Molecules and Cells

MiniResource

A practical handbook on single-cell RNA sequencing data quality control and downstream analysis

Gyeong Dae Kim†, Chaemin Lim†, and Jihwan Park*

School of Life Sciences, Gwangju Institute of Science and Technology (GIST), Gwangju 61005, Republic of Korea
*Corresponding author. jihwan.park@gist.ac.kr
https://doi.org/10.1016/j.mocell.2024.100103

ABSTRACT

Advancements in single-cell analysis have facilitated high-resolution observation of the transcriptome in individual cells. However, standards for obtaining high-quality cells and data analysis pipelines remain variable. Here, we provide the groundwork for improving the quality of single-cell analysis by delineating guidelines for selecting high-quality cells and considerations throughout the analysis. This review will streamline researchers' access to single-cell analysis and serve as a valuable guide for analysis.

Keywords: Downstream analysis, Quality control, Single-cell RNA sequencing

INTRODUCTION

The emergence of single-cell technology has revolutionized biomedical science, enabling comprehensive exploration of cellular heterogeneity, individual cell characteristics, and cell lineage trajectory (Kim and Park, 2021; Yoon et al., 2024). Despite these advancements, challenges persist due to batch effects arising from variations such as tissue storage, dissociation processes, and sequencing library preparation, leading to inconsistent results (Lee et al., 2021). Moreover, inherent technical hurdles yield highly sparse data, complicating result interpretation (Choi and Kim, 2019). In response, various computational tools and quality control measures have been proposed, yet the absence of consensus guidelines poses a significant challenge in applying these tools to highly variable experimental conditions and samples.

Additionally, downstream analysis involves several time-consuming steps, each demanding careful evaluation for result appropriateness. While various tools are available for these downstream analyses, including batch correction, dimension reduction, clustering, and cell-type annotation, those with limited experience in single-cell analysis still find it challenging to determine the appropriate tool for specific circumstances and conditions.

Beyond the studies benchmarking and explaining each step in quality control and downstream analysis, our objective is to address practical challenges by offering a comprehensive guideline for quality control and each stage of downstream analysis (Fig. 1). In particular, we focused on barcode-based single-cell RNA sequencing (scRNA-seq) techniques that are widely used, including droplet-, microwell-, and combinatorial barcoding-based methods. Consequently, we aim to enhance the reliability and reproducibility of commonly employed single-cell studies.

MAIN BODY

Considerations in Transcripts Quality Control

To ensure the reliability and quality of the analysis results, it is crucial to address artifact transcripts like ambient RNAs. For example, transcripts from damaged or apoptotic cells may leak out from cells during single-cell isolation, exist in the solution, and then potentially become encapsulated in droplets along with other cells. Besides these ambient RNAs, contamination between transcripts may arise by evaporation in plate-based protocols and from chimeric complementary DNA being called “barcode swap” due to incorrect binding between barcodes during sequencing (Maxwell et al., 2023; Wagener and Plennevaux, 2014; Yang et al., 2020). These transcripts complicate cell-type annotation by contaminating endogenous gene expression profiling and lead differences by ambient profiles rather than true biological differences. Hence, we should consider removing genes as artifact RNA in the following cases: (1) detection of cell-type–specific markers from other cell types, particularly those derived from cells with a higher proportion in the given tissue; (2) genes from cells displaying elevated levels of mitochondrial genes. Given that these cells, expressing high mitochondrial genes, are likely dead or dying, the transcripts may include RNAs originating from cell-free sources.

To remove ambient RNA contamination, several tools were developed. SoupX does not depend significantly on

† These authors contributed equally to this work as the first authors.

