Lab 12 Homework

Marina Puffer

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.

```
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
```

Sample size for each genotype:

```
table(expr$geno)
```

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

```
median(expr$exp[expr$geno=="A/A"])
```

```
[1] 31.24847
```

```
median(expr$exp[expr$geno=="A/G"])
```

```
[1] 25.06486
```

```
median(expr$exp[expr$geno=="G/G"])
```

```
[1] 20.07363
```

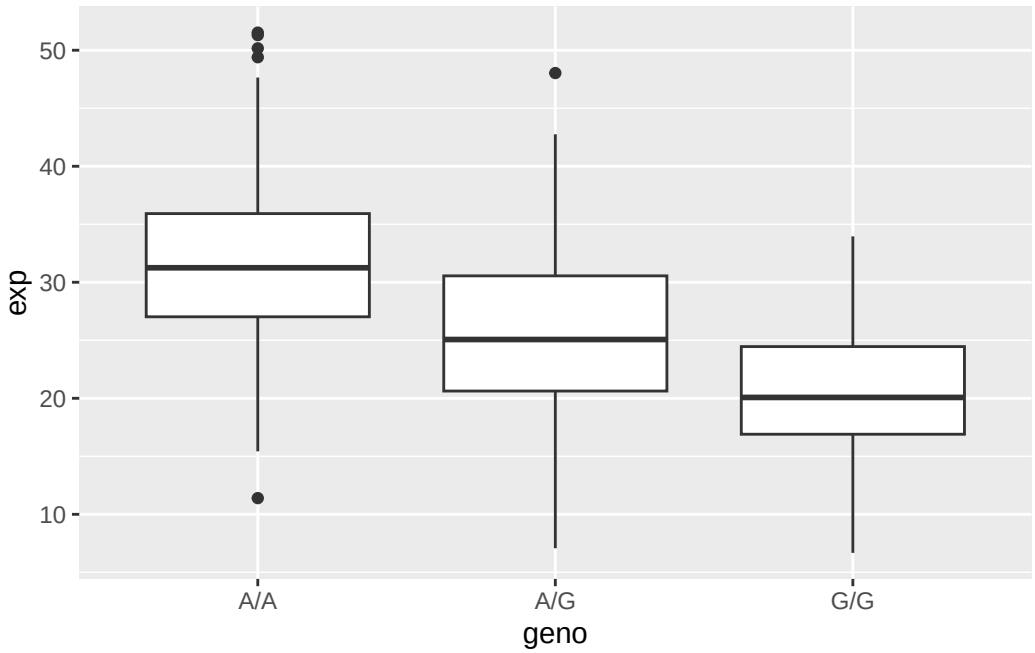
Median expression value for A/A is 31.25, A/G is 25.06, and G/G is 20.07.

```
library(ggplot2)
```

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? Make a boxplot

```
ggplot(expr)+aes(x=geno, y=exp)+geom_boxplot()
```



Median expression levels for the genotypes are shown above. The A/A allele is more highly expressed than G/G, so the G/G SNP likely decreases the expression of this gene.