Final Submission

Sabin Hart

2024-08-25

Libraries

```
library(tidyverse)
```

```
## -- 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
```

```
library(pheatmap)
library(ggthemes)
library(stringr)
library(rlist)
library(stargazer)
```

```
## Warning: package 'stargazer' was built under R version 4.0.5
```

```
##
## Please cite as:
##
##  Hlavac, Marek (2022). stargazer: Well-Formatted Regression and Summary Statistics Tables.
##  R package version 5.2.3. https://CRAN.R-project.org/package=stargazer
```

Data Import and Prep

```
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.
```

```
set.seed(07212001)
```

Data Cleaning

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

gene <- gene_data %>%
  select(-1, -AADAC) %>%
  select(sample(1:99, 15, replace = FALSE))

covs <- pheno_data_raw %>%
  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
  )) %>%
  select(sex, icu_status) %>%
  mutate(row_index = row_number())

# 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)
```

Summary Stats Table

```
# 2 additional continuous (3 total) and 1 additional categorical variable (3 total)

# continuous: apacheii, ferritin(ng/ml), age
```

```r
# categorical: stratify on sex; mechanical_ventilation, icu_status

categorical_table <- cleaned_data %>%
  select(sex, mechanical_ventilation, icu_status) %>%
  filter(sex != 'unknown') %>%
  group_by(sex) %>%
  summarize(mv_n = sum(mechanical_ventilation=='yes'),
            mv_p = round(mean(mechanical_ventilation=='yes')*100, 1),
            icu_n = sum(icu_status == 'yes'),
            icu_p = round(mean(icu_status == 'yes')*100, 1),
            c = n()) %>%
  mutate('Sex' = sex,
         `On Mechanical Ventilation` = paste0(mv_n, " (", mv_p, "%)"),
         `In ICU` = paste0(icu_n, " (", icu_p, "%)"),
         Count = c,
         .keep='none')

continuous_table <- cleaned_data %>%
  select(apacheii, `ferritin(ng/ml)`, age, sex) %>%
  filter(sex != 'unknown') %>%
  filter(apacheii != 'unknown') %>%
  filter(`ferritin(ng/ml)` != 'unknown') %>%
  mutate(age = as.numeric(age),
         apacheii = as.numeric(apacheii),
         `ferritin(ng/ml)` = as.numeric(`ferritin(ng/ml)`)) %>%
  group_by(sex) %>%
  summarize(a_mean = round(mean(age, na.rm = T), 1),
            a_sd = round(sd(age, na.rm = T), 1),
            ap_mean = round(mean(apacheii, na.rm = T), 1),
            ap_sd = round(sd(apacheii), 1),
            f_mean = round(mean(`ferritin(ng/ml)`, na.rm = T)),
            f_sd = round(sd(`ferritin(ng/ml)`))) %>%
  mutate('Sex' = sex,
         Age = paste0(a_mean, " (", a_sd, ")"),
         `Apache II` = paste0(ap_mean, " (", ap_sd, ")"),
         `Ferritin (ng/mL)` = paste0(f_mean, " (", f_sd, ")"),
         .keep = 'none')
```

