COVID-19 Severity Prediction from Genomic Expression

Emily Buckley (eb4ub), Tucker Cullen (stc7mg), Julia Pasco-Anderson (jgp2mk), Zack Thomas (zvt3fn)

1. Motivation

Since January 2020, the United States has reported over fifteen million COVID-19 cases with more than 285,000 resulting in death (CDC, 2020). With only one drug currently approved as a treatment and still so many deaths, new therapeutic options are needed. A better understanding of the pathophysiology of COVID-19 will allow for faster development of and more effective treatments. Thus, we aim to identify dysregulated genes caused by COVID-19. By identifying up-regulated and down-regulated genes in different levels of COVID-19 severity, we will be able to better understand the effect of COVID-19 and what genes are most involved. The identification of these dysregulated genes will aid in the development of treatments and gene therapies for COVID-19.

2. Background

With a death rate of 1.9% in the US and 1,562,000 people dead worldwide (Johns Hopkins University, 2020), COVID-19 has upended the world over the course of the last year. Although a few antiviral drugs, such as remdesivir, have been approved for use against COVID-19, the world still lacks a consistently effective treatment (Harvard Health, 2020). Furthermore, the virus has been found to impact different people in starkly different ways. Some people face horrible illness resulting in death, while others — an estimated 20% of COVID-19 patients (Citroner, 2020) — remain entirely asymptomatic. The explanation behind this phenomenon is unknown, but some progress has been made in identifying potential factors.

3. Related Work

3.1 A study done recently at the University of Washington found that viral load, age, and sex are all contributing factors to the severity of an individual's COVID-19 case (Lieberman et al., 2020). They sequenced the RNA of 430 SARS-CoV-2 infected patients and 54 negative controls and the differential gene expression based on a model using the negative binomial distribution. Results of the experiment show that SARS-CoV-2, the virus that causes COVID-19, leads to a significant interferon-driven antiviral response and decreased transcription of ribosomal proteins. Higher viral load corresponded with increased expression of interferon-responsive genes like ACE2 and decreased presence of B cells and neutrophils. Age had an effect on the levels of expressed chemokines and their associated genes, and females had increased B and NK cell-specific transcripts and decreased NFkB inhibitors compared to males (Lieberman et al., 2020).

3.2 Another study, conducted by researchers at Duke University, identified dysregulated gene expression unique to samples infected with SARS-CoV-2 when compared to negative controls, as well as samples infected with influenza, bacterial pneumonia, and seasonal coronavirus. A regression model based on the interferon-stimulated gene-driven panviral signature provides the possibility for early detection of COVID-19, influenza, and seasonal coronavirus. The upregulation of B-cell activation and Immunoglobulin genes in early symptomatic COVID-19 cases provided the basis for an improved gene expression signature to distinguish SARS-CoV-2, seasonal coronavirus, influenza, and bacterial infections (McClain et al., 2020). A Berlin study identified increased activated HLA-DR^{hi}CD11c^{hi}CD14^{+}

monocytes in patients with mild COVID-19 and neutrophil precursors and dysregulated myeloid cell responses in patients with severe COVID-19 infections (Schulte-Schrepping et al., 2020).

4. Claim/Target Task

Our goal is to identify the genes implicated in the human body's response to COVID-19. We will be applying machine learning methods to a RNA sequencing dataset generated by Lieberman et al. (Lieberman et al., 2020). Shotgun RNA sequencing was conducted on nasopharyngeal swabs collected from 430 SARS-CoV-2 positive individuals and 54 negative controls. Metadata from each individual is also available, including N1 cycle threshold (Ct) values that were obtained from RT-PCR SARS-CoV-2 tests. These Ct values can be used to categorize each individual's viral load as "low", "medium", or "high".

