Molecular Therapy
Methods & Clinical Development
Review

# An Introduction to the Analysis of Single-Cell RNA-Sequencing Data

Aisha A. AlJanahi,[1,2] Mark Danielsen,[2] and Cynthia E. Dunbar[1]

[1]Translational Stem Cell Biology Branch, NHLBI, NIH, Bethesda, MD, USA; [2]Department of Biochemistry and Molecular & Cellular Biology, Georgetown University Medical Center, Washington, DC, USA

**The recent development of single-cell RNA sequencing has deepened our understanding of the cell as a functional unit, providing new insights based on gene expression profiles of hundreds to hundreds of thousands of individual cells, and revealing new populations of cells with distinct gene expression profiles previously hidden within analyses of gene expression performed on bulk cell populations. However, appropriate analysis and utilization of the massive amounts of data generated from single-cell RNA sequencing experiments are challenging and require an understanding of the experimental and computational pathways taken between preparation of input cells and output of interpretable data. In this review, we will discuss the basic principles of these new technologies, focusing on concepts important in the analysis of single-cell RNA-sequencing data. Specifically, we summarize approaches to quality-control measures for determination of which single cells to include for further examination, methods of data normalization and scaling to overcome the relatively inefficient capture rate of mRNA from each cell, and clustering and visualization algorithms used for dimensional reduction of the data to a two-dimensional plot.**

Until recently, single-cell gene expression profiling was limited to studying several select transcripts from a few individual cells. High-throughput sequencing along with high-yield cell separation methods have paved the way to modern single-cell sequencing platforms such as Fluidigm C1, DropSeq, Chromium 10X, SCI-Seq, and many others that have been developed over the past decade (Table 1). These technologies are able to characterize the transcriptional profile of hundreds up to many thousands of single cells at a time. All rely on labeling mRNA molecules with DNA barcodes during reverse transcription and/or subsequent steps, which allows indexing of the transcripts back to their individual cells of origin. Although each method is unique in the way it separates cells and labels the mRNA molecules, they all rely on similar computational pipelines for the representation of the transcriptional profiles. In this review, we will discuss some of the most common algorithms used in these computational pipelines, using DropSeq as a primary example, because it is the most cost-effective and widely available single-cell gene expression platform (Table 1). However, these concepts are applicable to most single-cell sequencing platforms that use DNA barcodes as an approach to link mRNA transcripts to a single cell of origin.

Single-cell RNA sequencing has been uniquely valuable to gain insights into cellular heterogeneity in tissues and for identification of previously unknown cell types.[1–3] Single-cell technologies can also be used to define subpopulations within a known cell type by searching for differential gene expression patterns within the cell population of interest.[1,4] In addition, these technologies can effectively isolate the signal from rare cell populations, which would be hidden in output from bulk cell population RNA sequencing.[5–8] Moreover, the technology can be used to infer potentially useful markers, such as cell surface proteins, for cell types with no known markers. Because single-cell sequencing analysis is driven by clustering of cells based on their differentially expressed genes, the genes that drive the clustering can be examined as possible unique markers for the cell population of interest.[1,9] Lastly, single-cell sequencing can be employed in studies of cell lineage and the regulation of differentiation. For example, a population of stem cells can be induced to differentiate, and single-cell sequencing performed at series of time points can provide "snapshots" of the progression of differentiation. These snapshots can then be used to infer the trajectories that cells follow to reach each terminally differentiated state and the key genes that are differentially regulated at each branch point.[1,10–12]

Many of these applications rely on specialized algorithms that have been developed and made available by leading bioinformatics labs. However, in this review, we will focus on the basic quality-control and data normalization pipeline that all single-cell sequences must undergo before applying any specialized algorithms. We will also discuss simple cell clustering and visualization algorithms.

## Generation of Single-Cell Expression Datasets via Droplet Methodologies

Droplet-based single-cell gene expression approaches, including DropSeq[13] and the commercial 10X platform,[14–19] use microfluidic chips to isolate single cells along with single beads in oil-encapsulated droplets, using microfluidics to bring oil, beads, and cell suspensions together in such a way that each droplet contains at most a single cell.[20] The beads are coated with DNA oligos that are composed of a

https://doi.org/10.1016/j.omtm.2018.07.003.

**Correspondence:** Cynthia E. Dunbar, MD, Translational Stem Cell Biology Branch, NHLBI, NIH, Clinical Research Center, Room 4E-5132, 9000 Rockville Pike, Bethesda, MD 20892, USA.
**E-mail:** dunbarc@nhlbi.nih.gov

**Table 1. Widely Used Single-Cell Sequencing Methods**

| *Sequencing Method* | Starting Cell No. | Cell Separation | Notes | Cell Capture | Transcript Capture | Representative Library Prep Cost per Cell[a] |
|---|---|---|---|---|---|---|
| *Fluidigm C1*[b] | ~1,000 cells | cells capture in size-specific chambers | must know the size of cells of interest; allows for staining and imaging prior to cell rupture | 96- or 800-chamber units are available | an average of 6,606 genes/cell (no data on percentage) | $1.70 |
| *DropSeq* | ~150,000 cells/run | droplet-based separation | remains the most cost-effective and most customizable | ~5% of cells per run (approximately 7,000 cells) | ~10.7% of the cell's transcripts | $0.06 |
| *Chromium 10X* | ~1,700 cells/run | droplet-based separation | the most commercially successful method; almost fully automated | ~65% of cells per run (approximately 1,000 cells) | ~14% of the cell's transcripts | $0.10 |
| *SCI-Seq* | ~500,000 cells (depends on experimental design) | FACS sorter; cells are never singly isolated | combinatorial indexing of individual methanol-fixed permeable cells | 5%–10% of cells | ~10%–15% of the cell's transcripts | $0.05–$0.14[c] |

All of the methods require the establishment of a cell dissociation technique. The price is highly dependent on the number of cells sequenced, the desired depth of sequencing, and the sequencing platform used. For this table, the prices are at the lower end of the price range for single-cell library prep.
[a]As of July 2018.
[b]Based on the 800-chamber medium-size isolation unit.
[c]Dependent on how many cells are prepped for sequencing and how many doublets are tolerated.

