From healthy to diseased; using scVI to detect nonlinear gene expression changes on paths between healthy and sick cells in latent space

Yegor Kuznetsov[1] **Sophia Jannetty**[2]
[1]University of Washington Department of Computer Science
[2]University of Washington Department of Biology
yegor@uw.edu
jannetty@uw.edu

Abstract

Researcers comparing sick cell transcription profiles to healthy cell transcription profiles may miss important information about intermediate transitioning transcription profiles not represented in healthy or sick datasets. Without time-series data calibrated to degree of disease progression, analyzing changes in transcriptional profiles as diseases progress over time can be challenging. Here we investigate the possibility of generating and analyzing artificial transcription data for cells transitioning from a healthy to a sick state. Using the deep generative modeling technique scVI (Lopez et al. 2018), we embed published scRNAseq data from sick and healty cells into a learned 10-dimensional latent space. We then sample the latent space on paths between sick and healthy cells in latent space and use the model decoder to generate transcriptional profiles for cells that would exist at each location. We then analyze these transcriptomes and identify genes that have a nonlinear predicted trajectory on the path between healthy and sick. We propose future methods for investigating and validating this method.

1 Introduction

Differential expression analysis is a common technique for assessing differences between transcriptional profiles of healthy and sick cells(Anders and Huber 2010). However, differential expression analysis may miss transient changes in expression that occur as a cell transitions from healthy to sick (Figure 1). We aim to develop a technique that builds on recent advances in scRNAseq analysis to address this limitation.

Several recent papers have demonstrated how neural networks can be used to embed scRNA-seq data in low-dimensional latent space (Ding, Condon, and Shah 2018; Wang and Gu 2018; Grønbech et al. 2018; Eraslan et al. 2019). Emphasized utilities of these methods have included effective denoising (Eraslan et al. 2019), dimensionality reduction (Wang and Gu 2018; Ding, Condon, and Shah 2018), and visualization (Wang and Gu 2018). Several of these methods also provide decoders that allow for translation from latent space back into gene space. Grønbech et al. (2018) and Lopez et al. (2018) both discuss decoding latent-space representations of cells by using a nonlinear decoder to generate a posterior estimate of the expression of each gene in a cell. Our project is focused on investigating the possibility of using these tools to learn about changes in gene expression patterns as a cell transitions from healthy to sick.

Throughout this project we aim to explore the following enormous assumption; given a latent space in which healthy and sick cells cluster apart from each other (such that there are healthy regions and sick regions in the space), do decoded locations between these regions correspond to transcriptional

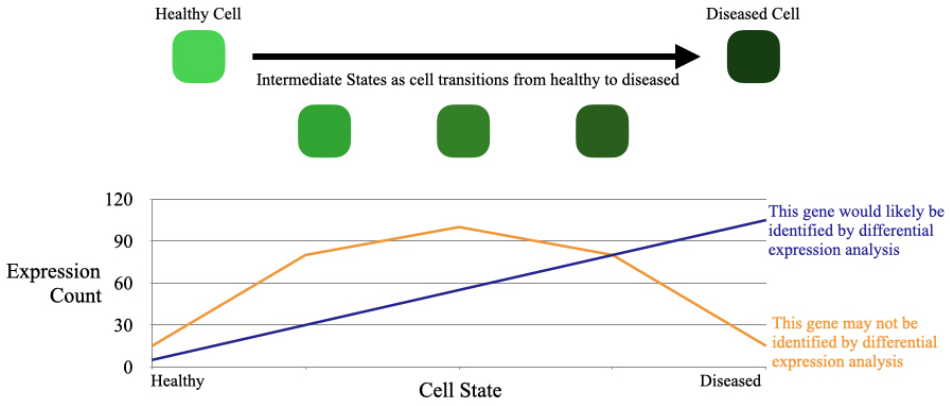

Figure 1: Project motivation. Differential expression analysis is effective at identifying genes that have different rates of transcription between healthy and sick cells. However, differential expression analysis may not identify genes that have transient changes in expression as cells transition from healthy to sick.

profiles of cells in transition between healthy and sick? Furthermore, do decoded locations closer to healthy regions than to sick regions correspond to transcriptional profiles of cells that are only slightly sick (and vice versa)? We begin by operating as though this is a fair assumption. We tried using this method using two separate datasets (both relevant to hematopoietic stem and progenitor cells, or HSPCs). For each, we trained the scVI model on scRNA-seq data with sick and healthy labels. Once scVI embedded each cell in latent space, we sampled the space between healthy and sick cells and used the model's nonlinear decoder to generate transcriptional profiles corresponding to each location. We then analyzed the model-predicted changes in gene expressions. Figure 2 gives a broad overview of our approach.

