Final Project

Xinyi Yu

August 21, 2025

1 Introduction

The dataset analyzed in this project was derived from the study of Overmyer et al., 2021, which profiled immune responses in COVID-19 patients and non-COVID controls. For the main analyses, we focus on the AAAS gene. The AAAS gene encodes a member of the WD-repeat family of regulatory proteins, which is part of the nuclear pore complex and anchored by NDC1. It plays an important role in the development of the peripheral and central nervous system. Mutations in this gene cause achalasia-addisonianism-alacrima syndrome (AAAS), also known as triple-A or Allgrove syndrome PubChem, n.d. Given its role in nuclear pore complex function and stress response, AAAS was selected as a candidate gene for this analysis due to its potential relevance for host-pathogen interactions and immune dysregulation in COVID-19.

2 Methods

2.1 Data Source

The dataset was obtained from the study by Overmyer et al., 2021, available through public repositories associated with the publication.

2.2 R Version and Packages

The analyses were conducted in R version 4.2.1. The main packages included tidyverse Wickham et al., 2019 for data manipulation and visualization, ggplot2 Wickham, 2016 for plotting, pheatmap Kolde, 2025 for heatmap generation, and kableExtra Zhu, 2021 for creating and formatting complex tables.

2.3 Clustering algorithm

Hierarchical clustering was performed on the gene expression data using the Euclidean distance and complete linkage to generate the heatmap.

3 Results

3.1 Table 1. Summary statistics by disease status

Table 1 shows the summary statistics by disease status. Continuous variables are summarized as mean with standard deviation and categorical variables as counts with percentages. Patients with COVID-19 tend to be older and have higher ferritin levels, and ICU cases are more common in this group. The sex distribution is relatively balanced. In general, the COVID-19 group appears clinically more severe, and future work can test whether these observed differences are statistically significant.

Table 1: Summary statistics by disease status

Variable	disease state: COVID-19	disease state: non-COVID-19	Overall
age	60.84 (16.15)	61.96 (15.36)	61.06 (15.94)
hospital_free	22.09 (16.62)	32.36 (15.09)	24.14 (16.79)
ferritin	932.76 (1094.04)	250.50 (238.21)	833.52 (1042.80)
sex	-	-	-
female	38 (38.0%)	13 (52.0%)	-
male	62 (62.0%)	12 (48.0%)	-
icu_status	-	-	-
no	50 (50.0%)	10 (40.0%)	-
yes	50 (50.0%)	15 (60.0%)	-

3.2 Histogram of Gene Expression

The histogram shows the distribution of AAAS expression levels in all samples. The x-axis represents AAAS expression levels, and the y-axis represents the number of samples. Expression values appear to be approximately symmetric, with most samples clustered between 10 and 20. This suggests that while there is variability, extreme high or low expression values are uncommon.

