Homework for Lab 12 - Thu 2.13

Jessica Le (PID: A17321021)

Section 4: Population Scale Analysis

One sample is obviously not enough to know what is happening in a population. You are interested in assessing genetic differences on a population scale. So, you processed about ~230 samples and did the normalization on a genome level. Now, you want to find whether there is any association of the 4 asthma-associated SNPs (rs8067378...) on ORMDL3 expression.

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

```
expr <- read.table("rs8067378_ENSG00000172057.6.txt")
head(expr)
```

```
   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
```

```
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 1.9.4     v tidyr     1.3.1
v purrr     1.0.4
-- 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 becor
```

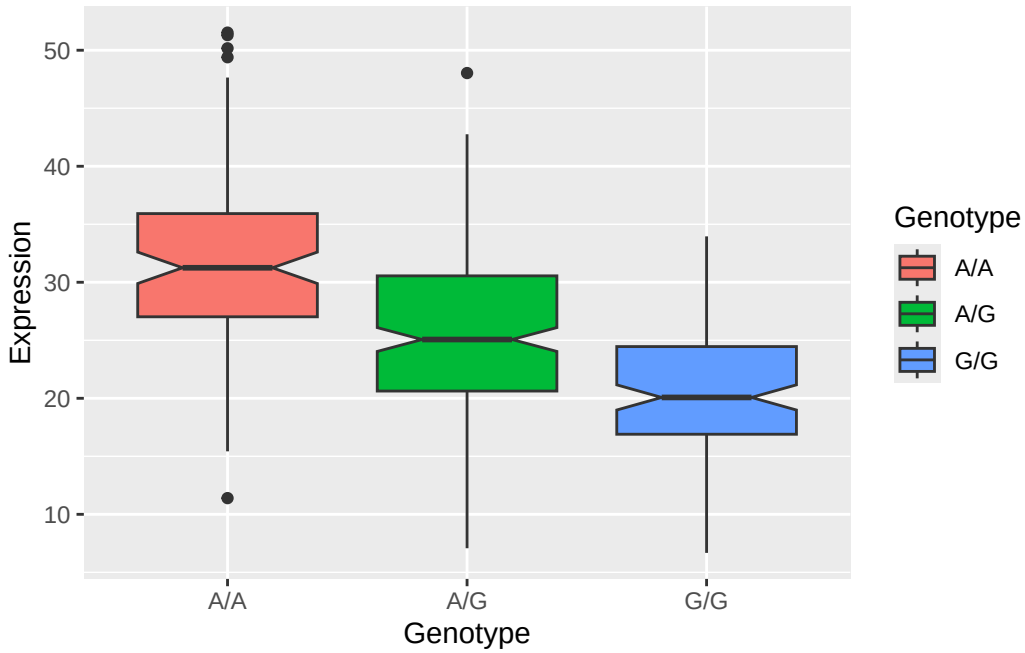
```
expr %>%
  group_by(geno) %>%
  summarize(median=median(exp))
```

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

The median expression levels (rounded to 2 decimal places) for the genotypes are the following: 31.25 for A/A, 25.06 for A/G, and 20.07 for G/G.

> Q14. 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)
ggplot(expr) + aes(x=geno, y=exp, fill=geno) +
  geom_boxplot(notch=TRUE,) +
  labs(x="Genotype", y="Expression", fill="Genotype")
```



From the plot, it can be inferred that the expression of homozygous A (A/A) is the most common for ORMDL3 expression followed by A/G, and homozygous G (G/G) is the least common. It can be inferred that SNP does effect expression of ORMDL3 because certain genotype can lead to relatively higher expression of ORMDL3 compared to others. In this case, it is shown that the genotype A/A expresses ORMDL3 more often than the G/G genotype.