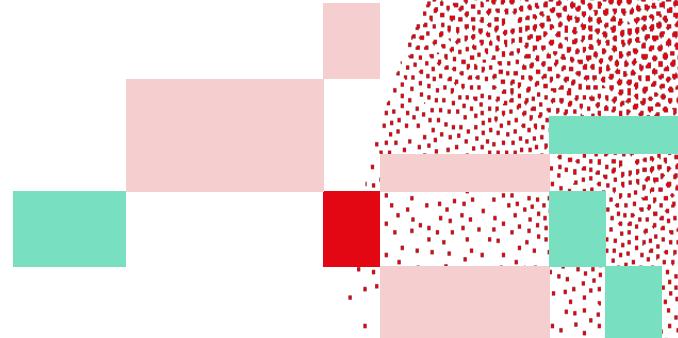




Swiss Institute of  
Bioinformatics



# Enrichment analysis

[tania.wyss@sib.swiss](mailto:tania.wyss@sib.swiss)

[gustavo.ruizbuendia@sib.swiss](mailto:gustavo.ruizbuendia@sib.swiss)

# Schedule

- **9:00 - 9:30**
- Introduction
- **9:30 – 10:30**
- Over-representation analysis
- Exercise
- **10:30ish** break
- **10:50 - 12:30**
- Method of gene set enrichment analysis
- Exercise
- **12:30ish - 13:30** lunch break
- **13:30 - 15:30**
- Visualization of enrichment results
- Exercise
- **15:30ish - 15:50** break
- **15:50 - 16:50**
- Ontologies and sources of gene sets
- Exercise
- **16:50 - 17:00** Feedback and end of day

# The Translational Data Science group



- Part of the **SIB Swiss Institute of Bioinformatics**
- Located at the AGORA Cancer Research Center in **Lausanne**
- Provides **the statistics, bioinformatics and computational expertise** to molecular biology and applied research labs.
- Participates in fundamental and translational research by providing expertise in **data analysis** of single-cell and bulk multi-omics, spatial transcriptomics, flow cytometry, etc

For core facility service inquiry: [nadine.fournier@sib.swiss](mailto:nadine.fournier@sib.swiss)

<https://agora-cancer.ch/scientific-platforms/translational-data-science-facility/>

<https://www.sib.swiss/raphael-gottardo-group>

# Tell us about yourself !

- Write your name and some keywords about yourself and/or your research into the shared doc, to share about yourself.
-  vevox poll



Photo by National Cancer Institute, Unsplash



Photo by Scott Graham, Unsplash

# Course material

- <https://sib-swiss.github.io/enrichment-analysis-training/>

The screenshot shows a website for 'Enrichment analysis'. At the top, there's a red header bar with the SIB logo and the text 'Enrichment analysis'. Below the header, on the left, is a sidebar with a navigation menu:

- Enrichment analysis
- Home
- Precourse preparations
- Course schedule
- Materials
- Exercises** (this item is highlighted in red)
- Bonus code
- Useful links

The main content area has a title 'Exercises' with a pencil icon. Below it is a text block: 'In this section, you will find the R code that we will use during the course. We will explain the code and output during correction of the exercises.' Further down, another section titled 'Source of data' contains text about a study by Ercolano et al 2020.

We will work with RNA sequencing data generated by [Ercolano et al 2020](#). This study described the transcriptomes of immune cells that are circulating in the blood of humans in healthy conditions. Different types of immune cells circulate in human blood. In this study, 2 cell types were included: Natural Killer (NK) cells and CD4+ T helper (Th) cells. These 2 types of cells have different functions: NK cells provide a rapid response in the innate immune response at the

- **Feedback:** survey at the end of the day about your opinion on this course (link sent by course organizer).

# Credits: 0.25 ECTS

- Please provide answers and R code for an additional exercise (eg 1 Word with answers and figures and 1 script file, or 1 file generated using Rmarkdown)  
<https://sib-swiss.github.io/enrichment-analysis-training/exercises/#extra-exercise-for-ects-credits>
- Sign up for credit by adding your name to the shared doc file
- Send answers to [tania.wyss@sib.swiss](mailto:tania.wyss@sib.swiss) within 1 week

# Questions and Exercises

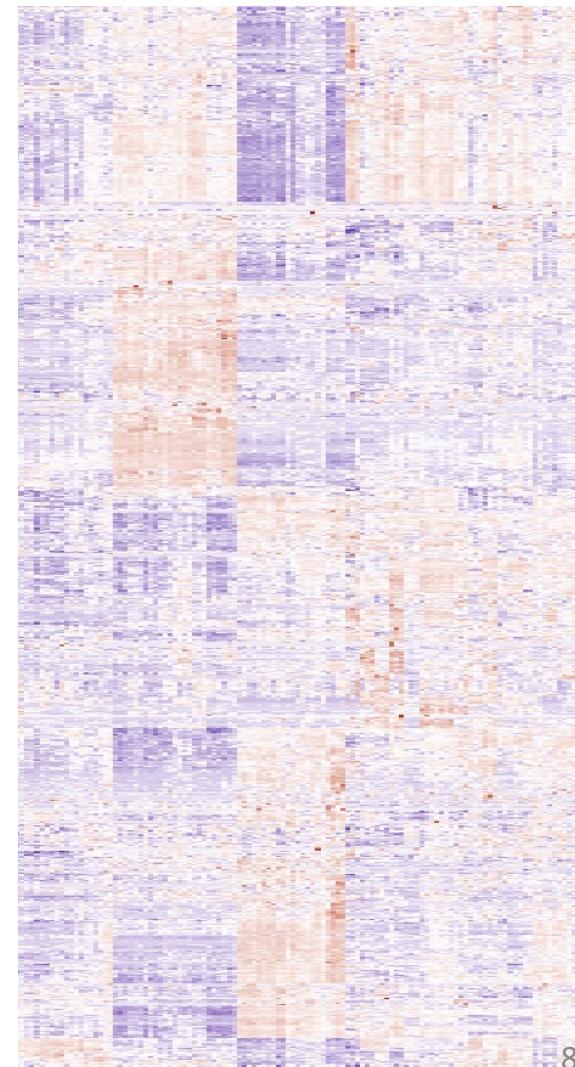
- Feel free to interrupt with questions by asking them directly or raising your (virtual) hand.
- Use the Q&A in shared doc (or Zoom chat), we will provide answers
- Add a  when you are done with the current exercise
- Exercises in R:
  - We will try to debug as much as possible
  - We are happy if you share your results or alternative code!



# Why do we perform enrichment analysis?

- Gene expression analysis yields hundreds to thousands of significant genes
  - We need to summarize the information provided by so many genes
  - Understand their biological relationships
  - Understand the genes' function (**functional analysis**)
  - Identify overarching biological processes or molecular pathways taking place in your system

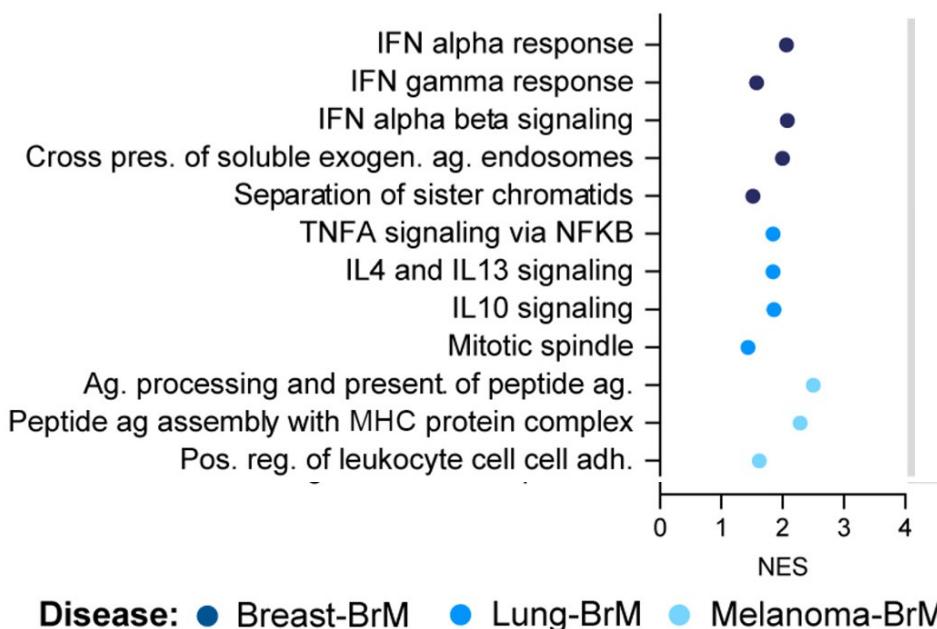
Some genes have similar expression pattern across samples



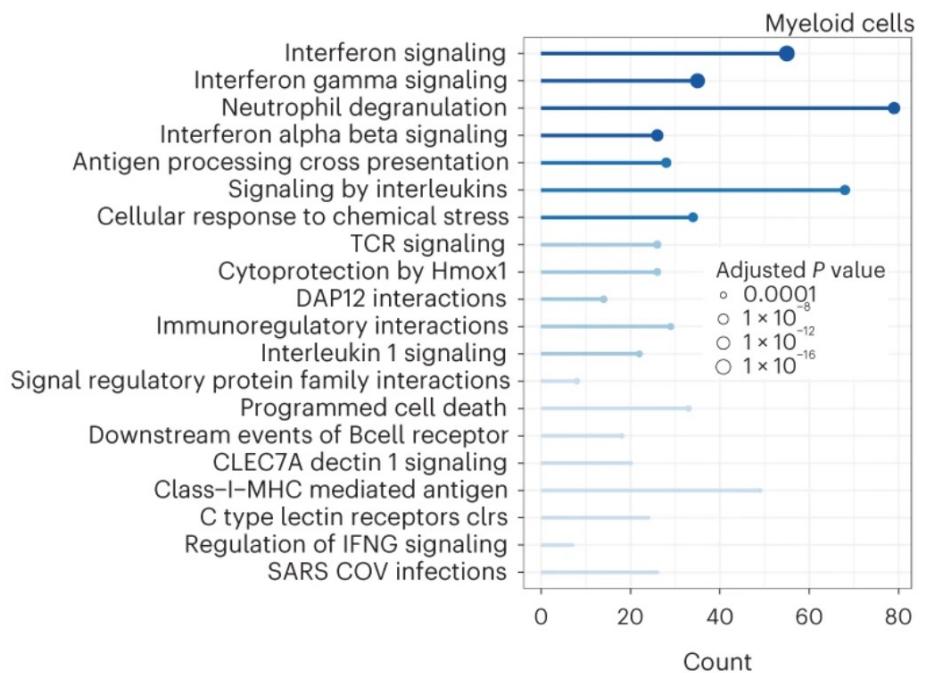
# Enrichment analysis in the literature – non-exhaustive examples

Often presented in *omics* studies

Different molecular alterations in vasculature of brain metastasis from different origins, compared to normal brain vasculature



Impact of a treatment on myeloid cells, pathways that could contribute to tumor growth limitation



Bulk RNAseq (GSEA)

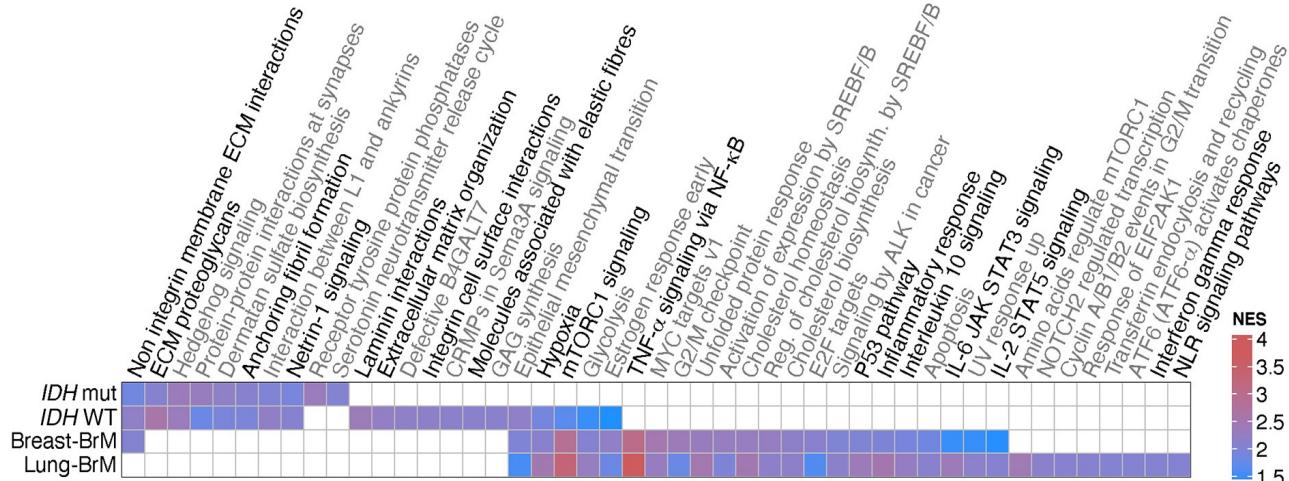
<https://doi.org/10.1016/j.ccell.2023.12.018>

