QBS 103 Final Project

Ruining Zhou

August 2025

1 Introduction

This project explores how the expression of different genes varies between different characteristics of patients. The data set analyzed in this project is derived from the study "Large-Scale Multi-omic Analysis of COVID-19 Severity" conducted by Overmyer et al. (2020). In this study, the authors performed RNA sequencing and high-resolution mass spectrometry on 128 blood samples collected from patients positive for COVID-19 and negative for COVID-19. They collected different stages and results of the disease. The authors identified 219 molecular characteristics that are very significant for the status and severity of COVID-19. This data set provides a comprehensive investigation of the molecular changes associated with the status and severity of COVID-19.

I chose AATF (Apoptosis Antagonizing Transcription Factor) in my main plot. The apoptosis-antagonizing transcription factor (AATF) participates in transcriptional regulation, cell cycle control, DNA damage reactions, and the implementation of cell death processes. It also interacts directly with nuclear hormone receptors, increasing their transactivation (Sharma, 2013). All of these processes may be highly relevant in COVID-19. The AATF gene plays an important role in regulating cell survival and stress responses. Examining the expression of AATF may provide insights into COVID-19.

This project focused on the different genes and key variables in the sample, including age (continuous), sex (male/female), and clinical status (ICU / non-ICU). To examine how gene expression varies across these variables, this project generated a series of outputs in R, including: (i) a table of summary statistics, (ii) a histogram of gene expression, (iii) a scatter plot of gene expression against the selected continuous covariate, (iv) a box plot of gene expression stratified by the two selected categorical covariates, (v) a heatmap of 10 genes, and (vi) a violin plot. These descriptive analyzes illustrate the relationship between gene expression and key demographic and clinical covariates in the data set.

2 Methods

For this project, the main analysis focused on the relationship between the gene expression and covariates. The analysis tool for this project is RStudio.

The data set analyzed in this project was obtained from the study "Large-Scale Multi-omic Analysis of COVID-19 Severity" (Overmyer et al., 2020). From the dataset, the main plot selected continuous covariate is age and two selected categorical covariates are sex (male/female) and status (ICU/NonICU).These covariates were used to examine variation in gene expression across demographics.

To explore the relationship between gene expression and covariates, this project generated a series of descriptive analyses using R packages including ggplot2 (H, 2016), pheatmap (Kolde, 2025), dplyr (Wickham, 2023), knitr (Yihui, 2015 Xie, 2025), kableExtra (Zhu, 2024), tibble (Müller, 2023), and stringr (Wickham, 2023). Specifically, this project produced: (i) a table of summary statistics separated by ICU status; (ii) a histogram of AATF expression values; (iii) a scatter plot of AATF expression vs. Age; (iv) a box plot of AATF expression separated by Sex and ICU Status; (v) a heatmap of 18 selected genes; (vi) a violin plot for AATF gene expression by Sex and Status These plots were created to show both the distributional features of gene expression and its relationship.

3 Results

3.1 Table of summary statistics

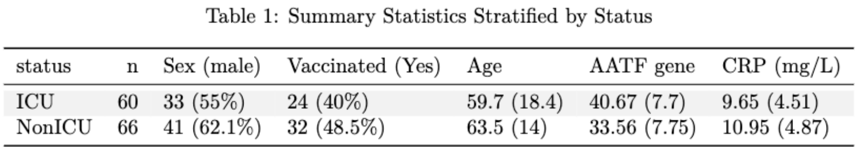

Table 1: Summary Statistics Stratified by Status

status	n	Sex (male)	Vaccinated (Yes)	Age	AATF gene	CRP (mg/L)
ICU	60	33 (55%)	24 (40%)	59.7 (18.4)	40.67 (7.7)	9.65 (4.51)
NonICU	66	41 (62.1%)	32 (48.5%)	63.5 (14)	33.56 (7.75)	10.95 (4.87)