```
## Warning: There was 1 warning in `mutate()`.
## i In argument: `age = as.numeric(age)`.
## Caused by warning:
## ! NAs introduced by coercion
```

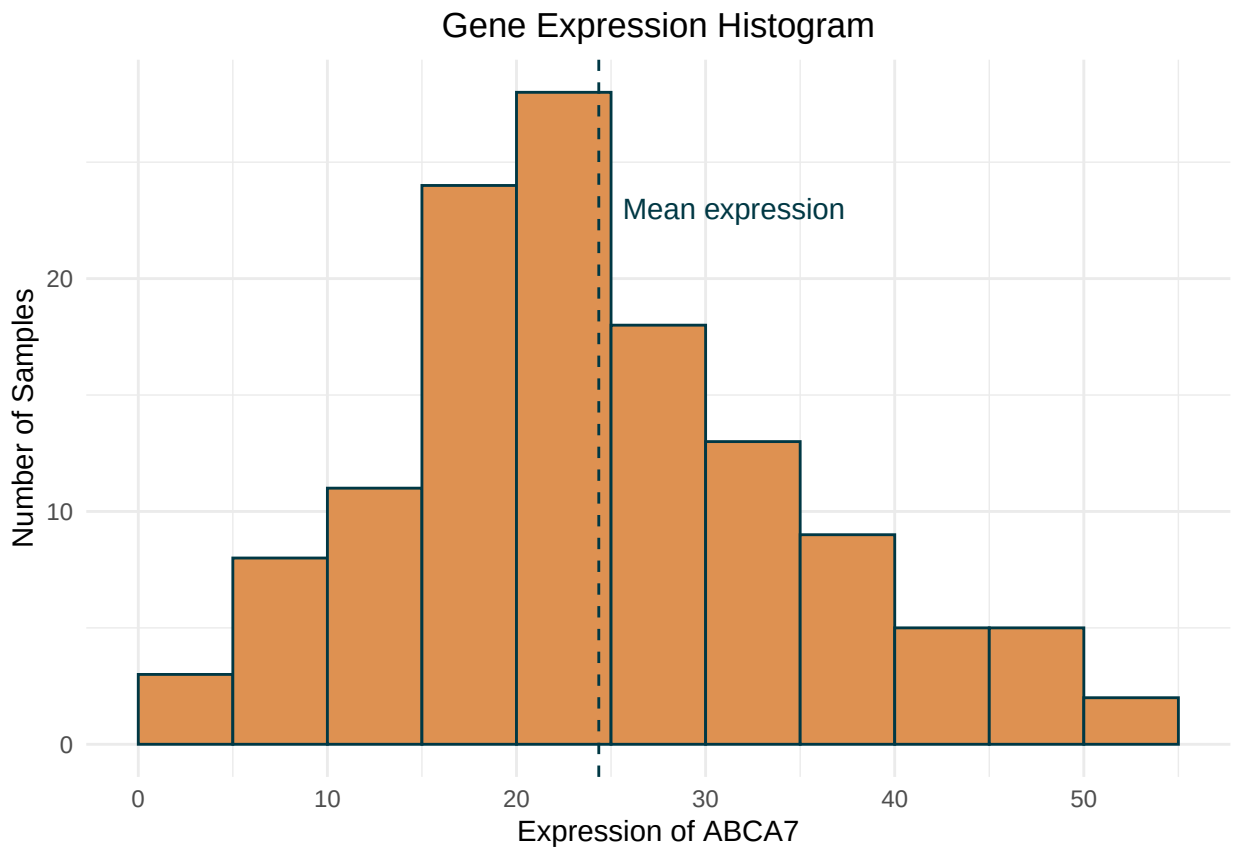
```r
summary_table <- left_join(continuous_table, categorical_table, by = "Sex")
rm(categorical_table, continuous_table)
stargazer(summary_table,
          type = 'latex',
          title = "Summary Statistics",
          summary = FALSE,
          out = 'plots/complete/summary_table.tex')
```

