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

Pathway result exploration

To further visualise the results of the pathway analysis, we run a Principal Component Analysis(PCA). This allows us to see where the samples lay within the data space and how the different pathways effect the position of the samples.

We can look at a PCA in two ways. We can look at all the found pathways, or we can look at the pathways that were significantly different between the control and the sjogren samples. The first option is highly dimensional, as over 900 pathways were found. The second option risks being too simplified, where we risk missing key components. Therefor we can look at both options.

We import the information gathered by the GSVA pathway analysis.

```
gsva_pathway_scores <- read.csv(".\\data\\pathway_data\\pathway_matrix_result.csv",
                                row.names = 1)
gsva_pathway_scores_sig <- read.csv(".\\data\\pathway_data\\sig_found_pathways.csv",
                                    row.names = 1)
phenotype_data <- read.csv(".\\data\\group_to_id.csv")
```

As PCA looks at the variance, highly correlated pathways will mean that one of those pathways doesn't bring new information into the PCA. We therefor choose to remove highly correlated pathways (>=0.95).

```
gsva_matrix <- t(gsva_pathway_scores)
corr_matrix <- cor(gsva_matrix)

correlated_values <- findCorrelation(corr_matrix, cutoff = 0.95)

dim(gsva_matrix)
```

```
## [1]  47 908
```

```
gsva_df <- as.data.frame(gsva_matrix[, -correlated_values])
dim(gsva_df)
```

