Class 12

Saba Heydari Seradj

Section 1. Proportion of G/G in a population

Downloaded a CSV file from Ensemble.

Here we read the CSV file:

```
mxl <- read.csv('genotypes.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
```

14.06% of the population are G|G.

Section 4. Population analysis

```
expr <- read.table('rs8067378_ENSG00000172057.6.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
```

How many samples do we have?

```
nrow(expr)
```

```
[1] 462
```

462 samples.

What is the sample size for each genotype?

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

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

A/A has 108 samples, A/G has 233, and G/G has 121.

```
summary(expr)
```

```
    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  
```

What is the median expression levels for each genotype?

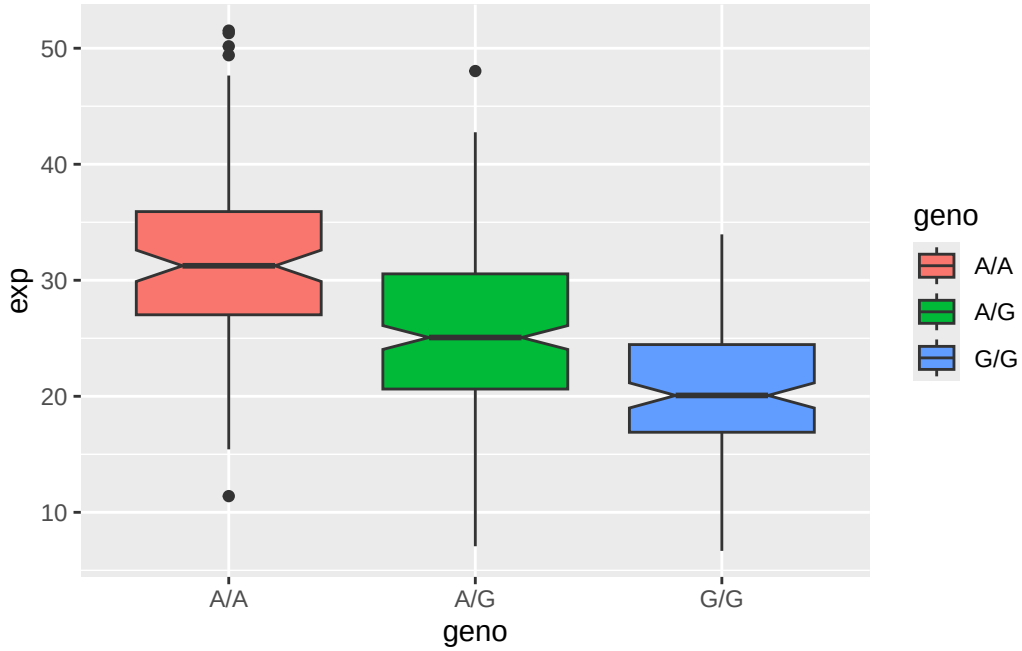
```
median_expression <- tapply(expr$exp, expr$geno, median, na.rm = TRUE)
median_expression
```

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

Let's make a boxplot!

```
library(ggplot2)

g <- ggplot(expr) + aes(geno, exp, fill=geno) + geom_boxplot(notch=T)
g
```



Q14

We can infer that the G/G SNP reduces expression of ORMDL3 as it has the lowest expression compared to other genotypes.