Book Chapter

A Practical Introduction to Single-Cell RNA-seq in Immuno-Oncology

Benoît Aliaga*, Matthieu Genais and Vera Pancaldi

Centre de Recherches en Cancérologie de Toulouse, Université de Toulouse, Inserm, CNRS, Université Toulouse III-Paul Sabatier, France

***Corresponding Author:** Benoît Aliaga, Centre de Recherches en Cancérologie de Toulouse, Université de Toulouse, Inserm, CNRS, Université Toulouse III-Paul Sabatier, Toulouse, France

Published **October 31, 2023**

How to cite this book chapter: Benoît Aliaga, Matthieu Genais, Vera Pancaldi. A Practical Introduction to Single-Cell RNA-seq in Immuno-Oncology. In: Hussein Fayyad Kazan, editor. Immunology and Cancer Biology. Hyderabad, India: Vide Leaf. 2023.

Acknowledgment: We thank Chloé Bessière, and Malvina Marku for providing interesting comments and improvements to this manuscript.

Abstract

Single-cell approaches are a major revolution in biology. With this technology, it becomes possible to sequence a tumor's transcriptome and dissect tumor heterogeneity. Studying the interaction between heterogeneous cancer and immune cells beyond population averages becomes accessible. This approach

is promising to improve immuno-oncology treatments for patients. To exploit its full potential, biologists need to understand the steps needed to perform these experiments, the main scRNA-seq analysis pipeline components, and the frequently used tools available in the literature for further computational downstream analysis and interpretation. Throughout the chapter, we will guide the reader to several available libraries and packages that can be used to perform these analyses. Inference of intercellular communication will be further explored in the context of immuno-oncology at the end of the chapter.

Introduction

Since the birth of DNA sequencing, first performed by Fred Sanger and his group in 1977, sequencing technologies have been revolutionized several times [1]. The first technology (Sanger sequencing) uses the chain termination method, which generates DNA fragments that elongate at different points using dye-dideoxynucleotides. Electrophoresis is employed to separate DNA based on size. A laser scanner will provide an electropherogram, from which we can read the DNA sequence. This technique was widely used, and it remains frequently used nowadays. However, this method has some limitations. Indeed, it can only sequence short pieces of DNA (300 to 1000 bp), and the sequence quality degrades after 700 to 900 bases. Moreover, it has major limitations in cost and time. The second generation of sequencing, called “Next Generation Sequencing (NGS) Technologies” appeared at the beginning of the 2000s. With these new sequencers, it became possible to generate millions of short reads in parallel; sequencing was quicker than the Sanger method, could be achieved at a lower cost, and could be performed on smaller quantities of DNA. NGS opened new opportunities to decipher the genomes and to study the transcriptomes, which had up to then been studied using array-based technologies, limiting the quantification of transcripts only based on specific sequence probes distributed along the genome. Even the NGS technologies have some limitations: it is necessary to prepare amplified sequencing libraries before sequencing amplified DNA clones, with these steps being time-consuming and amplification libraries being expensive.

Moreover, there are still unresolved issues in sequencing complex genomes with many repetitive regions, due to the difficulty of assembling short reads. For these reasons, new sequencers came to the market with a new technology that aimed to further reduce the price of sequencing, and simplify the library preparation. These sequencers employ Single Molecule Sequencing Technology [2]. A few years ago, single-cell technologies slowly emerged [1], making it possible to sequence the transcriptome at the single-cell level and allowing us to study tissues in unprecedented detail. Tissue heterogeneity, detection of rare subpopulations, trajectory inference, gene regulatory network inference, and cell-cell communication inference are examples of what we can do with this technology [3].

Cancer involves uncontrolled proliferation of specific cells but it has become apparent that this process involves a complex ecological system of interacting cells. The tumor is composed of several cell types including normal cells, fibroblasts, immune cells, endothelial cells, adipocytes, and cancer cells, which interact in a surrounding environment rich in signaling molecules, and the extracellular matrix. During oncogenesis, cells are fed by the blood vessels, which give them the necessary nutrients for their growth. Intercellular communication has a fundamental role in homeostasis and also in cancer. This communication allows the recruitment and modulation of the stromal and immune cells, cell fate decisions, proliferation, and migration. This crosstalk inside the tumor and in the surrounding environment will promote angiogenesis, immune-escape, pre-metastatic niche formation, metastasis, and drug resistance [4,5]. Cell-cell communication (CCC) can occur either through direct cell interactions, mediated by gap junctions, cell adhesion, and intercellular bridges (tunnel nanotubes), or indirectly via the release of soluble factors, such as cytokines, growth factors, and chemokines. Extracellular vesicles are an important mode of communication between cancer cells and the tumor microenvironment (TME). The immune cells in the TME are abundant, varied in types, phenotypes, and states (CD4 and CD8 T lymphocytes, naive T lymphocytes, B lymphocytes, macrophages, NK cells, etc.). In fact, it was observed that tumor-infiltrating lymphocytes are tightly related to tumor growth and patient prognosis [6–9].

Tumor heterogeneity and the composition of its microenvironment are major causes of treatment failure and cancer resistance. For example, immune checkpoint blockers only work in at most 30% of the patients (and very often much less), and understanding how to predict patients' responses has become a real priority [10]. Relapse can be explained by the acquired resistance mechanisms present in the subclones, which have self-renewing characteristics. They will stay quiescent until the selective pressure of treatment or immune response is gone.

Descriptions of the TME have therefore become essential, and they can be obtained by either bulk technologies combined with deconvolution approaches or, more recently, by applying single-cell approaches. Single-cell technology allows us to study tumor heterogeneity, as well as cell-cell communication involving all the cells in the TME.

Single-cell sequencing technologies provide a large amount of data and require bioinformatic skills and knowledge to design scRNA-seq experiments and analyze them appropriately. The raw data produced are not exploitable immediately and need to be pre-processed before being used to address biological questions. This chapter will give an overview of different aspects of computational analysis of single-cell RNA-seq datasets. We will describe the different steps from designing scRNA-seq experiments to pre-processing and analysis of the data, mentioning which tools are available for the different steps of the analysis. The list of methods and tools will not be exhaustive, as this is a field in constant expansion and new ones are frequently produced, but we hope this chapter will serve as a helpful introduction to the topic and allow interested researchers to identify more complete resources to acquire deeper knowledge or more detailed information.

Designing Single-Cell Experiments and Choosing the Best Single-Cell Technologies

Before performing single-cell sequencing, it is necessary to carefully design the experiment in a way that will ensure that data analysis will generate robust and trustworthy results. This part is crucial to reduce the technical noise and to objectively measure the biological effect that we want to study.

Firstly, it is important to make a clear plan considering the following points:

- The project's goal
- The biological question and its hypothesis
- The budget and time allocated to the project
- The tissue under study
- The number of samples, replicates, and cells that we expect to consider in our experiment
- Whether we are interested in studying gene expression, alternative splicing, or also in identifying rare cell subpopulations

Secondly, it is important to choose the right technology to answer the biological question, and sometimes, this can be difficult because several technologies have been developed (table 1). Currently, two main technologies are used in scRNA-seq: plate-based and droplet-based.

- For the plate-based, fluorescence-activated cell sorting (FACS) is necessary to deposit one cell in each well (plate 96 or 384) containing a hypotonic lysis buffer Triton-X100 where mRNA is separated [11]. After that, cell barcodes are added to each well, and libraries are made. The major advantage of these methods is that they can sequence full-length transcripts, so they can be used to study structural variations such as RNA fusion, mutations in transcripts, and detection of pseudogenes and splice variants at the single-cell level [12]. Smart-seq techniques are the most used plate based method.
- Droplet-based methods use microfluidics which allows the fabrication of devices with microchannels handling very small quantities of liquid in micro volumes. In scRNA-seq, isolated cells encapsulate both barcode-containing beads/hydrogels, and unique molecular identifiers (UMIs), such that after pooling, and sequencing, each read can be mapped back to its cell of origin. Drop-seq and Chromium by 10X are examples of these methods.

