Supplemental File of

clusterProfiler 4.0: A universal enrichment tool for functional and comparative study

Tianzhi Wu§, Erqiang Hu§, Meijun Chen, Pingfan Guo, Zehan Dai, Tingze Feng, Shuangbin Xu, Lang Zhou, Wenli Tang, Li Zhan, Xiaocong Fu, Shanshan Liu, Xiaochen Bo[*] and Guangchuang Yu[*]

[*]correspondence: Guangchuang Yu <gcyu1@smu.edu.cn> and Xiaochen Bo <boxc@bmi.ac.cn>

1 R packages that depends on clusterProfiler

The `clusterProfiler` library is one of the fundamental packages and it had been incorporated in more than twenty R packages (in CRAN or Bioconductor) to perform functional enrichment analysis for different topics, especially for cancer research.

```r
db <- utils::available.packages(repo=BiocManager::repositories())
pkgs <- tools::package_dependencies('clusterProfiler', db=db,
              which = c("Depends", "Imports"), reverse=TRUE)[[1]]
sort(pkgs)
```

```
##  [1] "AutoPipe"        "bioCancer"       "CEMiTool"        "CeTF"
##  [5] "conclus"         "DAPAR"           "debrowser"       "eegc"
##  [9] "enrichTF"        "esATAC"          "ExpHunterSuite"  "famat"
## [13] "fcoex"           "GDCRNATools"     "immcp"           "IRISFGM"
## [17] "maEndToEnd"      "MAGeCKFlute"     "methylGSA"       "miRspongeR"
## [21] "MoonlightR"      "multiSight"      "netboxr"         "PFP"
## [25] "recountWorkflow" "RNASeqR"         "RVA"             "signatureSearch"
## [29] "TCGAbiolinksGUI" "TCGAWorkflow"    "TimiRGeN"
```

Table S1: R packages that rely on clusterProfiler to perform functional analysis.

Package	Description
AutoPipe	Automated Transcriptome Classifier Pipeline: Comprehensive Transcriptome Analysis
bioCancer	Interactive Multi-Omics Cancers Data Visualization and Analysis
CEMiTool	Co-expression Modules identification Tool
CeTF	Coexpression for Transcription Factors using Regulatory Impact Factors and Partial Correlation and Information Theory analysis
conclus	ScRNA-seq Workflow CONCLUS - From CONsensus CLUSters To A Meaningful CONCLUSion
DAPAR	Tools for the Differential Analysis of Proteins Abundance with R
debrowser	Interactive Differential Expresion Analysis Browser
eegc	Engineering Evaluation by Gene Categorization (eegc)
enrichTF	Transcription Factors Enrichment Analysis
esATAC	An Easy-to-use Systematic pipeline for ATACseq data analysis
ExpHunterSuite	Package For The Comprehensive Analysis Of Transcriptomic Data
famat	Functional analysis of metabolic and transcriptomic data
fcoex	FCBF-based Co-Expression Networks for Single Cells
GDCRNATools	an R/Bioconductor package for integrative analysis of lncRNA, mRNA, and miRNA data in GDC
immcp	Candidate Prescriptions Discovery Based on Pathway Fingerprint
IRISFGM	Comprehensive Analysis of Gene Interactivity Networks Based on Single-Cell RNA-Seq
maEndToEnd	An end to end workflow for differential gene expression using Affymetrix microarrays
MAGeCKFlute	Integrative Analysis Pipeline for Pooled CRISPR Functional Genetic Screens
methylGSA	Gene Set Analysis Using the Outcome of Differential Methylation
miRspongeR	Identification and analysis of miRNA sponge interaction networks and modules
MoonlightR	Identify oncogenes and tumor suppressor genes from omics data
multiSight	Multi-omics Classification, Functional Enrichment and Network Inference analysis
netboxr	netboxr
PFP	Pathway Fingerprint Framework in R
recountWorkflow	recount workflow: accessing over 70,000 human RNA-seq samples with Bioconductor
RNASeqR	an R package for automated two-group RNA-Seq analysis workflow
RVA	RNAseq Visualization Automation
signatureSearch	Environment for Gene Expression Searching Combined with Functional Enrichment Analysis
TCGAbiolinksGUI	TCGAbiolinksGUI: A Graphical User Interface to analyze cancer molecular and clinical data
TCGAWorkflow	TCGA Workflow Analyze cancer genomics and epigenomics data using Bioconductor packages
TimiRGeN	Time sensitive microRNA-mRNA integration, analysis and network generation tool

