Gene function annotation and Gene set analysis

Paul D Thomas
University of Southern California
October 26, 2018

In this lecture

- Methods and online information sources for function annotation
  - Understand what you are getting from each source so you can use it wisely
  - Gene Ontology
  - Pathway databases
- Emphasis on understanding “computationally predicted” function annotations (homology)
  - These make up the bulk of available annotations

In this lecture

- What are function annotations?
- How are they created?
- How do I use them?
  - Enrichment analysis
- Where do I get them?
  - Download
  - Create new ones

Gene function annotation sources

- Gene Ontology (GO)
- Pathway databases
  - Reactome
  - PANTHER
  - BioCyc
  - KEGG, WikiPathways (less computable)

Gene Ontology

Ontology

A structured representation of biology, composed of:

- Classes
- Relations
- Definitions

+

Annotations

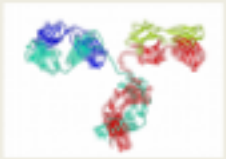
Statements about the functions of specific **gene products.**

3 aspects:

- Molecular function
- Biological process
- Cellular component

- Glutamine-tRNA ligase activity
- Translation
- Cytoplasm

QARS
Gln tRNA synthetase



- Antigen binding
- Adaptive immune response
- Extracellular

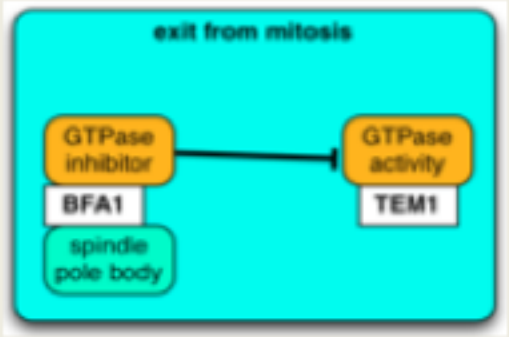
IGHA1
Immunoglobulin heavy constant alpha 1

=

Model of biology

Representation of current knowledge in a manner that is:

- Human understandable
- Machine computable



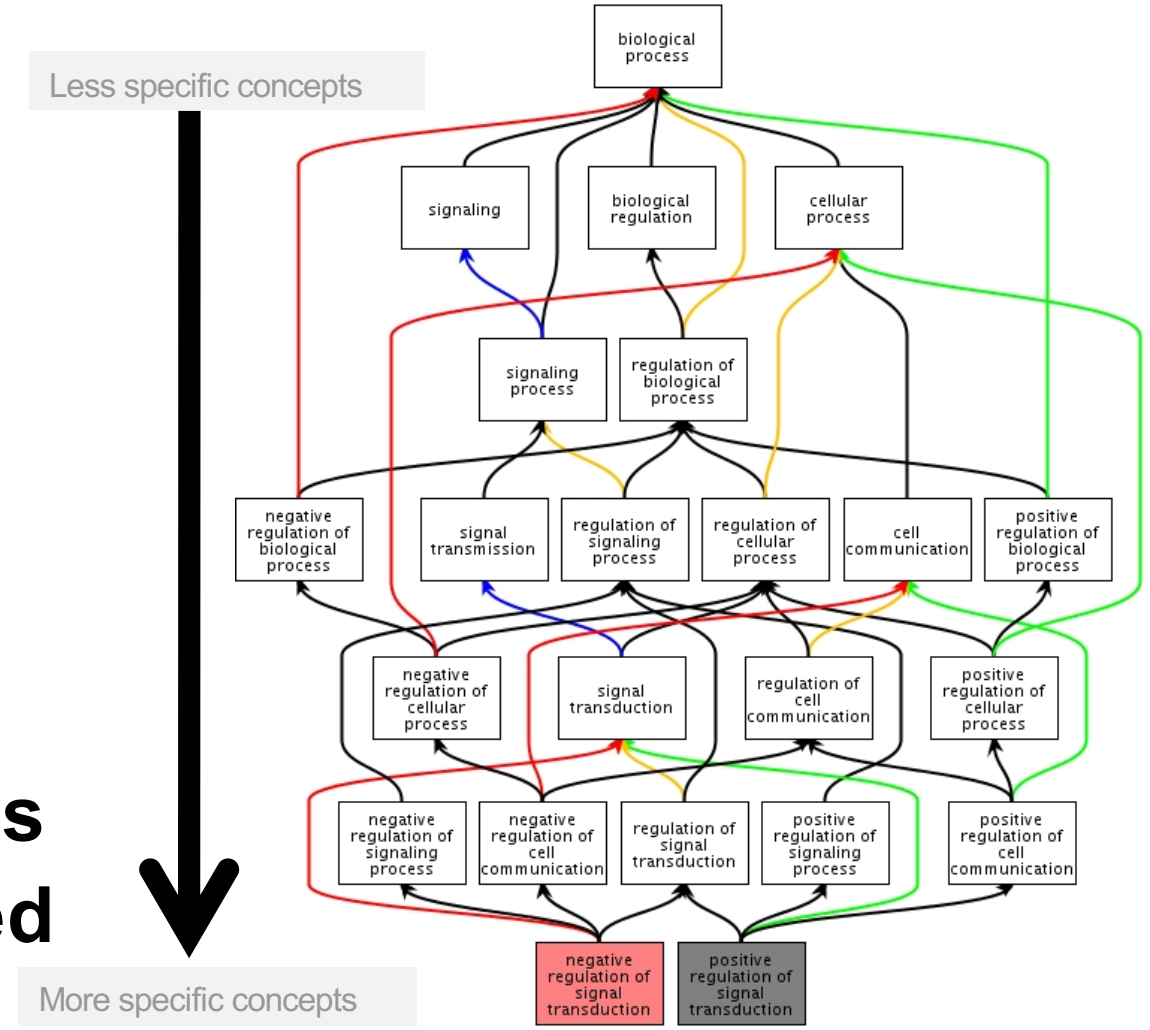
Gene Ontology

- Formal representation of biology knowledge domain, as it relates to genes and gene products (mostly proteins)
- Three knowledge domains:
  - Molecular function: what a gene product does with its direct physical interaction partners, e.g. protein kinase
  - Cellular component: where the protein is located when the function is carried out, e.g. plasma membrane
  - Biological process: “system” function carried out by multiple molecular functions working together in a regulated manner, e.g. pathways, cellular processes, organ functions, organism behavior
- Concepts are joined together by directional Relations: is_a, part_of, regulates

The Gene Ontology

- **A way to capture biological knowledge in a written and computable form**

- **A set of concepts and their relationships to each other arranged as a graph**



www.ebi.ac.uk/QuickGO

GO "annotations"

- An annotation is a statement about biological function
- Formally, an association between a gene* and a GO term describing its function

  Examples:
  Annotation 1: INSR + 'receptor activity'
  Annotation 2: INSR + 'plasma membrane'
  Annotation 3: INSR + 'insulin receptor signaling pathway'

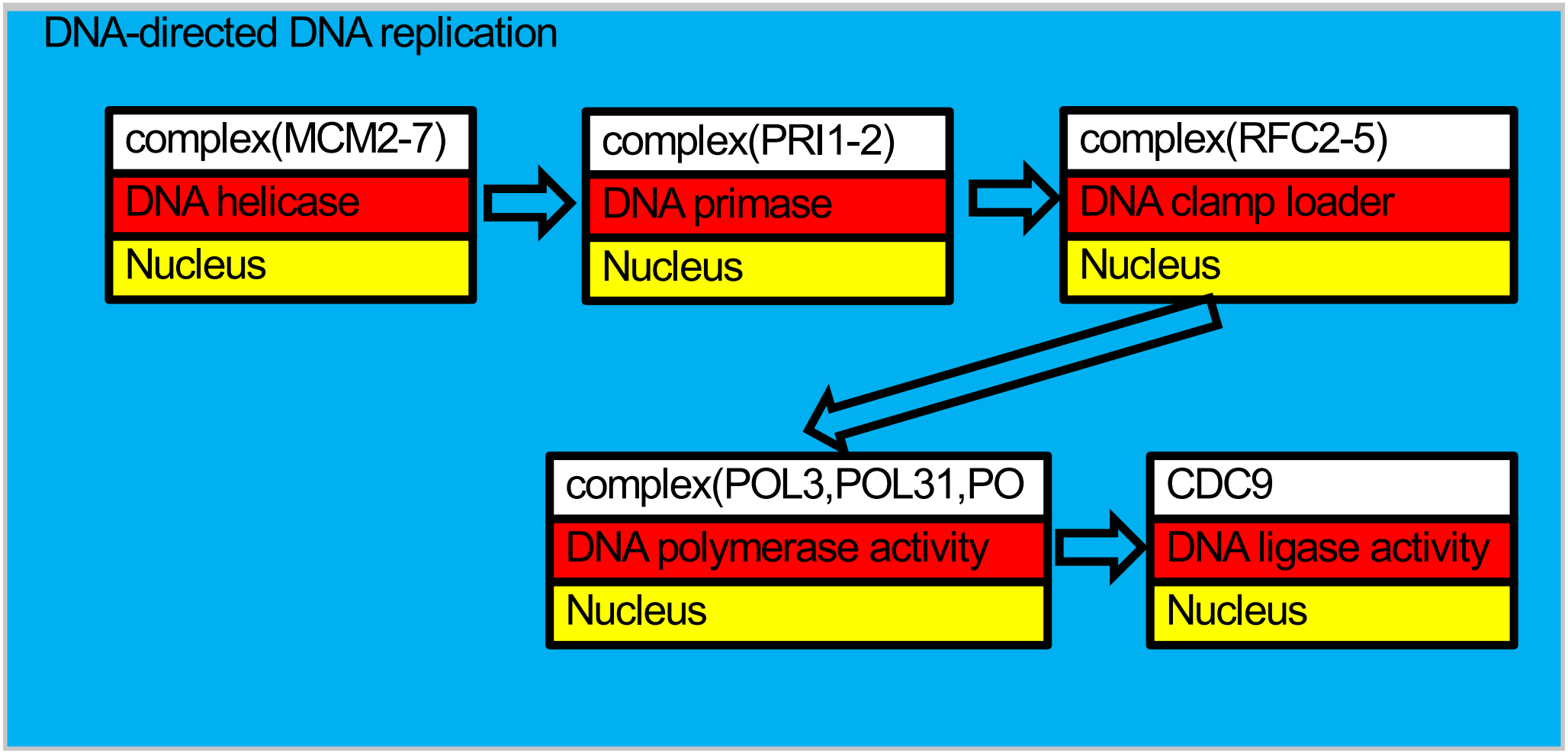
- Each annotation must be based on *evidence*, which is recorded as part of the annotation
  - Evidence code (type of evidence)
  - Reference (published journal article)

* Usually a gene product (protein or functional RNA), can also be a specific isoform or a macromolecular complex

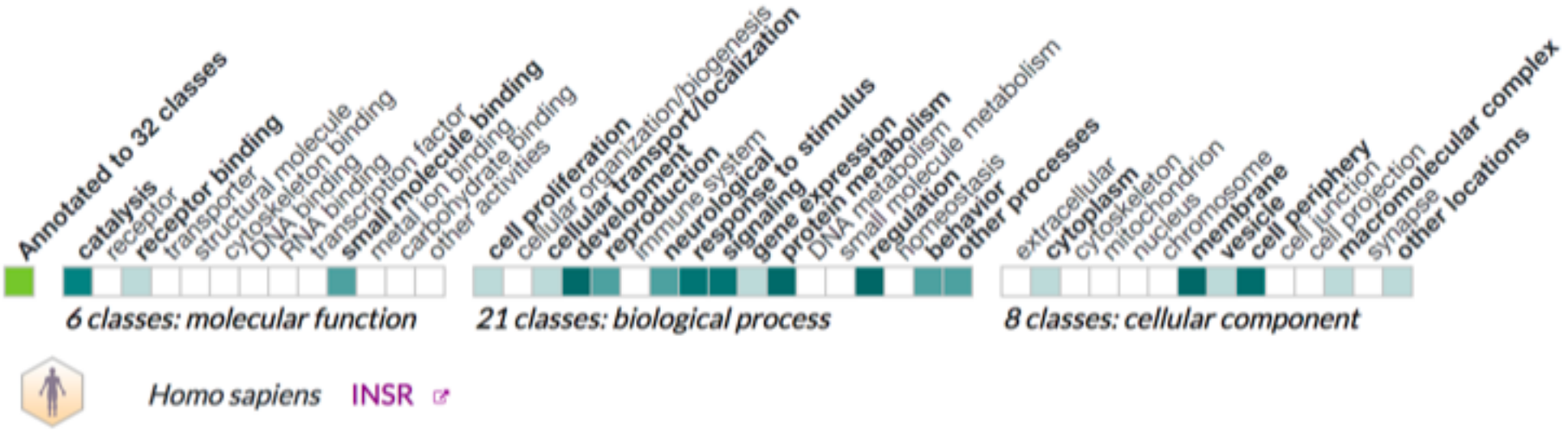
GO seeks to represent biological knowledge at many levels of organization

Tissue regeneration

Cell cycle



Human Insulin Receptor gene INSR



https://www.alliancegenome.org/gene/HGNC:6091

Human Insulin Receptor gene INSR

Homo sapiens INSR signaling

Term	Evidence	Based on	Reference
activation of MAPK activity	IMP		PMID:17001305
G protein-coupled receptor signaling pathway	IDA		PMID:9092559
insulin receptor signaling pathway	ISS	UniProtKB:P15127	PMID:19406747
	IDA		PMID:20455999
			PMID:6849137
			PMID:8440175
	TAS		Reactome:R-HSA-74752
insulin-activated receptor activity	IDA		PMID:6849137
			PMID:8440175

https://www.alliancegenome.org/gene/HGNC:6091

"Known" functions of genes: Where did this information come from?

MMP2 involved_in collagen catabolic process
ADAMTS2 involved_in collagen catabolic process
ADAMTS3 involved_in collagen catabolic process
…

Published papers

Biocuration

"GO annotations"
= computational statements about how genes function in biological systems


GO
Annotating cardiovascular genes


AspGD


PomBase


dictyBase


MGI


FlyBase


CGD


SGD


GENEONTOLOGY
Unifying Biology


EcoCyc


IntAct


InterPro


WormBase
The Biology and Genome of C. elegans.


PRO
Protein Ontology


tair


e!Ensembl


BERKELEY LAB


UniProt


Reactome


AgBase


GRAMENE


University of Maryland School of Medicine
INSTITUTE FOR GENOME SCIENCES


RGD


ZFIN


J. Craig Venter
INSTITUTE

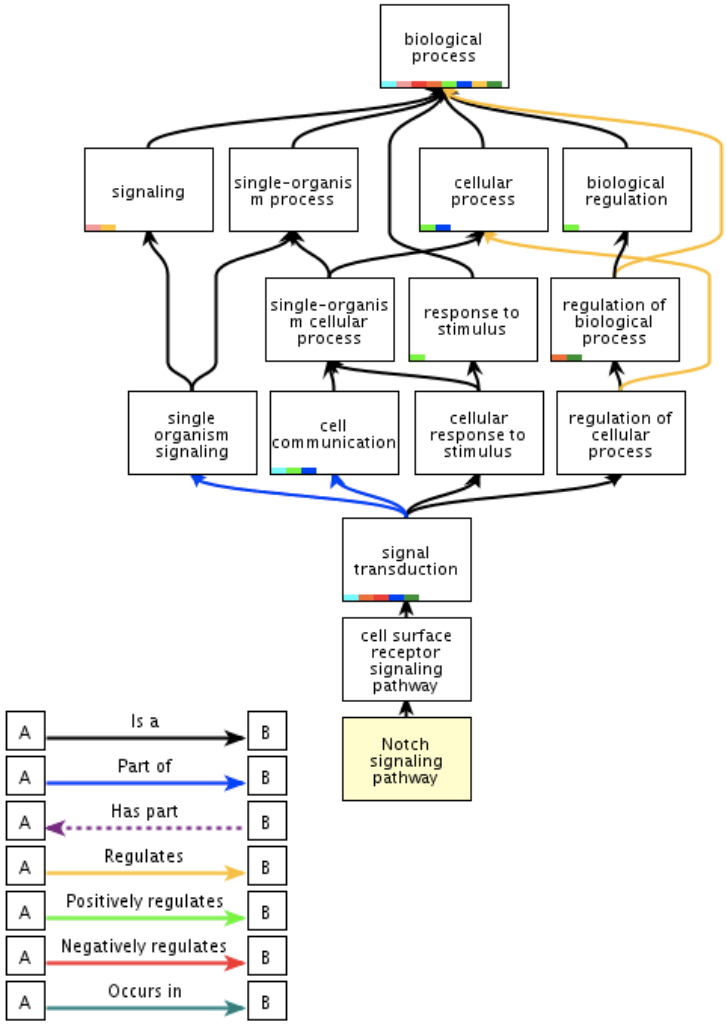

PANTHER
Classification System

Pathway representations

- Point of view from the molecular reaction
  - Generalized to include covalent and noncovalent (e.g. binding) reactions
- Concepts are reaction, molecule classes
- Relations are between molecule classes and reactions
  - Catalyst
  - Reactant
  - Product
- Top level structure provided by SBML, BioPAX
  - Systems modeling community vs. Genomics community

