Class 17

Joseph Girgiss (PID: A17388247)

Section 1: Proportion of G|G in a population

Downloaded a CSV file from Ensembl.

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) / 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 population.

```r
gbr <- read.csv("373522-SampleGenotypes-Homo_sapiens_Variation_Sample_rs8067378 (2).csv")

# find G|G proportion
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 explore 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? Let's read the data.

> 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.

```r
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
```

```r
nrow(expr)
```

```
[1] 462
```

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

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

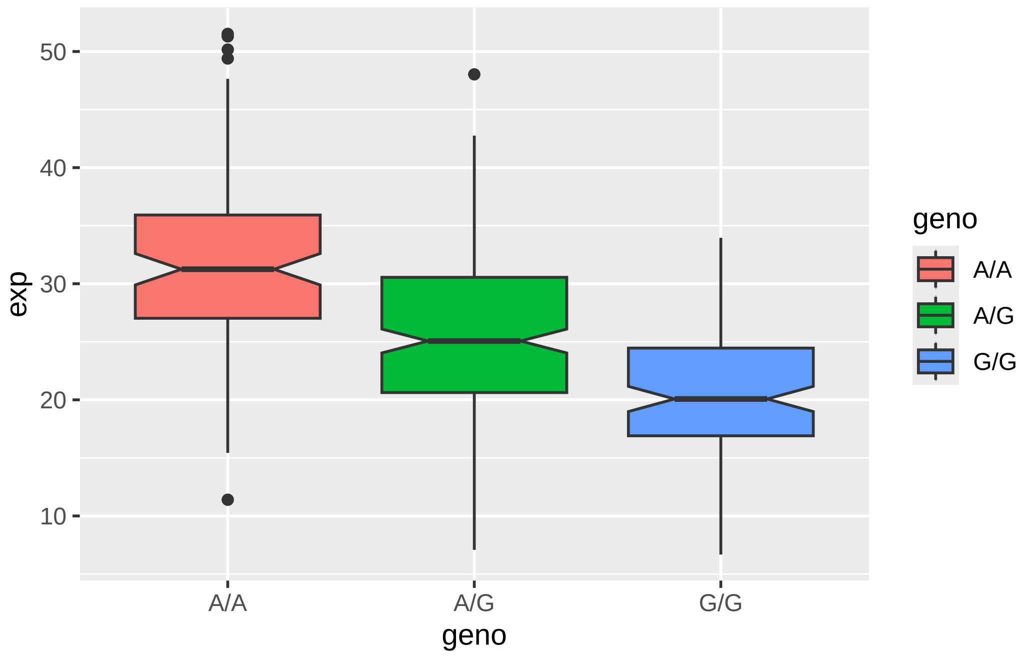
```
tapply(expr$exp, expr$geno, median)
```

```
     A/A      A/G      G/G
31.24847 25.06486 20.07363
```

Let's make a boxplot

> 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?

```
library(ggplot2)
ggplot(expr) + aes(geno, exp, fill=geno) +
  geom_boxplot(notch=TRUE)
```



The A/A genotype shows the highest expression of ORMDL3, while A/G and G/G show progressively lower expression levels. This pattern suggests a allelic effect of the SNP on gene expression, where each copy of the G allele is associated with reduced ORMDL3 expression. So, yes, SNP does effect the expression of ORMDL3.