Enrichment analyses

Irina Mohorianu (CSCI)

Lecture compiled from slides designed by Martin Morgan and Pascale Gaudet

Enrichment analysis. Overview

Lists of differentially expressed genes

↓

[1] Datasets
- GO datasets
- Pathway datasets

[2] Statistical approaches

[3] Interpretation of results

Databases

[1] Why do we perform enrichment analyses?

Too much information available for each gene of interest

PubMed: db of over 15 million citations

Basic search: rad51 → 3929 articles

Organism Limited search:

rad51 AND Human (organism) → 2488 articles

Disease Limited search:

rad51 AND cancer → 1909

[2] What type of information is available?

GO (Gene ontologies): BP (biological process)
CC (cellular component)
MF (molecular function)

Ontologies attach
FACTS to
KNOWLEDGE

GO datasets

GO (Gene ontologies): BP (biological process)
CC (cellular component)
MF (molecular function)

What is a Gene Ontology?

Gene annotation system
Controlled vocabulary that can be applied across organisms
Used to describe gene products and their interactions

or, more formal:

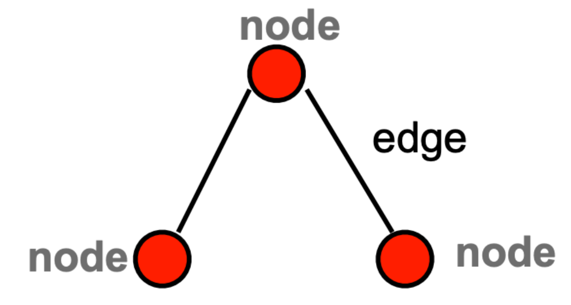
Ontologies provide controlled, consistent vocabularies to describe concepts and relationships, thereby enabling knowledge sharing
– Gruber 1993

GO datasets

What is a Gene Ontology?

GO = a collection of:
Terms
Definitions (spectrum of robustness/ consistency)
Logical relationships

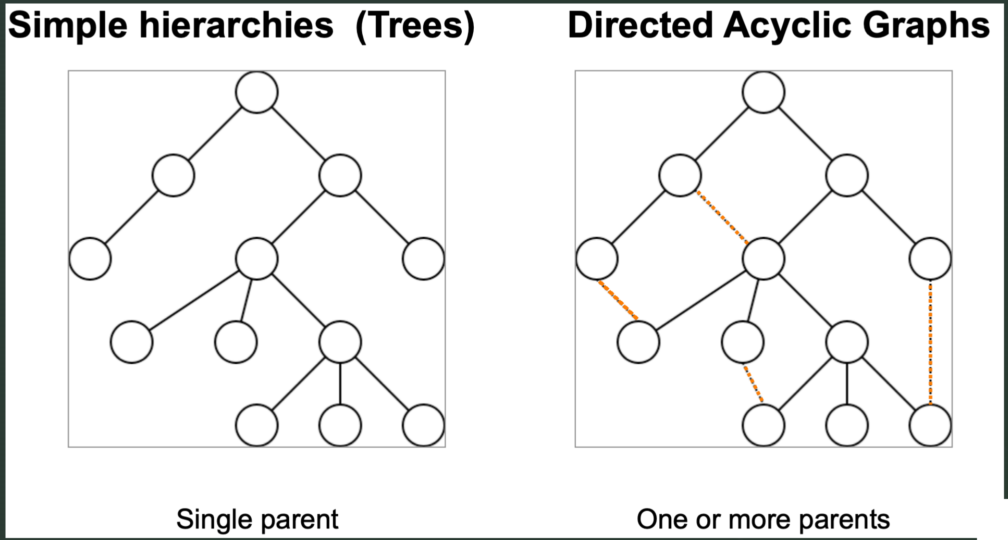
Structured as a graph:
Nodes = concepts in the ontology
Edges = relationships between the concepts



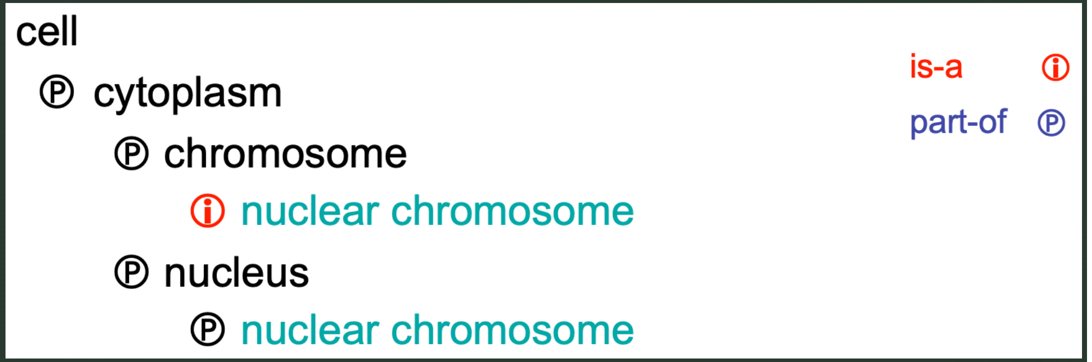
Type of relationships:
is-a
part-of
regulates

GO datasets

Expectation. Reality

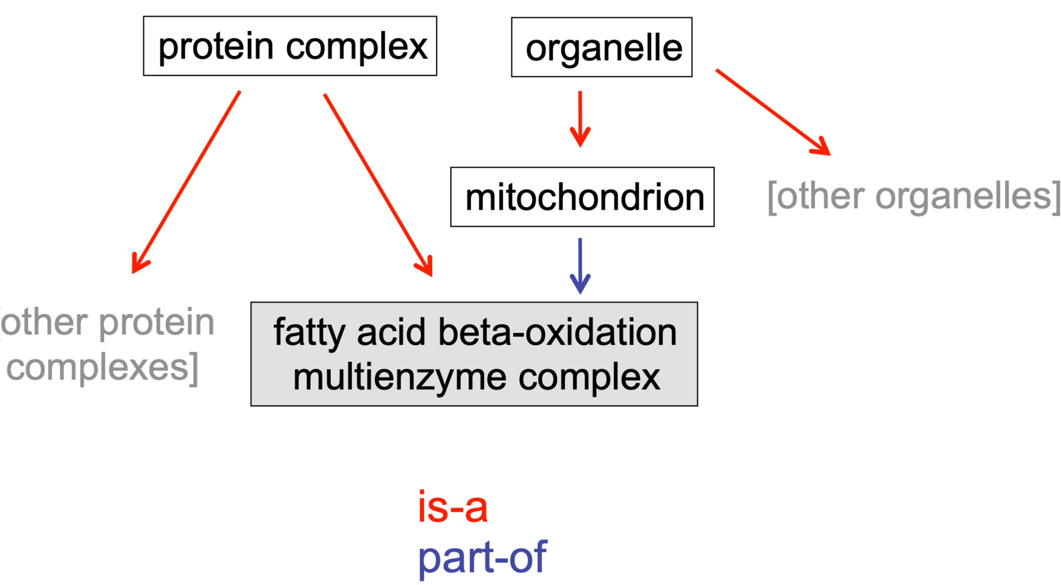


True path rule:
The path from a child term all the way up to its top-level parent(s) must always be true



cell
- Ⓟ cytoplasm
  - Ⓟ chromosome
    - ⓘ nuclear chromosome
  - Ⓟ nucleus
    - Ⓟ nuclear chromosome

is-a ⓘ
part-of Ⓟ


protein complex
organelle
mitochondrion
[other organelles]
[other protein complexes]
fatty acid beta-oxidation multienzyme complex
is-a
part-of

