JFAITH_QBS103_final_project

2025-07-12

Comparing Biomarkers to Covid-19

Data Wrangling

Read in files

```
gene_data <- read.csv("data/QBS103_GSE157103_genes.csv")
meta_data <- read.csv('data/QBS103_GSE157103_series_matrix-1.csv')
```

Gene isolation

I chose to look at the ABCF1 gene becaue of it's role in retulating immune responses. Source: https://journals.plos.org/plosone/article?id=10.1371/journal.pone.0175918

```
# Select ABCF1 gene
gene <- gene_data[which(gene_data$X=='ABCF1'), ]

# Get all column names except X
column_names <- names(gene)[-1]

#Melt dataframe
#referenced: https://tidyr.tidyverse.org/reference/pivot_longer.html
gene_long <- pivot_longer(gene, cols = all_of(column_names), names_to = "participant_id", values_to = "
```

Merge metadata & gene data

```
gene_dataset <- inner_join(gene_long, meta_data, by="participant_id")
# referenced: https://datascienceplus.com/merging-datasets-with-tidyverse/


# Check to see if every record in the gene data matched to the metadata table since inner joining
if (nrow(gene_long) == nrow(meta_data)) {
  print('All records matched and none were dropped through the merge')
} else (
  print('missing records')
)
```

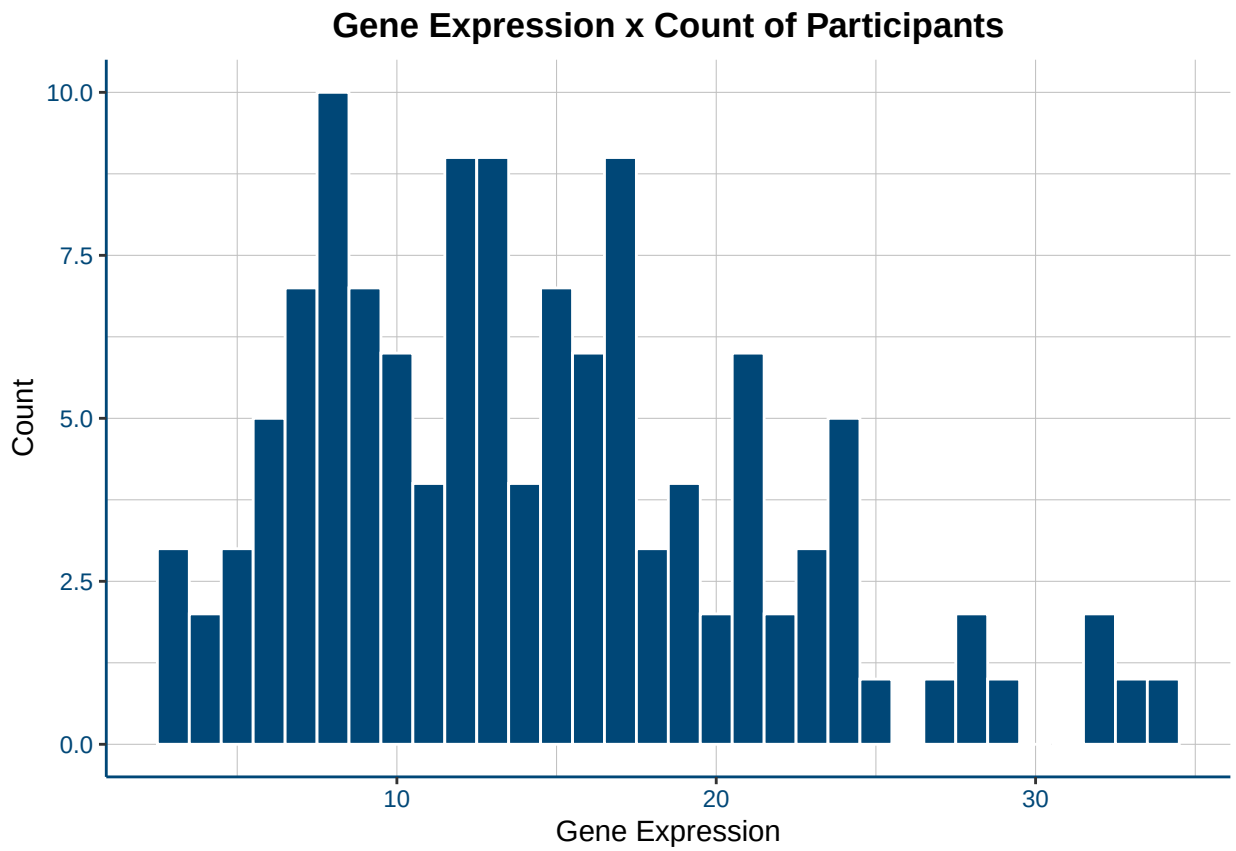
```
## [1] "All records matched and none were dropped through the merge"
```

