Final Submission Report

Lakshmi Deepthi Gangiredla

August 2024

Contents

1 Introduction

In this paper, we use the publicly available GSE157103 gene data set found in 'Large-Scale Multi-omic Analysis of COVID-19 Severity' to conduct an analysis on the gene expression of ABCA3 and how its expression differs based on different covariates such as age, sex, ICU status, and COVID-19 disease status. For background, the GSE157103 gene data set consists of RNA-seq and high-resolution mass spectrometry data of 128 different blood samples from COVID-19-positive and COVID-19-negative patients [4]. 219 different molecular features, including ABCA3 gene expression, that are known to be correlated with COVID-19 status and severity were mapped to understand how individuals respond to disease on a molecular level. Categorical and continuous characteristics regarding the patient health status and the blood sample such as patient age, charleson score, hospital free days, and ventilator free days were also documented in the dataset.

The ABCA3 gene is responsible for providing instructions to proteins involved in lung surfactant production. This surfactant can be found in the lining of lung tissue and comprises a mixture of phospholipids and proteins. The function of surfactant is to help humans breathe more easily by preventing the alveoli, air sacs in the lung, from sticking together after exhalation. In 'ABCA3 and LZTFL1 Polymorphisms and Risk of COVID-19 in the Czech Population', it was documented that ABCA3 plays a vital role in proper lung function and that it is highly plausible that lower levels of ABCA3 could be a contributor to COVID-19 susceptibility or severity [2]. In this analysis we will further investigate the gene expression of ABCA3 among the 128 samples found in this dataset.

2 Methods

Methods regarding sample collection and analysis of the GSE157103 gene data samples can be found in the paper "Large-Scale Multi-Omic Analysis of COVID-19 Severity". [4]

The entire analysis of this project was conducted in RStudio 2024.04.2+764 for macOS. The packages used for data wrangling anddata table formatting were dplyr and tidyverse [9] [8].

The analysis began by loading the BS103-GSE157103-series-matrix.csv and QBS103-GSE157103-genes.csv files found on the QBS 103 Canvas page into RStudio. Both CSV files were loaded into data frames, and the genes data was transposed and merged with the series matrix by participant ID to create a complete dataset that included both gene expression and co-variate data for each participant.

A summary table of summary statistics for age, hospital free days post-45 day follow-up, ventilator-free days, ICU status, and disease status stratified by sex was generated using the percent of unique categorical values, mean, sum, median, and IQR functions found in the tableone package [5]. Mean/Standard Deviation or Median/IQR statistics were selected to summarize the continuous variables after using the histogram function in base R to check the distribution of the values in each of the variables. Mean and standard deviation was used for age due to the normal distribution of the values, and median and IQR was used for hospital free days post-45 day follow-up and ventilator free days due to the non- normal distribution of values. All continuous variables were converted to numeric format for analysis. The table was formatted in LaTeX using the kableExtra package [6].

To generate the histogram of ABCA3 gene expression, scatter plot of ABCA3 gene expression by age, and boxplot of ABCA3 gene expression by sex and ICU status, the ggplot2 package was utilized [7]. Additionally, a Bee Swarm plot was generated using the ggplot2 and ggbeeswarm packages to create a plot of ABCA3 gene expression based on the interaction of sex and ICU status of patients [1].

The heat map of genes ABAT, ABCA1, ABCA10, ABCA12, ABCA13, ABCA2, ABCA3, ABCA4, ABCA5, ABCA6, and ABCA7 was generated using the pheatmap package in R [3]. The heat map was annotated by including tracking bars for patient sex and ICU status. The clustering algorithm used to cluster the heat map rows and columns by distance was the euclidean clustering method.

3 Results

3.1 Table of Summary Statistics

The table includes summary statistics of categorical variables ICU status and COVID-19 disease status, and continuous variables age, hospital free days post 45 day follow up, and ventilator free days grouped by patient sex (Table 1).

Table 1: Summary Table

	Female (N=51)	Male (N=74)	Unknown (N=1)
Age			
Mean (SD)	59.9 (18.3)	62.7 (14.7)	NA
Hospital Free Days			
Median [Min, Max]	34.0 [0, 44.0]	28.0 [0, 44.0]	30.0 [30.0, 30.0]
Ventilator Free Days			
Median [Min, Max]	28.0 [0, 28.0]	28.0 [0, 28.0]	28.0 [28.0, 28.0]
ICU_Status			
No	27 (52.9%)	33 (44.6%)	0 (0%)
Yes	24 (47.1%)	41 (55.4%)	1 (100%)
Disease_Status			
COVID-19	38 (74.5%)	62 (83.8%)	0 (0%)
Non-COVID-19	13 (25.5%)	12 (16.2%)	1 (100%)

3.2 Histogram of ABCA3 Gene Expression

The histogram of gene ABCA3 shows the frequency of gene expression values (Figure 1).

