My_Graphs

2024-07-25

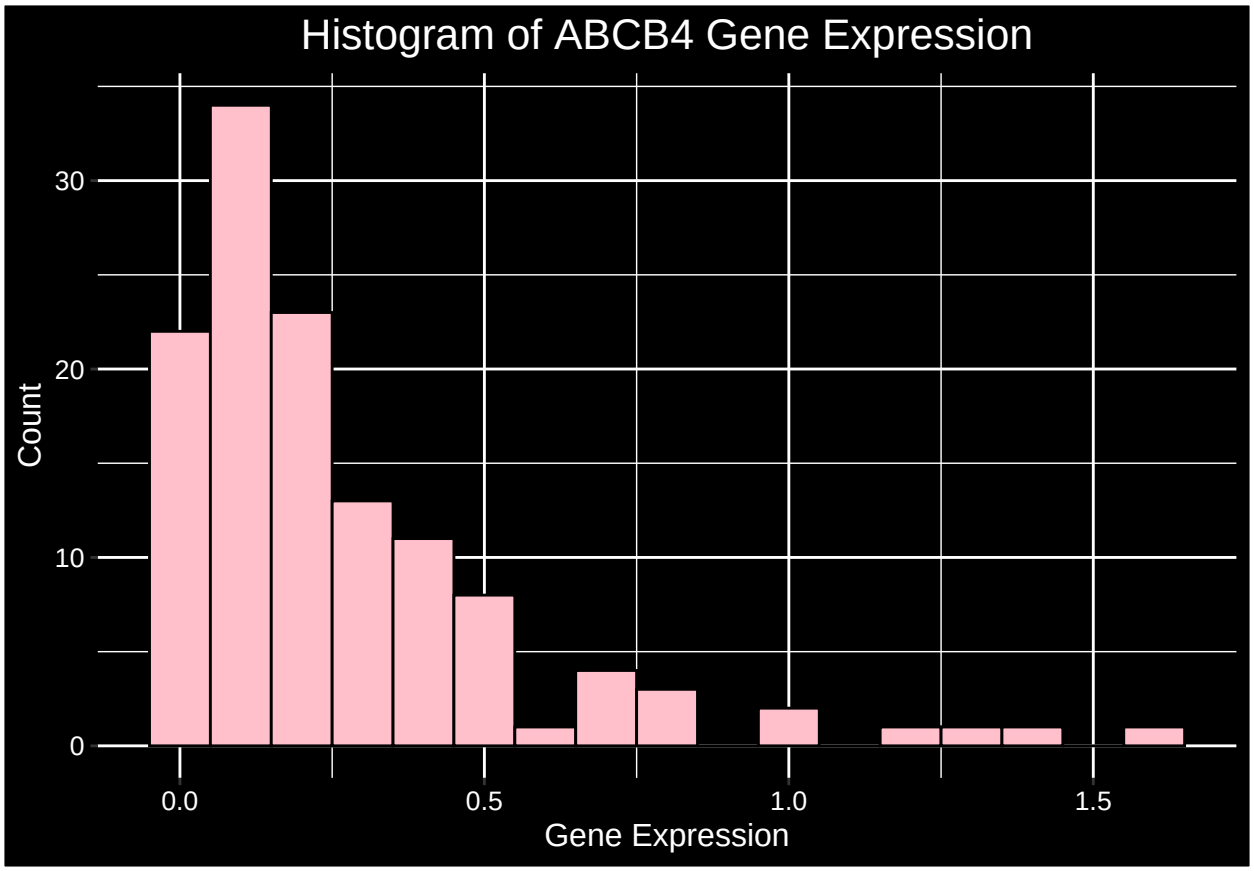
```
# Load packages and tidyverse includes ggplot2
library(tidyverse)
```

```
## -- 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   3.5.1     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
```

```
# Pulling Data from the two provided files.
# row.names = 1 to indicate that the first column of the CSV file is names for the data frame.
gene_express_df <- read.csv("C:/Users/casha/Downloads/QBS103_GSE157103_genes.csv", row.names = 1)
metadata_df <- read.csv("C:/Users/casha/Downloads/QBS103_GSE157103_series_matrix.csv")

# Selecting specific columns of interest
metadata_relevant <- metadata_df %>%
  select(participant_id, geo_accession, age, disease_status, sex)
#Continuous Covariant = age
#Categorical Covariants = disease_status, sex

# Extract expression data for the ABCB4 gene
abcb4_expression <- gene_express_df %>%
  rownames_to_column("gene") %>%
  filter(gene == "ABCB4") %>% #Keep ONLY ABCB4 rows

#wide format to long format
  pivot_longer(cols = -gene, names_to = "participant_id", values_to = "expression") %>%
  select(-gene)

# Merge expression data with metadata
merged_df <- abcb4_expression %>%
#https://www.datacamp.com/tutorial/merging-data-r: How to merge 2 datasets in R
  merge(metadata_relevant, by = "participant_id")

# Ensure correct data types
merged_df <- merged_df %>%
  mutate(age = as.numeric(age), #make sure age is numeric
         sex = factor(sex),
         disease_status = factor(disease_status))
```