Single-cell RNAseq (ORA)

<https://doi.org/10.1038/s43018-023-00668-y><sup>9</sup>

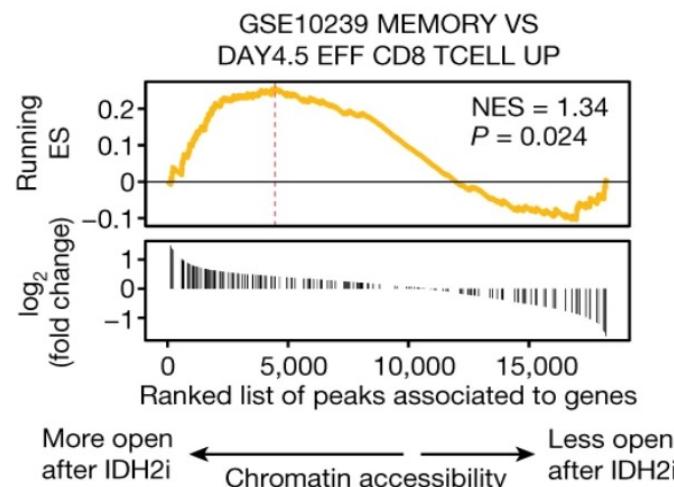
# Enrichment analysis in the literature – non-exhaustive examples

Neutrophils (immune cells) express different pathways depending on the brain tumor genotype (mut/WT) or origin (primary vs metastatic tumor)



Bulk RNAseq (GSEA)

<https://doi.org/10.1016/j.cell.2023.08.043>



Increased memory phenotype in immune cells exposed to a component.

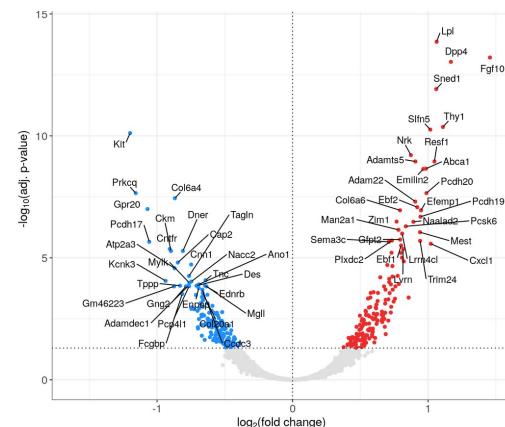
Bulk ATACseq (GSEA)

<https://doi.org/10.1038/s41586-023-06546-y>

# Enrichment analysis – input data

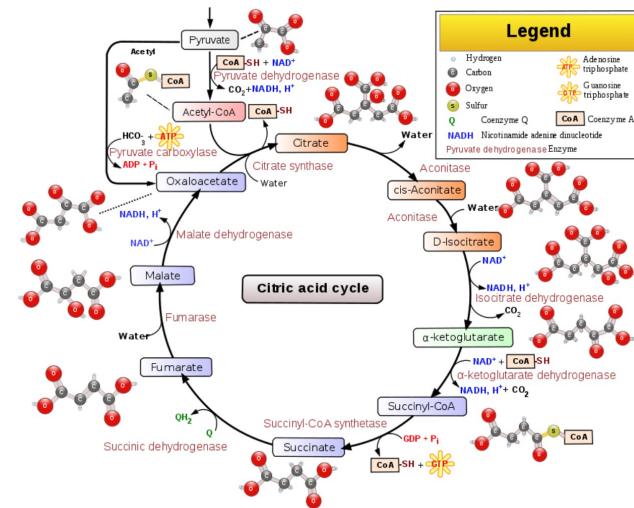
List of genes/proteins that are:

- Differentially expressed between 2 conditions
- Similar expression pattern across samples
- ...
- Either available as a list of gene symbols/IDs or with a score associated to each gene: e.g. T statistic or fold change



Database of gene/protein functional annotation

- Genes need to be grouped into gene sets/pathways/functional annotations.
- Consortia of researchers usually create these gene groupings/annotations



# Enrichment analysis - three major steps

- Obtain a gene/protein list from omics data
- Apply statistical methods to identify pathways enriched in the gene list relative to what is expected by chance
- Visualize and interpret the results

List of genes of interest

Statistical methods to determine enriched pathways

Create figures

# Enrichment analysis in non-model organisms

- Need **functional annotation** of genes: genes need to be grouped into pathways/functions.
- If not available, **convert your genes into the orthologs** of a closely related species that has such a database.
- Will require effort to find a gene functional annotation database. **All statistical analyses** are otherwise the same.

See Useful links:

<https://sib-swiss.github.io/enrichment-analysis-training/links/#tools-for-species-other-than-human-or-mouse>

# Approaches used in enrichment analysis

Test your gene list for enrichment of:

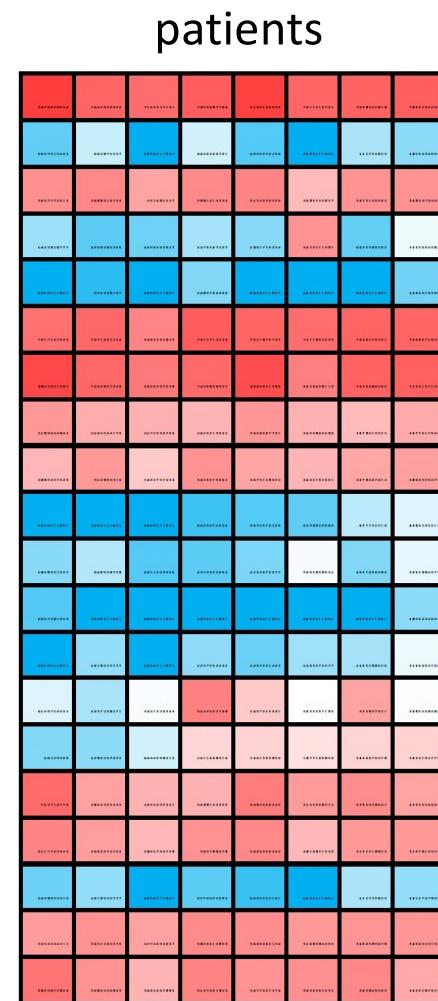
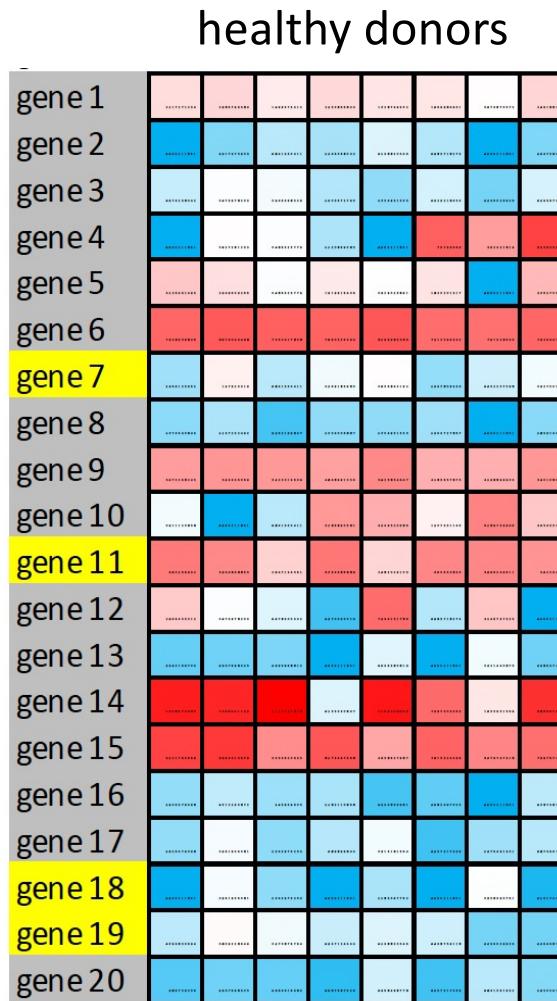
- Genes associated with a particular function or pathway (targeted)
- Genes annotated into a large collection of gene sets (exploratory)

Statistical methods available (covered today):

- over-representation analysis (ORA)
- gene set enrichment analysis (GSEA)

# Over-representation analysis (ORA)

Are the DE genes overlapping with the genes contained within the **yellow set**?



sort based  
on T-statistic or  
fold change  
and count  
significant genes  
based on p-value



gene 20	8.86
gene 7	7.98
gene 16	7.72
gene 8	6.78
gene 17	6.50
gene 1	6.47
gene 3	4.44
gene 19	1.54
gene 13	0.48
gene 2	0.39
gene 6	-0.06
gene 9	-0.11
gene 18	-0.55
gene 10	-2.40
gene 15	-3.67
gene 4	-5.09
gene 11	-6.52
gene 5	-6.59
gene 12	-6.60
gene 14	-6.65

gene 20	8.86
gene 7	7.98
gene 16	7.72
gene 8	6.78
gene 17	6.50
gene 1	6.47
gene 3	4.44
gene 19	1.54
gene 13	0.48
gene 2	0.39
gene 6	-0.06
gene 9	-0.11
gene 18	-0.55
gene 10	-2.40
gene 15	-3.67
gene 4	-5.09
gene 11	-6.52
gene 5	-6.59
gene 12	-6.60
gene 14	-6.65

Up-regulated & p<0.05  
Down-regulated & p<0.05

# Fisher's exact test

2 x 2 count table	Up-regulated	Not up-regulated	Total
Yellow	2	2	4
Not yellow	6	9	15
Total	8	11	19

contingency table

$H_0$ : The proportion of yellow genes up-regulated is the same as the proportion of yellow genes that are not up-regulated.

$H_1$ : The proportion of yellow genes up-regulated is not the same as the proportion of yellow genes that are not up-regulated.

# Fisher's exact test in R

```
> cont.table <- matrix(c(2, 2, 6, 9), ncol=2, byrow = T)  
> fisher.test(cont.table)
```

Fisher's Exact Test for Count Data

data: cont.table

p-value = 1

alternative hypothesis: true odds ratio is not equal to 1

95 percent confidence interval:

0.0842889 25.7046974

sample estimates:

odds ratio

1.467696

2 x 2 count table		Up-regulated	Not up-regulated	Total
Yellow		2	2	4
Not yellow		6	9	15
Total		8	11	19

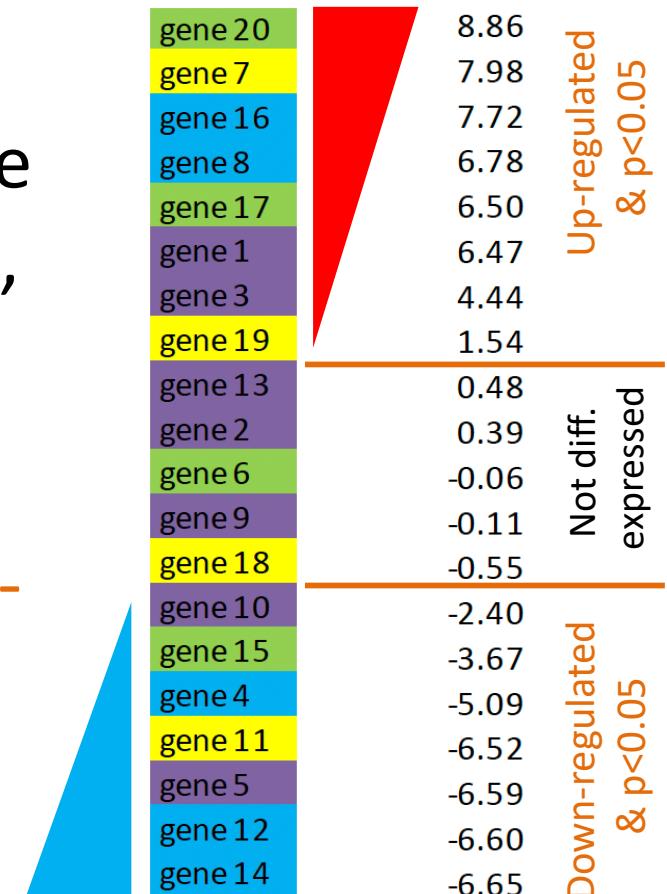
$$\begin{aligned}2/8 &= \\0.25\end{aligned}$$

$$\begin{aligned}2/12 &= \\0.167\end{aligned}$$

# Which gene sets are differentially expressed?

gene 20	8.86
gene 7	7.98
gene 16	7.72
gene 8	6.78
gene 17	6.50
gene 1	6.47
gene 3	4.44
gene 19	1.54
gene 13	0.48
gene 2	0.39
gene 6	-0.06
gene 9	-0.11
gene 18	-0.55
gene 10	-6.27
gene 15	-6.30
gene 4	-6.50
gene 11	-6.52
gene 5	-6.59
gene 12	-6.60
gene 14	-6.65

Run individual Fisher's exact tests for each gene set, **yellow, blue, purple, green**



⇒ Multiple tests need p-value adjustment.

