QBS 103 Final Project

Shuwen Hou

August 2025

Contents

1 Introduction

The dataset analyzed in this project originates from Overmyer et al. (2021), who performed a large-scale multi-omic study of COVID-19 severity. The data, available in the Gene Expression Omnibus (GSE157103), include whole blood RNA sequencing together with detailed clinical covariates such as age, sex, Charlson score, and ventilator-free days. For this analysis, we focus on the gene *ABCB1* as the primary feature of interest to explore its distribution and relationship with clinical covariates.

2 Methods

2.1 Data Source

The data for this analysis were obtained from the study by Overmyer et al. (2021). This dataset includes whole blood RNA sequencing and clinical metadata from patients with or without COVID-19.

2.2 Software and Packages

All analyses were conducted in `R` version 4.3.3 (R Core Team, 2024). The following packages were used: `tidyverse`, `ggplot2`, `pheatmap`, `dplyr`, `tidyr` (Wickham et al., 2019; Wickham, 2016; Kolde, 2025; Wickham et al., 2023, 2024).

2.3 Variables

The primary gene of interest was `ABCB1`. Continuous covariates include `charlson_score`, `age`, and `ventilator_free_days`. Categorical covariates included `disease_status`, `sex`, and `icu_status`.

2.4 Statistical Summaries and Graphics

Continuous variables were summarized as mean (standard deviation) and median [IQR]. Categorical variables were summarized as counts and percentages. Plots included histograms, scatter plots, and boxplots for the main gene of interest, as well as a heatmap of the top 10 genes and an additional `geom_ribbon` plot to display mean expression across Charlson score with standard error ribbons. Heatmap rows and columns were clustered using Euclidean distance and complete linkage.

3 Results

3.1 Table of Summary Statistics

Table 1: Summary statistics of covariates stratified by `disease_status`.

	COVID-19	non-COVID-19	Total
Continuous variables[mean(sd)/median(IQR)]			
Age	60.8 (16.1) / 62 [50.3, 73.8]	62.8 (15.6) / 65 [53, 75]	61.2 (16.0) / 62 [50.5, 74.0]
Charlson score	3.28 (2.48) / 3 [1, 5]	4.35 (2.42) / 4 [2.25, 6]	3.50 (2.49) / 3 [2, 5]
Ventilator-free days	19.81 (11.56) / 28 [10.5, 28]	22.42 (10.06) / 28 [24, 28]	20.35 (11.28) / 28 [12.8, 28]
Categorical variables			
Sex			
Female	38 (38.0%)	13 (50.0%)	51 (40.5%)
Male	62 (62.0%)	12 (46.2%)	74 (58.7%)
Unknown	0 (0.0%)	1 (3.8%)	1 (0.8%)
ICU status			
No	50 (50.0%)	10 (38.5%)	60 (47.6%)
Yes	50 (50.0%)	16 (61.5%)	66 (52.4%)

Result: As shown in Table 1, patients with COVID-19 had a slightly lower mean age compared to non-COVID-19 patients (60.8 vs. 62.8 years). Charlson scores were lower among COVID-19 patients than in non-COVID-19 patients (mean 3.28 vs. 4.35). Ventilator-free days were fewer in the COVID-19 group relative to the non-COVID-19 group (mean 19.81 vs. 22.42). The distribution of sex was similar across groups, while a greater proportion of non-COVID-19 patients required ICU admission (61.5% vs. 50.0%).

