Class 11: Genome Informatics

Kaitlyn Powell

Section 1: Proportion of G/G in population

Downloaded CSV file from emsemble < https://uswest.ensembl.org/Homo_sapiens/Variation/Sample?db=core;
39955106;v=rs8067378;vdb=variation;vf=105535077#373531_tablePanel

Here we read the CSV file:

```
mxl  <- read.csv("g:g data MXL.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 
```

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

```
     A|A      A|G      G|A      G|G 
0.343750 0.328125 0.187500 0.140625 
```

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

```r
gbr <- read.csv("g:g data GBR.csv")
```

Find proportion of G/G

```r
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      -
```

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

```
      A|A       A|G       G|A       G|G 
0.2527473 0.1868132 0.2637363 0.2967033 
```

This variant that is associcated with childhood asthma is more frequent in the GBR population.

Let’s now 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.

How many samples do we have?

```
expr <- read.table("gene expression data.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 rows are there in the dataset?

```
nrow(expr)
```

```
[1] 462
```

> 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

The sample size for the A/A genotype is 108 and the median expression level is 30. The sample size for the A/G genotype is 233 and the median expression level is 25. The sample size for the G/G genotype is 121and the median expression level is 20.

How many are there of each genotype?

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

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

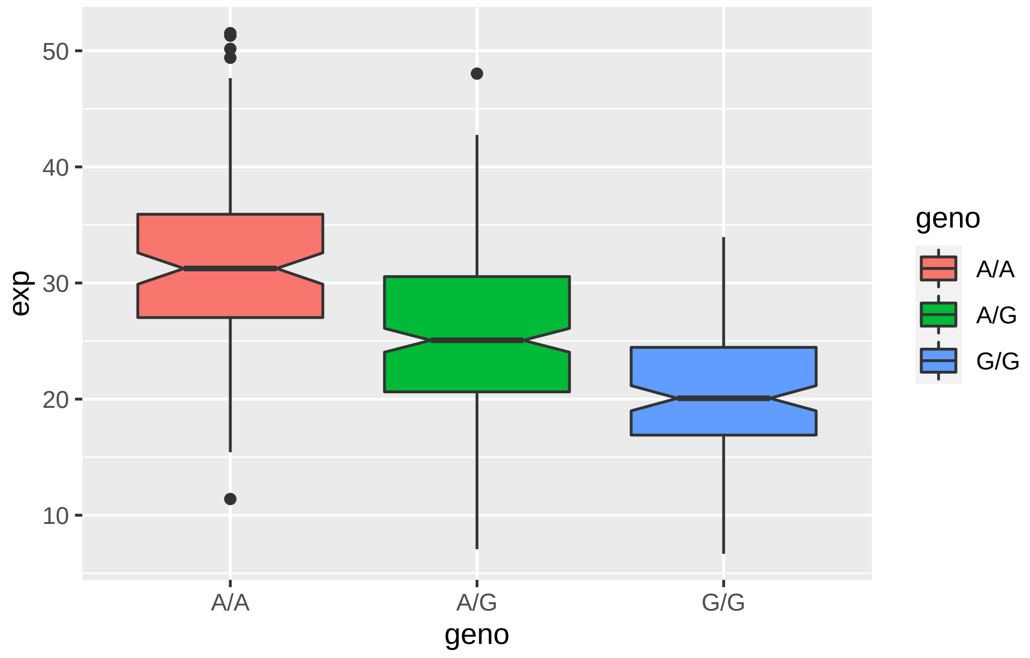
```
library(ggplot2)
```

> 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?

It can be inferred that the A/A genotype has a much higher level of expression of ORMDL3 than the G/G genotype.It does appear that the SNP does effect the expression of the ORMDL3 gene becasue the gentypes that contai nat least one A allele have a significantly higher expression level than the G/G genotype.

Let's make a boxplot.

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


exp
A/A
A/G
G/G
geno
geno
A/A
A/G
G/G