Pages: 1–10

Stratification of Systemic Lupus Erythematosus Patients Using Gene Expression Data to Reveal Expression of Distinct Immune Pathways

Aditi Deokar ADEOKAR@BU.EDU
Boston University Academy
Boston, MA

Abstract

Systemic lupus erythematosus (SLE) is the tenth leading cause of death in females 15-24 years old in the US. The diversity of symptoms and immune pathways expressed in SLE patients causes difficulties in treating SLE as well as in new clinical trials. This study used unsupervised learning on gene expression data from adult SLE patients to separate patients into clusters. The dimensionality of the gene expression data was reduced by three separate methods (PCA, UMAP, and a simple linear autoencoder) and the results from each of these methods were used to separate patients into six clusters with k-means clustering.

The clusters revealed three separate immune pathways in the SLE patients that caused SLE. These pathways were: (1) high interferon levels, (2) high autoantibody levels, and (3) dysregulation of the mitochondrial apoptosis pathway. The first two pathways have been extensively studied in SLE. However, mitochondrial apoptosis has not been investigated before to the best of our knowledge as a standalone cause of SLE, independent of autoantibody production, indicating that mitochondrial proteins could lead to a new set of therapeutic targets for SLE in future research.

Keywords: systemic lupus erythematosus, gene expression, unsupervised learning, interferon, autoantibody, apoptosis

1. Introduction

Systemic lupus erythematosus (SLE) is the tenth most common cause of death among females 15-24 years old in the US (Yen and Singh, 2018). SLE is one of many autoimmune diseases, which are diseases in which a patient's immune system mistakes parts of their own body as foreign, attacking their healthy organs and tissue (Lupus Foundation of America, 2020).

SLE can be driven by defects in the innate immune system and/or the adaptive immune system. SLE patients are often characterized by high levels of interferon-1, which causes inflammation in the innate immune system in response to viruses. In SLE, high interferon levels can be caused by a variety of factors, such as neutrophil extracellular traps (Bengtsson and Rönnblom, 2017). Most SLE patients also have high levels of autoantibodies, which are antibodies directed against self cells and are created by mature B cells (plasma cells) (Dema and Charles, 2016). Autoantibodies cause a much more targeted response than the innate immune system, but SLE patients can have a wide range of autoantibodies - one study found over 180 autoantibodies expressed in SLE patients (Yaniv et al., 2015). Some patients with lupus do not even have autoantibodies, and many of the autoantibodies in SLE are also found in other rheumatic diseases (Egner, 2000).

© A. Deokar.

The heterogeneity of lupus symptoms and immune pathways affected makes it difficult to treat, because different drugs work well on different patients. Merrill et al. (2017) found that certain standard drugs (anti-rheumatic drugs and immunosuppressants) affect immune pathways differently in interferon-low and interferon-high patients. While there is still debate on whether SLE is one disease or many (Agmon-Levin et al., 2012), it is clear that subdividing SLE patients into categories will help treat patients.

Previous studies have tackled this problem by dividing patients based on antibody levels (Artim-Esen et al., 2014), gene expression (Toro-Domínguez et al., 2018), and immune molecule levels (Hamilton et al., 2018). However, none of these studies have reached a consensus on the best subdivision of SLE. Guthridge et al. (2020) used all three of these factors to divide SLE patients into seven clusters with unsupervised machine learning. But, in practice, gathering these different types of patient data to categorize a patient into an SLE subdivision is infeasible.

In this study, we use unsupervised machine learning to categorize SLE patients using only gene expression data, which is more accessible than all three types of data combined. This would help determine if gene expression data alone reveals similar patterns in immune pathway expression as does its combination with antibody levels and immune molecule levels.

2. Data and Methods

We use a gene expression dataset available on GEO (accession number GSE138458) containing data collected by Guthridge et al. (2020). The data includes 336 samples in total, with 24 control patients and 198 SLE patients. 108 of the SLE patients have two or more samples taken. Data pre-normalized by Guthridge et al. (2020) is used, employing bgAdjust background correction, vst variance stabilizing transformation, and rank invariant normalization, and outlier removal (1 control and 5 SLE).

