Final Project

Nupoor Marwah

August 2025

Contents

1 Introduction

The global burden of death and disease caused by the COVID-19 pandemic is particularly striking, especially when considering that the primary cause of these high mortality rates is due to complications from acute respiratory distress syndrome (ARDS) related to coronavirus onset [1]. As a result, many recent studies have brought to light the notion that the intensity of illness with SARS-CoV-2, and its association with ARDS, is in many ways dictated by molecular and genetic factors of the human host [1]. It has also been found that level of gene expression for certain human physiologic processes is altered during COVID-19 infection, often playing a role in overall metabolic efficiency and function [1]. Thus, a dataset, originally generated by Overmyer et al. using high-resolution mass spectrometry and RNA sequencing, sought to understand how metabolic processes such as lipid transport and amino acid production were impacted by COVID-19 and associated ARDS. In this dataset generated by Overmyer et al., 128 blood samples were collected and sequenced from participants with and without COVID-19 infection and used to map their significance with coronavirus to dictate disease severity predictions and outcome prognoses [1].

Using the research conducted by Overmyer et al. as a basis, the aim of this study was to analyze a variety of genes from that dataset, specifically focusing on the AASDHPPT gene in terms of its level of expression based on disease state and to plot the gene against different covariates known to be associated with COVID-19 infection. The AASDHPPT gene was observed specifically due to its association with many comorbidities when its expression is low, which one can predict may lead to negative health outcomes if infection with COVID-19 is initiated [2]. The AASDHPPT gene, when having high expression, efficiently works to transport "apo" proteins (inactive proteins) to their active state so that fatty acid synthesis, peptide synthesis, and amino acid lysine production can take place [2].

2 Methods

2.1 R packaging

R was used as the only coding software to generate summary statistics and plots for this study. Certain packages were installed to enhance analysis. These packages included "data.table" for easier data manipulation purposes [3], "ggplot2" for creating vibrant and easy-to-read plots [4], "tidyverse" set of packages for data visualization and data transformation [5], "glue" for straight-forward interpretation with string variables [6[, and "kableExtra" for table output styling [7]. The data source was a paper published by Overmyer et al. in 2021 in the 12th edition of a journal called Cell Systems [1].

2.2 Gene Expression Visualization

To get a basic understanding of AASDHPPT gene expression among the 128 participants in the dataset, a histogram was created. Then, in order to understand how gene expression varied by a continuous variable, the covariate Charlson Score was plotted against the AASDHPPT gene in a scatterplot. Charlson Score is a numeric metric used to determine the likelihood of death within 10 years for an individual depending on how many concurrent comorbidities he/she has. This variable was chosen as the continuous variable because it can provide insight on how AASDHPPT expression may change depending on whether someone has COVID-19 or not. It is worth understanding if infection with COVID-19 may lower gene expression in those with multiple chronic comorbid conditions. From there, a boxplot was generated to look at how AASDHPPT expression differed by sex in those with COVID-19 compared to those without COVID-19.

From here, R v4.4.2 was used to generate a function that created similar histograms, scatterplots, and boxplots for two additional genes. A loop was implemented in R to generate the figures using the newly created function. The A2M gene was chosen because infection with COVID-19 and high Charlson Scores from comorbidities likely increase the risk (or prevalence) of chronic inflammation, which the A2M gene functions to regulate [8]. Thus, analysis on this gene's expression may provide insight on if it could be a potential biomarker for outcome severity. The AAAS gene was also chosen to analyze because it is related to regulating oxidative stress and initiating repair of DNA damage, factors that are also related to coronavirus infection and high Charlson Score [9].

A table was generated to lay out the summary statistics for all the covariates analyzed, as well as the categorical variable of ICU status, and the continuous variables Ferritin levels in ng/mL and number of ventilator-free days.

