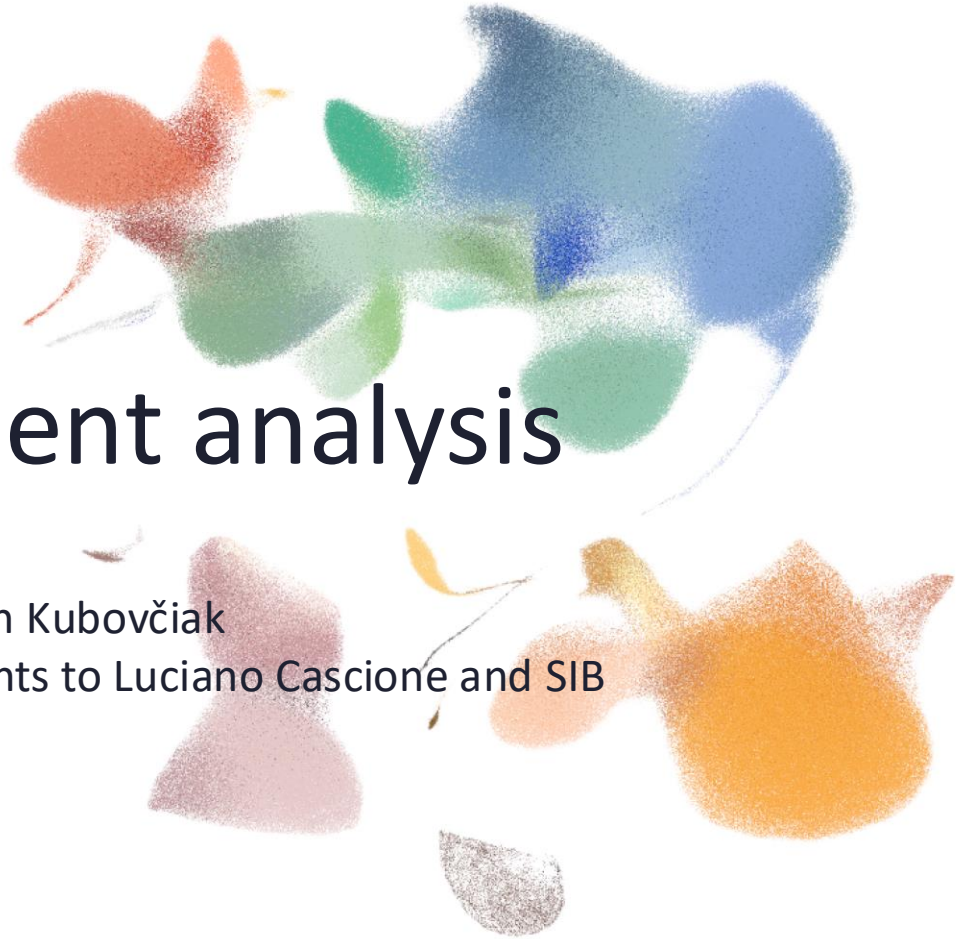
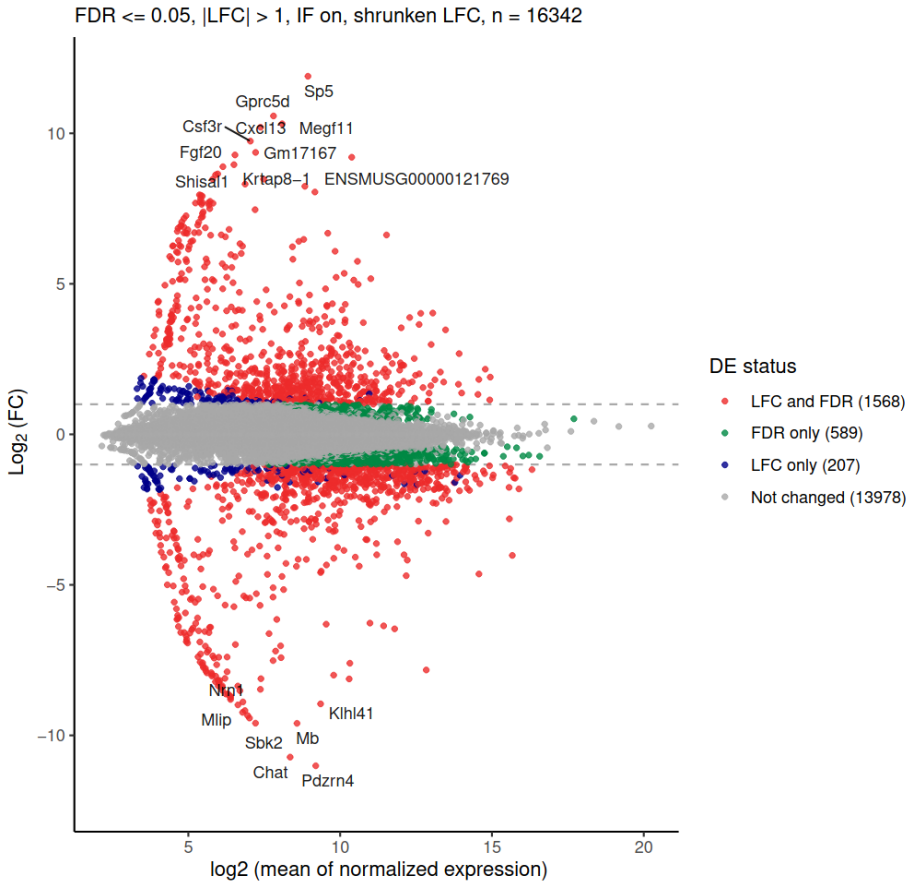


