week08

Vina Nguyen

2023-05-21

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.

This is the final file you got (https://bioboot.github.io/bggn213_W19/classˆmaterial/rs8067378_ENSG00000172057.6.txt). The first column is sample name, the second column is genotype and the third column are the expression values.

> 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.

How many samples do we have?

```
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
```

```
nrow(expr)
```

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

How many of each type?

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

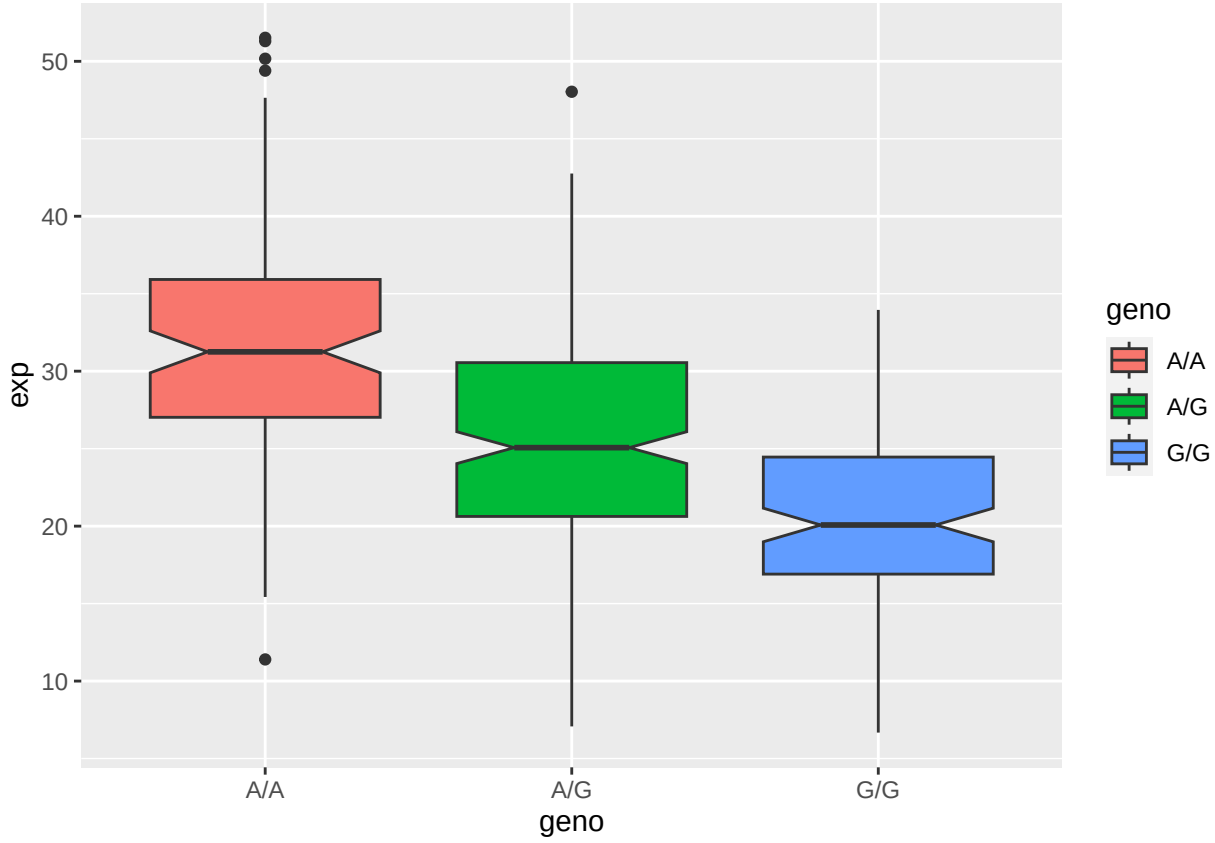
```
##
## A/A A/G G/G
## 108 233 121
```

```
library(ggplot2)
```

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?

Lets make a boxplot.

```
exp_geno <- ggplot(expr) +
  aes(geno, exp, fill=geno) +
  geom_boxplot(notch=TRUE)
exp_geno
```



Expression with AA is highest. Expression with GG is lowest. Having a GG in this location is associated with having a reduced expression of this gene. The SNP effects the expression of ORMDL3 since it is in the chromosome 17 location with G alleles.