Homework 12: Population Scale Analysis

Nicolò (PID: A18109144)

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 the file
pop <- read.table("rs8067378_ENSG00000172057.6.txt")
head(pop)
```

```
   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(pop)
```

```
[1] 462
```

Sample size for each genome

```
table(pop$geno)
```

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

A/A genotype median expression level

```
aa <- pop[which(pop$geno == "A/A"),]
head(aa)
```

```
    sample geno      exp
3  HG00361  A/A 31.32628
4  HG00135  A/A 34.11169
6  NA11993  A/A 32.89721
8  NA18498  A/A 47.64556
13 NA20585  A/A 30.71355
15 HG00235  A/A 25.44983
```

```
median(aa$exp)
```

```
[1] 31.24847
```

A/G genotype median expression level

```
ag <- pop[which(pop$geno == "A/G"),]
head(ag)
```

```
    sample geno      exp
1  HG00367  A/G 28.96038
2  NA20768  A/G 20.24449
7  HG00256  A/G 31.48736
10 HG00115  A/G 33.85374
11 NA20806  A/G 16.29854
12 HG00278  A/G 19.73450
```

```
median(ag$exp)
```

```
[1] 25.06486
```

G/G genotype median expression level

```
gg <- pop[which(pop$geno == "G/G"),]
head(gg)
```

```
    sample geno      exp
5  NA18870  G/G 18.25141
9  HG00327  G/G 17.67473
17 NA12546  G/G 18.55622
20 NA18488  G/G 23.10383
23 NA19214  G/G 30.94554
28 HG00112  G/G 21.14387
```

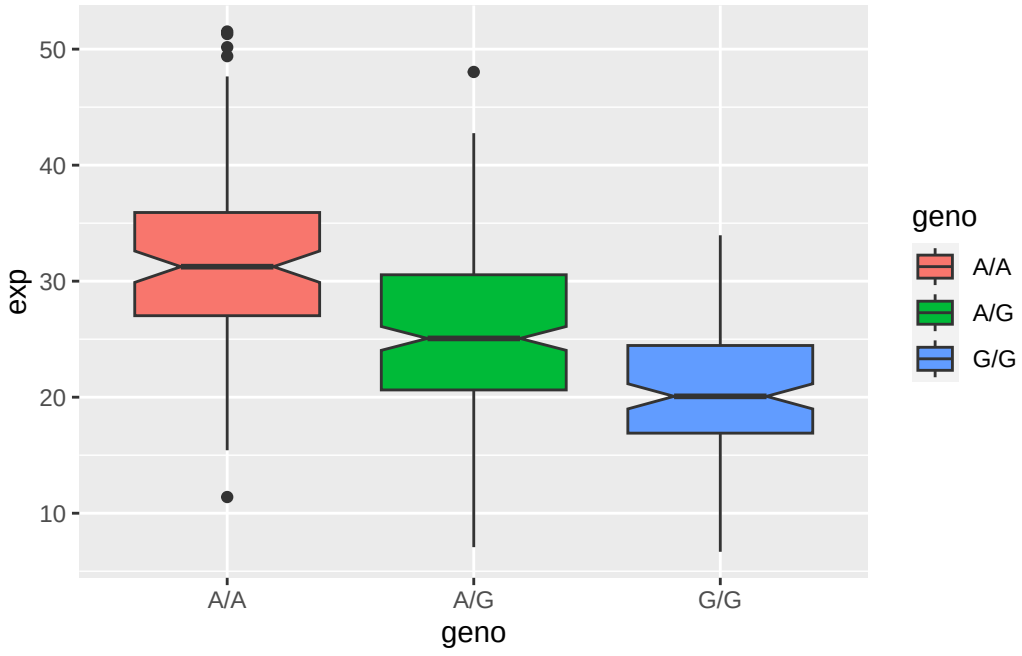
```
median(gg$exp)
```

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



By looking at the boxplot, it seems like the SNP does affect the ORMDL3 gene expression level, in particular the G allele seems to lower the expression level.