Class 11

Mady Welch

Section 4: Population Scale Analysis

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

```
txtfile <- "https://bioboot.github.io/bggn213_W19/class-material/rs8067378_ENSG00000172057
datatable <- read.table(txtfile)
head(datatable)
```

```
   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
```

```
summary(datatable)
```

```
    sample              geno                exp        
 Length:462         Length:462         Min.   : 6.675  
 Class :character   Class :character   1st Qu.:20.004  
 Mode  :character   Mode  :character   Median :25.116  
                                       Mean   :25.640  
                                       3rd Qu.:30.779  
                                       Max.   :51.518  
```

```
table(datatable$geno)
```

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

Genome Sample Sizes:

A/A = 108

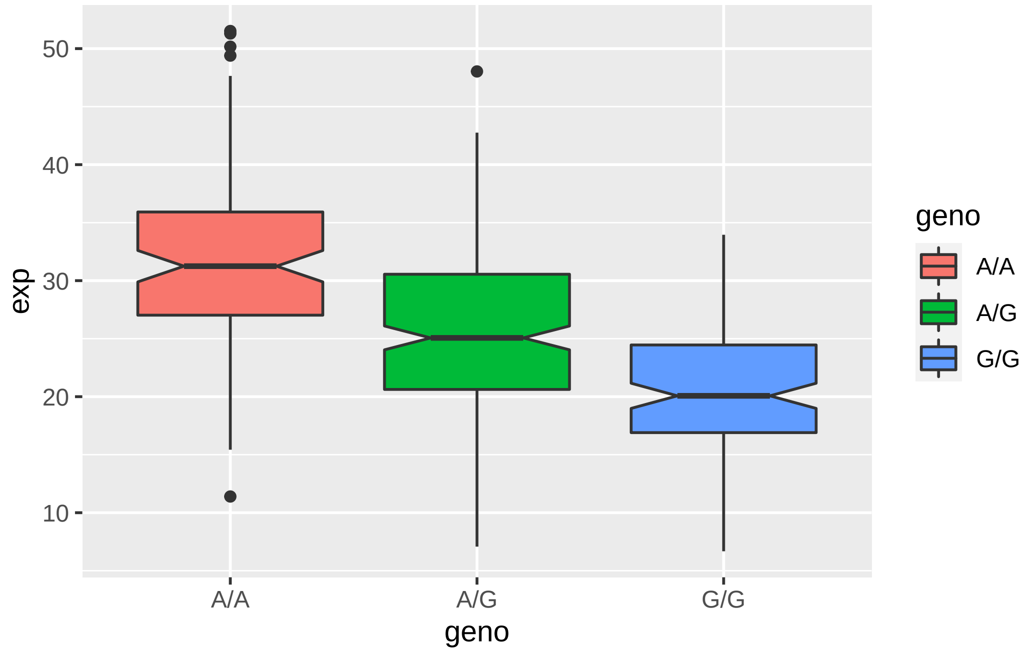
A/G = 233

G/G = 121

```r
library(ggplot2)

dataplot <- ggplot(datatable) +
  aes(x=geno, y=exp, fill=geno) +
  geom_boxplot(notch=TRUE)

dataplot
```



Median Expression Levels:

A/A ~ 32

A/G ~ 25

G/G ~ 20

> 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?

The median expression level for A/A is much higher than G/G, but the ranges for A/A and G/G overlap. We can most likely assume that the SNP does slightly affect the expression of ORMDL3.