# class 12 homework

Eli Sobel A69027989

## 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_F24/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. Open a new RMarkdown document in RStudio to answer the following two questions. Submit your resulting PDF report with your working code, output and narrative text answering Q13 and Q14 to GradeScope.

> 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.

The sample size is 108 for A/A, 233 for A/G, and 121 for G/G. The median expression levels per genotype are 31.25 for A/A, 25.06 for A/G, and 20.07 for G/G. Code provided below.

Read in data and examine the structure with `head()`:

```
expr <- read.table("https://bioboot.github.io/bggn213_F24/class-material/rs8067378_ENSG00000
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
```

Generate table of samples per genotype with `table()`:

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

```

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

Calculate median expression per genotype:

```r
# make sure tidyr and dplyr are installed via install.packages()
library(tidyr)
```

```
Warning: package 'tidyr' was built under R version 4.4.2
```

```r
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
```

```r
median_exp_per_geno <- expr %>%
  group_by(geno) %>%
  summarise(median_exp = median(exp))
median_exp_per_geno
```

```
# 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
```

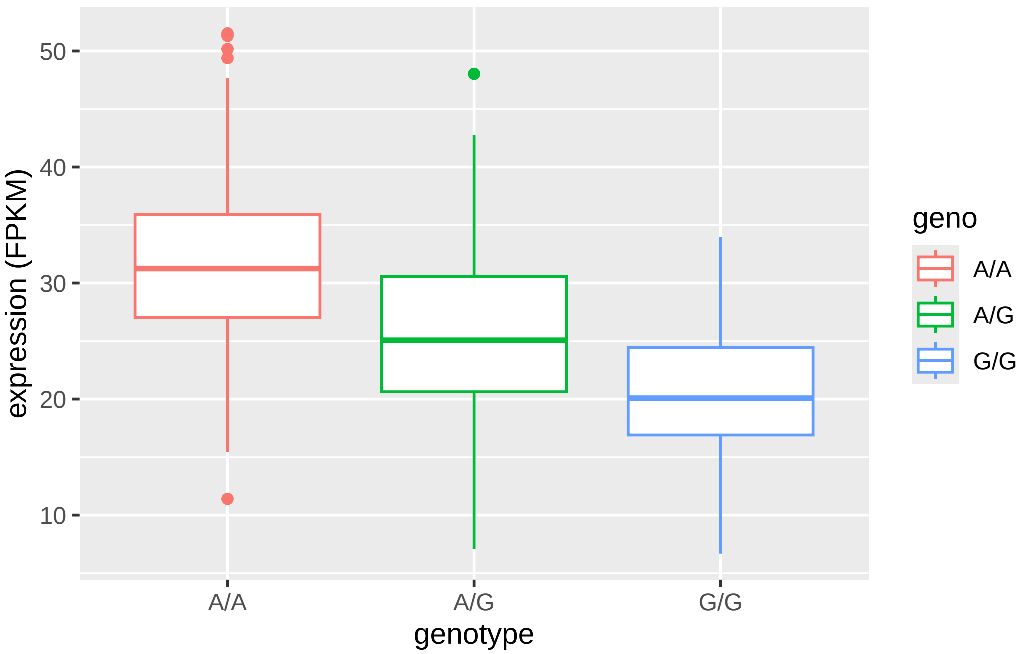
> Q14: Generate a boxplot with a box per genotype, what could you infer from the relative expression value between A/A and G/G displayed in this plot? Does the SNP effect the expression of ORMDL3?

You could infer from the relatively higher expression range of A/A compared to G/G, as well as the intermediate range of A/G, that the G->A SNP is associated with increased expression of ORMDL3. It is likely that there is a causal relationship between this SNP and ORMDL3 expression.

Use ggplot to make a box plot of expression by genotype:

```r
library(ggplot2)

x <- ggplot(expr, aes(geno, exp, col=geno)) + geom_boxplot() + labs(x = "genotype", y = "exp
x
```


expression (FPKM)
A/A
A/G
G/G
genotype
geno
A/A
A/G
G/G