Gene ontology enrichment for RNAseq analysis of strain LU439 samples L2, L3 and strain MAC101p6 (ClusterProfiler package)

Julia Lienard

2024-06-01

Use of the complete genome annotation with Blast2GO, Interpro, Mycobacteria database from LU439T1 to map the RNAseq data of all samples, including MAC101 to be able to compare the data. Common differentially expressed genes (DEGs) are the one shared between the 3 pairs or at least between the two strains (called extended common DEGs)

Required packages

```
library(tidyverse)
```

```
-- Attaching core tidyverse packages ------------------------ tidyverse 2.0.0 --
v dplyr     1.1.4     v readr     2.1.5
v forcats   1.0.0     v stringr   1.5.1
v ggplot2   3.5.1     v tibble    3.2.1
v lubridate 1.9.3     v tidyr     1.3.1
v purrr     1.0.2
-- Conflicts ------------------------------------------ tidyverse_conflicts() --
x dplyr::filter() masks stats::filter()
x dplyr::lag()    masks stats::lag()
i Use the conflicted package (<http://conflicted.r-lib.org/>) to force all conflicts to becor
```

```
if (!require("BiocManager", quietly = TRUE))
    install.packages("BiocManager")

BiocManager::install("clusterProfiler")
```

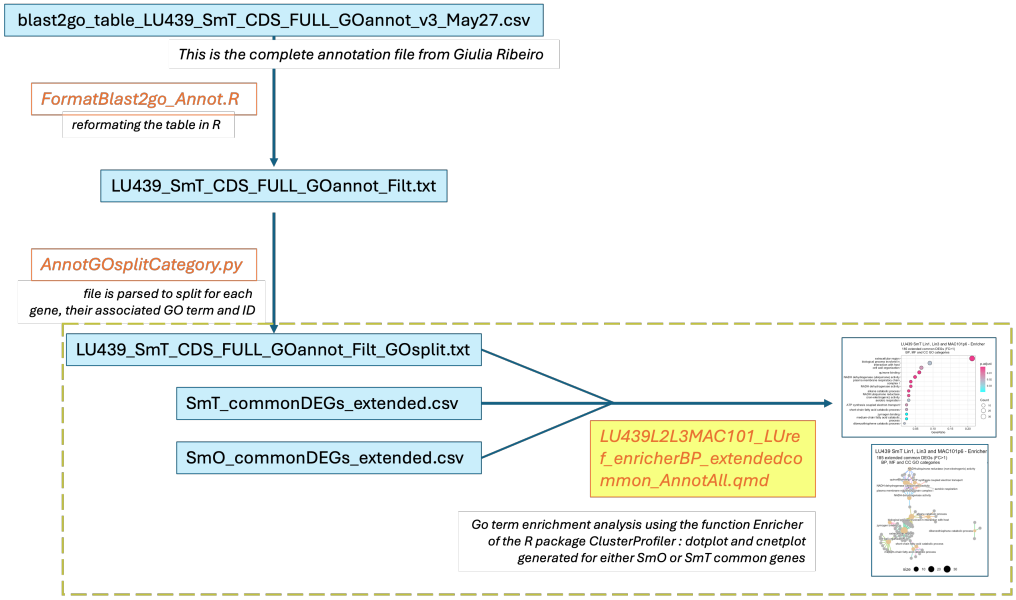
```
Bioconductor version 3.19 (BiocManager 1.30.25), R 4.4.1 (2024-06-14)

Warning: package(s) not installed when version(s) same as or greater than current; use
  `force = TRUE` to re-install: 'clusterProfiler'

Old packages: 'boot', 'foreign', 'MASS', 'nlme', 'ragg', 'RcppArmadillo',
  'survival'
```

```
library(clusterProfiler)
```

```
clusterProfiler v4.12.6 Learn more at https://yulab-smu.top/contribution-knowledge-mining/

Please cite:

S Xu, E Hu, Y Cai, Z Xie, X Luo, L Zhan, W Tang, Q Wang, B Liu, R Wang,
W Xie, T Wu, L Xie, G Yu. Using clusterProfiler to characterize
multiomics data. Nature Protocols. 2024, doi:10.1038/s41596-024-01020-z

Attaching package: 'clusterProfiler'

The following object is masked from 'package:purrr':

    simplify

The following object is masked from 'package:stats':

    filter
```

```
library(enrichplot)
```

