Population Scale Analysis

Meg Robinson

2/19/2022

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.

```
expr <- read.table("exp.txt")
head(expr)
```

```
##    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
```

```
nrow(expr)
```

```
## [1] 462
```

There are 462 individuals we have this data for. But how many for each genotype?

```
table(expr$geno)
```

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

```
summary(expr)
```

```
##     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
```

This 'summary()' function gives us the overall median expression levels, but we want to see it per genotype.

```
summary(expr$exp[expr$geno == "A/A"])
```

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

```
summary(expr$exp[expr$geno == "G/G"])

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

summary(expr$exp[expr$geno == "A/G"])

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

ANSWER: Genotype A/A has a median expression level of 31.25. Genotype G/G has a median expression level of 20.074. Genotype A/G has a median expression level of 25.065.

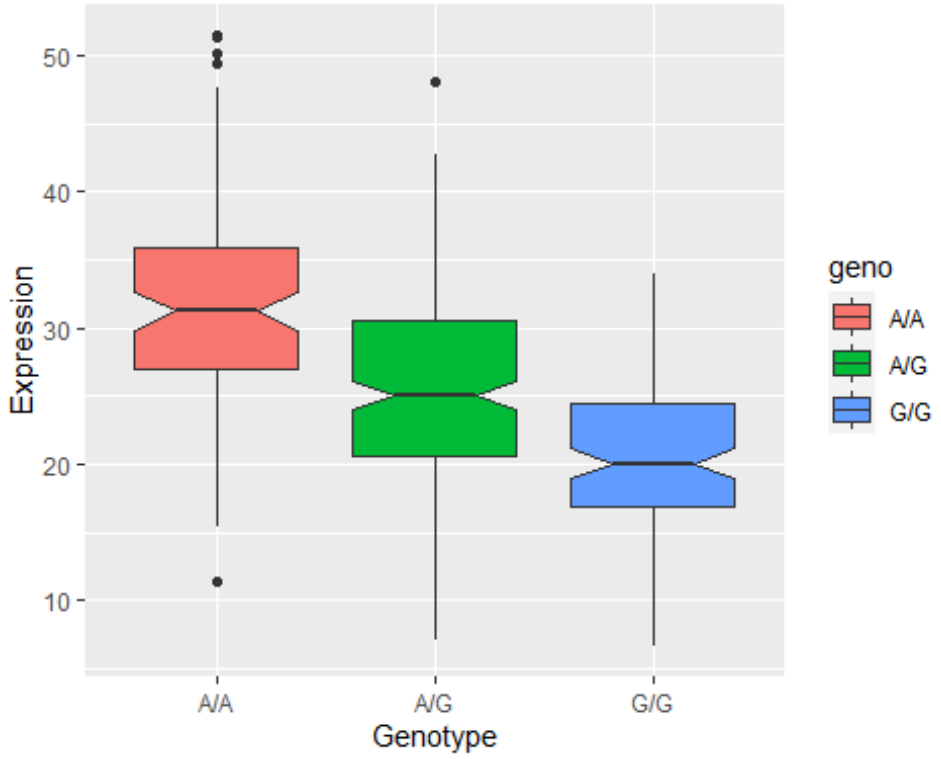
Lets make a summary figure to display these results.

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?

```
library(ggplot2)

bplot <- ggplot(expr) + aes(x=geno, y=exp, fill=geno) +
geom_boxplot(notch=TRUE) + xlab("Genotype") + ylab("Expression")

bplot
```



ANSWER: Based on this box plot, I can infer that the relative expression of G/G is markedly lower than the expression of A/A. Yes, it seems that SNP influences ORMDL3 gene expression (known for asthma)