QBS 103 Final Project Report

Ruiqi Li

August 19, 2025

1 Introduction

The dataset used in this project comes from Overmyer *et al.*, 2020, which profiled leukocyte gene expression in COVID-19 patients [2]. For my main visualizations, I focused on the **A1BG** gene. A1BG (alpha-1-B glycoprotein) is expressed in human plasma, and I investigated its relationship with several clinical covariates, including age (continuous), sex (categorical), and ICU status (categorical). Additional covariates such as ferritin and CRP levels were also used for descriptive statistics and multivariate visualization.

2 Methods

Data preprocessing and analysis were performed in R. Packages used included `ggplot2`, `dplyr`, `tidyr`, and `pheatmap` [3, 4, 5, 1]. Counts were $\log_2$-transformed with a pseudocount of 1 before visualization. For the heatmap, I selected the top-variance genes (n=20), standardized them by row z-scores, and clustered rows and columns using Euclidean distance with complete linkage.

3 Results

3.1 Table of summary statistics

Table 3.1 presents descriptive statistics stratified by ICU status. Categorical variables are reported as n (%), and continuous variables as mean (SD) or median [IQR].

Table 1: Summary statistics stratified by ICU status

	ICU (n = 66)	Non-ICU (n = 60)
Categorical variables		
Sex, female	24 (36.9%)	27 (45.0%)
Sex, male	41 (63.1%)	33 (55.0%)
Mechanical ventilation, Yes	46 (69.7%)	5 (8.3%)
Mechanical ventilation, No	20 (30.3%)	55 (91.7%)
Continuous variables		
Age, mean (SD)	63.5 (14.0)	58.7 (17.8)
Ferritin, median [IQR]	685 [325–1212]	401 [131–870]
CRP, mean (SD)	150 (106)	109 (94.4)

3.2 Histogram of gene

The histogram of A1BG expression values across participants is shown in Figure 1. The distribution is right-skewed, with most individuals exhibiting low expression.

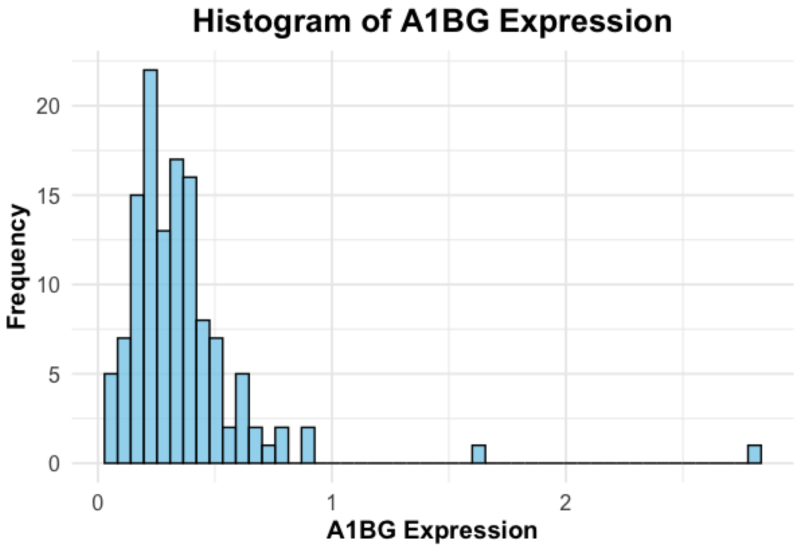

Figure 1: Histogram of A1BG expression across all participants.

3.3 Scatter plot of gene + continuous covariate

Figure 2 shows a scatter plot of A1BG versus age, stratified by ICU status and sex. There was no clear linear association between A1BG expression and age.

