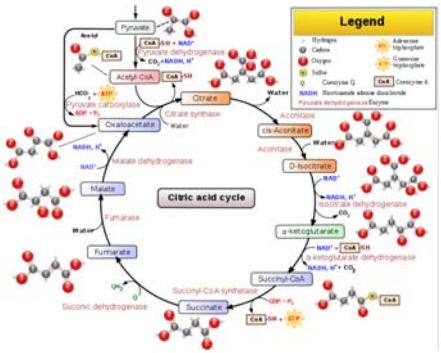
What is a gene set?



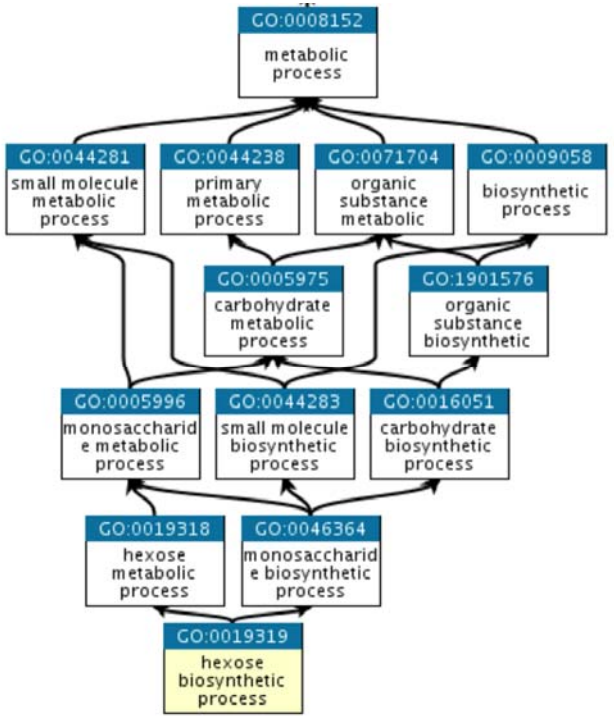
https://en.wikipedia.org/wiki/Citric_acid_cycle

- Genes working together in a pathway (e.g. energy release through Krebs cycle)
- Genes located in the same compartment in a cell (e.g. all proteins located in the cell nucleus)
- Proteins that are all regulated by a same transcription factor
- Custom gene list that comes from a publication and that are down-regulated in a mutant
- List of genes associated with a disease
- … etc!
- Several gene sets are grouped into Knowledge bases

Gene ontology

- http://geneontology.org/

- collaborative effort to address the need for consistent descriptions of gene products across databases
- GO Consortium: develop a comprehensive, computational model of biological systems, ranging from the molecular to the organism level, across the multiplicity of species in the tree of life
- GO terms = GO categorizations
- GO term: each with a name (DNA repair) and a unique accession number (GO:0005125)

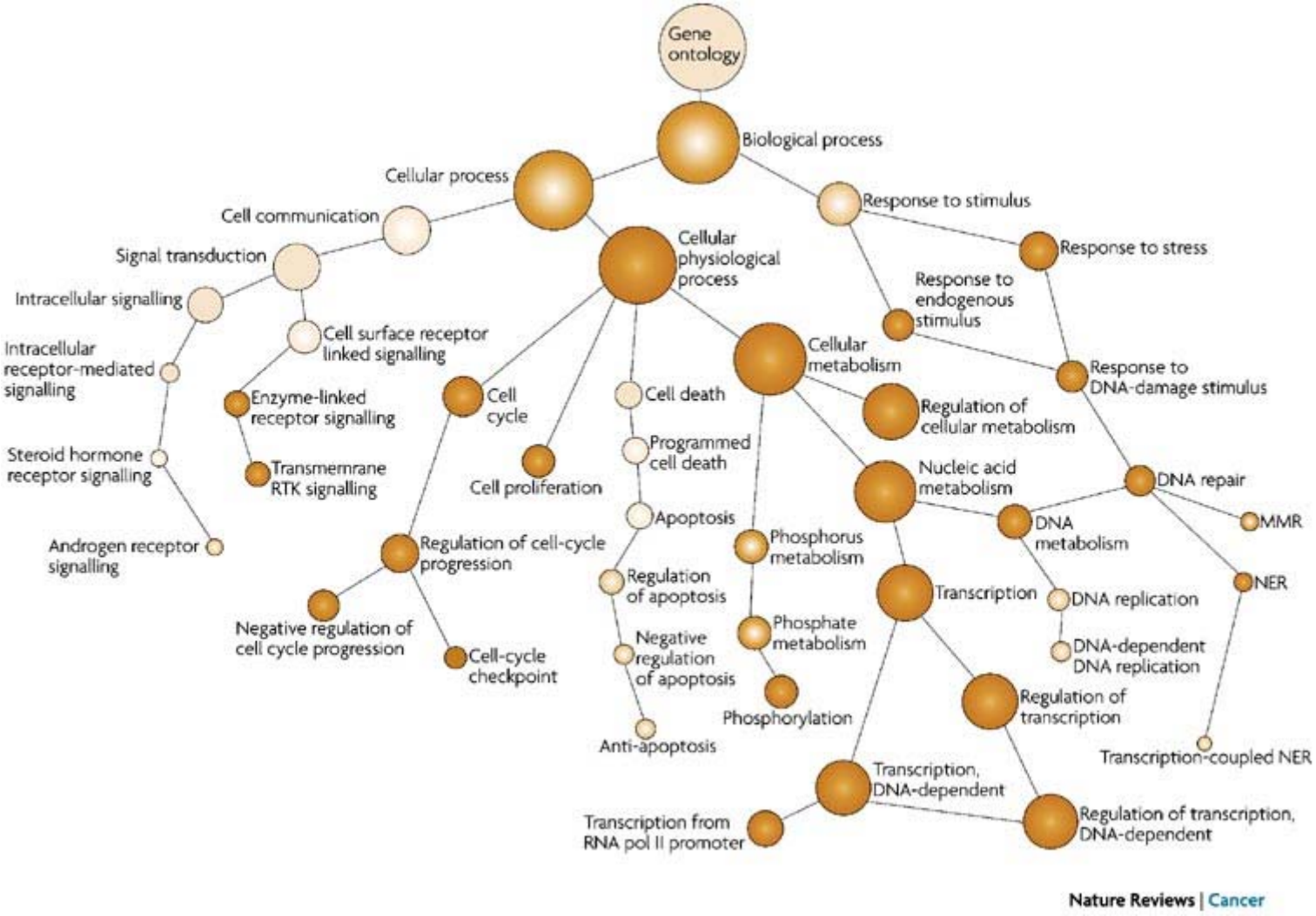


Gene ontology

GO ontologies: GO terms organized in 3 independent controlled vocabularies

- **Molecular function**: represents the biochemical activity of the gene product, such activities could include "ligand", "GTPase", and "transporter".
- **Cellular component**: refers to the location in the cell of the gene product. Cellular components could include "nucleus", "lysosome", and "plasma membrane".
- **Biological process**: refers to the biological role involving the gene or gene product, and could include "transcription", "signal transduction", and "apoptosis". A biological process generally involves a chemical or physical change of the starting material or input.

Gene ontology



KEGG

https://www.genome.jp/kegg/

KEGG PATHWAY Database

Wiring diagrams of molecular interactions, reactions and relations

KEGG2 PATHWAY BRITE MODULE KO GENES DISEASE DRUG COMPOUND

Select prefix: map | Organism | Enter keywords | Go | Help

[New pathway maps | Update history]