Notch signaling pathway in GO

Relations to more general classes

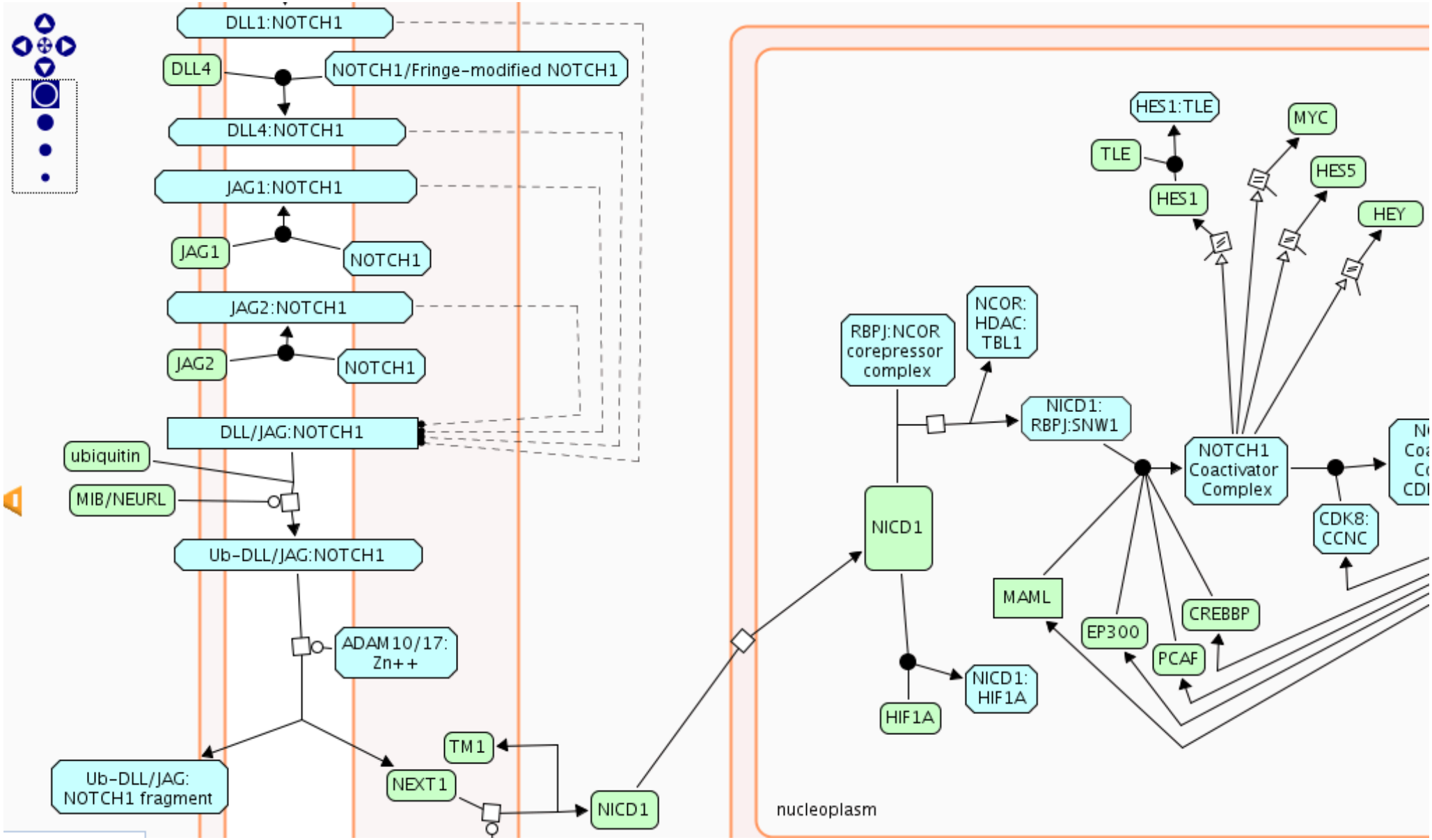


Relations to more specific classes

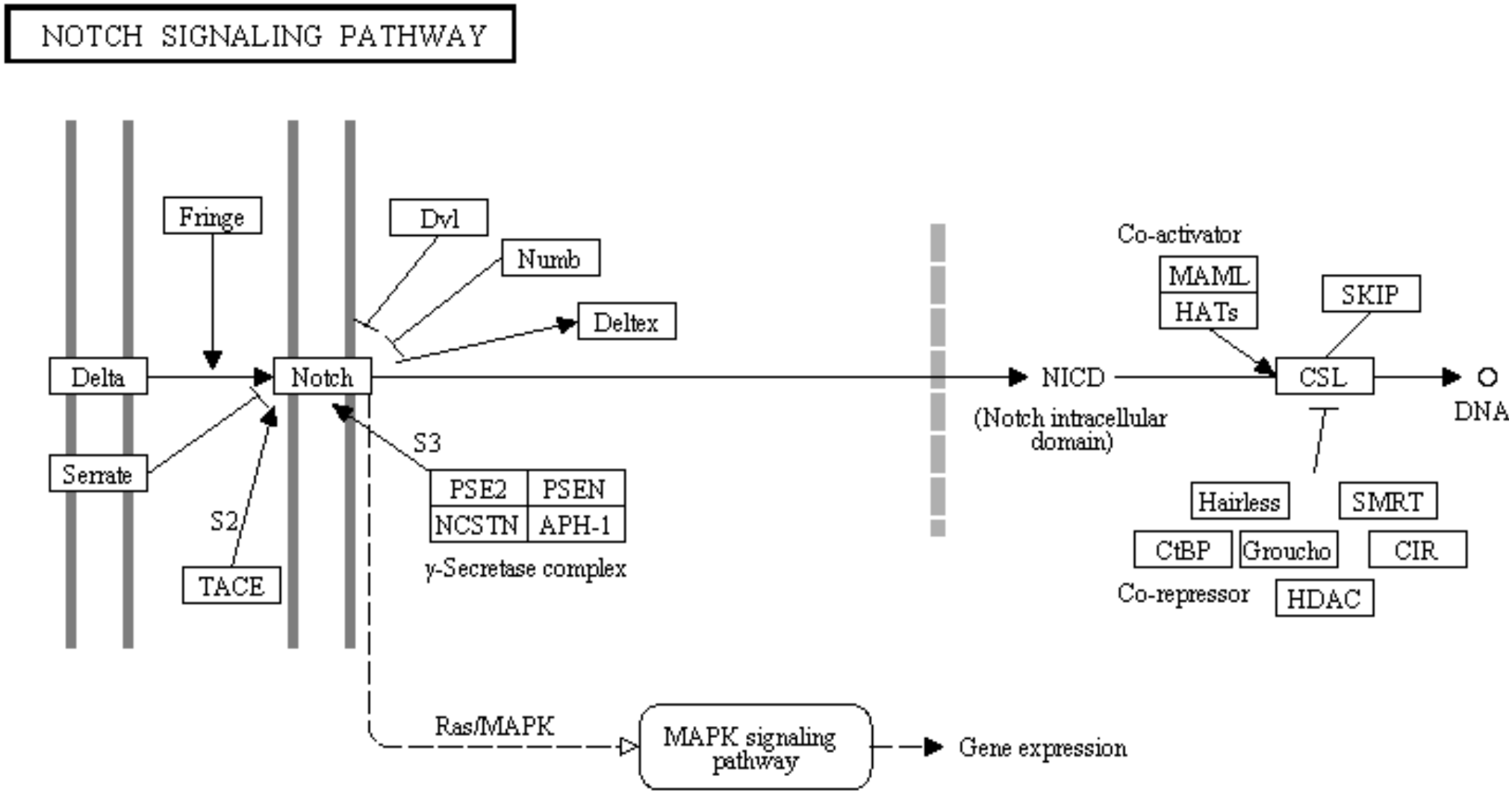
GO:0007219 Notch signaling pathway

- GO:0045746 negative regulation of Notch signaling pathway
- GO:0035333 Notch receptor processing, ligand-dependent
- GO:0061314 Notch signaling involved in heart development
- GO:0060853 Notch signaling pathway involved in arterial endothelial cell fate commitment
- GO:0060227 Notch signaling pathway involved in camera-type eye photoreceptor fate com
- GO:0021876 Notch signaling pathway involved in forebrain neuroblast division
- GO:0021880 Notch signaling pathway involved in forebrain neuron fate commitment
- GO:0003137 Notch signaling pathway involved in heart induction
- GO:2000796 Notch signaling pathway involved in negative regulation of venous endothelia
- GO:0003270 Notch signaling pathway involved in regulation of secondary heart field cardio
- GO:1902359 Notch signaling pathway involved in somitogenesis
- GO:0045747 positive regulation of Notch signaling pathway
- GO:0007221 positive regulation of transcription of Notch receptor target
- GO:0008593 regulation of Notch signaling pathway

Notch signaling in Reactome



Notch signaling in KEGG



GO vs. pathway representations

- GO
  - More comprehensive
  - More context (ontology parent classes)
- Pathway representations
  - More information about pathway steps
  - In Reactome, more information about molecular complex association/dissociation

GO annotations
know what you're getting

- Annotation is an association between
  - A gene/gene product
  - A Gene Ontology term

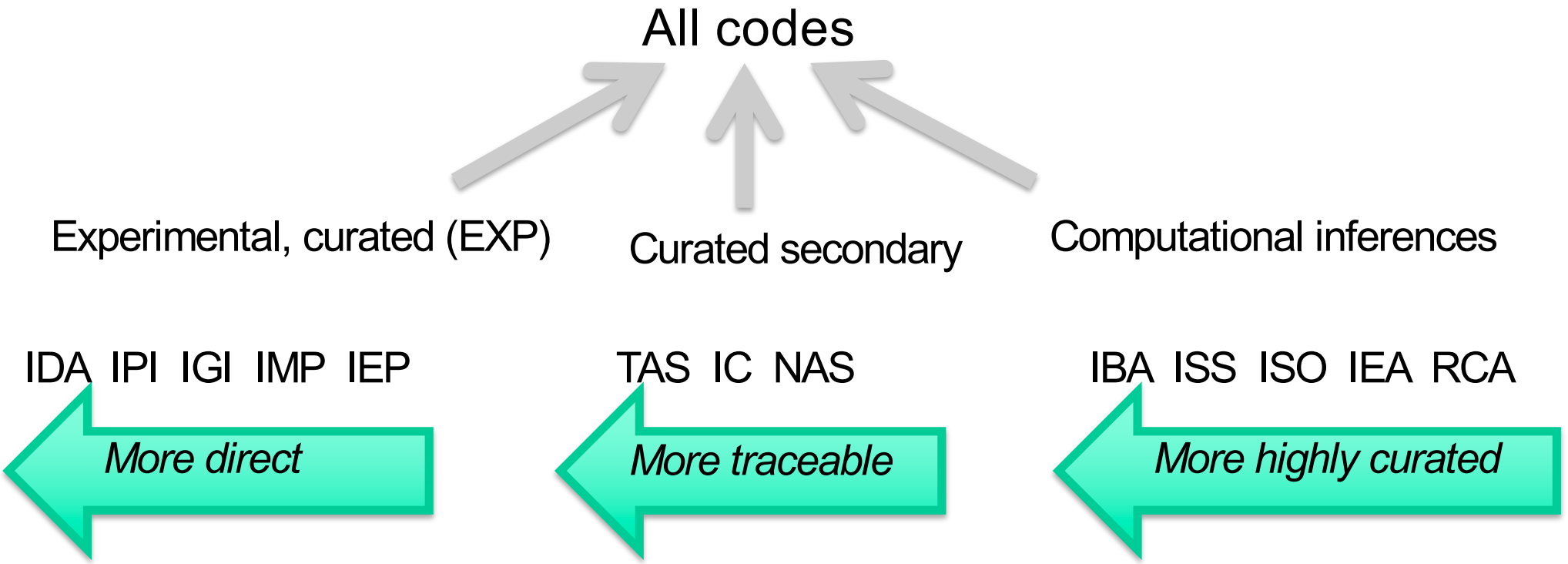
Annotation 1: INSR performs_function 'receptor activity'
Annotation 2: INSR located_in 'plasma membrane'
Annotation 3: INSR involved_in 'insulin receptor signaling pathway'

- But there is more information
  - Qualifier
  - Evidence code and evidence

Common qualifiers

- NOT
  - This is really important, it means that the gene product does NOT have a particular function
- contributes_to
  - This is usually used when a gene product is part of a complex that has a particular molecular function, but it is not the active subunit

GO evidence codes



http://www.geneontology.org/page/guide-go-evidence-codes

General advice for evidence codes

- Filter out less reliable annotations
  - high-throughput evidence codes (HTP*)
  - Large-scale computational predictions (RCA)
  - Expression pattern evidence (IEP)
- Inferred annotations are generally accurate, so include them
  - Based on homology, and curator-reviewed at varying levels

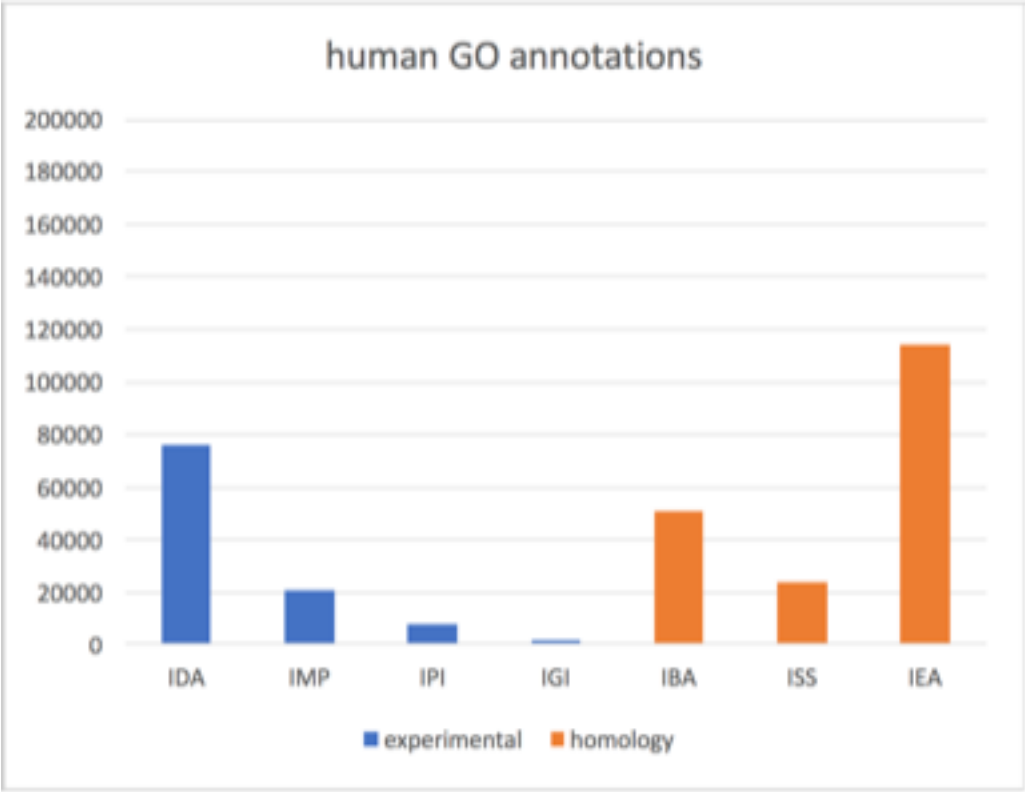
*and more specific HTP codes: HDA, HGI, HMP, HEP

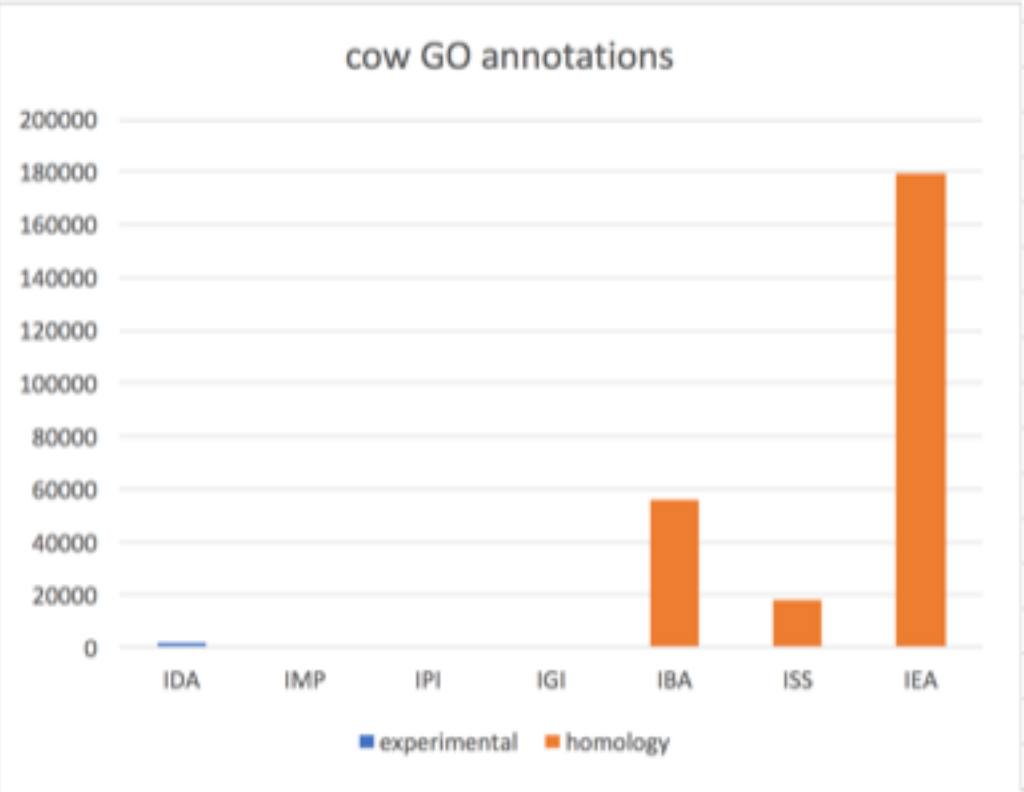
Homology inference

- Our knowledge of human genes is limited
  - Only ~22,000 of 115,000 papers used in GO annotations are on human genes
- How can we use phylogenetic inference to augment human gene annotations?

