frontiers | Frontiers in Genetics

TYPE Review
PUBLISHED 03 October 2022
DOI 10.3389/fgene.2022.994069

OPEN ACCESS

EDITED BY
Quan Zou,
University of Electronic Science and Technology of China, China

REVIEWED BY
Yanming Di,
Oregon State University, United States
Fu Yang,
Duke University, United States

*CORRESPONDENCE
Roberto Semeraro,
roberto.semeraro@unifi.it

SPECIALTY SECTION
This article was submitted to Computational Genomics, a section of the journal Frontiers in Genetics

RECEIVED 14 July 2022
ACCEPTED 15 September 2022
PUBLISHED 03 October 2022

CITATION
Carangelo G, Magi A and Semeraro R (2022), From multitude to singularity: An up-to-date overview of scRNA-seq data generation and analysis.
Front. Genet. 13:994069.
doi: 10.3389/fgene.2022.994069

From multitude to singularity: An up-to-date overview of scRNA-seq data generation and analysis

Giulia Carangelo[1], Alberto Magi[2] and Roberto Semeraro[3]*

[1]Department of Experimental and Clinical Biomedical Sciences "Mario Serio", University of Florence, Florence, Italy, [2]Department of Information Engineering, University of Florence, Florence, Italy, [3]Department of Experimental and Clinical Medicine, University of Florence, Florence, Italy

Single cell RNA sequencing (scRNA-seq) is today a common and powerful technology in biomedical research settings, allowing to profile the whole transcriptome of a very large number of individual cells and reveal the heterogeneity of complex clinical samples. Traditionally, cells have been classified by their morphology or by expression of certain proteins in functionally distinct settings. The advent of next generation sequencing (NGS) technologies paved the way for the detection and quantitative analysis of cellular content. In this context, transcriptome quantification techniques made their advent, starting from the bulk RNA sequencing, unable to dissect the heterogeneity of a sample, and moving to the first single cell techniques capable of analyzing a small number of cells (1–100), arriving at the current single cell techniques able to generate hundreds of thousands of cells. As experimental protocols have improved rapidly, computational workflows for processing the data have also been refined, opening up to novel methods capable of scaling computational times more favorably with the dataset size and making scRNA-seq much better suited for biomedical research. In this perspective, we will highlight the key technological and computational developments which have enabled the analysis of this growing data, making the scRNA-seq a handy tool in clinical applications.

KEYWORDS

single cell, RNA sequencing, transcriptomics, spatial transcriptomics, biomedical applications, technological evolution

1 Introduction

For many years researchers have tried to comprehend the complexity of tissues, organs and organisms (Grizzi and Chiriva-Internati, 2005). In order to gain this understanding, many studies have focused on cell characterization, redefining the cell as not only the structural but also the functional unit of life (Arendt et al., 2016).

Traditionally, cells have been classified by their morphology or by the expression of certain proteins in functionally distinct settings, but the advent of NGS techniques paved the way for the detection and quantitative analysis of cellular content (Mosmann et al.,

1986; Orkin, 2000; Poulin et al., 2016). The high amount of data generated in modern genomics and transcriptomics experiments permitted to better characterize the architecture of genomes and the complexity of the molecular mechanisms underlying cellular activity, allowing an increasingly more accurate and in-depth depiction of cell plasticity in dynamic processes such as development, differentiation and disease evolution (Sedlazeck et al., 2018; Stark et al., 2019).

Modern cellular and molecular biology knowledge is largely derived from RNA sequencing (RNA-seq) experiments. Over the last 20 years, the transcriptome quantification has shaped our understanding of mechanisms responsible for phenomena, such as the alternativeness of the mRNA splicing process, the regulation of gene expression by non-coding and enhancer RNAs respectively and the drug resistance in some types of cancer, becoming a common and powerful technology suitable for biomedical research (Wang et al., 2008; Morris and Mattick, 2014; Li et al., 2016; Marco-Puche et al., 2019).

The adaptation and evolution of RNA-seq has been driven by technological developments and resulted in a progressive increase of the analysis resolution. Starting from the so called "bulk" RNA-seq, capable of measuring the average gene expression levels of ensembles of millions of cells, we moved to the scRNA-seq that, by allowing to profile the transcriptome of single cells, has revealed rare cellular properties and biologically meaningful cell-to-cell variability, laying the groundwork for heterogeneity-oriented studies (Svensson et al., 2018; Li and Wang, 2021).

As experimental protocols have improved rapidly, computational workflows for processing the data have also been refined, taking into account the increased throughput of scRNA-seq experiments (Andrews et al., 2021). The current "standard" analysis pipeline consists of two main sections: preprocessing, including all the steps necessary to clean the data matrix from unwanted sources of information (quality control, normalization, data correction, feature selection and dimensionality reduction) and cell- and gene-level downstream analysis, used to extract biological insights and describe the underlying biological system. For each of these steps, computational biologists developed a range of methods which perform better in different tasks and settings, making the creation of generalizable workflows for single cell experiments analysis challenging.

In this perspective, we will present an overview of the computational workflow, arguing the tools available to proceed in each step and highlighting the key technological developments which have enabled the analysis of this ever-increasing amount of data, making the scRNA-seq a handy tool in biomedical research.

2 Single cell sequencing

The first studies of single cells date back to the early 90s and were motivated by incoming discoveries which highlighted cell plasticity in dynamic processes and the different functionality based on localization (Eberwine et al., 1992). The advent of NGS techniques opened up to the era of quantitative analysis of cellular content, although first transcriptomics techniques (bulk RNA-seq) were not able to survey the diversity of cell types in a sample (Hong et al., 2020). The scaling of technologies to profile large numbers of cells in parallel has been the key to driving single cell transcriptomics forward (see Figure 1).

2.1 Technical evolution

The first example of single cell transcriptomics is the study of a handful of mouse primordial germ cells by Tang et al. (2009). By manual modification of cDNA amplification protocols previously employed in microarray analyses, he captured and quantified for the first time the full-length cDNAs for 64% of the expressed genes of a single cell, without affecting the accuracy of the protocol, which was however very time consuming and limited to small numbers of atypically large cells.

In the wake of Tang et al., new different approaches were developed including the so-called tag sequencing methods. For instance, in 2011, Islam et al. quantified the transcriptome of 85 cells by means of single cell tagged reverse transcription (STRT) (Islam et al., 2011). In brief, the authors settled single cells into the wells of a 96-well PCR plate preloaded with lysis buffer and then added reverse transcription (RT) reagents to generate a first-strand cDNA. Next, a unique template-switching oligo (TSO) with a specific sequence (six-base) on its 3′ end and a universal primer sequence on the 5′ was added to each well triggering the RT template-switching mechanism which produces a cDNA molecule incorporating the sequence at the 3′ of the TSO.

The introduction of these "barcode" sequences allowed, for the first time, to assay many cells in parallel *via* multiplexed unbiased RNA-seq, although, in the STRT-seq method, full-length cDNA is amplified by template-switching, but only the 5′ end fragment is captured and sequenced. To overcome thislimitation, the full-length SMART-seq (Ramsköld et al., 2012) and SMART-seq2 (Picelli et al., 2013) protocols were developed by Ramsköld et al. and Picelli et al., in 2011 and 2013 respectively. Compared with existing tag based methods, SMART-seq has improved read coverage across transcripts, promoting a detailed analyses of alternative transcript isoforms and identification of single-nucleotide polymorphisms (SNP).

In sight of this, it is therefore necessary to clarify that it is possible to profile the transcriptome through full-length transcript analysis or by digital counting of either 3′ or 5′ ends. While the two methods carry similar levels of reproducibility, the latter methods consist in a cost-effective solution to quantify a high amount of transcripts at the expense of a large loss of information for each of these,

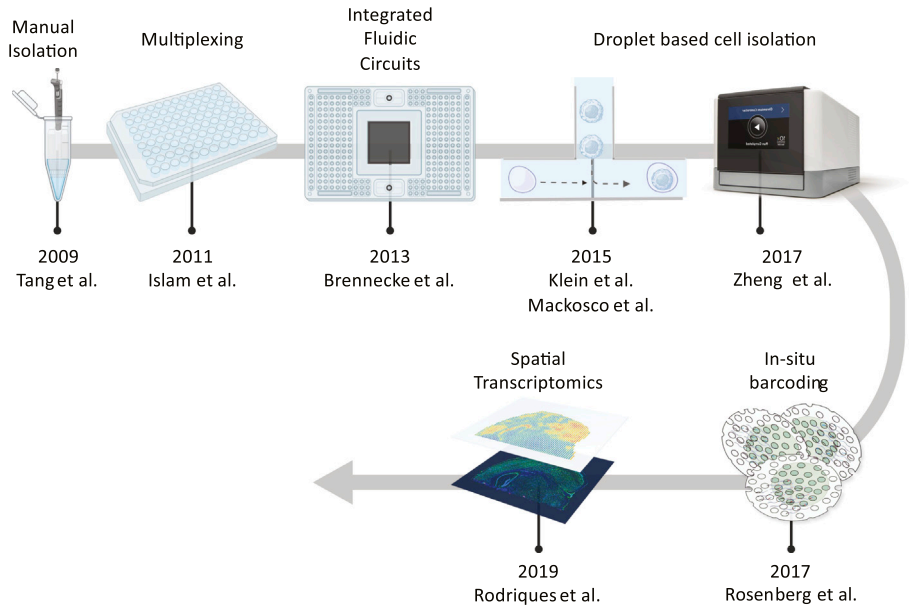

FIGURE 1
Noteworthy technologies that have allowed to profile large numbers of cells in parallel. Starting from manual isolation methods, a jump to ~100 cells was enabled by sample multiplexing and than the development of integrated fluidic circuits increased these numbers to an order of magnitude. Next, the introduction of nanodroplet technologies increased throughput even further to hundreds of thousands of cells, as for *in situ* barcoding which favoured the development of spatial methods.

contrary to the former which, by taking advantage of full-length transcripts entirety, allows the detection of splice variants and alternative transcripts, as well as genetic alterations in the transcribed fraction but for a lower number of cells.

A further application of SMART-seq2 protocols, although with some modifications (Egidio et al., 2014), is found also in the work of Brennecke et al. (2013). By means of an integrated fluidic circuit (IFC) method, implemented in the Fluidigm C1 system, they studied 96 cells isolated into individual reaction chambers and subjected to automatic staining, lysis, and sequencing in extraordinarily fast times and in a "passive" manner never seen before. In fact, the key feature of this technology is the design of microfluidics devices (or chips) that allow the sequential delivery of very small and precise volumes into tiny reaction chambers. However, a major limitation derives from the number of these chambers (96) which restrict the analysis to an equivalent number of cells, as for Brennecke in 2013. Some following large-scale studies made use of a large number of IFCs to create big data sets (Zeisel et al., 2015).

