# QBS 103 Final Project Report

Ruiqi Li

August 21, 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.
For dimensionality reduction, we performed principal component analysis (PCA) on the top-variance genes ($\log_2$-transformed, row-standardized), using Euclidean distance on the centered and scaled expression matrix (`prcomp` with `center=TRUE` and `scale.=TRUE`).

## 3 Results

### 3.1 Table of summary statistics

Table 3.1 summarizes baseline characteristics of the 126 participants stratified by ICU status at enrollment.
Categorical variables are reported as $n$ (%), computed within each stratum: sex (female/male) and receipt of mechanical ventilation (Yes/No).
Continuous variables are summarized as mean (SD) for approximately symmetric distributions (age, CRP) and as median [IQR] for right-skewed measures (ferritin).
Sex labels were standardized to *female/male* after trimming whitespace; ICU status was harmonized from the original *yes/no* field; and mechanical ventilation indicates any documented ventilation during the index admission.
Overall, the ICU group was older on average than the Non-ICU group (63.5 vs. 58.7 years) and exhibited higher inflammatory markers: ferritin median 685 [325–1212] vs. 401 [131–870] ng/mL, and CRP mean 150 vs. 109 mg/L. Mechanical ventilation was markedly more frequent among ICU

patients (69.7% vs. 8.3%). Sex distributions were broadly similar between strata (female: 36.9% in ICU vs. 45.0% in Non-ICU).

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 (Figure 1) shows a strongly right-skewed distribution with most observations concentrated at low values and a long, sparse upper tail. A small number of high-expression outliers are present, which widen the range but do not alter the overall pattern that the majority of samples exhibit modest expression. This shape suggests over-dispersion and potential zero/near-zero inflation that are typical for gene-level measurements; it also motivates the use of robust summaries (e.g., medians/IQRs) and transformations in downstream analyses.

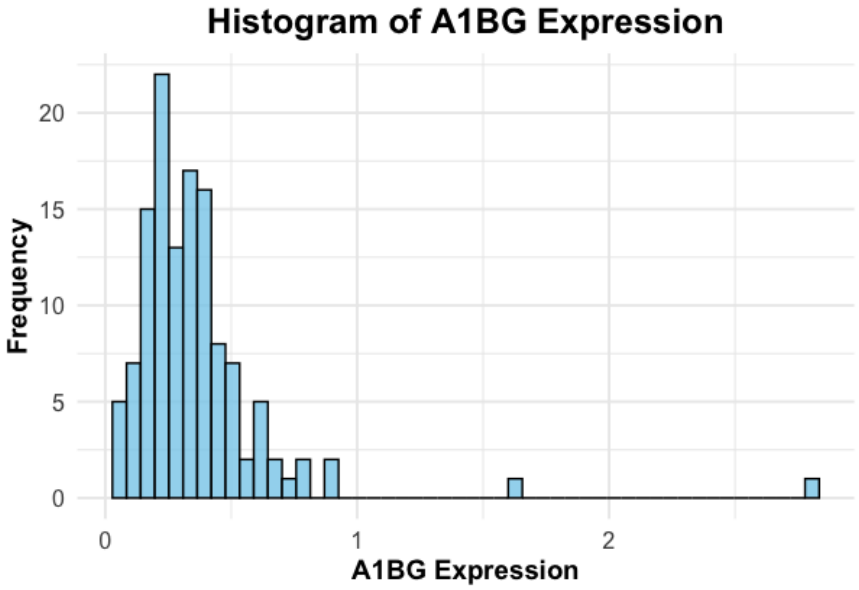


Figure 1: Histogram of A1BG expression across all participants.

### 3.3 Scatter plot of gene + continuous covariate

The A1BG–age scatterplot with a LOESS smoother (Figure 2) indicates no strong monotonic association between expression and age. The fitted curve is shallow and largely contained within the 95 percent confidence band, consistent with a weak or null relationship. Variability appears fairly constant across the age range, with a few isolated higher-expression points that do not align with a simple linear trend. Overall, these data do not provide evidence that A1BG expression changes meaningfully with age in this cohort.