2 Comparing clusterProfiler with other tools

Table S2: Comparing clusterProfiler with other tools

		Input and annotation						Method			Interpretation				
Software	Repo	Annotation	Supported organisms	ID conversion	Updated KEGG	External annotation data	Support GMT file	Algorithm	Selection of background set	Profile comparison	Output	Tidy interface	Support ggplot2	Visualization methods	Remove redundant terms
clusterProfiler	2	GO, KEGG, WikiPathways	plenty	Y	Y	Y	Y	ORA, GSEA	Y	Y	enrichResult, gseaResult, compareClusterResult (S4)	Y	Y	11	Y
DOSE	2	DisGeNE, DO, NCG	1	N	NA	N	N	ORA, GSEA	Y	N	enrichResult, gseaResult (S4)	Y	Y	11	Y
fgsea	2	NA	NA	Y	NA	Y	Y	ORA, GSEA	Y	N	data.table	Y	Y	2	N
ReactomePA	2	Reactome	7	N	NA	N	N	ORA, GSEA	Y	N	enrichResult, gseaResult (S4)	Y	Y	11	N
GOstats	2	NA	NA	N	NA	Y	N	ORA	Y	N	GOHyperGResult (S4)	N	N	1	N
enrichR	1	GO, KEGG, WikiPathways, BioCarta, Reactome, GEO, GeneSigDB, HPO, KEA, MSigDB, COVID-19 Related Gene Sets	5	Y	N	N	N	ORA	N	N	list	N	N	3	N
GSA	1	NA	NA	N	NA	Y	Y	Gene set analysis	N	N	GSA (S3)	N	N	1	N
globaltest	2	GO, KEGG, MSigDB, Anni	21	N	N	N	N	regression analysis	N	N	gt (S4)	N	N	2	N
gage	2	GO, KEGG	plenty	Y	Y	Y	N	GSEA	N	N	list	N	N	0	Y
gprofiler2	1	GO, KEGG, Reactome,WikiPathways, miRTarBase, TRANSFAC, Human Protein Atlas, protein complexes from CORUM, HPO	plenty	Y	N	Y	Y	ORA	N	Y	list	N	N	1	N
SPIA	2	KEGG	plenty	N	Y	N	N	Signaling Pathway Impact Analysis	Y	N	data.frame	Y	Y	1	N
safe	2	GO, KEGG, PFAM, Reactome	20	N	N	Y	N	ORA, Wilcoxon rank sum, Pearson’s chi-squared type statistic, t-statistic	N	N	SAFE (S4)	N	N	2	N
CePa	1	NCI_Nature, KEGG, BioCarta, Reactome	1	N	N	N	N	CePa	Y	N	cepa (S3)	N	N	3	N
GANPA	1	NA	NA	N	NA	Y	N	GANPA	N	N	.csv files	N	N	0	N
PADOG	2	KEGG	1	N	Y	Y	N	PADOG	N	Y	data.frame	Y	Y	0	N
ViSEAGO	2	GO	21	N	N	N	N	ORA, GSEA	N	Y	fgsea, enrich_GO_terms (S4)	N	N	2	N
GOGANPA	1	NA	NA	N	NA	Y	N	GO-Functional-Network-based Gene-Set-Analysis	N	N	.csv file	N	N	0	N
GSAR	2	NA	NA	N	NA	Y	N	two-sample Nnparametric multivariate test	N	N	list	N	N	0	N
netgsa	1	NA	NA	N	NA	Y	N	netgsa	N	N	list	N	N	3	N
sigPathway	2	NA	NA	N	NA	Y	N	GSEA, sigPathway	N	N	list	N	N	0	N
SeqGSEA	2	NA	NA	Y	NA	Y	Y	GSEA	N	N	SeqGeneSet (S4)	N	N	0	N
hypeR	2	MSigDB, KEGG, Reactome, MetaboAnalyst	11	N	N	Y	N	ORA, GSEA	Y	N	hyp (R6)	N	N	3	N
escape	2	MSigDB	11	N	N	Y	N	GSEA	N	N	data.frame	N	N	6	N
methylGSA	2	GO, KEGG, Reactome	1	Y	N	N	N	ORA, GSEA	N	N	data.frame	Y	Y	0	N
enrichTF	2	Transcription factor information	2	N	NA	N	N	t-tests, ORA	N	N	list	N	N	0	N
DEGraph	2	KEGG	plenty	N	Y	Y	N	t-tests	N	N	list	N	N	1	N
famat	2	GO, KEGG, Wikipathways, Reactome	1	N	Y	N	N	ORA	N	N	list	N	N	0	N

[1] Repo: 1 for CRAN and 2 for Bioconductor

[2] Supported organisms: ’NA’ for not applicable as there is no species annotation data internally supported by the package; ’plenty’ for hundreds or thousands species supported (mostly for KEGG and/or GO)

[3] Tidy interface: whether the output object can be processed directly using tidy tools such as dplyr

[4] Support ggplot2: whether the output object can be visualized directly using ggplot2 command

[5] Y for supported, N for not supported and NA for not applicable

3 Installation

To install `clusterProfiler` package, please enter the following command in R:

```r
if (!requireNamespace("BiocManager", quietly = TRUE))
    install.packages("BiocManager")
BiocManager::install("clusterProfiler")
```

