Swiss Institute of Bioinformatics

From matrix to biological insights

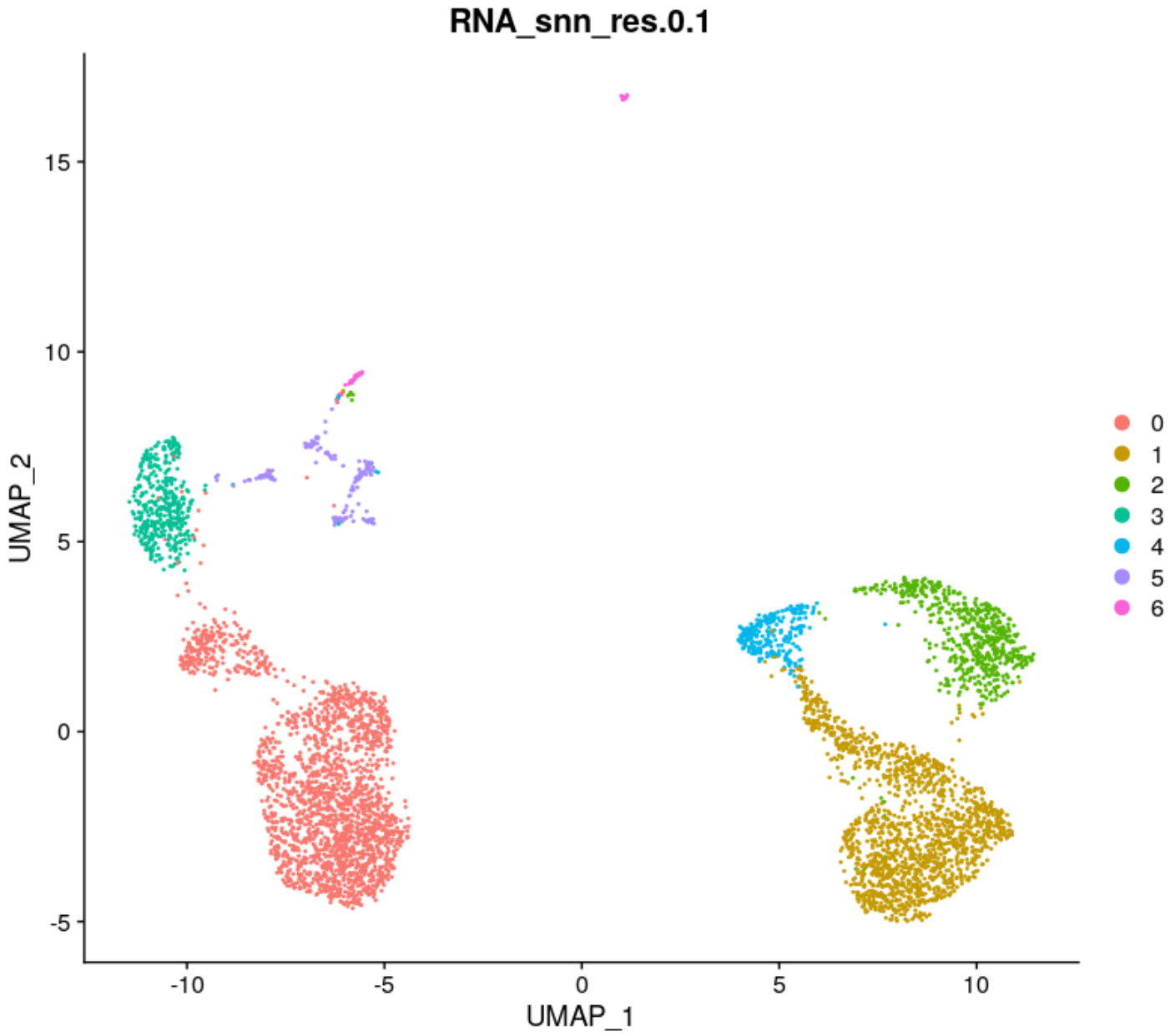
Luciano Cascione, PhD
Bioinformatics Core Unit

LUCIANO CASCIONE, PHD

BELLINZONA, Nov. 14TH 2025

elixir SWITZERLAND

Differential gene expression analysis



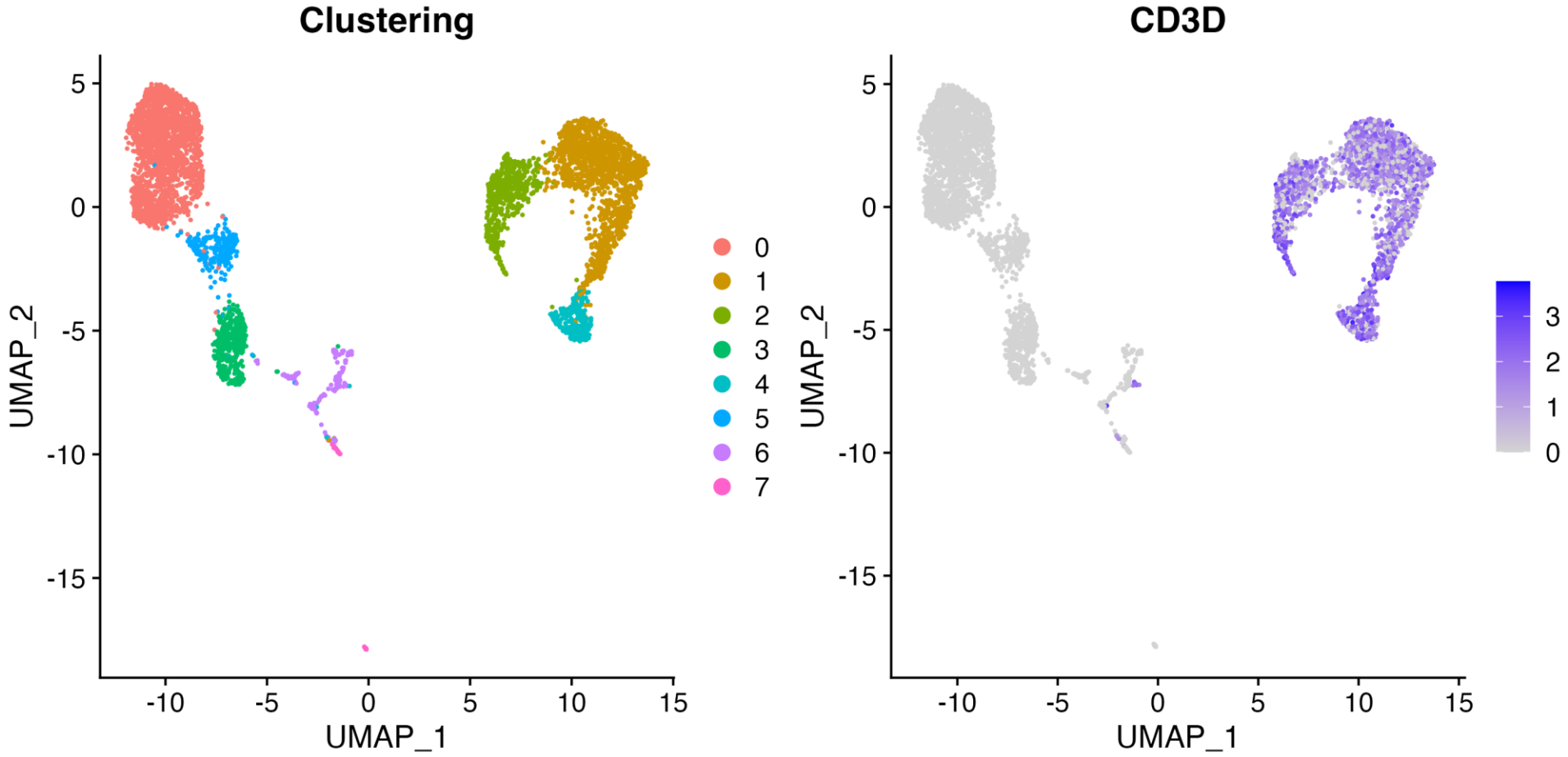
Two types of gene expression analysis

Marker gene identification: gene overexpressed by each cell type or cluster with the dataset => *It can help in cell type annotation*

Differential gene expression analysis: gene whose expression if modulated (up or down) by experimental condition within a cell type or cluster

Marker gene identification

Which genes are more (or less) expressed in one cell type that in the other



FindAllMarkers()

Finds genes that are DE between 1 cluster and all other cells.

