

Pathway Analysis

NGS workshop

Date: 17th May 2023...3rd hour

In this script, we will be using the “clusterProfiler” package to perform over-representation analysis and GSEA (demonstration) in R

References:

<http://bioconductor.org/packages/release/bioc/html/clusterProfiler.html>

<https://yulab-smu.github.io/clusterProfiler-book/index.html>

[https://www.cell.com/the-innovation/pdf/S2666-6758\(21\)00066-7.pdf](https://www.cell.com/the-innovation/pdf/S2666-6758(21)00066-7.pdf)

Required libraries

```
library(clusterProfiler)
library(dplyr)
library(ggplot2)
library(org.Hs.eg.db)
```

Read the table with the significant results created earlier

```
DEG05_symbol <- read.csv("counts/DEGs_5uMaza_treatment_significant.csv")
```

Pathway Analysis

GO Ontologies (ont)

Describe the roles of genes and gene products. The GO terms are organized into three ontologies.

1. Biological process (BP) : refers to the biological role involving the gene or gene product, such as translation, transcription, signal transduction, and cell cycle
2. Molecular function (MF): characterizes the biochemical activity of the gene product, such activities as ligand, transporter,
3. Cellular component (CC): indicates the location in the cell of the gene product

```
geneset <- as.character(DEG05_symbol$entrez)
#GO over-representation analysis
ora_go <- enrichGO(gene      = geneset,
                   universe   = NULL,
                   # all available genes in db
```

```

OrgDb          = org.Hs.eg.db,      # Hs: homo sapiens
ont            = "BP",              # One of MF, BP, CC
pAdjustMethod  = "BH",              # Benjamini & Hochberg (FDR)
pvalueCutoff   = 0.01,              # adjusted p-value cutoff
qvalueCutoff   = 0.05,
readable       = TRUE)

```

The enriched pathways are saved in the result of the ora_go object

[View the results using the head command](#)

```
head(ora_go@result)
```

```

      ID                                     Description GeneRatio
GO:0002181 GO:0002181      cytoplasmic translation      19/156
GO:0045071 GO:0045071 negative regulation of viral genome replication 12/156
GO:0019079 GO:0019079      viral genome replication      14/156
GO:0045069 GO:0045069      regulation of viral genome replication 12/156
GO:0048525 GO:0048525      negative regulation of viral process 12/156
GO:0051607 GO:0051607      defense response to virus      17/156
      BgRatio      pvalue      p.adjust      qvalue
GO:0002181 161/18903 7.248108e-17 1.625751e-13 1.570932e-13
GO:0045071  57/18903 3.340344e-14 3.746196e-11 3.619878e-11
GO:0019079 130/18903 3.614476e-12 2.702423e-09 2.611300e-09
GO:0045069  87/18903 6.777014e-12 3.800211e-09 3.672072e-09
GO:0048525  91/18903 1.171919e-11 5.257227e-09 5.079959e-09
GO:0051607 304/18903 6.020178e-10 2.028598e-07 1.960196e-07
                                     geneID
GO:0002181 RPL31/RPL3/RPL22/RPS6/RPS24/RPS2/RPL13A/RPL32/RPS3A/RPL7/RPS3/RPL29/RPS14/RPSA/RPS7/RPL4/RPS23/RPL12/RPS4X
GO:0045071 OAS1/OAS3/OAS2/IFIH1/BST2/RSAD2/IFITM3/MX1/RESF1/IFIT1/IFITM1/ISG15
GO:0019079 NUCKS1/OAS1/OAS3/OAS2/IFIH1/BST2/RSAD2/IFITM3/EEF1A1/MX1/RESF1/IFIT1/IFITM1/ISG15
GO:0045069 OAS1/OAS3/OAS2/IFIH1/BST2/RSAD2/IFITM3/MX1/RESF1/IFIT1/IFITM1/ISG15
GO:0048525 OAS1/OAS3/OAS2/IFIH1/BST2/RSAD2/IFITM3/MX1/RESF1/IFIT1/IFITM1/ISG15
GO:0051607 OAS1/NOP53/OAS3/OAS2/IFIH1/IFI6/BST2/RSAD2/PARP9/IFITM3/MX1/STAT2/DDX60L/MX2/IFIT1/IFITM1/ISG15
      Count
GO:0002181 19
GO:0045071 12
GO:0019079 14
GO:0045069 12
GO:0048525 12
GO:0051607 17

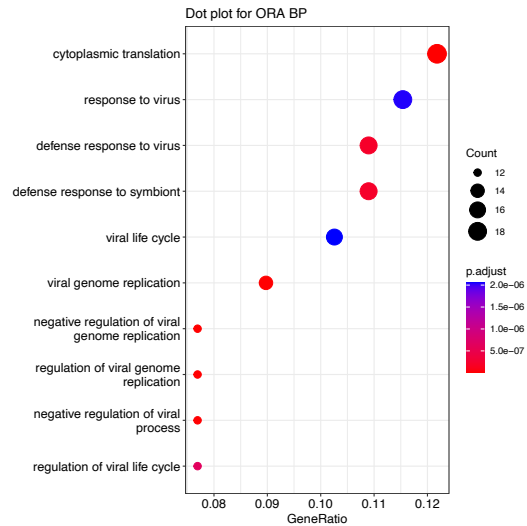
```