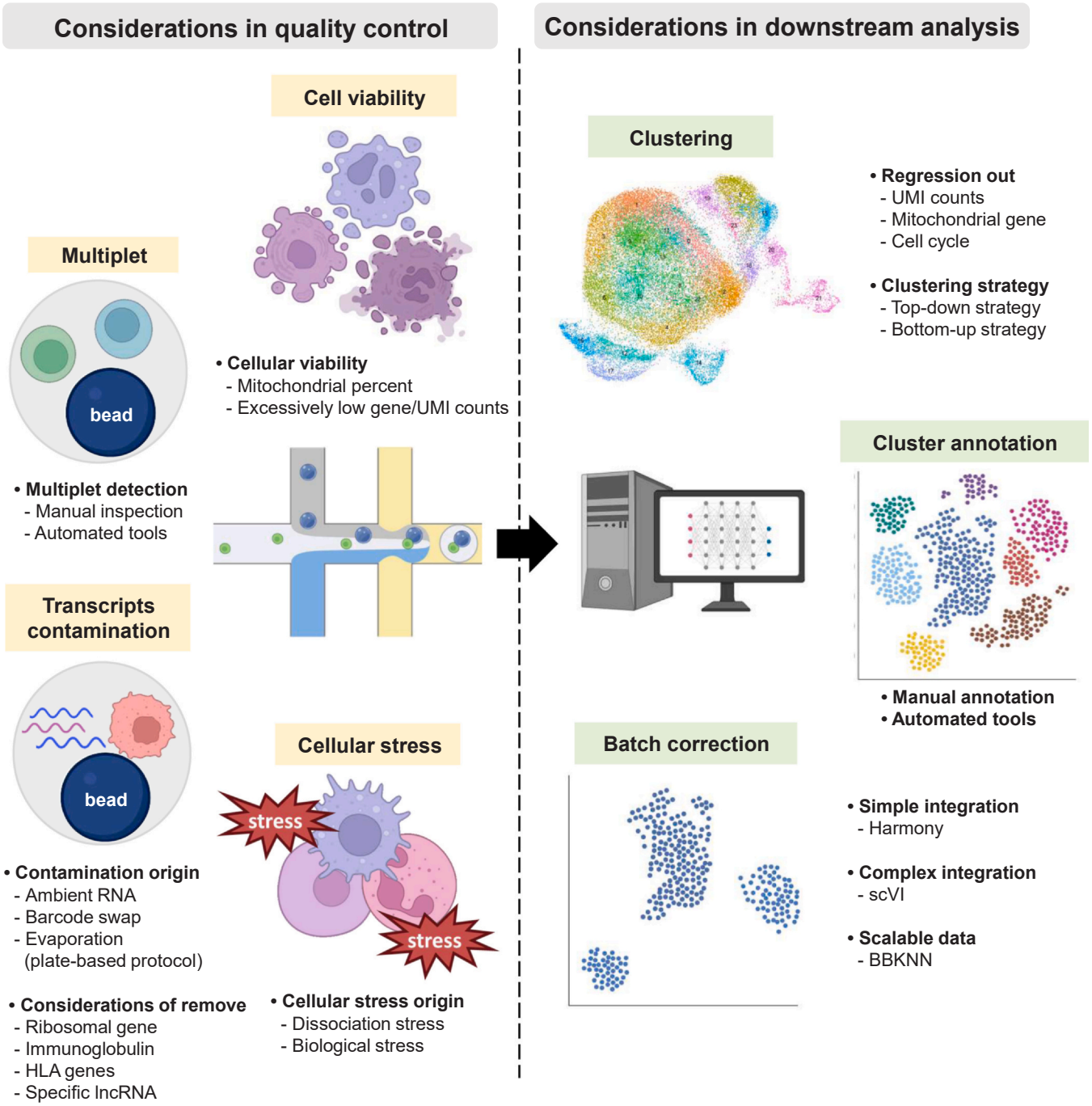

Fig. 1. Factors requiring quality control in single-cell RNA sequencing analysis and considerations in each factor (left). Stages of downstream analysis after quality control and considerations at each stage (right). lncRNA, long non-coding RNA.

precise pre-annotation, but it does require prior knowledge of the user's marker genes due to the necessity of manual input (Young and Behjati, 2020). Notably, SoupX performs much better with single-nucleus compared to single-cell data (Janssen et al., 2023). CellBender is suited for cleaning up and extracting the biological signal from noisy datasets, providing the most accurate estimation of background noise when compared to other tools (Fleming et al., 2023; Janssen et al., 2023).