Figure 1: The table of summary statistics separated by ICU Status

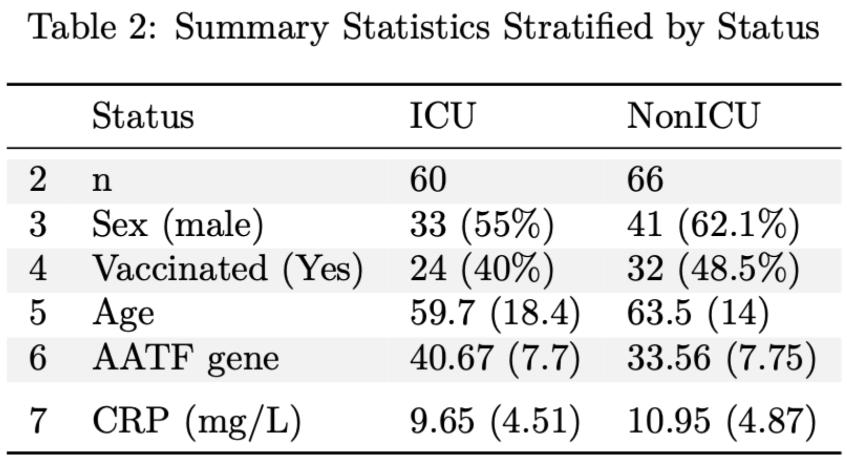

Table 2: Summary Statistics Stratified by Status

	Status	ICU	NonICU
2	n	60	66
3	Sex (male)	33 (55%)	41 (62.1%)
4	Vaccinated (Yes)	24 (40%)	32 (48.5%)
5	Age	59.7 (18.4)	63.5 (14)
6	AATF gene	40.67 (7.7)	33.56 (7.75)
7	CRP (mg/L)	9.65 (4.51)	10.95 (4.87)

Figure 2: The table of summary statistics separated by ICU Status 2

This summary statistics table is stratified by ICU status. It showed that ICU patients were generally younger with a mean age 59.7 versus a mean age 63.5 with NonICU patients. It had a higher average AATF gene expression which is 40.67 versus 33.56 compared to NonICU patients. The CRP levels were slightly lower in the ICU group which is 9.65 versus NonICU group is 10.95.

3.2 Histogram of gene

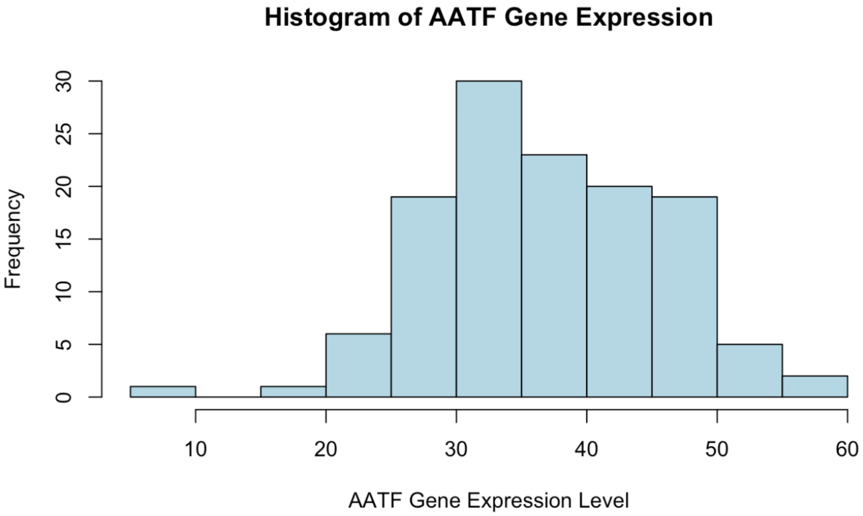

Figure 3: a histogram of AATF expression values

The histogram of AATF gene expression shows that expression values are roughly centered between 30 and 40 with most samples clustering in this range. The distribution is approximately symmetric with a couple lower and higher expression outliers. This suggests the variability is moderate across participants.

