue so that researchers can perform TI, quantify it, or compare it to other available methods to decide on the best approach for their dataset [121]. Recent advancements have led to the use of time-series data in place of snapshot data. Tempora is an upcoming algorithm that may prove to be more biologically relevant than previous methods as it uses biological pathway information and identifies time-dependent pathways for ordering and inferring time-series data [130].

Undoubtedly, TI is becoming a popular tool for studying biological processes like cellular differentiation, immune response, tumor progression, and resistance, but inferring trajectories alone cannot be reliable as it needs to be validated using supporting biological evidence. With better algorithms and more time-series data availability, TI can be used to predict pre-disease state of cells, which may help in the early detection of diseases.

4.2. Gene-level analysis

Gene-level analysis is an integral part of cell-level analysis for studying cellular structures and identities, but independently gene-level analysis of single cells reveals a more comprehensive inference of cellular pathways and regulatory networks. It involves DE analysis, pathway analysis, gene regulatory networks (GRN), and gene set analysis. We shall be discussing some of the important gene-level analysis and the information they have revealed.

4.2.1. Differential Expression (DE)

We have discussed DE testing in previous sections while discussing clustering, but at the gene-level, we focus more on the stochastic nature of gene expression, and distinctive signatures only observed at the single-cell level. Although principally the same as bulk DE analysis, DE for single-cell data was developed to deal with artifacts inherent to scRNA-seq such as multimodality, dropouts, and heterogeneity. Among popular DE tools for scRNA-seq, MAST uses a linear hurdle model to account for confounders, and DE is determined using the likelihood ratio test [115]. SCDE uses a Bayesian approach that incorporates an evidence-based approach to evaluate the likelihood of the average level of gene expression for individual cells and measure the fold changes [116]. These techniques have shown higher sensitivity to other techniques. In some cases, bulk DE methods, when used with gene-weights, exceeds performance on scRNA-seq specific DE methods but at the price of being computationally intensive. Apart from having low detection accuracy for true DE genes, there is also a lack of agreement between various available tools. This demands better tools that account for the multimodality of scRNA-seq data, its artifacts, and identifies true DE genes having biological relevance. Wang, Tianyu et al. performed a recent comparative analysis of DE tools that could guide researchers to evaluate DE tools, choose appropriate ones for their analysis, and improve upon existing techniques [131]. Trajectory-based DE methods have also been developed called tradeSeq, enabling DE analysis between-lineage and within-lineage, providing a continuous resolution of gene expression changes through a dynamic process [132].

DE studies allow us to identify distinct expression profiles of cellular pathways that help us understand the effect of perturbations and the underlying mechanisms of disease pathologies.

4.2.2. Gene Regulatory Network (GRN)

The above discussion on gene expression poses another question on how gene expression is regulated in cells. Stochastic gene expression observed amongst single-cells indicates that gene regulation that relies on transcription factors, signaling molecule, and co-factors is regulated in a specific way. Uncovering these GRNs will reveal the basis of gene expression stochasticity and provide mechanistic insights into normal and abnormal cellular phenotypes. Several tools have been developed

Table 3
Current and future applications of scRNA-seq analysis in major fields of biomedical sciences.[a]

Field of study	Scope of analysis	Applications
Immunology	(i) Clustering of regional immune cells (ii) Trajectory analysis of individual immune lineages	Identification of novel immune cell subtypes [137], revealing immune microenvironment across tissues [138], understanding regional immunity in tumors [139]. Build developmental trajectory of immune cells, and gain mechanistic insights [140]. Immunology studies with the aid of SCT will help us develop better and targeted immunotherapies.
Cancer biology	(i) Clustering of tumor microenvironment (ii) DE analysis of tumors (iii) Construction of gene regulatory maps	Researchers and clinicians have struggled with understanding tumor heterogeneity for a long time. scRNA-seq has revealed intra-tumor, inter-tumor heterogeneity, and rare tumor subpopulations [141]. It will help understand cell-cell interactions in the tumor ecosystem, tumor resistance, refractory, and recurrence mechanisms. It can help elucidate genetic and non-genetic mechanisms for cancer and help devise better treatment regimes.
Cell-cell communication studies	(i) Integrating the count matric generated from scRNA-seq with known ligand-receptor interaction matrix (ii) Construction of GRNs	Deciphering cell-cell interactions is crucial to understanding both cellular development and diseases. SCT has enabled the inference of ligand-receptor interactions at an unprecedented resolution. Employing SCT we can identify communication patterns and use them to predict functions of poorly studied pathways. Tools like NATMI [142] is being used to identify which cell-type pairs or cellular communities communicate more frequently or specifically, what ligand-receptor pairs are the most active within a network, and has offered insights into autocrine signaling in cell-cell communication.
Stem cell	(i) Trajectory analysis of stem cells (ii) DE analysis of progenitor cells	When used to construct hematopoietic lineage trees, SCT analysis revealed that the differentiation process is continuous instead of the traditional belief that it is a stepwise process [143]. Also helped reveal a novel pathway used by stem-cells for self-renewal [144]. Another group of researchers developed a scRNA-seq based CRISPER interference technique to

(continued on next page)

Table 3 (*continued*)

Field of study	Scope of analysis	Applications
		study transcription factors involved in human endoderm development that revealed underlying factors and effects of perturbation [145]. It can help devise novel methods for treating genetic diseases and developmental disorders and better stem-cell therapies.
Neurobiology	(i) Clustering of regional brain cells (ii) Pseudo temporal ordering (iii) DE analysis of partitioned brain functions	It has been used to identify neuronal subtypes [146], understanding cellular programs involved in early development, cell populations in neuronal diseases, tracking transcriptional landscape of aging, resolving regional cell type landscape, and other nuances of brain function [147]. scRNA-seq has led to rapid accumulation of normal and tumor brain cell data, which can be further used to understand brain function and diseases.
Infection biology	(i) Classification of cellular and viral transcriptomes (ii) Genome profiling of host cell during infection	The recent COVID-19 pandemic has drawn all the attention to infectious diseases and host-pathogen interactions. It has also made us highly aware of the lack of robust techniques to understand infection mechanisms and devise treatment strategies. scRNA-seq analysis can help uncover host-pathogen relationships, map infection progression, and help identify druggable targets. The study of human microbiome interaction with immune cells will indicate the pathological state that develops when it is perturbed [148,149].
Diagnostics	(i) DE analysis and Gene regulatory network construction for identification of novel diagnostic biomarkers.	scRNA-seq data is becoming increasingly available for pre-clinical and clinical samples of diseases. It enables the construction of high-resolution cellular maps for diseases, helps identify novel biomarkers, and unravels underlying disease mechanisms that aids in developing better treatment regimes. Used in Cerebrospinal Fluid Research and diagnostics [150], a molecular diagnostic test using scRNA-seq analysis identified two gene-sets involved in the autoimmune response that is suggestive of disease progression and could drive lupus nephritis treatment [151]. Researchers successfully identified and validated diagnostic and therapeutic biomarkers for rheumatoid arthritis for mice and humans, using scRNA-seq analysis and network tools [152]. In the future, it may also help in identifying dynamical network biomarkers that will help predict a pre-disease state.
Precision Medicine	(i) Longitudinal profiling with scRNA-seq. (ii) Pre- and post-treatment analysis	scRNA-seq helps identify the more notorious clones or malignant cells in cancer, thus tailored therapy could be designed to target this particular group of cells. Profiling patient samples can help monitor disease states, response to therapies, and mechanisms of resistance [153,154]. However, for this ambitious idea to become a reality, it would require more cost-effective and reliable analysis techniques than currently available.