Apart from ambient RNAs, specific studies have considered ribosomal genes, immunoglobulin genes, human leukocyte antigens genes, and specific long non-coding RNA (eg, metastasis-associated lung adenocarcinoma transcript 1) as elements that should be filtered out. This is because they can induce unwanted batch effects in downstream clustering steps due to their overabundant expression and uncertain origination from various cell types (Gharaie et al., 2023; Kunes et al., 2024; Smillie et al., 2019).

In addition, genes or cells associated with stress signatures are considered for removal since they can reduce the reliability of results. Stress-related genes are induced by various factors, such as sample storage and dissociation, and the values may vary following cellular structure and characteristics even in the same sample. To identify stress signatures, approximately 200 dissociation-related genes or stress-related genes have been suggested (Romanov et al., 2020; van den Brink et al., 2017). However, it is crucial to cautiously approach their removal, as stress-related gene expression can reflect biological response and disease status.

Considerations in Cellular Quality Control

A doublet or multiplet, where more than 1 cell is captured within a single droplet or microwell, arose as a technical artifact during the scRNA-seq library preparation process. The multiplet rate is influenced by the scRNA-seq platform and the number of loaded cells (Depasquale et al., 2019; Nguyen et al., 2018). For instance, 10x Genomics, which utilizes a droplet-based platform, reported that when 7,000 target cells are loaded, 378 multiplets are identified, constituting 5.4% of the total cells (10X_Genomics, 2022). Notably, this rate escalates to 7.6% when the number of target cells is increased to 10,000. In contrast, the BD rhapsody platform, which is based on a microwell-based system, exhibits significantly lower multiplet rates compared to 10x Genomics by inspecting multiplets through automated microscopy.

Several methods have been developed to filter out doublets, each employing distinct algorithmic approaches and offering unique advantages. Notably, Scrublet demonstrates scalability, enabling analysis of large datasets, while doubletCells exhibits strong statistical stability across varying cell and gene numbers (Lun et al., 2016b; Wolock et al., 2019; Xi and Li, 2021). In terms of accuracy and impact on downstream analyses like differential gene expression, clustering, and trajectory inference, DoubletFinder outperforms the other doublet-detection methods suggested in this paper (McGinnis et al., 2019; Xi and Li, 2021).

While these multiplet removal tools are useful, even the method with the highest multiplet-detection accuracy was relatively low at 0.537, and they exhibit substantial variation across different datasets (Xi and Li, 2021). Therefore, it is recommended to employ an appropriate combination of automated tools and manual inspection to account for the complexity of the conditions and samples. Cells co-expressing well-known markers of distinct cell types require careful scrutiny. In some instances, such co-expressing cells have been identified as representing transitional states (Park et al., 2018). However, other studies have opted to remove co-expressing cells due to concerns about doublets (Karademir et al., 2022).

After removing transcript contamination and multiplets, additional filtering is recommended to exclude cells with excessively high or low gene/unique molecular identifier (UMI) counts. High counts may indicate multiplet artifacts, whereas low counts indicate potential low-quality cells (Kim et al., 2022; Park et al., 2018).

Additionally, cells with a mitochondrial percentage exceeding 5% to 15% were excluded as considered low-quality cells (Luo et al., 2021; Sikkema et al., 2023). However, the criteria for removing cells based on mitochondrial percent can vary depending on factors such as species, sample types, and experimental conditions (Osorio and Cai, 2020; Subramanian et al., 2022). For instance, human samples often exhibit a higher percentage of mitochondrial genes compared to mice, and highly metabolically active tissues like kidneys may display robust expression of mitochondrial genes (Osorio and Cai, 2020; Uhlén et al., 2015).

