Barcelona Biomedical Research Park

CRG Centre de Regulació Genòmica

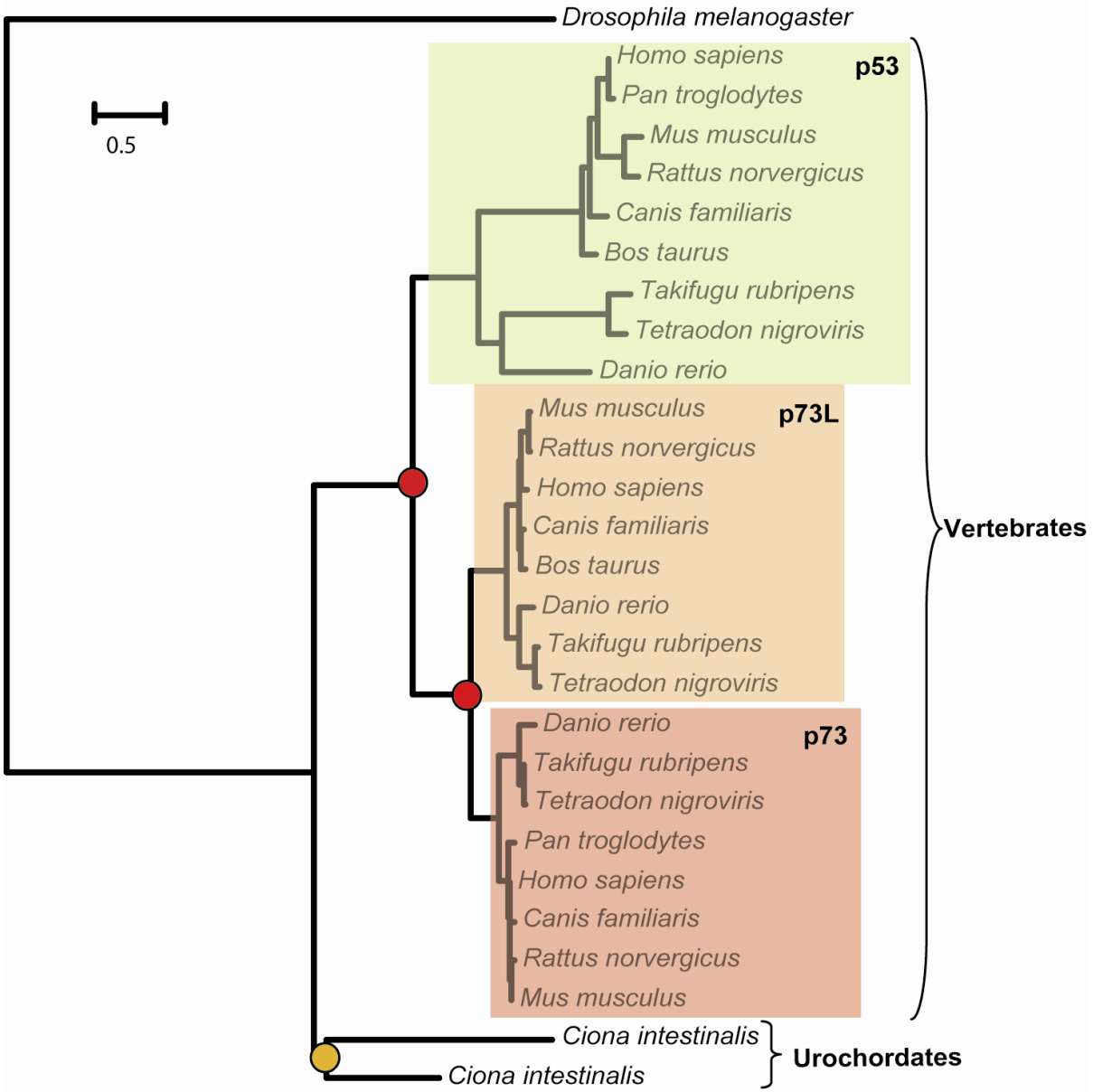
Orthology
Part II

Orthology prediction methods

Toni Gabaldón
Centre for Genomic Regulation (CRG), Barcelona

Classical approach: phylogenetic inference

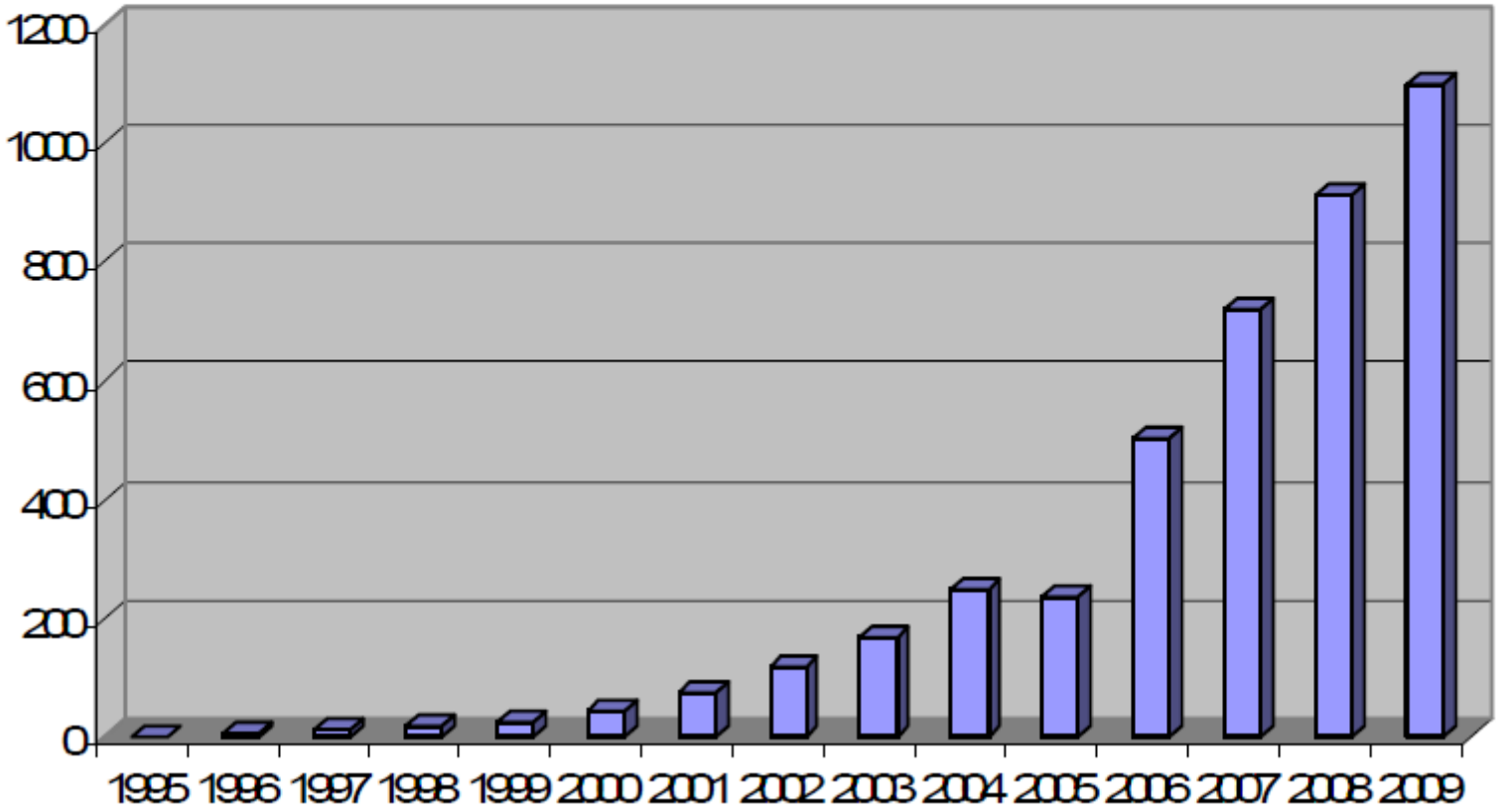
- Build a gene tree
- Compare to the species tree
- Infer duplications and speciation events
- Assign orthology and paralogy relationships accordingly

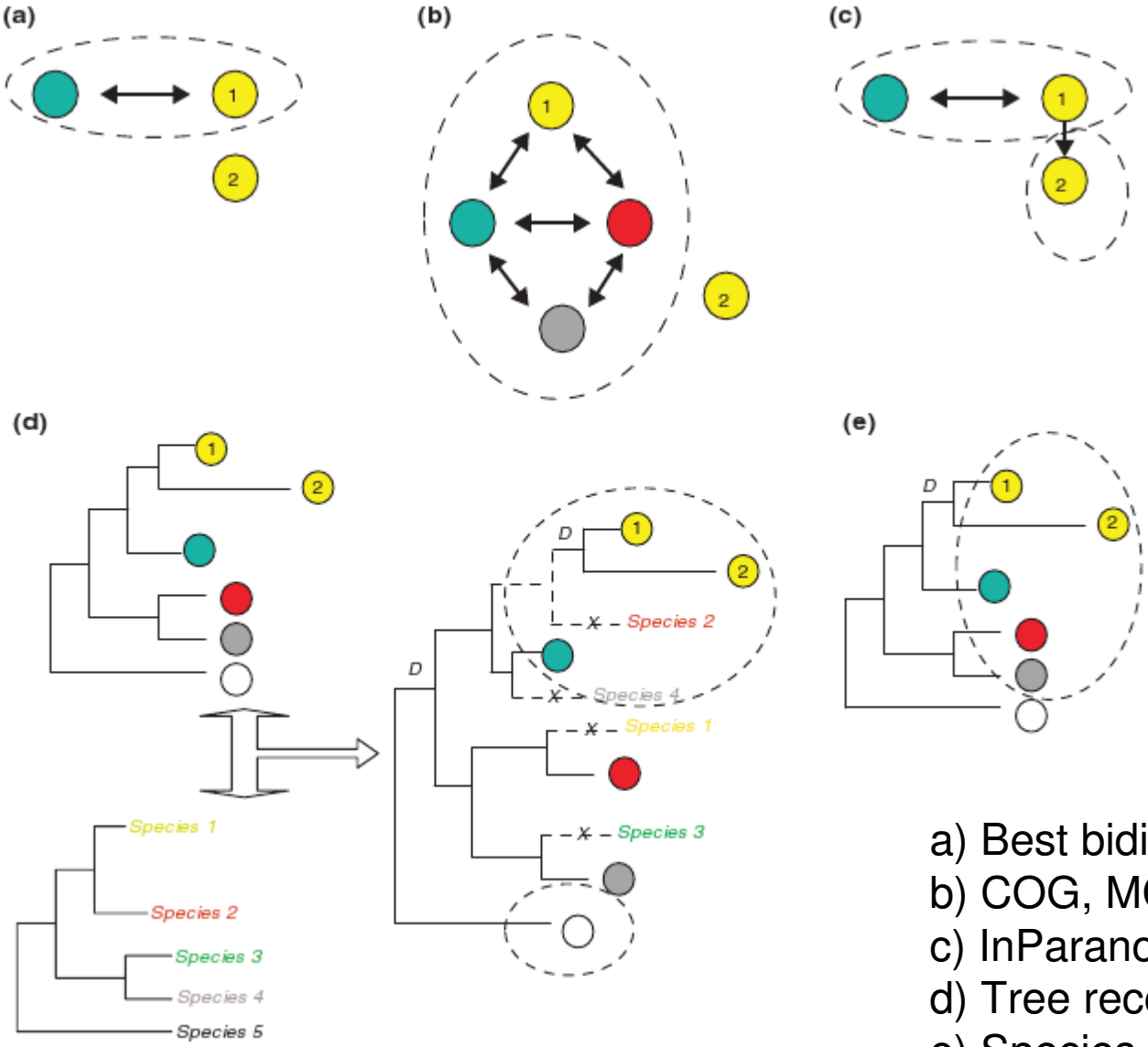

Drosophila melanogaster
0.5
Homo sapiens
Pan troglodytes
Mus musculus
Rattus norvergicus
Canis familiaris
Bos taurus
Takifugu rubripens
Tetraodon nigroviris
Danio rerio
p53
Mus musculus
Rattus norvergicus
Homo sapiens
Canis familiaris
Bos taurus
Danio rerio
Takifugu rubripens
Tetraodon nigroviris
p73L
Danio rerio
Takifugu rubripens
Tetraodon nigroviris
Pan troglodytes
Homo sapiens
Canis familiaris
Rattus norvergicus
Mus musculus
p73
Vertebrates
Ciona intestinalis
Ciona intestinalis
Urochordates

Going genome-wide scale:

Everything must be done automatic and “blind”

