Final Project

Xinqiao Li

August 2025

1 Introduction

The dataset used in this study originates from the large-scale multi-omic analysis of COVID-19 severity published in Cell Systems (2021) . It contains comprehensive clinical and demographic information of COVID-19 patients, including ICU status, use of mechanical ventilation, age, sex, and multiple inflammatory and coagulation biomarkers such as ferritin, C-reactive protein, and D-dimer. These variables are closely associated with disease severity and provide an important foundation for exploring molecular mechanisms.

In this study, I selected AASS as the primary gene of interest. AASS is a key enzyme in the lysine degradation pathway, catalyzing the conversion of lysine into α-aminoadipate semialdehyde. Lysine metabolism is not only linked to energy homeostasis but also plays potential roles in immune regulation and inflammatory responses. Previous studies have suggested that disturbances in amino acid metabolism are strongly associated with viral infections and dysregulation of host immune responses. Therefore, the expression of AASS may have important implications in the pathophysiology of COVID-19.

To systematically evaluate the relationship between AASS expression and clinical features, this study focuses on the following covariates: age, sex, icu status, use of mechanical ventilation, and several key biomarker levels. By integrating clinical and molecular information, this analysis aims to provide a more comprehensive understanding of the potential role of AASS in COVID-19 severity.

2 Methods

The data analyzed in this study was obtained from the COVID-19 multi-omic dataset described in Cell Systems (2021). This dataset includes gene expression data along with detailed clinical information such as demographic characteristics, ICU status, and mechanical ventilation use.

The data analyzed in this study was obtained from the COVID-19 multi-omic dataset described by [1]. This dataset includes gene expression profiles together with detailed clinical information such as demographic characteristics, ICU admission status, and use of mechanical ventilation.

All analyses were performed in **R version 4.3.1** [2]. The following R packages were used for data wrangling, visualization, and statistical analysis:

- `tidyverse`, which includes `dplyr` [3] and `purrr` [4], for data manipulation and preprocessing;
- `ggplot2` [5] for creating plots such as histograms, scatter plots, and boxplots;
- `pheatmap` [6] for clustered heatmap generation;

For unsupervised clustering in the heatmap analysis, hierarchical clustering was applied using Euclidean distance [7] as the distance metric and complete linkage as the clustering algorithm, consistent with methods discussed in class.

All plots were generated in R and exported in formats suitable for integration into this LaTeX report.

3 Results

3.1 Table of Summary Statistics

Table 1: Summary Statistics Stratified by ICU Status

Variable	Level	ICU = No	ICU = Yes
Continuous variables: Mean (SD)			
Ventilator-free days	—	25.95 (7.22)	15.17 (11.93)
Charlson score	—	3.11 (2.45)	3.82 (2.50)
Age	—	58.67 (17.82)	63.45 (14.00)
Categorical variables: n (%)			
Sex	Male	26 (45.6%)	24 (36.4%)
	Female	31 (54.4%)	42 (63.6%)
Mechanical ventilation	No	52 (91.2%)	20 (30.3%)
	Yes	5 (8.8%)	46 (69.7%)

Table Summary statistics stratified by ICU status presents the baseline characteristics of the study population stratified by ICU status.

For continuous covariates, patients admitted to the ICU had fewer ventilator-free days on average compared to non-ICU patients , and they also tended to be older (mean age 63.45 vs. 58.67 years). The Charlson comorbidity score was slightly higher among ICU patients (3.82 vs. 3.11), suggesting a greater burden of comorbidities in this group.

For categorical covariates, the distribution of sex varied across ICU groups. Among ICU patients, 36.4% were male and 63.6% were female, whereas in the non-ICU group, 45.6% were male and 54.4% were female. Use of mechanical

ventilation also differed substantially: 69.7% of ICU patients required mechanical ventilation compared with only 8.8% of non-ICU patients.

Overall, ICU patients were generally older, had fewer ventilator-free days, slightly higher comorbidity scores, and were more likely to receive mechanical ventilation compared to patients not admitted to the ICU.

