VEuPathDB.org
Eukaryotic Pathogen, Vector & Host Informatics Resources

Enrichment Analysis Tools in VEuPathDB

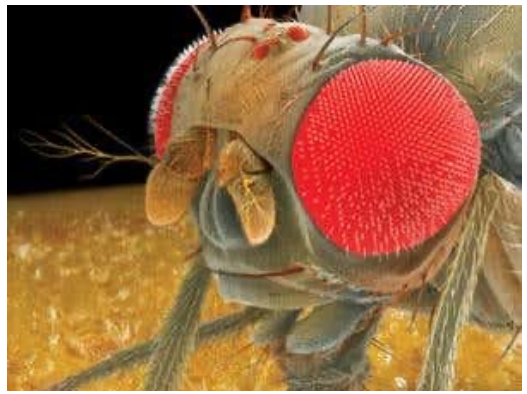
Stuart Brown
University of Pennsylvania

Evelina Basenko
University of Liverpool

Webinar Series: https://veupathdb.org/veupathdb/app/static-content/webinars.html

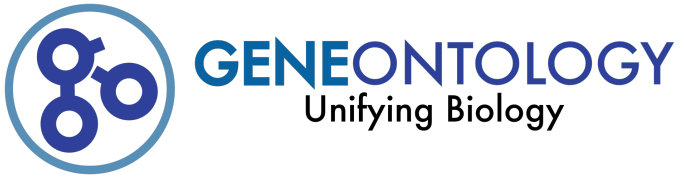
A Factual Resource for Gene Function

- Experimental knowledge obtained in one organism is often applicable to other organisms
- But, biology is complex, and functional information about genes is often difficult to transfer across organisms
- A small number of genetic “model organisms” including mouse, fruitfly, yeast, nematode, and *E. coli* have been extensively studied, often with mutations in every single gene.
- However, it is necessary to make an abstraction – to identify the general **biological process** involved in gene function, rather than some very organism-specific phenotype like pink eyes, number of bristles on the thorax, or a particular swimming behavior.



?

The Gene Ontology Consortium (GO)



"The mission of the GO Consortium is to develop an up-to-date, comprehensive, model of biological systems, from the molecular level to larger pathways, cellular and organism-level systems. The GO knowledgebase is the world's largest source of information on the functions of genes. This knowledge is a foundation for computational analysis of large-scale molecular biology and genetics experiments in biomedical research.

"The Gene Ontology (GO) consortium, began in **1998** when researchers studying the genome of three model organisms — *Drosophila melanogaster* (fruit fly), *Mus musculus* (mouse), and *Saccharomyces cerevisiae* (brewer's or baker's yeast) — agreed to work collaboratively on **a common classification scheme for gene function**."

