Enrichment analyses and results contextualisation with Knowledge Graphs

Summer School Multi-omics Data Analysis and Integration

Enrichment analyses

- [*Gene sets, pathways, metabolites*] <u>*enrichment*</u> analyses

↓

over-representation

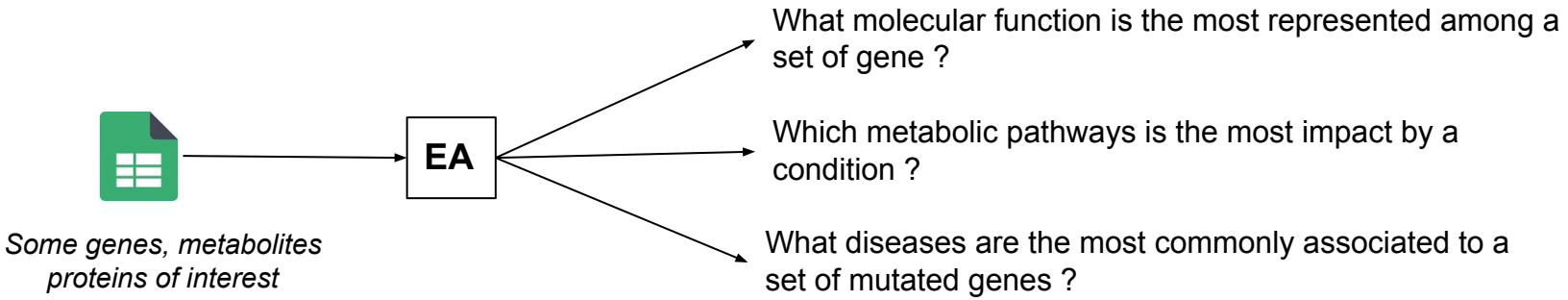
García-Campos, M.A. et al. 2015. Pathway Analysis: State of the Art. Front Physiol

Enrichment analyses

- [*Gene sets, pathways, metabolites*] *enrichment* analyses

over-representation

- Give directions for results interpretation



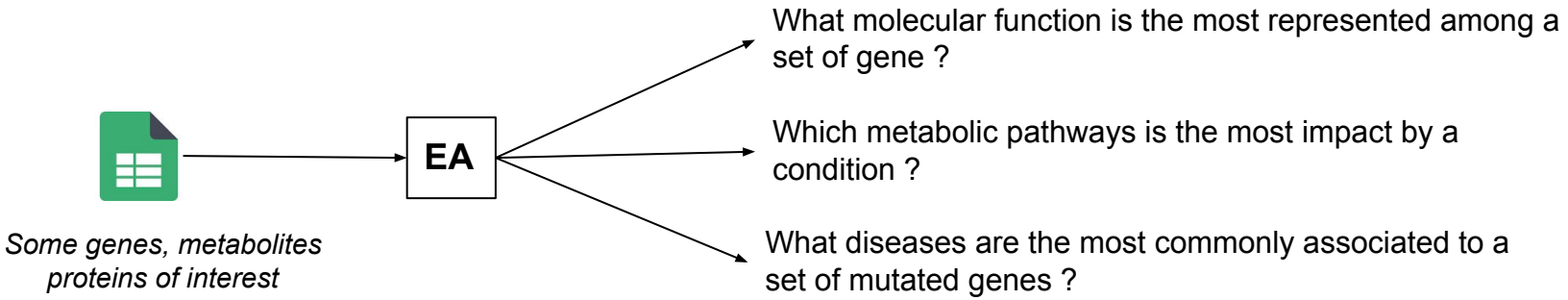
García-Campos, M.A. et al. 2015. Pathway Analysis: State of the Art. Front Physiol

Enrichment analyses

- [*Gene sets, pathways, metabolites*] *enrichment* analyses

over-representation

- Give directions for results interpretation

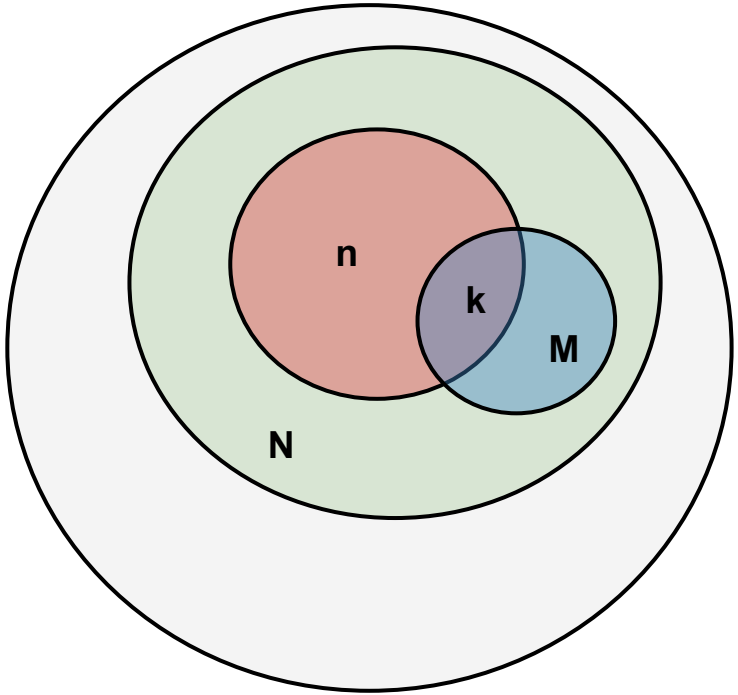


- Families of approaches:
  - Over-Representation Analysis (ORA)
  - Functional Class Scoring (eg. GSEA)
  - Topology-based methods

García-Campos, M.A. et al. 2015. Pathway Analysis: State of the Art. Front Physiol

ORA: Over-Representation Analysis

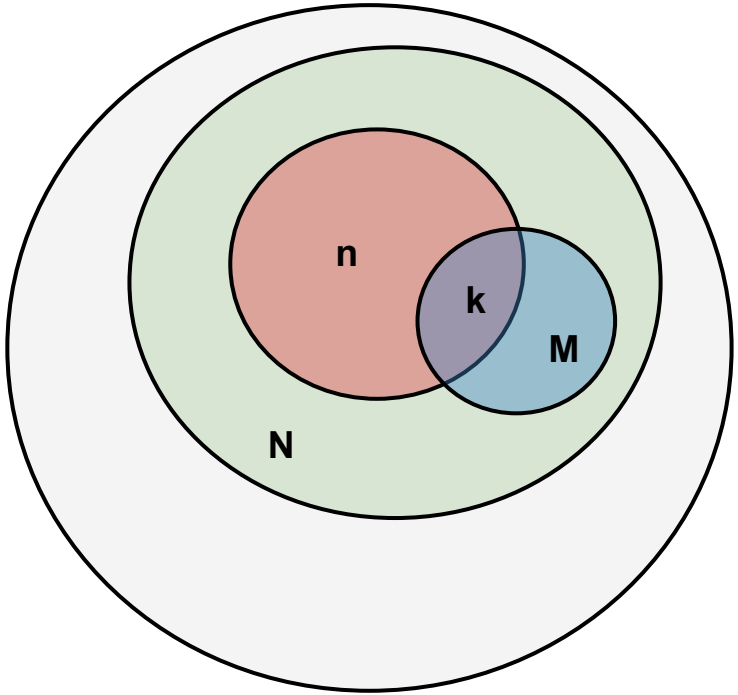
What does an ORA ? it compare **overlap** between **sets**.



- Sets of genes, proteins, metabolites, organisms, etc.
  - a Universe (size = **N**) - *or background set*
  - a set of interest (size = **n**)
  - a reference set (size = **M**) (share a common biological theme)
  - an overlap **k**

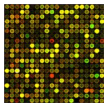
ORA: Over-Representation Analysis

What does an ORA ? it compare **overlap** between **sets**.



- Sets of genes, proteins, metabolites, organisms, etc.
  - a Universe (size = **N**) - *or background set*
  - a set of interest (size = **n**)
  - a reference set (size = **M**) (share a common biological theme)
  - an overlap **k**

In a classic RNA-seq analysis



Experiment → **DE analysis** → **GO terms**

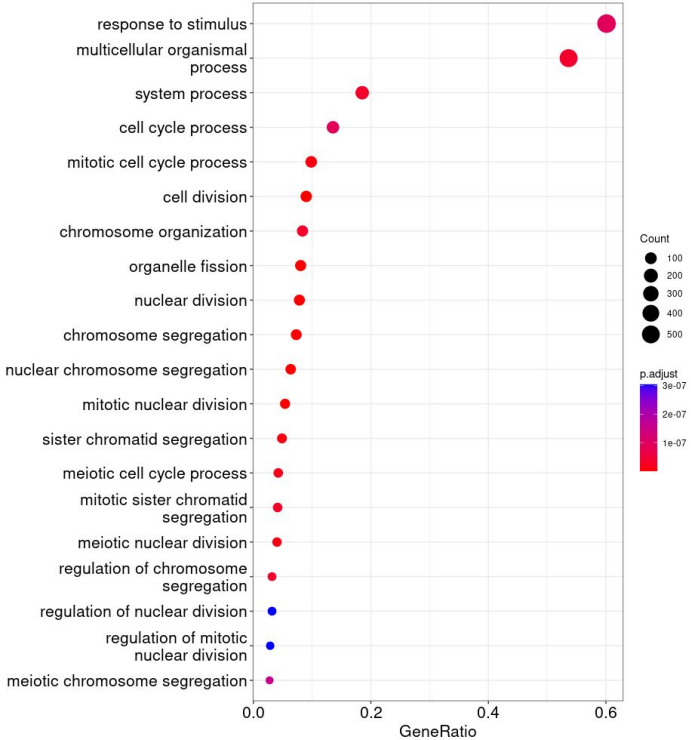
GENEONTOLOGY
Unifying Biology

  - **N** genes measured in the assay
  - **n** genes differentially expressed
  - **M** genes annotated to a GO term of interest
  - an overlap **k**

ORA: A practical example (1)

TCGA-BRCA: 5 Normal .vs. 5 Tumor samples → GDE analysis → 1068 DE genes

R packages for ORA: clusterProfiler, Enrichr



GENEONTOLOGY Unifying Biology

Standard GO (*Biological processes*) Enrichment analysis

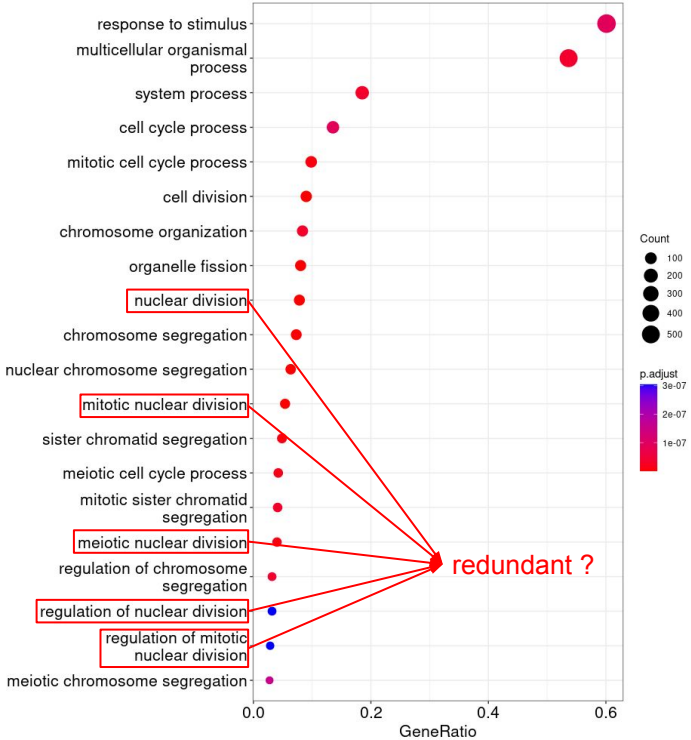
```
ego <- enrichGO(gene = DE.set,
      universe = universe,
      OrgDb = HS.annotation,
      ont = "BP",
      keyType = "SYMBOL",
      minGSSize = 1,
      maxGSSize = 100000,
      pAdjustMethod = "BH")
```

https://master.bioconductor.org/packages/release/bioc/vignettes/TCGAbiolinks/inst/doc/analysis.html

ORA: A practical example (1)

TCGA-BRCA: 5 Normal .vs. 5 Tumor samples → GDE analysis → 1068 DE genes

R packages for ORA: clusterProfiler, Enrichr



GENEONTOLOGY Unifying Biology

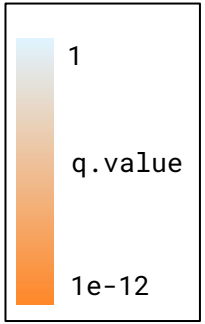
Standard GO (*Biological processes*) Enrichment analysis

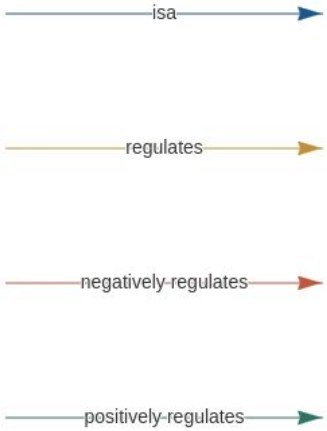
```
ego <- enrichGO(gene = DE.set,
      universe = universe,
      OrgDb = HS.annotation,
      ont = "BP",
      keyType = "SYMBOL",
      minGSSize = 1,
      maxGSSize = 100000,
      pAdjustMethod = "BH")
```

