Project Network-based Data Analysis

Annalisa Xamin

Contents

Data selection

The analysis will use the dataset GSE20437 obtained from GEO.The dataset is generated from Affymetrix HU133A microarrays and contains 42 tissue samples.

In detail, the data includes:

- 18 reduction mammoplasty (RM) breast epithelium samples,
- 18 histologically normal (HN) epithelial samples from breast cancer patients (9 ER+ and 9 ER-), and
- 6 histologically normal epithelial samples from prophylactic mastectomy patients.

Note that sample numbers correspond to individual patient samples.

```
# download the GSE20437 expression data series
#gse <- getGEO("GSE20437", destdir= './data/', getGPL = F)
```

```
# load the local copy of the data
gse <- getGEO(file = "./data/GSE20437_series_matrix.txt.gz", getGPL = FALSE)
```

```
# getGEO returns a list of expression objects, but...
length(gse)
```

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

```
# shows us there is only one object in it.
# We assign it to the same variable.
#gse <- gse[[1]] # run only if you download data and
# if you don't use the local copy
```

```
# extract metadata
metadata <- data.frame(gse@phenoData@data)
```

```
expr(gse[1])
```

```
## gse[1]
```

Annotation of probesets from a GEO dataset to gene symbols

For the later analysis, it is useful to annotate the probe sets of the GEO data set to gene symbol. To do that, first we extract the name of the probes and then, we perform the annotation using biomaRt.

```
id_to_annotate <- rownames(gse@assayData[["exprs"]])
# extract id to annotate

# annotation
#mart <- useMart("ENSEMBL_MART_ENSEMBL")
#mart <- useDataset("hsapiens_gene_ensembl", mart)
#annotLookup <- getBM(
#  mart = mart,
#  attributes = c(
#    "affy_hg_u133_plus_2",
#    "ensembl_gene_id",
#    "gene_biotype",
#    "external_gene_name",
#    "description"),
#  filter = "affy_hg_u133_plus_2",
#  values = id_to_annotate,
#  uniqueRows=TRUE)
```

```
#saveRDS(annotLookup, file = "output/annotLookup.rds")

# load the local copy to use when biomaRt is unavailable
annotLookup <- readRDS("output/annotLookup.rds")
```

```
head(annotLookup)
```

```
##   affy_hg_u133_plus_2 ensembl_gene_id   gene_biotype external_gene_name
## 1           203957_at ENSG00000278573     pseudogene
## 2         201104_x_at ENSG00000271254 protein_coding
## 3           203957_at ENSG00000278066     pseudogene
## 4           204673_at ENSG00000293532 protein_coding               MUC2
## 5           202801_at ENSG00000288516 protein_coding             PRKACA
## 6           203488_at ENSG00000288324 protein_coding             ADGRL1
```

```
##                                                                                                 description
## 1
## 2
## 3
## 4                                  mucin 2, oligomeric mucus/gel-forming [Source:HGNC Symbol;Acc:HGNC:7512]
## 5 protein kinase cAMP-activated catalytic subunit alpha [Source:HGNC Symbol;Acc:HGNC:9380]
## 6                adhesion G protein-coupled receptor L1 [Source:HGNC Symbol;Acc:HGNC:20973]
```

```
indicesLookup <- match(rownames(gse@assayData[["exprs"]]), annotLookup$affy_hg_u133_plus_2)
#head(annotLookup[indicesLookup, "external_gene_name"])
```

```
dftmp <- data.frame(rownames(gse@assayData[["exprs"]]),
                    annotLookup[indicesLookup,
                                c("affy_hg_u133_plus_2", "external_gene_name")])
head(dftmp, 20)
```

```
##      rownames.gse.assayData...exprs.... affy_hg_u133_plus_2 external_gene_name
## 2697                          1007_s_at           1007_s_at               DDR1
## 2072                            1053_at             1053_at               RFC2
## 2748                             117_at              117_at              HSPA6
## 5260                             121_at              121_at               PAX8
## 3793                          1255_g_at           1255_g_at    GUCA1ANB-GUCA1A
## 4330                            1294_at             1294_at               UBA7
## 5181                            1316_at             1316_at               THRA
## 295                             1320_at             1320_at             PTPN21
## 1843                          1405_i_at           1405_i_at               CCL5
## 1896                            1431_at             1431_at             CYP2E1
## 5223                            1438_at             1438_at              EPHB3
## 357                             1487_at             1487_at            ESRRAP1
## 742                           1494_f_at           1494_f_at             CYP2A6
## 107                           1598_g_at           1598_g_at               GAS6
## 1753                        160020_at           160020_at               MMP14
## 1918                            1729_at             1729_at              TRADD
## 2040                            1773_at             1773_at        CHURC1-FNTB
## 5218                             177_at              177_at               PLD1
## 1405                             179_at              179_at            UPK3BP1
## 2176                            1861_at             1861_at                BAD
```

```
length(rownames(gse@assayData[["exprs"]]))
```

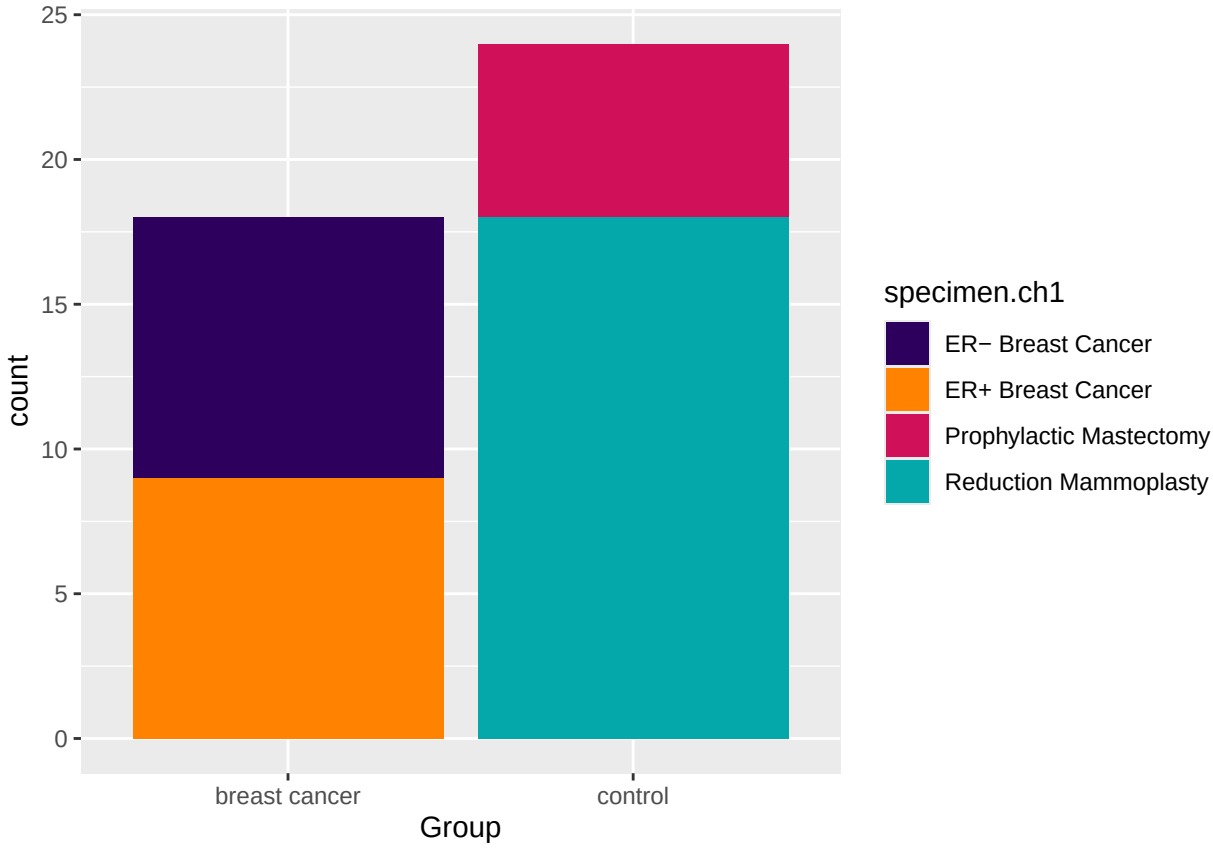
```
## [1] 22283
```

```
table(dftmp[,1] == dftmp[,2])
```

```
##
##  TRUE
## 20259
```

Explore the groups in the dataset

```
p <- ggplot(metadata, aes(x=disease.state.ch1, fill=specimen.ch1))+
  geom_bar()+
  scale_fill_manual(values = my_colors[c(1,4,6,7)])
p + labs(x = "Group")
```



Exploratory analysis

```
# show what we have:
show(gse)
```

```
## ExpressionSet (storageMode: lockedEnvironment)
## assayData: 22283 features, 42 samples
##   element names: exprs
## protocolData: none
## phenoData
##   sampleNames: GSM512539 GSM512540 ... GSM512580 (42 total)
##   varLabels: title geo_accession ... tissue:ch1 (38 total)
##   varMetadata: labelDescription
## featureData: none
## experimentData: use 'experimentData(object)'
##   pubMedIds: 20197764
## Annotation: GPL96
```

The actual expression data are accessible in the `exprs` section of `gse`, an Expression Set and the generic data class that BioConductor uses for expression data.

```
head(exprs(gse))
```

```
##           GSM512539 GSM512540 GSM512541 GSM512542 GSM512543 GSM512544 GSM512545 GSM512546 GSM512547 G
## 1007_s_at    2461.4    3435.7    1932.5    2377.7    3055.3    2978.1    2348.5    2963.9    2776.9
## 1053_at        26.7     159.0      31.2     140.7      69.9      98.5      37.0      59.9      86.7
## 117_at         82.6     243.4     150.2      95.1     209.3     103.4      91.2     168.4     162.7
```

```
## 121_at        942.3     897.5     840.8     870.9     685.4     791.8     886.5     954.2     843.1
## 1255_g_at      71.8      87.9      75.4      58.1      31.8      40.3      70.5      43.3      51.6
## 1294_at       630.2     571.4     346.3     679.9    1289.3     421.1     417.6     811.6     778.1
##           GSM512550 GSM512551 GSM512552 GSM512553 GSM512554 GSM512555 GSM512556 GSM512557 GSM512558 G
## 1007_s_at    3037.1    3545.8    3322.6    1963.7    3609.6    2078.9    4138.6    4260.7    2453.6
## 1053_at        82.9      97.7      69.7      82.0      45.6      84.5      31.7      37.4      82.4
## 117_at        113.5      80.0     186.4     106.6     145.6     144.4     133.6     278.6     173.0
## 121_at        912.2     911.6     862.4     705.0     984.6     853.8     846.8    1273.0     833.6
## 1255_g_at      53.7      30.5      15.2      42.5      76.6      88.2      90.6      65.8      25.8
## 1294_at       987.7     938.5     924.6     480.8    1054.1     632.0     448.0    1345.2    1248.9
##           GSM512561 GSM512562 GSM512563 GSM512564 GSM512565 GSM512566 GSM512567 GSM512568 GSM512569 G
## 1007_s_at    4340.1    3155.3    2390.3    2738.8    3233.1    2836.6    2915.4    3457.5    2798.7
## 1053_at        76.7     100.3     115.4      14.1      47.6      77.1      47.1      47.0      83.2
## 117_at        168.0      95.2      73.6     122.7     107.6     120.9     143.4      92.5      72.1
## 121_at        827.0     629.4     709.2     305.6     877.4     425.7     643.8     771.3     681.1
## 1255_g_at      87.9      44.6      59.3      12.0      82.1      59.2      62.2      28.3      97.6
## 1294_at      2218.1    1321.1     606.7    1709.9     980.8    1268.4     955.8    1157.5     888.6
##           GSM512572 GSM512573 GSM512574 GSM512575 GSM512576 GSM512577 GSM512578 GSM512579 GSM512580
## 1007_s_at    3669.5    3310.1    3942.2    4520.4    3596.1    2989.0    3164.5    2764.3    4258.5
## 1053_at        24.1       8.8      44.6      54.7      56.7      89.9      63.4      57.0      59.5
## 117_at        165.8     141.6      97.1     132.7     124.3     210.5     131.4      89.6     123.3
## 121_at        746.9    1090.3    1008.7     718.6     988.4     295.9     957.3     630.8     869.2
## 1255_g_at      53.0      39.9      11.0      50.2      60.0      34.3      33.5      61.7      50.4
## 1294_at      1138.5     483.0    1326.5    1179.4     668.3     863.2    1055.5    1287.6    1127.8
```

To conveniently access the data rows and columns present in `exprs(gse)`, this matrix is assigned to its own variable `ex`.

```
# exprs (gse) is a matrix that we can assign to its own variable, to
# conveniently access the data rows and columns
ex <- exprs(gse)
dim(ex) # 42 sample, 22283 genes
```

```
## [1] 22283    42
```

The dataset contains gene expression data of 22283 genes (rows) from 42 patients (columns).

Pre-processing

```
# Analyze value distributions
boxplot(ex, main = 'Boxplot of the data before normalization',
        xlab = "Samples",
        ylab = "Expression Value",
        varwidth = TRUE
        )
```

The boxplot shows that scaling is necessary. So, in this case, I try to apply a log transformation to the data.

```
ex2<-log(ex)
ex2 <- na.omit(as.matrix(ex2))
#dim(ex2) # 22283    42 same as before
boxplot(ex2, main = 'Boxplot of the data after applying a logarithmic transformation',
        xlab = "Samples",
        ylab = "Expression Value"
        )
```

From the boxplot after the log transformation, I can see that there is some variation in the median of the

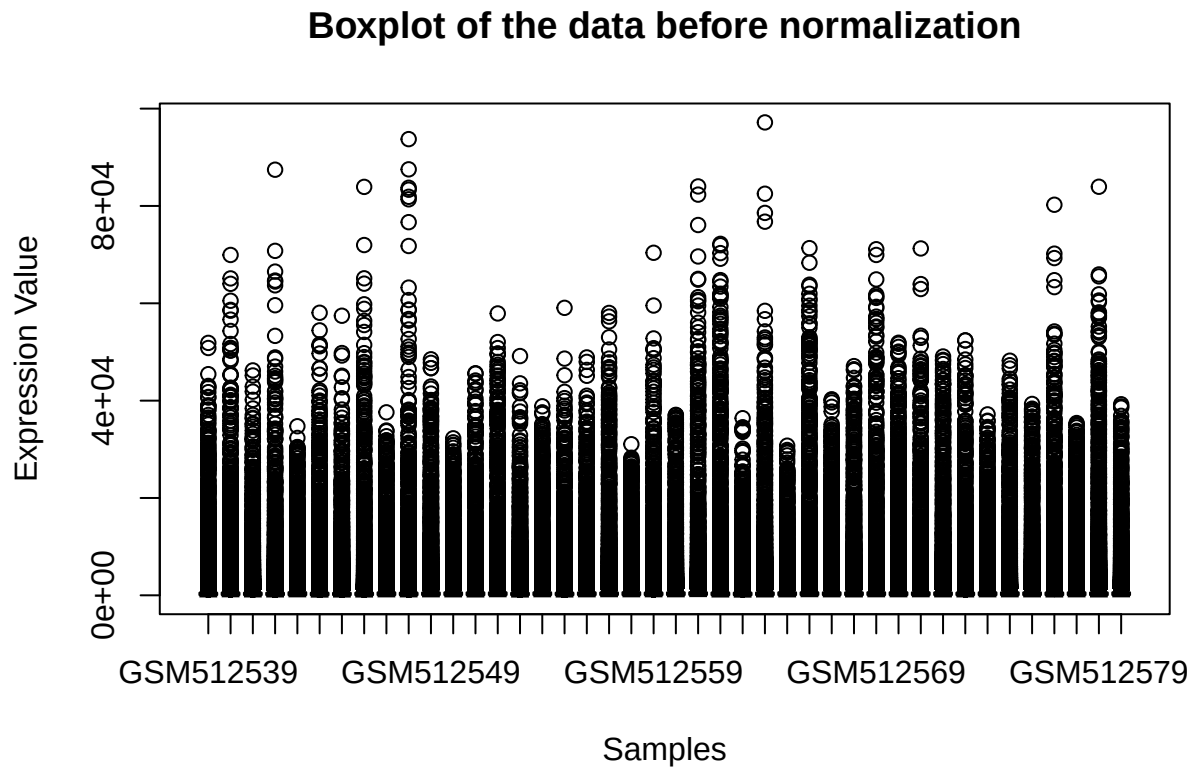


Figure 1: Boxplot of the data before normalization

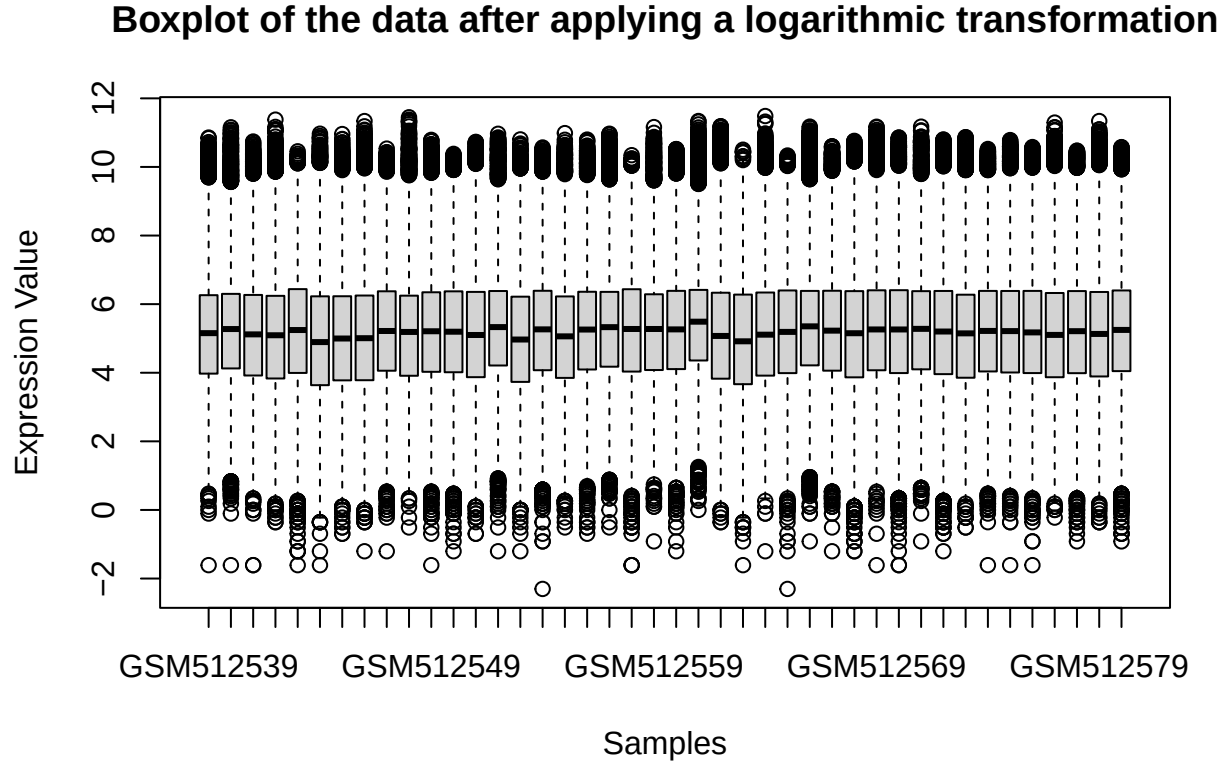


Figure 2: Boxplot of the data after applying a logarithmic transformation

samples. So, one of the simplest normalization strategies is to align the log values so that all channels have the same median. A convenient choice is zero so that positive or negative values reflect signals above or below the median for a particular channel.

```
normalized.log.ex=scale(log(ex))

# boxplot post median normalization
boxplot(normalized.log.ex,
        main = 'Boxplot of the data after median normalization',
        xlab = "Samples",
        ylab = "Expression Value")
```

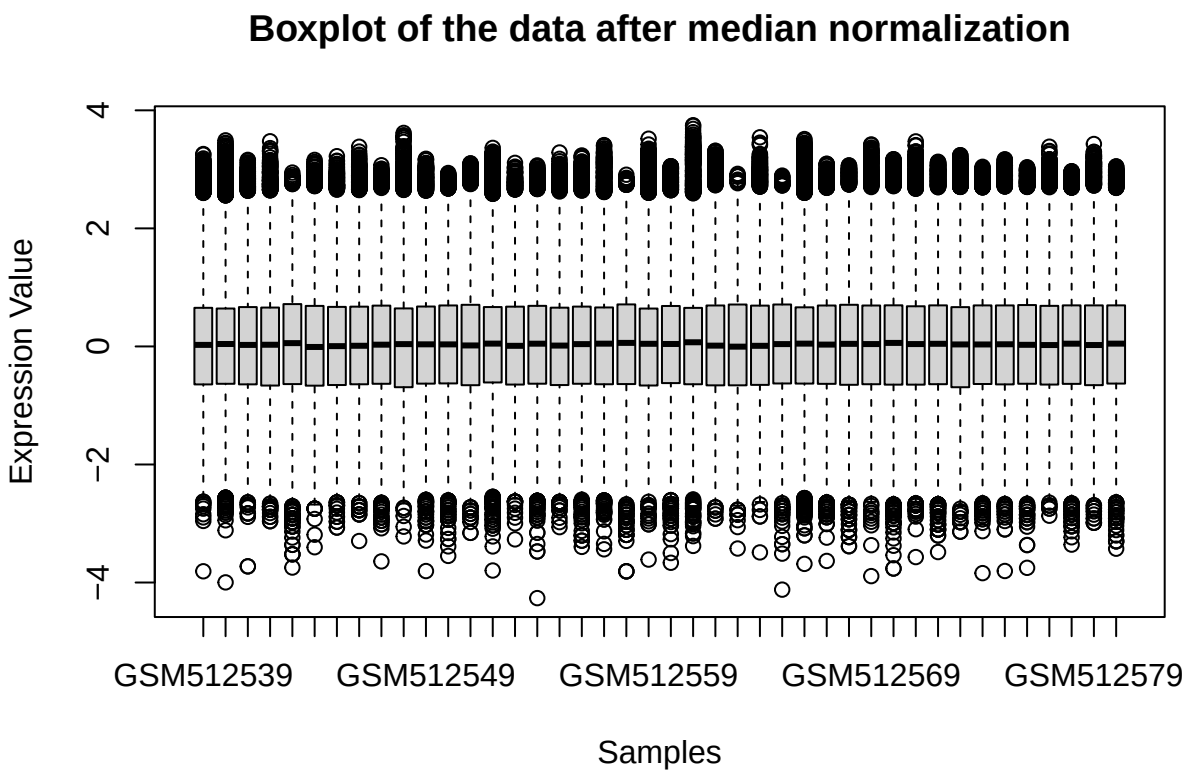


Figure 3: Boxplot of the data after normalization

PCA

PCA is a dimensionality reduction technique that allows to condense thousands of dimensions into just two or three.For the dataset's samples, the PCA scores display the coordinates in relation to these additional dimensions.

```
pca <- prcomp(t(normalized.log.ex))

summary(pca)
```

```
## Importance of components:
##                            PC1      PC2      PC3      PC4      PC5      PC6      PC7      PC8      PC9
## Standard deviation     21.6856 12.78817 11.18542 10.68629 10.28318 9.05796 8.85994 8.67632 8.31063 7
## Proportion of Variance  0.1798  0.06253  0.04783  0.04366  0.04043 0.03137 0.03001 0.02878 0.02641 0
## Cumulative Proportion   0.1798  0.24232  0.29016  0.33382  0.37425 0.40561 0.43563 0.46441 0.49081 0
##                           PC12    PC13    PC14    PC15    PC16    PC17    PC18    PC19    PC20    PC2
```

```
## Standard deviation     7.64884 7.42525 7.35447 7.25191 7.12698 7.03899 6.96990 6.86700 6.81828 6.726
## Proportion of Variance 0.02237 0.02108 0.02068 0.02011 0.01942 0.01894 0.01857 0.01803 0.01777 0.017
## Cumulative Proportion  0.56081 0.58189 0.60257 0.62267 0.64209 0.66104 0.67961 0.69764 0.71541 0.732
##                           PC24    PC25    PC26    PC27    PC28    PC29    PC30    PC31    PC32    PC
## Standard deviation     6.51084 6.39868 6.39565 6.35069 6.16713 6.14788 6.07658 6.01840 5.86441 5.806
## Proportion of Variance 0.01621 0.01565 0.01564 0.01542 0.01454 0.01445 0.01412 0.01385 0.01315 0.012
## Cumulative Proportion  0.78278 0.79843 0.81407 0.82949 0.84403 0.85848 0.87260 0.88645 0.89960 0.912
##                           PC36    PC37    PC38    PC39    PC40    PC41      PC42
## Standard deviation     5.46227 5.35021 5.28516 5.20027 5.10709 5.01262 4.456e-14
## Proportion of Variance 0.01141 0.01094 0.01068 0.01034 0.00997 0.00961 0.000e+00
## Cumulative Proportion  0.94846 0.95940 0.97008 0.98042 0.99039 1.00000 1.000e+00
```

```
#screeplot(pca)
```

To get the summary of the PCA and the plot showing the variance explained by the first 10 components, it is possible to use the functions commented in the chunks above.

