Gene function annotation and Gene set analysis

Paul D. Thomas, Ph.D.
University of Southern California
July 2015

In this lecture

- Introduction to ontologies 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
- Phylogenetic analysis of function
  - Importance of homology inference
- Basics of enrichment analysis of genomic data
  - Clues about biological interpretation of genomics experiments

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)

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

Protein function ontologies

- Gene Ontology (GO)
- Pathway Ontologies
  - 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 (from 200?)

is_a relations from the GO are NOT shown by Entrez

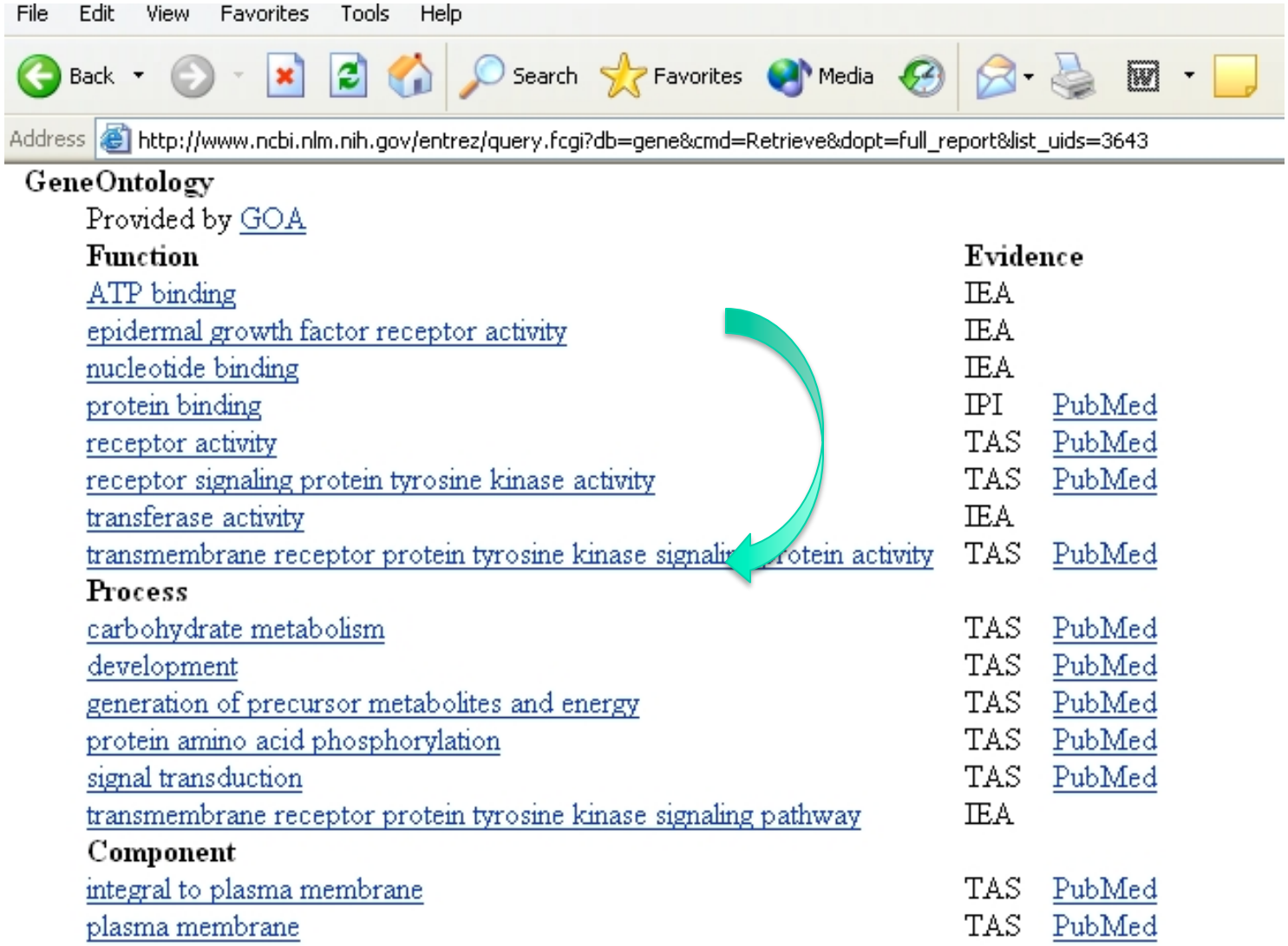
GeneOntology

Provided by GOA

Function	**Evidence**	
ATP binding	IEA	
epidermal growth factor receptor activity	IEA	
nucleotide binding	IEA	
protein binding	IPI	PubMed
receptor activity	TAS	PubMed
receptor signaling protein tyrosine kinase activity	TAS	PubMed
transferase activity	IEA	
transmembrane receptor protein tyrosine kinase signaling protein activity	TAS	PubMed
Process		
carbohydrate metabolism	TAS	PubMed
development	TAS	PubMed
generation of precursor metabolites and energy	TAS	PubMed
protein amino acid phosphorylation	TAS	PubMed
signal transduction	TAS	PubMed
transmembrane receptor protein tyrosine kinase signaling pathway	IEA	
Component		
integral to plasma membrane	TAS	PubMed
plasma membrane	TAS	PubMed

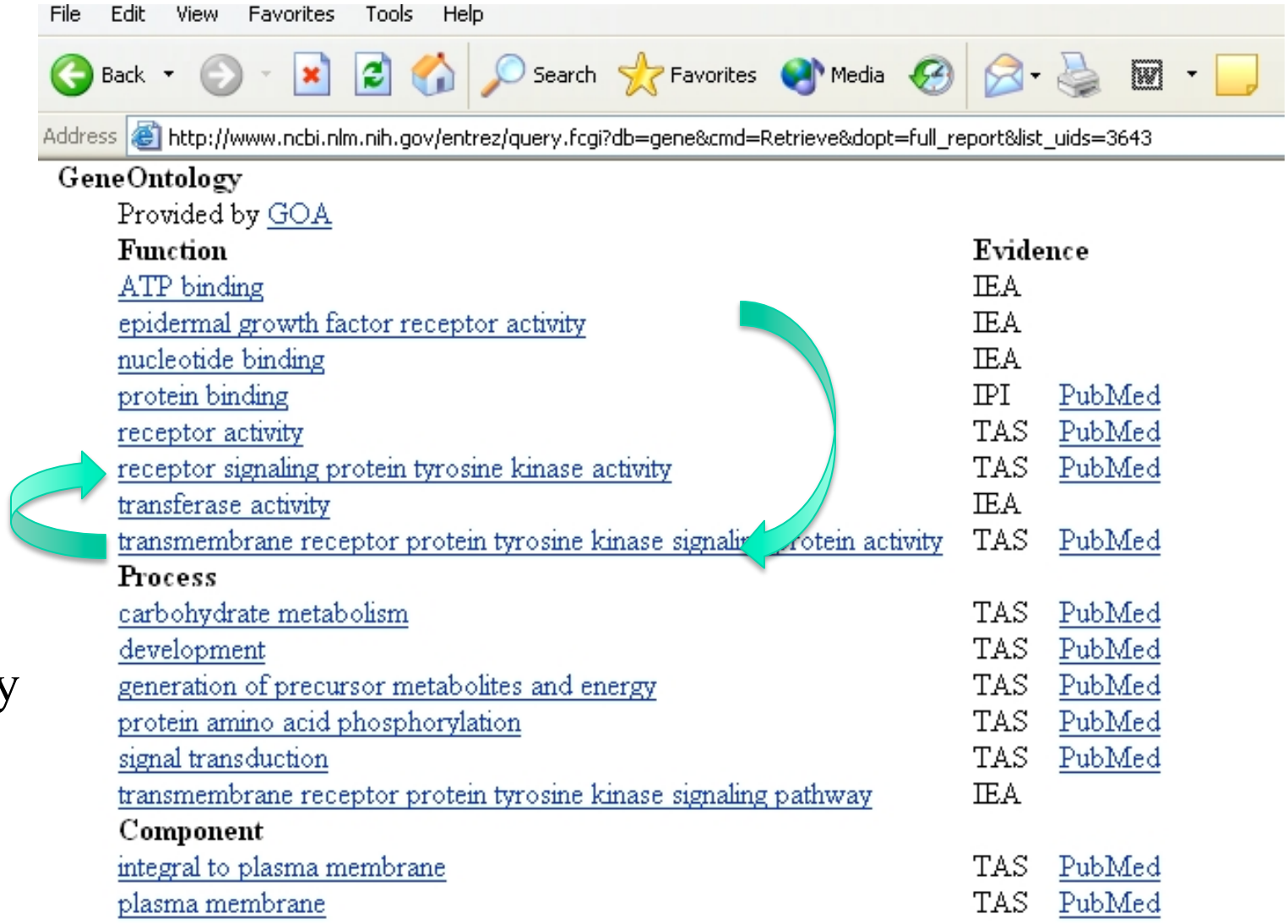
Entrez Gene: INSR (from 200?)

is_a relations from the GO are NOT shown by Entrez



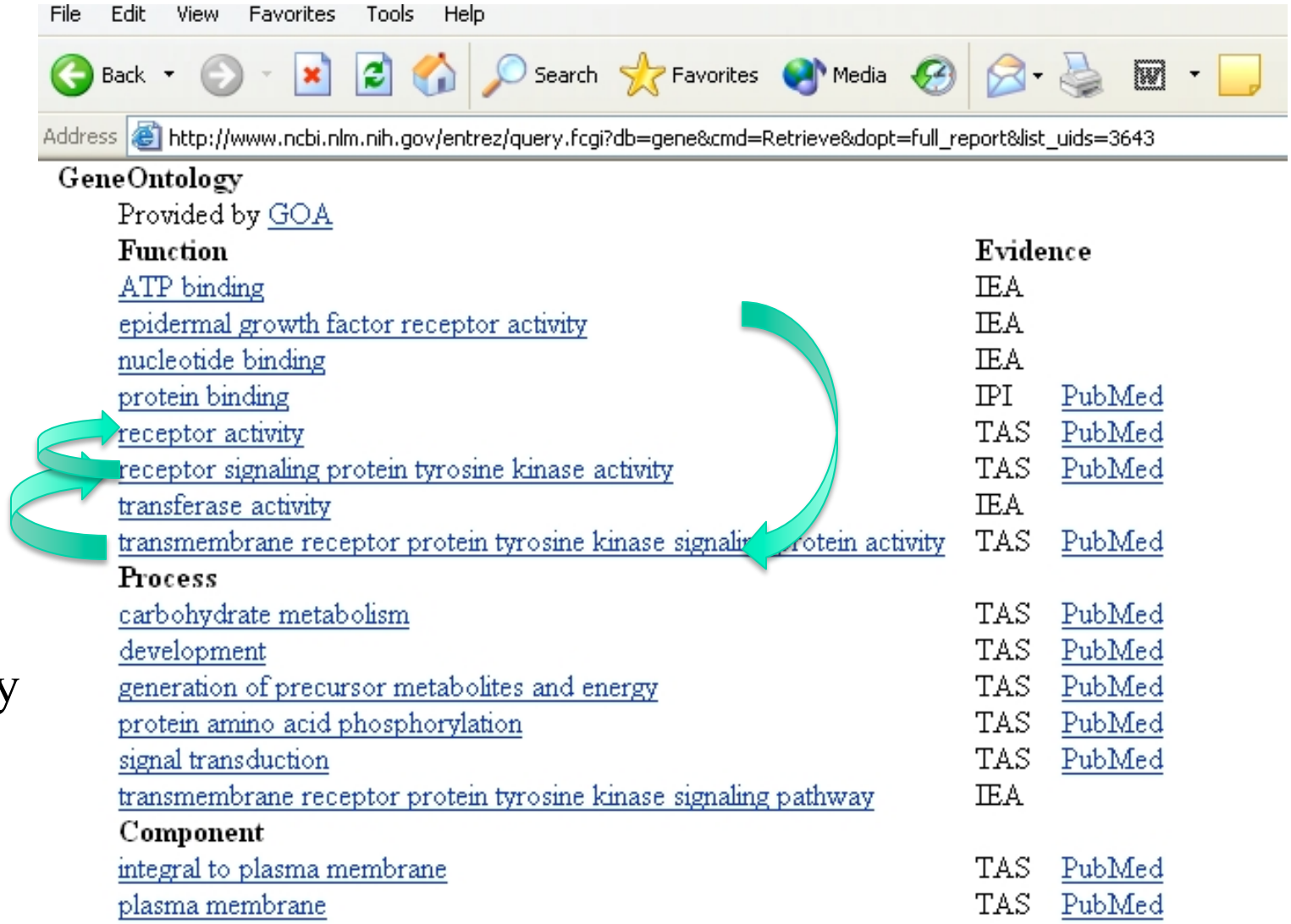
Entrez Gene: INSR (from 200?)

