class11

Amit Subramanian

Section 1. Proportion of G|G in MXL population

Let's read in our csv file we downloaded from Ensemble.

```
mxl <- read.csv("373522-SampleGenotypes-Homo_sapiens_Variation_Sample_rs8067378.csv")
head(mxl)
```

```
  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      -
```

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

```
A|A A|G G|A G|G
 23  17  24  27
```

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

```
     A|A      A|G      G|A      G|G
25.27473 18.68132 26.37363 29.67033
```

Let's take a look at another population, for example: GBR.

```
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 of GBR

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

```
A|A A|G G|A G|G
 25  19  26  30
```

We now know that the variant that is associated with childhood asthma is more frequent in the GBR population than the MXL population.

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.

> 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.

```
# This will tell us how many samples we have

expr <- read.table(url("https://bioboot.github.io/bimm143_S24/class-material/rs8067378_ENS
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
```

```
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(
    sample_size = n (),
    median_expression = median (exp, na.rm=TRUE)
  )
```

```
# A tibble: 3 x 3
  geno  sample_size median_expression
  <chr>       <int>             <dbl>
1 A/A           108              31.2
2 A/G           233              25.1
3 G/G           121              20.1
```

```
nrow(expr)
```

```
[1] 462
```

We need to find how many there are of each type using 'table()'.

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

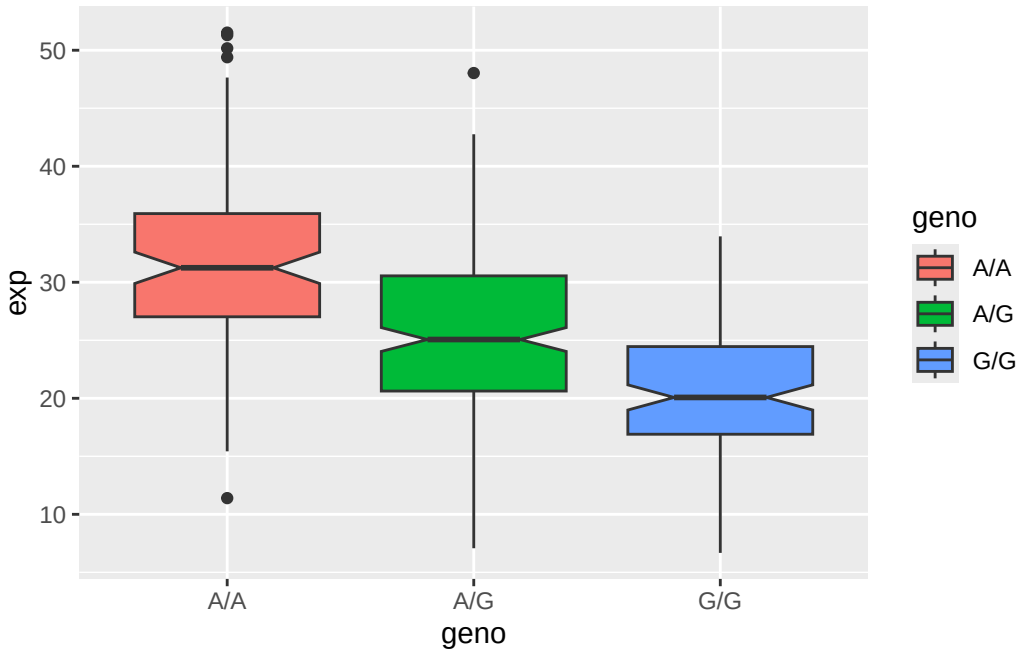
```
A/A A/G G/G
108 233 121
```

> 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?

Let's create a boxplot

```
library(ggplot2)
```

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



From this plot we are able to gather that the relative expression level of A|A is higher than the relative expression level of G|G. Additionally, the SNP does appear to affect the expression of ORMDL3.