Neuro-Oncology Advances
8(S1), vdaf047, 2026 | https://doi.org/10.1093/noajnl/vdaf047

Pitfalls in analysis and interpretation of single-cell RNA-seq data in cancer

Itay Tirosh

All author affiliations are listed at the end of the article.

Corresponding Author: Itay Tirosh, PhD, Department of Molecular Cell Biology, Weizmann Institute of Science, Rehovot, Israel (itay.tirosh@weizmann.ac.il).

Abstract

Single-cell and single-nuclei RNA-seq (sc/snRNA-seq) have become a central approach in cancer research, and their widespread impact has been enabled by various computational tools developed specifically for sc/snRNA-seq analysis. Nevertheless, proper analysis and interpretation of sc/snRNA-seq data requires significant expertise, and the inadequate utility of certain computational methods may lead to dubious results. To mitigate these issues, it is important to recognize the limitations of sc/snRNA-seq data, the assumptions of common methods, and to perform robust analysis. Here, I describe common pitfalls in cancer sc/snRNA-seq analysis and discuss ways to overcome them. Among others, this includes a discussion of potential errors in statistical analysis, in inference of chromosomal aberrations, in trajectory analysis, and in signature-based analysis of bulk RNA-seq data. This review may help readers to avoid common pitfalls and to perform informative analysis and careful interpretation of sc/snRNA-seq datasets in cancer.

Key Points

- Single cell omics methods are transforming cancer research.
- However, single cell data analysis is complex, requiring expertise and careful interpretations.
- Common pitfalls in single cell data analysis are identified and described in detail

Cancer, like most other diseases, occurs in the context of a tissue that is composed of diverse cells that interact in complex ways. While a great deal of research considers the entire tumor as a single entity and measures tumor-wide features, deep understanding of cancer biology requires methods that can uncover this cellular diversity. Accordingly, single-cell and single-nuclei RNA-seq (sc/snRNA-seq), as well as other single-cell methods, have risen in the last decade from a specialized technology of few expert labs to a widespread methodology in cancer research.[1]

The adoption of sc/snRNA-seq by hundreds of cancer research labs has advanced a revolution in cancer research, including the identification of diverse cancer cell states, extensive characterization of the tumor microenvironment, cell–cell interactions, and other discoveries. However, the extensive use of sc/snRNA-seq also introduces important challenges, especially in the context of computational analysis and their interpretation. Sc/snRNA-seq of even a single tumor generates an extensive amount of data—thousands of cells, each represented by thousands of transcripts. This amount of data requires computational approaches to process, normalize, visualize the data, and to perform various downstream analyses to address specific questions.[2] Moreover, sc/snRNA-seq is associated with high levels of noise, both systematic and stochastic noise, which highly complicates the analysis and the interpretation of sc/snRNA-seq data.

In the past, complex types of analysis were mostly performed by dedicated computational labs, often using lab-specific approaches. However, with the rapid expansion of sc/snRNA-seq, a great need emerged for publicly available tools that would enable noncomputational researchers to carry out sc/snRNA-seq analysis. This need has led to the development

of over a thousand computational methods specifically designed for sc/snRNA-seq data.[3] Accordingly, researchers can readily apply a diverse set of tools to their sc/snRNA-seq data and make various discoveries.

Some of these computational methods have become widely used and thereby have had a huge contribution to the field. Seurat,[4] Scanpy,[5] and similar packages provide a platform for complete analysis of sc/snRNA-seq data that integrates methods to tackle various tasks. Monocle,[6] Velocyto,[7] and similar methods implement trajectory analyses to infer relationships between different cell types and cell states that are identified in sc/snRNA-seq data. InferCNV,[8,9] CopyKat,[10] and other methods enable the inference of chromosomal aberrations from sc/snRNA-seq data. Harmony,[11] scANVI,[12] and other methods enable the integration of multiple batches or related data types. Finally, Cibersort,[13] MuSiC,[14] and other methods perform deconvolution of bulk samples based on sc/snRNA-seq datasets.

Collectively, these methods have had a tremendous positive impact. Nevertheless, the existence of such off-the-shelf tools also increases the risk that users will apply them without a complete understanding of the methods, their underlying assumptions, the parameters that could be tuned to specific datasets, the reliability and robustness of the results, and the potential pitfalls associated with these methods. Such expertise is difficult to obtain, and while certain aspects of "best practices" are included in the description of these software packages, additional and more nuanced information remains hidden from most users and is often known only to the most experienced users.

Here, I describe 7 common pitfalls in the analysis of sc/snRNA-seq data while highlighting implications and ways to overcome or mitigate them (Table 1). These pitfalls reflect inherent limitations of sc/snRNA-seq datasets (pitfall #1), incorrect use of statistical tests (pitfall #2), and oversimplified use of sc/snRNA-seq analysis tools (pitfalls #3–7).

Pitfall #1: Overlooking the Intrinsic Limitations of sc/snRNA-seq Data

While it is tempting to consider sc/snRNA-seq data as largely accurate and comparable to bulk RNA-seq data, it is important to acknowledge the limitations of sc/snRNA-seq data and to understand the implications for downstream analysis and interpretations.[15] Below I summarize 4 main limitations of sc/snRNA-seq and their implications.

Incomplete and Biased Capture of Cells (pitfall 1.1)

scRNA-seq analysis of tumors typically begins with sample acquisition, dissociation, and then cell capture by a microfluidic device. Each of these steps is associated with the loss of a considerable fraction of cells. Many cells do not survive through the processes of sample acquisition (eg, surgery) and dissociation, and additional cells are either not captured in the microfluidic device or alternatively are captured together with other cells, hence generating a doublet that should be excluded in downstream analysis. The exact fraction of cells that are successfully profiled varies extensively between experiments and by itself may not pose a problem, since typical scRNA-seq studies aim to profile only thousands of cells from samples that contain millions of cells.

