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

```r
library("DESeq2")
library("ggplot2")
library("factoextra")
library("ggfortify")
library("org.Hs.eg.db")
library("clusterProfiler")
```

```r
raw_counts <- read.csv(".\\data\\count_data\\merged_outrider_counts.txt",
                       sep = "\t")
phenotype_data <- read.csv(".\\data\\group_to_id.csv")

row.names(raw_counts) <- raw_counts$GeneID

raw_counts <- raw_counts[-c(1)]

# Remove zero count rows
raw_counts <- raw_counts[rowSums(raw_counts) > 0, ]
# # Remove genes where 75% of the samples have zero counts found
raw_counts <- raw_counts[rowMeans(raw_counts == 0) <= 0.75, ]

row.names(phenotype_data) <- phenotype_data$ID
phenotype_data$group <- as.factor(phenotype_data$group)
```

```r
# Construct dds object
dds <- DESeqDataSetFromMatrix(countData = raw_counts,
                              colData = phenotype_data,
                              design = ~ group)
```

DEseq does some normalisation for us. It does the following: - It estimates the size factors so that samples with different sequencings depths can be compared. - It estimates dispersion, giving an estimation of how scattered the data is - Negative binomial GLM fitting, and tests whether there's difference between control and patients with Sjogren

Here, as we are looking for differences between groups rather than between samples, this type of normalisation is prefered to the OUTRIDER normalisation, as OUTRIDER normalisation is done so that outliers are still notable.

```r
# run DEseq
dds <- DESeq(dds)
```

```
## estimating size factors

## estimating dispersions

## gene-wise dispersion estimates

## mean-dispersion relationship

## final dispersion estimates
```

```
## fitting model and testing

## -- replacing outliers and refitting for 2378 genes
## -- DESeq argument 'minReplicatesForReplace' = 7
## -- original counts are preserved in counts(dds)

## estimating dispersions

## fitting model and testing
```

```
res <- results(dds)

sig_res <- as.data.frame(res[which(res$padj <= 0.05 &
                                   abs(res$log2FoldChange) > 1), ])
```

PCA

For PCA of these results, we follow DESeq2 guidelines and perform a variance stabilising transformation. This will make sure the counts are homoscedastic, that is when the variance is constant instead of being higher when counts are higher.

The parameter 'blind' is set to TRUE so that information about the groups is not used in this transformation.

```
vsd_blind <- vst(dds, blind = TRUE)
```

We now run the PCA, but as we have many genes, we only run a PCA with a thousand genes. These genes are the genes that have the most variance. After that, a PCA is ran with R's prcomp function and is plotted.

```
# Get variance per rows
rv <- rowVars(assay(vsd_blind))
# Select the genes that are in the variance top 1000
select <- order(rv, decreasing = TRUE)[seq_len(min(1000, length(rv)))]
# Run PCA with the selected samples
pca_res_blind <- prcomp(t(assay(vsd_blind)[select, ]))
```

A thousand genes will create quite a few loadings, and so we choose to only display the top 20 loadings for visibility's sake.

```
contribution_object <- fviz_contrib(pca_res_blind,
                                    choice = "var", axes = 1:2, top = 20)

contributions <- contribution_object$data

top_contrib <- rownames(contributions[order(contributions$contrib,
                                            decreasing = TRUE), ][1:20, ])
```

Finally, we can plot the results of the PCA and add the loadings. When interpreting this plot, keep in mind we only took the 1000 genes that had the most variance into account. This means the genes that did not show much variantion, but that could play a role in differentiating sjogren from control patients could have been ignored. However, as we did do a differential expression analysis as well, these genes will come through there.

```
loadings <- pca_res_blind$rotation[top_contrib, 1:2]

autoplot(pca_res_blind, data = phenotype_data,
         col = "group", label.label = "ID", label = TRUE) +
  geom_segment(data = loadings, aes(x = 0, y = 0, xend = PC1, yend = PC2),
               arrow = arrow(length = unit(0.1, "in")),
               col = "brown") +
  geom_text(data = loadings, aes(x = PC1, y = PC2, label = top_contrib),
            nudge_y = 0.001, size = 3) +
  scale_x_continuous(expand = c(0.02, 0.02))
```

