Final

2024-08-26

Introduction

The data set used in this analysis includes gene expression profiles from patients with varying disease statuses, including COVID-19 and non-COVID-19 conditions. The data set consists of multiple gene expression measurements across different participants. For the primary analysis, the gene of interest is A1BG, A1BG is a gene encoding a plasma glycoprotein that has been implicated in various biological processes, including immune response modulation.

Methods

All analyses were conducted using R version 4.3.0. The following R packages were utilized:

- tidyverse (v2.0.0): For data manipulation and visualization.
- knitr (v1.46): For report generation and integrating R code within LaTeX.
- pheatmap (v1.0.12): For creating heatmaps with clustering of rows and columns.

For the heatmap visualization, hierarchical clustering was employed to identify patterns in gene expression across participants. The clustering was performed on both rows (genes) and columns (participants) using Euclidean distance as the metric and the complete linkage method. All data are from Large-Scale Multi-omic Analysis of COVID-19 Severity (https://pubmed.ncbi.nlm.nih.gov/33096026/).

Results

```
library(tidyverse)
library(knitr)
library(pheatmap)
setwd("/Users/zhangbingquan/Desktop/Fdn of data science/7.1-8.1")
genes <- read.csv("QBS103_GSE157103_genes.csv")
series <- read.csv("QBS103_GSE157103_series_matrix.csv",
                   na.strings = c("", "unknown"),
                   strip.white = T, stringsAsFactors = T)
genes_long <- genes %>% filter(X == 'A1BG') %>%
  gather(key = 'participant_id', value = 'expression', -X)
data <- inner_join(genes_long, series, by="participant_id")
data %>%
  filter(!is.na(sex)) %>%
  group_by(disease_status, sex, icu_status) %>%
  summarise(n=n(),
            fibrinogen.mean=mean(fibrinogen, na.rm=T),
```

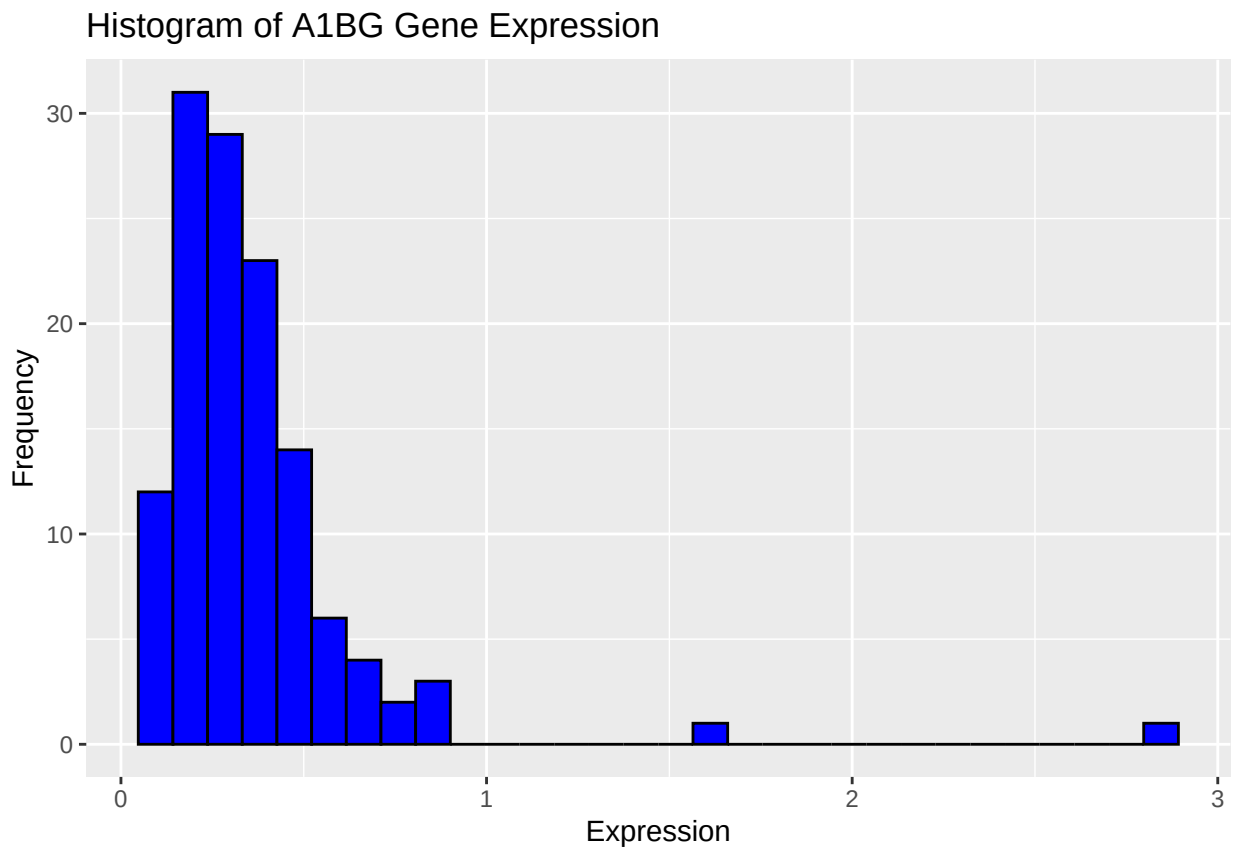
```
            fibrinogen.sd=sd(fibrinogen, na.rm=T),
            crp.mean=mean(crp.mg.l., na.rm=T),
            crp.sd=sd(crp.mg.l., na.rm=T),
            ) %>%
  kable(digits = 3, caption="Summary statistics")
```

