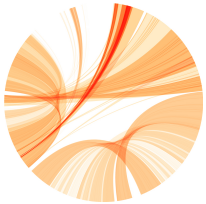
NGS2 course

Making Sense of Gene Lists

Stefan Wyder

May 2015

Universität Zürich[UZH]

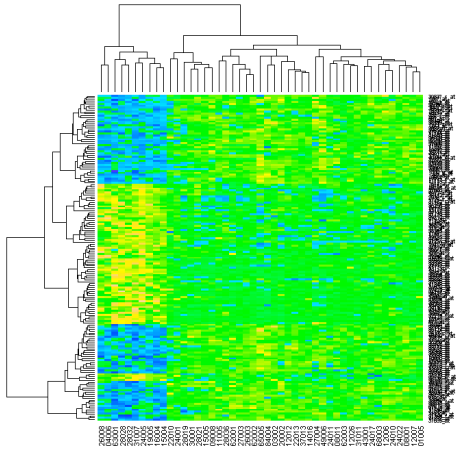


URPP Evolution in Action

Gene List Annotation

- You performed a genomic experiment and obtained a gene list
- Who wants to work through a list of hundreds of genes?
- What's next?

Your omics experiment (RNA-Seq, microarrays, proteomics, GWAS,...)



Gene list

MYB3
TLR4
TLR2
NFkB
DAG1
MEP1a
MEP1b
CYLD
USP40
APEH
USP3

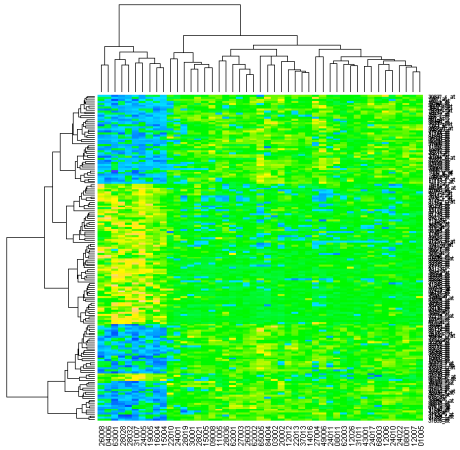
?

Gene List Annotation

- We test whether the differentially expressed genes in our experiment are enriched in some predefined gene lists.
- Based on previous knowledge

Previous Knowledge:
a collection of gene sets
(GO, Pathways)

Your experiment



Gene list

MYB3
TLR4
TLR2
NFkB
TEP1
MEP1a
MEP1b
CYLD

Gene list annotation

Biological Insight

Obtaining Biological Insight

- to summarize gene lists
- to help and speed up the interpretation of an experiment
- to gain mechanical insight
- to find regulated processes/pathways
- to find involved regulatory elements (TF, miRNA)
- to identify new members of a pathway
- to find similar experiments
- ....

Analysis based on gene lists is expected to be more robust and reproducible than single-gene analysis.

Enrichment Analysis

Over-Representation Analysis

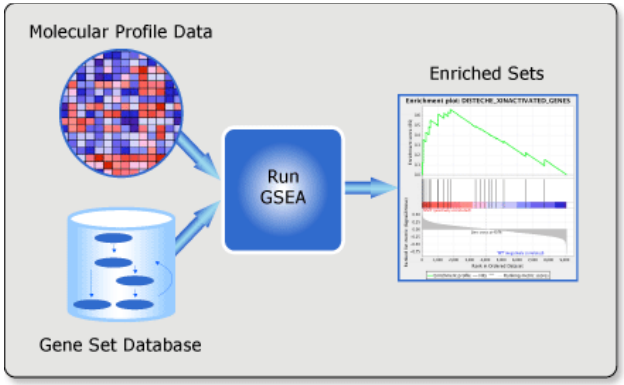
- hypergeometric aka Fisher's exact test
- input: 4 counts
- we need to set a cut-off a priori
- different results at different thresholds!

8	12
2	2412

Gene Set Enrichment Analysis (GSEA)

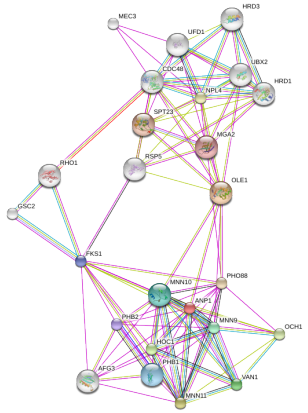
Subramanian et al. (2005) PNAS and many follow-up papers

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



Network Analysis

- covers also the less well understood portion of gene interactions
- based on co-expression, co-citation, PPI, ...
- example: STRING (http://www.string-db.org/)

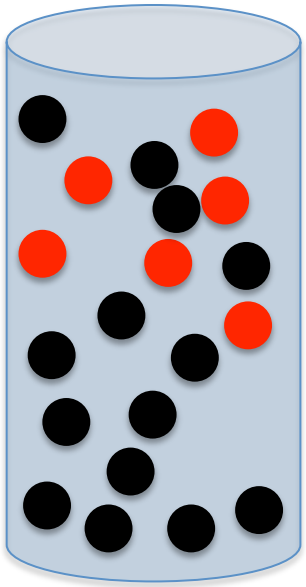


Over-Representation Analysis

5000 black and 10 red balls in an urn

each ball represents 1 gene

10 red balls ("Cytochromes")