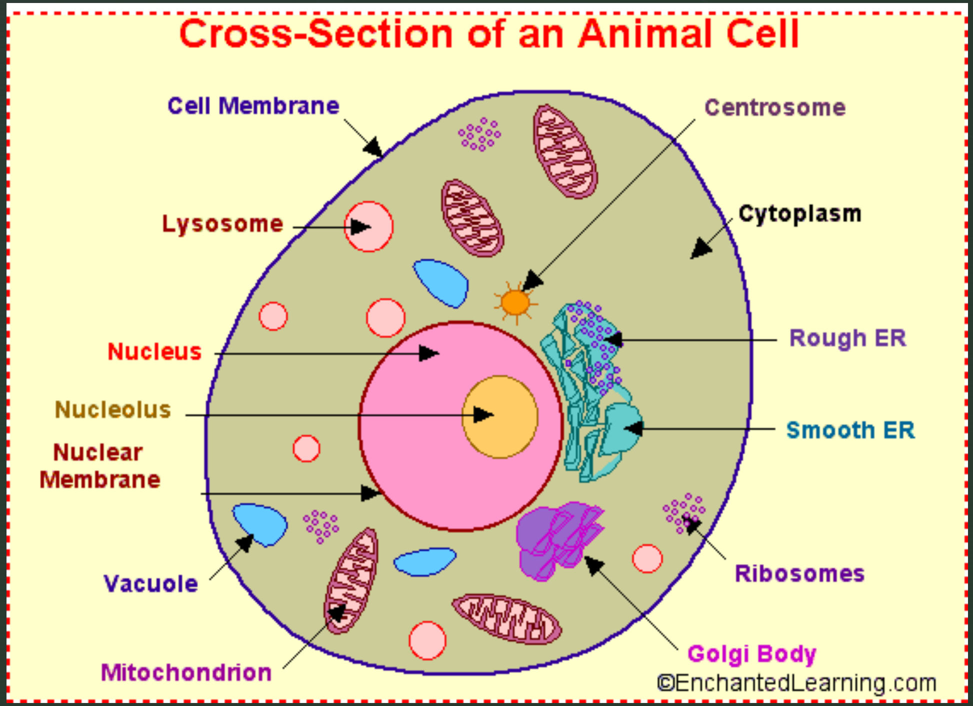
GO datasets

GO (Gene ontologies):

BP (biological process) :: what processes is it involved in?

CC (cellular component) :: where is it?

MF (molecular function) :: what does it do?



Cellular component

The locations relative to cellular structures in which a gene product performs a function, either cellular compartments (e.g., mitochondrion), or stable macromolecular complexes of which they are parts (e.g., the ribosome). Unlike the other aspects of GO, cellular component classes refer not to processes but rather a cellular anatomy.

http://geneontology.org/docs/ontology-documentation/

GO datasets

Molecular function

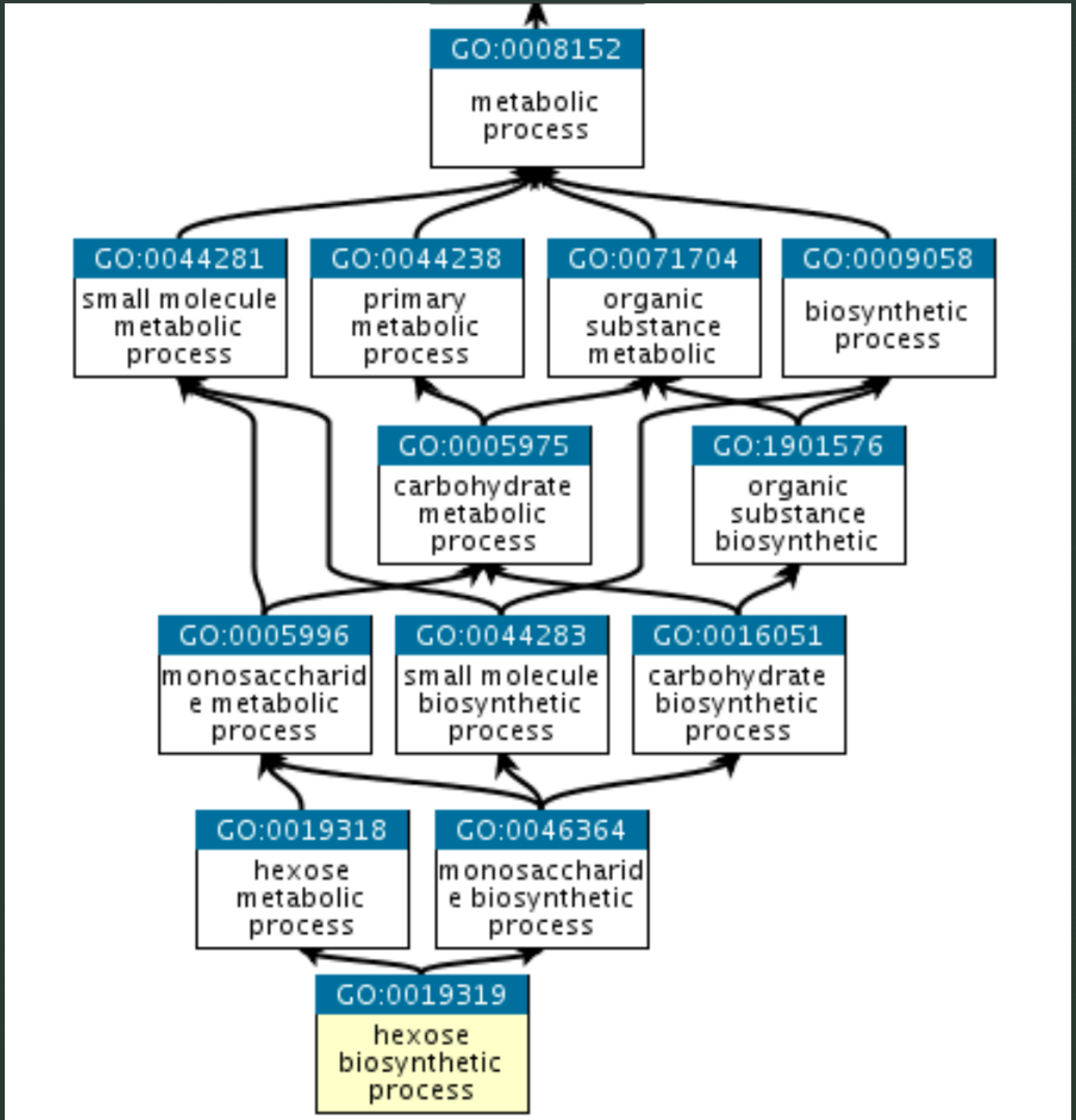
Molecular-level activities performed by gene products. Molecular function terms describe activities that occur at the molecular level, such as "catalysis" or "transport". GO molecular function terms represent activities rather than the entities (molecules or complexes) that perform the actions, and do not specify where, when, or in what context the action takes place. Molecular functions generally correspond to activities that can be performed by individual gene products (*i.e.* a protein or RNA), but some activities are performed by molecular complexes composed of multiple gene products. Examples of broad functional terms are *catalytic activity* and *transporter activity*; examples of narrower functional terms are *adenylate cyclase activity* or *Toll-like receptor binding*.

Biological process

The larger processes, or 'biological programs' accomplished by multiple molecular activities. Examples of broad biological process terms are *DNA repair* or *signal transduction*. Examples of more specific terms are *pyrimidine nucleobase biosynthetic process* or *glucose transmembrane transport*. Note that a biological process is not equivalent to a pathway. At present, the GO does not try to represent the dynamics or dependencies that would be required to fully describe a pathway.

http://geneontology.org/docs/ontology-documentation/

GO datasets


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

The three GO ontologies are is a disjoint, meaning that **no is a relations operate between terms from the different ontologies**. Other relationships e.g. **part of** and **regulates** operate across GO ontologies. E.g. the MF ‘cyclin-dependent protein kinase activity’ is part of the BP ‘cell cycle’.

term: gluconeogenesis
identifier: GO:0006094
definition: The formation of glucose from noncarbohydrate precursors, such as pyruvate, amino acids and glycerol.