```
##
```

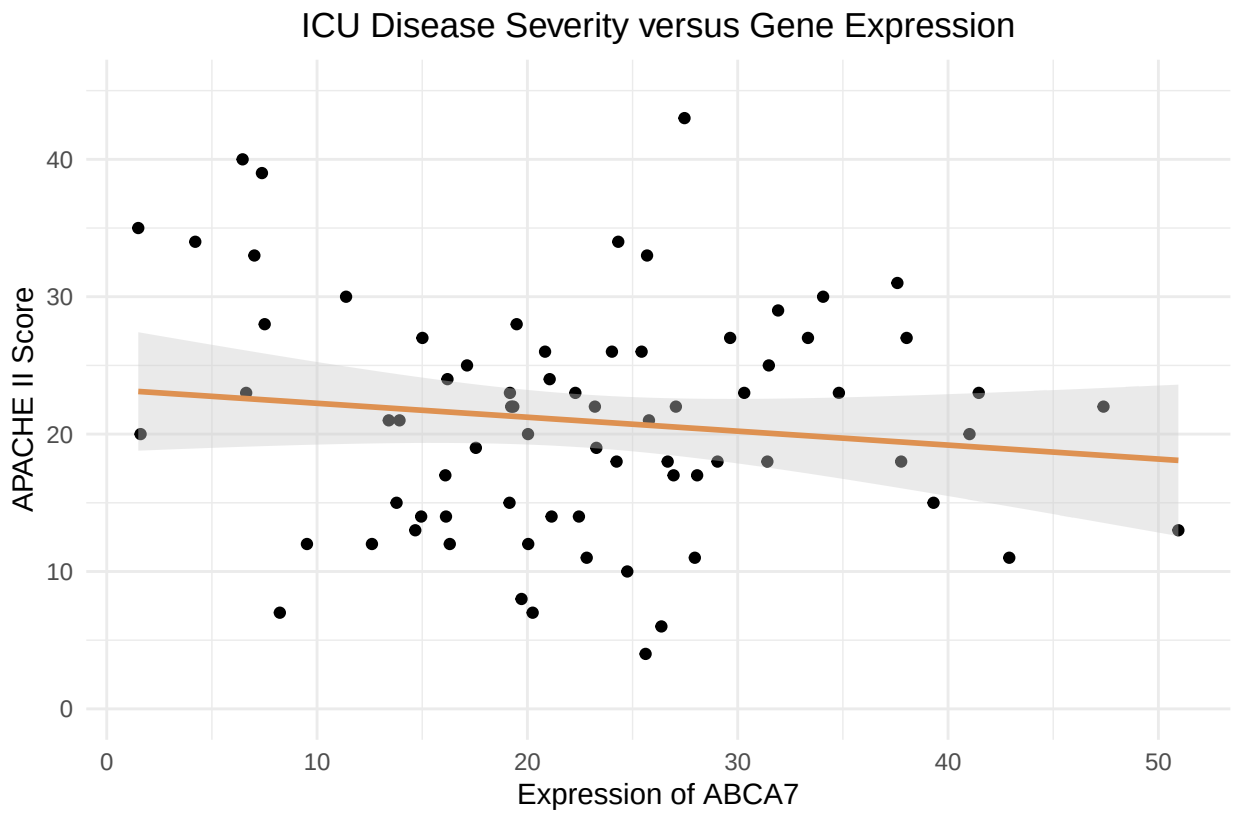
```
## % Table created by stargazer v.5.2.3 by Marek Hlavac, Social Policy Institute. E-mail: marek.hlavac a
## % Date and time: Sun, Aug 25, 2024 - 13:05:13
## \begin{table}[!htbp] \centering
##   \caption{Summary Statistics}
##   \label{}
## \begin{tabular}{@{\extracolsep{5pt}} cccccccc}
## \\[-1.8ex]\hline
## \hline \\[-1.8ex]
##  & Sex & Age & Apache II & Ferritin (ng/mL) & On Mechanical Ventilation & In ICU & Count \\
## \hline \\[-1.8ex]
## 1 & female & 66.5 (16.2) & 21.4 (8.9) & 668 (1087) & 16 (31.4\%) & 24 (47.1\%) & 51 \\
## 2 & male & 62.3 (12.1) & 20.7 (7.5) & 1073 (911) & 35 (47.3\%) & 41 (55.4\%) & 74 \\
## \hline \\[-1.8ex]
## \end{tabular}
## \end{table}
```

Final Plots

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

# create plot
plot1 <- ggplot(data = plot1_data, aes(x = ABCA7)) +
  geom_histogram(fill = '#DE9151', bins = 12, binwidth = 5, center = 2.5, color = "#003844") +
  geom_vline(xintercept = mean(plot1_data$ABCA7), linetype = 'dashed', color = '#003844') +
  annotate("text", x = 31.5, y = 23, color = '#003844', 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
```

