RNAseq using Deseq2 and Functional enrichment Analysis

2022-04-16

##RNAseq using Deseq2 and Functional enrichment Analysis
Dr. Amarinder Singh Thind and Simarpreet Kaur
Date : 18-19 April, 2022

```
##### Install packages, if not done before

# if (!requireNamespace("BiocManager", quietly = TRUE)) install.packages("BiocManager")
# BiocManager::install("DESeq2")
# BiocManager::install("biomaRt")
# BiocManager::install('PCAtools')
# BiocManager::install('EnhancedVolcano')
```

```
###################### load the raw count matrix ######################

#setwd("/Users/athind/Dropbox/RNAseq_using_DEseq2-april16/") #Path_to_working_directory

setwd("./")
rawcount<-read.table ("RawCount_input.csv",header=TRUE,  sep=",",  row.names=1)

## Replace NAs by zero and
rawcount <- round(rawcount)
rawcount[is.na(rawcount)] <- 0
```

```
###################### Data annotation  ##################################

anno <-read.table ("Annotation_of_samples_12_Samples_ALL.csv",header=TRUE,  sep=",", row.name
s = 1) ##In this case we have 3 coulmns (a) sample (b) Condition (c) batch
#rownames(anno) <- anno$sample  ##add rownames as sample name (if not already), because pca f
unction check rownames of anno == col of data matrix

#####################################################################
```

PCA plot for pre DE investigation

```
############ PCA plot for pre DE investigation ##############


library(PCAtools)
```

```
## Warning: package 'PCAtools' was built under R version 4.0.3
```

```
## Loading required package: ggplot2
```

```
## Warning: package 'ggplot2' was built under R version 4.0.5
```

```
## Loading required package: ggrepel
```

```
## Warning: package 'ggrepel' was built under R version 4.0.5
```

```
##
## Attaching package: 'PCAtools'
```

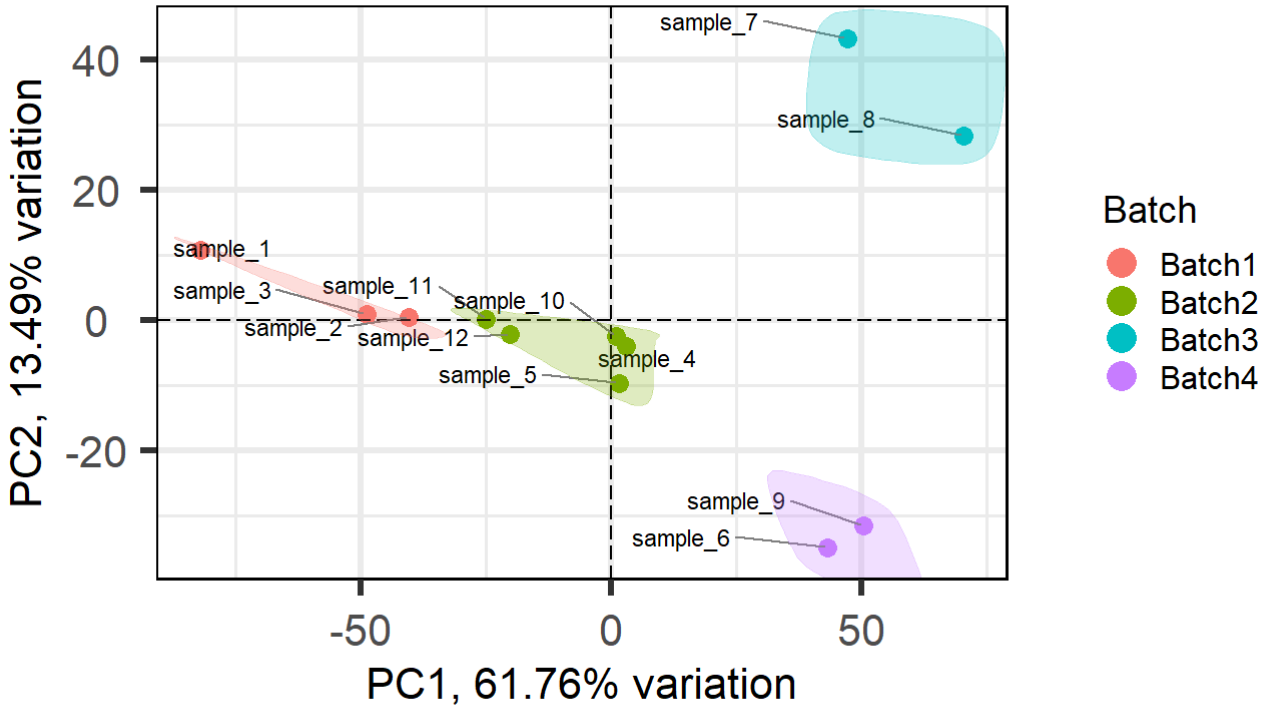
```
## The following objects are masked from 'package:stats':
##
##     biplot, screeplot
```

```
anno <- anno[match(colnames(rawcount), anno$Sample),] ## reordering anno row with colnmaes of rawcount
lograwcount <- as.matrix(log2(rawcount +1))  ## log transformation of rawcount for PCA plot

top1000.order <- head(order(matrixStats::rowVars(lograwcount), decreasing = TRUE), 1000)
p <- PCAtools::pca(mat = lograwcount[top1000.order,], metadata = anno, removeVar = 0.01)
```

```
## -- removing the lower 1% of variables based on variance
```

```
biplot(p,lab = paste0(p$metadata$Sample),
        colby = 'Batch',  #Sample #Batch #Condition #sex
        hline = 0, vline = 0,
        legendPosition = 'right',
        encircle = T )
```

```
## Registered S3 methods overwritten by 'ggalt':
##   method                  from
##   grid.draw.absoluteGrob  ggplot2
##   grobHeight.absoluteGrob ggplot2
##   grobWidth.absoluteGrob  ggplot2
##   grobX.absoluteGrob      ggplot2
##   grobY.absoluteGrob      ggplot2
```



Lets check combat normalization

```
#################################################################
################## Lets check combat normalization #############
############### SVA ######################

  #BiocManager::install("sva")

library('sva')
```

```
## Warning: package 'sva' was built under R version 4.0.3
```

```
## Loading required package: mgcv
```

```
## Loading required package: nlme
```

```
## This is mgcv 1.8-31. For overview type 'help("mgcv-package")'.
```

```
## Loading required package: genefilter
```

```
## Warning: package 'genefilter' was built under R version 4.0.3
```

```
## Loading required package: BiocParallel
```

```
## Warning: package 'BiocParallel' was built under R version 4.0.3
```

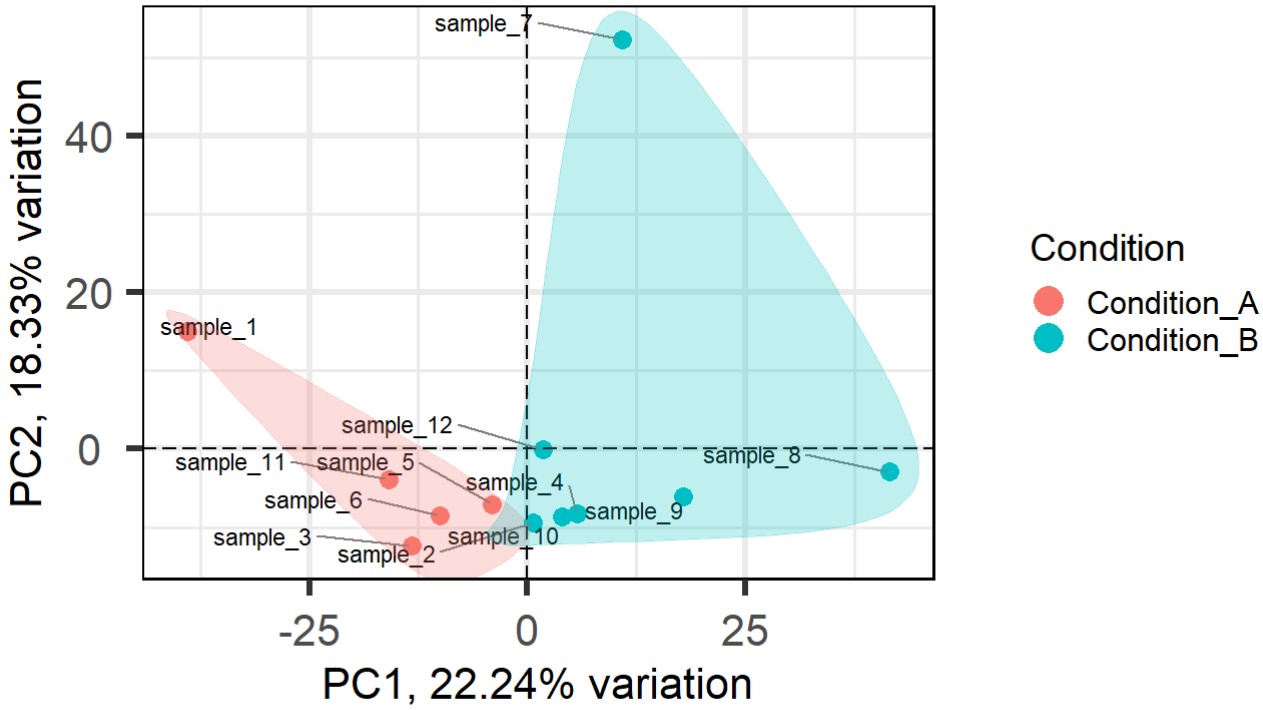
```
rawcount <- as.matrix(rawcount)
adjusted_counts <- ComBat_seq(rawcount, batch=anno$Batch, group=anno$Condition) ##In ComBat-seq, user may specify biological covariates, whose signals will be preserved in the adjusted data. I
```

```
## Found 4 batches
## Using full model in ComBat-seq.
## Adjusting for 1 covariate(s) or covariate level(s)
## Estimating dispersions
## Fitting the GLM model
## Shrinkage off - using GLM estimates for parameters
## Adjusting the data
```

```
nor_set <- as.matrix(log2(adjusted_counts+1)) ## log transformation of adjusted count
top1000.order <- head(order(matrixStats::rowVars(nor_set), decreasing = TRUE), 1000)
pp <- PCAtools::pca(mat =nor_set[top1000.order,] , metadata = anno, removeVar = 0.01)
```

```
## -- removing the lower 1% of variables based on variance
```

```
biplot(pp,
      lab = paste0(p$metadata$Sample),
      #colby = 'Batch',   #Batch_log', #Condition
      colby = 'Condition',
      hline = 0, vline = 0,
      legendPosition = 'right',encircle = T)
```




```
  ##### Do we suppose to remove any defaulty sample  #########

   ### subset raw and conditional data for defined pairs
   ##### Removing sample number 7 ##########

anno <- anno[!(anno$Sample == 'sample_7' | anno$Sample == 'sample_8'),]
rawcount <- as.data.frame(rawcount)
rawcount <- rawcount[,names(rawcount) %in% anno$Sample]

### Go back to PCA plot and check what happned
### perform combat normalization again after removal of sample

# Define conditions (for contrast) that you want to compare if you have more than one #contro
l #case
# This is pair-wise comparison, so only consider one pair at one time

firstC<-"Condition_A"       #case1 #case2 #case3 etc
SecondC <-"Condition_B"
p.threshold <- 0.05   ##define threshold for filtering
```

DESeq2 Analysis

```
################################ Create DESeq2 datasets ##############################
library(DESeq2)
```

```
## Warning: package 'DESeq2' was built under R version 4.0.3
```

```
## Loading required package: S4Vectors
```

```
## Warning: package 'S4Vectors' was built under R version 4.0.3
```

```
## Loading required package: stats4
```

```
## Loading required package: BiocGenerics
```

```
## Warning: package 'BiocGenerics' was built under R version 4.0.5
```

```
## Loading required package: parallel
```

```
##
## Attaching package: 'BiocGenerics'
```

```
## The following objects are masked from 'package:parallel':
##
##     clusterApply, clusterApplyLB, clusterCall, clusterEvalQ,
##     clusterExport, clusterMap, parApply, parCapply, parLapply,
##     parLapplyLB, parRapply, parSapply, parSapplyLB
```

```
## The following objects are masked from 'package:stats':
##
##     IQR, mad, sd, var, xtabs
```

```
## The following objects are masked from 'package:base':
##
##     anyDuplicated, append, as.data.frame, basename, cbind, colnames,
##     dirname, do.call, duplicated, eval, evalq, Filter, Find, get, grep,
##     grepl, intersect, is.unsorted, lapply, Map, mapply, match, mget,
##     order, paste, pmax, pmax.int, pmin, pmin.int, Position, rank,
##     rbind, Reduce, rownames, sapply, setdiff, sort, table, tapply,
##     union, unique, unsplit, which.max, which.min
```

```
##
## Attaching package: 'S4Vectors'
```

```
## The following object is masked from 'package:base':
##
##     expand.grid
```

```
## Loading required package: IRanges
```

```
## Warning: package 'IRanges' was built under R version 4.0.3
```

```
##
## Attaching package: 'IRanges'
```

```
## The following object is masked from 'package:nlme':
##
##     collapse
```

```
## The following object is masked from 'package:grDevices':
##
##     windows
```

```
## Loading required package: GenomicRanges
```

```
## Warning: package 'GenomicRanges' was built under R version 4.0.3
```

```
## Loading required package: GenomeInfoDb
```

```
## Warning: package 'GenomeInfoDb' was built under R version 4.0.5
```

```
## Loading required package: SummarizedExperiment
```

```
## Warning: package 'SummarizedExperiment' was built under R version 4.0.3
```

```
## Loading required package: MatrixGenerics
```

```
## Warning: package 'MatrixGenerics' was built under R version 4.0.3
```

```
## Loading required package: matrixStats
```

```
## Warning: package 'matrixStats' was built under R version 4.0.5
```

```
##
## Attaching package: 'matrixStats'
```

```
## The following objects are masked from 'package:genefilter':
##
##     rowSds, rowVars
```

```
##
## Attaching package: 'MatrixGenerics'
```

```
## The following objects are masked from 'package:matrixStats':
##
##     colAlls, colAnyNAs, colAnys, colAvgsPerRowSet, colCollapse,
##     colCounts, colCummaxs, colCummins, colCumprods, colCumsums,
##     colDiffs, colIQRDiffs, colIQRs, colLogSumExps, colMadDiffs,
##     colMads, colMaxs, colMeans2, colMedians, colMins, colOrderStats,
##     colProds, colQuantiles, colRanges, colRanks, colSdDiffs, colSds,
##     colSums2, colTabulates, colVarDiffs, colVars, colWeightedMads,
##     colWeightedMeans, colWeightedMedians, colWeightedSds,
##     colWeightedVars, rowAlls, rowAnyNAs, rowAnys, rowAvgsPerColSet,
##     rowCollapse, rowCounts, rowCummaxs, rowCummins, rowCumprods,
##     rowCumsums, rowDiffs, rowIQRDiffs, rowIQRs, rowLogSumExps,
##     rowMadDiffs, rowMads, rowMaxs, rowMeans2, rowMedians, rowMins,
##     rowOrderStats, rowProds, rowQuantiles, rowRanges, rowRanks,
##     rowSdDiffs, rowSds, rowSums2, rowTabulates, rowVarDiffs, rowVars,
##     rowWeightedMads, rowWeightedMeans, rowWeightedMedians,
##     rowWeightedSds, rowWeightedVars
```

```
## The following objects are masked from 'package:genefilter':
##
##     rowSds, rowVars
```

```
## Loading required package: Biobase
```

```
## Warning: package 'Biobase' was built under R version 4.0.3
```

```
## Welcome to Bioconductor
##
##     Vignettes contain introductory material; view with
##     'browseVignettes()'. To cite Bioconductor, see
##     'citation("Biobase")', and for packages 'citation("pkgname")'.
```

```
##
## Attaching package: 'Biobase'
```

```
## The following object is masked from 'package:MatrixGenerics':
##
##     rowMedians
```

```
## The following objects are masked from 'package:matrixStats':
##
##     anyMissing, rowMedians
```

```
##dds <- DESeqDataSetFromMatrix(countData = rawcount, colData = anno, design = ~Condition )
##rawcount
dds <- DESeqDataSetFromMatrix(countData = rawcount, colData = anno, design =  ~Batch+Conditio
n )  ###USE this one if you have extra col in anno data with Batch info
```