Our list of differentially expr. genes: 4/5 balls are red

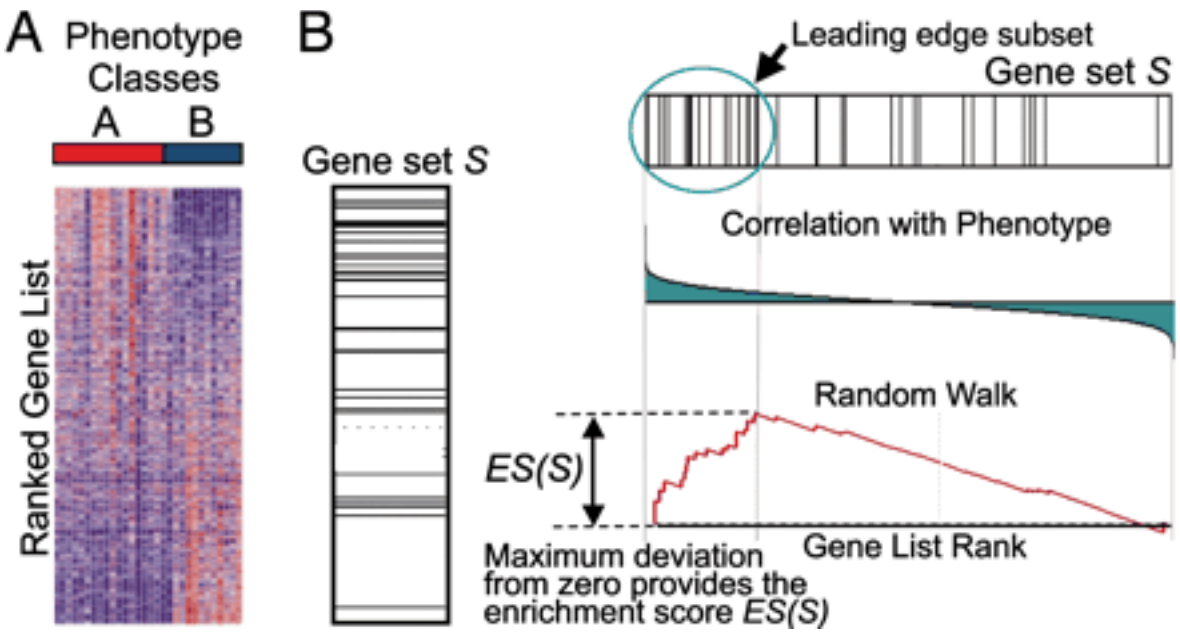
- CYP4F11
- CYP1A
- MEP1A
- CYP26B
- CYP3A43

What is the probability?
2x2 contigency table

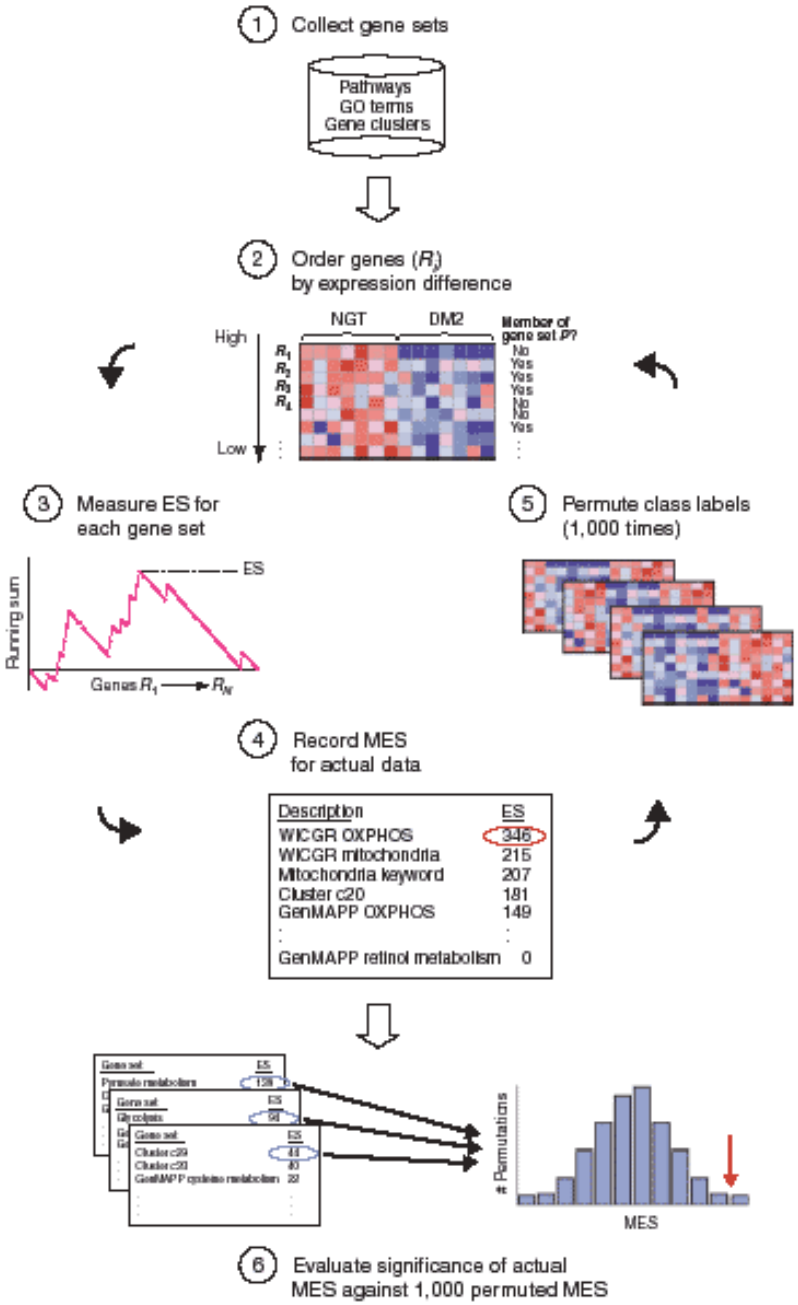
	Selected	Not
in category	4	6
not in category	1	4989

