Deep learning tackles single-cell analysis - A survey of deep learning for scRNA-seq analysis

Mario Flores1§, Zhentao Liu1, Tinghe Zhang1, Md Musaddaqui Hasib1, Yu-Chiao C

2021-11-23

Contents

–>

test text justification s

```
## Warning: package 'gtsummary' was built under R version 4.1.1

## #BlackLivesMatter

## Warning: package 'survival' was built under R version 4.1.1

##
## Attaching package: 'dplyr'

## The following objects are masked from 'package:stats':
##
##     filter, lag

## The following objects are masked from 'package:base':
##
##     intersect, setdiff, setequal, union

## Warning: package 'knitr' was built under R version 4.1.1

## Warning: package 'kableExtra' was built under R version 4.1.1
```

```
##
## Attaching package: 'kableExtra'
## The following object is masked from 'package:dplyr':
##
##     group_rows
## Warning: package 'formattable' was built under R version 4.1.1
```

About the authors

§ Corresponding authors

Mario Flores (mario.flores@utsa.edu)

Yidong Chen (cheny8@uthscsa.edu)

Yufei Huang (yuh119@pitt.edu)

Affiliations

1Department of Electrical and Computer Engineering, the University of Texas at San Antonio, San Antonio, TX 78249, USA

2Greehey Children's Cancer Research Institute, University of Texas Health San Antonio, San Antonio, TX 78229, USA

3Department of Population Health Sciences, University of Texas Health San Antonio, San Antonio, TX 78229, USA

4Department of Microbiology and Molecular Genetics, University of Pittsburgh, Pittsburgh, Pennsylvania, PA 15232, USA

5Department of Medicine, School of Medicine, University of Pittsburgh, PA 15232, USA

6UPMC Hillman Cancer Center, University of Pittsburgh, PA 15232, USA

Book Maintainer

Hello there :D

Any feedback and contributions will be appreciated. Please email me if you have any!

Mail: sumin.jo@utsa.edu

Website: Huang-AI4Medicine-Lab

About this book

This book is full version of our research paper [**Please add paper publish information here**].

Keywords deep learning; single-cell RNA-seq; imputation; dimention reduction; clustering; batch correction; cell type identification; functional prediction; visualization

Abstract

Since its selection as the method of the year in 2013, single-cell technologies have become mature enough to provide answers to complex research questions. However, together with the growth of single-cell profiling technologies, there has also been an increase of computational challenges to process the generated datasets. It's here that by effectively leveraging large data sets, Deep Learning (DL) is positioning as the first option for single-cell analyses. Here we provide a unified mathematical description of the DL methods used in single cell RNA sequencing (scRNA-Seq) followed with the survey of the most representative published DL algorithms for scRNA-Seq in the field.

Key Points

- Single cell RNA sequencing technology generate large collection of transcriptomic profiles of up to millions of cells, enabling biological investigation of hidden structures or cell types, predicting their effects or responses to treatment more precisely, or utilizing sub-population to address unanswered hypotheses.
- Current Deep Learning-based analysis approaches for single cell RNA seq data is systematically reviewed in this paper according to the challenge they address and their roles in the analysis pipeline.
- A unified mathematical description of the surveyed DL models is presented and the specific model features were discussed when reviewing each approach.

- A comprehensive summary of the evaluation metrics, comparison algorithms, and datasets by each approaches is presented.

Chapter 1

Introduction

Single cell sequencing technology has been a rapidly developing area to study genomics, transcriptomics, proteomics, metabolomics, and cellular interactions at the single cell level for cell-type identification, tissue composition and reprogramming (Lahnemann et al. 2020; Vitak et al. 2017) . Specifically, sequencing of the transcriptome of single cells, or single-cell RNA-sequencing (scRNA-seq), has become the dominant technology in many frontier research areas such as disease progression and drug discovery (Wolock, Lopez, and Klein 2019; Bost et al. 2020) . One particular area where scRNA-seq has made a tangible impact is cancer, where scRNA-seq is becoming a powerful tool for understanding invasion, intratumor heterogeneity, metastasis, epigenetic alterations, detecting rare cancer stem cells, and therapeutic response (refs). Currently, scRNA-seq is applied to develop personalized therapeutic strategies that are potentially useful in cancer diagnosis, therapy resistance during cancer progression, and the survival of patients (Kinker et al. 2020; Navin 2015). The scRNA-seq has also been adopted to combat COVID-19 to elucidate how the innate and adaptive host immune system miscommunicates resulting in worsening the immunopathology produced during this viral infection (Mannarapu, Dariya, and Bandapalli 2021; Wauters et al. 2021).

