Enrichment analyses and results contextualisation with Knowledge Graphs

Summer School Multi-omics Data Analysis and Integration

Maxime DELMAS - 07/09/2023

idiap
RESEARCH INSTITUTE

Enrichment analyses

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

~ over-representation

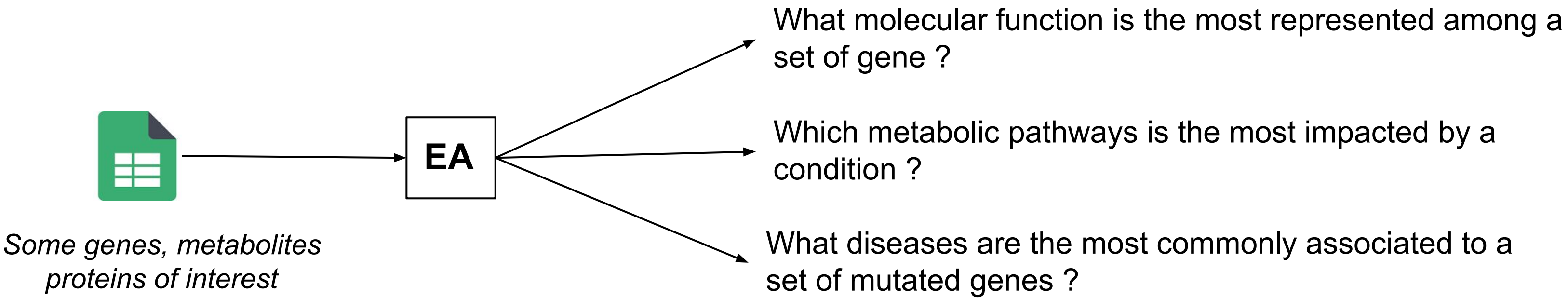
García-Campos, M.A. et al. 2015. Pathway Analysis: State of the Art. Front Physiol

Enrichment analyses

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

~ over-representation

- Give directions for results interpretation



EA

Some genes, metabolites proteins of interest

What molecular function is the most represented among a set of gene ?

Which metabolic pathways is the most impacted by a condition ?

What diseases are the most commonly associated to a set of mutated genes ?

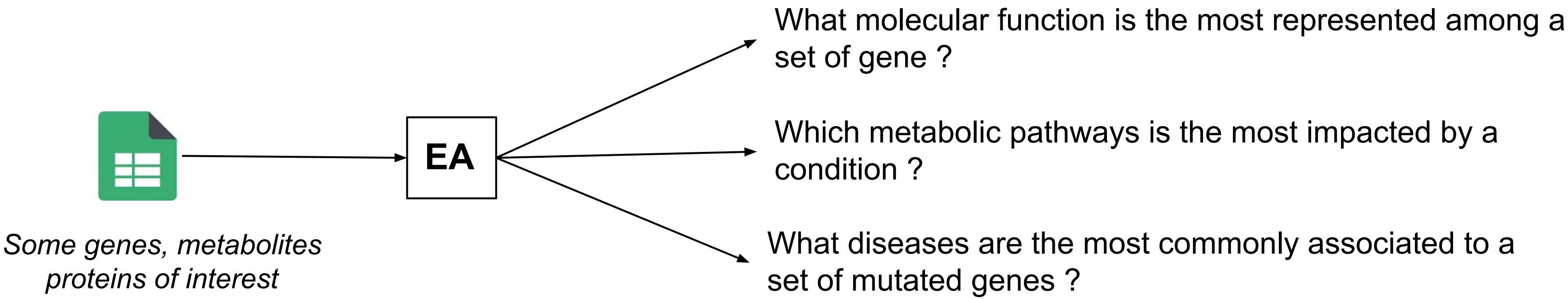
García-Campos, M.A. et al. 2015. Pathway Analysis: State of the Art. Front Physiol

Enrichment analyses

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

  ~ over-representation

- Give directions for results interpretation



Some genes, metabolites proteins of interest

What molecular function is the most represented among a set of gene ?

Which metabolic pathways is the most impacted by a condition ?

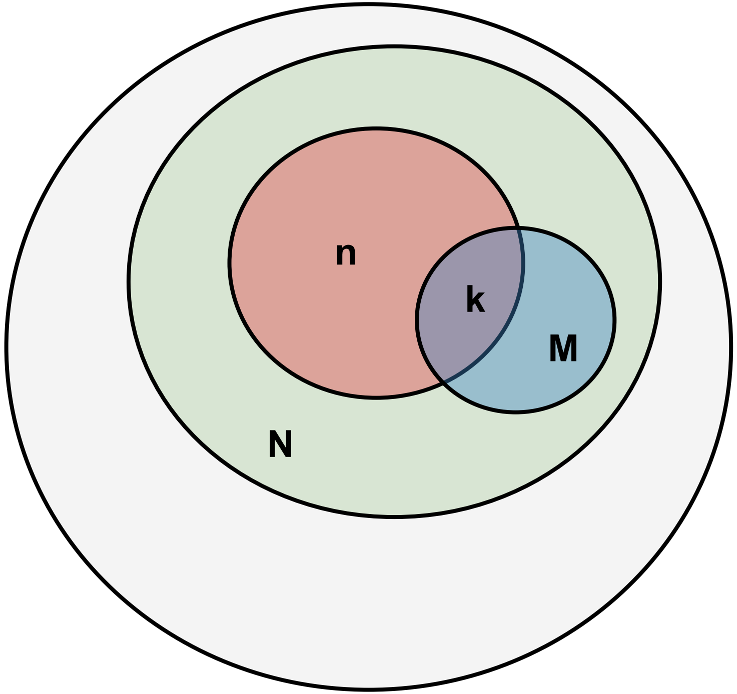
What diseases are the most commonly associated to a set of mutated genes ?

- 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

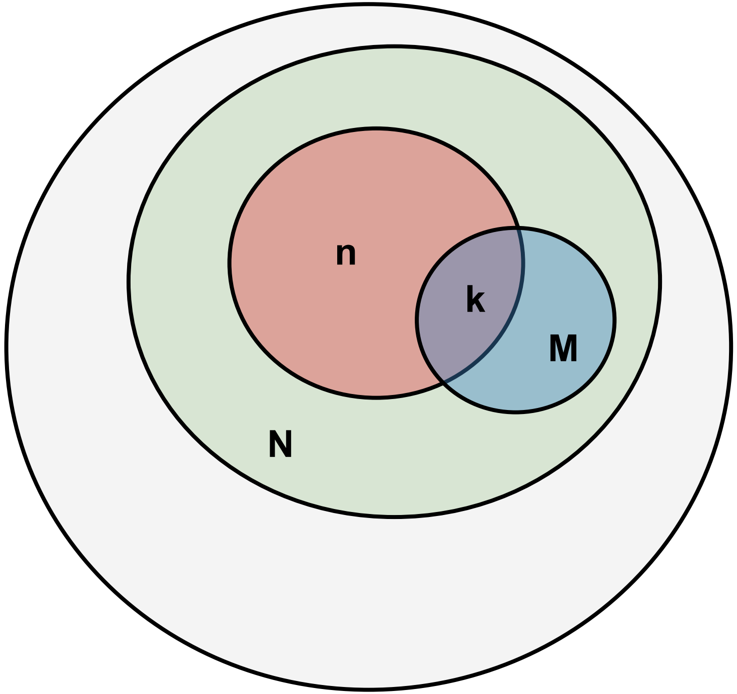
What does an ORA ? it compare **overlap** between **sets**.



- 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**

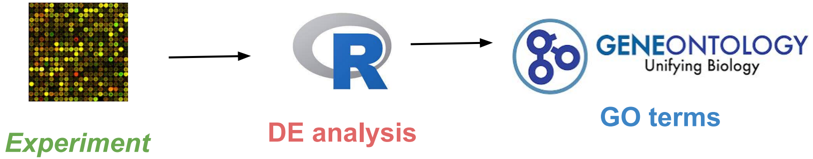
ORA: Over-Representation Analysis

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

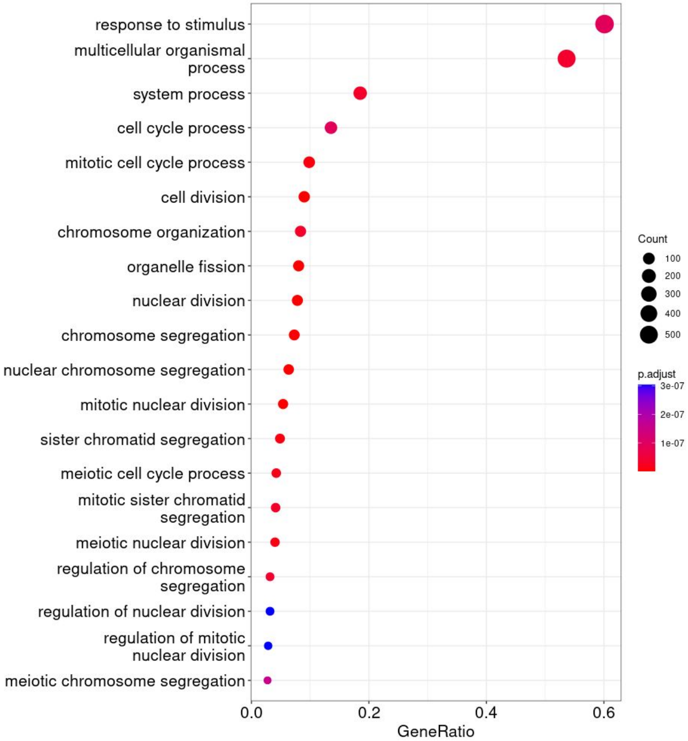
GO terms

- **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



Standard GO (*Biological processes*) Enrichment analysis

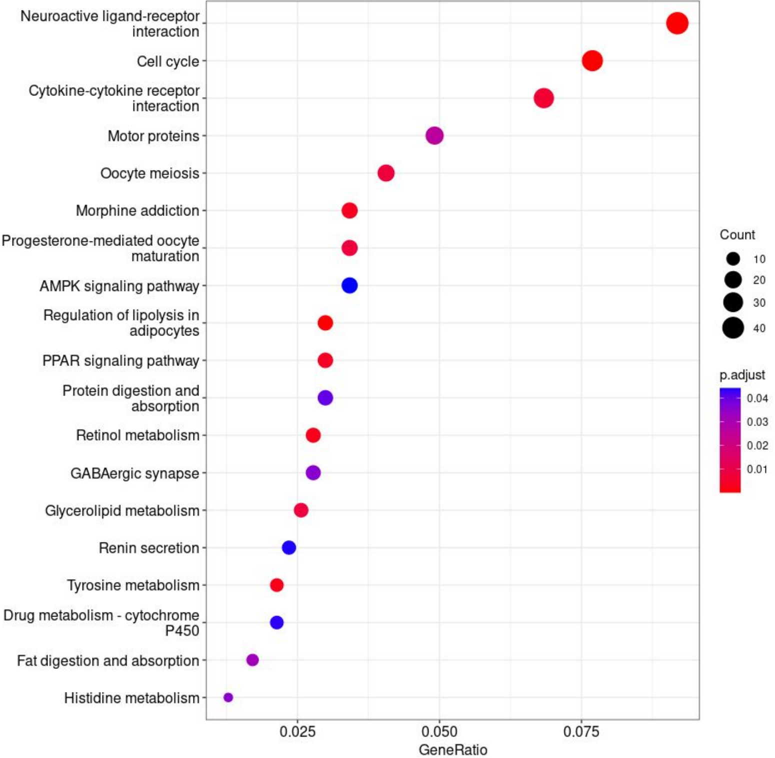
```
ego <- enrichGO(gene = DE.set,
      universe = universe,
      OrgDb = HS.annotation,
      ont = "BP",
      keyType = "SYMBOL",
      minGSSize = 1,
      maxGSSize = 100000,
      pAdjustMethod = "BH")
```

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

ORA: A practical example (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

clusterProfiler

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

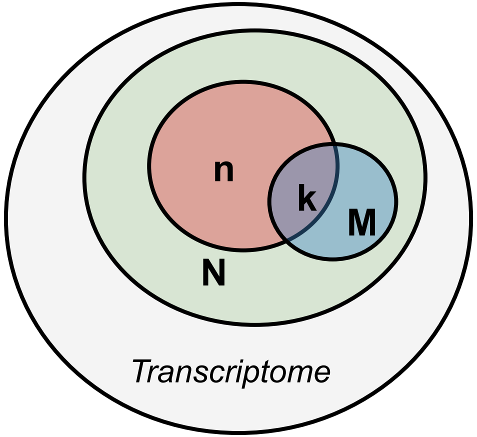
Wieder, C. et al. Pathway analysis in metabolomics: Recommendations for the use of over-representation analysis. PLoS Comput Biol 17

ORA: Back to the fundamentals

clusterProfiler

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

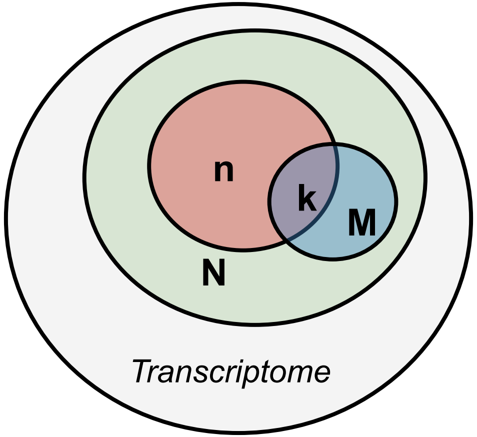
$$P\left(X \geq k\right) = 1 - \sum_{i=0}^{k-1} \frac{\binom{M}{i}\binom{N-M}{n-i}}{\binom{N}{n}}$$

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

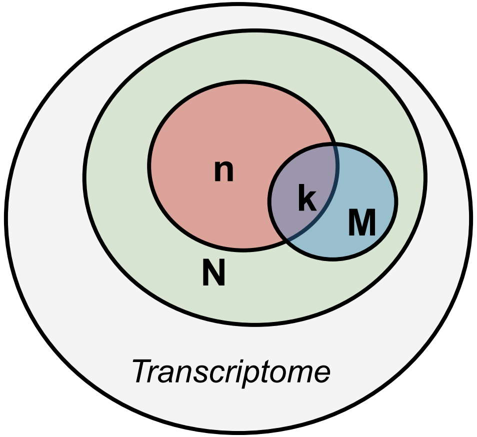
	in BP set	not in BP set
in Gene set	k = 22	923
not in Gene set	169	14454

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



Contingency table

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

N

Hypergeometric distribution

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

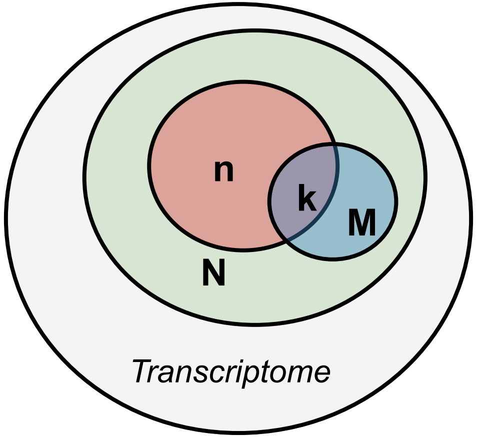
Wieder, C. et al. Pathway analysis in metabolomics: Recommendations for the use of over-representation analysis. PLoS Comput Biol 17

ORA: Back to the fundamentals

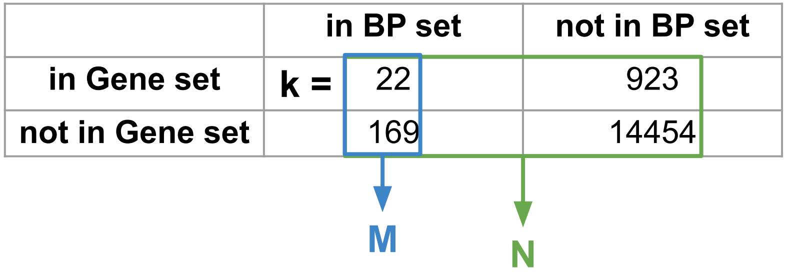
clusterProfiler

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

Hypergeometric distribution

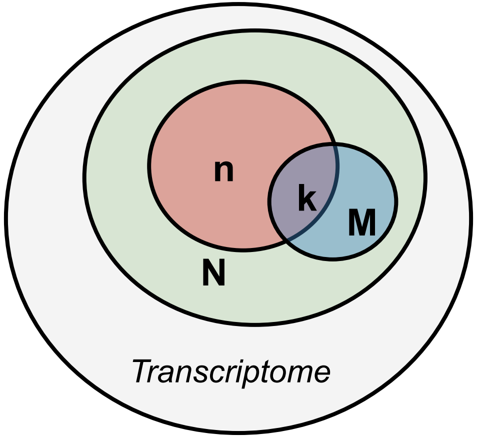
$$P\left(X \geq k\right) = 1 - \sum_{i=0}^{k-1} \frac{\binom{M}{i}\binom{N-M}{n-i}}{\binom{N}{n}}$$

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

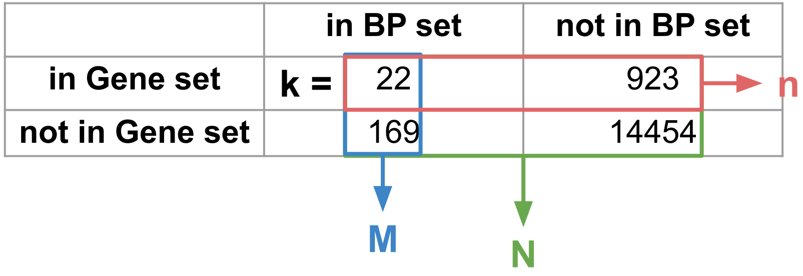
2 equivalent ways of representing and computing



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



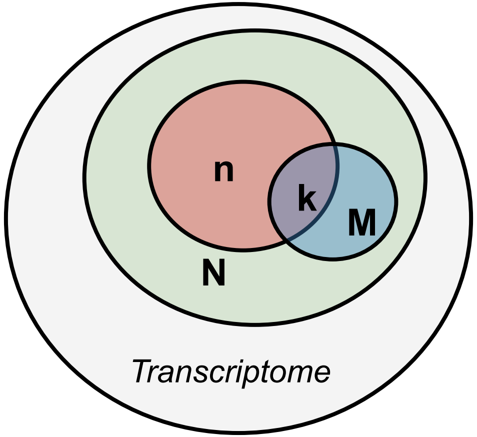
		in BP set	not in BP set
in Gene set	**k =**	22	923
not in Gene set		169	14454

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

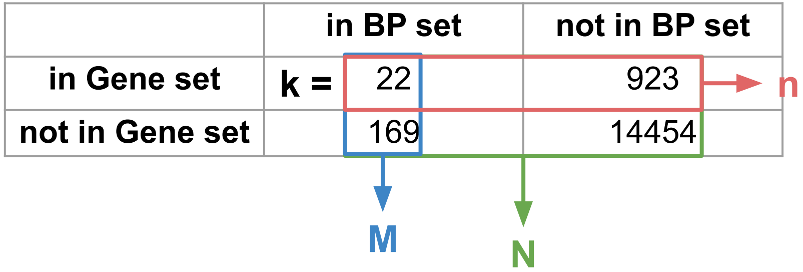
2 equivalent ways of representing and computing



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

Right-tailed fisher exact test

= *Testing Independence*

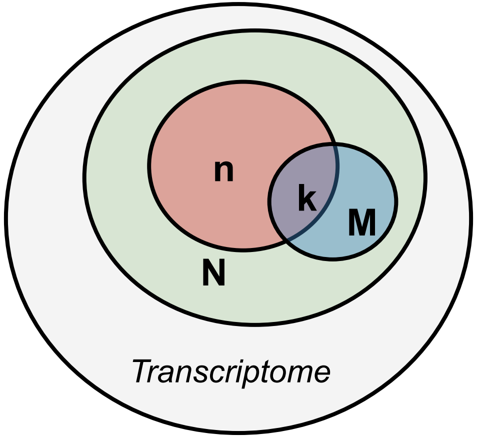
Wieder, C. et al. Pathway analysis in metabolomics: Recommendations for the use of over-representation analysis. PLoS Comput Biol 17

ORA: Back to the fundamentals

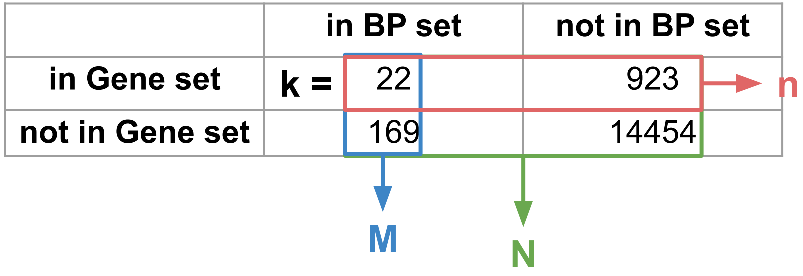
clusterProfiler

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

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

= *Testing Independence*

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

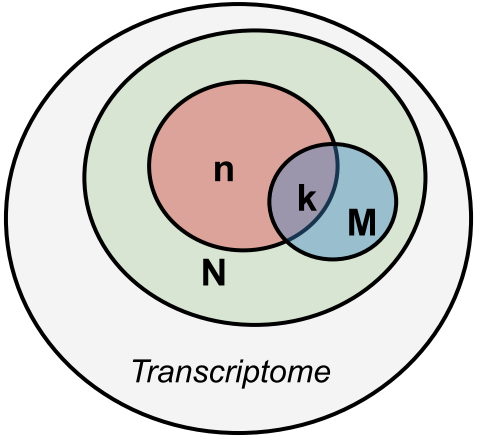
Wieder, C. et al. Pathway analysis in metabolomics: Recommendations for the use of over-representation analysis. PLoS Comput Biol 17

ORA: Back to the fundamentals

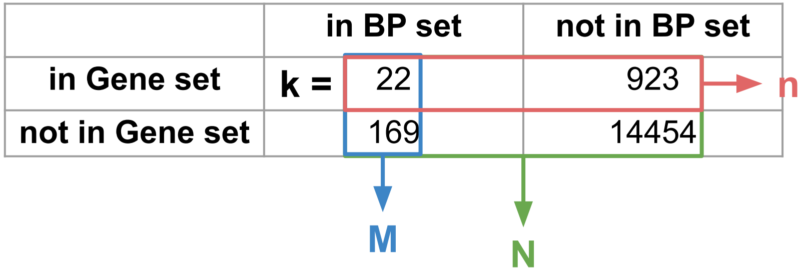
clusterProfiler

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

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

= *Testing Independence*

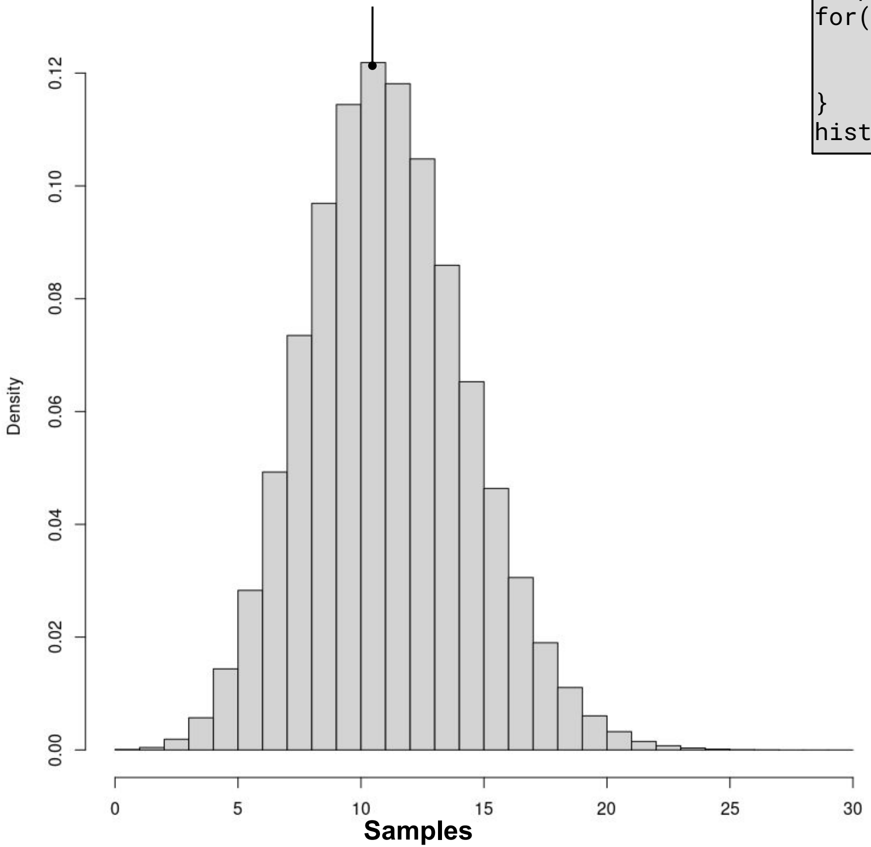
+ correction for multiple-tests

q.value

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

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

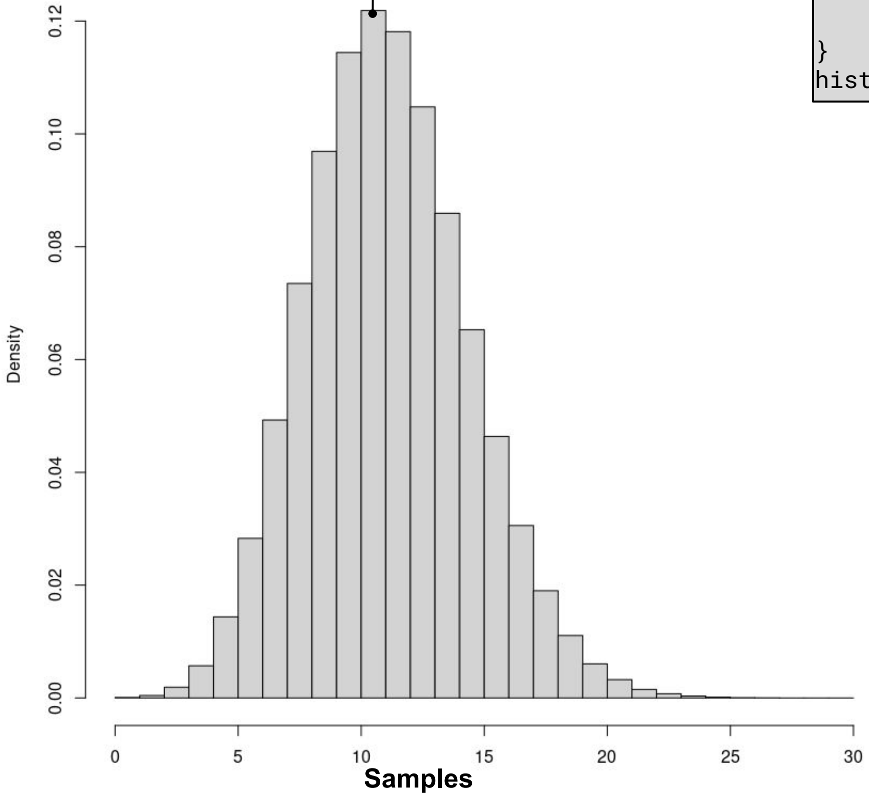
ORA: Intuition via random sampling



```
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))
```

