Analysis of Gene Expression in COVID-19 Patients

Prerana Dubey

August 21, 2025

1 Introduction

The dataset analyzed in this report was obtained from the Gene Expression Omnibus (GSE157103) [1]. The dataset consists of RNA-seq and high-resolution mass spectrometry data from 128 plasma and leukocyte samples collected from hospitalized patients with and without COVID-19 (n=102 and 26, respectively) across varying disease severities. This multi-omic resource enables systems-level analysis of molecular features associated with COVID-19 status and severity, supporting both biological discovery and outcome prediction. This dataset includes transcriptomic data from COVID-19 patients and healthy controls. For this project, the primary gene of interest is **ABHD17B**, which was explored across different covariates. The dataset includes demographic and clinical variables such as sex, disease status, and ventilation status. This analysis uses basic statistical summaries, visualizations, and clustering methods to examine associations between gene expression and patient characteristics.

2 Methods

Data were processed and analyzed using R (version 4.5.0). Key packages included `ggplot2` [2] for plots, `pheatmap` [3] for heatmap generation and kable package for table one generation [4]. The continuous variables of choice were age, ferritin (ng.ml.) and, Charlson score, and the categorical variables chosen were biological sex, ICU status and Disease status with respect to COVID-19 diagnosis. Continuous variables were summarized using mean (SD), while categorical variables were summarized as counts and percentages in Table 1. We also analysed different genes of interest by plotting them against three covariates. The gene of interest here was ABHD17B. Several plots like a histogram (ref. Fig. 2), scatterplot(ref. Fig. 3) and boxplot(ref. Fig. 4) were created to analyze the expression of this gene and its effects on the participants.

3 Results

In Table 1, we present baseline characteristics of the study population stratified by sex. Categorical variables, including ICU status and disease status, are summarized as counts with percentages, while continuous variables such as age, ferritin, and Charlson score are summarized as mean (SD). Overall, the table provides a clear comparison of demographic and clinical features between male and female participants.

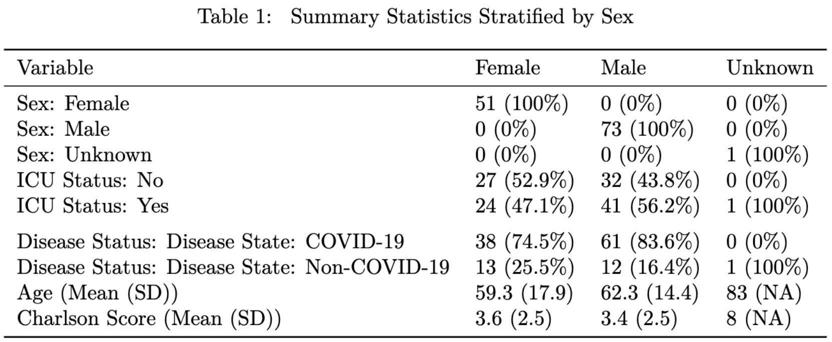

Table 1: Summary Statistics Stratified by Sex

Variable	Female	Male	Unknown
Sex: Female	51 (100%)	0 (0%)	0 (0%)
Sex: Male	0 (0%)	73 (100%)	0 (0%)
Sex: Unknown	0 (0%)	0 (0%)	1 (100%)
ICU Status: No	27 (52.9%)	32 (43.8%)	0 (0%)
ICU Status: Yes	24 (47.1%)	41 (56.2%)	1 (100%)
Disease Status: Disease State: COVID-19	38 (74.5%)	61 (83.6%)	0 (0%)
Disease Status: Disease State: Non-COVID-19	13 (25.5%)	12 (16.4%)	1 (100%)
Age (Mean (SD))	59.3 (17.9)	62.3 (14.4)	83 (NA)
Charlson Score (Mean (SD))	3.6 (2.5)	3.4 (2.5)	8 (NA)

Figure 1: Baseline characteristics of the study population stratified by sex. Categorical variables are presented as counts with percentages, and continuous variables are presented as mean (SD). *Note:* Missing values were excluded from percentage and mean (SD) calculations. Percentages are calculated within each sex stratum.

We plot several graphs to study the expression of the gene ABHD17B. The histogram in Figure 2 shows ABHD17B expression values (x) against the frequency of samples (y). The distribution appears to be normal where in the highest number of samples show an average gene expression of 7.5 as per the histogram.