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

```r
# Histogram of ABCB4 gene expression
ggplot(merged_df, aes(x = expression)) +
#https://www.datacamp.com/tutorial/make-histogram-basic-r
  geom_histogram(binwidth = 0.1, fill = "pink", color = "black") + #binwidth is how wide we want each b
#change labels
  labs(
    title = "Histogram of ABCB4 Gene Expression",
    x = "Gene Expression",
    y = "Count" #number which falls within THIS level of gene expression
  ) +
  theme(
    plot.background = element_rect(fill = "black"),
    panel.background = element_rect(fill = "black"),
    text = element_text(size = 12, color = "white"),
    plot.title = element_text(hjust = 0.5, size = 16, color = "white"),
    axis.text = element_text(color = "white") #changes the color of the numbers on gridlines
  )
```



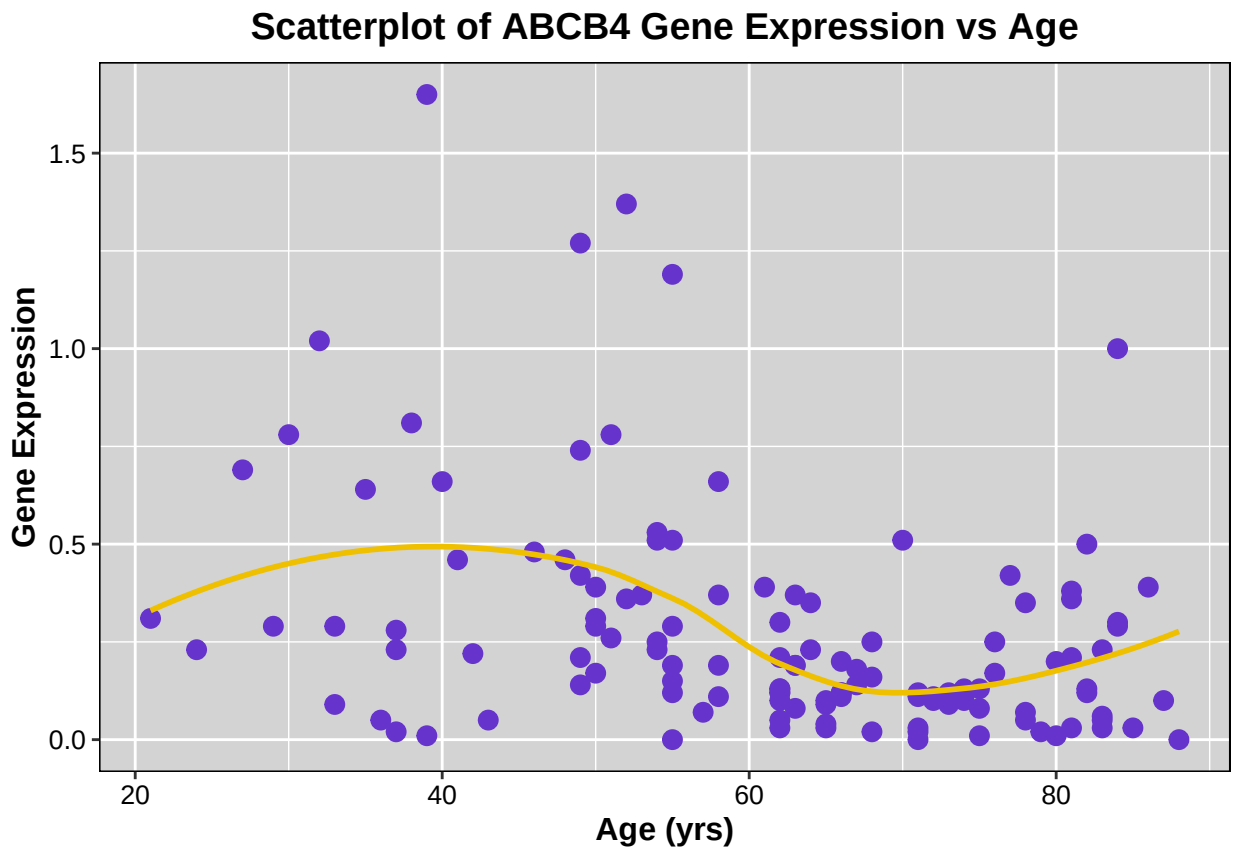

```r
# Scatterplot of ABCB4 expression vs. age
ggplot(merged_df, aes(x = age, y = expression)) +
  geom_point(color = "#6633CC", size = 3) +  # Adjusting color and size to see points better and I did
```

```
  geom_smooth(method = "loess", color = "#EFC000FF", se = FALSE) +  # Add a regression line https://sta
# (decided to go with a loess instead of lm since a loess line shows the higher average gene expression
  labs(
    title = "Scatterplot of ABCB4 Gene Expression vs Age",
    x = "Age (yrs)", #continuous covariate
    y = "Gene Expression"
  ) +
  theme(
    plot.background = element_rect(fill = "white"),
    panel.background = element_rect(fill = "lightgrey"),
    text = element_text(size = 12),
    plot.title = element_text(hjust = 0.5, face = "bold", size = 14),
    axis.title = element_text(face = "bold"),
    axis.text = element_text(color = "black"),
    panel.border = element_rect(color = "black", fill = NA)
#https://www.geeksforgeeks.org/add-panel-border-to-ggplot2-plot-in-r/
#don't fill border or no graph appears
#    panel.grid = element_line(color = "grey") : decided to remove as did not help but learned a new te
#https://r-charts.com/ggplot2/grid/
  )
```