To reproduce examples in this document, you need to install several extra packages:

```
install.packages(c("forcats", "ggplot2", "ggnewscale", "ggupset"))
BiocManager::install(c("org.Hs.eg.db", "enrichplot",
            "ChIPseeker", "TxDb.Hsapiens.UCSC.hg19.knownGene"))
```

4 Data sets

Several data sets were used in this document, including:

- `geneList` provided by the DOSE package
- `DE_GSE8057` provided by the clusterProfiler package
- `GSM1295076_CBX6_BF_ChipSeq_mergedReps_peaks.bed.gz` provided by the ChIPseeker package

The `geneList` was derived from the R package breastCancerMAINZ that contains 200 breast cancer samples, including 29 samples in grade I, 136 samples in grade II and 35 samples in grade III. The ratio of geometric mean of grade III samples versue geometric mean of grade I samples for each gene was computed. The `geneList` data set contains logarithm of these ratios (base 2).

The `DE_GSE8057` data set XXXXXXXXXX

The GSM1295076_CBX6_BF_ChipSeq_mergedReps_peaks.bed.gz file can be accessed via `ChIPseeker::getSampleFiles()[[4]]` or downloaded using the command `ChIPseeker::downloadGSMbedFiles("GSM1295076")`. Experimental design and protocols are provided in https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSM1295076.

5 Examples of using clusterProfiler

This session provides source codes to reproduce the figures presented in the manuscript.

5.1 GO enrichment analysis

```
library(clusterProfiler)
library(enrichplot)

## geneList for GSEA examples
data(geneList, package="DOSE")

## fold change > 2 as DE genes, for ORA examples
de <- names(geneList)[abs(geneList) > 2]


ego <- enrichGO(de, OrgDb = "org.Hs.eg.db", ont="BP", readable=TRUE)

## use simplify to remove redundant terms
ego2 <- simplify(ego, cutoff=0.7, by="p.adjust", select_fun=min)


## visualization
ego <- pairwise_termsim(ego)
ego2 <- pairwise_termsim(ego2)

p1 <- emapplot(ego, cex_label_category=.8, cex_line=.5) + coord_cartesian()
```

```
## Coordinate system already present. Adding new coordinate system, which will replace the existing one.
```

```
p2 <- emapplot(ego2, cex_label_category=.8, cex_line=.5) + coord_cartesian()
```

```
## Coordinate system already present. Adding new coordinate system, which will replace the existing one.
```

```
cowplot::plot_grid(p1, p2, labels=c("A", "B"), rel_widths=c(1, 1.2))
```

