

Swiss Institute
of
Bioinformatics

Enrichment analysis

June 24th 2022

tania.wyss@sib.swiss

gustavo.ruizbuendia@sib.swiss

Schedule

- **9:00 - 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



Swiss Institute of
Bioinformatics

[Let's collaborate](#)

Careers

Contact

Directory



Research infrastructure ▾

Scientific community ▾

About SIB ▾



Raphael Gottardo's group

The Translational Data Science (TDS) group focuses on developing novel computational tools, statistical methods and machine learning algorithms...

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

The Translational Data Science (TDS) group focuses on developing novel computational tools, statistical methods and machine learning algorithms for the analysis of high-throughput and high-dimensional datasets generated by novel assay technologies with applications in immunology, vaccine research and immunotherapy. We collaborate with multiple research groups locally, nationally and internationally to address important immunological problems through high-dimensional modeling and data integration.

Domains of activity:

Core facility and competence center, Biostatistics, Infectious diseases, Machine learning, Mathematical modeling, Next generation sequencing, Oncology, Personalized medicine, Single-cell biology, Transcriptomics, Vaccines

For core facility service inquiry: nadine.fournier@sib.swiss

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

Course material

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

The screenshot shows a website with a red header bar. On the left, there's a small logo and the text "Enrichment analysis". Below the header, on the left side, 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 RNA sequencing data from Ercolano et al 2020.

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

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](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 Monique Zahn)
- Send answers to tania.wyss@sib.swiss by July 1st 2022, 11:59pm

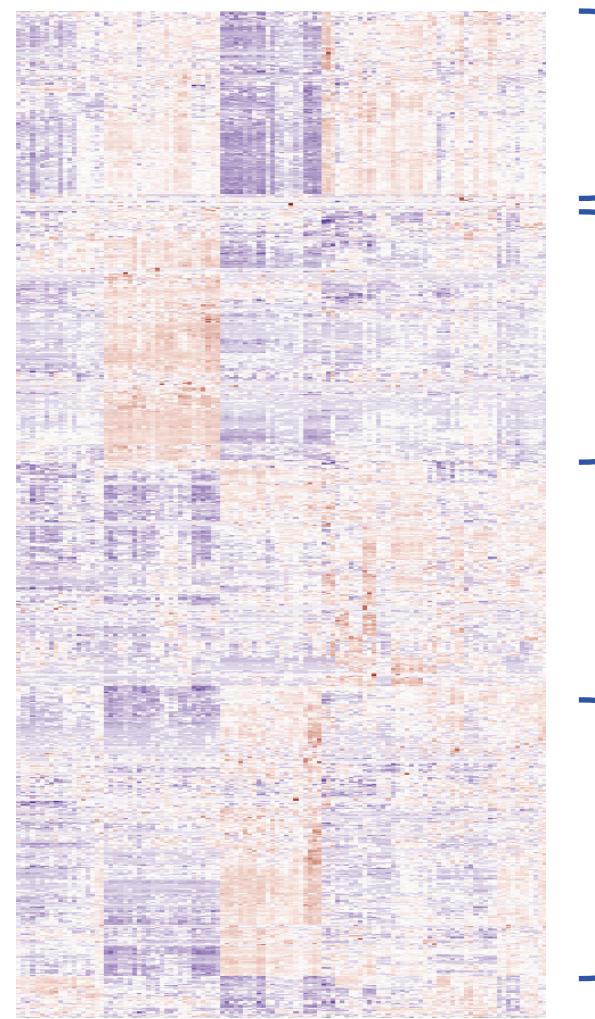
Questions and Exercises

- Feel free to interrupt with questions by asking them directly or raising your (virtual) hand.
- Use the chat or Q&A in google Doc, Gustavo and I will answer
- Exercises in R:
 - We will try to debug as much as possible
 - We are happy if you share your results or alternative code!
 - Computational power on RStudio cloud is limited, might crash



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

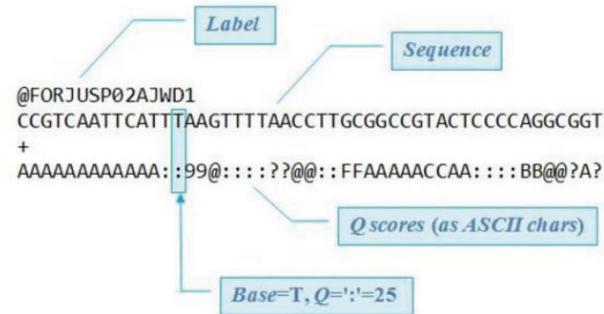


IVY GAP: <https://glioblastoma.alleninstitute.org/>

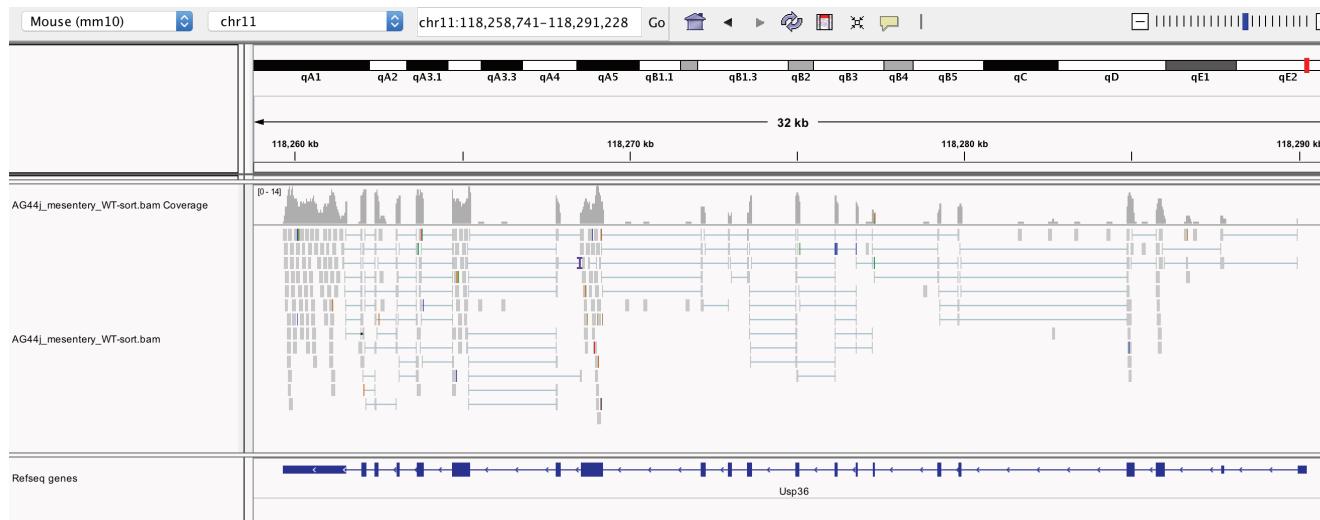
Typical RNA sequencing analysis workflow

fastq file:

```
@HWI-M01141:63:A4NDL:1:1101:14849:1418 1:N:0:TATAGCAGACACCGT  
NACGAAGGGTCAAGCGTTACTCGGAATTACTGGCGTAAGCGTGCCTAGGTGGTT/  
+  
#>>>A??FAA1BGGEGGAAFGCA@BFF1D2BCF/EEG/DBEE/E?GAEEFGAEFG1  
@HWI-M01141:63:A4NDL:1:1101:13802:1421 1:N:0:TATAGCAGACACCGT  
NACGGAGGGTCAAGCGTTAACCGAATTACTGGCGTAAGCGCACGCAGCGGTGTT/  
+  
#>>AAABBBABBGGGGGGGG?FHGGGGGGHHHHHHHHGGGGH  
@HWI-M01141:63:A4NDL:1:1101:15928:1426 1:  
NACGTAGGGTCCAGCGTTAACCGAATTACTGGCGTAAA/  
+  
#>>AABFB@FBGGGGGGGGGGHGGGGFHHHHHHHGGGGH  
@HWI-M01141:63:A4NDL:1:1101:14861:1431 1:  
NACGAAGGGTCAAGCGTTACTCGGAATTACTGGCGTAAA/  
+  
#>>AAAABBFABGGGGGGCEGHGGEFFHHHHHHGGGGH  
@HWI-M01141:63:A4NDL:1:1101:15264:1465 1:  
NACGTAGGGTGCAGCGTTCTCGGAATTACTGGCGTAAA/  
+
```



Filter quality
Align to ref. genome



count reads
→ per gene

Downstream
statistical analysis:
R: import
counts table

Differential gene expression analysis

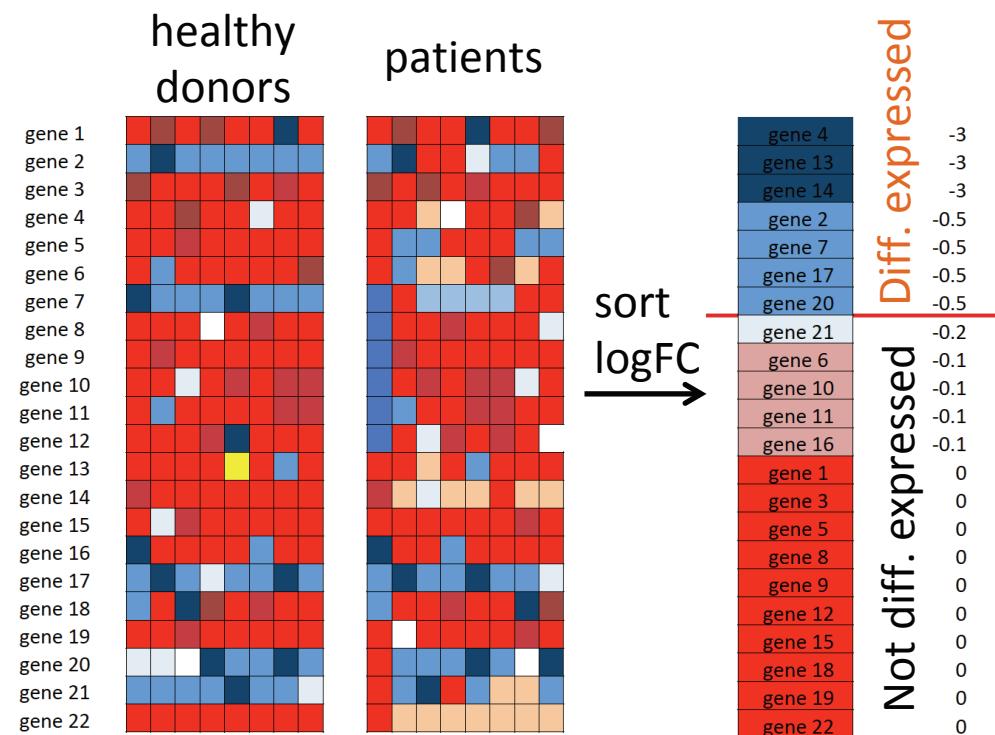
- Comparing 2 groups:

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

- Fold change in genomics:

$$\log_2 \text{of ratios} = \log \text{fold change}$$

$$\log(\pi_{i1}/\pi_{i2}) = \log(\pi_{i1}) - \log(\pi_{i2})$$



Differential gene expression analysis

- Comparing 2 groups:

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

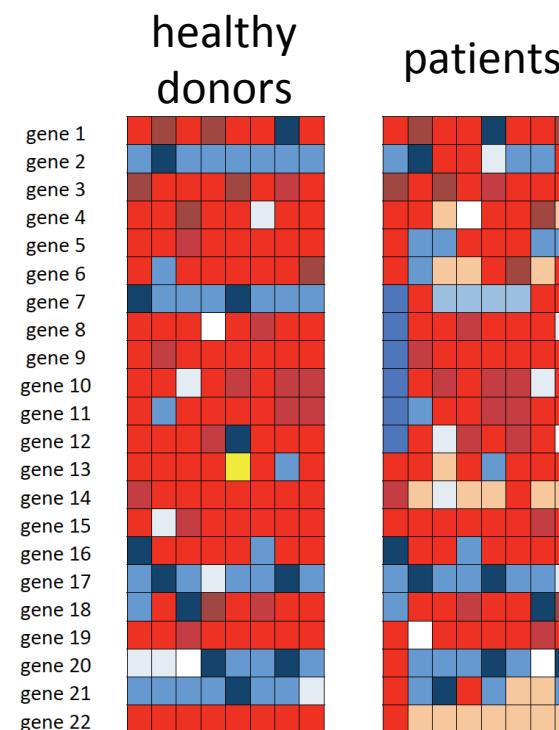
- T-test:

H_0 : Healthy donors and patients have similar gene i expression

$$H_{0i} : \pi_{i1} = \pi_{i2}$$

H_1 : Healthy donors and patients don't have a similar gene i expression

$$H_{1i} : \pi_{i1} \neq \pi_{i2}$$



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:

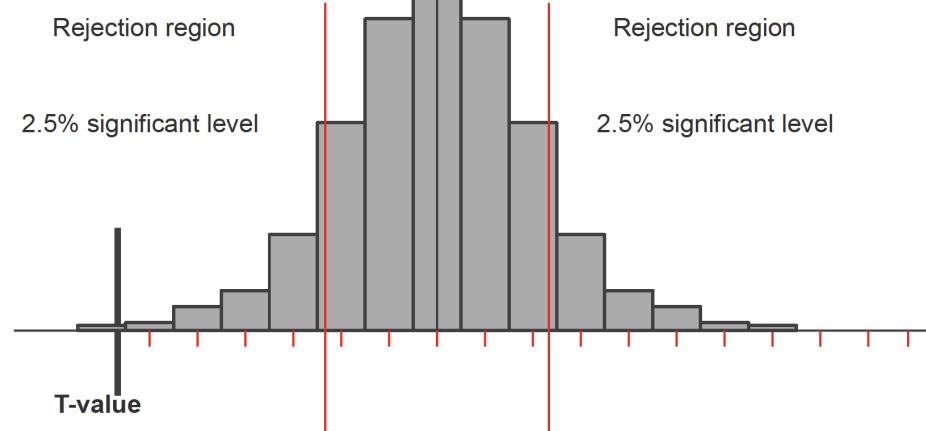
-8.908753 -4.234104

sample estimates:

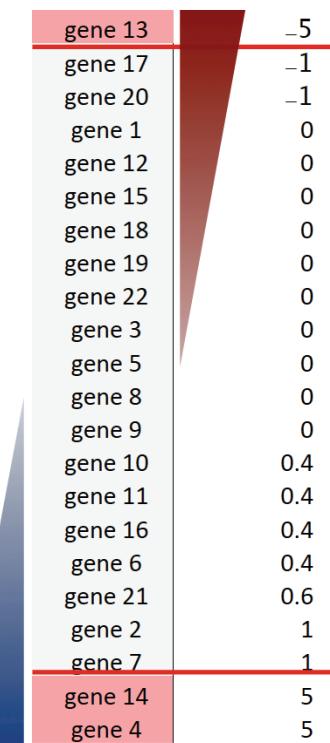
mean of x mean of y

6.00000 12.57143

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

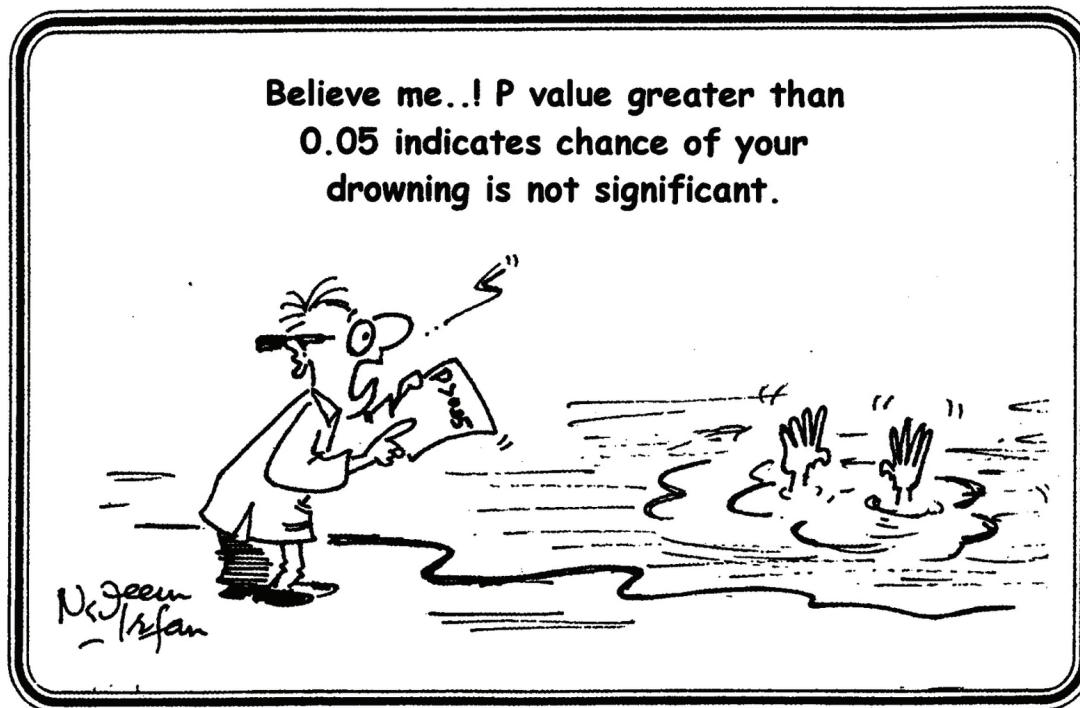


sort based
on T-statistic



What does $p < 0.05$ mean?

- It implies that it is acceptable to have a 5% probability to incorrectly reject the null hypothesis while it is correct.
- It means that if we repeat an experiment 20 times, we would reject the null hypothesis once because of random error.



P-value adjustment: what is it?

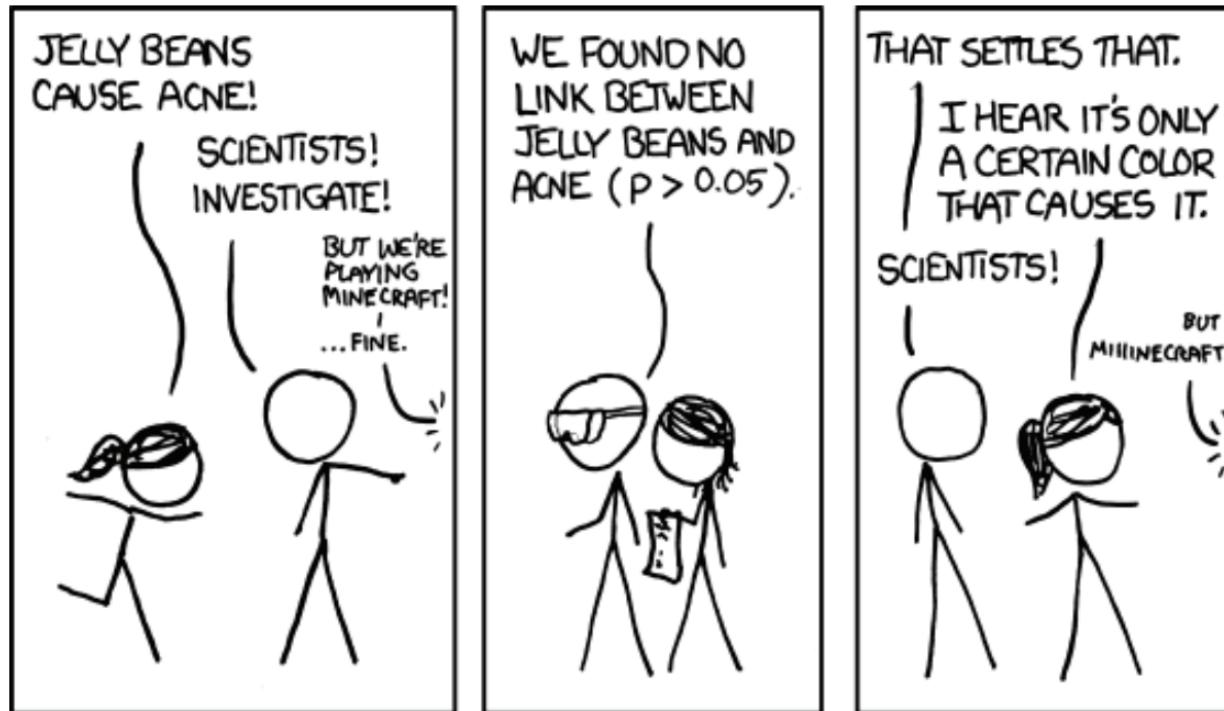


Photo by Patrick Fore on Unsplash

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
($P < 0.05$).


WE FOUND NO
LINK BETWEEN
MAUVE JELLY
BEANS AND ACNE
($P > 0.05$).

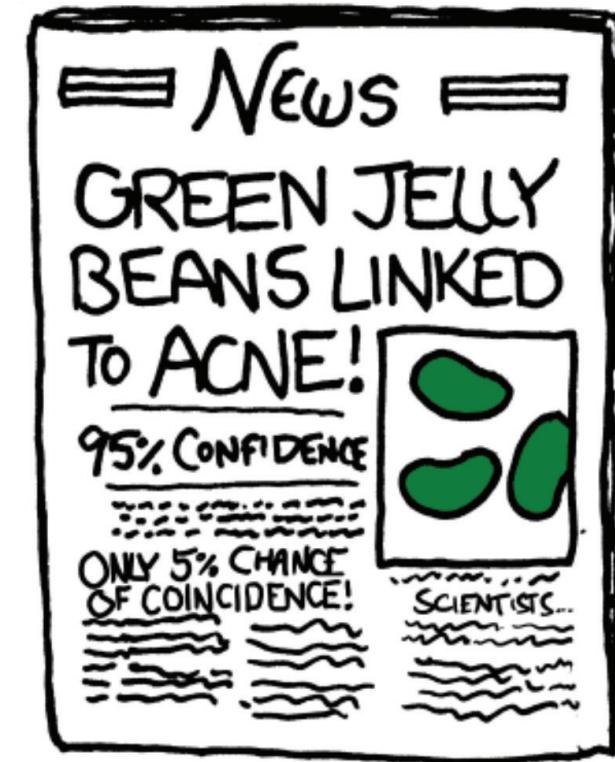

WE FOUND NO
LINK BETWEEN
BEIGE JELLY
BEANS AND ACNE
($P > 0.05$).


