

Class 11 pt2

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.

```
req <- read.table("https://bioboot.github.io/bimm143_F22/class-material/rs8067378_ENSG000000000000")
head(req)
```

| | 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 |

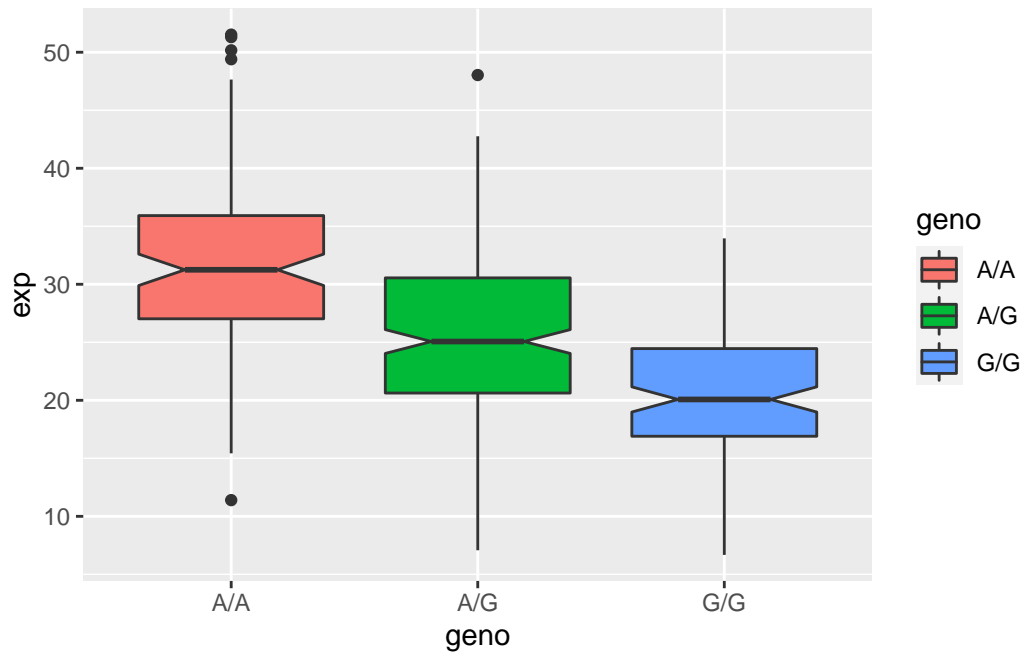
```
table(req$geno)
```

| A/A | A/G | G/G |
|-----|-----|-----|
| 108 | 233 | 121 |

The table function shows us the sample size for each phenotype in the dataset. The median value for A/A is 30, A/G is 25 and G/G is 20 as we will see in the boxplot generated in the next question.

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



While the boxplots do not overlap between A/A and G/G, their error bars do. This suggests that they may have different distributions that are somewhat similar. It is likely that the SNP affects the expression of ORMDL3.