one-sided Fisher's exact test
p-value = 4.03e-11

Gene Set Enrichment Analysis



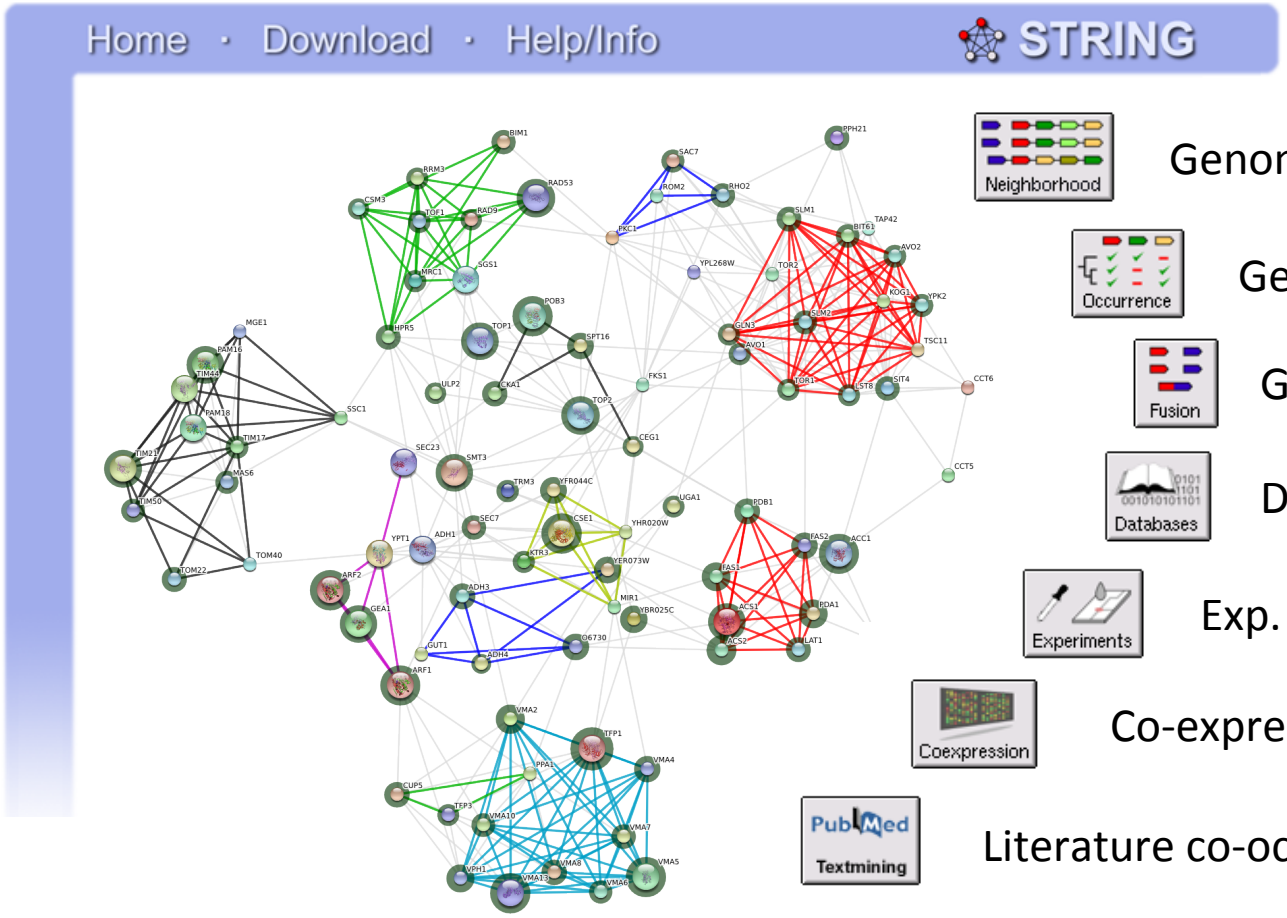
Subramanian et al. (2005) PNAS



Mootha et al. (2003) Nature Genetics

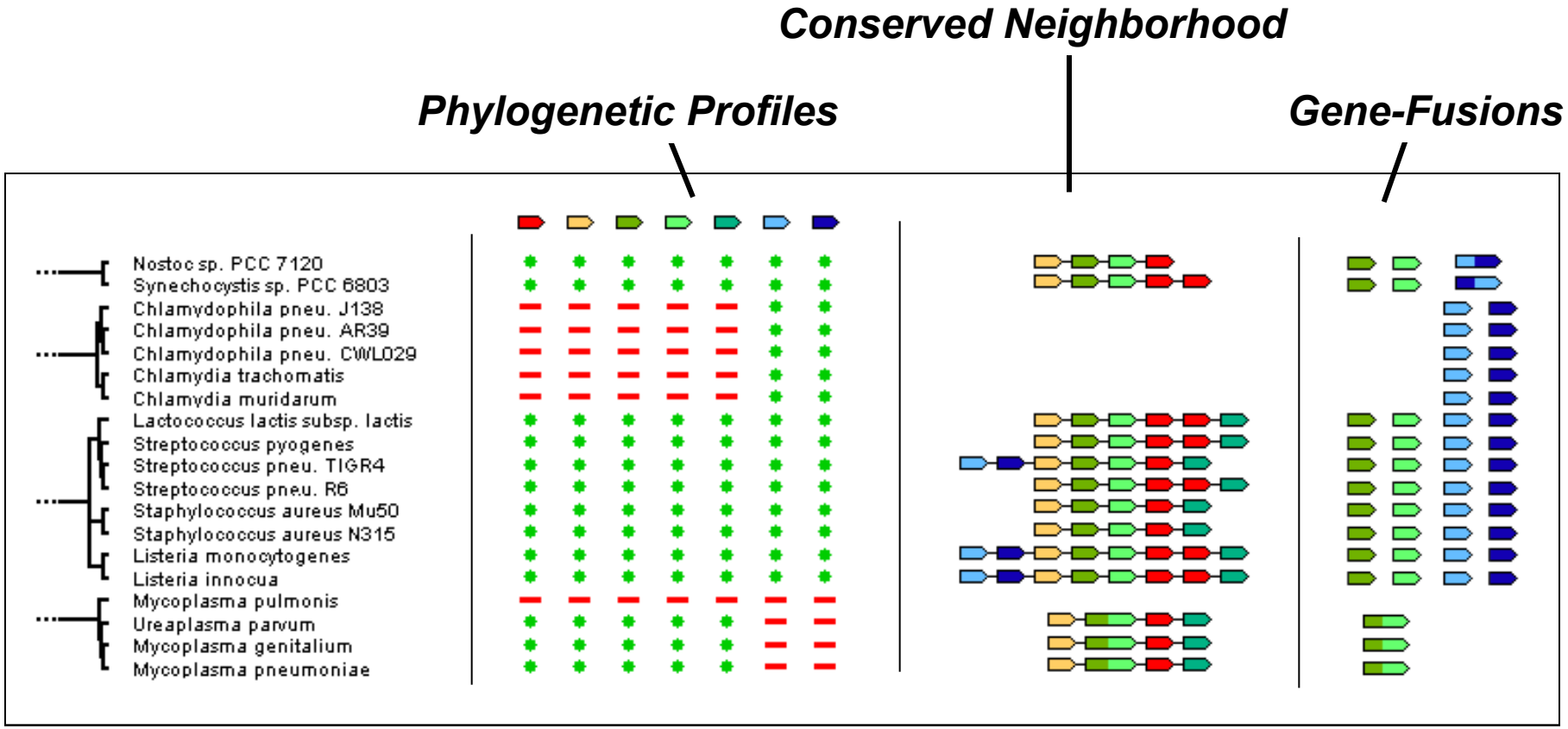
STRING

http://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*** *(>2000 species: Animals, Bact, Plants,…)*

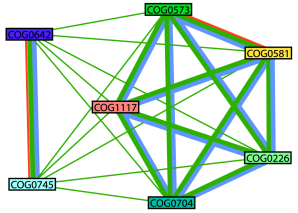
Interaction prediction from genome information