3.2 Histogram of Gene

The histogram of AASS expression provides an overview of its distribution across participants. As shown in Figure 1, most participants have expression values clustered between 0 and 0.4, while a smaller subset shows markedly higher expression levels extending beyond 0.6. This long right tail suggests the presence of outliers or a subgroup of individuals with elevated expression. Such skewness is important to recognize, as it indicates that standard parametric assumptions (e.g., normality) may not hold without transformation or appropriate statistical adjustments.

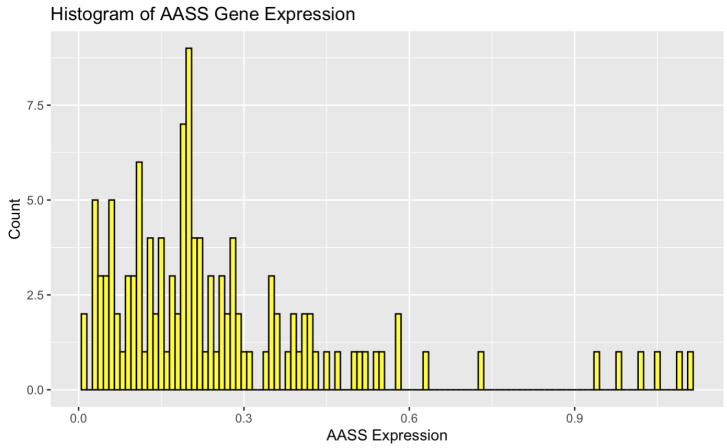

Figure 1: Histogram of AASS Gene Expression

3.3 Scatter Plot of Gene and Continuous Covariate

The scatter plot examines the relationship between AASS expression and age. As seen in Figure 2, points are widely scattered without a clear linear trend, so age does not strongly influence AASS expression levels in this dataset. However, there is substantial variability in expression at all ages, and a few older participants show high expression values. This observation implies that other biological or clinical factors, rather than chronological age alone, may better explain heterogeneity in gene expression.

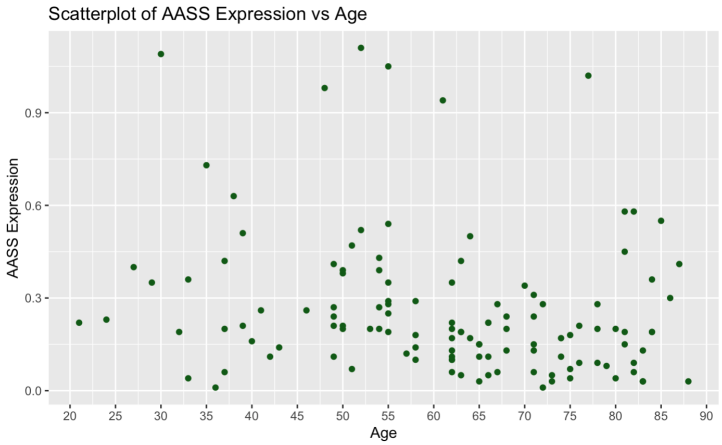

Figure 2: Scatterplot of AASS Expression VS Age

3.4 Boxplot of Gene Stratified by Two Categorical Covariates

Boxplots stratified by sex and ICU status allow direct comparison of subgroup distributions (Figure 3). Across both sexes, ICU participants tended to have lower median AASS expression compared with non-ICU participants. The interquartile ranges for ICU groups were narrower, so less variability within this subgroup. Additionally, several outliers are visible, particularly among non-ICU participants, so individuals with unusually high expression. These findings suggest that ICU status, instead of sex, is a stronger determinant of AASS expression differences.