5. Proposed Solution

Over 35,000 genes are included in the dataset, though it is likely that most are not relevant to SARS-CoV-2. Thus the first step in our analysis will be to reduce the dimensionality of the data. For this purpose, we will employ techniques such as principal component analysis (PCA), T-distributed stochastic neighbor embedding (tSNE), and uniform manifold approximation and projection (UMAP) (Siva, 2020). We will then use least absolute shrinkage and selection operator (LASSO) regression in order to determine what genes (features) are most important in predicting SARS-CoV-2 viral load. LASSO uses the L1 norm as a regularizer, which - unlike the L2 norm used in ridge regression - causes the regression coefficients for unimportant features to go to zero (Nagpal, 2017). This essentially removes unimportant variables from the model, which is favorable for this application since many irrelevant genes are included in the dataset. The coefficients learned through LASSO regression will show us which genes most closely correlate with increased viral load. These findings will aid in understanding the mechanisms behind COVID-19 and could potentially lead to new treatment pathways. Our proposed workflow is outlined in *Figure 1*.

6. Intuitive Figure

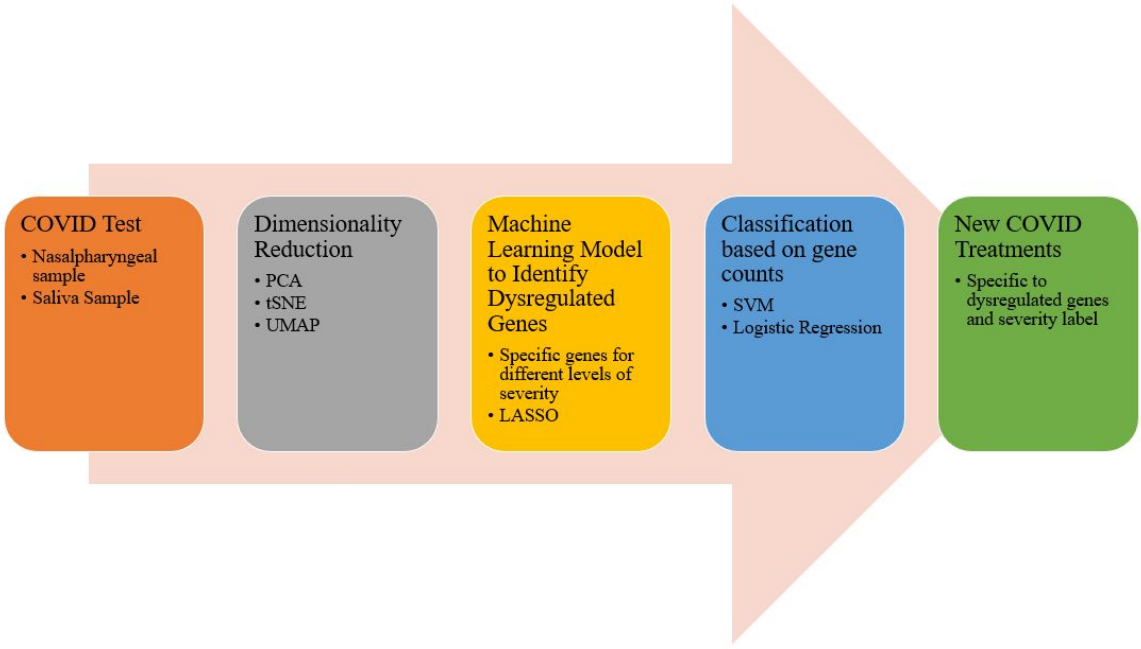

Figure 1. Workflow for determining dysregulated genes. First, a COVID-19 test will be performed. Next, the gene count data will be reduced and put through a well-trained model to predict severity. Finally, appropriate treatment will be determined based on the results.

7. Implementation

First, the data from the Lieberman et al. study was preprocessed and normalized, utilizing Pandas and log normalization, respectively. Preprocessing involved appending the cycle threshold values to the gene expression matrix and assigning a severity label to each sample based on the range in which the cycle threshold value fell, as defined in the Lieberman et al. study. Samples with unknown cycle threshold values were removed. DESeq normalization was attempted, but ultimately log normalization was used. Next, PCA was utilized for dimensionality reduction and identification of the most influential genes. PCA was conducted for all samples and then for positive samples only. tSNE and UMAP methods were then conducted to visualize clustering of the data. Like PCA, tSNE and UMAP were conducted on the dataset as a whole and then on just the COVID-19-positive samples. For each of tSNE and UMAP, several different parameters were tested so that those yielding the clearest clustering could be identified. For tSNE, perplexity and distance metrics were tested. For UMAP, distance metrics and neighbor numbers were tested. LASSO and linear regression were conducted to model viral load and identify the most predictive genes. Variance Threshold feature selection was utilized prior to running LASSO and linear regression in order to further reduce dataset size and emphasize the most influential genes. Elastic Net Regularization was attempted as an alternative to LASSO. Finally, SVM and logistic regression (in conjunction with PCA) were utilized to determine classification accuracy based on gene counts. Sklearn packages were utilized for PCA, tSNE, UMAP, LASSO, Elastic Net, VarianceThreshold, linear regression, SVC, and logistic regression. Matplotlib and seaborn provided the plots in this project. Numpy and Pandas were used throughout for data organization.

