Class11

Ryan Fong

2022-11-02

Section 1. Proportion on G/G in a population

Download a CSV file from Ensemble < https://uswest.ensembl.org/Homo_sapiens/Variation/Sample?db=core;r=17:39894595-39895595;v=rs8067378;vdb=variation;vf=105535077;sample=mxl#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      -
```

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

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

```
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.

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

Find proportion of G|G

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

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

The variant that is associated with childhood asthma is more frequent in GBR population than in MXL population.

Section 4: Population Scale Analysis

> 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
```

We want to check total genotype

```
nrow(expr)
```

```
## [1] 462
```

See how many of each type

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

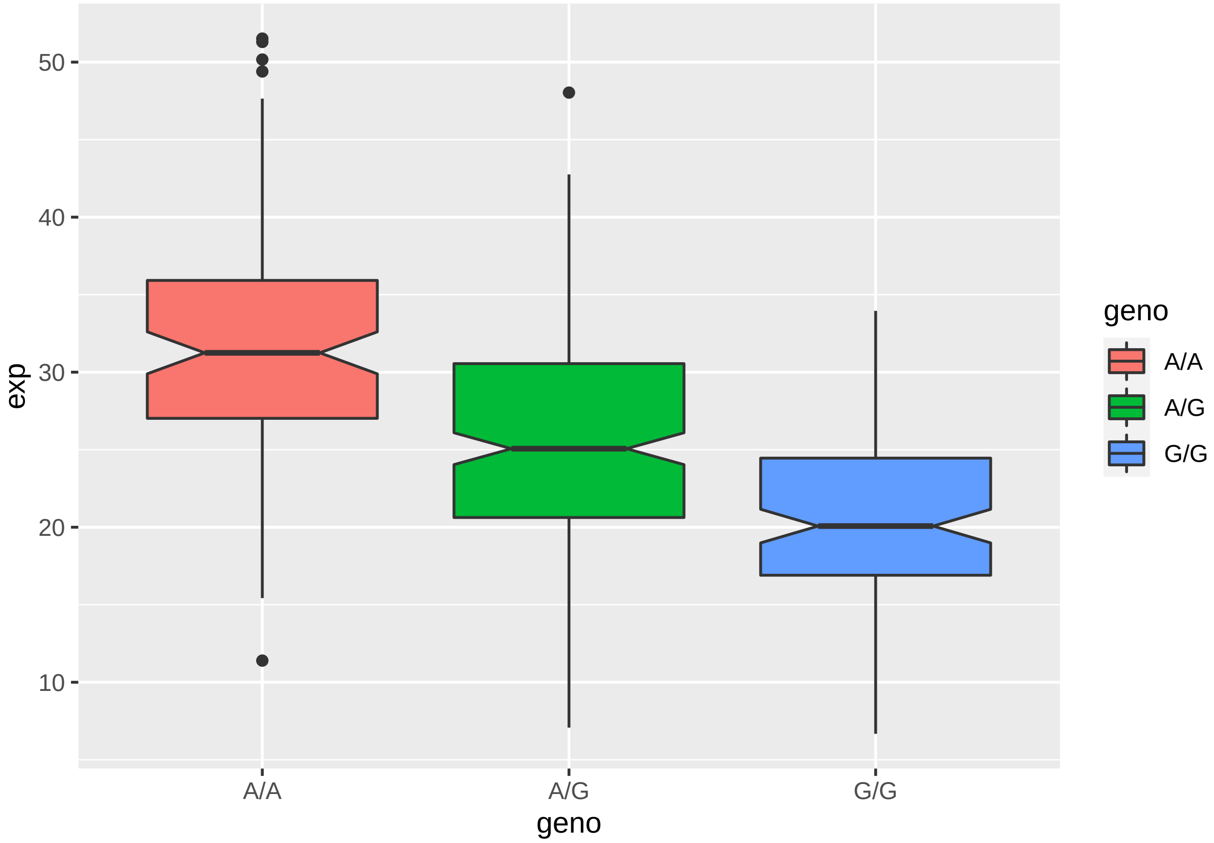
```
##
## A/A A/G G/G
## 108 233 121
```

ggplot2 wil first be loaded

```
library(ggplot2)
```

Now a boxplot will be made

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



From the boxplot, the median expression levels for the A/A genotype is roughly 32. For the A/G genotype, it is 25. For the G/G genotype it is 20

> 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?

I can infer from this relative expression value between A/A and G/G in this plot is that A/A has more expression when compared to G/G. SNP does have an effect on the expression of ORMDL3 because there are increases and decreases in expression in other genese from SNP.