poly(T) tail at the 3′ end for the capture of cellular mRNAs, and at the 5′ end both a cell barcode that is identical for every oligo coating an individual bead and a library of individual unique molecular identifier (UMI) barcodes of high diversity, each UMI different for every oligo on the bead (Figure 1A).[13,21,22] The transcripts from each individual cell captured and labeled by the DNA oligos attached to a bead within the droplets are reverse transcribed, amplified with PCR, and sequenced using a high-throughput platform, after breaking and pooling droplet contents.[23–28] The resulting sequences are aligned to a reference genome in order to annotate each transcript with its gene name. The cell barcodes on the aligned sequences allow for the computational linking of each gene transcript to its cell of origin. The number of copies of individual gene transcripts expressed in each individual cell is tallied using the UMIs, allowing the assembly of digital gene expression matrices (DGEs), which are tables of cell barcodes and gene counts.[13,21,29–31]

The insights feasible from such complicated sequencing data are only as good as the computational interpretation that follows. Most available single-cell sequencing algorithms do not have a graphic user interface. Thus, performing single-cell analysis requires knowledge of some programming languages in order to interact with the pre-established algorithms for aligning, clustering, and visualizing the data. In addition, in-depth knowledge of the biology of the cells of interest is essential to correctly interpret the data and make appropriate decisions based on quality-control measures. A bioinformatician with expertise in single-cell sequencing is able to generate analyses that can be used to make meaningful biological inferences, choosing appropriate cutoffs for the algorithms applied and avoiding misleading results.

### Quality-Control Metrics

Because droplet-based experiments can be considered to be thousands of separate experiments taking place on individual cells within individual droplets, it is essential to apply quality-control (QC) metrics designed to decide which of these individual droplet datasets is valid for further interpretation (Figure 2). QC can most effectively be performed on droplet-based datasets by applying a number of different parameters that detect unsuccessful droplets and exclude their data from further analyses.[13,32–37]

QC parameters are unique to each run because they are dependent on the cells or tissue being sequenced. A common QC metric is the number of transcripts tallied per cell, or the percent of transcripts per cell that align to the reference genome. Cells with transcript counts below or above a defined cutoff are marked as outliers and are not included in further analysis; these outlier cutoffs can be user-defined for each experiment (e.g., cells with less than 20 transcripts and more than 5,000 are removed from the analysis), or automatically applied by the program (e.g., cells with a sum of transcripts larger than 2 SDs from the mean are removed). The presence of a very large number of transcripts with one cell barcode may result from doublets (i.e., two or more cells suspended in one droplet), and such data are eliminated from the analysis. Conversely, a small number of transcripts per cell barcode is often an indicator of poor capture quality, which could be because of cell death, premature cell rupture, or the capture of random mRNA escaping from cells and floating in the cell suspension reagents.[13,32,34,36,37] Additional QC metrics can be applied, for instance, simply excluding all cells that express a specific gene to remove contaminating cells that are not of interest from an analysis, or can be more elaborate, for example, including only cells that have a specific ratio of two or more specific genes.[33,36]

When deciding on QC cutoffs, the diversity of the tissue being analyzed must be taken into account. For instance, when designing an experiment to study migrating cancer cells found in the blood, where the number of cancer cells is very low compared with the overall number of normal blood cells, the *counts of transcripts* QC metric must be adjusted. In this tissue, the dominant cells are blood cells, which are generally quiescent and have relatively low amounts of RNA compared with active cancer cells.[38] If all cells with a transcript

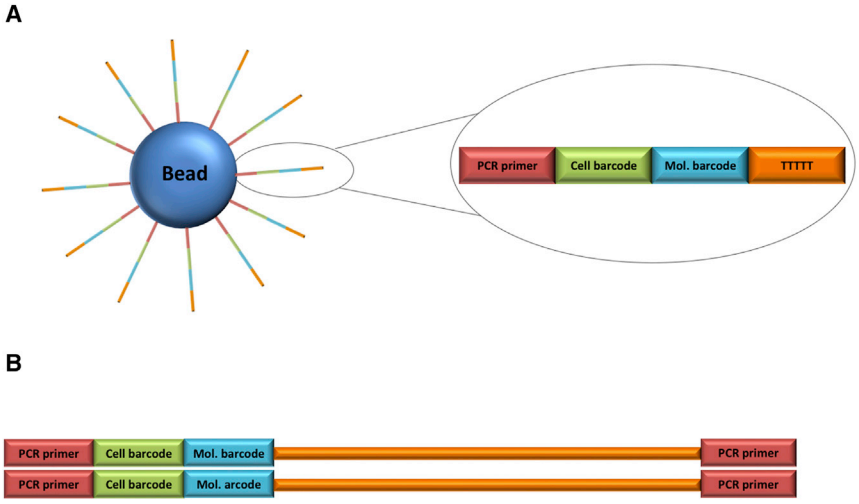


**Figure 1. The Structure of Drop-Seq Bead and Resulting Sequence Libraries**

(A) The structure of a DropSeq single-cell sequencing bead. The oligos extending from the bead have a PCR primer, a cell barcode that is unique to the bead to label each cell, a UMI that is unique to each individual oligo arm to allow unique labeling of each captured molecule, and a poly(T) tail to capture poly(A)-tailed mRNAs. (B) Structure of the sequencing ready library. Red: PCR primers which are also used as sequencing primers. Green and blue: the cellular and molecular barcodes from the bead. Orange: the captured transcript with the poly(A/T) tail.

count higher than 2 SDs from the mean are removed from the analysis, it could lead to the elimination of all cancer cells, mistaking them for doublets because of their high transcriptional activity compared with the much larger population of blood cells. Setting cutoffs appropriately may require spike in experiments prior to running experimental samples.

Another common QC metric is the number of mitochondrial gene transcripts.[32,33,35,39,40] High numbers of mitochondrial transcripts are indicators of cell stress,[41] and therefore cells with elevated mitochondrial gene expression are often not included in the analysis, because most experiments will not benefit from clustering cells based on stress levels. However, just as with *number of transcripts*, this parameter is highly dependent on the tissue type and the questions being investigated. For example, 30% of total mRNA in the heart is mitochondrial due to high energy needs of cardiomyocytes, compared with 5% or less in tissues with low energy demands.[42] For instance, 30% mitochondrial mRNA is representative of a healthy heart muscle cell, but would represent a stressed lymphocyte.

