

From Genes to Biology: The Gene Ontology / Pathway enrichment

Biol4230

Tues, April 10, 2018

Bill Pearson wrp@virginia.edu 4-2818 Pinn 6-057

- Gene Ontology (GO)
 - "Ontology" – a directed acyclic graph (DAG)
 - molecular function, biological process, cellular component
 - evidence and evidence codes
 - positives and negatives, missing data
- Function/Pathway enrichment analysis
 - do sets (subsets) of differentially expressed genes reflect a pathway?
 - Over representation analysis (ORA)
 - functional class scoring – GSEA (gene set enrichment analysis)

fasta.bioch.virginia.edu/biol4230

1

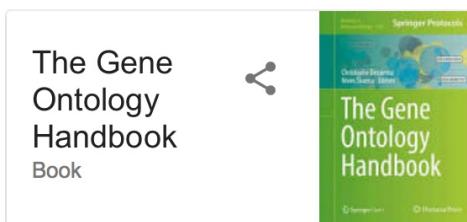
To learn more:

1. Harris, M. A. *et al.* The Gene Ontology (GO) database and informatics resource. *Nucleic Acids Res* **32**, D258–61 (2004).
www.geneontology.org
2. Rhee, S. Y., Wood, V., Dolinski, K. & Draghici, S. Use and misuse of the gene ontology annotations. *Nat Rev Genet* **9**, 509–515 (2008).
3. Nehrt, N. L., Clark, W. T., Radivojac, P. & Hahn, M. W. Testing the ortholog conjecture with comparative functional genomic data from mammals. *PLoS Comput Biol* **7**, e1002073 (2011).
4. Thomas, P. D. *et al.* On the Use of Gene Ontology Annotations to Assess Functional Similarity among Orthologs and Paralogs: A Short Report. *PLoS Comput Biol* **8**, e1002386 (2012).
5. Khatri, P., Sirota, M. & Butte, A. J. Ten years of pathway analysis: current approaches and outstanding challenges. *PLoS Comput Biol* **8**, e1002375 (2012).
6. Rhee, S. Y., Wood, V., Dolinski, K. & Draghici, S. Use and misuse of the gene ontology annotations. *Nat Rev Genet* **9**, 509–515 (2008).
7. Subramanian, A. *et al.* Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles. *Proc Natl Acad Sci U S A* **102**, 15545–15550 (2005).

fasta.bioch.virginia.edu/biol4230

2

To learn more:



This book is open access under a CC BY 4.0 license. This book provides a practical and self-contained overview of the Gene Ontology (GO), the leading project to organize biological knowledge on genes and their products across genomic resources. ... [Google Books](#)

Originally published: November 22, 2016

fasta.bioch.virginia.edu/biol4230

3

Gene Ontology/Enrichment Analysis

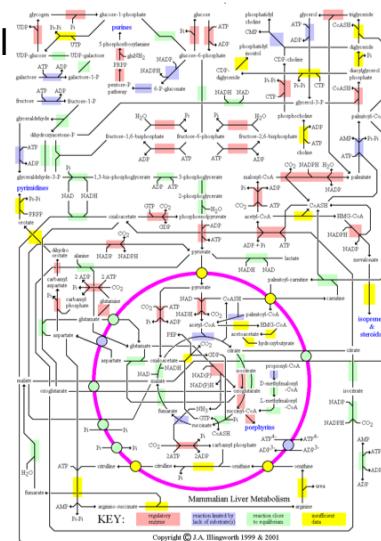
- I have a set of differentially expressed genes – what is happening to the cell?
- Gene Ontology (GO)
 - "Ontology" – a directed acyclic graph (DAG)
 - molecular function, biological process, cellular component
 - evidence and evidence codes
 - positives and negatives, missing data
 - One of many
- Function/Pathway enrichment analysis
 - do sets (subsets) of differentially expressed genes reflect a pathway?
 - Over representation analysis (ORA)
 - functional class scoring – GSEA (gene set enrichment analysis)

fasta.bioch.virginia.edu/biol4230

4

What is happening to the cell?

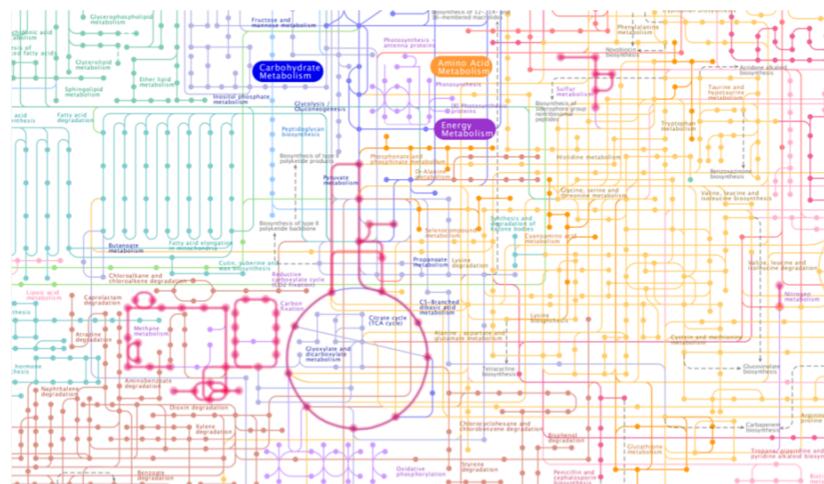
- Cellular functions are chemical
- Fundamental biochemical processes are lined chemical reactions: pathways
 - cell division: DNA replication, mitosis, segregation
 - metabolism: energy, amino-acids, detoxification
 - response to stimuli: signaling
- Some pathways are better understood than others



fasta.bioch.virginia.edu/biol4230

5

KEGG pathways (energy metabolism)



www.genome.jp/kegg/pathway/map/map01100.html

fasta.bioch.virginia.edu/biol4230

6

Differential gene expression on pathways

- Goal: to identify the (known) biological pathways that are activated during biological transitions
 - from stationary/resting to growth phase
 - from normal to cancerous
 - from pluripotent to lineage specific
 - in response to environmental stimuli
- Requirements:
 - list of genes turned on or off / up or down
 - (known) relationships between genes/proteins
 - Gene Ontology
 - shared pathways/processes KEGG/Reactome
- Measure of over-representation
 - hypergeometric (Fisher's Exact test)
 - permutation
 - GSEA

fasta.bioch.virginia.edu/biol4230

7

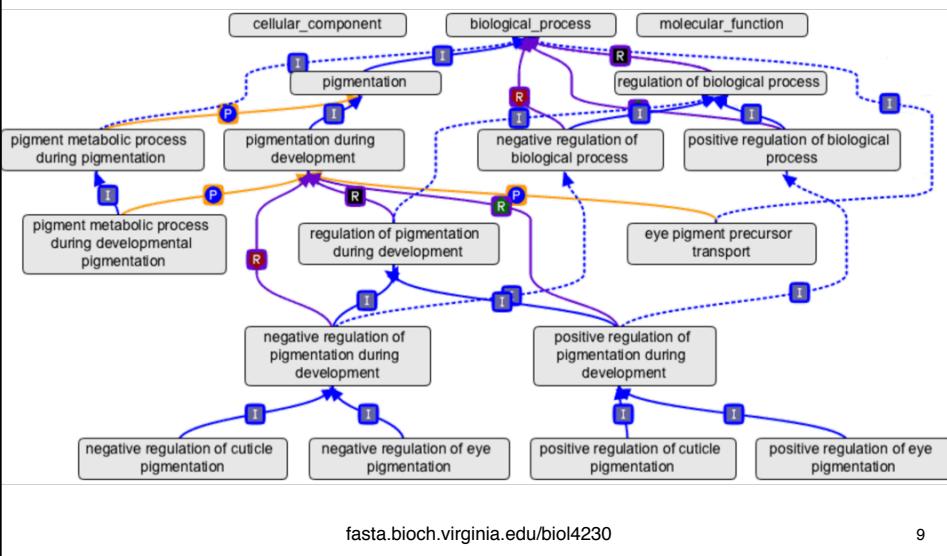
GO: The Gene Ontology (geneontology.org)

