The Gene Ontology's Reference Genome Project:
A Unified Framework for Functional Annotation across Species

The Reference Genome Group of the Gene Ontology Consortium

The Gene Ontology

An initiative to unify the representation of gene product attributes across all species

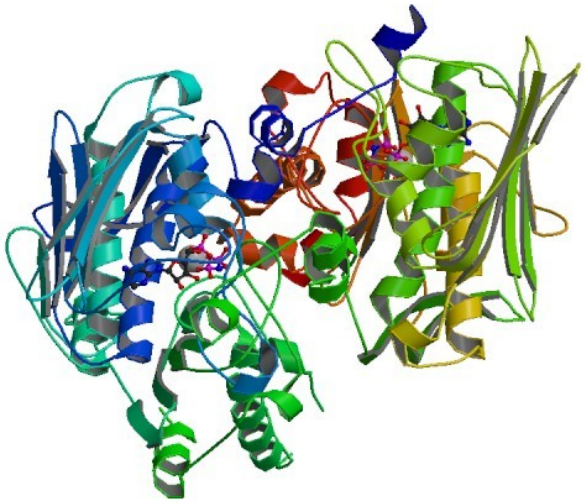
Three domains:

Cellular Component: the parts of a cell or its extracellular environment;

Molecular Function: the elemental activities of a gene product at the molecular level, such as binding or catalysis;

Biological Process: operations or sets of molecular events with a defined beginning and end, pertinent to the functioning of integrated living units

Gene Product annotation with GO terms

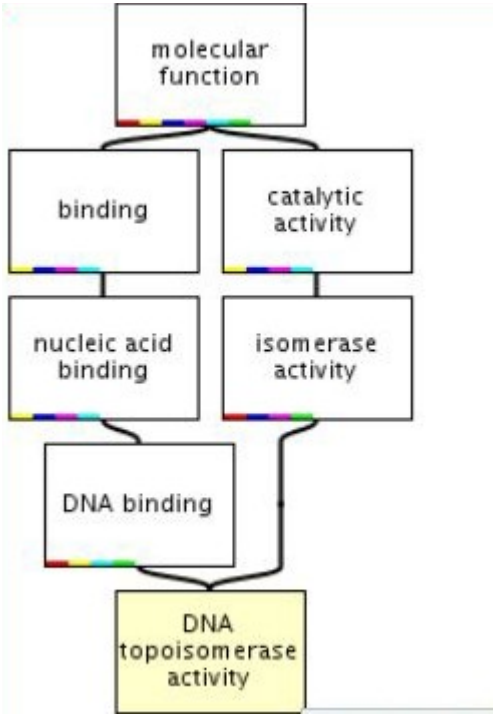


Human DNA topoisomerase IIA (P11388)

Cellular Component
- nucleus
- chromosome
- DNA topoisomerase complex

Molecular Function
- chromatin binding
- DNA topoisomerase activity
- DNA-dependent ATPase activity

Biological Process
- DNA replication
- DNA topological change
- DNA ligation
- DNA repair



GO:0003916 DNA topoisomerase activity

Definition

Catalysis of the transient cleavage and passage of individual DNA strands or double helices through one another, resulting a topological transformation in double-stranded DNA.

Children

I GO:0003917 DNA topoisomerase type I activity
I GO:0003918 DNA topoisomerase (ATP-hydrolyzing) activity

Organisms annotated

- *Arabidopsis thaliana*
- *Caenorhabditis elegans*
- *Danio rerio*
- *Dictyostelium discoideum*
- *Drosophila melanogaster*
- *Escherichia coli*
- *Gallus gallus*
- *Homo sapiens*
- *Mus musculus*
- *Rattus norvegicus*
- *Saccharomyces cerevisiae*
- *Schizosaccharomyces pombe*

Selection criteria

- The different organisms represent a wide range of the phylogenetic spectrum
- They are the basis of a significant body of scientific literature
- Supported by a reasonably sized community of researchers
- Provides an experimental system for the study of human disease, or for economically important activities such as agriculture
- Importantly, all of these organisms are supported by an established database that includes GO curators

Reference genome annotation process

1. Protein family is chosen for annotation and each Model Organism Database annotate gene products based on available experimental data

