Evolutionary relationships as a paradigm for integrating biological knowledge:

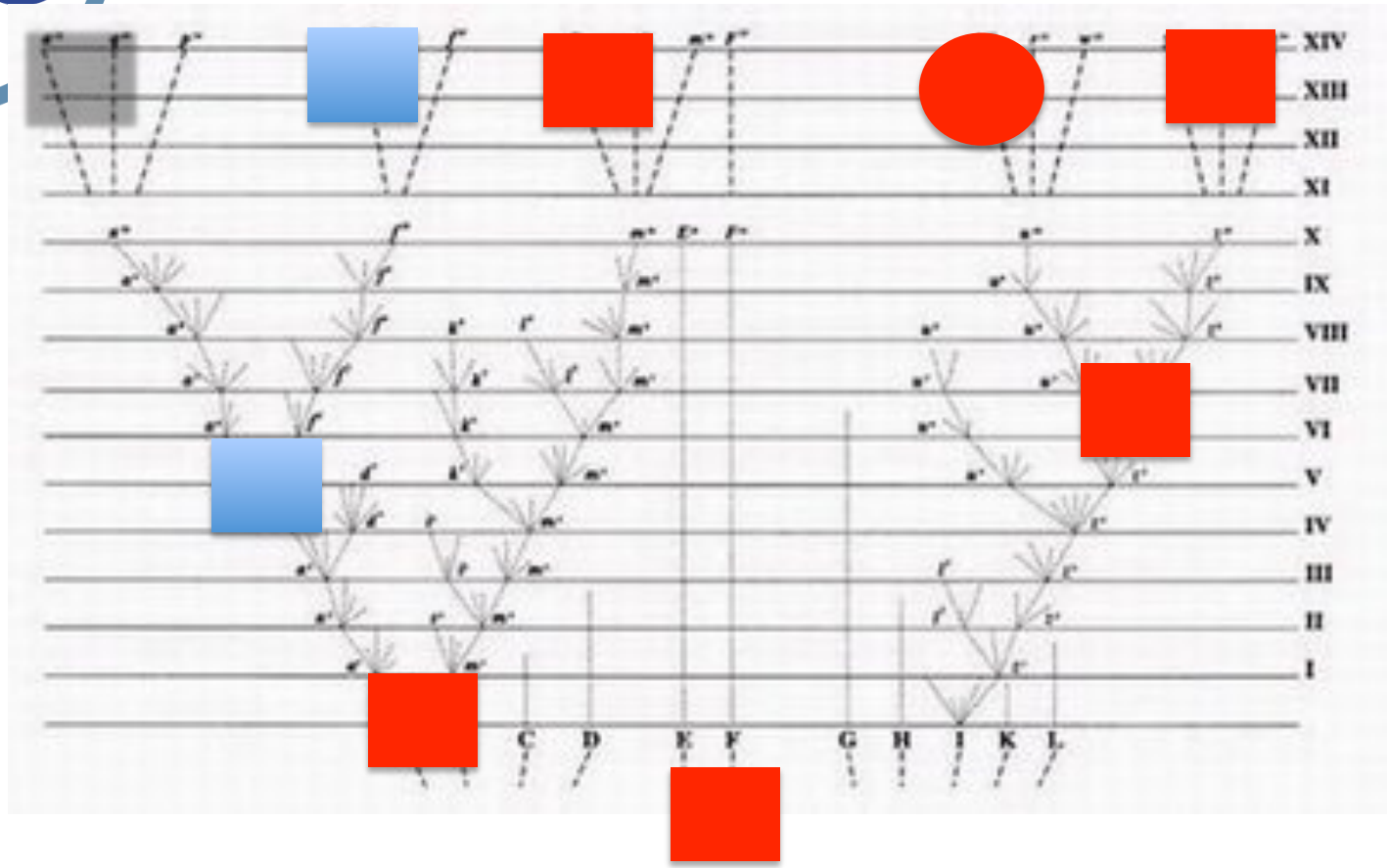
The GO Phylogenetic Annotation Project

Paul D. Thomas, Ph.D.
University of Southern California
Gene Ontology

Outline

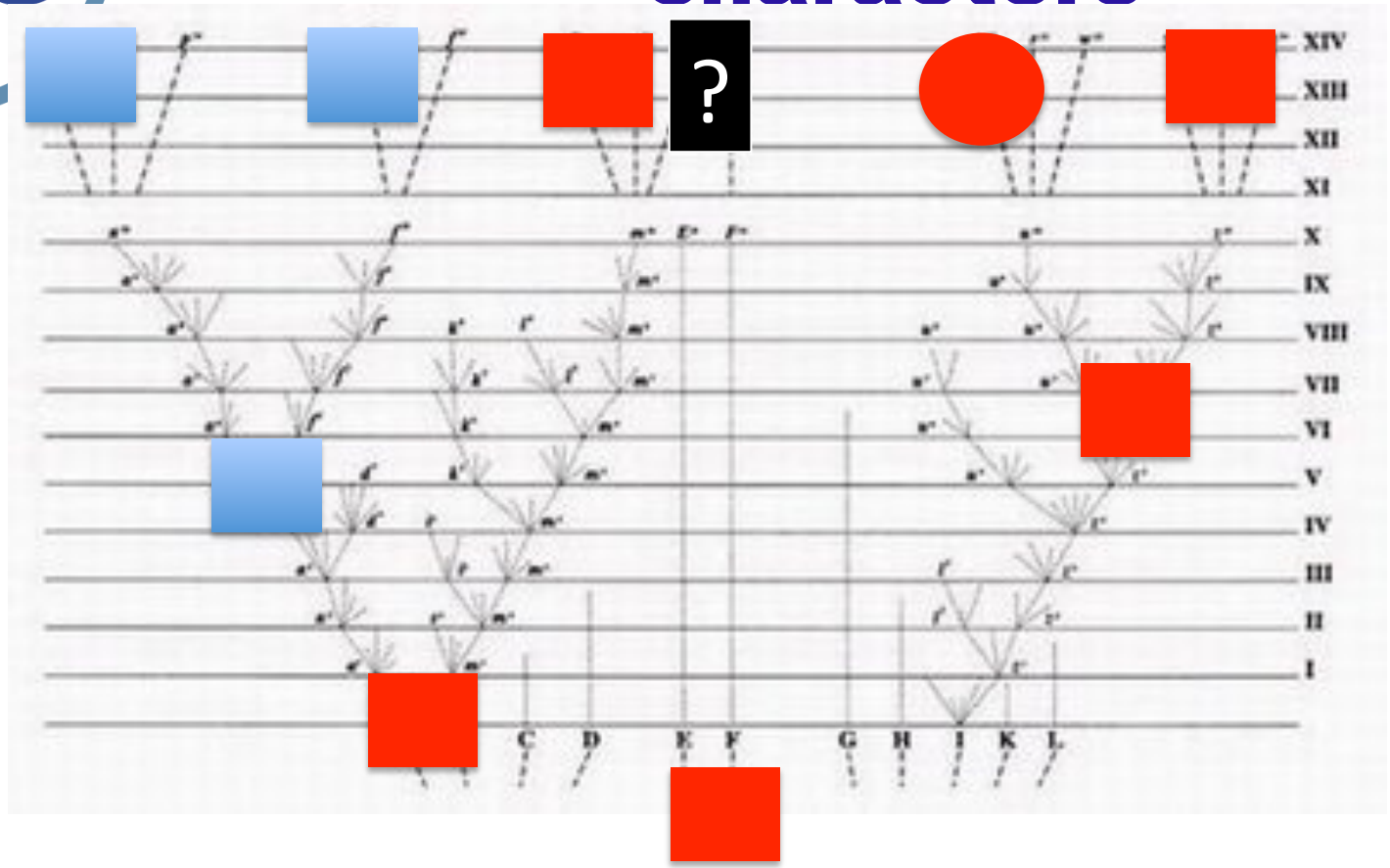
- Evolution and data integration
- GO Phylogenetic Annotation Project
  - Integrating information across multiple model organisms
  - Modeling evolution of the functions of related genes: molecular function, cellular component and biological process
  - Relationship to phenotype information

