Assignment-1 Accession number:GSE48556

Osteoarthritis: peripheral blood mononuclear cells
Analysis of peripheral blood mononuclear cells (PBMCs) from osteoarthritis(OA) patients .
Results provide insight into the feasibility of using gene expression profiling of PBMCs to detect the onset of osteoarthritis.

1) Download data:

- **Firstly, we have used three packages : GEOquerry , limma ,umap, ggplot2**

```
library(GEOquery)
library(ggplot2)
library(limma)
library(umap)
```

- **This code downloads the dataset from the GEO database with the accession number "GSE48556".**
  GSEMatrix=True indicates that we want to download the expression matrix of the dataset.
  AnnotGPL: True indicates that we want annotation information
  Next it checks if the dataset contains more than one dataset and extracts the dataset that corresponds to "GPL6947"
  If there is only one dataset then idx variable is set to 1

```
gset <- getGEO("GSE48556", GSEMatrix =TRUE, AnnotGPL=TRUE)
if (length(gset) > 1) idx <- grep("GPL6947", attr(gset, "names")) else idx <- 1
gset <- gset[[idx]]
```

2)EDA and preprocessing and pdata and fdata:

```
# process background correction and normalize the dataset.

exprs <-backgroundCorrect(exprs, method="normexp", offset=1)
exprs(gset) <- normalizeBetweenArrays(exprs(gset))
```

 # pdata and fdata and attributes.

```
pdata <- pData(gset)
fdata <- fData(gset)
sampleNames(gset)
Meta(gset)
```

- pData(gset) -> to access the pdata of the dataset.
- fData (gset) -> to acces the fdata of the dataset.
- sampleNames(gset) -> gives the colNames
- Meta(gset) -> to get the metadata and attributes information

 # EDA and Assigning groups and preprocessing the data

```
fvarLabels(gset) <- make.names(fvarLabels(gset))

# group membership for all samples
gsms <- paste0("00000000000000000000000000000000000000000000000000",
               "00000000000000000000000000000000000000000000000000",
               "00000011111111111111111111111111111111")
sml <- strsplit(gsms, split="")[[1]]

ex <- exprs(gset)
ex[which(ex <= 0)] <- NaN
# log2 transformation
exprs(gset) <- log2(ex)

exprs(gset) <- normalizeBetweenArrays(exprs(gset)) # normalize data

# assign samples to groups and set up design matrix
gs <- factor(sml)

groups <- make.names(c("osteoarthritis","control"))
levels(gs) <- groups
gset$group <- gs
design <- model.matrix(~group + 0, gset)
colnames(design) <- levels(gs)

nall <- nrow(gset)
gset <- gset[complete.cases(exprs(gset)), ]
```

3) log 2 transformation:

```
exprs(gset) <- log2(ex)
```

#The above code will do the log transformation of the data.

We can the effect on our dataset by plotting the boxplot before and after the transformation.
Usng the code. One can observe

```
boxplot(exprs(gset), main = "Original Data", xlab = "Samples", ylab = "Expression values")
```

Effects of log transformation:

- It is done to normalize the data and to reduce the impact of outliers and extreme values of our dataset.
- It also reduces the dynamic range of values and the distribution is made more symmetrical.

4) DEA,t -test holm correction and volcano plot.

 # the first line performs t test between the two groups osteoarthritis and control

```
t_test <- apply(gset, 1, function(x) t.test(x[gs=="osteoarthritis"], x[gs=="control"]))
```

