Final Project

Xinqiao Li

2025-08-19

```
# Table of summary statistics
meta <- read.csv("QBS103_GSE157103_series_matrix-1.csv", check.names = FALSE)
library(dplyr)
```

```
##
## Attaching package: 'dplyr'
```

```
## The following objects are masked from 'package:stats':
##
##     filter, lag
```

```
## The following objects are masked from 'package:base':
##
##     intersect, setdiff, setequal, union
```

```
meta_clean <- meta %>%
  dplyr::select(
    ventilator_free_days = `ventilator-free_days`,
    charlson_score,
    age,
    sex,
    icu_status,
    mechanical_ventilation
  ) %>%
  mutate(
    across(c(ventilator_free_days, charlson_score, age),
           ~ suppressWarnings(as.numeric(.)))

  ) %>%
  na.omit()

str(meta_clean)
```

```
## 'data.frame':    123 obs. of  6 variables:
##  $ ventilator_free_days  : num  0 28 28 28 23 28 0 0 2 28 ...
##  $ charlson_score        : num  0 2 2 1 1 7 7 2 1 2 ...
##  $ age                   : num  39 63 33 49 49 38 78 64 62 52 ...
##  $ sex                   : chr  " male" " male" " male" " male" ...
##  $ icu_status            : chr  " no" " no" " no" " no" ...
##  $ mechanical_ventilation: chr  " yes" " no" " no" " no" ...
##  - attr(*, "na.action")= 'omit' Named int [1:3] 6 86 104
##   ..- attr(*, "names")= chr [1:3] "6" "86" "104"
```

```
head(meta_clean)
```

```
##   ventilator_free_days charlson_score age     sex icu_status
## 1                    0              0  39    male         no
## 2                   28              2  63    male         no
## 3                   28              2  33    male         no
## 4                   28              1  49    male         no
## 5                   