Pathway Maps

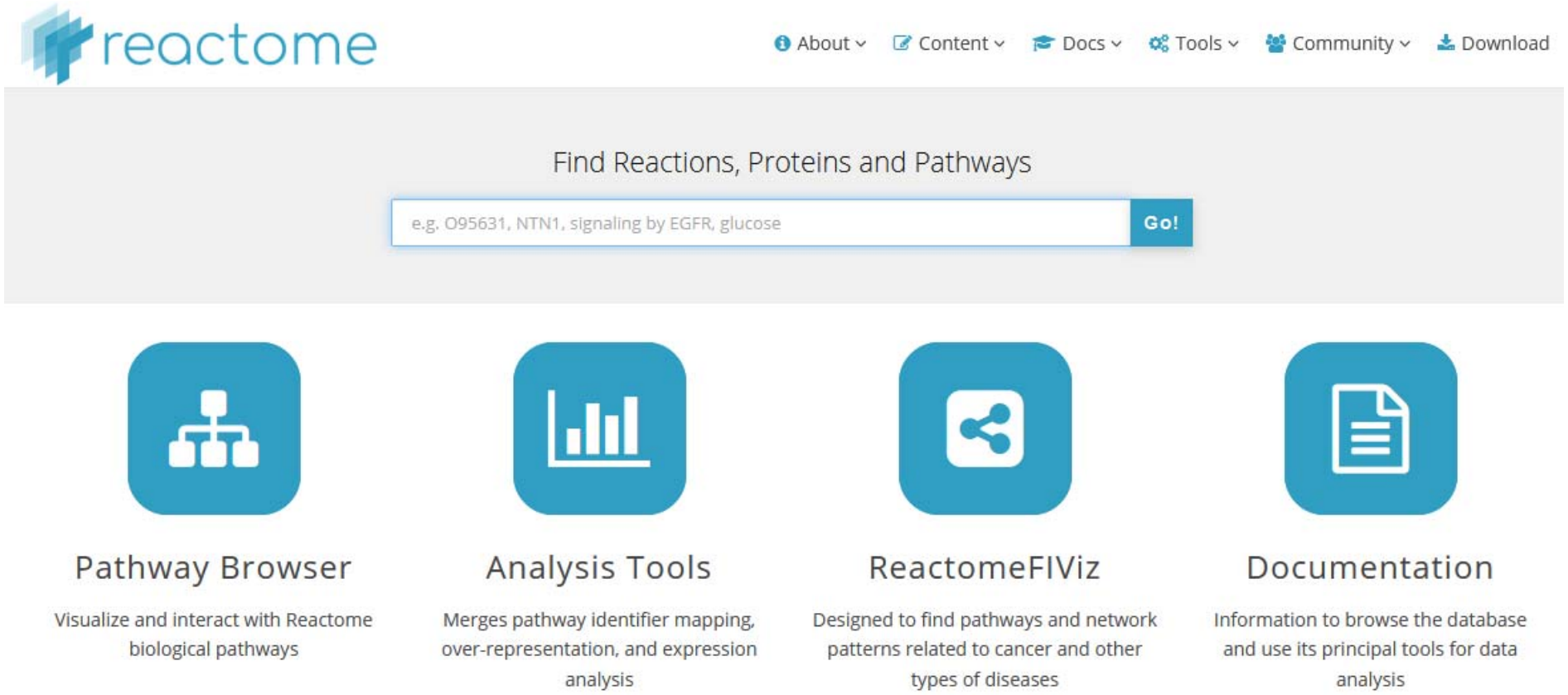
KEGG PATHWAY is a collection of manually drawn pathway maps representing our knowledge of the molecular interaction, reaction and relation networks for:

1. **Metabolism**
   Global/overview Carbohydrate Energy Lipid Nucleotide Amino acid Other amino Glycan
   Cofactor/vitamin Terpenoid/PK Other secondary metabolite Xenobiotics Chemical structure
2. **Genetic Information Processing**
3. **Environmental Information Processing**
4. **Cellular Processes**
5. **Organismal Systems**
6. **Human Diseases**
7. **Drug Development**

KEGG PATHWAY is the reference database for pathway mapping in **KEGG Mapper**.

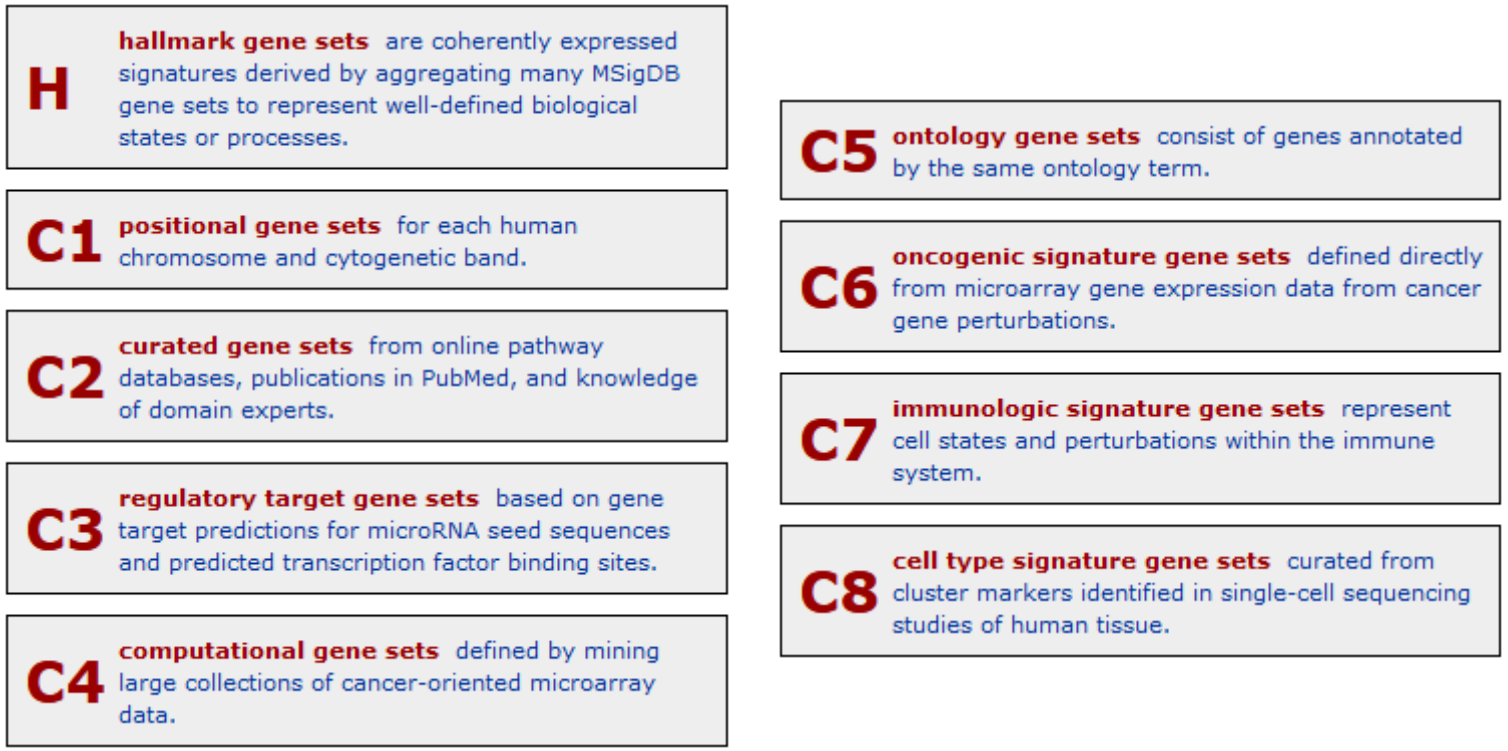
Reactome

https://reactome.org/



MSigDB

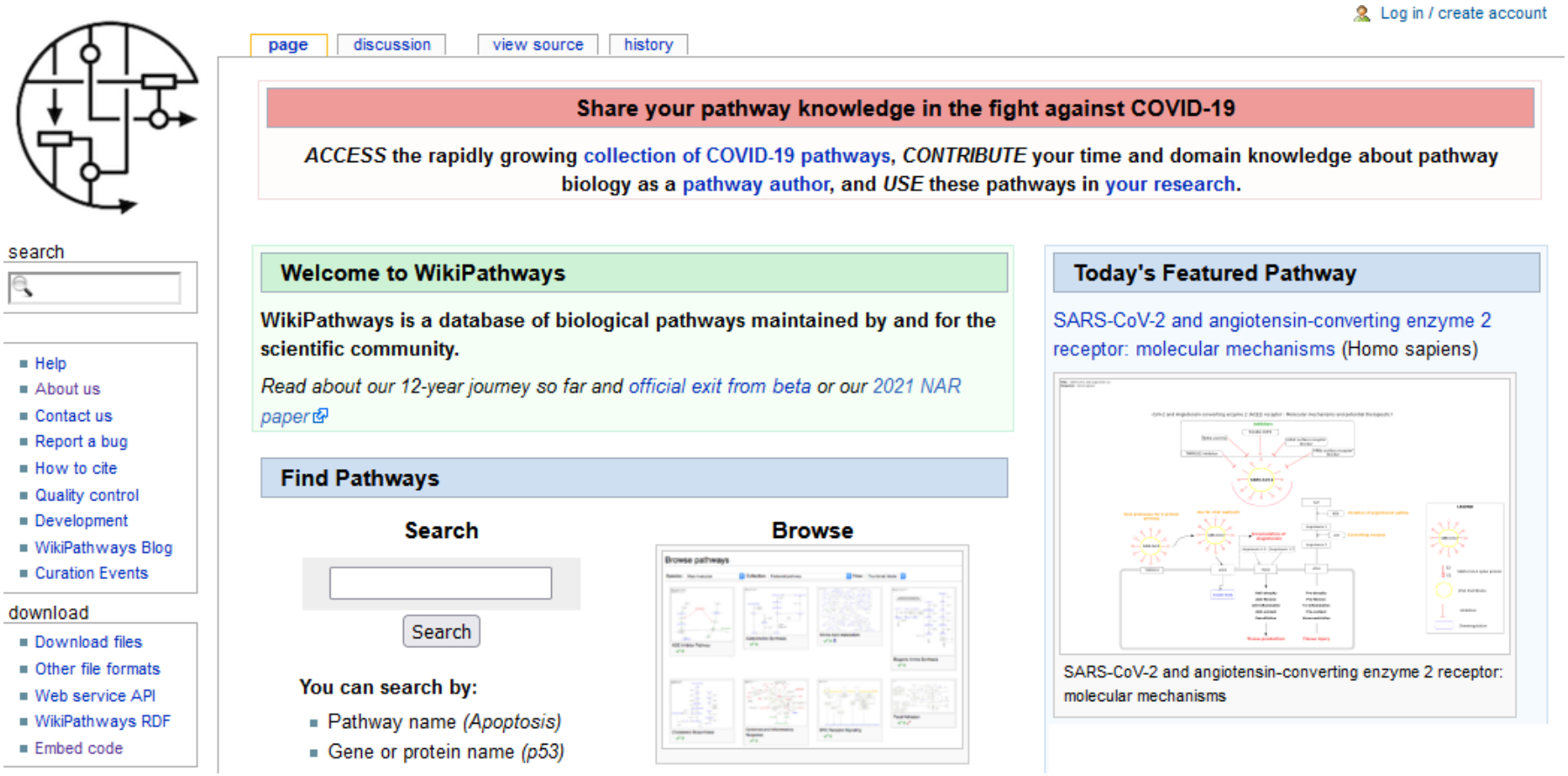
https://www.gsea-msigdb.org/gsea/msigdb/index.jsp



