Final Project

Eric Chen

2024-07-30

#1

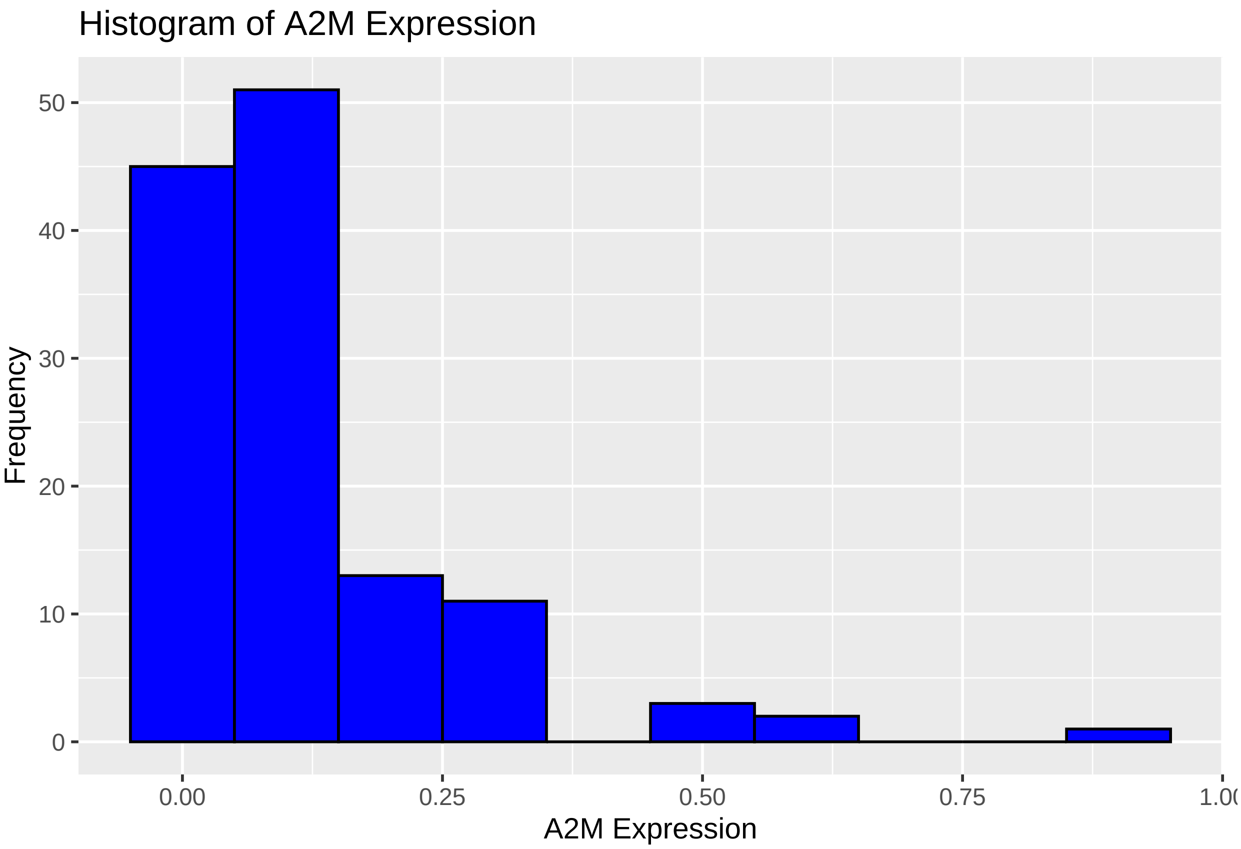
```r
# Load necessary libraries
library(tidyverse)
```

```
## -- Attaching core tidyverse packages ------------------------ tidyverse 2.0.0 --
## v dplyr     1.1.4     v readr     2.1.4
## v forcats   1.0.0     v stringr   1.5.0
## v ggplot2   3.4.4     v tibble    3.2.1
## v lubridate 1.9.3     v tidyr     1.3.0
## v purrr     1.0.1
## -- 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 error
```

```r
library(tidyr)

# Read the data
metadata <- read.csv("QBS103_GSE157103_series_matrix.csv")
gene_expression <- read.csv("QBS103_GSE157103_genes.csv")

# Clean whitespace in metadata
metadata <- metadata %>%
  mutate(across(everything(), ~ trimws(.)))

# Convert gene expression data to long format
gene_expression_long <- gene_expression %>%
  pivot_longer(cols = -X, names_to = "Sample", values_to = "Expression") %>%
  rename(Gene = X)

# Merge the data
merged_data <- left_join(gene_expression_long, metadata, by = c("Sample" = "participant_id"))

# Select a gene for analysis
selected_gene <- "A2M"
plot_data <- merged_data %>% filter(Gene == selected_gene)

# Convert continuous covariate (age) to numeric, handle non-numeric values
plot_data$age <- as.numeric(plot_data$age)
```