is_a relations from the GO are NOT shown by Entrez



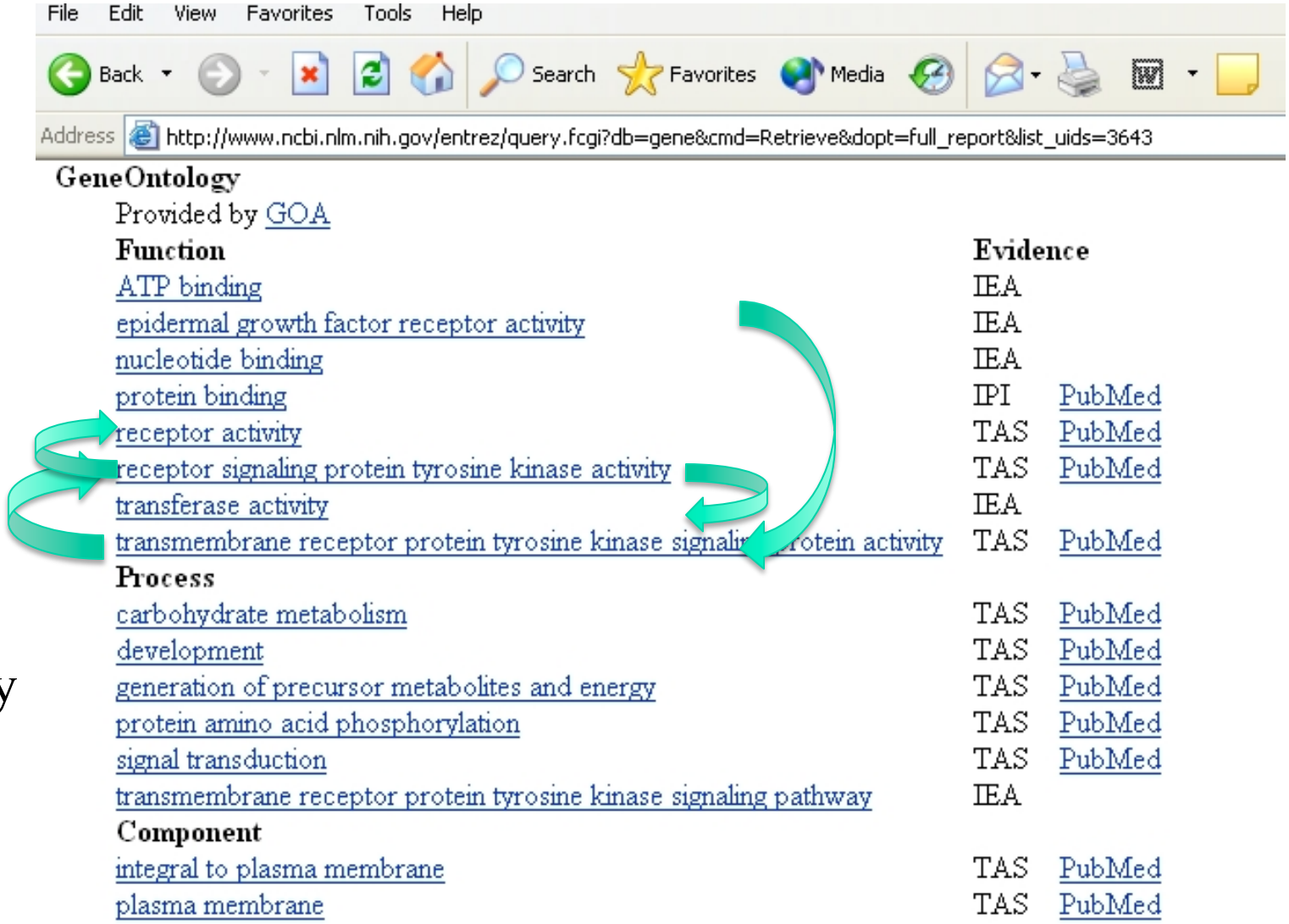
Entrez Gene: INSR (from 200?)

is_a relations from the GO are NOT shown by Entrez



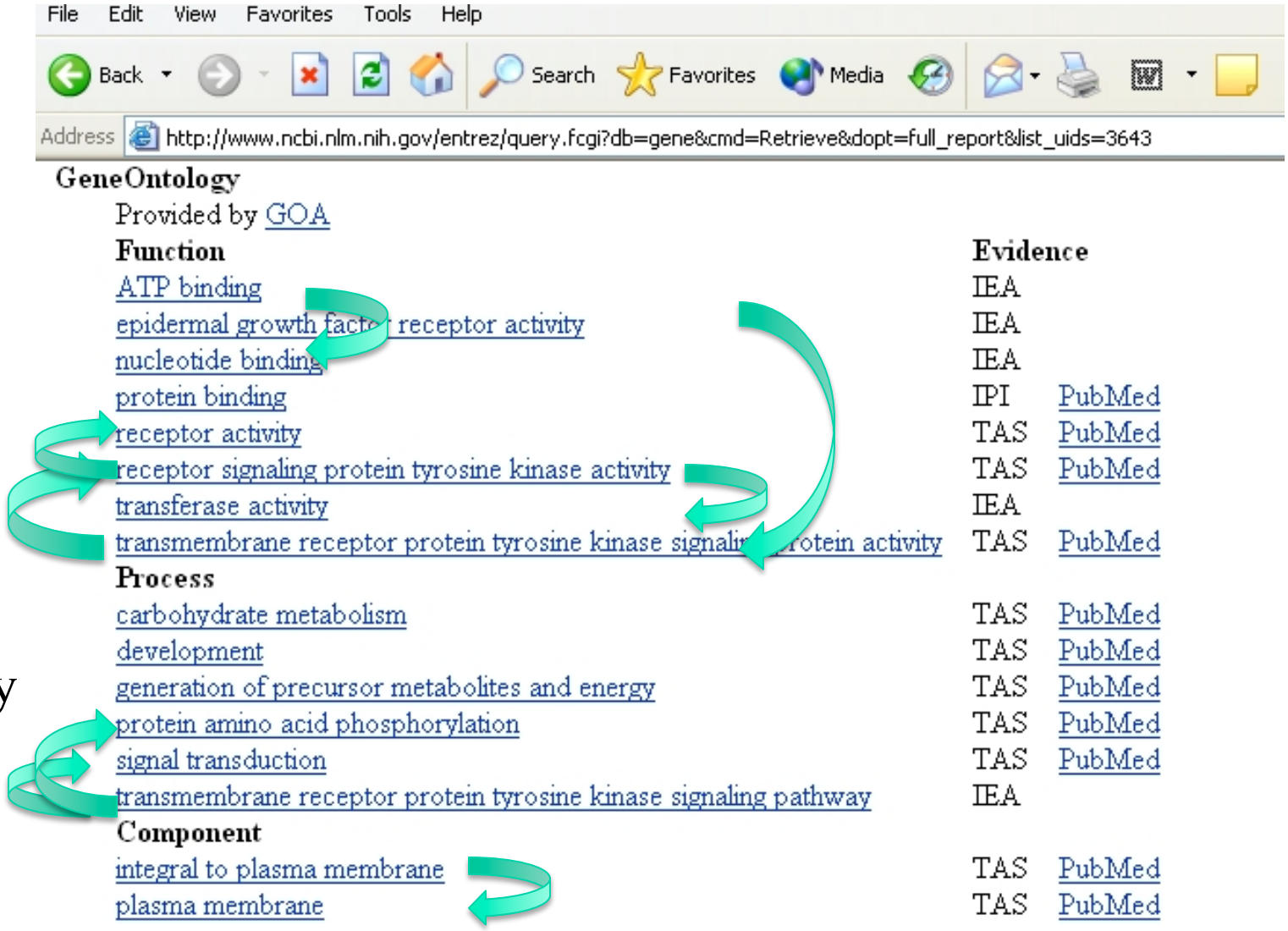
Entrez Gene: INSR (from 200?)

is_a relations from the GO are NOT shown by Entrez



Entrez Gene: INSR (from 200?)

