Class10

Ayse

2/18/2022

Genotype data from 1000 gemones

We need to determine frequence of different alleles in the MXL population

```
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)
```

```
##
##      A|A      A|G      G|A      G|G
## 0.343750 0.328125 0.187500 0.140625
```

In the GBR population

```
mxl <- read.csv("373522-SampleGenotypes-Homo_sapiens_Variation_Sample_rs8067378.csv")
head(mxl)
```

```
##   Sample..Male.Female.Unknown. Genotype..forward.strand. Population.s. Father
## 1                  HG00096 (M)                       A|A ALL, EUR, GBR      -
## 2                  HG00097 (F)                       G|A ALL, EUR, GBR      -
## 3                  HG00099 (F)                       G|G ALL, EUR, GBR      -
## 4                  HG00100 (F)                       A|A ALL, EUR, GBR      -
## 5                  HG00101 (M)                       A|A ALL, EUR, GBR      -
## 6                  HG00102 (F)                       A|A ALL, EUR, GBR       -
##   Mother
## 1      -
## 2      -
## 3      -
## 4      -
## 5      -
## 6      -
```

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

```
##
##       A|A       A|G       G|A       G|G
## 0.2527473 0.1868132 0.2637363 0.2967033
```

Homework

Q13:determine the sample size for each genotype and their corresponding median expression levels for each of these genotypes

sample size for A/A is 108, A/G is 233. Median expression level for A/A is 31.25, A/G is 25.06, and G/G is 20.07.

```
tbl <- read.table("rs8067378_ENSG00000172057.6.txt")
summary(tbl)
```

```
##      sample     geno          exp
##  HG00096:  1   A/A:108   Min.   : 6.675
##  HG00097:  1   A/G:233   1st Qu.:20.004
##  HG00099:  1   G/G:121   Median :25.116
##  HG00100:  1             Mean   :25.640
##  HG00101:  1             3rd Qu.:30.779
##  HG00102:  1             Max.   :51.518
##  (Other):456
```

```
gg <- median(tbl[tbl$geno == "G/G",]$exp)
aa <- median(tbl[tbl$geno == "A/A",]$exp)
ag <- median(tbl[tbl$geno == "A/G",]$exp)
gg
```

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

```
aa
```

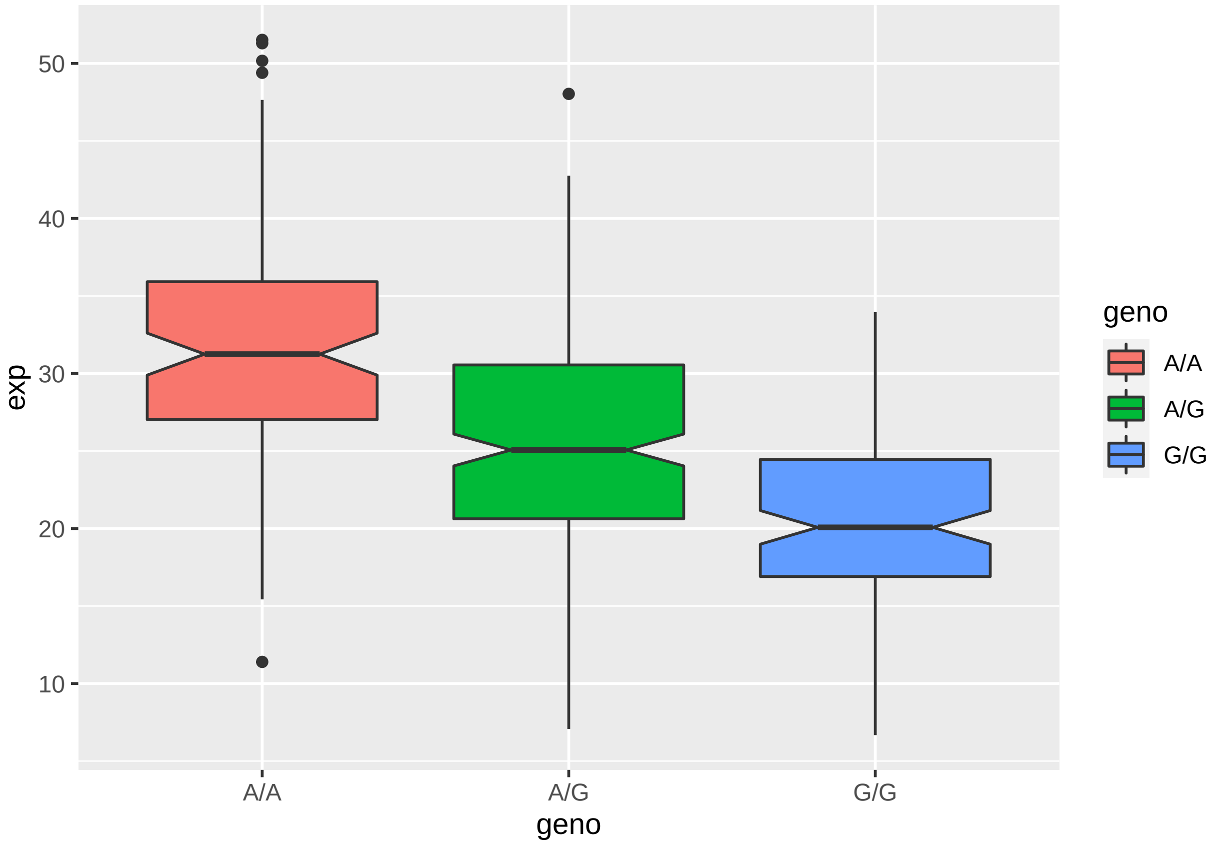
```
## [1] 31.24847
```

```
ag
```

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

boxplot for each genotype

```r
library(ggplot2)
bx <- ggplot(tbl, aes(geno, exp, fill = geno))+
         geom_boxplot(notch = "TRUE")
bx
```



Q14: 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? Expression of ORMDL3 is higher for the A/A geontype compared to the G/G genotype. The SNP may down-regulate expression of ORMDL3.