"genomic context"

quantify …

integrate …



networks

Other Interaction Sources

Interaction Databases

IntAct

BioGRID BETA

Human Protein Reference Database

Pathway Databases

Reactome

KEGG Kyoto Encyclopedia of Genes and Genomes

the Gene Ontology

PathwayInteractionDatabase

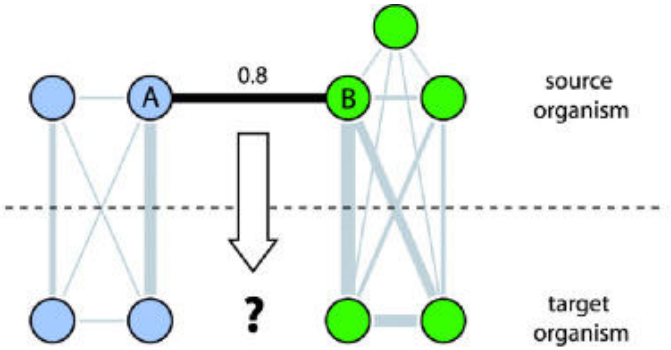
Automated Textmining

PubMed www.pubmed.gov

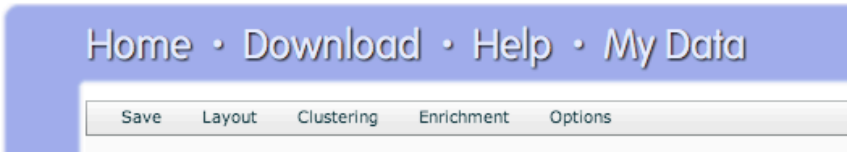
OMIM

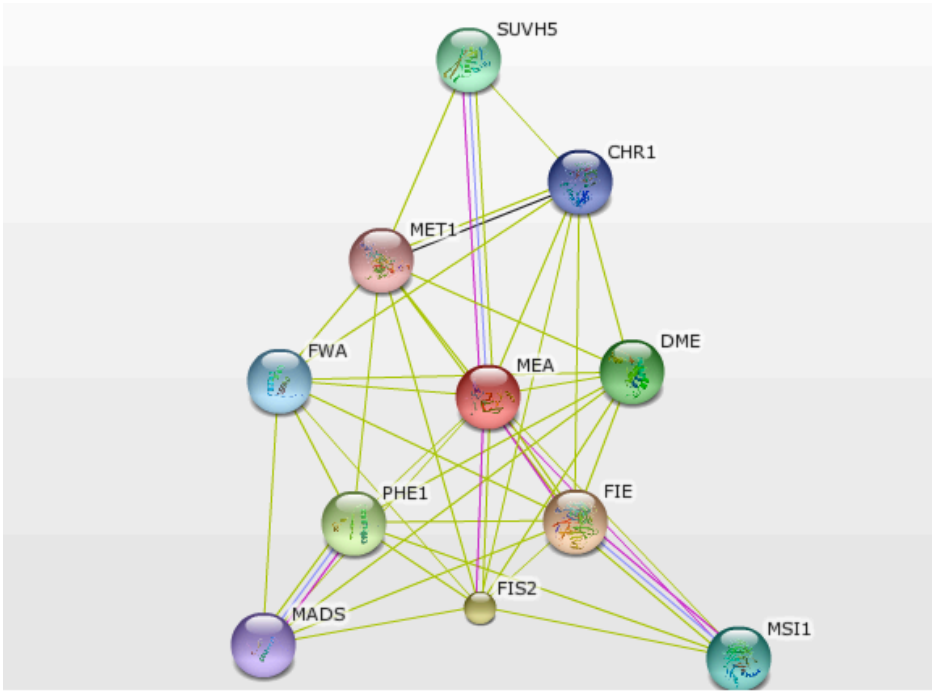
SGD

Interolog Transfer



Output





This is the **evidence view**. Different line colors represent the types of evidence for the associa

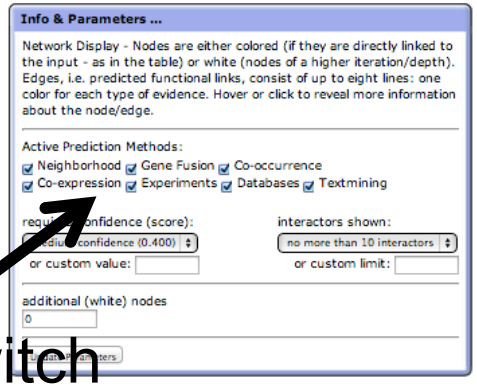
(requires Flash player 10 or better)

3 Views

Add more partners not in the input

Input

Pred. partners



Switch On/off channel

Clickable evidence

STRING / Network analysis

- can do more than gene list annotation:
  - Predicting gene function
  - Identifying candidates for an unknown enzyme in a pathway
  - Identifying new member genes of a biological process
  - Finding relevant literature

- ID mapper engine with large number of gene formats and synonyms

- STRING performs well compared with single-species databases

- R package to access STRING functionality from R
- available for download

Annotation Sources

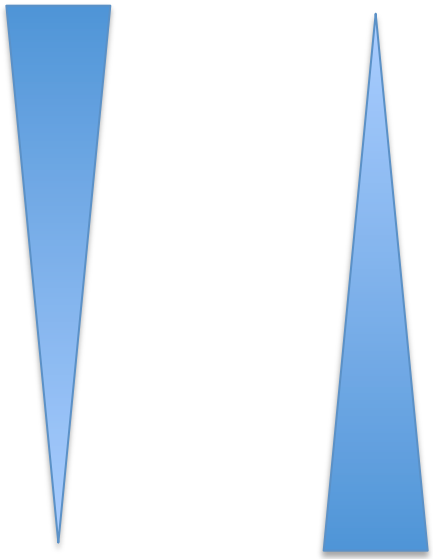
Pathways

KEGG, Reactome, BioCyc, ...

