Gene function annotation

Paul D. Thomas, Ph.D.
University of Southern California
October 2014

What is function annotation?

- The formal answer to the question: what does this gene do?
- The association between: **a description of biological function, in electronic form**, with **a biological sequence** (gene or gene product e.g. protein or functional RNA)

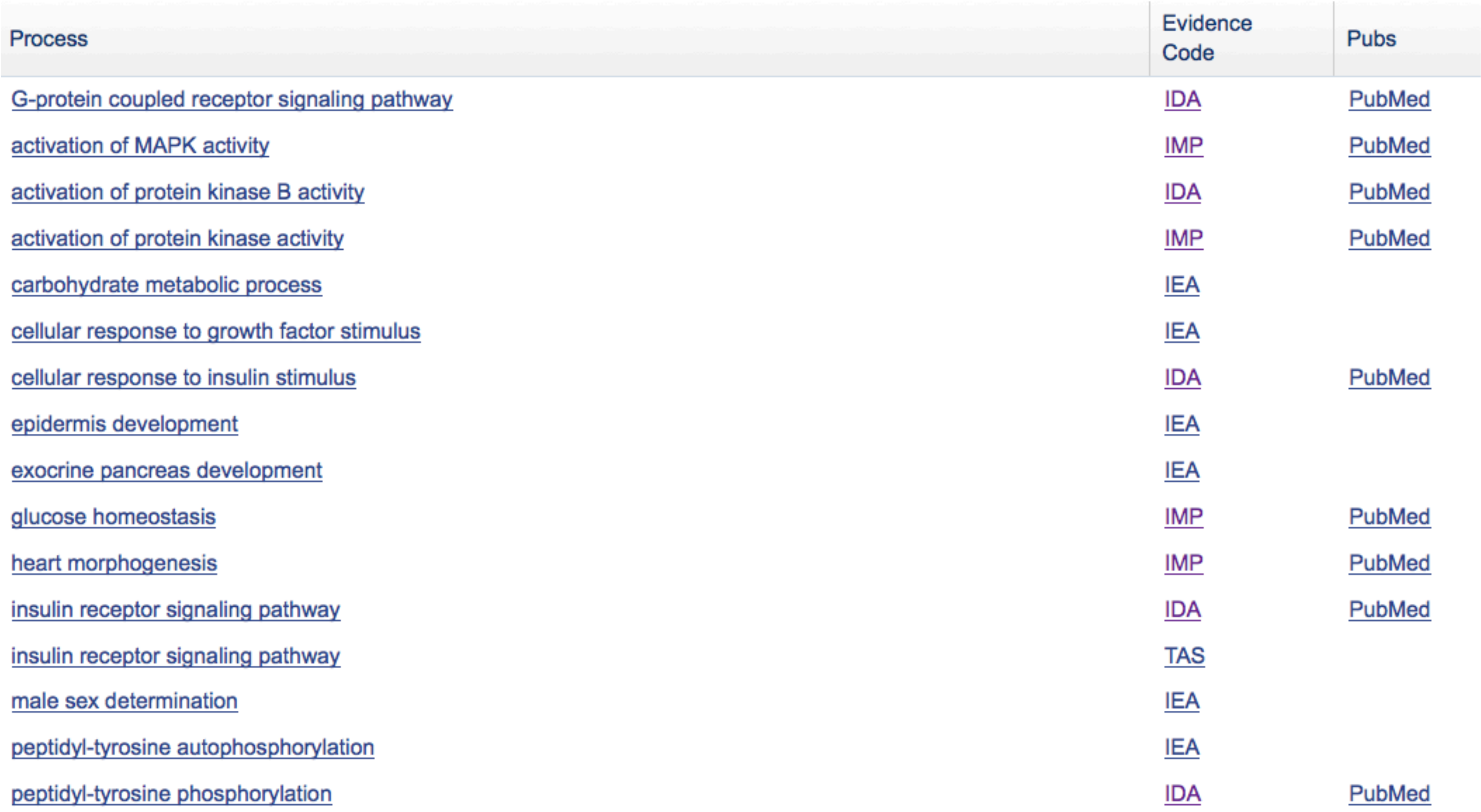
In this lecture

- Introduction to databases of *gene function*
- 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

Ontologies

- A formal structuring of knowledge
- Consists of concepts and relations
- Concept (entity, class, term): a class of things in the real world
  - Continuant (thing that exists)
  - Occurrent (process)
- Relation: a type of relationship between concepts
  - E.g. is_a, part_of

Entrez Gene: INSR



Process	Evidence Code	Pubs
G-protein coupled receptor signaling pathway	IDA	PubMed
activation of MAPK activity	IMP	PubMed
activation of protein kinase B activity	IDA	PubMed
activation of protein kinase activity	IMP	PubMed
carbohydrate metabolic process	IEA	
cellular response to growth factor stimulus	IEA	
cellular response to insulin stimulus	IDA	PubMed
epidermis development	IEA	
exocrine pancreas development	IEA	
glucose homeostasis	IMP	PubMed
heart morphogenesis	IMP	PubMed
insulin receptor signaling pathway	IDA	PubMed
insulin receptor signaling pathway	TAS	
male sex determination	IEA	
peptidyl-tyrosine autophosphorylation	IEA	
peptidyl-tyrosine phosphorylation	IDA	PubMed

Gene function annotation sources

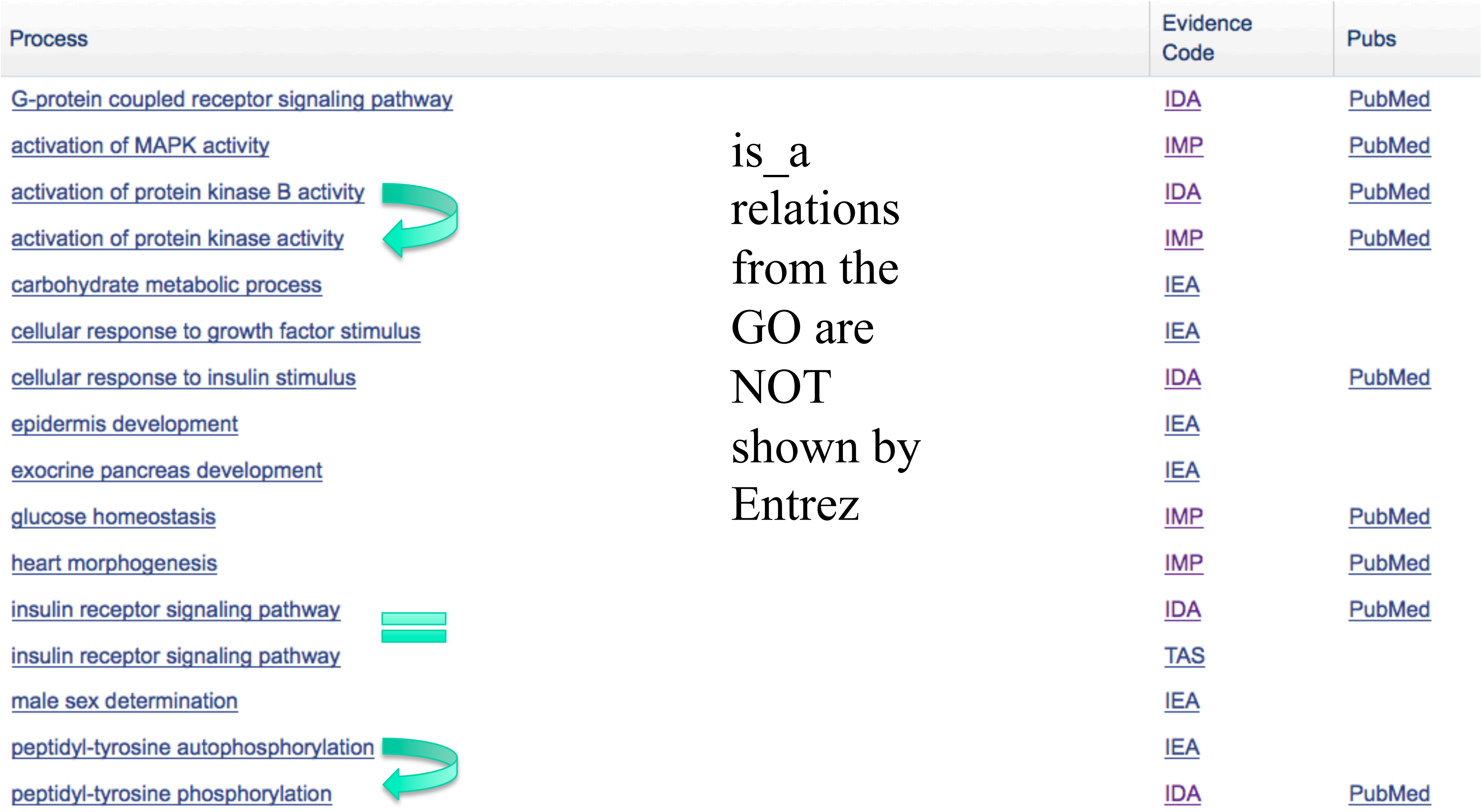
- Gene Ontology (GO)
- Pathway databases
  - Reactome
  - PANTHER
  - BioCyc
  - KEGG (kind of)

Thomas PD, Lewis SE, Mi H, Ontology annotation: mapping genomic regions to biological function, Curr. Opin. Biol. Chem.11:1-8 (2007)

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

Entrez Gene: INSR



Process	Evidence Code	Pubs
G-protein coupled receptor signaling pathway	IDA	PubMed
activation of MAPK activity	IMP	PubMed
activation of protein kinase B activity	IDA	PubMed
activation of protein kinase activity	IMP	PubMed
carbohydrate metabolic process	IEA	
cellular response to growth factor stimulus	IEA	
cellular response to insulin stimulus	IDA	PubMed
epidermis development	IEA	
exocrine pancreas development	IEA	
glucose homeostasis	IMP	PubMed
heart morphogenesis	IMP	PubMed
insulin receptor signaling pathway	IDA	PubMed
insulin receptor signaling pathway	TAS	
male sex determination	IEA	
peptidyl-tyrosine autophosphorylation	IEA	
peptidyl-tyrosine phosphorylation	IDA	PubMed

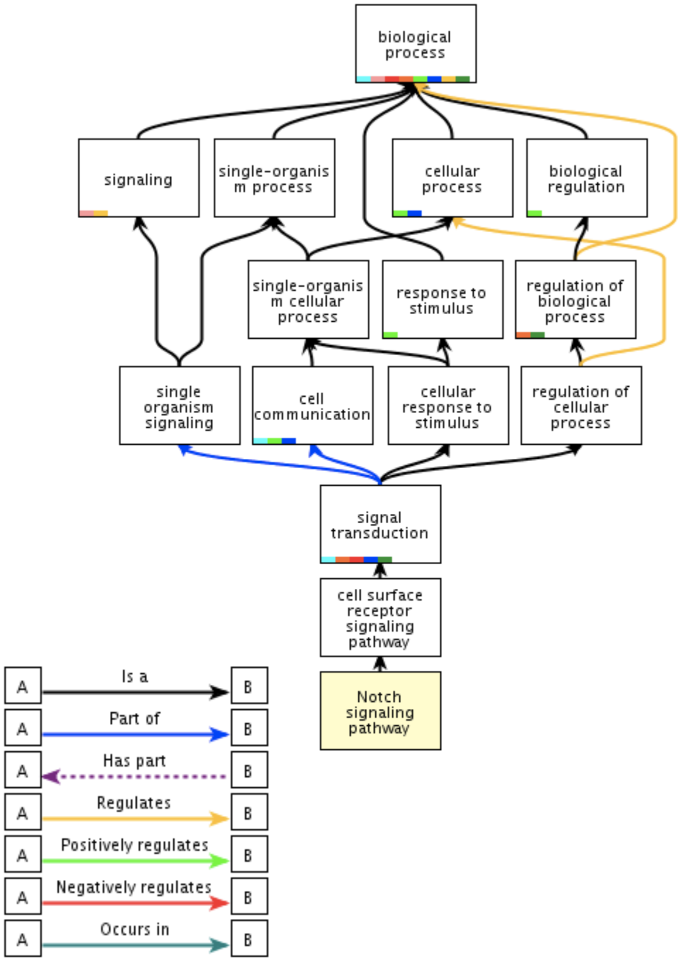
is_a relations from the GO are NOT shown by Entrez

Pathway representations

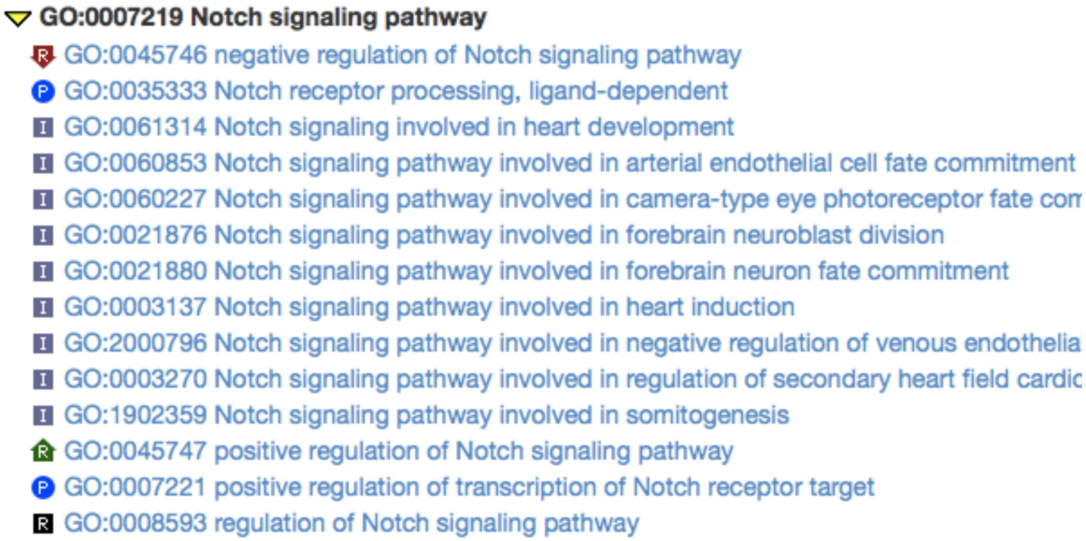
- Point of view from the molecular reaction
  - Generalized to include covalent and noncovalent (e.g. binding) reactions
- Concepts are reaction, molecule classes
- Relations are between molecule classes and reactions
  - Catalyst
  - Reactant
  - Product
- Top level structure provided by SBML, BioPAX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