Table 1: Summary statistics

disease_status	sex	icu_status	n	fibrinogen.mean	fibrinogen.sd	crp.mean	crp.sd
disease state: COVID-19	female	no	21	490.333	161.092	86.420	73.151
disease state: COVID-19	female	yes	17	505.143	147.622	168.644	121.219
disease state: COVID-19	male	no	28	613.167	200.404	149.759	106.715
disease state: COVID-19	male	yes	33	539.710	223.174	153.909	102.595
disease state: non-COVID-19	female	no	6	363.250	182.546	43.250	42.959
disease state: non-COVID-19	female	yes	7	402.833	184.493	95.920	80.220
disease state: non-COVID-19	male	no	4	317.500	84.146	17.700	11.879
disease state: non-COVID-19	male	yes	8	283.500	23.335	104.575	87.660

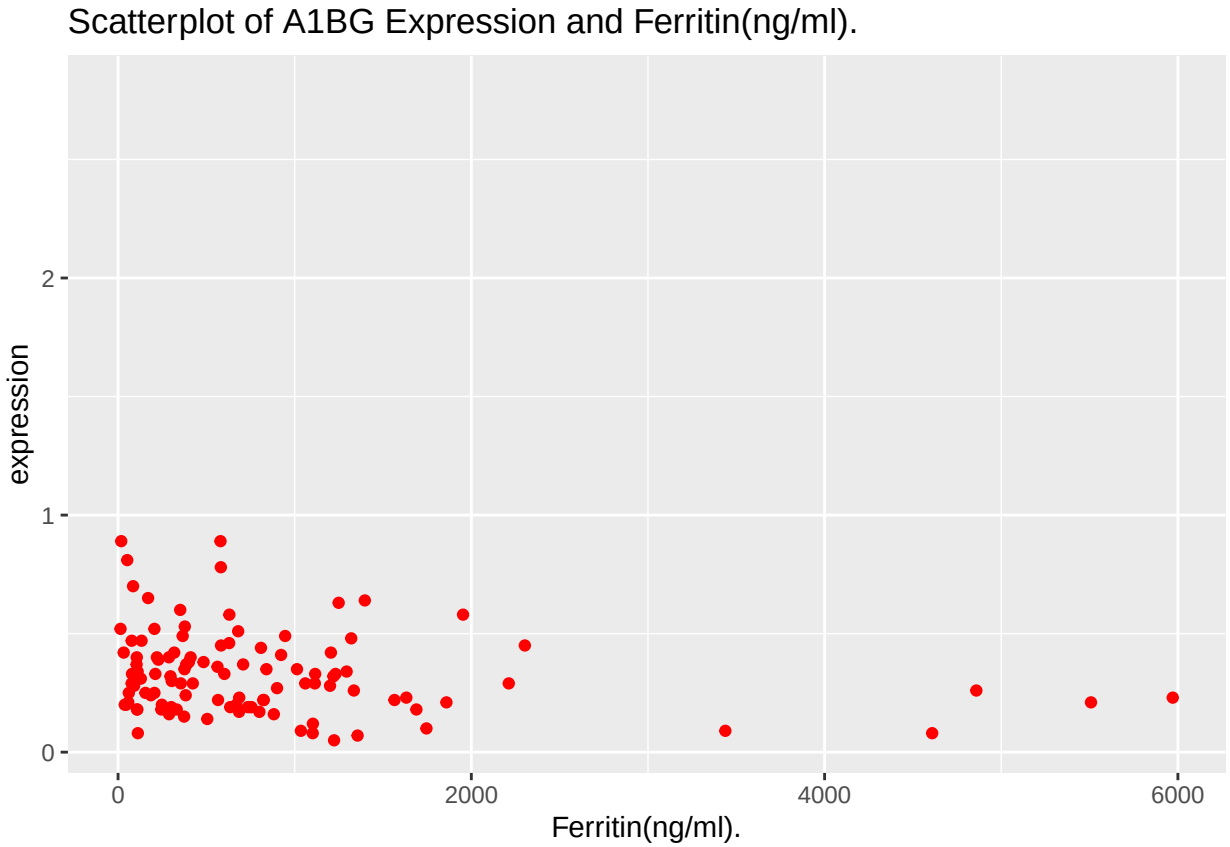
This table shows the mean and standard deviation of fibrinogen and Crp levels in COVID-19 and non-COVID-19 patients in different genders and ICU status. The data show that the overall fibrinogen and Crp levels of COVID-19 patients are significantly higher than those of non-COVID-19 patients. The mean Crp of male COVID-19 patients admitted to the ICU is 539.710, which is significantly higher than the mean of 283.500 for male non-COVID-19 patients admitted to the ICU. For Crp, the mean of female COVID-19 patients admitted to the ICU is 168.644, while that of female non-COVID-19 patients admitted to the ICU is 95.920, with a significant difference. In general, COVID-19 patients, especially those admitted to the ICU, have significantly higher inflammation and coagulation-related indicators than non-COVID-19 patients, regardless of gender.

```
ggplot(genes_long, aes(x = expression)) +
  geom_histogram(fill = "blue", color = "black", bins=30) +
  labs(title = "Histogram of A1BG Gene Expression", x = "Expression", y = "Frequency")
```