Depending on the goal of an experiment, a gene-specific QC metric can be added. Genes that are always present in very low quantities, and will never reach statistical significance between cell types, may be removed from the analysis to decrease the computational load.[32,33,36] This can be done by either setting a cutoff of gene count per cell (e.g., gene count in cell = <5 in all of the sequenced cells) or setting a cutoff for the count sum across all or a subset of cells (e.g., $\sum$count of gene across all of the sequenced cells = <300). Excluding such genes from the analysis will speed the computational process; however, some genes that are very slightly differentially expressed and contribute to the variance of the data might be lost.

### Data Normalization and Scaling

When analyzing sequencing data, normalization to eliminate batch effects is crucial if multiple sequencing runs are to be compared with each other. These batch effects can be caused by often unavoidable technical variations such as the duration samples were kept on ice, number of freeze-thaw cycles, method of RNA isolation, sequencing depth, etc.[43–45] Investigators should always strive to keep these variables as constant as possible between experiments and sequencing runs. However, droplet-based sequencing in addition consists of thousands of individual cell experiments, hence cell-specific biases must also be considered when normalizing, in order to be able to compare the expression of one cell to another.[46,47] A notable cell-specific bias is caused by mRNA capture efficiency, where the mRNA molecules are not captured by the bead at the same proportion in all droplets (Table 1). This is referred to as "dropout events," and it is the main cause for sparsity of data, which we will discuss further in the next paragraph. Furthermore, for bulk RNA sequencing, normalizing data involve comparing multiple batches of similar biological material (e.g., comparing blood cells with blood cells), but in single-cell sequencing the individual cells are not all of the same type. This requires adjusting the normalization parameters to retain cell-to-cell variability while eliminating technical noise caused by batch effects and cell-specific biases.[47]

The sparsity of data, due to the inefficiency of mRNA capture (i.e., at best, DropSeq is predicted to capture about 10% of each cell's mRNA[13]) poses the biggest challenge for the analysis of droplet single-cell sequencing data. The DGE matrix is expected to be mostly filled with zeros because of these dropout events.[48] Therefore, normalization and scaling are vital prior to interpreting the data.[44] Unfortunately, this requires making assumptions about the cells that can be biologically inaccurate. An accepted way to normalize the sequencing data is based on comparisons with housekeeping genes.[32] Based on literature and knowledge of the biological sample sequenced, a housekeeping gene is selected for normalization. The selected gene is assumed to be expressed at the same level in all cells, and the sequencing data are scaled to make the expression level of the selected housekeeping gene equal in all cells. However, the housekeeping gene method can be inaccurate because these genes are not always present in the same amount in different cell populations.[49,50] To avoid making the assumption that a housekeeping gene is present in an equal amount in all cells, the scaling can be based on all non-differentially expressed genes in all or some of the cells.[32,51] This

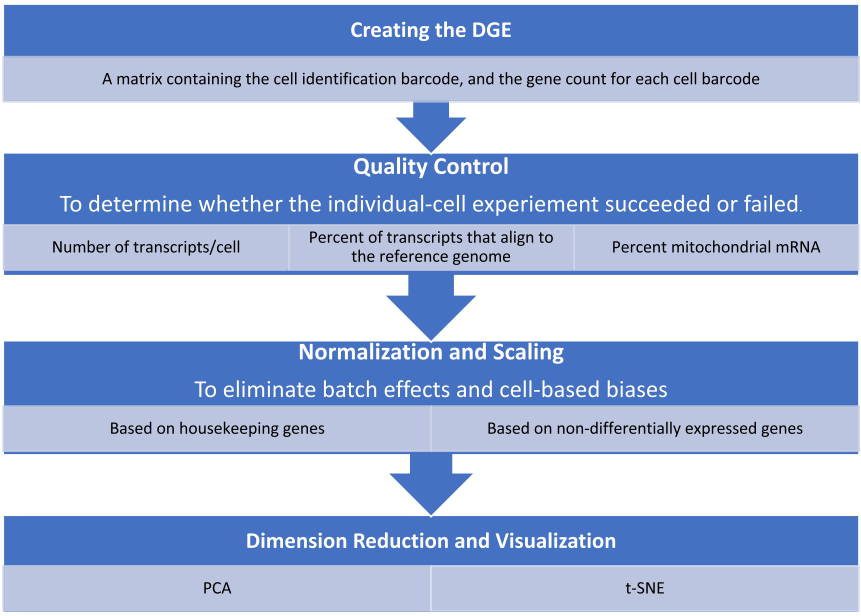


**Figure 2. Analysis Workflow**

approach assumes that all genes that are non-differentially expressed between cells are expressed equally in all cells, and it infers a scaling factor for each cell that is used to normalize transcript counts.[32]

### Dimension Reduction and Visualization

After normalization, an unbiased clustering algorithm can be used to determine which cells are closely related based on their gene expression profiles. Principal component analysis (PCA) is often the clustering algorithm of choice, because it is a relatively simple linear dimensionality reduction algorithm that can predict the relatedness of multidimensional data, or in this case, predict the relatedness of cells based solely on differential gene expression.[5,10,32,37,52–56] PCA merges the information from correlated genes into one "metagene" or principal component (PC). By definition, PC1 explains the greatest possible variance in the data and has the largest SD (e.g., for a specific experiment 30% of the variance between the cells is explained by genes that define PC1). PC2 explains the second greatest portion of variance in the data, and so on (e.g., an additional 20% of the variance between the cells will be due to genes that define PC2, and an extra 8% is attributed to the genes that define PC3). The PCs are ranked based on significance of explaining the data's variance, with PC1 being the highest-ranking PC. The lower ranking the PC, the less it contributes to explaining the variance of the data. Therefore, using the lower ranked PCs is generally not advantageous because it increases the computational load, yet barely adds any information to the representation of the biological variability of the cells. Thus, deciding how many PCs to use for visualization is important. This can be done using visual plotting methods. A simple example is the knee or elbow plot like the one shown in Figure 3, where the SD for each PC is plotted to represent the amount of variance encompassed within that PC.[37] As expected, the first PC contributes the most to the variance and has the highest SD. After the fifth PC, the contribution to the explanation of variance plateaus. That is where the "elbow" is determined to be, and it becomes the cutoff for PC inclusion in visualization.

t-Distributed stochastic neighbor embedding (t-SNE) is a common visualization approach.[57–59] It uses a machine learning algorithm that reduces dimensions and is well suited for embedding high-dimensional data into two- or three-dimensional space for visualization, without losing information about the relative distance between the plotted data points or, in this case, cells. For example, if the diversity of the cells was found to be well represented with seven PCs, then seven axes or dimensions are required to represent the cells. t-SNE will plot the cells on a two-dimensional plot in a way that maintains the seven-dimension relationship between cells, so that cells that are neighbors on a seven-dimension plot remain neighbors on a two-dimension plot. Whereas PCs analysis is linear, t-SNE is a non-linear dimensional reduction method.