Notch signaling pathway in GO

Relations to
more general classes



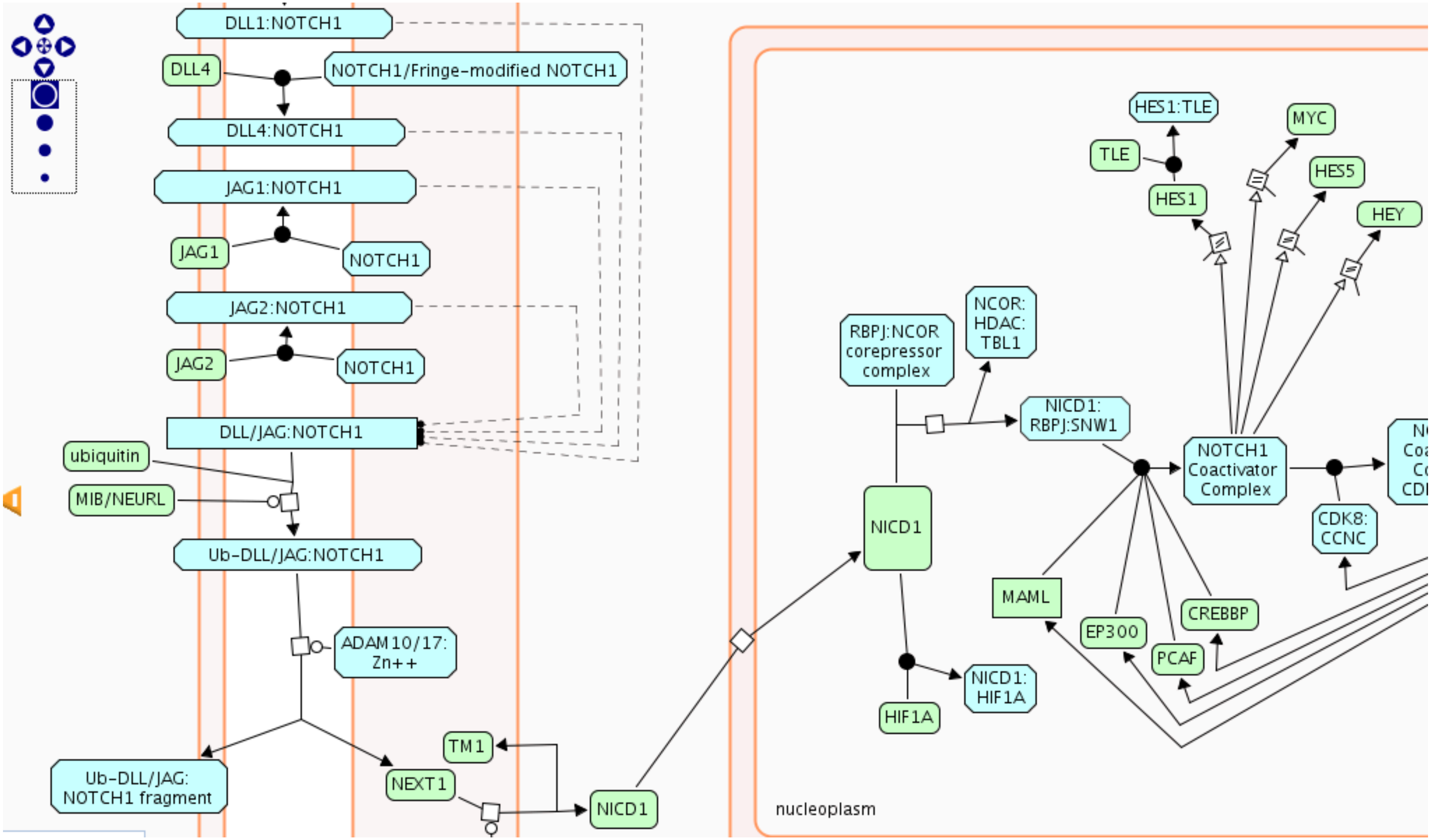
Relations to
more specific classes



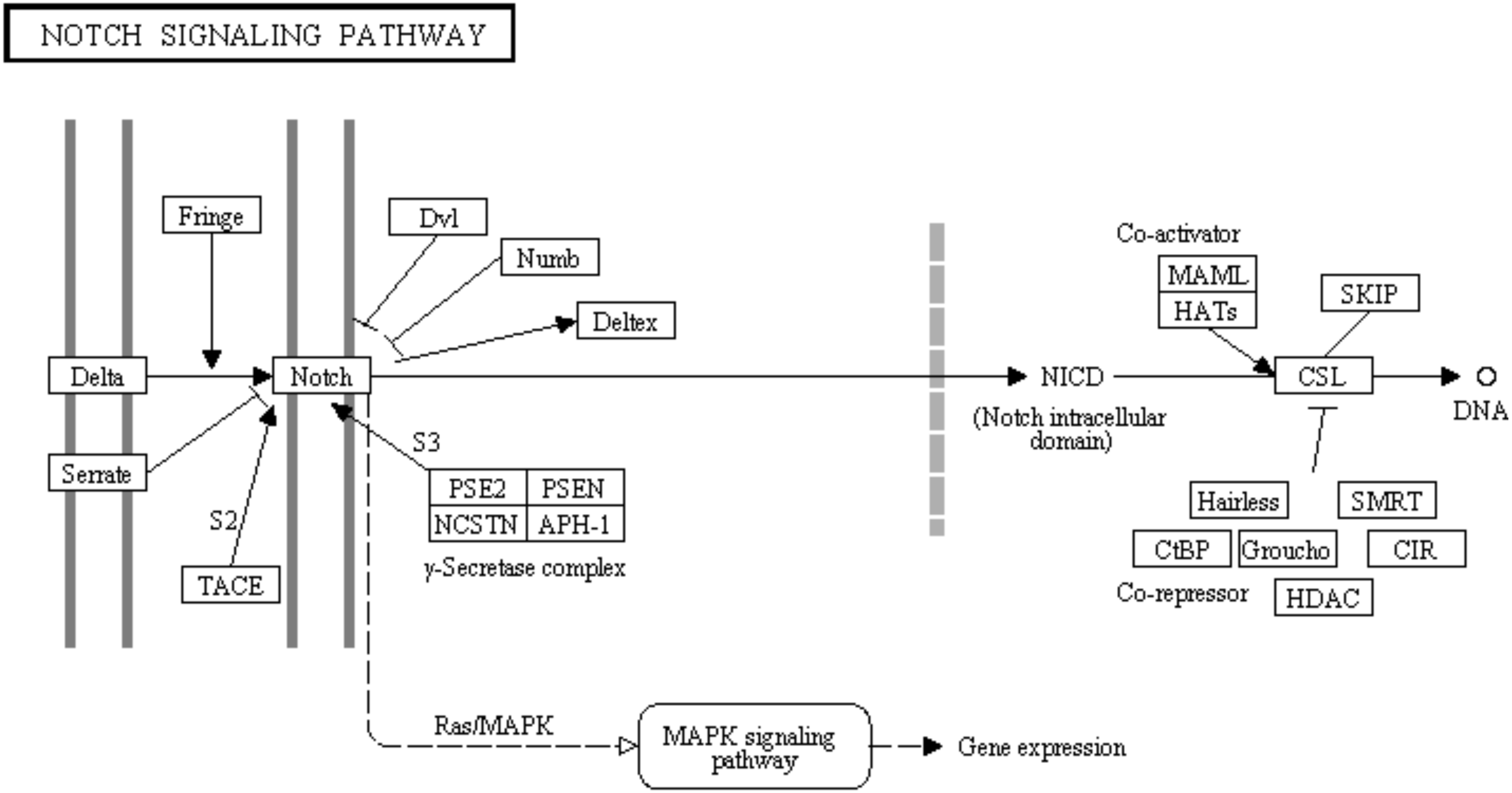
GO:0007219 Notch signaling pathway

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

Notch signaling in Reactome



Notch signaling in KEGG



GO vs. pathway representations

- GO is a simpler representation of molecular events, but has more biological context
- Pathway representations are more detailed at the molecular level, and can capture dependencies and temporal series

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

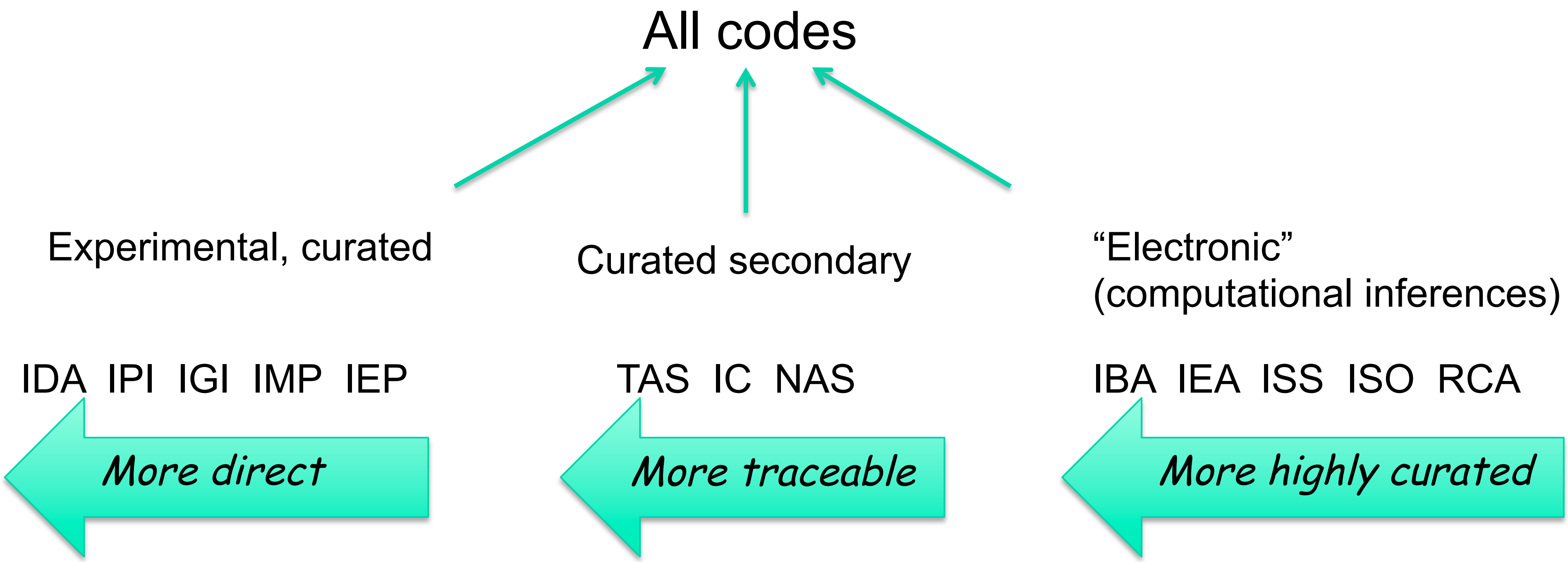
Common qualifiers

- 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

Evidence

- GO annotations are based on evidence, which is given a type (evidence code) and a reference (usually a PubMed identifier)
- Evidence types
  - Curated from the primary literature
    - EXP, IDA, IEP, IGI IMP, IPI
  - Curated from “secondary sources”
    - TAS, NAS, IC
  - Curated from homology inference
    - ISS, IBA
  - Uncurated
    - IEA, RCA

GO evidence codes



All codes

Experimental, curated

Curated secondary

"Electronic"
(computational inferences)

IDA IPI IGI IMP IEP

TAS IC NAS

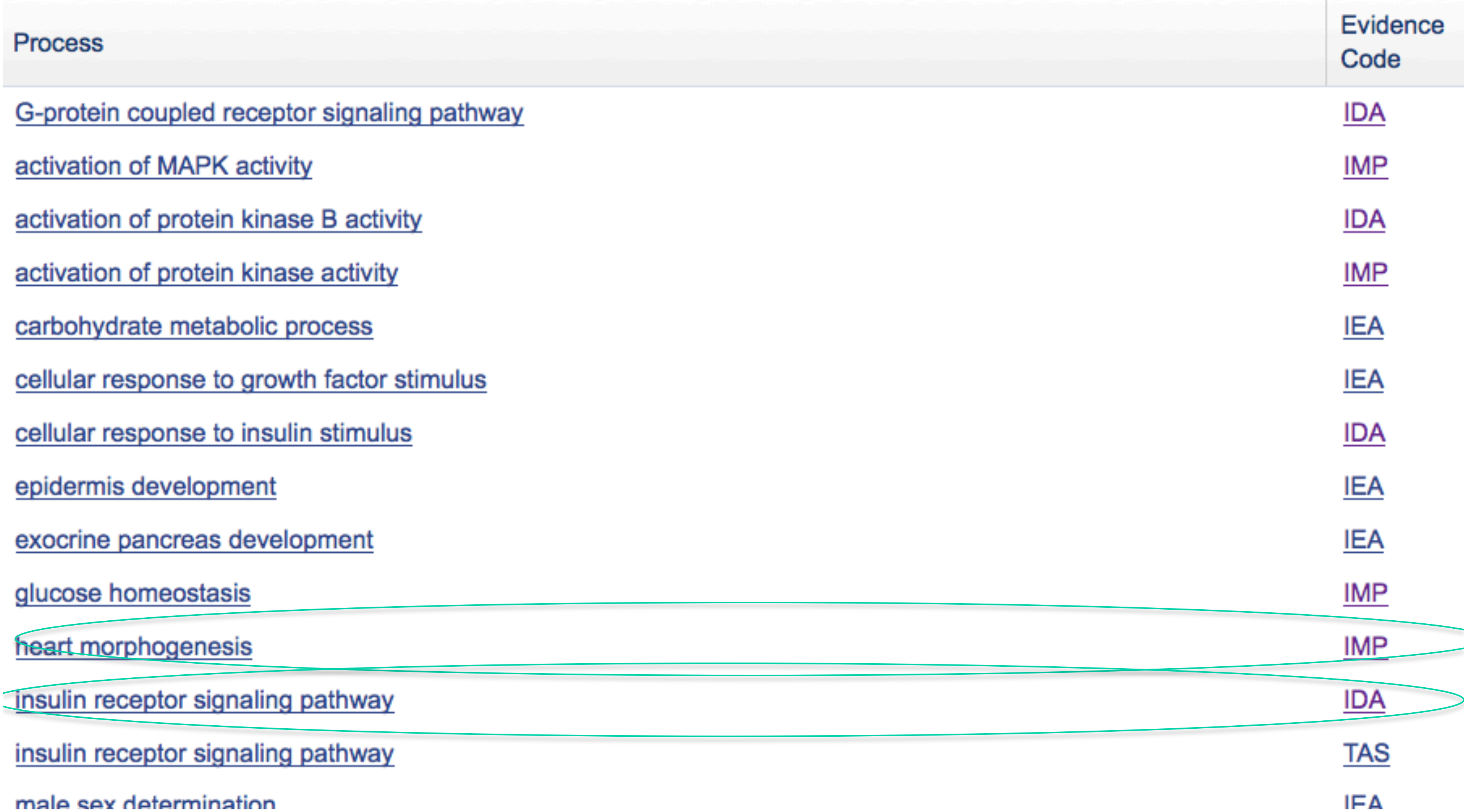
IBA IEA ISS ISO RCA

More direct

More traceable

More highly curated

IDA tends to be more "direct" than IMP, which can be a downstream causal effect



Process	Evidence Code
G-protein coupled receptor signaling pathway	IDA
activation of MAPK activity	IMP
activation of protein kinase B activity	IDA
activation of protein kinase activity	IMP
carbohydrate metabolic process	IEA
cellular response to growth factor stimulus	IEA
cellular response to insulin stimulus	IDA
epidermis development	IEA
exocrine pancreas development	IEA
glucose homeostasis	IMP
heart morphogenesis	IMP
insulin receptor signaling pathway	IDA
insulin receptor signaling pathway	TAS
male sex determination	IEA

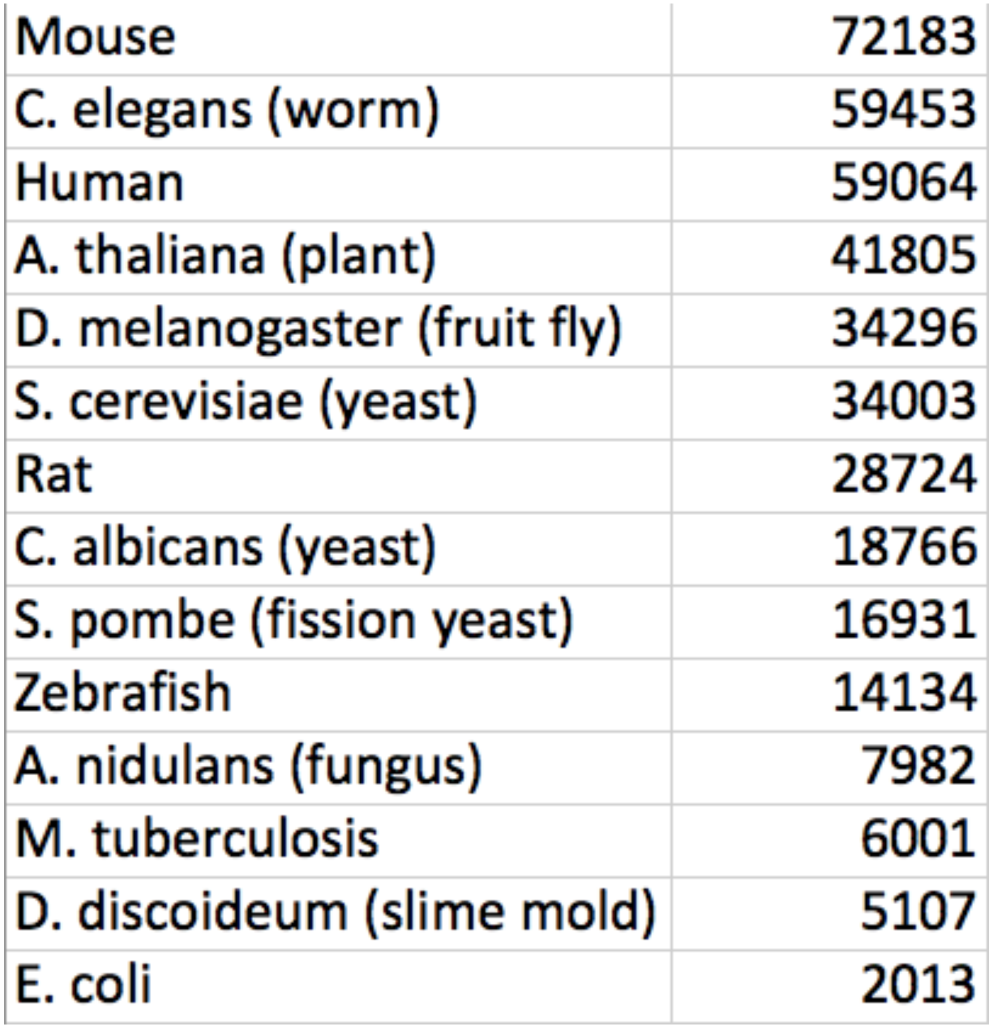
Experimental evidence codes

- Expert biologist reads a paper, and selects GO terms that best describe functions that are experimentally demonstrated in the paper
- GO database currently includes annotations from over 100,000 scientific papers
- Reference field links to paper and allows you to verify the annotation

Direct, literature-based annotation

- Function annotation **inference** based on direct evidence in the scientific literature
  - Experiment performed on that gene product itself
- Text mining and management (Textpresso)
  - Very active area of research
- Curator reads abstract or article and manually enters annotation
- GO annotation is performed at 12 different “model organism databases” and UniProt
- Two types:
  - Primary source: experimental paper (Evidence codes: IMP, IGI, IDA, IEP, IPI)
  - Secondary source: review article, introduction to another article, curator inference (TAS, NAS, IC)

GO experimental annotations cover a few major "model organisms"



Mouse	72183
C. elegans (worm)	59453
Human	59064
A. thaliana (plant)	41805
D. melanogaster (fruit fly)	34296
S. cerevisiae (yeast)	34003
Rat	28724
C. albicans (yeast)	18766
S. pombe (fission yeast)	16931
Zebrafish	14134
A. nidulans (fungus)	7982
M. tuberculosis	6001
D. discoideum (slime mold)	5107
E. coli	2013

Experimental evidence types

- “Experimental” evidence codes
  - IDA: inferred from direct assay
  - IGI: inferred from genetic interaction
  - IPI: inferred from protein interaction
  - IMP: inferred from mutant phenotype
  - IEP: inferred from expression pattern
  - EXP: inferred from experimental evidence
- Important distinctions
  - IDA, IGI, IPI: usually the most direct
  - IMP, IEP: can be indirect, downstream effects
  - IEP is used very cautiously by curators

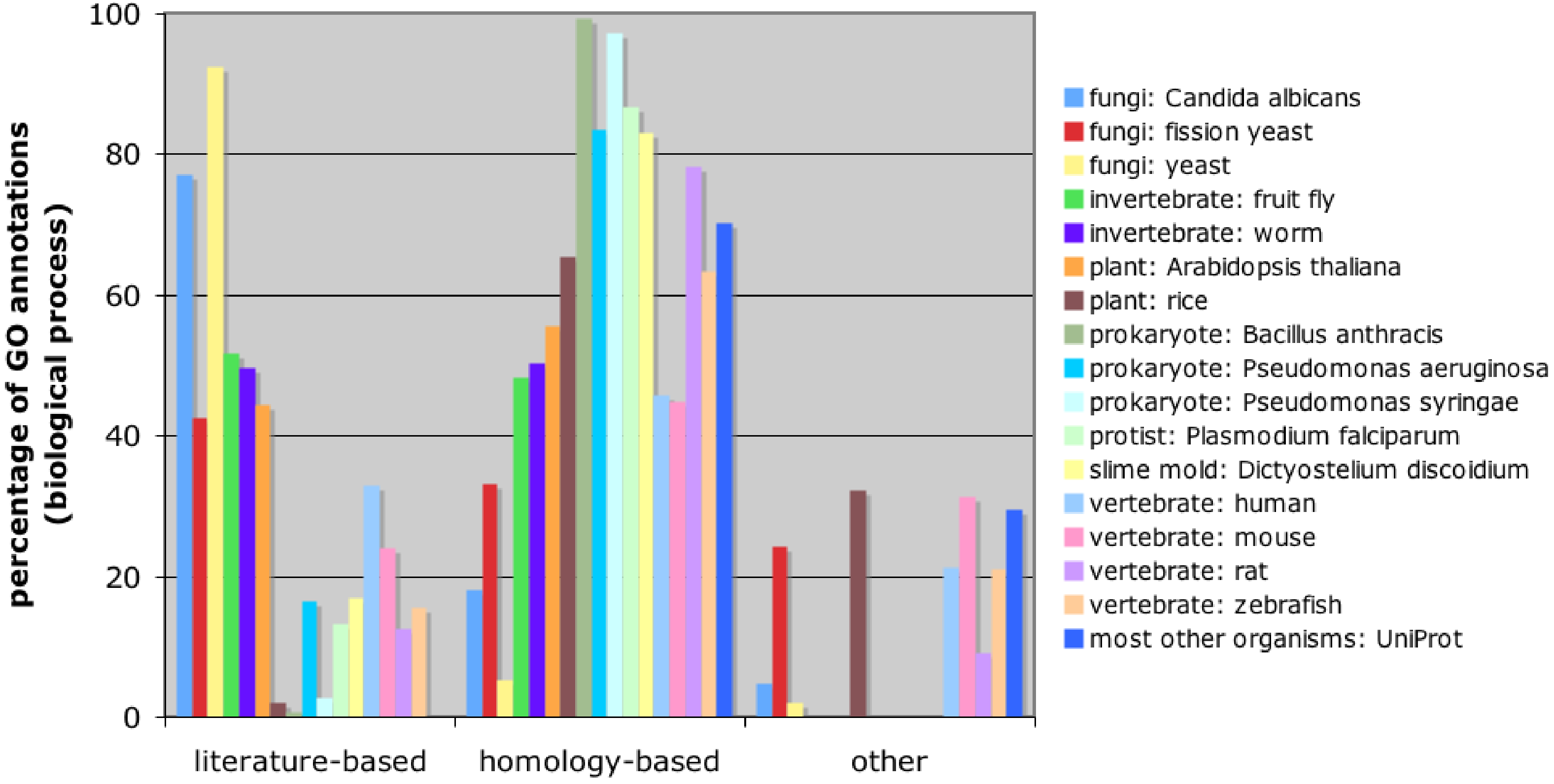
“Secondary” source annotations from literature