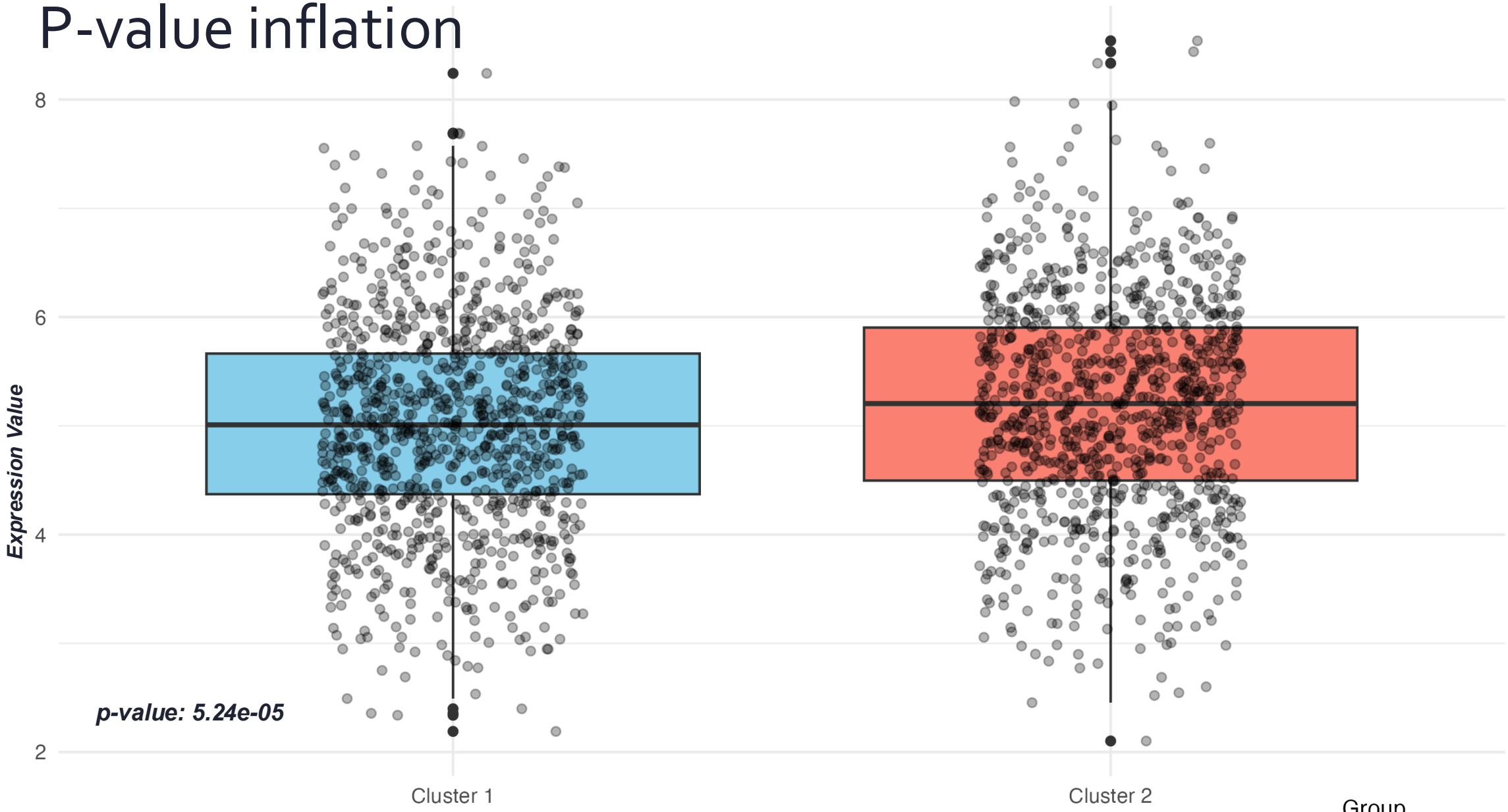
FindMarkers()

to perform pairwise DGE analysis, e.g. between cluster 1 and cluster 2

Typical Output

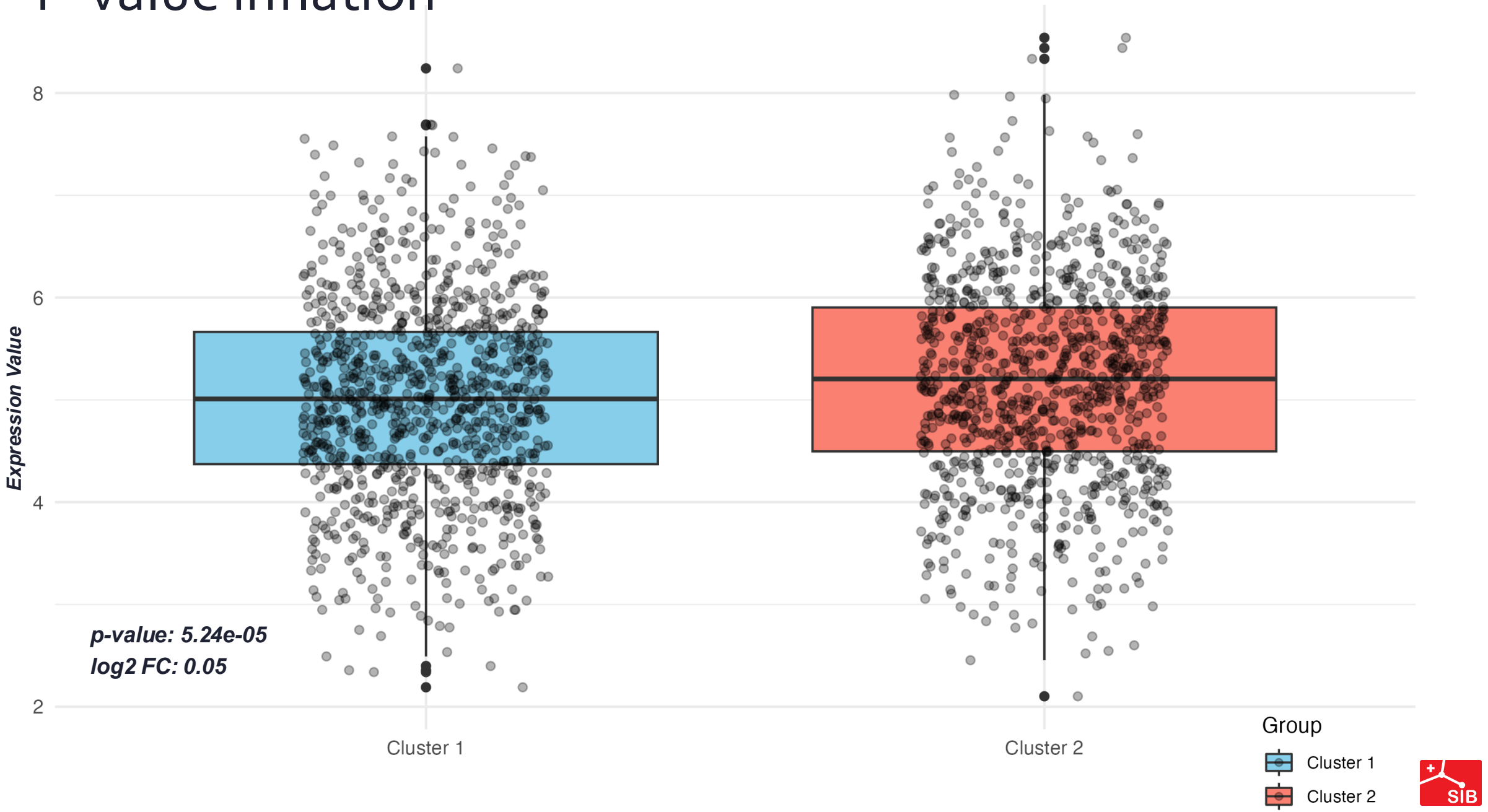
Gene Symbol	avg_log2FC	pct.1	pct.2	p_val	p_val_adj	cluster
CHI3L1	5.61	0.958	0.225	7.63E-255	1.97E-250	2
HLA-DRA	3.41	0.978	0.215	2.84E-253	7.32E-249	2
PTGFR	4.31	0.795	0.093	6.43E-244	1.66E-239	2
HLA-DRB5	3.54	0.818	0.097	2.10E-243	5.41E-239	2
GRIN2A	4.24	0.69	0.05	1.64E-235	4.22E-231	2
CDHR3	3.34	0.892	0.159	2.56E-229	6.59E-225	2
AKR1C3	3.16	0.955	0.254	5.75E-223	1.48E-218	2
KCNK15	3.76	0.897	0.187	4.88E-217	1.26E-212	2
HLA-DRB1	2.53	0.78	0.102	2.71E-212	6.98E-208	2
PLPP3	3.34	0.965	0.322	3.54E-210	9.11E-206	2