However, whether this "sampling" of cells is largely random or is biased toward particular cell types and cell states is highly important, as it relates to the common assumption that the observed distribution of cells in scRNA-seq data is proportional to the real distribution in the tumor tissue. Microfluidic cell capture is thought to be largely random, but the survival of cells throughout the experiment, and in particular during dissociation, often biases the observed cells against cell types or states with a short half-life (eg, neutrophils), those adhering to other cells (eg, brain endothelial cells) and those that do not survive dissociation due to complex morphologies (eg, neurons). For example, in scRNA-seq of glioma tissues, we typically detect very few (if any) neurons, endothelial cells, and even astrocytes.[9,16–18]

The observed distribution of cell types and states is expected to be more accurate in the context of snRNA-seq,[19] where dissociation is replaced by nuclei isolation, hence eliminating some sources of biased cell retention. Nevertheless, even with snRNA-seq, differences are observed between protocols[19] and hence a retention bias may still be expected. Another solution may be provided by emerging spatial transcriptomic and proteomic approaches that enable almost a complete detection of all cells in the examined tissue slide and hence less bias.[20,21] However, this advantage of spatial methods may be reduced by the common usage of small tissue samples for spatial analysis. In summary, biased cell retention depends on the exact protocol, but it is difficult to conclude that such bias does not exist (or that it is small) without direct evidence, which is challenging to obtain. For example, comparison of cell type frequencies between different methods that are unlikely to share the same bias (scRNA-seq, snRNA-seq, spatial methods, and even deconvolution of bulk RNA-seq) may help to assess such biases.

Different biases in cell retention constitute one form of batch effect that complicates the direct comparison of sc/snRNA-seq datasets from distinct platforms, labs, and researchers. Within a single cohort—profiled by the same researchers using a single protocol—it would seem reasonable to assume that biases in cell retention are largely constant and therefore that the observed distribution of cell types and states may be compared across samples (eg, patients), as done by most studies. Yet this distribution should typically not be assumed to reflect an accurate depiction of the original tissue.

Incomplete Capture of mRNAs (pitfall 1.2)

All sc/snRNA-seq protocols have limited sensitivity, such that only a fraction of the mRNAs in a cell are being sequenced. Sensitivity is often limited by insufficient sequencing depth. However, even when samples are sequenced at a high depth (ie, approaching saturation, such

Neuro-Oncology Advances

Table 1. Summary of Pitfalls and Potential Solutions or Mitigation Strategies

Pitfall in cancer scRNA-seq analysis	Potential solutions or mitigation strategies
1.1: Incomplete and biased capture of cells	- Avoid assumptions of complete/representative coverage - Compare cell frequencies within a single study/platform where biases are likely comparable - Prefer snRNA-seq and spatial transcriptomics over scRNA-seq, to avoid dissociation-based biases
1.2: Incomplete capture of mRNAs	- Focus on genes with medium/high average levels - Avoid assumptions that undetected genes are not expressed
1.3: Variability in number of transcripts per cell	- Test if clusters are driven by high/low number of transcripts per cell rather than by specific biology
1.4: Cells with spurious expression data	- Exclude cells with (i) limited complexity, (ii) high expression of mitochondrially encoded genes, (iii) very low expression of housekeeping genes (eg, encoding ribosomal proteins), or (iv) unexpected coexpression of signatures of 2 distinct cell types
2.1: Excessive statistical power in single-cell analysis	- Add a threshold for effect size (ie, fold-change) beyond the *P*-value thresholds.
2.2: Violating independence assumption in comparison across tumors	- In statistical analysis of multiple tumors, each tumor should be considered as one sample rather than as many independent samples. Yet, separate analysis can be performed for the frequency and expression profile of each cell type/state
3.1: Incorrect selection of reference cells for CNA inference	- Use multiple reference cell types and focus on consistent results - Identify and exclude spurious CNA inferences (eg, chr. 6p loss when using macrophage reference)
3.2: Interpretation of inferred CNA signal	- Focus on the most reliable apparent CNAs, based on (i) strong signals, (ii) expected patterns (eg, chr7/10 in GBM), (iii) covering chr. arms or full chr., and (iv) consistently observed in many cells (per tumor)
4.1: Determine number of clusters	- Avoid the assumption that there is a single optimal number of clusters, examine multiple methods and focus on robust patterns - First define only highly discrete clusters (likely reflecting cell types) and then analyze the (possibly continuous) diversity within each of them
4.2: Spurious clusters	- Carefully interpret (and potentially exclude) the clusters associated with (i) particularly high/low complexity, (ii) high expression of mitochondrially encoded genes, (iii) low expression of housekeeping genes, and (iv) unexpected coexpression of signatures for distinct cell types
4.3: Inference from UMAP plots	- Minimize the use of UMAP and rely more on heatmaps or other visualizations, including custom visualizations for specific questions
5: Spurious inference of cellular trajectories	- Avoid assumptions of unidirectional transitions without supporting evidence - Restrict trajectory analysis to cells of a single tumor and cell type, where direct transitions are more likely
6: Removing biological signal through batch correction	- Avoid the assumption that malignant cells from different patients should cluster together. - Also perform analysis without integration methods, especially by analyzing one tumor at a time, or a subset of tumors where batch effects are not apparent
7: Oversimplified inference from bulk datasets	- Focus the inference from bulk datasets on populations of cells that are frequent enough and that have gene sets with high specificity that can reliably be used for bulk deconvolution

that further sequencing will primarily involve resequencing of the same transcripts), the sensitivity remains limited, reflecting the inherent limitation of the protocol. Sensitivity differs between protocols, and in particular is lower for snRNA-seq than for scRNA-seq protocols, due to the loss of cytoplasmic mRNAs in the former. While sensitivity is typically difficult to quantify, it has been estimated as approximately 10% for standard protocols.[22] Coupled with the fact that, even for important genes, there are often few mRNA molecules in single cells at any given moment, this leads to very noisy measurement of individual genes by sc/snRNA-seq.

Notably, the impact of low sensitivity on sc/snRNA-seq analysis is particularly large for low-to-medium expressed genes. For example, *POU3F2* encodes a neural transcription factor that influences glioblastoma propagation potential.[23] Yet, due to its relatively low expression level, *POU3F2* is often undetected in sc/snRNA-seq data of glioblastoma cells, and even when it is detected, the observed variability across cells is less reliable than that of more highly expressed genes. For genes such as *POU3F2*, random sampling of ~10% of the mRNAs per cell could cause variation between one or even zero detected transcript in some cells and 3 or more detected transcripts in other cells, thus generating large apparent differences in gene expression (eg, when quantified by fold-change), purely due to a nonbiological effect. In contrast, for more highly expressed genes, noise due to sampling would be much smaller, for example, with some cells having 8–9 transcripts and others having 12–13 transcripts. It is therefore recommended to assess cellular states based on signatures of multiple genes rather than by individual

markers, hence averaging out the sampling noise of individual genes, and potentially to give more weight to relatively highly expressed genes in which sampling-based errors are smaller.