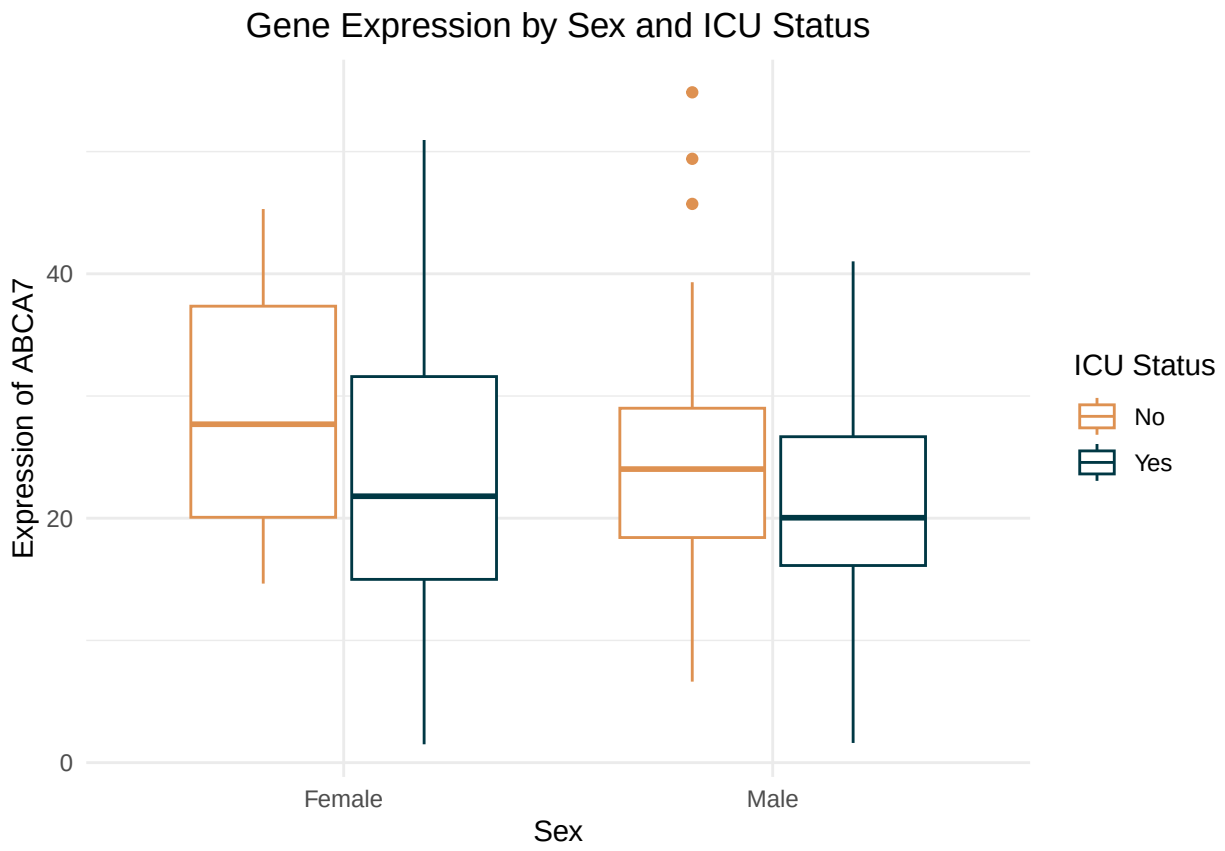



```
# data prep
plot2_data <- cleaned_data %>%
  filter(apacheii != 'unknown') %>%
  select(apacheii, ABCA7)

# create plot
plot2 <- ggplot(data = plot2_data, aes(x=ABCA7, y = as.numeric(`apacheii`))) +
  geom_point() +
  geom_smooth(method='lm', formula= y~x, color = '#DE9151', fill = "grey80") +
  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
```



Line of best fit and standard error

```
plot3_data <- cleaned_data %>%
  select(sex, icu_status, ABCA7) %>%
  filter(sex != 'unknown') %>%
  na.omit()


# create plot
plot3 <- ggplot(data = plot3_data, mapping = aes(y=ABCA7, x = sex, color = icu_status)) +
  geom_boxplot() +
  theme_minimal() +
  labs(title = paste0("Gene Expression by Sex and ICU Status"),
       y = "Expression of ABCA7",
       x = 'Sex') +
  theme(plot.title = element_text(hjust = 0.5)) +
  scale_colour_manual("ICU Status", values = c("#DE9151", "#003844"),
                      labels = c("No", "Yes")) +
  scale_x_discrete(labels = c("Female", "Male"))

plot3
```

