Population Scale Anlaysis [HW]

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.

This is the final file you got (https://bioboot.github.io/bggn213_W19/class-material/rs8067378_ENSG00000172057.6.txt) The first column is sample name, the second column is genotype and the third column are the expression values.

Open a new RMarkdown document in RStudio to answer the following two questions. Submit your resulting PDF report with your working code, output and narrative text answering Q13 and Q14 to GradeScope.

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
```

```
colnames(expr)
```

```
## [1] "sample" "geno"   "exp"
```

```
nrow(expr)
```

```
## [1] 462
```

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

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

There are 462 samples within the data. Within the dataset there are 108 A/A, 233 A/G, and 121 G/G

```
library(dplyr)
```

```
##
## Attaching package: 'dplyr'
```

```
## The following objects are masked from 'package:stats':
##
##     filter, lag
```

```
## The following objects are masked from 'package:base':
##
##     intersect, setdiff, setequal, union
```

Median expression levels

```
median_aa <- expr %>%
  filter(geno == 'A/A') %>%
  summarise(median_aa = median(exp)) %>%
  pull(median_aa)

median_aa
```

```
## [1] 31.24847
```

```
median_ag <- expr %>%
  filter(geno == 'A/G') %>%
  summarise(median_ag = median(exp)) %>%
  pull(median_ag)

median_ag
```

```
## [1] 25.06486
```

```
median_gg <- expr %>%
  filter(geno == 'G/G') %>%
  summarise(median_gg = median(exp)) %>%
  pull(median_gg)

median_aa
```

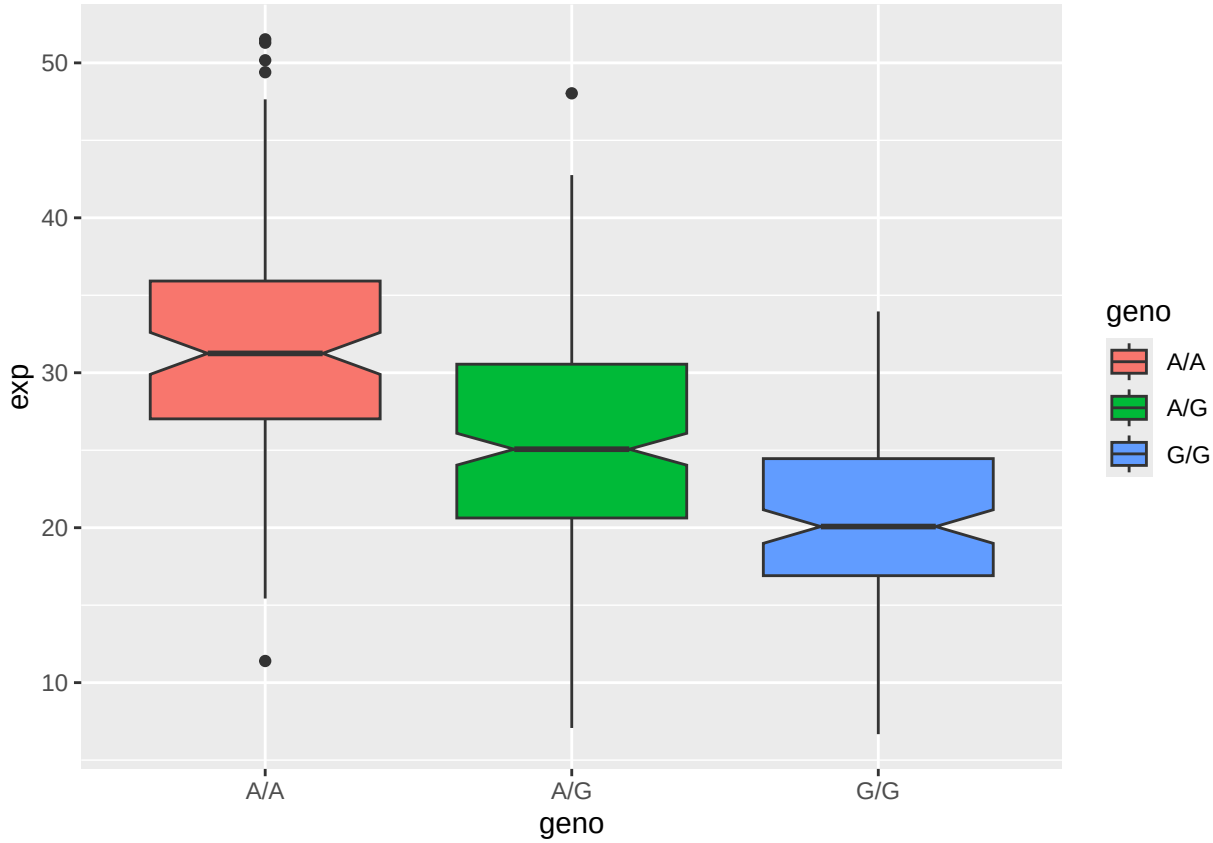
```
## [1] 31.24847
```

The median expression levels for A/A = 31.24847, A/G = 25.06486, and G/G = 31.24847

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(x = geno,
      y = exp,
      fill = geno) +
  geom_boxplot(notch=T)
```



Based on the boxplot we can see that when someone is homozygous for A (A/A), the expression levels are at a relatively high level as comapred to when someone is homoygous for G (G/G). Based on this inference we can tell that SNP does indeed effect the exoression of ORMLD3