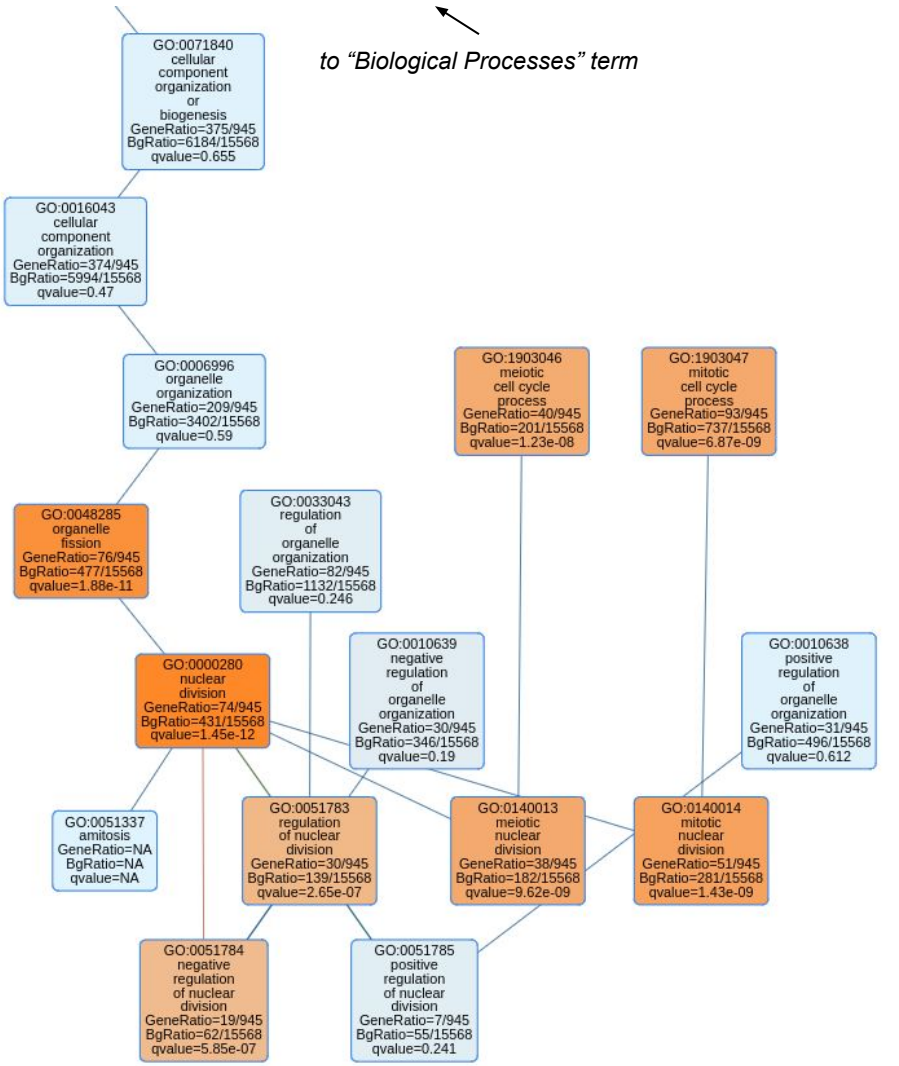
https://master.bioconductor.org/packages/release/bioc/vignettes/TCGAbiolinks/inst/doc/analysis.html

ORA: A practical example (1) - DAG view

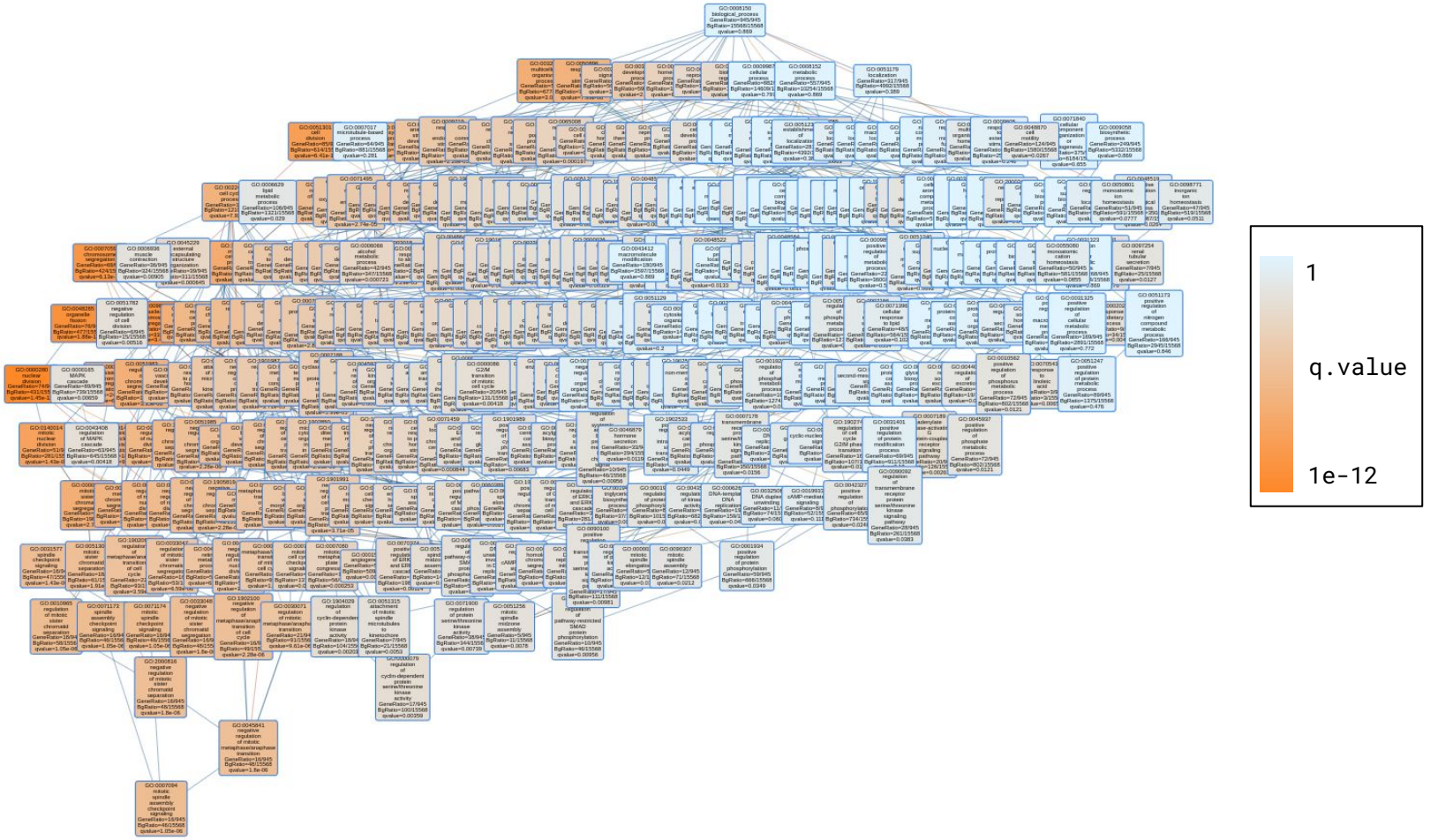
The Gene Ontology in a DAG (**Directed Acyclic** Graph)







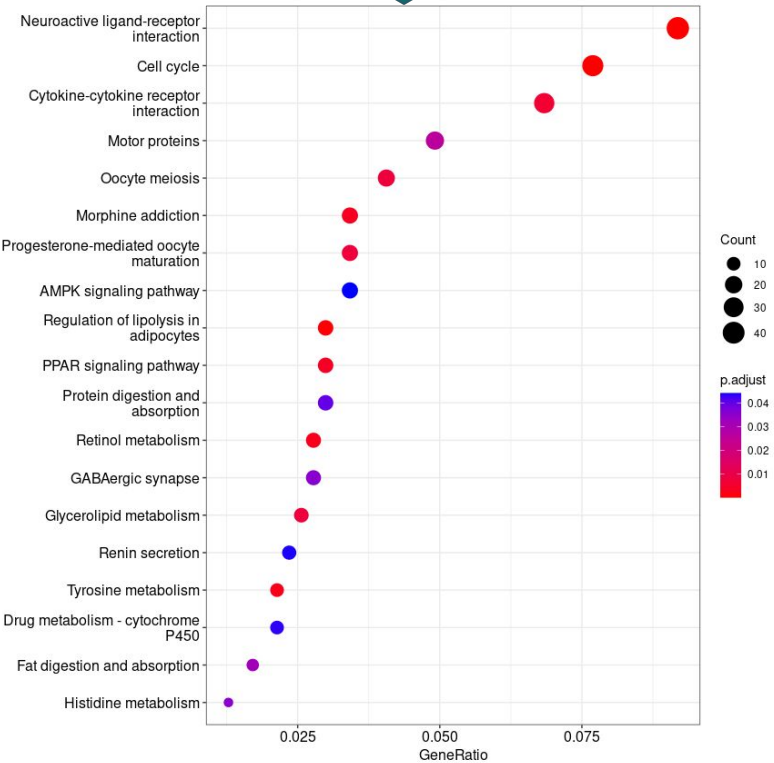
ORA: A practical example (1) - a broader DAG view



ORA: A practical example (2)

TCGA-BRCA: 5 Normal .vs. 5 Tumor samples → GDE analysis → 1068 DE genes

R packages for ORA: clusterProfiler, Enrichr



Standard KEGG (*Pathway*) Enrichment analysis

```
ekegg <- enrichKEGG(gene = DE.set2,
      organism = "hsa",
      keyType = "ncbi-geneid",
      pAdjustMethod = "BH",
      universe = universe2,
      use_internal_data = FALSE)
```

https://master.bioconductor.org/packages/release/bioc/vignettes/TCGAbiolinks/inst/doc/analysis.html

ORA: Back to the fundamentals

ID	Description	GeneRatio	BgRatio	p.value	p.adjust	q.value
GO:0006836	neurotransmitter transport	22/945	191/15568	0.002881322	0.04909408	0.04278198

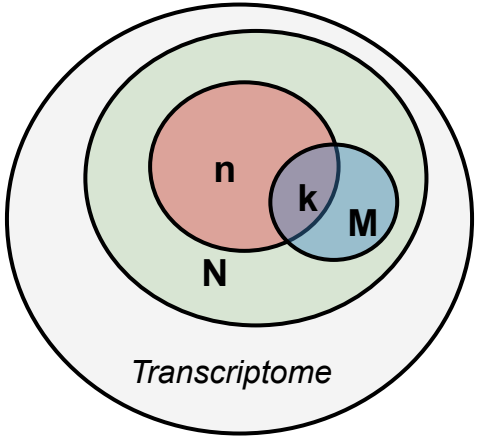
2 equivalent ways of representing and computing

Wieder, C. et al. Pathway analysis in metabolomics: Recommendations for the use of over-representation analysis. PLoS Comput Biol 17

ORA: Back to the fundamentals

ID	Description	GeneRatio	BgRatio	p.value	p.adjust	q.value
GO:0006836	neurotransmitter transport	22/945	191/15568	0.002881322	0.04909408	0.04278198

2 equivalent ways of representing and computing



Hypergeometric distribution

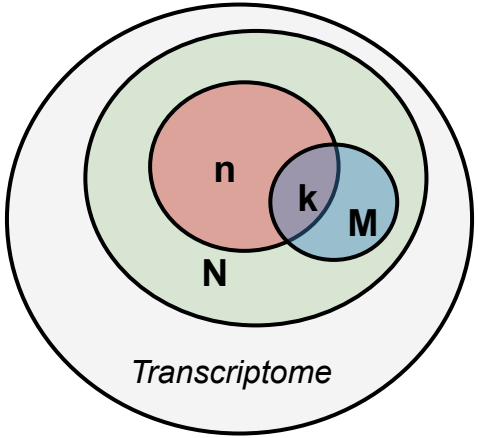
$$P\left(X \geq k\right) = 1 - \sum_{i=0}^{k-1} \frac{\binom{M}{i}\binom{N-M}{n-i}}{\binom{N}{n}}$$

Wieder, C. et al. Pathway analysis in metabolomics: Recommendations for the use of over-representation analysis. PLoS Comput Biol 17

ORA: Back to the fundamentals

ID	Description	GeneRatio	BgRatio	p.value	p.adjust	q.value
GO:0006836	neurotransmitter transport	22/945	191/15568	0.002881322	0.04909408	0.04278198

2 equivalent ways of representing and computing



Hypergeometric distribution

$$P\left(X \geq k\right) = 1 - \sum_{i=0}^{k-1} \frac{\binom{M}{i}\binom{N-M}{n-i}}{\binom{N}{n}}$$