No pathological processes
No experimental conditions
No evolutionary relationships
No gene products

GO datasets

Types of annotation:

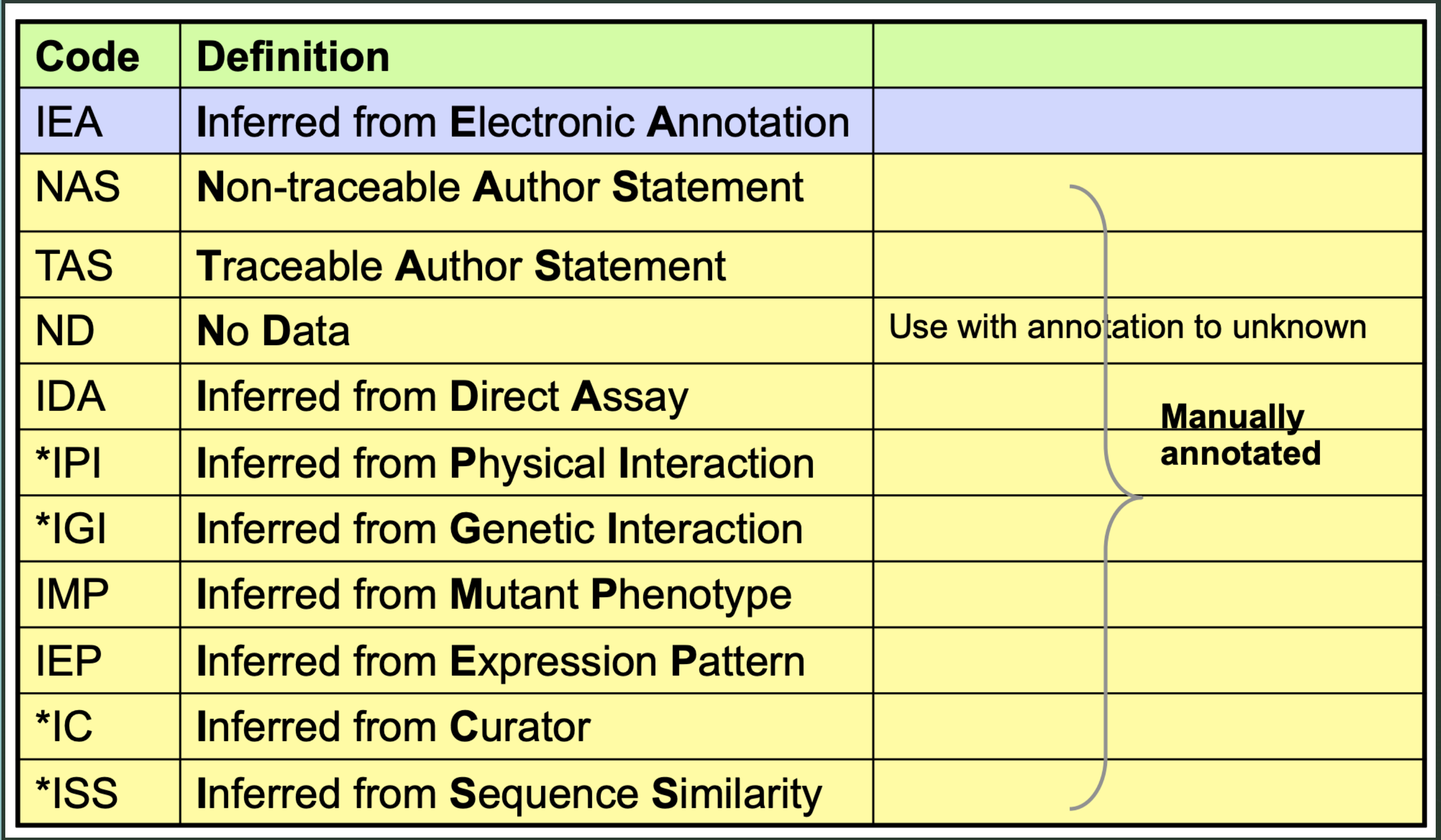
- **manual** annotation (manual curation)
    - High–quality, specific gene/gene productassociations derived from:
        - Peer-reviewed papers [evidence codes to grade evidence]
        - BUT – is very time consuming and requires trained biologists
        - Curators performs manual sequence similarity analyses to transfer annotations between highly similar gene products (BLAST, protein domain analysis)
- **automatic** annotation
    - Provides large-coverage
    - BUT – annotations tend to use high-level GO terms and provide little detail.

All annotations must:

[a] be attributed to a source

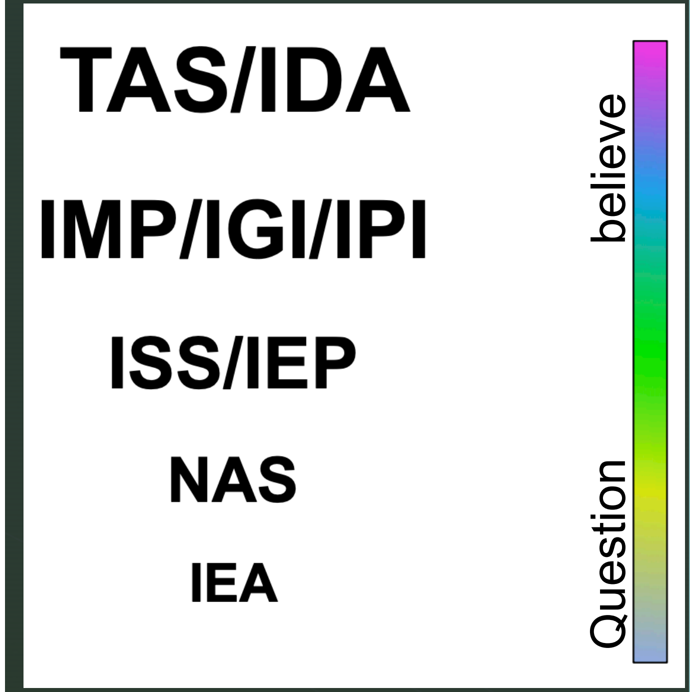
[b] indicate what evidence was found to support the GO term-gene/protein association