- Ontology relationships – Directed Acyclic Graph (DAG) of relationships
 - is-a
 - part-of / has-part
 - regulates / positively-regulates / negatively-regulates
- Hierarchical – three orthogonal hierarchies
 - molecular function
 - biological process
 - cellular location
 - (no sense of time, or developmental stage)
- Curated, with Evidence codes
 - experimental
 - similarity based (but curated)
 - IEA Inferred from Electronic Annotation (no human)
- Absence of activity/process annotation does NOT guarantee absence of activity/process

fasta.bioch.virginia.edu/biol4230

8

The Gene Ontology (GO)



The Gene Ontology (GO) : trees vs. DAGs

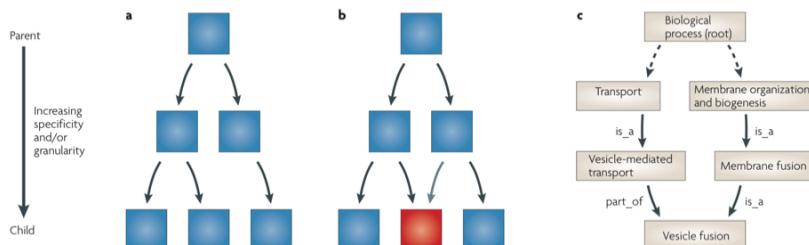


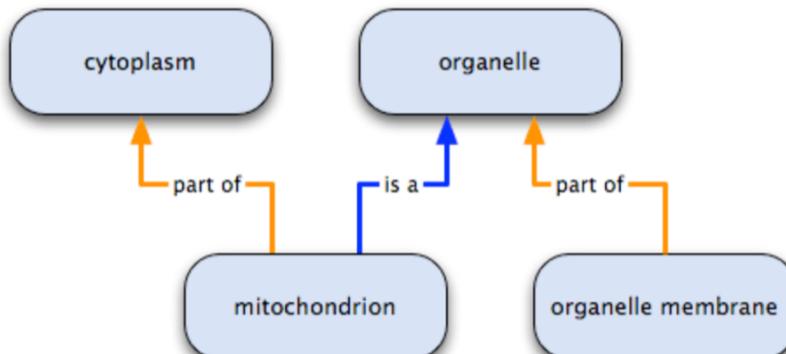
Figure 1 | simple trees versus directed acyclic graphs. Boxes represent nodes and arrows represent edges. **a** An example of a simple tree, in which each child has only one parent and the edges are directed, that is, there is a source (parent) and a destination (child) for each edge. **b** A directed acyclic graph (DAG), in which **each child can have one or more parents**. The node with multiple parents is coloured red and the additional edge is coloured grey. **c** An example of a node, vesicle fusion, in the biological process ontology with multiple parentage. The dashed edges indicate that there are other nodes not shown between the nodes and the root node (biological process). A root is a node with no incoming edges, and at least one leaf (also called a sink). A leaf node is a node with no outgoing edges, that is, a terminal node with no children (vesicle fusion). Similar to a simple tree, A DAG has directed edges and does not have cycles, that is, no path starts and ends at the same node, and will always have at least one root node. The depth of a node is the length of the longest path from the root to that node, whereas the height is the length of the longest path from that node to a leaf 41. *is_a* and *part_of* are types of relationships that link the terms in the GO ontology. More information about the relationships between GO terms are found online (An Introduction to the Gene Ontology).

Rhee, S. Y., et al. *Nat Rev Genet* **9**, 509–515 (2008).

fastabioch.virginia.edu/biol4230

10

Gene Ontology Relationships: is-a, part-of (regulates/up-regulates/down-regulates)

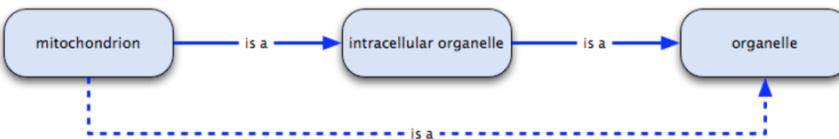


geneontology.org/page/ontology-relations

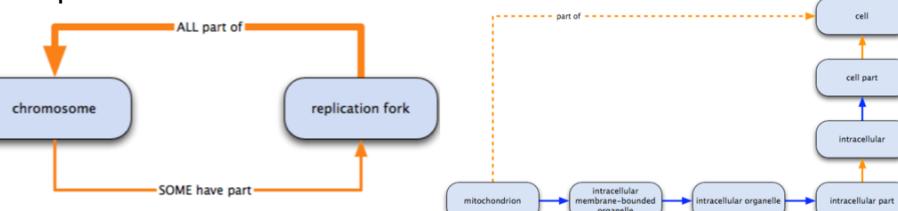
fasta.bioch.virginia.edu/biol4230

11

Gene Ontology Relationships: is-a:



part-of:



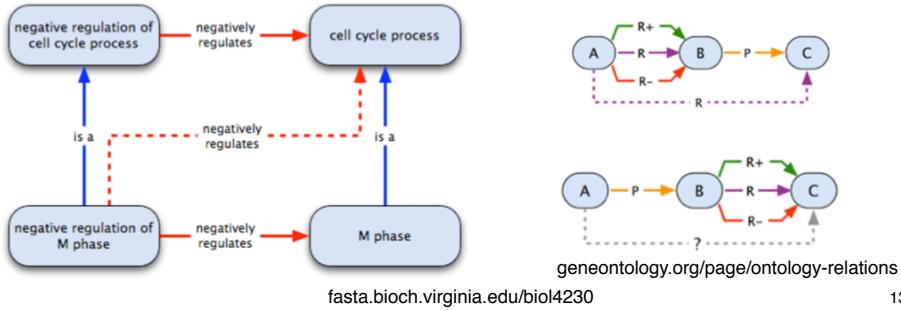
geneontology.org/page/ontology-relations

fasta.bioch.virginia.edu/biol4230

12

Gene Ontology Relationships: is-a, part-of, has-part, regulates

- is-a : identity (synonyms, reversible)
- part-of : sub-set, not reversible
- has-part: converse of part-of, not reversible
- regulates/up-regulates/down-regulates
- can be combined in logically consistent ways



fastabioch.virginia.edu/biol4230

13

Gene Ontology (GO) hierarchies for GSTM1_HUMAN

- Molecular function (chemistry)
 - glutathione binding, GST activity, enzyme binding, homodimerization, detoxification of nitrogen compound
- Biological process (pathway, function)
 - xenobiotic metabolic process, glutathione derivative biosynthetic process, small molecular metabolic process
- Cellular location
 - cytosol, cytoplasm