[a] Applications in different fields are not mutually exclusive.

for inferring GRN, some derived from bulk analysis methods and some designed explicitly for scRNA-seq data, but GRNs are incredibly complex to decipher. The SCENIC algorithm simultaneously constructs gene regulatory networks and performs clustering [133]. It bases the identification of stable cellular states on the activity of GRNs in each cell and is well suited for the discovery of cell states that are driven by transcription factors and cis-regulatory sequences. SCGRN is a supervised feature learning-based approach for inferring GRNs that employs three different ML techniques [134], is promising against unsupervised learning approaches used earlier. Another study by Qin et al. presents a toolkit, Scribe, for inferring causal GRNs from single-cell datasets that indicate that pseudotime data perform poorly compared to true time-series data [136]. Such insights will encourage newer studies that focus on deriving true biological insights from scRNA-seq data. Recently, researchers explored GRN inference algorithms and developed a framework called BEELINE, where they used synthetic networks with predictable trajectories, literature curated Boolean models, and diverse transcriptional regulatory networks to assess the accuracy of GRN methods [135]. This study can be used as a benchmark for selecting GRN algorithms or for developing new strategies.

The study of GRNs at single-cell resolution will contribute significantly to system biology approaches and help build more precise models. These models can further help discover dynamical network biomarkers that can predict disease and pre-disease states.

5. Applications and future prospects

The past decade has seen a momentous upsurge in SCT and its applications across neurology, microbiology, cell biology, molecular biology, immunology, cancer biology, bioinformatics, stem cell, biomedical sciences, and clinical and diagnostic applications. Recent research efforts evince its potential to be of use in translational research. To list all the applications is beyond the scope of this article. We list some of the major areas that have benefited from SCT are listed in Table 3.

scRNA transcriptomics can be used with multiple other technologies

to yield more comprehensive results for understanding cellular biology. Spatial transcriptomics is one such technology that preserves the spatial location of gene expression in cells during analysis, and when used with scRNA-seq, is an excellent way to study tissue microenvironment. The possibilities of what can be achieved with this technology are countless.

6. Conclusion

Single-cell RNA sequencing technology has zoomed in on cellular biology like never before. This review objectively points out that single-cell analysis is a multi-step process, and no one step is exclusive of the other. All the steps, only when carefully monitored in tandem, will give precise results. Leveraging SCT and multiple single-cell modalities at once bears a remarkable ingenuity in understanding complex cellular processes, capturing cellular heterogeneity, and disease states. It opens new frontiers of research. Efforts to characterize all cells in a human body, such as the Chan-Zuckerberg Initiative- Human Cell Atlas, serve as major reservoirs for researchers across several fields from biological sciences to computational sciences to come together and develop from. There is an increasing demand for better tools, techniques, analysis algorithms, and experimental validation measures that can rapidly materialize the vision of understanding biological processes at a single-cell resolution.

Declaration of Competing Interest

The authors do not have any conflicts of interests to declare.

Acknowledgments

This work was supported by the project “Genetic Analysis of Dermatological Disorders” (BT/PR5402/BID/7/408/2012 dated: 6/7/2017), Department of Biotechnology, Government of India.