### Considerations Regarding Data Generation Efficiency and Alternative Single-Cell Platforms

The computational approaches discussed in this review are focused on droplet-based separation methods such as DropSeq and Chromium 10X. However, because most single-cell sequencing platforms share the main principal of labeling mRNA from each cell with unique DNA barcodes, then computationally tracing back these molecules to their cells of origin, similar principles and algorithms can be used for datasets from other approaches, keeping in mind the type of technical variations or artifacts that may result from specific processes included in different platforms. These platforms and methods of cell separation, labeling, and DNA amplification have been recently reviewed in detail by Valihrach et al.,[60] where they discuss the basic principles of each platform and the advantages and pitfalls of the methods.

For example, in SCI-Seq,[61] cells are fixed with alcohol, making them permeable. The fixed cells are sorted using a flow cytometric sorter, dispensing specific numbers of cells into each well on a multiwall plate (Table 1). The mRNA of the cells in each well is barcoded with an oligo unique to that well via reverse transcription. The cells from all wells are then pooled, and another round of fluorescence-activated cell sorting (FACS) of cells into wells at lower density is performed, followed by adding a second unique well-specific barcode, creating a unique barcode combination for each cell. This process can be repeated again to reduce the chance of two cells being labeled with

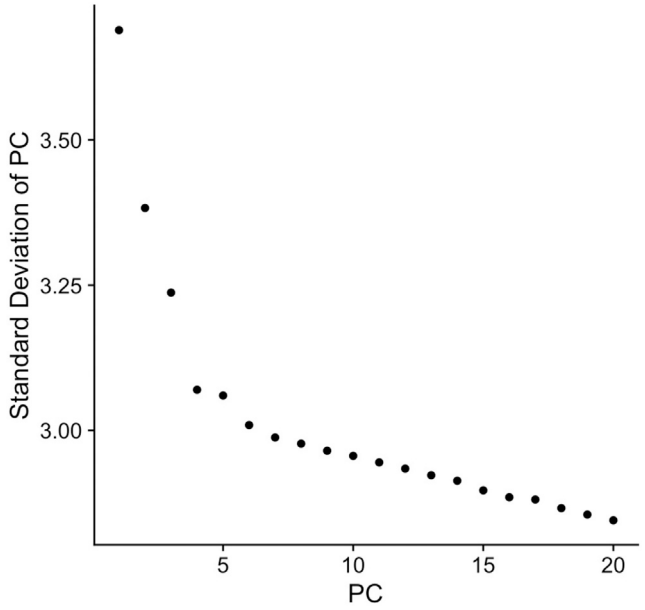


**Figure 3. Elbow Plot Analysis of Principle Components Variance**
A plot of the SD of each principal component, representing the amount of variance it contributes to the data. Here, the plot shows an elbow at around PC5. PC4 and PC6 are also valid choices for the PC cutoff.

the same barcode combinations. This combinatorial approach to single-cell labeling requires a specialized algorithm to create the DGE matrix, because in contrast with droplet-based methods, a single cell is not defined by a single barcode but instead by a unique combination of barcodes. It is worth noting that this approach, requiring at least two rounds of cell sorting, could be more stressful to cells and impact on gene expression.

Another example is the interpretation of the *number of transcripts/cell* parameter to exclude doublets, because each method is expected to produce a different rate of doublets. In the Fluidigm C1 system, where individual cells are captured by size-specific chambers, the doublet rate drops from 7% to 3% after microscopic examination of cells in the 96-chamber medium size isolation unit (Table 1).[62] The rate is not zero because cells are sometimes stacked on top of each other in the isolation chamber, making them look like single cells, and therefore can be missed by microscopy.[63] If the number of transcripts is significantly higher (e.g., more than 2 SDs higher than the mean) in more than 3% of the microscopically examined cells or 7% of non-examined cells, this could indicate a mixed cell population composed of a small fraction of transcriptionally active cells and a larger portion of quiescent cells, or it could be due to a high rate of true doublets, in which case the size of the size-specific isolation chambers might be inappropriate for the cell population being studied.

It should be noted that most popular single-cell analysis pipelines are driven by the most differentially expressed genes between cells. This is beneficial for finding gene markers for unknown populations.[64–66] However, if researchers aim to study cell types that are very similar, or find subpopulations within one major cell type, then those cells can be sorted prior to analysis in order to increase the number of cells of interest, thereby increasing the power of the analysis. Even though FACS has been shown to have a minimal effect on gene expression,[67,68] sorting prolongs the time that cells are not in optimal culture conditions and kept in a single-cell suspension, which could stress the cells and possibly alter mRNA and mitochondrial mRNA expression.[68–71] Also, passing of cells in small chambers, or through microfluidics or a cell sorter can cause shear stress and impact some cell types more than others in terms of causing cell stress or death, especially because the cells are vulnerable in a single-cell suspension.[72–74] Therefore, delicate cell types might be under-represented in droplet-based single-cell sequencing experiments, especially if the cells were sorted prior to single-cell isolation.

## Conclusions

In summary, we have discussed concepts important to applying analytic pipelines for the analysis of single-cell gene expression data, and specific parameters that change depending on cell types or condition. We also provide examples of some types of technical variation that need to be considered in order to adapt this pipeline to non-droplet-based methods. The pipeline starts with the creation of a DGE matrix, which contains gene counts in each cell, from the raw sequencing files. The rest of the analysis is applied on this matrix file. QC determines which cells to exclude from downstream analysis because of various reasons like the suspicion of doublets or cellular stress. Normalization and scaling are then performed to compensate for the sparsity of data because of the low mRNA capture rate. Then, dimension reduction is done based on the most differentially expressed genes. Finally, if done correctly, visualization of the data will result in plots showing the relatedness of each cell to its neighbor in two- or three-dimensional space.