2.3 Additional Plots

Next, a heatmap was created using the package "ComplexHeatmap" to look at the expression of 10 different genes [10]. Tracking bars were added for the categorical variables of sex and disease status (COVID-19 vs. non-COVID-19). The heatmap included clustered rows and columns. To cluster, the command 'cluster rows = TRUE' was used to ensure that participants were clustered in rows. The command 'cluster columns = TRUE' was used to ensure that the genes were clustered in the columns. Below is the final line of code used to create the heatmap with clustering:

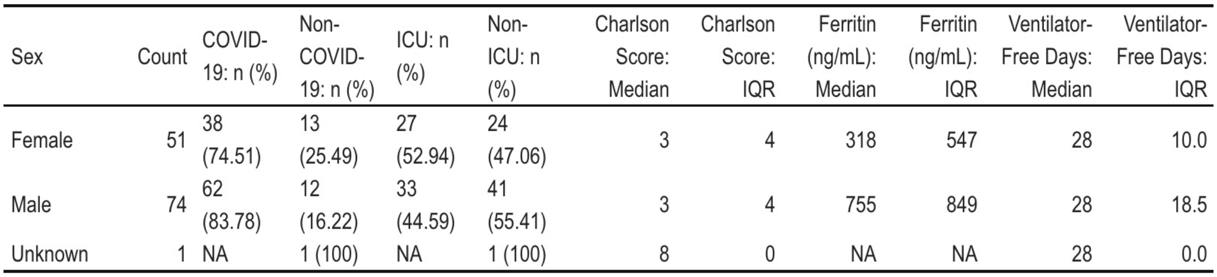
```
Heatmap(
  expression_mat,
  name = "Gene␣Expression",
  column_title = "Heatmap␣of␣the␣Expression␣of␣10␣Genes",
  right_annotation = hm_annot,
  cluster_rows = TRUE,
  cluster_columns = TRUE
```

Finally, a new type of plot, a PCA plot, was generated. A Principal Component Analysis plot, colloquially referred to as a PCA plot, was used to understand where the most variance in the data lied and if the covariate "sex" that was used against gene expression for the 10 previously selected genes was a major factor influencing variation in the data. The plot was colored by sex to visualize whether males or females had different clustering for gene expression. A similar methodology was employed for a PCA plot with the covariate "disease status".

3 Results

3.1 Table of Summary Statistics

Table 1: Multiple Covariates Stratified by Sex



Sex	Count	COVID-19: n (%)	Non-COVID-19: n (%)	ICU: n (%)	Non-ICU: n (%)	Charlson Score: Median	Charlson Score: IQR	Ferritin (ng/mL): Median	Ferritin (ng/mL): IQR	Ventilator-Free Days: Median	Ventilator-Free Days: IQR
Female	51	38 (74.51)	13 (25.49)	27 (52.94)	24 (47.06)	3	4	318	547	28	10.0
Male	74	62 (83.78)	12 (16.22)	33 (44.59)	41 (55.41)	3	4	755	849	28	18.5
Unknown	1	NA	1 (100)	NA	1 (100)	8	0	NA	NA	28	0.0

According to Table 1, among the 51 females, 38 (74.5%) were COVID-19 positive and 13 (25.5%) were COVID-19 negative. This was compared to 62 (83.8%) COVID-19 positive and 12 (16.2%) COVID-19 negative males. There was a higher percentage of females in the ICU compared to males (52.9% vs. 44.6%, respectively). Additionally, the median and interquartile range (IQR) for Charlson score was equal for males and females, with a median score of 3 and an IQR of 4. However, for the individual with unknown sex, their median Charlson Score was higher, at 8. Median Ferritin in ng/mL was higher for males than for females (755 vs. 318, respectively), and a similar trend was seen for Ferritin IQR (849 vs. 547). Median number of ventilator-free days was the same for males, females, and the unknown individual at 28 days. However, the IQR for the number of ventilator-free days was higher for males than for females (18.5 vs. 10.0, respectively).