3.2 Histogram of `ABCB1` Expression

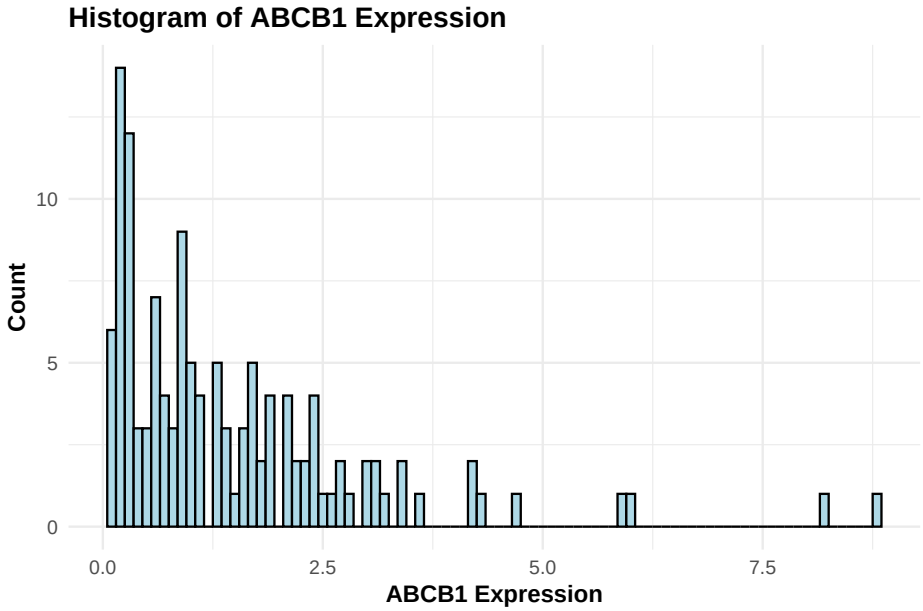

Figure 1: Histogram of `ABCB1` expression.

Result: As shown in Figure 1, the distribution of *ABCB1* expression was not uniform and tend to sit on the left, with most samples clustered at low expression levels and a small number having much higher expression. The long right tail indicated the presence of potential outliers of individuals with high *ABCB1* expression. This plot suggests that summary statistics such as the mean may be strongly influenced by a few high-expression cases.

3.3 Scatter Plot: `ABCB1` vs. Charlson Score

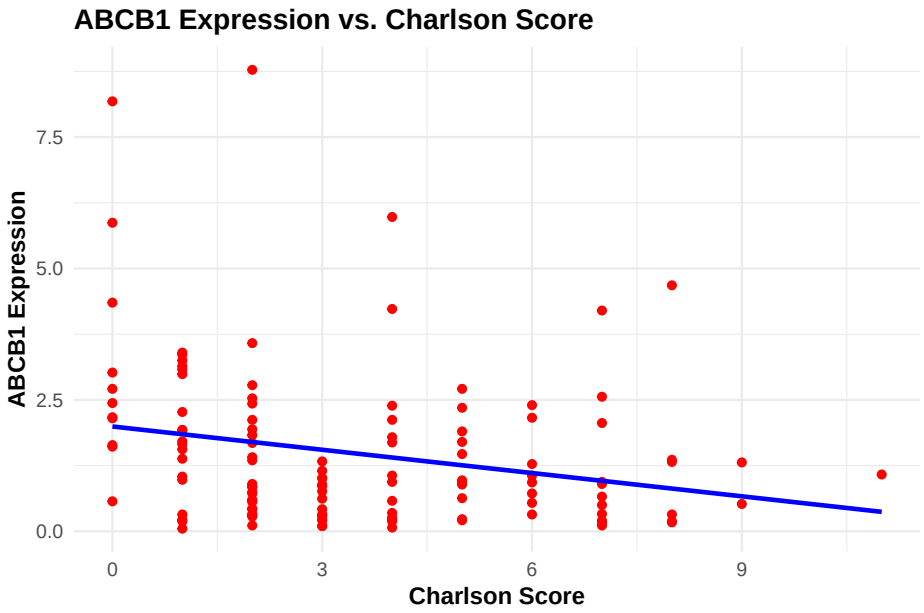

Figure 2: Scatter plot of `ABCB1` expression vs. Charlson score with fitted line.

Result: As shown in Figure 2, *ABCB1* expression did not appear to have a strong relationship with Charlson score. While the fitted regression line suggested a slight negative slope, the data showed wide variability in expression at each Charlson score value, with some outliers. This indicates that what measured by Charlson score was not a major cause of variation in *ABCB1* expression in this dataset.