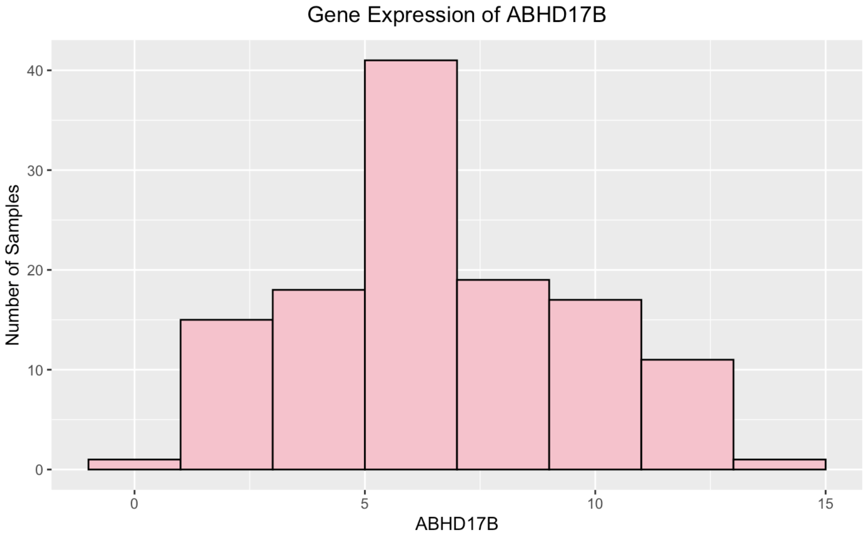

Figure 2: Histogram of ABHD17B expression: Each bar represents a range of expression values (width = 2.0).The height of the bar shows how many samples had expression values in that range.

Next, we plot the expression of gene ABHD17B against the age of all the participants using a scatterplot in Figure 3. The points are widely scattered without a clear upward or downward trend. The variability in expression appears similar across age groups (no obvious clustering by age). This suggests no strong linear relationship between age and ABHD17B expression. So There does not appear to be a clear association between age and ABHD17B gene expression in this dataset. Expression values are variable but relatively consistent across the age spectrum.

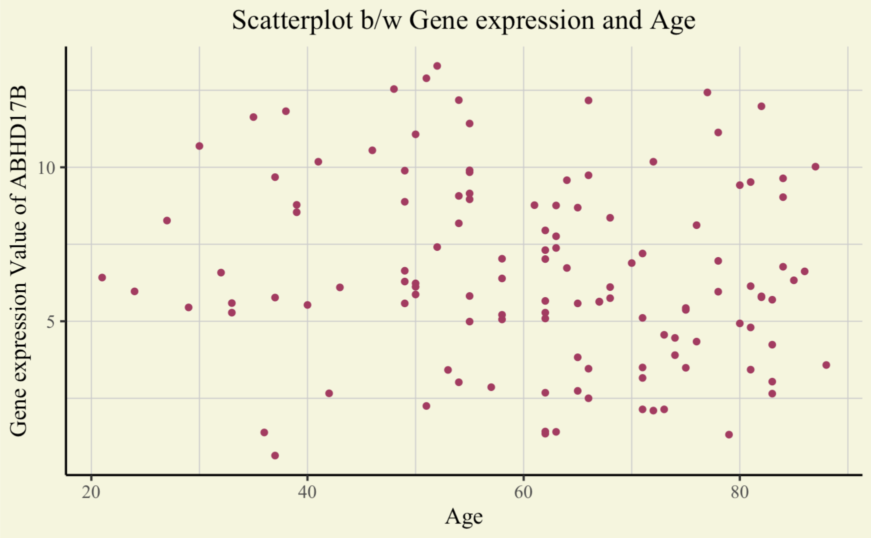

Figure 3: Scatter plot of ABHD17B expression vs. Age.

We also used a boxplot in Figure 4 to assess the distribution of ABHD17B gene expression levels, grouped by sex (male/female), and further split by disease status (COVID-19 vs Non-COVID-19). Each box shows the distribution of values (interquartile range), with the median marked by the horizontal line. In both males and females, COVID-19 patients display higher ABHD17B expression compared to non-COVID controls, with the effect most pronounced in females. This suggests that ABHD17B is upregulated in the context of COVID-19.

