```r
knitr::opts_chunk$set(warning = FALSE, echo = TRUE)
```

Preparing the dataset

Before the data is analysed by GSVA, a few things need to be as per the GSVA's manual:

- Counts need a valid gene identifier eg. Entrez, ENSEMBL, HGNZ...
- Duplicates need to be removed
- Counts need to be normalised
- A geneset collections needs to created

Sjogren data

To do this, we reead the count data and turn the identifiers into valid ENSEMBL identifiers. Along with this, we read the meta data for this dataset.

```r
phenotype_data <- read.csv(".\\data\\group_to_id.csv")
counts <- read.csv(".\\data\\normalized_data.txt", sep = " ")
```

```r
gene_ids <- sub("\\..*", "", counts$GeneID)

counts$ENSEMBL <- gene_ids
```

Here the duplicates are removed, we find that pseudoautosomal regions are usually causing the duplicates. As these regions have no valid counts, these rows can simply be removed without loss of data.

```r
# Duplicates are from pseudoautosomal regions, IDs are difficult to parse
# No counts found in these regions and thus they are removed
par_y_ids <- counts$GeneID[duplicated(counts$ENSEMBL)]

counts <- counts[!counts$GeneID %in% par_y_ids, ]

row.names(counts) <- counts$ENSEMBL
```

Finally, the counts are turned into a matrix and log-transformed as per the recommendations.

```r
counts_matrix <- data.matrix(counts[, -c(1, 49)])
```

```r
# using recommendations from:
# https://academic.oup.com/bib/article/22/1/545/5722384#225532952
# OUTRIDER data is normalized with FPKM, thus we take the log from that
normalized_matrix <- log2(1 + counts_matrix)
```

Pathways data

For the pathways, we use the pathways found in wikipathways. This is an open, collaborative platform dedicated to the curation of biological pathways.

```
wiki_set <- getGmt(".\\data\\wikipathways-20250210-gmt-Homo_sapiens.gmt",
                   geneIdType = EntrezIdentifier())

wiki_set_ens <- mapIdentifiers(wiki_set, ENSEMBLIdentifier("org.Hs.eg.db"))
```

GSVA analysis

With all the data collected, a GSVA run can be done and the results can be analysed. We use the same type of analysis used in the previously mentioned manual.

GSVA run

The run itself is broken into two parts. The first part is creating the GSVA parameters object. This object combines the normalized counts and the geneset collection, along with additional information such as which gene identifier is being used, the minimum size that a pathway will need to be, and the distribution used.

```
gsva_param <- gsvaParam(normalized_matrix, wiki_set_ens, kcdf = "Gaussian",
                        annotation = ENSEMBLIdentifier(), minSize = 2)
```

In the second part, the GSVA is run with the previously made parameter object.

```
gsva_result <- gsva(gsva_param)
```

GSVA results

For analysing the results, we perform a differential analysis on the GSVA results. This is done to get the signficance of the difference in pathway expression between the different samples.

```
# Create a design matrix for linear model
design_matrix <- model.matrix(~factor(phenotype_data$group))
colnames(design_matrix) <- c("ALL", "sjogrenVScontrol")
# Fit linear model
fit <- lmFit(gsva_result, design_matrix)
# Empirical bayes, calculate a prior distribution and use that
# to compute t-statistic
fit <- eBayes(fit)
# Summarizes result of model, p-value is adjusted using BH-method
res <- decideTests(fit, p.value = 0.05)
summary(res)
```

```
##        ALL sjogrenVScontrol
## Down     0                0
## NotSig 908              898
## Up       0               10
```

We adjust for the geneset size as this appears to have an effect on the performance of the model. When a geneset is smaller, higher amounts of variance is found.

```r
# Check to see if geneset size has effect on residual error
geneset_sizes <- geneSetSizes(gsva_result)

plot(geneset_sizes, fit$sigma, xlab = "gene sets sizes",
     ylab = "Standard deviation of residuals", las = 1, pch = ".", cex = 3)
```