8. Data Summary

To generate the data, RNA metagenomic sequencing was conducted on 484 individuals (Lieberman et al., 2020). 430 of these individuals tested positive for Sars-CoV-2, while the other 54 serve as controls. The PCR test results for the positive individuals are reported alongside each sample. These results take the form of cycle threshold values. Ct values indicate the number of cycles it takes to amplify DNA to detectable levels when conducting a polymerase chain reaction (PCR). A lower number signifies that it takes less time to achieve a detectable amount of viral DNA, meaning viral load (the amount of COVID-19 virus present in the body) is higher. These values were used to categorize the positive patients into three sub-categories: high, medium, and low viral load. An example of how the data is structured can be seen in *Figure 2*.

Patient ID	Raw Gene Counts							Ct	Viral Load
	A1BG	ABL2	ACE2	ACYP1	ADAD1	ACTR1	...		
POS_001	0	0	88	3	9	1	...	24.18	Low
POS_002	0	3	6	15	12	0	...	21.12	Medium
POS_003	0	4	70	3	2	1	...	18.09	High
NEG_001	0	0	0	0	2	1	...	n/a	Neg
...	...	...	...	...	...	...	...		...

Figure 2. Shows a sample of the dataset structure, with samples in the rows and raw gene counts in each column. Ct values and the corresponding categories are contained in additional columns.

9. Experimental Results

PCA was used both as a dimensionality reduction technique and a visualization technique. PCA identified the first 100 principal components as covering 77% of the variance. However, as can be seen in *Figure 3*, the first component accounts for about 40% of the total variance. The top 10 gene loadings for the first two components are reported in *Figure 4*. Finally, the first two components were used to visualize the data, as displayed in *Figure 5*. General clustering between positive and negative patient samples was observed.

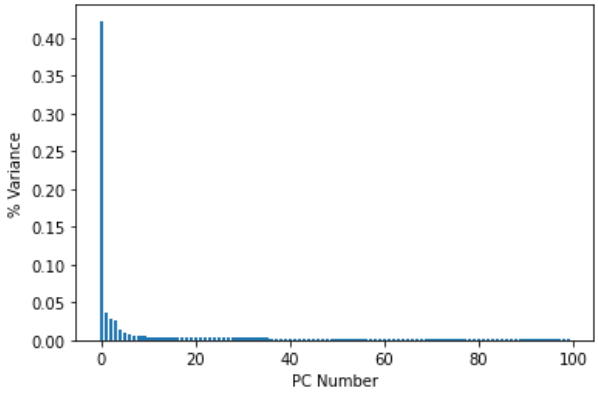

Figure 3. PCA variance coverage. Principal Components 1-100 and the percentage of variance they are each responsible for. PC1 is responsible for nearly a majority.

SMG1P3	0.014950	TREM1	0.039595
DNAH1	0.014439	SLC2A3	0.038223
DLEC1	0.014335	CXCR2	0.037308
JMJD1C	0.014108	IL1B	0.036995
UBE2H	0.013939	AQP9	0.036954
SMG1P1	0.013840	CR1	0.035688
DNAH5	0.013797	CSF3R	0.035568
NPIPB5	0.013769	NABP1	0.035395
DNAH10	0.013756	FFAR2	0.035182
CDHR3	0.013676	FPR2	0.035055

Figure 4. PCA gene loadings. The genes with the top 10 most significant loadings for PC1 (left) and PC2 (right).