3.2 AASDHPPT Gene Expression Analysis: Histogram of Gene

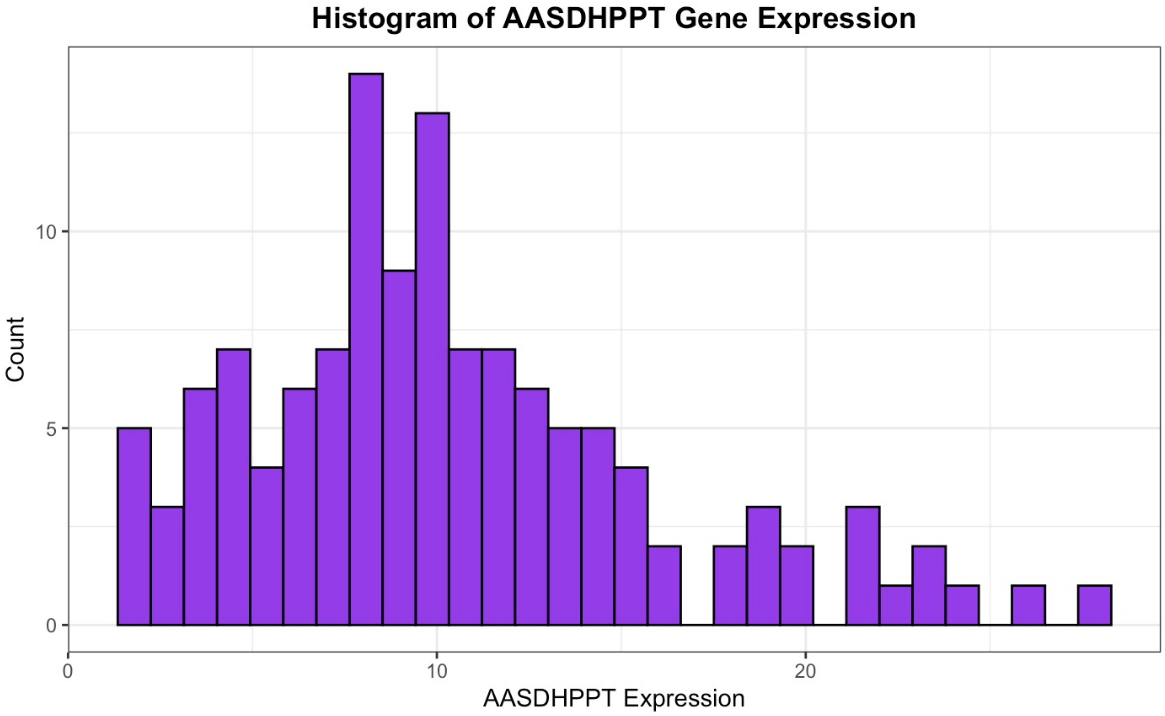


Figure 1: Histogram of AASDHPPT Gene Expression

When looking at the histogram of AASDHPPT gene expression by count as seen in Figure 1, there seemed to be a slight right skew, indicating that most individuals had a moderate level of gene expression, but there were a small number of individuals with a higher than normal or higher than expected level of gene expression. Most individuals had a gene expression value of around 10 units, but a select few had high levels of expression greater than or equal to 20 units.