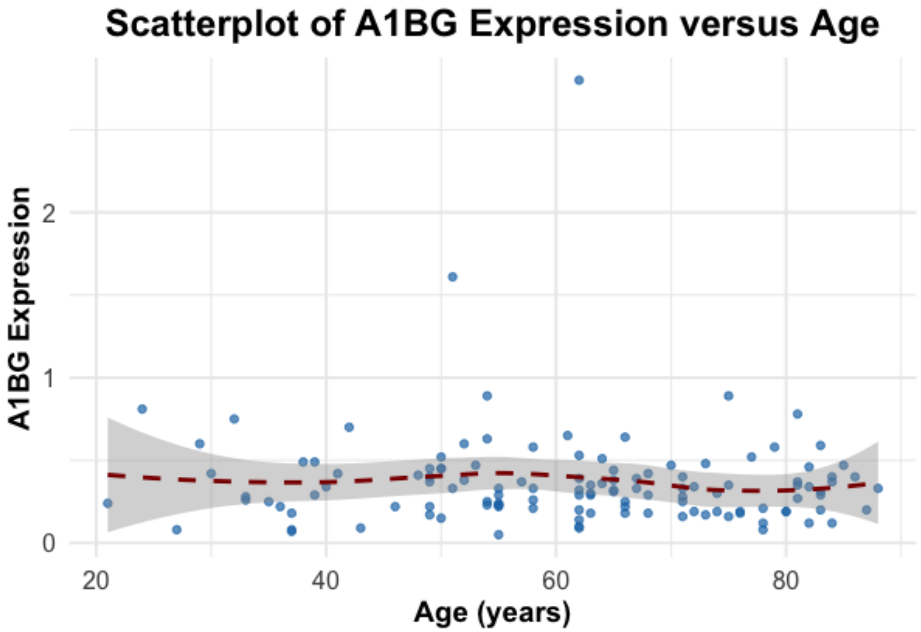


Figure 2: Scatter plot of A1BG expression vs. age with LOESS smoother; points colored by ICU status and shaped by sex.

### 3.4 Boxplot of gene stratified by two categorical covariates

The boxplots stratified by sex and ICU status (Figure 3) show broadly similar A1BG expression between females and males, with overlapping interquartile ranges and comparable medians. When further grouped by ICU status ("yes" vs "no"), median differences remain small; non-ICU participants tend to have slightly higher central tendency, but the spread overlaps substantially across groups. The "unknown" sex category has very limited data (essentially a single value), so it should not be over-interpreted. Taken together, these comparisons suggest that large sex-specific differences are unlikely, and any ICU-related shifts in A1BG expression—if present—are modest.

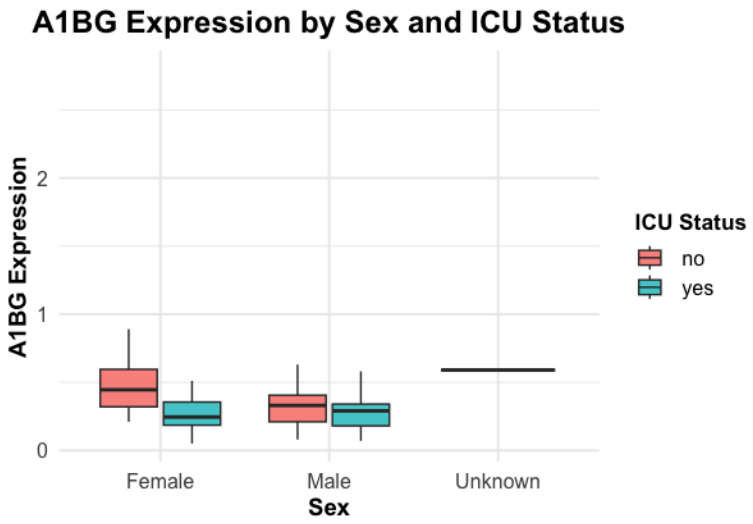


Figure 3: Boxplot of A1BG expression by sex and ICU status.

### 3.5 Heatmap

The heatmap of the top variance genes (Figure 4) summarizes multigene patterns after row z-scoring and hierarchical clustering. Column annotations reveal partial organization of samples by ICU status, with several gene blocks exhibiting coordinated up- or down-regulation among ICU subjects relative to non-ICU. Sex annotation shows no dominant block-level pattern, consistent with the boxplot findings. Because the visualization uses a subset of samples/genes and row scaling emphasizes relative (not absolute) differences, the heatmap highlights structure that is consistent with global disease-severity effects rather than single-gene changes.