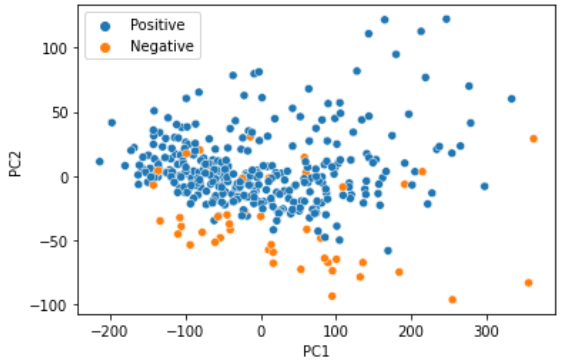

Figure 5. PCA plot. Plotting data against only PC1 and PC2 shows relatively clear clustering of positive and negative cases.

tSNE was used as an alternative to PCA to reduce the data to two dimensions for visualization. Of the distance metrics and perplexity values tested, the clearest clustering of positive patients versus negative patients occurred using a perplexity of 5 and distance metrics of Cosine and Correlation as seen in *Figure 6*. However, when testing the same set of hyperparameters, tSNE was not able to cluster just the positive patients by severity.

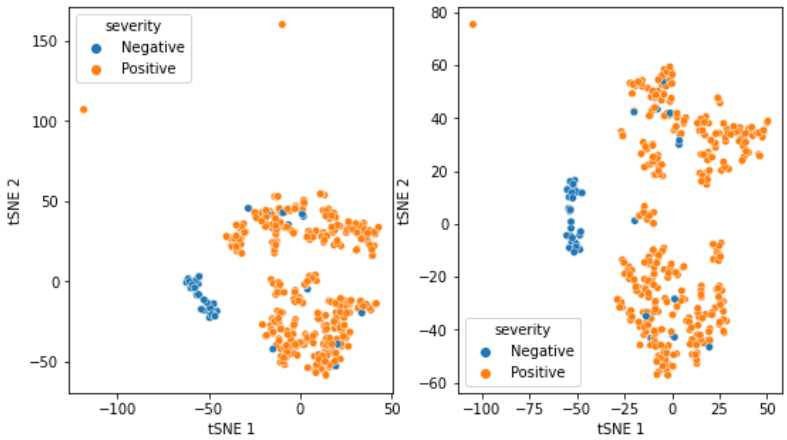

Figure 6. tSNE clustering positive and negative patients. While not a perfect separation, using a perplexity of 5 and distance metric of Cosine (left) or Correlation (right) clustered the two groups the best.

UMAP, much like tSNE, clustered positive versus negative fairly well but when applied to just the positive cases to cluster by severity, not distinct clusters formed as seen in *Figure 7*.

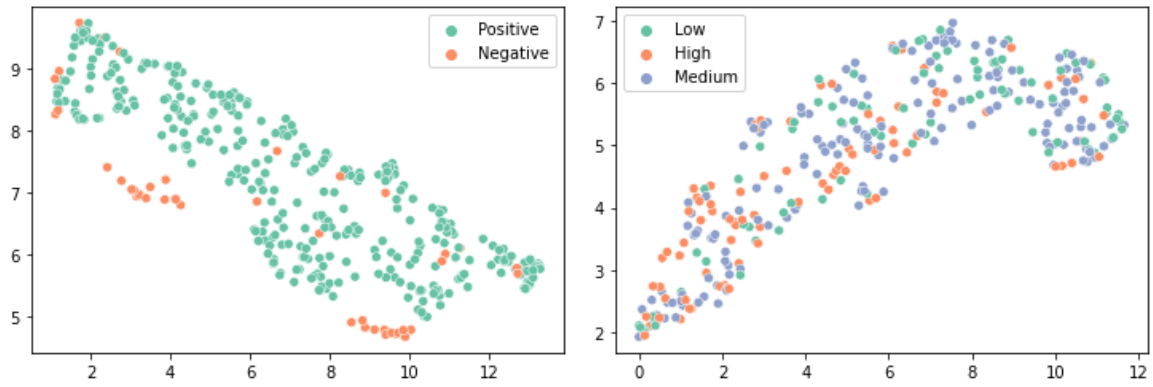

Figure 7. UMAP clustering. Distinct negative clustered can be seen when coloring for positive versus negative (left). However, no clear clusters formed when using the positive patients and coloring for severity (right).