# Enrichment analysis using R: one possibility among others

## clusterProfiler

platforms all rank 41 / 2140 support 1 5 / 2 3 in Bioc 11 years  
build ok updated < 1 week dependencies 125

DOI: [10.18129/B9.bioc.clusterProfiler](https://doi.org/10.18129/B9.bioc.clusterProfiler)  

### A universal enrichment tool for interpreting omics data

Bioconductor version: Release (3.15)

This package supports functional characteristics of both coding and non-coding genomics data for thousands of species with up-to-date gene annotation. It provides a universal interface for gene functional annotation from a variety of sources and thus can be applied in diverse scenarios. It provides a tidy interface to access, manipulate, and visualize enrichment results to help users achieve efficient data interpretation. Datasets obtained from multiple treatments and time points can be analyzed and compared in a single run, easily revealing functional consensus and differences among distinct conditions.

Author: Guangchuang Yu [aut, cre, cph] , Li-Gen Wang [ctb], Erqiang Hu [ctb], Xiao Luo [ctb], Meijun Chen [ctb], Giovanni Dall'Olio [ctb], Wanqian Wei [ctb]

Maintainer: Guangchuang Yu <[guangchuangyu@gmail.com](mailto:guangchuangyu@gmail.com)>

## Built-in functions for enrichment analysis

Built-in gene sets for human, mouse, yeast, etc

Built-in GO and KEGG (see later)

- <https://bioconductor.org/packages/release/bioc/html/clusterProfiler.html>
- G Yu, LG Wang, Y Han, QY He. clusterProfiler: an R package for comparing biological themes among gene clusters. *OMICS: A Journal of Integrative Biology* 2012, 16(5):284-287. [doi:\[10.1089/omi.2011.0118\]\(http://dx.doi.org/10.1089/omi.2011.0118\)](https://doi.org/10.1089/omi.2011.0118)
- Full vignette: <http://yulab-smu.top/clusterProfiler-book/>

# Functions for Fisher test and for ORA with R and clusterProfiler

Fisher exact test (package stats)

```
fisher.test(x, y = NULL, workspace = 200000, hybrid = FALSE,  
           hybridPars = c(expect = 5, percent = 80, Emin = 1),  
           control = list(), or = 1, alternative = "two.sided",  
           conf.int = TRUE, conf.level = 0.95,  
           simulate.p.value = FALSE, B = 2000)
```

enricher(): implementation of hypergeometric test (one-sided Fisher test) for user defined gene list and gene set collections (package clusterProfiler)

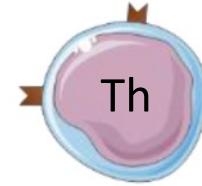
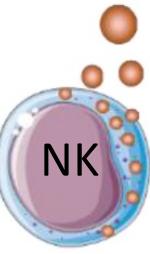
```
enricher(  
  gene,  
  pvalueCutoff = 0.05,  
  pAdjustMethod = "BH",  
  universe = NULL,  
  minGSSize = 10,  
  maxGSSize = 500,  
  qvalueCutoff = 0.2,  
  gson = NULL,  
  TERM2GENE,  
  TERM2NAME = NA  
)
```

TERM2GENE:  
A 2-column  
data frame

term	gene
GOBP_ADAPTIVE_IMMUNE_RESPONSE	ZC3H12A
GOBP_ADAPTIVE_IMMUNE_RESPONSE	ZNF683
GOBP_ADAPTIVE_IMMUNE_RESPONSE	ZP3
GOBP_HAIR_CELL_DIFFERENTIATION	ATOH1
GOBP_HAIR_CELL_DIFFERENTIATION	CDH23
GOBP_HAIR_CELL_DIFFERENTIATION	CLRN1

Eg genes that are markers of cell clusters of single-cell RNA seq





# Recap and exercise 1

- Once we have identified differentially expressed (DE) genes, we can use an over-representation analysis to determine whether or not the genes of a gene set of interest are over-represented among the DE genes or not.
- Exercise 1:**
- Results table of differential gene expression analysis between 2 human immune cell types, natural killer (NK) cells and CD4 T helper cells (Th):

ensembl_gene_id	symbol	logFC	t	P.Value	p.adj
ENSG00000000003	TSPAN6	-5.643604444	-4.67212847	4.260000e-05	7.358019e-04
ENSG00000000419	DPM1	-0.181898089	-1.10183079	2.780198e-01	5.176076e-01
ENSG00000000457	SCYL3	0.496987374	1.49103508	1.448691e-01	3.449889e-01
ENSG00000000460	C1orf112	1.121799095	1.44589945	1.570599e-01	3.630935e-01
ENSG00000000938	FGR	10.670687340	7.21234165	1.980000e-08	1.718657e-06
ENSG00000000971	CFH	-3.412927673	-2.78888655	8.480300e-03	4.610083e-02

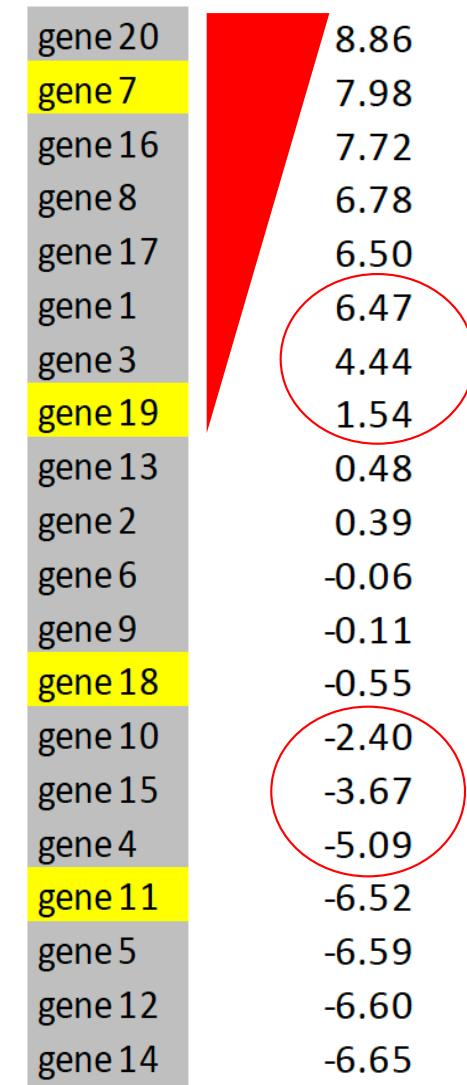
Positive logFC = higher in NK  
Negative logFC = lower in NK

- Run a **Fisher's exact test** to determine whether genes involved in the **adaptive immune response** are over-represented among the genes up-regulated in Th cells.

# Fisher's exact test is threshold-based

2 x 2 count table	Up-regulated	Not up-regulated	Total
Yellow	2	2	4
Not yellow	6	9	15
Total	8	11	19

Contingency table with count of genes,  
does not take into account the **magnitude**  
of the change of each gene.

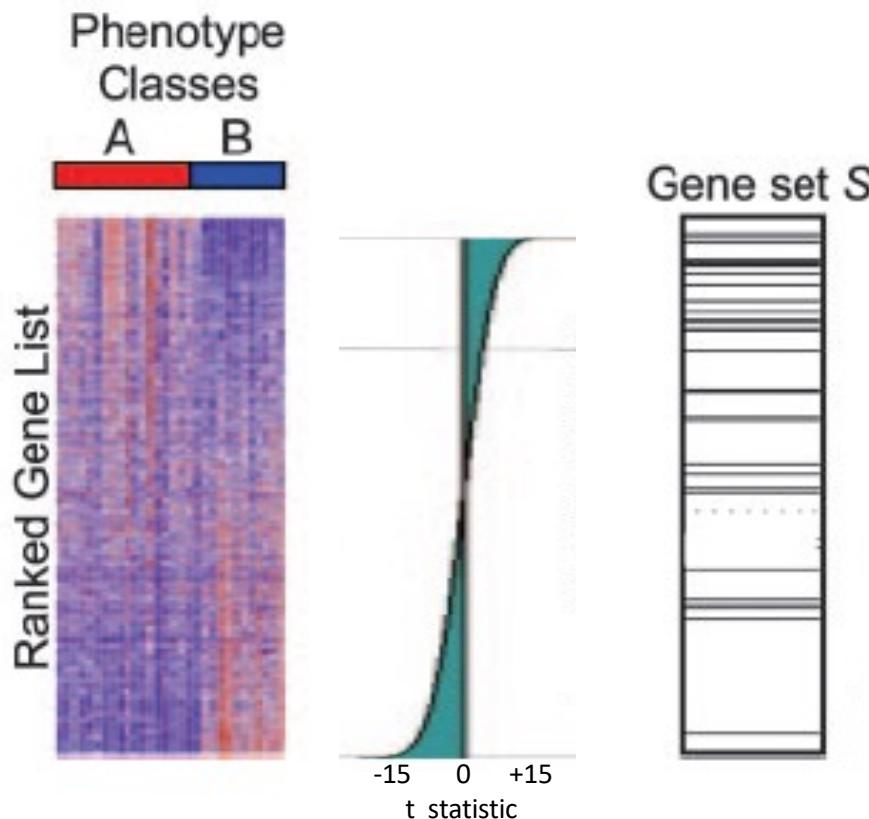


# Gene set enrichment analysis (GSEA)

- **Threshold-free:** the whole list of genes detected in the omics data is used.
- GSEA is a computational method that determines whether an *a priori* defined set of genes shows statistically significant, concordant differences between two biological states (MSigDB)
- Rank all genes based on score (eg t-statistic) and calculate an enrichment score (ES) that reflects the degree to which the members of a gene set are overrepresented at the top or bottom of the ranked genes.