These studies have led to a massive amount of scRNA-seq data deposited to public databases such as 10X Single-cell gene expression dataset, Human Cell Atlas, and Mouse Cell Atlas. Expressions of millions of cells from 18 species have been collected and deposited, waiting for further analysis. On the other hand, due to biological and technical factors, scRNA-seq data presents several analytical challenges related to its complex characteristics like missing expression values, high technical and biological variance, noise and sparse gene coverage, and elusive cell identities (Lahnemann et al. 2020) . These characteristics make it difficult to directly apply commonly used bulk RNA-seq data analysis techniques and have called for novel statistical approaches for scRNA-seq data cleaning and computational algorithms for data analysis and interpretation. To this end, specialized scRNA-seq analysis pipelines such as Seurat (Stuart et al. 2019) and Scanpy (Wolf, Angerer, and Theis 2018). along with a large collection of task-specific tools, have been developed to address the intricate technical and biological complexity of scRNA-seq data.

Recently, deep learning has demonstrated its significant advantages in natural language processing and speech and facial recognition with massive data (Srinivasan et al. 2020; Amodio et al. 2019; Lopez et al. 2018). Such advantages have initiated the application of DL in scRNA-seq data analysis as a competitive alternative to conventional machine learning approaches for uncovering cell clustering (Amodio et al. 2019; Eraslan et al. 2019) , cell type identification (Amodio et al. 2019; Xu et al. 2020), gene imputation (Arisdakessian et al. 2019; Tran et al. 2020; Petegrosso, Li, and Kuang 2020) , and batch correction (Abdelaal et al. 2019) in scRNA-seq analysis. Compared to conventional machine learning (ML) approaches, DL is more powerful in capturing complex features of high-dimensional scRNA-seq data. It is also more versatile , where a single model can be trained to address multiple tasks or adapted and transferred to different tasks. Moreover, the DL training scales more favorably with the number of cells in scRNA-seq data size, making it particularly attractive for handling the ever-increasing volume of single cell data. Indeed, the growing body of DL-based tools has demonstrated DL's exciting potential as a learning paradigm to significantly advance the tools we use to interrogate scRNA-seq data.

In this paper, we present a comprehensive review of the recent advances of DL methods for solving the present challenges in scRNA-seq data analysis (Table**??**) from the quality control, normalization/batch effect reduction, dimension reduction, visualization, feature selection, and data interpretation by surveying deep learning papers published up to April 2021. In order to maintain high quality for this review, we choose not to include any (bio)archival papers, although a proportion of these manuscripts contain important new findings that would be published after completing their peer-reviewed process. Previous efforts to review the recent advances in machine learning methods focused on efficient integration of single cell data (Picelli et al. 2013; Macosko et al. 2015) . A recent review of DL applications on single cell data has summarized 21 DL algorithms that might be deployed in single cell studies (Chen, Ning, and Shi 2019). It also evaluated the clustering and data correction effect of these DL algorithms using 11 datasets.

In this review, we focus more on the DL algorithms with a much detailed explanation and comparison. Further, to better understand the relationship of each surveyed DL model with the overall scRNA-seq analysis pipeline, we organize the surveys according to the challenge they address and discuss these DL models following the analysis pipeline. A unified mathematical description of the surveyed DL models is presented and the specific model features are discussed when reviewing each method. This will also shed light on the modeling connections among the surveyed DL methods and the recognization of the uniqueness of each model. Besides the models, we also summarize the evaluation matrics of these DL algorithms and compare the tools that integrate these DL algorithms. Access to these DL algorithms with the original research results, available datasets used by these methods are also listed to demonstrate the advantages and utility of the DL algorithms. We envision that this survey will serve as an important information portal for learning the application of DL for scRNA-seq analysis and inspire innovative use of DL to address a broader range of new challenges in emerging multi-omics and spatial single-cell sequencing.