Contingency table

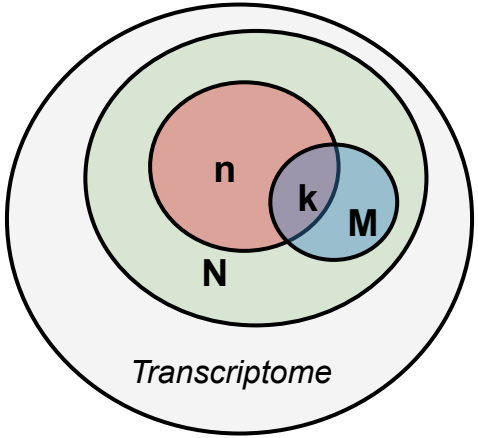
	in BP set	not in BP set
in Gene set	k = 22	923
not in Gene set	169	14454

Wieder, C. et al. Pathway analysis in metabolomics: Recommendations for the use of over-representation analysis. PLoS Comput Biol 17

ORA: Back to the fundamentals

ID	Description	GeneRatio	BgRatio	p.value	p.adjust	q.value
GO:0006836	neurotransmitter transport	22/945	191/15568	0.002881322	0.04909408	0.04278198

2 equivalent ways of representing and computing



Hypergeometric distribution

$$P\left(X \geq k\right) = 1 - \sum_{i=0}^{k-1} \frac{\binom{M}{i}\binom{N-M}{n-i}}{\binom{N}{n}}$$

Contingency table

	in BP set	not in BP set
in Gene set	k = 22	923
not in Gene set	169	14454

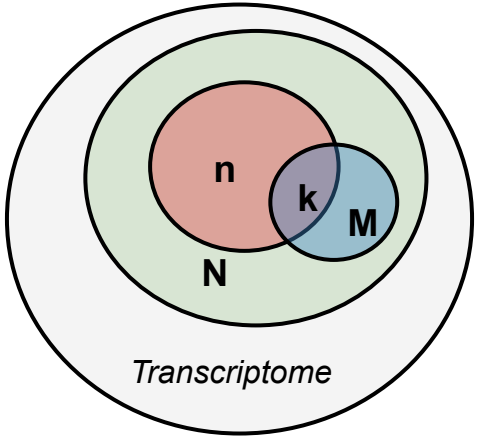
N

Wieder, C. et al. Pathway analysis in metabolomics: Recommendations for the use of over-representation analysis. PLoS Comput Biol 17

ORA: Back to the fundamentals

ID	Description	GeneRatio	BgRatio	p.value	p.adjust	q.value
GO:0006836	neurotransmitter transport	22/945	191/15568	0.002881322	0.04909408	0.04278198

2 equivalent ways of representing and computing



Hypergeometric distribution

$$P\left(X \geq k\right) = 1 - \sum_{i=0}^{k-1} \frac{\binom{M}{i}\binom{N-M}{n-i}}{\binom{N}{n}}$$

Contingency table

	in BP set	not in BP set
in Gene set	k = 22	923
not in Gene set	169	14454

M

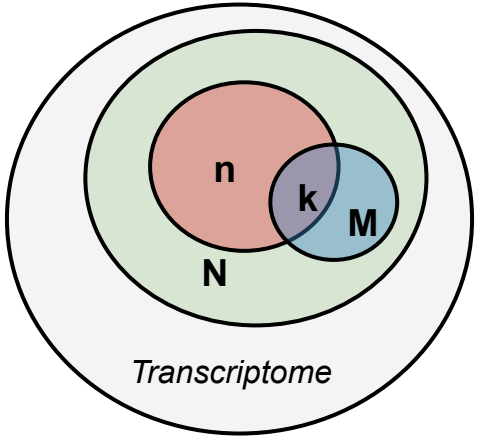
N

Wieder, C. et al. Pathway analysis in metabolomics: Recommendations for the use of over-representation analysis. PLoS Comput Biol 17

ORA: Back to the fundamentals

ID	Description	GeneRatio	BgRatio	p.value	p.adjust	q.value
GO:0006836	neurotransmitter transport	22/945	191/15568	0.002881322	0.04909408	0.04278198

2 equivalent ways of representing and computing



Hypergeometric distribution

$$P\left(X \geq k\right) = 1 - \sum_{i=0}^{k-1} \frac{\binom{M}{i}\binom{N-M}{n-i}}{\binom{N}{n}}$$

Contingency table

	in BP set	not in BP set
in Gene set	k = 22	923
not in Gene set	169	14454

n

M

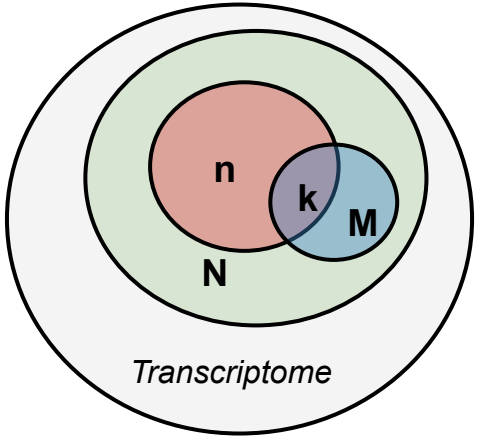
N

Wieder, C. et al. Pathway analysis in metabolomics: Recommendations for the use of over-representation analysis. PLoS Comput Biol 17

ORA: Back to the fundamentals

ID	Description	GeneRatio	BgRatio	p.value	p.adjust	q.value
GO:0006836	neurotransmitter transport	22/945	191/15568	0.002881322	0.04909408	0.04278198

2 equivalent ways of representing and computing



Hypergeometric distribution

$$P\left(X \geq k\right) = 1 - \sum_{i=0}^{k-1} \frac{\binom{M}{i}\binom{N-M}{n-i}}{\binom{N}{n}}$$

Contingency table

	in BP set	not in BP set
in Gene set	k = 22	923
not in Gene set	169	14454

n

M

N

Right-tailed fisher exact test

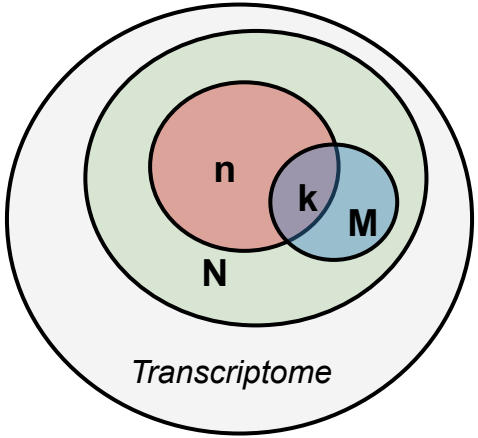
= *Testing Independence*

Wieder, C. et al. Pathway analysis in metabolomics: Recommendations for the use of over-representation analysis. PLoS Comput Biol 17

ORA: Back to the fundamentals

ID	Description	GeneRatio	BgRatio	p.value	p.adjust	q.value
GO:0006836	neurotransmitter transport	22/945	191/15568	0.002881322	0.04909408	0.04278198

2 equivalent ways of representing and computing



Contingency table

	in BP set	not in BP set
in Gene set	k = 22	923
not in Gene set	169	14454

n

M

N

Hypergeometric distribution → p.value=0.002881322 ← *Right-tailed fisher exact test*

= *Testing Independence*

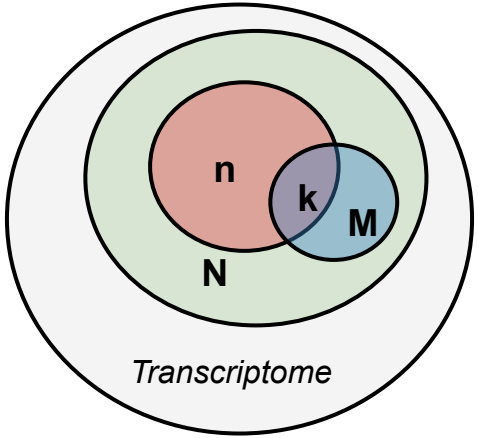
$$P\left(X \geq k\right) = 1 - \sum_{i=0}^{k-1} \frac{\binom{M}{i}\binom{N-M}{n-i}}{\binom{N}{n}}$$

Wieder, C. et al. Pathway analysis in metabolomics: Recommendations for the use of over-representation analysis. PLoS Comput Biol 17

ORA: Back to the fundamentals

ID	Description	GeneRatio	BgRatio	p.value	p.adjust	q.value
GO:0006836	neurotransmitter transport	22/945	191/15568	0.002881322	0.04909408	0.04278198

2 equivalent ways of representing and computing



Contingency table

	in BP set	not in BP set
in Gene set	k = 22	923
not in Gene set	169	14454

n

M

N

Hypergeometric distribution

$$P\left(X \geq k\right) = 1 - \sum_{i=0}^{k-1} \frac{\binom{M}{i}\binom{N-M}{n-i}}{\binom{N}{n}}$$

p.value=0.002881322

Right-tailed fisher exact test

= Testing Independence

+ correction for multiple-tests

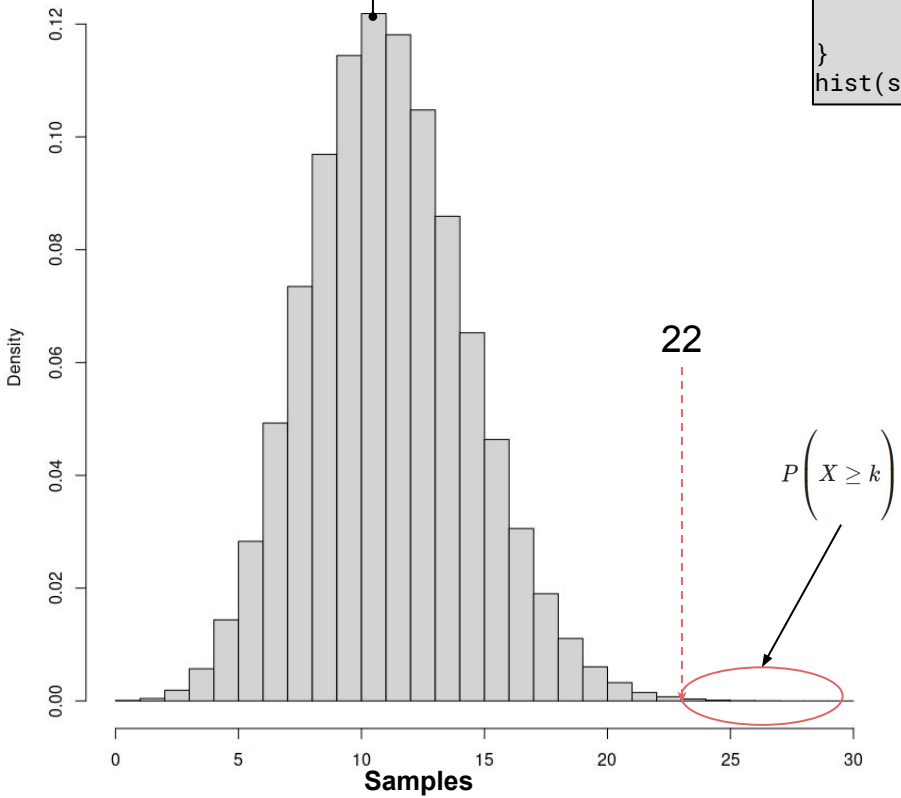
q.value

Wieder, C. et al. Pathway analysis in metabolomics: Recommendations for the use of over-representation analysis. PLoS Comput Biol 17

ORA: Intuition via random sampling

$$\frac{191}{15568} \times \frac{945}{15568} \times 15568 \approx 11.6$$

$P(\text{in Gene set})$ $P(\text{in BP set})$

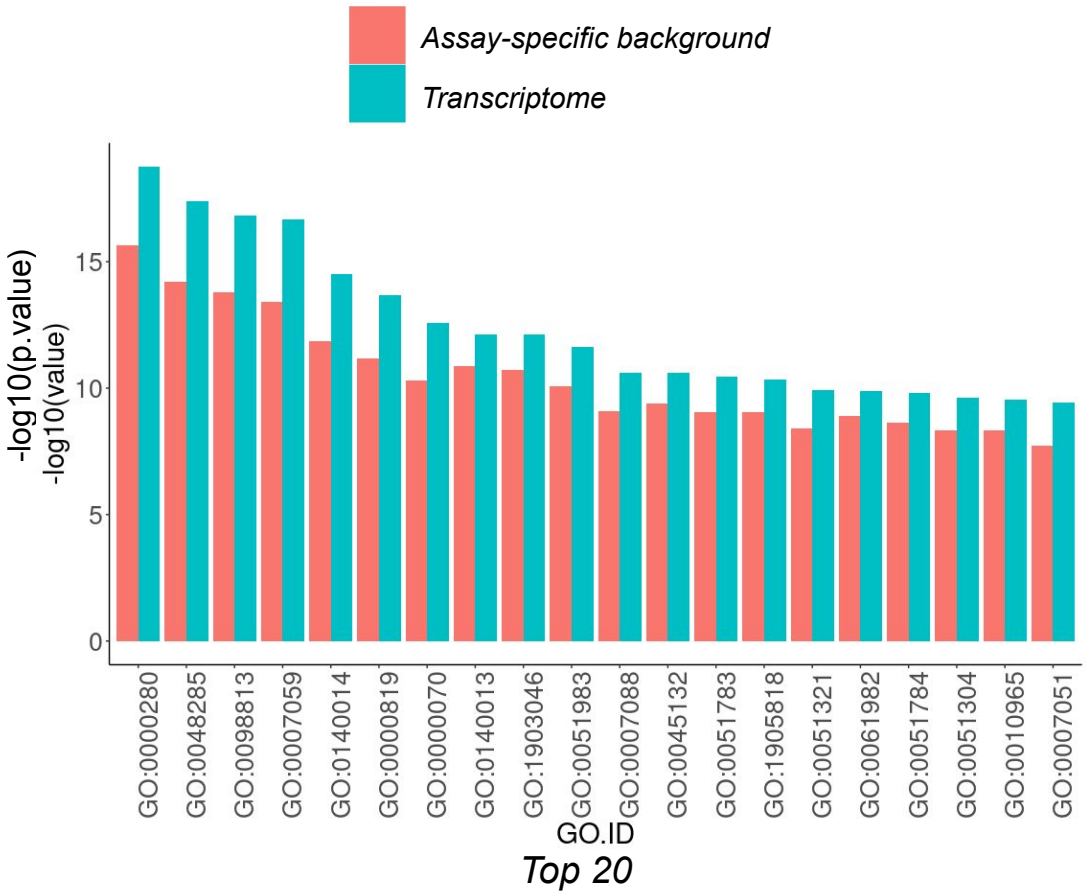



```
N_SAMPLES <- 1000000
samples <- vector(mode = "numeric", length = N_SAMPLES)
for(i in 1:N_SAMPLES){
      s <- sample(size = 945, x = universe, replace = F)
      samples[i] <- sum(s %in% GO.0006836)
}
hist(samples, freq = F, breaks = seq(0,30,1))
```