3.3 AASDHPPT Gene Expression Analysis: Scatterplot of Gene + Continuous Covariate

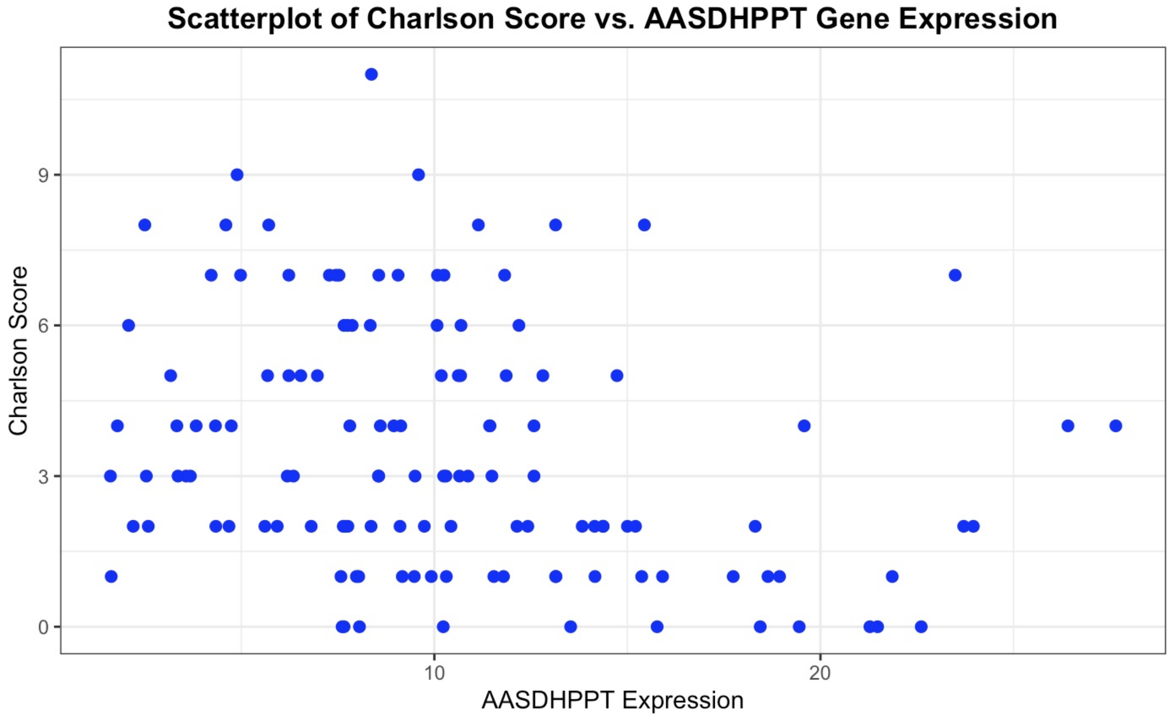


Figure 2: Scatterplot of AASDHPPT Gene Expression by Charlson Score

Based on the scatterplot of Charlson Score by AASDHPPT gene expression in Figure 2, while fairly interspersed and not a strong correlation, there looks to be a higher amount of clustering around a lower gene expression and higher Charlson Score. This makes sense, as it can be inferred that individuals with a greater number of comorbid conditions tend to have less efficient metabolic functioning. Though this would need to be explored further to concretely discern this, for the purposes of this study, it is a suitable assumption.