Darwin's species tree (1859)



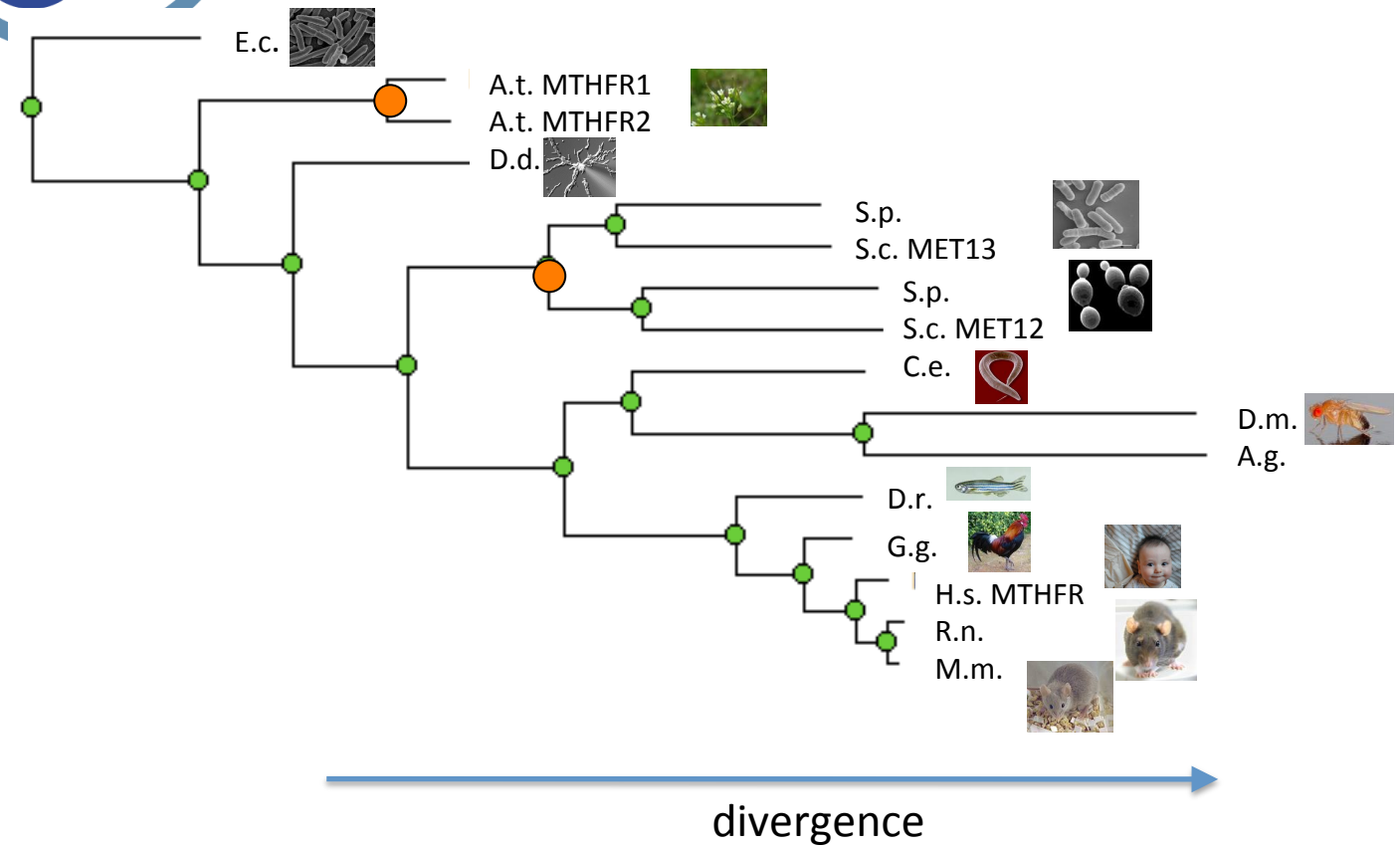
- **Characters in common are due to inheritance**
  - **Allows *inferences* about common ancestor**
  - **And specific changes along branches of a tree**