Several studies compared the two technologies and showed that plate-based methods allow more detection of genes per cell if we compare them with the droplet method. But these last methods quantified mRNA levels with less amplification noise due to the use of UMI [13]. In addition, Chromium detects more cell clusters than Smart-seq2, which on the other side detects more genes than Chromium [14]. A comparison of droplet methods like Chromium and drop-seq showed that Chromium has higher molecular sensitivity, and precision, and less technical noise [15].

In conclusion, plate-based methods must be considered if the main aim is to identify rare cell subpopulations or structural variation. On the contrary, if the goal is to study the heterogeneity of the tissue, Chromium is adequate.

Thirdly, we can start designing the experiment to ensure that we reduce the technical noise, which could have several origins, also depending on the single-cell technology used (batch effect, amplification bias, dropout, etc). This noise can confound downstream analysis and can be dealt with using two approaches:

- If we do a **balanced design**, samples, and replicates are sequenced in the same lanes on the flow cell (same conditions). Thus, it becomes possible to compare them and to be sure of the origin of the variation [16,17]. However, it is not always possible to do a balanced design, and in some cases, we do not have the choice to do a confounded design.
- In a **confounded design**, the samples and replicates are separated from the others (different lanes and flow cells), thus when we compare the measure between them, it becomes difficult to identify the source of the biological variation. In this case, several statistical methods exist to correct batch effects.

Table 1: the different single-cell technologies.

Technology	Isolation	Capacity (# of cells)	Coverage	UMI or spike-in	Advantages	Disadvantages	Year	Reference
Smart-seq	FACS	96 plates or 384 plates	Full-length	spike-in			2012	[18]
Smart-seq2	FACS	96 plates or 384 plates	Full-length	spike-in	● detect more genes ● alternative splicing	● capture a high proportion of mitochondrial genes ● cost	2013	[11,19]
Drop-seq	FACS		3’	UMI	● most cost-effective ● customizable		2015	[20]
Chromium	Droplet-based	1,000-10,000 cells	3’ or 5’	UMI	● cost	● Higher noise for mRNA with low expression levels ● dropout problem	2016	[21]

Data Analysis Pipeline Overview

Once the sequencing has been performed, a thorough data analysis will be key to extract biological information. This analysis will be divided into 3 steps (figure 1). The two first steps are common, and the third depends on the analysis goals defined in the experimental design.

- **Step 1:** *pre-processing*. In this stage, the sequence quality needs to be checked, and if it is not good, the sequences need to be trimmed. After that, the sequences will be aligned with the genome reference; if the mapping score is good, the analysis will follow step 2.
- **Step 2:** *main analysis*. This step is crucially dependent on the experimental design and can be divided into several subsets. For example, if it is a balanced or confounded design, it is necessary to correct the batch effects. The final result will be biased if an appropriate statistical method is not applied. This stage will close by the clustering, revealing the different subpopulations in the tissue/tumor.
- **Step 3:** can be denoted as *functional analysis*, depending on the biological question. It involves studying expression profiles at the gene or cell level. At the gene level, it is possible to study differential gene expression in different conditions (treated or not, for example) or use various tools for inferring the gene regulatory networks and identify pathways that are differentially enriched (through functional enrichment analysis). The trajectory inference (or pseudo-time) and the cell-cell interaction inference are examples of common analysis at the cell level.

In the following, we will provide an overview of different software frequently used for carrying out these steps.

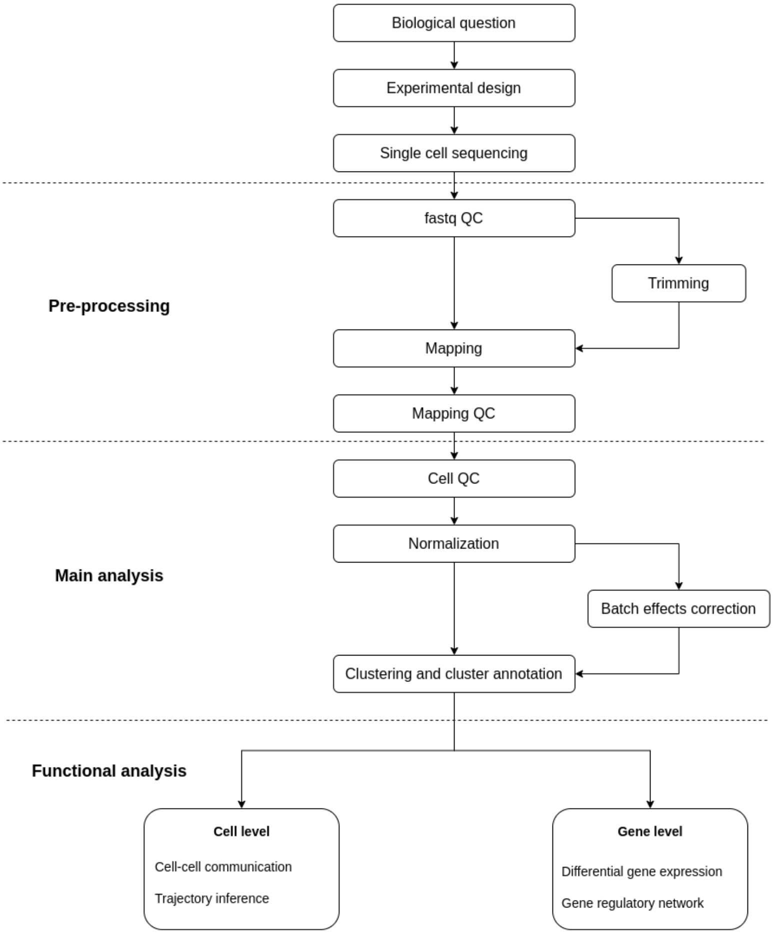

Figure 1: the different steps of single-cell analysis.

Pre-Processing

Raw Quality Check & Trimming

As for bulk RNA-seq, the scRNA-seq analysis starts by checking the quality of the raw sequences. Several phenomena, like sequencing errors, PCR artifacts, and contaminations, can degrade the final sequencing result during the wet lab part and sequencing. To detect them, tools will check the presence of adaptors, the GC content, duplicated reads, and overrepresented k-mers (explained in the mapping and quality check paragraph). It is well known that the sequencing quality decreases at the 3' end of the reads. Thus, these bases must be removed to improve the mappability. FastQC is the most used tool for this part, and it computes some statistics about the composition and quality of raw sequences [22]. These statistics include the following:

- Summary statistics
- Distribution of per-base sequence quality
- Distribution of quality scores per sequence
- Distribution of per-base N content
- Sequence length distribution
- Sequence duplication
- Distribution of overrepresented sequences

With an automated pipeline, it will become easy to run FASTQC on a large number of samples. But, the FASTQC reports are not easy to compare between them. With MultiQC, it becomes easier to compare the FASTQC reports and interpret them [23]. Moreover, MultiQC will generate an HTML file report. Several tools, such as Trimmomatic [24] and Cutadapt [25], exist to discard the low-quality reads, trim adaptor sequences, and detect contamination, and poor-quality bases.