LASSO Regression resulted in an R^2 value of 0.62. It was also used to identify the most important genes in the model. The genes with the ten largest magnitude beta coefficients are: PCSK5, CBLC, CCNB1, ZNF600, LRRC37A3, RPL7P19, KLK13, SNX10, LILRB4, and HDHD3. A plot of the LASSO predicted and actual Ct values for each of the testing samples can be seen in *Figure 8*.

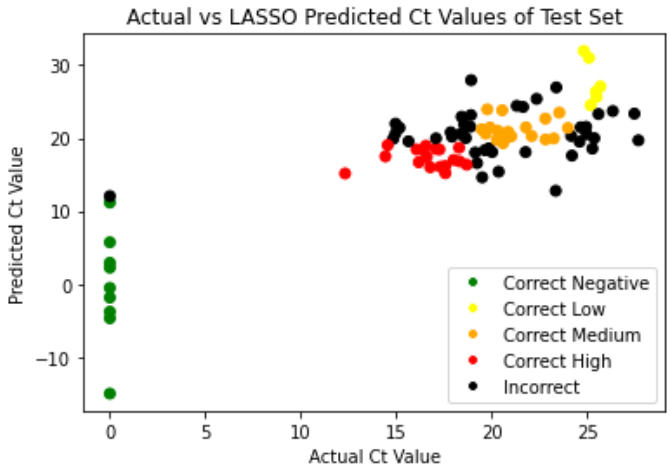

Figure 8. Shows the actual vs. LASSO predicted Ct values. The model fit the testing data with an R^2 value of 0.62. The testing samples that were incorrectly categorized by the model are shown in black.

Linear Regression was conducted in order to compare its performance with LASSO. Linear regression showed a similar fit on the test data as LASSO, with an R^2 value of 0.63.

Support Vector Classifier, like linear regression, was conducted to compare its performance with LASSO as little to no information about significant genes can be drawn from the SVC model. After testing multiple hyperparameters, the best accuracy obtained by the SVC was 0.63, very similarly to LASSO and linear regression. These results are displayed in *Figure 9*.

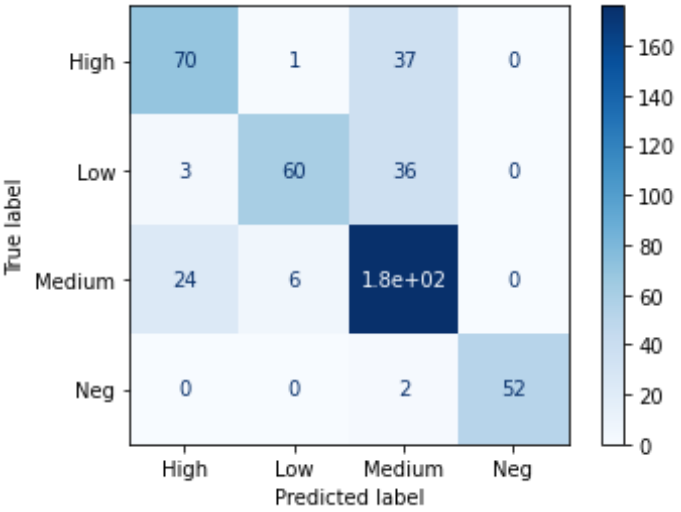

Figure 9. Confusion matrix for SVC model predictions
The model fit the testing data with an accuracy of 0.63.
Hyperparameters: {'kernel': 'linear', 'C': 0.001, 'degree': 1}

Logistic Regression was used in a similar manner to the Support Vector Classifier, to compare its performance to the regression models. After testing multiple hyperparameters, the best accuracy obtained by logistic regression was 0.65 and the classification results are displayed in *Figure 10*.

