

Swiss Institute
of
Bioinformatics

Enrichment analysis

December 3rd 2021

isabelle.dupanloup@sib.swiss

tania.wyss@sib.swiss

The Bioinformatics Core Facility at SIB



Home
People
Research
Projects
Publications
Services
Teaching
Resources
Partners
Contact

Welcome to **BCF-SIB**



About **BCF-SIB**

The Bioinformatics Core Facility (BCF) is a research and service group within the [SIB Swiss Institute of Bioinformatics](#). 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 microarrays, 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



Swiss Institute of
Bioinformatics

Let's collaborate

Careers Contact Directory Intranet



Research infrastructure ▾

Scientific community ▾

About SIB ▾



Home

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/
1XAmufwECKlEHibPnYcIQSYboADfowK1KG2RB
c3RcBUo/edit#heading=h.5xrppxpatnym](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

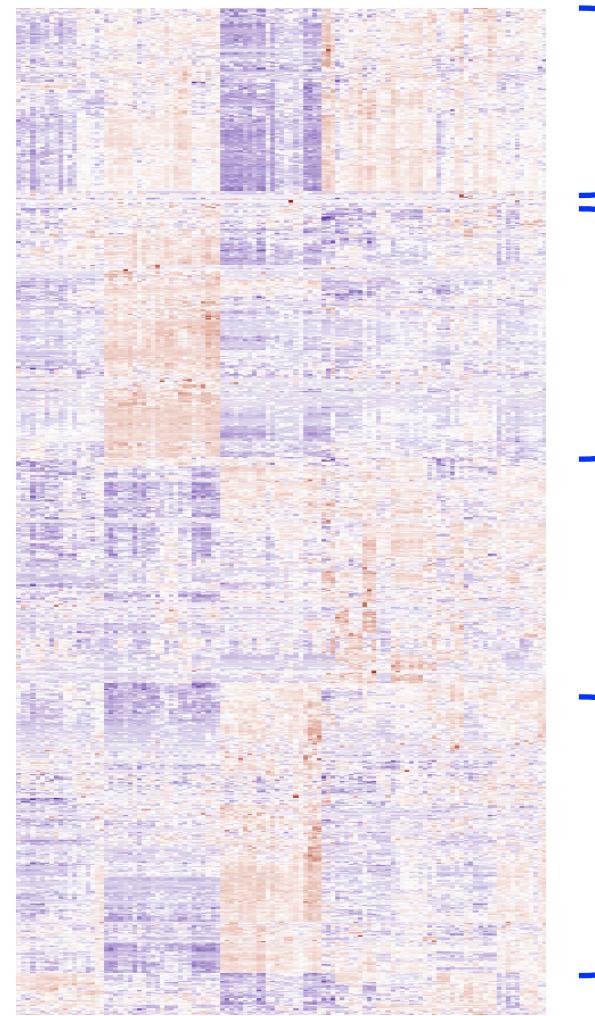


Course material

- Moodle:
- <https://edu.isb-sib.ch/course/view.php?id=521>
- 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/
1XAmufwECkIEHibPnYcIQSYboADfowK1KG2RBc3RcBUo/edit#heading=h.
5xrppxpatnym](https://docs.google.com/document/d/1XAmufwECkIEHibPnYcIQSYboADfowK1KG2RBc3RcBUo/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

