Class 11 Lab

Andres Sandoval

Section 1 - Proportion of G/G in a population

MXL - Mexican Ancestry in Los Angeles

Downloaded csv file from Ensemble.

Here we read the 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 function summarizes the data in the table
```

```
round(table(mxl$Genotype..forward.strand.) / nrow(mxl) *100, 2)
```

```
  A|A   A|G   G|A   G|G
34.38 32.81 18.75 14.06
```

Let's try examining another 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 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 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.

Q13

How many samples do you 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
```

How many individuals?

```
nrow(expr)
```

```
[1] 462
```

How many samples of each genotype?

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

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

BoxPlot section

```
library(ggplot2)
```

Now, you want to find whether there is any association of the 4 asthma-associated SNPs (rs8067378...) on ORMDL3 expression.

Let's make our boxplot with this data.

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

Q13. What is the median expression of each of the genotypes?

```
library(dplyr)
```

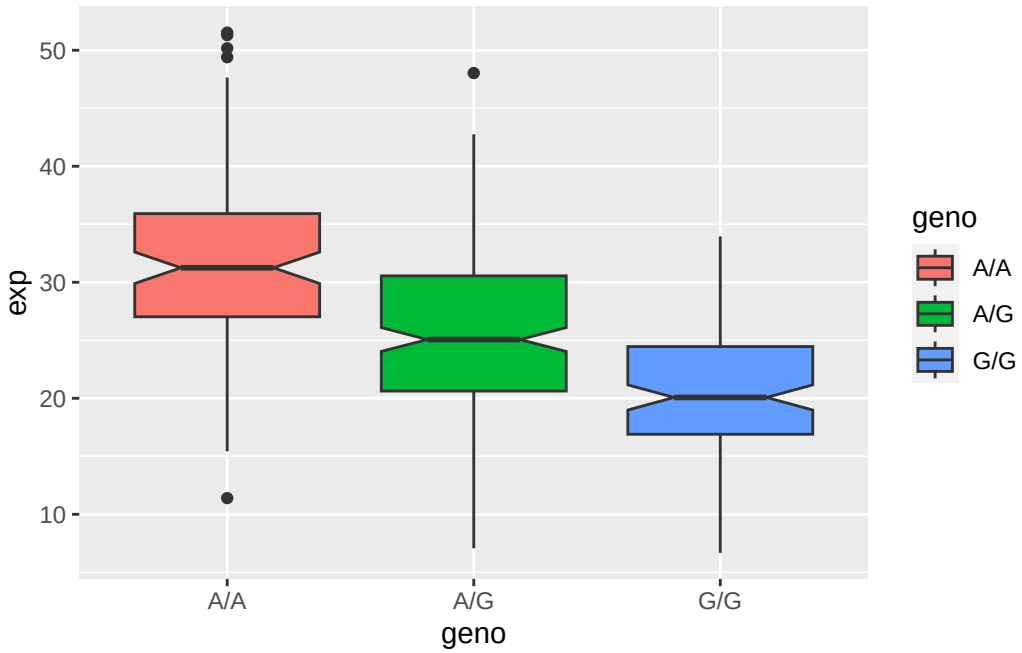
```
Attaching package: 'dplyr'

The following objects are masked from 'package:stats':

    filter, lag

The following objects are masked from 'package:base':

    intersect, setdiff, setequal, union
```

```
expr %>% group_by(geno) %>% summarize(median = median(exp))
```

```
# A tibble: 3 x 2
  geno  median
  <chr>  <dbl>
1 A/A     31.2
2 A/G     25.1
3 G/G     20.1
```

Q14. Boxplot

```
boxplot
```


50
40
30
20
10
exp
A/A
A/G
G/G
geno
geno
A/A
A/G
G/G