Variability in Number of Transcripts Per Cell (pitfall 1.3)

The previous section highlighted the limited overall sensitivity of sc/snRNA-seq data, but another problem is that such sensitivity is not the same for all cells. An overall sensitivity of 10% (ie, on average, 10% of the transcripts in a cell are sequenced), could still mean that some cells have only 2% sensitivity while others have 20% sensitivity. Such variable sensitivity within a dataset could lead to extensive apparent variability, in which many genes will be consistently detected in high-sensitivity cells, but not in low-sensitivity cells. As a result, sensitivity could become a dominant source of expression variability.

In typical sc/snRNA-seq datasets we cannot directly measure sensitivity, but we can evaluate a proxy for it—the number of detected transcripts (ie, UMIs) or genes per cell, which I refer to as the *complexity* of the cell library. Complexity is linked to the real number of molecules per cell and hence tends to be lower in smaller cell types, like lymphocytes. Conversely, complexity may be higher for proliferating cells, reflecting their increased mRNA production. Yet, even within a single nonproliferating cell type, complexity varies dramatically, suggesting that it may predominantly reflect the cell-specific sensitivity of sc/snRNA-seq. Notably, while standard quality control (QC) procedures remove cells with particularly high or particularly low complexity,[2] the allowed range of complexity is typically 10-fold or more, indicating that this is likely to remain a major source of variation after even stringent QC.

As noted above, complexity is driven both by technical factors (sc/snRNA-seq sensitivity) and by biological factors (amount of mRNA per cell), and in specific cases it is difficult to distinguish between them. Thus, the difference between clusters with low versus high complexity should be interpreted with caution, considering the possibility that technical factors likely contribute and possibly even dominate such clusters. It is therefore recommended to plot the complexity of cells and examine whether this variable may account for any of the results from sc/snRNA-seq analysis, including clustering as well as downstream results.

Cells With Spurious Expression Data (pitfall 1.4)

The previous section highlights the need to exclude cells with very low complexity. While this is done by all sc/snRNA-seq studies, the appropriate threshold for inclusion of cells is difficult to determine. There is an inherent trade-off between, on one hand, the inclination to retain more cells in order to increase the coverage of all populations, and on the other hand, the importance of excluding lower-quality cells in which noise and technical limitations have a larger impact on the data, possibly distorting downstream analysis. A typical choice for complexity threshold is 200 genes per cell, but at least in some settings a higher threshold may be beneficial, such as when focusing on cell types with high average complexity (eg, when analyzing malignant cells with an average complexity >3000 genes, we prefer to raise the threshold to 1000 genes or even higher than that[17,24]). The optimal choice for a threshold depends on the cell types being examined, as well as the exact platform, the sample quality, and the goals of downstream analyses. If the goals are well defined and can be quantified, then it may be possible to perform the analysis with multiple thresholds and identify the most efficient one, although in most cases this is not possible, and thresholds are determined in an ad hoc manner, while considering the observed distribution of complexity.

In addition to cells with very low complexity, other populations of cells may also reflect low quality or spurious data and should accordingly be removed. First, cells with high levels of mitochondrially encoded genes may reflect dying cells with low quality of their measured expression profiles. Second, in several cases we encountered suspicious expression profiles of subsets of cells in which the expression of housekeeping genes (eg, those encoding ribosomal proteins) is particularly low, while many other genes are detected as expressed although they are usually not expressed in the corresponding cell type.[9,25] This pattern of low housekeeping expression may indicate that in addition to mRNAs, there is ample sequencing of DNA fragments that are potentially primed by genomic polyA stretches, therefore decreasing the proportion of reads covering highly expressed genes (housekeeping genes) while increasing the proportion of reads covering nonexpressed genes.[26]

Finally, doublet cells may be incorrectly interpreted as reflecting unique cell states. While many studies use doublet removal methods, in our experience these methods often fail to capture some of the apparent doublets. Therefore, any cells highly expressing marker genes and signature gene sets for 2 distinct cell types (that are not expected to be coexpressed by known cell types) should be considered as potential doublets and interpreted with caution.

Pitfall #2: Power and Dependencies in Statistical Tests

Computational analyses often rely on calculated *P*-values to identify statistically significant patterns. In sc/snRNA-seq studies, when hundreds or thousands of cells are analyzed per tumor, 2 common issues may arise—one reflecting the outcome of high statistical power (pitfall 2.1) and the other reflecting an incorrect assumption of independence (pitfall 2.2).[27]

First, if we compare the levels of a certain gene between 2 tumors, then the large number of cells per tumor could mean that statistical power is extremely high. Hence, even minor expression differences could be deemed statistically significant, although they are unlikely to be biologically meaningful.[27] Moreover, across comparisons of multiple tumor pairs, the number of genes deemed as statistically significant could be driven by the statistical power (numbers of profiled cells) more than by the real biological differences. To avoid such issues, it is important to use not

only a *P*-value threshold, but also a fold-change threshold, or some other threshold that reflects *effect size* rather than statistical significance. For example, a 2-fold expression difference is a standard threshold in bulk RNA-seq analysis that can also be applied to sc/snRNA-seq (when comparing the averages of 2 populations of cells). When examining more subtle differences in cell states, it may be justified to decrease the threshold to 1.5-fold, but lower differences than that may be considered as spurious even if *P*-values appear highly significant.

Second, if we would like to compare 2 groups of tumors, for example, responders versus nonresponders for a certain treatment, then for standard statistical analysis (eg, *t* test) it is not valid to consider cells as the element that is being compared between the 2 groups.[27] Instead, the comparison should be done only across samples (one value per sample, reflecting pseudobulk levels, that is, the aggregate mRNA counts of all relevant cells in that sample) rather than across cells. In other words, the sample size in such analysis should be the same whether the samples were profiled by bulk RNA-seq or by sc/snRNA-seq. The reason for this is that single cells from the same tumor are not independent of one another—they all share certain tumor features—and statistical tests typically assume independence between all measurements in a group. Hence, sc/snRNA-seq provides the ability to identify distinct cell types and cell states, which then enables a more refined analysis for how each one of them varies between tumors (in terms of frequency or expression profile); yet it does not directly increase the sample size for comparisons between tumors, and such analysis remains limited by the number of analyzed tumors, regardless of the number of profiled cells.