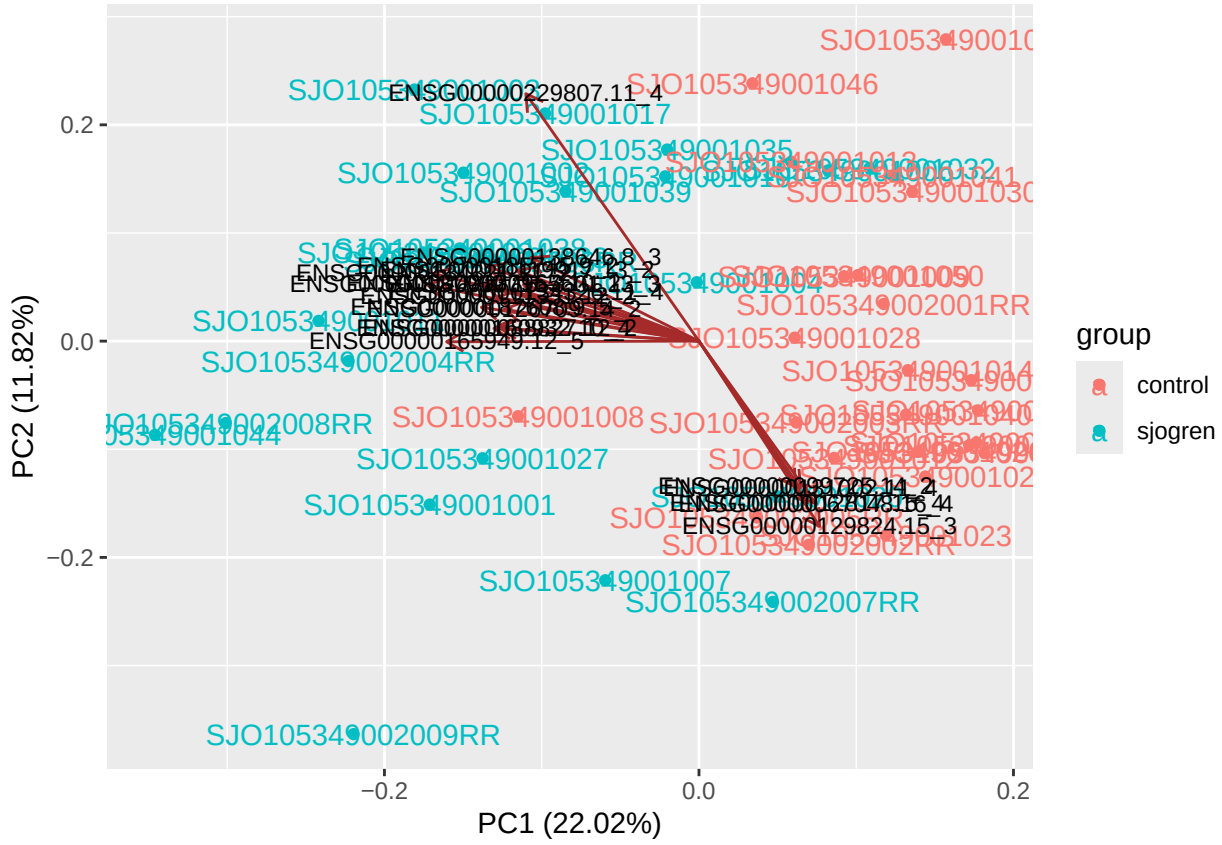


Figure 1: A PCA plot, the first component on the x-axis and the second component on the y-axis. There's a seperation between the sjogren (blue) samples and the control (pink) samples, though not all samples are equally seperated.