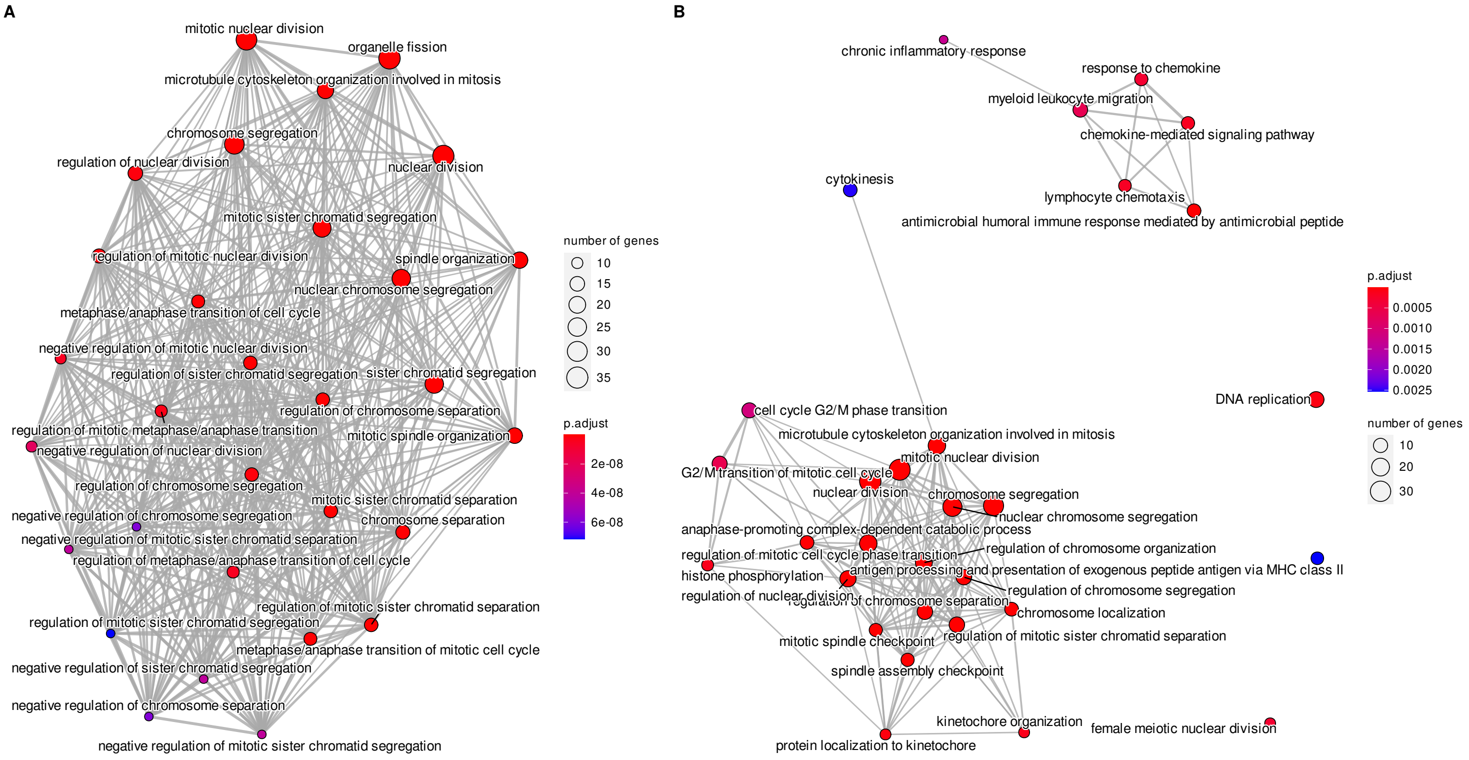

Fig. 1: Gene ontology enrichment analysis.

5.2 KEGG enrichment analysis

```
data(geneList, package="DOSE")
kk <- gseKEGG(geneList, organism = "hsa")

## sorted by absolute values of NES
kk2 <- arrange(kk, abs(NES))

## visualization
kp1 <- gseaplot2(kk2, 1:5, pvalue_table=F, base_size=14)
kp2 <- upsetplot(kk2, n=5)
cowplot::plot_grid(kp1, kp2, rel_widths=c(1, .5), labels=c("A", "B"))
```

5.3 Functional interpretation of genomic regions of interest

```
library(ChIPseeker)
## the file can be downloaded using `downloadGSMbedFiles("GSM1295076")`
file <- "GSM1295076_CBX6_BF_ChipSeq_mergedReps_peaks.bed.gz"
gr <- readPeakFile(file)

library(TxDb.Hsapiens.UCSC.hg19.knownGene)
TxDb <- TxDb.Hsapiens.UCSC.hg19.knownGene
genes <- seq2gene(gr, tssRegion=c(-1000, 1000), flankDistance = 3000, TxDb)

library(clusterProfiler)
## downloaded from https://maayanlab.cloud/Enrichr/geneSetLibrary?mode=text&libraryName=ENCODE_and_ChEA_Consens
encode <- read.gmt("ENCODE_and_ChEA_Consensus_TFs_from_ChIP-X.txt")
g <- bitr(genes, 'ENTREZID', 'SYMBOL', 'org.Hs.eg.db')
```

```
## Warning in bitr(genes, "ENTREZID", "SYMBOL", "org.Hs.eg.db"): 5.32% of input
## gene IDs are fail to map...
```