Enrichment analysis

Jan Kubovčiak
with acknowledgements to Luciano Cascione and SIB

From DEGs to biological insights



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

- test if differentially expressed genes are enriched in genes associated with a particular function

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

Gene set resources:

Gene Ontology -> https://geneontology.org

- Biological process
- Molecular function
- Cellular component

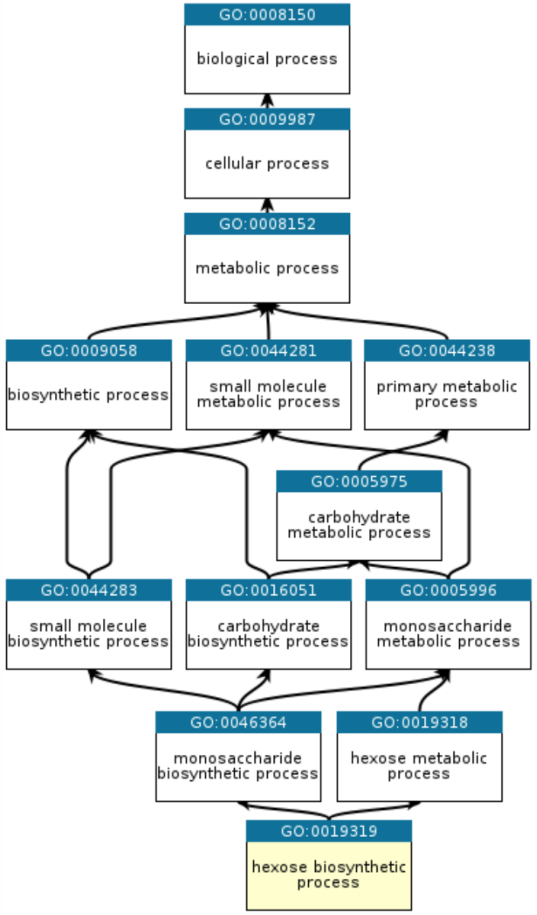
KEGG -> https://www.kegg.jp/kegg/pathway.html

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

- Hallmark gene sets
- Positional gene sets
- Immunological sets etc.