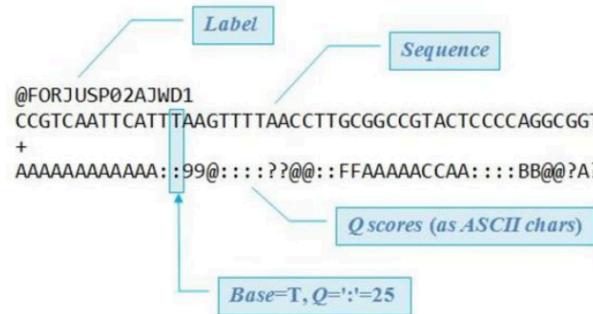


IVY GAP: <https://glioblastoma.alleninstitute.org/>

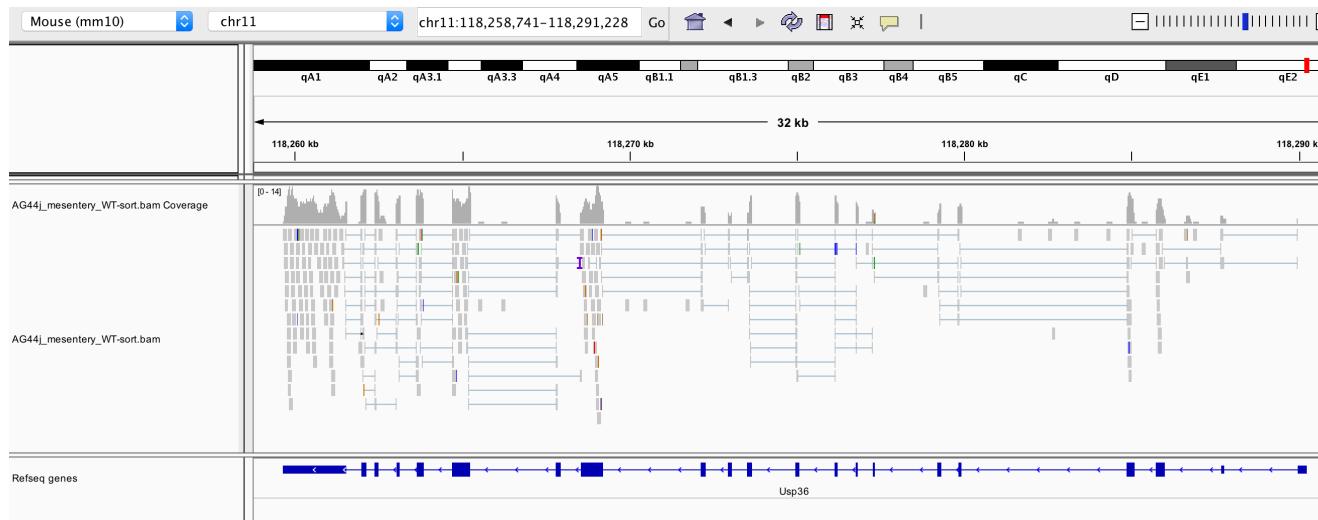
Typical RNA sequencing analysis workflow