3.4 AASDHPPT Gene Expression Analysis: Boxplot of Gene Stratified by 2 Categorical Covariates

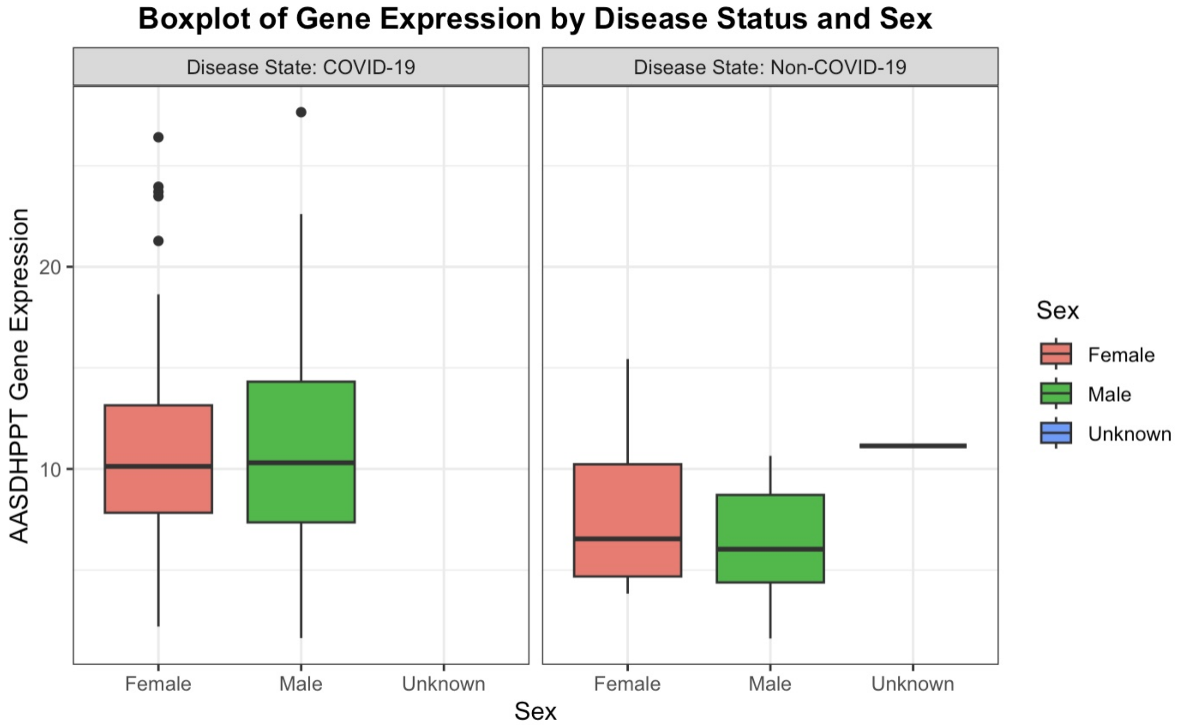


Figure 3: Boxplot of AASDHPPT Gene Expression Stratified by Disease Status and Sex

As seen in Figure 3, when stratifying by disease status, both females and males with COVID-19 had higher levels of AASDHPPT gene expression compared to females and males without COVID-19. Both sexes among the infected showed a median gene expression at approximately 10 units, both including outliers of individuals showing gene expression of 17 units or above. However, this can be contrasted with the overall lower median gene expression of 7 units in both uninfected sexes. Despite a slight right skew in uninfected females, overall level of expression remained similar in those without COVID-19.

3.5 Multi-Gene Expression Analysis: Heatmap

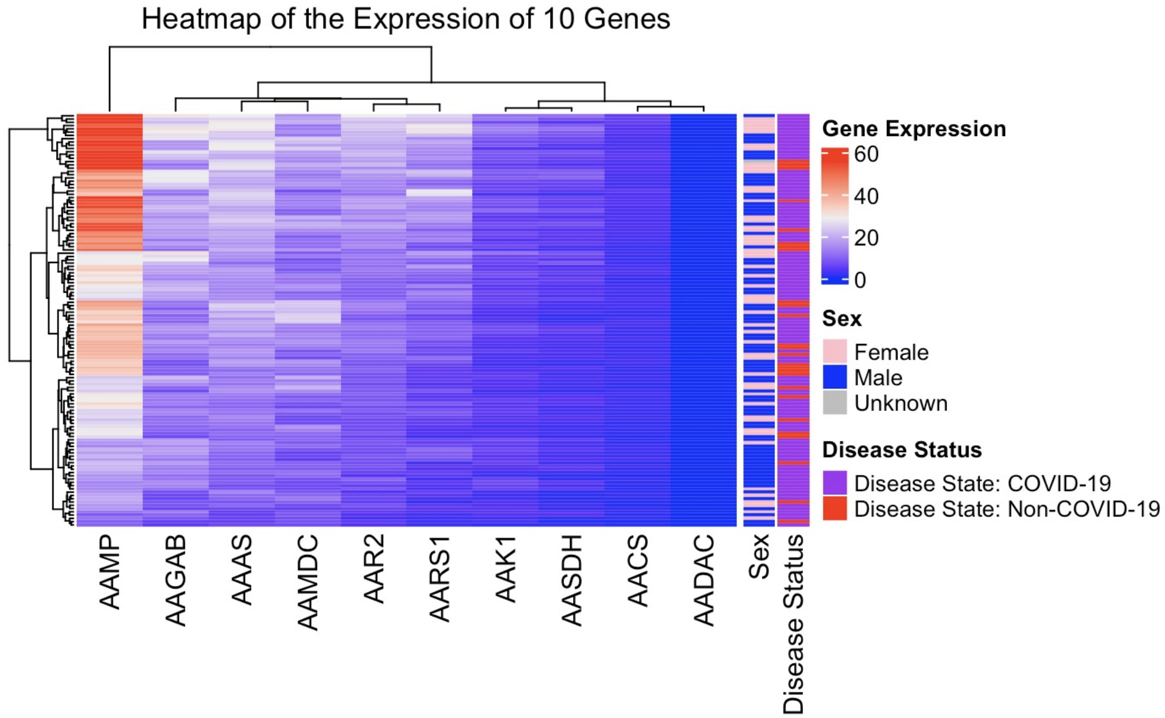


Figure 4: Clustered Heatmap of 10 Genes and Their Expression with Tracking Bars of Sex and Disease Status