References

[1] P. Angerer, L. Simon, S. Tritschler, F.A. Wolf, D. Fischer, F.J. Theis, Single cells make big data: new challenges and opportunities in transcriptomics, Curr. Opin. Syst. Biol. (2017), https://doi.org/10.1016/j.coisb.2017.07.004.
[2] B. Hwang, J.H. Lee, D. Bang, Single-cell RNA sequencing technologies and bioinformatics pipelines, Exp. Mol. Med. (2018), https://doi.org/10.1038/s12276-018-0071-8.
[3] S. Huang, Non-genetic heterogeneity of cells in development: more than just noise, Development (2009), https://doi.org/10.1242/dev.035139.
[4] N. Li, H. Clevers, Coexistence of quiescent and active adult stem cells in mammals, Science 80 (2010), https://doi.org/10.1126/science.1180794.
[5] B.D. Aevermann, M. Novotny, T. Bakken, J.A. Miller, A.D. Diehl, D. Osumi-Sutherland, R.S. Lasken, E.S. Lein, R.H. Scheuermann, Cell type discovery using single-cell transcriptomics: implications for ontological representation, Hum. Mol. Genet. (2018), https://doi.org/10.1093/hmg/ddy100.
[6] S. Linnarsson, S.A. Teichmann, Single-cell genomics: Coming of age, Genome Biol. (2016), https://doi.org/10.1186/s13059-016-0960-x.
[7] E. Shapiro, T. Biezuner, S. Linnarsson, Single-cell sequencing-based technologies will revolutionize whole-organism science, Nat. Rev. Genet. (2013), https://doi.org/10.1038/nrg3542.
[8] P. Angerer, L. Simon, S. Tritschler, F.A. Wolf, D. Fischer, F.J. Theis, Single cells make big data: new challenges and opportunities in transcriptomics, Curr. Opin. Syst. Biol. 4 (2017) 85–91, https://doi.org/10.1016/j.coisb.2017.07.004.
[9] G. Brady, M. Barbara, N.N. Iscove, Representative in vitro cDNA amplification from individual hemopoietic cells and colonies, Methods Mol. Cell. Biol. 2 (1990) 17–25, 08987750.
[10] J. Eberwine, H. Yeh, K. Miyashiro, Y. Cao, S. Nair, R. Finnell, M. Zettel, P. Coleman, Analysis of gene expression in single live neurons, Proc. Natl. Acad. Sci. U. S. A. (1992), https://doi.org/10.1073/pnas.89.7.3010.
[11] F. Tang, K. Lao, M.A. Surani, Development and applications of single-cell transcriptome analysis, Nat. Methods (2011), https://doi.org/10.1038/nmeth.1557.
[12] F. Tang, C. Barbacioru, Y. Wang, E. Nordman, C. Lee, N. Xu, X. Wang, J. Bodeau, B.B. Tuch, A. Siddiqui, K. Lao, M.A. Surani, mRNA-Seq whole-transcriptome analysis of a single cell, Nat. Methods (2009), https://doi.org/10.1038/nmeth.1315.
[13] D. Hebenstreit, Methods, challenges and potentials of single cell RNA-seq, Biology (Basel) (2012), https://doi.org/10.3390/biology1030658.
[14] D. Grün, A. Lyubimova, L. Kester, K. Wiebrands, O. Basak, N. Sasaki, H. Clevers, A. Van Oudenaarden, Single-cell messenger RNA sequencing reveals rare intestinal cell types, Nature (2015), https://doi.org/10.1038/nature14966.
[15] S. Petropoulos, D. Edsgärd, B. Reinius, Q. Deng, S.P. Panula, S. Codeluppi, A. Plaza Reyes, S. Linnarsson, R. Sandberg, F. Lanner, Single-cell RNA-seq reveals lineage and x chromosome dynamics in human preimplantation embryos, Cell (2016), https://doi.org/10.1016/j.cell.2016.03.023.
[16] A.A. Tu, T.M. Gierahn, B. Monian, D.M. Morgan, N.K. Mehta, B. Ruiter, W. G. Shreffler, A.K. Shalek, J.C. Love, TCR sequencing paired with massively parallel 3′ RNA-seq reveals clonotypic T cell signatures, Nat. Immunol. (2019), https://doi.org/10.1038/s41590-019-0544-5.
[17] R.J. Miragaia, T. Gomes, A. Chomka, L. Jardine, A. Riedel, A.N. Hegazy, N. Whibley, A. Tucci, X. Chen, I. Lindeman, G. Emerton, T. Krausgruber, J. Shields, M. Haniffa, F. Powrie, S.A. Teichmann, Single-cell transcriptomics of regulatory T cells reveals trajectories of tissue adaptation, Immunity (2019), https://doi.org/10.1016/j.immuni.2019.01.001.
[18] M.J.T. Stubbington, T. Lönnberg, V. Proserpio, S. Clare, A.O. Speak, G. Dougan, S. A. Teichmann, T cell fate and clonality inference from single-cell transcriptomes, Nat. Methods (2016), https://doi.org/10.1038/nmeth.3800.
[19] A.K. Shalek, R. Satija, X. Adiconis, R.S. Gertner, J.T. Gaublomme, R. Raychowdhury, S. Schwartz, N. Yosef, C. Malboeuf, D. Lu, J.J. Trombetta, D. Gennert, A. Gnirke, A. Goren, N. Hacohen, J.Z. Levin, H. Park, A. Regev, Single-cell transcriptomics reveals bimodality in expression and splicing in immune cells, Nature (2013), https://doi.org/10.1038/nature12172.
[20] J. Wagner, M.A. Rapsomaniki, S. Chevrier, T. Anzeneder, C. Langwieder, A. Dykgers, M. Rees, A. Ramaswamy, S. Muenst, S.D. Soysal, A. Jacobs, J. Windhager, K. Silina, M. van den Broek, K.J. Dedes, M. Rodríguez Martínez, W. P. Weber, B. Bodenmiller, A single-cell atlas of the tumor and immune ecosystem of human breast cancer, Cell (2019), https://doi.org/10.1016/j.cell.2019.03.005.
[21] J.M. Granja, S. Klemm, L.M. McGinnis, A.S. Kathiria, A. Mezger, M.R. Corces, B. Parks, E. Gars, M. Liedtke, G.X.Y. Zheng, H.Y. Chang, R. Majeti, W.J. Greenleaf, Single-cell multiomic analysis identifies regulatory programs in mixed-phenotype acute leukemia, Nat. Biotechnol. (2019), https://doi.org/10.1038/s41587-019-0332-7.
[22] C. Yao, H.W. Sun, N.E. Lacey, Y. Ji, E.A. Moseman, H.Y. Shih, E.F. Heuston, M. Kirby, S. Anderson, J. Cheng, O. Khan, R. Handon, J. Reilley, J. Fioravanti, J. Hu, S. Gossa, E.J. Wherry, L. Gattinoni, D.B. McGavern, J.J. O'Shea, P. L. Schwartzberg, T. Wu, Single-cell RNA-seq reveals TOX as a key regulator of CD8+ T cell persistence in chronic infection, Nat. Immunol. (2019), https://doi.org/10.1038/s41590-019-0403-4.
[23] S.M. Shaffer, M.C. Dunagin, S.R. Torborg, E.A. Torre, B. Emert, C. Krepler, M. Beqiri, K. Sproesser, P.A. Brafford, M. Xiao, E. Eggan, I.N. Anastopoulos, C. A. Vargas-Garcia, A. Singh, K.L. Nathanson, M. Herlyn, A. Raj, Rare cell variability and drug-induced reprogramming as a mode of cancer drug resistance, Nature (2017), https://doi.org/10.1038/nature22794.
[24] P. Yu, W. Lin, Single-cell transcriptome study as big data, Genom. Proteome. Bioinform. (2016), https://doi.org/10.1016/j.gpb.2016.01.005.
[25] J. Zheng, K. Wang, Emerging deep learning methods for single-cell RNA-seq data analysis, Quant. Biol. (2019), https://doi.org/10.1007/s40484-019-0189-2.
[26] R. Petegrosso, Z. Li, R. Kuang, Machine learning and statistical methods for clustering single-cell RNA-sequencing data, Brief. Bioinform. (2019), https://doi.org/10.1093/bib/bbz063.
[27] B. Vieth, S. Parekh, C. Ziegenhain, W. Enard, I. Hellmann, A systematic evaluation of single cell RNA-seq analysis pipelines, Nat. Commun. (2019), https://doi.org/10.1038/s41467-019-12266-7.
[28] G. Chen, B. Ning, T. Shi, Single-cell RNA-seq technologies and related computational data analysis, Front. Genet. (2019), https://doi.org/10.3389/fgene.2019.00317.
[29] K. Sheng, W. Cao, Y. Niu, Q. Deng, C. Zong, Effective detection of variation in single-cell transcriptomes using MATQ-seq, Nat. Methods (2017), https://doi.org/10.1038/nmeth.4145.
[30] S. Picelli, Å.K. Björklund, O.R. Faridani, S. Sagasser, G. Winberg, R. Sandberg, Smart-seq2 for sensitive full-length transcriptome profiling in single cells, Nat. Methods (2013), https://doi.org/10.1038/nmeth.2639.
[31] L.D. Goldstein, Y.J.J. Chen, J. Dunne, A. Mir, H. Hubschle, J. Guillory, W. Yuan, J. Zhang, J. Stinson, B. Jaiswal, K.B. Pahuja, I. Mann, T. Schaal, L. Chan, S. Anandakrishnan, C. Wah Lin, P. Espinoza, S. Husain, H. Shapiro, K. Swaminathan, S. Wei, M. Srinivasan, S. Seshagiri, Z. Modrusan, Massively parallel nanowell-based single-cell gene expression profiling, BMC Genomics (2017), https://doi.org/10.1186/s12864-017-3893-1.
[32] X. Fan, X. Zhang, X. Wu, H. Guo, Y. Hu, F. Tang, Y. Huang, Single-cell RNA-seq transcriptome analysis of linear and circular RNAs in mouse preimplantation embryos, Genome Biol. (2015), https://doi.org/10.1186/s13059-015-0706-1.
[33] S. Islam, U. Kjällquist, A. Moliner, P. Zajac, J.B. Fan, P. Lönnerberg, S. Linnarsson, Characterization of the single-cell transcriptional landscape by highly multiplex RNA-seq, Genome Res. (2011), https://doi.org/10.1101/gr.110882.110.
[34] S. Islam, U. Kjällquist, A. Moliner, P. Zajac, J.B. Fan, P. Lönnerberg, S. Linnarsson, Highly multiplexed and strand-specific single-cell RNA 5′ end sequencing, Nat. Protoc. (2012), https://doi.org/10.1038/nprot.2012.022.
[35] G.X.Y. Zheng, J.M. Terry, P. Belgrader, P. Ryvkin, Z.W. Bent, R. Wilson, S. B. Ziraldo, T.D. Wheeler, G.P. McDermott, J. Zhu, M.T. Gregory, J. Shuga, L. Montesclaros, J.G. Underwood, D.A. Masquelier, S.Y. Nishimura, M. Schnall-Levin, P.W. Wyatt, C.M. Hindson, R. Bharadwaj, A. Wong, K.D. Ness, L.W. Beppu, H.J. Deeg, C. McFarland, K.R. Loeb, W.J. Valente, N.G. Ericson, E.A. Stevens, J. P. Radich, T.S. Mikkelsen, B.J. Hindson, J.H. Bielas, Massively parallel digital

