frontiers | Frontiers in Genetics

TYPE Review
PUBLISHED 08 November 2023
DOI 10.3389/fgene.2023.1256991

OPEN ACCESS

EDITED BY
Jun Wan,
Indiana University, United States

REVIEWED BY
Yaqiang Cao,
National Institutes of Health (NIH), United States
Vasudha Srivastava,
University of California San Francisco, United States

*CORRESPONDENCE
Enrique Hernández-Lemus,
ehernandez@inmegen.gob.mx
Guillermo de Anda-Jáuregui,
gdeanda@inmegen.edu.mx

RECEIVED 11 July 2023
ACCEPTED 24 October 2023
PUBLISHED 08 November 2023

CITATION
Paas-Oliveros E, Hernández-Lemus E and de Anda-Jáuregui G (2023), Computational single cell oncology: state of the art.
Front. Genet. 14:1256991.
doi: 10.3389/fgene.2023.1256991

Computational single cell oncology: state of the art

Ernesto Paas-Oliveros[1], Enrique Hernández-Lemus[1,2]* and Guillermo de Anda-Jáuregui[1,2,3]*

[1]Computational Genomics Division, National Institute of Genomic Medicine, Mexico City, Mexico, [2]Center for Complexity Sciences, Universidad Nacional Autónoma de México, Mexico City, Mexico, [3]Investigadores por Mexico, Conahcyt, Mexico City, Mexico

Single cell computational analysis has emerged as a powerful tool in the field of oncology, enabling researchers to decipher the complex cellular heterogeneity that characterizes cancer. By leveraging computational algorithms and bioinformatics approaches, this methodology provides insights into the underlying genetic, epigenetic and transcriptomic variations among individual cancer cells. In this paper, we present a comprehensive overview of single cell computational analysis in oncology, discussing the key computational techniques employed for data processing, analysis, and interpretation. We explore the challenges associated with single cell data, including data quality control, normalization, dimensionality reduction, clustering, and trajectory inference. Furthermore, we highlight the applications of single cell computational analysis, including the identification of novel cell states, the characterization of tumor subtypes, the discovery of biomarkers, and the prediction of therapy response. Finally, we address the future directions and potential advancements in the field, including the development of machine learning and deep learning approaches for single cell analysis. Overall, this paper aims to provide a roadmap for researchers interested in leveraging computational methods to unlock the full potential of single cell analysis in understanding cancer biology with the goal of advancing precision oncology. For this purpose, we also include a notebook that instructs on how to apply the recommended tools in the Preprocessing and Quality Control section.

KEYWORDS

single cell transcriptomics, computational oncology, bioinformatics, cellular heterogeneity, best practices Frontiers

1 Introduction

Cancer, as a complex and multifaceted disease, continues to pose significant challenges to medical professionals and researchers worldwide. Traditionally, cancer has been studied at the tissue (also known as *bulk*) level, providing valuable insights into the overall behavior of tumors. However, this approach fails to capture the intrinsic heterogeneity that exists within tumors, leading to an incomplete understanding of the disease and hindering the development of targeted therapies.

In recent years, the advent of single cell analysis has revolutionized the field of oncology by enabling the characterization of individual cells within a tumor. This powerful technique allows researchers to dissect the tumor heterogeneity by unraveling cellular diversity, aiming to decipher the dynamic processes that underlie tumor progression, metastasis, and therapy resistance.

Single cell analysis involves the isolation, profiling, and sequencing of individual cells, providing researchers with high-resolution data on the genetic, epigenetic, transcriptomic, proteomic, and metabolic features of each cell. By unveiling the molecular landscape of tumors at the single cell level, this approach offers unprecedented insights into tumor evolution, clonal dynamics, and the cellular interactions that drive cancer development and response to therapy.

Through an in-depth examination of recent studies and cutting-edge advancements, we will highlight the immense potential of single cell analysis in driving personalized medicine and improving clinical outcomes in oncology. Moreover, we will explore the challenges and limitations of this technology, in particular those related to data analysis and interpretation, taking into account the technological biases, and the need for scalable and cost-effective methodologies.

Ultimately, the integration of single cell analysis into oncology research has the potential to revolutionize our understanding of cancer biology, foster the development of more precise diagnostics and therapies, and pave the way for personalized approaches that consider the unique cellular landscape of each patient's tumor.

1.1 Cellular heterogeneity in cancer

Hanahan and Weinberg (Hanahan, 2022) have outlined major hallmarks of cancer function, yet there is no single molecular pathway to attain these functions and there may be other mechanisms to be found. Among the many fragments of cancer's mechanistic puzzle, one important component of cancer complexity lies in the complex cellular environment within tumor tissues. In this regard, single cell experimental technologies, such as single cell RNA sequencing (scRNA-seq), may provide relevant clues to better understanding the molecular basis of the characteristic functional features of the tumor multicellular environment. Single cell analysis allow to study processes in the intersection between cell states and convergence to biological function. scRNA-seq, in particular, allows for the simultaneous profiling of genome expression for most cells in a tissue sample. The single cell transcriptome represents a middle ground to characterize biological pathways, shifting to molecular focus to chart the variability among individual molecular programs in order to infer possible functional phenotypes, even among rare cell types. As we grasp this focus, there is a continuous enhancement of computational algorithms and approaches. New features are added to the methods used in standardized single cell analysis. This is why an overview of the recent advances and how they can be applied to advance the understanding and treatment of cancer is of importance.

Two main aspects of cancer have been apparent since its first observation. Its nature as a malignant tumor and its almost unbending resilience. Nevertheless, only in the 1800's, with the advent of the microscope and Virchow's proposal for cancer as a disease of the cell, did we begin to understand the alterations in and around the cell that contribute to tumor proliferation and adaptability (Lonardo et al., 2015). These characteristics and focus have made us understand that malignant cells vary their state and function in various modalities: across the course of the disease, in their location on the tissue and in response to external insults, most importantly therapy.

Initially, the focus in cancer research was on karyotypic and mutational alterations, underscoring the evolutionary adaptiveness of cancer (Burrell et al., 2013). Today, we understand that heterogeneity can be observed in various molecular pathways of cells which can contribute to the proliferative and adaptive capabilities of the tumour. These include active metabolic programs (Kim and DeBerardinis, 2019), epigenetic configurations (Bell et al., 2019), transcriptional profiles (Kinker et al., 2019), exosomal disposition (Lee et al., 2022) and microbial interactions (Niño et al., 2022). Even more, cells surrounding the malignant tissue in solid cancers can be recruited by the tumor, can try to fight it and even influence the state and functions of malignant cells. Behaviors often observed in tumor infiltrating lymphocites (TILs), tumor associated macrophages (TAMs), cancer associated fibroblasts (CAFs) among others. Together with the tumor, these cells comprise the tumor micro-environment (TME) and its characterization and dynamics have been a subject of numerous studies, particularly since the advent of transcriptome sequencing technologies (Nam et al., 2021).

In this review we present a summary of the theoretical principles and the latest technologies of this framework and convey a landscape of the state of the art in applications for cancer. The search was further systematized by automated and prioritized bibliographic search.

2 The need for proper experimental designs for single cell analysis in oncology

Analyzing tumor samples at the single-cell level presents several experimental design challenges (Kolodziejczyk et al., 2015; Dal Molin and Di Camillo, 2019), each of which can significantly impact the quality and interpretability of the data. Some of the key challenges and considerations are as follows:

A first challenge in the design of single cell RNASeq experiments lies in representing the full complexity of the tissues/phenotypes in an unbiased way. Obtaining an adequate number of high-quality single cells from tumor tissues can be challenging (Birnbaum, 2018). Tumors often consist of a mixture of cancer cells, stromal cells, and immune cells (Guillaumet-Adkins et al., 2017; Miller et al., 2017). The sample size required depends on the research question, but it is crucial to ensure that the sample size is statistically meaningful (Davis et al., 2019; Su et al., 2020). One must also consider rare cell types: Some cancer subpopulations or rare cell types within tumors may be of particular interest, but these can be challenging to capture in sufficient numbers (Jiang et al., 2016; Xie et al., 2020). Furthermore, to ensure the reproducibility of findings, it is essential to collect and analyze multiple samples or replicate experiments (Skinnider et al., 2019; Zimmerman et al., 2021).

Another key issue is the identification of the best possible (or available) source of tissues. Fresh tissues are ideal for single-cell analysis as they preserve cellular viability and gene expression profiles. However, obtaining fresh samples can be logistically challenging, especially for certain cancer types or when dealing with clinical specimens. The alternatives here are the use of frozen or

fixed tissues. Frozen tissues are a valuable alternative when fresh samples are unavailable. They can preserve RNA and protein, but the freezing process can introduce artifacts and affect the quality of single-cell data (Slyper et al., 2020; Jiang et al., 2023). Fixed tissues are also useful but with relevant limitations: these can provide spatial information and allow the analysis of archival samples (Gerdes et al., 2013; Phan et al., 2021; Tian L. et al., 2023). However, fixation can alter cellular morphology and gene expression, making it less suitable for certain single-cell assays.

A third aspect to evaluate is how to balance Intra vs. Inter-Patient or Tumor Heterogeneity. Intra-tumor heterogeneity must be considered, tumors are often composed of subclones with distinct genetic and phenotypic characteristics. To capture intra-tumor heterogeneity, researchers need to profile multiple single cells from different regions within a tumor (Martelotto et al., 2014; Dong et al., 2021; Lo et al., 2023). On the other hand, different individuals are also quite heterogeneous even in analogous regions/organs/tissues, designs must deal with inter-patient heterogeneity because comparing single cells from different patients adds another layer of heterogeneity. It is essential to consider patient-to-patient variability when drawing conclusions about cancer biology (Yancovitz et al., 2012; Wang T. et al., 2022).

Aside from the purely biological/clinical issues of the experimental designs one need also consider technical decisions. For instance, many research questions in cancer biology require the integration of different data types, such as genomics, transcriptomics, proteomics, and epigenomics (Peng et al., 2020; Nam et al., 2021; Ma et al., 2022). Designing experiments that allow for the simultaneous profiling of multiple omics layers in the same single cells is technically challenging and researchers need to adequately ponder when doing so will add enough depth to their study to justify the additional costs and logistic complexities (Li et al., 2021; Dimitriu et al., 2022). Furthermore, analyzing multi-omic data from single cells often requires the development or application of specialized computational tools for data integration and interpretation (Hu et al., 2018; Rautenstrauch et al., 2022).

Hence, if we want to exploit single-cell analysis of cancers as a powerful approach to provide insights into tumor heterogeneity, clonal evolution, and therapy response, we need to carefully consider sample acquisition, preservation methods, and experimental design to address the unique challenges posed by single-cell studies (Baran-Gale et al., 2018; Lafzi et al., 2018; Nguyen et al., 2018). Collaboration between experimentalists and computational biologists is crucial to maximize the quality and utility of single-cell cancer data (Bacher and Kendziorski, 2016). Additionally, ongoing advancements in single-cell technologies and analytical methods are continually improving our ability to overcome these challenges and gain deeper insights into cancer biology. So one needs to be aware of ongoing developments in the field.

3 A primer on scRNA-seq analysis

The primary objective in transcriptome sequencing is to measure the number of RNA transcripts in the cytosol and nucleus of cells in a sample. There are various protocols that have been developed to achieve single-cell sequencing. They can differ in various steps of the process, and each of these steps can contribute to the customization of a specific experiment. In the following, we outline the general steps (See Figure 1), how they work and how each variation can help tackle different settings. References for the articles presenting these methods can be found in Table 1, so that they are not repeated in the text.

We start by explaining and exposing the main technologies for the experimental steps required for scRNA-seq. Naturally, to sequence RNA from individual cells after extracting the tissue of interest, one must physically isolate the cells. We will hence start with this necessary step.

3.1 Cell separation

Prior to isolating any cells, if they come from solid tissue, the cells must be dissociated. This is normally done with typsin, collagenase or and/or papain, although there are *in situ* methods available for spatial transcriptomics (Shah et al., 2017). Careful handling and special consideration for fragile cells must be taken into account during dissociation because the stress response can alter the transcriptional program (Lee et al., 2021).

Initially, cells were manually pipetted, which was time-consuming and defeated the purpose of obtaining an overview of all cells in a sample. In most methods, cells are isolated via fluorescence-activated cell-sorting (FACS), diffused into picowells, and piped away through microfluidics or reverse-emulsified with nano-droplets.

FACS-sorting can be used without a biomarker to randomly select cells from a solution but it is time consuming and it does not allow for very high throughput, as is the case with Smart-Seq and FLASH-seq protocols. It is to be noted that FACS is often used before in many workflows to select a population of interest or to exclude dead cells.

With the use of beads that bind to random cells and a picowell plate into which only a bead with a cell can fit, many cells can be isolated and enclosed to react, separated by a semipermeable membrane. This method achieves the highest throughput, like in the case of the Seq-Well platform, where around 88k cells can be captured in one run. Nevertheless it is prone to noise because of the high quantity of wells. To avoid this, micro-wells that are filled via microfluidics can be used. The Fluidigm C1 platform takes this approach but sacrifices a lot of throughput, filling only plates of 96 wells from each drop of solution.

The most popular method (Chromium 10x) uses oil or hydrogel droplets to encapsulate cells through reverse emulsion. A bead with many oligos that reacts with a lysed cell is also inside the droplet. Although this method offers high throughput and does not require a plate, there is a small possibility of duplicates in the droplets, which increases exponentially with the number of cells captured.

3.2 Library generation and sequencing

In the last year, there has been a lot of progress in the parallelization and efficacy of reverse transcription of RNA molecules, which are converted into cDNA that is able to be sequenced in NGS (next-generation sequencers).

The Chromium 10x v2 platform only transcribes from the 3' end, using an oligo dT for priming, and it skips the template

TABLE 1 Avaried sample of popular scRNA-seq platforms.

Protocol	Cell isolation	Time (h)	# of cells	Cost	Trans.	Reference
Chromium10x v2	Droplet	9	10k	$$$$	3'	Zheng et al. (2017)
Chromium10x v3	Droplet	9	10k	$$$$	Full	Simone et al. (2019)
InDrop	Droplet	24	>10k	$$$	3'	Zilionis (2017)
Smart-Seq2	FACS	25	384	$	Full	Picelli et al. (2013)
Smart-Seq3	FACS	10	384	$$	Full	Hagemann-Jensen et al. (2020)
FLASH-seq	FACS	4.5	384	$$	Full	Hahaut et al. (2022)
Fluidigm C1	Microfluidics	5	96	$$	Full	DeLaughter (2018)
Seq-Well	Picowell	10	88k	$$	Full	Aicher et al. (2019)

switching step due to the difficulties of performing this step inside a droplet. Template switching works by providing an alternate sequence for the reverse transcriptase (e.g., Superscript III) to switch to, in case it encounters a stopping sequence, then it latches back on to the transcript. This mechanism, and another priming oligo at the end of the transcript, guarantee a full read. Nowadays, due to molecular advances, full length transcripts can also be sequenced in high throughput platforms. A feature that facilitates the detection of SNPs, isoforms and allelic variants in the analysis.