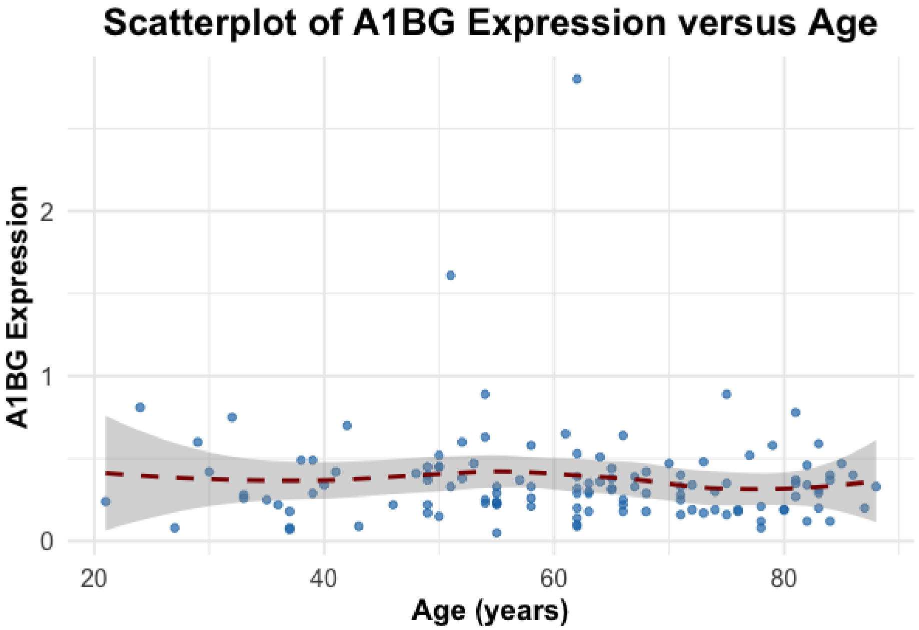

Figure 2: Scatter plot of A1BG expression vs. age with LOESS smoother; points colored by ICU status and shaped by sex.

3.4 Boxplot of gene stratified by two categorical covariates

A1BG expression stratified by sex and ICU status is shown in Figure 3. Median expression levels appeared similar between males and females, but some differences were observed by ICU status.

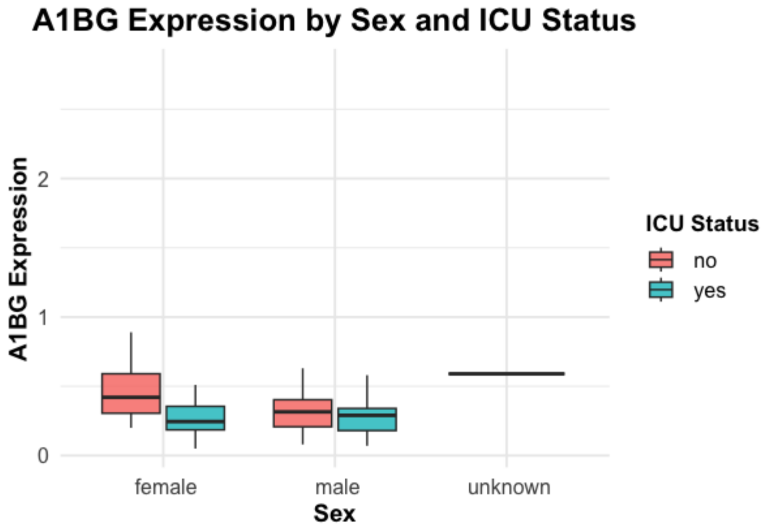

Figure 3: Boxplot of A1BG expression by sex and ICU status.

3.5 Heatmap

The clustered heatmap (Figure 4) of the top 20 variance genes reveals structured differences across patients. Tracking bars for sex and ICU status are shown above the columns.

