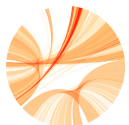


BIO634: making sense of gene lists



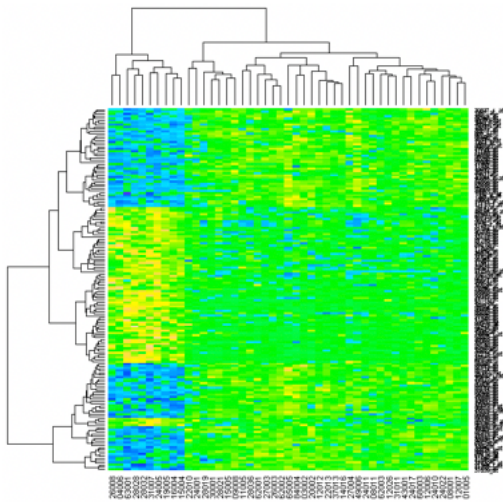
Adapted from **Stefan Wyder**
class on BIO634



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



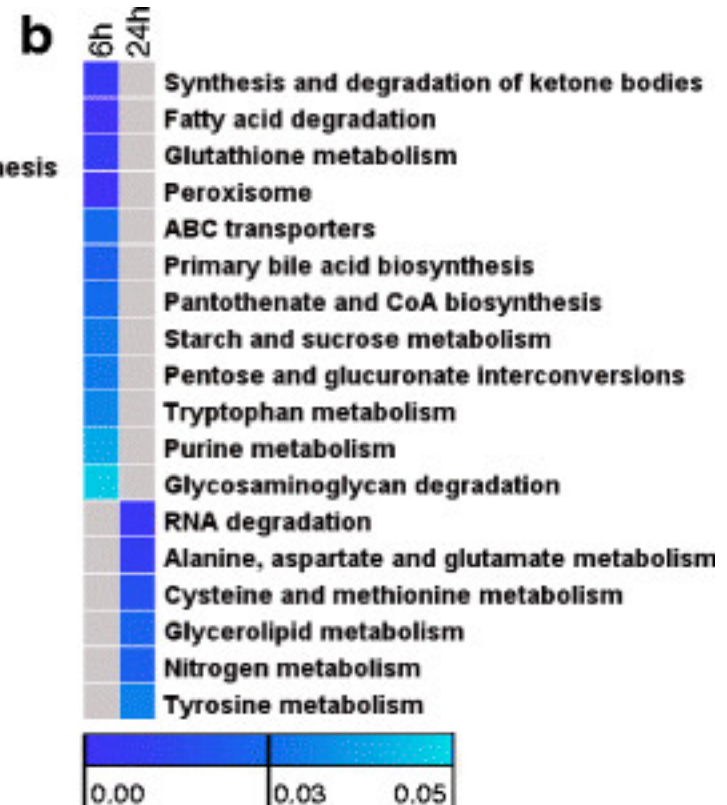
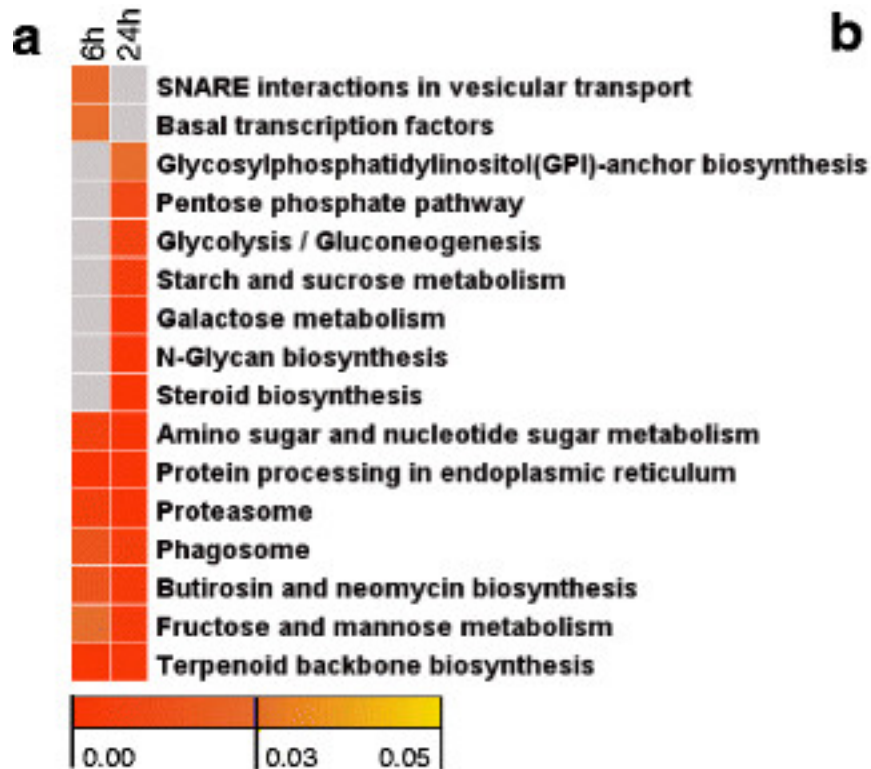
experimental data
RNA-seq, ChIP-seq, CLIP