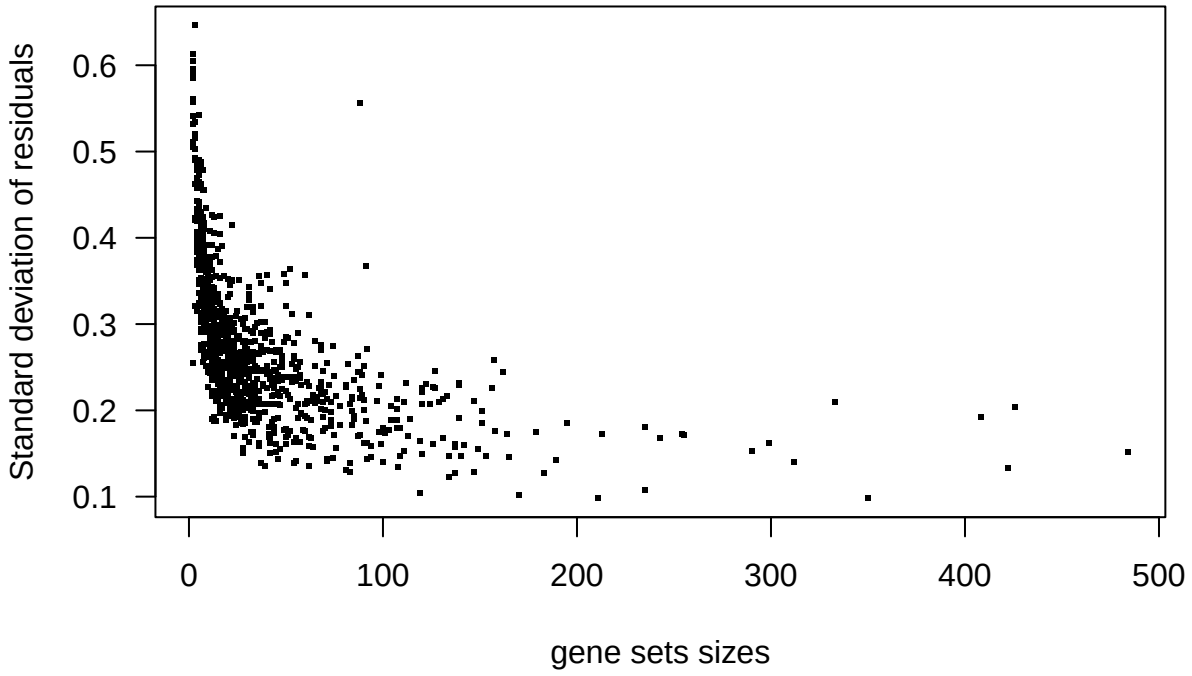


Figure 1: A scatter plot where all pathways are mapped. The x-axis represents the gene set size, and the y-axis represents the standard deviation of the residuals. As the size increases, the deviation decreases.

```r
fit <- eBayes(fit, trend = geneset_sizes)
# Summarizes result of model, p-value is adjusted using BH-method
res <- decideTests(fit, p.value = 0.05)
summary(res)
```

```
##        ALL sjogrenVScontrol
## Down     0                0
## NotSig 908              899
## Up       0                9
```

After re-fitting the model, now taking into account how geneset size can effect the results, we visualise the pathways that were found in a volcano plot.

```r
result_table <- topTable(fit, coef = 2, n = Inf)
result_table$dif_expressed <- "NOT SIGNIFICANT"
result_table$dif_expressed[result_table$adj.P.Val < 0.05] <- "SIGNIFICANT"

ggplot(data = result_table,
       aes(x = logFC, y = -log10(adj.P.Val), col = dif_expressed)) +
  geom_point() +
  geom_hline(yintercept = -log10(0.05),
             col = "black", linetype = "dotted") +
  theme_minimal()
```

