HW Class12 Pt.2 (Population analysis) [Q13 Q14 BoxPlot]

Nhi To (A18053310)

2024-02-17

Section 1. Proproportion og G/G in a population

Download a CSV file from Ensemble < https://useast.ensembl.org/Homo_sapiens/Variation/Sample?db=core;r=
39960098;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      -
```

```
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 a 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
```

This variant that is associated with childhood asthma is more frequent in the GBR population with the 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. 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. Hint: The read.table(), summary() and boxplot() functions will likely be useful here. There is an example R script online to be used ONLY if you are struggling in vein. Note that you can find the medium value from saving the output of the boxplot() function to an R object and examining this object. There is also the medium() and summary() function that you can use to check your understanding.

ANSWER: The sample size for A/A genotype is 108. The sample size for A/G genotype is 233. The sample size for G/G genotype is 121. The corresponding median expression level for A/A genotype is 31.24847, A/G is 25.06486, and G/G is 20.07363.

```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
genotype_median <- aggregate(exp ~ geno, data = expr, FUN = median)
genotype_median
```

```
  geno      exp
1  A/A 31.24847
2  A/G 25.06486
3  G/G 20.07363
```

```r
summary(expr)
```

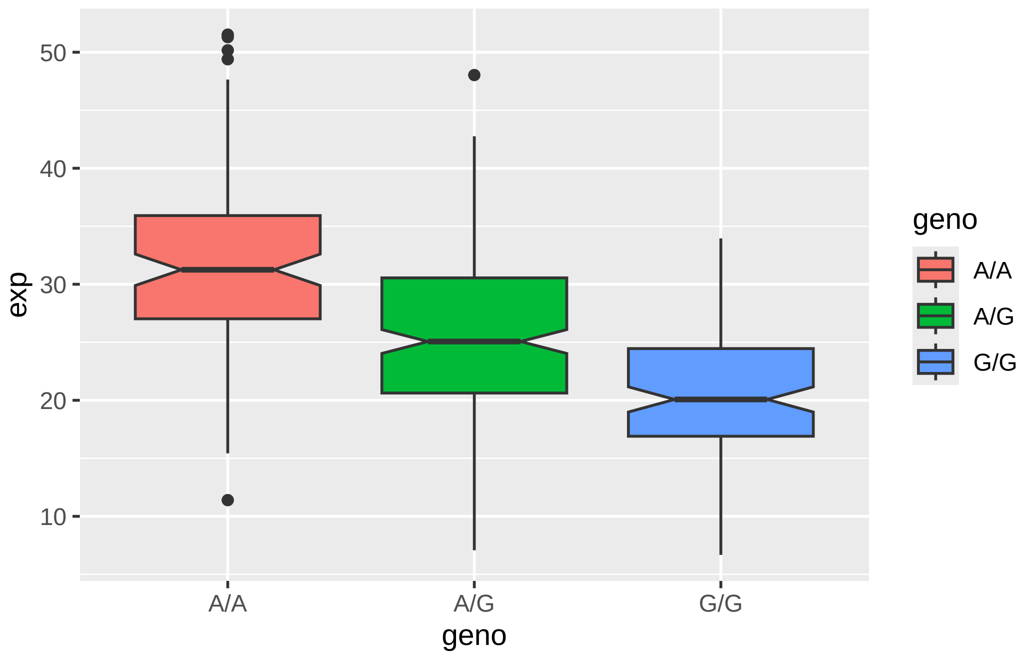
```
    sample              geno                exp        
 Length:462         Length:462         Min.   : 6.675  
 Class :character   Class :character   1st Qu.:20.004  
 Mode  :character   Mode  :character   Median :25.116  
                                       Mean   :25.640  
                                       3rd Qu.:30.779  
                                       Max.   :51.518  
```

> 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?
>
> **ANSWER**: The boxplot with a box per genotype is generated below. From the boxplot, we can visually see that the A/A genotype has the highest median expression of ORMDL3, while G/G has the lowest. Since the expression of ORMDL3 is lowest when the genotype is G/G, we can infer that the G allele is associated with the decrease of ORMDL3 expression. Yes, the SNP (single nucleotide polymorphism), has an effect on the expression of ORMDL3. Having an A allele will have some type of increasing effect on the expression of ORMDL3, while the G allele will have a decreasing effect. This suggests that SNP regulate the expression of ORMDL3.

```r
library(ggplot2)
```

Let's make a boxplot

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


50
40
30
20
10
exp
A/A
A/G
G/G
geno
geno
A/A
A/G
G/G