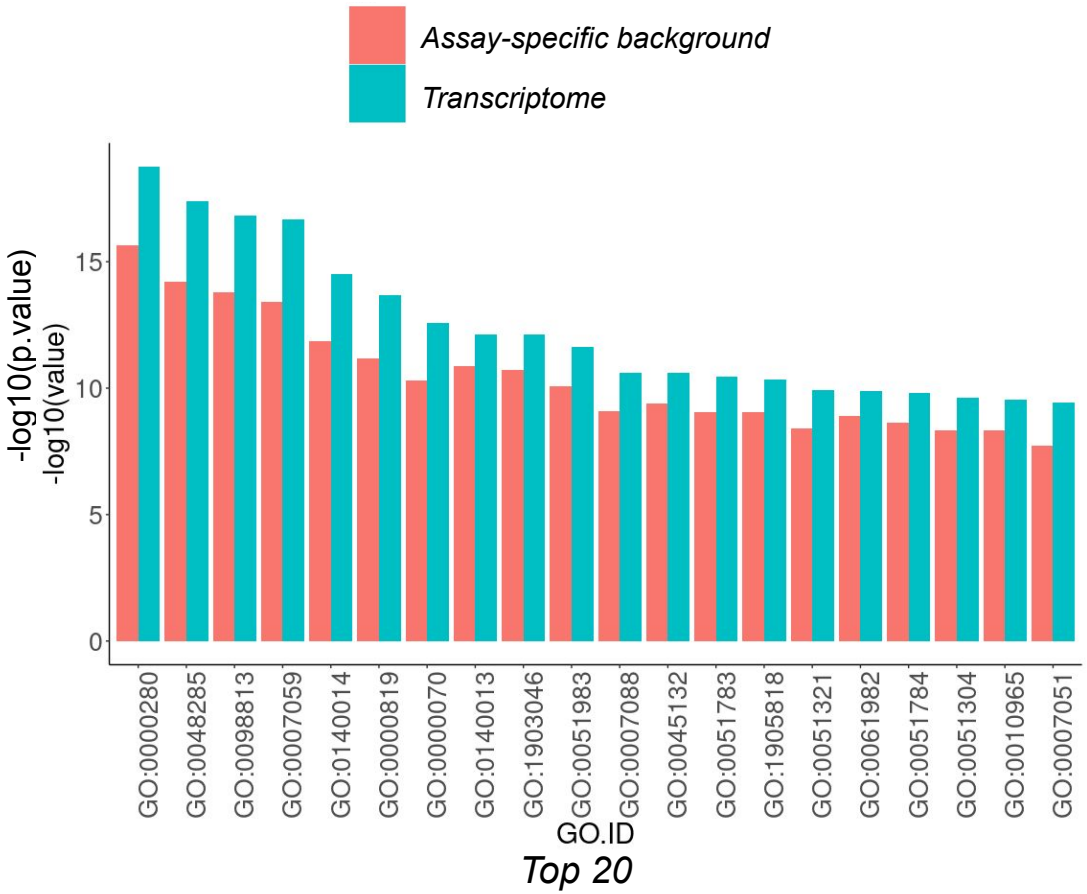
*What proportion of r**andom sets** sampled from the universe show more than 22 genes included in GO:0006836 ?*

```
estimate = 0.002816
```

ORA: The Impact of the *universe* definition



ORA: The Impact of the *universe* definition



Leads to overestimation of the p-value

+

Order ~ preserved

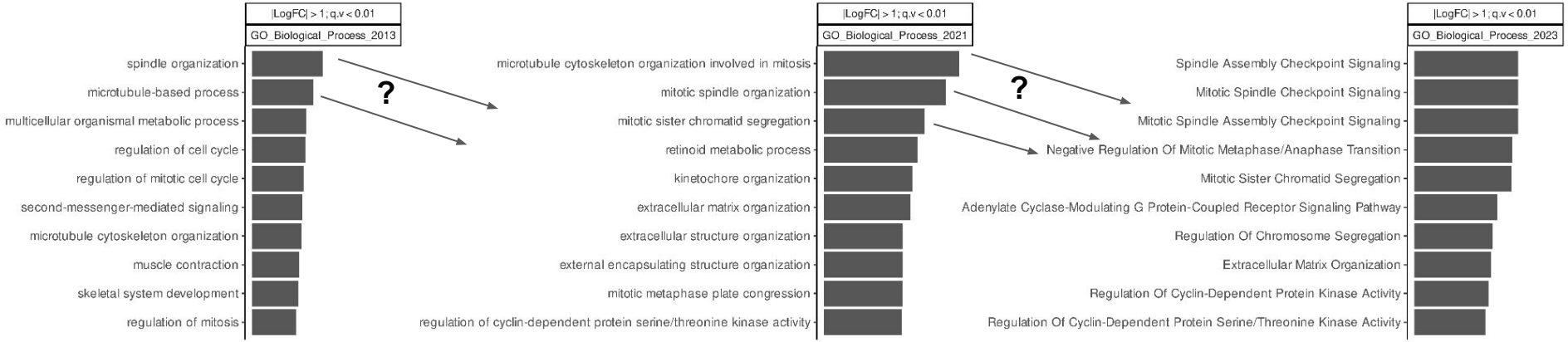
↓

Increase false positive enriched terms

133 vs 218 enriched GO terms
(q.value <= 1e.3)

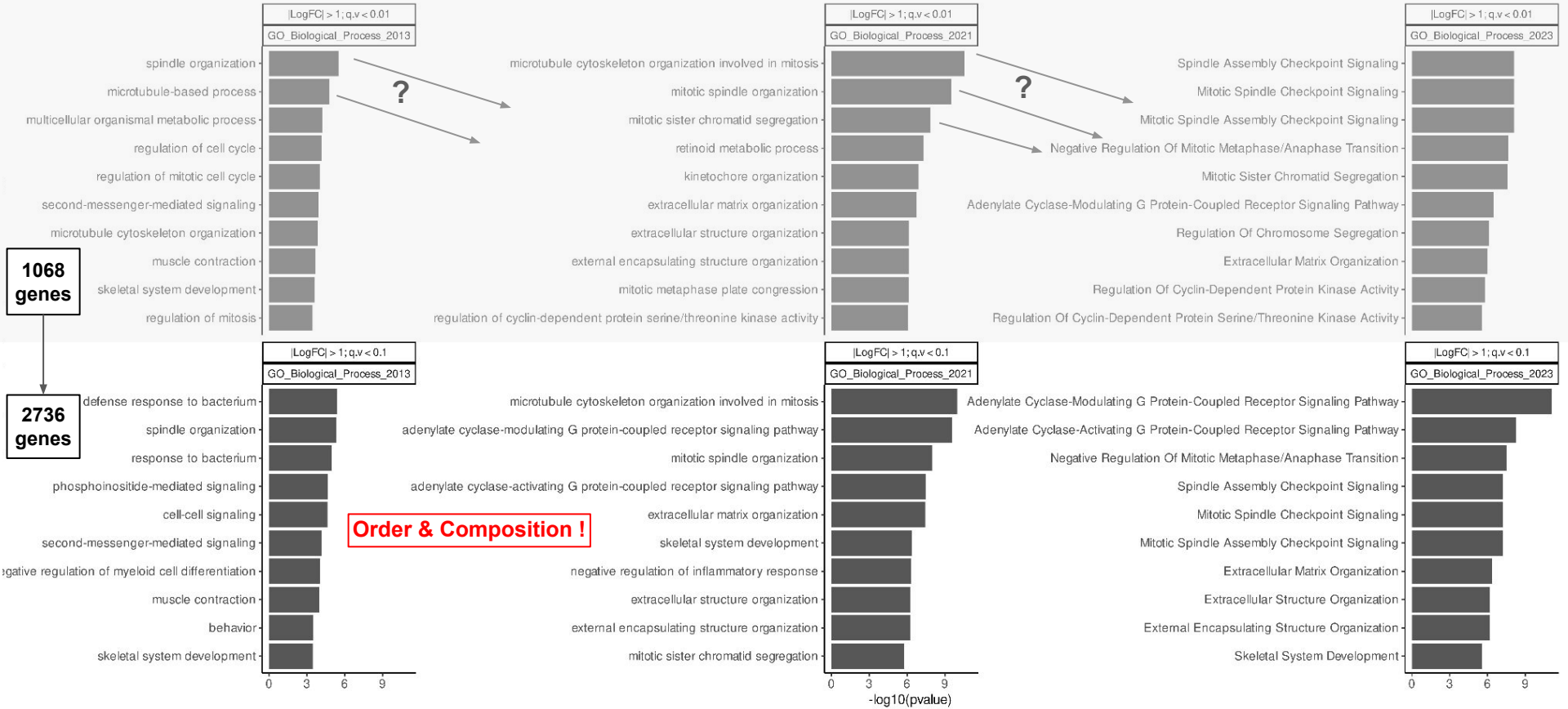
ORA: Impact of the database and gene set thresholds choices

Enrichr



ORA: Impact of the database and gene set thresholds choices

Enrichr



ORA: Impact of the database and gene set thresholds choices

How many significantly enriched Biological Processes ? (q.value < 0.01)

		GO BP 2013	**GO BP 2021**	**GO BP 2023**
2736 genes	**q.value < 0.1**	8	37	30
1068 genes	**q.value < 0.01**	4	44	40

Thresholds and database choices also have an impact of the number of enriched terms

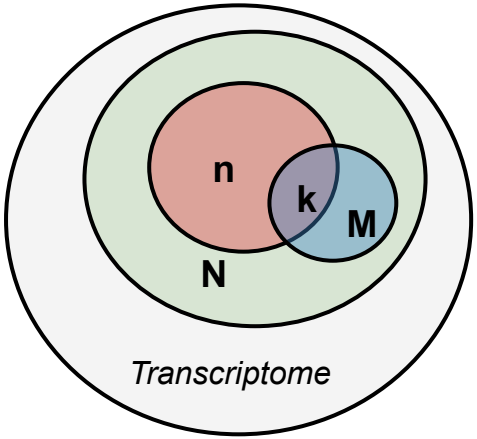
ORA: Several biases

- All parameters are important
  - a universe set (size = **N**)
  - a set of interest (size = **n**)
  - a reference set (size = **M**)

Unspecific background set can create false positives

Selection thresholds are important:
- Too large = noisy detection
- Too small = low detection

Reference database and versions can an impact of results



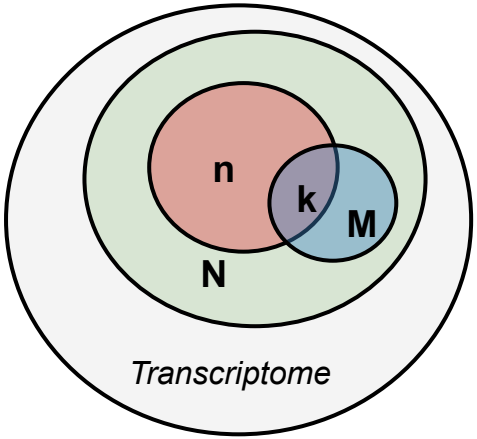
ORA: Several biases

- All parameters are important
  - a universe (size = **N**)
  - a set of interest (size = **n**)
  - a reference set (size = **M**)

Unspecific universe (*background set*) can create false positives

Selection thresholds are important:
- Too large = noisy detection
- Too small = low detection

Reference database and versions can an impact of results



Always specify the background set, the applied thresholds and the database version → Good results = **Reproducible** results

