QBS 103 Final Submission

Haoyang Cui

Aug. 29th, 2 p.m.

1 Data Select

We pick up the data from arge-Scale Multi-omic Analysis of COVID-19 Severit, Overmyer et al. 2021
For the data we use, we delete the participant _id NONCOVID_15_83y_unknown_ICU as we can not access its sex data, so we give up this one. And there is a wrong participant_id, I use setdiff function to find it and make the two id same. We use the tidyverse package, Wickham and RStudio 2023 to edit the data.

```r
genes <- read.csv("QBS103_GSE157103_genes.csv")
matrix <- read.csv("QBS103_GSE157103_series_matrix.csv")

id_genes <- colnames(genes)
id_genes <- id_genes[-1]
id_matrix <- matrix$participant_id
diff1 <- setdiff(id_genes, id_matrix)
diff2 <- setdiff(id_matrix, id_genes)
colnames(genes)[colnames(genes) == diff1] <- diff2
matrix <- matrix %>% rename(’procalcitonin.ng.ml.’ = ’procalcitonin.ng.ml..’)

unknown_id <- matrix$’participant_id’[matrix$sex == ’ unknown’]

matrix <- matrix[!matrix$participant_id %in% unknown_id, ]
genes[[unknown_id]] <- NULL
```

2 Table

2.1 Variable Choose

For the table, we generate 3 category variable: sex, disease_status and icu_status, and 3 continuous variable: age, ferritin and procalcitonin.

2.2 Table

From the table, Male patients were slightly higher than female patients in terms of COVID-19 infection rate, proportion of admission to ICU, and ferritin levels, while female patients had significantly higher procalcitonin levels than male patients.

Comparison of COVID-19 and non-COVID-19 patients

	Male	**Female**	**Overall**
	(N=74)	**(N=51)**	**(N=125)**
Category			
Disease Status			
COVID-19	62 (83.8%)	38 (74.5%)	100 (80%)
non-COVID-19	12 (16.2%)	13 (25.5%)	25 (20%)
Icu Status			
Yes	41(55.4%)	24 (47.1%)	65(52%)
No	33(44.6%)	27(52.9%)	60(48.0%)
Continuous			
Age			
Mean (SD)	62.3 (14.4)	59.3 (18.0)	61.1 (15.9)
Median [Min, Max]	63 [27, 88]	62 [21, 86]	62.0 [21.0, 88.0]
ferritin.ng.ml.			
Mean (SD)	993.3 (1013.0)	619.3 (1054.3)	833.5 (1042.8)
Median [Min, Max]	755 [58, 5971]	318 [14, 5508]	573 [14, 5971]
procalcitonin.ng.ml.			
Mean (SD)	2.47 (5.79)	3.94 (13.65)	3.08 (9.79)
Median [Min, Max]	0.69 [0.05, 36.00]	0.500 [0.05, 86.39]	0.5 [0.05, 86.39]

3 New plot type

To introduce more about the category relation, I use the faceting barplot named facet_grid in ggplot2 (Wickham et al. 2024), and add count number on it by using geom_text.

```r
selected_table %>%
  ggplot(aes(x = sex, fill = disease_status)) +
  geom_bar() +
  facet_grid(icu_status ~ .) +
  geom_text(aes(label = ..count..), stat = "count", position = position_stack(vjust = 0.5)) +
  scale_fill_manual(values = c("disease state: COVID-19" = "#f69F00", "disease state: non-COVID-19" = "#5654E9")) +
  labs(x = "sex", fill = "disease_status", title = "Occupation by ICU Status, Sex, and Disease Status") +
  theme_minimal()
```