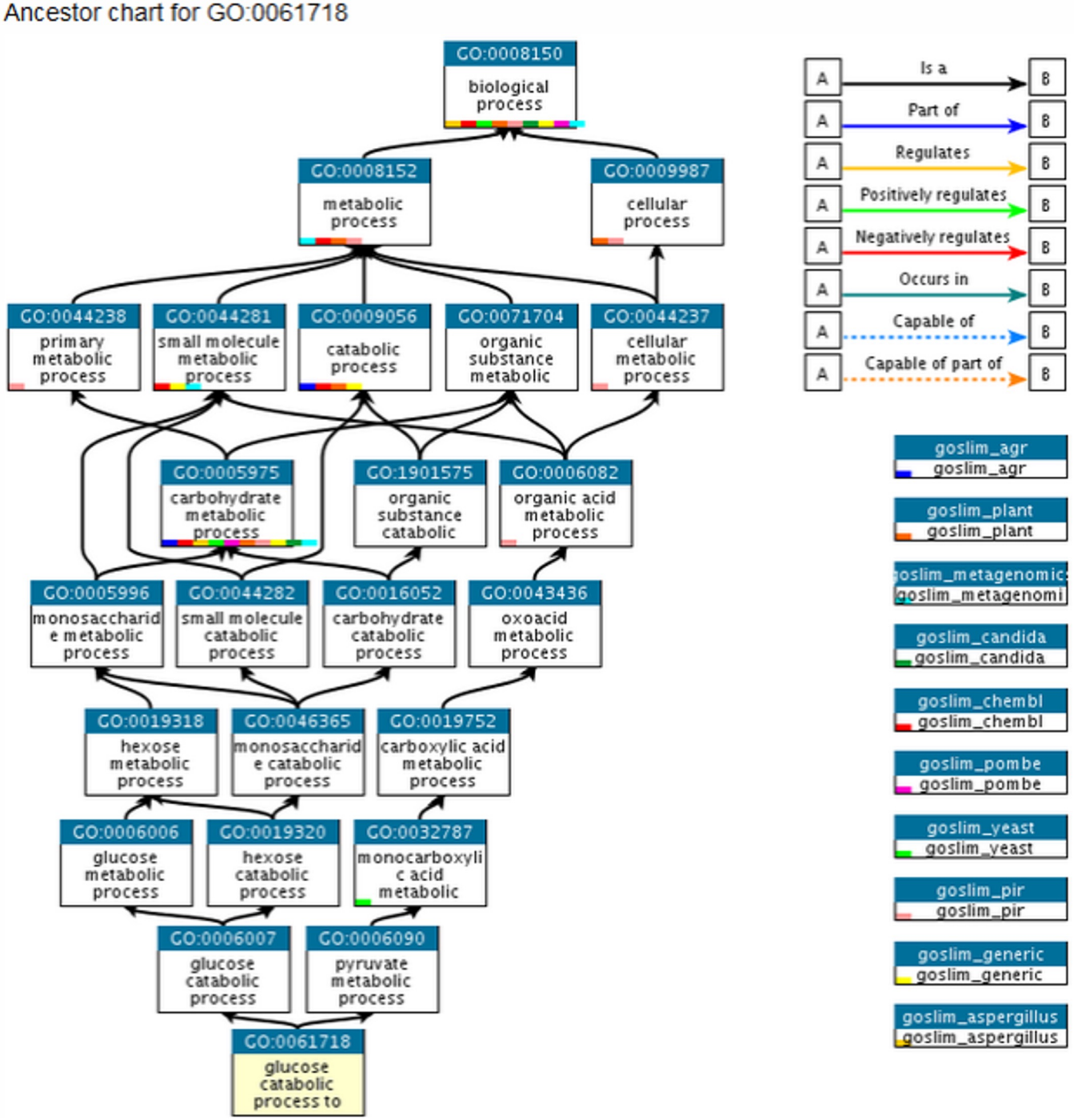
The GO now includes experimental evidence from several model organism databases including: Flybase, Mouse Genome Database (MGI), Rat Genome Database (RGD), Saccharomyces Genome Database(SGD), WormBase, Xenbase, and Zebrafish Information Network(ZFIN).

Ontology

- A defined vocabulary of clearly defined terms with a system of categories and relationships
- A method to name biological processes uniquely, systematically, and consistently.

- Ontologies are generally structured as a hierarchy from more general to more specific terms.
- This is particularly helpful in biology where we can define the function of a protein in a stepwise fashion:
  - Membrane protein > Transporter > Ion transporter > Calcium transporter > Voltage gated calcium transporter

An ontology for bio-process: glucose catabolism to pyruvate



https://www.ebi.ac.uk/QuickGO/term/GO:0061718

Slide created by Peter D'Eustachio, Reactome

3 GO Ontologies

- The GO is composed of 3 different ontologies

- **Bioprocess** describes processes like energy metabolism, reaction to stimuli, DNA repair, etc.

- **Molecular function** maps to steps in metabolic pathways, but also generalizes to activities of a protein such as “catalysis” or “transport”.
  - GO molecular function terms do not specify where, when, or in what context the action takes place.

- **Cellular Component** refers to a location in the cell such as membrane, nucleus, mitochondrion, ribosome, etc.

GO is based on experimental facts

- Every GO term annotated to a gene is based on traceable, evidence-based statements relating a specific gene product to specific terms to describe its biological role.

- Currently, the GO includes experimental findings from over 150,000 published papers, represented as over 700,000 experimentally-supported annotations.
- An additional 6 million annotations are computationally inferred – generally by protein sequence similarity.

Gene Lists

- Many genomics experiments produce a list of genes as a result
  - Differential gene expression
  - Genes (orthologs) that are restricted to a particular clade
  - Transcription factor binding (ChIPseq)
  - Protein-protein interaction
  - Targets of regulators
  - SNPs (mutations) associated with disease
  - SNPs that differentiate two strains or two species with different phenotypes

- How do we identify *INTERESTING* (rather than random) properties of a list of genes produced by an experiment or analysis?
    - The answer is: Gene Function Enrichment Analysis

Simple math for Enrichment

$$\text{Enrichment score} = \frac{\text{Fraction of genes in } \mathbf{list} \text{ with Function A}}{\text{Fraction of genes in } \mathbf{genome} \text{ with Function A}}$$

And then a statistical test to determine if this level of enrichment is "significantly" greater than you might expect by chance

[Note that it is necessary that all the genes in the list come from a single genome that contains GO annotations]

Fisher's Exact Test

"Use Fisher's exact test of independence when you have two nominal variables and you want to see whether the proportions of one variable are different depending on the value of the other variable. Use it when the sample size is small."