Completely sequenced genomes





a) Best bidirectional hits
b) COG, MCL-clustering approach
c) InParanoid
d) Tree reconciliation
e) Species-overlap (PhylomeDB)

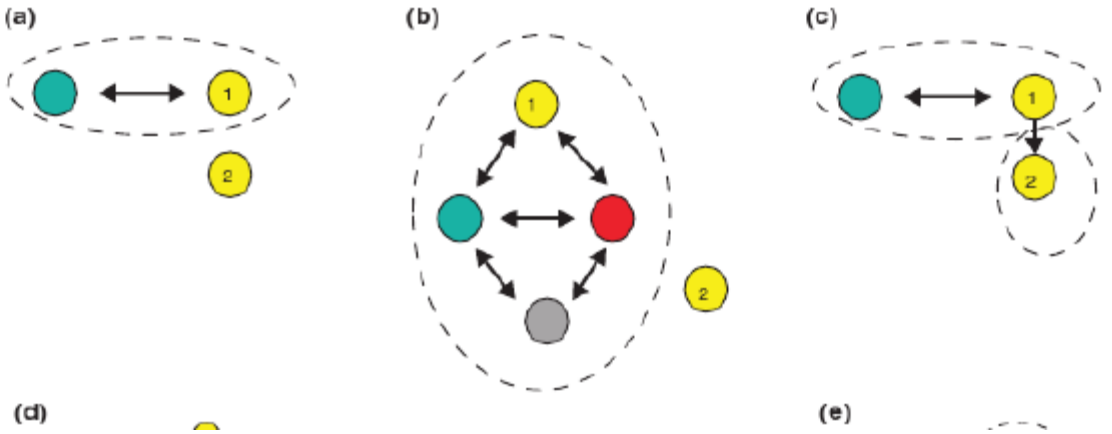
Gabaldón, T. *Genome Biology* (2008)

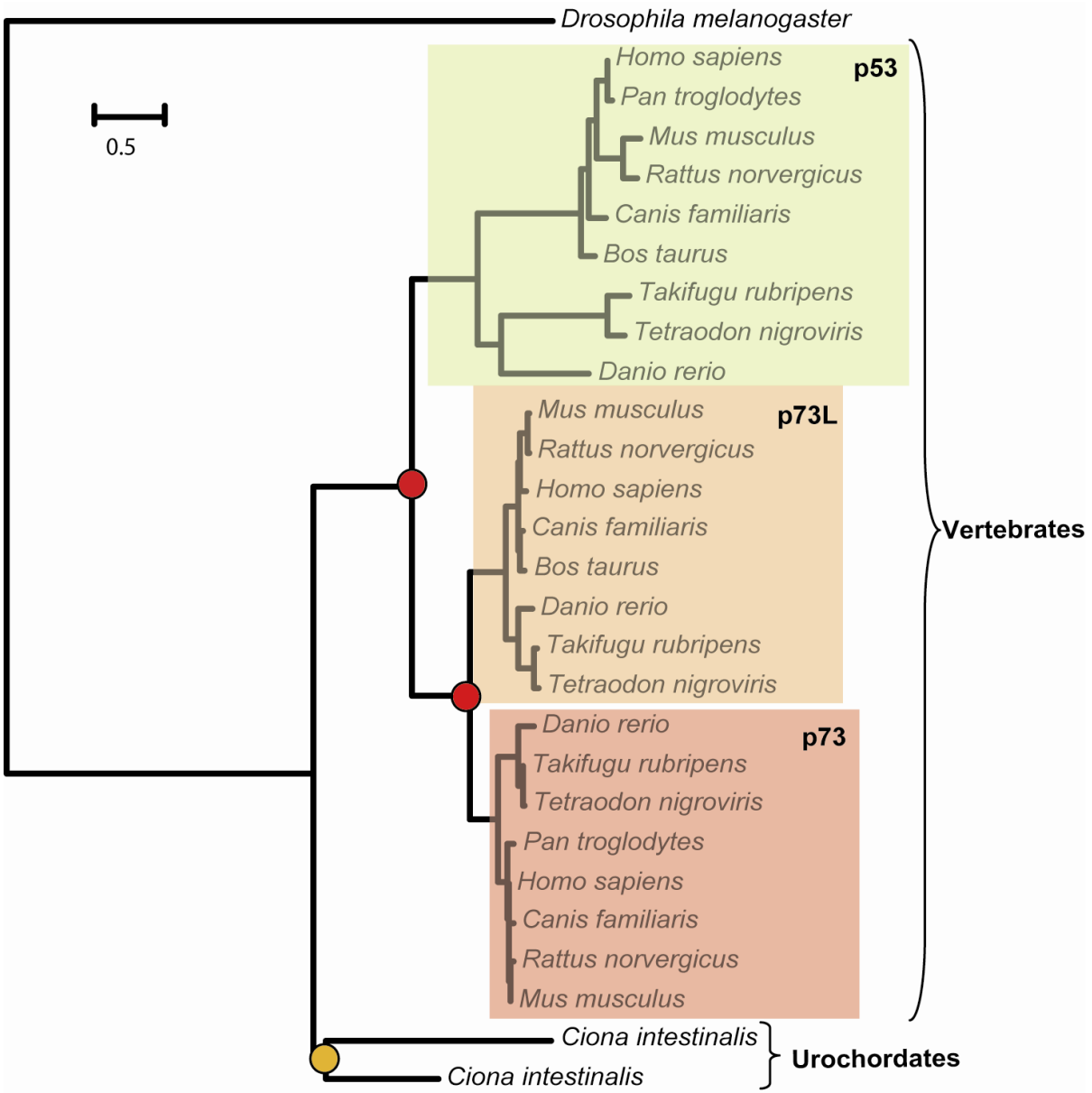
Similarity-based approaches (many more approaches):

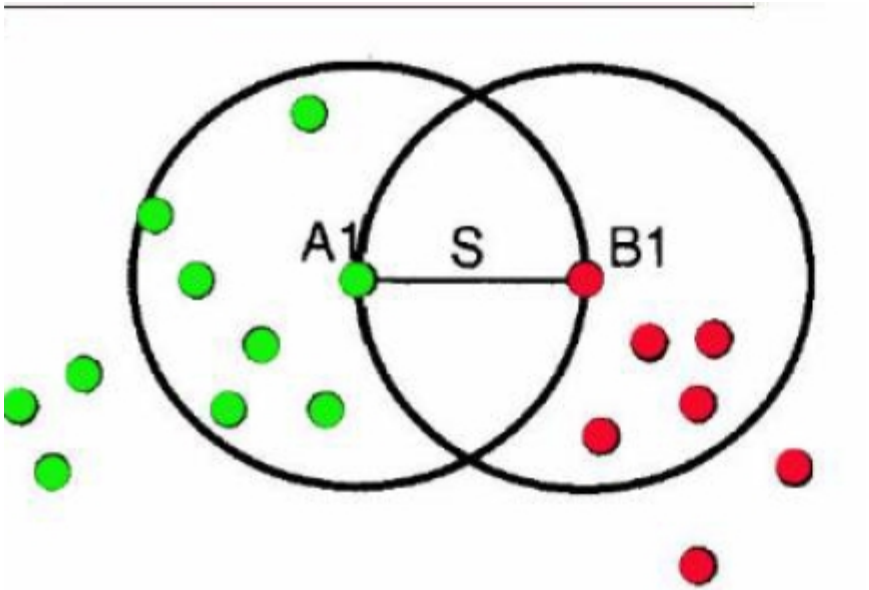
- **Best Reciprocal Hits**
- Detects all orthologies as one-to one. Highly affected by paralogy. Low rate of false positives but high rates of false negatives.

- The simplest and fastest method, still widely used

-

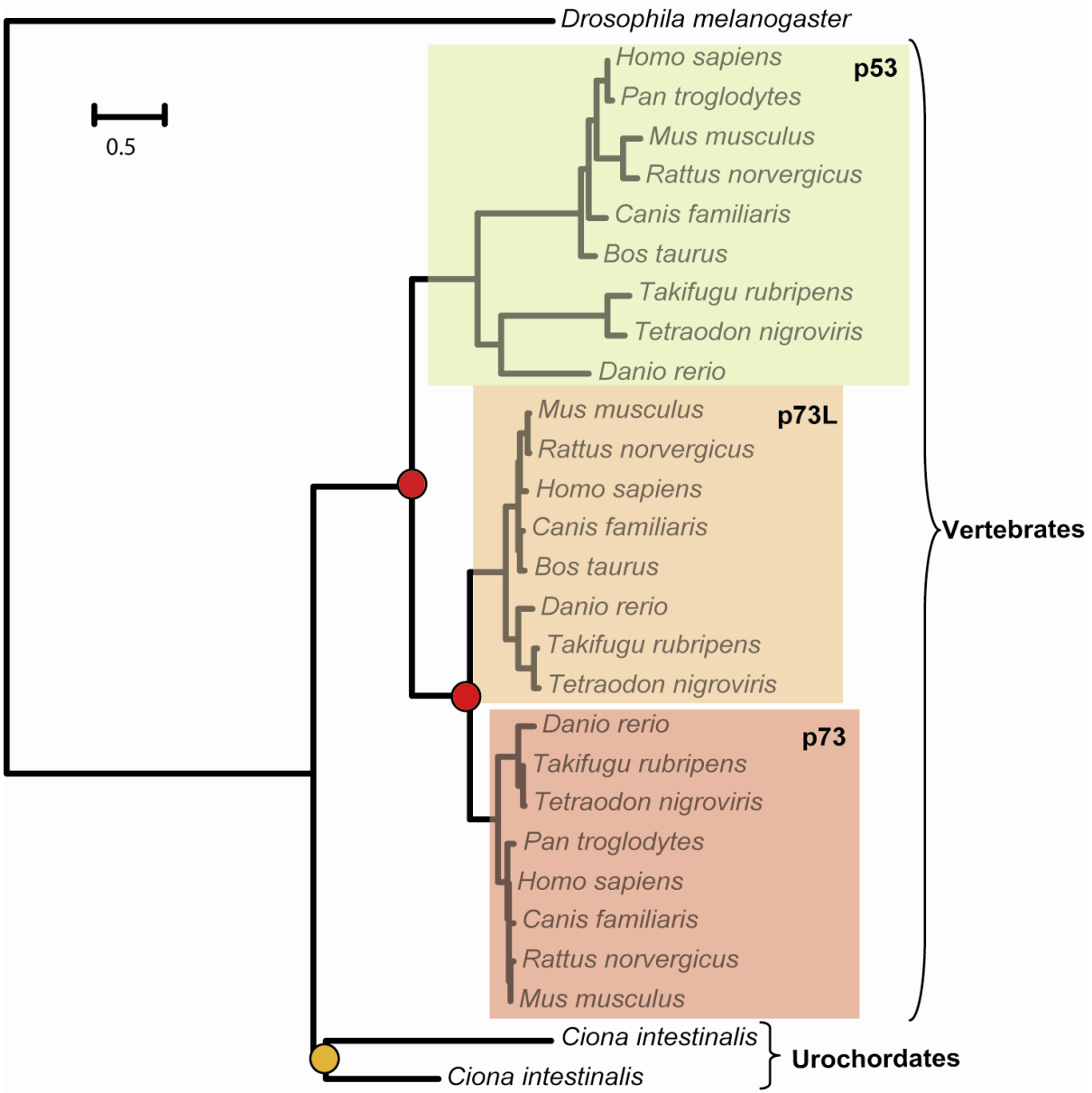



Drosophila melanogaster
0.5
Homo sapiens
Pan troglodytes
Mus musculus
Rattus norvergicus
Canis familiaris
Bos taurus
Takifugu rubripens
Tetraodon nigroviris
Danio rerio
p53
Mus musculus
Rattus norvergicus
Homo sapiens
Canis familiaris
Bos taurus
Danio rerio
Takifugu rubripens
Tetraodon nigroviris
p73L
Danio rerio
Takifugu rubripens
Tetraodon nigroviris
Pan troglodytes
Homo sapiens
Canis familiaris
Rattus norvergicus
Mus musculus
p73
Vertebrates
Ciona intestinalis
Ciona intestinalis
Urochordates



InParanoid

In-Paranoid.
Improved BRH to detect in-paralogs as well. Works well at the pairwise level. (multi-paranoid for multi-species comparisons


Drosophila melanogaster
0.5
Homo sapiens
Pan troglodytes
Mus musculus
Rattus norvergicus
Canis familiaris
Bos taurus
Takifugu rubripens
Tetraodon nigroviris
Danio rerio
p53
Mus musculus
Rattus norvergicus
Homo sapiens
Canis familiaris
Bos taurus
Danio rerio
Takifugu rubripens
Tetraodon nigroviris
p73L
Danio rerio
Takifugu rubripens
Tetraodon nigroviris
Pan troglodytes
Homo sapiens
Canis familiaris
Rattus norvergicus
Mus musculus
p73
Vertebrates
Ciona intestinalis
Ciona intestinalis
Urochordates