fastabioch.virginia.edu/biol4230

14

**Gene Ontology (GO) hierarchies for
GSTM1_HUMAN**

molecular function
biological process

amigo.geneontology.org/amigo/gene_product/UniProtKB:P09488

Total: 20; showing 11-20	Results count	10 20									
Gene/product	Gene/product name	Qualifier	Direct annotation	Annotation extension	Source	Taxon	Evidence	Evidence with	PANTHER family	Isoform	Reference
GSTM1	Glutathione S-transferase Mu 1		glutathione transferase activity		UniProtKB	Homo sapiens	IDA		glutathione s-transferase pthr1571		PMID:8373352
GSTM1	Glutathione S-transferase Mu 1		glutathione metabolic process		UniProtKB	Homo sapiens	IDA		glutathione s-transferase pthr1571		PMID:8373352
GSTM1	Glutathione S-transferase Mu 1		glutathione binding		UniProtKB	Homo sapiens	IDA		glutathione s-transferase pthr1571		PMID:8373352
GSTM1	Glutathione S-transferase Mu 1		glutathione transferase activity		UniProtKB	Homo sapiens	TAS		glutathione s-transferase pthr1571		PMID:8403204
GSTM1	Glutathione S-transferase Mu 1		cytosol		UniProtKB	Homo sapiens	TAS		glutathione s-transferase pthr1571		Reactome:REACT_6854
GSTM1	Glutathione S-transferase Mu 1		enzyme binding		UniProtKB	Homo sapiens	IPI	UniProtKB:P09488	glutathione s-transferase pthr1571		PMID:8373352
GSTM1	Glutathione S-transferase Mu 1		protein homodimerization activity		UniProtKB	Homo sapiens	IPI	UniProtKB:P09488	glutathione s-transferase pthr1571		PMID:8373352
GSTM1	Glutathione S-transferase Mu 1		xenobiotic metabolic process		UniProtKB	Homo sapiens	TAS		glutathione s-transferase pthr1571		Reactome:REACT_6959
GSTM1	Glutathione S-transferase Mu 1		glutathione derivative biosynthetic process		UniProtKB	Homo sapiens	TAS		glutathione s-transferase pthr1571		Reactome:REACT_6926
GSTM1	Glutathione S-transferase Mu 1		cellular detoxification of nitrogen compound		UniProtKB	Homo sapiens	IDA		glutathione s-transferase pthr1571		PMID:8373352

fasta.bioch.virginia.edu/biol4230

15

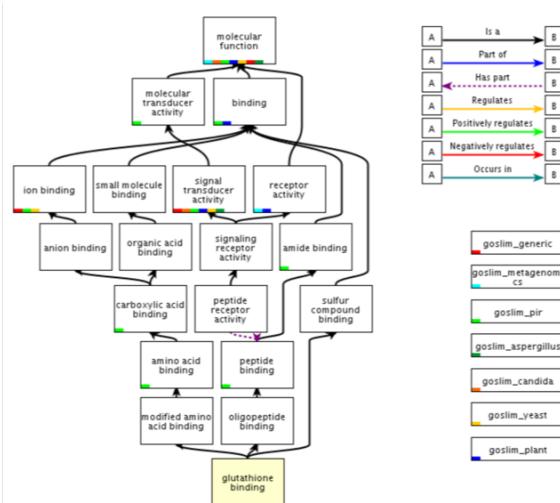
**Gene Ontology (GO) hierarchy for glutathione binding molecular function
(annotating the ontology, NOT a protein)**

Term Information	
Accession	GO:0043295
Name	glutathione binding
Ontology	molecular_function
Synonyms	None
Definition	Interacting selectively and non-covalently with glutathione; a tripeptide composed of the three amino acids cysteine, glutamic acid and glycine. Source: ISBN:0198506732, GOC:bf
Comment	None
History	See term history for GO:0043295 at QuickGO
Subset	None
Community	GN Add usage comments for this term on the GONUTS wiki.
Related	Link to all genes and gene products associated to glutathione binding. Link to all direct and indirect annotations to glutathione binding. Link to all direct and indirect annotations download (limited to first 10,000) for glutathione binding.

fasta.bioch.virginia.edu/biol4230

16

Gene Ontology (GO) hierarchy for glutathione binding molecular function



fasta.bioch.virginia.edu/biol4230

17

Gene Ontology (GO) hierarchy for xenobiotic metabolic process Biological Process

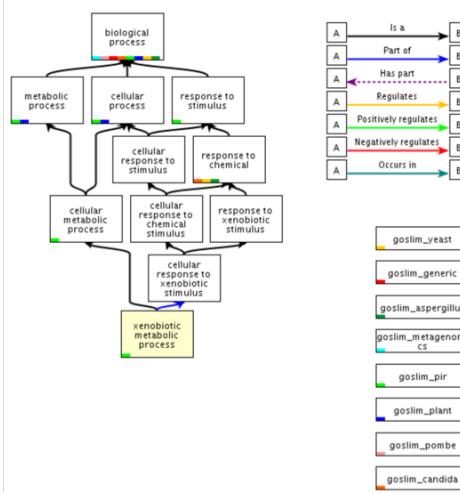
Term Information

Accession	GO:0006805
Name	xenobiotic metabolic process
Ontology	biological_process
Synonyms	xenobiotic metabolism
Definition	The chemical reactions and pathways involving a xenobiotic compound, a compound foreign to living organisms. Used of chemical compounds, e.g. a xenobiotic chemical, such as a pesticide. Source: GOC:cab2
Comment	None
History	See term history for GO:0006805 at QuickGO
Subset	gosubset_prok goslim_pir
Community	GN Add usage comments for this term on the GONUTS wiki.
Related	Link to all genes and gene products associated to xenobiotic metabolic process. Link to all direct and indirect annotations to xenobiotic metabolic process. Link to all direct and indirect annotations download (limited to first 10,000) for xenobiotic metabolic process.

fasta.bioch.virginia.edu/biol4230

18

Gene Ontology (GO) hierarchy for xenobiotic metabolic process Biological Process



QuickGO - <http://www.ebi.ac.uk/QuickGO>

fasta.bioch.virginia.edu/biol4230

19

Gene Ontology (GO) annotations have evidence codes

		molecular function		biological process									
Total: 20; showing 11-20	Results count 10	Gene/product	Gene/product name	Qualifier	Direct annotation	Annotation extension	Source	Taxon	Evidence	Evidence with	PANTHER family	Isoform	Reference
		GSTM1	Glutathione S-transferase Mu 1		glutathione transferase activity		UniProtKB	Homo sapiens	IDA		glutathione s-transferase pthr1571		PMID:8373352
		GSTM1	Glutathione S-transferase Mu 1		glutathione metabolic process		UniProtKB	Homo sapiens	IDA		glutathione s-transferase pthr1571		PMID:8373352
		GSTM1	Glutathione S-transferase Mu 1		glutathione binding		UniProtKB	Homo sapiens	IDA		glutathione s-transferase pthr1571		PMID:8373352
		GSTM1	Glutathione S-transferase Mu 1		glutathione transferase activity		UniProtKB	Homo sapiens	TAS		glutathione s-transferase pthr1571		PMID:8403204
		GSTM1	Glutathione S-transferase Mu 1		catalysis		UniProtKB	Homo sapiens	TAS		glutathione s-transferase pthr1571		Reactome:REACT_6854
		GSTM1	Glutathione S-transferase Mu 1		enzyme binding		UniProtKB	Homo sapiens	IP	UniProtKB:P09488	glutathione s-transferase pthr1571		PMID:8373352
		GSTM1	Glutathione S-transferase Mu 1		protein homodimerization activity		UniProtKB	Homo sapiens	IP	UniProtKB:P09488	glutathione s-transferase pthr1571		PMID:8373352
		GSTM1	Glutathione S-transferase Mu 1		xenobiotic metabolic process		UniProtKB	Homo sapiens	TAS		glutathione s-transferase pthr1571		Reactome:REACT_6959
		GSTM1	Glutathione S-transferase Mu 1		glutathione derivative biosynthetic process		UniProtKB	Homo sapiens	TAS		glutathione s-transferase pthr1571		Reactome:REACT_6926
		GSTM1	Glutathione S-transferase Mu 1		cellular detoxification of nitrogen compound		UniProtKB	Homo sapiens	IDA		glutathione s-transferase pthr1571		PMID:8373352

fasta.bioch.virginia.edu/biol4230

20

Gene Ontology entries have Evidence Codes

geneontology.org/page/guide-go-evidence-codes

Experimental:

- Inferred from Experiment (EXP)
- Inferred from Direct Assay (IDA)
- Inferred from Physical Interaction (IPI)
- Inferred from Mutant Phenotype (IMP)
- Inferred from Genetic Interaction (IGI)
- Inferred from Expression Pattern (IEP)

Literature based:

- Traceable author statement (TAS)

Computational (and someone looked at it)

- Inferred from Sequence or structural Similarity (ISS)
- Inferred from Sequence Orthology (ISO)
- Inferred from Sequence Alignment (ISA)
- Inferred from Sequence Model (ISM)
- Inferred from Genomic Context (IGC)
- Inferred from Biological aspect of Ancestor (IBA)
- Inferred from Biological aspect of Descendant (IBD)
- Inferred from Key Residues (IKR)
- Inferred from Rapid Divergence (IRD)
- Inferred from Reviewed Computational Analysis (RCA)

Computational (no human curation)

- Inferred from Electronic Annotation (IEA)

Evidence codes are **not** statements of the quality of the annotation. Within each evidence code classification, some methods produce annotations of higher confidence or greater specificity than other methods.... Thus evidence codes **cannot** be used as a measure of the quality of the annotation.

fasta.bioch.virginia.edu/biol4230

21

Gene Ontology coverage, by organism (2015)

Species	Source	Genes	Annots	non-IEA	Date
P. falciparum	GeneDB	2373	6250	6250	3/10/2015
E. coli	PortEco	3770	45842	13302	6/26/2014
D. melano.	FlyBase	14646	102825	90887	2/16/2015
B. taurus	GO/EBI	20466	163368	35893	3/31/2015
G. gallus	GO/EBI	12945	101588	15119	3/31/2015
Bos taurus	GO/EBI	17349	141466	33661	3/31/2015
C. lupus	GO/EBI	16016	123620	19392	3/31/2015
Human	GO/EBI	18963	366697	284606	3/31/2015
S. scrofa	GO/EBI	16811	121450	22559	3/31/2015
O. sativa	Gramene	41140	49282	49282	9/22/2009
Microbio	JCVI	56852	142146	142146	3/24/2011
M. musculus	MGI	24177	354620	255070	4/2/2015
R. norvegicus	RGD	26563	416902	255149	4/4/2015
S. pombe	PomBase	5382	39112	34278	03/25/2015
S. cerevisiae	SGD	6379	94252	48762	4/4/2015
A. thaliana	TAIR	30469	230073	184681	3/31/2015
C. elegans	WormBase	20318	134916	67739	9/30/2014
D. rerio	ZFIN	19655	167449	48985	4/6/2015
UniPr, no IEA	GO/EBI	148533	756506	756506	-
UniProt	GO/EBI	29516189	201248286	2114923	-

geneontology.org/page/download-annotations

22

Using GO to test functional conservation: A cautionary tale

NL Nehrt, WT Clark, P Radivojac, MW Hahn (2011) "Testing the ortholog conjecture with comparative functional genomic data from mammals" *PLOS Comp. Biol.* 7:e1002073

A common assumption in comparative genomics is that orthologous genes share greater functional similarity than do paralogous genes (the "ortholog conjecture"). Many methods used to computationally predict protein function are based on this assumption, even though it is largely untested. Here we present the first large-scale test of the ortholog conjecture using comparative functional genomic data from human and mouse. We use the experimentally derived functions of more than 8,900 genes, as well as an independent microarray dataset, to directly assess our ability to predict function using both orthologs and paralogs. Both datasets show that [paralogs are often a much better predictor of function than are orthologs, even at lower sequence identities](#). Among paralogs, those found within the same species are consistently more functionally similar than those found in a different species. We also find that paralogous pairs residing on the same chromosome are more functionally similar than those on different chromosomes, perhaps due to higher levels of interlocus gene conversion between these pairs. In addition to offering implications for the computational prediction of protein function, our results shed light on the relationship between sequence divergence and functional divergence. We conclude that the most important factor in the evolution of function is not amino acid sequence, but rather the cellular context in which proteins act.

fasta.bioch.virginia.edu/biol4230

23

Nehrt et al, (2011) Testing the ortholog conjecture...

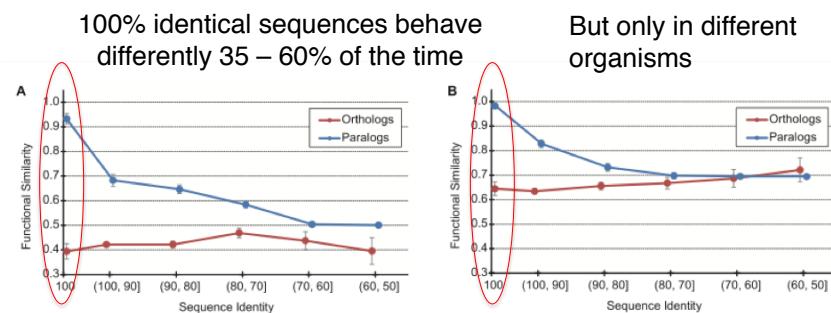


Figure 1. The relationship between functional similarity and sequence identity for human-mouse orthologs (red) and all paralogs (blue). Standard error bars are shown. (A) Biological Process ontology, (B) Molecular Function ontology.
doi:10.1371/journal.pcbi.1002073.g001

PLoS Comput Biol. 2011 7:e1002073.
Testing the ortholog conjecture with
comparative functional genomic data from
mammals. Nehrt NL, et al.

PLoS Comput Biol. 2012 8:e1002386
On the Use of Gene Ontology Annotations
to Assess Functional Similarity among
Orthologs and Paralogs: A Short Report.
P.D. Thomas et al

fasta.bioch.virginia.edu/biol4230

24

On the Use of Gene Ontology Annotations to Assess Functional Similarity among Orthologs and Paralogs: A Short Report

Thomas, et al. PLOS Comp. Biol. (2012) 8:e1002386

A recent paper (Nehrt et al., PLoS Comput. Biol. 7:e1002073, 2011) has proposed a metric for the “functional similarity” between two genes that uses only the Gene Ontology (GO) annotations directly derived from published experimental results. Applying this metric, the authors concluded that paralogous genes within the mouse genome or the human genome are more functionally similar on average than orthologous genes between these genomes, an unexpected result with broad implications if true. We suggest, based on both theoretical and empirical considerations, that this proposed metric should not be interpreted as a functional similarity, and therefore cannot be used to support any conclusions about the “ortholog conjecture” (or, more properly, the “ortholog functional conservation hypothesis”). First, we reexamine the case studies presented by Nehrt et al. as examples of orthologs with divergent functions, and come to a very different conclusion: they actually exemplify how GO annotations for orthologous genes provide complementary information about conserved biological functions. We then show that there is a global ascertainment bias in the experiment-based GO annotations for human and mouse genes: particular types of experiments tend to be performed in different model organisms. We conclude that the reported statistical differences in annotations between pairs of orthologous genes do not reflect differences in biological function, but rather complementarity in experimental approaches. Our results underscore two general considerations for researchers proposing novel types of analysis based on the GO: 1) that GO annotations are often incomplete, potentially in a biased manner, and subject to an “open world assumption” (absence of an annotation does not imply absence of a function), and 2) that conclusions drawn from a novel, large-scale GO analysis should whenever possible be supported by careful, in-depth examination of examples, to help ensure the conclusions have a justifiable biological basis.

