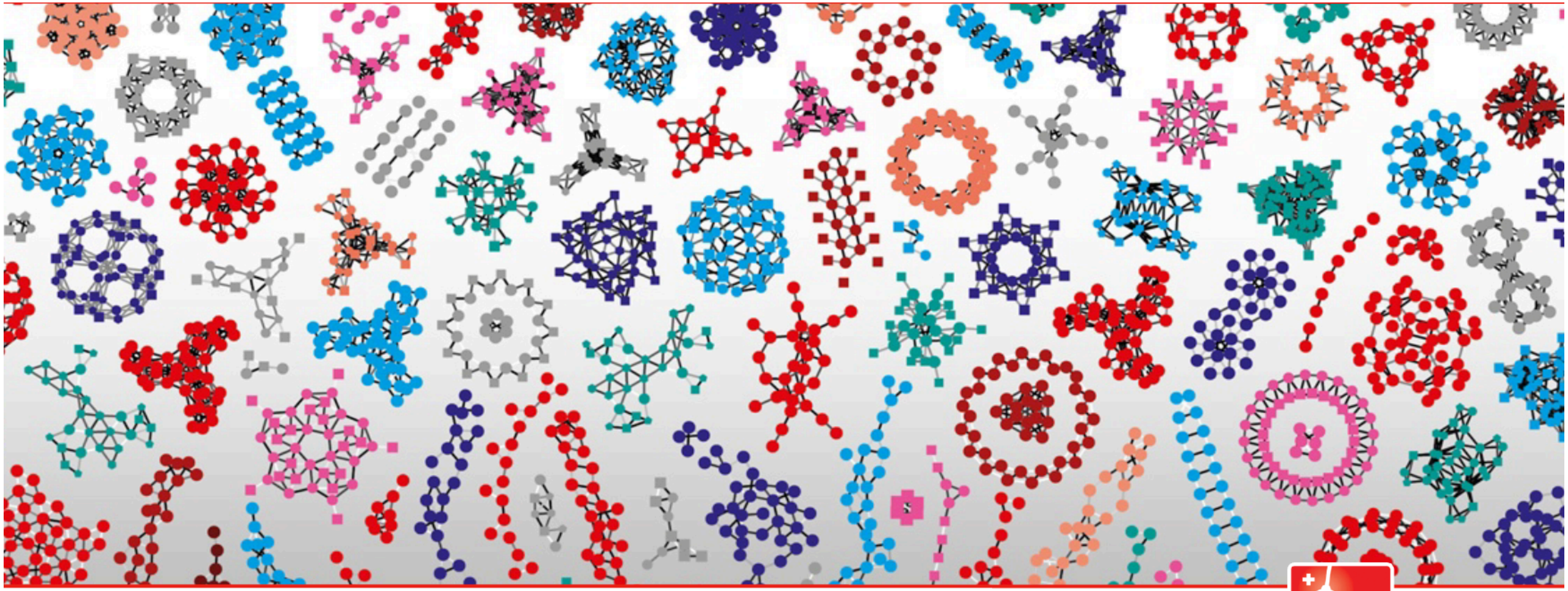


Enrichment analysis

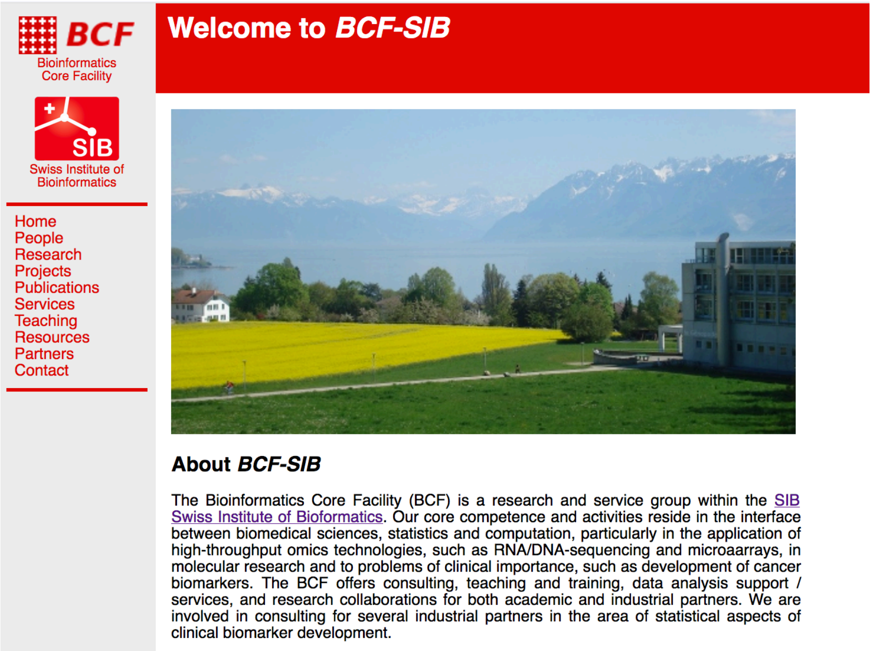
Swiss Institute of Bioinformatics

December 3rd 2021

isabelle.dupanloup@sib.swiss

tania.wyss@sib.swiss

The Bioinformatics Core Facility at SIB



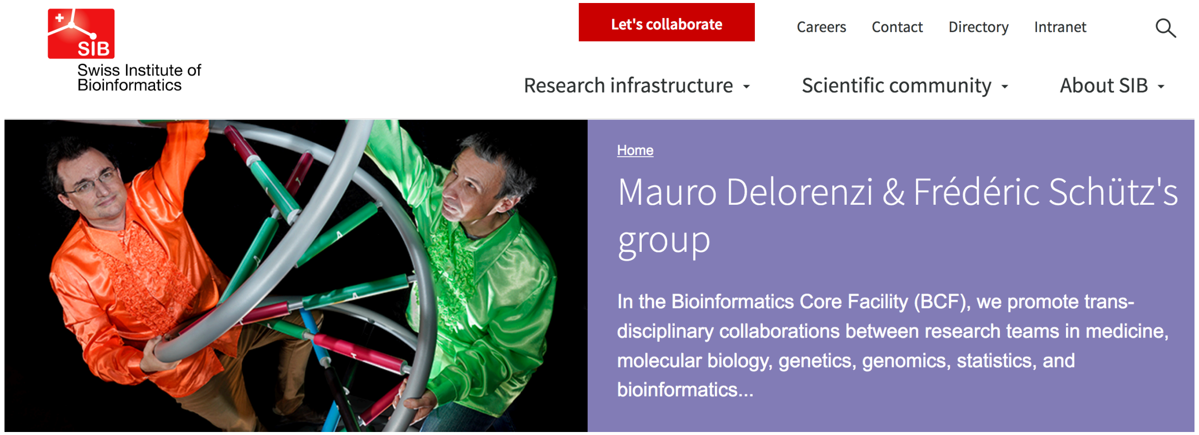
Welcome to *BCF-SIB*

About *BCF-SIB*

The Bioinformatics Core Facility (BCF) is a research and service group within the SIB Swiss Institute of Bioformatics. Our core competence and activities reside in the interface between biomedical sciences, statistics and computation, particularly in the application of high-throughput omics technologies, such as RNA/DNA-sequencing and microaarrays, in molecular research and to problems of clinical importance, such as development of cancer biomarkers. The BCF offers consulting, teaching and training, data analysis support / services, and research collaborations for both academic and industrial partners. We are involved in consulting for several industrial partners in the area of statistical aspects of clinical biomarker development.

https://bcf.sib.swiss

- Teaching and training
- Biostatistics support
- Collaboration



Mauro Delorenzi & Frédéric Schütz's group

In the Bioinformatics Core Facility (BCF), we promote trans-disciplinary collaborations between research teams in medicine, molecular biology, genetics, genomics, statistics, and bioinformatics...

https://www.sib.swiss/mauro-delorenzi-frederic-schutz-group

Schedule

- **9:00 - 10:30**
- Recall:
  - a. Differential expression
  - b. Statistical tests
- Exercise
- **10:30 - 10:45** break
- **10:45 - 12:30**
- Method of gene set enrichment analysis
- Exercise
- **12:30 - 13:30** lunch break
- **13:30 - 15:30**
- Ontologies and sources of gene sets
- Exercise
- **15:30 - 15:45** break
- **15:45 - 16:50**
- Visualization of enrichment results
- Exercise
- **16:50 - 17:00** Feedback and end of day

Credits: 0.25 ECTS

- Please provide results of exercises 2, 3 & 4 plus answers and R code for an additional exercise in a document (eg 1 Word with figures and 1 script file, or 1 file generated from Rmarkdown)
- Sign up for credit here:
- https://docs.google.com/document/d/1XAmufwECklEHibPnYcIQSYboADfowK1KG2RBc3RcBUo/edit#heading=h.5xrppxpatnym
- Send answers to tania.wyss@sib.swiss by December 10th 2021

First, tell us about yourself !

- What organism are you working on? What type of data are you analyzing?
- Write your name and some keywords about yourself and/or your research into the Google doc, to share about yourself.

Photo by National Cancer Institute, Unsplash

Photo by Scott Graham, Unsplash

Questions and Exercises

- Feel free to interrupt with questions by asking them directly or raising your hand.
- Can also use the chat or Q&A in google doc, Isabelle and I will answer

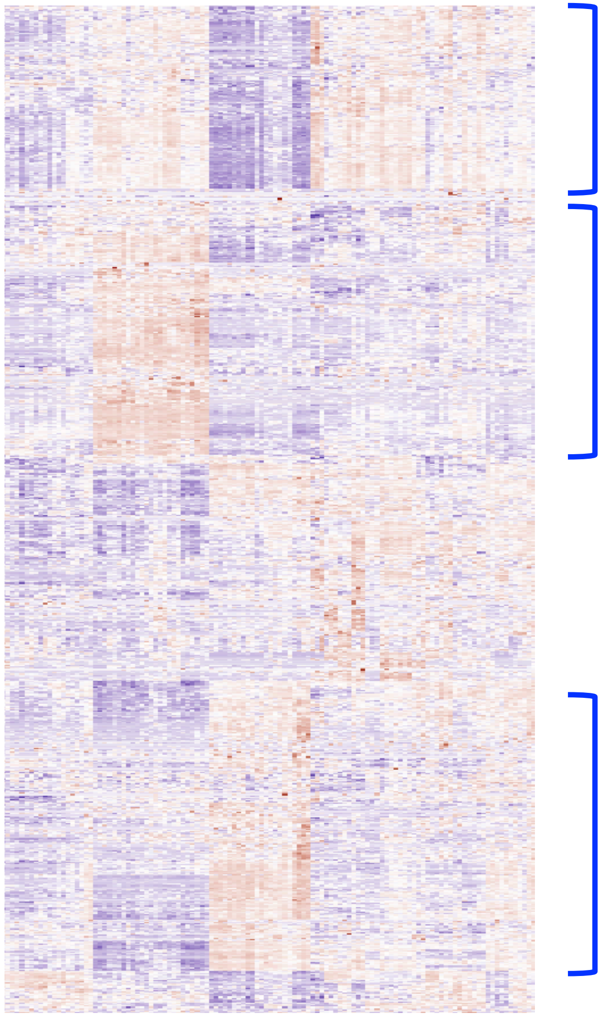
- Exercises in R:
  - We will try to debug as much as possible 👍
  - We are happy if you share your results!
  - Computational power on RStudio cloud is limited, might crash

Course material

- Moodle:
- https://edu.sib.swiss/course/view.php?id=550
- Login: enrich21
- Password: SIB-enrich21
- Feedback, survey at the end of the day.
- Additional links and answers to questions added to google doc:
- https://docs.google.com/document/d/1XAmufwECklEHibPnYcIQSYboADfowK1KG2RBc3RcBUo/edit#heading=h.5xrppxpatnym

Why do we perform enrichment analysis?

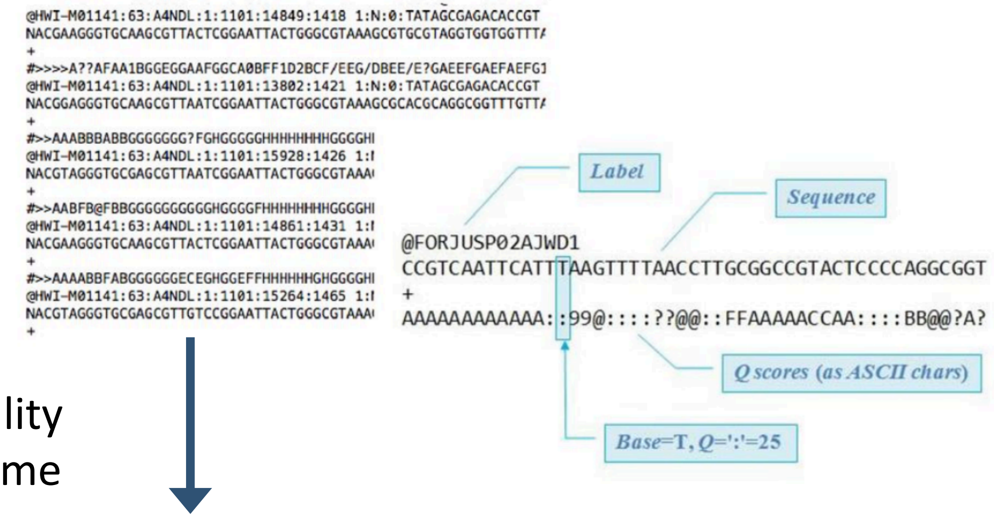
- Gene expression analysis yields hundreds to thousands of significant genes
  - We need to summarize the information provided by so many genes
  - Understand their biological relationships



IVY GAP: https://glioblastoma.alleninstitute.org/

Typical RNA sequencing analysis workflow

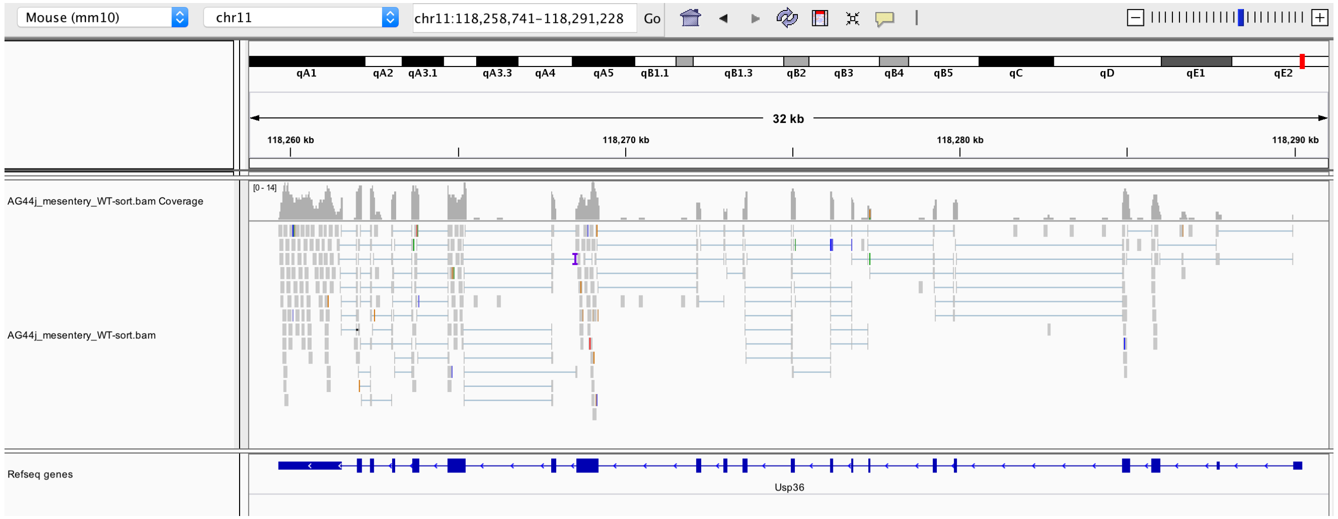
fastq file:



```
@HWI-M01141:63:A4NDL:1:1101:14849:1418 1:N:0:TATAGCGAGACACCGT
NACGAAGGGTGCAAGCGTTACTCGGAATTACTGGGCGTAAAGCGTGCGTAGGTGGTTTA
+
#>>>>A??AFAA1BGGEGGAAFGGCA0BFF1D2BCF/EEG/DBEE/E?GAEEFGAEFAEFG1
@HWI-M01141:63:A4NDL:1:1101:13802:1421 1:N:0:TATAGCGAGACACCGT
NACGGAGGGTGCAAGCGTTAATCGGAATTACTGGGCGTAAAGCGCACGCAGGCGGTTTGTTA
+
#>>AAABBBABBGGGGGGG?FGHGGGGGHHHHHHHHGGGGHI
@HWI-M01141:63:A4NDL:1:1101:15928:1426 1:I
NACGTAGGGTGCGAGCGTTAATCGGAATTACTGGGCGTAAA
+
#>>AABFB@FBBGGGGGGGGGGHGGGGFHHHHHHHHGGGGHI
@HWI-M01141:63:A4NDL:1:1101:14861:1431 1:I
NACGAAGGGTGCAAGCGTTACTCGGAATTACTGGGCGTAAA
+
#>>AAAABBFABGGGGGGECEGHGGEFFHHHHHHGHGGGGHI
@HWI-M01141:63:A4NDL:1:1101:15264:1465 1:I
NACGTAGGGTGCGAGCGTTGTCCGGAATTACTGGGCGTAAA
+
```

```
@FORJUSP02AJWD1
CCGTCAATTCATTTAAGTTTTAACCTTGCGGCCGTACTCCCCAGGCGGT
+
AAAAAAAAAAAA::99@::::??@@::FFAAAAACCAA::::BB@@?A?
```

Filter quality
Align to ref. genome



count reads
per gene

Downstream statistical analysis:
R: import counts table

Differential gene expression analysis

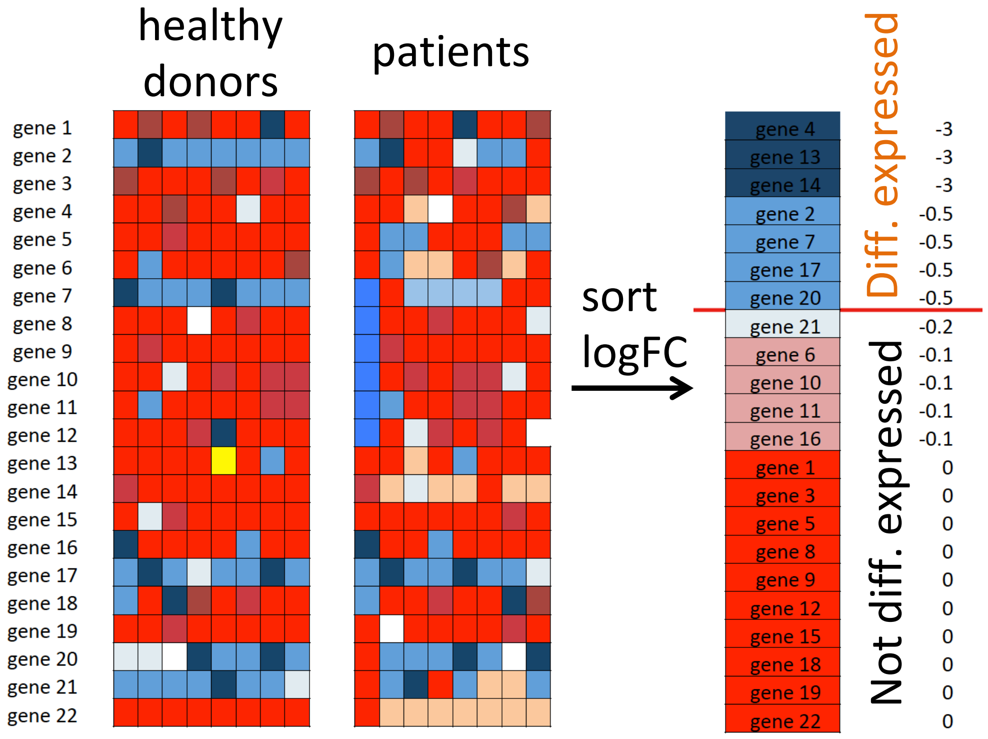
- Comparing 2 groups:

  For each gene i, is there a difference in expression between control and patients?