Note:

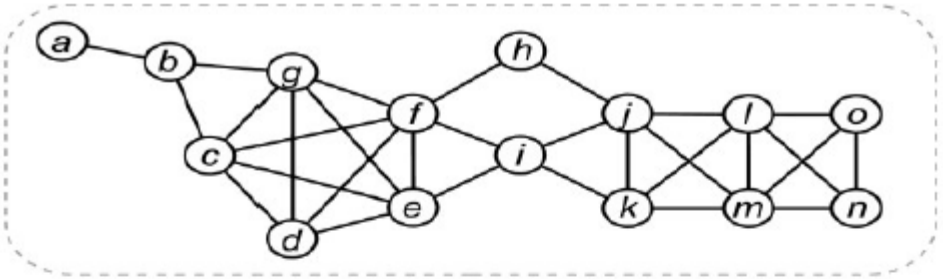
Definition of **in-** and **out-paralogues** require the specification of a given **speciation-node** of reference

COG-like
(used by many DBs like STRING)

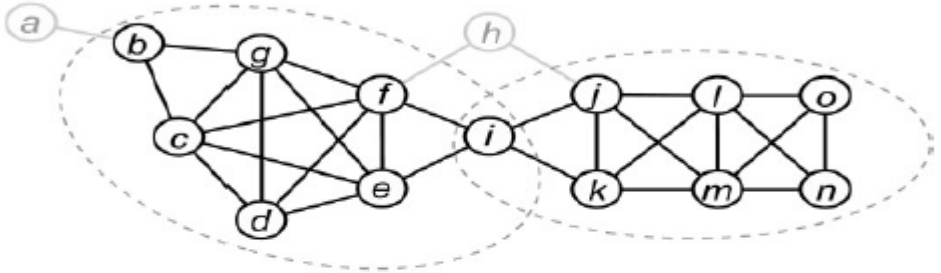
Exploits multi-species information.
Predicts clusters of orthologous groups (in-paralogs) not all pairs in a cluster are paralogs.

Can be used at different stringent levels

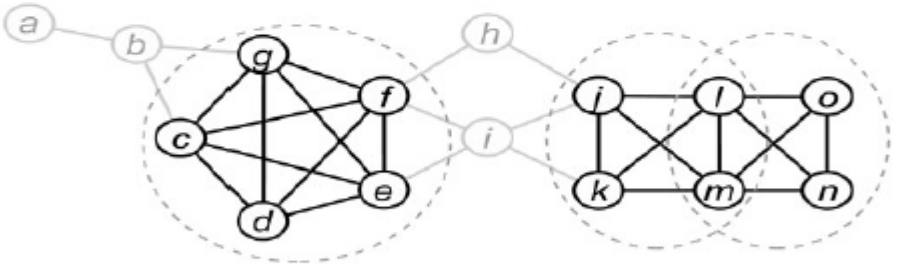
2



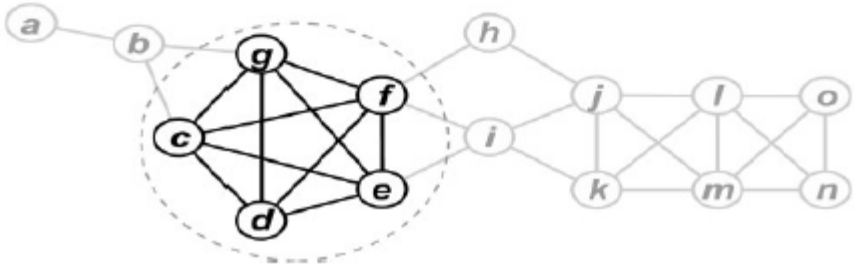
3



4



5



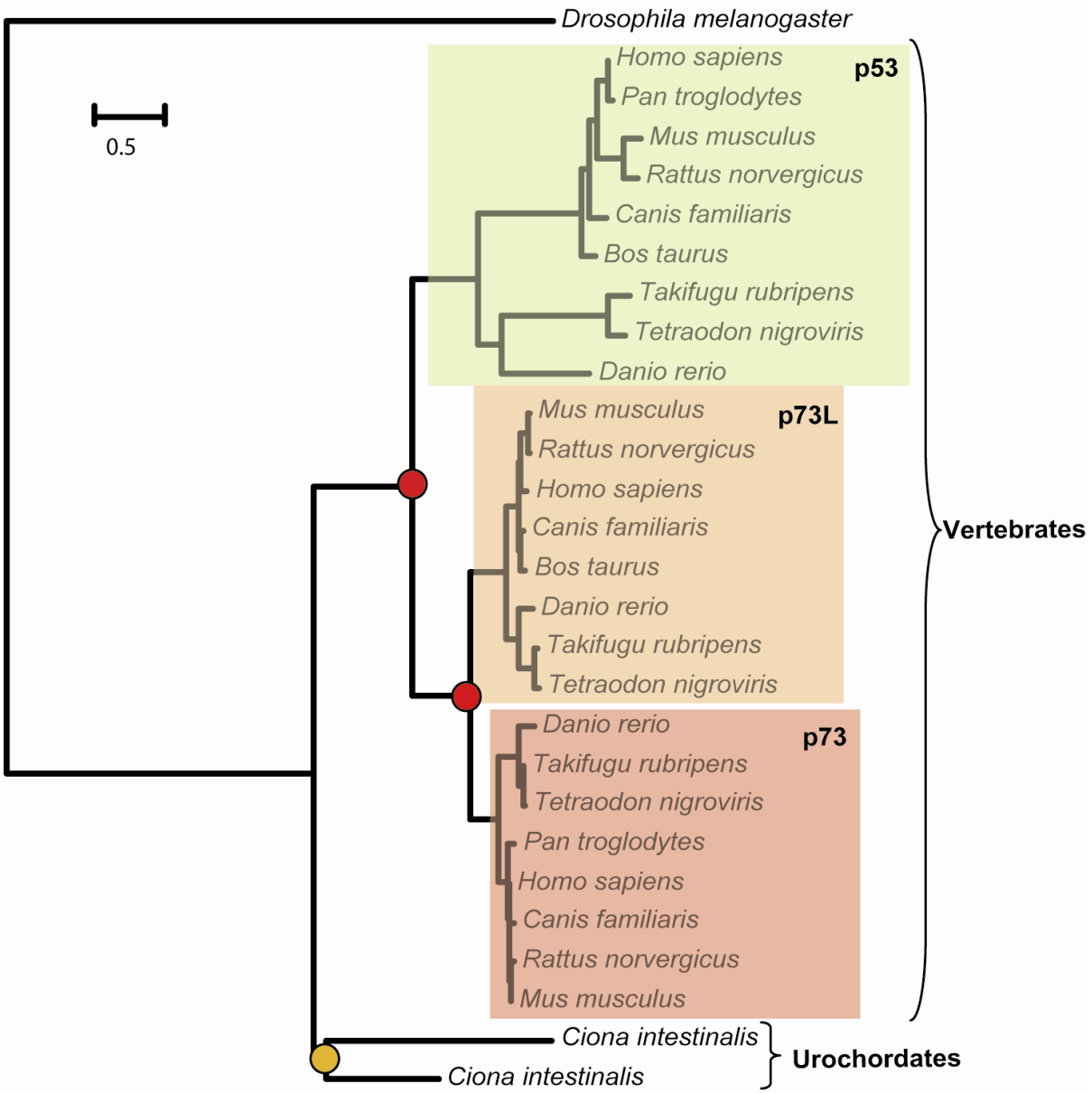
Clustering methods produce: orthologous groups

Equivalent to the earlier concept of sub-family

Orthologous groups = Group of sequences derived from a single gene in a common ancestor. They may include orthologs and in-paralogues.

Each orthologous group has implicit the specification of an ancestral species of reference (a speciation node).

How many orthologous groups? 3 at the level of vertebrates, 1 at the level of chordates



The definition of a reference ancestral species is just an approximation to the inherently hierarchical nature of gene family evolution: and is thus incomplete.

To alleviate this, many databases define orthologous groups at various hierarchical levels (e.g Metazoa, Vertebrates, Mammals, Primates)

Methods based on phylogeny where not used at a large scale due to limitations in computational power (phylogenetics is costly).

However, these has changed recently, fast pipelines and algorithms are available:

Ensembl trees, PhylomeDB, TreeFam, etc..

Review

Large-scale assignment of orthology: back to phylogenetics?

Toni Gabaldón

Bioinformatics and Genomics Program, Center for Genomic Regulation, Doctor Aiguader, 88, 08003 Barcelona, Spain.
Email: tgabaldon@crg.es

Published: 30 October 2008

Genome ***Biology*** 2008, **9**:235 (doi:10.1186/gb-2008-9-10-235)

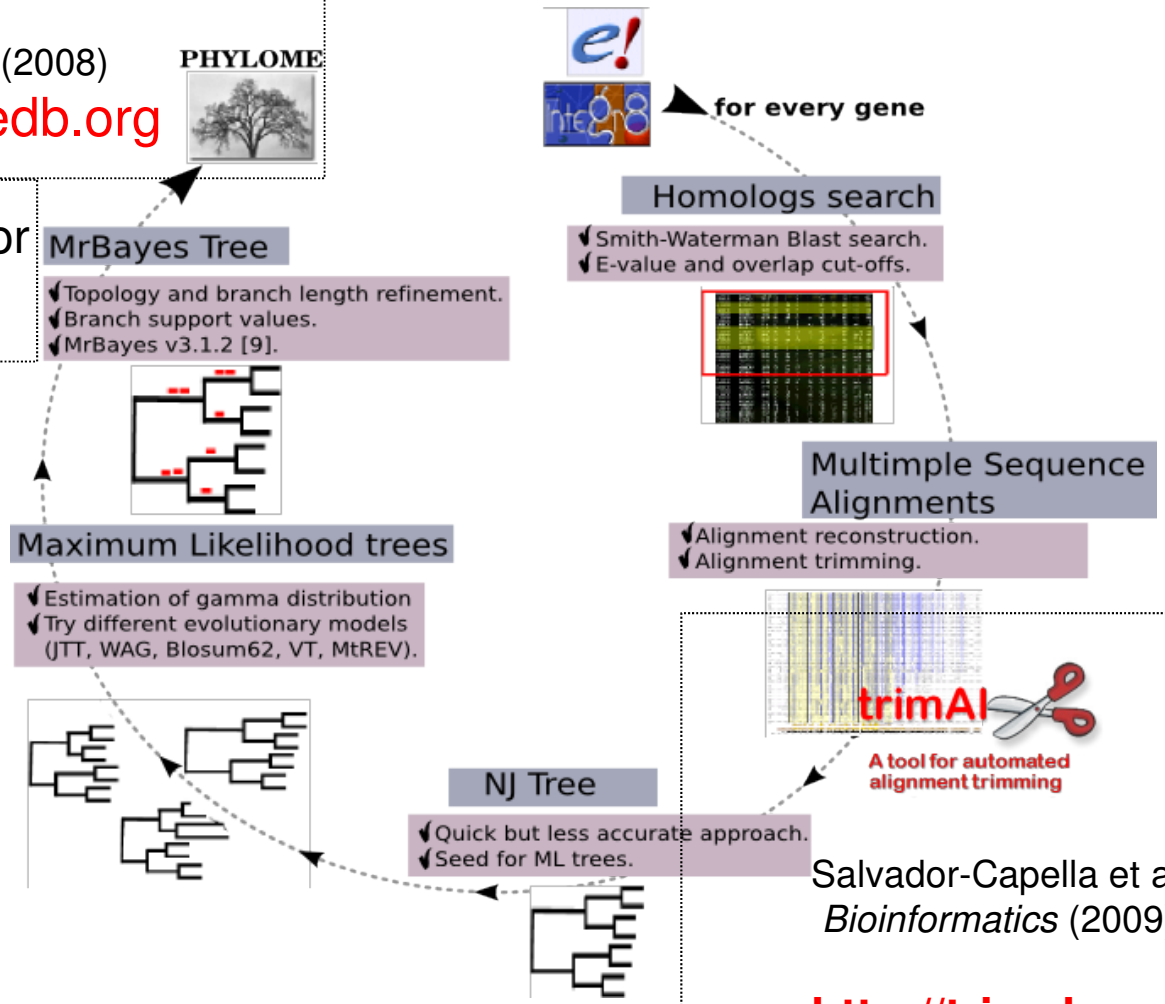
Abstract

Reliable orthology prediction is central to comparative genomics. Although orthology is defined by phylogenetic criteria, most automated prediction methods are based on pairwise sequence comparisons. Recently, automated phylogeny-based orthology prediction has emerged as a feasible alternative for genome-wide studies.