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
($P > 0.05$).

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

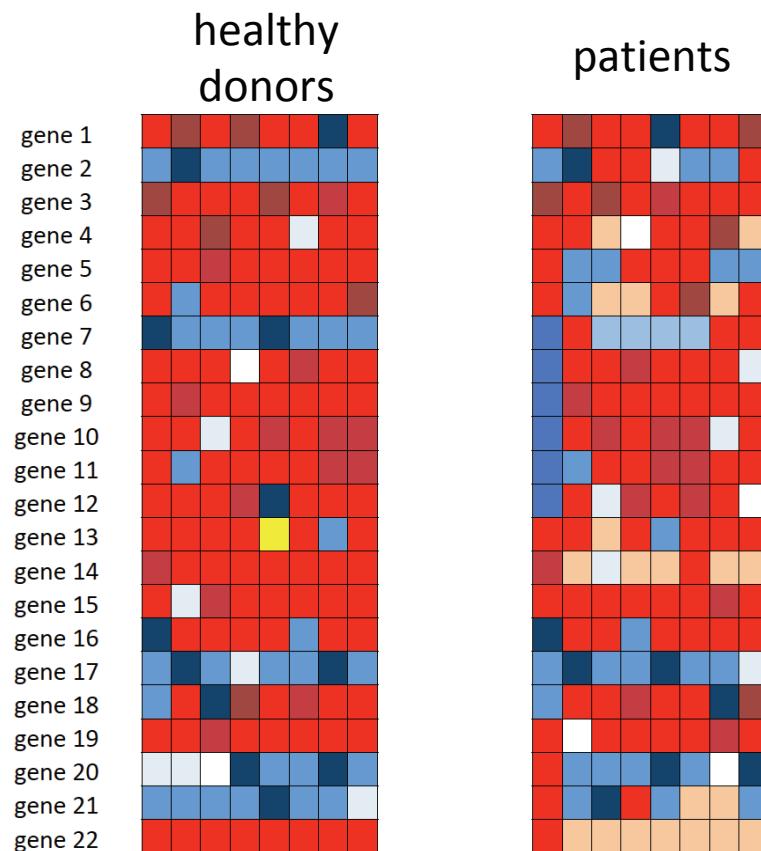
- **Benjamini-Hochberg procedure (BH, decreases the FDR)**
- Rank the p-values from smallest to largest, adjust less and less as the p-values get larger:
 $p\text{-value}_1*(n/1)$
 $p\text{-value}_2*(n/2)$
...
 $p\text{-value}_k*(n/k) = p\text{-value}_k*1$
n= total number of p-values (genes)
k= rank number of each p-value

Differential gene expression analysis using R

- Bioconductor

<https://bioconductor.org/>

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



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

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

RNA sequencing pipeline

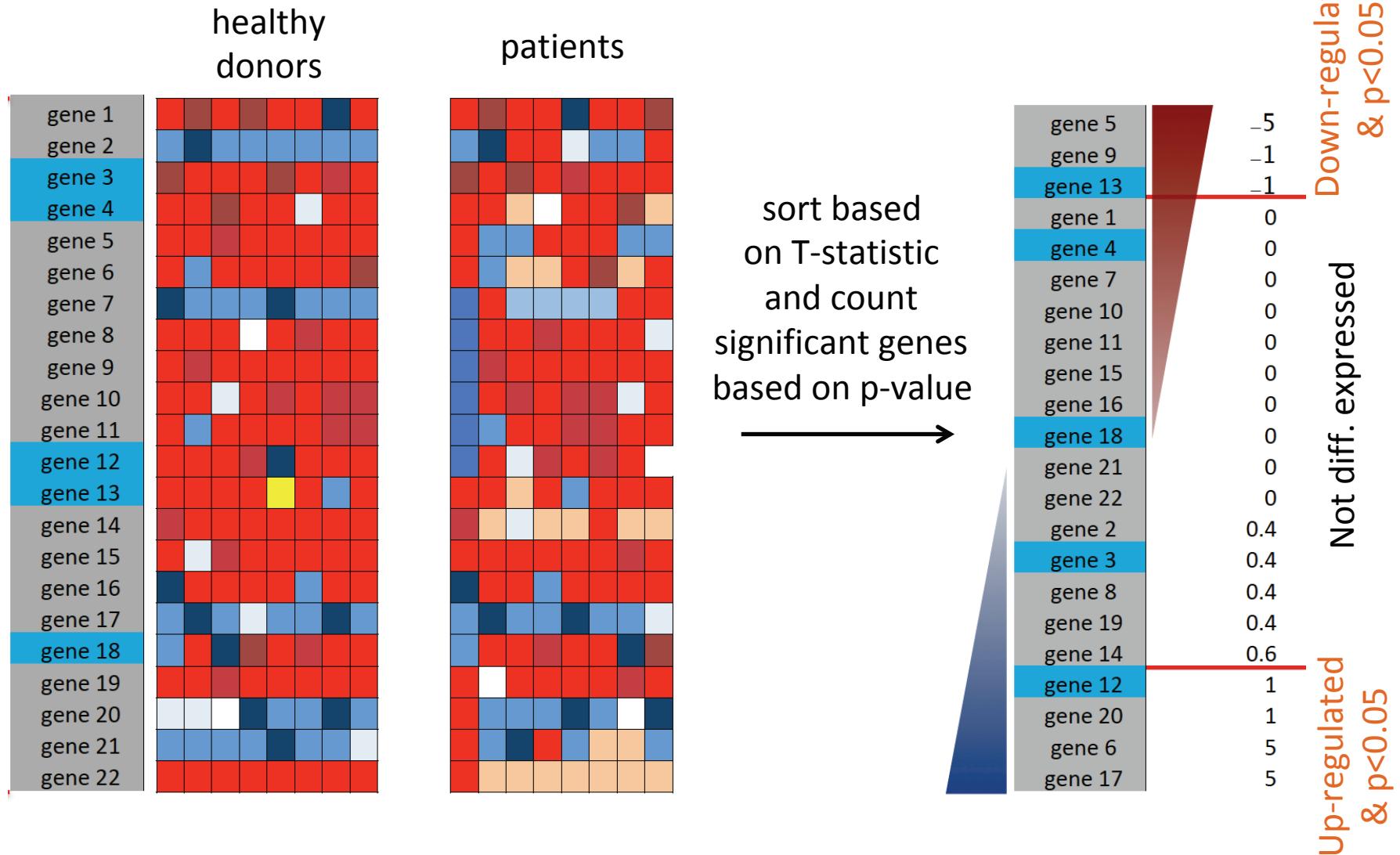
Differential expression analysis

Enrichment analysis

Several methods available, e.g.:

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

Are the genes belonging to the blue set differentially expressed?



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 that are not differentially expressed.

H_1 : The proportion of blue genes differentially expressed is not the same as the proportion of blue genes that are not differentially expressed.

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

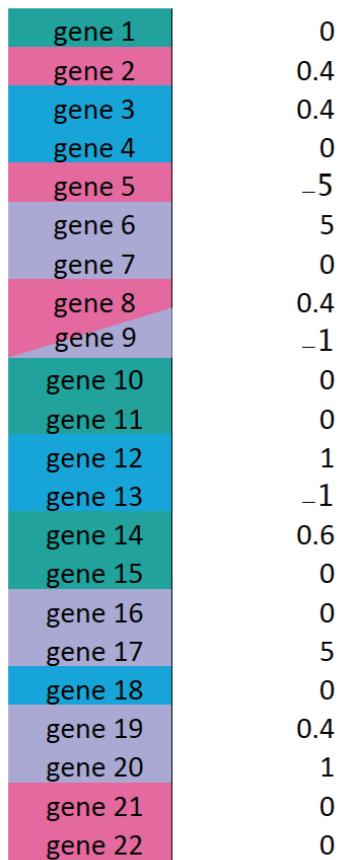
$$2/7 =$$

$$0.29$$

$$3/15 =$$

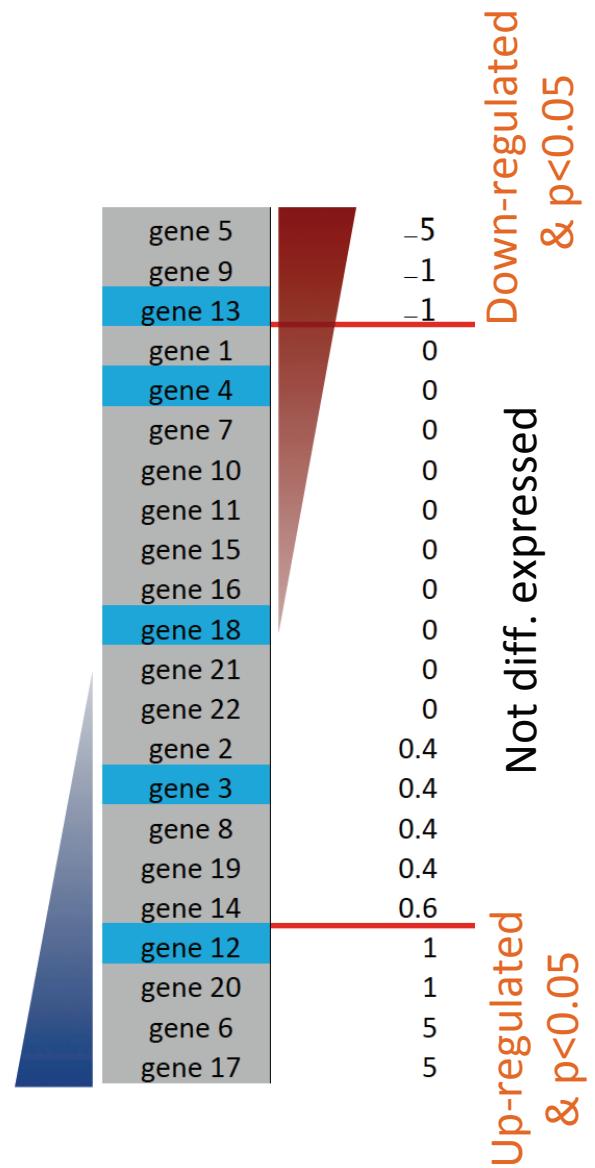
$$0.20$$

Which gene sets are differentially expressed?



