

NGS2 course

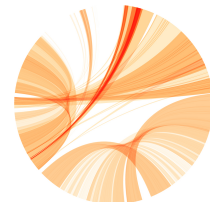
Making Sense of Gene Lists

Stefan Wyder

May 2015



**Universität
Zürich** ^{UZH}

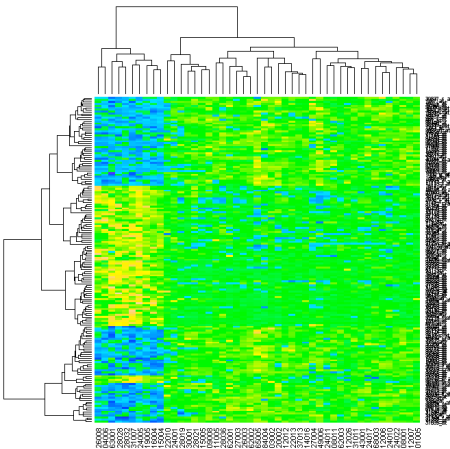


**URPP
Evolution
in Action**

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,...)



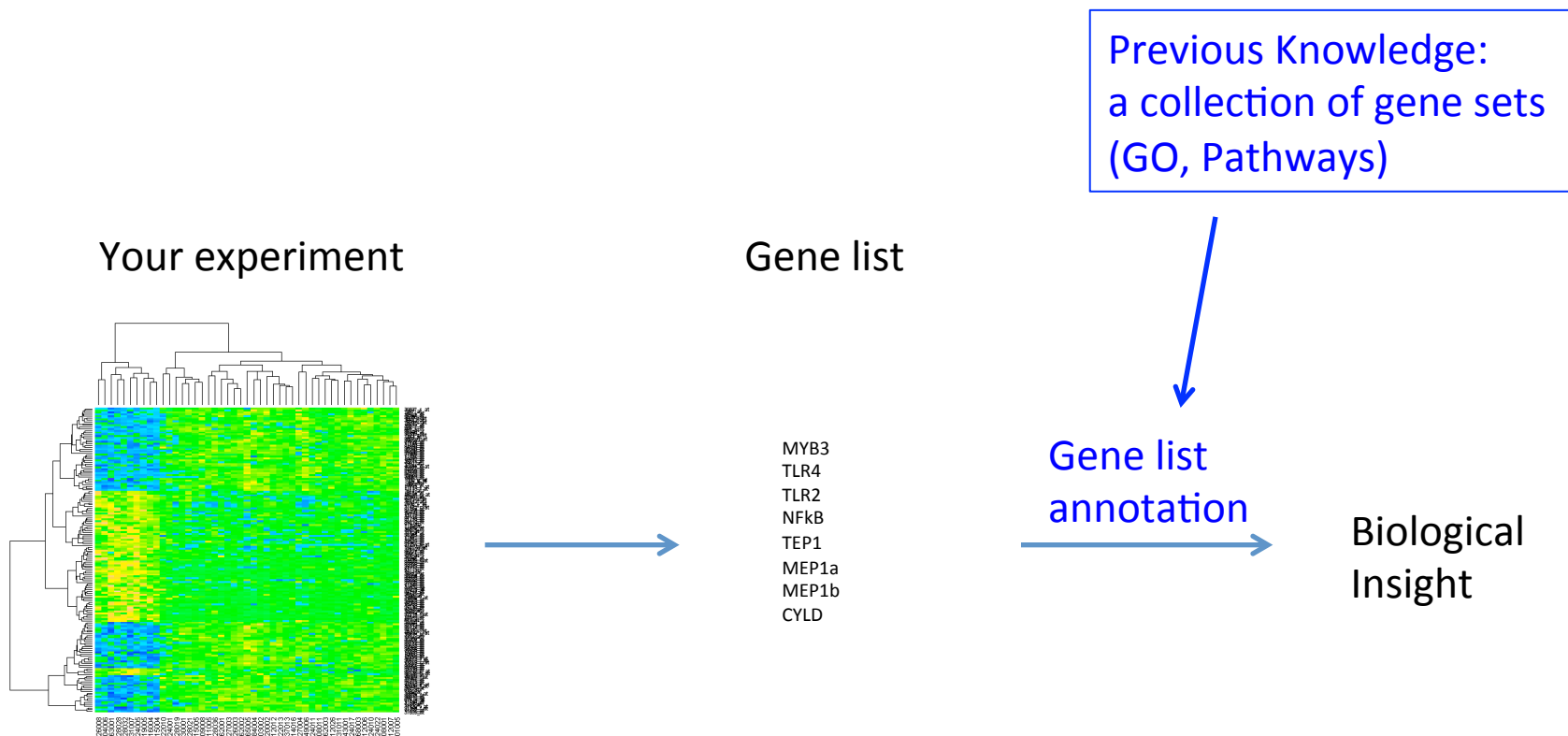
Gene list

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



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

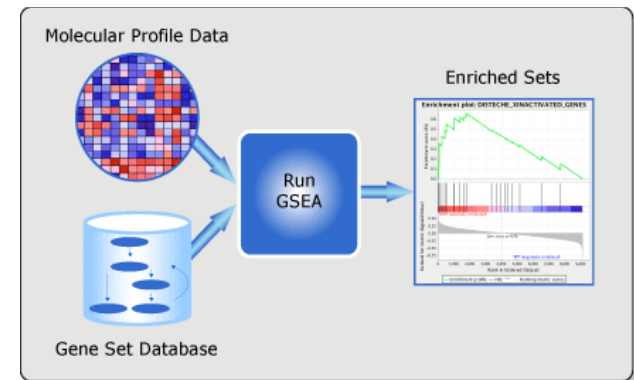
- hypergeometric aka Fisher's exact test
- input: 4 counts
- we need to set a cut-off a priori
- different results at different thresholds!

