Analysis of Data From ’Large-Scale Multi-omic Analysis of COVID-19 Severity’, Overmyer et al. 2021

QBS 103 Final Project
Project Author: Sabin Hart
Contact: sabin.d.hart.gr@dartmouth.edu

***Abstract*— This study performs a detailed analysis of data from Overmyer et al.’s 2021 ”Large-scale multi-omic analysis of COVID-19 severity,” focusing on the gene ABCA7. Using R, the study explores the relationship between ABCA7 expression and COVID-19 severity, ICU status, and demographic factors. Key findings include a negative correlation between ABCA7 expression and disease severity, variations in expression across sex and ICU status, and the importance of normalization in visualizing gene expression patterns. Summary statistics and visualizations are provided to support these findings.**

1. Introduction

The body of this study is built off the data and analysis of ”Large-scale multi-omic analysis of COVID-19 severity” (2021) by Overmyer et al [6]. The authors map 128 COVID-19 positive and negative blood samples with genotypical features, disease severity and symptomality. These data include the expression of nearly 100 genes as well as ICU status, some demographic features, and evaluative scores.

I focus my analysis primarily on the gene ABCA7. According to the National Institute of Health, ABCA7 is a member of the ABC1 subfamily of ATP-binding casette transporters, the only subfamily found exclusively in multicellular eukaryotes. Its exact function is not yet known ”...however, the expression pattern suggests a role in lipid homeostasis in cells of the immune system. [1]

2. Data and Methods

The data were collected, cleaned, and prepared using R Version 4.0.4. In addition to the Base R functionality, the following packages were used: tidyverse - for a variety of data processing and visualization techniques [8]; pheatmap - for visualizing the heatmaps [4]; ggthemes - for additional data visualization themes [2]; stringr - for better string operations [7]; rlist - for better list operations [5]; and stargazer - for outputting R data in LaTeX tables [3].

The data were provided in two parts: a gene expression chart and a phenotype information chart. By transposing and joining based on the participants’ IDs, an overall data table was created.

With the data, I create a histogram, which displays the distribution of ABCA7 expression. I also create a scatter plot between ABCA7 and disease severity, with a smoothed linear regression, and a boxplot showing the gene’s association with sex and ICU status.

In addition, I plot heatmaps of a subset of 15 genes against 15 participants with highest variance. Clustering is performed to better display results, using euclidean distance, or the square root of the sum of squared distances as the clustering value. Since these are genes and absolute expression can vary heavily, I present both a crude and a normalized heatmap. The normalization is such that the mean expression is zero and standard deviation is one.

The final plot is a density plot of the expression of a subset of 5 of the genes chosen for the heatmap. This shows their relative densities at different levels of expression. These plots are all displayed in Section 3.

With the data complete, the summary statistics are displayed in Table 1. Stratified by sex, results for the continuous variables (Age, Apache II Score, and Ferritin) are displayed as ’mean (standard deviation)’ while discrete variables (On Mechanical Ventilation and In ICU) are displayed with ’count (percent%)’.

3. Results

Figure 1 displays the ABCA7 Expression Histogram. It seems to be roughly normally distributed around its mean, 24.

Table 1: Summary Statistics

Sex	Age	Apache II	Ferritin (ng/mL)	Mechanical Ventilation	In ICU	Count
Female	66.5 (16.2)	21.4 (8.9)	668 (1087)	16 (31.4%)	24 (47.1%)	51
Male	62.3 (12.1)	20.7 (7.5)	1073 (911)	35 (47.3%)	41 (55.4%)	74