- Fold change in genomics:

  $\log_2$ of ratios = log fold change

$\log(\pi i1 / \pi i2) = \log(\pi i1) - \log(\pi i2)$

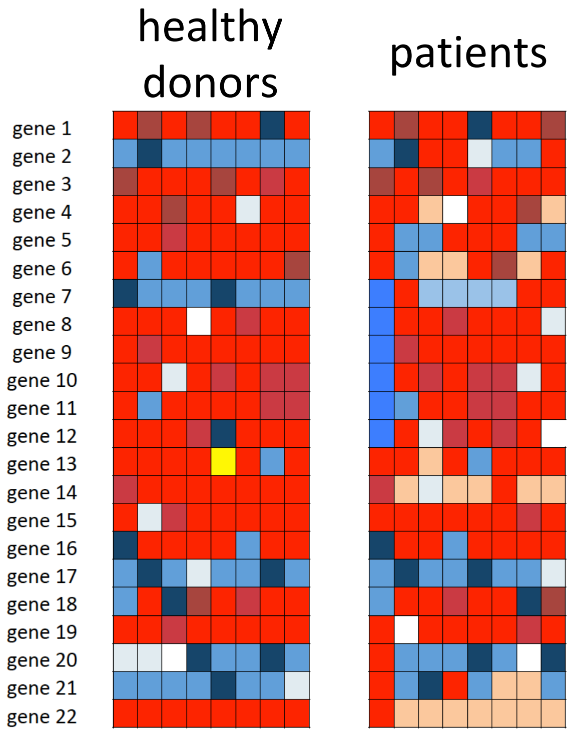


gene 4	-3
gene 13	-3
gene 14	-3
gene 2	-0.5
gene 7	-0.5
gene 17	-0.5
gene 20	-0.5
gene 21	-0.2
gene 6	-0.1
gene 10	-0.1
gene 11	-0.1
gene 16	-0.1
gene 1	0
gene 3	0
gene 5	0
gene 8	0
gene 9	0
gene 12	0
gene 15	0
gene 18	0
gene 19	0
gene 22	0

Differential gene expression analysis

- Comparing 2 groups:

  For each gene i, is there a significant difference in mean expression between control and patients?



- T-test:

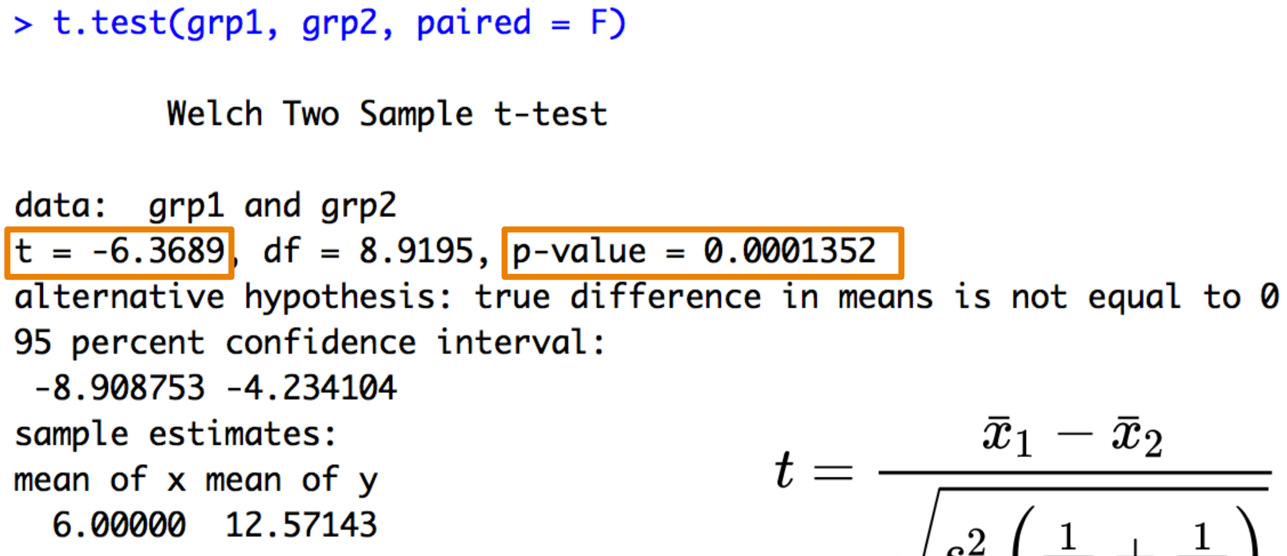
H0: Healthy donors and patients have similar gene I expression

$H0i : \pi i1 = \pi i2$

H1: Healthy donors and patients don’t have a similar gene i expression

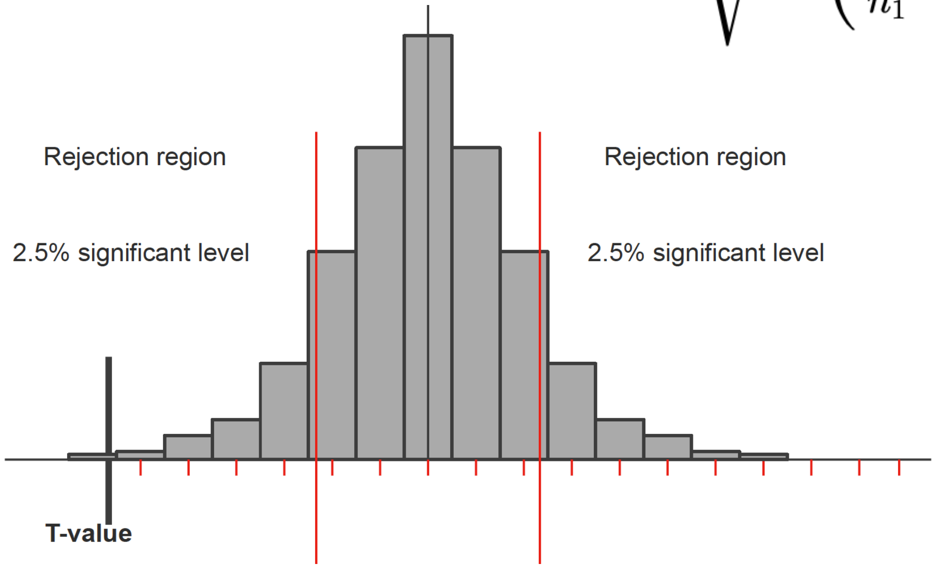
$H1i : \pi i1 \neq \pi i2$

T-test in R

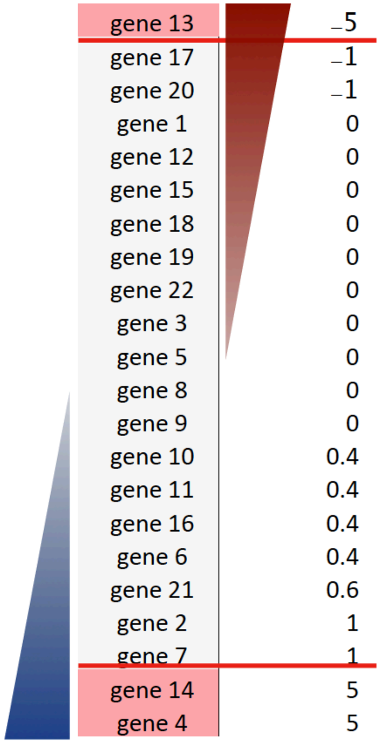


```
> t.test(grp1, grp2, paired = F)

        Welch Two Sample t-test

data:  grp1 and grp2
t = -6.3689, df = 8.9195, p-value = 0.0001352
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 -8.908753 -4.234104
sample estimates:
mean of x mean of y
  6.00000  12.57143
```

$$t = \frac{\bar{x}_1 - \bar{x}_2}{\sqrt{s^2\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}}$$



sort based on T-statistic

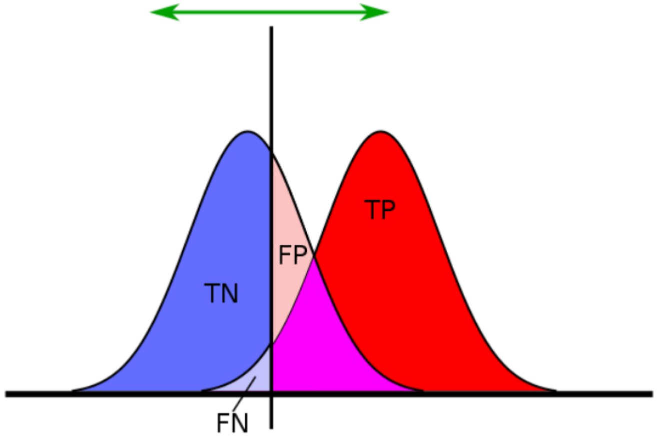


gene 13	–5
gene 17	–1
gene 20	–1
gene 1	0
gene 12	0
gene 15	0
gene 18	0
gene 19	0
gene 22	0
gene 3	0
gene 5	0
gene 8	0
gene 9	0
gene 10	0.4
gene 11	0.4
gene 16	0.4
gene 6	0.4
gene 21	0.6
gene 2	1
gene 7	1
gene 14	5
gene 4	5

What does $p < 0.05$ mean?

- It means that we suspect that the difference observed is not due to chance alone
- It means that if we repeat an experiment 20 times, we would reject the null hypothesis once because of random error

Decision Truth	H_0 not rejected (negative)	H_0 Rejected (positive)
H_0 is true (no signal in the data)	☺ specificity $1-\alpha$ True negative TN	X Type I error False Positive α
H_0 is false (there is something to find)	X Type II error False Negative β	☺ Power $1 - \beta$; sensitivity True Positive TP



P-value adjustment: what is it?

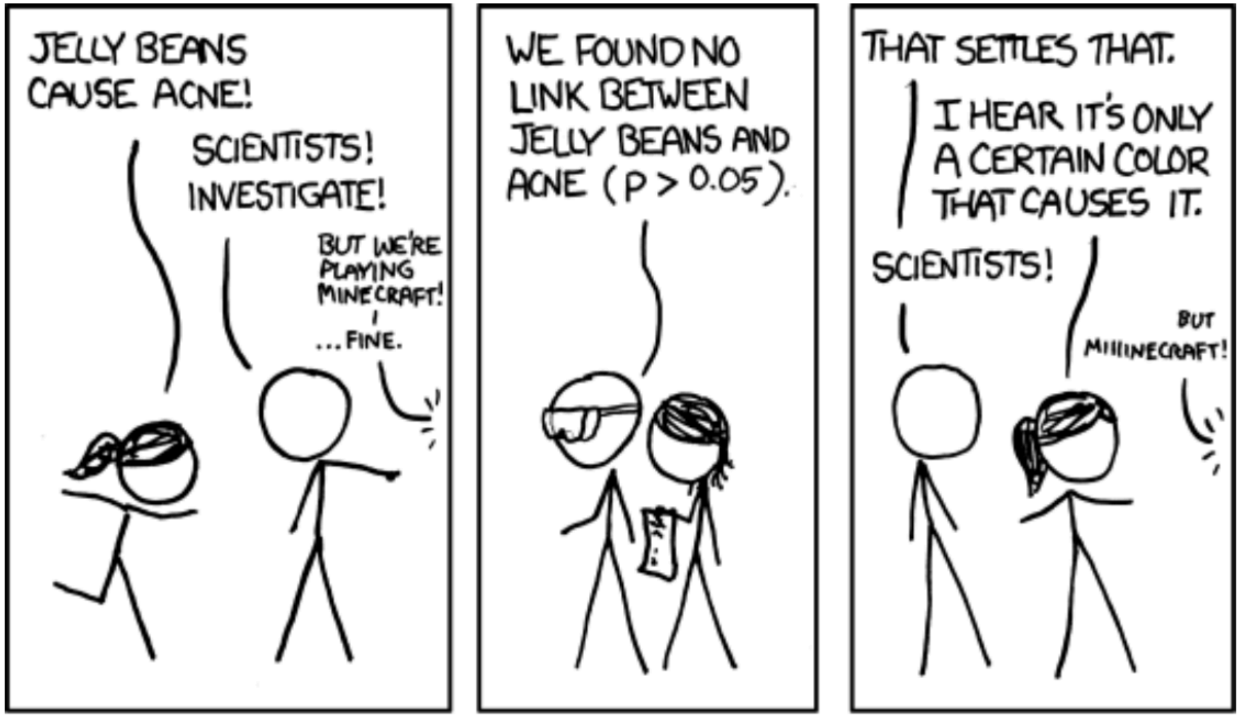
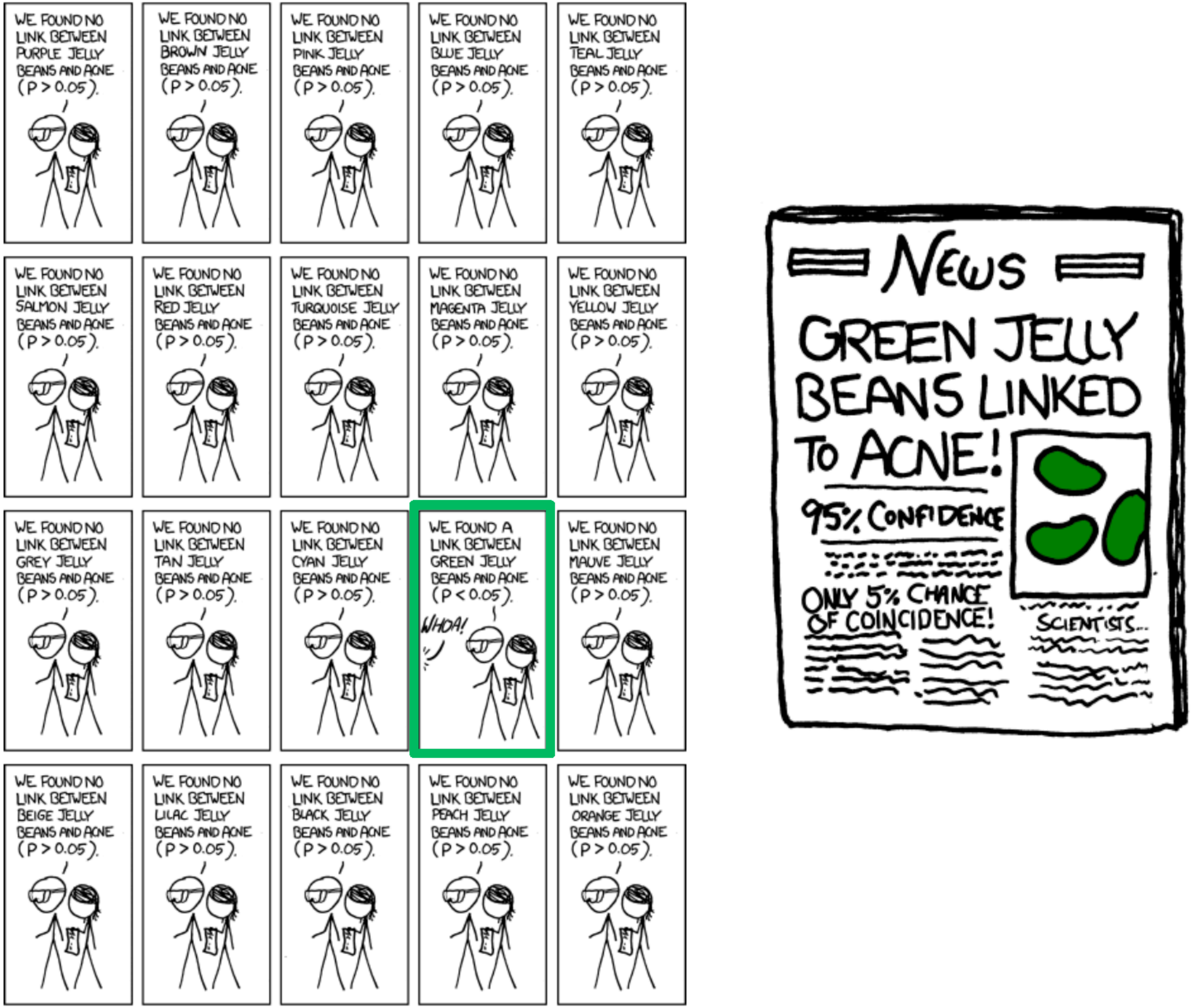

JELLY BEANS CAUSE ACNE!
SCIENTISTS! INVESTIGATE!
BUT WE'RE PLAYING MINECRAFT!
...FINE.
WE FOUND NO LINK BETWEEN JELLY BEANS AND ACNE (P > 0.05).
THAT SETTLES THAT.
I HEAR IT'S ONLY A CERTAIN COLOR THAT CAUSES IT.
SCIENTISTS!
BUT MIIINECRAFT!

Photo by Patrick Fore on Unsplash

Cartoon: https://xkcd.com/882/
Paper on p-value adjustment: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC6099145/


WE FOUND NO LINK BETWEEN PURPLE JELLY BEANS AND ACNE (P > 0.05).
WE FOUND NO LINK BETWEEN BROWN JELLY BEANS AND ACNE (P > 0.05).
WE FOUND NO LINK BETWEEN PINK JELLY BEANS AND ACNE (P > 0.05).
WE FOUND NO LINK BETWEEN BLUE JELLY BEANS AND ACNE (P > 0.05).
WE FOUND NO LINK BETWEEN TEAL JELLY BEANS AND ACNE (P > 0.05).
WE FOUND NO LINK BETWEEN SALMON JELLY BEANS AND ACNE (P > 0.05).
WE FOUND NO LINK BETWEEN RED JELLY BEANS AND ACNE (P > 0.05).
WE FOUND NO LINK BETWEEN TURQUOISE JELLY BEANS AND ACNE (P > 0.05).
WE FOUND NO LINK BETWEEN MAGENTA JELLY BEANS AND ACNE (P > 0.05).
WE FOUND NO LINK BETWEEN YELLOW JELLY BEANS AND ACNE (P > 0.05).
WE FOUND NO LINK BETWEEN GREY JELLY BEANS AND ACNE (P > 0.05).
WE FOUND NO LINK BETWEEN TAN JELLY BEANS AND ACNE (P > 0.05).
WE FOUND NO LINK BETWEEN CYAN JELLY BEANS AND ACNE (P > 0.05).
WE FOUND A LINK BETWEEN GREEN JELLY BEANS AND ACNE (P < 0.05).
WHOA!
WE FOUND NO LINK BETWEEN MAUVE JELLY BEANS AND ACNE (P > 0.05).
WE FOUND NO LINK BETWEEN BEIGE JELLY BEANS AND ACNE (P > 0.05).
WE FOUND NO LINK BETWEEN LILAC JELLY BEANS AND ACNE (P > 0.05).
WE FOUND NO LINK BETWEEN BLACK JELLY BEANS AND ACNE (P > 0.05).
WE FOUND NO LINK BETWEEN PEACH JELLY BEANS AND ACNE (P > 0.05).
WE FOUND NO LINK BETWEEN ORANGE JELLY BEANS AND ACNE (P > 0.05).
NEWS
GREEN JELLY BEANS LINKED TO ACNE!
95% CONFIDENCE
ONLY 5% CHANCE OF COINCIDENCE!
SCIENTISTS...