1.1 Related Work

Our method heavily builds upon pseudotime. Pseudotime (Reid and Wernisch 2016) is a method for estimating a cell's progress through a given transition. Pseudotime takes transcriptional data (scRNAseq data or microarray data) from cells assumed to be at various phases of a given transition. The method then embeds this data in a latent space with an enforced structure related to the temporal capture time of each cell (Reid and Wernisch 2016). The resulting latent dimension corresponds to pseudotime. Using this method, researchers can take a set of transcriptional profiles and order them by their phase in progression between one state and another. However, this method is both supervised (in that it imposes a time-related restraint on its latent space) and is only effective on datasets in which cells are at various stages of progression. This method cannot be used to generate predicted transcriptional profiles of cells at a certain phase of transition if cells in that phase of transition do not exist in the dataset.

Our method is also related to pathway analysis methods like TIPS (Zheng et al. 2021) and MetaCell (Baran et al. 2019). TIPS orders cells in pseudotime and then uses prior knowledge of the data to investigate changes in specific, hypothetically relevant genes' expressions as cells progress through pseudotime (Zheng et al. 2021). This method is an exciting progression of pseudotime, however it also can only draw observations about changes in gene expression levels captured within the dataset. Additionally, TIPS relies on prior knowledge to identify genes to investigate. MetaCell generates composite cell profiles from cells that cluster by similarity to minimize noise, allowing for analysis between different cell types. This method also has a supervised component (in that the genes used to calculate similarity between cells are selected by the researcher) and does not ensure that cells at different phases in a transition will be represented in seperate composite profiles.

Our work is, to our knowledge, unique in two ways. First, our method aims to identify transcriptional changes over transitional phases without assuming cells of each phase are represented in the dataset. We are assuming we have a starting point and an ending point, and are generating transcriptional profiles for intermediate transitional phases sampling the latent space between these starting and

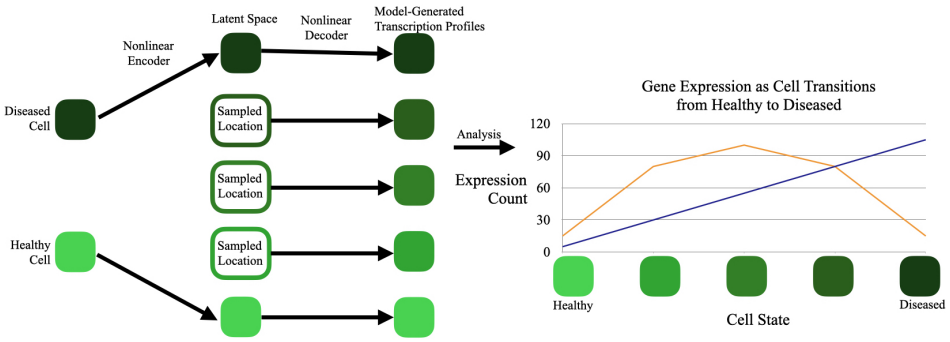

Figure 2: A schematic of our approach. We trained scVI on scRNA-seq data of healthy and diseased cells. We then sampled locations on the lines between healthy and diseased cells in latent space and used scVI's decoder to generate transcriptional profiles from these sampled locations. We then analyzed the changes in expression of each gene over time.

