Class 11: Genome Informatics and High Througput Sequencing

Aishwarya Ramesh

Section 1: Proportion of G/G in a Population

Reading 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
```

Roughly 14.06% are homozygous for asthma associated gene in MXL population.

Now let's look at a different population. Looking at GBR population.

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

FInd proportion of GG in GBR

```
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 childbood 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
```

We have 462 individuals in this 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.

Finding sample sizes

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

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

Finding median expression levels

```
gg_summary <- summary(expr$exp[expr$geno == 'G/G'])
ag_summary <- summary(expr$exp[expr$geno == 'A/G'])
aa_summary <- summary(expr$exp[expr$geno == 'A/A'])
gg_summary
```

```
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
  6.675  16.903  20.074  20.594  24.457  33.956
```

```
ag_summary
```

```
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
  7.075  20.626  25.065  25.397  30.552  48.034
```

```
aa_summary
```

```
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
  11.40   27.02   31.25   31.82   35.92   51.52
```

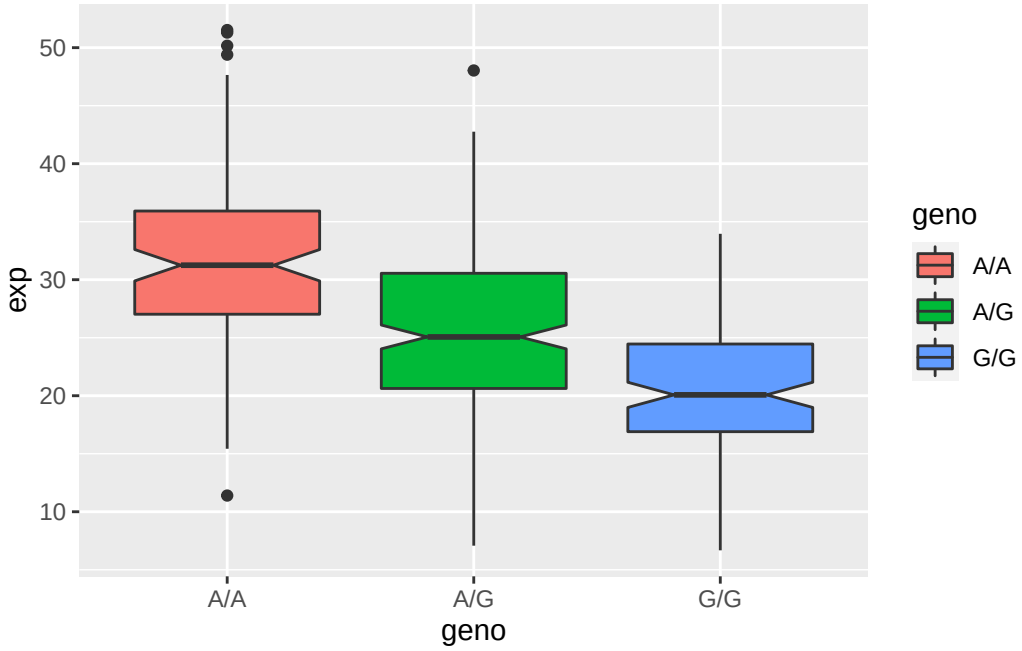
The sample size for A/A is 108, the sample size for A/G is 233, the sample size for G/G is 121. The median expression for G/G is 20.074. The median expression for A/G is 25.065. The median expression for A/A is 31.25.

```
library(ggplot2)
```

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?

Making boxplot

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



The expression is relatively higher for A/A compared to G/G. From this, we can conclude that having a G/G is associated with having a lower expression of this particular gene ORMDL3, and thus that there will be differential asthma outcomes for A/A as compared to G/G.