fasta.bioch.virginia.edu/biol4230

25

On the Use of Gene Ontology Annotations to Assess Functional Similarity among Orthologs and Paralogs: A Short Report

Thomas, et al. PLOS Comp. Biol. (2012) 8:e1002386

- MAP4K2 (Map kinase kinase kinase kinase)
 - 94% human mouse orthologous sequence identity; 5% orthologous functional similarity
 - functional similarity within mouse paralogs 69%
 - human proteins belong to intracellular protein kinase cascade and protein phosphorylation (kinase activity)
 - mouse Map4K2 only annotated as vesicle targeting
 - both protein are active in the same biological processes, but different processes annotated in different organisms
- Nuclear receptors
 - THRA/ThrA (thyroid hormone receptor) vs. estrogen receptors
 - again, paralogs annotated as more similar, because ligand-specific activities not consistent in human/mouse.
- Absence of activity/process annotation does not guarantee absence of activity/process.

fasta.bioch.virginia.edu/biol4230

26

On the Use of Gene Ontology Annotations to Assess Functional Similarity among Orthologs and Paralogs: A Short Report
Thomas, et al. PLOS Comp. Biol. (2012) 8:e1002386

- Testing the Ortholog Conjecture (Nerht, 2011) is wrong
- By focusing on the "highest quality" annotations (experiment based), Nerht discovered that similar experiments are done in the same organism (human, mouse), but the same experiment is often not done in two different organisms (why duplicate effort?)
- Absence of activity/process annotation does not guarantee absence of activity/process.
- Very few true negative annotations

fasta.bioch.virginia.edu/biol4230

27

GO: The Gene Ontology (geneontology.org)

- Ontology relationships – Directed Acyclic Graph (DAG) of relationships
 - is-a
 - part-of / has-part
 - regulates / positively-regulates / negatively-regulates
- Hierarchical – three orthogonal hierarchies
 - molecular function
 - biological process
 - cellular location
 - (no sense of time, or developmental stage)
- Curated, with Evidence codes
 - experimental
 - similarity based (but curated)
 - IEA Inferred from Electronic Annotation (no human)
- Absence of activity/process annotation does NOT guarantee absence of activity/process

fasta.bioch.virginia.edu/biol4230

28

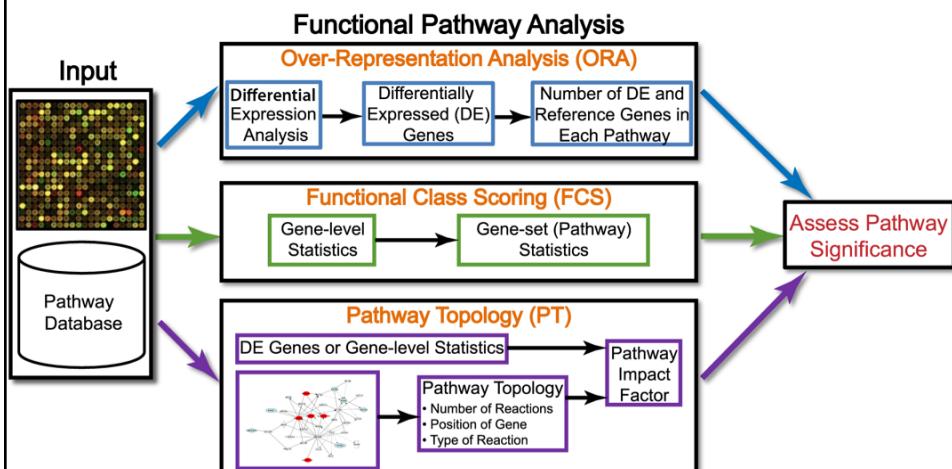
Gene Ontology/Enrichment Analysis

- I have a set of differentially expressed genes – what is happening to the cell?
- Gene Ontology (GO)
 - "Ontology" – a directed acyclic graph (DAG)
 - molecular function, biological process, cellular component
 - evidence and evidence codes
 - positives and negatives, missing data
 - One of many
- Function/Pathway enrichment analysis
 - do sets (subsets) of differentially expressed genes reflect a pathway?
 - Over representation analysis (ORA)
 - functional class scoring – GSEA (gene set enrichment analysis)

fasta.bioch.virginia.edu/biol4230

29

From Genes to Pathways: enrichment analysis



Khatri, et al. *PLoS Comput Biol* 8, e1002375 (2012).

30

Enrichment analysis

- Given a set of differentially expressed (up/down) genes
- And a set of Gene Ontology or Pathway relationships
- Can we use the differentially expressed genes to identify the biological process/pathway involved

fasta.bioch.virginia.edu/biol4230

31

GO/KEGG/PFAM enrichment

- are my 100's of candidates involved in similar process/pathways/functions?
- hypergeometric test for independence:

	difference significant	insignificant difference	total
in group:	k	m-k	m
not in group:	n-k	N+k-n-m	N-m
total:	n	N-n	N

$$P(X = k) = \frac{\binom{m}{k} \binom{N-m}{n-k}}{\binom{N}{n}}$$

$$\binom{a}{b} = \frac{a!}{b!(a-b)!}$$

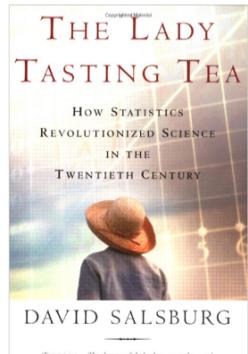
What should 'N' be?

- Total number of genes?
- Number of genes expressed?
- Number of genes up? down?

fasta.bioch.virginia.edu/biol4230

32

The significance of differences: Fisher's Exact Test



1. Around 1930, Muriel Bristol claimed, in a conversation with R. A. Fisher, that she could tell when milk was poured into tea, which was much preferable to tea being poured into milk.
2. Fisher chose to test this hypothesis by preparing 8 cups of tea, 4 tea first, 4 milk first, and asking Ms. Bristol to identify the 4 cups with tea first.
3. If she has no ability to identify milk first/tea first, then one expects her to be right 50% of the time (2 cups). But what if she was right for 3 of the 4 cups?

