Untitled

Saudat

25/05/2021

Install and load required packages

```r
#BiocManager::install("clusterProfiler", version = "3.8")
#BiocManager::install("pathview")
#BiocManager::install("enrichplot")
#BiocManager::install("ggnewscale")
#BiocManager::install("europepmc")
BiocManager::install("pathview")

library(clusterProfiler)
library(enrichplot)
# we use ggplot2 to add x axis labels (ex: ridgeplot)
library(ggplot2)
library(pathview)
```

Annotations

I'm using *D melanogaster* data, so I install and load the annotation "org.Dm.eg.db" below. See all annotations available here: http://bioconductor.org/packages/release/BiocViews.html#____OrgDb (there are 19 presently available).

```r
# SET THE DESIRED ORGANISM HERE
organism = "org.Dm.eg.db"
#BiocManager::install(organism, character.only = TRUE)
library(organism, character.only = TRUE)
organism = org.Dm.eg.db
```

#Prepare Input

```r
# reading in data from deseq2
df = read.csv("drosphila_example_de.csv", header=TRUE)
# we want the log2 fold change
original_gene_list <- df$log2FoldChange
# name the vector
names(original_gene_list) <- df$X
# 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)
```

Gene Set Enrichment function

```
gse <- gseGO(geneList=gene_list,
             ont ="ALL",
             keyType = "ENSEMBL",
             nPerm = 10000,
             minGSSize = 3,
             maxGSSize = 800,
             pvalueCutoff = 0.05,
             verbose = TRUE,
             OrgDb = organism,
             pAdjustMethod = "none")
```

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

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

```
## Warning in .GSEA(geneList = geneList, exponent = exponent, minGSSize =
## minGSSize, : We do not recommend using nPerm parameter incurrent and future
## releases
```

```
## Warning in fgsea(pathways = geneSets, stats = geneList, nperm = nPerm, minSize
## = minGSSize, : You are trying to run fgseaSimple. It is recommended to use
## fgseaMultilevel. To run fgseaMultilevel, you need to remove the nperm argument
## in the fgsea function call.
```

```
## Warning in preparePathwaysAndStats(pathways, stats, minSize, maxSize, gseaParam, : There are ties in
## The order of those tied genes will be arbitrary, which may produce unexpected results.
```

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

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

Output

##Table of results

```
head(gse)
```

```
##            ONTOLOGY         ID                                  Description
## GO:0031226       CC GO:0031226        intrinsic component of plasma membrane
## GO:0005887       CC GO:0005887         integral component of plasma membrane
## GO:0004888       MF GO:0004888    transmembrane signaling receptor activity
## GO:0007186       BP GO:0007186 G protein-coupled receptor signaling pathway
## GO:0004930       MF GO:0004930          G protein-coupled receptor activity
## GO:0019932       BP GO:0019932          second-messenger-mediated signaling
##            setSize enrichmentScore       NES       pvalue     p.adjust
## GO:0031226     466      -0.3990762 -1.620822 0.0001284852 0.0001284852
## GO:0005887     453      -0.4075260 -1.652176 0.0001292658 0.0001292658
## GO:0004888     305      -0.4181735 -1.637021 0.0001357036 0.0001357036
## GO:0007186     215      -0.5222521 -1.968425 0.0001435132 0.0001435132
## GO:0004930     111      -0.5825608 -2.022065 0.0001550868 0.0001550868
## GO:0019932     104      -0.5919972 -2.039066 0.0001561280 0.0001561280
##               qvalues rank                   leading_edge
## GO:0031226 0.09307353 1351  tags=22%, list=9%, signal=20%
## GO:0005887 0.09307353 1351  tags=22%, list=9%, signal=21%
## GO:0004888 0.09307353 1458 tags=21%, list=10%, signal=19%
## GO:0007186 0.09307353  885  tags=28%, list=6%, signal=27%
## GO:0004930 0.09307353 1016  tags=32%, list=7%, signal=30%
## GO:0019932 0.09307353  907  tags=26%, list=6%, signal=25%
##
## GO:0031226 FBgn0085420/FBgn0040507/FBgn0036278/FBgn0000037/FBgn0027843/FBgn0032006/FBgn0263916/FBgn0
## GO:0005887             FBgn0085420/FBgn0040507/FBgn0036278/FBgn0000037/FBgn0032006/FBgn0263916/FBgn0
## GO:0004888
## GO:0007186
## GO:0004930
## GO:0019932
```