However, using `ggplot2` and `factoextra` packages is possible to get a more concise and informative plot reporting the same information.

```
pcaVar <- get_eig(pca)
pcaVar <- pcaVar$variance.percent[1:10]
screeDf <- data.frame("Dimensions" = as.factor(seq(1,10)),
                      "Percentages" = pcaVar,
                      "Labels" = paste(round(pcaVar, 2), "%"))

p <- ggplot(data = screeDf, aes(x=Dimensions, y=Percentages))+
  geom_bar(stat = "identity", fill = "#d1105a")+
  geom_text(aes(label=Labels), vjust=-0.5, color="black", size=3.6)+
  ggtitle("Scree Plot")+
  ylab("Percentage of variance explained")+
  scale_x_discrete(labels = as.factor(seq(1,10)))
p
```

The scree plot shows that the first dimensions on the left are the more important because the percentage of variance explained by them is higher. The remaining principal components account for a very small proportion of the variability and are probably unimportant.

Let's try to plot the PCA, looking if we can see a separation between Control and Breast Cancer groups.

```
# draw PCA plot control VS breast cancer
group <- c(rep("cadetblue1",18), rep("red",18), rep("cadetblue1",6) )
plot(pca$x[,1], pca$x[,2], xlab="PCA1", ylab="PCA2", main="PCA for components 1 and 2", type="p", pch=1
text(pca$x[,1], pca$x[,2], rownames(pca$data), cex=0.75)
legend("topleft", col=c("cadetblue1","red"), legend = c("Control", "Breast Cancer"),
    pch = 20, bty='n', cex=.55)
```

Let's try to add the control subtypes. The vector group used in the PCA plot is based on the data. The samples corresponding to the colors are the following:

- **Light blue**: reduction mammoplasty (RM) breast epithelium samples
- **Red**: histologically normal (HN) epithelial samples from breast cancer patient
- **Purple**: histologically normal breast epithelium (NlEpi) from prophylactic mastectomy patient samples

```
# draw PCA plot
group <- c(rep("cadetblue1",18), rep("red",18), rep("purple",6) ) # vector of colors based on the order
```

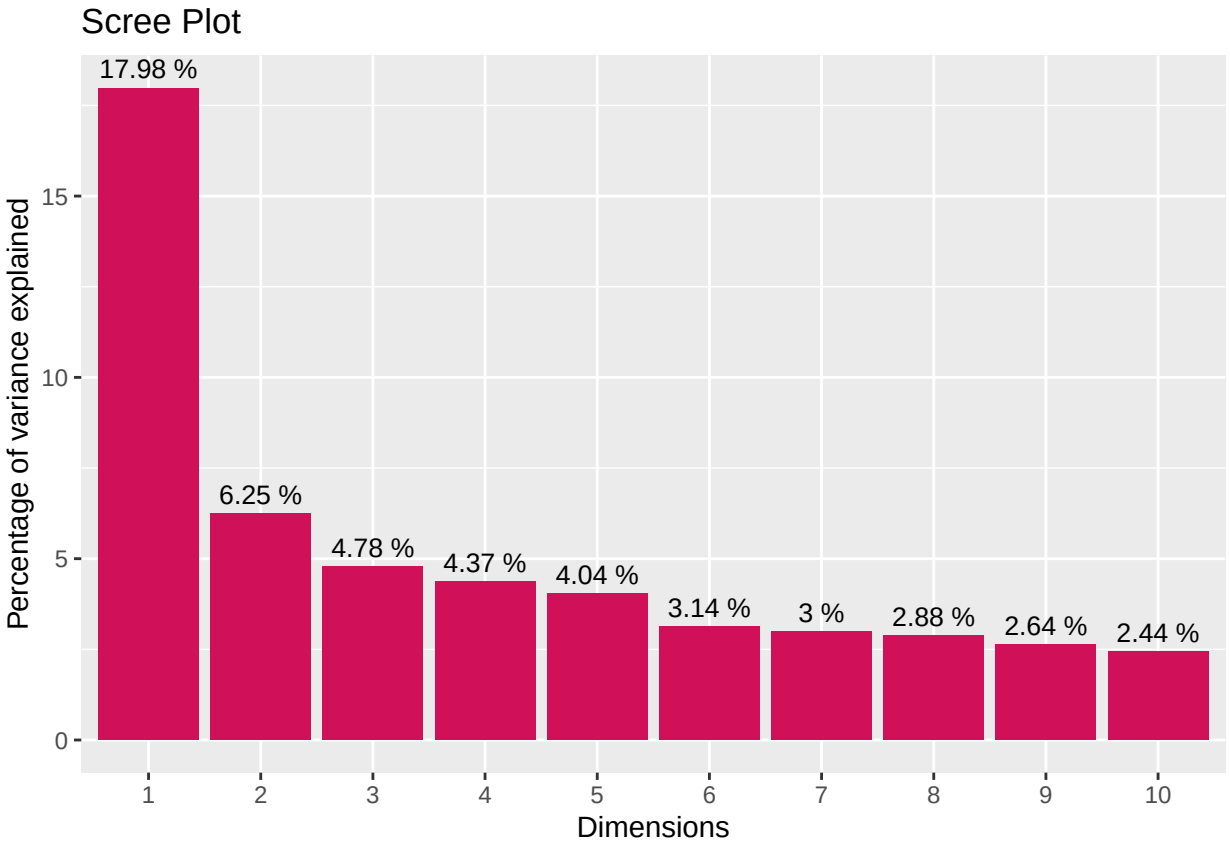


Figure 4: Scree Plot

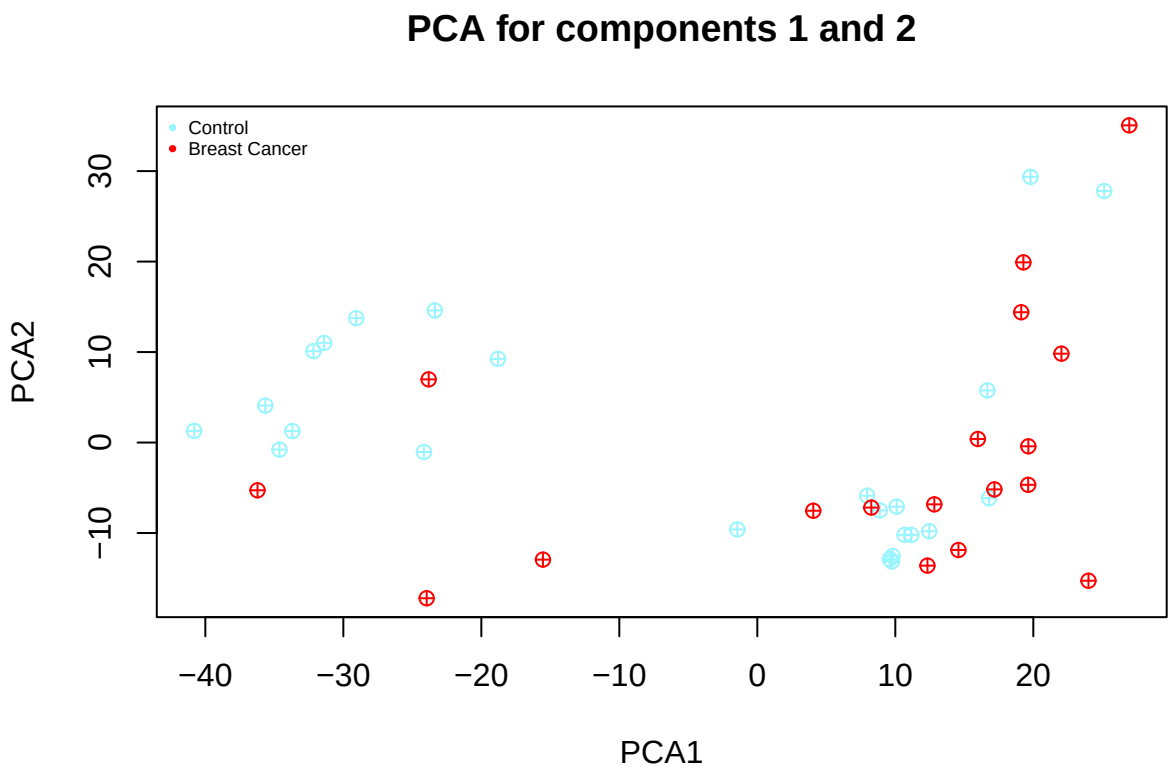


Figure 5: PCA analysis of 2 components showing breast cancer and control samples.

```
plot(pca$x[,1], pca$x[,2], xlab="PCA1", ylab="PCA2", main="PCA for components 1 and 2", type="p", pch=1
text(pca$x[,1], pca$x[,2], rownames(pca$data), cex=0.75)
legend("topleft", col=c("cadetblue1","red","purple"), legend = c("Reduction Mammoplasty", "Breast Cance
    pch = 20, bty='n', cex=.55)
```

Then, I try to see if there is a separation also inside different types of Breast Cancer.

```
# draw PCA plot with all subtypes
group <- c(rep(my_colors[7],18), rep(my_colors[4],9), rep(my_colors[1],9), rep(my_colors[6],6) ) # vect
plot(pca$x[,1], pca$x[,2], xlab="PCA1", ylab="PCA2", main="PCA for components 1 and 2", type="p", pch=1
text(pca$x[,1], pca$x[,2], rownames(pca$data), cex=0.75)
legend("topleft", col=c(my_colors[7],my_colors[4],my_colors[1],my_colors[6]), legend = c("Reduction Mam
    pch = 20, bty='n', cex=.55)
```

Interactive PCA plot

Let's try to explore an interactive PCA plot.

```
components<-pca[["x"]]
components<-data.frame(components)
type<-c(rep("RM", 18), rep("HN",18), rep("NlEpi",6))
components<-cbind(components, type )

fig <- plot_ly(components, x=~PC1, y=~PC2,
               color=type,colors=c('cadetblue1', 'red','purple'),
               type='scatter',mode='markers')
fig
```

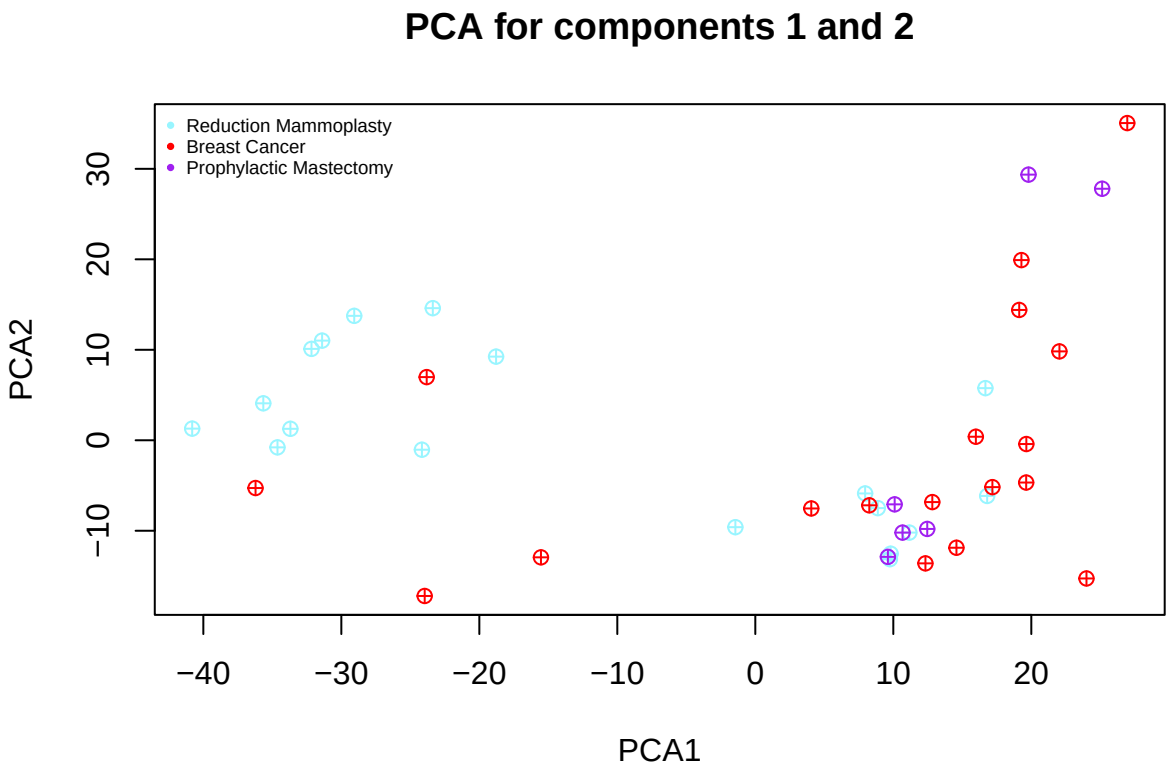


Figure 6: PCA analysis of 2 components showing breast cancer and the two subtypes of control samples.

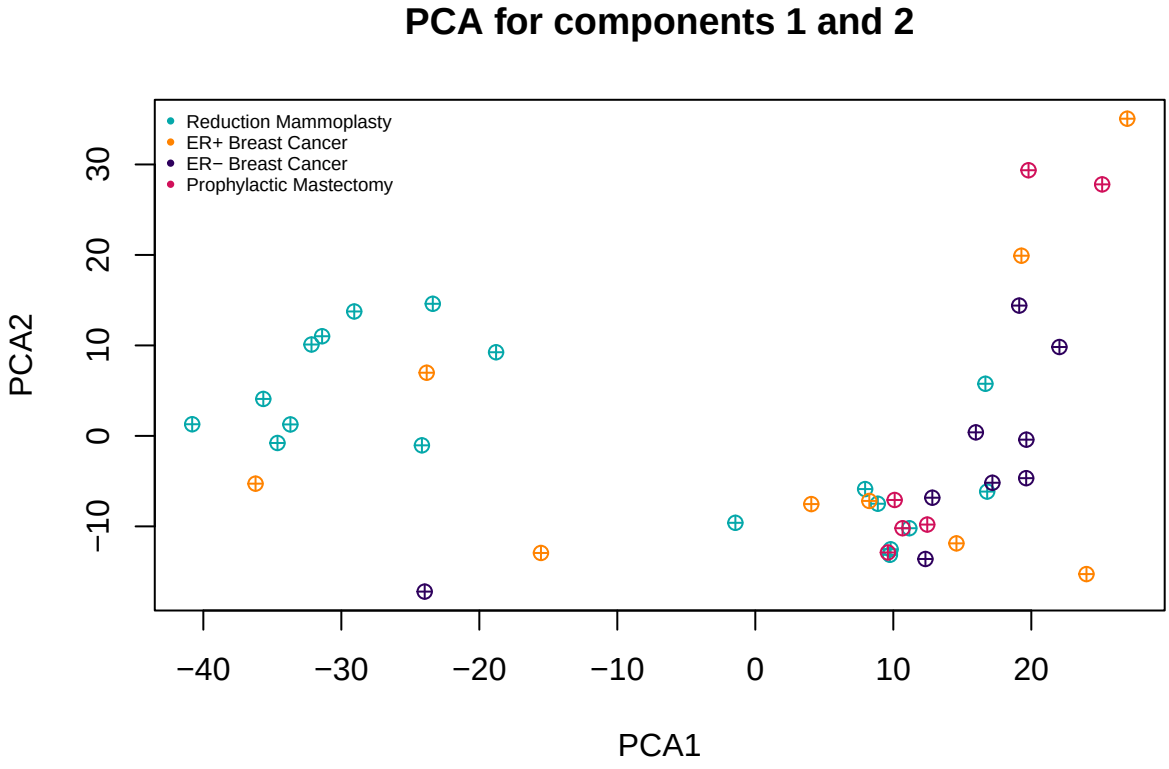


Figure 7: PCA analysis of 2 components, showing all samples specimen.

```
fig2 <- plot_ly(components, x=~PC1, y=~PC2, z=~PC3,
                color=type, colors=c('cadetblue1', 'red','purple'),
                mode='markers', marker = list(size = 4))
fig2
```

```
fig3 <- plot_ly(components, x=~PC1, y=~PC3,
                color=type, colors=c('cadetblue1', 'red','purple'),
                type='scatter',mode='markers')
fig3
```

UMAP

```
set.seed(123)
umap_result <- umap(t(normalized.log.ex))

umap_df <- as.data.frame(umap_result$layout)
colnames(umap_df) <- c("UMAP1", "UMAP2")

# Add sample types to the UMAP data frame
umap_df$type <- c(rep("RM", 18), rep("HN",18), rep("NlEpi",6))

ggplot(umap_df, aes(x = UMAP1, y = UMAP2, color = type)) +
  geom_point(size = 3, alpha = 0.8) +
  scale_color_manual(values = c("RM" = "cadetblue1",
                                "HN" = "red",
                                "NlEpi" = "purple")) +
  theme_minimal() +
  labs(title = "2D UMAP Projection of GSE20437", x = "UMAP1", y = "UMAP2")
```

Clustering

K-means

K-means is an unsupervised method that relies on the tuning parameter k (number of resulting clusters). This method is sensitive to outliers, so it is used for a first exploratory analysis.

k=2

Let's try to compute K-means by setting $k = 2$, since we know that there are 2 groups: breast cancer and control.

```
set.seed(1)
k <- 2 # number of clusters

kmeans_result2 <- kmeans(t(normalized.log.ex),k)
table(kmeans_result2$cluster) # tells how many samples were assigned to each cluster
```

```
##
##  1  2
## 14 28
```

The results are plotted with a PCA for a better visualization.

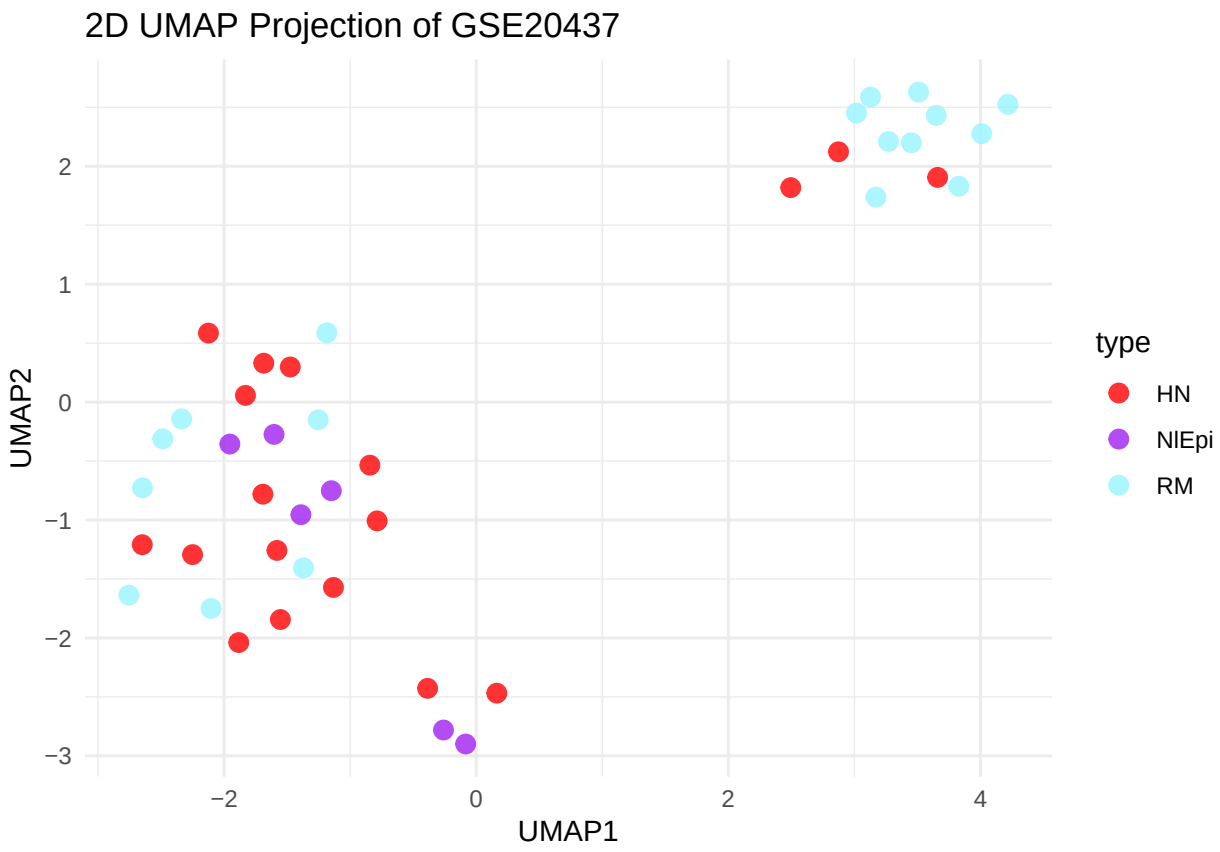


Figure 8: 2D UMAP

```r
# Create a data frame for plotting
cluster_df_k2 <- data.frame(
  PC1 = pca$x[,1],
  PC2 = pca$x[,2],
  Cluster = as.factor(kmeans_result2$cluster), # Cluster from K-means
  Disease = metadata$disease.state.ch1 # Disease state (Breast cancer or not)
)

enhanced_colors <- c("#1f77b4", "#ff7f0e")

ggplot(cluster_df_k2, aes(x = PC1, y = PC2, color = Cluster, shape = Disease)) +
  geom_point(size = 3, alpha = 0.8) +
  scale_color_manual(values = enhanced_colors) +  # Color code for clusters
  scale_shape_manual(values = c(16, 17)) +  # Shape code for disease state
  labs(title = "K-means Clustering of PCA Components",
       x = "PCA1", y = "PCA2",
       color = "Cluster", shape = "Disease State") +
  theme_minimal() +
  theme(legend.position = "top")
```