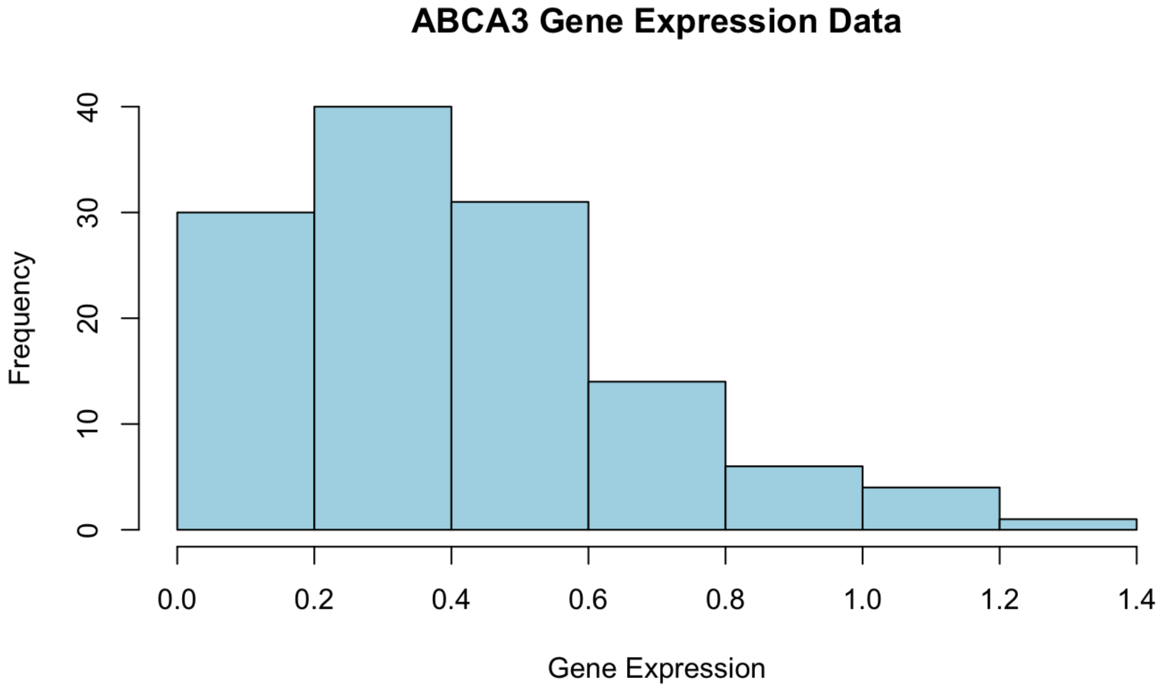

Figure 1: Histogram of ABCA3 gene expression

3.3 Scatter Plot of Gene Expression and Age

The scatter plot of ABCA3 gene expression by age is colored by patient age group (Figure 2). ABCA3 gene expression does not appear to be associated with age as there are no visible trends or clusters for any age group.

3.4 Heatmap of 10 Genes by Sex and ICU Status

The heatmap of genes ABAT, ABCA1, ABCA10, ABCA12, ABCA13, ABCA2, ABCA3, ABCA4, ABCA5, ABCA6, and ABCA7 shows gene expression values clustered for each patient in the dataset (Figure 3). The plot includes tracking bars for patient ICU status and sex. The histogram above does not appear to show any clear association between the gene expression of the different genes or categorical variables (ICU status and sex) among patients.

3.5 Box plot of Gene Stratified by Sex and ICU Status

The box plot shows ABCA3 gene expression by sex and ICU status (Figure 4). While it does not appear that sex is associated with ABCA3 gene expression, there appears to be higher expression of the gene among patients not in the ICU compared to those who are in the ICU.

3.6 Bee Swarm Plot of Gene by Sex and ICU Status Interaction

The bee swarm plot shows ABCA3 gene expression based on the interaction of sex and ICU status of patients (Figure 5). In this plot, it appears that gene expression of ABCA3 is highest among female and male patients who are not in the ICU compared to female and male patients who are in the ICU.