23              1  49    male         no
## 7                   28              7  38  female         no
##   mechanical_ventilation
## 1                    yes
## 2                     no
## 3                     no
## 4                     no
## 5                    yes
## 7                     no
```

```
cont_vars <- c("ventilator_free_days", "charlson_score", "age")
cat_vars  <- c("sex", "mechanical_ventilation")

fmt_mean_sd <- function(x) sprintf("%.2f (%.2f)",
                                   mean(x, na.rm = TRUE),
                                   sd(x,   na.rm = TRUE))

for (v in cont_vars) {
  x <- as.numeric(meta_clean[[v]])
  grp <- meta_clean$icu_status
  res <- tapply(x, grp, fmt_mean_sd)
  cat("Variable:", v, "\n")
  print(res)
  cat("\n")
}
```

```
## Variable: ventilator_free_days
##              no             yes
##  "25.95 (7.22)" "15.17 (11.93)"
##
## Variable: charlson_score
##            no           yes
## "3.11 (2.45)" "3.82 (2.50)"
##
## Variable: age
##              no             yes
## "58.67 (17.82)" "63.45 (14.00)"
```

```
for (v in cat_vars) {
  grp <- meta_clean$icu_status
  tab <- table(meta_clean[[v]], grp)
  prop <- prop.table(tab, margin = 2) * 100
  res <- paste0(tab, " (", sprintf("%.1f", prop), "%)")

  cat("Variable:", v, "\n")
  print(res)
  cat("\n")
}
```

```
## Variable: sex
## [1] "26 (45.6%)" "31 (54.4%)" "24 (36.4%)" "42 (63.6%)"
##
## Variable: mechanical_ventilation
## [1] "52 (91.2%)" "5 (8.8%)"   "20 (30.3%)" "46 (69.7%)"
```

```
#publication quality
library(ggplot2)
library(tidyverse)
```

```