Mapping and Quality Check

After doing the raw quality check and adaptor trimming, mapping is the next step. By definition, read mapping assigns each read to a specific location in the genome. As explained in the different *single-cell technologies*, in Chromium and droplet technology, we have three important objects, (i) a cDNA fragment that identifies an RNA transcript, (ii) a cell barcode for each cell, and (iii) a unique molecular identifier (UMI). Mapping of reads includes four steps:

- Aligning the reads to a reference genome
- Assigning reads to genes
- Cell barcode demultiplexing (allocate each read to a specific cell)

- UMI deduplication (count the number of unique RNA molecules)

These four steps will produce a cell expression matrix, which contains the counts of RNA molecules in each cell for each gene. Several tools have been developed for bulk and single cell RNA-seq. Thus, it is not easy to choose a good aligner for the analysis, and benchmarking studies could be helpful. Here, we will present the main read mapping software used in scRNA-seq, which could be useful in immuno-oncology projects.

Two main approaches exist for alignment. The first tools use splice-aware aligner algorithms. Genes in the human genome contain a lot of introns, and coding sequences are short. Thus, it becomes difficult to properly align reads to the genome. For example, reads can be mapped entirely within an exon or can be spanning two or more exons [26–28]. To overcome this difficulty, splice-aware aligners have been developed. They used the annotation file (GFF/GTF). In this way, STAR can detect the splice junctions and correctly map the read to the reference genome. HISAT2 [29], STARsolo [30], and CellRanger (10X read mapping software) are the most used splice-aware alignment tools in scRNA-seq.

The other methods are based on pseudo-alignment algorithms, which include four steps. Firstly, the reference transcriptome will be split into *k*-mers, and a De Bruijn graph will be constructed. *k*-mers are unique length *k* subsequences of a sequence. A De Bruijn graph is a directed graph in which vertices are *k*-mers, and edges represent overlaps between the *k*-mers. Through the graph, a path represents a sequence [31]. Secondly, the RNA-seq reads will be converted into *k*-mers. Thirdly, the software will use the *k*-mers to assign reads to a transcript or several transcripts. Finally, the reads will be counted for each transcript or for each gene. In single-cell two common tools use pseudo-alignment strategy: kallisto/BUStools [32] and Salmon/Alevin/Alevin-fry [33].

In bulk RNA-seq, STAR is one of the top-performing read mapping tools [26,34]. When 10X developed their technology, they also wrote a read mapping software derived from STAR, CellRanger, which is one of the most used software in the literature. This software uses STAR to perform the alignment while the transcript quantification part is done by the 10X proprietary algorithms. Alexander Dobin, the STAR's developer, decided to develop an extension of STAR, called STARsolo. CellRanger and STARsolo produce similar results [30]. Unlike CellRanger, STARSolo can take into account multi-gene reads (transcripts that align well to two or more genes), which is important to detect different classes of biologically important genes (e.g. paralogs) [30]. As written above, STARSolo uses the annotation file to recognize the splice junctions and to detect the spliced/unspliced transcripts. This information is important to perform RNA velocity studies to reconstruct pseudo-temporal trajectories of cell phenotypes starting from a cell mixture. STAR showed a better alignment rate and measured more abundance of the gene compared to Kallisto/BUStools [34]. By comparing 10x PBMC 3K data clustering results (annotation of cell types), the pipeline which used STAR and Kallisto annotated the same cell types but one cell type was lacking with Kallisto [34]. Kallisto has the advantage of being 4 times faster than STARSolo and the memory usage is 7.7 times less than the previous one [34].

Alignment files (bam) can contain biases, which are introduced during sequencing, sample preparation, and/or mapping algorithm. Thus, checking the quality of the read alignment is an important step. Thus, we will have an idea about the read alignment to the human genome and if the data fit with the expected outcome. The percentage of reads mapped to the reference genome (human) is a global indicator of the overall sequencing accuracy. A percentage above 90% in all samples indicates a very good mapping rate. Usually, we expect between 70 and 90% of reads mapped on the human genome. We also expect a small fraction of reads to map to multiple regions in the genome (multi-mapping reads). After these first read mapping quality statistics, we check to see where the reads are mapped. We expect more than 60% of reads in the exonic regions and between 20-30% in the intronic regions. If an equal distribution of reads mapping to intronic, exonic, and intergenic regions is present, this could be a sign of DNA contamination since mRNA from introns is normally quickly degraded.

Table 2: the different mapper frequently used in single-cell analysis.

Aligner	**Strategy**	**Advantage**	**Disadvantage**	**References**
STARsolo	Splice-aware	Precise	Slow	[30]
HISAT2	Splice-aware	Fast		[29]
CellRanger	Splice-aware	User-friendly	Proprietary software	No publication
Kallisto/BUStools	Pseudoalignment	Fast		[32]
Salmon/Alevin/Alevin-fry	Pseudoalignment	Fast		[33]

Main Analysis

Cell Quality Check

After ensuring that mapping quality is good, it is important to remove low-quality cells which can bias the analysis. In single-cell data, some metrics are used for quality control:

- the number of UMIs per cell, which represents the number of transcripts per cell
- the number of *features,* which represents the number of detected genes per cell
- the mitochondrial ratio, giving the percentage of reads coming from mitochondrial genes per cell (representing the living status of the cell)

In the literature, researchers use thresholds to filter the low-quality cells. The number of UMIs should be above 500 to have enough transcripts per cell and at least 250 genes must be detected per cell.

Traditionally, we see a high mitochondrial ratio in low UMIs and a low number of genes in cells, showing dying/damaged cells. A threshold <0.2 for this ratio is used to remove top damaged cells (except if high mitochondrial gene expression is expected in the experiment).

This first step of quality control was for cells, but more quality checks must be performed at the gene level. For example, we can remove genes that are expressed in less than 10 cells. This way we will keep living cells and expressed genes. Some literature is available to understand these thresholds [35,36]. Some data-driven methods exist to avoid choosing thresholds as they are often arbitrarily chosen [38]. Some code in R is available to be guided through the steps of quality control for single-cell experiments [1].

A possible additional step for quality control would be to remove doublets from the experiment. Technically, doublets are generated when two cells merge due to errors in cell sorting or capture, more often in droplet-based experiments. A benchmark for doublet detection was performed identifying DoubletFinder as the best in detection accuracy and also in computational efficiency (memory usage+time) [38].

Normalization

As written in the introduction, the experimental design and the sequencing can generate several technical biases. The variability in sequencing depth might be increased by technical factors like sequencing depth, amplification, gene length, and GC content. But they are not the only source of unwanted variation. The amount of RNA per cell can vary between cell cycle stages [39,40]. Hence, it becomes difficult to untangle the biological differences from the technical ones between samples. The goal of the normalization is to eliminate/reduce these technical biases so that we can preserve the biological signal in our transcriptomic data. Bulk RNA-seq developed normalization methods. However, these methods are not suitable for single cell transcriptomics. Indeed, scRNA-seq generates abundant zero-expression values [41]. If bulk normalization methods are used in scRNA-seq, it may be a source of overcorrection for lowly expressed genes [42,43]. To avoid this problem, specific normalization methods for scRNA-seq have been developed. They are based on the scRNA-seq technologies that have been developed: plate based which uses spike-ins and droplet technology which uses UMIs. Using a technology that uses UMI can reduce technical biases.

scRNA-seq normalization methods are divided into two steps: scaling and transformation. The aim of scaling is to apply a size factor to scale data. In other words, all counts for each cell are divided by a cell-specific factor. The main hypothesis is that the bias affects all genes equally with the expected mean count for that cell. By dividing the counts by the size factor, we can remove the bias. To conclude, the size factor for each cell represents the estimate of the relative bias [44]. Then, the number of counts becomes comparable across cells and is less related to technical variation. The goal of transformation is to reduce the skewness in the distribution of the normalized values. Log transformation is often used for this step.

