

# Class12

Sabrina Wu

2024-11-07

#Section 1. Proportion of G/G in a population

Downloaded a CSV file from Ensembl < [https://useast.ensembl.org/Homo\\_sapiens/Variation/Sample?db=core;r=17:39894595-39895595;v=rs8067378;vdb=variation;vf=959672880#373531\\_tablePanel](https://useast.ensembl.org/Homo_sapiens/Variation/Sample?db=core;r=17:39894595-39895595;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
```

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

```
gbr <- read.csv("373522-SampleGenotypes-Homo_sapiens_Variation_Sample_rs8067378.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 18.68 26.37 29.67
```

This variant that is associated with childhood asthma is more frequent in the GBR population than the MKL population.

Lets now idg into this further.

##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.

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.

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
```

```
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
```

```
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
```

```
medians <- expr %>%
  group_by(geno) %>%
  summarize(median(exp, na.rm=TRUE))
medians
```

```
## # A tibble: 3 x 2
##   geno 'median(exp, na.rm = TRUE)'
##   <chr>          <dbl>
## 1 A/A          31.2
## 2 A/G          25.1
## 3 G/G          20.1
```

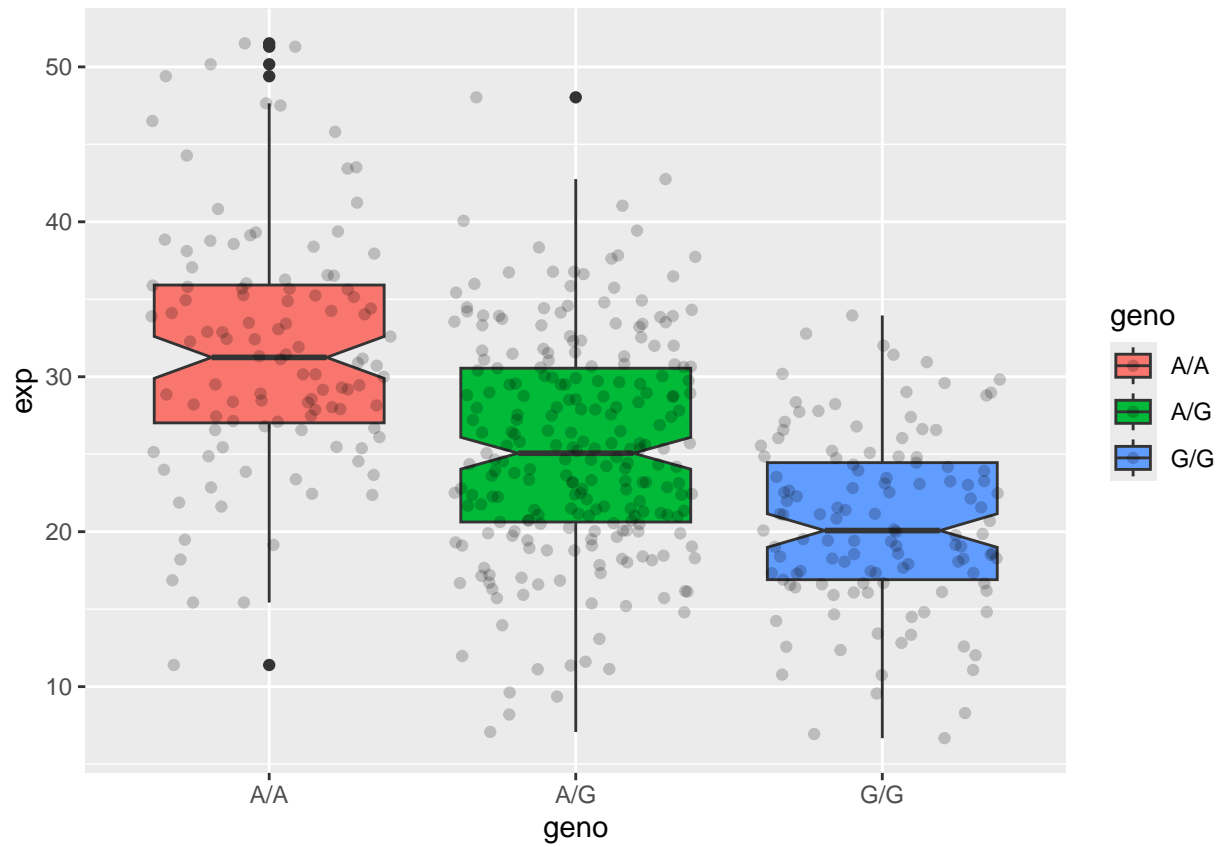
There are 462 samples total where 108 corresponding to A/A, 233 for A/G, and 121 for G/G. The median expression level of A/A is 31.24, A/G is 25.06, and G/G is 20.07.

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? Hint: An example boxplot is provided overleaf – yours does not need to be as polished as this one

```
library(ggplot2)
```

Let's make a boxplot

```
boxplot <- ggplot(expr) + aes(x=geno, y=exp, fill=geno) +
  geom_boxplot(notch=TRUE) +
  geom_jitter(alpha=0.2)
boxplot
```



There are more A/A expression than the other two types of genotypes since the median, and quartiles are higher on the boxplot. SNP does effect the expression of ORM DL3.