transcriptional profiling of single cells, Nat. Commun. (2017), https://doi.org/10.1038/ncomms14049.
[36] D.M. DeLaughter, The use of the fluidigm C1 for RNA expression analyses of single cells, Curr. Protoc. Mol. Biol. (2018), https://doi.org/10.1002/cpmb.55.
[37] E.Z. Macosko, A. Basu, R. Satija, J. Nemesh, K. Shekhar, M. Goldman, I. Tirosh, A. R. Bialas, N. Kamitaki, E.M. Martersteck, J.J. Trombetta, D.A. Weitz, J.R. Sanes, A.K. Shalek, A. Regev, S.A. McCarroll, Highly parallel genome-wide expression profiling of individual cells using nanoliter droplets, Cell (2015), https://doi.org/10.1016/j.cell.2015.05.002.
[38] A.M. Klein, L. Mazutis, I. Akartuna, N. Tallapragada, A. Veres, V. Li, L. Peshkin, D. A. Weitz, M.W. Kirschner, Droplet barcoding for single-cell transcriptomics applied to embryonic stem cells, Cell (2015), https://doi.org/10.1016/j.cell.2015.04.044.
[39] P. Hu, W. Zhang, H. Xin, G. Deng, Single cell isolation and analysis, Front. Cell Dev. Biol. (2016), https://doi.org/10.3389/fcell.2016.00116.
[40] A.E. Saliba, A.J. Westermann, S.A. Gorski, J. Vogel, Single-cell RNA-seq: advances and future challenges, Nucleic Acids Res. (2014), https://doi.org/10.1093/nar/gku555.
[41] V. Menon, Clustering single cells: a review of approaches on high-and low-depth single-cell RNA-seq data, Brief. Funct. Genomics (2018), https://doi.org/10.1093/bfgp/elx044.
[42] M.S. Cembrowski, Single-cell transcriptomics as a framework and roadmap for understanding the brain, J. Neurosci. Methods (2019), https://doi.org/10.1016/j.jneumeth.2019.108353.
[43] D.B. Weibel, G.M. Whitesides, Applications of microfluidics in chemical biology, Curr. Opin. Chem. Biol. (2006), https://doi.org/10.1016/j.cbpa.2006.10.016.
[44] J.S. Marcus, W.F. Anderson, S.R. Quake, Microfluidic single-cell mRNA isolation and analysis, Anal. Chem. (2006), https://doi.org/10.1021/ac0519460.
[45] X. Chen, S.A. Teichmann, K.B. Meyer, From tissues to cell types and back: single-cell gene expression analysis of tissue architecture, Annu. Rev. Biomed. Data Sci. (2018), https://doi.org/10.1146/annurev-biodatasci-080917-013452.
[46] C. Ziegenhain, B. Vieth, S. Parekh, B. Reinius, A. Guillaumet-Adkins, M. Smets, H. Leonhardt, H. Heyn, I. Hellmann, W. Enard, Comparative analysis of single-cell RNA sequencing methods, Mol. Cell (2017), https://doi.org/10.1016/j.molcel.2017.01.023.
[47] V. Espina, J.D. Wulfkuhle, V.S. Calvert, A. VanMeter, W. Zhou, G. Coukos, D. H. Geho, E.F. Petricoin, L.A. Liotta, Laser-capture microdissection, Nat. Protoc. (2006), https://doi.org/10.1038/nprot.2006.85.
[48] S. Nichterwitz, G. Chen, J. Aguila Benitez, M. Yilmaz, H. Storvall, M. Cao, R. Sandberg, Q. Deng, E. Hedlund, Laser capture microscopy coupled with smart-seq2 for precise spatial transcriptomic profiling, Nat. Commun. (2016), https://doi.org/10.1038/ncomms12139.
[49] T.M. Gierahn, M.H. Wadsworth, T.K. Hughes, B.D. Bryson, A. Butler, R. Satija, S. Fortune, J. Christopher Love, A.K. Shalek, Seq-well: portable, low-cost rna sequencing of single cells at high throughput, Nat. Methods 14 (2017) 395–398, https://doi.org/10.1038/nmeth.4179.
[50] J. Cao, J.S. Packer, V. Ramani, D.A. Cusanovich, C. Huynh, R. Daza, X. Qiu, C. Lee, S.N. Furlan, F.J. Steemers, A. Adey, R.H. Waterston, C. Trapnell, J. Shendure, Comprehensive single-cell transcriptional profiling of a multicellular organism, Science 80 (2017), https://doi.org/10.1126/science.aam8940.
[51] A.B. Rosenberg, C.M. Roco, R.A. Muscat, A. Kuchina, P. Sample, Z. Yao, L. T. Graybuck, D.J. Peeler, S. Mukherjee, W. Chen, S.H. Pun, D.L. Sellers, B. Tasic, G. Seelig, Single-cell profiling of the developing mouse brain and spinal cord with split-pool barcoding, Science 80 (360) (2018) 176–182, https://doi.org/10.1126/science.aam8999.
[52] B. Hwang, J.H. Lee, D. Bang, Single-cell RNA sequencing technologies and bioinformatics pipelines, Exp. Mol. Med. 50 (2018), https://doi.org/10.1038/s12276-018-0071-8.
[53] T. Kivioja, A. Vähärautio, K. Karlsson, M. Bonke, M. Enge, S. Linnarsson, J. Taipale, Counting absolute numbers of molecules using unique molecular identifiers, Nat. Methods (2012), https://doi.org/10.1038/nmeth.1778.
[54] S. Islam, A. Zeisel, S. Joost, G. La Manno, P. Zajac, M. Kasper, P. Lönnerberg, S. Linnarsson, Quantitative single-cell RNA-seq with unique molecular identifiers, Nat. Methods (2014), https://doi.org/10.1038/nmeth.2772.
[55] M.D. Luecken, F.J. Theis, Current best practices in single-cell RNA-seq analysis: a tutorial, Mol. Syst. Biol. (2019), https://doi.org/10.15252/msb.20188746.
[56] O. Stegle, S.A. Teichmann, J.C. Marioni, Computational and analytical challenges in single-cell transcriptomics, Nat. Rev. Genet. (2015), https://doi.org/10.1038/nrg3833.
[57] T. Ilicic, J.K. Kim, A.A. Kolodziejczyk, F.O. Bagger, D.J. McCarthy, J.C. Marioni, S.A. Teichmann, Classification of low quality cells from single-cell RNA-seq data, Genome Biol. (2016), https://doi.org/10.1186/s13059-016-0888-1.
[58] P. Brennecke, S. Anders, J.K. Kim, A.A. Kołodziejczyk, X. Zhang, V. Proserpio, B. Baying, V. Benes, S.A. Teichmann, J.C. Marioni, M.G. Heisler, Accounting for technical noise in single-cell RNA-seq experiments, Nat. Methods (2013), https://doi.org/10.1038/nmeth.2645.
[59] Chapter 6 Quality Control, Orchestrating Single-Cell Analysis with Bioconductor, (s.d.), https://osca.bioconductor.org/quality-control.html (accedit 6 agost 2020), 2020.
[60] C.A. Vallejos, D. Risso, A. Scialdone, S. Dudoit, J.C. Marioni, Normalizing single-cell RNA sequencing data: challenges and opportunities, Nat. Methods (2017), https://doi.org/10.1038/nmeth.4292.
[61] L. Wen, F. Tang, How to catch rare cell types, Nature (2015), https://doi.org/10.1038/nature15204.