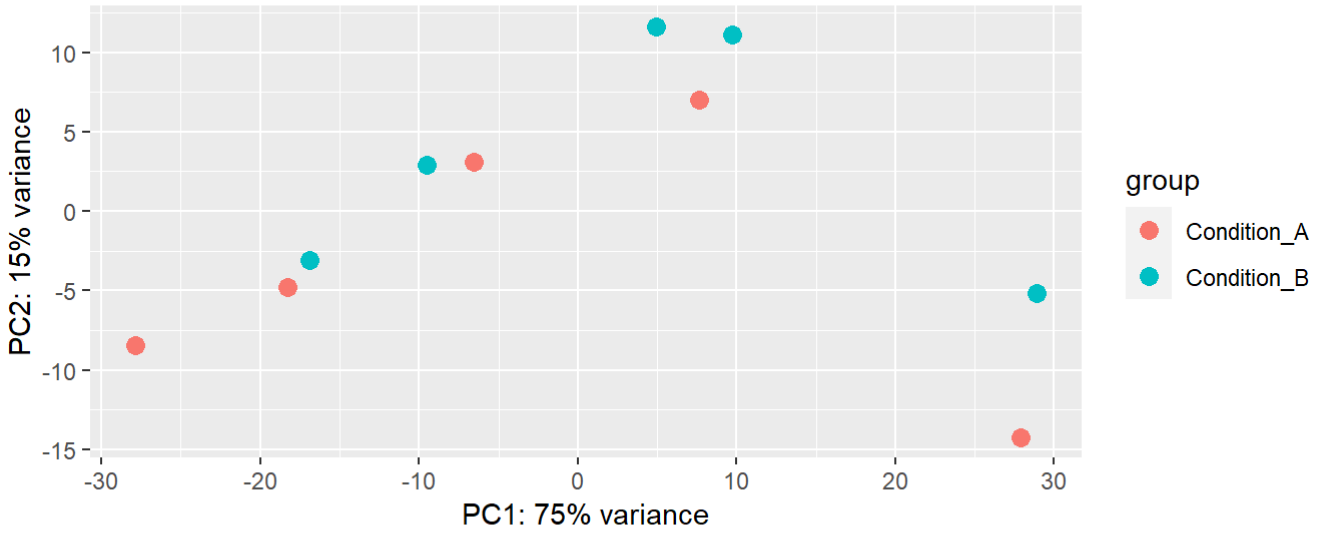
```
## converting counts to integer mode
```

```
## Warning in DESeqDataSet(se, design = design, ignoreRank): some variables in
## design formula are characters, converting to factors
```

```
#dds = DESeq2::DESeqDataSetFromMatrix(countData = adjusted_counts, colData = anno, design = ~ Condition)  ##https://github.com/zhangyuqing/ComBat-seq/issues/7

##When considering batch effects in group design, it takes into account the mean differences across batch,
##not necessarily the variance differences. ComBat-Seq is designed to address both mean and variance batch effects.
###In theory, no, you do not need to include batch as a covariate any more. However, you can always try both and evaluate the results.
```

When considering batch effects in group design, it takes into account the mean differences across batch, not necessarily the variance differences. ComBat-Seq is designed to address both mean and variance batch effects. In theory, no, you do not need to include batch as a covariate any more. However, you can always try both and evaluate the results.

```
#View(counts(dds))

dds <- estimateSizeFactors(dds)
normalized_counts <- counts(dds, normalized=TRUE)  ## extract normalization count after executing Deseq2 for visualization purpose
vst <- vst(dds, blind=TRUE)  ### Transform counts for data visualization #options (1) vst (2) rld
plotPCA(vst, intgroup="Condition")  ### Plot PCA
```



```
## Run DESEQ2
dds <- DESeq(dds)
```

```
## using pre-existing size factors
```

```
## estimating dispersions
```

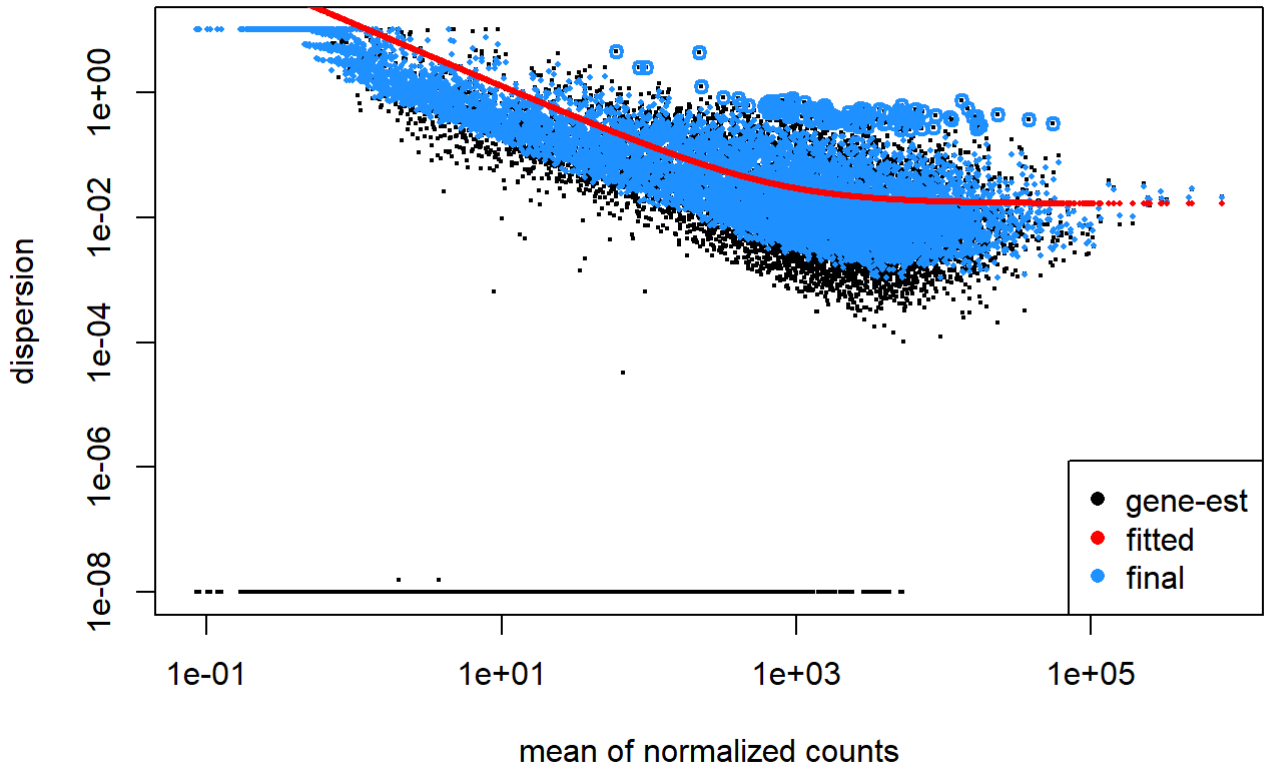
```
## gene-wise dispersion estimates
```

```
## mean-dispersion relationship
```

```
## final dispersion estimates
```

```
## fitting model and testing
```

```
##ensure your data is a good fit for the DESeq2 model
plotDispEsts(dds)
```

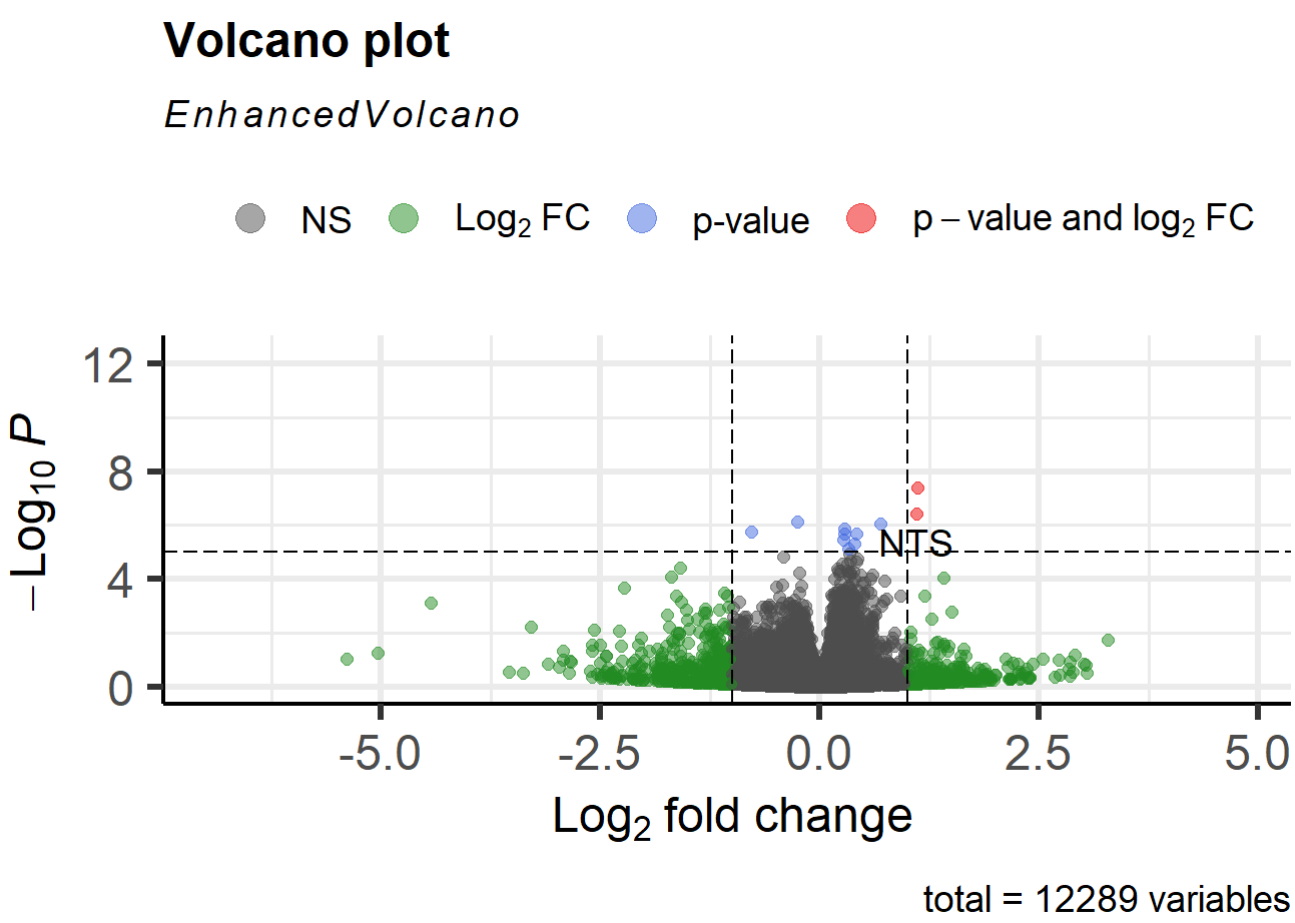


contrast based comparison

```
################## contrast based  comparison ##########################

#In case of multiple comparisons ## we need to change the contrast for every comparision
contrast<- c("Condition",firstC,SecondC)

res <- results(dds, contrast=contrast)  ## extract result dataframe
View(as.data.frame(res))

### Valcono plot
library(EnhancedVolcano)
```

```
## Warning: package 'EnhancedVolcano' was built under R version 4.0.3
```

```
EnhancedVolcano(res,
                lab = rownames(res),
                x = 'log2FoldChange',
                y = 'pvalue')   ## Default cut-off for log2FC is >|2| and for P value is 10e-6. USE  pCutoff = 10e-6, FCcutoff = 2.0
```



```
res$threshold <- as.logical(res$padj < p.threshold)  #Threshold defined earlier

nam <- paste('down_in',firstC, sep = '_')
#res$nam <- as.logical(res$log2FoldChange < 0)
res[, nam] <- as.logical(res$log2FoldChange < 0)

genes.deseq <- row.names(res)[which(res$threshold)]   ### list of gene with Padjust < defined threshold
genes_deseq2_sig <- res[which(res$threshold),]
```