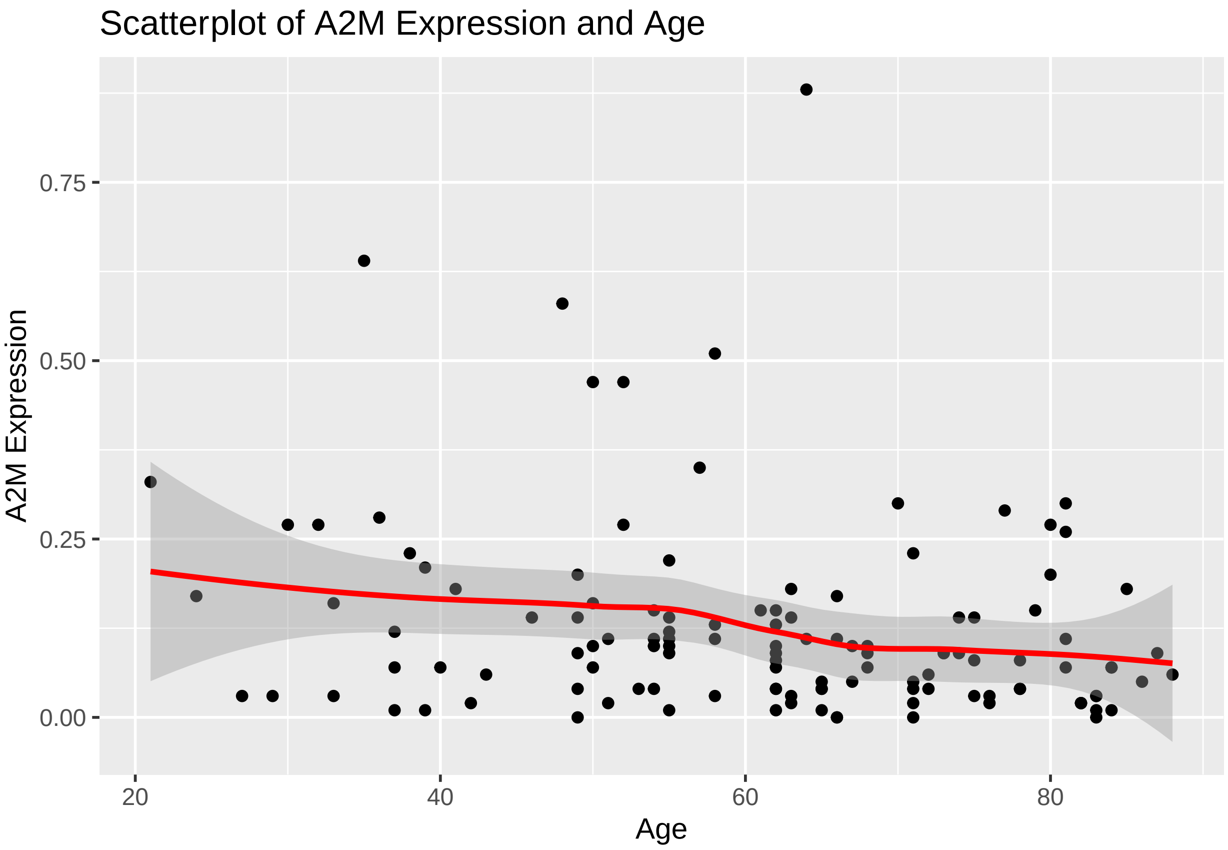
```
## Warning: NAs introduced by coercion
```

```
# Generate a histogram for gene expression
ggplot(plot_data, aes(x = Expression)) +
  geom_histogram(binwidth = 0.1, fill = "blue", color = "black") +  # Adjust binwidth
  labs(title = paste("Histogram of", selected_gene, "Expression"), x = paste(selected_gene, "Expression
```

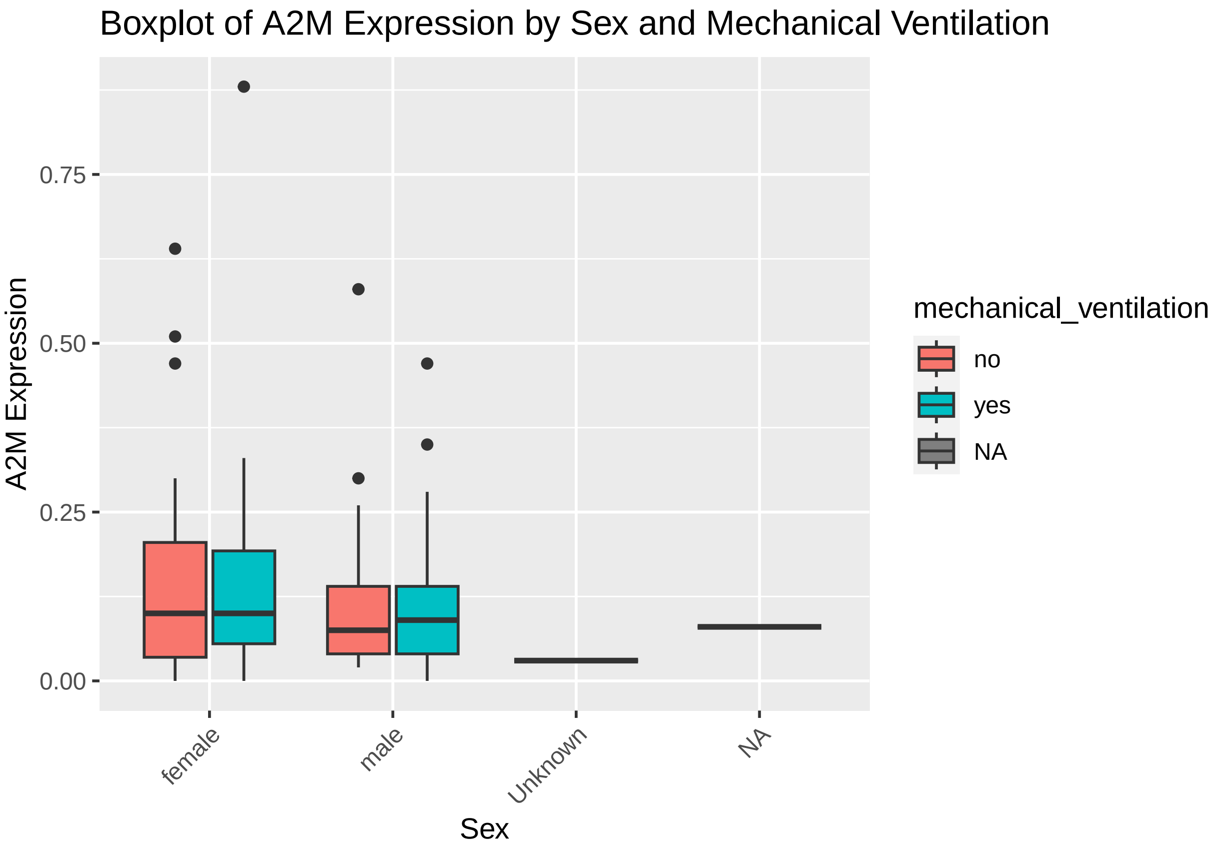


```
# Generate a scatter plot for gene expression and continuous covariate (age)
# Remove NA values from age
plot_data_scatter <- plot_data %>% filter(!is.na(age))

ggplot(plot_data_scatter, aes(x = age, y = Expression)) +
  geom_point() +
  geom_smooth(method = "loess", color = "red") +  # Adding a smoothed line
  labs(title = paste("Scatterplot of", selected_gene, "Expression and Age"), x = "Age", y = paste(selec
```

```
## ‘geom_smooth()‘ using formula = ’y ~ x’
```



```
# Generate a boxplot of gene expression separated by both categorical covariates (sex and mechanical_ve
# Handle 'unknown' value in sex variable
plot_data <- plot_data %>% mutate(sex = ifelse(sex == "unknown", "Unknown", sex))

ggplot(plot_data, aes(x = sex, y = Expression, fill = mechanical_ventilation)) +
  geom_boxplot() +
  labs(title = paste("Boxplot of", selected_gene, "Expression by Sex and Mechanical Ventilation"), x =
  theme(axis.text.x = element_text(angle = 45, hjust = 1))  # Rotate x-axis labels for better readabili
```