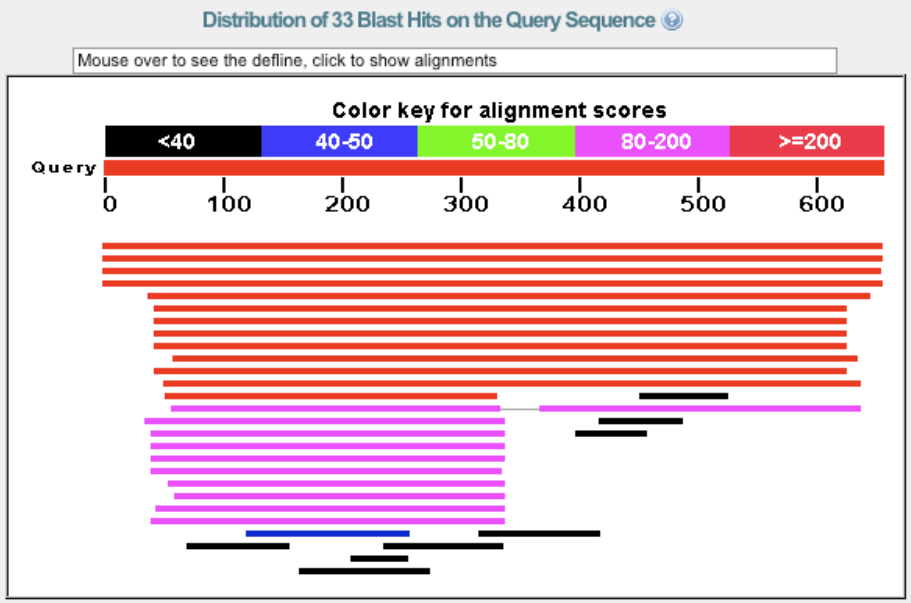
(103271)	Homo sapiens
(97150)	Mus musculus
(90192)	Fungi
(65153)	Viridiplantae
(58699)	Arabidopsis thaliana
(48555)	Rattus norvegicus
(46558)	Drosophila melanogaster
(45594)	Saccharomyces cerevisiae S288c
(33989)	Bacteria
(21772)	Danio rerio
(20589)	Schizosaccharomyces pombe
(19561)	Caenorhabditis elegans
(14426)	Escherichia coli K-12
(8572)	Candida albicans
(6808)	Dictyostelium discoideum
(6773)	Mycobacterium tuberculosis H37Rv
(4546)	Pseudomonas aeruginosa PAO1

Availability of experimental annotations depends on the organism





Inference using pairwise homology search (BLAST)



Sequences producing significant alignments:		Score (Bits)	E Value	
sp|P42898.3|MTHR_HUMAN	RecName: Full=Methylenetetrahydrofolat...	1360	0.0	G
sp|Q60HE5.1|MTHR_MACFA	RecName: Full=Methylenetetrahydrofolat...	1306	0.0	
sp|Q5I598.1|MTHR_BOVIN	RecName: Full=Methylenetetrahydrofolat...	1203	0.0	G
sp|Q9WU20.1|MTHR_MOUSE	RecName: Full=Methylenetetrahydrofolat...	1203	0.0	G
sp|Q17693.2|MTHR_CAEEL	RecName: Full=Probable methylenetetrah...	627	6e-179	G
sp|Q9SE94.1|MTHR1_MAIZE	RecName: Full=Methylenetetrahydrofola...	524	7e-148	G
sp|Q75HE6.1|MTHR_ORYSJ	RecName: Full=Probable methylenetetrah...	523	2e-147	G
sp|Q9SE60.1|MTHR1_ARATH	RecName: Full=Methylenetetrahydrofola...	515	4e-145	G
sp|O80585.2|MTHR2_ARATH	RecName: Full=Methylenetetrahydrofola...	512	3e-144	G
sp|Q10258.1|MTHR1_SCHPO	RecName: Full=Methylenetetrahydrofola...	468	5e-131	G
sp|P53128.2|MTHR2_YEAST	RecName: Full=Methylenetetrahydrofola...	462	3e-129	G
sp|O74927.2|MTHR2_SCHPO	RecName: Full=Methylenetetrahydrofola...	345	6e-94	G
sp|O67422.1|METF_AQUAE	RecName: Full=5,10-methylenetetrahydro...	213	3e-54	
sp|P46151.2|MTHR1_YEAST	RecName: Full=Methylenetetrahydrofola...	196	4e-49	G
sp|O54235.1|METF_STRLI	RecName: Full=5,10-methylenetetrahydro...	189	7e-47	
sp|P71319.1|METF_PECCC	RecName: Full=5,10-methylenetetrahydro...	162	9e-39	
sp|P11003.2|METF_SALTY	RecName: Full=5,10-methylenetetrahydro...	156	5e-37	
sp|P0AEZ1.1|METF_ECOLI	RecName: Full=5,10-methylenetetrahydro...	153	3e-36	

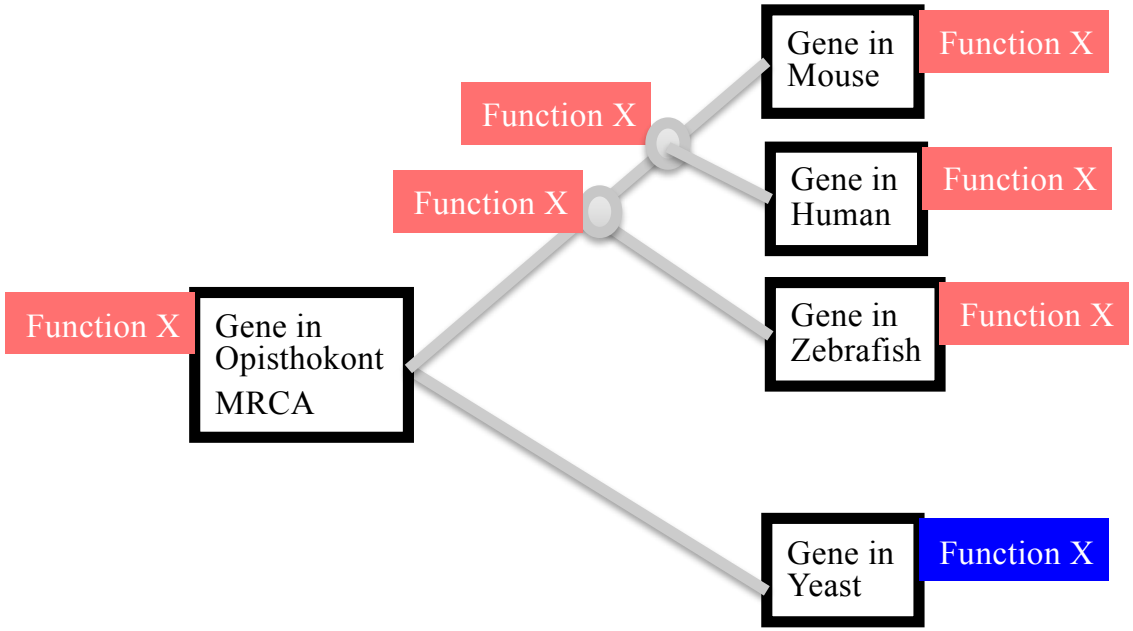
Significant hit to a yeast protein with a literature-based annotation.

This ID is in the evidence field

Homology is common ancestry

- Two sequences are similar **because** they are homologous (at least for relatively long, non-repetitive sequences, i.e. almost all genes)
- More properly, transitive annotation of function is inheritance!

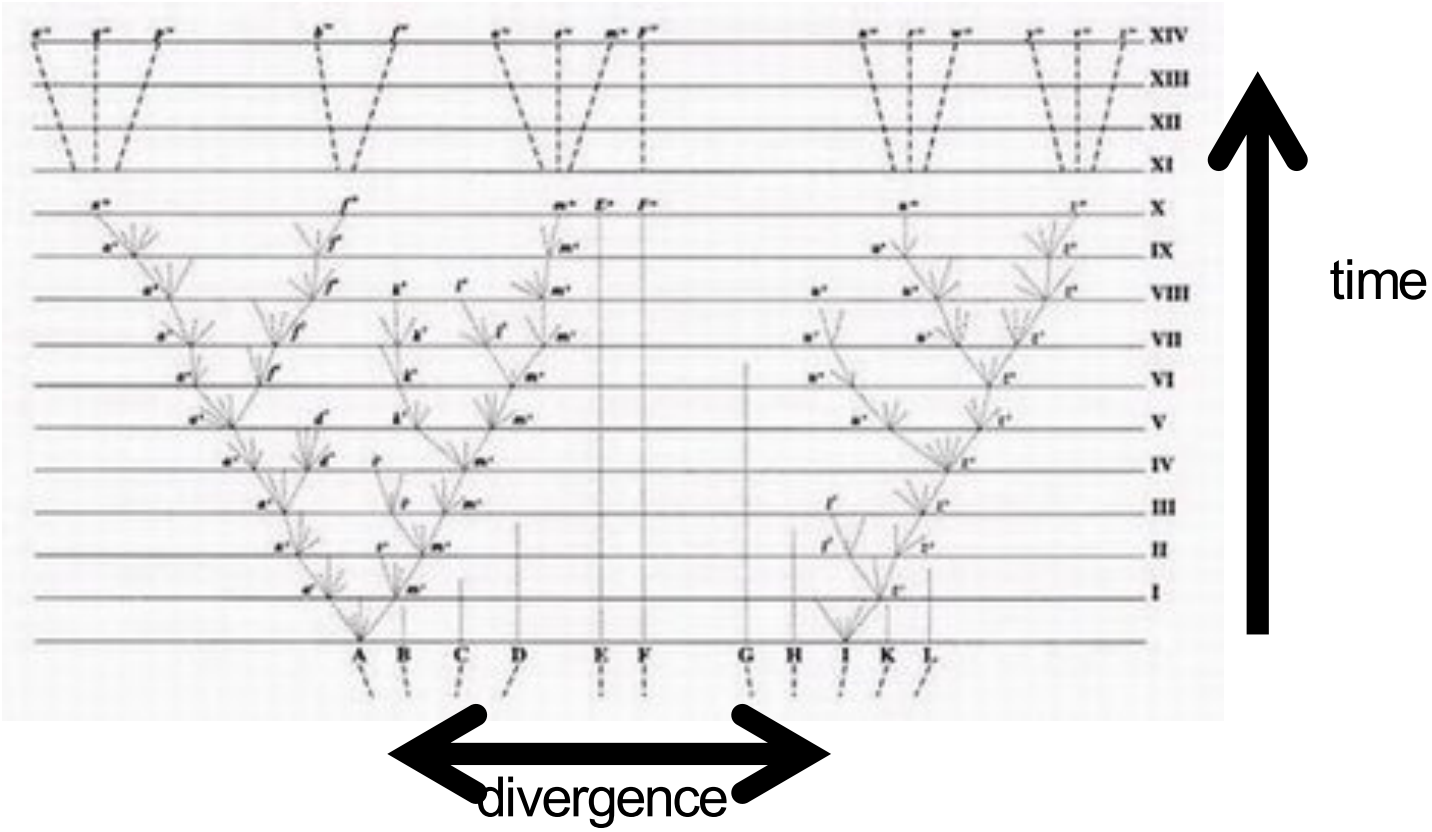
– related genes have a common function **because** their common ancestor had that function, which was inherited by its descendants



Darwin's tree of life

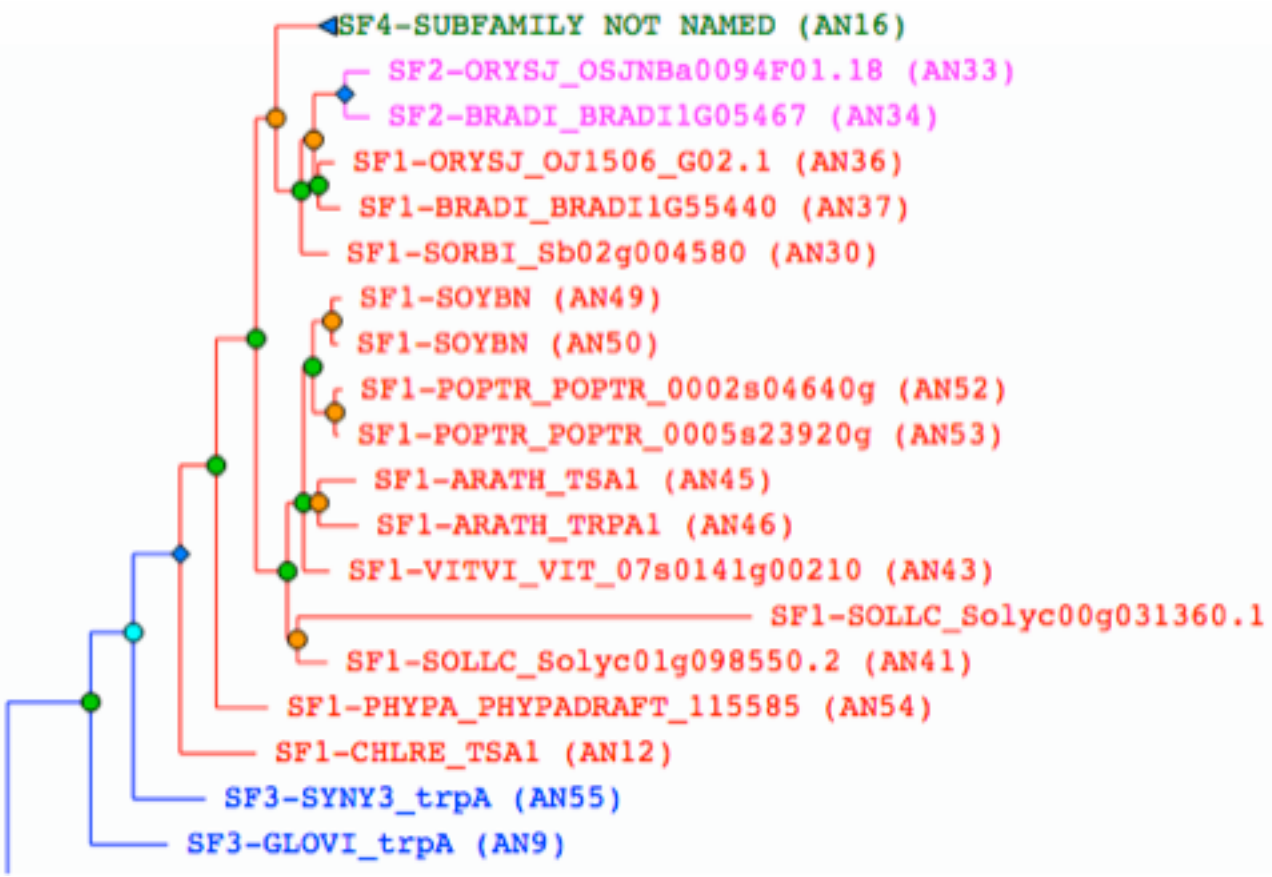
Species "split" into separate daughter species

Each daughter species changes *independently* over generations



- A "tree" arises from a repeated process of
  - Copying at a "point" in time
  - Divergence over time

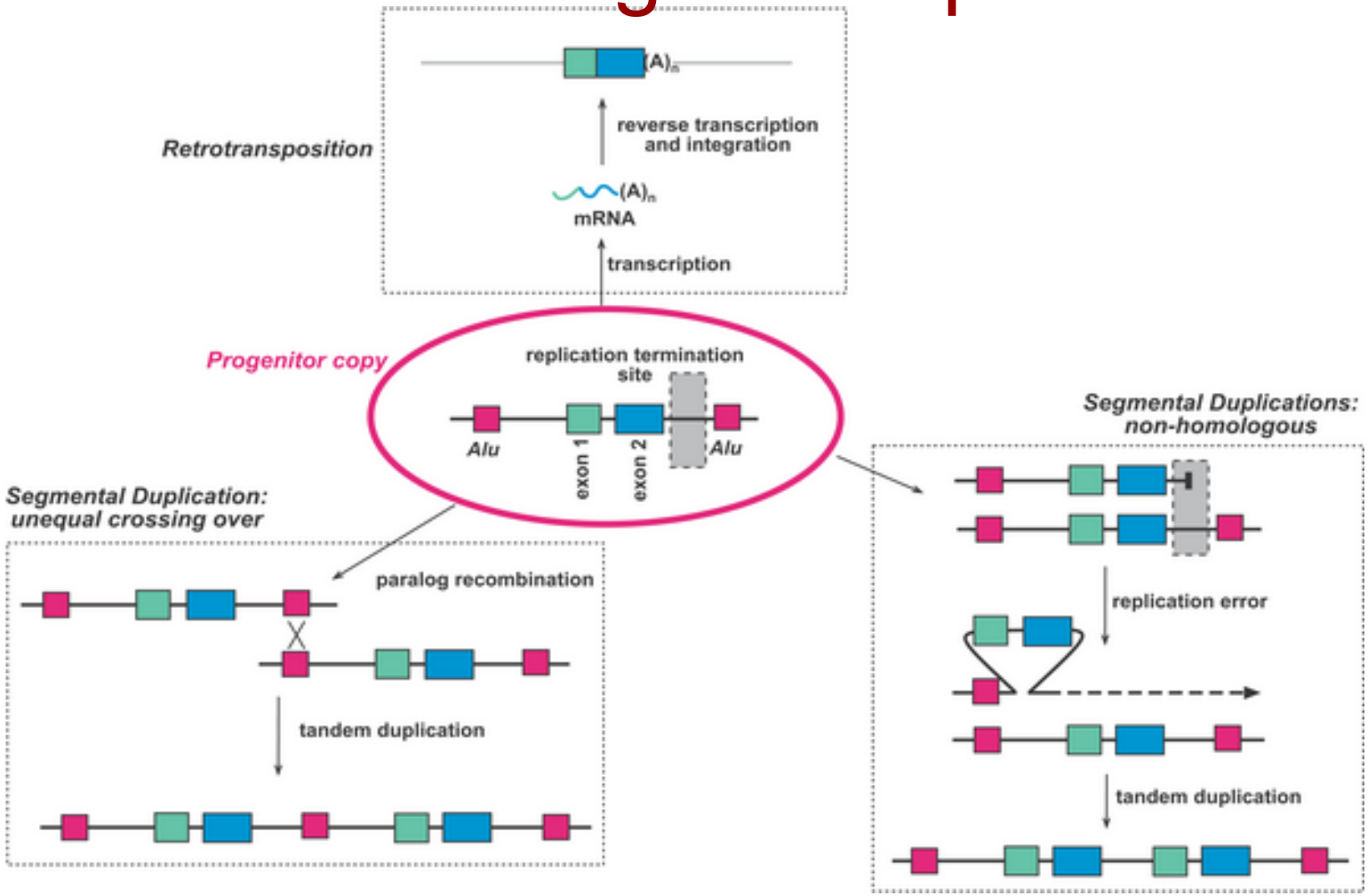
A “reconciled” gene tree



O	OSJNBa0094F01.18	Oryza sativa
B	BRADI1G05467	Brachypodium dis…
O	OJ1506_G02.1	Oryza sativa
B	BRADI1G55440	Brachypodium dis…
S	Sb02g004580	Sorghum bicolor
G		Glycine max
G		Glycine max
P	POPTR_0002s04640g	Populus trichoca…
P	POPTR_0005s23920g	Populus trichoca…
A	TSA1	Arabidopsis thal…
A	TRPA1	Arabidopsis thal…
V	VIT_07s0141g00210	Vitis vinifera
S	Solyc00g031360.1	Solanum lycopers…
S	Solyc01g098550.2	Solanum lycopers…
P	PHYPADRAFT_115585	Physcomitrella p…
C	TSA1	Chlamydomonas re…
B	trpA	Synechocystis
B	trpA	Gloeobacter viol…