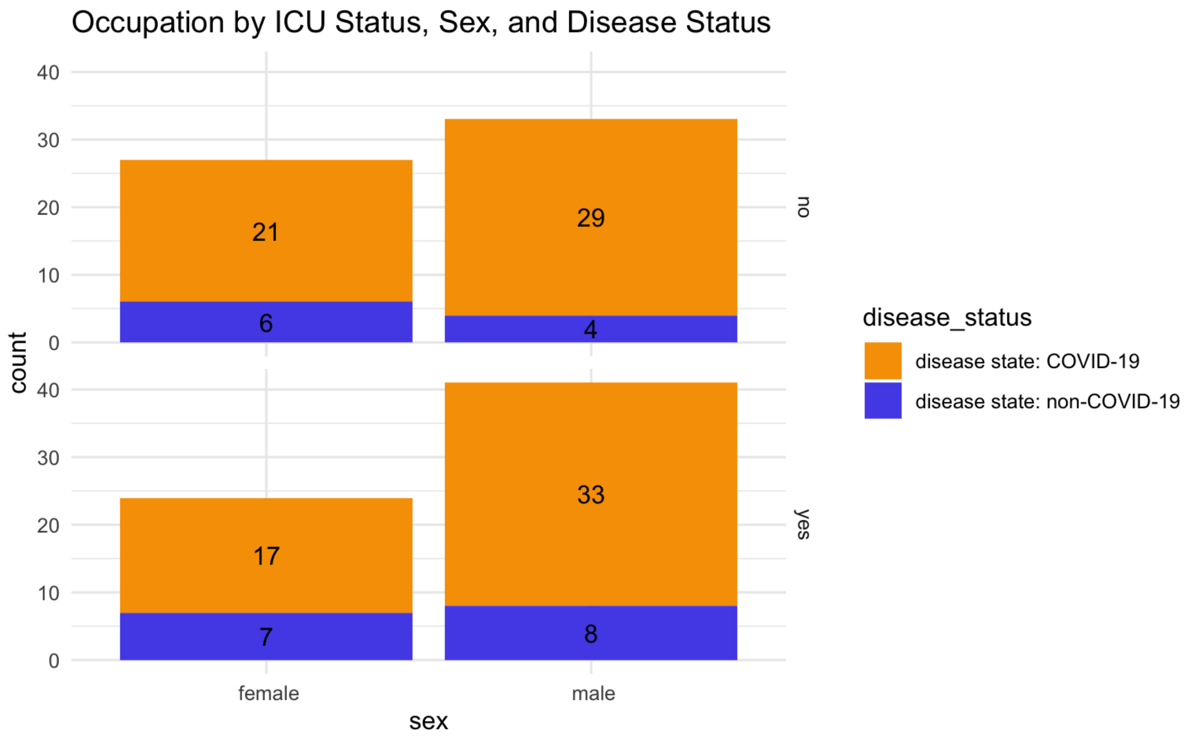

Figure 1: Bar plot with faceting

From figure 1, the COVID-19 infection rate in male patients was slightly higher than that in female patients, regardless of whether they were admitted to the ICU. And among all patients, the number of people with Covid-19 was much higher than that of non-COVID patients.

4 Histogram, Scatter plot, and Boxplot from submission 1

We still use the ggplot2 (Wickham et al. 2024) to observe the gene AAMP.