The reactions involved in reverse transcription are designed to deliver a cDNA molecule that can be easily sequenced and amplified through PCR. One such reaction that is becoming more widespread is the 'Unique Molecular Identifier' (UMI) barcoding. By attaching an average of a random sequence of 10 bases to the primer, it is almost guaranteed that every transcript has a unique barcode. This information can later be extracted to remove the amplification bias that occurs when cDNA is amplified via PCR. Barcoding was first used by the platforms that used droplet separation, because the amplification happened with all the transcripts from all the cells mixed up, but it is implemented in most newer platforms. Plate well technologies amplify within the well, so the amplification bias is almost linear. Nevertheless, UMIs are a molecular memory that does not rely on modelling and have been shown to correlate better to the actual genes in the library (Kivioja et al., 2012).

The library is then sequenced using a next-generation sequencing (NGS) platform, such as Illumina's NextSeq 500 or ThermoFisher's IonTorrent. Sequencing is performed in batch and the result is the first piece of digital information to be handled in the pipeline: A multiplexed FASTQ file.

3.3 Preprocessing and quality control

Starting from here, the workflow is entirely digital and can be run on a computer (See Figure 2). To illustrate the recommended methods in this review and to help with the setting up of an environment for running these advanced frameworks we provide a github repository with a notebook and a container in https://github.com/epaaso/comp-oncology.

To begin to analyze the batch FASTQ file, it need to be demultiplexed, that is the cell barcodes and UMI labels are extracted and the remaining sequence is annotated. That is the sequences are aligned and mapped to known genes, exons, introns or sequences of interest. This is achieved with the assistance of algorithms such as BLAST, RefSeq, or GenCode. Software like CellRanger which is designed to work with sequences generated by the 10x Chromium platform or STARSolo that is more general. Care must be taken in this step, because a faulty annotation would bias all the following analysis. When in doubt, these tools also come with possible quality control measures. There are two main annotations used for sequences, the Ensembl ID, whose main feature is that it is unique and the gene symbol, which is more closely related to its discovery or function. When employing different tools, the conversion from Ensembl ID to gene symbol or *vice versa* is often required, and this largely depends on the reference database used. Almost all reference datasets come from the Ensembl website, but older platforms use the legacy hg19 reference. Additionally, the database can consider not only genes but miRNAs, non-coding transcripts and others.

This annotation results in a quality control filtered BAM or SAM file, from which the repeated appearances of annotated sequences can be counted, to end up with a count matrix that has rows as cells and genes as columns or *vice versa*. This matrix can be used for downstream analysis to ask questions about the biological phenomenon, but usually some further quality control needs to be done.

In the following section, we present the most relevant data manipulation procedures that aim to provide us with the most biologically relevant yet computationally efficient dataset. This methods do not *a priori* look for a trend or ask a question of the experiment. However they can change or be coupled to the downstream analyses that can follow.

3.3.1 Filters and feature selection

Biologically, cells that have a high amount of mitochondrial genes in the substrate (around 5%–10%) are considered to be dead cells or cells that underwent too much stress. These cells are normally removed. Additionally, cells that are outliers with either too few or too many measured genes may be either dead cells or doublets/multiplets. There is a plethora of algorithms to account for various biological effects, the most accepted of which is the cell cycle correction (Barron and Li, 2016). Nevertheless, many corrections are controversial and applying one correction can hide the presence of another. Another important effect is the contamination of the

transcripts by ambient RNA, CellBender (Fleming et al., 2022) uses Bayesian modelling and neural networks to extract the signal from ambient RNA.

Even though the filtering steps reduce the dimension of the features by a certain amount, *feature selection* is done to account for these next issues: Due to the curse of dimensionality, where data gets sparser the more dimensions there are, the sampling capacity required to arrive to a statistically relevant result raises exponentially. Too many genes are redundant and considering all can lead to overfitting. For addressing these problems the most popular pipelines, like Seurat, search for *highly variable genes* (HVG), with the aid of generalized linear models. This is an example of an unsupervised approach, and there are more sophisticated ones available. A popular tool, SCMarker (Wang et al., 2019), creates gene expression modalities and filters out the genes in the sparse modalities. M3Drop (Andrews and Hemberg, 2018) takes into account the dropout distribution and filters the ones that go out from the distribution.

However, the selection process can have various effects on downstream analysis. That's why many algorithms developed nowadays are supervised based on the analysis to be performed. For example, Triku (Ascensión et al., 2022) uses a k-nearest-neighbour clustering approach to create an expression profile of a certain cluster of genes, and then selects for the genes that are most informative within each cluster. Since the induced analysis needs to be run to optimize selection, some greedy algorithms are used to speed up the process. For example, in genetic algorithms a certain set of features are selected and rated based on their suitability for downstream analysis. The generations that have the highest score continue to have features added to them. This approach is similar to decision trees.

In summary, the variety of feature selection algorithms address various concerns like efficiency, sparsity and dropouts. Being attached to the downstream analyses also contributes to their diversity. To choose the optimal method, care must be taken to examine the underlying hypotheses of the integrated methods in a downstream analysis. These methods also need to be prioritized based on the context of the data and the experiment.

While *feature selection* diminishes the amount of information to remove stochastic or design artifacts, *imputation* aims to do this by adding more information.

3.3.2 Imputation

A very debated topic is the occurrence of dropouts, which can occur due to various factors like incomplete reads, amplification errors or even transcriptional bursts (Dar et al., 2012). Through modeling and zero inflation, dropouts can be imputed or artifacts removed. Some authors suggest that most dropouts are not significant and attribute them to intrinsic stochasticity by adjusting for a negative binomial distribution (Silverman et al., 2020). Another important factor is that sparse matrices with large blocks of zeros, pose challenges when doing the calculations that many of the downstream analyses require. Accordingly, there are a lot of methods that take this into consideration. This can be done during the preprocessing phase or implemented in downstream analysis. There is an ongoing discussion on whether correction through imputation, smoothing, or no correction is the optimal solution (Hou et al., 2020) and the actual answer varies from case to case.

3.3.3 Normalization

Adjusting to a distribution is also useful for normalization, a practice of basic importance to correct technical variations that may be present in different reads. In the previous paragraph, we discussed cell counts. However, to account for differences in gene expression between cells or within a cell, the expression counts must be scaled so that high counts do not overshadow other expressions. Basic strategies like CPM (Counts Per Million), RPKM (Reads Per Kilobase Million), TPM (Transcripts Per Million), and FPKM (Fragments Per Kilobase Million) adjust based on all the read counts, using a global scale factor that can sometimes overcorrect and impair downstream analysis.

Another important factor addressed in normalization is the stability of variance, where the expression values are transformed by a function, such as log. This prevents high values from dominating the variation when comparing normalized counts. However, this approach also has drawbacks, like masking counts of rare cell populations. That's why more sophisticated methods scale with respect to subsets inferred from different concepts, like the cell or proximity clusters based on gene expression (Lun et al., 2016), or scale with respect to Pearson coefficients adjusted to a probability distribution, like the popular 'scTransform' (Ke et al., 2022) used in the Seurat pipeline. This method adjusts to a negative binomial distribution, as it can model the stochastic counting of events by that distribution. There are, however, many other proposals to obtain the correct distribution. Borella and collaborators, in PsiNorm (Borella et al., 2021), propose using the Pareto distribution due to the scale-free nature of many complex systems. Finally, a costly yet efficient alternative is spiking the cells with a small fraction of constantly expressed genes called spike-ins. ISnorm (Lin et al., 2020) suggests such a method.

Normalization, unlike feature selection, is not generally coupled to the downstream method, so the array of options is not as varied. Additionally, from the aspects that can be mitigated with normalization, sequencing depth and stability of variance considerations are essential. Hence, the list of features a technique considers is less varied than with feature selection. Nevertheless, the more sophisticated algorithms intersect in the variations they consider with another preprocessing step that is essential when comparing different samples: *batch effects*.

3.3.4 Batch effects and data integration

While normalization corrects for technical effects in a run of the sequencing pipeline, 'batch effects' account for variations that occur between different runs of each sample, donor, protocol, or sequencing platform. The main idea is to form a batch of cells that could have a common source of variation, referred to as a batch correlate. There is no agreed-upon method to integrate different datasets, and often steps taken during the normalization phase, mainly data transformation, can inadvertently mask biological effects when removing batch effects. That's why it is crucial to tailor the approach according to the specific experiments and the batch correlates one wishes to filter out. For example, when building the Human Lung Cell Atlas (Sikkema et al., 2022), it was essential not to use inter-individual variability as a batch correlate because capturing the diversity of cells under varying conditions was important. They employed the scANVI tool to integrate various samples.

Preliminarily, an ideal strategy to account for sample preparation errors would be to mix differently prepared samples in a sequencing run. However, this is expensive and not always possible when dealing with data banks. In this regard, the proposed algorithms help integrate different instrument runs and data from various sources. This integration provides a more complete picture of the cell profiles in a specific tissue or type of cancer.

The different correction techniques correlate errors with different scopes, such as systemic batch effects, clusters of similar cells, single cells, and even gene expression profiles. Nevertheless, zero inflation and gene expression distribution are variations also taken into account when normalizing, and the intersection of these factors must be considered when applying both. A good strategy is to apply normalization first because its effects are systematic. Additionally, you can define two different batches and correct them sequentially, being cautious not to overcorrect. A helpful heuristic to consider when determining the order is to correct for the source of variation with the highest impact first so that others are not hidden.

There is no consensus for a broad categorization of available methods, but Ryu et al. (2023) proposes one based on the underlying mathematical approach:

- Linear decomposition based models
- Similarity based methods in reduced dimension space
- Generative models using variational autoencoders

The first approach has been widely used in batch RNA-seq, like in the well-known '*removeBatchEffect*' function in limma (Ritchie et al., 2015). In general, one decomposes the expression matrix X into a sum of the corrected expression matrix G_C (or a factor matrix times their loadings $R_F \times D_F$) and a design matrix that defines, for instance, the batch groups, D_B times its loadings R_B ($X = G_C + D_B \times R_B$). One of the best performing and most used methods (it is the standard match correction method in the scanpy pipeline) in this category is ComBat (Johnson et al., 2007), which uses general lineal models, since optimization is done via the empirical Bayes approximation. Recently, the main author of ComBat developed an improved version named ComBat-seq (Zhang et al., 2020) which considers *zero inflation* and uses a negative binomial distribution that outputs transcript counts instead of a continuous variable.

The second category is better at considering cell variations that are not homogeneous for all cells in a batch and considers the single cell nature of the experiment in a more natural way. To be able to compare similarity it results very handy to have a lower dimensional representation of the expression profile of a cell. This can be achieved via dimensionality reduction methods (see next subsection). Many of the methods in this category do some sort of dimensionality reduction before starting to look for similarities and some even do the correction in the embedded dimension [like scANVI (Xu et al., 2021)], which can be a problem when wanting to perform other downstream analysis. A very handy guideline sheet for comparing the latest techniques and their features can be found in the supplementary material for (Ryu et al., 2023).

Mutually Nearest Neighbours [MNN, used in Seurat Haghverdi et al. (2018)] does not perform dimensionality reduction; instead looks for similarities in the cells of different batches directly with the use of the kNN algorithm. This ends up being computationally expensive, however there have been improvements to the method, such as fastMNN that does dimensionality reduction via PCA, but it still does not perform very good in benchmarks. An underlying assumption that is not considered in MNN is that the variations are at the cluster level. Methods like Harmony (Korsunsky et al., 2019), LIGER (Welch et al., 2019) or ScMerge (Lin et al., 2019) use clustering of cells and optimize for metrics related to these groups, like maximum diversity clustering (used in Harmony). Harmony has been tested in various benchmarks and even though it is not as sophisticated as the other similarity methods, it is comparable, however, and has good speed and cell type recovery.

The third approach makes use of the latest developments in neural networks to consider the possible non-linear nature of the batch effects. *Variational autoencoders* are used hence for this kind of data because they model a probability distribution with the help of neural networks. Broadly, one–*encoder*–network models a latent space probability distribution and a second–*decoder*–network outputs a generative model that tries to reconstruct the expression counts. In this way, a batch coefficient can be separated as a parameter of the distribution. scVI (Lopez et al., 2018), scANVI (Xu et al., 2021), DESC (Li et al., 2020) and scGEN (Lotfollahi et al., 2019) all use this methodology. scANVI is an improvement of scVI that uses cell type information and performs much better than scVI. The latent space representation saves computational power and the consideration of non-linearity allows the correction of a broader range of batch effects. This is why these methods are almost as fast as the linear decomposition ones and as effective as the similarity ones.

Data integration often uses *dimensionality reduction* as a first step, to represent the main features of a transcriptome efficiently and cancelling out the noise. This procedure when projected onto 2 or 3 dimensions also helps to visualize the cells, but there are various caveats to be considered in the next section.

3.3.5 Dimensionality reduction

In the life sciences, *dimensionality reduction* has almost always been done via principal component analysis (PCA). PCA has the advantage of scoring how much each feature contributes to every reduced component, but in high dimensions the difference in distances tends no to vary very much, and as it preserves *raw* euclidean distance it misses a lot of local structure. Nevertheless, because it focuses on maximizing variance, it is good in preserving global structure. That is why, PCA is still used prior to consider other dimensionality reduction methods that are computationally expensive. On the other hand, to recover local structure, used even in bulk sequencing analysis, *t-distributed stochastic neighborhood embedding* (t-SNE) orders data points by sampling from a distribution and attracting or repelling them if they are in the high dimensional neighborhood of other points. The clusters thus obtained have often (though not always!) been shown to coincide with actual cell types. Nevertheless, due to focus on locality, the global structure, that is to say, the position of a cluster with respect to another, is not conserved. Many variations of this underlying approach have been proposed. *Uniform Manifold Approximation and Projection* (UMAP) is one of them, it aims to maintain global structure by fitting the points to a high dimensional uniform manifold.

Nevertheless, it has been shown that its preservation of global structure is even less than the theoretical limit allowed for

embeddings of 2-3 dimensions. Which grows with a complexity of $\mathcal{O}(\sqrt{n})$ (Chari and Pachter, 2022). The search to maintain local and global structure and recent advances in big data have birthed many methods that outperform UMAP in its preservation of global structure. Algorithms like triMAP (Amid and Warmuth, 2019), PaCMAP (Huang H. et al., 2022), art-SNE (Kobak and Berens, 2019) perform fairly well to this end.