Nodes are colored according to evolutionary “copying” events:
speciation (green/blue), orange (duplication), cyan (horizontal transfer)

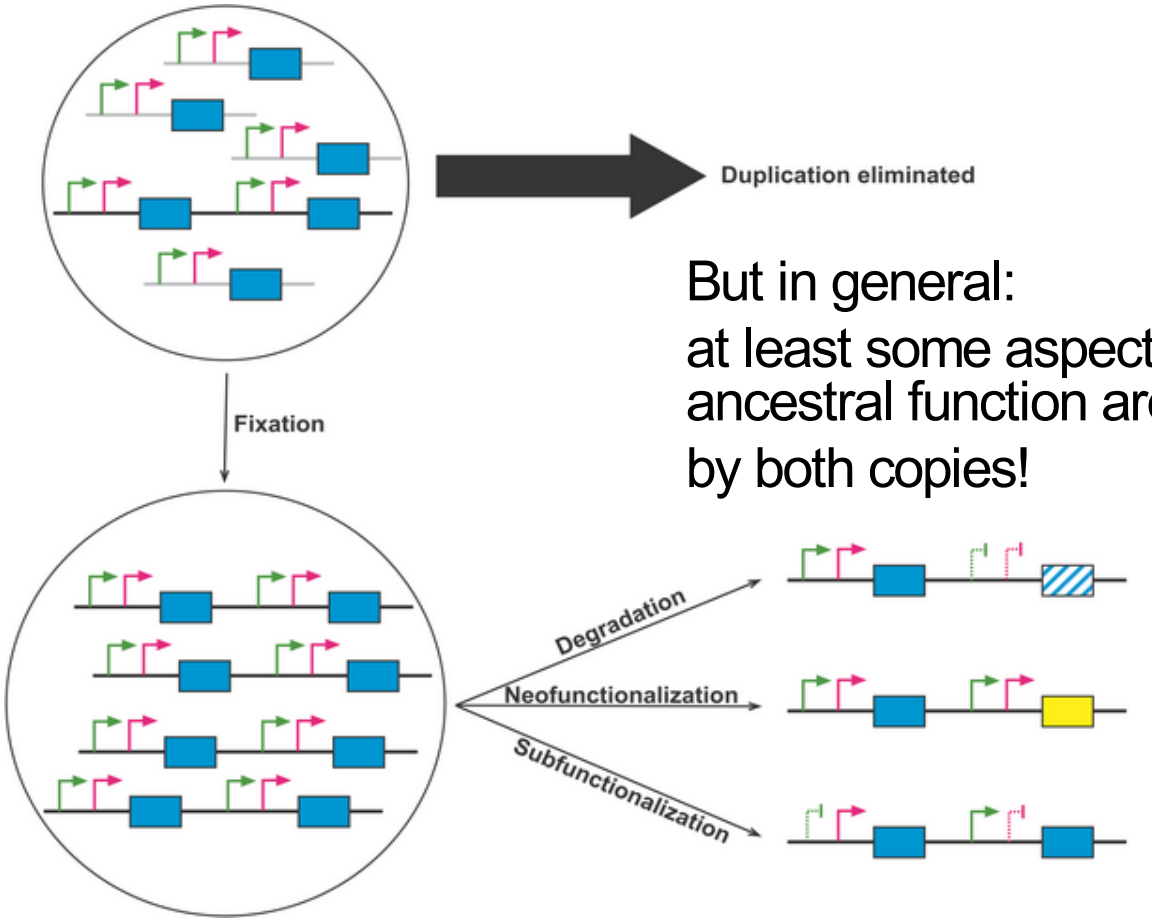
Mechanisms of gene duplication



Hurles M (2004) Gene Duplication: The Genomic Trade in Spare Parts. PLOS Biology 2(7): e206. https://doi.org/10.1371/journal.pbio.0020206
http://journals.plos.org/plosbiology/article?id=10.1371/journal.pbio.0020206

PLOS | BIOLOGY

Fate of duplicate genes

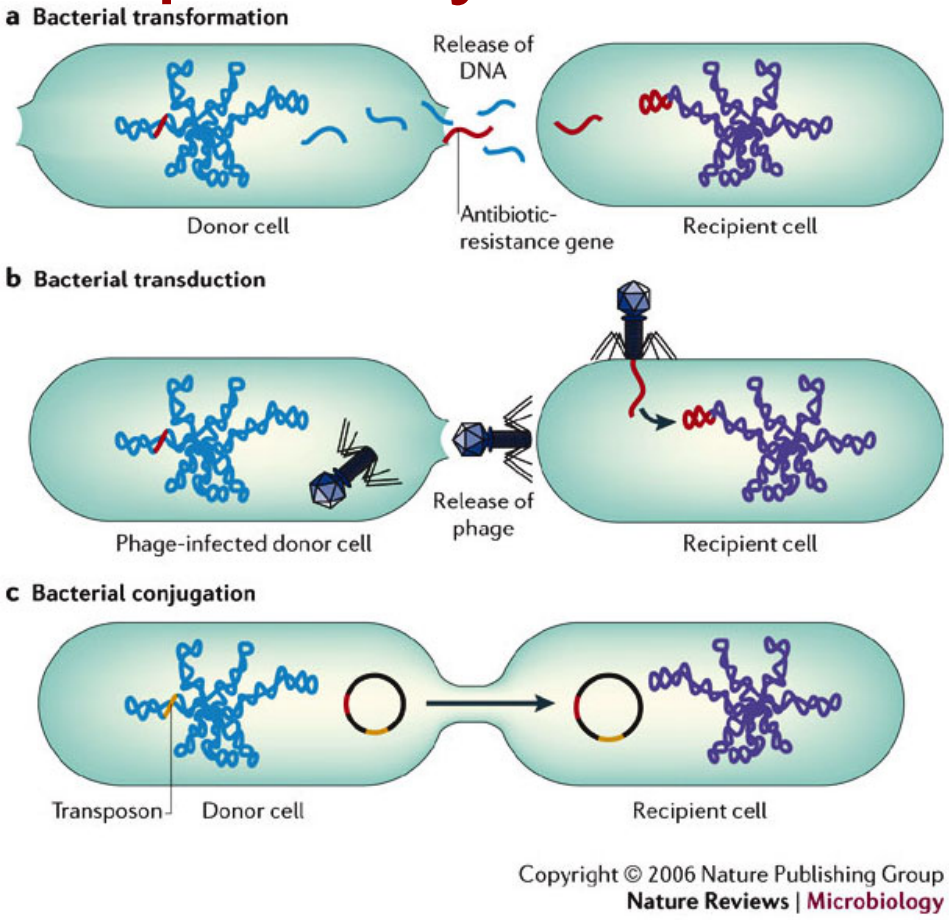


But in general:
at least some aspects of ancestral function are inherited by both copies!

Hurles M (2004) Gene Duplication: The Genomic Trade in Spare Parts. PLOS Biology 2(7): e206. https://doi.org/10.1371/journal.pbio.0020206
http://journals.plos.org/plosbiology/article?id=10.1371/journal.pbio.0020206

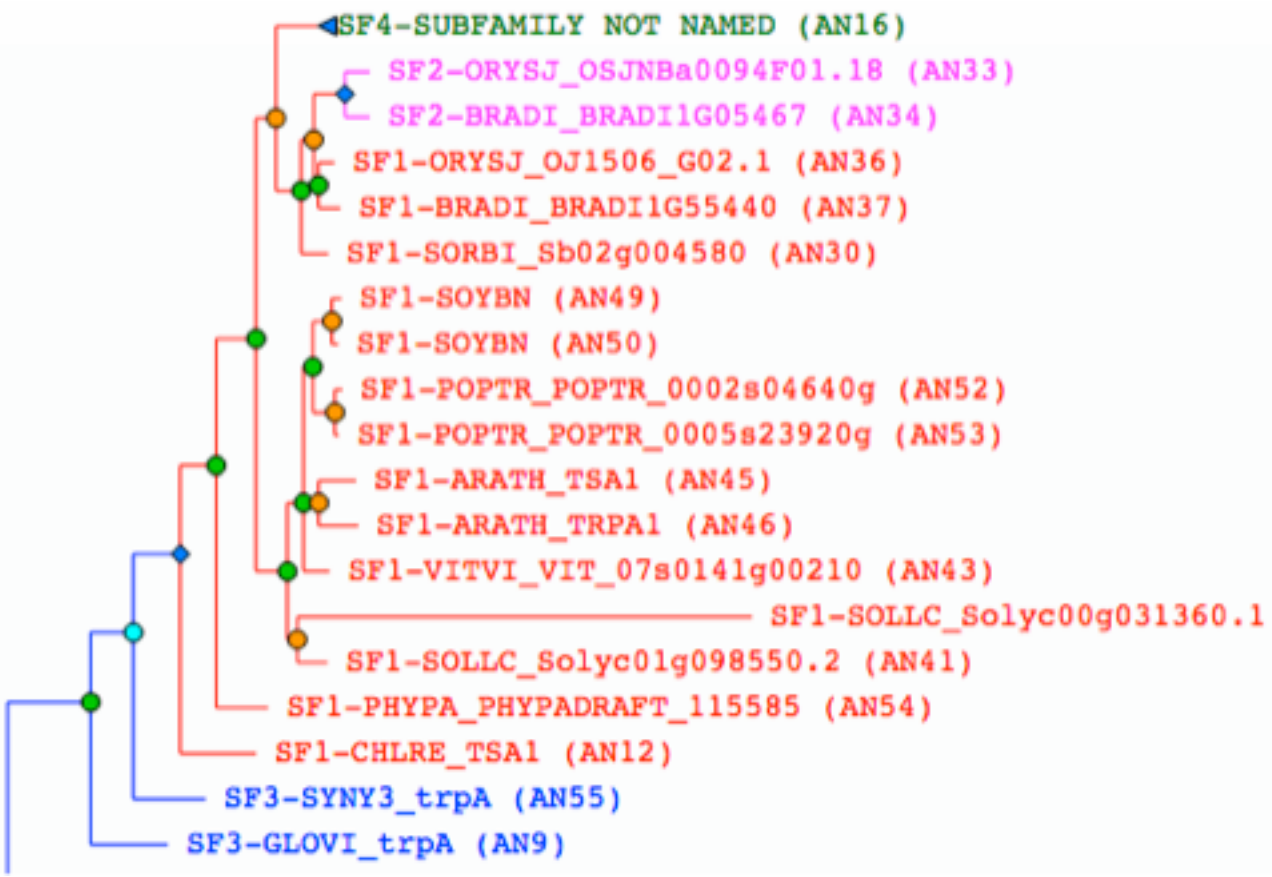
PLOS | BIOLOGY

Mechanisms of horizontal transfer in prokaryotes



Furuya EY and Lowy F (2006) Antimicrobial-resistant bacteria in the community setting *Nat Rev Microbiol.* **4**: 36–45 doi:10.1038/nrmicro1325

Gene family (focus on the leaves)



O	OSJNBa0094F01.18	Oryza sativa
B	BRADI1G05467	Brachypodium dis…
O	OJ1506_G02.1	Oryza sativa
B	BRADI1G55440	Brachypodium dis…
S	Sb02g004580	Sorghum bicolor
G		Glycine max
G		Glycine max
P	POPTR_0002s04640g	Populus trichoca…
P	POPTR_0005s23920g	Populus trichoca…
A	TSA1	Arabidopsis thal…
A	TRPA1	Arabidopsis thal…
V	VIT_07s0141g00210	Vitis vinifera
S	Solyc00g031360.1	Solanum lycopers…
S	Solyc01g098550.2	Solanum lycopers…
P	PHYPADRAFT_115585	Physcomitrella p…
C	TSA1	Chlamydomonas re…
B	trpA	Synechocystis
B	trpA	Gloeobacter viol…

A family is a group of homologous (related) sequences

GO annotation of protein families

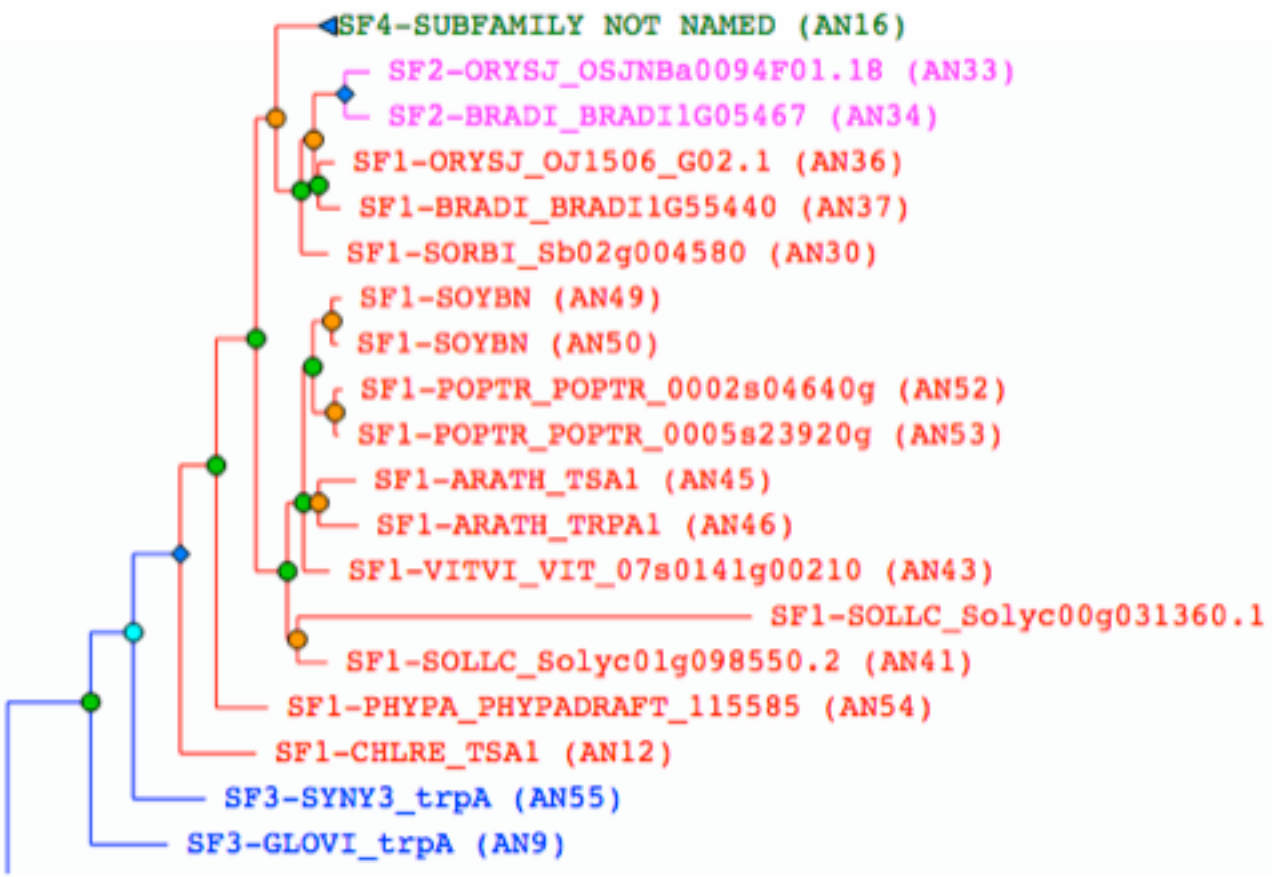
- Find functions that are broadly conserved among family members
- Annotate entire family with the corresponding GO terms

Manual GO annotation of predictive protein signatures: the InterPro approach to GO curation

Sarah Burge, Elizabeth Kelly, David Lonsdale, Prudence Mutowo-Muellenet, Craig McAnulla, Alex Mitchell, Amaia Sangrador-Vegas, Siew-Yit Yong, Nicola Mulder, Sarah Hunter

Database, Volume 2012, 1 January 2012, bar068, https://doi.org/10.1093/database/bar068

Gene tree (focus on the branch points)

