Class11

Juliette Bokor (PID: A16808121)

Section 1. Proportion of G/G in MXL population

Downloaded a CSV file from Ensemble

```
mxl <- read.csv("MXL_populationSNP.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
```

There are 9 samples with the genotype G|G

Section 4: Homework on 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.

```
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
```

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

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

```
## I need to find the median for each genotype
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
```

```
medians <- expr %>%
  group_by(geno)%>%
  summarize(median_exp = median(exp))
medians
```

```
# A tibble: 3 x 2
  geno  median_exp
  <chr>      <dbl>
1 A/A         31.2
2 A/G         25.1
3 G/G         20.1
```
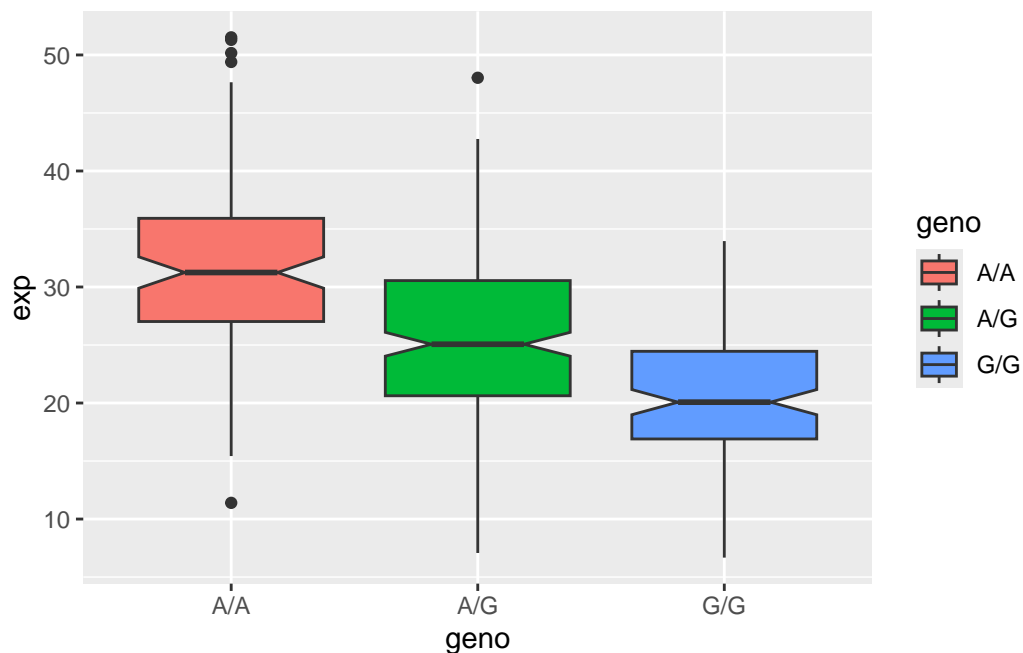
Q14: Generate a boxplot with a box per genotype, what could you infer from the relative expression value between A/A and G/G displayed in this plot? Does the SNP effect the expression of ORMDL3?

```
library(ggplot2)
exprplot <- ggplot(expr) +
  aes(x=geno, y=exp, fill=geno) +
  geom_boxplot(notch=TRUE)

exprplot
```

The relative expression of ORMDL3 is decreased with the G/G genotype in comparison to the A/A genotype. The SNP does affect the expression of ORMDL3.