3.4 Boxplot: `ABCB1` Expression by `sex` and `disease_status`

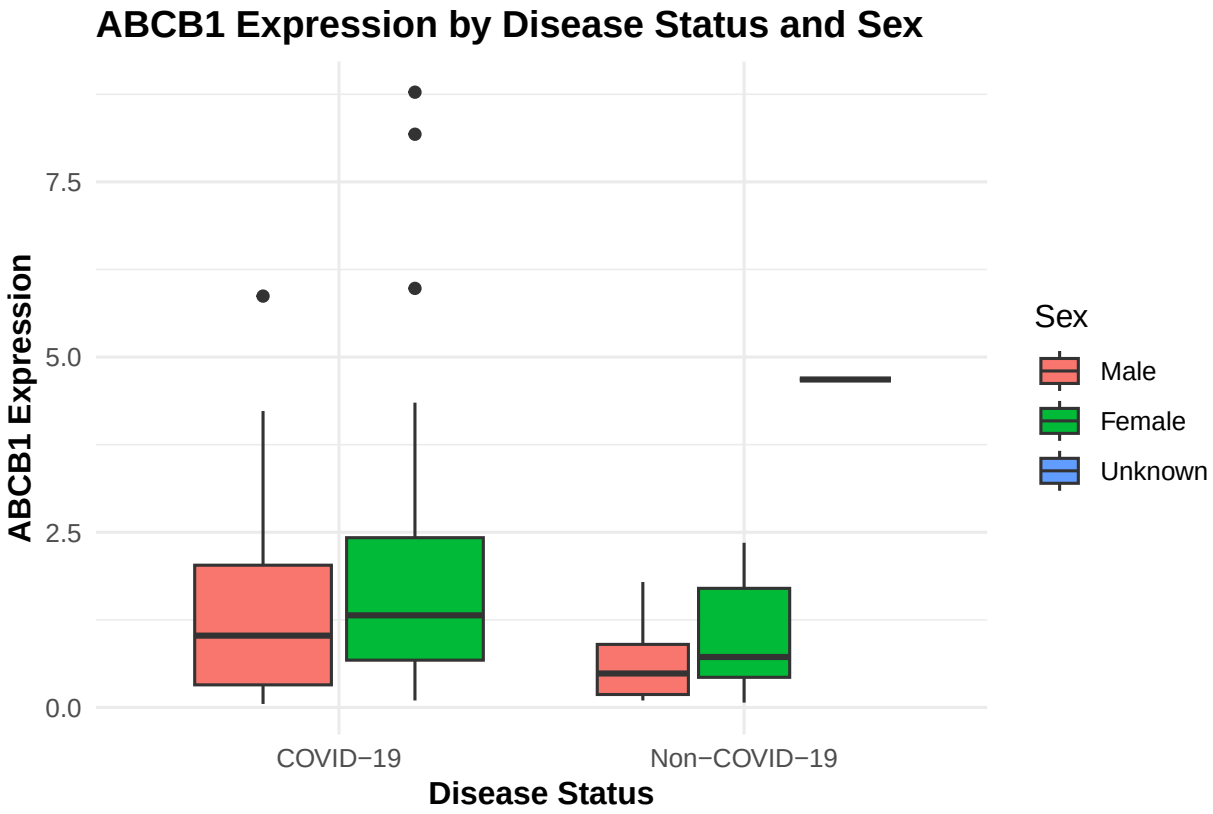

Figure 3: Boxplot of `ABCB1` expression stratified by `sex` and `disease_status`.

Result: As shown in Figure 3, *ABCB1* expression varied by both sex and disease status. Among males, median expression was lower in COVID-19 patients compared to non-COVID-19 patients, while expression among females showed less difference between the groups. The spread of expression values was generally greater in the COVID-19 group, with a few outliers appeared. These patterns suggest that sex may have some effects in the relationship between disease status and *ABCB1* expression.