 # it extracts p value from each t test using sapply function

```
p_value <- sapply(t_test, function(x) x$p.value)
```

 # performs log_fold_change

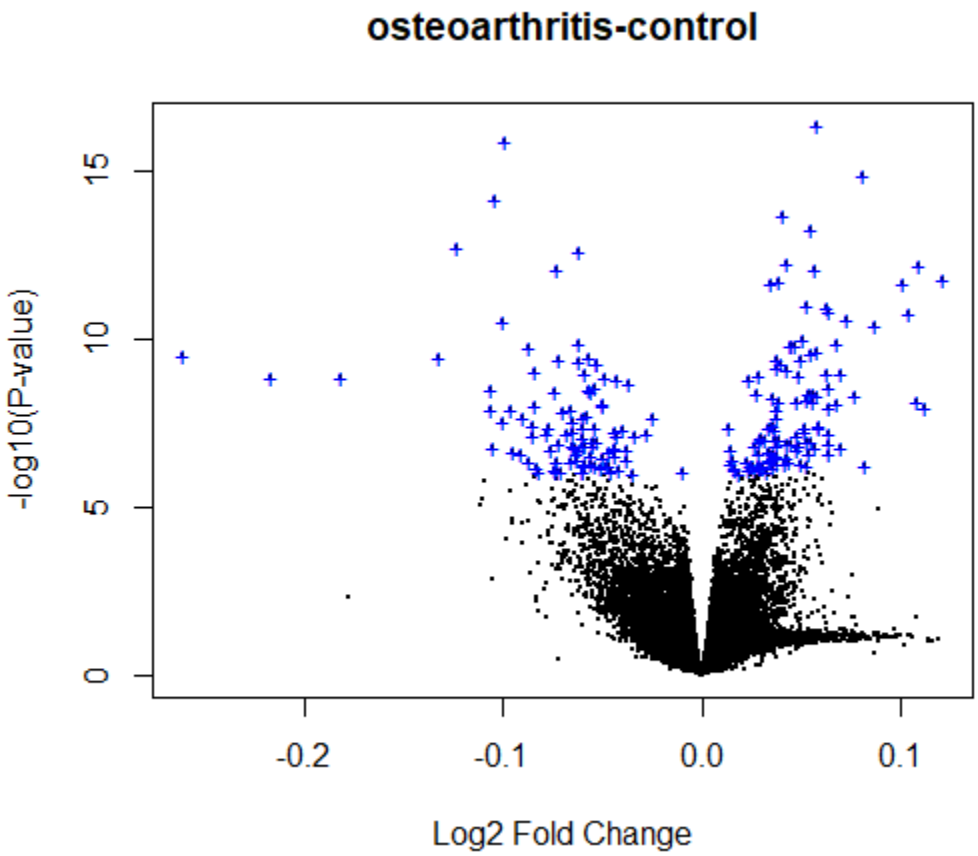
```
log_FoldChange<-apply(gset,1,function(x)log2(mean(x[gs=="osteoarthritis"])/mean(x[gs=="control"])))
```

 # using Holm correction we correct the p-value

```
holm_correction <- p.adjust(p_value, method = "holm")
```

 # Create a volcano plot to visualize the results

```
plot(log_FoldChange, -log10(holm_correction), pch=20, main="volcano_plot", xlab="log 2 fold change", ylab="-log10(pvalue)",
    xlim=c(-6,6), ylim=c(0,10), col=ifelse(abs(log_FoldChange) > 2 & holm_correction < 0.05, "red", "black"))
abline(h=-log10(0.05), col="blue", lty=2)
abline(v=c(-2,2), col="blue", lty=2)
```



5)DEA using limma package and volcano plot.

Volcano_plot: It is used to visualize the results of differentially gene expression analysis. The plot displays the statistical significance on the y -axis and the fold change on the x-axis.
It can identify the genes that are significantly differentially expressed and have large fold changes.

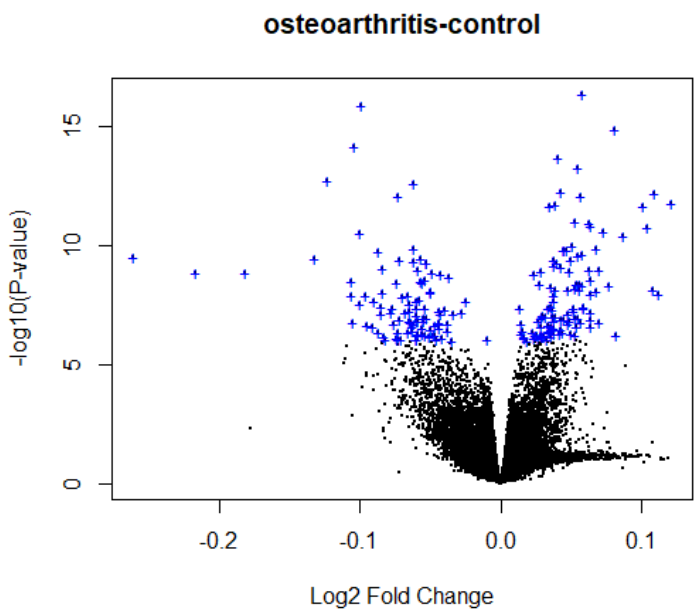
```
# using the limma package
fit  <- lmFit(gset, design)

#recalculate model coefficients to fit the linear model to pre processed expression above and recalculate the coeffecients
vector1 <- c(paste(groups[1],"-",groups[2],sep=""))
cont.matrix <- makeContrasts(contrasts=vector1, levels=design)
new_cordinate <- contrasts.fit(fit, cont.matrix)
```

