Lab11_inclass

Qianqian Tao

2023-05-16

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=17:39815101-39975102;v=rs8067378;vdb=variation;vf=105535077;sample=HG00109#373531_tablePanel (https://useast.ensembl.org/Homo_sapiens/Variation/Sample?db=core;r=17:39815101-39975102;v=rs8067378;vdb=variation;vf=105535077;sample=HG00109#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")
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.

Let's see if G|G versus A|G really affect the expression level of the gene 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.

How many samples do we have?

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

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

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

```
nrow(expr)
```

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

```
library(tidyverse)
```

```
## ── Attaching core tidyverse packages ──────────────────────── tidyverse 2.0.0 ──
## ✓ dplyr     1.0.8     ✓ readr     2.1.2
## ✓ forcats   1.0.0     ✓ stringr   1.4.0
## ✓ ggplot2   3.3.5     ✓ tibble    3.1.6
## ✓ lubridate 1.8.0     ✓ tidyr     1.2.0
## ✓ purrr     0.3.4
## ── Conflicts ────────────────────────────────────────── tidyverse_conflicts() ──
## x dplyr::filter() masks stats::filter()
## x dplyr::lag()    masks stats::lag()
## i Use the [conflicted package](http://conflicted.r-lib.org/) to force all conflicts t
o become errors
```

```
expr %>%
  group_by(geno) %>%
  summarize(medianExp = median(exp))
```

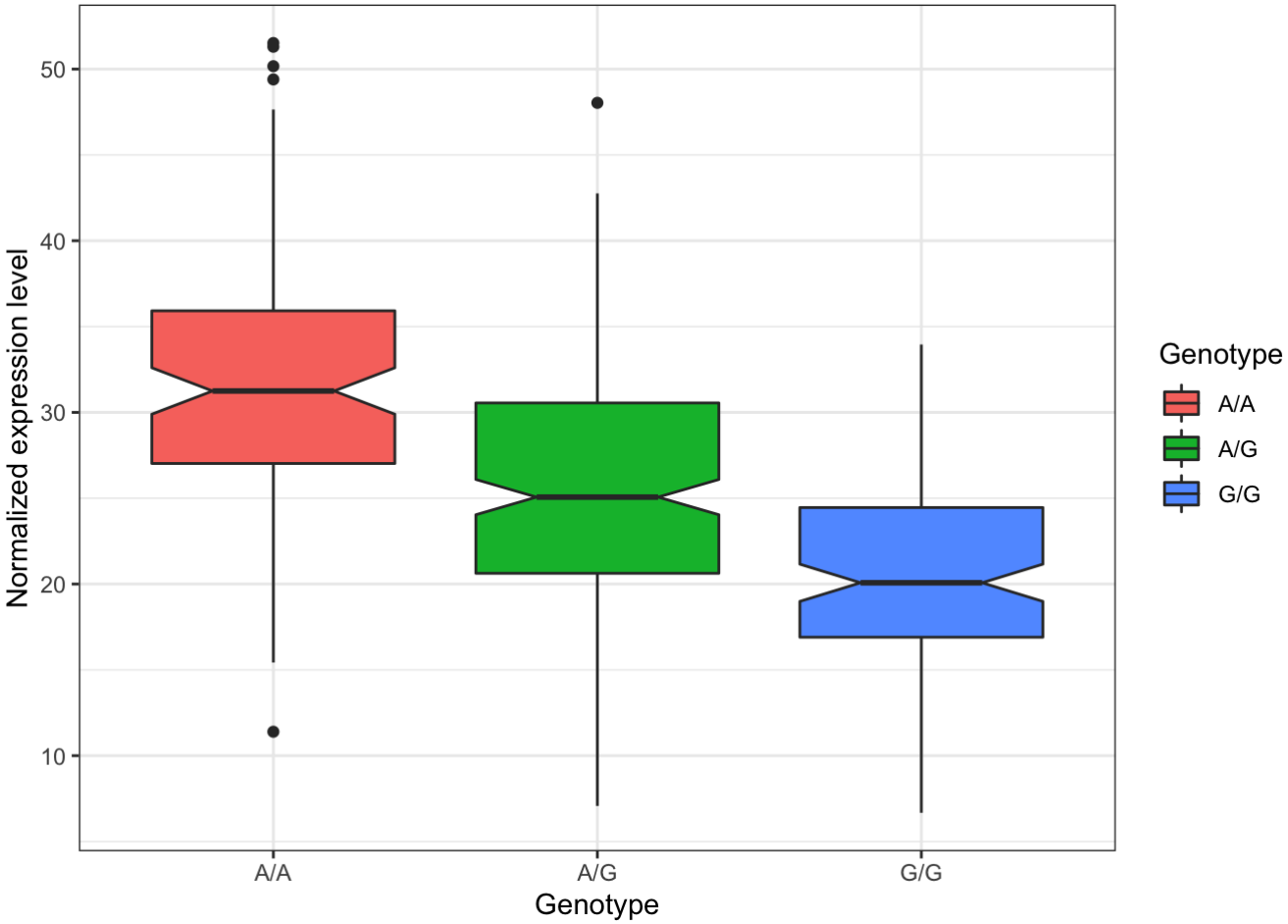
```
## # A tibble: 3 × 2
##   geno  medianExp
##   <chr>     <dbl>
## 1 A/A        31.2
## 2 A/G        25.1
## 3 G/G        20.1
```

We have 462 samples including 108 for A/A, 233 for A/G and 121 for G/G. The median expression level for A/A is 31.24847, A/G is 25.06486, and G/G for 20.07363.

> 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?

Let's make a boxplot

```
library(ggplot2)
ggplot(expr, aes(geno, exp, fill=geno))+
  geom_boxplot(notch=TRUE)+
  labs(x="Genotype", y="Normalized expression level", fill="Genotype")+
  theme_bw()
```



According to the boxplot, G/G is associated with lower gene expression levels compared to A/A. Therefore, we could conclude that this G/G SNP reduces the expression level of the gene *ORMDL3*.