8	12
2	2412

Gene Set Enrichment Analysis (GSEA)

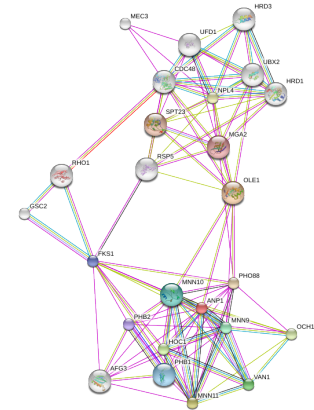
Subramanian et al. (2005) PNAS and many follow-up papers

- 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



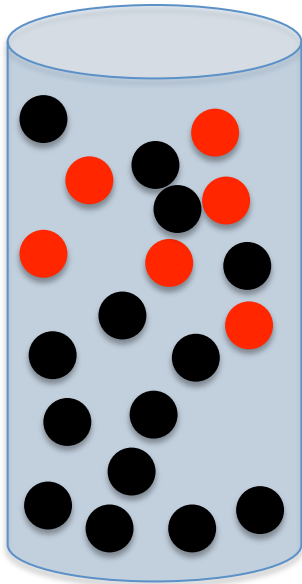
Network Analysis

- covers also the less well understood portion of gene interactions
- often inferred from co-expression data
- example: STRING (<http://www.string-db.org/>)
- combines info from co-expression, co-citation, PPI,



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

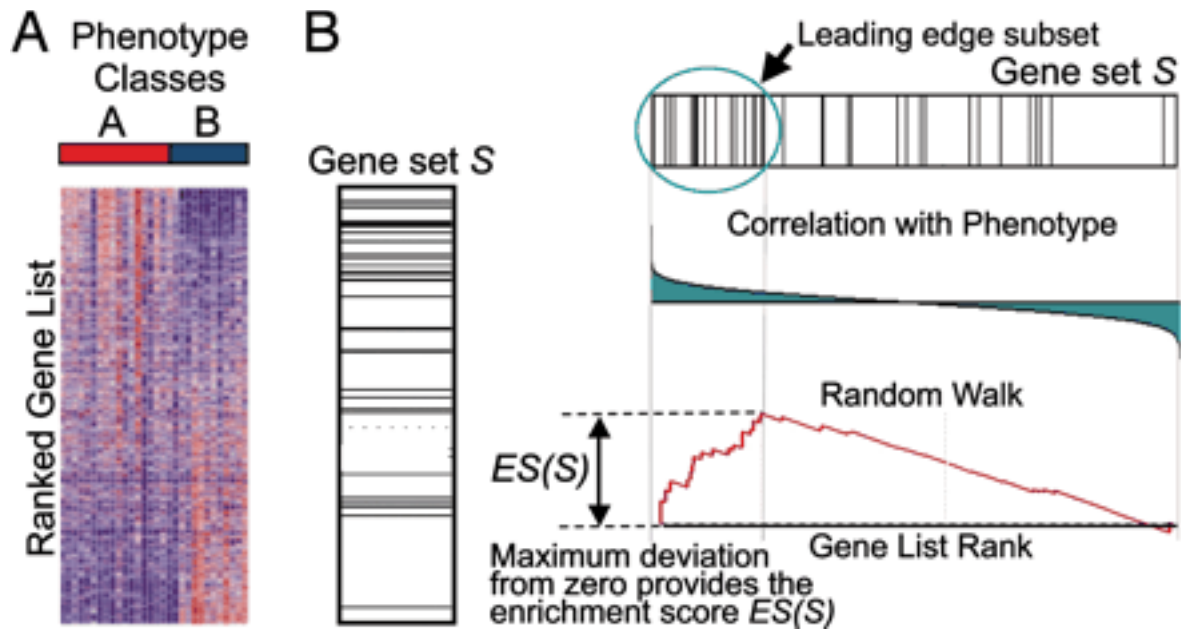
- CYP4F11
- CYP1A
- MEP1A
- CYP26B
- CYP3A43

