Week 19: Genome Informatics

Stefanie Hodapp (PID: A53300084)

12/6/2021

Population Scale Analysis

> 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.

```
# read this .txt file into R
df <- read.table("rs8067378_ENSG00000172057.6.txt")
head(df)
```

```
##    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
```

```
# determine the sample size for each genotype
table(df$geno)        # A/A: 108, A/G: 233, G/G: 121
```

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

```
# determine 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
```

```
# filter data for each genotype
df_AA <- filter(df, geno == "A/A")
head(df_AA)
```

```
##     sample geno      exp
## 3  HG00361  A/A 31.32628
## 4  HG00135  A/A 34.11169
## 6  NA11993  A/A 32.89721
## 8  NA18498  A/A 47.64556
## 13 NA20585  A/A 30.71355
## 15 HG00235  A/A 25.44983
```

```
df_AG <- filter(df, geno == "A/G")
head(df_AG)
```

```
##     sample geno      exp
## 1  HG00367  A/G 28.96038
## 2  NA20768  A/G 20.24449
## 7  HG00256  A/G 31.48736
## 10 HG00115  A/G 33.85374
## 11 NA20806  A/G 16.29854
## 12 HG00278  A/G 19.73450
```

```
df_GG <- filter(df, geno == "G/G")
head(df_GG)
```

```
##     sample geno      exp
## 5  NA18870  G/G 18.25141
## 9  HG00327  G/G 17.67473
## 17 NA12546  G/G 18.55622
## 20 NA18488  G/G 23.10383
## 23 NA19214  G/G 30.94554
## 28 HG00112  G/G 21.14387
```

```
# Calculate median expression levels for each genotype
summary(df_AA)
```

```
##     sample              geno                exp
##  Length:108         Length:108         Min.   :11.40
##  Class :character   Class :character   1st Qu.:27.02
##  Mode  :character   Mode  :character   Median :31.25
##                                        Mean   :31.82
##                                        3rd Qu.:35.92
##                                        Max.   :51.52
```

```
median(df_AA$exp)         # 31.25
```

```
## [1] 31.24847
```

```
summary(df_AG)
```

```
##     sample              geno                exp        
##  Length:233         Length:233         Min.   : 7.075  
##  Class :character   Class :character   1st Qu.:20.626  
##  Mode  :character   Mode  :character   Median :25.065  
##                                        Mean   :25.397  
##                                        3rd Qu.:30.552  
##                                        Max.   :48.034
```

```
median(df_AG$exp)         # 25.065
```

```
## [1] 25.06486
```

```
summary(df_GG)
```

```
##     sample              geno                exp        
##  Length:121         Length:121         Min.   : 6.675  
##  Class :character   Class :character   1st Qu.:16.903  
##  Mode  :character   Mode  :character   Median :20.074  
##                                        Mean   :20.594  
##                                        3rd Qu.:24.457  
##                                        Max.   :33.956
```

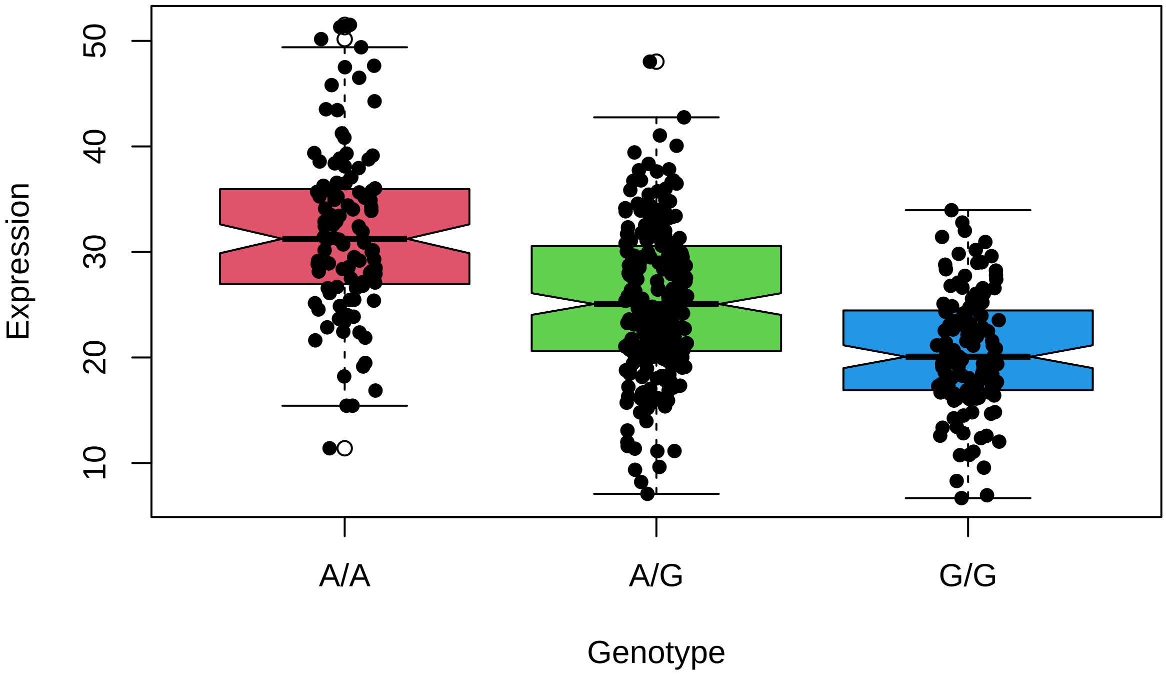
```
median(df_GG$exp)         # 20.074
```

```
## [1] 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?
>
> From this boxplot, you can infer that the genotype affects ORMDL3 expression, in which G/G corresponds with the least expression while A/A corresponds with the greatest expression. Based on these data, this SNP does effect the expression of ORMDL3.

```
boxplot(exp~geno, data=df, xlab="Genotype", ylab="Expression", col=2:4, notch = TRUE)

stripchart(exp~geno, data=df, method = "jitter", pch = 16, vertical = TRUE, add = TRUE)
```


Expression
50
40
30
20
10
A/A
A/G
G/G
Genotype