3.5 Heatmap

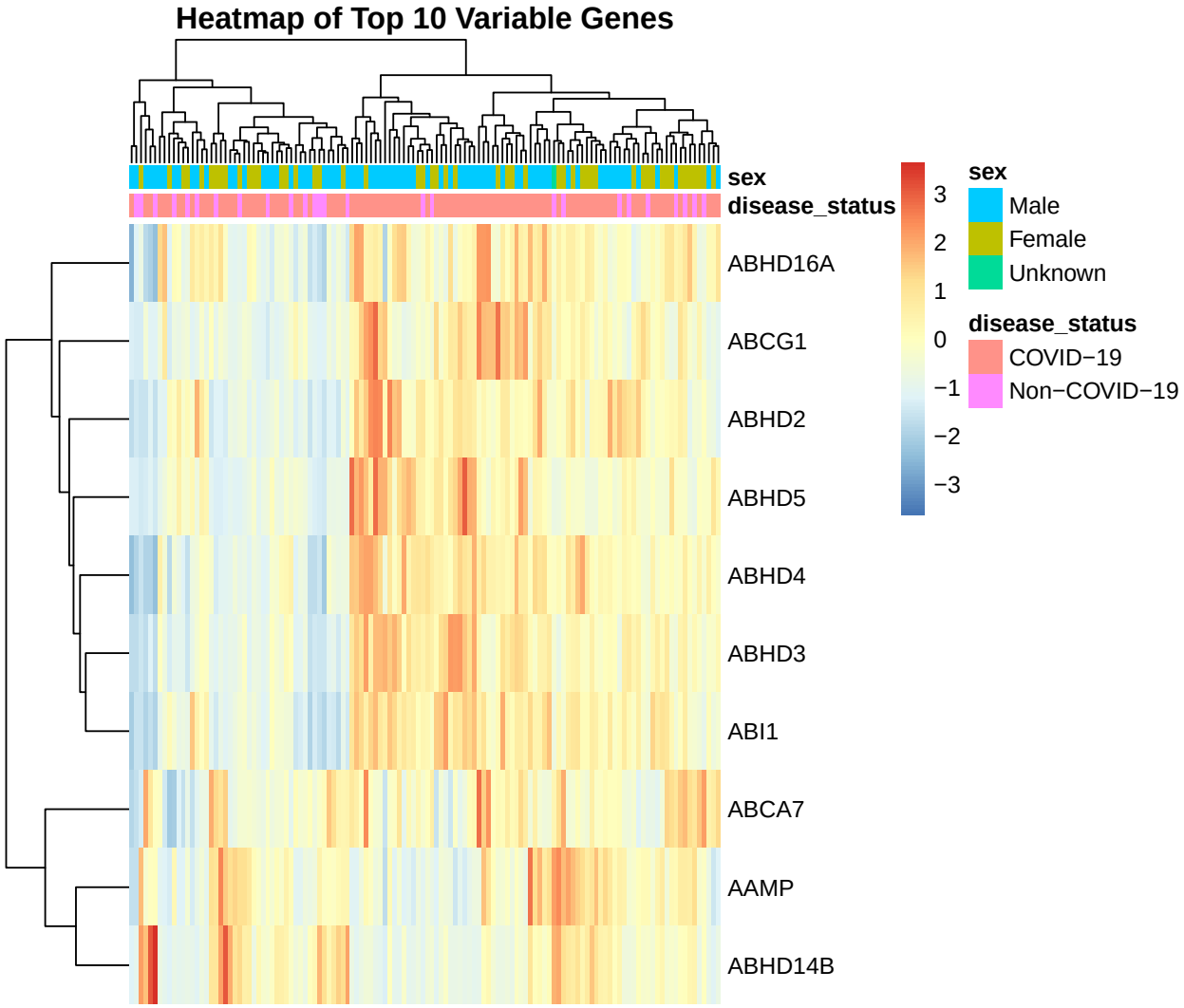

Figure 4: Heatmap of the top 10 genes with clustered rows and columns with annotation bars for `sex` and `disease_status`.

Result: As shown in Figure 4, the heatmap of the top 10 variable genes showed clear differences in expression patterns across samples. Many of the COVID-19 cases grouped together and appeared distinct from the non-COVID-19 controls, although there was still some overlap. Certain genes, such as *ABHD16A* and *ABCG1*, displayed noticeable changes in expression between the two groups. The annotation bars also highlight that clustering was partly related to disease status, while sex differences were less obvious.