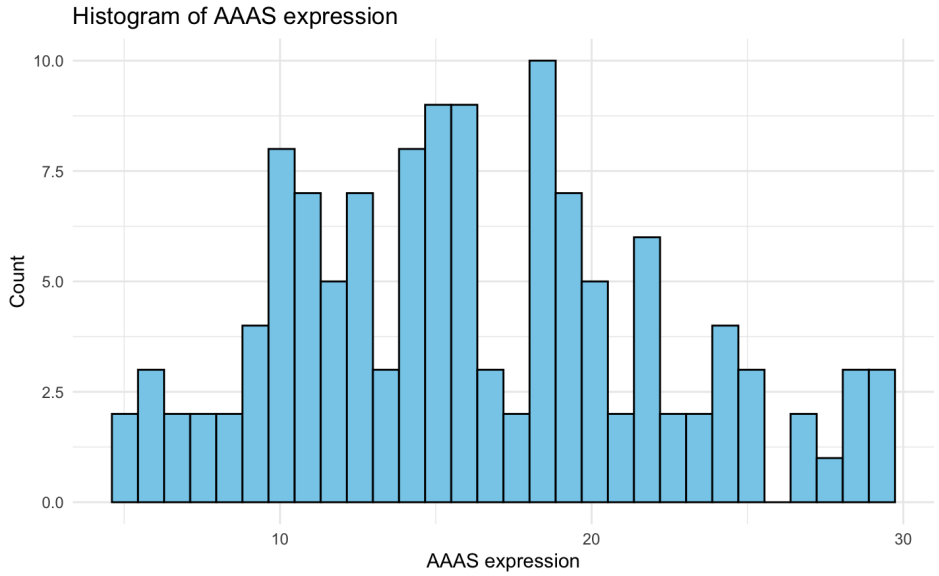

Figure 1: Histogram of Gene Expression

3.3 Scatter Plot of Gene Expression vs Age

The scatter plot illustrates the relationship between AAAS expression and age. Each point represents a sample. No clear linear association was observed, as the expression values appear scattered between age groups. This suggests that age alone may not strongly influence AAAS expression.

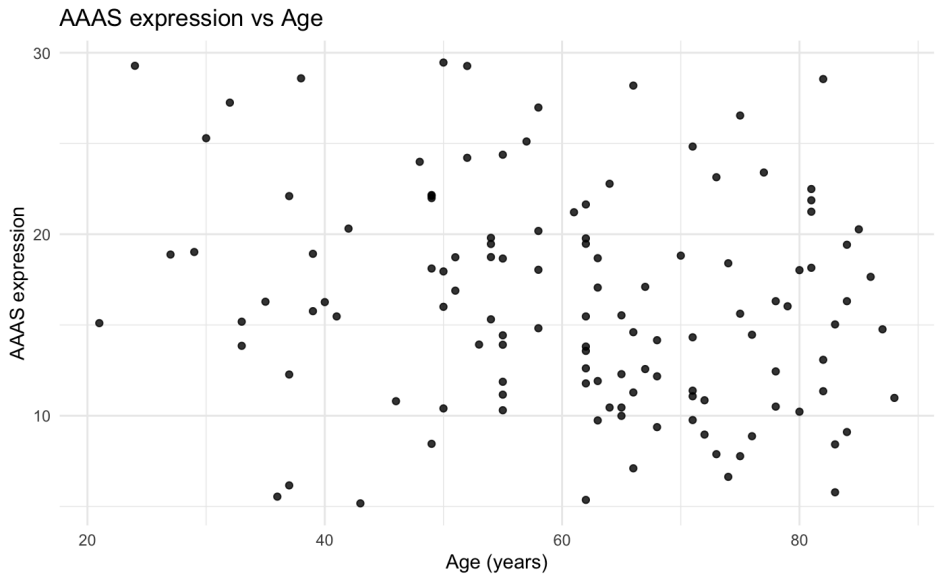

Figure 2: Scatter Plot of Gene Expression vs Age

3.4 Boxplot of gene stratified by 2 categorical covariates

This boxplot shows how AAAS expression varies by sex and disease status. The x-axis shows biological sex and colors indicate disease state. For both females and males, COVID-19 patients tend to show higher variability and slightly higher median expression. Non-COVID samples have more compact expression ranges. Among males, COVID-19 patients have a higher median expression of AAAS than Non-Covid 19 patients. All above might implies that neither sex nor disease status alone explains differences in AAAS expression, but interactions could be further explored.