## ── Attaching core tidyverse packages ──────────────────────── tidyverse 2.0.0 ──
## ✔ forcats   1.0.0     ✔ stringr   1.5.1
## ✔ lubridate 1.9.4     ✔ tibble    3.3.0
## ✔ purrr     1.0.4     ✔ tidyr     1.3.1
## ✔ readr     2.1.5
## ── Conflicts ────────────────────────────────────────── tidyverse_conflicts() ──
## ✖ dplyr::filter() masks stats::filter()
## ✖ dplyr::lag()    masks stats::lag()
## ℹ Use the conflicted package (<http://conflicted.r-lib.org/>) to force all conflic
ts to become errors
```

```
library(tidyr)
setwd("/Users/zaozao/Desktop/Xinqiao")
gene_expression <- read.csv("QBS103_GSE157103_genes.csv", row.names = 1, check.names
= FALSE)
meta <- read.csv("QBS103_GSE157103_series_matrix-1.csv", check.names = FALSE)

aass_expression <- as.data.frame(t(gene_expression["AASS", ]))
colnames(aass_expression) <- "AASS"
aass_expression$participant_id <- rownames(aass_expression)

meta$participant_id <- as.character(meta[[1]])

meta_with_aass <- merge(meta, aass_expression, by = "participant_id")

head(meta_with_aass)
```

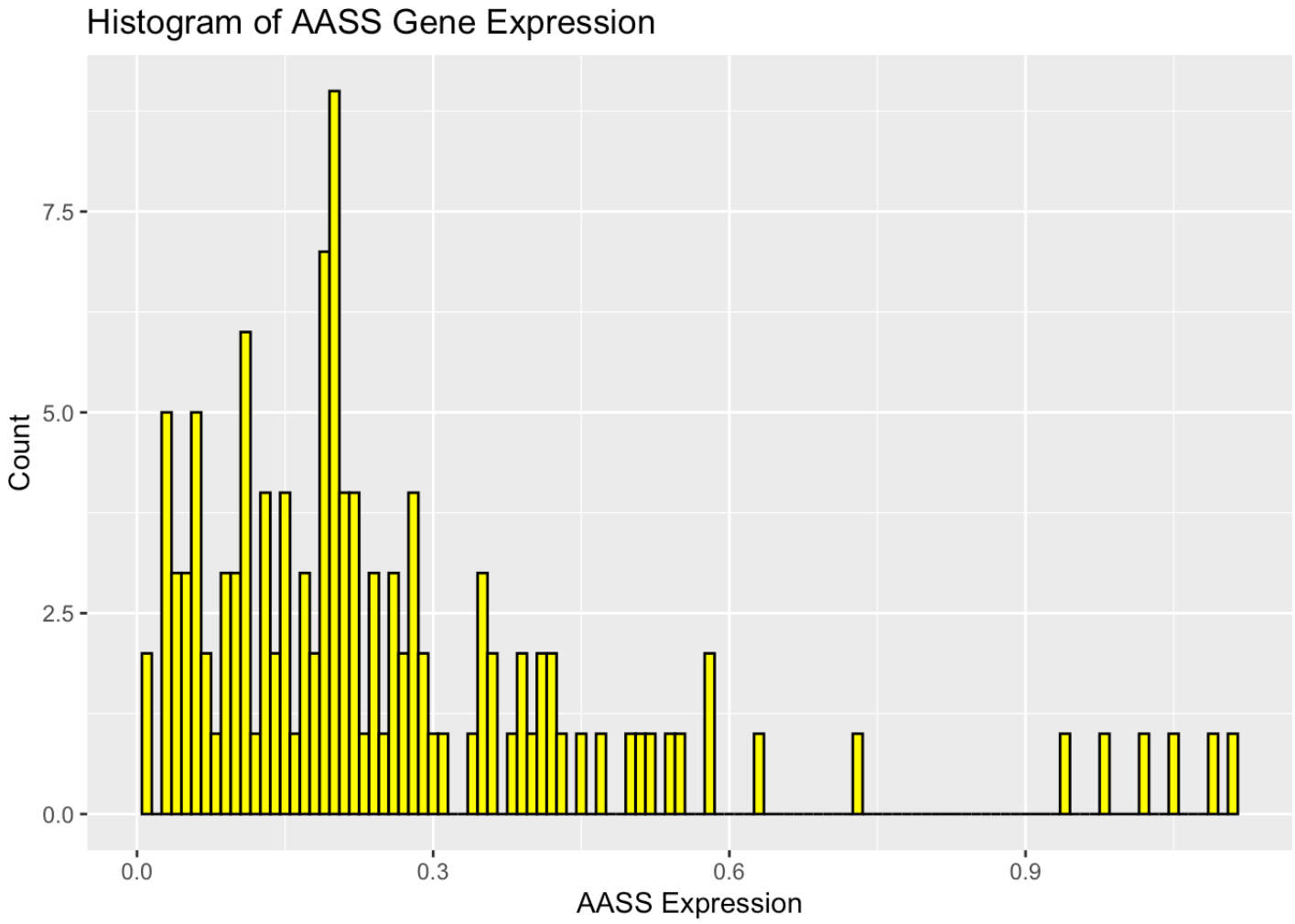
```
##               participant_id geo_accession                 status
## 1   COVID_01_39y_male_NonICU    GSM4753021 Public on Aug 29 2020
## 2   COVID_02_63y_male_NonICU    GSM4753022 Public on Aug 29 2020
## 3   COVID_03_33y_male_NonICU    GSM4753023 Public on Aug 29 2020
## 4   COVID_04_49y_male_NonICU    GSM4753024 Public on Aug 29 2020
## 5   COVID_05_49y_male_NonICU    GSM4753025 Public on Aug 29 2020
## 6 COVID_07_38y_female_NonICU    GSM4753027 Public on Aug 29 2020
##   !Sample_submission_date last_update_date type channel_count
## 1             Aug 28 2020      Aug 29 2020  SRA             1
## 2             Aug 28 2020      Aug 29 2020  SRA             