In 2015, the advent of microfluidic platforms bypassed this drawback thanks to the usage of nanoliter microreactor droplets which can encapsulate cells with no physical, and therefore numerical, restraints. The inDrop (Klein et al., 2015) and the Drop-seq (Macosko et al., 2015) protocols enter the scene with related commercial systems that allow to randomly capture cells in beads containing lysis buffer, RT reagents and barcoded oligonucleotide primers, so that mRNA is released from each cell and remains trapped in the bead to be barcoded during synthesis of cDNA. The two methods mainly differ in barcoding strategy and amplification technique, since the inDrop protocol uses hydrogel beads bearing poly(T) primers with defined barcodes and, after pooling, initiates linear amplification (IVT), contrary to Drop-seq which uses beads with random barcodes and amplifies through PCR. The random isolation of cells, however, comes with inherent limitations. Poisson statistics of cell capture to ensure that mostly single cells are isolated means there will always be large inefficiencies in terms of cell isolation, and the pool of barcodes will always have to be substantially larger than the number of cells captured to avoid barcode duplication. A large number of barcodes means the usage of very long and therefore expensive oligos. To reduce their synthesis costs, two different strategies are adopted by both methods: the combination of multiple shorter designed barcodes (e.g., 8–10 bases) into longer barcodes (e.g., 8 bases +22-base linker +10 bases = 40 bases), as for InDrop, or the synthesis of very long (e.g., 12 bases) random barcodes, as for DropSeq. This second procedure is simpler than the first and does not require any synthesized oligos for the barcodes. However, in the first approach barcodes can be designed to avoid biases and ensure that each sequence will be distinct.

The need for a large number of oligos was mitigated in 2017, through the advent of the combinatorial *in situ* barcoding

methods, when Rosenberg et al. introduced the split-pool ligation-based transcriptome sequencing (SPLiT-seq), a low-cost, scRNA-seq method that enables transcriptional profiling of hundreds of thousands of fixed cells or nuclei in a single experiment (Rosenberg et al., 2018). In brief, a suspension of formaldehyde-fixed cells or nuclei passes through four rounds of combinatorial barcoding. At the first round, cells are distributed in a 96-wells plate and labelled with a specific tag. Next, cells are pooled and subjected to another label-expanding round. So, in the third round, another portion is added, carrying with it a unique molecular identifier (UMI) specific for each transcript and also used in other tag-based methods, such as STRT-seq, InDrop and Drop-seq, to better quantify the native, unamplified transcript levels (Islam et al., 2014; Stegle et al., 2015). Finally, sequencing adapters are introduced by PCR and, subsequently, each transcriptome is assembled by combining reads containing the same four-barcode combination.

Along with SPLiT-seq, one of the most vastly used methods makes its entry. The 10x Genomics company presents a new system called Chromium, based on an inDrop-seq variant. Specifically, single cells, RT reagents, Gel Beads containing barcoded oligonucleotides, and oil are combined onto a microfluidic chip to form reaction vesicles called Gel Beads in Emulsion, or GEMs. GEMs are formed in parallel within the 8 microfluidic channels of the chip, allowing the user to process hundreds to hundreds of thousands of single cells in a single 7-min run, with a ~65% of capture efficiency (Zheng et al., 2017). Within each GEM reaction vesicle, a single cell is lysed, the Gel Bead is dissolved to free the identically barcoded RT oligonucleotides into solution, and reverse transcription of polyadenylated mRNA occurs. As a result, all cDNAs from a single cell will have the same barcode, allowing the sequencing reads to be mapped back to their single cells of origin. The scalability and robustness of the system has favored the rapid diffusion of this device and its acquisition by many research laboratories in the medical field. Another contribution to this field comes from the so-called spatial RNA sequencing (spRNA-seq). Introduced in 2019 to enable the understanding of how tumor cells can communicate with each other, escape the immune system, develop drug resistance and metastasize, it combines the strengths of the global transcriptional analysis of bulk RNA-seq and *in situ* hybridization, providing whole transcriptome data with spatial information. Two technologies are currently available by 10x Genomics and Nanostring Technologies, both using proprietary spatial gene expression slides on which to fix fresh-frozen or Formalin-Fixed Paraffin-Embedded (FFPE) tissue. The two technologies differ for slide functionalization. The 10x device contains oligo capture probes, similar to those coating the gel beads, and once the tissue is fixed, stained and imaged, it is permeabilized to release the RNA, captured by probes and subjected to on-slide cDNA synthesis (Ståhl et al., 2016; Rodriques et al., 2019). The Nanostring system, uses barcode-labeled probes and fluorescent markers to hybridize to mRNA targets and to establish tissue "geography" respectively. After the regions-of-interest (ROIs) are selected, the barcodes are released *via* UV exposure and collected from the ROIs on the tissue (Moses and Pachter, 2022).

The labeled RNAs, for both technologies, are then sequenced through standard NGS procedures.

The spRNA-seq is still in its early stages and there are several common challenges that limit its applications, including non-single cell resolution, relatively low sensitivity, high cost and labor-intensive process, but given its capacity to dissect intercellular subpopulations sensitively and spatially, it will inevitably become a fundamental area of research in both discovery and therapeutics.

2.2 Bioinformatic analysis

2.2.1 General information and workflow

The rapid technological evolution that allowed the parallel analysis of thousands of cells, promoting the spread of scRNA-seq techniques, was accompanied by the development of new data analysis pipelines capable of managing such a large amount of data. The mathematical representation of these massive datasets is an "expression" matrix, defined by the number of detected genes and observed cells respectively. The process aimed at its generation starts with the read quality check. The FastQ files outputted from the sequencer are evaluated by means of quality check tools, like FastQC (https://www.bioinformatics.babraham.ac.uk/projects/fastqc/), to undergo de-multiplexing, adapter trimming, alignment and count. Tailored pipelines such as Cell Ranger (Zheng et al., 2017), UMI-tools (Smith et al., 2017), scPipe (Tian et al., 2018) and zUMIs (Parekh et al., 2018), were developed to carry out these preliminary steps. Alternatively, researchers can build their own workflows by combining individual methods that address each of the aforementioned tasks (see Table 1). For instance, the STAR (Dobin et al., 2013) aligner implements the STARsolo algorithm suited to trim, align and count this kind of data in a very fast way (Brüning et al., 2022).

Moreover, if reads are UMI-tagged, only cell barcodes that represent intact individual cells are kept. The most unambiguous approach to assess emptiness is to calculate a dataset-specific threshold of the minimum number of UMIs required to consider a barcode as a cell (Zheng et al., 2017). Alternatively tools, such as EmptyDrops (Lun et al., 2016a), identify cell barcodes that significantly deviate from background levels of RNA present in empty wells. The resulting cells still show unwanted biases. All processes involved in bias removal define the so called "preprocessing" which consists in quality control, normalization, batch correction, feature selection and dimensionality reduction. All these steps are preparatory for the following expression analysis, used to extract biological insights and describe the underlying biological system (see Figure 2).

TABLE 1 Raw data processing tools.

	Name	Alignment	QC	Count	CC	PL	References
Pipelines	CellRanger	x	x	x	x	R/Python	Zheng et al. (2017)
	UMI-tools	x	x	x	x	Python	Smith et al. (2017)
	scPipe	x	x	x	x	C++/R	Tian et al. (2018)
	zUMIs	x	x	x	x	R/Perl	Parekh et al. (2018)
	dropEst	x	x	x	x	C++	Petukhov et al. (2018)
	Optimus	x	x	x	x	Python/C++	
Tools	STAR	x	x	x	x	C/C++	Dobin et al. (2013)
	HISAT2	x	-	-	-	C/C++	Kim et al. (2015)
	kallisto	-	-	x	-	C/C++	Bray et al. (2016)
	FastQC	-	x	-	-	Java	
	HTSeq	-	x	x	-	Python	Putri et al. (2022)
	featureCount	-	-	x	-	C	Liao et al. (2014)
	EmptyDrops	-	-	-	x	R	Lun et al. (2019)

QC, quality check; CC, cell calling; PL, programming language.

Also in this context, tailored pipelines and individual tools are available to perform each operation. Toolboxes, such as Scanpy (Wolf et al., 2018), SCell (Diaz et al., 2016), Seurat (Hao et al., 2021) and scater (McCarthy et al., 2017) allow to complete multiple tasks bypassing problems related to data format conversions, making the analysis simpler. On the other hand, it is important to remember that it is difficult for a tool with many functions to continue to represent the state of the art in all of them.

In this perspective, we will present an overview of the computational workflow, arguing the tools available to proceed in each step (see Table 2).

2.2.1.1 Quality control

Before analyzing the expression matrix, we must assess the uniqueness of each barcode and cell viability. To this end, it is important to keep in mind that some droplets might contain more than 1 cell or no cell at all, making it a doublet, multiplet or an empty droplet. Furthermore, cells can be dying or damaged during isolation, misrepresenting the sample composition. So, we need to filter out them.

A possible solution is to identify these cells by evaluating three aspects of the data: the number of counts per cell/barcode (count depth), the number of genes per cell/barcode, and the fraction of counts from mitochondrial genes per cell/barcode. The thresholds for these covariates are arbitrary based on the general characteristics of the data itself, but they allow us to filter out cells with low count depths, few detected genes and/or high fraction of mitochondrial counts, as those are considered damaged cells, and at the same time they allow to filter out cells with too high counts which are indicative of doublets or multiplets (Ilicic et al., 2016).

However, a misinterpretation of these covariates could lead to wrong filtering, since in some cases a deviation in one of these values may be related to a particular cell condition, such as heavy respiration (high mitochondrial counts), quiescence (low counts, few genes) and a larger size (high counts). Therefore, they should be considered jointly when univariate thresholding decisions are made, and these thresholds should be set as permissive as possible to avoid filtering out viable cell populations unintentionally.

For doublet detection, more precise methods were developed (Xiong et al., 2022). For instance, scrublet (Wolock et al., 2019) is able to discern "embedded" (same cell type) from "neotypic" (different cell types) doublets, assuming that among all observed transcriptomes, multiplets are relatively rare events and that all cell states contributing to doublets are also present as single cells elsewhere in the data.

Quality control can also include a gene filtering step, since genes expressed in few cells are non-informative of the cellular heterogeneity. The threshold is again arbitrary, but in principle it should scale with the number of cells in the dataset and the intended downstream analysis, because, based on that choice, for example, it could limiti the identification of small clusters that might actually carry valuable information about less represented cell population.

2.2.1.2 Normalization

By means of quality control we removed sources of unwanted and inaccurate information. However, the dataset is still affected by multiple biases due to technical and biological variability. Sources responsible for such events could be, for example, capture efficiency, amplification and incomplete library sequencing. The consequence is an alteration in the counts which make cells incomparable (Macosko et al., 2015).