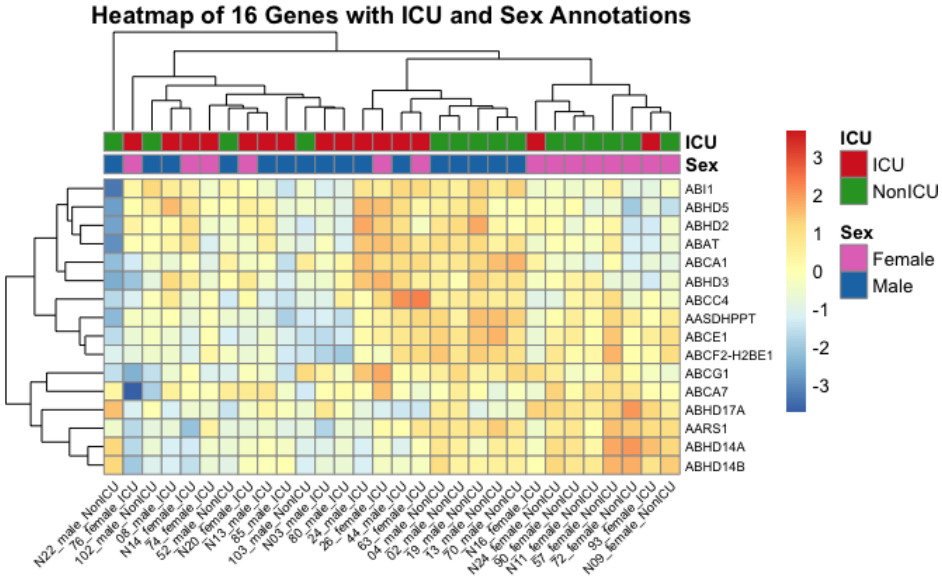


Figure 4: Heatmap of the top 16 variance genes. Rows and columns are clustered with Euclidean distance and complete linkage. Annotation bars indicate sex and ICU status.

### 3.6 New plot type (Hexbin)

To provide a new view of the relationship between age and gene expression beyond the standard scatterplot, I used a **hexbin plot** implemented in `ggplot2::geom_hex`. A hexbin plot partitions the 2D plane into small hexagonal cells and colours each cell by the number of observations it contains. This effectively solves overplotting when many points overlap and reveals the underlying *density structure* of the data.

**What is shown.** Each panel displays patients from one ICU stratum (ICU vs. NonICU). The **x-axis** is age (years) and the **y-axis** is A1BG expression. The **colour scale** (legend "Count") indicates how many patients fall into each hexagonal bin; darker/brighter cells correspond to higher local density.

**How it was produced.** From the merged analysis dataset, I filtered records with finite age and A1BG values, harmonised ICU labels, and then drew:

- `geom_hex(bins = 28)` to create hexagonal bins (28 bins per axis; larger values give finer resolution).
- `facet_wrap(~ ICU)` to show ICU and NonICU side by side for direct comparison.
- `scale_fill_viridis_c()` to apply a perceptually uniform colour scale for counts.

**What it adds.** Unlike a scatterplot that can obscure dense regions through overplotting, the hexbin view highlights *where* observations concentrate. In this dataset, most patients cluster around ages ∼55–80 with A1BG expression in the ∼0.2–0.6 range across both strata, while a few sparse high-expression cells appear only in the NonICU panel, indicating rare outliers. The side-by-side panels make it easy to spot any shift in the bulk density between ICU and NonICU groups.

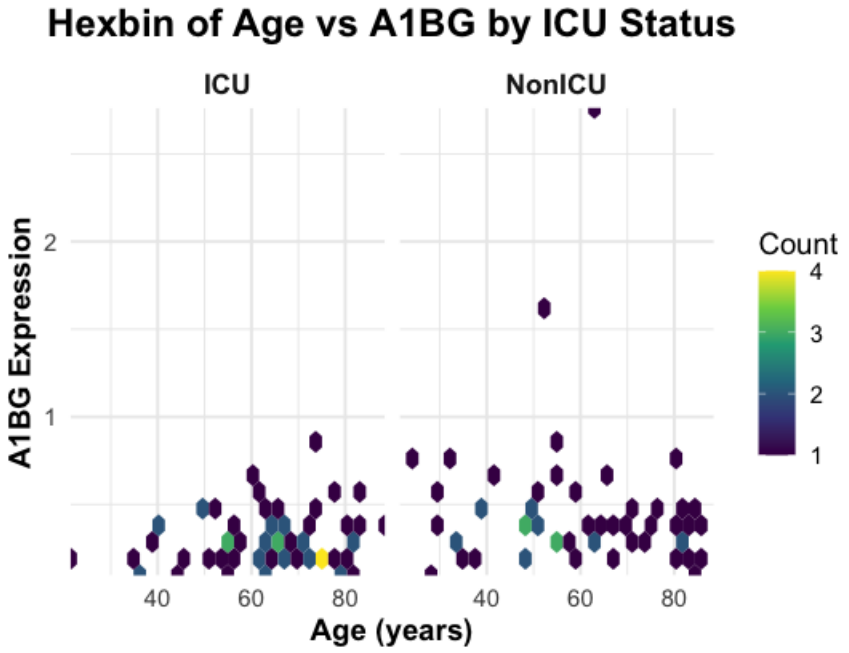


Figure 5: Hexbin of Age vs A1BG Expression, stratified by ICU status.

# 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, **??**). 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.