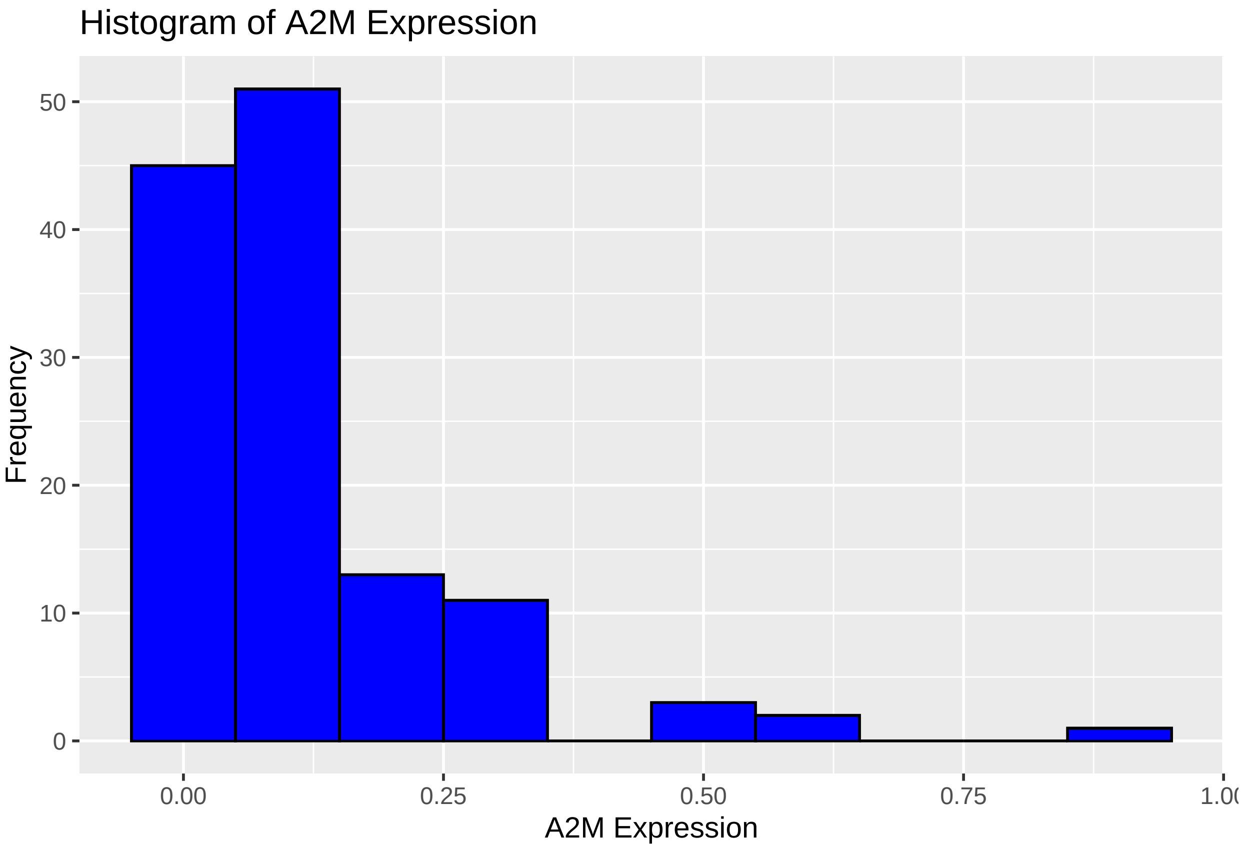


#2

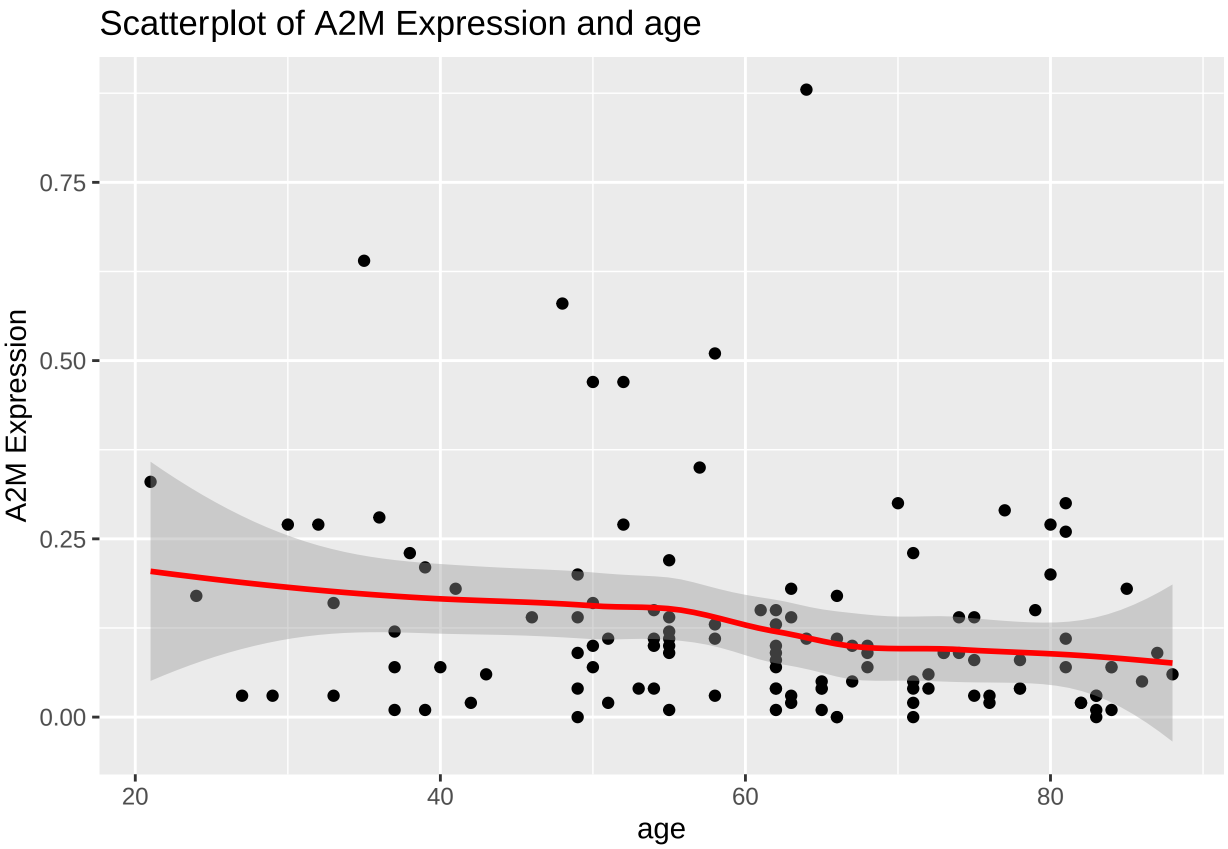
```
# Function to create plots
create_plots <- function(data, genes, continuous_covariate, categorical_covariate1, categorical_covaria
  for (gene in genes) {
    plot_data <- data %>% filter(Gene == gene)

    # Convert continuous covariate to numeric, handle non-numeric values
    plot_data[[continuous_covariate]] <- as.numeric(plot_data[[continuous_covariate]])

    # Histogram for gene expression
    print(
      ggplot(plot_data, aes(x = Expression)) +
        geom_histogram(binwidth = 0.1, fill = "blue", color = "black") +
        labs(title = paste("Histogram of", gene, "Expression"), x = paste(gene, "Expression"), y = "Fre
    )

    # Scatter plot for gene expression and continuous covariate
    plot_data_scatter <- plot_data %>% filter(!is.na(plot_data[[continuous_covariate]]))

    print(
      ggplot(plot_data_scatter, aes_string(x = continuous_covariate, y = "Expression")) +
        geom_point() +
        geom_smooth(method = "loess", color = "red") +
        labs(title = paste("Scatterplot of", gene, "Expression and", continuous_covariate), x = continu
    )

    # Boxplot of gene expression separated by both categorical covariates
    plot_data <- plot_data %>% mutate(!!categorical_covariate1 := ifelse(get(categorical_covariate1) ==
    print(
      ggplot(plot_data, aes_string(x = categorical_covariate1, y = "Expression", fill = categorical_cova
        geom_boxplot() +
        labs(title = paste("Boxplot of", gene, "Expression by", categorical_covariate1, "and", categori
        theme(axis.text.x = element_text(angle = 45, hjust = 1))
    )
  }
}

# Select additional genes
additional_genes <- c("A2M", "AARSD1", "ABHD2")

# Generate figures using the function
create_plots(data = merged_data, genes = additional_genes, continuous_covariate = "age", categorical_co
```

```
## Warning in create_plots(data = merged_data, genes = additional_genes,
## continuous_covariate = "age", : NAs introduced by coercion
```

