

# Class 18

Caitriona Brennan

#Section1 Proportin of G/G in a population

```
mxl <- read.csv("373531-SampleGenotypes-Homo_sapiens_Variation_Sample_rs8067378.csv")
head(mx1)
```

```
## 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 -
```

Use table to tell you the number of occurrences of each allele in the column

```
table(mx1$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. I picked the GBR population

```
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. 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 ORMDL3 expression.

How many samples do we have?

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

How many of each genotype?

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

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

To summarize the results in a figure lets make a boxplot

```
library(ggplot2)
```

The expression of ORMDL3 is lowered with this SNP

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