What is the probability?
2x2 contingency table

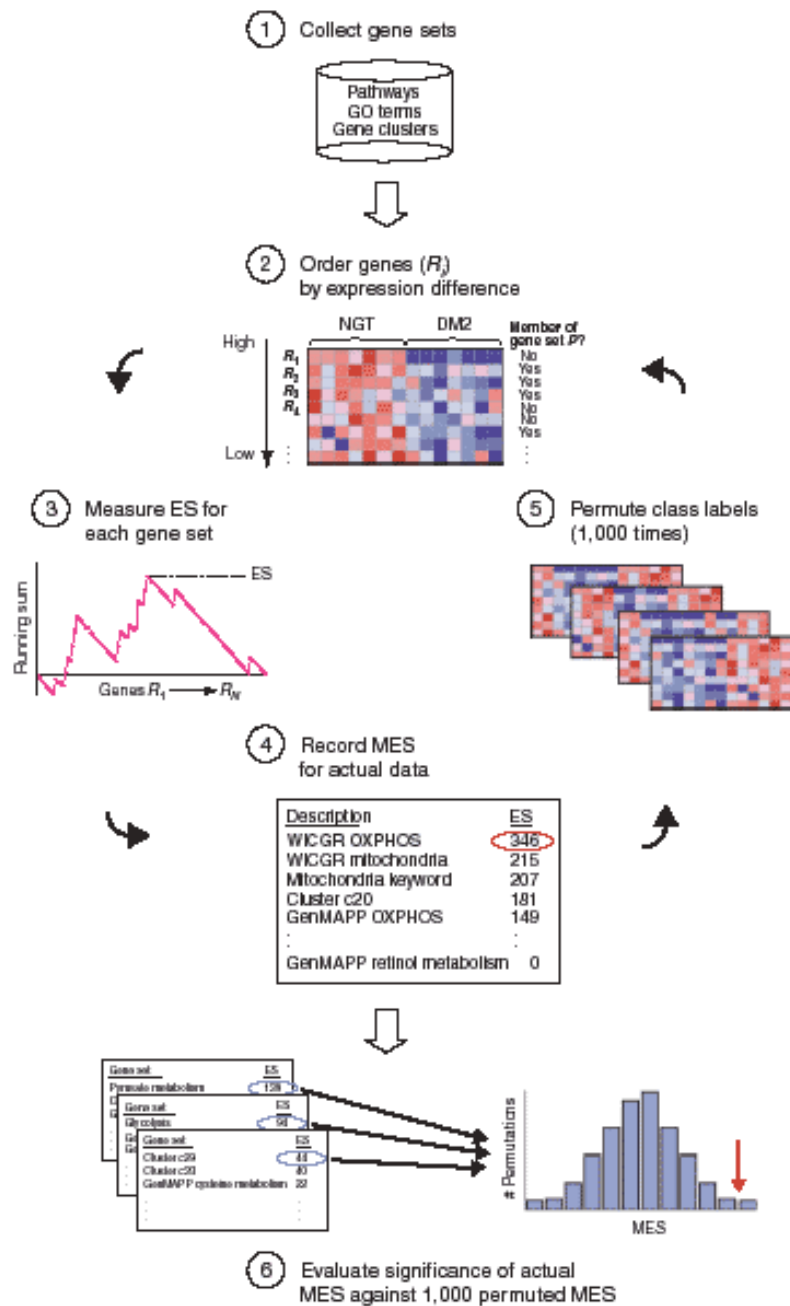
	Selected	Not
in category	4	6
not in category	1	4989