In the results clusterProfiler provides geneRatio (ratio of input genes that are annotated in a term) and BgRatio (ratio of all genes that are annotated in this term), pvalue
p.adjust qvalue

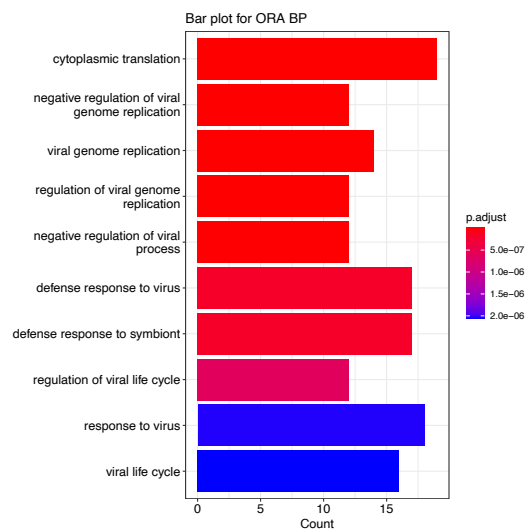
Visualize enriched GO terms

There are different ways to visualize the enriched pathway such as dot plot, bar plot, and concept network plot as follow:

```
dotplot(ora_go, showCategory = 10)+ ggtitle("Dot plot for ORA BP")
```



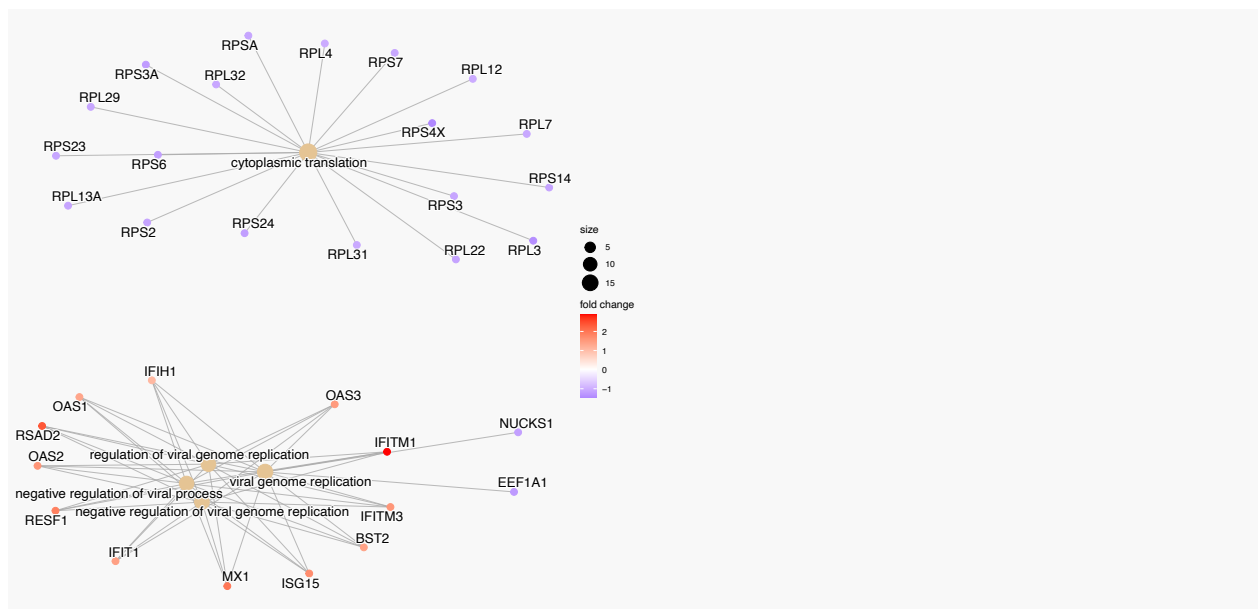
```
barplot(ora_go, showCategory = 10)+ ggtitle("Bar plot for ORA BP")
```



To display the relation of specific genes with the GO terms use cnetplot (concept network).