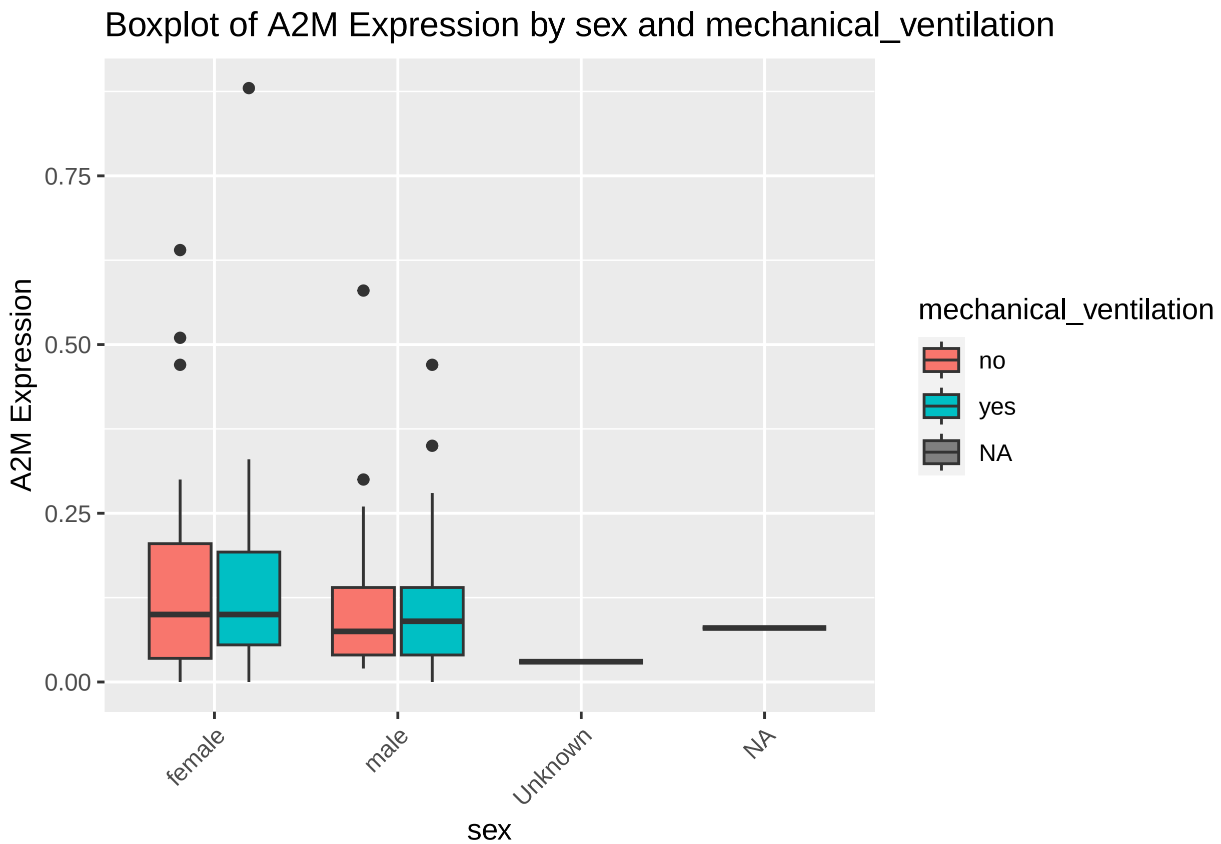
```
## Warning: ‘aes_string()‘ was deprecated in ggplot2 3.0.0.
## i Please use tidy evaluation idioms with ‘aes()‘.
## i See also ‘vignette("ggplot2-in-packages")‘ for more information.
## This warning is displayed once every 8 hours.
## Call ‘lifecycle::last_lifecycle_warnings()‘ to see where this warning was
## generated.
```

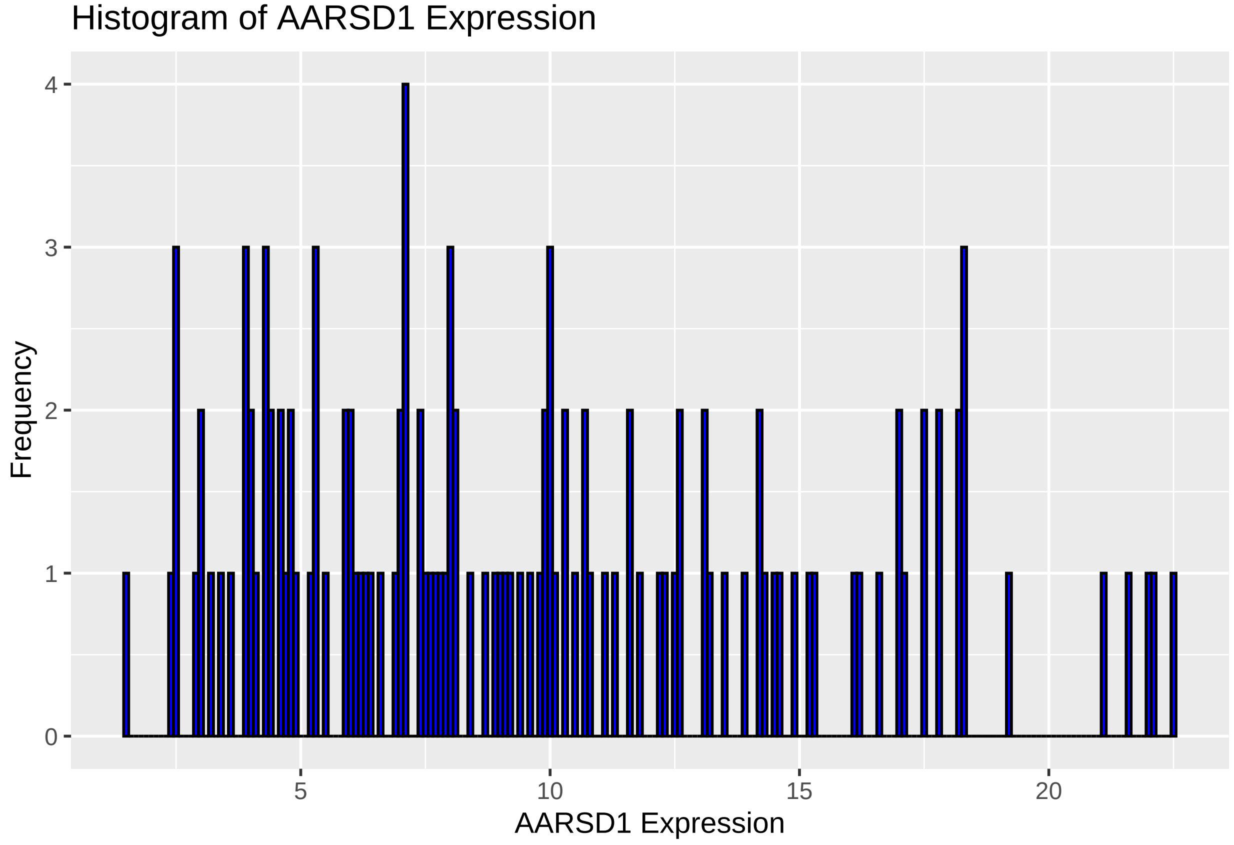


```
## 'geom_smooth()' using formula = 'y ~ x'
```