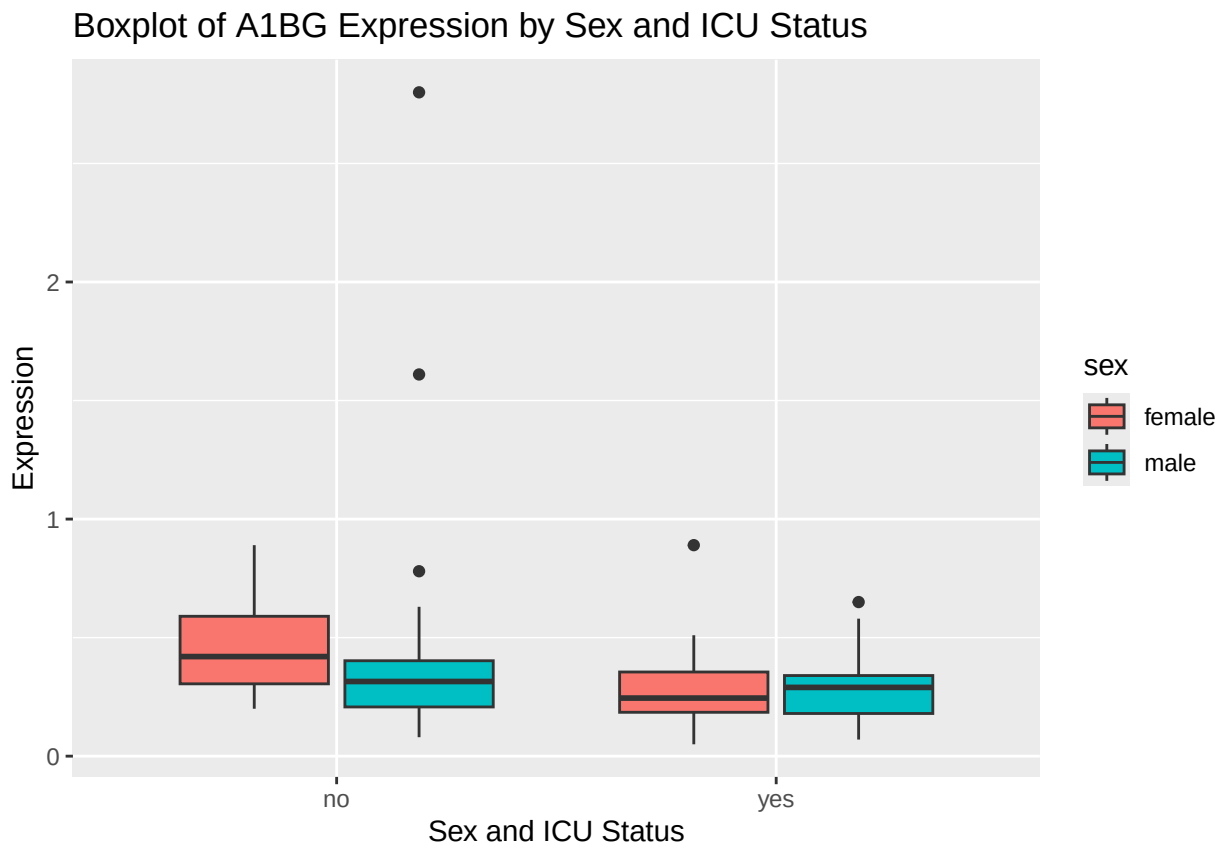
This histogram shows the distribution of A1BG gene expression levels. The horizontal axis represents the gene expression value, ranging from 0 to 3, and the vertical axis represents the corresponding frequency. Most gene expression values are concentrated between 0 and 1, especially in the area close to 0, where the frequency is the highest, indicating that the A1BG gene expression level of most samples is low, and the overall distribution is right-skewed.

```
ggplot(data, aes(x = ferritin.ng.ml., y = expression)) + geom_point(color = "red") +
labs(title = "Scatterplot of A1BG Expression and Ferritin(ng/ml).",
     x = "Ferritin(ng/ml).", y='expression')
```



The scatter plot suggests that the expression and ferritin are weakly negatively associated.Most of the data points are concentrated in the area where the ferritin concentration is less than 2000 ng/ml and the gene expression value is less than 1. There is no obvious upward trend in the gene expression value with the increase of ferritin concentration.