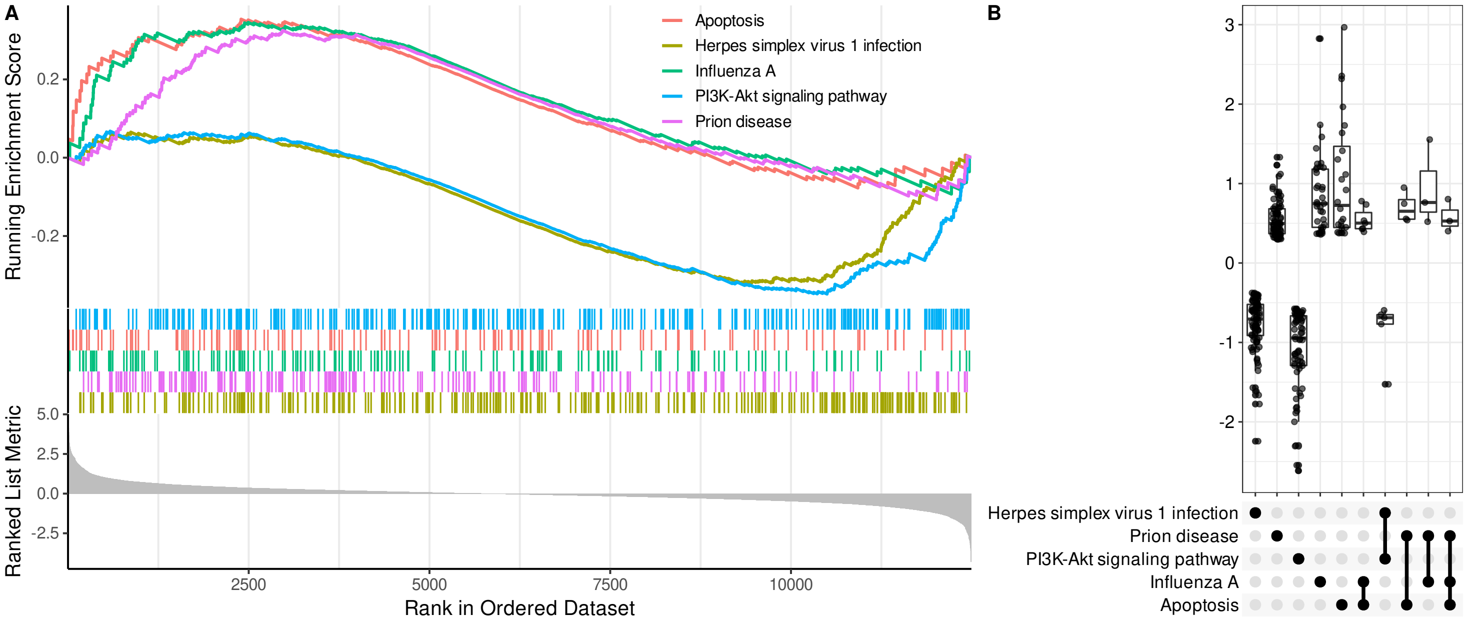

Fig. 2: KEGG pathway enrichment analysis.

```
x <- enricher(g$SYMBOL, TERM2GENE=encode)
enrichplot::cnetplot(x, cex_label_gene=0.6)
```

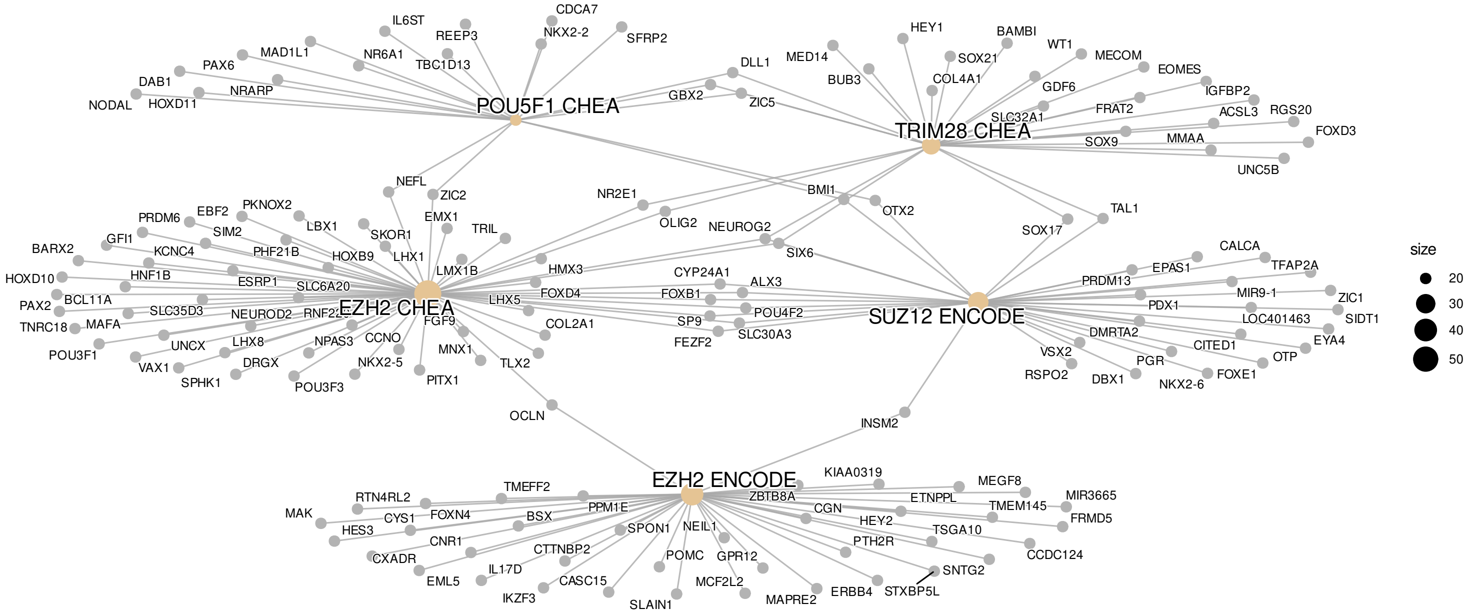

Fig. 3: Functional enrichment analysis of genomic regions of interest.