2. Protein family curators annotate ancestors proteins

3. MODs add inferred annotations to the gene products from the organism they annotate

Curation Priorities

Genes whose products are highly conserved during evolution,
e.g. the gyrase/topoisomerase II gene family

Genes whose human ortholog is known to be implicated in disease, e.g. the MutS homolog gene family (hereditary colorectal cancer)

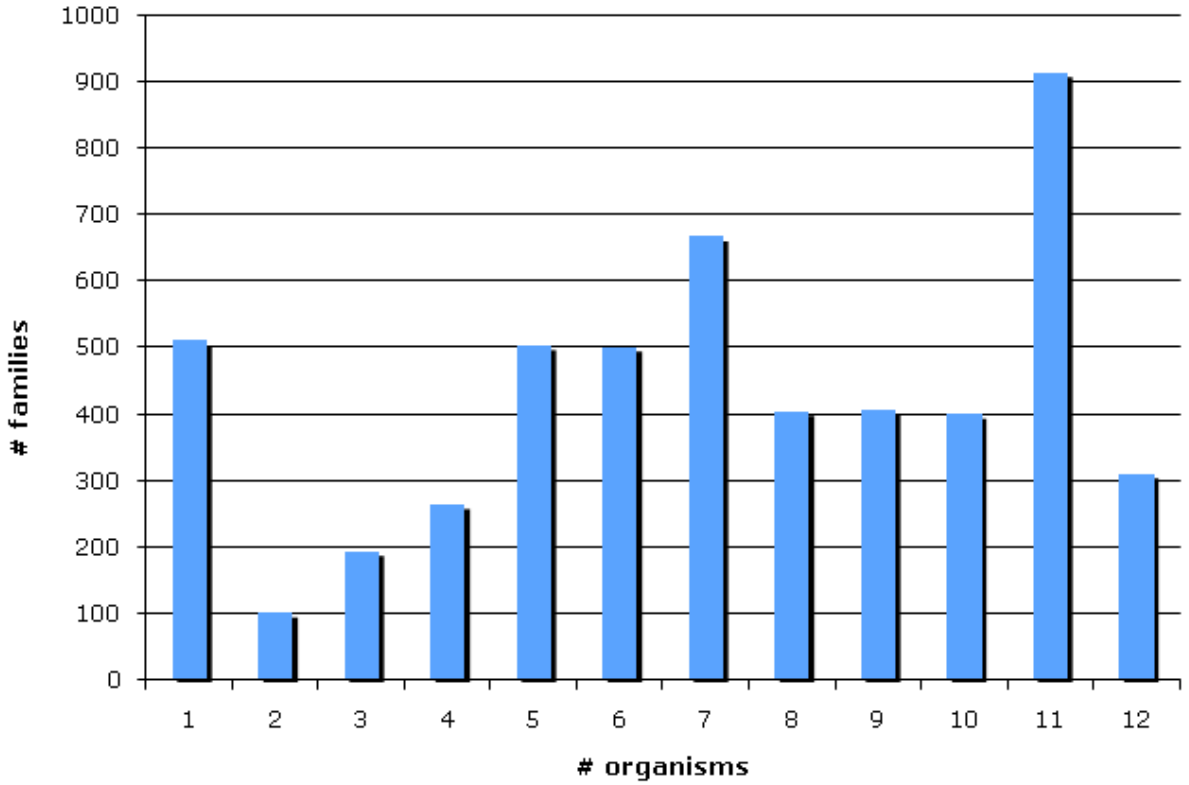
Genes whose products are involved in known biochemical and signaling pathways,
e.g. the PYGB phosphorylase that participates in glycogen degradation

Genes identified from recently published literature as having an important or new scientific impact,
e.g. POU5F1 (homeobox gene) that is important for stem cell function