geneA
geneB
geneC
geneD
...
geneX

annotation via previous knowledge
gene ontology, pathway analysis

**biological
insight**

Gene list annotation



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.

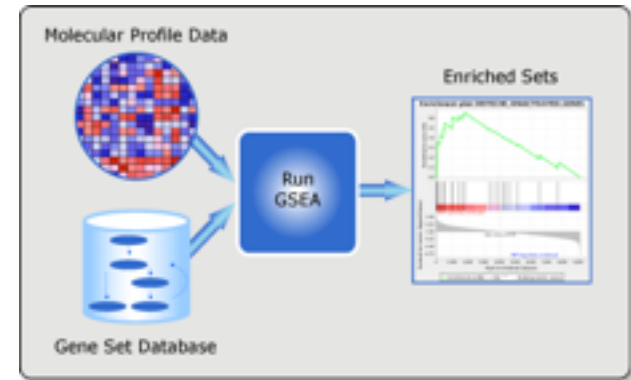
Enrichment analysis

1 Over-representation analysis

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

2 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



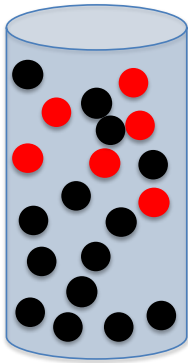
3 Network analysis

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



1 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 Ontology

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

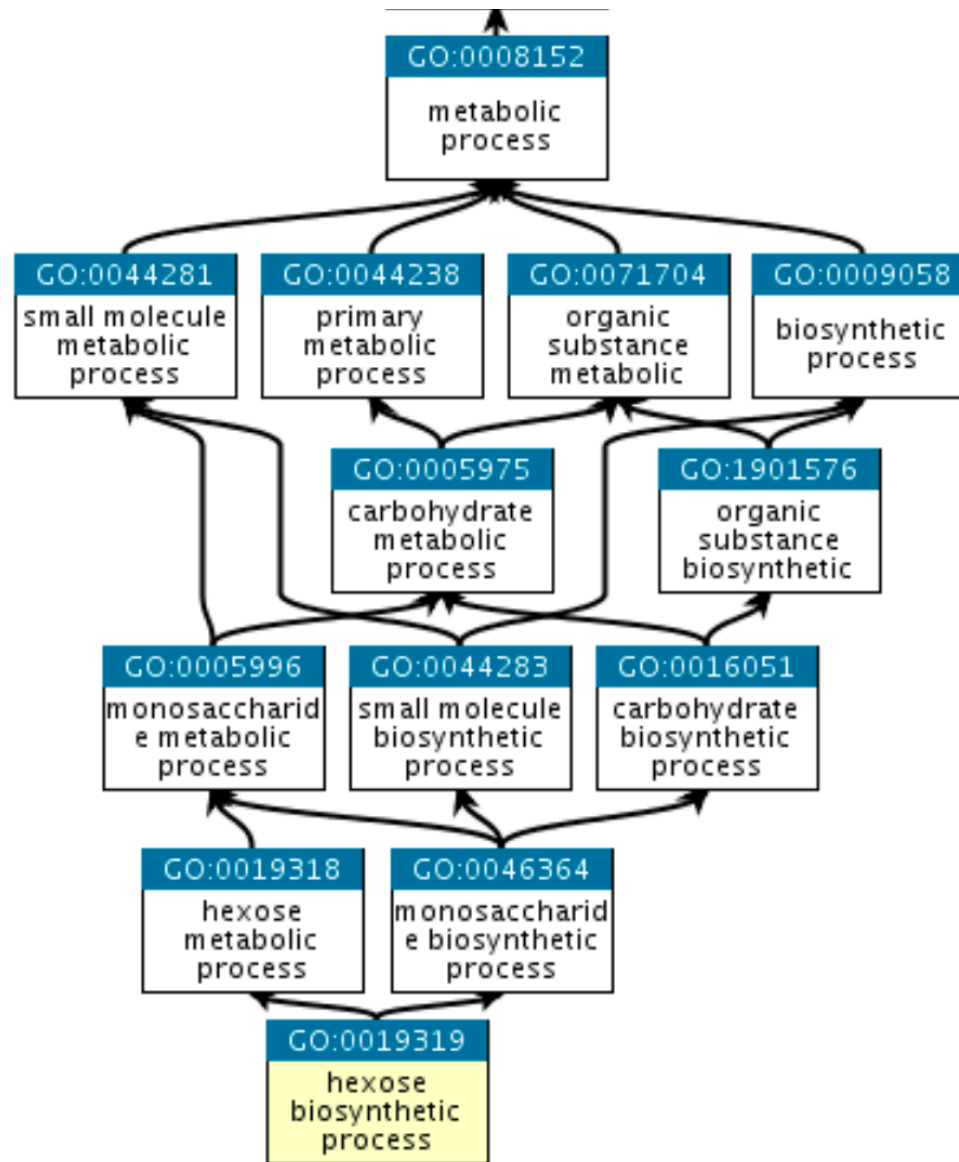
controlled vocabulary of terms

transparent (sources)

manually curated lists for model species

transfer to orthologs in other species (inferred annotation)

Gene Ontology



GO table view

Filter lineage gene product counts

Data source

No filter
ASAP
AspGD
CGD

Species

G. gallus
H. sapiens
M. grisea
M. musculus

Ancestors and Children

Inferred Tree View

Graph View

Other Views

Downloads

Mappings

- GO:0008150 biological process [24796 gene products]
 - GO:0008152 metabolic process [9742 gene products]
 - GO:0071704 organic substance metabolic process [8982 gene products]
 - GO:0043170 macromolecule metabolic process [7191 gene products]
 - GO:0044238 primary metabolic process [8588 gene products]
 - GO:0019538 protein metabolic process [4116 gene products]
 - GO:0006508 proteolysis [1284 gene products]
 - GO:0033619 membrane protein proteolysis [38 gene products]
 - GO:0045861 negative regulation of proteolysis [46 gene products]
 - GO:0045862 positive regulation of proteolysis [83 gene products]
 - GO:0035897 proteolysis in other organism [0 gene products]
 - GO:0051603 proteolysis involved in cellular protein catabolic process [406 gene products]
 - GO:0030162 regulation of proteolysis [490 gene products]
 - GO:0097264 self proteolysis [2 gene products]

