Class 11

Challana Tea

Section 4: Population Scale Analysis

One sample is obviously not enough to know what is happening in a population. You are interested in assessing genetic differences on a population scale. So, you processed about ~230 samples and did the normalization on a genome level. Now, you want to find whether there is any association of the 4 asthma-associated SNPs (rs8067378...) on ORMDL3 expression.

> 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.

```
# Read data into table
exp_data <- read.table('rs8067378_ENSG00000172057.6.csv')
head(exp_data)
```

```
   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
```

```
# Sample sizes of each genotype
table(exp_data$geno)
```

```

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

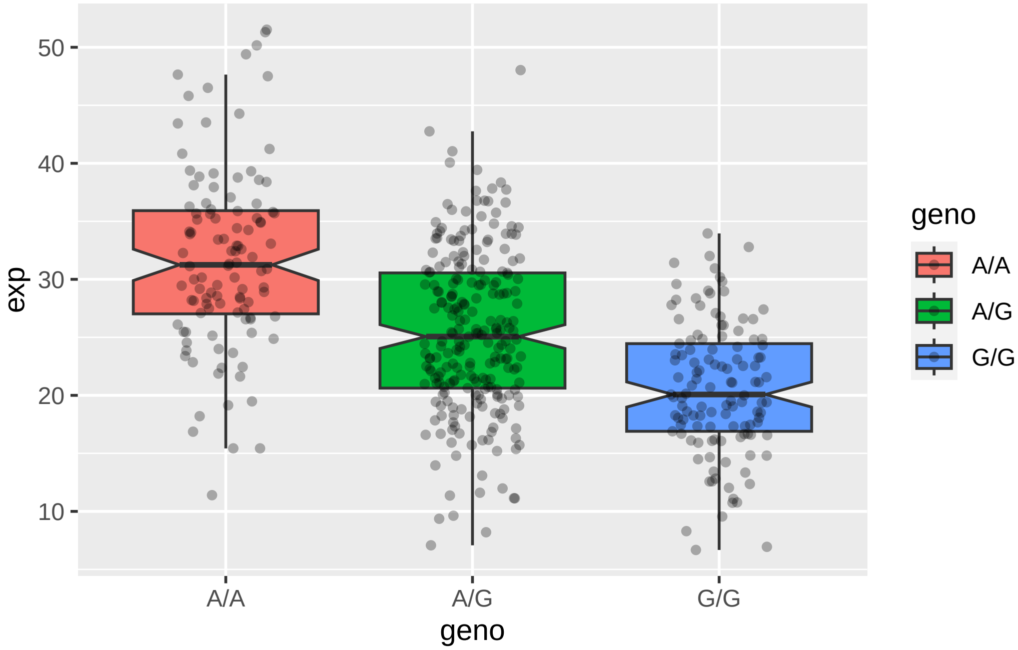
Find the median expression value of each genotype.

```
aggregate(exp_data$exp, list(exp_data$geno), median)
```

```
  Group.1        x
1     A/A 31.24847
2     A/G 25.06486
3     G/G 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)
p <- ggplot(exp_data) +
  aes(x=geno, y=exp, fill=geno) +
  geom_boxplot(notch=TRUE, outlier.shape=NA)
p + geom_jitter(shape=16, position=position_jitter(0.2), alpha = 0.3)
```



The SNP decreases overall expression of ORMDL3