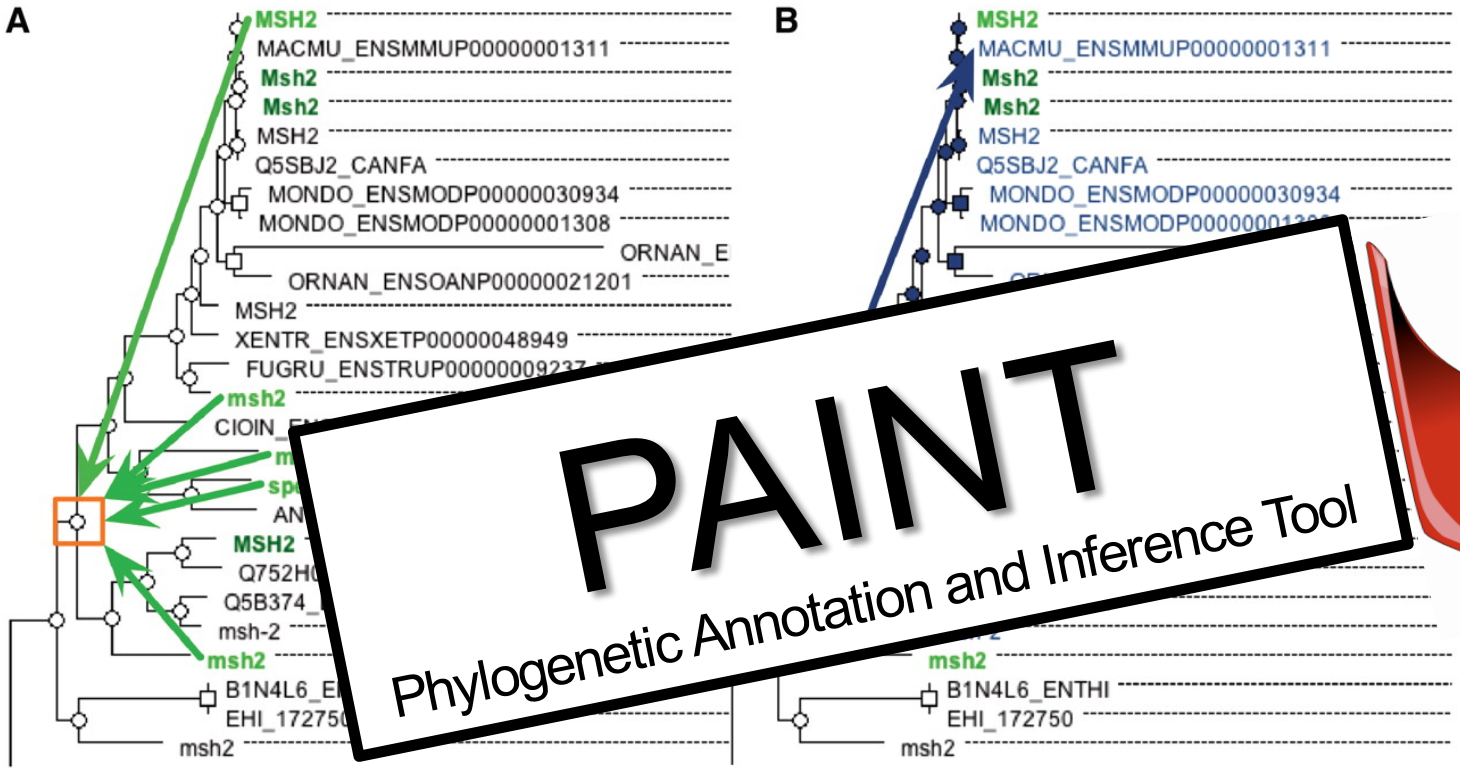


O	OSJNBa0094F01.18	Oryza sativa
B	BRADI1G05467	Brachypodium dis…
O	OJ1506_G02.1	Oryza sativa
B	BRADI1G55440	Brachypodium dis…
S	Sb02g004580	Sorghum bicolor
G		Glycine max
G		Glycine max
P	POPTR_0002s04640g	Populus trichoca…
P	POPTR_0005s23920g	Populus trichoca…
A	TSA1	Arabidopsis thal…
A	TRPA1	Arabidopsis thal…
V	VIT_07s0141g00210	Vitis vinifera
S	Solyc00g031360.1	Solanum lycopers…
S	Solyc01g098550.2	Solanum lycopers…
P	PHYPADRAFT_115585	Physcomitrella p…
C	TSA1	Chlamydomonas re…
B	trpA	Synechocystis
B	trpA	Gloeobacter viol…

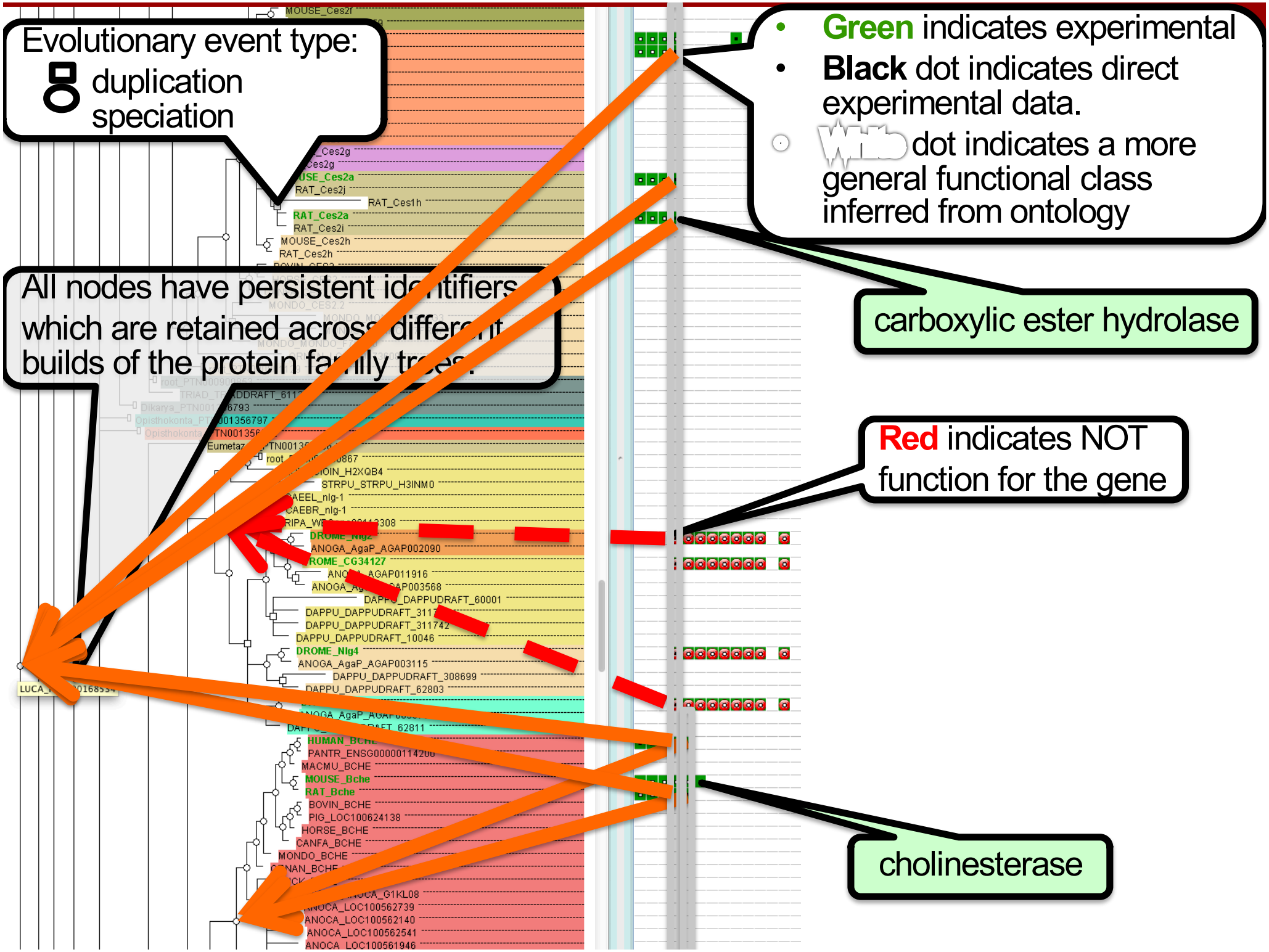
Function can change along any branch in the tree
But it is most likely to change after gene duplication

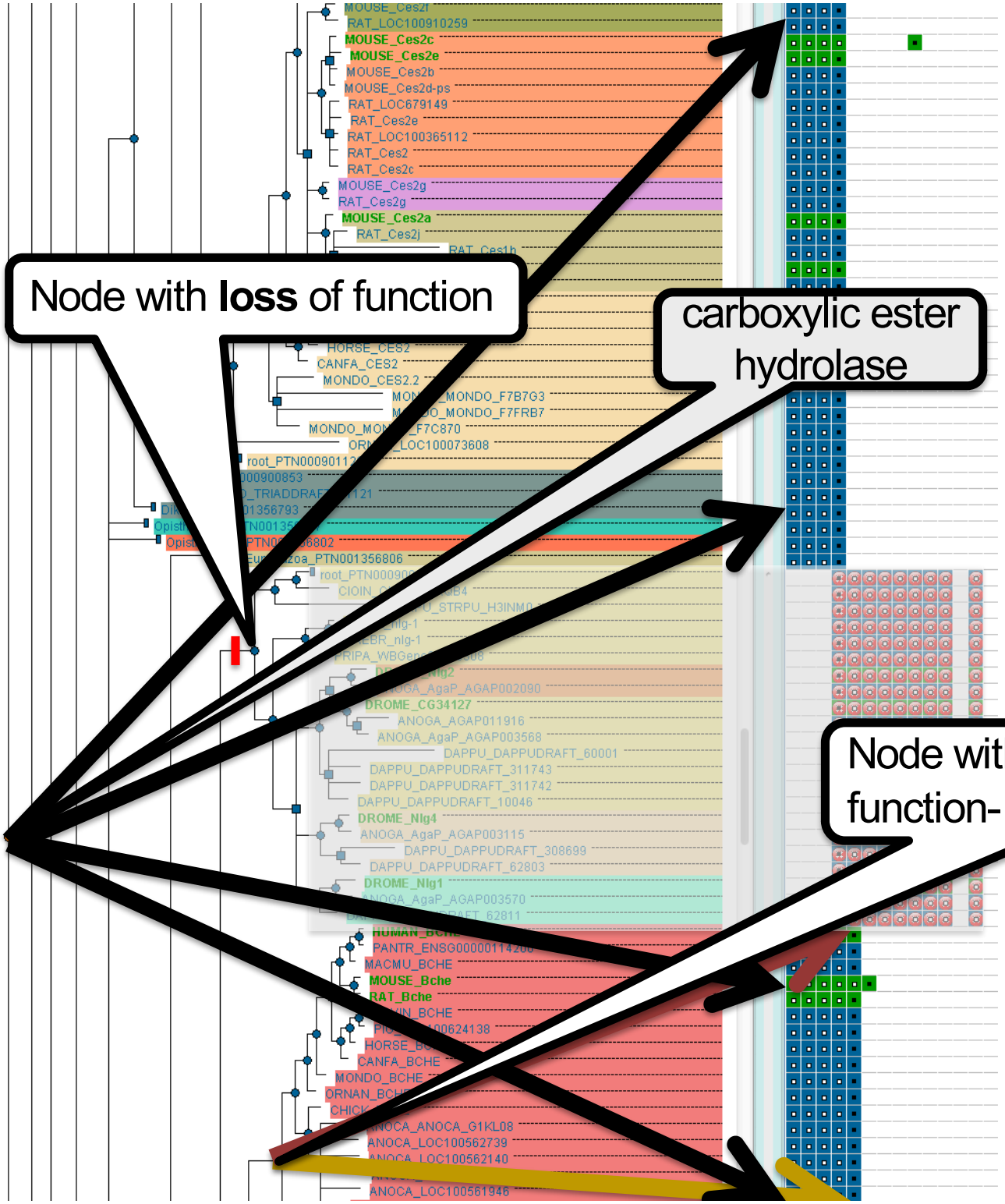
GO Phylogenetic Annotation Project

- Use reconciled gene trees
  - PANTHER, currently ~13,000 families covering ~1.1M genes in 104 species
- Manually review experimental GO annotations for related genes
- Build a model of branches in the gene tree where specific functions were gained and lost, that explains the distribution of functions found in modern-day genes



Gaudet, P., et al. (2011). Phylogenetic-based propagation of functional annotations within the Gene Ontology consortium. *Briefings in Bioinformatics*, 12(5), 449–62. doi:10.1093/bib/bbr042

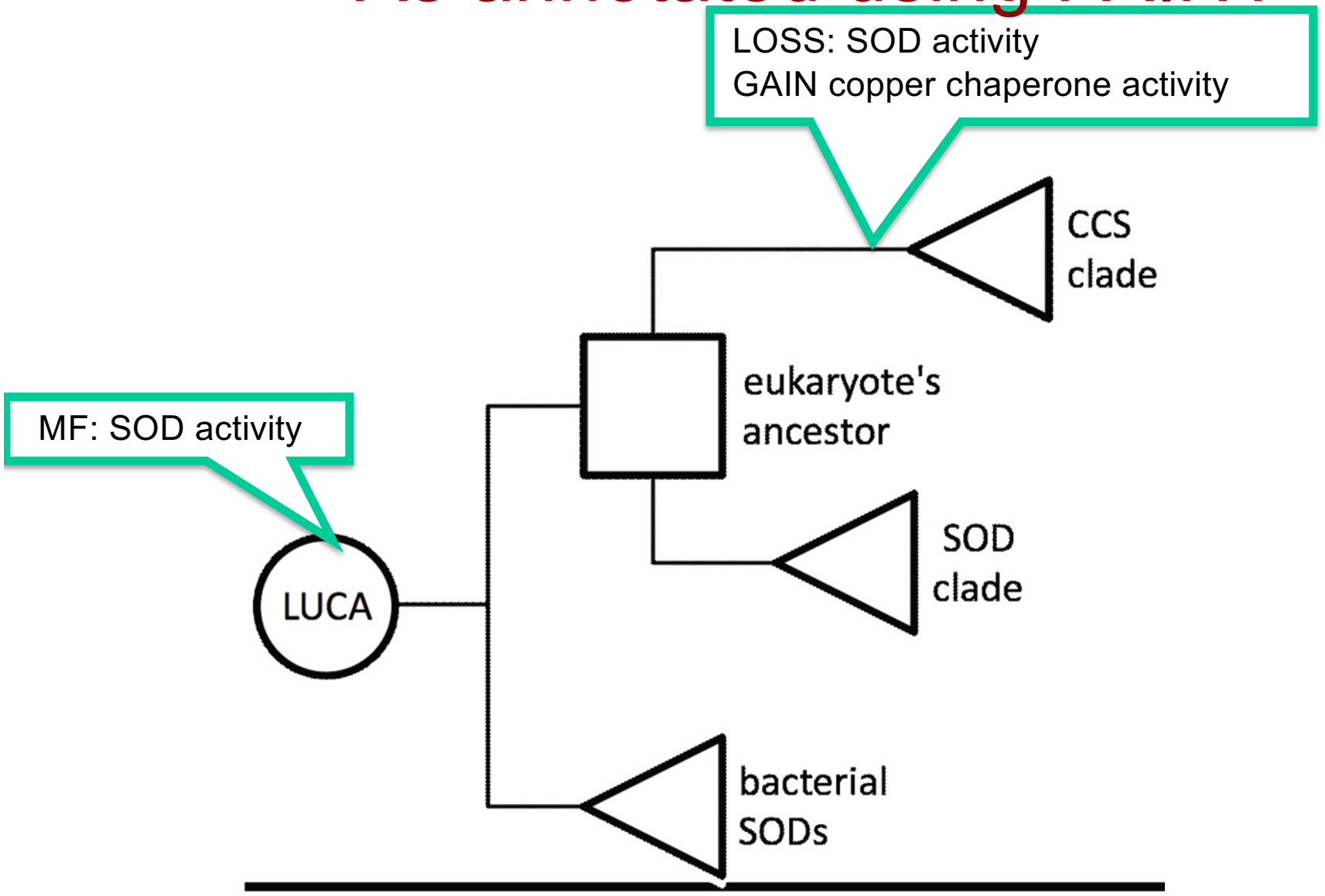

Evolutionary event type:
duplication
speciation
All nodes have persistent identifiers which are retained across different builds of the protein family trees.
Green indicates experimental
Black dot indicates direct experimental data.
dot indicates a more general functional class inferred from ontology
carboxylic ester hydrolase
Red indicates NOT function for the gene
cholinesterase
MOUSE_Ces2f
RAT_Ces2j
RAT_Ces1h
RAT_Ces2a
RAT_Ces2i
MOUSE_Ces2h
RAT_Ces2h
STRPU_STRPU_H3INM0
CAEEL_nlg-1
CAEBR_nlg-1
DROME_Nlg2
ANOGA_AgaP_AGAP002090
DROME_CG34127
DAPPU_DAPPUDRAFT_60001
DAPPU_DAPPUDRAFT_311742
DAPPU_DAPPUDRAFT_10046
DROME_Nlg4
ANOGA_AgaP_AGAP003115
DAPPU_DAPPUDRAFT_308699
DAPPU_DAPPUDRAFT_62803
HUMAN_BCHE
MACMU_BCHE
MOUSE_Bche
RAT_Bche
BOVIN_BCHE
PIG_LOC100624138
HORSE_BCHE
CANFA_BCHE
MONDO_BCHE
ANOCA_LOC100562739
ANOCA_LOC100562140
ANOCA_LOC100562541



- Inheritance of functions through tree, unless following a loss
- If a corresponding experimental annotation is not already present, these are PREDICTIONS

Gaudet, P., et al. (2011). Phylogenetic-based propagation of functional annotations within the Gene Ontology consortium. *Briefings in Bioinformatics*, 12(5), 449–62. doi:10.1093/bib/bbr042