Essentially they are all variations of weighing a low-dimensional graph by some nearness metric in the high dimensional space. art-SNE, for instance, runs the t-SNE algorithm with a low and high *perplexity* and takes a mean of the two runs (Perplexity is a measure of how many neighbours to consider as being near a certain point). It manages to preserve more global structure this way but is computationally expensive. tri-MAP was the first of many attempts to achieve recovering actual cell types, using triplets of points that have neighbouring points and randomly sampled far points in the triplet and connecting them to one another. It is fast but has been criticized by the community because its effectiveness depends mainly on pre-processing steps (Huang H. et al., 2022). The authors of PaCMAP, did a very thorough job of laying out the underlying mathematical approach in these nearness graph embeddings (Huang H. et al., 2022). They present a visualization to identify the algorithms that do this approach incorrectly called *rainbow plot*. Having this in mind they propose PacMAP, an algorithm that considers a nearness metric and medium near metric. Their method is fast and auto-adjusts its parameters, a choice that is not systematized in t-SNE and UMAP. Another graph embedding that performs very good by their measures is ForceAtlas2 but is not very time efficient and does not use neural networks. Fortunately, Both et al. (2023) have recently proposed a method that leverages neural networks and the information of the edges in the graph to speed up ForceAtlas2 significantly, especially in cases where there is a lot of distinguishable communities.

The power of deep neural networks is also being leveraged to do this embedding. They are suited because of their ability to handle large-scale high-dimensional data and to incorporate different factors, like batch correction in the same run. In benchmarks (Xiang et al., 2021) they are comparable or even better in speed and accuracy to the best non-linear methods. But there are not many other benchmarks that compare these approaches with the standard ones and the theory of why deep neural networks have managed to classify more accurately in, for example, images, than any of the linear decomposition or graph embedding methods is not standard.

A good way to benchmark the conservation of global structure is by building hierarchical clusters in a high dimensional dataset and checking if the visualization separates this clusters. This feature is the focus of downstream analysis that want to infer trajectories of differentiation in cells, an analysis that we will address later. While methods like PacMAP and art-SNE do conserve this structure sometimes, there is a mathematical argument for using a *hyperbolic space* as the embedding space. Hyperbolic geometry enables the embedding of complex hierarchical data in only two dimensions while preserving the pairwise distances between points in the hierarchy. PoincareMap (Klimovskaia et al., 2020) is a pioneering paper in this regard, that has been optimized for high-throughput and dropouts by Tian T. et al. (2023). Although they are not featured in global benchmarks they do their own with respect to the ones mentioned here and outperform all of them.

3.4 Downstream analyses using scRNA-seq

After applying these corrections and controls, comparability is often assured, and the richness of the data can be leveraged to find structure. Subsequently, this structure can be used to generate and explore hypotheses through analysis. However, it is good practice to check if the conclusions significantly change when excluding or altering the parameters of a quality control method. In the following section, we provide a guide to the most popular and latest downstream analyses used. An overview of these methods can be seen in Figure 3. The central concept in these downstream analyses is the idea of clusters, as they can be correlated with a cell type. This forms the basis for describing the heterogeneity of a tissue, comparing expression between different types, or inferring trajectories.

3.4.1 Clustering

To label every cell as pertaining to a phenotype or cell type, the visualization conferred by dimensionality reduction methods results insufficient, though it is often used as an intermediate step. The basic idea in clustering is to group together the cells that have similar gene expression profiles, frequently via an unsupervised approach. The main strategies that are used for this endeavor are:

- Clustering algorithms by distance
- Community detection
- Hierarchical analysis

However, embedding or dimensionality reduction is commonly used as an early step to then cluster in the reduced space, which can be 2D, 3D or even hyperbolic, and the approach can vary greatly as can be seen in the previous section. Additionally, when the embedding is done with neural networks one can embed into any space and include other factors like batch correction in the same process. Such is the case of scPhere (Ding and Regev, 2021), where they perform an embedding to hyperbolic 2D or 3D space and claim to solve cell crowding and better capture temporal trajectories.

In clustering algorithms the distance between the points is used to minimize inter-cluster distance or to find densely packed regions. The simplest approach for this is k-means clustering, however often it does not recover the *actual* cell types when there are several of them. There are also algorithms that search separate the points according to the differences in density like GiniClust (Jiang et al., 2016), which is an optimized version of the popular VDBSCAN (Ester et al., 1996).

Community detection methods often work with a KNN graph from processed data and infer communities via graph algorithms for finding modules (Alcalá-Corona et al., 2021). This approaches have been the most used because of their reduced complexity and because they only need to use neighbouring nodes for the computation. The scanpy and SEURAT3 pipelines used the Louvain algorithm (Blondel et al., 2008) for a while, but have defected to the Leiden algorithm because it is more efficient and overcomes a flaw of the Louvain algorithm wherein communities could be built that have disconnected components (Anuar et al., 2021). Another way of detecting communities is via spectral decomposition of the adjacency matrix. Continuous Non-negative Matrix Factorization by Puram et al. (2017) does this. A recent, very fast algorithm that uses simplified graphs for community detection via spectral decomposition is Secuer (Wei et al., 2022) it enjoys reduced

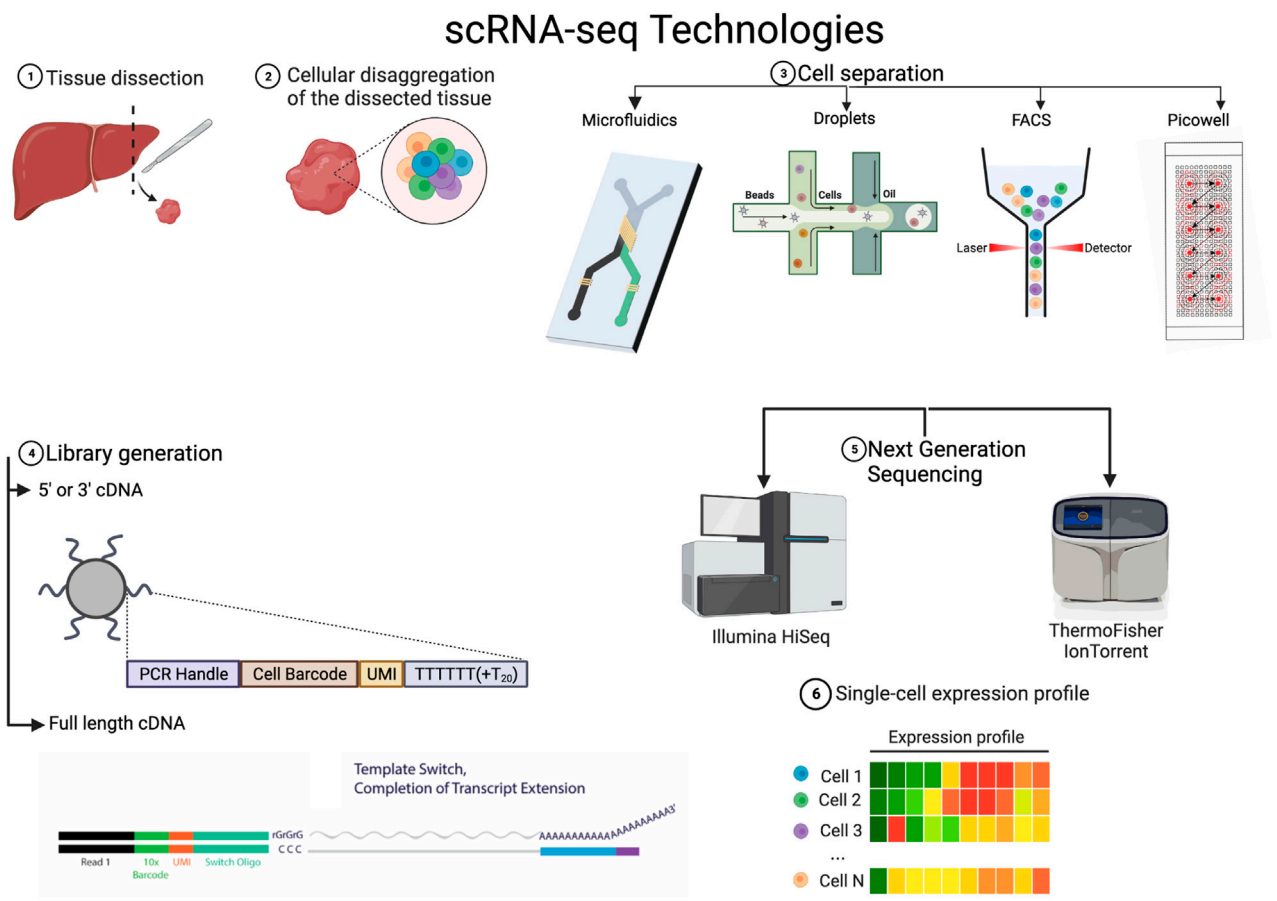

FIGURE 1
Variations on the technical steps for scRNA-seq 1 & 2 Tissue is dissected from the sample of interest and disaggregated. 3. There are various methods to separate the cells, it is to be noted that FACS is often used before in many workflows to select a population of interest or to exclude dead cells. 4. The library generation is done with reverse transcriptases of which there are various kinds but the main difference lies in whether they just transcribe in one direction or both. 5. The NGS platforms to actually sequence and convert to digital data are varied and there always new platforms being developed. 6. All the various technologies lead to a single cell expression profile that has broadly the form in the figure. Created with BioRender.com.

runtime and memory usage over one order of magnitude for datasets with more than 1 million cells.

There are also ensemble methods like single-cell aggregated (From Ensemble) clustering (SAFE) (Yang et al., 2018), Single-cell Aggregated Clustering via Mixture Model Ensemble (SAME) (Huh et al., 2019), single-cell graph partitioning ensemble (Sc-GPE) (Zhu et al., 2020) which use various methods and take a mean of all of them. These have however the problem of extended run time, and an accumulation of the different errors of the methods.

The latest clustering procedures mix the best parts of this algorithms in one process. Like scCAN that uses deep neural networks to perform dimensionality reduction, batch correction and community detection in the reduced space (Tran et al., 2022). scPhere (Ding and Regev, 2021) also does this, while focusing on the hierarchical aspect. As the methods using deep neural network are only just beginning to be used, there are not many benchmarks to really argue for their advantage. Nevertheless, Naitzat et al. (2020) gives a compelling argument, arguing that high dimensional data can be seen as topological manifolds that are transformed in each layer to separate the different categories into simpler lower-dimensional embeddings. The piece-wise linear activation function that these networks use, does a non-continuous transformation that manages to separate the structures better because it can break interconnected structures.

3.4.2 Cluster annotation

A precise definition of *cell type* from single cell analyses remains elusive to date, however the clusters obtained by the methods just mentioned can be assigned to a certain cell identity, e.g., a group of cells that may share various common features. Due to the gigantic variation between experiments, the recommended approach is to try to annotate the clusters automatically, then manually and lastly do a revision by experts.

To annotate the discrete clusters manually, pipelines like Seurat and scanpy resort to a basic differential expression measure, which uses a t-test o a Wilcoxon rank sum (non-parametric) to compare the expression of a gene (or gene set) among all clusters. This delivers the so called marker genes for each cluster which are then compared to known gene expression signatures for a specific cell type. Nevertheless, this approach can be flawed as, for example, surface protein expression does not directly imply that they are present in the surface, nor do they uniquely identify a cell type. That

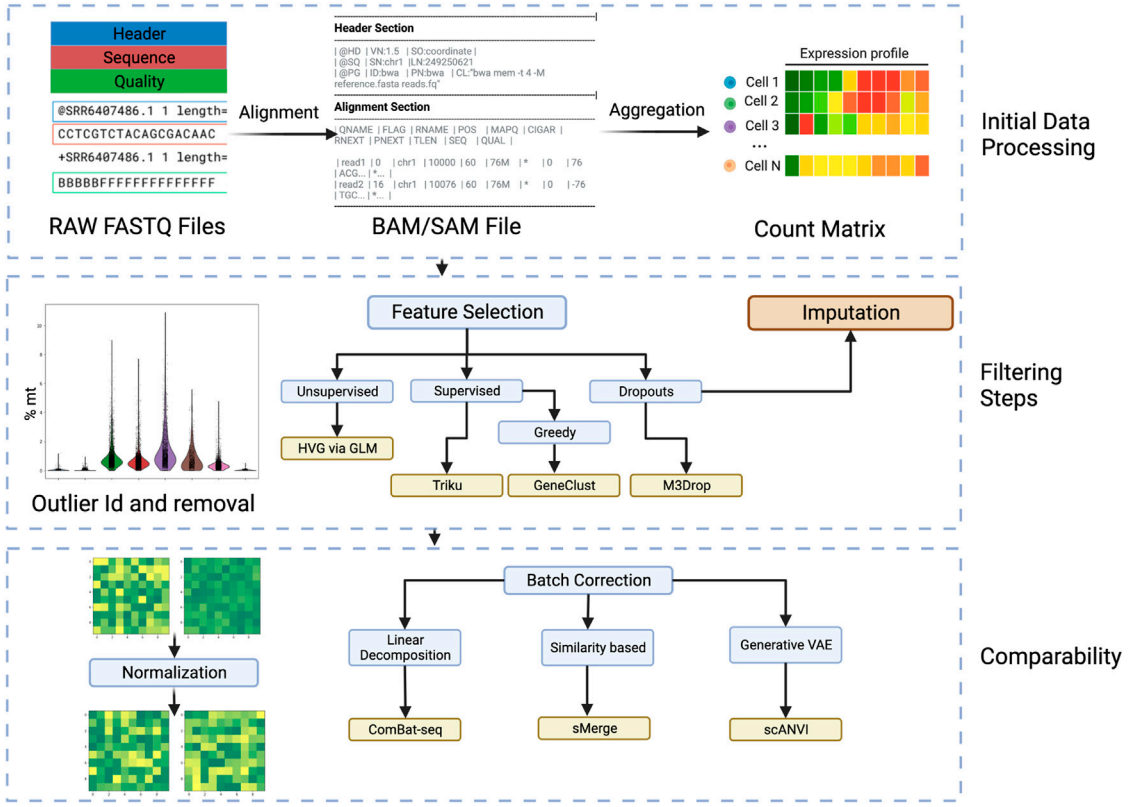

FIGURE 2
Workflow steps for going from raw sequencing files to batch invariant count matrices. Initial Data Processing: In this phase the FASTQ files are aligned to obtain a Sequence Alignment Map (SAM) or Binary Alignment Map (BAM). Afterwards, the aligned files are aggregated into the transcripts and cells to obtain a count matrix. The filtering steps follow, where cell outliers are identified, by, i.e., percentage of mitochondrial genes, and removed. Filtering steps: Passing to the gene scale, and minding the high sparsity and dimension of the data, the selection of features is very important and various considerations and variations can be considered. In particular, using a supervised approach takes more time but can be tailored to the analysis and is more specific than using an unsupervised approach. The dropouts are also considered in various feature selection algorithms. Another approach to consider dropouts is by imputation, but its usefulness is debated. Comparability: To remove the effects that happen due to technical noise and batch preparation, normalization and batch correction are very important. Most researched is the batch correction where several approaches can be taken. Created with BioRender.com.

is why latent embeddings of all the genes expressed are often used to automatically classify cells into annotated clusters.

The automated approach can be done with classifiers or by using reference datasets. Examples of classifiers that are trained on previously annotated data sets or atlases and that consider a large set of genes are CellTypist (Xu et al., 2023) and Clustifyr (Fu et al., 2020). References can be either individual samples of the data set or, ideally, well-curated existing atlases. Query-to-reference mapping can then be performed with methods such as scArches (Lotfollahi et al., 2022), Symphony (Kang et al., 2021) or Azimuth (Hao et al., 2021). Table 2 enlists some of their characteristics.