Use the value of the Log2 fold change to color the gene in the network with the log 2 fold change value

```
logFC_de <- DEG05_symbol$log2FoldChange
names(logFC_de) <- DEG05_symbol$entrez
cnetplot(ora_go, foldChange = logFC_de, circular = FALSE)
```



Gene set enrichment analysis (GSEA)

Demonstration only

In this exercise, we will explore the enrichment of BP Gene Ontology terms using GSEA

The dataset that we used in this workshop is reduced in size. GSEA was not significant. We are showing the steps to run GSEA in R using the data generated from the whole dataset.

If you would like to run this on your data generated in the course, please use DEGs_5uMaza_treatment_All.csv file and use the `pvalueCutoff = 1` for the code `GSEA.r`

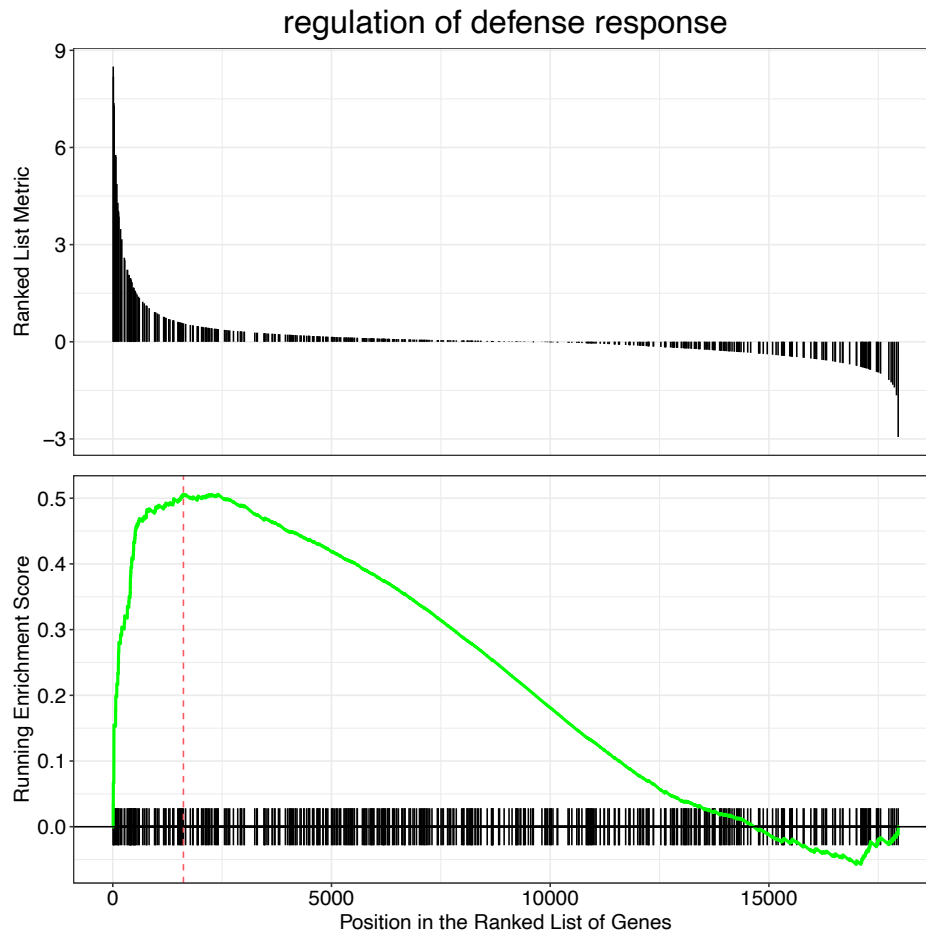
```
All_genes <- read.table("~/results/WholeDataSet_2Rep/DEGs_5uMaza_treatment_All.csv", sep=";", header=T)
logFC_all <- All_genes$log2FoldChange
names(logFC_all) <- All_genes$entrez
logFC_all_sorted <- sort(logFC_all, decreasing = TRUE)
```

results might vary depending on the laptop you are using.
`set.seed(123456)`

```
gsea_go <- gseGO(geneList      = logFC_all_sorted,
                 OrgDb         = org.Hs.eg.db,
                 ont            = "BP",
                 nPerm          = 1000,
                 minGSSize      = 100, #minimum gene set size (#genes in set)
                 maxGSSize      = 500,
                 pvalueCutoff    = 0.05,
                 verbose         = FALSE)
```