1
## 3             Aug 28 2020      Aug 29 2020  SRA             1
## 4             Aug 28 2020      Aug 29 2020  SRA             1
## 5             Aug 28 2020      Aug 29 2020  SRA             1
## 6             Aug 28 2020      Aug 29 2020  SRA             1
##               source_name_ch1 organism_ch1           disease_status age     sex
## 1 Leukocytes from whole blood Homo sapiens disease statena COVID-19  39    male
## 2 Leukocytes from whole blood Homo sapiens disease statena COVID-19  63    male
## 3 Leukocytes from whole blood Homo sapiens disease statena COVID-19  33    male
## 4 Leukocytes from whole blood Homo sapiens disease statena COVID-19  49    male
## 5 Leukocytes from whole blood Homo sapiens disease statena COVID-19  49    male
## 6 Leukocytes from whole blood Homo sapiens disease statena COVID-19  38  female
##   icu_status apacheii charlson_score mechanical_ventilation
## 1         no       15              0                   yes
## 2         no       na              2                    no
## 3         no       na              2                    no
## 4         no       na              1                    no
## 5         no       19              1                   yes
## 6         no       na              7                    no
##   ventilator-free_days hospital-free_days_post_45_day_followup ferritin(ng/ml)
## 1                    0                                      0             946
```

```