fastq file:

```
@HWI-M01141:63:A4NDL:1:1101:14849:1418 1:N:0:TATAGCGAGACACCGT  
NACGAAGGGTCAAGCGTTACTCGGAATTACTGGCGTAAGCGTGCCTAGGTGGTTT/  
+  
#>>>A??FAA1BGGEGGAAFGCA@BFF1D2BCF/EEG/DBEE/E?GAEEFGAEAFGJ  
@HWI-M01141:63:A4NDL:1:1101:13802:1421 1:N:0:TATAGCGAGACACCGT  
NACGGAGGGTCAAGCGTTAACCGAATTACTGGCGTAAGCGCACGCAGCGGTGTTT/  
+  
#>>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

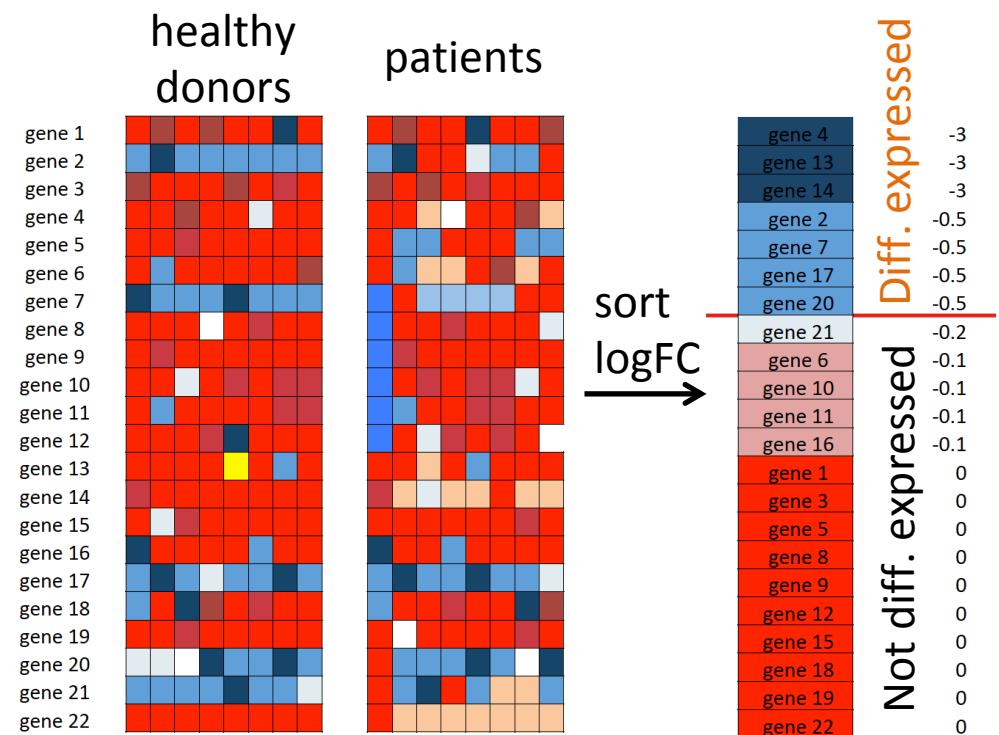
- Comparing 2 groups:

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?

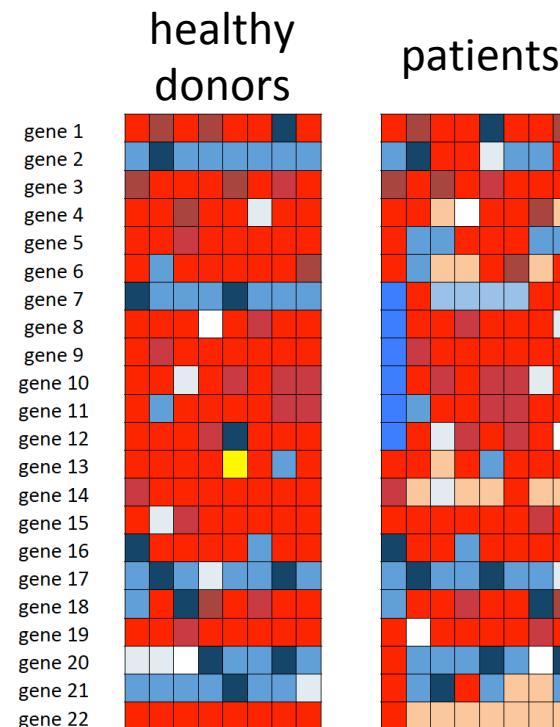
- T-test:

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:

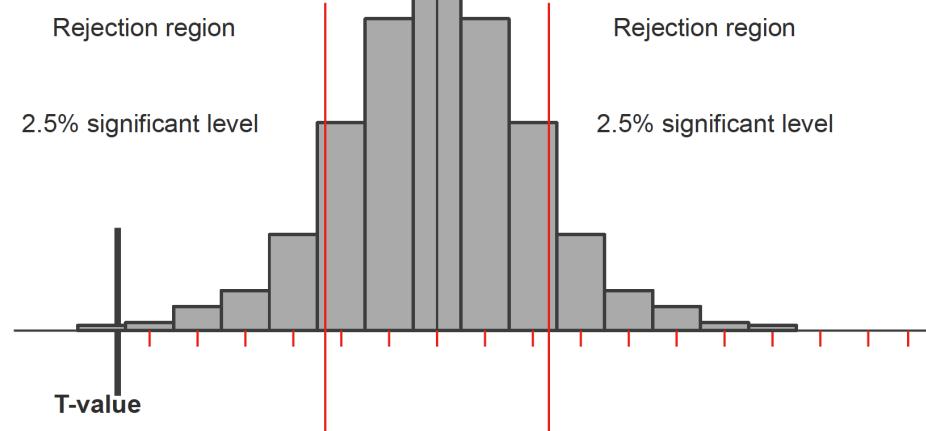
-8.908753 -4.234104

sample estimates:

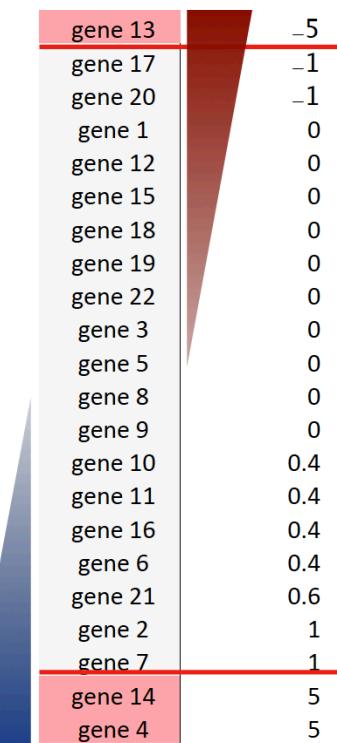
mean of x mean of y

6.00000 12.57143

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



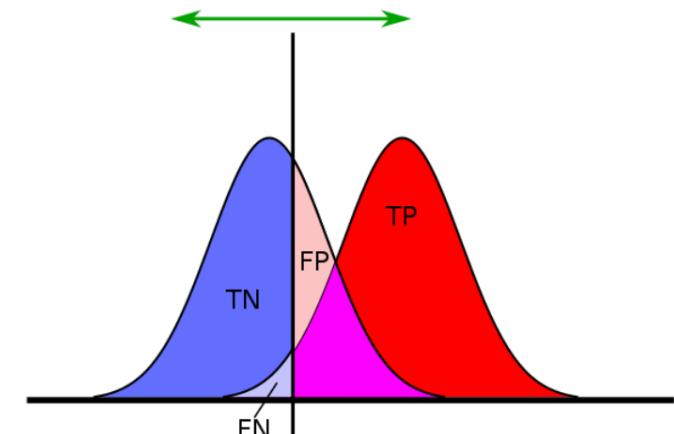
sort based
on T-statistic



What does $p < 0.05$ mean?

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



P-value adjustment: what is it?

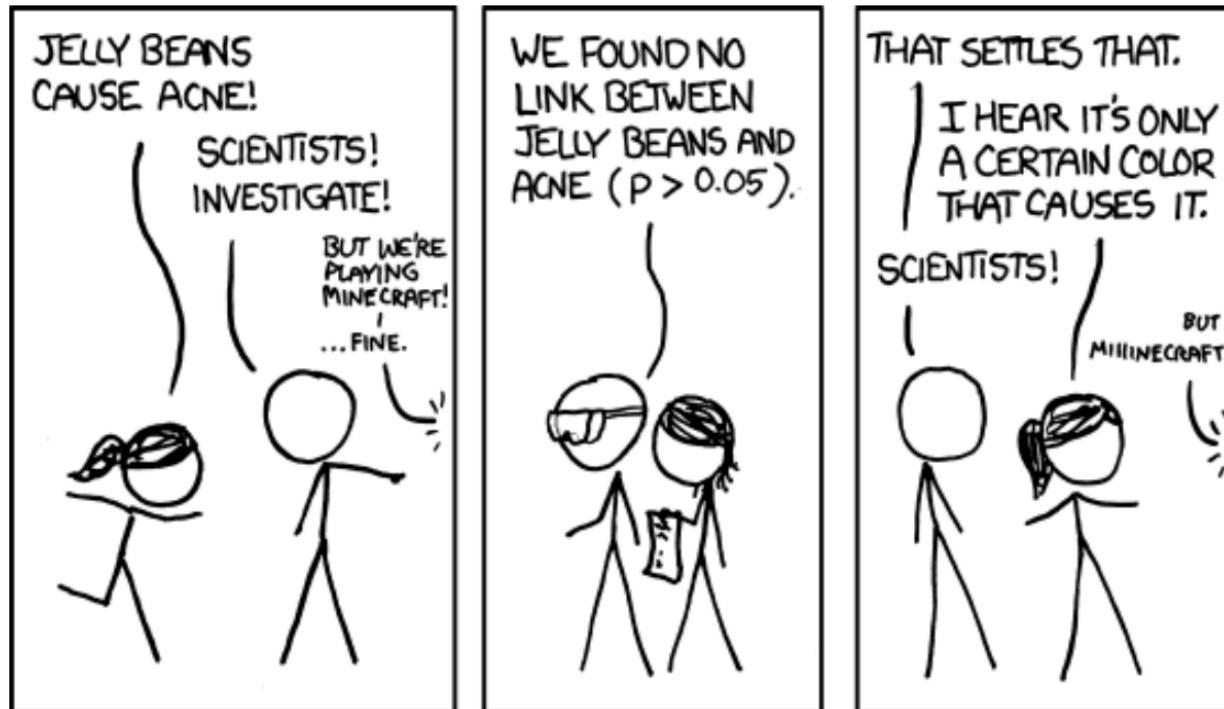


Photo by Patrick Fore on Unsplash

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
($P > 0.05$).

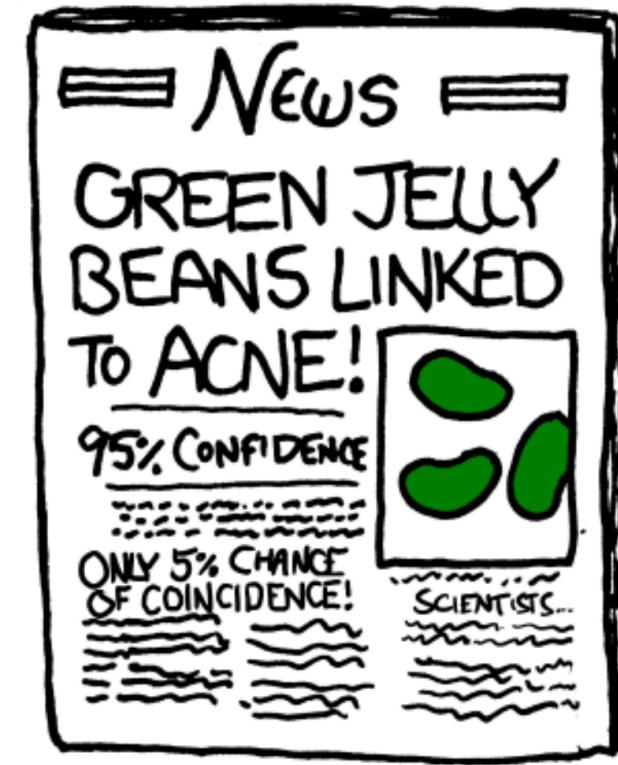

WE FOUND NO
