Class 12: Genome Informatics

Raquel Gonzalez (PID: A16207442)

2024-02-16

Section 1. Proportion og G/G in a population

Downloaded a CSV file from Ensemble < https://useast.ensembl.org/Homo_sapiens/Variation/Sample?db=core;r=17:39830003-39921005;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) * 100
```

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

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 thre is any association of the 4 asthma-associated SNPs (rs8067378…) on ORMDL3 expression.

How many samples do we have?

```
expr <- read.table("expression_genotype_results.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
```

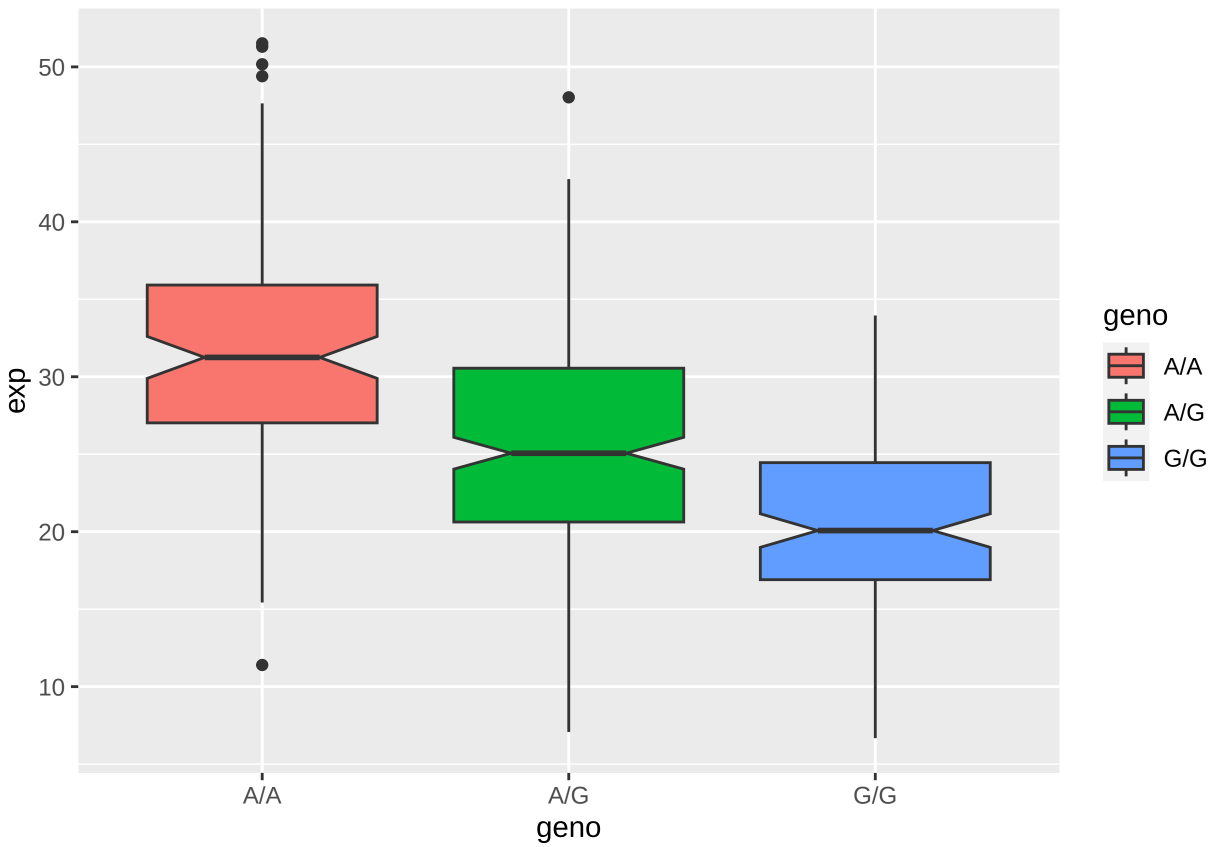
```
medians <- aggregate(exp ~ geno, data = expr, FUN = median)
medians
```

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

The sample size for A/A is 108 individuals, the sample size for A/G is 233 individuals, and the sample size for G/G is 121 individals. The median expression levels for A/A is 31.25, for A/G is 25.06, and for G/G is 20.07.

Let's make a boxplot to communicate our results.

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



The relative expression level from A/A suggests an increased expression of this gene, whereas the one from G/G suggests a decreased expression of this gene. The SNP does impact the expression of ORMDL3.