- TAS: traceable author statement
  - The author referenced another paper; these are being traced and replaced by primary annotations
- NAS: nontraceable author statement
  - The author did not reference another paper; these are no longer commonly used as evidence
- IC: inferred by curator
  - For example, a paper demonstrates transcription factor activity in a human cell; curator infers that it must function in the nucleus

“Electronic” evidence

- Important distinction: degree of manual review
  - RCA: no systematic review, mostly “guilt by association” methods
  - ISO: no review, but conservative rules for function inference for some 1:1 orthologs
  - ISS: review of pairwise homology and function, but no consistent rules
  - IEA: review of large lists of homologous proteins and selection of which terms to infer
  - IBA: review of ALL experimental annotations for each gene family and selection of which terms to infer by constructing explicit evolutionary model

Most GO annotations are based on homology (except for some yeasts)



Curr Opin Chem Biol. 2007 Feb;11(1):4-11. Epub 2007 Jan 5.

Ontology annotation: mapping genomic regions to biological function.

Thomas PD, Mi H, Lewis S.

Evolutionary Systems Biology Group, Artificial Intelligence Center, SRI International, Menlo Park, CA 94025, USA. paul.thomas@sri.com

Homology is still the most informative predictor of function

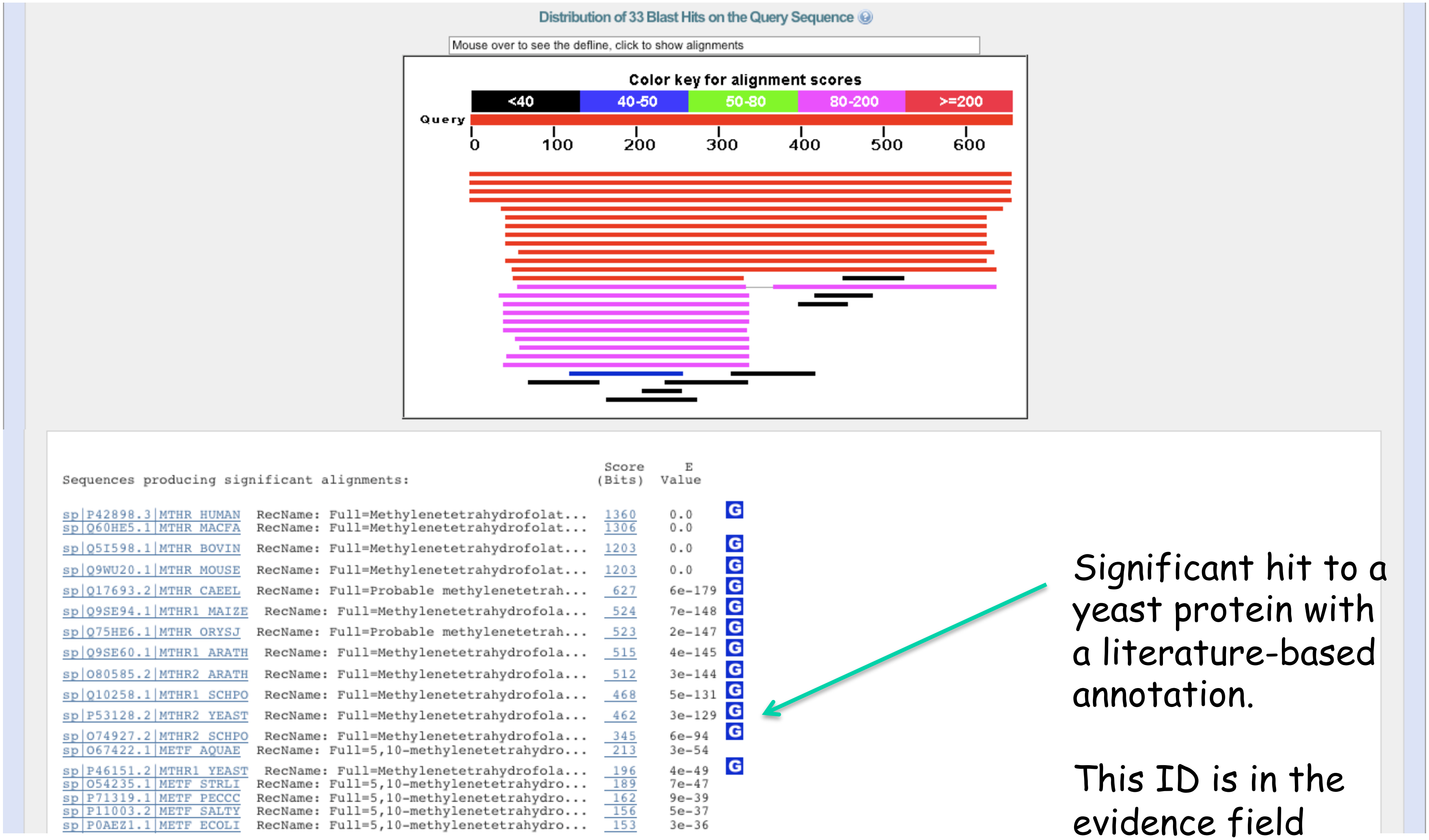
- Many "guilt by association" methods, e.g. protein interaction network analysis, gene co-expression, etc.
- In recent function prediction experiment (CAFA), homology still found to be major component of informative predictions
  - See BMC Bioinformatics 14:suppl 3 (2013), e.g. Hamp et al., Gillis et al.

Homology-based annotation

- “traditional” pairwise view
  - If two sequences are similar, they are likely to share some functions in common
  - So if I know the function of one gene, I can make inferences about the function of another gene
    - “transitive annotation” (ISS evidence code in GO)
  - Very commonly applied, in database search algorithms like BLAST, FASTA (e.g. Blast2GO)
  - This success has led to overinterpretation of its meaning by many casual users
    - A class of database search has become a metaphor, implying that **“genes have similar functions *because* they have similar sequences”**

ISS is based on pairwise sequence comparison:

example BLAST results for human MTHFR vs. SwissProt database



Distribution of 33 Blast Hits on the Query Sequence

Mouse over to see the defline, click to show alignments

```
                                                                      Score    E
Sequences producing significant alignments:                          (Bits)  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.

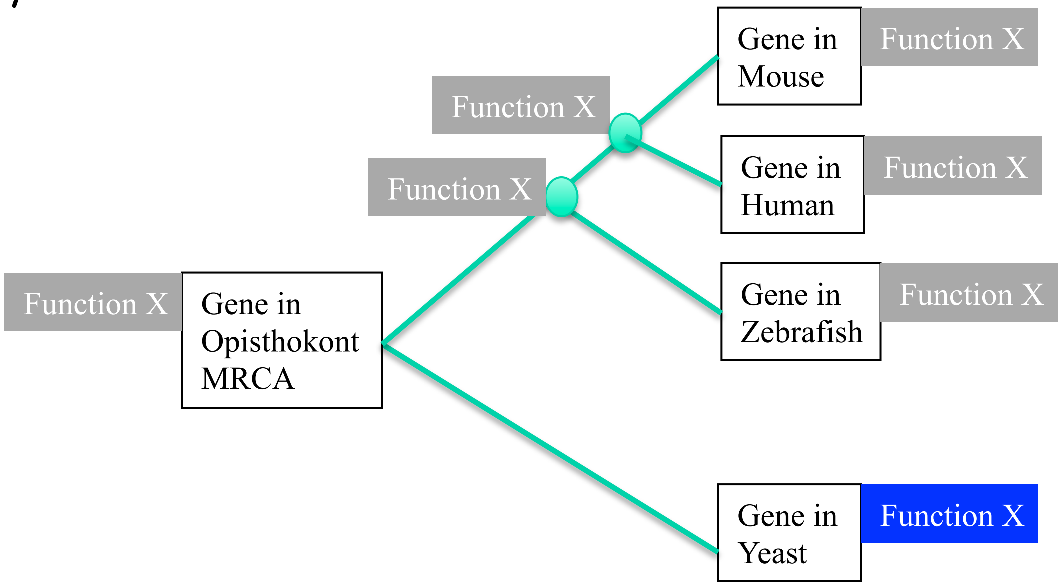
This ID is in the evidence field

Understanding what homology inference really is

- 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"

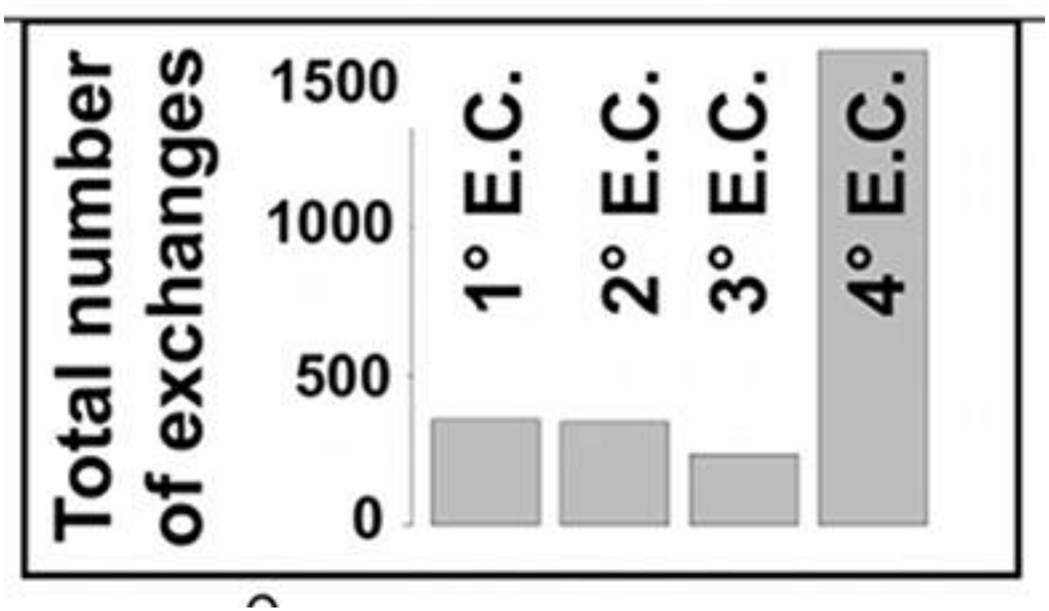
– not just an inference about one gene. It is also making inferences about
  - The most recent common ancestor (MRCA)
  - Continuous inheritance since the MRCA
  - Potential inheritance by other descendants of the MRCA



Fundamental challenge in using sequence similarity to annotate function (1): SEQUENCES of different genes (proteins) evolve at different rates

- Sequence divergence (e.g. BLAST score or E-value) cannot be simply converted to an evolutionary relationship
  - Score depends on time, selective constraints, length of gene/protein sequence, sequence composition
- Problem can be addressed using phylogenetic trees

Fundamental challenge in using sequence similarity to annotate function (2): Different GO functions in same protein family evolve at different rates



- Enzyme mechanism (1-3) evolves more slowly than substrate specificity (4)
- In general, no pairwise similarity threshold to reliably predict all different functions!
- Problem can be addressed by treating different functions independently

PLoS Comput Biol. 2012;8(3):e1002403. Epub 2012 Mar 1.

Exploring the evolution of novel enzyme functions within structurally defined protein superfamilies.

Furnham N, Sillitoe I, Holliday GL, Cuff AL, Laskowski RA, Orengo CA, Thornton JM.

EMBL-EBI, Wellcome Trust Genome Campus, Hinxton, Cambridge, United Kingdom. nickf@ebi.ac.uk

Using trees to get relationships between genes

- ISO: inferred from sequence orthology
  - From Ensembl Compara
  - Function annotations are NOT REVIEWED
    - For vertebrates: infers that all experimental annotations in any vertebrate are true of all vertebrates IF there is one-to-one gene orthology
    - For plants: infers that all experimental annotations in any plant are true of all plants IF orthology AND sequence identity > 60%.

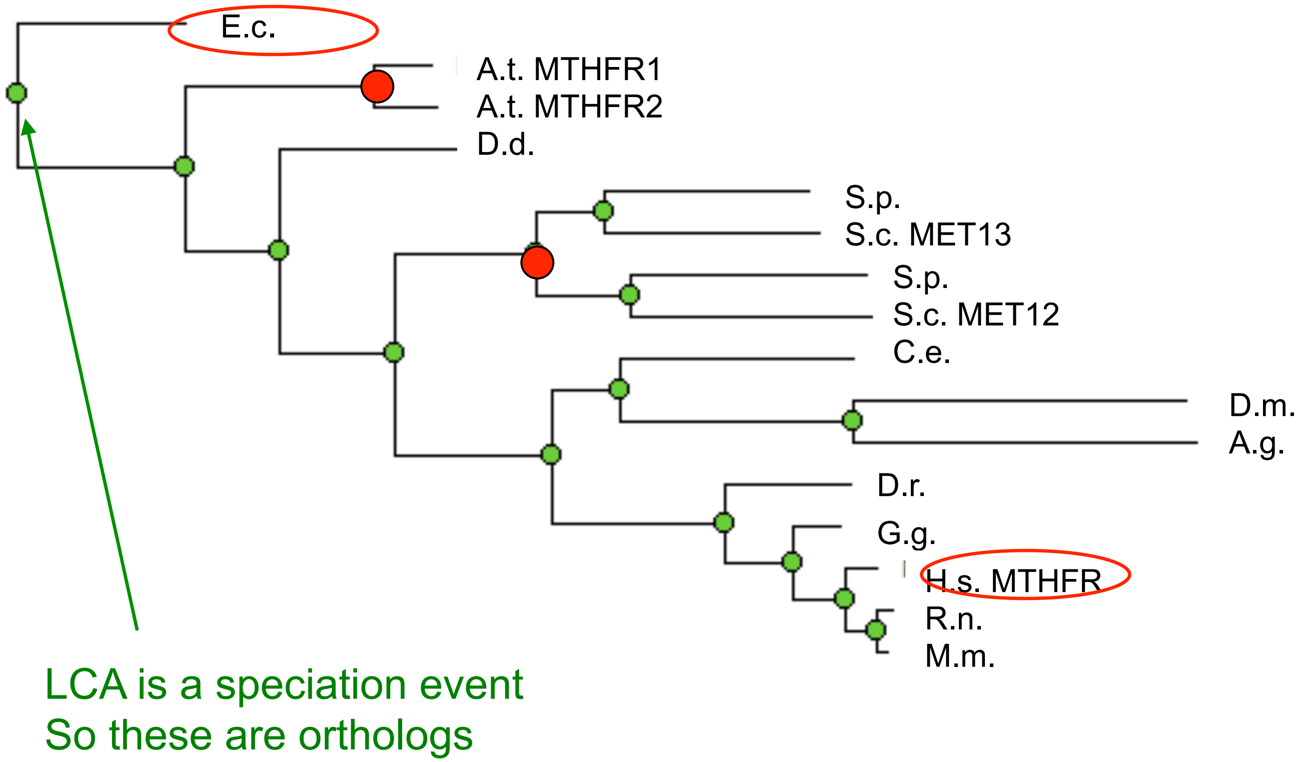
Understanding ISO: Concept of orthologs

- The term “orthologs” is often used to denote “the same gene” in different organisms but this is not techically correct, and can lead to confusion
- Defined by J. Fitch (Syst Zool 19:99, 1970)
- Orthologs share a MRCA immediately preceding a speciation event
  - i.e. they can be traced to a **single** gene in the most recent common ancestor population/species
- Paralogs share a MRCA immediately preceding a gene duplication event
  - i.e. they can be traced to a gene duplication event in the most recent common ancestor population/species, and can be traced to **distinct** ancestral genes in that species

Why orthology is confusing

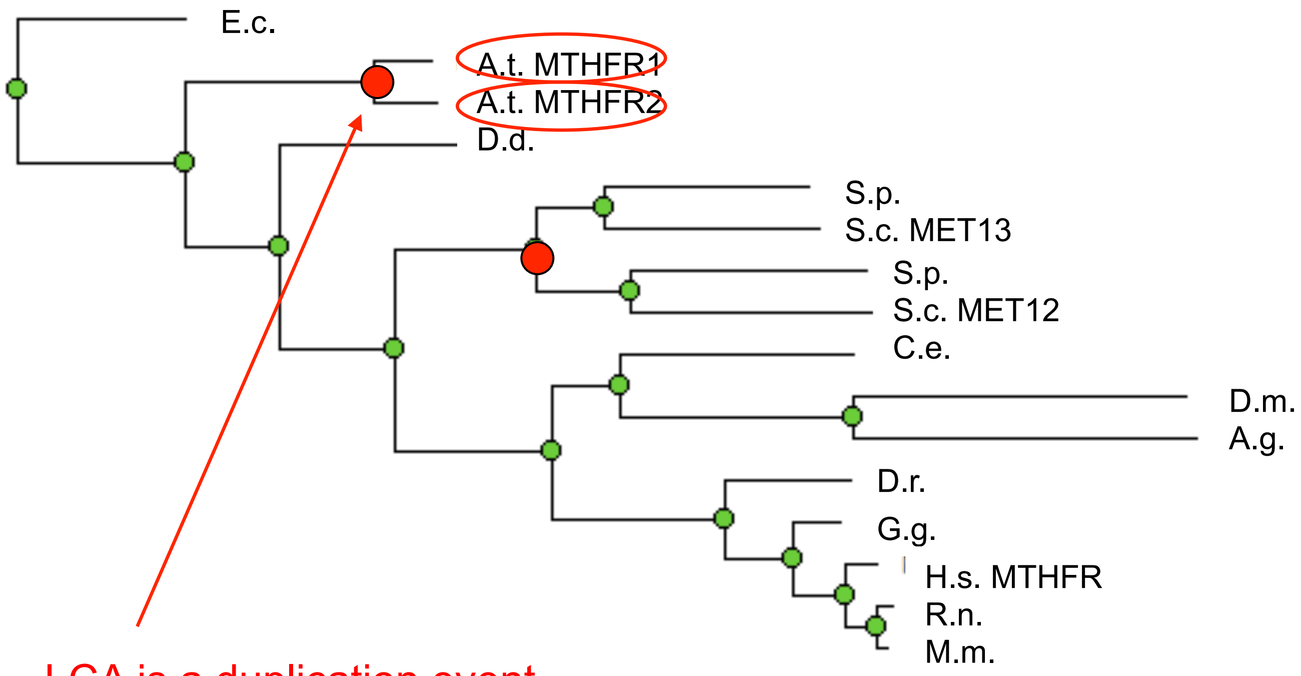
- It is a statement about an evolutionary relationship and not about gene function