LINK BETWEEN
BEIGE JELLY
BEANS AND ACNE
($P > 0.05$).


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 to control FDR)**
- Rank the p-values from smallest to largest, adjust less and less as the p-values get larger:

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

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

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

n = number of genes

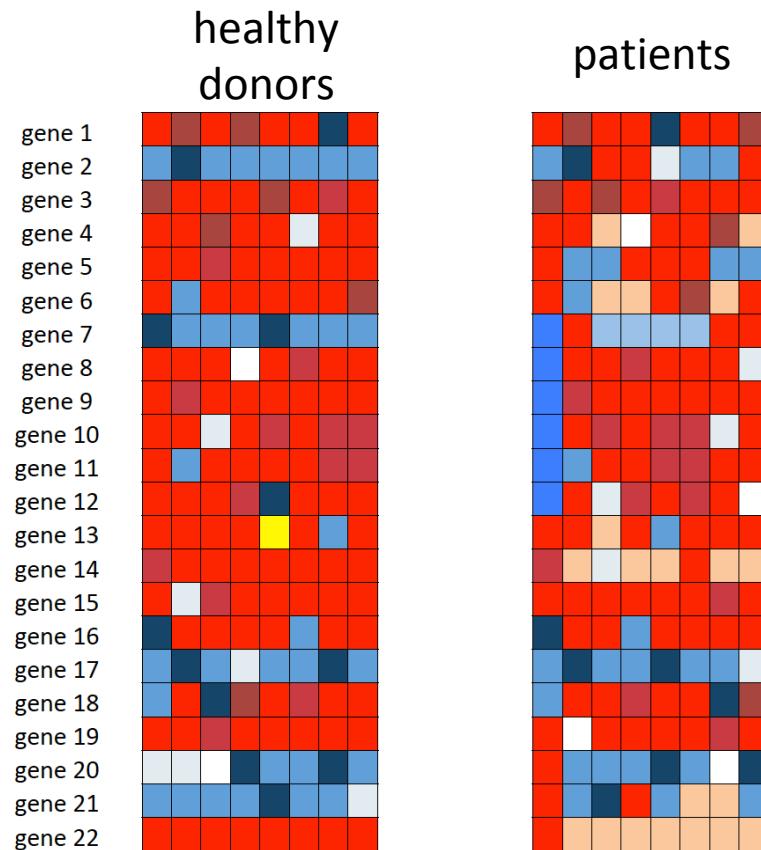
k = rank number

Differential gene expression analysis using R

- Bioconductor

<https://bioconductor.org/>

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



RStudio tour

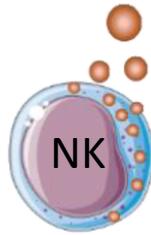
The screenshot shows the RStudio interface with the following components:

- Code Editor:** Displays an R script titled "ea_2020_script.R" with the following content:

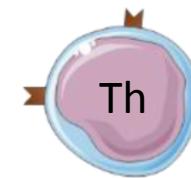
```
1 ##### Enrichment analysis course, SIB, June 26th 2020
2
3
4 # Enrichment analysis, SIB course, June 26th 2020
5
6 # load the packages needed for the R exercise
7 library(clusterProfiler)
8 library(org.Hs.eg.db)
9 library(pathview)
10 # library(biomaRt)
11
12 # Some reminders about the usage of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

