class 12

Melanda Aboueid (PID: A17473102)

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.

> 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 medium value from saving the output of the boxplot() function to an R object and examining this object. There is also the medium() and summary() function that you can use to check your understanding.

```
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
```

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

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

```
tapply(expr$exp, expr$geno, median)
```

```
     A/A      A/G      G/G
31.24847 25.06486 20.07363
```

```
summary(expr)
```

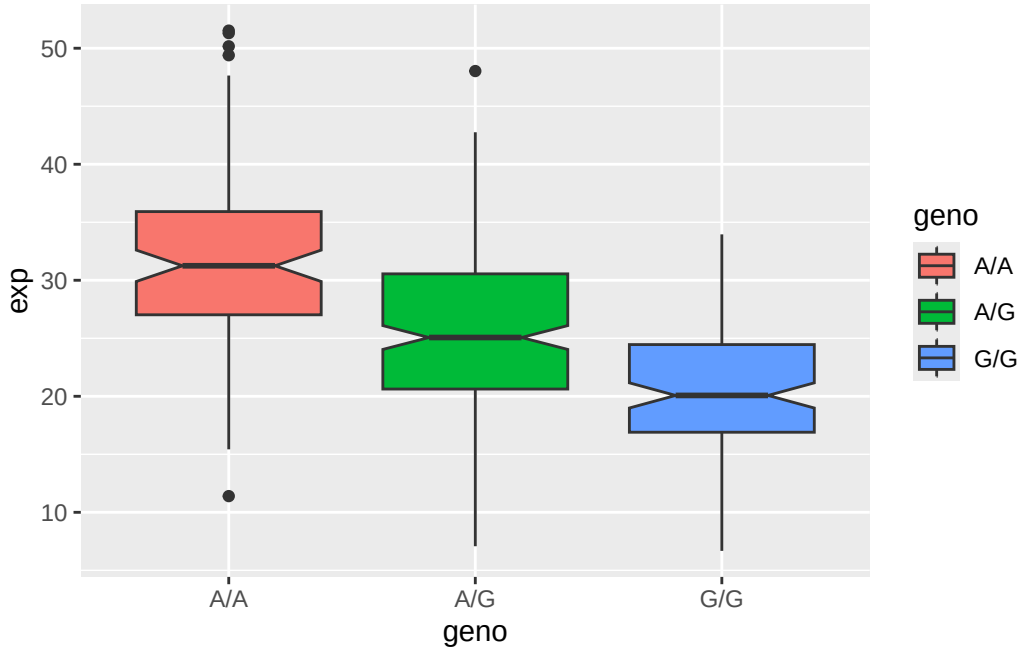
```
    sample              geno                exp
 Length:462         Length:462         Min.   : 6.675
 Class :character   Class :character   1st Qu.:20.004
 Mode  :character   Mode  :character   Median :25.116
                                       Mean   :25.640
                                       3rd Qu.:30.779
                                       Max.   :51.518
```

The dataset contained 462 samples total. The genotype counts were 108 A/A, 233 A/G, and 121 G/G. The median expression levels were 31.25 for A/A, 25.06 for A/G, and 20.07 for G/G. This shows that ORMDL3 expression is highest in A/A individuals and lowest in G/G individuals.

> 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 = TRUE)
```



The boxplot shows that individuals with the A/A genotype have the highest ORMDL3 expression, while G/G individuals have the lowest. This indicates that the rs8067378 SNP is associated with differences in ORMDL3 expression. Therefore, the SNP appears to affect gene expression.