These algorithms are in general use and are often included in an easy-to-use packages such as Seurat[13,37] (https://satijalab.org/seurat/), an R-based package that creates R objects compatible with other downstream algorithms; scran[51] (http://bioconductor.org/packages/release/bioc/html/scran.html), which also includes algorithms for cell-cycle assignment; ascend[75] (https://github.com/IMB-Computational-Genomics-Lab/ascend), which includes well-established and new algorithms providing a flexible analysis framework; and many others.[33,34,36,37,46,76–86] A few of these packages were reviewed by Yip et al.[87] comparing their accuracy and precision in detecting highly variable genes. These pipelines are usually followed by more specialized algorithms, depending on the purpose of the experiment. A few examples of specialized algorithms include Monocle (http://cole-trapnell-lab.github.io/monocle-release/), an algorithm designed to analyze differentiation trajectories of single cells;[11,12] SingleSplice (https://github.com/jw156605/SingleSplice), used to study alternative splicing in single-cell populations;[88] and OncoNEM (https://bitbucket.org/edith_ross/onconem/src), a tool to infer the hierarchical evolutionary relationships between tumor cells based on their somatic mutations.[89] Extensive collections of single-cell

sequencing tools and their applications can be found on websites such as https://github.com/seandavi/awesome-single-cell and https://www.scrna-tools.org/, which are updated frequently as new tools become available.

## CONFLICTS OF INTEREST

The authors have no relevant conflicts of interest to report.

## ACKNOWLEDGMENTS

We thank Dr. Stefan Cordes for illuminating discussions regarding SCI-Seq and single-cell sequencing, and Lauren Truitt for providing the Figure 3 image. This work was supported by the Intramural Program of the National Heart, Lung, and Blood Institute (C.E.D. and A.A.A.) and the Saudi Arabian Cultural Mission to the US (A.A.A.).