Plots normalized count of top 20 genes

```
############ Plots normalized count of top 20 genes ## sorted based on padjust and filter by |logFC| >=1

res$gene <- row.names(res)
View(as.data.frame(res))

# Order results by padj values

#library(dplyr)
library(tidyverse)
```

```
## Warning: package 'tidyverse' was built under R version 4.0.5
```

```
## -- Attaching packages --------------------------------------- tidyverse 1.3.1 --
```

```
## v tibble  3.1.6     v dplyr   1.0.8
## v tidyr   1.2.0     v stringr 1.4.0
## v readr   2.1.2     v forcats 0.5.1
## v purrr   0.3.4
```

```
## Warning: package 'tibble' was built under R version 4.0.5
```

```
## Warning: package 'tidyr' was built under R version 4.0.5
```

```
## Warning: package 'readr' was built under R version 4.0.5
```

```
## Warning: package 'purrr' was built under R version 4.0.5
```

```
## Warning: package 'dplyr' was built under R version 4.0.5
```

```
## Warning: package 'stringr' was built under R version 4.0.5
```

```
## Warning: package 'forcats' was built under R version 4.0.5
```

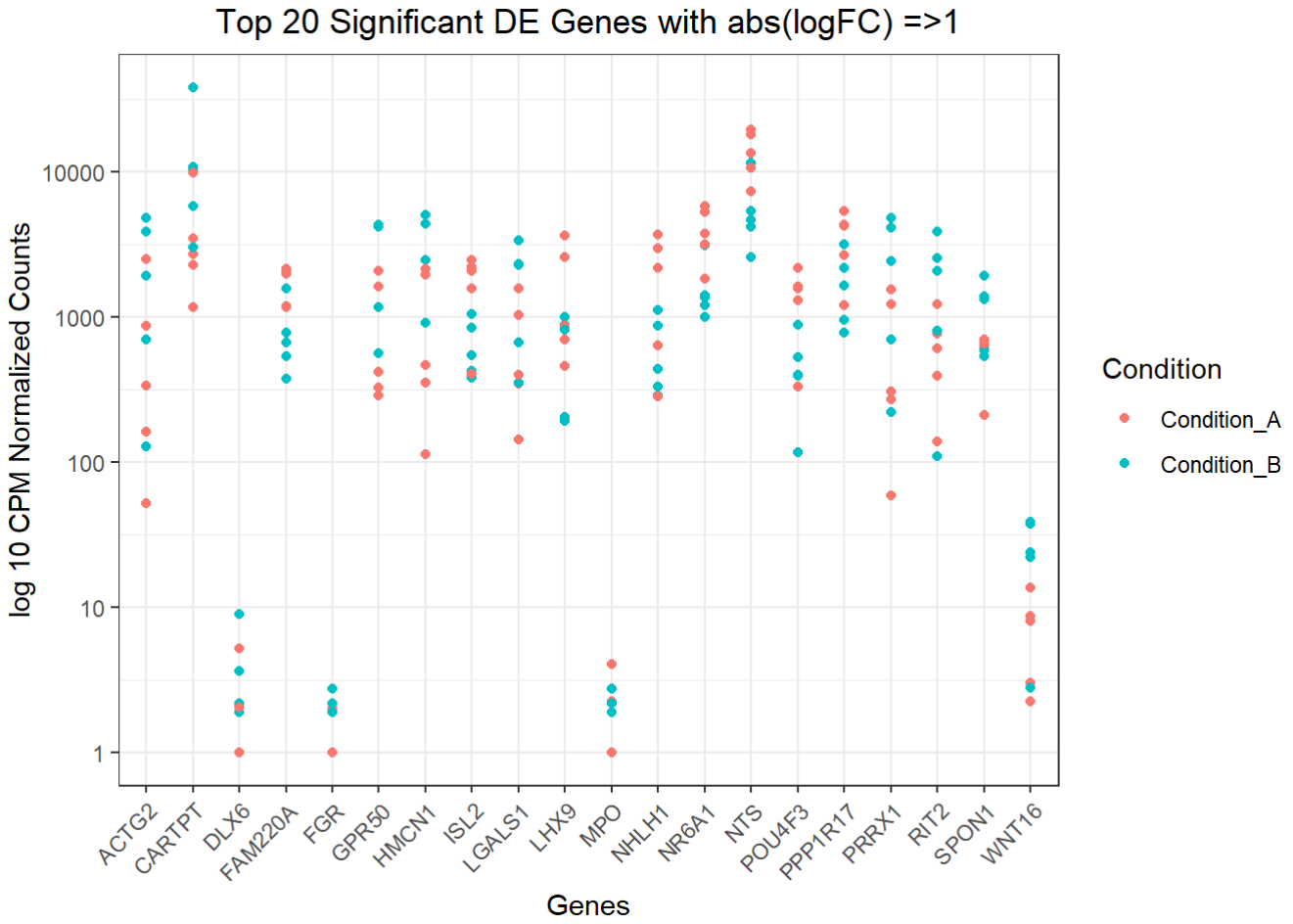
```
## -- Conflicts ---------------------------------------- tidyverse_conflicts() --
## x dplyr::collapse()        masks IRanges::collapse(), nlme::collapse()
## x dplyr::combine()         masks Biobase::combine(), BiocGenerics::combine()
## x dplyr::count()           masks matrixStats::count()
## x dplyr::desc()            masks IRanges::desc()
## x tidyr::expand()          masks S4Vectors::expand()
## x dplyr::filter()          masks stats::filter()
## x dplyr::first()           masks S4Vectors::first()
## x dplyr::lag()             masks stats::lag()
## x BiocGenerics::Position() masks ggplot2::Position(), base::Position()
## x purrr::reduce()          masks GenomicRanges::reduce(), IRanges::reduce()
## x dplyr::rename()          masks S4Vectors::rename()
## x dplyr::slice()           masks IRanges::slice()
## x readr::spec()            masks genefilter::spec()
```

```
top20 <- res %>%
  as.data.frame %>%
  arrange(padj) %>%     #Arrange rows by padj values
  filter(abs(log2FoldChange) >=1) %>%   #filter based on logFC
  pull(gene) %>%        #Extract character vector of ordered genes
  head(n=20)        #Extract the first 20 genes

top20_norm <- as.data.frame(normalized_counts[rownames(normalized_counts) %in% top20,])

top20_norm_v2 <- top20_norm ## will use later for heatmap

top20_norm <- (top20_norm+1) ## in later step to remove infinity bias due to log

top20_norm$gene <-  row.names(top20_norm)
top20_norm <- top20_norm %>%
  pivot_longer(!gene, names_to = "samplename", values_to = "normalized_counts") # Gathering t
he columns to have normalized counts to a single column
```

```
# Create tibbles including row names
mov10_meta <- anno %>%
  rownames_to_column(var="samplename") %>%
  as_tibble()

top20_norm <- inner_join(mov10_meta, top20_norm)
```

```
## Joining, by = "samplename"
```

```
################3
## plot using ggplot2

ggplot(top20_norm) +
  geom_point(aes(x = gene, y = normalized_counts, color = Condition)) +
  scale_y_log10() +
  xlab("Genes") +
  ylab("log 10 CPM Normalized Counts") +
  ggtitle("Top 20 Significant DE Genes with abs(logFC) =>1") +
  theme_bw() +
  theme(axis.text.x = element_text(angle = 45, hjust = 1)) +
  theme(plot.title = element_text(hjust = 0.5))
```



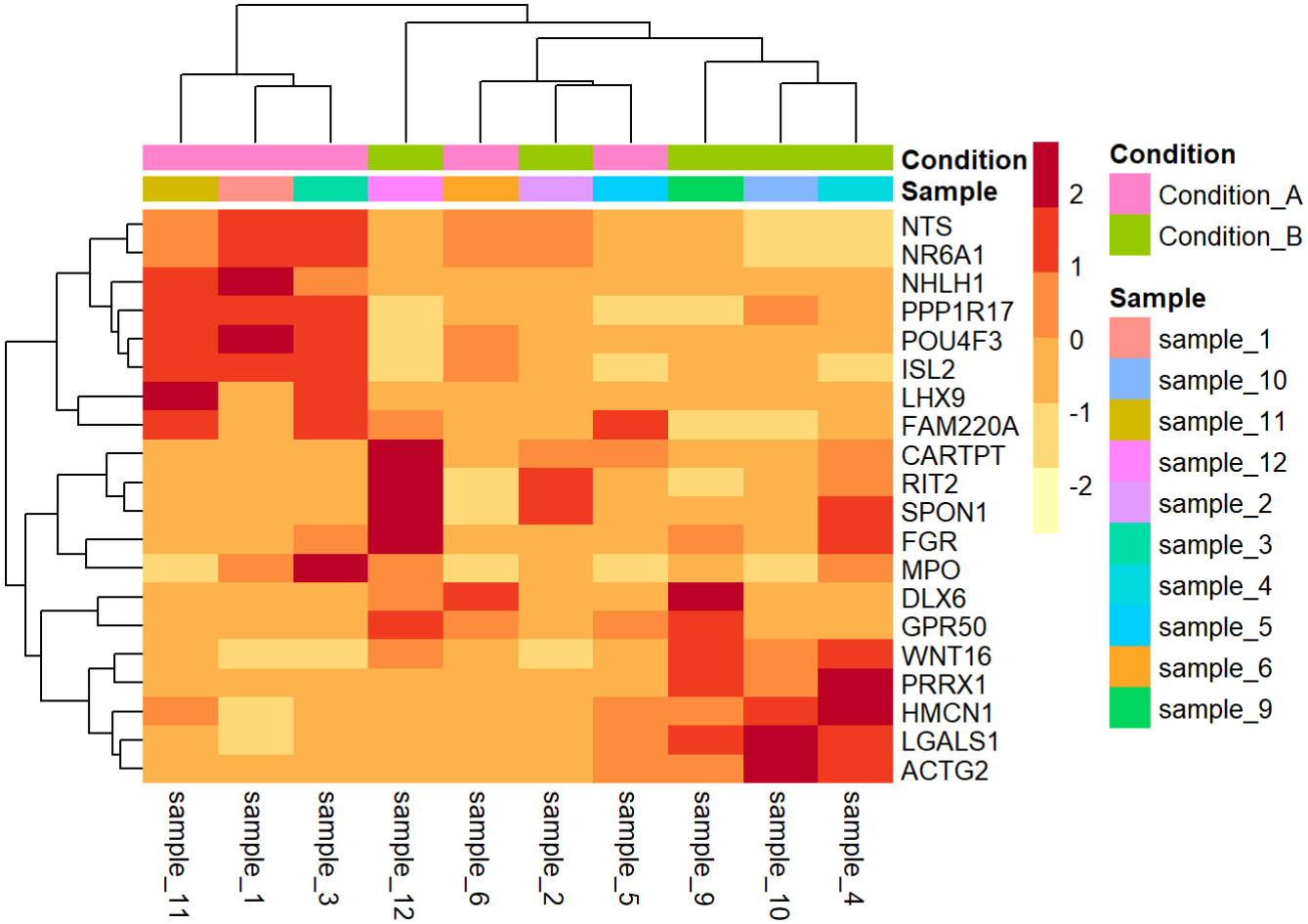

```
##################
library(RColorBrewer)
```

```
## Warning: package 'RColorBrewer' was built under R version 4.0.5
```

```
### Set a color palette
heat_colors <- brewer.pal(6, "YlOrRd")

### Run pheatmap
library(pheatmap)
```

```
## Warning: package 'pheatmap' was built under R version 4.0.5
```

```
pheatmap(top20_norm_v2 ,
         color = heat_colors,
         cluster_rows = T,
         show_rownames = T,
         annotation_col = anno[,1:2],
         border_color = NA,
         fontsize = 10,
         scale = "row",
         fontsize_row = 10,
         height = 20)
```




```
file <- paste('Deseq2_',firstC,'_v_',SecondC,'_results_significant_padj',p.threshold,'.csv',s
ep = '')
all_results <- paste('Deseq2_',firstC,'_v_',SecondC,'_all_results.csv',sep = '')

write.table(genes_deseq2_sig,all_results,sep = ",")  ## no LogFC threshold
```

```
library("biomaRt")
```

```
## Warning: package 'biomaRt' was built under R version 4.0.3
```

```
######################  Filter for coding genes (In case want to filter non-coding Genes) ###
#####################
#new_config <- httr::config(ssl_verifypeer = FALSE) ############For certificate error
#httr::set_config(new_config, override = FALSE)     ############For certificate error

### define the mart for h_sapiens

#ensembl_mart <- useMart(biomart="ensembl", dataset="hsapiens_gene_ensembl")  ## either this
 or following line
ensembl_mart <- useEnsembl(biomart = "ensembl", dataset = "hsapiens_gene_ensembl", mirror =
"asia")

#all_genes <- getBM(attributes = c( "hgnc_symbol","ensembl_gene_id","ensembl_gene_id_versio
n"),  mart =ensembl_mart)  ## etract df of verious types of ID

#################################################################3

### Add EntrezID column to results dataframe for easier downstream processing ####
genes_deseq2_sig <- as.data.frame(genes_deseq2_sig)
genes_deseq2_sig$hgnc_symbol = row.names(genes_deseq2_sig)  ## significant gene table from pr
evious DE analysis
row.names(genes_deseq2_sig) <- NULL


genes.deseq.entrezid <- getBM(attributes = c("hgnc_symbol", "entrezgene_id"), filters = "hgnc
_symbol", values = genes_deseq2_sig$hgnc_symbol, mart = ensembl_mart)
#genes.deseq.entrezid = as.data.frame(genes.deseq.entrezid) ## if not

merged <- merge(genes_deseq2_sig, genes.deseq.entrezid, by.x= "hgnc_symbol", by.y="hgnc_symbo
l")

##### You may want to filter genes based on LOGFC threshold

merged <- merged[(merged$log2FoldChange >=1 | merged$log2FoldChange <= -1),]
```