ORA: It can be any gene sets - WGCNA example

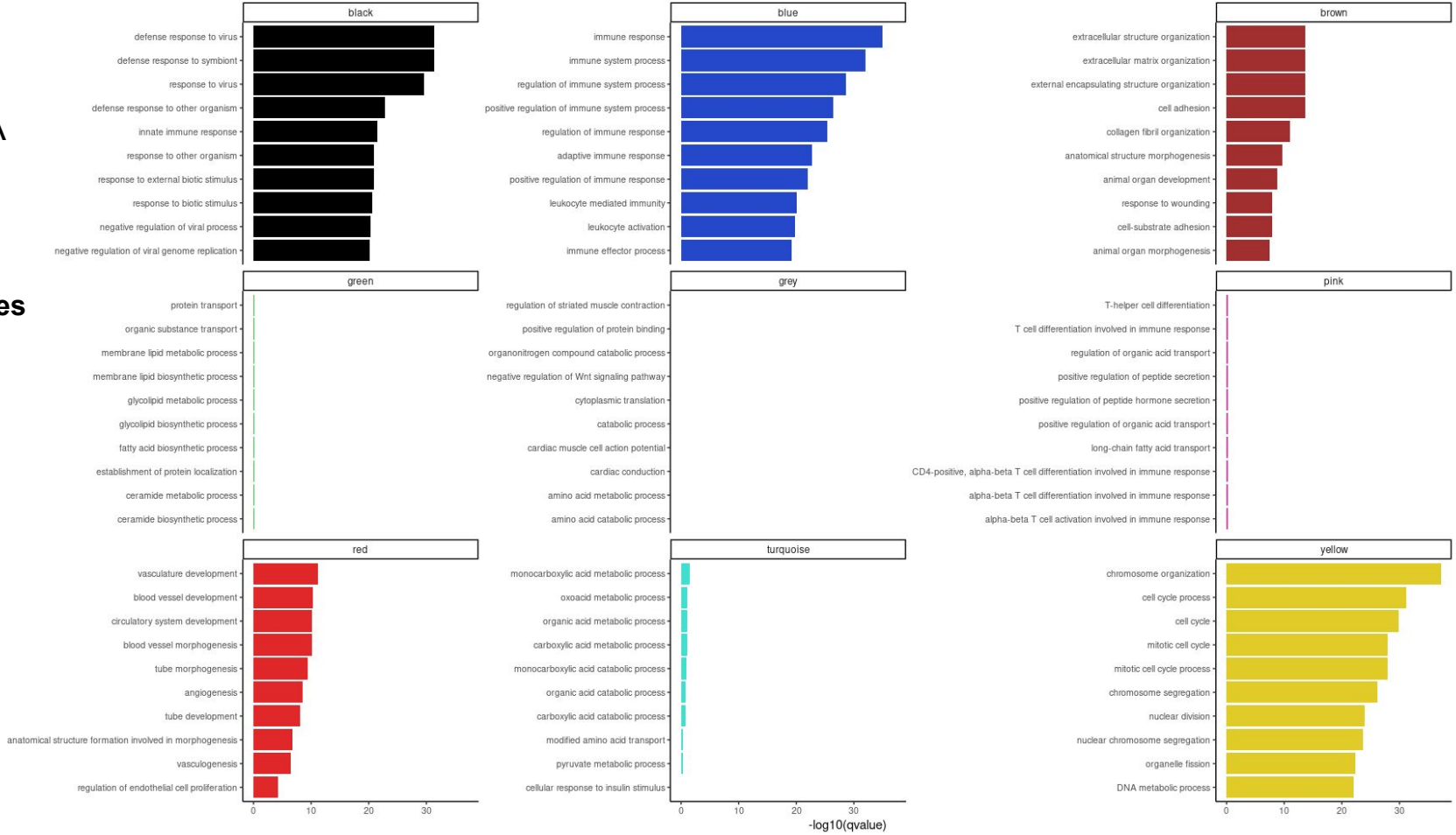
TP WGCNA

↓

9 gene modules

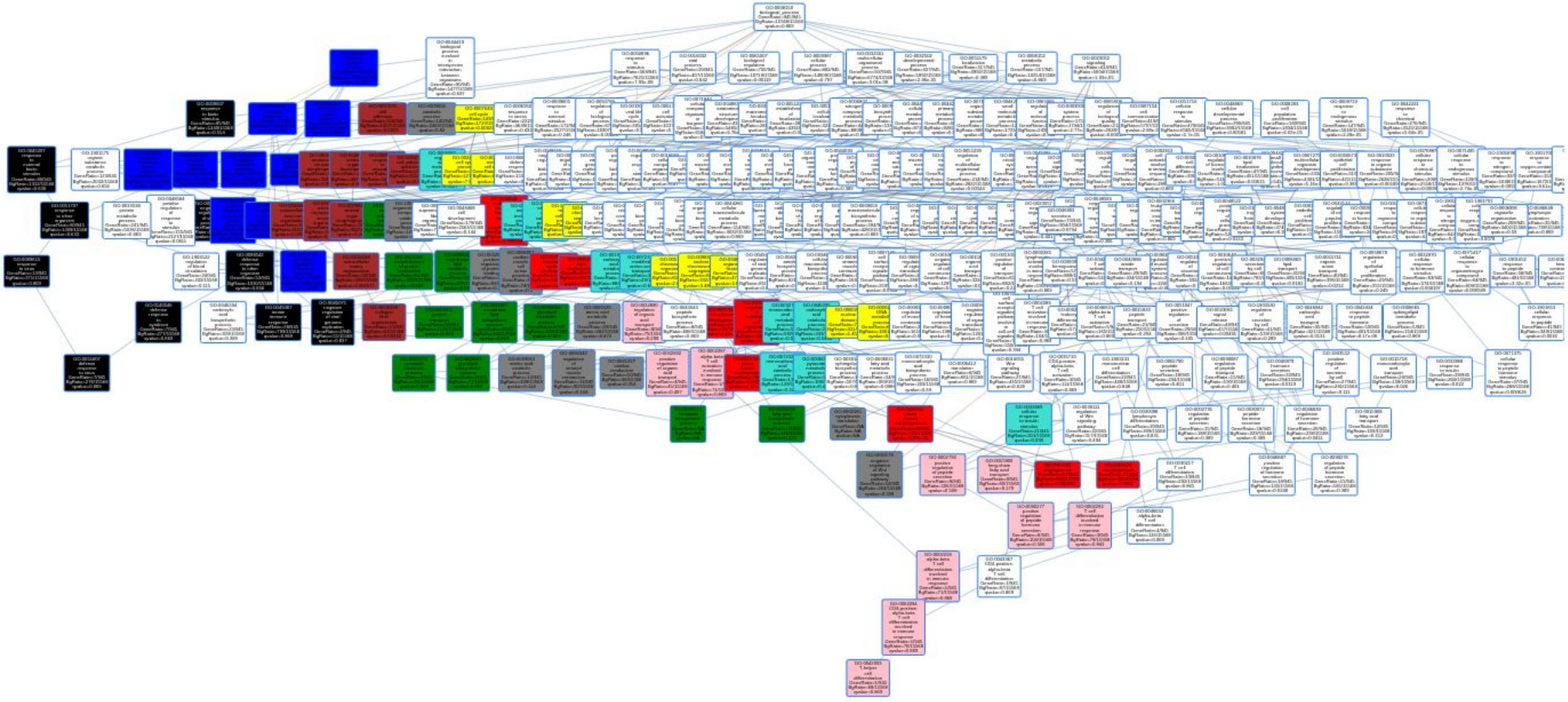
↓

Enrichment analysis on modules



ORA: It can be any gene sets - WGCNA example mapping on DAG

Mapping ofthe top 10 per modules on the GO BP DAG



GSEA: A Function scoring method

- Recall of the impact of the threshold (q.value & logFC) on the ORA results

→ **GSEA**

- All genes are not equivalent: sign and intensity of variation

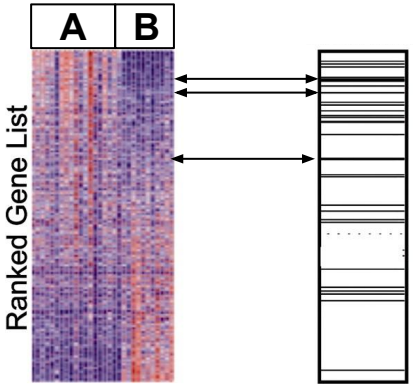
Subramanian, A. et al., 2005. Gene set enrichment analysis: A knowledge-based approach for interpreting genome-wide expression profiles.
James H Joly et al., 2019, Differential Gene Set Enrichment Analysis: a statistical approach to quantify the relative enrichment of two gene sets, *Bioinformatics*.

GSEA: A Function scoring method (1)

- Recall of the impact of the threshold (q.value & logFC) on the ORA results

→ **GSEA**

- All genes are not equivalent: sign and intensity of variation



possible metrics: Log2FC, signed p-values, etc.

Genes associated with a GO term

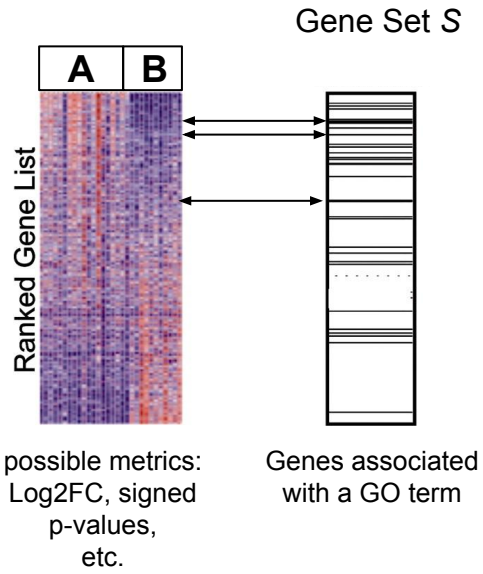
Subramanian, A. et al., 2005. Gene set enrichment analysis: A knowledge-based approach for interpreting genome-wide expression profiles.
James H Joly et al., 2019, Differential Gene Set Enrichment Analysis: a statistical approach to quantify the relative enrichment of two gene sets, *Bioinformatics*.

GSEA: A Function scoring method (1)

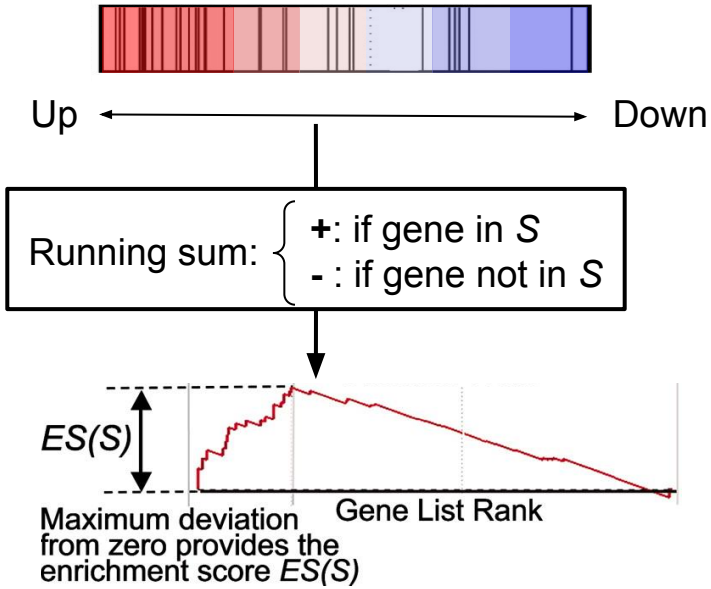
- Recall of the impact of the threshold (q.value & logFC) on the ORA results

→ **GSEA**

- All genes are not equivalent: sign and intensity of variation



Compute the **E**nrichment **S**core (ES)



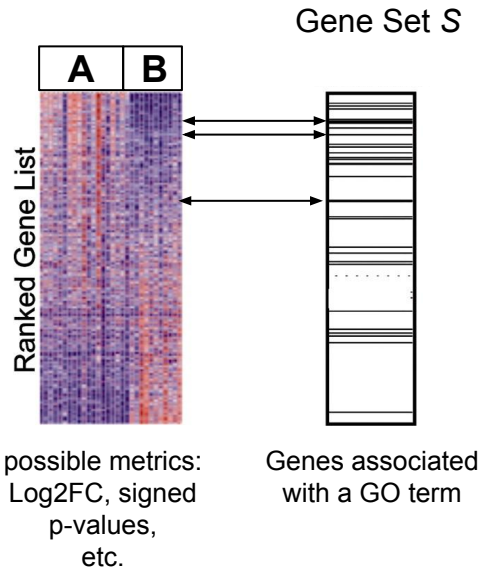
Subramanian, A. et al., 2005. Gene set enrichment analysis: A knowledge-based approach for interpreting genome-wide expression profiles.
James H Joly et al., 2019, Differential Gene Set Enrichment Analysis: a statistical approach to quantify the relative enrichment of two gene sets, *Bioinformatics*.