Run individual Fisher's exact tests for each gene set, **blue**, **pink**, **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>

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

Functions for Fisher test and for ORA 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)
```

enricher(): implementation of hypergeometric test (one-sided Fisher 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

RStudio tour

The screenshot shows the RStudio IDE interface. The top menu bar includes File, Edit, Code, View, Plots, Session, Build, Debug, Profile, Tools, and Help. The status bar indicates R 4.0.0.

In the code editor pane, two files are open: `ea_2020_script.R` and `EA_2020_Exercise_1.R`. The `ea_2020_script.R` file contains the following R code:

```
1 ##### Enrichment analysis course, SIB, June 26th 2020
2
3
4 # Enrichment analysis, SIB course, June 26th 2020
5
6 # load the packages needed for the R exercise
7 library(clusterProfiler)
8 library(org.Hs.eg.db)
9 library(pathview)
10 # library(biomart)
11
12 # Some reminders about the usage of R:
13
```

The `EA_2020_Exercise_1.R` file is currently active. The R Script dropdown in the bottom right corner is set to "R Script".

The Environment pane shows the Global Environment with objects like `convert_ens_sy...`, `GO_NK_Th`, and `NK_vs_Th`.

The Files pane displays the contents of the project folder `/cloud/project/`:

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

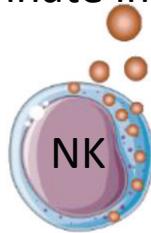
The Console pane shows the R startup message and a prompt:

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

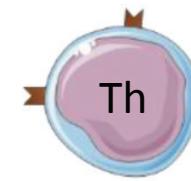
Type 'demo()' for some demos, 'help()' for on-line help, or
'help.start()' for an HTML browser interface to help.
Type 'q()' to quit R.

[Workspace loaded from /cloud/project/.RData]
> |
```

Innate immunity



Adaptive immunity



Recap and exercise 1

<https://www.mdpi.com/1420-3049/24/24/4530/html>

- 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.

RNA sequencing data from:

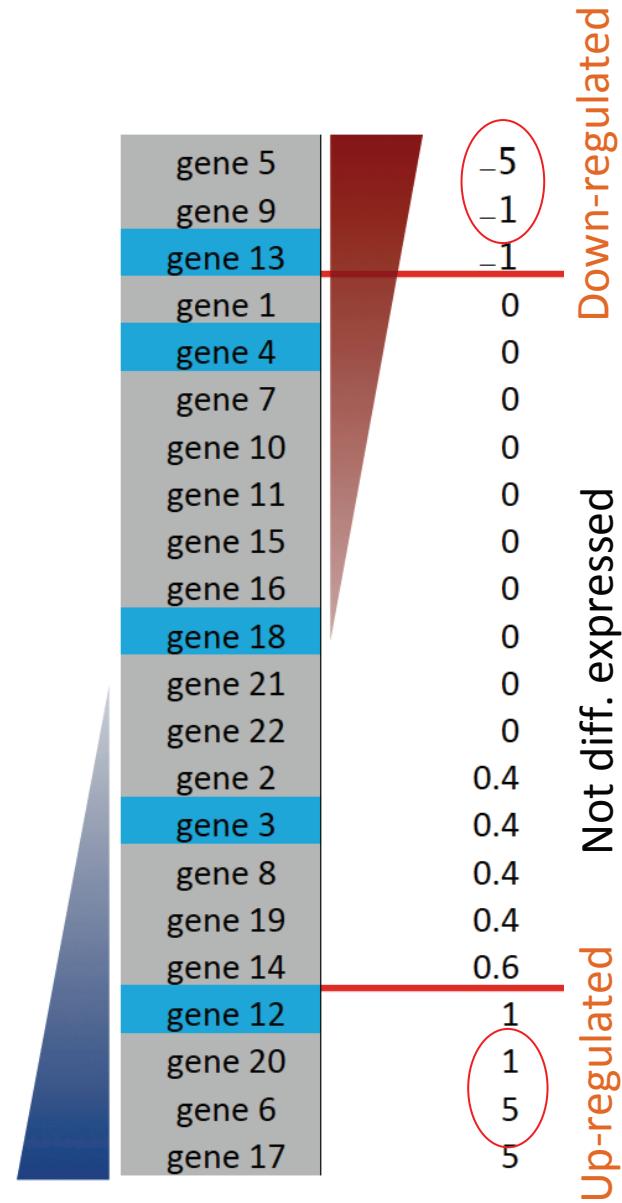
<https://jlb.onlinelibrary.wiley.com/doi/full/10.1002/JLB.5MA0120-209R?af=R>

<https://ashpublications.org/bloodadvances/article/3/22/3674/428873/CD56-as-a-marker-of-an-ILC1-like-population-with>

Fisher's exact test is threshold-based

2x2 count table	Differentially expressed	Not Differentially expressed	total
blue	2	3	5
Not blue	5	12	17
total	7	15	22

Contingency table with count of genes,
without taking into account the **magnitude**
of the change of each gene.

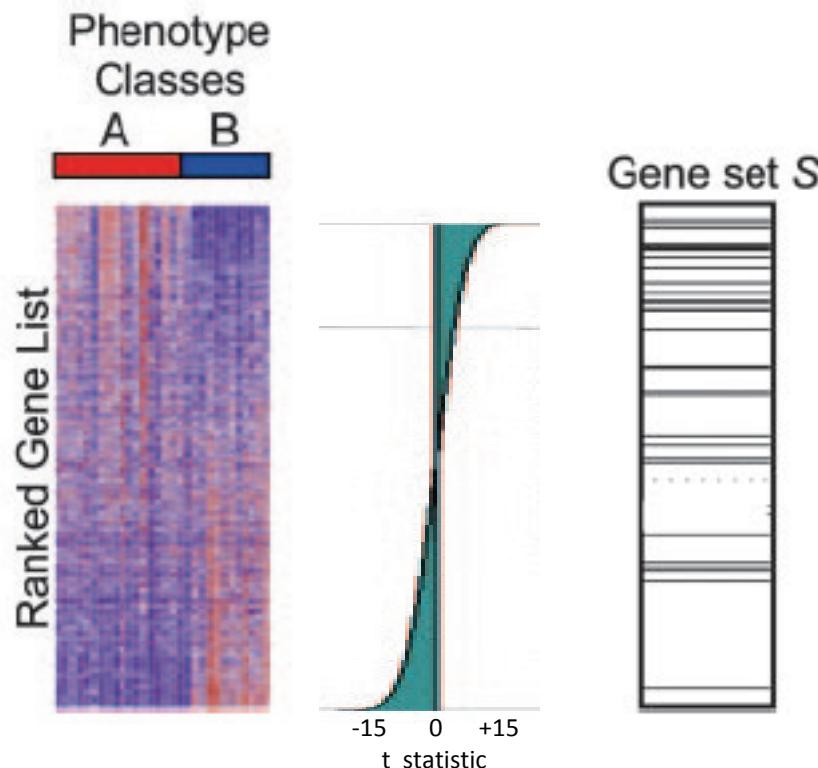


Gene set enrichment analysis (GSEA)

- **Threshold-free:** the whole list of genes detected in the RNA sequencing experiment 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.

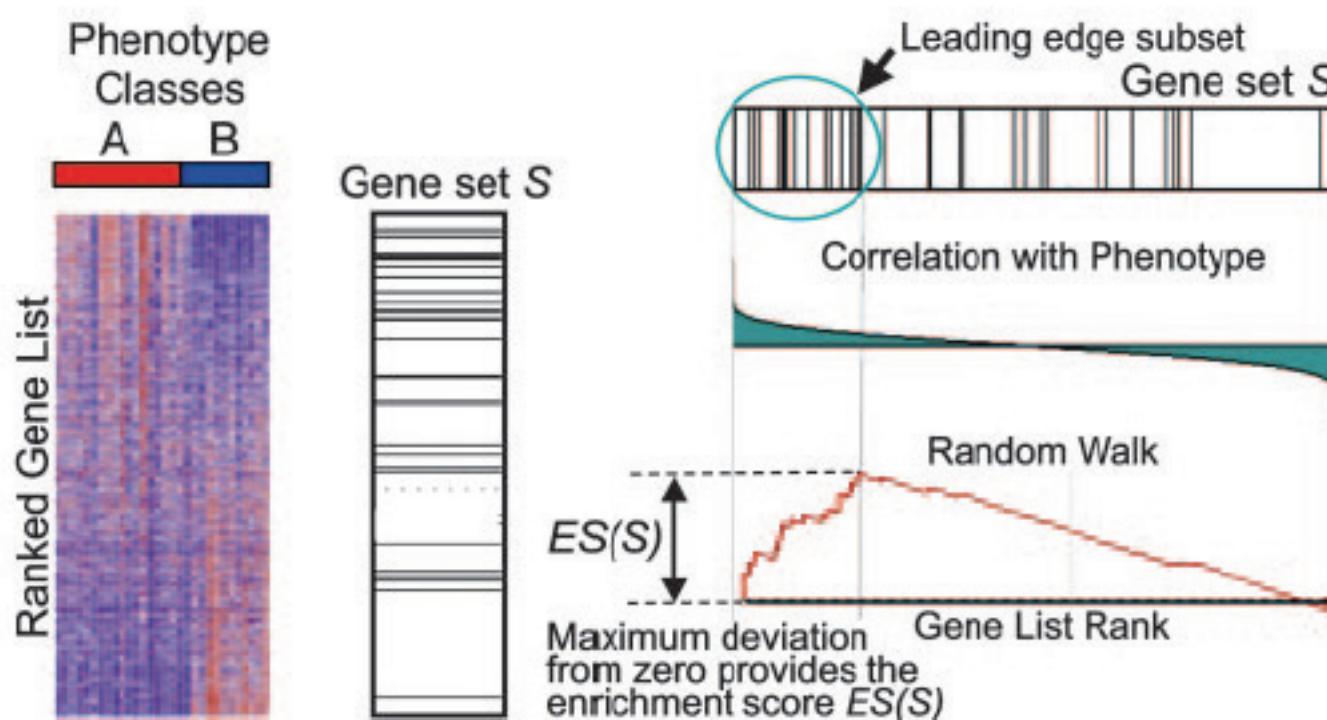
Method of GSEA

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)