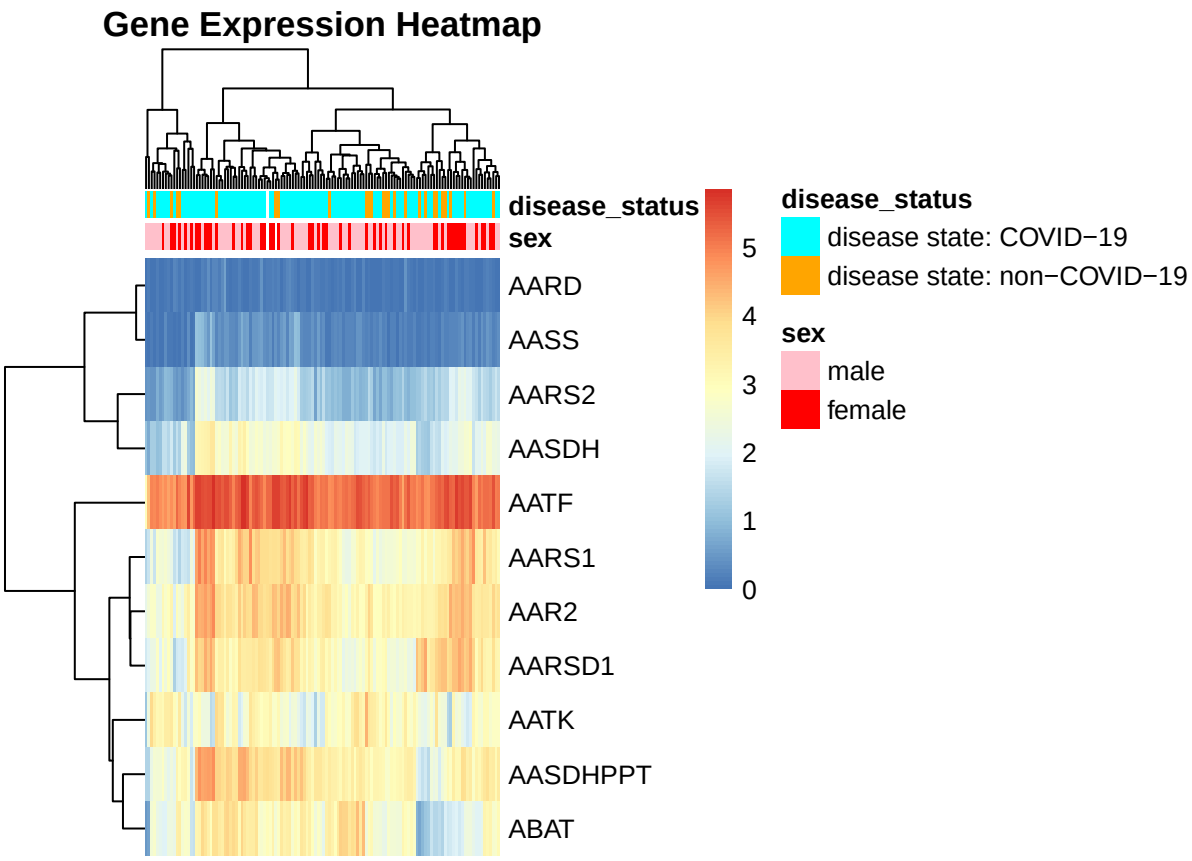
```
data %>% filter(!is.na(sex)) %>%
  ggplot(aes(x=icu_status, y=expression, fill=sex)) + geom_boxplot() +
  labs(title = "Boxplot of A1BG Expression by Sex and ICU Status",
           x = "Sex and ICU Status",
           y = "Expression")
```



This box plot shows the A1BG gene expression levels of different genders depending on whether they were admitted to the ICU. It suggests that the female has a higher average expression compared with male, and the expression with no icu status is higher compared with patients with icu status.

