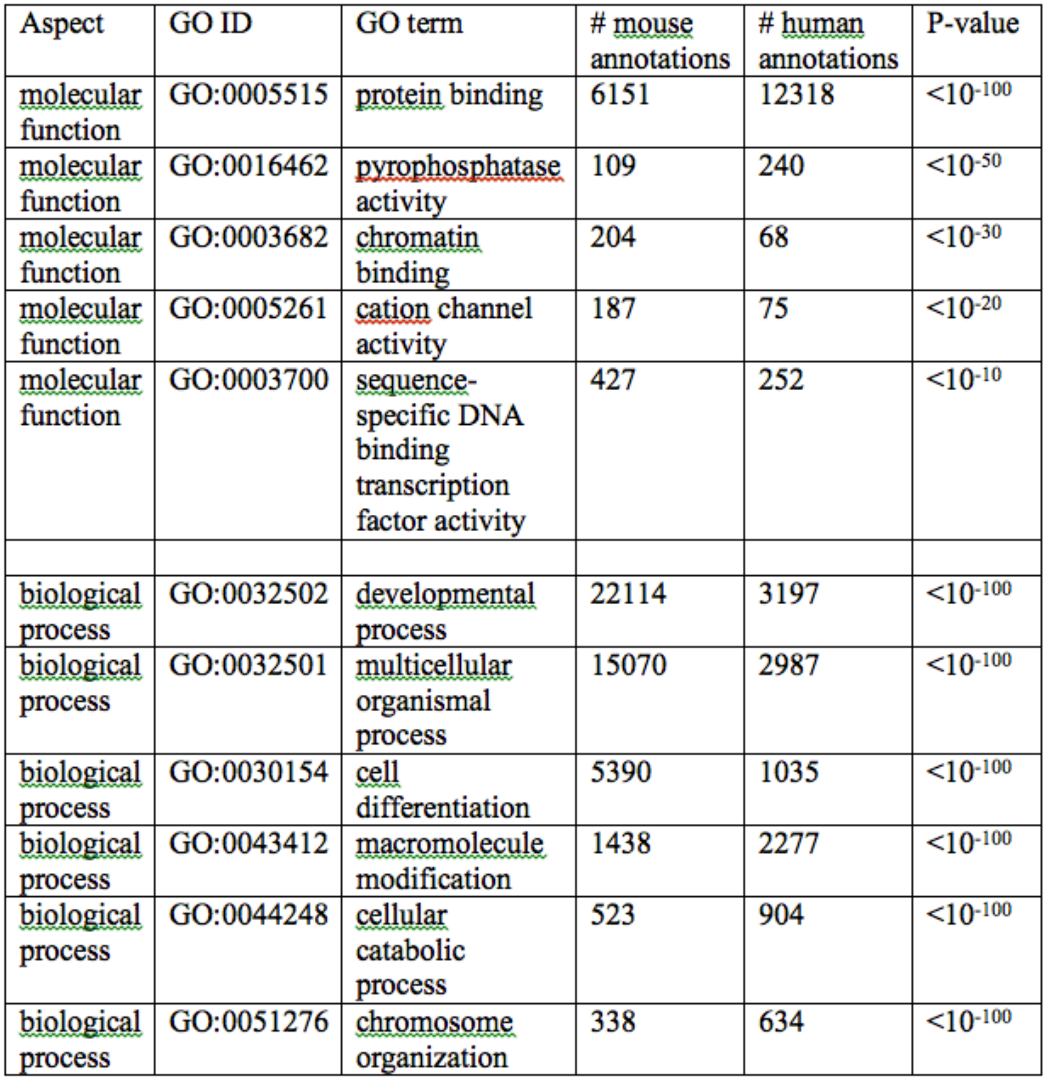
Basic problem formulation

- We have information about gene function from experiments in diverse organisms
- How do we integrate information about related genes to
  - Get a fuller picture of gene function
  - Annotate genes that have not been fully explored experimentally