## 2                   28                                        39             1060
## 3                   28                                        18             1335
## 4                   28                                        39              583
## 5                   23                                        27              800
## 6                   28                                        42              366
##   crp(mg/l) ddimer(mg/l_feu) procalcitonin(ng/ml) lactate(mmol/l) fibrinogen
## 1      73.1              1.3                   36             0.9        513
## 2        na             1.03                 0.37              na         na
## 3      53.2             1.48                 0.07              na        513
## 4     251.1             1.32                 0.98            0.87        949
## 5     355.8             0.69                 4.92            1.48        929
## 6        na             0.87                 0.06            1.17        478
##   sofa AASS
## 1    8 0.21
## 2   na 0.42
## 3   na 0.04
## 4   na 0.41
## 5    7 0.21
## 6   na 0.63
```

```
dim(meta_with_aass)
```

```
## [1] 124  26
```

```
library(ggplot2)
# 1. publication quality histogram
ggplot(meta_with_aass, aes(x = AASS)) +
  geom_histogram(binwidth = 0.01, fill = "yellow", color = "black") +
  labs(title = "Histogram of AASS Gene Expression",
       x = "AASS Expression",
       y = "Count")
```

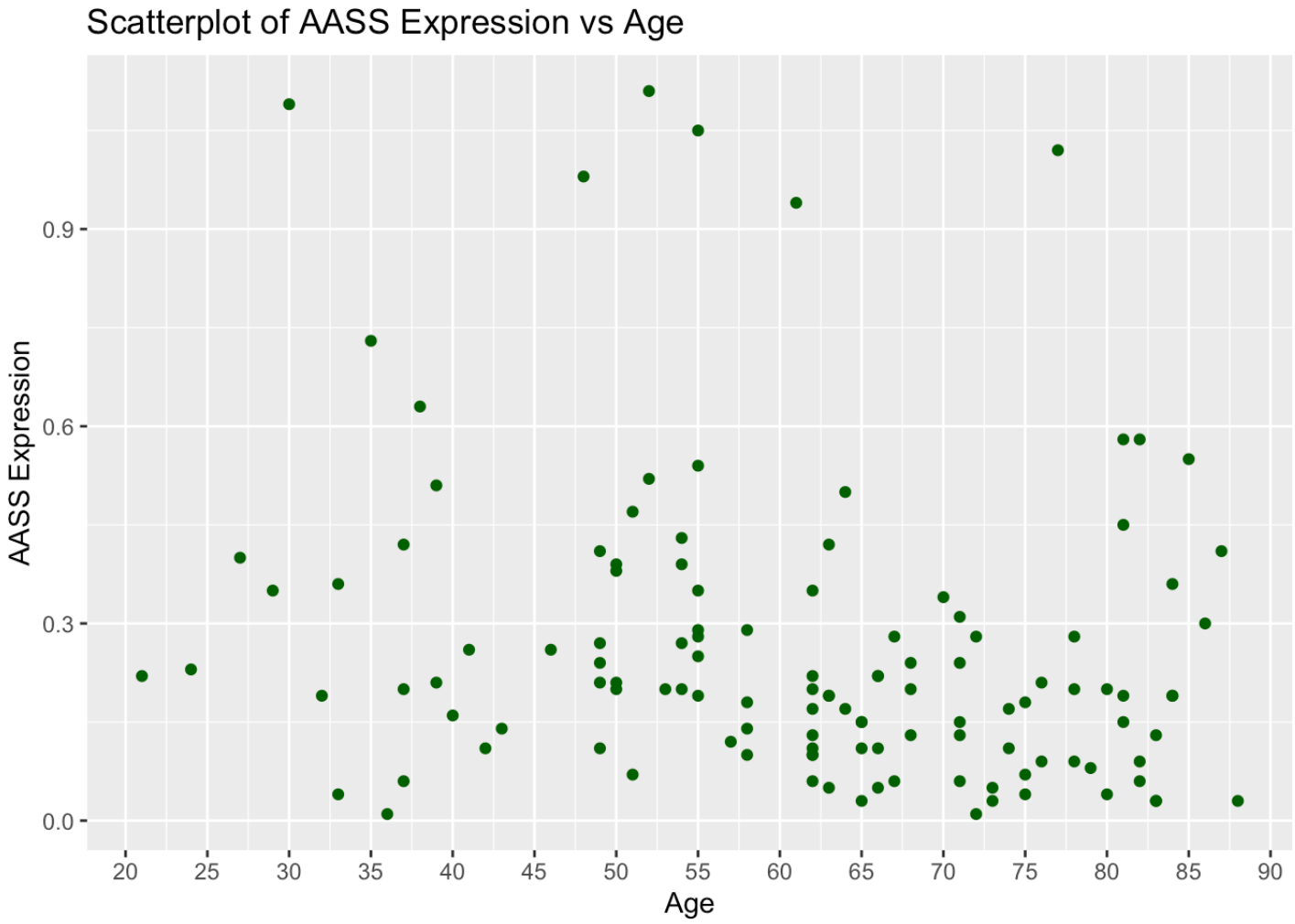



```
# 2. Scatterplot
meta_with_aass$age <- as.numeric(as.character(meta_with_aass$age))
```

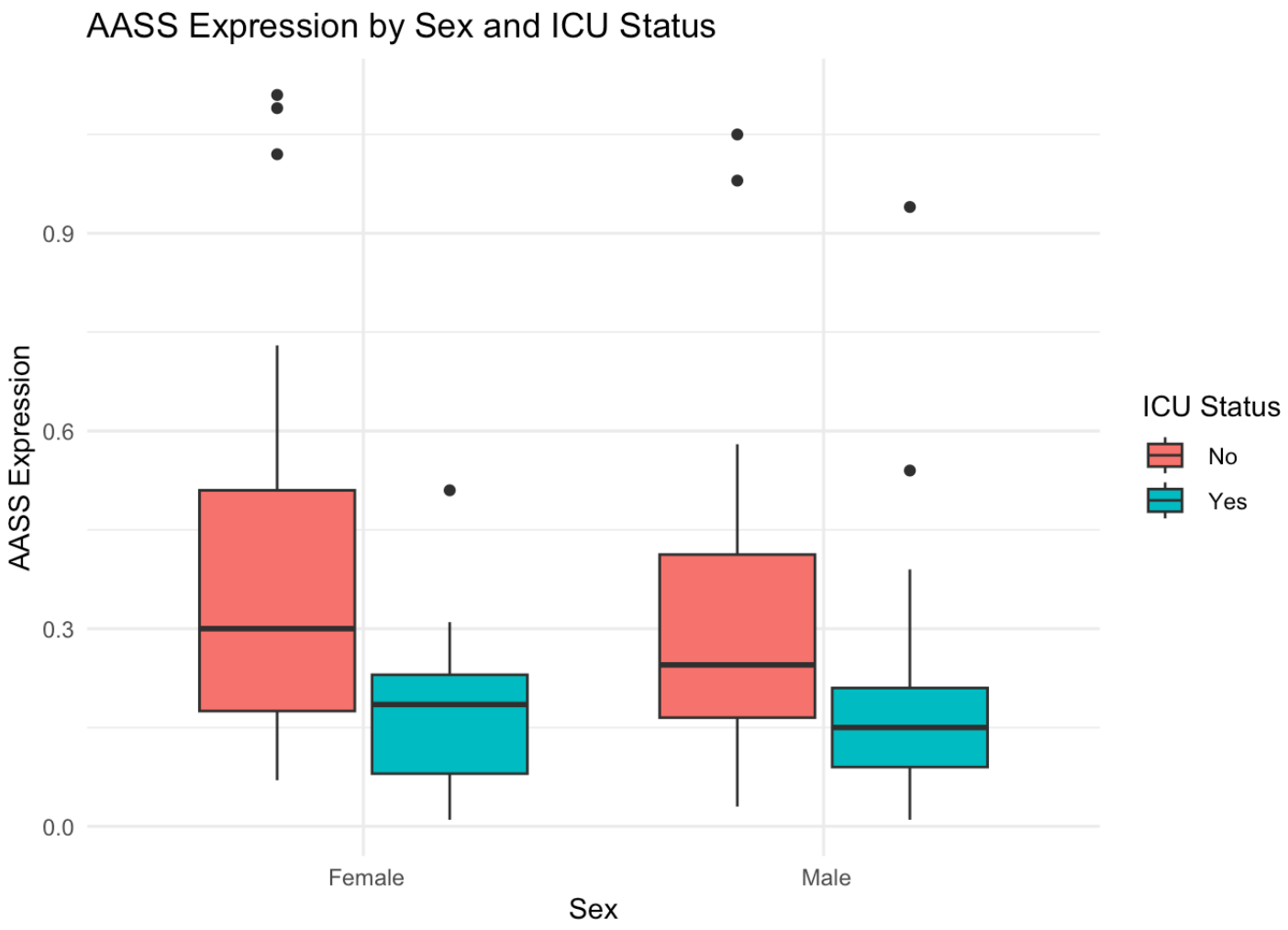
```
## Warning: NAs introduced by coercion
```

```
ggplot(meta_with_aass, aes(x = age, y = AASS)) +
  geom_point(color = "darkgreen") +

  scale_x_continuous(breaks = seq(10, 100, by = 5)) +
  labs(title = "Scatterplot of AASS Expression vs Age",
       x = "Age",
       y = "AASS Expression")
```