Filter for coding genes

```
######################  Filter for coding genes (In case want to filter non-coding Genes) ###
####################
library("biomaRt")

#new_config <- httr::config(ssl_verifypeer = FALSE) ############For certificate error
#httr::set_config(new_config, override = FALSE)     ############For certificate error

### define the mart for h_sapiens

#ensembl_mart <- useMart(biomart="ensembl", dataset="hsapiens_gene_ensembl")  ## either this
 or following line
ensembl_mart <- useEnsembl(biomart = "ensembl", dataset = "hsapiens_gene_ensembl", mirror =
"asia")

#all_genes <- getBM(attributes = c( "hgnc_symbol","ensembl_gene_id","ensembl_gene_id_versio
n"),  mart =ensembl_mart)  ## etract df of verious types of ID

#################################################################3

### Add EntrezID column to results dataframe for easier downstream processing ####
genes_deseq2_sig <- as.data.frame(genes_deseq2_sig)
genes_deseq2_sig$hgnc_symbol = row.names(genes_deseq2_sig)  ## significant gene table from pr
evious DE analysis
row.names(genes_deseq2_sig) <- NULL


genes.deseq.entrezid <- getBM(attributes = c("hgnc_symbol", "entrezgene_id"), filters = "hgnc
_symbol", values = genes_deseq2_sig$hgnc_symbol, mart = ensembl_mart)
#genes.deseq.entrezid = as.data.frame(genes.deseq.entrezid) ## if not

merged <- merge(genes_deseq2_sig, genes.deseq.entrezid, by.x= "hgnc_symbol", by.y="hgnc_symbo
l")

##### You may want to filter genes based on LOGFC threshold

merged <- merged[(merged$log2FoldChange >=1 | merged$log2FoldChange <= -1),]
```

```
######### Rank all genes based on their fold change #########

#BiocManager::install("clusterProfiler", force = TRUE)
#BiocManager::install("pathview", force = TRUE)
#BiocManager::install("enrichplot", force = TRUE)

library(clusterProfiler)
```

```
## Warning: package 'clusterProfiler' was built under R version 4.0.3
```

```
##
```

```
## clusterProfiler v3.18.1  For help: https://guangchuangyu.github.io/software/clusterProfile
r
##
## If you use clusterProfiler in published research, please cite:
## Guangchuang Yu, Li-Gen Wang, Yanyan Han, Qing-Yu He. clusterProfiler: an R package for com
paring biological themes among gene clusters. OMICS: A Journal of Integrative Biology. 2012,
16(5):284-287.
```

```
##
## Attaching package: 'clusterProfiler'
```

```
## The following object is masked from 'package:biomaRt':
##
##     select
```

```
## The following object is masked from 'package:purrr':
##
##     simplify
```

```
## The following object is masked from 'package:IRanges':
##
##     slice
```

```
## The following object is masked from 'package:S4Vectors':
##
##     rename
```

```
## The following object is masked from 'package:stats':
##
##     filter
```

```
library(enrichplot)
```

```
## Warning: package 'enrichplot' was built under R version 4.0.3
```

```
library(ggplot2)

# SET THE DESIRED ORGANISM HERE
organism = "org.Hs.eg.db"

#BiocManager::install(organism, character.only = TRUE, force = TRUE)

library(organism, character.only = TRUE)
```

```
## Loading required package: AnnotationDbi
```

```
## Warning: package 'AnnotationDbi' was built under R version 4.0.3
```

```
##
## Attaching package: 'AnnotationDbi'
```

```
## The following object is masked from 'package:clusterProfiler':
##
##     select
```

```
## The following object is masked from 'package:dplyr':
##
##     select
```

```
##
```

```
keytypes(org.Hs.eg.db)
```

```
##  [1] "ACCNUM"       "ALIAS"        "ENSEMBL"      "ENSEMBLPROT"  "ENSEMBLTRANS"
##  [6] "ENTREZID"     "ENZYME"       "EVIDENCE"     "EVIDENCEALL"  "GENENAME"
## [11] "GO"           "GOALL"        "IPI"          "MAP"          "OMIM"
## [16] "ONTOLOGY"     "ONTOLOGYALL"  "PATH"         "PFAM"         "PMID"
## [21] "PROSITE"      "REFSEQ"       "SYMBOL"       "UCSCKG"       "UNIGENE"
## [26] "UNIPROT"
```

```
#We will take the log2FoldChange value from previously saved significant results file
#Deseq2_case1_v_Control_results_significant.csv


df <- read.csv("Deseq2_case1_v_Control_results_significant_padj0.05.csv")
#df <- merged

# we want the log2 fold change
original_gene_list <- df$log2FoldChange
print(original_gene_list)
```

```
##  [1] -3.3746677 -5.5019061  2.4111152  3.0032625  2.1549852 -1.6592273
##  [7] -3.3088168 -0.9355305 -3.7523776 -1.8295745 -2.9292739  1.8946263
## [13]  1.9665047 -1.5365003  2.2806666  2.8573898  1.3294100 -1.6483657
## [19] -2.9063413  3.5811752 -2.5145863  2.1761636 -0.8358506  1.4580477
## [25] -1.3522671  3.4421976 -4.9974798 -2.3654838  1.4271523 -1.2074048
## [31]  1.4270554  1.8408182  1.2192076
```

```
# name the vector
names(original_gene_list) <- df$entrezgene_id

# omit any NA values
gene_list<-na.omit(original_gene_list)

# sort the list in decreasing order (required for clusterProfiler)
gene_list = sort(gene_list, decreasing = TRUE)

print(gene_list)
```

```
##       7869      11063       <NA>       5798        586      93166       6556
##  3.5811752  3.4421976  3.0032625  2.8573898  2.4111152  2.2806666  2.1761636
##       1301       4481       9235       1462      84679       3604       7305
##  2.1549852  1.9665047  1.8946263  1.8408182  1.4580477  1.4271523  1.4270554
##      26064      81671       6558      56204      79090       6646       4784
##  1.3294100  1.2192076 -0.8358506 -0.9355305 -1.2074048 -1.3522671 -1.5365003
##       6095      23604       2328       7049      56920       6401      22844
## -1.6483657 -1.6592273 -1.8295745 -2.3654838 -2.5145863 -2.9063413 -2.9292739
##      84649       6296      55711       7021        416
## -3.3088168 -3.3746677 -3.7523776 -4.9974798 -5.5019061
```

Gene Set Enrichment

```
#### Gene Set Enrichment #####

library(stats)

gse <- gseGO(geneList=gene_list,
             ont ="ALL",
             keyType = "ENTREZID",
             minGSSize = 3,
             maxGSSize = 800,
             pvalueCutoff = 0.05,
             verbose = TRUE,
             OrgDb = org.Hs.eg.db,
             pAdjustMethod = "none")
```

```
## preparing geneSet collections...
```

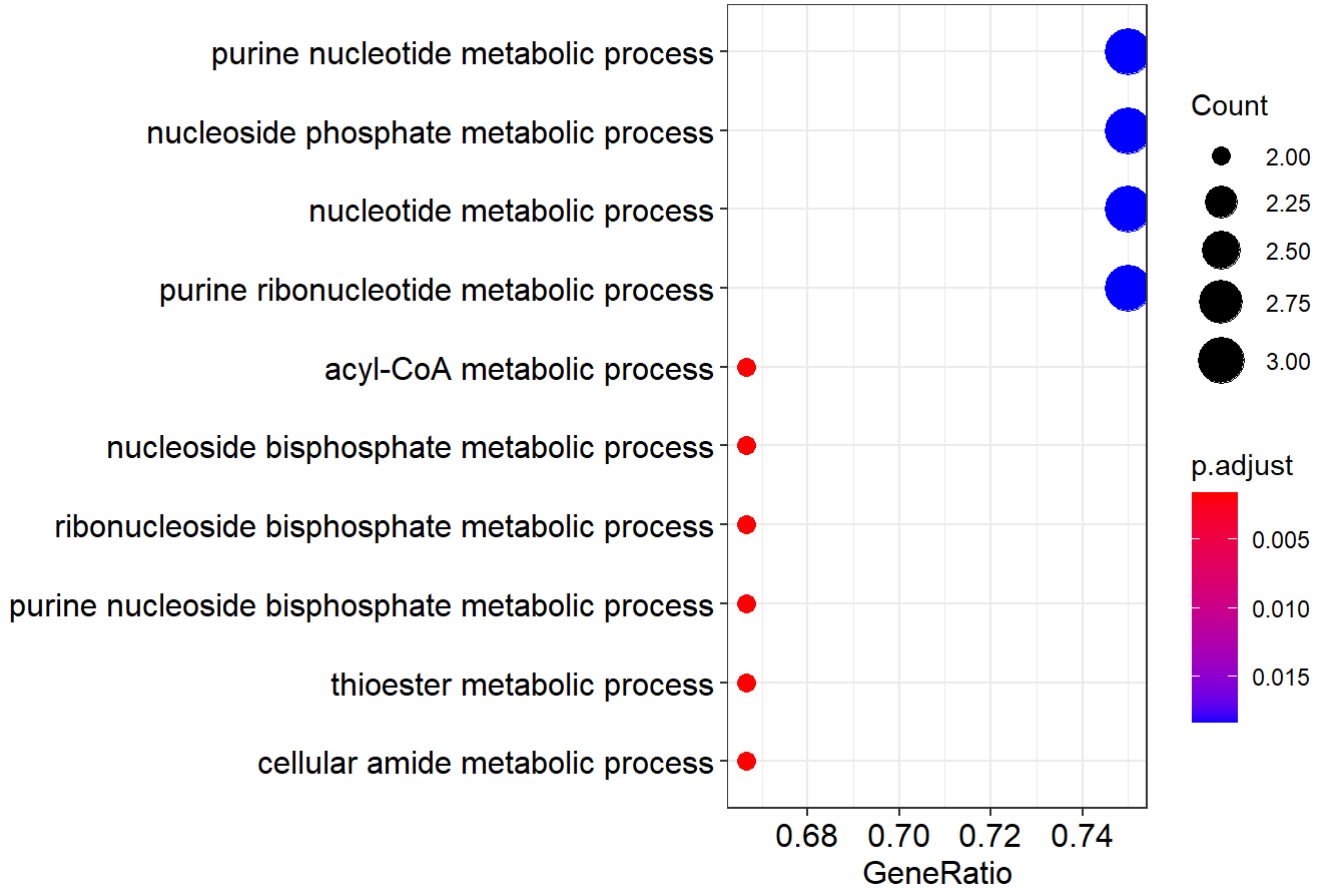
```
## GSEA analysis...
```

```
## leading edge analysis...
```

```
## done...
```

```
# require(DOSE)
dotplot(gse, showCategory=10, split=".sign", orderBy = "X")
```

```
## wrong orderBy parameter; set to default `orderBy = "x"`
```



We can see that in our dataset not a single value is enriched at a pvalue cut-off of 0.05.

Lets exlore other functions with a sample dataset and see what analysis we can do with a list of differrentially expressed genes from geneList dataset of DOSE package

##GO Enrichment Analysis of a gene set. Given a vector of genes, enrichGO function will return the enrichment GO categories after FDR control.

```
##GO Enrichment Analysis of a gene set.
##Given a vector of genes, enrichGO function will return the
##enrichment GO categories after FDR control.


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

```
## Warning: package 'GOSemSim' was built under R version 4.0.3
```

```
## GOSemSim v2.16.1  For help: https://guangchuangyu.github.io/GOSemSim
##
## If you use GOSemSim in published research, please cite:
##  [36m- [39m Guangchuang Yu. Gene Ontology Semantic Similarity Analysis Using GOSemSim. In:
Kidder B. (eds) Stem Cell Transcriptional Networks. Methods in Molecular Biology, 2020, 2117:
207-215. Humana, New York, NY. doi:10.1007/978-1-0716-0301-7_11
##  [36m- [39m Guangchuang Yu, Fei Li, Yide Qin, Xiaochen Bo, Yibo Wu, Shengqi Wang. GOSemSi
m: an R package for measuring semantic similarity among GO terms and gene products Bioinforma
tics 2010, 26(7):976-978. doi:10.1093/bioinformatics/btq064
```

```
library(ggnewscale)
```

```
## Warning: package 'ggnewscale' was built under R version 4.0.5
```

```
library(DOSE)
```

```
## Warning: package 'DOSE' was built under R version 4.0.3
```

```
## DOSE v3.16.0  For help: https://guangchuangyu.github.io/software/DOSE
##
## If you use DOSE in published research, please cite:
## Guangchuang Yu, Li-Gen Wang, Guang-Rong Yan, Qing-Yu He. DOSE: an R/Bioconductor package f
or Disease Ontology Semantic and Enrichment analysis. Bioinformatics 2015, 31(4):608-609
```

```
##
## Attaching package: 'DOSE'
```

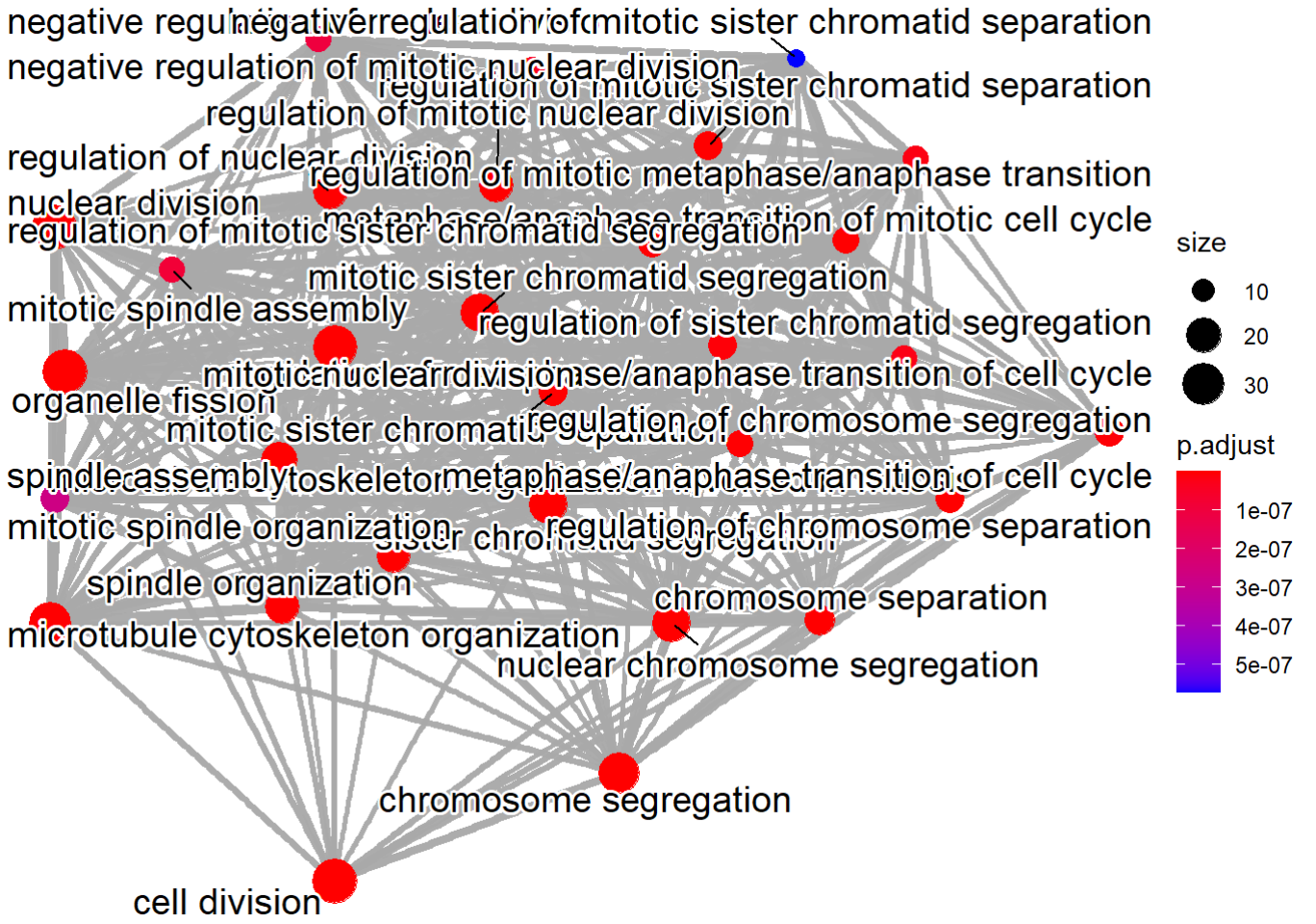
```
## The following objects are masked from 'package:GOSemSim':
##
##     clusterSim, geneSim, mclusterSim
```