## REFERENCES

1. Artegiani, B., Lyubimova, A., Muraro, M., van Es, J.H., van Oudenaarden, A., and Clevers, H. (2017). A single-cell RNA sequencing study reveals cellular and molecular dynamics of the hippocampal neurogenic niche. Cell Rep. *21*, 3271–3284.
2. Villani, A.-C., Satija, R., Reynolds, G., Sarkizova, S., Shekhar, K., Fletcher, J., Griesbeck, M., Butler, A., Zheng, S., Lazo, S., et al. (2017). Single-cell RNA-seq reveals new types of human blood dendritic cells, monocytes, and progenitors. Science *356*, eaah4573.
3. Glass, L.L., Calero-Nieto, F.J., Jawaid, W., Larraufie, P., Kay, R.G., Göttgens, B., Reimann, F., and Gribble, F.M. (2017). Single-cell RNA-sequencing reveals a distinct population of proglucagon-expressing cells specific to the mouse upper small intestine. Mol. Metab. *6*, 1296–1303.
4. Shalek, A.K., Satija, R., Adiconis, X., Gertner, R.S., Gaublomme, J.T., Raychowdhury, R., Schwartz, S., Yosef, N., Malboeuf, C., Lu, D., et al. (2013). Single-cell transcriptomics reveals bimodality in expression and splicing in immune cells. Nature *498*, 236–240.
5. Shalek, A.K., Satija, R., Shuga, J., Trombetta, J.J., Gennert, D., Lu, D., Chen, P., Gertner, R.S., Gaublomme, J.T., Yosef, N., et al. (2014). Single-cell RNA-seq reveals dynamic paracrine control of cellular variation. Nature *510*, 363–369.
6. Grün, D., Lyubimova, A., Kester, L., Wiebrands, K., Basak, O., Sasaki, N., Clevers, H., and van Oudenaarden, A. (2015). Single-cell messenger RNA sequencing reveals rare intestinal cell types. Nature *525*, 251–255.
7. Mahata, B., Zhang, X., Kolodziejczyk, A.A., Proserpio, V., Haim-Vilmovsky, L., Taylor, A.E., Hebenstreit, D., Dingler, F.A., Moignard, V., Göttgens, B., et al. (2014). Single-cell RNA sequencing reveals T helper cells synthesizing steroids de novo to contribute to immune homeostasis. Cell Rep. *7*, 1130–1142.
8. Torre, E., Dueck, H., Shaffer, S., Gospocic, J., Gupte, R., Bonasio, R., Kim, J., Murray, J., and Raj, A. (2018). Rare cell detection by single-cell RNA sequencing as guided by single-molecule RNA FISH. Cell Syst. *6*, 171–179.e5.
9. Zhao, X., Gao, S., Wu, Z., Kajigaya, S., Feng, X., Liu, Q., Townsley, D.M., Cooper, J., Chen, J., Keyvanfar, K., et al. (2017). Single-cell RNA-seq reveals a distinct transcriptome signature of aneuploid hematopoietic cells. Blood *130*, 2762–2773.
10. Treutlein, B., Brownfield, D.G., Wu, A.R., Neff, N.F., Mantalas, G.L., Espinoza, F.H., Desai, T.J., Krasnow, M.A., and Quake, S.R. (2014). Reconstructing lineage hierarchies of the distal lung epithelium using single-cell RNA-seq. Nature *509*, 371–375.
11. Trapnell, C., Cacchiarelli, D., Grimsby, J., Pokharel, P., Li, S., Morse, M., Lennon, N.J., Livak, K.J., Mikkelsen, T.S., and Rinn, J.L. (2014). The dynamics and regulators of cell fate decisions are revealed by pseudotemporal ordering of single cells. Nat. Biotechnol. *32*, 381–386.
12. Qiu, X., Mao, Q., Tang, Y., Wang, L., Chawla, R., Pliner, H.A., and Trapnell, C. (2017). Reversed graph embedding resolves complex single-cell trajectories. Nat. Methods *14*, 979–982.
13. Macosko, E.Z., Basu, A., Satija, R., Nemesh, J., Shekhar, K., Goldman, M., Tirosh, I., Bialas, A.R., Kamitaki, N., Martersteck, E.M., et al. (2015). Highly parallel genome-wide expression profiling of individual cells using nanoliter droplets. Cell *161*, 1202–1214.
14. Kitzman, J.O. (2016). Haplotypes drop by drop. Nat. Biotechnol. *34*, 296–298.
15. Zheng, G.X.Y., Lau, B.T., Schnall-Levin, M., Jarosz, M., Bell, J.M., Hindson, C.M., Kyriazopoulou-Panagiotopoulou, S., Masquelier, D.A., Merrill, L., Terry, J.M., et al. (2016). Haplotyping germline and cancer genomes with high-throughput linked-read sequencing. Nat. Biotechnol. *34*, 303–311.
16. Theilgaard-Mönch, K., Cowland, J., and Borregaard, N. (2001). Profiling of gene expression in individual hematopoietic cells by global mRNA amplification and slot blot analysis. J. Immunol. Methods *252*, 175–189.
17. Dahlin, J.S., Hamey, F.K., Pijuan-Sala, B., Shepherd, M., Lau, W.W.Y., Nestorowa, S., Weinreb, C., Wolock, S., Hannah, R., Diamanti, E., et al. (2018). A single-cell hematopoietic landscape resolves 8 lineage trajectories and defects in Kit mutant mice. Blood *131*, e1–e11.
18. Nguyen, Q.H., Lukowski, S.W., Chiu, H.S., Senabouth, A., Bruxner, T.J.C., Christ, A.N., Palpant, N.J., and Powell, J.E. (2018). Single-cell RNA-seq of human induced pluripotent stem cells reveals cellular heterogeneity and cell state transitions between subpopulations. Genome Res. *28*, 1053–1066.
19. Nguyen, Q.H., Pervolarakis, N., Blake, K., Ma, D., Davis, R.T., James, N., Phung, A.T., Willey, E., Kumar, R., Jabart, E., et al. (2018). Profiling human breast epithelial cells using single cell RNA sequencing identifies cell diversity. Nat. Commun. *9*, 2028.
20. Moon, H.-S., Je, K., Min, J.-W., Park, D., Han, K.-Y., Shin, S.H., Park, W.Y., Yoo, C.E., and Kim, S.H. (2018). Inertial-ordering-assisted droplet microfluidics for high-throughput single-cell RNA-sequencing. Lab Chip *18*, 775–784.
21. Klein, A.M., Mazutis, L., Akartuna, I., Tallapragada, N., Veres, A., Li, V., Peshkin, L., Weitz, D.A., and Kirschner, M.W. (2015). Droplet barcoding for single-cell transcriptomics applied to embryonic stem cells. Cell *161*, 1187–1201.
22. Hanson, W.M., Chen, Z., Jackson, L.K., Attaf, M., Sewell, A.K., Heemstra, J.M., and Phillips, J.D. (2016). Reversible oligonucleotide chain blocking enables bead capture and amplification of T-cell receptor α and β chain mRNAs. J. Am. Chem. Soc. *138*, 11073–11076.
23. Picelli, S., Faridani, O.R., Björklund, A.K., Winberg, G., Sagasser, S., and Sandberg, R. (2014). Full-length RNA-seq from single cells using Smart-seq2. Nat. Protoc. *9*, 171–181.
24. Goetz, J.J., and Trimarchi, J.M. (2012). Transcriptome sequencing of single cells with Smart-Seq. Nat. Biotechnol. *30*, 763–765.
25. Zhu, Y.Y., Machleder, E.M., Chenchik, A., Li, R., and Siebert, P.D. (2001). Reverse transcriptase template switching: a SMART approach for full-length cDNA library construction. Biotechniques *30*, 892–897.
26. 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.
27. Hashimshony, T., Senderovich, N., Avital, G., Klochendler, A., de Leeuw, Y., Anavy, L., Gennert, D., Li, S., Livak, K.J., Rozenblatt-Rosen, O., et al. (2016). CEL-Seq2: sensitive highly-multiplexed single-cell RNA-Seq. Genome Biol. *17*, 77.
28. Ramsköld, D., Luo, S., Wang, Y.-C., Li, R., Deng, Q., Faridani, O.R., Daniels, G.A., Khrebtukova, I., Loring, J.F., Laurent, L.C., et al. (2012). Full-length mRNA-Seq from single-cell levels of RNA and individual circulating tumor cells. Nat. Biotechnol. *30*, 777–782.
29. Liu, N., Jiang, Y., Xing, M., Zhao, B., Hou, J., Lim, M., Huang, J., Luo, X., and Han, L. (2018). Digital gene expression profiling analysis of aged mice under moxibustion treatment. Evid. Based Complement. Alternat. Med. *2018*, 4767328.
30. Islam, S., Kjällquist, U., Moliner, A., Zajac, P., Fan, J.B., Lönnerberg, P., and Linnarsson, S. (2011). Characterization of the single-cell transcriptional landscape by highly multiplex RNA-seq. Genome Res. *21*, 1160–1167.
31. Islam, S., Zeisel, A., Joost, S., La Manno, G., Zajac, P., Kasper, M., Lönnerberg, P., and Linnarsson, S. (2014). Quantitative single-cell RNA-seq with unique molecular identifiers. Nat. Methods *11*, 163–166.
32. Lun, A.T.L., McCarthy, D.J., and Marioni, J.C. (2016). A step-by-step workflow for low-level analysis of single-cell RNA-seq data with Bioconductor. F1000Res. *5*, 2122.

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

34. Zhao, C., Hu, S., Huo, X., and Zhang, Y. (2017). Dr.seq2: a quality control and analysis pipeline for parallel single cell transcriptome and epigenome data. PLoS ONE *12*, e0180583.

35. Haque, A., Engel, J., Teichmann, S.A., and Lönnberg, T. (2017). A practical guide to single-cell RNA-sequencing for biomedical research and clinical applications. Genome Med. *9*, 75.

36. Guo, M., Wang, H., Potter, S.S., Whitsett, J.A., and Xu, Y. (2015). SINCERA: a pipeline for single-cell RNA-seq profiling analysis. PLoS Comput. Biol. *11*, e1004575.

37. Satija, R., Farrell, J.A., Gennert, D., Schier, A.F., and Regev, A. (2015). Spatial reconstruction of single-cell gene expression data. Nat. Biotechnol. *33*, 495–502.

38. Darmanis, S., Sloan, S.A., Croote, D., Mignardi, M., Chernikova, S., Samghababi, P., Zhang, Y., Neff, N., Kowarsky, M., Caneda, C., et al. (2017). Single-cell RNA-seq analysis of infiltrating neoplastic cells at the migrating front of human glioblastoma. Cell Rep. *21*, 1399–1410.

