Class 10 HW

Vivian Cai

2/20/2022

Section 4: Population Scale Analysis [HOMEWORK]

```
# Read the data into R using the url on the class website
data = read.table("https://bioboot.github.io/bggn213_W22/class-material/rs8067378_ENSG00000172057.6.txt

# Take a look
summary(data)
```

```
##     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
```

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.

As shown below, there are **108 A|A, 233 A|G, and 121 G|G** genotypes in the dataset, and their respective median expressions are **31.2, 25.1, and 20.1** FPKM.

```
library(dplyr)
```

```
##
## Attaching package: 'dplyr'

## The following objects are masked from 'package:stats':
##
##     filter, lag

## The following objects are masked from 'package:base':
##
##     intersect, setdiff, setequal, union
```

```
data %>%
  group_by(geno) %>%
  summarise(length(sample), median(exp))
```

```
## # A tibble: 3 x 3
##   geno  `length(sample)` `median(exp)`
##   <chr>            <int>         <dbl>
## 1 A/A                108          31.2
## 2 A/G                233          25.1
## 3 G/G                121          20.1
```

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?

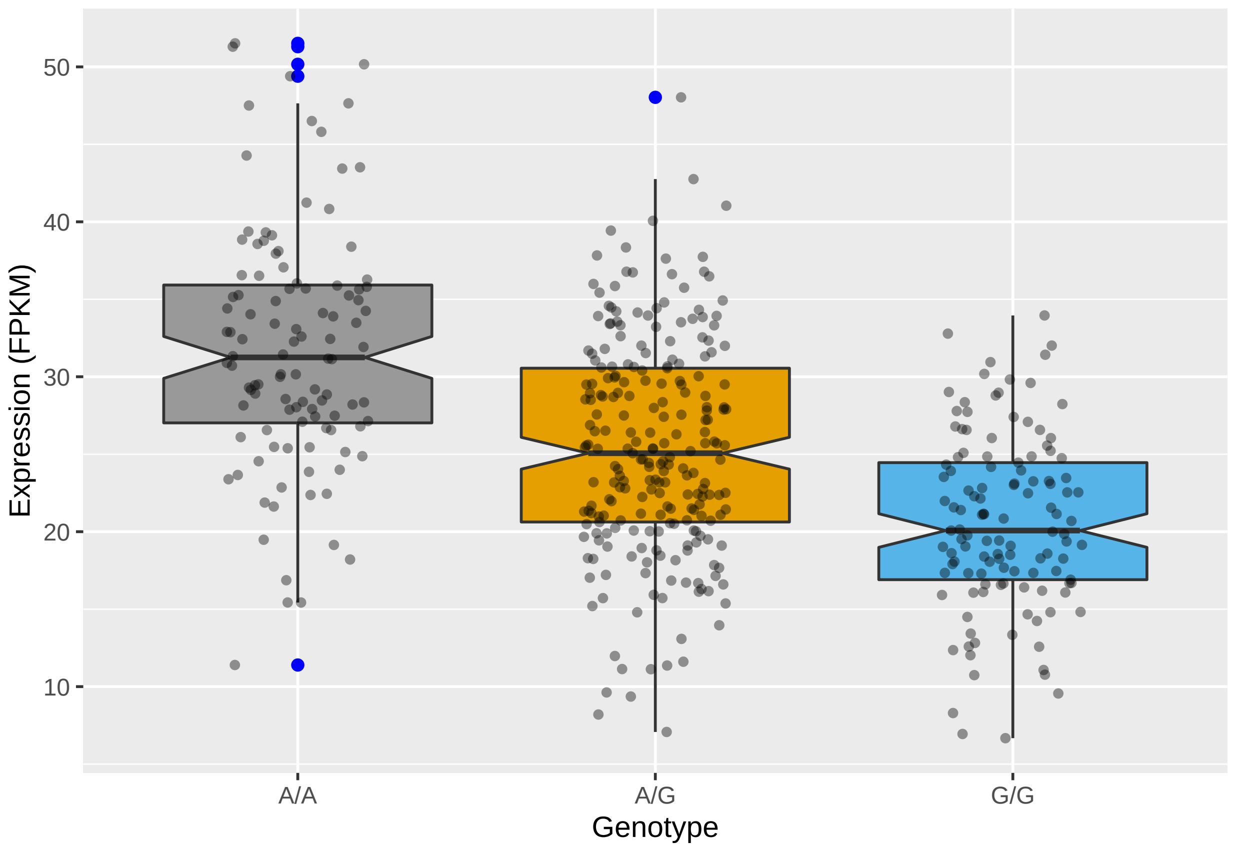
As shown below, the **G|G genotype has a lower expression level than the A|A genotype**. Even the heterozygous A|G shows a reduced expression level (the sample distribution is located at a generally lower level). Thus, **the SNP does affect ORMDL3 expression**. Whether or not this change is biologically relevant requires expreiemntal evidence.

```
library(ggplot2)

# converting our data$geno to a factor object
data$geno <- as.factor(data$geno)
head(data)
```

```
##    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
```

```
# Make boxplot with ggplot2
ggplot(data, aes(x = geno, y = exp, fill = geno)) +
         geom_boxplot(outlier.colour="blue", outlier.shape=16,
             outlier.size=2, notch=TRUE) +
         geom_jitter(shape=16, position=position_jitter(0.2), alpha = 0.4) +
         labs(x="Genotype", y = "Expression (FPKM)") +
         scale_fill_manual(values=c("#999999", "#E69F00", "#56B4E9")) +
         theme(legend.position="none")
```


50
40
30
20
10
Expression (FPKM)
A/A
A/G
G/G
Genotype