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

Gene Set Enrichment Analysis

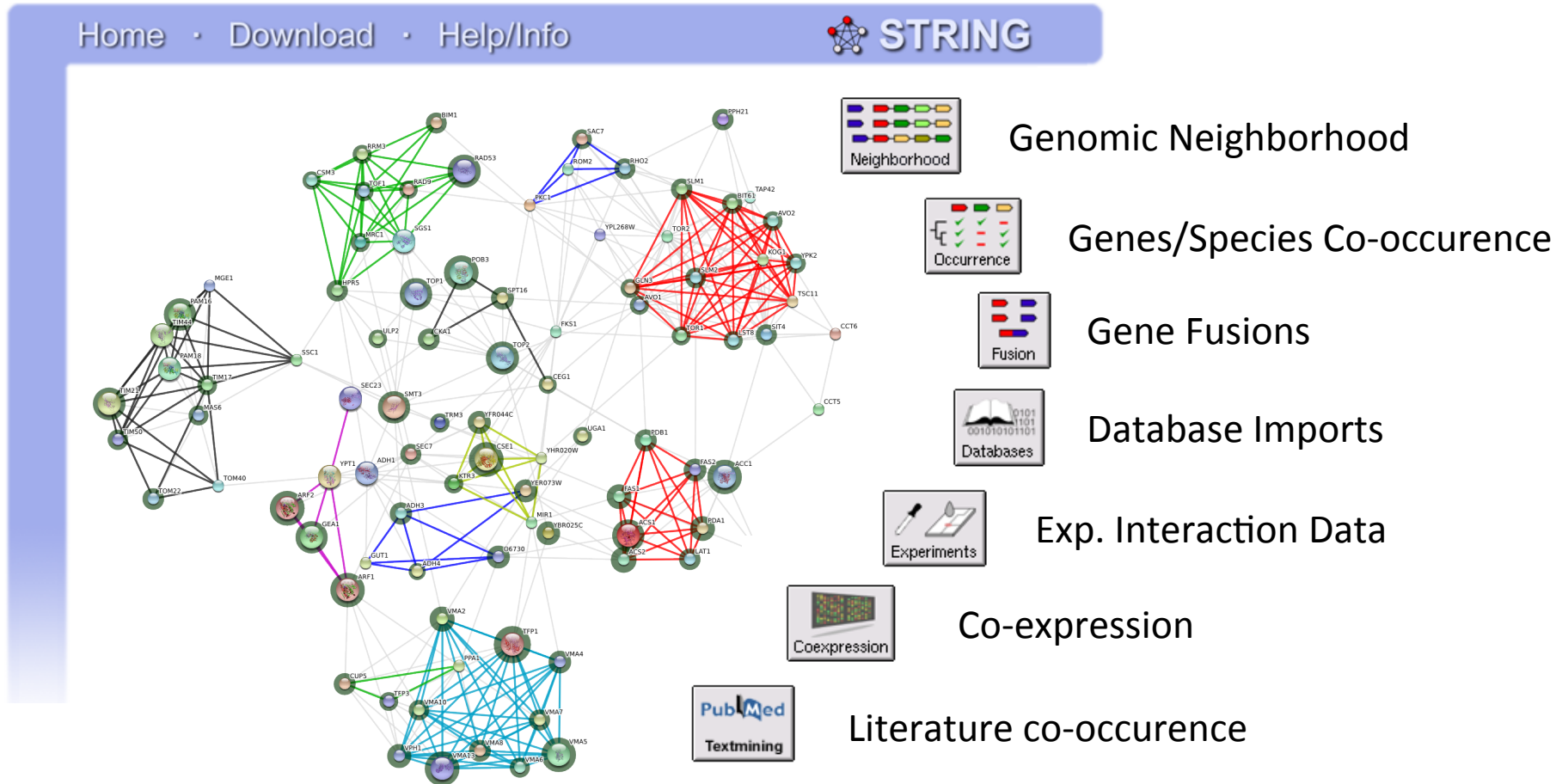


Subramanian et al. (2005) PNAS



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,...)***

109

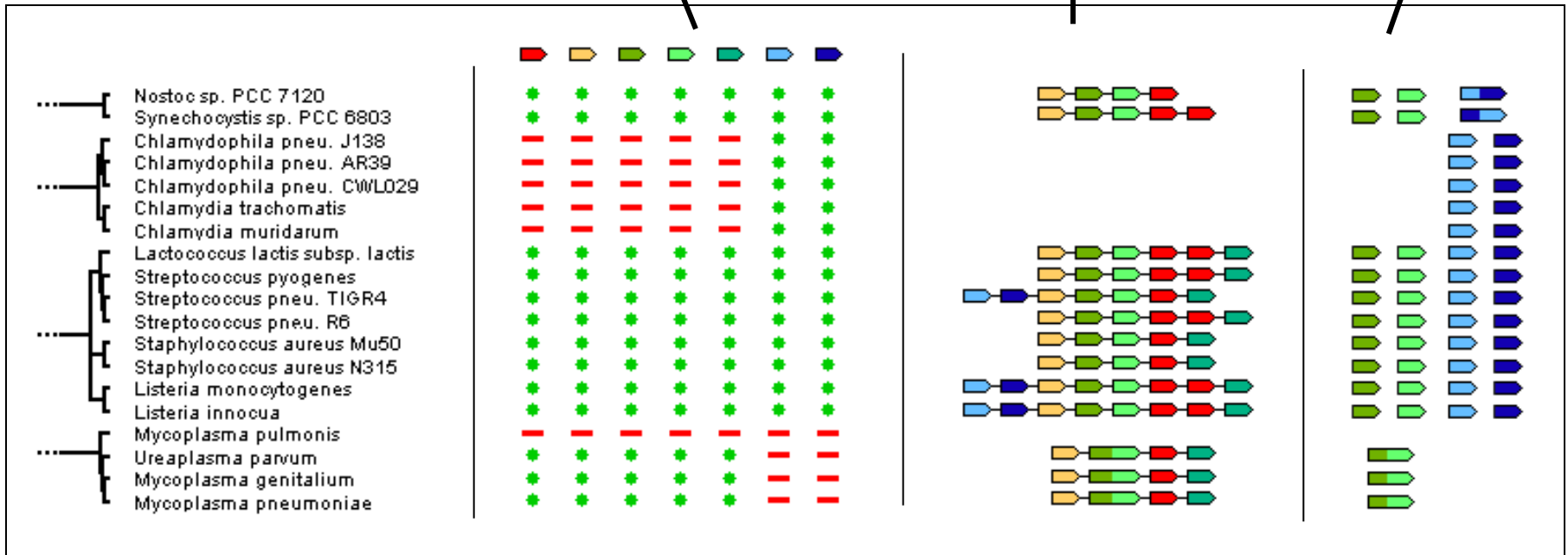
Interaction prediction from genome information

mainly bacteria

Conserved Neighborhood

Phylogenetic Profiles

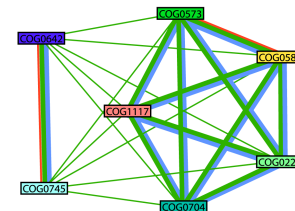
Gene-Fusions



“genomic context”

quantify ...

