Class 18 Worksheet

Julia Napoli

12/1/2021

Section 1. Proportion of G/G in a population

Downloaded a CSV file with desired data. Now we'll read this CSV file.

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

```
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 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.

How many samples do we have?

```
expr <- read.table("worksheet18file.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
```

Let's determine the sample size for each genotype.

```
sample_size <- table(expr$geno)

sample_size
```

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

```
sum(sample_size)
```

```
## [1] 462
```

Now let's find the median expression levels for each of these genotypes.

```
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
```

```
AA_rows <- filter(expr, expr$geno == "A/A")
summary(AA_rows$exp)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   11.40   27.02   31.25   31.82   35.92   51.52
```

```
AG_rows <- filter(expr, expr$geno == "A/G")
summary(AG_rows$exp)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   7.075  20.626  25.065  25.397  30.552  48.034
```

```
GG_rows <- filter(expr, expr$geno == "G/G")
summary(GG_rows$exp)
```

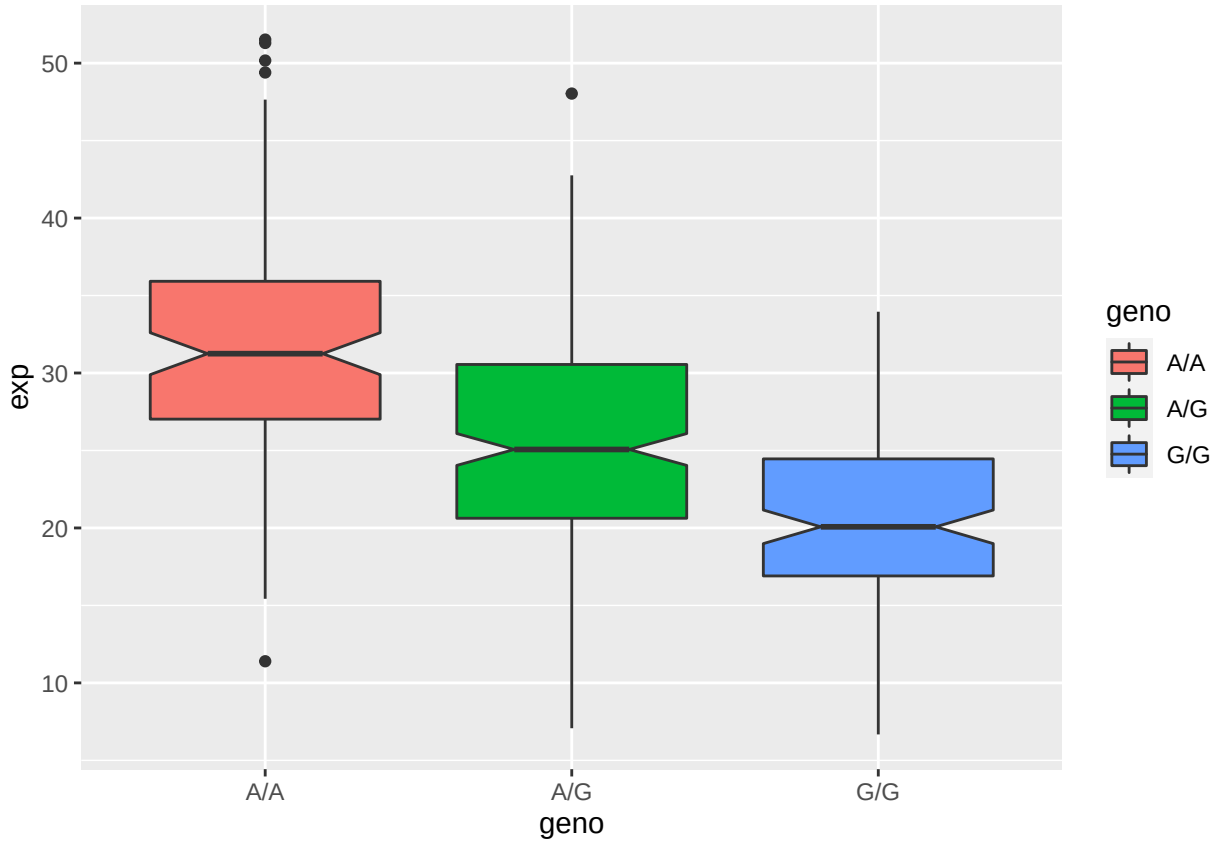
```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   6.675  16.903  20.074  20.594  24.457  33.956
```

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?

Let's make a boxplot with this data.

```
library(ggplot2)
```

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



You can infer that the expression value between the A/A and G/G genotypes are stastistically significantly different, since their IQRs are entirely visually distinct from one another.

Yes, the SNP does affect the expression of ORMDL3! The G/G phenotype is associated with having reduced expression of this gene in comparison to the other genotypes.