P-value inflation



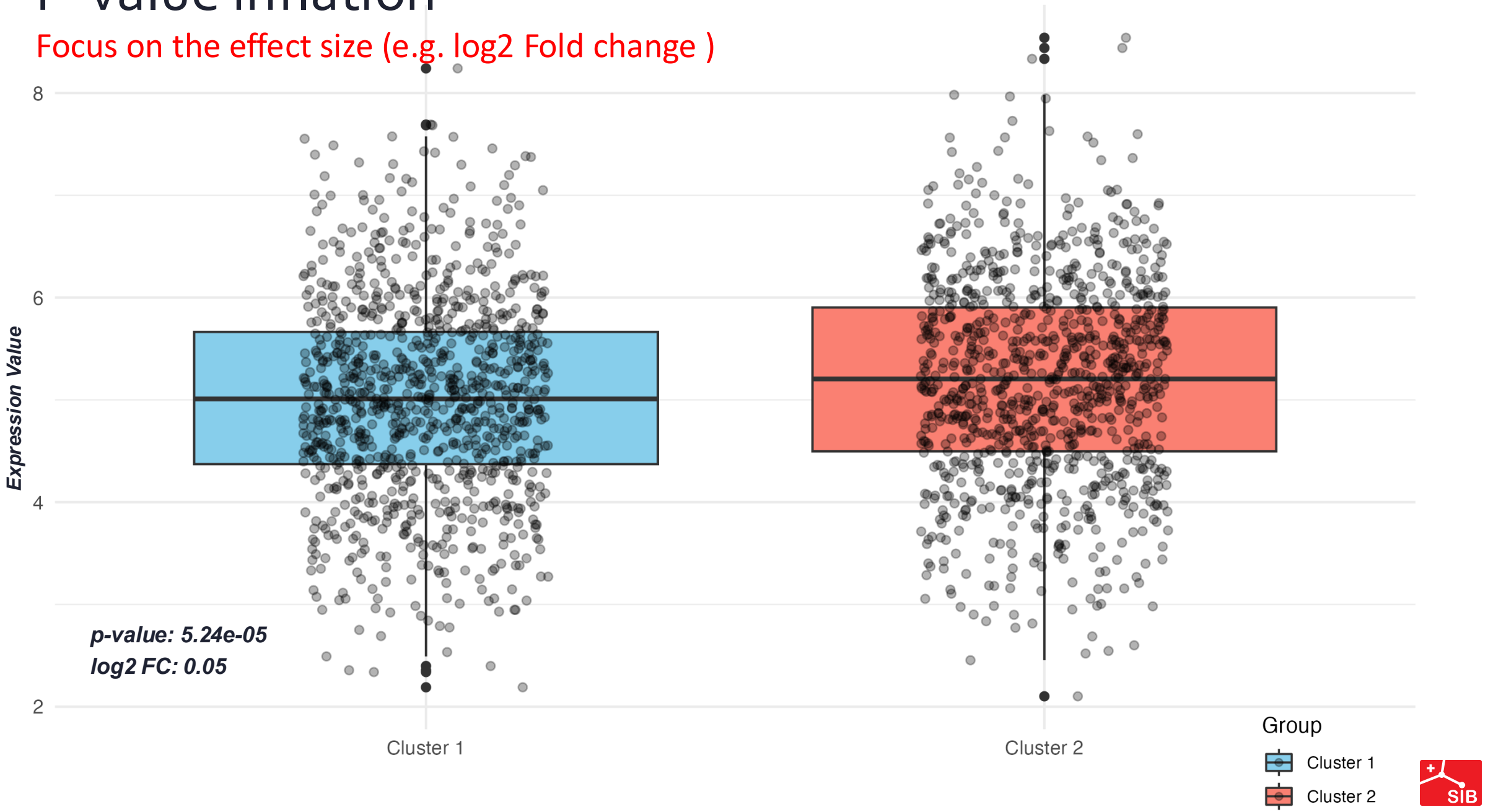
The probability of observing a test statistic at least as extreme as the one observed, assuming that the null hypothesis is true.

P-value inflation



P-value inflation

Focus on the effect size (e.g. log2 Fold change)



What is the ideal method

Analysis | Published: 26 February 2018

Bias, robustness and scalability in single-cell differential expression analysis

Charlotte Soneson & Mark D Robinson

Many methods have been used to determine differential gene expression from single-cell RNA (scRNA)-seq data. We evaluated 36 approaches using experimental and synthetic data and found considerable differences in the number and characteristics of the genes that are called differentially expressed. Prefiltering of lowly expressed genes has important effects, particularly for some of the methods developed for bulk RNA-seq data analysis. However, we found that bulk RNA-seq analysis methods do not generally

https://www.nature.com/articles/nmeth.4612

What is the ideal method

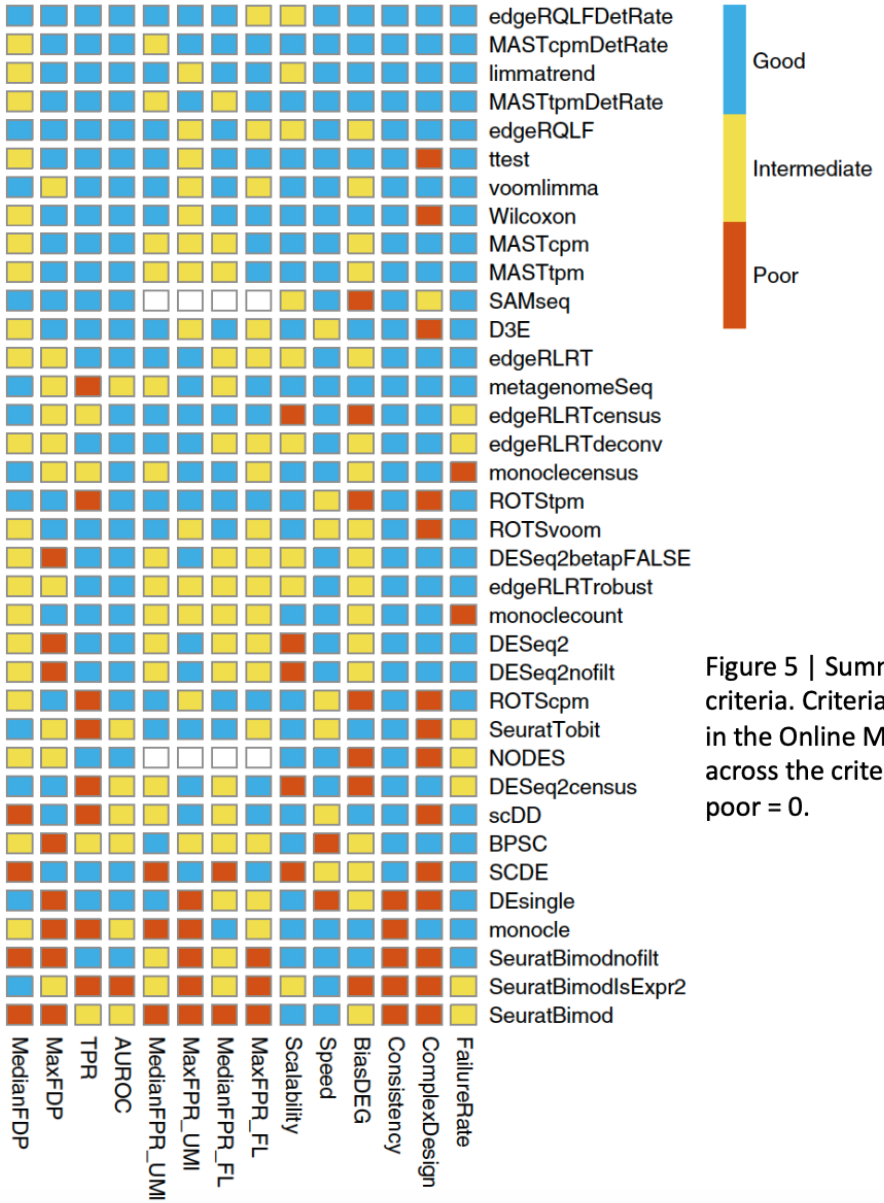


Figure 5 | Summary of DE method performance across all major evaluation criteria. Criteria and cutoff values for performance categories are available in the Online Methods. Methods are ranked by their average performance across the criteria, with the numerical encoding good = 2, intermediate = 1, poor = 0.

https://github.com/csoneson/conquer_comparison/blob/master/scripts/apply_edgeRQLFDetRate.R

