week08

May Wu PID:A59010588

12/1/2021

Section 1, proportion of G/G in a pupulation

download a csv file from ensemble: https://uswest.ensembl.org/Homo_sapiens/Variation/Sample?db=core;r=17:39898867-40018868;v=rs8067378;vdb=variation;vf=105535077#373531_tablePanel

```
mxl = read.csv("373531-SampleGenotypes-Homo_sapiens_Variation_Sample_rs8067378.csv")
head(mxl)
```

```
##   Sample..Male.Female.Unknown. Genotype..forward.strand. Population.s. Father
## 1                  NA19648 (F)                       A|A ALL, AMR, MXL      -
## 2                  NA19649 (M)                       G|G ALL, AMR, MXL      -
## 3                  NA19651 (F)                       A|A ALL, AMR, MXL      -
## 4                  NA19652 (M)                       G|G ALL, AMR, MXL      -
## 5                  NA19654 (F)                       G|G ALL, AMR, MXL      -
## 6                  NA19655 (M)                       A|G ALL, AMR, MXL      -
##   Mother
## 1      -
## 2      -
## 3      -
## 4      -
## 5      -
## 6      -
```

```
table(mxl$Genotype..forward.strand.)
```

```
##
## A|A A|G G|A G|G
##  22  21  12   9
```

```
table(mxl$Genotype..forward.strand.)/nrow(mxl)*100
```

```
##
##     A|A     A|G     G|A     G|G
## 34.3750 32.8125 18.7500 14.0625
```

```
gbr = read.csv("373522-SampleGenotypes-Homo_sapiens_Variation_Sample_rs8067378.csv")
```

```
round(table(gbr$Genotype..forward.strand.) / nrow(gbr) * 100,2)
```

```
##
##   A|A   A|G   G|A   G|G
## 25.27 18.68 26.37 29.67
```

This variant that is associated with childhood asthma is more frequent in the GBR population than the MKL population.

Let's dig into this further

```
expr = read.table("sample_data.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
```

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.

A: There are 462 data in total. Information for each genotype as follow:

- genotype; sample size; median
- A/A : 108; 31.25
- A/G : 233; 25.065
- G/G : 121; 20.074

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

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

```
summary(expr[expr$geno == "A/A",])
```

```
##     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
```

```
summary(expr[expr$geno == "A/G",])
```

```
##     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
```

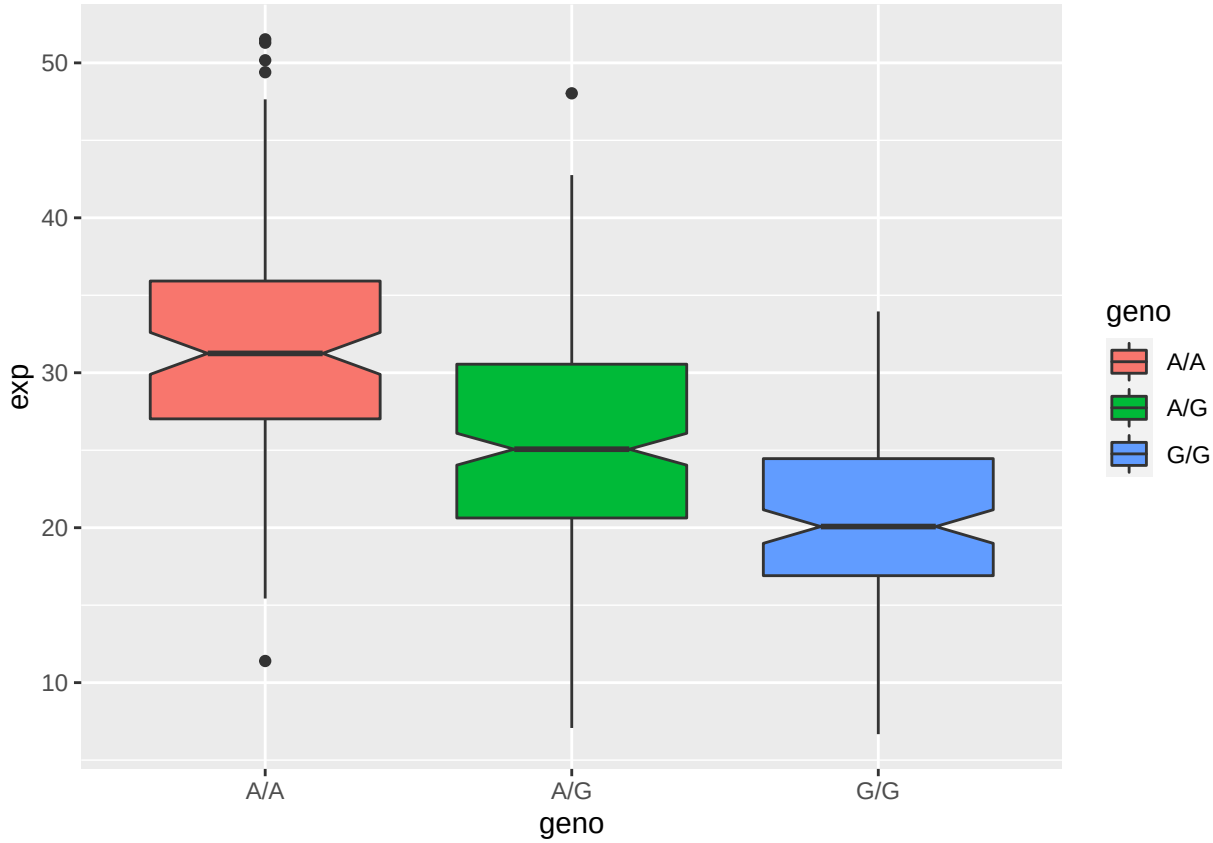
```
summary(expr[expr$geno == "G/G",])
```

```
##     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
```

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?

A: A/A and G/G has different expression level, which means this SNP does affect ORMDL3 expression level.

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


50
40
30
20
10
exp
A/A
A/G
G/G
geno
geno
A/A
A/G
G/G