As expected, since the PCA does not take different groups into account, there are a few genes that have quite some variance between samples but are not expressed differently in a signficant way between sjogren and control samples. As these genes do still effect the data, they could be batch effects and thus still have an interesting effect on the data.

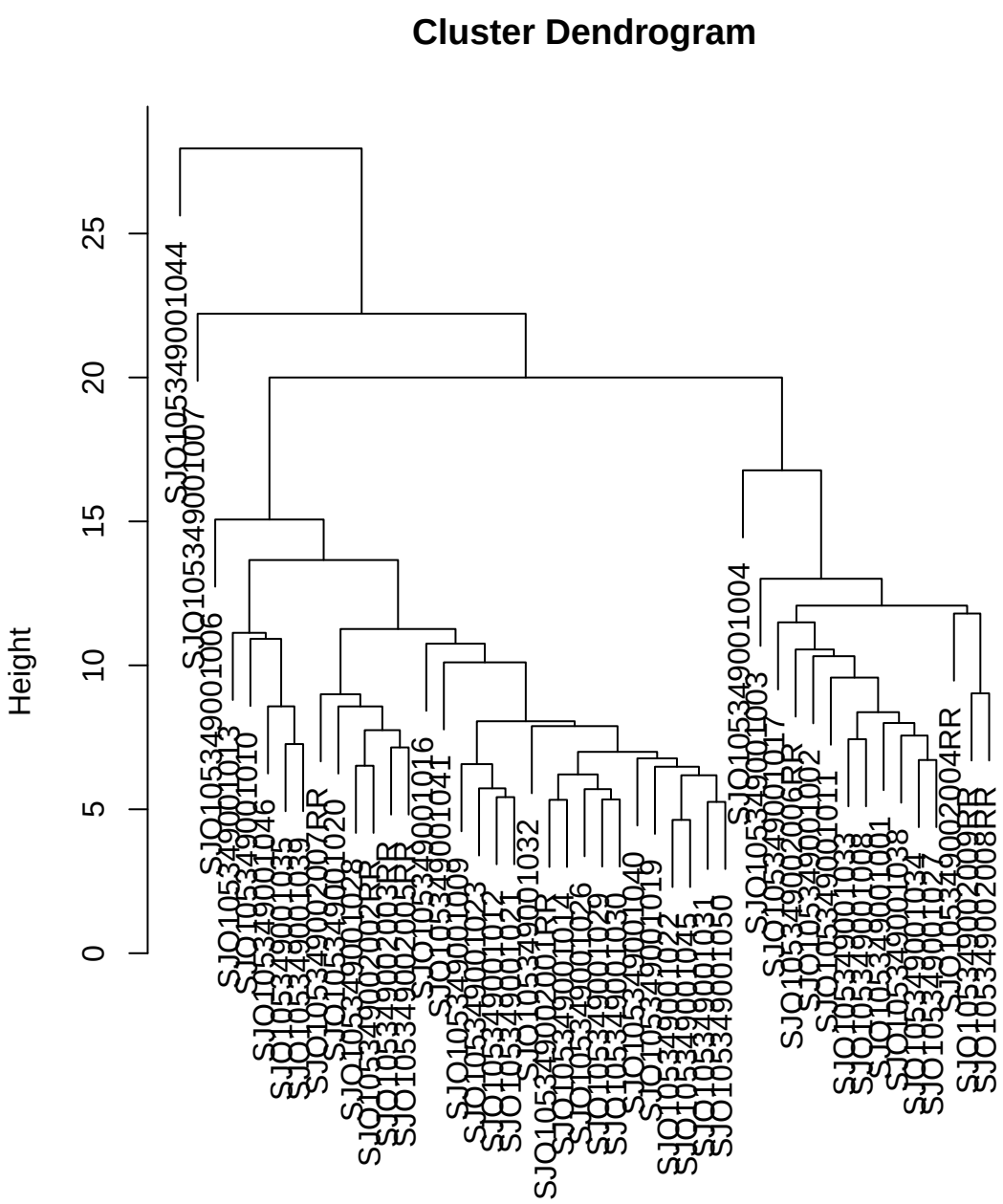
```
normalized_counts <- as.data.frame(assay(vsd_blind))
sig_gene_names <- row.names(sig_res)
significant_counts <- normalized_counts[sig_gene_names, ]
```

