Class19: Genome Informatics

Monica Lin (PID: A15524235)

11/30/2021

Section 4: Population Scale Analysis

One sample is obviously not enough to know what is happening in a population. We are interested in assessing genetic differences on a population scale.

```
expr <- read.table("Expression genotype results.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
```

Q13. How many samples do we have?
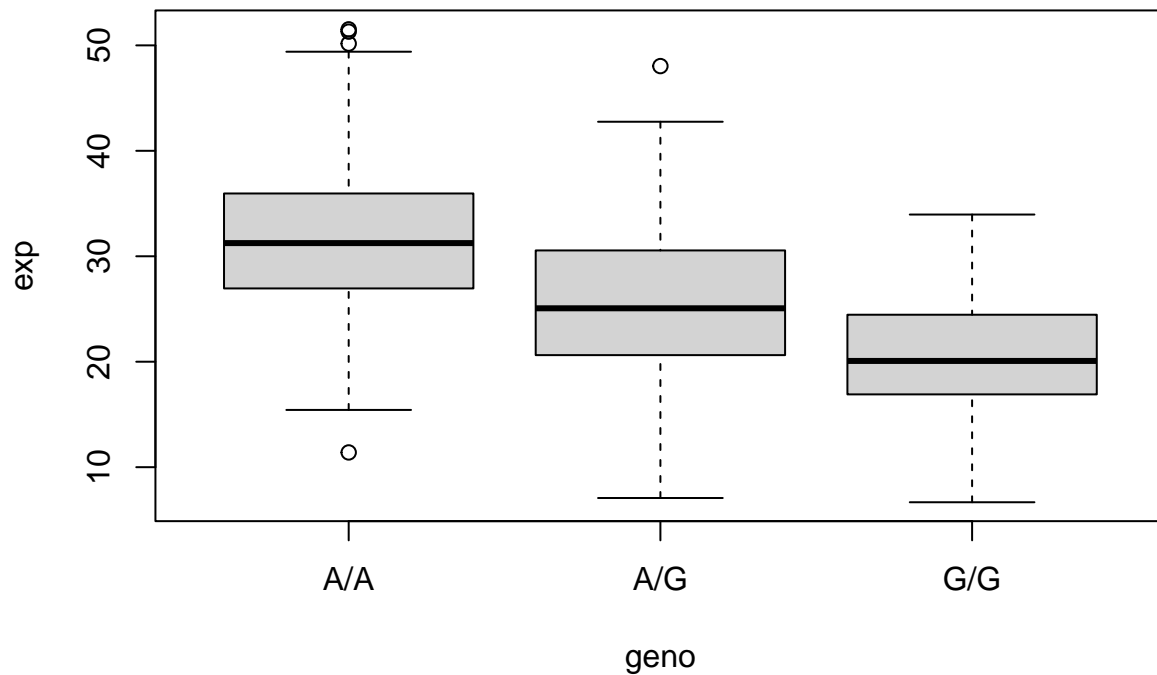
```
nrow(expr)
```

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

What are the sample sizes for each genotype?

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

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

What are the corresponding median expression levels for each genotype?

```
boxplot <- boxplot(exp~geno, data=expr)
```

```
# The third row displays the median values for each genotype. The second value
# designates to how many decimal places the statistics will round.

round(boxplot$stats, 2)
```

```
##        [,1]  [,2]  [,3]
## [1,] 15.43  7.08  6.67
## [2,] 26.95 20.63 16.90
## [3,] 31.25 25.06 20.07
## [4,] 35.96 30.55 24.46
## [5,] 49.40 42.76 33.96
```

The sample size of genotype A|A is 108, with a median value of 31.25. A|G has a sample size of 233 and a median value of 25.06. G|G has a sample size of 121 and a median value of 20.07.

