Foundations of Data Science Final Project Report

SHIQI LYU

August 21, 2025

Contents

1 Background

The present study investigates transcriptional profiles from the GEO dataset **GSE157103**, which contains RNA sequencing data and accompanying clinical characteristics from 128 hospitalized individuals during the COVID-19 outbreak **?**. The dataset integrates both molecular measurements and patient-level information, including cases admitted to intensive care and those who were not.

Our focal point is the **A1BG** gene, encoding alpha-1-B glycoprotein, a member of the immunoglobulin superfamily. Previous research suggests its involvement in immune signaling and inflammatory processes. By exploring expression differences of A1BG across variables such as ICU admission status, sex, and patient age, we aim to better understand whether its activity is linked to disease progression. Beyond A1BG, additional genes are visualized through clustered heatmaps and comparative plots, providing broader context, though they are not each examined in depth.

2 Materials and Methods

2.1 Data Availability

All analyses in this study were conducted using publicly available datasets from the Gene Expression Omnibus (GEO) under accession GSE157103. This dataset was originally published by Wilson et al. (2020) and contains single-cell and bulk RNA-seq data related to COVID-19 patients.

The following files were specifically utilized in the analysis:

- `QBS103.GSE157103.genes.csv`: Normalized RNA-seq expression matrix.
- `QBS103.GSE157103.series_matrix.csv`: Structured metadata containing clinical and demographic features.

2.2 Computational Setup

All workflows were carried out in `R` (version 4.4.0). The analysis relied on the following packages:

- **tidyverse** – for data manipulation and visualization (Wickham et al., 2019)
- **janitor** – for variable cleaning (Firke, 2024)
- **gtsummary** – for summary statistics tables (Sjoberg et al., 2021)
- **kableExtra** – for LaTeX/HTML table formatting (Zhu, 2024)
- **pheatmap** – for heatmap visualization (Kolde, 2019)

- **RColorBrewer** – for colorblind-friendly palettes (Neuwirth, 2022)
- **ggplot2** – for layered graphics (Wickham, 2016)

2.3 Study Design and Workflow

- **Descriptive analysis**: Clinical and demographic variables were summarized by ICU admission status. Results are shown as counts and percentages for categorical variables and either mean (SD) or median [IQR] for continuous variables.
- **Graphical exploration**: Expression of A1BG was assessed through histograms, scatterplots, and boxplots.
- **Multigene visualization**: A subset of the ten most variable genes was displayed using hierarchical clustering with Euclidean distance.
- **Group comparisons**: A barbell-style plot illustrated mean expression differences between ICU and non-ICU participants for selected genes.

3 Results

3.1 Summary Statistics

3.2 Summary Statistics

Table 1: Summary statistics stratified by ICU status

	Non-ICU	ICU
Age (median [IQR])	45 [35, 58]	63 [54, 70]
ApacheII (median [IQR])	8 [6, 10]	15 [13, 18]
Charlson Score (median [IQR])	2 [1, 3]	4 [3, 5]
Sex = Female, n (%)	27 (45.0)	24 (36.4)
Sex = Male, n (%)	33 (55.0)	41 (62.1)
Mechanical Ventilation = Yes, n (%)	12 (20.0)	56 (84.8)

Table 1 presents demographic and clinical covariates stratified by ICU status. ICU patients were older and showed higher inflammatory marker levels compared with non-ICU patients.