ORA: Intuition via random sampling

$$\underbrace{\frac{191}{15568}}_{P(\text{in BP set})} \times \underbrace{\frac{945}{15568}}_{P(\text{in Gene set})} \times 15568 \approx 11.6$$

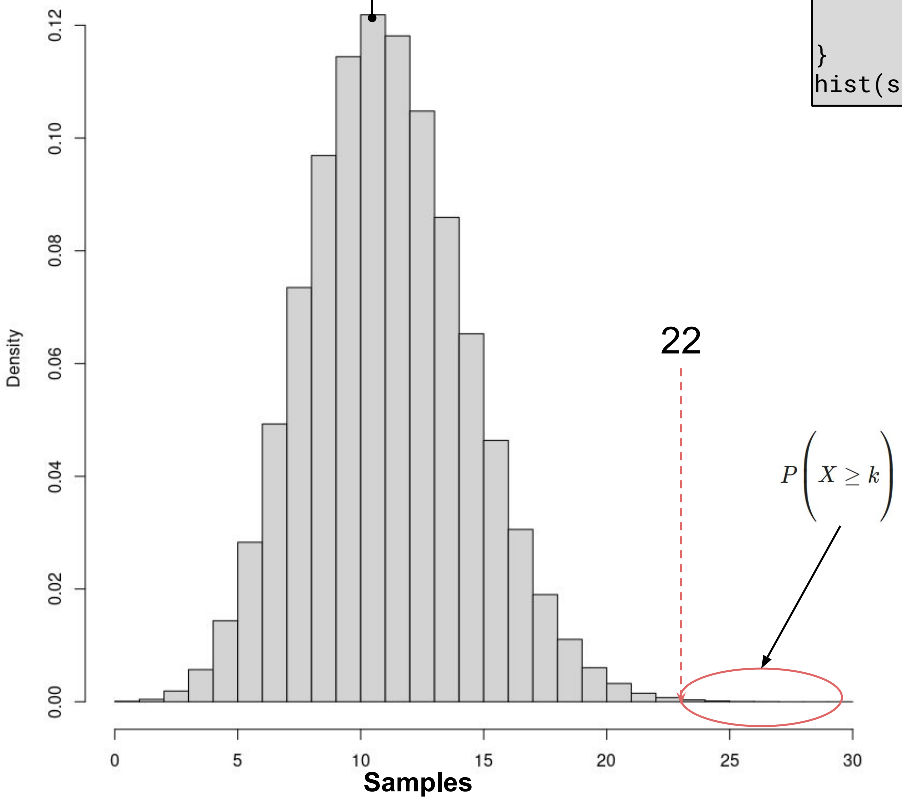


```
N_SAMPLES <- 1000000
samples <- vector(mode = "numeric", length = N_SAMPLES)
for(i in 1:N_SAMPLES){
      s <- sample(size = 945, x = universe, replace = F)
      samples[i] <- sum(s %in% GO.0006836)
}
hist(samples, freq = F, breaks = seq(0,30,1))
```

ORA: Intuition via random sampling

$$\underbrace{\frac{191}{15568}}_{P(\text{in BP set})} \times \underbrace{\frac{945}{15568}}_{P(\text{in Gene set})} \times 15568 \approx 11.6$$

```
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 ?*

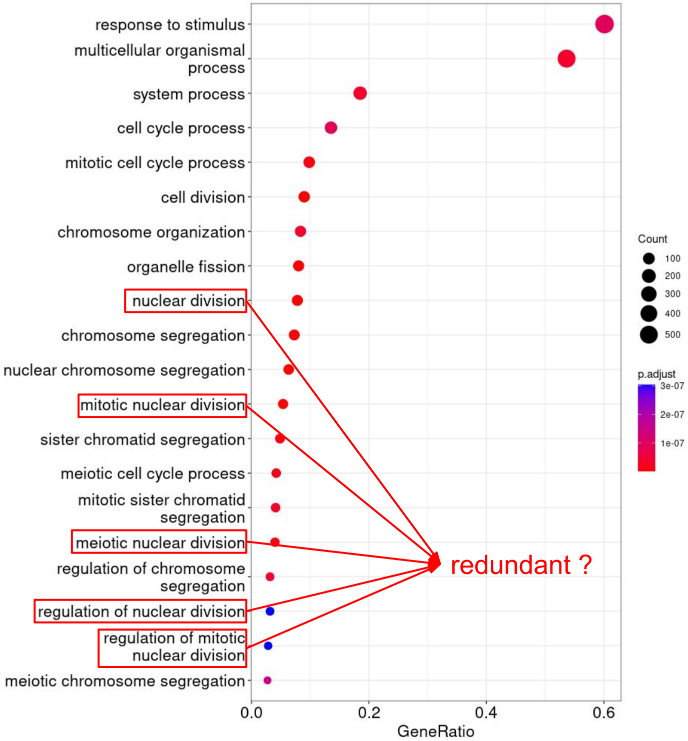
```
estimate = 0.002816
```

ORA: A practical example (1)

TCGA

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

R packages for ORA: clusterProfiler Enrichr



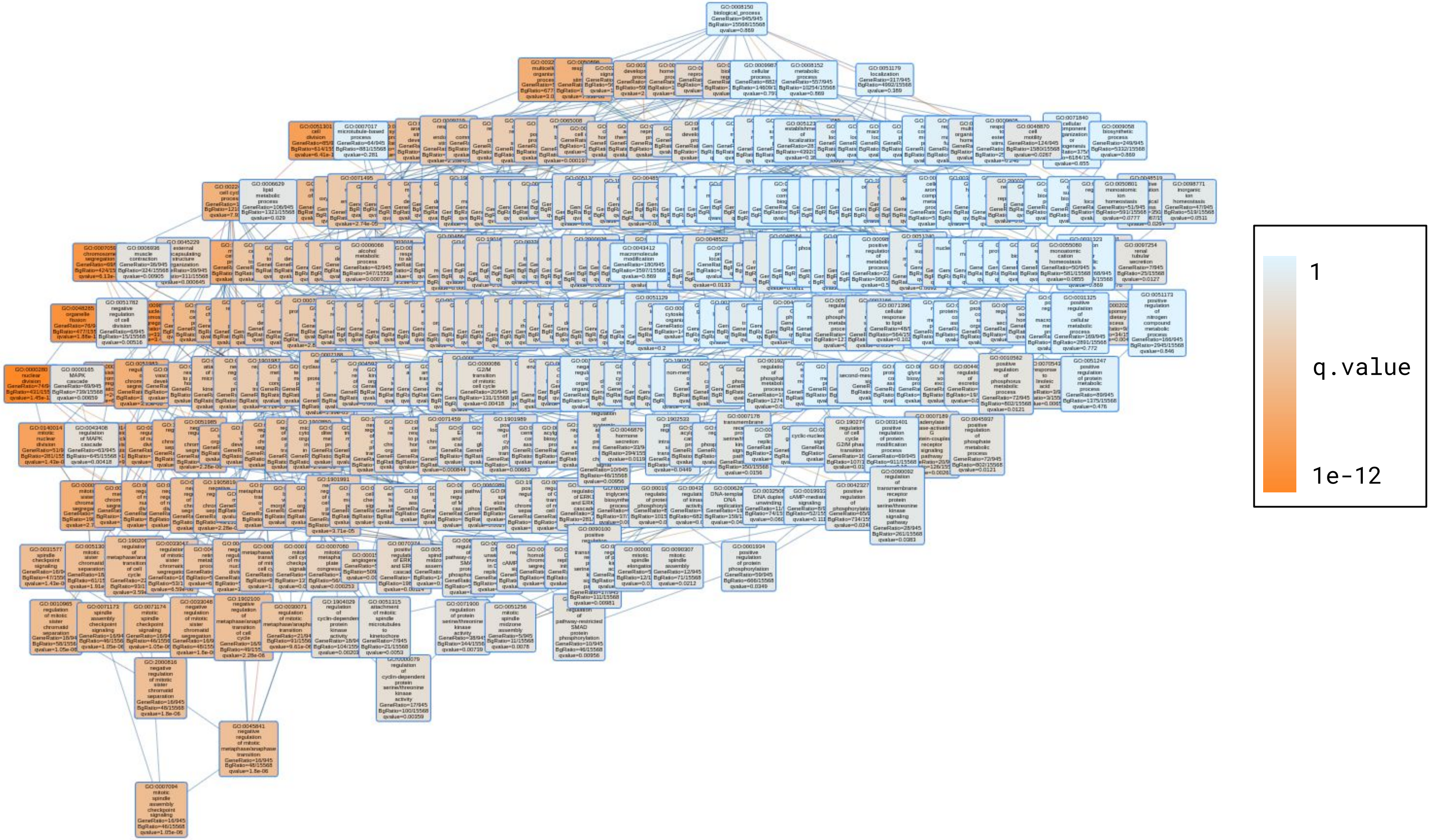
GENEONTOLOGY
Unifying Biology

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")
```

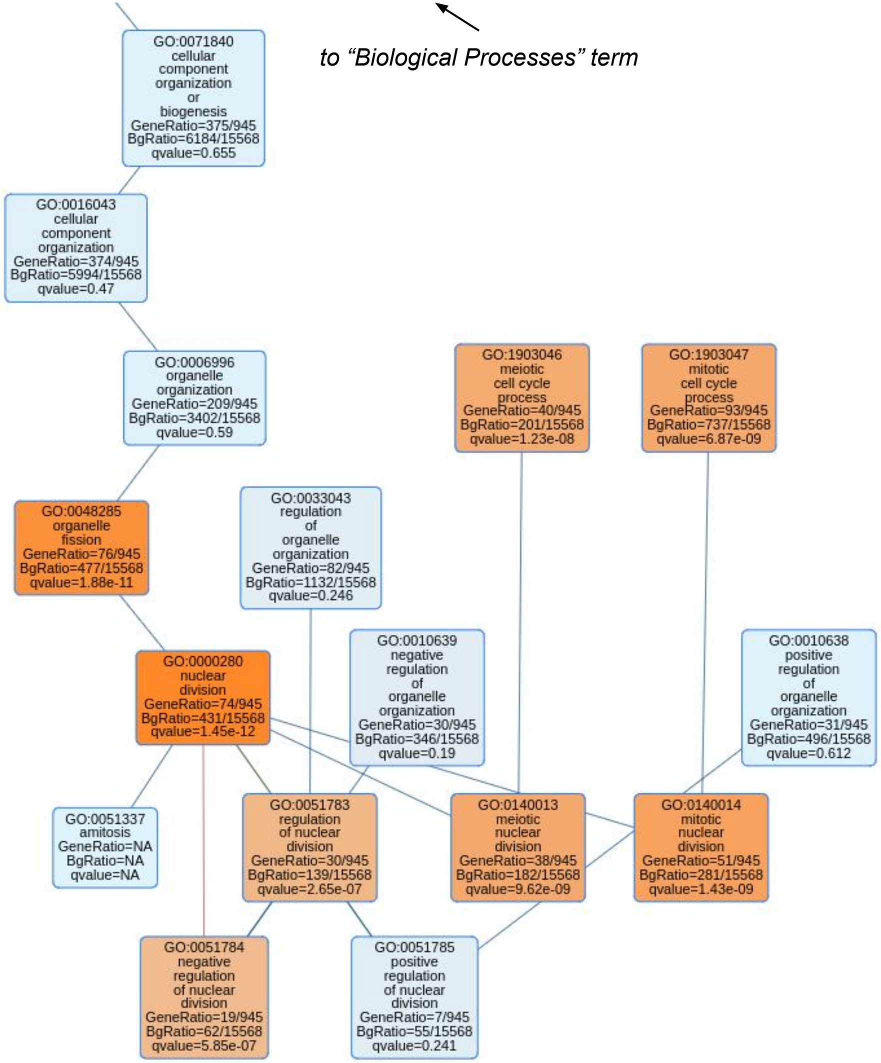
https://master.bioconductor.org/packages/release/bioc/vignettes/TCGAbiolinks/inst/doc/analysis.html

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

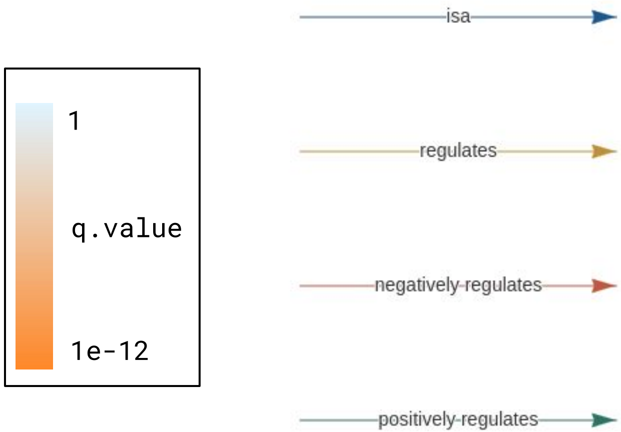


ORA: A practical example (1) - DAG view

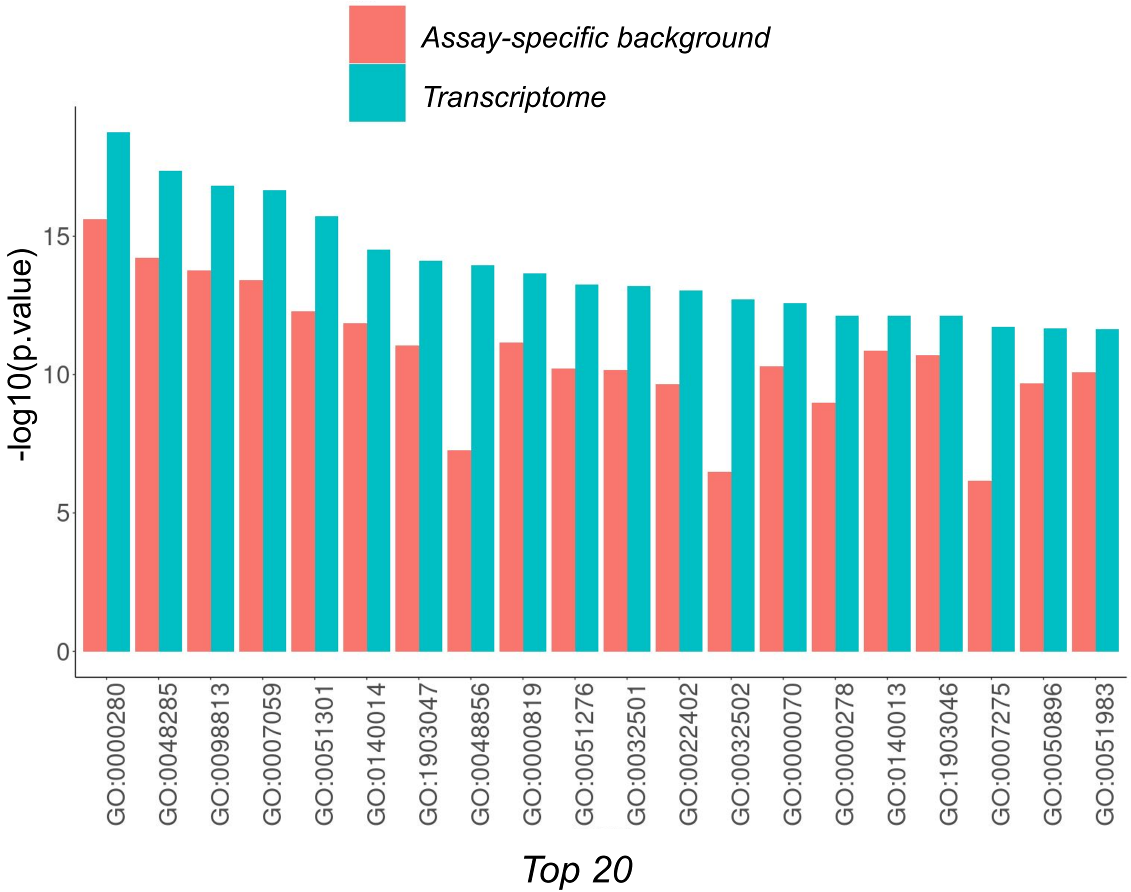
The Gene Ontology in a DAG (**Directed Acyclic** Graph)


GO:0071840 cellular component organization or biogenesis GeneRatio=375/945 BgRatio=6184/15568 qvalue=0.655
GO:0016043 cellular component organization GeneRatio=374/945 BgRatio=5994/15568 qvalue=0.47
GO:0006996 organelle organization GeneRatio=209/945 BgRatio=3402/15568 qvalue=0.59
GO:1903046 meiotic cell cycle process GeneRatio=40/945 BgRatio=201/15568 qvalue=1.23e-08
GO:1903047 mitotic cell cycle process GeneRatio=93/945 BgRatio=737/15568 qvalue=6.87e-09
GO:0048285 organelle fission GeneRatio=76/945 BgRatio=477/15568 qvalue=1.88e-11
GO:0033043 regulation of organelle organization GeneRatio=82/945 BgRatio=1132/15568 qvalue=0.246
GO:0000280 nuclear division GeneRatio=74/945 BgRatio=431/15568 qvalue=1.45e-12
GO:0010639 negative regulation of organelle organization GeneRatio=30/945 BgRatio=346/15568 qvalue=0.19
GO:0010638 positive regulation of organelle organization GeneRatio=31/945 BgRatio=496/15568 qvalue=0.612
GO:0051337 amitosis GeneRatio=NA BgRatio=NA qvalue=NA
GO:0051783 regulation of nuclear division GeneRatio=30/945 BgRatio=139/15568 qvalue=2.65e-07
GO:0140013 meiotic nuclear division GeneRatio=38/945 BgRatio=182/15568 qvalue=9.62e-09
GO:0140014 mitotic nuclear division GeneRatio=51/945 BgRatio=281/15568 qvalue=1.43e-09
GO:0051784 negative regulation of nuclear division GeneRatio=19/945 BgRatio=62/15568 qvalue=5.85e-07
GO:0051785 positive regulation of nuclear division GeneRatio=7/945 BgRatio=55/15568 qvalue=0.241

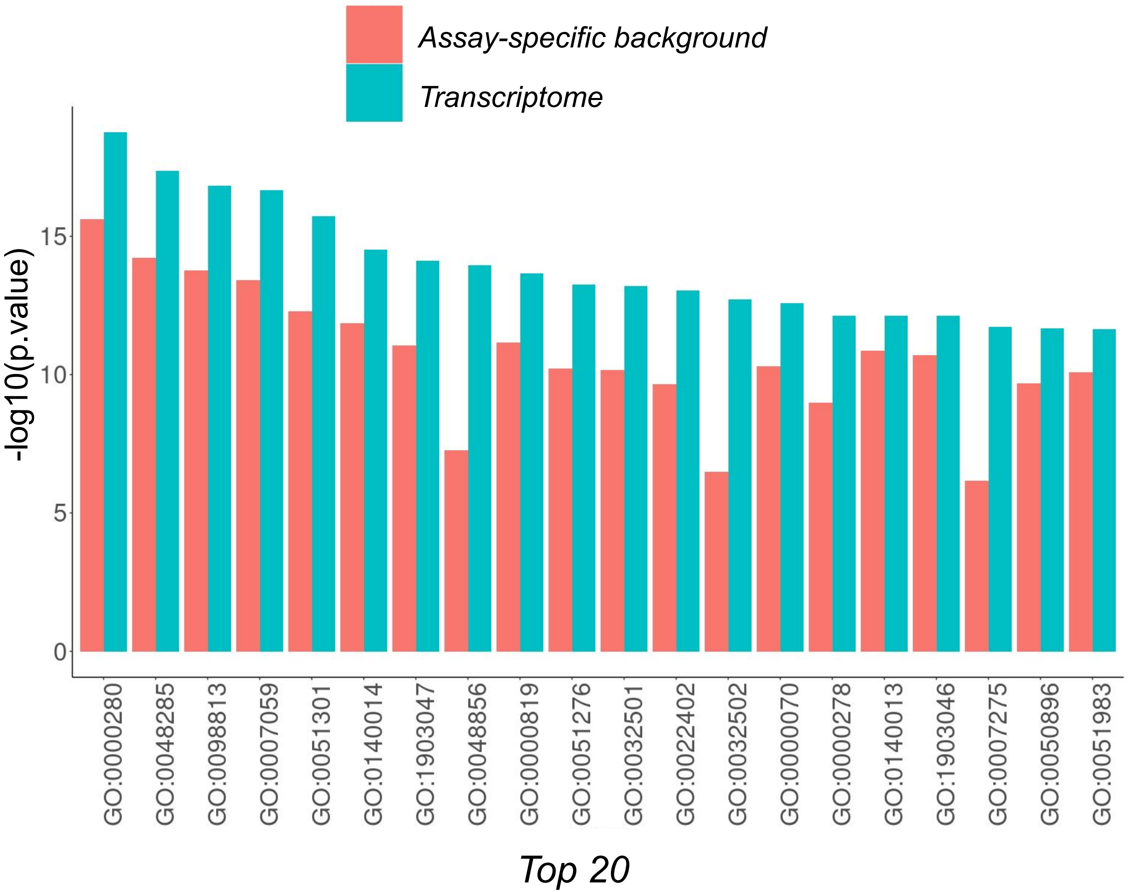
to “Biological Processes” term



ORA: The Impact of the *universe* definition



ORA: The Impact of the *universe* definition



Leads to overestimation of the p-value

+

Order ~ preserved

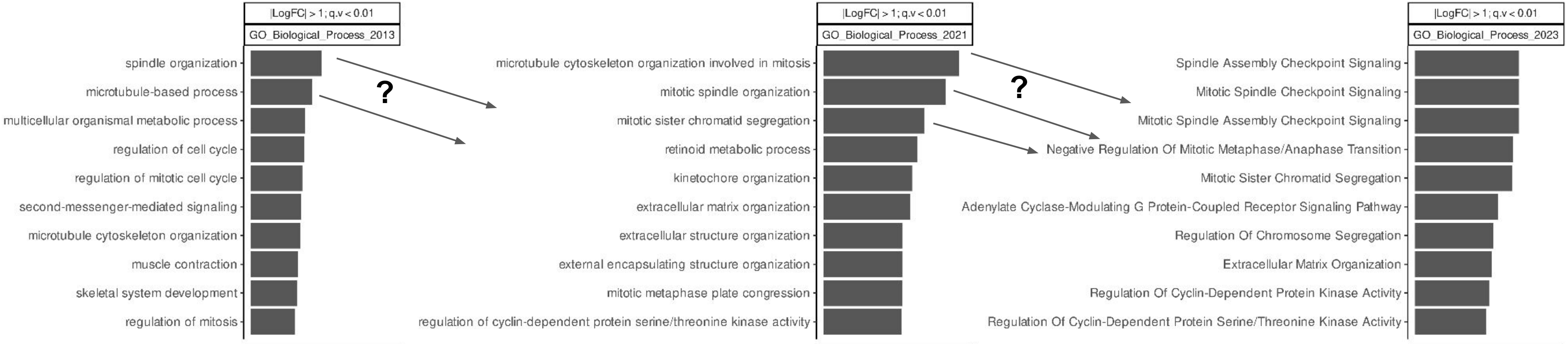
↓

Increase false positive enriched terms

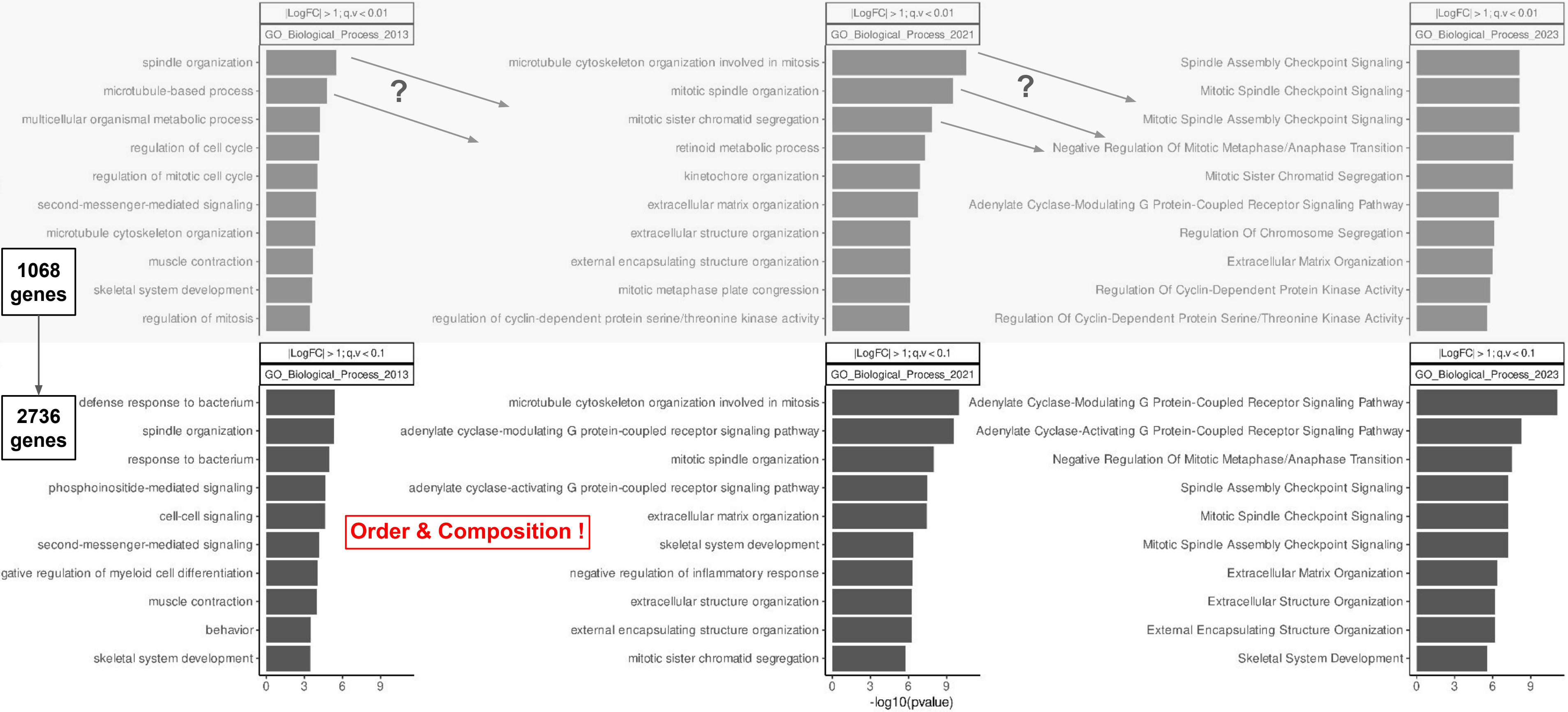
133 vs 218 enriched GO terms
(q.value <= 1e.3)

ORA: Impact of the database and gene set thresholds choices

Enrichr



ORA: Impact of the database and gene set thresholds choices



1068 genes

2736 genes

Order & Composition !