Reactome -> https://reactome.org/

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



Methods

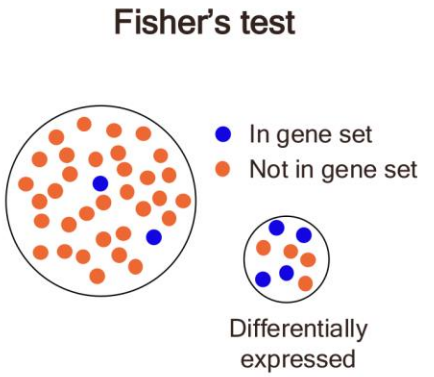
- Over-representation analysis (ORA)
  - Fisher's test of proportions of DEGs and gene set
- Gene set enrichment analysis (classical GSEA)
  - Enrichment score of gene set in ranked list of genes based on deregulation metric

Over-representation analysis (ORA)

```
> 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
```



H_0: The proportion of genes in the gene set is the same for both groups

H_a: The proportion of genes in the gene set is higher in the differentially expressed group

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

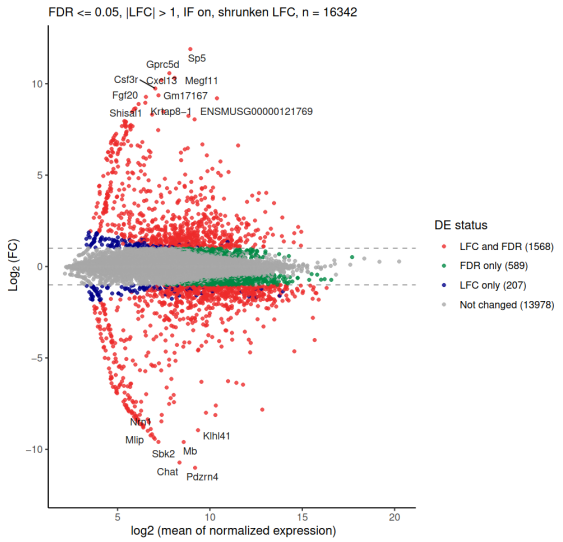
Which gene-sets are differentially expressed

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

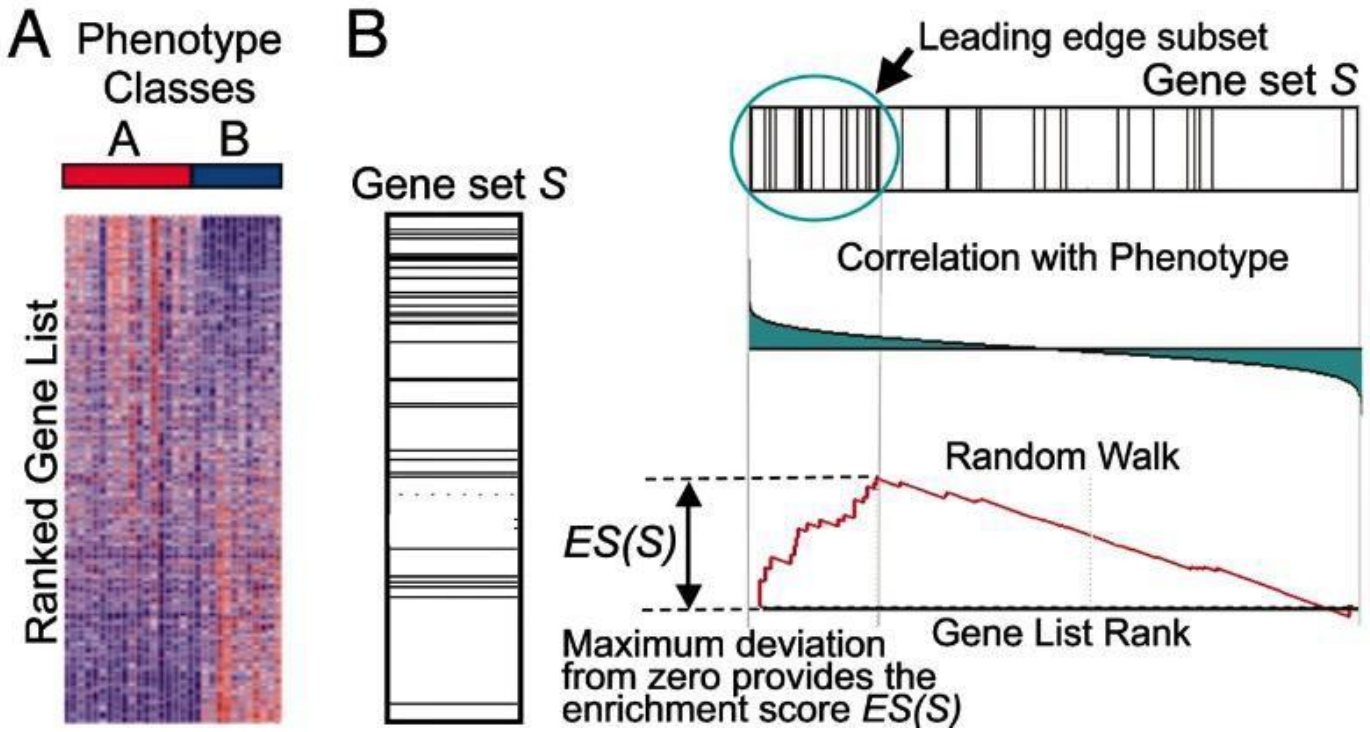
⇒Multiple tests need p-value adjustment.