Example: Annotations for human and mouse genes are largely complementary



Aspect	GO ID	GO term	# mouse annotations	# human annotations	P-value
molecular function	GO:0005515	protein binding	6151	12318	$<10^{-100}$
molecular function	GO:0016462	pyrophosphatase activity	109	240	$<10^{-50}$
molecular function	GO:0003682	chromatin binding	204	68	$<10^{-30}$
molecular function	GO:0005261	cation channel activity	187	75	$<10^{-20}$
molecular function	GO:0003700	sequence-specific DNA binding transcription factor activity	427	252	$<10^{-10}$
biological process	GO:0032502	developmental process	22114	3197	$<10^{-100}$
biological process	GO:0032501	multicellular organismal process	15070	2987	$<10^{-100}$
biological process	GO:0030154	cell differentiation	5390	1035	$<10^{-100}$
biological process	GO:0043412	macromolecule modification	1438	2277	$<10^{-100}$
biological process	GO:0044248	cellular catabolic process	523	904	$<10^{-100}$
biological process	GO:0051276	chromosome organization	338	634	$<10^{-100}$

“Transitive annotation”

- “ISS” GO evidence code: Inference from sequence similarity
- A class of database search algorithm (e.g. BLAST) has become a metaphor
  - Implies “genes have similar functions because they have similar sequences”

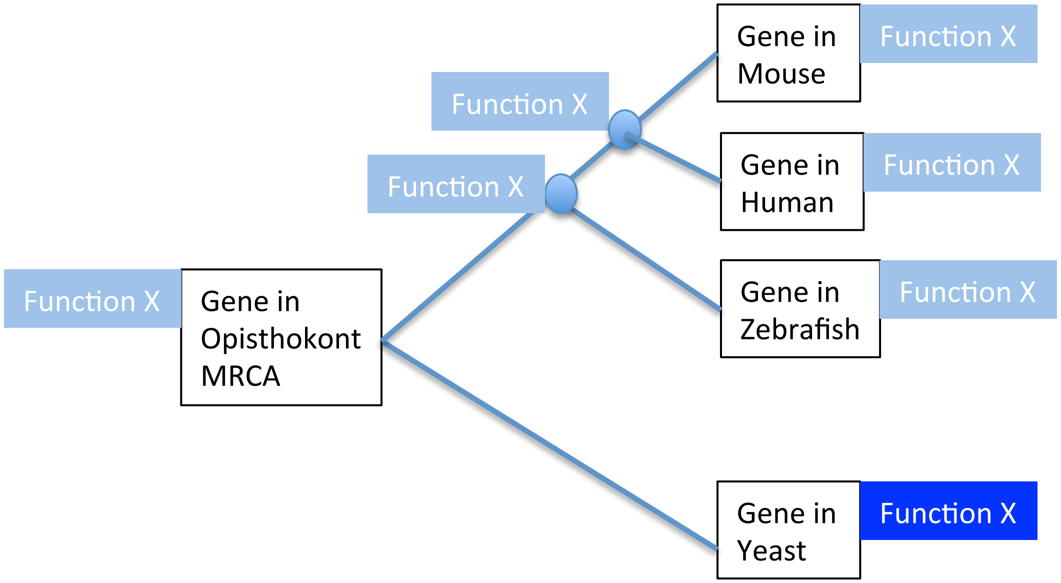
…AVSQPDE…

$P<10^{-100}$

…AVSNPDD…

What is transitive annotation?