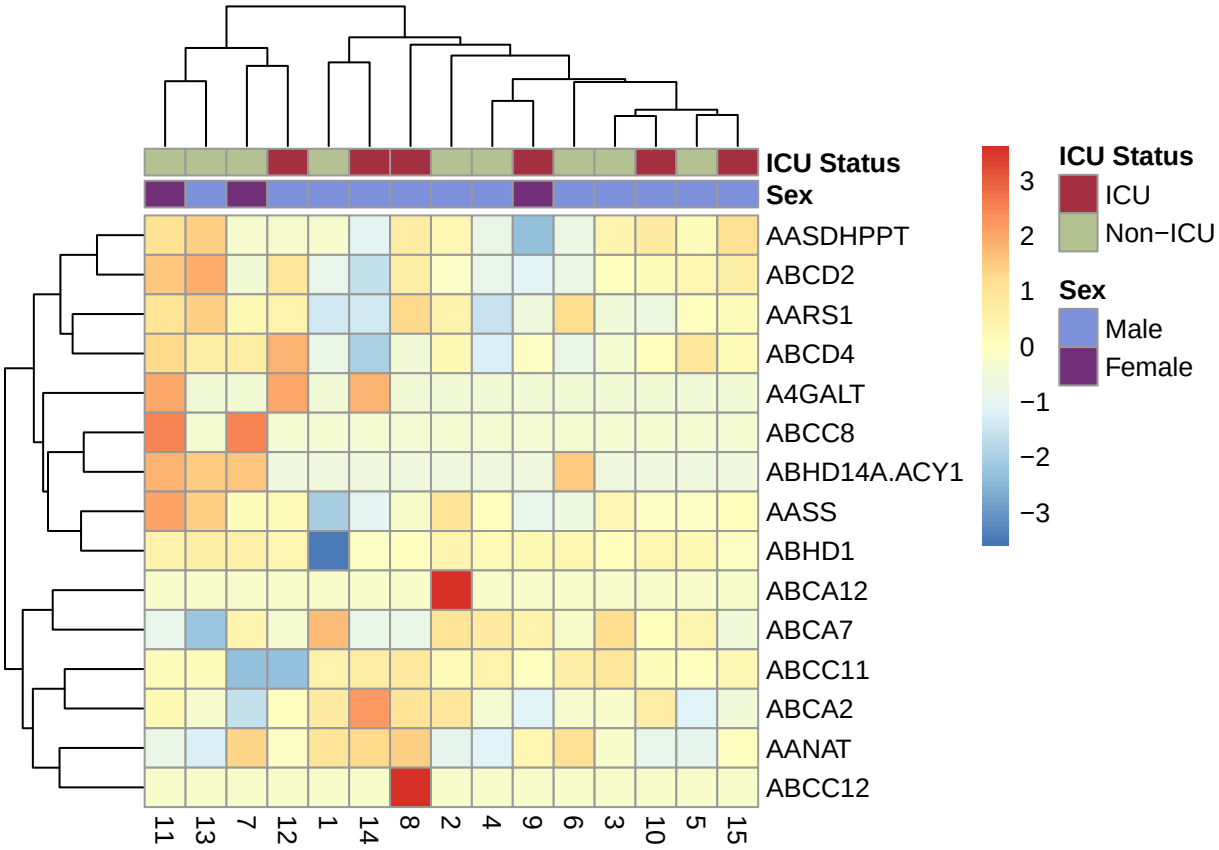



```
rm(plot1_data, plot2_data, plot3_data)
```

