Class 12 Lab

Derek Chang (PID: A16942232)

Population Scale:

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.

How many samples do we have:

Reading the Table:

```
expr <- read.table("DataClass12.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
```

```
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
```
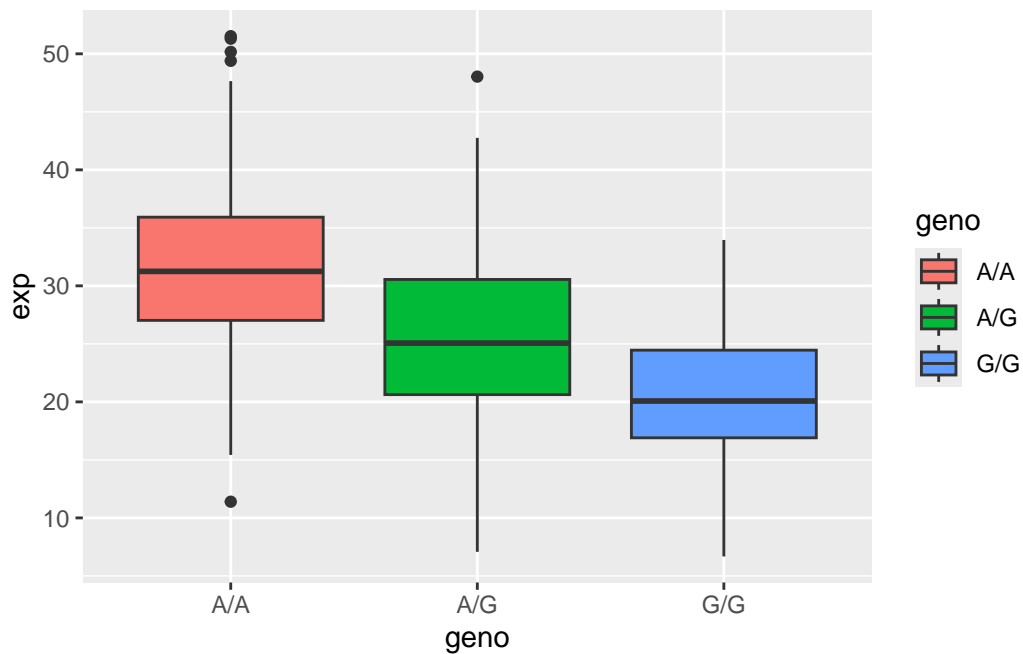
Reading the Sample size for each genotype:

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

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

Making a boxplot from expr data

```
library(ggplot2)
```

Making the boxplot using `ggplot`:

```
exprplot <- ggplot(expr, aes(geno, exp, fill = geno)) + geom_boxplot()
exprplot
```


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

```
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
```

```
medians <- expr %>% group_by(geno) %>% summarize(median_value = median(exp))
medians
```

```
# A tibble: 3 x 2
  geno   median_value
  <chr>         <dbl>
1 A/A            31.2
2 A/G            25.1
3 G/G            20.1
```

Given the table function for each genotype, We have 108 A/A, 233 A/G, and 121 G/G genotypes. The median for A/A is 31.24847, the median for A/G is 25.06486, and the median for G/G is 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?

Based on the boxplot, we can infer that A/A has a much higher relative expression value compared to G/G. Given this, it seems that the SNP does affect the expression levels of ORMDL3.