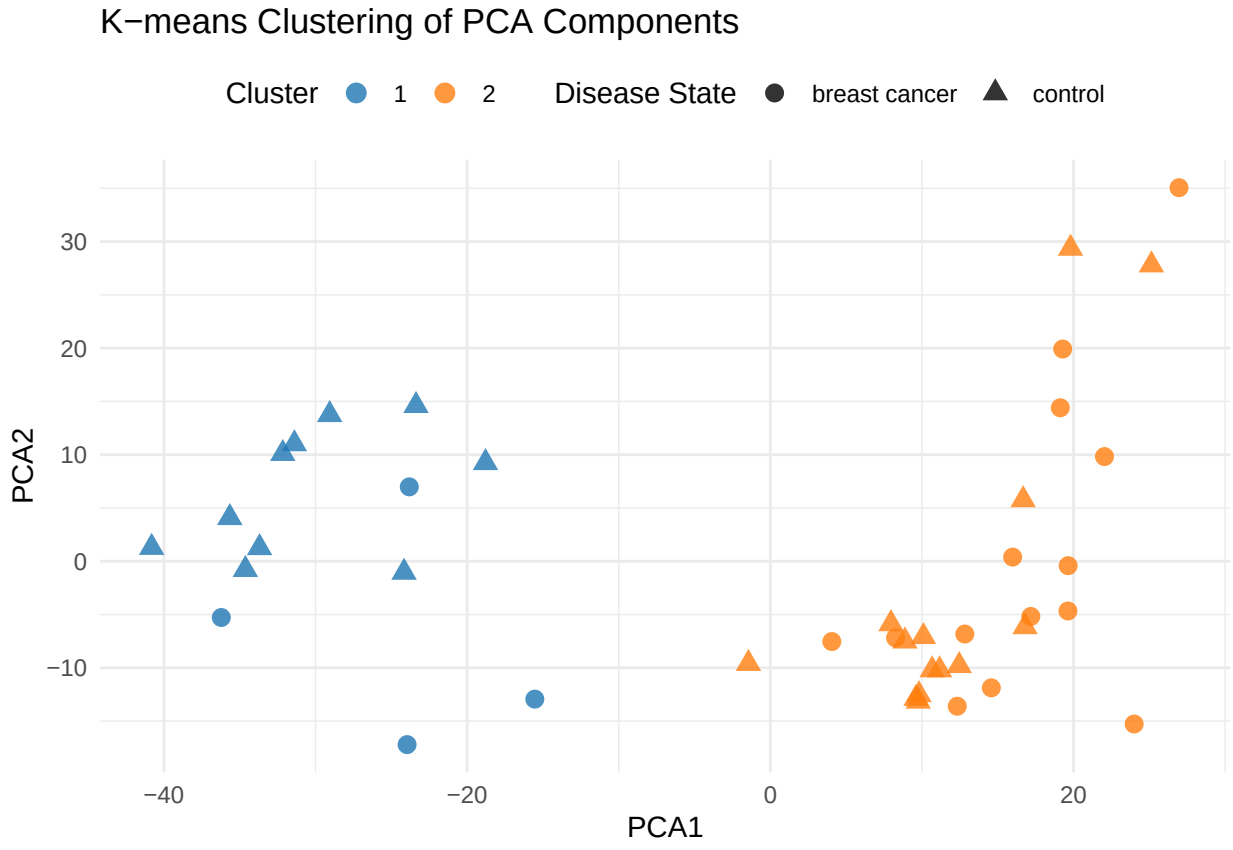


Figure 9: K-means clustering of PCA components 1 and 2 breast cancer and control samples.

```r
## Create list with the predictions
pred_k <- kmeans_result2$cluster
pred_k[which(pred_k==1)] <- "control"
pred_k[which(pred_k==2)] <- "breast cancer"
```

```r
pred_k <- factor(pred_k,levels = c("control", "breast cancer"))

true_labels <- ifelse(metadata$disease.state.ch1 == "breast cancer", 1, 0)
pred_labels <- ifelse(pred_k == "breast cancer", 1, 0)

table(true_labels)
```

```
## true_labels
##  0  1
## 24 18
```

```r
table(pred_labels)
```

```
## pred_labels
##  0  1
## 14 28
```

```r
# Compute ROC
roc_k <- roc(response = true_labels, predictor = pred_labels)
auc_k <- auc(roc_k)

## Accuracy
acc_k <- mean(pred_k==metadata$disease.state.ch1)

## Crete dataframe to store the accuracy results
res_df <- data.frame(Kmeans = c(acc_k, auc_k))
rownames(res_df) <- c("Accuracy", "AUC")
```

```r
plot(roc_k, main = paste("ROC Curve (AUC =", round(auc_k, 3), ")"), col = "#d62728", lwd = 2)
abline(a = 0, b = 1, lty = 2, col = "gray")
```

k=4

Let's try to increase the number of clusters to 4, as we know that there are 4 subtypes. The *control* group contains both Reduction mammoplasty (RM) breast epithelium samples and histologically normal breast epithelium (NlEpi) from prophylactic mastectomy patient samples; while the *breast cancer* group contains histologically normal (HN) epithelial samples from breast cancer patient both with ER+ and ER-.

```r
# K-means clustering with k = 4
set.seed(1)
k <- 4  # Number of clusters
kmeans_result_k4 <- kmeans(t(normalized.log.ex), k)
table(kmeans_result_k4$cluster)
```

```
##
##  1  2  3  4
##  4  7 10 21
```

```r
# Create a data frame for plotting
cluster_df_k4 <- data.frame(
  PC1 = pca$x[,1],
  PC2 = pca$x[,2],
  Cluster = as.factor(kmeans_result_k4$cluster), # Cluster from K-means
  Disease = metadata$disease.state.ch1, # Disease state (Breast cancer or not)
  Specimen = metadata$specimen.ch1  # Specimen type (e.g., control, cancer)
)
```

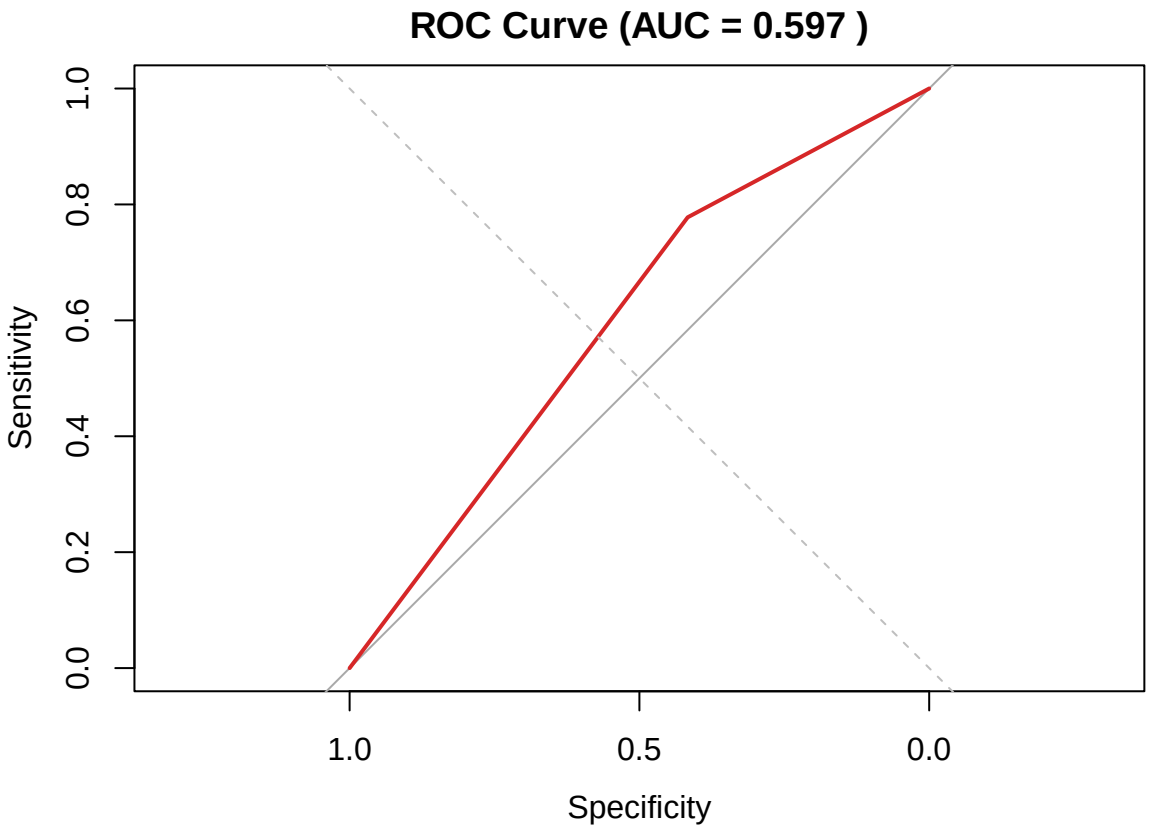


Figure 10: ROC curve K-means

```r
enhanced_colors <- c("#1f77b4", "#ff7f0e", "#2ca02c", "#d62728")
shape_values <- c(15, 16, 17, 18)  # Square, Circle, Triangle, Diamond

ggplot(cluster_df_k4, aes(x = PC1, y = PC2, color = Cluster, shape = Specimen)) +
  geom_point(size = 4, alpha = 0.8) +
  scale_color_manual(values = enhanced_colors) +
  scale_shape_manual(values = shape_values) +  # Different shapes for specimen types
  labs(title = "K-means Clustering of PCA Components (k=4)",
       x = "PCA1", y = "PCA2",
       color = "Cluster", shape = "Specimen Type") +
  theme_minimal() +
  theme(legend.position = "top")
```

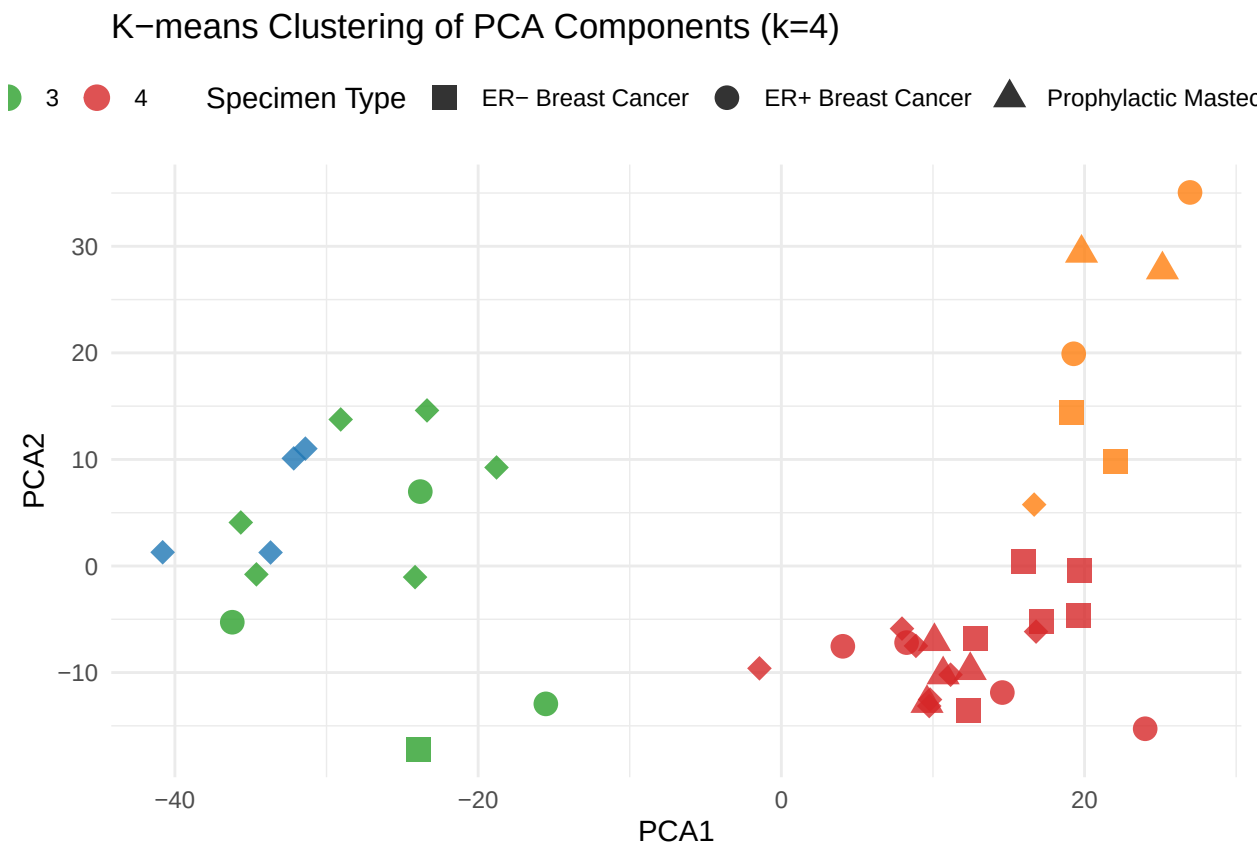


Figure 11: K-means clustering of PCA components 1 and 2 with k=4.

Hierarchical

Hierarchical clustering produces a dendrogram. At each step a distance matrix is computed and the points with the lower distance are clustered together. Then, the distance matrix is recomputed considering the new cluster as one point. In the end, the tree is cut to have the chosen number of clusters.

For this clustering method, distance between samples and between clusters must be computed. In this case, sample distances are computed with the *Euclidean* method, and the cluster distances with *average linkage*.

```r
dist_matrix <- dist(t(normalized.log.ex), method = "euclidean")
hc <- hclust(dist_matrix, method = "average")
```

```r
tree <- as.phylo(hc)

k_hc <- 2
hclusters <- cutree(hc, k = k_hc)
cluster_df <- data.frame(
  sample = names(hclusters),
  cluster = factor(hclusters)
)

annot <- data.frame(
  sample = colnames(normalized.log.ex),  # samples are columns
  disease = metadata$disease.state.ch1,
  specimen = metadata$specimen.ch1
)

tip_data <- merge(cluster_df, annot, by.x = "sample", by.y = "sample")

# Plot with samples' name
ggtree(tree, layout = "rectangular", branch.length = "none") %<+% annot +
  # Tip points colored by disease, shaped by specimen
  geom_tippoint(aes(color = disease, shape = specimen), size = 2) +
  scale_color_manual(values = c("control" = "#1b9e77", "breast cancer" = "#d95f02")) +
  scale_shape_manual(values = c(
    "Reduction Mammoplasty" = 15,
    "ER+ Breast Cancer" = 16,
    "ER- Breast Cancer" = 17,
    "Prophylactic Mastectomy" = 18
  )) +
  geom_tiplab(size = 1.7, angle = 85, hjust = 0.5, offset=-0.5)+ # name samples
  theme_tree2() +
  coord_flip() + scale_x_reverse() +
  ggtitle("Hierarchical Clustering Dendrogram") +
  theme(legend.position = "right")
```

```r
# Plot without samples' name
ggtree(tree, layout = "rectangular", branch.length = "none") %<+% annot +
  # Tip points colored by disease, shaped by specimen
  geom_tippoint(aes(color = disease, shape = specimen), size = 2) +
  scale_color_manual(values = c("control" = "#1b9e77", "breast cancer" = "#d95f02")) +
  scale_shape_manual(values = c(
    "Reduction Mammoplasty" = 15,
    "ER+ Breast Cancer" = 16,
    "ER- Breast Cancer" = 17,
    "Prophylactic Mastectomy" = 18
  )) +
  #geom_tiplab(size = 1.7, angle = 85, hjust = 0.5, offset=-0.5)+ # name samples
  theme_tree2() +
  coord_flip() + scale_x_reverse() +
  ggtitle("Hierarchical Clustering Dendrogram") +
  theme(legend.position = "right")
```

```r
# Predictions (label clusters as "control" and "breast cancer")
pred_h <- hclusters
```

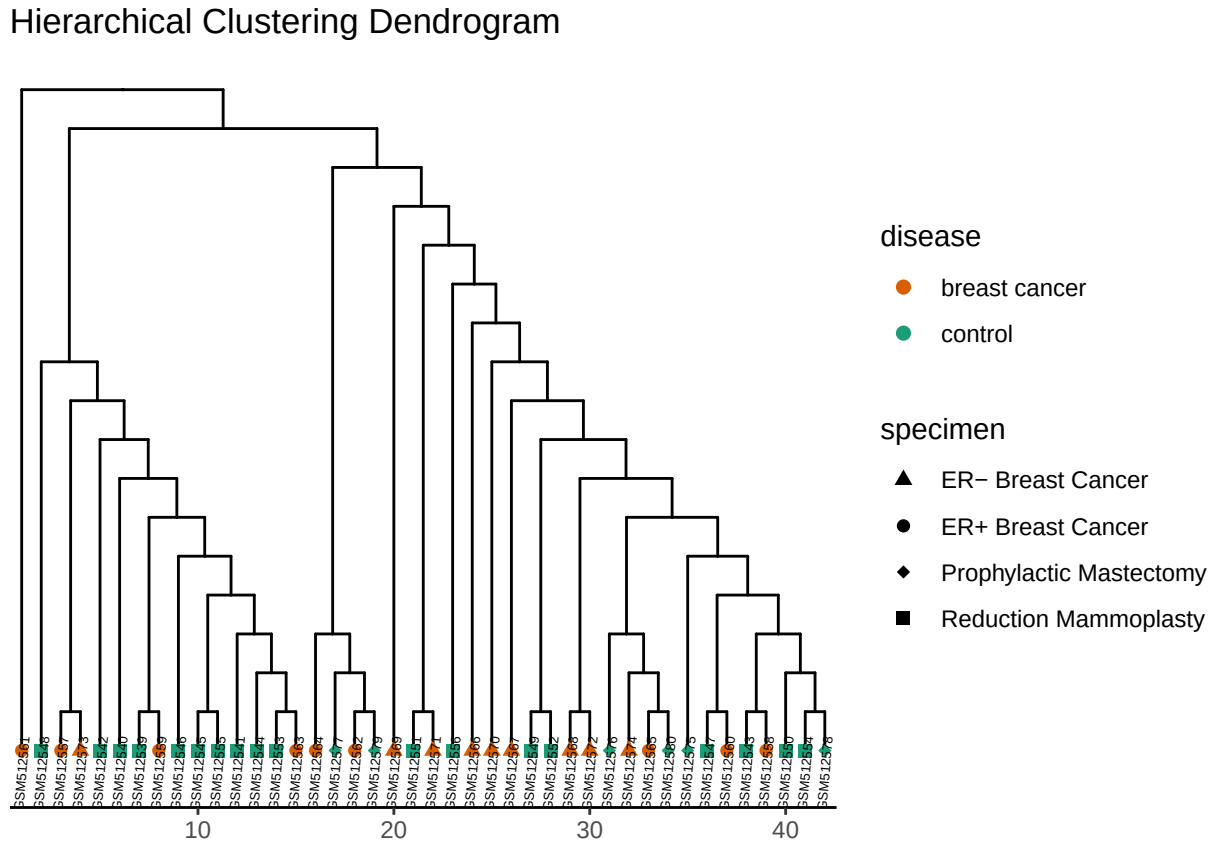


Figure 12: Hierarchical clustering dendrogram. The sample distances are computed with Euclidean distance, while the cluster distance with average linkage.

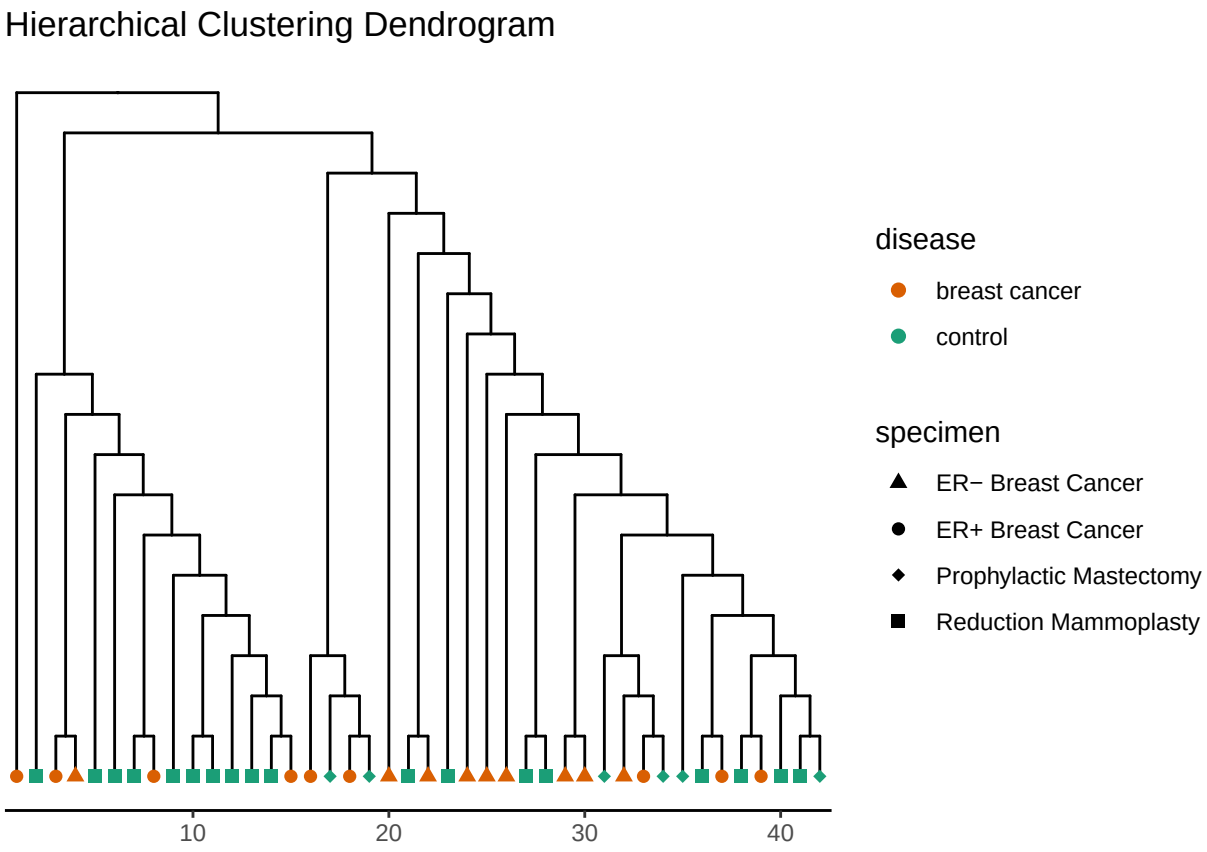


Figure 13: Hierarchical clustering dendrogram without samples' labels. The sample distances are computed with Euclidean distance, while the cluster distance with average linkage.

```
cluster1_majority <- names(which.max(table(pred_h[metadata$disease.state.ch1 == "control"])))
pred_h <- ifelse(hclusters == as.numeric(cluster1_majority), "control", "breast cancer")
pred_h <- factor(pred_h, levels = c("control", "breast cancer"))

# Accuracy
acc_h <- mean(pred_h == metadata$disease.state.ch1)

# AUC from ROC curve
true_labels_h <- ifelse(metadata$disease.state.ch1 == "breast cancer", 1, 0)
pred_labels_h <- ifelse(pred_h == "breast cancer", 1, 0)
roc_h <- roc(response = true_labels_h, predictor = pred_labels_h)
auc_h <- auc(roc_h)

# Update results table
res_df["Hierarchical"] <- c(acc_h, auc_h)
```

Random forest

Supervised methods require the data to be divided into a training and a test set.

```
set.seed(1)
# Extract expression data
expr_data <- t(normalized.log.ex)  # samples are rows
labels <- metadata$disease.state.ch1  # A factor with "control" and "breast cancer"

# 70% train, 30% test
train_index <- createDataPartition(labels, p = 0.7, list = FALSE)

# Split data
train_data <- expr_data[train_index, ]
test_data  <- expr_data[-train_index, ]

train_labels <- labels[train_index]
test_labels  <- labels[-train_index]
```

Random forest requires the tuning of 2 parameters: `ntree` and the `mtry`. The first is the number of trees to grow and the second is the number of features (genes) considered when building each tree. The best parameters are chosen comparing OOB errors of random forests fitted with different values of the two parameters.

```
set.seed(1234)
rf <- randomForest(x=train_data, y=as.factor(train_labels), ntree = 600)
plot(rf, main = "Random Forest")
```

The plot above shows how the error rate changes according to the number of trees. This allows to track how model accuracy improves (or doesn't) as more trees are added. The black line represent the overall OOB (out-of-bag) error rate, while the other colored lines report the class specific error rates. From the plot, we can see that a plateau in the OOB error is reached with a number of trees around 520.

```
control <- trainControl(method='cv',
                        search='grid')

mtry <- ncol(train_data)
tunegrid <- expand.grid(mtry = seq(1, min(50, mtry), by = 5))
```