The first method of normalization is LogNorm which is the default method in the Seurat package for scRNA-seq analysis [45]. The idea of this method is to measure the gene expression for each cell is normalized over the total expression. Thus, we can eliminate the effect of the sequencing depth variation between cells. Firstly, we compute the normalized gene expression value (x_i) of gene X in cell *i* (Eq. 1). The transformation is performed by the log (Eq. 2). scRNAseq has a lot of zero-expression data values and to avoid zero counts, in Eq. 2, we add 1.

$$x_i = \frac{The\ read\ count\ of\ gene\ X\ in\ cell\ i}{Total\ of\ counts\ of\ cell\ i} \times 10^4 \qquad \text{(Eq. 1)}$$

$$f(x_i) = ln(x_i + 1) \qquad \text{(Eq. 2)}$$

The LogNorm method is a global scale factor because it is applied on all genes. However, if this is not the case, these methods may fail to detect true differential expressed genes. Genes with weak to moderate expression tend to get overcorrected, while genes with high expression get undercorrected. To avoid this problem, two other methods have been developed.

SCnorm is a method [46] that uses quantile regression to estimate the dependence of transcript expression on sequencing depth for every gene. Genes with similar dependence are then grouped, and a second quantile regression is used to estimate scale factors within each group. Within-group adjustment for sequencing depth is then performed using the estimated scale factors to provide normalized estimates of expression. However, this method has a major problem. As written above scRNA-seq has a lot of zero-expression values and this issue is not taken into account by this method.

As for LogNorm, sctransform is a method developed in the Seurat package [47]. The authors proposed a novel statistical approach for the modeling, normalization, and variance stabilization of UMI count data for scRNA-seq. They observed a linear relationship between UMI counts and the number of genes detected in a cell. They showed that different groups of genes cannot be normalized by the same constant factor, representing an intrinsic challenge for scaling-factor-based normalization schemes, regardless of how the factors themselves are calculated. This method has three steps. Firstly, sctransform fits a generalized linear model (GLM) for each gene with UMI counts as the response variable and sequencing depth as the explanatory variable. This model describes the influence of technical noise on UMI counts. Secondly, sctransform uses the model parameter values and gene mean to learn global trends in the data. Thus, it is possible to perform independent regularizations for all parameters. Thirdly, the regularized regression parameters are used to define an affine function that transforms UMI counts into Pearson residuals. These residuals will inform us on how much the count is far from the true mean expression.

Batch Effect Correction & Integration

Very often people confuse normalization, batch effect, and data integration. These notions are different steps of pre-processing, but they are essentially different. To clarify, the goal of normalization is to target the variance from sequencing, like library preparation, amplification bias caused by gene length, GC content, *etc* [48]. Normalization is applied to the count matrix. This step does not correct the other sources of unwanted variation, which could stem from experimental design (sequencing platforms, sequencing lane, timing, reagents for example) and should be removed with batch effect correction [49]. We must distinguish three cases. Firstly, the correction of the samples from the same experiment. Secondly and thirdly, the correction between experiments performed in the same laboratory or between datasets from different laboratories. For the last two, we will need to perform data integration, which combines data from different sources and provides users with a unified view of them [50].

Several software has been developed to correct batch effects and to integrate data and it is based on three broad strategies:

- Regression-based correction
- Joint dimensionality reduction
- Joint dimensionality reduction and graph-based joint clustering

The first strategy uses regression-based correction. ComBat is a software that uses this strategy and it was the first method written to correct batch effects in microarray and bulk RNA-seq [51]. At the beginning of single-cell analysis, ComBat was used, but quickly, three main pitfalls were detected. Firstly, it does not account for differences in population composition. Secondly, it assumes the batch effect is additive. Thirdly, it is prone to overcorrection (in case of partial confounding). This method works well in small-medium datasets like microarray with similar cell type composition. Otherwise, it will fail in a large dataset with a complex mixture of cell types [49,50,52]. Another method must be used to correct the batch effect in scRNA-seq.

The second strategy uses joint dimensionality reduction (jDR). By definition, dimensionality reduction includes numerous methods for transforming a high-dimensional space, with a lot of variables or features, into a low-dimensional one, with few variables or features. This transformation will preserve the characteristic and/or structure of the data. These methods are applied to a dataset individually. In bioinformatics, we can have several datasets in our experiments. Joint dimensionality reduction will allow us to transform several data sets in a low dimensional space while preserving the specificities of each dataset. In other words, jDR methods use existing dimensionality reduction methods to apply multiple data sets [53,54].

Harmony is an example of a tool that uses the jDR strategy [55]. Firstly, Harmony will perform a Principal Component Analysis (PCA) to integrate the cells in low dimensional space and assign them to clusters. Secondly, the algorithm will compute the cluster centroids for each dataset. Thirdly it will apply a correction factor for each cluster. Finally, cells are rearranged into the cluster from the last correction. This workflow will be repeated until convergence is obtained, meaning that additional training will not improve batch correction. Mutual Nearest Neighbors (MNN) is an algorithm to correct batch effects with jDR strategy.

The MNN algorithm is inspired by the idea of K-Nearest Neighbors (KNN). This algorithm has 2 main assumptions [49].

Firstly, there is at least one cell population that is present in both batches. Secondly, the batch effect variation is much smaller than the biological effect variation between different cell types. The method tries to find the most similar cells (mutual neighbors) between the batches. Then, the algorithm will measure the difference between batches to quantify how strong the batch effect is. This information is used to scale the counts for the rest of the cells in the batches.

Seurat uses Canonical Correlation Analysis (CCA) [56]. In this method, the data from the batches are projected into a low-dimensional space. The algorithm maximizes correlation (or covariance) between the data sets from different batches. The dataset projections are correlated but they do not overlap well in low dimensional space. This problem is fixed with the Dynamic Time Warping (DTW) algorithm, which compares the similarity or calculates the distance between two or more arrays with different lengths. CCA data projection will be stretched and squeezed to align well between them.

These tools are frequently used in scRNA-seq analysis in immuno-oncology. Several studies compared and evaluated their efficiency. By testing different tools with five scenarios of batch effect correction and several datasets, Tran *et al*, showed that Harmony, and Seurat achieved good scores. On the other hand, Combat was the worst-performing method [57].

Feature Selection & Dimensionality Reduction

In data science, we often work with high-dimensional data. The dimension of a dataset corresponds to the number of attributes/features that exist in a dataset. For example, a table with 2 columns is a 2-dimensional dataset, which can be represented by a 2D plot. If we add another dimension, we will obtain a 3-dimensional space and a 3D plot. We can add as many dimensions as we want. High-dimensional datasets are common in genomics [58–61]. Having a high dimensional dataset leads to difficulties during analyses and visualization, leading to what is currently referred to as the ‘Curse of Dimensionality’. In scRNA-seq, the datasets have a high-dimensional space with *N* *(tens to hundreds normally)* samples, *M (thousands)* genes, and *P (thousands)* cells. This requires a lot of computational time, while some algorithms struggle with many dimensions. Dimensionality reduction (DR) describes the techniques that transform the data from a high-dimensional space into a low-dimensional space to overcome these difficulties. They are divided into two different groups: (i) linear and (ii) non-linear. In linear methods, the output (low dimension) of the system is proportional to the input (high dimension). This proportionality is achieved by the linear projection of the original data onto a low-dimensional space. In the case of non-linear methods, the output of the system is not proportional to the input. In scRNA-seq, we can use both methods.

Principal Component Analysis (PCA) is a linear method, which is a commonly used DR method. The PCA algorithm will find the first principal component with the largest variance in the data. Thus, it will seek the second component with the largest variance which is not correlated to the first component. This process will be repeated until the component reaches a threshold defined by the users.

The *t*-Stochastic Neighborhood Embedding (*t*-SNE) algorithm is a non-linear method for DR [62]. This algorithm is divided into 3 steps. Firstly, the algorithm will convert the Euclidean distances of a high dimensional space into a conditional probability that represents similarities. Secondly, the algorithm will create a low-dimensional space where the data will be represented, but on which we do not know the coordinates of our points. We are therefore going to randomly distribute the points over this new space. The rest is quite similar to the first step, we calculate the similarities of the points in the newly created space, but using a t-Student distribution and not Gaussian. Thirdly, to faithfully represent the points in the lower dimensional space, we would ideally like the similarity measures in the two spaces to be consistent. We, therefore, need to compare the similarities of points in the two spaces using the Kullback-Leibler (KL) measure.