39. Alles, J., Karaiskos, N., Praktiknjo, S.D., Grosswendt, S., Wahle, P., Ruffault, P.L., Ayoub, S., Schreyer, L., Boltengagen, A., Birchmeier, C., et al. (2017). Cell fixation and preservation for droplet-based single-cell transcriptomics. BMC Biol. *15*, 44.

40. Ilicic, T., Kim, J.K., Kolodziejczyk, A.A., Bagger, F.O., McCarthy, D.J., Marioni, J.C., and Teichmann, S.A. (2016). Classification of low quality cells from single-cell RNA-seq data. Genome Biol. *17*, 29.

41. Zhao, Q., Wang, J., Levichkin, I.V., Stasinopoulos, S., Ryan, M.T., and Hoogenraad, N.J. (2002). A mitochondrial specific stress response in mammalian cells. EMBO J. *21*, 4411–4419.

42. Mercer, T.R., Neph, S., Dinger, M.E., Crawford, J., Smith, M.A., Shearwood, A.M., Haugen, E., Bracken, C.P., Rackham, O., Stamatoyannopoulos, J.A., et al. (2011). The human mitochondrial transcriptome. Cell *146*, 645–658.

43. Rizzetto, S., Eltahla, A.A., Lin, P., Bull, R., Lloyd, A.R., Ho, J.W.K., Venturi, V., and Luciani, F. (2017). Impact of sequencing depth and read length on single cell RNA sequencing data of T cells. Sci. Rep. *7*, 12781.

44. Gao, S. (2018). Data analysis in single-cell transcriptome sequencing. Methods Mol. Biol. *1754*, 311–326.

45. Hicks, S.C., Teng, M., and Irizarry, R.A. (2015). On the widespread and critical impact of systematic bias and batch effects in single-cell RNA-seq data. bioRxiv. https://doi.org/10.1101/025528.

46. Jia, C., Hu, Y., Kelly, D., Kim, J., Li, M., and Zhang, N.R. (2017). Accounting for technical noise in differential expression analysis of single-cell RNA sequencing data. Nucleic Acids Res. *45*, 10978–10988.

47. Bacher, R., Chu, L.F., Leng, N., Gasch, A.P., Thomson, J.A., Stewart, R.M., Newton, M., and Kendziorski, C. (2017). SCnorm: robust normalization of single-cell RNA-seq data. Nat. Methods *14*, 584–586.

48. Hicks, S.C., Townes, F.W., Teng, M., and Irizarry, R.A. (2017). Missing data and technical variability in single-cell RNA-sequencing experiments. Biostatistics. , Published online November 6, 2017. https://doi.org/10.1093/biostatistics/kxx053.

49. Barber, R.D., Harmer, D.W., Coleman, R.A., and Clark, B.J. (2005). GAPDH as a housekeeping gene: analysis of GAPDH mRNA expression in a panel of 72 human tissues. Physiol. Genomics *21*, 389–395.

50. Chang, Y.-C., Ding, Y., Dong, L., Zhu, L.-J., Jensen, R.V., and Hsiao, L.-L. (2018). Differential expression patterns of housekeeping genes increase diagnostic and prognostic value in lung cancer. PeerJ *6*, e4719.

51. 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.

52. Hotelling, H. (1933). Analysis of a complex of statistical variables into principal components. J. Educ. Psychol. *24*, 417–441.

53. Sun, Z., Wang, C.-Y., Lawson, D.A., Kwek, S., Velozo, H.G., Owyong, M., Lai, M.D., Fong, L., Wilson, M., Su, H., et al. (2017). Single-cell RNA sequencing reveals gene expression signatures of breast cancer-associated endothelial cells. Oncotarget *9*, 10945–10961.

54. Hu, G., Huang, K., Hu, Y., Du, G., Xue, Z., Zhu, X., and Fan, G. (2016). Single-cell RNA-seq reveals distinct injury responses in different types of DRG sensory neurons. Sci. Rep. *6*, 31851.

55. Wang, M., Lin, F., Xing, K., and Liu, L. (2017). Random X-chromosome inactivation dynamics in vivo by single-cell RNA sequencing. BMC Genomics *18*, 90.

56. Dulken, B.W., Leeman, D.S., Boutet, S.C., Hebestreit, K., and Brunet, A. (2017). Single-cell transcriptomic analysis defines heterogeneity and transcriptional dynamics in the adult neural stem cell lineage. Cell Rep. *18*, 777–790.

57. van der Maaten, L.J.P., and Hinton, G.E. (2008). Visualizing high-dimensional data using t-SNE. J. Mach. Learn. Res. *9*, 2579–2605.

58. Herring, C.A., Chen, B., McKinley, E.T., and Lau, K.S. (2018). Single-cell computational strategies for lineage reconstruction in tissue systems. Cell. Mol. Gastroenterol. Hepatol. *5*, 539–548.

59. Amir, A.D., Davis, K.L., Tadmor, M.D., Simonds, E.F., Levine, J.H., Bendall, S.C., Shenfeld, D.K., Krishnaswamy, S., Nolan, G.P., and Pe'er, D. (2013). viSNE enables visualization of high dimensional single-cell data and reveals phenotypic heterogeneity of leukemia. Nat. Biotechnol. *31*, 545–552.

60. Valihrach, L., Androvic, P., and Kubista, M. (2018). Platforms for single-cell collection and analysis. Int. J. Mol. Sci. *19*, 22–24.

61. Cao, J., Packer, J.S., Ramani, V., Cusanovich, D.A., Huynh, C., Daza, R., Qiu, X., Lee, C., Furlan, S.N., Steemers, F.J., et al. (2017). Comprehensive single-cell transcriptional profiling of a multicellular organism. Science *357*, 661–667.

62. Fluidigm. (2016). Doublet rate and detection on the C1 IFCs. White Paper, http://info.fluidigm.com/rs/673-MRG-416/images/C1-Med-96-IFC-Redesign_wp_101-3328B1_FINAL.pdf.

63. Durruthy-Durruthy, R., and Ray, M. (2018). Using Fluidigm C1 to generate single-cell full-length cDNA libraries for mRNA sequencing. Methods Mol. Biol. *1706*, 199–221.