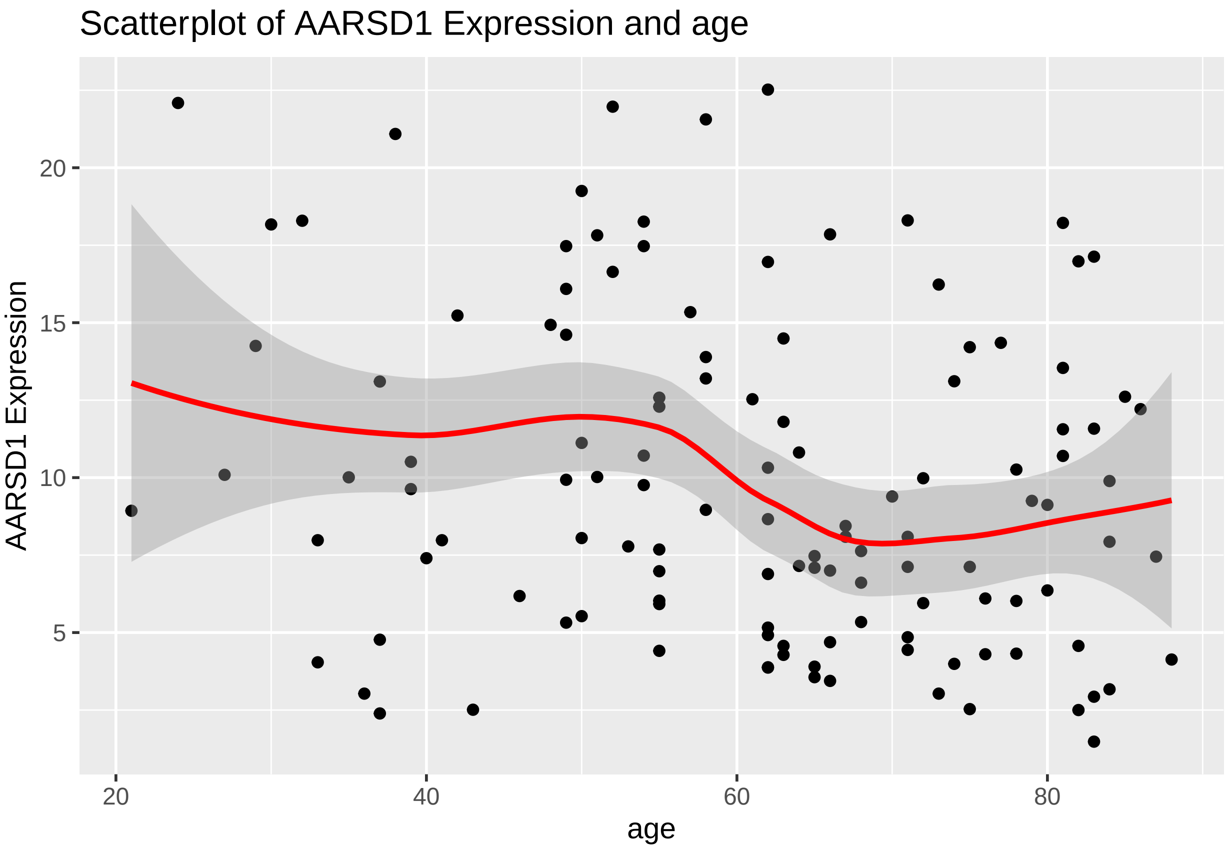


```
## Warning in create_plots(data = merged_data, genes = additional_genes,
## continuous_covariate = "age", : NAs introduced by coercion
```

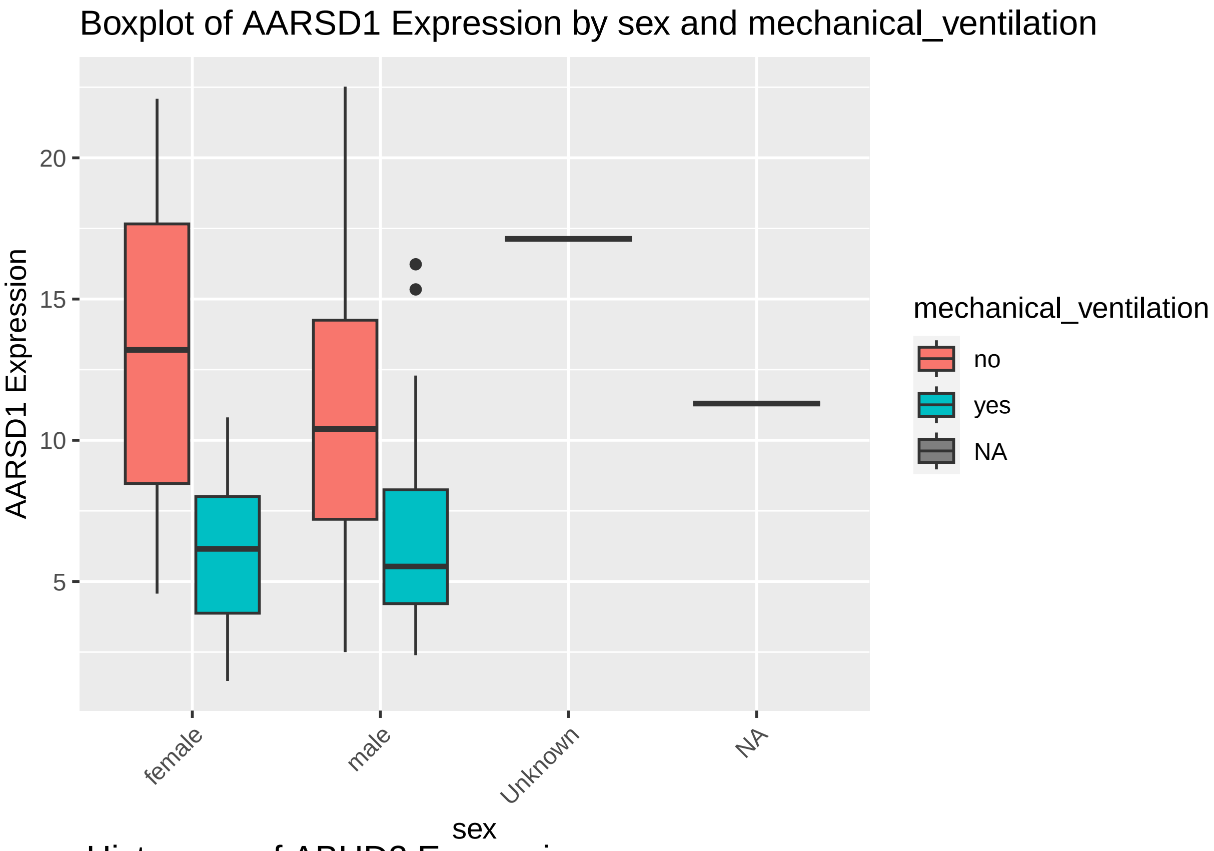

Boxplot of A2M Expression by sex and mechanical_ventilation
A2M Expression
0.75
0.50
0.25
0.00
female
male
Unknown
NA
sex
mechanical_ventilation
no
yes
NA