#computation of top significant values and Empirical Bayes moderation is applied to variable and improve the estimation of variance

```
new_cordinate <- eBayes(new_cordinate, 0.01)
new_cordinate <- eBayes(new_cordinate, 0.01)
top_table <- topTable(new_cordinate, adjust="holm", sort.by="B", number=250)
dT <- decideTests(new_cordinate, adjust.method="holm", p.value=0.05)
colnames(new_cordinate) # list contrast names
```

#Ploting the volcano plot

```
volcanoplot(new_cordinate, coef=1, main=colnames(new_cordinate)[1], pch=20,
        highlight=length(which(dT[1]!=0)), names=rep('+', nrow(new_cordinate)))
```



6) Choose a significant cutoff based on log(FC) and p-values and justify why you chose those values as the cutoff.

The cutoff value of **p-value =0.05** which indicates that there is a 5 percent change of obtaining a result as extreme as the one observed assuming the null hypothesis is true which is that we assume there is no difference between them.P-value less than 0.05 indicates that there is a significant difference between the two groups being compared.

The cutoff value of **logFC is -1 and 1 i.e -1=log(2) and 1= log(0.5)**
logFC is the measure of difference in the expression levels of genres.

logFC =-1 indicates that the expression of the gene is downregulated by 2-fold or the expression of the gene is upregulated by 2-fold . Generally , a logFC of 1 or -1 is the threshold for identifying DEGs in a study

7) Perform Enrichment analysis using the set of genes that you have obtained using the Gene set enrichment analysis method.

#First install the package clusterProfiler , org.Hs.eg.db and enrichplot using Biomanager

```
library(clusterProfiler)
library(enrichplot)
library(org.Hs.eg.db)
```

 # The differentially expressed genes are expressed in expressed_genes and by using bitr() function which convert the gene symbols in expressed_genes to Entrez _ds which are unique identifiers used in the enrichGO() function.

```
expressed_genes <- c(top_table$"Gene.symbol")
genes_mapped <- bitr(expressed_genes, fromType="SYMBOL", toType="ENTREZID", OrgDb=org.Hs.eg.db)
```

 # perform gene set enrichment analysis using enrichGO

```
ans <- enrichGO(
  gene=genes_mapped$ENTREZID,
  keyType = "ENTREZID",
  OrgDb=org.Hs.eg.db,
  ont="ALL",
  pvalueCutoff=0.05,
  qvalueCutoff=0.05,
  minGSSize=5,
  maxGSSize=500,

)
```

8) Explain the meaning of different parameters in your Gene set enrichment analysis code. Show the results of enrichment in various plots and make observations.

 # perform gene set enrichment analysis using enrichGO

```
ans <- enrichGO(
  gene=genes_mapped$ENTREZID,
  keyType = "ENTREZID",
  OrgDb=org.Hs.eg.db,
  ont="ALL",
  pvalueCutoff=0.05,
  qvalueCutoff=0.05,
```

