Class 12

Kristiana Wong A16281367

2024-02-19

##Proportion of homozygous allele in MXL population

```r
mxl <- read.csv("MXL.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      -
```

```r
table(mxl$Genotype..forward.strand.) / nrow(mxl) * 100
```

```
## 
##     A|A     A|G     G|A     G|G 
## 34.3750 32.8125 18.7500 14.0625
```

##Part 2. Population Scale Analysis Homework

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.

Hint: The read.table(), summary() and boxplot() functions will likely be useful here. There is an example R script online to be used ONLY if you are struggling in vein. Note that you can find the medium value from saving the output of the boxplot() function to an R object and examining this object. There is also the medium() and summary() function that you can use to check your understanding.

```r
ORMDL3 <- read.table("rs8067378_ENSG00000172057.6.txt")
head(ORMDL3)
```

```
##    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
```

```r
# Group by Genotype and calculate sample size and median expression
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
```

```r
summary_table <- ORMDL3 %>%
  group_by(geno) %>%
  summarise(
    SampleSize = n(),           # Count of samples for each genotype
    MedianExpression = median(exp, na.rm = TRUE)  # Median expression for each genotype
  )

# Display the summary table
print(summary_table)
```

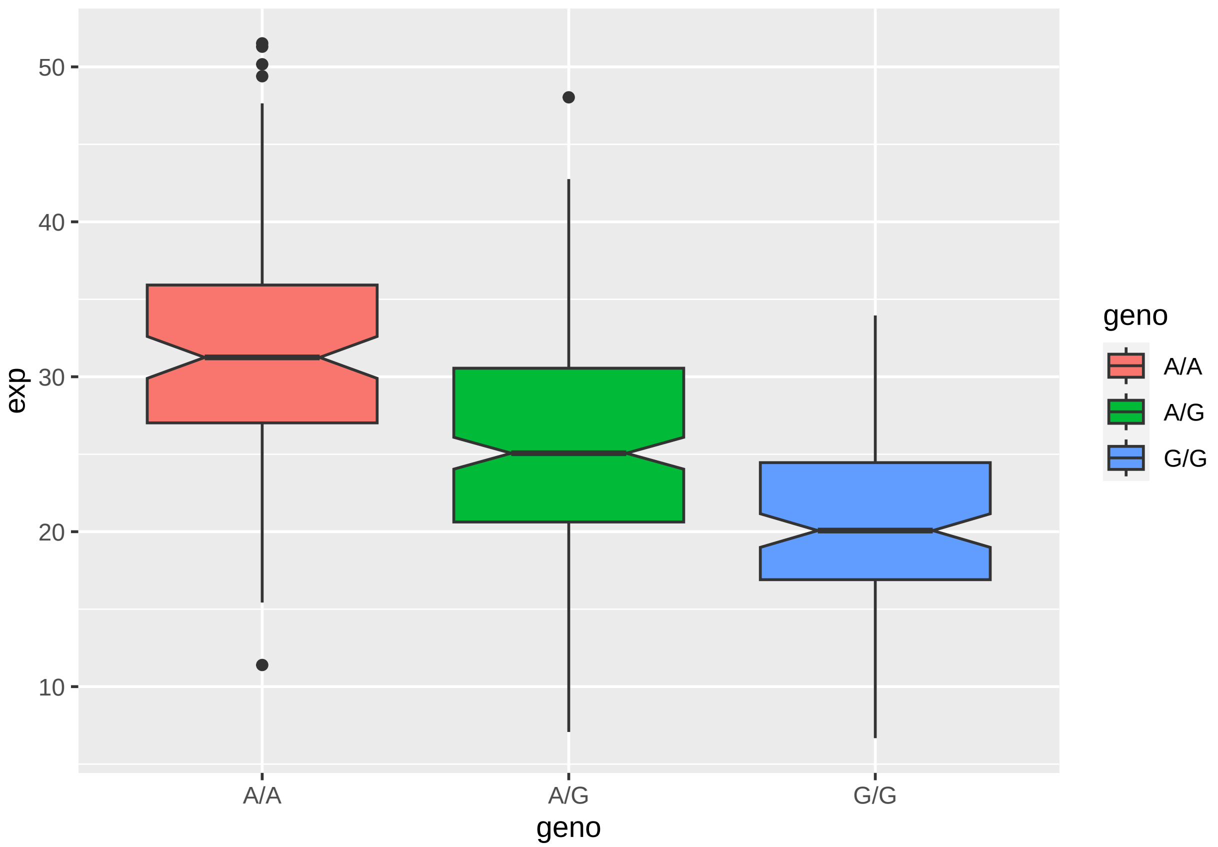
```
## # A tibble: 3 x 3
##   geno  SampleSize MedianExpression
##   <chr>      <int>            <dbl>
## 1 A/A          108             31.2
## 2 A/G          233             25.1
## 3 G/G          121             20.1
```

There are 108 A/A, 233 A/G, and 121 G/G genotypes in this sample. The median expression is 31.24%, 25.06%, and 20.07% respectively.

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?

```r
library(ggplot2)

ggplot(ORMDL3) +
  aes(geno, exp, fill = geno) +
  geom_boxplot(notch = T)
```



We can infer that A/A or the A allele is more highly expressed than G/G or G allele. Hence, from this, indeed the SNP does effect ORMDL3 expression.