Principle of the script



Loading the data

```
library(tidyverse)
# Opening the formatted genome annotation file (output table from
 ↪  FormatBlast2go.R):
Ref_GO <-
 ↪  read.delim("~/Desktop/Master/BINP39/RNAseq_visualization/LU439/1_data/1_genomeLU439T1_Bla
 ↪  header = TRUE, sep = "\t")

# Opening the RNAseq results analysis: common DEGs identified by Giulia
 ↪  Ribeiro for SmT samples:
Common_SmT_DEG <-
 ↪  read.table("~/Desktop/Master/BINP39/RNAseq_visualization/LU439_L2_L3_MAC101p6/REF_LU439T
 ↪  header = TRUE, sep = "\t")

# Opening the RNAseq results analysis: common DEGs identified by Giulia
 ↪  Ribeiro for SmO samples:
Common_SmO_DEG <-
 ↪  read.delim("~/Desktop/Master/BINP39/RNAseq_visualization/LU439_L2_L3_MAC101p6/REF_LU439T
```

Preparing the data

```
# Removing unwanted character in front of the pgap annotation
Common_SmT_DEG$ID <- gsub("cds-", "", as.character(Common_SmT_DEG$ID))
colnames(Common_SmT_DEG)[1] <- "pgap_ID"

Common_SmO_DEG$ID <- gsub("cds-", "", as.character(Common_SmO_DEG$ID))
colnames(Common_SmO_DEG)[1] <- "pgap_ID"

# Preparing the variables TERM2GENE, TERM2NAME, required for enricher
 ↪  function from the reference annotation filtered for Biological Process
 ↪  (BP)

TERM2GENE_BP <- Ref_GO |> filter(GO_category == "Biological Process") |>
  select(GO_ID, pgap_ID)

TERM2NAME_BP <- Ref_GO |> filter(GO_category == "Biological Process") |>
  select(GO_ID, GO_name)

Ref_GO_BP <- Ref_GO |> filter(GO_category == "Biological Process")
```

Extracting list of DEGs for enrichment analysis

Set to : Log2FoldChange $>$ 1 (all DEGs in the list have already p value $<$ 0.05)

```
# SmT
SmT_up_1 <- subset(Common_SmT_DEG, select = c(1))
names(SmT_up_1) <- NULL
SmT_up_1 <- SmT_up_1[,1]

# SmO
SmO_up_1 <- subset(Common_SmO_DEG, select = c(1))
names(SmO_up_1) <- NULL
SmO_up_1 <- SmO_up_1[,1]
```

Enrichment Biological Processes GO terms -SmT

```
library(clusterProfiler)
# Enrichment analysis
smt_BP <- enricher(gene = SmT_up_1,
                           pvalueCutoff = 0.05,
                           qvalueCutoff = 0.05,
                           pAdjustMethod = "fdr",
                           universe = Ref_GO_BP$pgap_ID ,
                           minGSSize = 5,
                           maxGSSize = 500,
                           TERM2GENE = TERM2GENE_BP,
                           TERM2NAME = TERM2NAME_BP)
```

Dotplot graph for enrichment analysis

```
# Plotting
library(enrichplot)

# setting the parameters for the plot
options(enrichplot.colours = c("violetred2","cyan"), font.size =10)
layout.params = list(layout = "kk")
cex.params = list(category_node = 1.5)

dotplot(smt_BP, showCategory=15, font.size = 10, label_format = 40) +
labs(title ="LU439 SmT Lin1, Lin3 and MAC101p6 - Enricher",
     subtitle = " 185 extended common DEGs (FC>1)
     Biological Process")
```