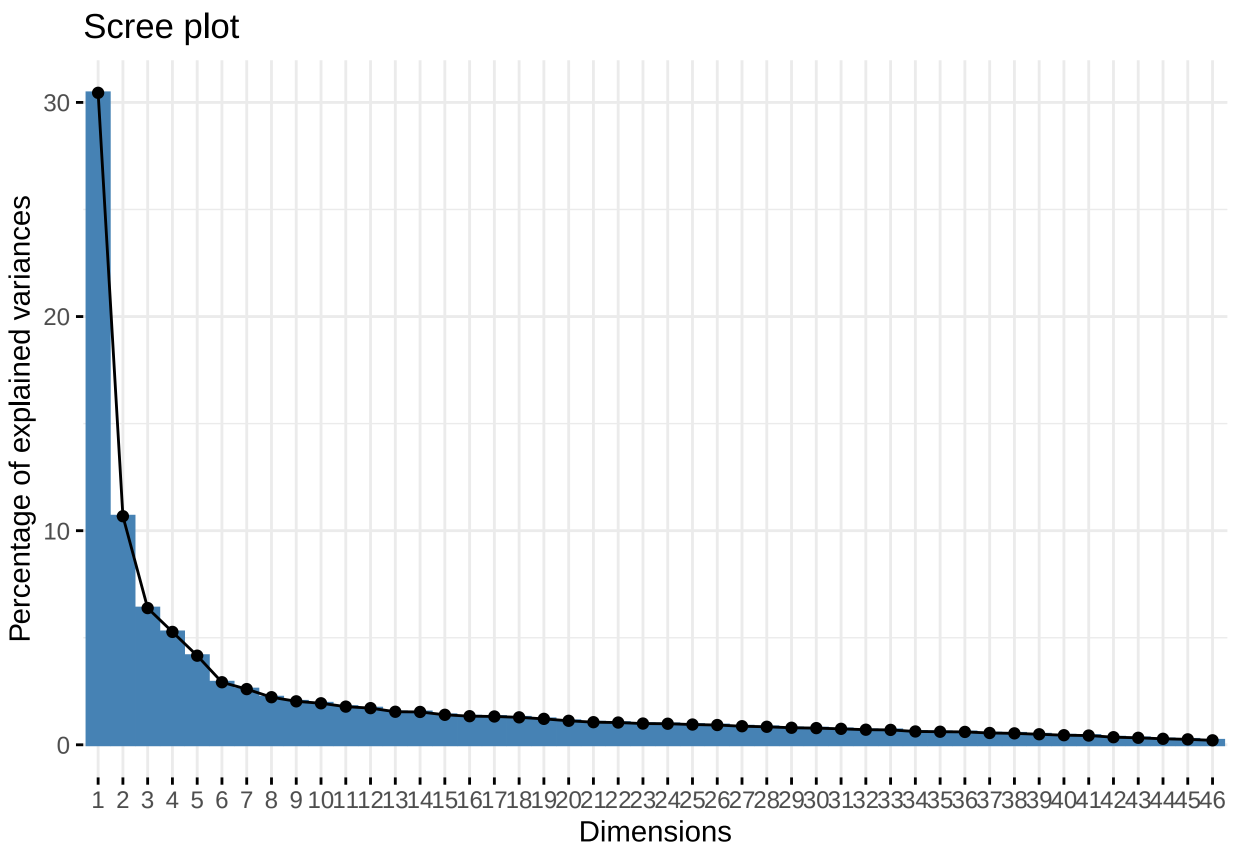
```
## [1]  47 883
```

To run the PCA itself, we center the data but we do not scale it as the data is already scaled for us. As the screeplot shows, we see the first component captures around 30% of the variance. Together with the second component, around 40% of the variance is captured. To capture 100% of the variance, 46 components are needed. This is to be expected as we only have 47 samples in total. Around the 6th component, we capture around 60% of the variance and we see the amount of variance captured after that gets quite small.

```
# We do not use scale as data is already similar in scale
# This prevents loss of information
res_pca_all_pathways <- prcomp(gsva_matrix, scale = FALSE, center = TRUE)
summary(res_pca_all_pathways)
```

```
## Importance of components:
##                           PC1    PC2     PC3     PC4     PC5     PC6     PC7
## Standard deviation     4.7458 2.8101 2.17323 1.97482 1.75461 1.46999 1.38708
## Proportion of Variance 0.3045 0.1068 0.06385 0.05272 0.04162 0.02921 0.02601
## Cumulative Proportion  0.3045 0.4112 0.47509 0.52781 0.56943 0.59864 0.62465
##                            PC8    PC9    PC10    PC11    PC12    PC13    PC14
## Standard deviation     1.28224 1.2254 1.19772 1.14837 1.12558 1.06838 1.06548
## Proportion of Variance 0.02223 0.0203 0.01939 0.01783 0.01713 0.01543 0.01535
## Cumulative Proportion  0.64688 0.6672 0.68657 0.70440 0.72153 0.73696 0.75231
##                           PC15    PC16    PC17    PC18    PC19    PC20    PC21
## Standard deviation     1.01806 0.99422 0.98913 0.97388 0.94666 0.91078 0.88403
## Proportion of Variance 0.01401 0.01336 0.01323 0.01282 0.01212 0.01121 0.01057
## Cumulative Proportion  0.76632 0.77968 0.79291 0.80573 0.81785 0.82906 0.83963
##                           PC22    PC23    PC24    PC25    PC26    PC27    PC28
## Standard deviation     0.87683 0.85662 0.85306 0.83692 0.82648 0.80114 0.78897
## Proportion of Variance 0.01039 0.00992 0.00984 0.00947 0.00923 0.00868 0.00842
## Cumulative Proportion  0.85002 0.85994 0.86978 0.87925 0.88848 0.89716 0.90557
##                           PC29    PC30    PC31    PC32    PC33    PC34    PC35
## Standard deviation     0.76746 0.75931 0.74259 0.72212 0.71753 0.67848 0.67208
## Proportion of Variance 0.00796 0.00779 0.00745 0.00705 0.00696 0.00622 0.00611
## Cumulative Proportion  0.91354 0.92133 0.92879 0.93584 0.94280 0.94902 0.95513
##                           PC36    PC37    PC38    PC39    PC40    PC41    PC42
## Standard deviation     0.66711 0.63905 0.62857 0.60359 0.57556 0.56572 0.51256
## Proportion of Variance 0.00602 0.00552 0.00534 0.00493 0.00448 0.00433 0.00355
## Cumulative Proportion  0.96114 0.96666 0.97200 0.97693 0.98141 0.98574 0.98929
##                           PC43   PC44    PC45    PC46      PC47
## Standard deviation     0.49209 0.4554 0.43403 0.39306 2.108e-15
## Proportion of Variance 0.00327 0.0028 0.00255 0.00209 0.000e+00
## Cumulative Proportion  0.99256 0.9954 0.99791 1.00000 1.000e+00
```

```
fviz_eig(res_pca_all_pathways, col.var = "darkblue", ncp = 46)
```

