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

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 exploiting these decoders 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 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 (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 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. We labeled the health score of each artifical cell as its distance from the sick end of the line (the healthy cell itself has a healthy score of 1, the sick cell has a healthy score of 0, an artifical cell located halfway on the line connecting the healthy and sick cells has a healthy score of .5). We then use Hotspot (DeTomaso and Yosef 2021) to

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).

1.1 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).

Current options for treating AA fall within three categories, each of which attempts to address pancytopenia in a different way. The first, bone marrow transplantation, aims to replace failing bone marrow with healthy donor marrow. The second, immunosuppression, aims to eliminate immune-mediated AA. Despite the success of immunosuppressant treatments, the mechanisms by which the immune system damages bone marrow remain unknown; the strongest evidence for immune-mediation of aplastic anemia is the effectiveness of immunosuppressant treatments (Young 2018). The third treatment category, stem-cell stimulation, aims to promote stem-cell regeneration within the patient directly.

Despite the severity of this disease, the mechanisms underlying its onset and the treatment options 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 differences between healthy and AA HSPCs 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. These studies have thus far used PCA to reduce dataset dimensionality, followed by graph-based clustering and likelihood ratio tests to identify significantly differentially expressed genes between clusters (Zhu et al. 2021). These studies have successfully identified differentially expressed genes between healthy and AA HSPCs. However, their adjustments for noise and batch effects are limited to their initial dimension reductions. This initial dimension reduction assumes a generalized linear model is sufficient for accurately mapping onto the low-dimensional manifold underlying the data. This is a particularly problematic assumption in papers like Tonglin et al. (2022) where samples were only taken from four patients (two healthy and two with AA). With such a low n of patients, it would be challenging to convincingly claim differences between AA and healthy clusters are due to disease status as opposed to being due to differences between individuals.

In this project we explored methods of learning differences between healthy and AA HSPCs using the scRNA-seq data published by Tonglin et al. (2022). We used computational methods to model both healthy and diseased HSPC transcriptome profiles with the goal of identifying perturbations capable of moving a diseased HSPC to a healthy cell or vice versa. Such identified perturbations could both suggest potential new therapies and help identify the mechanisms of AA onset and treatment. We have not yet reached this goal of identifying perturbations. However, our work sheds light on what methods are useful when working with data collected from a very small number of individuals.

1.2 Chronic Myelomonocytic Leukemia

2 Methods and abbreviated results

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 1) and by individual (see Figure 2).

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 1). We also noticed that cells from different individuals seemed

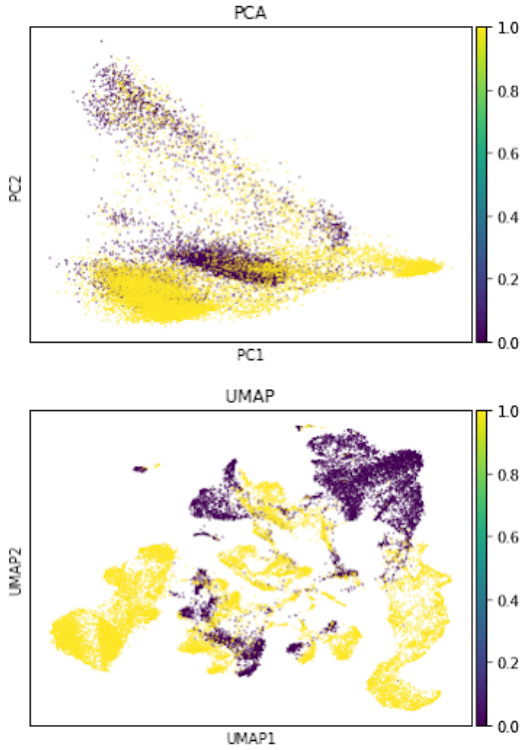

Figure 1: Principal Component and UMAP Projection visualizations colored by disease status. 0 is healthy, 1 is AA.

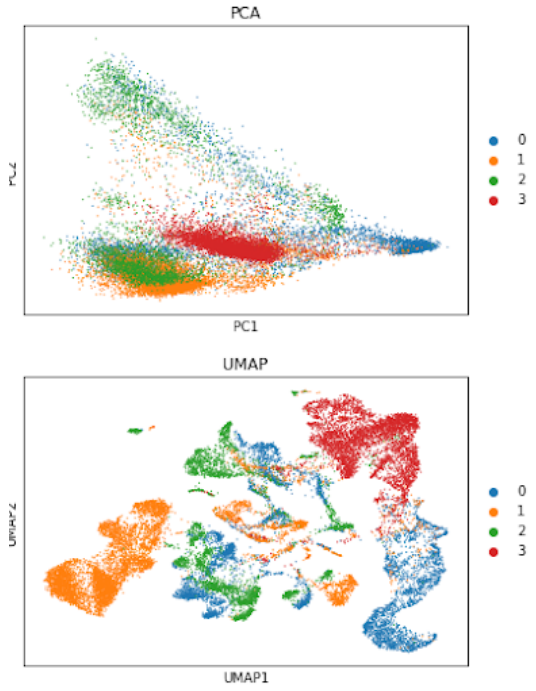

Figure 2: Principal Component and UMAP Projection visualizations colored by individual. Individuals 0 and 1 have AA. Individuals 2 and 3 are healthy controls.

to have very different transcriptional profiles (see Figure 2). 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.

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

DeTomaso, David, and Nir Yosef. 2021. "Hotspot identifies informative gene modules across modalities of single-cell genomics." *Cell Systems* 12, no. 5 (May): 446–456.e9. ISSN: 24054712, accessed December 7, 2022. https://doi.org/10.1016/j.cels.2021.04.005. https://linkinghub.elsevier.com/retrieve/pii/S2405471221001149.

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.

"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.

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/.