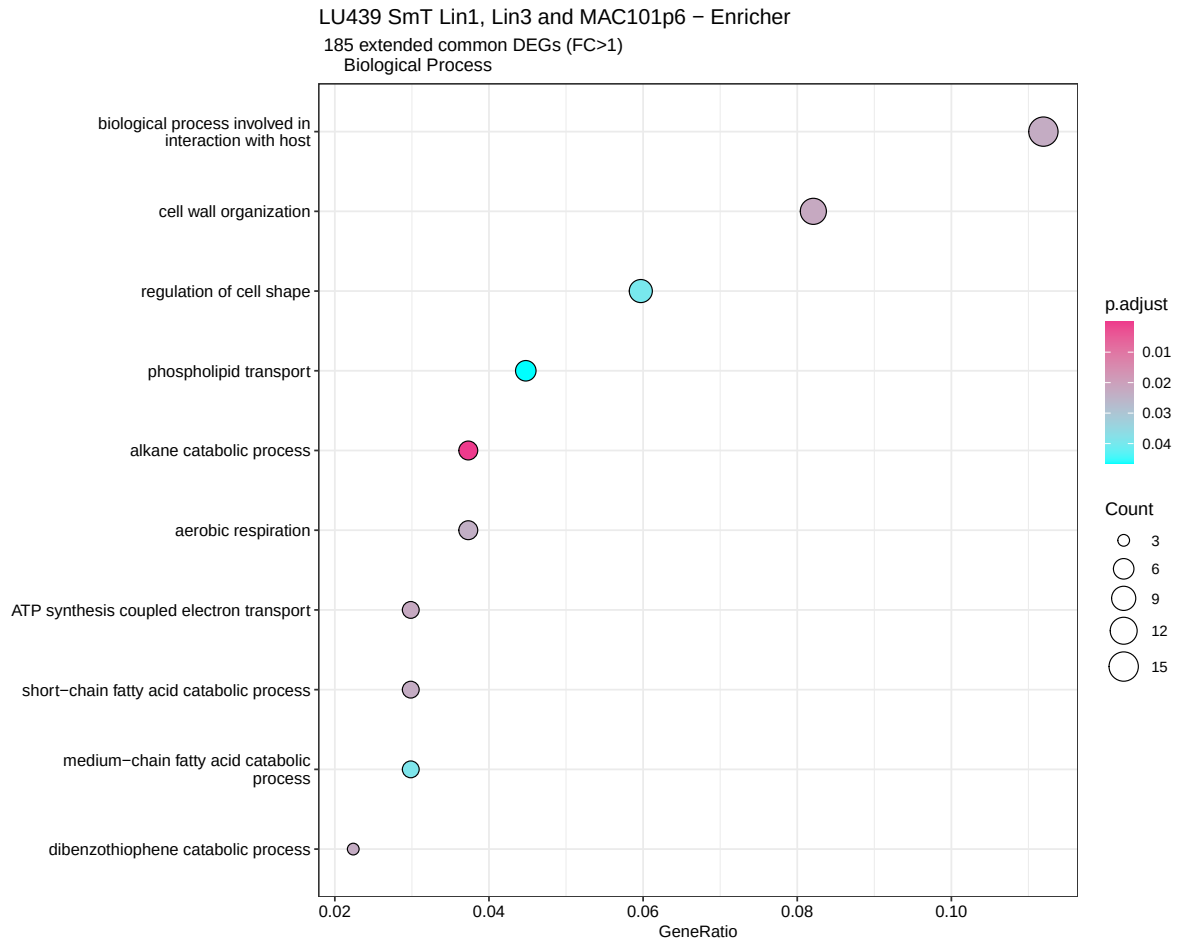


Figure 1: Dotplot for Enrichment Biological Processes GO terms -SmT

```
ggsave("LU439TL1L3MAC101p6_refLUT1_EnricherBPExtended_FC1_dotplot.pdf",
       plot = last_plot(),
       width = 6,
       height = 6,
       dpi = 300)
```

