Class 12: Sample Genomics

Christopher Levinger (A17390693)

Table of contents

Section 1: Proportion of Population

SampleGenotypes-Homo_sapiens_Variation_Sample_rs8067378.csv

Here we read the CSV file and determine the allele frequency.

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

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

Now let's look at another population for Great Britain.

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

Find 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 is more frequent among the GBR pop than the MXL.

Let's now dig 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. 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. How many samples do we 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
```

```
nrow(expr)
```

```
[1] 462
```

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

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

The sample size for the A/A genotype is 108, for A/G it is 233, and for G/G is is 121 as seen above.

```
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 %>%
  filter(geno == "G/G") %>%
  summarise(median_expression = median(exp))
```

```
  median_expression
1          20.07363
```

Answering the secon part of Question 13, the median expression level for the G/G genotype is 20.07363.

```
library(dplyr)

expr %>%
  filter(geno == "A/G") %>%
  summarise(median_expression = median(exp))
```

```
  median_expression
1          25.06486
```

The median expression level for the A/G heterozygote is 25.06486.

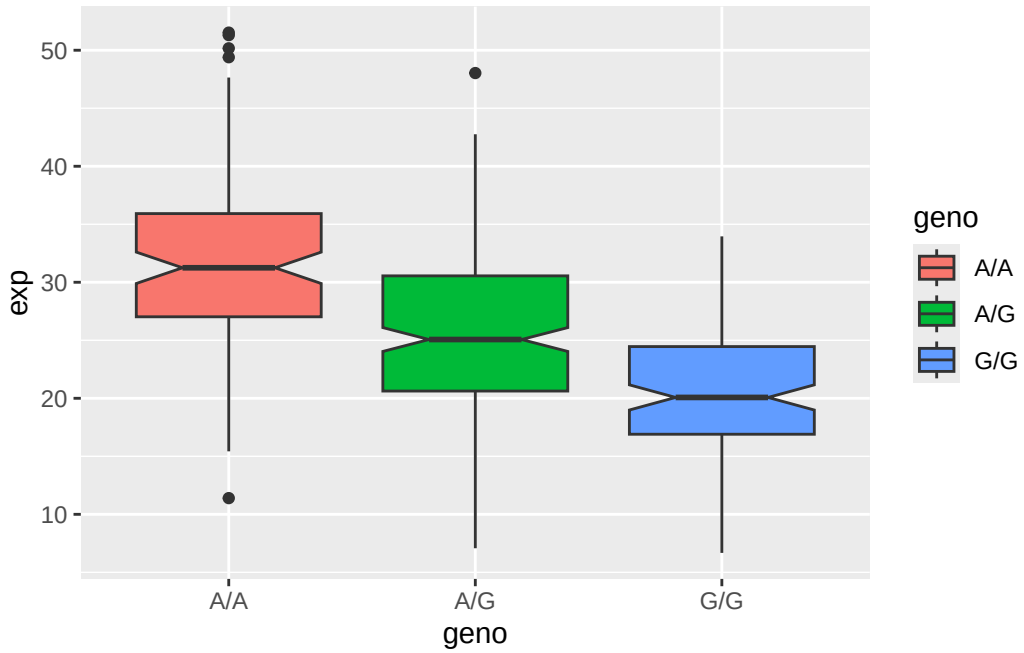
```
library(dplyr)

expr %>%
  filter(geno == "A/A") %>%
  summarise(median_expression = median(exp))
```

```
  median_expression
1          31.24847
```

The median expression level for the A/A homozygote non-asthma genotype is 31.24847.

> 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 make a boxplot with this data.

```
library(ggplot2)
ggplot(expr) + aes(geno, exp, fill=geno) +
  geom_boxplot(notch=TRUE)
```



We can infer from this boxplot that there are noticeable differences in the expression level of the key ORMDL3 marker of asthma between the A/A genotype and the G/G asthma genotype, whereby expression of this gene is significantly downregulated for this mutation leading to change in genotype. Looking at the relative medians, there is a difference of roughly 10 in the value for the expression of this gene between the two genotypes and the maximum value for the G/G data pool is only slightly above the median for the A/A homozygote. Thus, we can infer a significant decrease in the relative expression of ORMDL3 between the A/A and the G/G genotype and that such change of allele must clearly be related to ORMDL3 regulation and thereby potentially implicated in asthma. Clearly single nucleotide polymorphisms(SNP) affect the expression of the ORMDL3 potential marker for asthma. We see for this variable region of Chromosome 17 discussed previously that the normal A/A homozygote has the highest ORMDL3 expression level and that further mutations at this site result in continual decrease in its expression. For example, the change to the heterozygote A/G, as discussed more quantitatively in Question 13 results in a decrease in median expression level from 31.2 roughly to 25.1, yielding a difference of 6.1. With a further mutation to the G/G homozygote most implicated in the asthma pathology, we see the median expression level drop even further to an approximate 20.1. This a 5 point median decrease from the previous heterozygote and a 11.1 median decrease from the original homozygote of A/A. Thus, clearly for this SNP, which is most closely associated with the variant rs8067378, changes at this locus in the chromosome, especially when greater in quantity, have significant effects in dropping the expression of ORMDL3. It can implied that reduced expression from the SNP/point mutation to the G allele may be implicated in certain biological mechanisms driving asthma suseptibility.