##Dotplot

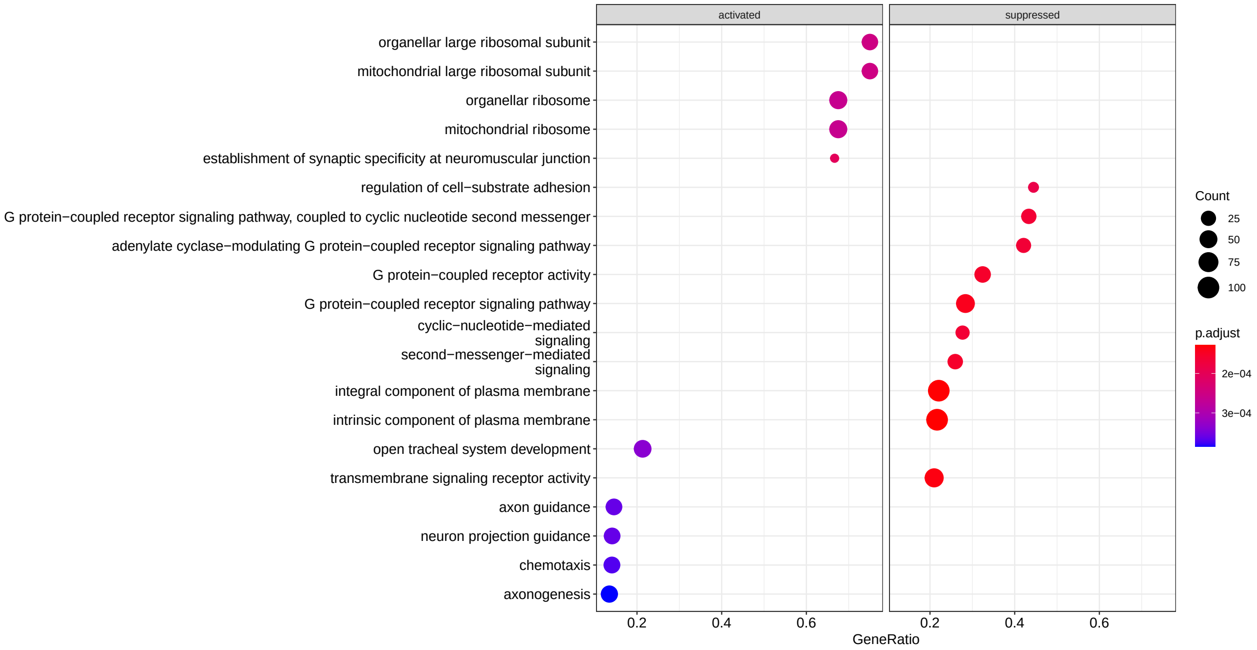
```r
require(DOSE)
```

```
## Loading required package: DOSE
```

```
## 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 for Disease
```

```r
dotplot(gse, showCategory=10, split=".sign") + facet_grid(.~.sign)
```