To further clarify this issue, imagine that it was valid to compare the cells rather than the samples. In this case, one would be able to deeply profile only 2 samples (A and B, one responder and one nonresponder, respectively), with thousands of cells per tumor, and from this data identify genes that are significantly associated with treatment response. Clearly, such analysis would be invalid given that any difference between A and B would be automatically considered as also associated with treatment response. Instead, the statistical power for the associations with treatment response should be derived only from the number of samples, not from the number of cells. Accordingly, even in the age of single-cell studies it remains challenging to find reliable associations with tumor or patient features (response, survival, etc.) as this requires profiling many tumor samples.

Pitfall #3: Incorrect Inference of Chromosomal Aberrations

A common step in computational analysis of tumors involves the inference of chromosomal copy-number aberrations (CNA), also referred to as copy-number variations (CNV) or aneuploidy.[8–10] Except for rare tumors that lack CNAs, this provides an effective approach to detect the malignant cells and distinguish them from nonmalignant cells of the tumor microenvironment. Moreover, this approach provides a genetic description of the malignant cells, highlighting their specific patterns of CNAs. Finally, CNA inference often enables the detection of multiple genetic subclones in the same tumor, thereby exposing intratumor genetic evolution.

CNA inference is implemented by multiple methods, including the widely used inferCNV,[8] a newer version called inferCNA,[17] copyKat,[10] and others. However, these different methods follow a similar process. First, expression values are combined across neighboring genes in each chromosomal region. Second, this signal is normalized by comparison to reference cells that should have a normal karyotype. Third, the normalized values are interpreted to determine whether CNAs exist or not. Below I describe major challenges with the latter 2 steps that have important implications for CNA inference.

Correct Selection of Reference Cells (pitfall 3.1)

The second step for CNA inference requires comparison of chromosomal signals from the analyzed cells to a reference set of cells with normal karyotype. Notably, since the patterns observed in reference cells are subtracted from the patterns of the analyzed cells, correct selection of reference cells is essential. If reference cells include malignant cells, by mistake, then this could hinder the detection of CNAs. Moreover, even nonmalignant cells may not serve as an optimal reference and may lead to errors in CNA inference, as further described below.

CNA inference is based on average gene expression patterns across chromosomal regions. The assumption in this approach is that different genes in the same region tend to have independent regulation of their expression and therefore their average expression should be largely balanced within a cell, especially when considering a large enough number of chromosomally adjacent genes (eg, 100 genes). The obvious exception is in cases of aneuploidy, where the entire chromosomal region is lost or gained, in which case all genes in the region are affected in the same way, and hence their average expression is also affected.

However, it is important to keep in mind that even in the absence of aneuploidy, sets of neighboring genes are not entirely independent in their regulation, both due to regulatory mechanisms that affect multiple adjacent genes, and also because the location of genes in the genome is not random, such that adjacent genes are often functionally related and similarly regulated.[28] Therefore, the average expression across chromosomal regions is determined not only by aneuploidy but also by more subtle patterns that are unique to each cell type and each cell state. A commonly encountered example is that myeloid cells tend to have high expression of many genes in a region of chromosome arm 6p that harbors the MHC locus.[29] Therefore, when myeloid cells are used as a reference, one might expect the reference expression pattern to be inflated in chromosome 6p, and this may lead to the spurious inference of loss of chromosome 6p in the malignant cells, only due to the identity of the reference cells.

Since each cell type will have its own typical chromosomal expression pattern, it follows that the optimal reference should be of cells from the same cell type as the

queried cells, but with a normal karyotype, and from the same batch of data. However, it is typically not possible to obtain such reference cells. If we attempt to select reference cells that are most similar to the malignant cells, then we run the risk of accidentally choosing malignant cells; if we select control cells from a healthy tissue, then we run the risk of having batch effects between the analyzed tumor and the healthy sample; and if we choose a very different cell type that is clearly nonmalignant (eg, macrophages in a brain tumor) then we run the risk that our CNA inference will be affected by the unique expression pattern of the reference cells (eg, high MHC expression).

We therefore recommend the use of multiple references in parallel, such that only CNAs identified in comparison to the different references will be considered as reliable. This conservative approach can be manually implemented by running inferCNV multiple times but can also be automated through a more advanced implementation that combines multiple references together (InferCNA[17]). In brain tumors, we often use both macrophages and endothelial cells from a single tumor as alternative references for the suspected malignant cells of that same tumor. However, this requires ~50 cells of each cell type, which we do not detect in every glioma and hence in some cases we use only one cell type or alternatively use reference cells from other tumors from the same batch of data.

Interpretation of Inferred CNA Signal (pitfall 3.2)

The last step in CNA inference is to interpret whether calculated signals are indicative of a real chromosomal event or of other (non-aneuploidy) causes of expression patterns, such as noise, suboptimal reference cells, and other forms of gene regulation. While there is no simple rule for making such a distinction, several guidelines may be proposed, as detailed below.

First, the strongest signals in CNA inference are typically linked to aneuploidy, while other signals tend to be weaker, so the simplest guideline is to consider only the strongest observed patterns as likely aneuploidy. Second, some CNAs may be expected based on prior knowledge, such as chromosome 7 gain and 10 loss in glioblastoma, or chromosome 1p and 19q loss in oligodendroglioma. Therefore, an inferred CNA profile that contains these apparent events may be deemed more reliable than one that lacks expected patterns.

Third, real chromosomal events tend to be either focal (one or only a few genes) or large-scale (entire chromosomal arms or entire chromosomes).[30] Focal events are typically not detected well by CNA inference, since we need to average the signal across many (eg, 100) adjacent genes, and therefore the most reliable inferred CNA patterns are those that appear to cover entire chromosomal arms or entire chromosomes, while those events that appear to cover smaller segments within chromosomes could potentially reflect other causes than aneuploidy.