is_a relations from the GO are NOT shown by Entrez

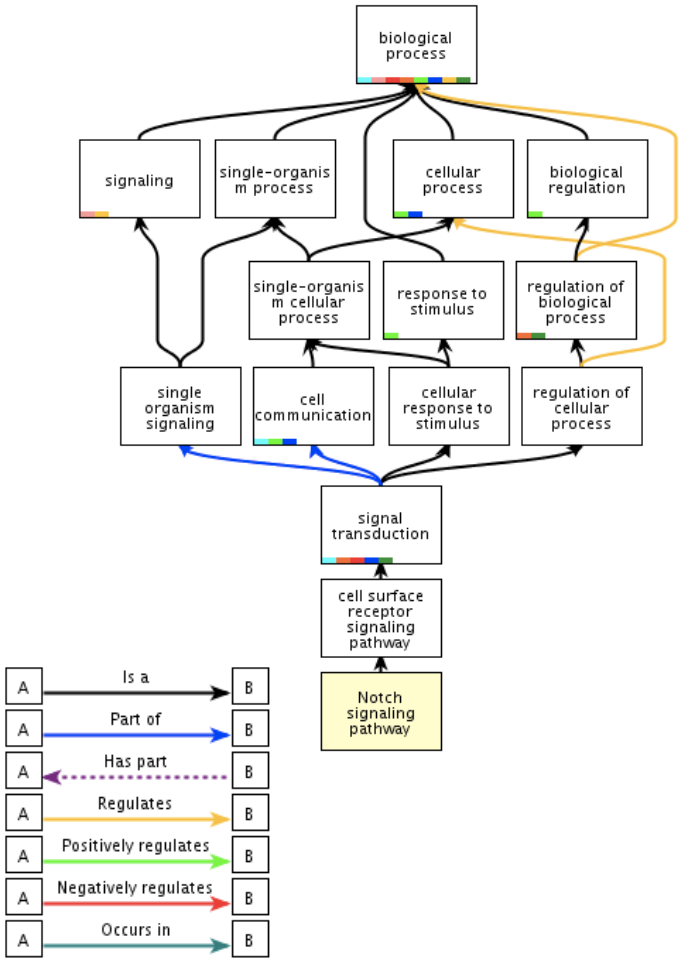


Pathway representations

- 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

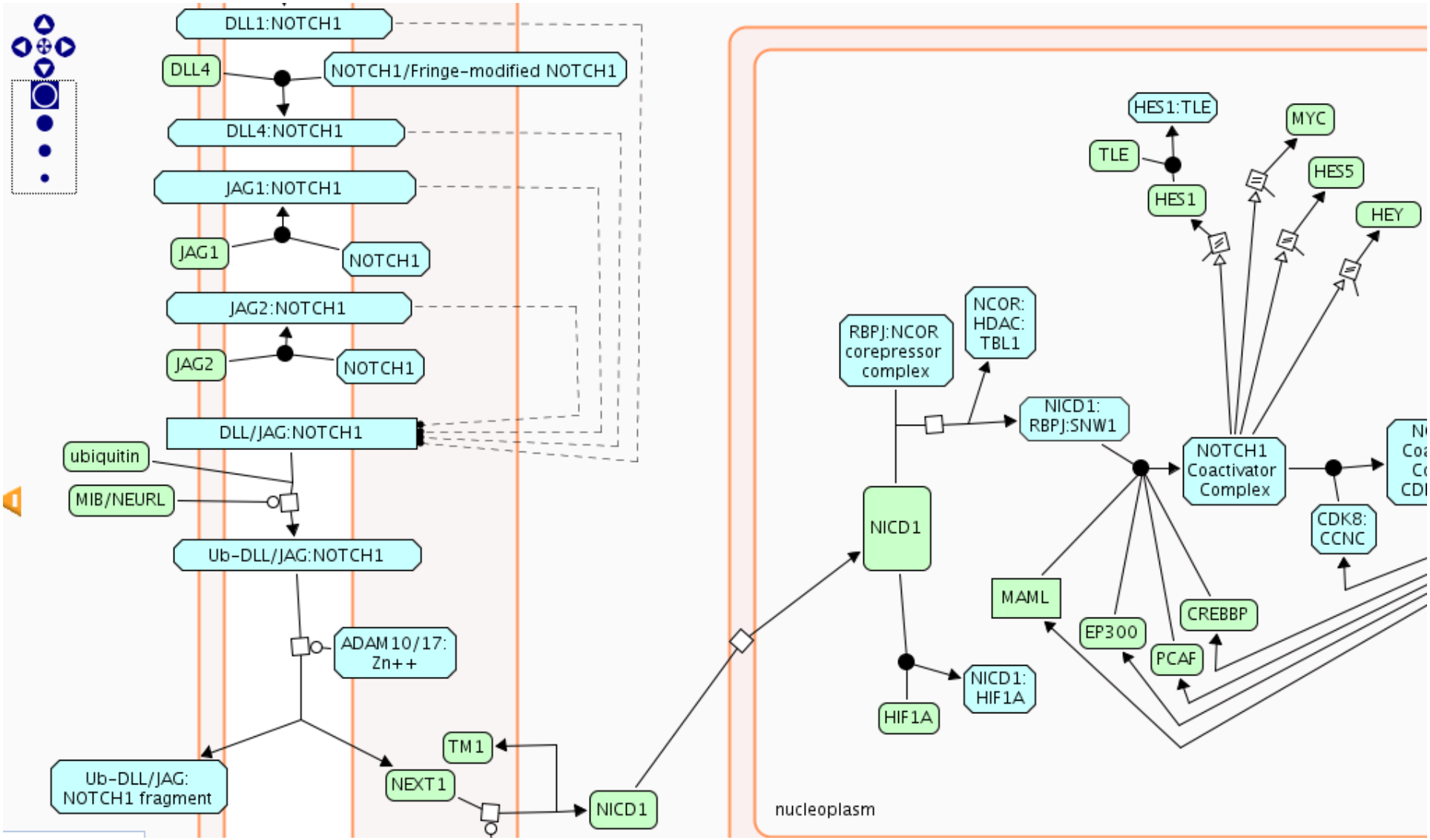


Relations to more specific classes

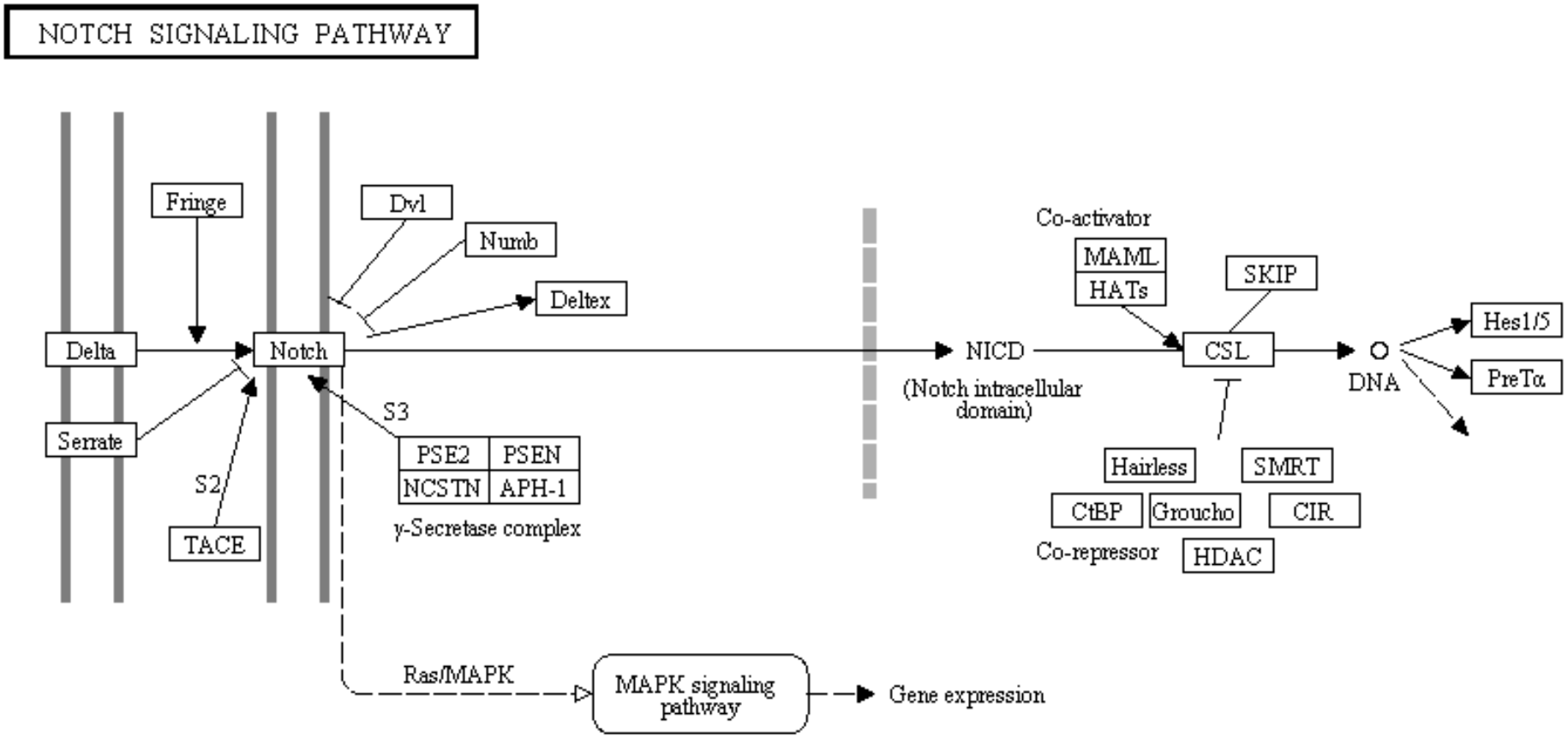
GO:0007219 Notch signaling pathway

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

What do the evidence codes mean?

- "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

What do the evidence codes mean?

- Reviewed homology evidence
  - ISS: inferred from sequence similarity
  - IBA: inferred from biological ancestry
- Important distinctions
  - ISS: pairwise, usually from BLAST, manually reviewed and assigned
    - No consistent rules
  - IBA: phylogenetic context, manually reviewed and assigned using phylogenetic annotation

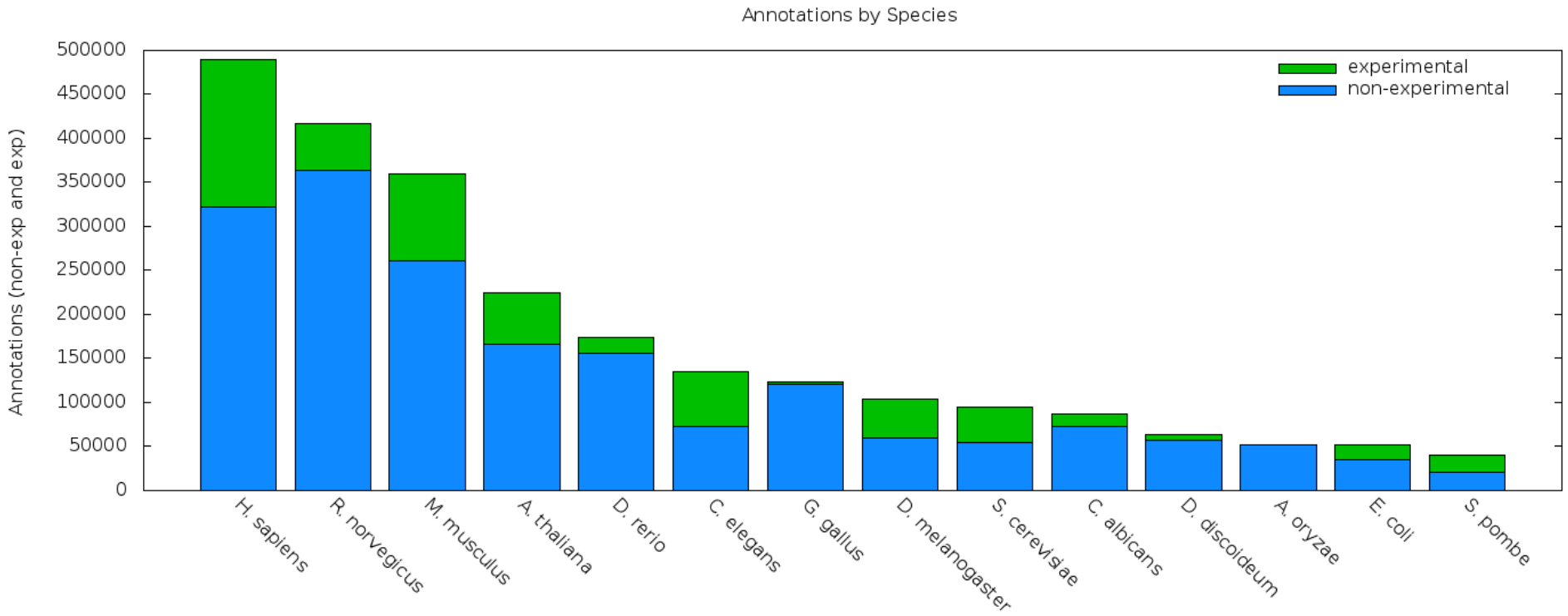
What do the evidence codes mean?

- Unreviewed
  - IEA: inferred from electronic annotation
    - These are primarily based on homology information
  - ISO: inferred from sequence orthology
- IEA annotations far outnumber any other type, two major types
  - 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

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 annotations as of July 1, 2015



- GO experimental annotations cover a few major "model organisms"
- Most GO annotations are based on homology

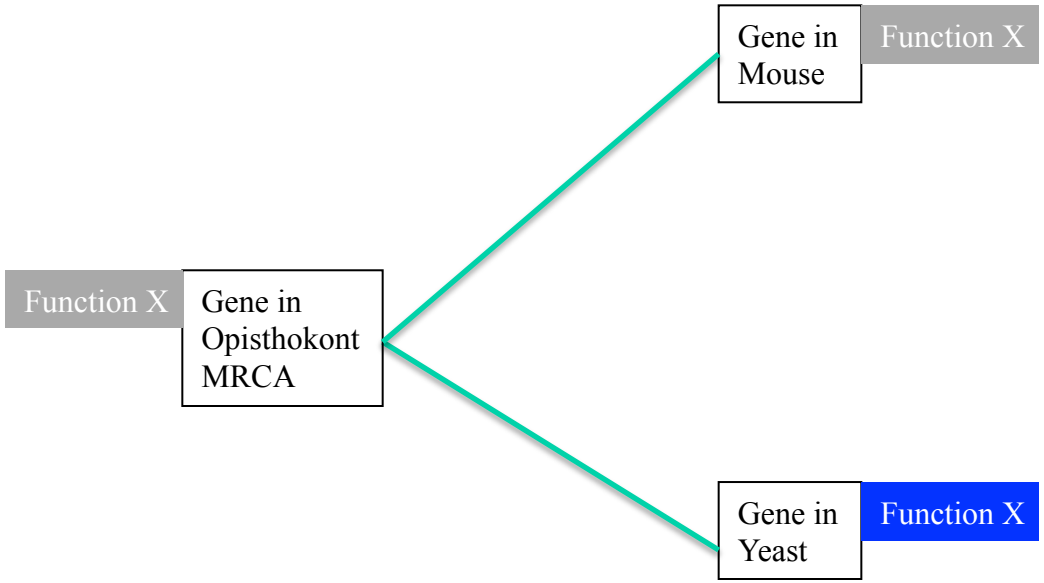
Homology is still the most informative predictor of function

- 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.

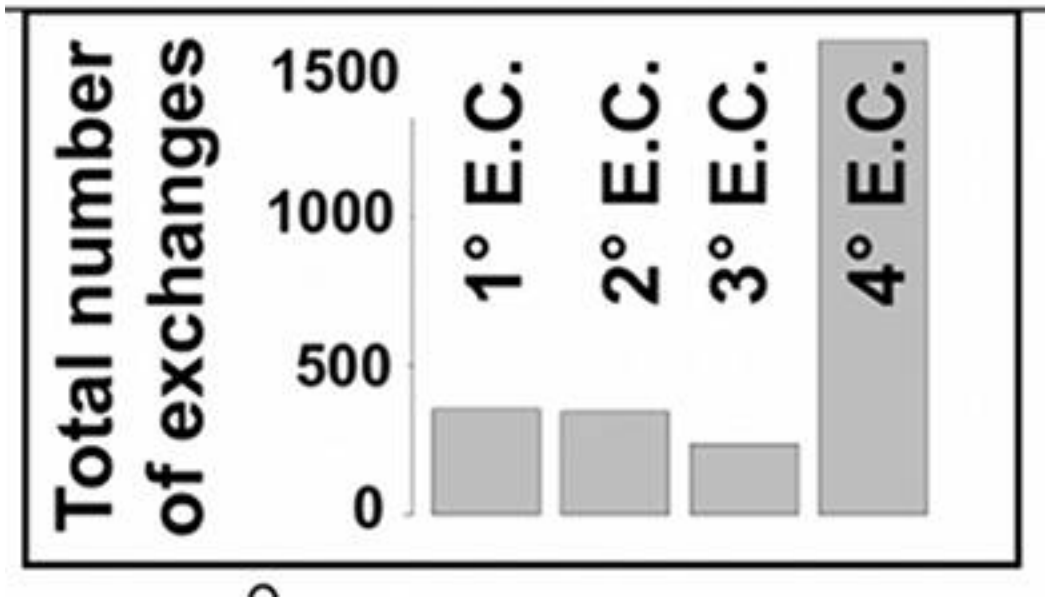
Homology-based annotation

- AKA "transitive annotation"
- An annotation of function of one gene is "transferred" (or "propagated") to another gene, based on homology between gene or protein sequences

Implicitly assumes that the last common ancestor of those two genes had that function, which was then subsequently inherited by both genes



N.B. Different GO functions evolve at different rates in protein families



Enzyme mechanism (1-3) evolves more slowly than substrate specificity (4)

In general, easier to predict by homology a more general GO class than a more specific one

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

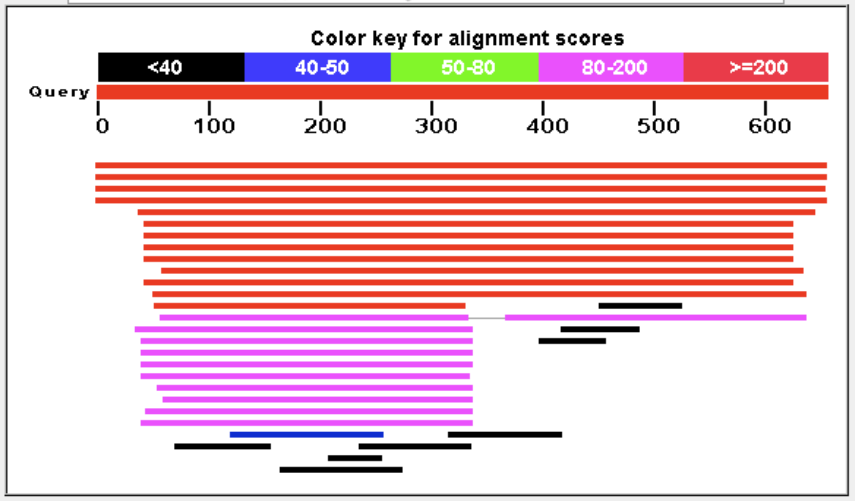
Exploring the evolution of novel enzyme functions within structurally defined protein superfamilies.