https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4707969/

WikiPathways

https://www.wikipathways.org/index.php/WikiPathways



GSEA of other gene sets in R

ClusterProfiler: GSEA for KEGG pathways

```
gseKEGG(geneList, organism = "hsa", keyType = "kegg", exponent = 1,
  nPerm = 1000, minGSSize = 10, maxGSSize = 500,
  pvalueCutoff = 0.05, pAdjustMethod = "BH", verbose = TRUE,
  use_internal_data = FALSE, seed = FALSE, by = "fgsea")
```

Import a .gmt file of gene sets and convert to format needed for clusterProfiler

```
read.gmt(gmtfile)
```

```
> head(term2gene_h)
                                ont    gene
1 HALLMARK_TNFA_SIGNALING_VIA_NFKB    JUNB
2 HALLMARK_TNFA_SIGNALING_VIA_NFKB   CXCL2
3 HALLMARK_TNFA_SIGNALING_VIA_NFKB    ATF3
4 HALLMARK_TNFA_SIGNALING_VIA_NFKB  NFKBIA
5 HALLMARK_TNFA_SIGNALING_VIA_NFKB TNFAIP3
6 HALLMARK_TNFA_SIGNALING_VIA_NFKB   PTGS2
```

conversion of gene ID types with clusterProfiler

```
bitr(geneID, fromType, toType, OrgDb, drop = TRUE)
```

Recap and exercise 3

- We have seen how to perform GSEA using the built-in GO gene sets. Please perform GSEA with the built-in KEGG pathways, as well as with the hallmark gene sets obtained from MSigDB.

- Exercise 3: use functions of clusterProfiler and data provided in Ex. 1, and hallmark gene sets downloaded from MSigDB

  - First convert the gene symbols to EntrezID to perform a GSEA of KEGG pathways (with argument minGSSize=30).

  - Are the majority of gene sets rather up-regulated or down-regulated?

  - Is there a KEGG immune-related gene set coming up? Is there a KEGG Natural killer gene set coming up?

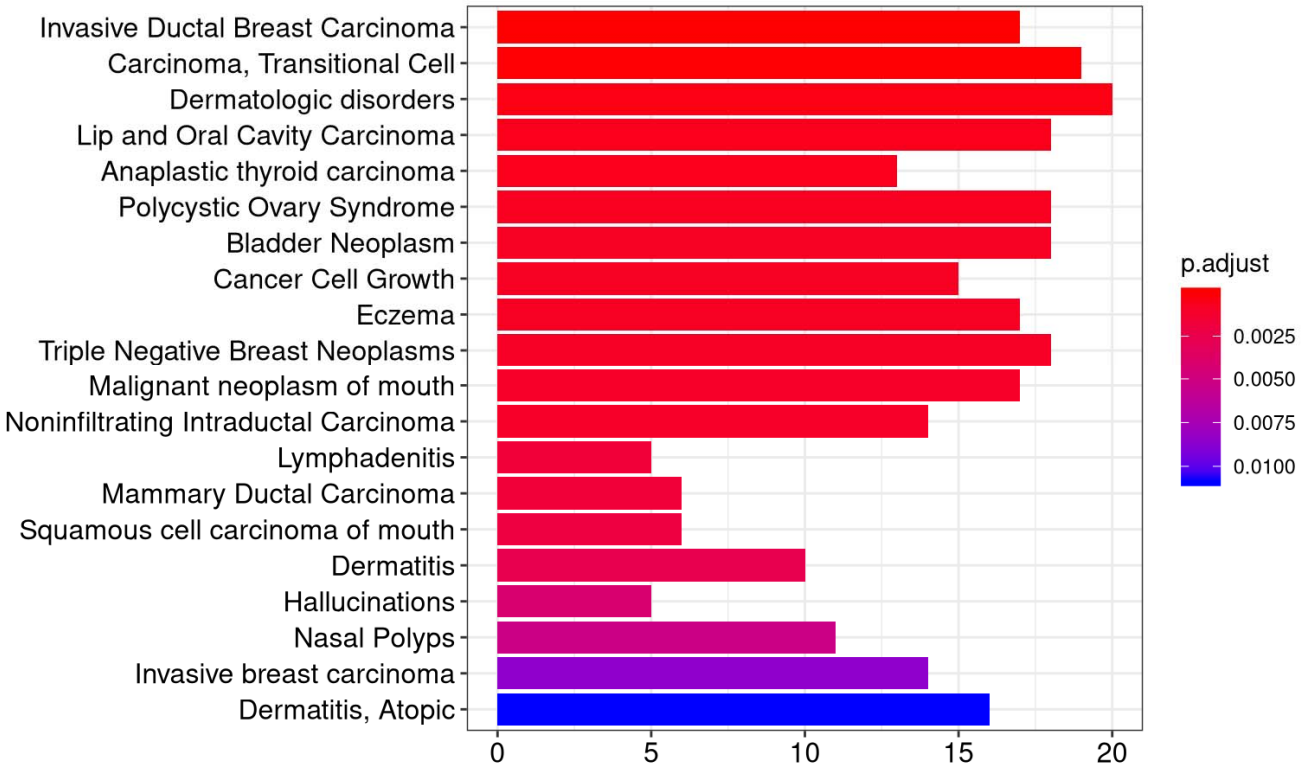
  - If you want to see which genes are included in one of the built-in KEGG pathways, where could you find this information?

  - Import the hallmark gene sets and run a GSEA. How many significant gene sets are there?

Visualization of Functional Enrichment Results

- barplot

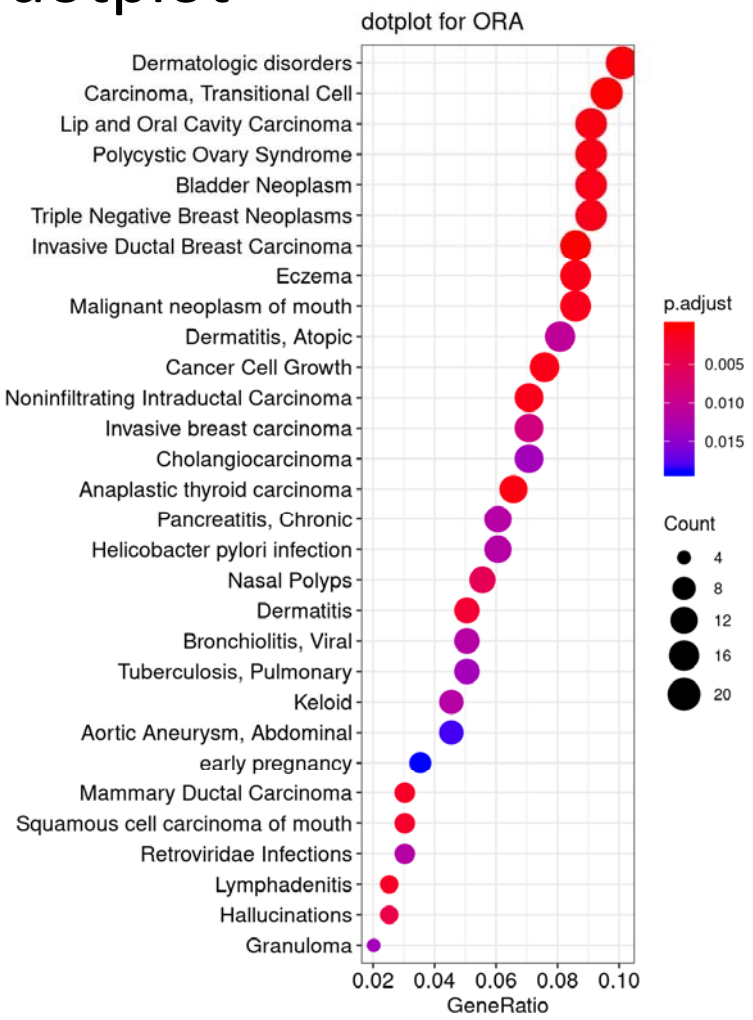
```
ego <- enrichGO(de, OrgDb='org.Hs.eg.db', ont="BP", keyType = "SYMBOL")
barplot(ego, showCategory=20)
```

