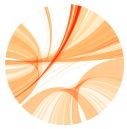
June 2019

BIO634: making sense of gene lists

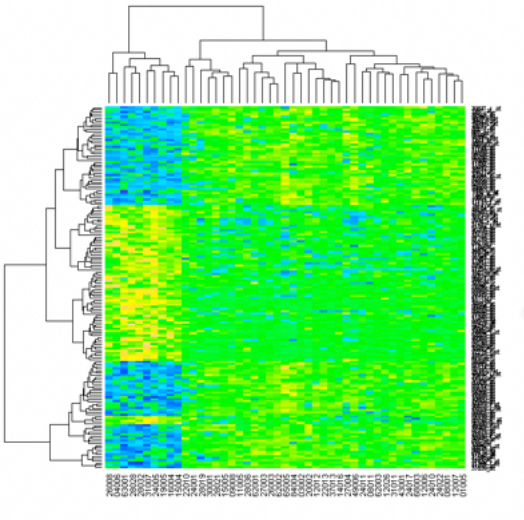
Adapted from **Stefan Wyder**
class on BIO634 (2018)



Universität
Zürich[UZH]

Gene list annotation

you performed a genomic experiment and obtained a gene list
hundreds of genes is too much, you would like to pinpoint interesting gene candidates



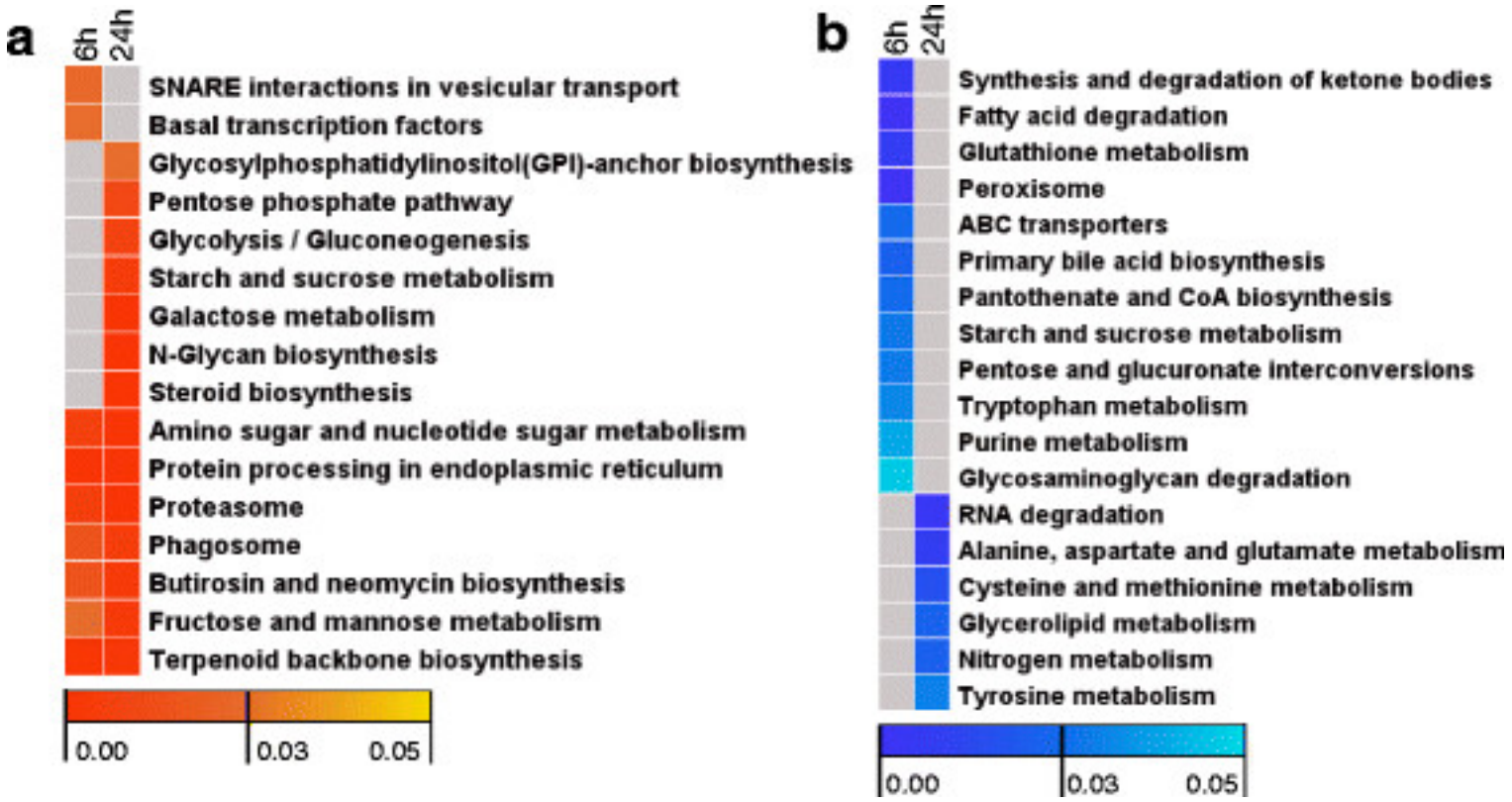
experimental data
RNA-seq, ChIP-seq, CLIP

