CLass10

Ethan Lai

2/18/2022

What proportion of the MXL sample population is G/G homozygous

```
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.) /nrow(mxl) *100
```

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

14.065% are G/G homozygous

#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. 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.

```
expr <- read.table("rs8067378_ENSG00000172057.6.txt")
exprTable<- table(expr$geno)
exprTable
```

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

108 A/A, 233 A/G, and 121 G/G

To find the medians:

```
exprdf<- data.frame(expr)
exprAA<- exprdf[exprdf[2]=="A/A",]
exprAG<- exprdf[exprdf[2]=="A/G",]
exprGG<- exprdf[exprdf[2]=="G/G",]
median (exprAA[,3])
```

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

```
median (exprAG[,3])
```

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

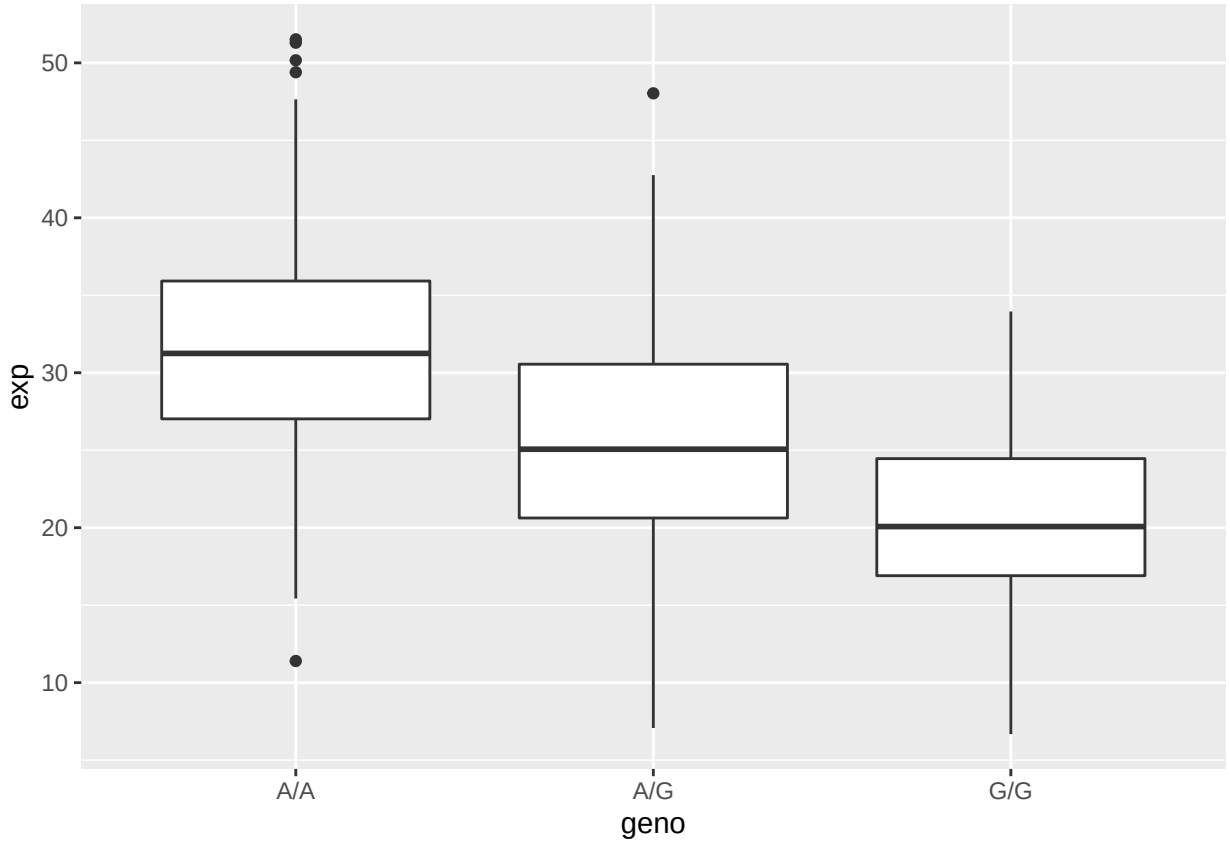
```
median (exprGG[,3])
```

```
## [1] 20.07363
```

So the medians for A/A, A/G, and G/G respectively are 31, 25, and 20.

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?

```
library(ggplot2)
ggplot(expr, aes(x=geno, y=exp)) + geom_boxplot()
```



G/G seems to have the lowest expression, with AA the highest, and heterozygous A/G intermediate between the two homozygous genotypes!