Strategies and Considerations in scRNA-seq Analysis

After quality control, several important considerations arise in the analysis pipeline. Typically, factors such as total UMIs per cell, mitochondrial gene percentage, and stress signatures can be selected for regression out during scaling analysis to address unwanted technical and biological variations derived from sequencing depth and cellular stress (Hafemeister and Satija, 2019). Furthermore, the cell cycle score is regarded as a confounding factor and regressed out to mitigate the effects of cell cycle heterogeneity (Luecken and Theis, 2019).

Dimensional reduction is performed to extract biological signals from the data, which requires users' decision to set a threshold. Recently, an unbiased scRNA-seq data analysis method, single-cell low-dimension embedding using effective noise subtraction, was developed, which reduces signal distortion and detects biological signals without manual tuning (Kim et al., 2024b). Moreover, determining the optimal resolution value for cell clustering is challenging, as it heavily relies on the unique characteristics of each dataset, the research purpose, and the specific cell types of interest to researchers. Hence, the following 2 types of strategies are recommended for determining clusters: The first approach employs a top-down strategy, classifying cells into the minimum number of main cell types and then further subclustering each main cell type. The second approach utilizes a bottom-up strategy, classifying cells into a large number of initial clusters and then merging clusters if a pair of clusters exhibits fewer than a certain number of differentially expressed genes (eg, 10 genes) (Kim et al., 2024a).

When integrating multiple datasets for unified analysis, identifying batch effects is crucial. Batch effects stem from technical and experimental variations rather than biological differences, potentially causing clusters to appear as distinct cell types even when they are actually the same. A recent paper benchmarked batch correction methods and indicated that their performance varies depending on the scalability, complexity, and availability of cell annotations within the dataset (Luecken et al., 2022). For example, Harmony is a valuable option for simple integration tasks involving distinct batch and biological structures (Korsunsky et al., 2019). However, for more complex integration tasks such as tissue or organ atlases, tools like single-cell variational inference are more suitable (Lopez et al., 2018). Additionally, BBKNN (batch balanced k nearest neighbours) has demonstrated excellent performance in handling scalable data concerning runtime and memory efficiency (Polański et al., 2020). While batch correction methods offer substantial robustness in mitigating unwanted variation, it is crucial to acknowledge that their application may not be universally effective. For example, in heterogeneous samples such as tumors or cases involving biologically meaningful differences in experimental conditions, improper correction of heterogeneity

could lead to unintended biases in the data analysis (Wu et al., 2021). Hence, it is strongly recommended to implement batch correction with careful consideration of the specific context and utmost caution.

Even after applying batch correction, uncertain clusters often remain. For instance, cell clusters of the same cell type may be segregated based on the total number of UMIs. This segregation can result from biological variance or technical bias, and thus stably expressed genes are utilized to identify the source of this segregation (Lin et al., 2019). Differential expression of stably expressed genes across cells, correlating with variations in UMI depth, suggests a technical effect, potentially due to pooling inefficiencies. To minimize technical effects induced by pooling, alternative normalization methods are employed (Lun et al., 2016a).

Cell-type annotation of clusters is typically performed manually, relying on established marker gene expression profiles within each cluster. However, this approach requires expertise and is often time-consuming (Pasquini et al., 2021). Furthermore, annotating cell types can be particularly challenging in 3 scenarios: (1) accurate annotation of immune cell types often requires both positive and negative markers (Ianevski et al., 2022). (2) Annotating novel cell types can be difficult when distinct marker genes are lacking. (3) Distinguishing between multiple subcell types that exhibit similar expression patterns of known marker genes is also challenging. In these cases, relying solely on specific marker genes may be insufficient for accurate annotation. Therefore, it is strongly recommended to adopt a combined strategy utilizing both manual expertise and automated annotation tools. Various automated cell-type annotation methods have been developed based on marker gene databases, correlation analysis, and supervised classification; further details are described in this benchmarking paper (Pasquini et al., 2021). Additionally, a Generative Pre-trained Transformers based approach has emerged, demonstrating high accuracy, low laboriousness, and consistency (Hou and Ji, 2024).