Inference of unknown characters



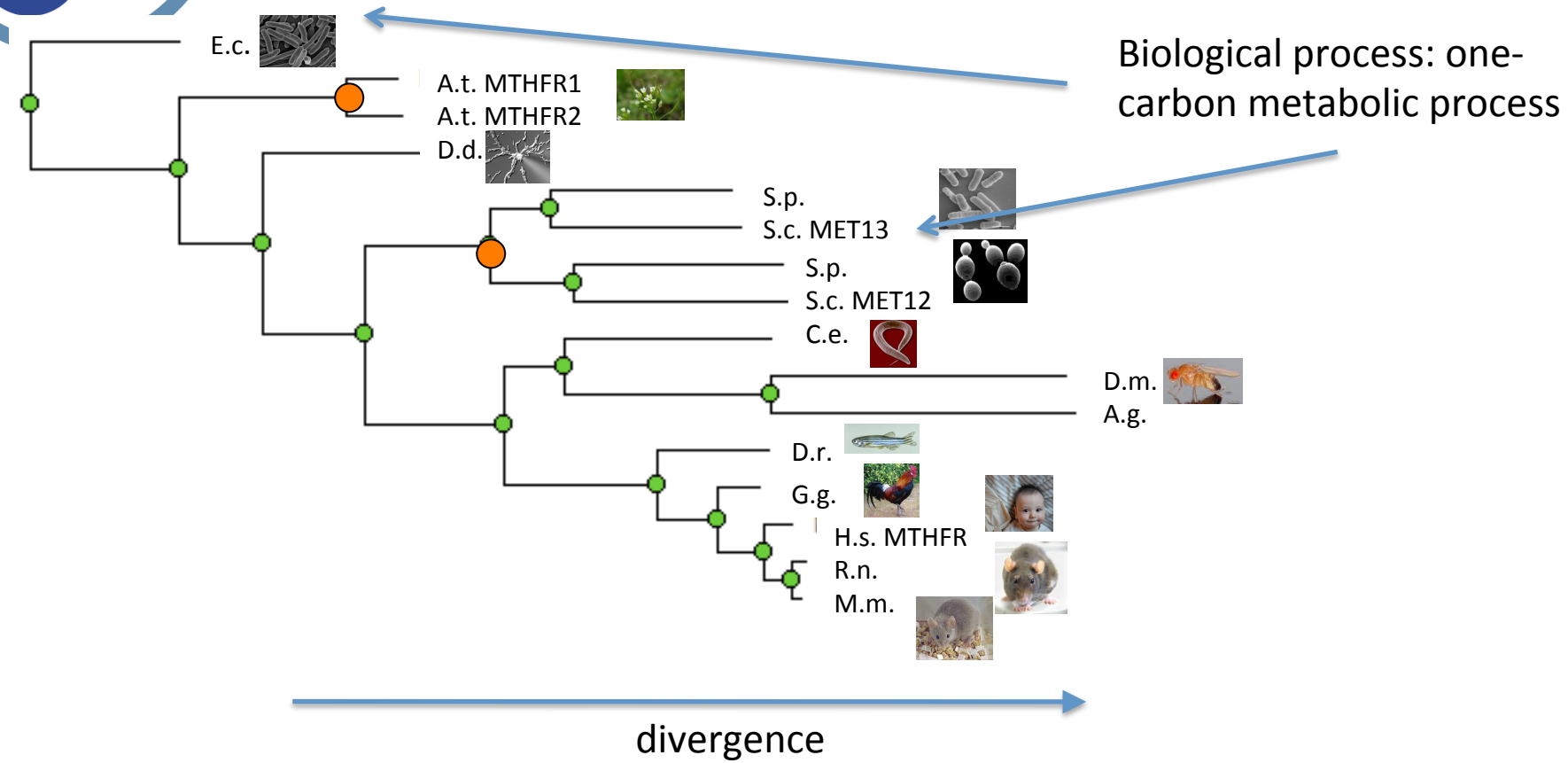
- **Inferred ancestral characters allow inference of unknown characters if we know the phylogeny**

We can do this today at the level of genes



- Complete genomes
  - Gene/genome sequences as characters for inferring tree and reconstructing ancestral states

Gene function as an evolutionary character



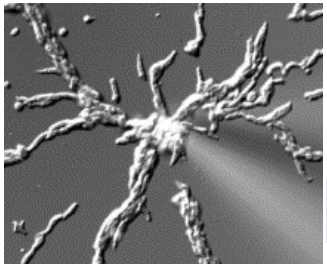
- Function likely present in common ancestor
- Human gene also likely to inherit the same function
- Can model both gain and loss of characters

Evolutionary framework

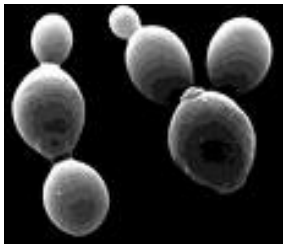
- **Evolutionary framework**
  - **Integrate information at points of common ancestry**
  - **Infer unknown character states of living organisms**
- **Both are inferences, and in many cases is vastly improved by biocuration**
- **Framework allows these inferences to be**
  - **Made**
  - **Recorded**
  - **Traced**
  - **Updated**







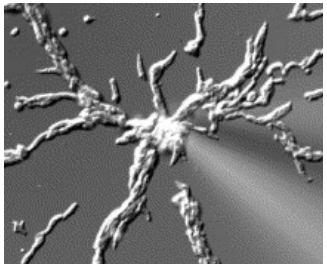
Gene Ontology Phylogenetic Annotation Project

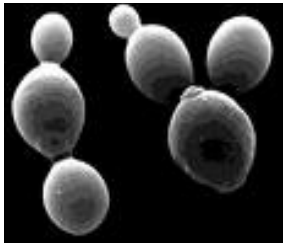











Phylogenetic Annotation Project

- GO database contains gene function information ("annotations") from experiments in over 120,000 published scientific papers
  - Primarily in 12 organisms
- Annotations are diverse, very incomplete, and can differ widely among related genes
- Goal: integrate the information from related genes into a model of gene function conservation and change
  - Allows inference of GO annotations by homology, especially for human genes
  - Allows synthesis of a view of how a given biological system is similar and different between model organisms

Integration of gene phylogenies with experimental information about gene function

- **Gene phylogenies**
  - **How are the genes among different model organisms related to each other? (orthologs, paralogs, in-paralogs, etc)**
  - **When did a given gene first appear, and in which extant organisms does it remain?**
- **Function**
  - **Which functions are conserved, and among which homologs?**