3.3 Scatter plot of gene + continuous covariate

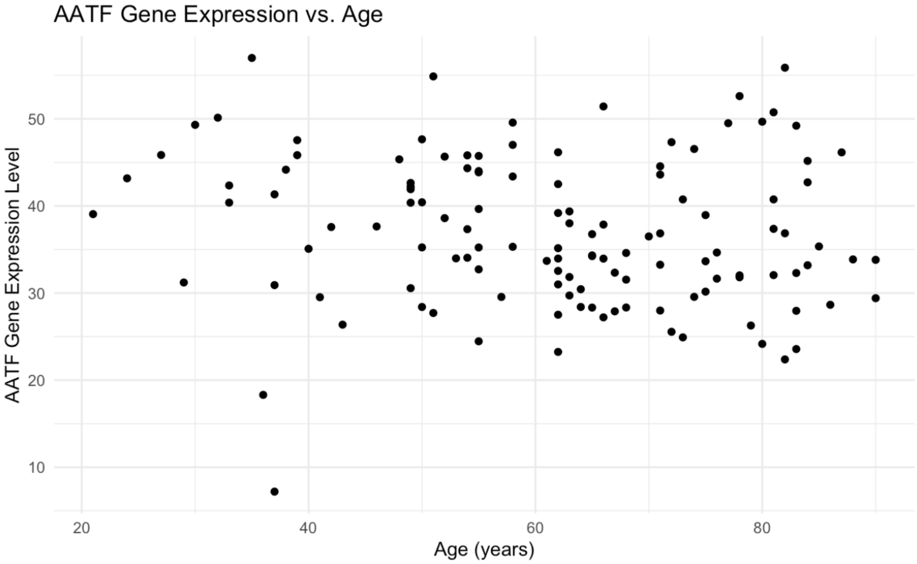

Figure 4: a scatter plot of AATF expression vs. Age

The scatterplot shows the relationship between AATF gene expression levels and age from the data set. There does not appear to be a clear linear trend in which the points are widely scattered across age groups. This suggests that there may not be a strong relationship between AATF gene expression levels and age.

3.4 Boxplot of gene stratified by 2 categorical covariates

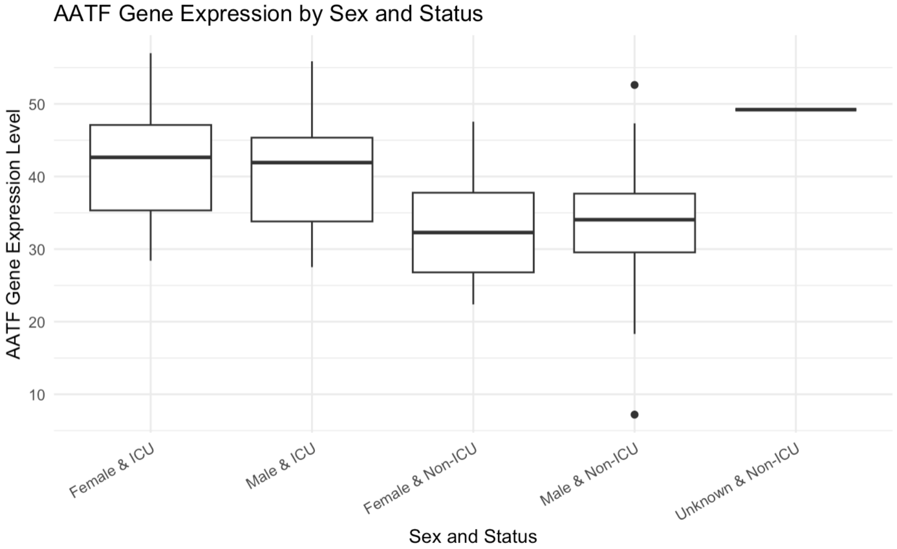

Figure 5: The box plot of AATF expression separated by Sex and ICU Status

The box plot shows the relationship between AATF expression separated by Sex and ICU Status. The trend might indicate that AATF is increased in ICU patients. The sex differences are modest which females in ICU having a slightly higher median than males.