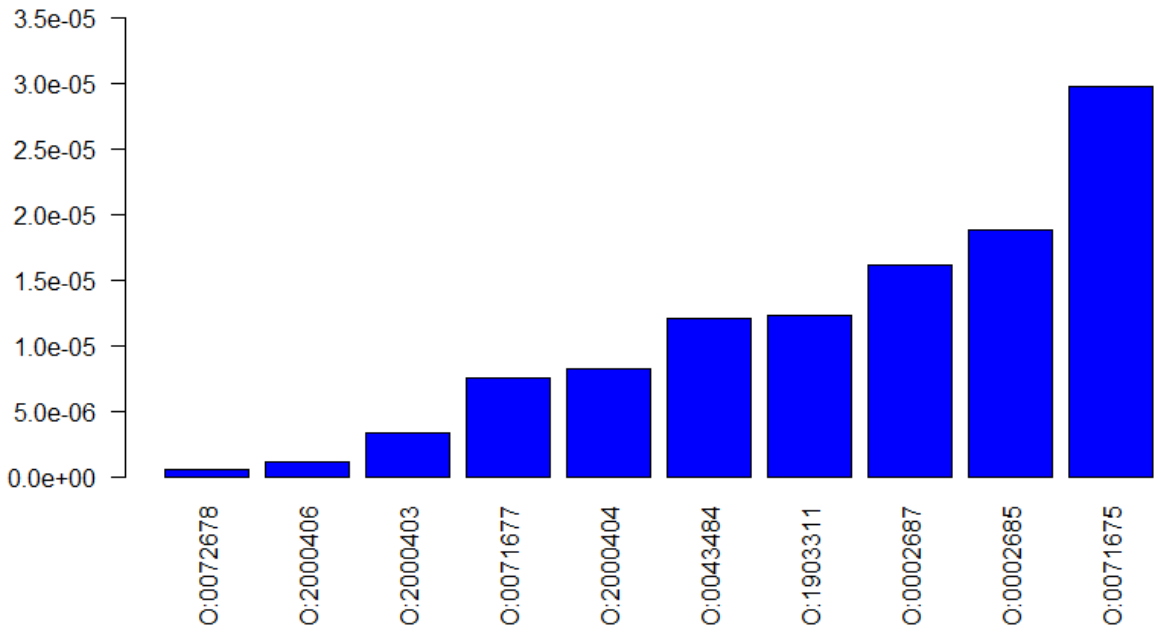
```
  minGSSize=5,
  maxGSSize=500,
)
```

Explanation of arguments:

gene=a character vector of gene IDs , it is set to genes_mapped$ENTREZID which is a vector of Entrez IDs of the expressed genes.
keyType: Is is set to ENTREZID to indicate that the input vector contains Entrez IDs
OrgDb: to signify the annotated package org.Hs.eg.db
Ont: the type of geneOntology which is used like BP (biological process) , " MF" (molecular function)
It is set to "ALL" to perform enrichment analysis for all available ontologies.
pvalueCutoff: the maximum adjusted p -value which is 0.05
qvalueCutoff : adjusted p value which is set to 0.05
minGSize: the minimum gene set allowed in the gene set which is set to 10
maxGSize: the maximum gene set allowed in the gene set which is set to 500

9)Observe and analyze the pathways which you obtained and make observations.
#Bar Graph

```
most_significant_10 <- ans[1:10]$pvalue
barplot(most_significant_10,names.arg =ans[1:10]$ID, las = 2, col = "blue", ylim = c(0,
max(most_significant_10)*1.2))
```



Observation_bar_plot: Bar plot is the most widely used method to visualize enriched terms. It depicts the enrichment scores (*e.g.* p values) and gene count or ratio as bar height and color

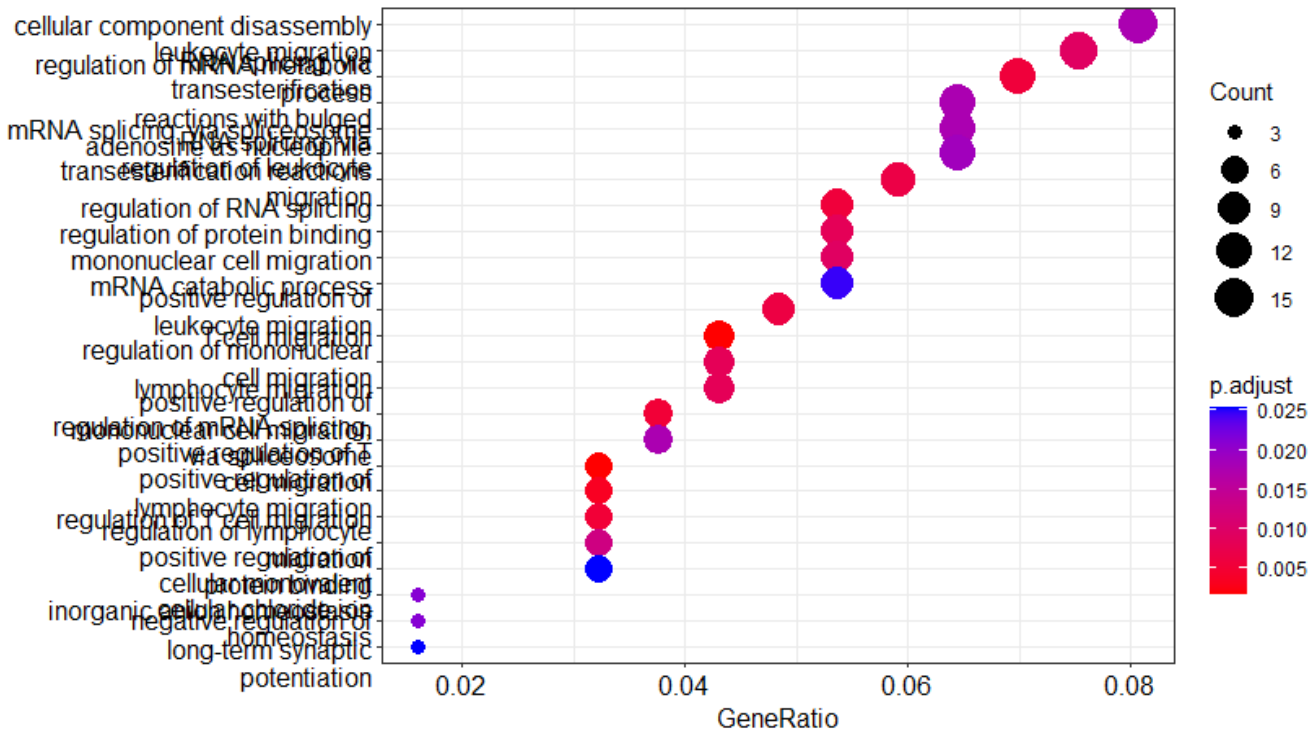
The 10 most significant expressed genes are taken from the ans which is the result of the enrichment test . In the plot we can see the gene IDs of 10 significantly expressed genes in x axis.