integrate ...



networks

Other Interaction Sources

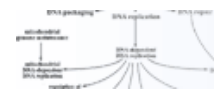
Interaction Databases



Pathway Databases



Reactome



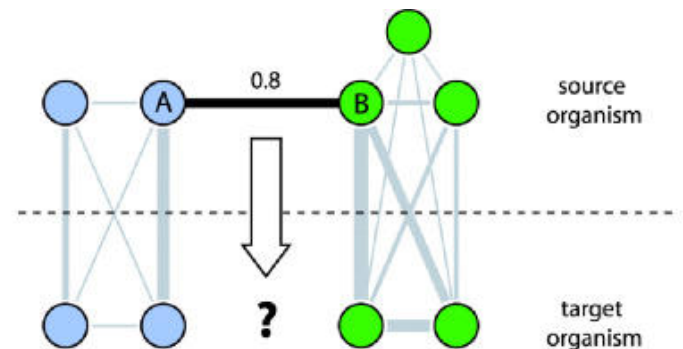
the Gene Ontology

PathwayInteractionDatabase

Automated Textmining



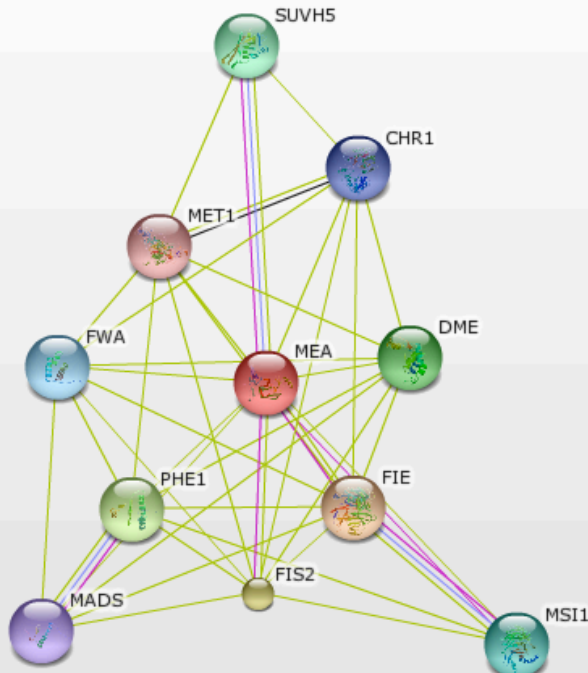
Interolog Transfer



Output

Home • Download • Help • My Data

Save Layout Clustering Enrichment Options



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



(requires Flash player 10 or better)

3 Views

Add more partners not in the input

Input

Your Input:

- MEA MEA (MEDEA); sequence-specific DNA binding / transcription factor; Encodes a putative transcription factor MEDEA (MEA) that negatively regulates seed development in the absence of fertilization. Mutations in this locus result in embryo lethality. MEA is a Polycomb group gene that is imprinted in the endosperm. The maternal allele is expressed and the paternal allele is silent. MEA is controlled by DEMETER (DME), a DNA glycosylase required to activate MEA expression, and MET1 (METHYLTRANSFERASE 1 (MET1)), which maintains CG methylation at the MEA locus. MEA is involved in the regulation of gene expression. (1689 aa) (*Arabidopsis thaliana*)

Predicted Functional Partners:

		Neighborhood	Gene Fusion	Co-occurrence	Experiments	Databases	Textmining	[Homology]	Score
FIE	FIE (FERTILIZATION-INDEPENDENT ENDOSPERM); nucleotide binding / transcription factor/ transcrip [...]								0.999
FIS2	FIS2 (FERTILIZATION INDEPENDENT SEED 2); transcription factor; Encodes a negative regulator of [...]								0.999
PHE1	PHE1 (PHERES1); DNA binding / transcription factor; Type I MADS-box protein, regulated by MEA a [...]								0.987
DME	DME (DEMETER); DNA N-glycosylase; DNA-(apurinic or apyrimidinic site) lyase; Encodes a DNA glyco [...]								0.984
SUVH5	SUVH5 (SU(VAR)5); H3K9me3-binding protein; Encodes a heterodimeric protein [...]								0.982
MSI1	MSI1 (MULTIDRUG RESISTANCE 1); protein; Encodes a multidrug resistance protein [...]								0.949
FWA	FWA; DNA binding / transcription factor; Encodes a transcription factor [...]								0.867
CHR1	CHR1 (CHROMATIN REMODELING 1); ATPase/ helicase; Protein is similar to SWI2/SNF2 chromatin remo [...]								0.790
MADS	MADS-box protein (AGL40); MADS-box protein (AGL40); FUNCTIONS IN- transcription factor activity [...]								0.786
MET1	MET1 (METHYLTRANSFERASE 1); methyltransferase; Encodes a cytosine methyltransferase MET1. Requi [...]								0.742

Views:



Info & Parameters ...

