Class19 Homework (Population analysis)

Bangyan Hu (PID: A15540189)

11/7/2021

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.

```
expr <- read.table("https://bioboot.github.io/bggn213_W19/class-material/rs8067378_ENSG00000172057.6.tx
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
```

```
bp <- boxplot(exp ~ geno, data=expr,plot=F)
bp
```

```
## $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
```

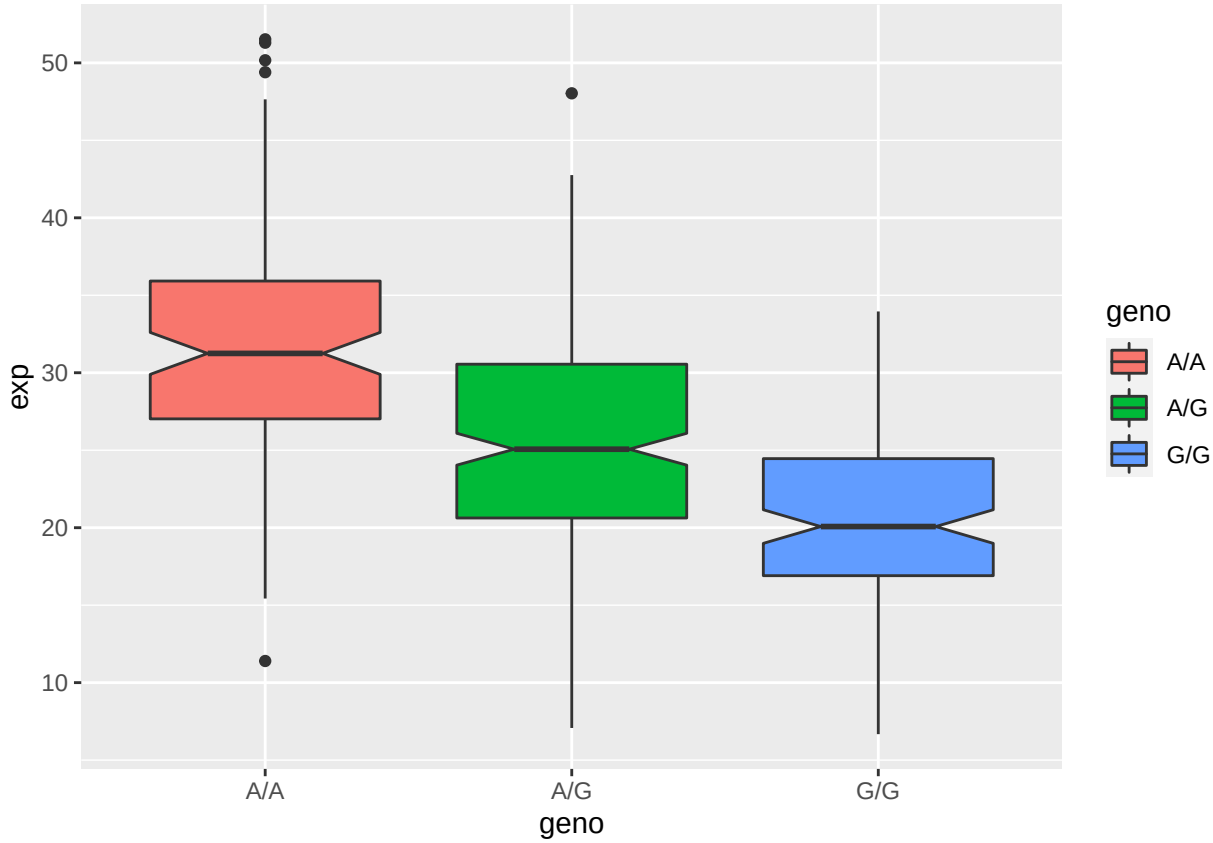
```
##
## $n
## [1] 108 233 121
##
## $conf
##          [,1]     [,2]     [,3]
## [1,] 29.87942 24.03742 18.98858
## [2,] 32.61753 26.09230 21.15868
##
## $out
## [1] 51.51787 50.16704 51.30170 11.39643 48.03410
##
## $group
## [1] 1 1 1 1 2
##
## $names
## [1] "A/A" "A/G" "G/G"
```

The sample size for each genotype and their corresponding median expression levels for each of these genotypes: A/A: sample size is 108, and corresponding median expression level is 31.24847. A/G: sample size is 233, and corresponding median expression level is 25.06486. G/G: sample size is 121, and corresponding median expression level 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?

plot them using ggplot.

```
library(ggplot2)
boxp <- ggplot(expr) + aes(geno, exp, fill=geno) + geom_boxplot(notch=TRUE)
boxp
```



Based on the boxplot demonstrated here, we can easily observe that the relative expression value between A/A and G/G are statistically different, as the relative expression value of A/A is significantly higher than that of G/G, which indicates that the SNP effect the expression of ORMDL3.