geneA
geneB
geneC
geneD
…
geneX

annotation via previous knowledge
gene ontology, pathway analysis

biological insight

Gene list annotation



PMID:26584608

Biological insight

interpretation of an experiment

find regulated processes / pathways

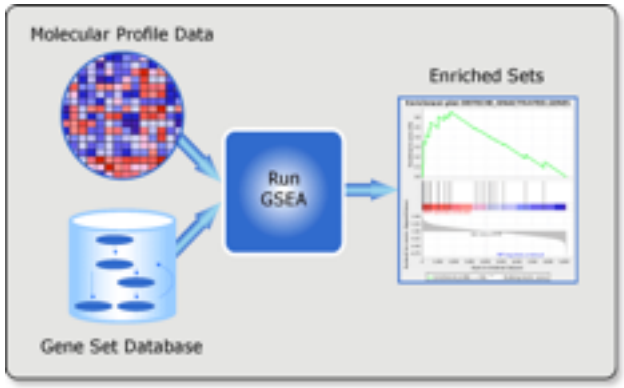
find involved regulatory elements, TF, RNA-binding proteins

identify new members of a pathway

find similar experiments

Analysis based on gene lists expected to be more robust and reproducible compared to single gene analysis.

Enrichment analysis



Over-representation analysis

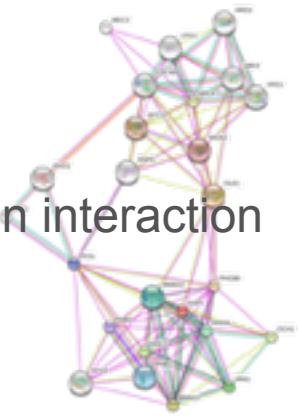
- hypergeometric / Fisher's exact test
- setting a cutoff a priori
- different results at different thresholds

GSEA, gene set enrichment analysis

- bypasses the need for a cutoff
- input: list of all measured genes ranked by some measure / effect size
- weak but consistent regulation of several members of a gene set can be detected

Network analysis

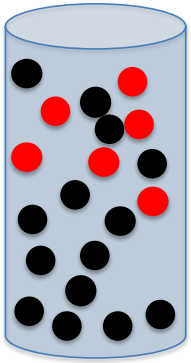
- also covers less understood parts of gene interactions
- often inferred from co-expression data
- string-db.org, combines co-expression, co-citation, protein-protein interaction



Over-representation analysis

5.000 black and 10 red “genes”

10 red “genes” are cytochromes



our list of differentially expressed genes

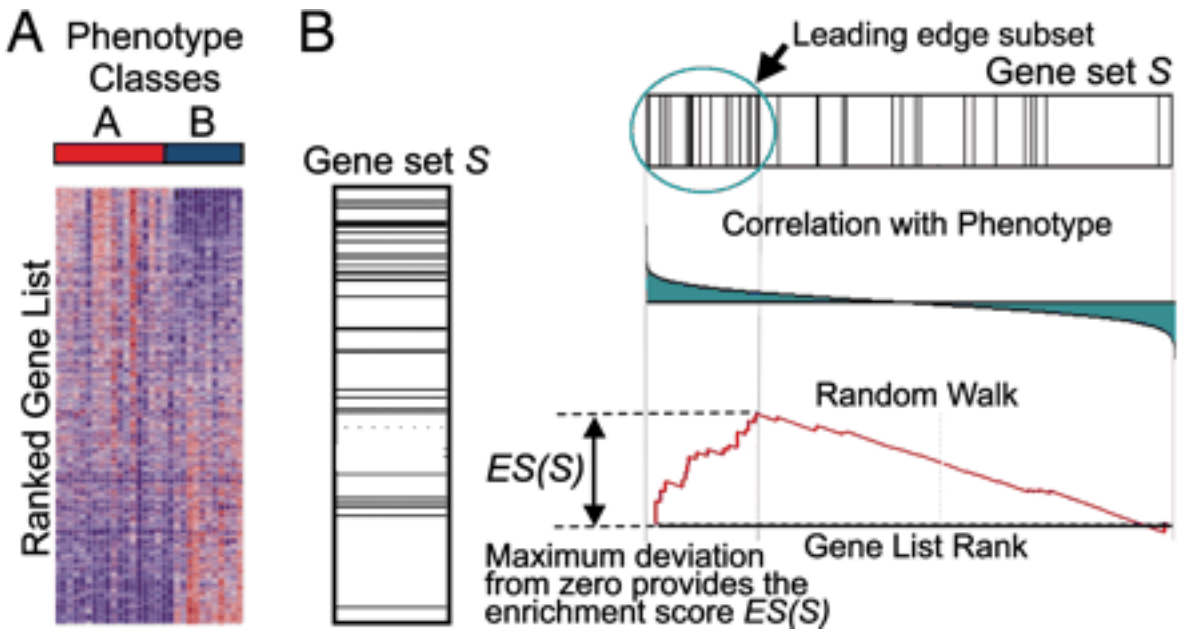
- CYP4F11
- CYP1A
- MEP1A
- CYP26B
- CYP3A43

what is the probability?