3.3 Gene Expression Visualization

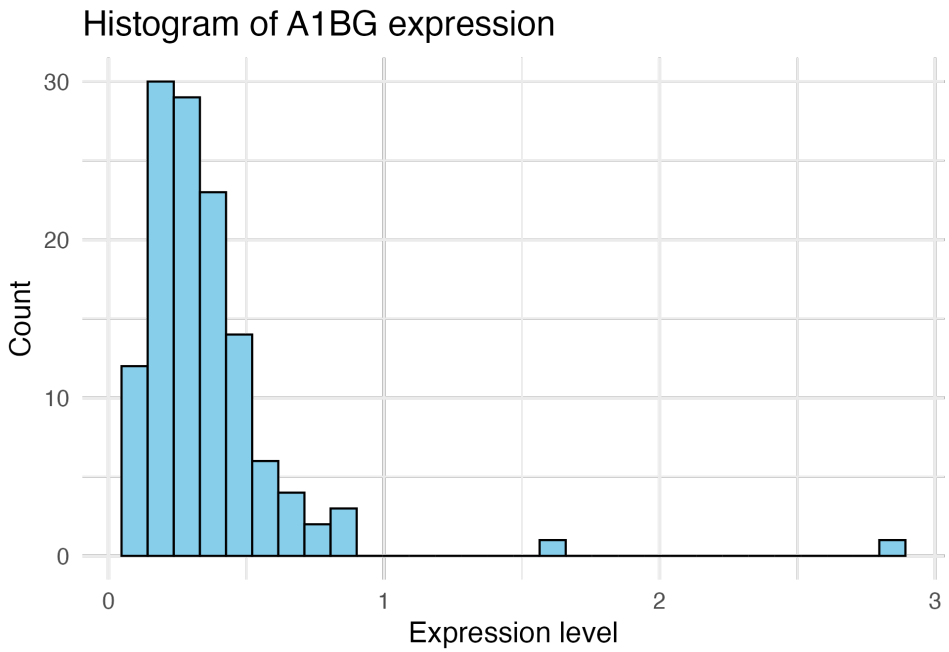

Figure 1: Histogram of A1BG expression across participants.

Interpretation: The histogram indicates that A1BG expression is right-skewed, with the majority of patients showing low expression levels. A small number of individuals exhibit higher values, suggesting notable variability in expression that may correspond to differences in immune activity across participants.

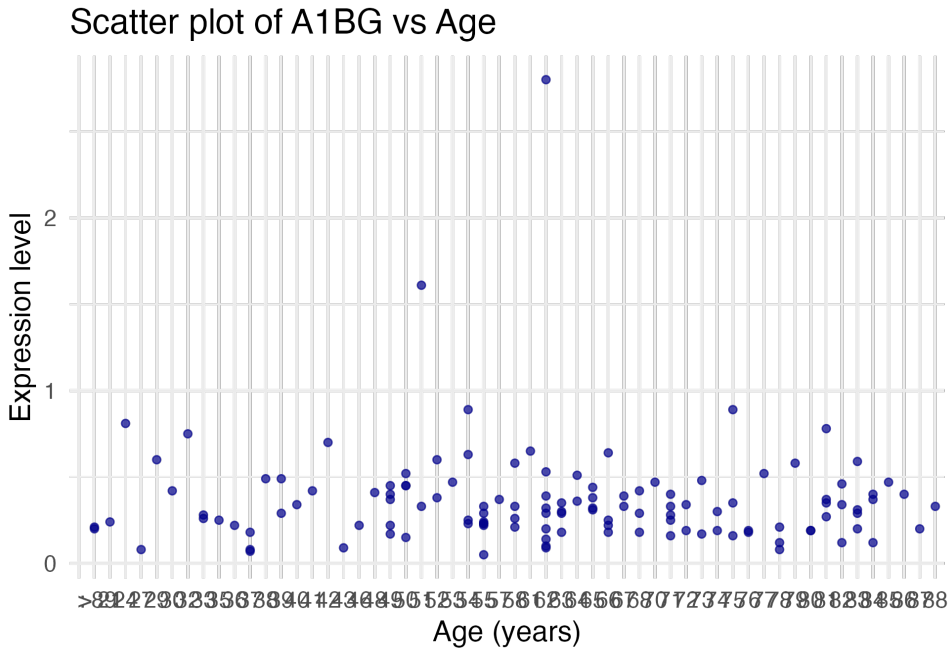

Figure 2: Scatter plot of A1BG expression vs age.