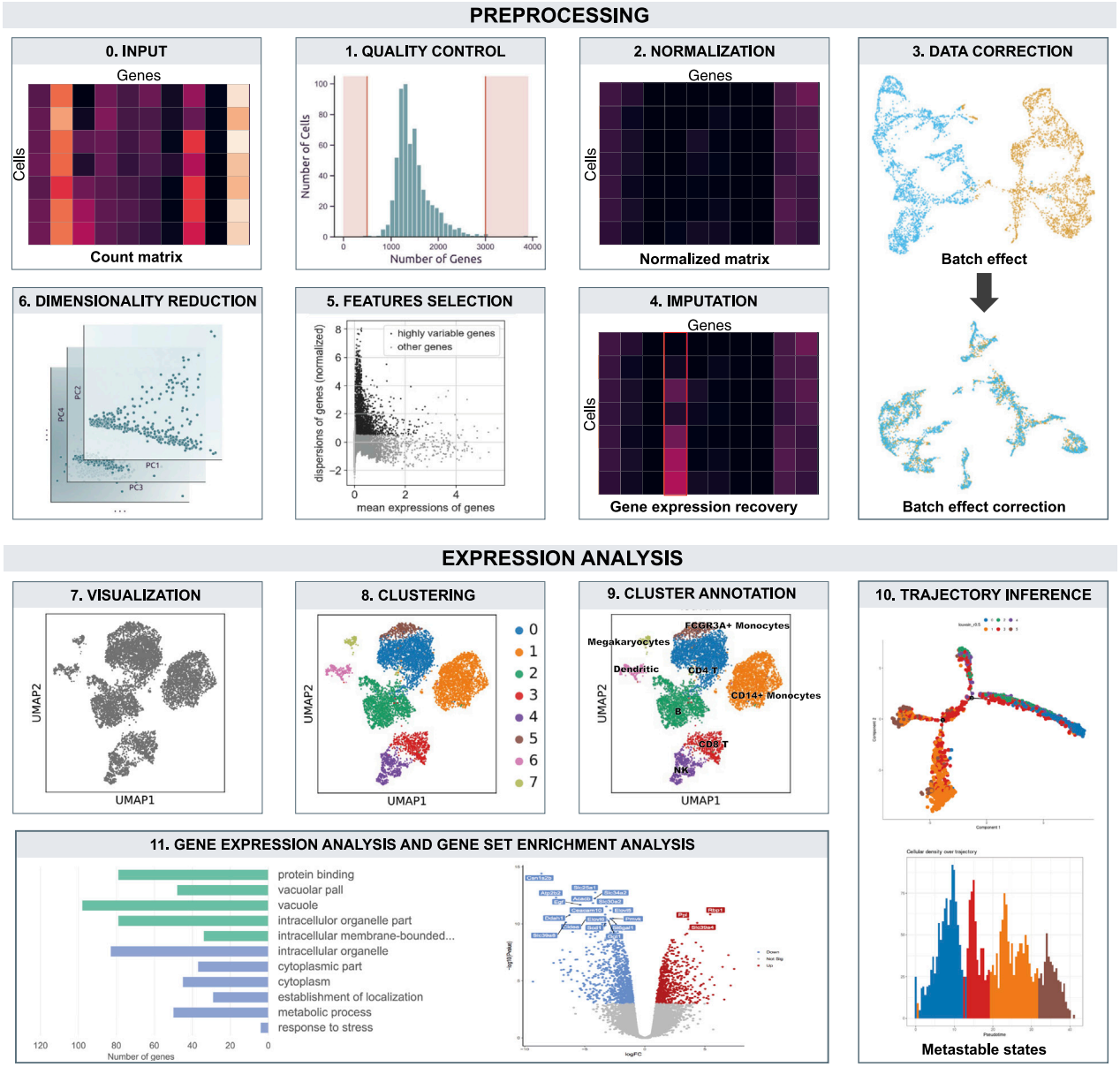

FIGURE 2
Overview of the workflow. The count matrix undergo preprocessing and expression analysis. Boxes are ordered according data analysis flow.

Normalization addresses this issue by e.g., scaling count data to obtain correct relative gene expression abundances between cells. Available methods can be linear or non-linear: a linear approach involves the estimation of size factors based on a linear regression over genes, while non-linear methods usually apply parametric modelling on count data and correlate technical and biological sources of variability to correct both (Lytal et al., 2020).

The most common normalization approach is the count depth scaling by "counts per million" (CPM), which operates by dividing gene counts by the total number of mapped reads per sample and multiplying by 1×10^6. CPM falls within linear global scaling normalization methods and assumes that all cells in the dataset initially contained an equal number of mRNA molecules (10^6) and count depth differences arise only due to sampling. Variations of this method scale the size factors with different factors of 10, or by the median count depth per cell in the dataset. Tools such as scran (Lun et al., 2016b) and Scanpy implement extensions of CPM approach. The former was proven to perform better than others in order to proceed with differential expression (DE) analysis (Vieth et al., 2019).

TABLE 2 Analysis tools.

		Preprocessing				Expression analysis				PL	
	Name	QC	N	BC	DR	V	C	DE	TI		References
Pipelines	CellRanger	x	x	x	x	x	x	x	-	R/Python	Zheng et al. (2017)
	Scanpy	x	x	x	x	x	x	x	x	Python	Wolf et al. (2018)
	Seurat	x	x	x	x	x	x	x	-	R	Hao et al. (2021)
	SCell	x	x	x	x	x	x	x	x	Matlab	Diaz et al. (2016)
	scater	x	x	x	x	x	x	x	x	R	McCarthy et al. (2017)
	Pagoda2	x	x	x	x	x	x	x	-	R	Lopez et al. (2018)
Tools	Doublet Finder	x	-	-	-	-	-	-	-	R	McGinnis et al. (2019)
	Scrublet	x	-	-	-	-	-	-	-	Python	Wolock et al. (2019)
	scds	x	-	-	-	-	-	-	-	R	Bais and Kostka (2020)
	scran	x	x	-	-	-	-	-	-	R	Bray et al. (2016)
	SCnorm	-	x	-	-	-	-	-	-	R	
	bioinfokit	-	x	-	-	-	-	-	-	R	Putri et al. (2022)
	ComBat	-	-	x	-	-	-	-	-	R	Johnson et al. (2007)
	mnnCorrect	-	-	x	-	-	-	-	-	R	Haghverdi et al. (2018)
	Harmony	-	-	x	-	-	-	-	-	R	Korsunsky et al. (2019)
	BBKNN	-	-	x	-	-	-	-	-	Python	Polański et al. (2020)
	SAUCIE	-	-	x	x	x	x	-	-	Python	Amodio et al. (2019)
	scVI	-	-	x	x	-	-	x	-	Python	Boyeau et al. (2019)
	PCA	-	-	-	x	-	-	-	-	Python	Pedregosa et al. (2011)
	t-SNE	-	-	-	x	x	-	-	-	Python/R	Van der Maaten and Hinton (2008)
	UMAP	-	-	-	x	x	-	-	-	Python/R	McInnes et al. (2018)
	Louvain	-	-	-	-	-	x	-	-	Python/R	Blondel et al. (2008)
	Leiden	-	-	-	-	-	x	-	-	Python/R	Traag et al. (2019)
	MAST	-	-	-	-	-	-	x	-	R	Finak et al. (2015)
	scCODE	-	-	-	-	-	-	x	-	R	Zou et al. (2022)
	Slingshot	-	-	-	-	-	-	-	x	R	Street et al. (2018)
	DPT	-	-	-	-	-	-	-	x	Python	Haghverdi et al. (2016)
	Whishbone	-	x	-	x	x	-	x	x	Python	Setty et al. (2016)
	Monocle2	-	x	x	x	x	x	x	x	R	Trapnell et al. (2014)
	Monocle3	-	x	x	x	x	x	x	x	R	Cao et al. (2019)
	velocyto	-	x	x	x	x	x	x	x	Python/R	La Manno et al. (2018)
	scVelo	-	x	x	x	x	x	x	x	Python	Bergen et al. (2020)

QC, quality check; N, normalization; BC, batch correction; DR, dimensionality reduction; V, visualization; C, clustering; DE, differential expression; TI, trajectory inference; PL, programming language.

For datasets with strong batch effects, non-linear methods were proven to be more reliable, particularly for plate-based scRNA-seq data, usually affected by batch effect between plates (Svensson et al., 2017).

For full-length sequencing protocols, methods which consider the gene length are more suitable. The most common is "transcripts per million" method (TPM), implemented, for example, in the bioinfokit toolbox (http://doi.org/10.5281/zenodo.3698145) (Putri et al., 2022).

Another crucial factor for normalization is the presence of synthetic spike-ins or UMIs as a means to correct for amplification bias. By adding known concentrations of external transcripts, called spike-ins, it is possible to evaluate the presence of technical artifacts, looking for differences between their observed and expected expression. By calculating a cell-specific factor that adjusts for the differences, and by applying that factor to endogenous genes, normalized expression estimates can be obtained. In spite of the promise, there are many challenges in getting spike-ins to work well, which

can result in inconsistent detections (Grün et al., 2014). Contrary to spike-ins, UMIs are easier to handle since they are attached to individual transcripts prior to PCR, making each molecule unique and allowing an absolute molecular count (Kivioja et al., 2011).

Also, genes can be normalized to make them comparable between cells. Gene counts can be scaled to have a zero mean and a unit variance (z-score), making genes equally weighted. The scaling is currently not a routine because sometimes it could be useful to give genes the same weight and sometimes not, due to the effect produced by an expression magnitude difference.

Normalized data should be log (x+1)-transformed for use with following analysis methods that assume data are normally distributed. Three main effects derive from this transformation: log values represent log fold changes (unit to measure expression), they become normally distributed, reducing the skewness of the data and finally, the mean-variance relationship typical of single cell data is mitigated (Brennecke et al., 2013).

2.2.1.3 Batch Correction

Through the normalization, we mitigated the sources of technical variability responsible for gene counts alterations. However, the dataset may still contains unwanted signals of technical and biological nature. In the latter category falls for e.g., the cell cycle effect, while in the former, the batch effect deriving from different experimental protocols or/and different plates.

In order to get rid of these biases, it is possible to proceed with data and batch correction. Currently, several tools can accomplish these tasks with different approaches (Chu et al., 2022). For example, in development-oriented studies regressing out the cell cycle effect could uncover the desired biological signals (Vento-Tormo et al., 2018; Büttner et al., 2019). To this end, methods such as Scanpy and Seurat implement functions to score the cell cycle phases and regress linearly their biological effect. Alternatively, tailored tools based on complex models, like f-scLVM (Buettner et al., 2017), are available. Sometimes, also the count bias produced by differences in cell size, if not enough corrected through normalization, could be further mitigated to emphasize development-related signals. In this situation, regressing both covariates at the same time could be the best solution to account for dependence between them.

Correcting for biological biases, however, it is not always necessary or useful, since they can be avoided through pondered experimental design or because they can relate to the biological process of interest. The same observation is in part valid also for those of technical nature. In fact, even in this case a clever experimental design allows to reduce their influence but, if present, they have no correlation with the biological signals, so they must be mitigated. This process, named batch correction, can be conducted between samples and cells of the same experiment through linear models, or among different datasets derived from multiple experimental settings through non-linear models.

One of the most common linear methods is ComBat (Johnson et al., 2007) which take into account the batch effect on mean and variance of the dataset, performing very well in most settings (Büttner et al., 2019).

If the differences in the datasets are more pronounced, linear models could confound the intra- and inter-technical and biological biases, and in this circumstances non-linear models implemented in tools such as Canonical Correlation Analysis (CCA) (Butler et al., 2018), Mutual Nearest Neighbors (MNN) (Haghverdi et al., 2018), Batch balanced kNN (BBKNN) (Polański et al., 2020) and Harmony (Korsunsky et al., 2019) have been proved to overcome the same issue and smooth out unwanted and misleading differences.

2.2.1.4 Imputation

The information stored in a single cell dataset has a very sparse nature. In mathematical terms, it translates into a matrix full of zeros. Many normalization approaches do not remove them, assuming that they represent missing values to account in calculations. However, reducing their number could reduce the noise, improving the estimation of gene-gene correlations (van Dijk et al., 2018).

Currently, many tools are available to achieve this task, and the best performing ones are mainly based on deep learning algorithms (Bao et al., 2022). In this category fall DeepImpute (Arisdakessian et al., 2019) and Deep Count Autoencoder network (DCA) (Eraslan et al., 2019). The first one uses highly correlated genes of the target genes to impute the missing values, while the second can capture the nonlinear gene-gene correlation. Their application proved to improve the performance in cell clustering, DE analysis and trajectory inference.

However, when applying expression recovery, one should take into consideration that no method is perfect. Thus, any method may over- or under-correct noise in the data. Indeed, false correlation signals have been reported as a result of expression recovery (Andrews and Hemberg, 2018).