Gene Ontology (GO)

Gene/Protein Networks

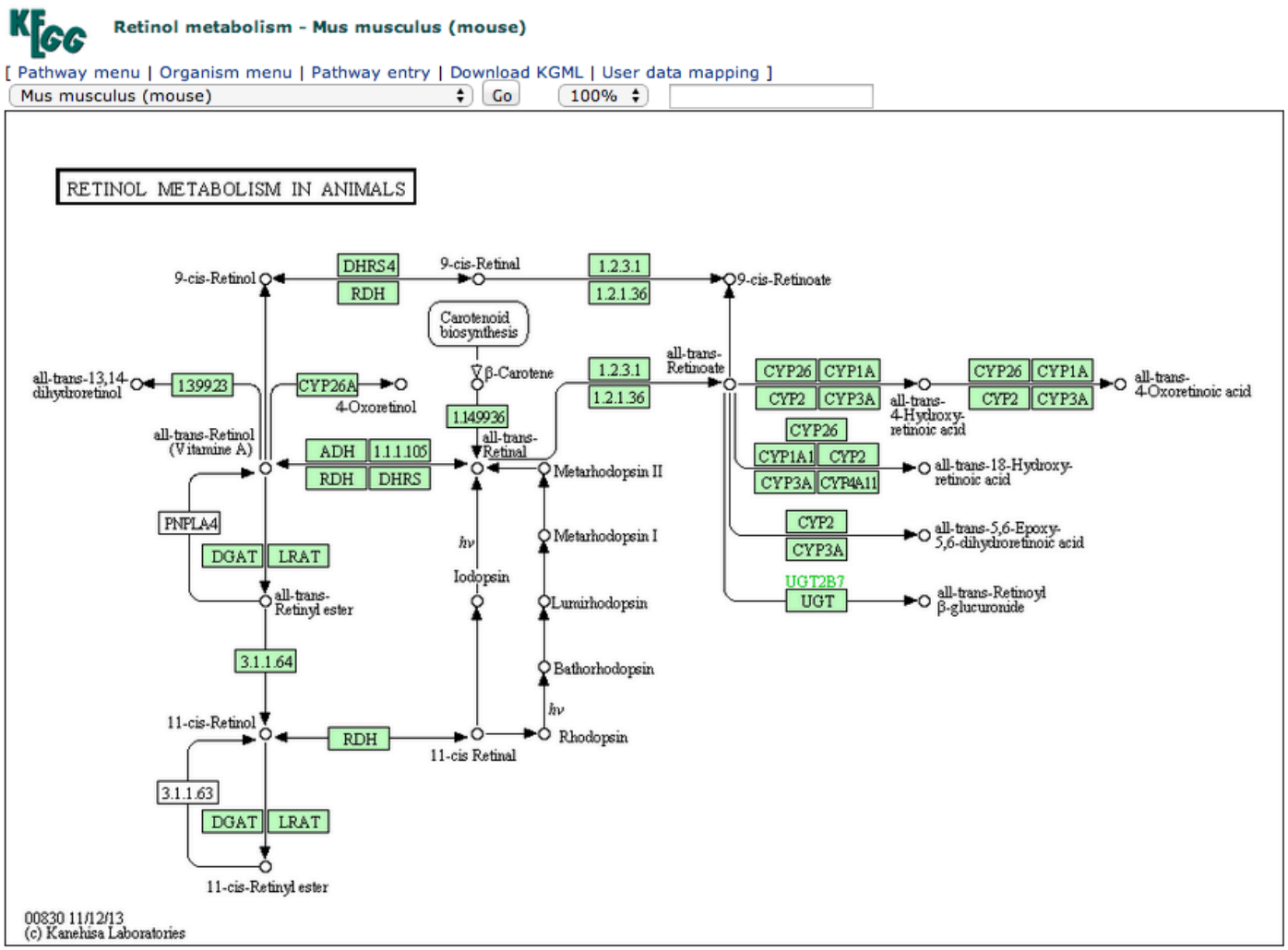
e.g. STRING



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) as of May 2015.

Example KEGG Pathway



Gene Ontology

Gene Ontology (GO)

http://www.geneontology.org/

- describes how gene products behave in a cellular context (BP, MF, C)
- controlled vocabulary of terms
- transparent (sources)
- manually curated lists for model species
- transfer to orthologs in other species (inferred annotation)

Example

murine ADAM10

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

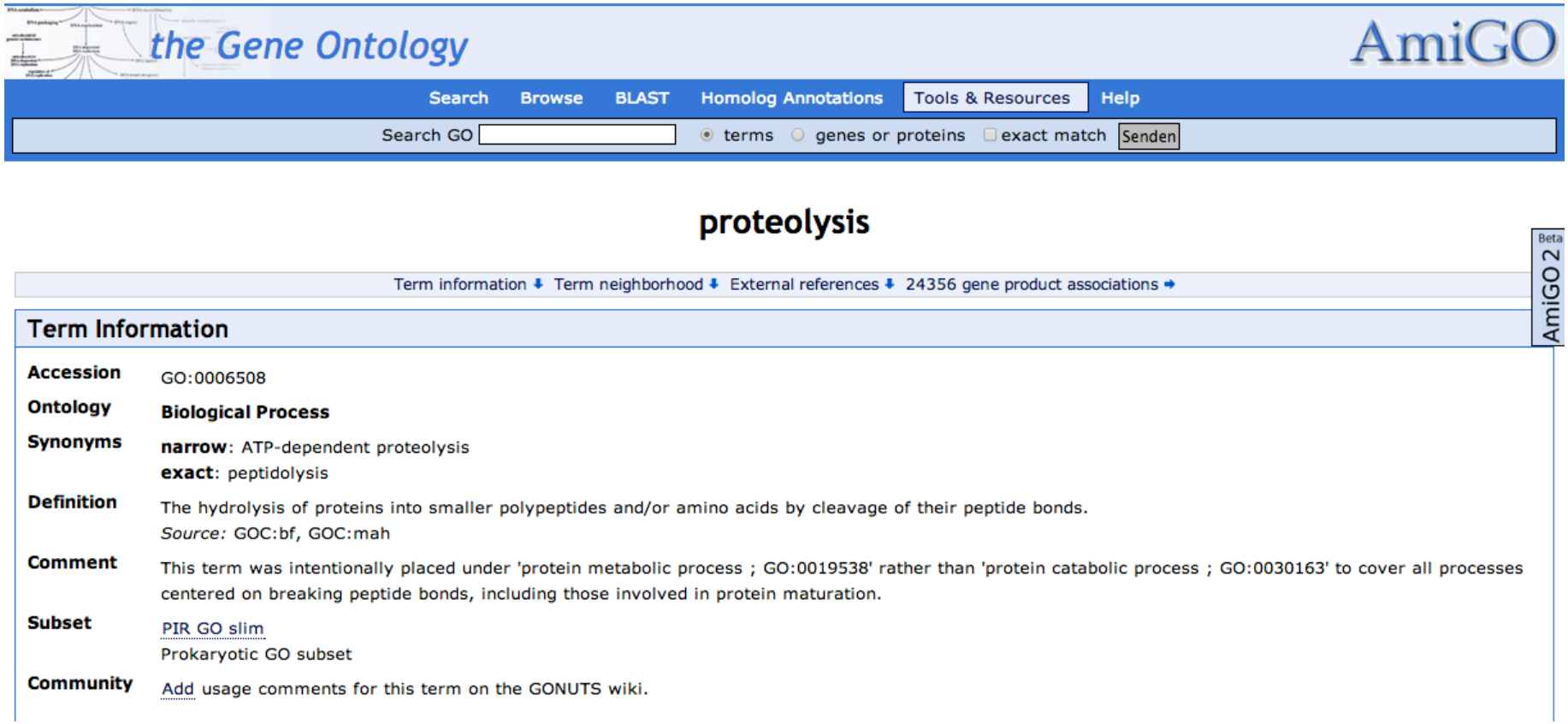
Biological Process
GO:0007220 Notch receptor processing
GO:0001701 in utero embryonic development
GO:0008284 positive regulation of cell proliferation
..