GSEA: A Function scoring method (1)

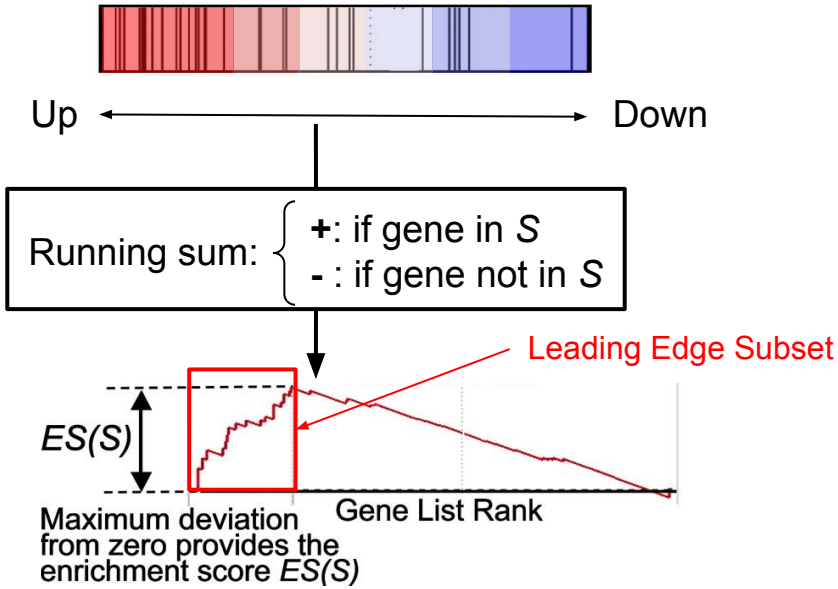
- Recall of the impact of the threshold (q.value & logFC) on the ORA results

→ **GSEA**

- All genes are not equivalent: sign and intensity of variation



Compute the **E**nrichment **S**core (ES)



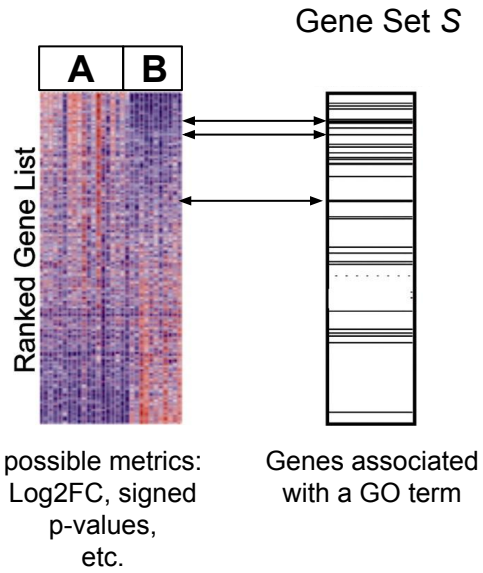
Subramanian, A. et al., 2005. Gene set enrichment analysis: A knowledge-based approach for interpreting genome-wide expression profiles.
James H Joly et al., 2019, Differential Gene Set Enrichment Analysis: a statistical approach to quantify the relative enrichment of two gene sets, *Bioinformatics*.

GSEA: A Function scoring method (1)

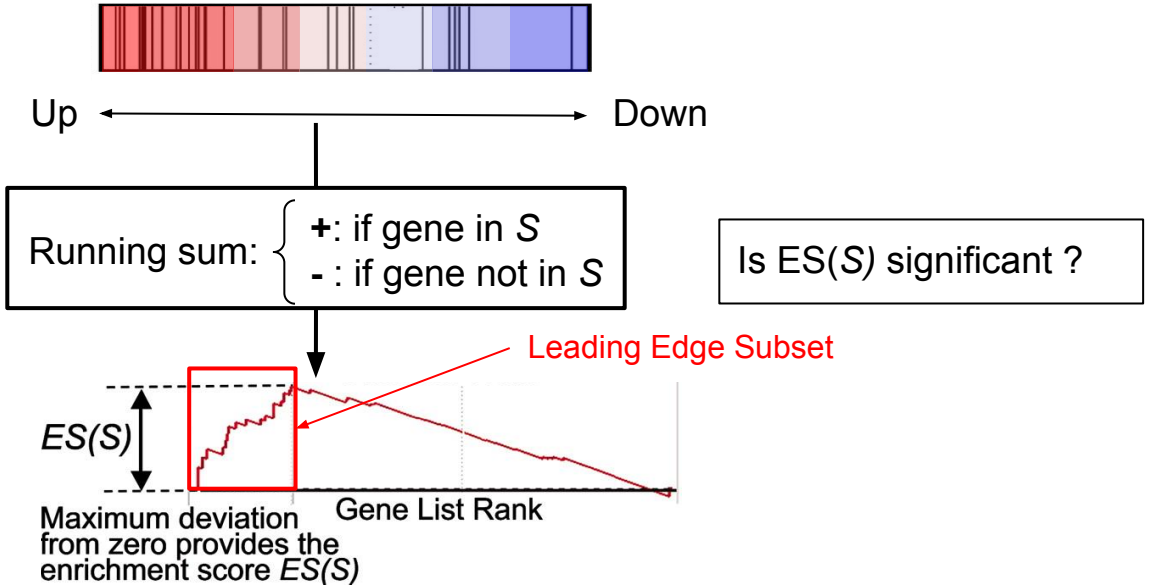
- Recall of the impact of the threshold (q.value & logFC) on the ORA results

→ **GSEA**

- All genes are not equivalent: sign and intensity of variation



Compute the **E**nrichment **S**core (ES)



Subramanian, A. et al., 2005. Gene set enrichment analysis: A knowledge-based approach for interpreting genome-wide expression profiles.
James H Joly et al., 2019, Differential Gene Set Enrichment Analysis: a statistical approach to quantify the relative enrichment of two gene sets, *Bioinformatics*.

GSEA: A Function scoring method (2)

Is ES(*S)* significant ?

- in unweighted settings (first version of GSEA): exact p-value estimation with KS-test

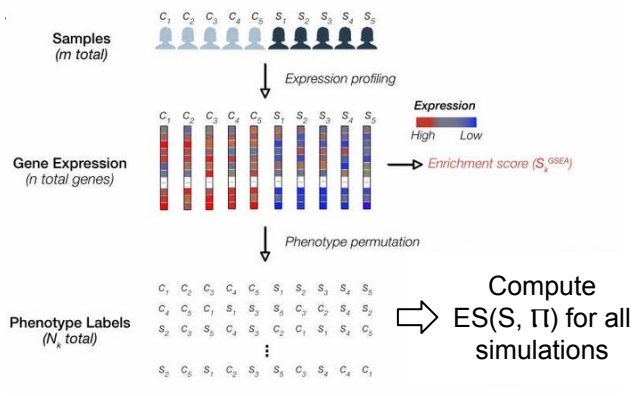
https://www.pathwaycommons.org/guide/primers/data_analysis/gsea/

GSEA: A Function scoring method (2)

Is ES(*S*) significant ?

- in unweighted settings (first version of GSEA): exact p-value estimation with KS-test
- in weighted settings (common): empirical estimation via permutation test (simulations Π)

Phenotype permutation
(better if enough sample)



Compute ES(S, Π) for all simulations

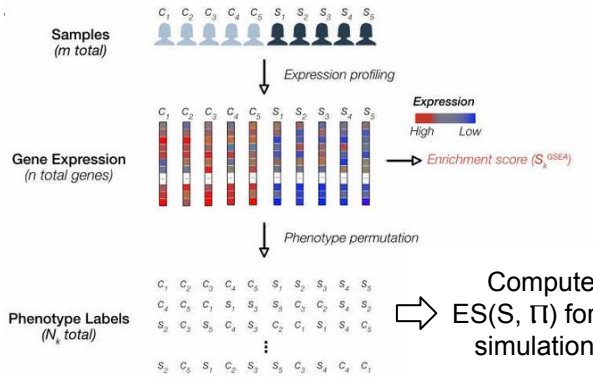
https://www.pathwaycommons.org/guide/primers/data_analysis/gsea/

GSEA: A Function scoring method (2)

Is ES(*S*) significant ?

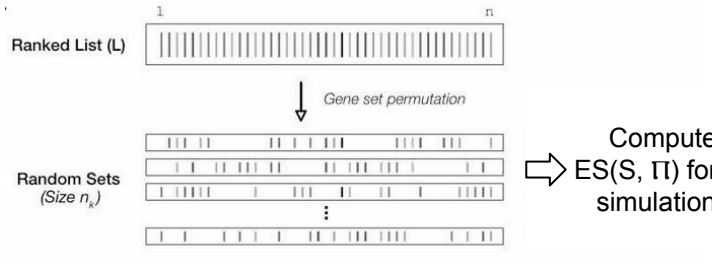
- in unweighted settings (first version of GSEA): exact p-value estimation with KS-test
- in weighted settings (common): empirical estimation via permutation test (simulations Π)

Phenotype permutation (better if enough sample)



Compute ES(S, Π) for all simulations

Gene set Permutation (most common but less efficient)



Compute ES(S, Π) for all simulations

https://www.pathwaycommons.org/guide/primers/data_analysis/gsea/

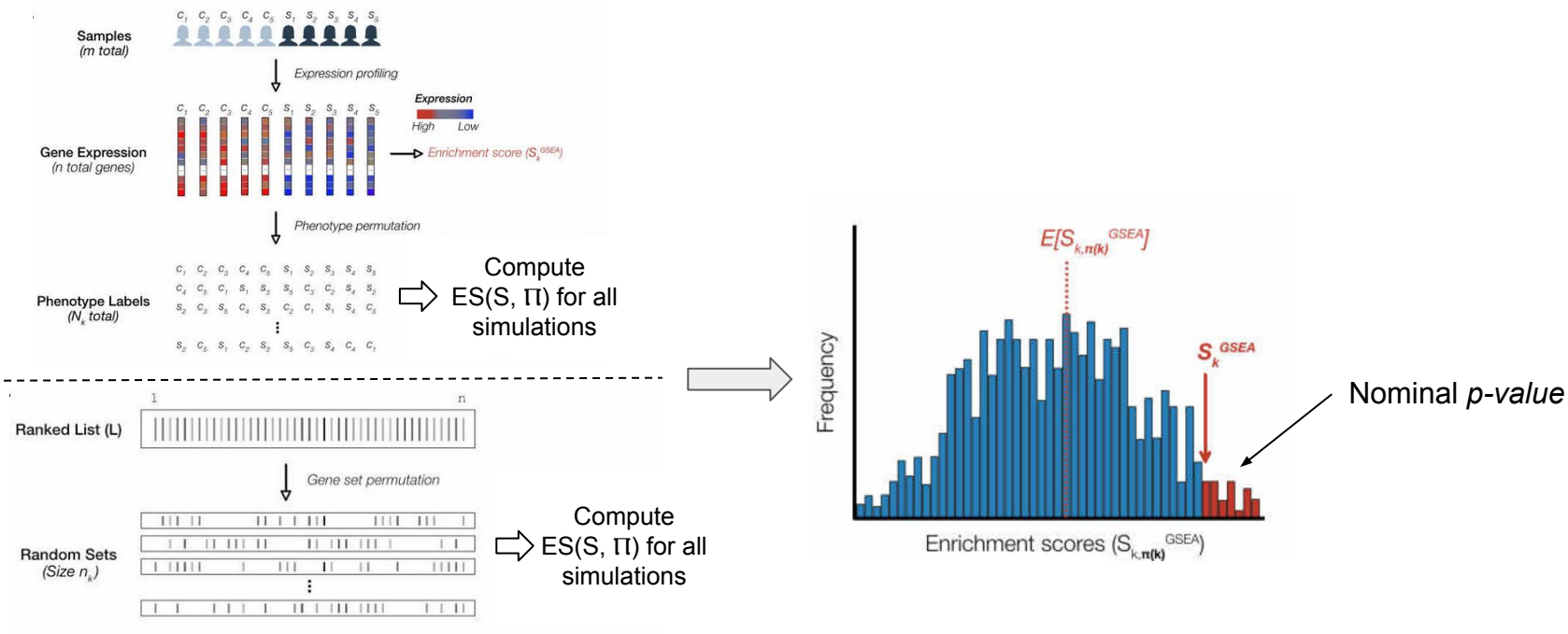
GSEA: A Function scoring method (2)

Is ES(*S*) significant ?

- in unweighted settings (first version of GSEA): exact p-value estimation with KS-test
- in weighted settings (common): empirical estimation via permutation test (simulations Π)

Phenotype permutation (better if enough sample)

Gene set Permutation (most common but less efficient)



https://www.pathwaycommons.org/guide/primers/data_analysis/gsea/

GSEA: A Function scoring method (3)

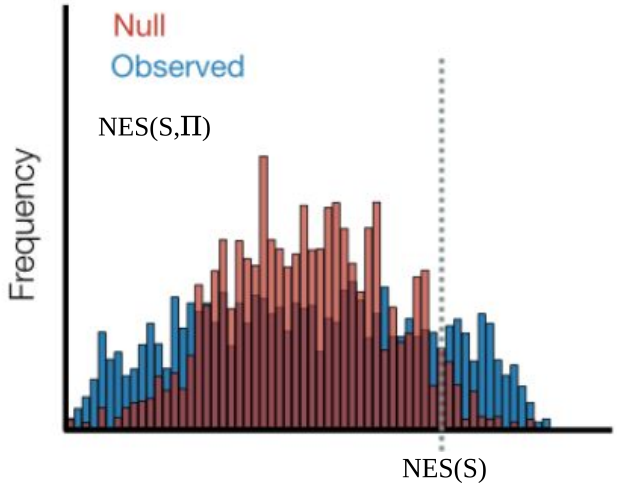
- How to account for Gene set size differences ?