In light of this, it is hard to assess if imputation will succeed in a particular application. A reasonable approach would be to impute for visualization and avoid it to generate hypothesis during exploratory data analysis.

2.2.1.5 Feature selection and dimensionality reduction

After proceeding with the “data cleaning” steps, a human scRNA-seq dataset can still contain up to 15,000 genes. Such a big and multidimensional object is, however, hard to manage and visualize. For these reasons, it is subjected to dimensionality reduction.

To go through this process it is important to keep in mind that many residual genes do not represent the data variability, which is a key feature to explore the heterogeneity of the sample, and so that we can consider them uninformative and ignorable. This process is called feature selection. A common way to reach

this result is to look for highly variable genes (HVGs) by binning them by their mean expression and preserving the ones with the highest mean-to-variance ratio in each bin (Brennecke et al., 2013). Methods such as Scanpy and Cell Ranger implement functions to define the HVGs starting from log-transformed data, while others like Seurat work on the raw counts. Typically, between 1,000 and 5,000 HVGs are selected to proceed with robust downstream analysis (Klein et al., 2015).

Their identification is crucial also to proceed with the following dimensionality reduction. Indeed, common methods like the Principal Component Analysis (PCA) (Pearson, 1901; Pedregosa et al., 2011) benefit from using HVGs to define the reduced components used to summarize the dataset features in a low-dimensional space. This is possible through a linear approach which transforms a set of correlated variables into a smaller number of uncorrelated variables, called principal components (PCs), preserving as much of the data's variation as possible. To determine the N most informative PCs, "elbow" heuristics or the permutation-test-based jackstraw method can be used (Chung and Storey, 2015; Macosko et al., 2015).

The PCA is a technique that comes from the field of linear algebra and can be used as a data preparation technique to create a projection of a dataset prior to fitting a model. Indeed, for complex datasets whose structure could not be captured by two or three PCs, non-linear combination methods such as t-distributed stochastic neighbour embedding (t-SNE) (Van der Maaten and Hinton, 2008) and Uniform Approximation and Projection (UMAP) (McInnes et al., 2018) perform better, taking advantage of PCA data.

2.2.1.6 Visualization and clustering

Non-linear methods are commonly used to create a two-dimensional plot summarizing an scRNA-seq dataset from a larger number of significant components. t-SNE and UMAP are two typical solutions to achieve this task and are implemented in almost all scRNA-seq data processing toolbox. t-SNE takes a high dimensional data set and reduces it to a low dimensional graph focusing on capturing local similarity at the expense of global structure. UMAP, instead, tends to favour fully connected representations of the dataset using a cell-cell nearest-neighbour network to then estimates a low dimensional embedding of the data. The latter is largely replacing the former, although different representations could give different insights. In this perspective, it is good to know that also diffusion maps and partition-based methods exists to visualize complex data in different manners and for different applications, e.g., diffusion maps are good to make inferences in trajectory analyses, while partition-based methods approximate the topology of the data using clusters to produce a simplified "coarse-grain" visualization of the data, useful with very large datasets.

The clustering is commonly performed with the Louvain (Blondel et al., 2008) and the Leiden (Traag et al., 2019) algorithms.

The aim of this step is to define groups of cells with similar expression profiles, because these groups could represent cell types, intermediate cell states or other interesting aspects of the data.

Both methods are based on K-Nearest Neighbour approach (KNN graph) where cells are represented as nodes in a graph, each connected to its K most similar cells, obtained using Euclidean distances on the PC-reduced expression space, so that densely sampled regions of expression space will be represented as densely connected regions in the plot (Zappia and Oshlack, 2018).

Clustering can also be performed at multiple resolutions to inspect data at different levels of detail (i.e., more clusters of smaller dimensions). Moreover, the resulting groups can be iteratively subclustered to allow the identification of cell states captured within the same cluster.

2.2.1.7 Cluster annotation

Once clusters have been defined, it is time to identify the represented cell populations. This can be done by defining their gene signatures through the identification of marker genes. To this end, DE testings are usually applied between two groups representing the cluster and the rest of the dataset. Next, simple statistical tests such as the Wilcoxon rank-sum test or the *t*-test are used to rank the derived genes by their difference in expression. The top-ranked genes from the respective test statistic are regarded as marker genes.

Clusters can be also annotated by comparing marker genes from the dataset with those from reference datasets *via* enrichment tests, the Jaccard index or other overlap statistics. Indeed, reference databases such as the mouse brain atlas (Zeisel et al., 2018) or the Human Cell Atlas (HCA) (Regev et al., 2017) are increasingly becoming available, facilitating cell identity annotation. Also automated methods like single cell NET (Tan and Cahan, 2019) are available to accomplish this step and speedup the annotation process, although a manual revision is always suggested due to the plasticity of cell states which sometimes could be confused with others.

2.2.1.8 Trajectory analysis and metastable states

Cell clustering sometimes is not the appropriate strategy to study a dataset. Many biological processes, characterizing a dataset, cannot be described through discrete classification but rather in a more continuous way (Tanay and Regev, 2017). To achieve this result we need to apply gene dynamic models capable of ordering cells along an axis defining the time process, also known as pseudotime. This type of approach is commonly used to study processes such as development and differentiation, and it is called Trajectory analysis.

Several methods are currently available to infer trajectories of increasing complexity, from simple linear or bifurcating paths to complex graphs, trees, or more intricated trajectories.

Usually, these algorithms take the reduced or corrected data as input in order to minimize technical variation and capture only the biological one, taking advantage also of HVGs, which are used to define the consecutive states derived from transcriptional distances from a root cell. None of the available methods has been shown to overperform the others for all kinds of trajectories, although different approaches benefit different ends, as shown in previous comparative studies (Saelens et al., 2019).

For instance, the tool Slingshot (Street et al., 2018) proved to perform better when inferring linear or multifurcating trajectories, contrary to the current state-of-the-art, Monocle2 (Trapnell et al., 2014), which gives best results in more complex and branched situations, along with its later version Monocle3 (Cao et al., 2019) and the Diffusion Pseudotimes (DPT) implemented in Scanpy (Haghverdi et al., 2016).

The aforementioned python toolbox offers also the chance to reconcile the information derived from clustering and trajectory inference, by means of the Partition-based graph abstraction (PAGA) algorithm (Wolf et al., 2019). In detail, using a statistical model for cell cluster interactions, PAGA places an edge between cluster nodes whose cells are more similar than expected, generating a map representing the static and dynamic nature of the data.

As trajectory inference deals with the way the cells in our sample change according to a pseudotime, it becomes possible to define the "preferential" transcriptomic states of the process evaluating the region density. Dense regions of cells represent the so called "metastable states" which can be visualized through histograms.

Unfortunately, few of the aforementioned methods include an evaluation of uncertainty in their model, so the predicted results should be confirmed with alternative approaches to avoid method bias (Griffiths et al., 2018). A common way to achieve this goal is to infer time dynamics by measuring relative abundances of exonic and intronic reads, representing spliced and unspliced transcripts. The change of their abundance, termed RNA velocity, allows to infer the direction in which each cell is moving in expression space along with an estimate of the rate of change, unlocking new ways to study cellular dynamics by granting access to not only the descriptive state of a cell, but also to its direction and speed of movement.

Currently, two modeling approaches exist, the originally proposed "steady-state" model adopted by velocyto (La Manno et al., 2018) and the subsequently extended dynamical model implemented in scVelo (Bergen et al., 2020). The former estimates velocities as the deviation of the observed ratio of unspliced to spliced mRNA from an inferred steady-state ratio, by leading sometimes to predicition errors if the central assumptions of a common splicing rate and the observation of the full splicing dynamics with steady-state mRNA levels are violated. The latter overcomes these limitations by generalizing velocity estimation to transient systems through the application of a likelihood-based dynamical model which solves the full transcriptional dynamics of splicing kinetics.

2.2.1.9 Gene expression analysis

Once the nature of each cluster is assessed, focusing on gene expression can give us a much broader idea on processes and mechanisms that differ among them. In this perspective, tools such as DE analysis and gene set enrichment analysis (GSEA) can help us investigate the molecular variability deriving from different experimental (medical treatment) or biological (different cell lines) conditions.

DE methods originate with bulk sequencing data analysis, where a few samples were compared to understand the molecular consequences of different experimental conditions. In single cell settings, the variables at stake increase as the number of cells under examination increases, due to cell-to-cell variability and biases such as dropout (Vallejos et al., 2017; Hicks et al., 2018). Tailored tools like MAST (Finak et al., 2015) or scCODE (Zou et al., 2022) are available to handle these features and perform DE on large single cell datasets in reasonable times, however, bulk DE tools, like DESeq2 (Love et al., 2014) and EdgeR (Robinson et al., 2010), have been proved to outperform some single cell counterparts if properly calibrated, but taking long times (Van den Berge et al., 2018). Uncorrected data are preferred for these applications, so it is crucial to account for confounding factors to perform a robust estimation of differentially expressed genes.

The testing result consists in a long list of genes differentially expressed between two or more conditions, sometimes hard to interpret in a meaningful way. To overcome this limitation, we can analyze them by grouping into sets based on shared characteristics, e.g., biological process and matabolic pathway. This approach, called GSEA, tests whether these characteristics are overrepresented in the candidate gene list and relies on the usage of curated databases such as the Gene Ontology (Ashburner et al., 2000; The Gene Ontology Consortium, 2017), KEGG (Kanehisa et al., 2017), String (https://string-db.org) and Reactome (Gillespie et al., 2022). Tools like gseapy (https://gseapy.readthedocs.io/en/latest/) and biomaRt (Durinck et al., 2009) are available to accomplish this task through multiple tests, querying the mentioned databases. Furthermore, novel algorithms (Vento-Tormo et al., 2018) allowed to proceed with paired ligand-receptor analyses which inspect the interaction between cell clusters.

2.3 Experimental design considerations

scRNA-seq has opened new avenues for the characterization of heterogeneity in a large variety of cellular systems, allowing to obtain transcriptome-wide data from individual cells. Although gene-expression profiling at single cell level has revealed an

unprecedented variety of cell types and subpopulations that were invisible with traditional experimental techniques, it introduced new challenges due to the intrinsic nature of the data.

Indeed, scRNA-seq datasets show increased variability, complex expression distributions and an abundance of zeros compared to those produced in "bulk" experiments, making challenging to create broadly applicable experimental designs. In light of this, each experiment requires the user to make informed decisions before to proceed with a pondered design, which have to satisfy three principles formalized by R. A. Fisher in 1935: replication, randomization and blocking (Box, 1980).

To prepare an experiment, respecting such principles, it is good to start with a balanced block design in which samples collected from multiple conditions are evenly distributed across plates and lanes of the sequencer in order to reduce technical variation and not confound it with the biological one (Baran-Gale et al., 2018). On the opposite, processing samples separately, isolating cells from each sample onto separate plates (one for sample) and sequencing them on separate lanes (one for sample), produces a confounded design affected by additional sources of technical variation associated with batch preparation of libraries or sequencing. In this context, balanced design allow to bypass the batch correction step in the computational analysis, reducing computational times and user intervention on data.

Experimental design considerations will also be affected by the various protocols and platforms available for scRNA-seq. For instance, full-length capture or 3′ methods offer different way to explore sample characteristics.