Cnetplot for clustering the enriched GO terms

```
cex.params = list(category_label = 0.5, gene_label = 0.4, font_face = 2)
cnetplot(smt_BP,
         node_label="category",
```

```
         showCategory=15,
         layout = "kk",
         colorEdge = TRUE,
         cex_category =0.5,
         cex.params = cex.params) +
labs(title ="LU439 SmT Lin1, Lin3 and MAC101p6 - Enricher",
     subtitle = " 185 extended common DEGs (FC>1)
     Biological Process") +
  guides(
    category = guide_colourbar(position = "top"),
    size   = guide_legend(position = "bottom")
  ) +
  theme(legend.position = "right")
```

```
Warning in cnetplot.enrichResult(x, ...): Use 'color.params = list(edge = your_value)' instea
 The colorEdge parameter will be removed in the next version.
```

```
Warning in cnetplot.enrichResult(x, ...): Use 'cex.params = list(category_node = your_value)
 The cex_category parameter will be removed in the next version.
```

```
Warning: Removed 50 rows containing missing values or values outside the scale range
(`geom_text_repel()`).
```

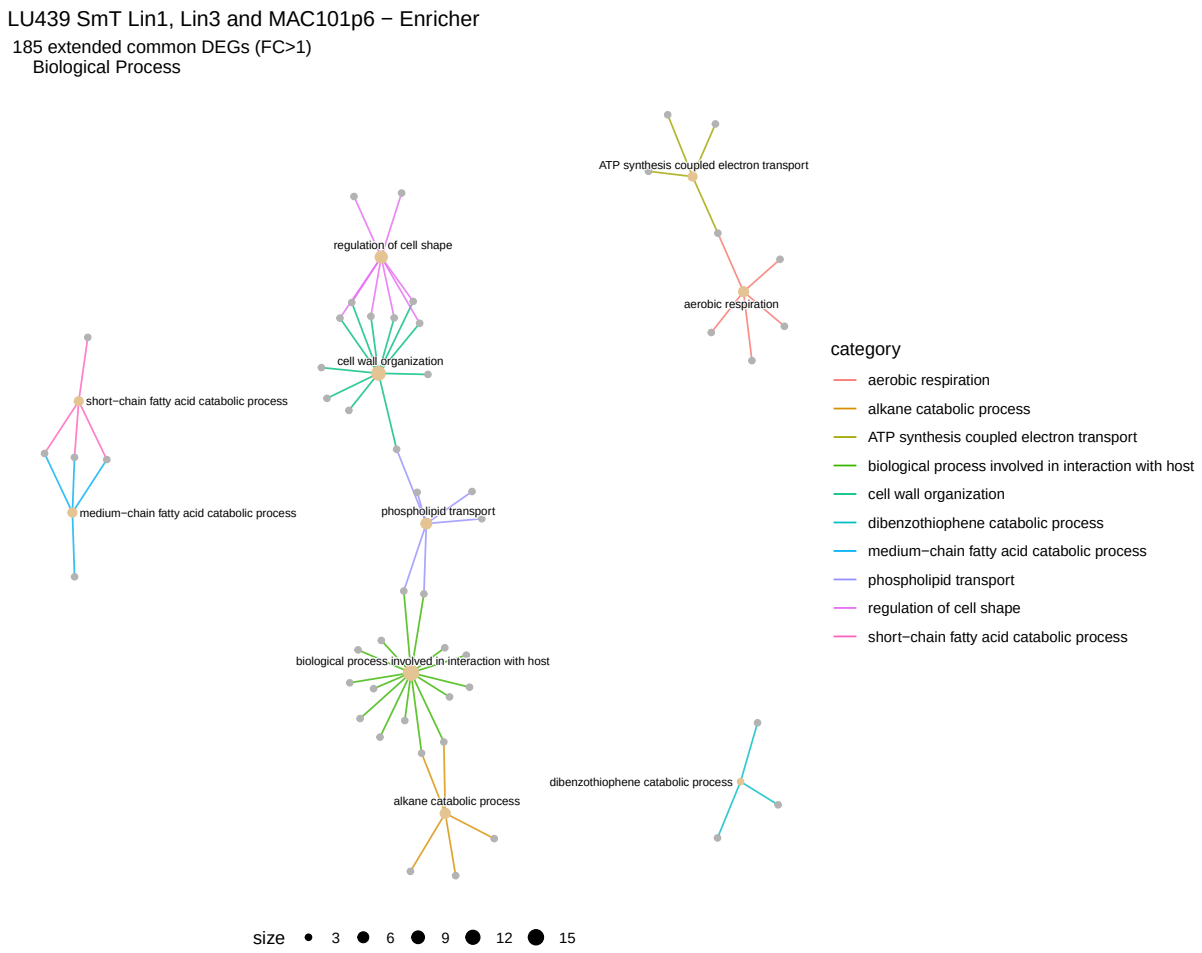


Figure 2: Cnetplot for clustering of the Enrichment Biological Processes GO terms -SmT

```
ggsave("LU439TL1L3MAC101p6_refLUT1_EnricherBPExtended_FC1_cnetplot.pdf", plot
↪  = last_plot(),
       width = 6,
       height = 6,
       dpi = 300)
```

```
Warning: Removed 50 rows containing missing values or values outside the scale range
(`geom_text_repel()`).
```

Enrichment Biological Processes GO terms -SmO

```
# Enrichment analysis
smo_BP <- enricher(gene = SmO_up_1,
                           pvalueCutoff = 0.05,
                           qvalueCutoff = 0.05,
                           pAdjustMethod = "fdr",
                           universe = Ref_GO_BP$pgap_ID ,
                           minGSSize = 5,
                           maxGSSize = 500,
                           TERM2GENE = TERM2GENE_BP,
                           TERM2NAME = TERM2NAME_BP)
```

Dotplot graph for enrichment analysis

```
# Plotting
dotplot(smo_BP, showCategory=12, font.size = 10, label_format = 40) +
labs(title ="LU439 SmO Lin2, Lin3 and MAC101p6 - Enricher",
     subtitle = " 181 extended common DEGs (FC>1)
     Biological Process")
```