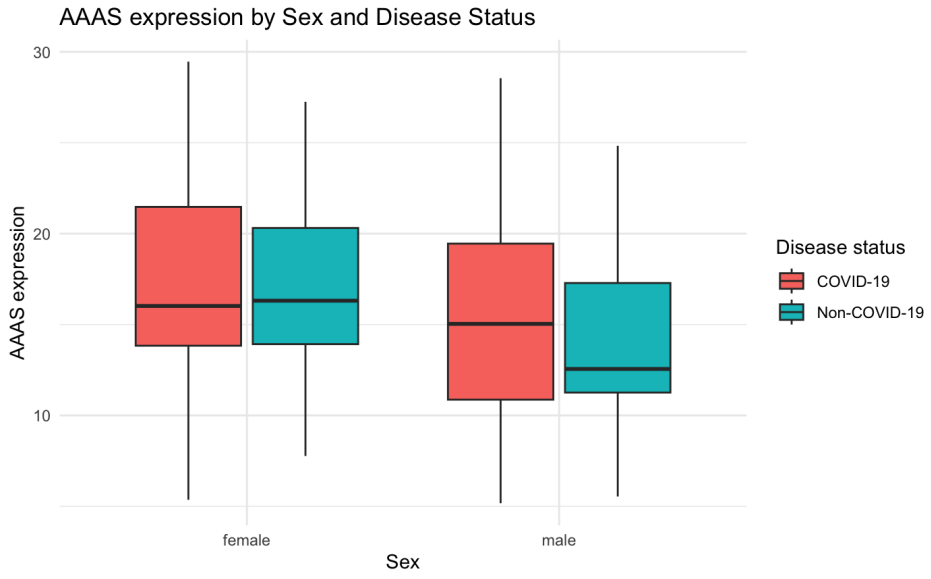

Figure 3: Boxplot of AAAS gene expression stratified by sex and disease status

3.5 Heatmap of top variable genes

This heatmap shows the expression of 20 genes with the largest variance in 20 randomly sampled samples. Red indicates higher expression (after log2 transformation), blue indicates lower expression, and yellow indicates intermediate levels. Genes with similar expression patterns are grouped together.

For example, ABCG1 shows low expression in most samples and is clustered with other low-expression genes such as ABHD13 and ABCC5. In contrast, ABHD3, ABHD5, and ABHD16A show much higher expression in some samples and are clustered together. The clustering of samples does not separate COVID-19 from non-COVID-19 or males from females. Instead, samples from different groups are mixed, which suggests that disease status and sex are not the main factors here. Some genes, such as ABHD3, also show strong variation between individuals, with some samples very high (dark red) and others much lower (light blue). This variation looks stronger than the effect of disease status or sex. In conclusion, the heatmap highlights two main patterns: one cluster of genes with consistently low expression, and another with variable or high expression peaks. Clinical variables do not drive the clustering, while individual differences and gene-specific behavior play a stronger role.

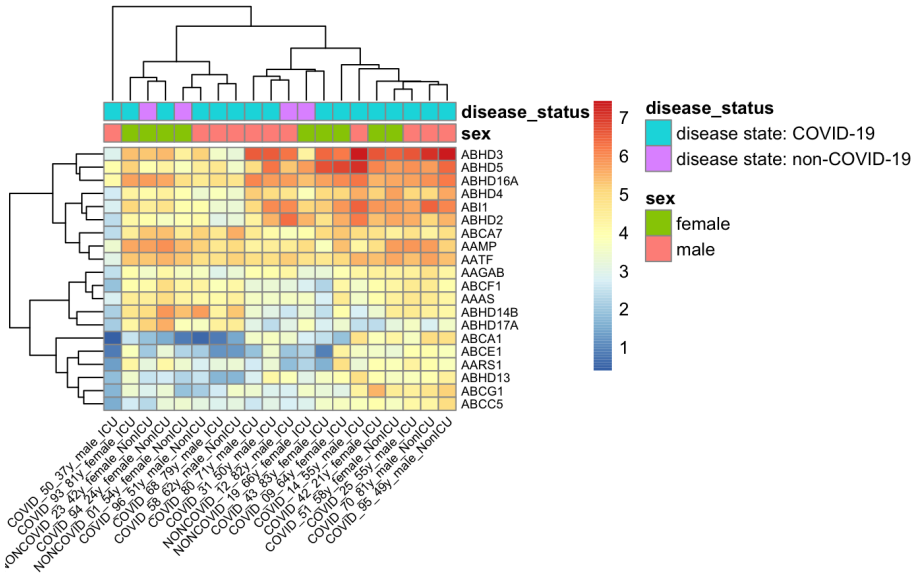

Figure 4: Heatmap of top variable genes

3.6 2D density of AAAS expression vs. Age

This 2D density plot shows the distribution of AAAS expression against age, stratified by COVID-19 status. In the COVID-19 group, there are visibly more samples, with most clustering in the lower-right area corresponding to older age and lower expression. However, some points show higher expression at younger ages. In the non-COVID group, the distribution is more dispersed across the plot, without a clear concentration. In conclusion, AAAS expression patterns appear more concentrated among COVID-19 patients, while non-COVID patients exhibit broader variability.