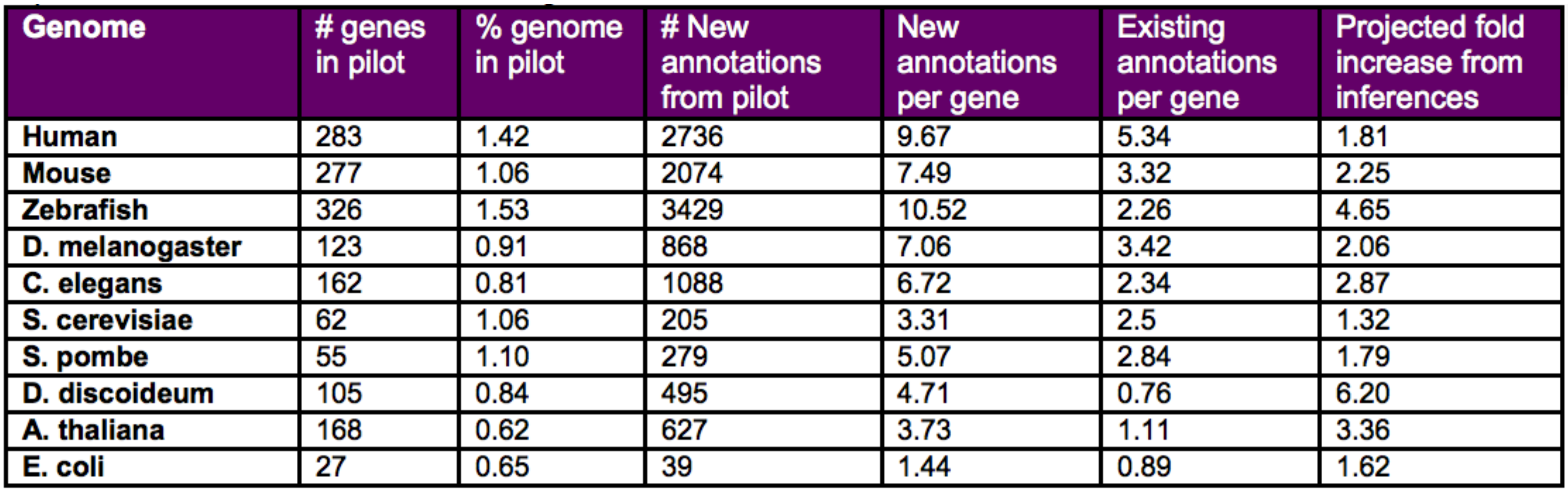
- More properly, transitive annotation of function is inheritance!
  - Two sequences are similar **because** they are homologous (at least for relatively long, non-repetitive sequences, i.e. almost all genes)
  - related genes have a common function because their common ancestor had that function, which 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



Transitive annotation using annotated ancestral genes

- For the Reference Genome Project, we want to be explicit about evolutionary inferences
  - 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
    - Annotation of a tree node means “this function evolved on the branch prior to this node”
    - A NOT annotation of a tree node means “this ancestral function was lost on the branch prior to this node”

Phylogenetic annotation pilot



Genome	# genes in pilot	% genome in pilot	# New annotations from pilot	New annotations per gene	Existing annotations per gene	Projected fold increase from inferences
Human	283	1.42	2736	9.67	5.34	1.81
Mouse	277	1.06	2074	7.49	3.32	2.25
Zebrafish	326	1.53	3429	10.52	2.26	4.65
D. melanogaster	123	0.91	868	7.06	3.42	2.06
C. elegans	162	0.81	1088	6.72	2.34	2.87
S. cerevisiae	62	1.06	205	3.31	2.5	1.32
S. pombe	55	1.10	279	5.07	2.84	1.79
D. discoideum	105	0.84	495	4.71	0.76	6.20
A. thaliana	168	0.62	627	3.73	1.11	3.36
E. coli	27	0.65	39	1.44	0.89	1.62

All annotations, including curator notes, available at pantree.org

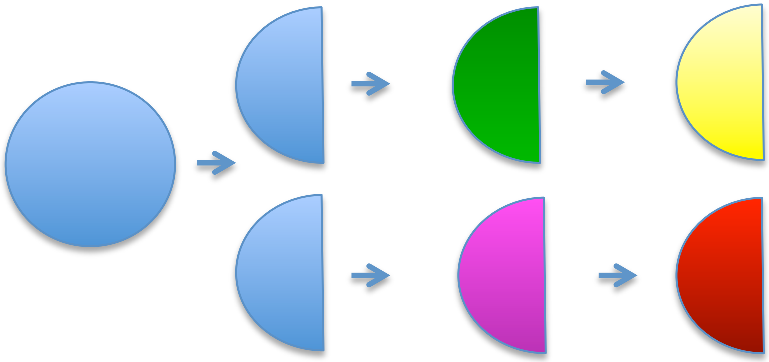
Protein families and function evolution: basics