As example, in an observational study setting, working with high numbers of cells could be the best solution to get insights on the transcriptional heterogeneity of the sample. To this aim, 3′ methods represent the best solution allowing to capture higher amounts of cells (100–1,00,000) and quantify their transcriptomes in a more simple and precise way, thanks to the usage of UMIs. On the other hand, to conduct more "in depth" observations or study genetic alterations (SNPs, structural variants) in the transcribed fraction, full-length approaches are more well suited, benefiting from a higher capture efficiency and a more precise information, but at cost of a minor number of cells (96–384). Therefore, more reads will be required for more refined tasks (Pollen et al., 2014; Wu et al., 2014), such as fully characterizing transcript structure, estimating the expression of rare isoforms, or distinguishing cells on the basis of subtle differences, while fewer reads but larger cell numbers may be preferred when mapping out a large population, searching for rare but distinct cell types, or pooling cells *in silico* to obtain average gene-expression clusters. According to this, if we design an experiment to search for a rare cell population, we have to take into account the number of cells that need to be sequenced to get such a population. This parameter can be estimated based on the expected heterogeneity of all cells in a sample, the minimum frequency expected of the rare cell type within the sample and the minimum number of cells of each type desired in the resulting data set.

In case no prior knowledge about the heterogeneity of the cell population is available, a practical solution is to perform the study with a high cell number and lower sequencing depth, and then perform pre-purification of the interested cells by fluorescence-activated cell sorting (FACS) with in-depth sequencing.

Another relevant difference between the two protocols relates to the UMIs usage. Indeed, full-length approaches make the inclusion of UMIs difficult, as each full-length transcript is fragmented following reverse transcription, and each fragment would need to be linked to the single UMI for that transcript. On the other hand, 3′ methods, like the 10x Genomics system, include a 10/12 bp UMI in each read at the beginning of the protocol, facilitating the molecule counting and the evaluation of sequencing saturation through the analysis of UMI duplicates. Moreover, the use of UMI has an impact on normalization procedure, since they are a consistent means to correct for amplification bias.Overall, several factors need to be considered before choosing a method for scRNA-seq. Whatever the design, it is always beneficial to record and retain information on as many factors as possible to facilitate downstream diagnostics.

3 Biomedical applications

Modern cellular and molecular biology knowledge is largely derived from RNA-seq experiments which allowed to understand the complexity of the dynamics responsible for metabolic alterations, fueling much discovery and innovation in the field of medicine over recent years.

The evolution of such techniques was driven by the development of protocols and devices capable of extracting transcriptomic information from an ever increasing number of single cells, laying the groundwork for heterogeneity-oriented studies.

The chance to dissect a sample in its composing cell lines opened up new perspectives in clinical studies oriented to the discovery of rare cell populations involved in the onset and evolution of diseases such as tumors. A proof of this assertion comes from Ramsköld et al., in 2012 and Patel et al., in 2014, which studied, for the first time (Ramsköld et al., 2012; Patel et al., 2014), the compositional architecture of melanoma and glioblastoma samples at single cell level. In the wake of them, an increasing number of studies and researchers have started exploiting the technique to successfully characterize cell populations in a variety of tumors (Dago et al., 2014; Ting et al., 2014; Puram et al., 2017; Zhao et al., 2020; Pal et al., 2021; Tian et al., 2022), defining their role into the disease process and their identity through the assignment of gene signatures (Young et al., 2018; Peired et al., 2020). Other contributions to the field comes from the integration of scRNA-seq and Copy Number Variant (CNV) detection. Tirosh et al., in 2016,

successfully applied this technique to get new insights on intra- and interindividual, spatial, functional and genomic heterogeneity in melanoma cells, as well as details related to the tumor microenvironment and the cells populating it, validating the presence of a dormant drug-resistant population (Tirosh et al., 2016).

Similarly, in 2018, Fan et al. took advantage of CNVs and Loss of Heterozygosis (LOH) to identify and characterize the transcriptional programs which drive the distinct genetic subclones in a tumor sample (Fan et al., 2018).

Also in the neurological field, the scRNA-seq succeeded, revealing the heterogenous nature of brain cells involved in Alzheimer's disease and the different outcomes related to their different gene expression patterns (Mathys et al., 2019). In this contest, Lodato et al. exploited single cell sequencing to identify Single Nucleotide Variants (SNVs) in neuronal cells, demonstrating how somatic mutations can be used to reconstruct the developmental lineage of neurons, which live for decades in a postmitotic state accumulating mutations responsible for the creation of nested lineage trees and the relative polyclonal architecture (Lodato et al., 2015).

While, for blood, liver and heart samples, the introduction of trajectory analyses have provided new insights on differentiation processes, allowing to trace the fate of progenitor cells revealing the plasticity of their transcriptome through the identification of new transitional cell states (Jia et al., 2018; Popescu et al., 2019; Liang et al., 2022). However, the regulatory networks driving these processes are more complex and characterized by confounding factors like redundancy and nonlinear cross talk between pathways, e.g., developmental and signaling factors in the immune system. An unbiased approach to elucidate such a circuits and their alterations are the perturbation studies, which, by making use of the massive parallelism of single cell technologies merged with CRISPR-mediated editing, allow to knockout multiple target genes simultaneously producing different cell responses useful to clarify the function of multiple factors and their interactions in tens of thousands of cells (Adamson et al., 2016; Dixit et al., 2016; Jaitin et al., 2016). To extend this application to the analysis of multiple unrelated individuals, new methods that harness natural genetic variation were developed. Tools like demuxlet (Kang et al., 2018) determine the sample identity of each droplet, using genotyping data (SNPs), to characterize inter-individual variation and cell-type-specific genetic control of gene expression. Similarly, Van der Wijst et al. used SNP data to characterize alterations of gene co-expression pathways, focusing also on celltype-specific expression quantitative trait loci (eQTLs) (van der Wijst et al., 2018), promoting a new way to identify genetic variants that impact regulatory networks.

Another hot topic is damage recovery, since a better understanding of these mechanisms could allow us to identify the players involved in success or fail of such processes, offering new hints in the development of better diagnostic tools, prognostic biomarkers and signaling pathways amenable to therapeutic targeting (Kirita et al., 2019; Melica et al., 2022).

4 Future perspectives and conclusion

Single cell RNA sequencing was proven to be a cutting-edge technology in life sciences over the past decade. This field is developing remarkably rapidly and numerous easily accessible commercial solutions capable of characterizing hundreds of thousands of cells in parallel in reasonable times at competitive costs are currently available, making scRNA-seq much better suited for biomedical research and for clinical applications.

The spread of these devices fueled much discovery and innovation also in the computational biology field, promoting the development of novel approaches to extract information from the data produced by such technologies, and algorithms capable of analyzing them, scaling computational times more favorably with the dataset size. Moreover, along with RNA profiling, single cell technologies are currently employed to acquire information about multiple types of molecules in parallel, promoting the so-called "multimodal profiling". In fact, today it is possible to integrate information related to chromatin accessibility (Cusanovich et al., 2015), methylation state (Angermueller et al., 2016), cell-surface proteins (Stoeckius et al., 2017), to reveal the full-scale complexity of biological systems. Also, the developmental trajectories can be studied in a more precise way by matching the single cell technologies with CRISPR-Cas9 based genome editing. Methods such as scGESTALT (Raj et al., 2018) and LINNAEUS (Spanjaard et al., 2018) allow to simultaneously characterize molecular identities and lineage histories of thousands of cells during development and disease through the analysis of lineage barcodes, generated by genome editing.

However, high-throughput techniques come with the expense of decreased molecule capture rates, and future methods need to better balance cell numbers with cell resolution. Furthermore, with the future development of new and better bioinformatic tools, the individual tool recommendations presented here will require updates, yet the general considerations regarding the stages of data processing should remain the same.

Spatial dimension of single cell transcriptomics also represents an exciting field because, although novel and more precise technologies are becoming available (Eng et al., 2019), it presents several common challenges that limit its applications, including non-single cell resolution, relatively low sensitivity, high cost and labor-intensive process.

In conclusion, we have presented a brief and concise overview of single cell RNA sequencing technology and its

applications. The continuous development of the technology will broaden its adoption in clinical and personalized medicine.

Author contributions

GC and RS wrote the manuscript and organized the figures. AM supervised the manuscript. All authors read and approved the final manuscript.

Funding

This work has been supported by the Associazione Italiana per la Ricerca sul Cancro (AIRC Investigator Grant 20307, "Third Generation Cancer Genomics").

Conflict of interest

The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Publisher's note

All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.

References

Adamson, B., Norman, T. M., Jost, M., Cho, M. Y., Nuñez, J. K., Chen, Y., et al. (2016). A multiplexed single-cell crispr screening platform enables systematic dissection of the unfolded protein response. *Cell* 167, 1867–1882.e21. doi:10.1016/j.cell.2016.11.048

Amodio, M., van Dijk, D., Srinivasan, K., Chen, W. S., Mohsen, H., Moon, K. R., et al. (2019). Exploring single-cell data with deep multitasking neural networks. *Nat. Methods* 16, 1139–1145. doi:10.1038/s41592-019-0576-7

Andrews, T. S., and Hemberg, M. (2018). False signals induced by single-cell imputation. *F1000Res.* 7, 1740. doi:10.12688/f1000research.16613.2

Andrews, T. S., Kiselev, V. Y., McCarthy, D., and Hemberg, M. (2021). Tutorial: guidelines for the computational analysis of single-cell rna sequencing data. *Nat. Protoc.* 16, 1–9. doi:10.1038/s41596-020-00409-w

Angermueller, C., Clark, S. J., Lee, H. J., Macaulay, I. C., Teng, M. J., Hu, T. X., et al. (2016). Parallel single-cell sequencing links transcriptional and epigenetic heterogeneity. *Nat. Methods* 13, 229–232. doi:10.1038/nmeth.3728

Arendt, D., Musser, J. M., Baker, C. V. H., Bergman, A., Cepko, C., Erwin, D. H., et al. (2016). The origin and evolution of cell types. *Nat. Rev. Genet.* 17, 744–757. doi:10.1038/nrg.2016.127

Arisdakessian, C., Poirion, O., Yunits, B., Zhu, X., and Garmire, L. X. (2019). Deepimpute: an accurate, fast, and scalable deep neural network method to impute single-cell rna-seq data. *Genome Biol.* 20, 211. doi:10.1186/s13059-019-1837-6

Ashburner, M., Ball, C. A., Blake, J. A., Botstein, D., Butler, H., Cherry, J. M., et al. (2000). Gene ontology: tool for the unification of biology. the gene ontology consortium. *Nat. Genet.* 25, 25–29. doi:10.1038/75556

Bais, A. S., and Kostka, D. (2020). scds: computational annotation of doublets in single-cell RNA sequencing data. *Bioinformatics* 36, 1150–1158. doi:10.1093/bioinformatics/btz698

Bao, S., Li, K., Yan, C., Zhang, Z., Qu, J., and Zhou, M. (2022). Deep learning-based advances and applications for single-cell RNA-sequencing data analysis. *Brief. Bioinform.* 23, bbab473. doi:10.1093/bib/bbab473

Baran-Gale, J., Chandra, T., and Kirschner, K. (2018). Experimental design for single-cell RNA sequencing. *Brief. Funct. Genomics* 17, 233–239. doi:10.1093/bfgp/elx035

Bergen, V., Lange, M., Peidli, S., Wolf, F. A., and Theis, F. J. (2020). Generalizing rna velocity to transient cell states through dynamical modeling. *Nat. Biotechnol.* 38, 1408–1414. doi:10.1038/s41587-020-0591-3