Our pipeline:

Huerta-Cepas et al.
Nucleic Acids Res. (2008)
www.phylomedb.org

ETE: Environment for Tree Exploration
ete.cgenomics.org



Salvador-Capella et al
Bioinformatics (2009).

http://trimal.cgenomics.org

Pipeline described in Huerta-Cepas et al Genome Biology (2007)

Phylogeny-based methods

- General procedure: reconstruct the evolution of a gene family (phylogenetics), detect duplication and speciation nodes and predict orthology and paralogy accordingly.
- Two main methods for predicting duplication and speciation nodes from a tree:

→ Species tree reconciliation (RIO, Ensembl)

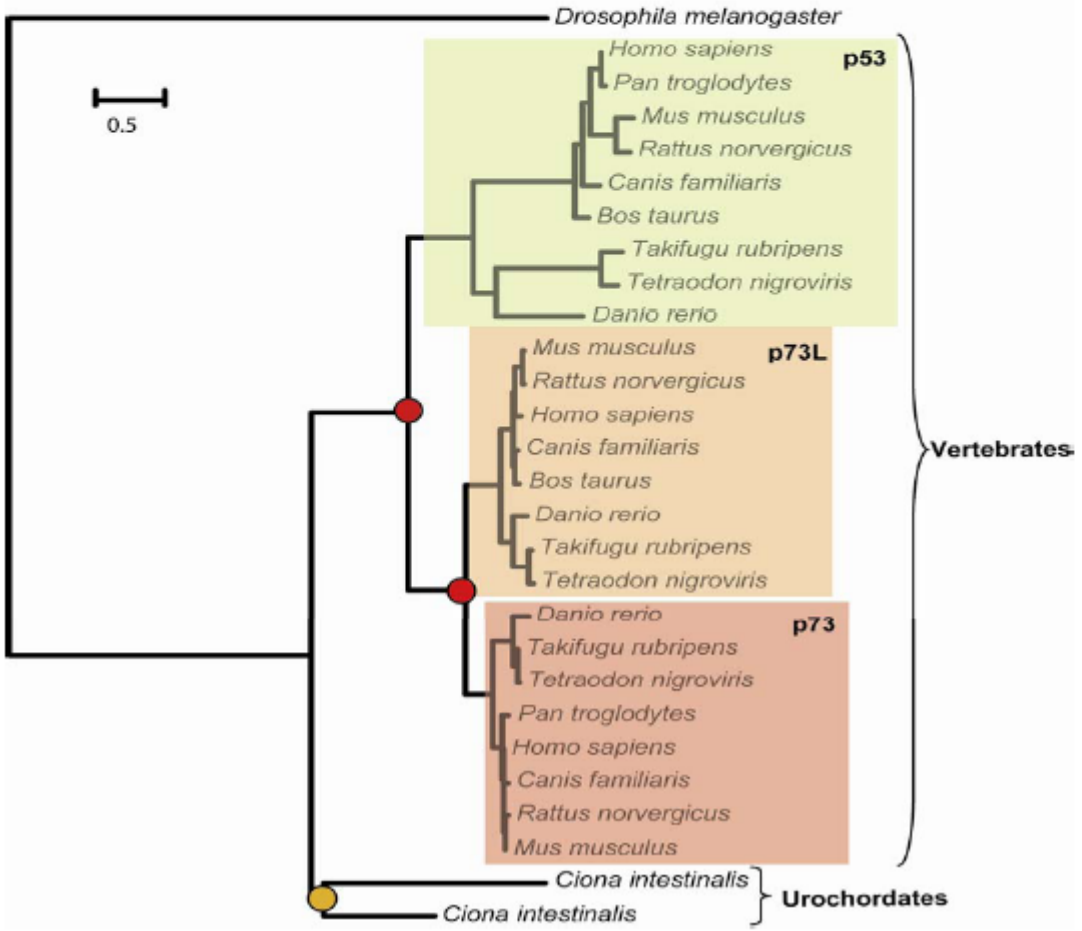
→ Species-overlap algorithms

Reconciliation with the species tree readily provides you information on speciation and duplication nodes in a tree

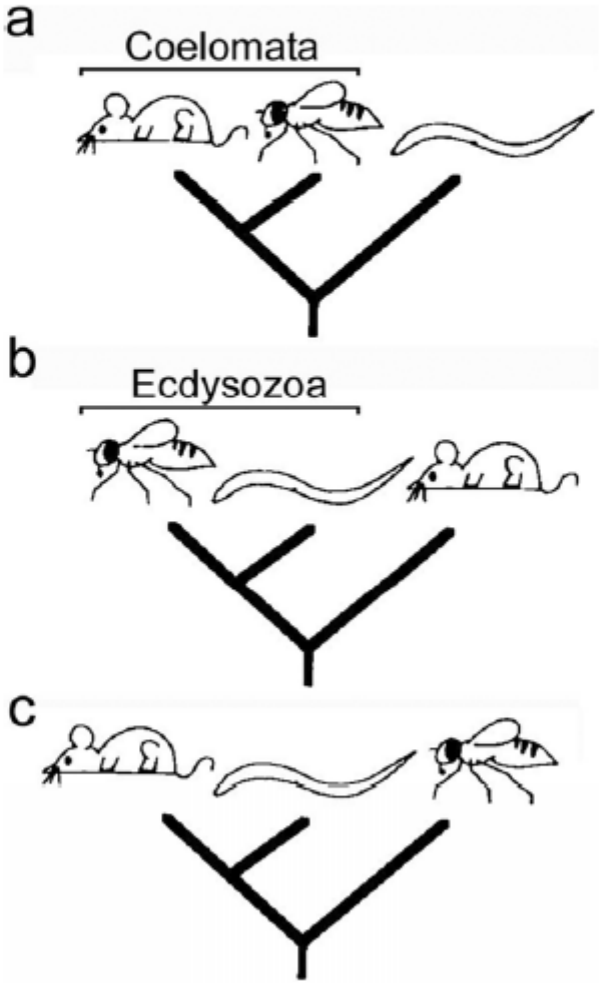
It works when these two assumptions are correct:

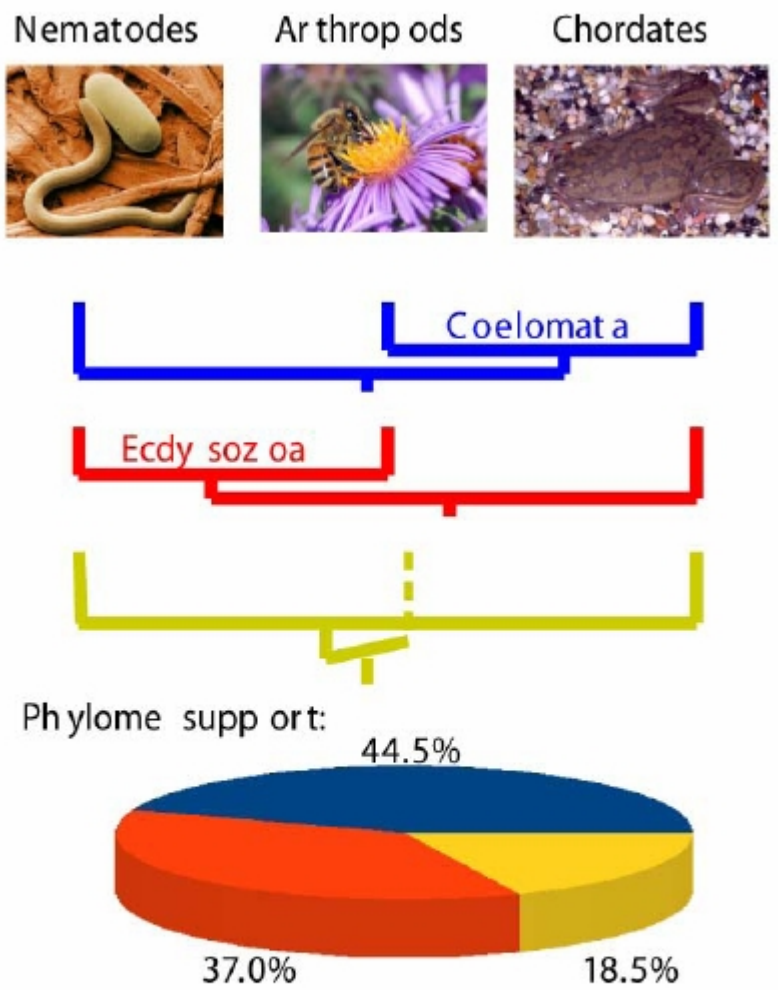
A) We know the true species tree

B) The gene tree is correct and reflects the species evolution



Uncertainty in species trees and topological variability in gene trees





What percentage of gene trees from the human phylome support each topology?