Network Display - Nodes are either colored (if they are directly linked to the input - as in the table) or white (nodes of a higher iteration/depth). Edges, i.e. predicted functional links, consist of up to eight lines; one color for each type of evidence. Hover or click to reveal more information about the node/edge.

Active Prediction Methods:

- ☒ Neighborhood ☒ Gene Fusion ☒ Co-occurrence
- ☒ Co-expression ☒ Experiments ☒ Databases ☒ Textmining

required confidence (score):

required confidence (0.400) +

or custom value:

interactors shown:

no more than 10 interactors +

or custom limit:

additional (white) nodes

0

show more partners

Clickable evidence

Switch
On/off
channel

STRING

- 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 understands a large number of gene formats
- 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



Level of
Detail



genes
annotated

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



Retinol metabolism - Mus musculus (mouse)

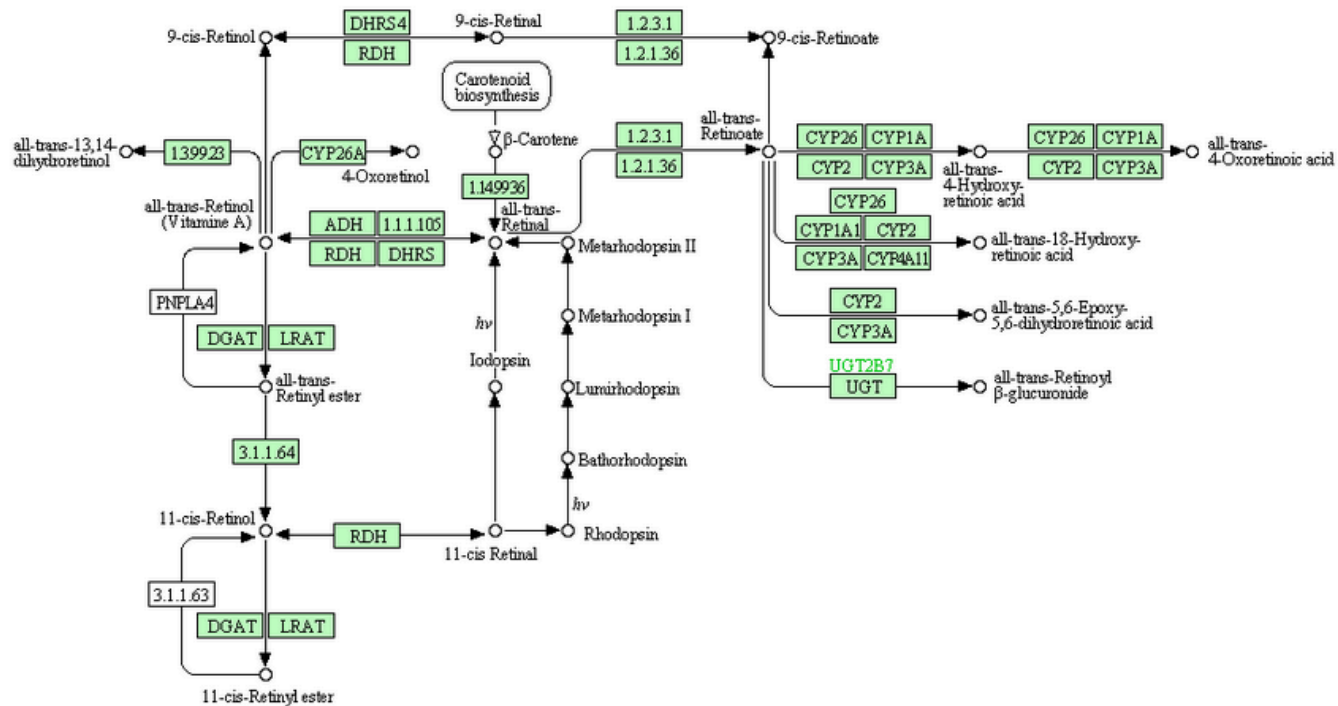
[[Pathway menu](#) | [Organism menu](#) | [Pathway entry](#) | [Download KGML](#) | [User data mapping](#)]

Mus musculus (mouse)

Go

100%

RETINOL METABOLISM IN ANIMALS



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

..