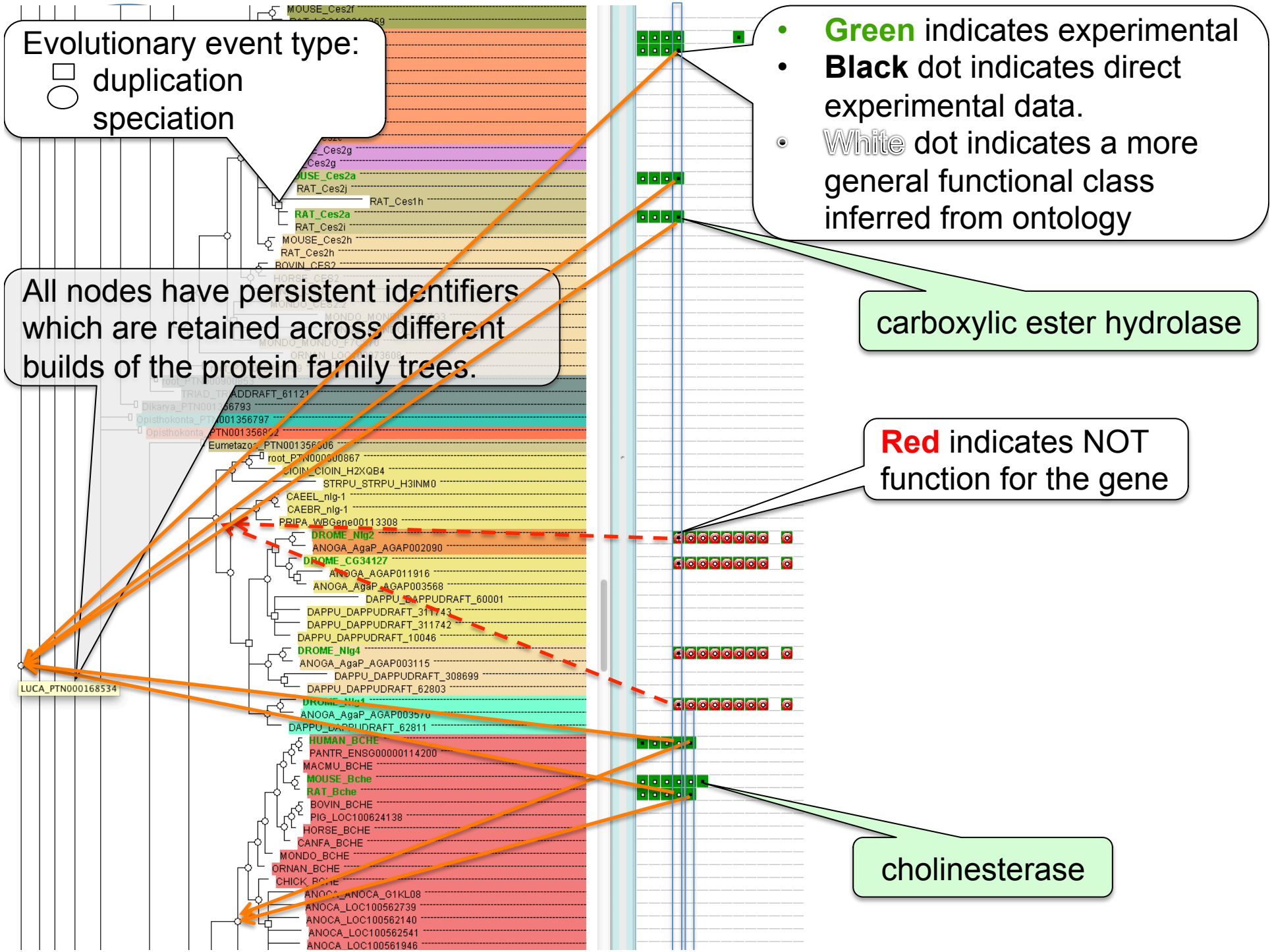
Use gene trees (and species tree) as basic data structure

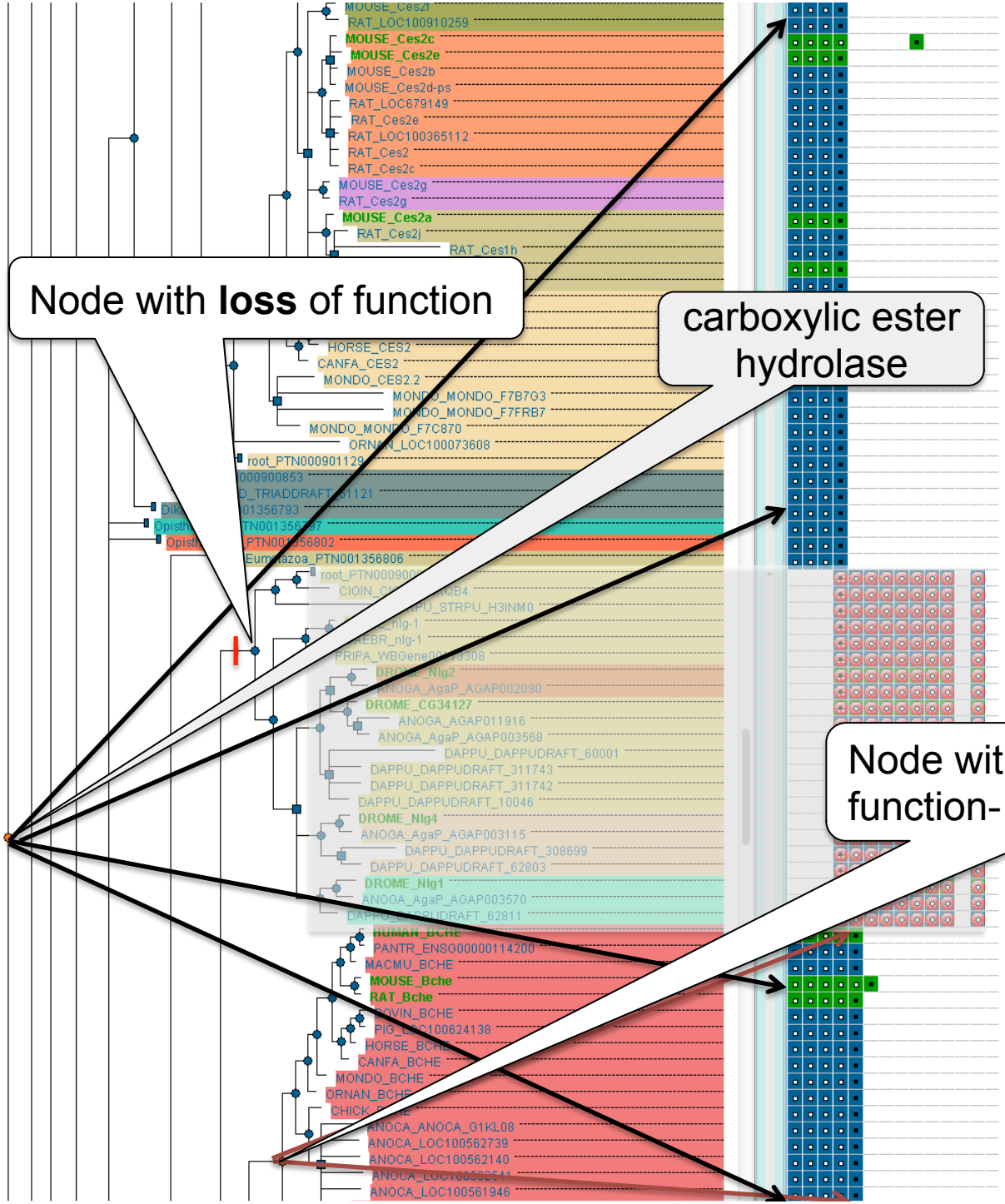
- **Orthologs are pairwise, so difficult to use in combination**
  - **for a gene family with 12 model organisms with only a single family member there are a minimum of 66 pairwise relations!**
  - **A tree is much simpler**
- **Trees allow explicit modeling of gain and loss of characters along tree**