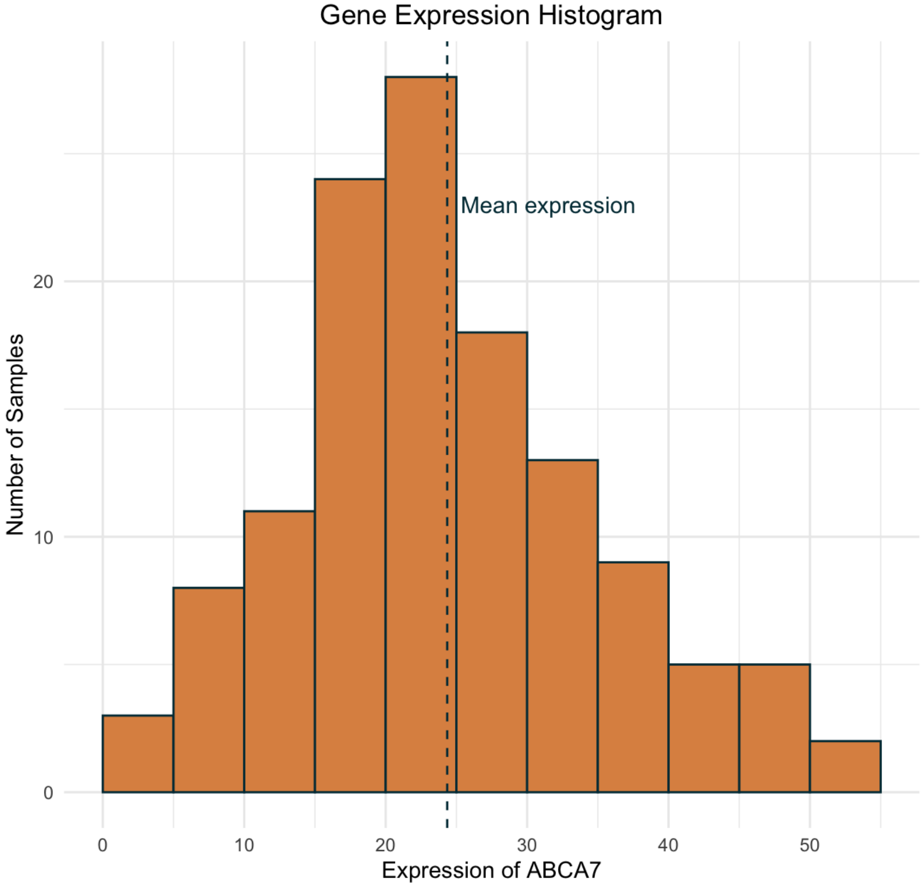

Figure 1: Expression Histogram of ABCA7

The Apache II score is a measure of disease severity where a higher score indicates greater medical duress. It has a negative correlation with ABCA7 expression. (Figure 2)

ABCA7 expression is appears reduced generally in those in the ICU as compared to those not, and seems to be generally lower in Males versus Females. As Figure 3 shows, there are more high outliers in men and more low outliers in women.

Without scaling, the genes appear to be highly clustered into three groups: five genes with an expression around 5, five around -5, and five around -15. Figure 4 displays this very clearly, with few recognizable patterns otherwise, although most of the outlying gene expressions are associated with Non-ICU patients.

The visual difference in Figure 5 displays the importance of scaling. Once normalized, the genes appear much less stratified and have no discernible patterns. Across Sex and ICU status there are no visible associations.

The final plot, displayed in Figure 6 is the density plot of a subset of five genes. Like the scaled heatmap, I normalize the gene expression to highlight relative differences rather than absolute. ABHD14A.ACY1 and ABC12 have narrow density bands arranged around the mean of zero while ABCD2, ABCD4, and ABCA2 have broader ranges with relatively lower density.

