class 11

Loretta Cheng

2023-05-30

#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=17:39900444-39901444;v=rs8069176;vdb=variation;vf=105553859#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
```

Let's look at a different population. I picked the GBR.

```
gbr <- read.csv("373522-SampleGenotypes-Homo_sapiens_Variation_Sample_rs8069176 (1).csv")
```

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 25.27 19.78 29.67
```

This variant that is associated with childhood asthma is frequence in the BGR population compared to MKL population.

Section 4: Population Scale Analysis

One sample is obviously not enough to know what is happening in a population. You are interested in assessing genetic differences on a population scale. So, you processed about ~230 samples and did the normalization on a genome level. Now, you want to find whetherthere is any association of the 4 asthma-associated SNPs (rs8067378…) on ORMDL3 expression.

How many samples do we have?

```
expr <- read.table("rs8067378_ENSG00000172057.6.txt",header=TRUE)
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
```

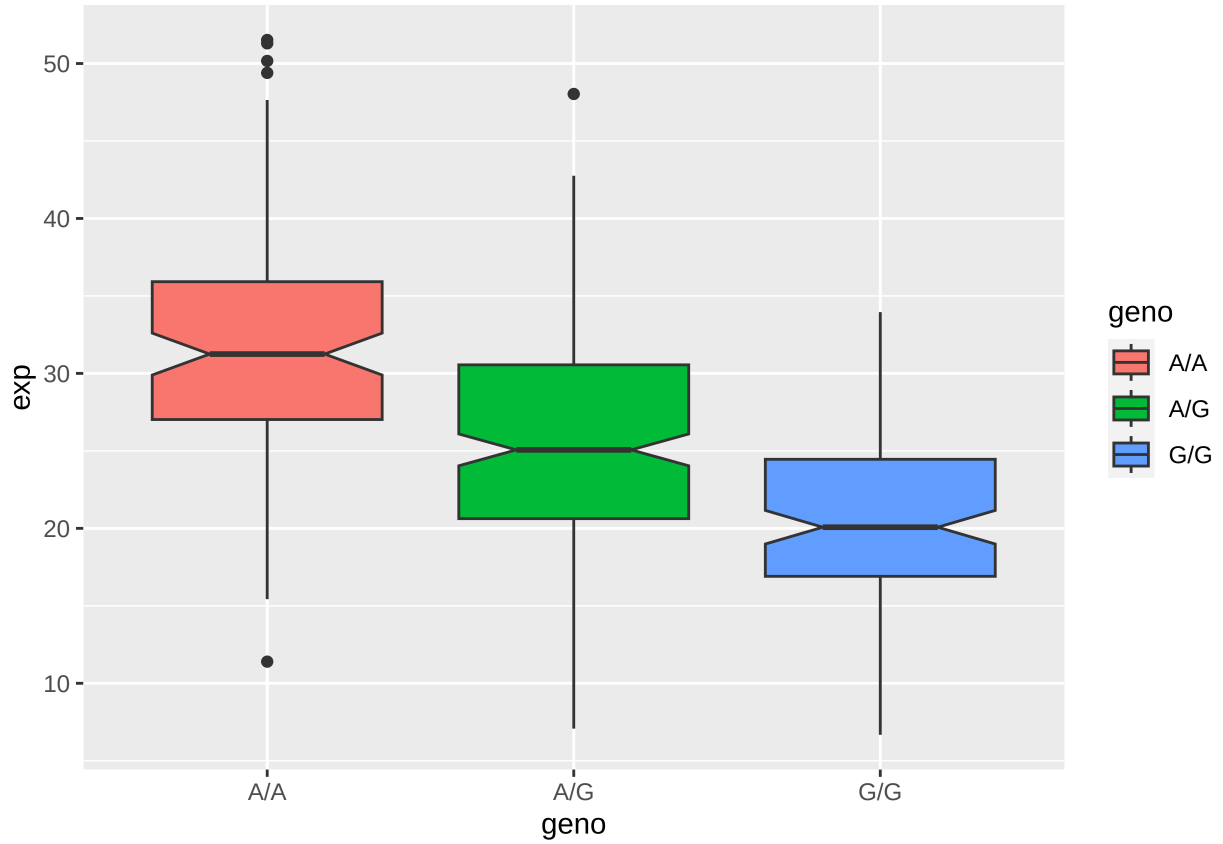
```
table(expr$geno)
```

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

```
library(ggplot2)
```

Q14. We are going to make a boxplot.

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



Looking at the boxplot per genotype, there is a higher expression of A/A genotype compared to G/G homozygous genotype. Yes, the SNP value effects the expression of ORMDL3.

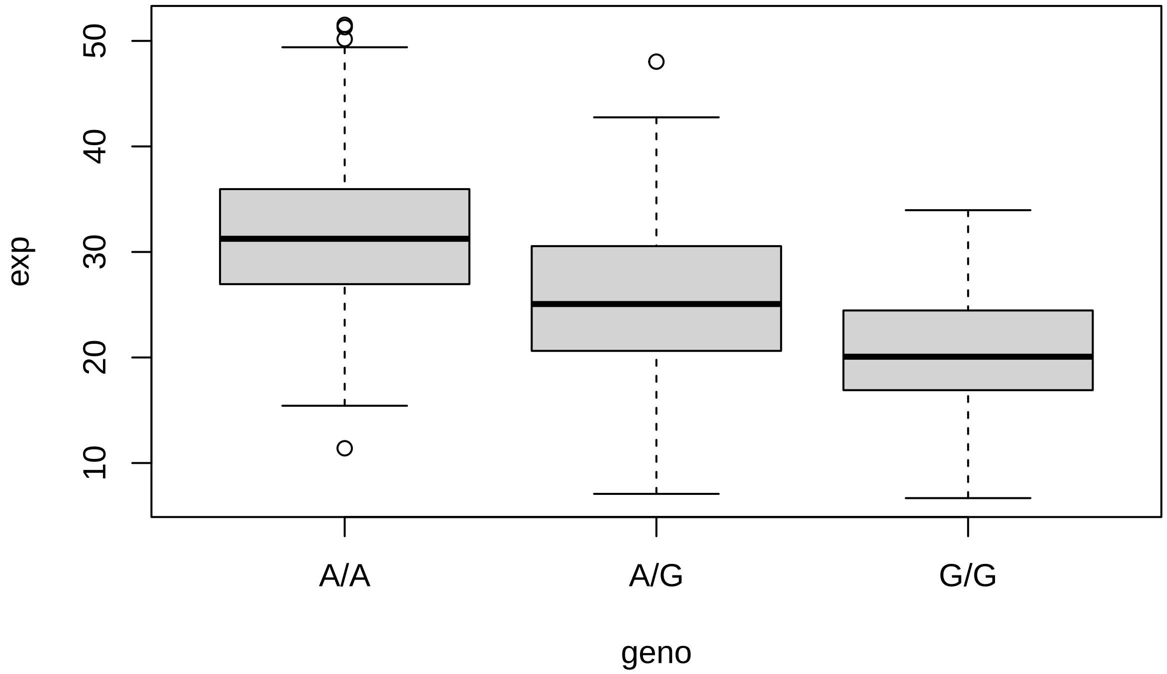
> Q13. Read this file into R and determine the sample size for each genotype and their corresponding median expression levels for each of these genotypes.

```
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
```

We are now going to make a boxplot of the data provided by the summary.

```
boxplot(exp ~ geno, data = expr)
```


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