The higher the bar , the more significant the gene set enrichments.

The ylim argument set the maximum value of the y-axis to 1.2 the max p value , ensuring that all bars are visible.

 # dot-plot

dotplot(ans, showCategory = 25)



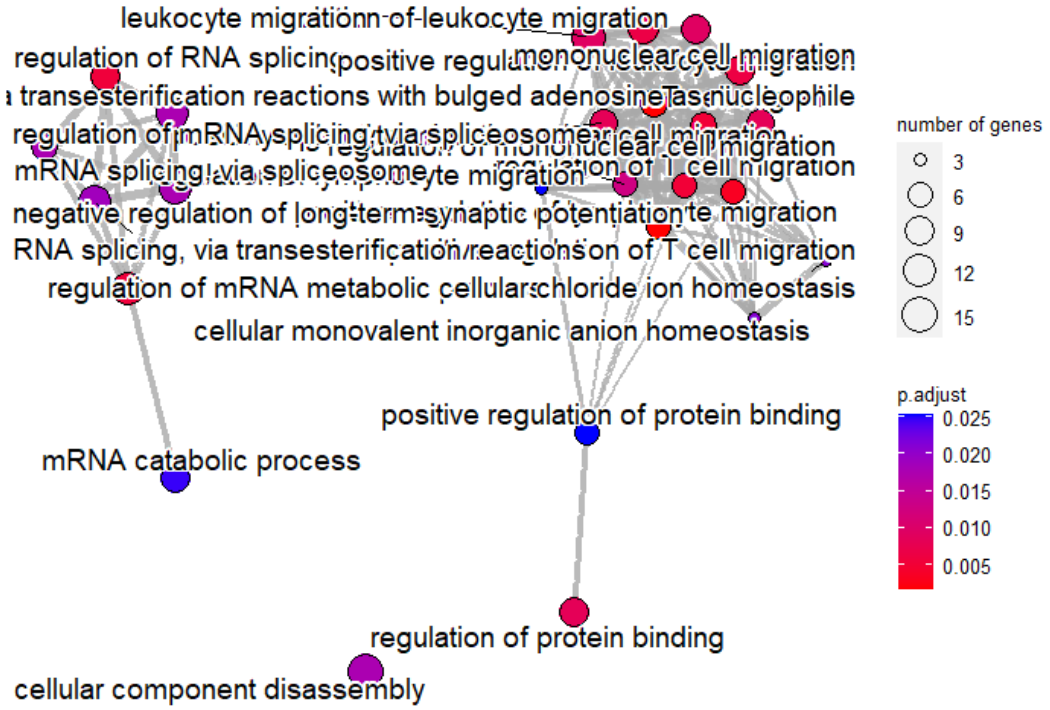
Observation_dotplot(): Dot plot is similar to bar plot with the capability to encode another score as dot size. Here , the x -axis represents the geneRatio.Each dot represents a gene set and the position on y -axis represents the enrichment score which represents how significantly the gene is expressed.

The showCategory =25 represents the limit of number of categories.