	selected	not-selected
red	4	6
black	1	4989

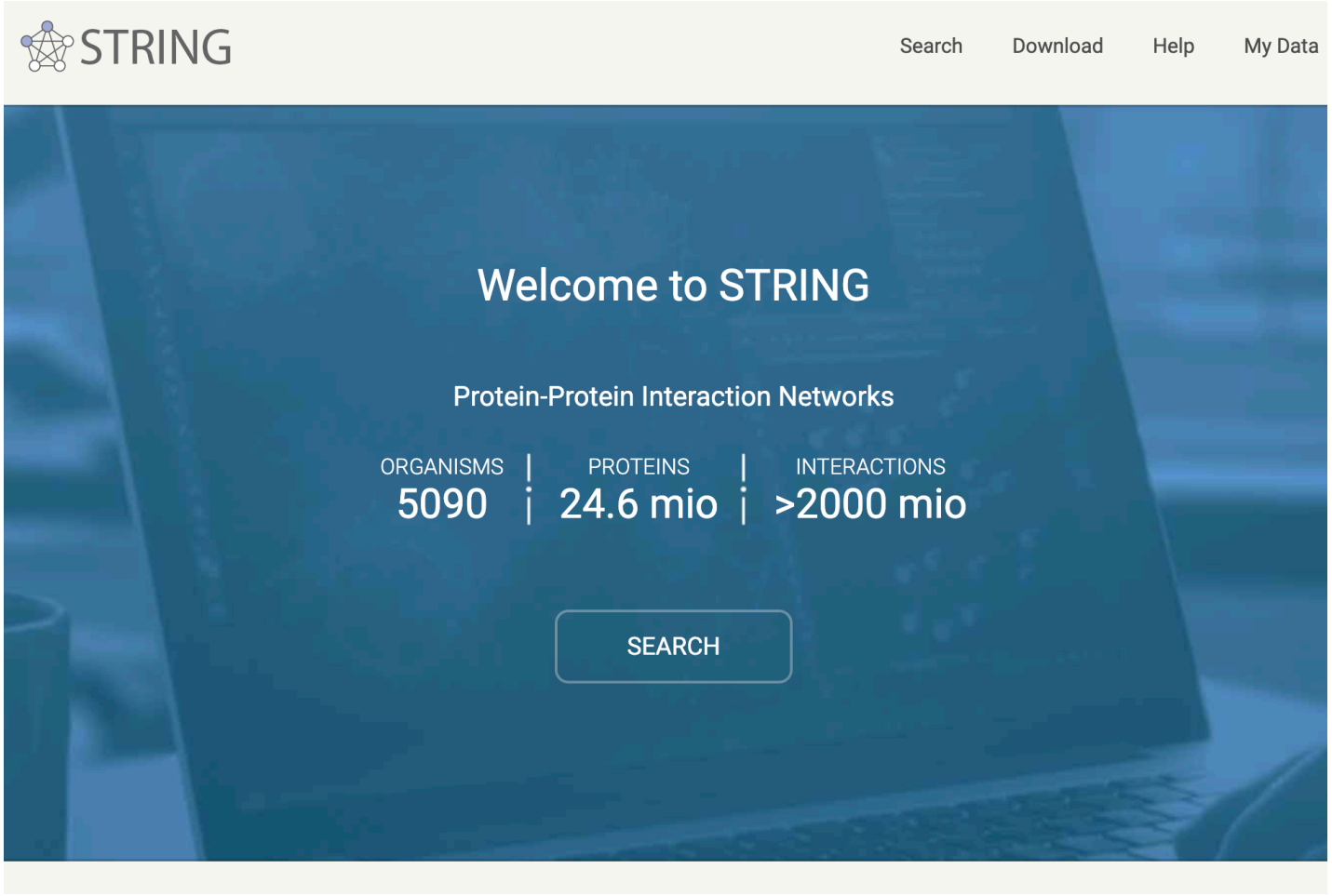
one sided Fisher’s exact test
p-value: 4.03e-11