[62] C. Hafemeister, R. Satija, Normalization and variance stabilization of single-cell RNA-seq data using regularized negative binomial regression, Genome Biol. (2019), https://doi.org/10.1186/s13059-019-1874-1.
[63] R. Bacher, L.F. Chu, N. Leng, A.P. Gasch, J.A. Thomson, R.M. Stewart, M. Newton, C. Kendziorski, SCnorm: robust normalization of single-cell RNA-seq data, Nat. Methods (2017), https://doi.org/10.1038/nmeth.4263.
[64] W.V. Li, J.J. Li, An accurate and robust imputation method scImpute for single-cell RNA-seq data, Nat. Commun. (2018), https://doi.org/10.1038/s41467-018-03405-7.
[65] D. van Dijk, R. Sharma, J. Nainys, K. Yim, P. Kathail, A.J. Carr, C. Burdziak, K. R. Moon, C.L. Chaffer, D. Pattabiraman, B. Bierie, L. Mazutis, G. Wolf, S. Krishnaswamy, D. Pe'er, Recovering gene interactions from single-cell data using data diffusion, Cell (2018), https://doi.org/10.1016/j.cell.2018.05.061.
[66] W. Gong, I.Y. Kwak, P. Pota, N. Koyano-Nakagawa, D.J. Garry, DrImpute: imputing dropout events in single cell RNA sequencing data, BMC Bioinformatics (2018), https://doi.org/10.1186/s12859-018-2226-y.
[67] M. Huang, J. Wang, E. Torre, H. Dueck, S. Shaffer, R. Bonasio, J.I. Murray, A. Raj, M. Li, N.R. Zhang, SAVER: gene expression recovery for single-cell RNA sequencing, Nat. Methods (2018), https://doi.org/10.1038/s41592-018-0033-z.
[68] S. Tracy, G.C. Yuan, R. Dries, RESCUE: imputing dropout events in single-cell RNA-sequencing data, BMC Bioinformatics (2019), https://doi.org/10.1186/s12859-019-2977-0.
[69] T. Peng, Q. Zhu, P. Yin, K. Tan, SCRABBLE: Single-cell RNA-seq imputation constrained by bulk RNA-seq data, Genome Biol. (2019), https://doi.org/10.1186/s13059-019-1681-8.
[70] P. Ye, W. Ye, C. Ye, S. Li, L. Ye, G. Ji, X. Wu, scHinter: imputing dropout events for single-cell RNA-seq data with limited sample size, Bioinformatics (2020), https://doi.org/10.1093/bioinformatics/btz627.
[71] P. Qiu, Embracing the dropouts in single-cell RNA-seq data, bioRxiv (2018), https://doi.org/10.1101/468025.
[72] V. Svensson, Droplet scRNA-seq is not zero-inflated, Nat. Biotechnol. (2020), https://doi.org/10.1038/s41587-019-0379-5.
[73] W. Tang, F. Bertaux, P. Thomas, C. Stefanelli, M. Saint, S. Marguerat, V. Shahrezaei, BayNorm: bayesian gene expression recovery, imputation and normalization for single-cell RNA-sequencing data, Bioinformatics (2020), https://doi.org/10.1093/bioinformatics/btz726.
[74] A.T.L. Lun, J.C. Marioni, Overcoming confounding plate effects in differential expression analyses of single-cell RNA-seq data, Biostatistics (2017), https://doi.org/10.1093/biostatistics/kxw055.
[75] M.B. Cole, D. Risso, A. Wagner, D. DeTomaso, J. Ngai, E. Purdom, S. Dudoit, N. Yosef, Performance assessment and selection of normalization procedures for single-cell RNA-seq, Cell Syst. (2019), https://doi.org/10.1016/j.cels.2019.03.010.
[76] P.Y. Tung, J.D. Blischak, C.J. Hsiao, D.A. Knowles, J.E. Burnett, J.K. Pritchard, Y. Gilad, Batch effects and the effective design of single-cell gene expression studies, Sci. Rep. (2017), https://doi.org/10.1038/srep39921.
[77] W. Chen, S. Zhang, J. Williams, B. Ju, B. Shaner, J. Easton, G. Wu, X. Chen, A comparison of methods accounting for batch effects in differential expression analysis of UMI count based single cell RNA sequencing, Comput. Struct. Biotechnol. J. (2020), https://doi.org/10.1016/j.csbj.2020.03.026.
[78] R. Satija, J.A. Farrell, D. Gennert, A.F. Schier, A. Regev, Spatial reconstruction of single-cell gene expression data, Nat. Biotechnol. (2015), https://doi.org/10.1038/nbt.3192.
[79] F.A. Wolf, P. Angerer, F.J. Theis, SCANPY: large-scale single-cell gene expression data analysis, Genome Biol. (2018), https://doi.org/10.1186/s13059-017-1382-0.
[80] S.R. Tyler, P.G. Rotti, X. Sun, Y. Yi, W. Xie, M.C. Winter, M.J. Flamme-Wiese, B. A. Tucker, R.F. Mullins, A.W. Norris, J.F. Engelhardt, PyMINEr finds gene and autocrine-paracrine networks from human islet scRNA-seq, Cell Rep. (2019), https://doi.org/10.1016/j.celrep.2019.01.063.
[81] V. Petukhov, J. Guo, N. Baryawno, N. Severe, D.T. Scadden, M.G. Samsonova, P. V. Kharchenko, dropEst: pipeline for accurate estimation of molecular counts in droplet-based single-cell RNA-seq experiments, Genome Biol. (2018), https://doi.org/10.1186/s13059-018-1449-6.
[82] R. Hillje, P.G. Pelicci, L. Luzi, Cerebro: interactive visualization of scRNA-seq data, Bioinformatics (2019), https://doi.org/10.1093/bioinformatics/btz877.
[83] K. Rue-Albrecht, F. Marini, C. Soneson, A.T.L. Lun, iSEE: interactive summarizedexperiment explorer, F1000Research (2018), https://doi.org/10.12688/f1000research.14966.1.
[84] C. Arisdakessian, O. Poirion, B. Yunits, X. Zhu, L.X. Garmire, DeepImpute: an accurate, fast, and scalable deep neural network method to impute single-cell RNA-seq data, Genome Biol. (2019), https://doi.org/10.1186/s13059-019-1837-6.
[85] T. Wang, T.S. Johnson, W. Shao, Z. Lu, B.R. Helm, J. Zhang, K. Huang, BERMUDA: a novel deep transfer learning method for single-cell RNA sequencing batch correction reveals hidden high-resolution cellular subtypes, Genome Biol. (2019), https://doi.org/10.1186/s13059-019-1764-6.
[86] L. Haghverdi, A.T.L. Lun, M.D. Morgan, J.C. Marioni, Batch effects in single-cell RNA-sequencing data are corrected by matching mutual nearest neighbors, Nat. Biotechnol. (2018), https://doi.org/10.1038/nbt.4091.
[87] E. Lin, S. Mukherjee, S. Kannan, A deep adversarial variational autoencoder model for dimensionality reduction in single-cell RNA sequencing analysis, BMC Bioinformatics 21 (2020) 64, https://doi.org/10.1186/s12859-020-3401-5.
[88] M. Amodio, D. van Dijk, K. Srinivasan, W.S. Chen, H. Mohsen, K.R. Moon, A. Campbell, Y. Zhao, X. Wang, M. Venkataswamy, A. Desai, V. Ravi, P. Kumar, R. Montgomery, G. Wolf, S. Krishnaswamy, Exploring single-cell data with deep