The Uniform Manifold Approximation and Projection (UMAP) algorithm is based on three assumptions about the data. Firstly, the data are uniformly distributed on the Riemannian manifold. Secondly, the Riemannian metric is locally constant, and finally, the manifold is locally connected. According to these assumptions, the manifold with fuzzy topology can be modeled. The UMAP algorithm has two main stages. The first stage involves constructing a weighted graph that encodes the local structure of the data. This is done by selecting a set of "landmark" points in the high-dimensional space and then calculating the distances between each point and its nearest neighbors. The distances are used to construct a weighted graph, where the nodes represent the data points and the edges represent the distances between them. The second stage involves finding a low-dimensional representation of the data that preserves the global structure of the graph. This is done by minimizing a cost function that measures the difference between the distances in the original high-dimensional space and the distances in the low-dimensional space. This optimization problem is solved using a technique called "stochastic gradient descent," which involves iteratively updating the low-dimensional representation in a way that reduces the cost function. UMAP has superior run-time performance compared with the *t*-SNE [63,64].

Clustering and Cell Type Annotation

These methods presented previously gather a set of learning algorithms whose goal is to group unlabeled data with similar properties. Thus, we obtain a cluster of different groups of cells whose cell types are unknown. We then need to assign cell types for each group. This step is a critical feature of scRNA-seq. Several tools Seurat [45], Monocle 3 [65], SCENIC [66] perform clustering with DR methods (t-SNE, UMAP) and cell type identification. In this step, we can identify rare cell types or subpopulations. In order to improve this identification, new tools, like scClassify [67], SingleCellNet [68], and Sincell [69] have been developed. Once the cell type assignment is done, we can start the downstream analysis, like cell trajectories or cell-cell communication inference.

An Example of Downstream Analysis: Cell-Cell Communication Inference and Analysis in TME

After clustering annotation, the *downstream analysis* will allow us to extract biological insights from the scRNA-seq data. As written in *Data analysis pipeline overview*, this analysis can be divided into two parts, which are cell- and gene-level. The cell-level analysis will use methods to characterize cellular structure like trajectory inference and cell-cell communication, while gene-level with differential gene expression and gene regulatory network will investigate molecular signals in the data. At the gene-level, with the differential gene expression approach ask the question is whether any genes are differentially expressed between two experimental conditions. Gene regulatory networks with scRNA-seq is the second method that we could perform at this level and it will be explained in the next chapter. At the cell-level, the clustering annotation cannot describe the whole cellular diversity. The observed heterogeneity is under continuous biological processes. By using trajectory inference methods, which use dynamic models of gene expression, it becomes possible to capture transitions between cell identities, and branching differentiation processes for example. Cell-cell communication (CCC) is the second type of analysis that can be performed at this level and as explained in the introduction, the rise of tools for inference of cell-cell communication from scRNA-seq has advanced the development of cancer immunotherapies. Here, we will explore more deeply CCC analysis with scRNA-seq data.

Cell-cell communication, also known as cell-cell interaction or intercellular communication, is essential for the development of multicellular systems [70]. Cells are able to receive and process many signals simultaneously which are from their immediate environment. But cells also send out messages to other cells close or far away. This intercellular communication requires coordination by soluble factors, associated membrane proteins, exosomes, and gap junction channels, for example. In the past, researchers thought that CCC was lost in cancer because cancer cells are disconnected from healthy cells but it is likely that communication is changed but is not lost. For example, in melanoma, malignant cells can deliver exosomes that create an environment for tumor cells to survive [71]. With scRNA-seq it's possible to infer CCC between TME, immune, and cancer cells. Several tools have been developed to infer CCC from scRNA-seq data and it can be difficult to choose a tool for our analysis. In this paragraph, we will explain the main ideas behind these tools, and after we will compare some of them.

All these CCC inference tools share a common input, the count matrix, which contains the transcript levels of each gene across different samples and cells. At the same time, the known interacting protein or ligand-receptor pairs in specialized databases like KEGG and Reactome are collected. This information is used to filter the count matrix, which will contain only the genes associated with the interacting proteins. This filtered table will be used for the CCC analysis which is divided into three steps [72]. Firstly, the expression levels of ligand-receptors pairs are used as inputs to compute a communication score by using a scoring function (function f(L, R), where L and R are the expression values of the ligand and the receptor). Secondly, an aggregation function will compute the communication scores between samples or cells. In the third step, the communication and aggregation scores are used to generate different graphics, like hierarchical and circle plots or network visualizations, which will facilitate the interpretation of the results. In this chapter, we do not explain the mathematical methods to compute the computation and aggregation score, but we refer the reader to a comprehensive review [73].

The inference of cell-cell communication from scRNA-seq data can help us understand the signaling alterations provoked by immune checkpoint blockers in the TME [74]. For example, CellPhoneDB was the first tool developed for this aim and became one of the most used tools in CCC [75,76]. In hepatocarcinoma and esophageal squamous carcinoma, CellPhoneDB found a potential reprogramming interaction from tumor cells to macrophages by the SPP1-CD44 axis [77,78]. This axis is involved in an immune checkpoint. Several studies with patients used CellPhoneDB to characterize CCC in ICB resistance and response. They identified enhanced signaling of

HAVCR3-LGALS9 (TIM3-Galectin9) in CD8+ T cells in non-responding and resistant patients [79–81]. CellPhoneDB is able to highlight the intercellular communication between immune cells and cancer cells, but it has some limitations. Indeed, CellPhoneDB takes into account only ligands and receptors, while it is known that some signaling cofactors in the sender or receiver cells can influence intracellular pathways. Alternative software, like CellChat, and NicheNet was developed to take this point into account. CellChat was used and showed promising results in different immunotherapy signaling studies [77,82–84]. NicheNet [85] is widely used when researchers want to investigate intercellular communication in the TME [74].

Conclusion: The Future of scRNA-seq in Immuno-Oncology

Single-cell RNA-seq is revolutionizing our perspective on the tumor microenvironment and driving innovative approaches in immuno-oncology research [86–88]. In this chapter, we have described the main parts of the computational analysis and given examples of downstream analysis. Understanding the different steps of this analysis is important for generating valuable experiments and trustworthy results. Unfortunately, single-cell analysis requires statistical knowledge and programming skills and can be difficult for biologists. In order to make scRNA-seq more accessible to a broader community of researchers and/or clinicians, some pipelines have been developed. scAmpi and Bullito are automated, flexible, and parallelizable pipelines [89,90]. These pipelines include all the steps and software described above. The main difference between the two is that scAmpi has been developed for clinical applications. pipeCom is a flexible R framework for pipeline comparison, which then chooses the best among the various tools [91].

As with everything, there are some limitations of these approaches that should be considered. For example, isolation of cells from solid tissues can introduce biases on the number of cells of each type that is captured and included in the data, so it is not advisable to assume that cell numbers obtained in scRNAseq experiments are directly proportional to those effectively present in the tissue. Also, defining cell types can be done based on expression of proteins on cell surfaces, which are not necessarily strongly correlated to the levels of the corresponding mRNAs. For this reason, additional technologies such as CITEseq [92] and INs-Seq [93] provide a multi-omic view of cells, detecting both cell surface proteins and transcripts on each cell. Similarly, the combination of transcriptomics with the identification of open chromatin regions can currently be performed on the same cell, like NEAT-seq [94] and smart3-seq [95]. It is also possible to perform genome and transcriptome single-cell approaches like G&T-seq [96] and scTrio-seq [97].

