Class18: Genome Informatics

Chloe J. Welch

11/30/2021

Section 1: Proportion of G/G in a Population

First, we began by downloading a CSV file from Ensembl: < https://uswest.ensembl.org/Homo_sapiens/Variation/Sample?db=core;r=17:39835097-39955098;v=rs8067378;vdb=variation;vf=105535077#373531_tablePanel >

We will now read the 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      -
```

What is the proportion of G/G?

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

```
##
##     A|A     A|G     G|A     G|G
## 34.3750 32.8125 18.7500 14.0625
```

Let's compare to another group. We will now download the data for the GBR (Great Britain) population: < https://uswest.ensembl.org/Homo_sapiens/Variation/Sample?db=core;r=17:39835097-39955098;v=rs8067378;vdb=variation;vf=105535077#373522_tablePanel >

```
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      -
```

What is the proportion of G/G?

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

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

The variant that is associated with childhood asthma is more frequent in the GBR population than in the MXL population. We will now explore this further.

Sections 2 and 3 were completed using Galaxy and the UCSC genome browser.

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?

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.

```
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
```

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

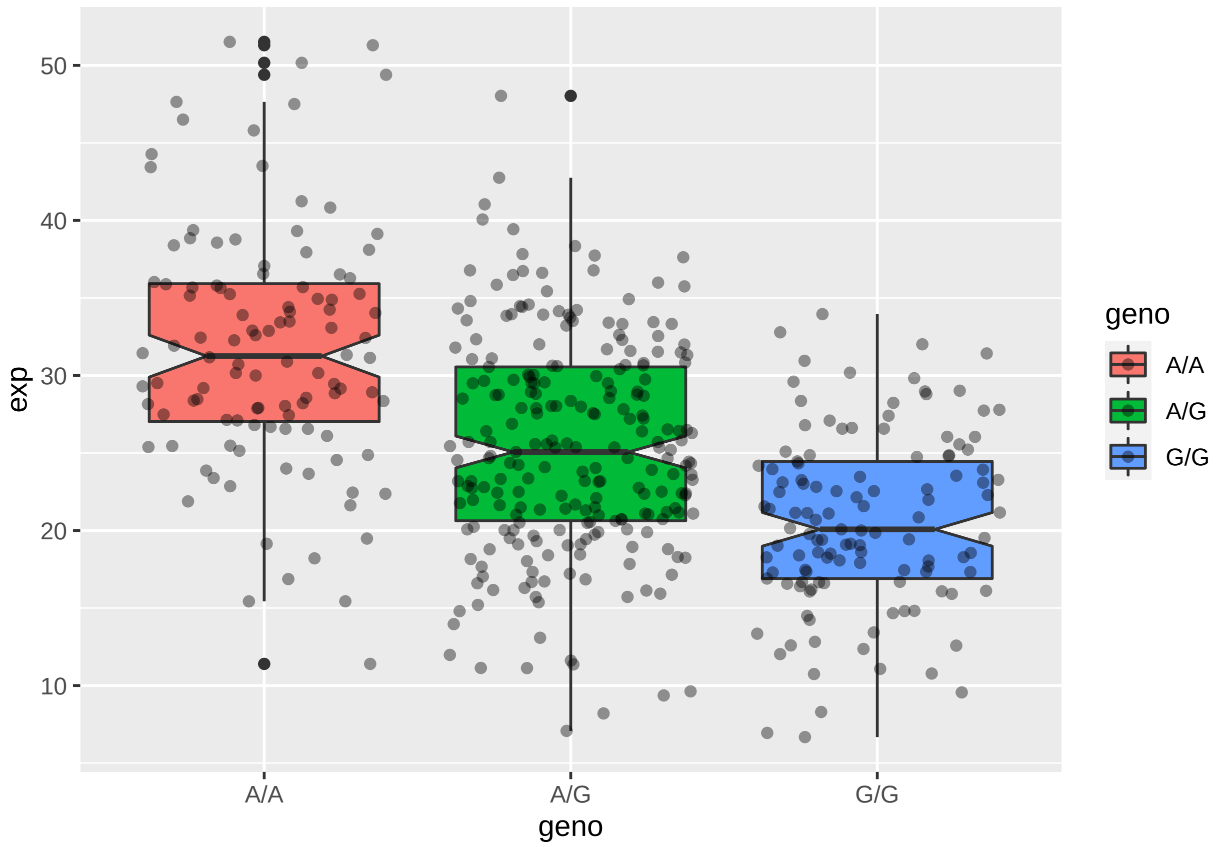
```
##
## A/A A/G G/G
## 108 233 121
```

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 affect the expression of ORMDL3?

Let's call `ggplot` so we can generate a summary figure.

```
library(ggplot2)
```

```
ggplot(expr) + aes(geno, exp, fill = geno) +
  geom_boxplot(notch = TRUE) +
  geom_jitter(alpha = 0.4)
```



Based on this plot, we can infer that the relative expression value of A/A is higher than that of G/G. Because we see variation in the expression levels between the three different genotypes, there could be association between the asthma-related SNPs and the ORMDL3 gene due to the genetic changes resulting in these different genotypes.