class12

Abraham Rachlin

Section 4: Population Scale Analysis

How many samples are there?

> 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.

```
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 
```

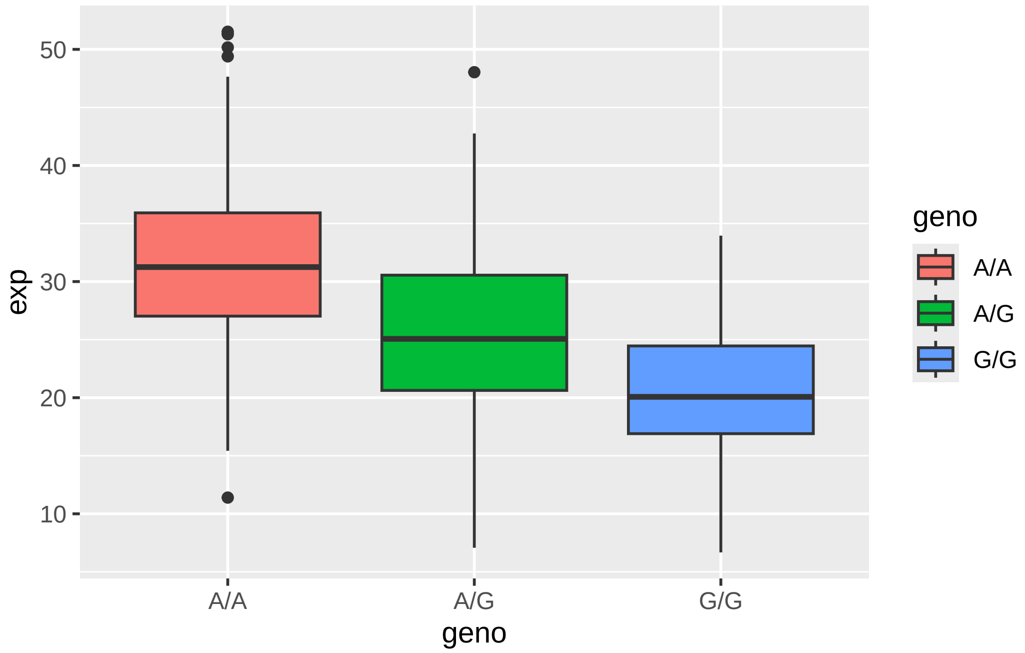
```r
library(ggplot2)
```

The median for A/A appears to be 31-32, the median for A/G appears to be 25, and the median for G/G appears to be around 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?

What you could infer about the expression value between A/A and G/G in the boxplot below is that A/A's expression valyue has a mean that is almost 10 values higher than that of G/G. G/G's maximum is not even A/A's minimum. Having a G/G does appeat to affect the expression of ORMDL3, as it is much lower than the others, therefore affecting its expression of the gene.

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


exp
A/A
A/G
G/G
geno
geno
A/A
A/G
G/G