Class 12 pt. 2: Population analysis

Morgan Black (PID A14904860)

Class 12 Section 4: Population scale analysis

```
expr <- read.table("rs8067378_ENSG00000172057.6.txt")
nrow(expr)
```

```
[1] 462
```

Q13: Determine the sample size for each genotype and their corresponding median expression levels for each of these genotypes

```
#How many data points are there for each genotype?
table(expr$geno)
```

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

```
#What's the median expression level for each of these genotypes?

#For the A/A genotype:
AA_geno <- subset(expr, geno == "A/A")
median(AA_geno$exp)
```

```
[1] 31.24847
```

```
#For the A/G genotype:
AG_geno <- subset(expr, geno == "A/G")
median(AG_geno$exp)
```
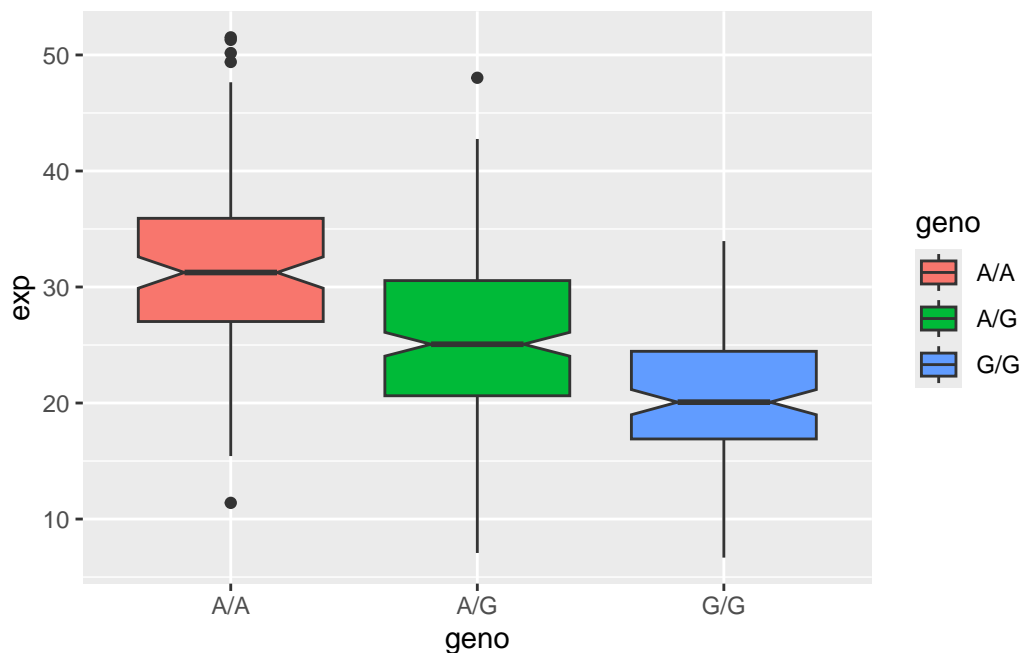
```
[1] 25.06486
```

```
#For the G/G genotype:
GG_geno <- subset(expr, geno == "G/G")
median(GG_geno$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?

```
library(ggplot2)
ggplot(expr) + aes(geno, exp, fill=geno) +
  geom_boxplot(notch=TRUE)
```

The relative expression levels are much lower in the G/G genotype than in the A/A genotype, showing that the G/G SNP clearly has an effect on ORMLD3 gene expression by leading to much lower expression in this population.