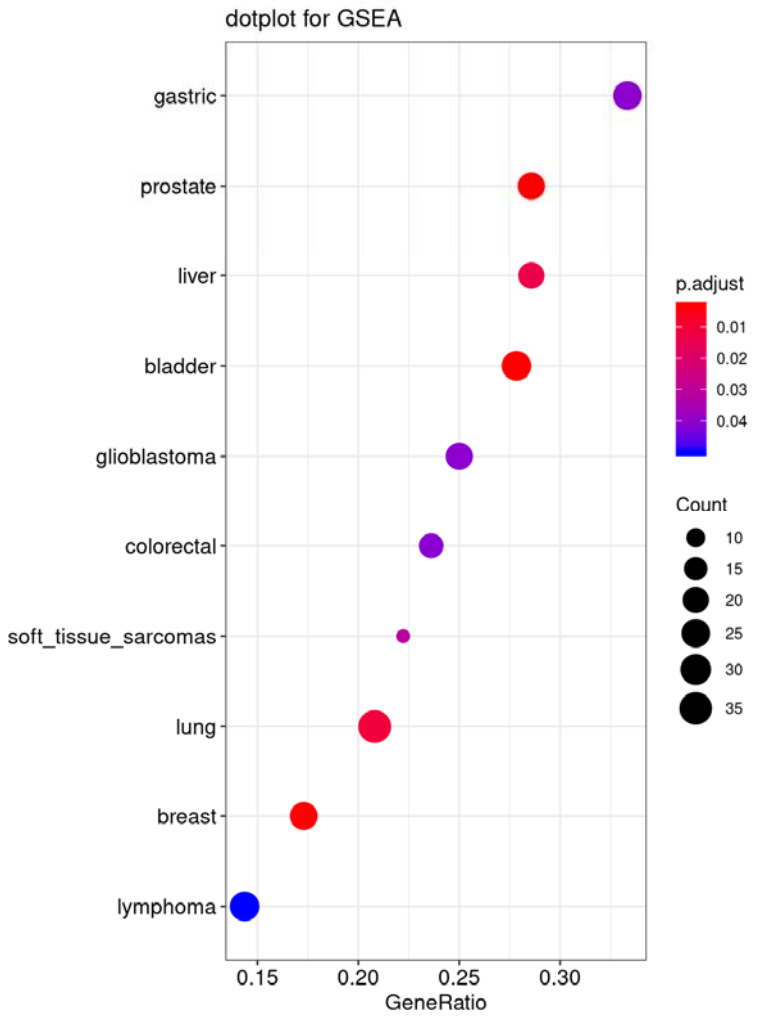


Visualization of Functional Enrichment Results

- dotplot

dotplot(ego, showCategory=20)

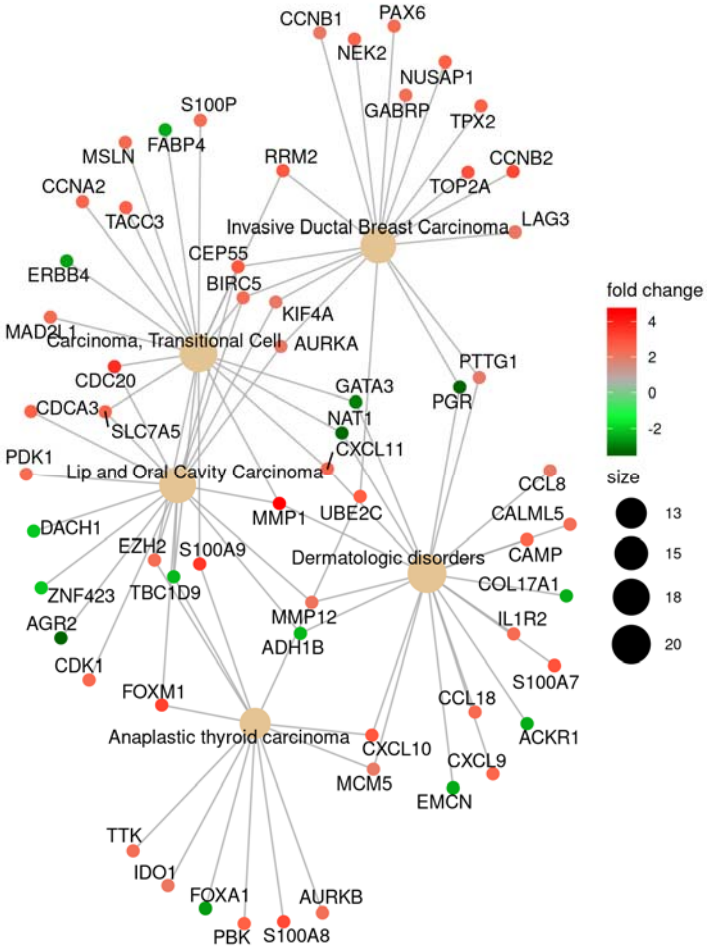




Visualization of Functional Enrichment Results

- cnetplot

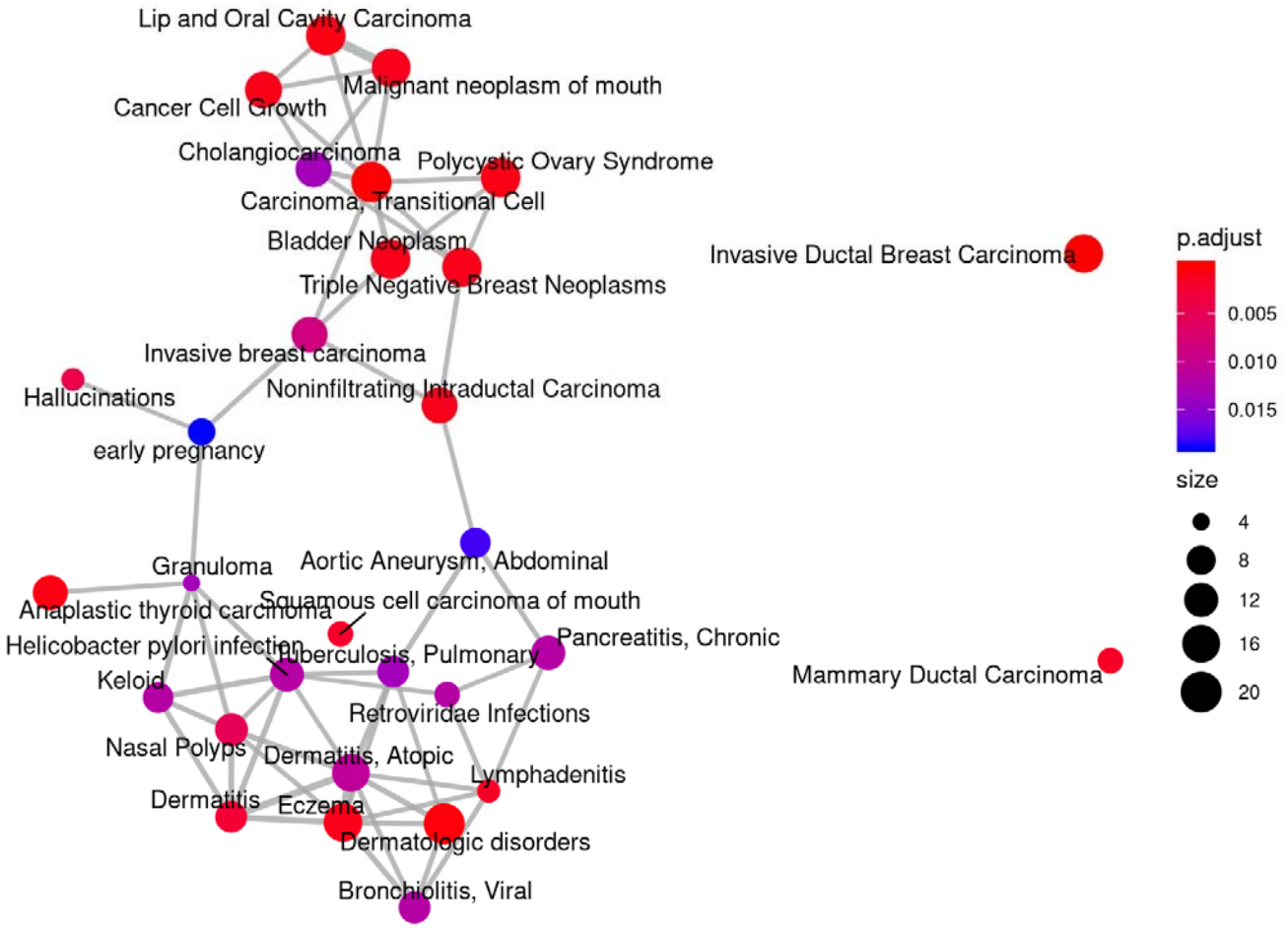
cnetplot(ego, categorySize="pvalue", foldChange=geneList)