 # enrichment map plot

pairwise_termsim_res <- pairwise_termsim(ans)

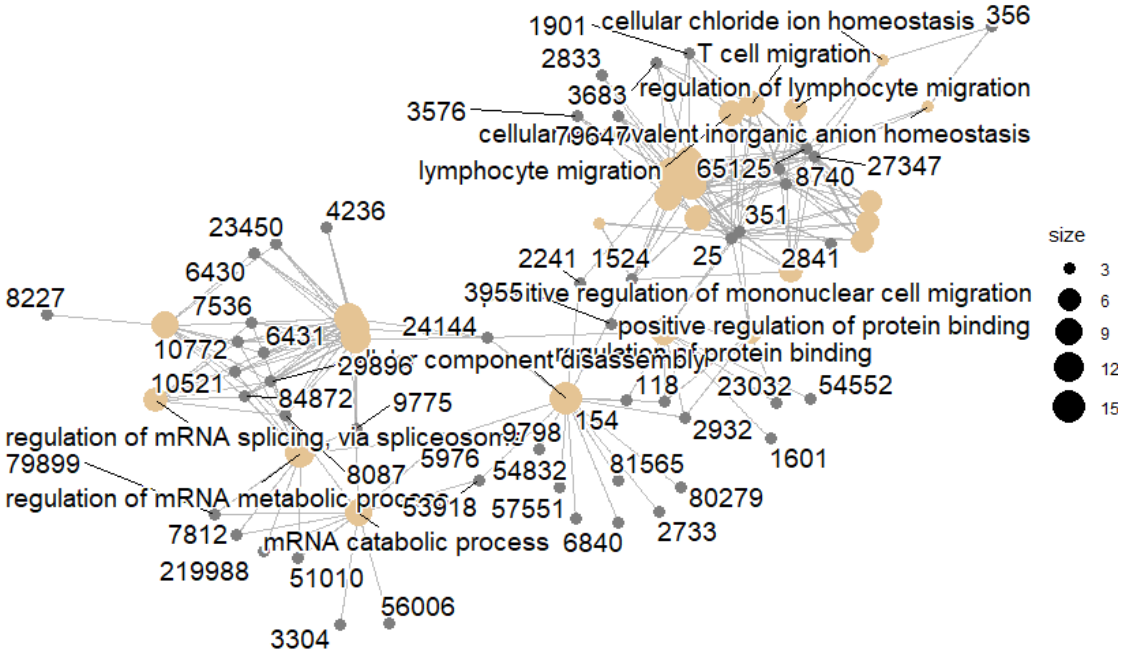
emapplot(pairwise_termsim_res, enrich=res, showCategory=25)



observation_enrichment_plot: Enrichment map organizes enriched terms into a network with edges connecting overlapping gene sets. In this way, mutually overlapping gene sets are tend to cluster together, making it easy to identify functional modules.Enrichment map plots are useful for identifying clusters of gene sets that share similar biological functions or pathways. These clusters are indicated by groups of nodes that are closely connected by thick edges.

