CLass 12

Mudit

Section 1. Proportion of G/G in a population

Downloaded a CSV file from Ensembl < https://useast.ensembl.org/Homo_sapiens/Location/View?db=core;r=39873367;v=rs12936231;vdb=variation;vf=959765854 >

Here we read this CSV file

```
mxl <- read.csv("373531-SampleGenotypes-Homo_sapiens_Variation_Sample_rs12936231.csv")

head(mxl)
```

```
  Sample..Male.Female.Unknown. Genotype..forward.strand. Population.s. Father
1                 NA19648 (F)                         C|C ALL, AMR, MXL      -
2                 NA19649 (M)                         G|G ALL, AMR, MXL      -
3                 NA19651 (F)                         C|C ALL, AMR, MXL      -
4                 NA19652 (M)                         G|G ALL, AMR, MXL      -
5                 NA19654 (F)                         G|G ALL, AMR, MXL      -
6                 NA19655 (M)                         C|G ALL, AMR, MXL      -
  Mother
1      -
2      -
3      -
4      -
5      -
6      -
```

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

```
C|C C|G G|C G|G
 22  21  12   9
```

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

```
    C|C     C|G     G|C     G|G
34.3750 32.8125 18.7500 14.0625
```

Now let's look at a different population. I picked GBR

```
gbr <- read.csv("373522-SampleGenotypes-Homo_sapiens_Variation_Sample_rs12936231.csv")

head(gbr)
```

```
  Sample..Male.Female.Unknown. Genotype..forward.strand. Population.s. Father
1                  HG00096 (M)                       C|C ALL, EUR, GBR      -
2                  HG00097 (F)                       G|C ALL, EUR, GBR      -
3                  HG00099 (F)                       G|G ALL, EUR, GBR      -
4                  HG00100 (F)                       C|C ALL, EUR, GBR      -
5                  HG00101 (M)                       C|C ALL, EUR, GBR      -
6                  HG00102 (F)                       C|C ALL, EUR, GBR      -
  Mother
1      -
2      -
3      -
4      -
5      -
6      -
```

```
table(gbr$Genotype..forward.strand.)
```

```
C|C C|G G|C G|G
 23  17  24  27
```

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

```
  C|C   C|G   G|C   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 in the MKL population.

Let's not dig into this further

Section 4: Population Scale Analysis

[HOMEWORK] 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. https://bioboot.github.io/bggn213_W19/class- material/rs8067378_ENSG00000172057.6.txt This is the final file you got (column is genotype and the third column are the expression values.). The first column is sample name, the second Open a new RMarkdown document in RStudio to answer the following two questions. Submit your resulting PDF report with your working code, output and narrative text answering Q13 and Q14 to GradeScope.

How many samples do we have?

```
expr <- read.table("Expression genotype results.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
```

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

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

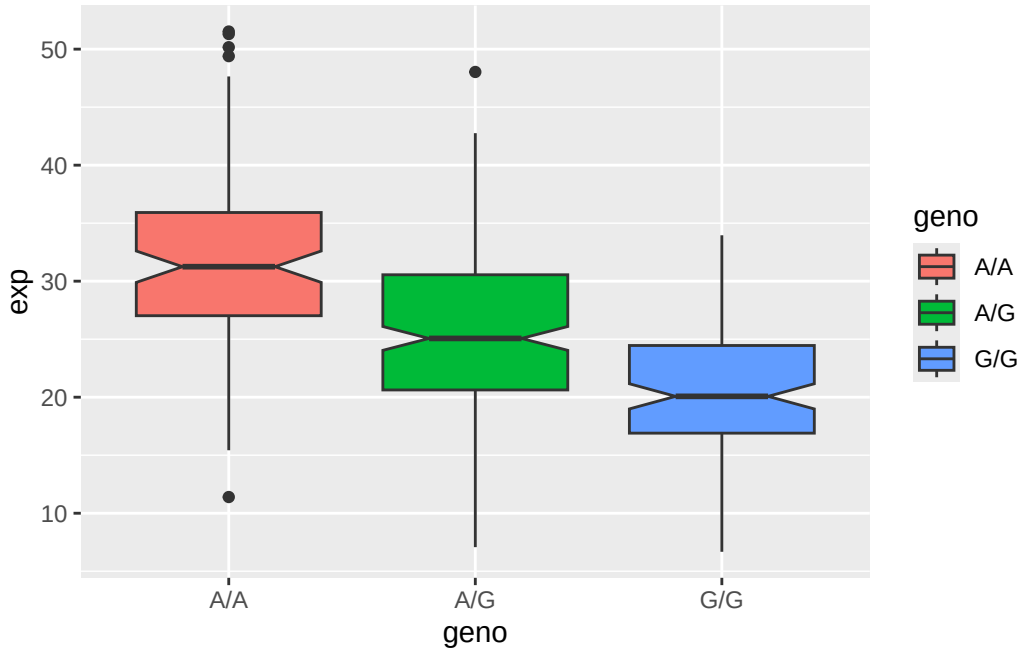
```
tapply(expr$exp, expr$geno, median)
```

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

```
library(ggplot2)
```

Lets make a boxplot

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



A/A has the highest median expression level, while G/G has the lowest.

difference in expression suggests that there may be an association between the genotype at this SNP and ORMDL3 expression levels, with A/A associated with higher expression and G/G with lower.