scVI enables effective scRNA-seq data preprocessing and highlights differences in HPSC transcriptional profiles of Aplastic Anemia patients and healthy donors

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

Aplastic anemia (AA) is a devastating illness that is challenging to study. Single-cell RNA-sequencing (scRNA-seq) datasets characterizing the transcriptional profiles of relevant cell types in AA patients have been published, but existing analyses of these datasets have been limited to clustering on dimensionalty-reduced data and performing differential expression analysis on seperate clusters. Here, we perform our own analysis on scRNA-seq data published by Tonglin et al. (2022). We begin by assessing the performance of a simple linear classifier model trained on the dimensionality-reduced data. We then use scVI to do something. We conclude that something.

1   Introduction

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. This is a curative treatment, but is limited by the prevalence of graft-versus-host disease and by its reliance on tissue donors (Young 2018). The second, immunosuppression, aims to eliminate immune-mediated AA. This category includes treatment with antilymphocyte globulin (combined with cyclosporine), which has a mild lymphocyte-depleting effect. This treatment leads to improved blood production in about 66% of patients (Bacigalupo 2017), however 30% to 60% of these patients experience relapse that requires years of continued cyclosporine therapy (Scheinberg et al. 2012). 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. One such treatment involves administrating eltrombopag, a synthetic mimic of thrombopoietin (a hormone that stimulates the production of platelets). This treatment has been shown in limited clinical trials to improve blood production in 80% of treated patients, though the mechanism of this treatment remains unknown (Young 2018).

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 depleted 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.

2 Methods and abbreviated results

2.1 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 20419 HSPCs (NUMBER AA and NUMBER healthy) with 33538 genes per cell. Sequencing was performed by 10x Genomics [CITE].

2.2 Data Preprocessing

We downloaded the data from the NCBI Gene Expression Omnibus [CITE https://www.ncbi.nlm.nih.gov/geo/GSE181989]. We used the python package Scanpy[CITE THIS] to filter out all genes that [INSERT GENE THRESHOLD] and all cells that [INSERT CELL THRESHOLD]. After filtering, we had [NUMBER] cells with [NUMBER] genes. [NUMBER] of these cells were from the two AA patients and [NUMBER] of these cells were from the healthy controls. DID WE DO ANY IMPUTATION HERE?

2.3 Principal Component and UMAP Visualizations

We used the Scanpy python package 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).

2.4 Linear Discriminant Analysis

We used the Scanpy python package 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.
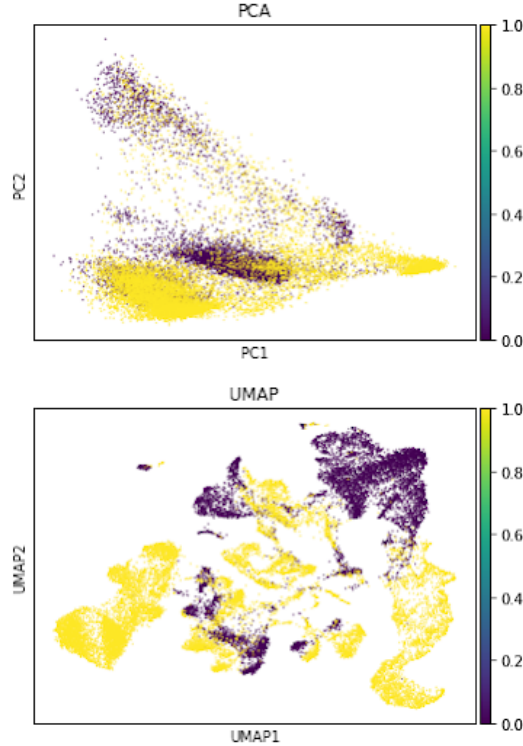
Figure 1: Principal Component and UMAP Projection visualizations colored by disease status. 0 is healthy, 1 is AA.

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.

2.5   scVI

We used the scvi-tools python package to perform scVI (CITE: https://docs.scvi-tools.org/en/stable/index.html) (Lopez et al.   2018).   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 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.
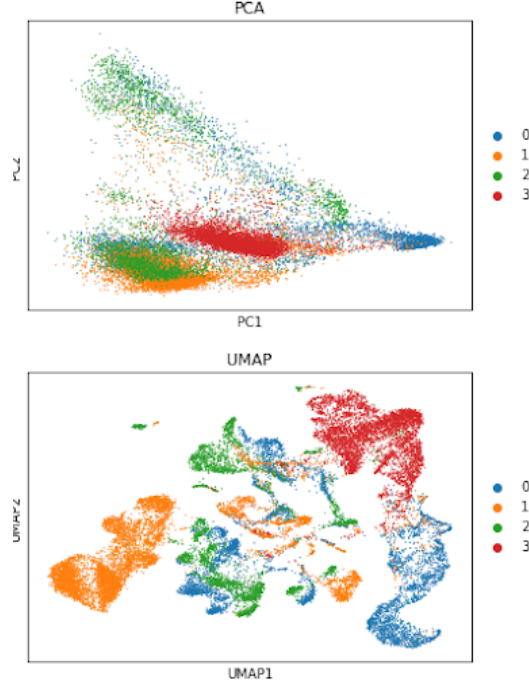
Figure 2: Principal Component and UMAP Projection visualizations colored by individual. Individuals 0 and 1 have AA. Individuals 2 and 3 are healthy controls.

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.

4

4   Conclusion

References

Bacigalupo, Andrea. 2017. "How I treat acquired aplastic anemia | Blood | American Society of Hematology." *Blood* (March 16, 2017). Accessed October 20, 2022. https://doi.org/10.1182/blood-2016-08-693481. https://ashpublications.org/blood/article/129/11/1428/35931/How-I-treat-acquired-aplastic-anemia.

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.

Scheinberg, Phillip, Olga Nunez, Barbara Weinstein, Priscila Scheinberg, Colin O. Wu, and Neal S. Young. 2012. "Activity of alemtuzumab monotherapy in treatment-naive, relapsed, and refractory severe acquired aplastic anemia." *Blood* 119, no. 2 (January 12, 2012): 345–354. ISSN: 0006-4971, accessed October 20, 2022. https://doi.org/10.1182/blood-2011-05-352328. https://doi.org/10.1182/blood-2011-05-352328.

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/.

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/.