# class18_genome

Jocelyn Olvera

12/1/2021

```
library(ggplot2)
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
```

```
expr <- read.table("../class18_genome/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
```

```
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
```

**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.**

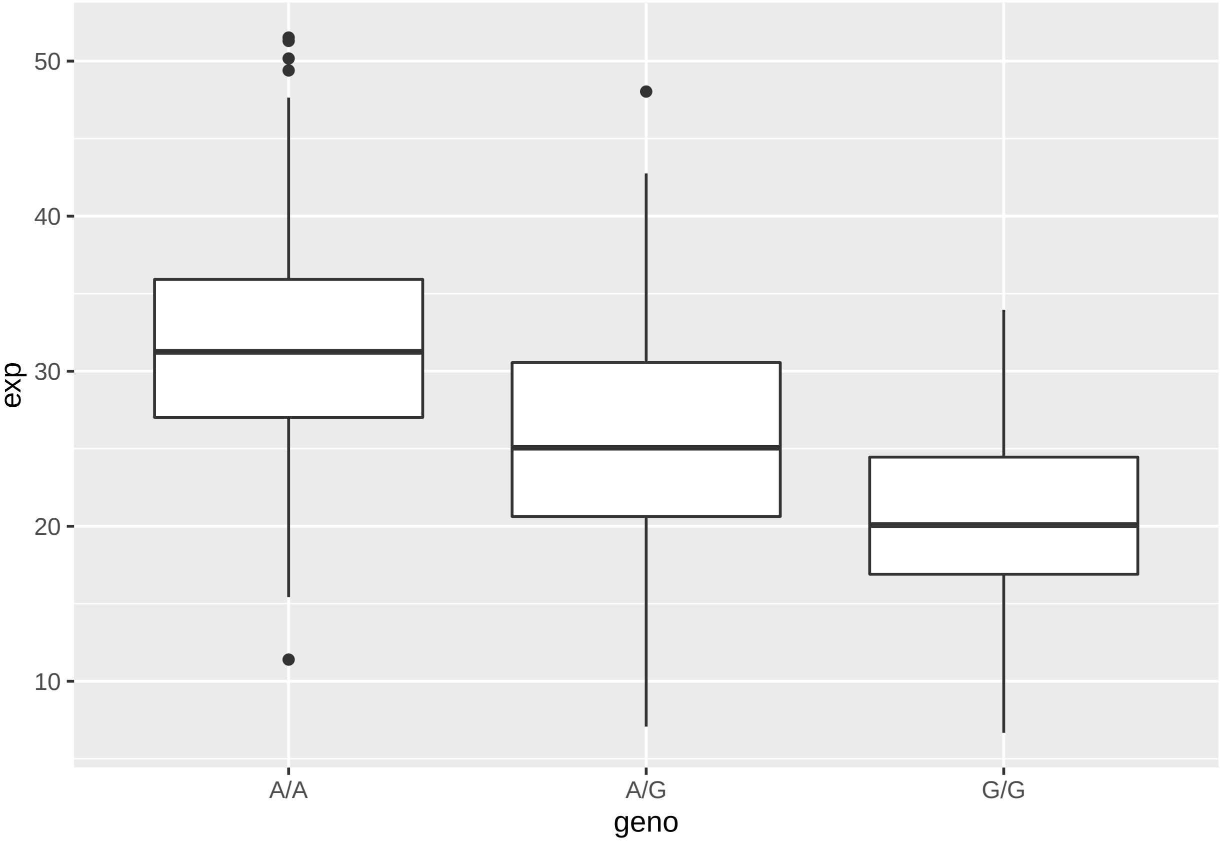
**Response:**

**A/A median = 31.24847**

**A/G median = 25.06486**

**G/G median = 20.07363**

```
ggplot(expr,
       aes(x = geno,
           y = exp)) +
  geom_boxplot()
```



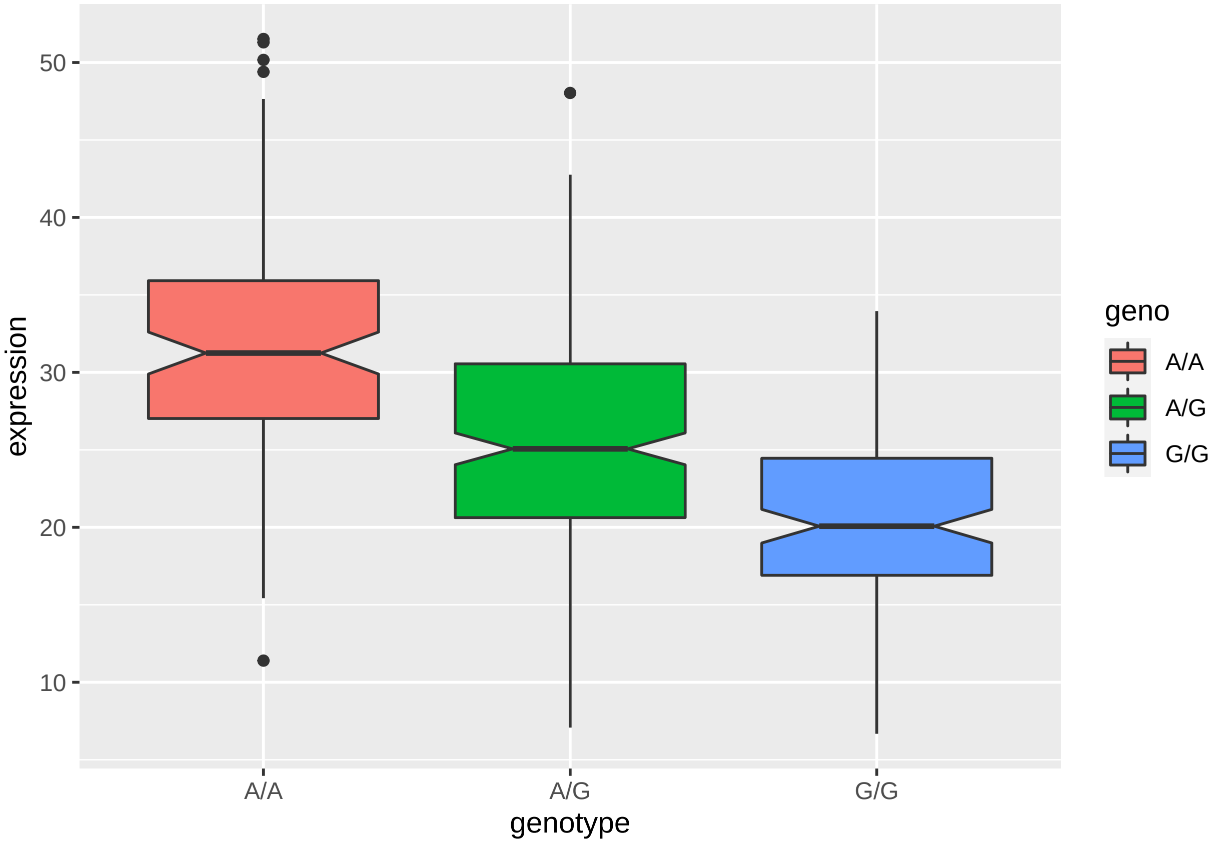
```
expr %>%
 group_by(geno) %>%
  summarise(Min = min(exp),
            Max = max(exp),
            Median = median(exp))
```

```
## # A tibble: 3 x 4
##   geno    Min   Max Median
##   <chr> <dbl> <dbl>  <dbl>
## 1 A/A   11.4   51.5   31.2
## 2 A/G    7.08  48.0   25.1
## 3 G/G    6.67  34.0   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? Hint: An example boxplot is provided overleaf – yours does not need to be as polished as this one.

**Response: A/A is more highly expressed than G/G in this gene. However, the presence of G/G is associated with a decreased expression of this gene. therefore, the SNP appears to effect the expression of the ORMDL3 gene and is inferred by the presence of G/G.**

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



### Questions #5 and 6

**Data downloaded from: https://uswest.ensembl.org/Homo_sapiens/Variation/Sample?db=core;r=17:39903525-39926526;v=rs8067378;vdb=variation;vf=105535077#373531_tablePanel**

### Section1: Proportion of G/G in a population

```
mxl <- read.csv("../class18_genome/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
```

# Total number of visuals in the dataset. Percentage of homozygous.

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

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

```
gbr <- read.csv("../class18_genome/373522-SampleGenotypes-Homo_sapiens_Variation_Sample_rs8067378.csv")
```

### Find proportion of GIG

```
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 asthama is more frequent in the GBR population thatn the MKL population.

##Lets now dig into this further.