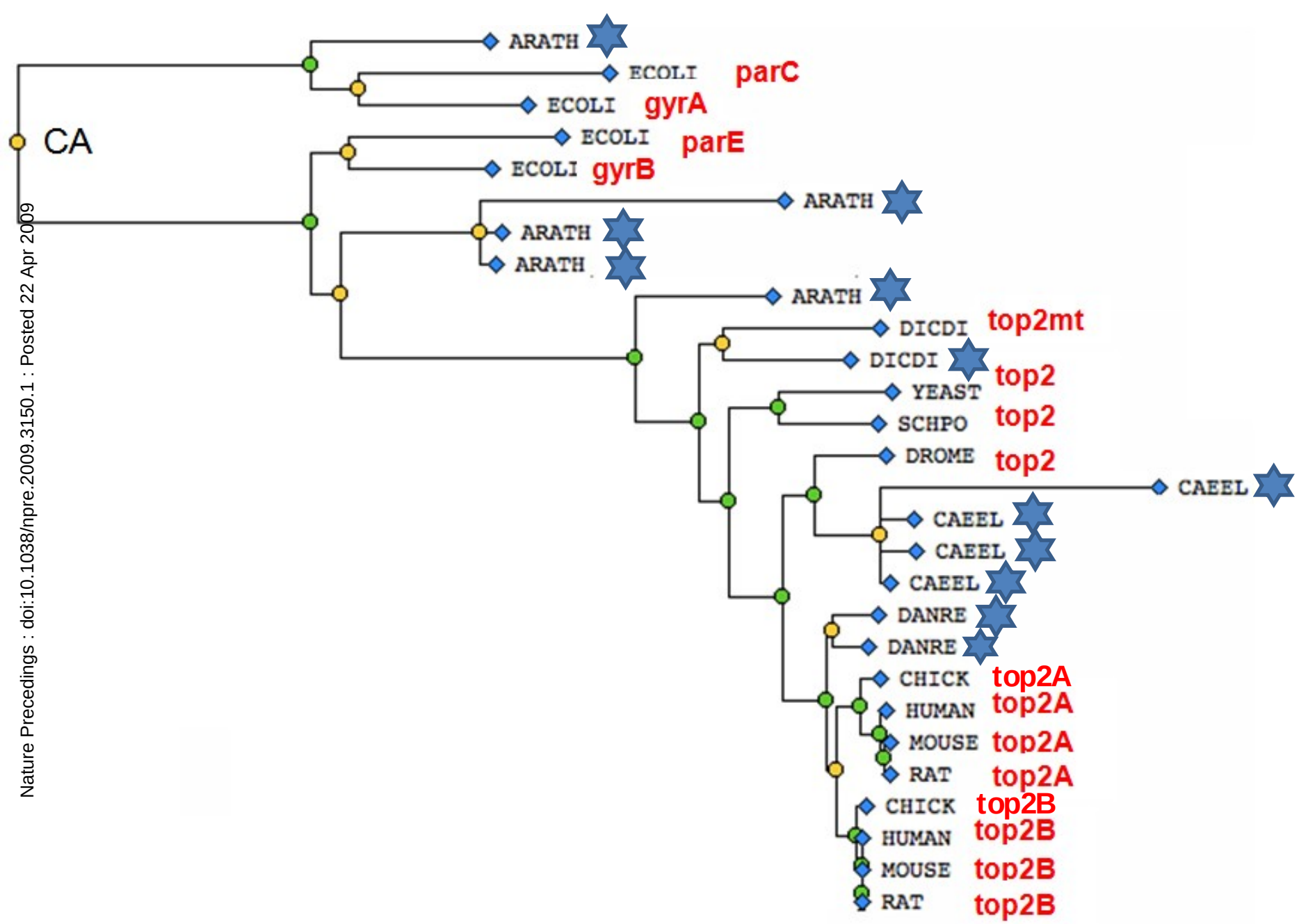
Establishing protein families

- PANTHER: Protein ANalysis THrough Evolutionary Relationships Classification System
- PANTHER version 7.beta.1:
  - 5198 families generated with sequences from 46 species