64. Cochain, C., Vafadarnejad, E., Arampatzi, P., Pelisek, J., Winkels, H., Ley, K., Wolf, D., Saliba, A.E., and Zernecke, A. (2018). Single-cell RNA-seq reveals the transcriptional landscape and heterogeneity of aortic macrophages in murine atherosclerosis. Circ. Res. *122*, 1661–1674.

65. Steuerman, Y., Cohen, M., Peshes-Yaloz, N., Valadarsky, L., Cohn, O., David, E., Frishberg, A., Mayo, L., Bacharach, E., Amit, I., and Gat-Viks, I. (2018). Dissection of influenza infection in vivo by single-cell RNA sequencing. Cell Syst. *6*, 679–691.e4.

66. Rodda, L.B., Lu, E., Bennett, M.L., Sokol, C.L., Wang, X., Luther, S.A., Barres, B.A., Luster, A.D., Ye, C.J., and Cyster, J.G. (2018). Single-cell RNA sequencing of lymph node stromal cells reveals niche-associated heterogeneity. Immunity *48*, 1014–1028.e6.

67. Beliakova-Bethell, N., Massanella, M., White, C., Lada, S., Du, P., Vaida, F., Blanco, J., Spina, C.A., and Woelk, C.H. (2014). The effect of cell subset isolation method on gene expression in leukocytes. Cytometry A *85*, 94–104.

68. Richardson, G.M., Lannigan, J., and Macara, I.G. (2015). Does FACS perturb gene expression? Cytometry A *87*, 166–175.

69. Chen, L., Lee, J.W., Chou, C.-L., Nair, A.V., Battistone, M.A., Păunescu, T.G., Merkulova, M., Breton, S., Verlander, J.W., Wall, S.M., et al. (2017). Transcriptomes of major renal collecting duct cell types in mouse identified by single-cell RNA-seq. Proc. Natl. Acad. Sci. USA *114*, E9989–E9998.

70. Jaff, N., Grankvist, R., Muhl, L., Chireh, A., Sandell, M., Jonsson, S., Arnberg, F., Eriksson, U., and Holmin, S. (2018). Transcriptomic analysis of the harvested endothelial cells in a swine model of mechanical thrombectomy. Neuroradiology *60*, 759–768.

71. Llufrio, E.M., Wang, L., Naser, F.J., and Patti, G.J. (2018). Sorting cells alters their redox state and cellular metabolome. Redox Biol. *16*, 381–387.

72. Vrtačnik, P., Kos, Š., Bustin, S.A., Marc, J., and Ostanek, B. (2014). Influence of trypsinization and alternative procedures for cell preparation before RNA extraction on RNA integrity. Anal. Biochem. *463*, 38–44.

73. Djukelic, M., Wixforth, A., and Westerhausen, C. (2017). Influence of neighboring adherent cells on laminar flow induced shear stress *in vitro*: a systematic study. Biomicrofluidics *11*, 024115.

74. Nathamgari, S.S.P., Dong, B., Zhou, F., Kang, W., Giraldo-Vela, J.P., McGuire, T., McNaughton, R.L., Sun, C., Kessler, J.A., and Espinosa, H.D. (2015). Isolating single cells in a neurosphere assay using inertial microfluidics. Lab Chip *15*, 4591–4597.

75. Senabouth, A., Lukowski, S.W., Alquicira, J., Andersen, S., Mei, X., Nguyen, Q.H., and Powell, J. (2017). ascend: R package for analysis of single cell RNA-seq data. bioRxiv 207704. https://doi.org/10.1101/207704.

76. 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.

77. Ritchie, M.E., Phipson, B., Wu, D., Hu, Y., Law, C.W., Shi, W., and Smyth, G.K. (2015). limma powers differential expression analyses for RNA-sequencing and microarray studies. Nucleic Acids Res. *43*, e47.

78. Zhou, Q., Su, X., Jing, G., Chen, S., and Ning, K. (2018). RNA-QC-chain: comprehensive and fast quality control for RNA-Seq data. BMC Genomics *19*, 144.

79. Brennecke, P., Anders, S., Kim, J.K., Kołodziejczyk, A.A., Zhang, X., Proserpio, V., Baying, B., Benes, V., Teichmann, S.A., Marioni, J.C., and Heisler, M.G. (2013). Accounting for technical noise in single-cell RNA-seq experiments. Nat. Methods *10*, 1093–1095.

80. Robinson, M.D., and Oshlack, A. (2010). A scaling normalization method for differential expression analysis of RNA-seq data. Genome Biol. *11*, R25.

81. Law, C.W., Chen, Y., Shi, W., and Smyth, G.K. (2014). voom: precision weights unlock linear model analysis tools for RNA-seq read counts. Genome Biol. *15*, R29.

82. Zhou, X., Lindsay, H., and Robinson, M.D. (2014). Robustly detecting differential expression in RNA sequencing data using observation weights. Nucleic Acids Res. *42*, e91.

83. Vallejos, C.A., Richardson, S., and Marioni, J.C. (2016). Beyond comparisons of means: understanding changes in gene expression at the single-cell level. Genome Biol. *17*, 70.

84. Katayama, S., Töhönen, V., Linnarsson, S., and Kere, J. (2013). SAMstrt: statistical test for differential expression in single-cell transcriptome with spike-in normalization. Bioinformatics *29*, 2943–2945.

85. Yip, S.H., Wang, P., Kocher, J.A., Sham, P.C., and Wang, J. (2017). Linnorm: improved statistical analysis for single cell RNA-seq expression data. Nucleic Acids Res. *45*, e179.

86. Iacono, G., Mereu, E., Guillaumet-Adkins, A., Corominas, R., Cuscó, I., Rodríguez-Esteban, G., Gut, M., Pérez-Jurado, L.A., Gut, I., and Heyn, H. (2017). bigSCale: an analytical framework for big-scale single-cell data. Genome Res. *28*, 878–890.

87. Yip, S.H., Sham, P.C., and Wang, J. (2018). Evaluation of tools for highly variable gene discovery from single-cell RNA-seq data. Brief. Bioinform. bby011. Published online February 21, 2018. https://doi.org/10.1093/bib/bby011.

88. Welch, J.D., Hu, Y., and Prins, J.F. (2016). Robust detection of alternative splicing in a population of single cells. Nucleic Acids Res. *44*, e73.

89. Ross, E.M., and Markowetz, F. (2016). OncoNEM: inferring tumor evolution from single-cell sequencing data. Genome Biol. *17*, 69.