Lookup of GO terms

AmiGO

<http://amigo.geneontology.org>

 *the Gene Ontology*

AmiGO





SearchBrowseBLASTHomolog AnnotationsTools & ResourcesHelp

Search GO

☒ terms☐ genes or proteins☐ exact match

Send

proteolysis


Term information  Term neighborhood  External references  24356 gene product associations 

Term Information

Accession	GO:0006508
Ontology	Biological Process
Synonyms	narrow: ATP-dependent proteolysis exact: peptidolysis
Definition	The hydrolysis of proteins into smaller polypeptides and/or amino acids by cleavage of their peptide bonds. <i>Source:</i> GOC:bf, GOC:mah
Comment	This term was intentionally placed under 'protein metabolic process ; GO:0019538' rather than 'protein catabolic process ; GO:0030163' to cover all processes centered on breaking peptide bonds, including those involved in protein maturation.
Subset	PIR GO slim Prokaryotic GO subset
Community	Add usage comments for this term on the GONUTS wiki.

GO Table View

GO:0006508 Proteolysis

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

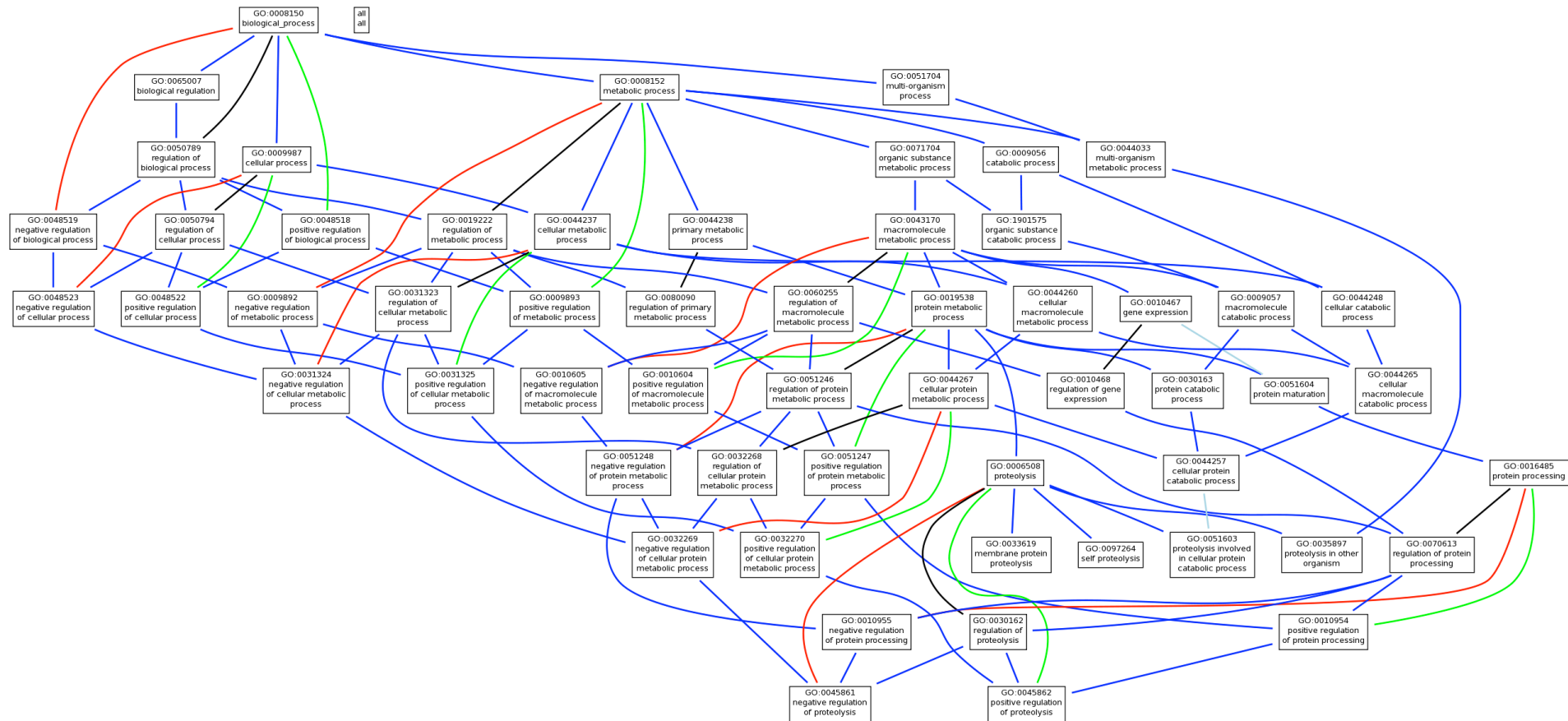
Downloads

Mappings

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

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)