```
## Warning: Removed 2 rows containing missing values or values outside the scale rang
e
## (`geom_point()`).
```




```
# 3. Boxplot

meta_with_aass <- meta_with_aass[meta_with_aass$icu_status != "_id", ]


ggplot(meta_with_aass, aes(x = sex, y = AASS, fill = icu_status)) +
  geom_boxplot() +
  labs(title = "AASS Expression by Sex and ICU Status",
       x = "Sex",
       y = "AASS Expression") +
  theme_minimal()+
  scale_x_discrete(labels = stringr::str_to_title)+
  scale_fill_discrete(labels = stringr::str_to_title, name = "ICU Status")
```




```
library(ggplot2)
library(dplyr)

df <- meta_with_aass %>%
  mutate(AASS = suppressWarnings(as.numeric(AASS))) %>%
  filter(is.finite(AASS))

ggplot(df, aes(x = AASS)) +
  geom_area(stat = "density", fill = "gold", alpha = 0.6, color = "black") +
  labs(
    title = "Distribution of AASS Gene Expression",
    x = "AASS expression",
    y = "Density"
  ) +
  theme_classic(base_size = 12) +
  theme(
    plot.title = element_text(face = "bold", hjust = 0),
    axis.title = element_text(face = "bold")
  )
```

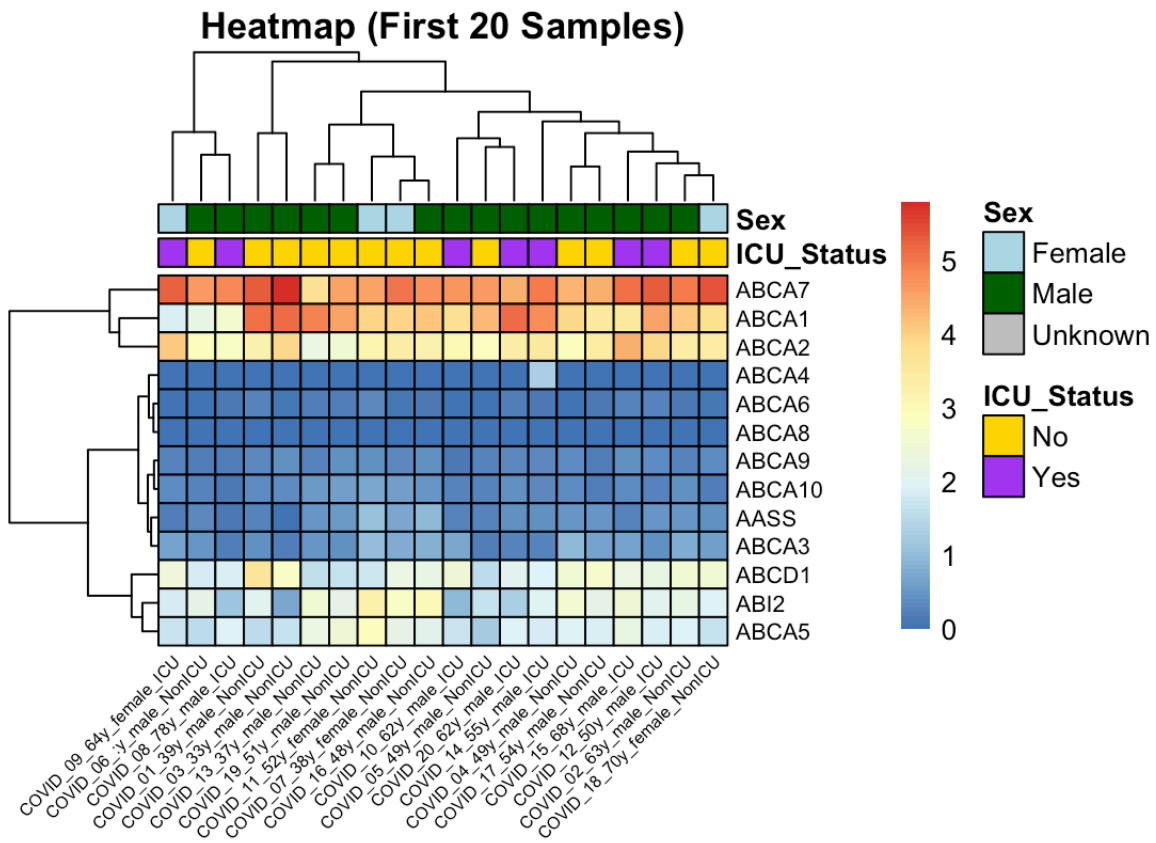

Distribution of AASS Gene Expression
Density
2
1
0
0.0
0.3
0.6
0.9
AASS expression

```
#heatmap
library(dplyr)
setwd("/Users/zaozao/Desktop/Xinqiao")
gene_expression <- read.csv("QBS103_GSE157103_genes.csv", row.names = 1, check.names
= FALSE)
meta <- read.csv("QBS103_GSE157103_series_matrix-1.csv", check.names = FALSE)

selected_genes <- c("AASS", "ABCD1", "ABI2", paste0("ABCA", 1:10))
expr_wide <- as.data.frame(t(gene_expression[selected_genes, , drop = FALSE]))
colnames(expr_wide) <- selected_genes

expr_wide$participant_id <- rownames(expr_wide)
expr_wide <- expr_wide[, c("participant_id", selected_genes)]

meta$participant_id <- colnames(gene_expression)
dataset <- merge(meta, expr_wide, by = "participant_id")

dataset[, selected_genes] <- lapply(dataset[, selected_genes], function(x) as.numeri
c(as.character(x)))

head(dataset)
```

```
##             participant_id geo_accession                status
## 1 COVID_01_39y_male_NonICU    GSM4753021 Public on Aug 29 2020
## 2 COVID_02_63y_male_NonICU    GSM4753022 Public on Aug 29 2020
## 3 COVID_03_33y_male_NonICU    GSM4753023 Public on Aug 29 2020
## 4 COVID_04_49y_male_NonICU    GSM4753024 Public on Aug 29 2020
## 5 COVID_05_49y_male_NonICU    GSM4753025 Public on Aug 29 2020
## 6  COVID_06_:y_male_NonICU    GSM4753026 Public on Aug 29 2020
##   !Sample_submission_date last_update_date type channel_count
## 1             Aug 28 2020      Aug 29 2020  SRA             1
## 2             Aug 28 