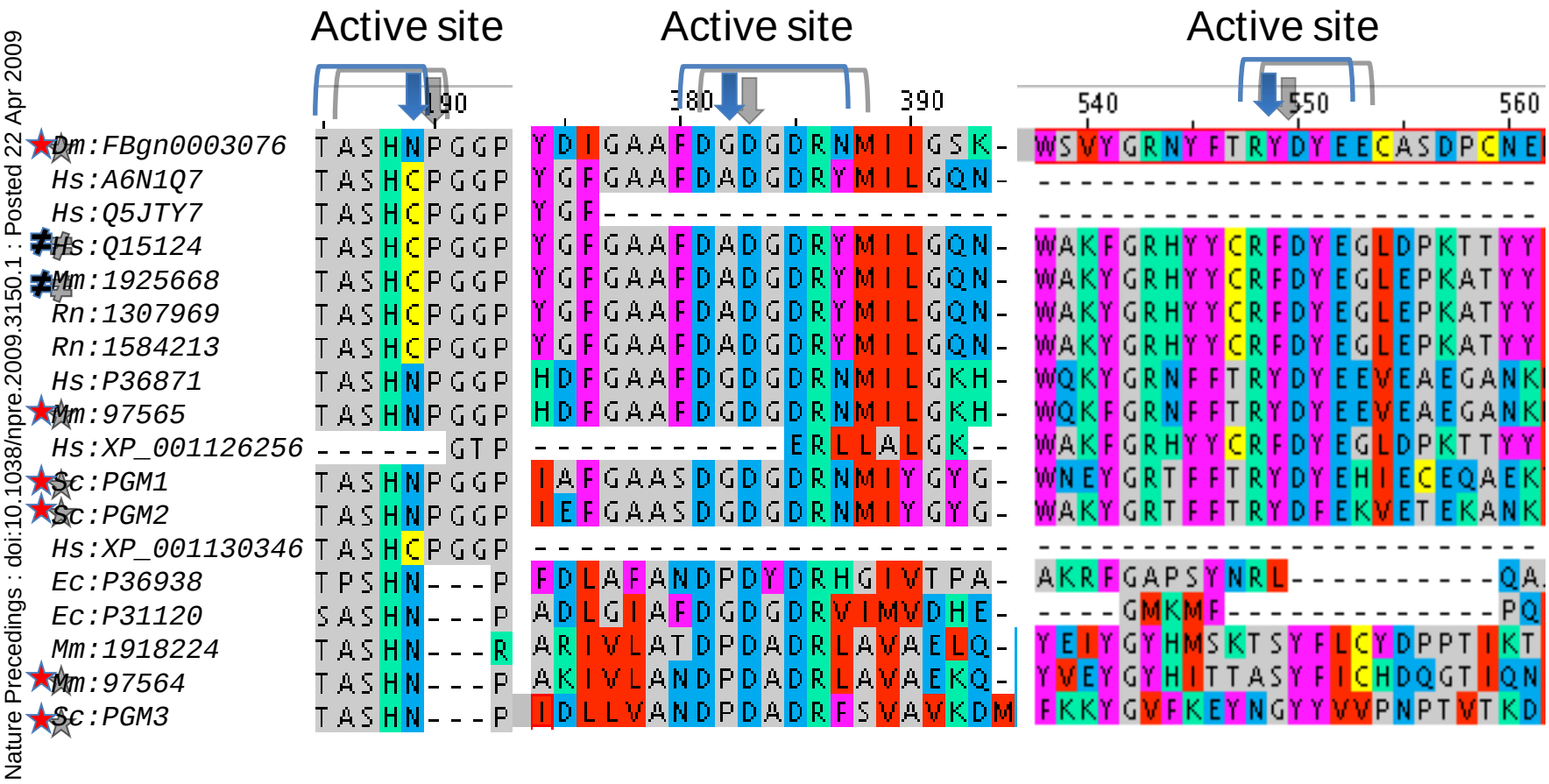
Distribution of PANTHER families across organisms



312 families represented in all 12 reference genomes
916 families are presents in all represented eukaryotes
4388 have members from at least four reference genomes