Fourth, we expect to find consistent CNA patterns across malignant cells in a tumor; some events may be shared by subsets of malignant cells (defining genetic subclone) while most detected events are typically shared across all malignant cells, reflecting early evolutionary events. Therefore, CNA patterns observed in large subsets of malignant cells may be considered as more reliable than patterns that are observed in individual (or in very few) cells, which may reflect dubious patterns due to noise or other non-aneuploidy effects.

Pitfall #4: Clustering and Its Visualization

After preprocessing, arguably the most basic step in single-cell analysis is clustering to detect different cell types as well as cell states. Accordingly, many clustering approaches have been developed and are used routinely.[31,32] Nevertheless, clustering results can heavily depend on the choice of clustering algorithm and its parameters. While a complete discussion of different clustering algorithms and their properties and challenges is beyond the scope of this review, the discrepancies between methods highlight the need to examine the robustness of clustering results across multiple methods and to carefully interpret inconsistent results.

While some methods automatically define the number of apparent clusters,[33] it is important to acknowledge that clustering with different resolutions (ie, number of clusters) often provides complementary views, such that a single optimal number may not exist (pitfall 4.1). One cause for the difficulty to define an appropriate number of clusters is the presence of continuous patterns of gene expression. While some groups of cells are highly distinct from one another (ie, without intermediate cells that resemble 2 or more groups to similar degrees), often we observe gradients of expression profiles. Such gradients may be decomposed into any number of clusters, depending on the methodology, although the exact number of defined clusters may have limited biological significance, only reflecting an oversimplification. It can therefore be advantageous to perform 2 subsequent steps of analysis: first, strictly define only robust and highly distinct clusters that likely reflect different cell types (while considering continuous patterns only as single clusters); and second, analyze the diversity within each cluster, using methodology that is better suited for continuous patterns, such as nonnegative matrix factorization or principal component analysis.[9,34,35]

As noted in previous sections, additional issues often arising during clustering include clusters that are driven by low complexity or quality, and clusters with coexpression of markers for 2 distinct cell types (reflecting doublets or cells highly affected by ambient RNA), both of which should be detected and excluded (pitfall 4.2). Finally, in addition to the methodology of clustering, it is also important to consider how clustering results are visualized, as data visualization often drives the process of discovery and hypothesis generation.

UMAP plots are widely used in the single-cell field as they provide a simple and elegant way to simultaneously visualize the separation between many clusters (ie, cell types). However, the ability to separate many clusters in a single 2D plot stems from the use of a nonlinear method in which the axes do not have a concrete biological meaning.[36] Therefore, distances between cells (or clusters) as well as

apparent shapes and trajectories in the UMAP plots cannot be reliably interpreted,[37] although many researchers intuitively trust and overinterpret such patterns (pitfall 4.3). Moreover, even for a single dataset, UMAP plots can look quite different, depending on their specific parameters. Hence, sc/snRNA-seq studies should ideally limit the use of UMAP plots, consider it only as an initial clustering visualization tool from which limited conclusions can be drawn, and design more focused follow-up analyses to address specific questions. Instead of UMAPs, examples of alternative visualizations that may be more directly amenable to interpretations include 2 main types of heatmaps: genes × cells heatmaps, showing the normalized expression levels of relevant genes across cells, while additional metadata features (eg, sample identity) may be included as aligned panels; and cells × cells heatmaps, showing correlations across cells, which are arranged by the clustering pattern.

Pitfall #5: Spurious Inference of Cellular Trajectories

When multiple cell states are identified for a given cell type, a natural follow-up question involves the relationship between those states. Does one state give rise to another state? A common approach to address such questions is to perform trajectory analysis, often also referred to as pseudotime analysis, in which the states may be assigned to different points along a theoretical time axis (pseudotime). Many methods have been developed for trajectory analysis, including multiple versions of Monocle,[6,38] Velocyto,[7] and others. The typical assumptions of such methods are that (1) cells dynamically transition between the states being examined and (2) the transitions are largely unidirectional and hence one state could be considered as preceding the other. Both of these assumptions are valid in many developmental systems, where cells differentiate from one state to another in a predictable manner.

However, these assumptions, and hence the use of trajectory methods, may not hold in other contexts such as in tumors. First, 2 states might not be connected by dynamic transitions if they reflect distinct genetic clones or cells from distinct tumors. Second, even if there are dynamic transitions, they might be bidirectional and cannot simply be described as one state preceding the other. Many studies have demonstrated the plasticity of cancer cells in which cells fluctuate between multiple states and hence do not conform to the hierarchical models that would be inferred from trajectory methods.[17,39–42]

Beyond these 2 generic assumptions, each trajectory method may have its own assumptions or limitations. For example, the widely used RNA Velocity method assumes that we can reliably estimate the fraction of unspliced mRNAs, per gene, and that this fraction primarily reflects the amount of recently generated nascent mRNAs compared to mature mRNAs.[7] However, the fraction of detected unspliced mRNAs may also be affected by the regulation of splicing, by DNA fragments that are sequenced along with the mRNA in sc/snRNA-seq, by the relative stability of different mRNAs and possibly by other processes. Given these assumptions, it is important to consider trajectory methods as generating hypotheses that require further validations. When validation is not possible, as is often the case in tumor studies, the predictions should be carefully considered.

Pitfall #6: Removing Biological Signal Through Batch Correction

Sc/snRNA-seq is prone to have batch effects, which complicate the integration of datasets from distinct platforms, labs, experimentalists, and even from different samples within a single project. It is therefore important to identify batch effects and correct them when possible. However, in cancer studies, where each patient may be considered as a separate batch, how can we distinguish between technical batch effects and genuine biological differences between patients? In extreme cases, for example, if all populations of malignant and nonmalignant cells are highly distinct between 2 patients, then one might infer that differences are primarily due to batch effects. But in other cases, where differences are more subtle and are seen primarily in certain populations (typically in malignant cells), we cannot confidently distinguish between biological and technical (batch) effects.