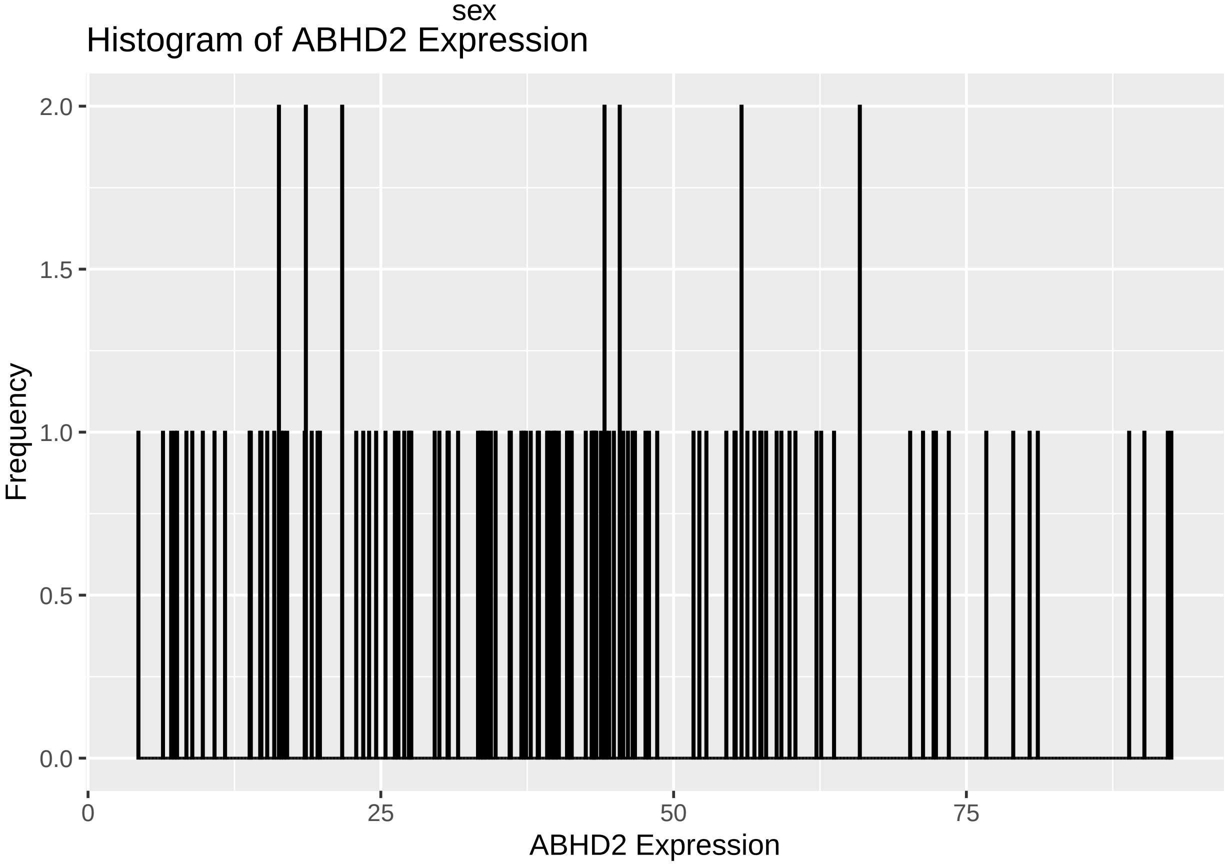

Histogram of AARSD1 Expression
Frequency
4
3
2
1
0
5
10
15
20
AARSD1 Expression

```
## `geom_smooth()` using formula = 'y ~ x'
```