Protein families

- Arise from copying and divergence
  - A tree is a natural way to represent this (Darwin)
- A family derives from a single common ancestor, and members retain ("conserve") sequence similarity due to functional constraint
- Proteins are modular: part or all of a protein may be copied and conserved, but a minimum functional unit must remain (a "domain")

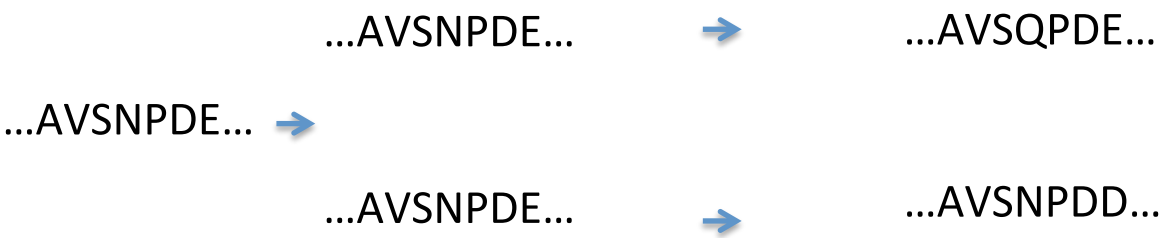
Representing evolution of related genes

- Start with Darwin's basic model:
  - Copying
    - An ancestral population splits into two separate populations
    - Each population is nearly identical at first
  - Divergence
    - Each population (copy) changes *independently* over generations
      - NATURAL SELECTION: adaptation to different environment



Representing evolution of related genes

- "Gene families"
- Add detail from population genetics/molecular evolution to apply to genes
  - Copying
    - An ancestral species splits into two separate species
      - SPECIATION
    - A gene is duplicated in one population and subsequently inherited
      - DUPLICATION
  - Divergence
    - Each copy (gene sequence) changes *independently* over generations
      - NATURAL SELECTION: sequence substitutions to adapt to new function/ role
      - NEUTRAL DRIFT: accumulation of "neutral" substitutions