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

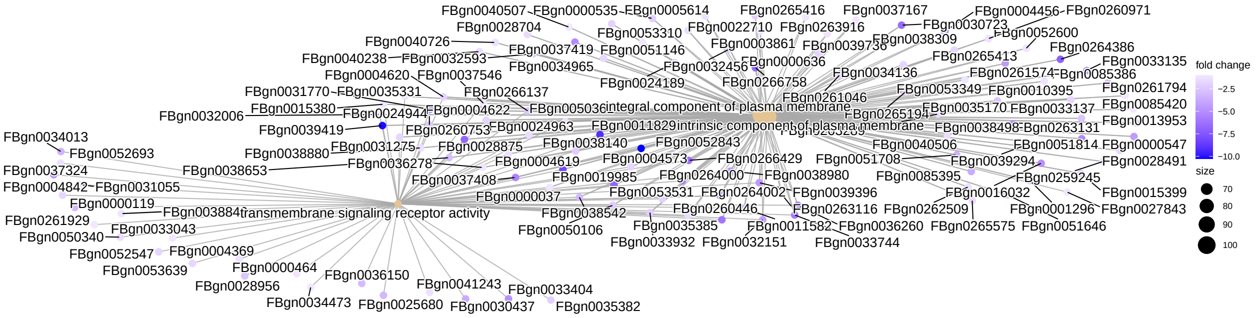


##Encrichment plot map:

```r
#emapplot(gse, showCategory = 10)
```

##Category Netplot

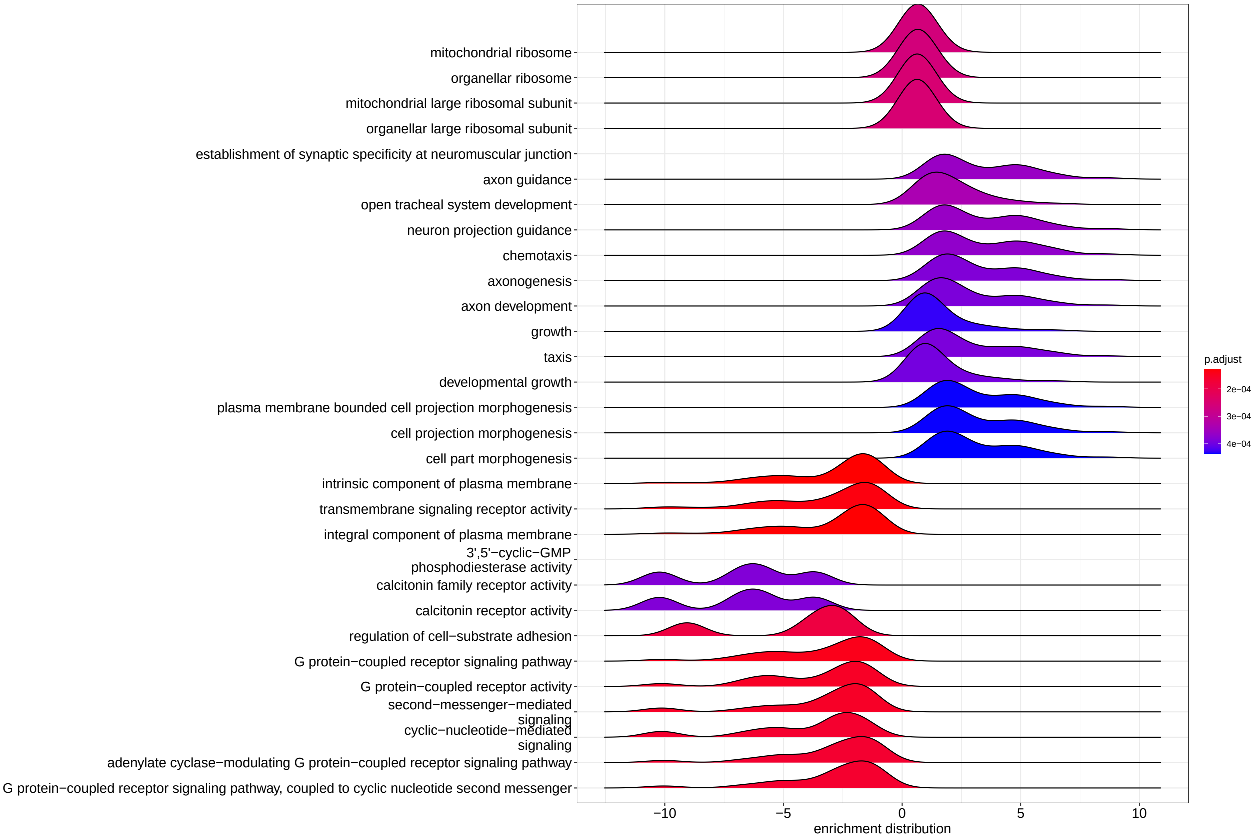
```r
# categorySize can be either 'pvalue' or 'geneNum'
cnetplot(gse, categorySize="pvalue", foldChange=gene_list, showCategory = 3)
```



## Ridgeplot Helpful to interpret up/down-regulated pathways.

```
ridgeplot(gse) + labs(x = "enrichment distribution")
```