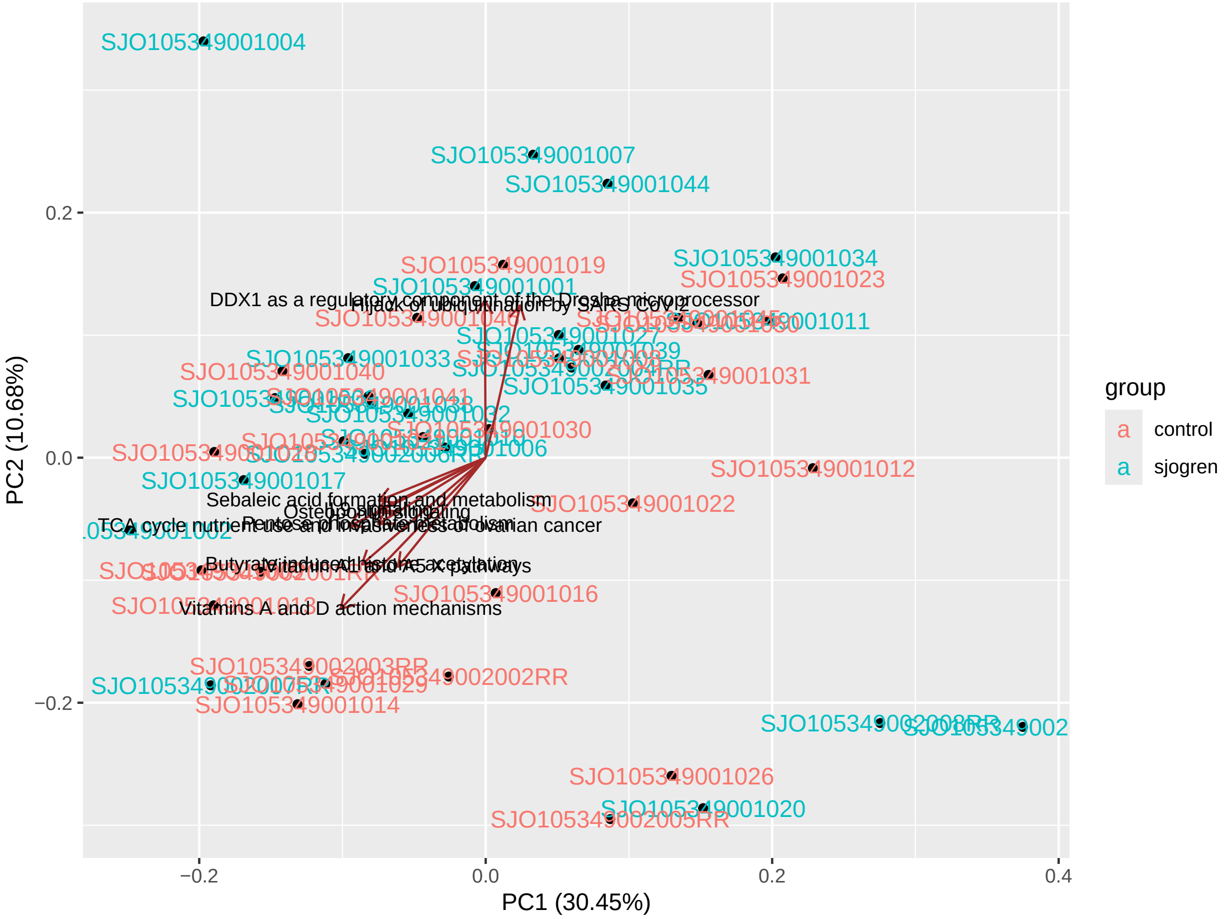


We gather the top contributing pathways to use as loadings in the PCA plot later on.

```
res_contrib <- get_pca_var(res_pca_all_pathways)$contrib

contribution_object <- fviz_contrib(res_pca_all_pathways,
                                    choice = "var", axes = 1:2, top = 10)

contributions <- contribution_object$data

top_contrib <- rownames(contributions[order(contributions$contrib,
                                            decreasing = TRUE), ][1:10, ])
```