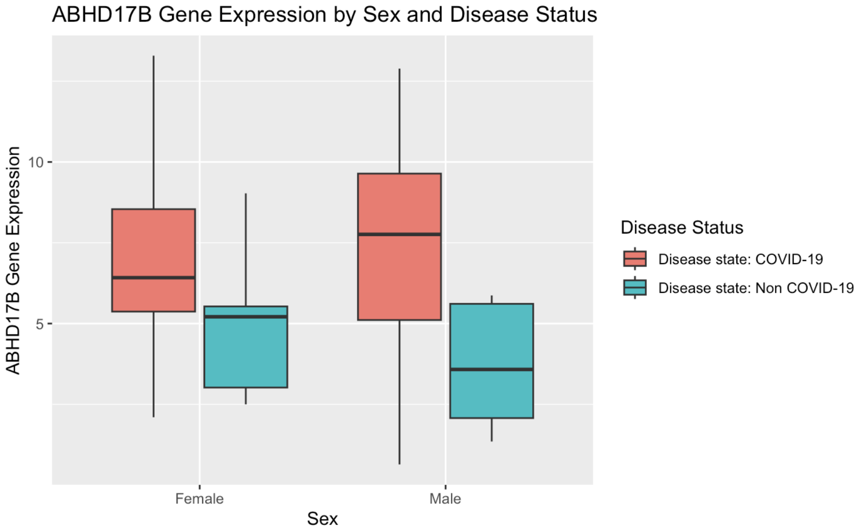

Figure 4: Boxplot of ABHD17B expression stratified by disease status and sex.

We then made use of a heat map to visualize patterns of gene expression across samples and identify relationships between genes and clinical factors. Heatmaps make it easy to see which genes are expressed at similar levels across patients, and whether certain genes cluster together, suggesting shared regulation or biological function. By clustering samples, we can see if patients with similar characteristics (like sex or ventilation status) tend to group together based on their gene expression, which could hint at biological differences.

This heatmap displays the expression patterns of several genes across different samples, with hierarchical clustering used to group samples and genes based on similarity in expression. The color gradient (blue to red) represents relative gene expression levels, with blue indicating lower expression and red indicating higher expression. The annotations on the left indicate sample characteristics: **Sex** (blue = male, orange = female, grey = unknown) and **Ventilation status** (pink = yes, purple = no). From the clustering, we can see distinct blocks of genes (e.g., AANAT, ABHD16A, ABHD14B) that show high expression in certain groups and low expression in others, suggesting that subsets of genes may be co-regulated or differentially expressed depending on biological or clinical factors. The annotation bars allow us to visually explore whether sex or ventilation status might correspond to distinct gene expression patterns.

The heatmap in Figure 5 shows a clustered heatmap of 10 selected genes. The rows represent participants and the columns represent genes. Tracking bars indicate sex and mechanical ventilation status.

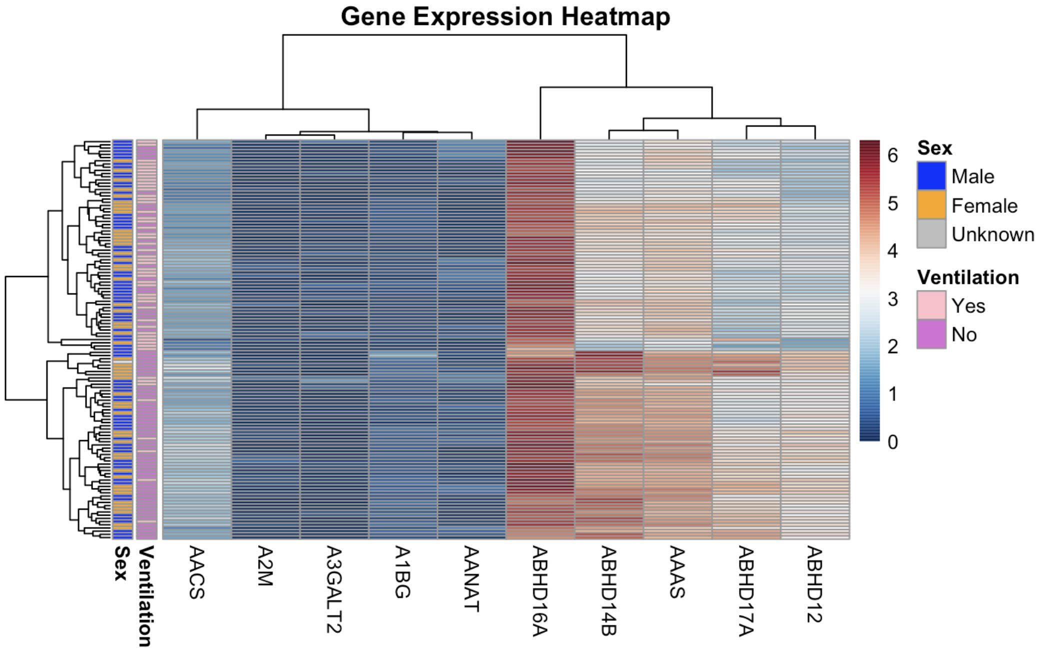

Figure 5: Heatmap of selected genes with annotation bars for sex and mechanical ventilation.