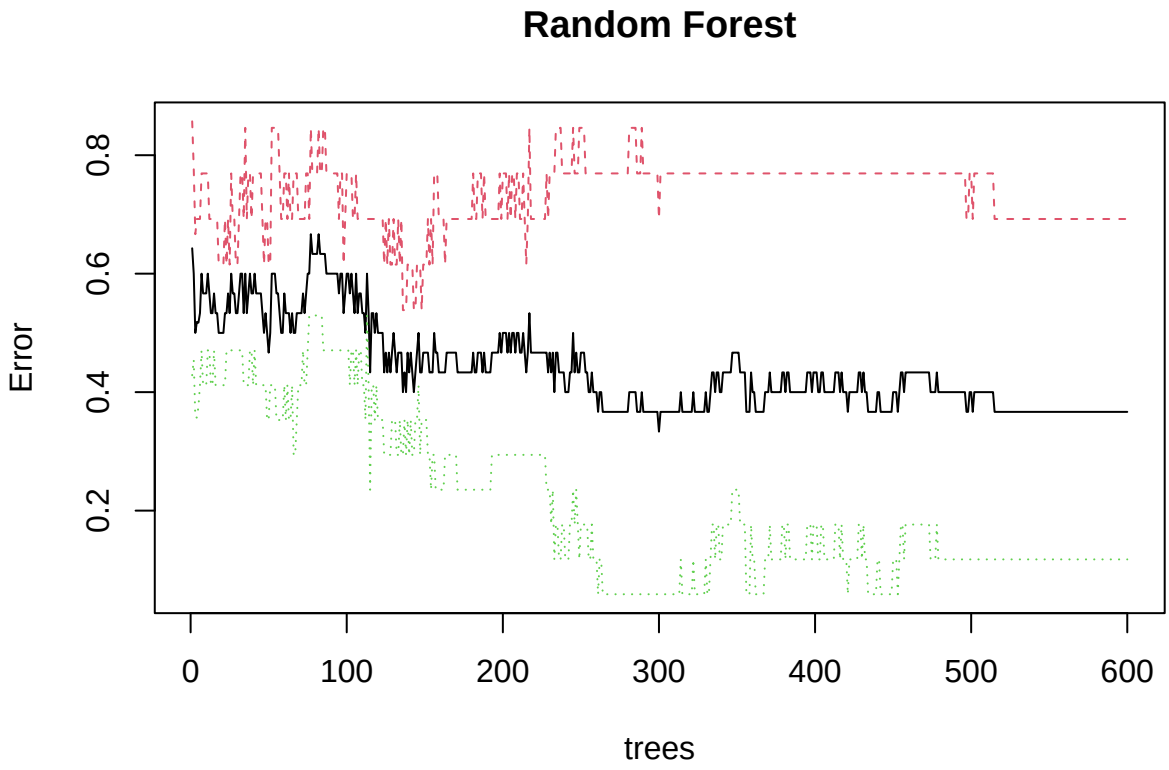


Figure 14: Plot showing how the error rate changes according to the number of trees.

```
rf_gridsearch <- train(x=train_data, y=train_labels,
                       method = 'rf',
                       ntree=520,
                       metric = 'Accuracy',
                       tuneGrid = tunegrid,
                       trControl = control)

mtry <- rf_gridsearch$bestTune$mtry
```

The best mtry results to be 41. So, a model is built on the training set with these parameters and then evaluated on the test set.

```
set.seed(1)

tunegrid <- expand.grid(.mtry=c(mtry))
rf <- train(x=train_data, y=train_labels,
            method = 'rf',
            ntree=520,
            metric = 'Accuracy',
            tuneGrid = tunegrid,
            trControl = control)
```

```
set.seed(1)

## Apply it on the test set
pred_rf <- predict(rf, test_data)
```

```r
## Accuracy
acc_rf <- mean(pred_rf==test_labels)
## Confusion matrix
table(pred_rf, test_labels)
```

```
##                test_labels
## pred_rf         breast cancer control
##   breast cancer             1       1
##   control                   4       6
```

```r
## AUC from ROC curve
prob_rf <- predict(rf, test_data, type = "prob")
roc_rf <- roc(test_labels, prob_rf$control)
# plot(roc_rf)
auc_rf <- auc(roc_rf)

## Update df
res_df["RandomForest"] <- c(acc_rf, auc_rf)
```

```r
rf <- randomForest(x=train_data, y=as.factor(train_labels), ntree = 520, mtry=mtry)
```

Feature selection

```r
# Get importance values
rf_importance <- importance(rf)
# Extract probe IDs in order of importance
ordered_probes <- rownames(rf_importance)[order(rf_importance[, 1], decreasing = TRUE)]
# Ensure correct mapping exists
probe_to_symbol <- annotLookup[, c("affy_hg_u133_plus_2", "external_gene_name")]
colnames(probe_to_symbol) <- c("probe", "symbol")
probe_symbols <- setNames(probe_to_symbol$symbol, probe_to_symbol$probe)

top_vars <- head(ordered_probes, 30)  # Show top 30 features
imp_df <- data.frame(
  probe = top_vars,
  importance = rf_importance[top_vars, 1],
  gene = ifelse(top_vars %in% names(probe_symbols), probe_symbols[top_vars], top_vars)
)

ggplot(imp_df, aes(x = reorder(gene, importance), y = importance)) +
  geom_bar(stat = "identity", fill = "#2e005d") +
  coord_flip() +
  labs(title = "Top 30 Variable Importance (Random Forest)", x = "Gene", y = "Importance") +
  theme_minimal(base_size = 12)
```

```r
expr_data_t <- t(expr_data)
top_expr <- expr_data_t[top_vars, , drop = FALSE]  # rows = probes, cols = samples

top_expr_df <- as.data.frame(top_expr)
top_expr_df$probe <- rownames(top_expr_df)

top_expr_long <- melt(top_expr_df, id.vars = "probe", variable.name = "sample", value.name = "expressio

# Add disease info
```

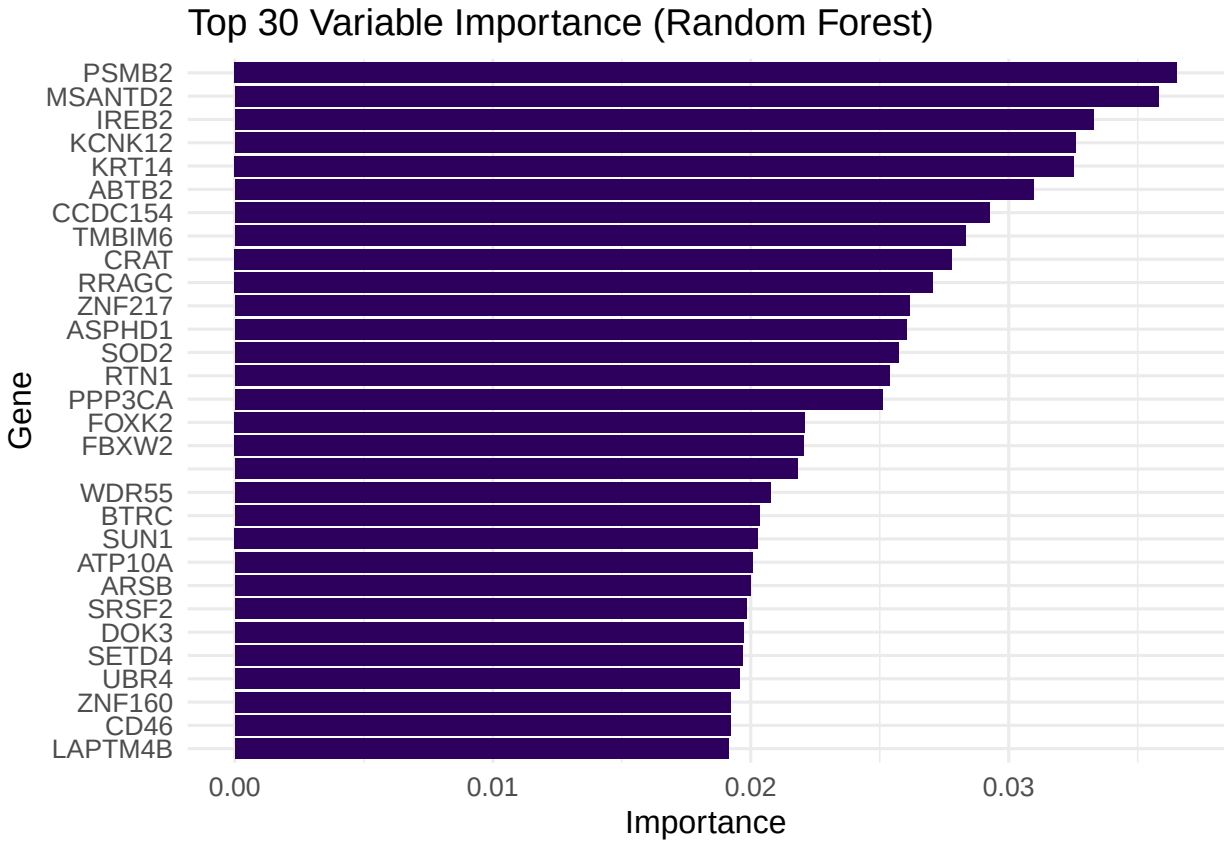


Figure 15: The plot shows the top 30 important variable according to random forest. The missing gene on the y-axis correspond to 221713_s_at. It is empty because it does not have a corresponding external gene name found in the annotation of biomaRt.

```r
top_expr_long$disease <- metadata$disease.state.ch1[match(top_expr_long$sample, metadata$geo_accession)]

# map probes to gene symbols
top_expr_long$gene <- ifelse(top_expr_long$probe %in% names(probe_symbols),
                             probe_symbols[top_expr_long$probe],
                             top_expr_long$probe)

ggplot(top_expr_long, aes(x = disease, y = expression, fill = disease)) +
  geom_boxplot(outlier.shape = NA) +
  facet_wrap(~ gene, scales = "free_y", ncol = 5) +
  theme_minimal(base_size = 12) +
  labs(title = "Expression of Top 30 Genes by Disease State",
       x = "Disease State", y = "Expression") +
  theme(axis.text.x = element_text(angle = 45, hjust = 1))
```

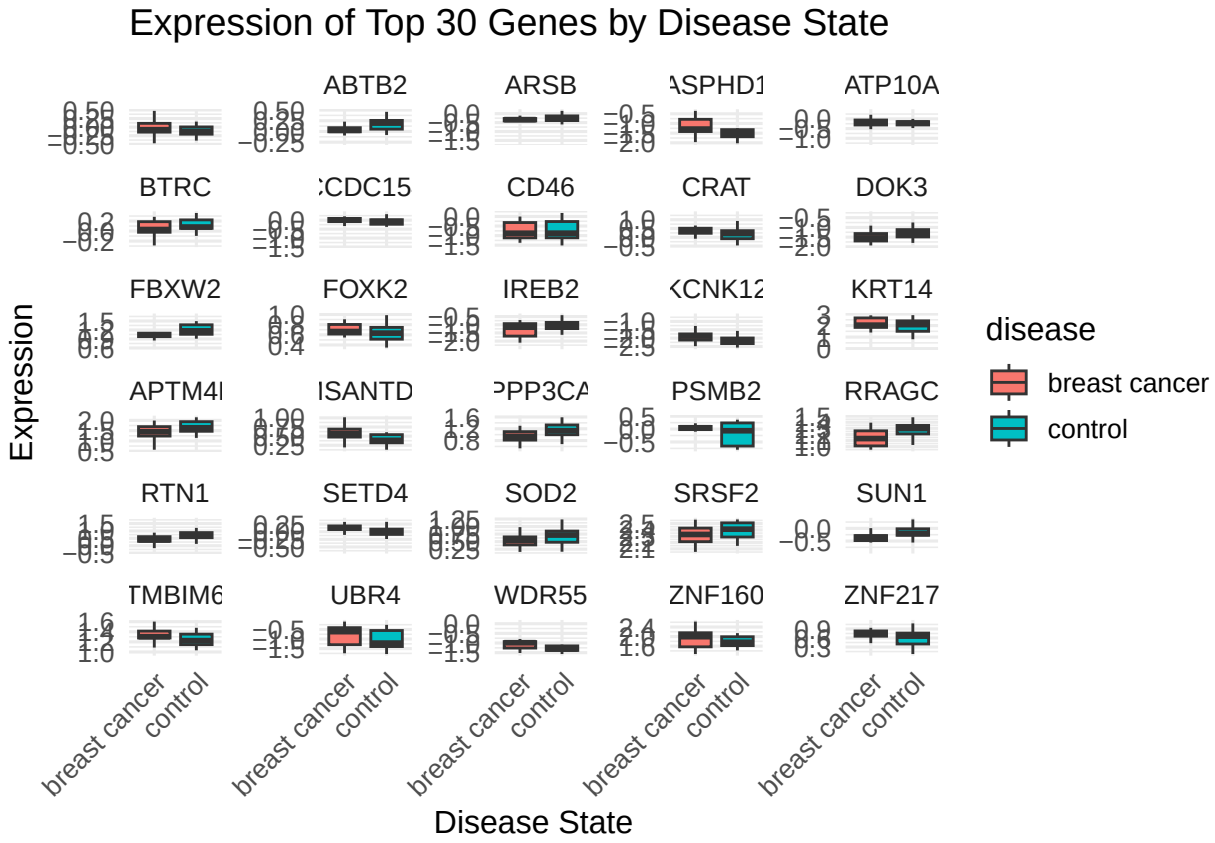


Figure 16: Plot showing how the expression of the top 30 genes changes in breast cancer and in the control.

Heatmap

```r
# Select top N important probes
top_n <- 30
top_probes <- imp_df$probe[1:top_n]

heatmap_data <- normalized.log.ex[top_probes, ]
# Add gene symbols
gene_symbols <- imp_df$gene[1:top_n]
```

```r
rownames(heatmap_data) <- gene_symbols

pheatmap(heatmap_data,
         scale = "row",  # normalize expression within each gene
         clustering_distance_rows = "euclidean",
         clustering_distance_cols = "euclidean",
         clustering_method = "complete",
         show_rownames = TRUE,
         show_colnames = TRUE,
         fontsize_row = 8,
         main = "Top Random Forest Genes Heatmap")
```

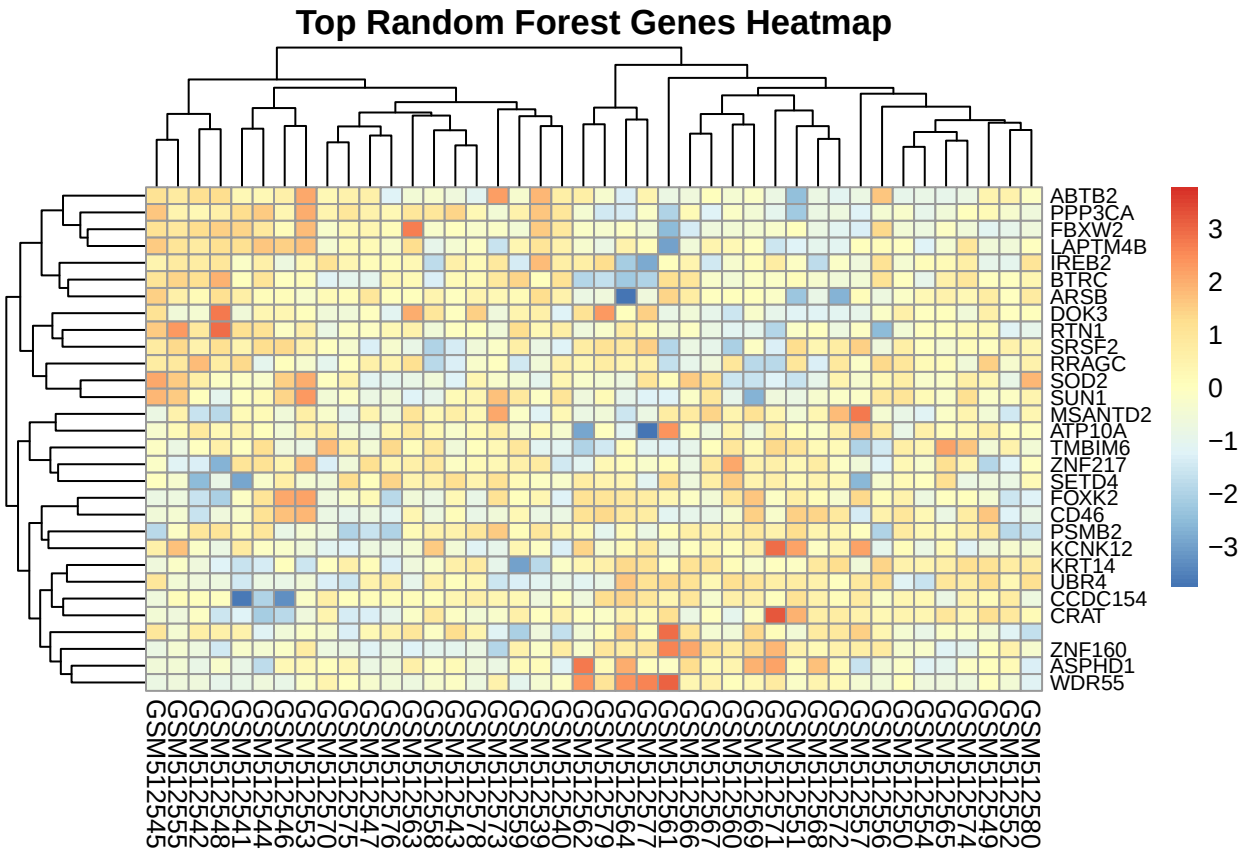


Figure 17: Heatmap of the most significant genes according to random forest.

```r
# Ensure sample labels are aligned with expression data
sample_types <- metadata$type
names(sample_types) <- colnames(normalized.log.ex)

# Subset for current samples
sample_types <- sample_types[colnames(heatmap_data)]

# Set rownames to match sample IDs
rownames(metadata) <- metadata$geo_accession
# Create annotation dataframe
annotation_col <- data.frame(Type = metadata[colnames(heatmap_data), "specimen.ch1"])
rownames(annotation_col) <- colnames(heatmap_data)
```

```r
# Convert to factor
annotation_col$Type <- factor(annotation_col$Type)
type_levels <- levels(annotation_col$Type)
palette_colors <- brewer.pal(n = length(type_levels), name = "Set2")  # or "Dark2", "Paired", etc.
names(palette_colors) <- type_levels

ann_colors <- list(Type = palette_colors)

pheatmap(heatmap_data,
         scale = "row",
         annotation_col = annotation_col,
         annotation_colors = ann_colors,
         show_colnames = TRUE,
         main = "Heatmap Top Random Forest Genes by Specimen")
```

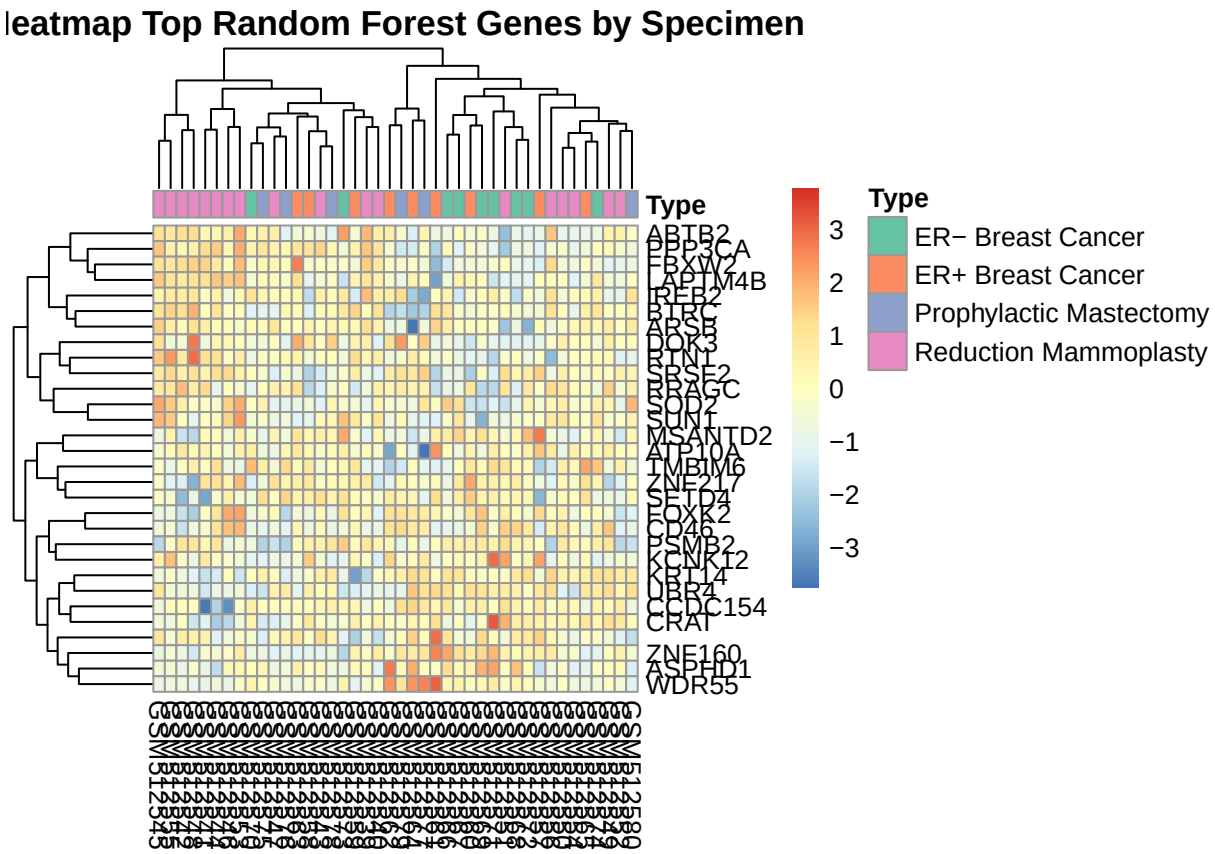


Figure 18: Heatmap of the most significant genes according to random forest, showing the specimen.

LDA

```r
# Convert group labels to factor
group_factor <- as.factor(metadata$disease.state.ch1)

# Perform t-tests
ttest <- genefilter::rowttests(as.matrix(t(expr_data)), group_factor)
ttest <- which(p.adjust(ttest$p.value)<0.1)
```

```r
## Reduce the original dataset
ttest_data <- expr_data[,ttest]
#dim(ttest_data) #42 2
ttest_train <- ttest_data[train_index, ]
#dim(ttest_train) #30 2
```

```r
set.seed(1)

## Fit the model with cv
control <- trainControl(method="cv", number=10)
lda_fit <- train(ttest_train, train_labels, method="lda",
                 metric="Accuracy", trControl=control)
print(lda_fit)
```

```
## Linear Discriminant Analysis
##
## 30 samples
##  2 predictor
##  2 classes: 'breast cancer', 'control'
##
## No pre-processing
## Resampling: Cross-Validated (10 fold)
## Summary of sample sizes: 27, 27, 27, 26, 27, 28, ...
## Resampling results:
##
##   Accuracy   Kappa
##   0.7583333  0.53
```

```r
## Apply it on the test set
pred_lda <- predict(lda_fit, test_data)
## Accuracy
acc_lda <- mean(pred_lda==test_labels)

## AUC from ROC curve
pred_lda <- predict(lda_fit, test_data, type = "prob")
roc_lda <- roc(test_labels, pred_lda[,"breast cancer"])
auc_lda <- auc(roc_lda)
#plot(roc_rf)
auc_lda <- auc(roc_lda)


## Update results table
res_df["LDA"] <- c(acc_lda, auc_lda)

#print(lda_fit)
```

LASSO

```r
set.seed(1)

## fit Lasso with cv
control <- trainControl(method="cv", number=10)
tunegrid <- expand.grid(alpha=1,
```

```
                        lambda=seq(.0001,1,by=.001))
lasso_fit <- train(train_data, train_labels, method="glmnet", family="binomial",
                   tuneGrid = tunegrid,
                   metric="Accuracy", trControl=control)
#print(lasso_fit)

# Apply it on the test set
pred_lasso <- predict(lasso_fit, test_data)

## Accuracy
acc_lasso <- mean(pred_lasso==test_labels)

pred_lasso <- predict(lasso_fit, test_data, type = "prob")
roc_lasso <- roc(test_labels, pred_lasso[,"breast cancer"])
auc_lasso <- auc(roc_lasso)


## Update results table
res_df["Lasso"] <- c(acc_lasso, auc_lasso)
```

```
results <- lasso_fit$results

ggplot(results, aes(x = lambda, y = Accuracy)) +
  geom_line(color = "steelblue") +
  geom_point(color = "darkred") +
  labs(title = "Lasso: Accuracy vs. Lambda",
       x = expression(lambda),
       y = "Accuracy") +
  theme_minimal()
```

```
best_lambda <- lasso_fit$bestTune$lambda
print(best_lambda)
```

```
## [1] 0.2021
```

SCUDO

The goal of this method is to avoid batch effects and obtain a method that is repeatable and reproducible.