The clustered heatmap in Figure 4 showed a pronounced level of AAMP gene expression in certain samples when individuals were COVID-19 positive, as evidenced by the bright red in the top left corner. This makes sense, as the AAMP gene is involved in human immune responses, such as cell invasion and migration during illness [11]. Thus, it seems sensible that blood samples showed high levels of AAMP gene expression during infection with coronavirus. This effect seemed to be prominent in both males and females. This information suggests that the AAMP gene may be associated with COVID-19 disease and could be a good biomarker for COVID-19 detection in future studies. The AASDH, AACS, and AADAC gene expression consistently showed low levels of expression across the samples. These genes likely were not associated with COVID-19 disease status. The AAGAB, AAAS, AAMDC, AAR2, AARS1, and AAK1 showed intermediate levels of gene expression across the participants, with expression levels higher than AASDH, AACS, and AADAC, but lower than AAMP. The clustering across rows was used to show how certain samples shared similar levels of gene expression, meanwhile the clustering across columns indicated similarities among genes in terms of level of expression, with the most pronounced variability in expression on the left-most side to the least visible variability in expression on the far right. In terms of the annotation tracking bars, there was a high amount of variability among males and females, indicating that sex was not a strong predictor of expression level for these 10 genes. Alternatively, from the disease status bar, COVID-19 positive participant samples were found primarily in clusters that showed high AAMP gene expression. The opposite trend remained true for COVID-19 negative participant samples; samples of non-COVID-19 were found mostly in clusters with low gene expression. This shows that disease status may play a role in the variation in gene expression, while sex does not. It should be noted, though, that this association was seen on an individual gene level and cannot necessarily be applied more broadly across many different genes. However, additional analyses and statistical tests would need to be conducted in follow-ups to this study to confirm this.