Lastly, we used a violin plot in Figure 6 to compare the distribution of a gene's expression between two groups (COVID-19 vs. non-COVID-19). It combines the features of a boxplot (showing median and spread) with a density plot (showing how the data are distributed across values).

The violin plot in Figure 6 compares AANAT gene expression between COVID-19 patients and non-COVID-19 patients. The wider parts of the violin show where most values are concentrated, and the boxplot inside shows the median and spread. We can see that the COVID-19 group has a wider range and some higher expression values compared to the non-COVID-19 group, which is more tightly clustered at lower levels. This suggests that AANAT expression may be influenced by disease status.

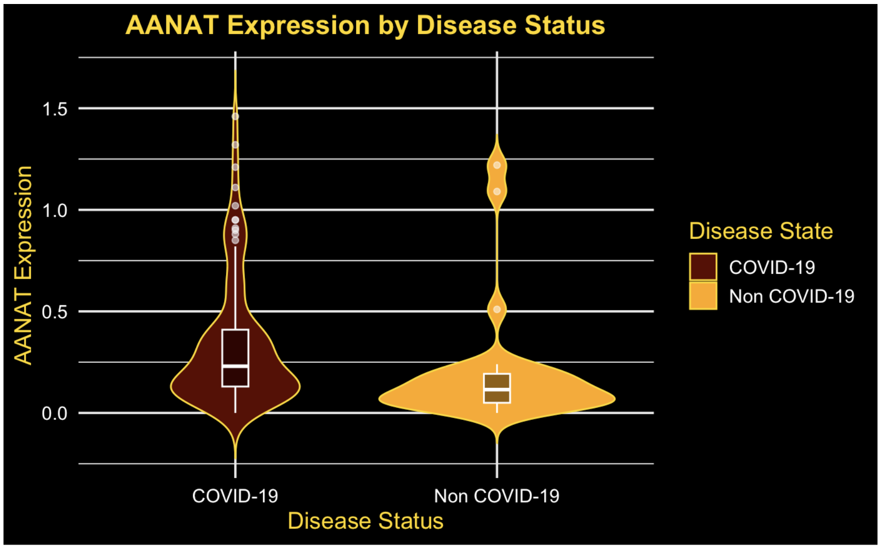

Figure 6: Violin plot of AANAT gene expression by disease status.

4 Discussion

In this project, we explored gene expression patterns in COVID-19 and non-COVID-19 patients using several visualization methods. The main gene we focused on, **ABHD17B**, did not show a strong relationship with age, but it did appear to be expressed at higher levels in COVID-19 patients, especially in females. This suggests that ABHD17B may be influenced by disease status rather than demographic factors like age. The heatmap also showed some clustering by ventilation status, which could mean that disease severity has an effect on overall gene expression. Finally, the violin plot of AANAT expression indicated more variability and slightly higher expression in COVID-19 patients compared to controls.

5 Conclusion

Overall, this analysis showed that gene expression can differ between COVID-19 and non-COVID-19 patients, with some evidence that both ABHD17B and AANAT may be involved. However, this was a small and exploratory project, so the results should be interpreted with caution. The patterns we observed provide ideas for future studies but do not confirm any direct biological role of these genes in COVID-19. More advanced statistical testing and larger datasets would be needed to draw stronger conclusions.

References

[1] Emily Clough and Tanya Barrett. The gene expression omnibus database. In *Statistical Genomics: Methods and Protocols*, pages 93–110. Springer, 2016.

[2] Hadley Wickham. *ggplot2: Elegant Graphics for Data Analysis.* Springer-Verlag New York, 2016.

[3] Raivo Kolde. *pheatmap: Pretty Heatmaps*, 2025. R package version 1.0.13.

[4] Hao Zhu. *kableExtra: Construct Complex Table with 'kable' and Pipe Syntax*, 2024. R package version 1.4.0.