It is evident that a good annotation depends on the quality of the reference data. That is why endeavors such as *the Human Cell Atlas* https://www.humancellatlas.org/ are paramount to have the most biologically relevant annotations. Similarly, various tools are being developed to upload batch invariant data to this atlas such as Symphony and scPhere (Ding and Regev, 2021).

3.4.3 Tumor cell classification

As can be seen from above, the best way to annotate cells from a tissue is to use a reference atlas. This is nevertheless a problem for neoplastic cells, for they have chromosomal and genetic aberrations and also an altered transcriptomic fingerprint. Additionally, cells in the tumor microenvironment have an altered phenotype, even though they are not neoplastic. To detect neoplastic cells there are many approaches that leverage the underlying molecular aberrations such as: transcript fusions, mutations, virus insertion, copy number aberrations and transcript splicing aberrations. Although there are different techniques (e.g., genome sequencing, CITE-seq, FISH) that are ideal to detect each one of these, efforts have been mode to infer these aberrations from transcriptome sequencing exclusively.

A widely used tool to infer copy number aberrations from scRNA-seq is inferCNV, which looks for large clusters of differential expression located in near chromosomal regions compared against a normal dataset. There is not a single publication that was dedicated to this tool, but various articles by the same group that used this method (e.g., Puram et al., 2017). It is part of a greater endeavour to understand cancer cells from transcriptomic data called trinityCTAT. The majority of the methods in this framework are however designed for bulk RNA-seq. An alternative that uses Bayesian modelling to also infer copy number aberrations is copyKAT (Gao et al., 2021) and it is not limited to large regions.

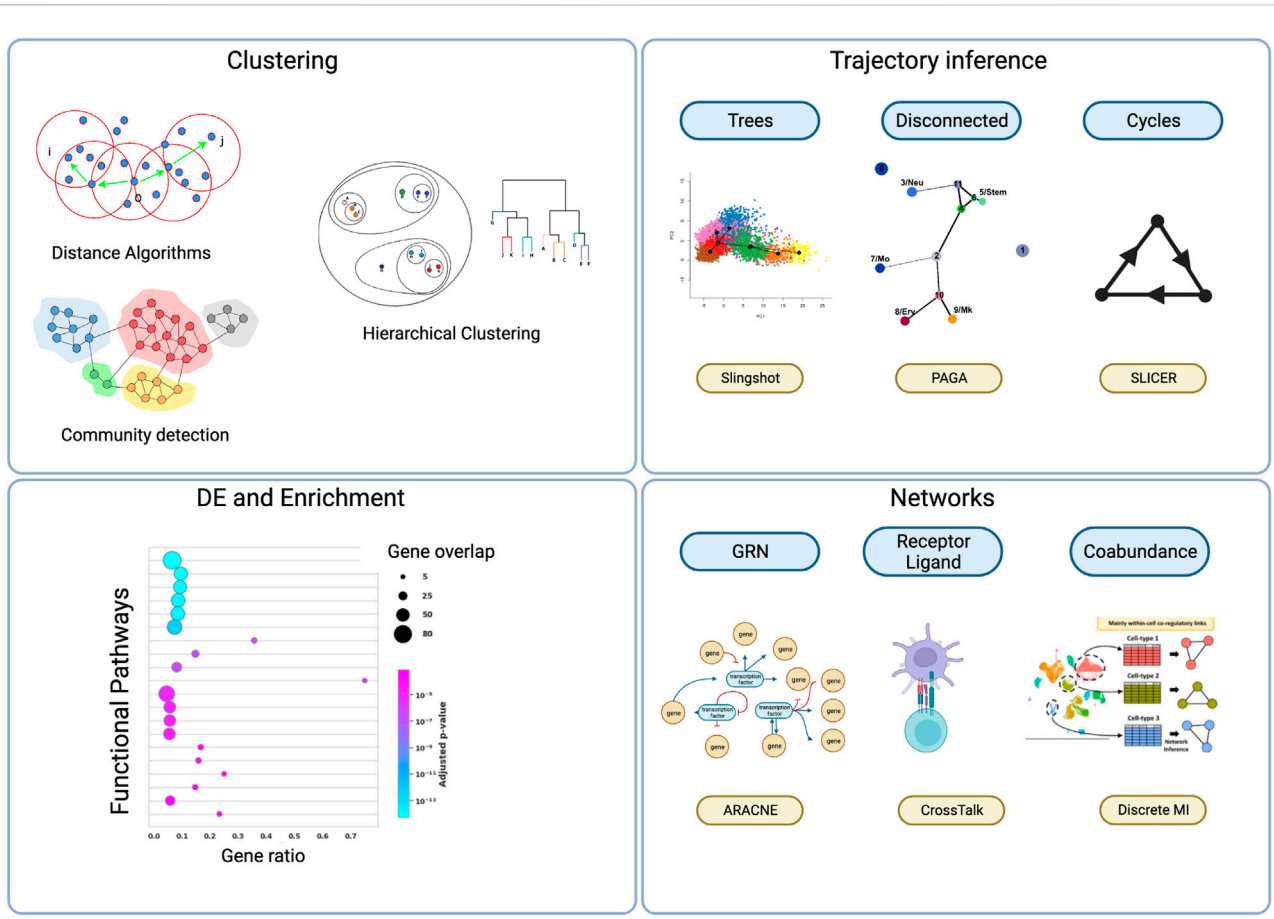

FIGURE 3
The main strategies used to analyze scRNA-seq count matrices. Clustering: By grouping cells based on their genetic expression cell types can be guessed or functional phenotypes can be grouped together. There is a plethora of methods to do this, but the main ones can be grouped by algorithms that consider the distances between the points in the feature space; community detection algorithms that leverage an inferred network from the distances of the points; hierarchical configurations in the consideration of the building of a cluster. Trajectory inference: Single cell data is static, but due to the abundance of cells in various states a trajectory of differentiation can be inferred. These trajectories can be very different depending on the experiment and there are various tools that consider the different cases. DE and Enrichment: Differential expression is done in single-cell over aggregates of cells. There methods to compare pairwise or over multiple clusters. With the differentially expressed genes, an enrichment can be run to look for enriched functional pathways. Networks: The use of networks can encompass various types of interactions between cells and genes. There are various methodologies to infer the networks, the most common one for interaction between cells is the expression or presence (via CITE-seq) of receptors in 1 cell and ligands in the other. A more general way to look at interactions can be through coabundance, which does not exclude exosome or other factors. On the gene level, gene regulatory networks (GRN's) are broadly used to model the effect of transcription factors and to look for hubs. Created with BioRender.com.

Transcript fusion is trickier to detect, because the aberrations are not as big. However unique split-mapped reads and discordant read pairs can be drawn upon in the annotating step when having full-length transcripts. scFusion (Jin et al., 2022) uses this but also applies a deep-learning model and a statistical model to filter out the abundant chimeras that arise from selecting the reads pointed out above. Thus achieving a very low rate of false positives, as they demonstrate with the help of T-cell sequencing that can only have fusions in the V(D)J-region of its TCR domain and also with spike-ins. Mutation and virus insertion detection are not very developed in scRNA-seq data.

Aside from the focus on these molecular aberrations, the reference approach of the previous section can also be leveraged to use reference datasets that have previously marked neoplastic cells. This is the way in which the developers (Dohmen et al., 2022) of ikarus define a gene signature based on reference data. After defining a gene signature from ranked gene sets differences, they train a logistic regression model to classify cells as being normal or tumor-like. There is also a movement to identify alterations of diseased tissue, that leverages the annotations that are produced the aforementioned scArches. The article by Dann et al. (2023) shows a possible framework to do this, but applied to COVID-19 data.

3.4.4 Trajectory inference

Trajectory Inference (TI), alternatively referred to as *pseudo-temporal ordering*, describes one common approach to identify the underlying dynamic cellular processes. While clustering effectively forms distinct groups of cell types and subtypes, it does not consider the variability arising from dynamic cellular processes such as transient cell states in cell differentiation, cell cycles, or environmental influences. TI addresses this limitation by arranging cells along a continuous path that minimizes transcriptional alterations between consecutive cell pairs. This arrangement, known as *pseudotime* (a one-dimensional

TABLE 2 Overview of cell cluster annotation methods and their characteristics.

Method	Category	Description	Pros	Cons
CellTypist	Classifier-based	A pre-trained classifier that uses machine learning algorithms for cell-type annotation, based on a reference atlas of known cell types; considers a large set of genes	Efficient, automatic annotation, generalizable	Affected by classifier type and training data quality; difficult to assess, may require manual verification
Clustifyr	Classifier-based	A pre-trained classifier that uses nearest centroid classification; it is trained on previously annotated datasets or atlases and considers a large set of genes	Efficient, automatic annotation, generalizable	Affected by classifier type and training data quality; difficult to assess, may require manual verification
scArches	Reference mapping	A method that leverages autoencoders for integrating and mapping query datasets to existing annotated single-cell references, allowing for label transfer on the resulting joint embedding	Automatic annotation, integrates heterogeneous datasets	Affected by reference data quality, model, and dataset suitability, may require manual verification
Symphony	Reference mapping	A method that uses mutual nearest neighbors and graph-based signal propagation for mapping to existing annotated single-cell references, enabling label transfer on the resulting joint embedding	Automatic annotation, scalable, robust to batch effects	Affected by reference data quality, model, and dataset suitability, may require manual verification
Azimuth	Reference mapping	A web-based tool that uses Seurat v4 for reference-based mapping and label transfer; performs label transfer on the resulting joint embedding by finding nearest neighbors in the reference data	Automatic annotation, user-friendly interface, handles diverse datasets	Affected by reference data quality, model, and dataset suitability, may require manual verification

TABLE 3 Overview of applications of scRNAseq for cancer research.

Technology	Experiment	References
Annotating	Identification of cell identities in various tissues, examining the heterogeneity of the TME and its effects on the course of the disease	Tirosh et al. (2016)
Annotating	Showed how the activation of hallmarks can be achieved in many ways	Kinker et al. (2019)
Annotating	Proposed an algorithm to predict the complexity of the development of a neoplasm	Woo et al. (2019)
Trajectory Inference	Detailed transitions happening in the TME from ulcerative colitis to UC-associated colon cancer	Wang et al. (2021)
Trajectory Inference	Localized two distinct transcriptional trajectories in Willms cancer	Young et al. (2018)
Trajectory Inference	Conducted trajectory inference analyses on infiltrating T cells in cases of liver cancer	Zhang et al. (2019)
Trajectory Inference	Detected transitions between cellular states in small-cell lung cancer	Guo et al. (2018)
Differential Expression	Developed an algorithm (HEART) to detect differentially expressed genes in cancer	Yuan et al. (2022)
Differential Expression	Analysed the deregulation of angiogenesis in two types of bone cancer	Feleke et al. (2022)
Gene Regulatory Networks	Described a complex handling of EMT by a network of transcription factors	Nam et al. (2021)
Gene Regulatory Networks	Showed advantage of single-cell over bulk transcriptomics in identifying a population of cells in melanoma	Wouters et al. (2020)
Gene Regulatory Networks	Found stemness related populations in hepatocellular carcinoma	Ho et al. (2019)
Cellular Interactions	Attempted to find pan-cancer interactions; found that a subset of tumor-associated macrophages may regulate the abundance of dysfunctional T cells through cytokine/chemokine signaling	Hong et al. (2021)
Cellular Interactions	Probed the TME surrounding co-opted vessels in lung cancer metastasis; inferred interactions through the expression of receptors and ligands, suggesting a putative involvement of macrophage subtypes in tumor-vessel cooption	Teuwen et al. (2021)
Cellular Interactions	Found that tumor-associated macrophages suppress tumor T cell infiltration and TIGIT-NECTIN2 interaction regulates the immunosuppressive environment	Ho et al. (2021)

manifold), signifies the progression of a cell through its dynamic processes, as measured by the transcriptional changes that occur during a biological process.

When considering which approach to use, an important factor is the expected trajectory of cell differentiation. One type of cell can differentiate into multiple types of cells (multi-branching) or just two (bifurcation). There can also be a dedifferentiation process (cycles), and a group of cells might not have a common progeny with another group of cells (disconnected). This is referred to as the topology of the trajectory (represented by a graph), and there are algorithms that can try to deduce it, or you can specify the topology. Another biological consideration is the specification of a starting cell and/or end cell. When the expected topology is unknown, trajectories and downstream hypotheses should be confirmed by multiple trajectory inference methods using different underlying assumptions. A thorough and updated resource, {dynverse} http://

guidelines.dynverse.org/, even has a decision tree to help decide what analysis to use. Outstanding mentions include PAGA (Wolf et al., 2019) for free trajectories, PAGA Tree (op. cit) and Slingshot (Street et al., 2018) for tree like trajectories.

The inferred trajectories may, however, not coincide with actual biologic entities. That is why pseudotime measures that take advantage of the biological information available mey help in adhering to the biology. *RNA-velocity* (Manno et al., 2018) relies on the presence of spliced vs. unspliced RNA and places a cell later or sooner in time according to where in the spectrum with respect to other cells it lies on. It is however limited to data that has been fully sequenced and assumes constant splicing rates, which can be verified by how the splicing rates are distributed. The most accepted tool for this inference is CellRank (Lange et al., 2022) with the use of the scVelo algorithm (Bergen et al., 2020). Other biological factors can be incorporated via lineage tracing, wherein various factors like naturally occurring genetic mutations, Cas9 perturbation data among other things are used. Tools that implement this are Cassiopea (Jones et al., 2020) and LineageOT (Forrow and Schiebinger, 2021).

3.4.5 Differential expression and gene set enrichment

With the annotation of clusters, there is already rich information about the heterogeneity of the sample, but the depth of the data can be further explored to look for variations at the gene level. *Differential gene expression* can be used to propose biological targets, check for differences in treatment, and support further downstream analysis. Additionally, a more accurate list of marker genes for cluster annotation could be obtained from more sophisticated methods. The expression of a gene can be plotted as a gradient to see how it changes along the population, but quantification needs to account for the various effects of cell variance, sample variance, and methodological variance when comparing genes. There are two broad approaches to consider these variations: the *pseudo-bulk* and the *individual cell approach*. Pseudo-bulks aggregate the gene expressions of all cells in labels (clusters) and compare expressions of genes across labels by taking advantage of methods already used in bulk RNA-seq. The best ranking and most widely used (Heumos et al., 2023) are DeSeq2 (Love et al., 2014), limma (Ritchie et al., 2015) and edgeR (Robinson et al., 2010). They can also be weighted by ZINB-Wave (Risso et al., 2018) to consider non biological zeroes and the stochasticity of the data.