Visualization of Functional Enrichment Results

- Enrichment map

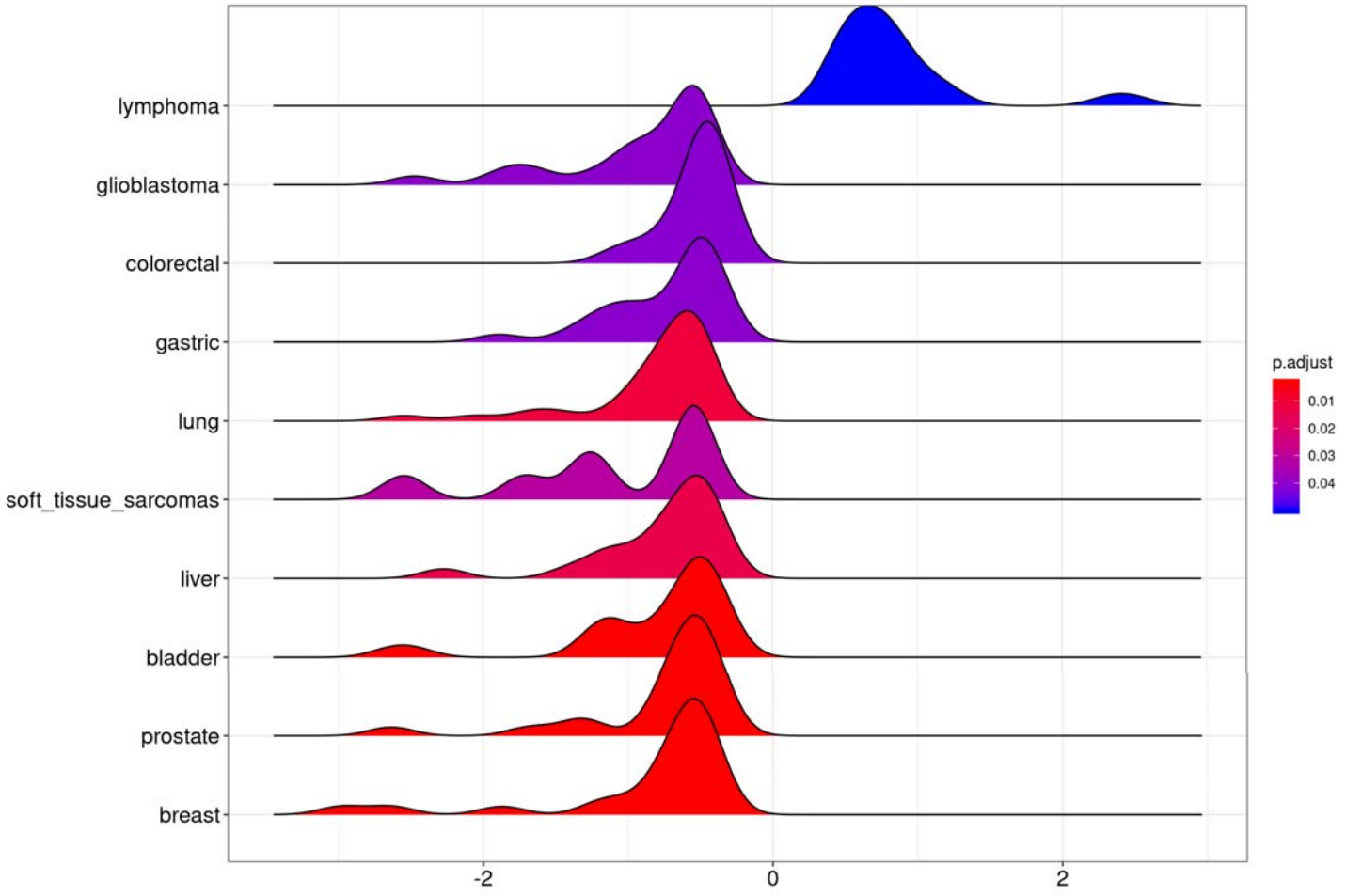
emapplot(ego)



Visualization of Functional Enrichment Results

```
ggo <- gseGO(gl, ont="BP")
ridgeplot(ggo)
```