Gene Set Enrichment Analysis

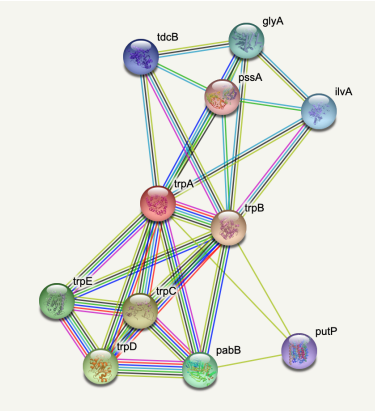


Subramanian et al. (2005) PNAS

string-db.org



string-db.org

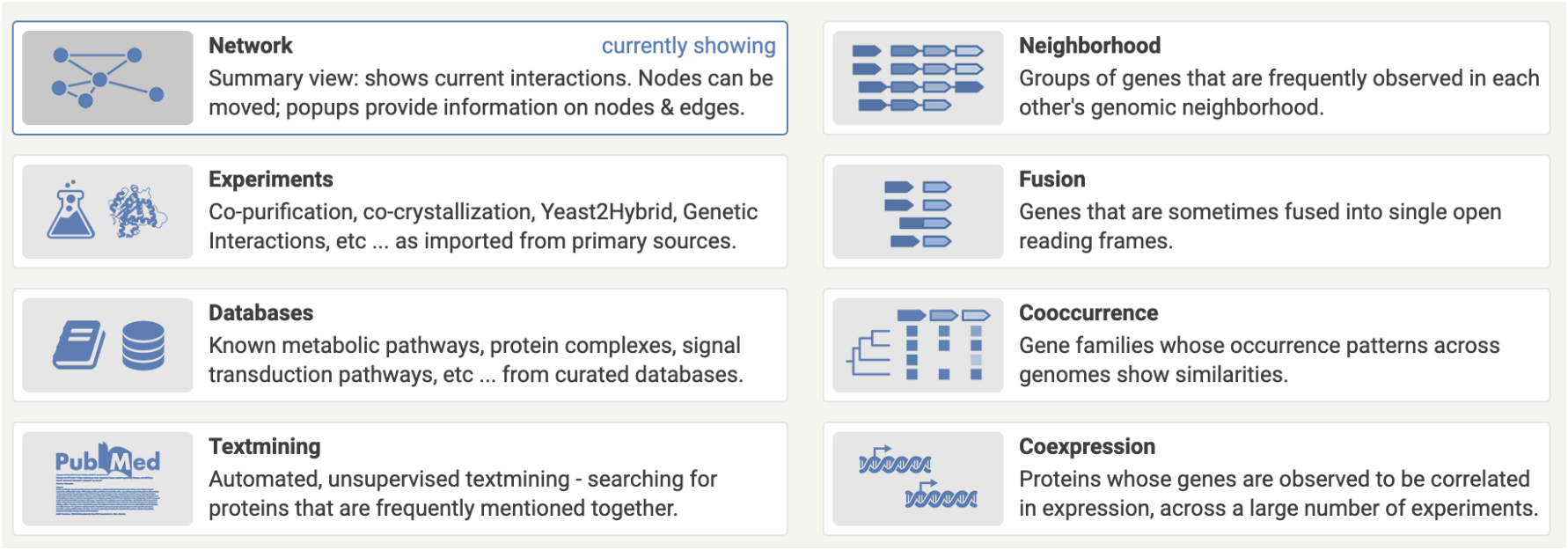


functional association networks (physical or functional interactions)

focus on useability and speed

integrated scoring scheme (each interaction has confidence score)

information transfer between species
(>5000 species: Animals, Bacteria, Plants)



string-db.org

More than gene-list annotation

- predict gene function
- identify candidates for an unknown enzyme in a pathway
- identify new member genes of a biological process
- find relevant literature

STRING performs well compared with single-species databases

R package to access STRING functionality from R

Other interaction sources

Interaction Databases

IntAct

BioGRID

Human Protein Reference Database

Pathway Databases

Reactome

KEGG

the Gene Ontology

PathwayInteractionDatabase

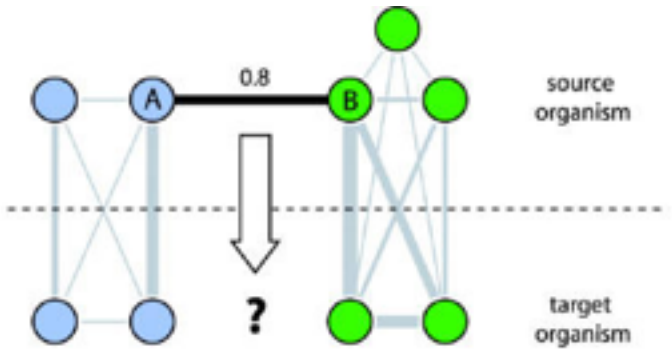
Automated Textmining

PubMed www.pubmed.gov

OMIM

SGD

Interolog Transfer

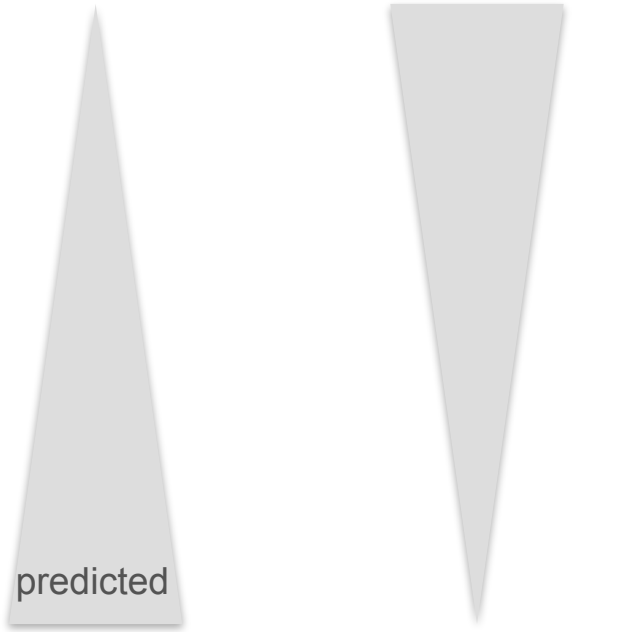


Annotation sources

Pathways
KEGG, Reactome, BioCyc

Gene Ontology

Networks
STRING



Pathways

pathway maps (aka reaction networks / wiring diagrams) represent experimental knowledge on metabolism and various other functions of the cell and the organism