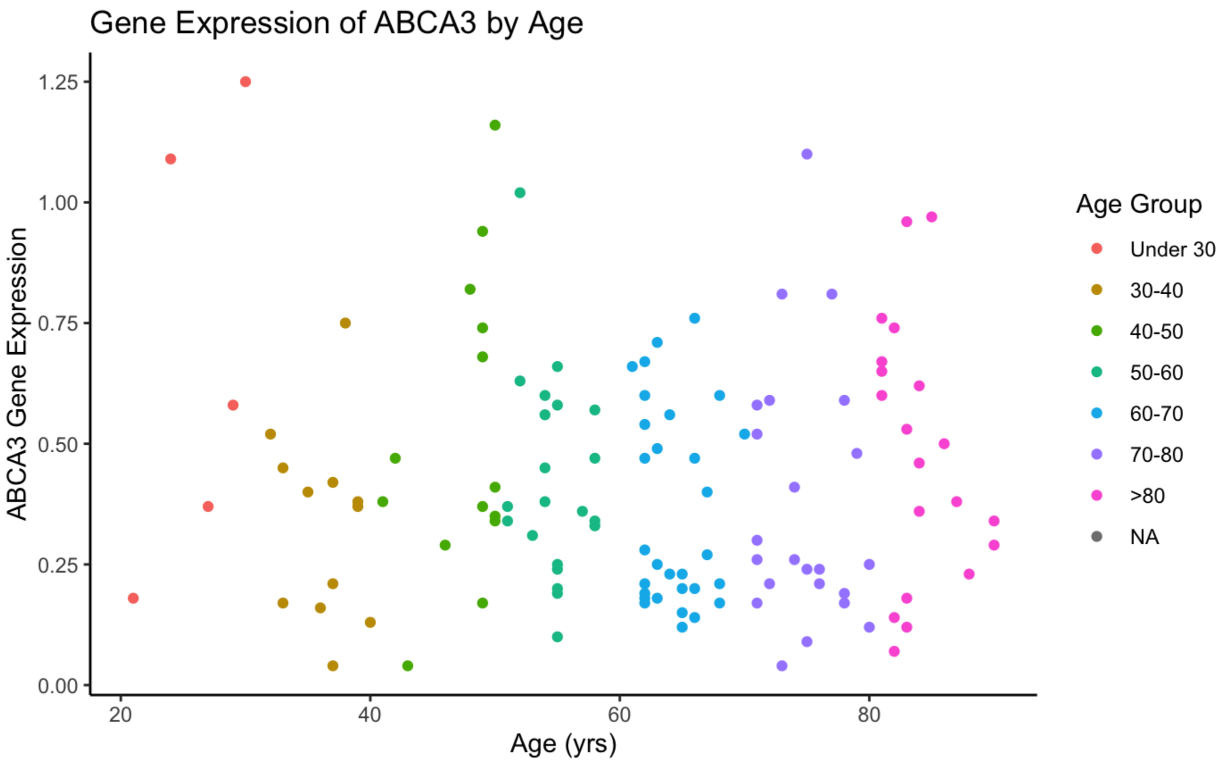

Figure 2: Scatter Plot of Gene Expression and Age

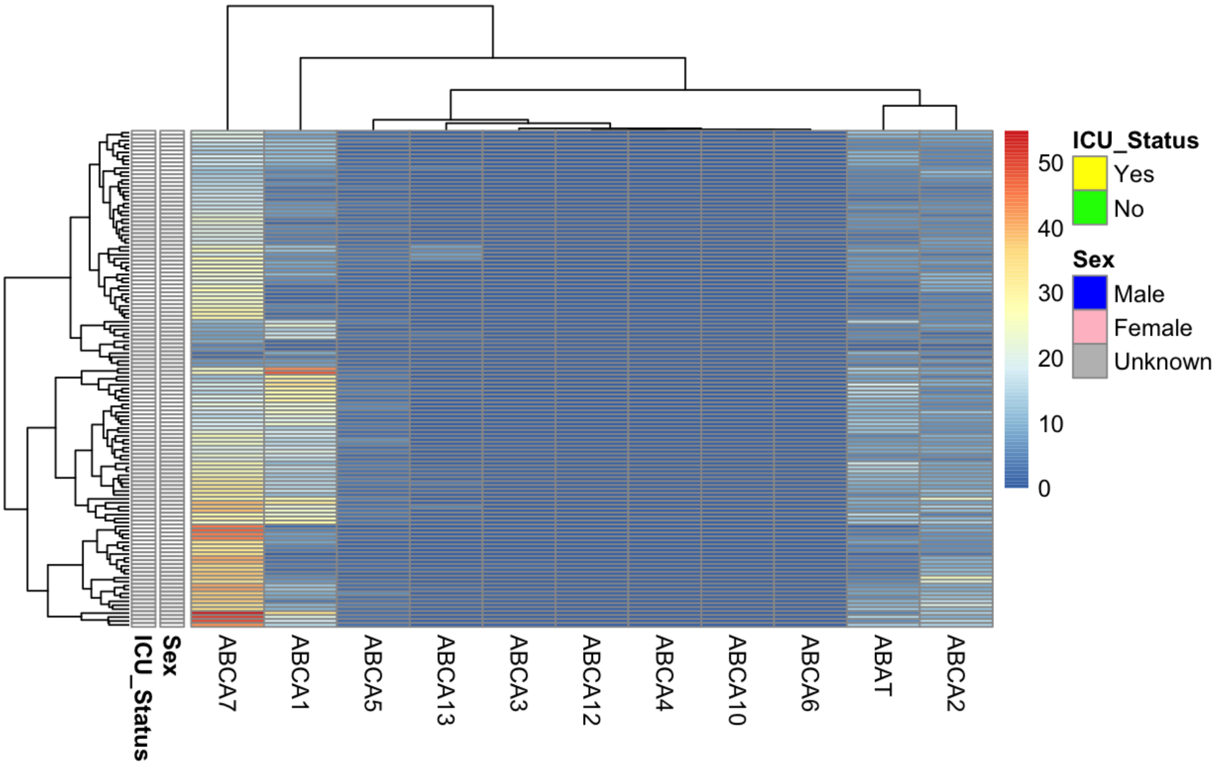

Figure 3: Heatmap of genes ABAT, ABCA1, ABCA10, ABCA12, ABCA13, ABCA2, ABCA3, ABCA4, ABCA5, ABCA6, and ABCA7