The variation across cells has been modelled with various distributions including GLM, GAM and Hurdle models, as well as non parametric models. There exist methods that perform comparisons against many labels at once but they are very costly in computational resources and do not perform much better, so we will stick to the bimodal models. The most popular and successful one is MAST (Finak et al., 2015). It uses generalized linear hurdle models that consider the zero counts. It is less time consuming than pseudo-bulk methods with weights. Sadly it has been demonstrated to underestimate the variability of gene expression and have a tendency to misclassify highly expressed genes as exhibiting differential expression (Squair et al., 2021), when compared to the pseudo-bulk methods. A good candidate that considers the cell-level, does not misclassify highly expressed genes and is much faster than some similar methods is NEBULA (He et al., 2021). It has also been benchmarked against MAST and other popular methods and has resulted the best overall in several metrics.

Building on top of differential expression, to be able to hint at more functional aspects of the tissue, the enrichment of a set of genes or gene profiles can be searched for in an enrichment cluster. To this end, enrichment frameworks such as decoupleR (i Mompel et al., 2022) provide access to different databases and methods in a single tool. Another proponent that works well with the scanpy framework is GSEApy (Fang et al., 2022) which leverages the GO database. Enrichment methods developed for bulk transcriptomics can be applied to scRNA-seq, but some single-cell-based methods, such as Pagoda (Fan et al., 2016), might outperform them. Although cluster analysis falls short in revealing the continuous range of states and the gene expression programs (GEPs) that are shared across various cell types, scAAnet, an autoencoder for single-cell non-linear archetypal analysis, has the ability to detect GEPs and deduce the proportional activity of each GEP among different cells (Wang and Zhao, 2022).

3.4.6 Networks

As has been shown in the previous sections, the use of methodologies that use graphs as mathematical objects is extensive. There are, however, ways to use networks in single-cell analysis that leverage many of their properties, such as their mesoscopic quantities and the ability to model dynamic processes with their help. Chief among these are the inference of cell-cell communication and gene regulatory networks (GRN's).

Inference of cell-cell communication is mainly done by differential expression of ligands and receptors in clusters. However, the extracellular matrix, transporters, physical interactions, and secreted vesicles can also be taken into account (Türei et al., 2021). A recent review (Dimitrov et al., 2022) considers the methods that just use ligand-receptor interaction and finds that the libraries they use do not have much overlap. These tools use varied basic statistical inferences, and cross-talk weighs the scores with the probability of autocrine signaling. The authors recommend using their tool LIANA, which provides an overall ranking for several combinations of methods. Additionally, there are frameworks that go on to infer inter-cellular signalling and functions from the receptor-ligand interactions like NicheNet (Browaeys et al., 2020). The performance of this tools depends on the tissue, and their approach seems to favor co-localized interactions. It is recommended to support this inferences with spatial transcriptomics. The networks obtained thusly can then be analyzed for connectivity, hubs and dynamic changes.

Gene regulatory networks draw inspiration from intracellular regulations such as transcription factors, second messengers, enhancers, and promoters. They use measures of co-expression, either in a snapshot or over time, to infer a connection between two genes and ultimately construct a network. The literature contains various measures of co-expression, although it is accepted that linear correlations cannot capture the complexity of the regulations occurring within cells. Two leading candidates are *Mutual Information* and *Spearman correlation*. Mutual information requires many samples to reconstruct the probability distribution functions but is the measure that can generally capture these regulations more comprehensively because a low score in mutual information indicates statistical independence between genes, which

cannot be ascertained with other measures. The actual biological information is very complex to decipher, as the presence of co-expression does not necessarily mean a direct regulation from one gene to another. The pioneering work by Margolin et al. (Margolin et al., 2006) that proposed mutual information as a viable candidate for modelling gene expression continues to be applied in various algorithms. One that is very scalable is ARACNEap (Lachmann et al., 2016), which uses and adjustable discretization of the gene expression values to reduce computation time. Care has to be taken when implementing this algorithms to single-cell data because of the sparsity. Nevertheless, a correlation has often been observed between sets of interconnected genes and a physiological function. Frameworks like Epoch (Su et al., 2022) or CellOracle (Kamimoto et al., 2023), take advantage of this fact to propose alterations in the expression of hubs of communities that can direct differentiation to another cell type.

Some other projects that are widely used are SCODE (Matsumoto et al., 2017), PIDC (Chan et al., 2017), SCENIC (Aibar et al., 2017), though they have been shown to perform poorly (Chen and Mar 2018). There have also been attempts to leverage neural networks to infer GRN's like with ScGRNs (Turki and Taguchi, 2020). To assess the performance of this emergent algorithms, BEELINE (Pratapa et al., 2020) proposes a framework to evaluate them, with the help of literature curated Boolean networks, predictable trajectories and other means.

4 Selected applications in cancer

The elucidating analyses that can be applied with single-cell transcriptomics have been used in all kinds of experiments to explore the intricacies of cancer. We have outlined important papers for every topic in Table 3. First, just by *annotating* the cell identities in various tissues, the heterogeneity of the TME and its effects on the course of the disease has grown. For example, while bulk sequencing classifies melanoma as MITF-high or AXL-high, at the single-cell level, every tumor contains malignant cells corresponding to both states (Tirosh et al., 2016). Also, in a study by Kinker et al. (Kinker et al., 2019), it is shown how the activation of hallmarks can be achieved in various ways. Researchers have observed that both EMT and senescence are associated with precise phenotypes and well-defined regulators during development and wound healing. However, in the context of tumors and cancer cell lines, these researchers have observed only partial phenotypes and limited dependence on these regulators. Kinker et al. (2019), on the other hand, puts a limit on the expression diversity observed in patient samples. In contrast, Woo and others (Woo et al., 2019) proposed an algorithm for predicting the complexity of neoplasm development as a prognostic marker based on the composition of the TME.

A malignant tumour can have multiple differentiation and dedifferentiation processes happening in its cells as well as in its surroundings. That is why *trajectory inference* has been used to detail the kind of transitions that happen in the TME. In a study by (Wang et al., 2021) the precise cellular composition and developmental trajectory from ulcerative colitis (UC) to UC-associated colon cancer was analyzed, and it was predicted that CD74, CLCA1, and DPEP1 played a potential role in disease progression. (Young et al., 2018). localized two distinct transcriptional trajectories in Wilms cancer. These trajectories correspond to the development of nephrogenic rest cells and Wilms cancer cells, respectively, and provide support for the hypothesis that Wilms tumor cells arise due to anomalies in fetal nephrogenesis originating from cells of the urethric bud. Similarly, trajectory inference analyses conducted on infiltrating T cells in cases of liver cancer (Zhang et al., 2019) and small-cell lung cancer, like from (Guo et al., 2018) have detected transitions between cellular states, specifically between the proliferating/activated state and the exhausted state.

The cellular heterogeneity in cancer poses a challenge to *differential expression* analysis. That is why (Yuan et al., 2022) developed an algorithm to detect differentially expressed genes in cancer called HEART. With this tool they identified several potential blood based biomarkers associated with colorectal cancer metastasis. Another study by (Feleke et al., 2022) analysed the deregulation of angiogenesis in two types of bone cancer; giant cell tumor bone andosteosarcoma. It found that the deregulation of the different VEGF factors is tissue specific and can be used as target treatment.

Gene regulatory networks can provide another way of identifying the functionality of the cells in cancer and their possible evolution. (Nam et al., 2021). describes a complex handling of EMT by a network of transcription factors such as SNAI1, SNAI2, ZEB1, TWIST1 and other regulators. (Wouters et al., 2020) in turn, show how single-cell has an advantage over bulk transcriptomics, by identifying a population of cells in melanoma that had been thought of as 2 cell types, *melanocyte* and *mesenchymal*, was confirmed as just one intermediate state with both expression programs. (Ho et al., 2019) found stemness related populations in hepatocellular carcinoma.

Cellular interactions are extremely important especially because the tumor has the ability to shape its TME. (Hong et al., 2021) tried to find pan-cancer interactions and found that a subset of tumor-associated macrophages (TAM), PLTP + C1QC + TAMs, may regulate the abundance of dysfunctional T cells through cytokine/chemokine signaling. More specifically (Teuwen et al., 2021) probed the TME surrounding co-opted vessels in lung cancer metastasis. Transcriptomic results, with the inference of interactions through the expression of receptors and ligands, may suggest a putative involvement of macrophage subtypes in tumor-vessel cooption. Also there are various advancements understanding immunoedition. (Ho et al., 2021) found that tumor-associated macrophages suppress tumor T cell infiltration and TIGIT-NECTIN2 interaction regulates the immunosuppressive environment.

Taken together all these applications are part of a new approach to cancer where the heterogeneity in the TME is paramount. Be it by finding rare cell types, considering various interactions or describing new differentiation pathways. One could think this applications help mostly for precision medicine, but the understanding of the physiology has also advanced because of this technology.

4.1 Single cell approaches are marking a difference in oncology studies

Single-cell technologies have had a transformative impact on cancer research by enabling researchers to delve into the heterogeneity of tumors at an unprecedented level of detail

(Ortega et al., 2017; Fan et al., 2018; Levitin et al., 2018; Ding et al., 2020; Wu F. et al., 2021). Some critical applications of single-cell technologies in cancer research and how their continued use is expected to further transform the field are shown below:

Single-cell RNA sequencing (scRNA-seq) has revealed the immense heterogeneity within tumors, identifying various cell types and transcriptional states (Nieto et al., 2021; Blise et al., 2022; Li C. et al., 2023). Researchers have used this technology to dissect clonal evolution (Losic et al., 2020; Miles et al., 2020; Morita et al., 2020; Nam et al., 2021), identifying driver mutations (Li et al., 2012; Roerink et al., 2018; Huang Z. et al., 2022), and tracking the emergence of drug-resistant subclones (Prieto-Vila et al., 2019; Liu L. et al., 2023; Li X. et al., 2023). Continued use of single-cell technologies will provide deeper insights into the evolution of tumors over time. This understanding is crucial for developing personalized treatment strategies and targeting therapy-resistant cell populations (Ding et al., 2020; Wang X. et al., 2022).

Single-cell approaches have been also instrumental in characterizing the tumor microenvironment, identifying different immune cell populations, and deciphering their functional states (Guo et al., 2018; Yuan et al., 2019; Van der Leun et al., 2020; Ren et al., 2021). This has led to discoveries related to immune evasion mechanisms in cancer (Sun et al., 2021; Liu W. et al., 2023). Ongoing use of single-cell technologies will contribute to the development of more effective immunotherapies (Gohil et al., 2021; Heinrich et al., 2021). Researchers will be able to design therapies that target specific immune cell subsets or reverse immunosuppressive signals within the tumor microenvironment (Davis-Marcisak et al., 2021).

Single-cell techniques have as well enabled the identification of rare and previously overlooked cell types within tumors, such as cancer stem cells or metastasis-initiating cells (Lawson et al., 2015; Kester and Van Oudenaarden, 2018; Orrapin et al., 2023). These discoveries have profound implications for understanding tumor initiation and progression. Continued use will likely uncover even rarer cell types and their roles in cancer. Targeting these cell populations could lead to novel therapeutic strategies.

Single-cell genomics has provided insights into the molecular mechanisms underlying drug resistance (Eyler et al., 2020). By profiling single cells, researchers have identified subpopulations with distinct resistance mechanisms (Tirier et al., 2021). The ongoing application of single-cell technologies will facilitate the development of more effective targeted therapies and strategies to overcome drug resistance. Precision medicine will become increasingly tailored to individual patients based on their tumor's unique molecular profile.

ScRNASeq analysis of circulating tumor cells (CTCs) and cell-free DNA (cfDNA) has enabled early cancer detection and monitoring (Eyler et al., 2020). This has implications for cancer screening and tracking treatment responses. As single-cell technologies continue to improve, the sensitivity and specificity of liquid biopsies will increase (Li et al., 2019; Lim et al., 2019; Pei et al., 2020). This non-invasive approach may become a routine part of cancer diagnosis and treatment monitoring (Li et al., 2022; Zhou et al., 2022).

One must also consider how related approaches such as single-cell epigenomic profiling have revealed epigenetic alterations in cancer cells that drive gene expression changes (Pierce et al., 2021; Casado-Pelaez et al., 2022). Spatial profiling techniques provide insights into the spatial organization of cells within tumors (Wu S. Z. et al., 2021). Combining single-cell transcriptomics with epigenomics and spatial data will offer a holistic view of the tumor (Ogbeide et al., 2022; Preissl et al., 2023). This integrated approach will elucidate the regulatory networks that govern cancer cell behavior and potentially identify new therapeutic targets.

Hence, single-cell technologies have already revolutionized cancer research by providing a deeper understanding of tumor heterogeneity, immune responses, and drug resistance mechanisms. Their continued use is expected to drive further discoveries, leading to more precise diagnostics, targeted therapies, and personalized treatment approaches. As these technologies become more accessible and sophisticated, they hold the potential to transform cancer research and patient care in the years to come (Suvà and Tirosh, 2019; Janiszewska et al., 2020; Wu F. et al., 2021).

5 Conclusion

We have presented here the state of the art in approaching the study of cancer through the means of single cell transcriptome sequencing. A summary of the latest methods and technologies to carry out this experiments and some interesting applications. Though there is a lack of golden standards to use, there is a lot of ingenuity in the new methods being developed and there is a constant effort to benchmark them independently.

The applications that continue to arise thanks to this technology go from building an atlas of the possible expression profiles in all types of cancers, through proposal of prognostic markers, elucidation of therapy resistance, explanation of alternative cancer hallmarks mechanisms, verification of cell lines and organoid simulations, inference of mechanisms of immune edition to proposal of targeted therapeutic agents.

scRNA-seq lies at the middle of all the possible molecular pathways that can be sequenced and can be greatly enhanced by aggregating it with spatial information and the other omics.

Author contributions

EP-O: Conceptualization, Investigation, Writing–original draft. EH-L: Conceptualization, Investigation, Writing–review and editing. GA-J: Conceptualization, Investigation, Writing–review and editing.

Funding

The author(s) declare financial support was received for the research, authorship, and/or publication of this article. This work has been supported by Intramural Funds from the National Institute of Genomic Medicine, Project 494-2022.

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

References

Aibar, S., González-Blas, C. B., Moerman, T., Huynh-Thu, V. A., Imrichova, H., Hulselmans, G., et al. (2017). Scenic: single-cell regulatory network inference and clustering. *Nat. Methods* 14, 1083–1086. doi:10.1038/nmeth.4463

Aicher, T. P., et al. (2019). Single cell methods, sequencing and proteomics. *Nat. Methods* 1979, 111–132.