```
data(geneList)
View(geneList)
gene <- names(geneList)[abs(geneList) > 2]
ego <- enrichGO(gene  = gene,
                universe      = names(geneList),
                OrgDb         = org.Hs.eg.db,
                ont           = "BP",
                pAdjustMethod = "BH",
                pvalueCutoff  = 0.01,
                qvalueCutoff  = 0.05,
                readable      = TRUE)

##### Visualization of enrichGO ######
d <- godata('org.Hs.eg.db', ont="BP")
```

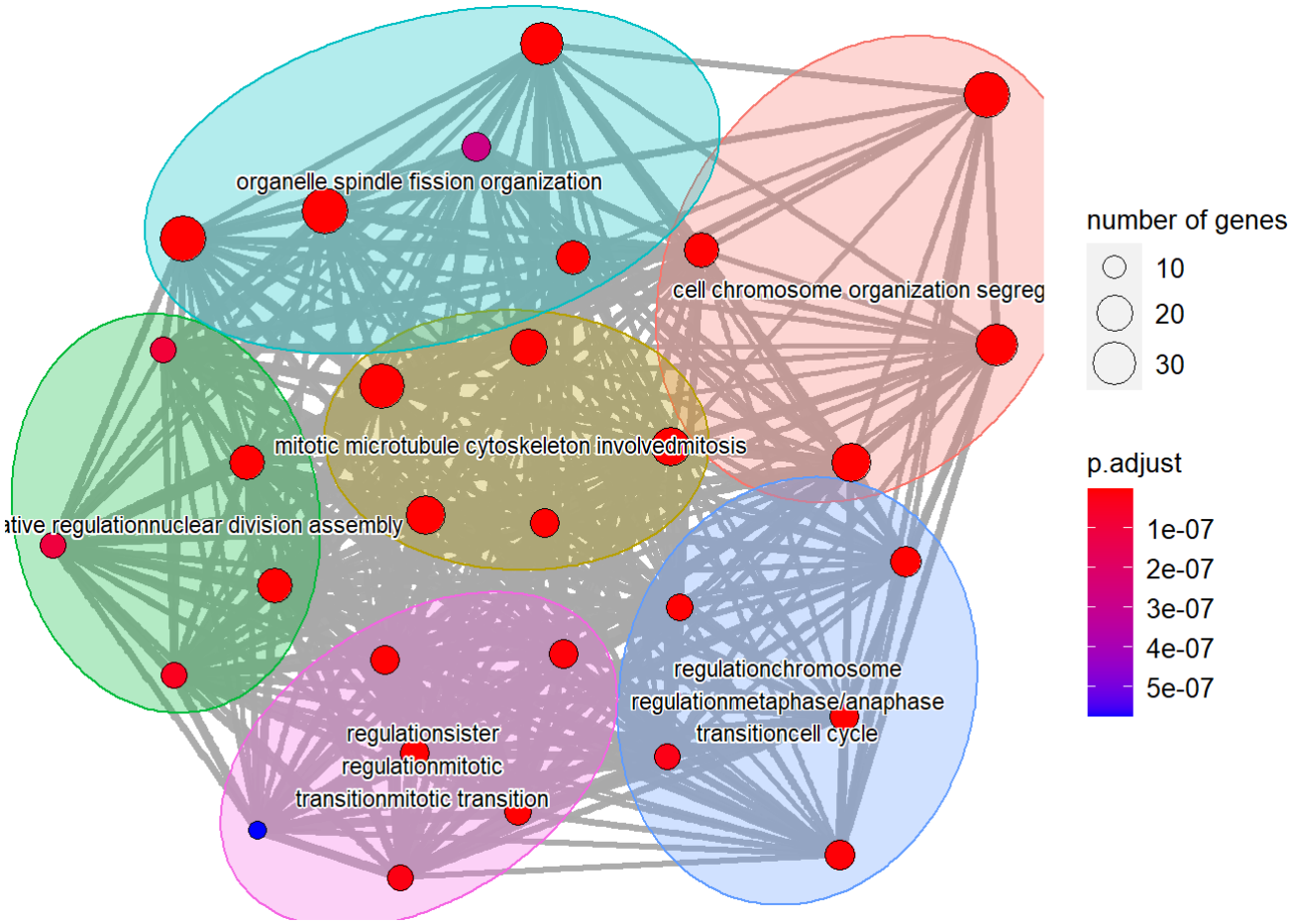
```
## preparing gene to GO mapping data...
```

```
## preparing IC data...
```

```
ego2 <- pairwise_termsim(ego, method="Wang", semData = d)
emapplot(ego2)
```




```
emapplot_cluster(ego2)
```




```
### Try GO with all different ont methods parameter
## BP = Biological Processes, CC= Cellular component, MF = Molecular functions

###In the following example, we selected fold change above 1 as the differential genes
##and analyzing their disease association.
```

In the following example, we selected fold change above 1 as the differential genes and analyzing their disease association.

enrich DO

```
#### enrich DO #####

library(ggupset)
```

```
## Warning: package 'ggupset' was built under R version 4.0.5
```

```
gene = names(geneList)[abs(geneList) > 1.5]
head(gene)
```

```
## [1] "4312"  "8318"  "10874" "55143" "55388" "991"
```

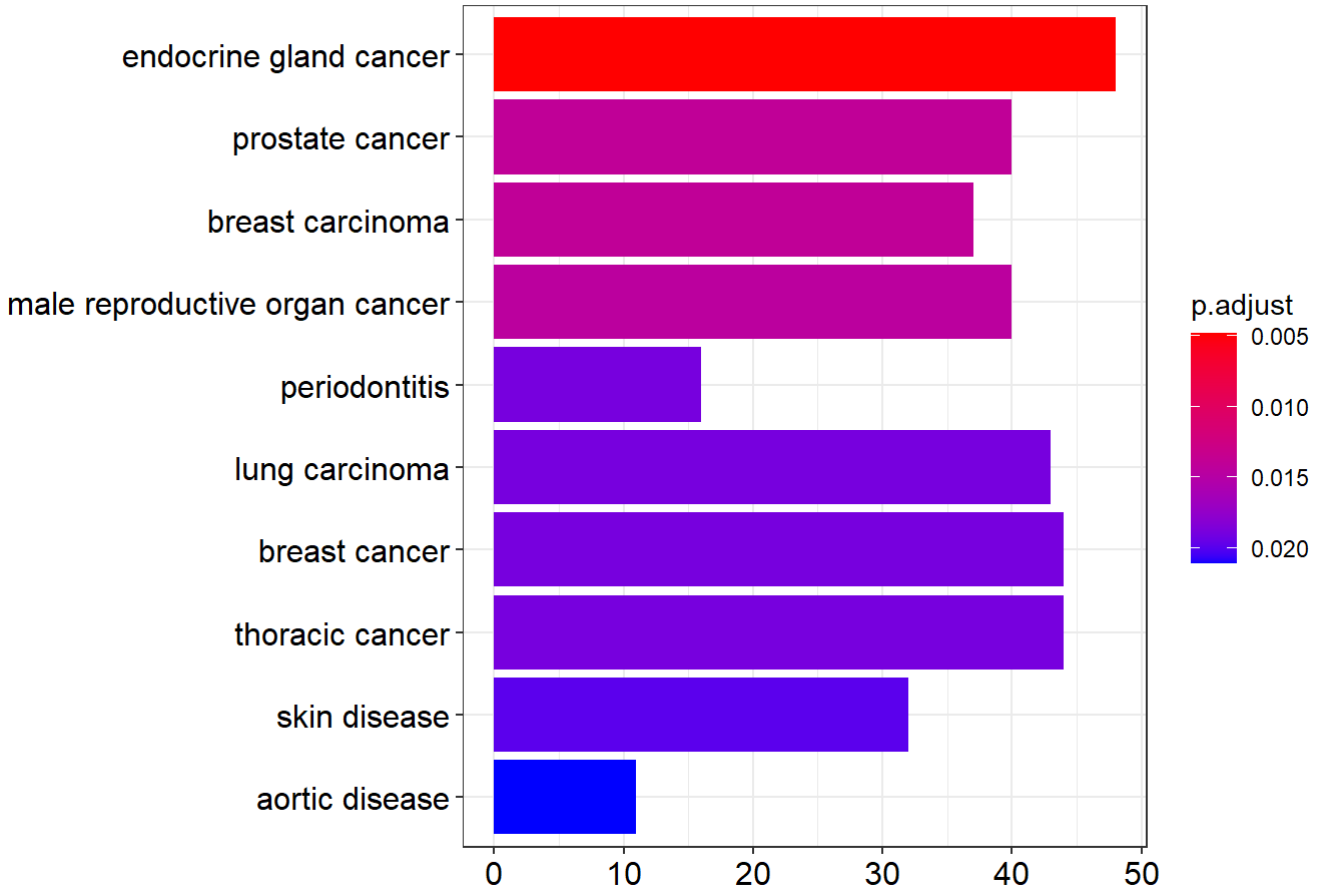
```
X = enrichDO(gene,ont = "DO",
             pvalueCutoff=0.05,
             pAdjustMethod = "BH",
             universe = names(geneList),
             minGSSize     = 5,
             maxGSSize     = 500,
             qvalueCutoff  = 0.05,
             readable      = FALSE)

#The readable is a logical parameter,
#indicates whether the entrezgene IDs will mapping to gene symbols or not
head(X)
```

```
##                    ID                     Description GeneRatio  BgRatio
## DOID:170     DOID:170         endocrine gland cancer    48/331 472/6268
## DOID:10283 DOID:10283                 prostate cancer    40/331 394/6268
## DOID:3459   DOID:3459               breast carcinoma    37/331 357/6268
## DOID:3856   DOID:3856 male reproductive organ cancer    40/331 404/6268
## DOID:824     DOID:824                   periodontitis    16/331 109/6268
## DOID:3905   DOID:3905                 lung carcinoma    43/331 465/6268
##                  pvalue    p.adjust      qvalue
## DOID:170   5.662129e-06 0.004784499 0.003826407
## DOID:10283 3.859157e-05 0.013921739 0.011133923
## DOID:3459  4.942629e-05 0.013921739 0.011133923
## DOID:3856  6.821467e-05 0.014410349 0.011524689
## DOID:824   1.699304e-04 0.018859464 0.015082872
## DOID:3905  1.749754e-04 0.018859464 0.015082872
##
geneID
## DOID:170   10874/7153/1381/6241/11065/10232/332/6286/2146/10112/891/9232/4171/993/5347/431
8/3576/1515/4821/8836/3159/7980/5888/333/898/9768/4288/3551/2152/9590/185/7043/3357/2952/532
7/3667/1634/1287/4582/7122/3479/4680/6424/80310/652/8839/9547/1524
## DOID:10283                                          4312/6280/6279/597/3627/332/6286/2146/
4321/4521/891/5347/4102/4318/701/3576/79852/10321/6352/4288/3551/2152/247/2952/3487/367/3667/
4128/4582/563/3679/4117/7031/3479/6424/10451/80310/652/4036/10551
## DOID:3459                                                              4312/6280/6279/7153/475
1/890/4085/332/6286/6790/891/9232/10855/4171/5347/4318/701/2633/3576/9636/898/8792/4288/2952/
4982/4128/4582/7031/3479/771/4250/2066/3169/10647/5304/5241/10551
## DOID:3856                                          4312/6280/6279/597/3627/332/6286/2146/
4321/4521/891/5347/4102/4318/701/3576/79852/10321/6352/4288/3551/2152/247/2952/3487/367/3667/
4128/4582/563/3679/4117/7031/3479/6424/10451/80310/652/4036/10551
## DOID:824
4312/6279/820/7850/4321/3595/4318/4069/3576/1493/6352/8842/185/2952/5327/4982
## DOID:3905                        4312/6280/2305/9133/6279/7153/6278/6241/55165/11065/814
0/10232/332/6286/3002/9212/4521/891/4171/9928/8061/4318/3576/1978/1894/7980/7083/898/6352/884
2/4288/2152/2697/2952/3572/4582/7049/563/3479/1846/3117/2532/2922
##            Count
## DOID:170      48
## DOID:10283    40
## DOID:3459     37
## DOID:3856     40
## DOID:824      16
## DOID:3905     43
```

