

# HW 12: Section 4

Carly Chang (A16843962)

## 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. Hint: The `read.table()`, `summary()` and `boxplot()` functions will likely be useful here. There is an example R script online to be used ONLY if you are struggling in vein. Note that you can find the median value from saving the output of the `boxplot()` function to an R object and examining this object. There is also the `median()` and `summary()` function that you can use to check your understanding.

```
expr <- read.table("rs8067378_ENSG00000172057.6.txt") # get file
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) # number of rows (samples)
```

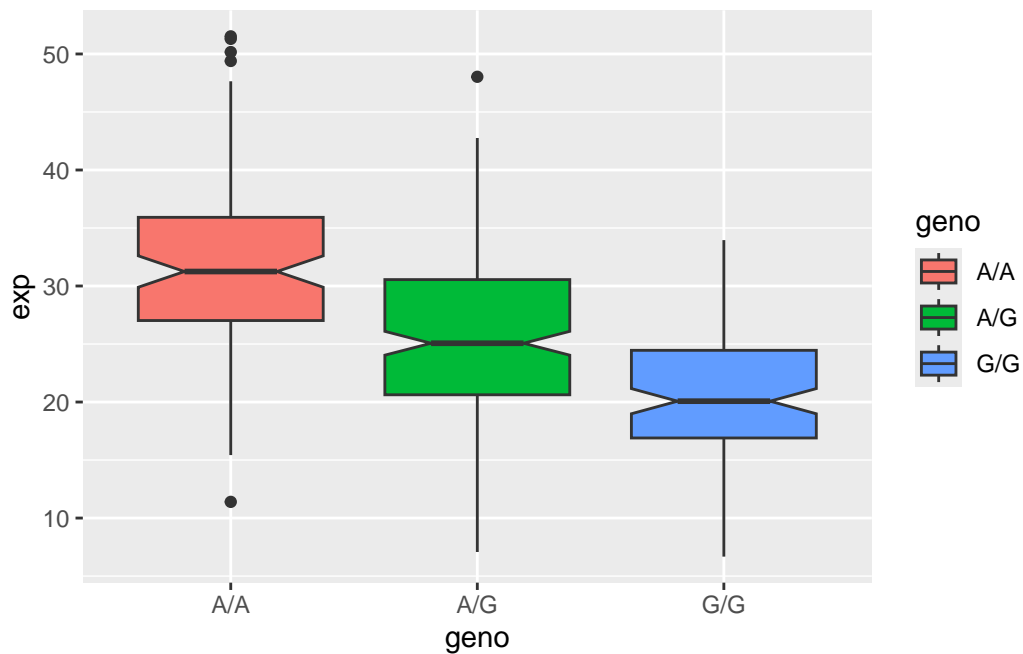
```
[1] 462
```

```
table(expr$geno) # number of each genotype
```

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

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? Hint: An example boxplot is provided overleaf – yours does not need to be as polished as this one.

```
library(ggplot2)  
ggplot(expr) +  
  aes(geno, exp, fill=geno) +  
  geom_boxplot(notch=T)
```



If notches don't overlap, then the difference is usually significant. G/G and A/A expression levels are quite different, indicating that the SNP affects the expression of ORMDL3.