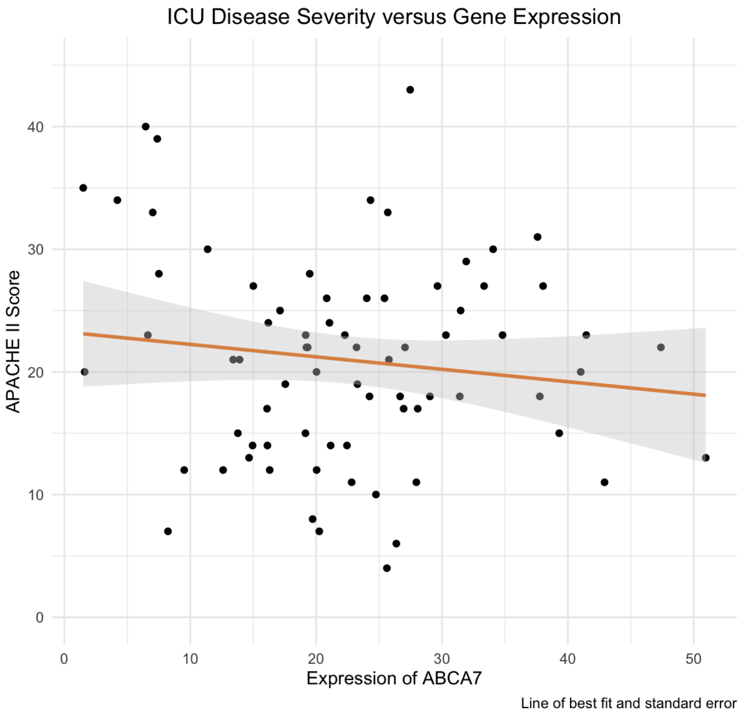

Figure 2: Scatterplot of ABCA7 vs. Apache II Score

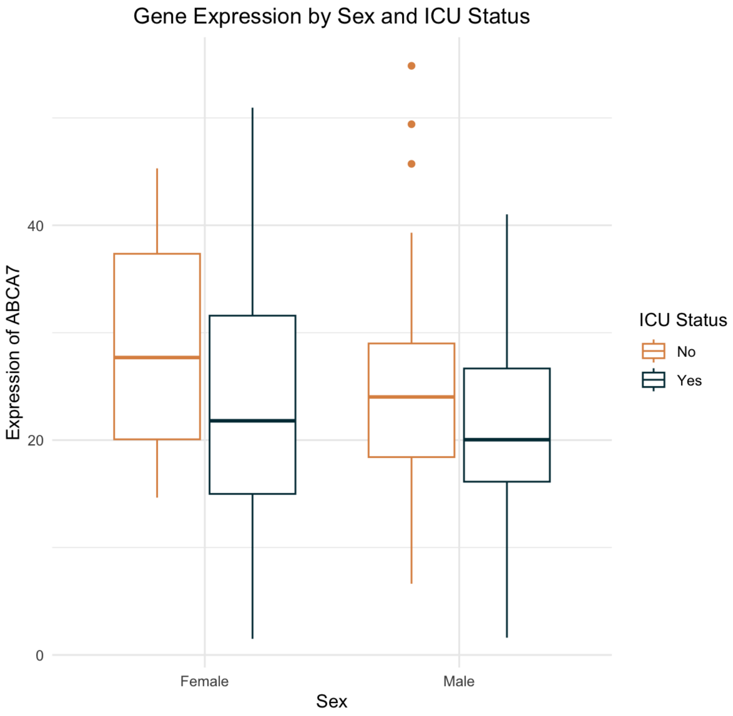

Figure 3: Boxplot of ABCA7 vs. Sex and ICU Status

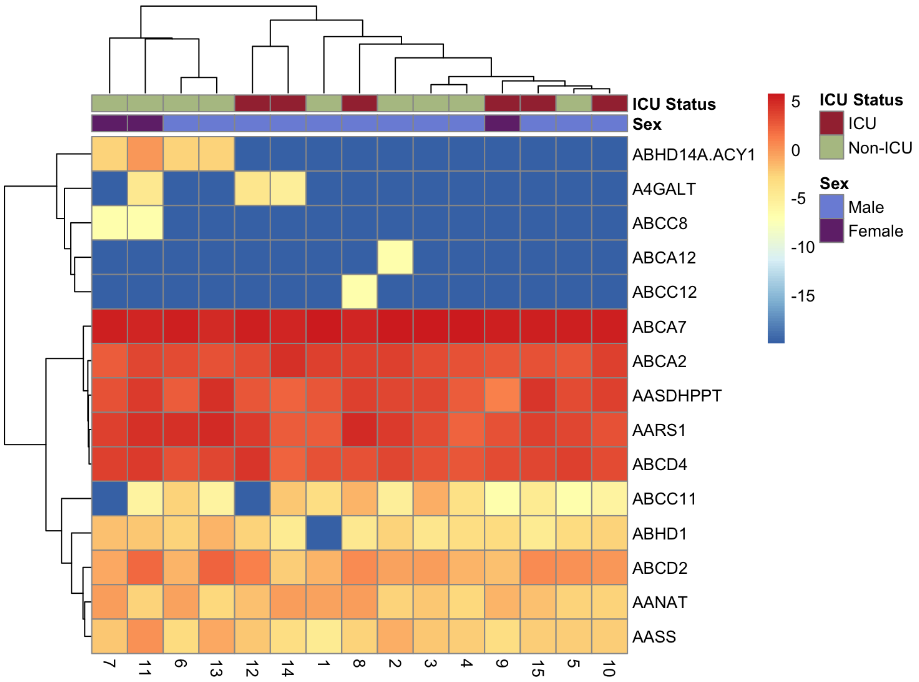

Figure 4: Unscaled Heatmap

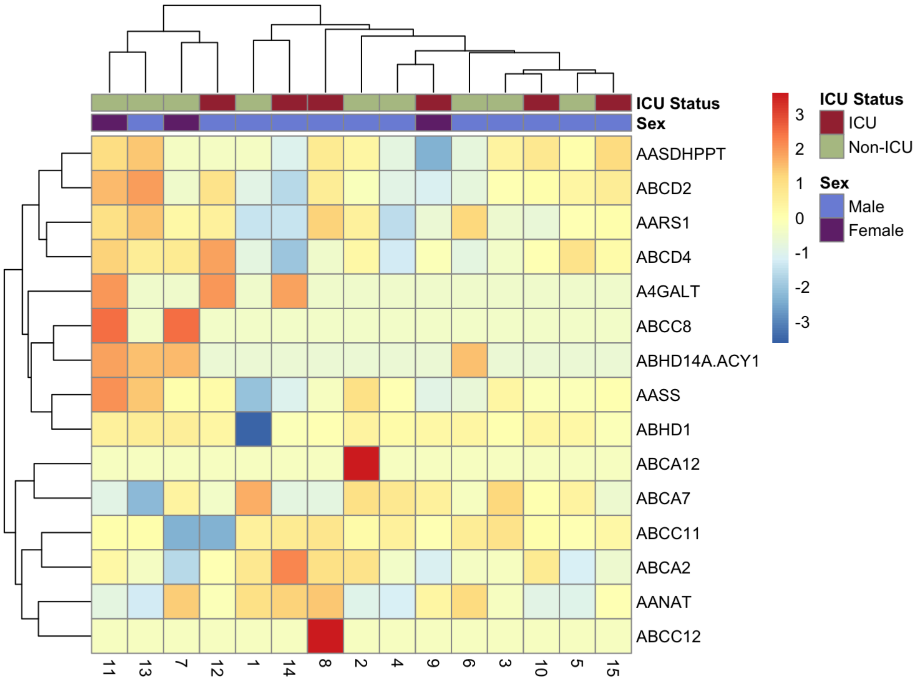

Figure 5: Scaled Heatmap

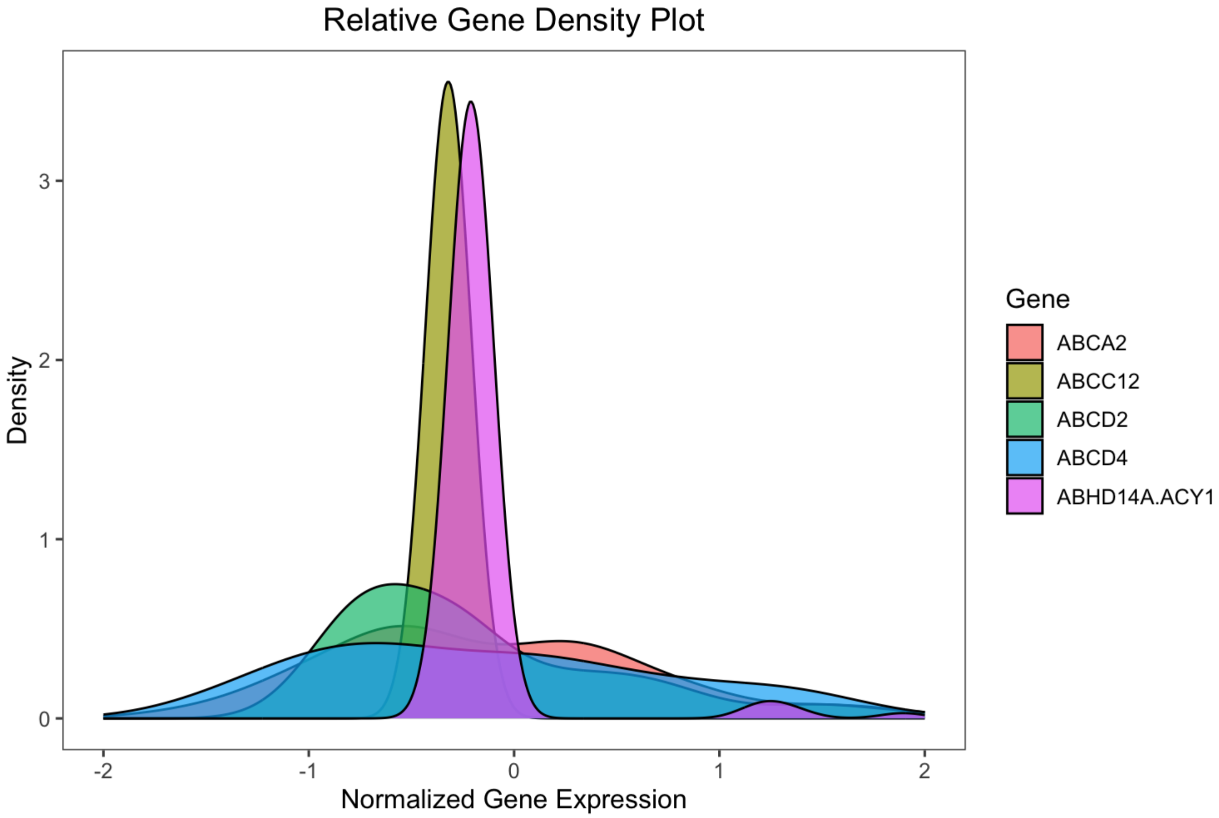

Figure 6: Density Plot of Gene Subset

References

[1] ABCA7 ATP binding cassette subfamily A member 7 [Homo sapiens (human)] - Gene - NCBI — ncbi.nlm.nih.gov. https://www.ncbi.nlm.nih.gov/gene/10347, August 2024. [Accessed 27-08-2024].

