Submission 1 Script

Sabin Hart

2024-07-25

Library Imports

```
## -- Attaching core tidyverse packages ------------------------ tidyverse 2.0.0 --
## v dplyr     1.1.4     v readr     2.1.5
## v forcats   1.0.0     v stringr   1.5.1
## v ggplot2   3.4.4     v tibble    3.2.1
## v lubridate 1.9.3     v tidyr     1.3.1
## v purrr     1.0.2
## -- Conflicts ------------------------------------------ tidyverse_conflicts() --
## x dplyr::filter() masks stats::filter()
## x dplyr::lag()    masks stats::lag()
## i Use the conflicted package (<http://conflicted.r-lib.org/>) to force all conflicts to become errors
```

Data Import

```
gene_data_raw <- read_csv("./final-data/QBS103_GSE157103_genes.csv")
```

```
## New names:
## Rows: 100 Columns: 127
## -- Column specification
## -------------------------------------------------------- Delimiter: "," chr
## (1): ...1 dbl (126): COVID_01_39y_male_NonICU, COVID_02_63y_male_NonICU,
## COVID_03_33y_...
## i Use `spec()` to retrieve the full column specification for this data. i
## Specify the column types or set `show_col_types = FALSE` to quiet this message.
## * `` -> `...1`
```

```
pheno_data_raw <- read_csv("./final-data/QBS103_GSE157103_series_matrix.csv")
```

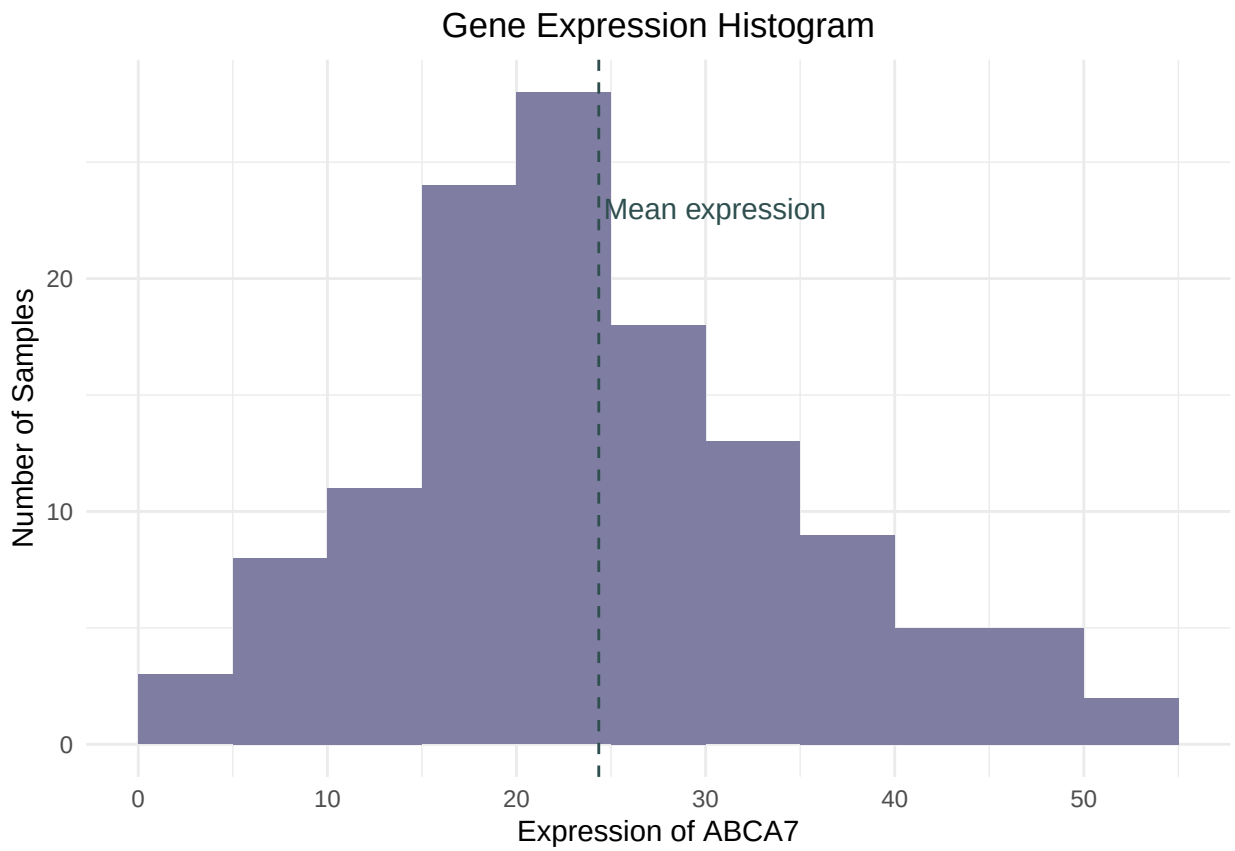
```
## Rows: 126 Columns: 25
## -- Column specification --------------------------------------------------------
## Delimiter: ","
## chr (21): participant_id, geo_accession, status, !Sample_submission_date, la...
## dbl  (4): channel_count, charlson_score, ventilator-free_days, hospital-free...
##
## i Use `spec()` to retrieve the full column specification for this data.
## i Specify the column types or set `show_col_types = FALSE` to quiet this message.
```

