class12_HW

Thrisha Praveen

2025-02-18

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.

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

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
```

How many in total?

```
nrow(expr)
```

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

How many of each genotype?

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

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

> 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.

```
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
```

```
expr %>%
  group_by(geno) %>%
  summarize(median_exp = median(exp, na.rm = TRUE))
```

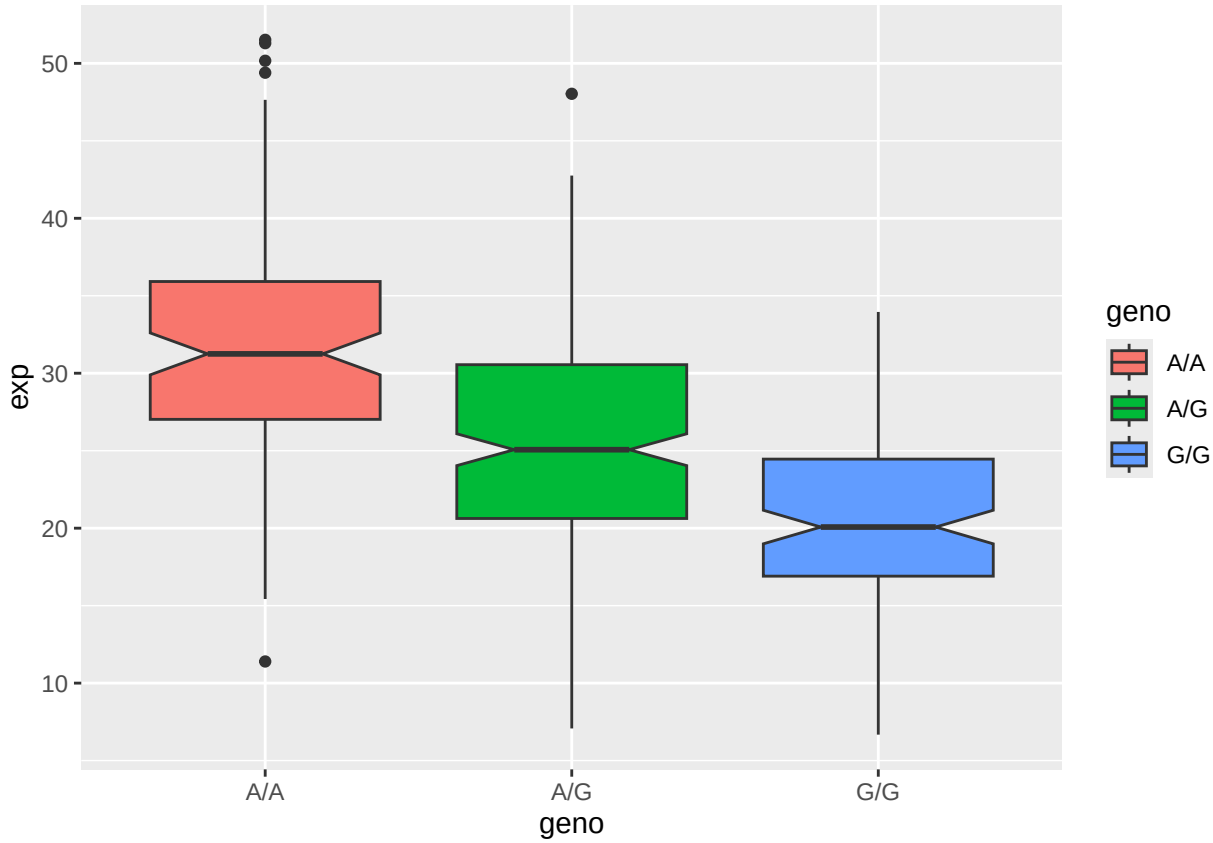
```
## # A tibble: 3 x 2
##   geno  median_exp
##   <chr>      <dbl>
## 1 A/A         31.2
## 2 A/G         25.1
## 3 G/G         20.1
```

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?

```
library(ggplot2)
```

Making a boxplot, grouped by genotype:

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



>The A/A genotype is more highly expressed than the G/G genotype in this sample. Thus, the SNP does affect ORMDL3 expression with a decrease in comparison to expression levels with the wild-type (A/A) genotype.