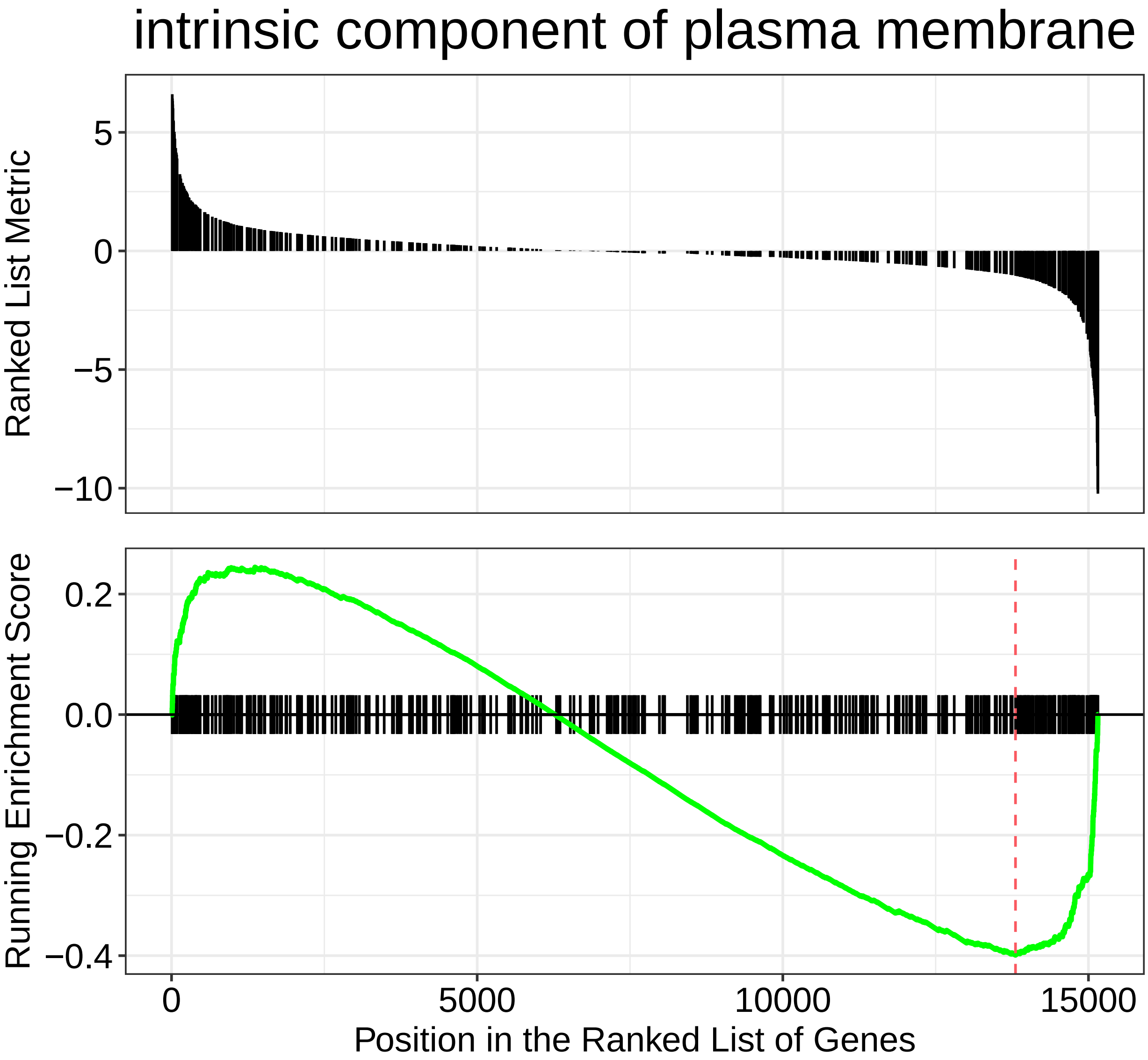
```
## Picking joint bandwidth of 0.77
```



## GSEA Plot

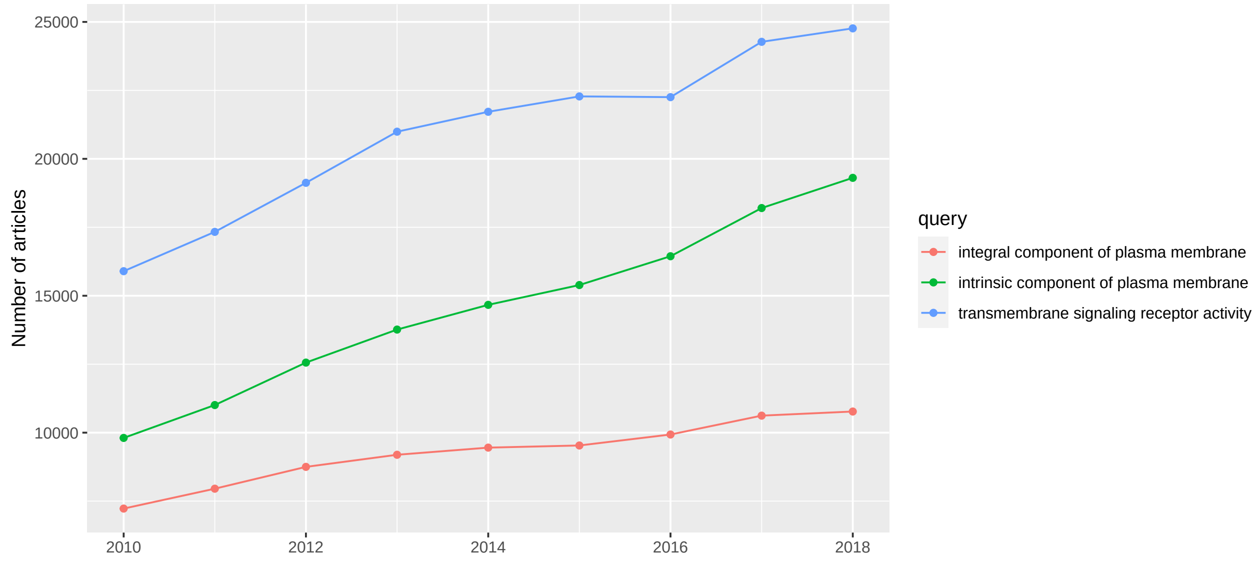
Traditional method for visualizing GSEA result.

```
# Use the `Gene Set` param for the index in the title, and as the value for geneSetId
gseaplot(gse, by = "all", title = gse$Description[1], geneSetID = 1)
```