5.4 Comparison for different conditions

```
## downloaded from https://wikipathways-data.wmcloud.org/current/gmt/
gmt <- 'wikipathways-20210310-gmt-Homo_sapiens.gmt'
wp <- read.gmt.wp(gmt)

data(DE_GSE8057)

xx <- compareCluster(Gene~time+treatment, data=DE_GSE8057, fun = enricher,
            TERM2GENE=wp[,c("wpid", "gene")], TERM2NAME=wp[,c("wpid", "name")])
```

```
pp <- dotplot(xx, x="time") + facet_grid(~treatment) +
    aes(x=fct_relevel(time, c('0h', '2h', '6h', '24h'))) + xlab(NULL)
```

```
print(pp)
```

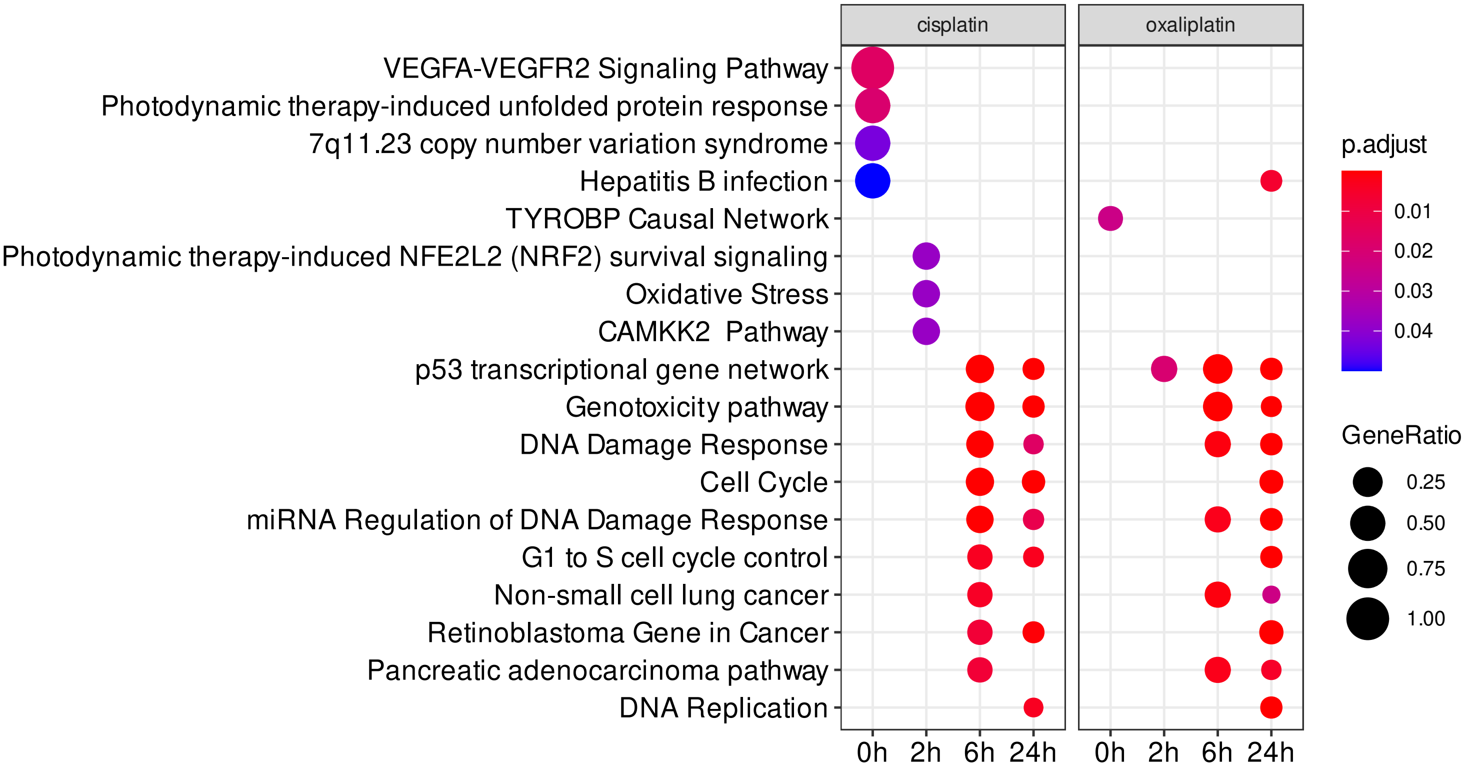

Fig. 4: Comparing functional profiles among different levels of conditions.

5.5 Visualization using ggplot2

```
library(forcats)
library(ggplot2)

ewp <- GSEA(geneList, TERM2GENE=wp[,c("wpid", "gene")], TERM2NAME=wp[,c("wpid", "name")])

ewp2 <- arrange(ewp, abs(NES)) %>%
        group_by(sign(NES)) %>%
        slice(1:5)
ego3 <- mutate(ego, richFactor = Count / as.numeric(sub("/\\d+", "", BgRatio)))

g1 <- ggplot(ego3, showCategory = 10,
  aes(richFactor, fct_reorder(Description, richFactor))) +
  geom_segment(aes(xend=0, yend = Description)) +
  geom_point(aes(color=p.adjust, size = Count)) +
  scale_color_viridis_c(guide=guide_colorbar(reverse=TRUE)) +
  scale_size_continuous(range=c(2, 10)) +
  theme_dose(12) +
  xlab("Rich Factor") +
  ylab(NULL) +
  ggtitle("Biological Processes")

g2 <- ggplot(ewp2, showCategory=10,
        aes(NES, fct_reorder(Description, NES), fill=qvalues)) +
    geom_col() +
    scale_fill_continuous(low='red', high='blue',
                        guide=guide_colorbar(reverse=TRUE)) +
    theme_dose(12) +
    xlab("Normalized Enrichment Score") +
    ylab(NULL) +
    ggtitle("WikiPathways")

cowplot::plot_grid(g1, g2, labels=c("A", "B"))
```