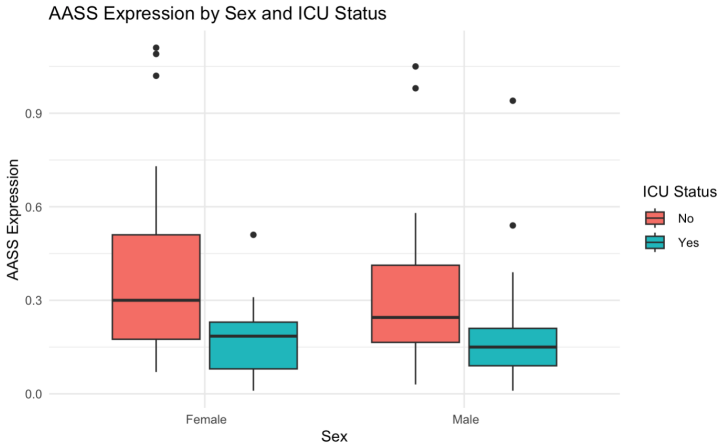

Figure 3: Boxplots of AASS Expression Stratified by Sex and ICU Status

3.5 Heatmap

The heatmap in Figure 4 visualizes expression patterns for the first 20 samples across several genes, including AASS. The clustering shows distinct groups of genes, inclding ABCA family showing correlated expression. Sample-level annotations for sex and ICU status reveal block structures, where groups of patients share similar expression profiles. Such clustering may reflect underlying biological processes, disease severity, or treatment effects. This figure demonstrates the value of multivariate visualization for detecting coordinated gene activity and subgroup heterogeneity that may not be evident from single-variable plots.

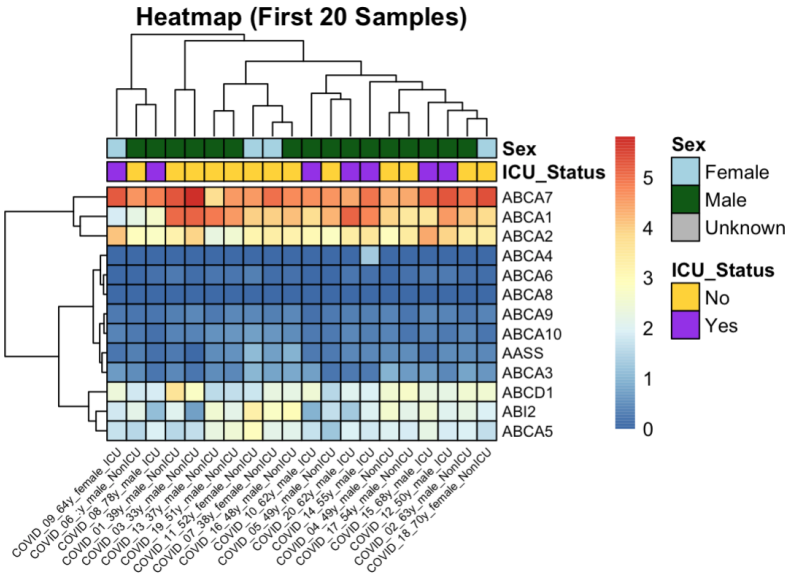

Figure 4: Heatmap of Selected Genes across 20 Samples

3.6 My own plot– Densityplot

The distribution of AASS expression was right-skewed, with many participants having low values and a small number having very high values.(Figure 5). This figure shows the density of AASS expression across all participants. The shaded area indicates where the data points are more dense. Most participants have low to moderate expression, while a small group has much higher expression. This plot is useful because it gives a clear picture of the overall pattern and highlights outliers that may be important for further study.

This shape suggests that most people show similar expression, but a few stand out. Looking at the overall distribution helps us see patterns that a single

value cannot show. It is useful because it summarizes the data and shows both common levels and outliers.

Figure 5: Distribution Of AASS Gene Expression (Area Plot).

References

[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, 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. 12(1):23–40.e7.

[2] R Core Team. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria, 2023.

[3] Hadley Wickham, Romain François, Lionel Henry, and Kirill Müller. *dplyr: A Grammar of Data Manipulation*, 2023. R package version 1.1.2.

[4] Hadley Wickham and Lionel Henry. *purrr: Functional Programming Tools*, 2023. R package version 1.0.2.

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

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

[7] Brian S. Everitt, Sabine Landau, Morven Leese, and Daniel Stahl. *Cluster Analysis*. Wiley, 5th edition, 2011.