This method compares signatures of different individuals. Each signature is sorted for the expression value. The most important genes for each signature are the most and the less expressed. Then signatures are compared and a map is constructed, where clusters can be seen if they are closely connected.

```
set.seed(1)
## Apply SCUDO on the training set
scudo_train <- scudoTrain(t(train_data), groups = as.factor(train_labels),
                          nTop = 25, nBottom = 25, alpha = 0.05)
scudo_train
```

```
## Object of class ScudoResults
## Result of scudoTrain
##
## Number of samples      :  30
## Number of groups       :  2
##     breast cancer :  13 samples
##     control :  17 samples
```

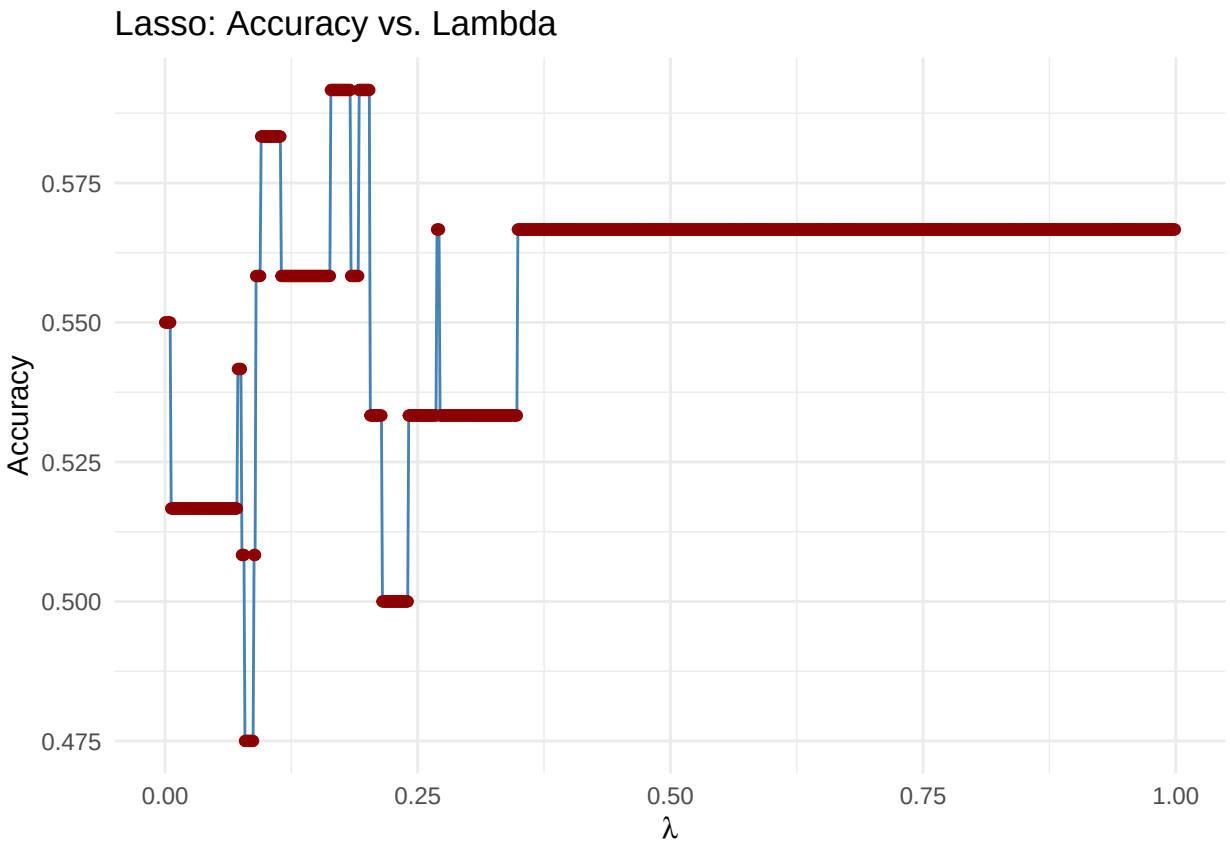


Figure 19: Plot showing how LASSO accuracy changes according to the lambda parameter chosen.

```
## upSignatures length    :  25
## downSignatures length  :  25
## Fold-changes           :  computed
##     grouped            :  No
## Feature selection      :  performed
##     Test               :  Wilcoxon rank sum test
##     p-value cutoff     :  0.05
##     p.adjust method    :  none
##     Selected features  :  1489
```

```r
## perform validation using testing samples
scudo_test <- scudoTest(scudo_train, t(test_data), as.factor(test_labels),
                        nTop = 25, nBottom = 25)

## Plot the results
testNet <- scudoNetwork(scudo_test, N = 0.2)
scudoPlot(testNet, vertex.label = NA)
```

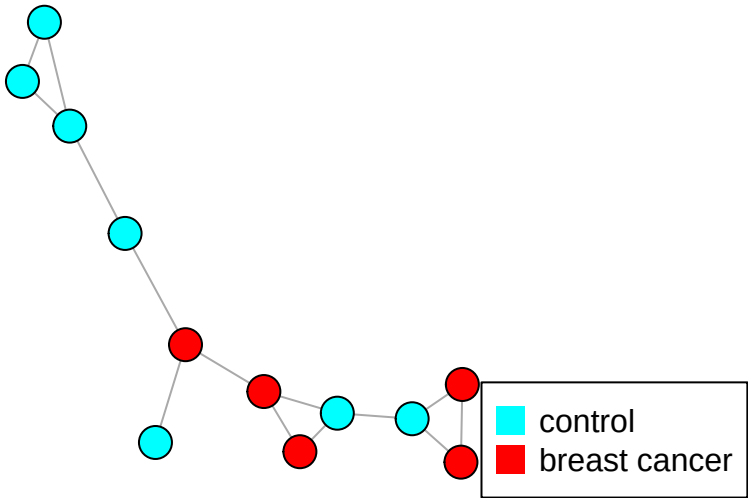


Figure 20: SCUDO network

```r
## Classification
pred_scudo <- scudoClassify(t(train_data), t(test_data),
                            nTop = 25, nBottom = 25, N = 0.2,
                            trainGroups = as.factor(train_labels),
                            featureSel = FALSE)
## Accuracy
acc_scudo <- mean(pred_scudo$predicted==test_labels)

## AUC from ROC curve
roc_scudo <- roc(test_labels, pred_scudo$scores[,2])
# plot(roc_rf)
auc_scudo <- auc(roc_scudo)

## Update results table
res_df["Scudo"] <- c(acc_scudo, auc_scudo)
```

Model comparison

The models fitted before are compared by accuracy and AUC.

```
res_df <- as.data.frame(res_df)
res_df$metric <- rownames(res_df)

res_long <- pivot_longer(res_df,
                         cols = -metric,
                         names_to = "Model",
                         values_to = "Value")
#order models by AUC
res_long$Model <- factor(res_long$Model,
                         levels = res_long %>%
                           filter(metric == "AUC") %>%
                           arrange(desc(Value)) %>%
                           pull(Model))

# Plot with value labels
ggplot(res_long, aes(x = Model, y = Value, fill = metric)) +
  geom_bar(stat = "identity", position = position_dodge(width = 0.9)) +
  geom_text(aes(label = round(Value, 2)),
            position = position_dodge(width = 0.9),
            vjust = -0.3, size = 3) +
  scale_fill_manual(values = c("Accuracy" = "#1b9e77", "AUC" = "#d95f02")) +
  labs(title = "Model Performance: Accuracy vs AUC",
       x = "Model",
       y = "Metric Value",
       fill = "Metric") +
  ylim(0, 1.1) +  # add a little headroom for the labels
  theme_minimal() +
  theme(axis.text.x = element_text(angle = 45, hjust = 1))
```

Functional enrichment analysis

The functional enrichment analysis is performed with `gprofiler2`, with the objective of finding out the molecular functions more represented among the most important genes. For this purpose, the genes with highest importance score according to the random forest model fitted before are considered.

```
geneList <- head(ordered_probes, 30)
gost_res <- gost(query = geneList,
                 organism = "hsapiens",
                 ordered_query = FALSE, multi_query = FALSE, significant = FALSE,
                 exclude_iea = FALSE, measure_underrepresentation = FALSE, evcodes = FALSE,
                 user_threshold = 0.05, correction_method = "g_SCS",
                 domain_scope = "annotated", custom_bg = NULL, numeric_ns = "",
                 sources = NULL, as_short_link = FALSE)

gost_df <- gost_res$result[c("term_id", "term_name", "p_value", "significant")]
gost_df <- gost_df[order(gost_df$p_value),]
gost_df <- gost_df[which(gost_df$significant==TRUE),]

# head(gost_df)
```

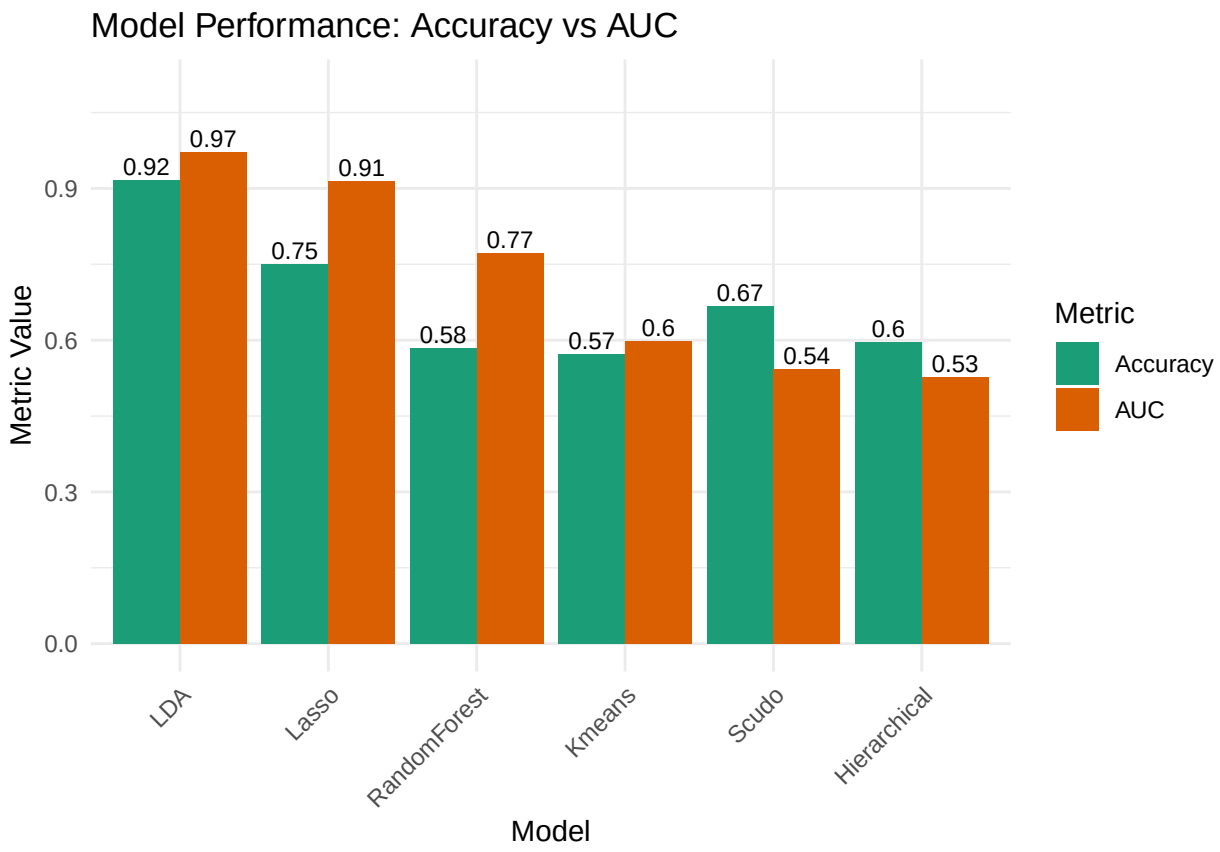


Figure 21: Comparison of the models' performances.

```r
# visualize results using a Manhattan plot
p <- gostplot(gost_res, capped = TRUE, interactive = FALSE)
publish_gostplot(p, filename = NULL)
```

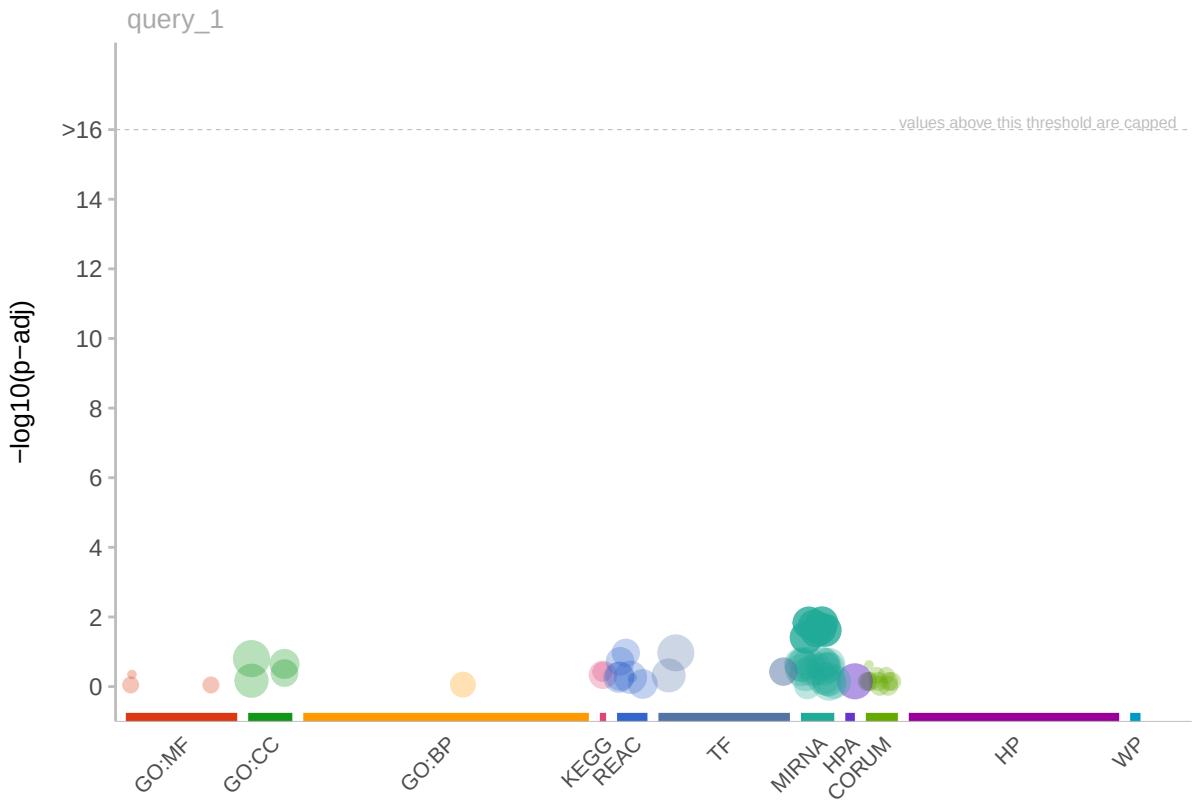


Figure 22: Results of the functional enrichment analysis

```r
publish_gosttable(gost_res, highlight_terms = gost_df$term_id[1:10],
                  use_colors = TRUE, filename = NULL,
                  show_columns = c("term_id", "term_name", "p_value"))
```

Network-based Analysis

The network analysis is performed with `pathfindR`, using the *KEGG*, *Gene Ontology* and *Reactome* databases. For this analysis, the 100 genes with the lowest p-value obtained with a t-test are taken into consideration. This list of genes has 23 genes in common with the one of the most important genes used for the functional enrichment.

```r
# Map probe IDs to Ensembl gene IDs
probe_to_ensembl <- setNames(annotLookup$ensembl_gene_id, annotLookup$affy_hg_u133_plus_2)

new_colnames <- probe_to_ensembl[colnames(expr_data)]

# Warn if there are unmatched probe IDs
unmatched <- is.na(new_colnames)
if (any(unmatched)) {
  warning(sum(unmatched), " probe IDs in expr_data were not found in annotLookup and will be removed.")
}
```

id	term_id	term_name		p_value
1	MIRNA:hsa−mir−3129−5p	hsa−mir−3129−5p		1.5e−02
2	MIRNA:hsa−mir−199b−3p	hsa−mir−199b−3p		1.5e−02
3	MIRNA:hsa−mir−199a−3p	hsa−mir−199a−3p		1.8e−02
4	MIRNA:hsa−mir−936	hsa−mir−936		2.2e−02
5	MIRNA:hsa−mir−4786−3p	hsa−mir−4786−3p		2.4e−02
6	MIRNA:hsa−mir−1263	hsa−mir−1263		2.6e−02
7	MIRNA:hsa−mir−3925−5p	hsa−mir−3925−5p		3.8e−02

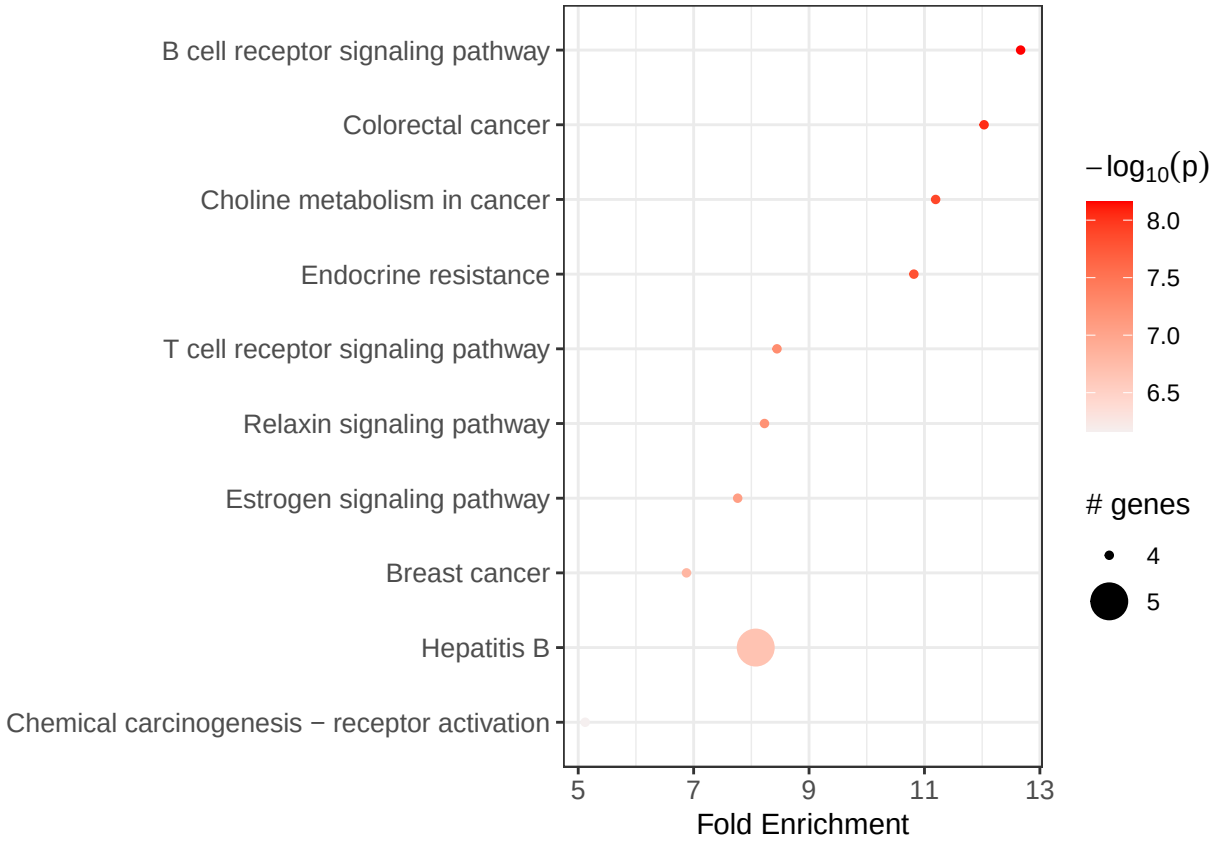
g:Profiler (biit.cs.ut.ee/gprofiler)

Figure 23: Results of the functional enrichment analysis

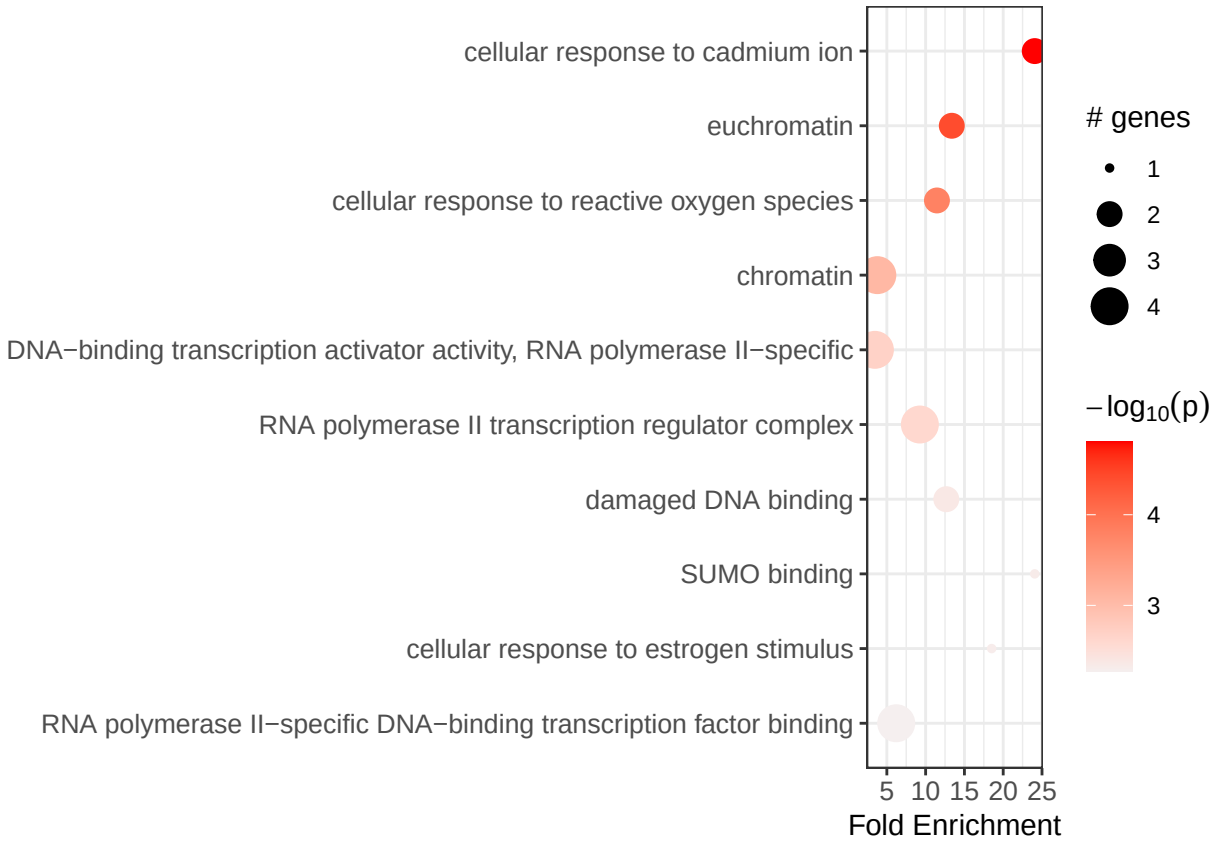
```
expr_data_new <- expr_data[, !unmatched]
colnames(expr_data_new) <- new_colnames[!unmatched]

# duplicated Ensembl IDs
expr_t <- t(expr_data_new)
expr_df <- data.frame(expr_t, ensembl_id = colnames(expr_data_new))
collapsed <- aggregate(. ~ ensembl_id, data = expr_df, FUN = mean)
rownames(collapsed) <- collapsed$ensembl_id
collapsed$ensembl_id <- NULL
expr_data_collapsed <- t(collapsed)

# Run t-test across two groups
tt <- genefilter::rowttests(as.matrix(t(expr_data_collapsed)), group_factor)
# Sort by p-value
tt <- tt[order(tt$p.value), ]

# Extract top 100 genes by p-value
top_n <- 100
geneList_tt <- data.frame(
  ENSEMBL = rownames(tt)[1:top_n],
  p.value = tt$p.value[1:top_n]
  # p.adjust = p.adjust(tt$p.value[1:top_n])
)

# Remove any trailing version numbers in Ensembl IDs (e.g., ENSG00000001234.1 -> ENSG00000001234)
gene_ids_split <- unlist(strsplit(as.character(geneList_tt$ENSEMBL), ".", fixed = TRUE))
geneList_tt$ENSEMBL <- gene_ids_split[startsWith(gene_ids_split, "ENS")]


# Annotate Ensembl IDs with gene symbols
gene_symbol_df <- bitr(geneList_tt$ENSEMBL,
                       fromType = "ENSEMBL",
                       toType = "SYMBOL",
                       OrgDb = org.Hs.eg.db)

geneList_tt <- merge(gene_symbol_df, geneList_tt)
```

```
# dim(geneList_tt) #105   3
```