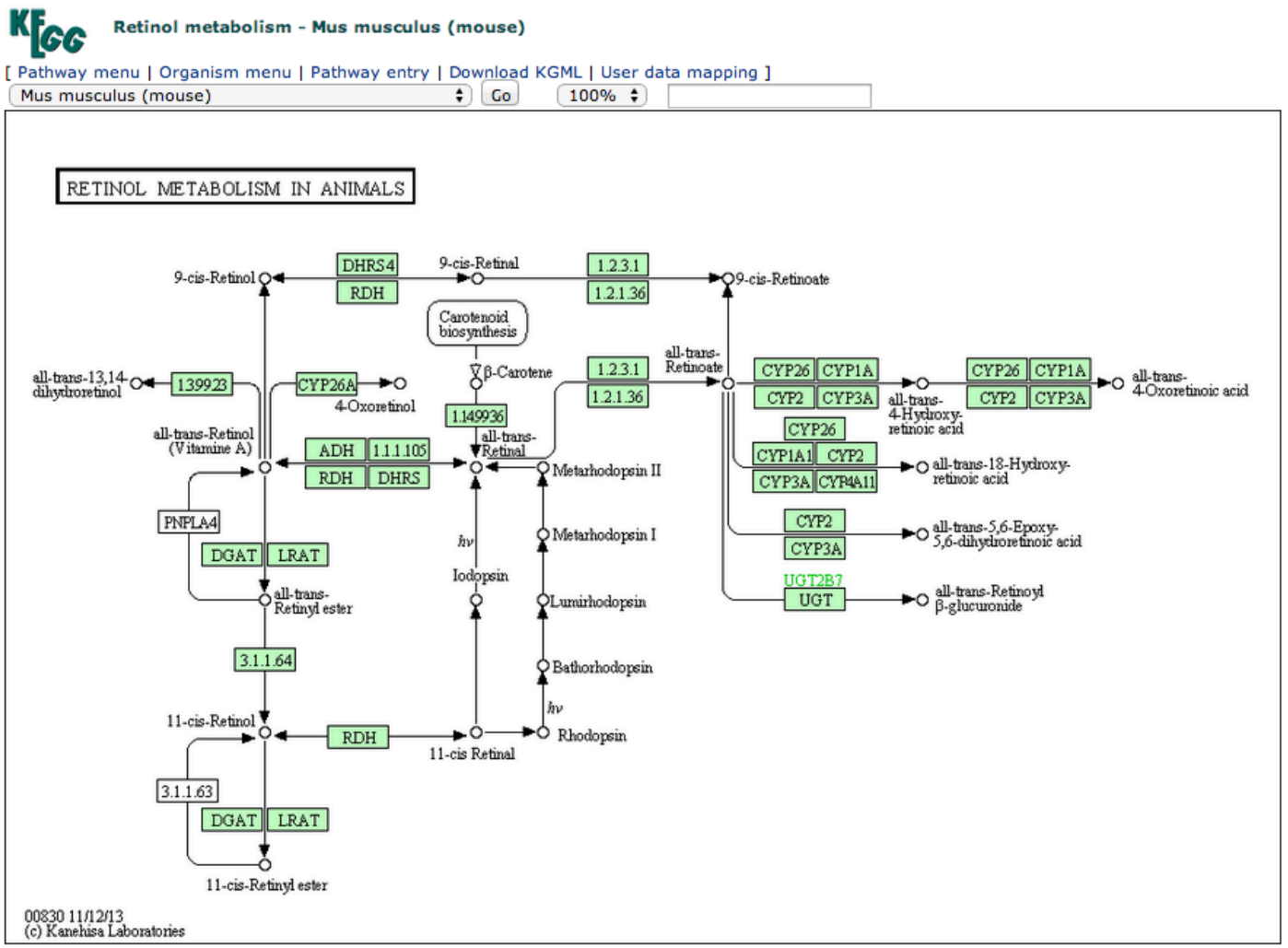
manually curated

the main databases are **KEGG** and **Reactome**

KEGG is free to use over the web but file download requires subscription

KEGG covers > 3.800 species (Archae, Bacteria, Plants, Animals) and Reactome covers 20 species (mostly mammals + fly + plants + E.coli)

KEGG: example



Gene Ontology

describes how gene products behave in a cellular context:
biological process, cellular component, molecular function

controlled vocabulary of terms

transparent (sources)

manually curated lists for model species

transfer to orthologs in other species (inferred annotation)