Given the high dimensionality nature of gene expression data, dimension reduction techniques are required. Prior work has used unsupervised learning following dimension reduction of gene expression data for other diseases such as cancer (Shi and Luo, 2010). The dimensionality of our 47,323 gene data is reduced using three separate methods: *Principal Component Analysis* (PCA), *Uniform Manifold Approximation and Projection* (UMAP), and a simple *autoencoder* (AE), which are intended to minimize the effects of random variation on the unsupervised clustering model in different ways.

With both PCA and UMAP, 200 reduced features are selected. In PCA, these explain 96.29% of the variance in the original 47,323 genes. UMAP is a nonlinear model (unlike PCA) similar to t-SNE, used for visualization as well as nonlinear dimension reduction (McInnes et al., 2020). The autoencoder aims to reduce the loss of information between the original inputs (genes) and the decoded output of the same dimension. AEs with both linear and sigmoid activation functions are validated and the linear AE is found to perform much better after 100 epochs (validation loss 0.068) than the sigmoid AE (validation loss 48.28). So, we select the 1000 encoded components from the linear AE for subsequent clustering.

The three datasets with reduced features are then used for k-means clustering. To determine the best number of clusters, we use distortion score, silhouette score, and Calinski-

Harabasz score. All these metrics are found to converge on 6 clusters for each dataset. k-means clustering is then used to derive 6 clusters from each dataset.

For visualization and interpretation of the clusters, we use 27 pre-existing modules created by Chaussabel et al. (2008). Each module represents a group of genes with a common function. These are used to calculate module scores for the three datasets. Module scores for each cluster represent the percentage of genes in each module that were significantly upregulated (i.e., overexpressed) or downregulated (i.e., underexpressed) in that cluster as compared to the controls, based on a two-tailed t-test ($p < 0.05$).

3. Results and Discussion

Figure 1 shows heatmaps generated from the module scores for the 3 feature-reduced datasets. These heatmaps show the percentage of underexpressed (brown) or overexpressed (purple) genes for SLE patients as compared to the controls.

The clusters originating from the PCA and UMAP dimensionality reductions show very similar patterns in the upregulated and downregulated modules, while the clusters originating from the AE mostly show a consistent level of increased or decreased gene expression across all modules, excepting cluster 5.

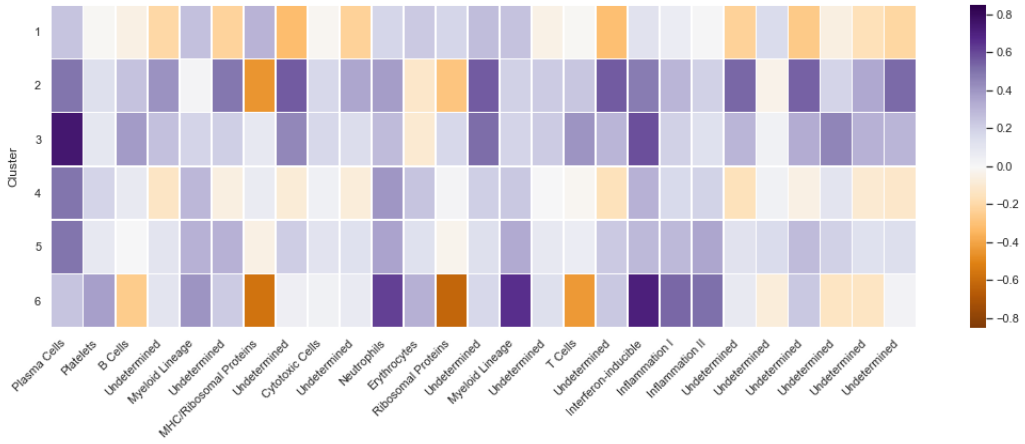
The patients in the clusters created from the PCA and UMAP dimensionality reduction techniques and AE cluster 5 can be designated as belonging to one of three groups: (a) interferon-driven SLE, (b) autoantibody-driven SLE, and (c) SLE caused by mitochondrial apoptosis. The first two groups substantiate results from prior literature, while the third group presents a pathway that suggests a novel cause of SLE.