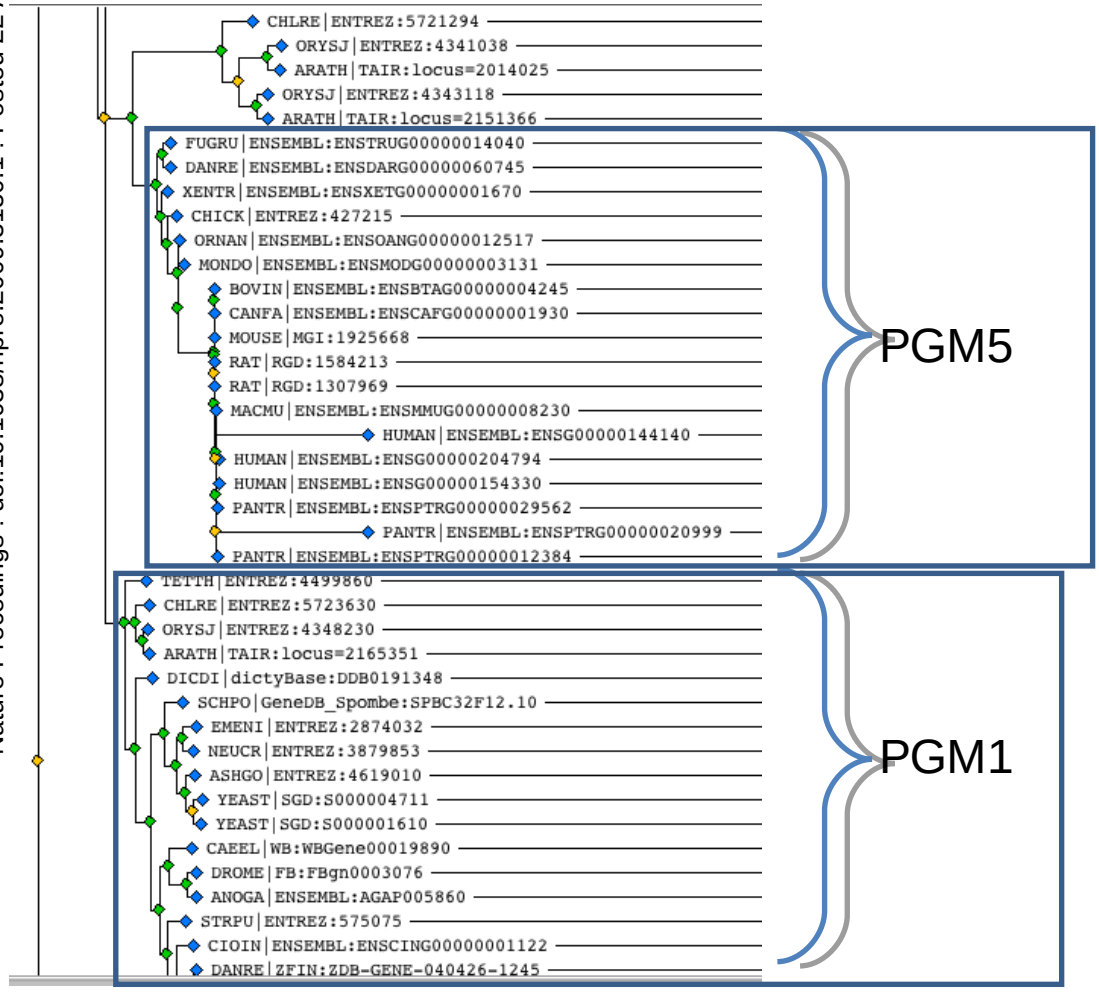
Phosphohexomutase family



phosphoglucomutase activity

Phosphohexomutase family

PTHR22573



NOT
phosphoglucomutase activity

phosphoglucomutase activity