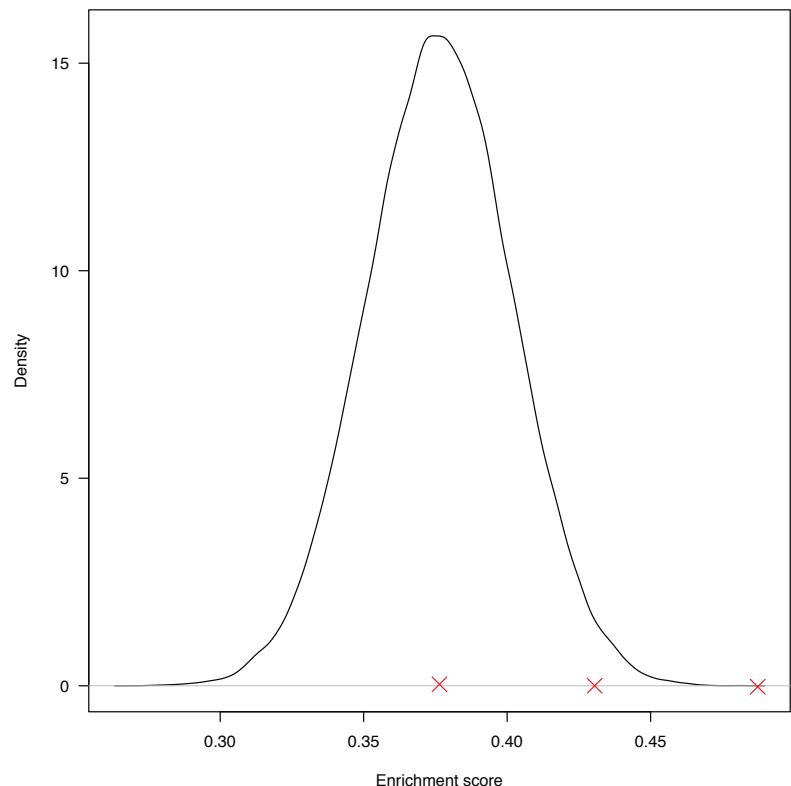
Method of GSEA



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



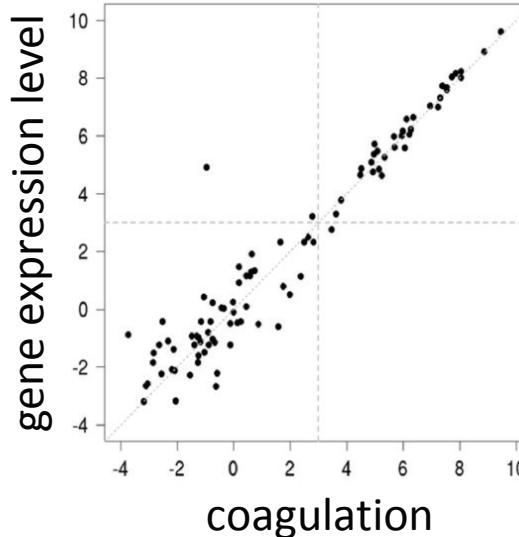
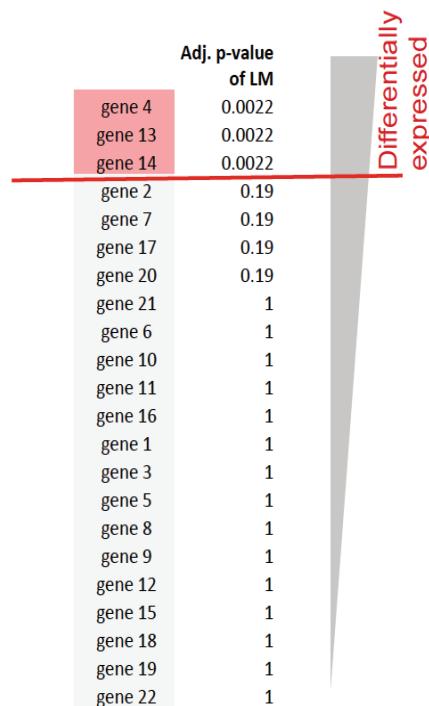
$$\text{NES} = \frac{\text{actual ES}}{\text{mean(ESs against all permutations of the dataset)}}$$

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

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 coefficients or p-value of linear model