## PubMed trend of enriched terms Plots the number/proportion of publications trend based on the query result from PubMed Central.

```
terms <- gse$Description[1:3]
pmcplot(terms, 2010:2018, proportion=FALSE)
```



# KEGG Gene Set Enrichment Analysis ## Prepare Input

```
# Convert gene IDs for gseKEGG function
# We will lose some genes here because not all IDs will be converted
ids<-bitr(names(original_gene_list), fromType = "ENSEMBL", toType = "ENTREZID", OrgDb=organism)
```

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

```
## Warning in bitr(names(original_gene_list), fromType = "ENSEMBL", toType =
## "ENTREZID", : 22.16% of input gene IDs are fail to map...
```

```
# remove duplicate IDS (here I use "ENSEMBL", but it should be whatever was selected as keyType)
dedup_ids = ids[!duplicated(ids[c("ENSEMBL")]),]
# Create a new dataframe df2 which has only the genes which were successfully mapped using the bitr fun
df2 = df[df$X %in% dedup_ids$ENSEMBL,]
# Create a new column in df2 with the corresponding ENTREZ IDs
df2$Y = dedup_ids$ENTREZID
# Create a vector of the gene unuiverse
kegg_gene_list <- df2$log2FoldChange
# Name vector with ENTREZ ids
names(kegg_gene_list) <- df2$Y
# omit any NA values
kegg_gene_list<-na.omit(kegg_gene_list)
# sort the list in decreasing order (required for clusterProfiler)
kegg_gene_list = sort(kegg_gene_list, decreasing = TRUE)
```

```
kegg_organism = "dme"
kk2 <- gseKEGG(geneList     = kegg_gene_list,
               organism     = kegg_organism,
               nPerm        = 10000,
               minGSSize    = 3,
               maxGSSize    = 800,
               pvalueCutoff = 0.05,
               pAdjustMethod = "none",
               keyType       = "ncbi-geneid")
```

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

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

## preparing geneSet collections...

## GSEA analysis...

## Warning in .GSEA(geneList = geneList, exponent = exponent, minGSSize =
## minGSSize, : We do not recommend using nPerm parameter incurrent and future
## releases

## Warning in fgsea(pathways = geneSets, stats = geneList, nperm = nPerm, minSize
## = minGSSize, : You are trying to run fgseaSimple. It is recommended to use
## fgseaMultilevel. To run fgseaMultilevel, you need to remove the nperm argument
## in the fgsea function call.

## Warning in preparePathwaysAndStats(pathways, stats, minSize, maxSize, gseaParam, : There are ties in
## The order of those tied genes will be arbitrary, which may produce unexpected results.

## Warning in preparePathwaysAndStats(pathways, stats, minSize, maxSize,
## gseaParam, : There are duplicate gene names, fgsea may produce unexpected
## results.