[2] Jeffrey B. Arnold. *ggthemes: Extra Themes, Scales and Geoms for 'ggplot2'*, 2022. R package version 4.2.4.

[3] Marek Hlavac. *stargazer: Well-Formatted Regression and Summary Statistics Tables*, 2018. R package version 5.2.3.

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

[5] Ren Kun. *rlist: A Toolbox for Non-Tabular Data Manipulation*, 2021. R package version 0.4.6.1.

[6] 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, Yuchen He, Amit Chopra, Hau C Chieng, Anupama Tiwari, Marc A Judson, Brett Paulson, Dain R Brademan, Yunyun Zhu, Lia R Serrano, Vanessa Linke, Lisa A Drake, Alejandro P Adam, Bradford S Schwartz, Harold A Singer, Scott Swanson, Deane F Mosher, Ron Stewart, Joshua J Coon, and Ariel Jaitovich. Large-scale multi-omic analysis of COVID-19 severity. *Cell Syst.*, 12(1):23–40.e7, January 2021.

[7] Hadley Wickham. *stringr: Simple, Consistent Wrappers for Common String Operations*, 2023. R package version 1.5.0.

[8] Hadley Wickham et al. *tidyverse: Easily Install and Load the 'Tidyverse'*, 2023. R package version 2.0.0.