Class 12

Vanesa Fernandez

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.

Exploring the data

```
## reading the data

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
```

Q. How many samples do we have?

```
##How many total samples?
nrow(expr)
```

```
[1] 462
```

Q. How many of each genotype?

```
sample_size <- table(expr$geno)
print(sample_size)
```

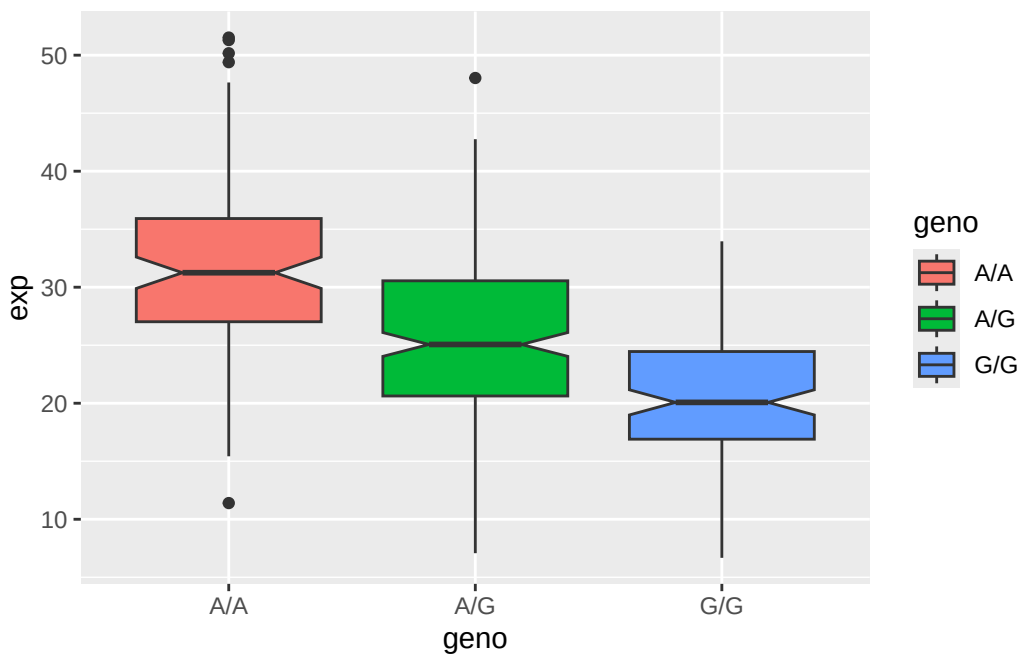
```
A/A A/G G/G
108 233 121
```

Making a figure to communicate the results

```
library(ggplot2)
```

Visualizing the data – Let's make a boxplot where we compare genotype vs expression levels

```
##helps to open the data table in another window to retrieve the name of the columns for x, y
ggplot(expr) +
  aes(x=geno, y= exp, fill=geno) + ## coloring filling by genotype
  geom_boxplot(notch=TRUE) ## returns a list that contains a variety of statistics, including
```



So far, with this plain plot and without running stats we can see that G/G genotype is associated with a reduced expression of ORMDL3 gene.

To extract the medians from the boxplot created above, use the boxplot function separately to save it as an object:

```
# Create a boxplot and save the output to an object
boxplot_stats <- boxplot(exp ~ geno, data = expr, plot = FALSE) ## creates the boxplot but d
```

```
# Extract the median values - 3rd row
medians <- boxplot_stats$stats[3, ] ## boxplot_stats$stats is a matrix where each row corres

print(medians)
```

```
[1] 31.24847 25.06486 20.07363
```

```
# Combine medians and sample sizes into a data frame
results <- data.frame(
  Genotype = names(sample_size), # Genotype names from the sample_size table
  SampleSize = as.numeric(sample_size), # Sample sizes converted to numeric
  MedianExpression = medians
)

print(results)
```

```
  Genotype SampleSize MedianExpression
1      A/A        108         31.24847
2      A/G        233         25.06486
3      G/G        121         20.07363
```

Alternatively …

```
#Extract medians using the `median()` function
median_values <- tapply(expr$exp, expr$geno, median)

print(median_values)
```

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