Methods of p-value adjustment

- Bonferroni: the alpha level is divided by the total number of tests
- if we run k=20 tests:

0.05/k = 0.05/20=0.0025

Good for small number of tests but too conservative for thousands of genes

- Benjamini-Hochberg procedure (BH to control FDR)
- Rank the p-values from smallest to largest, adjust less and less as the p-values get larger:

$\text{p-value}_1 * n/1$

$\text{p-value}_2 * n/2$

$\text{p-value}_k * n/k$

n= number of genes

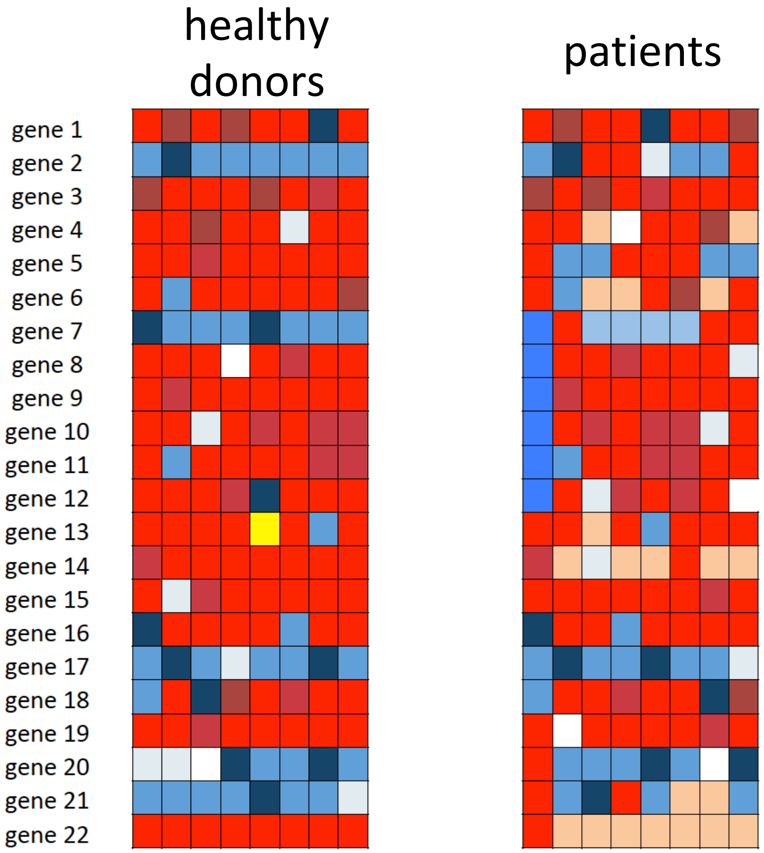
k= rank number

Differential gene expression analysis using R

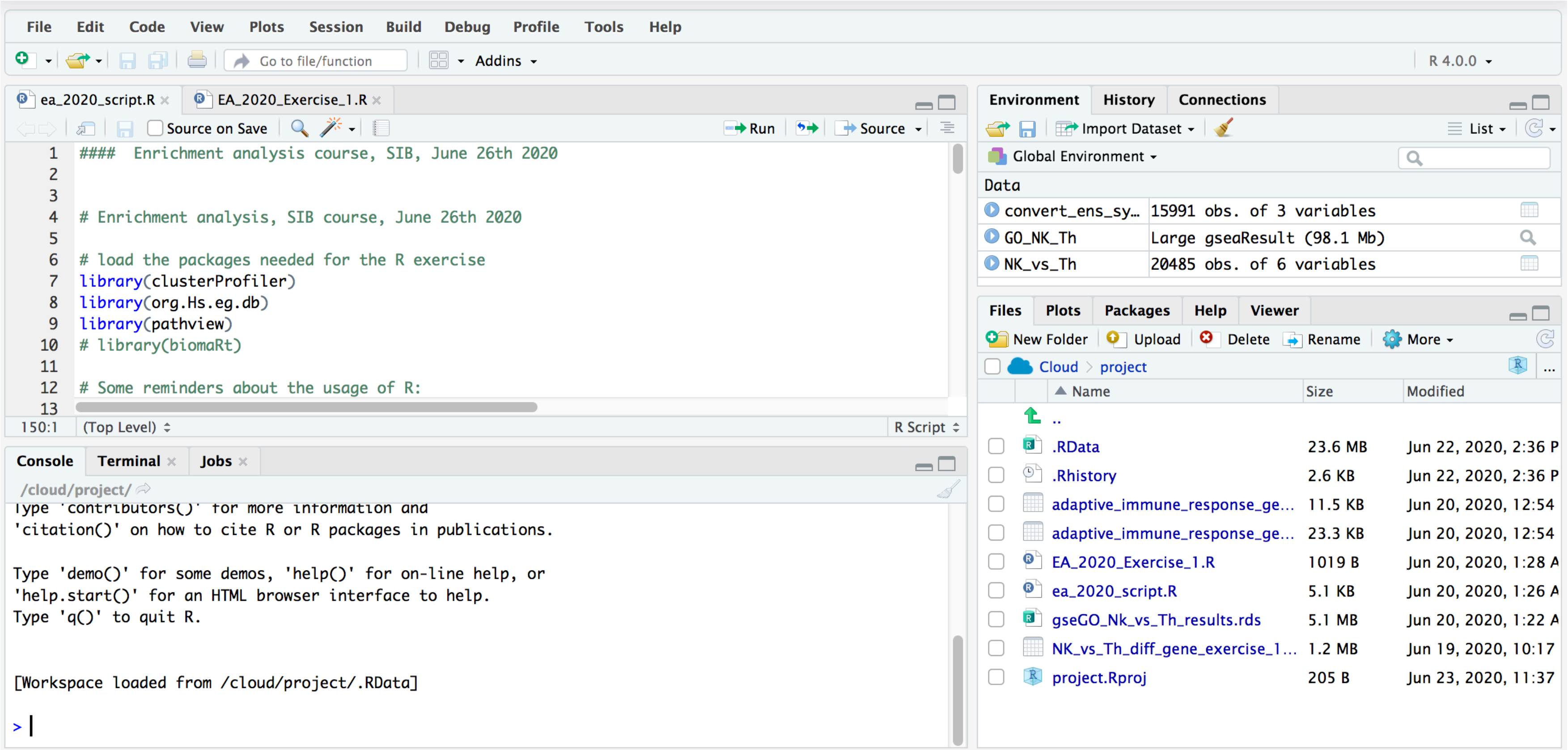
- Bioconductor

https://bioconductor.org/

- Several packages :
  - limma: t-test
  - DESeq2: Wald test
  - edgeR: exact test



RStudio tour



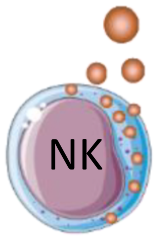
```
####  Enrichment analysis course, SIB, June 26th 2020


# Enrichment analysis, SIB course, June 26th 2020

# load the packages needed for the R exercise
library(clusterProfiler)
library(org.Hs.eg.db)
library(pathview)
# library(biomaRt)

# Some reminders about the usage of R:
```

```
Type 'contributors()' for more information and
'citation()' on how to cite R or R packages in publications.

Type 'demo()' for some demos, 'help()' for on-line help, or
'help.start()' for an HTML browser interface to help.
Type 'q()' to quit R.


[Workspace loaded from /cloud/project/.RData]

>
```

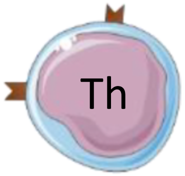
Data

convert_ens_sy…	15991 obs. of 3 variables
GO_NK_Th	Large gseaResult (98.1 Mb)
NK_vs_Th	20485 obs. of 6 variables

Name	Size	Modified
..		
.RData	23.6 MB	Jun 22, 2020, 2:36 P
.Rhistory	2.6 KB	Jun 22, 2020, 2:36 P
adaptive_immune_response_ge…	11.5 KB	Jun 20, 2020, 12:54
adaptive_immune_response_ge…	23.3 KB	Jun 20, 2020, 12:54
EA_2020_Exercise_1.R	1019 B	Jun 20, 2020, 1:28 A
ea_2020_script.R	5.1 KB	Jun 20, 2020, 1:26 A
gseGO_Nk_vs_Th_results.rds	5.1 MB	Jun 20, 2020, 1:22 A
NK_vs_Th_diff_gene_exercise_1…	1.2 MB	Jun 19, 2020, 10:17
project.Rproj	205 B	Jun 23, 2020, 11:37



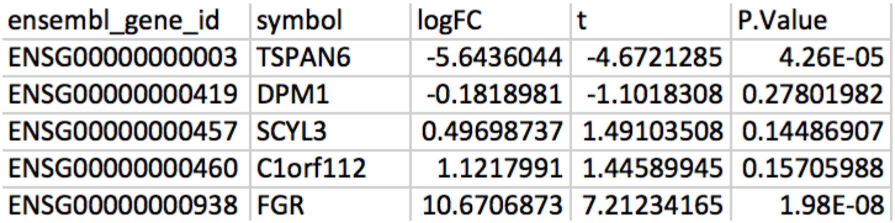
Recap and exercise 1



https://www.mdpi.com/1420-3049/24/24/4530/htm

- Differential gene expression analysis typically involves calculating fold change, running a statistical test to compare gene expression between 2 conditions, and adjusting the p-value.

- 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):
  - Is the gene CPS1 significantly differentially expressed between NK and Th cells?
  - How many genes are up-regulated and down-regulated in NK after BH adjustment?
  - Is the gene CPS1 still significant after BH adjustment?

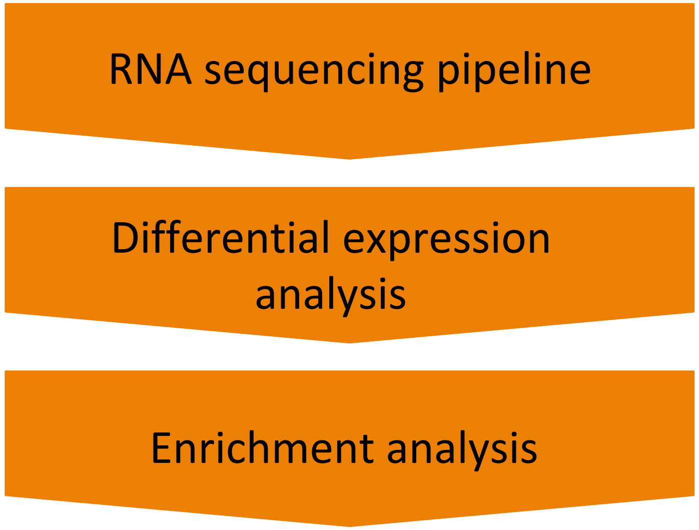


ensembl_gene_id	symbol	logFC	t	P.Value
ENSG00000000003	TSPAN6	-5.6436044	-4.6721285	4.26E-05
ENSG00000000419	DPM1	-0.1818981	-1.1018308	0.27801982
ENSG00000000457	SCYL3	0.49698737	1.49103508	0.14486907
ENSG00000000460	C1orf112	1.1217991	1.44589945	0.15705988
ENSG00000000938	FGR	10.6706873	7.21234165	1.98E-08

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

RNA sequencing data from:
https://jlb.onlinelibrary.wiley.com/doi/full/10.1002/JLB.5MA0120-209R?af=R
https://ashpublications.org/bloodadvances/article/3/22/3674/428873/CD56-as-a-marker-of-an-ILC1-like-population-with

Once we have identified DE genes, what do we do?



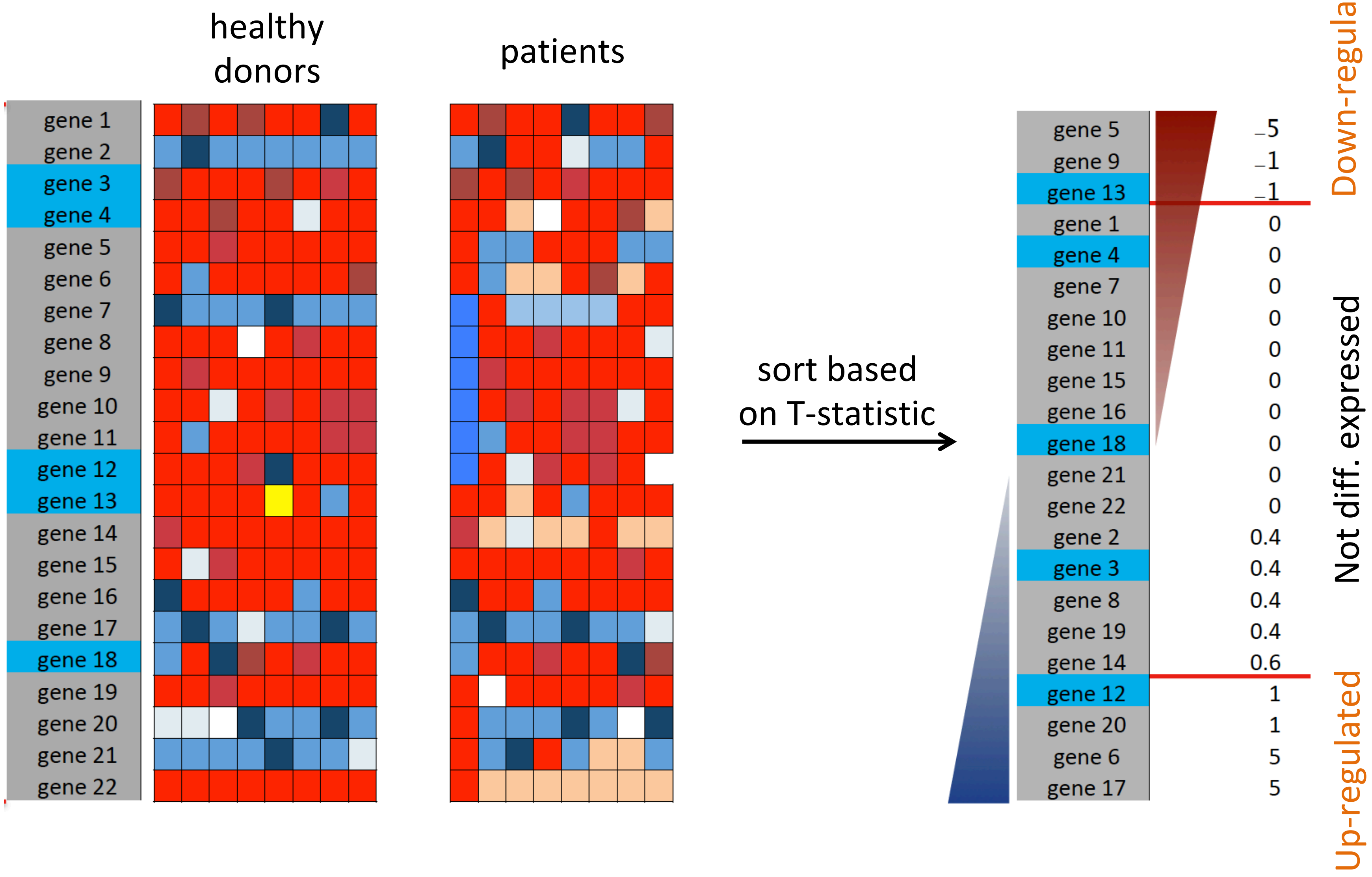
Several methods available, *e.g.*:

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

Goal: to gain biologically-meaningful insights from long gene lists

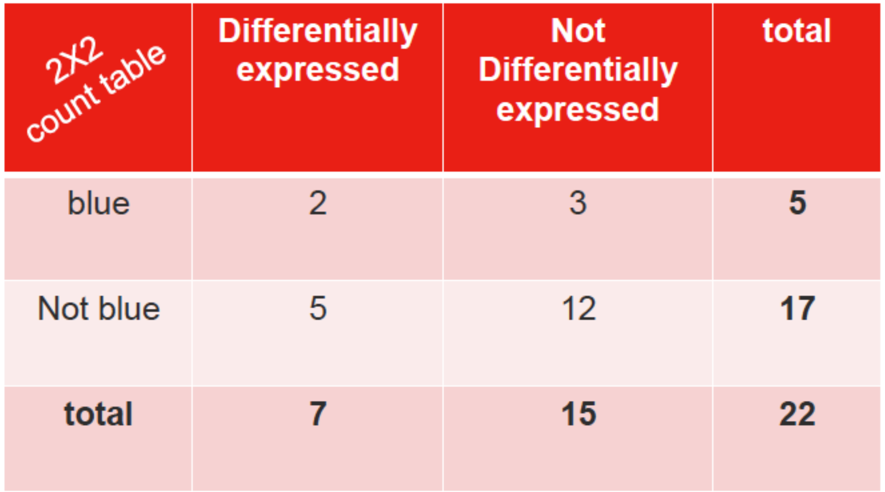
- test if differentially expressed genes are enriched in genes associated with a particular function
- approaches: test a small number of gene sets, or a large collection of gene sets

Are the genes belonging to the blue set differentially expressed?



gene 5	-5
gene 9	-1
gene 13	-1
gene 1	0
gene 4	0
gene 7	0
gene 10	0
gene 11	0
gene 15	0
gene 16	0
gene 18	0
gene 21	0
gene 22	0
gene 2	0.4
gene 3	0.4
gene 8	0.4
gene 19	0.4
gene 14	0.6
gene 12	1
gene 20	1
gene 6	5
gene 17	5

Fisher’s exact test



2X2 count table	Differentially expressed	Not Differentially expressed	total
blue	2	3	**5**
Not blue	5	12	**17**
total	**7**	**15**	**22**

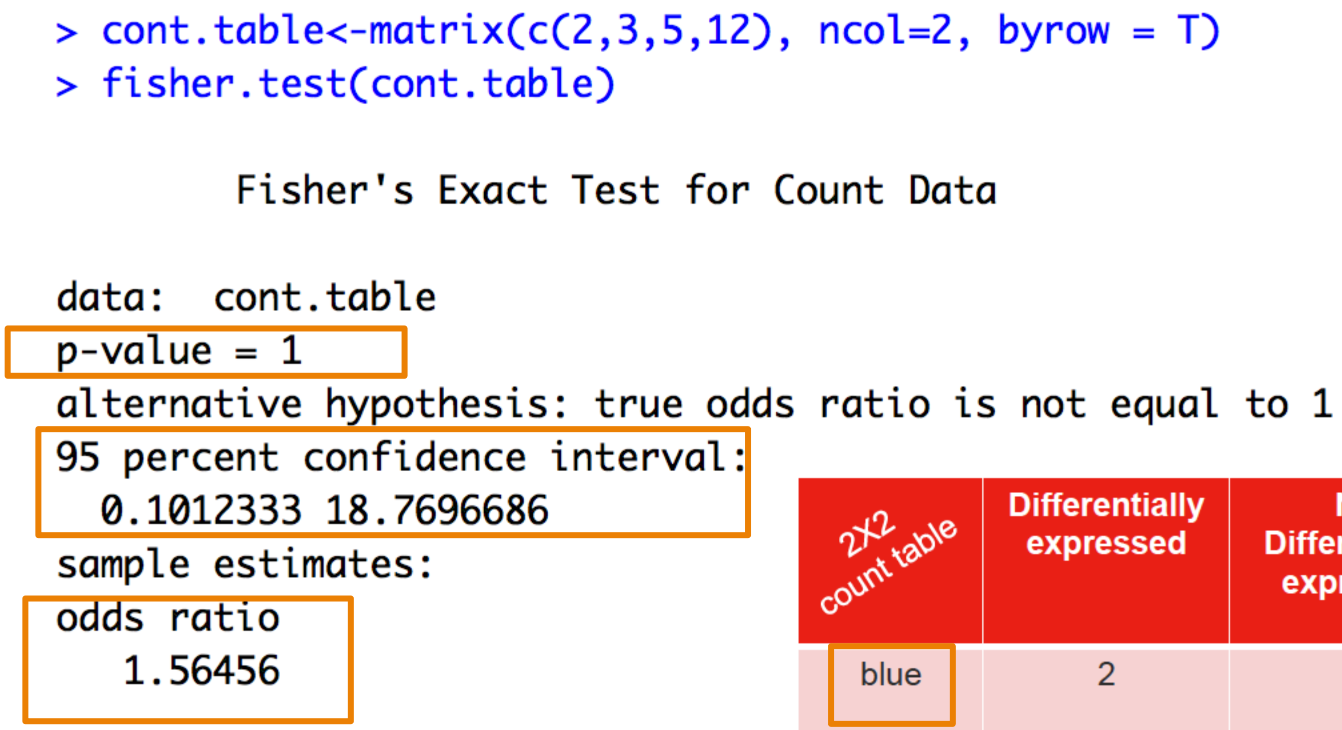
contingency table

H_0: The proportion of blue genes differentially expressed is the same as the proportion of blue genes in non-differentially expressed genes

H_1: The proportion of blue genes differentially expressed is not the same as the proportion of blue genes in non-differentially expressed genes

http://mengnote.blogspot.com/2012/12/calculate-correct-hypergeometric-p.html

Fisher’s exact test in R



```
> cont.table<-matrix(c(2,3,5,12), 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:
  0.1012333 18.7696686
sample estimates:
odds ratio
   1.56456
```

2X2 count table	Differentially expressed	Not Differentially expressed	total
blue	2	3	**5**
Not blue	5	12	**17**
total	**7**	**15**	**22**

2/7 =
0.29

3/15 =
0.20

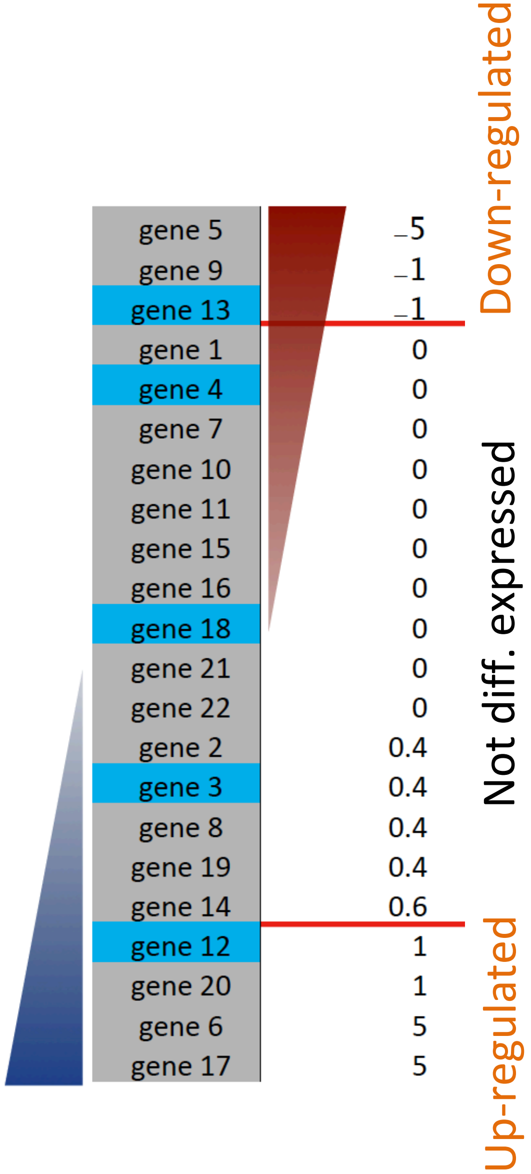
Which gene sets are differentially expressed?