⇒Fisher test is threshold-based



Gene-Set Enrichment Analysis

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

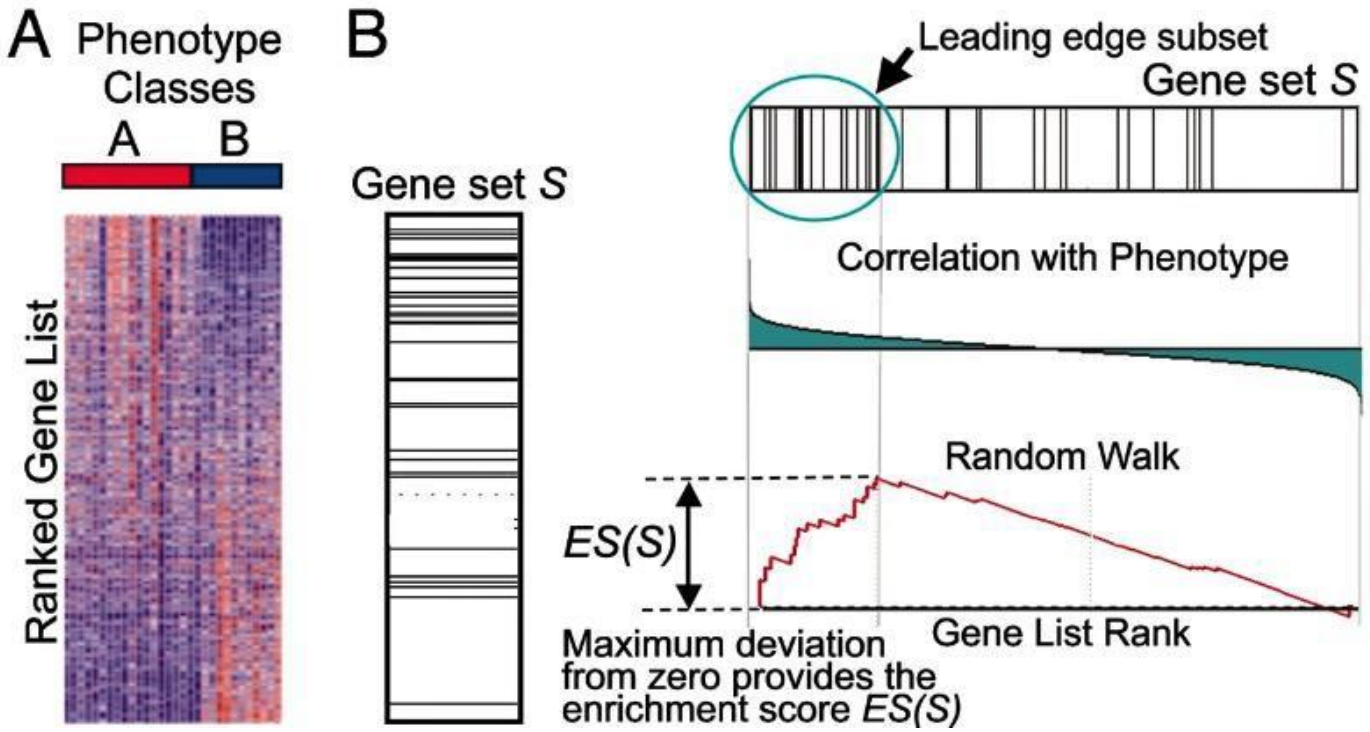


Genes are ranked based on their phenotypes.