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

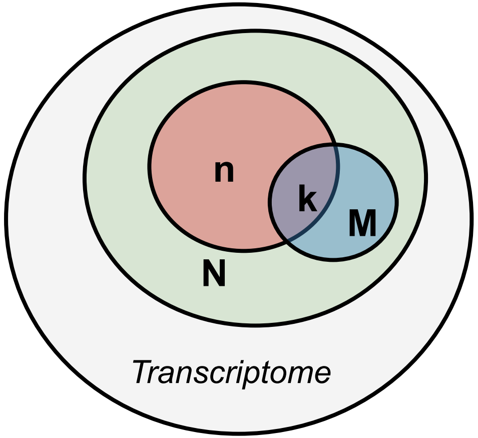
Unspecific universe (*background set*) can create false positives

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

Selection thresholds are important:

- Too large = noisy detection
- Too small = low detection

Reference database and versions can an impact of results



ORA: Several biases

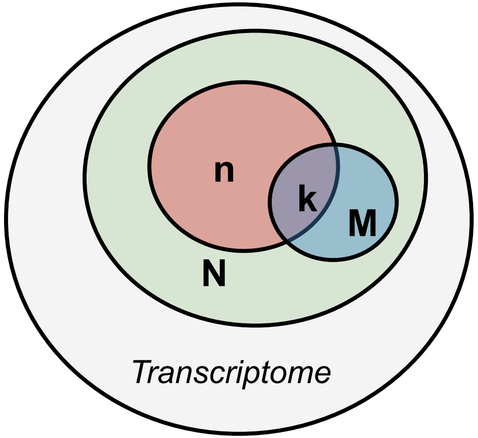
Unspecific universe (*background set*) can create false positives

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

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

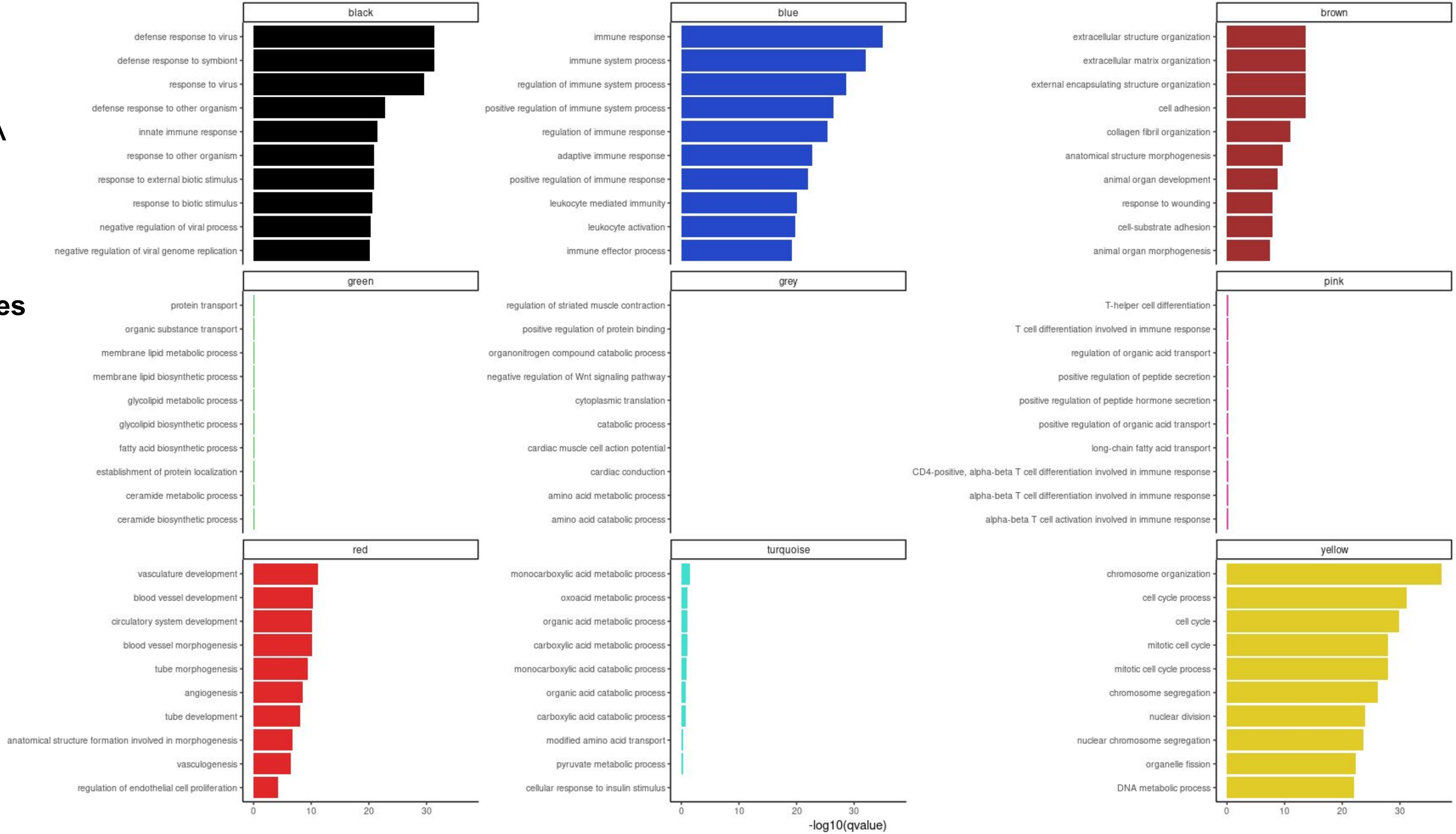
TP WGCNA

↓

9 gene modules

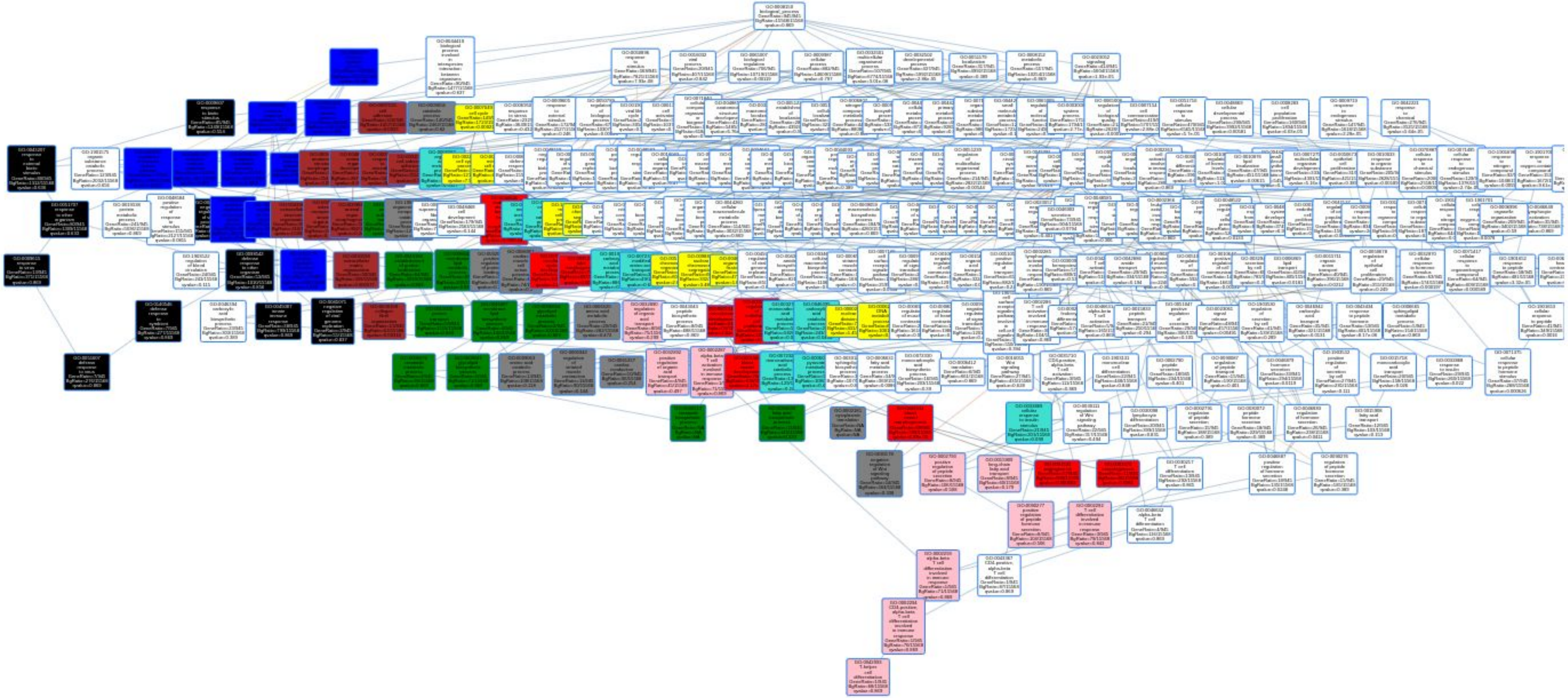
↓

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

→ **GSEA**

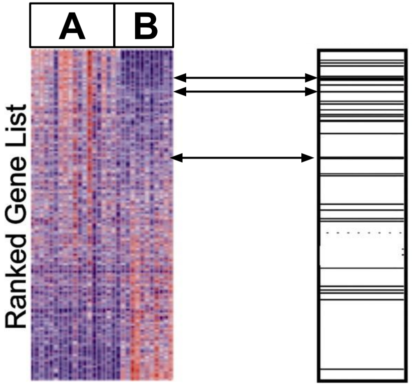
Subramanian, A. et al., 2005. Gene set enrichment analysis: A knowledge-based approach for interpreting genome-wide expression profiles.
James H Joly et al., 2019, Differential Gene Set Enrichment Analysis: a statistical approach to quantify the relative enrichment of two gene sets, *Bioinformatics*.

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.

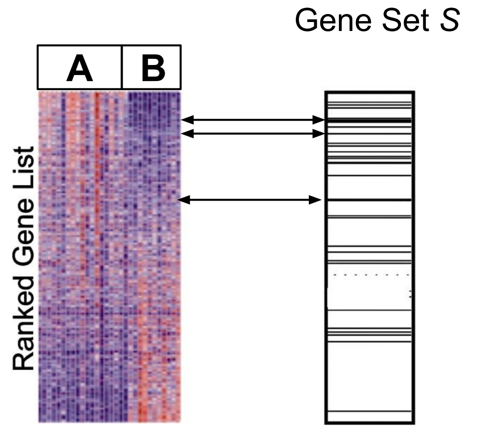
Genes associated
with a GO term

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

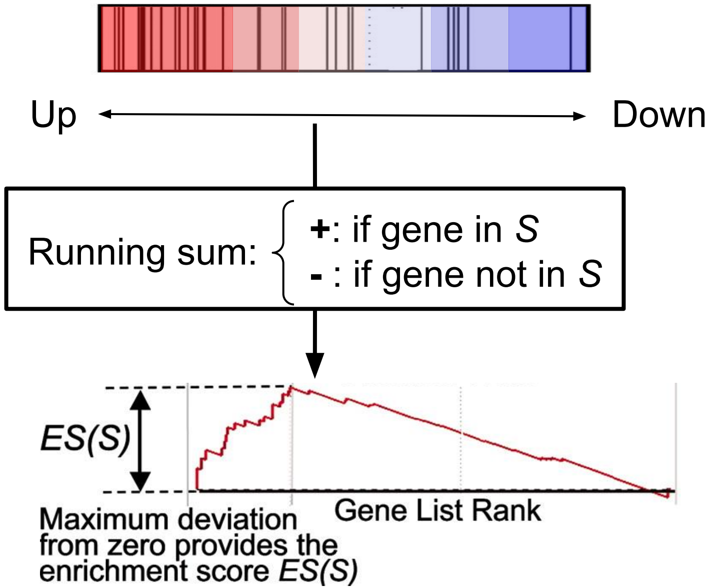
→ **GSEA**



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

Genes associated with a GO term

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


Up
Down
Running sum:
+: if gene in *S*
- : if gene not in *S*
ES(S)
Gene List Rank
Maximum deviation from zero provides the enrichment score *ES(S)*

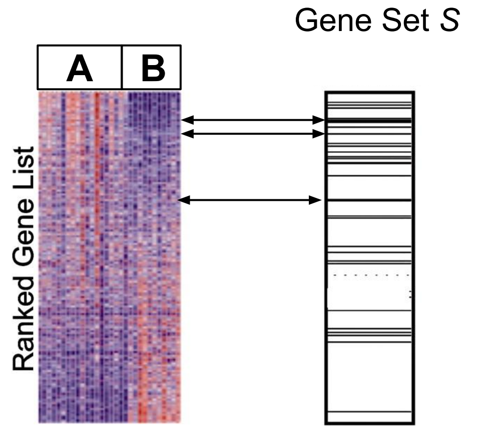
Subramanian, A. et al., 2005. Gene set enrichment analysis: A knowledge-based approach for interpreting genome-wide expression profiles.
James H Joly et al., 2019, Differential Gene Set Enrichment Analysis: a statistical approach to quantify the relative enrichment of two gene sets, *Bioinformatics*.

GSEA: A Function scoring method (1)

- Recall of the impact of the threshold (q.value & logFC) on the ORA results
- All genes are not equivalent: sign and intensity of variation

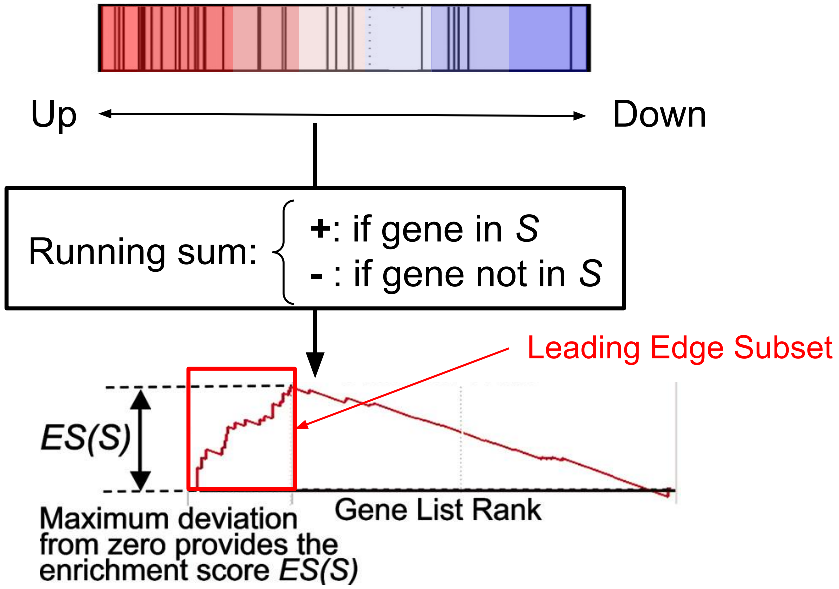
→ **GSEA**

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



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

Genes associated with a GO term

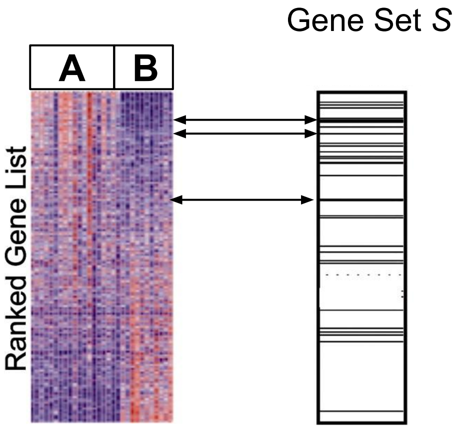

Up
Down
Running sum:
+: if gene in *S*
- : if gene not in *S*
Leading Edge Subset
ES(S)
Gene List Rank
Maximum deviation from zero provides the enrichment score *ES(S)*

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

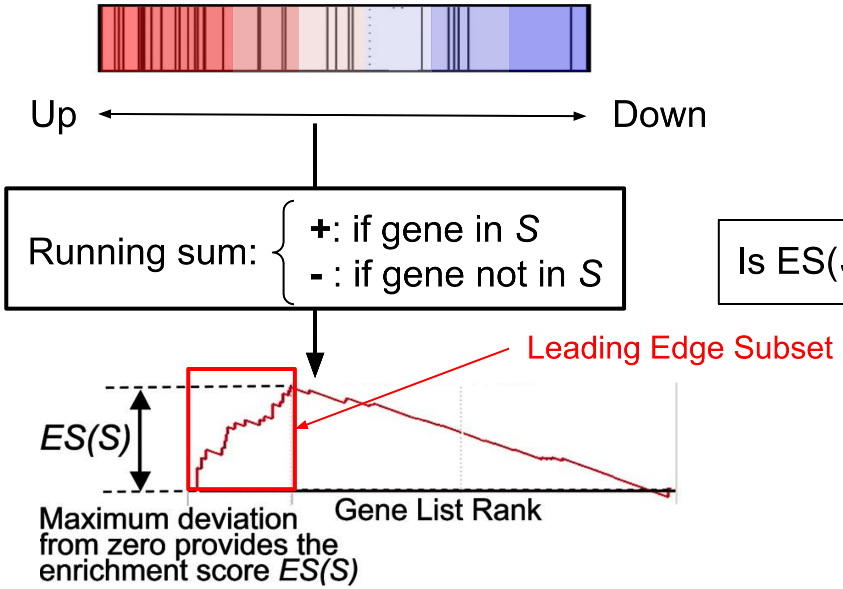
→ **GSEA**



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

Genes associated
with a GO term

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



Running sum: **+**: if gene in *S*
- : if gene not in *S*

Maximum deviation from zero provides the enrichment score *ES(S)*

Is ES(*S*) significant ?

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

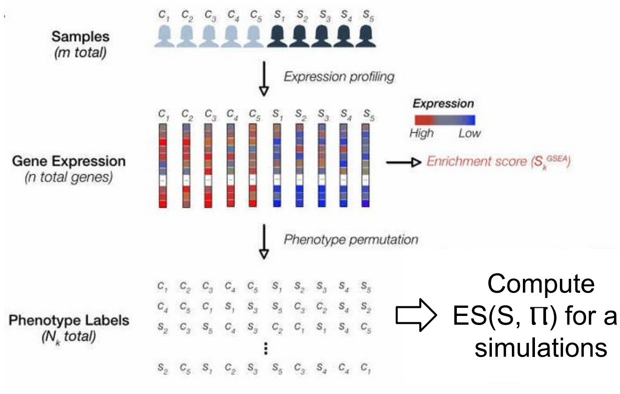
https://www.pathwaycommons.org/guide/primers/data_analysis/gsea/

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)


Samples
(m total)
C1 C2 C3 C4 C5 S1 S2 S3 S4 S5
Expression profiling
Expression
High Low
Gene Expression
(n total genes)
C1 C2 C3 C4 C5 S1 S2 S3 S4 S5
Enrichment score (S_k^GSEA)
Phenotype permutation
Phenotype Labels
(N_k total)
C1 C2 C3 C4 C5 S1 S2 S3 S4 S5
C4 C5 C1 S1 S3 S5 C3 C2 S4 S2
S2 C3 S5 C4 S3 C2 C1 S1 S4 C5
⋮
S2 C5 S1 C2 S3 S5 C3 S4 C4 C1

Compute
ES(S, Π) for all
simulations

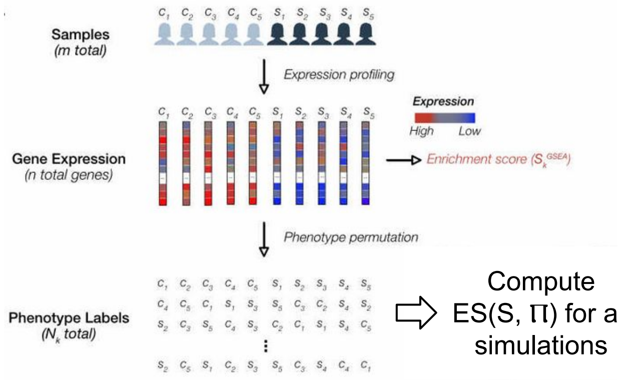
https://www.pathwaycommons.org/guide/primers/data_analysis/gsea/

GSEA: A Function scoring method (2)

Is ES(*S)* significant ?