Alcalá-Corona, S. A., Sandoval-Motta, S., Espinal-Enriquez, J., and Hernandez-Lemus, E. (2021). Modularity in biological networks. *Front. Genet.* 12, 701331. doi:10.3389/fgene.2021.701331

Amid, E., and Warmuth, M. K. (2019). *Trimap: large-scale dimensionality reduction using triplets. arXiv preprint arXiv:1910.00204.*

Andrews, T. S., and Hemberg, M. (2018). M3drop: dropout-based feature selection for scrnaseq. *Bioinformatics* 35, 2865–2867. doi:10.1093/bioinformatics/bty1044

Anuar, S. H. H., Abas, Z. A., Yunos, N. M., Zaki, N. H. M., Hashim, N. A., Mokhtar, M. F., et al. (2021). Comparison between louvain and leiden algorithm for network structure: a review. *J. Phys. Conf. Ser.* 2129, 012028. doi:10.1088/1742-6596/2129/1/012028

Ascensión, A., Ibáñez-Solé, O., Inza, I., Izeta, A., and Araúzo-Bravo, M. J. (2022). Triku: a feature selection method based on nearest neighbors for single-cell data. *GigaScience* 11, giac017. doi:10.1093/gigascience/giac017

Bacher, R., and Kendziorski, C. (2016). Design and computational analysis of single-cell rna-sequencing experiments. *Genome Biol.* 17, 63–14. doi:10.1186/s13059-016-0927-y

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

Barron, M., and Li, J. (2016). Identifying and removing the cell-cycle effect from single-cell rna-sequencing data. *Sci. Rep.* 6, 33892. doi:10.1038/srep33892

Bell, C. C., Fennell, K. A., Chan, Y.-C., Rambow, F., Yeung, M. M., Vassiliadis, D., et al. (2019). Targeting enhancer switching overcomes non-genetic drug resistance in acute myeloid leukaemia. *Nat. Commun.* 10, 2723. doi:10.1038/s41467-019-10652-9

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

Birnbaum, K. D. (2018). Power in numbers: single-cell rna-seq strategies to dissect complex tissues. *Annu. Rev. Genet.* 52, 203–221. doi:10.1146/annurev-genet-120417-031247

Blise, K. E., Sivagnanam, S., Banik, G. L., Coussens, L. M., and Goecks, J. (2022). Single-cell spatial architectures associated with clinical outcome in head and neck squamous cell carcinoma. *NPJ Precis. Oncol.* 6, 10. doi:10.1038/s41698-022-00253-z

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

Borella, M., Martello, G., Risso, D., and Romualdi, C. (2021). Psinorm: a scalable normalization for single-cell rna-seq data. *Bioinformatics* 38, 164–172. doi:10.1093/bioinformatics/btab641

Both, C., Dehmamy, N., Yu, R., and Barabási, A.-L. (2023). Accelerating network layouts using graph neural networks. *Nat. Commun.* 14, 1560. doi:10.1038/s41467-023-37189-2

Browaeys, R., Saelens, W., and Saeys, Y. (2020). Nichenet: modeling intercellular communication by linking ligands to target genes. *Nat. Methods* 17, 159–162. doi:10.1038/s41592-019-0667-5

Burrell, R. A., McGranahan, N., Bartek, J., and Swanton, C. (2013). The causes and consequences of genetic heterogeneity in cancer evolution. *Nature* 501, 338–345. doi:10.1038/nature12625

Casado-Pelaez, M., Bueno-Costa, A., and Esteller, M. (2022). Single cell cancer epigenetics. *Trends Cancer* 8, 820–838. doi:10.1016/j.trecan.2022.06.005

Chan, T. E., Stumpf, M. P., and Babtie, A. C. (2017). Gene regulatory network inference from single-cell data using multivariate information measures. *Cell Syst.* 5, 251–267. doi:10.1016/j.cels.2017.08.014

Chari, T., and Pachter, L. (2022). *The specious art of single-cell genomics*. bioRxiv. doi:10.1101/2021.08.25.457696

Chen, S., and Mar, J. C. (2018). Evaluating methods of inferring gene regulatory networks highlights their lack of performance for single cell gene expression data. *BMC Bioinforma.* 19, 232. doi:10.1186/s12859-018-2217-z

Dal Molin, A., and Di Camillo, B. (2019). How to design a single-cell rna-sequencing experiment: pitfalls, challenges and perspectives. *Briefings Bioinforma.* 20, 1384–1394. doi:10.1093/bib/bby007

Dann, E., Cujba, A.-M., Oliver, A. J., Meyer, K. B., Teichmann, S. A., and Marioni, J. C. (2023). Precise identification of cell states altered in disease using healthy single-cell references. *Nat. Genet.* 1. doi:10.1038/s41588-023-01523-7

Dar, R. D., Razooky, B. S., Singh, A., Trimeloni, T. V., McCollum, J. M., Cox, C. D., et al. (2012). Transcriptional burst frequency and burst size are equally modulated across the human genome. *Proc. Natl. Acad. Sci.* 109, 17454–17459. doi:10.1073/pnas.1213530109

Davis, A., Gao, R., and Navin, N. E. (2019). Scopit: sample size calculations for single-cell sequencing experiments. *BMC Bioinforma.* 20, 566–6. doi:10.1186/s12859-019-3167-9

Davis-Marcisak, E. F., Deshpande, A., Stein-O'Brien, G. L., Ho, W. J., Laheru, D., Jaffee, E. M., et al. (2021). From bench to bedside: single-cell analysis for cancer immunotherapy. *Cancer Cell* 39, 1062–1080. doi:10.1016/j.ccell.2021.07.004

DeLaughter, D. M. (2018). The Use of the fluidigm C1 for RNA expression analyses of aingle cells. *Curr. Protoc. Mol. Biology* 122, e55.

Dimitriu, M. A., Lazar-Contes, I., Roszkowski, M., and Mansuy, I. M. (2022). Single-cell multiomics techniques: from conception to applications. *Front. Cell Dev. Biol.* 10, 854317. doi:10.3389/fcell.2022.854317

Dimitrov, D., Türei, D., Garrido-Rodriguez, M., Burmedi, P. L., Nagai, J. S., Boys, C., et al. (2022). Comparison of methods and resources for cell-cell communication inference from single-cell rna-seq data. *Nat. Commun.* 13, 3224. doi:10.1038/s41467-022-30755-0

Ding, J., and Regev, A. (2021). Deep generative model embedding of single-cell rna-seq profiles on hyperspheres and hyperbolic spaces. *Nat. Commun.* 12, 2554. doi:10.1038/s41467-021-22851-4

Ding, S., Chen, X., and Shen, K. (2020). Single-cell rna sequencing in breast cancer: understanding tumor heterogeneity and paving roads to individualized therapy. *Cancer Commun.* 40, 329–344. doi:10.1002/cac2.12078

Dohmen, J., Baranovskii, A., Ronen, J., Uyar, B., Franke, V., and Akalin, A. (2022). Identifying tumor cells at the single-cell level using machine learning. *Genome Biol.* 23, 123. doi:10.1186/s13059-022-02683-1

Dong, X., Wang, F., Liu, C., Ling, J., Jia, X., Shen, F., et al. (2021). Single-cell analysis reveals the intra-tumor heterogeneity and identifies mlxipl as a biomarker in the cellular trajectory of hepatocellular carcinoma. *Cell death Discov.* 7, 14. doi:10.1038/s41420-021-00403-5

Ester, M., Kriegel, H.-P., Sander, J., and Xu, X. (1996). A density-based algorithm for discovering clusters in large spatial databases with noise. *KDD'96* 1996, 226–231.

Eyler, C. E., Matsunaga, H., Hovestadt, V., Vantine, S. J., van Galen, P., and Bernstein, B. E. (2020). Single-cell lineage analysis reveals genetic and epigenetic interplay in glioblastoma drug resistance. *Genome Biol.* 21, 174–221. doi:10.1186/s13059-020-02085-1

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

Fan, J., Salathia, N., Liu, R., Kaeser, G. E., Yung, Y. C., Herman, J. L., et al. (2016). Characterizing transcriptional heterogeneity through pathway and gene set overdispersion analysis. *Nat. Methods* 13, 241–244. doi:10.1038/nmeth.3734

Fang, Z., Liu, X., and Peltz, G. (2022). GSEApy: a comprehensive package for performing gene set enrichment analysis in Python. *Bioinformatics* 39, btac757. doi:10.1093/bioinformatics/btac757

Feleke, M., Feng, W., Song, D., Li, H., Rothzerg, E., Wei, Q., et al. (2022). Single-cell rna sequencing reveals differential expression of egfl7 and vegf in giant-cell tumor of bone and osteosarcoma. *Exp. Biol. Med.* 247, 1214–1227. doi:10.1177/15353702221088238

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

Fleming, S. J., Chaffin, M. D., Arduini, A., Akkad, A.-D., Banks, E., Marioni, J. C., et al. (2022). *Unsupervised removal of systematic background noise from droplet-based single-cell experiments using cellbender*. bioRxiv. doi:10.1101/791699

Forrow, A., and Schiebinger, G. (2021). Lineageot is a unified framework for lineage tracing and trajectory inference. *Nat. Commun.* 12, 4940. doi:10.1038/s41467-021-25133-1

Fu, R., Gillen, A. E., Sheridan, R. M., Tian, C., Daya, M., Hao, Y., et al. (2020). clustifyr: an r package for automated single-cell rna sequencing cluster classification. *F1000Research* 9, 223. doi:10.12688/f1000research.22969.2

Gao, R., Bai, S., Henderson, Y. C., Lin, Y., Schalck, A., Yan, Y., et al. (2021). Delineating copy number and clonal substructure in human tumors from single-cell transcriptomes. *Nat. Biotechnol.* 39, 599–608. doi:10.1038/s41587-020-00795-2

Gerdes, M. J., Sevinsky, C. J., Sood, A., Adak, S., Bello, M. O., Bordwell, A., et al. (2013). Highly multiplexed single-cell analysis of formalin-fixed, paraffin-embedded cancer tissue. *Proc. Natl. Acad. Sci.* 110, 11982–11987. doi:10.1073/pnas.1300136110

Gohil, S. H., Iorgulescu, J. B., Braun, D. A., Keskin, D. B., and Livak, K. J. (2021). Applying high-dimensional single-cell technologies to the analysis of cancer immunotherapy. *Nat. Rev. Clin. Oncol.* 18, 244–256. doi:10.1038/s41571-020-00449-x

Guillaumet-Adkins, A., Rodríguez-Esteban, G., Mereu, E., Mendez-Lago, M., Jaitin, D. A., Villanueva, A., et al. (2017). Single-cell transcriptome conservation in cryopreserved cells and tissues. *Genome Biol.* 18, 45–15. doi:10.1186/s13059-017-1171-9

Guo, X., Zhang, Y., Zheng, L., Zheng, C., Song, J., Zhang, Q., et al. (2018). Global characterization of t cells in non-small-cell lung cancer by single-cell sequencing. *Nat. Med.* 24, 978–985. doi:10.1038/s41591-018-0045-3

Hagemann-Jensen, M., Ziegenhain, C., Chen, P., Ramsköld, D., Hendriks, G-J., Larsson, A. J. M., et al. (2020). Single-cell RNA counting at allele and isoform resolution using Smart-seq3. *Nat. Biotechnol.* 38, 708–714.

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

Hahaut, V., Pavlinic, D., Carbone, W., Schuierer, S., Balmer, P., Quinodoz, M., et al. (2022). Fast and highly sensitive full-length single-cell RNA sequencing using FLASH-seq. *Nat. Biotechnol.* 40, 1447–1451.

Hanahan, D. (2022). Hallmarks of cancer: new dimensions. *Cancer Discov.* 12, 31–46. doi:10.1158/2159-8290.cd-21-1059

Hao, Y., Hao, S., Andersen-Nissen, E., Mauck, W. M., 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

He, L., Davila-Velderrain, J., Sumida, T. S., Hafler, D. A., Kellis, M., and Kulminski, A. M. (2021). Nebula is a fast negative binomial mixed model for differential or co-expression analysis of large-scale multi-subject single-cell data. *Commun. Biol.* 4, 629. doi:10.1038/s42003-021-02146-6

Heinrich, S., Craig, A. J., Ma, L., Heinrich, B., Greten, T. F., and Wang, X. W. (2021). Understanding tumour cell heterogeneity and its implication for immunotherapy in liver cancer using single-cell analysis. *J. Hepatology* 74, 700–715. doi:10.1016/j.jhep.2020.11.036

Heumos, L., Schaar, A. C., Lance, C., Litinetskaya, A., Drost, F., Zappia, L., et al. (2023). Best practices for single-cell analysis across modalities. *Nat. Rev. Genet.* 1, 550–572. doi:10.1038/s41576-023-00586-w

Ho, D. W.-H., Tsui, Y.-M., Chan, L.-K., Sze, K. M.-F., Zhang, X., Cheu, J. W.-S., et al. (2021). Single-cell rna sequencing shows the immunosuppressive landscape and tumor heterogeneity of hbv-associated hepatocellular carcinoma. *Nat. Commun.* 12, 3684. doi:10.1038/s41467-021-24010-1

Ho, D. W.-H., Tsui, Y.-M., Sze, K. M.-F., Chan, L.-K., Cheung, T.-T., Lee, E., et al. (2019). Single-cell transcriptomics reveals the landscape of intra-tumoral heterogeneity and stemness-related subpopulations in liver cancer. *Cancer Lett.* 459, 176–185. doi:10.1016/j.canlet.2019.06.002

Hong, F., Meng, Q., Zhang, W., Zheng, R., Li, X., Cheng, T., et al. (2021). Single-cell analysis of the pan-cancer immune microenvironment and sctime portal. *Cancer Immunol. Res.* 9, 939–951. doi:10.1158/2326-6066.cir-20-1026

Hou, W., Ji, Z., Ji, H., and Hicks, S. C. (2020). A systematic evaluation of single-cell rna-sequencing imputation methods. *Genome Biol.* 21, 218. doi:10.1186/s13059-020-02132-x

Hu, Y., An, Q., Sheu, K., Trejo, B., Fan, S., and Guo, Y. (2018). Single cell multi-omics technology: methodology and application. *Front. Cell Dev. Biol.* 6, 28. doi:10.3389/fcell.2018.00028

Huang, H., Wang, Y., Rudin, C., and Browne, E. P. (2022a). Towards a comprehensive evaluation of dimension reduction methods for transcriptomic data visualization. *Commun. Biol.* 5, 719. doi:10.1038/s42003-022-03628-x

Huang, Z., Sun, S., Lee, M., Maslov, A. Y., Shi, M., Waldman, S., et al. (2022b). Single-cell analysis of somatic mutations in human bronchial epithelial cells in relation to aging and smoking. *Nat. Genet.* 54, 492–498. doi:10.1038/s41588-022-01035-w

Huh, R., Yang, Y., Jiang, Y., Shen, Y., and Li, Y. (2019). Same-clustering: single-cell aggregated clustering via mixture model ensemble. *Nucleic Acids Res.* 48, 86–95. doi:10.1093/nar/gkz959

i Mompel, P. B., Santiago, J. V., Braunger, J., Geiss, C., Dimitrov, D., Müller-Dott, S., et al. (2022). decoupler: ensemble of computational methods to infer biological activities from omics data. *Bioinforma. Adv.* 2, vbac016. doi:10.1093/bioadv/vbac016

Janiszewska, M., Primi, M. C., and Izard, T. (2020). Cell adhesion in cancer: beyond the migration of single cells. *J. Biol. Chem.* 295, 2495–2505. doi:10.1074/jbc.REV119.007759

Jiang, A., Lehnert, K., Reid, S. J., Handley, R. R., Jacobsen, J. C., Rudiger, S. R., et al. (2023). *Isolated nuclei from frozen tissue are the superior source for single cell rna-seq compared with whole cells*, 2023–2102. bioRxiv.

