Lab 17 pt2 Population Analysis (Q13 Q14 box plot)

Blinda Sui (PID: A17117043)

2025-12-01

Section 4: Population Scale Analysis [HOMEWORK]

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
```

Answer of sample size for each genotype

```
# Sample size per genotype
table(expr$geno)
```

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

Answer of median expression levels

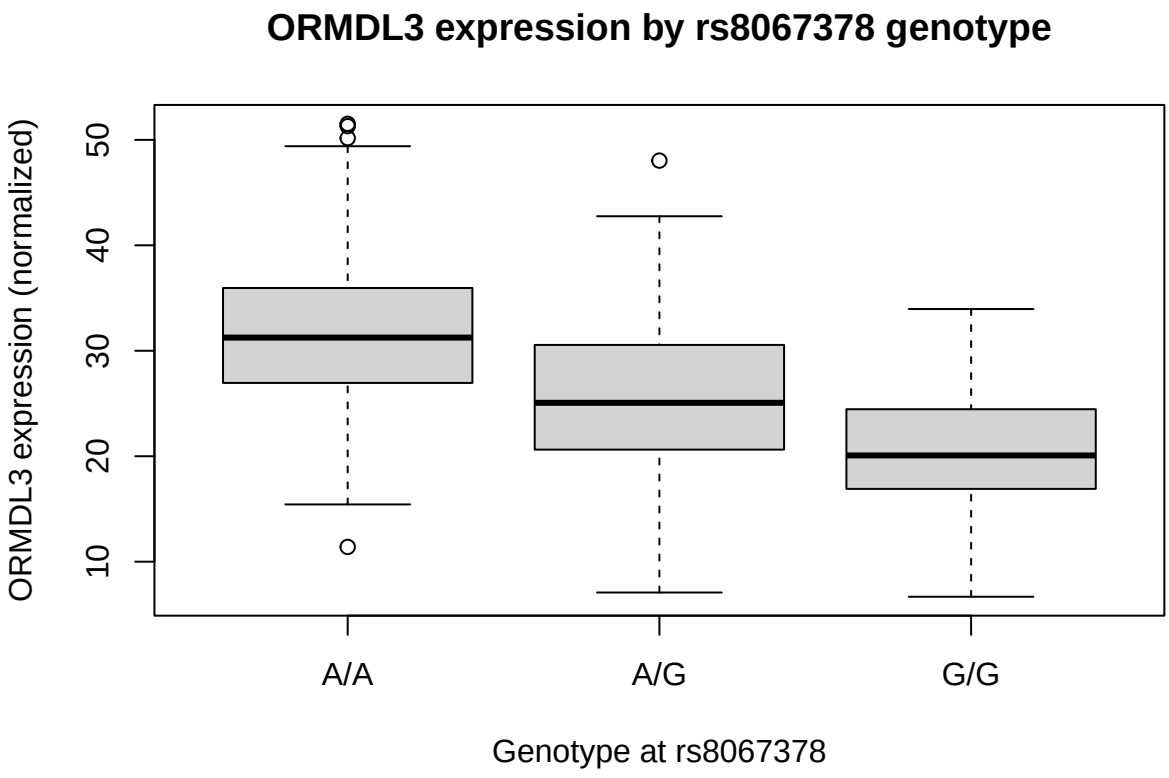
```
# Median expression per genotype
tapply(expr$exp, expr$geno, median)
```

```
##      A/A      A/G      G/G
## 31.24847 25.06486 20.07363
```

Sample size for A/A genotype is 108, A/G genotype is 233, G/G genotype is 121. Median of A/A genotype is 31.25, median of A/G genotype is 25.06, median of G/G genotype is 20.07.

> 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?

```r
# Alternative using boxplot statistics
bp <- boxplot(exp ~ geno, data = expr,
              xlab = "Genotype at rs8067378",
              ylab = "ORMDL3 expression (normalized)",
              main = "ORMDL3 expression by rs8067378 genotype")
```



Based on the boxplot: A/A has the highest expression values, A/G is intermediate, and G/G has the lowest expression values. So, A/A > A/G > G/G in terms of ORMDL3 expression (medians ~31 vs 25 vs 20). So expression of A/A show higher ORMDL3 expression than G/G, with A/G in between. This suggests that the asthma-risk allele is associated with increased ORMDL3 expression, so the SNP likely affects gene regulation.