3.1. Interferon-driven SLE

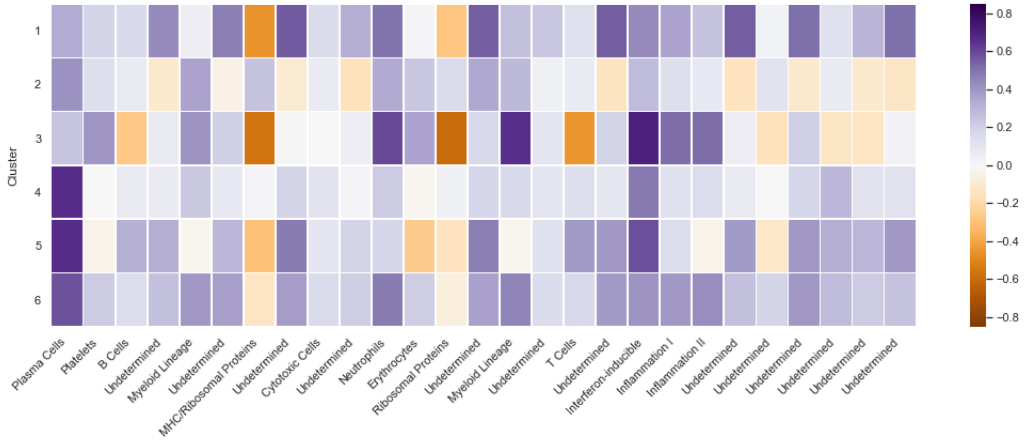
In lupus, type 1 interferon levels are often elevated, which can lead to inflammation and tissue damage caused by the innate immune system (Crow, 2014). PCA cluster 6 and UMAP cluster 3 in Figure 1 both display substantial upregulation of genes related to interferons and inflammation. These two clusters validate the patterns also observed in Guthridge et al. (2020)'s clusters 1, 4 and 6. All of these upregulated genes are related to the innate immune response. Those patients also have underexpressed B and T cells and normal expression of plasma cells, which would all be overexpressed if production of autoantibodies by plasma cells was the main reason for autoimmunity, rather than interferon levels.

3.2. Autoantibody-driven SLE

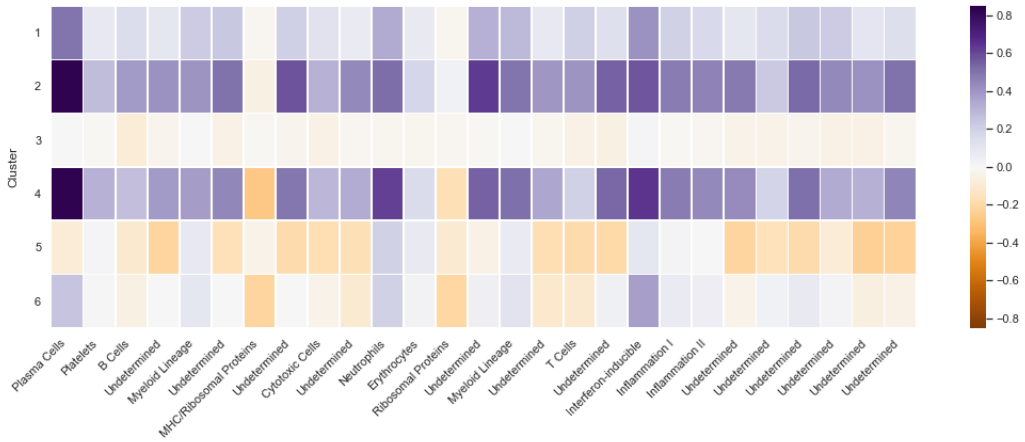
Many of the other PCA and UMAP clusters displayed upregulation of antibody-producing plasma cells; particularly PCA clusters 2, 3, 4 and 5 and UMAP clusters 4, 5, and 6. Guthridge et al. (2020) observed a similar trend, where their clusters 2, 3, and 5 had higher T cell, B cell, and plasma cell related expression. While autoantibodies are known to be common in SLE, the diversity of autoantibodies (as discussed in Yaniv et al. (2015)) means that there is still work to be done understanding what is different among the antibodies produced in these four PCA clusters and three UMAP clusters. Some of these differences might come from genes used to create the clusters that were not included in the modules used for the heatmap visualization.