  - Orthologs may be doing different things in their respective species
- It is a pairwise definition, yet "ortholog group" or "ortholog cluster" are common terms
  - Orthology is NOT TRANSITIVE
    - An ortholog cluster may contain pairs that are paralogs!
- Proposed solutions are also complicated
  - One solution is to ignore any cases except "one-to-one orthologs" where no gene duplication occurs, but this misses many functionally similar genes
    - All current ISO annotations are from one-to-one orthology
  - Another solution is to allow "close paralogs" ("in-paralogs", Sonnhammer) into the cluster.

Orthology
only defined for PAIRS of genes



Two genes are orthologs if their LCA was a speciation event

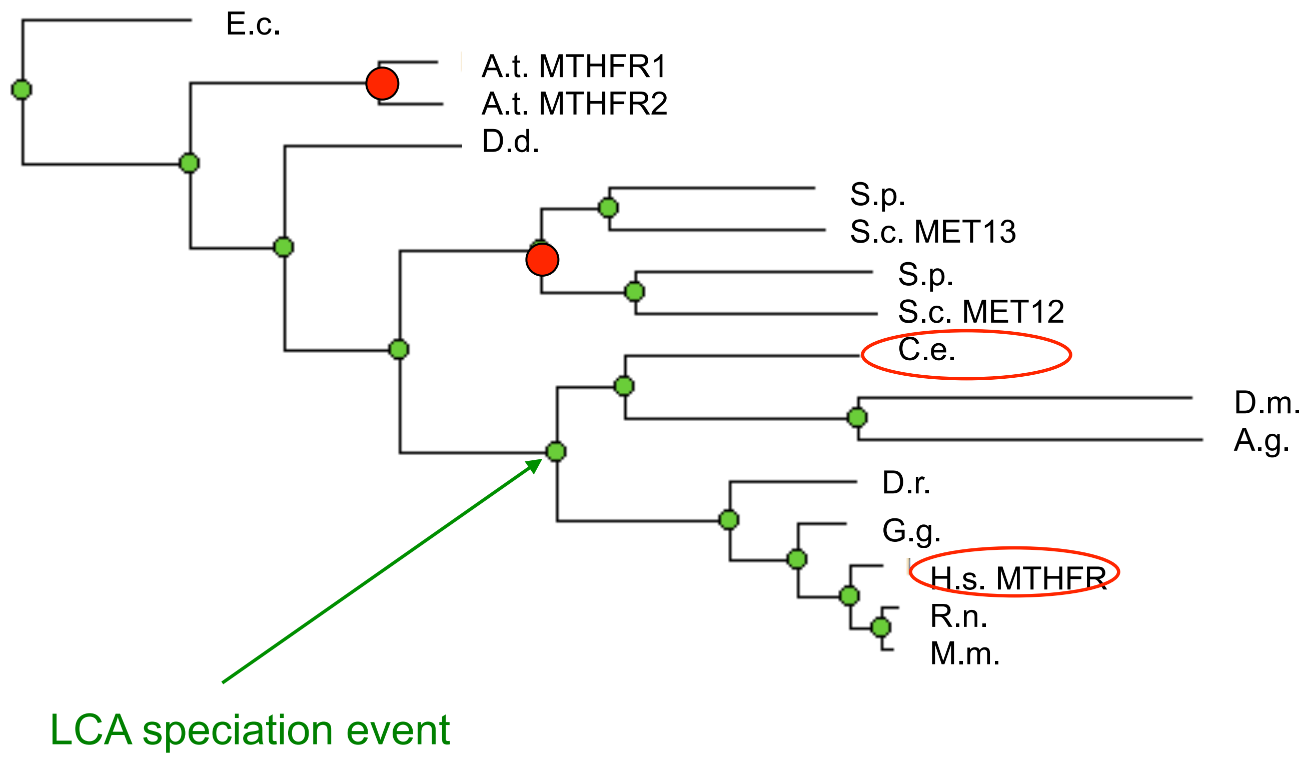
Paralogy
only defined for PAIRS of genes



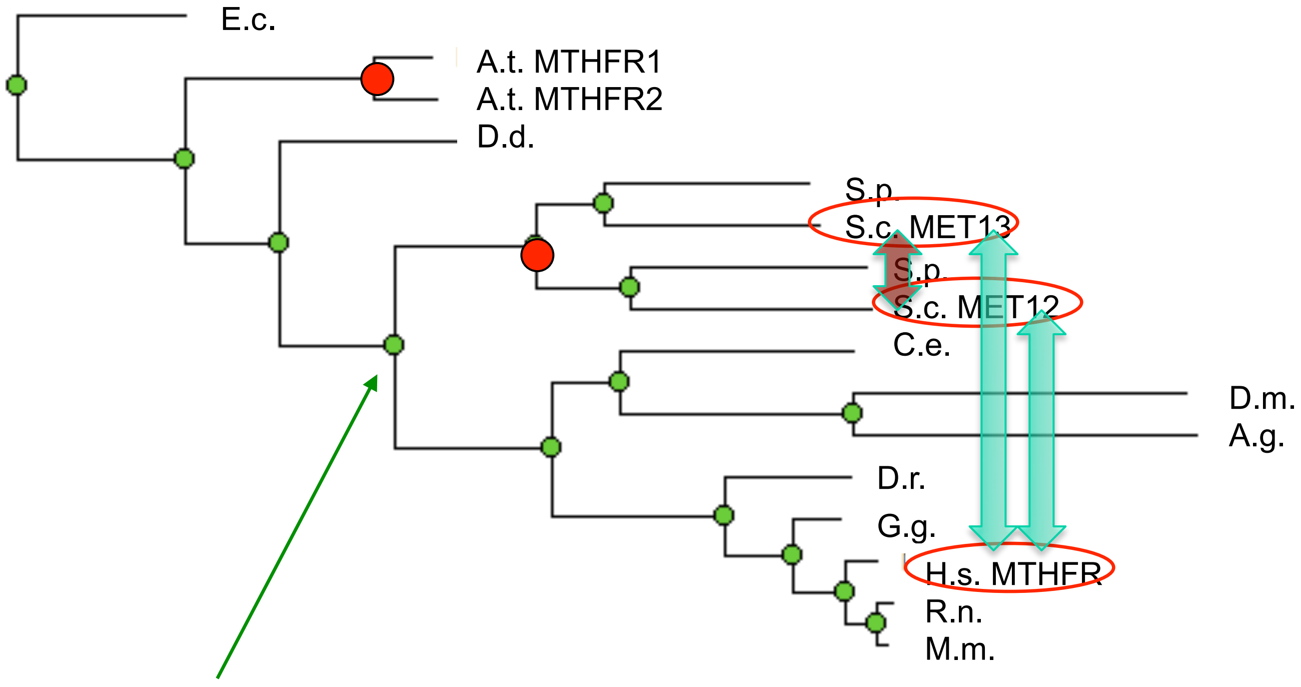
LCA is a duplication event
So these are paralogs

Two genes are paralogs if their LCA was a duplication event

Orthology is simple when there are no duplications following speciation



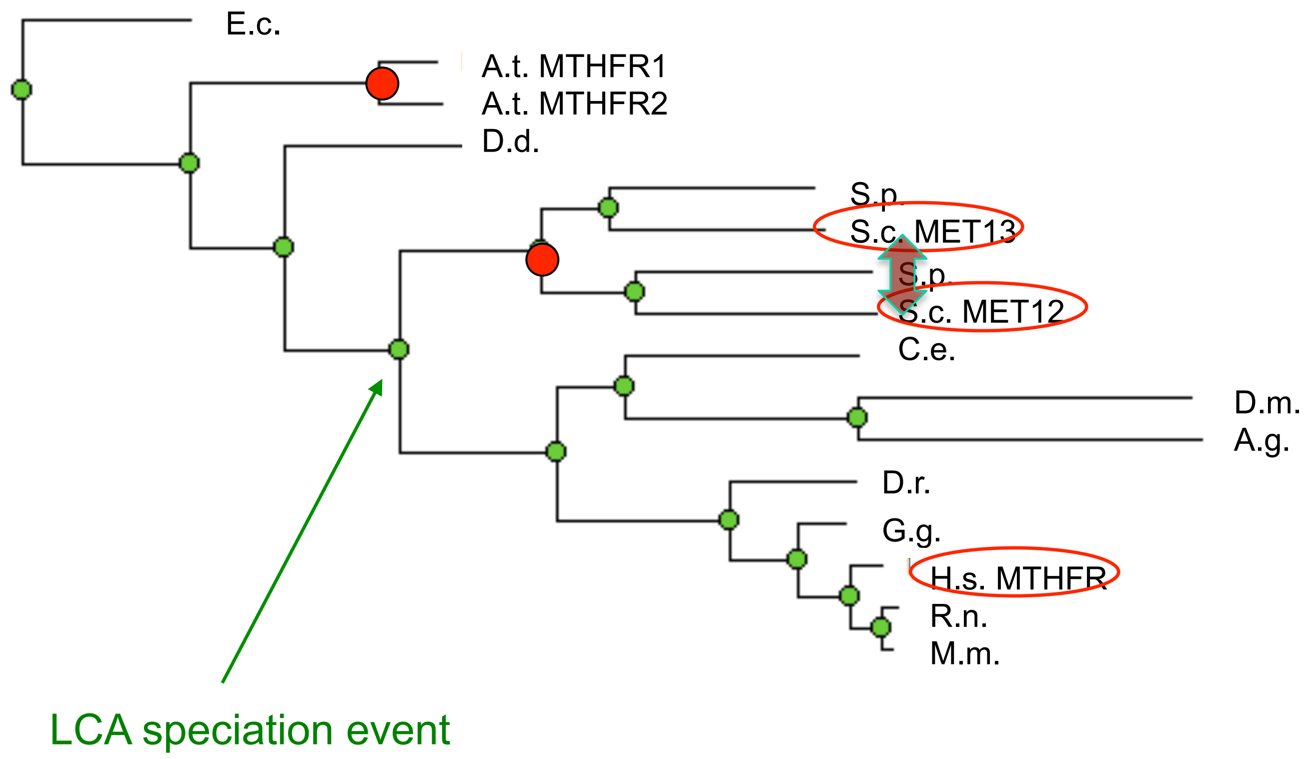
Orthology gets more complicated when there are duplications following speciation



LCA speciation event

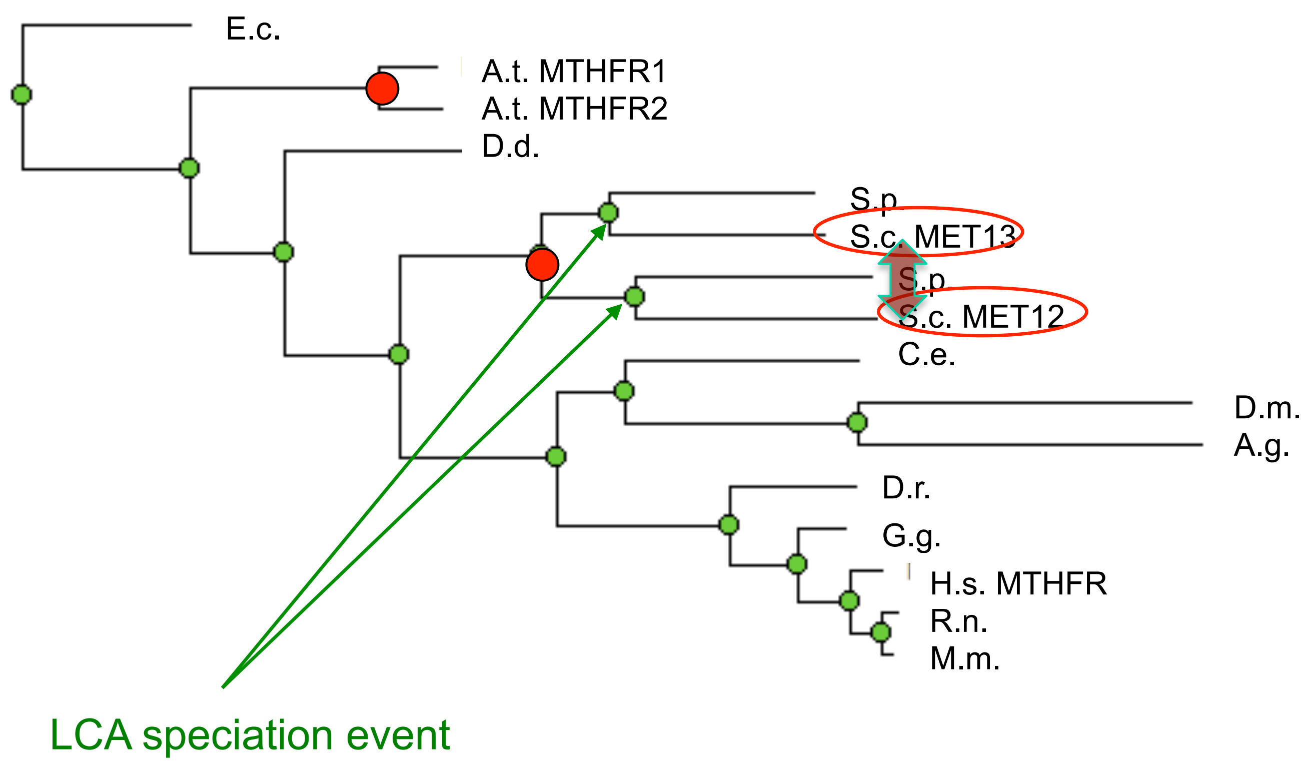
H.s. MTHFR has two orthologs in yeast
And these two orthologs are paralogs of each other

These genes are “in paralogs” with respect to each other when comparing to animal genomes



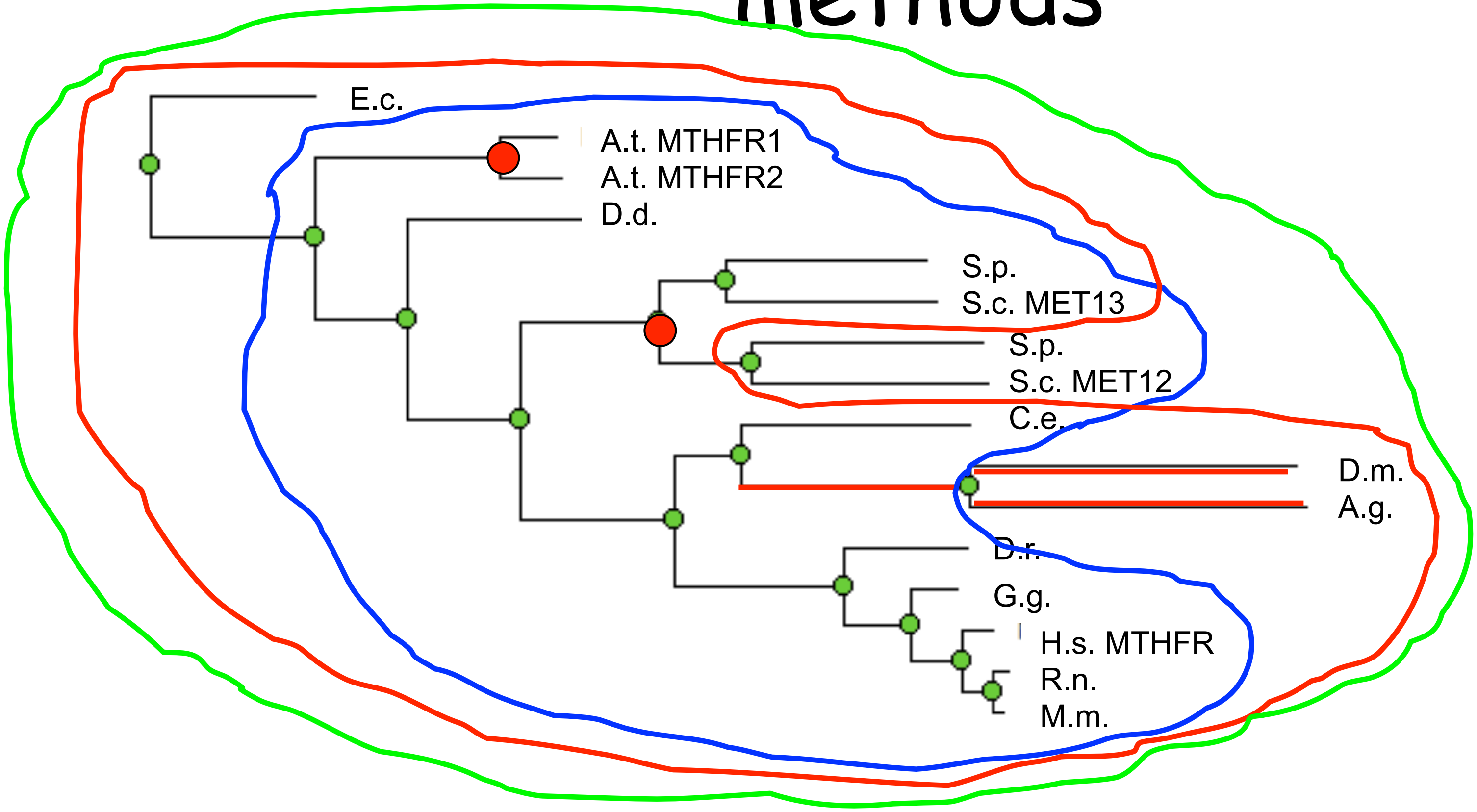
H.s. MTHFR has two orthologs in yeast
And these two orthologs are paralogs of each other

But these same genes are "out paralogs" with respect to each other when comparing fungal genomes



H.s. MTHFR has two orthologs in yeast
And these two orthologs are paralogs of each other

Clusters from different "orthology" methods



- OrthoMCL in red; PhiGs in blue; InParanoid in green
- An "ortholog cluster" is made by one or more "slices" through the protein family tree

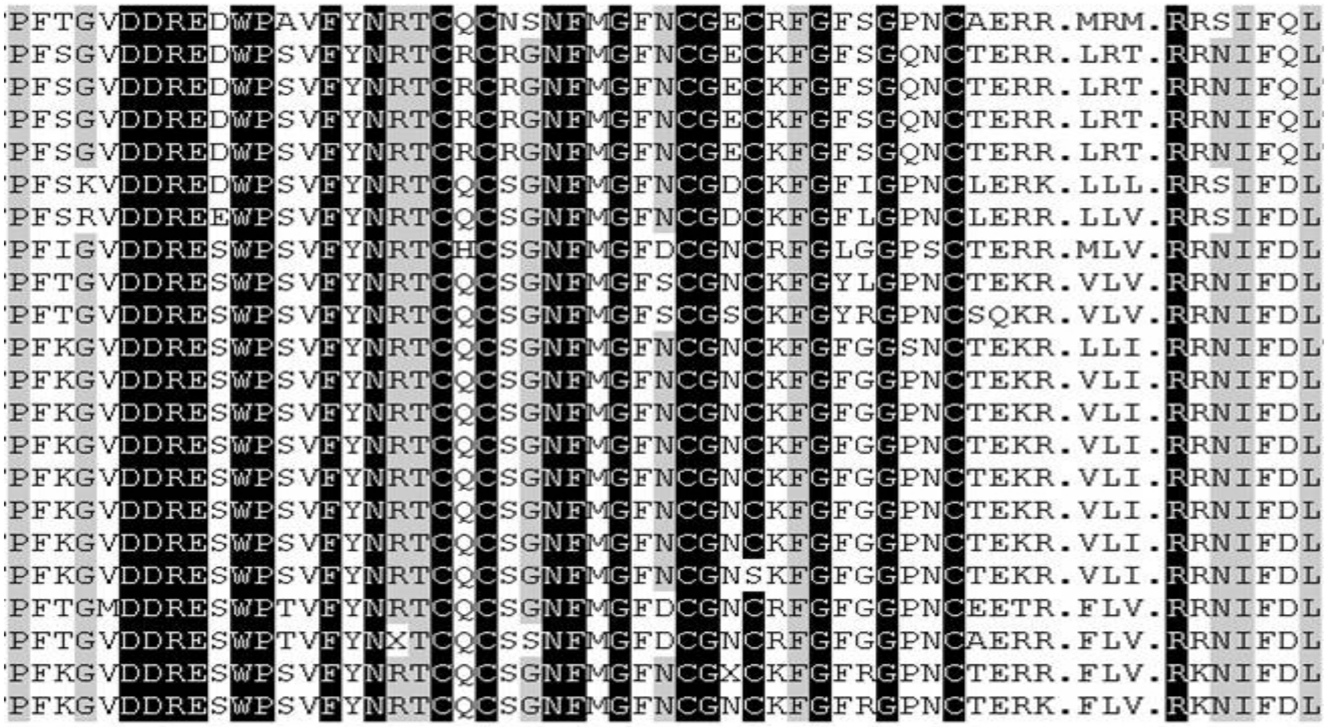
IEA annotations have multiple sources

- IEA annotations far outnumber any other type
- Two major sources
  - Swiss-Prot keywords, mapped to GO terms
    - Assigned manually, or by unreviewed sequence similarity
    - No evidence trail
  - InterPro models, mapped to GO terms manually
    - Assigned manually to families of related sequences, not to individual sequences

IEA annotations: InterPro

- InterproScan is among most highly-used automatic method
- Combines most popular web resources into one package
- Most of these are homology-based, searching a library of Hidden Markov Models (HMMs)
- Two distinct types of model
  - Domain-based (e.g. Pfam, SMART, Superfamily)
    - Model divergent groups usually with relatively ancient common ancestor
    - Domain shuffling has often occurred since this ancestor
    - Useful for seeing modular architecture
    - Will often predict only very general function, conserved since MRCA of module
  - Subfamily-based (e.g. PANTHER, TIGRFAMs, PRINTS)
    - Model groups that are more closely related (relatively recent ancestor or less divergent phylogenetic groups)
    - Domain shuffling has generally not occurred since this ancestor
    - Can predict much more specific functions

HMM: “generative model”, first-order, learn “hidden” states and probabilities

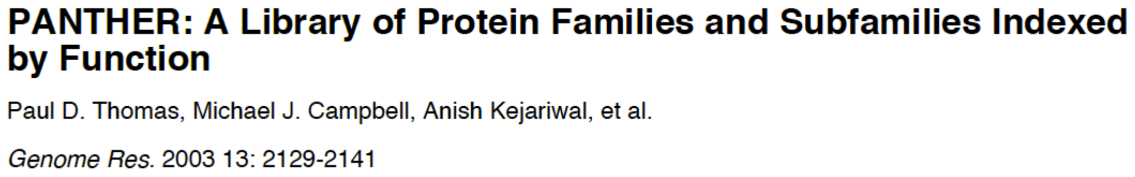