Integration of multiple types of biological knowledge

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

- Software tool: PAINT
  - Build explicit model of function evolution


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.
White dot indicates a more general functional class inferred from ontology
carboxylic ester hydrolase
Red indicates NOT function for the gene
cholinesterase
LUCA_PTN000168534
MOUSE_Ces2f
RAT_Ces2g
MOUSE_Ces2a
RAT_Ces2j
RAT_Ces1h
RAT_Ces2a
RAT_Ces2i
MOUSE_Ces2h
RAT_Ces2h
BOVIN_CES2
root_PTN000000867
CIOIN_CIOIN_H2XQB4
STRPU_STRPU_H3INM0
CAEEL_nlg-1
CAEBR_nlg-1
PRIPA_WBGene00113308
DROME_Nlg2
ANOGA_AgaP_AGAP002090
DROME_CG34127
ANOGA_AGAP011916
ANOGA_AgaP_AGAP003568
DAPPU_DAPPUDRAFT_60001
DAPPU_DAPPUDRAFT_311743
DAPPU_DAPPUDRAFT_311742
DAPPU_DAPPUDRAFT_10046
DROME_Nlg4
ANOGA_AgaP_AGAP003115
DAPPU_DAPPUDRAFT_308699
DAPPU_DAPPUDRAFT_62803
HUMAN_BCHE
PANTR_ENSG00000114200
MACMU_BCHE
MOUSE_Bche
RAT_Bche
BOVIN_BCHE
PIG_LOC100624138
HORSE_BCHE
CANFA_BCHE
MONDO_BCHE
ORNAN_BCHE
ANOCA_ANOCA_G1KL08
ANOCA_LOC100562739
ANOCA_LOC100562140
ANOCA_LOC100562541



- PAINTed nodes –
  - 3 steps carried out by curator
  - Gain & Loss of function
- Inferred By Descendants
  - Experimental annotations provide evidence
- Inferred by Ancestry
  - Propagation to unannotated leaves

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

PAINT Overall Progress

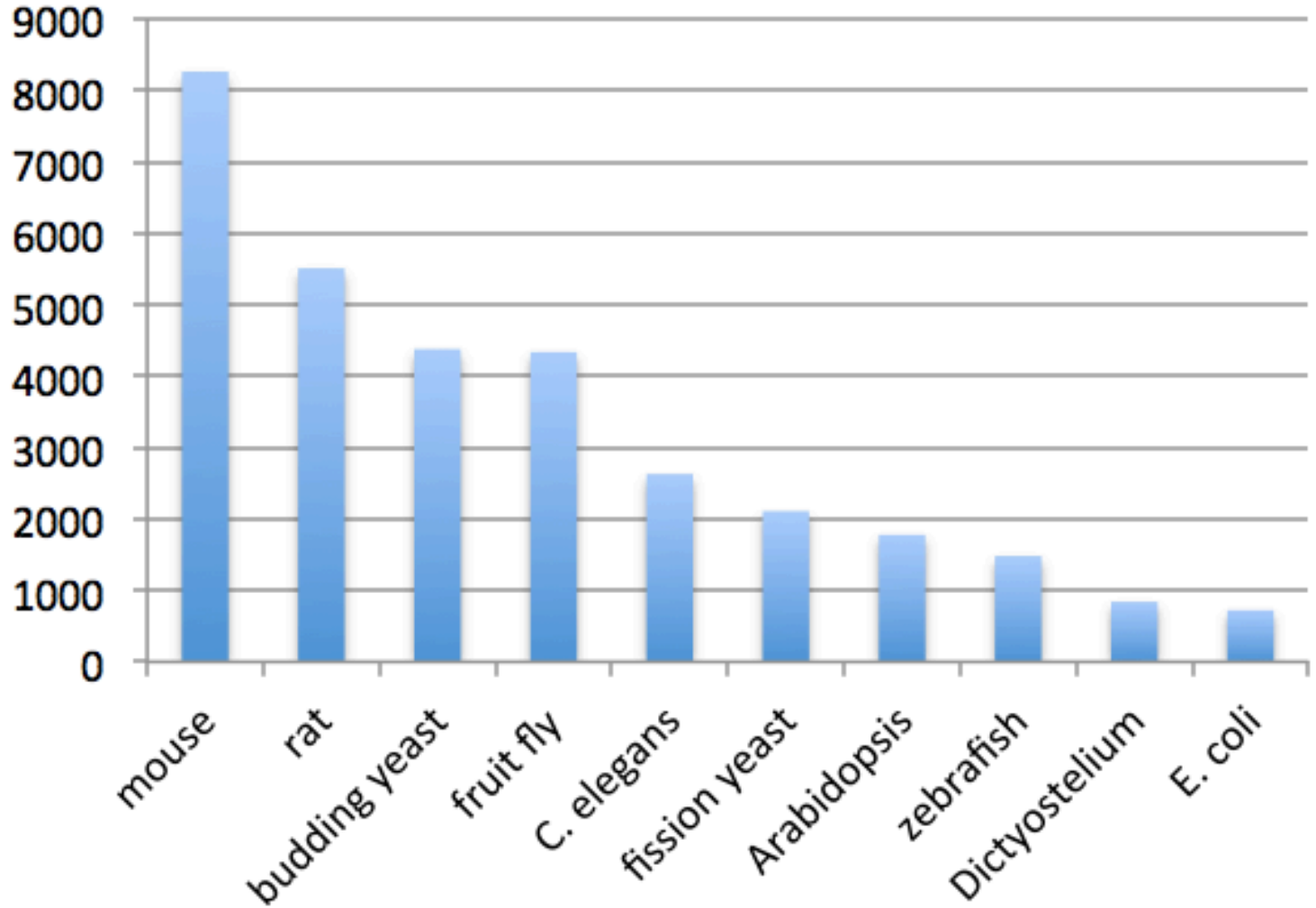
	8/25/2015
Total # families	1914
Total # sequences	306,293
Total # sequences with inferred annotations	264,322

Total progress measured in terms of
# of human genes covered in annotated families: 5976

Information from model organisms more than doubles the number of GO annotations for human genes

	PAINT annotations		Literature curation	
	genes	annotations	genes	annotations
Human	5118	20720	4227	17044
Mouse	5870	25248	3350	22382
Fly	2992	11762	2000	11624
Worm	3470	14842	2042	9680
Yeast	1286	3825	1685	9420
MOD (12)	44630	190153	21550	125863
Non-MOD (92)	219692	951995	2326	6576

