Class 11

Erica Sanchez (A15787505)

2024-05-12

```r
mxl <- read.csv("SampleGenotypes.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      -
```

Section 1 - proportion of (G/G) in a population

```r
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
```

Section 4 – 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 the dataset
filegp <- read.table("https://bioboot.github.io/bggn213_W19/class-material/rs8067378_ENSG0
```

```
summary(filegp)
```

```
    sample              geno                exp
 Length:462         Length:462         Min.   : 6.675
 Class :character   Class :character   1st Qu.:20.004
 Mode  :character   Mode  :character   Median :25.116
                                       Mean   :25.640
                                       3rd Qu.:30.779
                                       Max.   :51.518
```

```
str(filegp)
```

```
'data.frame':    462 obs. of  3 variables:
 $ sample: chr  "HG00367" "NA20768" "HG00361" "HG00135" ...
 $ geno  : chr  "A/G" "A/G" "A/A" "A/A" ...
 $ exp   : num  29 20.2 31.3 34.1 18.3 ...
```

```
# calculate the median expression level for each genotype
medians <- tapply(filegp$exp, filegp$geno, median)
print(medians)
```

```
     A/A      A/G      G/G
31.24847 25.06486 20.07363
```

```
# calculate the sample size for each genotype
sample_sizes <- tapply(filegp$geno, filegp$geno, length)
print(sample_sizes)
```

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

> 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?

```
boxplot(exp ~ geno, data = filegp, xlab = "geno", ylab = "exp", main = "exp of ORMDL3 by g
```
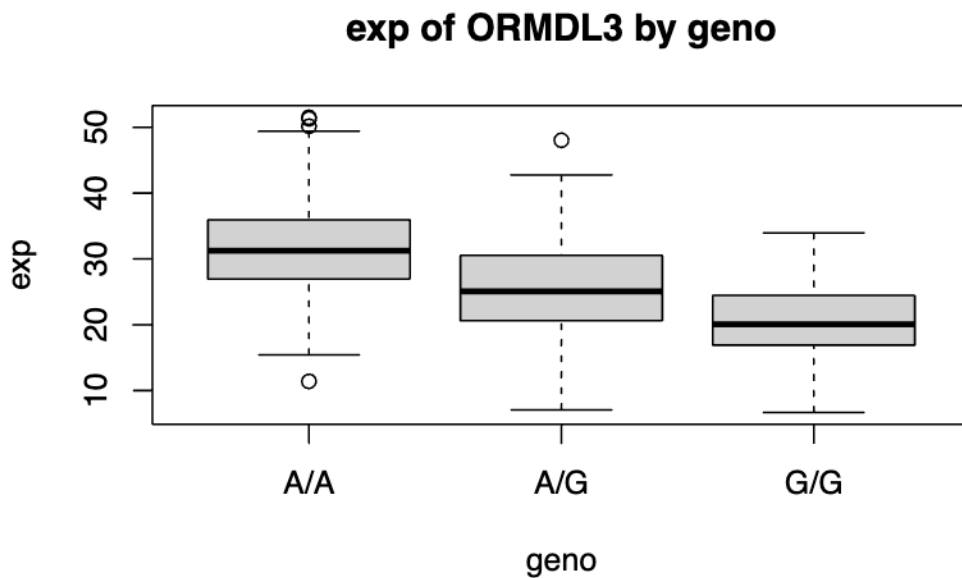

The SNP affects the expression of ORMDL3 because the boxplot shows the median expression
levels of A/A are higher than the median expression levels of G/G.