Blondel, V. D., Guillaume, J.-L., Lambiotte, R., and Lefebvre, E. (2008). Fast unfolding of communities in large networks. *J. Stat. Mech.* 2008, P10008. doi:10.1088/1742-5468/2008/10/p10008

Box, J. F. (1980). Ra fisher and the design of experiments, 1922–1926. *Am. Stat.* 34, 1–7. doi:10.2307/2682986

Boyeau, P., Lopez, R., Regier, J., Gayoso, A., Jordan, M. I., and Yosef, N. (2019). Deep generative models for detecting differential expression in single cells. *bioRxiv*. doi:10.1101/794289

Bray, N. L., Pimentel, H., Melsted, P., and Pachter, L. (2016). Near-optimal probabilistic rna-seq quantification. *Nat. Biotechnol.* 34, 525–527. doi:10.1038/nbt.3519

Brennecke, P., Anders, S., Kim, J. K., Kołodziejczyk, A. A., Zhang, X., Proserpio, V., et al. (2013). Accounting for technical noise in single-cell rna-seq experiments. *Nat. Methods* 10, 1093–1095. doi:10.1038/nmeth.2645

Brüning, R. S., Tombor, L., Schulz, M. H., Dimmeler, S., and John, D. (2022). Comparative analysis of common alignment tools for single-cell RNA sequencing. *Gigascience* 11, giac001. doi:10.1093/gigascience/giac001

Buettner, F., Pratanwanich, N., McCarthy, D. J., Marioni, J. C., and Stegle, O. (2017). f-sclvm: scalable and versatile factor analysis for single-cell RNA-seq. *Genome Biol.* 18, 212. doi:10.1186/s13059-017-1334-8

Butler, A., Hoffman, P., Smibert, P., Papalexi, E., and Satija, R. (2018). Integrating single-cell transcriptomic data across different conditions, technologies, and species. *Nat. Biotechnol.* 36, 411–420. doi:10.1038/nbt.4096

Büttner, M., Miao, Z., Wolf, F. A., Teichmann, S. A., and Theis, F. J. (2019). A test metric for assessing single-cell RNA-seq batch correction. *Nat. Methods* 16, 43–49. doi:10.1038/s41592-018-0254-1

Cao, J., Spielmann, M., Qiu, X., Huang, X., Ibrahim, D. M., Hill, A. J., et al. (2019). The single-cell transcriptional landscape of mammalian organogenesis. *Nature* 566, 496–502. doi:10.1038/s41586-019-0969-x

Chu, S.-K., Zhao, S., Shyr, Y., and Liu, Q. (2022). Comprehensive evaluation of noise reduction methods for single-cell RNA sequencing data. *Brief. Bioinform.* 23, bbab565. doi:10.1093/bib/bbab565

Chung, N. C., and Storey, J. D. (2015). Statistical significance of variables driving systematic variation in high-dimensional data. *Bioinformatics* 31, 545–554. doi:10.1093/bioinformatics/btu674

Cusanovich, D. A., Daza, R., Adey, A., Pliner, H. A., Christiansen, L., Gunderson, K. L., et al. (2015). Multiplex single cell profiling of chromatin accessibility by combinatorial cellular indexing. *Science* 348, 910–914. doi:10.1126/science.aab1601

Dago, A. E., Stepansky, A., Carlsson, A., Luttgen, M., Kendall, J., Baslan, T., et al. (2014). Rapid phenotypic and genomic change in response to therapeutic pressure in prostate cancer inferred by high content analysis of single circulating tumor cells. *PLoS One* 9, e101777. doi:10.1371/journal.pone.0101777

Diaz, A., Liu, S. J., Sandoval, C., Pollen, A., Nowakowski, T. J., Lim, D. A., et al. (2016). Scell: integrated analysis of single-cell RNA-seq data. *Bioinformatics* 32, 2219–2220. doi:10.1093/bioinformatics/btw201

Dixit, A., Parnas, O., Li, B., Chen, J., Fulco, C. P., Jerby-Arnon, L., et al. (2016). Perturb-seq: Dissecting molecular circuits with scalable single-cell RNA profiling of pooled genetic screens. *Cell* 167, 1853–1866.e17. doi:10.1016/j.cell.2016.11.038

Dobin, A., Davis, C. A., Schlesinger, F., Drenkow, J., Zaleski, C., Jha, S., et al. (2013). Star: ultrafast universal RNA-seq aligner. *Bioinformatics* 29, 15–21. doi:10.1093/bioinformatics/bts635

Durinck, S., Spellman, P. T., Birney, E., and Huber, W. (2009). Mapping identifiers for the integration of genomic datasets with the r/bioconductor package biomart. *Nat. Protoc.* 4, 1184–1191. doi:10.1038/nprot.2009.97

Eberwine, J., Yeh, H., Miyashiro, K., Cao, Y., Nair, S., Finnell, R., et al. (1992). Analysis of gene expression in single live neurons. *Proc. Natl. Acad. Sci. U. S. A.* 89, 3010–3014. doi:10.1073/pnas.89.7.3010

Egidio, C., Ooi, A., Holcomb, I., Ruff, D., Boutet, S., Wang, J., et al. (2014). A method for detecting protein expression in single cells using the c1™ single-cell auto prep system (tech2p.874). *J. Immunol.* 192, 135.5.

Eng, C.-H. L., Lawson, M., Zhu, Q., Dries, R., Koulena, N., Takei, Y., et al. (2019). Transcriptome-scale super-resolved imaging in tissues by RNA seqfish. *Nature* 568, 235–239. doi:10.1038/s41586-019-1049-y

Eraslan, G., Simon, L. M., Mircea, M., Mueller, N. S., and Theis, F. J. (2019). Single-cell rna-seq denoising using a deep count autoencoder. *Nat. Commun.* 10, 390. doi:10.1038/s41467-018-07931-2

Fan, J., Lee, H.-O., Lee, S., Ryu, D.-E., Lee, S., Xue, C., et al. (2018). Linking transcriptional and genetic tumor heterogeneity through allele analysis of single-cell RNA-seq data. *Genome Res.* 28, 1217–1227. doi:10.1101/gr.228080.117

Finak, G., McDavid, A., Yajima, M., Deng, J., Gersuk, V., Shalek, A. K., et al. (2015). Mast: a flexible statistical framework for assessing transcriptional changes and characterizing heterogeneity in single-cell RNA sequencing data. *Genome Biol.* 16, 278. doi:10.1186/s13059-015-0844-5

Gillespie, M., Jassal, B., Stephan, R., Milacic, M., Rothfels, K., Senff-Ribeiro, A., et al. (2022). The reactome pathway knowledgebase 2022. *Nucleic Acids Res.* 50, D687–D692. doi:10.1093/nar/gkab1028

Griffiths, J. A., Scialdone, A., and Marioni, J. C. (2018). Using single-cell genomics to understand developmental processes and cell fate decisions. *Mol. Syst. Biol.* 14, e8046. doi:10.15252/msb.20178046

Grizzi, F., and Chiriva-Internati, M. (2005). The complexity of anatomical systems. *Theor. Biol. Med. Model.* 2, 26. doi:10.1186/1742-4682-2-26

Grün, D., Kester, L., and van Oudenaarden, A. (2014). Validation of noise models for single-cell transcriptomics. *Nat. Methods* 11, 637–640. doi:10.1038/nmeth.2930

Haghverdi, L., Büttner, M., Wolf, F. A., Buettner, F., and Theis, F. J. (2016). Diffusion pseudotime robustly reconstructs lineage branching. *Nat. Methods* 13, 845–848. doi:10.1038/nmeth.3971

Haghverdi, L., Lun, A. T. L., Morgan, M. D., and Marioni, J. C. (2018). Batch effects in single-cell RNA-sequencing data are corrected by matching mutual nearest neighbors. *Nat. Biotechnol.* 36, 421–427. doi:10.1038/nbt.4091

Hao, Y., Hao, S., Andersen-Nissen, E., Mauck, W. M., 3rd, Zheng, S., Butler, A., et al. (2021). Integrated analysis of multimodal single-cell data. *Cell* 184, 3573–3587.e29. doi:10.1016/j.cell.2021.04.048

Hicks, S. C., Townes, F. W., Teng, M., and Irizarry, R. A. (2018). Missing data and technical variability in single-cell RNA-sequencing experiments. *Biostatistics* 19, 562–578. doi:10.1093/biostatistics/kxx053

Hong, M., Tao, S., Zhang, L., Diao, L.-T., Huang, X., Huang, S., et al. (2020). RNA sequencing: new technologies and applications in cancer research. *J. Hematol. Oncol.* 13, 166. doi:10.1186/s13045-020-01005-x

Ilicic, T., Kim, J. K., Kolodziejczyk, A. A., Bagger, F. O., McCarthy, D. J., Marioni, J. C., et al. (2016). Classification of low quality cells from single-cell RNA-seq data. *Genome Biol.* 17, 29. doi:10.1186/s13059-016-0888-1

Islam, S., Kjällquist, U., Moliner, A., Zajac, P., Fan, J.-B., Lönnerberg, P., et al. (2011). Characterization of the single-cell transcriptional landscape by highly multiplex RNA-seq. *Genome Res.* 21, 1160–1167. doi:10.1101/gr.110882.110

Islam, S., Zeisel, A., Joost, S., La Manno, G., Zajac, P., Kasper, M., et al. (2014). Quantitative single-cell RNA-seq with unique molecular identifiers. *Nat. Methods* 11, 163–166. doi:10.1038/nmeth.2772

Jaitin, D. A., Weiner, A., Yofe, I., Lara-Astiaso, D., Keren-Shaul, H., David, E., et al. (2016). Dissecting immune circuits by linking crispr-pooled screens with single-cell RNA-seq. *Cell* 167, 1883–1896.e15. doi:10.1016/j.cell.2016.11.039

Jia, G., Preussner, J., Chen, X., Guenther, S., Yuan, X., Yekelchyk, M., et al. (2018). Single cell RNA-seq and atac-seq analysis of cardiac progenitor cell transition states and lineage settlement. *Nat. Commun.* 9, 4877. doi:10.1038/s41467-018-07307-6

Johnson, W. E., Li, C., and Rabinovic, A. (2007). Adjusting batch effects in microarray expression data using empirical bayes methods. *Biostatistics* 8, 118–127. doi:10.1093/biostatistics/kxj037

Kanehisa, M., Furumichi, M., Tanabe, M., Sato, Y., and Morishima, K. (2017). Kegg: new perspectives on genomes, pathways, diseases and drugs. *Nucleic Acids Res.* 45, D353–D361. doi:10.1093/nar/gkw1092

Kang, H. M., Subramaniam, M., Targ, S., Nguyen, M., Maliskova, L., McCarthy, E., et al. (2018). Multiplexed droplet single-cell RNA-sequencing using natural genetic variation. *Nat. Biotechnol.* 36, 89–94. doi:10.1038/nbt.4042

Kim, D., Langmead, B., and Salzberg, S. L. (2015). Hisat: a fast spliced aligner with low memory requirements. *Nat. Methods* 12, 357–360. doi:10.1038/nmeth.3317

Kirita, Y., Chang-Panesso, M., and Humphreys, B. D. (2019). Recent insights into kidney injury and repair from transcriptomic analyses. *Nephron* 143, 162–165. doi:10.1159/000500638

Kivioja, T., Vähärautio, A., Karlsson, K., Bonke, M., Enge, M., Linnarsson, S., et al. (2011). Counting absolute numbers of molecules using unique molecular identifiers. *Nat. Methods* 9, 72–74. doi:10.1038/nmeth.1778