Contribution from each model organism to increase in human gene annotations



Relies on well-developed infrastructure

- **Up-to-date sets of genes across model organisms and other key taxa (UniProt)**
- **Building of phylogenetic trees for all relevant gene families**
- **Stable tree node identifiers across updated versions**
- **Tracing of experimental evidence and updating if evidence changes**

Synthesis to infer conservation and divergence of a pathway

- **Bring together information about as many genes as possible in each system**
- **How similar is each model organism to the homologous human system?**
- **Specifically, where is it similar and where is it different?**

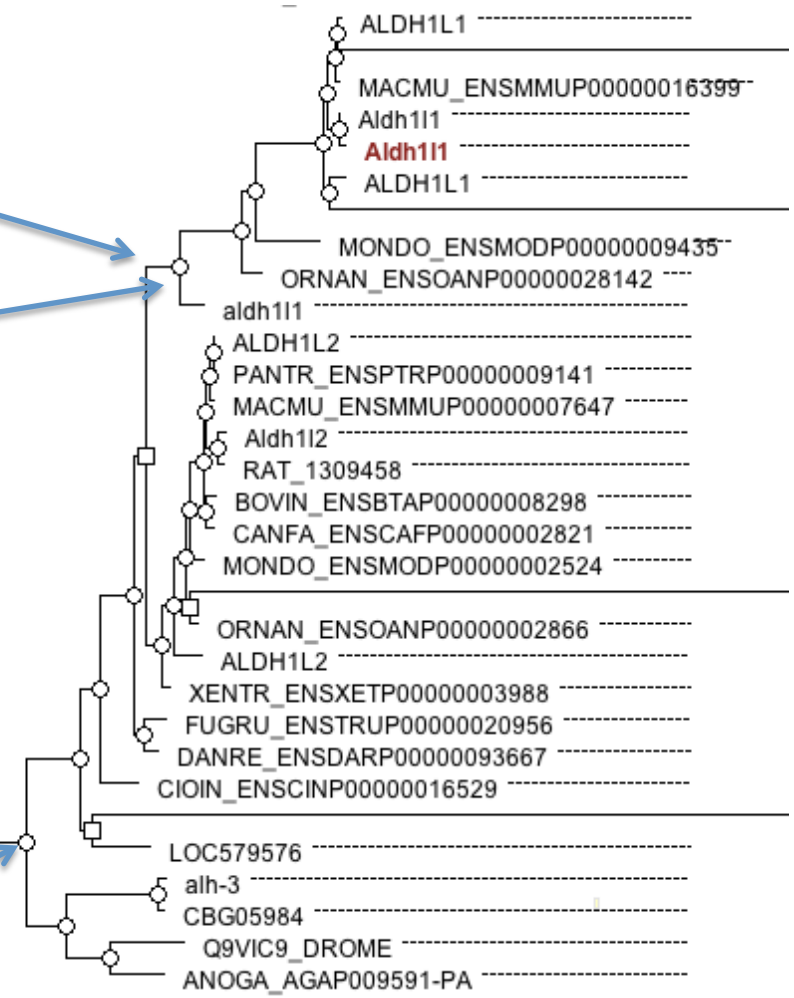
Appearance of a “new gene” and new function by gene duplication

Formyltetrahydrofolate dehydrogenase (mitochondrial) appears via duplication of the cytosolic gene

Gene present in tetrapod common ancestor

Formyltetrahydrofolate dehydrogenase (cytosolic) appears via duplication of a different dehydrogenase

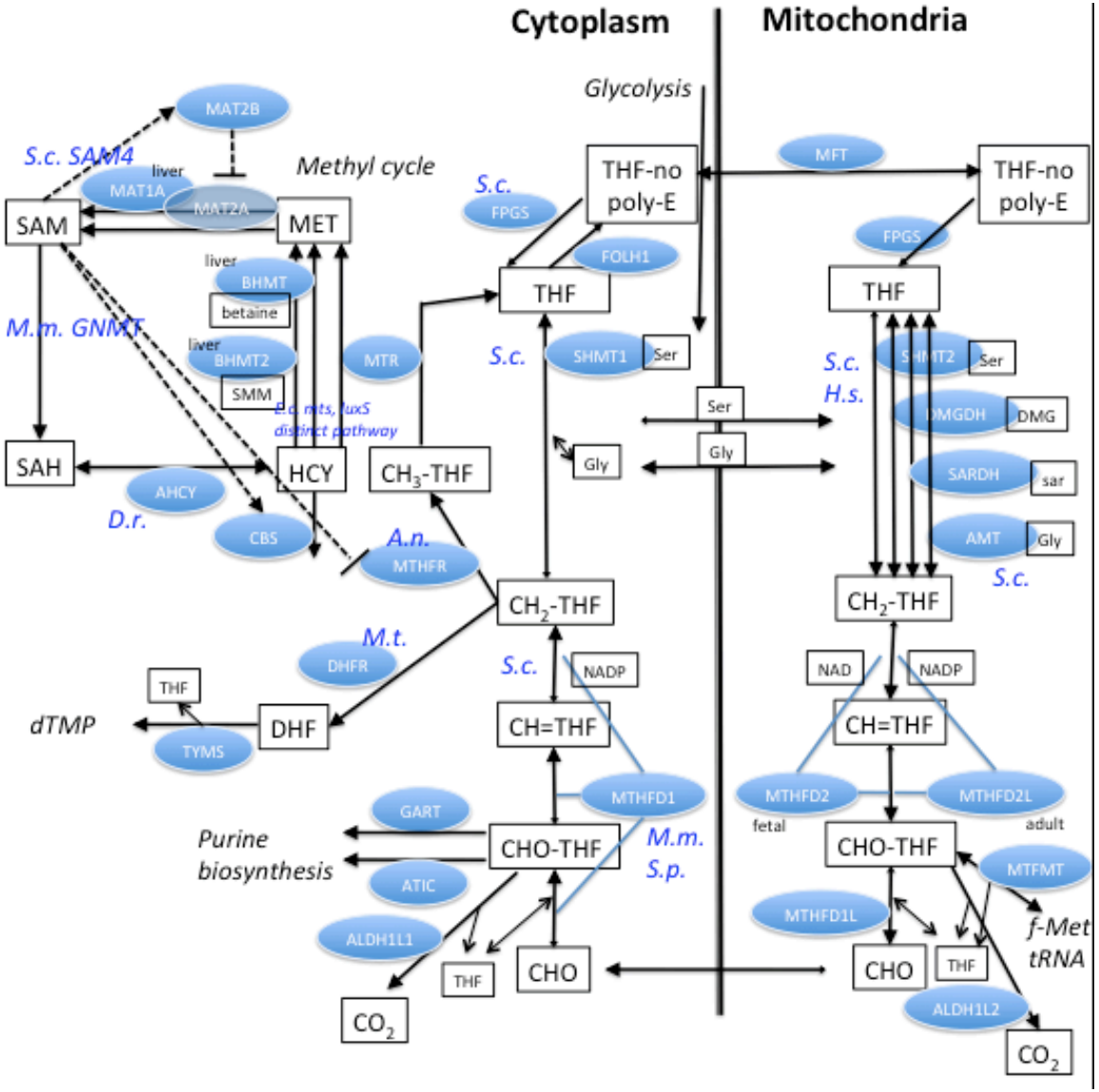
Gene present in metazoan common ancestor