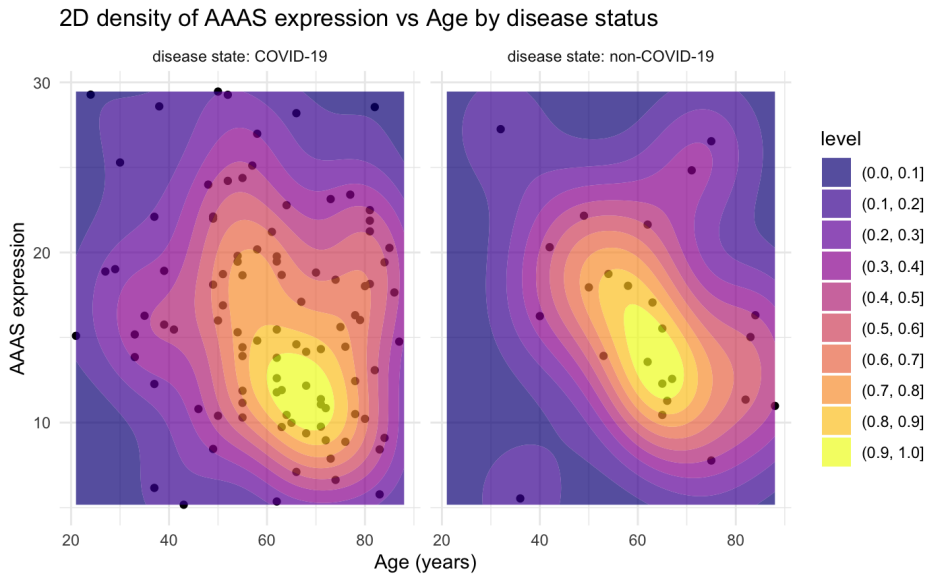

Figure 5: 2D density of AAAS expression vs. Age

3.7 Conclusion

In summary, the analysis provides an overview of both patient characteristics and gene expression patterns. Patients with COVID-19 tend to be older, have higher ferritin levels, and more ICU admissions, suggesting a clinically more severe profile. For the AAAS gene, its expression distribution was fairly symmetric, and no strong linear relationship with age was observed. Stratification by sex and disease status indicated that neither factor alone fully explained the expression differences, though interactions may be important. The heatmap of the top variable genes showed clusters of genes with consistently low expression and others with high variability, but clustering did not clearly separate patients by COVID-19 status or sex. Finally, the 2D density plot demonstrated that COVID-19 patients were more concentrated in older age with lower AAAS expression. However, non-COVID patients had a more dispersed distribution. Together, these findings suggest that individual variability may play a larger role in gene expression differences than clinical grouping alone, and further statistical testing could help confirm these observations.

References

Kolde, R. (2025). *Pheatmap: Pretty heatmaps* [R package version 1.0.13]. https://CRAN.R-project.org/package=pheatmap

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

PubChem. (n.d.). AAAS - aladin WD repeat nucleoporin (human). Retrieved August 21, 2025, from https://pubchem.ncbi.nlm.nih.gov/gene/AAAS/human

Wickham, H. (2016). *Ggplot2: Elegant graphics for data analysis*. Springer-Verlag New York. https://ggplot2.tidyverse.org

Wickham, H., Averick, M., Bryan, J., Chang, W., McGowan, L. D., François, R., Grolemund, G., Hayes, A., Henry, L., Hester, J., Kuhn, M., Pedersen, T. L., Miller, E., Bache, S. M., Müller, K., Ooms, J., Robinson, D., Seidel, D. P., Spinu, V., … Yutani, H. (2019). Welcome to the tidyverse. *Journal of Open Source Software*, *4*(43), 1686. https://doi.org/10.21105/joss.01686

Zhu, H. (2021). *Kableextra: Construct complex table with 'kable' and pipe syntax* [R package version 1.3.4]. https://CRAN.R-project.org/package=kableExtra