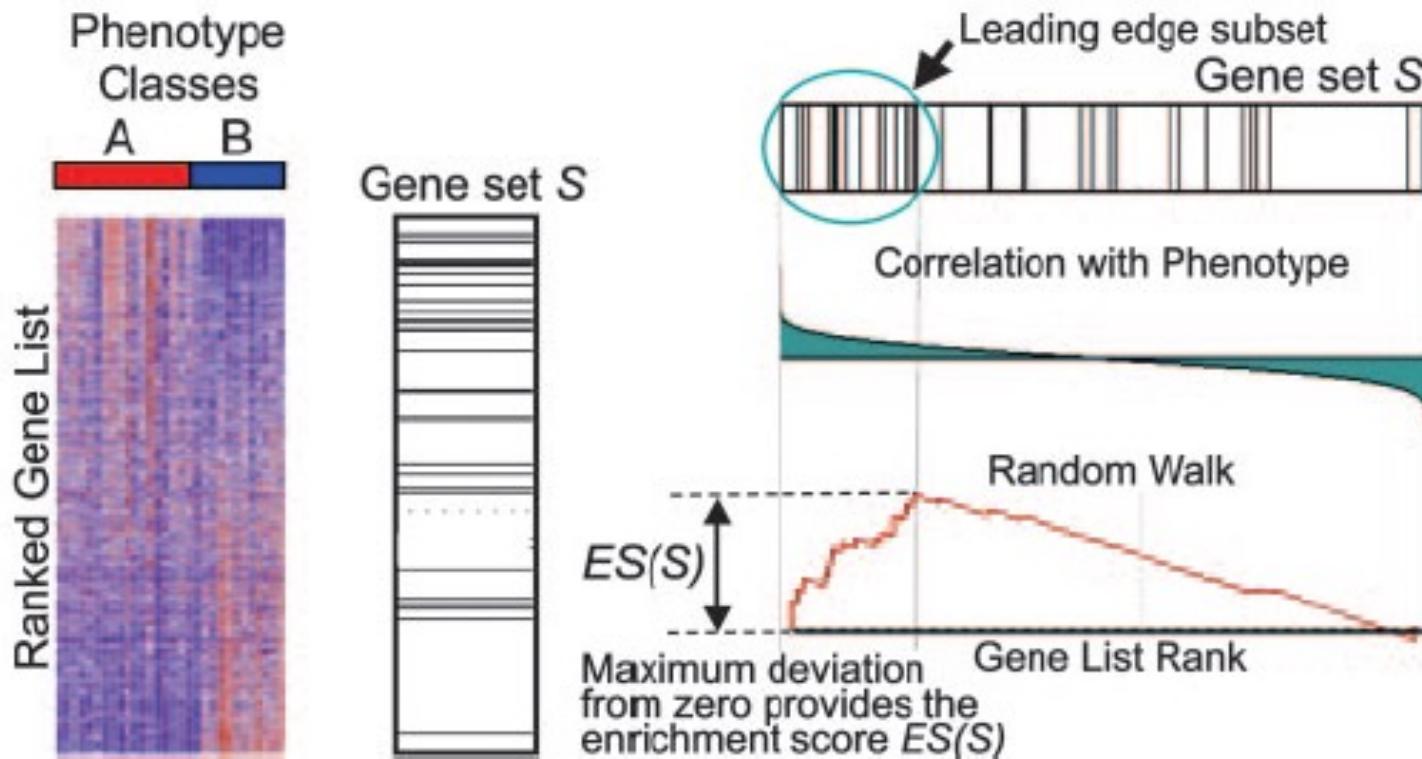
# Method of GSEA

Goal: determine whether the members of a gene set **S** are randomly distributed throughout a ranked gene list or if they are located at the top or bottom of the ranked gene lists



1. Sort the genes based on the t statistic (=weight)

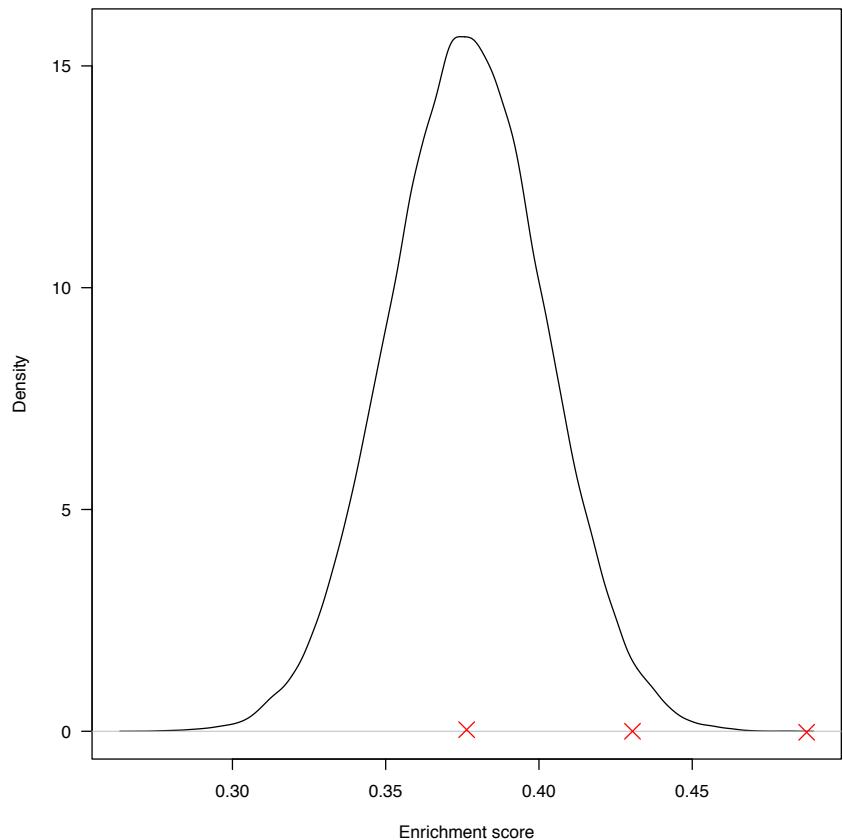
# Method of GSEA



1. Sort the genes based on the t statistic (=weight)
2. Calculate enrichment score ES using weight. The ES for a set is the maximum value reached (pos. or neg.)

# Method of GSEA

1. Sort the genes based on the t statistic (=weight)
2. Calculate enrichment score ES using weight. The ES for a set is the maximum value reached (pos. or neg.)
3. Perform permutations of samples and/or genes to recalculate random ES scores
4. Calculate Normalized ES (NES) and estimate p-value of each gene set based on randomized ES scores
5. Adjust p-value



$$\text{NES} = \frac{\text{actual ES}}{\text{mean(ESs against all permutations of the dataset)}}$$

Do not forget p-value  
adjustment if more than 1  
gene set is tested!

NES: 1    NES: 1.16    NES: 1.32  
p: 0.5    p: 0.05    p: 0.001

# Functions for GSEA with clusterProfiler

GSEA(): GSEA of user-defined gene sets using  
all ranked genes

```
GSEA(  
  geneList,  
  exponent = 1,  
  minGSSize = 10,  
  maxGSSize = 500,  
  eps = 1e-10,  
  pvalueCutoff = 0.05,  
  pAdjustMethod = "BH",  
  TERM2GENE,  
  TERM2NAME = NA,  
  verbose = TRUE,  
  seed = FALSE,  
  by = "fgsea",  
  ...  
)
```

gseGO(): GSEA of GO gene sets using  
all ranked genes

```
gseGO(  
  geneList,  
  ont = "BP",  
  OrgDb,  
  keyType = "ENTREZID",  
  exponent = 1,  
  minGSSize = 10,  
  maxGSSize = 500,  
  eps = 1e-10,  
  pvalueCutoff = 0.05,  
  pAdjustMethod = "BH",  
  verbose = TRUE,  
  seed = FALSE,  
  by = "fgsea",  
  ...  
)
```

# Bioconductor orgDb packages



Bioconductor  
OPEN SOURCE SOFTWARE FOR BIOINFORMATICS

About   Learn   Packages   Developers

<a href="#"><u>org.Sc.sgd.db</u></a>	Bioconductor Package Maintainer	Genome wide annotation for Yeast	42
<a href="#"><u>org.Ce.eg.db</u></a>	Bioconductor Package Maintainer	Genome wide annotation for Worm	45
<a href="#"><u>org.Bt.eg.db</u></a>	Bioconductor Package Maintainer	Genome wide annotation for Bovine	48
<a href="#"><u>org.Ss.eg.db</u></a>	Bioconductor Package Maintainer	Genome wide annotation for Pig	50
<a href="#"><u>org.Gg.eg.db</u></a>	Bioconductor Package Maintainer	Genome wide annotation for Chicken	51
<a href="#"><u>org.Cf.eg.db</u></a>	Bioconductor Package Maintainer	Genome wide annotation for Canine	52
<a href="#"><u>org.Mmu.eg.db</u></a>	Bioconductor Package Maintainer	Genome wide annotation for Rhesus	53
<a href="#"><u>org.Xl.eg.db</u></a>	Bioconductor Package Maintainer	Genome wide annotation for Xenopus	60



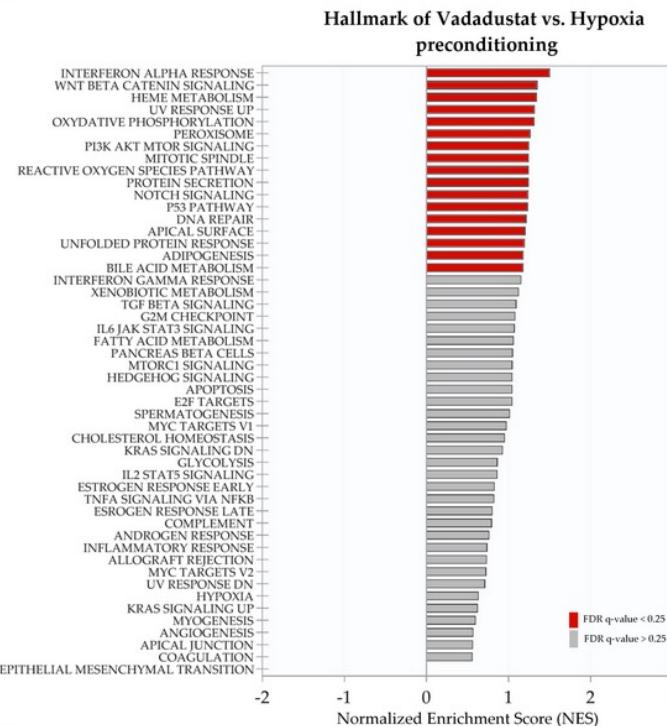
# Recap and exercise 2

- Fisher test is a threshold-based method, while GSEA is a threshold-free enrichment method. Both can be used for single or multiple gene sets.
- Exercise 2: use functions of `clusterProfiler` and data provided in Ex. 1
  - Run a GSEA for the Gene Ontology gene sets (more details on this collection later)
  - Explore the results: how many gene sets are significant? Are the gene sets up-regulated or down-regulated in NK cells?

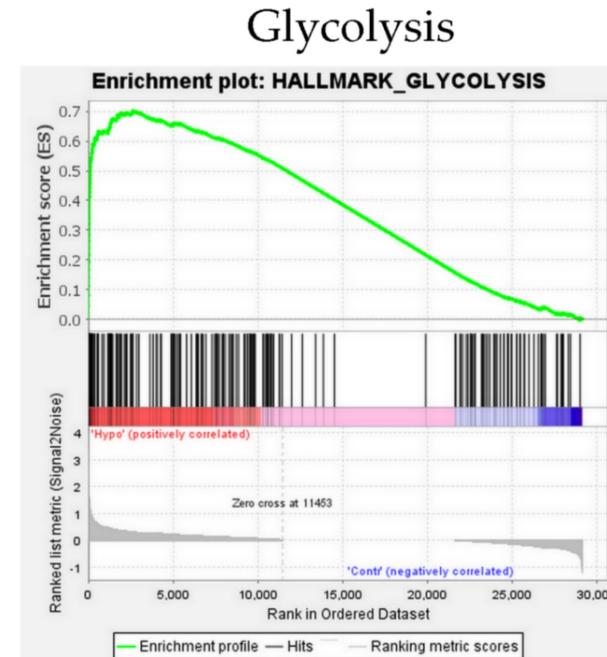
# Visualization of enrichment results

There are many options, here some common ones:

Barplot of NES for many pathways:



GSEA plot for a single pathway:



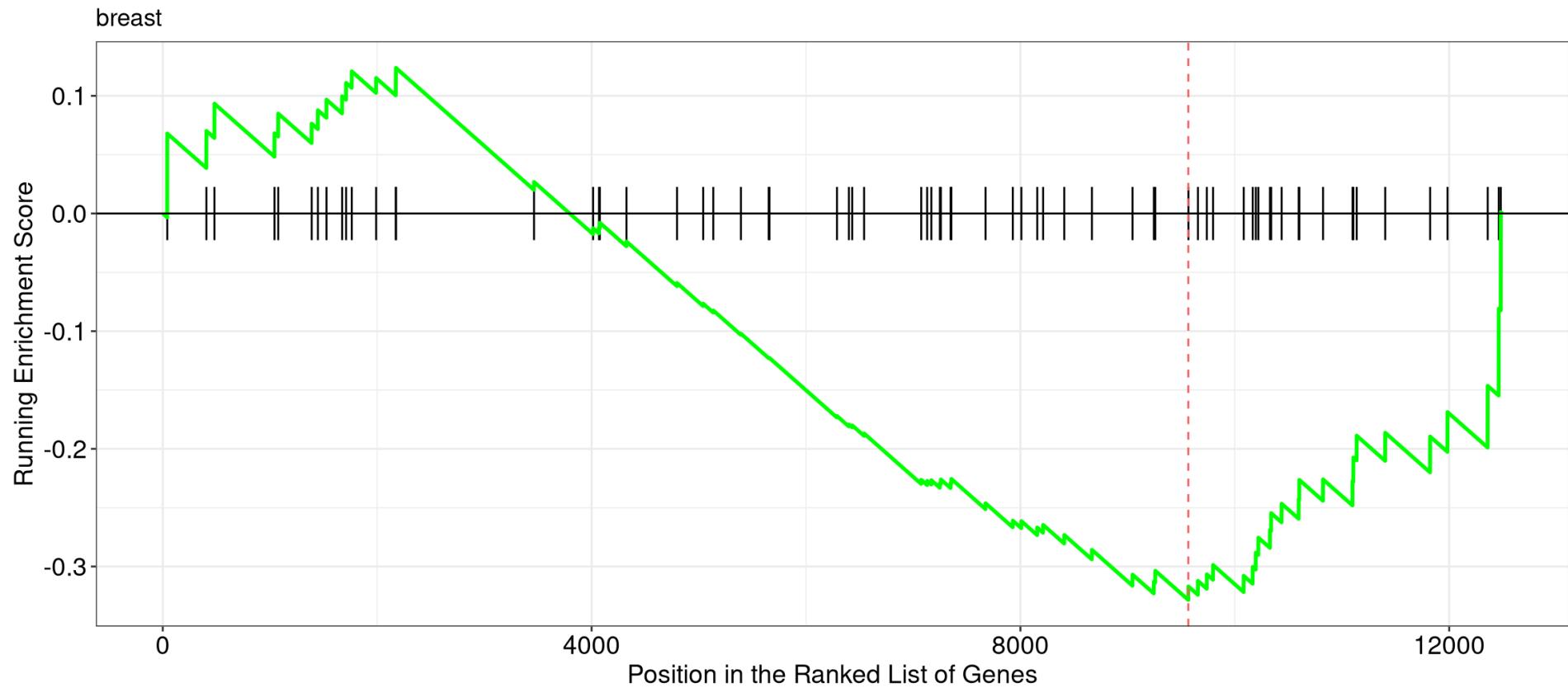
Via enrichplot: package for visualization using clusterProfiler objects

<https://www.bioconductor.org/packages/release/bioc/html/enrichplot.html>

# Visualizations available in clusterProfiler

GSEA plot (or barcode plot; for gseaResult objects)

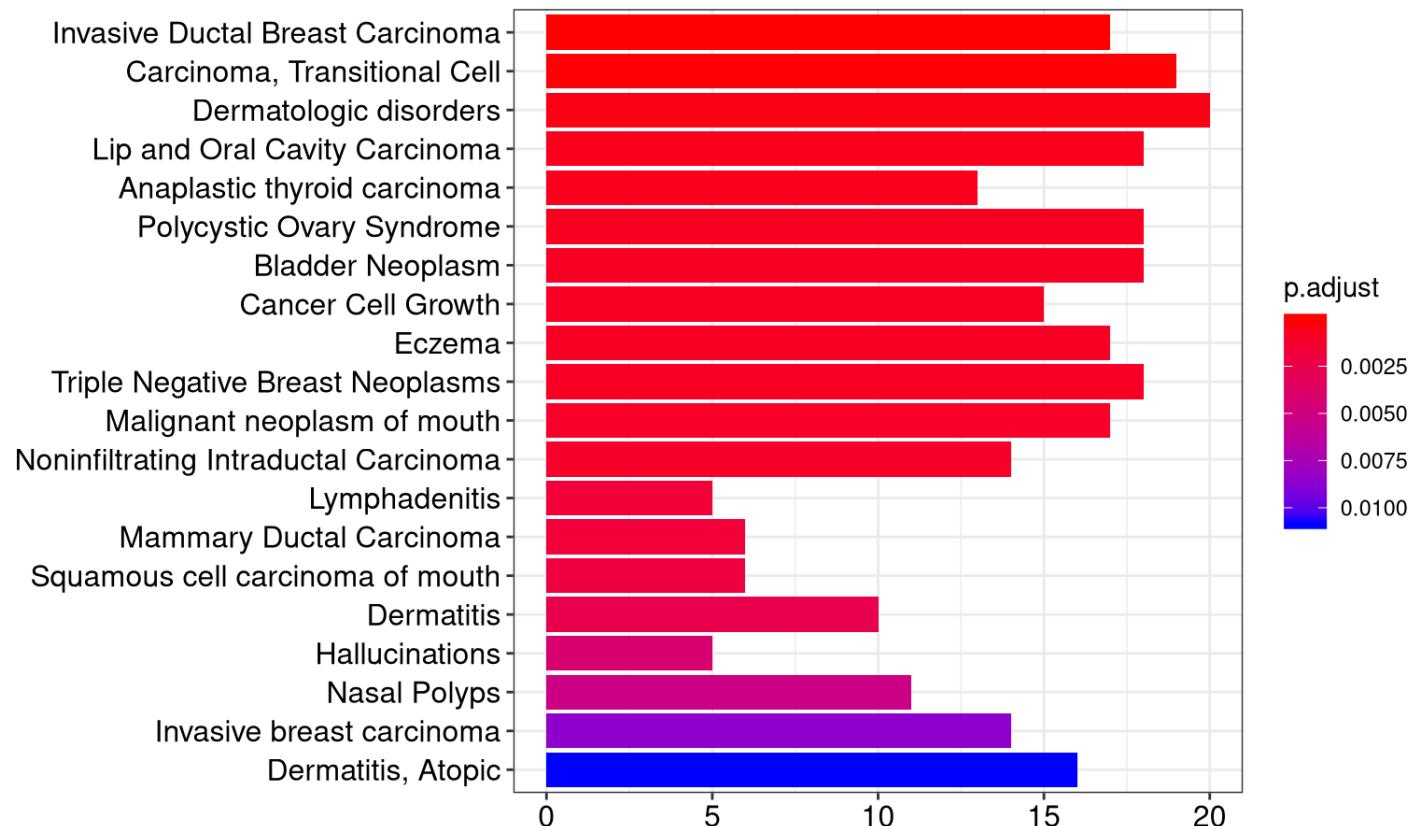
```
> gseaplot(h_NK_vs_Th, geneSetID = "breast", title=" breast")
```



# Visualizations available in clusterProfiler

**barplot** (from graphics package but works on enrichResult objects)

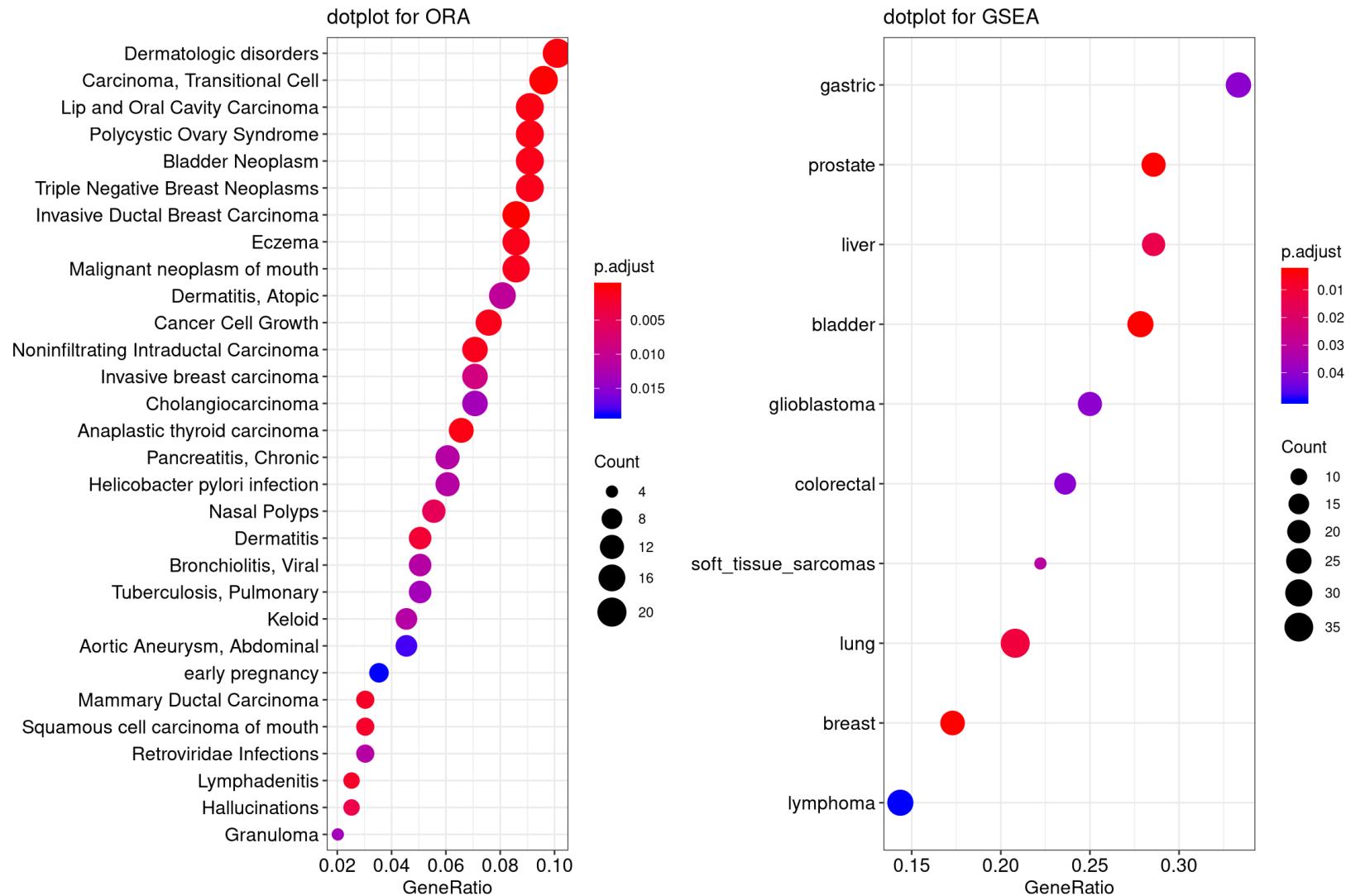
```
> ego <- enrichGO(de, OrgDb=org.Hs.eg.db, ont="BP", keyType = "SYMBOL")
> barplot(ego, showCategory=20)
```



# Visualizations available in clusterProfiler

## dotplot

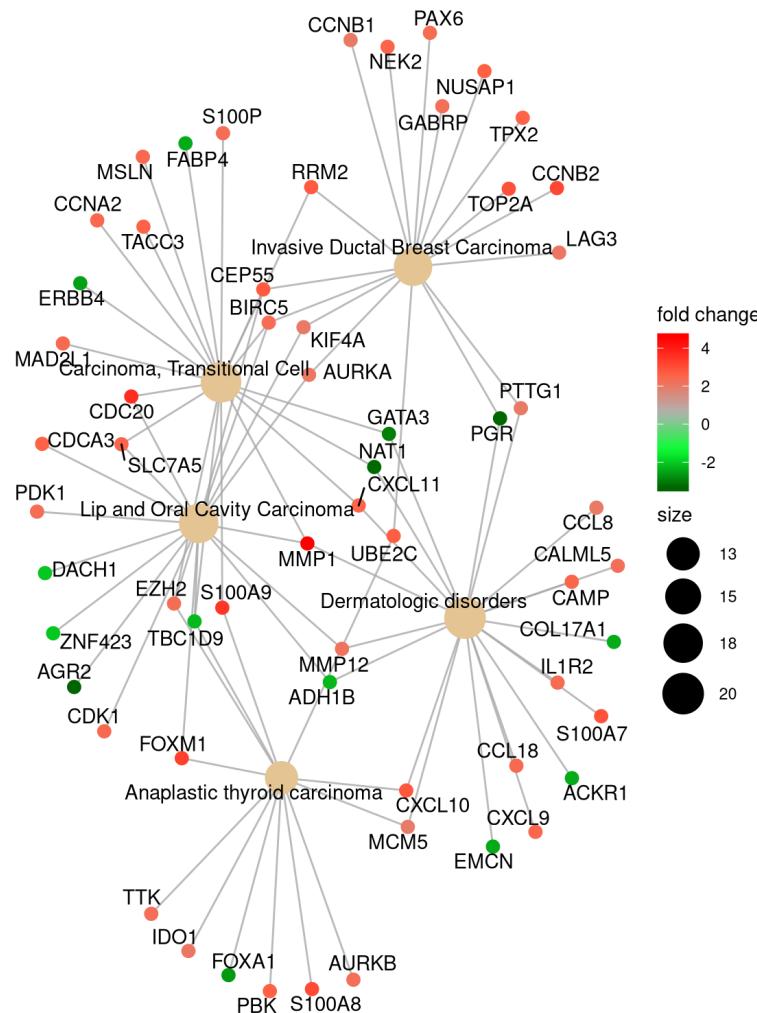
```
> ego <- enrichGO(de)
> dotplot(ego, showCategory=20)
```