3.6 Multi-Gene Expression Analysis: PCA Plot

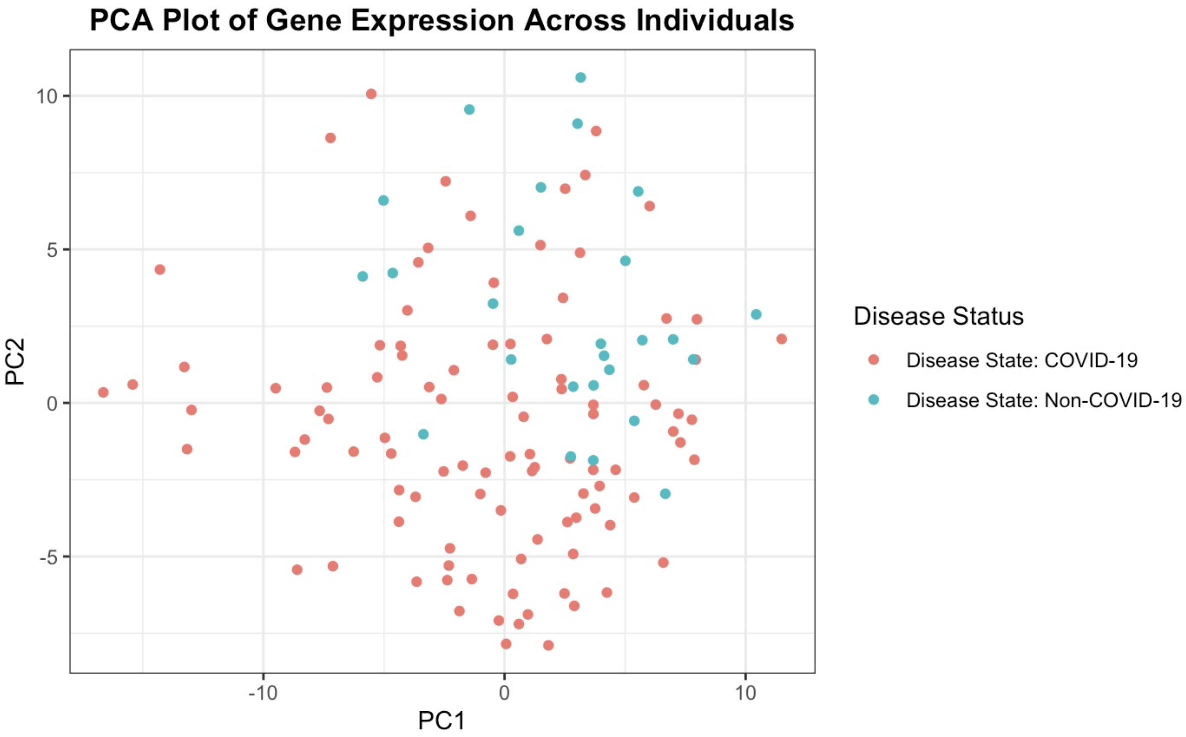


Figure 5: PCA Plot of Gene Expression Based on Disease Status

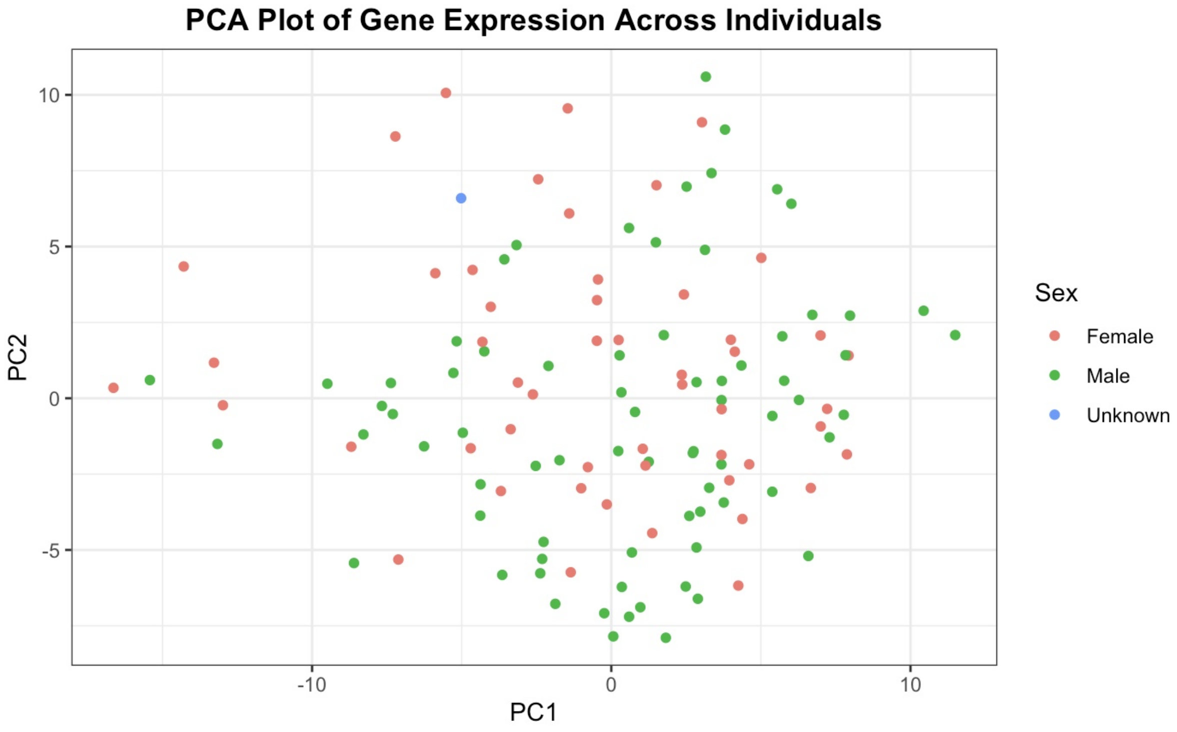


Figure 6: PCA Plot of Gene Expression Based on Sex

Lastly, in Figure 5, it is important to understand what a PCA plot is in order to interpret it. Here, each point represented a singular participant's sample. Red dots indicated individuals with a positive COVID-19 disease state and blue dots indicated individuals with a negative disease state. PC1 was the first principal component, which showed the location of the highest variance in gene expression. PC2 was the second principal component, which showed the location for the second highest variance in gene expression. Because the dots were scattered with no clear distinction in separation between those in the diseased and non-diseased states, disease status was not the primary source of variation. While the association between disease and variation existed on the single gene level as evidenced in Figure 4, the PCA map here suggests that this was not a pattern seen across multiple genes. A similar PCA plot was generated in Figure 6 to determine if sex as the other covariate could explain the variance seen in gene expression. Similar to the findings from Figure 5, there was no clear separation between the dots for each sex, indicating that sex was also not a main driver for the variance seen in the dataset. There are likely some other factors at play that were not accounted for in this analysis or in the original dataset by Overmyer et al.