[Handbook of Biological Statistics, John H. McDonald]

In the case of gene enrichment, we are testing if the fraction of genes with Function A in a gene list is different than the fraction in the whole genome

If 6/10 genes in a list have Function A, how often (how unlikely is it) that 6/10 are found for Function A when 10 genes are randomly drawn from the genome?

The null hypothesis is that the relative proportions are NOT different.

Note: this is equivalent to a hypergeometric test

$$p = 1 - \sum_{i=0}^{k-1} \frac{\binom{M}{i}\binom{N-M}{n-i}}{\binom{N}{n}}$$

Poker analogy

- If you are dealt 5 cards from a standard deck…
- How likely are you to get 2 Aces?

- Your hand has function “A” for 2/5 cards = 0.4
- The deck has function “A” for 4/52 cards = 0.0769

=Enrichment = 5.2 fold

But what is the exact probability to get this hand?
You have to compute all possible hands: all possible combinations of 52 cards, and then count all combinations that contain 2 Aces.

Formula for Fisher's Exact Test

(for those who like formulas)

$$p = ((a+b)!(c+d)!(a+c)!(b+d)!)/a!b!c!d!N!$$

Where 'a,' 'b,' 'c' and 'd' are the individual frequencies and and 'N' is the total frequency.

Given a gene list from an experiment and a list of all genes with a given function from a genome.

a = genes that are provided in both lists *(genes with function A in your list)*
b = genes that are provided in the first list and not the second *(genes in list not funct. A)*
c = genes that are provided in the second list and not the first *(genes of funct A not in list)*
d = genes that are not provided in either list *(the rest of the genome)*

https://en.wikipedia.org/wiki/Fisher%27s_exact_test

Multiple testing

- If you make the enrichment test for a large number of different functions, you are likely to hit one that seems significant just by chance
  - This is especially true for functions that are rare in the genome

- Statistical tests must be corrected for multiple testing either by Bonferonni (very strict) or Benjamini-Hochberg false discovery rate (FDR) (a bit less strict). These give an “adjusted p-value”, which will always be less significant than the original p-value.

GO SLIM

- To reduce the number of functions being tested (and the multiple testing penalty), GO has many subsets known as GO SLIM sets.
- These represent targeted groups of functions, often for a specific organism
- The SLIM set may avoid the lowest, most specific terms
- A SLIM set may focus on just one area of function or metabolism
- Anyone can make a SLIM subset of terms for their own use
  - This works in any statistical test just like the regular GO, it just limits the number of different terms that will be tested