- Environment Browser:** Shows the global environment with objects like "convert_ens_sy..." (15991 obs. of 3 variables), "GO_NK_Th" (Large gseaResult (98.1 Mb)), and "NK_vs_Th" (20485 obs. of 6 variables).
- File Browser:** Shows the contents of a project named "project" in the Cloud. The files listed are:

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



- 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

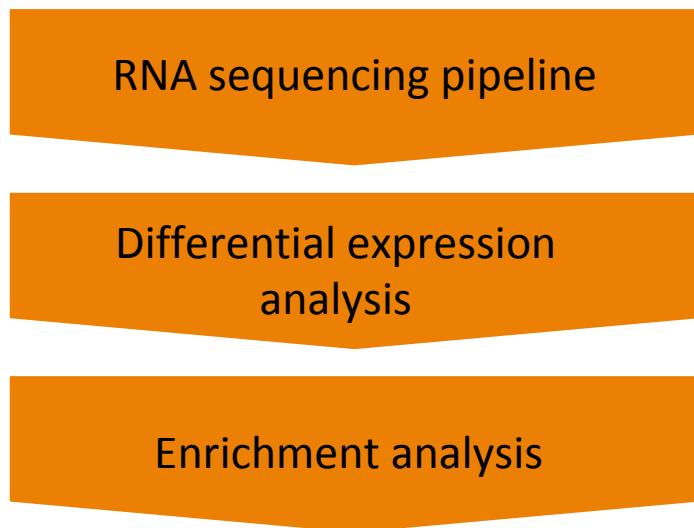
Positive logFC = higher in NK
Negative logFC = lower in NK

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?