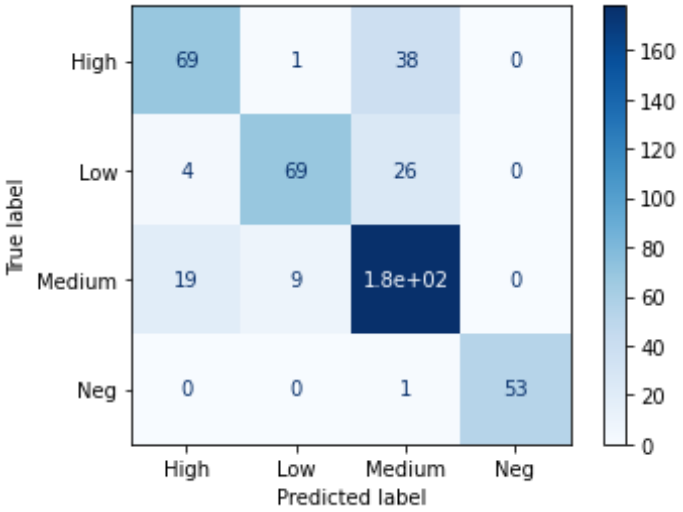

Figure 10. Confusion matrix for logistic regression model predictions The model fit the testing data with an accuracy of 0.65. Hyperparameters: {'penalty': 'l2', 'C': 0.001, 'solver': 'lbfgs'}

10. Experimental Analysis

The most significant findings in our analysis come from the LASSO regression step. Since LASSO regression uses the L1 norm as the regularizer, it performs feature selection in addition to regression (Nagpal, 2017). It does this by reducing the coefficients of unimportant features to zero, effectively eliminating them from the model. Likewise, LASSO assigns the largest magnitude coefficients to the most important features. The gene that was assigned the largest coefficient in our LASSO model is proprotein convertase subtilisin kexin 5 (PCSK5). PCSK5 has been linked to COVID-19 in other studies as well (Mick et al., 2020; Muus et al., 2020). More specifically, proprotein convertases like PCSK5 have been shown to play a role in coronavirus S-protein priming (Izaguirre, 2019; Millet & Whittaker, 2015). The fact that our model identified PCSK5 is promising. However, among the other genes with the top ten largest magnitude coefficients, no others have been linked to coronavirus in previous literature. Additionally, the genes identified in the Lieberman et al. paper - where we acquired the dataset - were not identified as important by our model (their coefficients were reduced to zero by LASSO) (Lieberman et al., 2020). This is contrary to our expectations, and could be due to having too many genes in the data, causing these particular genes to get overly suppressed in the model.

While the classification accuracies of our SVC and logistic regression models were not exceptional (at 63% and 65% respectively), they do show that it is possible to classify COVID-19 severity based on gene expression. Additionally, the models had particularly good success in distinguishing negative patients from positive patients (as can be seen in the confusion matrices in *Figures 9 and 10*). The models struggled, however, when classifying positive patients into sub-categories. This observation is

backed up by our tSNE and UMAP visualizations, which show clear clusters between positive and negative COVID-19 patients, but no clustering within the positive patients.

11. Conclusion and Future Work

The LASSO model can be used to identify critical COVID-19 genes such as PCSK5 via feature selection. However, this method is quite limited as it did not give any weight to genes that were marked as differentially expressed by Lieberman et al. Additionally, classifiers such as SVC and logistic regression can be used so somewhat accurately (0.63) classify COVID-19 severity based on gene expression.

Future work includes using the subset of differentially expressed genes as identified by Lieberman et al. to train the different models (LASSO, linear regression, SVC, logistic regression) and testing the accuracy based on those genes. Additionally, in this study, feature selection was only done with LASSO, whereas PCA was applied to the data prior to training the SVC and logistic regression models. Therefore, using the set of genes identified by feature selection to train the classifiers rather than PCA could potentially yield increased accuracy. Finally, the methods described in this manuscript could be applied to any raw gene count dataset to identify genes of interest in dysregulated states.