🟩	gene 1	0
🟥	gene 2	0.4
🟦	gene 3	0.4
🟦	gene 4	0
🟥	gene 5	−5
🟪	gene 6	5
🟪	gene 7	0
🟥	gene 8	0.4
🟥🟪	gene 9	−1
🟩	gene 10	0
🟩	gene 11	0
🟦	gene 12	1
🟦	gene 13	−1
🟩	gene 14	0.6
🟩	gene 15	0
🟪	gene 16	0
🟪	gene 17	5
🟦	gene 18	0
🟪	gene 19	0.4
🟪	gene 20	1
🟥	gene 21	0
🟥	gene 22	0

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

⇒Multiple tests need p-value adjustment.

⇒But Fisher test is threshold-based.



	gene 5	−5
	gene 9	−1
🟦	gene 13	−1
	gene 1	0
🟦	gene 4	0
	gene 7	0
	gene 10	0
	gene 11	0
	gene 15	0
	gene 16	0
🟦	gene 18	0
	gene 21	0
	gene 22	0
	gene 2	0.4
🟦	gene 3	0.4
	gene 8	0.4
	gene 19	0.4
	gene 14	0.6
🟦	gene 12	1
	gene 20	1
	gene 6	5
	gene 17	5

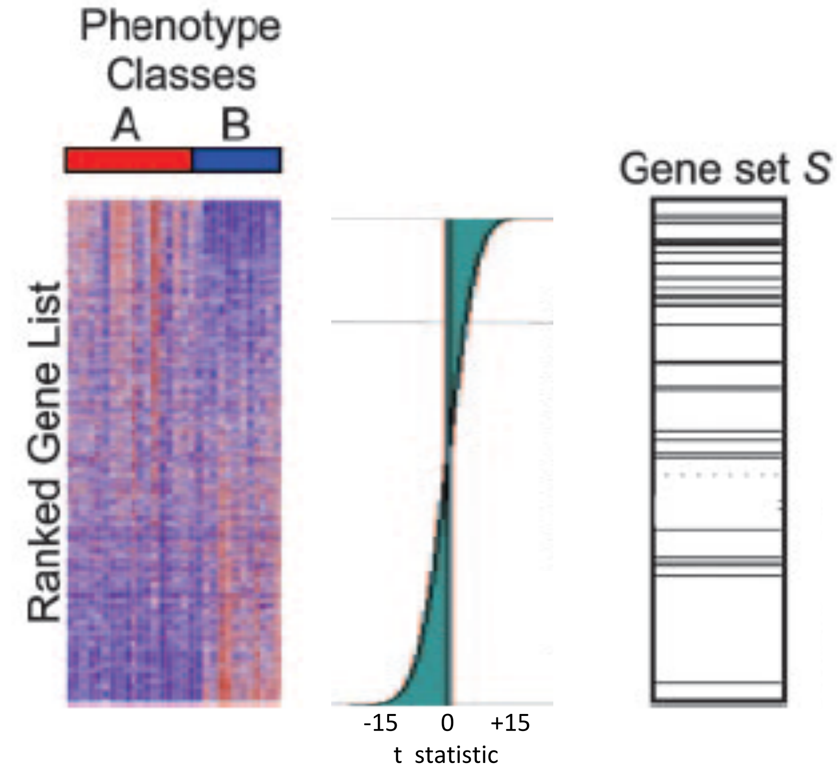
Gene set enrichment analysis (GSEA)

- 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)
- Threshold-free: the whole list of genes detected in the RNA sequencing experiment is used.
- 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.

Subramanian et al PNAS 2005. Gene set enrichment analysis: A knowledge-based approach for interpreting genome-wide expression profiles

Method of GSEA

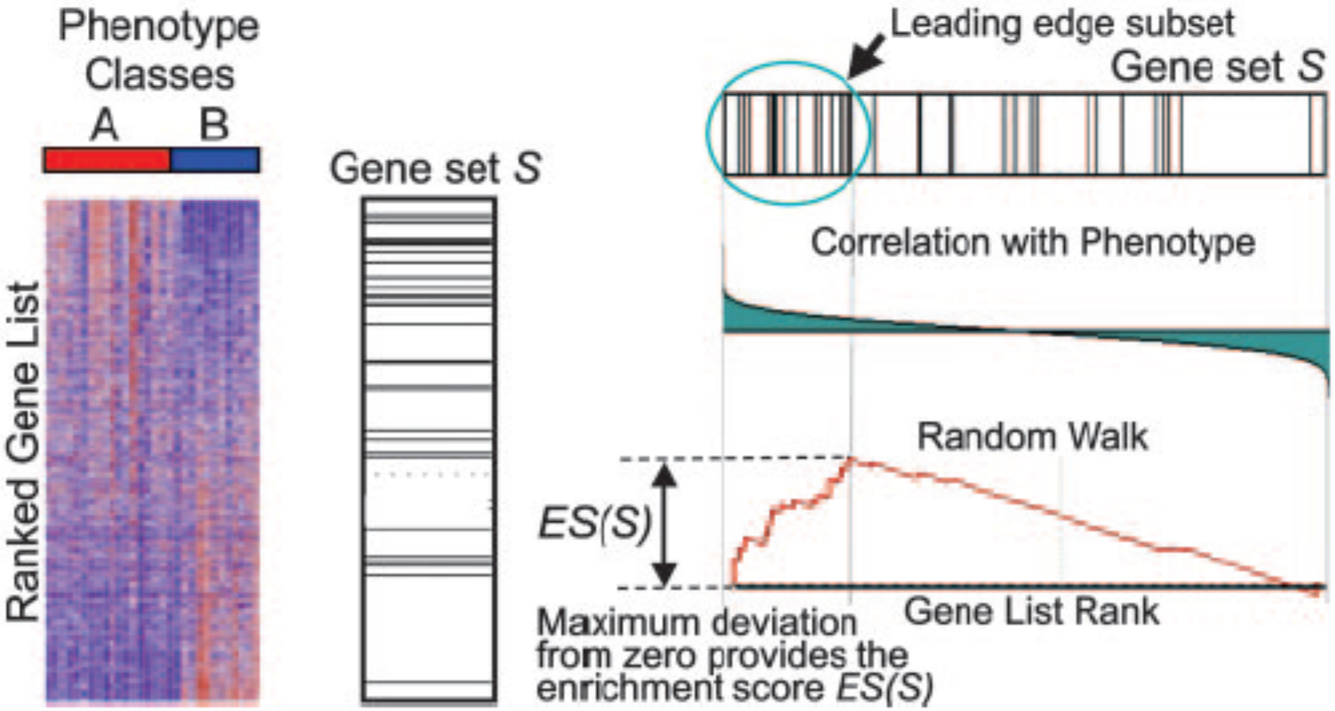
Goal: determine whether the members of a gene set S are randomly distributed throughout a ranked gene list or if they are located at the top or bottom of the ranked gene lists



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

Subramanian et al PNAS 2005. Gene set enrichment analysis: A knowledge-based approach for interpreting genome-wide expression profiles

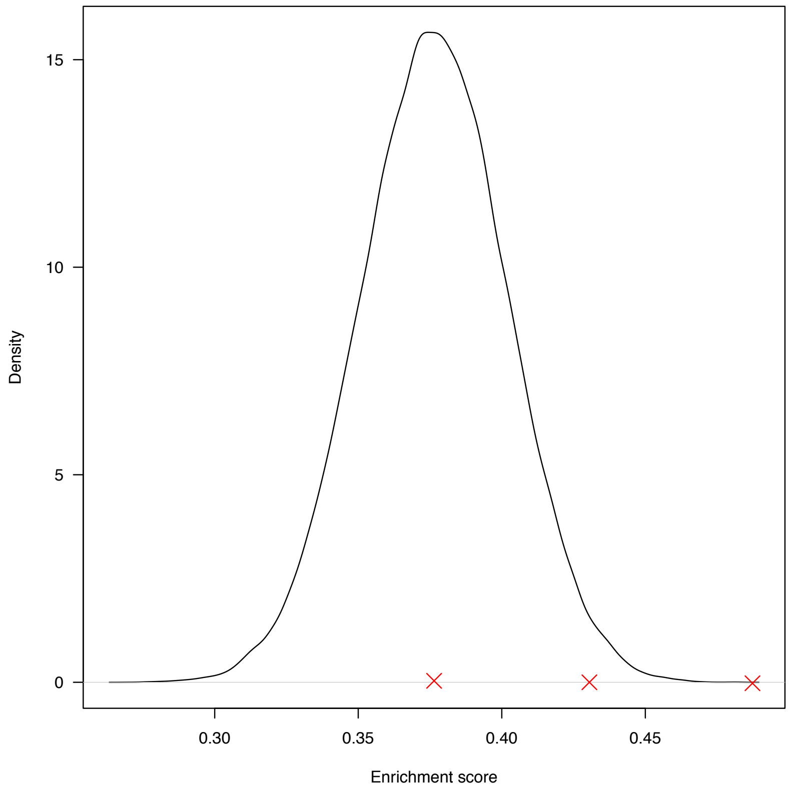
Method of GSEA



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



NES: 1
p: 0.5

NES: 1.16
p: 0.05

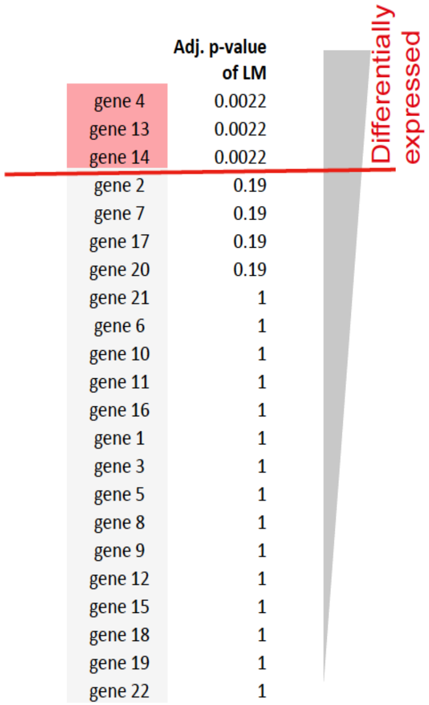
NES: 1.32
p: 0.001

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

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

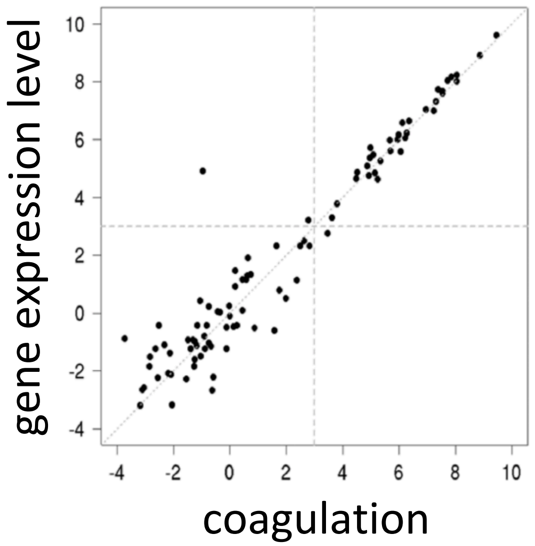
Apply GSEA to any type of data or score

- Use t-statistic from paired t-test
- Use F statistic of one way or two way ANOVA
- Use p-value of linear model



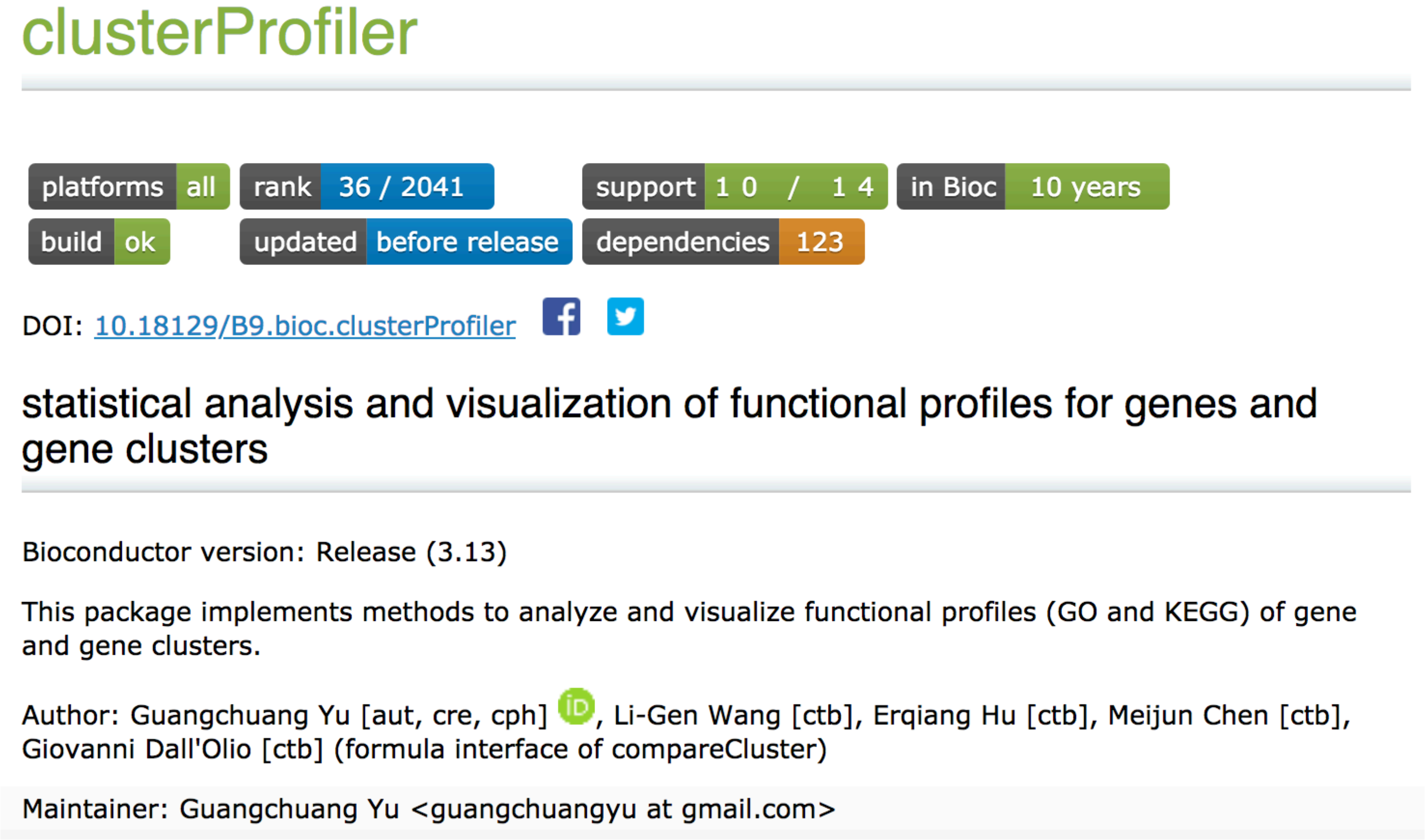
	Adj. p-value of LM
gene 4	0.0022
gene 13	0.0022
gene 14	0.0022
gene 2	0.19
gene 7	0.19
gene 17	0.19
gene 20	0.19
gene 21	1
gene 6	1
gene 10	1
gene 11	1
gene 16	1
gene 1	1
gene 3	1
gene 5	1
gene 8	1
gene 9	1
gene 12	1
gene 15	1
gene 18	1
gene 19	1
gene 22	1

Differentially expressed



GSEA for linear model implemented in romer() function of the limma package

GSEA using R: one possibility among many



clusterProfiler

platforms all | rank 36 / 2041 | support 10 / 14 | in Bioc 10 years

build ok | updated before release | dependencies 123

DOI: 10.18129/B9.bioc.clusterProfiler

statistical analysis and visualization of functional profiles for genes and gene clusters

Bioconductor version: Release (3.13)

This package implements methods to analyze and visualize functional profiles (GO and KEGG) of gene and gene clusters.