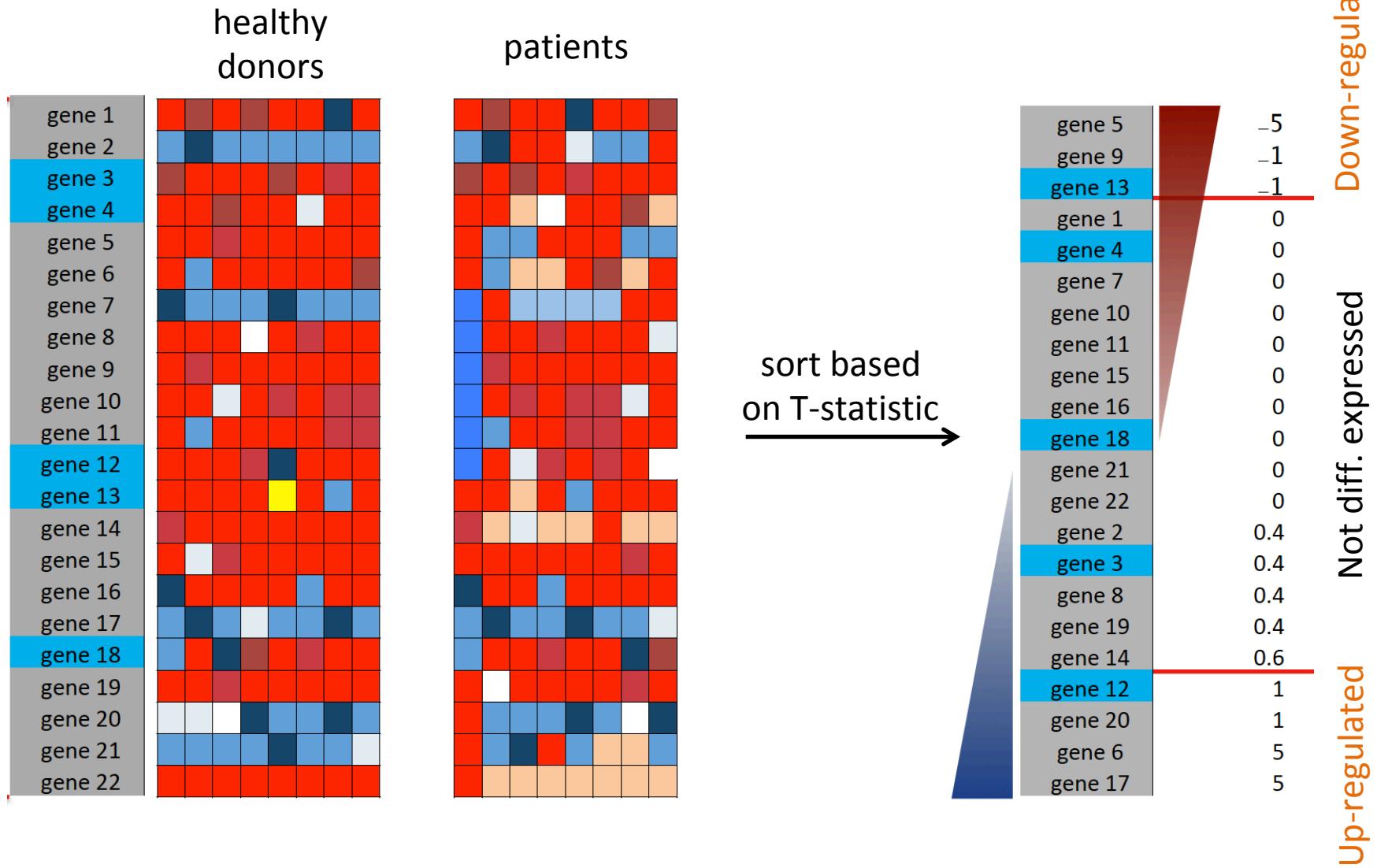
Several methods available, *e.g.*:

- 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?