In conclusion, single-cell approaches can be helpful to study the TME, but the spatial information and context of the cells are lost. Recently the development of spatial omics techniques at the single-cell level and computational methods to analyze them are revolutionizing biology once again [98]. Nowadays, it is possible to combine spatial transcriptomics with scRNA-seq datasets to infer spatial cell-cell communication [99–102]. With the progress of machine learning, new computational methods will be developed to integrate these datasets to understand how cell-cell communication influences the fate of cells in tumors.

References

1. Heather JM, Chain B. The sequence of sequencers: The history of sequencing DNA. Genomics. 2016; 107: 1–8.
2. Bleidorn C. Third generation sequencing: technology and its potential impact on evolutionary biodiversity research. Syst. Biodivers. 2016; 14: 1–8.
3. Dal Molin A, Di Camillo B. How to design a single-cell RNA-sequencing experiment: pitfalls, challenges and perspectives. Brief. Bioinform. 2018; 1–11.
4. Roghayyeh Baghban, Leila Roshangar, Rana Jahanban-Esfahlan, Khaled Seidi, Abbas Ebrahimi-Kalan, et al. Tumor microenvironment complexity and therapeutic implications at a glance. Cell Commun. Signal. 2020; 18: 59.
5. Salemme V, Centonze G, Cavallo F, Defilippi P, Conti L. The Crosstalk Between Tumor Cells and the Immune Microenvironment in Breast Cancer: Implications for

Immunotherapy. Front. Oncol. 2021; 11: 610303.
6. Brummel K, Eerkens AL, De Bruyn M, Nijman HW. Tumour-infiltrating lymphocytes: from prognosis to treatment selection. Br. J. Cancer. 2023; 128: 451–458.
7. Fridman WH, Pagès F, Sautès-Fridman C, Galon J. The immune contexture in human tumours: impact on clinical outcome. Nat. Rev. Cancer. 2012; 12: 298–306.
8. F Pagès, J Galon, MC Dieu-Nosjean, E Tartour, C Sautès-Fridman, et al. Immune infiltration in human tumors: a prognostic factor that should not be ignored. Oncogene. 2010; 29: 1093–1102.
9. Zhu N, Hou J. Assessing immune infiltration and the tumor microenvironment for the diagnosis and prognosis of sarcoma. Cancer Cell Int. 2020; 20: 577.
10. L Castelo-Branco, G Morgan, A Prelaj, M Scheffler, H Canhão, et al. Challenges and knowledge gaps with immune checkpoint inhibitors monotherapy in the management of patients with non-small-cell lung cancer: a survey of oncologist perceptions. ESMO Open. 2023; 8: 100764.
11. Simone Picelli, Omid R Faridani, Åsa K Björklund, Gösta Winberg, Sven Sagasser, et al. Full-length RNA-seq from single cells using Smart-seq2. Nat. Protoc. 2014; 9: 171–181.
12. Victoria Probst, Arman Simonyan, Felix Pacheco, Yuliu Guo, Finn Cilius Nielsen, et al. Benchmarking full-length transcript single cell mRNA sequencing protocols. BMC Genomics. 2022; 23: 860.
13. Christoph Ziegenhain, Beate Vieth, Swati Parekh, Björn Reinius, Amy Guillaumet-Adkins, et al. Comparative Analysis of Single-Cell RNA Sequencing Methods. Mol. Cell. 2017; 65: 631-643.e4.
14. Wang X, He Y, Zhang Q, Ren X, Zhang Z. Direct Comparative Analyses of 10X Genomics Chromium and Smart-seq2. Genomics Proteomics Bioinformatics. 2021; 19: 253-266.
15. Xiannian Zhang, Tianqi Li, Feng Liu, Yaqi Chen, Jiacheng Yao, et al. Comparative Analysis of Droplet-Based Ultra-High-Throughput Single-Cell RNA-Seq Systems. Mol. Cell. 2019; 73: 130-142.e5.
16. Baran-Gale J, Chandra T, Kirschner K. Experimental design

for single-cell RNA sequencing. Brief. Funct. Genomics. 2018; 17: 233–239.

17. Hicks SC, Teng M, Irizarry RA. Missing data and technical variability in single-cell RNA-sequencing experiments. Biostatistics. 2018; 562–578.
18. Goetz JJ, Trimarchi JM. Transcriptome sequencing of single cells with Smart-Seq. Nat. Biotechnol. 2012; 30: 763–765.
19. Simone Picelli, Åsa K Björklund, Omid R Faridani, Sven Sagasser, Gösta Winberg, et al. Smart-seq2 for sensitive full-length transcriptome profiling in single cells. Nat. Methods. 2013; 10: 1096–1098.
20. Evan Z Macosko, Anindita Basu, Rahul Satija, James Nemesh, Karthik Shekhar, et al. Highly parallel genome-wide expression profiling of individual cells using nanoliter droplets. Cell. 2015; 161: 1202–1214.
21. Grace XY Zheng, Jessica M Terry, Phillip Belgrader, Paul Ryvkin, Zachary W Bent, et al. Massively parallel digital transcriptional profiling of single cells. Nat. Commun. 2017; 8.
22. Andrews S. FastQC. 2012.
23. Ewels P, Magnusson M, Lundin S, Käller M. MultiQC: summarize analysis results for multiple tools and samples in a single report. Bioinformatics. 2016; 32: 3047–3048.
24. Bolger AM, Lohse M, Usadel B. Trimmomatic: a flexible trimmer for Illumina sequence data. Bioinformatics. 2014; 30: 2114–2120.
25. Martin M. Cutadapt removes adapter sequences from high-throughput sequencing reads. EMBnet.journal. 2011; 17: 10.
26. Giacomo Baruzzo, Katharina E Hayer, Eun Ji Kim, Barbara Di Camillo, Garret A FitzGerald, et al. Simulation-based comprehensive benchmarking of RNA-seq aligners. Nat. Methods. 2017; 14: 135–139.
27. Leonard D Goldstein, Yi Cao, Gregoire Pau, Michael Lawrence,Thomas D Wu, et al. Prediction and Quantification of Splice Events from RNA-Seq Data. PLOS ONE. 2016; 11: e0156132.
28. Pär G Engström, Tamara Steijger, Botond Sipos, Gregory R Grant, André Kahles, et al. Systematic evaluation of spliced alignment programs for RNA-seq data. Nat. Methods. 2013; 10: 1185–1191.

29. Kim D, Paggi JM, Park C, Bennett C, Salzberg SL. Graph-based genome alignment and genotyping with HISAT2 and HISAT-genotype. Nat. Biotechnol. 2019; 37: 907–915.
30. Kaminow B, Yunusov D, Dobin A, Spring C. STARsolo : accurate , fast and versatile mapping / quantification of single-cell and single-nucleus RNA-seq data. 2021; 1–35.
31. Compeau PEC, Pevzner PA, Tesler G. How to apply de Bruijn graphs to genome assembly. Nat. Biotechnol. 2011; 29: 987–991.
32. Páll Melsted, A. Sina Booeshaghi, Lauren Liu, Fan Gao, Lambda Lu, et al. Modular, efficient and constant-memory single-cell RNA-seq preprocessing. Nat. Biotechnol. 2021; 39: 813–818.
33. Srivastava A, Malik L, Sarkar H, Patro R. A Bayesian framework for inter-cellular information sharing improves dscRNA-seq quantification. Bioinformatics. 2020; 36: i292–i299.
34. Du Y, Huang Q, Arisdakessian C, Garmire LX. Evaluation of STAR and kallisto on single cell RNA-seq data alignment. G3 Genes Genomes Genet. 2020; 10: 1775–1783.
35. Imad Abugessaisa, Akira Hasegawa, Shuhei Noguchi, Melissa Cardon, Kazuhide Watanabe, et al. SkewC: Identifying cells with skewed gene body coverage in single-cell RNA sequencing data. iScience. 2022; 25: 103777.
36. Tomislav Ilicic, Jong Kyoung Kim, Aleksandra A Kolodziejczyk, Frederik Otzen Bagger, Davis James, et al. Classification of low quality cells from single-cell RNA-seq data. Genome Biol. 2016; 17: 1–15.
37. Ariel A Hippen, Matias M Falco, Lukas M Weber, Erdogan Pekcan Erkan, Kaiyang Zhang, et al. miQC: An adaptive probabilistic framework for quality control of single-cell RNA-sequencing data. PLOS Comput. Biol. 2021; 17: e1009290.
38. Xi NM, Li JJ. Benchmarking Computational Doublet-Detection Methods for Single-Cell RNA Sequencing Data. Cell Syst. 2021; 12: 176-194.e6.
39. Georgi K Marinov, Brian A Williams, Ken McCue, Gary P Schroth, Jason Gertz, et al. From single-cell to cell-pool transcriptomes: Stochasticity in gene expression and RNA splicing. Genome Res. 2014; 24: 496–510.