Author: Guangchuang Yu [aut, cre, cph], Li-Gen Wang [ctb], Erqiang Hu [ctb], Meijun Chen [ctb], Giovanni Dall'Olio [ctb] (formula interface of compareCluster)

Maintainer: Guangchuang Yu <guangchuangyu at gmail.com>

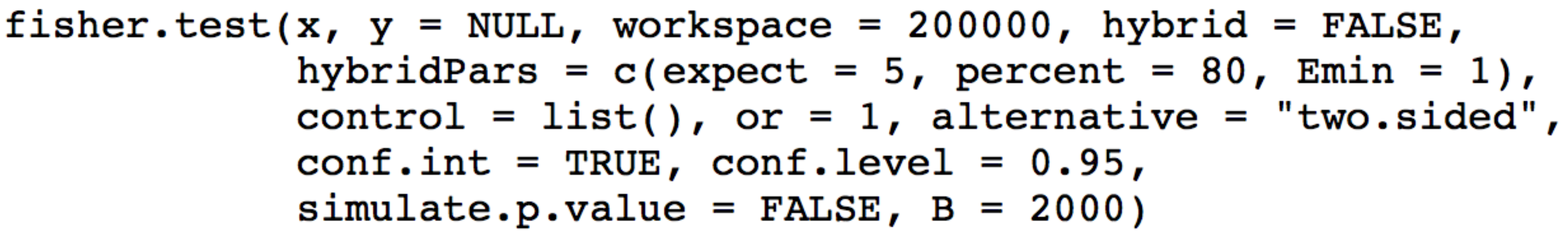
Built-in gene sets for human, mouse, yeast, etc

Built-in GO and KEGG (see later)

- ***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)
- Full vignette: http://yulab-smu.top/clusterProfiler-book/

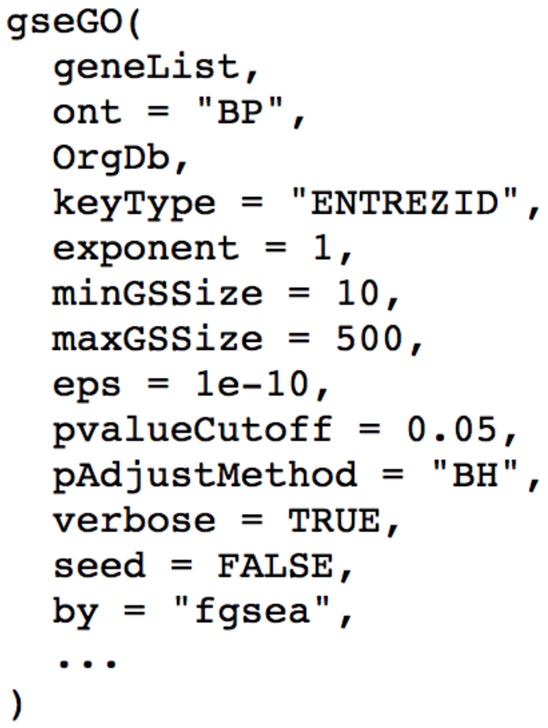
Functions for Fisher test and for enrichment analysis with clusterProfiler

Fisher exact test (package stats)



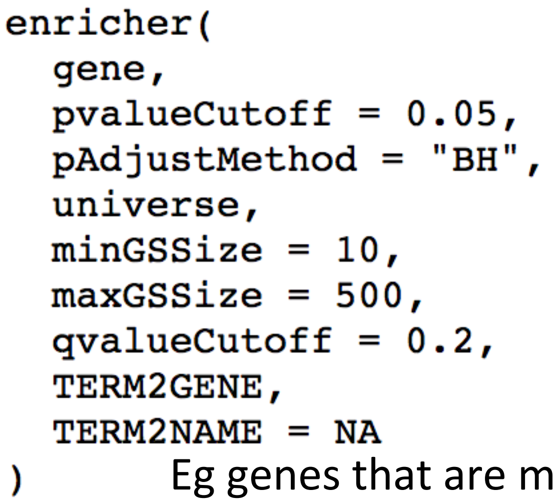
```
fisher.test(x, y = NULL, workspace = 200000, hybrid = FALSE,
            hybridPars = c(expect = 5, percent = 80, Emin = 1),
            control = list(), or = 1, alternative = "two.sided",
            conf.int = TRUE, conf.level = 0.95,
            simulate.p.value = FALSE, B = 2000)
```

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



```
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",
  ...
)
```

enricher(): similar to Fisher’s exact test, for user defined gene list and gene set annotations (package clusterProfiler)



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

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

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. Remember to use p-value adjustment if multiple Fisher tests are used.

- Exercise 2: use functions of clusterProfiler and data provided in Ex. 1

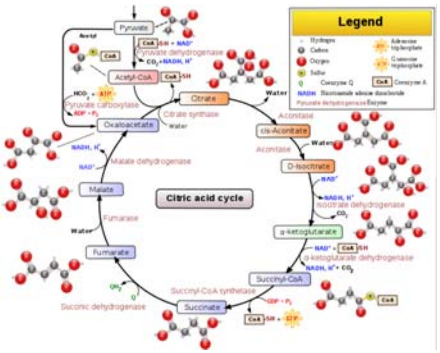
  - Is the adaptive immune response gene set significantly enriched in genes up-regulated in NK vs Th?

  - How many GO gene sets are significant after GSEA (use minGSSize=30) ?

  - Is the adaptive immune response gene set significant? Up-reg. or down-reg.?

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

What is a gene set?



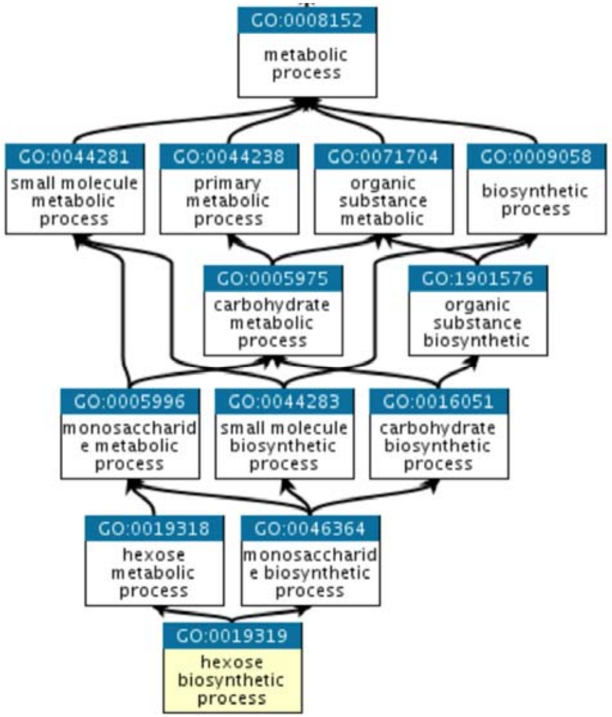
https://en.wikipedia.org/wiki/Citric_acid_cycle

- 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)

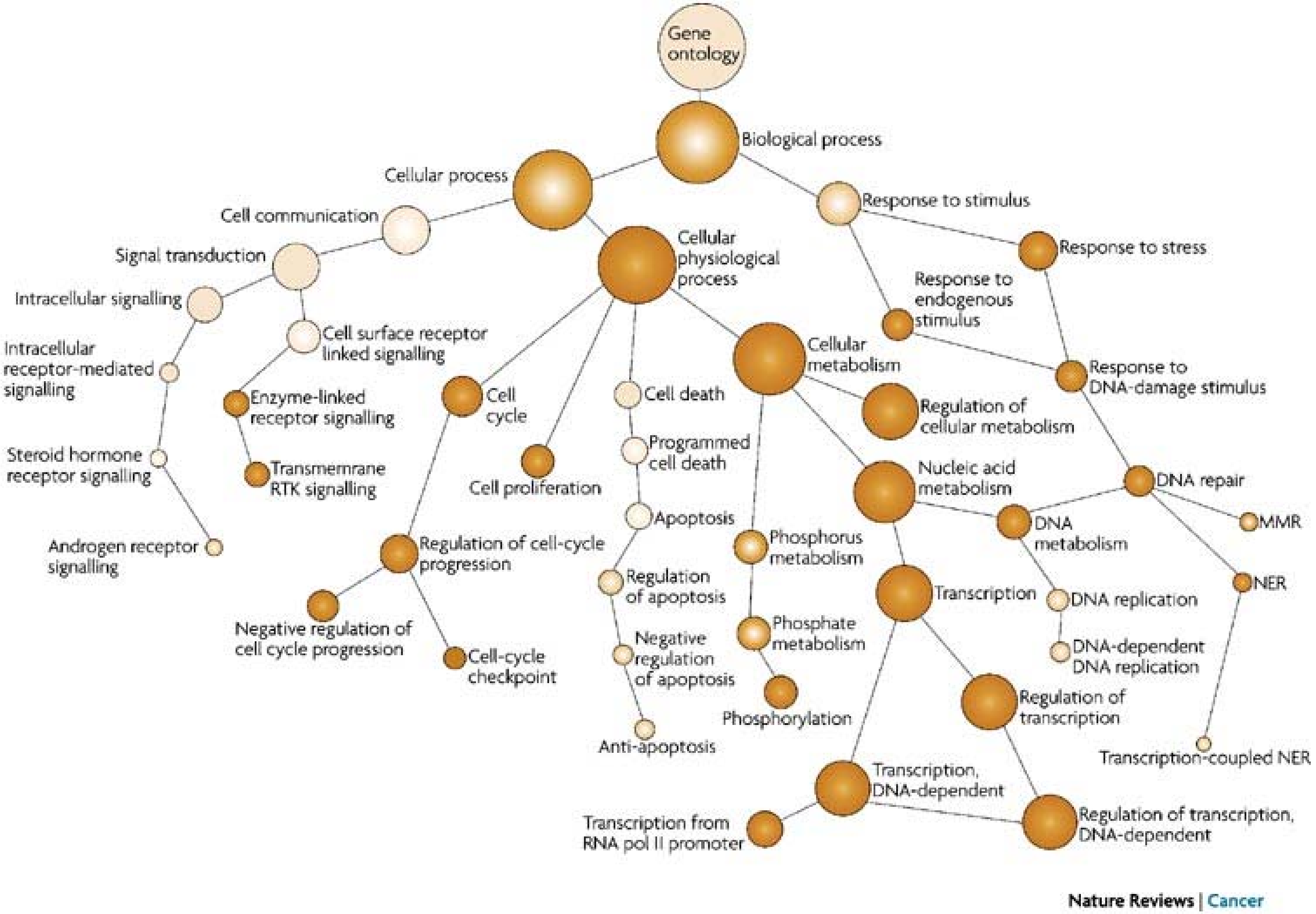

GO:0008152 metabolic process
GO:0044281 small molecule metabolic process
GO:0044238 primary metabolic process
GO:0071704 organic substance metabolic
GO:0009058 biosynthetic process
GO:0005975 carbohydrate metabolic process
GO:1901576 organic substance biosynthetic
GO:0005996 monosaccharide metabolic process
GO:0044283 small molecule biosynthetic process
GO:0016051 carbohydrate biosynthetic process
GO:0019318 hexose metabolic process
GO:0046364 monosaccharide biosynthetic process
GO:0019319 hexose biosynthetic process

Gene ontology

GO ontologies: GO terms organized in 3 independent controlled vocabularies

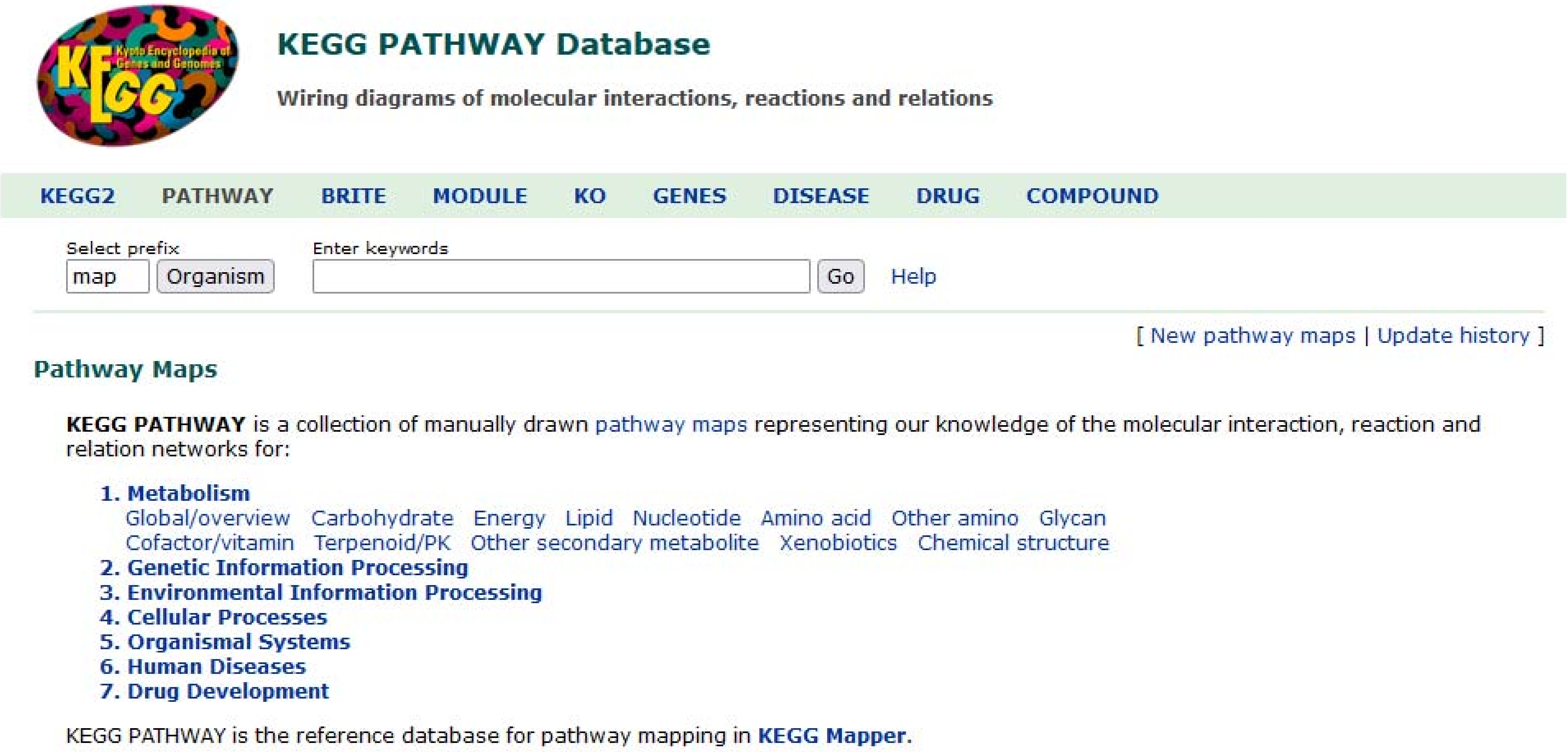
- **Molecular function**: represents the biochemical activity of the gene product, such activities could include "ligand", "GTPase", and "transporter".
- **Cellular component**: refers to the location in the cell of the gene product. Cellular components could include "nucleus", "lysosome", and "plasma 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


Gene ontology
Biological process
Cellular process
Response to stimulus
Cell communication
Response to stress
Signal transduction
Cellular physiological process
Response to endogenous stimulus
Intracellular signalling
Cell surface receptor linked signalling
Cellular metabolism
Response to DNA-damage stimulus
Intracellular receptor-mediated signalling
Enzyme-linked receptor signalling
Cell cycle
Cell death
Regulation of cellular metabolism
Programmed cell death
Steroid hormone receptor signalling
Transmemrane RTK signalling
Cell proliferation
Nucleic acid metabolism
DNA repair
Apoptosis
DNA metabolism
MMR
Androgen receptor signalling
Regulation of cell-cycle progression
Phosphorus metabolism
Regulation of apoptosis
Transcription
DNA replication
NER
Negative regulation of cell cycle progression
Phosphate metabolism
Negative regulation of apoptosis
DNA-dependent DNA replication
Cell-cycle checkpoint
Regulation of transcription
Phosphorylation
Anti-apoptosis
Transcription-coupled NER
Transcription, DNA-dependent
Regulation of transcription, DNA-dependent
Transcription from RNA pol II promoter
Nature Reviews | Cancer

KEGG

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



KEGG PATHWAY Database

Wiring diagrams of molecular interactions, reactions and relations

Pathway Maps

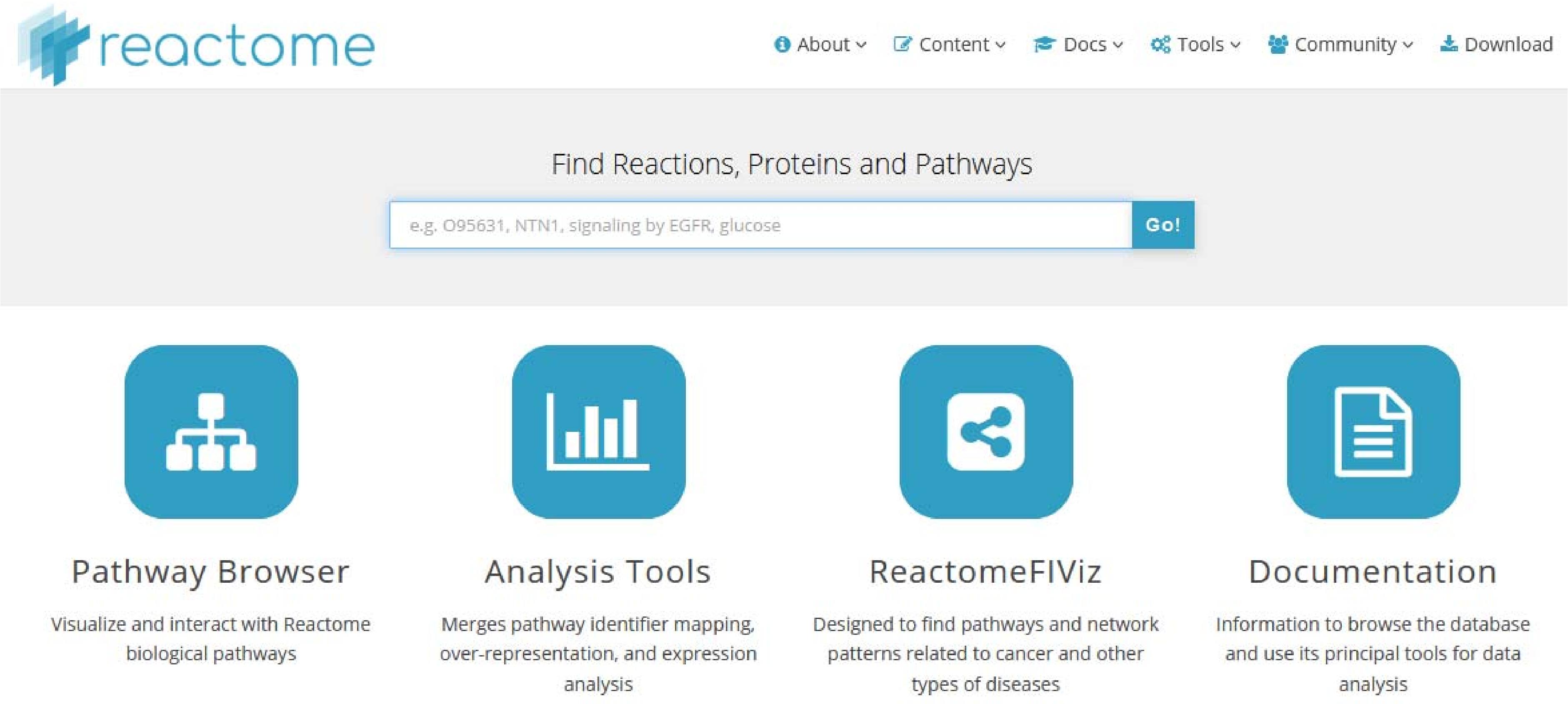
KEGG PATHWAY is a collection of manually drawn pathway maps representing our knowledge of the molecular interaction, reaction and relation networks for:

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/



Find Reactions, Proteins and Pathways

Pathway Browser

Visualize and interact with Reactome biological pathways

Analysis Tools

Merges pathway identifier mapping, over-representation, and expression analysis

ReactomeFIViz

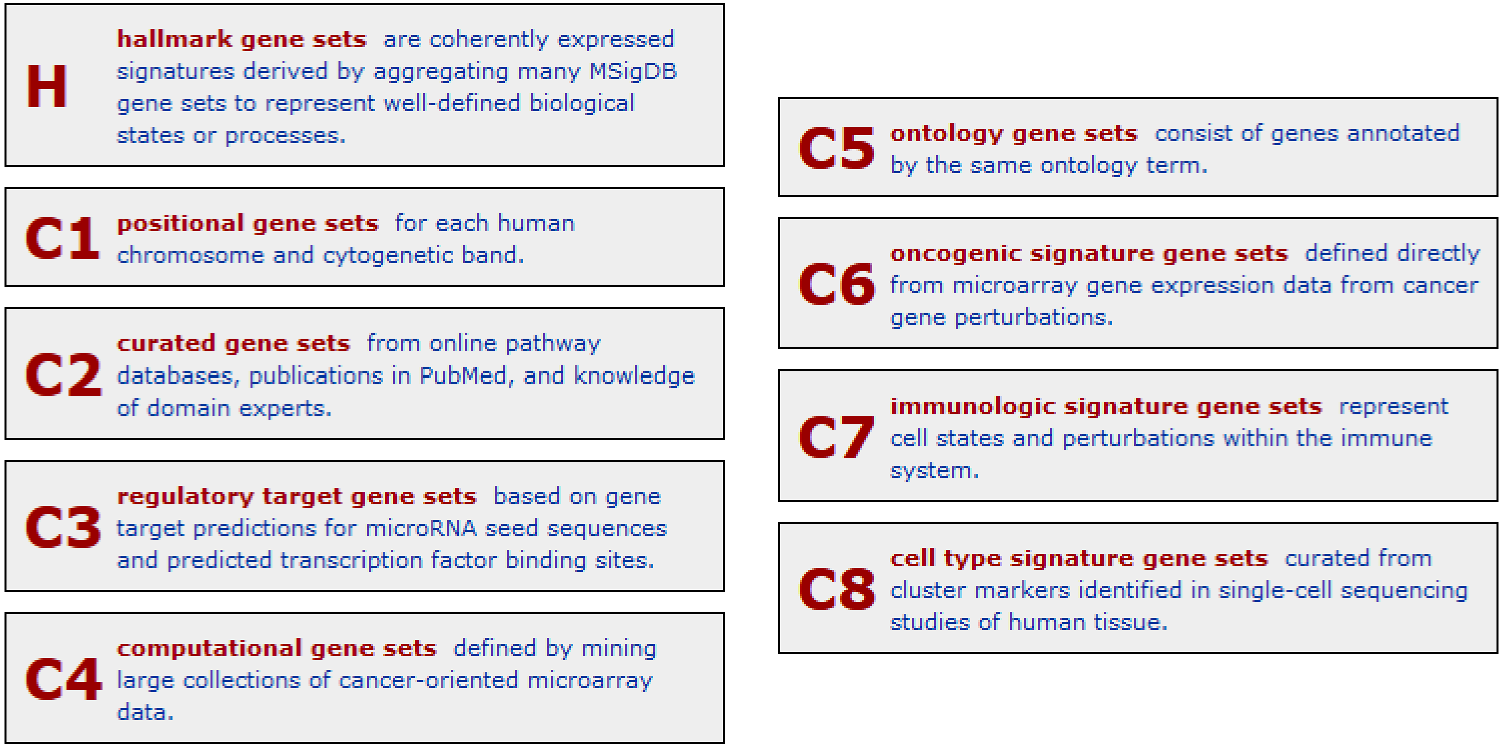
Designed to find pathways and network patterns related to cancer and other types of diseases

Documentation

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.