 # category netplot

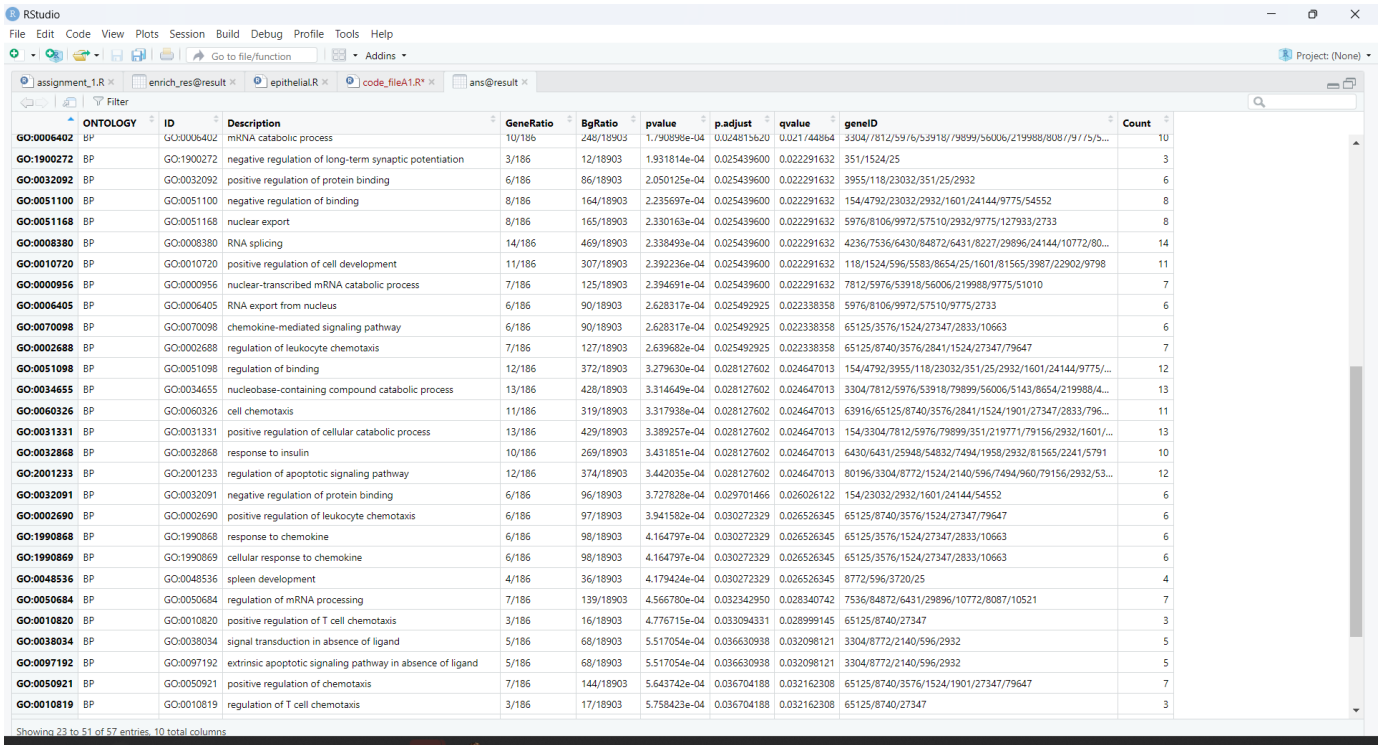
cnetplot(ans, foldChange = de_genes_ranked, showCategory = 25)



Observation_category netplot: The cnetplot depicts the linkages of genes and biological concepts (e.g. GO terms or KEGG pathways) as a network (helpful to see which genes are involved in enriched pathways and genes that may belong to multiple annotation categories).
The above plot shows a network of categories with clusters representing individual categories and edges representing the relationships between them.
The size of each node is proportional to the number of genes in the category which the color indicates the level of the enrichment with darker color represents lower p-value meaning high enrichment score.

