# Lab 12 Genome Informatics HW Pop Analysis

Daniel Gurholt (PID: A16767491)

### Section 4: Population Scale Analysis [HOMEWORK]

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/classmaterial/rs8067378_ENSG00000172057.6.txt ). The first column is sample name, the second column is genotype and the third column are the expression values.

> 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.

#Reading in genotype data

```
genedat<- read.table("gene data.txt")
head(genedat)
```

```
   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
```

#Determining sample size for the samples

```
nrow(genedat)
```

```
[1] 462
```

There is a total of 462 gene samples in this given data set.

```
table(genedat$geno)
```

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

There are 108 gene samples that are homozygous A|A, 233 gene samples that are heterozygous A|G, and 121 samples homozygous for G|G.

#Determining medians for each of the 3 genotypes present in the dataset

```
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
```

```
AA<- genedat %>%
  filter(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
```

```
summary(AA)
```

```
    sample              geno                exp        
 Length:108         Length:108         Min.   :11.40  
 Class :character   Class :character   1st Qu.:27.02  
 Mode  :character   Mode  :character   Median :31.25  
                                       Mean   :31.82  
                                       3rd Qu.:35.92  
                                       Max.   :51.52  
```

The median expression levels for the A/A genotype is 31.25

```
library(dplyr)
AG<- genedat %>%
  filter(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
```

```
summary(AG)
```

```
    sample              geno                exp        
 Length:233         Length:233         Min.   : 7.075  
 Class :character   Class :character   1st Qu.:20.626  
 Mode  :character   Mode  :character   Median :25.065  
                                       Mean   :25.397  
                                       3rd Qu.:30.552  
                                       Max.   :48.034  
```

The median expression levels for the A/G genotype is 25.065

```
library(dplyr)
GG<- genedat %>%
  filter(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
```

```
summary(GG)
```

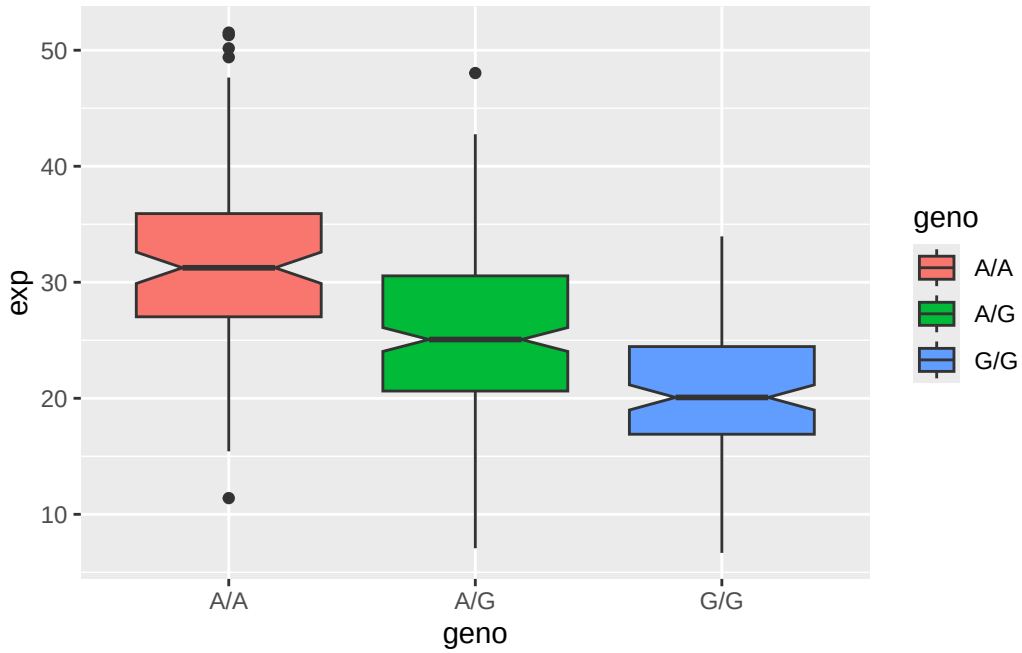
```
    sample              geno                exp
 Length:121         Length:121         Min.   : 6.675
 Class :character   Class :character   1st Qu.:16.903
 Mode  :character   Mode  :character   Median :20.074
                                       Mean   :20.594
                                       3rd Qu.:24.457
                                       Max.   :33.956
```

The median expression levels for the G/G genotype is 20.074

> 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)

Box<- ggplot(genedat) +
  aes(x=geno, y=exp, fill=geno) +
  geom_boxplot(notch=T)
Box
```




Looking at the resulting boxplot shows us that having the G/G genotype leads to an overall less expression of ORMDL3 than if the A/A genotype is present since expression levels are much higher in A/A compared to G/G. From this boxplot, I could conclude that the SNP does effect the expression of ORMDL3 depending on if a certain genotype is expressed or not.