Superoxide dismutase tree
As annotated using PAINT

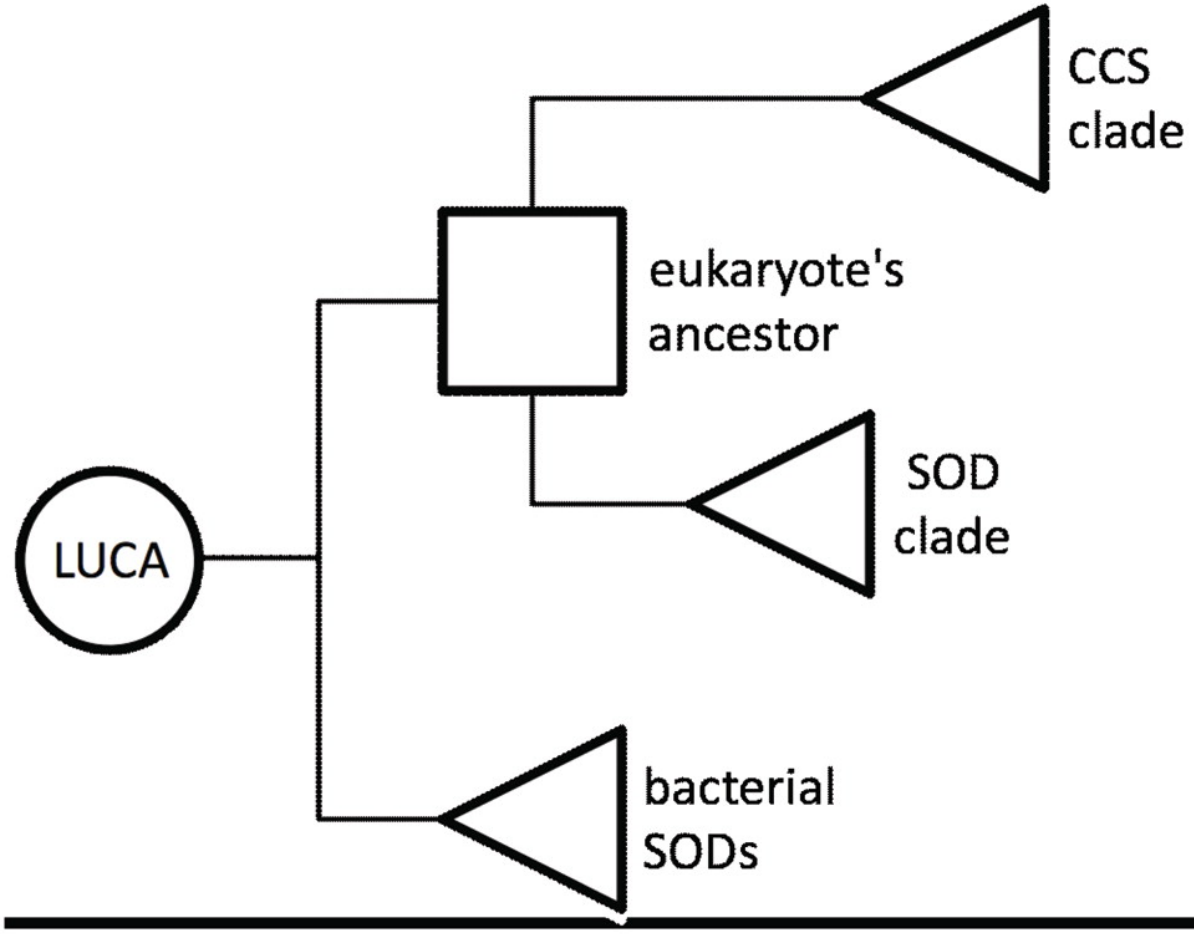


Superoxide dismutase family

Most members of this family are in the SOD clades

So the family is annotated by InterPro2GO as SOD activity

This is statistically accurate, but is incorrect for CCS clade members



Bottom line

- Experimental evidence codes remain the “gold standard”
  - BUT only available for a small subset of well-studied organisms
  - NOTE: some experimental codes indicate weaker evidence such as HTP and IEP
- Homology-inferred annotations are generally accurate
  - Especially IBA annotations

Where to get the data (for annotated genomes)

- GO annotations
  - Gene Ontology http://geneontology.org/
- Pathway data in SBML format
  - Pathway Commons https://www.pathwaycommons.org/

- For any analysis, make sure you note the version number and download date, as these resources are always being updated and analysis results may change from version to version

For unannotated genomes

- Use InterProScan to get InterPro2GO annotations
  - https://www.ebi.ac.uk/interpro/download.html
- Use TreeGrafter to get GO Phylogenetic Annotation project annotations
  - https://github.com/pantherdb/TreeGrafter

GO enrichment analysis

- Introduction to enrichment analysis
- Annotation sets
- Types of statistical tests
- Overrepresentation analysis using GO/PANTHER
  - Overrepresentation test
  - Enrichment test

Interpreting high-throughput "omics"/genomics experiments

- You've done a genome-wide experiment
  - Disease association study over 100 M distinct genomic variant sites
  - RNA-seq experiment that quantitates changes in tens of thousands of genes/splice forms
  - Etc.
- How do you interpret a huge number of individual measurements, in terms of the underlying biology?
- The main approach is "enrichment analysis", AKA "pathway analysis"

Enrichment analysis

- Uses ***known information*** about gene function
  - are any statistical trends in the kinds of FUNCTIONS of the genes that are changed in the experiment?
- Hypothesis: genes in the same biological subsystem ("module" or "pathway") tend to be coordinately regulated, or have similar biological effects when perturbed

Common enrichment analysis variations

- Different statistical tests
  - Require different data
- Different "annotation sets"
  - Appropriate sets depend on biological question, but most "omics" data analysis looks for correlated changes across groups of genes that may function together: pathways and GO biological processes
- How do they compare?
- If there are differences, don't just choose the one that you'd prefer to be true, examine the results to understand them

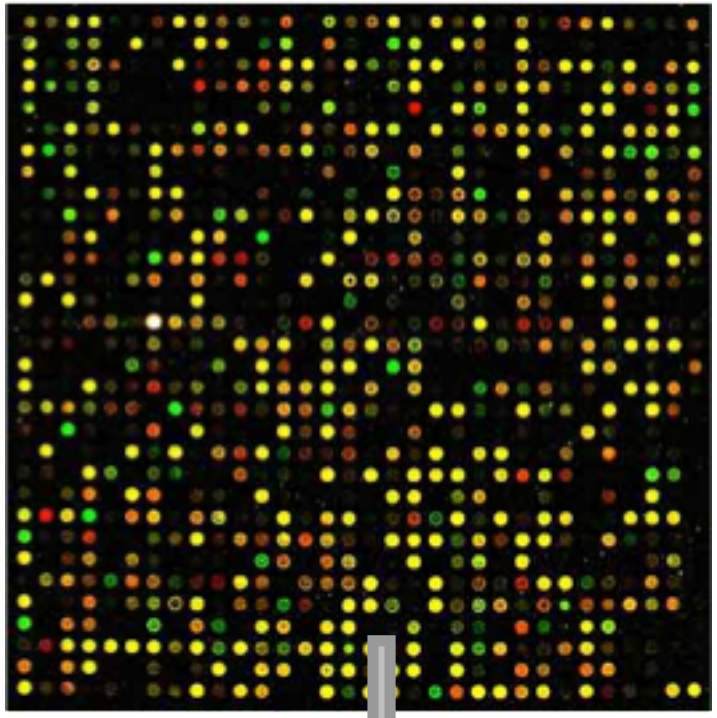
Two main types of test

- "Overrepresentation"
  - In my list of genes, are any functional classes found more often than expected?
- "Enrichment" (e.g. GSEA)
  - No list. For every gene in a large-scale experiment, a value is measured and computed.
  - Do the genes in a particular functional class have a distribution of values that is different from the expected distribution?

Overrepresentation test

- Input
  - A list of genes of interest
  - Optional but recommended: a “reference” list of genes from which the first list was chosen from
    - E.g. all genes with measurable expression in the experiment
- Output
  - Enrichment/depletion: which classes (e.g. pathways) show more (fewer) genes in the list than expected by chance
  - P-value: the probability that the observed enrichment/depletion is significantly different from the null hypothesis of NO ENRICHMENT/DEPLETION

Overrepresentation test



Your gene list of interest

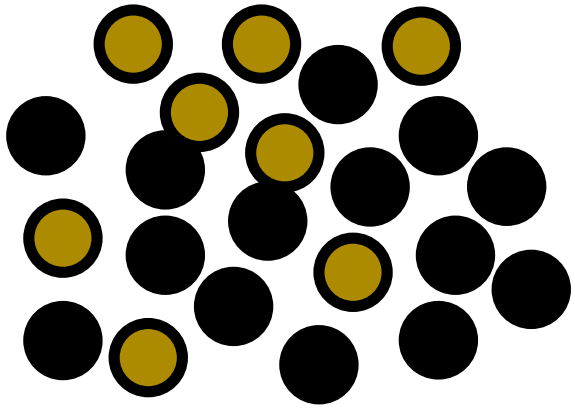
Reference gene list

Need to define:

Gene list(s) of interest
Reference gene list

Overrepresentation Test

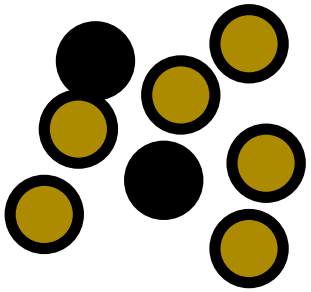
Reference gene list



Genes annotated with a given GO term
Genes not annotated with a given GO term

Is the given annotation class over- or under- represented compared to a reference?

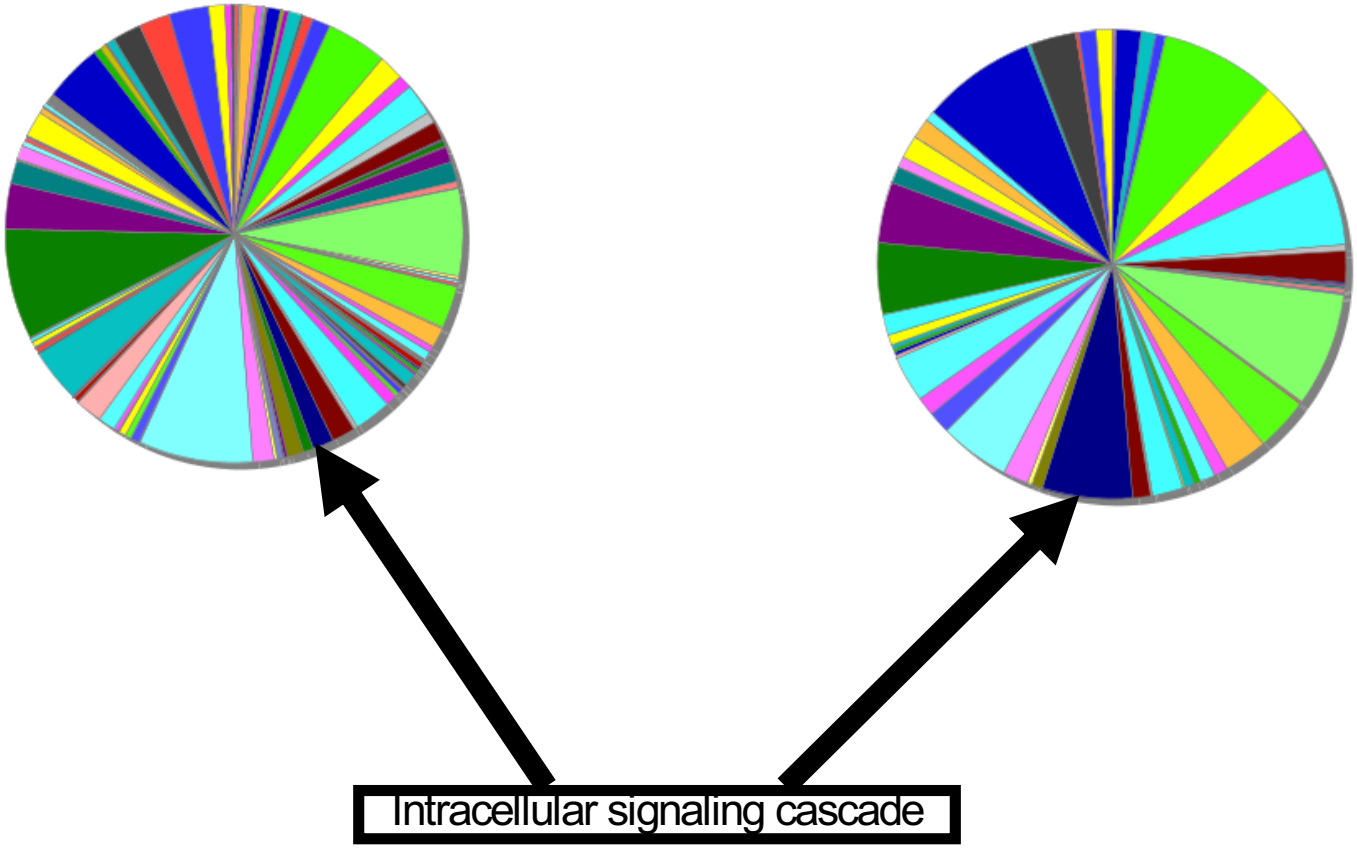
Your gene list of interest



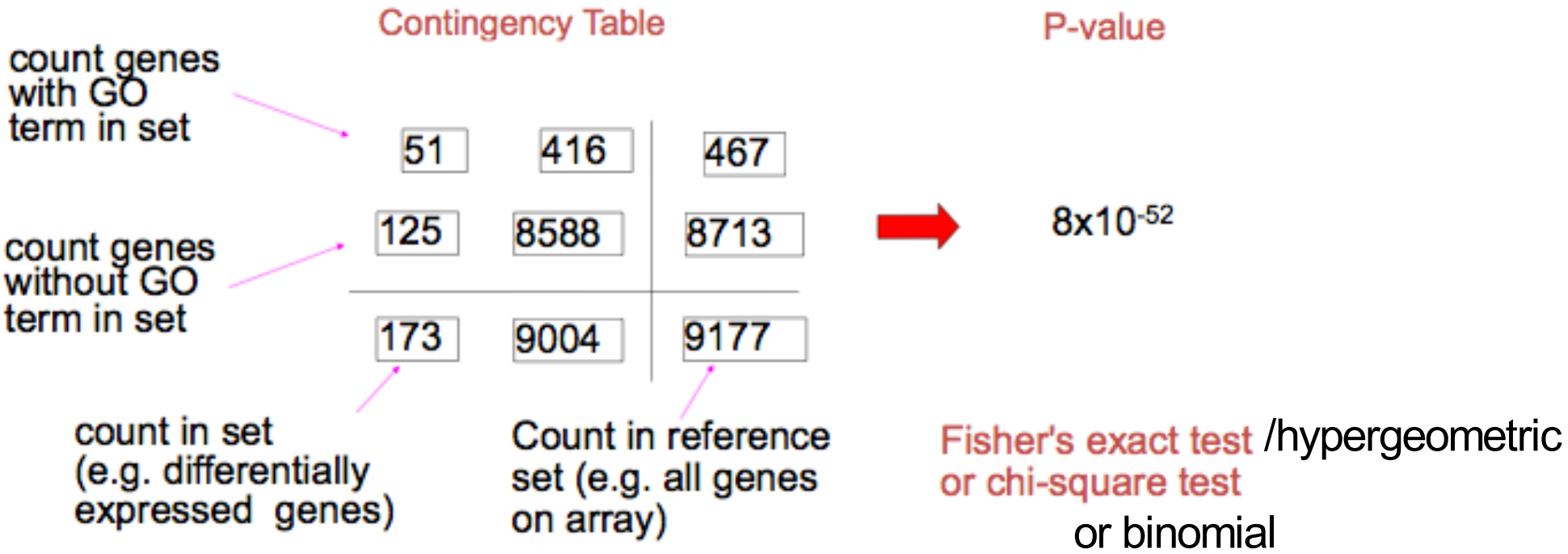
For each annotation term, are there more (overrepresented) or less (underrepresented) in the list of interest than expected by chance?

All genes on the reference gene list

Genes upregulated in tumor sample



Over (under) representation test example



Gene Ontology Consortium | Home | Documentation | Downloads | Community | Tools | About

Enrichment analysis

CG10469
RpL14
RpL31
RpL28
CG8791

biological process

Drosophila melanogaster

Submit

Advanced options / Help

Powered by PANTHER

Statistics

Actual RNA-seq experiment in Drosophila comparing wildtype to a Piwi mutant (this list is of genes down more than 2-fold in the mutant)

gy Consortium

ucts...

