Section 4. Population Scale Analysis

Joshua Cheung

02/16/2022

In these exercises we are interested in assessing genetic differences on a population scale. So, we processed about ~230 samples and did the normalization on a genome level. Now, you want to find whether there is any association of the asthma-associated SNPs (rs8067378) on ORMDL3 expression.

We first read the final file we downloaded as follows:

```
# We define the object expr.
expr <- read.table("rs8067378_ENSG00000172057.6.txt")
# We now preview the first 6 rows.
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
```

We see that the first, second, and third columns are the sample name, genotype, and expression values respectively.

We now find how many total samples there are using the nrow() function.

```
nrow(expr)
```

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

So there are 462 sample total.

> **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. Hint: The read.table(), summary() and boxplot() functions will likely be useful here. There is an example R script online to be used ONLY if you are struggling in vein. Note that you can find the medium value from saving the output of the boxplot() function to an R object and examining this object. There is also the medium() and summary() function that you can use to check your understanding.**

We wish to find out how many of each type of genotype there are.

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

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

Thus to answer the first part of question 13, we see that the sample size for the A|A, A|G, and G|G genotypes are 108, 233, and 121 respectively.

We now want to find the corresponding median expression levels for each of these genotypes. We first save the genotype column and expression level value column from expr as objects in R.

```
# We first save the genotype and expression level value columns from expr.
genotype <- expr$geno
expvalue <- expr$exp
```

Now we can use the tapply() function to compute median expression level value for each genotype as follows:

```
# We use the tapply() function.
round(tapply(expvalue, INDEX=genotype, FUN=median), digits=2)
```

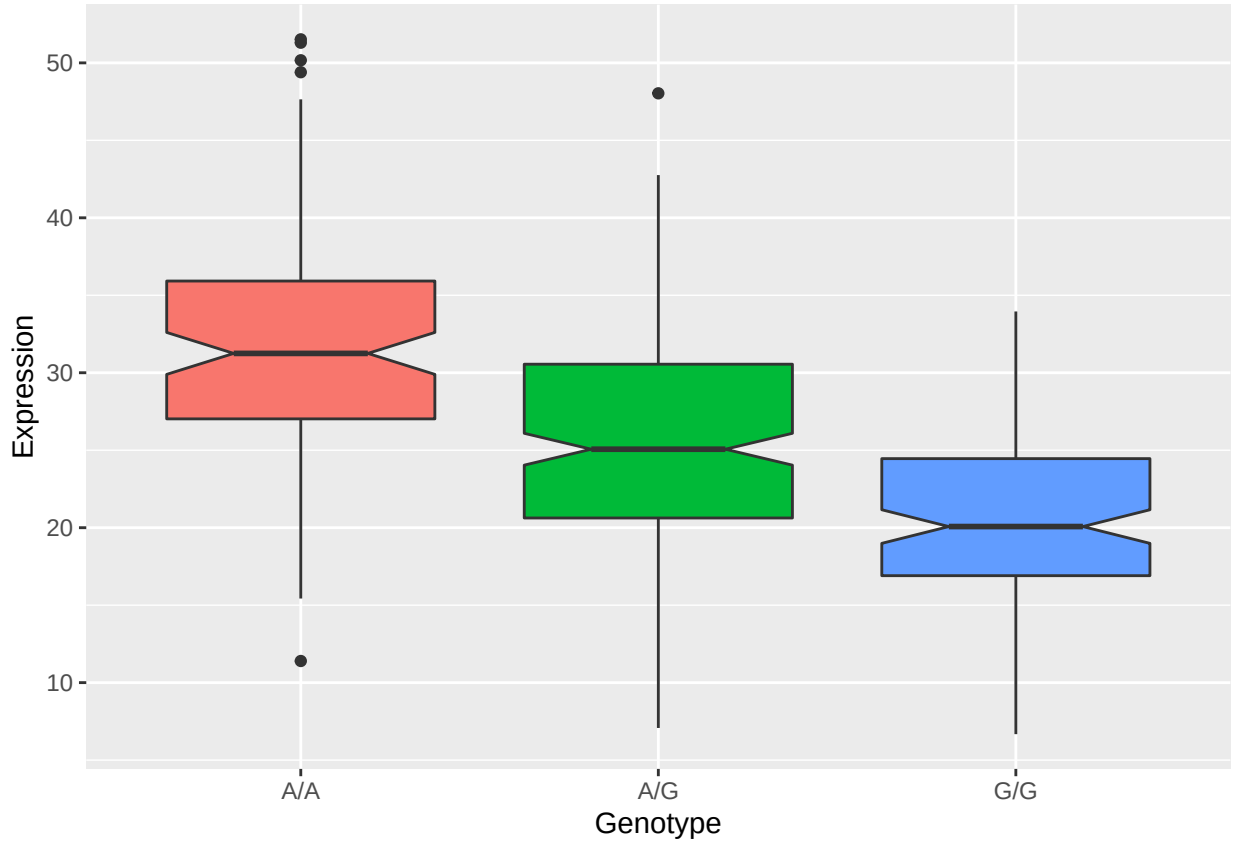
```
##   A/A   A/G   G/G
## 31.25 25.06 20.07
```

Thus to answer the second part of the question 13, the median expression level values for the A|A, A|G, and G|G genotypes are 31.25, 25.06, and 20.07 respectively.

> **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? Hint: An example boxplot is provided overleaf – yours does not need to be as polished as this one.**

```
# We call the ggplot2 package.
library(ggplot2)

# Now we create a box plot.
ggplot(expr) + aes(geno, exp, fill=geno) +
  labs(x="Genotype",
       y="Expression") +
  geom_boxplot(notch=TRUE, show.legend=FALSE)
```



To answer question 14, we see that the expression level values of the G|G genotype is clearly different from the other genotypes, especially the A|A genotype. More specifically, we can infer that the expression value is generally lower for the G|G phenotype than it is for the A|A phenotype. Thus having a G|G genotype at this SNP location in the genome is definitely associated with having a reduced expression levels of the ORMDL3 gene relative to the A|A genotype.