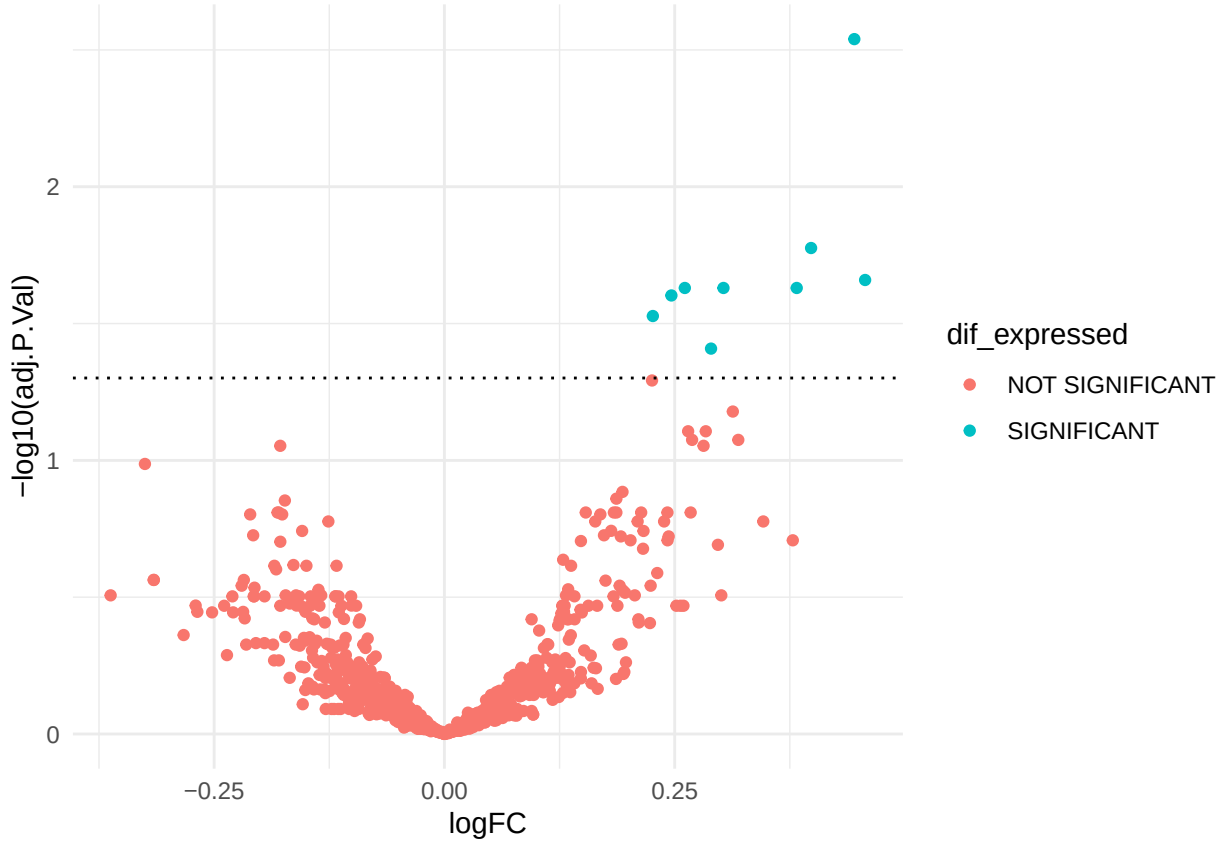