Download ontology

Gene Ontology: the framework for the model of biology. The GO defines concepts/classes used to describe gene function, and relationships between these concepts. It classifies functions along three aspects:

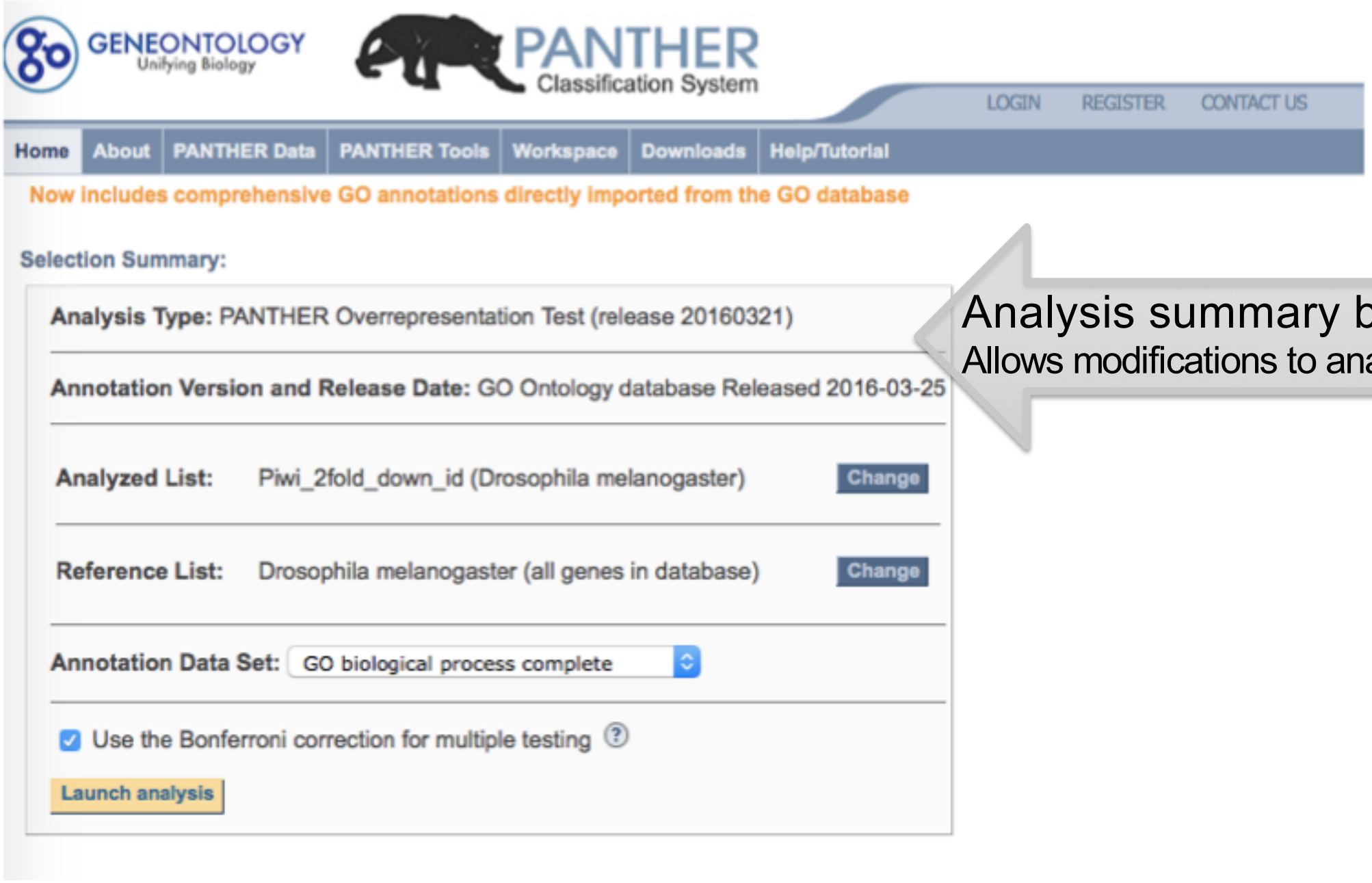
molecular function

Annotations

Download annotations (standard files)

Filter and download (customizable files <10k lines)

GO annotations: the model of biology. Annotations are statements describing the functions of specific genes, using concepts in the Gene Ontology. The simplest and most common annotation links one gene to one function, e.g. FZD4 + Wnt


GENEONTOLOGY
Unifying Biology
PANTHER
Classification System
LOGIN
REGISTER
CONTACT US
Home
About
PANTHER Data
PANTHER Tools
Workspace
Downloads
Help/Tutorial
Now includes comprehensive GO annotations directly imported from the GO database
Selection Summary:
Analysis Type: PANTHER Overrepresentation Test (release 20160321)
Annotation Version and Release Date: GO Ontology database Released 2016-03-25
Analyzed List: Piwi_2fold_down_id (Drosophila melanogaster)
Change
Reference List: Drosophila melanogaster (all genes in database)
Change
Annotation Data Set: GO biological process complete
Use the Bonferroni correction for multiple testing
Launch analysis
Analysis summary bo
Allows modifications to anal

Overrepresentation test results

Displaying only results with P<0.05; click here to display all results

GO biological process complete	Piwi_ref (REF) #	Piwi_2fold_down_id (▽ Hierarchy NEW!) #	expected	Fold Enrichment	+/-	P value
vitelline membrane formation involved in chorion	10	7	.47	14.89	+	1.44E-03
↳vitelline membrane formation	10	7	.47	14.89	+	1.44E-03
↳extracellular matrix assembly		7	.70	9.93	+	2.01E-02
↳egg coat formation		7	.47	14.89	+	1.44E-03
↳cellular component assembly involved in morphogenesis			6.25	3.84	+	7.08E-05
↳chorion-containing eggshell formation			4.89	5.52	+	3.42E-09
↳ovarian follicle cell development	271	34	12.74	2.67	+	6.04E-04
↳columnar/cuboidal epithelial cell development				2.66	+	6.56E-04
↳columnar/cuboidal epithelial cell differentiation				2.64	+	7.74E-04
↳epithelial cell differentiation	319	35	14.99	2.33	+	8.17E-03
↳epithelial cell development	29				+	2.74E-03
↳eggshell formation	105	27	4.93	5.47	+	4.24E-09
Unclassified	1013	53	47.61	1.11	+	0.00E00
cellular localization	591					4.07E-02
cellular protein modification process	659	8	30.97	.26	-	8.60E-04

in the reference list

observed

#expected

Fold enrichment

Overrepresent (+) or under-representation (-)

P value

Analysis Summary: Please report in publication

Analysis Type: PANTHER Overrepresentation Test (release 20160321)

Annotation Version and Release Date: GO Ontology database Released 2016-03-25

Analyzed List: Piwi_2fold_down_id (Drosophila melanogaster) Change

Reference List: Drosophila melanogaster (all genes in database) Change

Annotation Data Set: GO biological process complete

Use the Bonferroni correction for multiple testing

Add/change list

Results

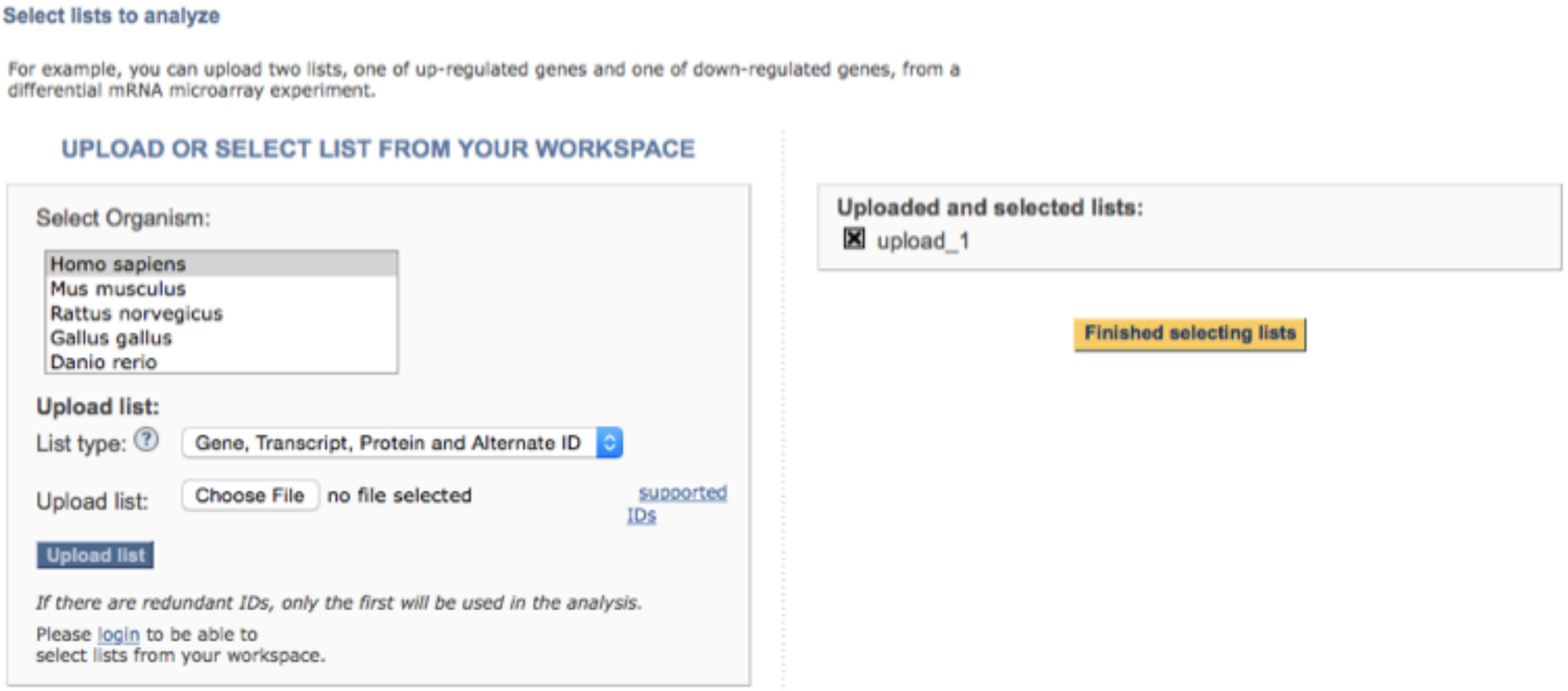
	Reference list	Piwi_2fold_down_id
Mapped IDs:	13690	300
Unmapped IDs:	0	125

Export results

Displaying only results with P<0.05; click here to display all results

	Drosophila melanogaster (REF)	Piwi_2fold_down_id (▽ Hierarchy NEW!)				
GO biological process complete	#	#	expected	Fold Enrichment	+/-	P value
vitelline membrane formation involved in chorion-containing eggshell formation	14	7	.31	22.82	+	1.04E-04

List can be changed, and up to 3 other lists can be added to the analysis



Analysis Summary: Please report in publication

Analysis Type: PANTHER Overrepresentation Test (release 20160321)

Annotation Version and Release Date: GO Ontology database Released 2016-03-25

Analyzed List: Piwi_2fold_down_id (Drosophila melanogaster) Change

Reference List: Drosophila melanogaster (all genes in database) Change

Annotation Data Set: GO biological process complete

Use the Bonferroni correction for multiple testing

Change reference list

Results

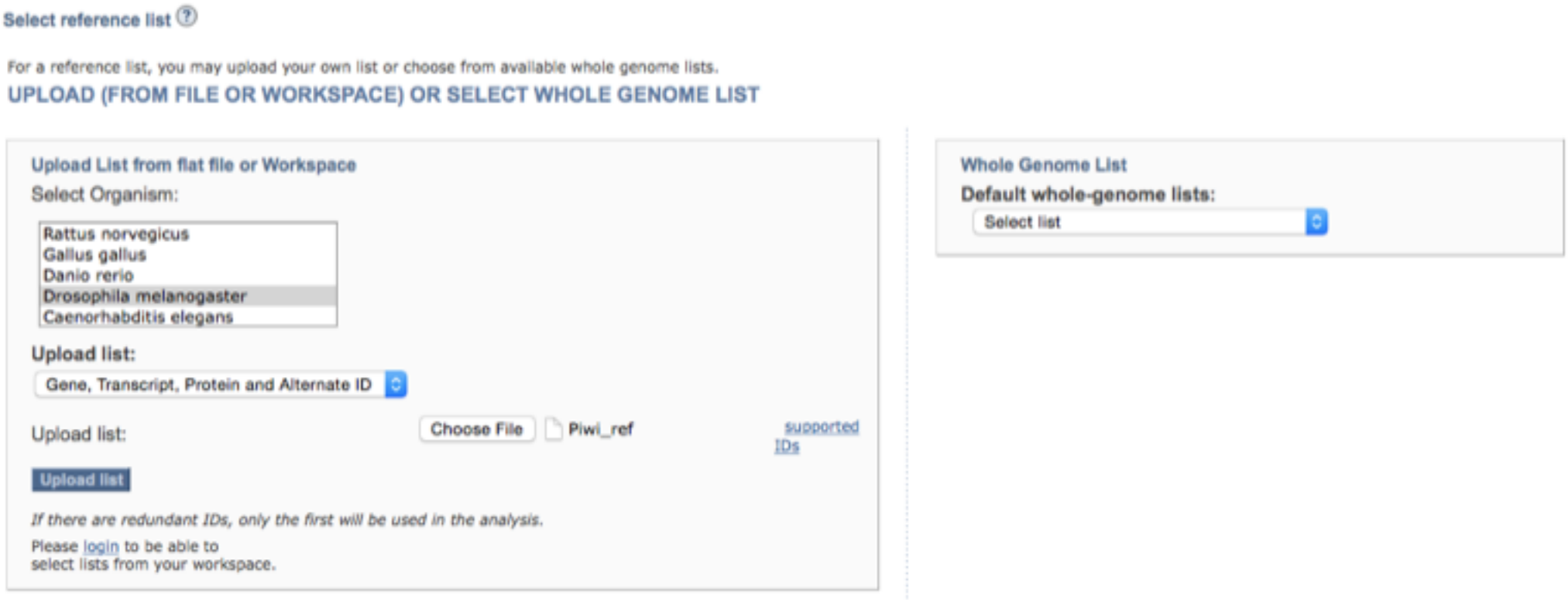
	Reference list	Piwi_2fold_down_id
Mapped IDs:	13690	300
Unmapped IDs:	0	125

Export results

Displaying only results with P<0.05; click here to display all results

	Drosophila melanogaster (REF)	Piwi_2fold_down_id (▽ Hierarchy NEW!)				
GO biological process complete	#	#	expected	Fold Enrichment	+/-	P value
vitelline membrane formation involved in chorion-containing eggshell formation	14	7	.31	22.82	+	1.04E-04

Reference list can be changed



Analysis Summary: Please report in publication

Analysis Type: PANTHER Overrepresentation Test (release 20160321)

Annotation Version and Release Date: GO Ontology database Released 2016-03-25

Analyzed List: Piwi_2fold_down_id (Drosophila melanogaster) Change

Reference List: Piwi_ref (Drosophila melanogaster) Change
There are duplicate IDs in the file. The unique set of IDs will be used.

Annotation Data Set: GO biological process complete

Use the Bonferroni correction for multiple testing

Results

	Reference list	Piwi_2fold_down_id
Mapped IDs:	6383	300
Unmapped IDs:	1529	125

Export results

Displaying only results with P<0.05; click here to display all results

	Piwi_ref (REF)	Piwi_2fold_down_id (▼ Hierarchy NEW!)				
GO biological process complete	#	#	expected	Fold Enrichment	+/-	P value
vitelline membrane formation involved in chorion-containing eggshell formation	10	7	.47	14.89	+	1.44E-03
↳vitelline membrane formation	10	7	.47	14.89	+	1.44E-03

Analysis Summary: Please report in publication

Analysis Type: PANTHER Overrepresentation Test (release 20160321)

Annotation Version and Release Date: GO Ontology database Released 2016-03-25

Analyzed List: Piwi_2fold_down_id (Drosophila melanogaster) Change

Reference List: Piwi_ref (Drosophila melanogaster) Change
There are duplicate IDs in the file. The unique set of IDs will be used.

Annotation Data Set: GO biological process complete

- PANTHER Pathways
- PANTHER GO-Slim Molecular Function
- PANTHER GO-Slim Biological Process
- PANTHER GO-Slim Cellular Component
- PANTHER Protein Class
- GO molecular function complete
- GO biological process complete
- GO cellular component complete

Use the Bonferroni

Change annotation set

Results

	Reference list	Piwi_2fold_down_id
Mapped IDs:	6383	300
Unmapped IDs:	1529	125

Export results

Displaying only results with P<0.05; click here to display all results

	Piwi_ref (REF)	Piwi_2fold_down_id (▽ Hierarchy NEW!)				
GO biological process complete	#	#	expected	Fold Enrichment	+/-	P value
vitelline membrane formation involved in chorion-containing eggshell formation	10	7	.47	14.89	+	1.44E-03
↳vitelline membrane formation	10	7	.47	14.89	+	1.44E-03

Displaying only results with P<0.05; click here to display all results

	Piwi_ref (REF)	Piwi_2fold_down_id (▽ Hierarchy NEW! ?)				
GO cellular component complete	#	#	expected	Fold Enrichment	+/-	P value
chorion	26	19	1.22	15.55	+	4.08E-14
↳external encapsulating structure	34	22	1.60	13.77	+	1.72E-15
cytosolic large ribosomal subunit	48	11	2.26	4.88	+	1.29E-02
↳cytosolic part	109	16	5.12	3.12	+	4.39E-02
↳intracellular part	3448	109	162.06	.67	-	3.22E-07
↳intracellular	3485	111	163.79	.68	-	3.95E-07
↳intracellular organelle part	1763	40	82.86	.48	-	1.33E-06
↳intracellular organelle	2870	93	134.89	.69	-	3.10E-04
↳organelle	2895	96	136.06	.71	-	9.56E-04
↳organelle part	1795	41	84.36	.49	-	1.12E-06
↳macromolecular complex	1765	48	82.95	.58	-	8.22E-04
↳cytosolic ribosome	83	16	3.90	4.10	+	1.66E-03
Unclassified	1460	88	68.62	1.28	+	0.00E00
endomembrane system	586	10	27.54	.36	-	3.83E-02
nucleoplasm	336	2	15.79	< 0.2	-	8.21E-03
↳nuclear lumen	514	8	24.16	.33	-	4.90E-02
↳intracellular organelle lumen	655	11	30.78	.36	-	1.08E-02
↳organelle lumen	655	11	30.78	.36	-	1.08E-02
↳membrane-enclosed lumen	664	11	31.21	.35	-	7.91E-03
↳nuclear part	797	8	37.46	.21	-	7.66E-07
↳nucleus	1505	35	70.73	.49	-	7.58E-05

Analysis Summary: Please report in publication

Analysis Type: PANTHER Overrepresentation Test (release 20150430)

Annotation Version and Release Date: GO Ontology database Released 2015-08-06

Analyzed List: upload_1 (Drosophila melanogaster) Change

Reference List: Drosophila melanogaster (all genes in database) Change

Annotation Data Set: GO biological process complete

Use the Bonferroni correction for multiple testing

Results

	Reference list	upload_1
Mapped IDs:	13690	300
Unmapped IDs:	0	125

Export results

View unmapped ID's

Displaying only results with P<0.05; click here to display all results

	Drosophila melanogaster (REF)	upload_1				
GO biological process complete	#	#	expected	Fold Enrichment	+/-	P value
vitelline membrane formation involved in chorion-containing eggshell formation	14	7	.31	> 5	+	9.69E-05
vitelline membrane formation	14	7	.31	> 5	+	9.69E-05
egg coat formation	14	7	.31	> 5	+	9.69E-05
extracellular matrix assembly	17	7	.37	> 5	+	3.57E-04
chorion-containing eggshell formation	124	27	2.72	> 5	+	3.53E-15

Unmapped IDs:

ID
CG7722
CG13114
snoRNA:Me28S-A982b
snoRNA:Me28S-C788a
snoRNA:Me28S-C3227b
snoRNA:Me28S-A992
CG8539
CG32972
CG2052
Fcp26Ac
CG15324
CG13636
CG11381
CG32774
snoRNA:Or-CD10
snoRNA:Me18S-U1356b