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

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",  
  ...  
)
```

TERM2GENE:

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

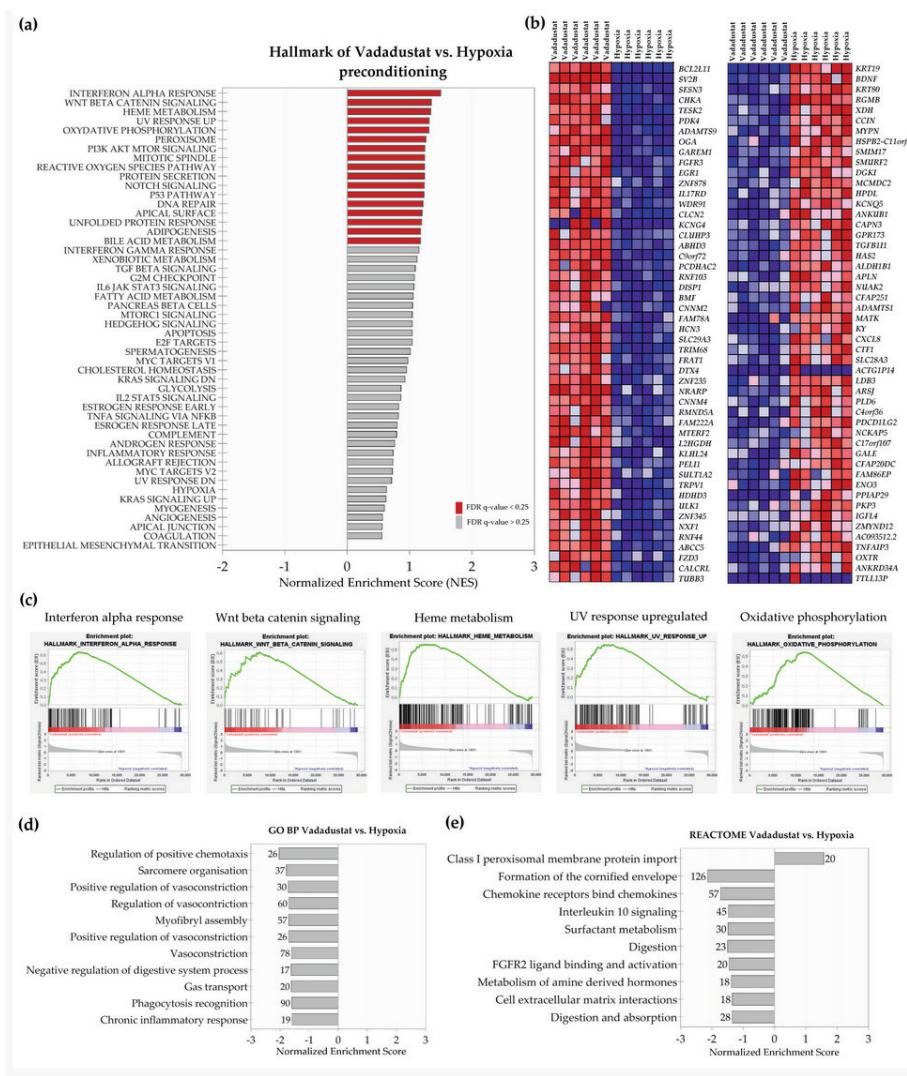
gseGO(): GSEA of GO gene sets using all ranked genes

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

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?

How to show the results of an enrichment analysis?



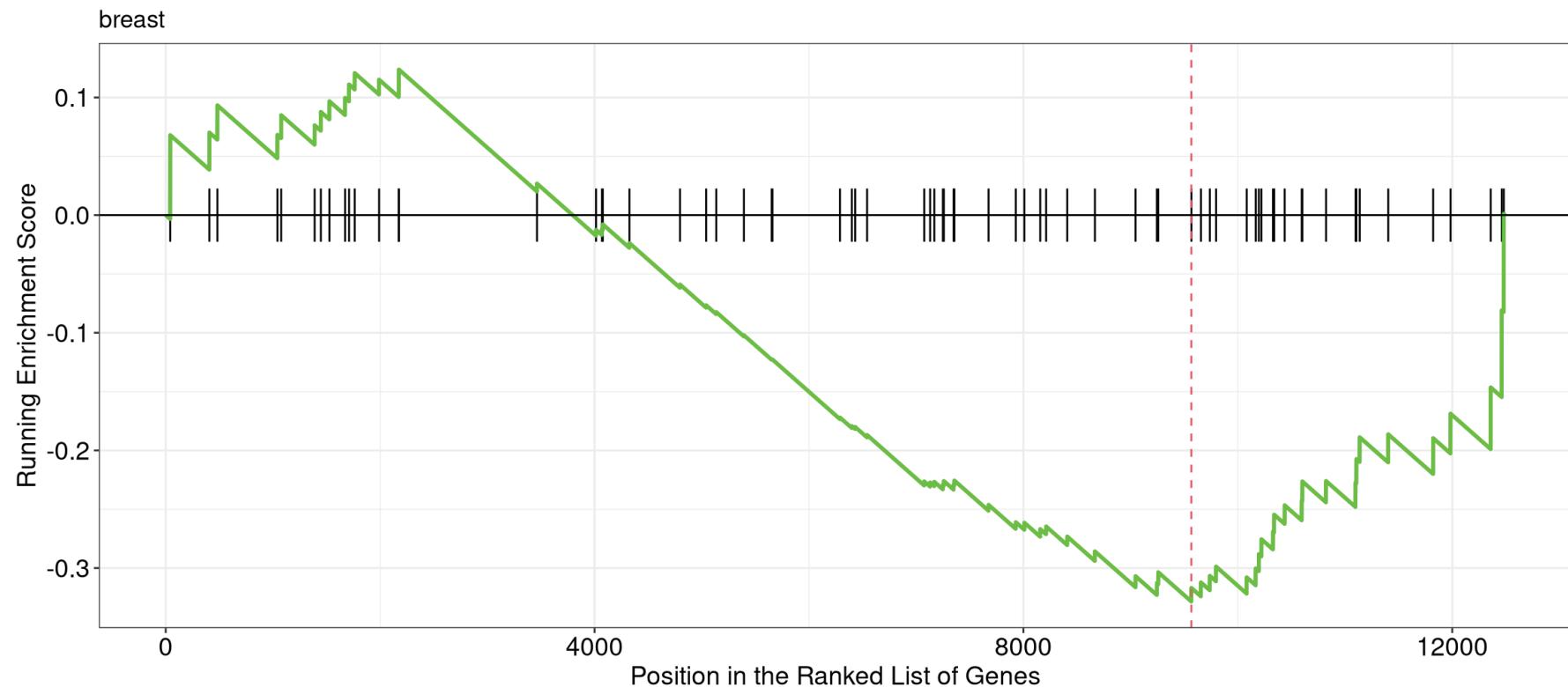
Zielniok et al 2021

Int. J. Mol. Sci. **2021**, *22*(15), 8160; <https://doi.org/10.3390/ijms22158160>

Visualizations available in clusterProfiler

- barcode plot

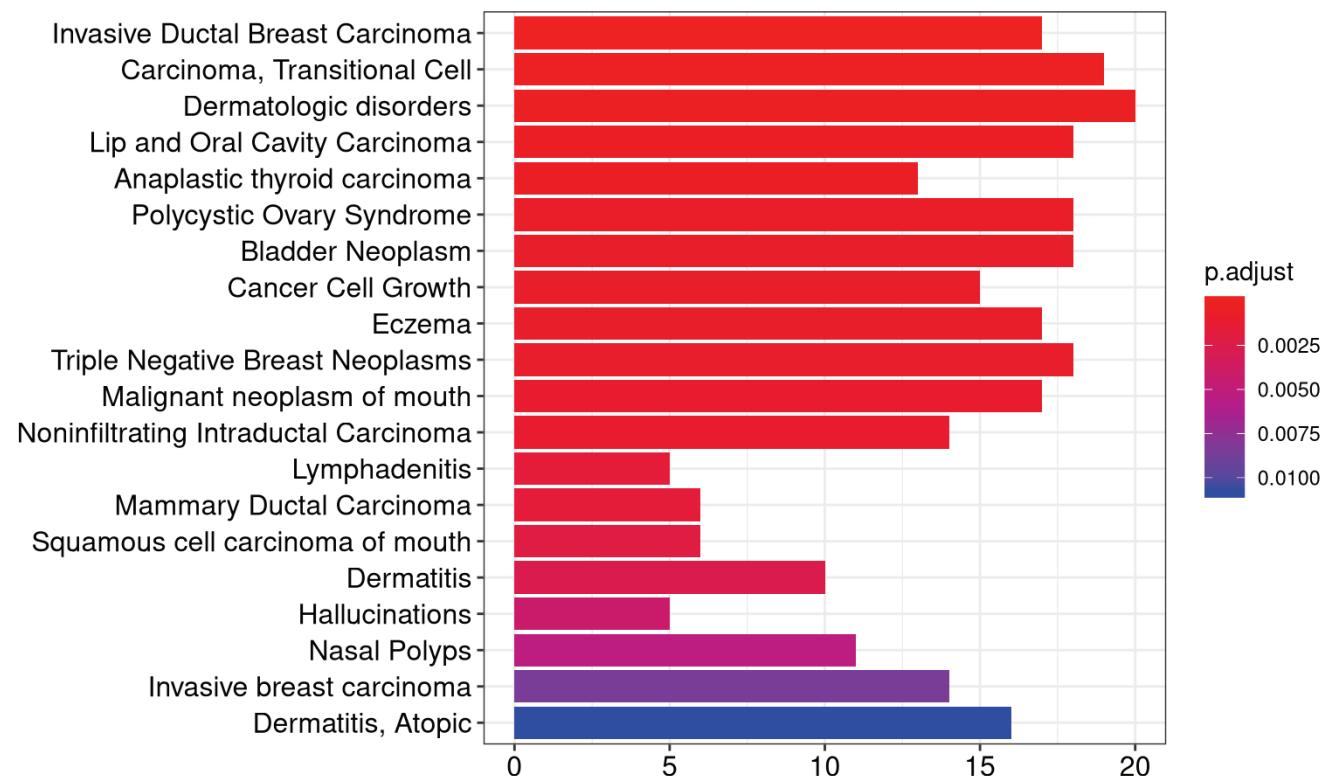
```
gseaplot(h_NK_vs_Th, geneSetID =  
"BREAST", title=" BREAST")
```



Visualizations available in clusterProfiler

- barplot

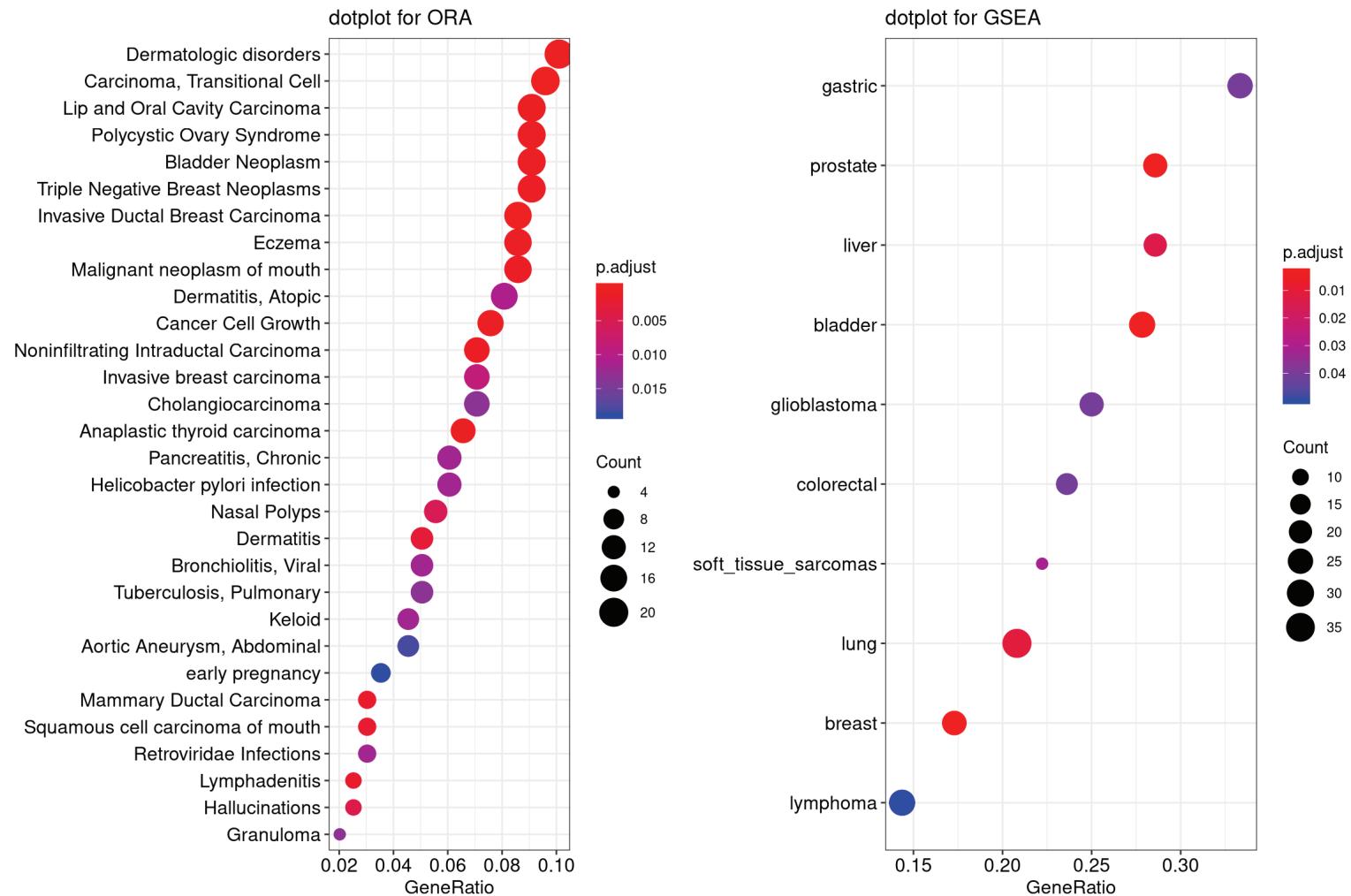
```
ego <- enrichGO(de, OrgDb='org.Hs.eg.db', ont="BP", keyType = "SYMBOL")
barplot(ego, showCategory=20)
```



Visualizations available in clusterProfiler

- dotplot

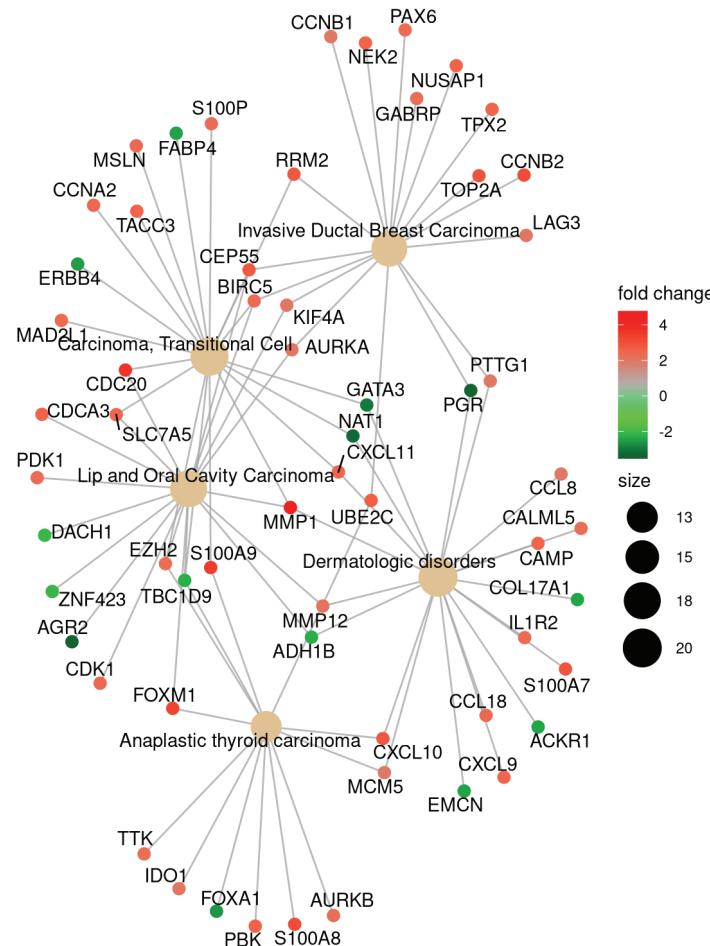
```
ego <- enrichGO(de)
dotplot(ego, showCategory=20)
```



Visualizations available in clusterProfiler

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

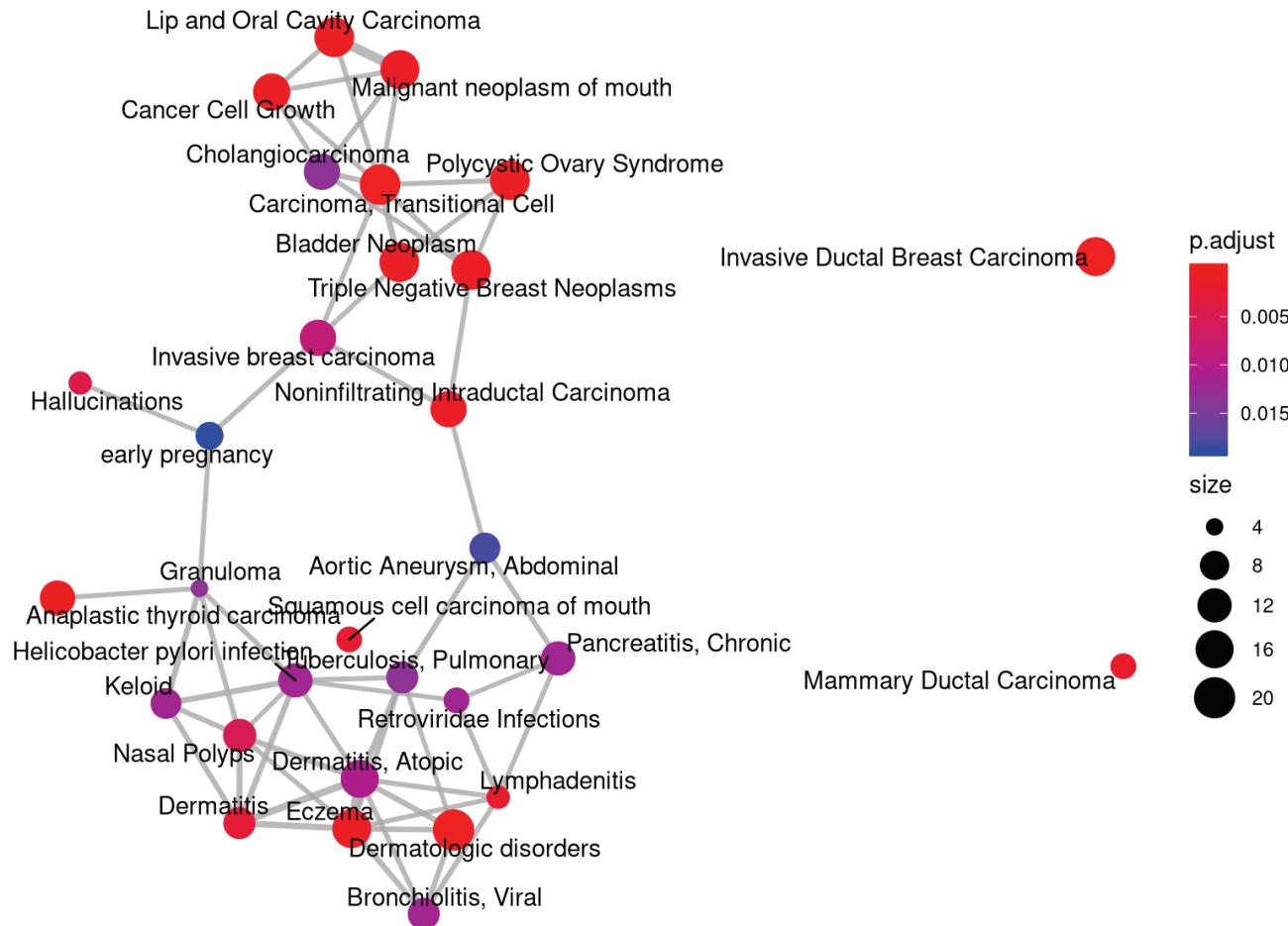
- Gene-concept network



Visualizations available in clusterProfiler

- Enrichment map

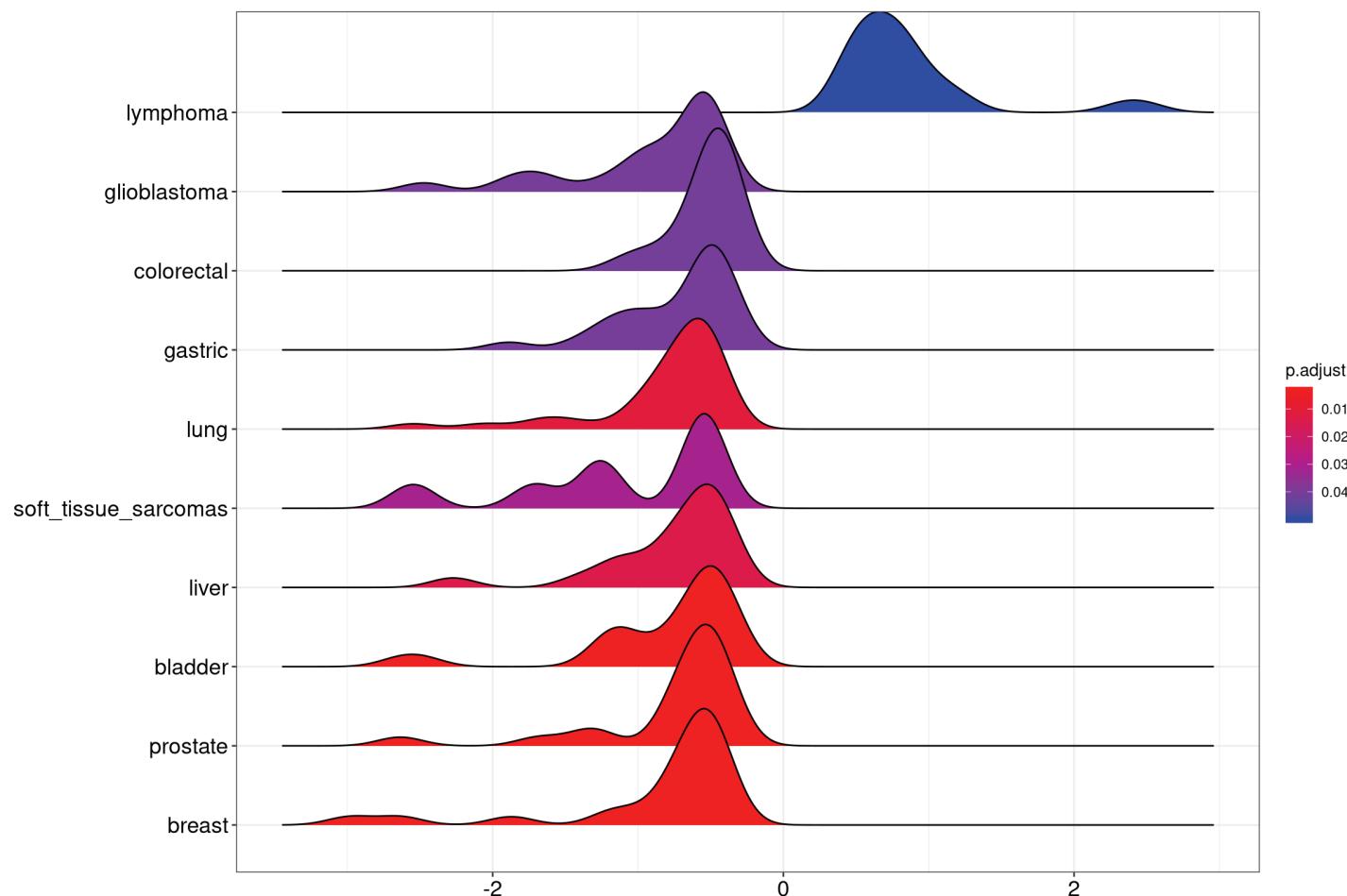
```
ego <- enrichGO(de)
emapplot(ego)
```



Visualizations available in clusterProfiler

- Ridgeplot

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



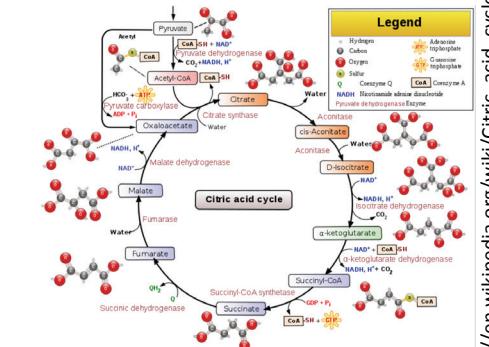
Recap and Exercise 3

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 $-\log_{10}(p\text{-value})$ 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?



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

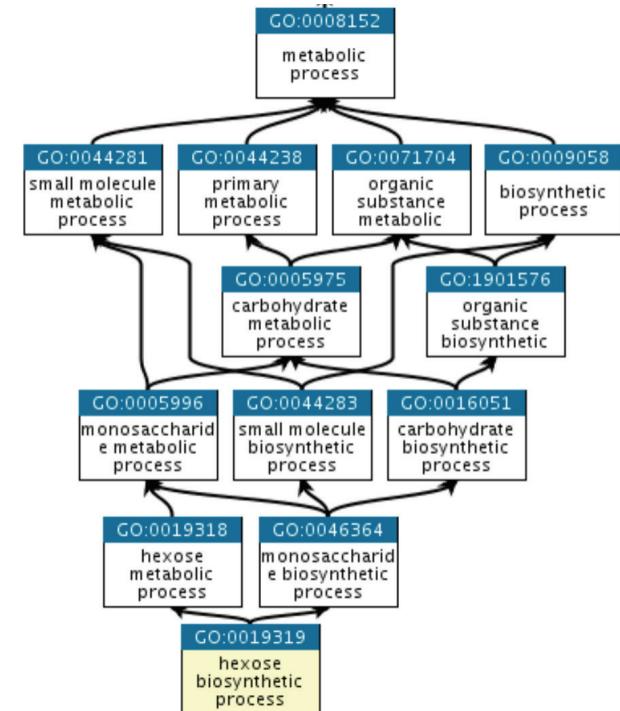