Gene Ontology example

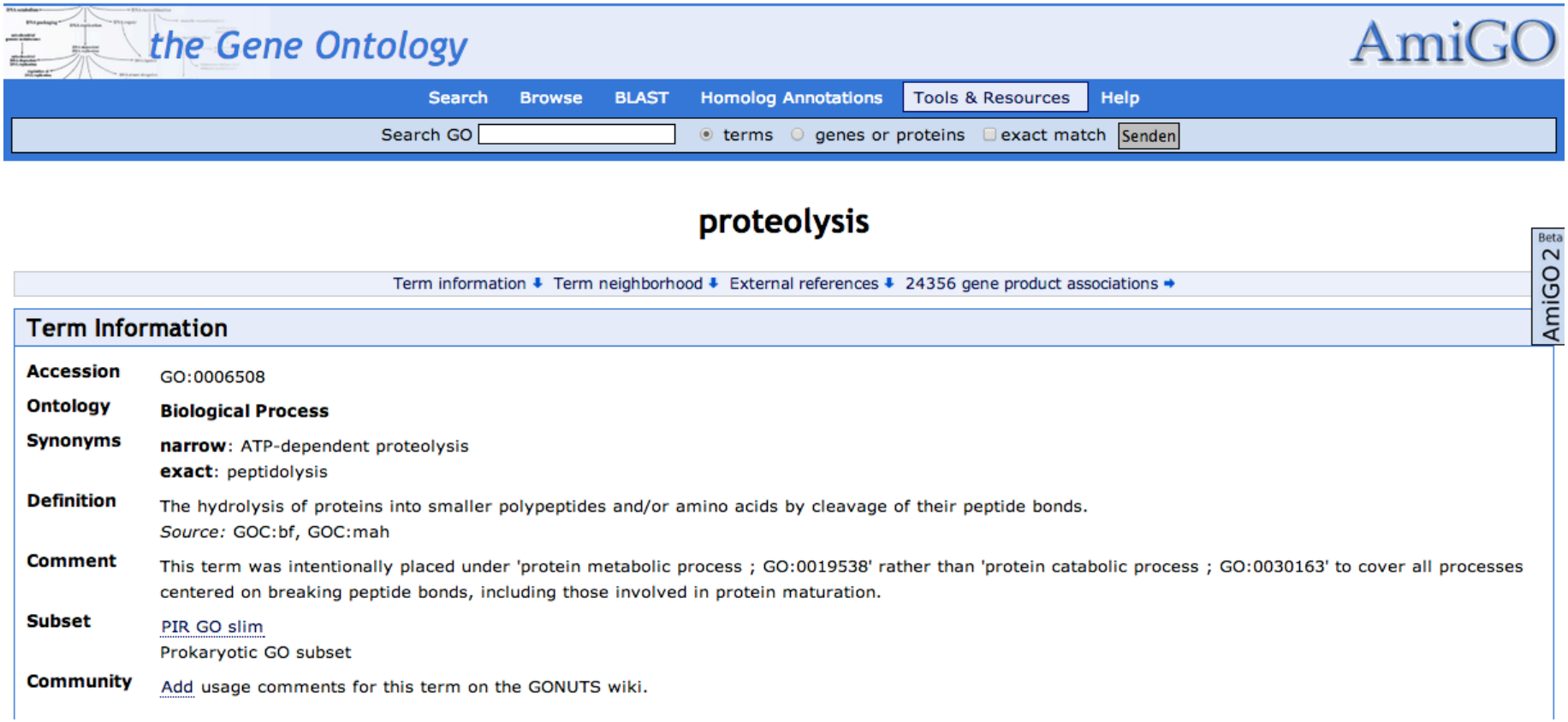
murine ADAM10

Molecular Function
GO:0008237 metallopeptidase activity
GO:0042169 SH2 domain binding
..

Biological Process
GO:0007220 Notch receptor processing
GO:0001701 in utero embryonic development
GO:0008284 positive regulation of cell proliferation
..

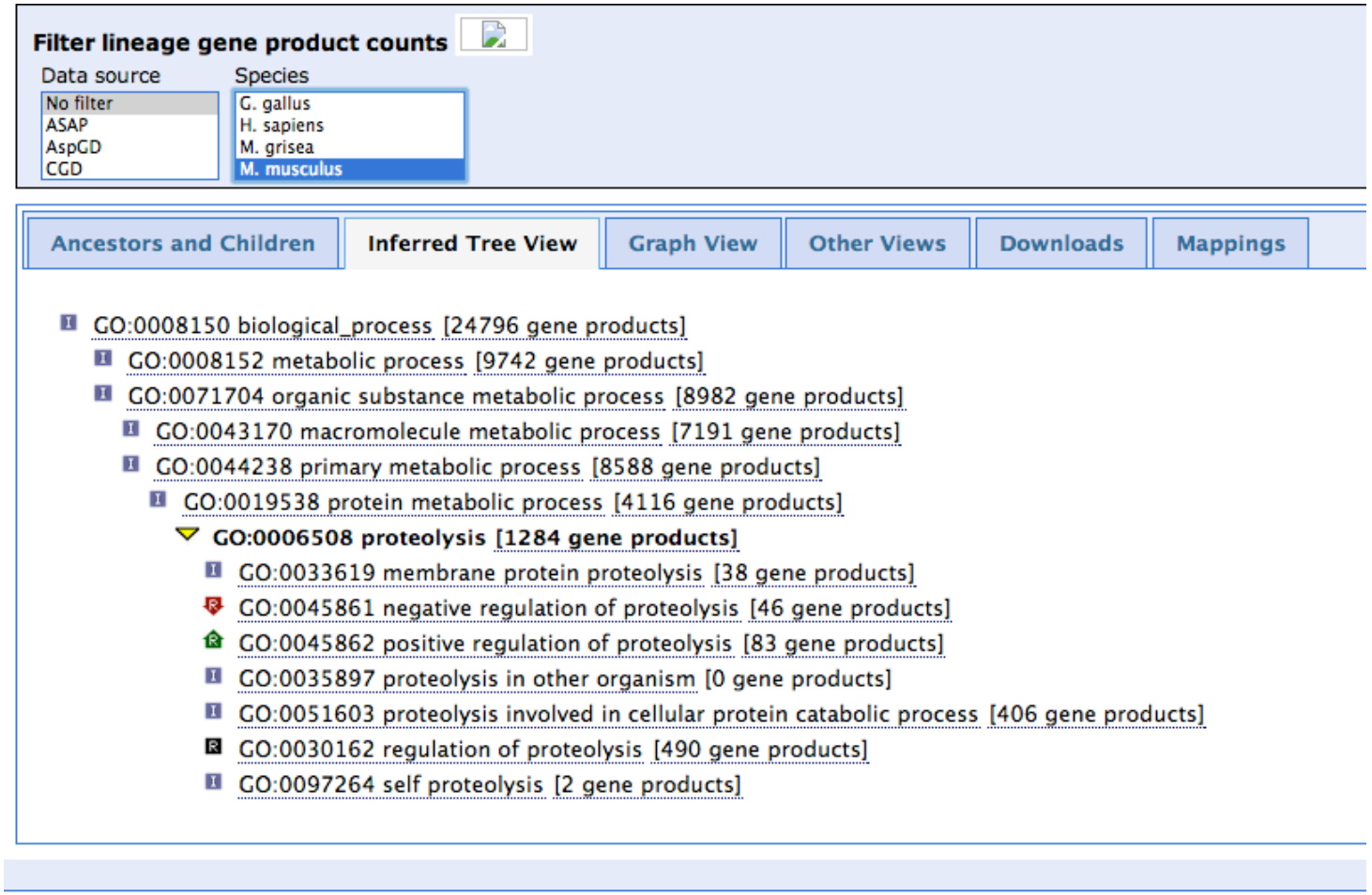
Cellular Compartment
GO:0005794 Golgi apparatus
GO:0009986 cell surface