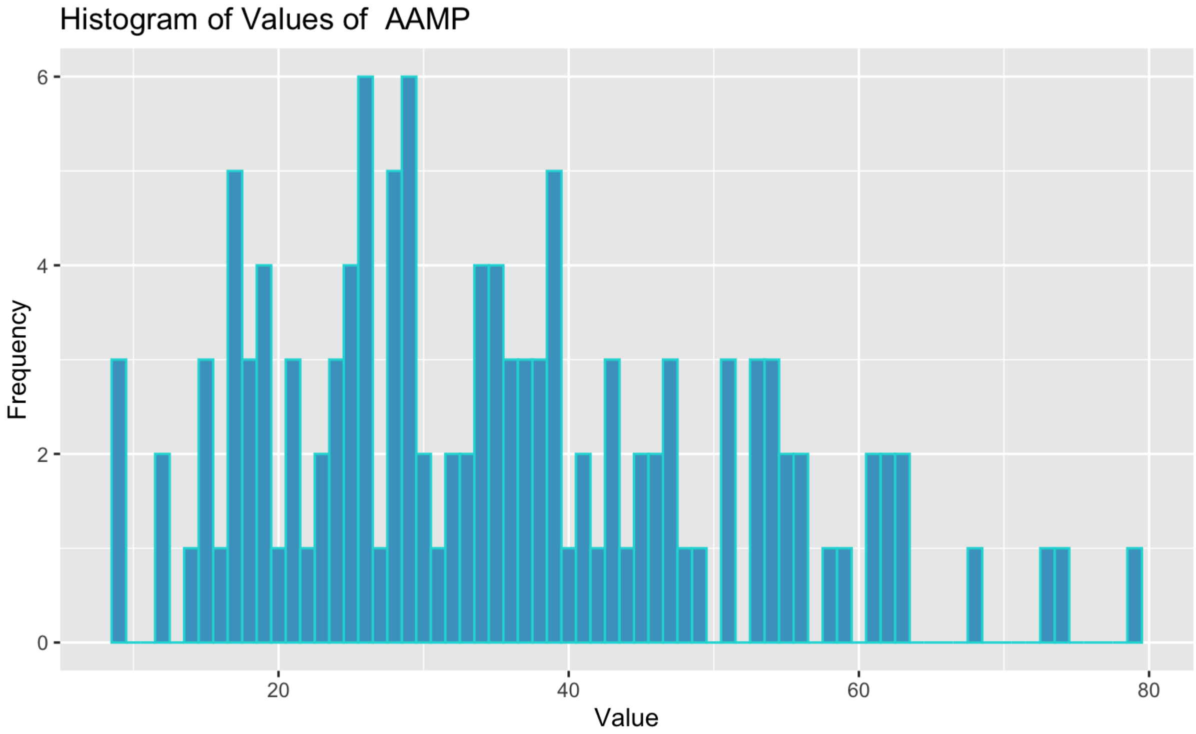

Figure 2: Histogram for AAMP

From figure 2, the gene value of AAMP has a relatively uniform distribution in different intervals, with the highest frequency appearing between about 20 and 40. The gene values in these intervals appear more frequently, with 5 to 6 occurrences in each interval. It shows a relatively dispersed distribution pattern with no obvious central trend.