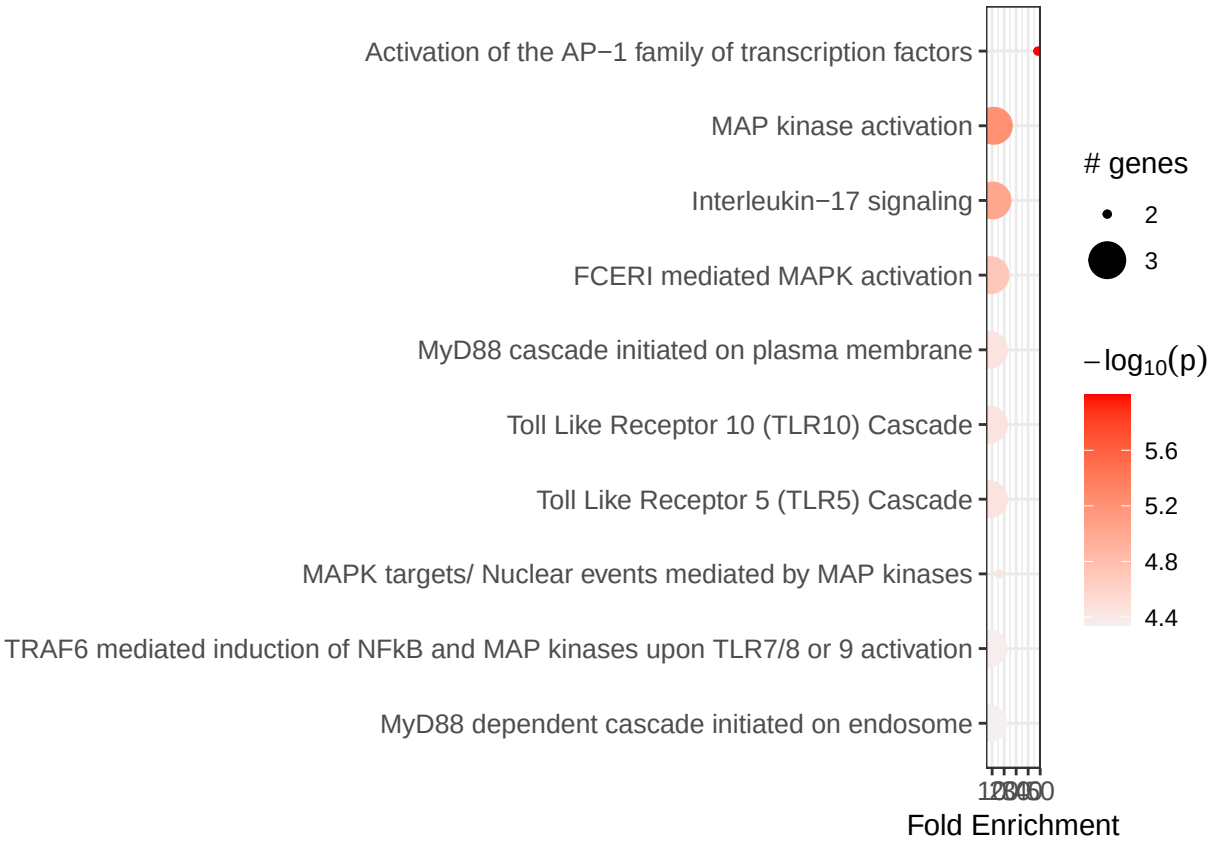
```
net_KEGG <- run_pathfindR(
  geneList_tt[c("SYMBOL", "p.value")],
  iterations = 1,
  gene_sets = "KEGG",
  silent_option = FALSE
)
```




```
net_GO <- run_pathfindR(
  geneList_tt[c("SYMBOL", "p.value")],
  iterations = 1,
  gene_sets = "GO-All",
  silent_option = FALSE
)
```




```
net_Reactome <- run_pathfindR(geneList_tt[c("SYMBOL", "p.value")],
                              iterations = 1,
                              gene_sets = "Reactome",
                              silent_option = FALSE)
```




```
enrichment_chart(net_KEGG)
```

```
enrichment_chart(net_GO)
```

```
enrichment_chart(net_Reactome)
```

```
term_gene_graph(net_Reactome, num_terms = 7, use_description = TRUE)
```

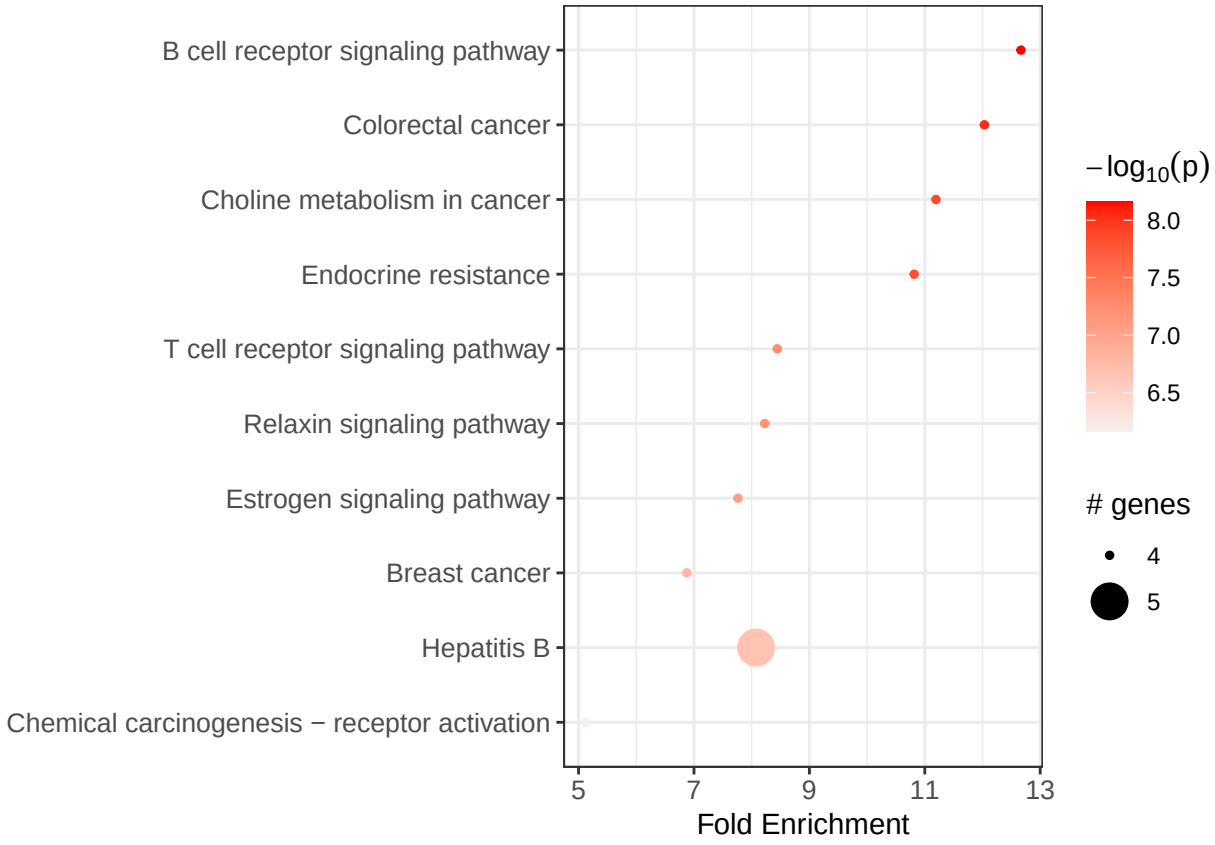


Figure 24: Results of the biological network analysis

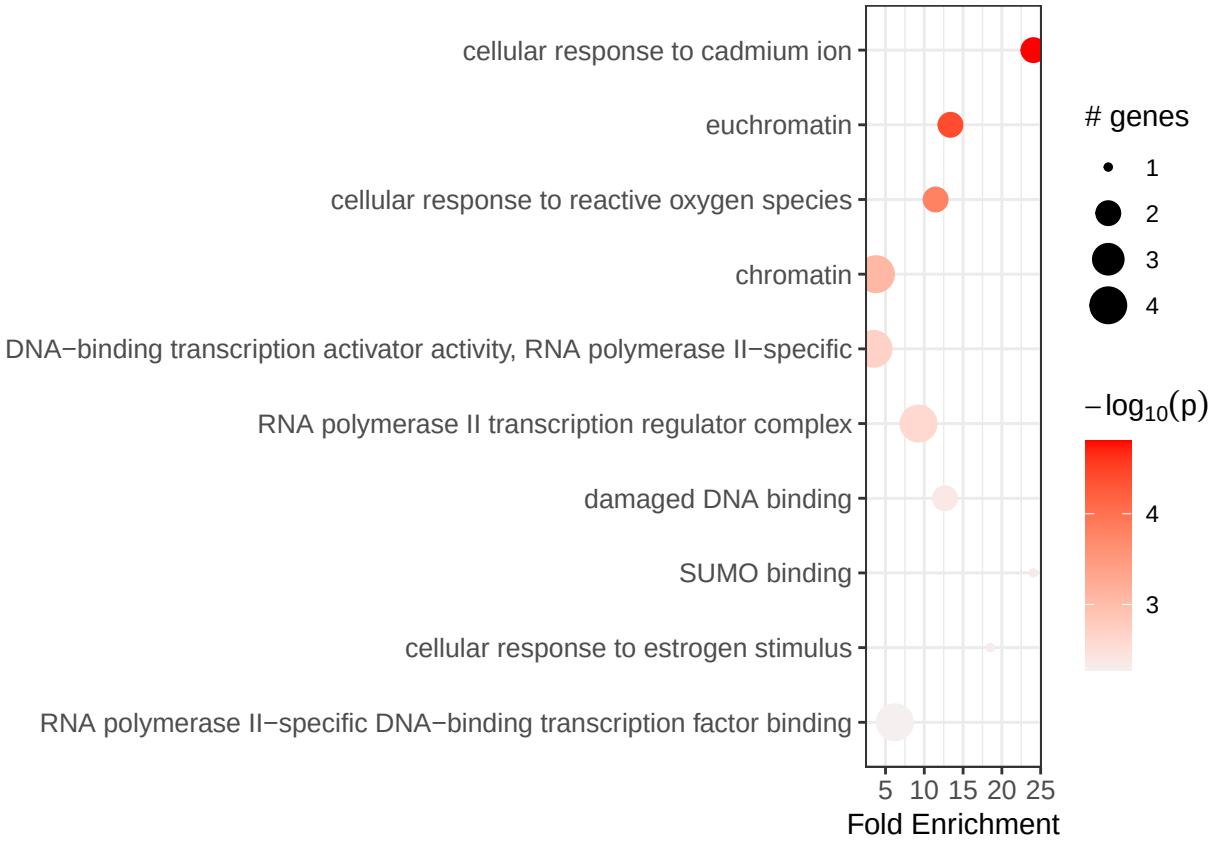


Figure 25: Results of the biological network analysis

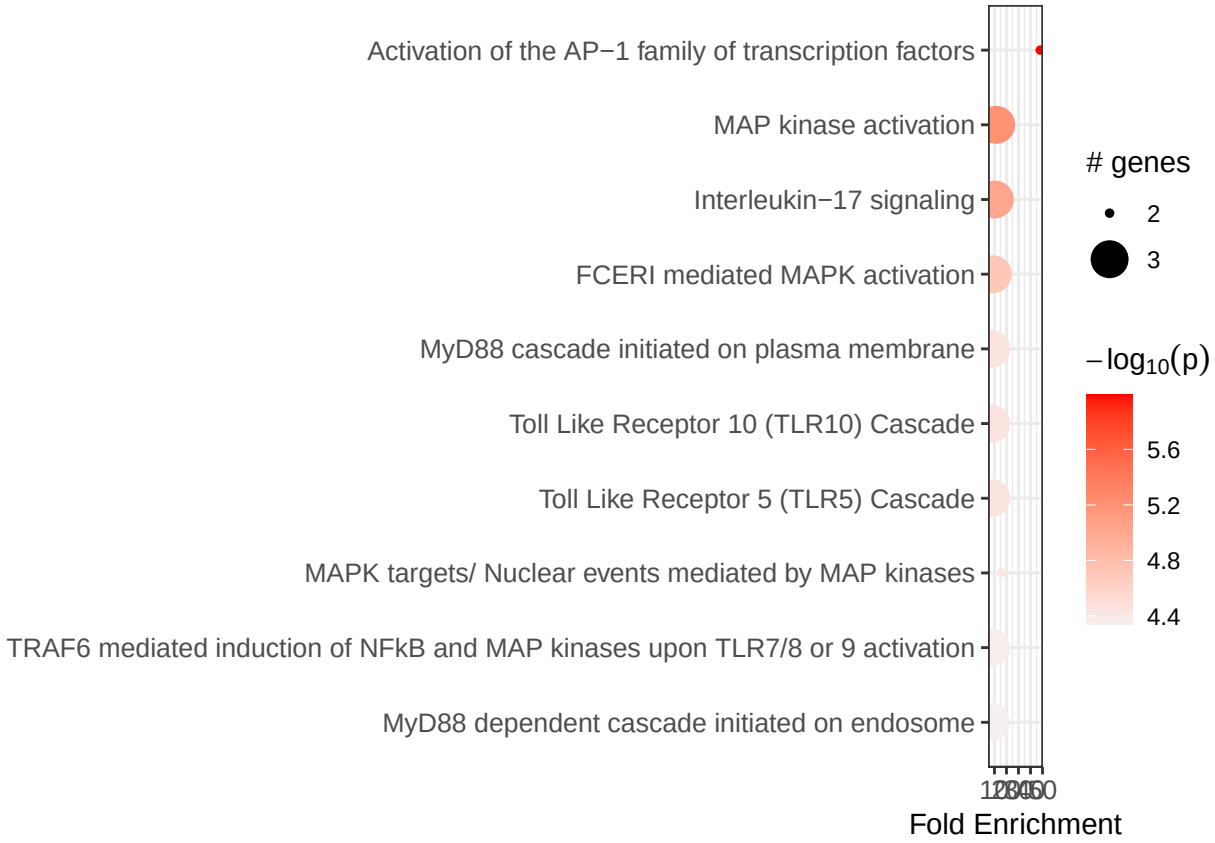
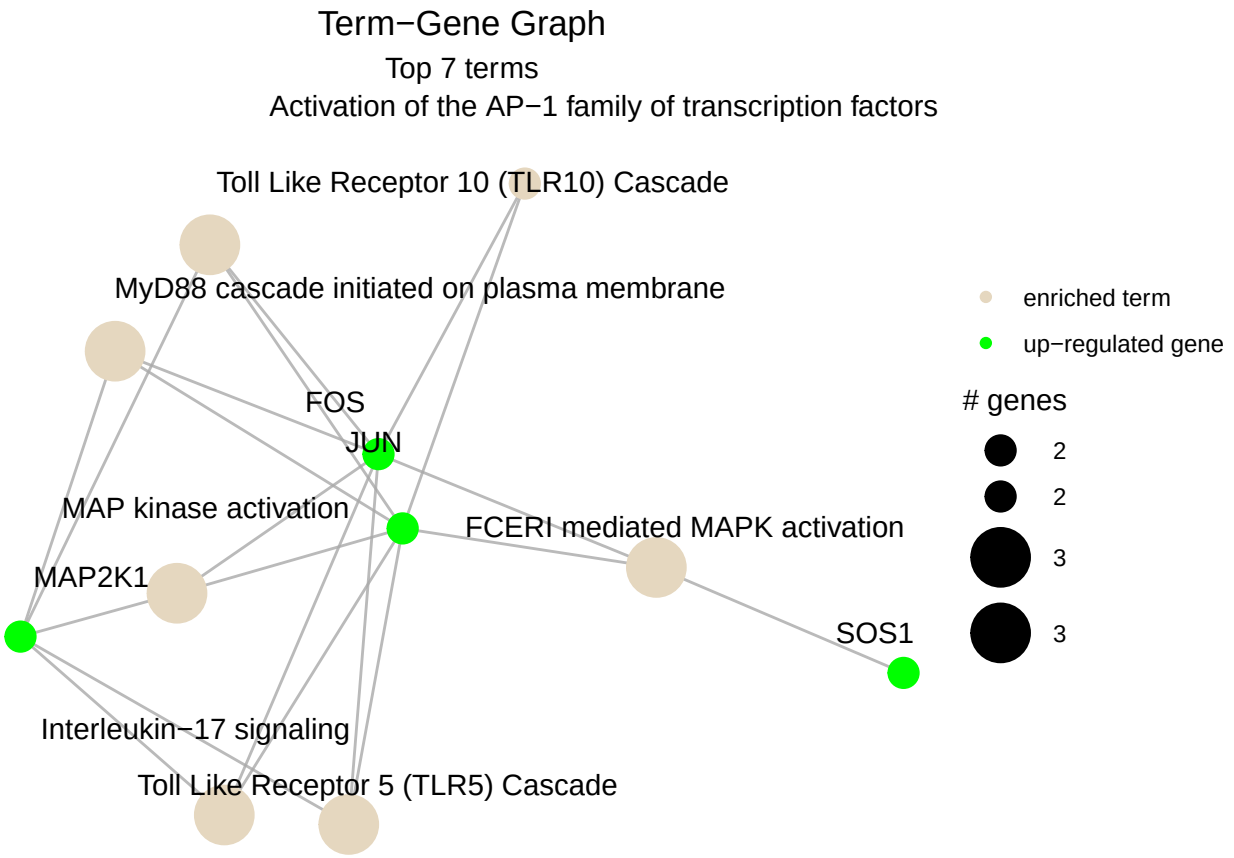
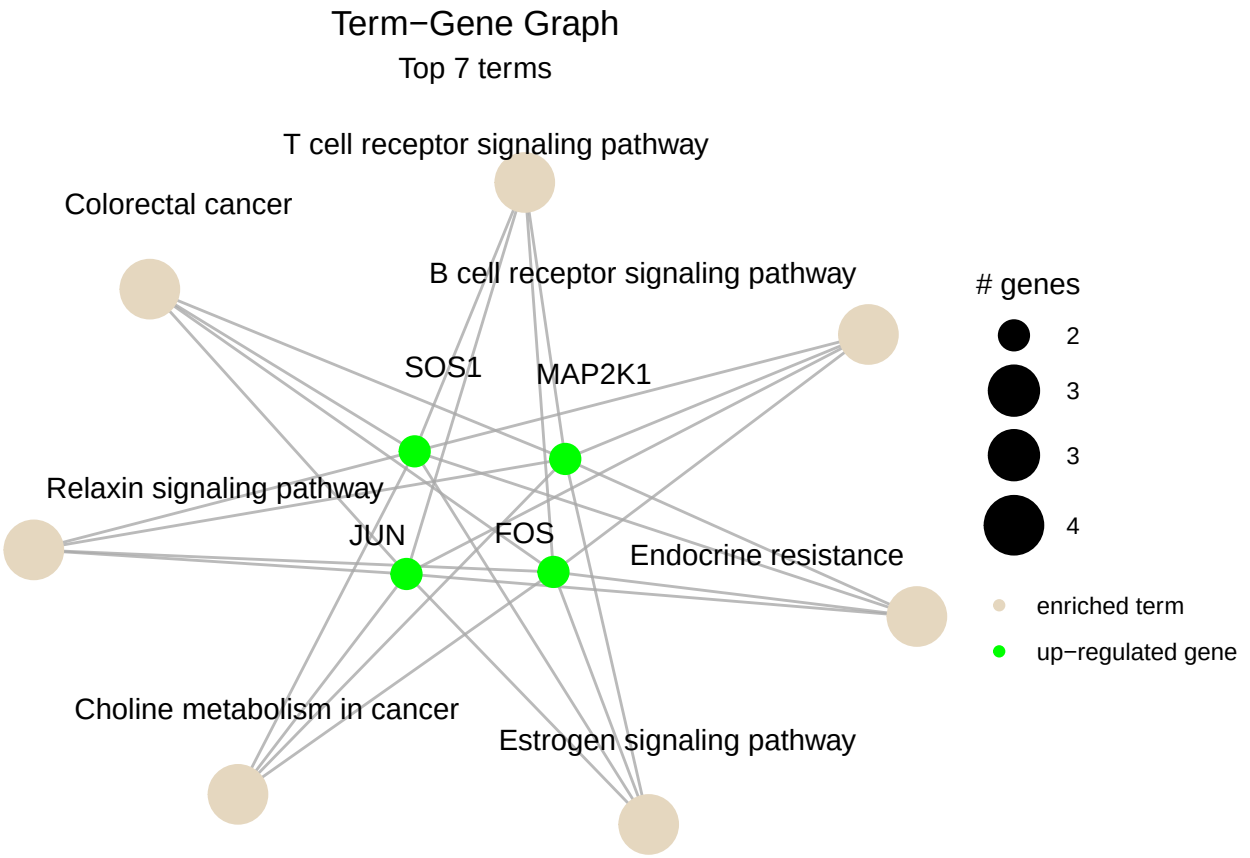


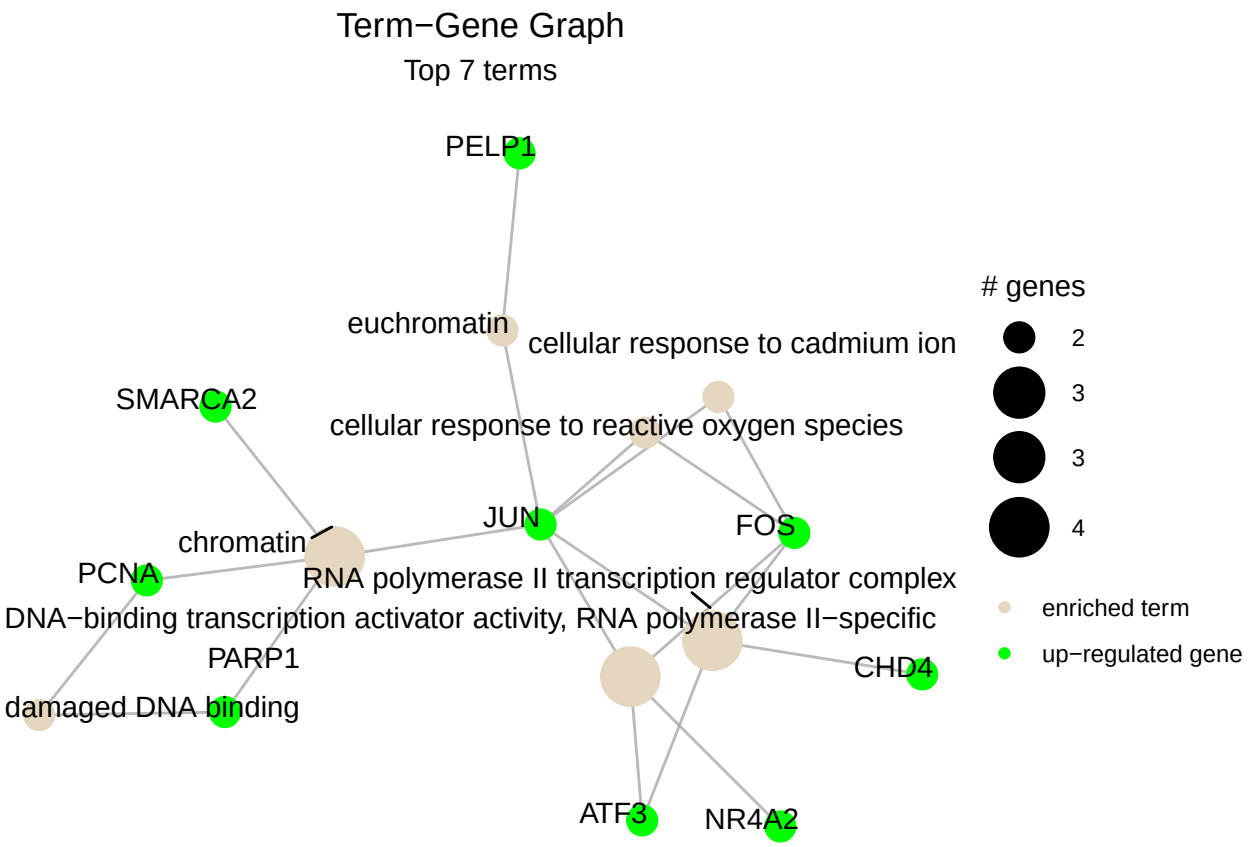
Figure 26: Results of the biological network analysis



