

Week 8 Lab Session

Yash Garodia

17/02/2022

Section 1. Proportion of G/G in a population

Downloaded a csv file from ensemble < https://uswest.ensembl.org/Homo_sapiens/Variation/Sample?db=core;r=17:39580097-40210098;v=rs8067378;vdb=variation;vf=105535077#373531_tablePanel >

```
#read.csv
mxl <- read.csv("W8Genomics.csv")
#identifying number of mxls,
table(mx1$Genotype..forward.strand.)
```

```
##
## A|A A|G G|A G|G
## 22 21 12 9
```

Each person (each genotype) is a row. So we can use the nrow function to find the total number of genotypes.

```
table(mx1$Genotype..forward.strand.) / nrow(mx1)
```

```
##
##      A|A      A|G      G|A      G|G
## 0.343750 0.328125 0.187500 0.140625
```

We can do the same thing for the british population(GBR).

```
gbr <- read.csv("373522-SampleGenotypes-Homo_sapiens_Variation_Sample_rs8067378.csv")
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 asthma is more frequent in the GBR population over the MXL population. Lets now dig into this further.

Section 4: Population Scale Analysis [HOMEWORK]

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 whether there is any association of the 4 asthma-associated SNPs (rs8067378...) on ORM3 expression. This is the final file you got (https://bioboot.github.io/bgg213_W19/class-material/rs8067378_ENSG00000172057.6.txt). The first column is sample name, the second column is genotype and the third column are the expression values. Open a new RMarkdown document in RStudio to answer the following two questions. Submit your resulting PDF report with your working code, output and narrative text answering Q13 and Q14 to GradeScope.

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 median value from saving the output of the `boxplot()` function to an R object and examining this object. There is also the `median()` and `summary()` function that you can use to check your understanding.

How many samples do we have?

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

```
library(dplyr)
```

```
## Warning: package 'dplyr' was built under R version 4.0.2
```

```
##
## 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
```

```
nrow(expr)
```

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

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

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

```
expr_AA <- expr %>% filter(expr$geno == "A/A")  
expr_AG <- expr %>% filter(expr$geno == "A/G")  
expr_GG <- expr %>% filter(expr$geno == "G/G")  
median(expr_AA$exp)
```

```
## [1] 31.24847
```

```
median(expr_AG$exp)
```

```
## [1] 25.06486
```

```
median(expr_GG$exp)
```

```
## [1] 20.07363
```

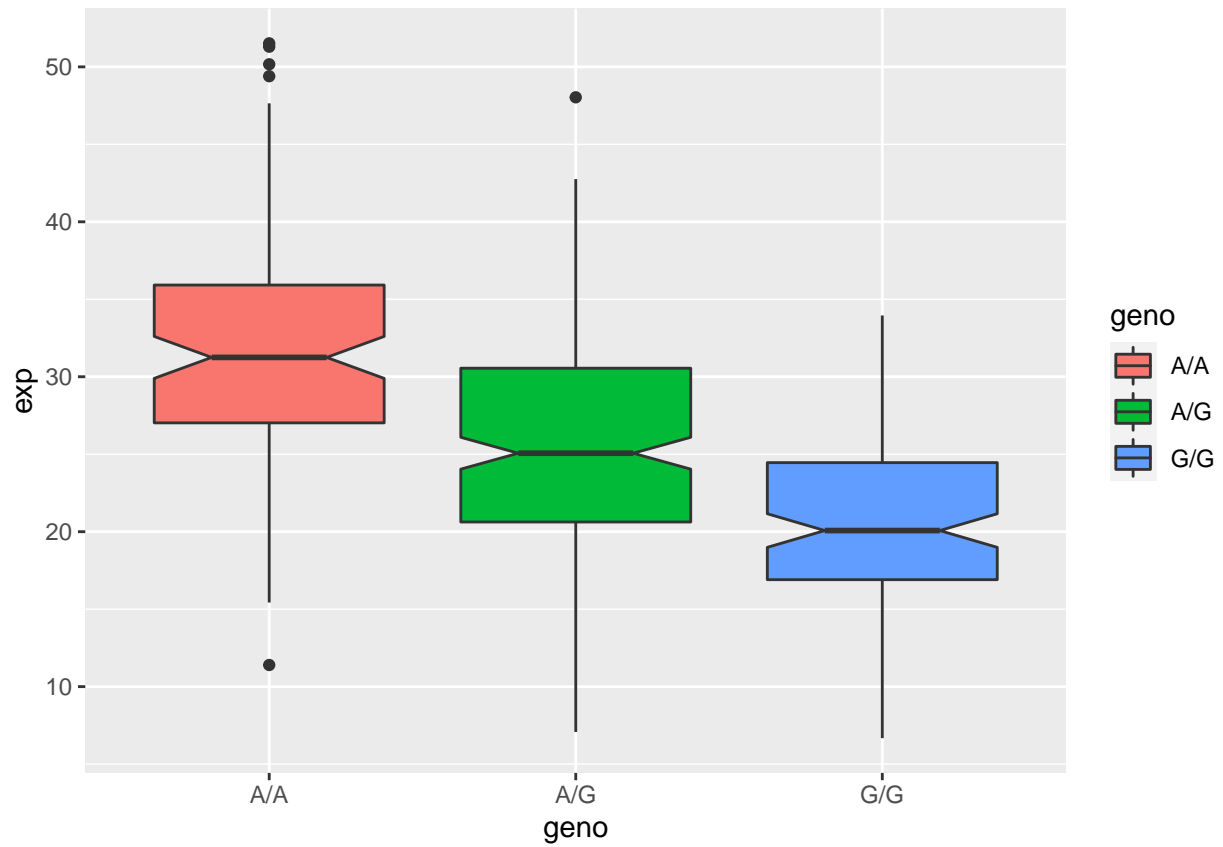
We have 108 A/A cases, 233 A/G cases and 121 G/G cases. A total of 462 cases. Median expression level of A/A is 31.248, of A/G is 25.064, of G/G is 20.074

Lets make a boxplot

```
library(ggplot2)
```

```
## Warning: package 'ggplot2' was built under R version 4.0.2
```

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



Clearly, we can see that presence of G/G in this gene is greatly associated with a reduction in gene expression.