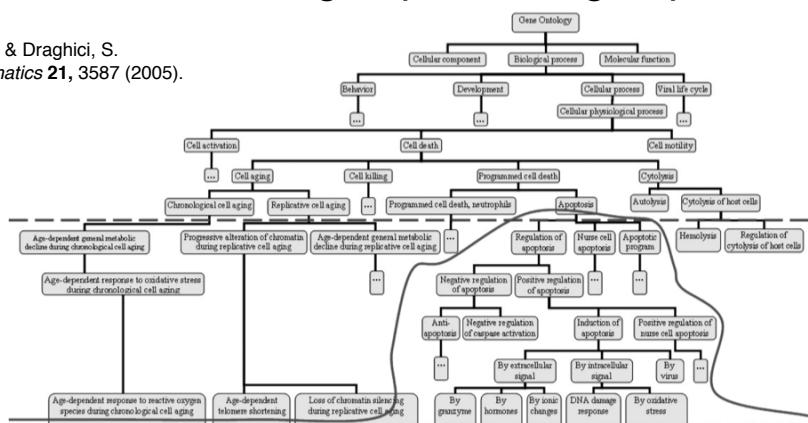
```
> fisher.test(matrix(c(4,0,0,4),nrow=2),
+             alternative='greater')
Fisher's Exact Test for Count Data
data: matrix(c(4, 0, 0, 4), nrow = 2)
p-value = 0.01427
alternative hypothesis: true odds ratio is not equal to 1
```

fasta.bioch.virginia.edu/biol4230

33

Enrichment: In group / Not in group

Khatri, P. & Draghici, S.
Bioinformatics 21, 3587 (2005).



	significant	insignificant	total
in group:	k	m-k	m
not in group:	n-k	N+k-n-m	N-m
total:	n	N-n	N

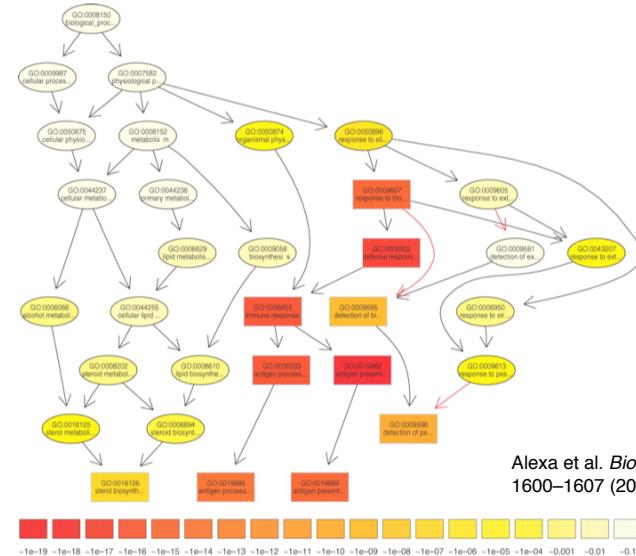
What should 'N' be?

- Total number of genes?
- Number of genes expressed?
- Number of genes up? down?

fasta.bioch.virginia.edu/biol4230

34

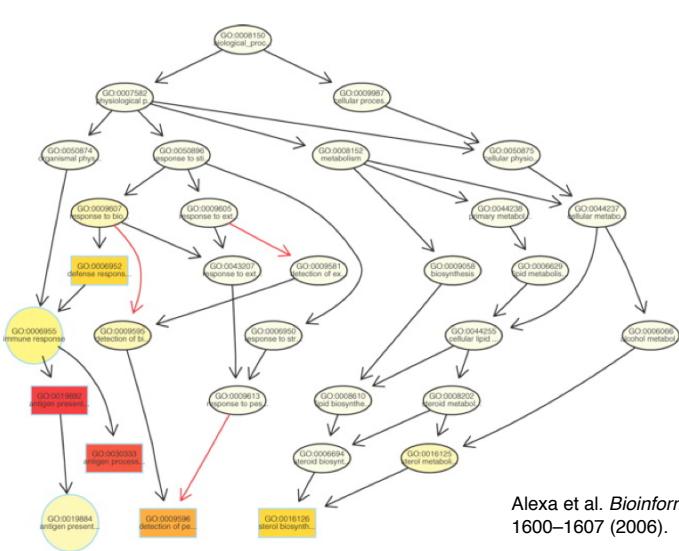
Many levels of GO annotation:



fasta.bioch.virginia.edu/biol4230

35

Correcting for multiple inheritance



fasta.bioch.virginia.edu/biol4230

36

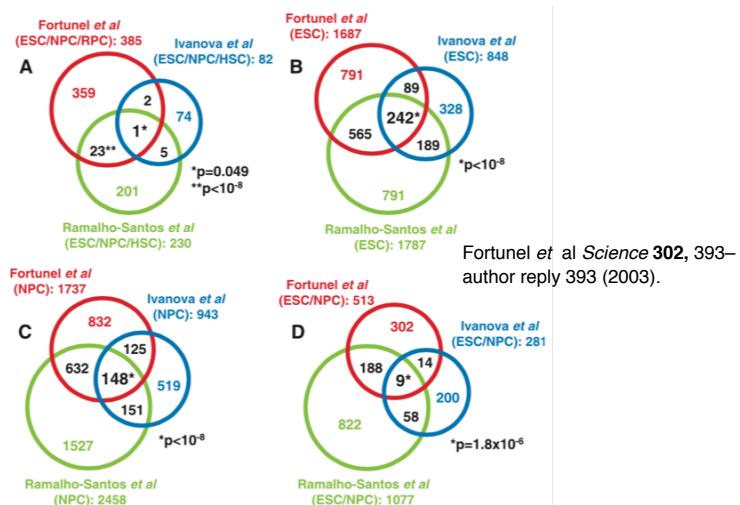
From Genes to Pathways: enrichment analysis

- over-representation analysis (ORA)
 - expected vs.. observed #s of DEGs that share:
 - a GO term
 - a KEGG/Reactome/IPA pathway
 - TF/cis-regulatory promoter elements
 - miRNA targets in 3' UTR
 - disease associations (GWAS, etc)
- hundreds of tools for this, differing by environment, statistics, database, visualization
- one favorite: GOrilla
 - <http://cbl-gorilla.cs.technion.ac.il/>

fasta.bioch.virginia.edu/biol4230

37

Over Representation Analysis - Reproducibility



(A) "Stemness" genes. (B) ESC-enriched genes (C) NPC-enriched genes. (D) Overlap of "stemness" genes—two types of stem cell (ESC/NPC)-enriched genes

fasta.bioch.virginia.edu/biol4230

38

Issues with Over Representation Analysis (ORA)

1. arbitrary significance thresholds for inclusion
2. Differential Expression magnitude/directionality not considered
3. sensitive to choice of background “universe”
 - all genes, genes on chip, or genes with sufficient signal that could possibly be called DEG?
4. correlation between genes ignored
5. correlation/cross-talk between pathways

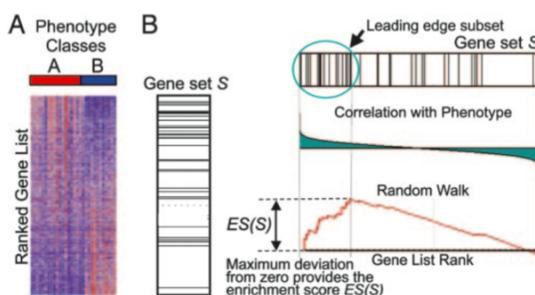
Functional Class Scoring (FCS) methods fix #1-3

fasta.bioch.virginia.edu/biol4230

39

FCS: Gene Set Enrichment Analysis (GSEA)

Given an *a priori* defined set of genes S (e.g., genes encoding products in a metabolic pathway, located in the same cytogenetic band, or sharing the same GO category), the goal of GSEA is to determine whether the members of S are randomly distributed throughout list L or primarily found at the top or bottom.



Subramanian, A. et al. . PNAS
102, 15545–15550 (2005).