3.6 Ribbon Plot

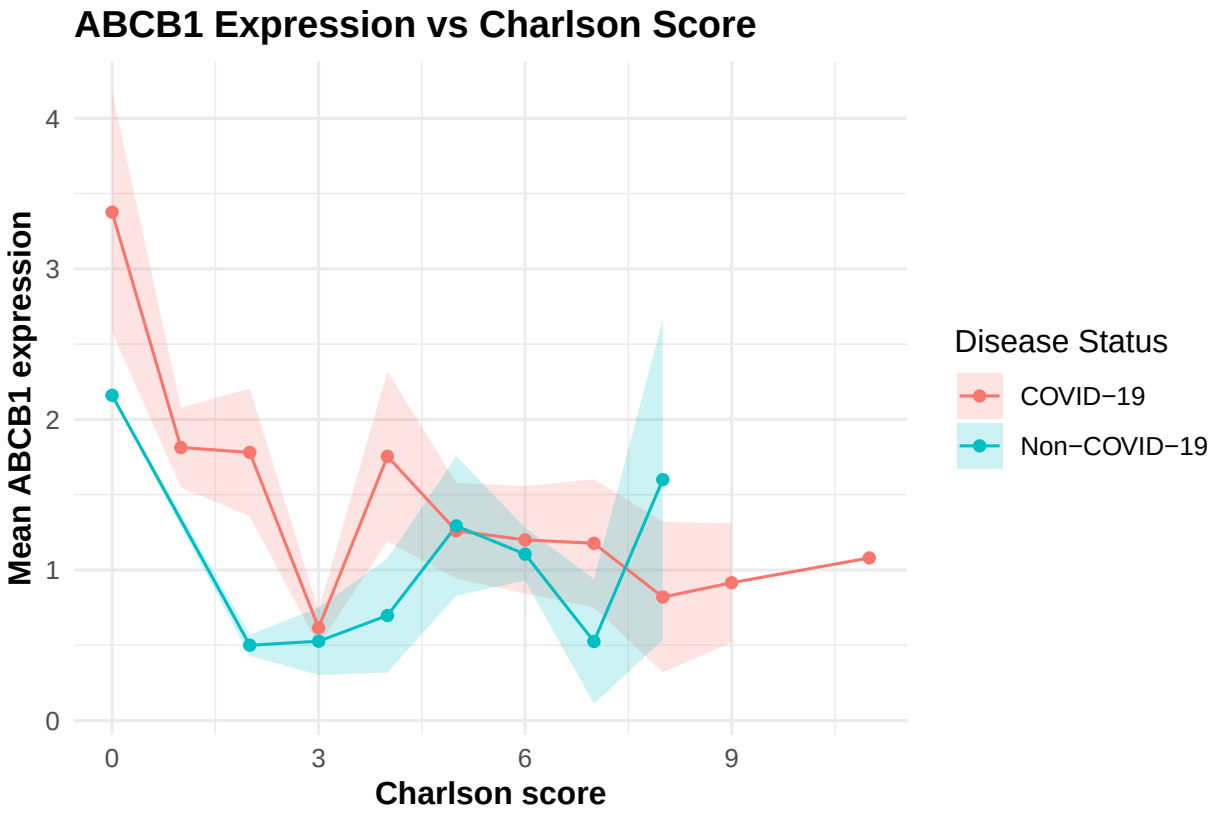

Figure 5: Ribbon plot of mean `ABCB1` expression and `charlson_score`

Result: As shown in Figure 5, mean *ABCB1* expression tended to decrease as Charlson score increased, although this is not very clear and obvious. COVID-19 patients started with higher expression at low Charlson scores, but their levels dropped quickly to a lower level. Non-COVID-19 patients showed lower initial expression, with smaller changes across scores. The shaded ribbons illustrate wide variability, suggesting considerable overlap between the two groups.

4 References

References

Kolde, R. (2025). *pheatmap: Pretty Heatmaps*. R package version 1.0.13.

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. (2021). Large-Scale Multi-omic Analysis of COVID-19 Severity. *Cell Systems*, 12(1):23–40.e7.

R Core Team (2024). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria.

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

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., Takahashi, K., Vaughan, D., Wilke, C., Woo, K., and Yutani, H. (2019). Welcome to the tidyverse. *Journal of Open Source Software*, 4(43):1686.

Wickham, H., François, R., Henry, L., Müller, K., and Vaughan, D. (2023). *dplyr: A Grammar of Data Manipulation*. R package version 1.1.4.

Wickham, H., Vaughan, D., and Girlich, M. (2024). *tidyr: Tidy Messy Data*. R package version 1.3.1.