multitasking neural networks, Nat. Methods (2019), https://doi.org/10.1038/s41592-019-0576-7.
[89] V.Y. Kiselev, A. Yiu, M. Hemberg, Scmap: projection of single-cell RNA-seq data across data sets, Nat. Methods (2018), https://doi.org/10.1038/nmeth.4644.
[90] H. Todorov, R. Cannoodt, W. Saelens, Y. Saeys, TinGa: fast and flexible trajectory inference with growing neural gas, Bioinformatics (2020), https://doi.org/10.1093/bioinformatics/btaa463.
[91] Z. Liu, H. Lou, K. Xie, H. Wang, N. Chen, O.M. Aparicio, M.Q. Zhang, R. Jiang, T. Chen, Reconstructing cell cycle pseudo time-series via single-cell transcriptome data, Nat. Commun. (2017), https://doi.org/10.1038/s41467-017-00039-z.
[92] R. Petegrosso, Z. Li, R. Kuang, Machine learning and statistical methods for clustering single-cell RNA-sequencing data, Brief. Bioinform. (2019), https://doi.org/10.1093/bib/bbz063.
[93] P. Lin, M. Troup, J.W.K. Ho, CIDR: ultrafast and accurate clustering through imputation for single-cell RNA-seq data, Genome Biol. (2017), https://doi.org/10.1186/s13059-017-1188-0.
[94] A. Butler, P. Hoffman, P. Smibert, E. Papalexi, R. Satija, Integrating single-cell transcriptomic data across different conditions, technologies, and species, Nat. Biotechnol. (2018), https://doi.org/10.1038/nbt.4096.
[95] L. McInnes, J. Healy, N. Saul, L. Großberger, UMAP: uniform manifold approximation and projection, J. Open Source Softw. (2018), https://doi.org/10.21105/joss.00861.
[96] E. Pierson, C. Yau, ZIFA: dimensionality reduction for zero-inflated single-cell gene expression analysis, Genome Biol. (2015), https://doi.org/10.1186/s13059-015-0805-z.
[97] D. Risso, F. Perraudeau, S. Gribkova, S. Dudoit, J.P. Vert, A general and flexible method for signal extraction from single-cell RNA-seq data, Nat. Commun. (2018), https://doi.org/10.1038/s41467-017-02554-5.
[98] Data Portal, Human Cell Atlas, (s.d.), https://www.humancellatlas.org/data-portal/ (accedit 18 març 2020), 2020.
[99] Single Cell Portal, (s.d.). https://singlecell.broadinstitute.org/single_cell (accedit 18 març 2020). 2020.
[100] Home, < Single Cell Expression Atlas < EMBL-EBI (s.d.). https://www.ebi.ac.uk/gxa/sc/home (accedit 18 març 2020), 2020.
[101] Samples, PanglaoDB (s.d.). https://panglaodb.se/samples.html?species=human&protocol=all protocols&sort=mostrecent (accedit 18 març 2020), 2020.
[102] M. Eugenio, R.L. Karp, G. Guo, P. Robson, A.H. Hart, L. Trippa, G.C. Yuan, Bifurcation analysis of single-cell gene expression data reveals epigenetic landscape, Proc. Natl. Acad. Sci. U. S. A. (2014), https://doi.org/10.1073/pnas.1408993111.
[103] L. Yang, J. Liu, Q. Lu, A.D. Riggs, X. Wu, SAIC: an iterative clustering approach for analysis of single cell RNA-seq data, BMC Genomics (2017), https://doi.org/10.1186/s12864-017-4019-5.
[104] H. Zhang, C.A.A. Lee, Z. Li, J.R. Garbe, C.R. Eide, R. Petegrosso, R. Kuang, J. Tolar, A multitask clustering approach for single-cell RNA-seq analysis in recessive dystrophic epidermolysis bullosa, PLoS Comput. Biol. (2018), https://doi.org/10.1371/journal.pcbi.1006053.
[105] D.A. du Verle, S. Yotsukura, S. Nomura, H. Aburatani, K. Tsuda, CellTree: an R/bioconductor package to infer the hierarchical structure of cell populations from single-cell RNA-seq data, BMC Bioinformatics (2016), https://doi.org/10.1186/s12859-016-1175-6.
[106] J.M. Zhang, J. Fan, H.C. Fan, D. Rosenfeld, D.N. Tse, An interpretable framework for clustering single-cell RNA-seq datasets, BMC Bioinformatics (2018), https://doi.org/10.1186/s12859-018-2092-7.
[107] V.A. Traag, L. Waltman, N.J. van Eck, From Louvain to Leiden: guaranteeing well-connected communities, Sci. Rep. (2019), https://doi.org/10.1038/s41598-019-41695-z.
[108] C. Xu, Z. Su, Identification of cell types from single-cell transcriptomes using a novel clustering method, Bioinformatics (2015), https://doi.org/10.1093/bioinformatics/btv088.
[109] E. Azizi, A.J. Carr, G. Plitas, A.E. Cornish, C. Konopacki, S. Prabhakaran, J. Nainys, K. Wu, V. Kiseliovas, M. Setty, K. Choi, R.M. Fromme, P. Dao, P. T. McKenney, R.C. Wasti, K. Kadaveru, L. Mazutis, A.Y. Rudensky, D. Pe'er, Single-cell map of diverse immune phenotypes in the breast tumor microenvironment, Cell (2018), https://doi.org/10.1016/j.cell.2018.05.060.
[110] Z. Ji, H. Ji, TSCAN: pseudo-time reconstruction and evaluation in single-cell RNA-seq analysis, Nucleic Acids Res. (2016), https://doi.org/10.1093/nar/gkw430.
[111] X. Qiu, Q. Mao, Y. Tang, L. Wang, R. Chawla, H.A. Pliner, C. Trapnell, Reversed graph embedding resolves complex single-cell trajectories, Nat. Methods (2017), https://doi.org/10.1038/nmeth.4402.
[112] T. Tian, J. Wan, Q. Song, Z. Wei, Clustering single-cell RNA-seq data with a model-based deep learning approach, Nat. Mach. Intell. (2019), https://doi.org/10.1038/s42256-019-0037-0.
[113] Y. Yang, R. Huh, H.W. Culpepper, Y. Lin, M.I. Love, Y. Li, SAFE-clustering: single-cell aggregated (from Ensemble) clustering for single-cell RNA-seq data, Bioinformatics (2019), https://doi.org/10.1093/bioinformatics/bty793.
[114] A.W. Zhang, C.O. Flanagan, E.A. Chavez, J.L.P. Lim, N. Ceglia, A. Mcpherson, M. Wiens, P. Walters, T. Chan, B. Hewitson, D. Lai, A. Mottok, C. Sarkozy, L. Chong, T. Aoki, X. Wang, A.P. Weng, J.N. Mcalpine, S. Aparicio, C. Steidl, K. R. Campbell, S.P. Shah, RNA-seq for tumor microenvironment profiling, Nat. Methods (2019), https://doi.org/10.1038/s41592-019-0529-1.
[115] G. Finak, A. McDavid, M. Yajima, J. Deng, V. Gersuk, A.K. Shalek, C.K. Slichter, H. W. Miller, M.J. McElrath, M. Prlic, P.S. Linsley, R. Gottardo, MAST: a flexible statistical framework for assessing transcriptional changes and characterizing heterogeneity in single-cell RNA sequencing data, Genome Biol. (2015), https://doi.org/10.1186/s13059-015-0844-5.