## leading edge analysis...

## done...
```

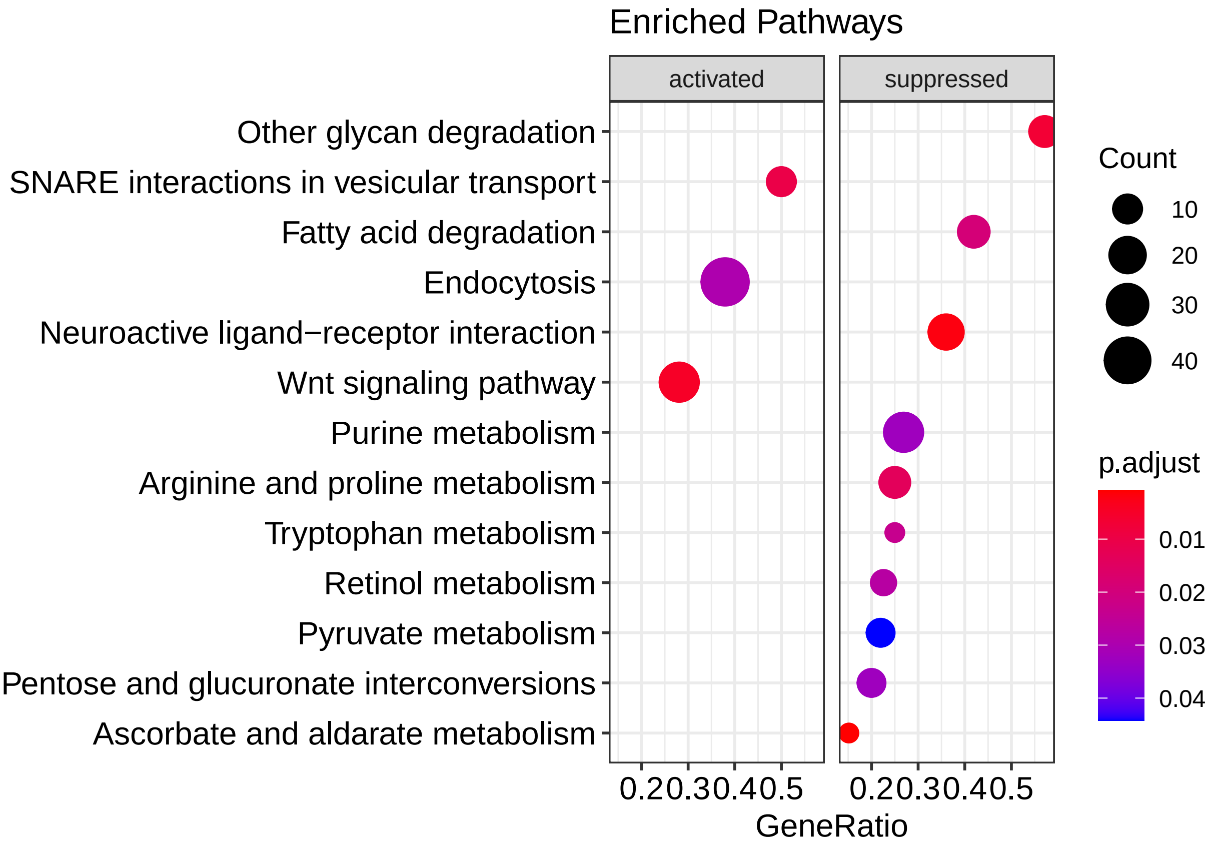
```r
head(kk2, 10)
```

```
##                 ID                               Description setSize
## dme00053 dme00053         Ascorbate and aldarate metabolism      33
## dme04080 dme04080   Neuroactive ligand-receptor interaction      50
## dme04310 dme04310                     Wnt signaling pathway      89
## dme00511 dme00511                  Other glycan degradation      21
## dme04130 dme04130 SNARE interactions in vesicular transport      20
## dme00330 dme00330           Arginine and proline metabolism      48
## dme00071 dme00071                    Fatty acid degradation      31
## dme00380 dme00380                     Tryptophan metabolism      20
## dme00830 dme00830                        Retinol metabolism      31
## dme04144 dme04144                               Endocytosis     116
##          enrichmentScore       NES      pvalue    p.adjust   qvalues rank
## dme00053      -0.6527139 -1.845014 0.000855432 0.000855432 0.1062537   51
## dme04080      -0.5753988 -1.759158 0.001997337 0.001997337 0.1240451 1183
## dme04310       0.4372485  1.604369 0.004685777 0.004685777 0.1940076 2043
## dme00511      -0.6770366 -1.737394 0.006877094 0.006877094 0.2135519 1371
## dme04130       0.6308289  1.695692 0.010321101 0.010321101 0.2563979 2521
## dme00330      -0.5315110 -1.615484 0.013664389 0.013664389 0.2828768 1978
## dme00071      -0.5753450 -1.601933 0.019118154 0.019118154 0.3371340 2579
## dme00380      -0.6327135 -1.604785 0.023573201 0.023573201 0.3371340 1250
## dme00830      -0.5591285 -1.556781 0.027385463 0.027385463 0.3371340 1749
## dme04144       0.3588310  1.372592 0.029507229 0.029507229 0.3371340 2930
##                            leading_edge
## dme00053  tags=15%, list=0%, signal=15%
## dme04080  tags=36%, list=9%, signal=33%
## dme04310 tags=28%, list=16%, signal=24%
## dme00511 tags=57%, list=11%, signal=51%
## dme04130 tags=50%, list=20%, signal=40%
## dme00330 tags=25%, list=16%, signal=21%
## dme00071 tags=42%, list=21%, signal=33%
## dme00380 tags=25%, list=10%, signal=23%
## dme00830 tags=23%, list=14%, signal=19%
## dme04144 tags=37%, list=23%, signal=29%
##
## dme00053
## dme04080
## dme04310
## dme00511
## dme04130
## dme00330
## dme00071
## dme00380
## dme00830
## dme04144 44921/42852/32791/44920/50022/39572/41551/40036/42841/42160/41917/47408/44263/37218/40250/44
```

Dotplot

