Class 11: Population Analysis

Peter Shamasha (A15857589)

Population Scale Analysis

About ~230 samples have been procesed and and normalizized on a genome level. Now, we want to find whether there is any association between the 4 asthma-associated SNPs on ORMDL3 expression.

> Q13: Read this file into R and determine the sample size for each genotype and their corresponding median expression levels for each of these genotypes.

```r
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
```

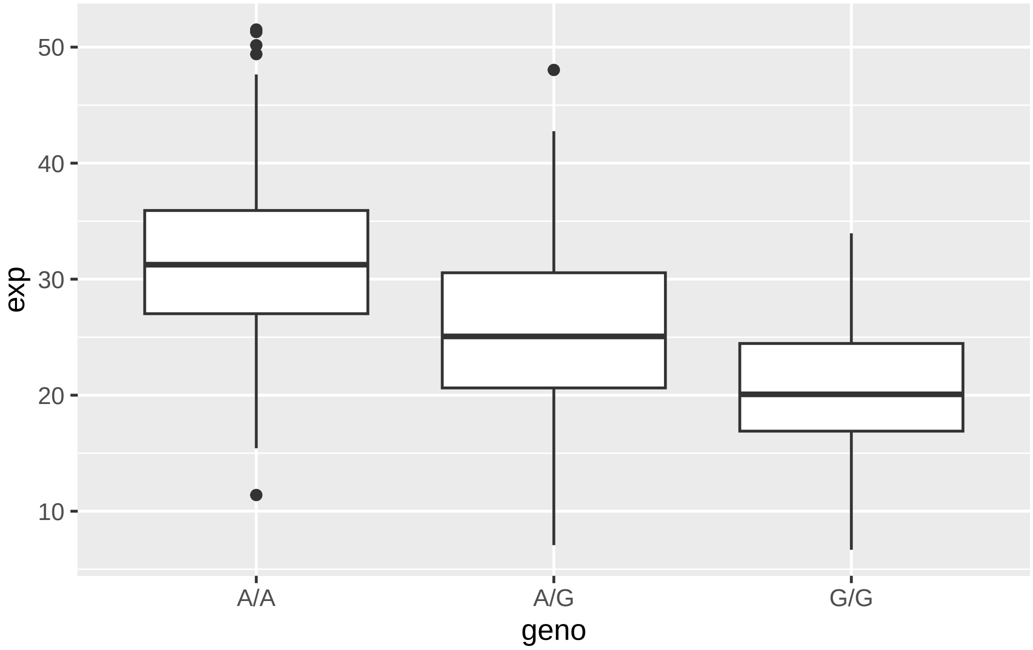
In order to determine the sample size for genotype, we can use the `table` function and the `$` syntax with the `geno` column in the table in order to table the samples sizes for each genotype.

```r
#Sample size of each genotype
table(expr$geno)
```

```

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

In order to find the median expression of each of the genotypes, we can make a boxplot using ggplot.

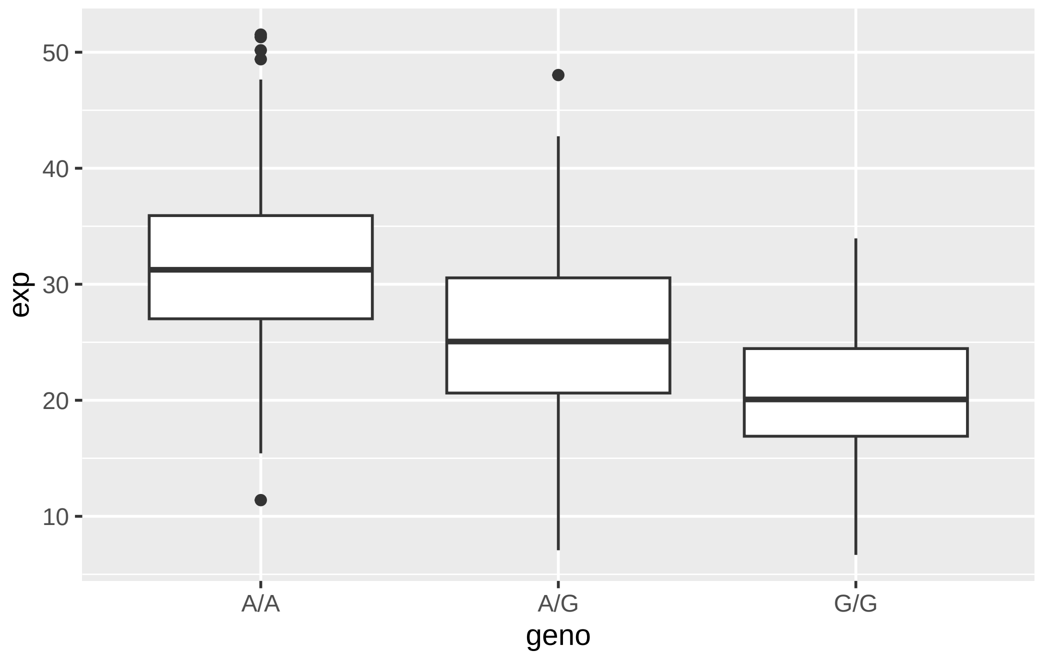
```
library(ggplot2)

ggplot(expr)+aes(x=geno, y=exp)+
  geom_boxplot()
```



> 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?

I will reuse the boxplot from the previous question

```
ggplot(expr)+aes(x=geno, y=exp)+
  geom_boxplot()
```



From this boxplot, we can infer that having the “A” alleles results in higher expression of ORMDL3