Class 8: Genome Informatics

Katherine Wong (A16162648)

12/4/2021

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.

```
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
```

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

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

Sample Size for A/A: 108, A/G= 233, G/G=121. Median for A/A = 31.24847. Median for A/G = 25.06486. Median for G/G = 20.07363.

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

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

```
median(expr[expr$geno == "A/G", "exp"])
```

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

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

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

```
library(ggplot2)
```

Let's make a boxplot

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