```
#setReadable function helps to convert entrezgene IDs to gene symbols
X <- setReadable(X, 'org.Hs.eg.db')
head(X)
```

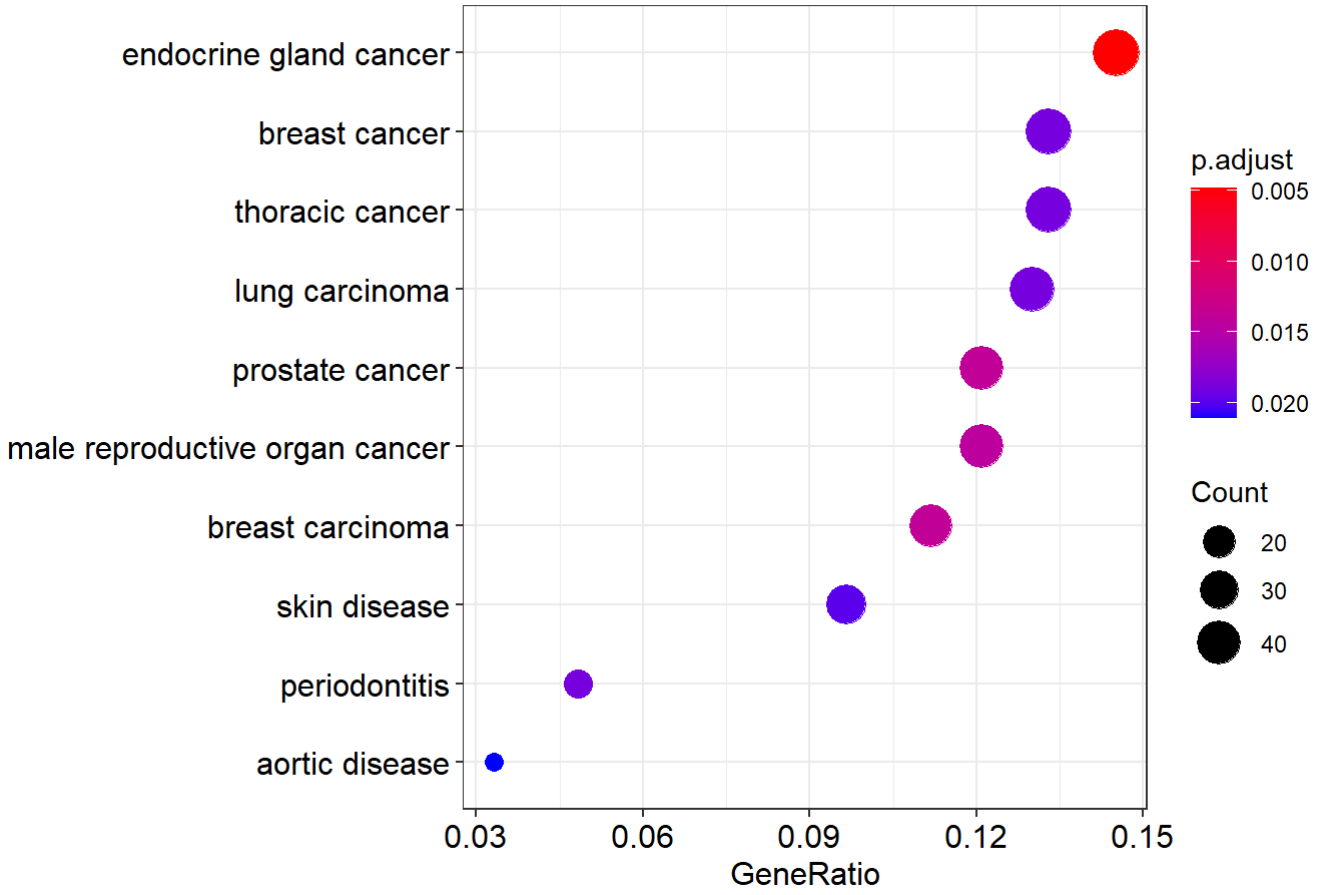
```
##                    ID                    Description GeneRatio  BgRatio
## DOID:170     DOID:170         endocrine gland cancer    48/331 472/6268
## DOID:10283 DOID:10283                 prostate cancer    40/331 394/6268
## DOID:3459   DOID:3459               breast carcinoma    37/331 357/6268
## DOID:3856   DOID:3856 male reproductive organ cancer    40/331 404/6268
## DOID:824     DOID:824                   periodontitis    16/331 109/6268
## DOID:3905   DOID:3905                 lung carcinoma    43/331 465/6268
##                  pvalue    p.adjust      qvalue
## DOID:170   5.662129e-06 0.004784499 0.003826407
## DOID:10283 3.859157e-05 0.013921739 0.011133923
## DOID:3459  4.942629e-05 0.013921739 0.011133923
## DOID:3856  6.821467e-05 0.014410349 0.011524689
## DOID:824   1.699304e-04 0.018859464 0.015082872
## DOID:3905  1.749754e-04 0.018859464 0.015082872
##
geneID
## DOID:170   NMU/TOP2A/CRABP1/RRM2/UBE2C/MSLN/BIRC5/S100P/EZH2/KIF20A/CCNB1/PTTG1/MCM2/CDC25
A/PLK1/MMP9/CXCL8/CTSV/NKX2-2/GGH/HMGA1/TFPI2/RAD51/APLP1/CCNE1/PCLAF/MKI67/IKBKB/F3/AKAP12/A
GTR1/TGFB3/HTR2B/GSTT1/PLAT/IRS1/DCN/COL4A5/MUC1/CLDN5/IGF1/CEACAM6/SFRP4/PDGFD/BMP4/CCN5/CXC
L14/CX3CR1
## DOID:10283                                                          MMP1/S100A9/S100A8/BCL2A1/CXCL
10/BIRC5/S100P/EZH2/MMP12/NUDT1/CCNB1/PLK1/MAGEA3/MMP9/BUB1B/CXCL8/EPHX3/CRISP3/CCL5/MKI67/IK
BKB/F3/ALOX15B/GSTT1/IGFBP4/AR/IRS1/MAOA/MUC1/AZGP1/ITGA7/MAK/TFF1/IGF1/SFRP4/VAV3/PDGFD/BMP
4/LRP2/AGR2
## DOID:3459                                                                MMP1/S100A9/S100A8/TOP2
A/NEK2/CCNA2/MAD2L1/BIRC5/S100P/AURKA/CCNB1/PTTG1/HPSE/MCM2/PLK1/MMP9/BUB1B/GBP1/CXCL8/ISG15/
CCNE1/TNFRSF11A/MKI67/GSTT1/TNFRSF11B/MAOA/MUC1/TFF1/IGF1/CA12/SCGB2A2/ERBB4/FOXA1/SCGB1D2/PI
P/PGR/AGR2
## DOID:3856                                                          MMP1/S100A9/S100A8/BCL2A1/CXCL
10/BIRC5/S100P/EZH2/MMP12/NUDT1/CCNB1/PLK1/MAGEA3/MMP9/BUB1B/CXCL8/EPHX3/CRISP3/CCL5/MKI67/IK
BKB/F3/ALOX15B/GSTT1/IGFBP4/AR/IRS1/MAOA/MUC1/AZGP1/ITGA7/MAK/TFF1/IGF1/SFRP4/VAV3/PDGFD/BMP
4/LRP2/AGR2
## DOID:824
MMP1/S100A8/CAMP/IL1R2/MMP12/IL12RB2/MMP9/LYZ/CXCL8/CTLA4/CCL5/PROM1/AGTR1/GSTT1/PLAT/TNFRSF1
1B
## DOID:3905                              MMP1/S100A9/FOXM1/CCNB2/S100A8/TOP2A/S100A7/RRM2/CEP5
5/UBE2C/SLC7A5/MSLN/BIRC5/S100P/GZMB/AURKB/NUDT1/CCNB1/MCM2/KIF14/FOSL1/MMP9/CXCL8/EIF4EBP1/E
CT2/TFPI2/TK1/CCNE1/CCL5/PROM1/MKI67/F3/GJA1/GSTT1/IL6ST/MUC1/TGFBR3/AZGP1/IGF1/DUSP4/HLA-DQA
1/ACKR1/GRP
##            Count
## DOID:170      48
## DOID:10283    40
## DOID:3459     37
## DOID:3856     40
## DOID:824      16
## DOID:3905     43
```

```
## Visualization of enrichDO results ##
barplot(X, showCategory=10)
```




```
dotplot(X)
```

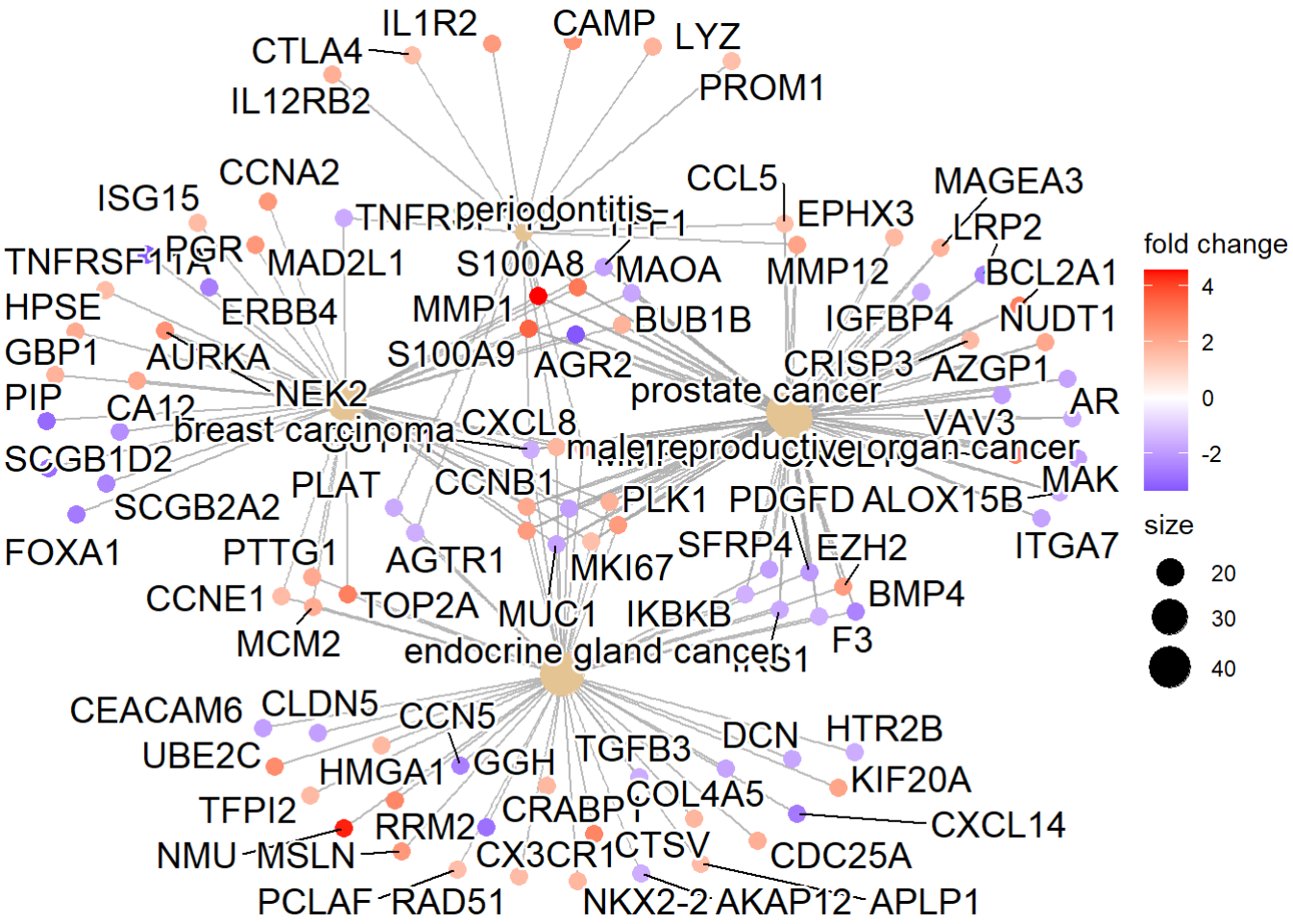
```
## wrong orderBy parameter; set to default `orderBy = "x"`
```



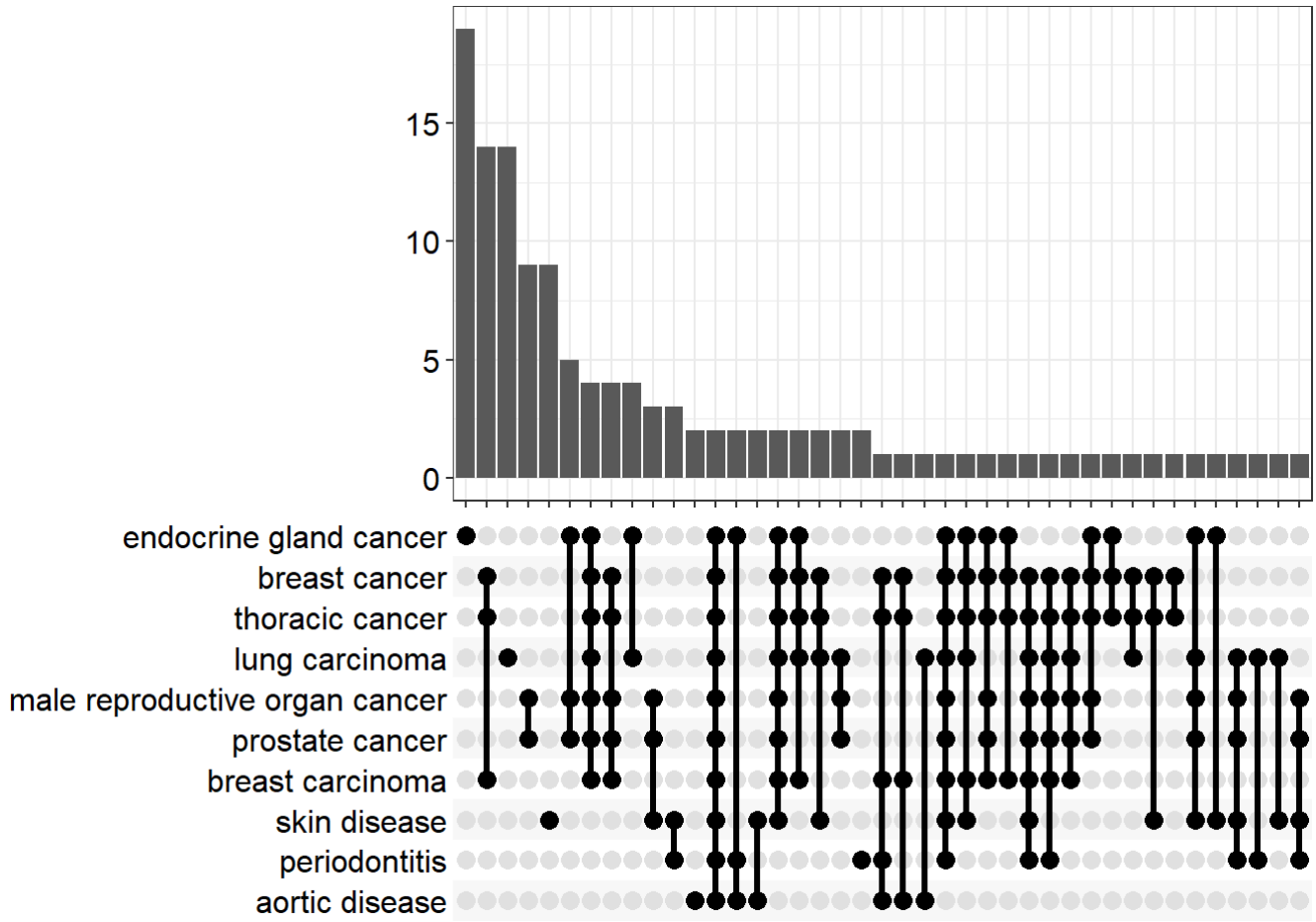
###Multiple annotation categories