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.

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.

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.

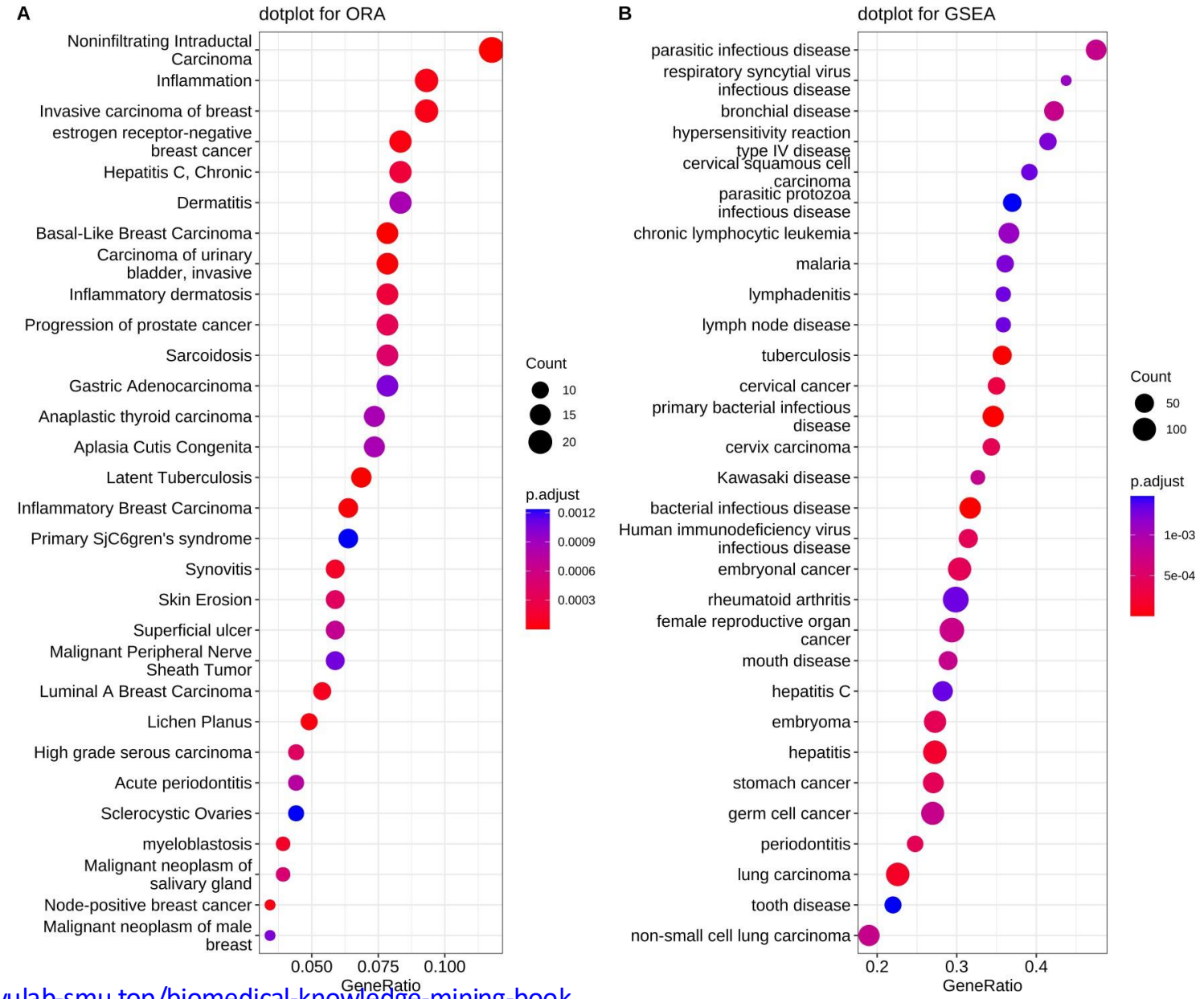
clusterProfiler

A universal enrichment tool for interpreting omics data

	KEGG	GO terms	custom
ORA	enrichKEGG()	enrichGO()	enricher()
GSEA	gseKEGG()	gseGO()	GSEA()