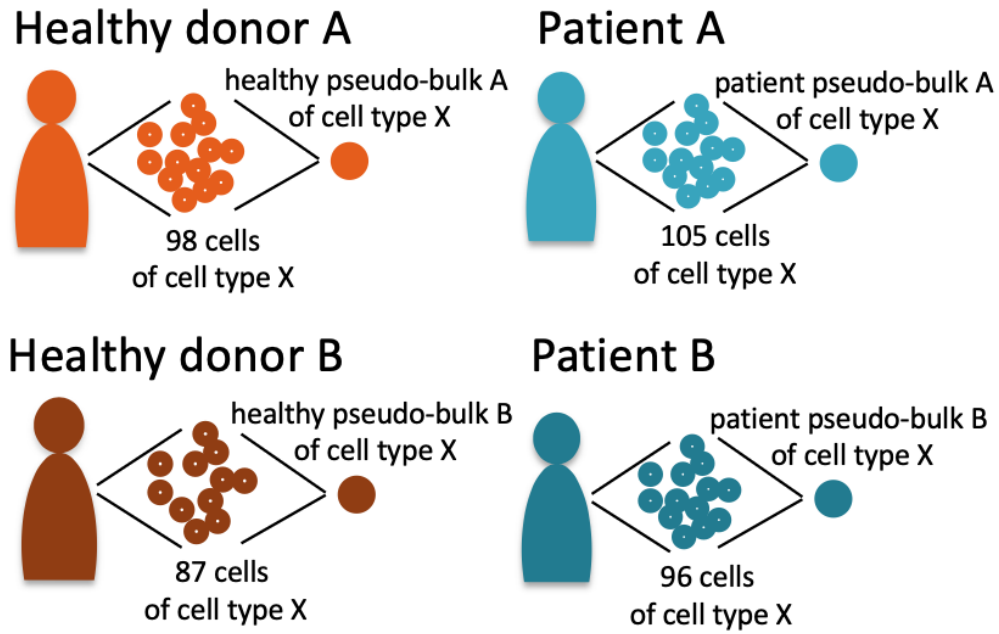
Limma/edgeR: old but gold

Methods designed for microarray and bulk RNAseq analysis

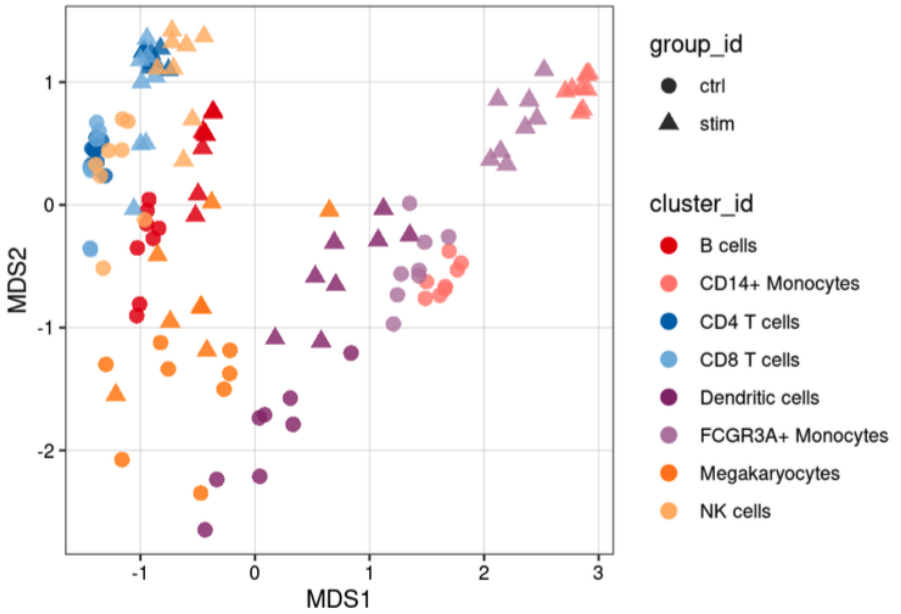
- Can be used to include batch effects in model as covariates
- Compare more than 2 groups: e.g. ANOVA(F-test)
- Can be used to analyze factorial design such as genotype x treatment

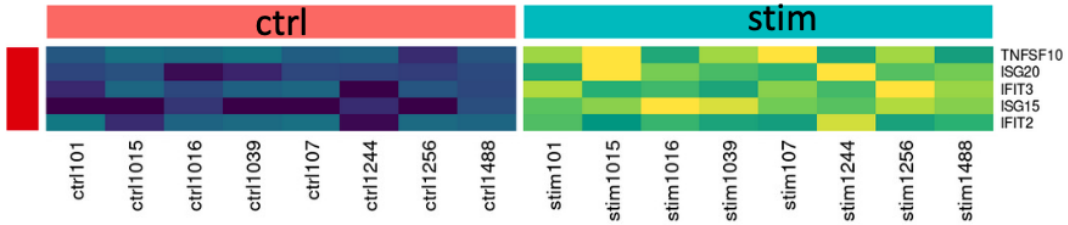
Analysis with limma and example of model with covariate:
https://ucdavis-bioinformatics-training.github.io/2018-June-RNA-Seq-Workshop/thursday/DE.html

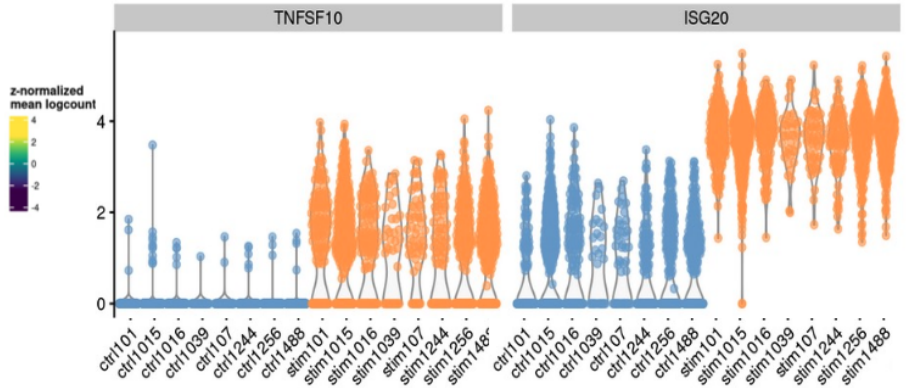
Pseudo-bulk DE analysis: muscat



One pseudo-bulk per cell type per sample



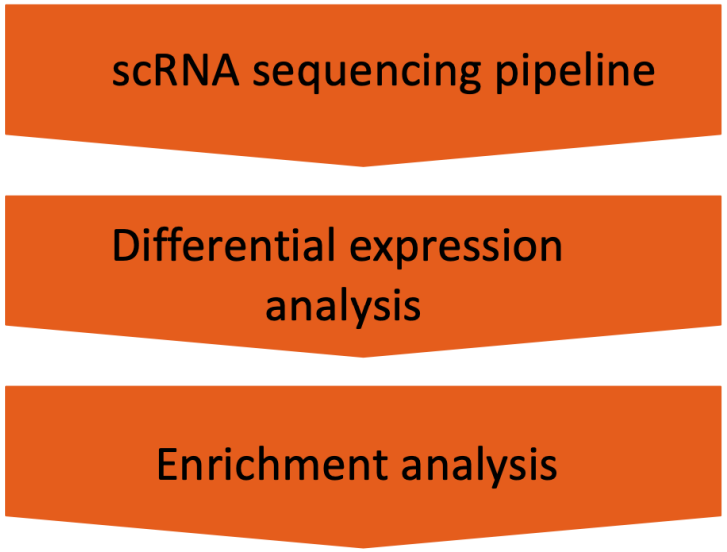




https://www.bioconductor.org/packages/release/bioc/vignettes/muscat/inst/doc/analysis.html

In the practical exercise: scuttle to aggregate counts + limma

From matrix to biological insights



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

What is a gene set?

A gene set is an unordered collection of genes that are functionally related.

- 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 that contain SNPs associated with a disease
- ...etc!
- Several gene sets are grouped into Knowledge bases
- A pathway can be interpreted as a gene set by ignoring functional relationships among genes

On-line Reseources:

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

The database containing several types of gene set lists:

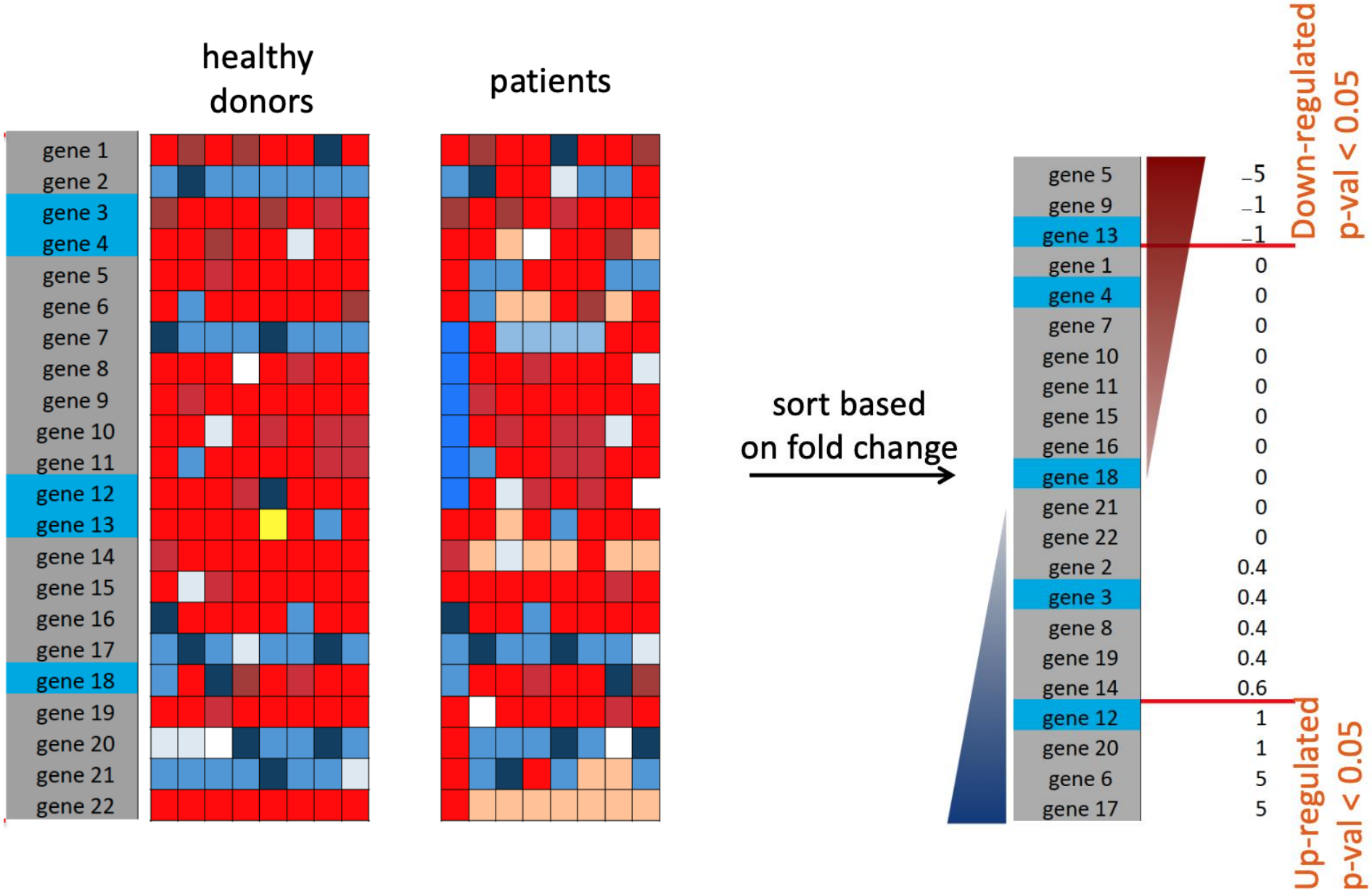
- Hallmark of cancer
- Positional Gene Set
- Published gene sets

KEGG -> https://www.kegg.jp/kegg/pathway.html

Reactome -> https://reactome.org/

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

Is the blue gene-set "modulated"



Fisher's exact test

```
> 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**

Fisher's exact test

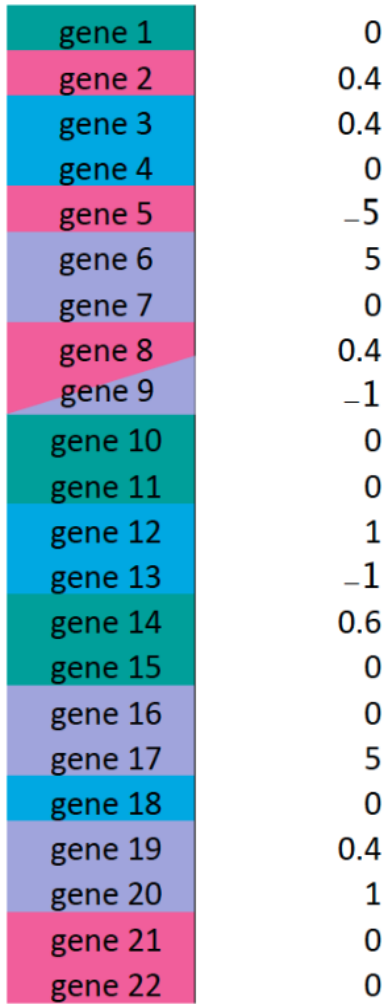
```
> cont.table<-matrix(c(2,3,5,12), ncol=2, byrow = T)
> fisher.test(cont.table)

        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**

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.

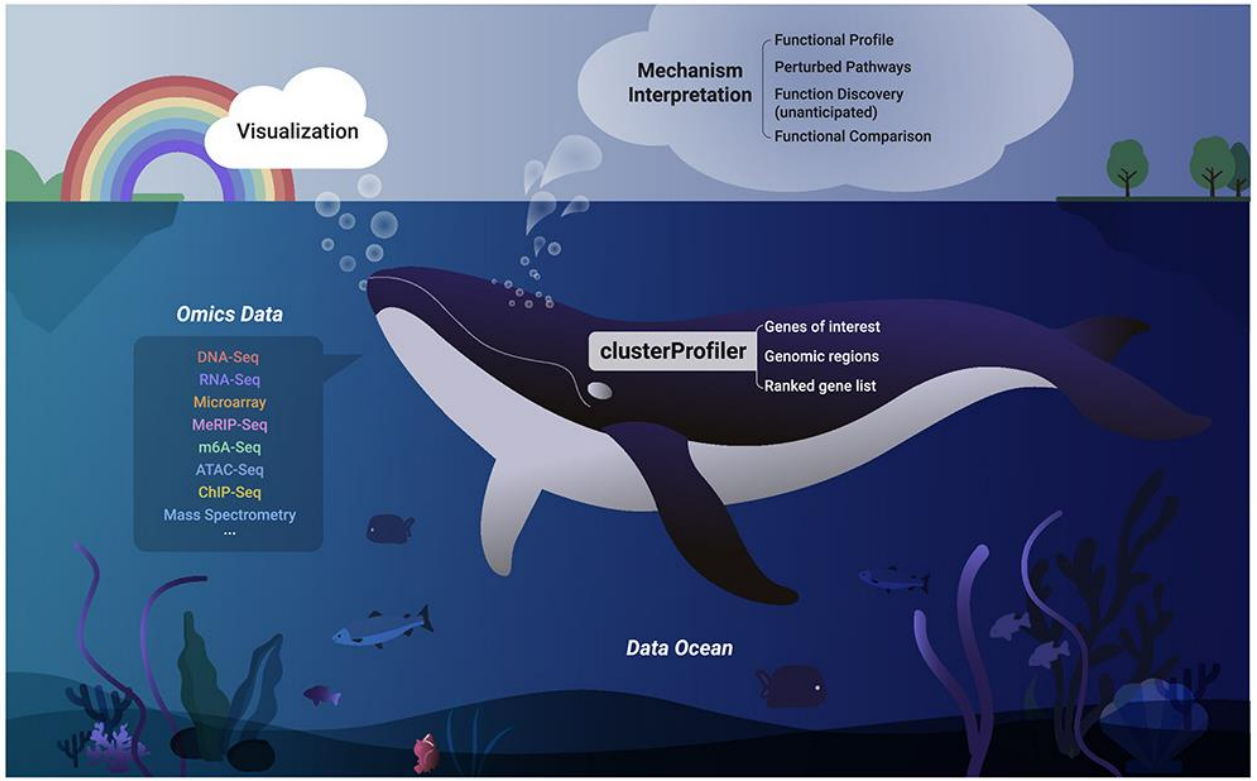
⇒Fisher test is threshold-based

clusterProfiler

A universal enrichment tool for interpreting omics data

platforms all | rank 36 / 2300 | support 1 5 / 1 8 | in Bioc 13.5 years | build ok | updated < 3 months | dependencies 132

DOI: 10.18129/B9.bioc.clusterProfiler



Wu T, et al. (2021). “clusterProfiler 4.0: A universal enrichment tool for interpreting omics data.” *The Innovation*, **2**(3), 100141. doi:10.1016/j.xinn.2021.100141.