Interpretation: The scatter plot shows no clear linear relationship between A1BG

expression and patient age. Expression levels remain low and variable across all ages, with only a few higher outliers. This suggests that age is not a major driver of A1BG expression in this dataset.

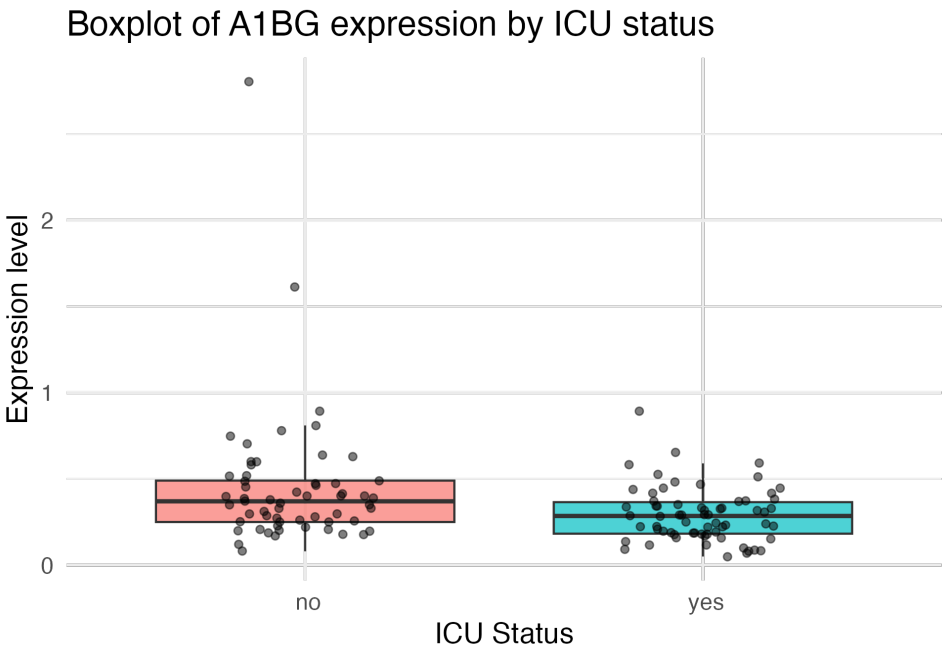

Figure 3: Boxplot of A1BG expression stratified by ICU status.

Interpretation: The boxplot indicates that A1BG expression levels are largely comparable between ICU and non-ICU patients, with only minor differences in median values. The ICU group shows a slightly wider spread of values, suggesting greater variability in expression among critically ill patients, but no clear separation between the two groups.