3.5 Heatmap

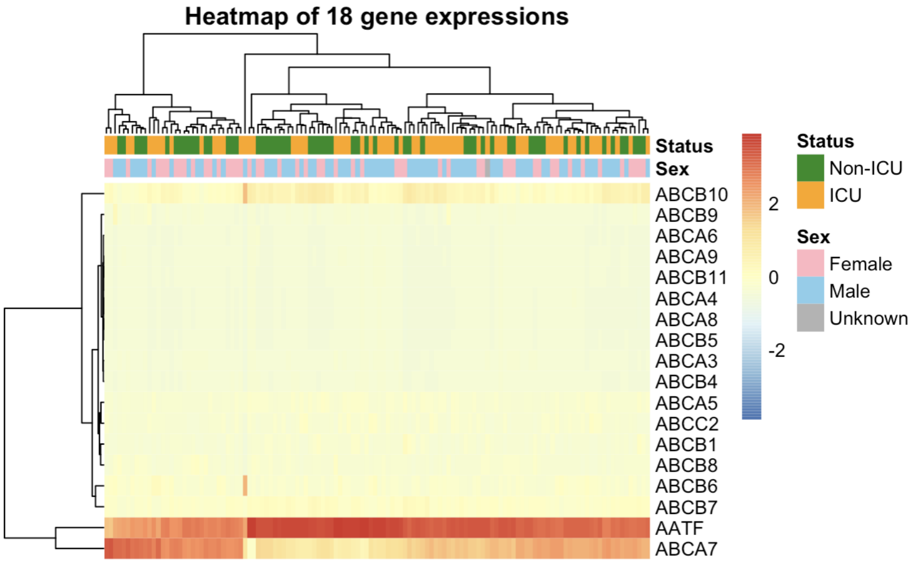

Figure 6: The heatmap of 18 selected genes

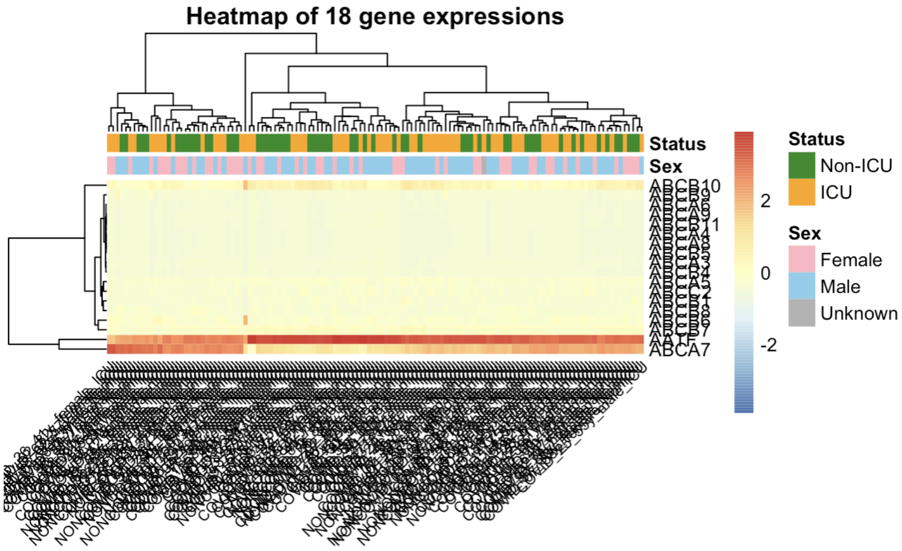

Figure 7: The heatmap with column name

The heatmap includes 18 selected gene expressions that show clustering of patients by ICU status and sex. In this row scaled heatmap, AATF gene is the dominant signal which is consistently warmer across most samples. The ABCA7 gene is near baseline with small variance but no clear group pattern which implying that it contributes little to class differentiation. Overall, the panel provides high within-gene relative elevation for AATF gene as well as flat behavior for ABCA7 gene.