GO:0006508 Proteolysis

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

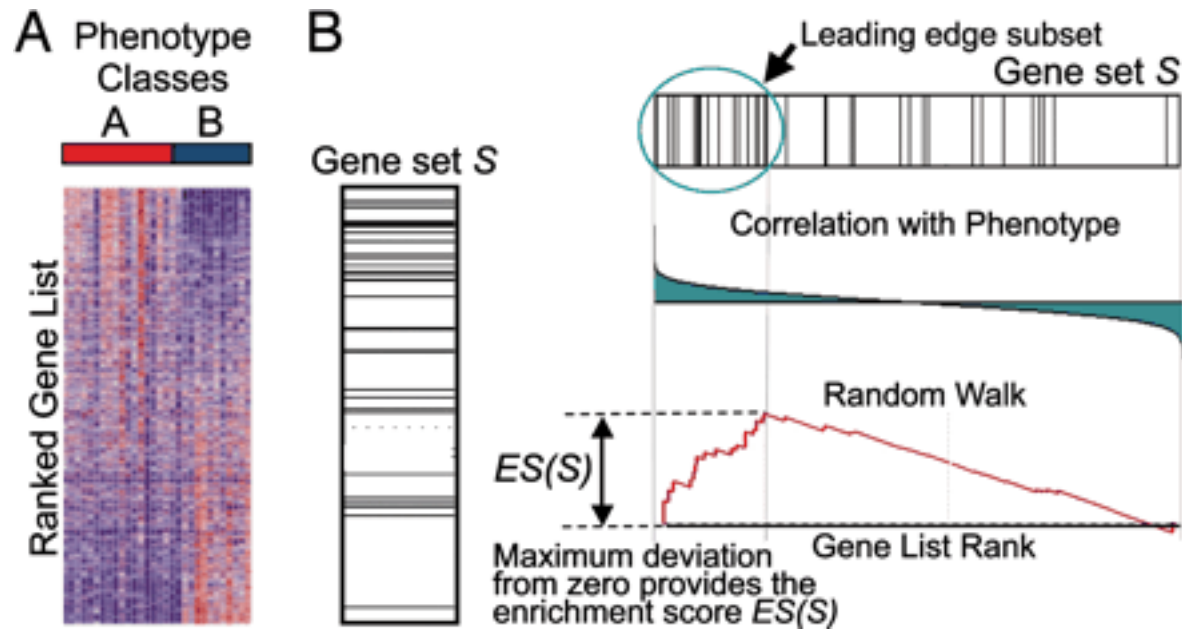
..

Cellular Compartment

GO:0005794 Golgi apparatus

GO:0009986 cell surface

2 Gene Set Enrichment Analysis



Subramanian et al. (2005) PNAS