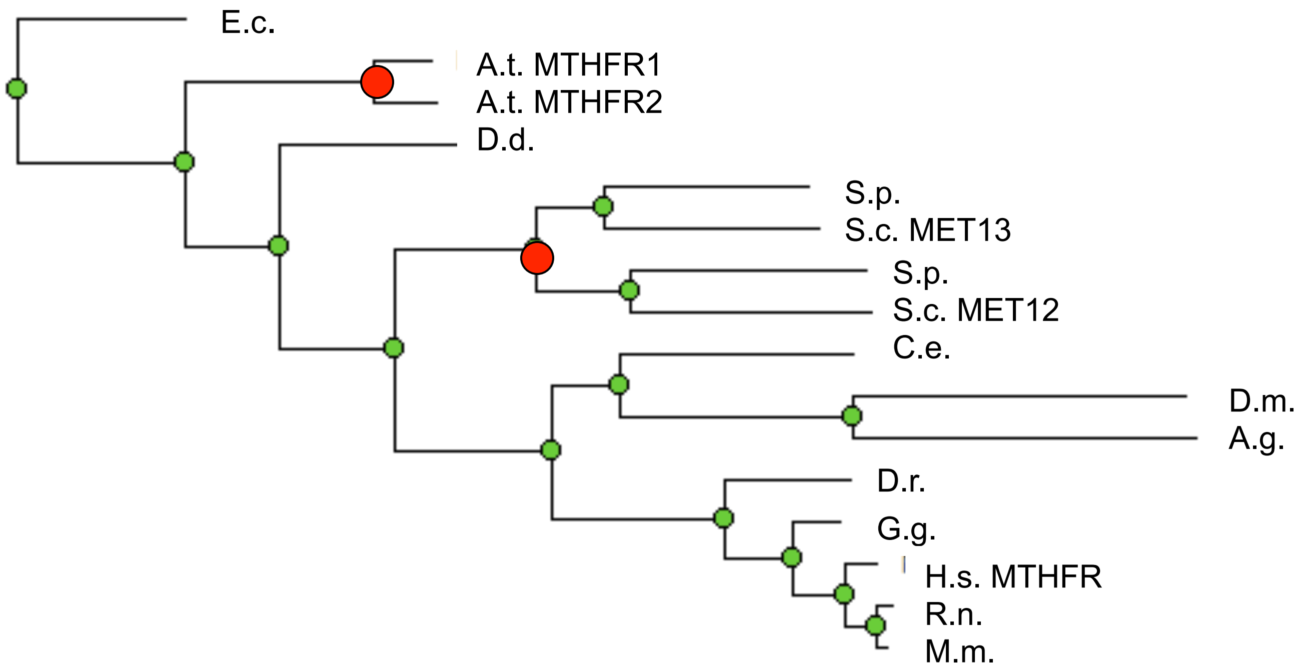


...AVSNPDE... → ...AVSQPDE...

...AVSNPDE... →

...AVSNPDE... → ...AVSNPDD...

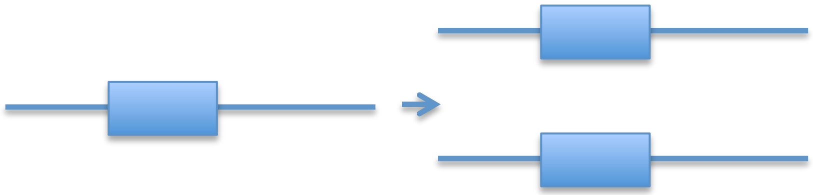
A gene tree



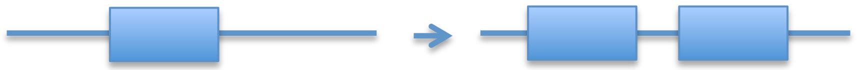
- Branch lengths: rate of sequence evolution
  - For neutral changes this can often act as a “molecular clock”
  - Non-neutral changes will speed up the rate of evolution

How does this relate to gene function?

- Copying
  - Speciation: one gene in each genome; two different species/genomes
  - Gene duplication: two copies in each genome with redundant function
- Divergence
  - Both copies begin with same function so are likely to retain at least some aspects of that ancestral function
  - Divergence more likely for gene duplication than speciation
    - Extra gene free from inherited functional constraints