Benefits

Increased depth of annotations

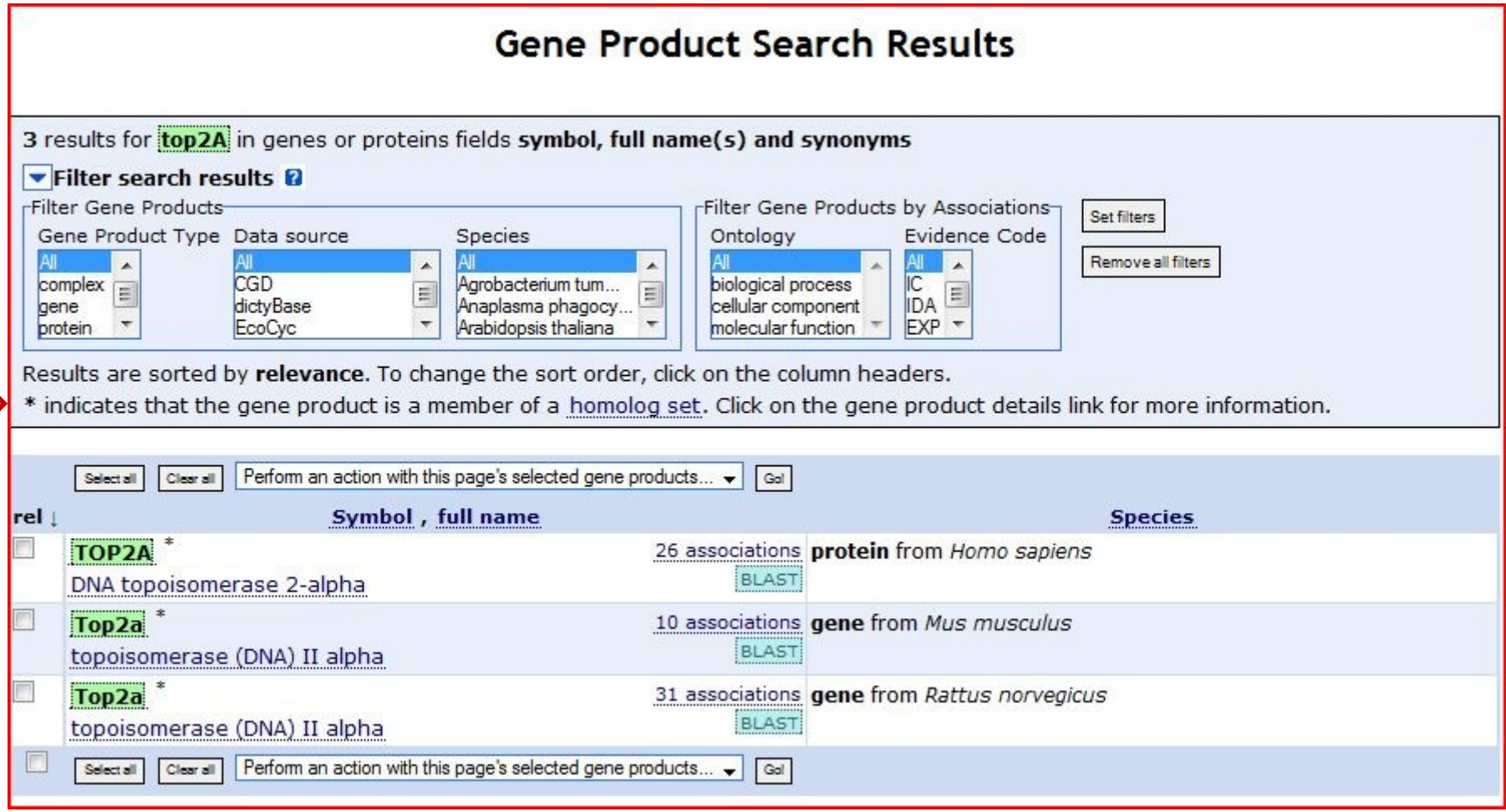
- about 2-fold increase in Information Content (IC)
- graph coverage increased by ~50%