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

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

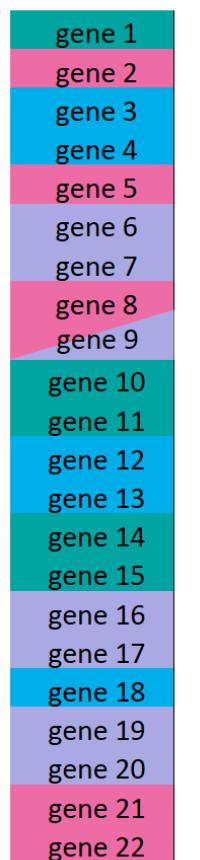
1.56456

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

$$2/7 = \\ 0.29$$

$$3/15 = \\ 0.20$$

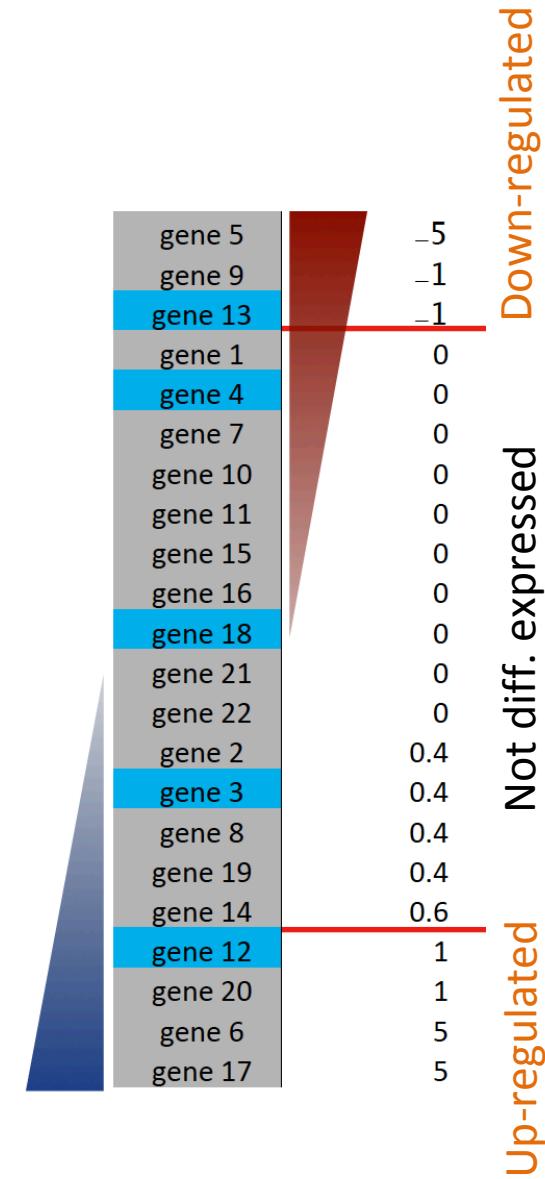
Which gene sets are differentially expressed?



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

⇒ Multiple tests need **p-value adjustment**.

⇒ But Fisher test is **threshold-based**.

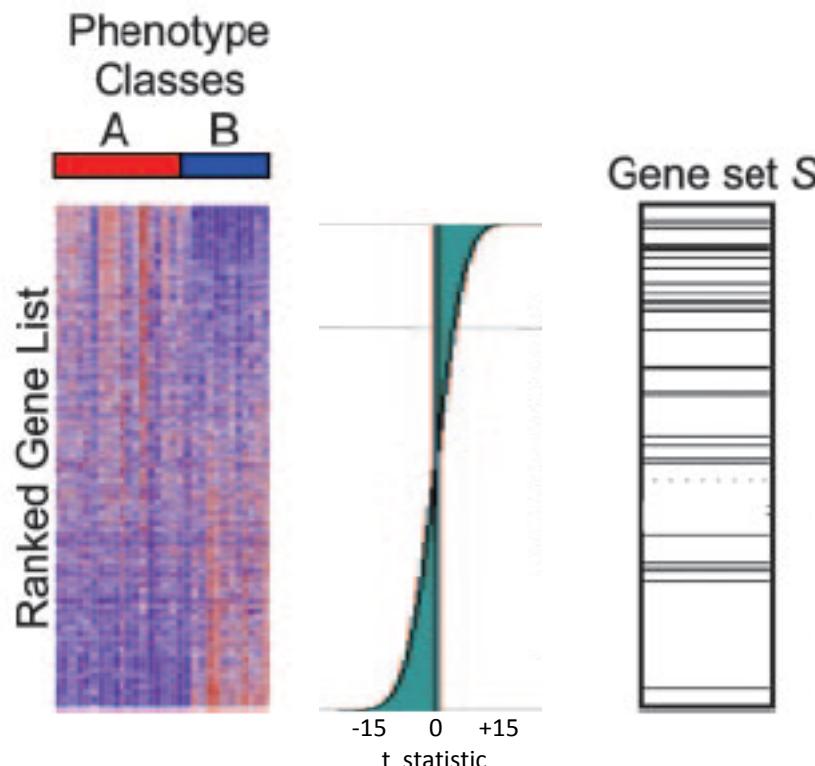


Gene set enrichment analysis (GSEA)

- 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.