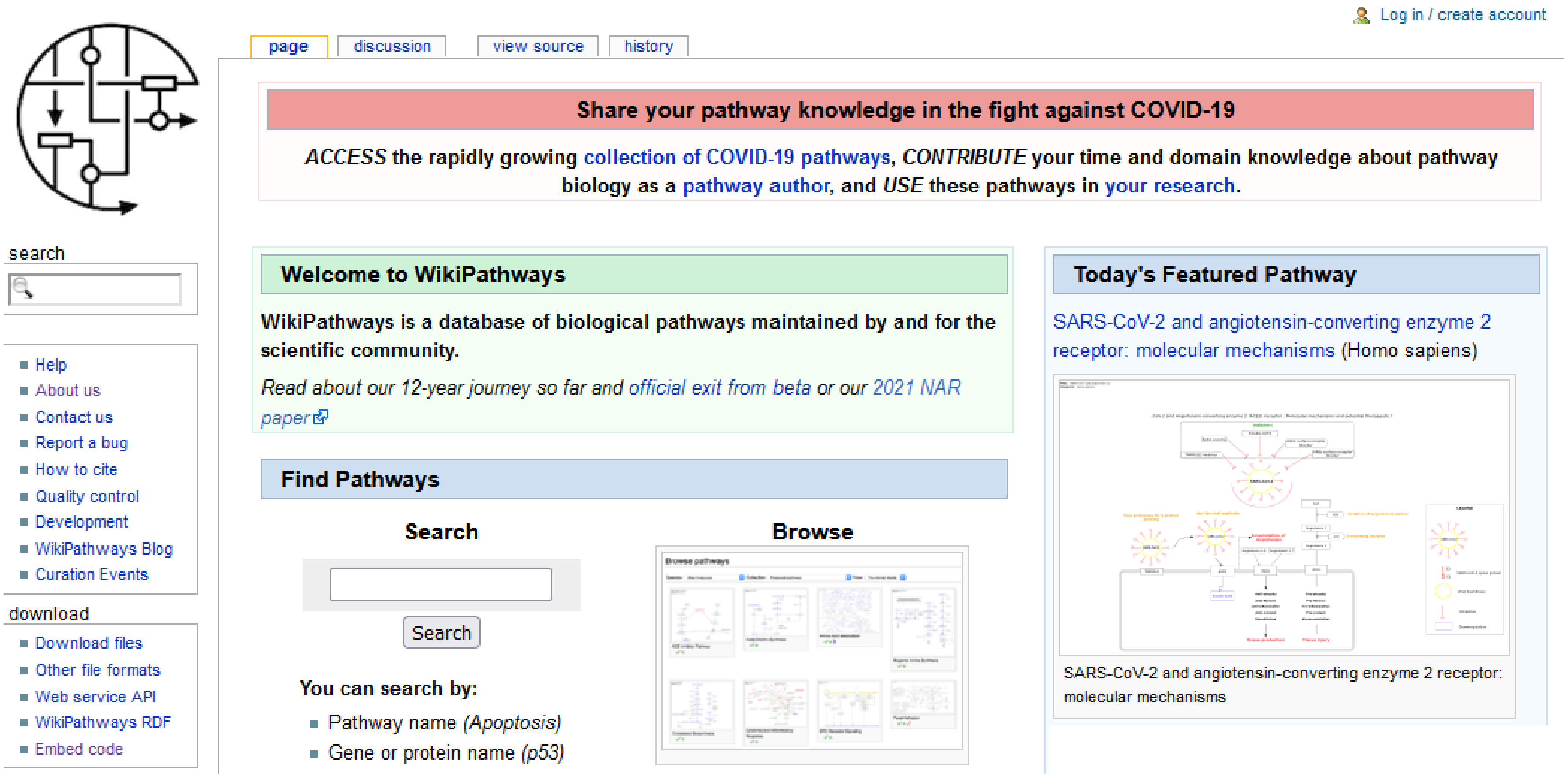
C7 **immunologic signature gene sets** represent cell states and perturbations within the immune system.

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

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

WikiPathways

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



- 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

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)*

Browse

Today's Featured Pathway

SARS-CoV-2 and angiotensin-converting enzyme 2 receptor: molecular mechanisms (Homo sapiens)

SARS-CoV-2 and angiotensin-converting enzyme 2 receptor: molecular mechanisms

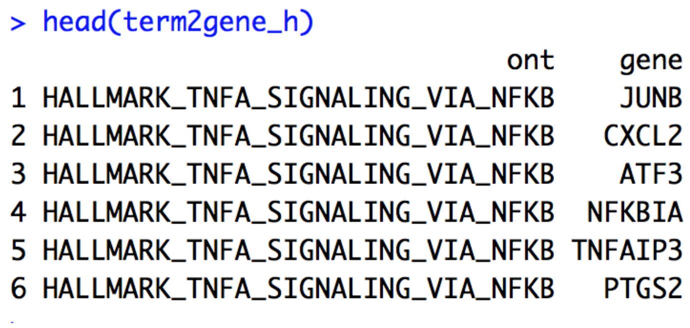
GSEA of other gene sets in R

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?

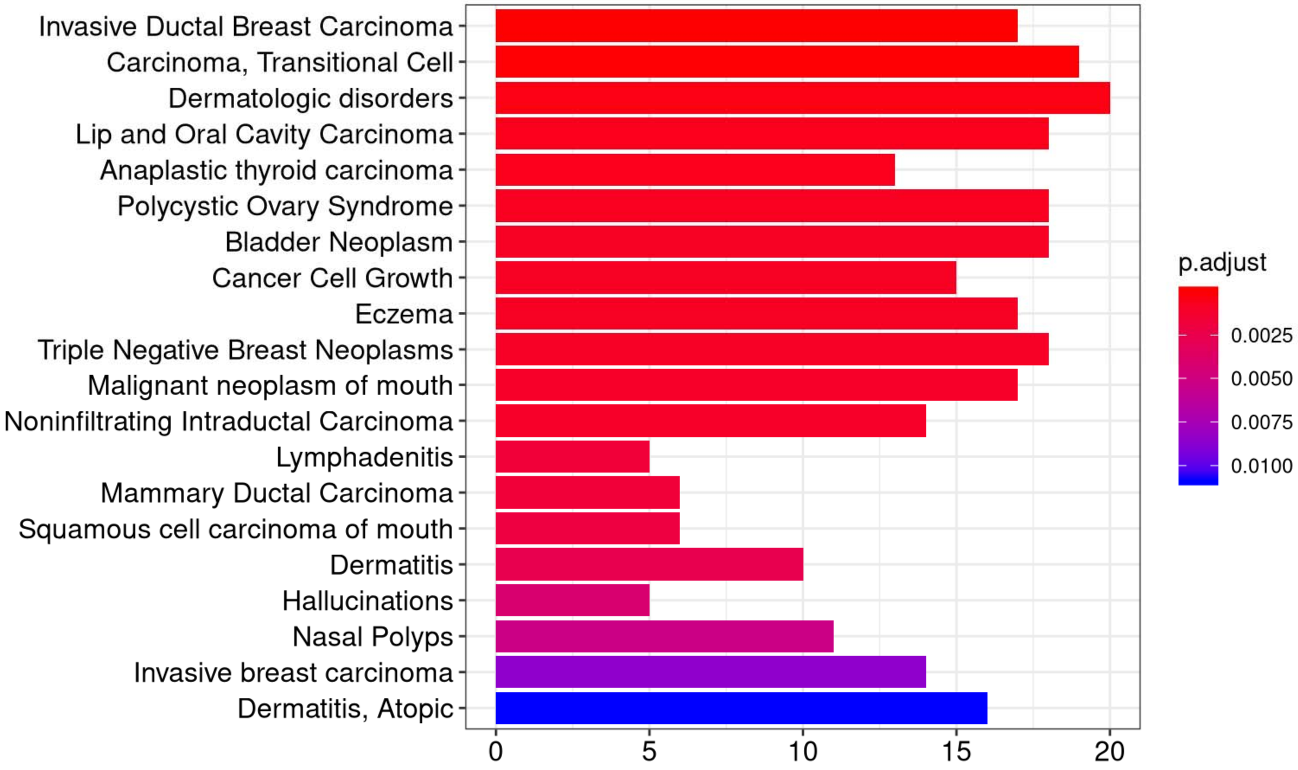
  - If you want to see which genes are included in one of the built-in KEGG pathways, where could you find this information?

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

Visualization of Functional Enrichment Results

- barplot

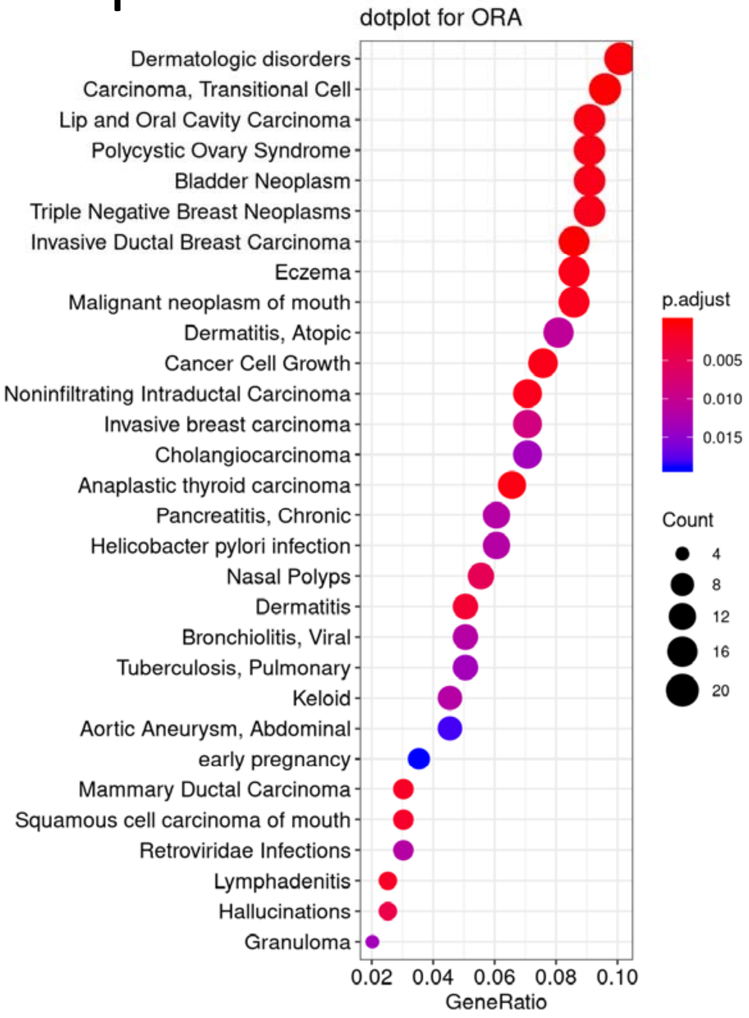
```
ego <- enrichGO(de, OrgDb='org.Hs.eg.db', ont="BP", keyType = "SYMBOL")
barplot(ego, showCategory=20)
```

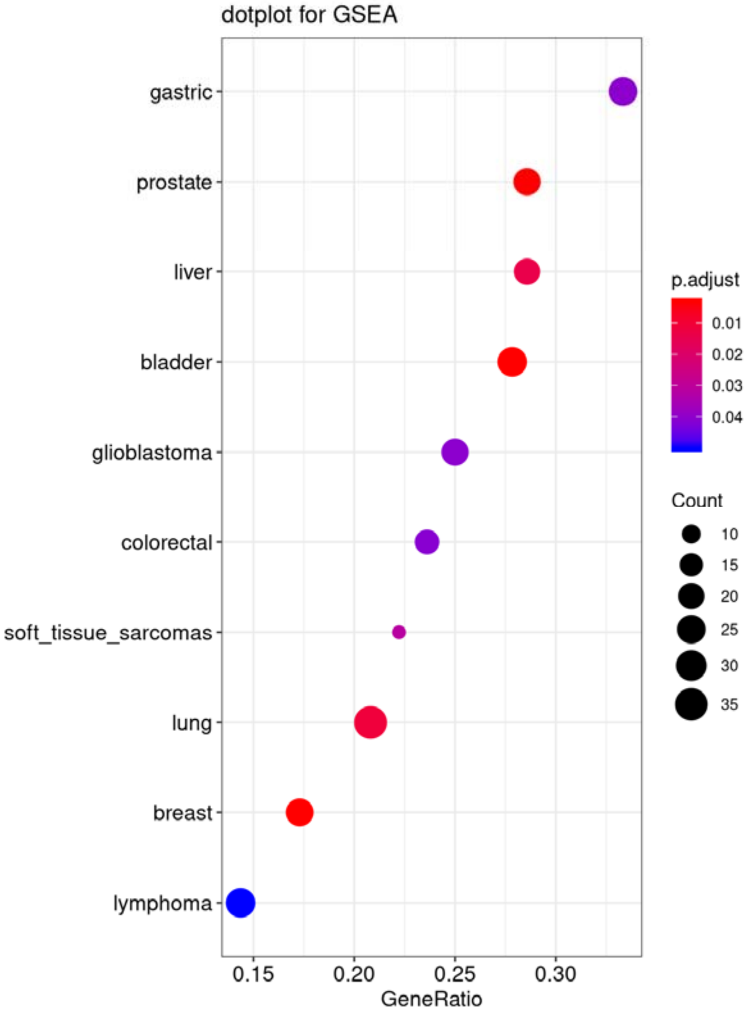


Visualization of Functional Enrichment Results

- dotplot

```
dotplot(ego, showCategory=20)
```

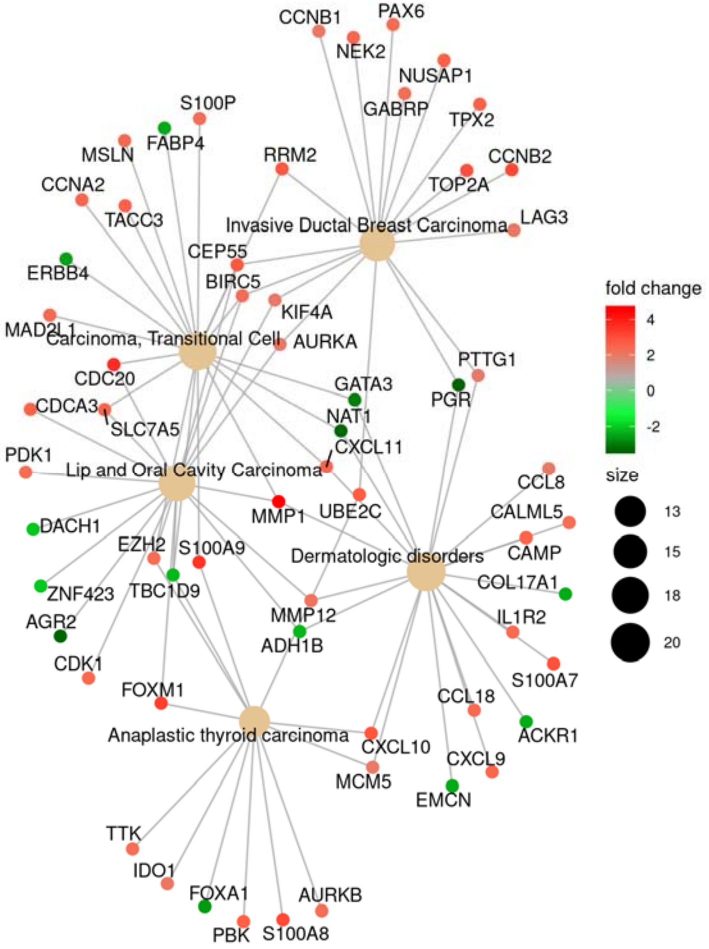




Visualization of Functional Enrichment Results

- cnetplot

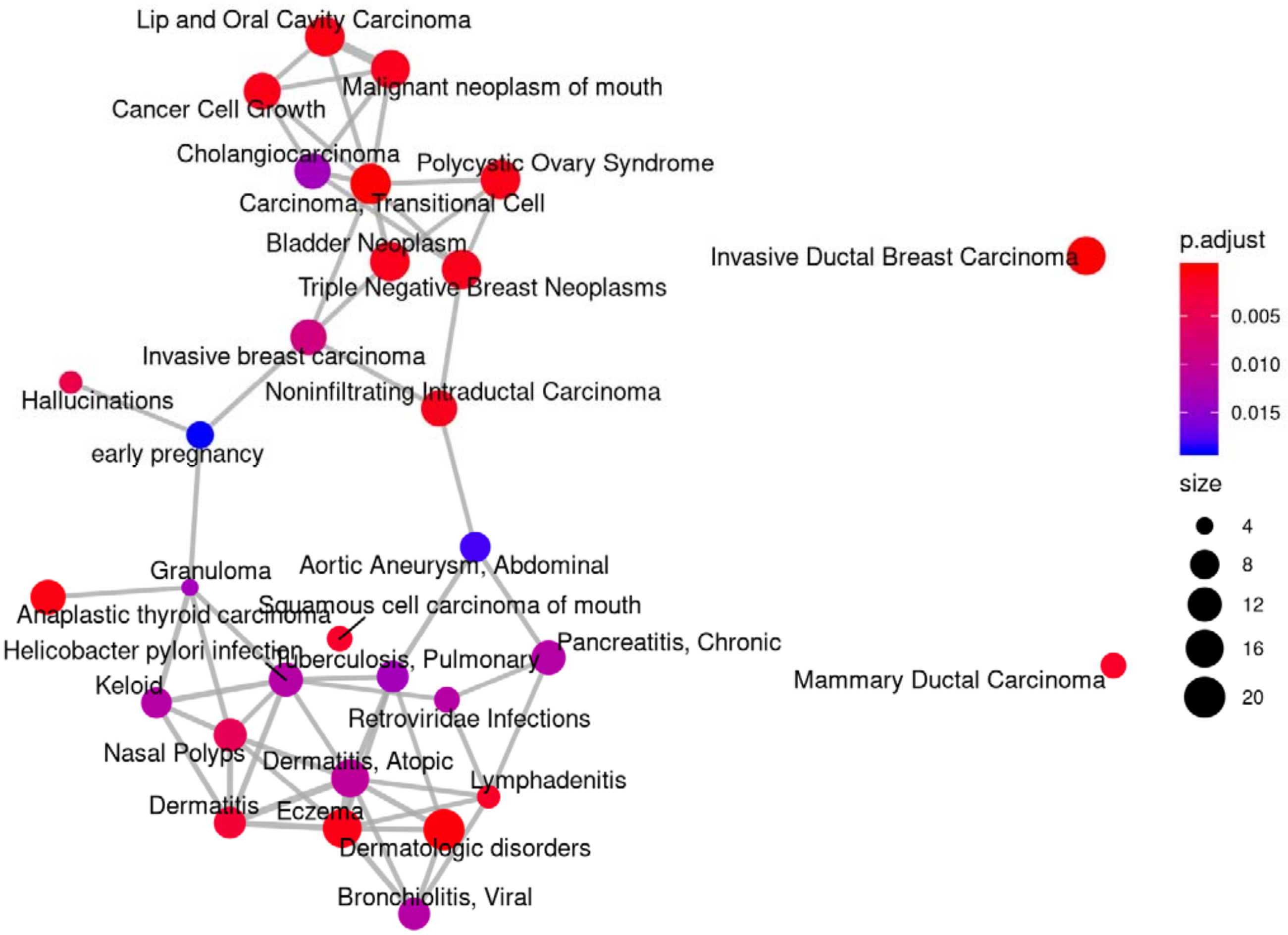
```
cnetplot(ego, categorySize="pvalue", foldChange=geneList)
```