Jiang, L., Chen, H., Pinello, L., and Yuan, G.-C. (2016). Giniclust: detecting rare cell types from single-cell gene expression data with gini index. *Genome Biol.* 17, 144–213. doi:10.1186/s13059-016-1010-4

Jin, Z., Huang, W., Shen, N., Li, J., Wang, X., Dong, J., et al. (2022). Single-cell gene fusion detection by scFusion. *Nat. Commun.* 13, 1084. doi:10.1038/s41467-022-28661-6

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

Jones, M. G., Khodaverdian, A., Quinn, J. J., Chan, M. M., Hussmann, J. A., Wang, R., et al. (2020). Inference of single-cell phylogenies from lineage tracing data using cassiopeia. *Genome Biol.* 21, 92. doi:10.1186/s13059-020-02000-8

Kamimoto, K., Stringa, B., Hoffmann, C. M., Jindal, K., Solnica-Krezel, L., and Morris, S. A. (2023). Dissecting cell identity via network inference and *in silico* gene perturbation. *Nature* 614, 742–751. doi:10.1038/s41586-022-05688-9

Kang, J. B., Nathan, A., Weinand, K., Zhang, F., Millard, N., Rumker, L., et al. (2021). Efficient and precise single-cell reference atlas mapping with symphony. *Nat. Commun.* 12, 5890. doi:10.1038/s41467-021-25957-x

Ke, M., Elshenawy, B., Sheldon, H., Arora, A., and Buffa, F. M. (2022). Single cell rna-sequencing: a powerful yet still challenging technology to study cellular heterogeneity. *BioEssays* 44, 2200084. doi:10.1002/bies.202200084

Kester, L., and Van Oudenaarden, A. (2018). Single-cell transcriptomics meets lineage tracing. *Cell stem Cell* 23, 166–179. doi:10.1016/j.stem.2018.04.014

Kim, J., and DeBerardinis, R. J. (2019). Mechanisms and implications of metabolic heterogeneity in cancer. *Cell Metab.* 30, 434–446. doi:10.1016/j.cmet.2019.08.013

Kinker, G. S., Greenwald, A. C., Tal, R., Orlova, Z., Cuoco, M. S., McFarland, J. M., et al. (2019). *Pan-cancer single cell rna-seq uncovers recurring programs of cellular heterogeneity*. bioRxiv. doi:10.1101/807552

Kivioja, T., Vähärautio, A., Karlsson, K., Bonke, M., Enge, M., Linnarsson, S., et al. (2012). Counting absolute numbers of molecules using unique molecular identifiers. *Nat. Methods* 9, 72–74. doi:10.1038/nmeth.1778

Klimovskaia, A., Lopez-Paz, D., Bottou, L., and Nickel, M. (2020). Poincaré maps for analyzing complex hierarchies in single-cell data. *Nat. Commun.* 11, 2966. doi:10.1038/s41467-020-16822-4

Kobak, D., and Berens, P. (2019). The art of using t-sne for single-cell transcriptomics. *Nat. Commun.* 10, 5416. doi:10.1038/s41467-019-13056-x

Kolodziejczyk, A. A., Kim, J. K., Svensson, V., Marioni, J. C., and Teichmann, S. A. (2015). The technology and biology of single-cell rna sequencing. *Mol. Cell* 58, 610–620. doi:10.1016/j.molcel.2015.04.005

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

Lachmann, A., Giorgi, F. M., Lopez, G., and Califano, A. (2016). ARACNe-AP: gene network reverse engineering through adaptive partitioning inference of mutual information. *Bioinformatics* 32, 2233–2235. doi:10.1093/bioinformatics/btw216

Lafzi, A., Moutinho, C., Picelli, S., and Heyn, H. (2018). Tutorial: guidelines for the experimental design of single-cell rna sequencing studies. *Nat. Protoc.* 13, 2742–2757. doi:10.1038/s41596-018-0073-y

Lange, M., Bergen, V., Klein, M., Setty, M., Reuter, B., Bakhti, M., et al. (2022). Cellrank for directed single-cell fate mapping. *Nat. Methods* 19, 159–170. doi:10.1038/s41592-021-01346-6

Lawson, D. A., Bhakta, N. R., Kessenbrock, K., Prummel, K. D., Yu, Y., Takai, K., et al. (2015). Single-cell analysis reveals a stem-cell program in human metastatic breast cancer cells. *Nature* 526, 131–135. doi:10.1038/nature15260

Lee, A. H., Koh, I. L., and Dawson, M. R. (2022). The role of exosome heterogeneity in epithelial ovarian cancer. *Adv. Cancer Biol. - Metastasis* 4, 100040. doi:10.1016/j.adcanc.2022.100040

Lee, T.-J., Wu, T., Kim, Y.-J., Park, J.-H., Lee, D. S., and Bhang, S. H. (2021). Alternative method for trypsin-based cell dissociation using poly (amino ester) coating and ph 6.0 pbs. *J. Bioact. Compatible Polym.* 36, 77–89. doi:10.1177/0883911520981710

Levitin, H. M., Yuan, J., and Sims, P. A. (2018). Single-cell transcriptomic analysis of tumor heterogeneity. *Trends cancer* 4, 264–268. doi:10.1016/j.trecan.2018.02.003

Li, C., Guan, R., Li, W., Wei, D., Cao, S., Xu, C., et al. (2023a). Single-cell rna sequencing reveals tumor immune microenvironment in human hypopharygeal squamous cell carcinoma and lymphatic metastasis. *Front. Immunol.* 14, 1168191. doi:10.3389/fimmu.2023.1168191

Li, W., Liu, J.-B., Hou, L.-K., Yu, F., Zhang, J., Wu, W., et al. (2022). Liquid biopsy in lung cancer: significance in diagnostics, prediction, and treatment monitoring. *Mol. cancer* 21, 25. doi:10.1186/s12943-022-01505-z

Li, X., Poire, A., Jeong, K. J., Zhang, D., Chen, G., Sun, C., et al. (2023b). Single-cell trajectory analysis reveals a cd9 positive state to contribute to exit from stem cell-like and embryonic diapause states and transit to drug-resistant states. *Cell Death Discov.* 9, 285. doi:10.1038/s41420-023-01586-9

Li, X., Wang, K., Lyu, Y., Pan, H., Zhang, J., Stambolian, D., et al. (2020). Deep learning enables accurate clustering with batch effect removal in single-cell rna-seq analysis. *Nat. Commun.* 11, 2338. doi:10.1038/s41467-020-15851-3

Li, Y., Ma, L., Wu, D., and Chen, G. (2021). Advances in bulk and single-cell multi-omics approaches for systems biology and precision medicine. *Briefings Bioinforma.* 22, bbab024. doi:10.1093/bib/bbab024

Li, Y., Xu, X., Song, L., Hou, Y., Li, Z., Tsang, S., et al. (2012). Single-cell sequencing analysis characterizes common and cell-lineage-specific mutations in a muscle-invasive bladder cancer. *Gigascience* 1, 12–14. doi:10.1186/2047-217X-1-12

Li, Z., Wang, Z., Tang, Y., Lu, X., Chen, J., Dong, Y., et al. (2019). Liquid biopsy-based single-cell metabolic phenotyping of lung cancer patients for informative diagnostics. *Nat. Commun.* 10, 3856. doi:10.1038/s41467-019-11808-3

Lim, S. B., Di Lee, W., Vasudevan, J., Lim, W.-T., and Lim, C. T. (2019). Liquid biopsy: one cell at a time. *NPJ Precis. Oncol.* 3, 23. doi:10.1038/s41698-019-0095-0

Lin, L., Song, M., Jiang, Y., Zhao, X., Wang, H., and Zhang, L. (2020). Normalizing single-cell rna sequencing data with internal spike-in-like genes. *NAR Genomics Bioinforma.* 2, lqaa059. doi:10.1093/nargab/lqaa059

Lin, Y., Ghazanfar, S., Wang, K. Y. X., Gagnon-Bartsch, J. A., Lo, K. K., Su, X., et al. (2019). Scmerge leverages factor analysis, stable expression, and pseudoreplication to merge multiple single-cell rna-seq datasets. *Proc. Natl. Acad. Sci.* 116, 9775–9784. doi:10.1073/pnas.1820006116

Liu, L., Zhang, Q., Wang, C., Guo, H., Mukwaya, V., Chen, R., et al. (2023a). Single-cell diagnosis of cancer drug resistance through the differential endocytosis of nanoparticles between drug-resistant and drug-sensitive cancer cells. *ACS Nano* 17, 19372–19386. doi:10.1021/acsnano.3c07030

Liu, W., Hu, H., Shao, Z., Lv, X., Zhang, Z., Deng, X., et al. (2023b). Characterizing the tumor microenvironment at the single-cell level reveals a novel immune evasion mechanism in osteosarcoma. *Bone Res.* 11, 4. doi:10.1038/s41413-022-00237-6

Lo, Y.-C., Liu, Y., Kammersgaard, M., Koladiya, A., Keyes, T. J., and Davis, K. L. (2023). Single-cell technologies uncover intra-tumor heterogeneity in childhood cancers. *Seminars Immunopathol.* 45, 61–69. doi:10.1007/s00281-022-00981-1

Lonardo, A. D., Nasi, S., and Pulciani, S. (2015). Cancer: we should not forget the past. *J. Cancer* 6, 29–39. doi:10.7150/jca.10336

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

Losic, B., Craig, A. J., Villacorta-Martin, C., Martins-Filho, S. N., Akers, N., Chen, X., et al. (2020). Intratumoral heterogeneity and clonal evolution in liver cancer. *Nat. Commun.* 11, 291. doi:10.1038/s41467-019-14050-z

Lotfollahi, M., Naghipourfar, M., Luecken, M. D., Khajavi, M., Büttner, M., Wagenstetter, M., et al. (2022). Mapping single-cell data to reference atlases by transfer learning. *Nat. Biotechnol.* 40, 121–130. doi:10.1038/s41587-021-01001-7

Lotfollahi, M., Wolf, F. A., and Theis, F. J. (2019). Scgen predicts single-cell perturbation responses. *Nat. Methods* 16, 715–721. doi:10.1038/s41592-019-0494-8

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. (2016). 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

Ma, A., Xin, G., and Ma, Q. (2022). The use of single-cell multi-omics in immuno-oncology. *Nat. Commun.* 13, 2728. doi:10.1038/s41467-022-30549-4

Manno, G. L., 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

Margolin, A. A., Nemenman, I., Basso, K., Klein, U., Wiggins, C., Stolovitzky, G., et al. (2006). ARACNE: an algorithm for the reconstruction of gene regulatory networks in a mammalian cellular context. *BMC Bioinforma.* 7, S7. doi:10.1186/1471-2105-7-s1-s7

Martelotto, L. G., Ng, C. K., Piscuoglio, S., Weigelt, B., and Reis-Filho, J. S. (2014). Breast cancer intra-tumor heterogeneity. *Breast Cancer Res.* 16, 210–211. doi:10.1186/bcr3658

Matsumoto, H., Kiryu, H., Furusawa, C., Ko, M. S., Ko, S. B., Gouda, N., et al. (2017). Scode: an efficient regulatory network inference algorithm from single-cell rna-seq during differentiation. *Bioinformatics* 33, 2314–2321. doi:10.1093/bioinformatics/btx194

Miles, L. A., Bowman, R. L., Merlinsky, T. R., Csete, I. S., Ooi, A. T., Durruthy-Durruthy, R., et al. (2020). Single-cell mutation analysis of clonal evolution in myeloid malignancies. *Nature* 587, 477–482. doi:10.1038/s41586-020-2864-x

Miller, A., Nagy, C., Knapp, B., Laengle, J., Ponweiser, E., Groeger, M., et al. (2017). Exploring metabolic configurations of single cells within complex tissue microenvironments. *Cell metab.* 26, 788–800. doi:10.1016/j.cmet.2017.08.014

Morita, K., Wang, F., Jahn, K., Hu, T., Tanaka, T., Sasaki, Y., et al. (2020). Clonal evolution of acute myeloid leukemia revealed by high-throughput single-cell genomics. *Nat. Commun.* 11, 5327. doi:10.1038/s41467-020-19119-8

Naitzat, G., Zhitnikov, A., and Lim, L.-H. (2020). *Topology of deep neural networks. arXiv*. doi:10.48550/arxiv.2004.06093

Nam, A. S., Chaligne, R., and Landau, D. A. (2021). Integrating genetic and non-genetic determinants of cancer evolution by single-cell multi-omics. *Nat. Rev. Genet.* 22, 3–18. doi:10.1038/s41576-020-0265-5

Nguyen, Q. H., Pervolarakis, N., Nee, K., and Kessenbrock, K. (2018). Experimental considerations for single-cell rna sequencing approaches. *Front. Cell Dev. Biol.* 6, 108. doi:10.3389/fcell.2018.00108

Nieto, P., Elosua-Bayes, M., Trincado, J. L., Marchese, D., Massoni-Badosa, R., Salvany, M., et al. (2021). A single-cell tumor immune atlas for precision oncology. *Genome Res.* 31, 1913–1926. doi:10.1101/gr.273300.120

Niño, J. L. G., Wu, H., LaCourse, K. D., Kempchinsky, A. G., Baryiames, A., Barber, B., et al. (2022). Effect of the intratumoral microbiota on spatial and cellular heterogeneity in cancer. *Nature* 611, 810–817. doi:10.1038/s41586-022-05435-0

Ogbeide, S., Giannese, F., Mincarelli, L., and Macaulay, I. C. (2022). Into the multiverse: advances in single-cell multiomic profiling. *Trends Genet.* 38, 831–843. doi:10.1016/j.tig.2022.03.015

Orrapin, S., Thongkumkoon, P., Udomruk, S., Moonmuang, S., Sutthitthasakul, S., Yongpitakwattana, P., et al. (2023). Deciphering the biology of circulating tumor cells through single-cell rna sequencing: implications for precision medicine in cancer. *Int. J. Mol. Sci.* 24, 12337. doi:10.3390/ijms241512337

Ortega, M. A., Poirion, O., Zhu, X., Huang, S., Wolfgruber, T. K., Sebra, R., et al. (2017). Using single-cell multiple omics approaches to resolve tumor heterogeneity. *Clin. Transl. Med.* 6, 46–16. doi:10.1186/s40169-017-0177-y

Pei, H., Li, L., Han, Z., Wang, Y., and Tang, B. (2020). Recent advances in microfluidic technologies for circulating tumor cells: enrichment, single-cell analysis, and liquid biopsy for clinical applications. *Lab a Chip* 20, 3854–3875. doi:10.1039/d0lc00577k

Peng, A., Mao, X., Zhong, J., Fan, S., and Hu, Y. (2020). Single-cell multi-omics and its prospective application in cancer biology. *Proteomics* 20, 1900271. doi:10.1002/pmic.201900271

Phan, H. V., van Gent, M., Drayman, N., Basu, A., Gack, M. U., and Tay, S. (2021). High-throughput rna sequencing of paraformaldehyde-fixed single cells. *Nat. Commun.* 12, 5636. doi:10.1038/s41467-021-25871-2

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.