# Visualizations available in clusterProfiler

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

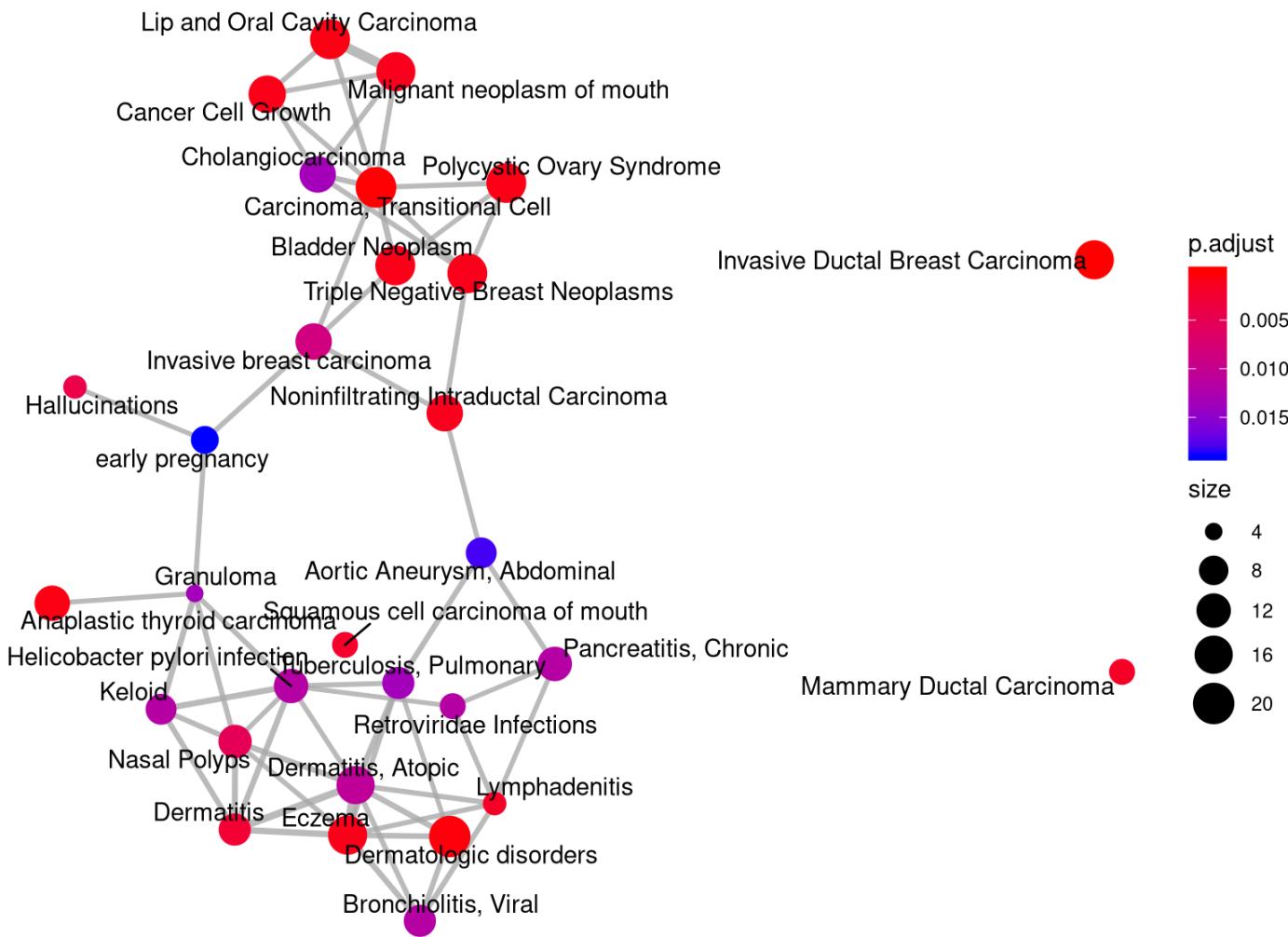
- Gene-concept network



# Visualizations available in clusterProfiler

- Enrichment map

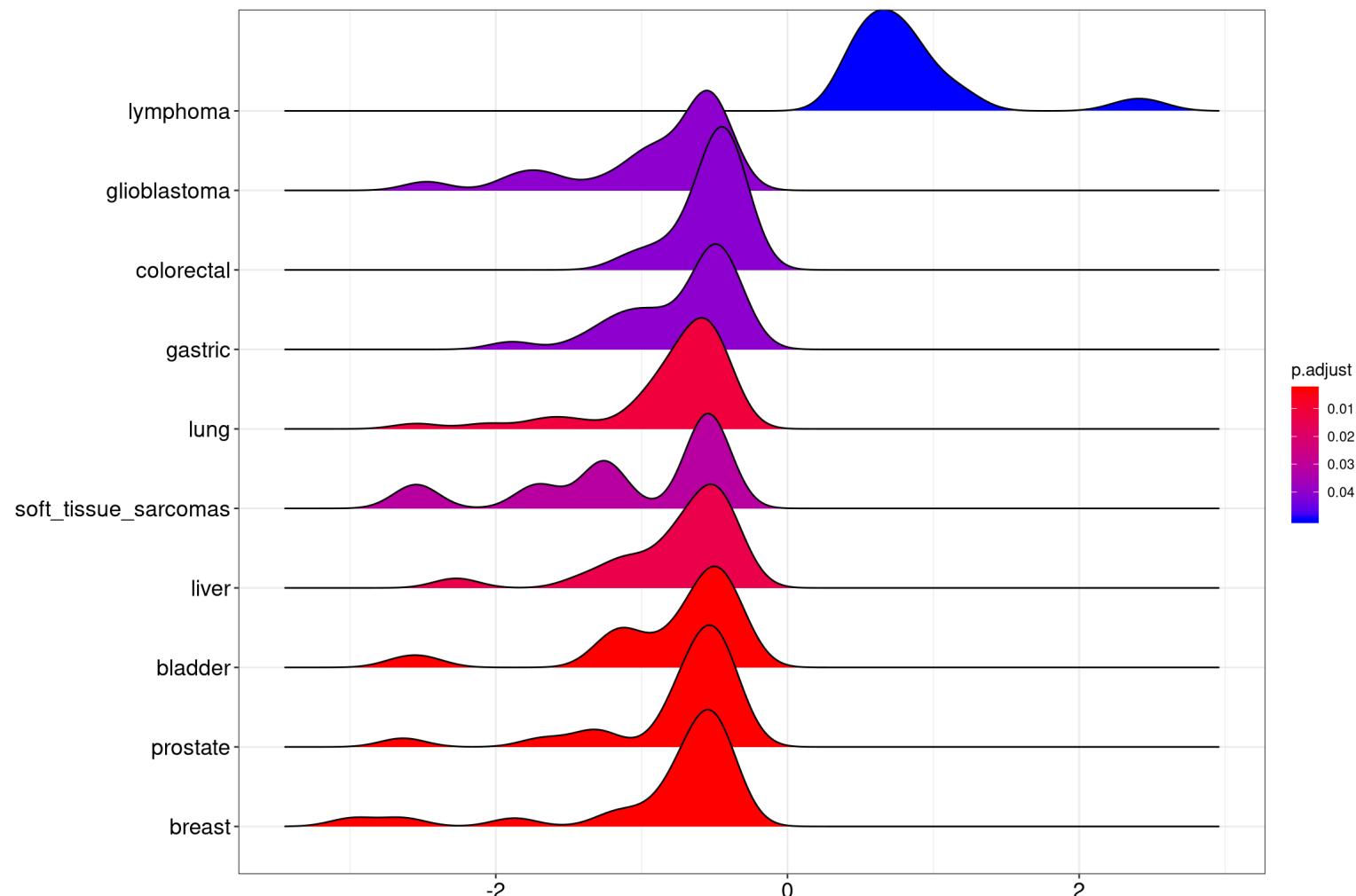
```
> ego <- enrichGO(de)  
> emapplot(ego)
```



# Visualizations available in clusterProfiler

- Ridgeplot

```
> ego <- gseGO(de)  
> ridgeplot(ego)
```





# Recap and Exercise 3

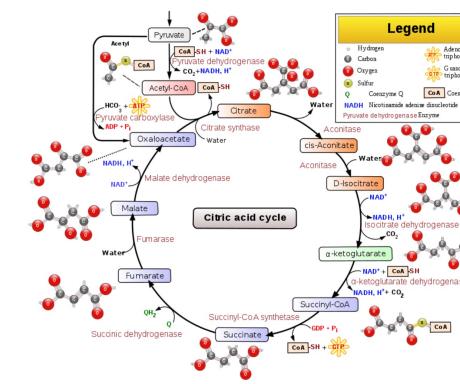
Several visualization methods can be used to represent the results, either for single gene sets (barcode plot) or for several gene sets (barplots, etc).

## Exercise 3: Create figures for the enrichment results:

- barplot of  $-\log_{10}(\text{p-value})$  of the p-values of the top 10 GO gene sets, or of positive and negative NES values
- Enrichment maps, gene-concept networks, ridge plots, etc

# What is a gene set?

- 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 SNPs associated with a disease
- ... etc!
- Several gene sets are grouped into Knowledge bases

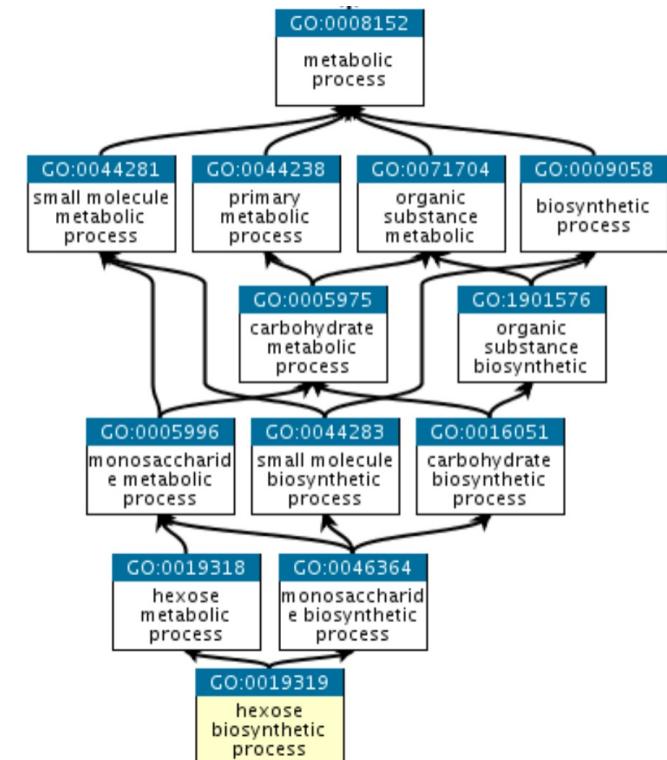


# Gene ontology

- <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)



The Gene Ontology (GO) knowledgebase is the world's largest source of information on the functions of genes.

