This section will walk you through analyzing differences in gene expression of gastric corpus biopsies, both normal and atrophic. Download the transcriptomic zipped folder and extract it. Go step by step through the R commands to determine the significantly up and down regulated genes. (If you use a non-biolinux machine, in R there are three packages that will need to be installed: tximport, DESeq2, and tidyverse if not already done so.) The commands also assume your working directing is in the transcriptomics folder you extracted. Answer the following questions using the last CSV file created and the correlation plot:

1. How many genes were significantly different (adj-p < 0.05)?

Using instructions in R:

There are 4 genes significantly different.

1. What does it mean when the log2 expression difference is less than 0?

When the log2 expression difference (or log2FoldChange) is less than 0, it means the gene is **downregulated** in the 'atrophic' condition relative to the 'control' condition. The log2FoldChange is calculated as log2. A negative value indicates that the numerator (atrophic expression) is smaller than the denominator (control expression).

1. How many genes were upregulated? Downregulated?

There **3 upregulated** genes and **1 downregulated** gene.

1. How many genes are left if you look at differential expression greater than log2 differences greater than 2? How about less than -2?

There **were 2 genes greater than log2** and **were 2 genes lesser than log2**.

1. Using the correlation plot from the R code, is there a single sample that may influence the overall reliability of the analysis? Explain how you decided on your answer.

Yes, based on the correlation plot, the sample **atr02** appears to be a potential outlier that may influence the overall reliability of the analysis. A high correlation value (close to 1) indicates that two samples have very similar gene expression profiles, while a lower value suggests a difference. When you look at the row for atr02, its correlation values with the other samples (e.g., atr01, atr03, gas01, gas02) are consistently lower than the correlation values among the other samples. This suggests that atr02's expression pattern is less like the rest of the dataset, which could be due to a technical artifact or a unique biological feature, making it an outlier.

For answering these questions, the following R code was used:

# Read the CSV file into a data frame

df <- read.csv("condition\_Atrophic\_vs\_Control.csv")

# Filter for significant genes (adj-p < 0.05)

significant\_genes <- df[df$padj < 0.05, ]

# Filter the data frame for genes with padj < 0.05 and count the rows

count\_significant\_genes <- sum(df$padj < 0.05, na.rm = TRUE)

# Print the result

print(count\_significant\_genes)

# Count upregulated genes (log2FoldChange > 0)

upregulated\_genes <- sum(significant\_genes$log2FoldChange > 0, na.rm = TRUE)

# Count downregulated genes (log2FoldChange < 0)

downregulated\_genes <- sum(significant\_genes$log2FoldChange < 0, na.rm = TRUE)

# Print the results

print(paste("Number of upregulated genes:", upregulated\_genes))

print(paste("Number of downregulated genes:", downregulated\_genes))

# Count genes with a log2FoldChange > 2

count\_log2\_gt\_2 <- sum(significant\_genes$log2FoldChange > 2, na.rm = TRUE)

# Count genes with a log2FoldChange < -2

count\_log2\_lt\_neg\_2 <- sum(significant\_genes$log2FoldChange < -2, na.rm = TRUE)

# Print the results

print(paste("Number of genes with log2FoldChange > 2:", count\_log2\_gt\_2))

print(paste("Number of genes with log2FoldChange < -2:", count\_log2\_lt\_neg\_2))