- no P value/FDR threshold
- more sensitive than hypergeometric tests
- statistics calculated by permutation testing

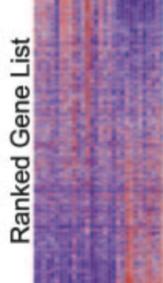
fasta.bioch.virginia.edu/biol4230

40

FCS: Gene Set Enrichment Analysis (GSEA)

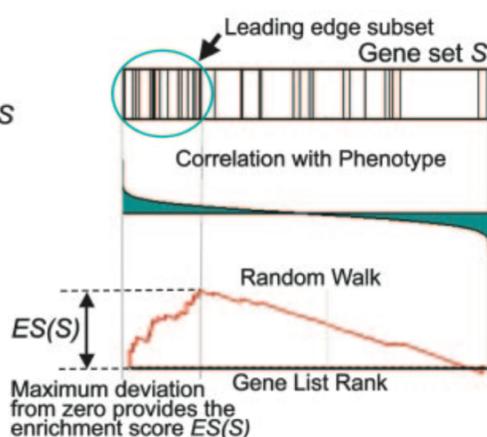
A Phenotype Classes

A B



B

Gene set S



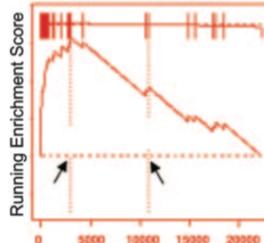
Subramanian, A. et al. . PNAS
102, 15545–15550 (2005).

fasta.bioch.virginia.edu/biol4230

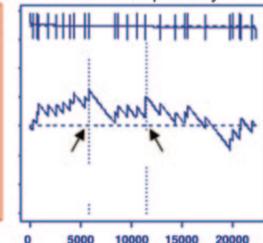
41

FCS: Gene Set Enrichment Analysis (GSEA)

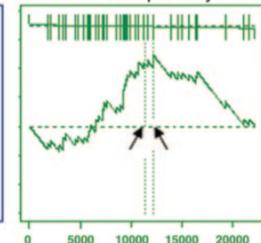
S1: chrX inactive



S2: vitcb pathway



S3: nkt pathway



The distribution of three gene sets, from the C2 functional collection, in the list of genes in the male female lymphoblastoid cell line example ranked by their correlation with gender: S1, a set of chromosome X inactivation genes; S2, a pathway describing vitamin c import into neurons; S3, related to chemokine receptors expressed by T helper cells. Shown are plots of the running sum for the three gene sets: S1 is significantly enriched in females as expected, S2 is randomly distributed and scores poorly, and S3 is not enriched at the top of the list but is nonrandom, so it scores well. Arrows show the location of the maximum enrichment score and the point where the correlation (signal-to-noise ratio) crosses zero. The new method reduces the significance of sets like S3.

Subramanian, A. et al. . PNAS
102, 15545–15550 (2005).

fasta.bioch.virginia.edu/biol4230

42

FCS: Gene Set Enrichment Analysis (GSEA)

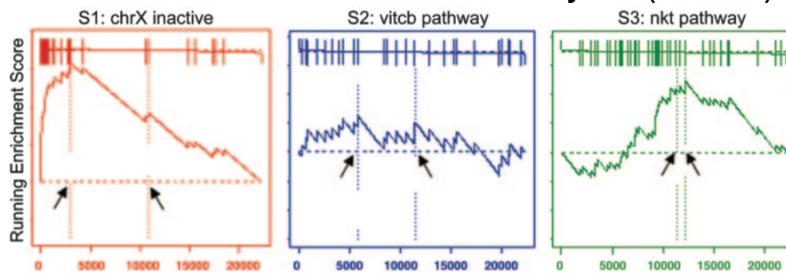


Table 1. P value comparison of gene sets by using original and new methods

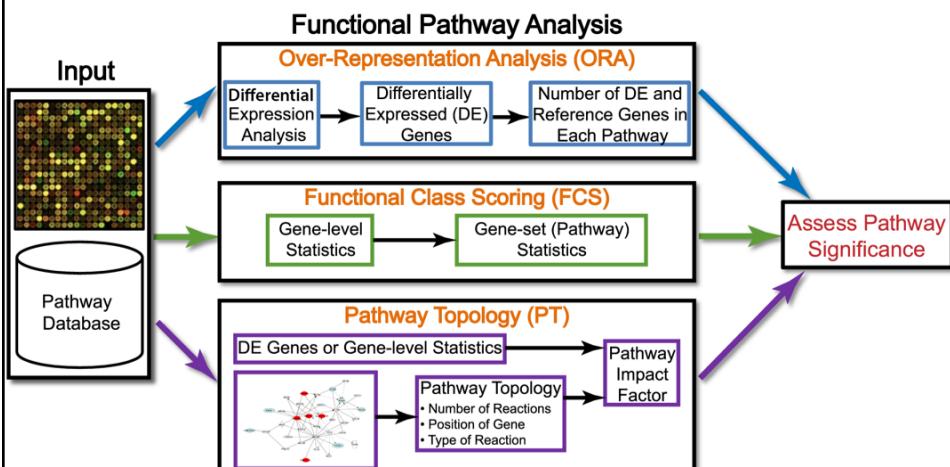
Gene set	Original method nominal P value	New method nominal P value
S1: chrX inactive	0.007	<0.001
S2: vitcb pathway	0.51	0.38
S3: nkt pathway	0.023	0.54

Subramanian, A. et al. . *PNAS*
102, 15545–15550 (2005).

fasta.bioch.virginia.edu/biol4230

43

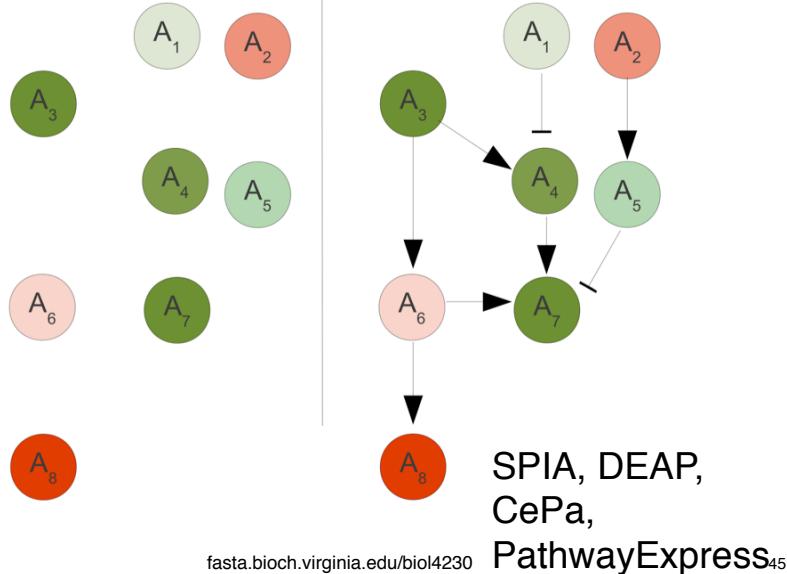
from genes to pathways: enrichment analysis



fasta.bioch.virginia.edu/biol4230

44

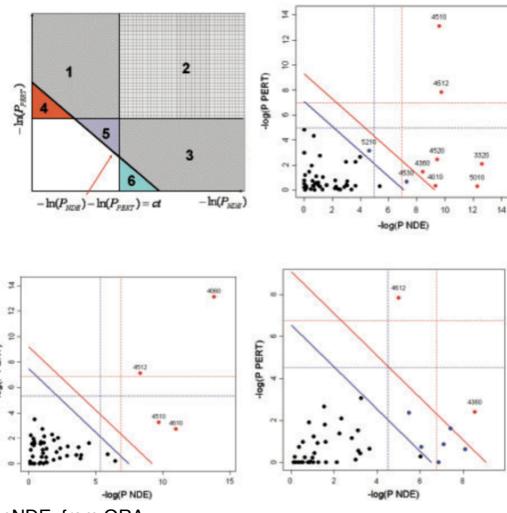
Pathway Topology: PT vs ORA set enrichment vs. pathway impact



SPIA – Signaling Pathway Impact Analysis

The X-axis shows the over-representation evidence, while the Y-axis shows the perturbation evidence. In the top-left plot, areas 2, 3 and 6 together will include pathways that meet the over-representation criterion ($PNDE <\alpha$). Areas 1, 2 and 4 together will include pathways that meet the perturbation criterion ($PPERT <\alpha$). Areas 1, 2, 3 and 5 will include the pathways that meet the combined SPIA criteria ($PG <\alpha$). Note how SPIA results are different from a mere logical operation between the two criteria (OR would be areas 1, 2, 3, 4 and 6; AND would be area 2).