```
PFTGVDDREDWPAVFYNRTCQCNSNFMGFNCGECRFGFSGPNCAERR.MRM.RRSIFQL
PFSGVDDREDWPSVFYNRTCRCRGNFMGFNCGECKFGFSGQNCTERR.LRT.RRNIFQL
PFSGVDDREDWPSVFYNRTCRCRGNFMGFNCGECKFGFSGQNCTERR.LRT.RRNIFQL
PFSGVDDREDWPSVFYNRTCRCRGNFMGFNCGECKFGFSGQNCTERR.LRT.RRNIFQL
PFSGVDDREDWPSVFYNRTCRCRGNFMGFNCGECKFGFSGQNCTERR.LRT.RRNIFQL
PFSKVDDREDWPSVFYNRTCQCSGNFMGFNCGDCKFGFIGPNCLERK.LLL.RRSIFDL
PFSRVDDREEWPSVFYNRTCQCSGNFMGFNCGDCKFGFLGPNCLERR.LLV.RRSIFDL
PFIGVDDRESWPSVFYNRTCHCSGNFMGFDCGNCRFGLGGPSCTERR.MLV.RRNIFDL
PFTGVDDRESWPSVFYNRTCQCSGNFMGFSCGNCKFGYLGPNCTEKR.VLV.RRNIFDL
PFTGVDDRESWPSVFYNRTCQCSGNFMGFSCGSCKFGYRGPNCSQKR.VLV.RRNIFDL
PFKGVDDRESWPSVFYNRTCQCSGNFMGFNCGNCKFGFGGSNCTEKR.LLI.RRNIFDL
PFKGVDDRESWPSVFYNRTCQCSGNFMGFNCGNCKFGFGGPNCTEKR.VLI.RRNIFDL
PFKGVDDRESWPSVFYNRTCQCSGNFMGFNCGNCKFGFGGPNCTEKR.VLI.RRNIFDL
PFKGVDDRESWPSVFYNRTCQCSGNFMGFNCGNCKFGFGGPNCTEKR.VLI.RRNIFDL
PFKGVDDRESWPSVFYNRTCQCSGNFMGFNCGNCKFGFGGPNCTEKR.VLI.RRNIFDL
PFKGVDDRESWPSVFYNRTCQCSGNFMGFNCGNCKFGFGGPNCTEKR.VLI.RRNIFDL
PFKGVDDRESWPSVFYNRTCQCSGNFMGFNCGNCKFGFGGPNCTEKR.VLI.RRNIFDL
PFKGVDDRESWPSVFYNRTCQCSGNFMGFNCGNSKFGFGGPNCTEKR.VLI.RRNIFDL
PFTGMDDRESWPTVFYNRTCQCSGNFMGFDCGNCRFGFGGPNCEETR.FLV.RRNIFDL
PFTGVDDRESWPTVFYNXTCQCSSNFMGFDCGNCRFGFGGPNCAERR.FLV.RRNIFDL
PFKGVDDRESWPSVFYNRTCQCSGNFMGFNCGXCKFGFRGPNCTERR.FLV.RKNIFDL
PFKGVDDRESWPSVFYNRTCQCSGNFMGFNCGNCKFGFRGPNCTERK.FLV.RKNIFDL
```

Mammalian tyrosinases excerpted from
an alignment spanning vertebrates

Profile-based annotation

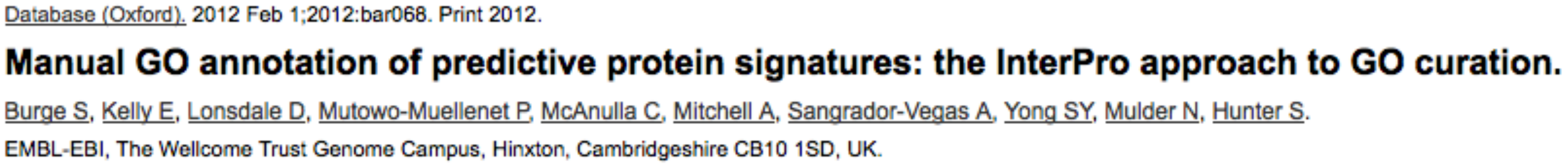
- Define a group of homologous sequences
  - Family/domain (e.g. Pfam)
  - Subfamily (e.g. PANTHER)
- For most methods, build an HMM to recognize members of the homologous group
- Annotate the group with functions/processes all known members have in common



PANTHER: A Library of Protein Families and Subfamilies Indexed by Function

Paul D. Thomas, Michael J. Campbell, Anish Kejariwal, et al.

Genome Res. 2003 13: 2129-2141



Database (Oxford). 2012 Feb 1;2012:bar068. Print 2012.

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

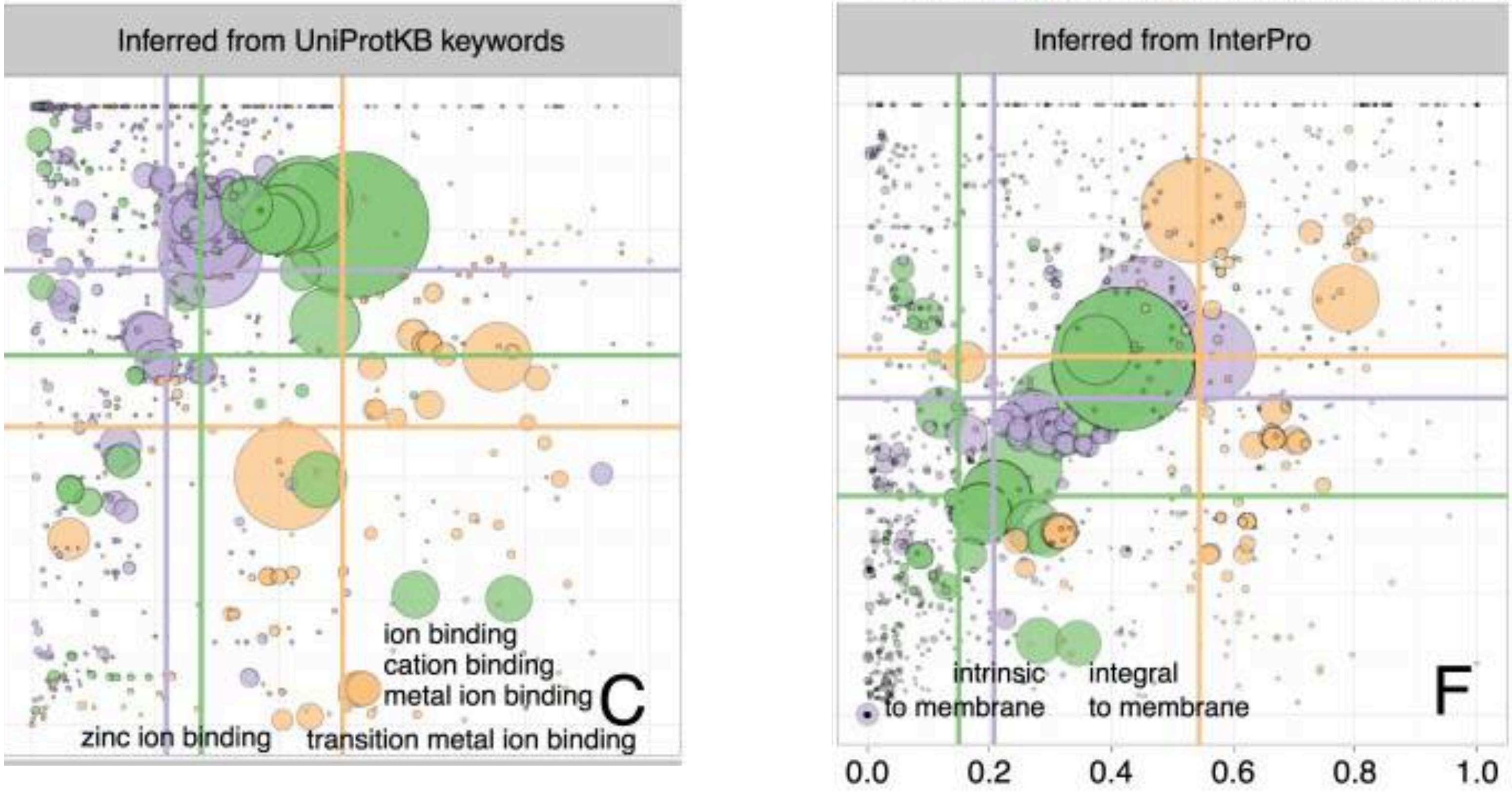
Burge S, Kelly E, Lonsdale D, Mutowo-Muellenet P, McAnulla C, Mitchell A, Sangrador-Vegas A, Yong SY, Mulder N, Hunter S.

EMBL-EBI, The Wellcome Trust Genome Campus, Hinxton, Cambridgeshire CB10 1SD, UK.

Profile-based annotation

- Driven by sequence relationships first, function later
  - Generally works well for molecular function
    - Sometimes loses specificity, depending on the approach
  - Loses specificity especially for biological process largely because of
    - co-option into new processes during evolution
    - Domain shuffling

IEA: keywords are more reliable than InterPro



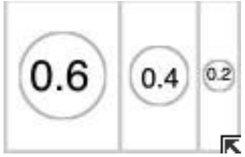
PLoS Comput Biol. 2012 May;8(5):e1002533. Epub 2012 May 31.

Quality of computationally inferred gene ontology annotations.

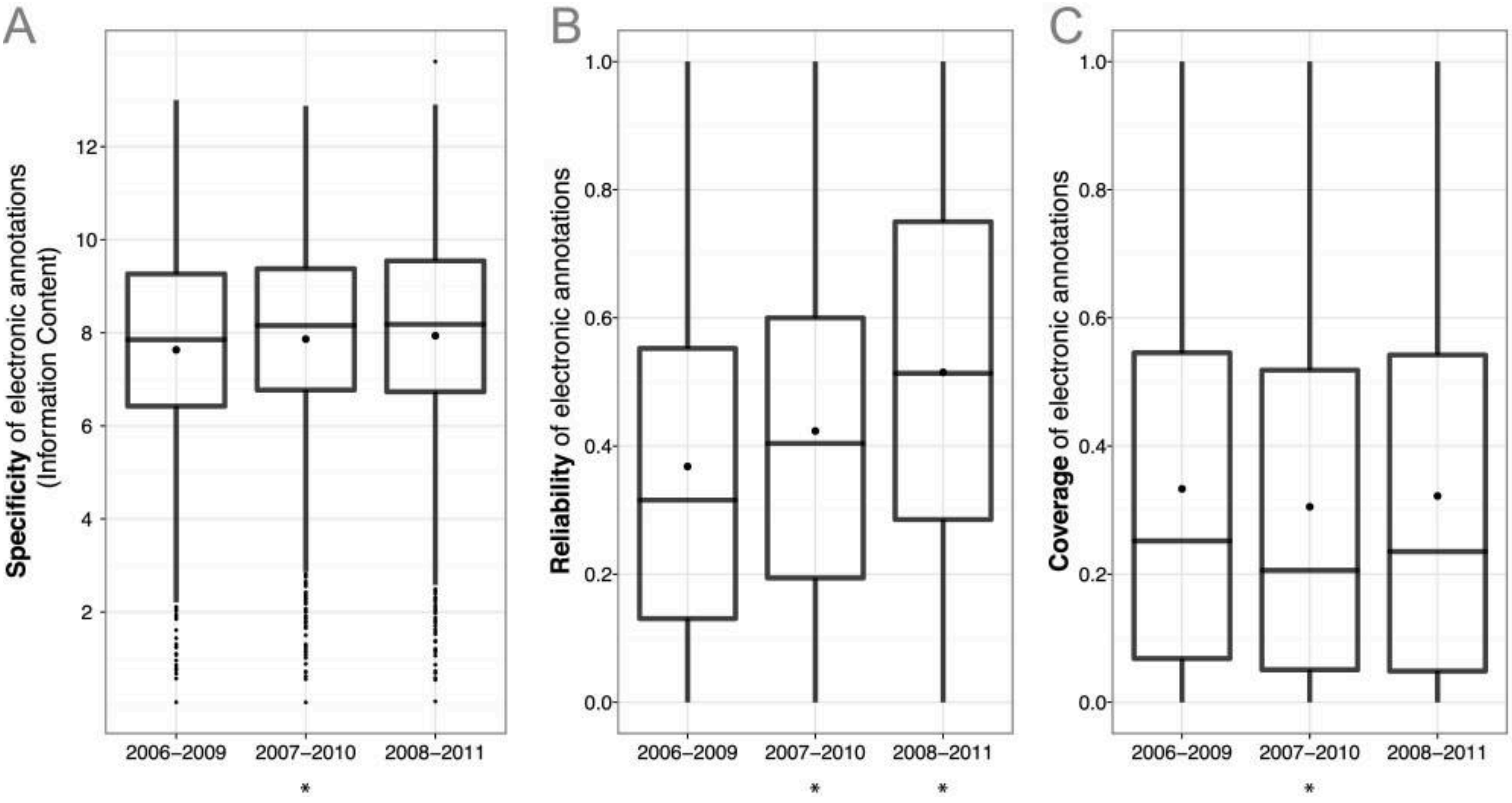
Skunca N, Altenhoff A, Dessimoz C.

Ruđer Bošković Institute, Division of Electronics, Zagreb, Croatia.

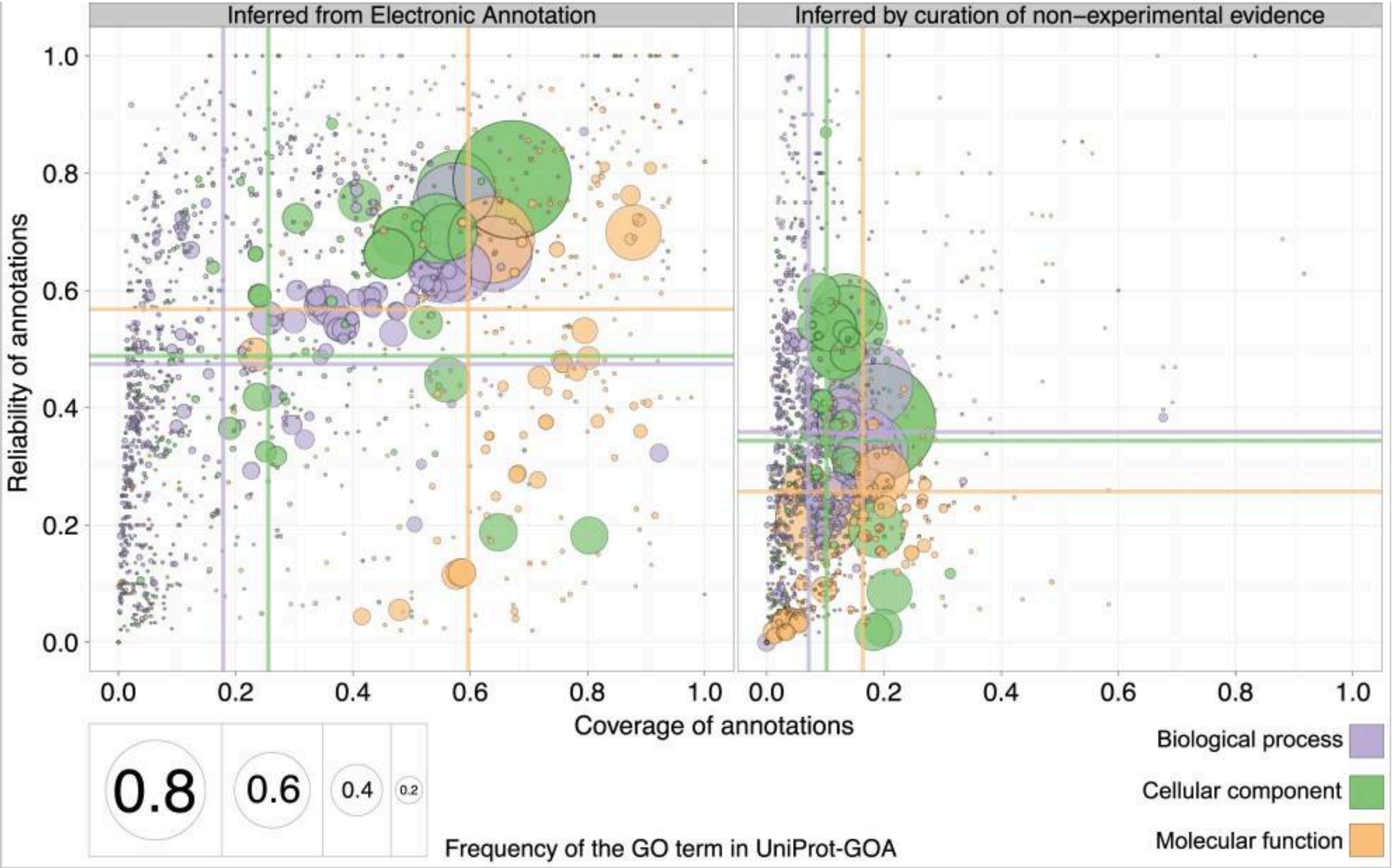
🟪 Biologica
process



IEAs have become more specific and more reliable



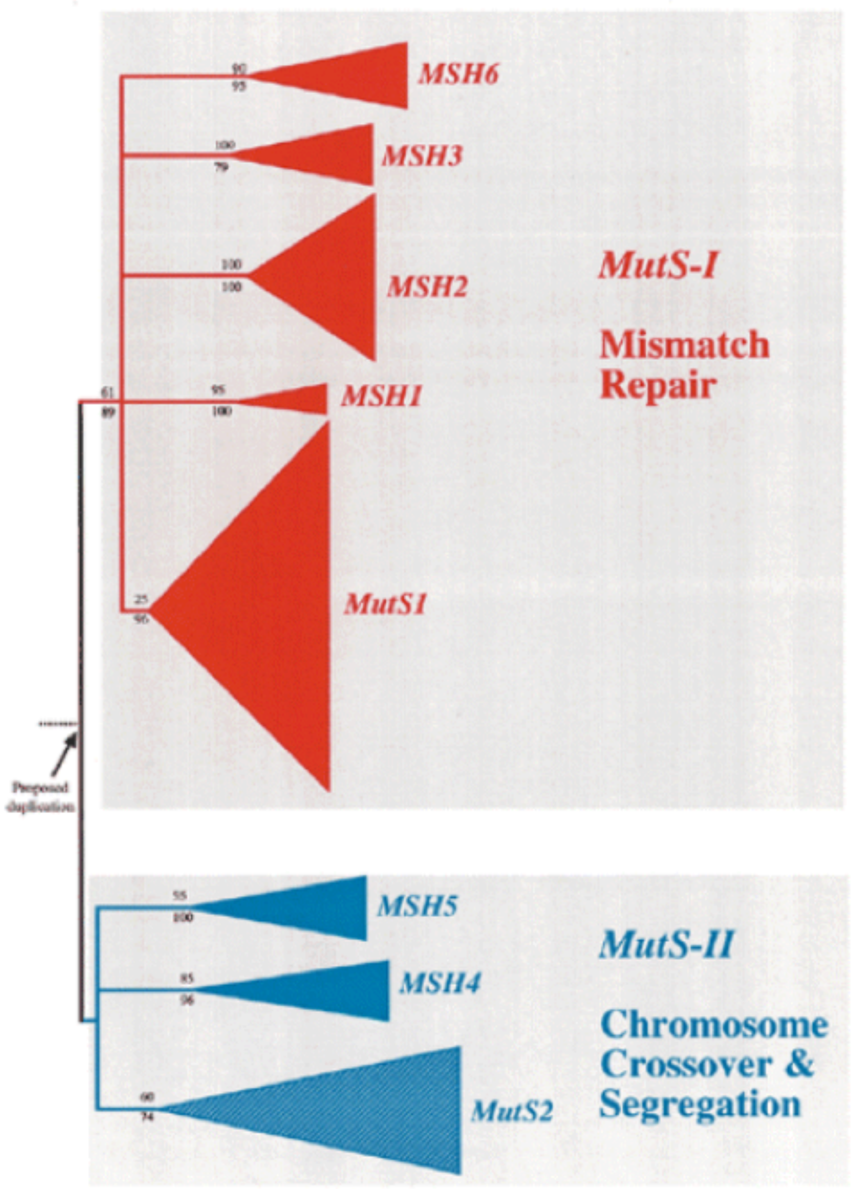
IEA is more reliable than ISS+IC



IBA: inferred annotations using manually annotated ancestral genes

- New effort within GO Consortium
  - Currently covers ~10% of genes in 85 genomes, growing daily
- Review ALL experimental annotations for ALL genes in a gene family
- Build explicit models of function evolution
  - Use “evolutionary reasoning”: descendants generally share a character because they inherited it from a common ancestor
    - Infer the function of an ancestor from knowledge about its descendants
    - Infer the function of uncharacterized descendants from inference about its ancestor
  - Create a model of evolution of function for every gene family
    - Gains of function
    - Losses of function

“Phylogenomic” function annotation



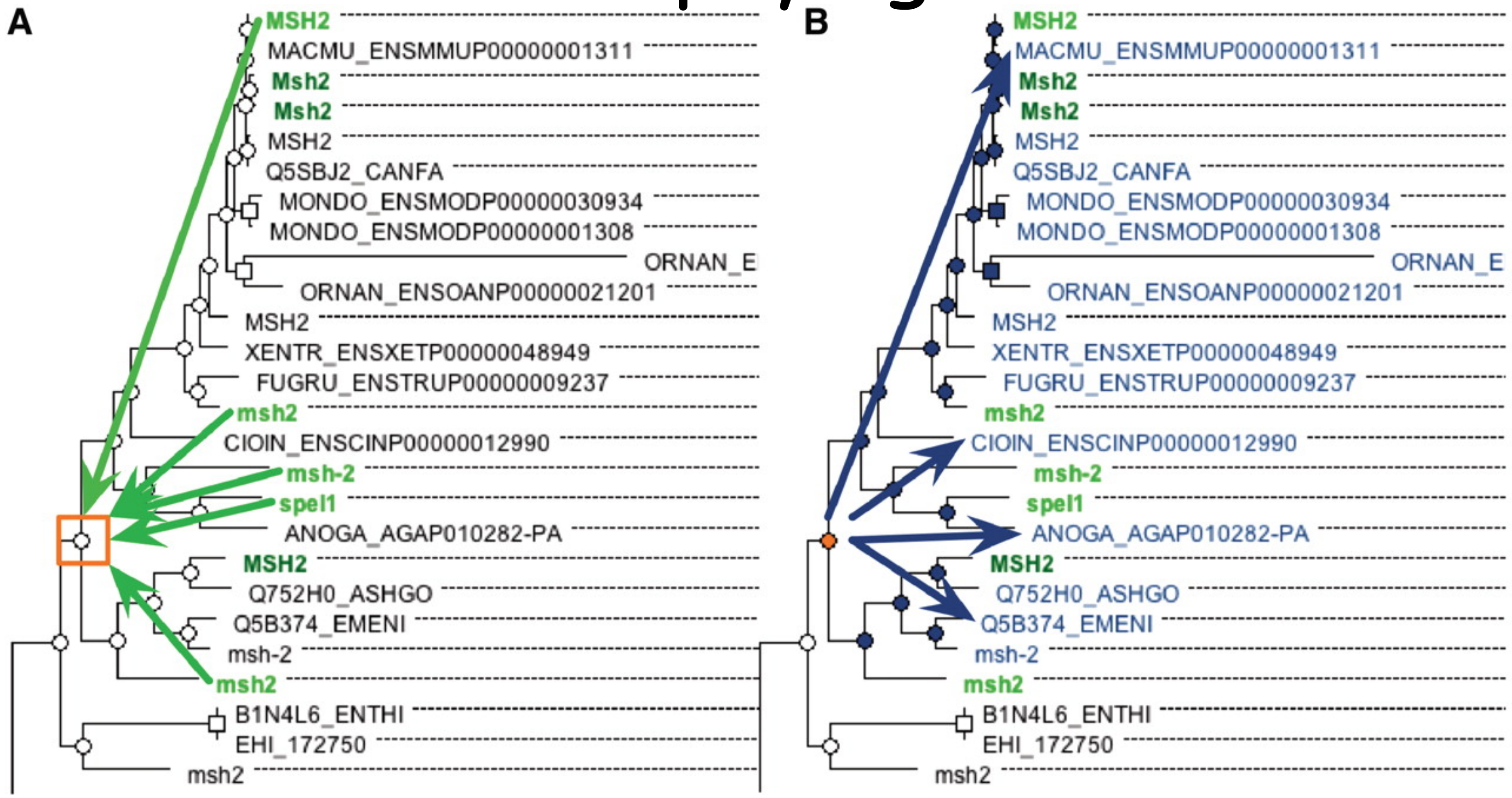