Visualization of Functional Enrichment Results

- Enrichment map

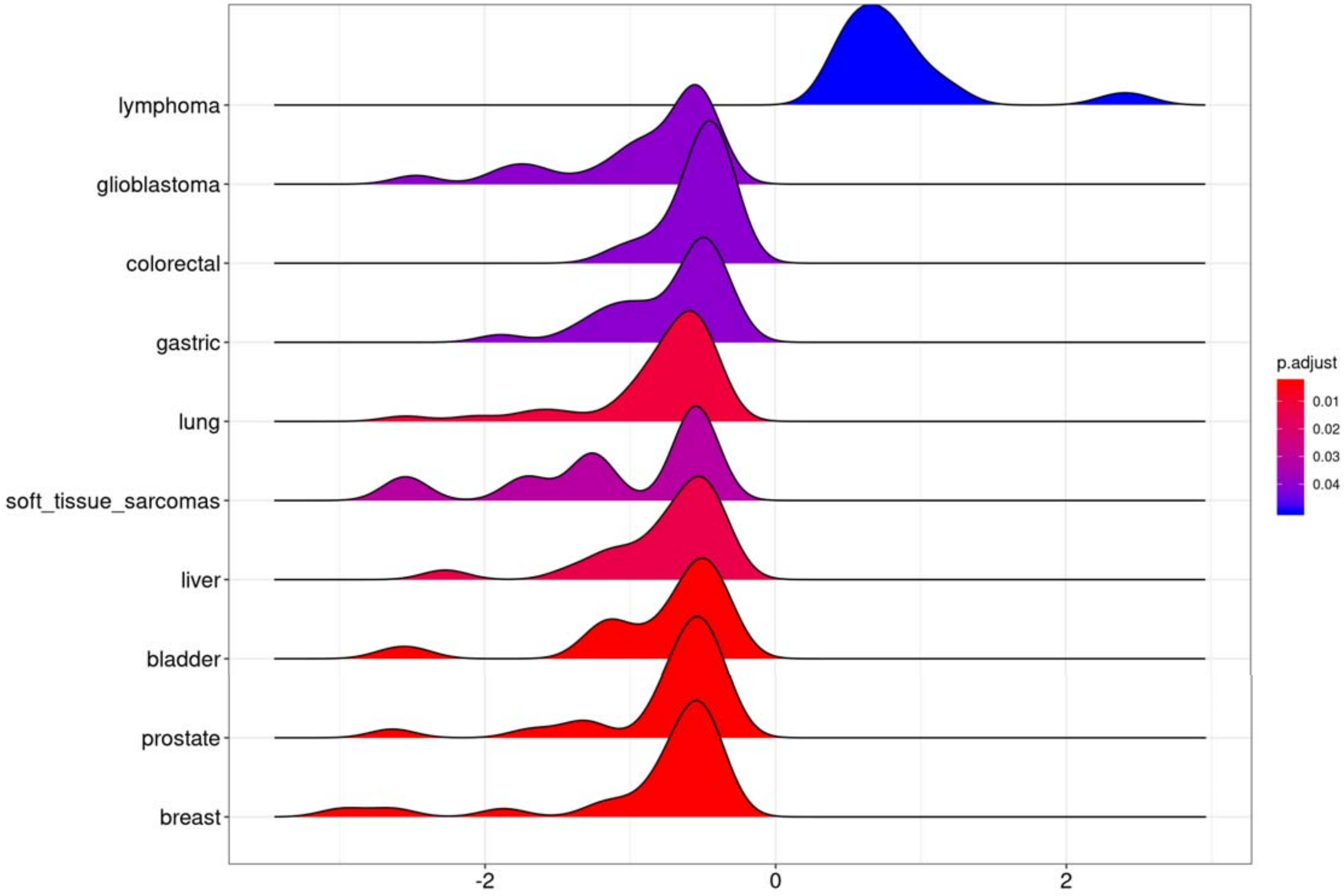
emapplot(ego)


Lip and Oral Cavity Carcinoma
Malignant neoplasm of mouth
Cancer Cell Growth
Cholangiocarcinoma
Polycystic Ovary Syndrome
Carcinoma, Transitional Cell
Bladder Neoplasm
Triple Negative Breast Neoplasms
Invasive Ductal Breast Carcinoma
Invasive breast carcinoma
Hallucinations
Noninfiltrating Intraductal Carcinoma
early pregnancy
Granuloma
Aortic Aneurysm, Abdominal
Anaplastic thyroid carcinoma
Squamous cell carcinoma of mouth
Helicobacter pylori infection
Tuberculosis, Pulmonary
Pancreatitis, Chronic
Mammary Ductal Carcinoma
Keloid
Retroviridae Infections
Nasal Polyps
Dermatitis, Atopic
Lymphadenitis
Dermatitis
Eczema
Dermatologic disorders
Bronchiolitis, Viral
p.adjust
0.005
0.010
0.015
size
4
8
12
16
20

Visualization of Functional Enrichment Results

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

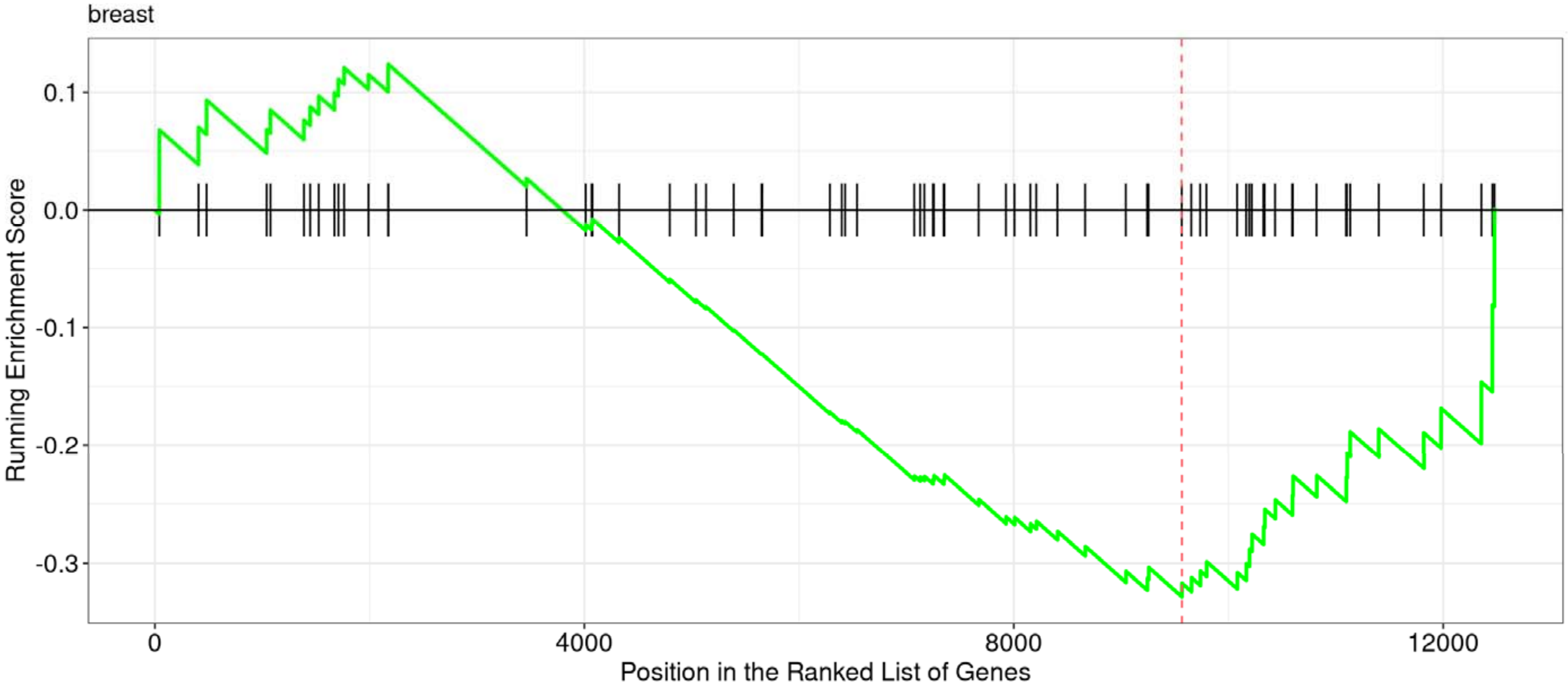
- Ridgeplot



Visualization of Functional Enrichment Results

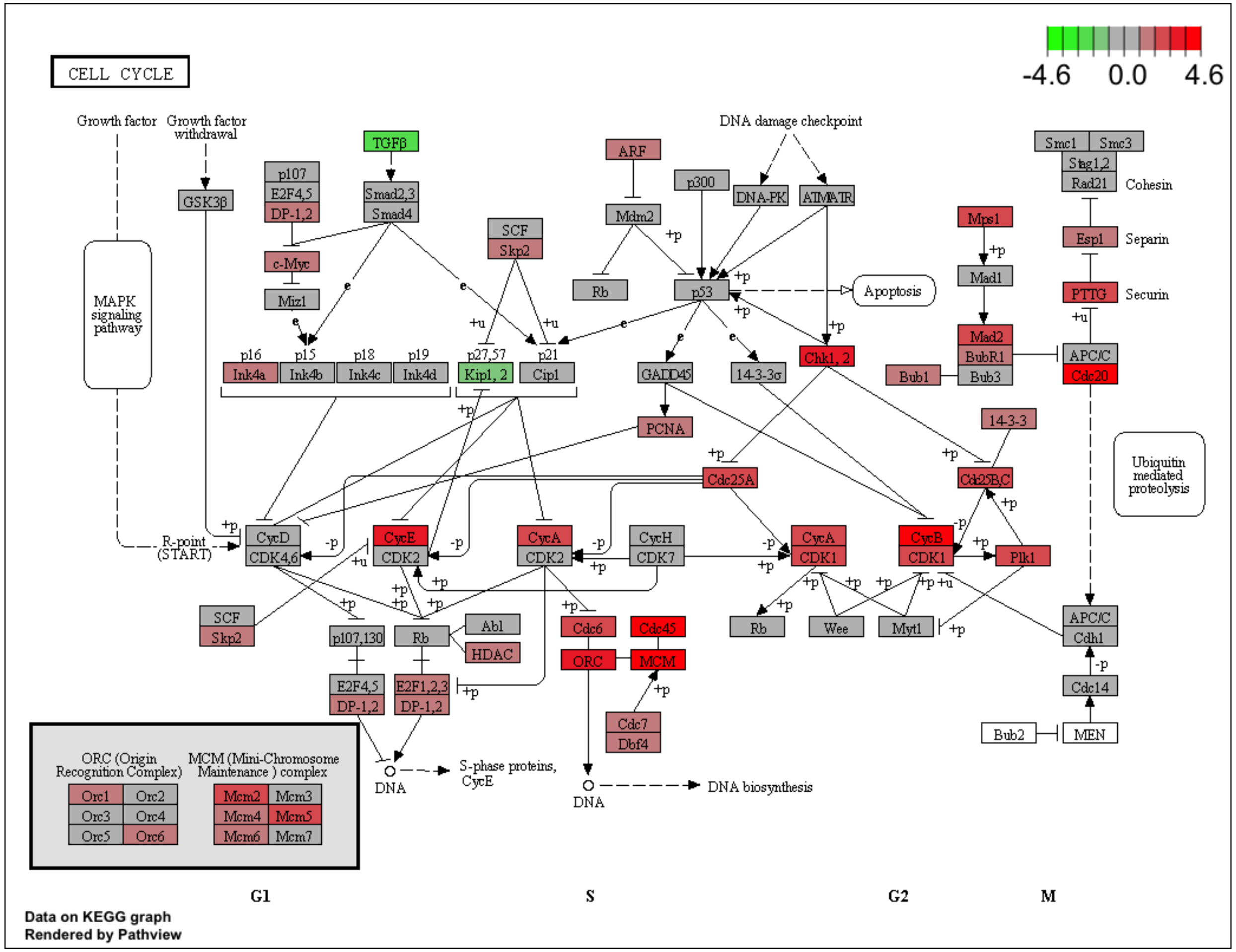
- visualizing GSEA result

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



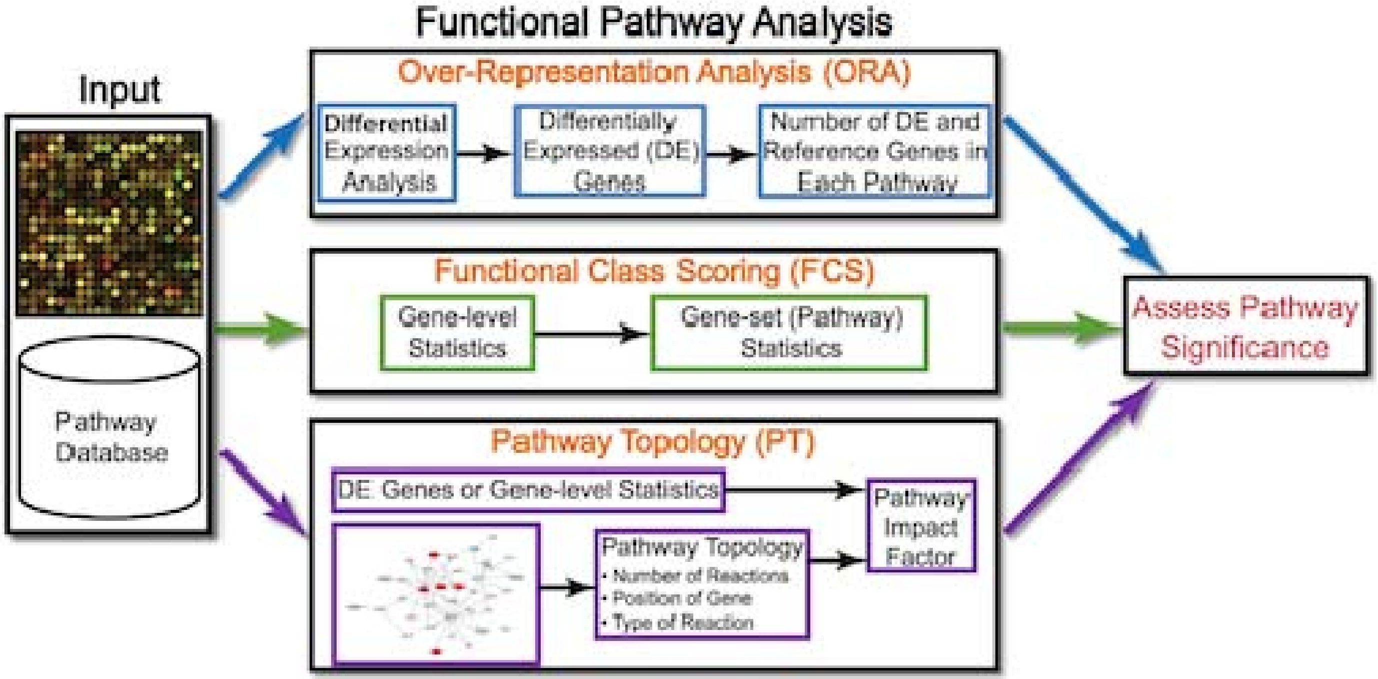
Visualization of Functional Enrichment Results

- pathview



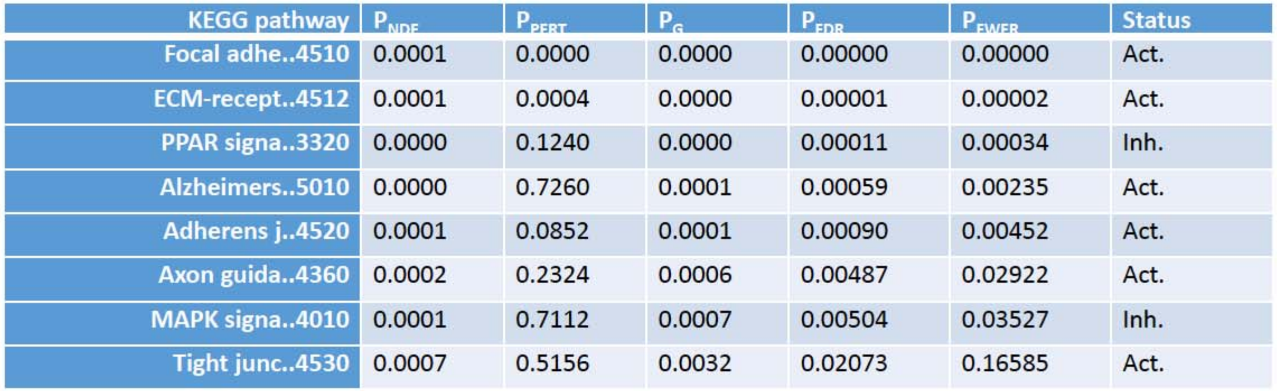
Data on KEGG graph
Rendered by Pathview

Functional analysis


Functional Pathway Analysis
Input
Pathway Database
Over-Representation Analysis (ORA)
Differential Expression Analysis
Differentially Expressed (DE) Genes
Number of DE and Reference Genes in Each Pathway
Functional Class Scoring (FCS)
Gene-level Statistics
Gene-set (Pathway) Statistics
Pathway Topology (PT)
DE Genes or Gene-level Statistics
Pathway Topology
• Number of Reactions
• Position of Gene
• Type of Reaction
Pathway Impact Factor
Assess Pathway Significance