GO datasets

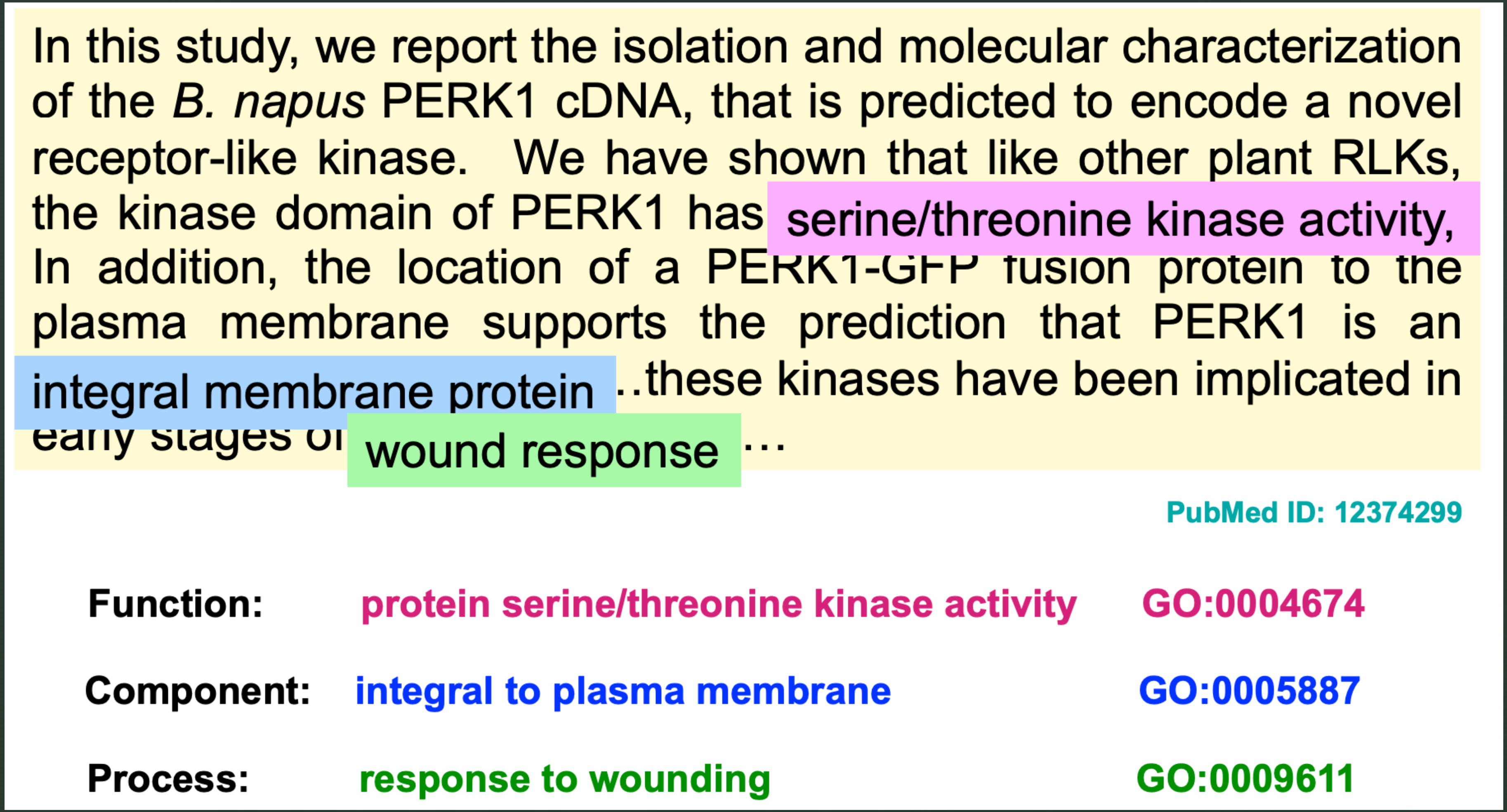


Code	Definition	
IEA	Inferred from Electronic Annotation	
NAS	Non-traceable Author Statement	
TAS	Traceable Author Statement	
ND	No Data	Use with annotation to unknown
IDA	Inferred from Direct Assay	
*IPI	Inferred from Physical Interaction	
*IGI	Inferred from Genetic Interaction	
IMP	Inferred from Mutant Phenotype	
IEP	Inferred from Expression Pattern	
*IC	Inferred from Curator	
*ISS	Inferred from Sequence Similarity	

Manually annotated



GO datasets

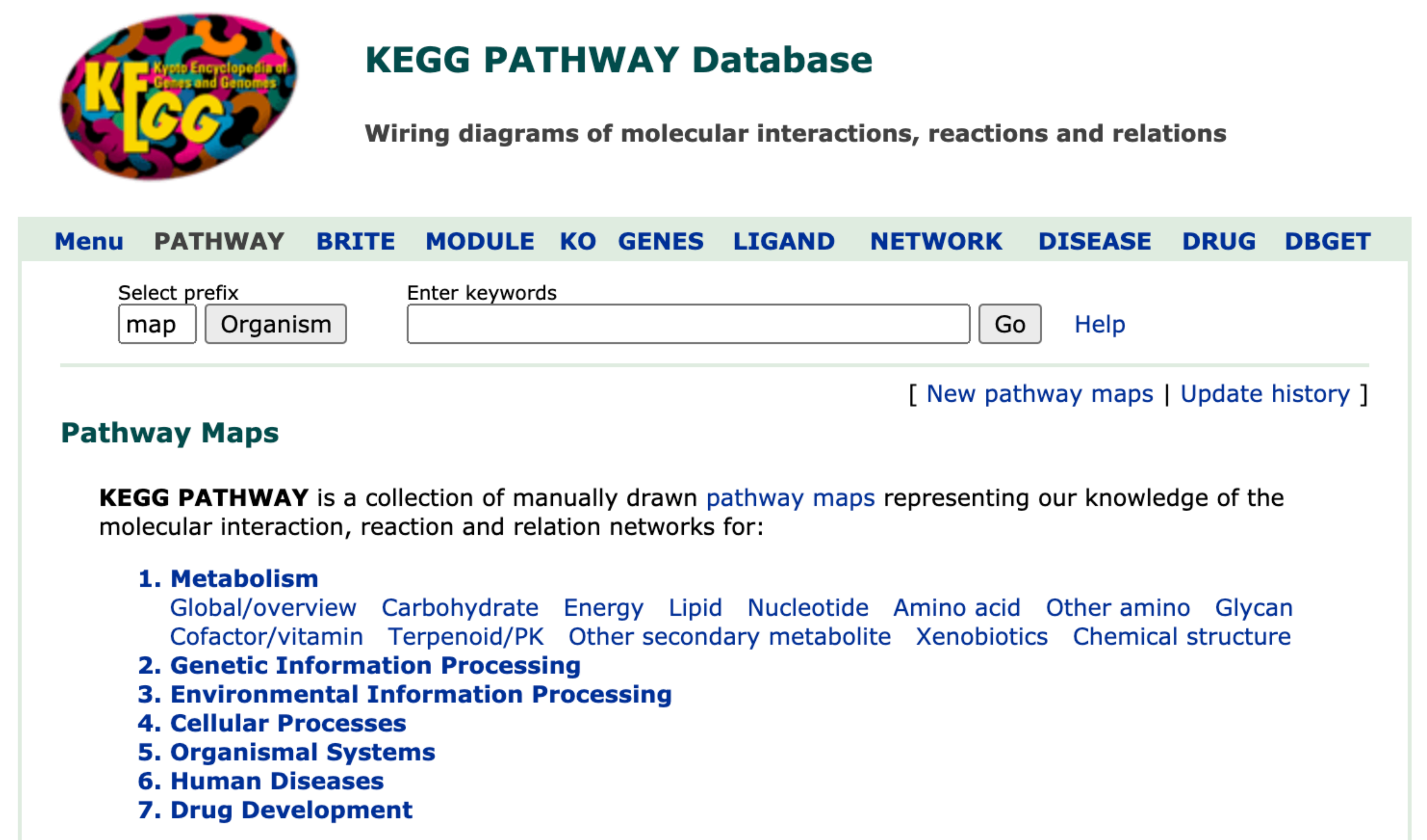


In this study, we report the isolation and molecular characterization of the *B. napus* PERK1 cDNA, that is predicted to encode a novel receptor-like kinase. We have shown that like other plant RLKs, the kinase domain of PERK1 has serine/threonine kinase activity, In addition, the location of a PERK1-GFP fusion protein to the plasma membrane supports the prediction that PERK1 is an integral membrane protein …these kinases have been implicated in early stages of wound response …

PubMed ID: 12374299

Function:	protein serine/threonine kinase activity	GO:0004674
Component:	integral to plasma membrane	GO:0005887
Process:	response to wounding	GO:0009611

Pathway datasets. KEGG
Kyoto Encyclopedia of Genes and Genomes



Kyoto Encyclopedia of Genes and Genomes
KEGG
Menu PATHWAY BRITE MODULE KO GENES LIGAND NETWORK DISEASE DRUG DBGET
Select prefix
map
Organism
Enter keywords
Go
Help
[New pathway maps | Update history]