(a) PCA heatmap



(b) UMAP heatmap



(c) Autoencoder heatmap

Figure 1: Comparative gene underexpression or overexpression for gene expression modules across clusters.

Brant et al. (2020), who grouped lupus patients based on their correlation between gene expression and disease activity, found one cluster where neutrophil levels correlated to disease activity and one where lymphocyte levels correlated to disease activity. Since neutrophil extracellular traps are one way that interferon levels become elevated, their neutrophil-correlated group might correspond to our high-interferon group, and their lymphocyte-correlated group might correspond to our antibody-driven group. More analysis should be done on disease activity correlation in our data to confirm this.

3.3. SLE caused by mitochondrial apoptosis

PCA clusters 1 and 4, UMAP cluster 2, and Autoencoder cluster 5 display a different pattern from many of the other clusters. In these clusters, many of the modules labeled as *Undetermined* by Chaussabel et al. (2008) were underexpressed. A closer look at the genes in these *Undetermined* modules reveals that they include mitochondrial ribosomal proteins, mitochondrial elongation factors, and proteins in the cAMP-signaling pathway. Mitochondrial ribosomal proteins, in addition to their ribosomal functions, are involved in apoptotic (programmed cell death) pathways (Kim et al., 2017), and cAMP signaling regulates mitochondrial apoptosis (Valsecchi et al., 2013). Apoptosis is known to be a factor in SLE, but mainly because ineffective clearance of apoptotic cells can expose B and T cells to intracellular material, leading to the creation of autoantibodies against this intracellular material (Mevorach, 2003).

We suggest that for the patients in these clusters, dysregulation of mitochondrial pathways or signaling from outside molecules, possibly lymphocytes, could cause mitochondrial apoptotic pathways to become activated in healthy cells, destroying healthy cells as is characteristic of SLE. These healthy cells would have a range of gene expression of mitochondrial proteins, but the cells with higher expression of the proteins would activate the apoptotic pathway and die. Only cells with lower expression levels would survive, so lower expression levels were found in our study. These lower expression levels would also impair mitochondrial functions, which has been observed to be true in SLE patients (Leishangthem et al., 2016).

Cluster 7 from the Guthridge et al. (2020) study also had low expression of mitochondrial respiration and mitochondrial stress genes (not discussed in their study). The discovery of this cluster of patients using two completely different machine learning approaches corroborates the idea that the mitochondrial apoptotic pathway is a novel cause for SLE. Future studies should investigate to a further extent the mitochondrial apoptotic pathway in SLE patients as a reason for destruction of self cells in addition to a way that autoantibodies are produced.

4. Conclusion and Future Work

In this study, we separated SLE patients into clusters based on their gene expression data using unsupervised learning. The data was collected by Guthridge et al. (2020), who clustered patients using antibody levels and immune phenotyping in addition to gene expression levels. We used only gene expression data and used entirely different methods from their study, in order to determine whether we would find similar clusters of patients. The dimensionality of the gene expression data was first reduced by three separate methods (PCA,

UMAP, and a simple linear autoencoder) and the results from each of these methods were used to separate patients into six clusters with k-means clustering. These clusters revealed there were three separate immune pathways in the SLE patients causing SLE. These pathways were 1) high interferon levels, 2) high autoantibody levels, and 3) dysregulation of the mitochondrial apoptosis pathway. All three of these pathways were present in Guthridge et al. (2020)'s clusters, but to our knowledge this study is the first to propose mitochondrial apoptosis as a standalone cause of SLE, independent of autoantibody production. Future studies should investigate to a further extent the mitochondrial apoptotic pathway in SLE patients as a reason for destruction of self cells in addition to a way that autoantibodies are produced and investigate mitochondrial proteins as possible therapeutic targets for SLE.