Visualizations

Histogram for gene expression

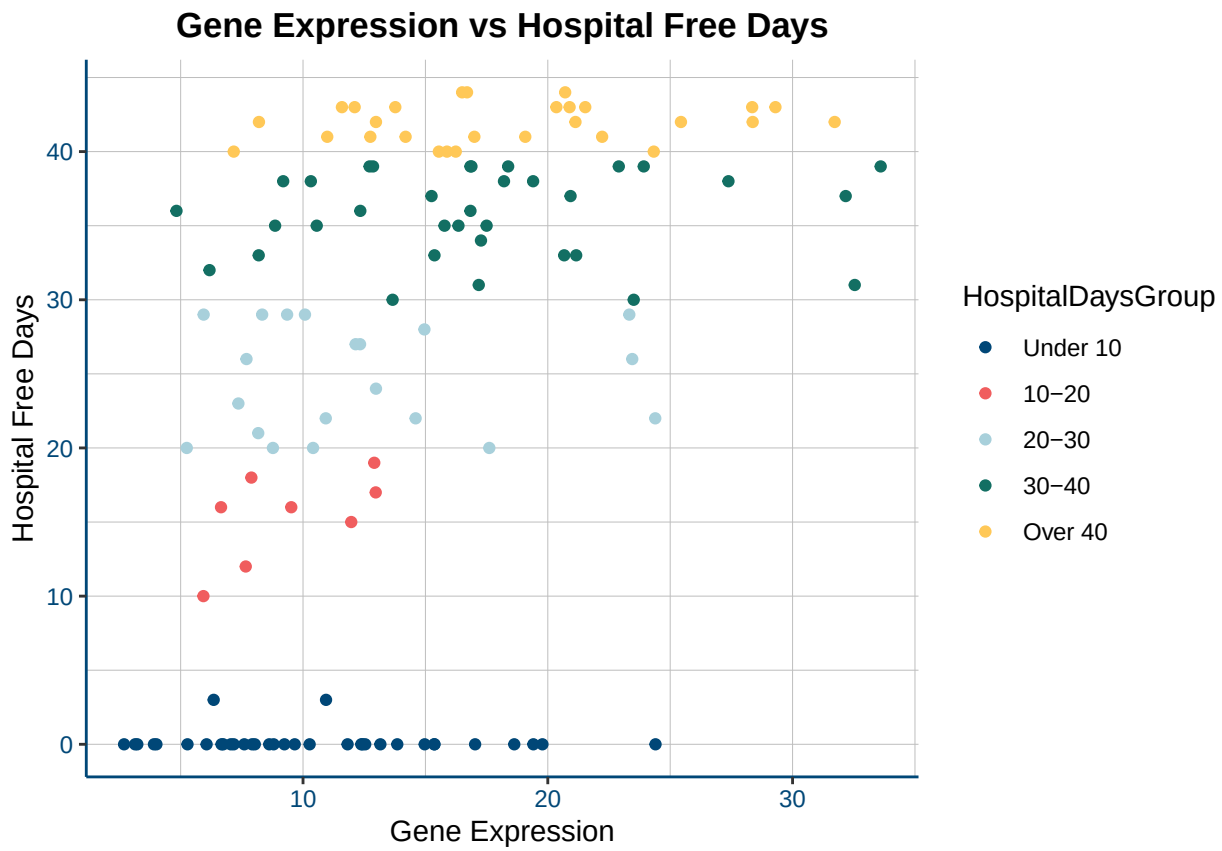
```
# Adding custom color category by days in the hospital
gene_dataset$HospitalDaysGroup <- cut(
  gene_dataset$hospital.free_days_post_45_day_followup,
  breaks = c(0, 10, 20, 30, 40, 50),
  labels = c('Under 10', '10-20', '20-30', '30-40', 'Over 40'),
  right = FALSE
)
```

```
#Custom Color Palette
colorPalette <- c('#004777', '#F05D5E', '#A8D0DB', '#136F63', '#FFC857', '#9DD9D2')

myTheme <- theme(
        panel.border = element_blank(),
        panel.grid.major = element_line(colour="grey", linewidth = rel(.1)),
        panel.grid.minor= element_line(colour="grey", linewidth = rel(.1)),
        #panel.grid.minor = element_blank(),
        # Define my axis
        plot.title = element_text(colour = "black", hjust = .5, face='bold'),
        axis.line = element_line(colour = '#004777', linewidth = rel(1)),
        axis.title.x = element_text(color='black'),
        #axis.title.y = element_text(color='black'),
        axis.text = element_text(color='#004777'),
        # Set plot background
        panel.background = element_blank()
        )


ggplot(gene_dataset, aes(x = expression)) +
  geom_histogram(binwidth=1, color="white", fill='#004777') +
  labs(title="Gene Expression x Count of Participants", x = 'Gene Expression',y = 'Count') +
  myTheme
```