4 References

1. Overmyer, K. A., Shishkova, E., Miller, I. J., Balnis, J., Bernstein, M. N., Peters-Clarke, T. M., Meyer, J. G., Quan, Q., Muehlbauer, L. K., Trujillo, E. A., He, Y., Chopra, A., Chieng, H. C., Tiwari, A., Judson, M. A., Paulson, B., Brademan, D. R., Zhu, Y., Serrano, L. R., ... Jaitovich, A. (2021). Large-Scale Multi-omic Analysis of COVID-19 Severity. Cell Systems, 12(1), 23-40.e7. https://doi.org/10.1016/j.cels.2020.10.003

2. Norden, P. R., Wedan, R. J., Longenecker, J. Z., Preston, S. E. J., Graber, N., Pentecost, O. A., Canfield, M., McLaughlin, E., Nowinski, S. M. (2024). Mitochondrial Phosphopantetheinylation is Required for Oxidative Function (p. 2024.05.09.592977). bioRxiv. https://doi.org/10.1101/2024.05.09.592977

3. Barrett T, Dowle M, Srinivasan A, Gorecki J, Chirico M, Hocking T, Schwendinger B (2024). data.table:Extension of ‘data.frame‘. R package version 1.16.4, ¡https://CRAN.R project.org/package=data.table.

4. H. Wickham. ggplot2: Elegant Graphics for Data Analysis. Springer-Verlag New York, 2016.

5. Wickham H, Averick M, Bryan J, Chang W, McGowan LD, François R, Grolemund G, Hayes A, Henry L, Hester J, Kuhn M, Pedersen TL, Miller E, Bache SM, Müller K, Ooms J, Robinson D, Seidel DP, Spinu V, Takahashi K, Vaughan D, Wilke C, Woo K, Yutani H (2019). “Welcome to the tidyverse.” Journal of Open Source Software, *4*(43), 1686. doi:10.21105/joss.01686 ¡https://doi.org/10.21105/joss.01686¿.

6. Hester J, Bryan J (2024). glue: Interpreted String Literals. R package version 1.8.0, ¡https://CRAN.R-project.org/package=glue.

7. Zhu H (2024). kableExtra: Construct Complex Table with ’kable’ and Pipe Syntax. R package version 1.4.0, ¡https://CRAN.R-project.org/package=kableExtra¿.

8. Mocchegiani, E., Malavolta, M. (2007). Zinc Dyshomeostasis, Ageing and Neurodegeneration: Implications of A2M and Inflammatory Gene Polymorphisms. Journal of Alzheimer’s Disease, 12(1), 101–109. https://doi.org/10.3233/JAD-2007-12110

9. Jühlen, Ramona, Jan Idkowiak, Angela E. Taylor, et al. “Role of ALADIN in Human Adrenocortical Cells for Oxidative Stress Response and Steroidogenesis.” PLoS ONE 10, no. 4 (2015): e0124582. https://doi.org/10.1371/journal.pone.0124582.

10. Gu, Z. (2016) Complex heatmaps reveal patterns and correlations in multidimensional genomic data. Bioinformatics.

11. Yao, S., Shi, F., Mu, N., Li, X., Ma, G., Wang, Y., Sun, X., Liu, X., Su, L. (2021). Angio-associated migratory cell protein (AAMP) interacts with cell division cycle 42 (CDC42) and enhances migration and invasion in human non-small cell lung cancer cells. Cancer Letters, 502, 1–8. https://doi.org/10.1016/j.canlet.2020.11.050

12. https://www.overleaf.com/learn