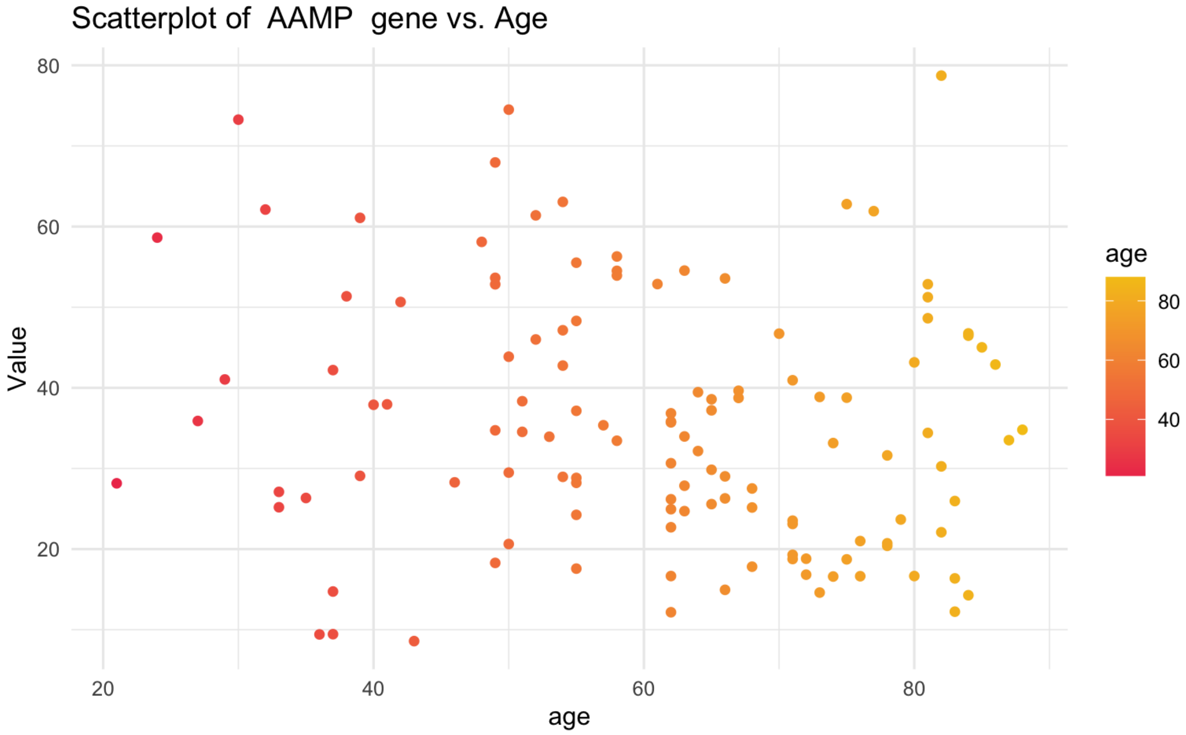

Figure 3: Scatterplot for AAMP

From figure 3, AAMP gene values were distributed across age groups, but there was no obvious linear trend, indicating that there may not be a direct correlation between AAMP values and age.
However, we can observe that there is a certain clustering of AAMP gene values in the middle-aged group, especially in areas with lower gene values.

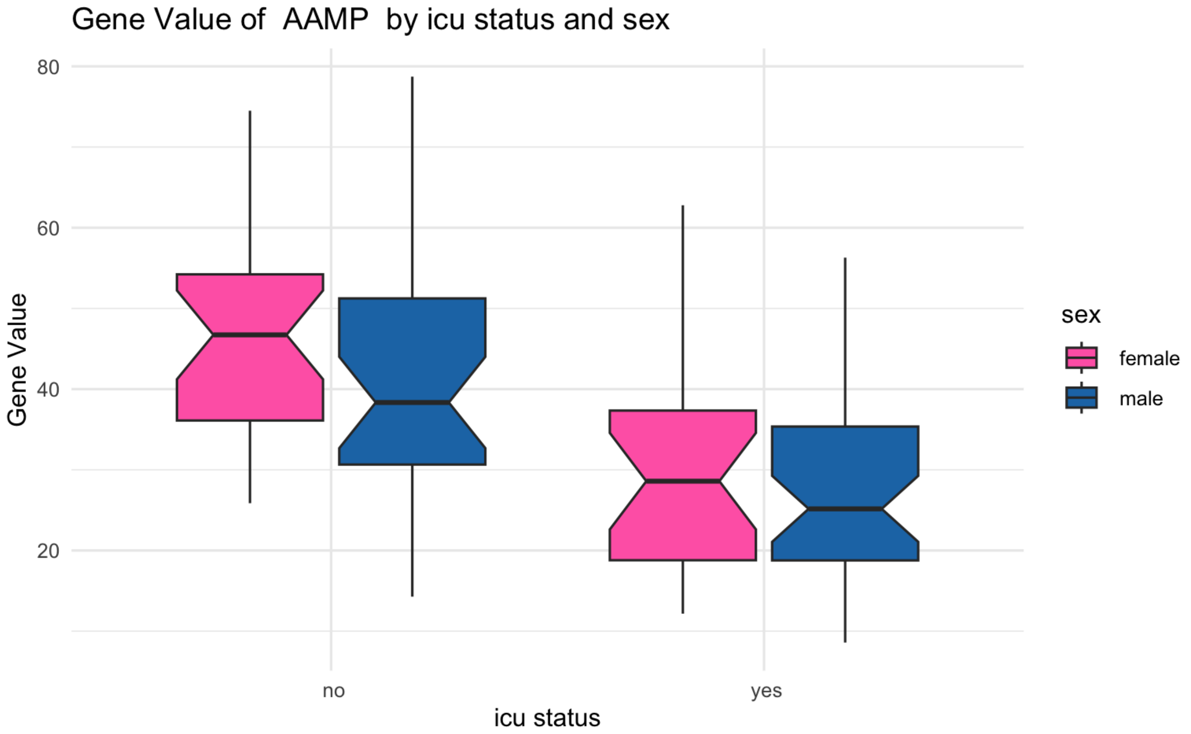

Figure 4: Boxplot for AAMP

From figure 4, there was no significant difference in the gene values between the sexes. The median gene values for patients admitted to the ICU tended to be slightly lower.

5 Heatmap

To build the heatmap, we use the package pheatmap (Kolde 2019), labeled the icu_status and sex as annotation.