Chapter 2

Overview of scRNA-seq processing pipeline

Various scRNA-seq techniques (like SMART-seq, Drop-seq, and 10X genomics sequencing (Eisenstein 2020; Vallejos, Marioni, and Richardson 2015) are available nowadays with their sets of advantages and disadvantages. Despite the differences in the scRNA-seq techniques, the data content and processing steps of scRNA-seq data are quite standard and conventional. A typical scRNA-seq dataset consists of three files: genes quantified (gene IDs), cells quantified (cellular barcode), and a count matrix (number of cells x number of genes), irrespective of the technology or pipeline used. A series of essential steps in scRNA-seq data processing pipeline and optional tools for each step with both ML and DL approaches are illustrated in Fig.2.1.

With the advantage of identifying each cell and unique molecular identifiers (UMIs) for expressions of each gene in a single cell, scRNA-seq data are embedded with increased technical noise and biases [23]. **Quality control (QC)** is the first and the key step to filter out dead cells, double-cells, or cells with failed chemistry or other technical artifacts. The most commonly adopted three QC covariates include the number of counts (count depth) per barcode identifying each cell, the number of genes per barcode, and the fraction of counts from mitochondrial genes per barcode (Eisenstein 2020).

Normalization is designed to eliminate imbalanced sampling, cell differentiation, viability, and many other factors. Approaches tailored for scRNA-seq have been developed including the Bayesian-based method coupled with spike-in, or BASiCS (Vallejos, Marioni, and Richardson 2015), deconvolution approach, scran (Lun, Bach, and Marioni 2016), and sctransfrom in Seurat where regularized Negative Binomial regression was proposed (Hafemeister and Satija 2019). Two important steps, batch correction and imputation, will be carried out if required by the analysis:

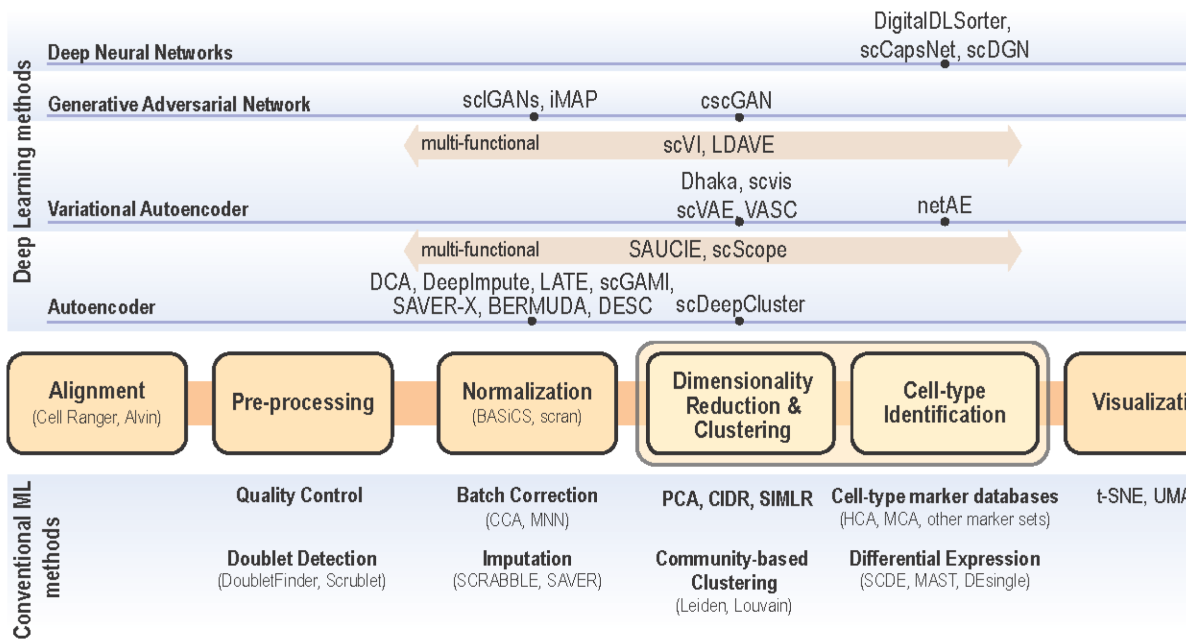

Figure 2.1: Single cell data analysis steps for both conventional ML methods (bottom) and DL methods (top). Depending on the input data and analysis objectives, major scRNA-se analysis steps are illustrated in the center flow chart. The convential ML approaches along with optional analysis modules are presented below each analysis step. Deep learning approaches are categorized as neural network models (DNN, CNN, CapsNet, and DANN), Generalive Adversarial Network (GAN), Variational Autoencoder, and Autoencoder. For each DL approach, optional algorithms are listed on top of each step in the pipeline.