KEGG PATHWAY Database

Wiring diagrams of molecular interactions, reactions and relations

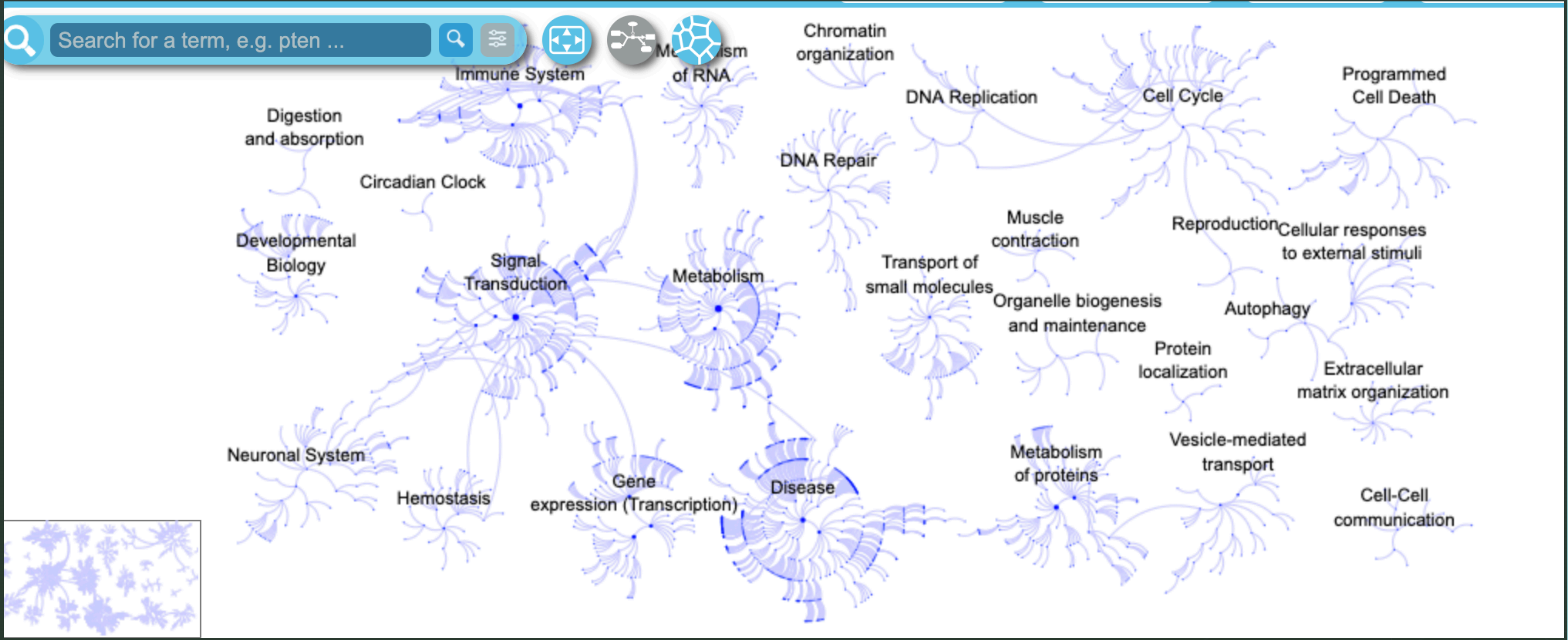
Pathway Maps

KEGG PATHWAY is a collection of manually drawn pathway maps representing our knowledge of the molecular interaction, reaction and relation networks for:

1. **Metabolism**
   Global/overview Carbohydrate Energy Lipid Nucleotide Amino acid Other amino Glycan
   Cofactor/vitamin Terpenoid/PK Other secondary metabolite Xenobiotics Chemical structure
2. **Genetic Information Processing**
3. **Environmental Information Processing**
4. **Cellular Processes**
5. **Organismal Systems**
6. **Human Diseases**
7. **Drug Development**

https://www.genome.jp/kegg/pathway.html

Pathway datasets. REACTOME



https://reactome.org/PathwayBrowser/

Enrichment analyses

Is expression of genes in a gene set associated with experimental condition?
E.g., Are there unusually many up-regulated genes in the gene set?

Many methods, a review is Kharti et al., 2012.

- [a] Over-representation analysis (ORA) – are differentially expressed (DE) genes in the set more common than expected?
- [b] Functional class scoring (FCS) – summarize statistic of DE of genes in a set, and compare to null
- [c] Issues with sequence data?
- [d] Issues with single-cell data?

Ten Years of Pathway Analysis: Current Approaches and Outstanding Challenges
Purvesh Khatri ,Marina Sirota,Atul J. Butte
https://journals.plos.org/ploscompbiol/article?id=10.1371/journal.pcbi.1002375

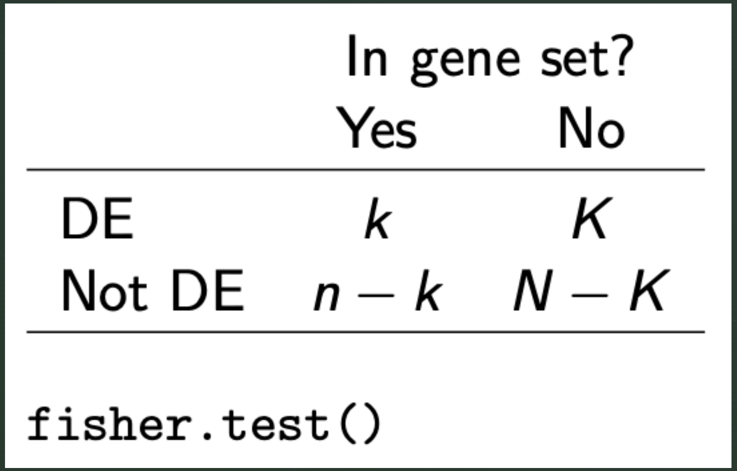
Enrichment analyses
Hypergeometric tests

[1] Classify each gene as “differentially expressed” DE or not,
e.g., based on adj $p < 0.05$ or |log2(FC) >= 0.5

[2] Are DE genes in the set more common than DE genes not in the set?

[3] Fisher hypergeometric test. GOstats; limma::goana()
Conditional hypergeometric to accommodate GO DAG, GOstats

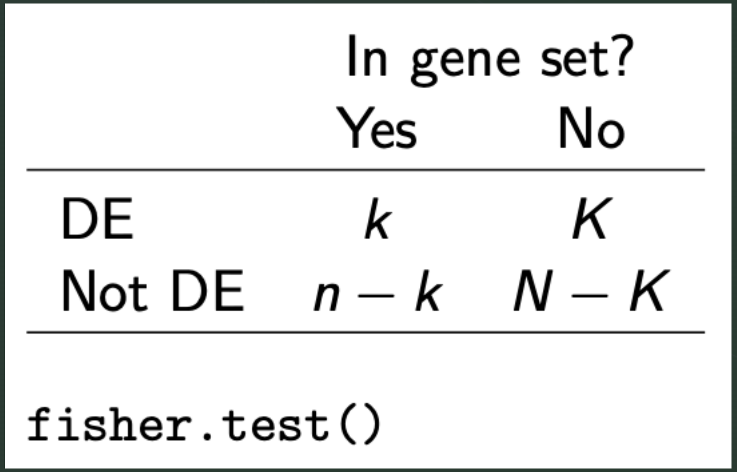
Con: artificial division into two groups (DE vs. not DE)
The number and identity of genes will depend on arbitrary thresholds e.g. p value thr, FC thr



	In gene set?	
	Yes	No
DE	k	K
Not DE	$n-k$	$N-K$

```
fisher.test()
```

Enrichment analyses
Hypergeometric tests



	In gene set?	
	Yes	No
DE	k	K
Not DE	$n-k$	$N-K$

```
fisher.test()
```

org: pval = 1

expressed: pval = 1.333e-05

What is your (gene) universe?
define the not DE set

```
All genes in the organism

              Yes          No
DE            10           90            ☹
Not DE        100          900

All genes expressed in the sample
              Yes          No
DE            10           90            ☺
Not DE        10           790
```

Enrichment analyses
Enrichment score

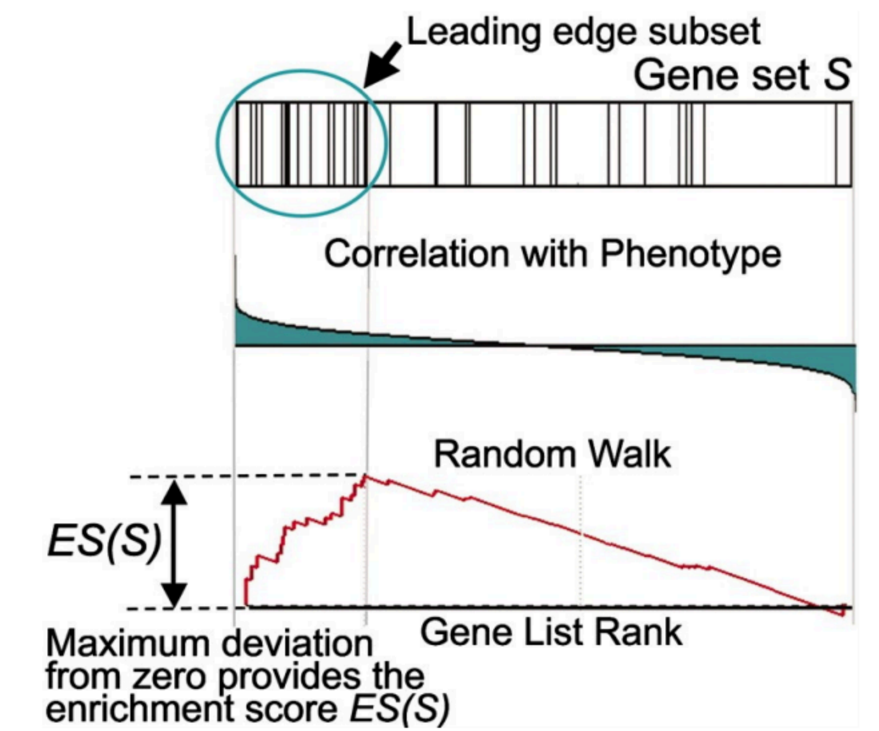
Mootha et al., 2003; modified
Subramanian et al., 2005.