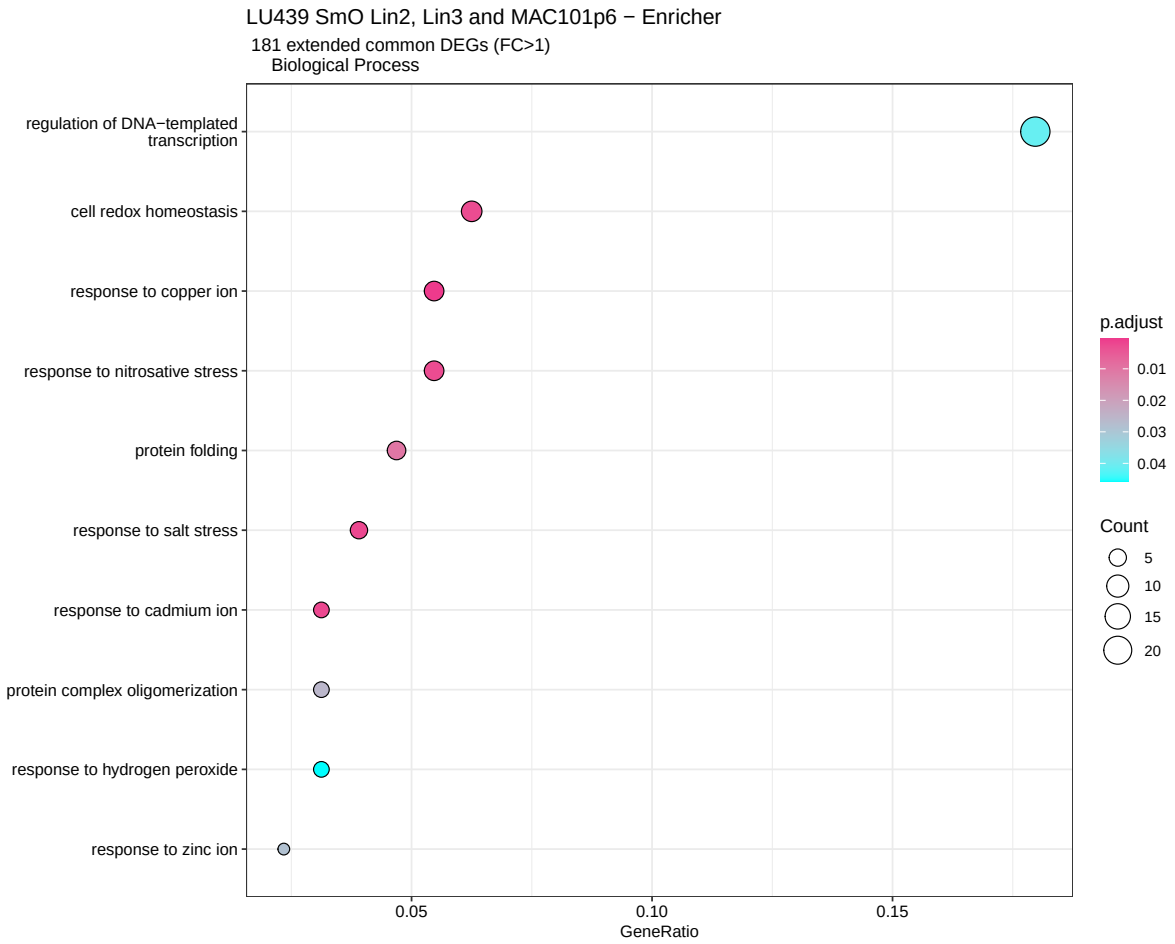


Figure 3: Dotplot for Enrichment Biological Processes GO terms -SmO

```
ggsave("LU439O2O3MAC101p6_refLUT1_EnricherBPExtended_FC1_dotplot.pdf", plot =
↪  last_plot(),
       width = 5.5,
       height = 6,
       dpi = 300)
```

Cnetplot for clustering the enriched GO terms

```
cex.params = list(category_label = 0.5, gene_label = 0.4, font_face = 2)
cnetplot(smo_BP,
         node_label="category",
```

```
         showCategory=15,
         layout = "kk",
         colorEdge = TRUE,
         cex_category = 0.5,
         cex.params = cex.params) +
labs(title ="LU439 SmO Lin2, Lin3 and MAC101p6 - Enricher",
     subtitle = " 181 extended common3 DEGs (FC>1)
     Biological Process") +
  guides(
    category = guide_colourbar(position = "top"),
    size   = guide_legend(position = "bottom")
  ) +
  theme(legend.position = "right")
```

```
Warning in cnetplot.enrichResult(x, ...): Use 'color.params = list(edge = your_value)' instea
 The colorEdge parameter will be removed in the next version.
```

```
Warning in cnetplot.enrichResult(x, ...): Use 'cex.params = list(category_node = your_value)
 The cex_category parameter will be removed in the next version.
```

```
Warning: Removed 44 rows containing missing values or values outside the scale range
(`geom_text_repel()`).
```