Increased breadth of annotations

Improvement to the GO

- improvements of terms
- improvements of definitions
- addition of synonyms

Visualization: AmiGO



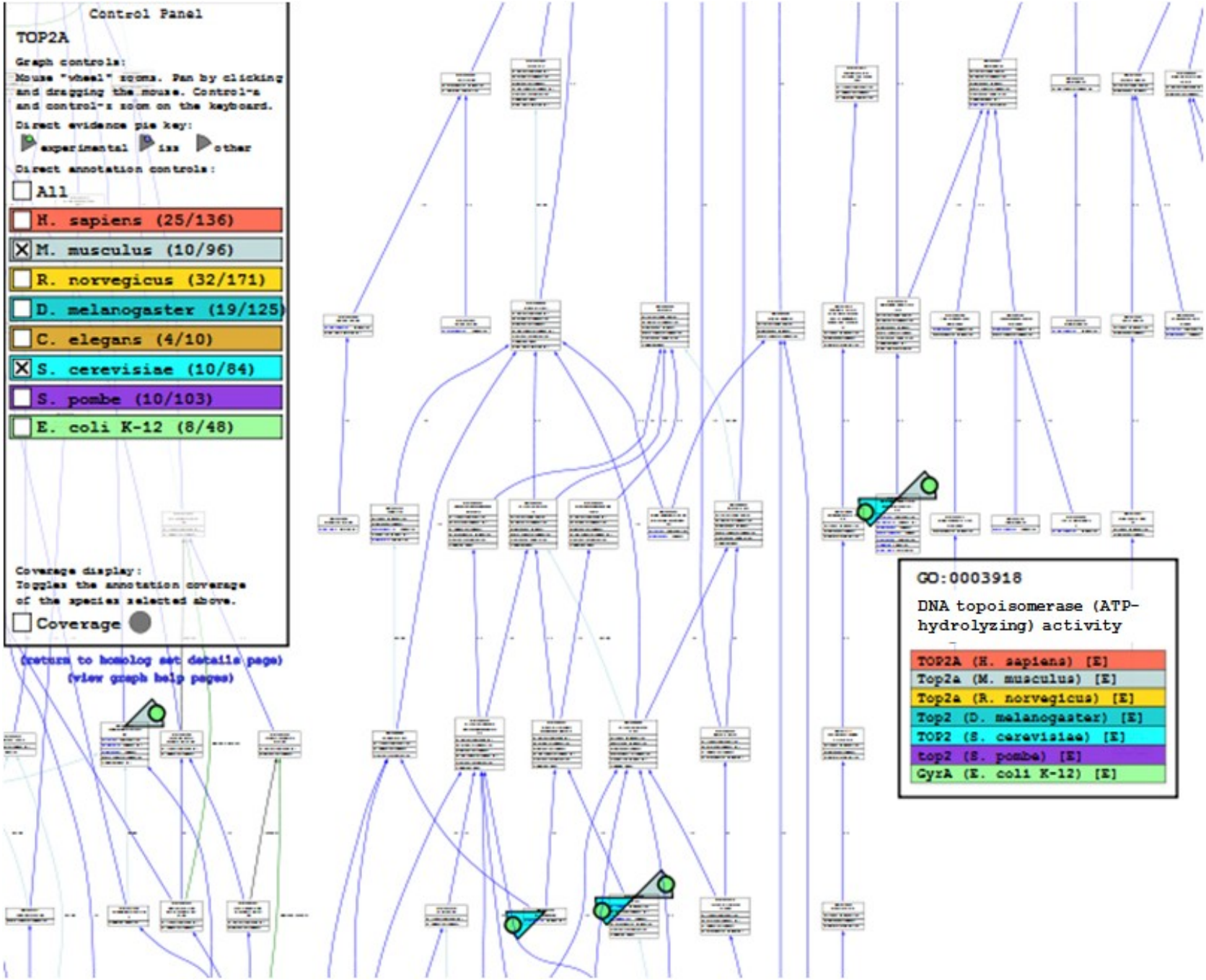
http://goweb-dev.stanford.edu/cgi-bin/amigo/search.cgi

TOP2A

Gene product information ⬇ Peptide sequence ⬇ Sequence information ⬇ 26 term associations ➡

Information

Symbol	TOP2A
Name(s)	DNA topoisomerase 2-alpha
Type	**protein**
Species	*Homo sapiens* *(human)*
Synonyms	IPI00218753 IPI00414101 IPI00478232 IPI00879004 TOP2 TOP2A TOP2A_HUMAN
Database	UniProtKB, UniProtKB:P11388
Sequence	View sequence; use as BLAST query sequence
Ref Genome	Homology under TOP2A (M. musculus S. cerevisiae S. pombe D. discoideum C. elegans H. sapiens R. norvegicus E. coli D. melanogaster)



Credits

Nature Precedings : doi:10.1038/npre.2009.3150.1 : Posted 22 Apr 2009

Annotators

Pascale Gaudet , Petra Fey, Rex Chisholm (dictyBase)

Tanya Berardini, Donghui Li (TAIR)

Emily Dimmer, Rachael Huntley (GOA), Ruth C. Lovering, Varsha K. Khodiyar (UCL)

Stacia R. Engel (SGD)

David P. Hill, Li Ni (MGI)

Doug Howe (ZFIN)

Jim Hu, Deborah A. Siegel (E.coliWiki)

Kimberly Van Auken , Ranjana Kishore (WormBase)

Fiona McCarthy (AgBase)

Victoria Petri (RGD)

Susan Tweedie (FlyBase)

Valerie Wood (GeneDB)