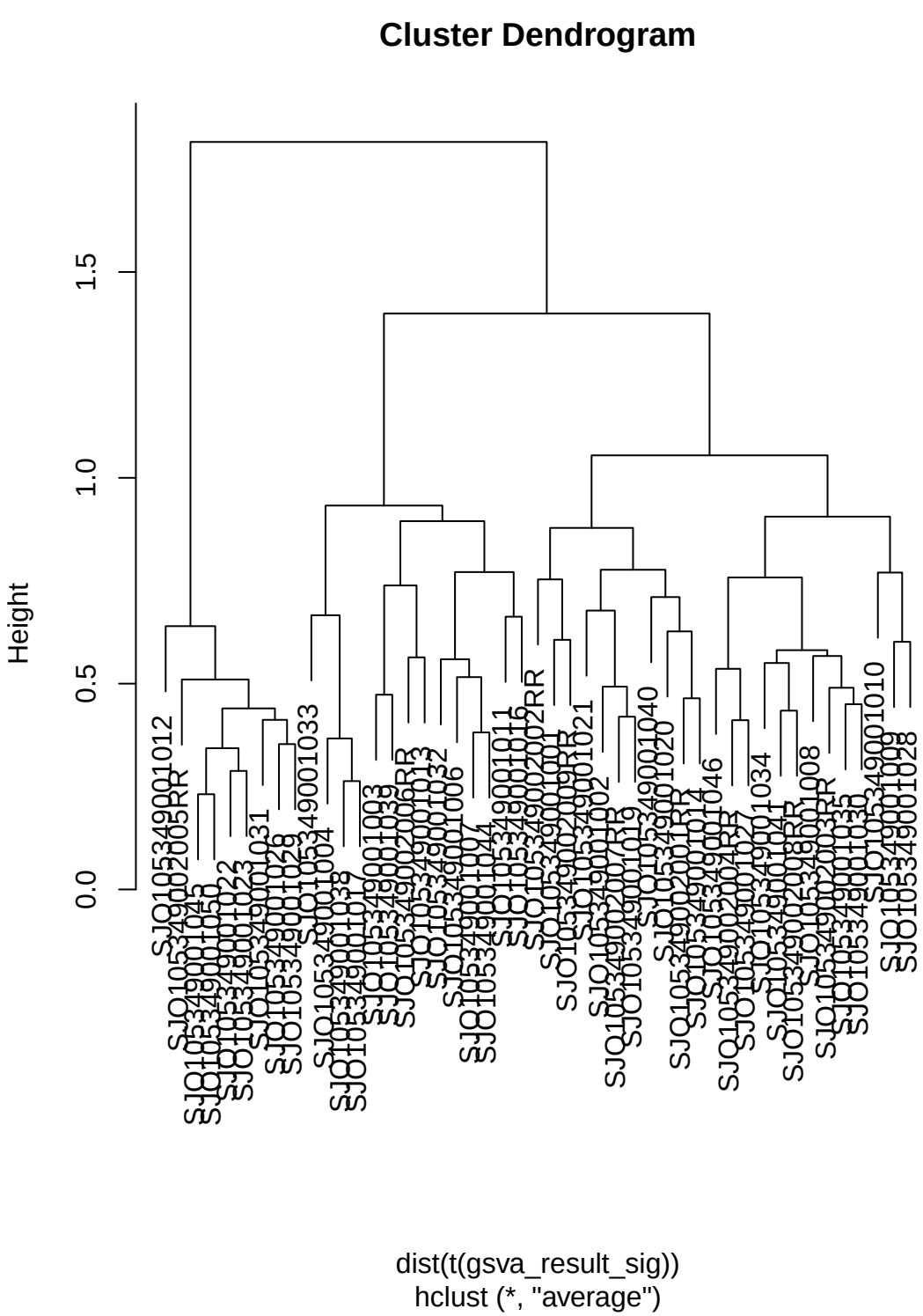