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

ISS is pairwise:

example BLAST results for human MTHFR vs. SwissProt database



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

ISO: Orthologs (vs. paralogs)

- 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

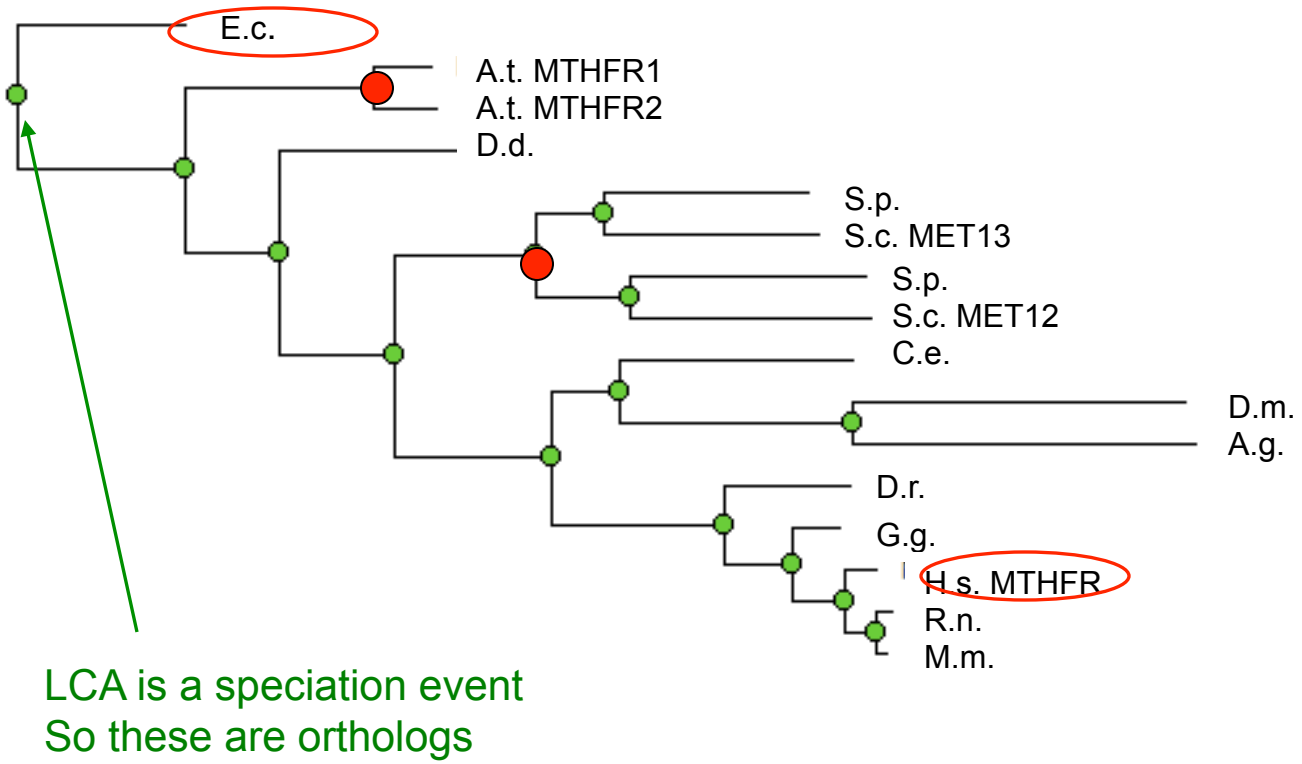
Why orthology is confusing

- 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!

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.

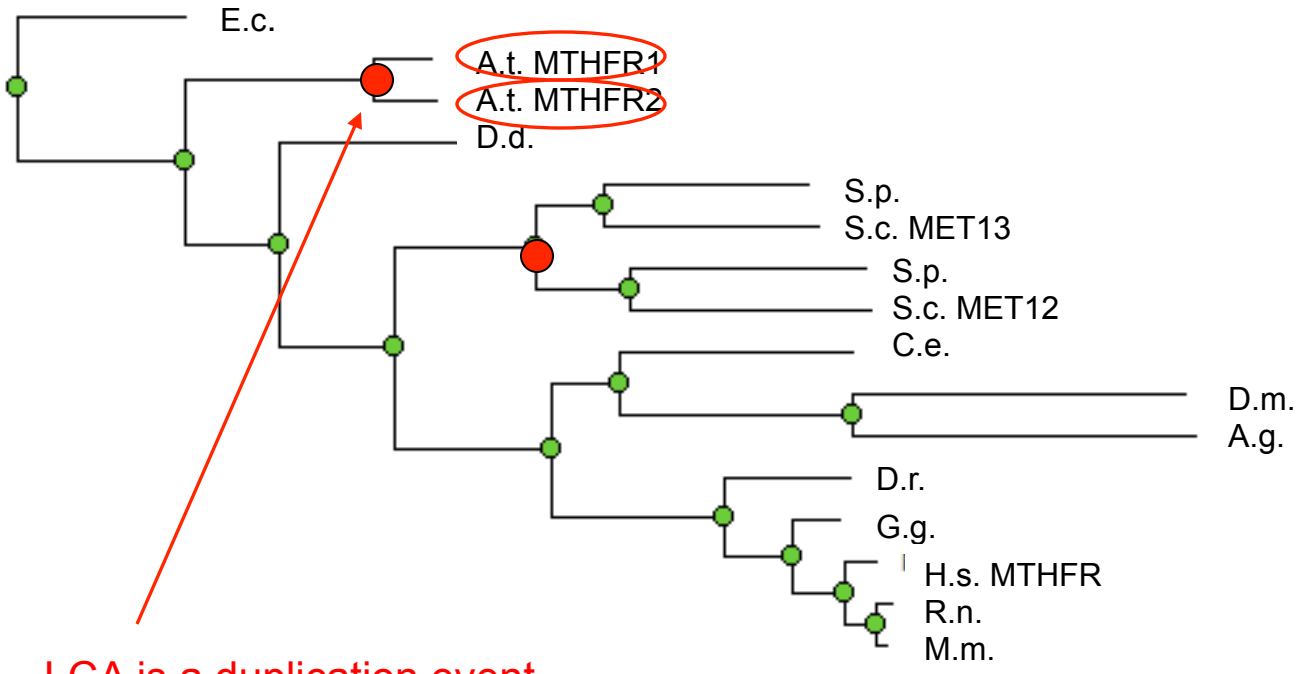
Orthology
only defined for PAIRS of genes



LCA is a speciation event
So these are orthologs

Two genes are orthologs if their LCA was a speciation event

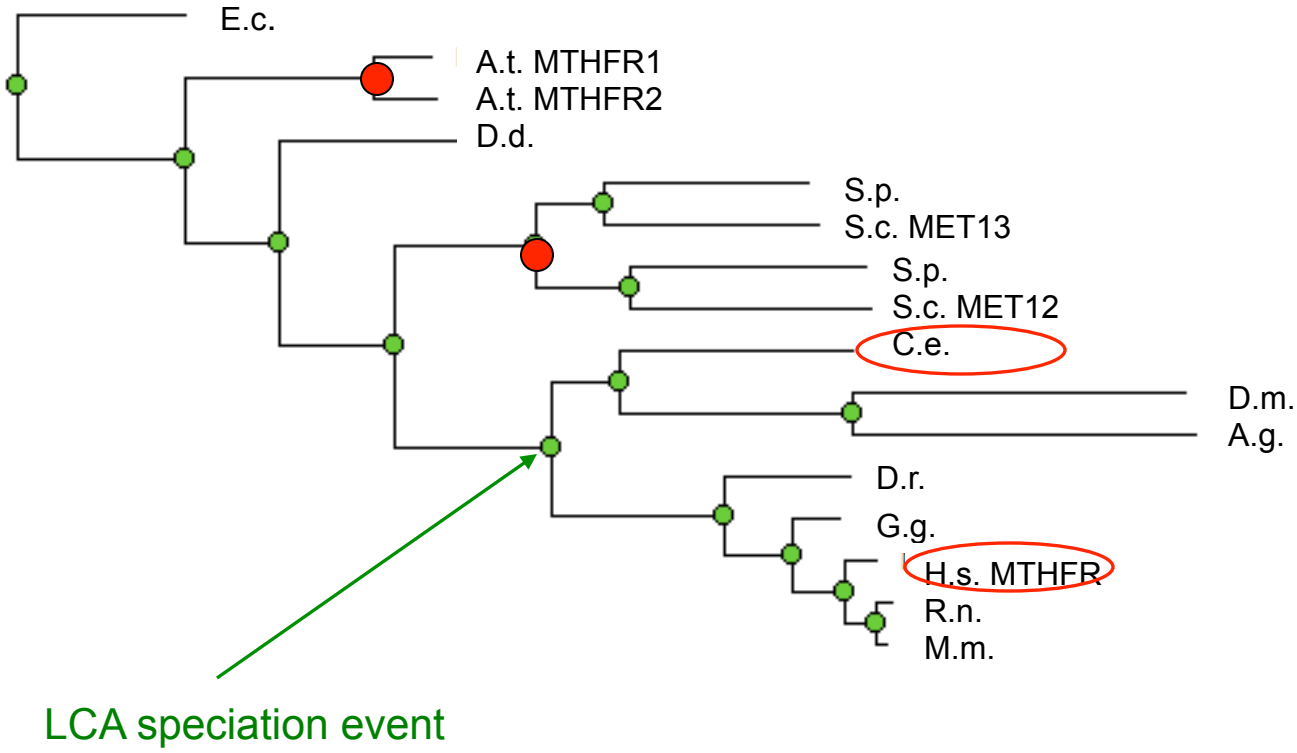
Paralogy
only defined for PAIRS of genes



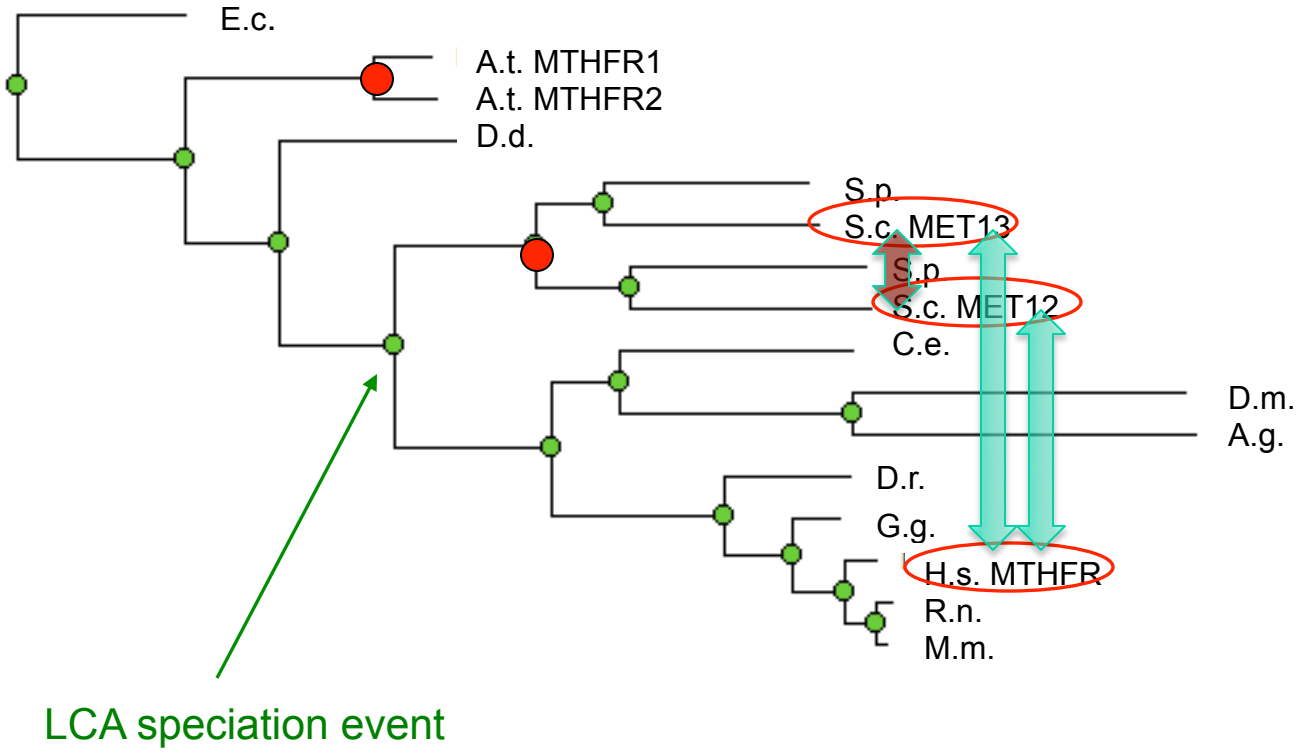
LCA is a duplication event
So these are paralogs