```
dotplot(kk2, showCategory = 10, title = "Enriched Pathways" , split=".sign") + facet_grid(.~.sign)
```

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

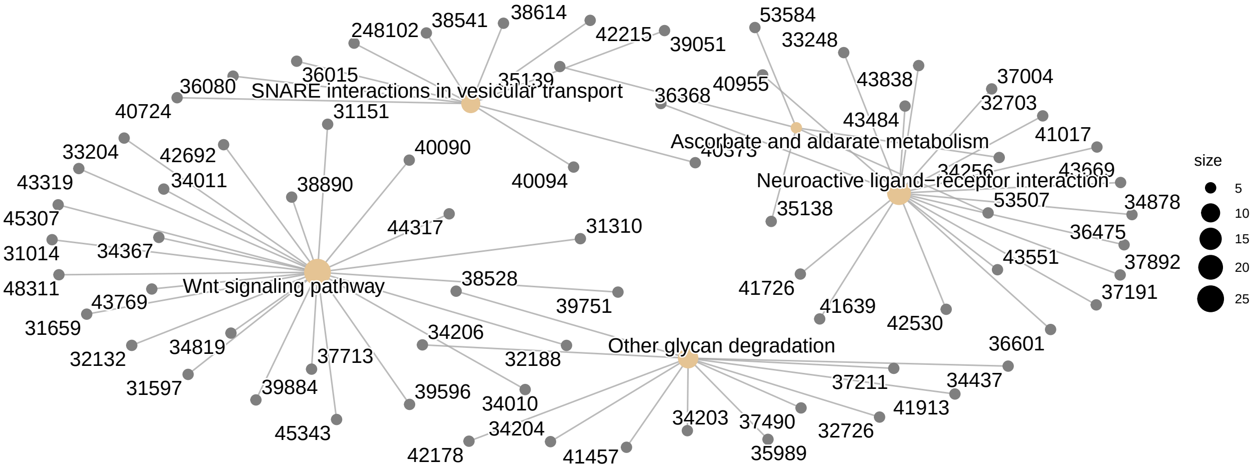


## Encrichment plot map:

```
#emapplot(kk2)
```

Category Netplot:

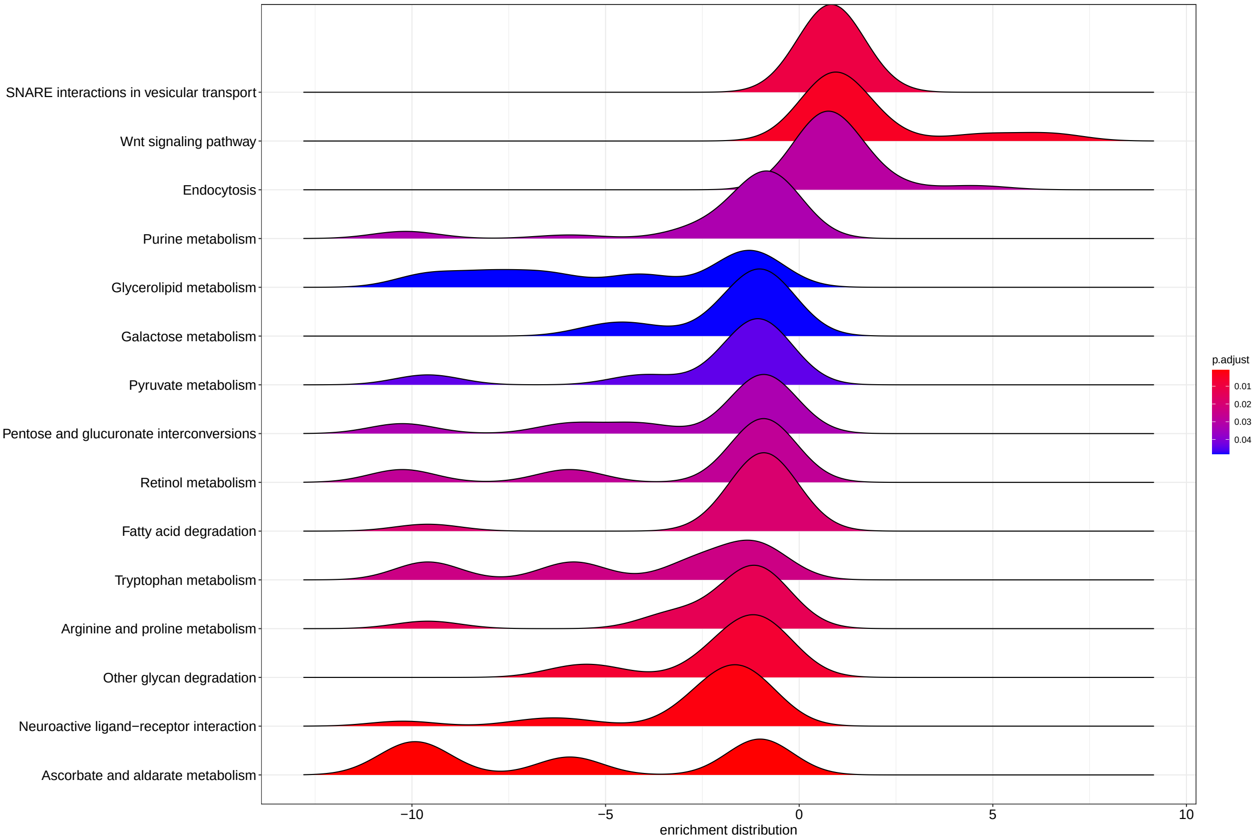
```r
# categorySize can be either 'pvalue' or 'geneNum'
cnetplot(kk2, categorySize="pvalue", foldChange=gene_list)
```



## Ridgeplot Helpful to interpret up/down-regulated pathways.

