class12

Jaewon Kim

Section 1. Proportion of G/G in a population

We downloaded file from https://useast.ensembl.org/Homo_sapiens/Variation/Sample?db=core;g=ENSG00000073605;r=17:39904595-39919854;v=rs8067378;vdb=variation;vf=959672880#373531_tablePanel

Here we read 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.) / nrow(mxl)
```

```
     A|A      A|G      G|A      G|G 
0.343750 0.328125 0.187500 0.140625 
```

Now let's look at different population GBR.

```
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      -
```

```
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 than the MKL population.

Section 4. Let's analyze gene expression

Q.13 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")

nrow(expr)
```

```
[1] 462
```

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

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

```
calc_mean <- function (input) {
  for (i in 1:length(input)) {
    filter <- input[i]
    x <- expr$geno == filter
    print(paste(filter, ":", median(expr[x, "exp"])))
  }
}

input <- c("A/A", "A/G", "G/G")
calc_mean(input)
```

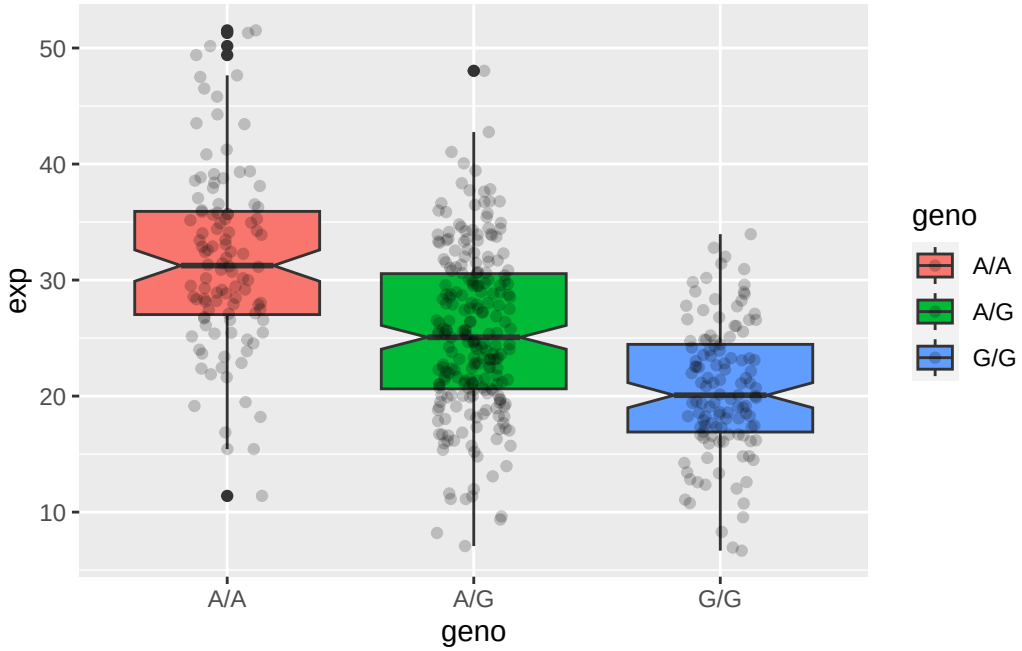
```
[1] "A/A : 31.248475"
[1] "A/G : 25.06486"
[1] "G/G : 20.07363"
```

There are total 462 samples, where 108 were "A/A", 233 were "A/G", and 121 were "G/G". Median expressions were 31.25, 25.06, 20.07, respectively.

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?

```
library(ggplot2)

ggplot(expr) +
  aes(x = geno, y = exp, fill = geno) +
  geom_boxplot(notch = TRUE) +
  geom_point(alpha = 0.2, position = position_jitter(w = 0.15))
```



Lower quartile of A/A (wt) is higher than upper quartile of G/G (mt) expression and median of A/A is higher than G/G, meaning that SNP does effect the expression of ORMDL3. Heterogeneous A/G having expression level between A/A and G/G further supports that SNP from A to G supress gene expression.