Two genes are paralogs if their LCA was a duplication event

Orthology is simple when there are no duplications following speciation

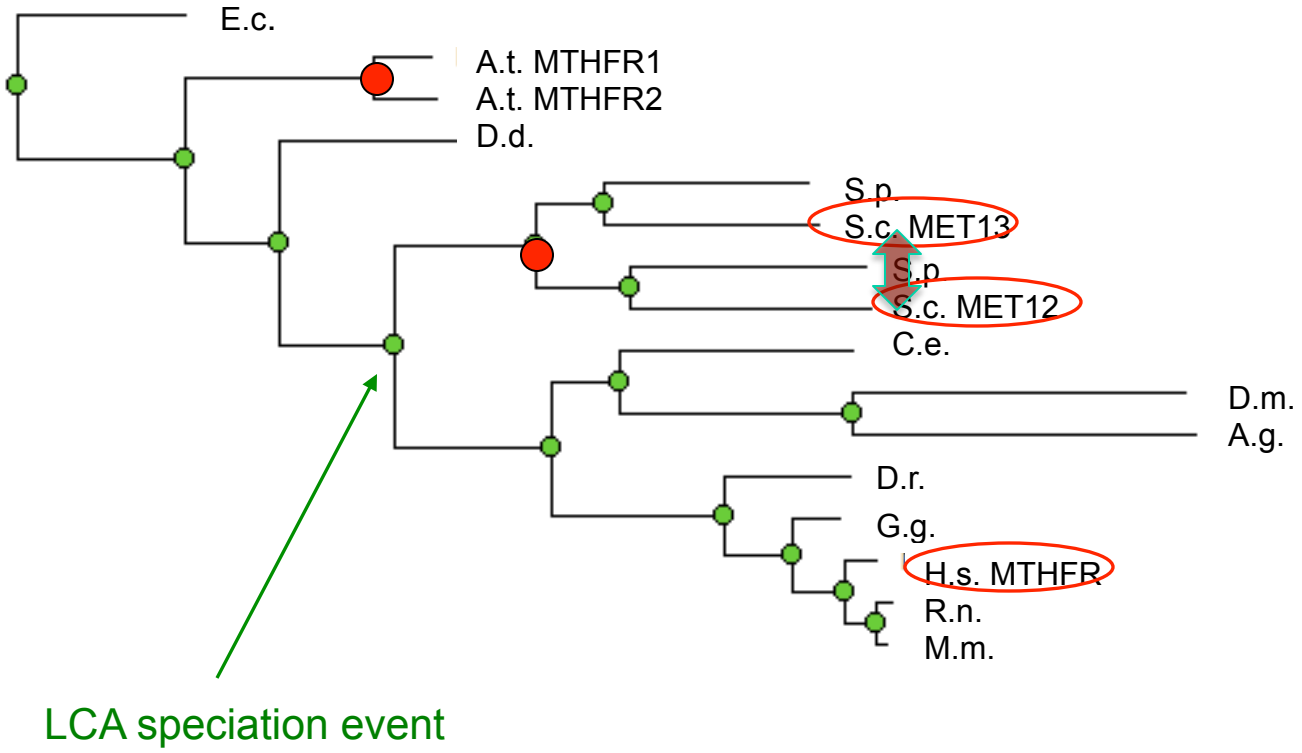


Orthology gets more complicated when there are duplications following speciation



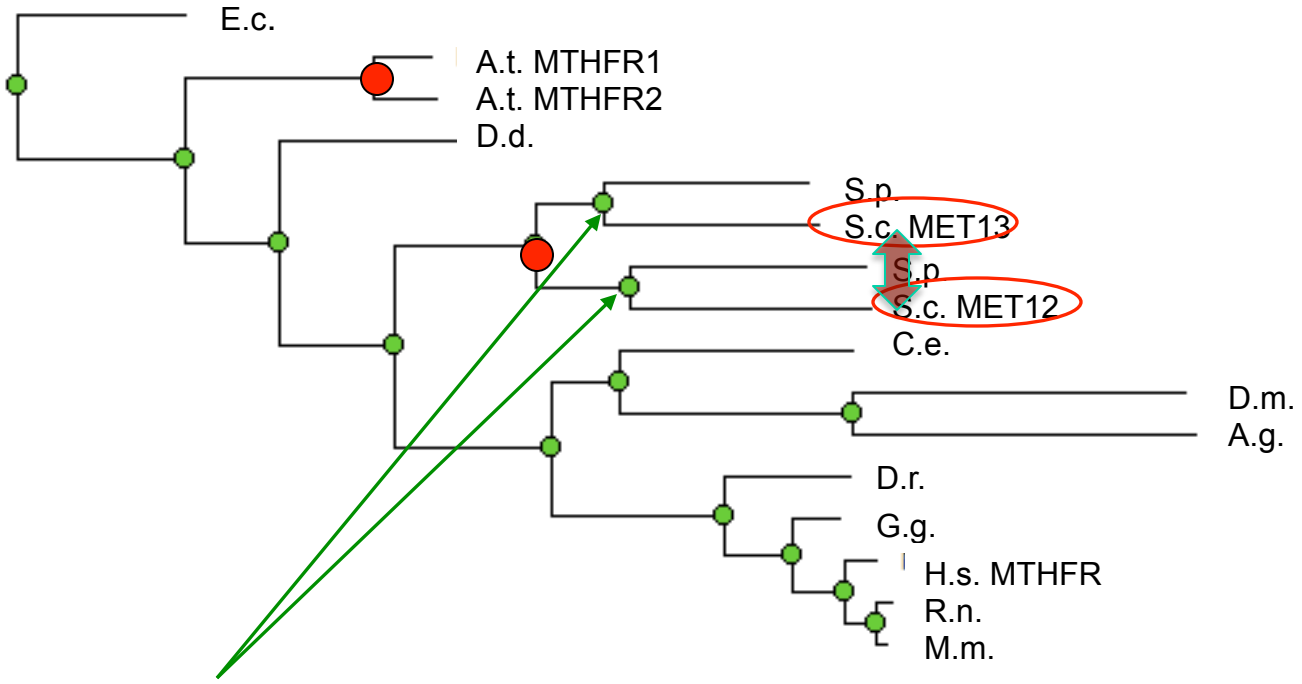
H.s. MTHFR has two orthologs in yeast
And these two orthologs are paralogs of each other

These genes are "in paralogs" with respect to each other AND the human ortholog



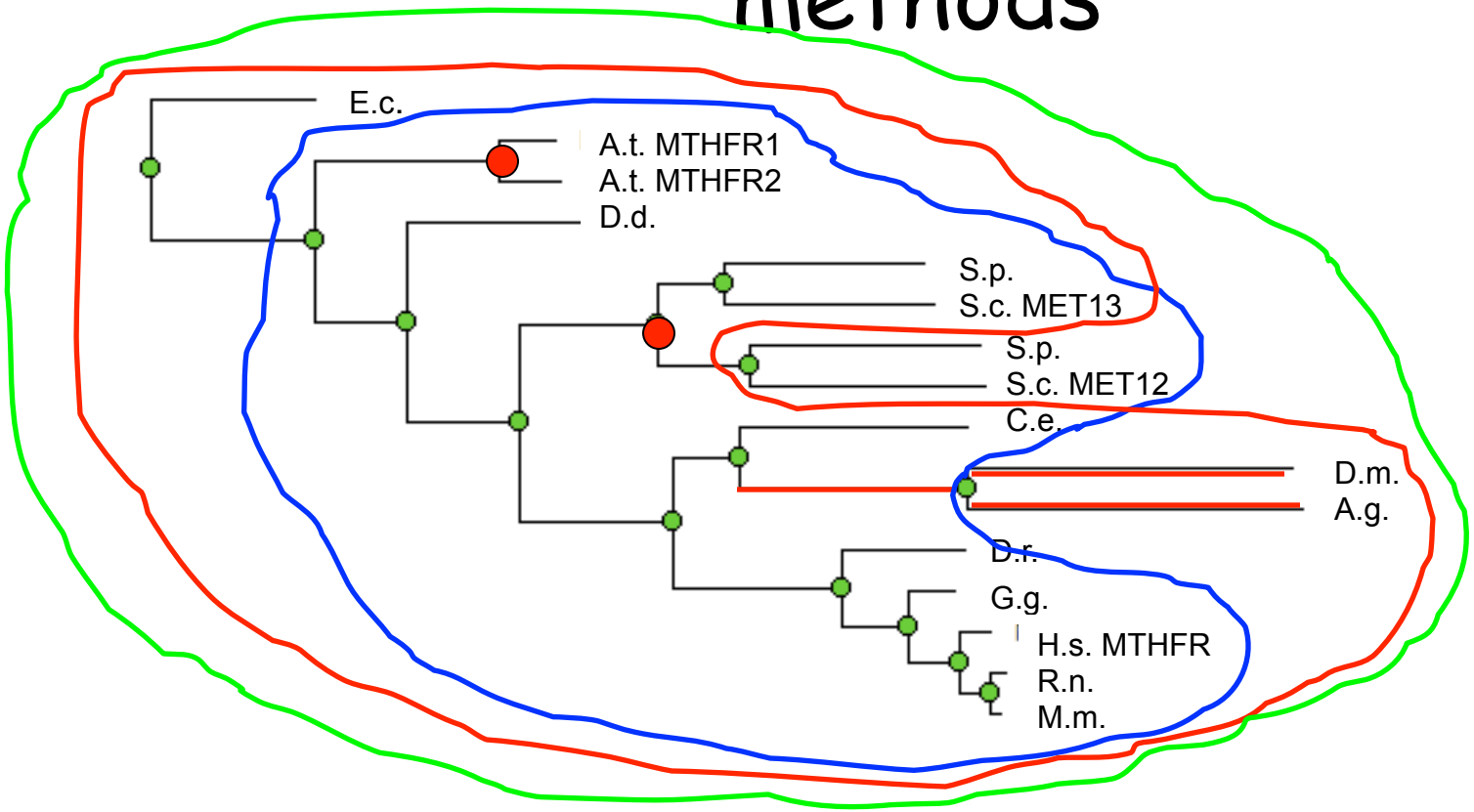
H.s. MTHFR has two orthologs in yeast
And these two orthologs are paralogs of each other

But these same genes are "out paralogs" with respect to each other



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

Clusters from different "orthology" methods

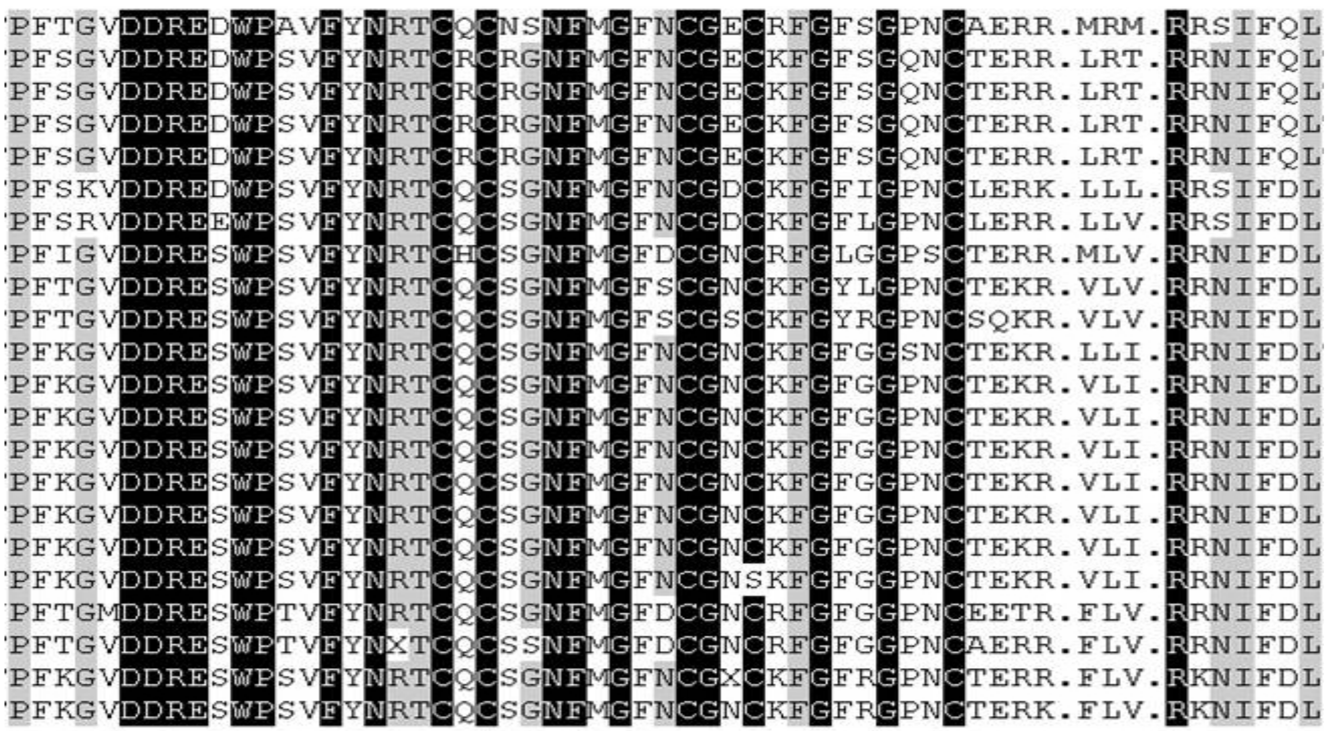


- OrthoMCL in red; PhiGs in blue; InParanoid in green
- An "ortholog cluster" is made by one or more "slices" through the protein family tree

IEA annotations: 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



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

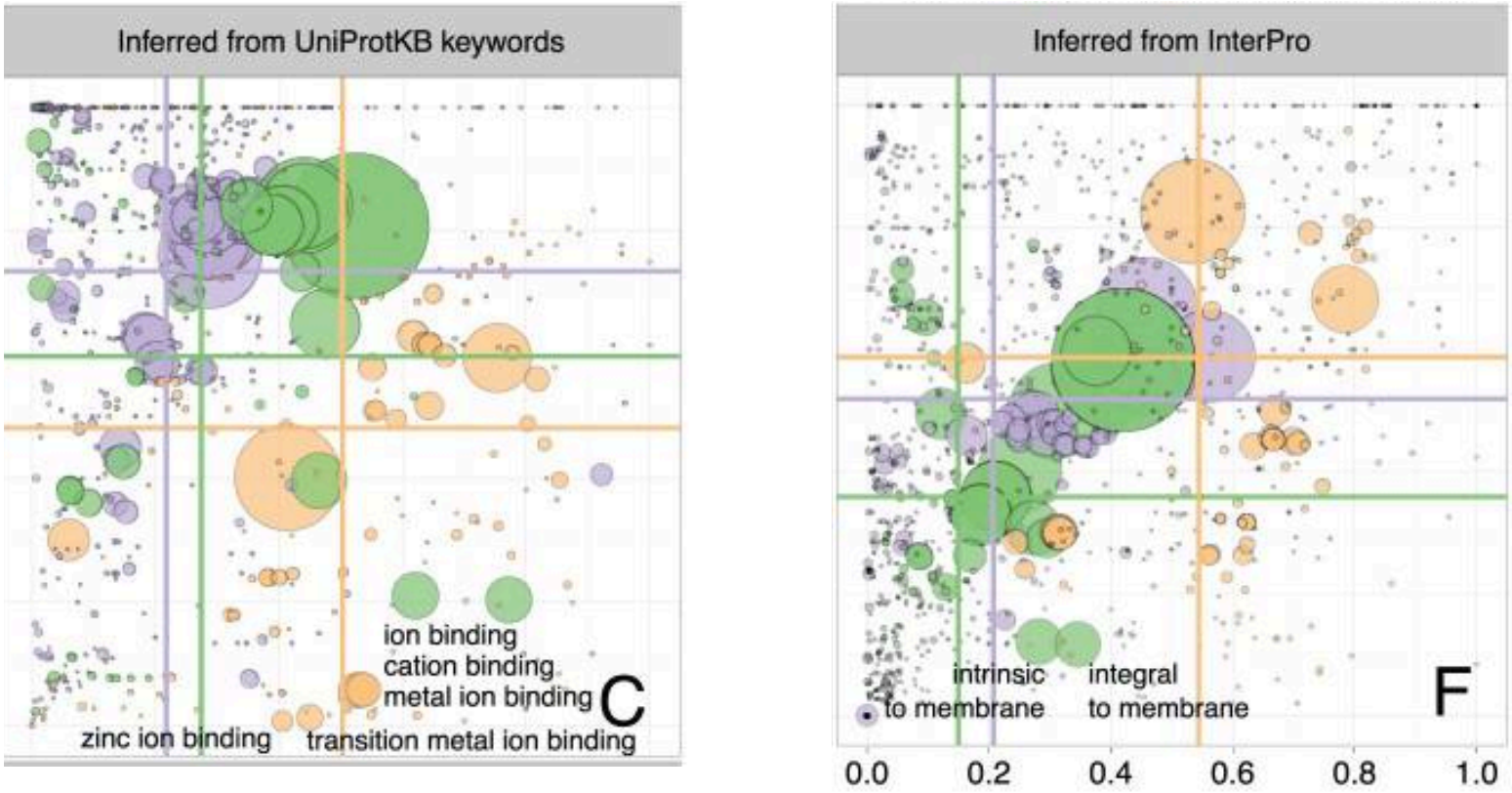
Database (Oxford). 2012 Feb 1;2012:bar068. Print 2012.