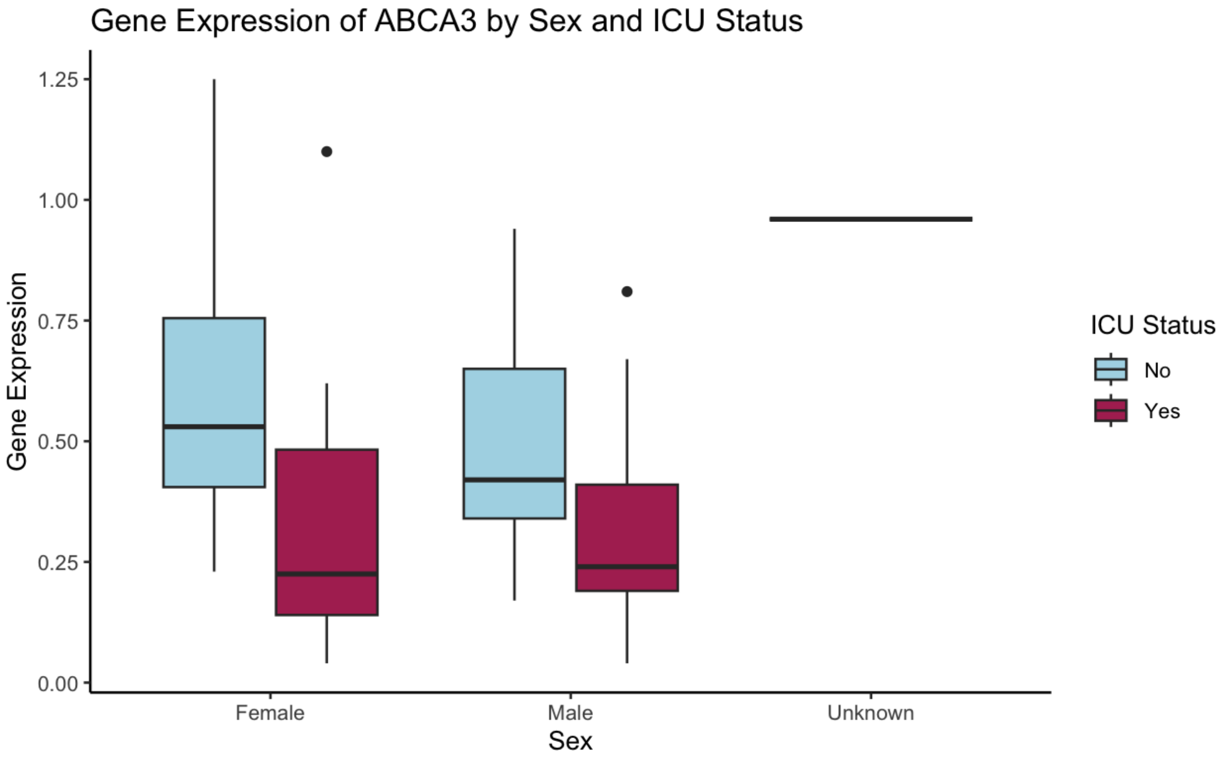

Figure 4: Box plot of Gene by Sex and ICU Status

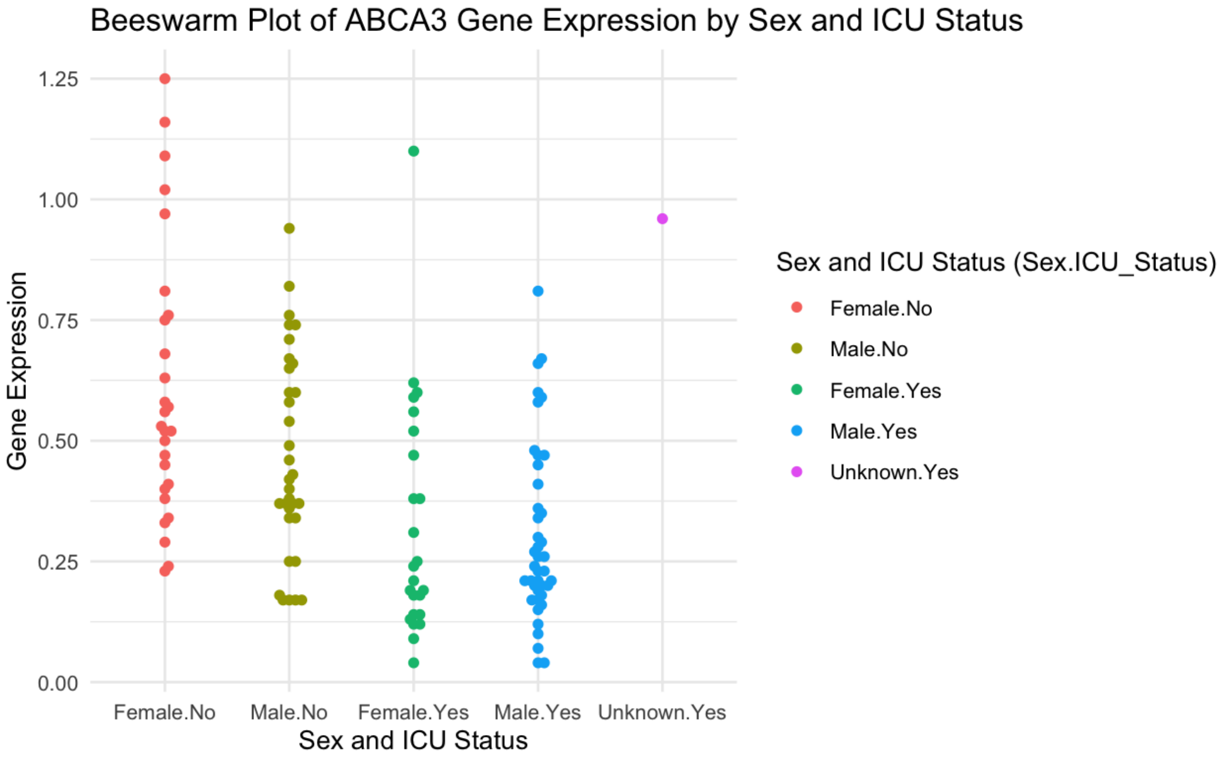

Figure 5: Bee Swarm Plot of Gene by Sex and ICU Status

References

[1] Ron Eklund. Beeswarm function - rdocumentation, 2021. https://www.rdocumentation.org/packages/beeswarm/versions/0.4.0/topics/beeswarm.

[2] Jaroslav A. Hubacek, Tom Philipp, Vera Adamkova, Ondrej Majek, and Ladislav Dusek. Abca3 and lztfl1 polymorphisms and risk of covid-19 in the czech population. *Physiological Research*, 72:539, 2023.

[3] Raivo Kolde. pheatmap: Pretty heatmaps, 2018. R package version 1.0.12.

[4] 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 Systems*, 12:23–40.e7, 1 2021.

[5] Benjamin Rich. table1: Tables of descriptive statistics in html, 2023. R package version 1.4.3.

[6] Thomas Travison, Timothy Tsai, Will Beasley, Yihui Xie, and Guangchuang Yu. Package 'kableextra' type package title construct complex table with 'kable' and pipe syntax. 2024.

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

[8] 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.

[9] Hadley Wickham, Romain François, Lionel Henry, and Kirill Müller. dplyr: A grammar of data manipulation, 2022. https://dplyr.tidyverse.org, https://github.com/tidyverse/dplyr.