

Lab 11: Introduction to Genome Informatics Lab

Jordan Laxa

2023-02-20

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.

```
gene <- read.csv("gene.txt")
library(stringr)
gene[c('number', 'sample', 'genotype', 'expression')] <- str_split_fixed(gene$sample.geno.exp, ' ', 4)
gene <- gene[c('number', 'sample', 'genotype', 'expression')]
gene = subset(gene, select = -c(number))
gene$expression <- as.numeric(gene$expression)
summary(gene)
```

```
##      sample      genotype      expression
## 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
```

```
geneAG <- subset(gene, genotype=="A/G")
length(which(gene$genotype=="A/G"))
```

```
## [1] 233
```

```
summary(geneAG$expression)
```

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

```
geneGG <- subset(gene, genotype=="G/G")
length(which(gene$genotype=="G/G"))
```

```
## [1] 121
```

```
summary(geneGG$expression)
```

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

```
geneAA <- subset(gene, genotype=="A/A")
length(which(gene$genotype=="A/A"))
```

```
## [1] 108
```

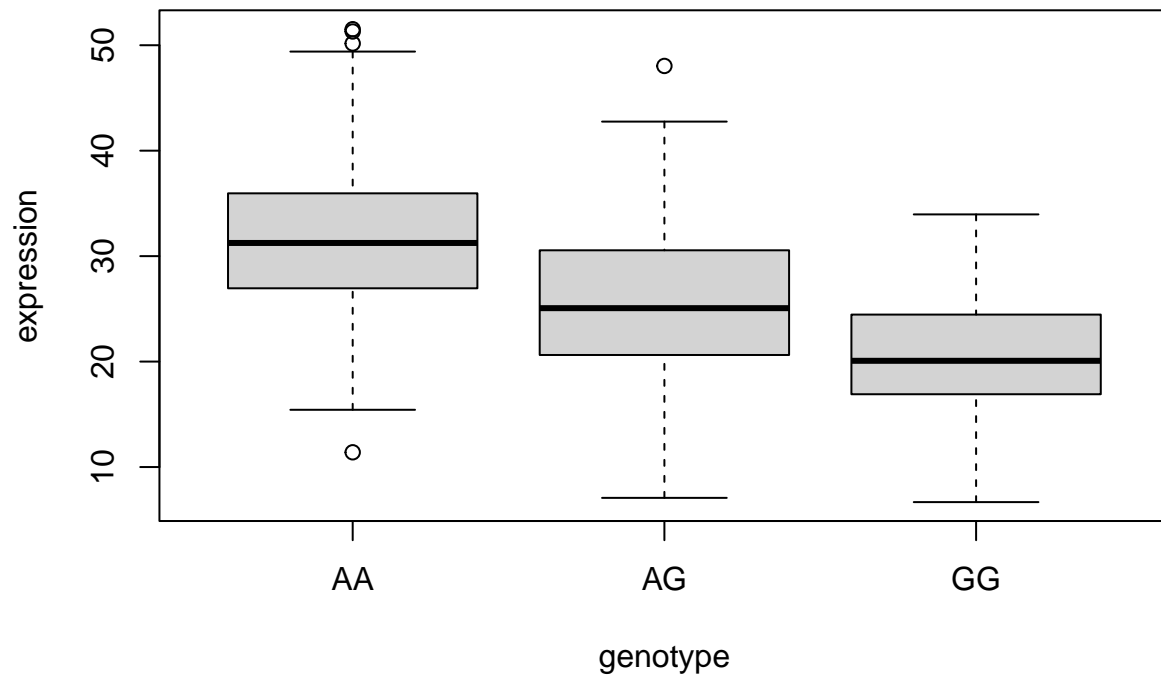
```
summary(geneAA$expression)
```

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

Q13A. The sample size for genotype A/G is 233 and it's median expression level is 25.065. The sample size for genotype G/G is 121 and it's median expression level is 20.074. The sample size for genotype A/A is 108 and it's median expression level is 31.25.

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 ORM DL3?

```
boxplot(geneAA$expression, geneAG$expression, geneGG$expression, xlab = "genotype", ylab = "expression")
```



Q14A. The median expression value of G/G is 20.074, which is relatively lower than the median expression value of A/A, 31.25. The SNP decreases the expression of ORM DL3.