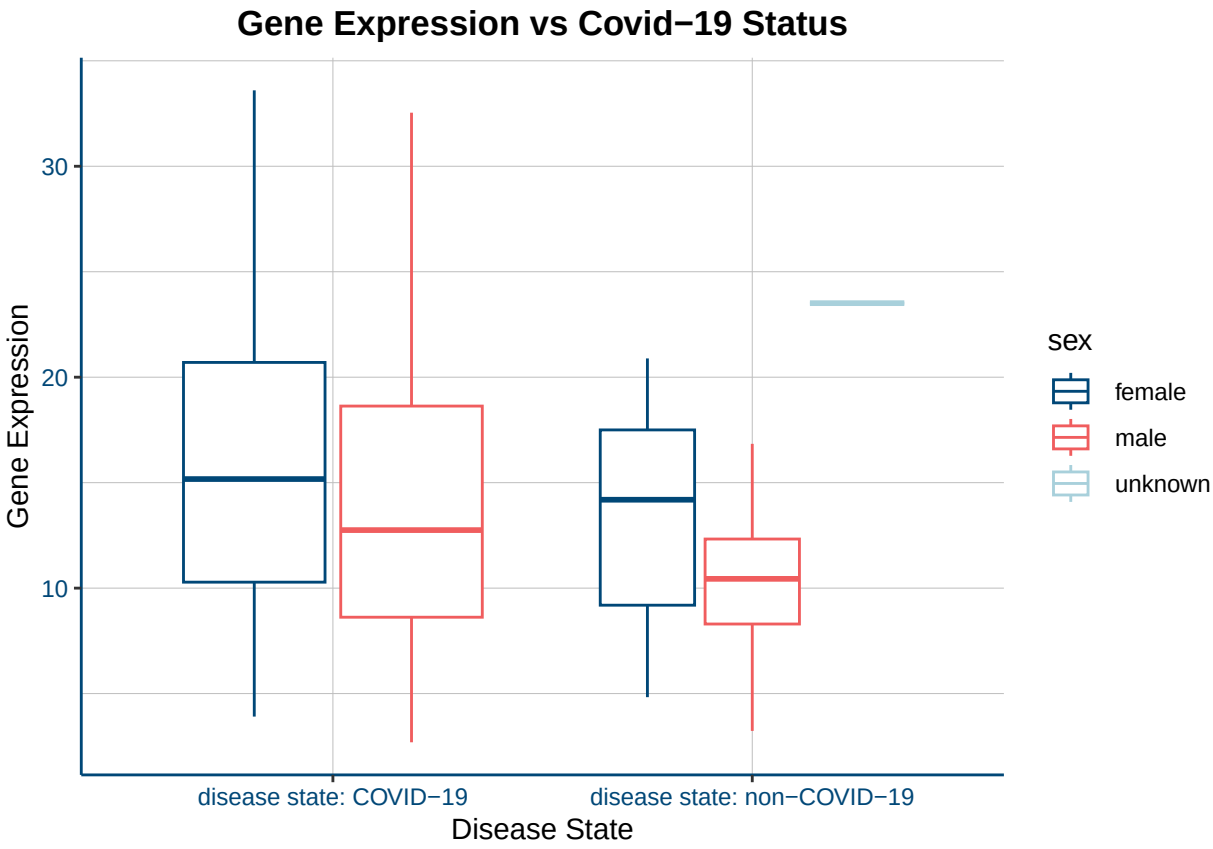
Scatterplot for gene expression and continuous covariate

```
ggplot(gene_dataset, aes(x = expression, y = hospital.free_days_post_45_day_followup, color = HospitalDa
  geom_point() +
  labs(title="Gene Expression vs Hospital Free Days", x = 'Gene Expression',y = 'Hospital Free Days')+
  scale_color_manual(values = colorPalette)+
  myTheme
```



Boxplot of gene expression separated by both categorical covariates

```
ggplot(gene_dataset, aes(x = disease_status, y = expression, color=sex)) +
  geom_boxplot() +
  labs(title="Gene Expression vs Covid-19 Status", x = 'Disease State',y = 'Gene Expression')+
  scale_color_manual(values = colorPalette) +
  myTheme
```


Gene Expression vs Covid−19 Status
Gene Expression
30
20
10
disease state: COVID−19
disease state: non−COVID−19
Disease State
sex
female
male
unknown