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



Retinol metabolism - *Mus musculus* (mouse)

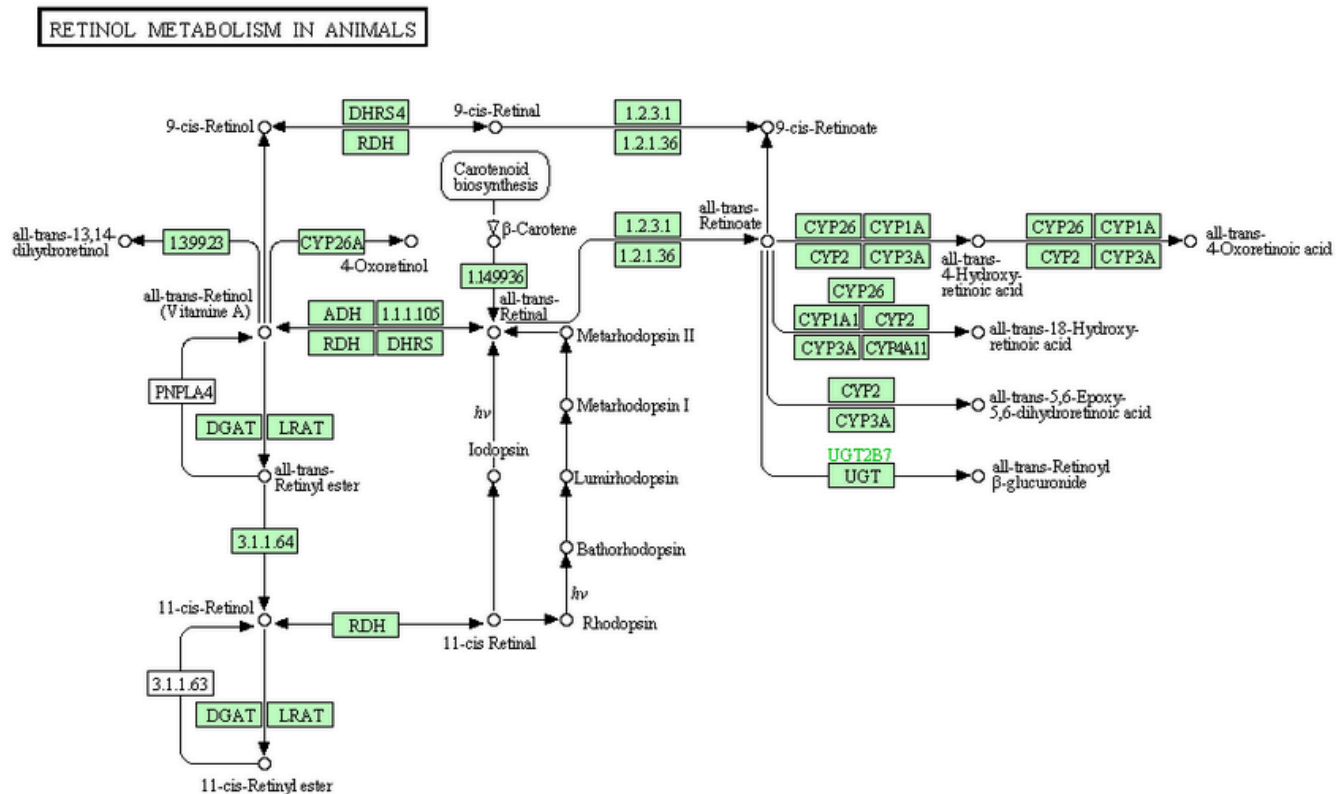
[\[Pathway menu | Organism menu | Pathway entry | Download KGML | User data mapping \]](#)