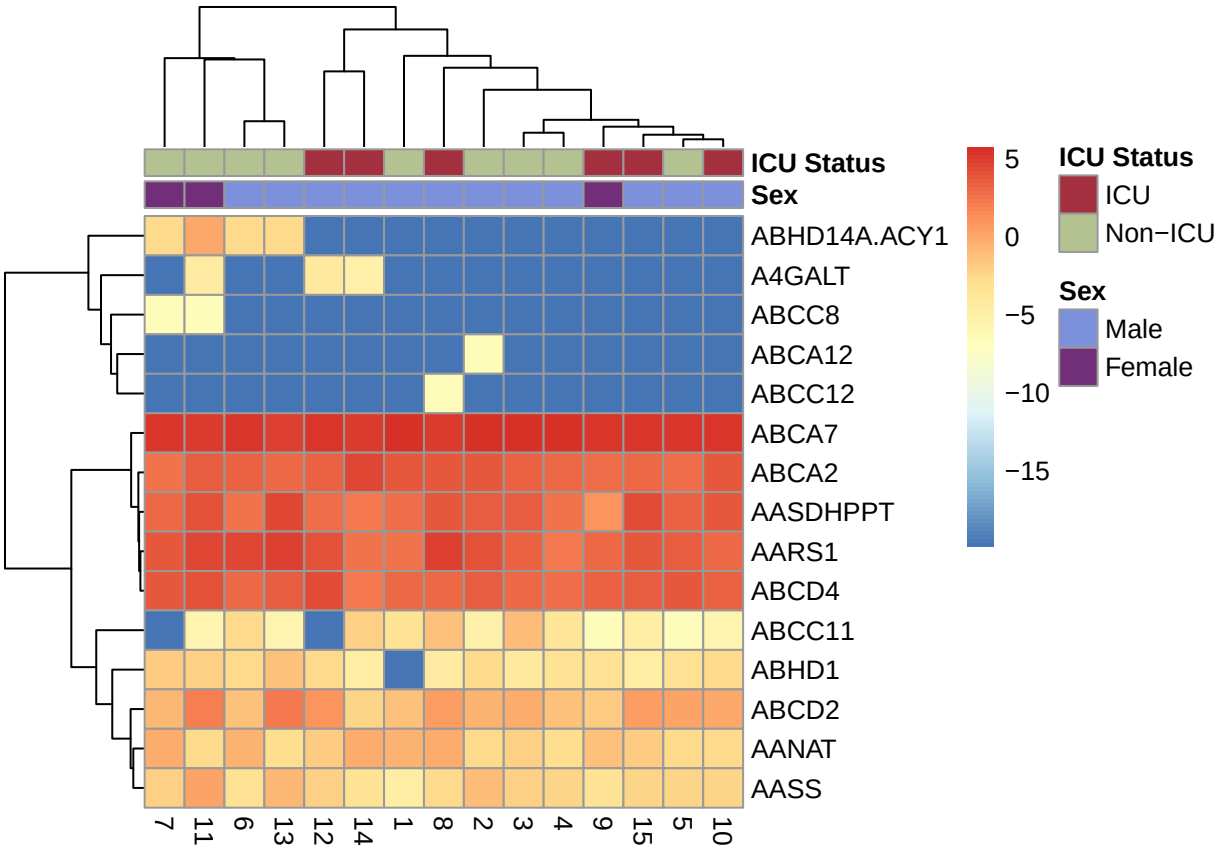
Heatmaps

```
# data prep
variances <- apply(gene, MARGIN = 1,FUN = var)
genes <- gene[order(variances,decreasing = T),]
genes[genes == 0] <- 0.000001
log2.genes <- log2(genes)
log2.genes <- log2.genes %>%
  mutate(row_index = row_number())

# collect gene data with annotation cols and rename for clarity
heatmap_data <- left_join(log2.genes, covs, by = "row_index") %>%
  select(-row_index) %>%
  rename('ICU Status' = icu_status,
         'Sex'  = sex)

# correct spelling for professionalism
annotations <- heatmap_data[, c("Sex", "ICU Status")]
annotations[annotations=='male'] <- 'Male'
annotations[annotations=='female'] <- 'Female'
annotations[annotations=='no'] <- 'Non-ICU'
annotations[annotations=='yes'] <- 'ICU'
```

```
annotation_colors <- list(
  Sex = c("Male" = "#7C90DB", "Female" = "#712F79"),  # Example colors for sex
  `ICU Status` = c("ICU" = "#A4303F", "Non-ICU" = "#B4C292")  # Example colors for icu_status
)

scaled_heatmap <- pheatmap(t(heatmap_data[1:15,1:15]),
                    annotation_col = annotations,
                    annotation_colors = annotation_colors,
                    scale = "row")
```




```
unscaled_heatmap <- pheatmap(t(heatmap_data[1:15,1:15]),
                             annotation_col = annotations,
                             annotation_colors = annotation_colors)
```

