Solutions for the exercises

Exercise 1: What is the read count of the gene "STPG1" for the CVID patient number 8?

```
df["STPG1",8]
134
```

Exercise 2: what is the median value of the gene "FUCA2"?

```
median(tdf$FUCA2) # or tdf %>% pull(FUCA2) %>% median()
174.5
```

Exercise 3: compute the IQR for the gene ANKIB1 only for the HD

```
tdf %>% filter(Group == "HD") %>% pull(ANKIB1) %>% IQR()
# OR
IQR(tdf$ANKIB1[Group=="HD"])
387
```

Exercise 4: identify the top two genes that are upregulated and the top two genes that are downregulated in CVID compared to the HD based on the log2FC.

```
sumdf[order(sumdf$log2FC),]    # Sort based on the log2FC to see better
```

Down: SEMA3F, DPM1, UP: CFH, HS3ST1

Exercise 5: how many genes are differentially expressed if you use a significance level of 0.05?

```
sum(sumdf$pvals<0.05)
13
```

Exercise 6: how many genes are differentially expressed if you use a significance level of 0.05 for the Bonferroni corrected p-values?

```
sum(sumdf$pvals_BF<0.05)
9
```

Exercise 7: how many genes are differentially expressed if you use a significance level of 0.05 for the Benjamini-Hochberg corrected p-values?

```
sum(sumdf$pvals_BH<0.05)
13
```

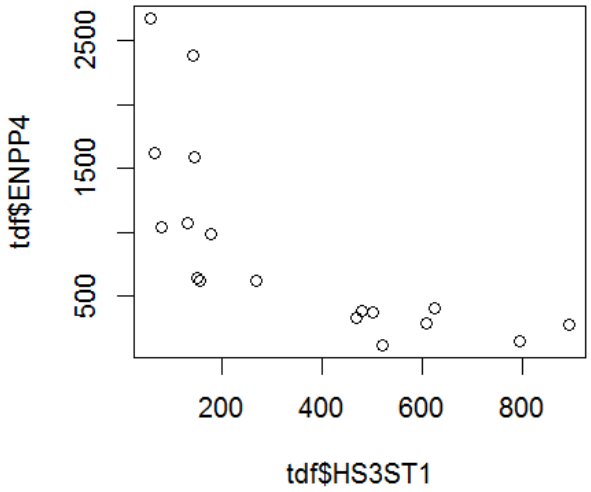
Exercise 8: find which pair of genes that show the strongest negative correlation. Make a scatter plot of these two genes.

Print the sorted correlation coefficient. At the bottom of this list, you will see that the genes HS3ST1 and ENPP4 have a correlation coefficient of -0.91. You can also use the code below:

```
cor_mat =cor(tdf,method="spearman")
which(cor_mat==min(cor_mat),arr.ind = TRUE)
```

Make a plot by:

```
plot(tdf$HS3ST1,tdf$ENPP4)
```



Exercise 9: which explanatory variable(s) are included in the "best model"?

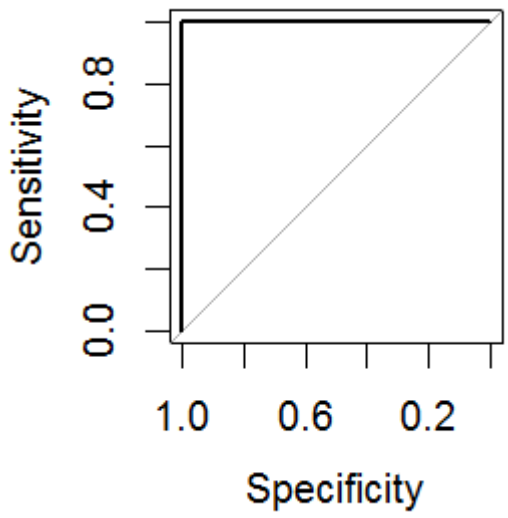
Only ANKIB1

Exercise 10: Select another gene from the list and create an ROC curve. What results did you get? Could your gene be a potential biomarker for CVID?

The answer depends on which gene you selected…

Exercise 11: Generate an ROC curve based on the gene FGR and propose an appropriate cutoff value to use in order to discriminate between the HD and CVID patients.

```
library(pROC)
roc1=roc(Group, tdf$FGR)
plot(roc1)
```



We could arbitrarily select a cutoff around 7000, which separates the groups:

```
stripchart(tdf$FGR~Group,vertical = T, method="jitter",ylab="Read count",main="FGR")
abline(h=7000,col=”red”)
```

Exercise 12: Compare the BH-adjusted p-values (last two columns) from the t-tests and the NB-regression. What is the major difference?

NB-regression generates much lower p-values and identifies one additional gene (CD99) that is differentially expressed.

Optional tutorial Hierarchical clustering

Exercise 1: Pick another pair of genes and compute the Euclidean distance. Make sure that the value you get corresponds to the value in the distance matrix (h).

For example: distance between SCYL3 and FGR

```
p=tdf$SCYL3
q=tdf$FGR
sqrt(sum((p-q)^2))
40559.32
```

Exercise 2: Find a pair of genes that have a very short distance to each other? Remember this pair and see if you can find this pair in the dendrogram that we will generate later.

For example, MYH16 and AOC1 have a distance of only 9. This is due to that they have such low counts:

```
data.frame(tdf$MYH16,tdf$AOC1)
```

Exercise 3: Can you find your pair of genes, from exercise 2, in the dendrogram above?

Yes, ACO1 cluster with MYH16 (and some other genes with low expression)

Exercise 4: Look at the count data (in "df") and try to understand why the gene FGR is so far away from the other genes in the dendrogram.

Because its expression (the counts) is so much higher compared to all other genes.

Exercise 5: Compare all three dendrograms that you have generated so far and identify differences.

The dendrogram based on the original counts deviates a lot from the other two.

Exercise 6: Do the CVID patient and HD form two distinct clusters? Can you find an individual that has a gene profile that deviates a lot from the other?

Yes, they group very well, except for the healthy donor number 4, which is far away from the other. This person seems to have a gene profile that deviates a lot from the other.

Exercise 7: Try the code below, using the original data, and explain why the heatmap looks very different from the one above.

Since the gene FGR has so much higher expression compared to the other genes, a heatmap based on the original count will only highlight this gene because all other genes have a relatively low expression.