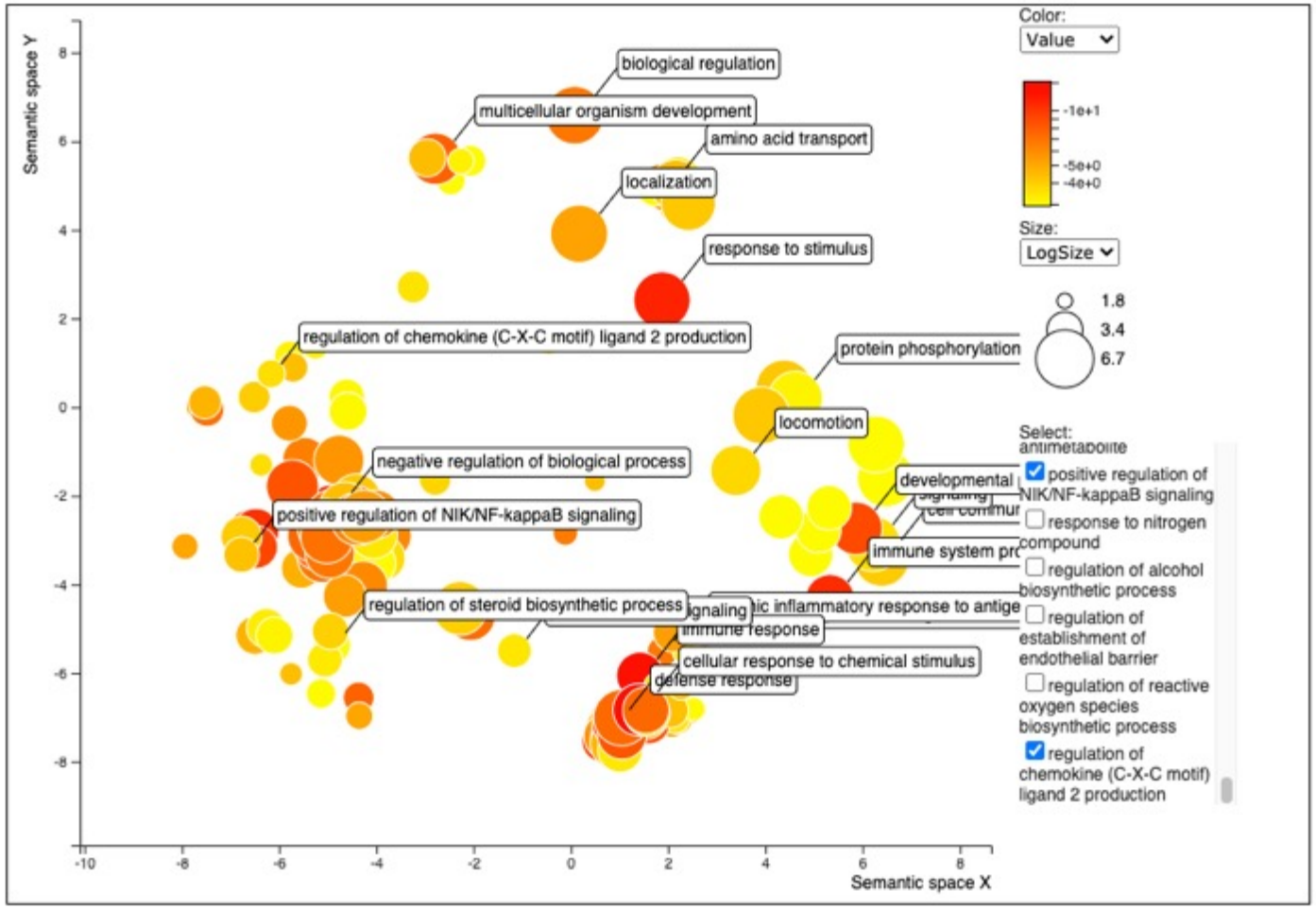
Subset name	Maintainer	File name	OBO format	OWL format	json format
GO slim AGR subset	Developed by GO Consortium for the Alliance of Genomes Resources	goslim_agr	obo	owl	json
Generic GO subset	GO Consortium	goslim_generic	obo	owl	json
***Aspergillus* subset**	*Aspergillus* Genome Data	goslim_aspergillus	obo	owl	json
***Candida albicans* subset**	*Candida* Genome Database	goslim_candida	obo	owl	json
***Drosophila* subset**	FlyBase	goslim_drosophila	obo	owl	json
Chembl Drug Target subset	ChEMBL	goslim_chembl	obo	owl	json
Metagenomics subset	InterPro group	goslim_metagenomic	obo	owl	json
Mouse GO slim	Mouse Genome Informatics	goslim_mouse	obo	owl	json
Plant subset	The *Arabidopsis* Information Resource	goslim_plant	obo	owl	json
Protein Information Resource subset	PIR	goslim_pir	obo	owl	json
***Schizosaccharomyces pombe* subset**	PomBase	goslim_pombe	obo	owl	json
Yeast subset	*Saccharomyces* Genome Database	goslim_yeast	obo	owl	json