Pathway analysis results on the Colorectal cancer (top right), LaborC (bottom left) and Vessels (bottom right) datasets. Each pathway is represented by a point. Pathways above the oblique red line are significant at 5% after Bonferroni correction, while those above the oblique blue line are significant at 5% after FDR correction. The vertical and horizontal thresholds represent the same corrections for the two types of evidence considered individually. Note that for the Colorectal cancer dataset (top right), the colorectal cancer pathway (ID = 5210) is only significant according to the combined evidence but not so according to any individual evidence PNDE or PPERT.



pNDE: from ORA
pPERT: from perturbation

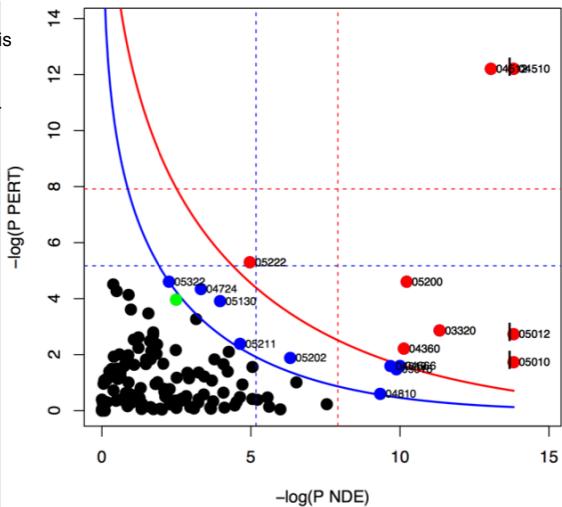
Tarca et al. *Bioinformatics* 25,
75–82 (2009).

SPIA – Signaling Pathway Impact Analysis

Figure 3: SPIA evidence plot for the colorectal cancer dataset. Each pathway is represented by one dot. The pathways at the right of the red curve are significant after Bonferroni correction of the global p-values, p_G , obtained by combining the pPERT and pNDE using the normal inversion method. The pathways at the right of the blue curve line are significant after a FDR correction of the global p-values, p_G .

The green dot shows the KEGG:05210 colon cancer pathway. This pathway is marginally significant ($\text{FDR} < 0.05$) with "normal inversion" combination of PERT and NDE, but not significant with Fisher's method.

pNDE: from ORA
pPERT: from perturbation



<http://www.bioconductor.org/packages/release/bioc/vignettes/SPIA/inst/doc/SPIA.pdf>

fasta.bioch.virginia.edu/biol4230

47

pathway crosstalk yields false positives:

A

rank	pathway	p(fdr)
1	Parkinson's disease	2.0e-06
2	Alzheimer's disease	3.6e-06
3	Huntington's disease	3.4e-05
4	Leishmaniasis	0.0003
5	Phagosome	0.0006
6	Cell cycle	0.0011
7	Oocyte meiosis	0.0016
8	Cardiac muscle contraction	0.0016
9	Toll-like receptor	0.0018
10	PPAR signaling pathway	0.0018
11	Chemokine signaling pathway	0.0154
12	Lysosome	0.0211
13	B cell receptor	0.0252
14	Systemic lupus erythematosus	0.0252
15	Compl. and coag. cascades	0.0342
16	Cytokine-cytokine rec. inter.	0.0346
17	Chagas disease	0.0466
18	Progest. med. oocyte matur.	0.0530
19	Fe epsilon RI signaling pathway	0.0548
20	Leukocyte transendoth. migr.	0.0548

B

rank	pathway	p(fdr)
1	Mitochondrial Activity	8.1e-10
2	Phagosome	9.3e-09
3	Cellcycl+Oocytome	5.8e-08
4	PPAR signaling pathway	0.001
5	Compl. C.C.+Systemic L.E.	0.002
6	* Cytok.-cytok. rec. int.	0.043
7	Toll-like receptor signaling	0.051
8	MAPK signaling pathway	0.115
9	B-cell receptor signaling	0.145
10	Lysosome	0.187
11	Nat. killer cell med. cytox.	0.187
12	* Cell cycle	0.229
13	Calcium signaling pathway	0.229
14	Cell adhesion molecules	0.258
15	NOD-like receptor signaling	0.258
16	Vasc. smooth muscle contr.	0.424
17	Dilated cardiomyopathy	0.424
18	* Oocyte meiosis	0.432
19	Type I diabetes mellitus	0.432
20	Wnt signaling pathway	0.476

The results of the ORA analysis in the fat remodeling experiment for the comparison between days 3 and 0, before (A) and after (B) correction for crosstalk effects. All P-values are FDR corrected. The lines show the significance thresholds: (blue) 0.01, (yellow) 0.05. Pathways highlighted in red represent pathways not related to the phenomenon in analysis, while pathways highlighted in green are those for which we know, with reasonable confidence, are involved in the given phenomenon. The white background indicates pathways for which we do not have conclusive information on their involvement (or lack of) with the phenomenon in analysis. (A) The top 20 pathways resulting from classical ORA before correction for crosstalk. The top four pathways are not related to fat remodeling. (B) The top 20 pathways after correction for crosstalk. Pathways ranked 1, 3, and 5 are modules that are functioning independently of the rest of their pathways in this particular condition. Starred pathways are pathways edited by removing such modules. Note the lack of any obvious false positive above the significance threshold(s).

Donato, M. et al. *Genome Res* 23, 1885–1893 (2013)

fasta.bioch.virginia.edu/biol4230

48

Functional analysis: ORA, FC, PT

- Methods assume independence, but pathways and GO DAGs are anything but independent
 - statistics may be too generous (false positives)
 - statistics may be too strict (false negatives)
- What is the right control?
 - try different approaches?
 - compare to other published datasets?
 - do "positive control" on well understood pathways
- All methods need experimental confirmation
 - find a drug that blocks the pathway
 - ablate a gene (or genes) in the pathway

Function/Pathway Enrichment

- Function/Pathway enrichment analysis
 - do sets (subsets) of differentially expressed genes reflect a pathway?
- Over Representation Analysis (ORA)
 - Fisher exact test, hypergeometric
 - competitive vs. self-contained tests
- Functional Class Scoring (FTS)
 - GSEA : Gene Set Enrichment Analysis
- Pathway Topology (PT)
 - SPIA : Signaling Pathway Impact Analysis
- What are the right "controls"?

Gene Ontology/Enrichment Analysis

- Biological function in the cell: Pathways (chemistry)
- Gene Ontology (GO)
 - "Ontology" – a directed acyclic graph (DAG)
 - molecular function, biological process, cellular component
 - evidence and evidence codes
 - positives and negatives, missing data
 - One of many
- Function/Pathway enrichment analysis
 - do sets (subsets) of differentially expressed genes reflect a pathway?