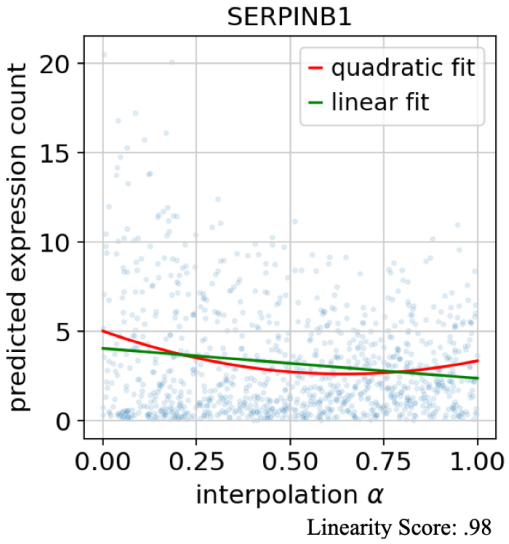

Figure 3: An example of a highly ranked gene in our analysis generated from the Ferrall-Fairbanks et al. (2022) Leukemia data. SERPINB1 is a neutrophil serine protease inhibitor that protects the cell from proteases released in the cytoplasm during stress or infection.

ending points. Second, our method is entirely unsupervised. We aim to detect genes that undergo transcriptional changes over a transition with no prior knowledge of the cells or the transition in place.

1.2 Datasets

For this project, we worked with two different datasets; an Aplastic Anemia dataset published by Tonglin et al. (2022), and a Chronic Myelomonocytic Leukemia dataset published by Ferrall-Fairbanks et al. (2022).

1.2.1 Aplastic Anemia

Aplastic anemia (AA) is a condition in which a patient's bone marrow fails to form enough red blood cells, white blood cells, and platelets, resulting in pancytopenia (Young 2018). Pathophysiological mechanisms of this disease include direct damage to bone marrow (commonly from chemotherapy), germ line loss-of-function mutations that interfere with blood-cell precursor DNA repair pathways, and autoimmune attack (Young 2018).

Despite the severity of this disease, the mechanisms underlying its onset remain obscure due to limitations in experimental techniques. It is thought that T cells may target hematopoietic stem and progenitor cells (HSPCs), which are blood cell precursor cells, in patients with immune-mediated AA (Tonglin et al. 2022). Understanding the transcriptional differences an HSPCs undergoes as it progresses from healthy to AA could uncover the cause of the T cell targeting. However HSPCs are severely depleated in AA patients, making them a difficult target to study (Zhu et al. 2021). Recent studies like Tonglin et al. (2022) and Zhu et al. (2021) have used single-cell RNA-sequencing (scRNA-seq) to perform differential expression analysis between healthy and AA patient HSPCs. However, these analyses will only identify genes that have different expression levels between sick and healthy cells. We aim to identify potential transient changes in gene expression that characterize an HSPC's transition from healthy to AA.

1.2.2 Chronic Myelomonocytic Leukemia

We selected the Chronic Myelomonocytic Leukemia data published by Ferrall-Fairbanks et al. (2022) because it was a large dataset (41,521 healthy cells from 5 individuals and 120,543 sick cells from 26 individuals) and because it was entirely cells from a very specific cell type (HSPCs expressing CD34). We wanted a large number of individuals represented in the dataset while making sure the greatest differences between cells would be disease status as opposed to cell type. The similarity of cell type to our AA data was also convenient in our overall familiarity with the cells.

1.2.3 Limitations of our Datasets and Potential Future Dataset of Interest

We began our project inspired to pursue research relevant to Aplastic Anemia. However the focus of our project has shifted and unfortunately the datasets we used are not optimal for investigating our new technique. The Tonglin et al. (2022) data is too small to confidently train a nonlinear encoder or decoder. The Ferrall-Fairbanks et al. (2022) data is from cancer patients, and cancer is a poor example of a disease in which cells exist on a spectrum of sickness. Cells can turn from healthy to cancerous, but once a cell is cancerous all its descendents will be cancerous. It is therefore most likely that the cancer cells in the dataset never transitioned from healthy to cancerous, making this a bad choice of datasets for investigating our new method.

The ideal dataset would be time-series cell-type-specific transcriptional data from cells undergoing a well-characterized progression. Time-course differentiation data could be perfect as we could train our model on data from the first and last time points and compare our model-generated intermediate transcriptional profiles to the transcriptional profiles of the cells at intermediate time points. We have identified the data published Lu et al. (2020) as a potential dataset of interest for future investigation of this method. This dataset would need to be filtered for a specific cell type, but it is in-vitro expression data with many replicates and multiple time points.

