Class 11: Extra Credit

Aparajita Pranjal

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.

Reading in file:

```r
data_new <- read.table("datafile.txt")
head(data_new)
```

```
   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
```

Counting sample size of each genotype:

```r
table(data_new$geno)
```

```

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

Obtaining median expression levels of each genotype:

```r
AA <- subset(data_new, geno=="A/A")
AG <- subset(data_new, geno=="A/G" | geno=="G/A")
GG <- subset(data_new, geno=="G/G")
```

```
median(AA$exp)
```

```
[1] 31.24847
```

```
median(AG$exp)
```

```
[1] 25.06486
```

```
median(GG$exp)
```

```
[1] 20.07363
```

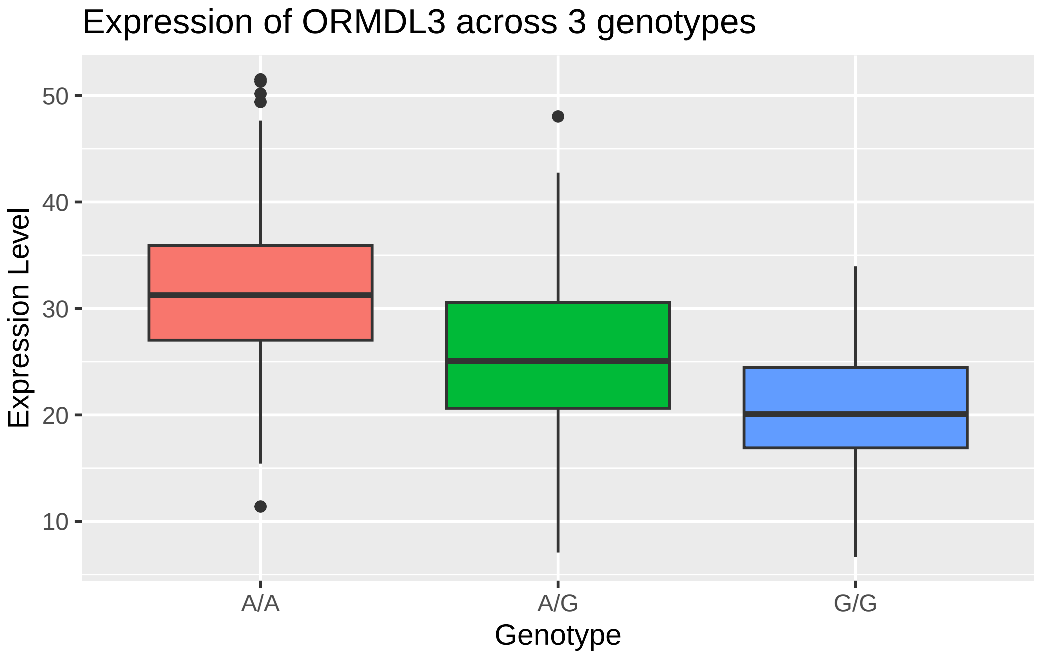
Q14: Generate a boxplot with a box per genotype, what could you infer from the relative expression value between A/A and G/G displayed in this plot? Does the SNP effect the expression of ORMDL3?

Generating boxplot using ggplot:

```
library(ggplot2)

p <- ggplot(data_new) +
  aes(data_new$geno, data_new$exp, fill = data_new$geno) +
  geom_boxplot()

p + ggtitle("Expression of ORMDL3 across 3 genotypes") +
    xlab("Genotype") +
    ylab("Expression Level") +
    theme(legend.position="none")
```



The SNP does affect the expression level of ORMDL3 as the median value for A/A is 31.2 and G/G is 20.1. A difference of about 11 is significant for gene expression.