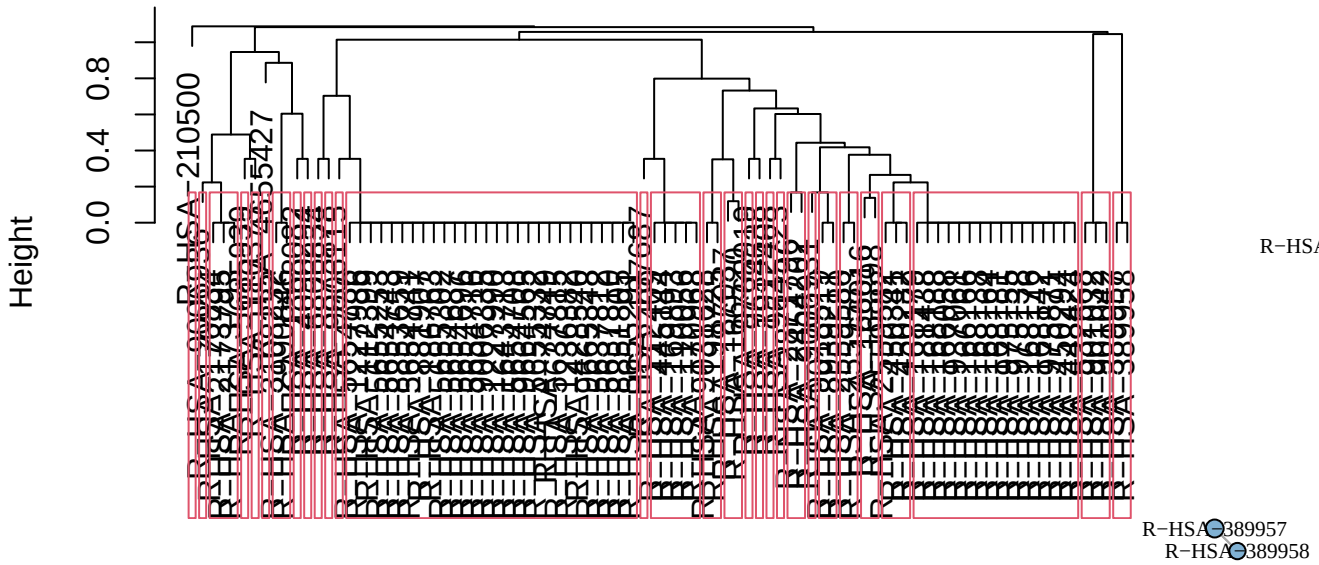

```
term_gene_graph(net_KEGG, num_terms = 7, use_description = TRUE)
```




```
term_gene_graph(net_GO, num_terms = 7, use_description = TRUE)
```




```
## cluster enriched terms
cluster_enriched_terms(net_Reactome)
```



	ID	Term_Description	Fold_E
1	R-HSA-450341	Activation of the AP-1 family of transcription factors	4
8	R-HSA-450282	MAPK targets/ Nuclear events mediated by MAP kinases	
33	R-HSA-2559582	Senescence-Associated Secretory Phenotype (SASP)	
2	R-HSA-450294	MAP kinase activation	
3	R-HSA-448424	Interleukin-17 signaling	
5	R-HSA-975871	MyD88 cascade initiated on plasma membrane	
6	R-HSA-168142	Toll Like Receptor 10 (TLR10) Cascade	
7	R-HSA-168176	Toll Like Receptor 5 (TLR5) Cascade	
9	R-HSA-975138	TRAF6 mediated induction of NFkB and MAP kinases upon TLR7/8 or 9 activation	
10	R-HSA-975155	MyD88 dependent cascade initiated on endosome	
11	R-HSA-168181	Toll Like Receptor 7/8 (TLR7/8) Cascade	
12	R-HSA-168164	Toll Like Receptor 3 (TLR3) Cascade	
13	R-HSA-168138	Toll Like Receptor 9 (TLR9) Cascade	
14	R-HSA-166166	MyD88-independent TLR4 cascade	
15	R-HSA-937061	TRIF (TICAM1)-mediated TLR4 signaling	
17	R-HSA-166058	MyD88:MAL(TIRAP) cascade initiated on plasma membrane	
19	R-HSA-168188	Toll Like Receptor TLR6:TLR2 Cascade	
22	R-HSA-181438	Toll Like Receptor 2 (TLR2) Cascade	
23	R-HSA-168179	Toll Like Receptor TLR1:TLR2 Cascade	
4	R-HSA-2871796	FCERI mediated MAPK activation	
31	R-HSA-2454202	Fc epsilon receptor (FCERI) signaling	
16	R-HSA-3108232	SUMO E3 ligases SUMOylate target proteins	
21	R-HSA-2990846	SUMOylation	
18	R-HSA-187037	Signaling by NTRK1 (TRKA)	
24	R-HSA-166520	Signaling by NTRKs	
20	R-HSA-2173795	Downregulation of SMAD2/3:SMAD4 transcriptional activity	
29	R-HSA-2173793	Transcriptional activity of SMAD2/SMAD3:SMAD4 heterotrimer	

```
## 36  R-HSA-170834                                    Signaling by TGF-beta Receptor Complex
## 25  R-HSA-166016                                       Toll Like Receptor 4 (TLR4) Cascade
## 28  R-HSA-168898                                             Toll-like Receptor Cascades
## 26 R-HSA-6806834                                                      Signaling by MET
## 27  R-HSA-187687                                                  Signalling to ERKs
## 30 R-HSA-2559580                                       Oxidative Stress Induced Senescence
## 37 R-HSA-2559583                                                 Cellular Senescence
## 32 R-HSA-9018519                                  Estrogen-dependent gene expression
## 43 R-HSA-8939211                                            ESR-mediated signaling
## 34 R-HSA-8851805                                     MET activates RAS signaling
## 35 R-HSA-5655291                                   Signaling by FGFR4 in disease
## 39 R-HSA-5637810                                Constitutive Signaling by EGFRvIII
## 40 R-HSA-5637812                                   Signaling by EGFRvIII in Cancer
## 41 R-HSA-9665348                                   Signaling by ERBB2 ECD mutants
## 42  R-HSA-428540                                              Activation of RAC1
## 46 R-HSA-1236382 Constitutive Signaling by Ligand-Responsive EGFR Cancer Variants
## 47 R-HSA-5637815          Signaling by Ligand-Responsive EGFR Variants in Cancer
## 50   R-HSA-74749                                              Signal attenuation
## 51 R-HSA-5655332                                   Signaling by FGFR3 in disease
## 52 R-HSA-9664565                                   Signaling by ERBB2 KD Mutants
## 54 R-HSA-5654708                          Downstream signaling of activated FGFR3
## 55 R-HSA-1643713                                     Signaling by EGFR in Cancer
## 56 R-HSA-1227990                                    Signaling by ERBB2 in Cancer
## 57 R-HSA-9006335                                    Signaling by Erythropoietin
## 62 R-HSA-5654716                          Downstream signaling of activated FGFR4
## 65 R-HSA-5654696                          Downstream signaling of activated FGFR2
## 66 R-HSA-5654687                          Downstream signaling of activated FGFR1
## 70 R-HSA-5655302                                   Signaling by FGFR1 in disease
## 71  R-HSA-186763                                 Downstream signal transduction
## 73  R-HSA-881907                  Gastrin-CREB signalling pathway via PKC and MAPK
## 74 R-HSA-5654741                                              Signaling by FGFR3
## 76 R-HSA-8853659                                                   RET signaling
## 77 R-HSA-5654743                                              Signaling by FGFR4
## 80 R-HSA-5655253                                   Signaling by FGFR2 in disease
## 85  R-HSA-112399                                         IRS-mediated signalling
## 86 R-HSA-1227986                                              Signaling by ERBB2
## 89 R-HSA-5654736                                              Signaling by FGFR1
## 38 R-HSA-9031628                                   NGF-stimulated transcription
## 61  R-HSA-198725 Nuclear Events (kinase and transcription factor activation)
## 44 R-HSA-9006936                                 Signaling by TGFB family members
## 45 R-HSA-5685939                                      HDR through MMEJ (alt-NHEJ)
## 48  R-HSA-901032                           ER Quality Control Compartment (ERQC)
## 59  R-HSA-901042                                    Calnexin/calreticulin cycle
## 68  R-HSA-532668 N-glycan trimming in the ER and Calnexin/Calreticulin cycle
## 49  R-HSA-110056                                       MAPK3 (ERK1) activation
## 60  R-HSA-170968                                      Frs2-mediated activation
## 64  R-HSA-169893                                 Prolonged ERK activation events
## 81  R-HSA-445144                                       Signal transduction by L1
## 87  R-HSA-112409                                RAF-independent MAPK1/3 activation
## 53 R-HSA-4655427                          SUMOylation of DNA methylation proteins
## 58 R-HSA-9006115                                       Signaling by NTRK2 (TRKB)
## 63 R-HSA-1368082                                 RORA activates gene expression
## 67 R-HSA-9006931                                 Signaling by Nuclear Receptors
## 69  R-HSA-389957                Prefoldin mediated transfer of substrate  to CCT/TriC
```

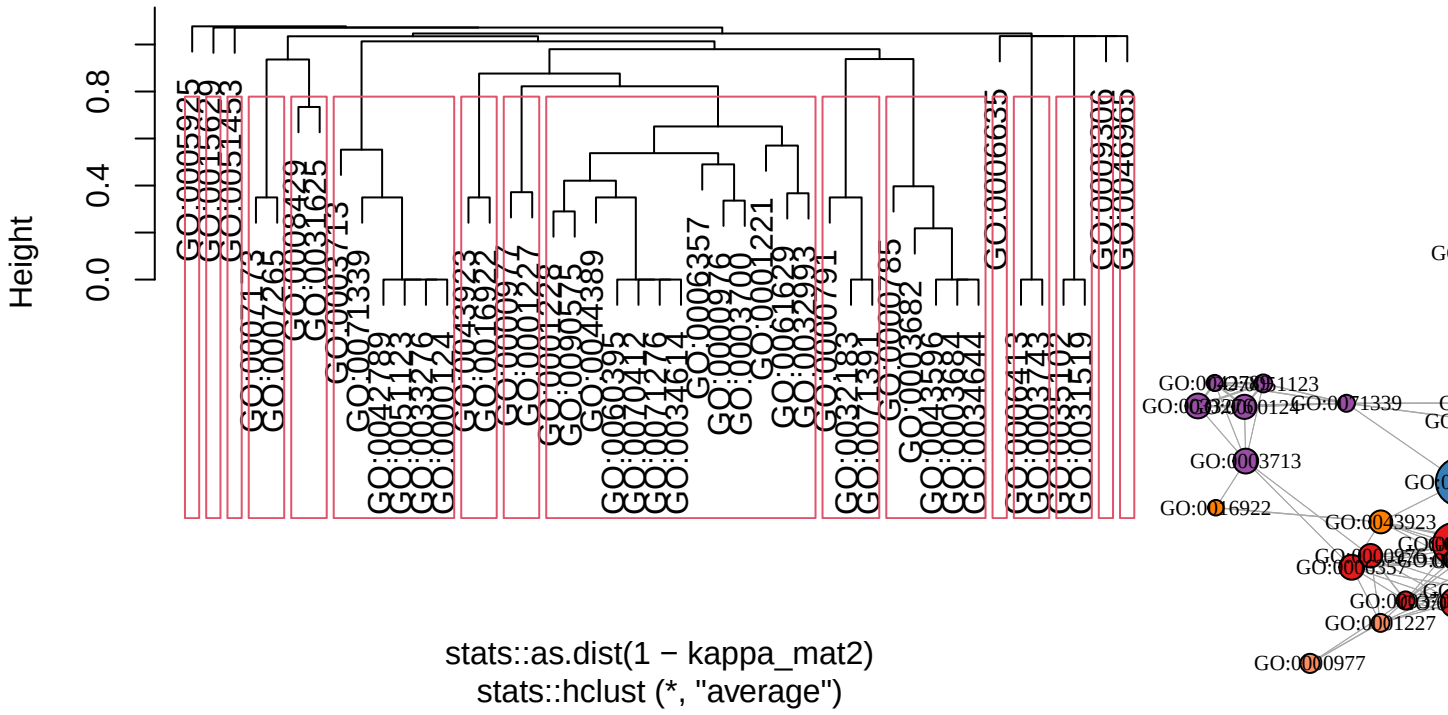
```
## 78  R-HSA-389958          Cooperation of Prefoldin and TriC/CCT  in actin and tubulin folding
## 72  R-HSA-110373 Resolution of AP sites via the multiple-nucleotide patch replacement pathway      1
## 75 R-HSA-8875878                                                 MET promotes cell motility
## 79 R-HSA-1912408                                    Pre-NOTCH Transcription and Translation      1
## 82 R-HSA-9768919                                  NPAS4 regulates expression of target genes    1
## 83 R-HSA-9645723                                         Diseases of programmed cell death
## 84 R-HSA-4090294                                   SUMOylation of intracellular receptors      1
## 88  R-HSA-210500                                  Glutamate Neurotransmitter Release Cycle      1
## 90 R-HSA-9634638       Estrogen-dependent nuclear events downstream of ESR-membrane signaling      1
##    occurrence support     lowest_p    highest_p                        Up_regulated Down_regulated C
## 1           1   0.040 1.010918e-06 1.010918e-06                            FOS, JUN
## 8           1   0.030 3.781458e-05 3.781458e-05                            FOS, JUN
## 33          1   0.005 1.160838e-03 1.160838e-03                            FOS, JUN
## 2           1   0.010 6.163032e-06 6.163032e-06                     FOS, JUN, MAP2K1
## 3           1   0.010 9.477220e-06 9.477220e-06                     FOS, JUN, MAP2K1
## 5           1   0.010 3.567643e-05 3.567643e-05                     FOS, JUN, MAP2K1
## 6           1   0.010 3.567643e-05 3.567643e-05                     FOS, JUN, MAP2K1
## 7           1   0.010 3.567643e-05 3.567643e-05                     FOS, JUN, MAP2K1
## 9           1   0.010 4.372557e-05 4.372557e-05                     FOS, JUN, MAP2K1
## 10          1   0.010 4.548581e-05 4.548581e-05                     FOS, JUN, MAP2K1
## 11          1   0.010 4.729853e-05 4.729853e-05                     FOS, JUN, MAP2K1
## 12          1   0.010 5.108554e-05 5.108554e-05                     FOS, JUN, MAP2K1
## 13          1   0.010 5.306191e-05 5.306191e-05                     FOS, JUN, MAP2K1
## 14          1   0.010 5.933519e-05 5.933519e-05                     FOS, JUN, MAP2K1
## 15          1   0.010 5.933519e-05 5.933519e-05                     FOS, JUN, MAP2K1
## 17          1   0.010 6.854303e-05 6.854303e-05                     FOS, JUN, MAP2K1
## 19          1   0.010 6.854303e-05 6.854303e-05                     FOS, JUN, MAP2K1
## 22          1   0.010 7.612032e-05 7.612032e-05                     FOS, JUN, MAP2K1
## 23          1   0.010 7.612032e-05 7.612032e-05                     FOS, JUN, MAP2K1
## 4           1   0.010 1.895904e-05 1.895904e-05                       FOS, JUN, SOS1
## 31          1   0.010 4.400761e-04 4.400761e-04                FOS, JUN, PSMB2, SOS1
## 16          1   0.020 6.580722e-05 6.580722e-05       NR4A2, PARP1, PCNA, PHC2, RORA
## 21          1   0.020 7.597991e-05 7.597991e-05       NR4A2, PARP1, PCNA, PHC2, RORA
## 18          1   0.005 6.854303e-05 6.854303e-05        CHD4, FOS, FOSB, MAP2K1, SOS1
## 24          1   0.005 1.276167e-04 1.276167e-04 BDNF, CHD4, FOS, FOSB, MAP2K1, SOS1
## 20          1   0.015 7.555895e-05 7.555895e-05                          PARP1, SNW1
## 29          1   0.010 3.493625e-04 3.493625e-04                          PARP1, SNW1
## 36          1   0.005 2.239119e-03 2.239119e-03                          PARP1, SNW1
## 25          1   0.010 1.908991e-04 1.908991e-04              CD180, FOS, JUN, MAP2K1
## 28          1   0.010 3.263808e-04 3.263808e-04        CD180, FOS, JUN, LGMN, MAP2K1
## 26          1   0.010 2.659266e-04 2.659266e-04                         SOS1, TNS4
## 27          1   0.005 2.809849e-04 2.809849e-04                       MAP2K1, SOS1
## 30          1   0.010 4.020427e-04 4.020427e-04                     FOS, JUN, PHC2
## 37          1   0.010 3.130768e-03 3.130768e-03                     FOS, JUN, PHC2
## 32          1   0.020 4.407019e-04 4.407019e-04                     FOS, FOSB, JUN
## 43          1   0.020 4.195628e-03 4.195628e-03                     FOS, FOSB, JUN
## 34          1   0.005 2.096000e-03 2.096000e-03                               SOS1
## 35          1   0.005 2.096000e-03 2.096000e-03                               SOS1
## 39          1   0.005 4.000511e-03 4.000511e-03                               SOS1
## 40          1   0.005 4.000511e-03 4.000511e-03                               SOS1
## 41          1   0.005 4.000511e-03 4.000511e-03                               SOS1
## 42          1   0.010 4.190517e-03 4.190517e-03                               SOS1
## 46          1   0.005 6.513581e-03 6.513581e-03                               SOS1
## 47          1   0.005 6.513581e-03 6.513581e-03                               SOS1
```

```
## 50          1   0.005 7.995835e-03 7.995835e-03                    SOS1
## 51          1   0.005 8.797491e-03 8.797491e-03                    SOS1
## 52          1   0.005 1.051005e-02 1.051005e-02                    SOS1
## 54          1   0.005 1.142329e-02 1.142329e-02                    SOS1
## 55          1   0.005 1.142329e-02 1.142329e-02                    SOS1
## 56          1   0.005 1.142329e-02 1.142329e-02                    SOS1
## 57          1   0.005 1.142329e-02 1.142329e-02                    SOS1
## 62          1   0.005 1.336367e-02 1.336367e-02                    SOS1
## 65          1   0.005 1.655889e-02 1.655889e-02                    SOS1
## 66          1   0.005 1.769984e-02 1.769984e-02                    SOS1
## 70          1   0.005 2.534176e-02 2.534176e-02                    SOS1
## 71          1   0.010 2.573930e-02 2.573930e-02                    SOS1
## 73          1   0.005 2.714740e-02 2.714740e-02                    SOS1
## 74          1   0.005 2.967429e-02 2.967429e-02                    SOS1
## 76          1   0.005 3.119421e-02 3.119421e-02                    SOS1
## 77          1   0.005 3.119421e-02 3.119421e-02                    SOS1
## 80          1   0.005 3.434762e-02 3.434762e-02                    SOS1
## 85          1   0.005 4.289333e-02 4.289333e-02                    SOS1
## 86          1   0.005 4.471594e-02 4.471594e-02                    SOS1
## 89          1   0.005 4.847457e-02 4.847457e-02                    SOS1
## 38          1   0.025 3.190319e-03 3.190319e-03          CHD4, FOS, FOSB
## 61          1   0.005 1.282843e-02 1.282843e-02          CHD4, FOS, FOSB
## 44          1   0.005 5.164939e-03 5.164939e-03       BMP2, PARP1, SNW1
## 45          1   0.005 6.283923e-03 6.283923e-03                   PARP1
## 48          1   0.005 7.236885e-03 7.236885e-03                  RNF139
## 59          1   0.005 1.142329e-02 1.142329e-02                  RNF139
## 68          1   0.005 2.135022e-02 2.135022e-02                  RNF139
## 49          1   0.005 7.995835e-03 7.995835e-03                  MAP2K1
## 60          1   0.005 1.172308e-02 1.172308e-02                  MAP2K1
## 64          1   0.005 1.615792e-02 1.615792e-02                  MAP2K1
## 81          1   0.005 3.724137e-02 3.724137e-02                  MAP2K1
## 87          1   0.005 4.485111e-02 4.485111e-02                  MAP2K1
## 53          1   0.005 1.141993e-02 1.141993e-02                    PHC2
## 58          1   0.005 1.142329e-02 1.142329e-02               BDNF, SOS1
## 63          1   0.005 1.455697e-02 1.455697e-02                    RORA
## 67          1   0.010 1.880366e-02 1.880366e-02 ARL4C, FOS, FOSB, JUN
## 69          1   0.005 2.225638e-02 2.225638e-02                    CCT8
## 78          1   0.005 3.143671e-02 3.143671e-02                    CCT8
## 72          1   0.010 2.624106e-02 2.624106e-02          PARP1, PCNA
## 75          1   0.010 3.119421e-02 3.119421e-02                    TNS4
## 79          1   0.010 3.275198e-02 3.275198e-02               JUN, SNW1
## 82          1   0.005 4.095826e-02 4.095826e-02               BDNF, FOS
## 83          1   0.005 4.110854e-02 4.110854e-02                     JUN
## 84          1   0.005 4.132894e-02 4.132894e-02          NR4A2, RORA
## 88          1   0.005 4.485111e-02 4.485111e-02     ARL6IP5, SLC38A2
## 90          1   0.005 4.891983e-02 4.891983e-02                     FOS
##            Status
## 1  Representative
## 8          Member
## 33         Member
## 2  Representative
## 3          Member
## 5          Member
## 6          Member
```

```
## 7          Member
## 9          Member
## 10         Member
## 11         Member
## 12         Member
## 13         Member
## 14         Member
## 15         Member
## 17         Member
## 19         Member
## 22         Member
## 23         Member
## 4  Representative
## 31         Member
## 16 Representative
## 21         Member
## 18 Representative
## 24         Member
## 20 Representative
## 29         Member
## 36         Member
## 25 Representative
## 28         Member
## 26 Representative
## 27 Representative
## 30 Representative
## 37         Member
## 32 Representative
## 43         Member
## 34 Representative
## 35         Member
## 39         Member
## 40         Member
## 41         Member
## 42         Member
## 46         Member
## 47         Member
## 50         Member
## 51         Member
## 52         Member
## 54         Member
## 55         Member
## 56         Member
## 57         Member
## 62         Member
## 65         Member
## 66         Member
## 70         Member
## 71         Member
## 73         Member
## 74         Member
## 76         Member
## 77         Member
## 80         Member
```

```
## 85         Member
## 86         Member
## 89         Member
## 38 Representative
## 61         Member
## 44 Representative
## 45 Representative
## 48 Representative
## 59         Member
## 68         Member
## 49 Representative
## 60         Member
## 64         Member
## 81         Member
## 87         Member
## 53 Representative
## 58 Representative
## 63 Representative
## 67 Representative
## 69 Representative
## 78         Member
## 72 Representative
## 75 Representative
## 79 Representative
## 82 Representative
## 83 Representative
## 84 Representative
## 88 Representative
## 90 Representative
```

```
## cluster enriched terms
cluster_enriched_terms(net_GO)
```