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.

IEA: keywords are more reliable than InterPro



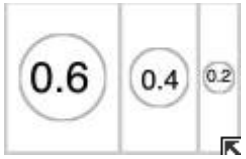
PLoS Comput Biol. 2012 May;8(5):e1002533. Epub 2012 May 31.

Quality of computationally inferred gene ontology annotations.

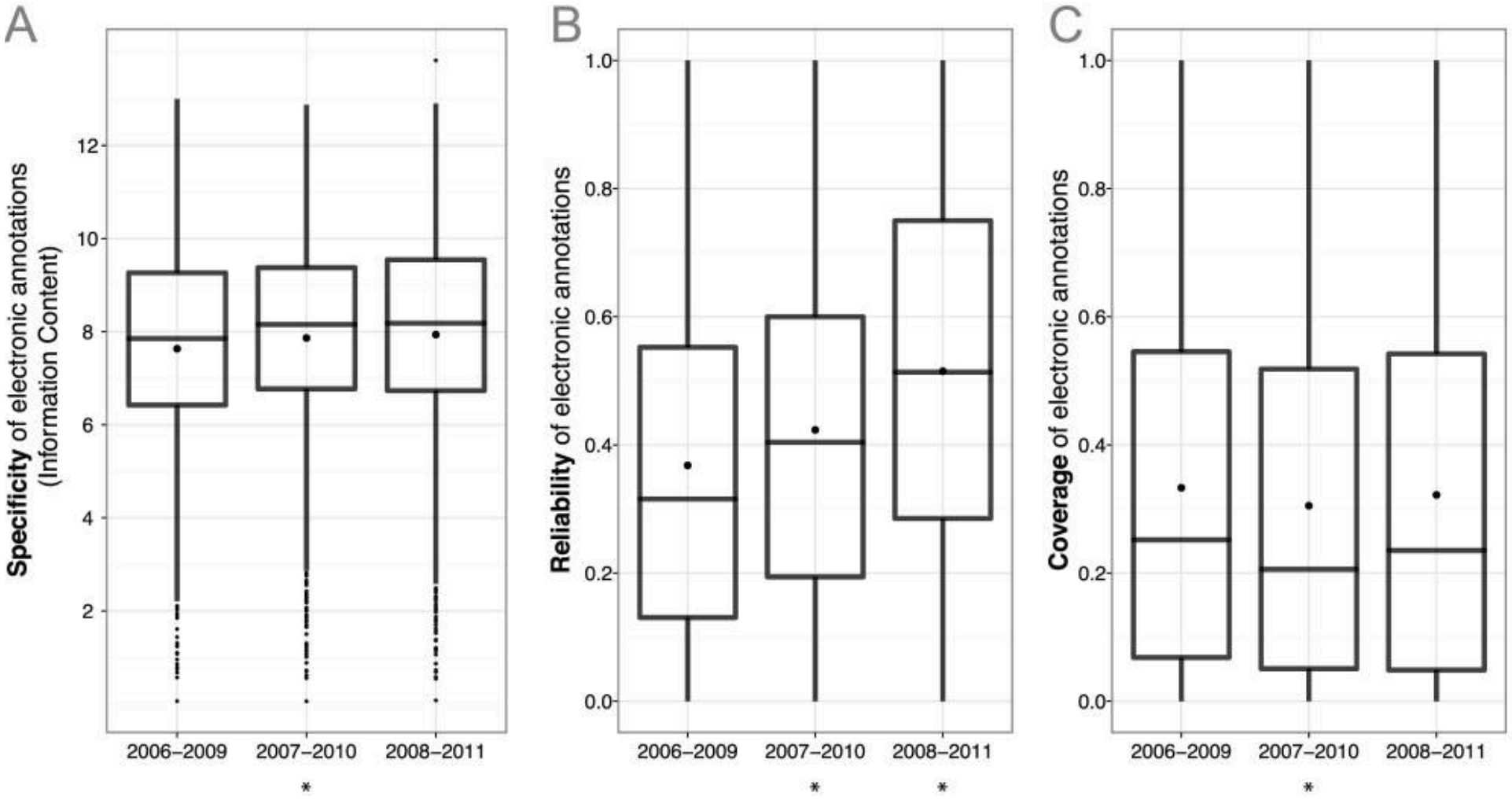
Skunca N, Altenhoff A, Dessimoz C.