40. Andrews TS, Kiselev VY, McCarthy D, Hemberg M. Tutorial: guidelines for the computational analysis of single-cell RNA sequencing data. Nat. Protoc. 2020; 16: 1–9.
41. Bacher R, Kendziorski C. Design and computational analysis of single-cell RNA-sequencing experiments. Genome Biol. 2016; 17: 1–14.
42. Vallejos CA, Risso D, Scialdone A, Dudoit S, Marioni JC. Normalizing single-cell RNA sequencing data: challenges and opportunities. Nat. Methods. 2017; 14: 565–571.
43. Chen G, Ning B, Shi T. Single-cell RNA-seq technologies and related computational data analysis. Front. Genet. 2019; 10: 1–13.
44. Anders S, Huber W. Differential expression analysis for sequence count data. Genome Biol. 2010; 11: R106.
45. Satija R, Farrell JA, Gennert D, Schier AF, Regev A. Spatial reconstruction of single-cell gene expression data. Nat. Biotechnol. 2015; 33: 495–502.
46. Rhonda Bacher, Li-Fang Chu, Ning Leng, Audrey P Gasch, James A Thomson, et al. SCnorm: robust normalization of single-cell RNA-seq data. Nat. Methods. 2017; 14: 584–586.
47. Hafemeister C, Satija R. Normalization and variance stabilization of single-cell RNA-seq data using regularized negative binomial regression. Genome Biol. 2019; 20: 296.
48. Cheng Jia, Yu Hu, Derek Kelly, Junhyong Kim, Mingyao Li, et al. Accounting for technical noise in differential expression analysis of single-cell RNA sequencing data. Nucleic Acids Res. 2017; 45: 10978–10988.
49. Haghverdi L, Lun ATL, Morgan MD, Marioni JC. Batch effects in single-cell RNA-sequencing data are corrected by matching mutual nearest neighbors. Nat. Biotechnol. 2018; 36: 421–427.
50. Luecken MD, Theis FJ. Current best practices in single-cell RNA-seq analysis: a tutorial. Mol. Syst. Biol. 2019; 15.
51. Johnson WE, Li C, Rabinovic A. Adjusting batch effects in microarray expression data using empirical Bayes methods. Biostatistics. 2007; 8: 118–127.
52. Hoa Thi Nhu Tran, Kok Siong Ang, Marion Chevrier, Xiaomeng Zhang, Nicole Yee Shin Lee, et al. A benchmark of batch-effect correction methods for single-cell RNA sequencing data. Genome Biol. 2020; 21: 12.

53. Laura Cantini, Pooya Zakeri, Celine Hernandez, Aurelien Naldi, Denis Thieffry, et al. Benchmarking joint multi-omics dimensionality reduction approaches for the study of cancer. Nat. Commun. 2021; 12: 124.
54. Wei Liu, Xu Liao, Yi Yang, Huazhen Lin, Joe Yeong, et al. Joint dimension reduction and clustering analysis of single-cell RNA-seq and spatial transcriptomics data. Nucleic Acids Res. 2022; 50: e72–e72.
55. Ilya Korsunsky, Nghia Millard, Jean Fan, Kamil Slowikowski, Fan Zhang, et al. Fast, sensitive and accurate integration of single-cell data with Harmony. Nat. Methods. 2019; 16: 1289–1296.
56. Butler A, Hoffman P, Smibert P, Papalexi E, Satija R. Integrating single-cell transcriptomic data across different conditions, technologies, and species. Nat. Biotechnol. 2018; 36: 411–420.
57. Chazarra-Gil R, van Dongen S, Kiselev VY, Hemberg M. Flexible comparison of batch correction methods for single-cell RNA-seq using BatchBench. Nucleic Acids Res. 2012; 49: e42–e42.
58. Ding J, Condon A, Shah SP. Interpretable dimensionality reduction of single cell transcriptome data with deep generative models. Nat. Commun. 2018; 9.
59. Huang H, Wang Y, Rudin C, Browne EP. Towards a comprehensive evaluation of dimension reduction methods for transcriptomic data visualization. Commun. Biol. 2022; 5: 719.
60. Chen Meng, Oana A Zeleznik, Gerhard G Thallinger, Bernhard Kuster, Amin M Gholami, et al. Dimension reduction techniques for the integrative analysis of multi-omics data. Brief. Bioinform. 2016; 17: 628–641.
61. Quackenbush J. Extracting biology from high-dimensional biological data. J. Exp. Biol. 2007; 210: 1507–1517.
62. Maaten L. van der & Hinton, G. Visualizing Data using t-SNE. J. Mach. Learn. Res. 2008; 9: 2579–2605.
63. Kobak D, Berens P. The art of using t-SNE for single-cell transcriptomics. Nat. Commun. 2019; 10.
64. McInnes L, Healy J, Melville J. UMAP: Uniform manifold approximation and projection for dimension reduction. arXiv 2018.

65. Junyue Cao, Malte Spielmann, Xiaojie Qiu, Xingfan Huang, Daniel M Ibrahim, et al. The single-cell transcriptional landscape of mammalian organogenesis. Nature. 2019; 566: 496–502.
66. Sara Aibar, Carmen Bravo González-Blas, Thomas Moerman, Vân Anh Huynh-Thu, Hana Imrichova, et al. SCENIC: single-cell regulatory network inference and clustering. Nat. Methods. 2017; 14: 1083–1086.
67. Yingxin Lin, Yue Cao, Hani Jieun Kim, Agus Salim, Terence P Speed, et al. scClassify: sample size estimation and multiscale classification of cells using single and multiple reference. Mol. Syst. Biol. 2020; 16.
68. Tan Y, Cahan P. SingleCellNet: A Computational Tool to Classify Single Cell RNA-Seq Data Across Platforms and Across Species. Cell Syst. 2019; 9: 207-213.e2.
69. Juliá M, Telenti A, Rausell A. Sincell: an R/Bioconductor package for statistical assessment of cell-state hierarchies from single-cell RNA-seq. Bioinformatics. 2015; 31: 3380–3382.
70. Singer SJ. Intercellular Communication and Cell-Cell Adhesion. Science. 1992; 255: 1671–1677.
71. Cassidy L Bland, Christina N Byrne-Hoffman, Audry Fernandez, Stephanie L Rellick, Wentao Deng, et al. Exosomes derived from B16F0 melanoma cells alter the transcriptome of cytotoxic T cells that impacts mitochondrial respiration. FEBS J. 2018; 285: 1033–1050.
72. Armingol E, Officer A, Harismendy O, Lewis NE. Deciphering cell–cell interactions and communication from gene expression. Nat. Rev. Genet. 2021; 22: 71–88.
73. Lihong Peng, Feixiang Wang, Zhao Wang, Jingwei Tan, Li Huang, et al. Cell–cell communication inference and analysis in the tumour microenvironments from single-cell transcriptomics: data resources and computational strategies. Brief. Bioinform. 2022; 23: bbac234.
74. Bridges K, Miller-Jensen K. Mapping and Validation of scRNA-Seq-Derived Cell-Cell Communication Networks in the Tumor Microenvironment. Front. Immunol. 2022; 13: 885267.
75. Efremova M, Vento-Tormo M, Teichmann SA, Vento-Tormo R. CellPhoneDB: inferring cell–cell communication

from combined expression of multi-subunit ligand–receptor complexes. Nat. Protoc. 2020; 15: 1484–1506.