We can use the function `gseaplot()` to visualize the results of the GSEA.

```
# Plot the GSEA results for the 4th enriched pathway  
gseaplot(gsea_go, geneSetID = 1, title = gsea_go$Description[4])
```



```
sessionInfo()
```

```
## R version 4.2.1 (2022-06-23)  
Platform: x86_64-pc-linux-gnu (64-bit)  
Running under: Ubuntu 20.04.6 LTS
```

```
Matrix products: default
```

```
BLAS: /mnt/service/software/packages/r/R-4.2.1/lib/R/lib/libRblas.so
```

```
LAPACK: /mnt/service/software/packages/r/R-4.2.1/lib/R/lib/libRlapack.so
```

```
locale:
```

```

[1] LC_CTYPE=en_GB.UTF-8      LC_NUMERIC=C               LC_TIME=en_GB.UTF-
8
[4] LC_COLLATE=en_GB.UTF-8    LC_MONETARY=en_GB.UTF-8   LC_MESSAGES=en_GB.UTF-8
[7] LC_PAPER=en_GB.UTF-8     LC_NAME=C                 LC_ADDRESS=C
[10] LC_TELEPHONE=C           LC_MEASUREMENT=en_GB.UTF-8 LC_IDENTIFICATION=C

```

attached base packages:

```

[1] stats4      stats      graphics  grDevices  utils      datasets  methods    base

```

other attached packages:

```

[1] clusterProfiler_4.6.2      EnhancedVolcano_1.16.0    ggrepel_0.9.3
[4] org.Hs.eg.db_3.16.0        AnnotationDbi_1.60.2      RColorBrewer_1.1-3
[7] DESeq2_1.38.0              SummarizedExperiment_1.26.1 Biobase_2.58.0
[10] MatrixGenerics_1.8.1      matrixStats_0.62.0        GenomicRanges_1.48.0
[13] GenomeInfoDb_1.34.9        IRanges_2.32.0            S4Vectors_0.36.2
[16] BiocGenerics_0.44.0        ggplot2_3.4.2             dplyr_1.1.1

```

loaded via a namespace (and not attached):

```

[1] ggnewscale_0.4.8          fgsea_1.24.0              colorspace_2.1-0
[4] ggtree_3.6.2              gson_0.1.0                qvalue_2.30.0
[7] XVector_0.38.0            aplot_0.1.10             farver_2.1.1
[10] graphlayouts_0.8.4        bit64_4.0.5              fansi_1.0.4
[13] scatterpie_0.1.8          codetools_0.2-19         splines_4.2.1
[16] cachem_1.0.7              GOSeqSim_2.24.0          geneplotter_1.76.0
[19] polyclip_1.10-4           jsonlite_1.8.4           annotate_1.76.0
[22] GO.db_3.16.0              png_0.1-8                ggforce_0.4.1
[25] compiler_4.2.1            httr_1.4.5               Matrix_1.5-4
[28] fastmap_1.1.1             lazyeval_0.2.2           cli_3.6.1
[31] tweenr_2.0.2              tools_4.2.1              igraph_1.4.2
[34] gtable_0.3.3              glue_1.6.2               GenomeInfoDbData_1.2.9
[37] reshape2_1.4.4            fastmatch_1.1-3          Rcpp_1.0.10
[40] enrichplot_1.18.4         vctrs_0.6.1              Biostrings_2.66.0
[43] ape_5.7-1                 nlme_3.1-162             ggraph_2.1.0
[46] stringr_1.5.0             lifecycle_1.0.3          XML_3.99-0.11
[49] DOSE_3.24.2               zlibbioc_1.44.0          MASS_7.3-60
[52] scales_1.2.1              tidygraph_1.2.3          parallel_4.2.1
[55] memoise_2.0.1             gridExtra_2.3            downloader_0.4
[58] ggfun_0.0.9               HDO.db_0.99.1            yulab.utils_0.0.6
[61] stringi_1.7.12            RSQLite_2.3.1            genefilter_1.80.0
[64] tidytree_0.4.2            BiocParallel_1.32.6      rlang_1.1.0
[67] pkgconfig_2.0.3           bitops_1.0-7             lattice_0.21-8
[70] purrr_1.0.1              treeio_1.22.0            patchwork_1.1.2
[73] labeling_0.4.2            cowplot_1.1.1            shadowtext_0.1.2
[76] bit_4.0.5                 tidyselect_1.2.0         plyr_1.8.8
[79] magrittr_2.0.3            R6_2.5.1                 generics_0.1.3
[82] DelayedArray_0.22.0       DBI_1.1.3                pillar_1.9.0
[85] withr_2.5.0               survival_3.5-5           KEGGREST_1.38.0
[88] RCurl_1.98-1.12          tibble_3.2.1             crayon_1.5.2
[91] utf8_1.2.3                viridis_0.6.2            locfit_1.5-9.6
[94] grid_4.2.1                data.table_1.14.8        blob_1.2.4
[97] digest_0.6.31            xtable_1.8-4             tidyr_1.3.0
[100] gridGraphics_0.5-1        munsell_0.5.0            viridisLite_0.4.1
[103] ggplotify_0.1.0

```