For example, we typically find that nonmalignant cell types cluster together across tumors, while malignant cells form patient-specific clusters.[9,43,44] This pattern suggests that technical batch effects may be limited in those cases, and that the malignant cells from different tumors are genuinely biologically different and hence should remain as separate clusters. However, if we apply batch correction methods then the malignant cells might be considered as a single-cell type and hence their biological differences would be at least partially removed. Since batch correction methods will be unable to fully decouple biological from technical effects, they may generate 2 types of errors: they might not remove certain technical effects, and perhaps more importantly, they would likely remove certain biological effects.[45,46]

A lot of effort is currently invested in developing integration methods that will combine data from different sources while removing batch effects. While such methods are indeed improving, it is important to acknowledge that these methods will not be able to fully distinguish biological from technical effects that differ between tumor samples, and therefore that data integration comes with a cost of potentially removing biological signals. This does not imply that integration methods should be avoided, but rather that their limitations and trade-offs should be appreciated. When batch effects appear to be limited, it is possible to perform the analysis in both ways (with and without batch correction) and examine whether the main results are robust to this choice. When performing batch correction, it should not be assumed that, after correction, cells from multiple tumors can be analyzed together without any signal reflecting the patient or sample identity,

as there would often still be residual patient-specific signals that could drive the clustering of cells.

An alternative approach to batch correction is to analyze each tumor separately, only comparing between cells from the same tumor, hence avoiding batch effects.[35] The patterns of diversity observed in each tumor (eg, intratumor signatures of specific subpopulations) can then be compared across tumors to detect recurrent patterns as well as those that are unique to certain tumors. This approach is efficient for the analysis of intratumor heterogeneity, although it is less appropriate to address certain questions about intertumor differences.

Pitfall #7: Oversimplified Inference From Bulk Datasets

Once gene expression signatures have been defined from the analysis of sc/snRNA-seq data (eg, reflecting novel subpopulations of cells), a common approach involves scoring of bulk datasets by these signatures to detect the patients that harbor the respective subpopulations of cells. This could be a powerful approach to extend sc/snRNA-seq analysis of a small cohort with limited clinical annotations into a large cohort with improved clinical annotations, such as TCGA cohorts. However, it is important to understand the limitations of this approach.

Most genes are not uniquely expressed by only one cell type and even the genes that are known markers of a particular cell type are often expressed by other cells. For example, CD4 is a well-known marker for CD4+ T-cells, but while its expression within T-cells is restricted to CD4+ T-cells, this gene is also expressed by macrophages and dendritic cells. Hence, scoring tumors for CD4 expression would not provide a reliable estimate for the abundance of CD4+T-cells as it would be confounded by the abundance of macrophages and dendritic cells.

Most studies score bulk tumors for gene sets (ie, signatures) rather than specific genes (eg, CD4) and hence may consider the issue of nonspecific expression as insignificant. In theory, if all genes in the signature are primarily expressed by the same cell type/state and each has a distinct secondary expression profile, then the secondary expression may average out across the different genes such that the combined signature will accurately capture only the intended cell type. However, in practice, the secondary expression profiles are often correlated across the genes in a signature such that they may strongly confound the signature score. For example, signatures for one type of proliferating cells (eg, proliferating T-cells) will often be confounded by expression of many cell cycle genes by other proliferating cells (eg, proliferating malignant cells); signatures for interferon-responsive fibroblasts will be confounded by expression of the same interferon-response genes by myeloid, malignant, and endothelial cells; and stress/activation signatures in one cell type (eg, JUN, FOS, ATF3 and their target genes) will be confounded by stress/activation of other cell types.

A common example of the confounding effect of stroma on malignant signatures involves epithelial-mesenchymal transitions (EMT).[47] Analysis of malignant cells within tumors often detects subpopulations that upregulate EMT-related genes, allowing the derivation of malignant EMT signatures.[35,44,48] However, since such analysis is typically restricted to malignant cells, it masks the expression of the same EMT genes by other cell types. Indeed, EMT genes are highly expressed by fibroblasts and pericytes, and to a lower degree also by endothelial cells and macrophages.[48] It is therefore difficult to distinguish whether high expression of EMT genes in a bulk sample reflects a malignant EMT-related state or high abundance of fibroblasts and pericytes. In fact, in head and neck cancer, the "mesenchymal" TCGA subtype appears to represent primarily the abundance of fibroblasts,[48] while in glioblastoma, the mesenchymal subtype appears to reflect the combined signal of mesenchymal-like cancer cells and macrophages/microglia.[49]

Nevertheless, careful selection of the EMT-related genes that are most specific to the malignant cells can improve the specificity of the signature to malignant cells.[48] It is therefore recommended that for the purpose of scoring bulk samples, gene signatures should not be used directly as derived from sc/snRNA-seq analysis but rather they should be pruned for the most specific genes, when comparing the intended cells to all other cell types that may be present in the bulk samples that will be examined (including cell types that might be excluded by dissociation and hence absent from certain scRNA-seq datasets). Such comparison could be done through the 3CA website,[50] which integrates sc/snRNA-seq data from more than 100 studies to derive the expression of each gene across distinct cell types in each tumor type.[35]

Importantly, the ability to derive reliable conclusions from scoring of bulk samples depends not only on the specificity of the genes in the signatures but also on the overall abundance of the respective cell type. In particular, for cell types with an overall low abundance in tumors, such as dendritic cells or endothelial cells, signatures of specific cell states are unlikely to perform well in bulk analysis.

For example, assume that the average frequency of endothelial cells in a certain cohort is 4%, and that by scRNA-seq analysis we identify 2 subsets of endothelial cells (E_1 and E_2) with equal frequencies. Even if we find a signature within endothelial cells that is completely unique to E_1 (relative to E_2) than its E_1 expression would reflect on average 2% of the cells per tumor. Now consider one of the genes (G) in that signature. Since G is likely expressed at some level by at least one of the more common cell types in those tumors, the bulk expression of G would likely not be dominated by its expression in endothelial cells. Even if G is expressed at a 10-fold lower level in malignant cells than in E_1 endothelial cells, then if the malignant cells constitute 40% of the cells in the tumor, the total expression of G in malignant cells would, on average, be twice higher in malignant than in endothelial cells. Therefore, G levels would not serve as a reliable estimate for E_1 cells, as it would be highly affected by the abundance and expression profiles of malignant cells. The same issue would likely hold for all genes in the E_1 signature.