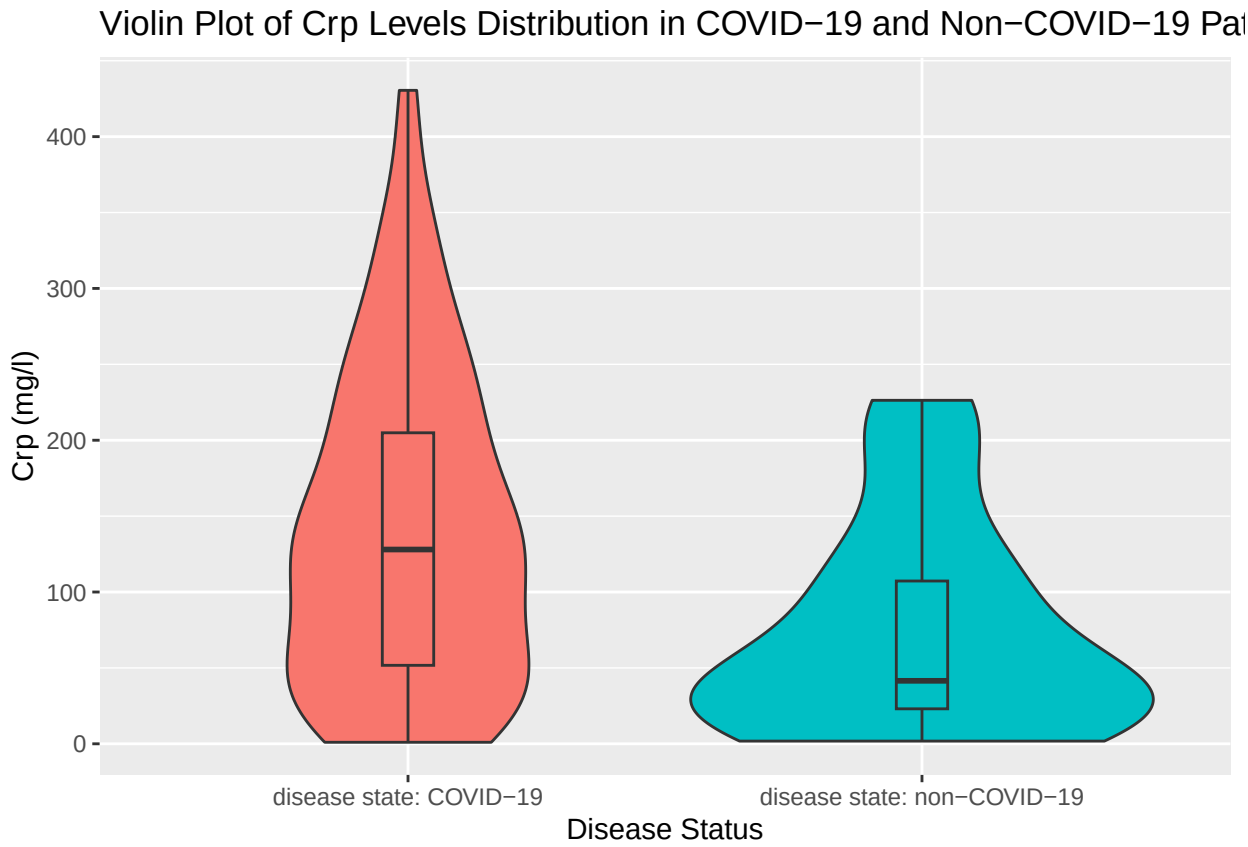
```
df <- genes
rownames(df) <- df$X
df <- df[20:30, -1]
df_log <- log2(df + 1)
annotation <- data.frame(participant_id = factor(series$participant_id),
                         sex = factor(series$sex),
                         disease_status = factor(series$disease_status))
rownames(annotation) <- annotation[,1]
annotation <- annotation[,-1]

cols <- list(sex = c('male' = 'pink', 'female' = 'red'),
                     disease_status = c('disease state: COVID-19' = 'cyan',
                                        'disease state: non-COVID-19' = 'orange'))
pheatmap(df_log,
        clusters_cols = T,
        cluster_rows = T,
        border_color = 'white',
        annotation_col = annotation,
        annotation_colors = cols,
        show_colnames = FALSE,
        main = "Gene Expression Heatmap")
```



The heatmap use 11 genes across all samples, with both rows and columns clustered, provides insights into the expression patterns of these genes. In order to reduce the degree of dispersion, the data were log2 processed.The clustering indicates that genes and samples share similar expression profiles, suggesting potential co-regulation or similar biological conditions.There was a clear grouping pattern in gene expression between COVID-19 patients and non-COVID-19 patients.