Data cleaning and merging

```
# flipping gene data
gene_data <- gene_data_raw %>%
  rename("gene_name" = 1) %>%
  column_to_rownames("gene_name") %>%
  t() %>%
  data.frame() %>%
  rownames_to_column("participant_id")

# combine data
cleaned_data <- gene_data %>%
  merge(pheno_data_raw, by='participant_id') %>%
  mutate(participant_id = as.numeric(sub(".*_(\\d+)_.*", "\\1", participant_id))) %>%
  mutate(participant_id = case_when(
    grepl('non', disease_status, ignore.case = TRUE) ~ participant_id + 200,
    TRUE ~ participant_id
  )) %>%
  arrange(participant_id)

rm(gene_data, gene_data_raw, pheno_data_raw)
```

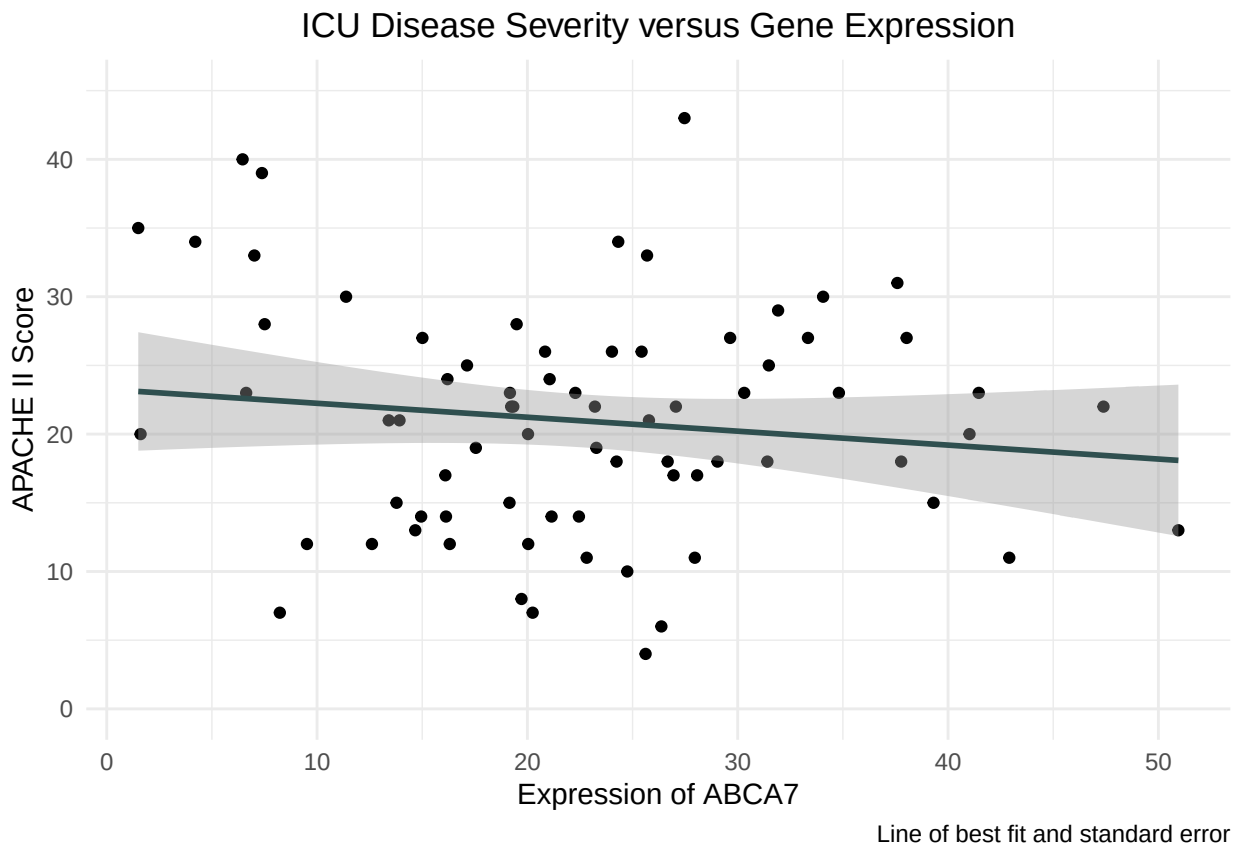
plot 1

```
# data prep
plot1_data <- cleaned_data %>%
  select('ABCA7')

# create plot
plot1 <- ggplot(data = plot1_data, aes(x = ABCA7)) +
  geom_histogram(fill = '#7f7da2', bins = 12, binwidth = 5, center = 2.5) +
  geom_vline(xintercept = mean(plot1_data$ABCA7), linetype = 'dashed', color = 'darkslategrey') +
  annotate("text", x = 30.5, y = 23, color = 'darkslategrey', label = 'Mean expression') +
  theme_minimal() +
  labs(title = "Gene Expression Histogram",
       x = "Expression of ABCA7",
       y = "Number of Samples") +
  theme(plot.title = element_text(hjust = 0.5)) +
  scale_x_continuous(breaks = seq(0, 70, 10))
plot1
```