Collectively, this review offers useful and practical guidelines for quality control at each stage of analysis. We anticipate that this work will enhance the reliability and reproducibility of single-cell studies.

FUNDING AND SUPPORT

This work was supported GIST-CNUH Research Collaboration grant and GIST-MIT Research collaboration grant funded by the GIST in 2024, and the National Research Foundation of Korea (NRF), funded by the Korean government (RS-2024-00335026).

AUTHOR CONTRIBUTIONS

G.D.K., C.L., and J.P. wrote the manuscript. All authors critically evaluated and approved the manuscript.

DECLARATION OF COMPETING INTERESTS

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

ACKNOWLEDGMENTS

We thank all of Dr. Park's laboratory members for their helpful discussion and comments.

ORCID

Gyeong Dae Kim: https://orcid.org/0000-0002-9288-6341
Chaemin Lim: https://orcid.org/0009-0001-3959-8738
Jihwan Park: https://orcid.org/0000-0002-5728-912X

Received May 1, 2024
Revised July 19, 2024
Accepted July 29, 2024
Available online 31 July 2024.

REFERENCES

10X_GENOMICS. (2022). Chromium Single Cell 3′ Reagent Kits User Guide (v3.1 Chemistry) (Online). https://www.10xgenomics.com/support/single-cell-gene-expression/documentation/steps/library-prep/chromium-single-cell-3-reagent-kits-user-guide-v-3-1-chemistry.

Choi, Y.H., and Kim, J.K. (2019). Dissecting cellular heterogeneity using single-cell RNA sequencing. Mol. Cells, *42*, 189-199.

Depasquale, E.A., Schnell, D.J., VAN Camp, P.-J., Valiente-Alandï, Í., Blaxall, B.C., Grimes, H.L., Singh, H., and Salomonis, N. (2019). DoubletDecon: deconvoluting doublets from single-cell RNA-sequencing data. Cell Rep. *29*, 1718-1727.e8.

Fleming, S.J., Chaffin, M.D., Arduini, A., Akkad, A.-D., Banks, E., Marioni, J.C., Philippakis, A.A., Ellinor, P.T., and Babadi, M. (2023). Unsupervised removal of systematic background noise from droplet-based single-cell experiments using CellBender. Nat. Methods, *20*, 1323-1335.

Gharaie, S., Lee, K., Noller, K., Lo, E.K., Miller, B., Jung, H.J., Newman-Rivera, A.M., Kurzhagen, J.T., Singla, N., Welling, P.A., et al. (2023). Single cell and spatial transcriptomics analysis of kidney double negative T lymphocytes in normal and ischemic mouse kidneys. Sci. Rep. *13*, 20888.

Hafemeister, C., and Satija, R. (2019). Normalization and variance stabilization of single-cell RNA-seq data using regularized negative binomial regression. Genome Biol. *20*, 296.

Hou, W., and Ji, Z. (2024). Assessing GPT-4 for cell type annotation in single-cell RNA-seq analysis. Nat. Methods, *21*, 1462-1465.

Ianevski, A., Giri, A.K., and Aittokallio, T. (2022). Fully-automated and ultra-fast cell-type identification using specific marker combinations from single-cell transcriptomic data. Nat. Commun. *13*, 1246.

Janssen, P., Kliesmete, Z., Vieth, B., Adiconis, X., Simmons, S., Marshall, J., Mccabe, C., Heyn, H., Levin, J.Z., Enard, W., et al. (2023). The effect of background noise and its removal on the analysis of single-cell expression data. Genome Biol. *24*, 140.

Karademir, D., Todorova, V., Ebner, L.J.A., Samardzija, M., and Grimm, C. (2022). Single-cell RNA sequencing of the retina in a model of retinitis pigmentosa reveals early responses to degeneration in rods and cones. BMC Biol. *20*, 86.