```
#gene may belong to multiple annotation categories,
#we developed cnetplot function to extract the complex association between genes and diseases
cnetplot(X, categorySize="pvalue", foldChange=geneList)
```

```
## Warning: ggrepel: 3 unlabeled data points (too many overlaps). Consider
## increasing max.overlaps
```

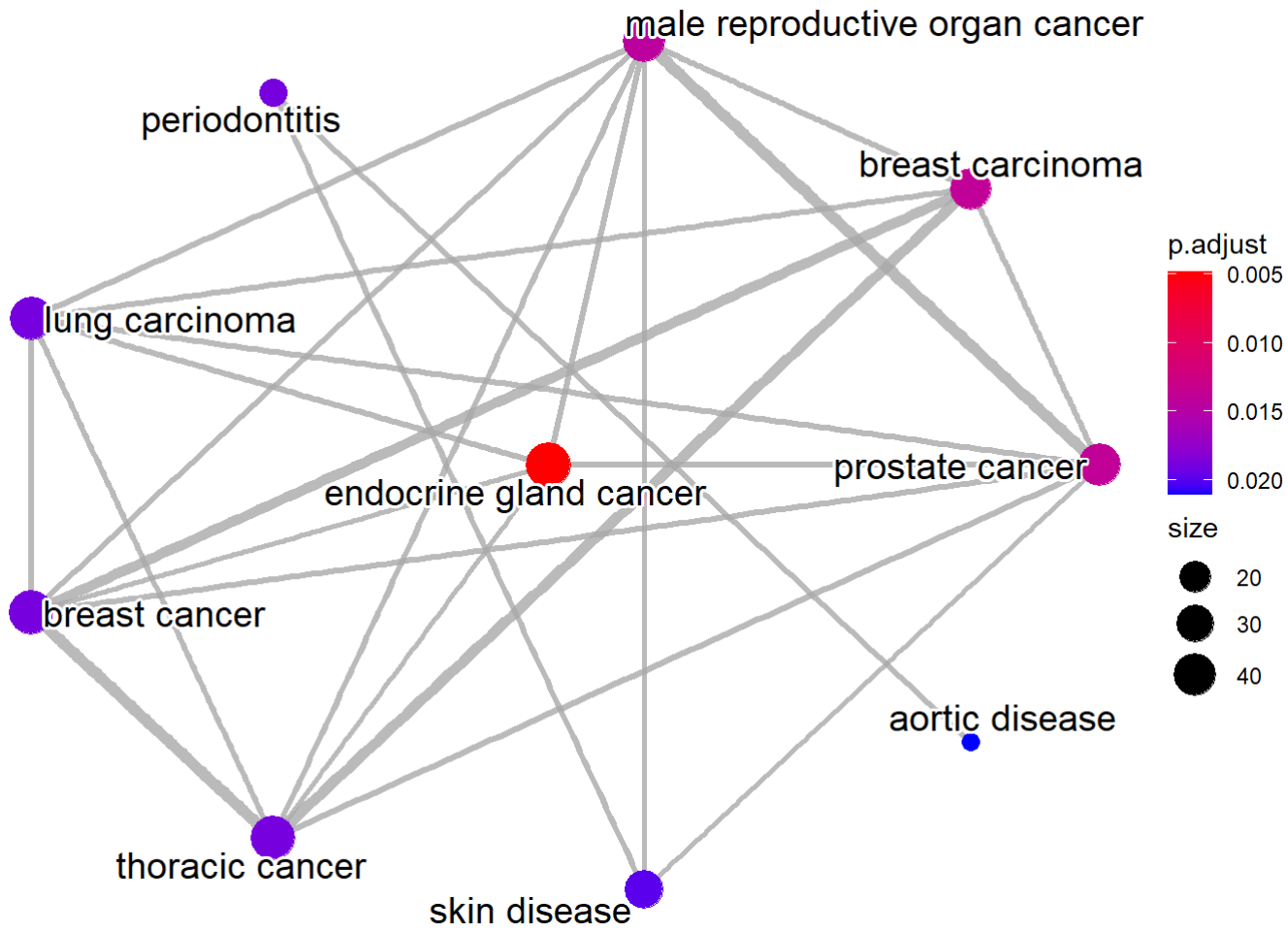



```
#upsetplot is an alternative to
#cnetplot for visualizing the complex association between genes and diseases.

upsetplot(X)
```



###Enrichment Map

```
## Enrichment Map ##
###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.

X2 <- pairwise_termsim(X)
emapplot(X2, showCategory = 10, layout = "star")
```



###Enrichment of NCG (Network of Cancer Gene)

```
#Network of Cancer Gene (NCG)3 is a manually curated repository of cancer genes.
#NCG release 5.0 (Aug. 2015) collects 1,571 cancer genes from 175 published studies.
#DOSE supports analyzing gene list and
#determine whether they are enriched in genes known to be mutated in a given cancer type.

######## enrichNCG function ####

gene2 <- names(geneList)[abs(geneList) < 3]
ncg <- enrichNCG(gene2)
head(ncg)
```

```
##                                                                    ID
## pan-cancer_paediatric                           pan-cancer_paediatric
## triple_negative_breast_cancer             triple_negative_breast_cancer
## breast_cancer                                         breast_cancer
## soft_tissue_sarcoma                                 soft_tissue_sarcoma
## paediatric_high-grade_glioma               paediatric_high-grade_glioma
## pancreatic_cancer_(all_histologies) pancreatic_cancer_(all_histologies)
##                                                          Description
## pan-cancer_paediatric                           pan-cancer_paediatric
## triple_negative_breast_cancer             triple_negative_breast_cancer
## breast_cancer                                         breast_cancer
## soft_tissue_sarcoma                                 soft_tissue_sarcoma
## paediatric_high-grade_glioma               paediatric_high-grade_glioma
## pancreatic_cancer_(all_histologies) pancreatic_cancer_(all_histologies)
##                                     GeneRatio  BgRatio       pvalue
## pan-cancer_paediatric                161/1782 182/2372 2.748816e-06
## triple_negative_breast_cancer         71/1782  75/2372 6.564667e-06
## breast_cancer                        146/1782 171/2372 5.249102e-04
## soft_tissue_sarcoma                   26/1782  26/2372 5.633144e-04
## paediatric_high-grade_glioma          25/1782  25/2372 7.524752e-04
## pancreatic_cancer_(all_histologies)   39/1782  41/2372 7.825494e-04
##                                        p.adjust       qvalue
## pan-cancer_paediatric               0.0002226541 0.0001504615
## triple_negative_breast_cancer       0.0002658690 0.0001796646
## breast_cancer                       0.0105644165 0.0071390469
## soft_tissue_sarcoma                 0.0105644165 0.0071390469
## paediatric_high-grade_glioma        0.0105644165 0.0071390469
## pancreatic_cancer_(all_histologies) 0.0105644165 0.0071390469
##
geneID
## pan-cancer_paediatric               2146/55353/4609/1029/3575/22806/3418/3066/2120/30012/8
67/7468/7545/3195/865/64109/4613/613/11177/7490/238/10736/10054/5771/4893/140885/1785/9760/34
17/6597/6476/9126/4869/10320/7307/80204/1050/8028/2312/6608/896/894/2196/4849/7023/5093/5079/
5293/5727/55181/171017/51322/5781/3718/55294/60/673/8085/5897/4851/51176/1108/7764/10664/609
8/2332/2201/6495/3845/7015/1441/2782/64919/4298/23512/8239/29102/6929/8021/6134/6598/4209/529
0/22941/8726/207/3717/2033/10716/4928/6932/694/5156/10019/6886/9968/7080/2623/7874/1654/4149/
3020/23219/55252/55729/10735/5728/4853/23451/51341/387/3206/6146/79718/2624/63035/3815/17102
3/23269/25/9839/23592/5896/7403/2260/54880/3716/9203/57178/6777/5789/4297/29072/90/546/120/25
836/8289/4345/9611/5925/4763/1997/1499/7157/3399/5295/1387/4602/51564/1027/4005/2322/2078/67
8/6403/55709/1277/7494/64061/2625
## triple_negative_breast_cancer
6790/898/4609/1029/1789/4436/2120/867/7128/1788/1030/7490/2271/238/675/2047/4914/1316/5291/52
93/5781/55294/8085/4851/4170/3845/355/1616/4854/5290/207/2033/4233/29110/2903/5979/5728/4853/
2624/3815/10000/7403/2260/55193/472/5789/4297/2065/4286/8626/8405/8289/10499/55164/5925/4763/
23405/1499/4921/7157/5295/1387/2078/324/7248/7048/22894/3480/2045/2066/2625
## breast_cancer
4751/701/898/639/29028/4609/7399/1029/1520/4436/83990/11200/10849/2072/4771/865/999/1788/2619
1/1030/10801/83737/6262/1956/672/8590/675/4893/6597/8202/2778/208/51412/896/2132/677/4849/422
1/65220/2854/55294/673/4193/8085/4851/57127/841/3265/7764/10664/9721/3845/3956/868/9175/6602/
11174/8239/9860/6954/5290/1523/207/2033/2334/3782/8312/9514/5156/186/54897/71/79728/545/143/2
064/4089/8471/8314/91/5289/1021/10735/5979/5728/4853/23451/9439/6738/387/55770/79718/4301/171
023/23013/51135/80243/4292/149076/10983/6103/7403/54880/4916/55193/9203/1635/1495/2309/472/50
76/2909/5789/4297/2065/29072/2263/546/8289/2874/9611/5925/6416/4763/7157/4088/23152/5295/679
4/1387/4602/1027/5737/324/595/7188/4681/4214/7494/2099/3480/4485/2891/6926/3169/2625
## soft_tissue_sarcoma
```

```
999/6850/4914/4342/2185/55294/2041/4851/2044/4058/5290/4486/5297/5728/3815/2324/7403/546/592
5/4763/1499/7157/5159/2045/3667/2066
## paediatric_high-grade_glioma
4609/1029/1019/4613/1030/1956/4914/896/894/673/8493/5290/4233/5156/1021/63035/54880/4916/90/5
46/4763/7157/5295/595/4915
## pancreatic_cancer_(all_histologies)
1029/4771/8997/7159/2011/6597/7307/3710/6710/55294/7091/3845/23654/7046/3096/4089/91/8241/545
49/92/23451/63035/7403/55193/23309/472/800/29072/23077/23499/8289/54894/6416/7157/4088/182/70
48/2199/26960
##                                    Count
## pan-cancer_paediatric                161
## triple_negative_breast_cancer         71
## breast_cancer                        146
## soft_tissue_sarcoma                   26
## paediatric_high-grade_glioma          25
## pancreatic_cancer_(all_histologies)   39
```

###Disease Gene Association

```
##The enrichment analysis of disease-gene associations is supported by the enrichDGN function
##to determine whether the genes have associations with any known diseases
#### gene disease association #####

dgn <- enrichDGN(gene)
head(dgn)
```

```
##                ID                                Description GeneRatio   BgRatio
## C0010278 C0010278                        Craniosynostosis    43/497 488/21671
## C0853879 C0853879             Invasive carcinoma of breast    42/497 473/21671
## C4733092 C4733092 estrogen receptor-negative breast cancer    34/497 356/21671
## C3642347 C3642347              Basal-Like Breast Carcinoma    28/497 245/21671
## C3642345 C3642345               Luminal A Breast Carcinoma    22/497 153/21671
## C0036202 C0036202                               Sarcoidosis    36/497 413/21671
##                pvalue     p.adjust       qvalue
## C0010278 4.609534e-14 2.267976e-10 1.636811e-10
## C0853879 7.105190e-14 2.267976e-10 1.636811e-10
## C4733092 2.446675e-12 4.864593e-09 3.510804e-09
## C3642347 3.047991e-12 4.864593e-09 3.510804e-09
## C3642345 7.034749e-12 8.438458e-09 6.090082e-09
## C0036202 7.930882e-12 8.438458e-09 6.090082e-09
##
geneID
## C0010278 4312/8318/6280/1062/6279/6278/3627/820/27299/6362/81620/2146/3002/29968/990/4318/
4069/3576/6890/23594/26279/1493/6352/4998/2152/2697/185/4330/5327/4982/1300/3667/2200/9607/35
72/563/7031/3479/6424/1846/3117/1308/2625
## C0853879        4312/7153/6278/9787/9582/51203/890/983/5080/2146/1111/9232/10855/4171/666
4/4102/2173/4318/701/3576/1978/8836/53335/1894/7980/8792/8842/2151/185/2952/367/4982/4582/692
6/3479/1602/23158/2066/3169/5304/2625/5241
## C4733092                                          2305/6278/79733/6241/81930/81620/2146/362
0/29968/11004/8061/3576/1894/2491/7083/8792/214/5327/367/4982/3667/4582/27324/3479/1846/8012
9/4137/8839/3169/1408/5304/2625/5241/10551
## C3642347                                                                          2305/106
2/4605/9833/7368/11065/10232/55765/5163/2146/2568/3620/6790/6664/29127/2173/4318/3576/3159/87
92/6663/27324/3479/1846/18/3169/2625/5241
## C3642345
2305/9833/7153/55355/1111/3161/4318/3576/2001/6663/4288/2152/185/4128/4582/27324/80129/3169/5
304/8614/2625/5241
## C0036202                                4312/6280/6279/10403/3627/6373/4283/27299/6362/300
2/4321/6355/6364/29851/4318/5004/4069/3576/26227/6890/6352/4485/23541/185/7043/6863/2952/498
2/25802/4582/2053/3479/3117/2167/80736/1524
##          Count
## C0010278    43
## C0853879    42
## C4733092    34
## C3642347    28
## C3642345    22
## C0036202    36
```

Gene set enrichment analysis