Functions for over-representation analyses

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)
```

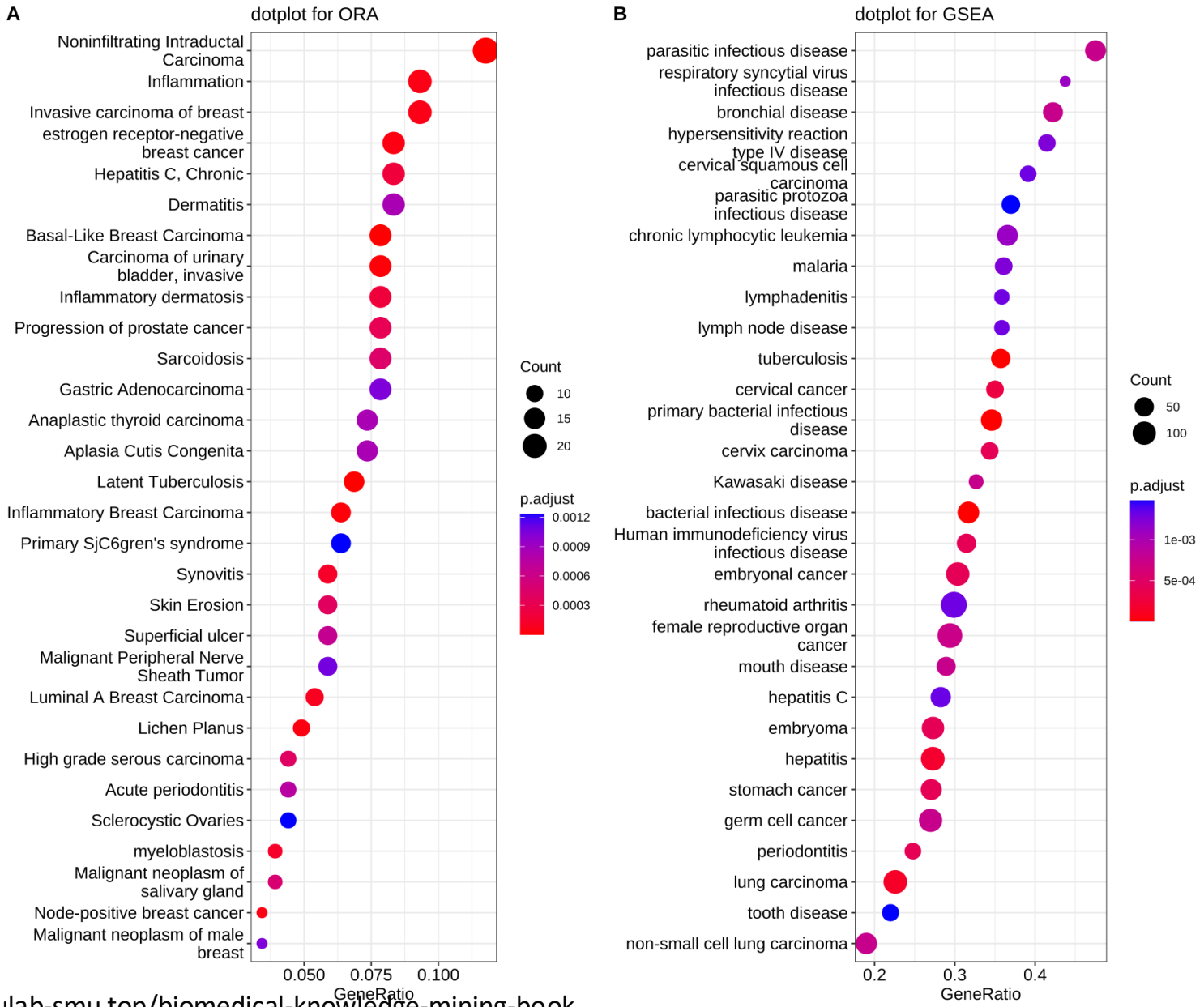
Over-representation analysis (similar to Fisher test) for built-in GO gene sets:

```
enrichGO(gene, OrgDb, keyType = "ENTREZID", ont = "MF",
  pvalueCutoff = 0.05, pAdjustMethod = "BH", universe,
  qvalueCutoff = 0.2, minGSSize = 10, maxGSSize = 500,
  readable = FALSE, pool = FALSE)
```

enricher(): similar enrichGO() but for user defined gene sets

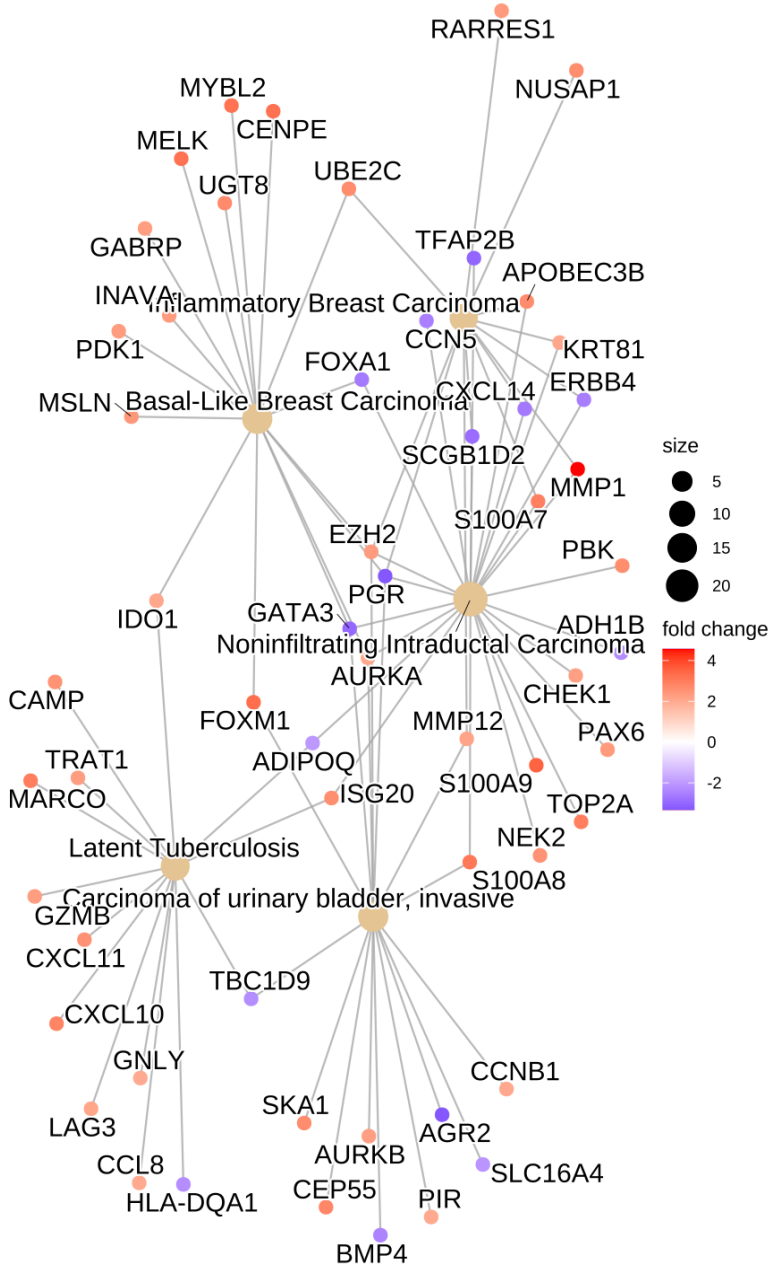
```
enricher(gene, pvalueCutoff = 0.05, pAdjustMethod = "BH", universe,
  minGSSize = 10, maxGSSize = 500, qvalueCutoff = 0.2, TERM2GENE,
  TERM2NAME = NA)
```

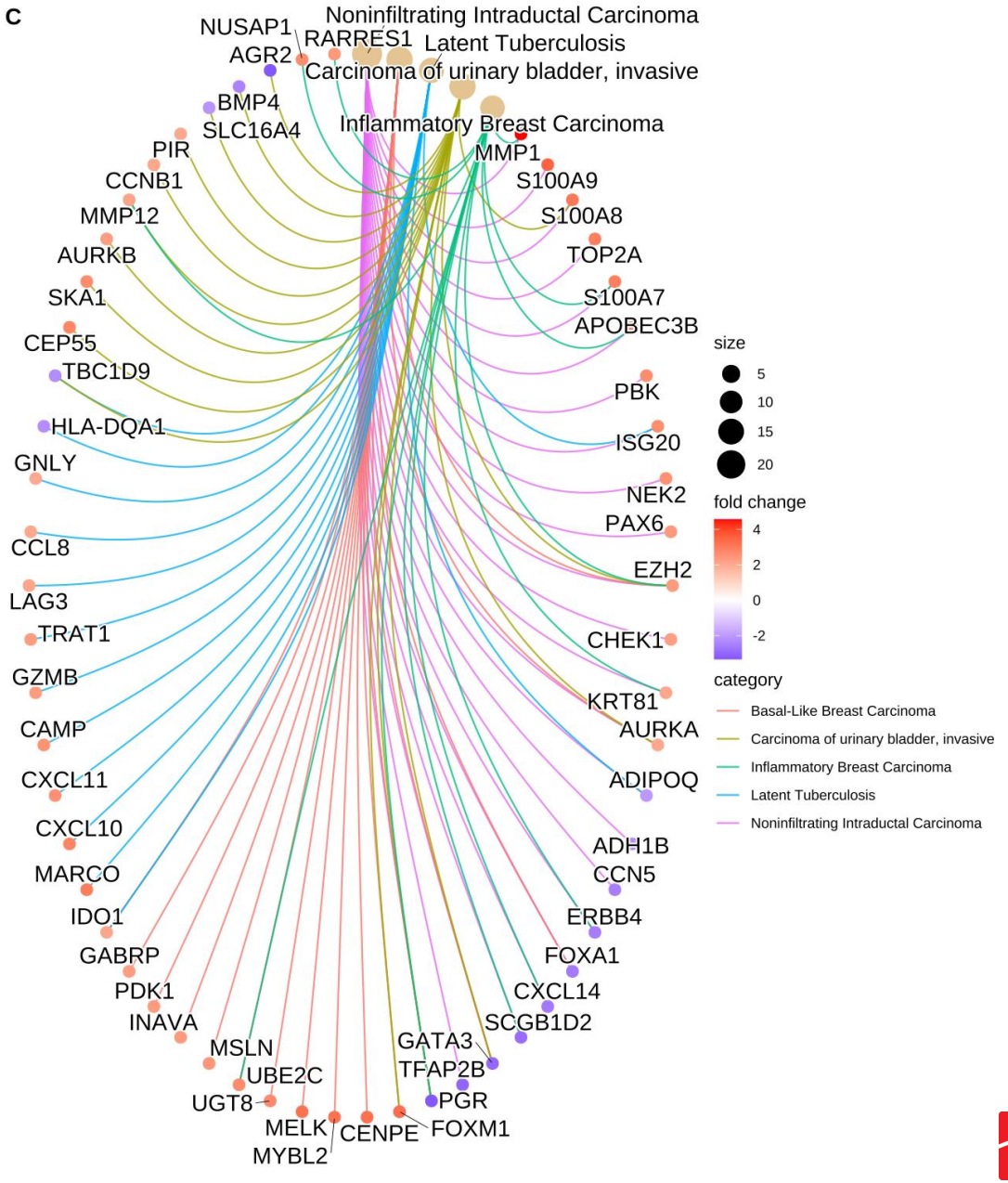
Visualization



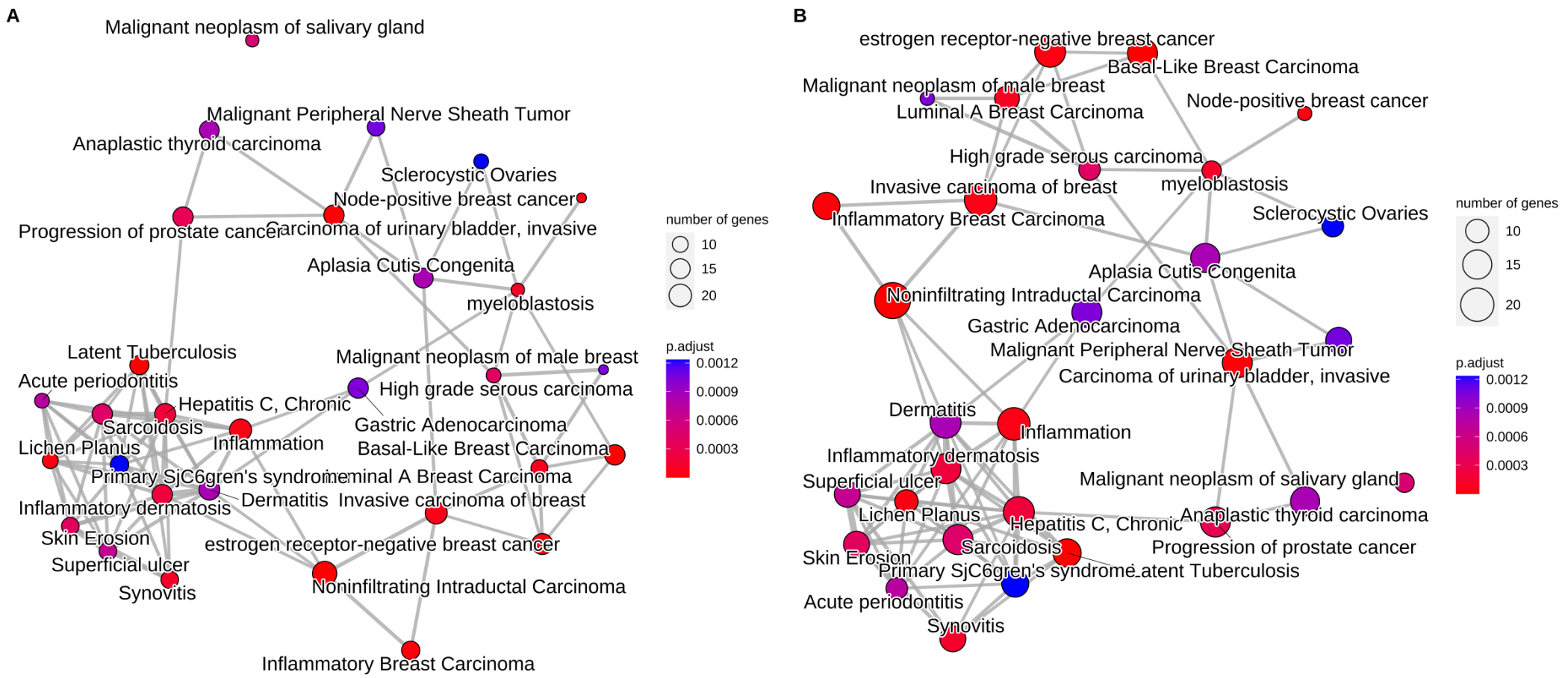
Adapted from: https://yulab-smu.top/biomedical-knowledge-mining-book

Gene-Cocept Network





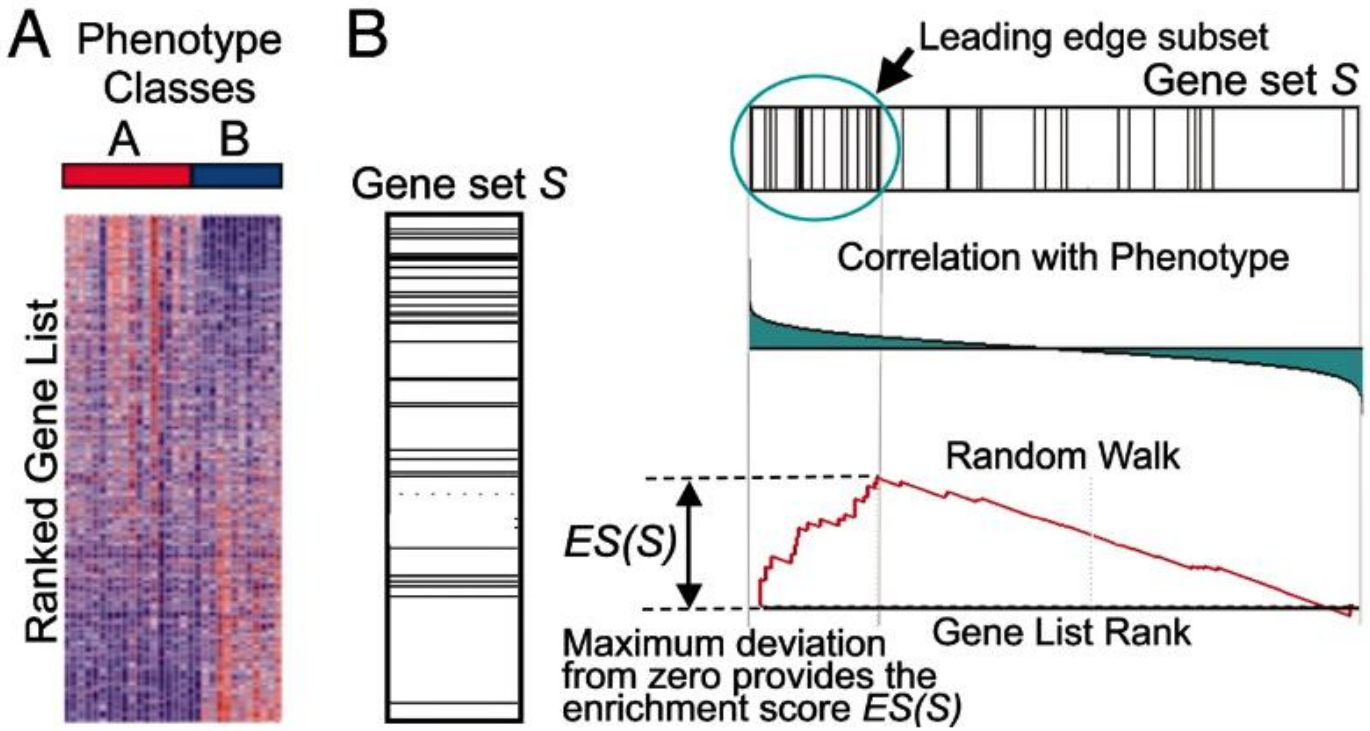
Enrichment Map



Edges connect overlapping gene sets. In this way, overlapping gene sets tend to cluster together, making it easy to identify functional module.

Gene-Set Enrichment Analysis

ORA fails to detect situations where all genes in a predefined set change in a small but coordinated way



Genes are ranked based on their phenotypes.

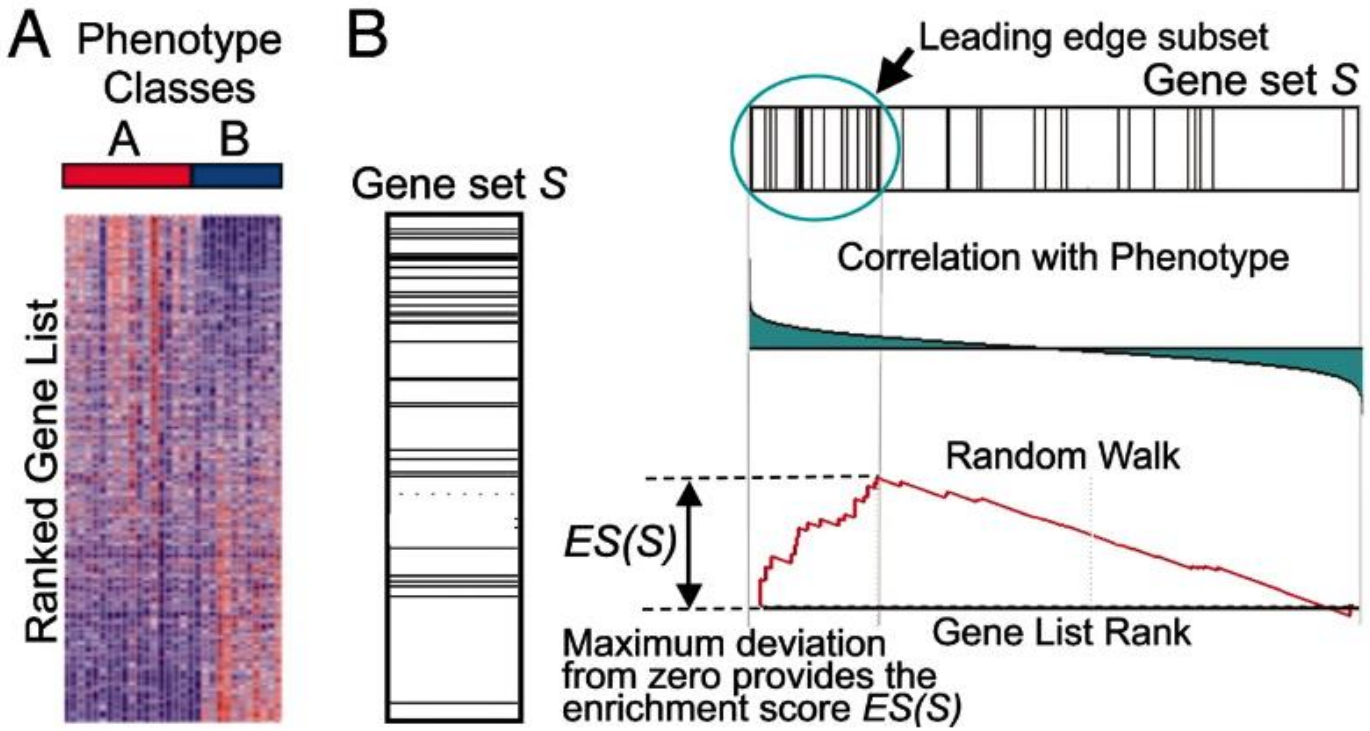
Given apriori defined set of gene *S*, the goal of GSEA is to determine whether the members of *S* are randomly distributed throughout the ranked gene list (*L*) or primarily found at the top or bottom.

clusterProfiler:
gseGO(): GSEA of GO terms using all ranked genes
gseKEGG(): GSEA of KEGG pathways using all ranked genes
GSEA(): GSEA of custom gene set collection using all ranked genes

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

Gene-Set Enrichment Analysis

ORA fails to detect situations where all genes in a predefined set change in a small but coordinated way



Genes are **ranked** based on their phenotypes.

Given apriori defined set of gene *S*, the goal of GSEA is to determine whether the members of *S* are randomly distributed throughout the ranked gene list (*L*) or primarily found at the top or bottom.

clusterProfiler:
gseGO(): GSEA of GO terms using all ranked genes