speciation



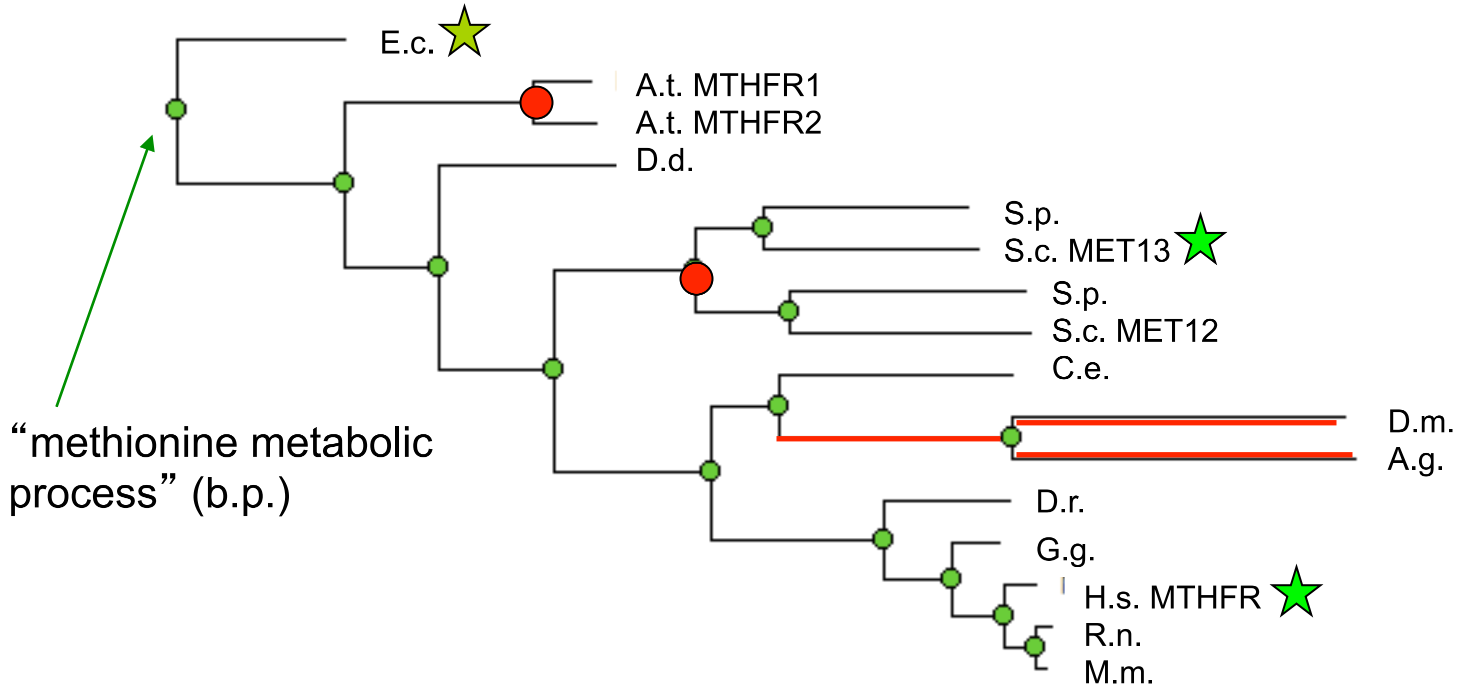
duplication

Gene duplication and functional novelty

- “Neofunctionalization” model
  - One copy retains ancestral function
  - One copy adapts to new function
    - More diverged copy often recognizable as having larger branch length
- “Subfunctionalization” model
  - Ancestral gene has at least two functions/ specificities
  - Each copy adapts to “specialize” in a subset of the ancestral functions