One-carbon pathway

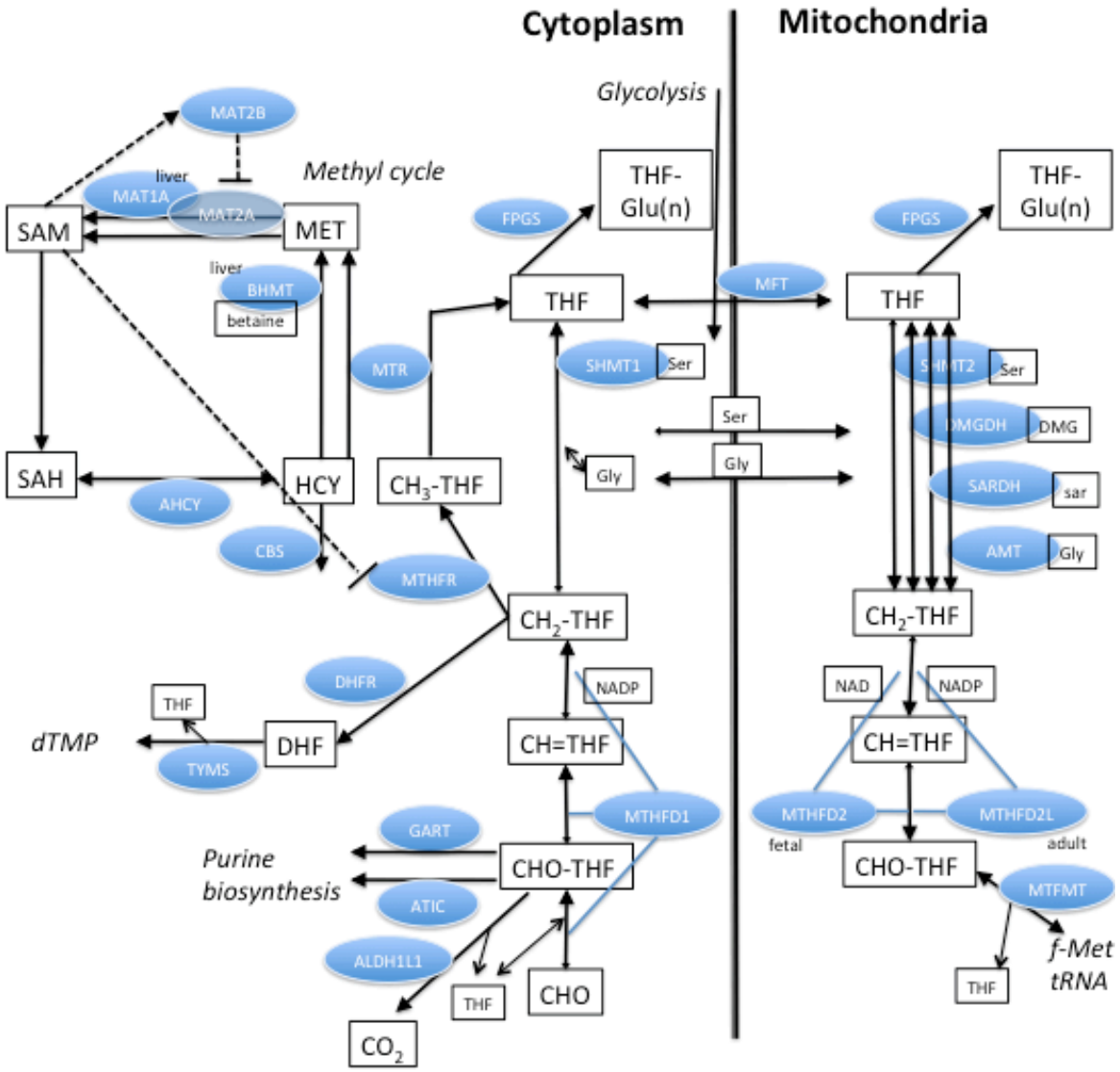
Consensus human pathway

From Nicholas Marini, Rine lab



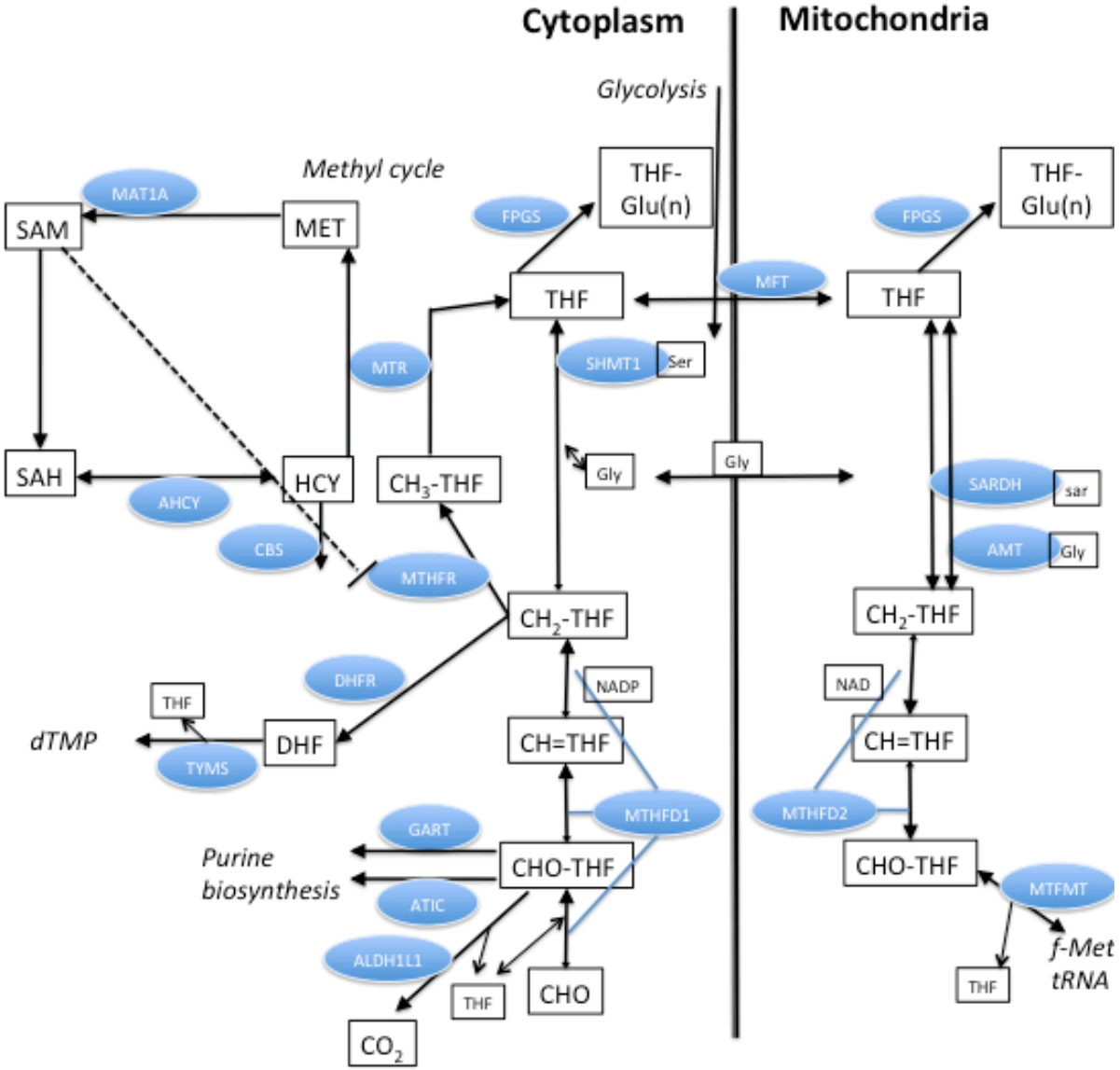
One-carbon pathway

Inferred pathway:
Vertebrate
common ancestor



One-carbon pathway

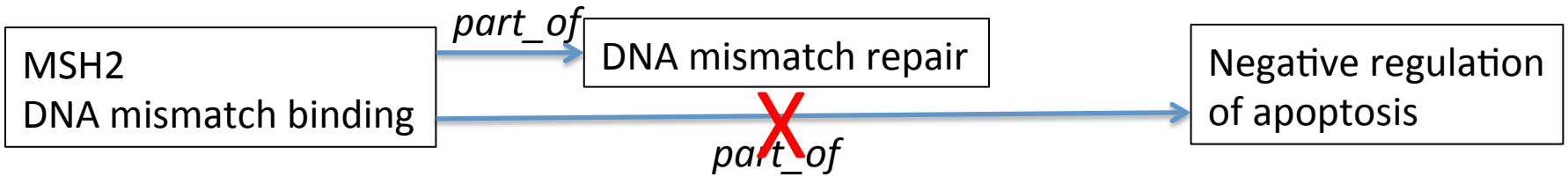
Inferred pathway:
Metazoan
common ancestor



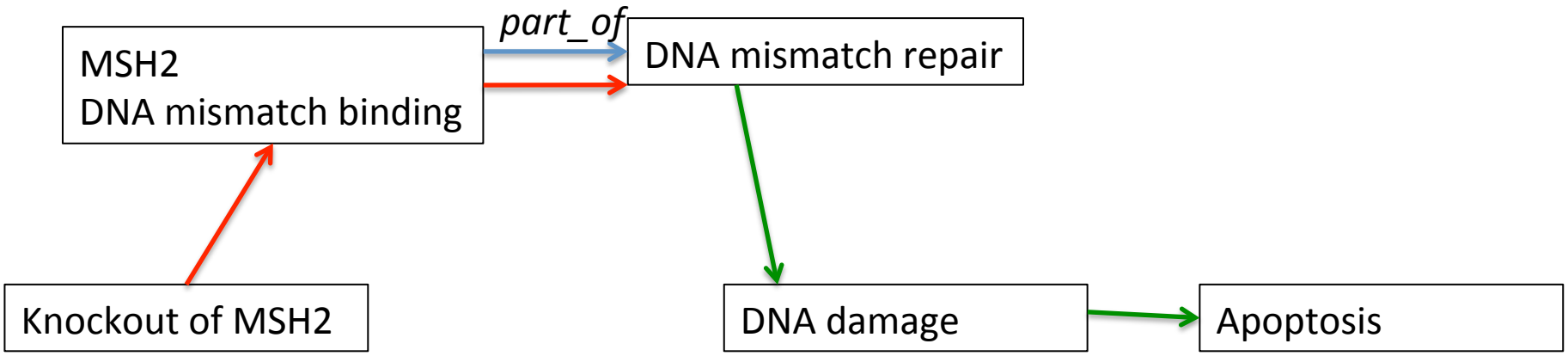
GO phylogenetic annotation currently makes only limited use of phenotype data

- **Many GO annotations are derived from experimentally observed phenotypes**
- **Phylogenetic annotation to date does not usually build them into the evolutionary model**
  - **Aim of GO is to capture "normal biology"**

Example of phenotype-derived GO annotation



Example: can the conserved process explain the phenotype?



Perturbations can be traced through network of “normal” causal relationships

We need to develop the tools to make computational representations of these causal connections!

Summary

- **The GO Consortium has developed a general infrastructure for inferring and annotating the evolution of any biological "character"**