Mus musculus (mouse)

Go

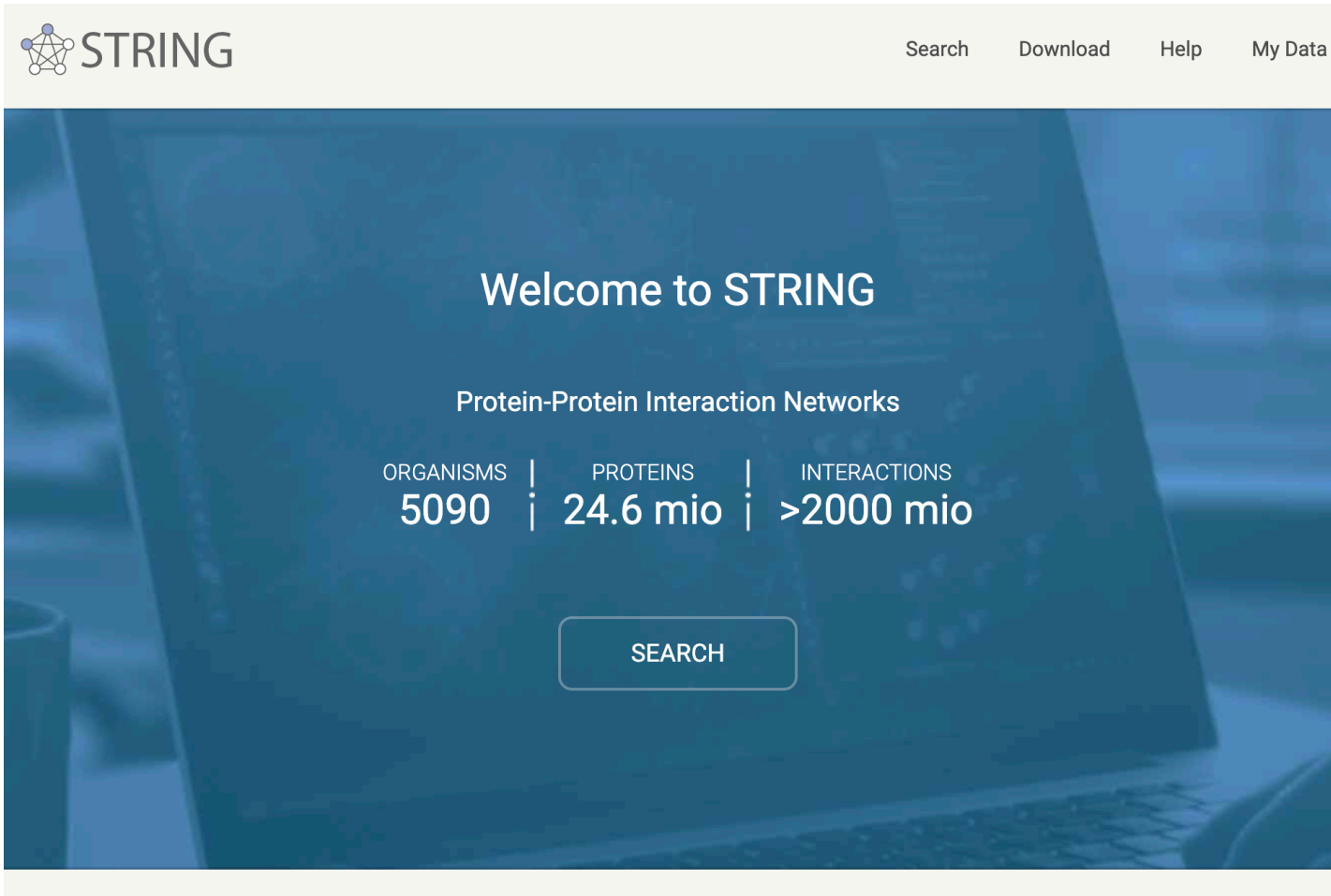
100%


0



00830 11/12/13
(c) Kanehisa Laboratories

3 Network analysis: string-db.org

The image is a screenshot of the STRING database homepage. At the top, there is a navigation bar with the STRING logo on the left and links for 'Search', 'Download', 'Help', and 'My Data' on the right. The main content area has a blue background with a blurred image of a laptop. In the center, it says 'Welcome to STRING' followed by 'Protein-Protein Interaction Networks'. Below this, there are three statistics: 'ORGANISMS 5090', 'PROTEINS 24.6 mio', and 'INTERACTIONS >2000 mio'. At the bottom center, there is a white 'SEARCH' button.

 **STRING**

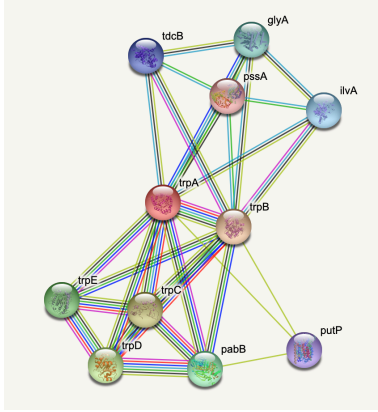
[Search](#) [Download](#) [Help](#) [My Data](#)

Welcome to STRING

Protein-Protein Interaction Networks

ORGANISMS	PROTEINS	INTERACTIONS
5090	24.6 mio	>2000 mio

[SEARCH](#)



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)



Network

currently showing

Summary view: shows current interactions. Nodes can be moved; popups provide information on nodes & edges.