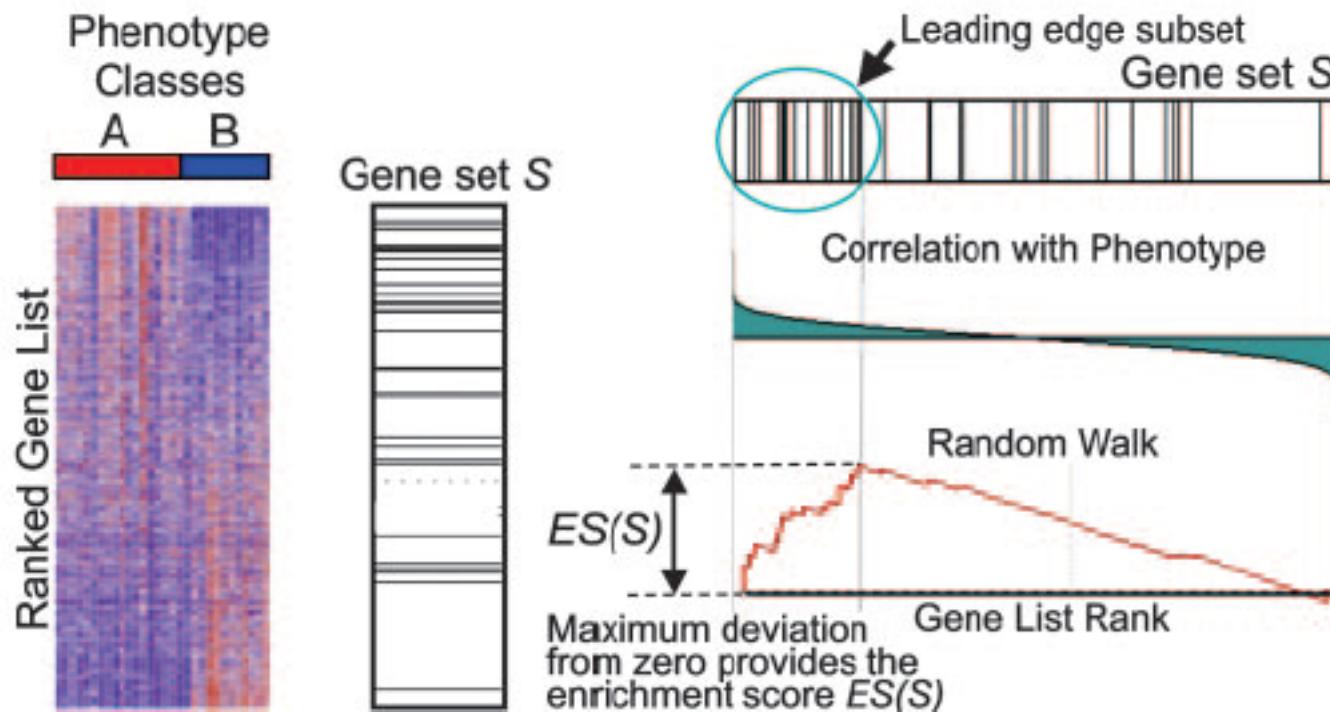
Method of GSEA

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)

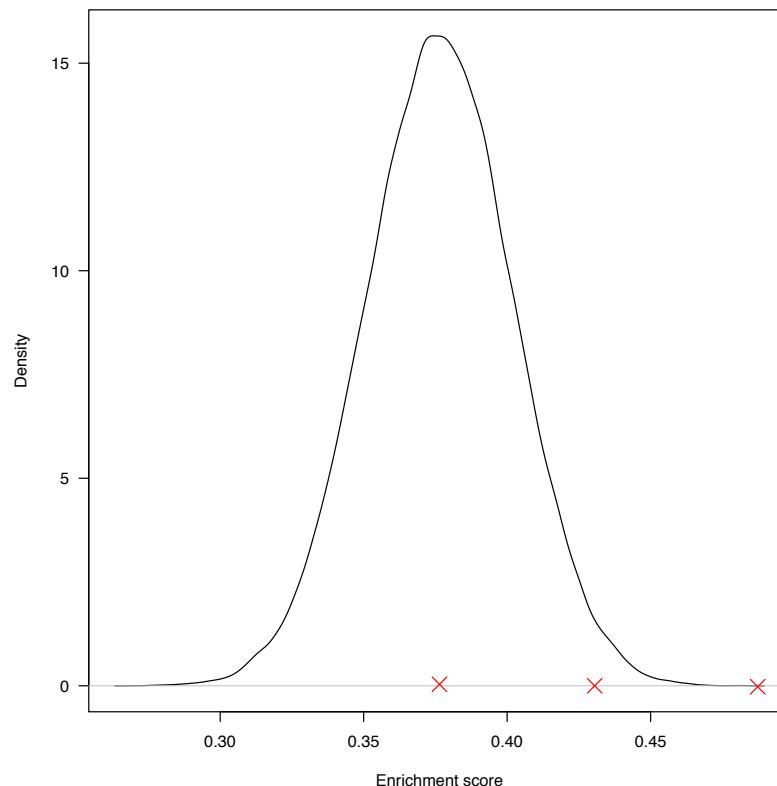
Method of GSEA



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



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

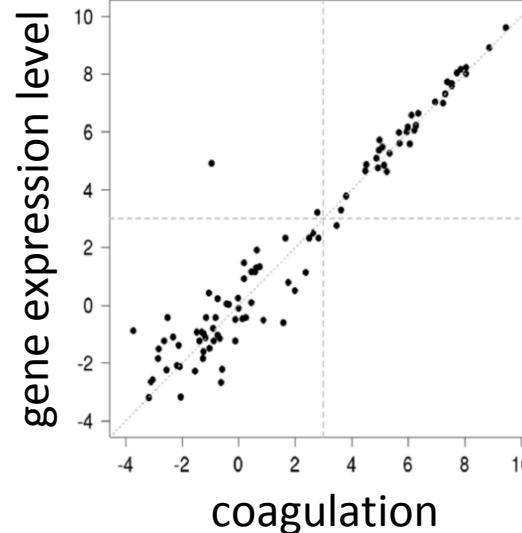
Do not forget p-value
adjustment if more than 1
gene set is tested!

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 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



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](https://doi.org/10.18129/B9.bioc.clusterProfiler) [f](#) [t](#)

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:\[10.1089/omi.2011.0118\]](https://doi.org/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?