Lab 11: Genome Informatics

Kevin Tan (PID: A16774162)

Section 1. Proportion of G/G in a population

Downloaded a CSV file from ensemble.

Here we read this CSV file:

```r
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      -
```

```r
table(mxl$Genotype..forward.strand.)
```

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

```r
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.

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

Find proportion of G|G

```r
table(gbr$Genotype..forward.strand.) / nrow(mxl) * 100
```

```
    A|A     A|G     G|A     G|G
35.9375 26.5625 37.5000 42.1875
```

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

lets now dig into this further.

Section 4: Population Scale Analysis

How many samples do we have?

```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
```

2

```
nrow(expr)
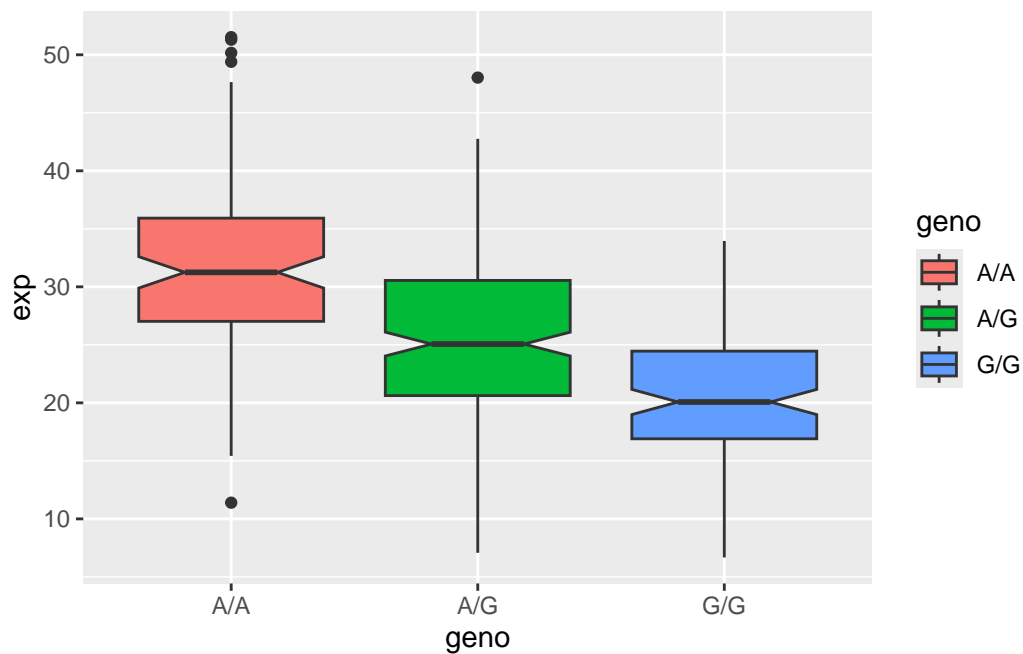```

[1] 462

Sample size of each genotype

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

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

```
library(ggplot2)
```

Lets make a boxplot

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

```r
df <- data.frame(expr)
```

```r
analysis <- function(x){
  med <- median(df$exp[df$geno == x])
  med
}
```

Median expression levels of each genotype

```r
analysis("A/A")
```

```
[1] 31.24847
```

```r
analysis("A/G")
```

```
[1] 25.06486
```

```r
analysis("G/G")
```

```
[1] 20.07363
```