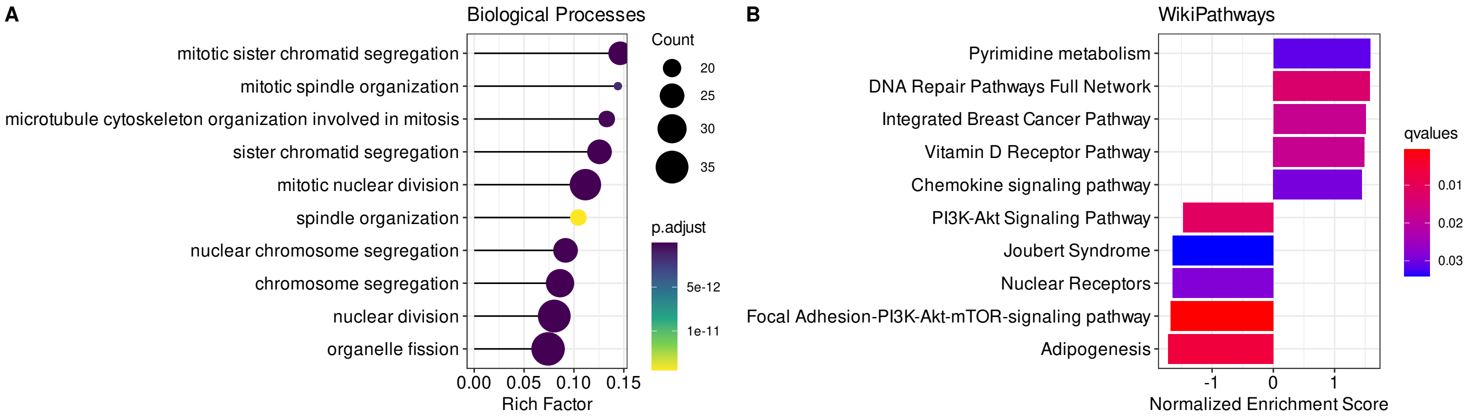

Fig. 5: Visualization enrichment results using ggplot2.

NOTE: source codes and datasets to produce this file can be obtained online[1].

6 Session information

Here is the output of `sessionInfo()` of the system on which the Supplemental file was compiled:

```
## R version 4.1.0 (2021-05-18)
## Platform: x86_64-pc-linux-gnu (64-bit)
## Running under: Arch Linux
##
## Matrix products: default
## BLAS:   /usr/lib/libblas.so.3.9.1
## LAPACK: /usr/lib/liblapack.so.3.9.1
##
## locale:
##  [1] LC_CTYPE=en_US.UTF-8       LC_NUMERIC=C
##  [3] LC_TIME=en_US.UTF-8        LC_COLLATE=en_US.UTF-8
##  [5] LC_MONETARY=en_US.UTF-8    LC_MESSAGES=en_US.UTF-8
##  [7] LC_PAPER=en_US.UTF-8       LC_NAME=C
##  [9] LC_ADDRESS=C               LC_TELEPHONE=C
## [11] LC_MEASUREMENT=en_US.UTF-8 LC_IDENTIFICATION=C
##
## attached base packages:
## [1] stats4    parallel  stats     graphics  grDevices utils     datasets
## [8] methods   base
##
## other attached packages:
##  [1] TxDb.Hsapiens.UCSC.hg19.knownGene_3.2.2
##  [2] GenomicFeatures_1.44.0
##  [3] GenomicRanges_1.44.0
##  [4] GenomeInfoDb_1.28.0
##  [5] ChIPseeker_1.28.3
##  [6] forcats_0.5.1
##  [7] ggplot2_3.3.3
##  [8] enrichplot_1.12.0
##  [9] clusterProfiler_4.0.0
```