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

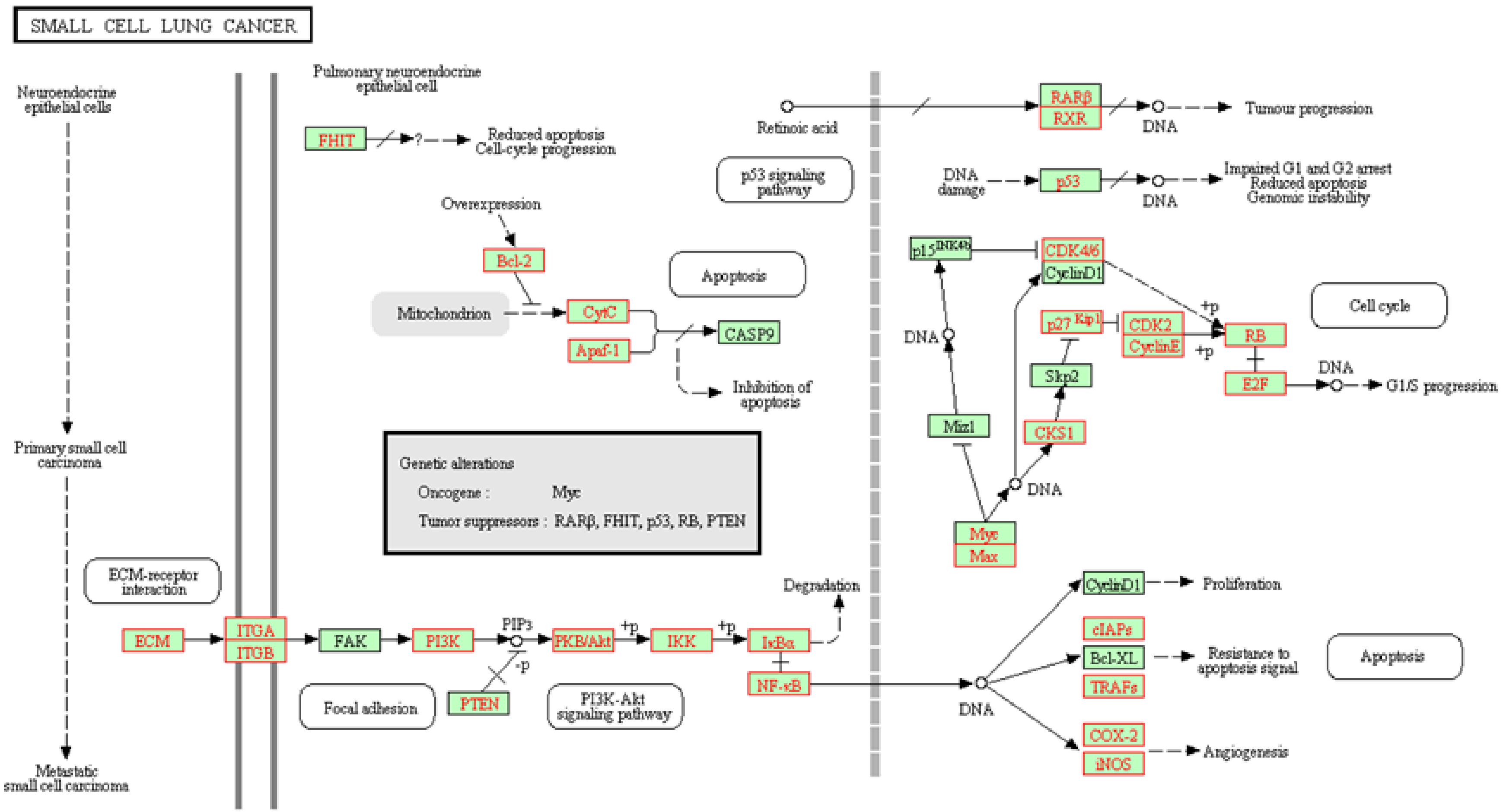


KEGG pathway	P_{NDE}	P_{PERT}	P_G	P_{FDR}	P_{FWER}	Status
Focal adhe..4510	0.0001	0.0000	0.0000	0.00000	0.00000	Act.
ECM-recept..4512	0.0001	0.0004	0.0000	0.00001	0.00002	Act.
PPAR signa..3320	0.0000	0.1240	0.0000	0.00011	0.00034	Inh.
Alzheimers..5010	0.0000	0.7260	0.0001	0.00059	0.00235	Act.
Adherens 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)

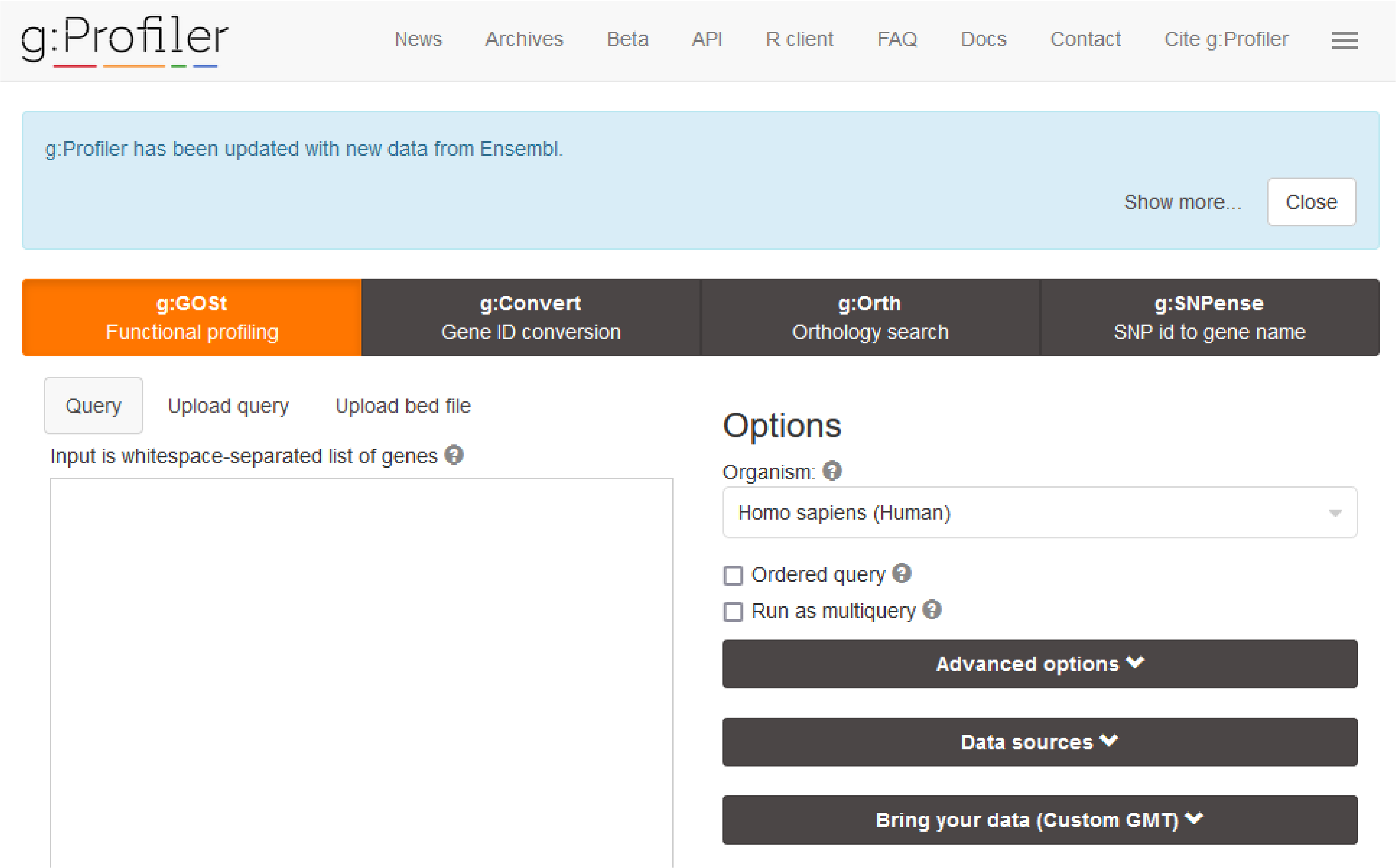
https://bioconductor.org/packages/release/bioc/html/SPIA.html

Functional analysis: **Pathway topology tools**



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

Additional resources for functional analysis



g:Profiler

g:Profiler has been updated with new data from Ensembl.

Input is whitespace-separated list of genes

Options

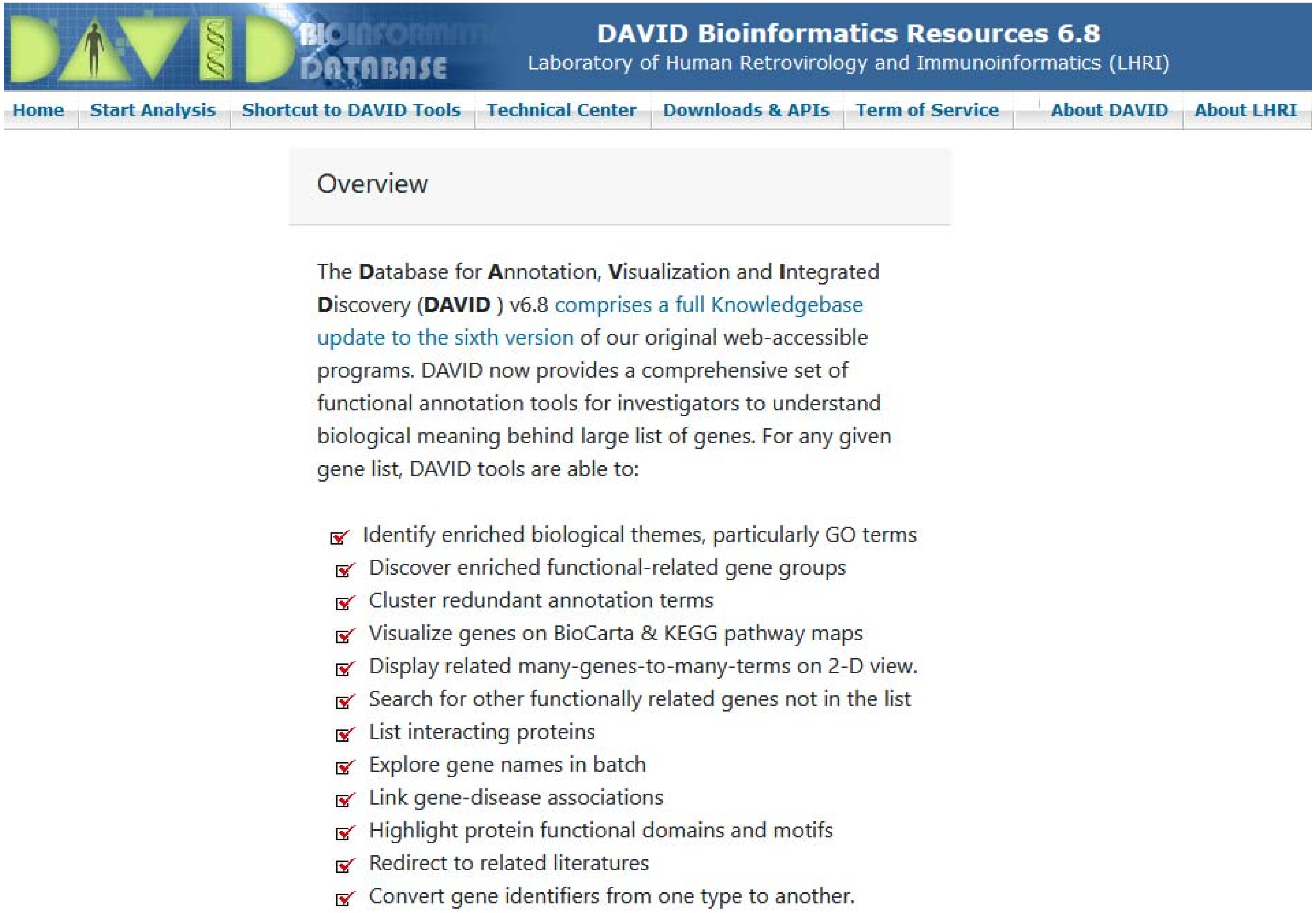
Organism: Homo sapiens (Human)

□ Ordered query

□ Run as multiquery

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

Additional resources for functional analysis



DAVID Bioinformatics Resources 6.8

Laboratory of Human Retrovirology and Immunoinformatics (LHRI)

Overview

The **D**atabase for **A**nnotation, **V**isualization and **I**ntegrated **D**iscovery (**DAVID**) v6.8 comprises a full Knowledgebase update to the sixth version of our original web-accessible programs. DAVID now provides a comprehensive set of functional annotation tools for investigators to understand biological meaning behind large list of genes. For any given gene list, DAVID tools are able to:

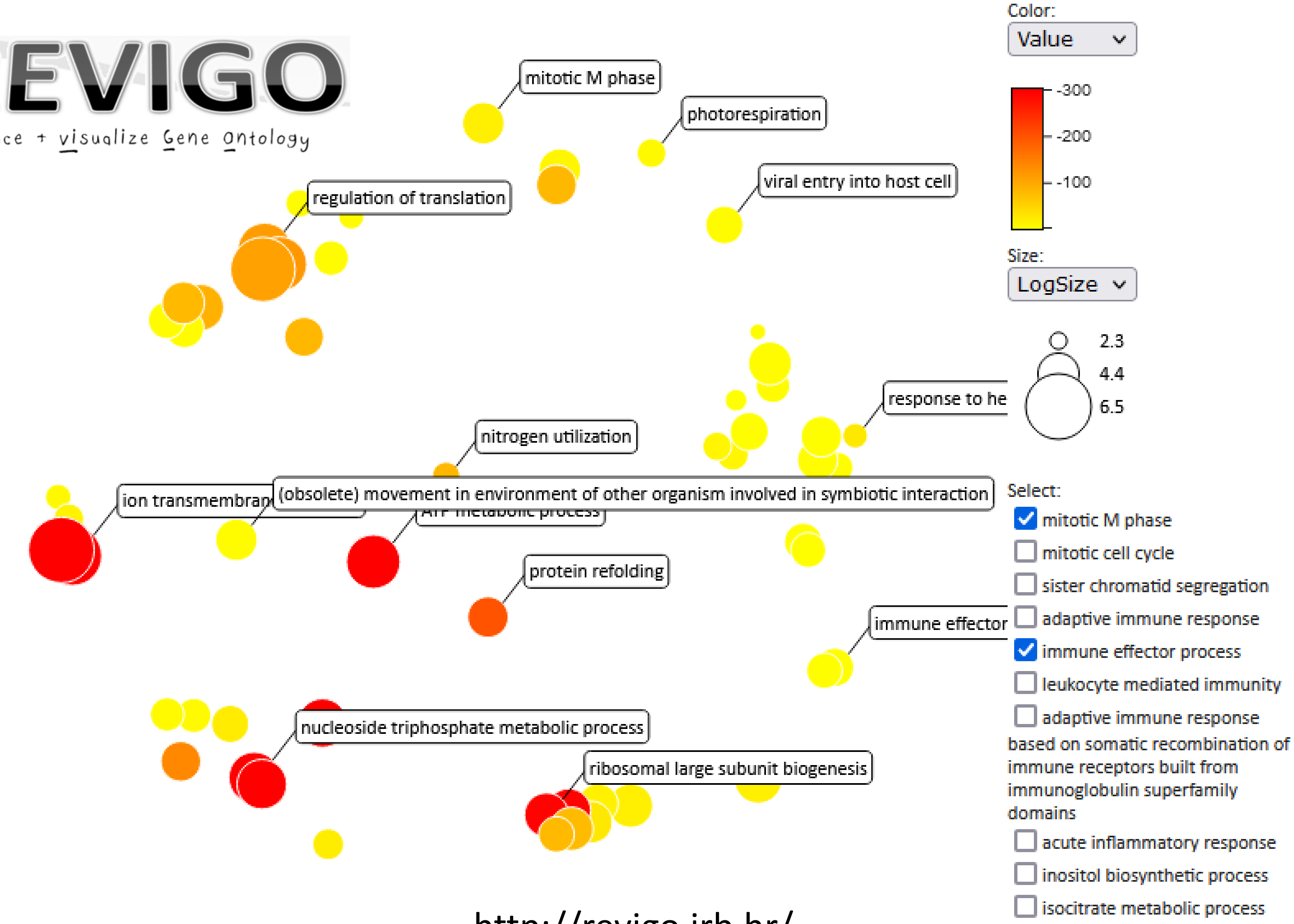
- ☑ Identify enriched biological themes, particularly GO terms
- ☑ Discover enriched functional-related gene groups
- ☑ Cluster redundant annotation terms
- ☑ Visualize genes on BioCarta & KEGG pathway maps
- ☑ Display related many-genes-to-many-terms on 2-D view.
- ☑ Search for other functionally related genes not in the list
- ☑ List interacting proteins
- ☑ Explore gene names in batch
- ☑ Link gene-disease associations
- ☑ Highlight protein functional domains and motifs
- ☑ Redirect to related literatures
- ☑ Convert gene identifiers from one type to another.

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

Additional resources for functional analysis

REVIGO

reduce + visualize Gene ontology


mitotic M phase
photorespiration
viral entry into host cell
regulation of translation
response to he
nitrogen utilization
ion transmembran
(obsolete) movement in environment of other organism involved in symbiotic interaction
ATP metabolic process
protein refolding
immune effector
nucleoside triphosphate metabolic process
ribosomal large subunit biogenesis
-300
-200
-100
2.3
4.4
6.5

Color: Value

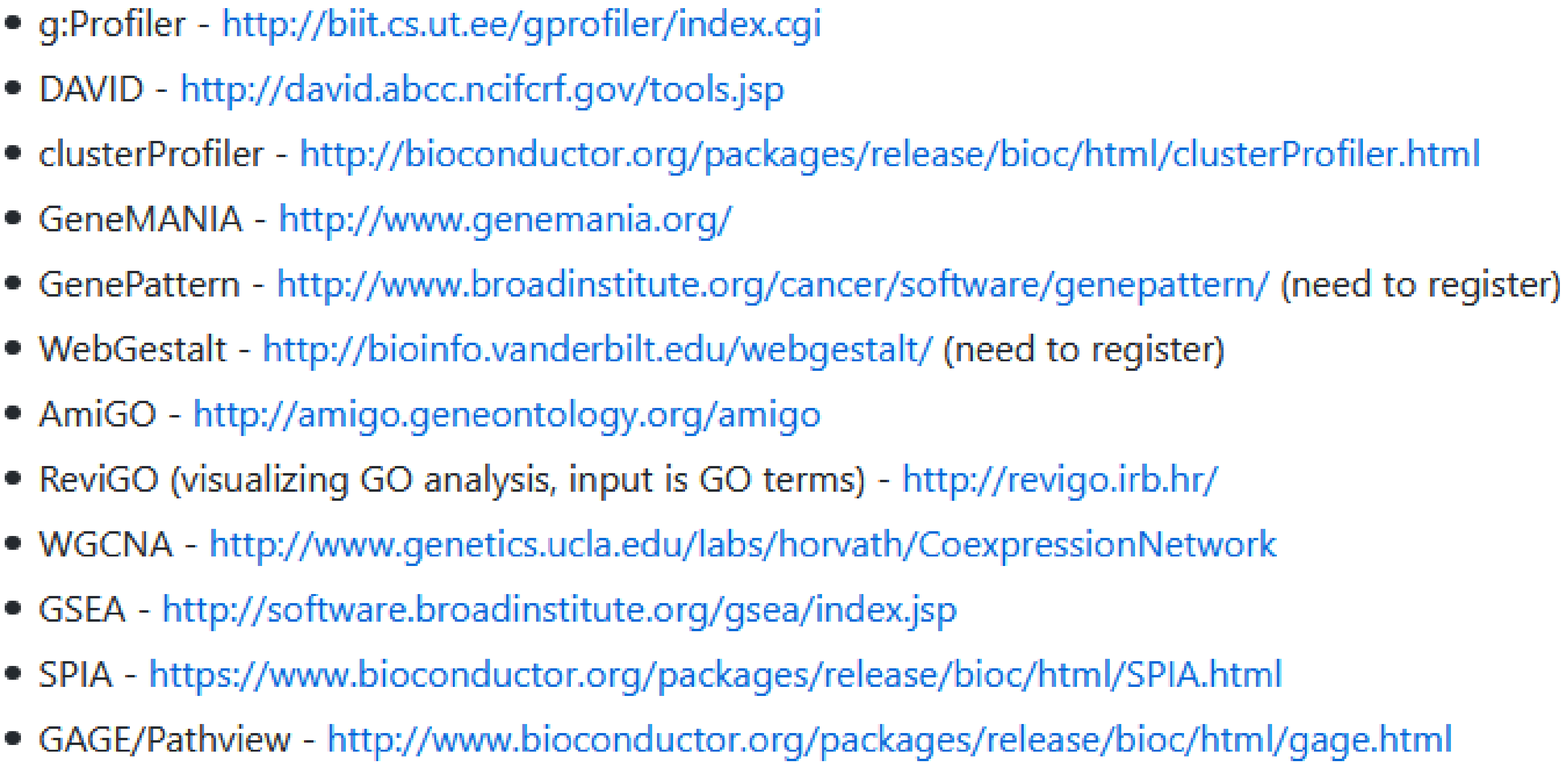
Size: LogSize

Select:

- ☑ mitotic M phase
- □ mitotic cell cycle
- □ sister chromatid segregation
- □ adaptive immune response
- ☑ immune effector process
- □ leukocyte mediated immunity
- □ adaptive immune response based on somatic recombination of immune receptors built from immunoglobulin superfamily domains
- □ acute inflammatory response
- □ inositol biosynthetic process
- □ isocitrate metabolic process

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