- **Batch Correction** is a common source of technical variation in high-throughput sequencing experiments due to variant experimental conditions such as technicians and experimental time, imposing a major challenge in scRNA-seq data analysis. Batch effect correction algorithms include detection of mutual nearest neighbors (MNNs) (Haghverdi et al. 2018) , canonical correlation analysis (CCA) with Seurat (Butler et al. 2018), and Hormony algorithm through cell-type representation (Korsunsky et al. 2019).
- **Imputation** step is necessary to handle high sparsity data matrix, due to missing value or dropout in scRNA-seq data analysis. Several tools have been developed to "impute" zero values in scRNA-seq data, such as SCRABBLE (Peng et al. 2019), SAVER (Huang et al. 2018) and scImpute (Li and Li 2018). Dimensionality reduction and visualization are essential steps to represent biological meaningful variation and high dimensionality with significantly reduced computational cost. Dimensionality reduction methods, such as PCA, are widely used in scRNA-seq data analysis to achieve that purpose. More advanced nonlinear approaches that preserve the topological structure and avoid overcrowding in lower dimension representation, such as LLE (Roweis and Saul 2000) (used in SLICER (Welch, Hartemink, and Prins 2016)), tSNE (Linderman et al. 2019), and UMAP (Becht et al. 2018) have also been developed and adopted as a standard in single-cell data visualization.

Dimensionality reduction and visualization are essential steps to represent biologically meaningful variations and high dimensionality with significantly reduced computational cost. Dimensionality reduction methods, such as principal component analysis (PCA), are widely used in scRNA-seq data analysis to achieve that purpose. More advanced nonlinear approaches that preserve the topological structure and avoid overcrowding in lower dimension representation, such as LLE [38] (used in SLICER [39]), tSNE [40], and UMAP [41], have also been developed and adopted as a standard in single-cell data visualization.

Clustering analysis is a key step to identify cell subpopulations or distinct cell types to unravel the extent of heterogeneity and their associated cell-type-specific markers. Unsupervised clustering is frequently used here to categorize cells into clusters by their similarity often taken the aforementioned dimensionality-reduced representations as input, such as community detection algorithm Louvain (Subelj and Bajec 2011) and Leiden (Traag, Waltman, and Eck 2019) , or data-driven dimensionality reduction followed with k-Means cluster by SIMLR (Wang et al. 2017) .

Feature selection is another important step in single-cell RNA-seq analysis is to select a subset of genes, or features, for cell-type identification and functional enrichment of each cluster. This step is achieved by differential expression analysis designed for scRNA-seq, such as MAST that used linear model fitting and likelihood ratio testing (Finak et al. 2015); SCDE that adopted a Bayesian approach with a Negative Binomial model for gene expression and Poisson process for dropouts (Kharchenko, Silberstein, and Scadden 2014), or DEsingle that utilized a Zero-Inflated Negative Binomial model to estimate the dropouts (Miao et al. 2018).

Besides these key steps, downstream analysis can include cell type identification, coexpression analysis, prediction of perturbation response, where DL has also been applied. Other advanced analyses including trajectory inference and velocity and pseudotime analysis are not discussed here because most of the approaches on these topics are non-DL based.

Chapter 3

Overview of common deep learning models for scRNA-seq analysis

We start our review by introducing the general formulations of widely used deep learning models. As most of the tasks including batch correction, dimensionality reduction, imputation, and clustering are unsupervised learning tasks, we will give special attention to unsupervised models including variational autoencoder (VAE), the autoencoder (AE), or generative adversarial networks (GAN). We will also discuss the general supervised and transfer learning formulations, which find their applications in cell type predictions and functional studies. We will discuss these models in the context of scRNA-seq, detailing the different features and training strategies of each model and bringing attention to their uniqueness.

