Class12 Homework

Jordan Prych A17080226

Table of contents

```
knitr::opts_chunk$set(echo=TRUE)
```

Section 1. Proportiona of G/G in a Population

Downloaded CSV file from Ensemble

Here we read this CSV file to determine allele frequency

```
MXL <- read.csv("373531-SampleGenotypes-Homo_sapiens_Variation_Sample_rs8067378 (1).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      -
```

```
MXL$Genotype..forward.strand.
```

```
 [1] "A|A" "G|G" "A|A" "G|G" "G|G" "A|G" "A|G" "A|A" "A|G" "A|A" "G|A" "A|A"
[13] "A|A" "G|G" "A|A" "A|G" "A|G" "A|G" "A|G" "G|A" "A|G" "G|G" "G|G" "G|A"
[25] "G|G" "A|G" "A|A" "A|A" "A|G" "A|A" "A|G" "G|A" "G|G" "A|A" "A|A" "A|A"
[37] "G|A" "A|G" "A|G" "A|G" "A|A" "G|A" "A|G" "G|A" "G|A" "A|A" "A|A" "A|G"
[49] "A|A" "A|A" "A|G" "A|G" "A|A" "G|A" "A|A" "G|A" "A|G" "A|A" "G|A" "A|G"
[61] "G|G" "A|A" "G|A" "A|G"
```

```
table(MXL$Genotype..forward.strand.)
```

```

A|A A|G G|A G|G 
 22  21  12   9 
```

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

```

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

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.

> 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.

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
```

There are 462 individuals(this is the sample size).

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

```

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

Let's find the median expression levels for each genotype from the boxplot below.

```
median <- tapply(expr$exp, expr$geno, median)
median
```

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

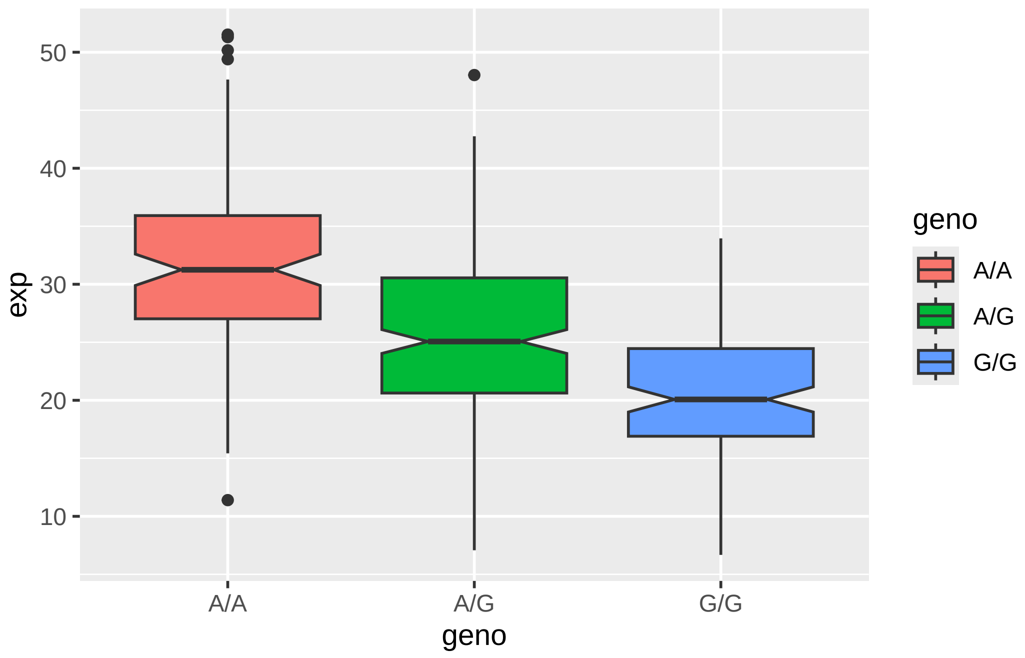
> Q14: Generate a boxplot with a box per genotype, what could you infer from the relative expression value between A/A and G/G displayed in this plot? Does the SNP effect the expression of ORMDL3?

From this boxplot, we can infer that having a G/G genotype results in decreased expression compared to a A/A genotype. Therefore, the SNP does effect the expression of ORMDL3, since a change in the nucleotide from an A to a G results in overall decreased expression of ORMDL3.

Let's make a boxplot of this data:

```
library(ggplot2)

bp <- ggplot(expr) + aes(x=geno, y=exp, fill=geno) + geom_boxplot(notch=TRUE)
bp
```


50
40
30
20
10
exp
A/A
A/G
G/G
geno
geno
A/A
A/G
G/G