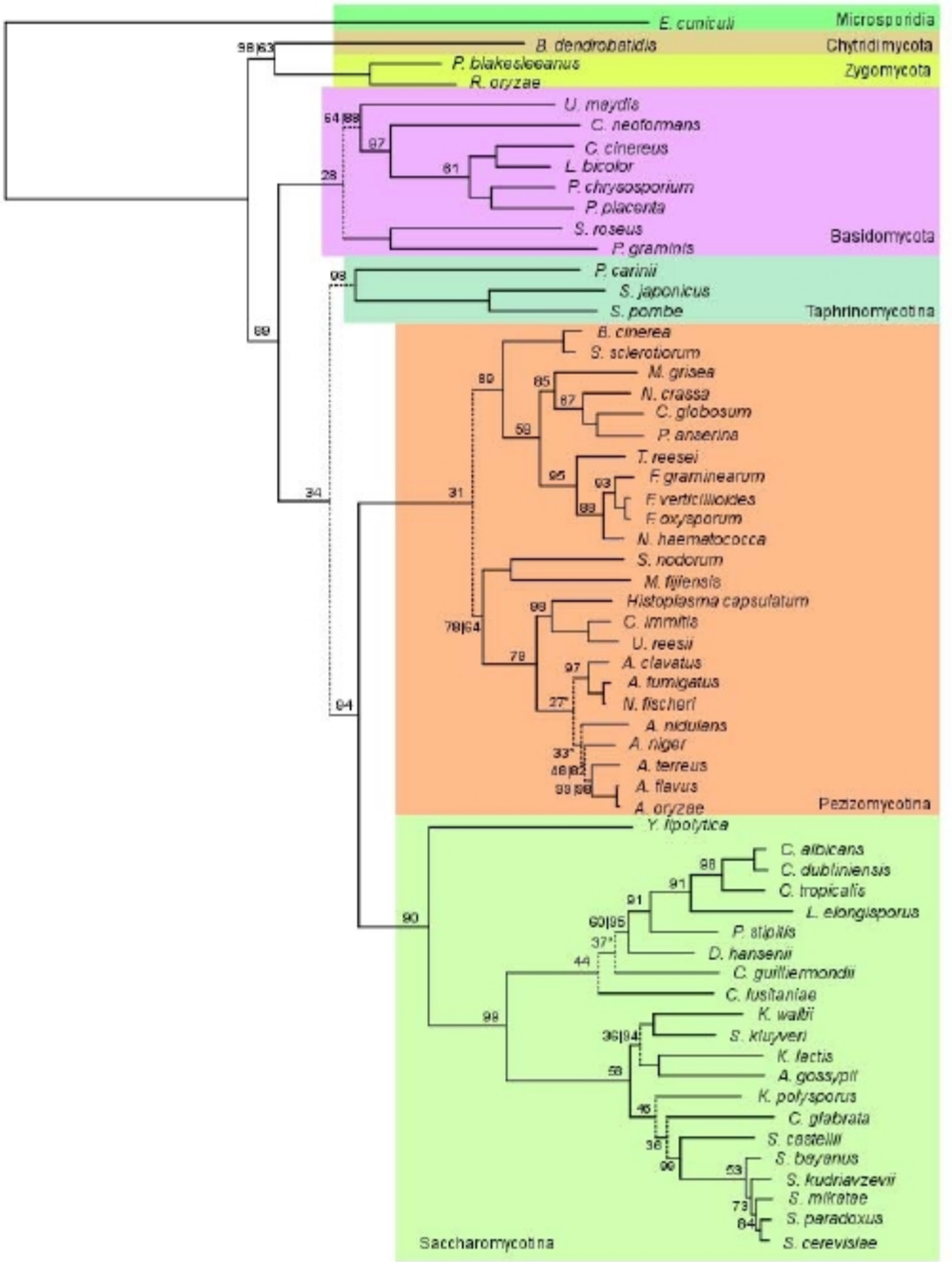
Similar results for

Primates
Rodents
laurasatheria

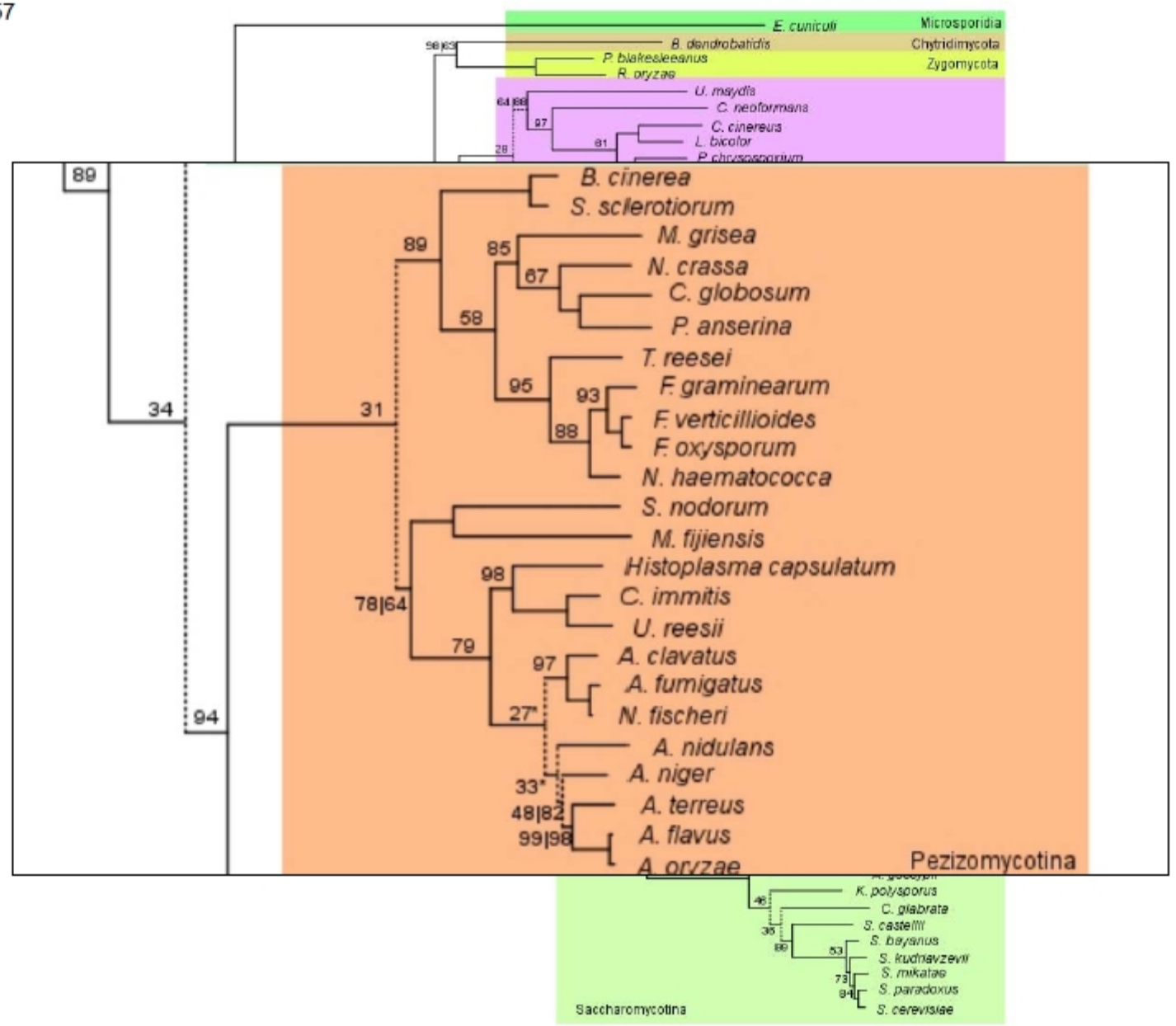
The tree vs the forest:

Comparison of a fungal species tree with the topological variability of the fungal phylome

Marcet-Houben M and Gabaldón T, 2009
PLoS ONE 4(2): e4357



Marcet-Houben M, Gabaldón T, 2009
PLoS ONE 4(2): e4357

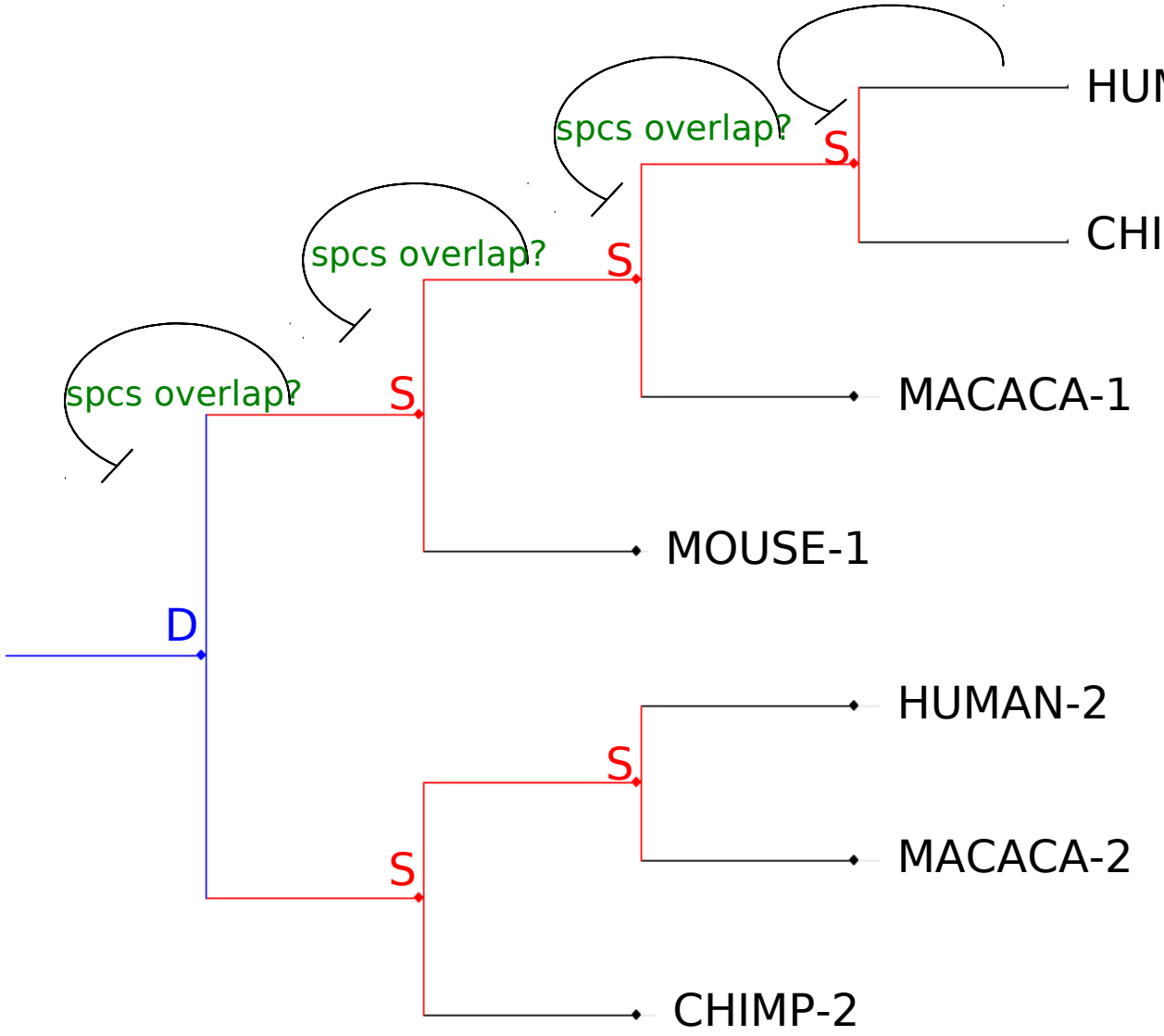


This large-degree of topological variability might be in part due to phylogenetic artifacts, insuficient phylogenetic signal, etc. But also to real evolutionary processes that render a gene tree different from a species tree: lineage sorting, gene conversion, etc

In any case: strict interpretation of gene and species trees will result in many incorrect predictions

To deal with topological variability we implemented a species-overlap algorithm

(described in Huerta-Cepas et al. (2007) The human phylome. Genome Biology)



Our algorithm

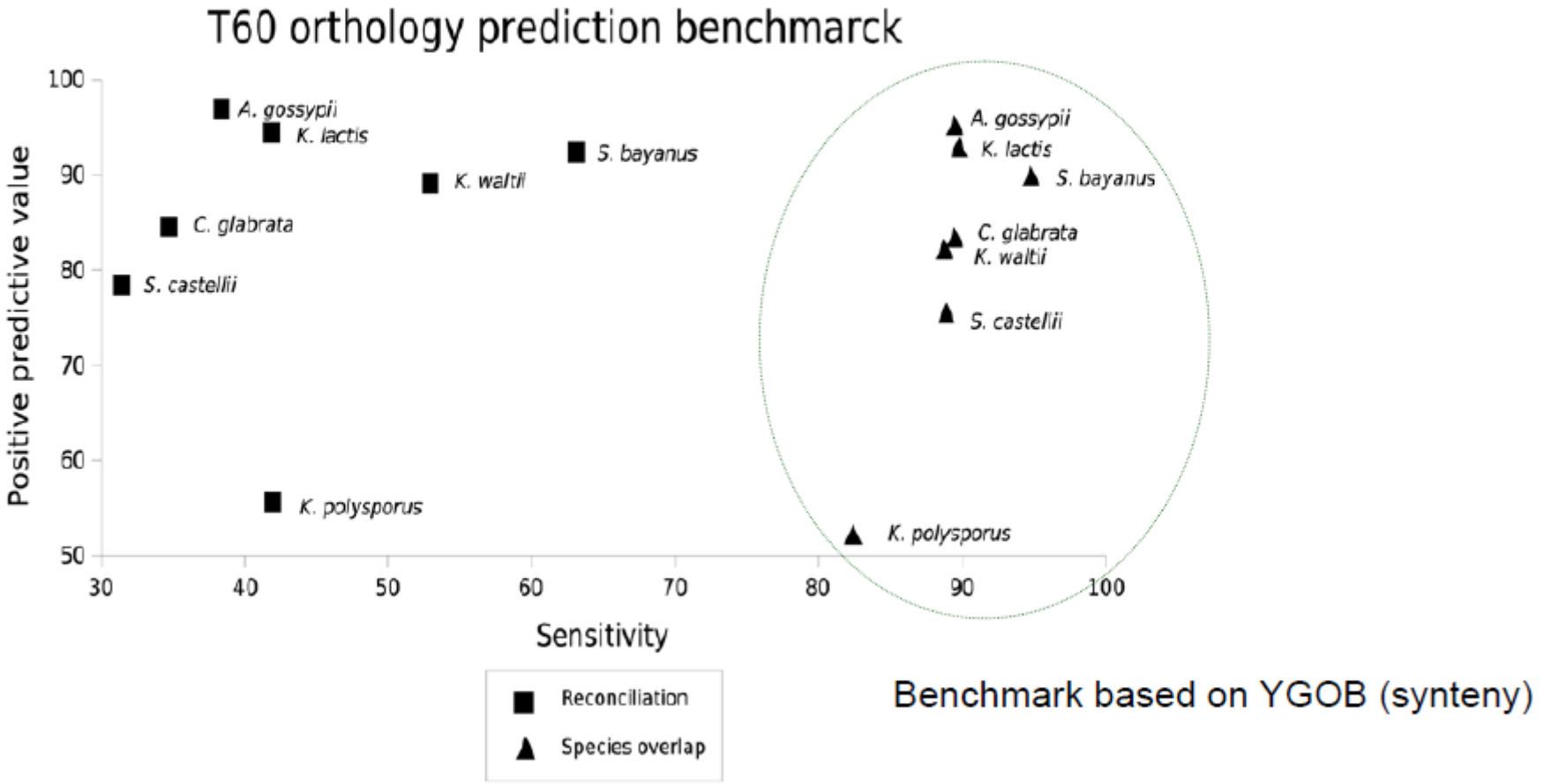
- We calculate a **species overlap score** for every node.

Species common to both partitions / sum of the species in both partitions

- We only need a rough species tree to set an outgroup.