76. Roser Vento-Tormo, Mirjana Efremova, Rachel A Botting, Margherita Y Turco, Miquel Vento-Tormo, et al. Single-cell reconstruction of the early maternal–fetal interface in humans. Nature. 2018; 563: 347–353.
77. Zhencong Chen, Mengnan Zhao, Jiaqi Liang, Zhengyang Hu, Yiwei Huang, et al. Dissecting the single-cell transcriptome network underlying esophagus non-malignant tissues and esophageal squamous cell carcinoma. eBioMedicine. 2021; 69: 103459.
78. Lulu Liu, Ruyi Zhang, Jingwen Deng, Xiaomeng Dai, Xudong Zhu, et al. Construction of TME and Identification of crosstalk between malignant cells and macrophages by SPP1 in hepatocellular carcinoma. Cancer Immunol. Immunother. 2022; 71: 121–136.
79. Ayse Bassez, Hanne Vos, Laurien Van Dyck, Giuseppe Floris, Ingrid Arijs, et al. A single-cell map of intratumoral changes during anti-PD1 treatment of patients with breast cancer. Nat. Med. 2021; 27: 820–832.
80. Kevin Bi, Meng Xiao He, Ziad Bakouny, Abhay Kanodia, Sara Napolitano, et al. Tumor and immune reprogramming during immunotherapy in advanced renal cell carcinoma. Cancer Cell. 2021; 39: 649-661.e5.
81. Yi-Quan Jiang, Zi-Xian Wang, Ming Zhong, Lu-Jun Shen, Xue Han, et al. Investigating Mechanisms of Response or Resistance to Immune Checkpoint Inhibitors by Analyzing Cell-Cell Communications in Tumors Before and After Programmed Cell Death-1 (PD-1) Targeted Therapy: An Integrative Analysis Using Single-cell RNA and Bulk-RNA Sequencing Data. OncoImmunology. 2021; 10: 1908010.
82. Alexander H Lee, Lu Sun, Aaron Y Mochizuki, Jeremy G Reynoso, Joey Orpilla, et al. Neoadjuvant PD-1 blockade induces T cell and cDC1 activation but fails to overcome the immunosuppressive tumor associated macrophages in recurrent glioblastoma. Nat. Commun. 2021; 12: 6938.
83. Park Y, Jeong J, Seong S, Kim W. In Silico Evaluation of Natural Compounds for an Acidic Extracellular Environment in Human Breast Cancer. Cells. 2021; 10: 2673.
84. Honghao Yin, Rui Guo, Huanyu Zhang, Songyi Liu, Yuehua

Gong, et al. A Dynamic Transcriptome Map of Different Tissue Microenvironment Cells Identified During Gastric Cancer Development Using Single-Cell RNA Sequencing. Front. Immunol. 2021; 12: 728169.

85. Browaeys R, Saelens W, Saeys Y. NicheNet: modeling intercellular communication by linking ligands to target genes. Nat. Methods. 2020; 17: 159–162.
86. Bai Z, Su G, Fan R. Single-cell Analysis Technologies for Immuno-oncology Research: from Mechanistic Delineation to Biomarker Discovery. Genomics Proteomics Bioinformatics. 2021; 19: 191–207.
87. Ma A, Xin G, Ma Q. The use of single-cell multi-omics in immuno-oncology. Nat. Commun. 2022; 13: 2728.
88. Fatima Valdes-Mora, Kristina Handler, Andrew M K Law, Robert Salomon, Samantha R Oakes, et al. Single-cell transcriptomics in cancer immunobiology: The future of precision oncology. Front. Immunol. 2018; 9.
89. Anne Bertolini, Michael Prummer, Mustafa Anil Tuncel, Ulrike Menzel, María Lourdes Rosano-González, et al. scAmpi—A versatile pipeline for single-cell RNA-seq analysis from basics to clinics. PLOS Comput. Biol. 2022; 18: e1010097.
90. Luis García-Jimeno, Coral Fustero-Torre, María José Jiménez-Santos, Gonzalo Gómez-López, Tomás Di Domenico, et al. bollito: a flexible pipeline for comprehensive single-cell RNA-seq analyses. Bioinformatics. 2022; 38: 1155–1156.
91. Germain PL, Sonrel A, Robinson MD. pipeComp, a general framework for the evaluation of computational pipelines, reveals performant single cell RNA-seq preprocessing tools. Genome Biol. 2020; 21: 227.
92. Marlon Stoeckius, Christoph Hafemeister, William Stephenson, Brian Houck-Loomis, Pratip K Chattopadhyay, et al. Simultaneous epitope and transcriptome measurement in single cells. Nat. Methods. 2017; 14: 865–868.
93. Yonatan Katzenelenbogen, Fadi Sheban, Adam Yalin, Ido Yofe, Dmitry Svetlichnyy, et al. Coupled scRNA-Seq and Intracellular Protein Activity Reveal an Immunosuppressive Role of TREM2 in Cancer. Cell. 2020; 182: 872-885.e19.
94. Amy F Chen, Benjamin Parks, Arwa S Kathiria, Benjamin

Ober-Reynolds, Jorg J Goronzy, et al. NEAT-seq: simultaneous profiling of intra-nuclear proteins, chromatin accessibility and gene expression in single cells. Nat. Methods. 2022; 19: 547–553.

95. Cheng H. Smart3-ATAC: a highly sensitive method for joint accessibility and full-length transcriptome analysis in single cells. 2021. Available online at: http://biorxiv.org/lookup/doi/10.1101/2021.12.02.470912
96. Iain C Macaulay, Wilfried Haerty, Parveen Kumar, Yang I Li, Tim Xiaoming Hu, et al. G&T-seq: parallel sequencing of single-cell genomes and transcriptomes. Nat. Methods. 2015; 12: 519–522.
97. Yu Hou, Huahu Guo, Chen Cao, Xianlong Li, Boqiang Hu, et al. Single-cell triple omics sequencing reveals genetic, epigenetic, and transcriptomic heterogeneity in hepatocellular carcinomas. Cell Res. 2016; 26: 304–319.
98. Marx V. Method of the Year: spatially resolved transcriptomics. Nat. Methods. 2021; 18: 9–14.
99. Chiara Baccin, Jude Al-Sabah, Lars Velten, Patrick M Helbling, Florian Grünschläger, et al. Combined single-cell and spatial transcriptomics reveal the molecular, cellular and spatial bone marrow niche organization. Nat. Cell Biol. 2020; 22: 38–48.
100. Cang Z, Nie Q. Inferring spatial and signaling relationships between cells from single cell transcriptomic data. Nat. Commun. 2020; 11: 1–13.
101. Bin Li, Wen Zhang, Chuang Guo, Hao Xu, Longfei Li, et al. Benchmarking spatial and single-cell transcriptomics integration methods for transcript distribution prediction and cell type deconvolution. Nat. Methods. 2022; 19: 662–670.
102. Yan L, Sun X. Benchmarking and integration of methods for deconvoluting spatial transcriptomic data. Bioinformatics. 2023; 39: btac805.