2 Methods and abbreviated results

we visualized clustering of the sick and healthy labels in latent space using UMAP projections. We then drew lines between sick individuals and healthy individuals. We selected evenly distributed points along these lines and decoded each location into its own distinct transcriptional profile. For each gene, we plotted each point's predicted expression given its interpolation alpha (the interpolation alpha being each point's distance from the sick end of the line such that the healthy cell itself has an interpolation alpha of 1, the sick cell has an interpolation alpha of 0, an artifical cell located halfway on the line connecting the healthy and sick cells has an interpolation alpha of .5, see).

We selected 100 random healthy cells and 100 random diseased cells and drew 100 lines connecting these cells'points in latent space. We randomly sampled one point in the latent space on each line and decoded these locations into transcriptional profiles. We then made plots for each gene showing the transcriptional trajectories seen across these 100 different paths through latent space. We fit linear and quadratic lines to each of these plots and calculated the sum of residuals for the

2.1 Aplastic Anemia data analysis

2.1.1 The Data

For this study we used the scRNA-seq data published by Tonglin et al. (2022). These data were from bone marrow samples taken from two healthy donors and two AA patients. These data comprise 23685 HSPCs (14163 AA and 9522 healthy) with 33538 genes per cell. Sequencing was performed by 10x Genomics ("Single Cell Gene Expression," n.d.).

2.1.2 Data Preprocessing

We downloaded the data from the NCBI Gene Expression Omnibus (https://www.ncbi.nlm.nih.gov/geo/ , GEO accession: GSE181989). We used the python package Scanpy (Wolf, Angerer, and Theis 2018) to filter out all genes that were in fewer than 3 cells and all cells that had expression of fewer than 200 genes. After filtering, we had 22505 cells with 20059 genes. [NUMBER] of these cells were from the two AA patients and [NUMBER] of these cells were from the two healthy controls.

2.1.3 Principal Component and UMAP Visualizations

We used the Scanpy python package (Wolf, Angerer, and Theis 2018) to make these visualizations. For the PC projection, we projected the data onto the first two principal components which captured 10.6% and 2.7% of the overall variance of the dataset respectively. We colored each cell by disease status (see Figure 4) and by individual (see Figure 5).

2.1.4 Linear Discriminant Analysis

We used the Scanpy python package (Wolf, Angerer, and Theis 2018) to perform linear discriminant analysis (LDA) on the first 50 principal components of our data. We used 5-fold cross validation to assess the model's prediction accuracy. Our model accurately predicted the disease status of 98.2% of the cells in the training data and 98.0% of the cells in the test data.

To assess whether LDA was distinguishing between transcriptome profiles of diseased vs healthy individuals or just differentiating between the individuals in general, we retrained the model using data from just two individuals (one sick and one diseased). We then used this model to predict the disease status of the other two individuals. We started by again using the first 50 principal components. Our model accurately predicted the disease status 48.5% of the test cells. We switched the individuals that were used to train/test the model and found the resulting model accurately predicted the disease status of 67% of the test cells. We also performed this analysis using the first 5 principal components, which yielded 56.6% and 60.4% testing accuracy for each training/test data pair.

2.1.5 scVI

We used the scvi-tools python package to perform scVI (Lopez et al. 2018; Gayoso et al. 2022). INSERT INFORMATION ABOUT THE TECNICAL DETAILS HERE.

3 Discussion and Results

3.1 Visualization reveals cells do not cluster by disease status

Existing AA scRNA-seq data analysis uses linear dimensionality reduction techniques as a preprocessing step prior to clustering. We began our analysis by following these same steps and assessing whether such preprocessing was sufficient for generating a dataset capable of training a simple but generalizable linear classifier. We first visualized our data following a tutorial provided by the Scanpy[CITE] python package and qualitatively noticed that the diseased and healthy cells do not seem to cluster distinctly (see Figure 4). We also noticed that cells from different individuals seemed to have very different transcriptional profiles (see Figure 5). This led us to hypothesize that a linear model would either perform poorly when predicting disease status or perform well by learning to distinguish between individuals instead of distinguishing between disease status.