```
ggplot(data, aes(x=disease_status, y=crp.mg.l., fill=disease_status)) +
  geom_violin(show.legend = F) +
  geom_boxplot(width=0.1, show.legend = F) +
  labs(x="Disease Status", y="Crp (mg/l)",
       title="Violin Plot of Crp Levels Distribution in COVID-19 and Non-COVID-19 Patients")
```



This violin plot shows the distribution of C-reactive protein (Crp) levels between COVID-19 and non-COVID-19 patients. The plot indicates that Crp levels are generally higher and more widely distributed among COVID-19 patients, with a noticeably higher median compared to non-COVID-19 patients, suggesting significantly elevated inflammation markers in those with COVID-19.

References

[1] tidyverse: Wickham, H., et al. (2019). "Welcome to the Tidyverse." Journal of Open Source Software, 4(43), 1686. DOI: 10.21105/joss.01686.

[2] knitr: Xie, Y. (2015). "Dynamic Documents with R and knitr." 2nd edition. Chapman; Hall/CRC.

[3] pheatmap: Kolde, R. (2019). "pheatmap: Pretty Heatmaps." R package version 1.0.12. CRAN.

[4] Overmyer KA, Shishkova E, Miller IJ, et al.(2020) Large-Scale Multi-omic Analysis of COVID-19 Severity. Cell Syst. 2021;12(1):23-40.e7. doi:10.1016/j.cels.