2020      Aug 29 2020  SRA             1
## 3             Aug 28 2020      Aug 29 2020  SRA             1
## 4             Aug 28 2020      Aug 29 2020  SRA             1
## 5             Aug 28 2020      Aug 29 2020  SRA             1
## 6             Aug 28 2020      Aug 29 2020  SRA             1
##               source_name_ch1 organism_ch1          disease_status age   sex
## 1 Leukocytes from whole blood Homo sapiens disease statena COVID-19  39  male
## 2 Leukocytes from whole blood Homo sapiens disease statena COVID-19  63  male
## 3 Leukocytes from whole blood Homo sapiens disease statena COVID-19  33  male
## 4 Leukocytes from whole blood Homo sapiens disease statena COVID-19  49  male
## 5 Leukocytes from whole blood Homo sapiens disease statena COVID-19  49  male
## 6 Leukocytes from whole blood Homo sapiens disease statena COVID-19  na  male
##   icu_status apacheii charlson_score mechanical_ventilation
## 1         no       15              0                   yes
## 2         no       na              2                    no
```

```
## 3         no       na              2                     no
## 4         no       na              1                     no
## 5         no       19              1                    yes
## 6         no       na              1                     no
##   ventilator-free_days hospital-free_days_post_45_day_followup ferritin(ng/ml)
## 1                    0                                       0             946
## 2                   28                                      39            1060
## 3                   28                                      18            1335
## 4                   28                                      39             583
## 5                   23                                      27             800
## 6                   28                                      36             563
##   crp(mg/l) ddimer(mg/l_feu) procalcitonin(ng/ml) lactate(mmol/l) fibrinogen
## 1      73.1              1.3                   36             0.9        513
## 2        na             1.03                 0.37              na         na
## 3      53.2             1.48                 0.07              na        513
## 4     251.1             1.32                 0.98            0.87        949
## 5     355.8             0.69                 4.92            1.48        929
## 6     129.1               na                 0.67            0.86        769
##   sofa AASS ABCD1 ABI2 ABCA1 ABCA2 ABCA3 ABCA4 ABCA5 ABCA6 ABCA7 ABCA8 ABCA9
## 1    8 0.21 11.26 3.22 32.30  8.47  0.37  0.01  1.86  0.19 39.31  