Kim, G.D., Shin, S.-I., Jung, S.W., An, H., Choi, S.Y., Eun, M., Jun, C.-D., Lee, S., and Park, J. (2024a). Cell type– and age-specific expression of lncRNAs across Kidney cell types. J. Am. Soc. Nephrol. *35*.

Kim, H., Chang, W., Chae, S.J., Park, J.-E., Seo, M., and Kim, J.K. (2024b). scLENS: data-driven signal detection for unbiased scRNA-seq data analysis. Nat. Commun. *15*, 3575.

Kim, J., and Park, J. (2021). Single-cell transcriptomics: a novel precision medicine technique in nephrology. Korean J. Internal Med. *36*, 479.

Kim, J.W., Nam, S.A., Yi, J., Kim, J.Y., Lee, J.Y., Park, S.Y., Sen, T., Choi, Y.M., Lee, J.Y., and Kim, H.L. (2022). Kidney decellularized extracellular matrix enhanced the vascularization and maturation of human kidney organoids. Adv. Sci. *9*, 2103526.

Korsunsky, I., Millard, N., Fan, J., Slowikowski, K., Zhang, F., Wei, K., Baglaenko, Y., Brenner, M., Loh, P.-R., and Raychaudhuri, S. (2019). Fast, sensitive and accurate integration of single-cell data with Harmony. Nat. Methods, *16*, 1289-1296.

Kunes, R.Z., Walle, T., Land, M., Nawy, T., and Pe'er, D. (2024). Supervised discovery of interpretable gene programs from single-cell data. Nat. Biotechnol. *42*, 1084-1095.

Lee, S., Kim, J., and Park, J.-E. (2021). Single-cell toolkits opening a new era for cell engineering. Mol. Cells, *44*, 127-135.

Lin, Y., Ghazanfar, S., Strbenac, D., Wang, A., Patrick, E., Lin, D.M., Speed, T., Yang, J.Y.H., and Yang, P. (2019). Evaluating stably expressed genes in single cells. Gigascience, *8*, Article giz106.

Lopez, R., Regier, J., Cole, M.B., Jordan, M.I., and Yosef, N. (2018). Deep generative modeling for single-cell transcriptomics. Nat. Methods, *15*, 1053-1058.

Luecken, M.D., Bþttner, M., Chaichoompu, K., Danese, A., Interlandi, M., Mueller, M.F., Strobl, D.C., Zappia, L., Dugas, M., Colomë-Tatchë, M., et al. (2022). Benchmarking atlas-level data integration in single-cell genomics. Nat. Methods, *19*, 41-50.

Luecken, M.D., and Theis, F.J. (2019). Current best practices in single-cell RNA-seq analysis: a tutorial. Mol. Syst. Biol. *15*, Article e8746.

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.

Lun, A.T., 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.

Luo, H., Xia, X., Kim, G.D., Liu, Y., Xue, Z., Zhang, L., Shu, Y., Yang, T., Chen, Y., and Zhang, S. (2021). Characterizing dedifferentiation of thyroid cancer by integrated analysis. Sci. Adv. *7*, eabf3657.

Maxwell, C.B., Sandhu, J.K., Cao, T.H., Mccann, G.P., Ng, L.L., and Jones, D.J. (2023). The edge effect in high-throughput proteomics: a cautionary tale. J. Am. Soc. Mass Spectrom. *34*, 1065-1072.

McGinnis, C.S., Murrow, L.M., and Gartner, Z.J. (2019). DoubletFinder: doublet detection in single-cell RNA sequencing data using artificial nearest neighbors. Cell Syst. *8*, 329-337.e4.

Nguyen, A., Khoo, W.H., Moran, I., Croucher, P.I., and Phan, T.G. (2018). Single cell RNA sequencing of rare immune cell populations. Front. Immunol. *9*, 1553.