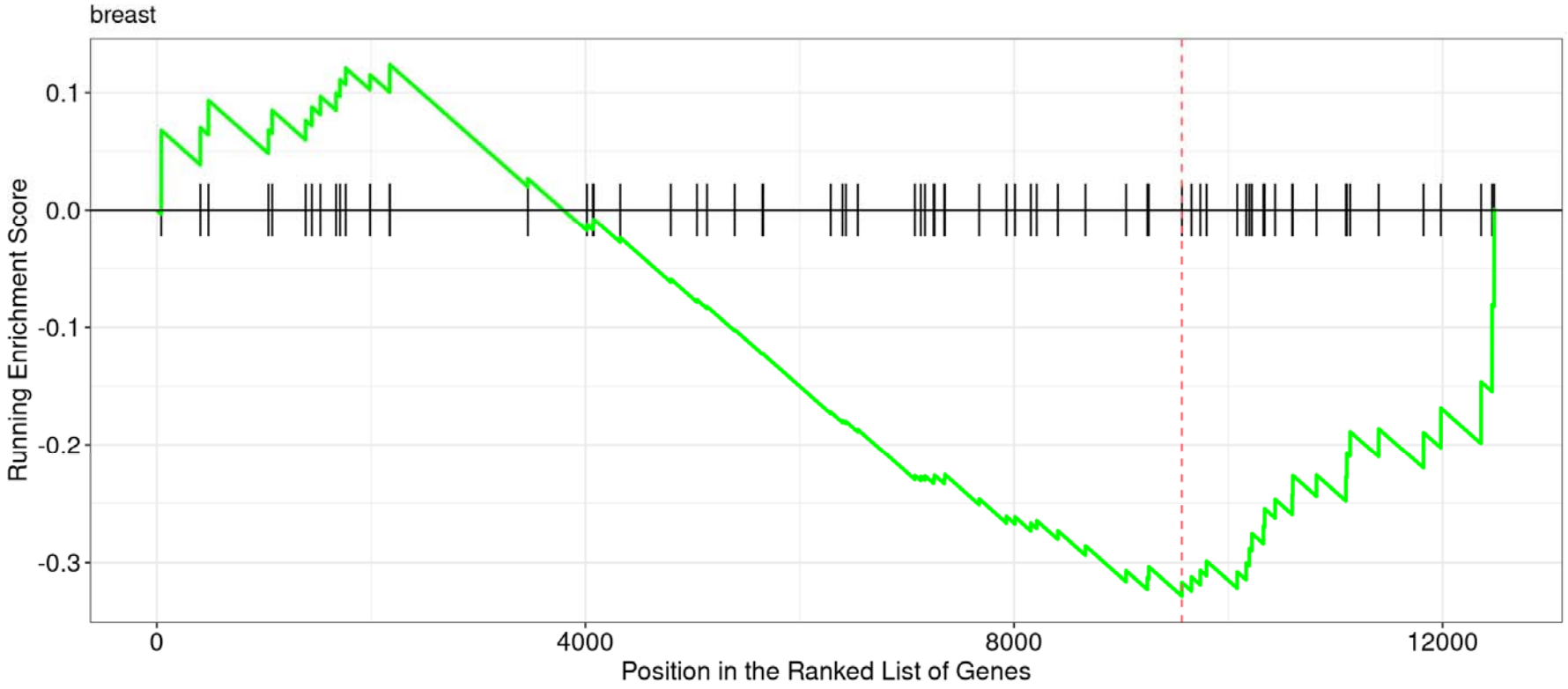
- Ridgeplot



Visualization of Functional Enrichment Results

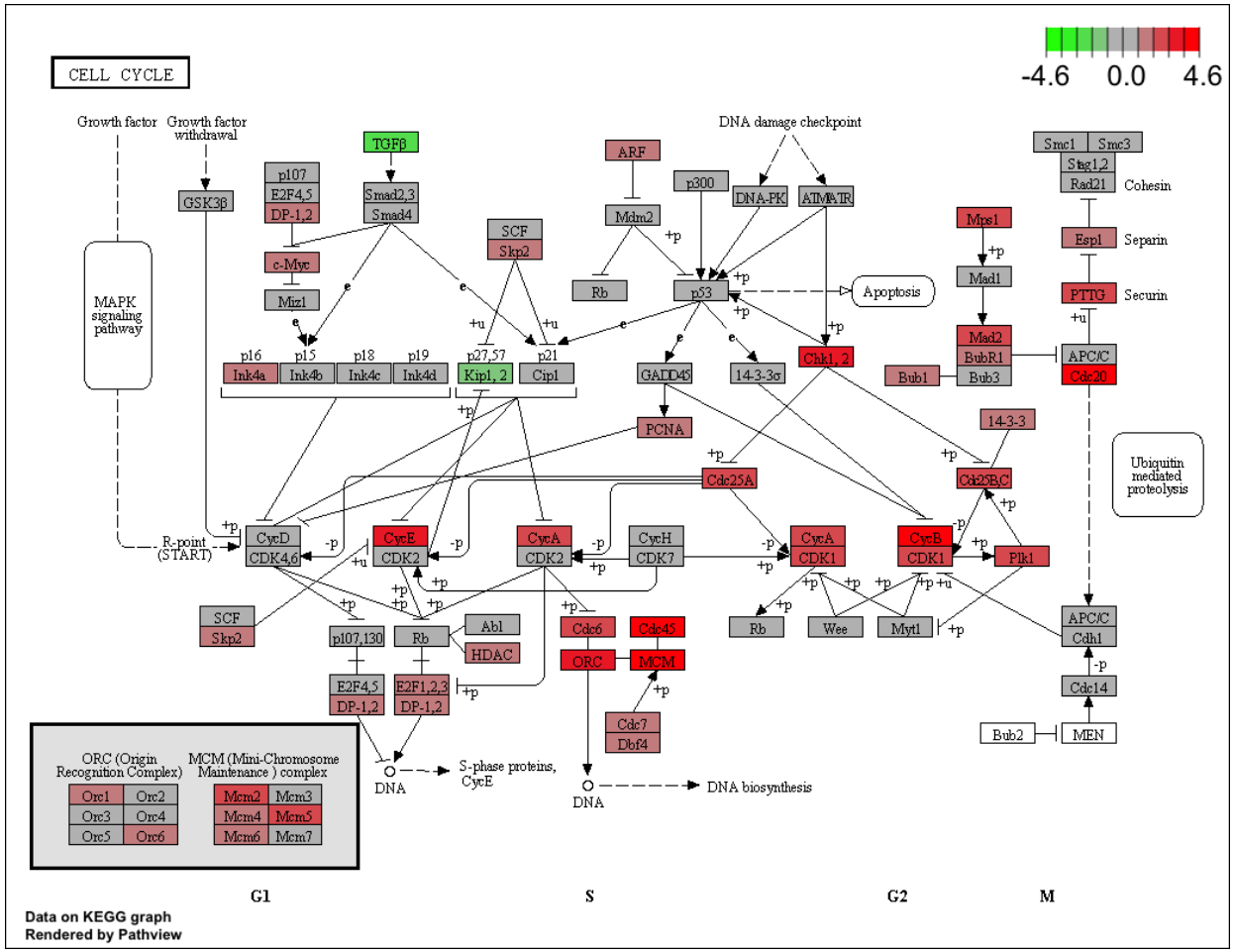
- visualizing GSEA result

gseaplot(h_NK_vs_Th, geneSetID = “BREAST", title=" BREAST")

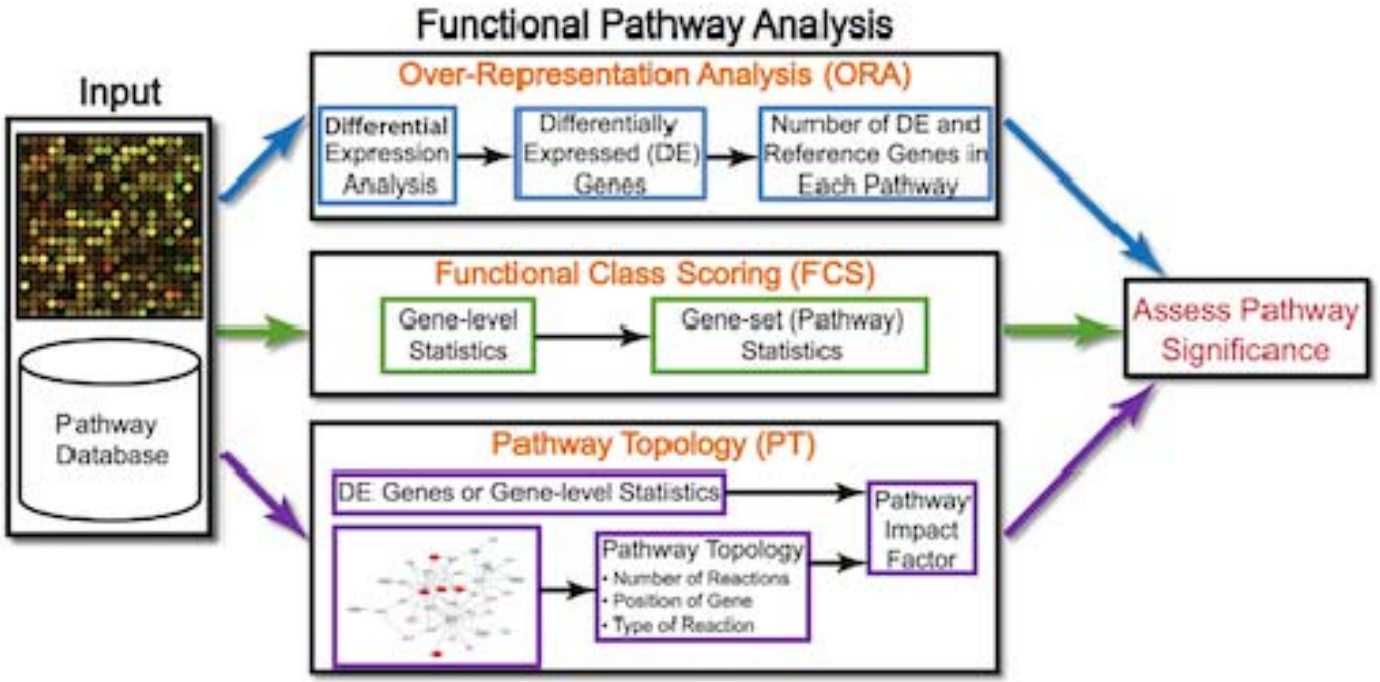


Visualization of Functional Enrichment Results

- pathview



Functional analysis



Functional analysis: **Pathway topology tools**

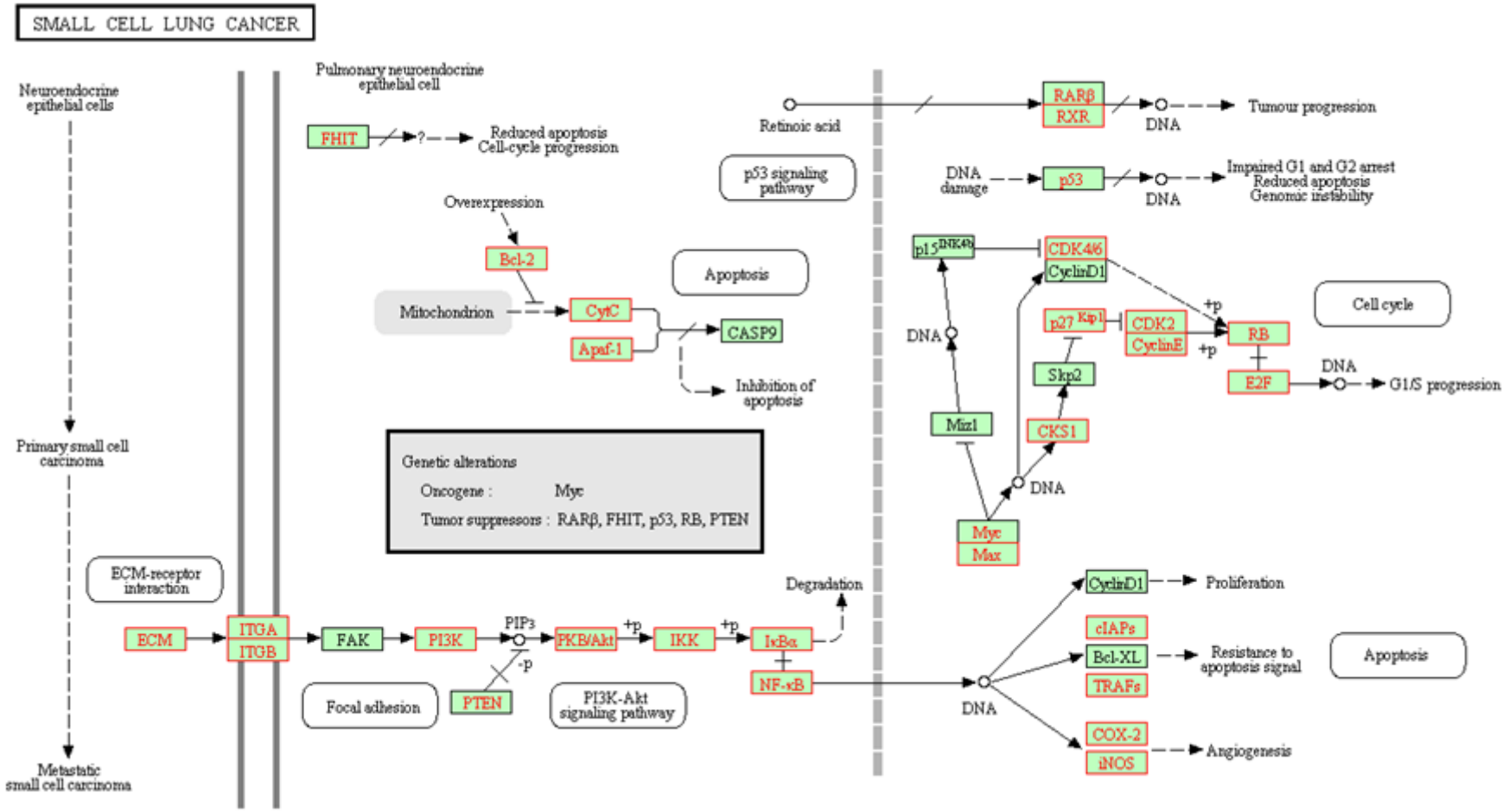
Signaling pathway impact analysis (SPIA)
Identification of dys-regulated pathways: taking into account gene interaction information + fold changes and adjusted p-values from differential expression analysis