Klein, A. M., Mazutis, L., Akartuna, I., Tallapragada, N., Veres, A., Li, V., et al. (2015). Droplet barcoding for single-cell transcriptomics applied to embryonic stem cells. *Cell* 161, 1187–1201. doi:10.1016/j.cell.2015.04.044

Korsunsky, I., Millard, N., Fan, J., Slowikowski, K., Zhang, F., Wei, K., et al. (2019). Fast, sensitive and accurate integration of single-cell data with harmony. *Nat. Methods* 16, 1289–1296. doi:10.1038/s41592-019-0619-0

La Manno, G., Soldatov, R., Zeisel, A., Braun, E., Hochgerner, H., Petukhov, V., et al. (2018). RNA velocity of single cells. *Nature* 560, 494–498. doi:10.1038/s41586-018-0414-6

Li, X., and Wang, C.-Y. (2021). From bulk, single-cell to spatial RNA sequencing. *Int. J. Oral Sci.* 13, 36. doi:10.1038/s41368-021-00146-0

Li, W., Notani, D., and Rosenfeld, M. G. (2016). Enhancers as non-coding RNA transcription units: Recent insights and future perspectives. *Nat. Rev. Genet.* 17, 207–223. doi:10.1038/nrg.2016.4

Liang, Y., Kaneko, K., Xin, B., Lee, J., Sun, X., Zhang, K., et al. (2022). Temporal analyses of postnatal liver development and maturation by single-cell transcriptomics. *Dev. Cell* 57, 398–414.e5. doi:10.1016/j.devcel.2022.01.004

Liao, Y., Smyth, G. K., and Shi, W. (2014). featurecounts: an efficient general purpose program for assigning sequence reads to genomic features. *Bioinformatics* 30, 923–930. doi:10.1093/bioinformatics/btt656

Lodato, M. A., Woodworth, M. B., Lee, S., Evrony, G. D., Mehta, B. K., Karger, A., et al. (2015). Somatic mutation in single human neurons tracks developmental and transcriptional history. *Science* 350, 94–98. doi:10.1126/science.aab1785

Lopez, R., Regier, J., Cole, M. B., Jordan, M. I., and Yosef, N. (2018). Deep generative modeling for single-cell transcriptomics. *Nat. Methods* 15, 1053–1058. doi:10.1038/s41592-018-0229-2

Love, M. I., Huber, W., and Anders, S. (2014). Moderated estimation of fold change and dispersion for RNA-seq data with deseq2. *Genome Biol.* 15, 550. doi:10.1186/s13059-014-0550-8

Lun, A. T. L., Bach, K., and Marioni, J. C. (2016a). Pooling across cells to normalize single-cell RNA sequencing data with many zero counts. *Genome Biol.* 17, 75. doi:10.1186/s13059-016-0947-7

Lun, A. T. L., McCarthy, D. J., and Marioni, J. C. (2016b). A step-by-step workflow for low-level analysis of single-cell RNA-seq data with bioconductor. *F1000Res.* 5, 2122. doi:10.12688/f1000research.9501.2

Lun, A. T. L., Riesenfeld, S., Andrews, T., Dao, T. P., Gomes, T., and Marioni, J. C.participants in the 1st Human Cell Atlas Jamboree (2019). Emptydrops: distinguishing cells from empty droplets in droplet-based single-cell RNA sequencing data. *Genome Biol.* 20, 63. doi:10.1186/s13059-019-1662-y

Lytal, N., Ran, D., and An, L. (2020). Normalization methods on single-cell RNA-seq data: An empirical survey. *Front. Genet.* 11, 41. doi:10.3389/fgene.2020.00041

Macosko, E. Z., Basu, A., Satija, R., Nemesh, J., Shekhar, K., Goldman, M., et al. (2015). Highly parallel genome-wide expression profiling of individual cells using nanoliter droplets. *Cell* 161, 1202–1214. doi:10.1016/j.cell.2015.05.002

Marco-Puche, G., Lois, S., Benítez, J., and Trivino, J. C. (2019). RNA-seq perspectives to improve clinical diagnosis. *Front. Genet.* 10, 1152. doi:10.3389/fgene.2019.01152

Mathys, H., Davila-Velderrain, J., Peng, Z., Gao, F., Mohammadi, S., Young, J. Z., et al. (2019). Single-cell transcriptomic analysis of alzheimer's disease. *Nature* 570, 332–337. doi:10.1038/s41586-019-1195-2

McCarthy, D. J., Campbell, K. R., Lun, A. T. L., and Wills, Q. F. (2017). Scater: pre-processing, quality control, normalization and visualization of single-cell RNA-seq data in r. *Bioinformatics* 33, 1179–1186. doi:10.1093/bioinformatics/btw777

McGinnis, C. S., Murrow, L. M., and Gartner, Z. J. (2019). Doubletfinder: Doublet detection in single-cell RNA sequencing data using artificial nearest neighbors. *Cell Syst.* 8, 329–337.e4. doi:10.1016/j.cels.2019.03.003

McInnes, L., Healy, J., Saul, N., and Großberger, L. (2018). Umap: Uniform manifold approximation and projection. *J. Open Source Softw.* 3, 861. doi:10.21105/joss.00861

Melica, M. E., Antonelli, G., Semeraro, R., Angelotti, M. L., Lugli, G., Landini, S., et al. (2022). Differentiation of crescent-forming kidney progenitor cells into podocytes attenuates severe glomerulonephritis in mice. *Sci. Transl. Med.* 14, eabg3277. doi:10.1126/scitranslmed.abg3277

Morris, K. V., and Mattick, J. S. (2014). The rise of regulatory RNA. *Nat. Rev. Genet.* 15, 423–437. doi:10.1038/nrg3722

Moses, L., and Pachter, L. (2022). Museum of spatial transcriptomics. *Nat. Methods* 19, 534–546. doi:10.1038/s41592-022-01409-2

Mosmann, T. R., Cherwinski, H., Bond, M. W., Giedlin, M. A., and Coffman, R. L. (1986). Two types of murine helper t cell clone. i. definition according to profiles of lymphokine activities and secreted proteins. *J. Immunol.* 136, 2348–2357.

Orkin, S. H. (2000). Diversification of haematopoietic stem cells to specific lineages. *Nat. Rev. Genet.* 1, 57–64. doi:10.1038/35049577

Pal, B., Chen, Y., Vaillant, F., Capaldo, B. D., Joyce, R., Song, X., et al. (2021). A single-cell RNA expression atlas of normal, preneoplastic and tumorigenic states in the human breast. *EMBO J.* 40, e107333. doi:10.15252/embj.2020107333

Parekh, S., Ziegenhain, C., Vieth, B., Enard, W., and Hellmann, I. (2018). zumis - a fast and flexible pipeline to process RNA sequencing data with umis. *Gigascience* 7, giy059. doi:10.1093/gigascience/giy059

Patel, A. P., Tirosh, I., Trombetta, J. J., Shalek, A. K., Gillespie, S. M., Wakimoto, H., et al. (2014). Single-cell RNA-seq highlights intratumoral heterogeneity in primary glioblastoma. *Science* 344, 1396–1401. doi:10.1126/science.1254257

Pearson, K. (1901). Liii. on lines and planes of closest fit to systems of points in space. *Lond. Edinb. Dublin philosophical Mag. J. Sci.* 2, 559–572. doi:10.1080/14786440109462720

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., et al. (2011). Scikit-learn: Machine learning in python. *J. Mach. Learn. Res.* 12, 2825–2830.

Peired, A. J., Antonelli, G., Angelotti, M. L., Allinovi, M., Guzzi, F., Sisti, A., et al. (2020). Acute kidney injury promotes development of papillary renal cell adenoma and carcinoma from renal progenitor cells. *Sci. Transl. Med.* 12, eaaw6003. doi:10.1126/scitranslmed.aaw6003

Petukhov, V., Guo, J., Baryawno, N., Severe, N., Scadden, D. T., Samsonova, M. G., et al. (2018). dropest: pipeline for accurate estimation of molecular counts in droplet-based single-cell RNA-seq experiments. *Genome Biol.* 19, 78. doi:10.1186/s13059-018-1449-6

Picelli, S., Björklund, Å. K., Faridani, O. R., Sagasser, S., Winberg, G., and Sandberg, R. (2013). Smart-seq2 for sensitive full-length transcriptome profiling in single cells. *Nat. Methods* 10, 1096–1098. doi:10.1038/nmeth.2639

Polański, K., Young, M. D., Miao, Z., Meyer, K. B., Teichmann, S. A., and Park, J.-E. (2020). Bbknn: fast batch alignment of single cell transcriptomes. *Bioinformatics* 36, 964–965. doi:10.1093/bioinformatics/btz625

Pollen, A. A., Nowakowski, T. J., Shuga, J., Wang, X., Leyrat, A. A., Lui, J. H., et al. (2014). Low-coverage single-cell mRNA sequencing reveals cellular heterogeneity and activated signaling pathways in developing cerebral cortex. *Nat. Biotechnol.* 32, 1053–1058. doi:10.1038/nbt.2967

Popescu, D.-M., Botting, R. A., Stephenson, E., Green, K., Webb, S., Jardine, L., et al. (2019). Decoding human fetal liver haematopoiesis. *Nature* 574, 365–371. doi:10.1038/s41586-019-1652-y

Poulin, J.-F., Tasic, B., Hjerling-Leffler, J., Trimarchi, J. M., and Awatramani, R. (2016). Disentangling neural cell diversity using single-cell transcriptomics. *Nat. Neurosci.* 19, 1131–1141. doi:10.1038/nn.4366

Puram, S. V., Tirosh, I., Parikh, A. S., Patel, A. P., Yizhak, K., Gillespie, S., et al. (2017). Single-cell transcriptomic analysis of primary and metastatic tumor ecosystems in head and neck cancer. *Cell* 171, 1611–1624.e24. doi:10.1016/j.cell.2017.10.044

Putri, G. H., Anders, S., Pyl, P. T., Pimanda, J. E., and Zanini, F. (2022). Analysing high-throughput sequencing data in python with htseq 2.0. *Bioinformatics* 38, 2943–2945. doi:10.1093/bioinformatics/btac166

Raj, B., Wagner, D. E., McKenna, A., Pandey, S., Klein, A. M., Shendure, J., et al. (2018). Simultaneous single-cell profiling of lineages and cell types in the vertebrate brain. *Nat. Biotechnol.* 36, 442–450. doi:10.1038/nbt.4103

Ramsköld, D., Luo, S., Wang, Y.-C., Li, R., Deng, Q., Faridani, O. R., et al. (2012). Full-length mRNA-seq from single-cell levels of RNA and individual circulating tumor cells. *Nat. Biotechnol.* 30, 777–782. doi:10.1038/nbt.2282

Regev, A., Teichmann, S. A., Lander, E. S., Amit, I., Benoist, C., Birney, E., et al. (2017). The human cell atlas. *Elife* 6, e27041. doi:10.7554/eLife.27041

Robinson, M. D., McCarthy, D. J., and Smyth, G. K. (2010). edger: a bioconductor package for differential expression analysis of digital gene expression data. *Bioinformatics* 26, 139–140. doi:10.1093/bioinformatics/btp616

Rodriques, S. G., Stickels, R. R., Goeva, A., Martin, C. A., Murray, E., Vanderburg, C. R., et al. (2019). Slide-seq: A scalable technology for measuring genome-wide expression at high spatial resolution. *Science* 363, 1463–1467. doi:10.1126/science.aaw1219