Heatmap

For another visualisiton, simply to explore the data, a heatmap is created. As there many counted genes, we only create a heatmap using the significantly expressed genes. We can then see how the samples cluster together based on the expression of the genes. We except to see to see the same pattern we find in the pathway analysis, a few controls clustering together very well but remaining controls mixing in with the sjogren samples.

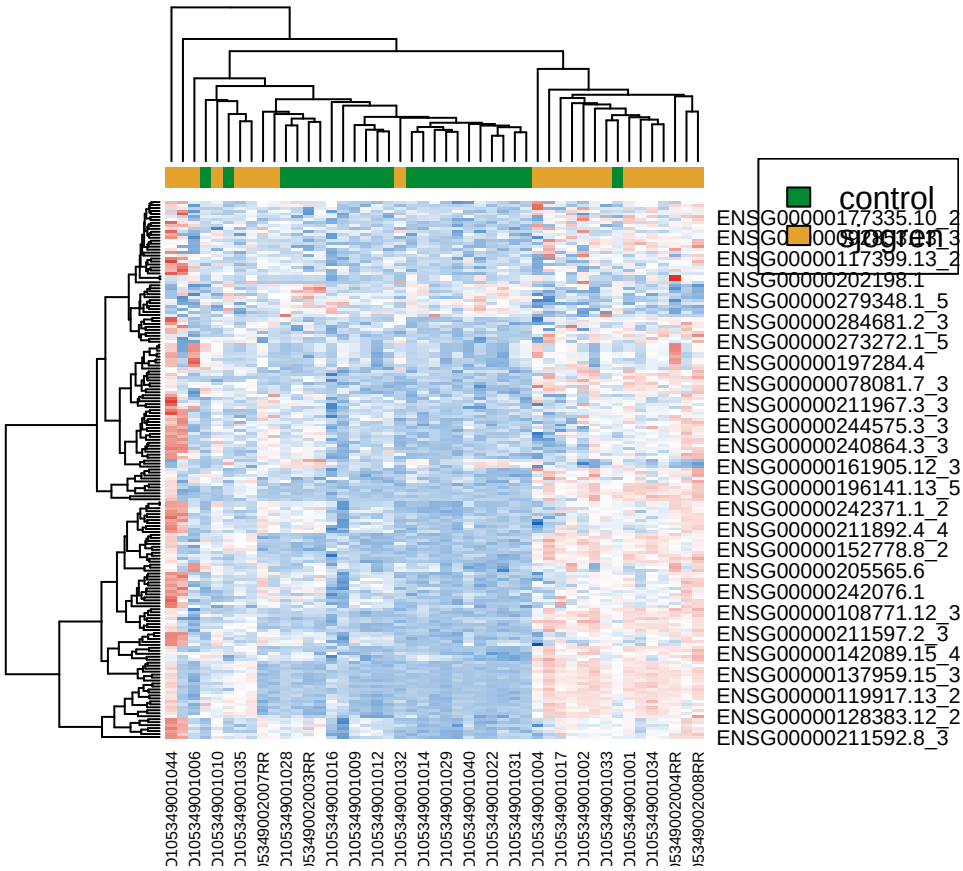
We use the same clustering method used in the gsva analysis.

