

# Class 12: Introduction to Genome Informatics Lab

Yu (Ericsson) Cao (PID: A16421048)

2024-02-16

## 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(mx1$Genotype..forward.strand.) / nrow(mx1) * 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. We picked the GBR.

```
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 associated with childhood asthma is more frequent in the GBR population than the MXL population.

Let's now dig 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.

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
```

```
library(ggplot2)
```

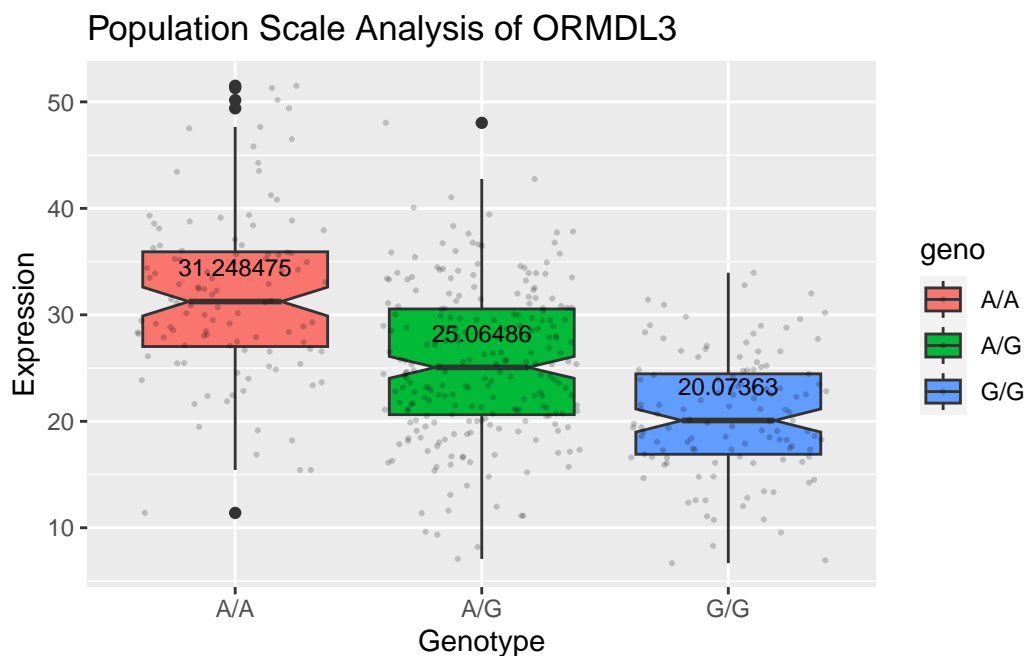
Let's make a boxplot

```
library(plyr)
```

```
p_meds <- ddply(expr, .(geno), summarise, med = median(exp)) #median displayed
```

```
ggplot(expr) + aes(geno, exp, fill=geno) +
  xlab("Genotype") +
  ylab("Expression") +
```

```
ggtitle("Population Scale Analysis of ORMDL3") +
geom_boxplot(notch=T) +
geom_jitter(color="black", size=0.4, alpha=0.2) +
geom_text(data = p_meds, aes(x = geno, y = med, label = med),
          size = 3, vjust = -1.5)
```



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.

A:

```
##Sample size of each genotype
table(expr$geno)
```

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

Their corresponding median expression levels are: A/A: 31.248, A/G: 25.065, G/G: 20.074.

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?

A: From this data, we see a trend where the presence of the G allele is linked with a decrease in gene expression, or the A allele is associated with an increase in gene expression. Individuals with the A/A genotype have the highest expression, and people with the A/G genotype have a medium level, and those with the G/G genotype have a relatively low expression.

This suggests that the SNP could indeed be affecting its expression. There is a potential dose-dependent relationship where two copies of the G allele (G/G genotype) leading to a reduction in expression compared to one copy (A/G genotype) or none (A/A genotype). Or it could be the vice versa, which an A allele leads to an increase in expression.