```r
ridgeplot(kk2) + labs(x = "enrichment distribution")
```

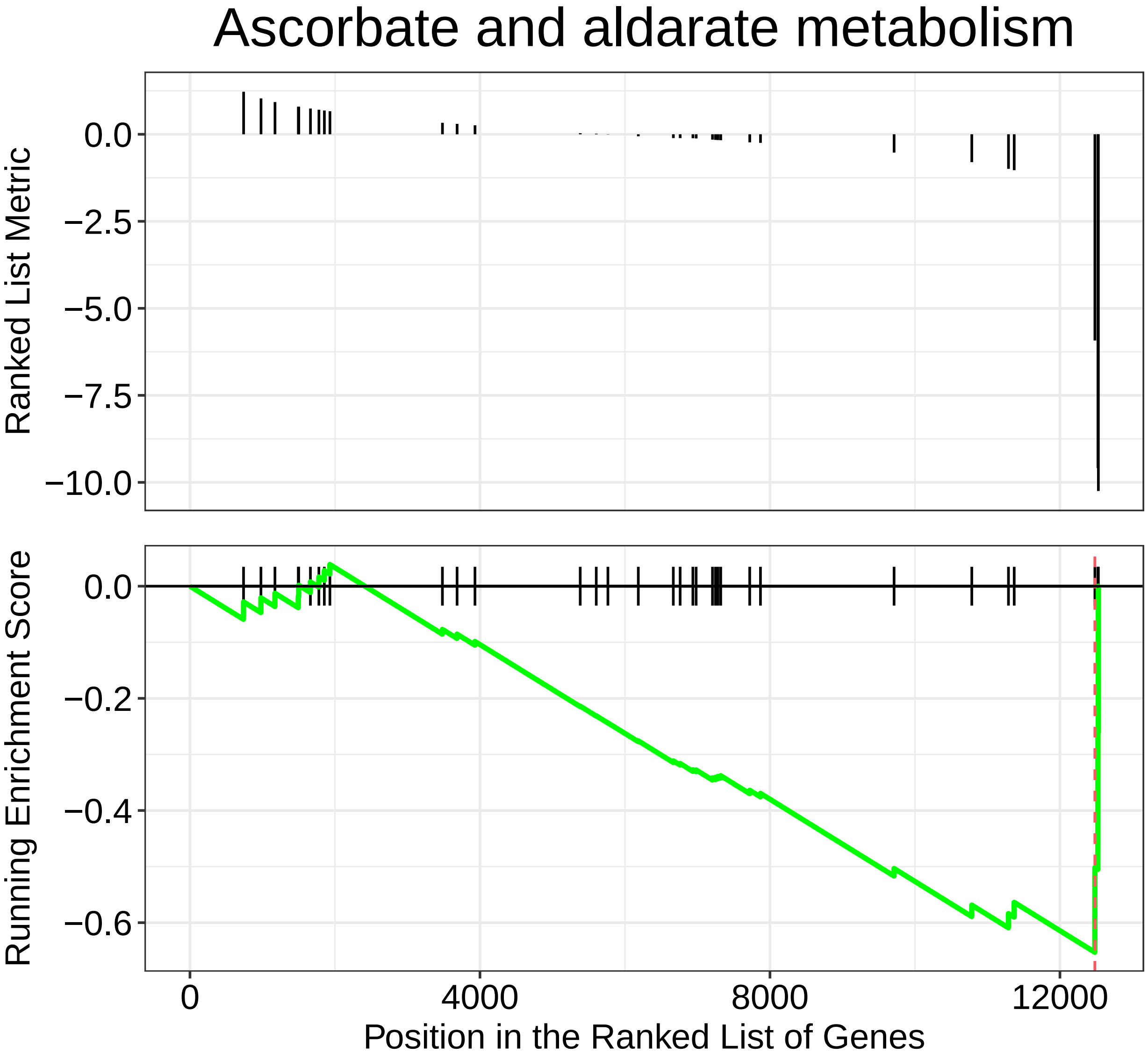
```
## Picking joint bandwidth of 0.849
```



# GSEA Plot

Traditional method for visualizing GSEA result.

```r
# Use the `Gene Set` param for the index in the title, and as the value for geneSetId
gseaplot(kk2, by = "all", title = kk2$Description[1], geneSetID = 1)
```



#Pathview

```r
# Produce the native KEGG plot (PNG)
dme <- pathview(gene.data=kegg_gene_list, pathway.id="dme04130", species = kegg_organism)
# Produce a different plot (PDF) (not displayed here)
dme <- pathview(gene.data=kegg_gene_list, pathway.id="dme04130", species = kegg_organism, kegg.native =
```