HW10 Q13-14

Sam Altshuler (PID: A59010373)

2/21/2022

Section 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. 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.
  - The sample size for the A/A genotype is n = 108, the sample size for G/A is n = 233, and for G/G is n = 121. The median expression levels for each genotype is 31.248, 25.064, and 20.074 (A/A, A/G, and G/G respectively).

```
library(dplyr)
```

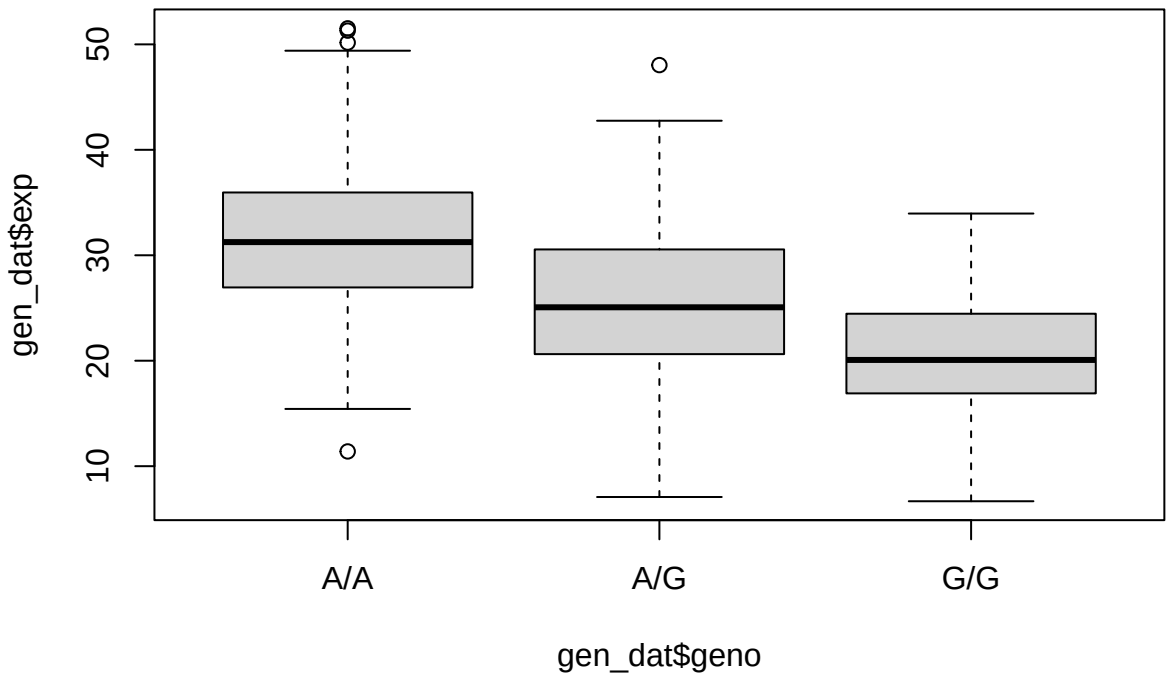
```
##
## Attaching package: 'dplyr'

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

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

```
gen_dat <- read.table(url("https://bioboot.github.io/bggn213_W19/class-material/rs8067378_ENSG0000017205

gen_box <- boxplot(gen_dat$exp ~ gen_dat$geno)
```




```
gen_box
```

```
## $stats
##          [,1]     [,2]     [,3]
## [1,] 15.42908  7.07505  6.67482
## [2,] 26.95022 20.62572 16.90256
## [3,] 31.24847 25.06486 20.07363
## [4,] 35.95503 30.55183 24.45672
## [5,] 49.39612 42.75662 33.95602
##
## $n
## [1] 108 233 121
##
## $conf
##          [,1]     [,2]     [,3]
## [1,] 29.87942 24.03742 18.98858
## [2,] 32.61753 26.09230 21.15868
##
## $out
## [1] 51.51787 50.16704 51.30170 11.39643 48.03410
##
## $group
## [1] 1 1 1 1 2
##
## $names
## [1] "A/A" "A/G" "G/G"
```

- 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?
  - From the boxplot below, it appears that the SNP does negatively affect the expression of ORMDL3. The G/G expression data is visually much lower than its A/A counterpart. This can also be seen in the answer above due to the differences in median expression levels.

```
library(ggplot2)

ggplot(gen_dat, aes(x= geno, y = exp)) +
  geom_boxplot()+
  labs(title = "Expression levels for ORMDL3 SNP")+
  xlab("Genotype") +
  ylab("Expression Level")
```