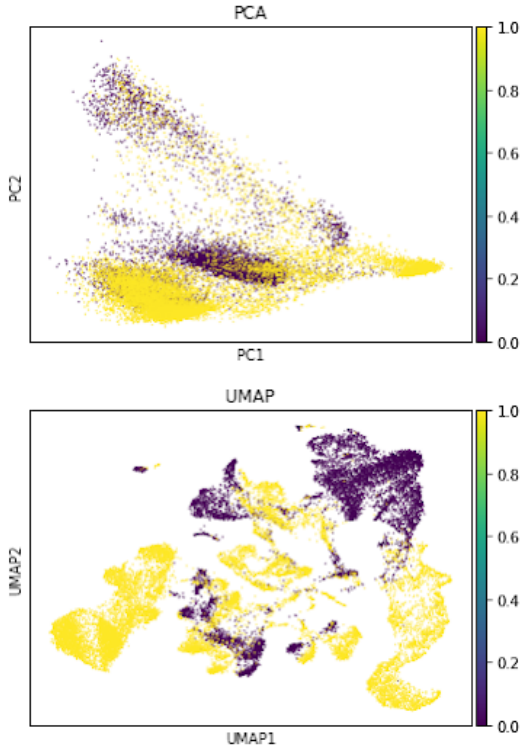

Figure 4: Principal Component and UMAP Projection visualizations colored by disease status. 0 is healthy, 1 is AA.

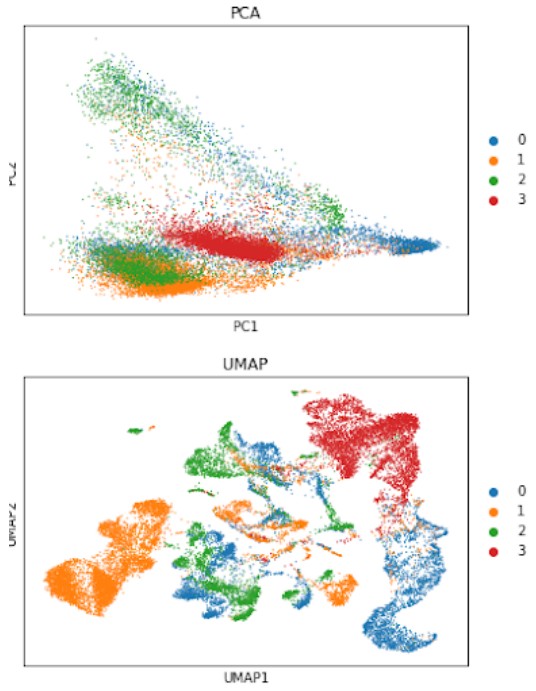

Figure 5: Principal Component and UMAP Projection visualizations colored by individual. Individuals 0 and 1 have AA. Individuals 2 and 3 are healthy controls.

3.2 LDA is insufficient for predicting disease status

We performed five-fold cross validation on an LDA model trained on the first 50 principal components of the dataset and found the prediction accuracy to be an outlandishly high 98%. We subsequently hypothesised that this accuracy could be attributed to the model learning to distingish each individual as opposed to the model distinguishing between healthy and AA individuals. To test this hypothesis, we used the same model hyperparameters but trained our model on data from exclusively one healthy and one AA individual. Our hypothesis was that if the model is truly learning to distinguish between illness status, a model trained on data from representative AA patients and healthy individuals should be able to accurately predict the disease status of cells from other AA and healthy patients. Conversely, if the model is learning to distingish individuals, a model trained on some individuals should not accurately predict the disease status of other individuals. Because our dataset only contains data from four individuals, we trained the model on data from two individuals (one sick and one AA) and tested the model using the other two patients' data. We then swapped the test and the training data and again assessed prediction accuracy. Our measured prediction accuracy was 48.5% and 67%. This decrease in accuracy supported our hypothesis that, even though it may be able to distinguish individuals, a simple linear model is not sufficient to reliably distinguish between AA and healthy cells.

We next considered whether larger principal components contained the dimensions along which the AA and healthy cells separated most distinctly and whether small variance introduced by the smaller principal components may be making the task of performing LDA more challenging. To investigate this possibility, we ran this analysis again using only the first five principal components of the dataset. This analysis yielded a prediction accuracies of 56.6% and 60.4%, leading us to conclude that even the largest principal components encompass variances unrelated to disease status.

3.3 scVI does something I hope

We discovered scVI through our class reading and wanted to investigate its performance on the Tonglin et al. (2022) data. We trained scVI on the Tonglin et al. (2022) data using the scvi-tools python packge (Gayoso et al. 2022) with the intention of exploring the differences in AA cell distributions vs healthy cell distributions in latent space. We next wanted to build a simple classifier using this latent representation in order to create a path between the two cell-type distributions.