Abdelaal, T., L. Michielsen, D. Cats, D. Hoogduin, H. Mei, M. J. T. Reinders, and A. Mahfouz. 2019. "A Comparison of Automatic Cell Identification Methods for Single-Cell RNA Sequencing Data." Journal Article. *Genome Biol* 20 (1): 194. https://doi.org/10.1186/s13059-019-1795-z.

Amodio, M., D. van Dijk, K. Srinivasan, W. S. Chen, H. Mohsen, K. R. Moon, A. Campbell, et al. 2019. "Exploring Single-Cell Data with Deep Multitasking Neural Networks." Journal Article. *Nat Methods* 16 (11): 1139–45. https://doi.org/10.1038/s41592-019-0576-7.

Arisdakessian, C., O. Poirion, B. Yunits, X. Zhu, and L. X. Garmire. 2019. "DeepImpute: An Accurate, Fast, and Scalable Deep Neural Network Method to Impute Single-Cell RNA-Seq Data." Journal Article. *Genome Biol* 20 (1): 211. https://doi.org/10.1186/s13059-019-1837-6.

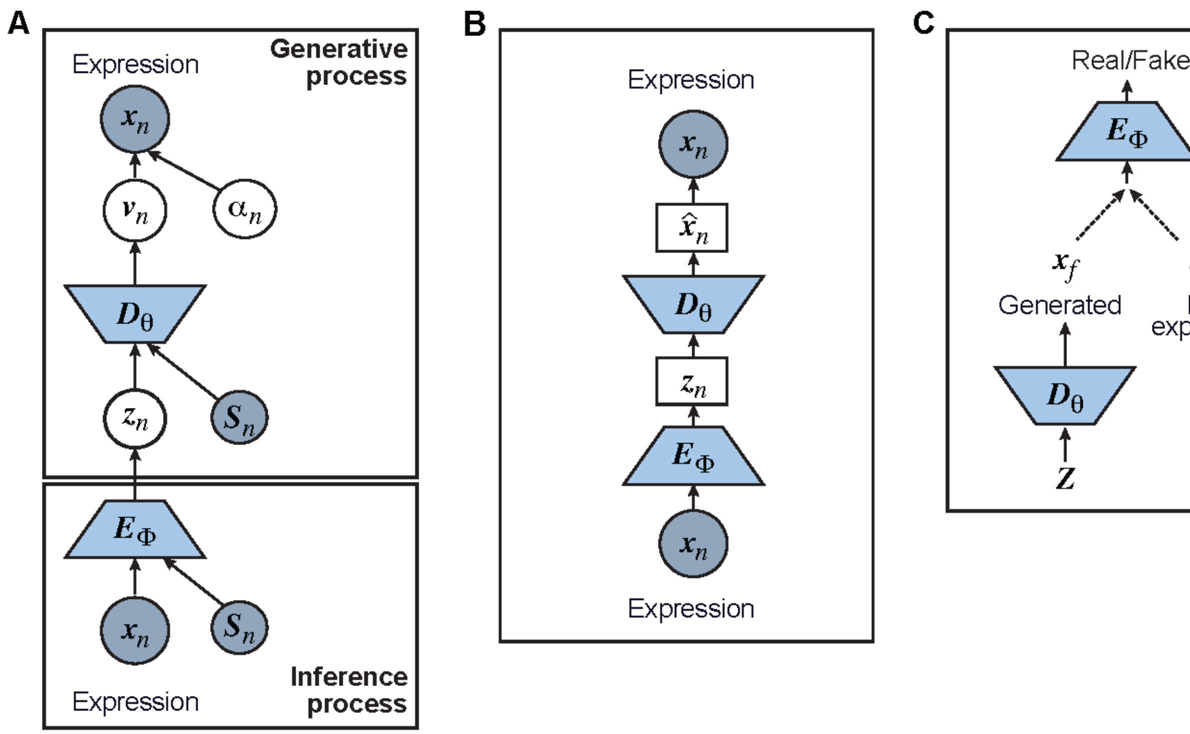

Figure 3.1: Graphical models of the surveyed DL models including A) Variational Autoencoder (VAE); B) Autoencoder (AE); and C) Generative Adversarial Network (GAN)

Becht, E., L. McInnes, J. Healy, C. A. Dutertre, I. W. H. Kwok, L. G. Ng, F. Ginhoux, and E. W. Newell. 2018. "Dimensionality Reduction for Visualizing Single-Cell Data Using UMAP." Journal Article. *Nat Biotechnol*. https://doi.org/10.1038/nbt.4314.