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




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

```
heatmap(count_matrix, ColSideColors = phenotype_data$colour,
        col = col, cexRow = 1, Colv = as.dendrogram(clust_obj))
legend(x = "topright", legend = c("control", "sjogren"),
       fill = c("#008B33", "#e4a22f"))
```



First, in the heatmap, two sjogren samples split off at the very start. These samples appear to have genes that are very strongly upregulated. After these two samples have split off, we find a that one group of sjogren samples split off rather well, but others mix more with the control.

Finding potentially interesting genes

To find potentially interesting genes, we can look at the top contributing genes and the significantly expressed genes. We already noted how some of the top contributing genes are not signifcatly different between groups but they could help explain batch effects which are also interesting to take into consideration.

From the top 20 contributing genes in the PCA, 14 of them are also significantly different between sjogren and control samples.

```
contributing_res <- sig_res[top_contrib, ]
sum(complete.cases(contributing_res))
```

```
## [1] 14
```

The remaining 6 are not different between the groups, so let's take a look at them to see what these genes are about. Searching information on OMIM, we find the following:

- Five out of the six genes that are either X-linked or Y-linked.
- The gene that is not linked to either the X or Y chromosomes, MX1, is linked to a protein that combats viral infections.

MX1 has been studied in relation to Sjogren's disease, and there is believe that it is associated with the disease. However, it is also seen in patients with other auto immune disease. As the controls are not healthy individuals and are instead taken from the immunology department of the UMCG, this could be a reason it did not show up as significant.

Another source of information can be finding variants in the found genes. We have the VIP results, and thus let's look at how many of theses genes have variants associated with them. As VIP uses HGNC symbols, we convert our found genes symbols from ENSEMBL to HGNC.

```
gene_ids <- sub("\\..*", "", top_contrib)
hgnc_contrib <- bitr(gene_ids,
                     fromType = "ENSEMBL",
                     toType = "SYMBOL",
                     OrgDb = "org.Hs.eg.db")
```

```
## 'select()' returned 1:1 mapping between keys and columns
```

One of the genes is not related to a HGNC symbol, a quick search of this IDs finds that this gene is a pseudo gene and is related to PRKY.

Following that, we can connect the two results using the following code:

```
vip_result <- read.csv(".\\data\\merged_all_variants.csv")
vip_contrib <- vip_result[vip_result$HGNC_SYMBOL %in% hgnc_contrib$SYMBOL, ]
```

Only one gene that is found in the signifcant genes and is in the top contributing genes, also has a found variant. The SIGLEC1 gene with two found variants in sample J26. This gene is not found in any signifcantly expressed pathways. It is associated with assessment of type I interferon activity. This gene is known to be associated with Sjogren's disease, specifically with a subtype that indicates extraglandular manifestations.

From the remaining 13 genes, they all appear to be related to interferons. This relation can be either being induced or stimulated by interferons. Interferons are well-known to have a role in Sjogren's disease.

IFI44L, IFIT1, and IFIT3 are known to be regulated by type 1 interferons. However, we don't find any variants in these genes. Methylation of the area might help explain the differences we see in expression. We can not confirm this, as it is out the scope of this research, but we can point to the following study. The study shows that Sjogren associated differential methylated positions can be found within these genes. We find several other genes in this study that are showing up in our analysis as well. Amongst them are MX1, CMPK2, LY6E, HERC5, and RSAD2.

Vip variants

Putting aside the results of PCA and just looking at the variants we find in VIP that we also find in the DEG, we can find four unique genes, amongst them the previously discuseed SIGLEC1 gene.

We discuss the rest of the found variants and their genes here.

ANK1, two mutations found in one sample, not known to be related to Sjogren. Predicted to be pathogenic by VIP, but ClinVar estimates it to benign. Mutations in this gene have been known to be related to Spherocytosis, type 1 which is a disorder that effects red blood cells.

PSMA3, one mutation found in one sample, is not known to be related to Sjogren specifically but does appear to be overexpressed in auto-immune diseases including Sjogren.

FCGBP, one mutation found in seven samples, and the gene has been found in a study that looked for SNPs related to Sjogren.