Class 12: Genomic Information

Emily Chen (PID:A16925878)

Sequence 1. Proportion of G/G in a population

Downloaded CSV file from Ensemble < https://useast.ensembl.org/Homo_sapiens/Variation/Sample?db=core;r
40054336;v=rs8067378;vdb=variation;vf=959672880#373531_tablePanel

Here we read this CSV 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      -
```

```
mxl$Genotype..forward.strand.
```

```
 [1] "A|A" "G|G" "A|A" "G|G" "G|G" "A|G" "A|G" "A|A" "A|G" "A|A" "G|A" "A|A"
[13] "A|A" "G|G" "A|A" "A|G" "A|G" "A|G" "A|G" "G|A" "A|G" "G|G" "G|G" "G|A"
[25] "G|G" "A|G" "A|A" "A|A" "A|G" "A|A" "A|G" "G|A" "G|G" "A|A" "A|A" "A|A"
```

```
[37] "G|A" "A|G" "A|G" "A|G" "A|A" "G|A" "A|G" "G|A" "G|A" "A|A" "A|A" "A|G"
[49] "A|A" "A|A" "A|G" "A|G" "A|A" "G|A" "A|A" "G|A" "A|G" "A|A" "G|A" "A|G"
[61] "G|G" "A|A" "G|A" "A|G"
```

we can use the `table()` to generate a table that will tell us the numbers of the different genotypes

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

```
A|A A|G G|A G|G 
 22  21  12   9 
```

To find the proportions we have to divide it but the number of rows. We would multiple it b y 100 to get a percentage

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

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

14% of the MXL community have the genotype G|G

Now let's look at a different population. We will be looking at the British in England and Scotland

```
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      -
```

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

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

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

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

~30% of the GBR community have the genotype G|G

This variant that is associated with childhood asthma is more frequent in the GBR population than thre MKL population.

Let's now dig into 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.

How many sampels do we have?

```r
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
```

```r
nrow(expr)
```

```
[1] 462
```

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

```

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

```r
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
```

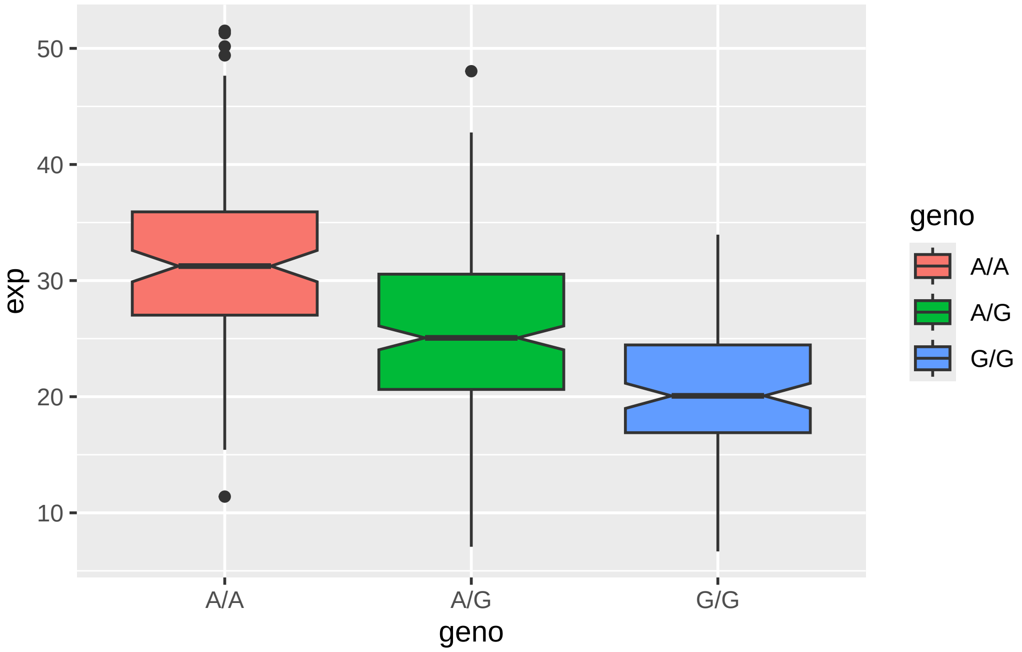
```r
expr %>%
  group_by(geno) %>%
  summarise(Median_Exp = median(exp))
```

```
# A tibble: 3 x 2
  geno  Median_Exp
  <chr>      <dbl>
1 A/A         31.2
2 A/G         25.1
3 G/G         20.1
```

> Q13. What is the sample size for each genotype and their corresponding median expression levels for each of these genotypes?

There are 462 samples in this data set, with 108 having the genotype A/A, 233 samples having genotype A/G, and 121 having the genotype G/G. The median expression of genotype A/A is 31.25, 25 for genotype A/G, and 20.1 for genotype G/G.

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



```
expr %>%
  group_by(geno) %>%
  summarise(Avg_Exp = mean(exp))
```

```
# A tibble: 3 x 2
  geno  Avg_Exp
  <chr>   <dbl>
1 A/A      31.8
2 A/G      25.4
3 G/G      20.6
```

> 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?

Looking at the box plot we can see that genotype A|A has a higher relative expression value than genotype G|G. If we were to look at the mean relative expression value for both these genotypes, A|A has a higher means than G|G. Yes this SNP effect the expression level of ORMDL3 because as we can see the individuals with the genotype A|A have a greater expression of ORMDL3 than individuals with the genotype G|G.