Gene ontology

- <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.



Not covered today: SetRank (cran), GOSemSim (bioconductor), Revigo (<http://revigo.irb.hr/>)

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>

Bi-directional eg mTOR signaling

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

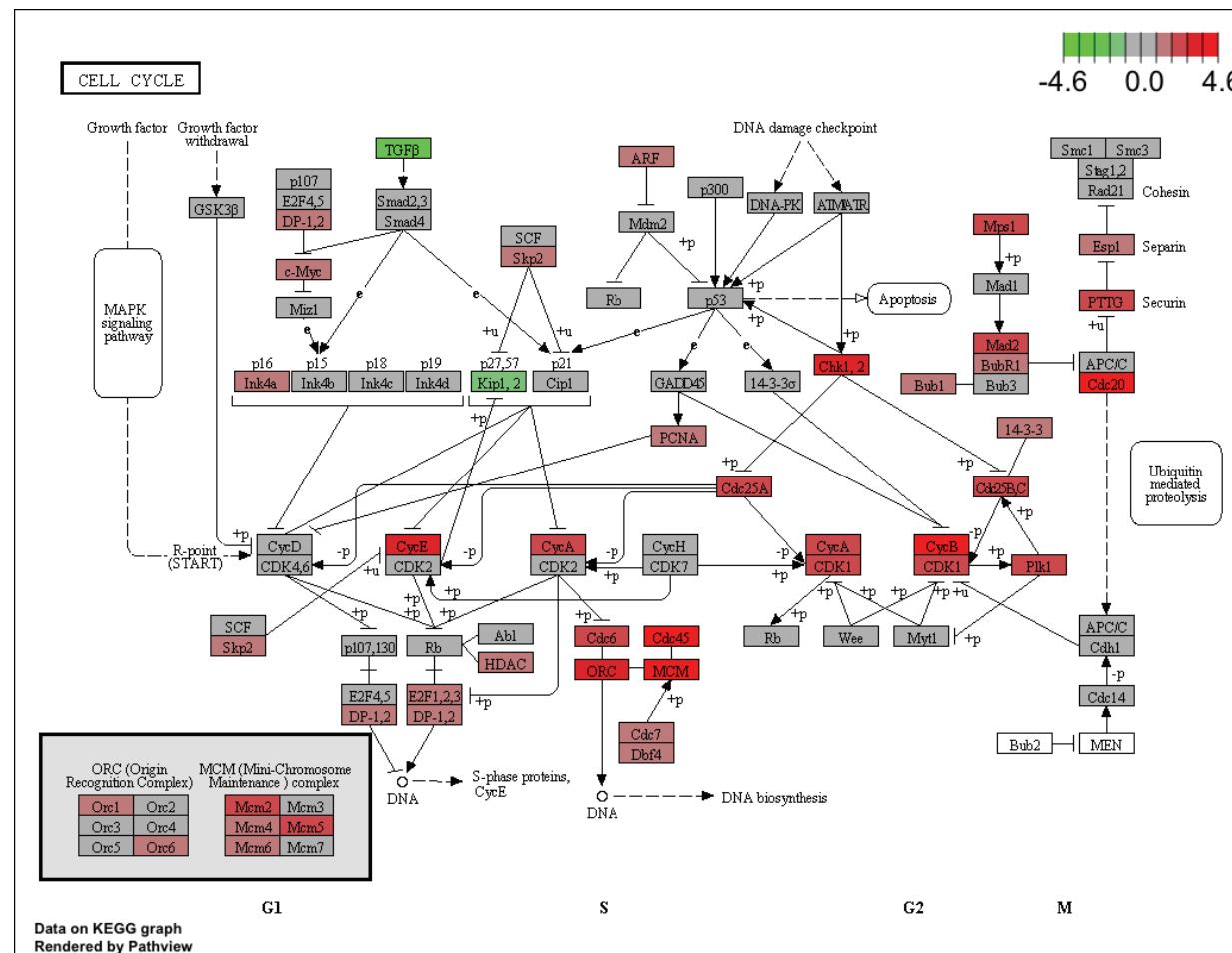
6. Human Diseases

7. Drug Development

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

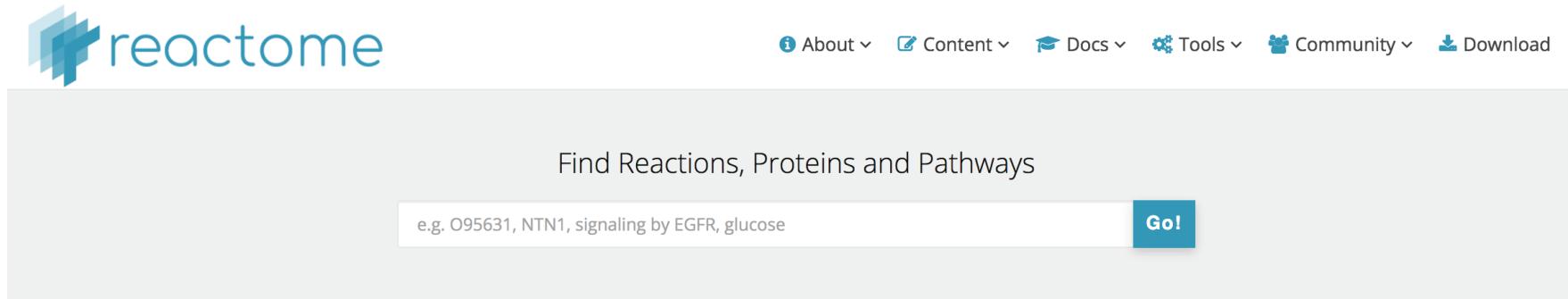
Visualizations available - pathview package

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

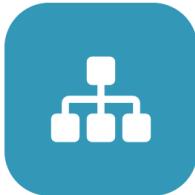
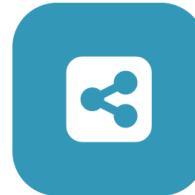


Reactome

<https://reactome.org/>



The screenshot shows the Reactome website homepage. At the top left is the Reactome logo. To its right is a navigation bar with links: About, Content, Docs, Tools, Community, and Download. Below the navigation bar is a search bar with the placeholder text "Find Reactions, Proteins and Pathways" and a "Go!" button. Underneath the search bar is a search input field containing the text "e.g. O95631, NTN1, signaling by EGFR, glucose". Below the search area are four large blue icons with white symbols: a tree-like structure for the Pathway Browser, a bar chart for Analysis Tools, a network graph for ReactomeFIViz, and a document for Documentation. Each icon has a corresponding title and a brief description below it.

			
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:

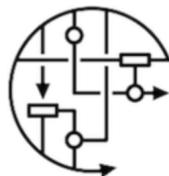
<https://www.gsea-msigdb.org/gsea/downloads.jsp>

The Hallmark collection:

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

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
- Frit 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

You can search by:

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

Browse