gseKEGG(): GSEA of KEGG pathways using all ranked genes
GSEA(): GSEA of custom gene set collection using all ranked genes

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

How Can I Rank the Genes?

How can I rank the genes?

Research Article | Open access | Published: 12 May 2017

Ranking metrics in gene set enrichment analysis: do they matter?

Joanna Zyla, Michal Marczyk, January Weiner & Joanna Polanska

BMC Bioinformatics **18**, Article number: 256 (2017) | Cite this article

45k Accesses | **43** Citations | **22** Altmetric | Metrics

Abstract

Background

There exist many methods for describing the complex relation between changes of gene expression in molecular pathways or gene ontologies under different experimental conditions. Among them, Gene Set Enrichment Analysis seems to be one of the most commonly used (over 10,000 citations). An important parameter, which could affect the final result, is the choice of a metric for the ranking of genes. Applying a default ranking metric may lead to poor results.

How can I rank the genes?

Research Article | Open access | Published: 12 May 2017

Ranking metrics in gene set enrichment analysis: do they matter?

Joanna Zyla, Michal Marczyk, January Weiner & Joanna Polanska

BMC Bioinformatics **18**, Article number: 256 (2017) | Cite this article

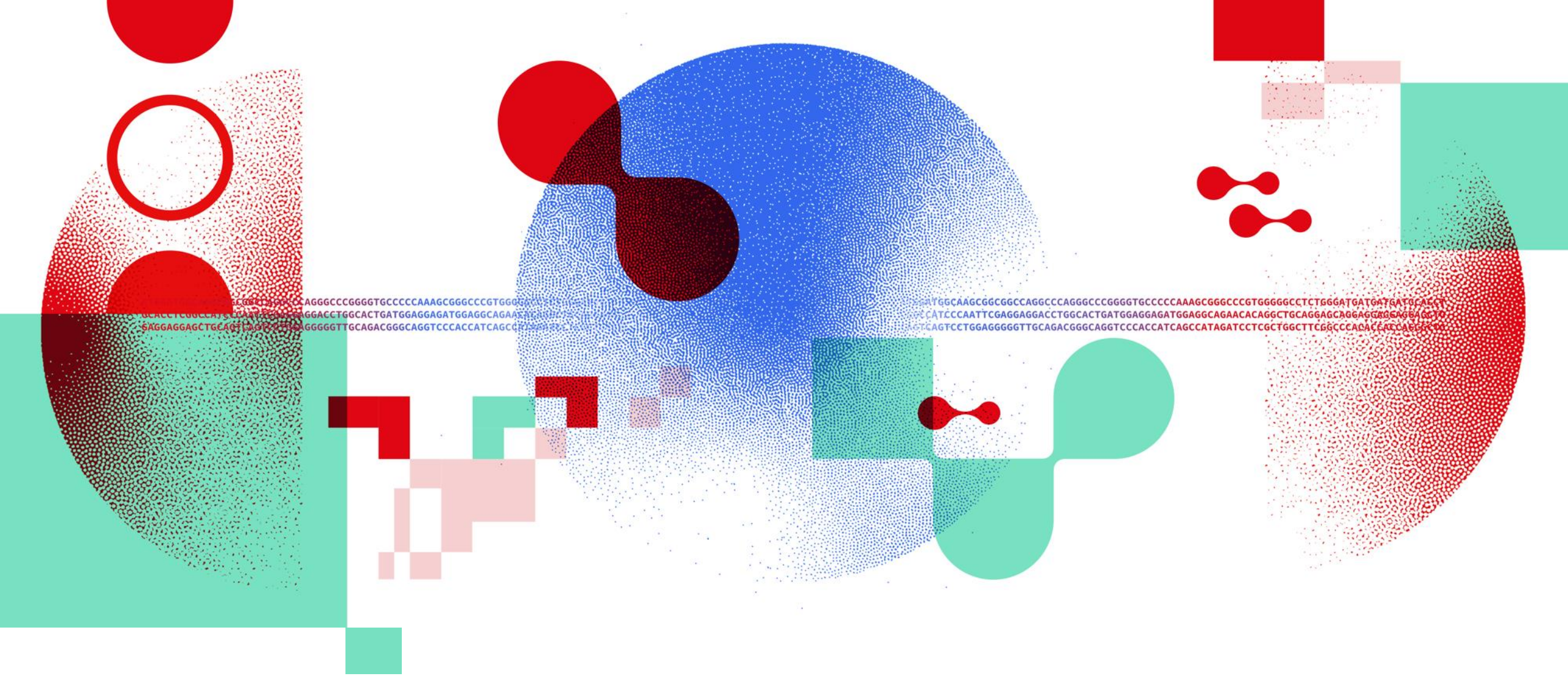
45k Accesses | **43** Citations | **22** Altmetric | Metrics

Abstract

Background

There exist many methods for describing the complex relation between changes of gene expression in molecular pathways or gene ontologies under different experimental conditions. Among them, Gene Set Enrichment Analysis seems to be one of the most commonly used (over 10,000 citations). An important parameter, which could affect the final result, is the choice of a metric for the ranking of genes. Applying a default ranking metric may lead to poor results.

```
My ranking metric: sign(log2FC) * -log10(p-value)
```



Thank you

DATA SCIENTISTS FOR LIFE

sib.swiss