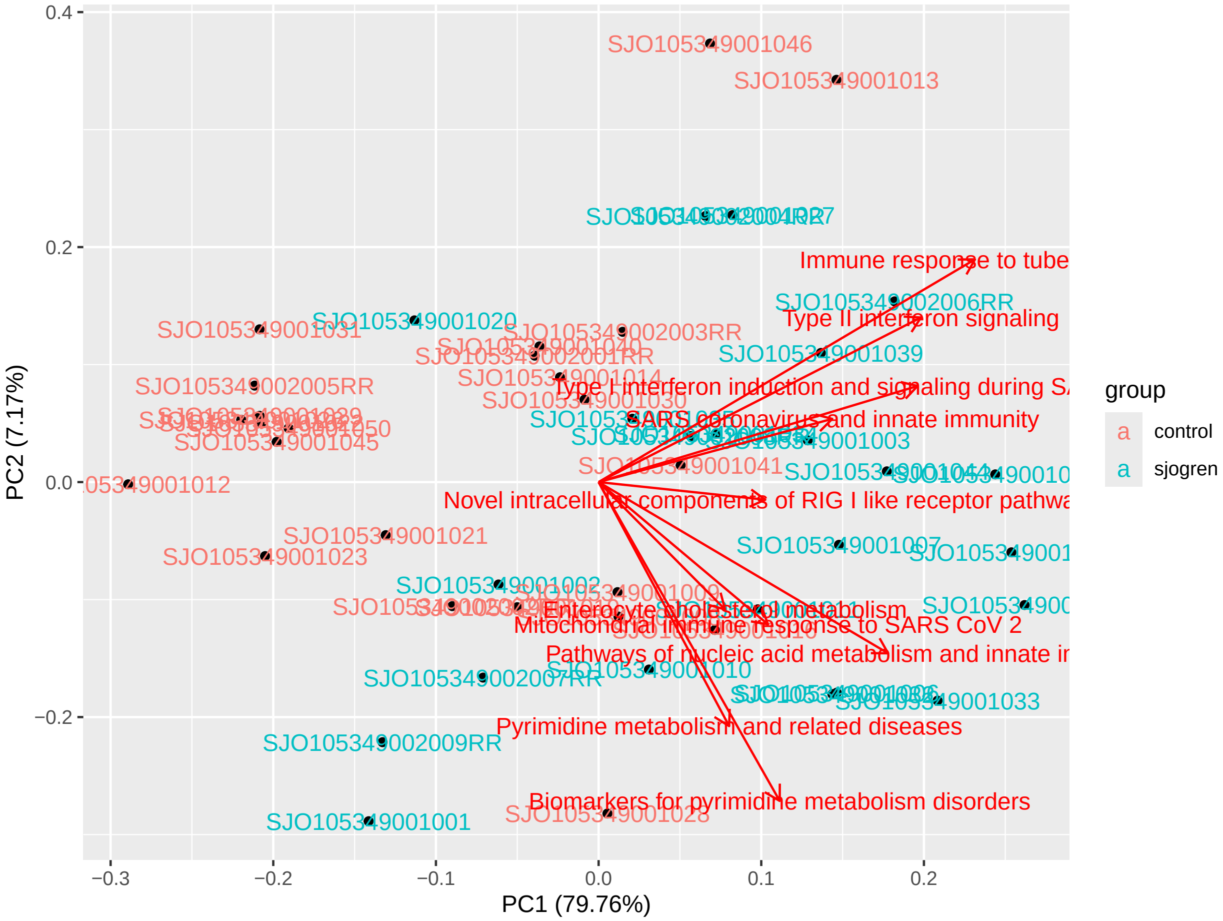
When we plot the PCA, we see very little seperation from control to sjogren. This is likely because we are looking much at total variance, not so much the variance across the different groups. This means batch effects may play a greater role in these results.

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

autoplot(res_pca_all_pathways, data = phenotype_data
         , label = TRUE, label.label = "ID",
         label.colour = "group") +
    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 = gsub("\\%.*", "", top_contrib)),
          nudge_y = 0.001, size = 3) +
  scale_x_continuous(expand = c(0.02, 0.02))
```



For a similar look into the data, we can choose results that were filtered on significance already. This leaves a lot of details behind but it will give a good visualistion to go along with the heatmap created in the gsva analysis.

```
gsva_matrix <- t(gsva_pathway_scores_sig)
corr_matrix <- cor(gsva_matrix)

correlated_values <- findCorrelation(corr_matrix, cutoff = 0.95)

gsva_df <- as.data.frame(gsva_matrix[, -correlated_values])

clust_obj <- hclust(dist(gsva_matrix), method = "average")

cluster_groups <- cutree(clust_obj, k = 2)
phenotype_data$cluster_id <- as.factor(cluster_groups)

res_pca_sig_pathways <- prcomp(gsva_matrix, scale = FALSE, center = TRUE)

autoplot(res_pca_sig_pathways, data = phenotype_data,
         label = TRUE, label.label = "ID",
        label.colour = "group", loadings = TRUE, loadings.label = TRUE)
```



As expected, we see a better seperation of control and sjogren samples, but there remain samples that aren't able to be seperated from a group. This was expected, it is heterogenous disease after all, but now we see the way in which the pathways affect the position of the samples, and we can see how certain samples are related to one another.