3.6 Violin plot

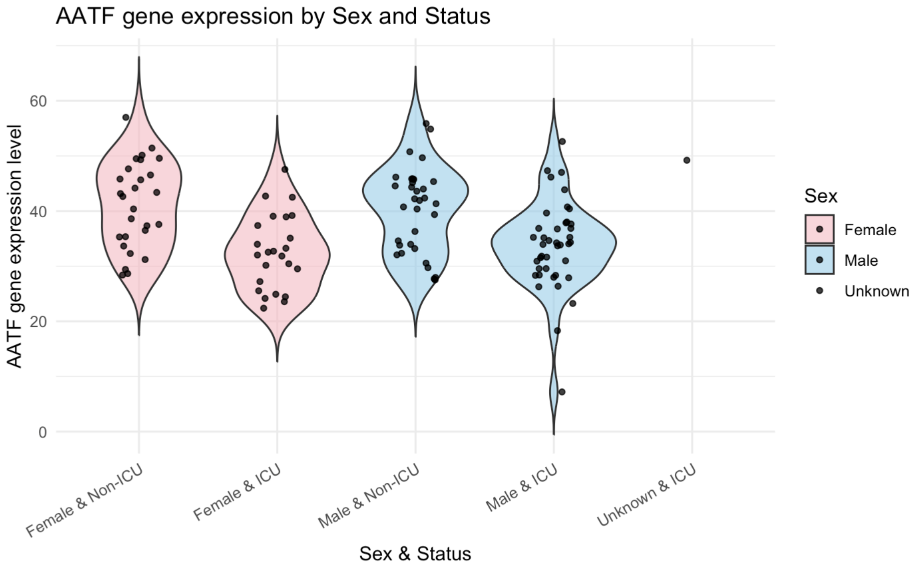

Figure 8: a violin plot for AATF gene expression by Sex and Status

The violin plot of AATF gene expression by sex and ICU status shows higher average expression in ICU patients compared to Non-ICU patients across both males and females. This result is similar to the box plot. The distributions also suggest that females tend to have slightly higher expression levels than males with both ICU group and Non-ICU group.

4 References

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

Kolde R (2025). $_{p}heatmap : PrettyHeatmaps_{.}Rpackageversion1.0.13, < https : //CRAN.R - project.org/package = pheatmap > .$

Müller K, Wickham H (2023). ${}_{t}ibble : SimpleDataFrames_{.}Rpackageversion3.2.1, < https : //CRAN.R - project.org/package = tibble >$.

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., … Jaitovich, A. (2021). Large-Scale Multi-omic Analysis of COVID-19 Severity. Cell systems, 12(1), 23–40.e7. https://doi.org/10.1016/j.cels.2020.10.003

Sharma, M. Apoptosis-antagonizing transcription factor (AATF) gene silencing: role in induction of apoptosis and down-regulation of estrogen receptor in breast cancer cells. Biotechnol Lett 35, 1561–1570 (2013). https://doi.org/10.1007/s10529-013-1257-8

Wickham H, François R, Henry L, Müller K, Vaughan D (2023). ${}_{d}plyr : AGrammarofDataManipulation_{.}Rpackageversion1.1.4, < https : //CRAN.R- project.org/package = dplyr >$.

Wickham H (2023). ${}_{s}tringr : Simple, ConsistentWrappersforCommonStringOperations_{.}Rpackageversi$
$https : //CRAN.R - project.org/package = stringr >$.

Xie Y (2025). ${}_{k}nitr : AGeneral-PurposePackageforDynamicReportGenerationinR_{.}Rpackageversion1.5$
$https : //yihui.org/knitr/ >$.

Yihui Xie (2015) Dynamic Documents with R and knitr. 2nd edition. Chapman and Hall/CRC. ISBN 978-1498716963

Yihui Xie (2014) knitr: A Comprehensive Tool for Reproducible Research in R. In Victoria Stodden, Friedrich Leisch and Roger D. Peng, editors, Implementing Reproducible Computational Research. Chapman and Hall/CRC. ISBN 978-1466561595

Zhu H (2024). ${}_{k}ableExtra : ConstructComplexTablewith'kable'andPipeSyntax_{.}Rpackageversion1.4.0, <$
$https : //CRAN.R - project.org/package = kableExtra >$.