In summary, bulk expression of a gene, or a gene set, reflects the aggregate expression of many cell types, and given that almost all genes are expressed by multiple cell

types, we must be careful in ascribing bulk expression profiles to any specific cell type. Bulk expression is most informative for highly specific genes that are expressed by common cell types but is difficult to interpret for genes expressed by multiple cell types and/or by low-frequency cell types. Instead of directly scoring bulk tumors with specific sc/snRNA-seq-derived signatures, many researchers use specialized deconvolution methods. While deconvolution methods like CIBERSORT[13] attempt to correct for the confounding effects noted above, their accuracy still depends on the specificity of expression signatures and the frequency of the corresponding cells. Therefore, deconvolution results should be interpreted with caution when applied to subtle cell state, to signatures with limited specificity (eg, proliferating cells, interferon-response and stress-response programs, and EMT), and to rare populations of cells.

Concluding Remarks

Each of the pitfalls described above is commonly encountered in cancer sc/snRNA-seq analysis, and mitigating those pitfalls is important in order to produce reliable results. In addition, there are other potential pitfalls and common mistakes that were not (or only minimally) covered here, such as overinterpretation of ambient RNA[51] and unrecognized doublet cells.[52] Overall, the complexity of sc/snRNA-seq analysis and the considerable potential for dubious results suggest that researchers should be cautious in their interpretations, acquire deep understanding of the computational methods that they use, explore their data with multiple approaches and parameters in order to reach robust conclusions, and perform their own QC and necessary experiments for validation of the various conclusions that are drawn. While this review is focused on sc/snRNA-seq, many of the issues described above will also be relevant for related technologies, such as single-cell ATAC-seq and spatial transcriptomics. More generally, as biomedical research is becoming more dependent on large datasets and their complex analysis, it is imperative that data analysis and interpretations of the results are done properly. This will ensure that we derive the most benefit from our datasets and that computational approaches will have an immense impact on biology and medicine.

Keywords

Single cell RNA-seq | Intratumor heterogeneity | Data analysis

Funding

This work was supported by an ERC (European Research Council) consolidator grant [101044318]. I.T. is the incumbent of the Dr. Celia Zwillenberg-Fridman and Dr. Lutz Zwillenberg Career Development Chair, and is supported by the Zuckerman STEM Leadership Program.

Supplement sponsorship

This article appears as part of the supplement "Single-Cell Technologies," sponsored by the Princess Margaret Cancer Research Centre and the Toronto Western Hospital Division of Neurosurgery.

Conflict of interest statement

I.T. is an advisory board member of Immunitas Therapeutics, and a scientific co-founder and advisory board member of Cellyrix Therapeutics.

Affiliations

Department of Molecular Cell Biology, Weizmann Institute of Science, Rehovot, Israel (I.T.)

References

1. Tirosh I, Suva ML. Cancer cell states: lessons from ten years of single-cell RNA-sequencing of human tumors. *Cancer Cell*. 2024;42(9):1497–1506.
2. Heumos L, Schaar AC, Lance C, et al.; Single-cell Best Practices Consortium. Best practices for single-cell analysis across modalities. *Nat Rev Genet*. 2023;24(8):550–572.
3. Zappia L, Theis FJ. Over 1000 tools reveal trends in the single-cell RNA-seq analysis landscape. *Genome Biol*. 2021;22(1):301.
4. Hao Y, Stuart T, Kowalski MH, et al. Dictionary learning for integrative, multimodal and scalable single-cell analysis. *Nat Biotechnol*. 2024;42(2):293–304.
5. Wolf FA, Angerer P, Theis FJ. SCANPY: large-scale single-cell gene expression data analysis. *Genome Biol*. 2018;19(1):15.
6. Cao J, Spielmann M, Qiu X, et al. The single-cell transcriptional landscape of mammalian organogenesis. *Nature*. 2019;566(7745):496–502.
7. La Manno G, Soldatov R, Zeisel A, et al. RNA velocity of single cells. *Nature*. 2018;560(7719):494–498.
8. inferCNV of the Trinity CTAT Project. https://github.com/broadinstitute/inferCNV
9. Tirosh I, Venteicher AS, Hebert C, et al. Single-cell RNA-seq supports a developmental hierarchy in human oligodendroglioma. *Nature*. 2016;539(7628):309–313.
10. Gao R, Bai S, Henderson YC, et al. Delineating copy number and clonal substructure in human tumors from single-cell transcriptomes. *Nat Biotechnol*. 2021;39(5):599–608.
11. Korsunsky I, Millard N, Fan J, et al. Fast, sensitive and accurate integration of single-cell data with Harmony. *Nat Methods*. 2019;16(12):1289–1296.
12. Xu C, Lopez R, Mehlman E, et al. Probabilistic harmonization and annotation of single-cell transcriptomics data with deep generative models. *Mol Syst Biol*. 2021;17(1):e9620.
13. Chen B, Khodadoust MS, Liu CL, Newman AM, Alizadeh AA (2018). Profiling tumor infiltrating immune cells with CIBERSORT. In: von Stechow L, ed. *Cancer Systems Biology. Methods in Molecular Biology*, vol 1711. New York, NY: Humana Press. https://doi.org/10.1007/978-1-4939-7493-1_12
14. Wang X, Park J, Susztak K, Zhang NR, Li M. Bulk tissue cell type deconvolution with multi-subject single-cell expression reference. *Nat Commun*. 2019;10(1):380.
15. Kharchenko PV. The triumphs and limitations of computational methods for scRNA-seq. *Nat Methods*. 2021;18(7):723–732.
16. Patel AP, Tirosh I, Trombetta JJ, et al. Single-cell RNA-seq highlights intratumoral heterogeneity in primary glioblastoma. *Science*. 2014;344(6190):1396–1401.