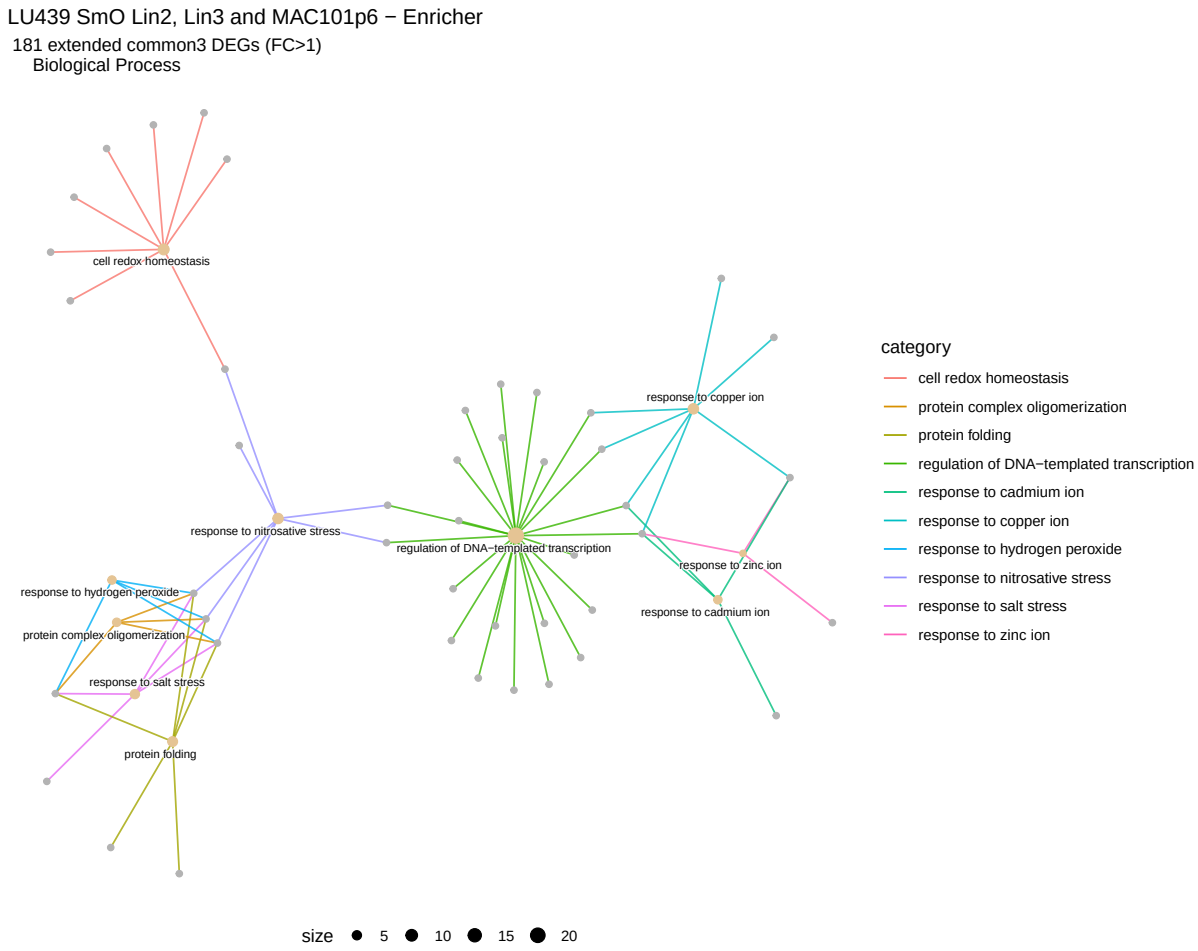


Figure 4: Cnetplot for clustering of the Enrichment Biological Processes GO terms -SmO

```
ggsave("LU439O2O3MAC101p6_refLUT1_EnricherBPExtended_FC1_cnetplot.pdf", plot
↪  = last_plot(),
       width = 5,
       height = 5,
       dpi = 300)
```

```
Warning: Removed 44 rows containing missing values or values outside the scale range
(`geom_text_repel()`).
```