# 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.

# KEGG

<https://www.genome.jp/kegg/pathway.html>

KEGG Databases Mapper Auto annotation Kanehisa Lab



## KEGG PATHWAY Database

Wiring diagrams of molecular interactions, reactions and relations

KEGG2 PATHWAY BRITE MODULE KO GENES COMPOUND DISEASE DRUG

Select prefix   Enter keywords   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

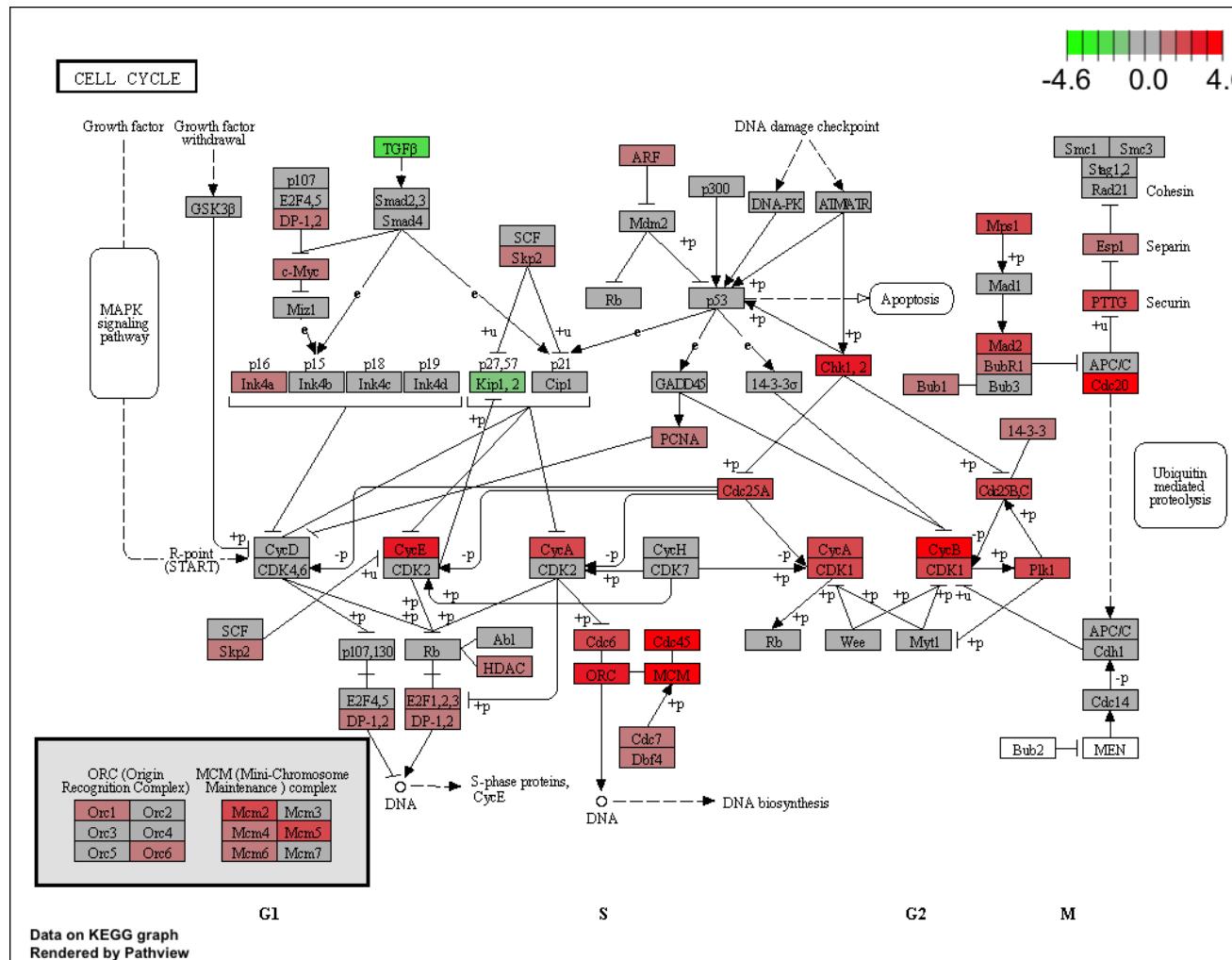
#### 7. Drug Development

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

# Visualization for KEGG pathways

## pathview package

```
> pathview(gene.data = geneList, pathway.id = "hsa04110", species = "hsa",
  limit = list(gene=max(abs(geneList)), cpd=1))
```



# Reactome

<https://reactome.org/>

The screenshot shows the Reactome website homepage. At the top left is the Reactome logo. At the top right are navigation links: About, Content, Docs, Tools, Community, and Download. Below the header is a search bar with the placeholder "Find Reactions, Proteins and Pathways" and a "Go!" button. The main content area features four large blue icons with white symbols: a network graph for the Pathway Browser, a bar chart for Analysis Tools, two interconnected nodes for ReactomeFIViz, and a document for Documentation.

Pathway Browser	Analysis Tools	ReactomeFIViz	Documentation
Visualize and interact with Reactome biological pathways	Merges pathway identifier mapping, over-representation, and expression analysis	Designed to find pathways and network patterns related to cancer and other types of diseases	Information to browse the database and use its principal tools for data analysis

# MSigDB

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

**H** **hallmark gene sets** are coherently expressed signatures derived by aggregating many MSigDB gene sets to represent well-defined biological states or processes.

**C1** **positional gene sets** for each human chromosome and cytogenetic band.

**C2** **curated gene sets** from online pathway databases, publications in PubMed, and knowledge of domain experts.

**C3** **regulatory target gene sets** based on gene target predictions for microRNA seed sequences and predicted transcription factor binding sites.

**C4** **computational gene sets** defined by mining large collections of cancer-oriented microarray data.

**C5** **ontology gene sets** consist of genes annotated by the same ontology term.

**C6** **oncogenic signature gene sets** defined directly from microarray gene expression data from cancer gene perturbations.

**C7** **immunologic signature gene sets** represent cell states and perturbations within the immune system.

**C8** **cell type signature gene sets** curated from cluster markers identified in single-cell sequencing studies of human tissue.

Download gmt files with version number:

<https://www.gsea-msigdb.org/gsea/downloads.jsp>

The Hallmark collection:

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

# msigdbr package

## Homologues for other species

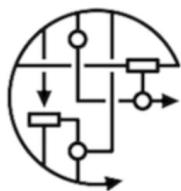
```
msigdbr_species()  
#> # A tibble: 20 × 2  
#>   species_name      species_common_name  
#>   <chr>                <chr>  
#> 1 Anolis carolinensis Carolina anole, green anole  
#> 2 Bos taurus        bovine, cattle, cow, dairy cow, domestic cattle, domes...  
#> 3 Caenorhabditis elegans <NA>  
#> 4 Canis lupus familiaris dog, dogs  
#> 5 Danio rerio       leopard danio, zebra danio, zebra fish, zebrafish  
#> 6 Drosophila melanogaster fruit fly  
#> 7 Equus caballus    domestic horse, equine, horse  
#> 8 Felis catus       cat, cats, domestic cat  
#> 9 Gallus gallus    bantam, chicken, chickens, Gallus domesticus  
#> 10 Homo sapiens    human  
#> 11 Macaca mulatta  rhesus macaque, rhesus macaques, Rhesus monkey, rhesus...  
#> 12 Monodelphis domestica gray short-tailed opossum  
#> 13 Mus musculus    house mouse, mouse  
#> 14 Ornithorhynchus anatinus  duck-billed platypus, duckbill platypus, platypus  
#> 15 Pan troglodytes chimpanzee
```

## Helper function to view available collections

```
msigdbr_collections()  
#> # A tibble: 23 × 3  
#>   gs_cat     gs_subcat      num_genesets  
#>   <chr>     <chr>          <int>  
#> 1 C1         ""              299  
#> 2 C2         "CPG"           3384  
#> 3 C2         "CP"             29  
#> 4 C2         "CP:BIOCARTA"    292  
#> 5 C2         "CP:KEGG"         186  
#> 6 C2         "CP:PID"          196  
#> 7 C2         "CP:REACTOME"      1615  
#> 8 C2         "CP:WIKIPATHWAYS"   664  
#> 9 C3         "MIR:MIRDB"        2377  
#> 10 C3        "MIR:MIRDB"        2377  
#> 11 C3        "MIR:MIRDB"        2377  
#> 12 C3        "MIR:MIRDB"        2377  
#> 13 C3        "MIR:MIRDB"        2377  
#> 14 C3        "MIR:MIRDB"        2377  
#> 15 C3        "MIR:MIRDB"        2377  
#> 16 C3        "MIR:MIRDB"        2377  
#> 17 C3        "MIR:MIRDB"        2377  
#> 18 C3        "MIR:MIRDB"        2377  
#> 19 C3        "MIR:MIRDB"        2377  
#> 20 C3        "MIR:MIRDB"        2377  
#> 21 C3        "MIR:MIRDB"        2377  
#> 22 C3        "MIR:MIRDB"        2377  
#> 23 C3        "MIR:MIRDB"        2377
```

# WikiPathways

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



search



- Help
- About us
- Contact us
- Report a bug
- How to cite
- Quality control
- Development
- WikiPathways Blog
- Curation Events

download

- Download files
- Other file formats
- Web service API
- WikiPathways RDF
- Embed code

activity

- Browse pathways
- Recent changes
- New releases
- New pathways
- Edit pathways

page discussion view source history Log in / create account

Share your pathway knowledge in the fight against COVID-19

ACCESS the rapidly growing collection of COVID-19 pathways, CONTRIBUTE your time and domain knowledge about pathway biology as a pathway author, and USE these pathways in your research.

Welcome to WikiPathways

WikiPathways is a database of biological pathways maintained by and for the scientific community. Read about our 12-year journey so far and [official exit from beta](#) or our [2021 NAR paper](#).

Find Pathways

Search

You can search by:

- Pathway name (*Apoptosis*)
- Gene or protein name (*p53*)
- Any page content (*cancer*)

Browse

Browse pathways

Browse by species and category

Today's Featured Pathway

COVID-19, thrombosis and anticoagulation (Homo sapiens)

Curator of the Week

Kristina Hanspers (Gladstone Institutes)

Get Pathways

Download

Growth

The screenshot shows the main interface of the WikiPathways website. At the top, there are navigation links for 'page', 'discussion', 'view source', and 'history'. On the right, there is a 'Log in / create account' link. A prominent red banner in the center says 'Share your pathway knowledge in the fight against COVID-19' and provides instructions for contributing. Below this, a green box welcomes users to 'Welcome to WikiPathways' and describes it as a database of biological pathways. It also links to the 2021 NAR paper. The main content area is divided into several sections: 'Find Pathways' with a search bar and options to search by pathway name, gene/protein name, or page content; 'Browse' with a grid of pathway thumbnails; 'Today's Featured Pathway' showing a complex network diagram for 'COVID-19, thrombosis and anticoagulation (Homo sapiens)'; 'Curator of the Week' featuring Kristina Hanspers from Gladstone Institutes; and 'Get Pathways' with download options and a growth chart showing the number of genes over time.

# GSEA of other gene sets in R

KEGG: ClusterProfiler built-in function for GSEA of 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")
```

User-defined gene set collection: Import a .gmt file of gene sets and convert to TERM2GENE data frame needed for clusterProfiler: `read.gmt(gmtfile)`

Converts a gmt text file with 1 gene set per line to a 2-column data frame:

```
!loads/h.all.v2024.1.Hs.symbols.gmt ◊  
h.all.v2024.1.Hs.symbols.gmt  
HALLMARKADIPOGENESIS https://www.gsea-msigdb.org/gsea/msigdb/human/geneset/HALLMARKADIPOGENESIS ABCA1 ABCB8 ACA2 ACADL  
HALLMARKALLOGRAFTREJECTION https://www.gsea-msigdb.org/gsea/msigdb/human/geneset/HALLMARKALLOGRAFTREJECTION AARS1 ABCE1  
HALLMARKANDROGENRESPONSE https://www.gsea-msigdb.org/gsea/msigdb/human/geneset/HALLMARKANDROGENRESPONSE ABCC4 ABHD2 AC  
HALLMARKANGIogenesis https://www.gsea-msigdb.org/gsea/msigdb/human/geneset/HALLMARKANGIogenesis APOM APP CCND2 COL3A1 CO  
HALLMARKAPICALJUNCTION https://www.gsea-msigdb.org/gsea/msigdb/human/geneset/HALLMARKAPICALJUNCTION ACTA1 ACTB ACTC1  
HALLMARKAPICALSURFACE https://www.gsea-msigdb.org/gsea/msigdb/human/geneset/HALLMARKAPICALSURFACE ADAM10 ADIPOR2 AFAP1L2 AF  
HALLMARKAPOPTOSIS https://www.gsea-msigdb.org/gsea/msigdb/human/geneset/HALLMARKAPOPTOSIS ADD1 AIMM3 ANKH ANXA1 AF  
HALLMARKBILEACIDMETABOLISM https://www.gsea-msigdb.org/gsea/msigdb/human/geneset/HALLMARKBILEACIDMETABOLISM ABCA1 ABCA2  
HALLMARKCHOLESTEROLHOMEOSTASIS https://www.gsea-msigdb.org/gsea/msigdb/human/geneset/HALLMARKCHOLESTEROLHOMEOSTASIS ABCA2  
HALLMARKCOAGULATION https://www.gsea-msigdb.org/gsea/msigdb/human/geneset/HALLMARKCOAGULATION A2M AC0X2 ADAM9 ANG ANXA1  
HALLMARKCOMPLEMENT https://www.gsea-msigdb.org/gsea/msigdb/human/geneset/HALLMARKCOMPLEMENT ACTN2 ADAM9 ADRA2B AKAP10 AN  
HALLMARKDNAREPAIR https://www.gsea-msigdb.org/gsea/msigdb/human/geneset/HALLMARKDNAREPAIR AAAS ADA ADCY6 ADMR1 AGO4  
HALLMARKE2FTARGETS https://www.gsea-msigdb.org/gsea/msigdb/human/geneset/HALLMARKE2FTARGETS AK2 ANP32E ASF1A ASF1B AT
```