[1] https://github.com/YuLab-SMU/supplemental-clusterProfiler-v4

```
## [10] DOSE_3.18.0
## [11] org.Hs.eg.db_3.13.0
## [12] AnnotationDbi_1.54.0
## [13] IRanges_2.26.0
## [14] S4Vectors_0.30.0
## [15] Biobase_2.52.0
## [16] BiocGenerics_0.38.0
## [17] kableExtra_1.3.4
## [18] magrittr_2.0.1
## [19] conflicted_1.0.4
## [20] rvcheck_0.1.8
## [21] wget_0.0.1
## [22] rmarkdown_2.8
##
## loaded via a namespace (and not attached):
##   [1] utf8_1.2.1                  R.utils_2.10.1
##   [3] tidyselect_1.1.1            RSQLite_2.2.7
##   [5] grid_4.1.0                  BiocParallel_1.26.0
##   [7] scatterpie_0.1.6            munsell_0.5.0
##   [9] codetools_0.2-18            withr_2.4.2
##  [11] colorspace_2.0-1            GOSemSim_2.18.0
##  [13] filelock_1.0.2              knitr_1.33
##  [15] uuid_0.1-4                  rstudioapi_0.13
##  [17] scholar_0.2.1               labeling_0.4.2
##  [19] MatrixGenerics_1.4.0        rcmdcheck_1.3.3
##  [21] GenomeInfoDbData_1.2.6      polyclip_1.10-0
##  [23] rhub_1.1.1                  bit64_4.0.5
##  [25] farver_2.1.0                rprojroot_2.0.2
##  [27] downloader_0.4              vctrs_0.3.8
##  [29] treeio_1.16.0               generics_0.1.0
##  [31] xfun_0.23                   BiocFileCache_2.0.0
##  [33] R6_2.5.0                    graphlayouts_0.7.1
##  [35] ypages_0.0.1                RJSONIO_1.3-1.4
##  [37] bitops_1.0-7                cachem_1.0.5
##  [39] fgsea_1.18.0                DelayedArray_0.18.0
##  [41] assertthat_0.2.1            BiocIO_1.2.0
##  [43] scales_1.1.1                ggraph_2.0.5
##  [45] gtable_0.3.0                processx_3.5.2
##  [47] tidygraph_1.2.0             rlang_0.4.11
##  [49] whoami_1.3.0                systemfonts_1.0.2
##  [51] splines_4.1.0               rtracklayer_1.52.0
##  [53] lazyeval_0.2.2              BiocManager_1.30.15
##  [55] yaml_2.2.1                  reshape2_1.4.4
##  [57] qvalue_2.24.0               tools_4.1.0
##  [59] bookdown_0.22               xopen_1.0.0
##  [61] gplots_3.1.1                ellipsis_0.3.2
##  [63] RColorBrewer_1.1-2          Rcpp_1.0.6
##  [65] plyr_1.8.6                  progress_1.2.2
##  [67] zlibbioc_1.38.0             purrr_0.3.4
##  [69] RCurl_1.98-1.3              ps_1.6.0
##  [71] rCharts_0.4.5               prettyunits_1.1.1
##  [73] viridis_0.6.1               cowplot_1.1.1
##  [75] SummarizedExperiment_1.22.0 ggrepel_0.9.1
##  [77] data.table_1.14.0           DO.db_2.9
##  [79] whisker_0.4                 ggnewscale_0.4.5
##  [81] R.cache_0.15.0              matrixStats_0.58.0
##  [83] hms_1.1.0                   patchwork_1.1.1
##  [85] evaluate_0.14               XML_3.99-0.6
##  [87] gridExtra_2.3               ggupset_0.3.0
##  [89] compiler_4.1.0              biomaRt_2.48.0
##  [91] tibble_3.1.2                KernSmooth_2.23-20
##  [93] crayon_1.4.1                shadowtext_0.0.8
##  [95] R.oo_1.24.0                 htmltools_0.5.1.1
##  [97] tidyr_1.1.3                 aplot_0.0.6
##  [99] DBI_1.1.1                   tweenr_1.0.2
## [101] dbplyr_2.1.1                MASS_7.3-54
## [103] rappdirs_0.3.3              boot_1.3-28
## [105] dlstats_0.1.4               Matrix_1.3-3
## [107] badger_0.1.0                cli_2.5.0
## [109] R.methodsS3_1.8.1           igraph_1.2.6
## [111] pkgconfig_2.0.3             GenomicAlignments_1.28.0
## [113] xml2_1.3.2                  ggtree_3.0.0
## [115] svglite_2.0.0               webshot_0.5.2
## [117] XVector_0.32.0              rematch_1.0.1
## [119] rvest_1.0.0                 stringr_1.4.0
## [121] callr_3.7.0                 digest_0.6.27
## [123] Biostrings_2.60.0           fastmatch_1.1-0
## [125] tidytree_0.3.4              restfulr_0.0.13
## [127] curl_4.3.1                  gtools_3.8.2
## [129] Rsamtools_2.8.0             rjson_0.2.20
## [131] lifecycle_1.0.0             nlme_3.1-152
## [133] jsonlite_1.7.2              desc_1.3.0
## [135] viridisLite_0.4.0           fansi_0.4.2
## [137] pillar_1.6.1                lattice_0.20-44
## [139] plotrix_3.8-1               KEGGREST_1.32.0
## [141] fastmap_1.1.0               httr_1.4.2
## [143] pkgbuild_1.2.0              GO.db_3.13.0
## [145] parsedate_1.2.1             glue_1.4.2
## [147] png_0.1-7                   bit_4.0.4
## [149] ggforce_0.3.3               stringi_1.6.2
## [151] blob_1.2.1                  caTools_1.18.2
## [153] memoise_2.0.0               dplyr_1.0.6
## [155] ape_5.5
```