The species-overlap algorithm (**PhylomeDB**) is highly accurate and less affected by gene tree/ species tree artifacts than tree-reconciliation

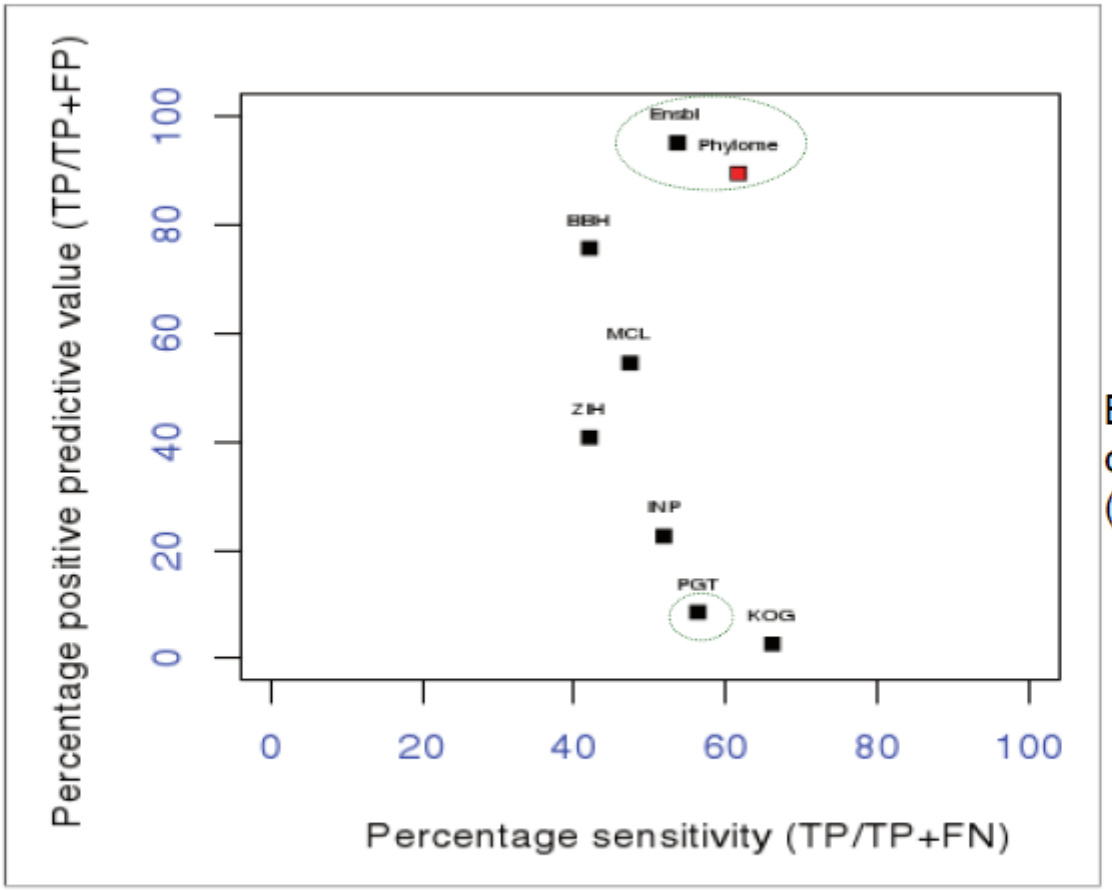
Tree reconciliation / species overlap
Marcet-Houben and Gabaldón. *PLoS ONE* (2009)



Benchmark based on YGOB (synteny)

Figure 2. Comparison of different orthology inference algorithms. The synteny based and manually curated orthology predictions available at YGOB database [18] is taken as a golden set to compute the number of true positives (TP), false positives (FP) and false negatives (FN) yielded by each method. For each method, the sensitivity S = TP/(TP+FN) and the positive predictive value P = TP/(TP+FP) are computed.
doi:10.1371/journal.pone.0004357.g002

The species-overlap algorithm (**PhylomeDB**) is highly accurate and less affected by gene tree/ species tree artifacts
than tree-reconciliation



Benchmark based on curated dataset (Hulsen et al.)

Blast based / phylogeny-based

Huerta-Cepas et al. *Genome Biology* (2007)

www.phylomedb.org

user:
pass
login

HOME BROWSE PHYLOMES DOWNLOADS FAQ HELP ABOUT

YBL058W Search! Blast Search! Select phylomes

Welcome to PhylomeDB.

PhylomeDB is a public database for complete collections of gene phylogenies (phylomes). It allows users to **interactively explore the evolutionary history of genes** through the visualization of phylogenetic trees and multiple sequence alignments. Moreover, phylomeDB provides genome-wide orthology and paralogy predictions which are based on the analysis of the phylogenetic trees. The automated pipeline used to reconstruct trees **aims at providing a high-quality phylogenetic analysis of different genomes** , including Maximum Likelihood or Bayesian tree inference, alignment trimming and evolutionary model testing. PhylomeDB includes also a public **download section with the complete set of trees, alignments and orthology predictions.**

DB phylome

user:
pass
login

HOME BROWSE PHYLOMES DOWNLOADS FAQ HELP ABOUT

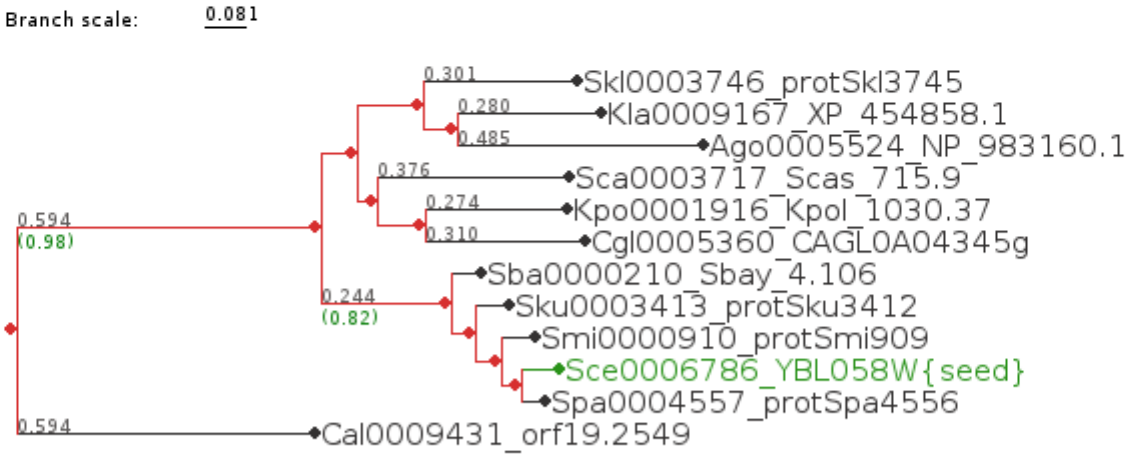
YBL058W | Search! | Blast Search! | Select phylomes

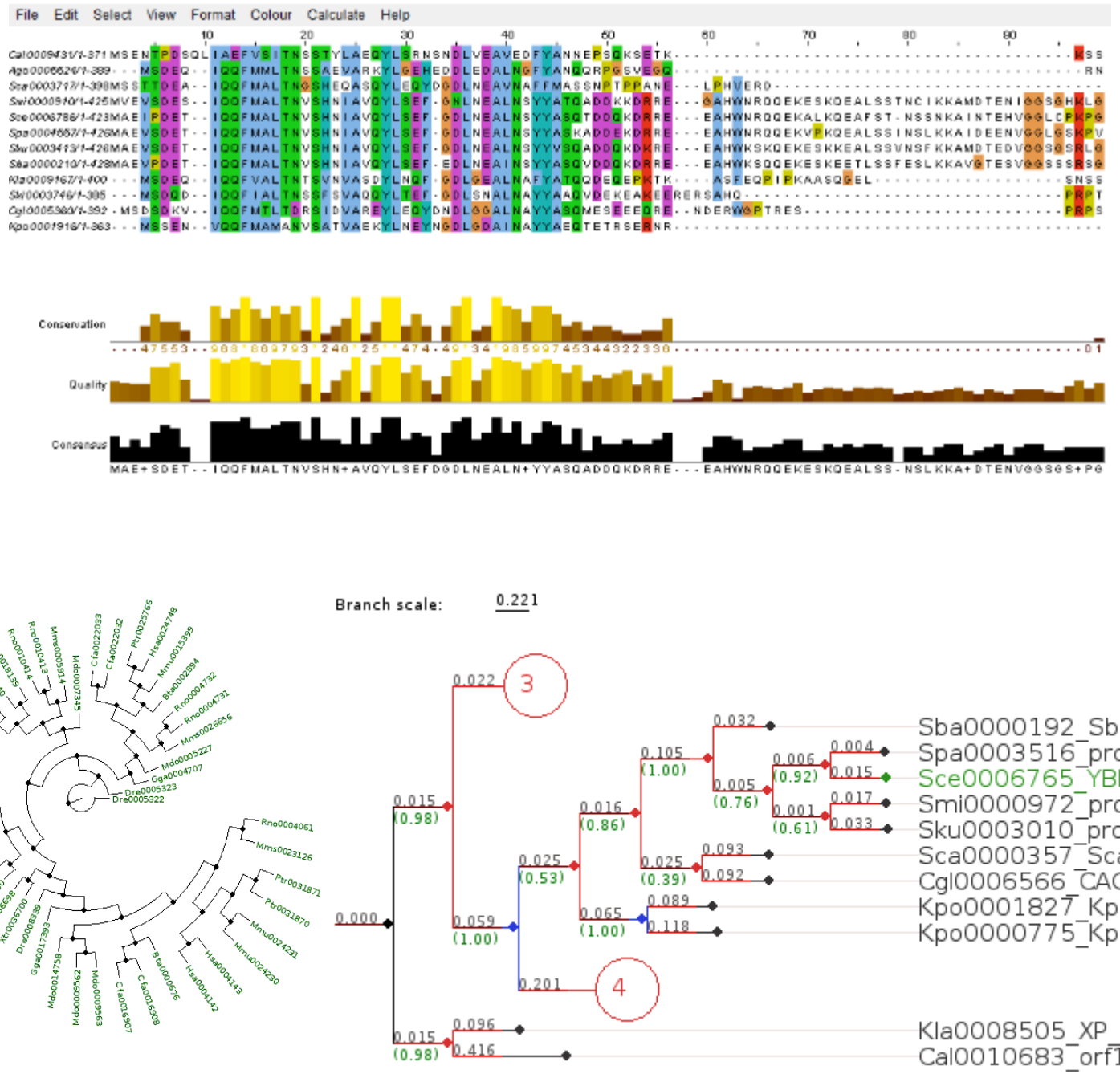
- YBL058W
  - Info
  - Orthologs
  - Seed Trees (4)
    - SceP12a
    - SceP21
    - SceP12b
    - SceP60
  - Collateral Trees (4)
    - Hsa0028724
    - Hsa0028192
    - Hsa0018651
    - Hsa0016629

Tree: Sce0006786 (in phylome SceP12a)

Tree model: WAG (lk=-5156.930000) | -- Download -- | -- Visualize (requires Java)--

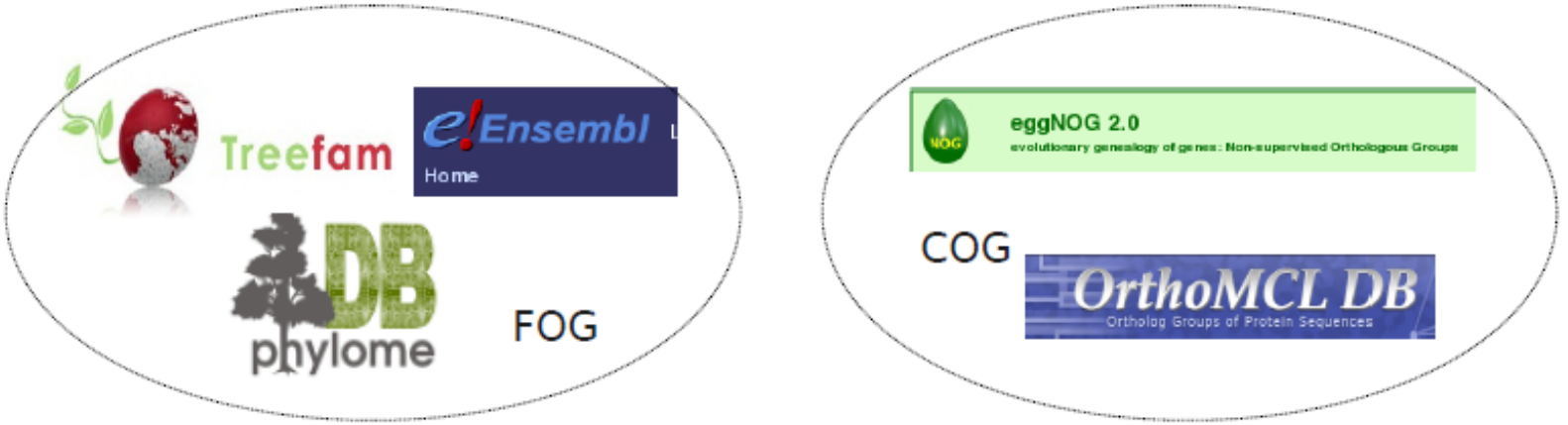
Tree Tools and Actions: Colapse tool | Search in tree | Clear