Figure 2: A volcano plot displaying which pathways are found to be significant. The y-axis represent the significane of the pathway and the x represents the difference of pathway expression between control and sjogren samples. The dotted line represents the spot where signifance is equal to 0.05.

For the next steps, we look at heatmaps and the clusters around it. We create two heatmaps, one with R's own heatmap function and another with GGplot. This is done because R includes a dendogram, whereas GGplot's map does not include this feature, but ggplot's plot are generally a little clearer. This can later be narrowed donw to simply using one heatmap.

First hierarichal clustering is done, the distance method used here is eulcidean and the linkage method is average.

```r
clust_obj <- hclust(dist(t(gsva_result_sig)), method = "average")
plot(clust_obj)
```


Cluster Dendrogram
Height
0.0
0.5
1.0
1.5
dist(t(gsva_result_sig))
hclust (*, "average")

Following such, the heatmap from base R is created. In both heatmaps, control samples are given a orange colour and the sjogren samples are given a green colour. The more up regulated a pathway is, the darker the red, the more down regulated a pathway is, the darker the blue.

```
phenotype_data$colour <- "#008B33"
phenotype_data$colour[phenotype_data$group == "sjogren"] <- "#e4a22f"
col <- colorRampPalette(c("#005AB5", "white", "#DC3220"))(256)

heatmap(gsva_result_sig, ColSideColors = phenotype_data$colour,
        col =  col, Colv = as.dendrogram(clust_obj), cexRow = 0.5)
legend(x = "topright", legend = c("control", "sjogren"),
       fill = c("#008B33", "#e4a22f"))
```