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

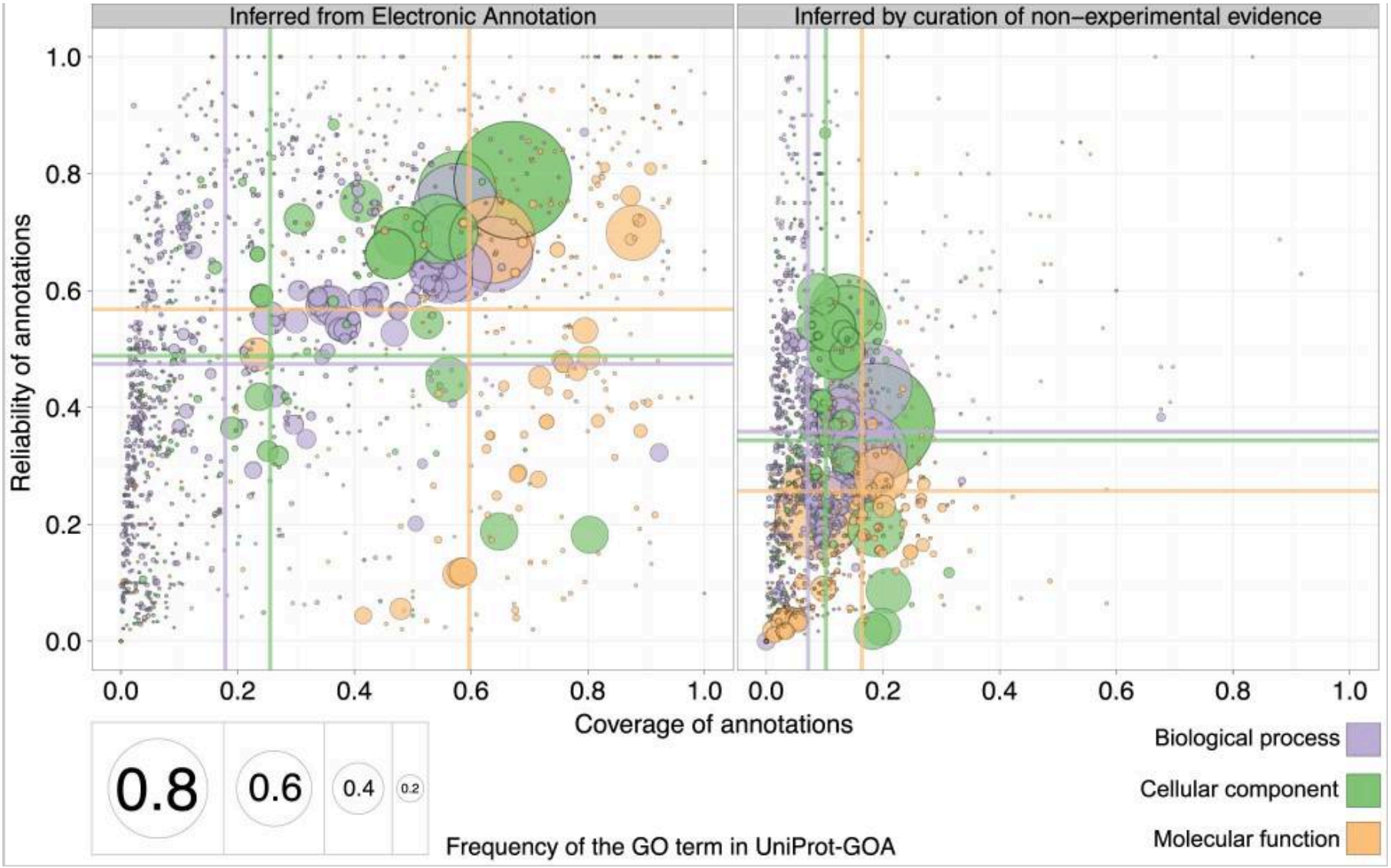
Biological process



IEAs have become more specific and more reliable



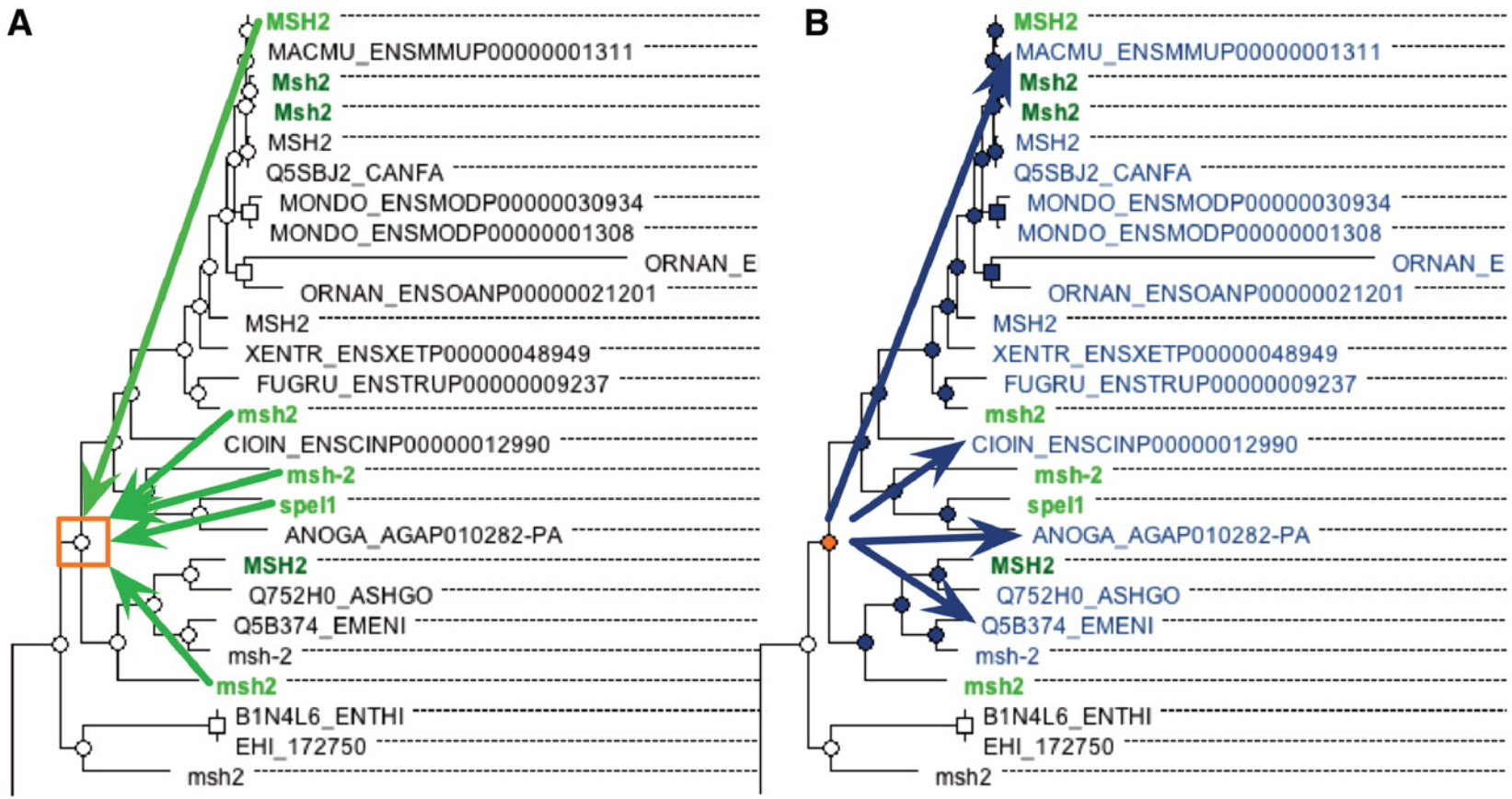
IEA is more reliable than ISS+IC



Profile-based annotation e.g. InterPro2GO

- 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

IBA: Use multiple pieces of evidence in a phylogenetic tree

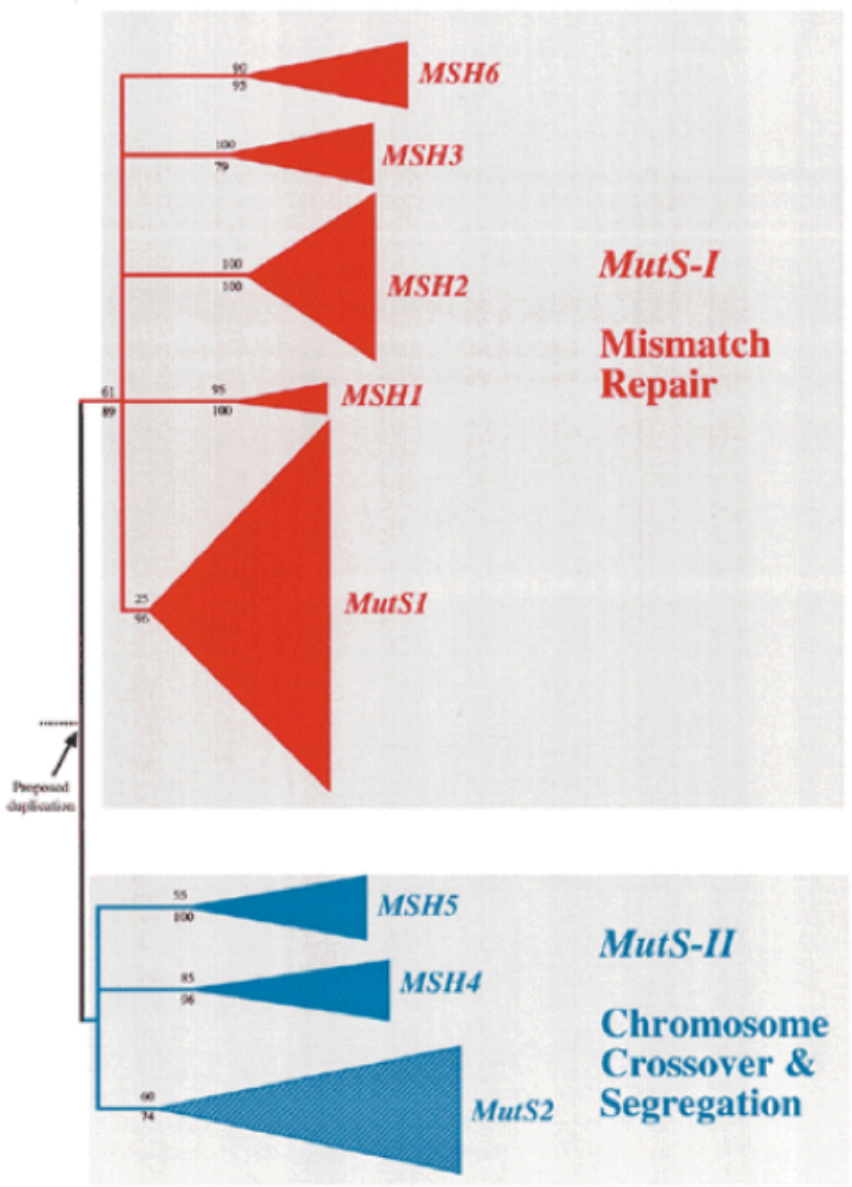


Integration of experimental GO annotations from different models (curated)

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

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

"Phylogenomic" function annotation

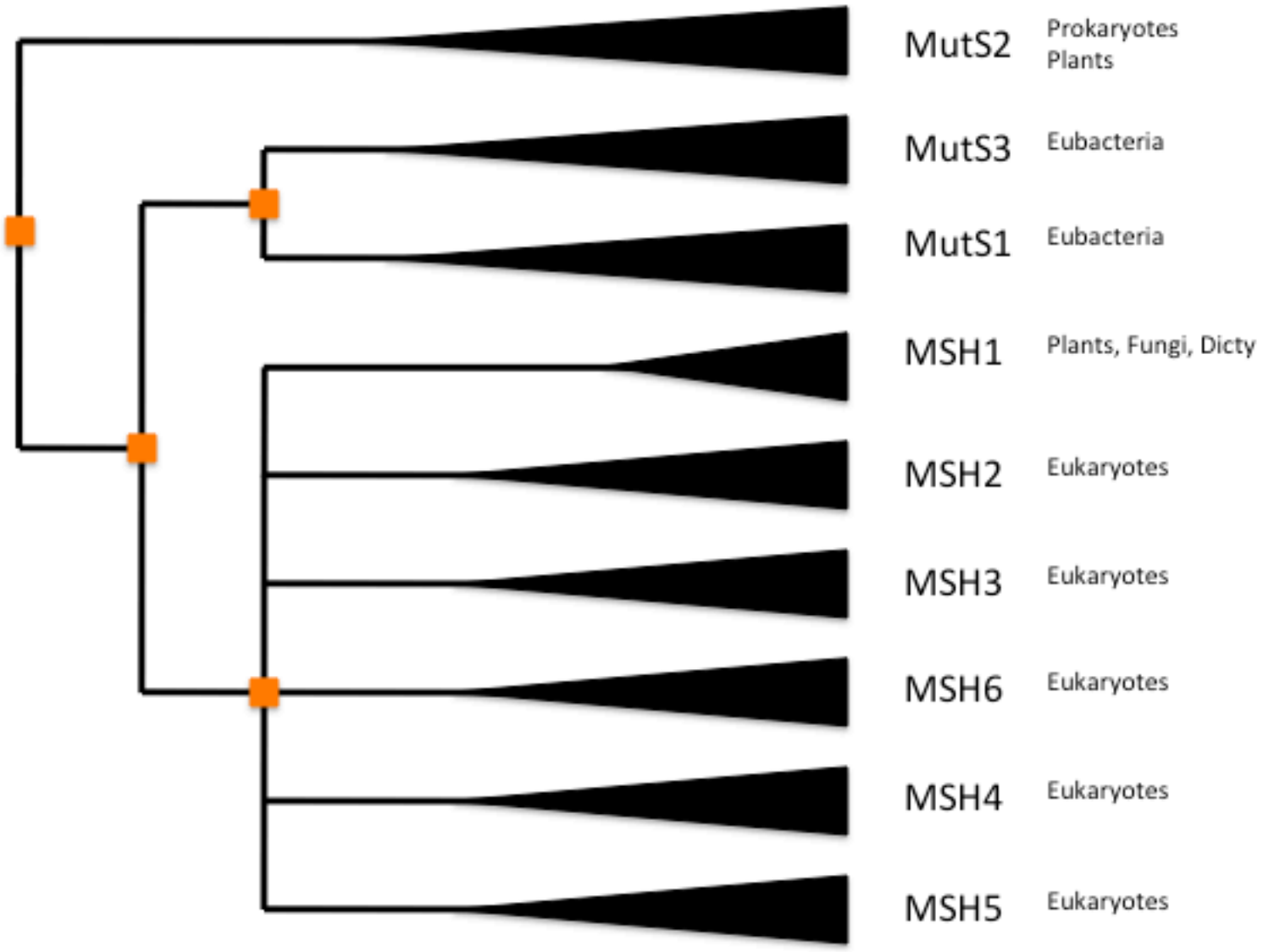


- View known data in the context of phylogenetic tree
- Infer subfamilies that share function

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

A phylogenomic study of the MutS family of proteins.

Eisen JA.

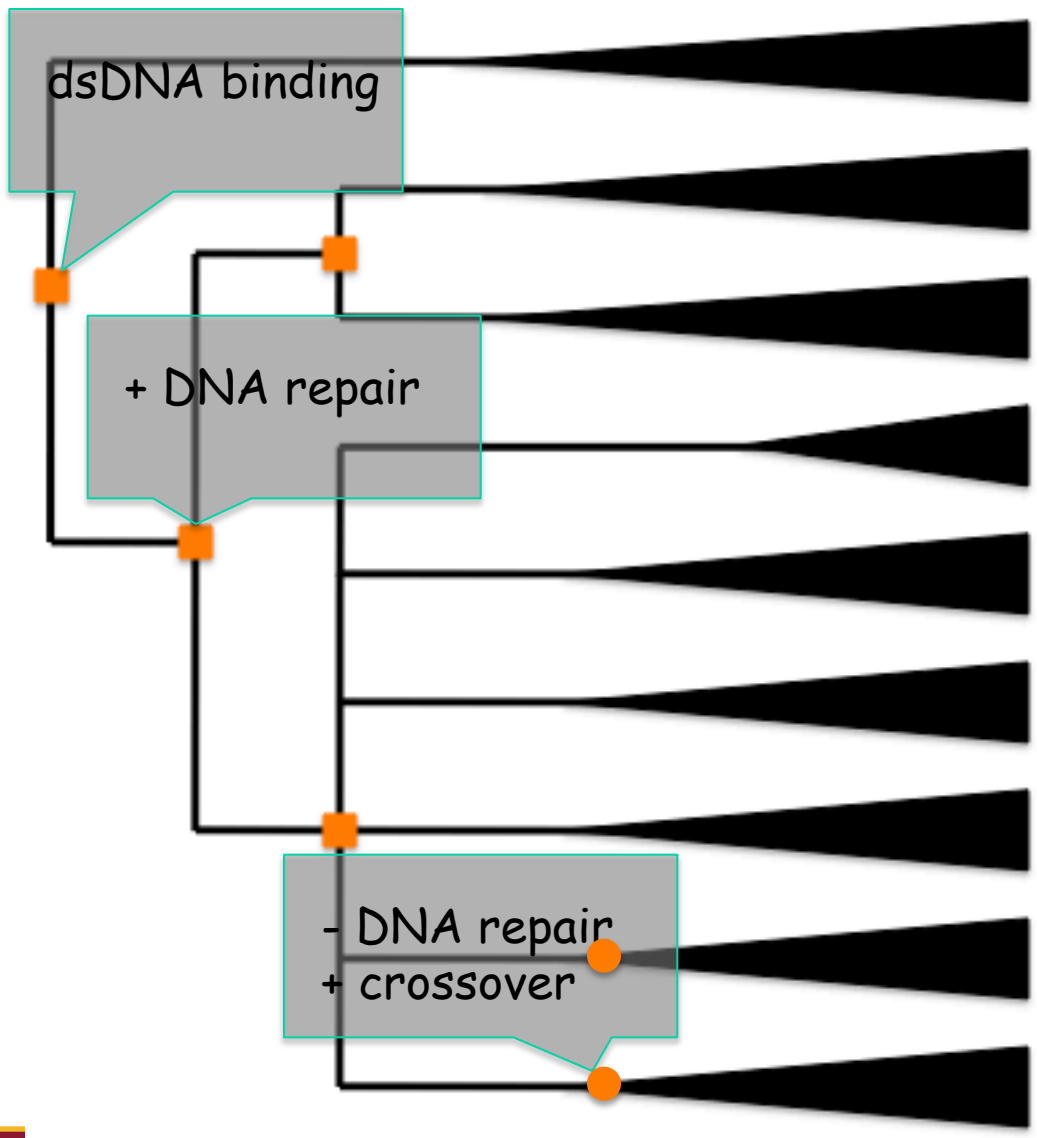


Keck School of Medicine of USC

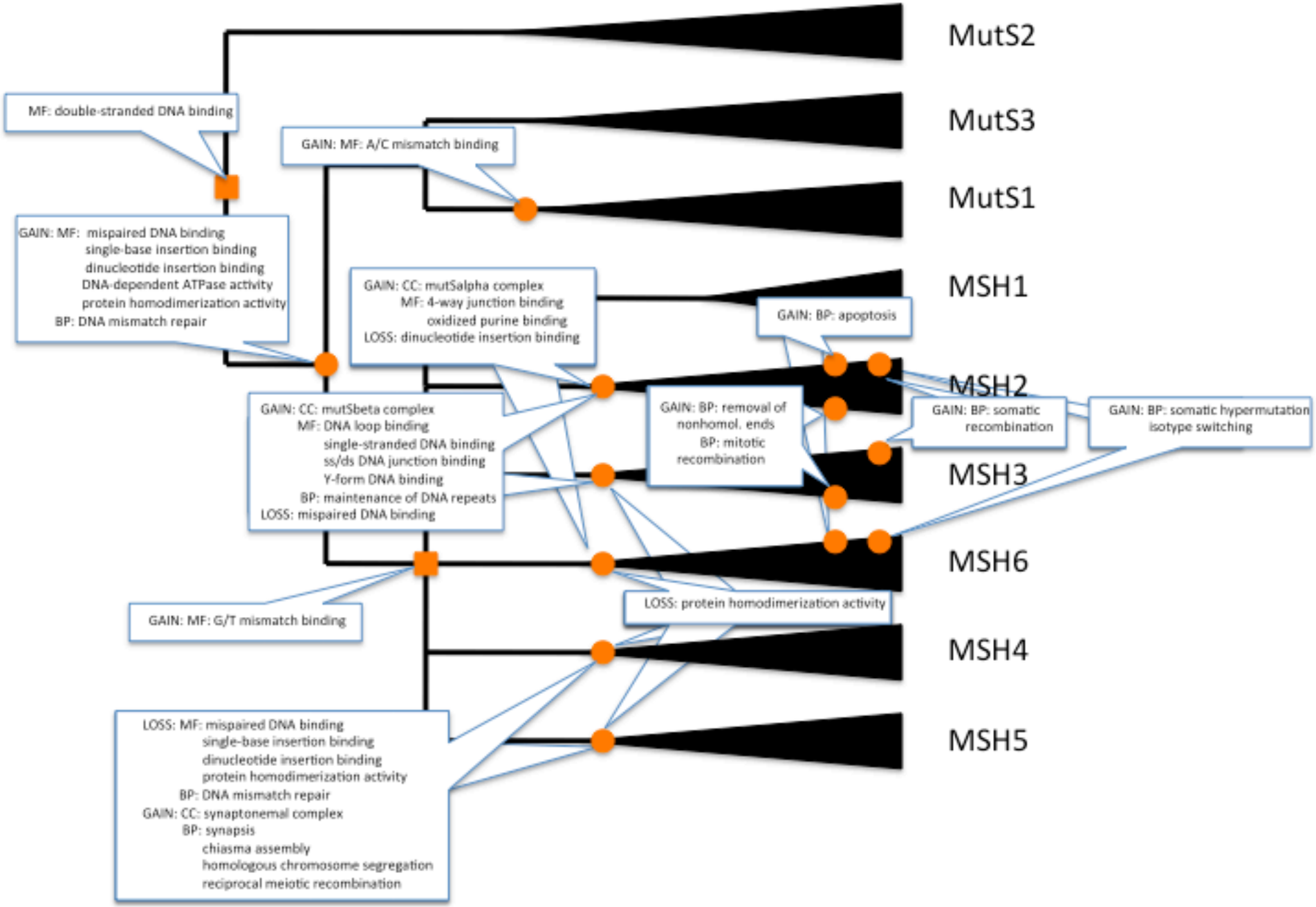
Nucleic Acids Res. 2007;35(22):7591-603. Epub 2007 Oct 26.