References

N. Agmon-Levin, M. Mosca, M. Petri, and Y. Shoenfeld. Systemic lupus erythematosus one disease or many? *Autoimmunity Reviews*, 11(8):593–595, 2012. ISSN 15689972. doi: 10.1016/j.autrev.2011.10.020. URL http://dx.doi.org/10.1016/j.autrev.2011.10.020.

Bahar Artim-Esen, Erhan Çene, Yasemin Şahinkaya, Semra Ertan, Özlem Pehlivan, Sevil Kamali, Ahmet Gül, Lale Öcal, Orhan Aral, and Murat Inanç. Cluster analysis of autoantibodies in 852 patients with systemic lupus erythematosus from a single center. *Journal of Rheumatology*, 41(7):1304–1310, 2014. ISSN 14992752. doi: 10.3899/jrheum.130984.

Anders A. Bengtsson and Lars Rönnblom. Role of interferons in SLE. *Best Practice and Research: Clinical Rheumatology*, 31(3):415–428, 2017. ISSN 15321770. doi: 10.1016/j.berh.2017.10.003.

Elizabeth J. Brant, Edward A. Rietman, Giannoula Lakka Klement, Marco Cavaglia, and Jack A. Tuszynski. Personalized therapy design for systemic lupus erythematosus based on the analysis of protein-protein interaction networks. *PLoS ONE*, 15(3):1–16, 2020. ISSN 19326203. doi: 10.1371/journal.pone.0226883. URL http://dx.doi.org/10.1371/journal.pone.0226883.

Damien Chaussabel, Charles Quinn, Jing Shen, Pinakeen Patel, Casey Glaser, Nicole Baldwin, Dorothee Stichweh, Derek Blankenship, Lei Li, Indira Munagala, Lynda Bennett, Florence Allantaz, Asuncion Mejias, Monica Ardura, Ellen Kaizer, Laurence Monnet, Windy Allman, Henry Randall, Diane Johnson, Aimee Lanier, Marilynn Punaro, Knut M. Wittkowski, Perrin White, Joseph Fay, Goran Klintmalm, Octavio Ramilo, A. Karolina Palucka, Jacques Banchereau, and Virginia Pascual. A Modular Analysis Framework for Blood Genomics Studies: Application to Systemic Lupus Erythematosus. *Immunity*, 29(1):150–164, 2008. ISSN 10747613. doi: 10.1016/j.immuni.2008.05.012.

Mary K. Crow. Type I Interferon in the Pathogenesis of Lupus. *The Journal of Immunology*, 192(12):5459–5468, 2014. ISSN 0022-1767. doi: 10.4049/jimmunol.1002795.

Barbara Dema and Nicolas Charles. Autoantibodies in SLE: Specificities, Isotypes and Receptors. *Antibodies*, 5(1):2, 2016. ISSN 2073-4468. doi: 10.3390/antib5010002.

William Egner. The use of laboratory tests in the diagnosis of SLE. *Journal of Clinical Pathology*, 53(6):424–432, 2000. ISSN 00219746. doi: 10.1136/jcp.53.6.424.

Joel M. Guthridge, Rufei Lu, Ly Thi Hai Tran, Cristina Arriens, Teresa Aberle, Stan Kamp, Melissa E. Munroe, Nicolas Dominguez, Timothy Gross, Wade DeJager, Susan R. Macwana, Rebecka L. Bourn, Stephen Apel, Aikaterini Thanou, Hua Chen, Eliza F. Chakravarty, Joan T. Merrill, and Judith A. James. Adults with systemic lupus exhibit distinct molecular phenotypes in a cross-sectional study. *EClinicalMedicine*, 20:100291, 2020. ISSN 25895370. doi: 10.1016/j.eclinm.2020.100291. URL https://doi.org/10.1016/j.eclinm.2020.100291.

Jennie A. Hamilton, Qi Wu, PingAr Yang, Bao Luo, Shanrun Liu, Jun Li, Alexa L. Mattheyses, Ignacio Sanz, W. Winn Chatham, Hui-Chen Hsu, and John D. Mountz. Cutting Edge: Intracellular IFN-β and Distinct Type I IFN Expression Patterns in Circulating Systemic Lupus Erythematosus B Cells. *The Journal of Immunology*, 201(8):2203–2208, 2018. ISSN 0022-1767. doi: 10.4049/jimmunol.1800791.