[116] P.V. Kharchenko, L. Silberstein, D.T. Scadden, Bayesian approach to single-cell differential expression analysis, Nat. Methods (2014), https://doi.org/10.1038/nmeth.2967.
[117] Y. Ma, S. Sun, X. Shang, E.T. Keller, M. Chen, X. Zhou, Integrative differential expression and gene set enrichment analysis using summary statistics for scRNA-seq studies, Nat. Commun. (2020), https://doi.org/10.1038/s41467-020-15298-6.
[118] H.A. Pliner, J. Shendure, C. Trapnell, Supervised classification enables rapid annotation of cell atlases, Nat. Methods (2019), https://doi.org/10.1038/s41592-019-0535-3.
[119] X. Shao, J. Liao, X. Lu, R. Xue, N. Ai, X. Fan, scCATCH: automatic annotation on cell types of Clusters from single-cell RNA sequencing data, iScience (2020), https://doi.org/10.1016/j.isci.2020.100882.
[120] L. Jiang, Rare cell type detection, en, Methods Mol. Biol. (2019), https://doi.org/10.1007/978-1-4939-9057-3_5.
[121] W. Saelens, R. Cannoodt, H. Todorov, Y. Saeys, A comparison of single-cell trajectory inference methods, Nat. Biotechnol. (2019), https://doi.org/10.1038/s41587-019-0071-9.
[122] C. Trapnell, D. Cacchiarelli, J. Grimsby, P. Pokharel, S. Li, M. Morse, N.J. Lennon, K.J. Livak, T.S. Mikkelsen, J.L. Rinn, The dynamics and regulators of cell fate decisions are revealed by pseudotemporal ordering of single cells, Nat. Biotechnol. (2014), https://doi.org/10.1038/nbt.2859.
[123] M. Setty, M.D. Tadmor, S. Reich-Zeliger, O. Angel, T.M. Salame, P. Kathail, K. Choi, S. Bendall, N. Friedman, D. Pe'Er, Wishbone identifies bifurcating developmental trajectories from single-cell data, Nat. Biotechnol. (2016), https://doi.org/10.1038/nbt.3569.
[124] G. Giecold, E. Marco, S.P. Garcia, L. Trippa, G.C. Yuan, Robust lineage reconstruction from high-dimensional single-cell data, Nucleic Acids Res. (2016), https://doi.org/10.1093/nar/gkw452.
[125] J. Chen, L. Rénia, F. Ginhoux, Constructing cell lineages from single-cell transcriptomes, Mol. Asp. Med. (2018), https://doi.org/10.1016/j.mam.2017.10.004.
[126] K. Campbell, C. Yau, Ouija: incorporating prior knowledge in single-cell trajectory learning using Bayesian nonlinear factor analysis, bioRxiv (2016), https://doi.org/10.1101/060442.
[127] R. Gonzalo Parra, N. Papadopoulos, L. Ahumada-Arranz, J. El Kholtei, N. Mottelson, Y. Horokhovsky, B. Treutlein, J. Soeding, Reconstructing complex lineage trees from scRNA-seq data using MERLoT, Nucleic Acids Res. (2019), https://doi.org/10.1093/nar/gkz706.
[128] A. Scialdone, K.N. Natarajan, L.R. Saraiva, V. Proserpio, S.A. Teichmann, O. Stegle, J.C. Marioni, F. Buettner, Computational assignment of cell-cycle stage from single-cell transcriptome data, Methods (2015), https://doi.org/10.1016/j.ymeth.2015.06.021.
[129] H. Matsumoto, H. Kiryu, C. Furusawa, M.S.H. Ko, S.B.H. Ko, N. Gouda, T. Hayashi, I. Nikaido, SCODE: an efficient regulatory network inference algorithm from single-cell RNA-Seq during differentiation, Bioinformatics. 33 (2017) 2314–2321, https://doi.org/10.1093/bioinformatics/btx194.
[130] T.N. Tran, G. Bader, Tempora: Cell Trajectory Inference Using Time-Series Single-Cell RNA Sequencing Data, bioRxiv, 2019, https://doi.org/10.1101/846907.
[131] T. Wang, B. Li, C.E. Nelson, S. Nabavi, Comparative analysis of differential gene expression analysis tools for single-cell RNA sequencing data, BMC Bioinformatics (2019), https://doi.org/10.1186/s12859-019-2599-6.
[132] K. Van den Berge, H. Roux de Bézieux, K. Street, W. Saelens, R. Cannoodt, Y. Saeys, S. Dudoit, L. Clement, Trajectory-based differential expression analysis for single-cell sequencing data, Nat. Commun. (2020), https://doi.org/10.1038/s41467-020-14766-3.
[133] S. Aibar, C.B. González-Blas, T. Moerman, V.A. Huynh-Thu, H. Imrichova, G. Hulselmans, F. Rambow, J.C. Marine, P. Geurts, J. Aerts, J. Van Den Oord, Z. K. Atak, J. Wouters, S. Aerts, SCENIC: single-cell regulatory network inference and clustering, Nat. Methods (2017), https://doi.org/10.1038/nmeth.4463.
[134] T. Turki, Y.H. Taguchi, SCGRNs: Novel supervised inference of single-cell gene regulatory networks of complex diseases, Comput. Biol. Med. (2020), https://doi.org/10.1016/j.compbiomed.2020.103656.
[135] A. Pratapa, A.P. Jalihal, J.N. Law, A. Bharadwaj, T.M. Murali, Benchmarking algorithms for gene regulatory network inference from single-cell transcriptomic data, Nat. Methods (2020), https://doi.org/10.1038/s41592-019-0690-6.
[136] X. Qiu, A. Rahimzamani, L. Wang, B. Ren, Q. Mao, T. Durham, J.L. McFaline-Figueroa, L. Saunders, C. Trapnell, S. Kannan, Inferring causal gene regulatory networks from coupled single-cell expression dynamics using scribe, Cell Syst. (2020), https://doi.org/10.1016/j.cels.2020.02.003.
[137] P. Savas, B. Virassamy, C. Ye, A. Salim, C.P. Mintoff, F. Caramia, R. Salgado, D. J. Byrne, Z.L. Teo, S. Dushyanthen, A. Byrne, L. Wein, S.J. Luen, C. Poliness, S. S. Nightingale, A.S. Skandarajah, D.E. Gyorki, C.M. Thornton, P.A. Beavis, S. B. Fox, P.K. Darcy, T.P. Speed, L.K. MacKay, P.J. Neeson, S. Loi, Single-cell profiling of breast cancer T cells reveals a tissue-resident memory subset associated with improved prognosis, Nat. Med. (2018), https://doi.org/10.1038/s41591-018-0078-7.
[138] E. Papalexi, R. Satija, Single-cell RNA sequencing to explore immune cell heterogeneity, Nat. Rev. Immunol. (2018), https://doi.org/10.1038/nri.2017.76.
[139] X. Yu, Y.A. Chen, J.R. Conejo-Garcia, C.H. Chung, X. Wang, Estimation of immune cell content in tumor using single-cell RNA-seq reference data, BMC Cancer (2019), https://doi.org/10.1186/s12885-019-5927-3.
[140] A.L. Roy, Transcriptional regulation in the immune system: one cell at a time, Front. Immunol. 10 (2019) 1355, https://doi.org/10.3389/fimmu.2019.01355.