Gene set enrichment (GSEA)

Gene ID	P-value
Gene 1	1.54e-5
Gene 2	4.20e-2
Gene 3	2.34e-7
Gene 4	0.00
Gene 5	1.09e-18
Gene 6	0.00
….	
….	
….	
….	
….	
….	
Gene 19,997	7.54e-12
Gene 19,998	4.31e-5
Gene 19,999	2.62e-2
Gene 20,000	1.29e-5

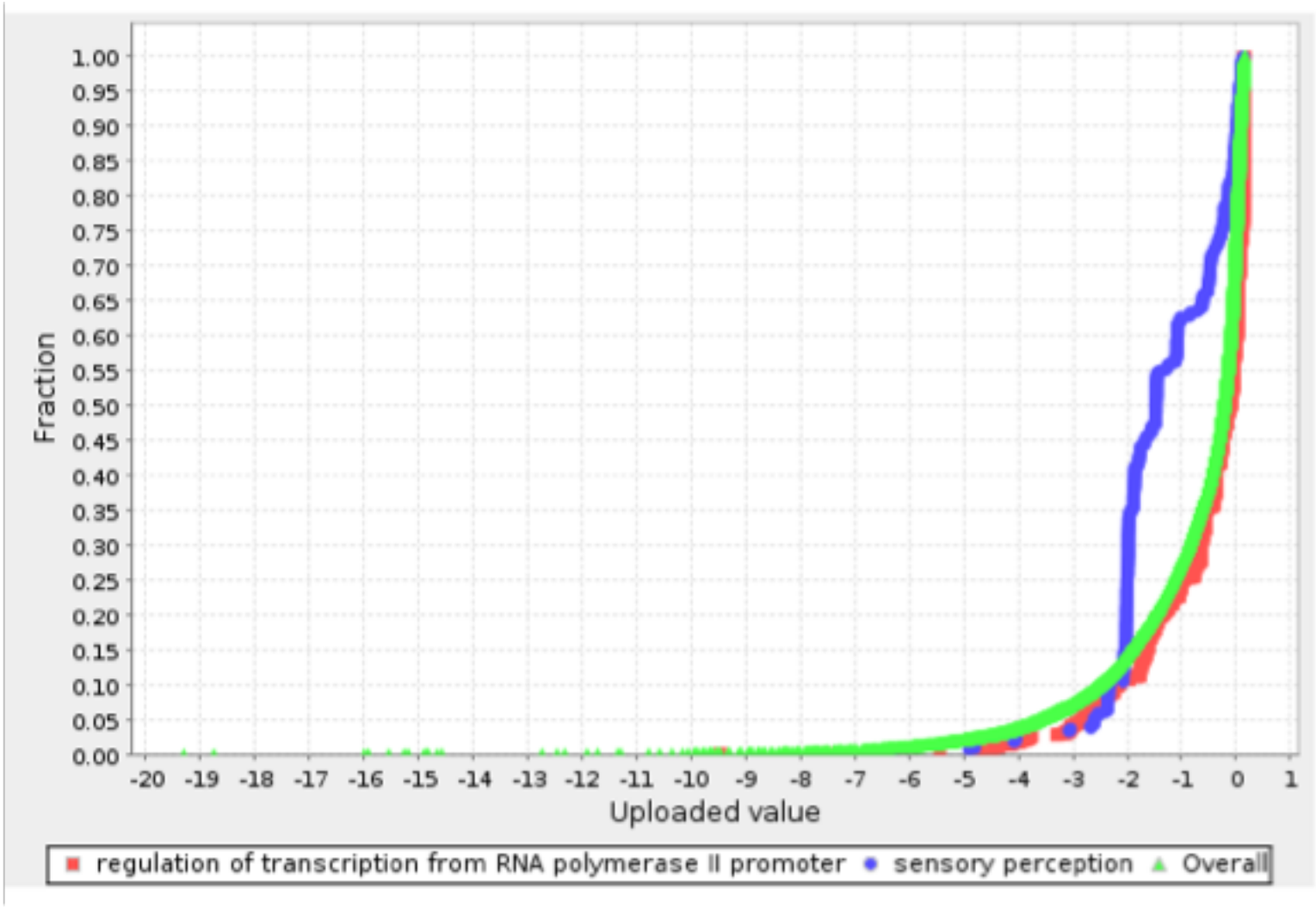
Gene ID	P-value
Gene 5	1.09e-18
….	
Gene 19,997	7.54e-12
….	
Gene 3	2.34e-7
….	
Gene 20,000	1.29e-5
Gene 1	1.54e-5
….	
Gene 19,998	4.31e-5
….	
Gene 19,999	2.62e-2
Gene 2	4.20e-2
….	
Gene 4	0.00
Gene 6	0.00

Gene ID	P-value
….	
Gene 3	2.34e-7
….	
Gene 19,998	4.31e-5
….	
Gene 2	4.20e-2
….	

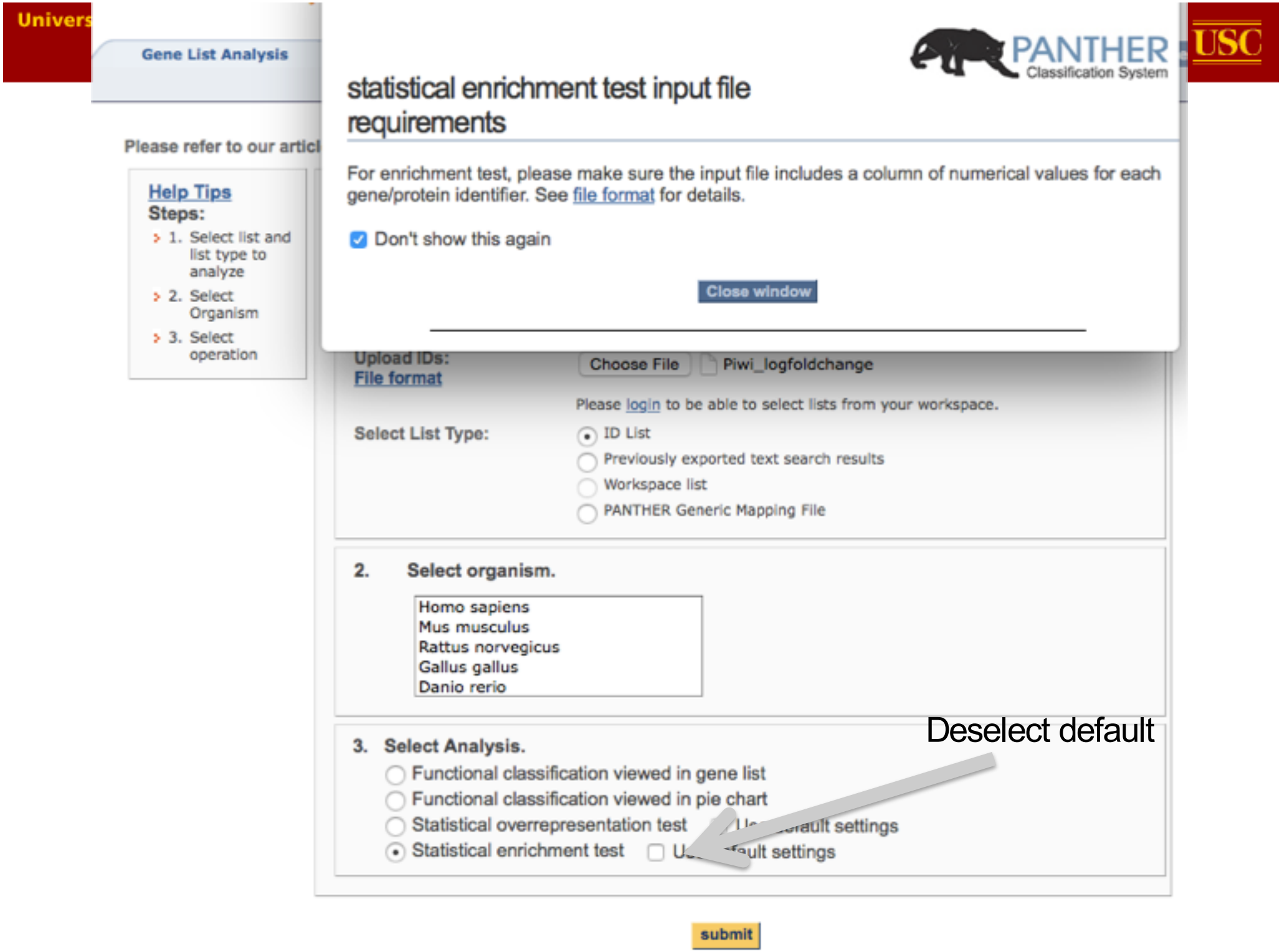
Enrichment test

- Input
  - A list of genes (as many as possible, to get good statistics!) and a quantitative value for each gene (e.g. fold change)
- Output
  - The probability that the distribution of values for the **genes in a given GO class** was drawn randomly from the distribution of values for **all genes**

Gene set enrichment (GSEA)



Statistically compares distribution of values for genes in a given annotation class, with distribution for all genes


Univers
USC
Gene List Analysis
PANTHER Classification System
statistical enrichment test input file requirements
For enrichment test, please make sure the input file includes a column of numerical values for each gene/protein identifier. See file format for details.
Don't show this again
Close window
Please refer to our articl
Help Tips
Steps:
1. Select list and list type to analyze
2. Select Organism
3. Select operation
Upload IDs:
File format
Choose File
Piwi_logfoldchange
Please login to be able to select lists from your workspace.
Select List Type:
ID List
Previously exported text search results
Workspace list
PANTHER Generic Mapping File
2. Select organism.
Homo sapiens
Mus musculus
Rattus norvegicus
Gallus gallus
Danio rerio
3. Select Analysis.
Functional classification viewed in gene list
Functional classification viewed in pie chart
Statistical overrepresentation test
Statistical enrichment test
Deselect default
submit

Enrichment test summary

Analysis Summary: Please report in publication

Analysis Type: PANTHER Enrichment Test (release 20141219)

Annotation Version and Release Date: GO Ontology database Released 2016-03-25

Analyzed List: Piwi_logfoldchange (Drosophila melanogaster) Change

There are duplicate IDs in the file. For duplicates, the first id/value pair in the file will be used.

Annotation Data Set: GO biological process complete

☑ Use the Bonferroni correction for multiple testing

Results

Analysis details:

Mapped IDs: 6383

Unmapped IDs: 1529

Graph selected categories | Export results

Displaying only results with P<0.05; click here to display all results (Hierarchy NEW!)

GO biological process complete	#	+/-	P value
translation (GO:0006412)	268	-	0.00E00
↳cellular macromolecule biosynthetic process (GO:0034645)	625	-	9.65E-12
↳macromolecule biosynthetic process (GO:0009059)	628	-	1.50E-11
↳organic substance metabolic process (GO:0071704)	2460	-	7.91E-06
↳metabolic process (GO:0008152)	2777	-	7.99E-07
↳organic substance biosynthetic process (GO:1901576)	887	-	0.00E00
↳biosynthetic process (GO:0009058)	927	-	0.00E00
↳cellular biosynthetic process (GO:0044249)	877	-	0.00E00
↳cellular metabolic process (GO:0044237)	2291	-	6.17E-06
↳gene expression (GO:0010467)	802	-	0.00E00
↳protein metabolic process (GO:0019538)	1170	-	1.25E-02
↳primary metabolic process (GO:0044238)	2301	-	3.12E-07
↳peptide biosynthetic process (GO:0043043)	273	-	0.00E00

mitochondrial electron transport, NADH to ubiquinone (GO:0006120)	22	-	2.52E-03
↳mitochondrial ATP synthesis coupled electron transport (GO:0042775)	40	-	7.62E-07
↳ATP synthesis coupled electron transport (GO:0042773)	42	-	7.46E-07
↳oxidative phosphorylation (GO:0006119)	46	-	9.76E-08
↳generation of precursor metabolites and energy (GO:0006091)	106	-	2.54E-13
↳cellular metabolic process (GO:0044237)	2291	-	6.17E-06
↳metabolic process (GO:0008152)	2777	-	7.99E-07
DNA repair (GO:0006281)	116	+	3.24E-03
↳cellular metabolic process (GO:0044237)	2291	-	6.17E-06
↳metabolic process (GO:0008152)	2777	-	7.99E-07
↳organic substance metabolic process (GO:0071704)	2460	-	7.91E-06
↳nucleobase-containing compound metabolic process (GO:0006139)	972	-	4.74E-02

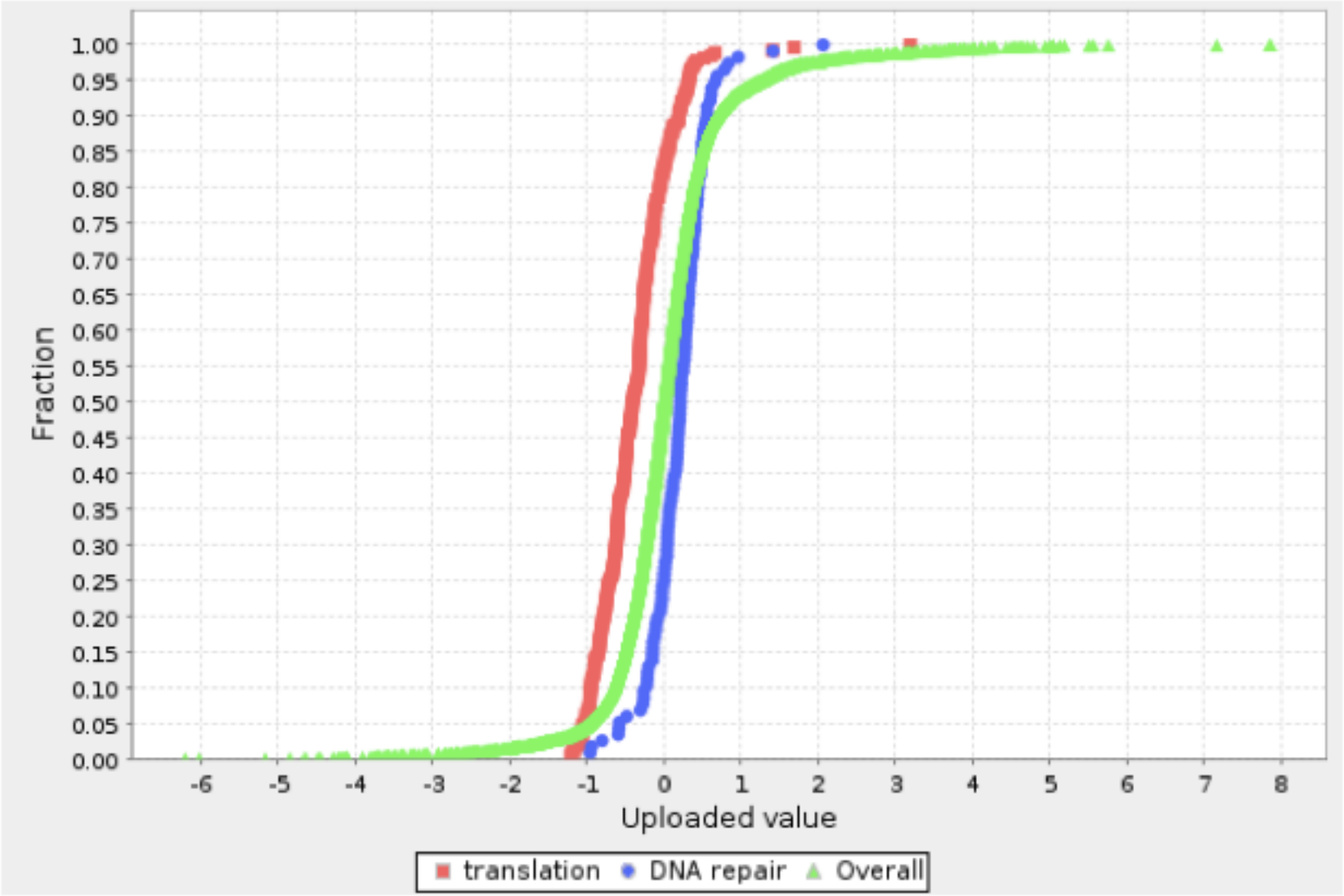
Graphing distribution for different classes helps interpret results

Graph selected categories | Export results

Displaying only results with P<0.05; click here to display all results (Δ Hierarchy NEW!)

GO biological process complete	#	+/-	P value
translation (GO:0006412)	268	-	0.00E00
↳cellular macromolecule biosynthetic process (GO:0034645)	625	-	9.65E-12
↳macromolecule biosynthetic process (GO:0009059)	628	-	1.50E-11
↳organic substance metabolic process (GO:0071704)	2460	-	7.91E-06
↳metabolic process (GO:0008152)	2777	-	7.99E-07
↳organic substance biosynthetic process (GO:1901576)	887	-	0.00E00
↳biosynthetic process (GO:0009058)	927	-	0.00E00
↳cellular biosynthetic process (GO:0044249)	877	-	0.00E00
↳cellular metabolic process (GO:0044237)	2291	-	6.17E-06
↳gene expression (GO:0010467)	802	-	0.00E00
↳protein metabolic process (GO:0019538)	1170	-	1.25E-02

Graphing distribution for different classes helps interpret results



Summary of best practices

- Enable others to reproduce your results
  - Report version of data, and tool
  - And provide data, of course
- Improving analysis (general)
  - Make sure GO annotations are up-to-date
  - For most tools, analysis is gene-centric– ensure that your data are also for individual genes (not splice forms, etc)
  - Check input identifiers that did not map to the database
    - These will be ignored in the analysis
    - Can these be fixed using alternative identifiers?
  - Are enriched classes related? (consider GO structure)
  - Consider ALL results, not just the ones you want to see
    - Explore the genes in enriched classes that are unexpected

For overrepresentation tests

- Use appropriate reference list (what could have been observed)
- Fold enrichment can be more informative than P-value, as long as the P-value is significant
  - P-value can depend on size of the gene set

For enrichment tests

- Upload quantitative values for as many genes as possible
- Graph distributions for enriched classes to help interpretation