17. Neftel C, Laffy J, Filbin MG, et al. An integrative model of cellular states, plasticity, and genetics for glioblastoma. *Cell*. 2019;178(4):835–849.e21.
18. Filbin MG, Tirosh I, Hovestadt V, et al. Developmental and oncogenic programs in H3K27M gliomas dissected by single-cell RNA-seq. *Science*. 2018;360(6386):331–335.
19. Slyper M, Porter CBM, Ashenberg O, et al. A single-cell and single-nucleus RNA-seq toolbox for fresh and frozen human tumors. *Nat Med*. 2020;26(5):792–802.
20. Chen J, Larsson L, Swarbrick A, Lundeberg J. Spatial landscapes of cancers: insights and opportunities. *Nat Rev Clin Oncol*. 2024;21(9):660–674.
21. Greenwald AC, Darnell NG, Hoefflin R, et al. Integrative spatial analysis reveals a multi-layered organization of glioblastoma. *Cell*. 2024;187(10):2485–2501.e26.
22. Islam S, Zeisel A, Joost S, et al. Quantitative single-cell RNA-seq with unique molecular identifiers. *Nat Methods*. 2014;11(2):163–166.
23. Suva ML, Rheinbay E, Gillespie SM, et al. Reconstructing and reprogramming the tumor-propagating potential of glioblastoma stem-like cells. *Cell*. 2014;157(3):580–594.
24. Puram SV, Mints M, Pal A, et al. Cellular states are coupled to genomic and viral heterogeneity in HPV-related oropharyngeal carcinoma. *Nat Genet*. 2023;55(4):640–650.
25. Tirosh I, Izar B, Prakadan SM, et al. Dissecting the multicellular ecosystem of metastatic melanoma by single-cell RNA-seq. *Science*. 2016;352(6282):189–196.
26. Liu H, Hu K, O'Connor K, Kelliher MA, Zhu LJ. CleanUpRNAseq: an R/Bioconductor package for detecting and correcting DNA contamination in RNA-seq data. *BioTech (Basel)*. 2024;13(3):30.
27. Wang M, Long Q. Addressing common misuses and pitfalls of *P* values in biomedical research. *Cancer Res*. 2022;82(15):2674–2677.
28. Ribeiro DM, Ziyani C, Delaneau O. Shared regulation and functional relevance of local gene co-expression revealed by single cell analysis. *Commun Biol*. 2022;5(1):876.
29. Kulski JK, Suzuki S, Shiina T. Human leukocyte antigen super-locus: nexus of genomic supergenes, SNPs, indels, transcripts, and haplotypes. *Hum Genome Var*. 2022;9(1):49.
30. Ben-David U, Amon A. Context is everything: aneuploidy in cancer. *Nat Rev Genet*. 2020;21(1):44–62.
31. Kiselev VY, Andrews TS, Hemberg M. Challenges in unsupervised clustering of single-cell RNA-seq data. *Nat Rev Genet*. 2019;20(5):273–282.
32. Waltman L, van Eck NJ. A smart local moving algorithm for large-scale modularity-based community detection. *The European Physical Journal B*. 2013;86:471.
33. Shahapure KR, Nicholas C. Cluster Quality Analysis Using Silhouette Score. *2020 IEEE 7th International Conference on Data Science and Advanced Analytics (DSAA)*, Sydney, NSW, Australia, 2020, pp. 747–748. doi: 10.1109/DSAA49011.2020.00096.
34. Kotliar D, Veres A, Nagy MA, et al. Identifying gene expression programs of cell-type identity and cellular activity with single-cell RNA-Seq. *Elife*. 2019;8:e43803.
35. Gavish A, Tyler M, Greenwald AC, et al. Hallmarks of transcriptional intratumour heterogeneity across a thousand tumours. *Nature*. 2023;618(7965):598–606.
36. McInnes L, Healy J, Saul N, Großberger L. UMAP: uniform manifold approximation and projection for dimension reduction. *ArXiv*. 2018;3:861.
37. Chari T, Pachter L. The specious art of single-cell genomics. *PLoS Comput Biol*. 2023;19(8):e1011288.
38. Trapnell C, Cacchiarelli D, Grimsby J, et al. The dynamics and regulators of cell fate decisions are revealed by pseudotemporal ordering of single cells. *Nat Biotechnol*. 2014;32(4):381–386.
39. Bhat GR, Sethi I, Sadida HQ, et al. Cancer cell plasticity: from cellular, molecular, and genetic mechanisms to tumor heterogeneity and drug resistance. *Cancer Metastasis Rev*. 2024;43(1):197–228.
40. Chaligne R, Gaiti F, Silverbush D, et al. Epigenetic encoding, heritability and plasticity of glioma transcriptional cell states. *Nat Genet*. 2021;53(10):1469–1479.
41. Perez-Gonzalez A, Bevant K, Blanpain C. Cancer cell plasticity during tumor progression, metastasis and response to therapy. *Nat Cancer*. 2023;4(8):1063–1082.
42. Rambow F, Marine JC, Goding CR. Melanoma plasticity and phenotypic diversity: therapeutic barriers and opportunities. *Genes Dev*. 2019;33(19-20):1295–1318.
43. Izar B, Tirosh I, Stover EH, et al. A single-cell landscape of high-grade serous ovarian cancer. *Nat Med*. 2020;26(8):1271–1279.
44. Puram S, Tirosh I, Parikh A, et al. Single-cell transcriptomic analysis of primary and metastatic tumor ecosystems in head and neck cancer. *Cell*. 2017;171(7):1611.
45. Emmanúel Antonsson S, Melsted P. Batch correction methods used in single cell RNA-sequencing analyses are often poorly calibrated. *bioRxiv*. 2024:2024.2003.2019.585562.
46. Tyler SR, Guccione E, Schadt EE. Erasure of biologically meaningful signal by unsupervised scRNAseq batch-correction methods. *bioRxiv*. 2023:2021.2011.2015.468733.
47. Dongre A, Weinberg RA. New insights into the mechanisms of epithelial-mesenchymal transition and implications for cancer. *Nat Rev Mol Cell Biol*. 2019;20(2):69–84.
48. Tyler M, Tirosh I. Decoupling epithelial-mesenchymal transitions from stromal profiles by integrative expression analysis. *Nat Commun*. 2021;12(1):2592.
49. Hara T, Chanoch-Myers R, Mathewson ND, et al. Interactions between cancer cells and immune cells drive transitions to mesenchymal-like states in glioblastoma. *Cancer Cell*. 2021;39(6):779–792.e11.
50. *Curated Cancer Cell Atlas*.
51. Young MD, Behjati S. SoupX removes ambient RNA contamination from droplet-based single-cell RNA sequencing data. *GigaScience*. 2020;9(12):giaa151.
52. Xi NM, Li JJ. Benchmarking computational doublet-detection methods for single-cell RNA sequencing data. *Cell Systems*. 2021;12(176):176–194.e6.