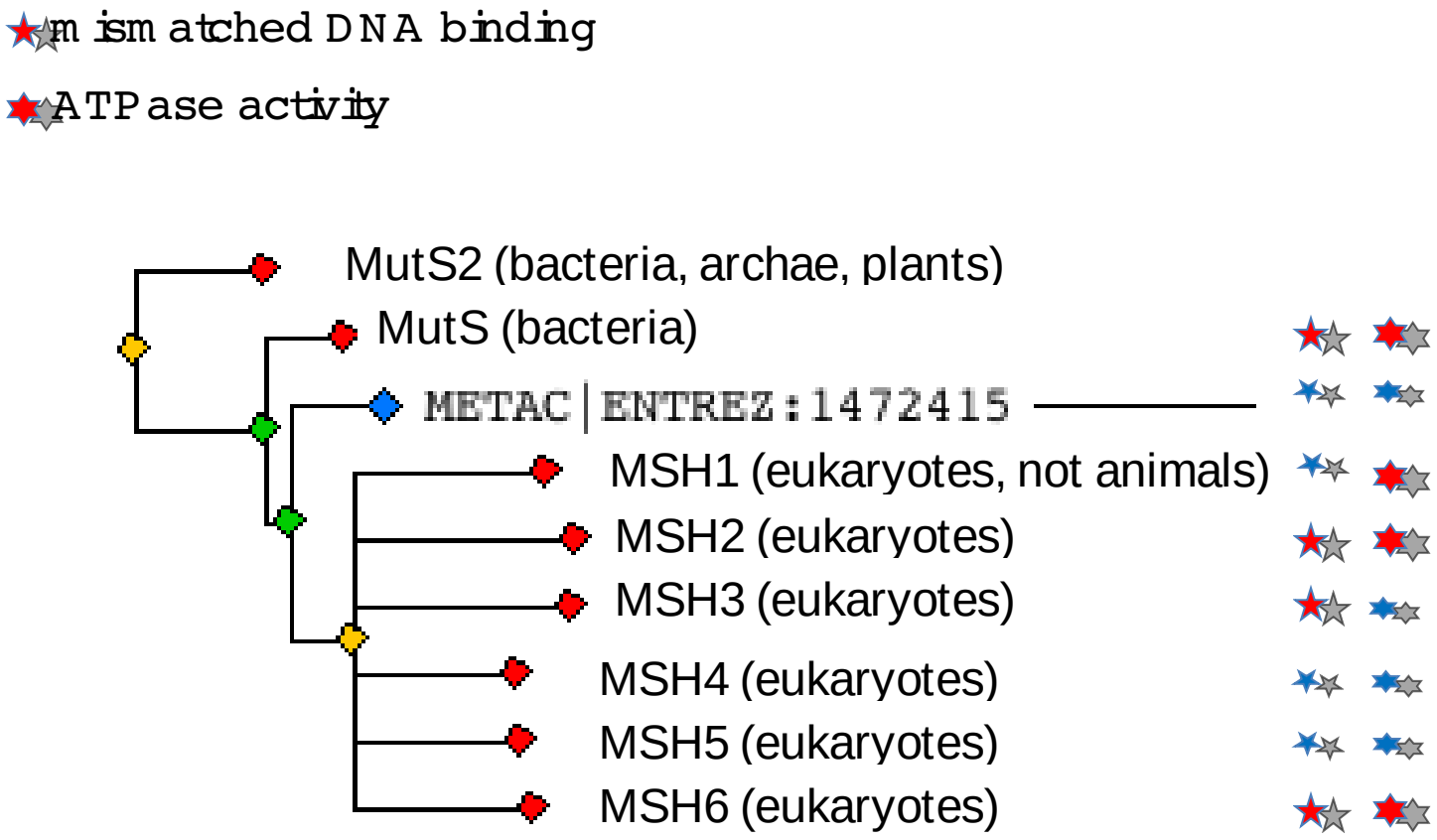
Computational staff representatives:

Siddhartha Basu (dictyBase), Seth Carbon (BBOP), Mary Dolan (MGI), and Chris Mungall (BBOP)

Protein families generated by: Kara Dolinski (PPOD), Michael S. Livstone (PPOD), and Paul Thomas (PANTHER)

PIs of the GO Consortium: Michael Ashburner (FlyBase), Judy Blake (MGI), Mike Cherry (SGD), and

MutS superfamily (PTHR11361)



Propagate both terms to all proteins in superfamily