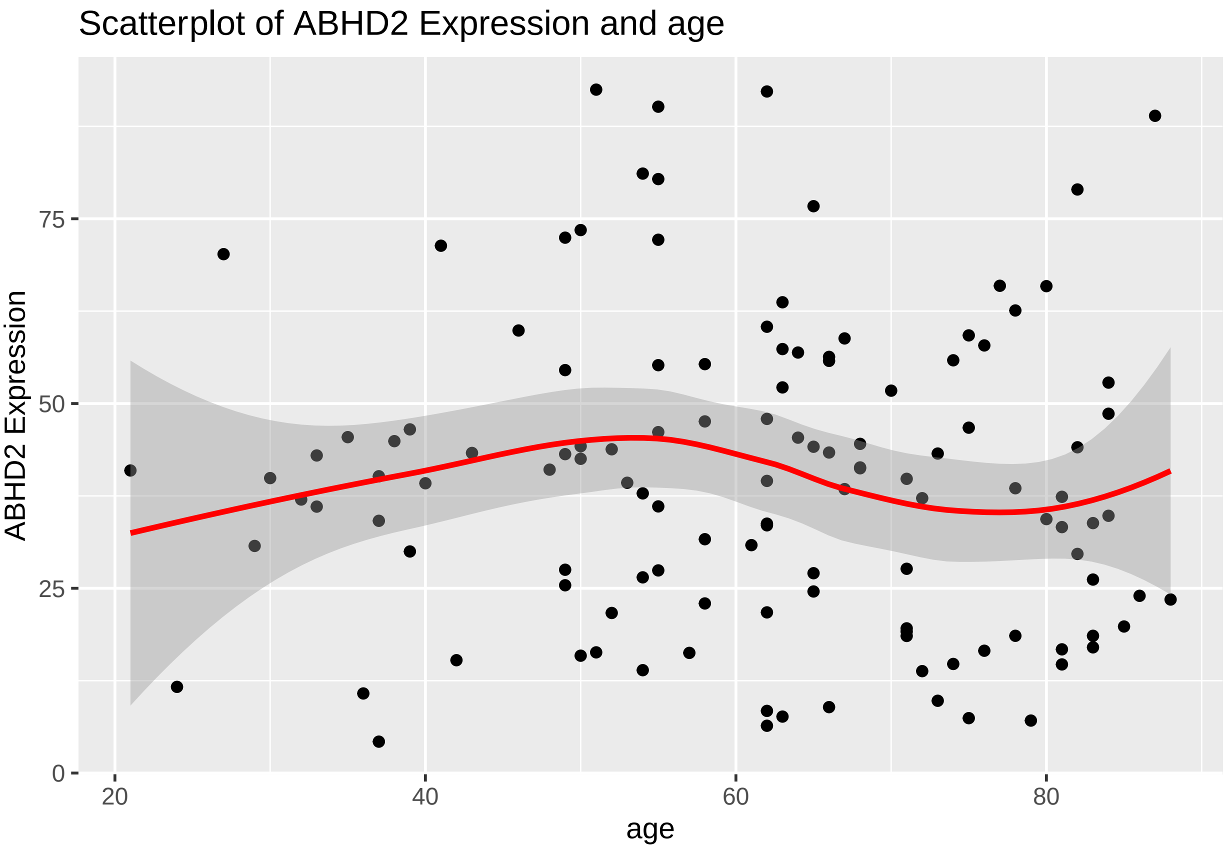


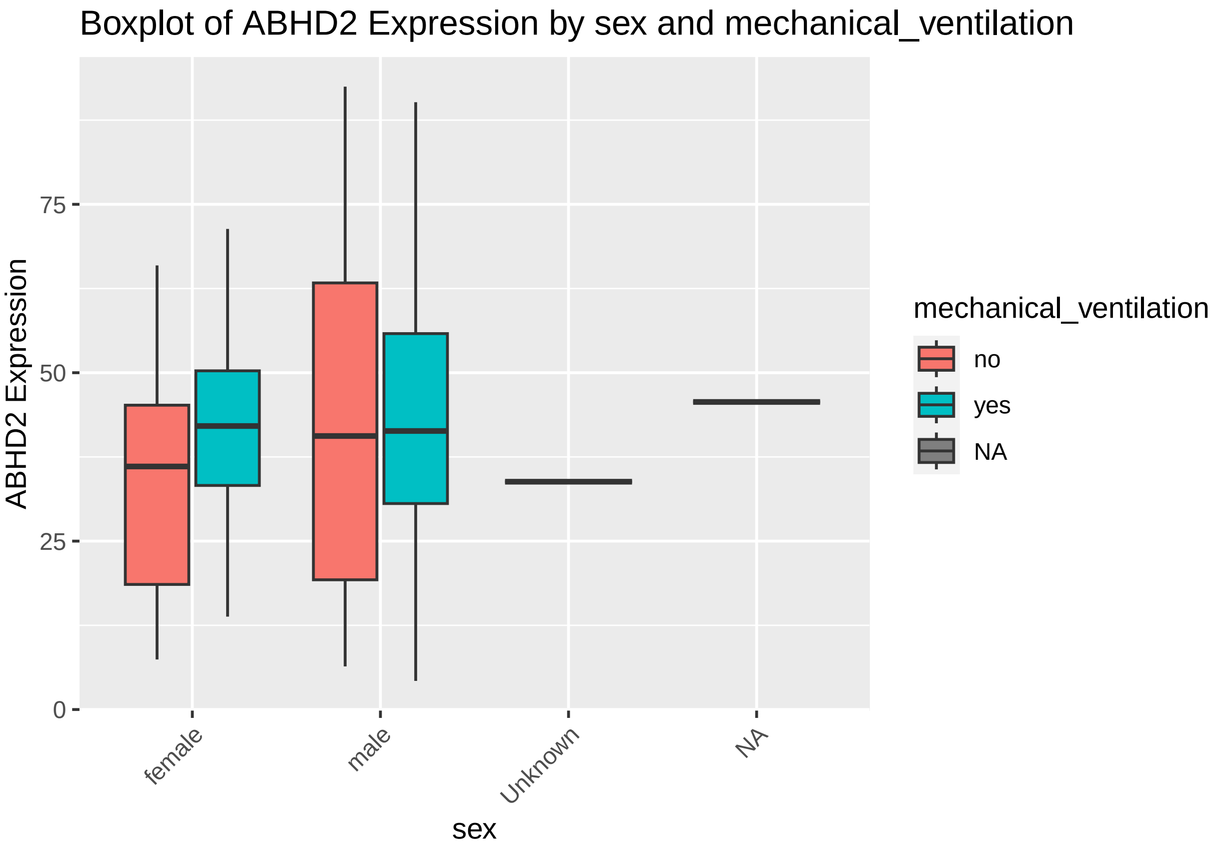
```
## Warning in create_plots(data = merged_data, genes = additional_genes,
## continuous_covariate = "age", : NAs introduced by coercion
```


Boxplot of AARSD1 Expression by sex and mechanical_ventilation
AARSD1 Expression
20
15
10
5
female
male
Unknown
NA
sex
mechanical_ventilation
no
yes
NA


Histogram of ABHD2 Expression
Frequency
2.0
1.5
1.0
0.5
0.0
0
25
50
75
ABHD2 Expression

```
## `geom_smooth()` using formula = 'y ~ x'
```




Boxplot of ABHD2 Expression by sex and mechanical_ventilation
ABHD2 Expression
0
25
50
75
female
male
Unknown
NA
sex
mechanical_ventilation
no
yes
NA