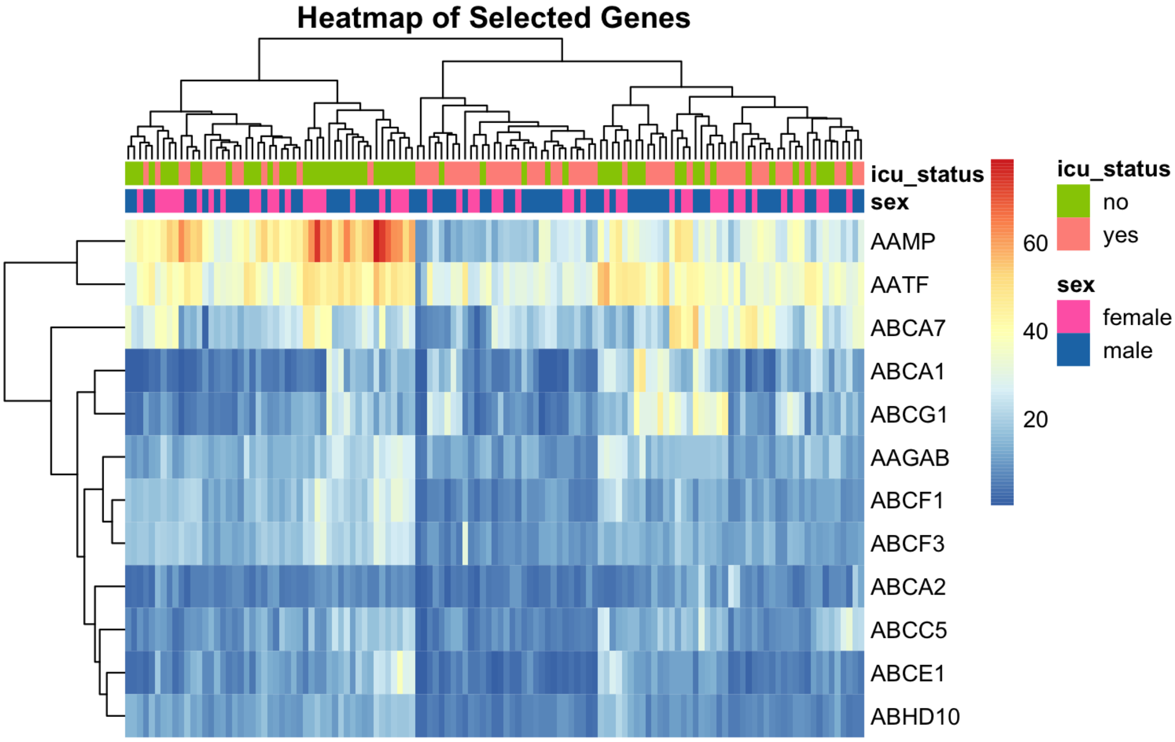

Figure 5: Heatmap

The results of figure 5 show that gene expression varies greatly among different individuals, especially in gene regions with high gene values.

References

Kolde, Raivo (2019). *pheatmap: Pretty Heatmaps*. R package version 1.0.12. Comprehensive R Archive Network (CRAN). DOI: 10.32614/CRAN.package.pheatmap. URL: https://CRAN.R-project.org/package=pheatmap.

Overmyer, Katherine A. et al. (2021). "Large-Scale Multi-omic Analysis of COVID-19 Severity". In: *Cell Systems* 12 (1). ISSN: 24054720. DOI: 10.1016/j.cels.2020.10.003.

Wickham, Hadley and RStudio (2023). *tidyverse: Easily Install and Load the 'Tidyverse'*. R package version 2.0.0. Comprehensive R Archive Network (CRAN). DOI: 10.32614/CRAN.package.tidyverse. URL: https://tidyverse.tidyverse.org.

Wickham, Hadley et al. (2024). *ggplot2: Create Elegant Data Visualisations Using the Grammar of Graphics*. R package version 3.5.1. Comprehensive R Archive Network (CRAN). DOI: 10.32614/CRAN.package.ggplot2. URL: https://ggplot2.tidyverse.org.