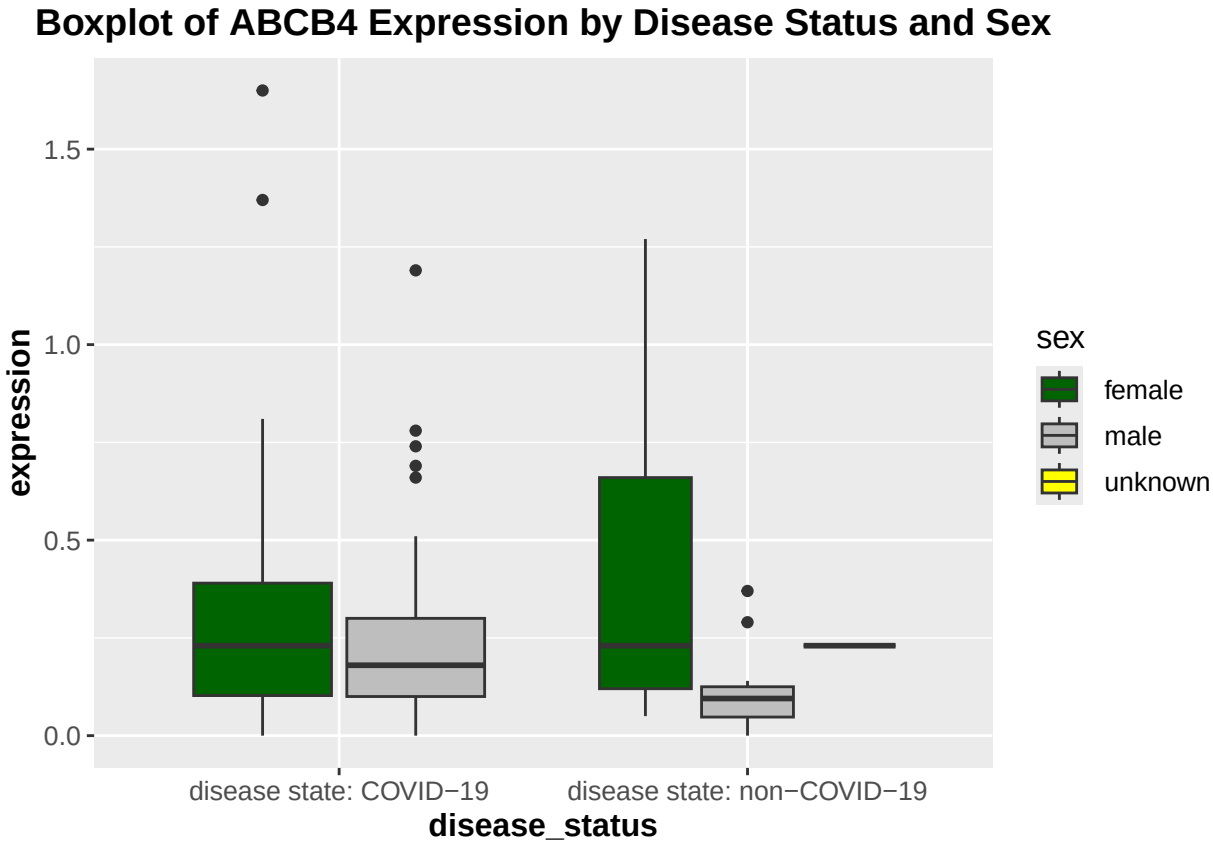
```
## `geom_smooth()` using formula = 'y ~ x'
```

```
## Warning: Removed 2 rows containing non-finite outside the scale range
## (`stat_smooth()`).
```

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




```
#Boxplot
ggplot(merged_df,aes(x = disease_status, y = expression, fill = sex)) +
#Add box plot
geom_boxplot() +
#Define colors: Have to use three colors because sex is female, male, and unknown
scale_fill_manual(values = c('darkgreen', 'grey', 'yellow')) +
    labs(
    title = "Boxplot of ABCB4 Expression by Disease Status and Sex"
  ) +
  theme(
    text = element_text(size = 12),
    plot.title = element_text(hjust = 0.5, face = "bold", size = 14),
    axis.title = element_text(face = "bold")
  )
```


Boxplot of ABCB4 Expression by Disease Status and Sex
expression
1.5
1.0
0.5
0.0
disease state: COVID−19
disease state: non−COVID−19
disease_status
sex
female
male
unknown