Class17_Genome_informatics

Zixuan Zeng (A16142927)

Table of contents

Question 13

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: The sample size for AA, AG, and GG genotypes are 108, 233, and 121, respectively. The median expression levels for AA, AG, and GG genotypes are 31.25, 25.065, and 20.074, respectively.

```
data <- read.table("rs8067378_ENSG00000172057.6.txt")
data_summary <- summary(data)
AA_median <- summary(data[data$geno == "A/A", "exp"])
AA_size <- nrow(data[data$geno == "A/A", ])
print(AA_median)
```

```
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max. 
  11.40   27.02   31.25   31.82   35.92   51.52 
```

```
print(AA_size)
```

```
[1] 108
```

```
AG_median <- summary(data[data$geno == "A/G", "exp"])
AG_size <- nrow(data[data$geno == "A/G", ])
print(AG_median)
```

```
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
  7.075  20.626  25.065  25.397  30.552  48.034
```

```r
print(AG_size)
```

```
[1] 233
```

```r
GG_median <- summary(data[data$geno == "G/G", "exp"])
GG_size <- nrow(data[data$geno == "G/G", ])
print(GG_median)
```

```
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
  6.675  16.903  20.074  20.594  24.457  33.956
```

```r
print(GG_size)
```

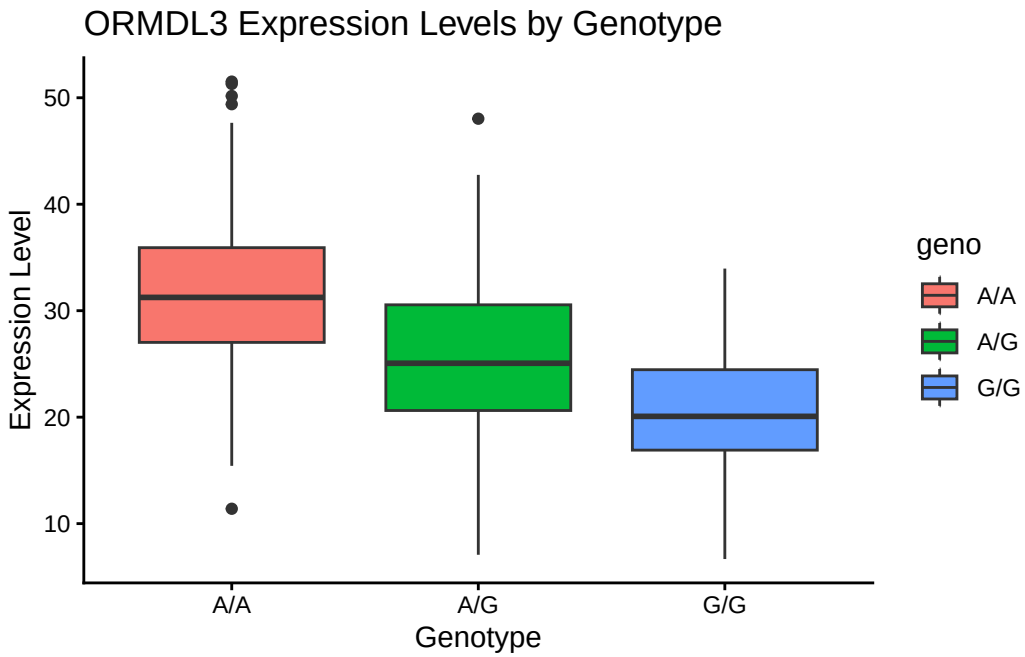
```
[1] 121
```

Question 14

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: From the boxplot, we can infer that the expression level of ORMDL3 is highest for the A/A genotype and lowest for G/G genotype. The result suggests that this SNP is likely to affect the expression of ORMDL3. The result from a two sided t-test (confidence level = 0.95, p-value < 2.2e-16) also suggests a significant difference in expression levels between the A/A and G/G genotypes, further supporting the conclusion that the SNP affects ORMDL3 expression.

```r
library(ggplot2)
ggplot(data) + aes(geno, exp, fill = geno) + geom_boxplot() +
  labs(title = "ORMDL3 Expression Levels by Genotype",
       x = "Genotype", y = "Expression Level") + theme_classic()
```




```
t.test(exp ~ geno, alternative = ("two.sided"), data = data[data$geno %in% c("A/A", "G/G"), ]
```

```
    Welch Two Sample t-test

data:  exp by geno
t = 12.214, df = 191.65, p-value < 2.2e-16
alternative hypothesis: true difference in means between group A/A and group G/G is not equal
95 percent confidence interval:
  9.412243 13.037619
sample estimates:
mean in group A/A mean in group G/G
         31.81864          20.59371
```