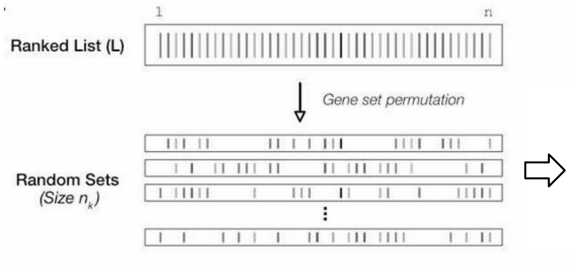
- in unweighted settings (first version of GSEA): exact p-value estimation with KS-test
- in weighted settings (common): empirical estimation via permutation test (simulations Π)

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/

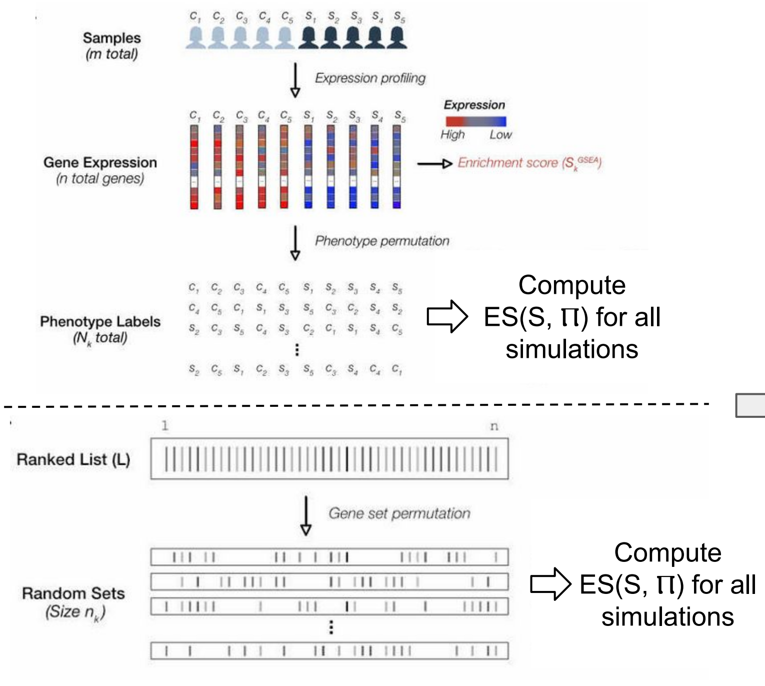
GSEA: A Function scoring method (2)

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)

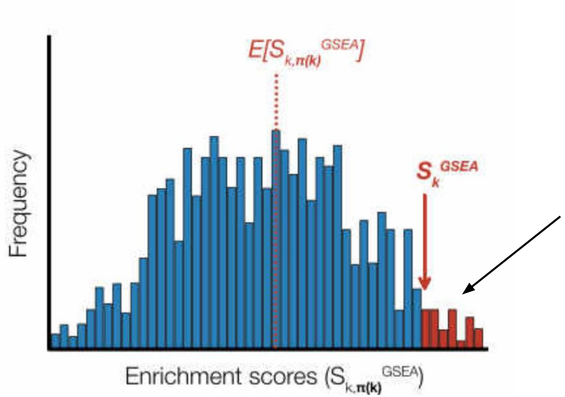
Gene set Permutation (most common but less efficient)



Samples
(m total)
c_1 c_2 c_3 c_4 c_5 s_1 s_2 s_3 s_4 s_5
Expression profiling
Expression
High
Low
Gene Expression
(n total genes)
Enrichment score (S_k^{GSEA})
Phenotype permutation
Phenotype Labels
(N_k total)
c_1 c_2 c_3 c_4 c_5 s_1 s_2 s_3 s_4 s_5
c_4 c_5 c_1 s_1 s_3 s_5 c_3 c_2 s_4 s_2
s_2 c_3 s_5 c_4 s_3 c_2 c_1 s_1 s_4 c_5
⋮
s_2 c_5 s_1 c_2 s_3 s_5 c_3 s_4 c_4 c_1
1
n
Ranked List (L)
Gene set permutation
Random Sets
(Size n_k)
⋮

Compute ES(S, Π) for all simulations

Compute ES(S, Π) for all simulations



Nominal *p-value*

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

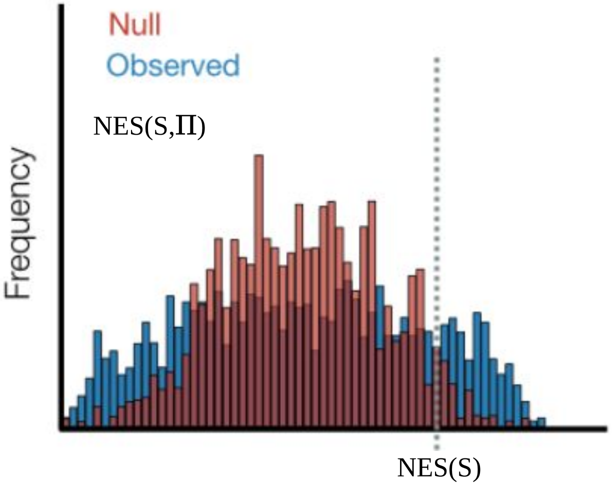
GSEA: A Function scoring method (3)

- How to account for Gene set size differences ?

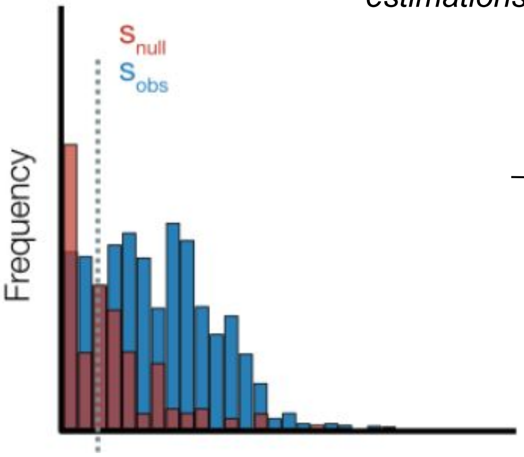
NES (**N**ormalised **E**nrichment **S**core)

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

- Multiple-test correction: FDR estimation



estimations separated by NES signs

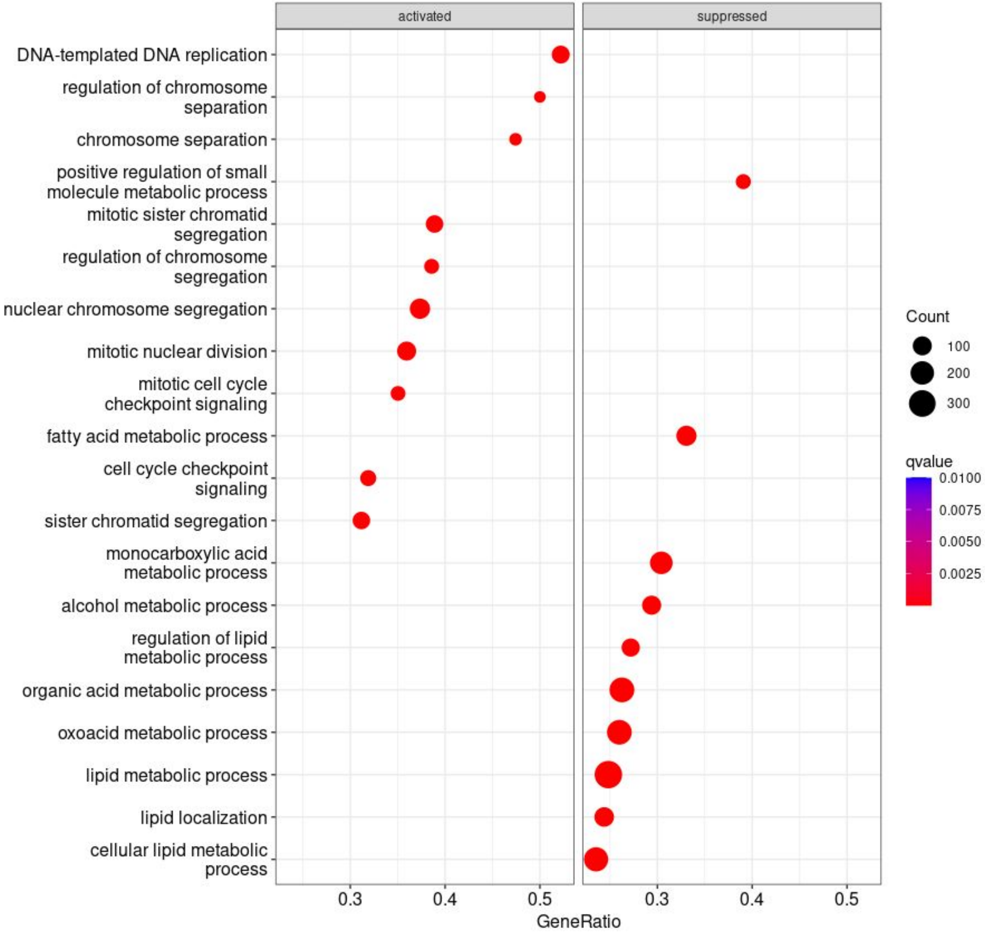


$\longrightarrow \hat{FDR}$

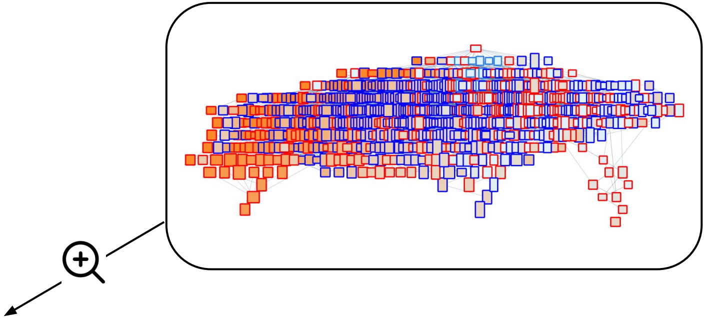
https://www.pathwaycommons.org/guide/primers/data_analysis/gsea/

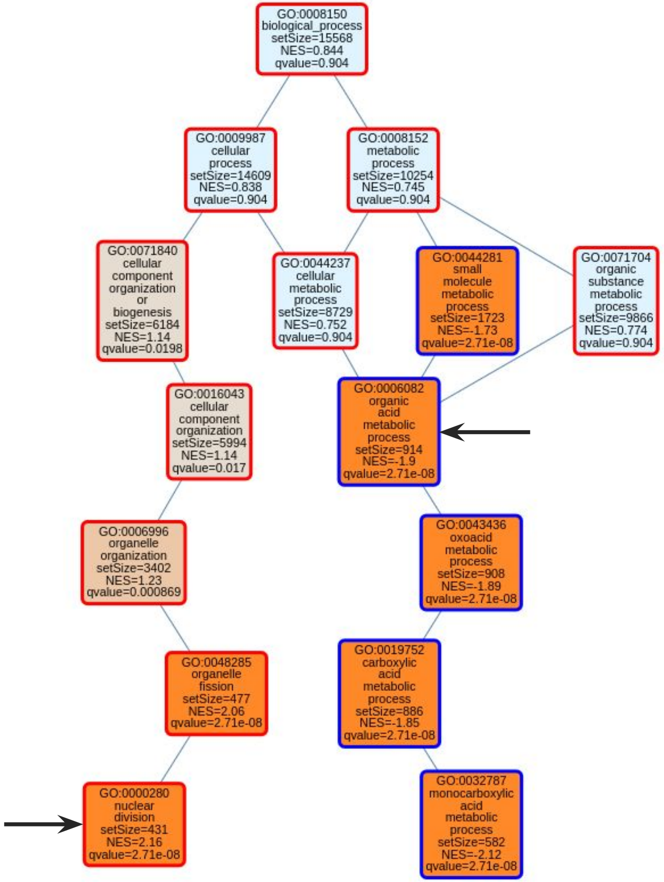
GSEA: A Function scoring method - example

- 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 !

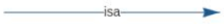


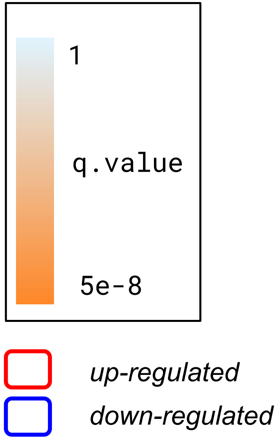
GSEA: Visualisation - example



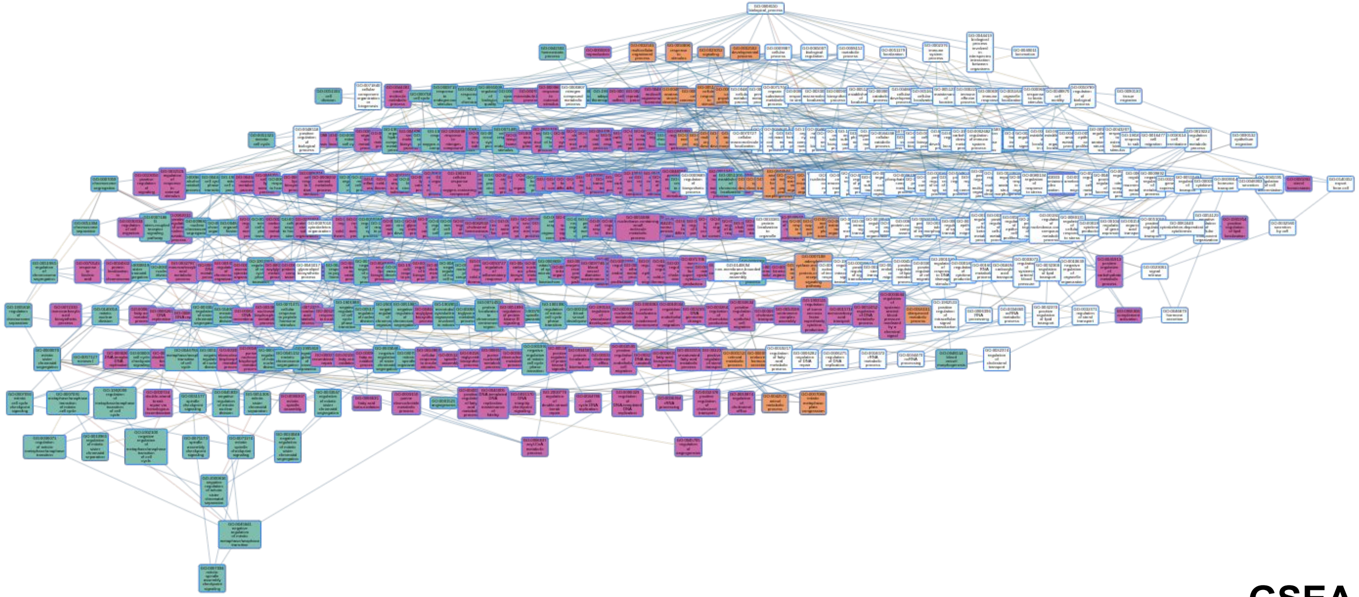


Up and Down regulated branch of the DAG





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 !

Enrichement Analyses - 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.

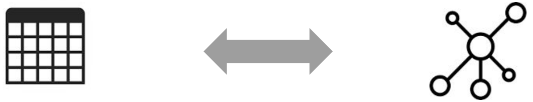
- GSEA is not affected by thresholding are give more weight to the most discriminant genes

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

Topological methods

Extend contextualisation with Biomedical Knowledge Graph

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



Attributes/Properties

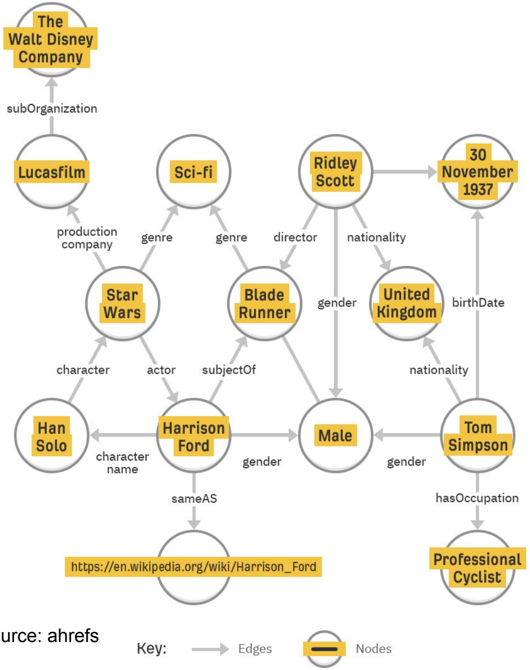
Relations/Paths

Different questions, different visualisation, different methods

What’s a Knowledge Graph ?

Connect the knowledge ⇨ **Knowledge Graphs**

Google knowledge graph

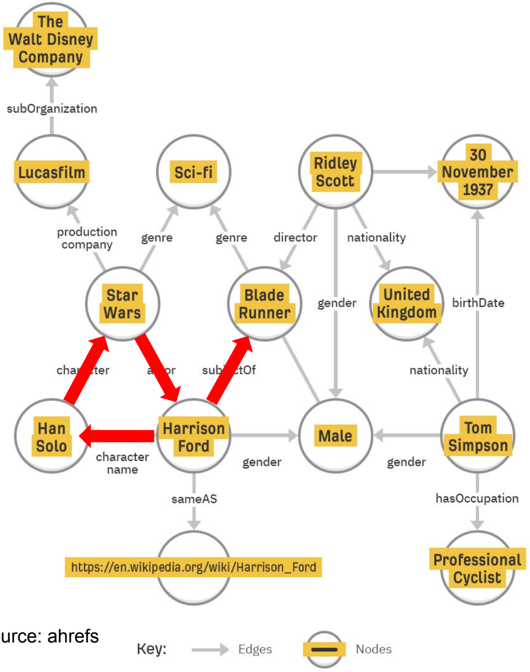


source: ahrefs

What’s a Knowledge Graph ?

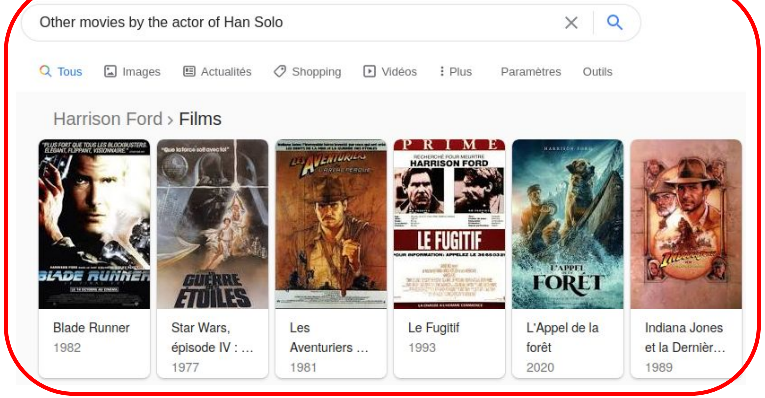
Connect the knowledge → **Knowledge Graphs**

Google knowledge graph



source: ahrefs

Complex Information Retrieval


Other movies by the actor of Han Solo

Harrison Ford › Films

Blade Runner
1982

Star Wars, épisode IV : ...
1977

Les Aventuriers ...
1981

Le Fugitif
1993

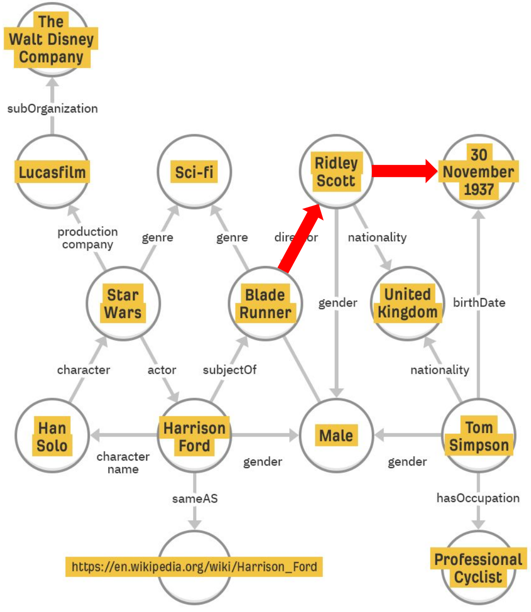
L'Appel de la forêt
2020

Indiana Jones et la Dernièr...
1989

What's a Knowledge Graph ?

Connect the knowledge ⇨ **Knowledge Graphs**

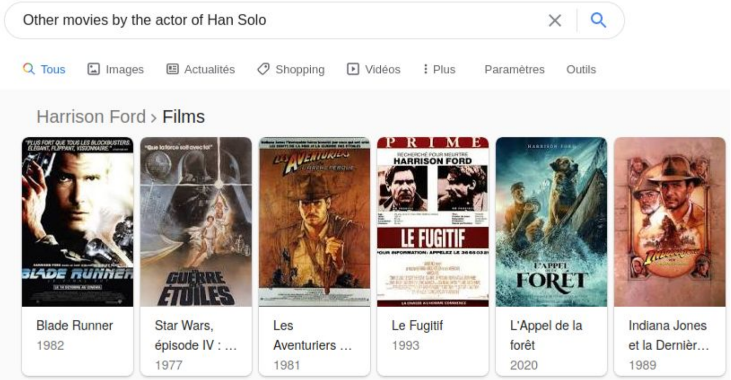
Google knowledge graph



source: ahrefs

Key: Edges Nodes

Complex Information Retrieval


Other movies by the actor of Han Solo

Harrison Ford › Films

Blade Runner
1982

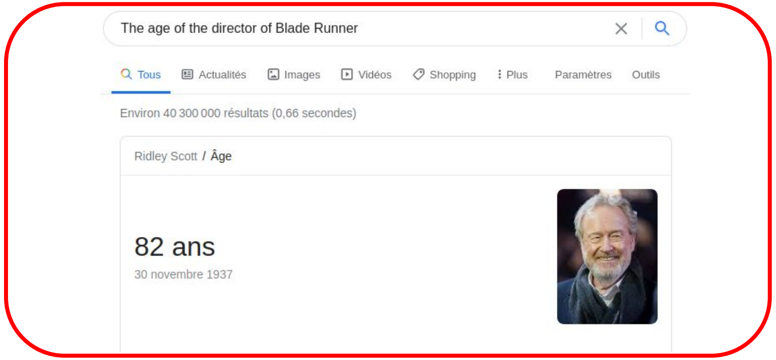
Star Wars, épisode IV : ...
1977

Les Aventuriers ...
1981

Le Fugitif
1993

L'Appel de la forêt
2020

Indiana Jones et la Dernièr...
1989


The age of the director of Blade Runner

Environ 40 300 000 résultats (0,66 secondes)

Ridley Scott / Âge

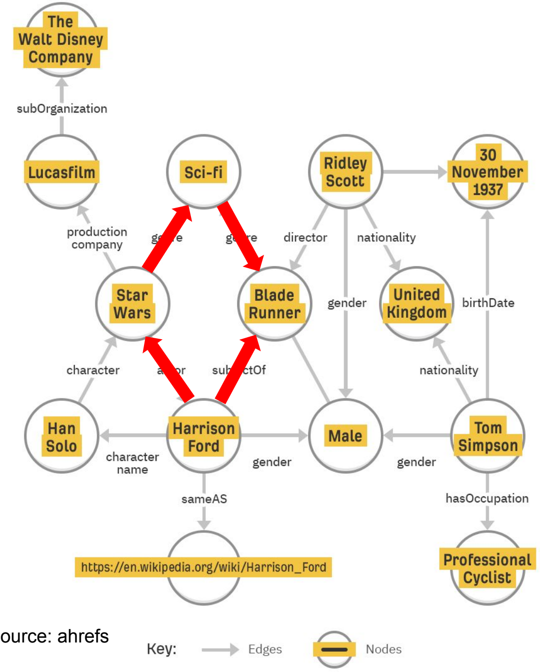
82 ans

30 novembre 1937

What’s a Knowledge Graph ?

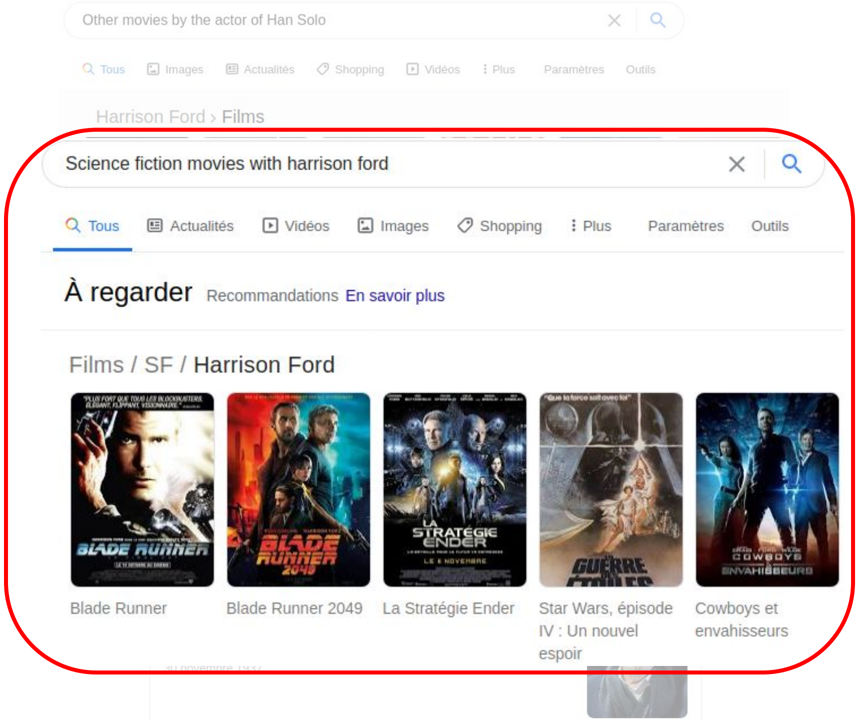
Connect the knowledge ⇨ **Knowledge Graphs**

Google knowledge graph



source: ahrefs

Complex Information Retrieval



Other movies by the actor of Han Solo

Tous Images Actualités Shopping Vidéos Plus Paramètres Outils

Harrison Ford › Films

Science fiction movies with harrison ford

Tous Actualités Vidéos Images Shopping Plus Paramètres Outils

À regarder Recommandations En savoir plus

Films / SF / Harrison Ford

Blade Runner	Blade Runner 2049	La Stratégie Ender	Star Wars, épisode IV : Un nouvel espoir	Cowboys et envahisseurs

Extend contextualisation with Biomedical Knowledge Graph

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

Attributes/Properties

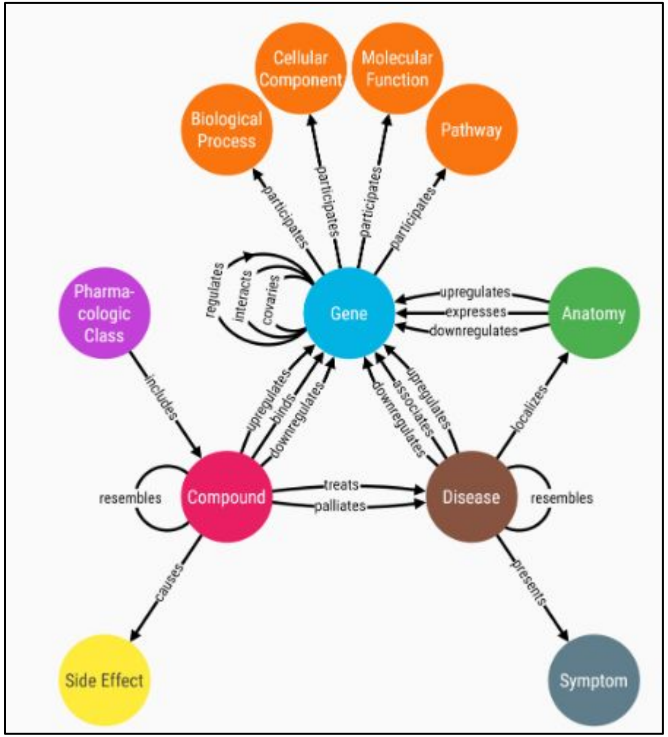
Relations/Paths

Different questions, different visualisation, different methods

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

- Different Model: RDF (in Semantic Web) and **LPG** (Labeled Property Graph)
- Efficient for complex information extraction
- Examples: Hetionet, Wikidata, PharmKG, FORUM, etc.