4 Conclusion

scRNA-seq data is challenging to analyze due to the inherent noise, batch effects, dropout rates, and differences between individuals. The challenge of accounting for person-to-person variability is amplified when there are a small number of individuals in your dataset. The Tonglin et al. (2022) data gave us an opportunity to experiment with a low-individual scRNA-seq dataset while working towards answering mechanistic questions about a devistating disease. Our investigations revealed that simple linear models are insufficient for discriminating between disease states with such a low number of individuals. Something about scVI.

References

Anders, Simon, and Wolfgang Huber. 2010. "Differential expression analysis for sequence count data." *Genome Biology* 11, no. 10 (October 27, 2010): R106. ISSN: 1474-760X, accessed December 16, 2022. https://doi.org/10.1186/gb-2010-11-10-r106. https://doi.org/10.1186/gb-2010-11-10-r106.

Baran, Yael, Akhiad Bercovich, Arnau Sebe-Pedros, Yaniv Lubling, Amir Giladi, Elad Chomsky, Zohar Meir, Michael Hoichman, Aviezer Lifshitz, and Amos Tanay. 2019. "MetaCell: analysis of single-cell RNA-seq data using K-nn graph partitions." *Genome Biology* 20, no. 1 (October 11, 2019): 206. ISSN: 1474-760X. https://doi.org/10.1186/s13059-019-1812-2.

Ding, Jiarui, Anne Condon, and Sohrab P. Shah. 2018. "Interpretable dimensionality reduction of single cell transcriptome data with deep generative models." Number: 1 Publisher: Nature Publishing Group, *Nature Communications* 9, no. 1 (May 21, 2018): 2002. ISSN: 2041-1723, accessed December 6, 2022. https://doi.org/10.1038/s41467-018-04368-5. https://www.nature.com/articles/s41467-018-04368-5.

Eraslan, Gökcen, Lukas M. Simon, Maria Mircea, Nikola S. Mueller, and Fabian J. Theis. 2019. "Single-cell RNA-seq denoising using a deep count autoencoder." Number: 1 Publisher: Nature Publishing Group, *Nature Communications* 10, no. 1 (January 23, 2019): 390. ISSN: 2041-1723, accessed December 6, 2022. https://doi.org/10.1038/s41467-018-07931-2. https://www.nature.com/articles/s41467-018-07931-2.

Ferrall-Fairbanks, Meghan C., Abhishek Dhawan, Brian Johnson, Hannah Newman, Virginia Volpe, Christopher Letson, Markus Ball, et al. 2022. "Progenitor Hierarchy of Chronic Myelomonocytic Leukemia Identifies Inflammatory Monocytic-Biased Trajectory Linked to Worse Outcomes." *Blood Cancer Discovery* 3, no. 6 (November 2, 2022): 536–553. ISSN: 2643-3230, accessed December 6, 2022. https://doi.org/10.1158/2643-3230.BCD-21-0217. https://doi.org/10.1158/2643-3230.BCD-21-0217.

Gayoso, Adam, Romain Lopez, Galen Xing, Pierre Boyeau, Valeh Valiollah Pour Amiri, Justin Hong, Katherine Wu, et al. 2022. "A Python library for probabilistic analysis of single-cell omics data." Number: 2 Publisher: Nature Publishing Group, *Nature Biotechnology* 40, no. 2 (February): 163–166. ISSN: 1546-1696, accessed December 2, 2022. https://doi.org/10.1038/s41587-021-01206-w. https://www.nature.com/articles/s41587-021-01206-w.

Grønbech, Christopher H., Maximillian F. Vording, Pascal N. Timshel, Capser K. Sønderby, Tune H. Pers, and Ole Winther. 2018. *scVAE: Variational auto-encoders for single-cell gene expression datas.* Pages: 318295 Section: New Results, May 16, 2018. Accessed December 6, 2022. https://doi.org/10.1101/318295. https://www.biorxiv.org/content/10.1101/318295v1.