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

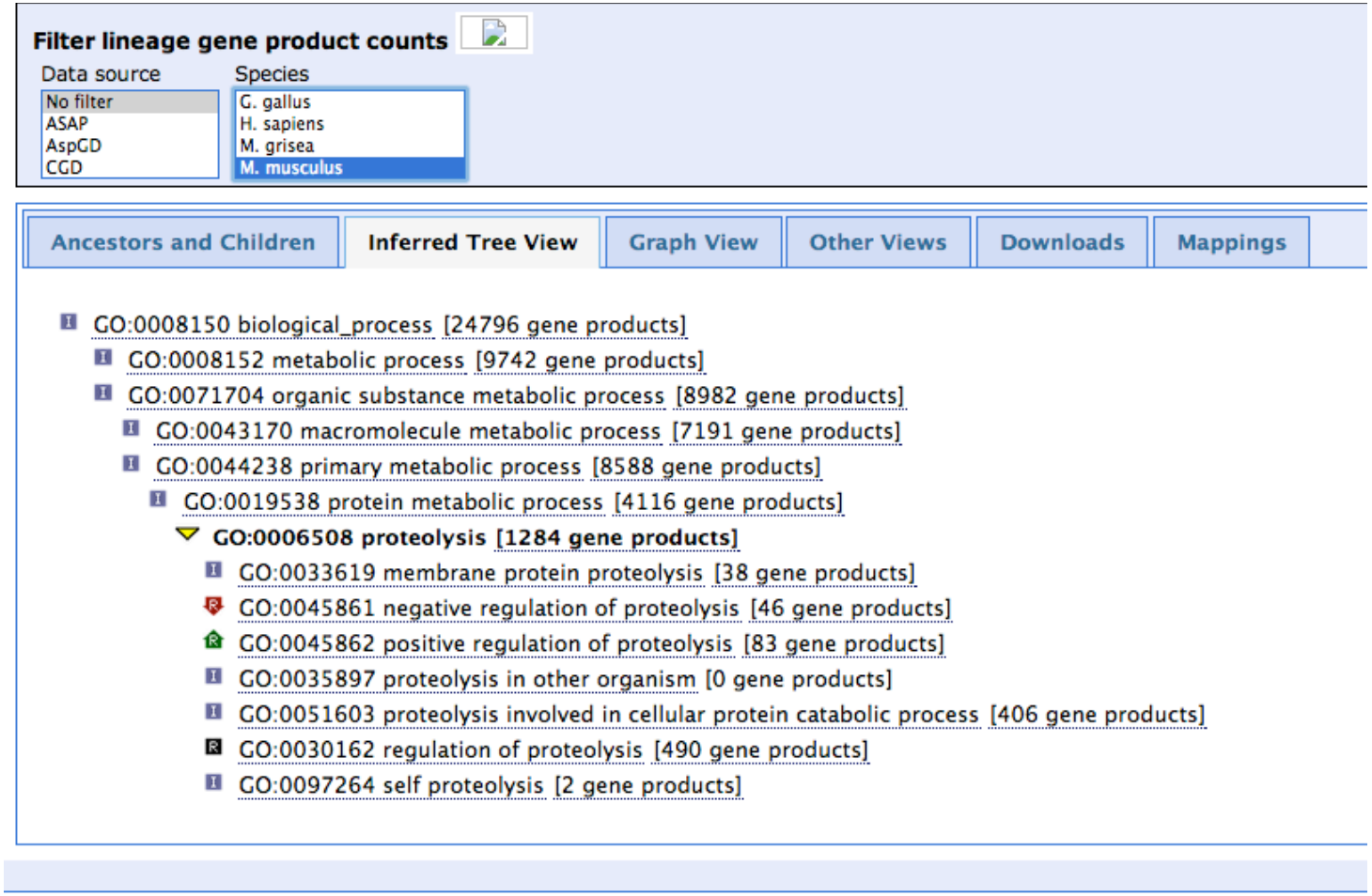
http://www.uniprot.org/uniprot/O35598

Lookup of GO terms

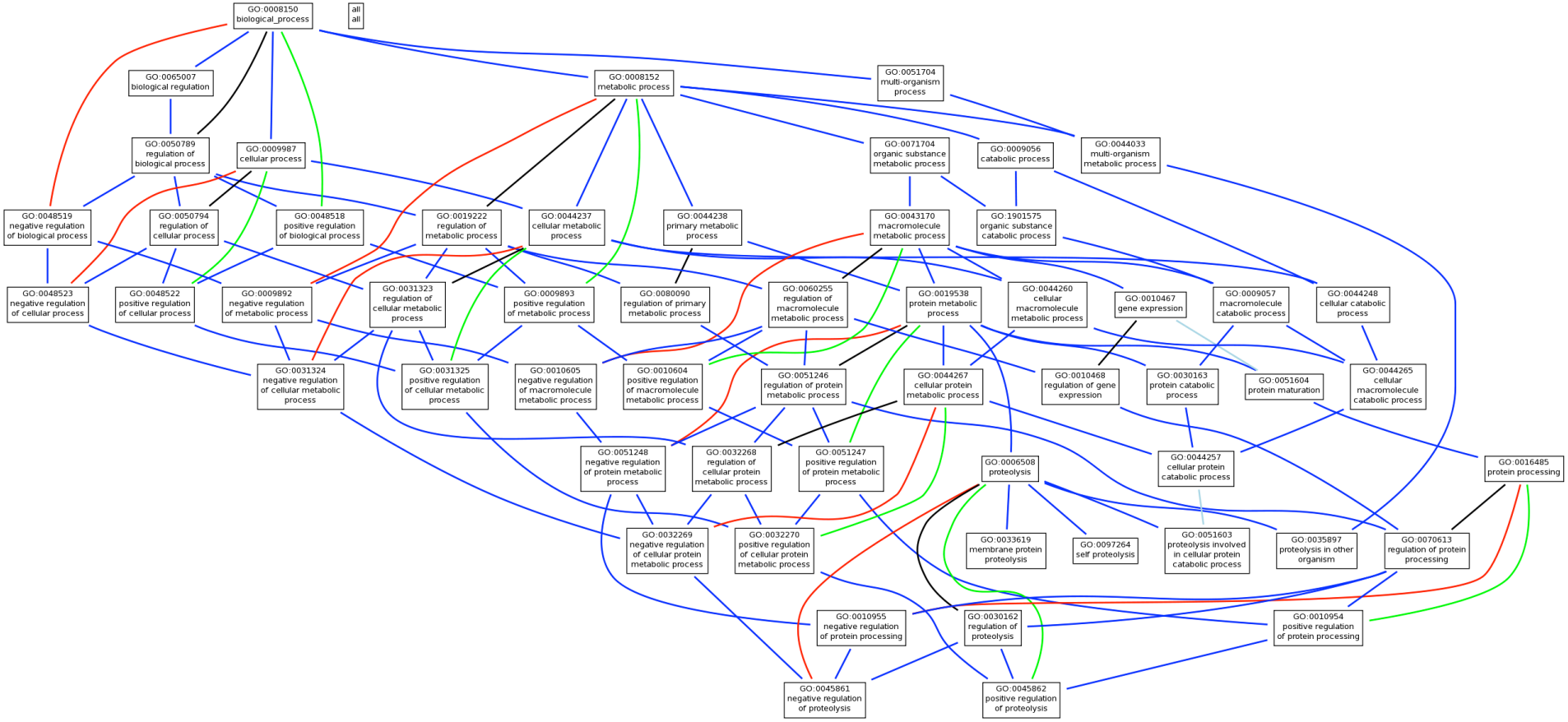
AmiGO
http://amigo.geneontology.org



GO Table View