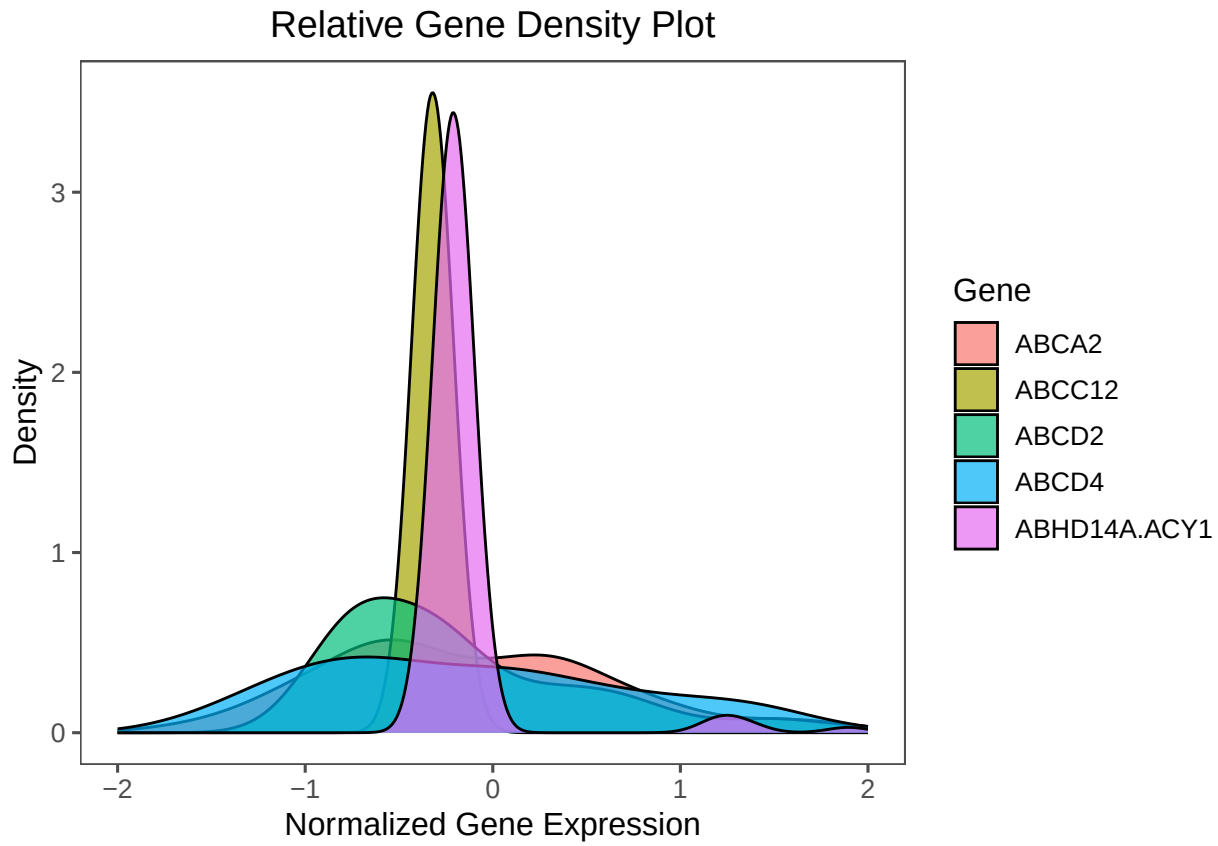



```
rm(genes, heatmap_data, log2.genes, variances, covs, annotations, annotation_colors)
```

Density Plot

```
# prepare data
density_data <- gene[11:15] %>%
  scale() %>%
  as.data.frame() %>%
  pivot_longer(everything(),
               names_to = "Gene")


# plot
density_plot <- ggplot(density_data, aes(value)) +
  geom_density(aes(fill = factor(Gene)), alpha = 0.7) +
  labs(title="Relative Gene Density Plot",
       x="Normalized Gene Expression",
       y = "Density",
       fill="Gene") +
  theme_few() +
  theme(plot.title = element_text(hjust = 0.5)) +
  scale_x_continuous(limits = c(-2, 2))
density_plot
```

```
## Warning: Removed 32 rows containing non-finite values (‘stat_density()‘).
```




```
rm(density_data)
```

Save Plots

```
## All Commented out to avoid re-saving

# ggsave("histogram.png",
#       plot=plot1,
#       path="plots/complete/",
#       width = 6,
#       height = 5)

# ggsave("scatter.png",
#       plot=plot2,
#       path="plots/complete/",
#       width = 6,
#       height = 5)

# ggsave("boxplot.png",
#       plot=plot3,
#       path="plots/complete/",
#       width = 6,
#       height = 5)

# ggsave("density.png",
```

```
#       plot=density_plot,
#       path="plots/complete/",
#       width = 6,
#       height = 5)

# ggsave("unscaled_heatmap.png",
#       plot=unscaled_heatmap,
#       path="plots/complete/",
#       width = 6,
#       height = 5)

# ggsave("scaled_heatmap.png",
#       plot=scaled_heatmap,
#       path="plots/complete/",
#       width = 6,
#       height = 5)
```