Interpret the enrichment results

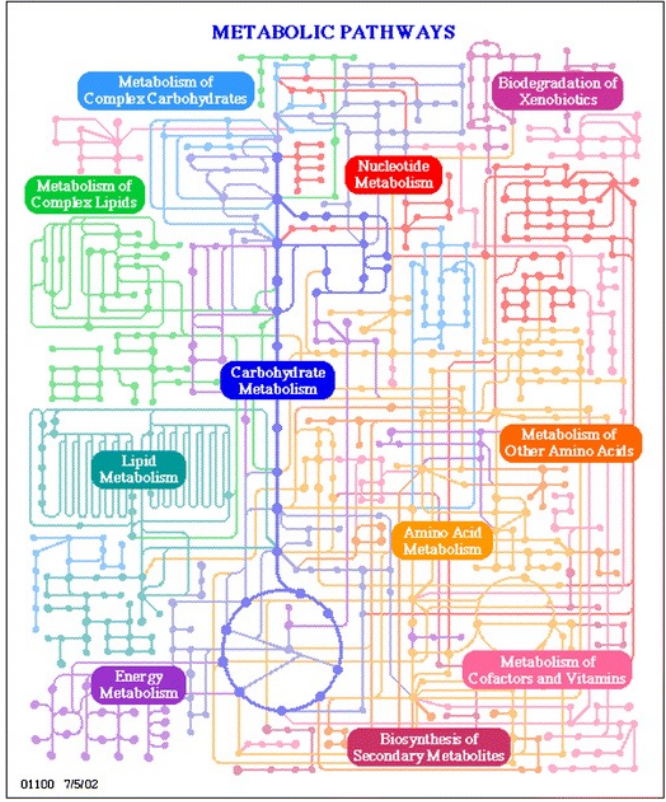
- Making an enrichment test for a gene list is now a fairly simple set of button clicks
- However, understanding the meaning of the result requires domain expertise
- Is this set of functions relevant to the process that constructed the list?
- Are the p-values strong enough to support an argument for the importance of a function in this experiment?
  - Can a larger or smaller gene list be used? – perhaps at a higher or lower fold change
  - Can a more restricted GO SLIM be used to emphasize relevant functions
- Sometimes the “enriched” terms are more general or more specific than the terms used in the community specializing in an organism

Cluster terms by meaning

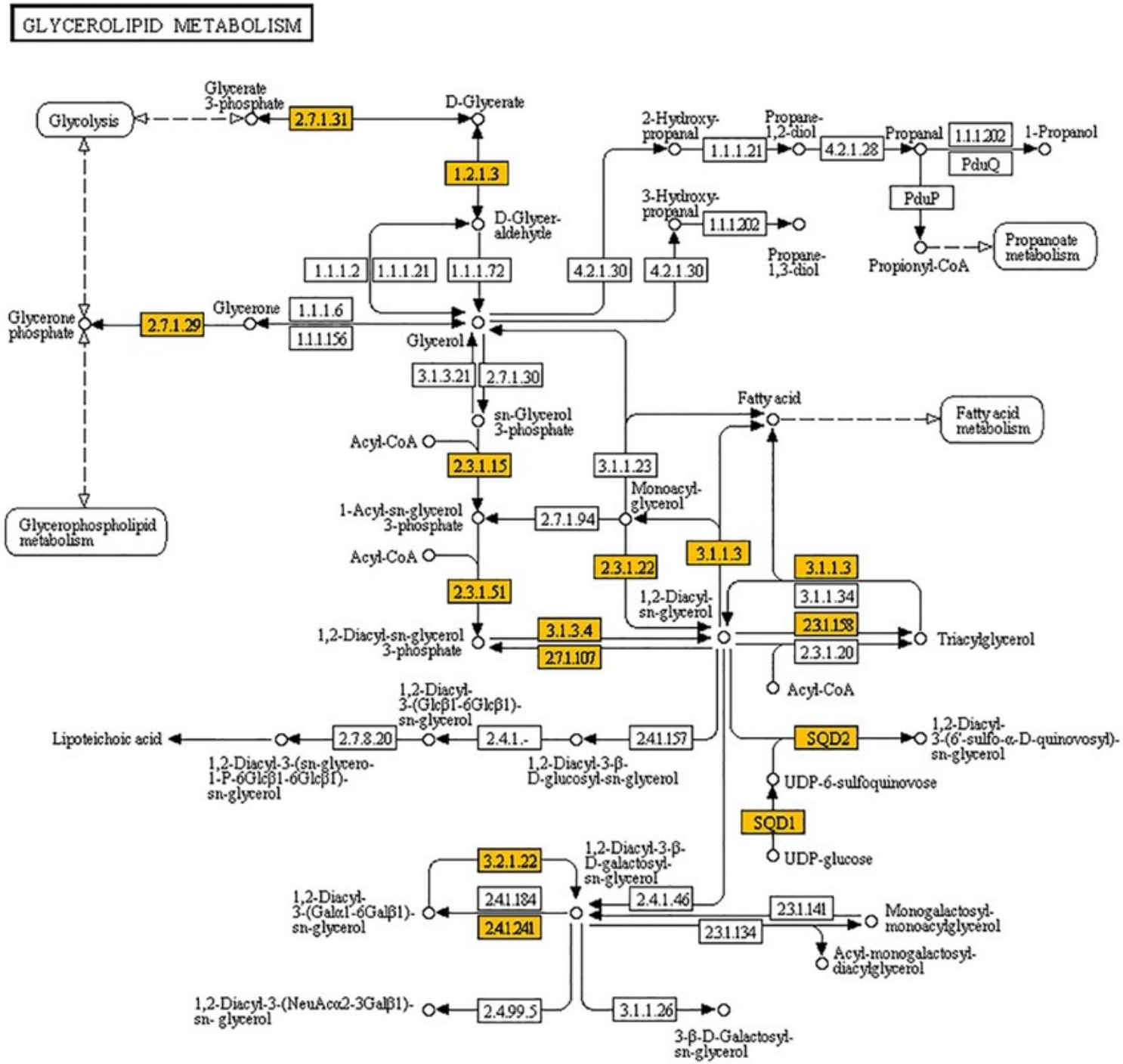
- REVIGO makes a “word cloud” clustering of a group of ontology terms discovered by an enrichment test.
- A cluster of related terms suggests an important process.
- Some of the terms in a cluster may have stronger p-values than others, which makes the whole cluster more interesting