Osorio, D., and Cai, J.J. (2020). Systematic determination of the mitochondrial proportion in human and mice tissues for single-cell RNA-sequencing data quality control. Bioinformatics, *37*, 963-967.

Park, J., Shrestha, R., Qiu, C., Kondo, A., Huang, S., Werth, M., Li, M., Barasch, J., and Susztãk, K. (2018). Single-cell transcriptomics of the mouse kidney reveals potential cellular targets of kidney disease. Science, *360*, 758-763.

Pasquini, G., Arias, J.E.R., Schæfer, P., and Busskamp, V. (2021). Automated methods for cell type annotation on scRNA-seq data. Comput. Struct. Biotechnol. J. *19*, 961-969.

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

Romanov, R.A., Tretiakov, E.O., Kastriti, M.E., Zupancic, M., Hæring, M., Korchynska, S., Popadin, K., Benevento, M., Rebernik, P., Lallemend, F., et al. (2020). Molecular design of hypothalamus development. Nature, *582*, 246-252.

Sikkema, L., Ramïrez-Suãstegui, C., Strobl, D.C., Gillett, T.E., Zappia, L., Madissoon, E., Markov, N.S., Zaragosi, L.-E., Ji, Y., and Ansari, M. (2023). An integrated cell atlas of the lung in health and disease. Nat. Med. *29*, 1563-1577.

Smillie, C.S., Biton, M., Ordovas-Montanes, J., Sullivan, K.M., Burgin, G., Graham, D.B., Herbst, R.H., Rogel, N., Slyper, M., Waldman, J., et al. (2019). Intra- and inter-cellular rewiring of the human colon during ulcerative colitis. Cell, *178*, 714-730.e22.

Subramanian, A., Alperovich, M., Yang, Y., and Li, B. (2022). Biology-inspired data-driven quality control for scientific discovery in single-cell transcriptomics. Genome Biol. *23*, 267.

Uhlën, M., Fagerberg, L., Hallstrøm, B.M., Lindskog, C., Oksvold, P., Mardinoglu, A., Sivertsson, Å., Kampf, C., Sjøstedt, E., and Asplund, A. (2015). Tissue-based map of the human proteome. Science, *347*, 1260419.

van den Brink, S.C., Sage, F., Vërtesy, Á., Spanjaard, B., Peterson-Maduro, J., Baron, C.S., Robin, C., and Van Oudenaarden, A. (2017). Single-cell sequencing reveals dissociation-induced gene expression in tissue subpopulations. Nat. Methods, *14*, 935-936.

Wagener, J., and Plennevaux, C. (2014). Eppendorf 96-well cell culture plate–a simple method of minimizing the edge effect in cell-based assays. Eppendorf Appl. Note, 326.

Wolock, S.L., Lopez, R., and Klein, A.M. (2019). Scrublet: computational identification of cell doublets in single-cell transcriptomic data. Cell Syst. *8*, 281-291.e9.

Wu, F., Fan, J., He, Y., Xiong, A., Yu, J., Li, Y., Zhang, Y., Zhao, W., Zhou, F., and Li, W. (2021). Single-cell profiling of tumor heterogeneity and the microenvironment in advanced non-small cell lung cancer. Nat. Commun. *12*, 2540.

Xi, N.M., and Li, J.J. (2021). Benchmarking computational doublet-detection methods for single-cell RNA sequencing data. Cell Syst. *12*, 176-194.e6.

Yang, S., Corbett, S.E., Koga, Y., Wang, Z., Johnson, W.E., Yajima, M., and Campbell, J.D. (2020). Decontamination of ambient RNA in single-cell RNA-seq with DecontX. Genome Biol. *21*, 57.

Yoon, B., Kim, H., Jung, S.W., and Park, J. (2024). Single-cell lineage tracing approaches to track kidney cell development and maintenance. Kidney Int. *105*, 1186-1199.

Young, M.D., and Behjati, S. (2020). SoupX removes ambient RNA contamination from droplet-based single-cell RNA sequencing data. Gigascience, *9*, giaa151.