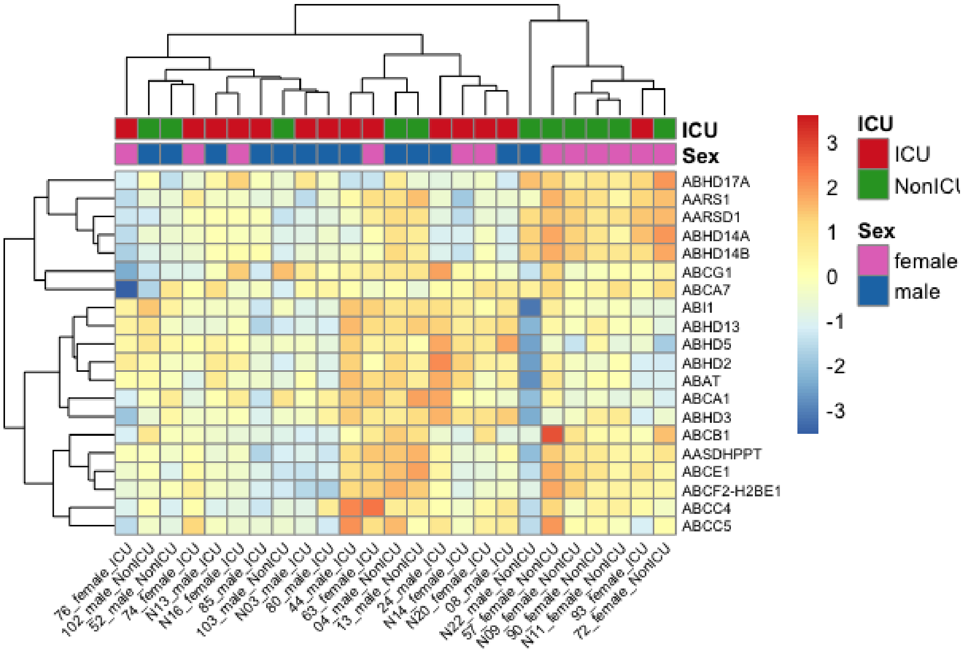

Figure 4: Heatmap of the top 20 variance genes. Rows and columns are clustered with Euclidean distance and complete linkage. Annotation bars indicate sex and ICU status.

3.6 New plot type (PCA)

To provide an alternative visualization, I performed principal component analysis (PCA) of the same top-variance genes (Figure 5). PC1 and PC2 explained a large proportion of the variance, and partial separation by ICU status was observed.

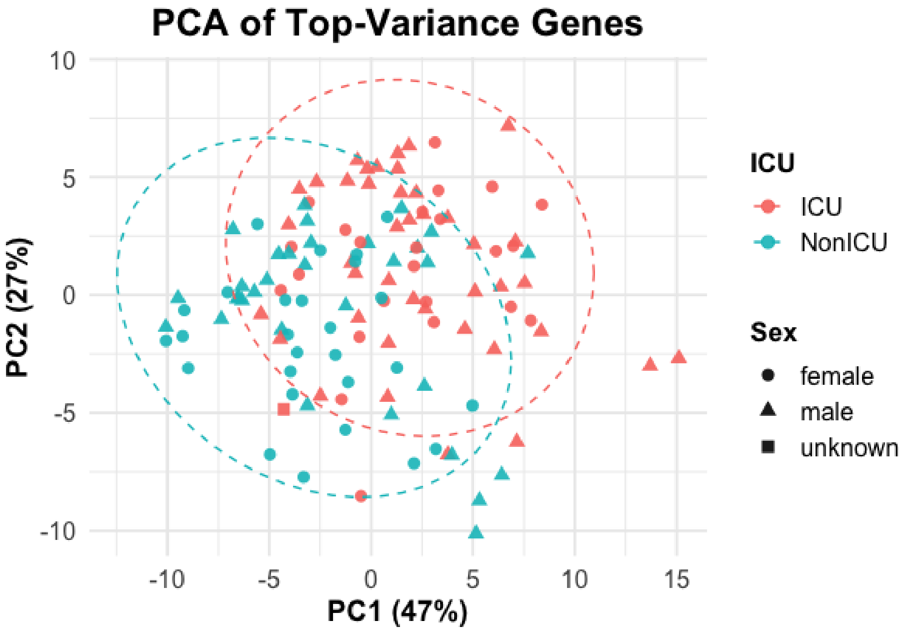

Figure 5: PCA of top-variance genes. Points are colored by ICU status and shaped by sex. Ellipses indicate group variance.

4 Discussion

Overall, A1BG expression did not show strong associations with age or sex. However, patterns across multiple high-variance genes suggest global transcriptional changes in ICU patients (Figures 4, 5). The PCA and heatmap provide complementary perspectives: the PCA summarizes global variance across samples, while the heatmap highlights gene-level differences.

5 References

References

[1] Kolde, R. *pheatmap: Pretty Heatmaps*, 2019. R package version 1.0.12.

[2] 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., Linke, V., Drake, L. A., Adam, A. P., Schwartz, B. S., Singer, H. A., Swanson, S., Mosher, D. F., Stewart, R., Coon, J. J., and Jaitovich, A. Large-scale multi-omic analysis of COVID-19 severity. 23–40.e7.

[3] Wickham, H. *ggplot2: Elegant Graphics for Data Analysis*, 2016. R package version 3.4.0.

[4] Wickham, H., François, R., Henry, L., and Müller, K. *dplyr: A Grammar of Data Manipulation*, 2023. R package version 1.1.2.

[5] Wickham, H., and Girlich, M. *tidyr: Tidy Messy Data*, 2025. R package version 1.3.1.