Brian Kegerreis, Michelle D. Catalina, Prathyusha Bachali, Nicholas S. Geraci, Adam C. Labonte, Chen Zeng, Nathaniel Stearrett, Keith A. Crandall, Peter E. Lipsky, and Amrie C. Grammer. Machine learning approaches to predict lupus disease activity from gene expression data. *Scientific Reports*, 9(1):1–12, 2019. ISSN 20452322. doi: 10.1038/s41598-019-45989-0. URL http://dx.doi.org/10.1038/s41598-019-45989-0.

Hyun-Jung Kim, Priyanka Maiti, and Antoni Barrientos. Mitochondrial ribosomes in cancer. *Seminars in Cancer Biology*, 47(3):67–81, dec 2017. ISSN 1044579X. doi: 10.1016/j.semcancer.2017.04.004. URL https://linkinghub.elsevier.com/retrieve/pii/S1044579X17300962.

B. D. Leishangthem, A. Sharma, and Archana Bhatnagar. Role of altered mitochondria functions in the pathogenesis of systemic lupus erythematosus. *Lupus*, 25(3):272–281, 2016. ISSN 14770962. doi: 10.1177/0961203315605370.

Lupus Foundation of America. What is lupus?, 2020. URL https://www.lupus.org/resources/what-is-lupus.

Leland McInnes, John Healy, and James Melville. UMAP: Uniform Manifold Approximation and Projection for Dimension Reduction, 2020. URL http://arxiv.org/abs/1802.03426.

Joan T. Merrill, Fred Immermann, Maryann Whitley, Tianhui Zhou, Andrew Hill, Margot O'Toole, Padmalatha Reddy, Marek Honczarenko, Aikaterini Thanou, Joe Rawdon, Joel M. Guthridge, Judith A. James, and Sudhakar Sridharan. The Biomarkers of Lupus Disease Study: A Bold Approach May Mitigate Interference of Background Immunosuppressants in Clinical Trials. *Arthritis and Rheumatology*, 69(6):1257–1266, 2017. ISSN 23265205. doi: 10.1002/art.40086.

Dror Mevorach. Systemic Lupus Erythematosus and Apoptosis. *Clinical reviews in allergy & immunology*, 25:49–59, 2003.

Michelle Petri, Wei Fu, Ann Ranger, Norm Allaire, Patrick Cullen, Laurence S. Magder, and Yuji Zhang. Association between changes in gene signatures expression and disease activity among patients with systemic lupus erythematosus. *BMC Medical Genomics*, 12 (1):1–9, 2019. ISSN 17558794. doi: 10.1186/s12920-018-0468-1.

Jinlong Shi and Zhigang Luo. Nonlinear dimensionality reduction of gene expression data for visualization and clustering analysis of cancer tissue samples. *Computers in Biology and Medicine*, 40(8):723 – 732, 2010. ISSN 0010-4825. doi: https://doi.org/10.1016/j.compbiomed.2010.06.007. URL http://www.sciencedirect.com/science/article/pii/S0010482510000958.

Daniel Toro-Domínguez, Jordi Martorell-Marugán, Daniel Goldman, Michelle Petri, Pedro Carmona-Sáez, and Marta E. Alarcón-Riquelme. Stratification of Systemic Lupus Erythematosus Patients Into Three Groups of Disease Activity Progression According to Longitudinal Gene Expression. *Arthritis and Rheumatology*, 70(12):2025–2035, 2018. ISSN 23265205. doi: 10.1002/art.40653.

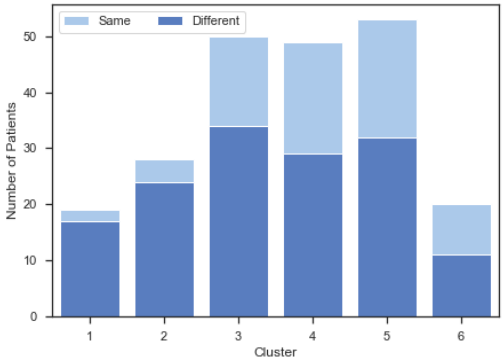
Federica Valsecchi, Lavoisier S. Ramos-Espiritu, Jochen Buck, Lonny R. Levin, and Giovanni Manfredi. cAMP and mitochondria. *Physiology*, 28(3):199–209, 2013. ISSN 15489213. doi: 10.1152/physiol.00004.2013.