Bost, P., A. Giladi, Y. Liu, Y. Bendjelal, G. Xu, E. David, R. Blecher-Gonen, et al. 2020. "Host-Viral Infection Maps Reveal Signatures of Severe COVID-19 Patients." Journal Article. *Cell* 181 (7): 1475–1488 e12. https://doi.org/10.1016/j.cell.2020.05.006.

Butler, A., P. Hoffman, P. Smibert, E. Papalexi, and R. Satija. 2018. "Integrating Single-Cell Transcriptomic Data Across Different Conditions, Technologies, and Species." Journal Article. *Nat Biotechnol* 36 (5): 411–20. https://doi.org/10.1038/nbt.4096.

Chen, G., B. Ning, and T. Shi. 2019. "Single-Cell RNA-Seq Technologies and Related Computational Data Analysis." Journal Article. *Front Genet* 10: 317. https://doi.org/10.3389/fgene.2019.00317.

Eisenstein, M. 2020. "Single-Cell RNA-Seq Analysis Software Providers Scramble to Offer Solutions." Journal Article. *Nat Biotechnol* 38 (3): 254–57. https://doi.org/10.1038/s41587-020-0449-8.

Eraslan, G., L. M. Simon, M. Mircea, N. S. Mueller, and F. J. Theis. 2019. "Single-Cell RNA-Seq Denoising Using a Deep Count Autoencoder." Journal Article. *Nat Commun* 10 (1): 390. https://doi.org/10.1038/s41467-018-07931-2.

Finak, G., A. McDavid, M. Yajima, J. Deng, V. Gersuk, A. K. Shalek, C. K. Slichter, et al. 2015. "MAST: A Flexible Statistical Framework for Assessing Transcriptional Changes and Characterizing Heterogeneity in Single-Cell RNA Sequencing Data." Journal Article. *Genome Biol* 16: 278. https://doi.org/10.1186/s13059-015-0844-5.

Hafemeister, C., and R. Satija. 2019. "Normalization and Variance Stabilization of Single-Cell RNA-Seq Data Using Regularized Negative Binomial Regression." Journal Article. *Genome Biol* 20 (1): 296. https://doi.org/10.1186/s13059-019-1874-1.

Haghverdi, L., A. T. L. Lun, M. D. Morgan, and J. C. Marioni. 2018. "Batch Effects in Single-Cell RNA-Sequencing Data Are Corrected by Matching Mutual Nearest Neighbors." Journal Article. *Nat Biotechnol* 36 (5): 421–27. https://doi.org/10.1038/nbt.4091.

Huang, M., J. Wang, E. Torre, H. Dueck, S. Shaffer, R. Bonasio, J. I. Murray, A. Raj, M. Li, and N. R. Zhang. 2018. "SAVER: Gene Expression Recovery for Single-Cell RNA Sequencing." Journal Article. *Nat Methods* 15 (7): 539–42. https://doi.org/10.1038/s41592-018-0033-z.

Kharchenko, P. V., L. Silberstein, and D. T. Scadden. 2014. "Bayesian Approach to Single-Cell Differential Expression Analysis." Journal Article. *Nat Methods*

11 (7): 740–42. https://doi.org/10.1038/nmeth.2967.

Kinker, G. S., A. C. Greenwald, R. Tal, Z. Orlova, M. S. Cuoco, J. M. McFarland, A. Warren, et al. 2020. "Pan-Cancer Single-Cell RNA-Seq Identifies Recurring Programs of Cellular Heterogeneity." Journal Article. *Nat Genet* 52 (11): 1208–18. https://doi.org/10.1038/s41588-020-00726-6.

Korsunsky, I., N. Millard, J. Fan, K. Slowikowski, F. Zhang, K. Wei, Y. Baglaenko, M. Brenner, P. R. Loh, and S. Raychaudhuri. 2019. "Fast, Sensitive and Accurate Integration of Single-Cell Data with Harmony." Journal Article. *Nat Methods* 16 (12): 1289–96. https://doi.org/10.1038/s41592-019-0619-0.

