class 11

Vidisha Marwaha (PID: A16677246)

Section 1. proportion of G/G in a population

Downloaded a CSV file from Ensemble < https://useast.ensembl.org/Homo_sapiens/Variation/Sample?db=core
39955100;v=rs8067378;vdb=variation;vf=959672880#373531_tablePanel >

Here we read the CSV file -

```
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 
```

Now lets look at a different population. I picked the GBR.

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

Find proportion of G/G

```
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 MXL population.

Lets dig into this further.

Section 4: Population Scale Analysis

[HOMEWORK] One sample is obviously not enough to know what is happening in a population. You are interested in assessing genetic differences on a population scale. So, you processed about ~230 samples and did the normalization on a genome level. Now, you want to find whether there is any association of the 4 asthma-associated SNPs (rs8067378…) on ORMDL3 expression. This is the final file you got. The first column is sample name, the second column is genotype and the third column are the expression values. Open a new RMarkdown document in RStudio to answer the following two questions. Submit your resulting PDF report with your working code, output and narrative text answering Q13 and Q14 to GradeScope.

How many samples do we have?

```
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
```

```
nrow(expr)
```

```
[1] 462
```

The sample sizes for each genotype is -

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

```

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

```
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  
```

The median expression levels for each of these genotypes is -

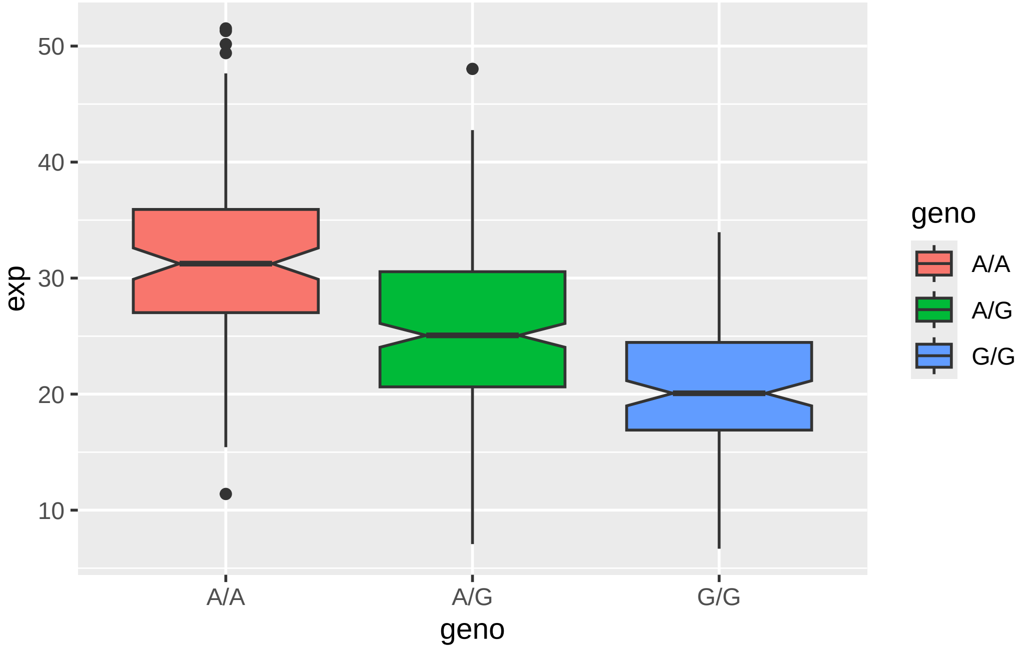
```
genotype <- expr$geno
expression <- expr[, 3]
genotype_data <- split(expression, genotype)
median_expression <- sapply(genotype_data, median)
median_expression
```

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

```r
library(ggplot2)
```

Let’s make a boxplot

```r
ggplot(expr) + aes(geno, exp, fill=geno) +
 geom_boxplot(notch=T)
```



The G|G has lower expression compared to A|A according to the boxplot. Having a G|G in this location is associated to having reduced expression of ORMDL3. The SNP does effect the expression of ORMDL3.