Gal Yaniv, Gilad Twig, Dana Ben Ami Shor, Ariel Furer, Yaniv Sherer, Oshry Mozes, Orna Komisar, Einat Slonimsky, Eyal Klang, Eyal Lotan, Mike Welt, Ibrahim Marai, Avi Shina, Howard Amital, and Yehuda Shoenfeld. A volcanic explosion of autoantibodies in systemic lupus erythematosus: A diversity of 180 different antibodies found in SLE patients. *Autoimmunity Reviews*, 14(1):75–79, 2015. ISSN 18730183. doi: 10.1016/j.autrev.2014.10.003. URL http://dx.doi.org/10.1016/j.autrev.2014.10.003.

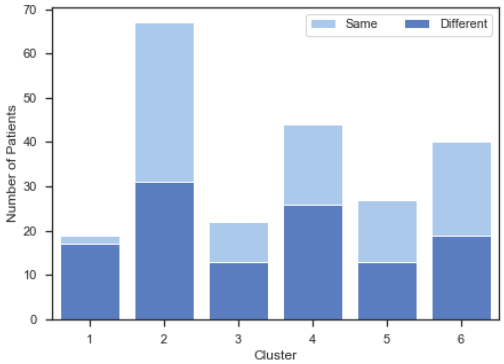
Eric Y Yen and Ram R Singh. Lupus – An Unrecognized Leading Cause of Death in Young Women: Population-based Study Using Nationwide Death Certificates, 2000–2015. *Arthritis and Rheumatology*, 70(8):1251–1255, 2018. doi: 10.1002/art.40512.

Appendix A. Analyzing Patients with Multiple Samples

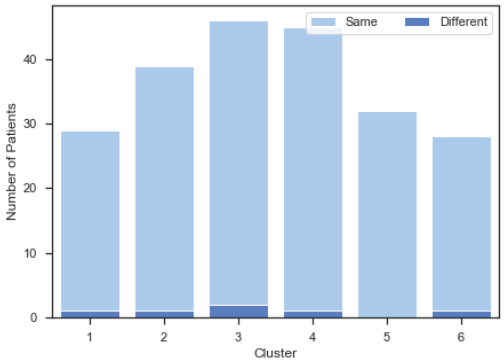
For patients who had multiple samples taken, k-means following the autoencoder classified them into the same cluster 97.3% of the time, while k-means following PCA and UMAP classified them into the same cluster 32.9% and 45.7% of the time respectively (Figure 2). While gene expression data is correlated with SLE disease activity (Kegerreis et al., 2019; Toro-Domínguez et al., 2018), Petri et al. (2019) found that the majority of gene expression signatures were stable in patients over time. This suggests that the autoencoder's dimensionality reduction may have emphasized the stable gene expression signatures, causing them to be a major factor in the clustering, but that PCA and UMAP, which aimed to preserve more of the variance in the data, did not maintain the data from genes whose expression was stable over time. Many of these more stable genes might not have been related to the immune system, so they were not included in Chaussabel et al. (2008)'s coexpression modules. Thus, the more variable modules that were in the heatmap would have shown a



(a) Distribution in PCA clusters



(b) Distribution in UMAP clusters



(c) Distribution in Autoencoder clusters

Figure 2: Patients in each cluster, of those who had multiple samples taken, who were put in the same cluster or different cluster for (a) PCA clusters, (b) UMAP clusters, (c) Autoencoder clusters.

lot of variation between the patients in each cluster, causing the clusters in the autoencoder heatmap to show a more consistent level of expression across all genes in the modules. Further analysis should be done to determine the level of variation in gene expression in the modules for the autoencoder clusters in comparison to the PCA and UMAP clusters, and to determine whether the PCA and UMAP clusters correlated to disease activity more than the autoencoder clusters, which this idea would imply.