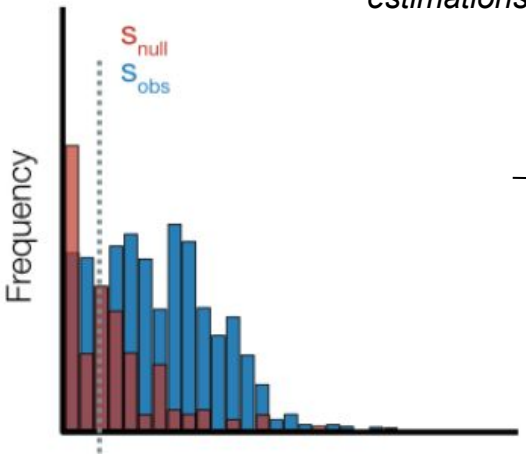
NES (**N**ormalised **E**nrichment **S**core)

$$\mathrm{NES(S)} = \frac{ES(S)}{E[ES(S,\Pi]}$$ ← Gives the direction of regulation + correct for gene set size + signed

- Multiple-test correction: FDR estimation

estimations separated by NES signs



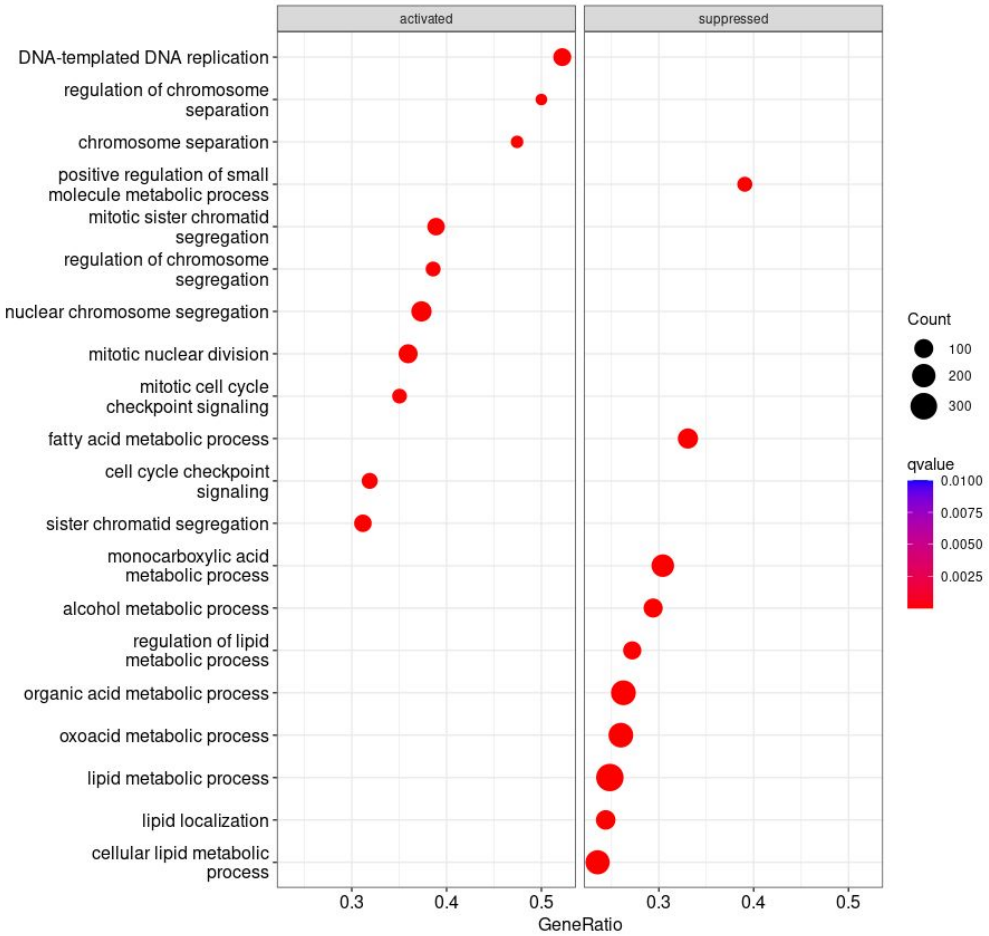


→ $F\hat{D}R$

https://www.pathwaycommons.org/guide/primers/data_analysis/gsea/

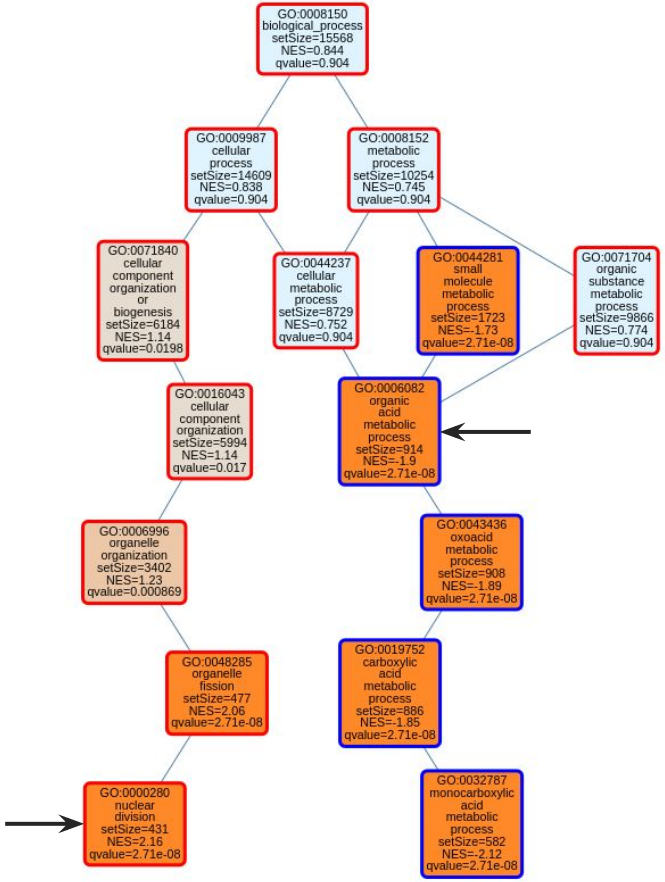
GSEA: A Function scoring method - example

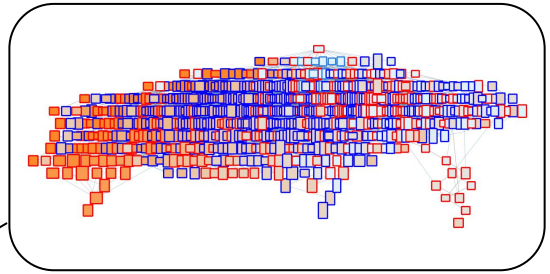
- No need of a cutoff on qvalue or LogFC, just a ranking metric !
- Results are separated between over and under expression BP
- Leading Edge subset can help to identify key actors
- However, same biases apply for database choices !

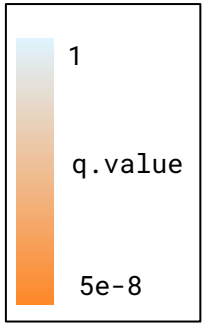


GSEA: Visualisation - example

Up and Down regulated branch of the DAG



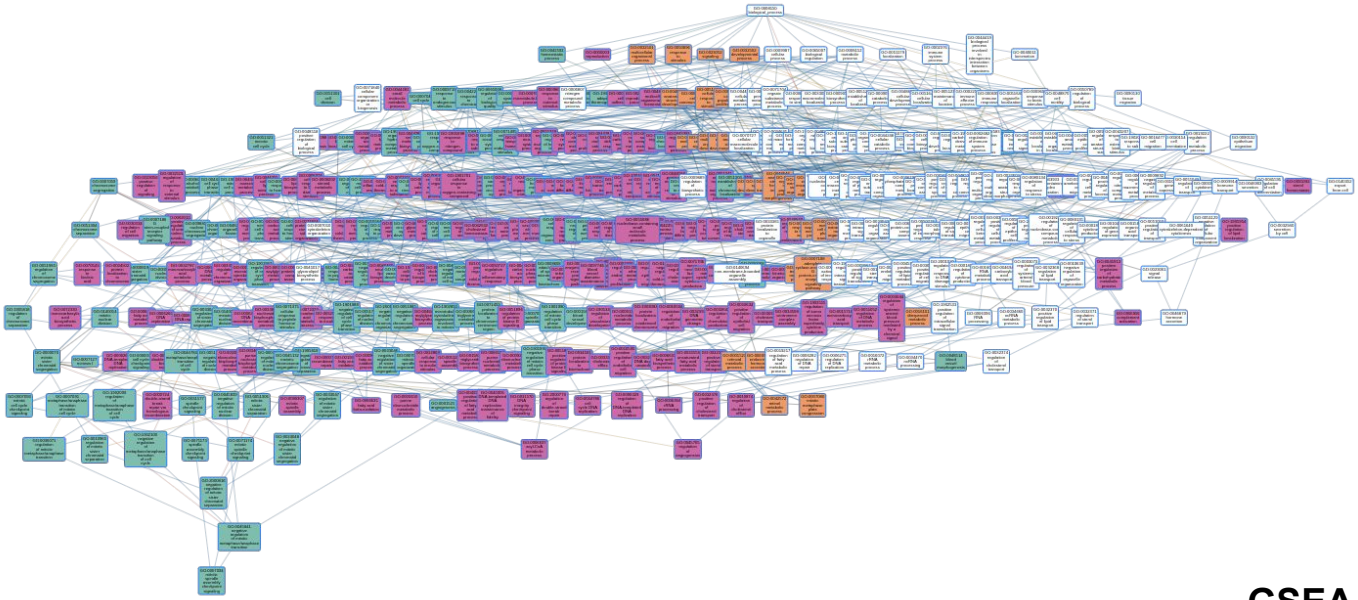




up-regulated

down-regulated

ORA vs GSEA: A visual comparison on the GO DAG graph



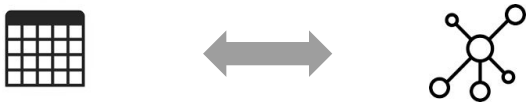
q.value < 1e-3

- **Only GSEA**
- **Both ORA and GSEA**
- **Only ORA**

GSEA is more sensitive !

Extend contextualisation with Biomedical Knowledge Graph

What is a Graph ?
A graph is defined by a set of **nodes** and **edges**

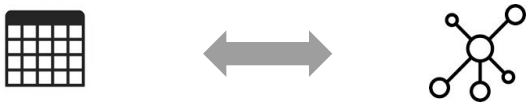


Attributes/Properties Relations/Paths

Different questions, different visualisation, different methods

Extend contextualisation with Biomedical Knowledge Graph

What is a Graph ?
A graph is defined by a set of **nodes** and **edges**



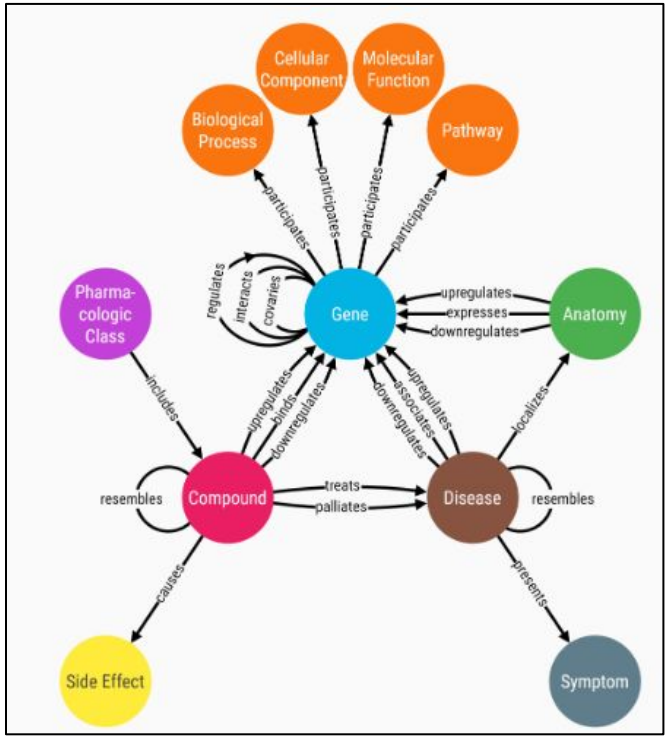
Attributes/Properties Relations/Paths

Different questions, different visualisation, different methods

What is a Biomedical Knowledge Graph ?
It's a graph describing biomedical entities and their relations

- Different Model: RDF (in Semantic Web) and LPG (Labeled Property Graph)
- Relations are stored at the individual record level
- Efficient for complex information extraction
- Examples: Hetionet, Wikidata, PharmKG, FORUM, etc.