Experiments

Co-purification, co-crystallization, Yeast2Hybrid, Genetic Interactions, etc ... as imported from primary sources.



Databases

Known metabolic pathways, protein complexes, signal transduction pathways, etc ... from curated databases.



Textmining

Automated, unsupervised textmining - searching for proteins that are frequently mentioned together.



Neighborhood

Groups of genes that are frequently observed in each other's genomic neighborhood.



Fusion

Genes that are sometimes fused into single open reading frames.



Cooccurrence

Gene families whose occurrence patterns across genomes show similarities.



Coexpression

Proteins whose genes are observed to be correlated in expression, across a large number of experiments.

Annotation sources

Pathways
KEGG, Reactome, BioCyc

Gene Ontology

Networks
STRING



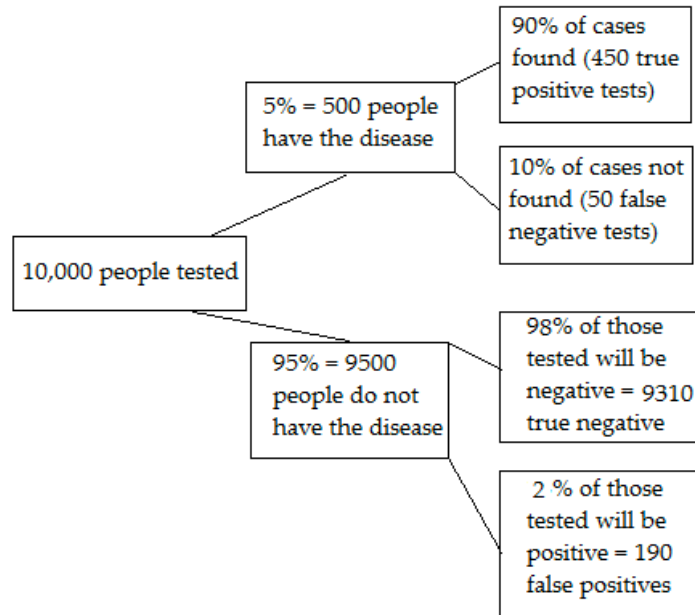
genes annotated



detail level

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 proteomics), 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