KEGG pathway	P_{NDE}	P_{PERT}	P_G	P_{FDR}	P_{FWER}	Status
Focal adhe..4510	0.0001	0.0000	0.0000	0.00000	0.00000	Act.
ECM-recept..4512	0.0001	0.0004	0.0000	0.00001	0.00002	Act.
PPAR signa..3320	0.0000	0.1240	0.0000	0.00011	0.00034	Inh.
Alzheimers..5010	0.0000	0.7260	0.0001	0.00059	0.00235	Act.
Adherens j..4520	0.0001	0.0852	0.0001	0.00090	0.00452	Act.
Axon guida..4360	0.0002	0.2324	0.0006	0.00487	0.02922	Act.
MAPK signa..4010	0.0001	0.7112	0.0007	0.00504	0.03527	Inh.
Tight junc..4530	0.0007	0.5156	0.0032	0.02073	0.16585	Act.

$P_{NDE} = P(X \geq N_{DE} \mid H_0)$
P_{PERT}: probability to observe a larger perturbation than observed
P_G: combination of P_{NDE} and P_{PERT}
P_{FDR}: adjusted FDR p-value
P_{FWER}: adjusted FDR p-value (more conservative)

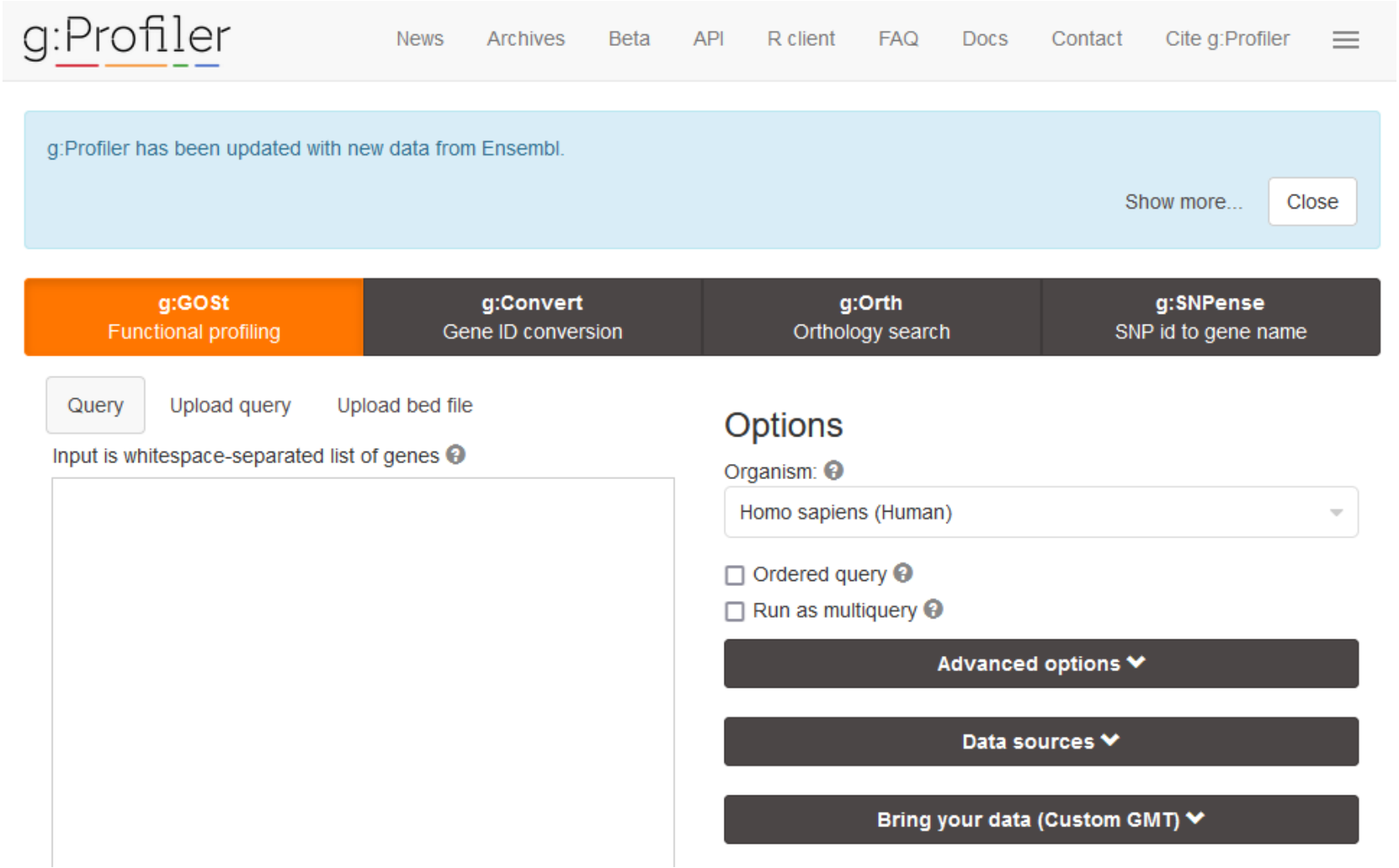
https://bioconductor.org/packages/release/bioc/html/SPIA.html

Functional analysis: **Pathway topology tools**



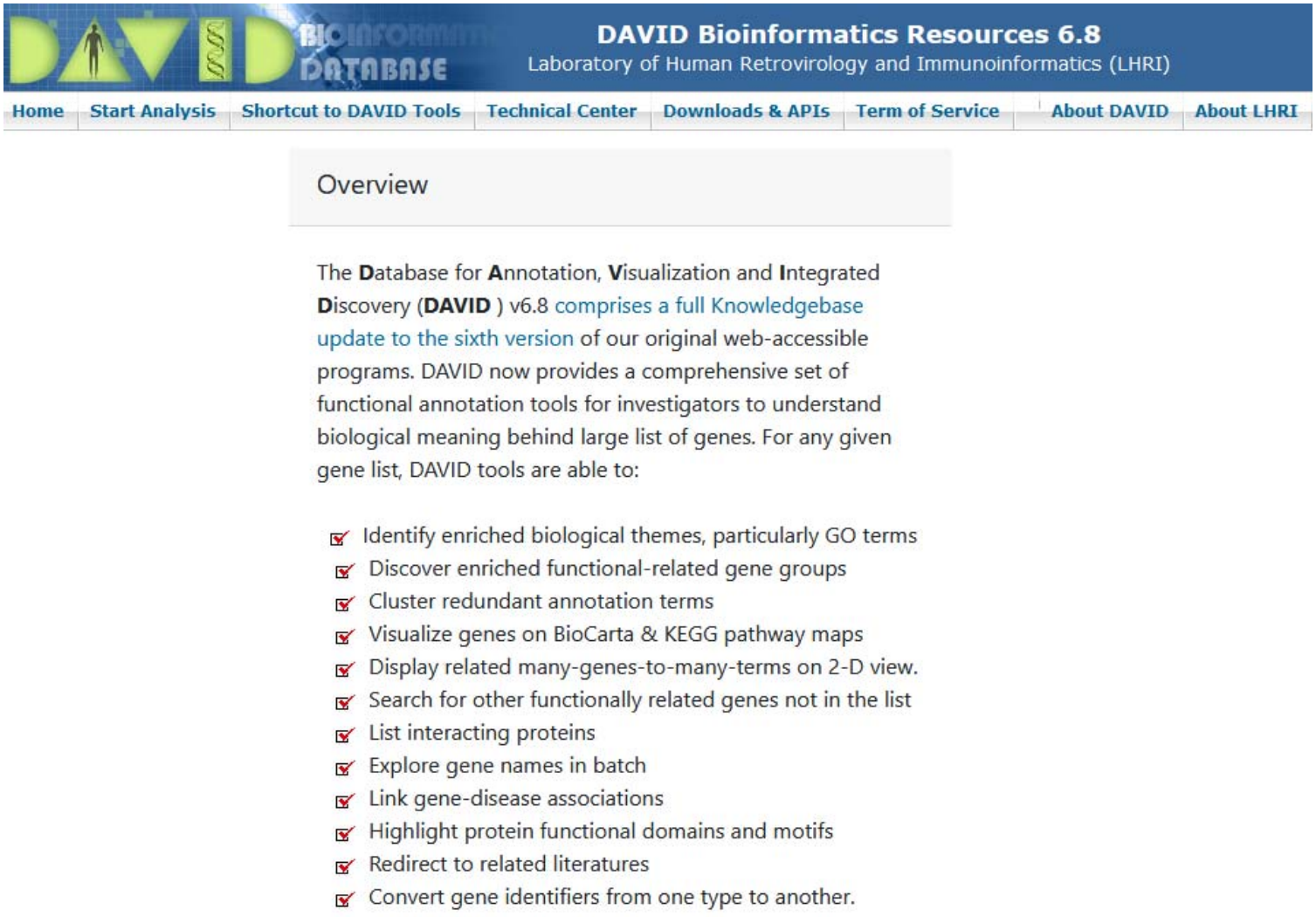
https://bioconductor.org/packages/release/bioc/html/SPIA.html