- View known data in the context of phylogenetic tree
- Infer subfamilies that share function

Nucleic Acids Res. 1998 Sep 15;26(18):4291-300.

A phylogenomic study of the MutS family of proteins.

Eisen JA.

IBA: Use multiple pieces of evidence in a phylogenetic tree

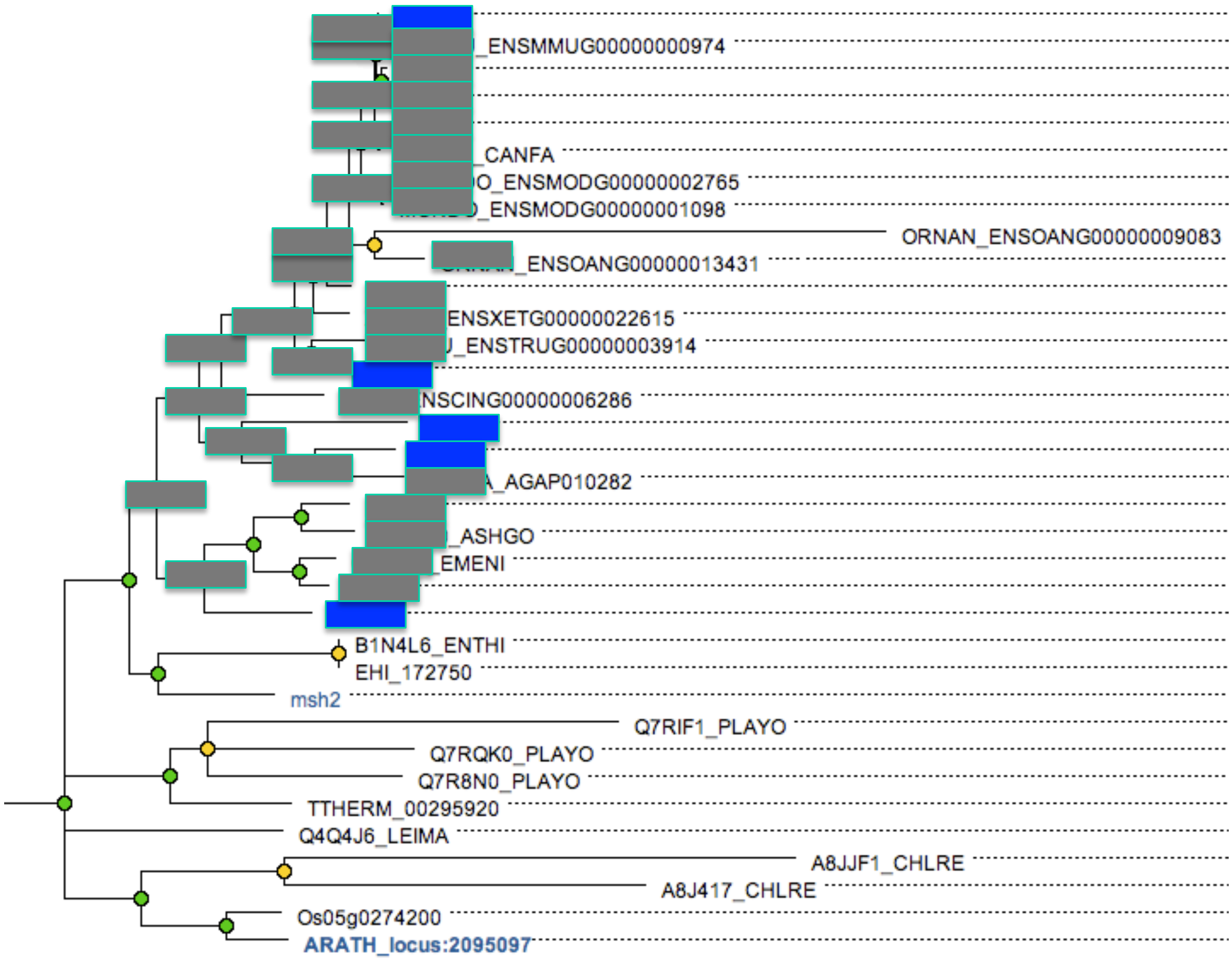


Integration of experimental GO annotations from different models (curated)

Inheritance of inferred ancestral annotations to annotate extant genes (automatic)

Gaudet P et al. Brief Bioinform 2011;12:449-462

Example annotation: maintenance of DNA repeat elements

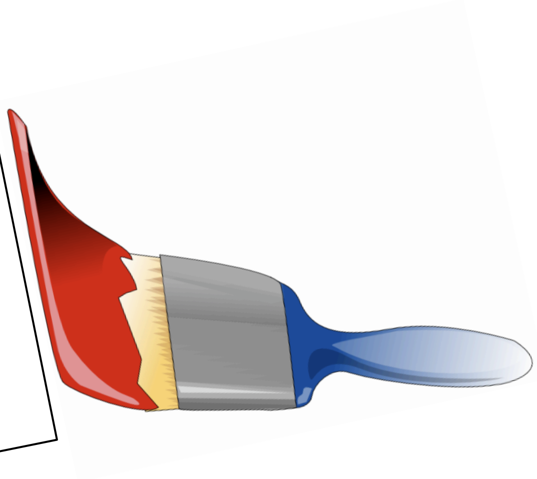


IBA: software-assisted manual annotation

- Need to view tree, annotations and additional relevant information
- Need to annotate trees with function gain and loss events

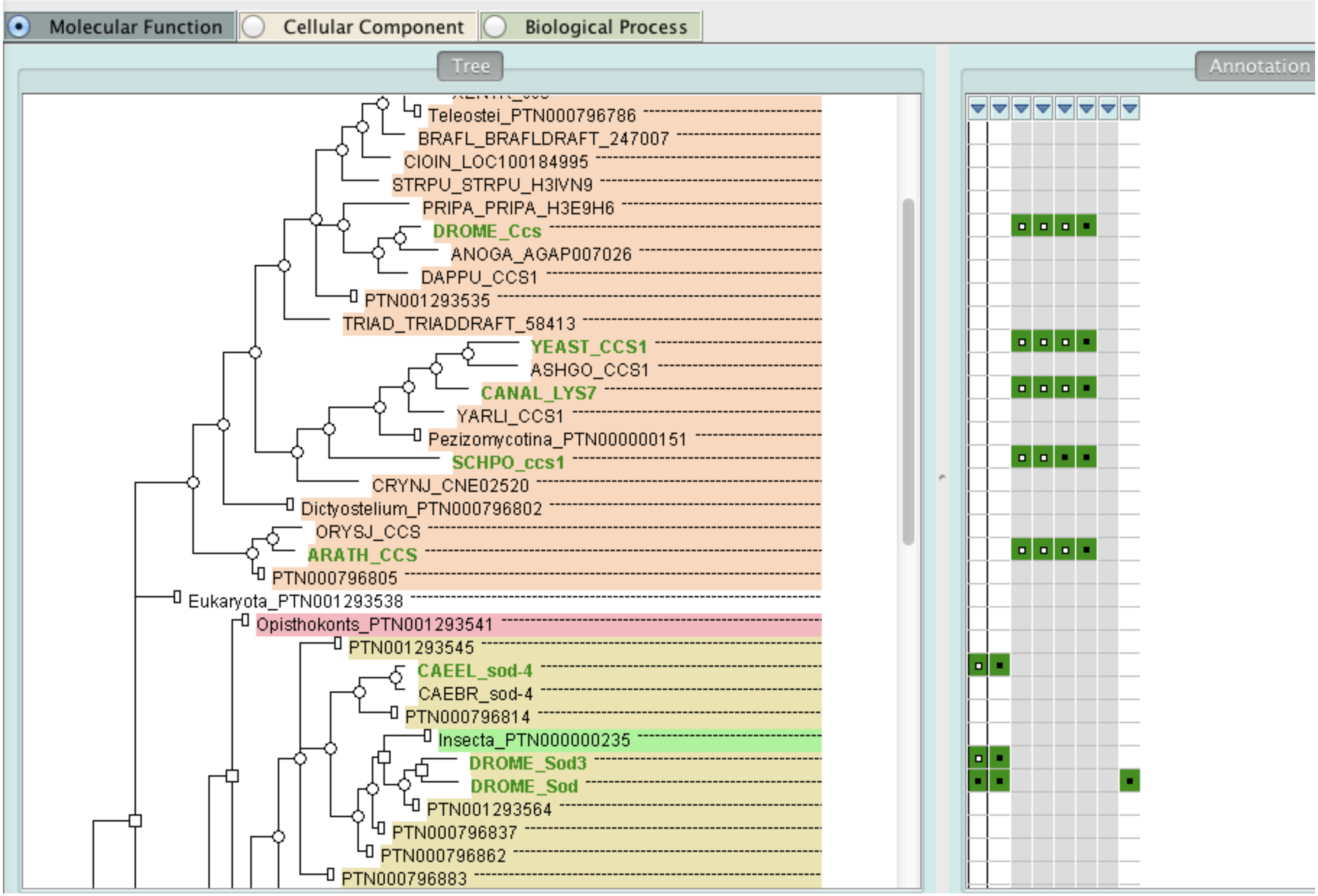
PAINT

Phylogenetic Annotation and Inference Tool

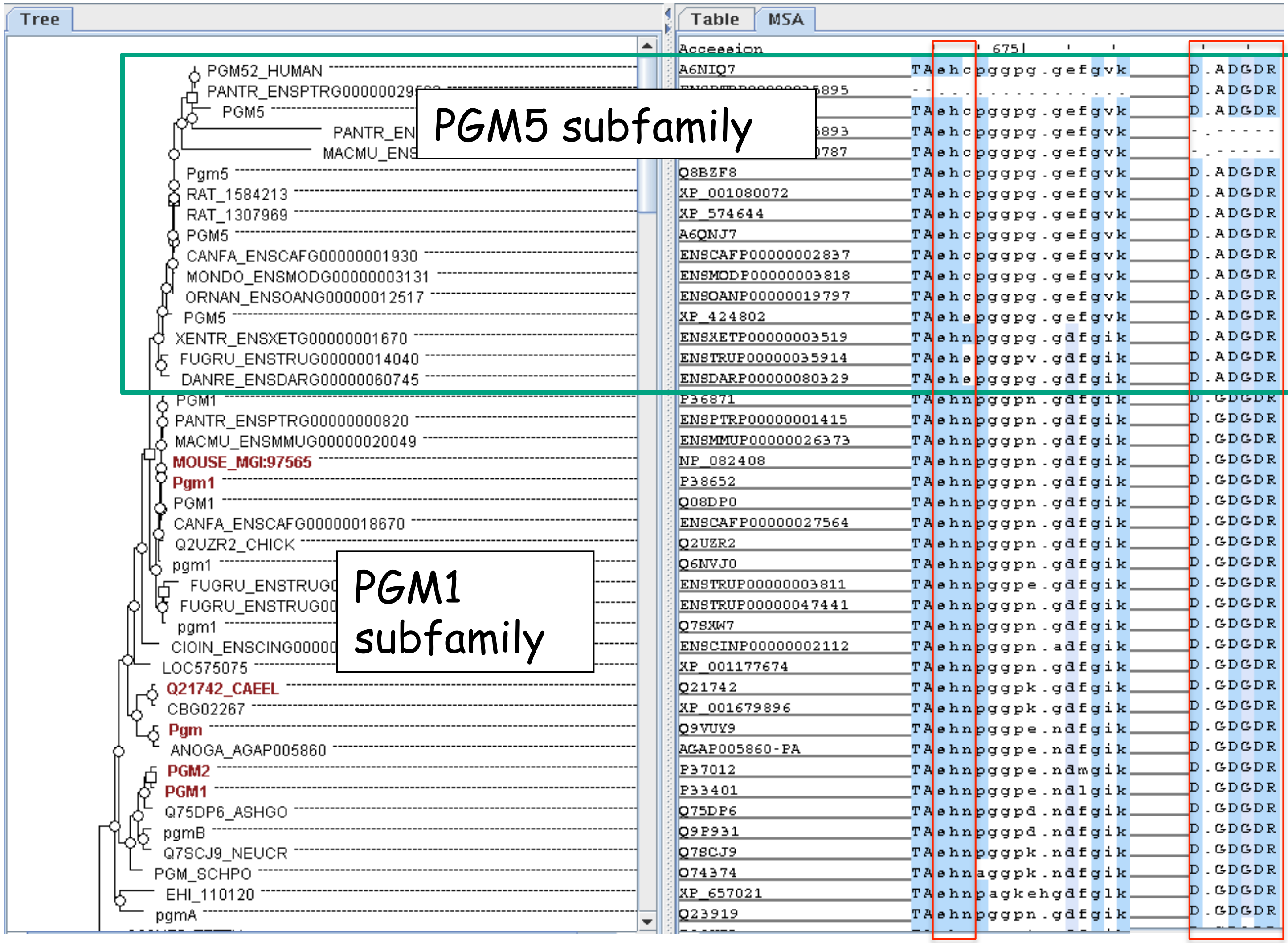


Integration of multiple types of biological knowledge

- GO annotations (from literature)
- Sequence feature annotations
  - Domains
  - Active sites
  - Modification sites
- Tree branch lengths


● Molecular Function
○ Cellular Component
○ Biological Process
Tree
Annotation
Teleostei_PTN000796786
BRAFL_BRAFLDRAFT_247007
CIOIN_LOC100184995
STRPU_STRPU_H3IVN9
PRIPA_PRIPA_H3E9H6
DROME_Ccs
ANOGA_AGAP007026
DAPPU_CCS1
PTN001293535
TRIAD_TRIADDRAFT_58413
YEAST_CCS1
ASHGO_CCS1
CANAL_LYS7
YARLI_CCS1
Pezizomycotina_PTN000000151
SCHPO_ccs1
CRYNJ_CNE02520
Dictyostelium_PTN000796802
ORYSJ_CCS
ARATH_CCS
PTN000796805
Eukaryota_PTN001293538
Opisthokonts_PTN001293541
PTN001293545
CAEEL_sod-4
CAEBR_sod-4
PTN000796814
Insecta_PTN000000235
DROME_Sod3
DROME_Sod
PTN001293564
PTN000796837
PTN000796862
PTN000796883

Evidence from specific protein sites



PGM5 subfamily

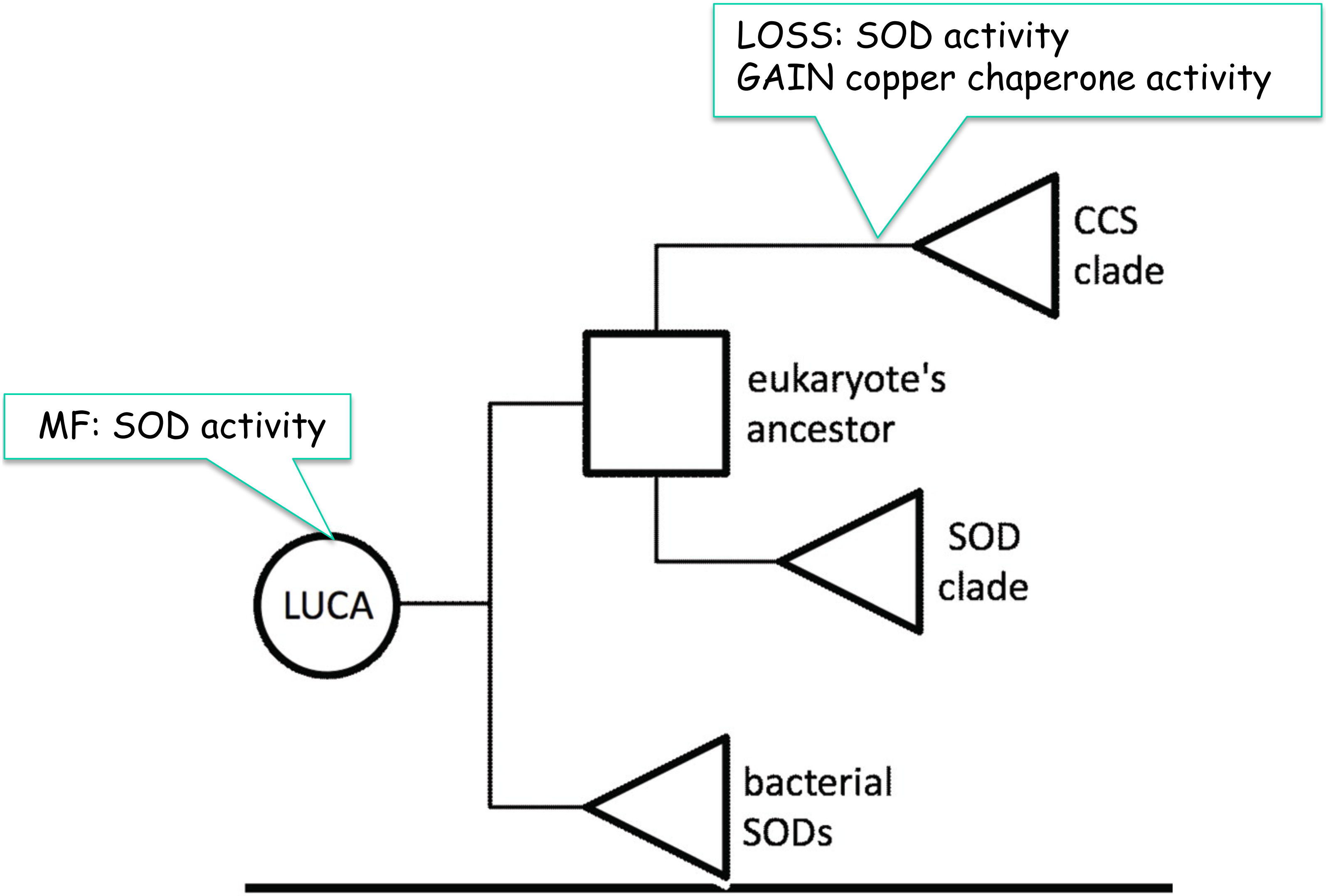
PGM1 subfamily

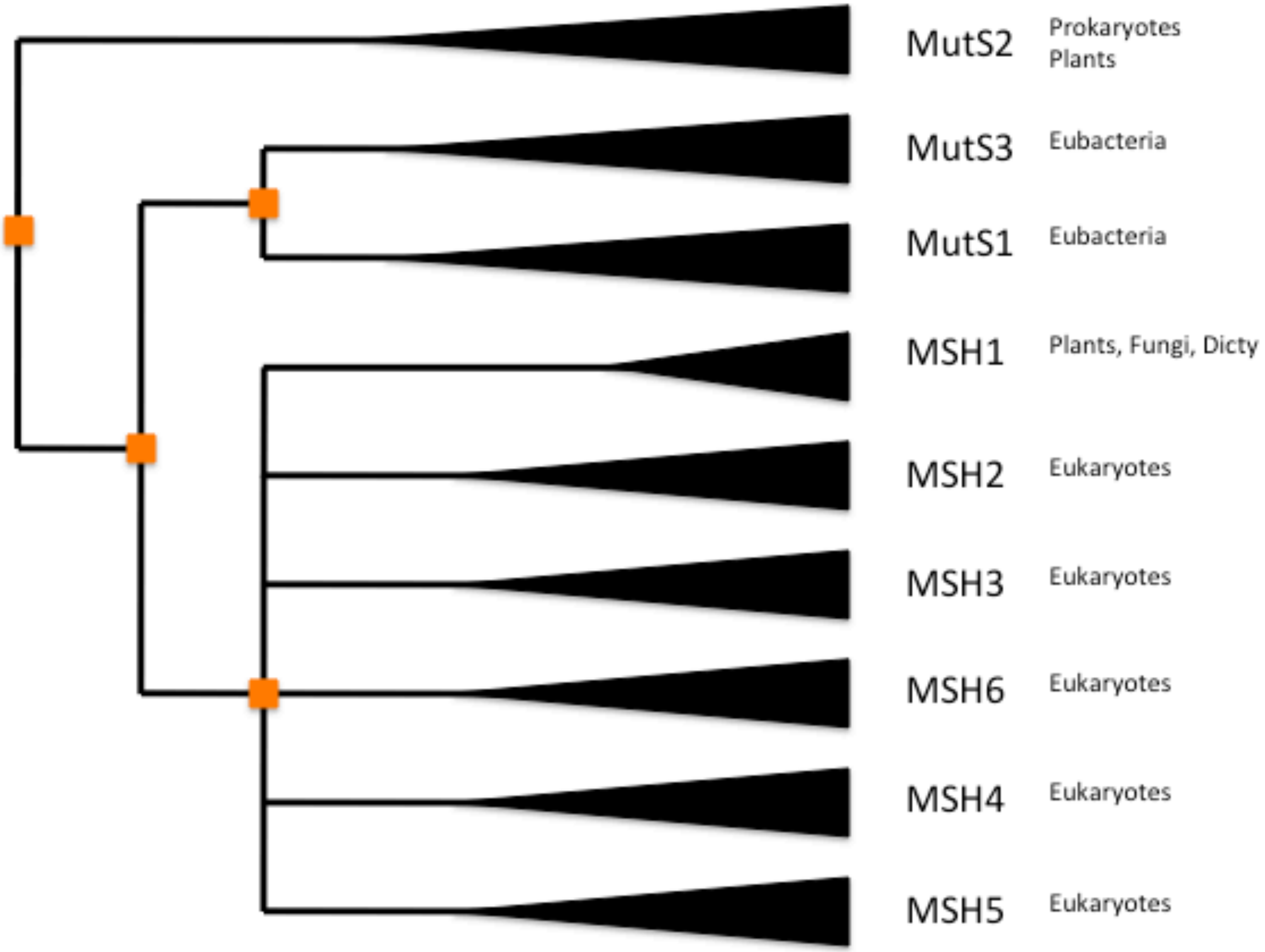
Accession	675	
A6NIQ7	TAehcpggpg.gefgvk	D.ADGDR
[illegible]5895	--..............	D.ADGDR
	TAehcpggpg.gefgvk	D.ADGDR
5893	TAehcpggpg.gefgvk	-.------
0787	TAehcpggpg.gefgvk	-.------
Q8BZF8	TAehcpggpg.gefgvk	D.ADGDR
XP_001080072	TAehcpggpg.gefgvk	D.ADGDR
XP_574644	TAehcpggpg.gefgvk	D.ADGDR
A6QNJ7	TAehcpggpg.gefgvk	D.ADGDR
ENSCAFP00000002837	TAehcpggpg.gefgvk	D.ADGDR
ENSMODP00000003818	TAehcpggpg.gefgvk	D.ADGDR
ENSOANP00000019797	TAehcpggpg.gefgvk	D.ADGDR
XP_424802	TAehepggpg.gefgvk	D.ADGDR
ENSXETP00000003519	TAehnpggpg.gdfgik	D.ADGDR
ENSTRUP00000035914	TAehepggpv.gdfgik	D.ADGDR
ENSDARP00000080329	TAehepggpg.gdfgik	D.ADGDR
P36871	TAehnpggpn.gdfgik	D.GDGDR
ENSPTRP00000001415	TAehnpggpn.gdfgik	D.GDGDR
ENSMMUP00000026373	TAehnpggpn.gdfgik	D.GDGDR
NP_082408	TAehnpggpn.gdfgik	D.GDGDR
P38652	TAehnpggpn.gdfgik	D.GDGDR
Q08DP0	TAehnpggpn.gdfgik	D.GDGDR
ENSCAFP00000027564	TAehnpggpn.gdfgik	D.GDGDR
Q2UZR2	TAehnpggpn.gdfgik	D.GDGDR
Q6NVJ0	TAehnpggpn.gdfgik	D.GDGDR
ENSTRUP00000003811	TAehnpggpe.gdfgik	D.GDGDR
ENSTRUP00000047441	TAehnpggpn.gdfgik	D.GDGDR
Q7SXW7	TAehnpggpn.gdfgik	D.GDGDR
ENSCINP00000002112	TAehnpggpn.adfgik	D.GDGDR
XP_001177674	TAehnpggpn.gdfgik	D.GDGDR
Q21742	TAehnpggpk.gdfgik	D.GDGDR
XP_001679896	TAehnpggpk.gdfgik	D.GDGDR
Q9VUY9	TAehnpggpe.ndfgik	D.GDGDR
AGAP005860-PA	TAehnpggpe.ndfgik	D.GDGDR
P37012	TAehnpggpe.ndmgik	D.GDGDR
P33401	TAehnpggpe.ndlgik	D.GDGDR
Q75DP6	TAehnpggpd.ndfgik	D.GDGDR
Q9P931	TAehnpggpd.ndfgik	D.GDGDR
Q7SCJ9	TAehnpggpk.ndfgik	D.GDGDR
O74374	TAehnaggpk.ndfgik	D.GDGDR
XP_657021	TAehnpagkehgdfglk	D.GDGDR
Q23919	TAehnpggpn.gdfgik	D.GDGDR

Curated active site information from CDD (cd03085)

- **phosphoglucomutase activity** ***LOSS*** **phosphoglucomutase activity (PGM5 subfamily)**

IBA: Loss of function can be annotated

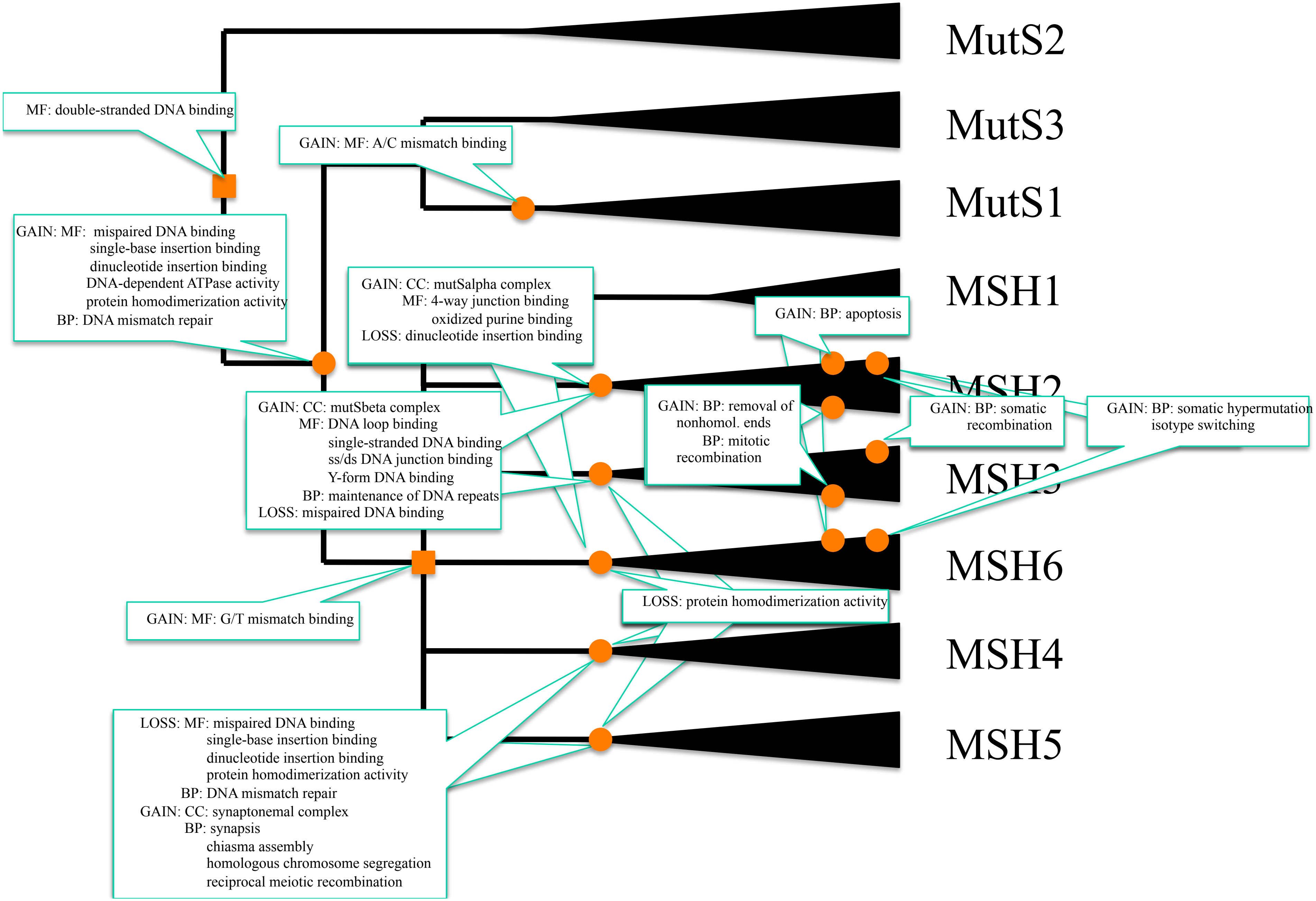




Nucleic Acids Res. 2007;35(22):7591-603. Epub 2007 Oct 26.

The origins and early evolution of DNA mismatch repair genes--multiple horizontal gene transfers and co-evolution.

Lin Z, Nei M, Ma H.



MF: double-stranded DNA binding