Q14. Let's make a boxplot, with a box per genotype.

```
library(ggplot2)
```

```
ggplot(expr) + aes(x=geno, exp, fill=geno) +
  geom_boxplot(notch=TRUE) + xlab("Genotype") + ylab("Expression")
```
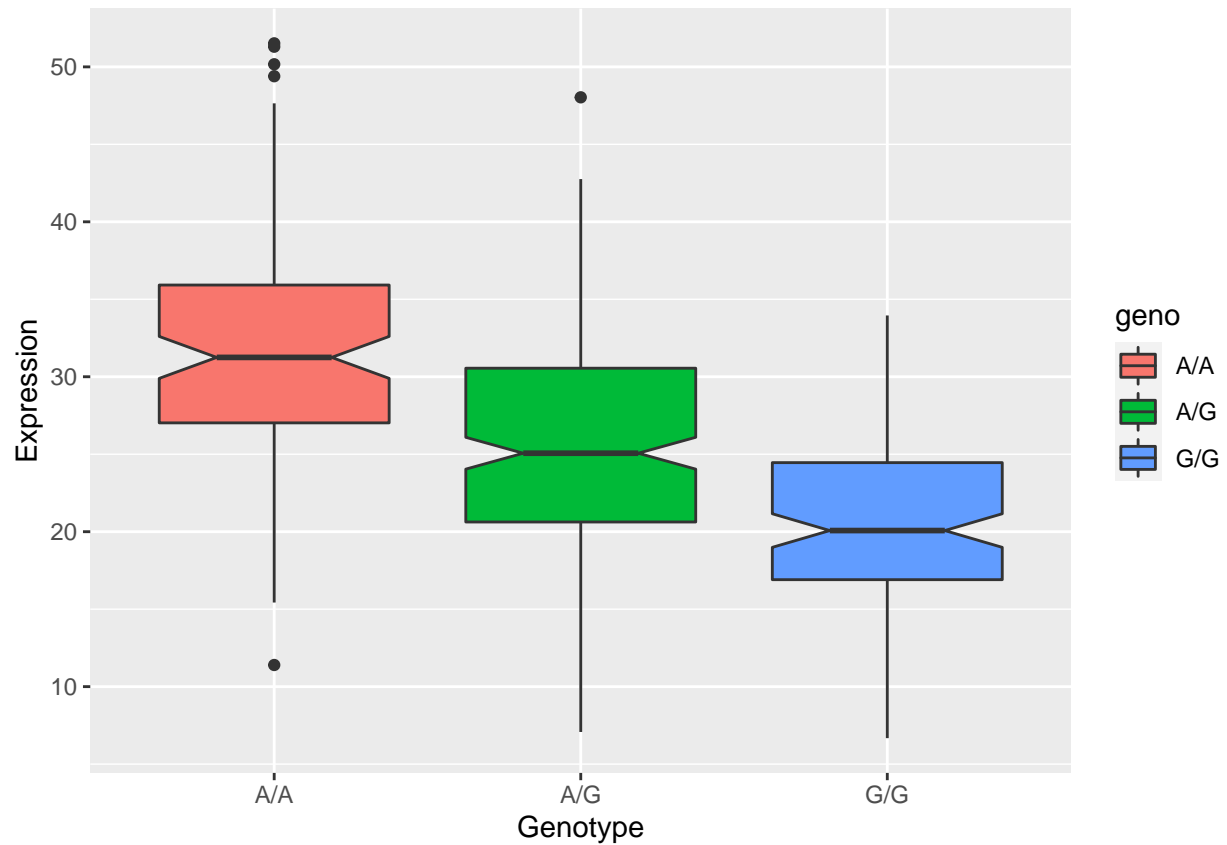
What could you infer from the relative expression value between A|A and G|G displayed in this plot? Does the SNP affect the expression of ORMDL3?

The homozygous G|G genotype is down-expressed in relation to the A|A and A|G genotypes. In other words, the A|A genotype is upregulated while G|G is downregulated. The SNP lowers the expression of ORMDL3.