Class12

Kyle Wittkop (A18592410)

Section 4: Population Scale Analysis

How many Samples do we have?

```
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
```

```
nrow(expr)
```

```
[1] 462
```

```
table(expr$geno)
```

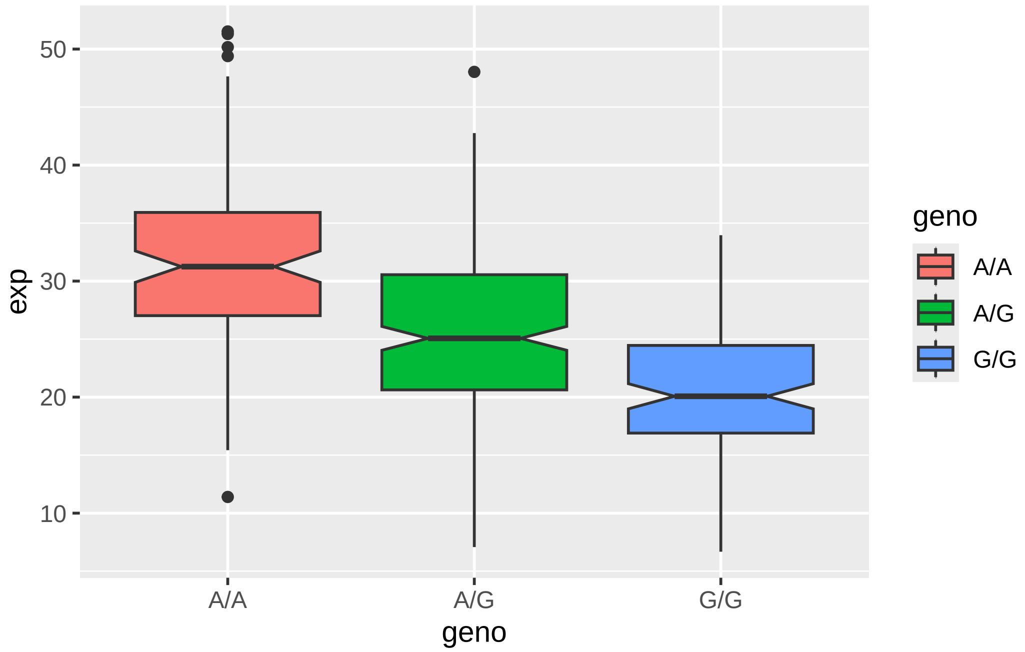
```

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

> 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

```
library(ggplot2)
```

Lets make a boxplot

```
ggplot(expr) + aes(geno, exp, fill=geno) +
  geom_boxplot(notch=TRUE)
```



> 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?

As seen in the box plot, individuals with the A/A genotype have the highest median expression, while those with the G/G genotype have the lowest median expression. This shows a trend where the presence of the A allele is associated with increased ORMDL3 expression, while the G allele is associated with lower expression. From this pattern, we can infer that the SNP does affect the expression of ORMDL3.