[1] Sort genes by log fold change

[2] Calculate running sum: incremented when gene in set, decremented when not.

[3] Maximum of the running sum is enrichment score ES; large ES means that genes in set are toward top of list.

[4] Permuting subject labels for significance


Leading edge subset
Gene set *S*
Correlation with Phenotype
Random Walk
ES(S)
Gene List Rank
Maximum deviation from zero provides the enrichment score *ES(S)*

Subramanian et al., 2005, fig 1.

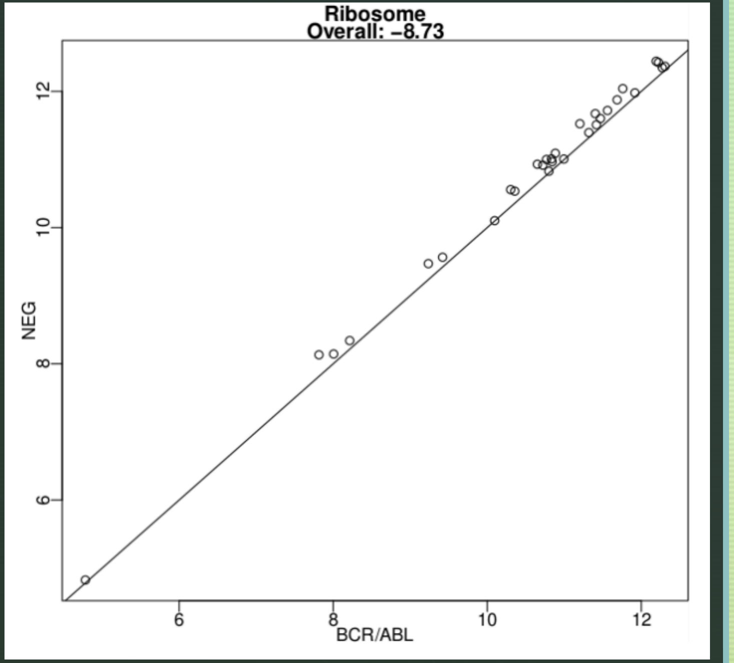
Gene set enrichment analysis: A knowledge-based approach for interpreting genome-wide expression profiles
Aravind Subramanian et al 2005 https://www.pnas.org/content/102/43/15545
PGC-1alpha-responsive genes involved in oxidative phosphorylation are coordinately downregulated in human diabetes
Vamsi K Mootha https://www.nature.com/articles/ng1180

Enrichment analyses
statistical test + permutation of labels

(developed on microarrays)

[1] Summarize t (or other) statistic across genes in each set
[2] Test for significance by permuting the subject labels

pro: Much more straight-forward to implement



The mean plot for the Ribosome pathway. Each point represents a gene in the pathway and the x-value is determined by the mean expression in the BCR/ABL group while the y-value is determined by the mean in the NEG group.

http://bioconductor.org/packages/release/bioc/vignettes/Category/inst/doc/Category.pdf

Enrichment analyses
Competitive vs self contained null hypothesis

[a] **Competitive null**: The genes in the gene set do not have stronger association with the subject condition than other genes. (Approach 1, 2)

[b] **Self-contained null**: The genes in the gene set do not have any association with the subject condition. Assessing individual sets. (Approach 3)

Remarks:
The self-contained null is closer to actual question of interest
Permuting subjects (rather than genes) is appropriate

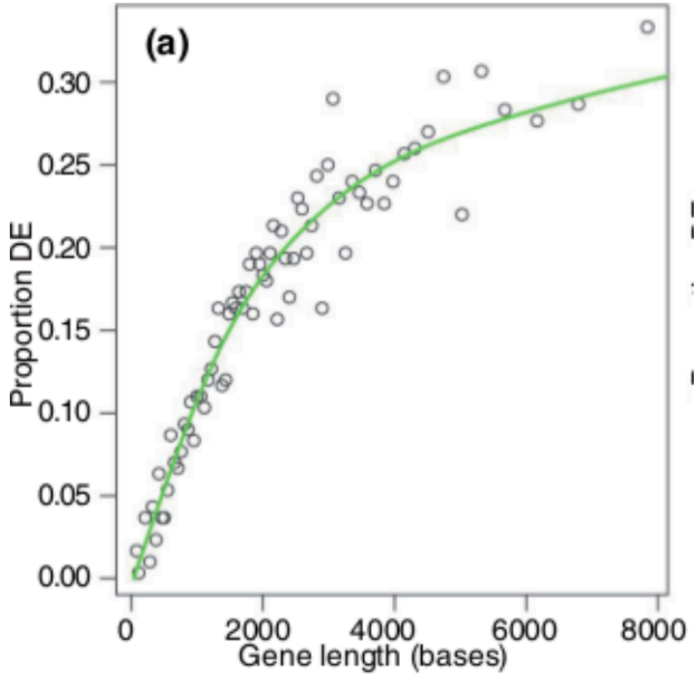
Goeman & Buhlmann, 2007, Bioinformatics 23.8: 980-987.

Enrichment analyses
Gene length normalisation

All else being equal, long genes have higher abundances than short genes

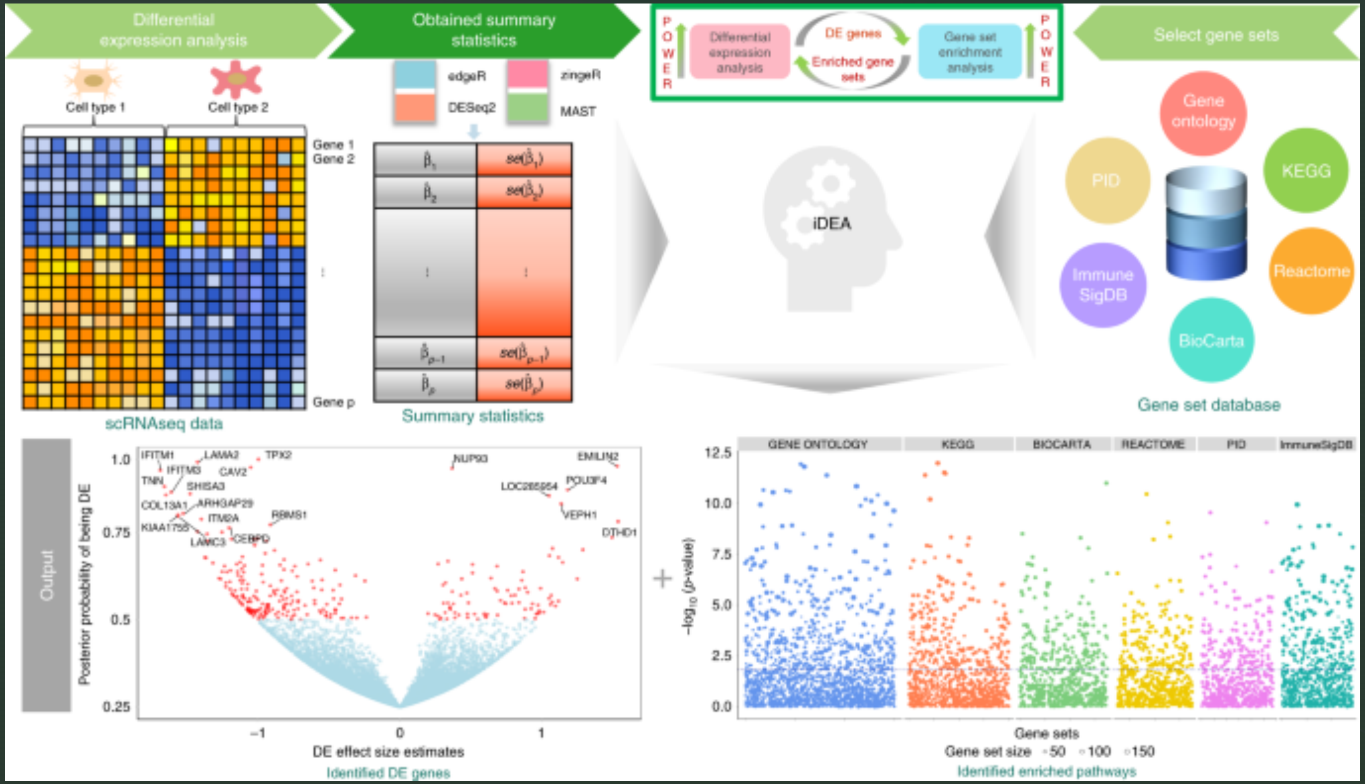
Per-gene p values proportional to gene size

Revise the comments on the gene length normalisation [and impact on DE, and enrichment analyses]



DE genes vs. transcript length. Points: bins of 300 genes. Line: fitted probability weighting function.

