Class 12 HW: Population analysis

Lily Huynh (PID: A16929651)

2025-02-13

Table of contents

Section 4: Population Scale Analysis 1

Section 4: Population Scale Analysis

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.

Ans. A/A sample size: 108, median expression levels (A/A): 31.24847; A/G sample size: 233, median expression levels (A/G): 25.06486; G/G sample size: 121, median expression levels (G/G): 20.07363

```
sample.table <- read.table("rs8067378_ENSG00000172057.6.txt", header=TRUE)

head(sample.table)
```

```
   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
```

```
summary(sample.table)
```

```
   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
```

I will use the `table()` function to figure out the sample size of each genotype.

```
sample.size.table <- table(sample.table$geno)
sample.size.table
```

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

Then I will make sort the data into different tables based on the genotype.

```
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
```

```
genoAA <- sample.table %>%
  filter(geno == "A/A")
head(genoAA)
```

```
    sample geno      exp
3  HG00361  A/A 31.32628
4  HG00135  A/A 34.11169
6  NA11993  A/A 32.89721
8  NA18498  A/A 47.64556
13 NA20585  A/A 30.71355
15 HG00235  A/A 25.44983
```

```
genoAG <- sample.table %>%
  filter(geno == "A/G")
head(genoAG)
```

```
    sample geno      exp
1  HG00367  A/G 28.96038
2  NA20768  A/G 20.24449
7  HG00256  A/G 31.48736
10 HG00115  A/G 33.85374
11 NA20806  A/G 16.29854
12 HG00278  A/G 19.73450
```

```
genoGG <- sample.table %>%
  filter(geno == "G/G")
head(genoGG)
```

```
    sample geno      exp
5  NA18870  G/G 18.25141
9  HG00327  G/G 17.67473
17 NA12546  G/G 18.55622
20 NA18488  G/G 23.10383
23 NA19214  G/G 30.94554
28 HG00112  G/G 21.14387
```

Finally, I will calculate the median for each genotype

```
median(genoAA$exp)
```

```
[1] 31.24847
```

```
median(genoAG$exp)
```

```
[1] 25.06486
```

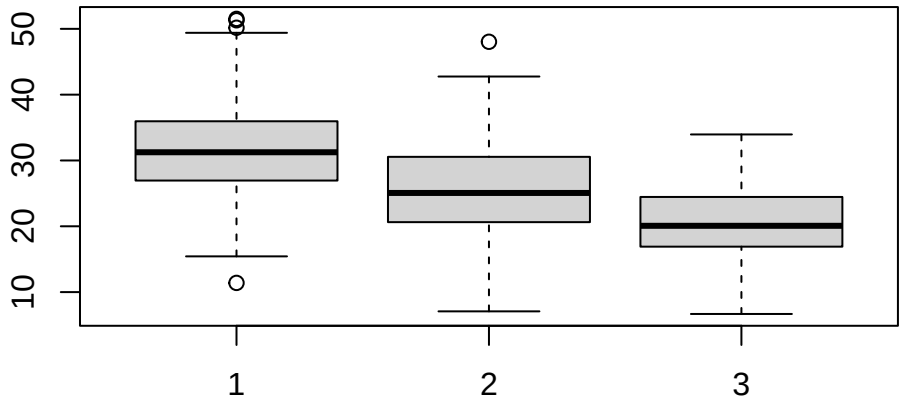
```
median(genoGG$exp)
```

```
[1] 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?
>
> Ans. The relative expression value of G/G is lower compared to A/A. Therefore, alleles with G will have a lower expression compared to alleles with A. Based on the boxplot, I can see that the SNP does effect the expression of ORMDL3 because when A/A is changed to A/G, the expression levels are lower.

I created two different boxplots, using the function `boxplot()` and `ggplot()`.

```
boxplot(genoAA$exp, genoAG$exp, genoGG$exp)
```



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