Class 12

Isabel Hui - A16887852

Section 1: Proportion of G/G in population

Downloaded a CSV file from Ensemble https://useast.ensembl.org/Homo_sapiens/Variation/Sample?db=core;r=17:39895045-39895146;v=rs8067378;vdb=variation;vf=959672880;sample=HG00109

Now we can read this CSV file.

```
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      -
```

```
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 [HOMEWORK]

One sample is not enough to know what is happening in a population.

How many samples do we have?

```
expr <- read.table("rs8067378.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
```

> 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.

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

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

```
library(dplyr)
```

```
Attaching package: 'dplyr'

The following objects are masked from 'package:stats':

    filter, lag
```

```
The following objects are masked from 'package:base':

    intersect, setdiff, setequal, union
```

```
df <- expr
str(df)
```

```
'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 ...
```

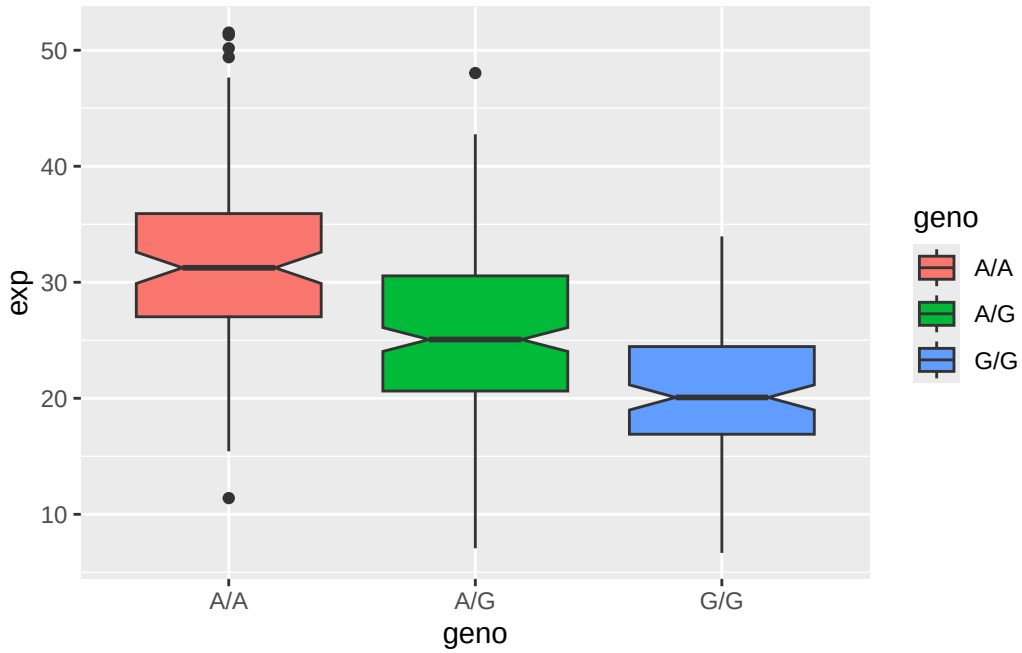
```
summary_df <- df %>%
  group_by(geno) %>%
  summarise(
    sample_size = n(),
    median_expression = median(exp, na.rm = TRUE)
  )

summary_df
```

```
# A tibble: 3 x 3
  geno  sample_size median_expression
  <chr>       <int>             <dbl>
1 A/A           108              31.2
2 A/G           233              25.1
3 G/G           121              20.1
```

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?

```
library(ggplot2)
```

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



The box plot does show differing levels of expression between the A/A and G/G genotype. The SNP does effect the expression of ORMDL3, as a G/G genotype (blue box) can be seen to significantly decrease expression as the mean and top and bottom quartile all lie below the values of the A/A genotype expression.