Lopez, Romain, Jeffrey Regier, Michael B. Cole, Michael I. Jordan, and Nir Yosef. 2018. "Deep generative modeling for single-cell transcriptomics." Number: 12 Publisher: Nature Publishing Group, *Nature Methods* 15, no. 12 (December): 1053–1058. ISSN: 1548-7105, accessed November 21, 2022. https://doi.org/10.1038/s41592-018-0229-2. https://www.nature.com/articles/s41592-018-0229-2.

Lu, Yufeng, Fion Shiau, Wenyang Yi, Suying Lu, Qian Wu, Joel D. Pearson, Alyssa Kallman, et al. 2020. "Single-Cell Analysis of Human Retina Identifies Evolutionarily Conserved and Species-Specific Mechanisms Controlling Development." *Developmental cell* 53, no. 4 (May 18, 2020): 473–491.e9. ISSN: 1534-5807, accessed December 16, 2022. https://doi.org/10.1016/j.devcel.2020.04.009. https://www.ncbi.nlm.nih.gov/pmc/articles/PMC8015270/.

Reid, John E., and Lorenz Wernisch. 2016. "Pseudotime estimation: deconfounding single cell time series." *Bioinformatics* 32, no. 19 (October 1, 2016): 2973–2980. ISSN: 1367-4803, accessed December 16, 2022. https://doi.org/10.1093/bioinformatics/btw372. https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5039927/.

"Single Cell Gene Expression." 10x Genomics. n.d. Accessed December 2, 2022. https://www.10xgenomics.com/products/single-cell-gene-expression.

Tonglin, Hu, Zhao Yanna, Yu Xiaoling, Gao Ruilan, and Yin Liming. 2022. "Single-Cell RNA-Seq of Bone Marrow Cells in Aplastic Anemia." *Frontiers in Genetics* 12 (January 3, 2022): 745483. ISSN: 1664-8021, accessed October 20, 2022. https://doi.org/10.3389/fgene.2021.745483. https://www.ncbi.nlm.nih.gov/pmc/articles/PMC8762313/.

Wang, Dongfang, and Jin Gu. 2018. "VASC: Dimension Reduction and Visualization of Single-cell RNA-seq Data by Deep Variational Autoencoder." *Genomics, Proteomics & Bioinformatics,* Bioinformatics Commons (II), 16, no. 5 (October 1, 2018): 320–331. ISSN: 1672-0229, accessed December 6, 2022. https://doi.org/10.1016/j.gpb.2018.08.003. https://www.sciencedirect.com/science/article/pii/S167202291830439X.

Wolf, F. Alexander, Philipp Angerer, and Fabian J. Theis. 2018. "SCANPY: large-scale single-cell gene expression data analysis." *Genome Biology* 19, no. 1 (February 6, 2018): 15. ISSN: 1474-760X, accessed December 2, 2022. https://doi.org/10.1186/s13059-017-1382-0. https://doi.org/10.1186/s13059-017-1382-0.

Young, Neal S. 2018. "Aplastic Anemia." Publisher: Massachusetts Medical Society _eprint: https://doi.org/10.1056/NEJMra1413485, *New England Journal of Medicine* 379, no. 17 (October 25, 2018): 1643–1656. ISSN: 0028-4793, accessed October 20, 2022. https://doi.org/10.1056/NEJMra1413485. https://doi.org/10.1056/NEJMra1413485.

Zheng, Zihan, Xin Qiu, Haiyang Wu, Ling Chang, Xiangyu Tang, Liyun Zou, Jingyi Li, et al. 2021. "TIPS: trajectory inference of pathway significance through pseudotime comparison for functional assessment of single-cell RNAseq data." *Briefings in Bioinformatics* 22, no. 5 (April 29, 2021): bbab124. ISSN: 1467-5463, accessed December 16, 2022. https://doi.org/10.1093/bib/bbab124. https://www.ncbi.nlm.nih.gov/pmc/articles/PMC8425418/.

Zhu, Caiying, Yu Lian, Chenchen Wang, Peng Wu, Xuan Li, Yan Gao, Sibin Fan, et al. 2021. "Single-cell transcriptomics dissects hematopoietic cell destruction and T-cell engagement in aplastic anemia." *Blood* 138, no. 1 (July 8, 2021): 23–33. ISSN: 0006-4971, accessed October 20, 2022. https://doi.org/10.1182/blood.2020008966. https://www.ncbi.nlm.nih.gov/pmc/articles/PMC8349468/.