Pierce, S. E., Granja, J. M., and Greenleaf, W. J. (2021). High-throughput single-cell chromatin accessibility crispr screens enable unbiased identification of regulatory networks in cancer. *Nat. Commun.* 12, 2969. doi:10.1038/s41467-021-23213-w

Pratapa, A., Jalihal, A. P., Law, J. N., Bharadwaj, A., and Murali, T. M. (2020). Benchmarking algorithms for gene regulatory network inference from single-cell transcriptomic data. *Nat. Methods* 17, 147–154. doi:10.1038/s41592-019-0690-6

Preissl, S., Gaulton, K. J., and Ren, B. (2023). Characterizing cis-regulatory elements using single-cell epigenomics. *Nat. Rev. Genet.* 24, 21–43. doi:10.1038/s41576-022-00509-1

Prieto-Vila, M., Usuba, W., Takahashi, R.-u., Shimomura, I., Sasaki, H., Ochiya, T., et al. (2019). Single-cell analysis reveals a preexisting drug-resistant subpopulation in the luminal breast cancer subtype. *Cancer Res.* 79, 4412–4425. doi:10.1158/0008-5472.CAN-19-0122

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. doi:10.1016/j.cell.2017.10.044

Rautenstrauch, P., Vlot, A. H. C., Saran, S., and Ohler, U. (2022). Intricacies of single-cell multi-omics data integration. *Trends Genet.* 38, 128–139. doi:10.1016/j.tig.2021.08.012

Ren, X., Zhang, L., Zhang, Y., Li, Z., Siemers, N., and Zhang, Z. (2021). Insights gained from single-cell analysis of immune cells in the tumor microenvironment. *Annu. Rev. Immunol.* 39, 583–609. doi:10.1146/annurev-immunol-110519-071134

Risso, D., Perraudeau, F., Gribkova, S., Dudoit, S., and Vert, J.-P. (2018). A general and flexible method for signal extraction from single-cell rna-seq data. *Nat. Commun.* 9, 284. doi:10.1038/s41467-017-02554-5

Ritchie, M. E., Phipson, B., Wu, D., Hu, Y., Law, C. W., Shi, W., et al. (2015). Limma powers differential expression analyses for rna-sequencing and microarray studies. *Nucleic Acids Res.* 43, e47. doi:10.1093/nar/gkv007

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

Roerink, S. F., Sasaki, N., Lee-Six, H., Young, M. D., Alexandrov, L. B., Behjati, S., et al. (2018). Intra-tumour diversification in colorectal cancer at the single-cell level. *Nature* 556, 457–462. doi:10.1038/s41586-018-0024-3

Ryu, Y., Han, G. H., Jung, E., and Hwang, D. (2023). Integration of single-cell rna-seq datasets: a review of computational methods. *Mol. Cells* 46, 106–119. doi:10.14348/molcells.2023.0009

Shah, S., Lubeck, E., Zhou, W., and Cai, L. (2017). Seqfish accurately detects transcripts in single cells and reveals robust spatial organization in the hippocampus. *Neuron* 94, 752–758. doi:10.1016/j.neuron.2017.05.008

Sikkema, L., Strobl, D., Zappia, L., Madissoon, E., Markov, N., Zaragosi, L., et al. (2022). *An integrated cell atlas of the human lung in health and disease*. bioRxiv. doi:10.1101/2022.03.10.483747

Silverman, J. D., Roche, K., Mukherjee, S., and David, L. A. (2020). *Naught all zeros in sequence count data are the same*. bioRxiv. doi:10.1101/477794

Simone, M. D., Rossetti, G., and Pagani, M. (2019). Single cell methods, sequencing and proteomics. *Methods Mol. Biology* 1979, 87–110.

Skinnider, M. A., Squair, J. W., and Foster, L. J. (2019). Evaluating measures of association for single-cell transcriptomics. *Nat. methods* 16, 381–386. doi:10.1038/s41592-019-0372-4

Slyper, M., Porter, C. B., Ashenberg, O., Waldman, J., Drokhlyansky, E., Wakiro, I., et al. (2020). A single-cell and single-nucleus rna-seq toolbox for fresh and frozen human tumors. *Nat. Med.* 26, 792–802. doi:10.1038/s41591-020-0844-1

Squair, J. W., Gautier, M., Kathe, C., Anderson, M. A., James, N. D., Hutson, T. H., et al. (2021). Confronting false discoveries in single-cell differential expression. *Nat. Commun.* 12, 5692. doi:10.1038/s41467-021-25960-2

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

Su, E. Y., Spangler, A., Bian, Q., Kasamoto, J. Y., and Cahan, P. (2022). Reconstruction of dynamic regulatory networks reveals signaling-induced topology changes associated with germ layer specification. *Stem Cell Rep.* 17, 427–442. doi:10.1016/j.stemcr.2021.12.018

Su, K., Wu, Z., and Wu, H. (2020). Simulation, power evaluation and sample size recommendation for single-cell rna-seq. *Bioinformatics* 36, 4860–4868. doi:10.1093/bioinformatics/btaa607

Sun, Y.-F., Wu, L., Liu, S.-P., Jiang, M.-M., Hu, B., Zhou, K.-Q., et al. (2021). Dissecting spatial heterogeneity and the immune-evasion mechanism of ctcs by single-cell rna-seq in hepatocellular carcinoma. *Nat. Commun.* 12, 4091. doi:10.1038/s41467-021-24386-0

Suvà, M. L., and Tirosh, I. (2019). Single-cell rna sequencing in cancer: lessons learned and emerging challenges. *Mol. Cell* 75, 7–12. doi:10.1016/j.molcel.2019.05.003

Teuwen, L.-A., Rooij, L. P. D., Cuypers, A., Rohlenova, K., Dumas, S. J., García-Caballero, M., et al. (2021). Tumor vessel co-option probed by single-cell analysis. *Cell Rep.* 35, 109253. doi:10.1016/j.celrep.2021.109253

Tian, L., Chen, F., and Macosko, E. Z. (2023a). The expanding vistas of spatial transcriptomics. *Nat. Biotechnol.* 41, 773–782. doi:10.1038/s41587-022-01448-2

Tian, T., Zhong, C., Lin, X., Wei, Z., and Hakonarson, H. (2023b). Complex hierarchical structures in single-cell genomics data unveiled by deep hyperbolic manifold learning. *Genome Res.* 33, 232–246. doi:10.1101/gr.277068.122

Tirier, S. M., Mallm, J.-P., Steiger, S., Poos, A. M., Awwad, M. H., Giesen, N., et al. (2021). Subclone-specific microenvironmental impact and drug response in refractory multiple myeloma revealed by single-cell transcriptomics. *Nat. Commun.* 12, 6960. doi:10.1038/s41467-021-26951-z

Tirosh, I., Venteicher, A. S., Hebert, C., Escalante, L. E., Patel, A. P., Yizhak, K., et al. (2016). Single-cell rna-seq supports a developmental hierarchy in human oligodendroglioma. *Nature* 539, 309–313. doi:10.1038/nature20123

Tran, B., Tran, D., Nguyen, H., Ro, S., and Nguyen, T. (2022). sccan: single-cell clustering using autoencoder and network fusion. *Sci. Rep.* 12, 10267. doi:10.1038/s41598-022-14218-6

Türei, D., Valdeolivas, A., Gul, L., Palacio-Escat, N., Klein, M., Ivanova, O., et al. (2021). Integrated intra- and intercellular signaling knowledge for multicellular omics analysis. *Mol. Syst. Biol.* 17, e9923. doi:10.15252/msb.20209923

Turki, T., and Taguchi, Y. H. (2020). Scgrns: novel supervised inference of single-cell gene regulatory networks of complex diseases. *Comput. Biol. Med.* 118, 103656. doi:10.1016/j.compbiomed.2020.103656

Van der Leun, A. M., Thommen, D. S., and Schumacher, T. N. (2020). Cd8+ t cell states in human cancer: insights from single-cell analysis. *Nat. Rev. Cancer* 20, 218–232. doi:10.1038/s41568-019-0235-4

Wang, F., Liang, S., Kumar, T., Navin, N., and Chen, K. (2019). Scmarker: *ab initio* marker selection for single cell transcriptome profiling. *PLoS Comput. Biol.* 15, e1007445. doi:10.1371/journal.pcbi.1007445

Wang, Q., Wang, Z., Zhang, Z., Zhang, W., Zhang, M., Shen, Z., et al. (2021). Landscape of cell heterogeneity and evolutionary trajectory in ulcerative colitis-associated colon cancer revealed by single-cell rna sequencing. *Chin. J. Cancer Res.* 33, 271–288. doi:10.21147/j.issn.1000-9604.2021.02.13

Wang, T., Shi, J., Li, L., Zhou, X., Zhang, H., Zhang, X., et al. (2022a). Single-cell transcriptome analysis reveals inter-tumor heterogeneity in bilateral papillary thyroid carcinoma. *Front. Immunol.* 13, 840811. doi:10.3389/fimmu.2022.840811

Wang, X., Xu, Y., Sun, Q., Zhou, X., Ma, W., Wu, J., et al. (2022b). New insights from the single-cell level: tumor associated macrophages heterogeneity and personalized therapy. *Biomed. Pharmacother.* 153, 113343. doi:10.1016/j.biopha.2022.113343

Wang, Y., and Zhao, H. (2022). Non-linear archetypal analysis of single-cell rna-seq data by deep autoencoders. *PLoS Comput. Biol.* 18, e1010025. doi:10.1371/journal.pcbi.1010025

Wei, N., Nie, Y., Liu, L., Zheng, X., and Wu, H.-J. (2022). Secuer: ultrafast, scalable and accurate clustering of single-cell rna-seq data. *PLOS Comput. Biol.* 18, e1010753. doi:10.1371/journal.pcbi.1010753

Welch, J. D., Kozareva, V., Ferreira, A., Vanderburg, C., Martin, C., and Macosko, E. Z. (2019). Single-cell multi-omic integration compares and contrasts features of brain cell identity. *Cell* 177, 1873–1887. doi:10.1016/j.cell.2019.05.006

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

Woo, J., Winterhoff, B. J., Starr, T. K., Aliferis, C., and Wang, J. (2019). *De novo* prediction of cell-type complexity in single-cell rna-seq and tumor microenvironments. *Life Sci. Alliance* 2, e201900443. doi:10.26508/lsa.201900443

Wouters, J., Kalender-Atak, Z., Minnoye, L., Spanier, K. I., Waegeneer, M. D., González-Blas, C. B., et al. (2020). Robust gene expression programs underlie recurrent cell states and phenotype switching in melanoma. *Nat. Cell Biol.* 22, 986–998. doi:10.1038/s41556-020-0547-3

Wu, F., Fan, J., He, Y., Xiong, A., Yu, J., Li, Y., et al. (2021a). Single-cell profiling of tumor heterogeneity and the microenvironment in advanced non-small cell lung cancer. *Nat. Commun.* 12, 2540. doi:10.1038/s41467-021-22801-0

Wu, S. Z., Al-Eryani, G., Roden, D. L., Junankar, S., Harvey, K., Andersson, A., et al. (2021b). A single-cell and spatially resolved atlas of human breast cancers. *Nat. Genet.* 53, 1334–1347. doi:10.1038/s41588-021-00911-1

Xiang, R., Wang, W., Yang, L., Wang, S., Xu, C., and Chen, X. (2021). A comparison for dimensionality reduction methods of single-cell rna-seq data. *Front. Genet.* 12, 646936. doi:10.3389/fgene.2021.646936

Xie, K., Huang, Y., Zeng, F., Liu, Z., and Chen, T. (2020). scaide: clustering of large-scale single-cell rna-seq data reveals putative and rare cell types. *NAR genomics Bioinforma.* 2, lqaa082. doi:10.1093/nargab/lqaa082

Xu, C., Lopez, R., Mehlman, E., Regier, J., Jordan, M. I., and Yosef, N. (2021). Probabilistic harmonization and annotation of single-cell transcriptomics data with deep generative models. *Mol. Syst. Biol.* 17, e9620. doi:10.15252/msb.20209620

Xu, C., Prete, M., Webb, S., Jardine, L., Stewart, B., Hoo, R., et al. (2023). *Automatic cell type harmonization and integration across human cell atlas datasets*. BiorXiv. doi:10.1101/2023.05.01.538994

Yancovitz, M., Litterman, A., Yoon, J., Ng, E., Shapiro, R. L., Berman, R. S., et al. (2012). Intra-and inter-tumor heterogeneity of brafv600e mutations in primary and metastatic melanoma. *PloS one* 7, e29336. doi:10.1371/journal.pone.0029336

Yang, Y., Huh, R., Culpepper, H. W., Lin, Y., Love, M. I., and Li, Y. (2018). Safe-clustering: single-cell aggregated (from ensemble) clustering for single-cell rna-seq data. *Bioinformatics* 35, 1269–1277. doi:10.1093/bioinformatics/bty793

Young, M. D., Mitchell, T. J., Braga, F. A. V., 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

Yuan, H., Yan, M., Zhang, G., Liu, W., Deng, C., Liao, G., et al. (2019). Cancersea: a cancer single-cell state atlas. *Nucleic acids Res.* 47, D900-D908–D908. doi:10.1093/nar/gky939

Yuan, S., Stewart, K. S., Yang, Y., Abdusselamoglu, M. D., Parigi, S. M., Feinberg, T. Y., et al. (2022). Ras drives malignancy through stem cell crosstalk with the microenvironment. *Nature* 612, 555–563. doi:10.1038/s41586-022-05475-6

Zhang, Q., He, Y., Luo, N., Patel, S. J., Han, Y., Gao, R., et al. (2019). Landscape and dynamics of single immune cells in hepatocellular carcinoma. *Cell* 179, 829–845. doi:10.1016/j.cell.2019.10.003

Zhang, Y., Parmigiani, G., and Johnson, W. E. (2020). Combat-seq: batch effect adjustment for rna-seq count data. *NAR Genomics Bioinforma.* 2, lqaa078. doi:10.1093/nargab/lqaa078

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.

Zhou, H., Zhu, L., Song, J., Wang, G., Li, P., Li, W., et al. (2022). Liquid biopsy at the frontier of detection, prognosis and progression monitoring in colorectal cancer. *Mol. Cancer* 21, 86. doi:10.1186/s12943-022-01556-2

Zhu, X., Li, J., Li, H.-D., Xie, M., and Wang, J. (2020). Sc-gpe: a graph partitioning-based cluster ensemble method for single-cell. *Front. Genet.* 11, 604790. doi:10.3389/fgene.2020.604790

Zilionis, R., Nainys, J., Veres, A., Savova, V., Zemmour, D., Klein, A. M., et al. (2017). Single-cell barcoding and sequencing using droplet microfluidics. *Nat. Protoc.* 12, 44–73.

Zimmerman, K. D., Espeland, M. A., and Langefeld, C. D. (2021). A practical solution to pseudoreplication bias in single-cell studies. *Nat. Commun.* 12, 738. doi:10.1038/s41467-021-21038-1