AmiGO: lookup of GO terms



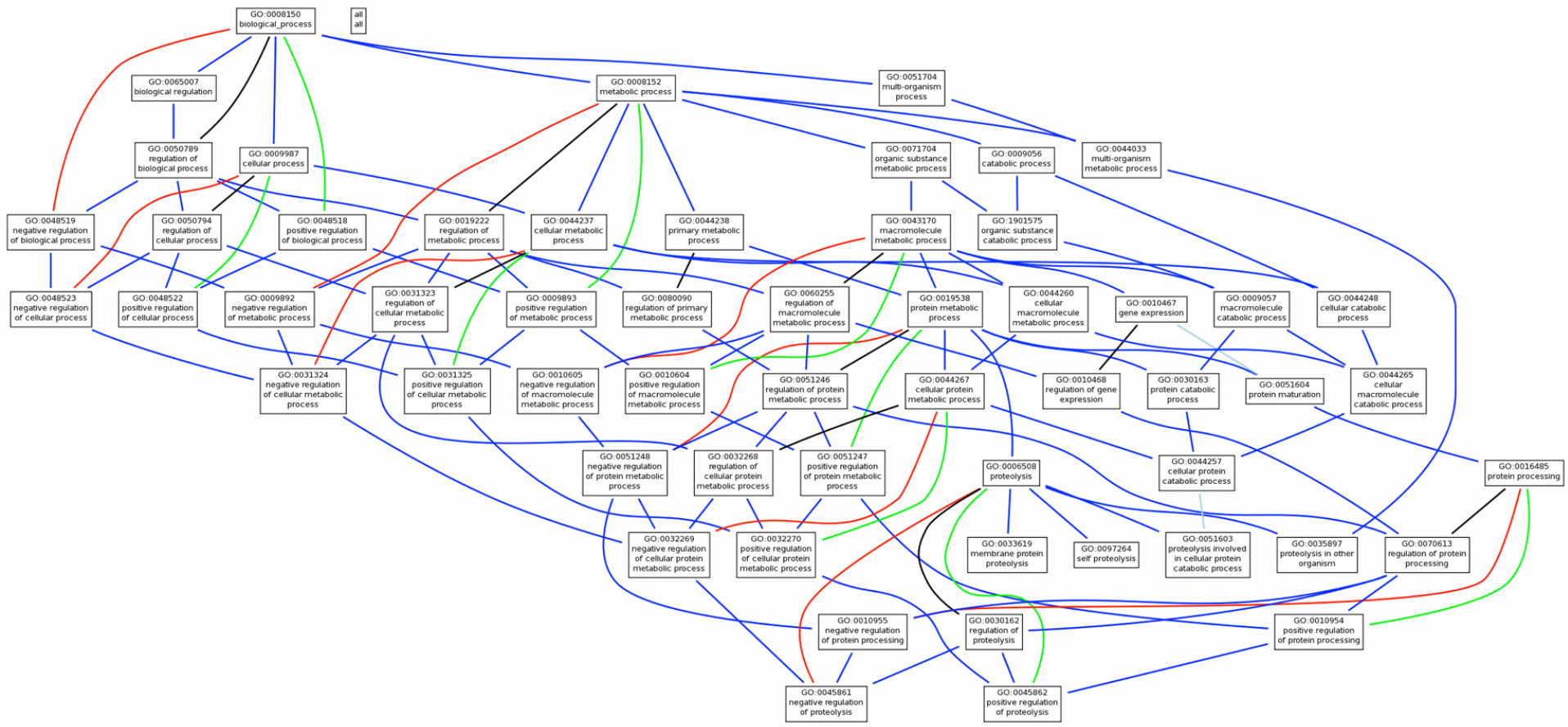
http://amigo.geneontology.org

GO table view



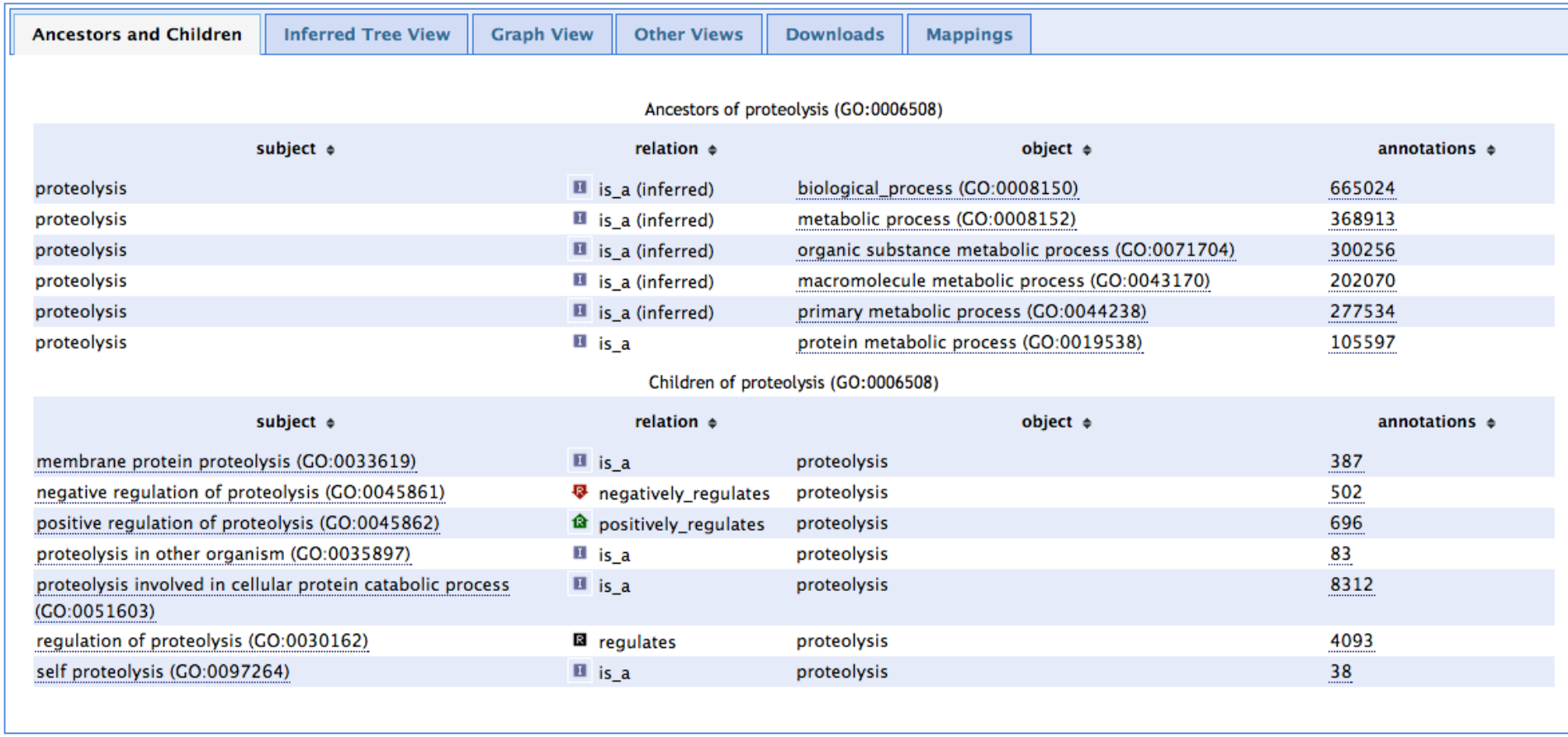
GO:0006508 Proteolysis

GO table view



GO:0006508 Proteolysis

GO ancestors and children



Ancestors and Children | Inferred Tree View | Graph View | Other Views | Downloads | Mappings

Ancestors of proteolysis (GO:0006508)

subject	relation	object	annotations
proteolysis	is_a (inferred)	biological_process (GO:0008150)	665024
proteolysis	is_a (inferred)	metabolic process (GO:0008152)	368913
proteolysis	is_a (inferred)	organic substance metabolic process (GO:0071704)	300256
proteolysis	is_a (inferred)	macromolecule metabolic process (GO:0043170)	202070
proteolysis	is_a (inferred)	primary metabolic process (GO:0044238)	277534
proteolysis	is_a	protein metabolic process (GO:0019538)	105597

Children of proteolysis (GO:0006508)

subject	relation	object	annotations
membrane protein proteolysis (GO:0033619)	is_a	proteolysis	387
negative regulation of proteolysis (GO:0045861)	negatively_regulates	proteolysis	502
positive regulation of proteolysis (GO:0045862)	positively_regulates	proteolysis	696
proteolysis in other organism (GO:0035897)	is_a	proteolysis	83
proteolysis involved in cellular protein catabolic process (GO:0051603)	is_a	proteolysis	8312
regulation of proteolysis (GO:0030162)	regulates	proteolysis	4093
self proteolysis (GO:0097264)	is_a	proteolysis	38

GO:0006508 Proteolysis

Experimental design

Experimental design is crucial for good chances of interpretability of results

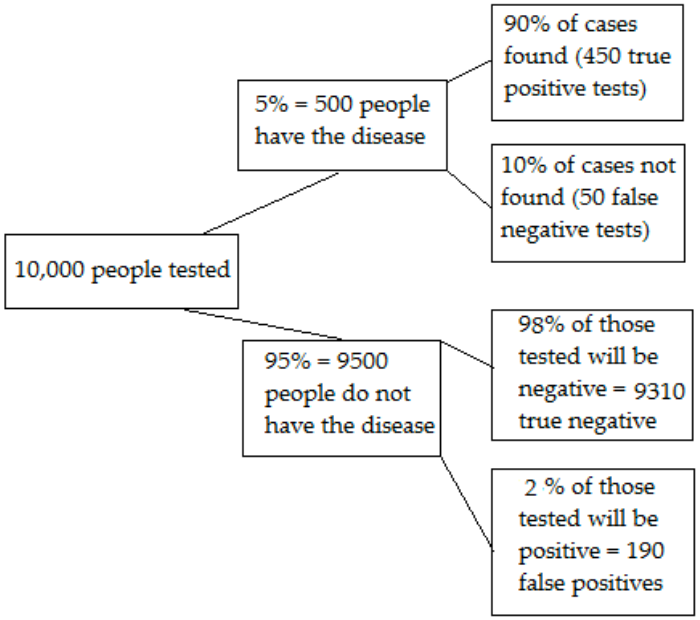
Randomize samples/treatments across lanes / flow cells

Multiple tissues/cell types/stages pooled in a sample -> complex and difficult to understand the ongoing processes

- e.g. observed changes can simply be due to changes in relative abundance of different cell types independant of regulation

FDR: false discovery rate

"if you repeat a test enough times, you're going to find an effect…but that effect may not actually exist"



The FDR approach adjusts the p-value for a series of tests. A p-value gives you the probability of a false positive on a single test; If you're running a large number of tests from small samples (which are common in fields like genomics and protoemics), you should use q-values instead.

- A p-value of 5% means that 5% of all tests will result in false positives.
- A q-value of 5% means that 5% of *significant* results will be false positives.

The procedure to control the FDR, using q-values, is called the Benjamini-Hochberg procedure, named after Benjamini and Hochberg (1995), who first described it.

Summary

Gene list annotation with Pathways and Gene Ontology can help to obtain biological insight

A Over-Representation Analysis
B Gene Set Enrichment Analysis, GSEA
C Network Analysis

Biological interpretation requires broad knowledge of physiology & biochemistry and is often the most difficult and time-consuming step of an experiment

Even experts can usually not make sense of all the significantly enriched processes/ pathways in well understood biological systems

Good experiments start with good experimental design, think of possible confounders