Class 12- Homework

Renee Zuhars (PID: A17329856)

Table of contents

Section 4: Population Scale Analysis

Question 13:

> Read this file into R and determine the sample size for each genotype and their corresponding median expression levels for each of these genotypes.

```
data <- "rs8067378_ENSG00000172057.6.txt"

df_data <- read.table(data) # reading the downloaded data into R as a table

head(df_data) # Viewing the dataset to see what the row/column names are
```

```
   sample geno      exp
1 HG00367  A/G 28.96038
2 NA20768  A/G 20.24449
3 HG00361  A/A 31.32628
4 HG00135  A/A 34.11169
5 NA18870  G/G 18.25141
6 NA11993  A/A 32.89721
```

```
nrow(df_data) # The number of rows in this dataset indicates how many samples there are.
```

```
[1] 462
```

```
table(df_data$geno) # Seeing the table for each genotype will indicate how much of the sample
```

```
A/A A/G G/G
108 233 121
```

```
library(dplyr) # Using tidyverse to sort first by genotype, and then by expression level- sho
```

```
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
```

```
df_data %>%
  group_by(geno) %>%
  summarise(median_exp = median(exp))
```

```
# A tibble: 3 x 2
  geno  median_exp
  <chr>      <dbl>
1 A/A         31.2
2 A/G         25.1
3 G/G         20.1
```

There are 462 total samples. Of these samples, 108 represent an A/A genotype, 233 represent an A/G genotype, and 121 represent a G/G genotype.

The median levels of expression are as follows: A/A: 31.24847
A/G: 25.06486
G/G: 20.07363

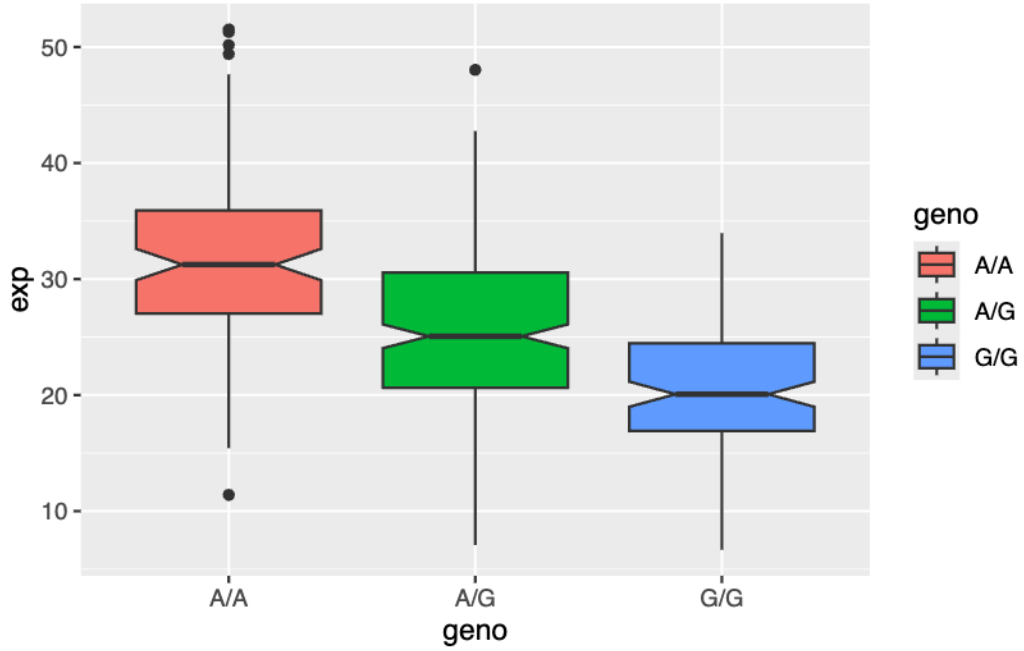
Question 14:

Generate a boxplot with a box per genotype, what could you infer from the relative expression value between A/A and G/G displayed in this plot? Does the SNP effect the expression of ORMDL3?

```
library(ggplot2)

# Using ggplot to create a box plot for the downloaded data.frame

ggplot(df_data) +
  aes(x = geno, y = exp, fill = geno) +
  geom_boxplot(notch = TRUE)
```




```
# Below recommended by chatGPT: a quick statistical test to see if the genotype at rs8067378

# Results are significant if the p value < 0.05 - indicated by this test as * (<0.05), ** (<

anova_result <- aov(exp ~ geno, data = df_data)
summary(anova_result)
```

```
             Df Sum Sq Mean Sq F value Pr(>F)
```

```
geno          2   7218    3609   75.66 <2e-16 ***
Residuals   459  21893      48
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

From the boxplot above, it is seen that A/A is expressed at significantly higher levels than G/G. We may be able to infer from this that G/G is the recessive and more rare genotype to have. We could also infer that an A/A genotype is more advantageous to have, as natural selection has selected for it in a larger population of individuals.

In accordance to a suggestion from ChatGPT, I ran a quick statistical test. Because the test returned a p-value of less than 0.001 (which is less than the usual threshold of <0.05), it can be assumed that having a different genotype at this location does affect expression of the gene ORMDL3.