Browse by species and category

Get Pathways

Download

Growth

Pathways



Today's Featured Pathway

COVID-19, thrombosis and anticoagulation (*Homo sapiens*)

COVID-19, thrombosis and anticoagulation

Curator of the Week



Kristina Hanspers (Gladstone Institutes)

Bioconductor

- **GO.db** : A set of annotation maps describing the entire Gene Ontology assembled using data from GO
- **gskb: mouse**
 - mm_GO: gene sets from Gene Ontology for mouse (*Mus musculus*)
 - mm_location: Gene sets based on chromosomal location
 - mm_metabolic: metabolic pathways
 - mm_miRNA: Target genes of microRNAs, predicted or experimentally verified
 - mm_pathway: Curated pathways
 - mm_TF: Transcription factor target genes.
 - mm_other
- **KEGG.db**: A set of annotation maps for KEGG assembled using data from KEGG

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 (or biomaRt package)

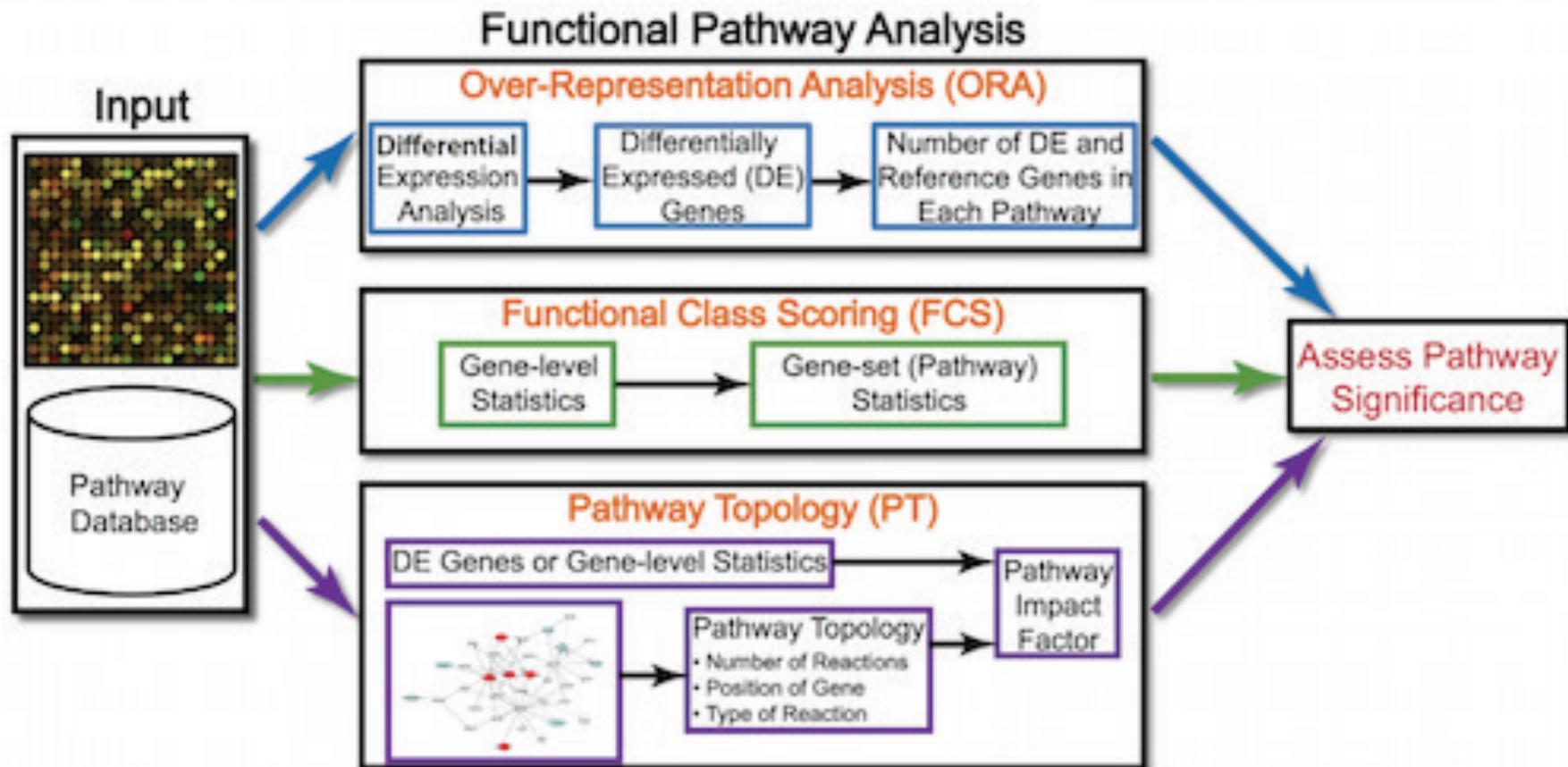
```
bitr(geneID, fromType, toType, OrgDb, drop = TRUE)
```

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 to perform a GSEA of KEGG pathways (with argument minGSSize=30).
 - Explore the results. Is there a KEGG immune-related gene set coming up? Is there a KEGG Natural killer gene set coming up?
 - Import the hallmark gene sets and run a GSEA. How many significant gene sets are there?

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	0.000000	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

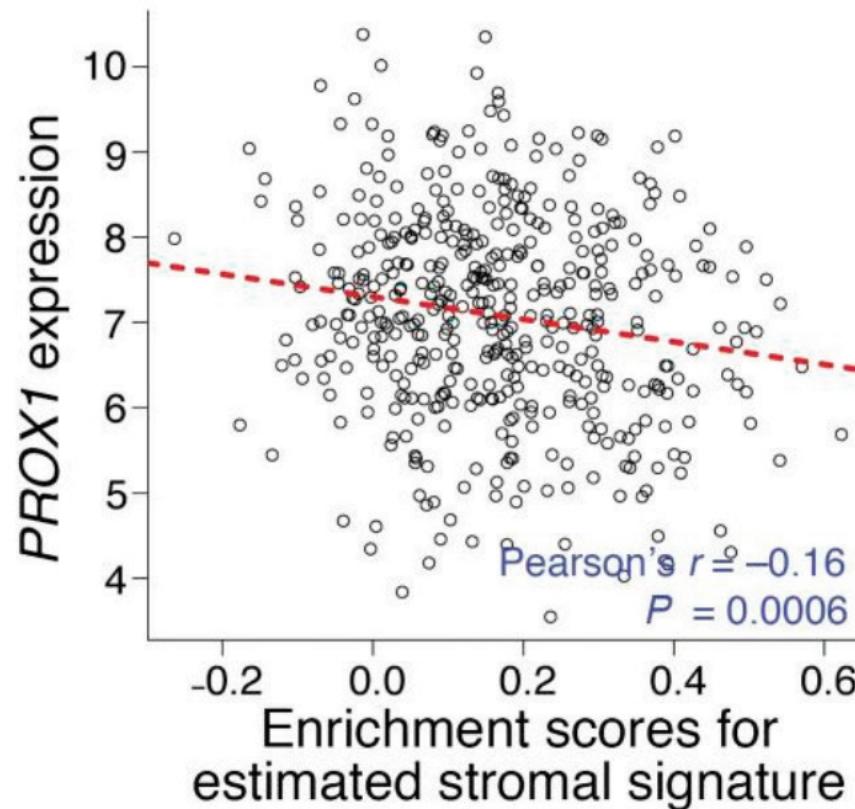
P_G: combination of P_{NDE} and P_{PERT}

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](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 Monique Zahn)
- Send answers to tania.wyss@sib.swiss by July 1st 2022, 11:59pm

Thank you for your attention!

Please fill in the **feedback** sent by Monique Zahn.

We thank Isabelle Dupanloup and Linda Dib for providing course material