GAIN: MF: A/C mismatch binding

GAIN: MF: mispaired DNA binding
single-base insertion binding
dinucleotide insertion binding
DNA-dependent ATPase activity
protein homodimerization activity
BP: DNA mismatch repair

GAIN: CC: mutSalpha complex
MF: 4-way junction binding
oxidized purine binding
LOSS: dinucleotide insertion binding

GAIN: BP: apoptosis

GAIN: CC: mutSbeta complex
MF: DNA loop binding
single-stranded DNA binding
ss/ds DNA junction binding
Y-form DNA binding
BP: maintenance of DNA repeats
LOSS: mispaired DNA binding

GAIN: BP: removal of
nonhomol. ends
BP: mitotic
recombination

GAIN: BP: somatic
recombination

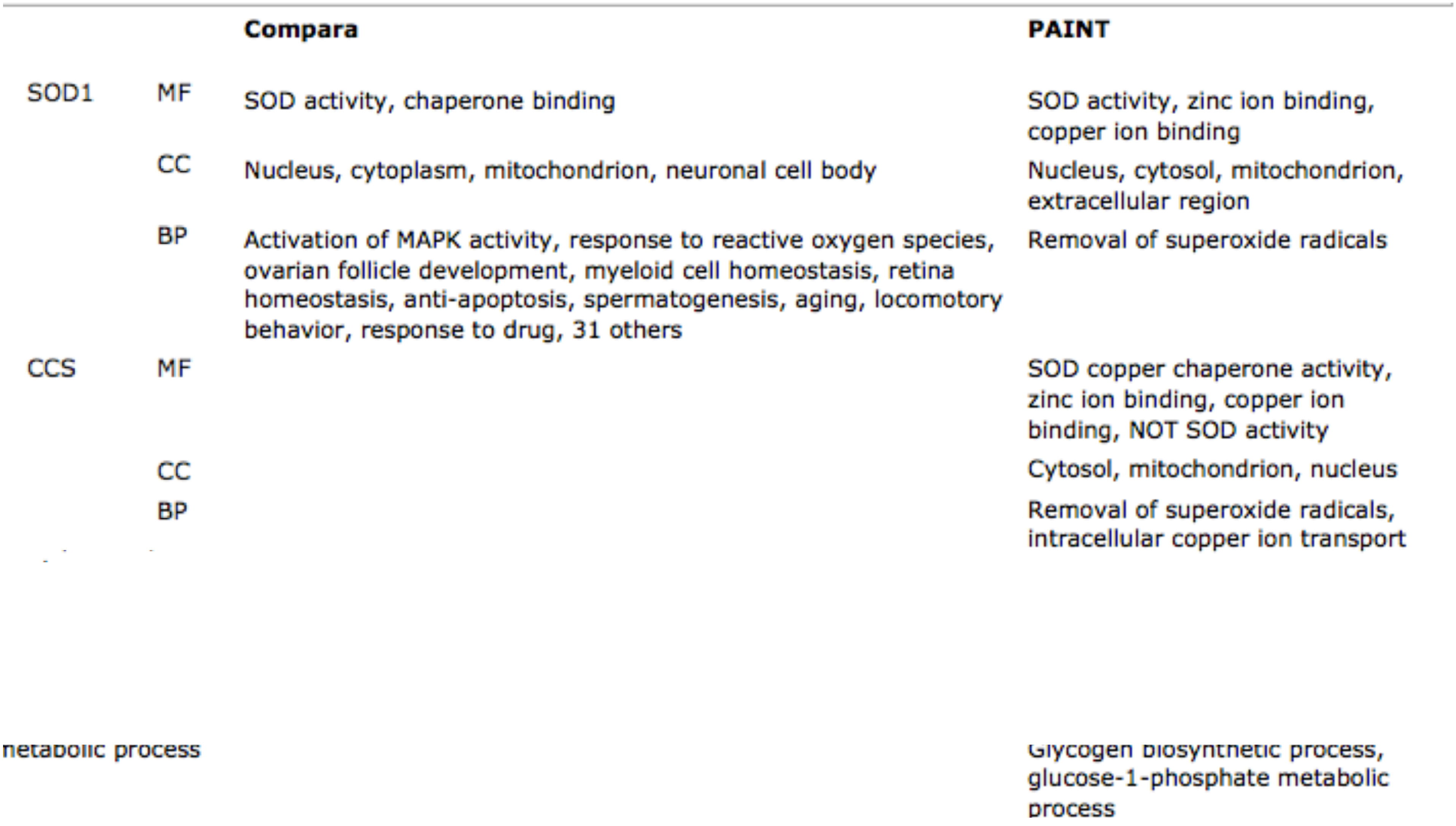
GAIN: BP: somatic hypermutation
isotype switching

GAIN: MF: G/T mismatch binding

LOSS: protein homodimerization activity

LOSS: MF: mispaired DNA binding
single-base insertion binding
dinucleotide insertion binding
protein homodimerization activity
BP: DNA mismatch repair
GAIN: CC: synaptonemal complex
BP: synapsis
chiasma assembly
homologous chromosome segregation
reciprocal meiotic recombination

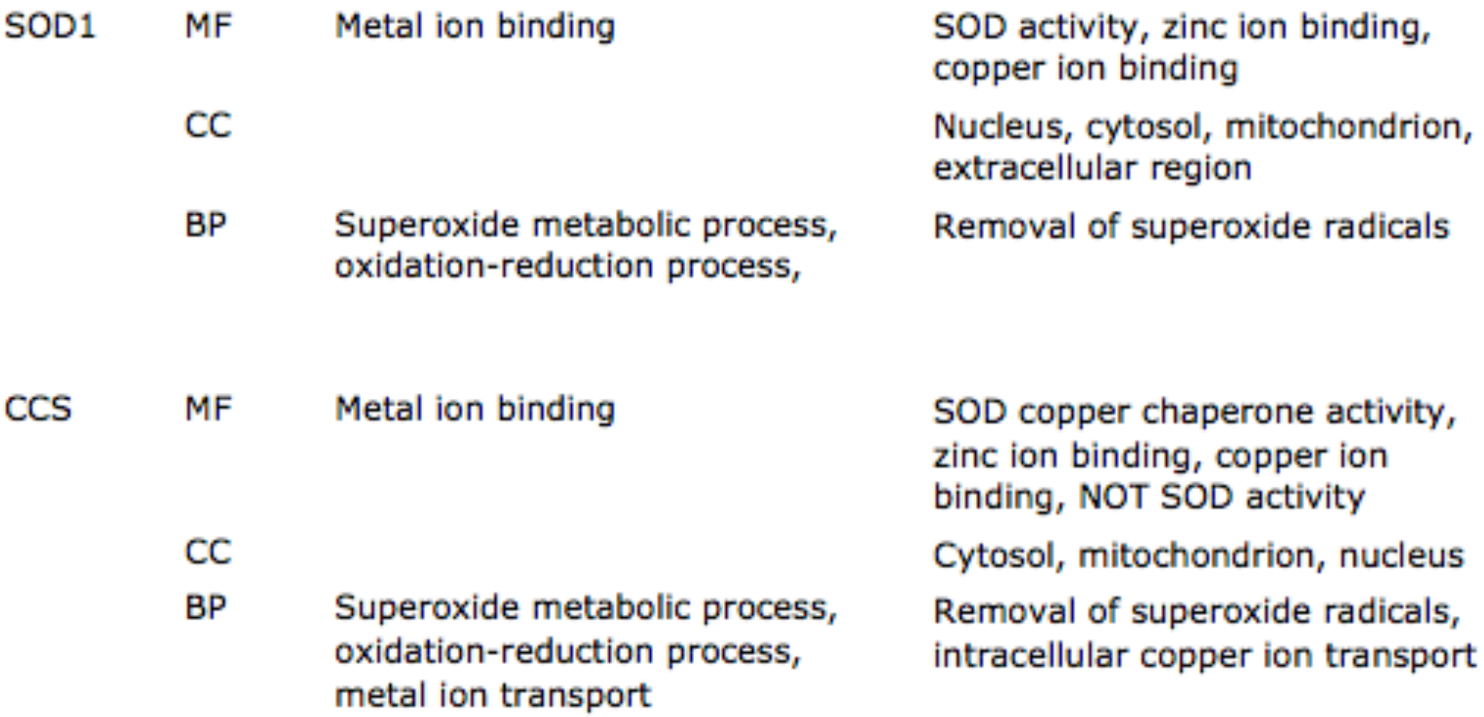
IBA vs ISO for SOD family
Only most informative annotations are propagated
Inferences can be made from non-vertebrate homologs



		Compara	PAINT
SOD1	MF	SOD activity, chaperone binding	SOD activity, zinc ion binding, copper ion binding
	CC	Nucleus, cytoplasm, mitochondrion, neuronal cell body	Nucleus, cytosol, mitochondrion, extracellular region
	BP	Activation of MAPK activity, response to reactive oxygen species, ovarian follicle development, myeloid cell homeostasis, retina homeostasis, anti-apoptosis, spermatogenesis, aging, locomotory behavior, response to drug, 31 others	Removal of superoxide radicals
CCS	MF		SOD copper chaperone activity, zinc ion binding, copper ion binding, NOT SOD activity
	CC		Cytosol, mitochondrion, nucleus
	BP		Removal of superoxide radicals, intracellular copper ion transport
metabolic process			Glycogen biosynthetic process, glucose-1-phosphate metabolic process

IBA vs IEA (InterPro) for SOD family

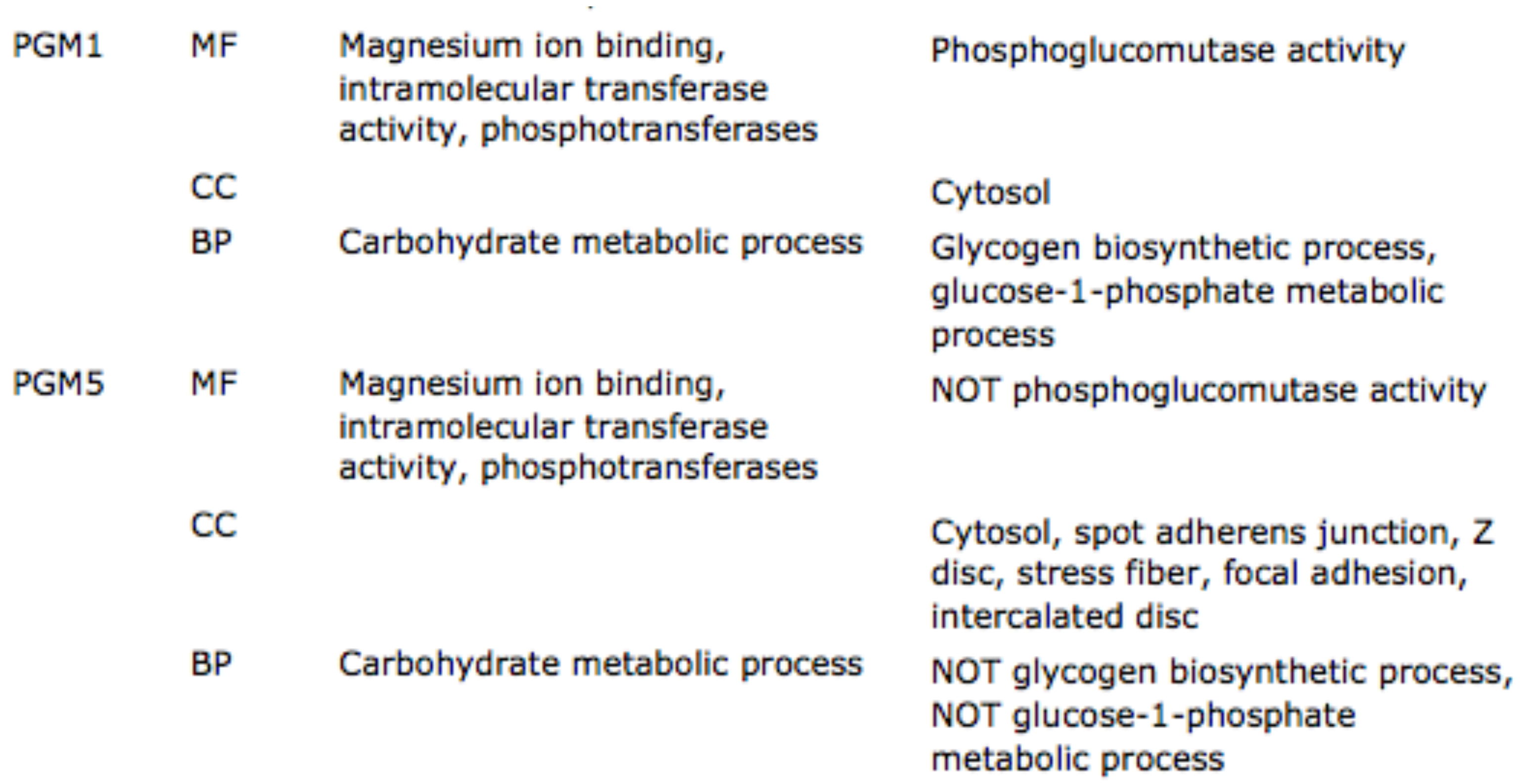
Higher specificity



SOD1	MF	Metal ion binding	SOD activity, zinc ion binding, copper ion binding
	CC		Nucleus, cytosol, mitochondrion, extracellular region
	BP	Superoxide metabolic process, oxidation-reduction process,	Removal of superoxide radicals
CCS	MF	Metal ion binding	SOD copper chaperone activity, zinc ion binding, copper ion binding, NOT SOD activity
	CC		Cytosol, mitochondrion, nucleus
	BP	Superoxide metabolic process, oxidation-reduction process, metal ion transport	Removal of superoxide radicals, intracellular copper ion transport

IBA vs IEA (InterPro) for PGM family
Higher specificity

Fewer false positive predictions



PGM1	MF	Magnesium ion binding, intramolecular transferase activity, phosphotransferases	Phosphoglucomutase activity
	CC		Cytosol
	BP	Carbohydrate metabolic process	Glycogen biosynthetic process, glucose-1-phosphate metabolic process
PGM5	MF	Magnesium ion binding, intramolecular transferase activity, phosphotransferases	NOT phosphoglucomutase activity
	CC		Cytosol, spot adherens junction, Z disc, stress fiber, focal adhesion, intercalated disc
	BP	Carbohydrate metabolic process	NOT glycogen biosynthetic process, NOT glucose-1-phosphate metabolic process

Bottom line

- Experimental evidence codes remain the “gold standard”
  - BUT only available for a small subset of well-studied organisms
  - NOTE: be aware of indirect effects annotated from IMP and IEP, you may want to filter these for some applications
- The next most reliable and specific tier is IBA, followed by IEA, then followed by ISS and IC
- If you want a more concise “summary” list of GO annotations, use IBA

Where to get the data

- GO annotations
  - Gene Ontology website
- Pathway data in SBML format
  - Pathway Commons website

- 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