REFERENCES

CDC. (2020, March 28). *Coronavirus Disease 2019 (COVID-19) in the U.S.* Centers for Disease Control and Prevention. https://www.cdc.gov/covid-data-tracker

Citroner, G. (2020, September 22). *1 in 5 COVID-19 Cases Are Asymptomatic but Can Spread the Disease*. Healthline. https://www.healthline.com/health-news/20-percent-of-people-with-covid-19-are-asymptomatic-but-can-spread-the-disease

Harvard Health. (2020). *Treatments for COVID-19*. Harvard Health. https://www.health.harvard.edu/diseases-and-conditions/treatments-for-covid-19

Izaguirre, G. (2019). The Proteolytic Regulation of Virus Cell Entry by Furin and Other Proprotein Convertases. *Viruses*, *11*(9). https://doi.org/10.3390/v11090837

Johns Hopkins University. (2020). *Home*. Johns Hopkins Coronavirus Resource Center. https://coronavirus.jhu.edu/

Lieberman, N. A. P., Peddu, V., Xie, H., Shrestha, L., Huang, M.-L., Mears, M. C., Cajimat, M. N., Bente, D. A., Shi, P.-Y., Bovier, F., Roychoudhury, P., Jerome, K. R., Moscona, A., Porotto, M., & Greninger, A. L. (2020). In vivo antiviral host transcriptional response to SARS-CoV-2 by viral load, sex, and age. *PLOS Biology*, *18*(9), e3000849. https://doi.org/10.1371/journal.pbio.3000849

McClain, M. T., Constantine, F. J., Henao, R., Liu, Y., Tsalik, E. L., Burke, T. W., Steinbrink, J. M., Petzold, E., Nicholson, B. P., Rolfe, R., Kraft, B. D., Kelly, M. S., Sempowski, G. D., Denny, T. N., Ginsburg, G. S., & Woods, C. W. (2020). Dysregulated transcriptional responses to SARS-CoV-2 in the periphery support novel diagnostic approaches. *MedRxiv*. https://doi.org/10.1101/2020.07.20.20155507

Mick, E., Kamm, J., Pisco, A. O., Ratnasiri, K., Babik, J. M., Calfee, C. S., Castañeda, G., DeRisi, J. L., Detweiler, A. M., Hao, S., Kangelaris, K. N., Kumar, G. R., Li, L. M., Mann, S. A., Neff, N., Prasad, P. A., Serpa, P. H., Shah, S. J., Spottiswoode, N., … Langelier, C. (2020). Upper airway gene expression differentiates COVID-19 from other acute respiratory illnesses and reveals suppression of innate immune responses by SARS-CoV-2. *MedRxiv*. https://doi.org/10.1101/2020.05.18.20105171

Millet, J. K., & Whittaker, G. R. (2015). Host cell proteases: Critical determinants of coronavirus tropism and pathogenesis. *Virus Research*, *202*, 120–134. https://doi.org/10.1016/j.virusres.2014.11.021

Muus, C., Luecken, M. D., Eraslan, G., Waghray, A., Heimberg, G., Sikkema, L., Kobayashi, Y., Vaishnav, E. D., Subramanian, A., Smilie, C., Jagadeesh, K., Duong, E. T., Fiskin, E., Triglia, E. T., Ansari, M., Cai, P., Lin, B., Buchanan, J., Chen, S., … The NHLBI LungMAP Consortium, and T. H. C. A. L. B. N. (2020). Integrated analyses of single-cell atlases reveal age, gender, and smoking status associations with cell type-specific expression of mediators of SARS-CoV-2 viral entry and highlights inflammatory programs in putative target cells. *BioRxiv*, 2020.04.19.049254. https://doi.org/10.1101/2020.04.19.049254

Nagpal, A. (2017, October 14). *L1 and L2 Regularization Methods*. Medium. https://towardsdatascience.com/l1-and-l2-regularization-methods-ce25e7fc831c

Schulte-Schrepping, J., Reusch, N., Paclik, D., Baßler, K., Schlickeiser, S., Zhang, B., Krämer, B., Krammer, T., Brumhard, S., Bonaguro, L., De Domenico, E., Wendisch, D., Grasshoff, M., Kapellos, T. S., Beckstette, M., Pecht, T., Saglam, A., Dietrich, O., Mei, H. E., … Sander, L. E. (2020). Severe COVID-19 Is Marked by a Dysregulated Myeloid Cell Compartment. *Cell*,

182(6), 1419-1440.e23. https://doi.org/10.1016/j.cell.2020.08.001

Siva, S. (2020, October 23). *Dimensionality Reduction for Data Visualization: PCA vs TSNE vs UMAP*. Medium. https://towardsdatascience.com/dimensionality-reduction-for-data-visualization-pca-vs-tsne-vs-umap-be4aa7b1cb29