File Edit Select View Format Colour Calculate Help
Conservation
Quality
Consensus
Branch scale: 0.221
Sba0000192_Sbay_4.88
Spa0003516_protSpa3515
Sce0006765_YBL039C{seed}
Smi0000972_protSmi971
Sku0003010_protSku3009
Sca0000357_Scas_495.5
Cgl0006566_CAGL0F04433g
Kpo0001827_Kpol_1028.96
Kpo0000775_Kpol_1013.78
Kla0008505_XP_454196.1
Cal0010683_orf19.3941

MetaPhOrs

(Meta-Phylogeny-Based-Orthologs)



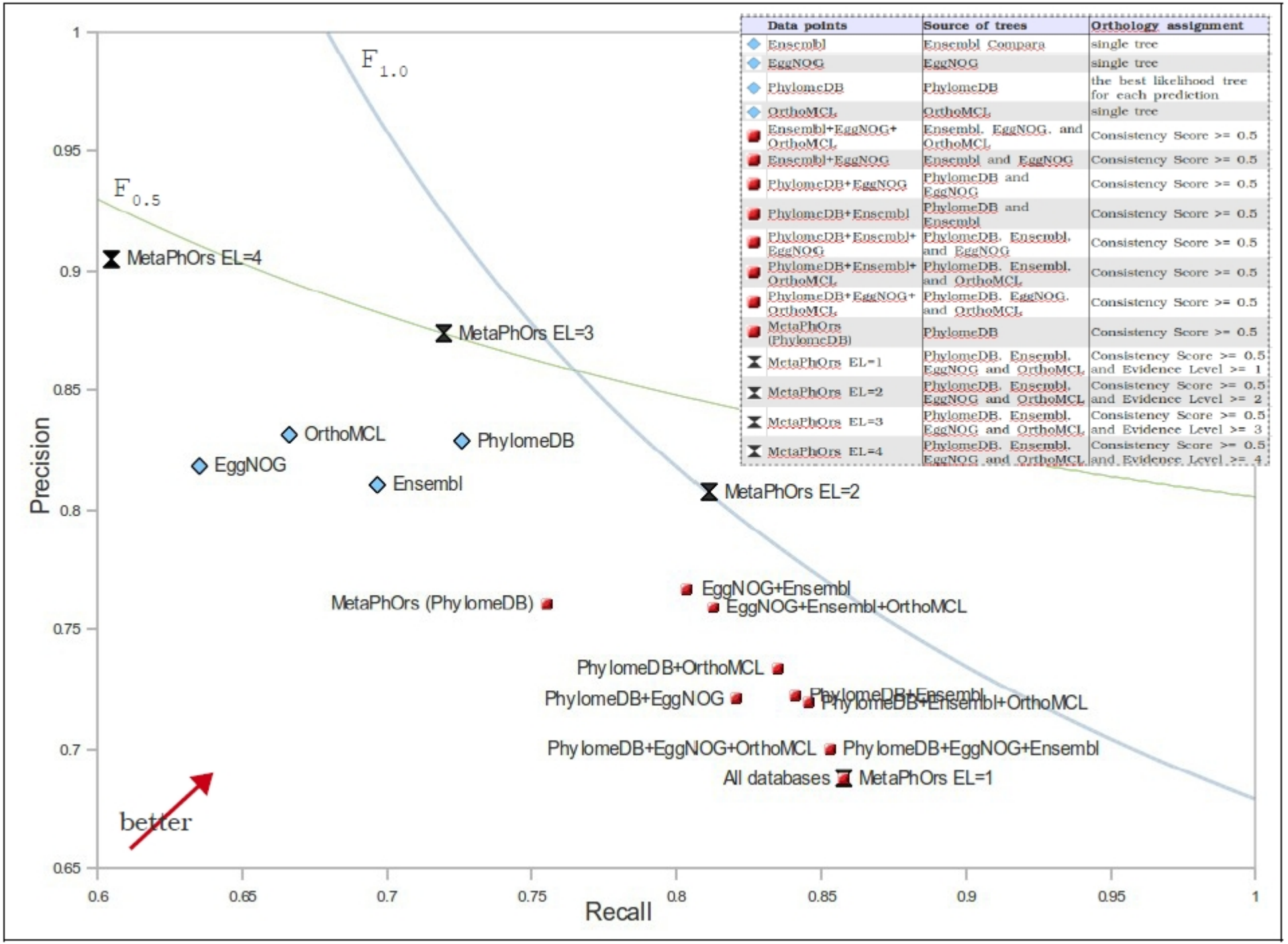
Use existing tree repositories

Reconstruct trees for orthologous groups

Integrate and use consistency across datasets as a proxy of reliability

result: phylogeny-based predictions across 800 genomes with a confidence score

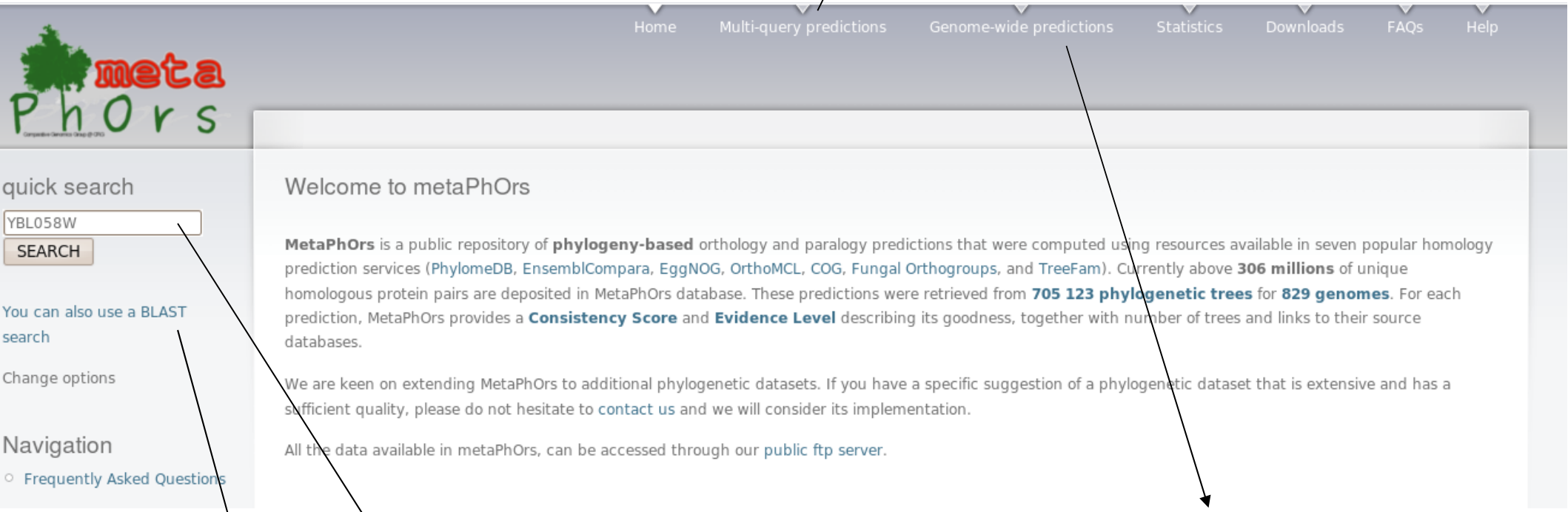
Pryszcz et. al. (NAR, 2011)



Data points	Source of trees	Orthology assignment
Ensembl	Ensembl Compara	single tree
EggNOG	EggNOG	single tree
PhylomeDB	PhylomeDB	the best likelihood tree for each prediction
OrthoMCL	OrthoMCL	single tree
Ensembl+EggNOG+OrthoMCL	Ensembl, EggNOG, and OrthoMCL	Consistency Score >= 0.5
Ensembl+EggNOG	Ensembl and EggNOG	Consistency Score >= 0.5
PhylomeDB+EggNOG	PhylomeDB and EggNOG	Consistency Score >= 0.5
PhylomeDB+Ensembl	PhylomeDB and Ensembl	Consistency Score >= 0.5
PhylomeDB+Ensembl+EggNOG	PhylomeDB, Ensembl, and EggNOG	Consistency Score >= 0.5
PhylomeDB+Ensembl+OrthoMCL	PhylomeDB, Ensembl, and OrthoMCL	Consistency Score >= 0.5
PhylomeDB+EggNOG+OrthoMCL	PhylomeDB, EggNOG, and OrthoMCL	Consistency Score >= 0.5
MetaPhOrs (PhylomeDB)	PhylomeDB	Consistency Score >= 0.5
MetaPhOrs EL=1	PhylomeDB, Ensembl, EggNOG and OrthoMCL	Consistency Score >= 0.5 and Evidence Level >= 1
MetaPhOrs EL=2	PhylomeDB, Ensembl, EggNOG and OrthoMCL	Consistency Score >= 0.5 and Evidence Level >= 2
MetaPhOrs EL=3	PhylomeDB, Ensembl, EggNOG and OrthoMCL	Consistency Score >= 0.5 and Evidence Level >= 3
MetaPhOrs EL=4	PhylomeDB, Ensembl, EggNOG and OrthoMCL	Consistency Score >= 0.5 and Evidence Level >= 4

http://orthology.phylomedb.org

Give me all orthologs for a list of IDs

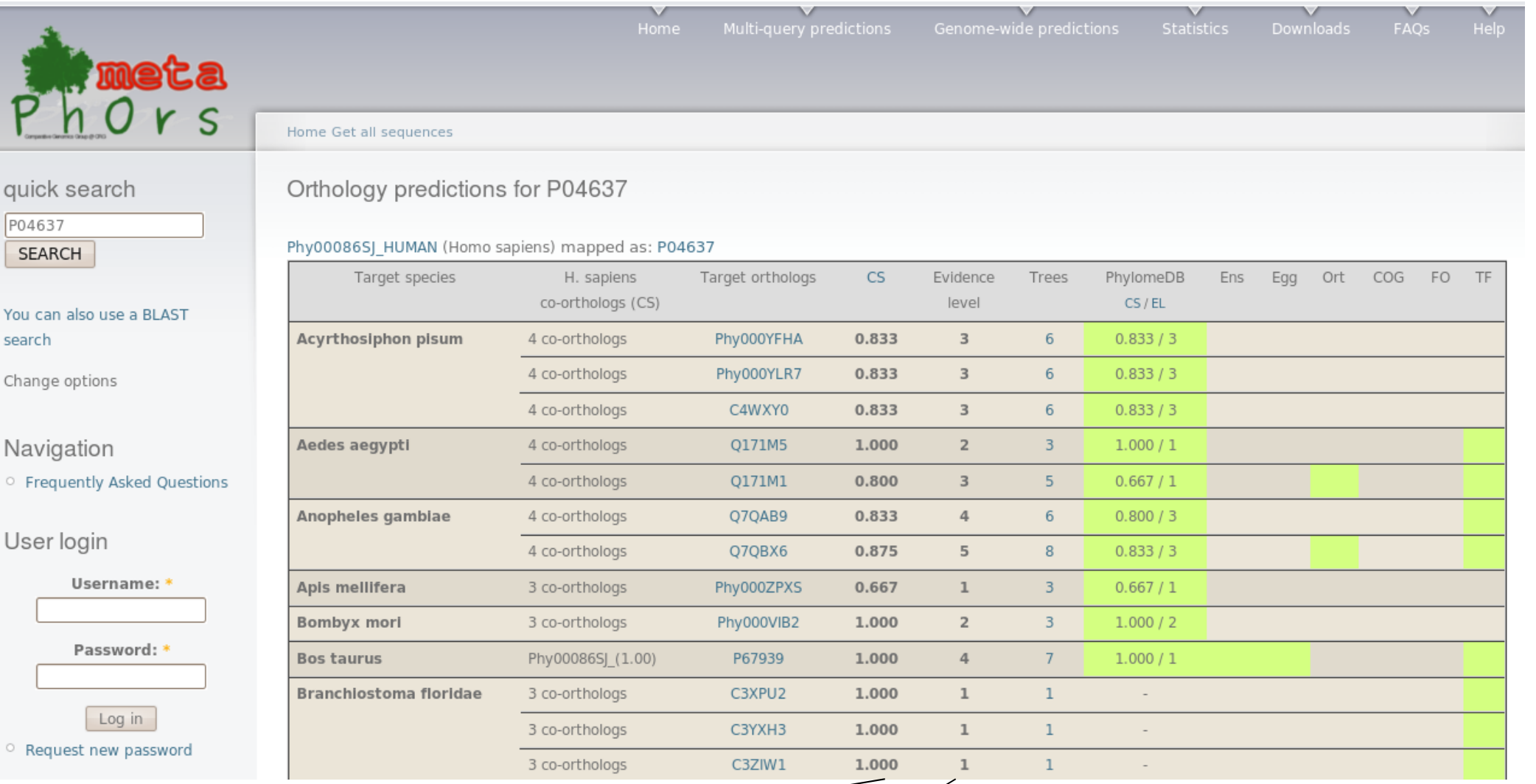


Give me all orthologs between Human and Mouse

Give me all orthologs to TP53

Blast my sequence and give me its orthologs

* Where it says **orthologs**, you can place **paralogs** instead!