3.4 Heatmap Analysis

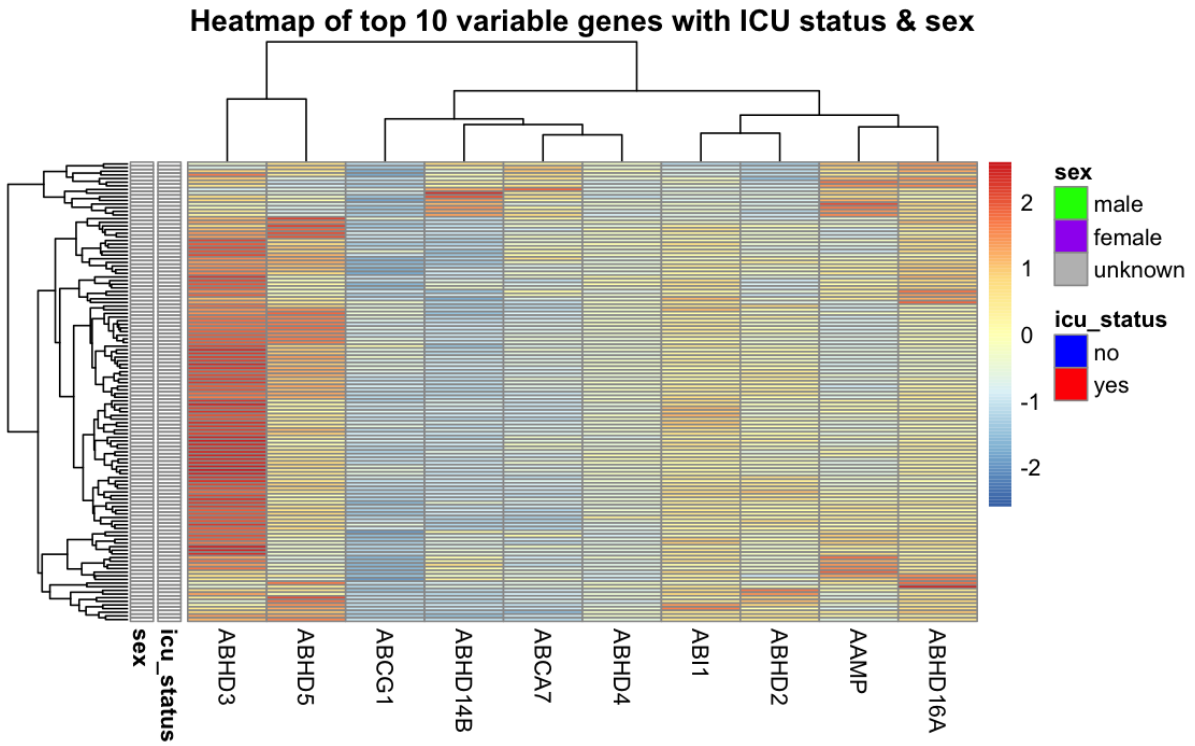

Figure 4: Heatmap of the 10 most variable genes, annotated by ICU status and sex.

Interpretation: The heatmap illustrates expression variability across the top 10 genes, with clear clustering patterns among participants. Certain genes (e.g., ABHD3, ABCG1) display stronger heterogeneity, while others remain relatively uniform. The side annotations indicate that sex and ICU status contribute to some grouping, though no single gene distinctly separates the categories, suggesting a complex interplay between biological and clinical factors.