```
####### Gene set enrichment analysis GSEA Plot ######
gsecc <- gseGO(geneList=geneList, ont="CC", OrgDb=org.Hs.eg.db, verbose=F)
```

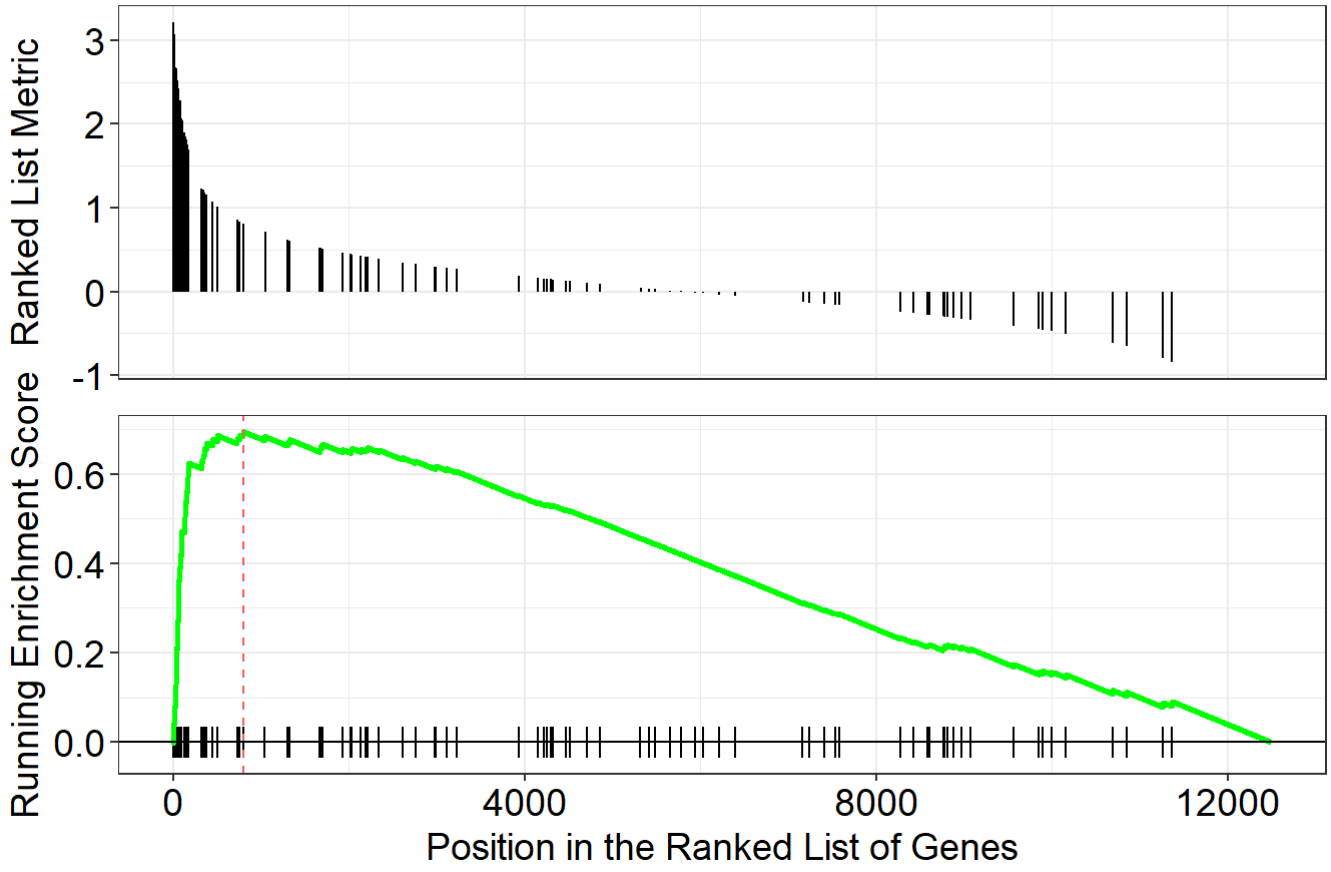
```
## Warning in fgseaMultilevel(...): For some pathways, in reality P-values are less
## than 1e-10. You can set the `eps` argument to zero for better estimation.
```

```
head(summary(gsecc))
```

```
## Warning in summary(gsecc): summary method to convert the object to data.frame is
## deprecated, please use as.data.frame instead.
```

```
##                    ID                                Description setSize
## GO:0000775 GO:0000775           chromosome, centromeric region     158
## GO:0000776 GO:0000776                                kinetochore     109
## GO:0000777 GO:0000777         condensed chromosome kinetochore      81
## GO:0000779 GO:0000779 condensed chromosome, centromeric region      95
## GO:0000793 GO:0000793                     condensed chromosome     167
## GO:0005819 GO:0005819                                  spindle     288
##            enrichmentScore      NES pvalue     p.adjust qvalues rank
## GO:0000775       0.6225504 2.707748  1e-10 7.666667e-09   6e-09  511
## GO:0000776       0.6570788 2.685598  1e-10 7.666667e-09   6e-09  449
## GO:0000777       0.6976827 2.737700  1e-10 7.666667e-09   6e-09  759
## GO:0000779       0.6945448 2.796205  1e-10 7.666667e-09   6e-09  798
## GO:0000793       0.5923226 2.585214  1e-10 7.666667e-09   6e-09 2215
## GO:0005819       0.4792198 2.252574  1e-10 7.666667e-09   6e-09  437
##                              leading_edge
## GO:0000775  tags=22%, list=4%, signal=22%
## GO:0000776  tags=25%, list=4%, signal=24%
## GO:0000777  tags=31%, list=6%, signal=29%
## GO:0000779  tags=33%, list=6%, signal=31%
## GO:0000793 tags=39%, list=18%, signal=32%
## GO:0005819  tags=15%, list=3%, signal=15%
##
core_enrichment
## GO:0000775
55143/1062/10403/55355/220134/4751/79019/55839/54821/4085/81930/81620/332/7272/64151/9212/679
0/891/11004/5347/701/11130/79682/57405/10615/79075/2491/11339/3070/9918/1058/699/1063/55055/1
051
## GO:0000776
1062/10403/55355/220134/4751/79019/55839/54821/4085/81930/81620/332/7272/9212/891/11004/5347/
701/11130/79682/57405/10615/2491/1058/699/1063/55055
## GO:0000777
1062/10403/55355/220134/4751/79019/55839/54821/4085/81620/332/891/11004/5347/701/11130/79682/
57405/10615/1058/699/1063/55055/79980/9735
## GO:0000779
1062/10403/55355/220134/4751/79019/55839/54821/4085/81620/332/64151/9212/6790/891/11004/5347/
701/11130/79682/57405/10615/9918/1058/699/1063/55055/1051/79980/9735/23310
## GO:0000793 1062/10403/7153/23397/55355/220134/4751/79019/55839/54821/4085/81620/332/64151/
9212/1111/6790/891/11004/5347/701/11130/79682/57405/10615/5888/4288/9918/1058/699/1063/55055/
641/1051/54892/3148/79980/9735/23310/10051/1104/23481/5885/7283/92822/54908/10592/6839/23212/
3014/5905/3619/11335/7273/9770/8940/79677/672/79902/55320/3297/675/5119/9793/79172
## GO:0005819
55143/991/9493/1062/259266/9787/220134/51203/22974/10460/4751/983/4085/81930/332/3832/7272/92
12/9055/3833/146909/10112/6790/891/24137/9928/11004/79801/990/5347/29127/701/10615/1894/9700/
56992/10733/54801/54959/29899/994/1063/26271
```

```
gseaplot(gsecc, geneSetID="GO:0000779")
```



KEGG Enrichment Analysis

```
###### KEGG Enrichment Analysis #######

library(clusterProfiler)

## KEGG pathway over-representation analysis
data(geneList, package="DOSE")
gene <- names(geneList)[abs(geneList) > 2]
```

```
kk <- enrichKEGG(gene         = gene,
                 organism     = 'hsa',
                 pvalueCutoff = 0.05)
```

```
## Reading KEGG annotation online:
##
## Reading KEGG annotation online:
```

```
head(kk)
```

```
##               ID                                                       Description
## hsa04110 hsa04110                                                        Cell cycle
## hsa04114 hsa04114                                                     Oocyte meiosis
## hsa04218 hsa04218                                                Cellular senescence
## hsa04061 hsa04061 Viral protein interaction with cytokine and cytokine receptor
## hsa03320 hsa03320                                             PPAR signaling pathway
## hsa04914 hsa04914                          Progesterone-mediated oocyte maturation
##          GeneRatio  BgRatio       pvalue     p.adjust       qvalue
## hsa04110     11/94 126/8142 1.829412e-07 3.841764e-05 3.774365e-05
## hsa04114     10/94 131/8142 2.368439e-06 2.486861e-04 2.443231e-04
## hsa04218     10/94 156/8142 1.135672e-05 7.949704e-04 7.810235e-04
## hsa04061      8/94 100/8142 1.821466e-05 9.562698e-04 9.394931e-04
## hsa03320      7/94  75/8142 2.285993e-05 9.601169e-04 9.432728e-04
## hsa04914      7/94 102/8142 1.651911e-04 5.781690e-03 5.680256e-03
##                                                      geneID Count
## hsa04110 8318/991/9133/890/983/4085/7272/1111/891/4174/9232    11
## hsa04114    991/9133/983/4085/51806/6790/891/9232/3708/5241    10
## hsa04218     2305/4605/9133/890/983/51806/1111/891/776/3708    10
## hsa04061           3627/10563/6373/4283/6362/6355/9547/1524     8
## hsa03320                 4312/9415/9370/5105/2167/3158/5346     7
## hsa04914                    9133/890/983/4085/6790/891/5241     7
```

```
## KEGG module over-representation analysis
#KEGG Module is a collection of manually defined function units. In some situation,
#KEGG Modules have a more straightforward interpretation

mkk <- enrichMKEGG(gene = gene,
                   organism = 'hsa',
                   pvalueCutoff = 1,
                   qvalueCutoff = 1)
```

```
## Reading KEGG annotation online:
##
## Reading KEGG annotation online:
```

```
head(mkk)
```

```
##            ID                                                Description
## M00912 M00912       NAD biosynthesis, tryptophan => quinolinate => NAD
## M00095 M00095           C5 isoprenoid biosynthesis, mevalonate pathway
## M00053 M00053 Pyrimidine deoxyribonucleotide biosynthesis, CDP => dCTP
## M00938 M00938 Pyrimidine deoxyribonucleotide biosynthesis, UDP => dTTP
## M00003 M00003             Gluconeogenesis, oxaloacetate => fructose-6P
## M00049 M00049      Adenine ribonucleotide biosynthesis, IMP => ADP,ATP
##        GeneRatio BgRatio      pvalue   p.adjust     qvalue     geneID Count
## M00912       2/9  12/831 0.006511179 0.03906707 0.03426936 23475/3620     2
## M00095       1/9  10/831 0.103710201 0.18875552 0.16557502       3158     1
## M00053       1/9  11/831 0.113535546 0.18875552 0.16557502       6241     1
## M00938       1/9  14/831 0.142439710 0.18875552 0.16557502       6241     1
## M00003       1/9  18/831 0.179674425 0.18875552 0.16557502       5105     1
## M00049       1/9  19/831 0.188755520 0.18875552 0.16557502      26289     1
```

```
## KEGG module gene set enrichment analysis ##

mkk2 <- gseMKEGG(geneList = geneList,
                 organism = 'hsa',
                 pvalueCutoff = 1)
```

```
## preparing geneSet collections...
```

```
## GSEA analysis...
```

```
## leading edge analysis...
```

```
## done...
```

```
head(mkk2)
```

```
##            ID                                                        Description
## M00001 M00001        Glycolysis (Embden-Meyerhof pathway), glucose => pyruvate
## M00002 M00002         Glycolysis, core module involving three-carbon compounds
## M00035 M00035                                          Methionine degradation
## M00938 M00938         Pyrimidine deoxyribonucleotide biosynthesis, UDP => dTTP
## M00009 M00009                              Citrate cycle (TCA cycle, Krebs cycle)
## M00104 M00104 Bile acid biosynthesis, cholesterol => cholate/chenodeoxycholate
##        setSize enrichmentScore       NES      pvalue  p.adjust   qvalues rank
## M00001      24       0.5739036  1.771863 0.005062727 0.1569445 0.1438880 2886
## M00002      11       0.6421781  1.599342 0.024822030 0.2699189 0.2474639 1381
## M00035      10       0.6784636  1.619691 0.027622470 0.2699189 0.2474639 1555
## M00938      10       0.6648004  1.587073 0.034828249 0.2699189 0.2474639  648
## M00009      22       0.4504911  1.370023 0.100238663 0.6214797 0.5697777 3514
## M00104      10      -0.5876900 -1.346806 0.125441696 0.6481154 0.5941975  961
##                          leading_edge
## M00001 tags=54%, list=23%, signal=42%
## M00002 tags=55%, list=11%, signal=49%
## M00035 tags=50%, list=12%, signal=44%
## M00938  tags=40%, list=5%, signal=38%
## M00009 tags=50%, list=28%, signal=36%
## M00104  tags=50%, list=8%, signal=46%
##                                                        core_enrichment
## M00001 5214/3101/2821/7167/2597/5230/2023/5223/5315/3099/5232/2027/5211
## M00002                                    7167/2597/5230/2023/5223/5315
## M00035                                           875/1789/191/1788/1786
## M00938                                             6241/7298/4830/1841
## M00009            3418/50/4190/3419/2271/3421/55753/3417/1431/6389/4191
## M00104                                     6342/10998/1581/3295/8309
```

Visualize enriched KEGG pathways

```
####### Visualize enriched KEGG pathways #########

## To view the KEGG pathway,use the browseKEGG function,
#which will open a web browser and highlight enriched genes.

browseKEGG(kk, 'hsa04110')

###use the pathview() function from the pathview to visualize enriched KEGG
##pathways identified by the clusterProfiler package

library("pathview")
```

```
## Warning: package 'pathview' was built under R version 4.0.3
```

```
## ##############################################################################
## Pathview is an open source software package distributed under GNU General
## Public License version 3 (GPLv3). Details of GPLv3 is available at
## http://www.gnu.org/licenses/gpl-3.0.html. Particullary, users are required to
## formally cite the original Pathview paper (not just mention it) in publications
## or products. For details, do citation("pathview") within R.
##
## The pathview downloads and uses KEGG data. Non-academic uses may require a KEGG
## license agreement (details at http://www.kegg.jp/kegg/legal.html).
## ##############################################################################
```

```
hsa04110 <- pathview(gene.data  = geneList,
                     pathway.id = "hsa04110",
                     species    = "hsa",
                     limit      = list(gene=max(abs(geneList)), cpd=1))
```

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

```
## Info: Working in directory C:/Users/simar/OneDrive/Desktop/Practice/RNAseq_using_DEseq2
```

```
## Info: Writing image file hsa04110.pathview.png
```

Check the image file hsa04110.pathview.png in your working directory, for the KEGG pathway image.