  - **Integrate information at points of common ancestry**
  - **Infer unknown character states of living organisms**
- **Evolutionary information can help identify similarities and differences between a model and a human system at the level of biological pathways/processes**
- **Could potentially be extended to / integrated with phenotype information**
  - **Capture computational model of normal biology and how perturbations can result in particular phenotypes**

Acknowledgments

- **PANTHER/USC**
  - **John Casagrande**
  - **Sagar Poudel**
  - **Huaiyu Mi**
  - **Anushya Muruganujan**
- **BBOP/LBNL**
  - **Suzanna Lewis**
  - **Chris Mungall**
  - **Ed Lee**
- **GO Phylogenetic curators**
  - **Mark Feuermann, Swiss-Prot**
  - **Pascale Gaudet, NextProt**
  - **Huaiyu Mi, USC**
  - **Karen Christie, MGI/ Jackson Lab**
  - **Donghui Li, TAIR**

- UniProt
  - Alan de Sousa
  - Maria Martin
  - Claire O'Donovan
- Other GO PIs
  - Judy Blake
  - Mike Cherry
  - Suzanna Lewis
  - Paul Sternberg

- Nicholas Marini, Jasper Rine

- Funding
  - HG002273 (GO)
  - GM081084 (Phylogenetic Annotation infrastructure development)