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

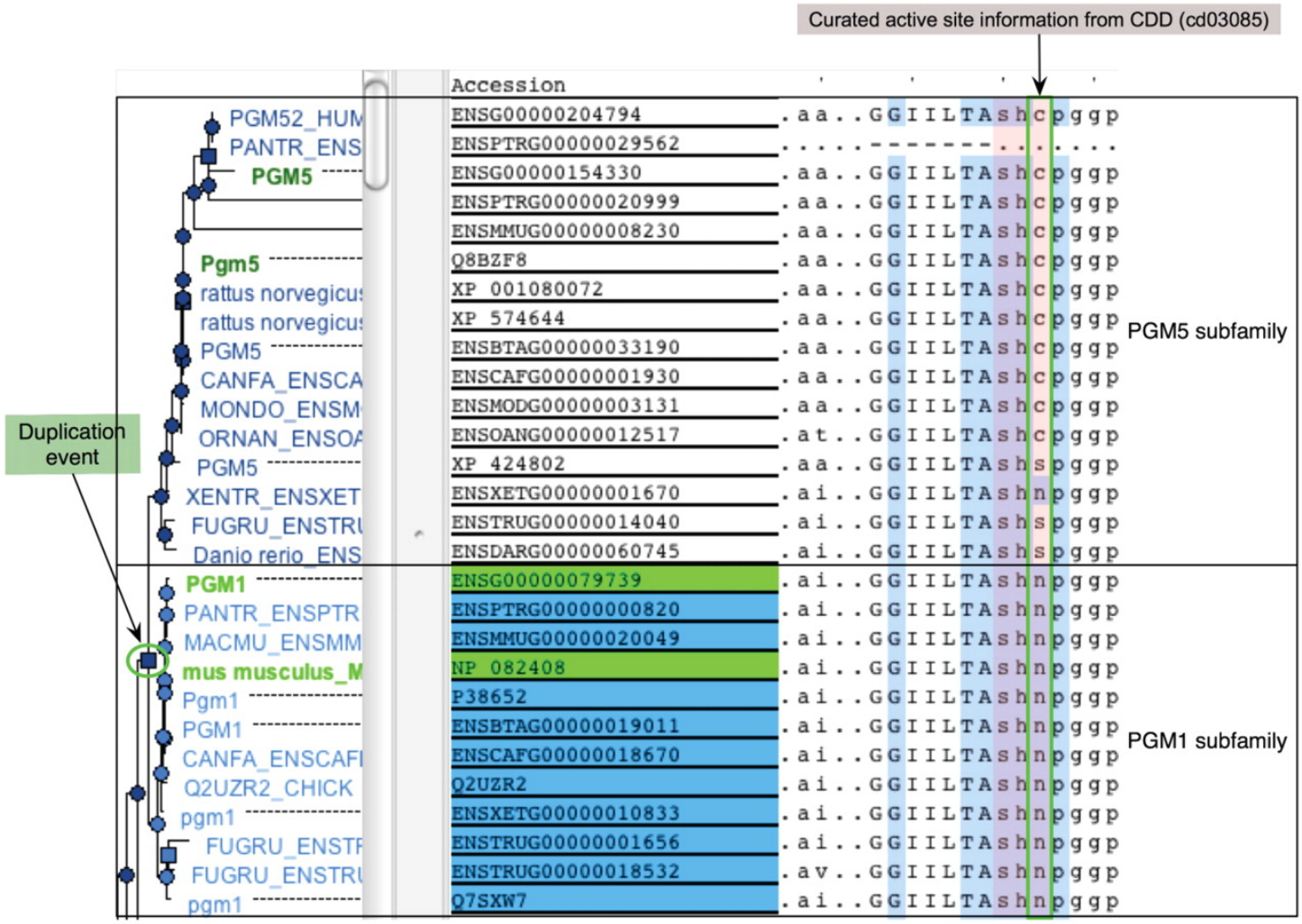
Lin Z, Nei M, Ma H.



		Inherited annotations
MutS2	Prokaryotes Plants	dsDNA binding
MutS3	Eubacteria	dsDNA binding DNA repair
MutS1	Eubacteria	dsDNA binding DNA repair
MSH1	Plants, Fungi, Dicty	dsDNA binding DNA repair
MSH2	Eukaryotes	dsDNA binding DNA repair
MSH3	Eukaryotes	dsDNA binding DNA repair
MSH6	Eukaryotes	dsDNA binding DNA repair
MSH4	Eukaryotes	dsDNA binding crossover
MSH5	Eukaryotes	dsDNA binding crossover


MutS2
MutS3
MutS1
MSH1
MSH2
MSH3
MSH6
MSH4
MSH5
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: Loss of function can be annotated
Here, evidence is substitution at active site

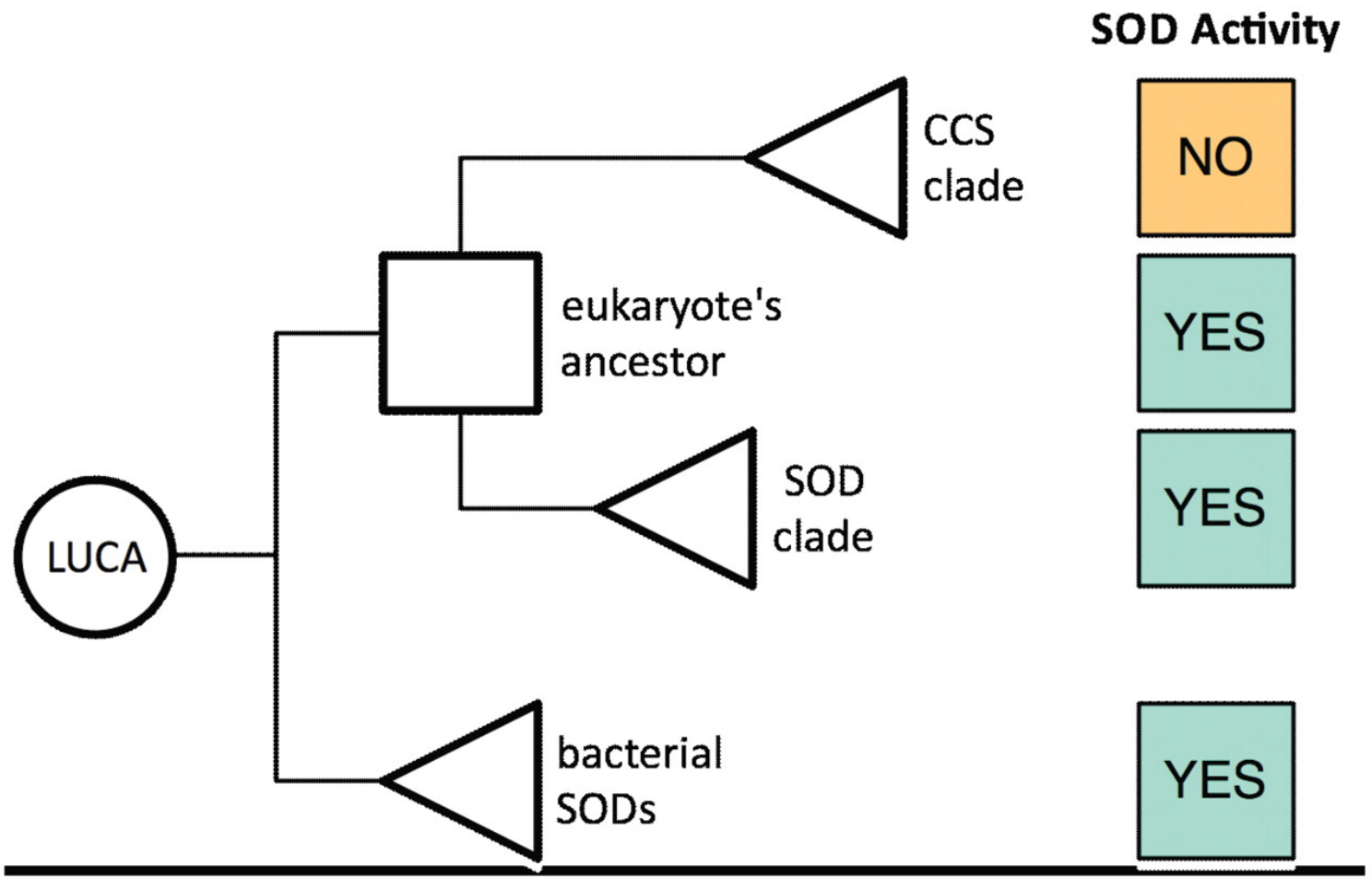


Gaudet P et al. Brief Bioinform 2011;12:449-462

IEA vs. IBA
PGM family

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

IBA: Loss of function can be annotated
Here, evidence is overall divergence



IEA vs. IBA
SOD family

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

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

Where to get the data

- GO annotations
  - GO 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

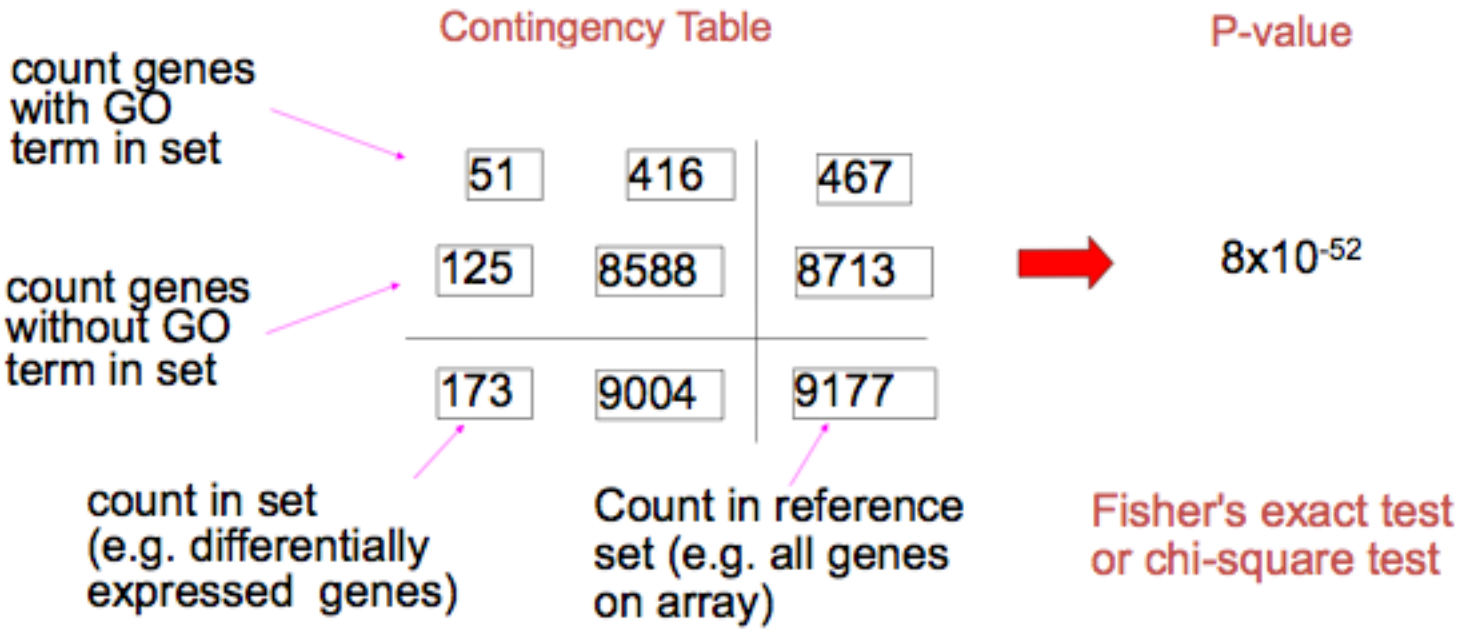
Enrichment analysis

- Used to find a biological interpretation of a large-scale "genomics" experiment
- Uses known information about gene function to see if there 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

Two main types of test

- "Overrepresentation"
  - Given a list of genes
  - Are some functional classes over (or under) represented in the list compared to random expectation?
- "Enrichment"
  - Given a list of genes and a quantitative value for each gene (e.g. fold change)
  - Does the distribution of values for genes in a functional class differ significantly from the expected distribution for random genes?

Over (under) representation test example



Enrichment tests

- Different statistical tests
  - For overrepresentation, have to choose threshold for defining list
- 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

Annotation sets

- Appropriate sets depend on biological question
  - most "omics" data analysis looks for correlated changes across groups of genes that may function together: pathways and GO biological processes
- Different sources
  - Gene Ontology
    - Molecular function, cellular component, biological process
  - Pathways
    - PANTHER, KEGG, Pathway Commons
- Different versions of same set
  - Always use most recent sets: most complete and accurate