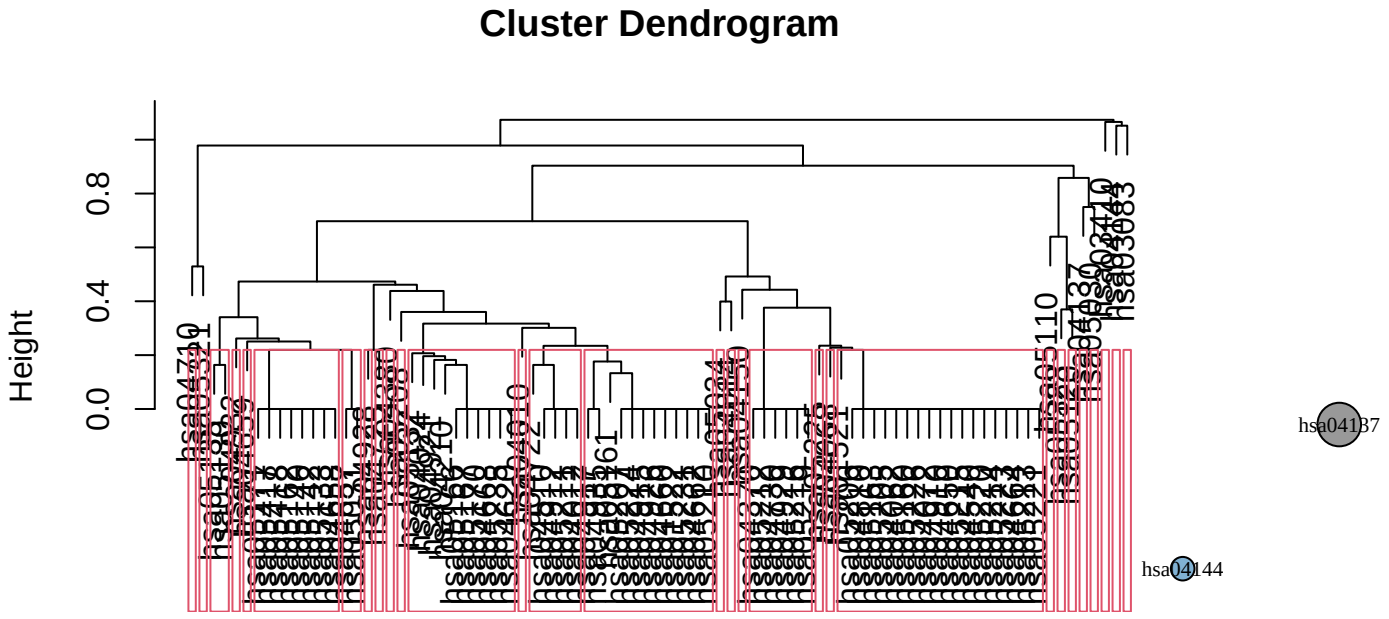
	ID	Term_Description	Fold_E
1	GO:0071276	cellular response to cadmium ion	1
3	GO:0034614	cellular response to reactive oxygen species	1
5	GO:0001228	DNA-binding transcription activator activity, RNA polymerase II-specific	1
6	GO:0090575	RNA polymerase II transcription regulator complex	1
10	GO:0061629	RNA polymerase II-specific DNA-binding transcription factor binding	1
11	GO:0006357	regulation of transcription by RNA polymerase II	1
15	GO:0000976	transcription cis-regulatory region binding	1
27	GO:0044389	ubiquitin-like protein ligase binding	1
28	GO:0070412	R-SMAD binding	1
29	GO:0060395	SMAD protein signal transduction	1
31	GO:0003700	DNA-binding transcription factor activity	1
34	GO:0001221	transcription coregulator binding	1
42	GO:0032993	protein-DNA complex	1
2	GO:0000791	euchromatin	1
8	GO:0032183	SUMO binding	1
9	GO:0071391	cellular response to estrogen stimulus	1
4	GO:0000785	chromatin	1
7	GO:0003684	damaged DNA binding	1
18	GO:0034644	cellular response to UV	1
19	GO:0043596	nuclear replication fork	4
22	GO:0003682	chromatin binding	1
12	GO:0033276	transcription factor TFTC complex	1
13	GO:0003713	transcription coactivator activity	1
14	GO:0000124	SAGA complex	1
33	GO:0051123	RNA polymerase II preinitiation complex assembly	1
36	GO:0071339	MLL1 complex	1
41	GO:0042789	mRNA transcription by RNA polymerase II	1

```
## 16 GO:0043923                                positive regulation by host of viral transcription       2
## 44 GO:0016922                                                    nuclear receptor binding
## 17 GO:0009306                                                           protein secretion
## 20 GO:0051453                                                 regulation of intracellular pH       3
## 21 GO:0046965                                          nuclear retinoid X receptor binding       2
## 23 GO:0015629                                                        actin cytoskeleton
## 24 GO:0035102                                                             PRC1 complex
## 45 GO:0031519                                                      PcG protein complex
## 25 GO:0000977 RNA polymerase II transcription regulatory region sequence-specific DNA binding
## 32 GO:0001227        DNA-binding transcription repressor activity, RNA polymerase II-specific
## 26 GO:0006413                                                    translational initiation
## 35 GO:0003743                                          translation initiation factor activity
## 30 GO:0006635                                                 fatty acid beta-oxidation
## 37 GO:0005925                                                           focal adhesion
## 38 GO:0008429                                          phosphatidylethanolamine binding       4
## 40 GO:0031625                                          ubiquitin protein ligase binding
## 39 GO:0007173                                epidermal growth factor receptor signaling pathway
## 43 GO:0007265                                                 Ras protein signal transduction
##    occurrence support     lowest_p    highest_p                        Up_regulated Down_regulated
## 1           1   0.030 1.555881e-05 1.555881e-05                             FOS, JUN
## 3           1   0.010 1.564513e-04 1.564513e-04                             FOS, JUN
## 5           1   0.020 2.000264e-03 2.000264e-03                 ATF3, FOS, JUN, NR4A2
## 6           1   0.010 2.471538e-03 2.471538e-03                  ATF3, CHD4, FOS, JUN
## 10          1   0.025 5.302986e-03 5.302986e-03                 PARP1, CHD4, FOS, JUN
## 11          1   0.025 5.548761e-03 5.548761e-03     ATF3, FOS, JUN, SOS1, TAF9, SNW1
## 15          1   0.025 7.458025e-03 7.458025e-03 ATF3, FOS, JUN, SMARCA2, TAF9, KLF11
## 27          1   0.005 2.097948e-02 2.097948e-02                                 JUN
## 28          1   0.035 2.159843e-02 2.159843e-02                             FOS, JUN
## 29          1   0.030 2.159843e-02 2.159843e-02                             FOS, JUN
## 31          1   0.005 2.269117e-02 2.269117e-02          ATF3, FOS, JUN, RORA, KLF11
## 34          1   0.015 2.684455e-02 2.684455e-02                      CHD4, FOS, RORA
## 42          1   0.020 3.886455e-02 3.886455e-02                           PARP1, FOS
## 2           1   0.015 3.898019e-05 3.898019e-05                           JUN, PELP1
## 8           1   0.005 4.632740e-03 4.632740e-03                                PELP1
## 9           1   0.005 4.818333e-03 4.818333e-03                                PELP1
## 4           1   0.010 8.499192e-04 8.499192e-04            PARP1, JUN, PCNA, SMARCA2
## 7           1   0.005 4.142410e-03 4.142410e-03                          PARP1, PCNA
## 18          1   0.005 8.462080e-03 8.462080e-03                          PARP1, PCNA
## 19          1   0.015 8.490352e-03 8.490352e-03                          PARP1, PCNA
## 22          1   0.005 1.026335e-02 1.026335e-02                 PARP1, PCNA, PELP1
## 12          1   0.010 5.621057e-03 5.621057e-03                                TAF9
## 13          1   0.005 5.640036e-03 5.640036e-03                SMARCA2, TAF9, SNW1
## 14          1   0.010 6.485453e-03 6.485453e-03                                TAF9
## 33          1   0.005 2.359913e-02 2.359913e-02                                TAF9
## 36          1   0.005 2.869420e-02 2.869420e-02                          TAF9, PELP1
## 41          1   0.005 3.461204e-02 3.461204e-02                                TAF9
## 16          1   0.025 8.399215e-03 8.399215e-03                           JUN, SNW1
## 44          1   0.005 3.886455e-02 3.886455e-02                                SNW1
## 17          1   0.005 8.399215e-03 8.399215e-03                               PDIA4
## 20          1   0.005 9.365115e-03 9.365115e-03                  SLC9A6, SLC26A6
## 21          1   0.005 9.724166e-03 9.724166e-03                               NR4A2
## 23          1   0.005 1.227182e-02 1.227182e-02                  CAPZB, AHNAK
## 24          1   0.005 1.404270e-02 1.404270e-02                               PHC2
## 45          1   0.005 4.254086e-02 4.254086e-02                               PHC2
```

```
## 25          1   0.005 1.842677e-02 1.842677e-02                  ATF3, RORA, ZBTB5
## 32          1   0.005 2.358436e-02 2.358436e-02                   ATF3, JUN, ZBTB5
## 26          1   0.005 2.006100e-02 2.006100e-02                               EIF1
## 35          1   0.005 2.684455e-02 2.684455e-02                               EIF1
## 30          1   0.005 2.166460e-02 2.166460e-02                              DECR1
## 37          1   0.005 3.001706e-02 3.001706e-02              MAP2K1, KLF11, TNS4
## 38          1   0.005 3.052565e-02 3.052565e-02                  NF1, GABARAPL2
## 40          1   0.005 3.417007e-02 3.417007e-02      JUN, NAE1, DAZAP2, GABARAPL2
## 39          1   0.005 3.060534e-02 3.060534e-02                              SOS1
## 43          1   0.005 3.886455e-02 3.886455e-02                       NF1, SOS1
##            Status
## 1  Representative
## 3          Member
## 5          Member
## 6          Member
## 10         Member
## 11         Member
## 15         Member
## 27         Member
## 28         Member
## 29         Member
## 31         Member
## 34         Member
## 42         Member
## 2  Representative
## 8          Member
## 9          Member
## 4  Representative
## 7          Member
## 18         Member
## 19         Member
## 22         Member
## 12 Representative
## 13         Member
## 14         Member
## 33         Member
## 36         Member
## 41         Member
## 16 Representative
## 44         Member
## 17 Representative
## 20 Representative
## 21 Representative
## 23 Representative
## 24 Representative
## 45         Member
## 25 Representative
## 32         Member
## 26 Representative
## 35         Member
## 30 Representative
## 37 Representative
## 38 Representative
## 40         Member
```

```
## 39 Representative
## 43         Member
```

```
## cluster enriched terms
cluster_enriched_terms(net_KEGG)
```



stats::as.dist(1 − kappa_mat2)
stats::hclust (*, "average")

```
##          ID                                                Term_Description Fold_Enrichment occurrence su
## 1  hsa04662                          B cell receptor signaling pathway       12.666293          1   0
## 2  hsa05210                                      Colorectal cancer       12.032979          1   0
## 3  hsa05231                           Choline metabolism in cancer       11.193469          1   0
## 4  hsa01522                                  Endocrine resistance       10.816161          1   0
## 5  hsa04660                          T cell receptor signaling pathway        8.444196          1   0
## 6  hsa04926                                Relaxin signaling pathway        8.227678          1   0
## 7  hsa04915                             Estrogen signaling pathway        7.763212          1   0
## 8  hsa05224                                       Breast cancer        6.875988          1   0
## 9  hsa05161                                         Hepatitis B        8.075825          1   0
## 10 hsa05207                 Chemical carcinogenesis - receptor activation        5.120416          1   0
## 14 hsa04917                              Prolactin signaling pathway       10.939072          1   0
## 22 hsa04935                 Growth hormone synthesis, secretion and action        6.747465          1   0
## 11 hsa05208               Chemical carcinogenesis - reactive oxygen species        7.481645          1   0
## 12 hsa05211                                      Renal cell carcinoma       11.644818          1   0
## 15 hsa04012                                    ErbB signaling pathway        9.890119          1   0
## 18 hsa04912                                    GnRH signaling pathway        8.594985          1   0
## 23 hsa04722                          Neurotrophin signaling pathway        8.751257          1   0
## 31 hsa04510                                       Focal adhesion        3.881606          1   0
## 13 hsa04380                                Osteoclast differentiation        9.704015          1   0
## 16 hsa05235           PD-L1 expression and PD-1 checkpoint pathway in cancer        8.804619          1   0
## 19 hsa04620                          Toll-like receptor signaling pathway        7.443080          1   0
## 20 hsa04668                                       TNF signaling pathway        7.148304          1   0
```

```
## 24 hsa04210                                             Apoptosis        7.890478          1   0
## 26 hsa04921                                Oxytocin signaling pathway    7.520612          1   0
## 28 hsa05170                      Human immunodeficiency virus 1 infection 4.078976          1   0
## 29 hsa05167           Kaposi sarcoma-associated herpesvirus infection    4.078976          1   0
## 32 hsa05166                      Human T-cell leukemia virus 1 infection 3.609894          1   0
## 34 hsa04024                                   cAMP signaling pathway    4.673001          1   0
## 37 hsa05132                                    Salmonella infection     4.336209          1   0
## 17 hsa04010                                   MAPK signaling pathway    5.469536          1   0
## 21 hsa04928       Parathyroid hormone synthesis, secretion and action   6.875988          1   0
## 25 hsa05135                                      Yersinia infection     7.701106          1   0
## 54 hsa05130                    Pathogenic Escherichia coli infection    4.197551          1   0
## 27 hsa04137                                   Mitophagy - animal        8.112120          1   0
## 30 hsa05133                                           Pertussis         6.779143          1   0
## 36 hsa04658                        Th1 and Th2 cell differentiation     5.662578          1   0
## 39 hsa05142                                        Chagas disease       4.962053          1   0
## 44 hsa05140                                       Leishmaniasis         7.078223          1   0
## 47 hsa05162                                            Measles          3.850553          1   0
## 50 hsa05418                   Fluid shear stress and atherosclerosis    3.646357          1   0
## 56 hsa05417                               Lipid and atherosclerosis     2.506871          1   0
## 60 hsa05171                          Coronavirus disease - COVID-19     2.187814          1   0
## 33 hsa05323                                   Rheumatoid arthritis     11.884423         1   0
## 35 hsa04657                                IL-17 signaling pathway      8.594985          1   0
## 58 hsa05031                                  Amphetamine addiction     11.644818         1   0
## 38 hsa05213                                     Endometrial cancer     8.913318          1   0
## 41 hsa05221                                  Acute myeloid leukemia    8.021986          1   0
## 42 hsa04664                        Fc epsilon RI signaling pathway     7.520612          1   0
## 43 hsa05223                             Non-small cell lung cancer     7.292714          1   0
## 45 hsa05214                                            Glioma           6.875988          1   0
## 46 hsa01521                  EGFR tyrosine kinase inhibitor resistance 10.313982         1   0
## 48 hsa05220                               Chronic myeloid leukemia     6.684988          1   0
## 49 hsa04540                                        Gap junction        6.504313          1   0
## 53 hsa05215                                      Prostate cancer       5.289221          1   0
## 57 hsa04650                  Natural killer cell mediated cytotoxicity 4.336209          1   0
## 61 hsa04910                               Insulin signaling pathway    3.731156          1   0
## 63 hsa04072                       Phospholipase D signaling pathway    3.513278          1   0
## 65 hsa05226                                      Gastric cancer        3.413611          1   0
## 66 hsa05160                                        Hepatitis C         3.365868          1   0
## 70 hsa04062                          Chemokine signaling pathway       2.766202          1   0
## 71 hsa05205                             Proteoglycans in cancer        2.533259          1   0
## 72 hsa05163                        Human cytomegalovirus infection     2.418689          1   0
## 76 hsa04810                        Regulation of actin cytoskeleton    2.228329          1   0
## 79 hsa05206                                    MicroRNAs in cancer     2.111049          1   0
## 40 hsa04659                              Th17 cell differentiation     7.148304          1   0
## 51 hsa04932                      Non-alcoholic fatty liver disease     5.084357          1   0
## 52 hsa04725                                  Cholinergic synapse       5.596734          1   0
## 55 hsa05034                                        Alcoholism          9.531072          1   0
## 59 hsa04068                               FoxO signaling pathway       5.966766          1   0
## 62 hsa04710                                     Circadian rhythm       8.021986          1   0
## 64 hsa04150                               mTOR signaling pathway       6.975640          1   0
## 67 hsa05225                             Hepatocellular carcinoma       4.598591          1   0
## 68 hsa04144                                         Endocytosis        2.177915          1   0
## 69 hsa05110                              Vibrio cholerae infection     11.459980         1   0
## 73 hsa05030                                      Cocaine addiction     15.361249         1   0
## 74 hsa04014                                  Ras signaling pathway      5.596734          1   0
```

```
## 75 hsa05216                                                Thyroid cancer        7.078223         1   0
## 80 hsa05219                                                Bladder cancer        6.333147         1   0
## 82 hsa04929                                                 GnRH secretion        5.596734         1   0
## 83 hsa04730                                          Long-term depression        4.718815         1   0
## 84 hsa05212                                             Pancreatic cancer        3.437994         1   0
## 85 hsa04370                                        VEGF signaling pathway        4.628069         1   0
## 77 hsa03410                                         Base excision repair       11.459980         1   0
## 78 hsa05120 Epithelial cell signaling in Helicobacter pylori infection        7.183868         1   0
## 81 hsa05321                                  Inflammatory bowel disease       12.447909         1   0
## 86 hsa03083                                 Polycomb repressive complex        3.342494         1   0
##       highest_p                          Up_regulated Down_regulated Cluster         Status
## 1  6.905773e-09                 FOS, JUN, MAP2K1, SOS1                     1 Representative
## 2  8.982017e-09                 MAP2K1, JUN, FOS, SOS1                     1         Member
## 3  1.300342e-08                 SOS1, MAP2K1, FOS, JUN                     1         Member
## 4  1.549294e-08                 MAP2K1, SOS1, JUN, FOS                     1         Member
## 5  5.464336e-08                 SOS1, FOS, JUN, MAP2K1                     1         Member
## 6  6.234488e-08                 MAP2K1, SOS1, FOS, JUN                     1         Member
## 7  8.371276e-08                 FOS, JUN, MAP2K1, SOS1                     1         Member
## 8  1.547563e-07                 MAP2K1, SOS1, FOS, JUN                     1         Member
## 9  2.120405e-07           MAP2K1, FOS, JUN, PCNA, SOS1                     1         Member
## 10 6.847810e-07                 MAP2K1, SOS1, FOS, JUN                     1         Member
## 14 1.521920e-06                      MAP2K1, SOS1, FOS                     1         Member
## 22 1.083429e-05                      MAP2K1, FOS, SOS1                     1         Member
## 11 7.814757e-07 SOS1, MAP2K1, NDUFS4, SLC26A6, FOS, JUN                     2 Representative
## 12 1.178650e-06                      SOS1, JUN, MAP2K1                     3 Representative
## 15 2.296120e-06                      SOS1, JUN, MAP2K1                     3         Member
## 18 4.064075e-06                      MAP2K1, SOS1, JUN                     3         Member
## 23 1.211518e-05                 BDNF, MAP2K1, SOS1, JUN                     3         Member
## 31 1.002226e-04                      MAP2K1, SOS1, JUN                     3         Member
## 13 1.421425e-06           MAP2K1, FOS, FOSB, JUN, ACP5                     4 Representative
## 16 3.685060e-06                        FOS, JUN, MAP2K1                     5 Representative
## 19 7.284486e-06                        MAP2K1, FOS, JUN                     5         Member
## 20 8.579078e-06                        MAP2K1, FOS, JUN                     5         Member
## 24 1.840080e-05                 PARP1, JUN, FOS, MAP2K1                     5         Member
## 26 2.232993e-05                 MAP2K1, RGS2, JUN, FOS                     5         Member
## 28 8.215644e-05                      MAP2K1, FOS, JUN                     5         Member
## 29 8.215644e-05                      MAP2K1, FOS, JUN                     5         Member
## 32 1.340175e-04                      JUN, FOS, MAP2K1                     5         Member
## 34 1.508424e-04                 MAP2K1, BDNF, FOS, JUN                     5         Member
## 37 2.034223e-04                MAP2K1, FOS, JUN, AHNAK                     5         Member
## 17 3.765718e-06       MAP2K1, NF1, SOS1, FOS, JUN, BDNF                     6 Representative
## 21 1.003856e-05                    MAP2K1, FOS, NR4A2                     7 Representative
## 25 2.029434e-05                WIPF2, FOS, JUN, MAP2K1                     8 Representative
## 54 1.418065e-03                        WIPF2, FOS, JUN                     8         Member
## 27 7.794969e-05                 HUWE1, JUN, GABARAPL2                     9 Representative
## 30 9.791464e-05                              FOS, JUN                    10 Representative
## 36 1.690497e-04                              JUN, FOS                    10         Member
## 39 2.521494e-04                              FOS, JUN                    10         Member
## 44 4.777869e-04                              FOS, JUN                    10         Member
## 47 5.423993e-04                              FOS, JUN                    10         Member
## 50 6.392467e-04                              FOS, JUN                    10         Member
## 56 1.973474e-03                              FOS, JUN                    10         Member
## 60 2.969307e-03                              FOS, JUN                    10         Member
## 33 1.460682e-04               ATP6V1B2, ACP5, FOS, JUN                    11 Representative
```

```
## 35 1.630940e-04                    FOS, FOSB, JUN 12 Representative
## 58 2.318603e-03                    FOS, FOSB, JUN 12         Member
## 38 2.370222e-04                      SOS1, MAP2K1 13 Representative
## 41 3.266741e-04                      SOS1, MAP2K1 13         Member
## 42 3.974727e-04                      MAP2K1, SOS1 13         Member
## 43 4.364034e-04                      SOS1, MAP2K1 13         Member
## 45 5.216970e-04                      SOS1, MAP2K1 13         Member
## 46 5.216970e-04                 SOS1, MAP2K1, NF1 13         Member
## 48 5.682075e-04                      SOS1, MAP2K1 13         Member
## 49 6.173919e-04                      MAP2K1, SOS1 13         Member
## 53 1.153775e-03                      SOS1, MAP2K1 13         Member
## 57 2.099645e-03                      MAP2K1, SOS1 13         Member
## 61 3.298431e-03                      SOS1, MAP2K1 13         Member
## 63 3.951112e-03                      MAP2K1, SOS1 13         Member
## 65 4.307240e-03                      MAP2K1, SOS1 13         Member
## 66 4.492997e-03                      SOS1, MAP2K1 13         Member
## 70 8.083059e-03                      MAP2K1, SOS1 13         Member
## 71 1.051158e-02                      MAP2K1, SOS1 13         Member
## 72 1.206780e-02                      SOS1, MAP2K1 13         Member
## 76 1.540705e-02                      MAP2K1, SOS1 13         Member
## 79 1.809731e-02                      MAP2K1, SOS1 13         Member
## 40 2.849262e-04                   JUN, FOS, RORA 14 Representative
## 51 7.965741e-04                  FOS, JUN, NDUFS4 15 Representative
## 52 9.727579e-04                       MAP2K1, FOS 16 Representative
## 55 1.580091e-03          BDNF, MAP2K1, FOSB, SOS1 17 Representative
## 59 2.721418e-03           SOS1, MAP2K1, GABARAPL2 18 Representative
## 62 3.381028e-03                              RORA 19 Representative
## 64 4.038243e-03       MAP2K1, RRAGC, SOS1, ATP6V1B2 20 Representative
## 67 5.943426e-03              MAP2K1, SOS1, SMARCA2 21 Representative
## 68 5.976460e-03                       CAPZB, WIPF2 22 Representative
## 69 6.687367e-03                   ATP6V1B2, PDIA4 23 Representative
## 73 1.397083e-02                   JUN, FOSB, BDNF 24 Representative
## 74 1.519560e-02       MAP2K1, SOS1, BDNF, NF1, SHOC2 25 Representative
## 75 1.521454e-02                            MAP2K1 26 Representative
## 80 2.537892e-02                            MAP2K1 26         Member
## 82 3.257027e-02                            MAP2K1 26         Member
## 83 4.592287e-02                            MAP2K1 26         Member
## 84 4.662832e-02                            MAP2K1 26         Member
## 85 4.775134e-02                            MAP2K1 26         Member
## 77 1.667903e-02                      PCNA, PARP1 27 Representative
## 78 1.714746e-02                     JUN, ATP6V1B2 28 Representative
## 81 3.196102e-02                JUN, IL18RAP, RORA 29 Representative
## 86 4.933907e-02                              PHC2 30 Representative
```