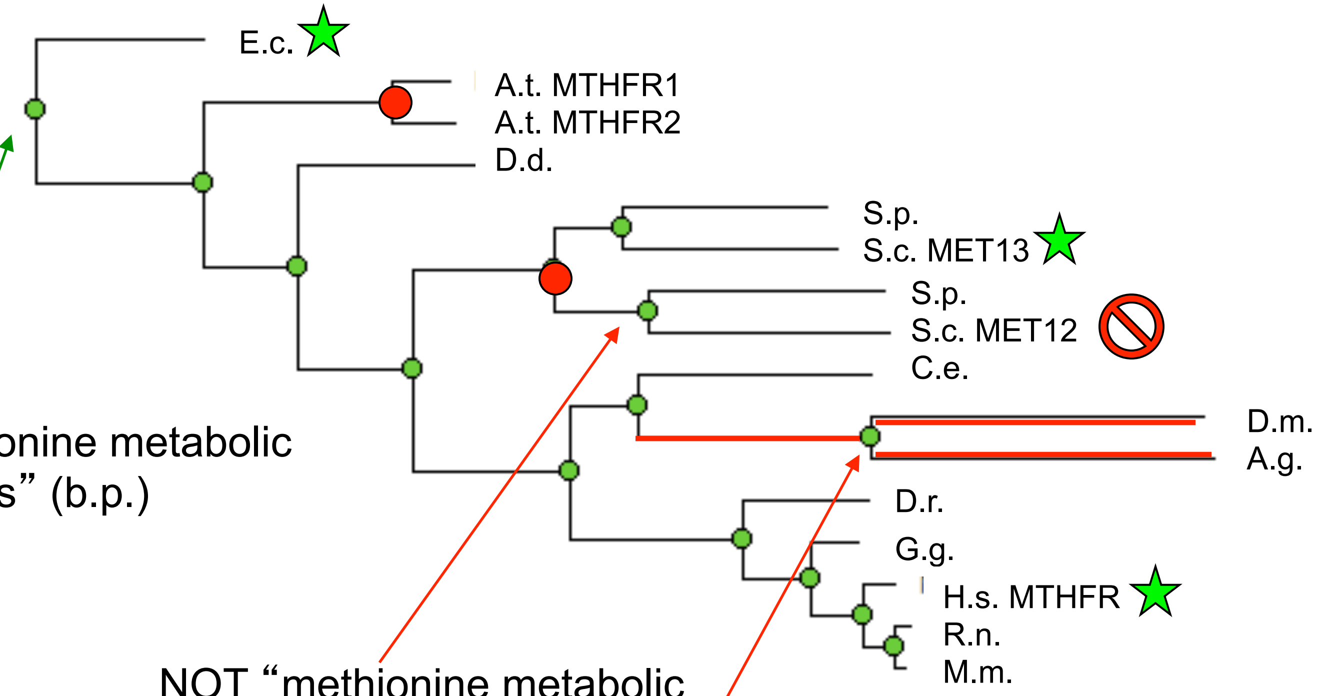
Homology inference in a tree

inheritance and divergence of function



Homology inference in a tree

inheritance and divergence of function



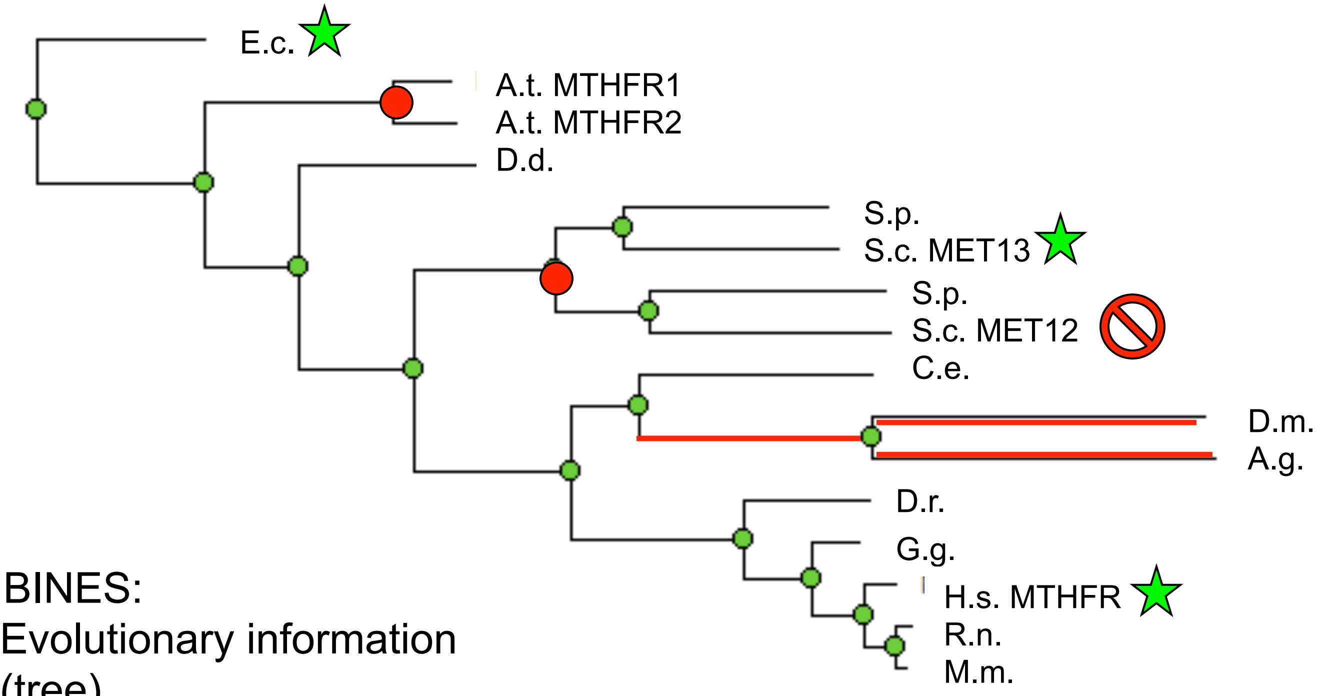
"methionine metabolic process" (b.p.)

NOT "methionine metabolic process" (b.p.)

NOT "methionine metabolic process" (b.p.)?

Homology inference in a tree

inheritance and divergence of function



COMBINES:

1. Evolutionary information (tree)
2. Experimental knowledge (GO annotations from literature)
3. Organism-specific biological knowledge (curators)

Orthologs and 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