Example of Hetionet



How to request a Knowledge Graph (in Neo4J) **- LPG**

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) **- LPG**

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) **- LPG**

neo4j Cypher

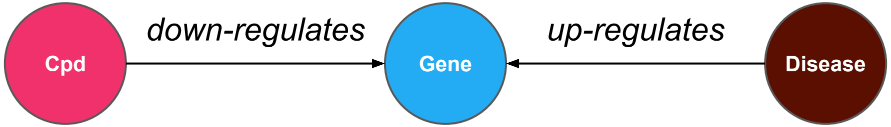
3 main clauses:

- 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:

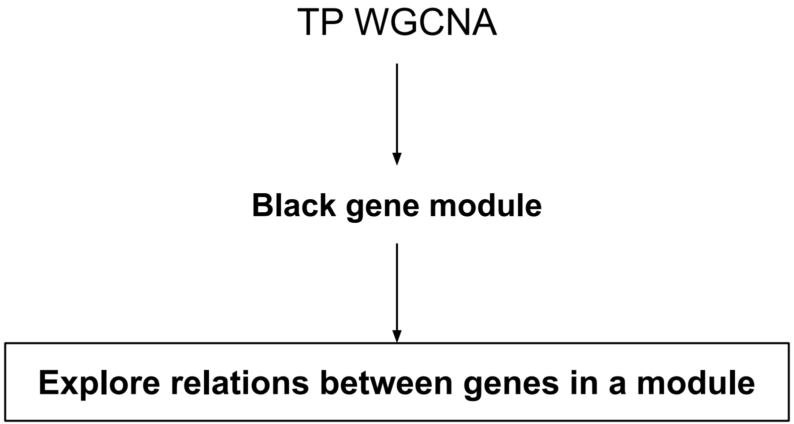
nodes: `(variable:Label)` *(Label is optional)*

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

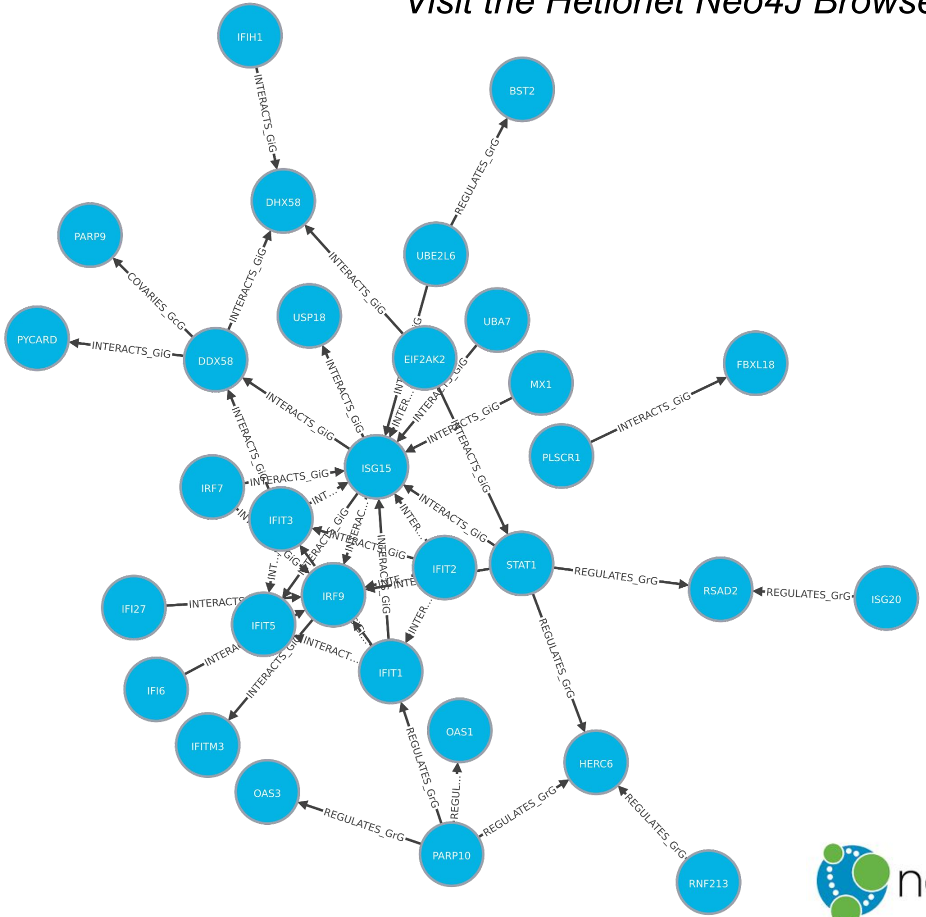


TP WGCNA

Black gene module

Explore relations between genes in a module

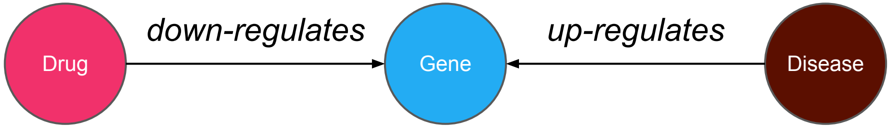
Visit the Hetionet Neo4J Browser



Extend contextualisation with Biomedical Knowledge Graph

A more complex path

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



Drug

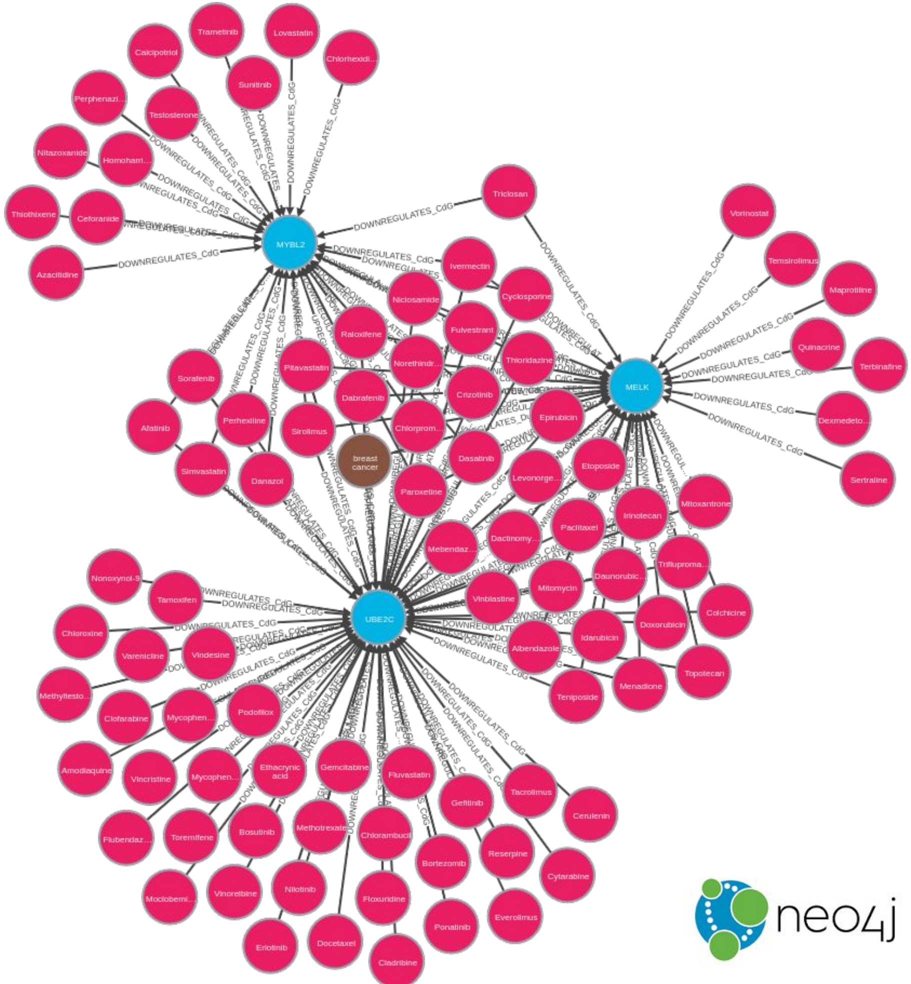
down-regulates

Gene

up-regulates

Disease

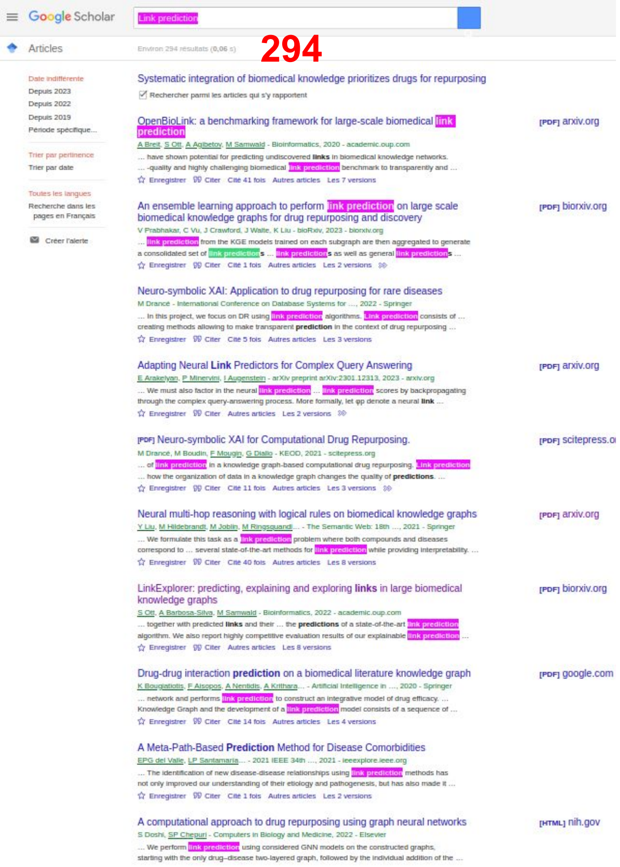
Drugs
Thioridazine
Doxorubicin
Dabrafenib
Teniposide



Extend contextualisation with Biomedical Knowledge Graph

Biomedical Knowledge Graphs as a resource to train link-prediction systems

neo4j



Environ 294 résultats (0,06 s)

294

Systematic integration of biomedical knowledge prioritizes drugs for repurposing

☑ Rechercher parmi les articles qui s'y rapportent

OpenBioLink: a benchmarking framework for large-scale biomedical link prediction — [PDF] arxiv.org
A Breit, S Ott, A Agibetov, M Samwald - Bioinformatics, 2020 - academic.oup.com
... have shown potential for predicting undiscovered **links** in biomedical knowledge networks.
... -quality and highly challenging biomedical link prediction benchmark to transparently and ...
☆ Enregistrer Citer Cité 41 fois Autres articles Les 7 versions

An ensemble learning approach to perform link prediction on large scale biomedical knowledge graphs for drug repurposing and discovery — [PDF] biorxiv.org
V Prabhakar, C Vu, J Crawford, J Waite, K Liu - bioRxiv, 2023 - biorxiv.org
... link prediction from the KGE models trained on each subgraph are then aggregated to generate a consolidated set of link predictions ... link predictions as well as general link predictions ...
☆ Enregistrer Citer Cité 1 fois Autres articles Les 2 versions

Neuro-symbolic XAI: Application to drug repurposing for rare diseases
M Drancé - International Conference on Database Systems for ..., 2022 - Springer
... In this project, we focus on DR using link prediction algorithms. Link prediction consists of ...
creating methods allowing to make transparent **prediction** in the context of drug repurposing ...
☆ Enregistrer Citer Cité 5 fois Autres articles Les 3 versions

Adapting Neural **Link** Predictors for Complex Query Answering — [PDF] arxiv.org
E Arakelyan, P Minervini, I Augenstein - arXiv preprint arXiv:2301.12313, 2023 - arxiv.org
... We must also factor in the neural link prediction ... link prediction scores by backpropagating through the complex query-answering process. More formally, let φp denote a neural **link** ...
☆ Enregistrer Citer Autres articles Les 2 versions

[PDF] Neuro-symbolic XAI for Computational Drug Repurposing. — [PDF] scitepress.o
M Drancé, M Boudin, F Mougin, G Diallo - KEOD, 2021 - scitepress.org
... of link prediction in a knowledge graph-based computational drug repurposing. Link prediction
... how the organization of data in a knowledge graph changes the quality of **predictions**. ...
☆ Enregistrer Citer Cité 11 fois Autres articles Les 3 versions

Neural multi-hop reasoning with logical rules on biomedical knowledge graphs — [PDF] arxiv.org
Y Liu, M Hildebrandt, M Joblin, M Ringsquandl... - The Semantic Web: 18th ..., 2021 - Springer
... We formulate this task as a link prediction problem where both compounds and diseases correspond to ... several state-of-the-art methods for link prediction while providing interpretability. ...
☆ Enregistrer Citer Cité 40 fois Autres articles Les 8 versions

LinkExplorer: predicting, explaining and exploring **links** in large biomedical knowledge graphs — [PDF] biorxiv.org
S Ott, A Barbosa-Silva, M Samwald - Bioinformatics, 2022 - academic.oup.com
... together with predicted **links** and their ... the **predictions** of a state-of-the-art link prediction algorithm. We also report highly competitive evaluation results of our explainable link prediction ...
☆ Enregistrer Citer Autres articles Les 8 versions

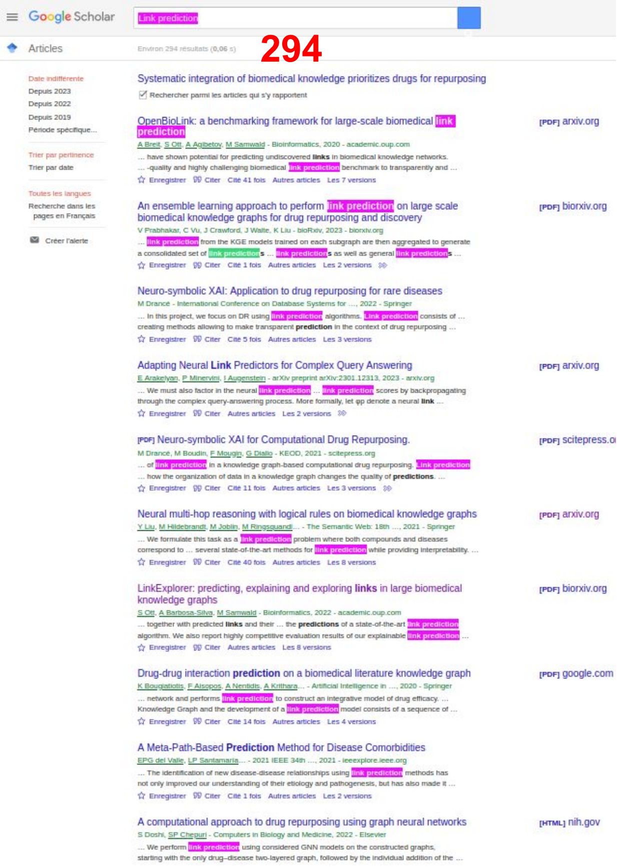
Drug-drug interaction **prediction** on a biomedical literature knowledge graph — [PDF] google.com
K Bougiatiotis, F Aisopos, A Nentidis, A Krithara... - Artificial Intelligence in ..., 2020 - Springer
... network and performs link prediction to construct an integrative model of drug efficacy. ...
Knowledge Graph and the development of a link prediction model consists of a sequence of ...
☆ Enregistrer Citer Cité 14 fois Autres articles Les 4 versions

A Meta-Path-Based **Prediction** Method for Disease Comorbidities
EPG del Valle, LP Santamaria... - 2021 IEEE 34th ..., 2021 - ieeexplore.ieee.org
... The identification of new disease-disease relationships using link prediction methods has not only improved our understanding of their etiology and pathogenesis, but has also made it ...
☆ Enregistrer Citer Cité 1 fois Autres articles Les 2 versions

A computational approach to drug repurposing using graph neural networks — [HTML] nih.gov
S Doshi, SP Chepuri - Computers in Biology and Medicine, 2022 - Elsevier
... We perform link prediction using considered GNN models on the constructed graphs, starting with the only drug–disease two-layered graph, followed by the individual addition of the ...

Extend contextualisation with Biomedical Knowledge Graph

Biomedical Knowledge Graphs as a resource to train link-prediction systems



Google Scholar | Link prediction

Articles | Environ 294 résultats (0,06 s) **294**

Date indifférente
Depuis 2023
Depuis 2022
Depuis 2019
Période spécifique...

Trier par pertinence
Trier par date

Toutes les langues
Recherche dans les pages en Français

Créer l'alerte

Systematic integration of biomedical knowledge prioritizes drugs for repurposing
☑ Rechercher parmi les articles qui s'y rapportent

OpenBioLink: a benchmarking framework for large-scale biomedical link prediction — [PDF] arxiv.org
A Breit, S Ott, A Agibetov, M Samwald - Bioinformatics, 2020 - academic.oup.com
… have shown potential for predicting undiscovered **links** in biomedical knowledge networks. … -quality and highly challenging biomedical link prediction benchmark to transparently and …
☆ Enregistrer Citer Cité 41 fois Autres articles Les 7 versions

An ensemble learning approach to perform link prediction on large scale biomedical knowledge graphs for drug repurposing and discovery — [PDF] biorxiv.org
V Prabhakar, C Vu, J Crawford, J Waite, K Liu - bioRxiv, 2023 - biorxiv.org
… link prediction from the KGE models trained on each subgraph are then aggregated to generate a consolidated set of link predictions … link predictions as well as general link predictions …
☆ Enregistrer Citer Cité 1 fois Autres articles Les 2 versions

Neuro-symbolic XAI: Application to drug repurposing for rare diseases
M Drancé - International Conference on Database Systems for …, 2022 - Springer
… In this project, we focus on DR using link prediction algorithms. Link prediction consists of … creating methods allowing to make transparent **prediction** in the context of drug repurposing …
☆ Enregistrer Citer Cité 5 fois Autres articles Les 3 versions

Adapting Neural **Link** Predictors for Complex Query Answering — [PDF] arxiv.org
E Arakelyan, P Minervini, I Augenstein - arXiv preprint arXiv:2301.12313, 2023 - arxiv.org
… We must also factor in the neural link prediction … link prediction scores by backpropagating through the complex query-answering process. More formally, let φp denote a neural **link** …
☆ Enregistrer Citer Autres articles Les 2 versions

[PDF] Neuro-symbolic XAI for Computational Drug Repurposing. — [PDF] scitepress.o
M Drancé, M Boudin, F Mougin, G Diallo - KEOD, 2021 - scitepress.org
… of link prediction in a knowledge graph-based computational drug repurposing. Link prediction … how the organization of data in a knowledge graph changes the quality of **predictions**. …
☆ Enregistrer Citer Cité 11 fois Autres articles Les 3 versions

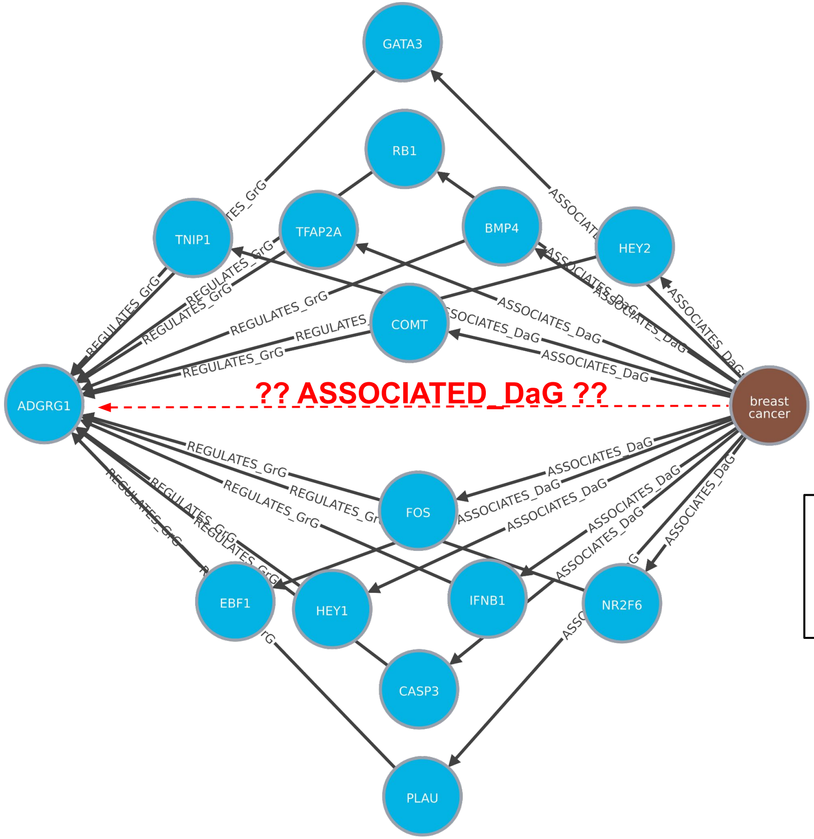
Neural multi-hop reasoning with logical rules on biomedical knowledge graphs — [PDF] arxiv.org
Y Liu, M Hildebrandt, M Joblin, M Ringsquandl… - The Semantic Web: 18th …, 2021 - Springer
… We formulate this task as a link prediction problem where both compounds and diseases correspond to … several state-of-the-art methods for link prediction while providing interpretability. …
☆ Enregistrer Citer Cité 40 fois Autres articles Les 8 versions

LinkExplorer: predicting, explaining and exploring **links** in large biomedical knowledge graphs — [PDF] biorxiv.org
S Ott, A Barbosa-Silva, M Samwald - Bioinformatics, 2022 - academic.oup.com
… together with predicted **links** and their … the **predictions** of a state-of-the-art link prediction algorithm. We also report highly competitive evaluation results of our explainable link prediction …
☆ Enregistrer Citer Autres articles Les 8 versions

Drug-drug interaction **prediction** on a biomedical literature knowledge graph — [PDF] google.com
K Bougiatiotis, F Aisopos, A Nentidis, A Krithara… - Artificial Intelligence in …, 2020 - Springer
… network and performs link prediction to construct an integrative model of drug efficacy. … Knowledge Graph and the development of a link prediction model consists of a sequence of …
☆ Enregistrer Citer Cité 14 fois Autres articles Les 4 versions

A Meta-Path-Based **Prediction** Method for Disease Comorbidities
EPG del Valle, LP Santamaria… - 2021 IEEE 34th …, 2021 - ieeexplore.ieee.org
… The identification of new disease-disease relationships using link prediction methods has not only improved our understanding of their etiology and pathogenesis, but has also made it …
☆ Enregistrer Citer Cité 1 fois Autres articles Les 2 versions

A computational approach to drug repurposing using graph neural networks — [HTML] nih.gov
S Doshi, SP Chepuri - Computers in Biology and Medicine, 2022 - Elsevier
… We perform link prediction using considered GNN models on the constructed graphs, starting with the only drug–disease two-layered graph, followed by the individual addition of the …



E.g Adamic Adar

$$A(x,y) = \sum_{u \in N(x) \cap N(y)} \frac{1}{\log |N(u)|}$$

"Friend of a friend"

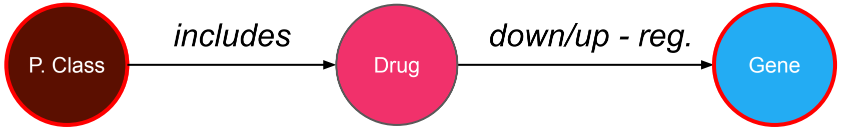
+ Embeddings, rule-based, Supervised, etc.