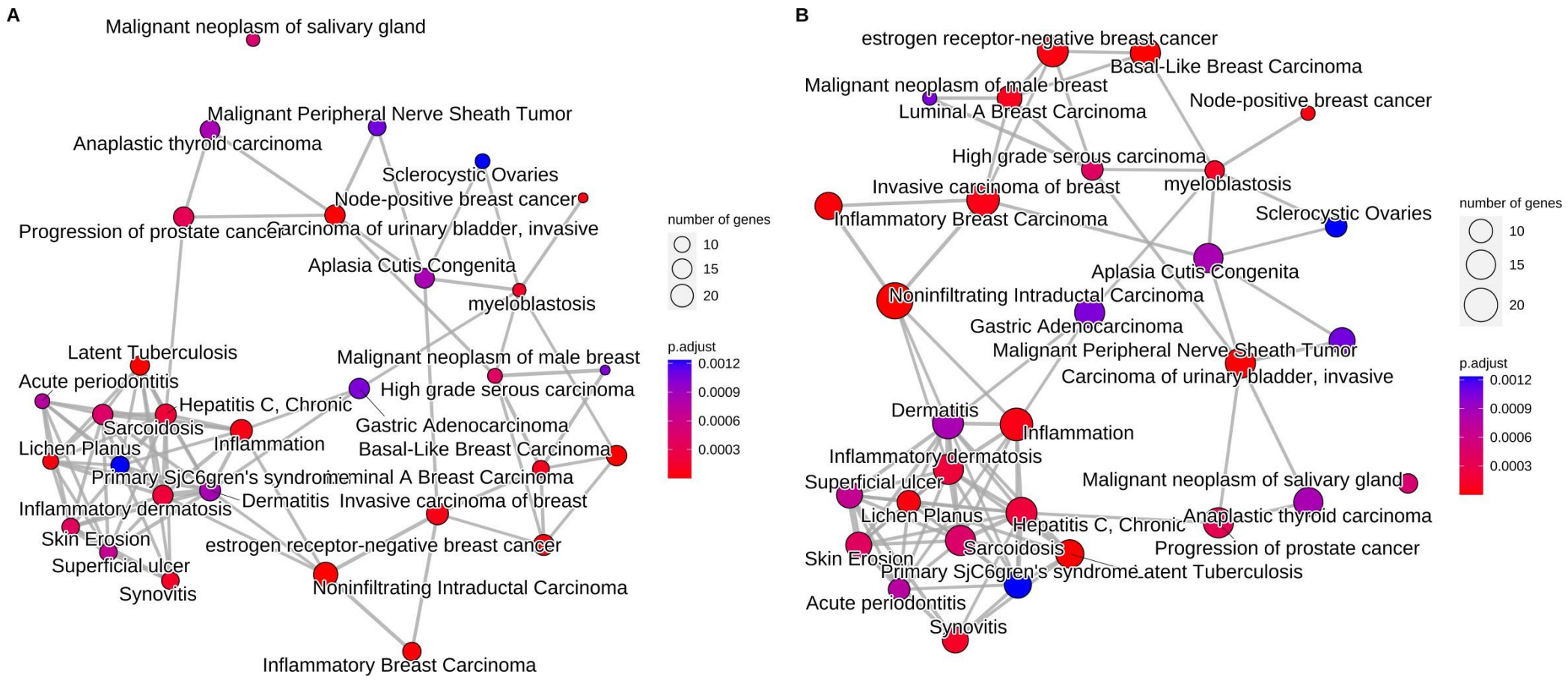
Wu T, et al. (2021). “clusterProfiler 4.0: A universal enrichment tool for interpreting omics data.” *The Innovation*, **2**(3), 100141. doi:10.1016/j.xinn.2021.100141.

Visualization



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

Enrichment Map



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

Thanks for your attention!