Lahnemann, D., J. Koster, E. Szczurek, D. J. McCarthy, S. C. Hicks, M. D. Robinson, C. A. Vallejos, et al. 2020. "Eleven Grand Challenges in Single-Cell Data Science." Journal Article. *Genome Biol* 21 (1): 31. https://doi.org/10.1186/s13059-020-1926-6.

Li, W. V., and J. J. Li. 2018. "An Accurate and Robust Imputation Method scImpute for Single-Cell RNA-Seq Data." Journal Article. *Nat Commun* 9 (1): 997. https://doi.org/10.1038/s41467-018-03405-7.

Linderman, G. C., M. Rachh, J. G. Hoskins, S. Steinerberger, and Y. Kluger. 2019. "Fast Interpolation-Based t-SNE for Improved Visualization of Single-Cell RNA-Seq Data." Journal Article. *Nat Methods* 16 (3): 243–45. https://doi.org/10.1038/s41592-018-0308-4.

Lopez, R., J. Regier, M. B. Cole, M. I. Jordan, and N. Yosef. 2018. "Deep Generative Modeling for Single-Cell Transcriptomics." Journal Article. *Nat Methods* 15 (12): 1053–58. https://doi.org/10.1038/s41592-018-0229-2.

Lun, A. T., K. Bach, and J. C. Marioni. 2016. "Pooling Across Cells to Normalize Single-Cell RNA Sequencing Data with Many Zero Counts." Journal Article. *Genome Biol* 17: 75. https://doi.org/10.1186/s13059-016-0947-7.

Macosko, E. Z., A. Basu, R. Satija, J. Nemesh, K. Shekhar, M. Goldman, I. Tirosh, et al. 2015. "Highly Parallel Genome-Wide Expression Profiling of Individual Cells Using Nanoliter Droplets." Journal Article. *Cell* 161 (5): 1202–14. https://doi.org/10.1016/j.cell.2015.05.002.

Mannarapu, M., B. Dariya, and O. R. Bandapalli. 2021. "Application of Single-Cell Sequencing Technologies in Pancreatic Cancer." Journal Article. *Mol Cell Biochem* 476 (6): 2429–37. https://doi.org/10.1007/s11010-021-04095-4.

Miao, Z., K. Deng, X. Wang, and X. Zhang. 2018. "DEsingle for Detecting Three Types of Differential Expression in Single-Cell RNA-Seq Data." Journal Article. *Bioinformatics* 34 (18): 3223–24. https://doi.org/10.1093/bioinformatics/bty332.

Navin, N. E. 2015. "The First Five Years of Single-Cell Cancer Genomics and Beyond." Journal Article. *Genome Res* 25 (10): 1499–1507. https://doi.org/10.1101/gr.191098.115.

Peng, T., Q. Zhu, P. Yin, and K. Tan. 2019. "SCRABBLE: Single-Cell RNA-Seq Imputation Constrained by Bulk RNA-Seq Data." Journal Article. *Genome Biol* 20 (1): 88. https://doi.org/10.1186/s13059-019-1681-8.

Petegrosso, R., Z. Li, and R. Kuang. 2020. "Machine Learning and Statistical Methods for Clustering Single-Cell RNA-Sequencing Data." Journal Article. *Brief Bioinform* 21 (4): 1209–23. https://doi.org/10.1093/bib/bbz063.

Picelli, S., A. K. Bjorklund, O. R. Faridani, S. Sagasser, G. Winberg, and R. Sandberg. 2013. "Smart-Seq2 for Sensitive Full-Length Transcriptome Profiling in Single Cells." Journal Article. *Nat Methods* 10 (11): 1096–98. https://doi.org/10.1038/nmeth.2639.

Roweis, S. T., and L. K. Saul. 2000. "Nonlinear Dimensionality Reduction by Locally Linear Embedding." Journal Article. *Science* 290 (5500): 2323–26. https://doi.org/10.1126/science.290.5500.2323.

