Class 11: Genome Informatics

Olivia Chu

2023-02-15

Section 4: Population Scale Analysis

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.

Hint: The read.table(), summary() and boxplot() functions will likely be useful here. There is an example R script online to be used ONLY if you are struggling in vein. Note that you can find the medium value from saving the output of the boxplot() function to an R object and examining this object. There is also the medium() and summary() function that you can use to check your understanding.

```
expr <- read.table("rs8067378_ENSG00000172057.6.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
```

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

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

There are 108 A/A samples, 233 A/G samples, and 121 G/G samples.

```
library(dplyr)
```

```
##
## Attaching package: 'dplyr'
## The following objects are masked from 'package:stats':
##
##     filter, lag
```

```
## The following objects are masked from 'package:base':
##
##     intersect, setdiff, setequal, union
```

```
aa <- expr %>%
  filter(geno == "A/A")

median(aa$exp)
```

```
## [1] 31.24847
```

The median expression level for the genotype A/A is 31.2.

```
ag <- expr %>%
  filter(geno == "A/G")

median(ag$exp)
```

```
## [1] 25.06486
```

The median expression level for the genotype A/G is 25.1.

```
gg <- expr %>%
  filter(geno == "G/G")

median(gg$exp)
```

```
## [1] 20.07363
```

The median expression level for the genotype G/G is 20.1.

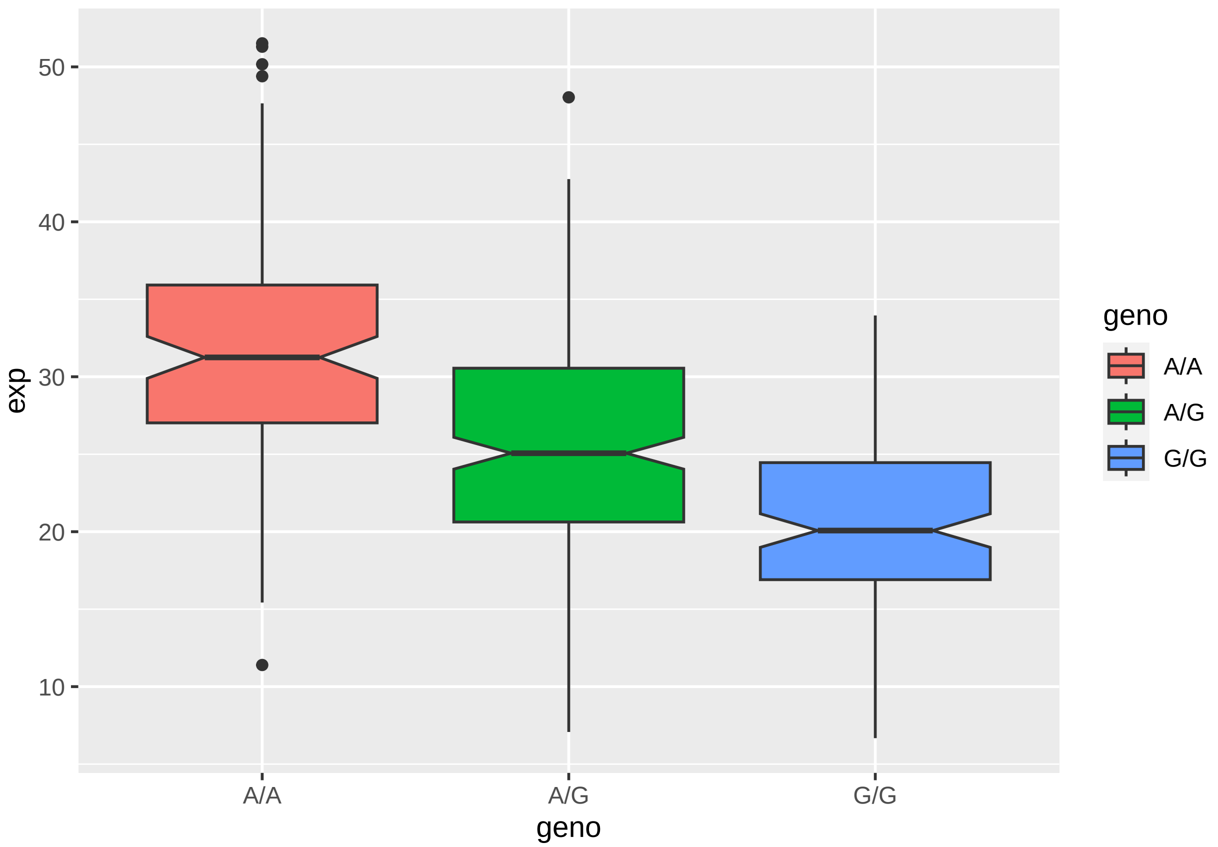
> Q14: Generate a boxplot with a box per genotype, what could you infer from the relative expression value between A/A and G/G displayed in this plot? Does the SNP effect the expression of ORMDL3?

Hint: An example boxplot is provided overleaf – yours does not need to be as polished as this one.

```
library(ggplot2)
```

Let's make a boxplot.

```
ggplot(expr) + aes(x=geno, y=exp, fill=geno) +
  geom_boxplot(notch=TRUE)
```



From this boxplot, we can observe that the expression level of the A/A genotype is higher than that of the expression level of the genotype G/G. From this, we can infer that having the G/G genotype in this location is associated with having reduced expression of this particular gene. In other words, because there is less expression of the G/G phenotype (compared to A/A and A/G), we can assume that the SNP effect does effect expression of ORMDL3.