Additional resources for functional analysis



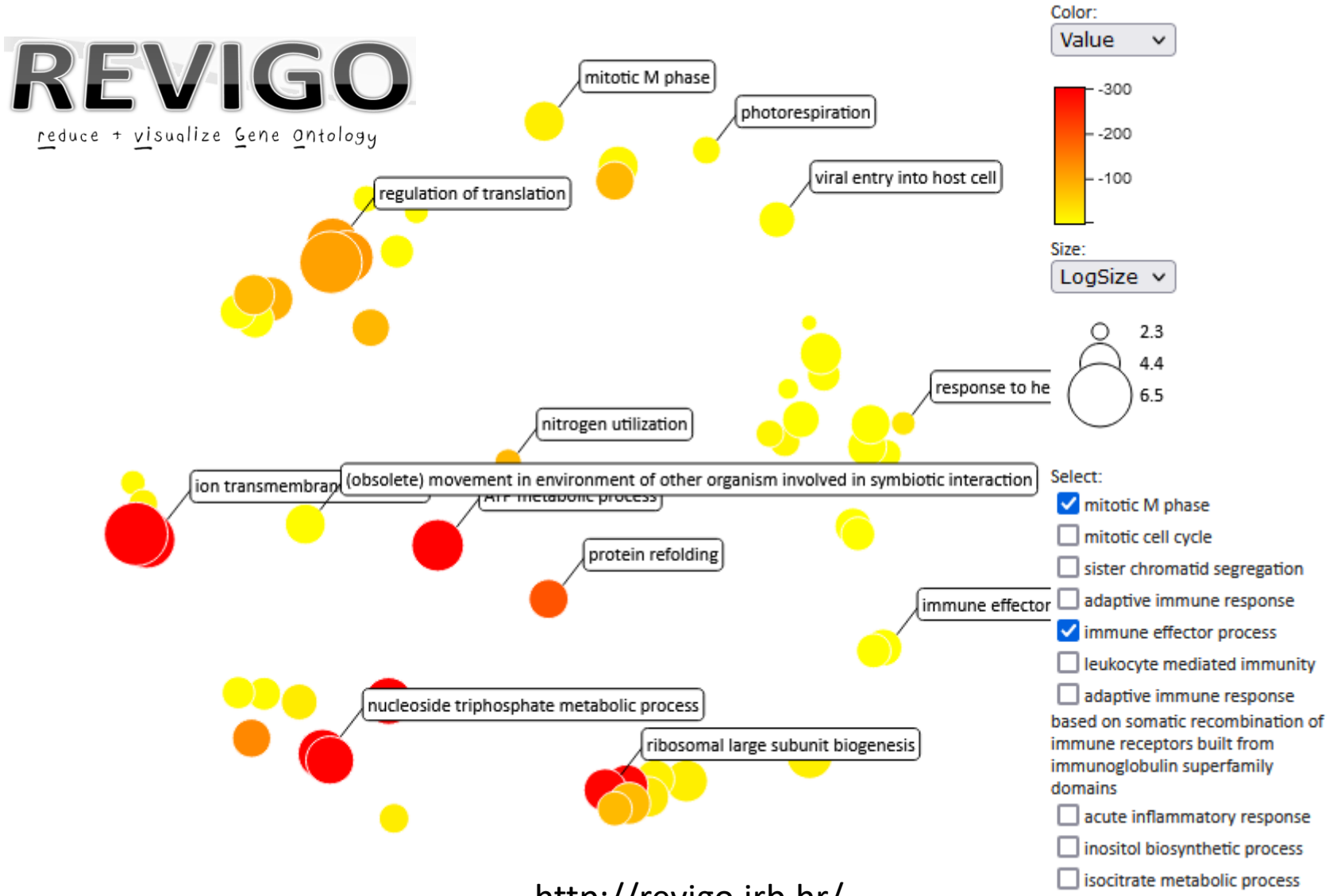
https://biit.cs.ut.ee/gprofiler/gost

Additional resources for functional analysis



https://david.ncifcrf.gov/home.jsp

Additional resources for functional analysis



http://revigo.irb.hr/

Additional resources for functional analysis

- g:Profiler - http://biit.cs.ut.ee/gprofiler/index.cgi
- DAVID - http://david.abcc.ncifcrf.gov/tools.jsp
- clusterProfiler - http://bioconductor.org/packages/release/bioc/html/clusterProfiler.html
- GeneMANIA - http://www.genemania.org/
- GenePattern - http://www.broadinstitute.org/cancer/software/genepattern/ (need to register)
- WebGestalt - http://bioinfo.vanderbilt.edu/webgestalt/ (need to register)
- AmiGO - http://amigo.geneontology.org/amigo
- ReviGO (visualizing GO analysis, input is GO terms) - http://revigo.irb.hr/
- WGCNA - http://www.genetics.ucla.edu/labs/horvath/CoexpressionNetwork
- GSEA - http://software.broadinstitute.org/gsea/index.jsp
- SPIA - https://www.bioconductor.org/packages/release/bioc/html/SPIA.html
- GAGE/Pathview - http://www.bioconductor.org/packages/release/bioc/html/gage.html

Recap and Exercise 4

- We have seen several types of visualization methods of functional enrichment results

Exercise 4: create the following figures:

- barplot of –log10(p-value) of top 10 GO p-values
- GSEA plot for HALLMARK MTORC1 SIGNALING
- pathview map for KEGG Natural Killer mediated cytotoxicity (optional: with none-significant genes in grey)

Some links

- Contact Tania if you wish to discuss enrichment analysis of your data more specifically:
  - tania.wyss@sib.swiss
- Contact the head of the Bioinformatics Core Facility if you need more extensive biostatics support:
  - mauro.delorenzi@sib.swiss

Links :

limma (for gene expression analysis and also includes functions for enrichment analysis):
https://www.bioconductor.org/packages/devel/bioc/vignettes/limma/inst/doc/usersguide.pdf

edgeR:
https://www.bioconductor.org/packages/release/bioc/vignettes/edgeR/inst/doc/edgeRUsersGuide.pdf

DESeq2:
http://bioconductor.org/packages/devel/bioc/vignettes/DESeq2/inst/doc/DESeq2.html

clusterProfiler:
https://yulab-smu.github.io/clusterProfiler-book/

bioconductor, introduction and structure
https://ivanek.github.io/analysisOfGenomicsDataWithR/02_IntroToBioc_html.html

online tool for overrepresentation analysis
http://www.pantherdb.org/

Credits: 0.25 ECTS

- Please provide results of exercises 2, 3 & 4 and answers to the following questions in a document:
  - Perform GSEA of the NK vs Th data using the Reactome gene sets downloaded on the MSigDB website (use minGSSize=30)
  - How many gene sets are significantly enriched? Generate an ordered barplot of the NES of all genesets, and generate a barcode plot for the gene set with the lowest NES

- Sign up for credit here:
  https://docs.google.com/document/d/1OT_1KDwr-7xKxwoNefKAnDTp4HPMr4UdNm2p6hmL-JI/edit#
- Send results to tania.wyss@sib.swiss

Thank you for your attention!

Please fill in the feedback available on the Moodle page:

https://edu.sib.swiss/course/view.php?id=550

Login: enrich21

Password: SIB-enrich21

We thank Linda Dib for providing course material