Confidence score [0-1] = fraction of independent trees that support this association

Evidence level

Check the trees

“*Estoy enganchado al metaphors como un drogata al caballo--y hoy parece que tienen el servidor colgado--porfa diselo a quien se encargue porque necesito mirar cosas ahi.*”

Our best feedback ever.

(Received last week from a famous Immunologist.)

¿With over 30 orthology databases, based on various methods, which ones to choose?

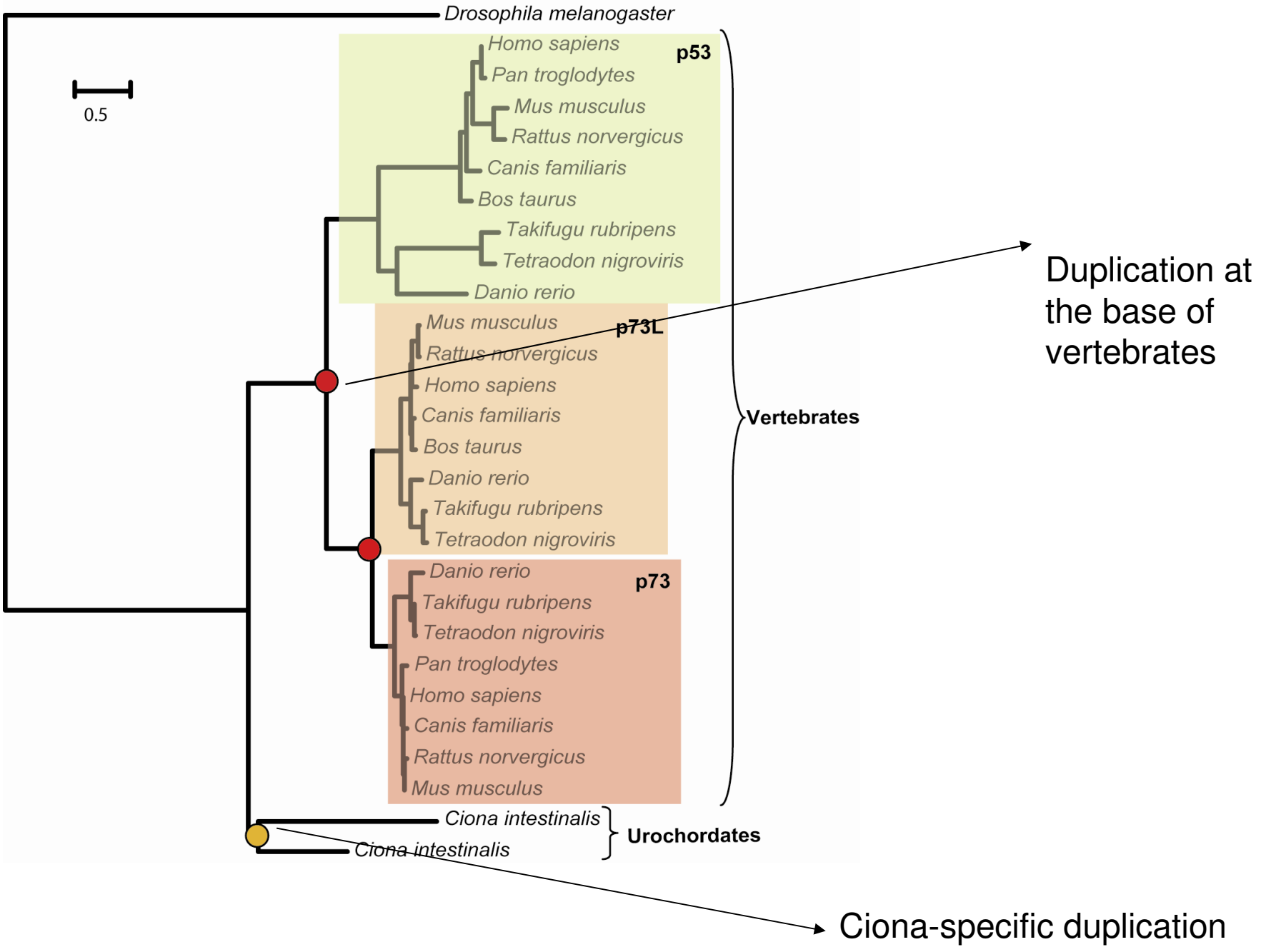
- Different taxonomic focuses
- Different methodologies
- Different outputs (pairwise relationships, groups, etc)
- Different interfaces
- Different accuracies (how to benchmark this?)

What about paralogy?

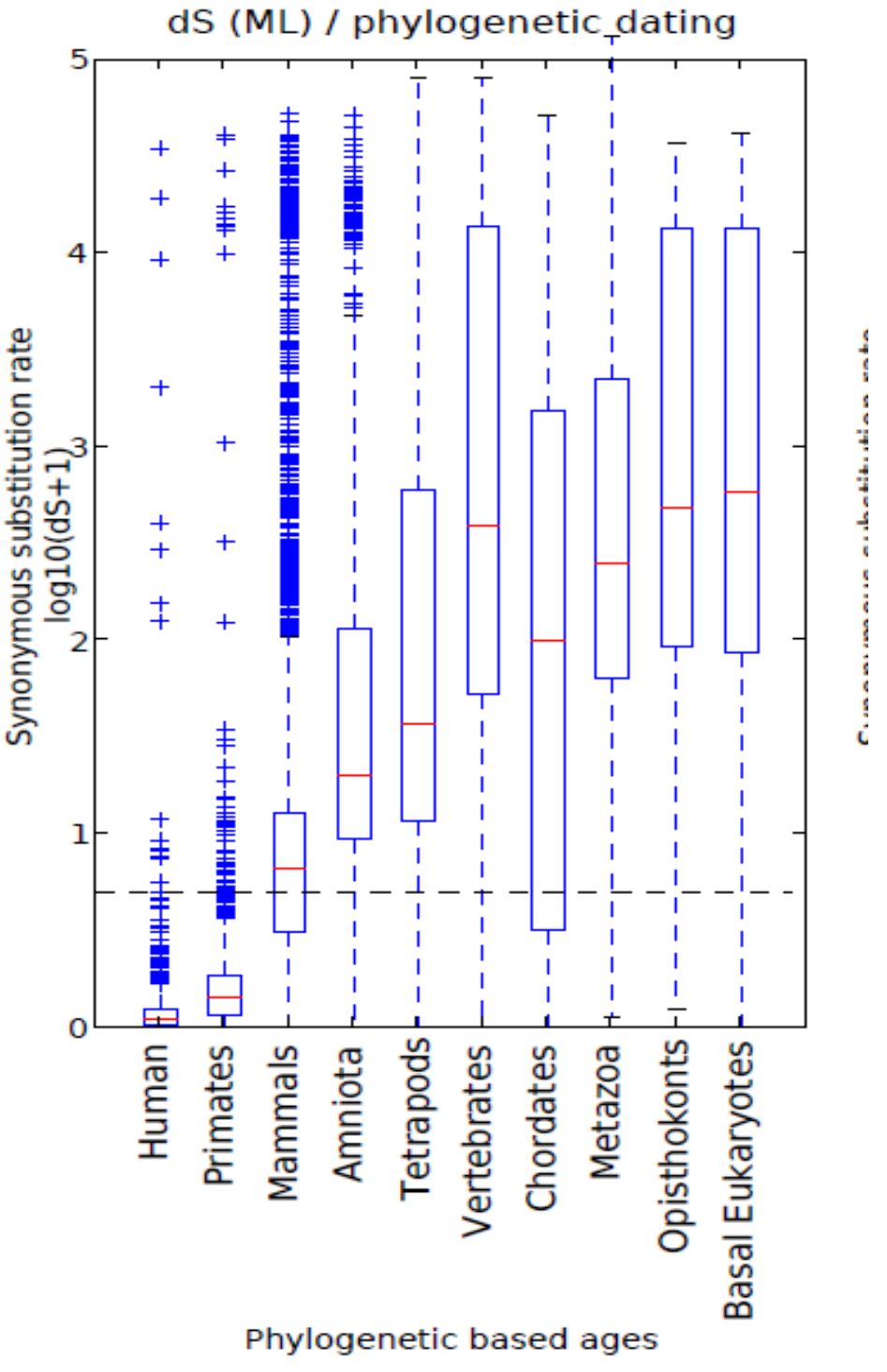
Most pairwise methods focus on orthologs, only in-paralogs are taken into account sometimes.

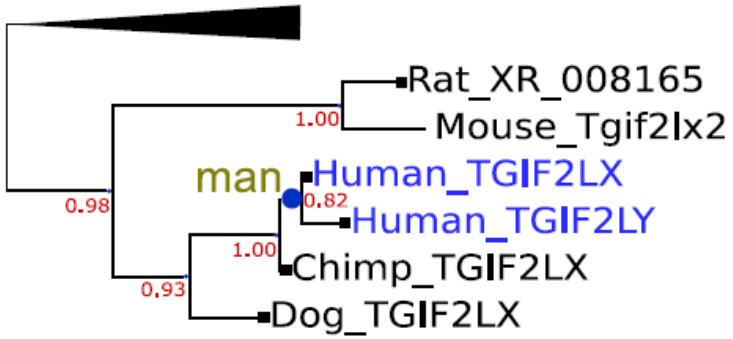
Phylogeny-based methods readily inform both on orthology and paralogy.

They also provide information on the possible date of the duplication (topological dating)


Drosophila melanogaster
0.5
Homo sapiens
Pan troglodytes
Mus musculus
Rattus norvergicus
Canis familiaris
Bos taurus
Takifugu rubripens
Tetraodon nigroviris
Danio rerio
p53
Mus musculus
Rattus norvergicus
Homo sapiens
Canis familiaris
Bos taurus
Danio rerio
Takifugu rubripens
Tetraodon nigroviris
p73L
Danio rerio
Takifugu rubripens
Tetraodon nigroviris
Pan troglodytes
Homo sapiens
Canis familiaris
Rattus norvergicus
Mus musculus
p73
Vertebrates
Ciona intestinalis
Ciona intestinalis
Urochordates
Duplication at the base of vertebrates
Ciona-specific duplication

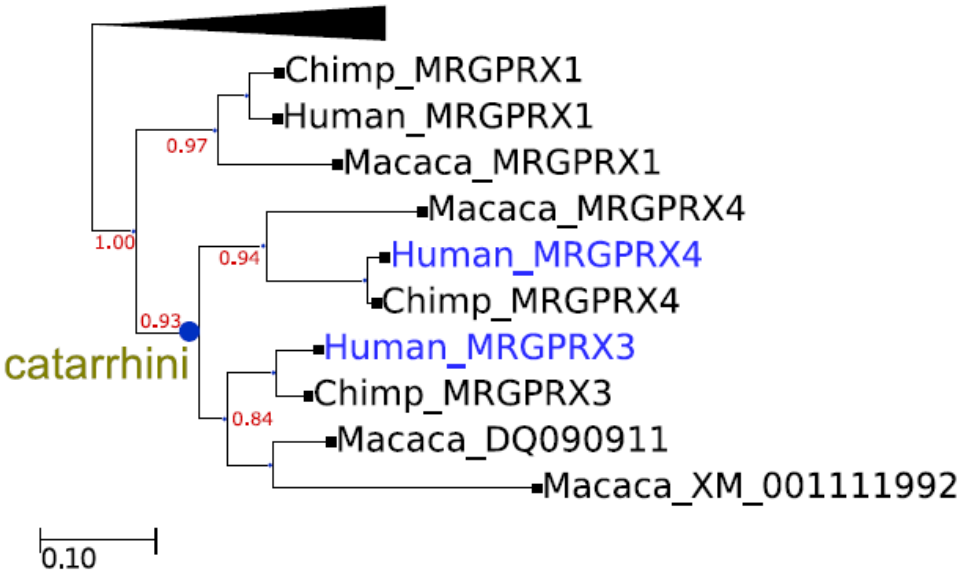
Comparison of topological dating vs dS





dS (ML): 0.24
dS (YN): 0.20
dS (NG): 0.19

C)



dS (ML): 0.09
dS (YN): 0.09
dS (NG): 0.10

D)