0.00  0.27
## 2   na 0.42  4.65 3.68 15.84  9.49  0.71  0.00  2.81  0.11 30.42  0.00  0.20
## 3   na 0.04  5.83 0.67 34.38 14.24  0.17  0.00  2.17  0.07 54.85  0.00  0.33
## 4   na 0.41  4.80 4.99 14.24  6.37  0.94  0.00  2.94  0.02 18.91  0.01  0.30
## 5    7 0.21  1.93 2.12 18.39  5.90  0.17  0.00  1.38  0.03 23.28  0.00  0.21
## 6   na 0.26  2.56 3.47  3.64  6.18  0.43  0.00  1.89  0.03 23.43  0.00  0.15
##   ABCA10
## 1   0.32
## 2   0.37
## 3   0.29
## 4   0.31
## 5   0.19
## 6   0.22
```

```
#heatmap
library(dplyr)
library(pheatmap)
library(stringr)


selected_genes <- c("AASS","ABCD1","ABI2", paste0("ABCA", 1:10))
gvar <- apply(gene_expression, MARGIN = 1, FUN = var)
gene_expression2 <- gene_expression[order(gvar, decreasing = TRUE), ]
log2.expr <- log2(gene_expression2 + 1)
mat_log <- log2.expr[selected_genes, , drop = FALSE]
mat_log <- log2.expr[selected_genes, 1:20, drop = FALSE]


ids <- colnames(mat_log)
idx <- match(ids, dataset$participant_id)

annotationData <- data.frame(
  row.names  = ids,
  ICU_Status = factor(str_to_title(trimws(dataset$icu_status[idx])),
                      levels = c("No","Yes")),
  Sex        = factor(str_to_title(trimws(dataset$sex[idx])),
                      levels = c("Female","Male","Unknown"))
)

annotationColors <- list(
  ICU_Status = c("No" = "gold", "Yes" = "purple"),
  Sex        = c("Female" = "lightblue", "Male" = "darkgreen", "Unknown" = "grey")
)

pheatmap(mat_log,
         cluster_rows = TRUE, cluster_cols = TRUE,
         clustering_distance_rows = "euclidean",
         clustering_distance_cols = "euclidean",
         annotation_colors = annotationColors,
         annotation_col = annotationData,
         main = "Heatmap (First 20 Samples)",
         border_color = "black",
         show_colnames = TRUE,
         angle_col = 45,
         fontsize_col = 6,
         fontsize_row = 8,
         cellwidth = 10,
         cellheight = 10)
```


Heatmap (First 20 Samples)
Sex
ICU_Status
ABCA7
ABCA1
ABCA2
ABCA4
ABCA6
ABCA8
ABCA9
ABCA10
AASS
ABCA3
ABCD1
ABI2
ABCA5
5
4
3
2
1
0
Sex
Female
Male
Unknown
ICU_Status
No
Yes
COVID_09_64y_female_ICU
COVID_06_:y_male_NonICU
COVID_08_78y_male_ICU
COVID_01_39y_male_NonICU
COVID_03_33y_male_NonICU
COVID_13_37y_male_NonICU
COVID_19_51y_male_NonICU
COVID_11_52y_female_NonICU
COVID_07_38y_female_NonICU
COVID_16_48y_male_NonICU
COVID_10_62y_male_ICU
COVID_05_49y_male_NonICU
COVID_20_62y_male_ICU
COVID_14_55y_male_ICU
COVID_04_49y_male_NonICU
COVID_17_54y_male_NonICU
COVID_15_68y_male_ICU
COVID_12_50y_male_ICU
COVID_02_63y_male_NonICU
COVID_18_70y_female_NonICU