Class 12: Q13-Q14 Homework

Pamelina Lo (SID: A16735368)

2024-11-07

Section 1. Proportion of G/G in a population

Downloaded a CSV file from Ensemble < https://useast.ensembl.org/Homo_sapiens/Variation/Sample?db=core;r=17:39780097-40010098;v=rs8067378;vdb=variation;vf=959672880#373531_tablePanel

Here we 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
```

Lets look at a different population. I picked the GBR (Great Britian).

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

```
##   Sample..Male.Female.Unknown. Genotype..forward.strand. Population.s. Father
## 1                 HG00096 (M)                       A|A ALL, EUR, GBR      -
## 2                 HG00097 (F)                       G|A ALL, EUR, GBR      -
## 3                 HG00099 (F)                       G|G ALL, EUR, GBR      -
```

```
## 4                  HG00100 (F)                        A|A ALL, EUR, GBR      -
## 5                  HG00101 (M)                        A|A ALL, EUR, GBR      -
## 6                  HG00102 (F)                        A|A ALL, EUR, GBR      -
##   Mother
## 1      -
## 2      -
## 3      -
## 4      -
## 5      -
## 6      -
```

Find proportion of G|G

```
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 assoicated with childhood asthma is more frequent in the GBR population than the MKL population.

Section 4: 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("https://bioboot.github.io/bimm143_F24/class-material/rs8067378_ENSG00000172057.6.tx
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
```

Sample Size:

```
nrow(expr)
```

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

Sample size and Median expression levels 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
```

```r
genotype_summary <- expr %>%
  group_by(expr$geno) %>%
  summarise(
    sample_size = n(),
    median_expression = median(exp, na.rm = FALSE)
  )

print(genotype_summary)
```

```
## # A tibble: 3 x 3
##   `expr$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
```

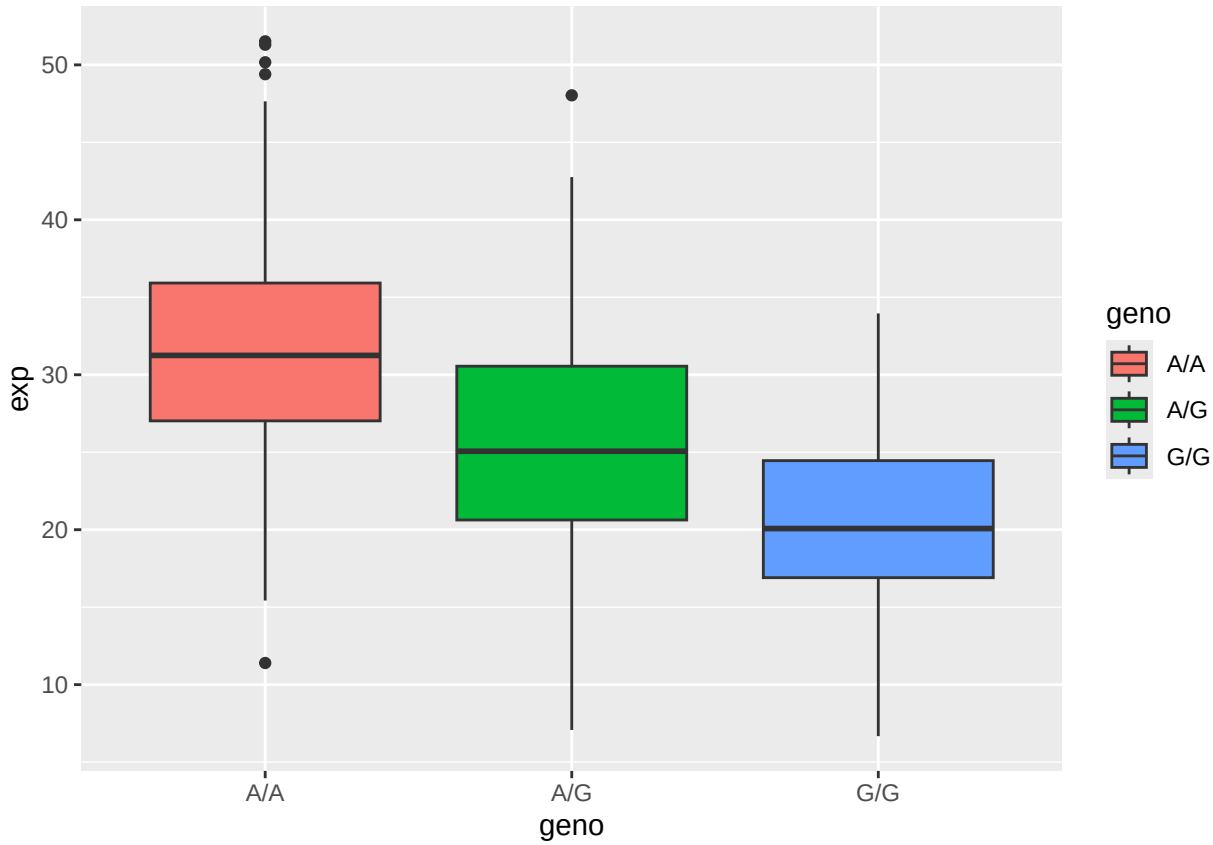
The sample sizes: A|A = 108 , A|G = 233, G|G = 121 Median Expression: A|A = 31.25 , A|G = 25.06 , G|G = 20.07

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?

I can infer that the relative expression value is that there is more expression of the A|A genotype, than G|G because, based on the plots, A|A appears to be higher than G|G plot because its median is larger and the distribution of the plot is more spread out. The SNP does effect the expression of ORMDL3 because having G|G genotype in this location associates to a reduced expression on this gene.

```r
library(ggplot2)
```

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


50
40
30
20
10
exp
A/A
A/G
G/G
geno
geno
A/A
A/G
G/G