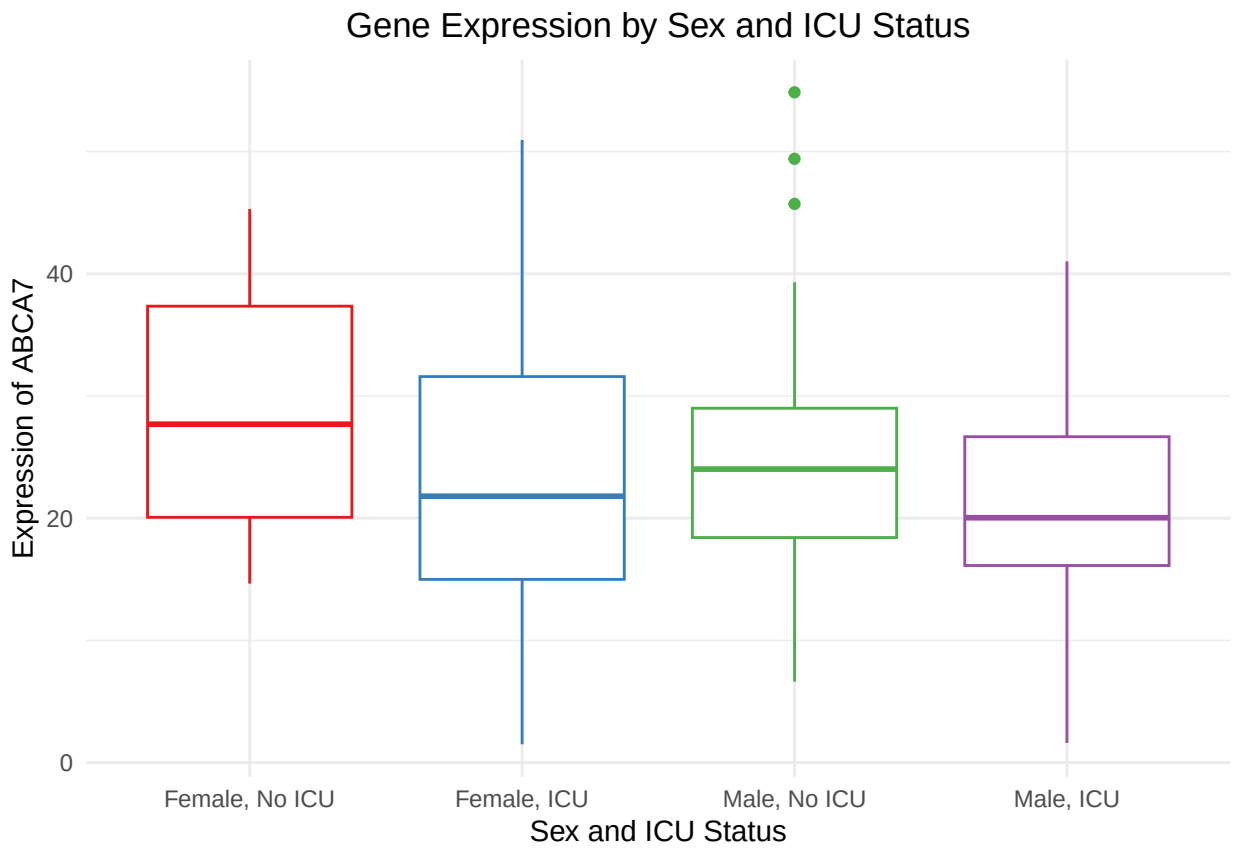
plot 2

```
# data prep
plot2_data <- cleaned_data %>%
  filter(apacheii != 'unknown') %>%
  select(apacheii, ABCA7)

# create plot
plot2 <- ggplot(data = plot2_data, aes(x=ABCA7, y = as.numeric(`apacheii`))) +
  geom_point() +
  geom_smooth(method='lm', formula= y~x, color = 'darkslategray') +
  theme_minimal() +
  labs(title = "ICU Disease Severity versus Gene Expression",
       x = "Expression of ABCA7",
       y = "APACHE II Score",
       caption = "Line of best fit and standard error") +
  theme(plot.title = element_text(hjust = 0.5)) +
  scale_y_continuous(limits = c(0, 45))

plot2
```



plot 3

```
# data prep
plot3_data <- cleaned_data %>%
  mutate(sexXicu_status = case_when(
    sex == 'male' & icu_status == 'no' ~ 'male_no',
    sex == 'female' & icu_status == 'no' ~ 'female_no',
    sex == 'male' & icu_status == 'yes' ~ 'male_yes',
    sex == 'female' & icu_status == 'yes' ~ 'female_yes')) %>%
  select(sexXicu_status, sex, icu_status, ABCA7) %>%
  filter(!is.na(sexXicu_status))

# create plot
plot3 <- ggplot(data = plot3_data, aes(x = sexXicu_status, y = ABCA7, color = sexXicu_status)) +
  geom_boxplot() +
  theme_minimal() +
  labs(title = "Gene Expression by Sex and ICU Status",
       y = "Expression of ABCA7",
       x = "Sex and ICU Status") +
  theme(plot.title = element_text(hjust = 0.5),
        legend.position = "none") +
  scale_x_discrete(labels=(c("Female, No ICU", "Female, ICU", "Male, No ICU", 'Male, ICU'))) +
  scale_colour_brewer(palette = "Set1")

plot3
```



save plots

```
# plot 1
# ggsave('gene_histo.pdf', plot=plot1, device = 'pdf', path = '../plots')

# plot 2
# ggsave('gene_scatter.pdf', plot=plot2, device = 'pdf', path = '../plots')

# plot 3
# ggsave('gene_boxplot.pdf', plot=plot3, device = 'pdf', path = '../plots')

# commented out to not resave on knitting
```