Srinivasan, S., A. Leshchyk, N. T. Johnson, and D. Korkin. 2020. "A Hybrid Deep Clustering Approach for Robust Cell Type Profiling Using Single-Cell RNA-Seq Data." Journal Article. *RNA* 26 (10): 1303–19. https://doi.org/10.1261/rna.074427.119.

Stuart, T., A. Butler, P. Hoffman, C. Hafemeister, E. Papalexi, 3rd Mauck W. M., Y. Hao, M. Stoeckius, P. Smibert, and R. Satija. 2019. "Comprehensive Integration of Single-Cell Data." Journal Article. *Cell* 177 (7): 1888–1902 e21. https://doi.org/10.1016/j.cell.2019.05.031.

Subelj, L., and M. Bajec. 2011. "Unfolding Communities in Large Complex Networks: Combining Defensive and Offensive Label Propagation for Core Extraction." Journal Article. *Phys Rev E Stat Nonlin Soft Matter Phys* 83 (3 Pt 2): 036103. https://doi.org/10.1103/PhysRevE.83.036103.

Traag, V. A., L. Waltman, and N. J. van Eck. 2019. "From Louvain to Leiden: Guaranteeing Well-Connected Communities." Journal Article. *Sci Rep* 9 (1): 5233. https://doi.org/10.1038/s41598-019-41695-z.

Tran, H. T. N., K. S. Ang, M. Chevrier, X. Zhang, N. Y. S. Lee, M. Goh, and J. Chen. 2020. "A Benchmark of Batch-Effect Correction Methods for Single-Cell RNA Sequencing Data." Journal Article. *Genome Biol* 21 (1): 12. https://doi.org/10.1186/s13059-019-1850-9.

Vallejos, C. A., J. C. Marioni, and S. Richardson. 2015. "BASiCS: Bayesian Analysis of Single-Cell Sequencing Data." Journal Article. *PLoS Comput Biol* 11 (6): e1004333. https://doi.org/10.1371/journal.pcbi.1004333.

Vitak, S. A., K. A. Torkenczy, J. L. Rosenkrantz, A. J. Fields, L. Christiansen, M. H. Wong, L. Carbone, F. J. Steemers, and A. Adey. 2017. "Sequencing Thousands of Single-Cell Genomes with Combinatorial Indexing." Journal Article. *Nat Methods* 14 (3): 302–8. https://doi.org/10.1038/nmeth.4154.

Wang, B., J. Zhu, E. Pierson, D. Ramazzotti, and S. Batzoglou. 2017. "Visualization and Analysis of Single-Cell RNA-Seq Data by Kernel-Based Similarity Learning." Journal Article. *Nat Methods* 14 (4): 414–16. https://doi.org/10.1038/nmeth.4207.

Wauters, E., P. Van Mol, A. D. Garg, S. Jansen, Y. Van Herck, L. Vanderbeke, A. Bassez, et al. 2021. "Discriminating Mild from Critical COVID-19 by Innate and Adaptive Immune Single-Cell Profiling of Bronchoalveolar Lavages." Journal Article. *Cell Res* 31 (3): 272–90. https://doi.org/10.1038/s41422-020-00455-9.

Welch, J. D., A. J. Hartemink, and J. F. Prins. 2016. "SLICER: Inferring Branched, Nonlinear Cellular Trajectories from Single Cell RNA-Seq Data." Journal Article. *Genome Biol* 17 (1): 106. https://doi.org/10.1186/s13059-016-0975-3.

Wolf, F. A., P. Angerer, and F. J. Theis. 2018. "SCANPY: Large-Scale Single-Cell Gene Expression Data Analysis." Journal Article. *Genome Biol* 19 (1): 15. https://doi.org/10.1186/s13059-017-1382-0.

Wolock, S. L., R. Lopez, and A. M. Klein. 2019. "Scrublet: Computational Identification of Cell Doublets in Single-Cell Transcriptomic Data." Journal Article. *Cell Syst* 8 (4): 281–291 e9. https://doi.org/10.1016/j.cels.2018.11.005.

Xu, Y., Z. Zhang, L. You, J. Liu, Z. Fan, and X. Zhou. 2020. "scIGANs: Single-Cell RNA-Seq Imputation Using Generative Adversarial Networks." Journal Article. *Nucleic Acids Res* 48 (15): e85. https://doi.org/10.1093/nar/gkaa506.