GO:0006508 Proteolysis

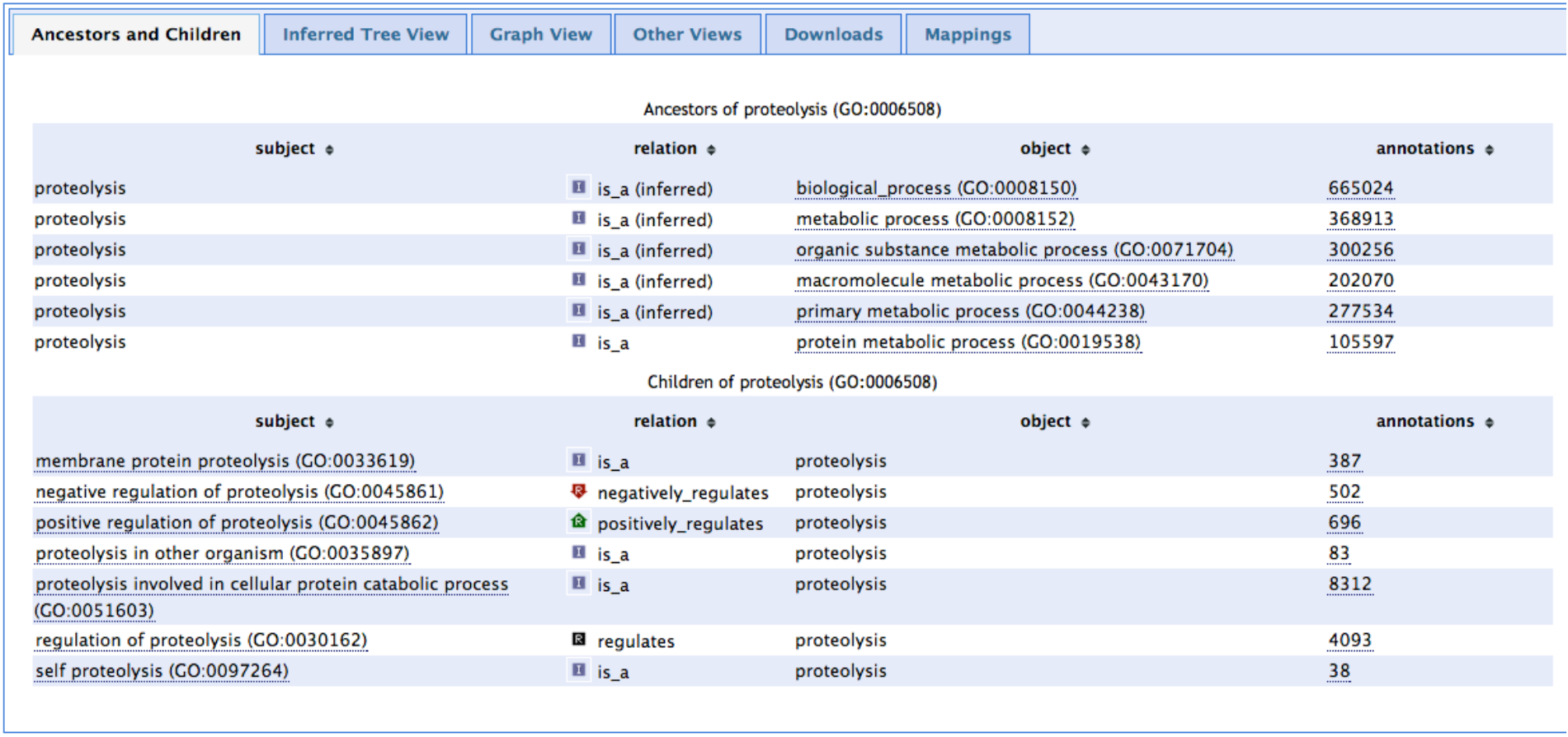


Graphical View
GO:0006508 Proteolysis



Ancestors and Children

AmiGO
http://amigo.geneontology.org



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

Ancestors of proteolysis (GO:0006508)

