Class 12: Introduction to Genome Informatics

AUTHOR
Nundini Varshney (PID: A16867985)

Section 4: Population Scale Analysis [HOMEWORK]

One sample is obviously not enough to know what is happening in a population. You are interested in assessing genetic differences on a population scale. So, you processed about ~230 samples and did the normalization on a genome level. Now, you want to find whether there is any association of the 4 asthma-associated SNPs (rs8067378...) on ORMDL3 expression.

> 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.

```
#First, I read the file into R

genstats <- read.table("textfile.txt", row.names = 1)
head(genstats)
```

```
   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
```

Sample size for each genotype?

```
#How many rows are there?
nrow(genstats)
```

```
[1] 462
```

```
#What is the sample size for each genotype?
table(genstats$geno)
```

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

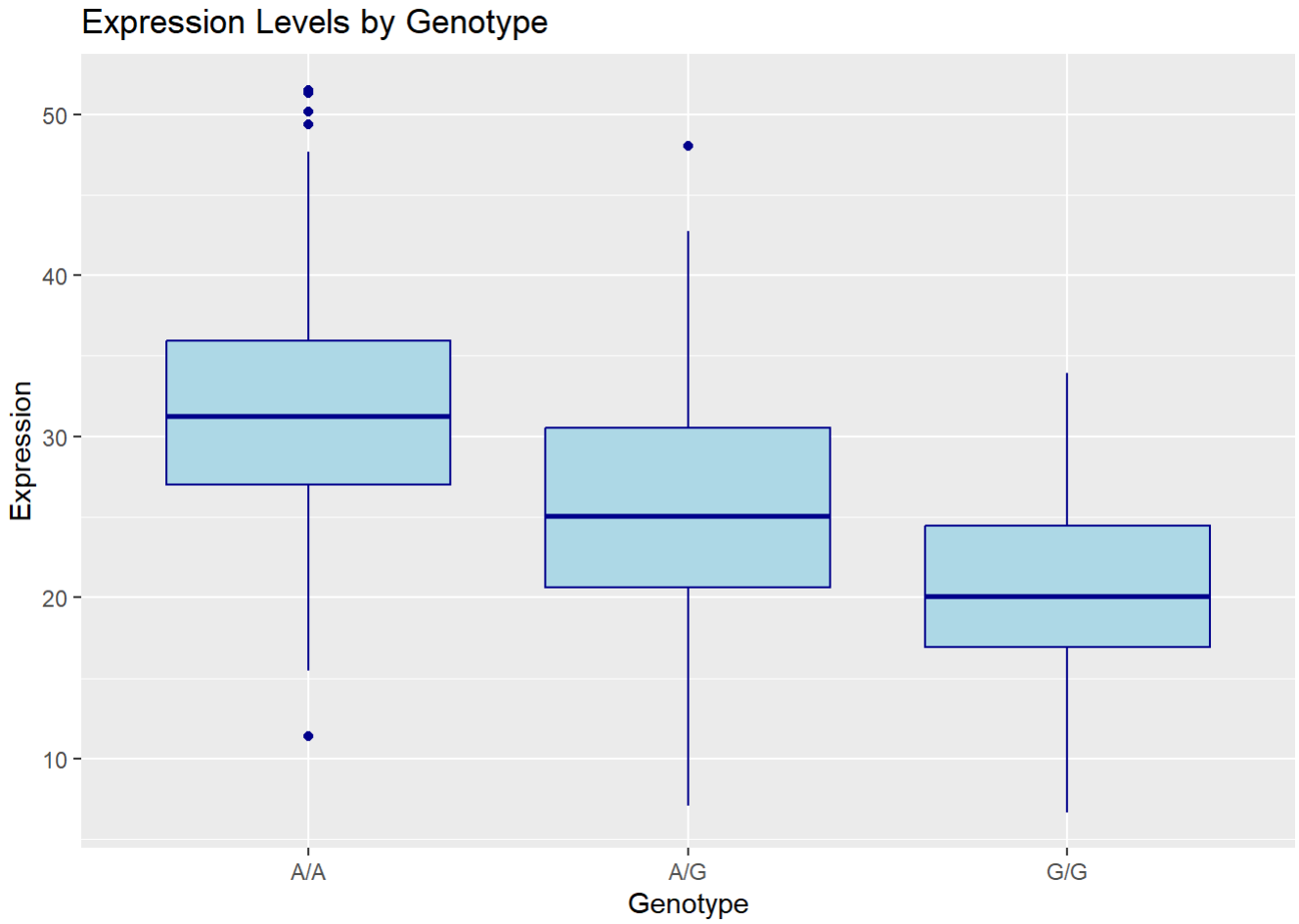
Median for each of the three genotypes?

```r
genstats <- read.table("textfile.txt", row.names = 1)

# Calculate the median expression level for each genotype
medians <- tapply(genstats$exp, genstats$geno, median)

# Print the median values for each type
print(medians)
```

```
     A/A      A/G      G/G
31.24847 25.06486 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?

```r
# Generate boxplot
library(ggplot2)
ggplot(genstats, aes(x = geno, y = exp)) +
  geom_boxplot(fill = "lightblue", color = "darkblue") +
  labs(x = "Genotype", y = "Expression",
       title = "Expression Levels by Genotype")
```



From this boxplot, we can infer that the A/A genotype has a higher expression level than the G/G genotype because visually speaking, the box and its median line is higher in comparison to G/G's. The SNP does affect the expression of ORMDL3.