Example of Hetionet



How to request a Knowledge Graph (in Neo4J)

neo4j Cypher

3 main clauses:

- MATCH: Specify the graph pattern
- WHERE: Add restrictions to the nodes or edges properties
- RETURN: Define what is included in the results

How to request a Knowledge Graph (in Neo4J)

neo4j Cypher

3 main clauses:

- MATCH: Specify the graph pattern
- WHERE: Add restrictions to the nodes or edges properties
- RETURN: Define what is included in the results

How to write:

nodes: `(variable:Label)` *(Label is optional)*

edges: `-[variable:Label]->`

How to request a Knowledge Graph (in Neo4J)

neo4j Cypher

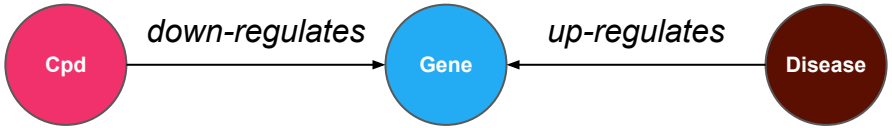
3 main clauses:

- MATCH: Specify the graph pattern
- WHERE: Add restrictions to the nodes or edges properties
- RETURN: Define what is included in the results

How to write:

nodes: `(variable:Label)` *(Label is optional)*

edges: `-[variable:Label]->`




```
MATCH (c:Compound)-[r1:DOWNREGULATES_CdG]->(g:Gene)<-[r2:UPREGULATES_DuG]-(d:Disease)
WHERE g.name IN [ “BRCA1”, “BRCA2”, … ]
RETURN c, r1, g, r2, d
```

Extend contextualisation with Biomedical Knowledge Graph

TP WGCNA

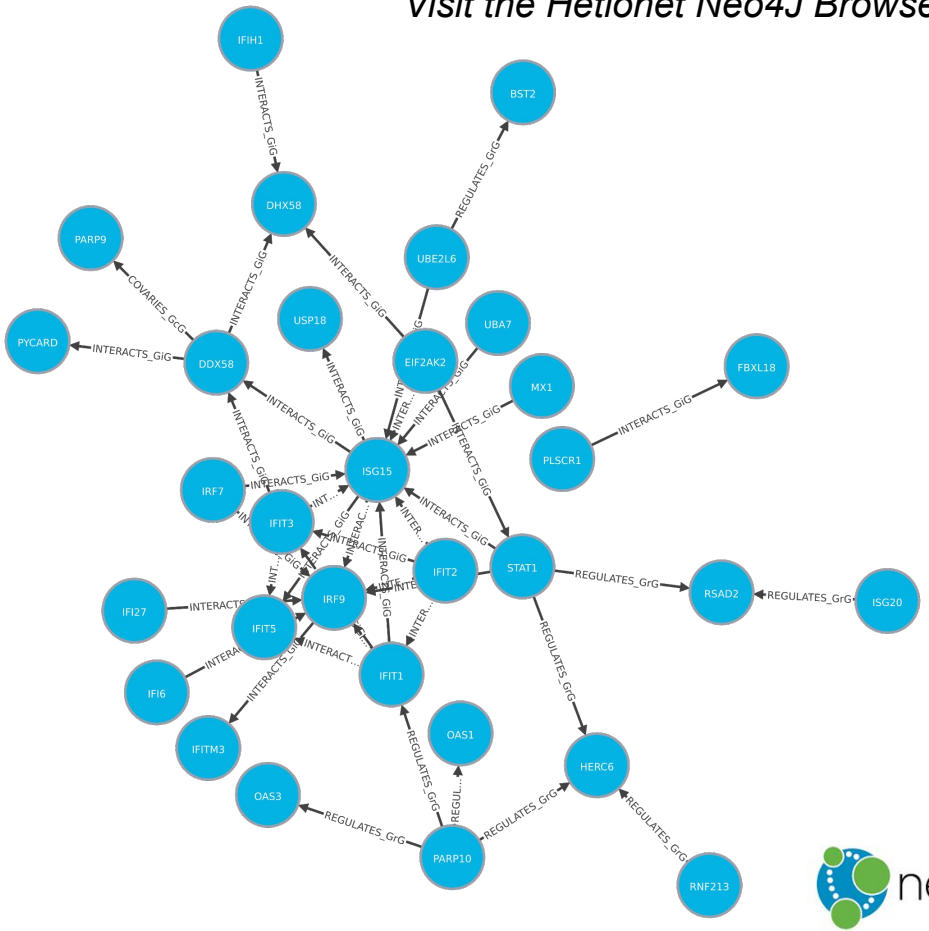
↓

Black gene module

↓

Explore relations between genes in a module

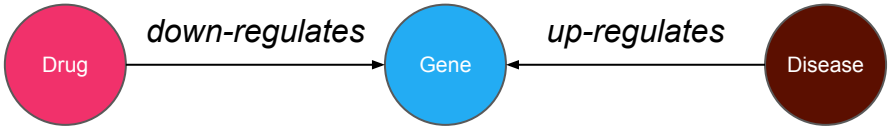
Visit the Hetionet Neo4J Browser



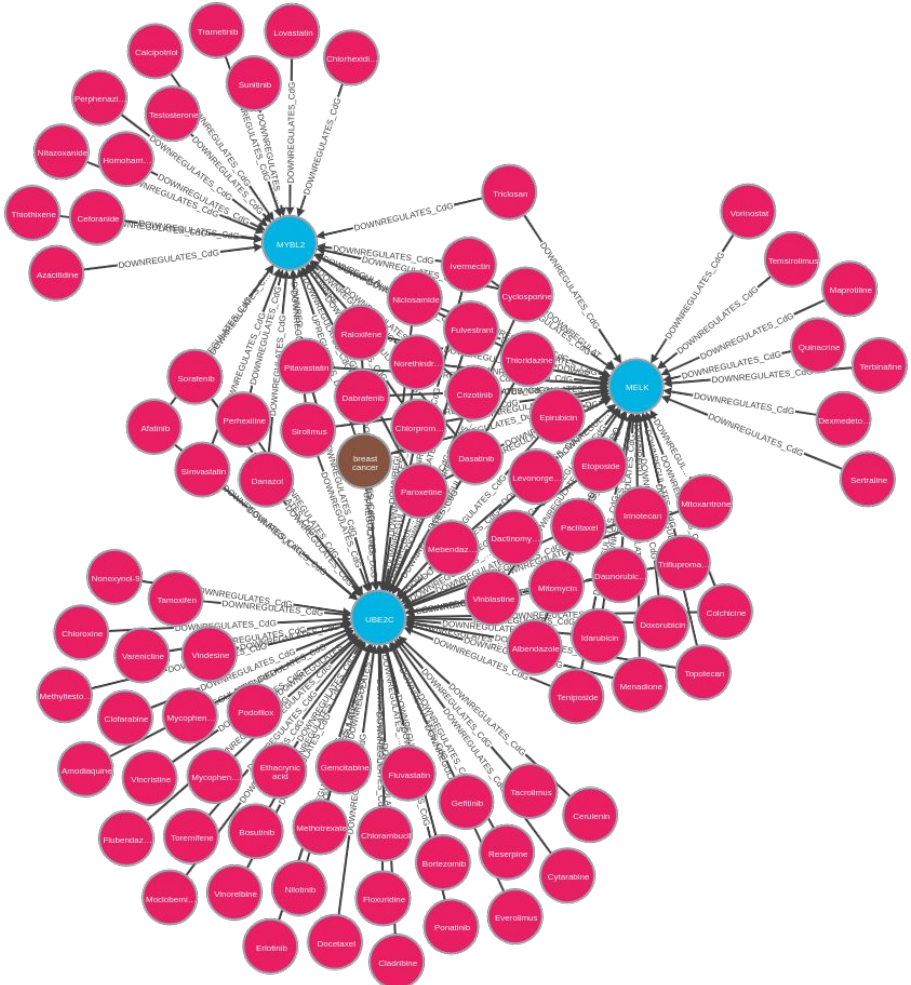
Extend contextualisation with Biomedical Knowledge Graph

A more complex path

By selecting only the up-regulated genes (LogFC > 5)



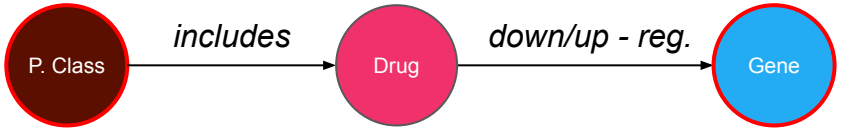
Drugs
Thioridazine
Doxorubicin
Dabrafenib
Teniposide

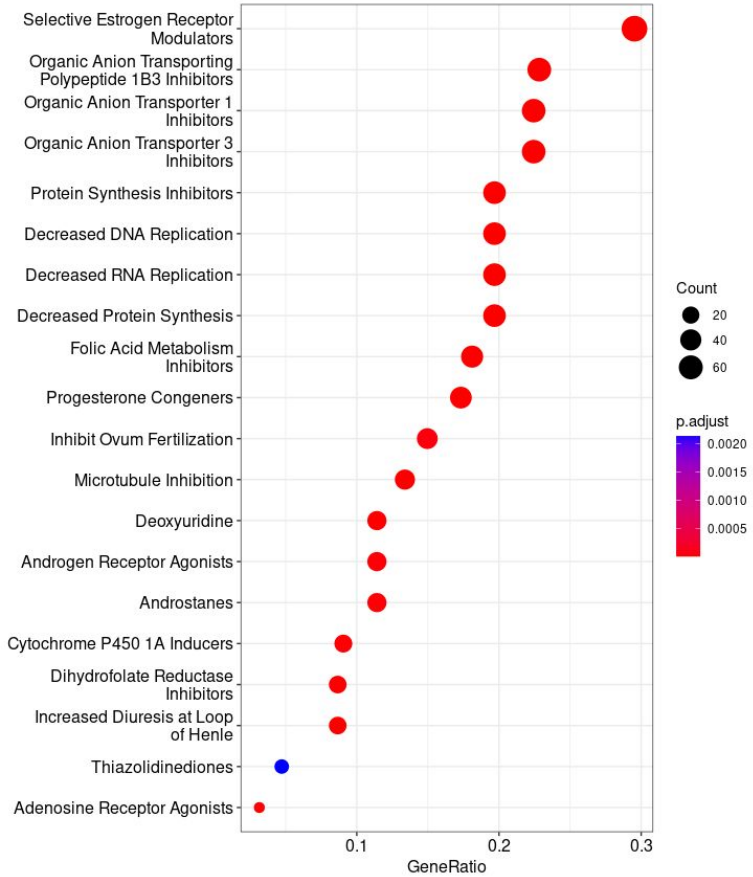


Extend contextualisation with Biomedical Knowledge Graph

Use a Biomedical Knowledge Graph to build a Enrichment custom background set

What class of drugs in enriched for their relation with the set of genes of interest ? (ORA)





Contextualisation of results: Conclusion

- Enrichment analyses (ORA or GSEA) are a powerful tool to suggest direction of interpretation and hypotheses

- ORA are simples and universals, but results can be affected by several biased: threshold, databases, universe.

- GSEA is not affected by thresholding are give more weight to the most discriminant genes

- Several biases remain:
  - Internal structure of pathway / interconnection between entities in a pathway → **Topological methods**
  - Overlap / interconnections between pathway
  - What about gene variants ?

- Biomedical KG can help to explore new connections between entities and with other entities

- They can also be used for building a custom background set.

- Build your own Biomedical KG ! Use **BioCypher**

The End

Ressources

- Enrichment analysis
  - Biblio & Resources
    - Wieder, C. et al. Pathway analysis in metabolomics: Recommendations for the use of over-representation analysis. PLoS Comput Biol 17
    - https://colab.research.google.com/drive/18pLzc_pv7Fpclotx4byYh9qMDjtnyG_u?usp=sharing
    - García-Campos, M.A. et al. 2015. Pathway Analysis: State of the Art. Front Physiol
    - Subramanian, A. et al., 2005. Gene set enrichment analysis: A knowledge-based approach for interpreting genome-wide expression profiles.
    - James H Joly et al., 2019, Differential Gene Set Enrichment Analysis: a statistical approach to quantify the relative enrichment of two gene sets, Bioinformatics.
    - https://www.pathwaycommons.org/guide/primers/data_analysis/gsea/
- Biomedical KG & Co.
  - LPG
    - *Hetionet*: https://het.io/about
    - *Drug Repurposing Knowledge Graph (DRKG)*: https://github.com/gnn4dr/DRKG
    - *BioKG*: https://github.com/dsi-bdi/biokg
    - PharmKG: https://academic.oup.com/bib/article/22/4/bbaa344/6042240
  - Web-Semantic
    - MetaNetX: https://www.metanetx.org/
    - Wikidata: https://www.wikidata.org/wiki/Wikidata:Main_Page
    - DisGeNeT: https://www.disgenet.org/
    - Rhea: https://www.rhea-db.org/
    - UniProt: https://www.uniprot.org/help/uniprotkb
  - Other resources
    - *Cypher Cheat Sheet*: https://neo4j.com/docs/cypher-cheat-sheet/5/auradb-enterprise/
    - *BioCypher*: https://biocypher.org/
    - Web-semantic MOOC: https://www.fun-mooc.fr/fr/cours/web-semantique-et-web-de-donnees/
    - Neo4J: https://www.youtube.com/channel/UCvze3hU6OZBkB1vkhH2lH9Q

Waagmeester, A. et al., 2020. Wikidata as a knowledge graph for the life sciences. eLife 9, e52614.