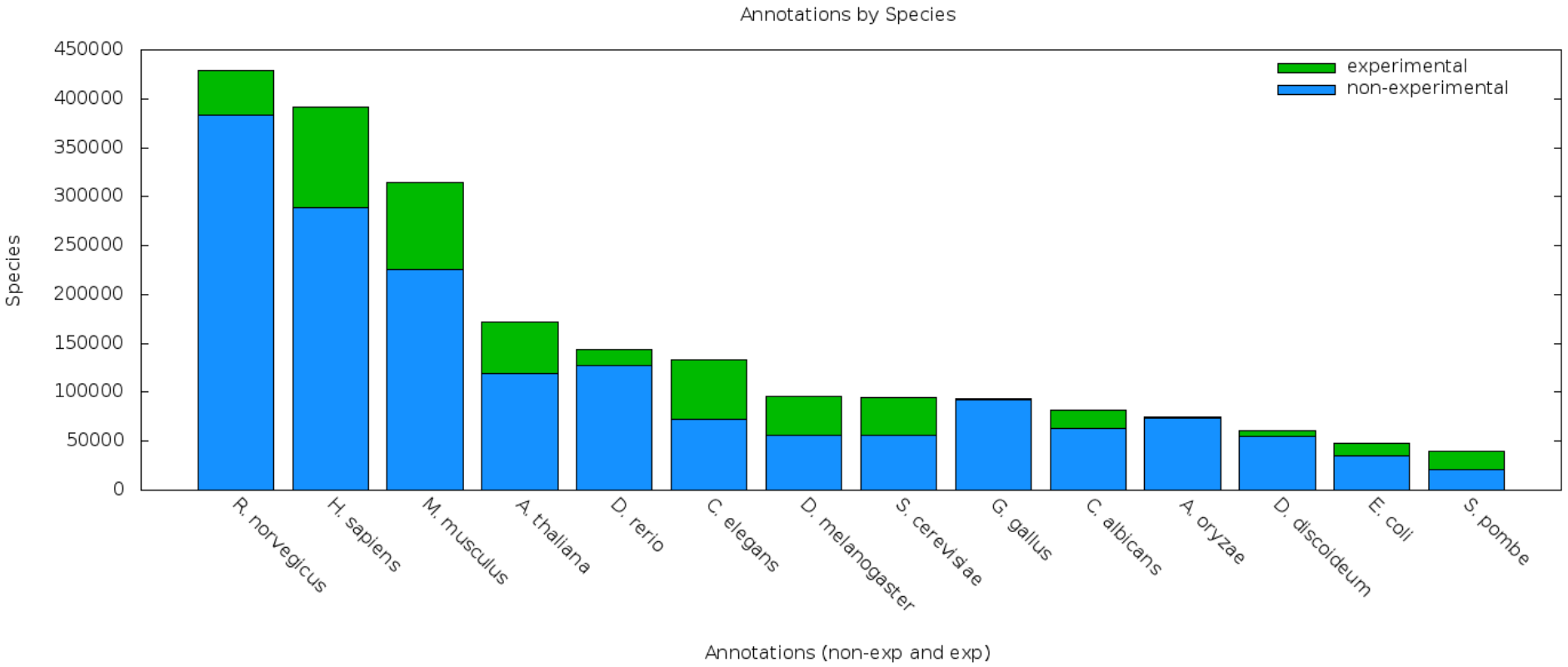
subject	relation	object	annotations
proteolysis	is_a (inferred)	biological_process (GO:0008150)	665024
proteolysis	is_a (inferred)	metabolic process (GO:0008152)	368913
proteolysis	is_a (inferred)	organic substance metabolic process (GO:0071704)	300256
proteolysis	is_a (inferred)	macromolecule metabolic 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 statistics

Even in model organisms only a minority of genes has experimental GO annotation



http://beta.geneontology.org/user-story-tags/everybody?page=5

False Discovery Rate (FDR)

Significance (alpha) level: probability of rejecting the null hypothesis given that it is true

Therefore at 5% significance level: for 100 tests where all null hypotheses are true, the expected number of incorrect rejections is 5

tests	incorrect rejections
100	5
10,000	500

Multiple Testing Correction

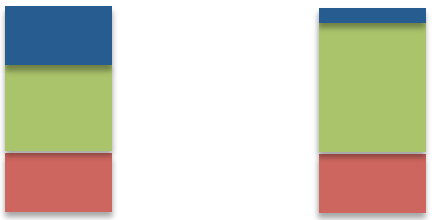
- Bonferroni
- False Discovery Rate (FDR): If we find 100 genes significantly differentially expressed at a 5% FDR, we expect at most 5 false discoveries in the list.

Experimental design

Interpretability depends mostly on appropriate experimental design!

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)



Group A Group B

Blood example during pregnancy

Summary

- Gene list annotation with Pathways and Gene Ontology can help to obtain biological insight.
- 3 main methods: 1. Over-Representation Analysis, 2. Gene Set Enrichment Analysis (GSEA), 3. 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

URLs & Tips

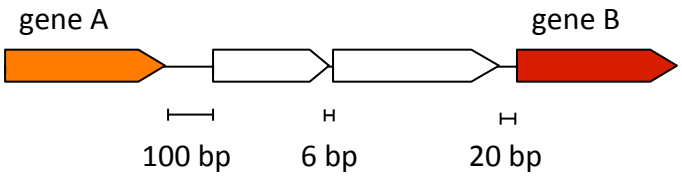
Main general Annotation Sources

- Gene Ontology (http://www.geneontology.org/)
  - AmiGO: http://amigo.geneontology.org
  - QuickGO: http://www.ebi.ac.uk/QuickGO/
  - Compilation of GO Tools: http://www.geneontology.org/GO.tools.shtml
- KEGG (http://www.genome.jp/kegg)
- Reactome (http://www.reactome.org)

- Most pathway databases offer also tools to colorize genes of interest on pathways
- Pathway analysis can also be done in R/bioconductor, see http://www.bioconductor.org/packages/release/BiocViews.html#___Pathways

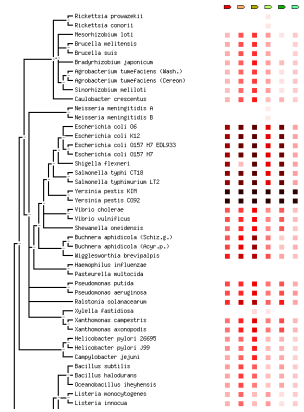
The raw score regimes

Neighborhood



raw score: sum of intergenic distances

Phylogenetic profiles



- "similarity profiles"
- singular value decomposition

raw score: euklidian distance

filter: downweigh scores
for homologous pairs

Fusion

raw score: constant (0.99)

experimental interactions

- two-hydrid, TAP, annotated complexes, …
- topology-based analysis: who with whom, how many other partners?

raw score: various (usually 'uniqueness' of interaction).

Co-expression

- download all microarray datasets
  for a given species
- data normalization (spatial correction)

raw score: pairwise pearson-correlation coefficient

Textmining

- download all PubMed abstracts
- identify proteins in the abstracts
- search for co-mentioned pairs

raw score: log-odds score