Enrichment analyses on single cell data

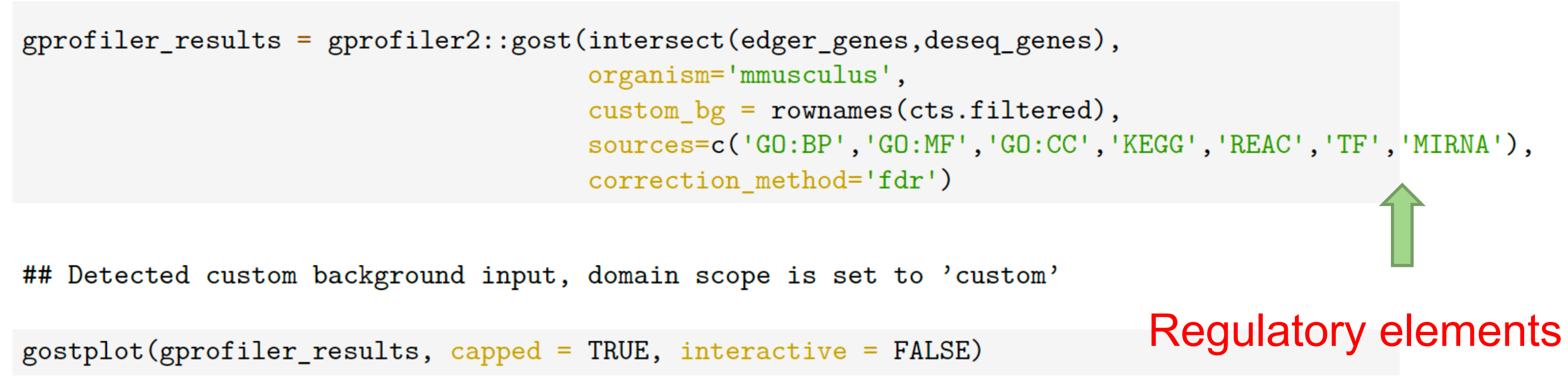


iDEA is designed to jointly model all genes together for integrative differential expression (DE) analysis and gene set enrichment (GSE) analysis. iDEA requires input association summary statistics from existing scRNA-seq DE methods in terms of the DE effect size estimate and its standard error (top left panels). iDEA also requires a pre-defined set of gene sets that we have compiled and pruned for use with the software (top right panels). With these two inputs, iDEA performs joint DE and GSE analysis through a Bayesian hierarchical model. For each gene set, iDEA outputs a p-value for testing whether the gene set is enriched with DE genes (bottom right panel) for GSE analysis.

Integrative differential expression and gene set enrichment analysis using summary statistics for scRNA-seq studies, Ma et al 2020 https://www.nature.com/articles/s41467-020-15298-6

Enrichment analyses
g:profiler

Enrichment analysis on the Yang et al 2019 data (mRNAseq practical)

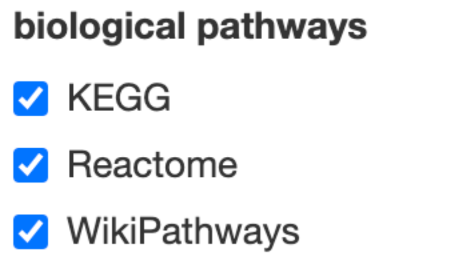


```
gprofiler_results = gprofiler2::gost(intersect(edger_genes,deseq_genes),
                                     organism='mmusculus',
                                     custom_bg = rownames(cts.filtered),
                                     sources=c('GO:BP','GO:MF','GO:CC','KEGG','REAC','TF','MIRNA'),
                                     correction_method='fdr')
```

```
## Detected custom background input, domain scope is set to ’custom’
```

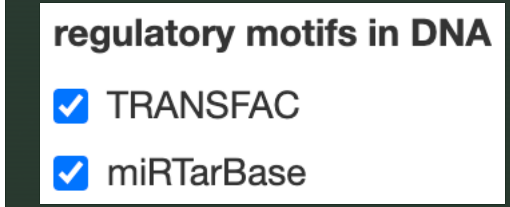
```
gostplot(gprofiler_results, capped = TRUE, interactive = FALSE)
```