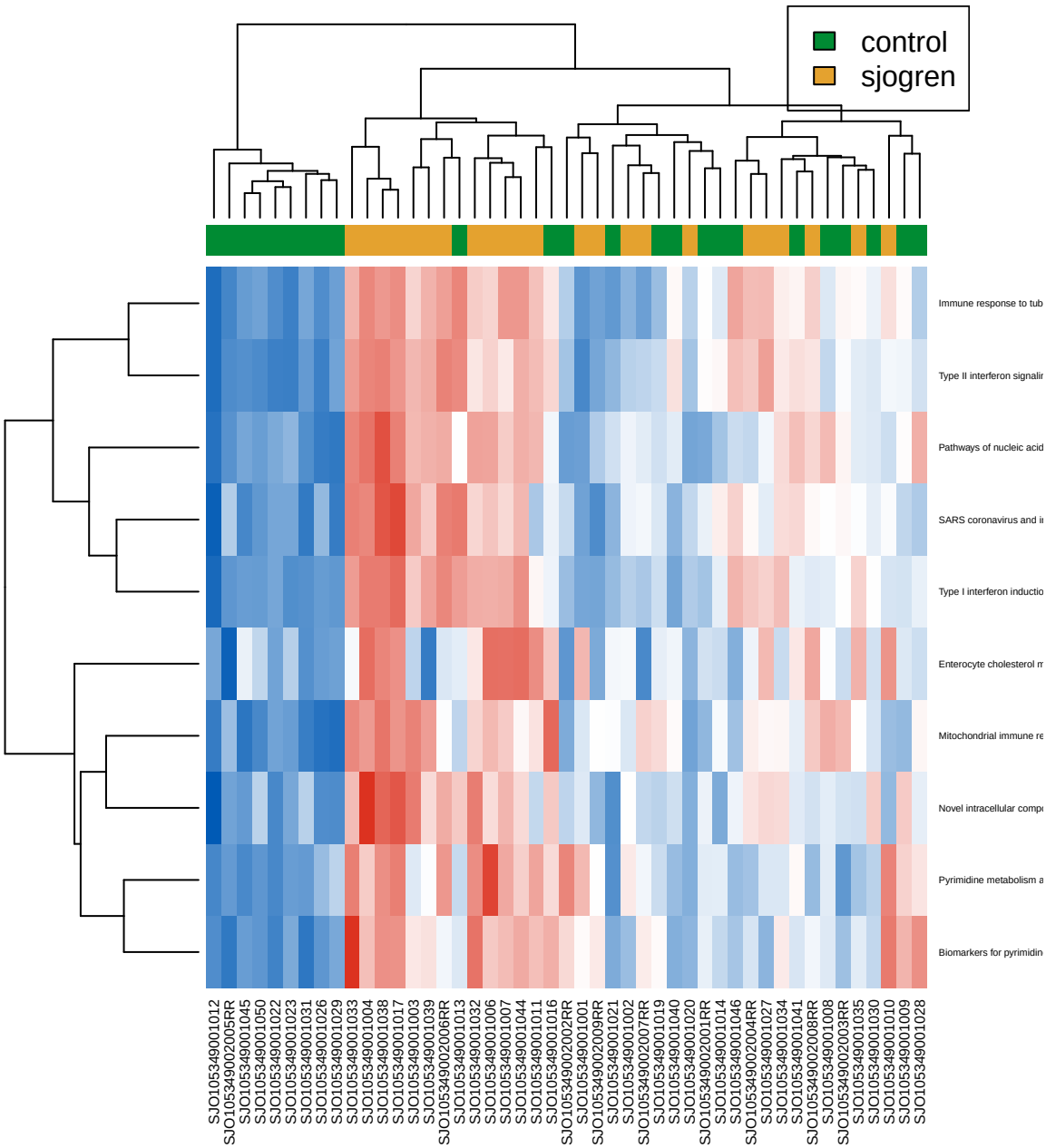


Figure 3: A base R eatmap displaying the samples and signifcant pathways found in the GSVA. Control samples are given a orange colour and the sjogren samples are given a green colour. The more up regulated a pathway is, the darker the red, the more downregulated a pathway is, the darker the blue.

For the ggplot heatmap, the data is transformed into a long format. Due to ggplot's lack of dendogram, we reorder the samples by using the earlier created cluster object.

```
result_table <- topTable(fit, coef = 2)
gsva_result_sig <- gsva_result[rownames(result_table), ]
rownames(gsva_result_sig) <- gsub("\\%.*", "", rownames(gsva_result_sig))
# Reorder by cluster
gsva_result_sig <- gsva_result_sig[, clust_obj$order]
phenotype_data <- phenotype_data[clust_obj$order, ]

long_gsva_result <- melt(gsva_result_sig)
colnames(long_gsva_result) <- c("pathways", "ID", "pathway_score")
```

Finally the heatmap is generated.

```
ggplot(long_gsva_result, aes(x = pathways, y = ID, fill = pathway_score)) +
  geom_tile() +
  scale_fill_gradient2(high = "#DC3220",
                       mid = "#f3f3f3",
                       low = "#005AB5") +
  theme(axis.text.x = element_text(angle = 90, vjust = 0.5, hjust = 1),
        axis.text.y = element_text(colour = phenotype_data$colour))
```