```
> 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 (or biomaRt package)

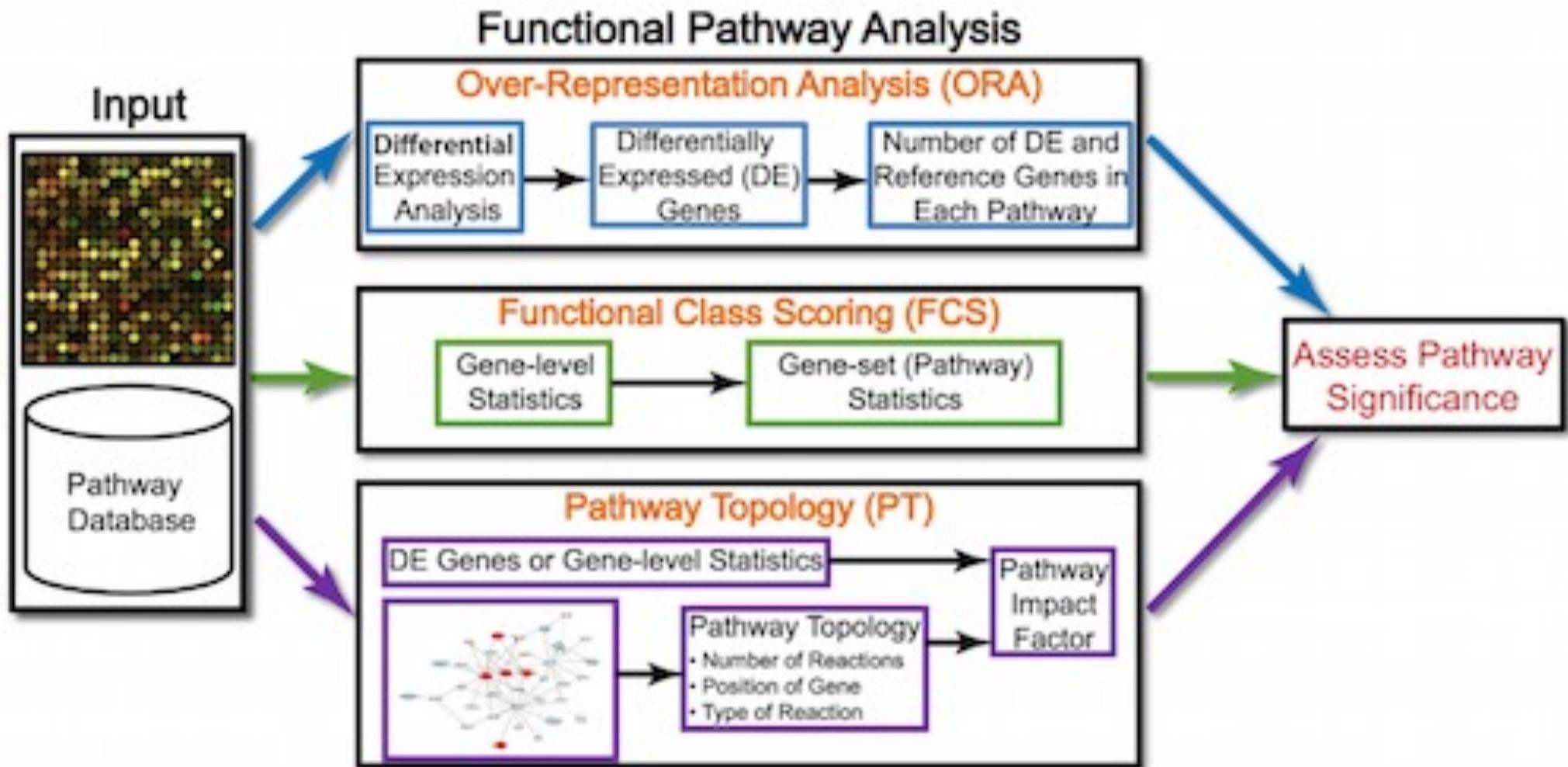
```
bitr(geneID, fromType, toType, OrgDb, drop = TRUE)
```

biomaRt: <https://bioconductor.org/packages/release/bioc/html/biomaRt.html>

# Recap and exercise 4

- 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 4: use functions of clusterProfiler and data provided in Ex. 1, and hallmark gene sets downloaded from MSigDB
  - First convert the gene symbols to EntrezID, then perform a GSEA of KEGG pathways (with argument minGSSize=30).
  - Explore the results. Is there an immune-related gene set coming up? Is there a Natural killer gene set coming up?
  - Using msigdbr, obtain a TERM2GENE data.frame of the Hallmark gene sets and run a GSEA. How many significant gene sets are there?

# Enrichment/functional analysis - summary



# 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.

$$P_{NDE} = P(X \geq N_{DE} | H_0)$$

$P_{PERT}$ : probability to observe a larger perturbation than observed

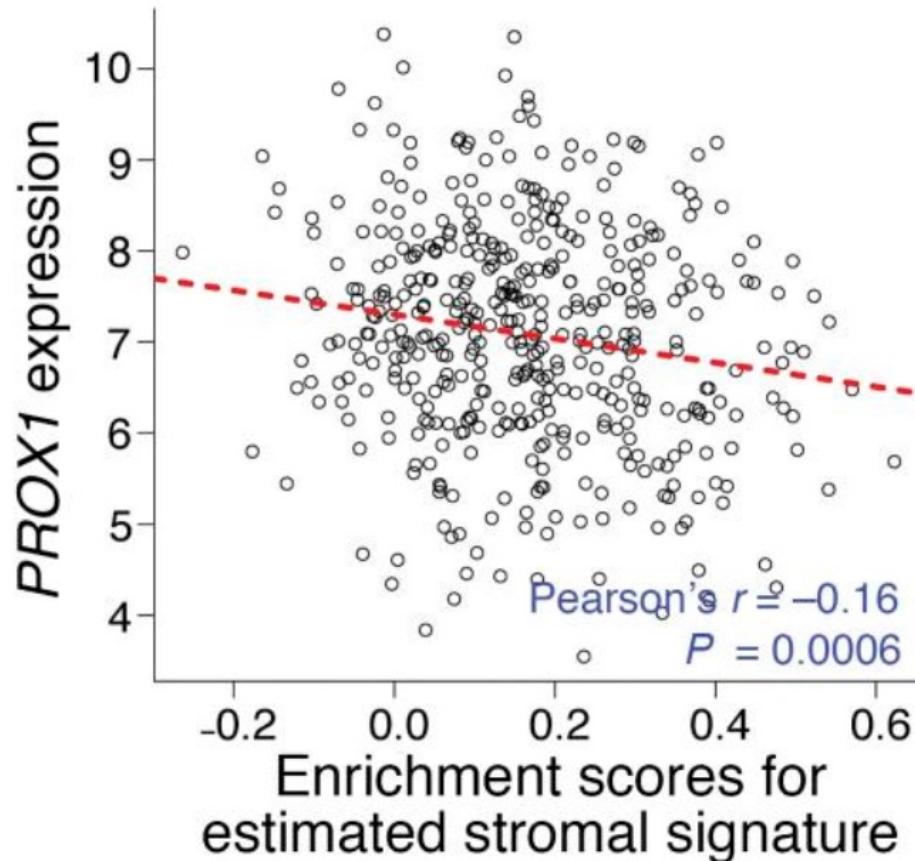
$P_G$ : combination of  $P_{NDE}$  and  $P_{PERT}$

$P_{FDR}$ : adjusted FDR p-value

$P_{FWER}$ : adjusted FDR p-value (more conservative)

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

# Single-sample gene set variation analysis



GSVA:

<https://bioconductor.org/packages/release/bioc/html/GSVA.html>

<https://www.jci.org/articles/view/129558>

# Credits: 0.25 ECTS

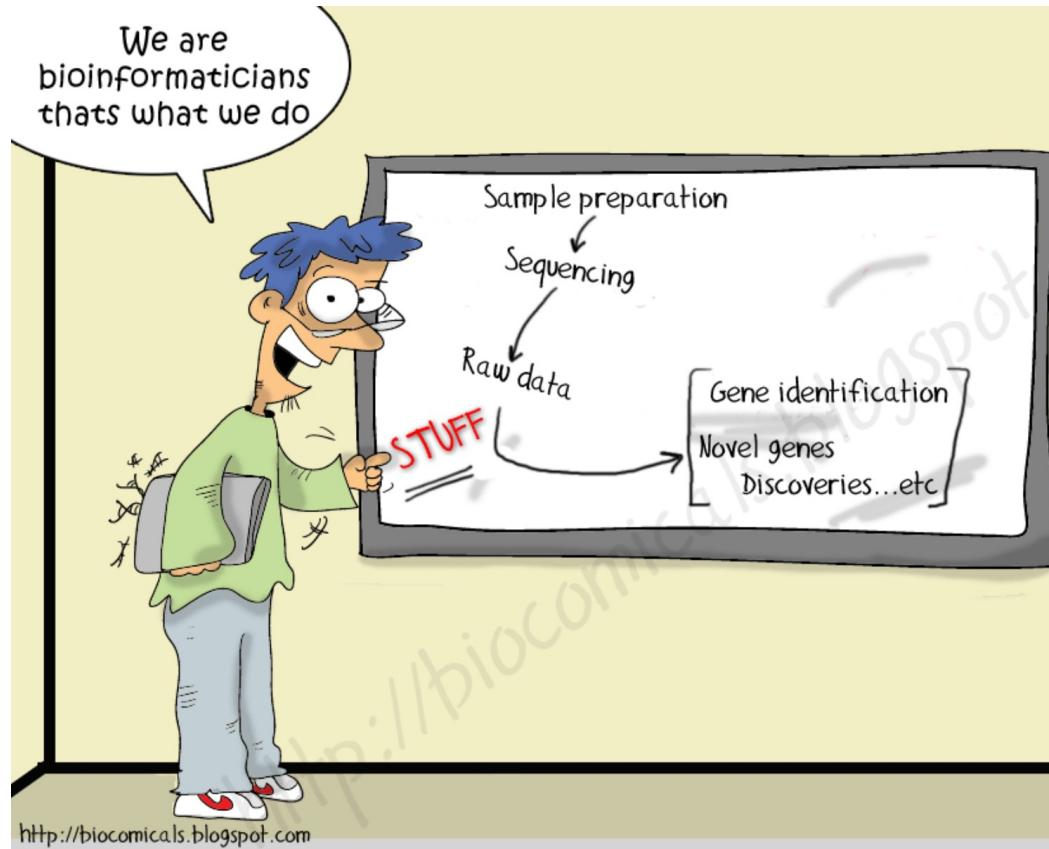
- Please provide answers and R code for an additional exercise (eg 1 Word with answers and figures and 1 script file, or 1 file generated from Rmarkdown)

<https://sib-swiss.github.io/enrichment-analysis-training/exercises/#extra-exercise-for-ects-credits>

Sign up for credit by adding your name to the google Doc file (email sent by course organizer)

- Send answers to [tania.wyss@sib.swiss](mailto:tania.wyss@sib.swiss) within 1 week

# Thank you for your attention!



Please fill in the **feedback** sent by the course organizer.  
We thank Isabelle Dupanloup and Linda Dib for  
providing course material.