 # To view different pathways in gene enrichment analysis

```
View(ans@result)
```



	ONTOLOGY	ID	Description	GeneRatio	BgRatio	pvalue	p.adjust	qvalue	geneID	Count
GO:0006402	BP	GO:0006402	mRNA catabolic process	10/186	248/18903	1.790898e-04	0.024815620	0.021744864	3304/7812/5976/53918/79899/56006/219988/8087/9775/5...	10
GO:1900272	BP	GO:1900272	negative regulation of long-term synaptic potentiation	3/186	12/18903	1.931814e-04	0.025439600	0.022291632	351/1524/25	3
GO:0032092	BP	GO:0032092	positive regulation of protein binding	6/186	86/18903	2.050125e-04	0.025439600	0.022291632	3955/118/23032/351/25/2932	6
GO:0051100	BP	GO:0051100	negative regulation of binding	8/186	164/18903	2.235697e-04	0.025439600	0.022291632	154/4792/23032/2932/1601/24144/9775/54552	8
GO:0051168	BP	GO:0051168	nuclear export	8/186	165/18903	2.330163e-04	0.025439600	0.022291632	5976/8106/9972/57510/2932/9775/127933/2733	8
GO:0008380	BP	GO:0008380	RNA splicing	14/186	469/18903	2.338493e-04	0.025439600	0.022291632	4236/7536/6430/84872/6431/8227/29896/24144/10772/80...	14
GO:0010720	BP	GO:0010720	positive regulation of cell development	11/186	307/18903	2.392236e-04	0.025439600	0.022291632	118/1524/596/5583/8654/25/1601/81565/3987/22902/9798	11
GO:0000956	BP	GO:0000956	nuclear-transcribed mRNA catabolic process	7/186	125/18903	2.394691e-04	0.025439600	0.022291632	7812/5976/53918/56006/219988/9775/51010	7
GO:0006405	BP	GO:0006405	RNA export from nucleus	6/186	90/18903	2.628317e-04	0.025492925	0.022338358	5976/8106/9972/57510/9775/2733	6
GO:0070098	BP	GO:0070098	chemokine-mediated signaling pathway	6/186	90/18903	2.628317e-04	0.025492925	0.022338358	65125/3576/1524/27347/2833/10663	6
GO:0002688	BP	GO:0002688	regulation of leukocyte chemotaxis	7/186	127/18903	2.639682e-04	0.025492925	0.022338358	65125/8740/3576/2841/1524/27347/79647	7
GO:0051098	BP	GO:0051098	regulation of binding	12/186	372/18903	3.279630e-04	0.028127602	0.024647013	154/4792/3955/118/23032/351/25/2932/1601/24144/9775/...	12
GO:0034655	BP	GO:0034655	nucleobase-containing compound catabolic process	13/186	428/18903	3.314649e-04	0.028127602	0.024647013	3304/7812/5976/53918/79899/56006/5143/8654/219988/4...	13
GO:0060326	BP	GO:0060326	cell chemotaxis	11/186	319/18903	3.317938e-04	0.028127602	0.024647013	63916/65125/8740/3576/2841/1524/1901/27347/2833/796...	11
GO:0031331	BP	GO:0031331	positive regulation of cellular catabolic process	13/186	429/18903	3.389257e-04	0.028127602	0.024647013	154/3304/7812/5976/79899/351/219771/79156/2932/1601/...	13
GO:0032868	BP	GO:0032868	response to insulin	10/186	269/18903	3.431851e-04	0.028127602	0.024647013	6430/6431/25948/54832/7494/1958/2932/81565/2241/5791	10
GO:2001233	BP	GO:2001233	regulation of apoptotic signaling pathway	12/186	374/18903	3.442035e-04	0.028127602	0.024647013	80196/3304/8772/1524/2140/596/7494/960/79156/2932/53...	12
GO:0032091	BP	GO:0032091	negative regulation of protein binding	6/186	96/18903	3.727828e-04	0.029701466	0.026026122	154/23032/2932/1601/24144/54552	6
GO:0002690	BP	GO:0002690	positive regulation of leukocyte chemotaxis	6/186	97/18903	3.941582e-04	0.030272329	0.026526345	65125/8740/3576/1524/27347/79647	6
GO:1990868	BP	GO:1990868	response to chemokine	6/186	98/18903	4.164797e-04	0.030272329	0.026526345	65125/3576/1524/27347/2833/10663	6
GO:1990869	BP	GO:1990869	cellular response to chemokine	6/186	98/18903	4.164797e-04	0.030272329	0.026526345	65125/3576/1524/27347/2833/10663	6
GO:0048536	BP	GO:0048536	spleen development	4/186	36/18903	4.179424e-04	0.030272329	0.026526345	8772/596/3720/25	4
GO:0050684	BP	GO:0050684	regulation of mRNA processing	7/186	139/18903	4.566780e-04	0.032342950	0.028340742	7536/84872/6431/29896/10772/8087/10521	7
GO:0010820	BP	GO:0010820	positive regulation of T cell chemotaxis	3/186	16/18903	4.776715e-04	0.033094331	0.028999145	65125/8740/27347	3
GO:0038034	BP	GO:0038034	signal transduction in absence of ligand	5/186	68/18903	5.517054e-04	0.036630938	0.032098121	3304/8772/2140/596/2932	5
GO:0097192	BP	GO:0097192	extrinsic apoptotic signaling pathway in absence of ligand	5/186	68/18903	5.517054e-04	0.036630938	0.032098121	3304/8772/2140/596/2932	5
GO:0050921	BP	GO:0050921	positive regulation of chemotaxis	7/186	144/18903	5.643742e-04	0.036704188	0.032162308	65125/8740/3576/1524/1901/27347/79647	7
GO:0010819	BP	GO:0010819	regulation of T cell chemotaxis	3/186	17/18903	5.758423e-04	0.036704188	0.032162308	65125/8740/27347	3

Observations: There are 57 entries in the gene enrichment test. Here we can see the geneRation , geneID and count of genes in the pathway. These 57 geneIDs passed the significance cutoffs ans size criteria in the gene expression analysis.
The p -value and q-value are also present which represent the enrichment of a gene in a particular biological pathways. It also represents the Ontogoly which is in this case is BP : biological process