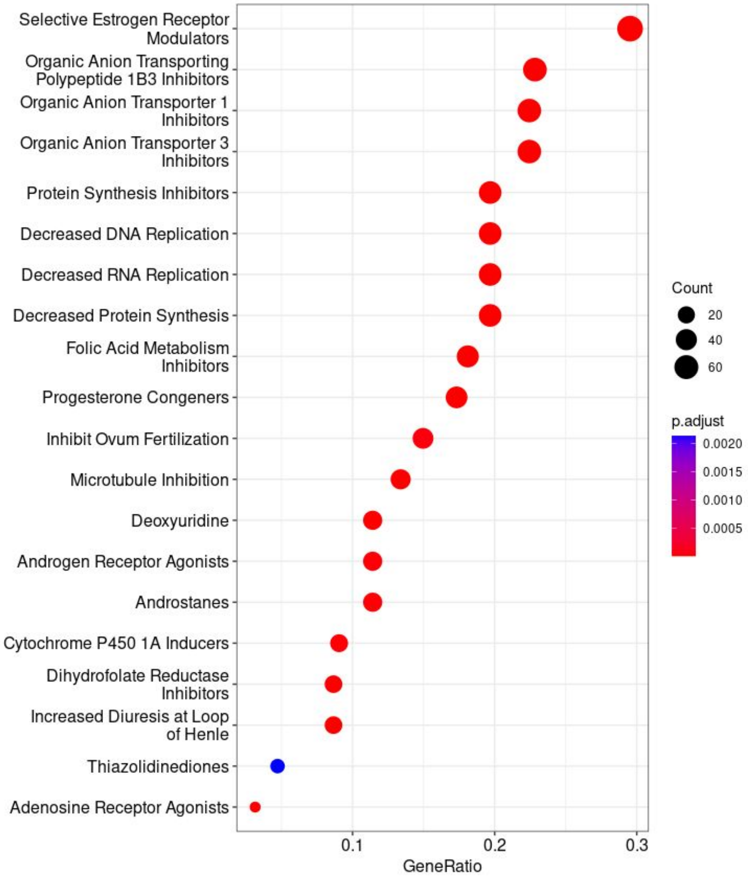
Extend contextualisation with Biomedical Knowledge Graph

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)

R + neo4j

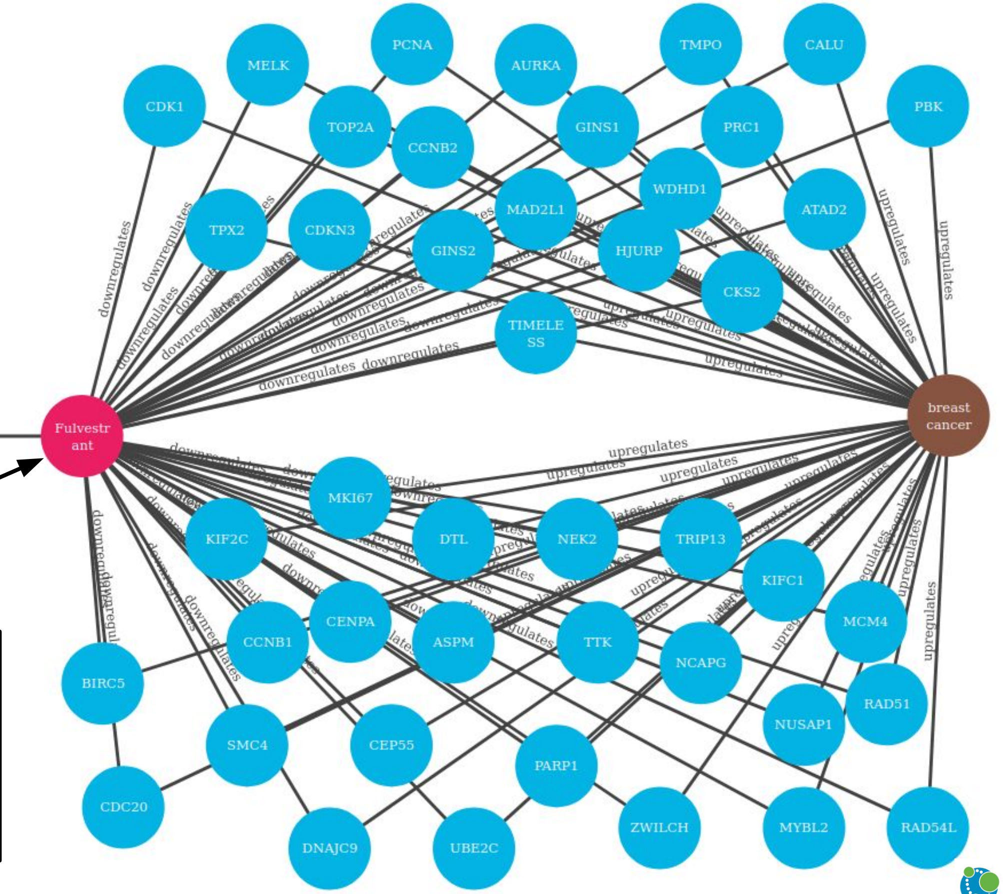


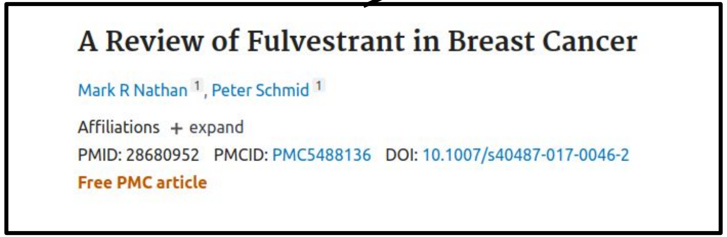


Extend contextualisation with Biomedical Knowledge Graph

Connectivity search

Give it a try: https://het.io/search





A Review of Fulvestrant in Breast Cancer

Mark R Nathan[1], Peter Schmid[1]

Affiliations + expand

PMID: 28680952 PMCID: PMC5488136 DOI: 10.1007/s40487-017-0046-2

Free PMC article

From LPG to Semantic Web

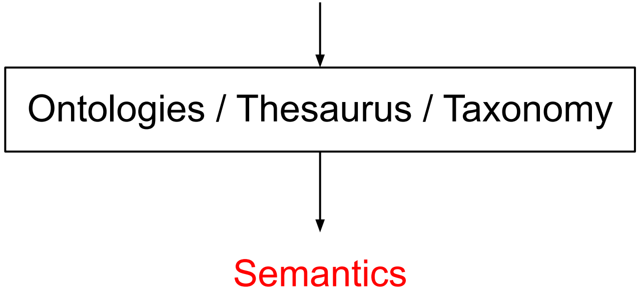
Labeled Property Graph (LPG) - eg. Neo4J

- Efficient extraction of relations (or paths) between entities
- The graph is flexible

Q1: *But what if we want to use relations beyond what is stored in Hetionet ?*
Like information from UniProt, Rhea, Wikidata, etc.

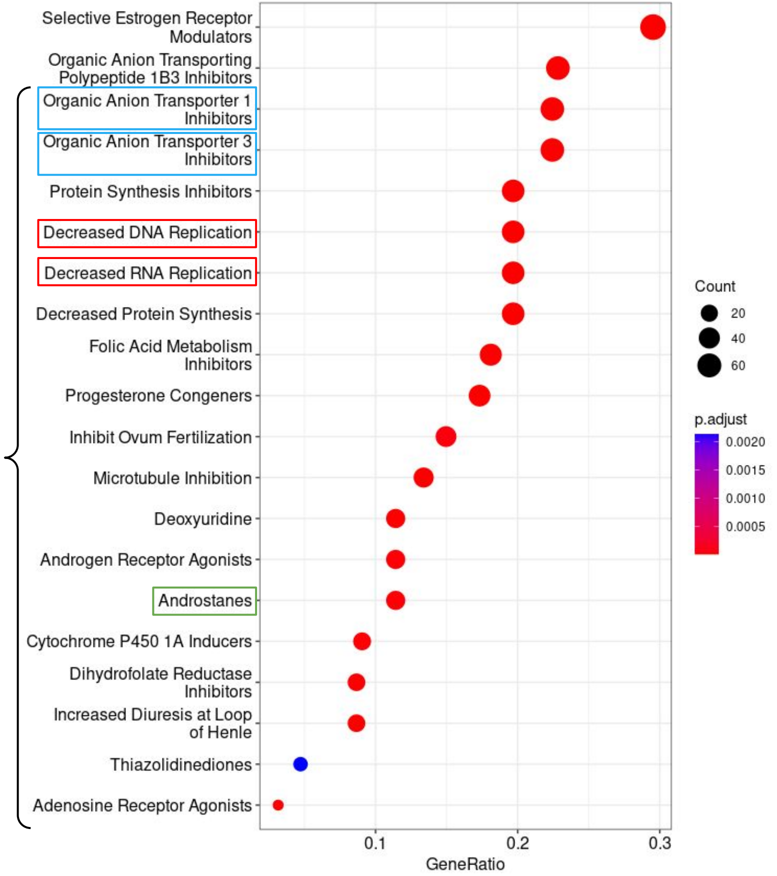
Q2: *What if we want to reason on the graph ?*

We would need something like the Gene **Ontology**, but for pharmacological classes !



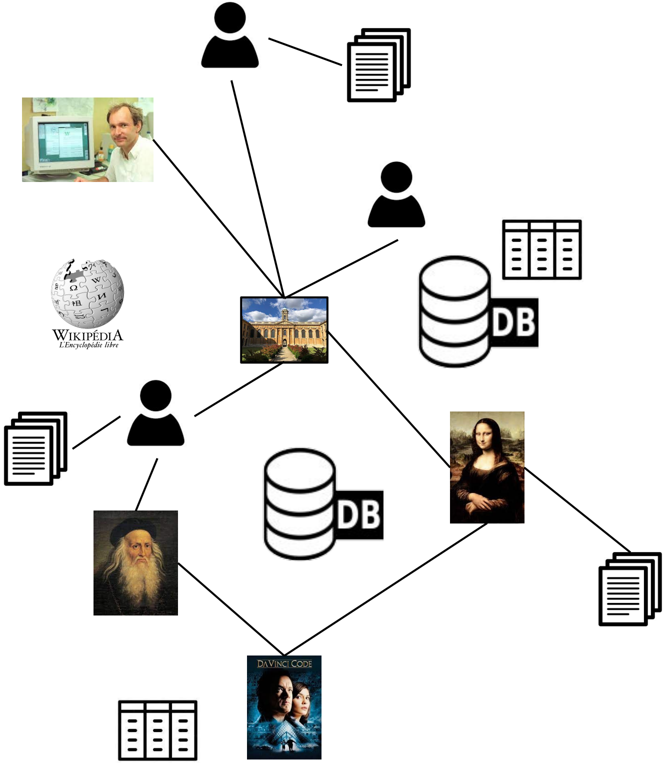
Ontologies / Thesaurus / Taxonomy

Semantics



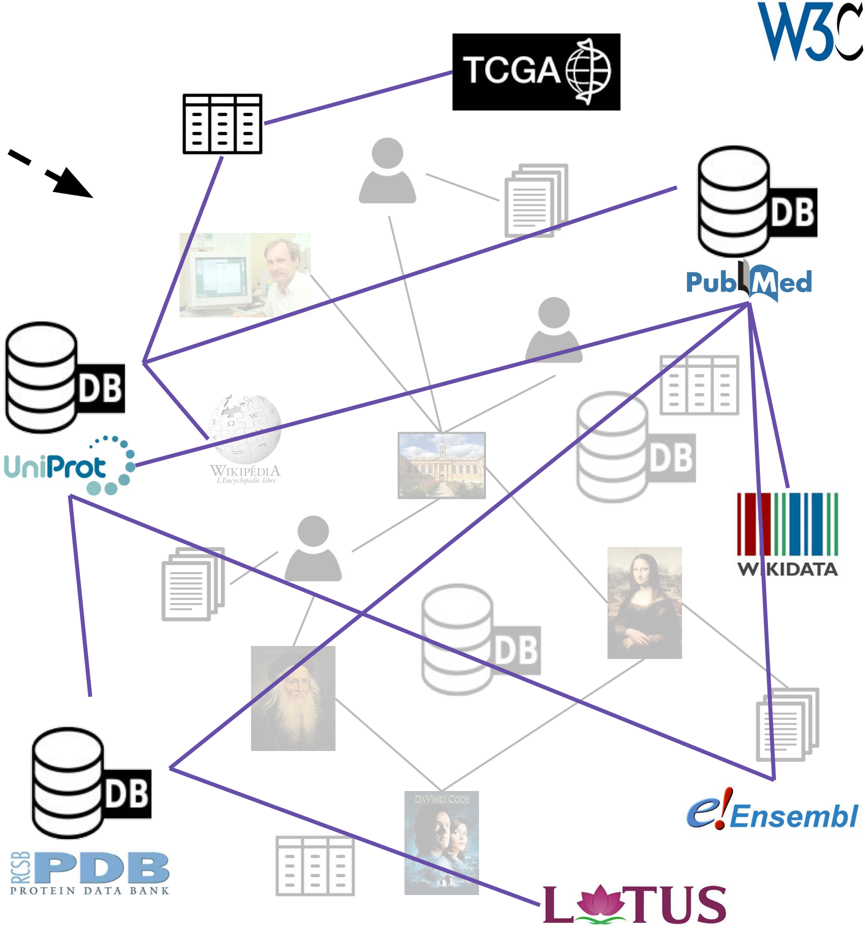
Semantic Web

Semantic Web



Semantic Web

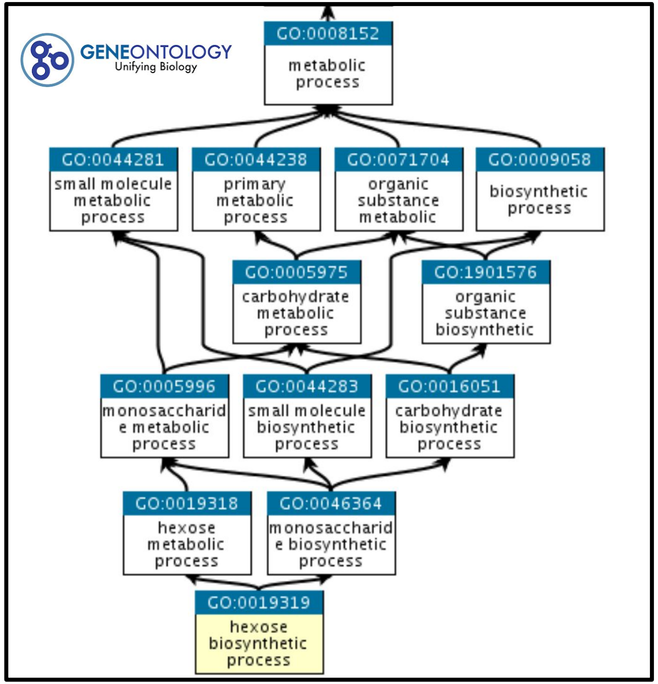
Semantic Web

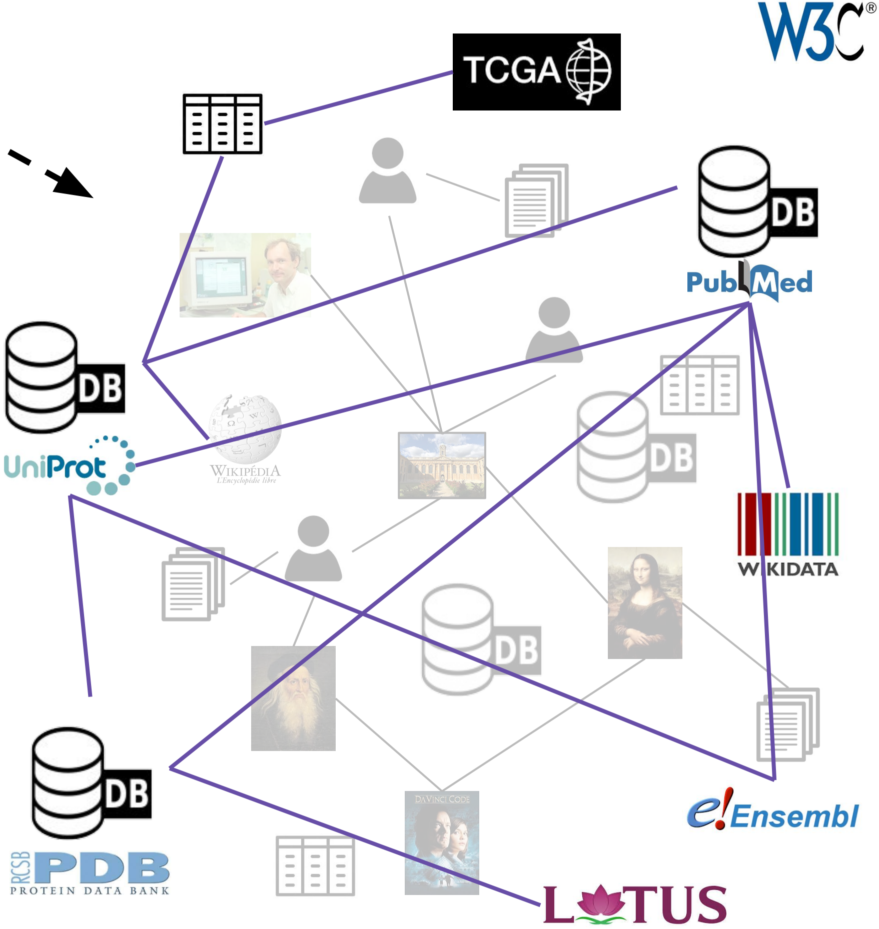


A collection of resources

Semantic Web

Semantic Web

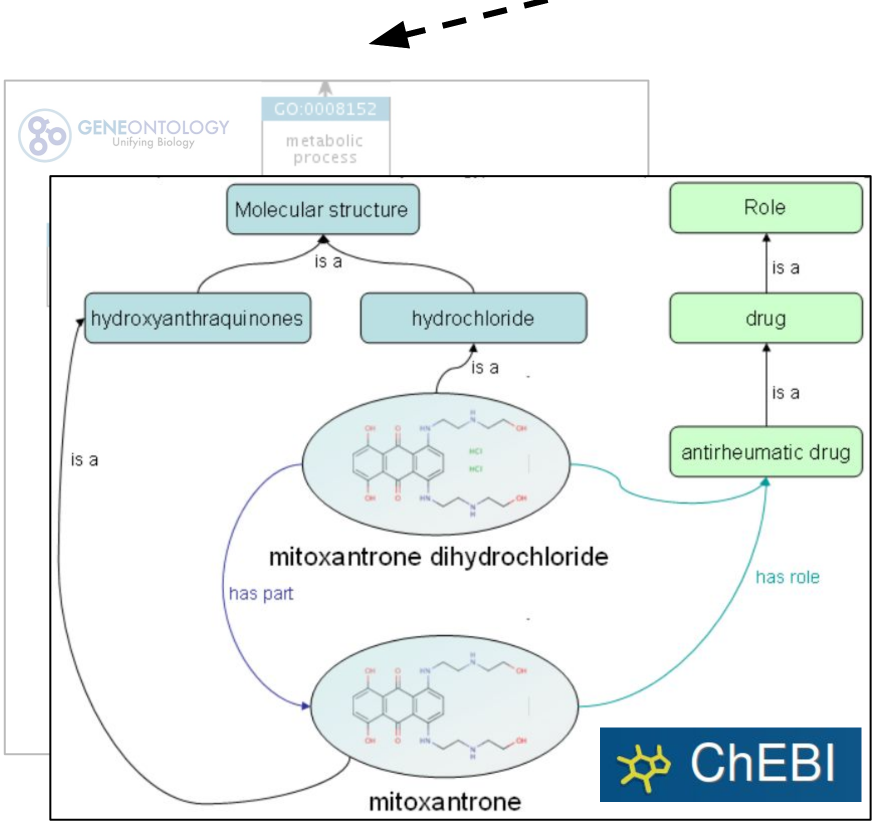

GENEONTOLOGY
Unifying Biology
GO:0008152
metabolic process
GO:0044281
small molecule metabolic process
GO:0044238
primary metabolic process
GO:0071704
organic substance metabolic
GO:0009058
biosynthetic process
GO:0005975
carbohydrate metabolic process
GO:1901576
organic substance biosynthetic
GO:0005996
monosaccharide metabolic process
GO:0044283
small molecule biosynthetic process
GO:0016051
carbohydrate biosynthetic process
GO:0019318
hexose metabolic process
GO:0046364
monosaccharide biosynthetic process
GO:0019319
hexose biosynthetic process