Regulatory elements



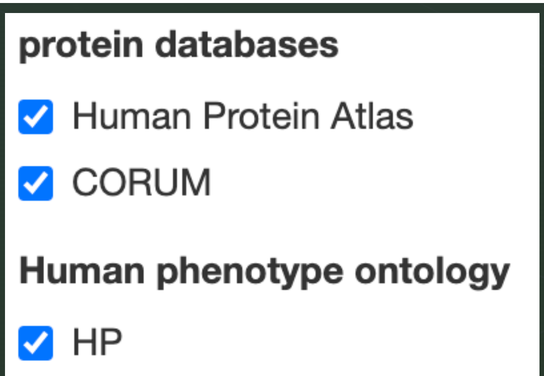
biological pathways

- ☑ KEGG
- ☑ Reactome
- ☑ WikiPathways



regulatory motifs in DNA

- ☑ TRANSFAC
- ☑ miRTarBase



protein databases

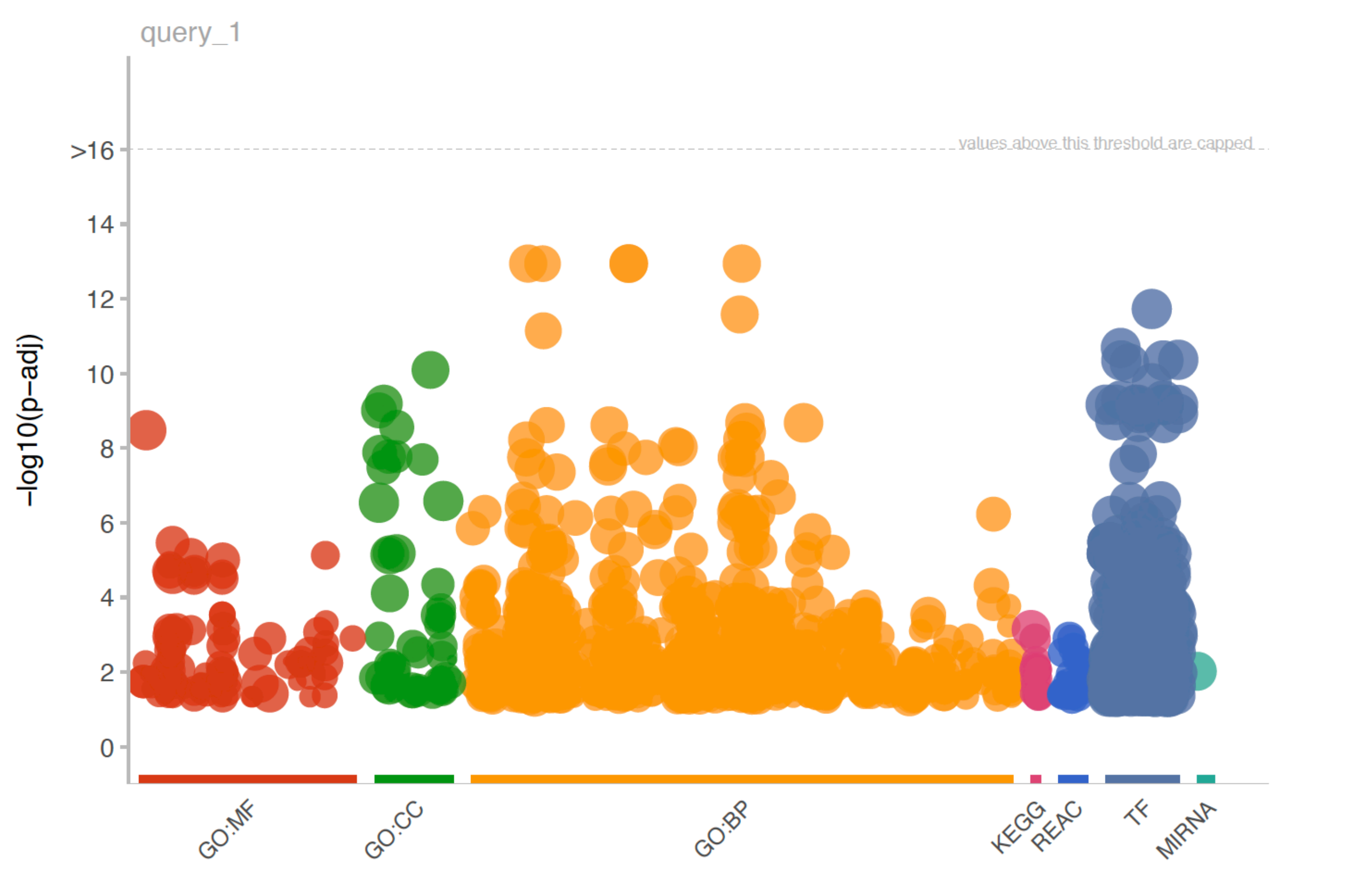
- ☑ Human Protein Atlas
- ☑ CORUM

Human phenotype ontology

- ☑ HP

https://biit.cs.ut.ee/gprofiler/gost

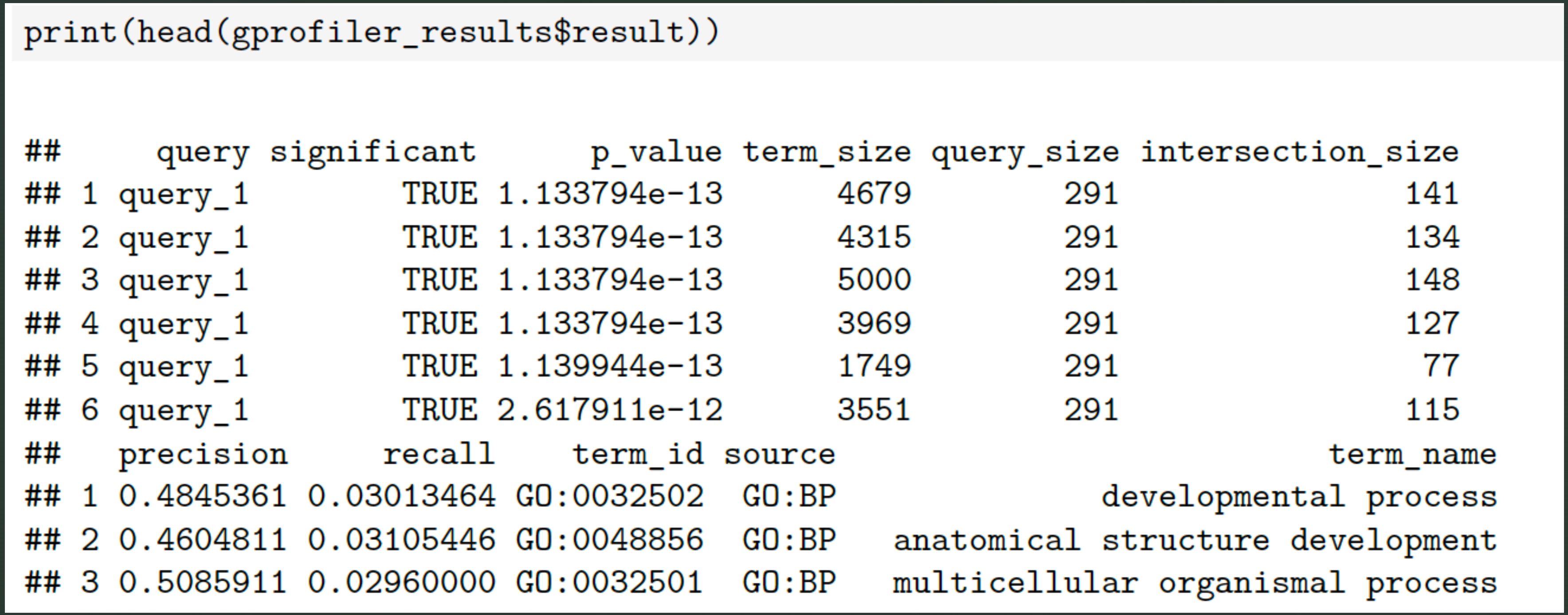
Enrichment analyses
g:profiler



Each entry is a term.

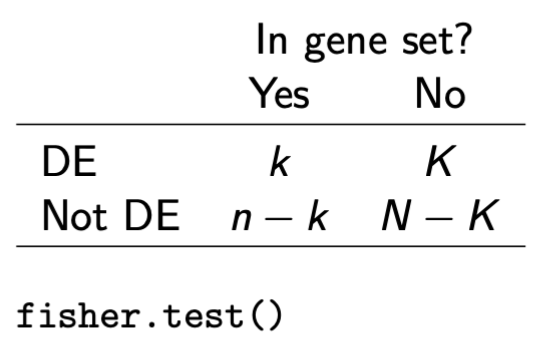
From a regulation perspective we see an enrichment of TFs (regulation at transcriptional level); the other samples present in the Yang et al 2019 data look at methylation levels.

Enrichment analyses



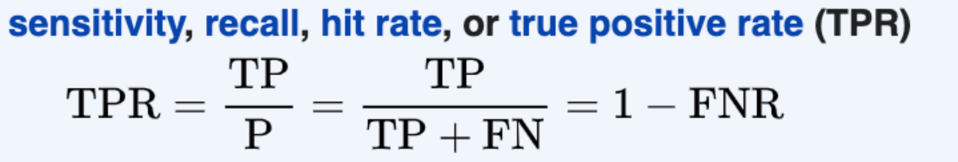
```
print(head(gprofiler_results$result))
```

```
##     query significant      p_value term_size query_size intersection_size
## 1 query_1        TRUE 1.133794e-13      4679        291               141
## 2 query_1        TRUE 1.133794e-13      4315        291               134
## 3 query_1        TRUE 1.133794e-13      5000        291               148
## 4 query_1        TRUE 1.133794e-13      3969        291               127
## 5 query_1        TRUE 1.139944e-13      1749        291                77
## 6 query_1        TRUE 2.617911e-12      3551        291               115
##   precision     recall    term_id source                        term_name
## 1 0.4845361 0.03013464 GO:0032502  GO:BP            developmental process
## 2 0.4604811 0.03105446 GO:0048856  GO:BP anatomical structure development
## 3 0.5085911 0.02960000 GO:0032501  GO:BP multicellular organismal process
```



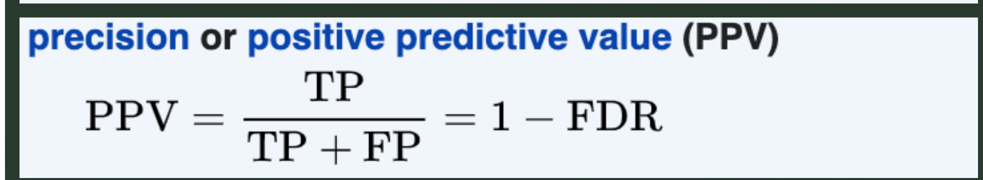
	In gene set?	
	Yes	No
DE	k	K
Not DE	$n-k$	$N-K$

`fisher.test()`



sensitivity, **recall**, **hit rate**, or **true positive rate** (TPR)

$$\mathrm{TPR} = \frac{\mathrm{TP}}{\mathrm{P}} = \frac{\mathrm{TP}}{\mathrm{TP}+\mathrm{FN}} = 1 - \mathrm{FNR}$$



precision or **positive predictive value** (PPV)

$$\mathrm{PPV} = \frac{\mathrm{TP}}{\mathrm{TP}+\mathrm{FP}} = 1 - \mathrm{FDR}$$