Rosenberg, A. B., Roco, C. M., Muscat, R. A., Kuchina, A., Sample, P., Yao, Z., et al. (2018). Single-cell profiling of the developing mouse brain and spinal cord with split-pool barcoding. *Science* 360, 176–182. doi:10.1126/science.aam8999

Saelens, W., Cannoodt, R., Todorov, H., and Saeys, Y. (2019). A comparison of single-cell trajectory inference methods. *Nat. Biotechnol.* 37, 547–554. doi:10.1038/s41587-019-0071-9

Sedlazeck, F. J., Lee, H., Darby, C. A., and Schatz, M. C. (2018). Piercing the dark matter: bioinformatics of long-range sequencing and mapping. *Nat. Rev. Genet.* 19, 329–346. doi:10.1038/s41576-018-0003-4

Setty, M., Tadmor, M. D., Reich-Zeliger, S., Angel, O., Salame, T. M., Kathail, P., et al. (2016). Wishbone identifies bifurcating developmental trajectories from single-cell data. *Nat. Biotechnol.* 34, 637–645. doi:10.1038/nbt.3569

Smith, T., Heger, A., and Sudbery, I. (2017). Umi-tools: modeling sequencing errors in unique molecular identifiers to improve quantification accuracy. *Genome Res.* 27, 491–499. doi:10.1101/gr.209601.116

Spanjaard, B., Hu, B., Mitic, N., Olivares-Chauvet, P., Janjuha, S., Ninov, N., et al. (2018). Simultaneous lineage tracing and cell-type identification using crispr-cas9-induced genetic scars. *Nat. Biotechnol.* 36, 469–473. doi:10.1038/nbt.4124

Ståhl, P. L., Salmén, F., Vickovic, S., Lundmark, A., Navarro, J. F., Magnusson, J., et al. (2016). Visualization and analysis of gene expression in tissue sections by spatial transcriptomics. *Science* 353, 78–82. doi:10.1126/science.aaf2403

Stark, R., Grzelak, M., and Hadfield, J. (2019). RNA sequencing: the teenage years. *Nat. Rev. Genet.* 20, 631–656. doi:10.1038/s41576-019-0150-2

Stegle, O., Teichmann, S. A., and Marioni, J. C. (2015). Computational and analytical challenges in single-cell transcriptomics. *Nat. Rev. Genet.* 16, 133–145. doi:10.1038/nrg3833

Stoeckius, M., Hafemeister, C., Stephenson, W., Houck-Loomis, B., Chattopadhyay, P. K., Swerdlow, H., et al. (2017). Simultaneous epitope and transcriptome measurement in single cells. *Nat. Methods* 14, 865–868. doi:10.1038/nmeth.4380

Street, K., Risso, D., Fletcher, R. B., Das, D., Ngai, J., Yosef, N., et al. (2018). Slingshot: cell lineage and pseudotime inference for single-cell transcriptomics. *BMC Genomics* 19, 477. doi:10.1186/s12864-018-4772-0

Svensson, V., Natarajan, K. N., Ly, L.-H., Miragaia, R. J., Labalette, C., Macaulay, I. C., et al. (2017). Power analysis of single-cell RNA-sequencing experiments. *Nat. Methods* 14, 381–387. doi:10.1038/nmeth.4220

Svensson, V., Vento-Tormo, R., and Teichmann, S. A. (2018). Exponential scaling of single-cell RNA-seq in the past decade. *Nat. Protoc.* 13, 599–604. doi:10.1038/nprot.2017.149

Tan, Y., and Cahan, P. (2019). Singlecellnet: A computational tool to classify single cell RNA-seq data across platforms and across species. *Cell Syst.* 9, 207–213.e2. doi:10.1016/j.cels.2019.06.004

Tanay, A., and Regev, A. (2017). Scaling single-cell genomics from phenomenology to mechanism. *Nature* 541, 331–338. doi:10.1038/nature21350

Tang, F., Barbacioru, C., Wang, Y., Nordman, E., Lee, C., Xu, N., et al. (2009). mRNA-seq whole-transcriptome analysis of a single cell. *Nat. Methods* 6, 377–382. doi:10.1038/nmeth.1315

The Gene Ontology Consortium (2017). Expansion of the gene ontology knowledgebase and resources. *Nucleic Acids Res.* 45, D331–D338. doi:10.1093/nar/gkw1108

Tian, L., Su, S., Dong, X., Amann-Zalcenstein, D., Biben, C., Seidi, A., et al. (2018). scpipe: A flexible r/bioconductor preprocessing pipeline for single-cell RNA-sequencing data. *PLoS Comput. Biol.* 14, e1006361. doi:10.1371/journal.pcbi.1006361

Tian, Y., Carpp, L. N., Miller, H. E. R., Zager, M., Newell, E. W., and Gottardo, R. (2022). Single-cell immunology of sars-cov-2 infection. *Nat. Biotechnol.* 40, 30–41. doi:10.1038/s41587-021-01131-y

Ting, D. T., Wittner, B. S., Ligorio, M., Vincent Jordan, N., Shah, A. M., Miyamoto, D. T., et al. (2014). Single-cell RNA sequencing identifies extracellular matrix gene expression by pancreatic circulating tumor cells. *Cell Rep.* 8, 1905–1918. doi:10.1016/j.celrep.2014.08.029

Tirosh, I., Izar, B., Prakadan, S. M., Wadsworth, M. H., 2nd, Treacy, D., Trombetta, J. J., et al. (2016). Dissecting the multicellular ecosystem of metastatic melanoma by single-cell RNA-seq. *Science* 352, 189–196. doi:10.1126/science.aad0501

Traag, V. A., Waltman, L., and van Eck, N. J. (2019). From louvain to leiden: guaranteeing well-connected communities. *Sci. Rep.* 9, 5233. doi:10.1038/s41598-019-41695-z

Trapnell, C., Cacchiarelli, D., Grimsby, J., Pokharel, P., Li, S., Morse, M., et al. (2014). The dynamics and regulators of cell fate decisions are revealed by pseudotemporal ordering of single cells. *Nat. Biotechnol.* 32, 381–386. doi:10.1038/nbt.2859

Vallejos, C. A., Risso, D., Scialdone, A., Dudoit, S., and Marioni, J. C. (2017). Normalizing single-cell RNA sequencing data: challenges and opportunities. *Nat. Methods* 14, 565–571. doi:10.1038/nmeth.4292

Van den Berge, K., Perraudeau, F., Soneson, C., Love, M. I., Risso, D., Vert, J.-P., et al. (2018). Observation weights unlock bulk RNA-seq tools for zero inflation and single-cell applications. *Genome Biol.* 19, 24. doi:10.1186/s13059-018-1406-4

Van der Maaten, L., and Hinton, G. (2008). Visualizing data using t-sne. *J. Mach. Learn. Res.* 9, 2579–2605.

van der Wijst, M. G. P., Brugge, H., de Vries, D. H., Deelen, P., Swertz, M. A., Franke, L., et al. (2018).Single-cell RNA sequencing identifies celltype-specific cis-eqtls and co-expression qtls. *Nat. Genet.* 50, 493–497. doi:10.1038/s41588-018-0089-9

van Dijk, D., Sharma, R., Nainys, J., Yim, K., Kathail, P., Carr, A. J., et al. (2018). Recovering gene interactions from single-cell data using data diffusion. *Cell* 174, 716–729.e27. doi:10.1016/j.cell.2018.05.061

Vento-Tormo, R., Efremova, M., Botting, R. A., Turco, M. Y., Vento-Tormo, M., Meyer, K. B., et al. (2018). Single-cell reconstruction of the early maternal-fetal interface in humans. *Nature* 563, 347–353. doi:10.1038/s41586-018-0698-6

Vieth, B., Parekh, S., Ziegenhain, C., Enard, W., and Hellmann, I. (2019). A systematic evaluation of single cell RNA-seq analysis pipelines. *Nat. Commun.* 10, 4667. doi:10.1038/s41467-019-12266-7

Wang, E. T., Sandberg, R., Luo, S., Khrebtukova, I., Zhang, L., Mayr, C., et al. (2008). Alternative isoform regulation in human tissue transcriptomes. *Nature* 456, 470–476. doi:10.1038/nature07509

Wolf, F. A., Angerer, P., and Theis, F. J. (2018). Scanpy: large-scale single-cell gene expression data analysis. *Genome Biol.* 19, 15. doi:10.1186/s13059-017-1382-0

Wolf, F. A., Hamey, F. K., Plass, M., Solana, J., Dahlin, J. S., Göttgens, B., et al. (2019). Paga: graph abstraction reconciles clustering with trajectory inference through a topology preserving map of single cells. *Genome Biol.* 20, 59. doi:10.1186/s13059-019-1663-x

Wolock, S. L., Lopez, R., and Klein, A. M. (2019). Scrublet: Computational identification of cell doublets in single-cell transcriptomic data. *Cell Syst.* 8, 281–291.e9. doi:10.1016/j.cels.2018.11.005

Wu, A. R., Neff, N. F., Kalisky, T., Dalerba, P., Treutlein, B., Rothenberg, M. E., et al. (2014). Quantitative assessment of single-cell RNA-sequencing methods. *Nat. Methods* 11, 41–46. doi:10.1038/nmeth.2694

Xiong, K.-X., Zhou, H.-L., Lin, C., Yin, J.-H., Kristiansen, K., Yang, H.-M., et al. (2022). Chord: an ensemble machine learning algorithm to identify doublets in single-cell RNA sequencing data. *Commun. Biol.* 5, 510. doi:10.1038/s42003-022-03476-9

Young, M. D., Mitchell, T. J., Vieira Braga, F. A., Tran, M. G. B., Stewart, B. J., Ferdinand, J. R., et al. (2018). Single-cell transcriptomes from human kidneys reveal the cellular identity of renal tumors. *Science* 361, 594–599. doi:10.1126/science.aat1699

Zappia, L., and Oshlack, A. (2018). Clustering trees: a visualization for evaluating clusterings at multiple resolutions. *Gigascience* 7. doi:10.1093/gigascience/giy083

Zeisel, A., Muñoz-Manchado, A. B., Codeluppi, S., Lönnerberg, P., La Manno, G., Juréus, A., et al. (2015). Brain structure. cell types in the mouse cortex and hippocampus revealed by single-cell RNA-seq. *Science* 347, 1138–1142. doi:10.1126/science.aaa1934

Zeisel, A., Hochgerner, H., Lönnerberg, P., Johnsson, A., Memic, F., van der Zwan, J., et al. (2018). Molecular architecture of the mouse nervous system. *Cell* 174, 999–1014.e22. doi:10.1016/j.cell.2018.06.021

Zhao, J., Guo, C., Xiong, F., Yu, J., Ge, J., Wang, H., et al. (2020). Single cell RNA-seq reveals the landscape of tumor and infiltrating immune cells in nasopharyngeal carcinoma. *Cancer Lett.* 477, 131–143. doi:10.1016/j.canlet.2020.02.010

Zheng, G. X. Y., Terry, J. M., Belgrader, P., Ryvkin, P., Bent, Z. W., Wilson, R., et al. (2017). Massively parallel digital transcriptional profiling of single cells. *Nat. Commun.* 8, 14049. doi:10.1038/ncomms14049

Zou, J., Deng, F., Wang, M., Zhang, Z., Liu, Z., Zhang, X., et al. (2022). sccode: an r package for data-specific differentially expressed gene detection on single-cell RNA-sequencing data. *Brief. Bioinform.* doi:10.1093/bib/bbac180