[141] M.L. Suvà, I. Tirosh, Single-cell RNA sequencing in cancer: lessons learned and emerging challenges, Mol. Cell (2019), https://doi.org/10.1016/j.molcel.2019.05.003.
[142] R. Hou, E. Denisenko, H.T. Ong, J.A. Ramilowski, A.R.R. Forrest, Predicting cell-to-cell communication networks using NATMI, Nat. Commun. (2020), https://doi.org/10.1038/s41467-020-18873-z.
[143] I.C. Macaulay, V. Svensson, C. Labalette, L. Ferreira, F. Hamey, T. Voet, S. A. Teichmann, A. Cvejic, Single-cell rna-sequencing reveals a continuous spectrum of differentiation in hematopoietic cells, Cell Rep. (2016), https://doi.org/10.1016/j.celrep.2015.12.082.
[144] K.S. Yan, C.Y. Janda, J. Chang, G.X.Y. Zheng, K.A. Larkin, V.C. Luca, L.A. Chia, A. T. Mah, A. Han, J.M. Terry, A. Ootani, K. Roelf, M. Lee, J. Yuan, X. Li, C.R. Bolen, J. Wilhelmy, P.S. Davies, H. Ueno, R.J. Von Furstenberg, P. Belgrader, S. B. Ziraldo, H. Ordonez, S.J. Henning, M.H. Wong, M.P. Snyder, I.L. Weissman, A. J. Hsueh, T.S. Mikkelsen, K.C. Garcia, C.J. Kuo, Non-equivalence of Wnt and R-spondin ligands during Lgr5 + intestinal stem-cell self-renewal, Nature (2017), https://doi.org/10.1038/nature22313.
[145] R.M.J. Genga, E.M. Kernfeld, K.M. Parsi, T.J. Parsons, M.J. Ziller, R. Maehr, Single-cell RNA-sequencing-based CRISPRi screening resolves molecular drivers of early human endoderm development, Cell Rep. (2019), https://doi.org/10.1016/j.celrep.2019.03.076.
[146] S. Darmanis, S.A. Sloan, Y. Zhang, M. Enge, C. Caneda, L.M. Shuer, M.G. H. Gephart, B.A. Barres, S.R. Quake, A survey of human brain transcriptome diversity at the single cell level, Proc. Natl. Acad. Sci. U. S. A. (2015), https://doi.org/10.1073/pnas.1507125112.
[147] Q. Mu, Y. Chen, J. Wang, Deciphering brain complexity using single-cell sequencing, Genom. Proteome. Bioinform. (2019), https://doi.org/10.1016/j.gpb.2018.07.007.
[148] A.C. Tolonen, R.J. Xavier, Dissecting the human microbiome with single-cell genomics, Genome Med. (2017), https://doi.org/10.1186/s13073-017-0448-7.
[149] P.M. Strzelecka, A.M. Ranzoni, A. Cvejic, Dissecting human disease with single-cell omics: application in model systems and in the clinic, DMM Dis. Model. Mech. (2018), https://doi.org/10.1242/dmm.036525.
[150] T.V. Lanz, A.K. Pröbstel, I. Mildenberger, M. Platten, L. Schirmer, Single-cell high-throughput technologies in cerebrospinal fluid research and diagnostics, Front. Immunol. (2019), https://doi.org/10.3389/fimmu.2019.01302.
[151] E. Der, H. Suryawanshi, P. Morozov, M. Kustagi, B. Goilav, S. Ranabathou, P. Izmirly, R. Clancy, H.M. Belmont, M. Koenigsberg, M. Mokrzycki, H. Rominieki, J.A. Graham, J.P. Rocca, N. Bornkamp, N. Jordan, E. Schulte, M. Wu, J. Pullman, K. Slowikowski, S. Raychaudhuri, J. Guthridge, J. James, J. Buyon, T. Tuschl, C. Putterman, J. Anolik, W. Apruzzese, A. Arazi, C. Berthier, M. Brenner, J. Buyon, R. Clancy, S. Connery, M. Cunningham, M. Dall'Era, A. Davidson, E. Der, A. Fava, C. Fonseka, R. Furie, D. Goldman, R. Gupta, J. Guthridge, N. Hacohen, D. Hildeman, P. Hoover, R. Hsu, J. James, R. Kado, K. Kalunian, D. Kamen, M. Kretzler, H. Maecker, E. Massarotti, W. McCune, M. McMahon, M. Park, F. Payan-Schober, W. Pendergraft, M. Petri, M. Pichavant, C. Putterman, D. Rao, S. Raychaudhuri, K. Slowikowski, H. Suryawanshi, T. Tuschl, P. Utz, D. Waguespack, D. Wofsy, F. Zhang, Tubular cell and keratinocyte single-cell transcriptomics applied to lupus nephritis reveal type I IFN and fibrosis relevant pathways, Nat. Immunol. (2019), https://doi.org/10.1038/s41590-019-0386-1.
[152] D.R. Gawel, J. Serra-Musach, S. Lilja, J. Aagesen, A. Arenas, B. Asking, M. Bengnér, J. Björkander, S. Biggs, J. Ernerudh, H. Hjortswang, J.E. Karlsson, M. Köpsen, E.J. Lee, A. Lentini, X. Li, M. Magnusson, D. Martínez-Enguita, A. Matussek, C.E. Nestor, S. Schäfer, O. Seifert, C. Sonmez, H. Stjernman, A. Tjärnberg, S. Wu, K. Åkesson, A.K. Shalek, M. Stenmarker, H. Zhang, M. Gustafsson, M. Benson, A validated single-cell-based strategy to identify diagnostic and therapeutic targets in complex diseases, Genome Med. (2019), https://doi.org/10.1186/s13073-019-0657-3.
[153] A.K. Shalek, M. Benson, Single-cell analyses to tailor treatments, Sci. Transl. Med. (2017), https://doi.org/10.1126/scitranslmed.aan4730.
[154] A.J. Wilk, A. Rustagi, N.Q. Zhao, J. Roque, G.J. Martínez-Colón, J.L. McKechnie, G.T. Ivison, T. Ranganath, R. Vergara, T. Hollis, L.J. Simpson, P. Grant, A. Subramanian, A.J. Rogers, C.A. Blish, A single-cell atlas of the peripheral immune response in patients with severe COVID-19, Nat. Med. 26 (2020) 1070–1076, https://doi.org/10.1038/s41591-020-0944-y.