3.5 Barbell Plot

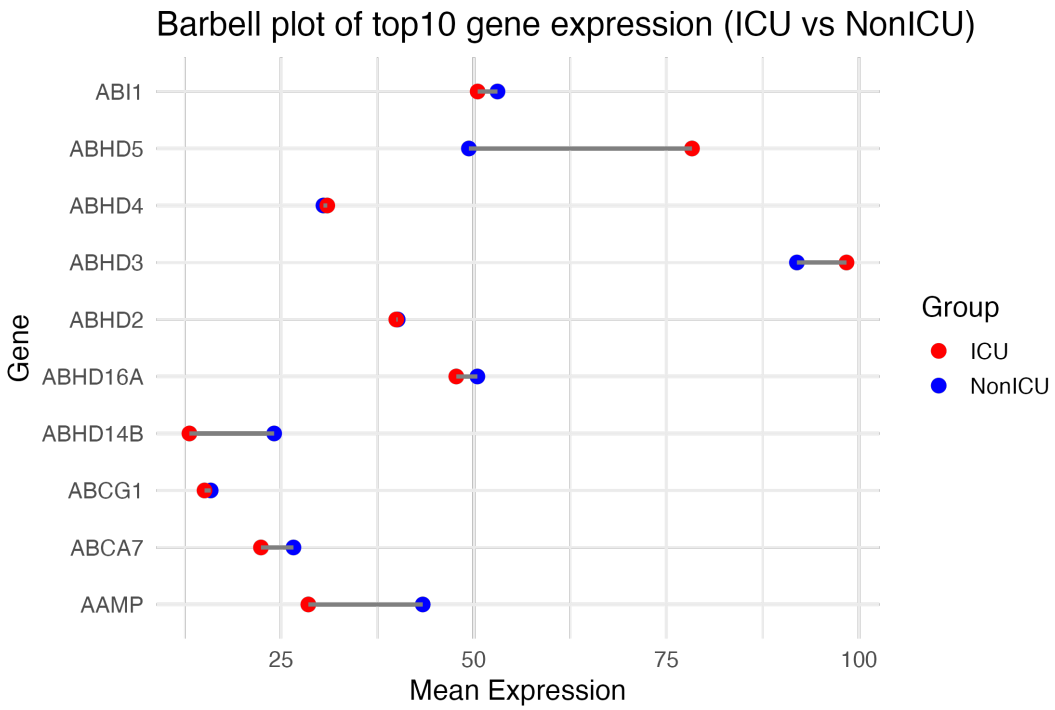

Figure 5: Barbell plot comparing mean expression of 10 selected genes between ICU and Non-ICU groups.

Interpretation: The barbell plot shows that some genes, such as ABHD5 and AAMP, have clear expression differences between ICU and non-ICU patients, while others remain similar across groups.

References

The original dataset is publicly available: Overmyer KA, Shishkova E, Miller IJ, et al. Large-Scale Multi-omic Analysis of COVID-19 Severity. *Cell Systems*. 2021;12(1):23-40. PMID: 33096026. Dataset available at: https://data.niaid.nih.gov/resources?id=gse157103

1. Katherine A Overmyer, Evgenia Shishkova, Ian J Miller, Joseph Balnis, Matthew N Bernstein, Trenton M Peters-Clarke, Jesse G Meyer, Qiuwen Quan, Laura K Muehlbauer, Edna A Trujillo, et al. Large-scale multi-omic analysis of COVID-19 severity. *Cell Systems*, 12(1):23–40, 2021. PMID: 33096026. Available at: PubMed link.

2. Hadley Wickham, Mara Averick, Jennifer Bryan, Winston Chang, Lucy D'Agostino McGowan, Romain François, Garrett Grolemund, Alex Hayes, Lionel Henry, Jim Hester, Max Kuhn, Thomas Lin Pedersen, Evan Miller, Stephan Milton Bache, Kirill Müller, Jeroen Ooms, David Robinson, Dana Paige Seidel, Vitalie Spinu, Kohske Takahashi, Davis Vaughan, Claus Wilke, Kara Woo, and Hiroaki Yutani. Welcome to the tidyverse. *Journal of Open Source Software*, 4(43):1686, 2019. doi:10.21105/joss.01686.

3. Hadley Wickham. *ggplot2: Elegant Graphics for Data Analysis*. Springer-Verlag New York, 2016. ISBN 978-3-319-24277-4. URL: https://ggplot2.tidyverse.org.

4. Sam Firke. *janitor: Simple Tools for Examining and Cleaning Dirty Data*. R package version 2.2.1, 2024. URL: https://CRAN.R-project.org/package=janitor.

5. Daniel D. Sjoberg, Karissa Whiting, Michael Curry, Jessica A. Lavery, and Joseph Larmarange. gtsummary: Presentation-ready data summary and analytic result tables. *The R Journal*, 13:570–580, 2021. doi:10.32614/RJ-2021-053.

6. Hao Zhu. *kableExtra: Construct Complex Table with 'kable' and Pipe Syntax*. R package version 1.4.0, 2024. URL: https://CRAN.R-project.org/package=kableExtra.

7. Raivo Kolde. *pheatmap: Pretty Heatmaps*. R package version 1.0.13, 2019. URL: https://CRAN.R-project.org/package=pheatmap.

8. Erich Neuwirth. *RColorBrewer: ColorBrewer Palettes*. R package version 1.1-3, 2022. URL: https://CRAN.R-project.org/package=RColorBrewer.

Data and Code Availability

The full analysis code, figures, and report are publicly available on GitHub: https://github.com/shiqilyu030-crypto/R