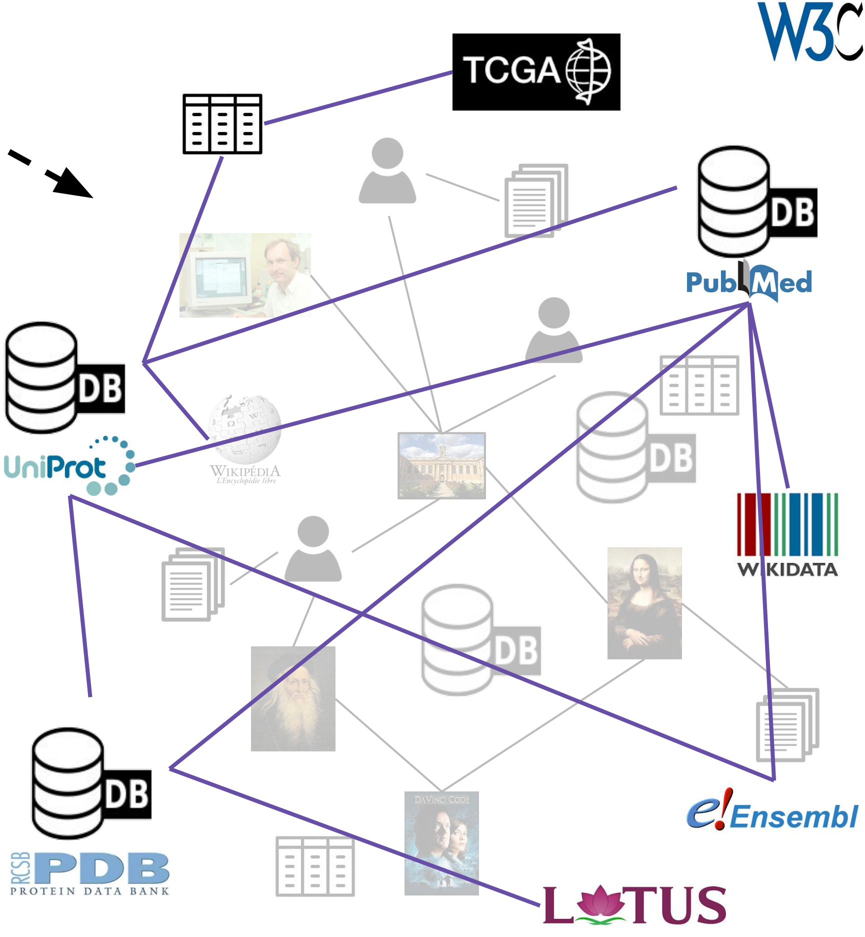


A collection of resources

Semantic Web

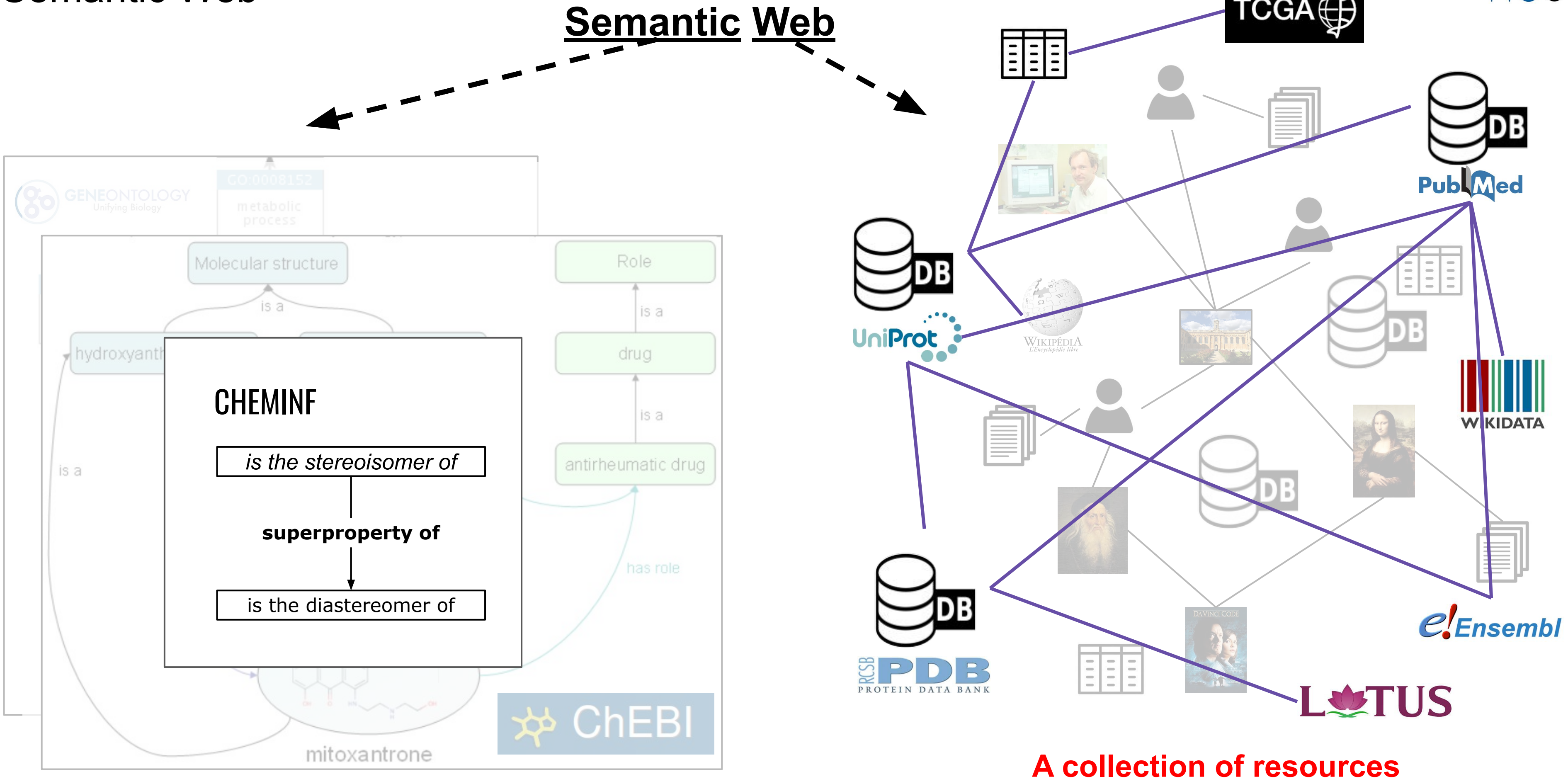
Semantic Web





A collection of resources

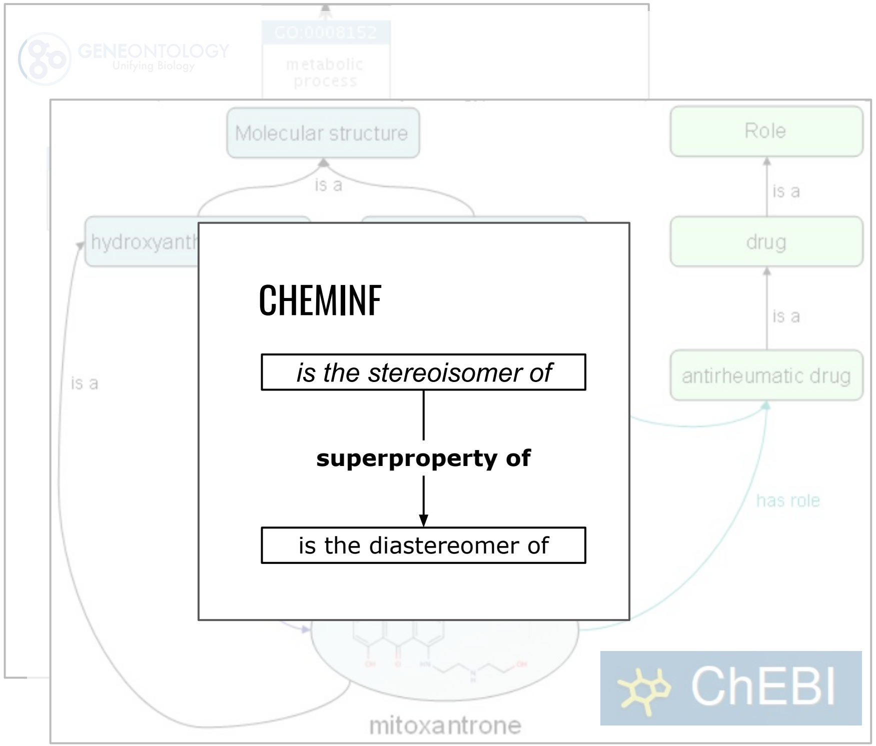
Semantic Web



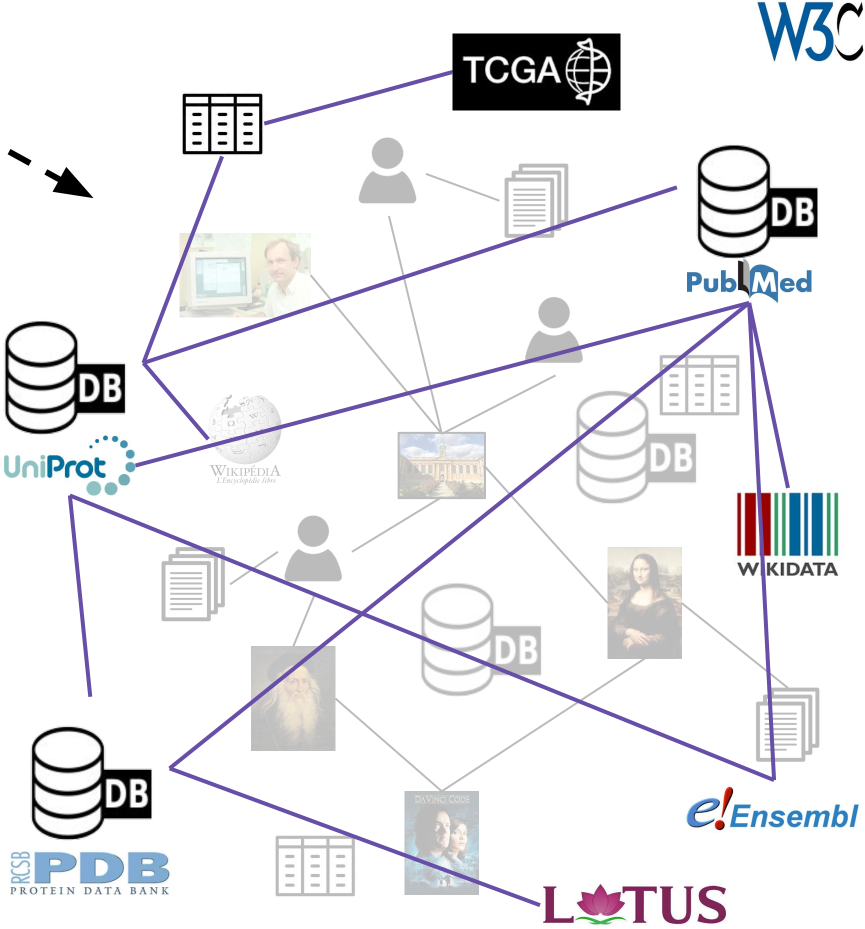
A collection of resources

Semantic Web

Semantic Web



A Semantic description of entities and relations

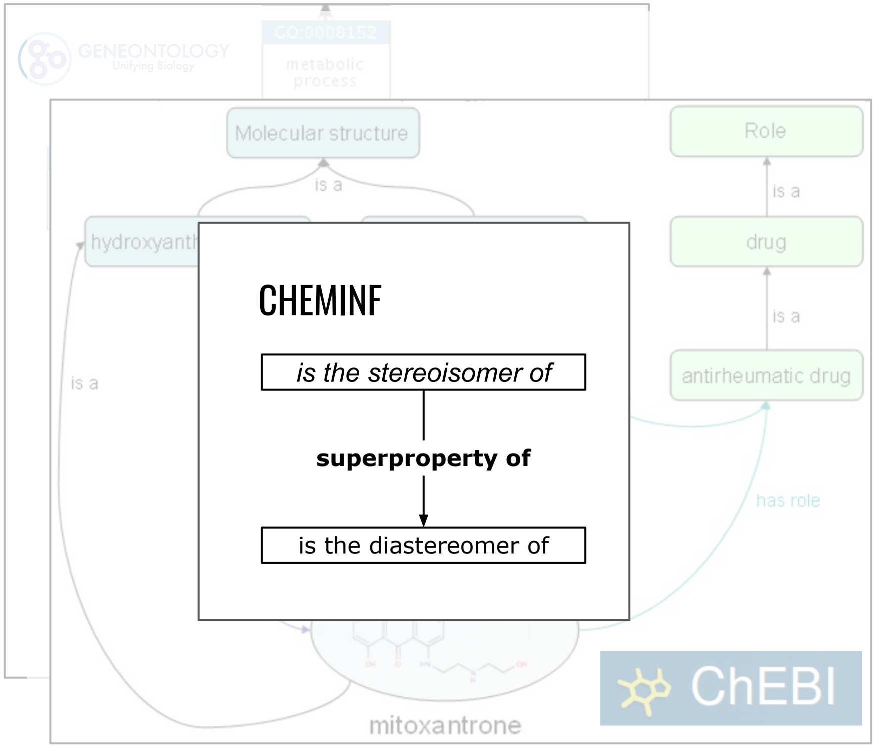


A collection of resources

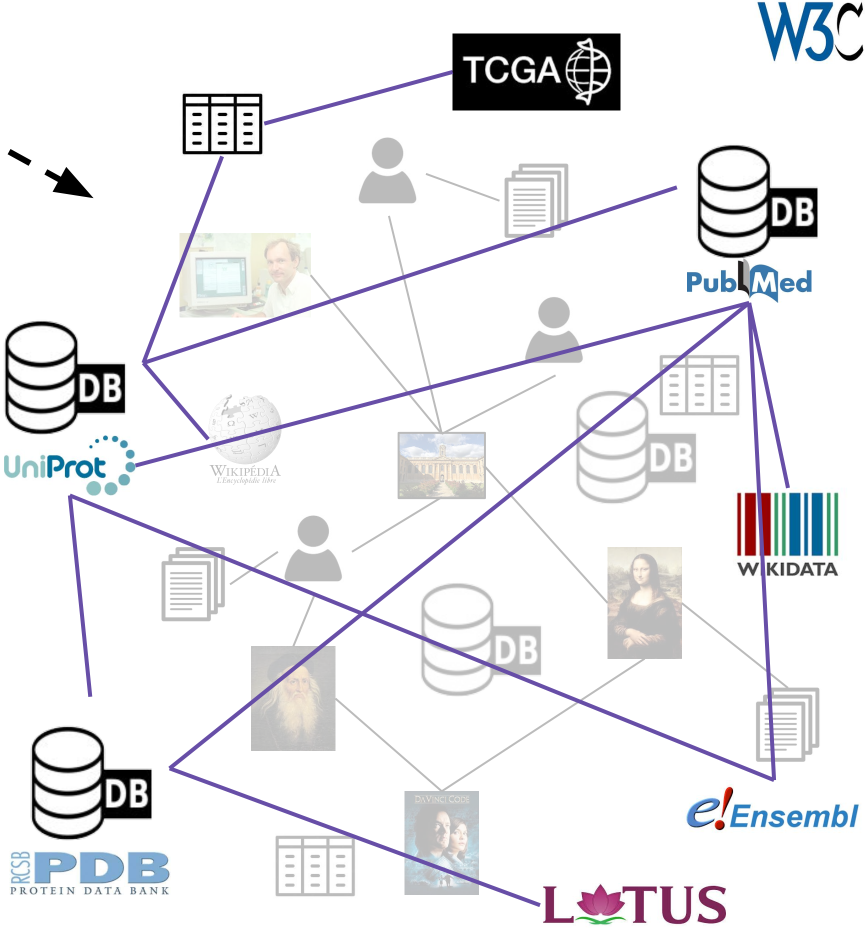
Semantic Web

Semantic Web

Bringing meaning



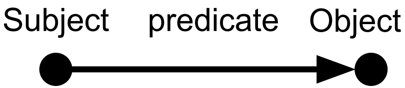
A Semantic description of entities and relations



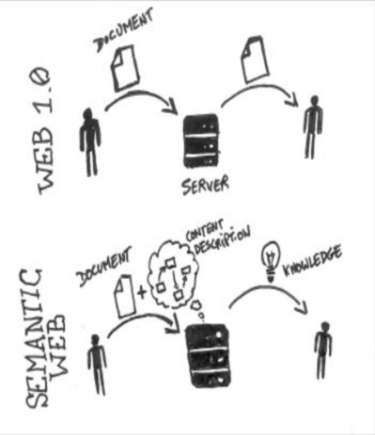
A collection of resources

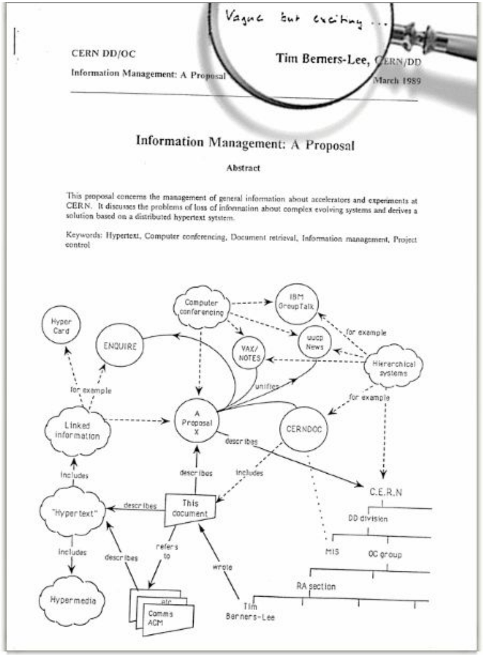
Semantic Web: technical introduction

A common formalism:



Subject predicate Object





Vague but exciting ...

CERN DD/OC

Information Management: A Proposal

Tim Berners-Lee, CERN/DD

March 1989

Information Management: A Proposal

Abstract

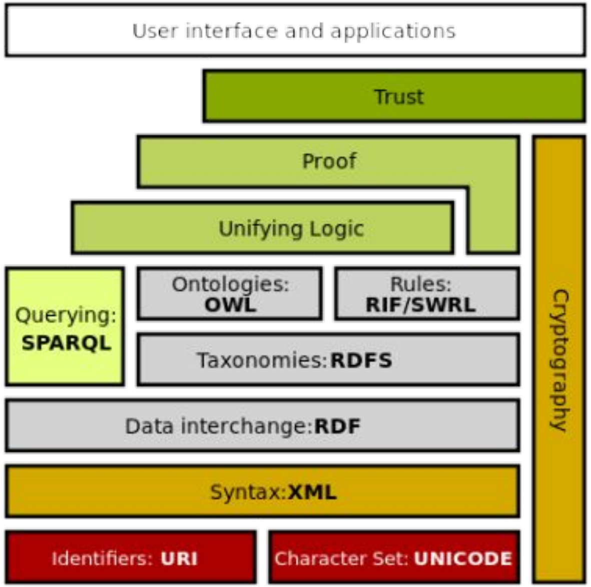
This proposal concerns the management of general information about accelerators and experiments at CERN. It discusses the problems of loss of information about complex evolving systems and derives a solution based on a distributed hypertext sytstem.

Keywords: Hypertext, Computer conferencing, Document retrieval, Information management, Project control

BEHIND THE SCENES

A stack of technologies

RDF



User interface and applications

Trust

Proof

Unifying Logic

Querying: **SPARQL**

Ontologies: **OWL**

Rules: **RIF/SWRL**

Taxonomies: **RDFS**

Data interchange: **RDF**

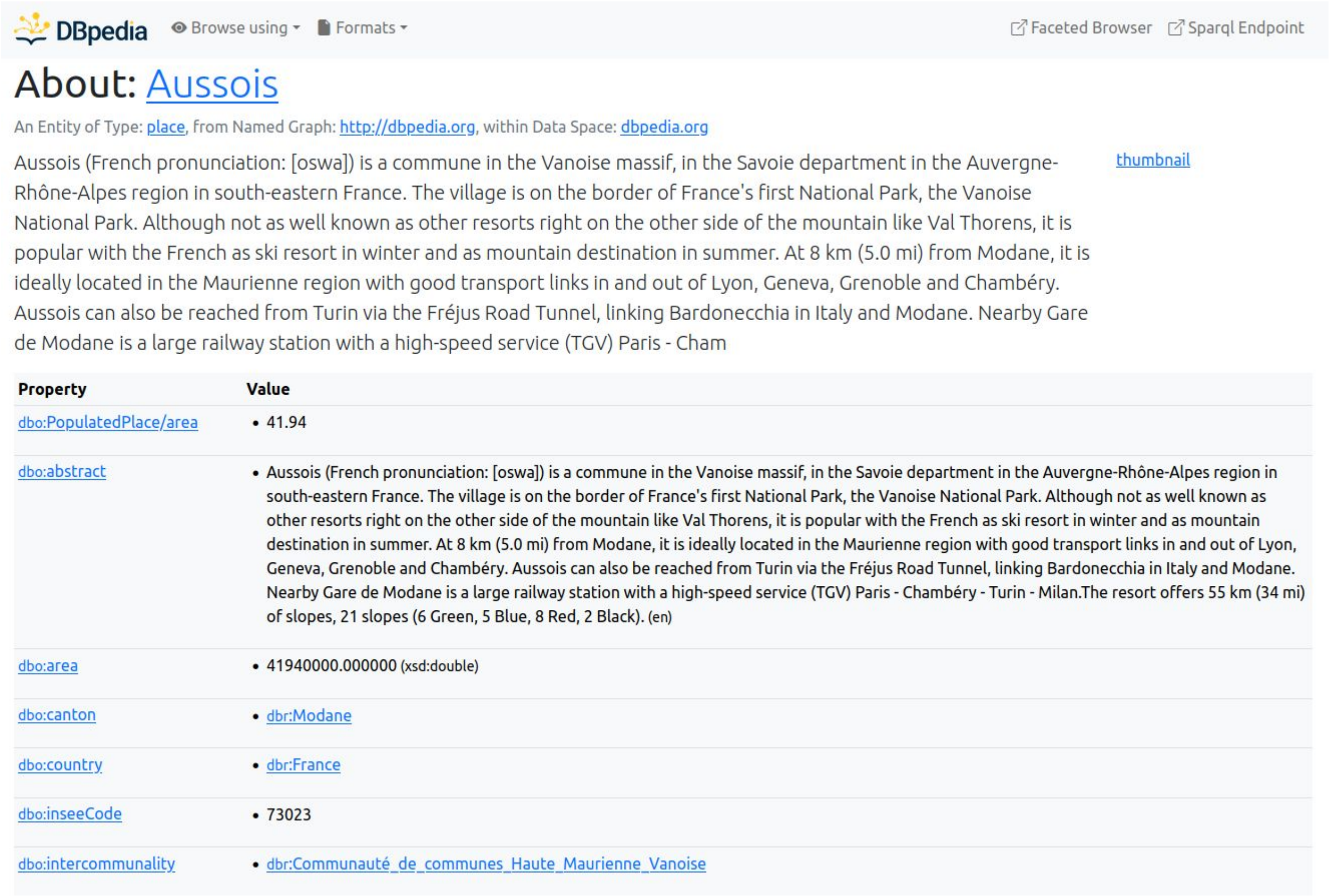
Syntax: **XML**

Identifiers: **URI**

Character Set: **UNICODE**

Cryptography

Even Aussois is described in a Knowledge Graph …



About: Aussois

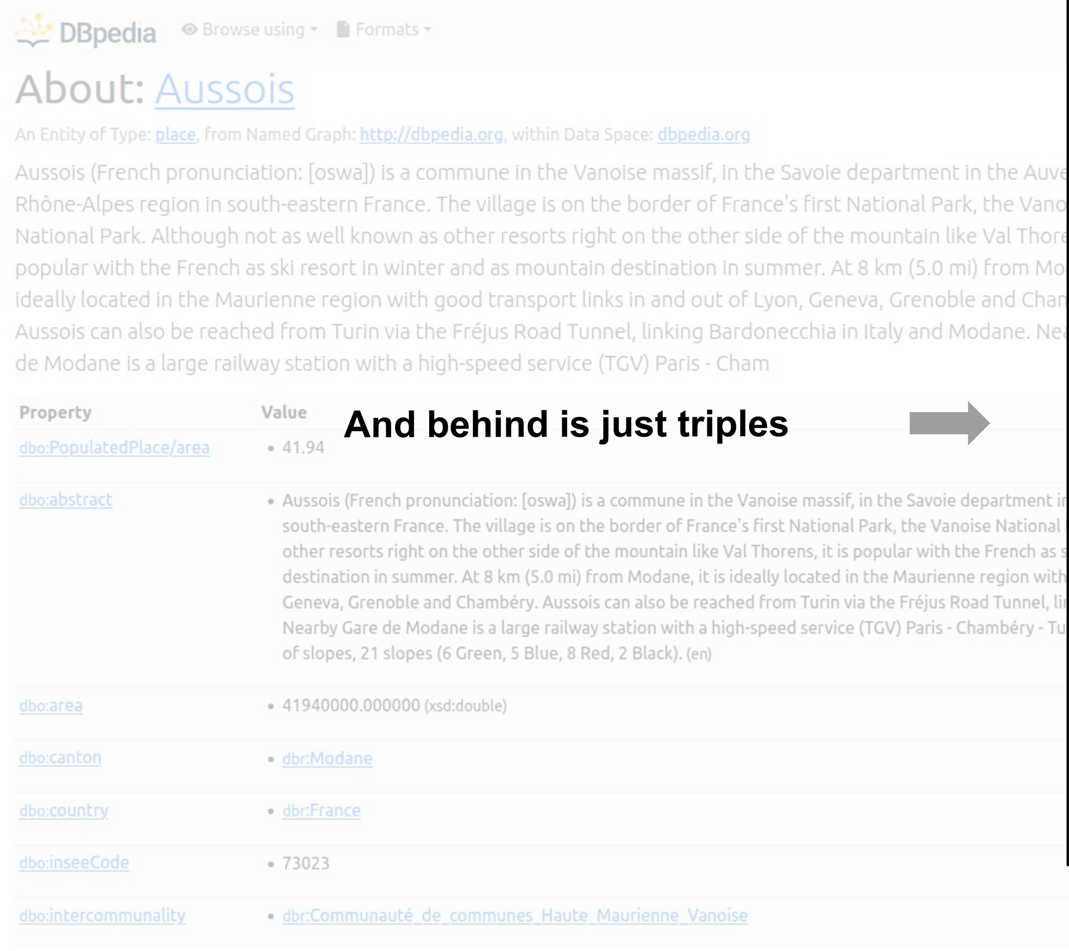
An Entity of Type: place, from Named Graph: http://dbpedia.org, within Data Space: dbpedia.org

Aussois (French pronunciation: [oswa]) is a commune in the Vanoise massif, in the Savoie department in the Auvergne-Rhône-Alpes region in south-eastern France. The village is on the border of France's first National Park, the Vanoise National Park. Although not as well known as other resorts right on the other side of the mountain like Val Thorens, it is popular with the French as ski resort in winter and as mountain destination in summer. At 8 km (5.0 mi) from Modane, it is ideally located in the Maurienne region with good transport links in and out of Lyon, Geneva, Grenoble and Chambéry. Aussois can also be reached from Turin via the Fréjus Road Tunnel, linking Bardonecchia in Italy and Modane. Nearby Gare de Modane is a large railway station with a high-speed service (TGV) Paris - Cham

thumbnail

Property	Value
dbo:PopulatedPlace/area	• 41.94
dbo:abstract	• Aussois (French pronunciation: [oswa]) is a commune in the Vanoise massif, in the Savoie department in the Auvergne-Rhône-Alpes region in south-eastern France. The village is on the border of France's first National Park, the Vanoise National Park. Although not as well known as other resorts right on the other side of the mountain like Val Thorens, it is popular with the French as ski resort in winter and as mountain destination in summer. At 8 km (5.0 mi) from Modane, it is ideally located in the Maurienne region with good transport links in and out of Lyon, Geneva, Grenoble and Chambéry. Aussois can also be reached from Turin via the Fréjus Road Tunnel, linking Bardonecchia in Italy and Modane. Nearby Gare de Modane is a large railway station with a high-speed service (TGV) Paris - Chambéry - Turin - Milan.The resort offers 55 km (34 mi) of slopes, 21 slopes (6 Green, 5 Blue, 8 Red, 2 Black). (en)
dbo:area	• 41940000.000000 (xsd:double)
dbo:canton	• dbr:Modane
dbo:country	• dbr:France
dbo:inseeCode	• 73023
dbo:intercommunality	• dbr:Communauté_de_communes_Haute_Maurienne_Vanoise

Even Aussois is described in a Knowledge Graph …



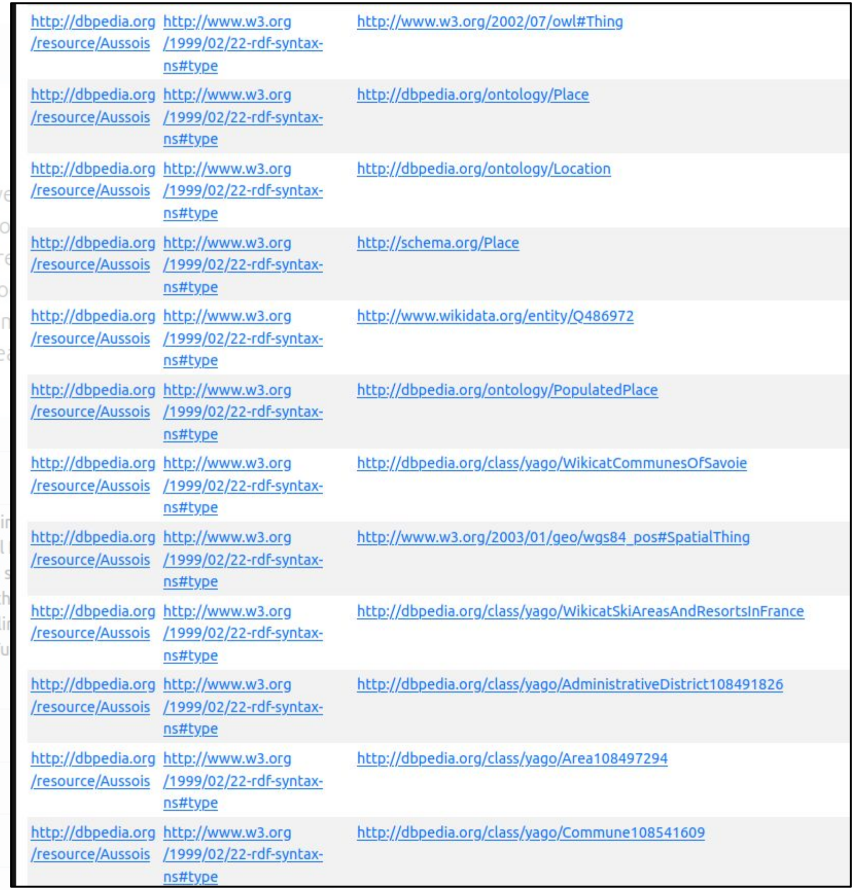
About: Aussois

An Entity of Type: place, from Named Graph: http://dbpedia.org, within Data Space: dbpedia.org

Aussois (French pronunciation: [oswa]) is a commune in the Vanoise massif, in the Savoie department in the Auve
Rhône-Alpes region in south-eastern France. The village is on the border of France's first National Park, the Vano
National Park. Although not as well known as other resorts right on the other side of the mountain like Val Thore
popular with the French as ski resort in winter and as mountain destination in summer. At 8 km (5.0 mi) from Mo
ideally located in the Maurienne region with good transport links in and out of Lyon, Geneva, Grenoble and Chan
Aussois can also be reached from Turin via the Fréjus Road Tunnel, linking Bardonecchia in Italy and Modane. Ne
de Modane is a large railway station with a high-speed service (TGV) Paris - Cham