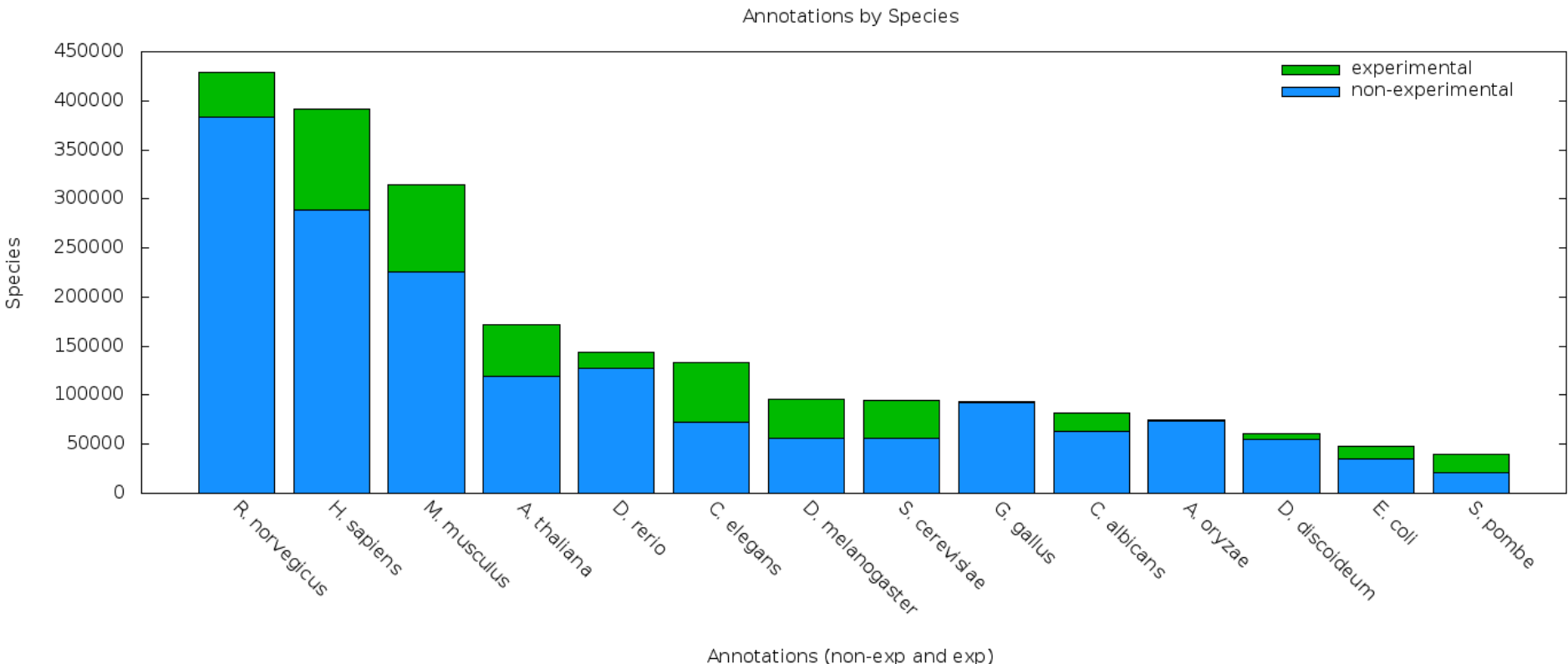
subject	relation	object	annotations
proteolysis	is_a (inferred)	biological process (GO:0008150)	665024
proteolysis	is_a (inferred)	metabolic process (GO:0008152)	368913
proteolysis	is_a (inferred)	organic substance metabolic process (GO:0071704)	300256
proteolysis	is_a (inferred)	macromolecule metabolic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



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

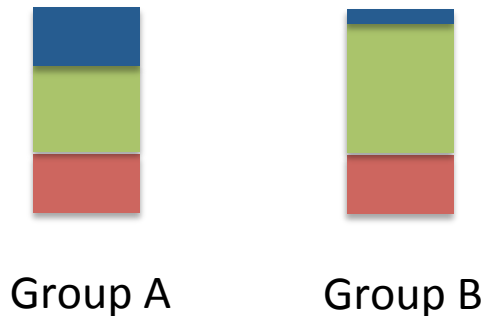
- Bonferroni
- False Discovery Rate (FDR): If we find 100 genes significantly differentially expressed at a 5% FDR, we expect at most 5 false discoveries in the list.

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 independent of regulation)



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

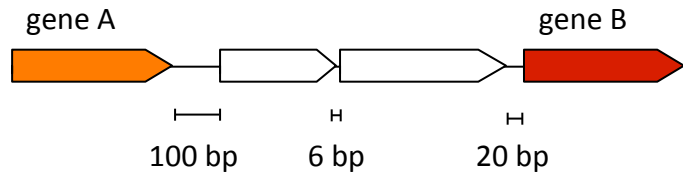
URLs & Tips

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

Fusion



raw score: constant (0.99)

experimental interactions

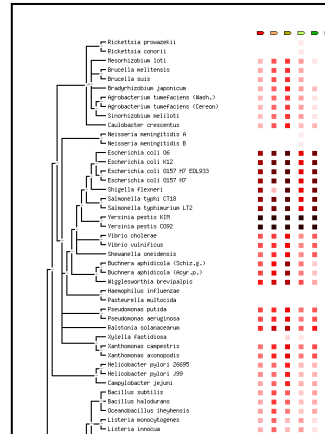
- two-hybrid, TAP, annotated complexes, ...
- topology-based analysis: who with whom, how many other partners?

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



Phylogenetic profiles

- “similarity profiles”
- singular value decomposition

raw score: euclidean distance

filter: downweigh scores for homologous pairs

Textmining

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

raw score: log-odds score