Metabolic Pathways



- Gene lists can also be examined for enrichment of terms from metabolic pathways (such as KEGG, Reactome, BioCyc)
- Each pathway can be treated as a function term, so all the genes/proteins in that pathway form a gene set.
- Enrichment is again calculated as the proportion of genes in a list from a given pathway compared to the proportion in that pathway for the whole genome.
- When an enriched pathway is found, a nice figure can be generated highlighting the genes in the list on the pathway map.


GLYCEROLIPID METABOLISM
Glycolysis
Glycerate 3-phosphate
2.7.1.31
D-Glycerate
1.2.1.3
D-Glycer-aldehyde
1.1.1.2
1.1.1.21
1.1.1.72
2-Hydroxy-propanal
1.1.1.21
Propane-1,2-diol
4.2.1.28
Propanal
1.1.1.202
1-Propanol
PduQ
PduP
3-Hydroxy-propanal
1.1.1.202
Propane-1,3-diol
4.2.1.30
4.2.1.30
Propionyl-CoA
Propanoate metabolism
Glycerone phosphate
2.7.1.29
Glycerone
1.1.1.6
1.1.1.156
Glycerol
3.1.3.21
2.7.1.30
sn-Glycerol 3-phosphate
Acyl-CoA
2.3.1.15
1-Acyl-sn-glycerol 3-phosphate
2.7.1.94
Monoacyl-glycerol
3.1.1.23
Fatty acid
Fatty acid metabolism
Glycerophospholipid metabolism
Acyl-CoA
2.3.1.22
3.1.1.3
2.3.1.51
1,2-Diacyl-sn-glycerol
1,2-Diacyl-sn-glycerol 3-phosphate
3.1.3.4
2.7.1.107
3.1.1.3
3.1.1.34
2.3.1.158
2.3.1.20
Triacylglycerol
Acyl-CoA
1,2-Diacyl-3-(Glcβ1-6Glcβ1)-sn-glycerol
Lipoteichoic acid
2.7.8.20
2.4.1.-
2.4.1.157
1,2-Diacyl-3-(sn-glycero-1-P-6Glcβ1-6Glcβ1)-sn-glycerol
1,2-Diacyl-3-β-D-glucosyl-sn-glycerol
SQD2
1,2-Diacyl-3-(6'-sulfo-α-D-quinovosyl)-sn-glycerol
UDP-6-sulfoquinovose
SQD1
UDP-glucose
1,2-Diacyl-3-β-D-galactosyl-sn-glycerol
3.2.1.22
2.4.1.46
1,2-Diacyl-3-(Galα1-6Galβ1)-sn-glycerol
2.4.1.184
2.4.1.241
2.3.1.141
2.3.1.134
Monogalactosyl-monoacylglycerol
Acyl-monogalactosyl-diacylglycerol
1,2-Diacyl-3-(NeuAcα2-3Galβ1)-sn- glycerol
2.4.99.5
3.1.1.26
3-β-D-Galactosyl-sn-glycerol

Strengths and Weaknesses of Gene Enrichment

+ Enrichment can find functional patterns in a list that is too large to understand by just reading gene names
+ Enrichment is flexible for your choice of functional gene sets
- Enrichment requires a well annotated genome
- Enrichment does not work well for very large or very small gene lists
- Enrichment is not sensitive for small changes in a large number of genes from a functional category or pathway
- Enrichment relies on a gene list that is usually constructed with an. arbitrary cutoff (p-value or fold change)