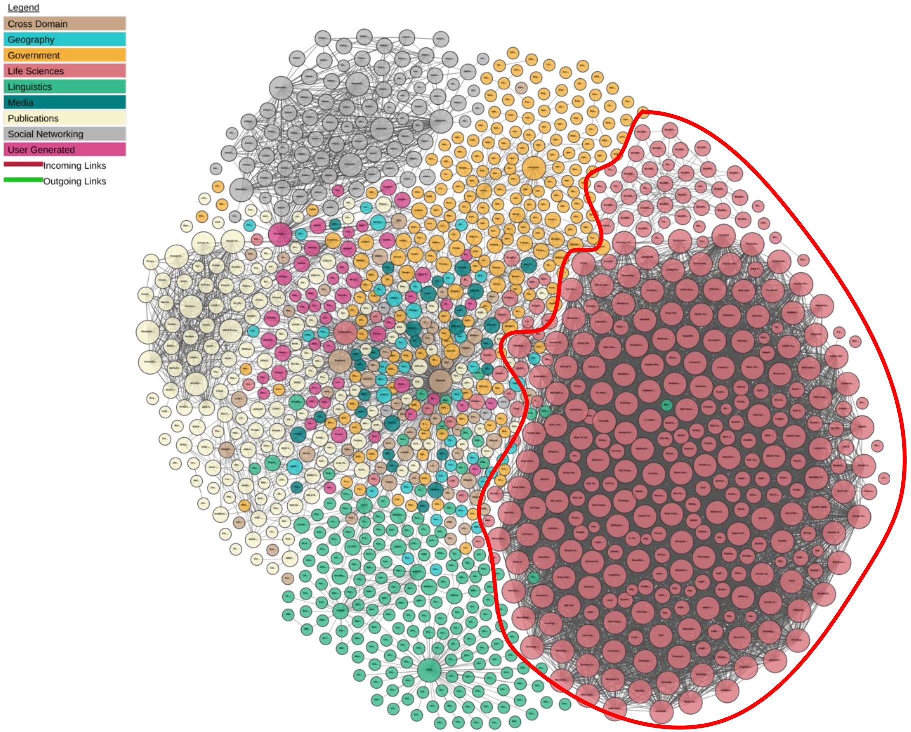
Property	Value
dbo:PopulatedPlace/area	• 41.94
dbo:abstract	• Aussois (French pronunciation: [oswa]) is a commune in the Vanoise massif, in the Savoie department i south-eastern France. The village is on the border of France's first National Park, the Vanoise National other resorts right on the other side of the mountain like Val Thorens, it is popular with the French as s destination in summer. At 8 km (5.0 mi) from Modane, it is ideally located in the Maurienne region with Geneva, Grenoble and Chambéry. Aussois can also be reached from Turin via the Fréjus Road Tunnel, li Nearby Gare de Modane is a large railway station with a high-speed service (TGV) Paris - Chambéry - Tu of slopes, 21 slopes (6 Green, 5 Blue, 8 Red, 2 Black). (en)
dbo:area	• 41940000.000000 (xsd:double)
dbo:canton	• dbr:Modane
dbo:country	• dbr:France
dbo:inseeCode	• 73023
dbo:intercommunality	• dbr:Communauté_de_communes_Haute_Maurienne_Vanoise

And behind is just triples →



http://dbpedia.org/resource/Aussois	http://www.w3.org/1999/02/22-rdf-syntax-ns#type	http://www.w3.org/2002/07/owl#Thing
http://dbpedia.org/resource/Aussois	http://www.w3.org/1999/02/22-rdf-syntax-ns#type	http://dbpedia.org/ontology/Place
http://dbpedia.org/resource/Aussois	http://www.w3.org/1999/02/22-rdf-syntax-ns#type	http://dbpedia.org/ontology/Location
http://dbpedia.org/resource/Aussois	http://www.w3.org/1999/02/22-rdf-syntax-ns#type	http://schema.org/Place
http://dbpedia.org/resource/Aussois	http://www.w3.org/1999/02/22-rdf-syntax-ns#type	http://www.wikidata.org/entity/Q486972
http://dbpedia.org/resource/Aussois	http://www.w3.org/1999/02/22-rdf-syntax-ns#type	http://dbpedia.org/ontology/PopulatedPlace
http://dbpedia.org/resource/Aussois	http://www.w3.org/1999/02/22-rdf-syntax-ns#type	http://dbpedia.org/class/yago/WikicatCommunesOfSavoie
http://dbpedia.org/resource/Aussois	http://www.w3.org/1999/02/22-rdf-syntax-ns#type	http://www.w3.org/2003/01/geo/wgs84_pos#SpatialThing
http://dbpedia.org/resource/Aussois	http://www.w3.org/1999/02/22-rdf-syntax-ns#type	http://dbpedia.org/class/yago/WikicatSkiAreasAndResortsInFrance
http://dbpedia.org/resource/Aussois	http://www.w3.org/1999/02/22-rdf-syntax-ns#type	http://dbpedia.org/class/yago/AdministrativeDistrict108491826
http://dbpedia.org/resource/Aussois	http://www.w3.org/1999/02/22-rdf-syntax-ns#type	http://dbpedia.org/class/yago/Area108497294
http://dbpedia.org/resource/Aussois	http://www.w3.org/1999/02/22-rdf-syntax-ns#type	http://dbpedia.org/class/yago/Commune108541609

The Web of Life Sciences



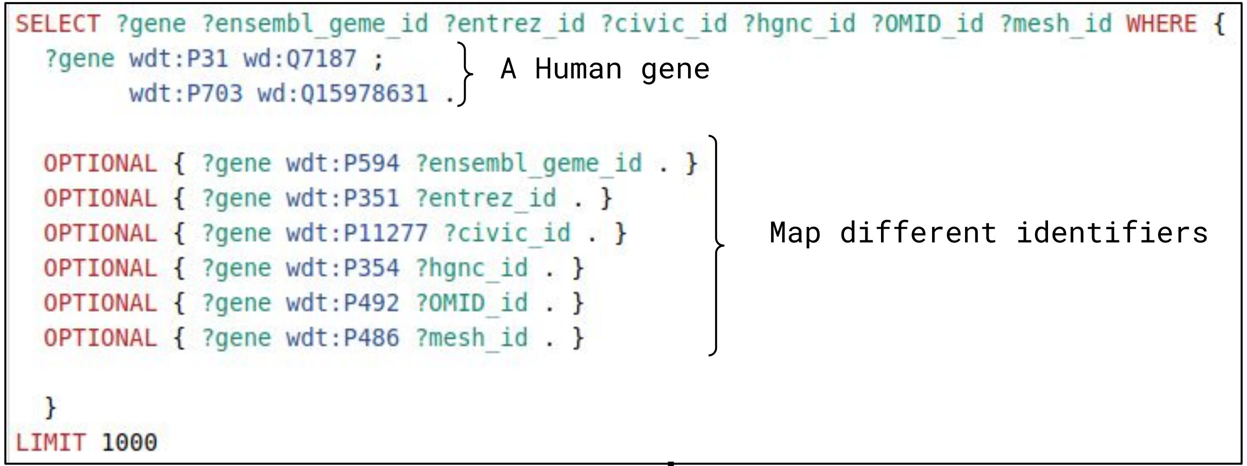
Life science datasets

MeSH	222,843,516
UniProt	188,410,158,927 entries
FORVM	8,780,371,866 entries
WikiPathways Pathways for the People	38,733,142 entries
DisGeNET	198,762,693 entries
Rhea SIB	1,366,976 entries

...

Data Sharing
+
decentralisation

The Web of Life Sciences: As a identifier mapping tool



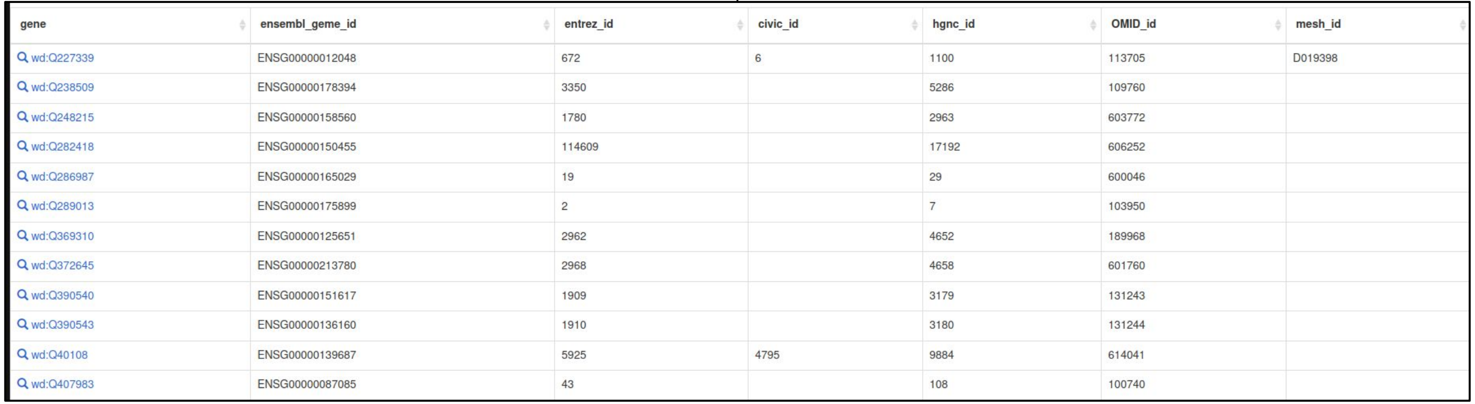
```
SELECT ?gene ?ensembl_geme_id ?entrez_id ?civic_id ?hgnc_id ?OMID_id ?mesh_id WHERE {
  ?gene wdt:P31 wd:Q7187 ;
        wdt:P703 wd:Q15978631 .

  OPTIONAL { ?gene wdt:P594 ?ensembl_geme_id . }
  OPTIONAL { ?gene wdt:P351 ?entrez_id . }
  OPTIONAL { ?gene wdt:P11277 ?civic_id . }
  OPTIONAL { ?gene wdt:P354 ?hgnc_id . }
  OPTIONAL { ?gene wdt:P492 ?OMID_id . }
  OPTIONAL { ?gene wdt:P486 ?mesh_id . }

  }
LIMIT 1000
```

A Human gene

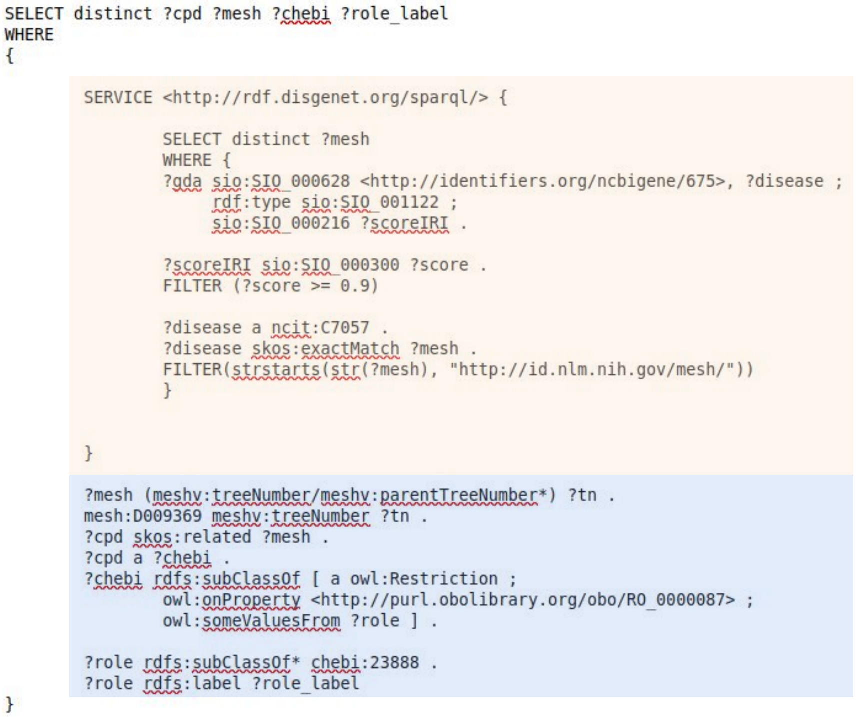
Map different identifiers



gene	ensembl_geme_id	entrez_id	civic_id	hgnc_id	OMID_id	mesh_id
wd:Q227339	ENSG00000012048	672	6	1100	113705	D019398
wd:Q238509	ENSG00000178394	3350		5286	109760	
wd:Q248215	ENSG00000158560	1780		2963	603772	
wd:Q282418	ENSG00000150455	114609		17192	606252	
wd:Q286987	ENSG00000165029	19		29	600046	
wd:Q289013	ENSG00000175899	2		7	103950	
wd:Q369310	ENSG00000125651	2962		4652	189968	
wd:Q372645	ENSG00000213780	2968		4658	601760	
wd:Q390540	ENSG00000151617	1909		3179	131243	
wd:Q390543	ENSG00000136160	1910		3180	131244	
wd:Q40108	ENSG00000139687	5925	4795	9884	614041	
wd:Q407983	ENSG00000087085	43		108	100740	

The Web of Life Sciences: A federated query example

Find compounds used as drugs for diseases caused by mutations on BRCA2



```
SELECT distinct ?cpd ?mesh ?chebi ?role_label
WHERE
{
        SERVICE <http://rdf.disgenet.org/sparql/> {

                SELECT distinct ?mesh
                WHERE {
                ?gda sio:SIO_000628 <http://identifiers.org/ncbigene/675>, ?disease ;
                     rdf:type sio:SIO_001122 ;
                     sio:SIO_000216 ?scoreIRI .

                ?scoreIRI sio:SIO_000300 ?score .
                FILTER (?score >= 0.9)

                ?disease a ncit:C7057 .
                ?disease skos:exactMatch ?mesh .
                FILTER(strstarts(str(?mesh), "http://id.nlm.nih.gov/mesh/"))
                }


        }
        ?mesh (meshv:treeNumber/meshv:parentTreeNumber*) ?tn .
        mesh:D009369 meshv:treeNumber ?tn .
        ?cpd skos:related ?mesh .
        ?cpd a ?chebi .
        ?chebi rdfs:subClassOf [ a owl:Restriction ;
                owl:onProperty <http://purl.obolibrary.org/obo/RO_0000087> ;
                owl:someValuesFrom ?role ] .

        ?role rdfs:subClassOf* chebi:23888 .
        ?role rdfs:label ?role_label
}
```

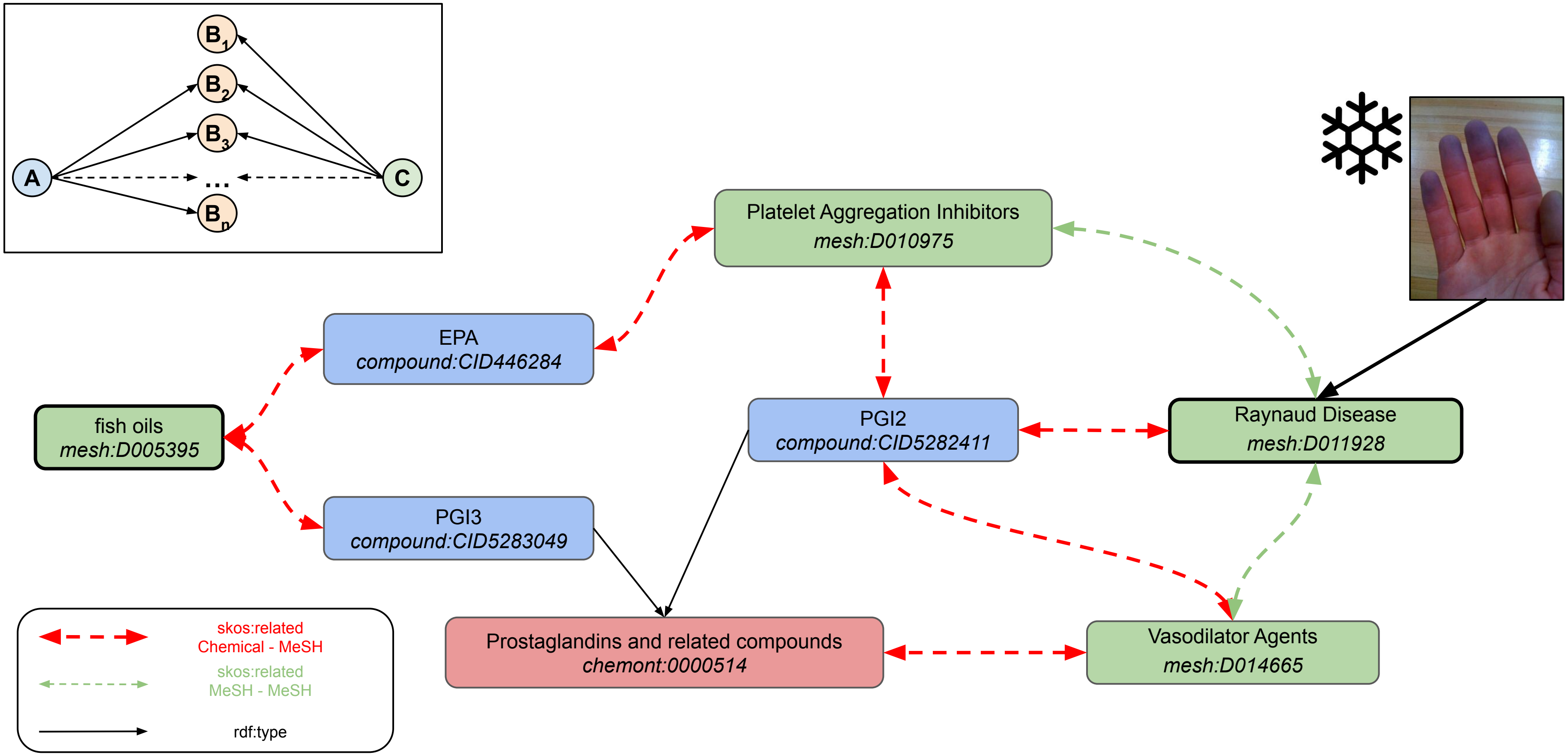
DisGeNET

Get all disease for which BRCA2 is a *biomarker*

FORVM

Find all related compounds that are classified as *drug*

The Web of Life Sciences: An example of Literature discovery



Swanson, D.R., 1986. Fish Oil, Raynaud's Syndrome, and Undiscovered Public Knowledge. Perspectives in Biology and Medicine

Biomedical Knowledge Graph

- Biomedical KG can help to explore new connections between entities
- They can also be used for building a custom background set in enrichment.
- Semantic Web act as a bridge between biomedical databases on the Web
  - An unified framework to describe entities and their relations
  - Integrates vocabulary, ontologies for a semantic description
- Need to understand the **schema** of the KG, before requesting
- Build your own Biomedical KG ! Use **BioCypher**

biocypher

a unifying framework for
biomedical knowledge graphs

The End

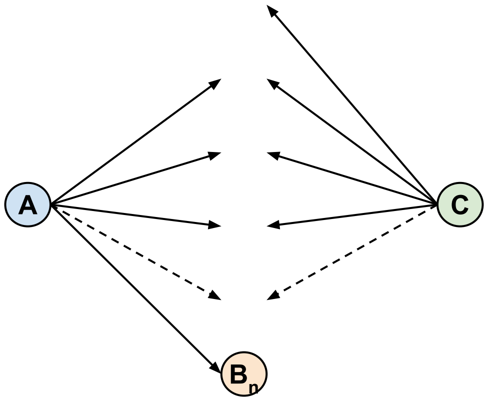
Resources

- 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

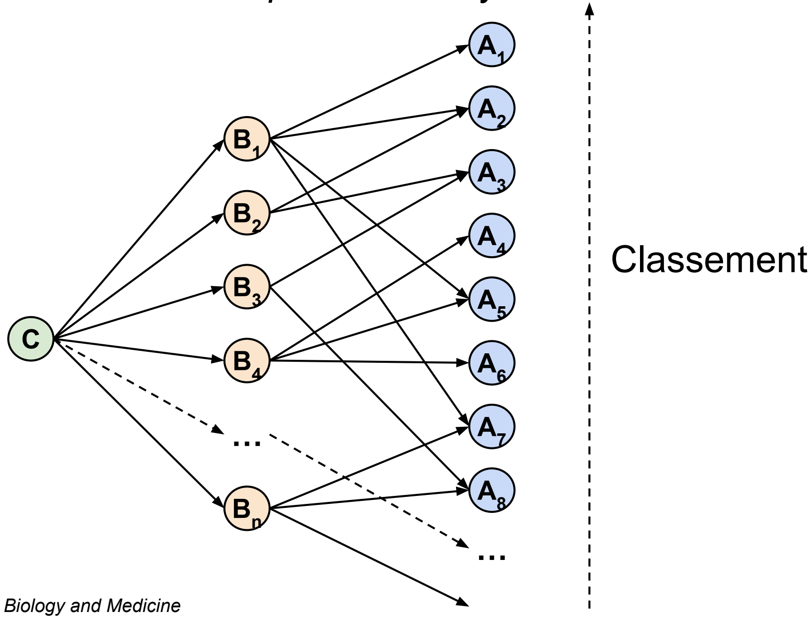
Waagmeester, A. et al., 2020. Wikidata as a knowledge graph for the life sciences. eLife 9, e52614.

Resources

- https://www.wikidata.org/wiki/Wikidata:SPARQL_query_service/queries/examples
- https://www.wikipathways.org/sparql.html
- Others Enrichment analysis methods:
    - https://pubmed.ncbi.nlm.nih.gov/14693814/
    - https://pubmed.ncbi.nlm.nih.gov/15647293/
    - https://pubmed.ncbi.nlm.nih.gov/15941488/



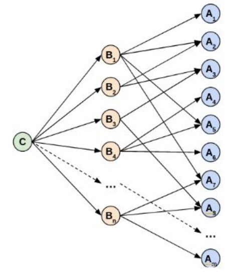
Open discovery

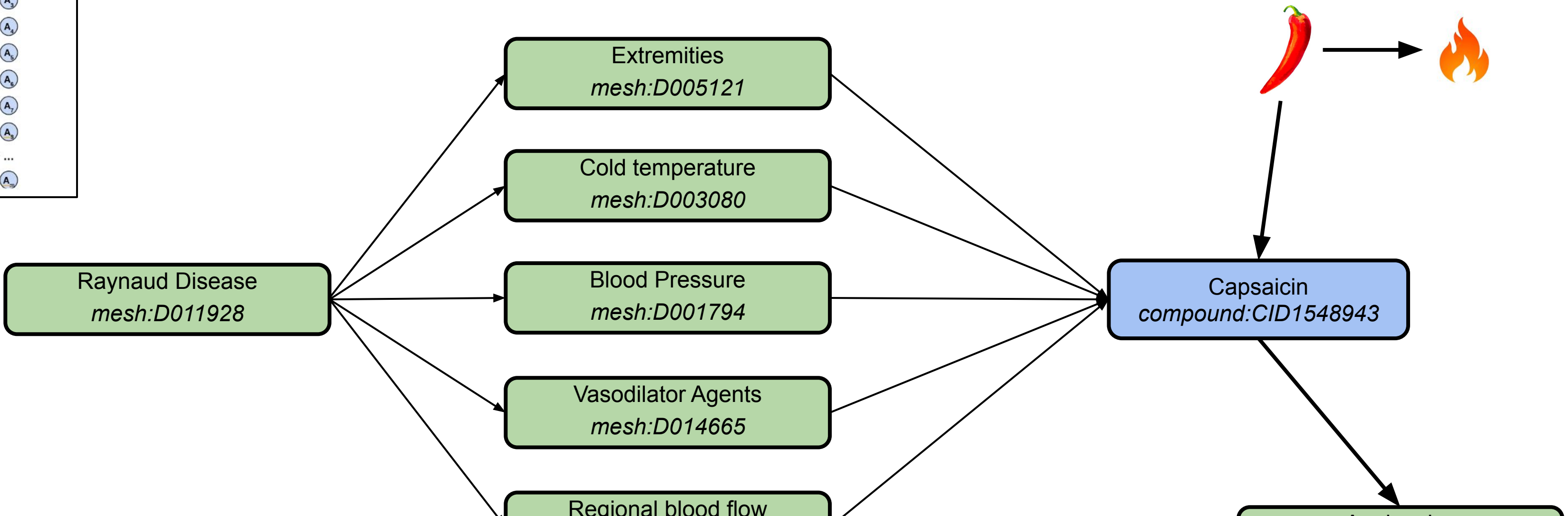


Swanson, D.R., 1986. Fish Oil, Raynaud’s Syndrome, and Undiscovered Public Knowledge. Perspectives in Biology and Medicine

S. Henry and B. T. McInnes. Literature Based Discovery : Models, methods, and trends. Journal of Biomedical Informatics, 74 :20–32, Oct. 2017.

Open Discovery





Raynaud Disease
mesh:D011928

Extremities
mesh:D005121

Cold temperature
mesh:D003080

Blood Pressure
mesh:D001794

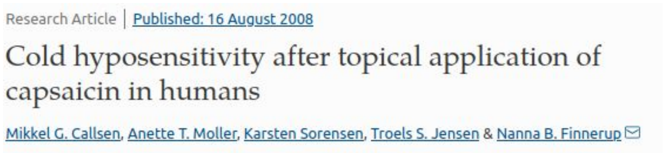
Vasodilator Agents
mesh:D014665

Regional blood flow
mesh:D012039

…
+61 concepts

Capsaicin
compound:CID1548943

Analgesics
mesh:D000700



Research Article | Published: 16 August 2008

Cold hyposensitivity after topical application of capsaicin in humans

Mikkel G. Callsen, Anette T. Moller, Karsten Sorensen, Troels S. Jensen & Nanna B. Finnerup