Section4: Population Scale Analysis

Ellice Wang (PID: A16882742)

2025-02-13

Table of contents

Q13: read expression data R

> determine the sample size for each genotype and their corresponding median expression levels for each of these genotypes

The median expression levels for each genotype are: A/A: 31.25 A/G: 25.065 G/G: 20.074

```
# import dplyr and ggplot
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
```

```
library(ggplot2)
```

```
url <- "https://bioboot.github.io/bimm143_W25/class-material/rs8067378_ENSG00000172057.6.txt"

exp_data <- read.table(url)
head(exp_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
```

```
unique(exp_data$geno)
```

```
[1] "A/G" "A/A" "G/G"
```

```
A_G <- exp_data[exp_data$geno == "A/G",]
summary(A_G)
```

```
    sample              geno                exp        
 Length:233         Length:233         Min.   : 7.075  
 Class :character   Class :character   1st Qu.:20.626  
 Mode  :character   Mode  :character   Median :25.065  
                                       Mean   :25.397  
                                       3rd Qu.:30.552  
                                       Max.   :48.034  
```

```
A_A <- exp_data[exp_data$geno == "A/A",]
summary(A_A)
```

```
    sample              geno                exp        
 Length:108         Length:108         Min.   :11.40  
 Class :character   Class :character   1st Qu.:27.02  
 Mode  :character   Mode  :character   Median :31.25  
                                       Mean   :31.82  
                                       3rd Qu.:35.92  
                                       Max.   :51.52  
```

```r
G_G <- exp_data[exp_data$geno == "G/G",]
summary(G_G)
```

```
    sample              geno                exp        
 Length:121         Length:121         Min.   : 6.675  
 Class :character   Class :character   1st Qu.:16.903  
 Mode  :character   Mode  :character   Median :20.074  
                                       Mean   :20.594  
                                       3rd Qu.:24.457  
                                       Max.   :33.956  
```

Q14: boxplot

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?

```r
ggplot(exp_data, aes(x=geno, y=exp, fill = geno)) +
  geom_boxplot(outliers = FALSE, notch = TRUE) +
  geom_jitter(color="black", position=position_jitter(0.2), alpha=0.3)
```