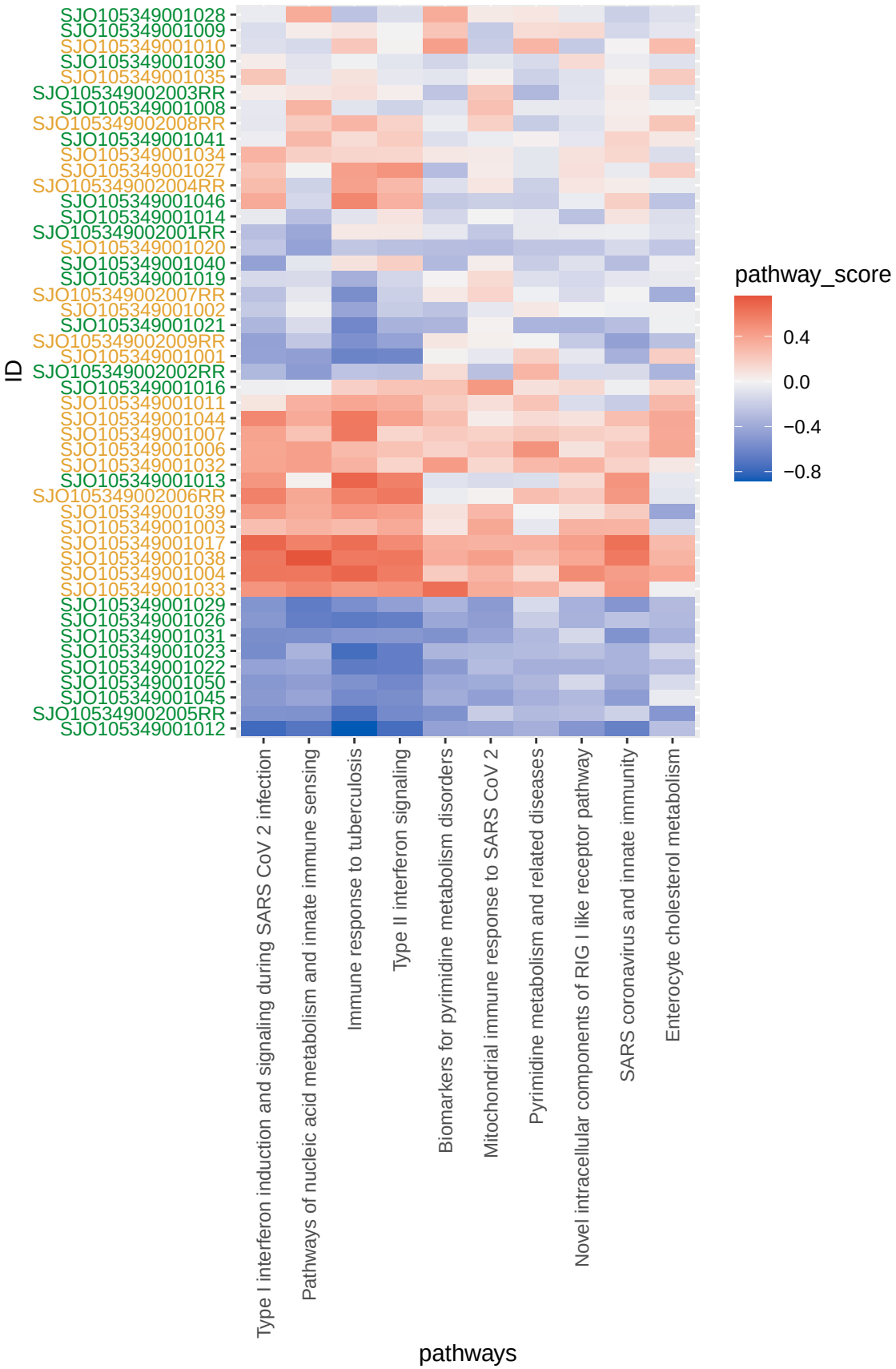


Figure 4: A GGplot heatmap displaying the samples and signifcant pathways found in the GSVA. Control samples are given a orange colour and the sjogren samples are given a green colour. The more up regulated a pathway is, the darker the red, the more downregulated a pathway is, the darker the blue.

From these heatmaps, the following things are seen:

- Several Sjogren samples are more closely related to control samples when looking at pathway expression. This lines up to what we know of Sjogren, which is a highly heterogenous disease. Thus, we don't expect all sjogren samples to map closely together.
- Pathways in Sjogren samples are often more up-regulated when compared to the control samples
- There are quite some COVID pathways being seen as signifcant. This might be because COVID pathways are over-represented in the WikiPathways dataset.

For further analysis, we can look at the genes found in the pathways that are significantly different. We turn the ENSEMBL ideas into HGNC symbols. We then writes this to a .csv, this file can be used to search the VIP results.

```
genes_in_pathways <- mapIdentifiers(wiki_set_ens[rownames(gsva_result_sig)],
                                    SymbolIdentifier("org.Hs.eg.db"))

genes_in_pathways <- geneIds(genes_in_pathways[rownames(gsva_result_sig)])
```

```
# List to long format data frame
genes_in_pathways <- stack(genes_in_pathways)
# Write to file for further analysis
write.csv(genes_in_pathways, "found_genes.csv", row.names = FALSE)
```