

Orthology

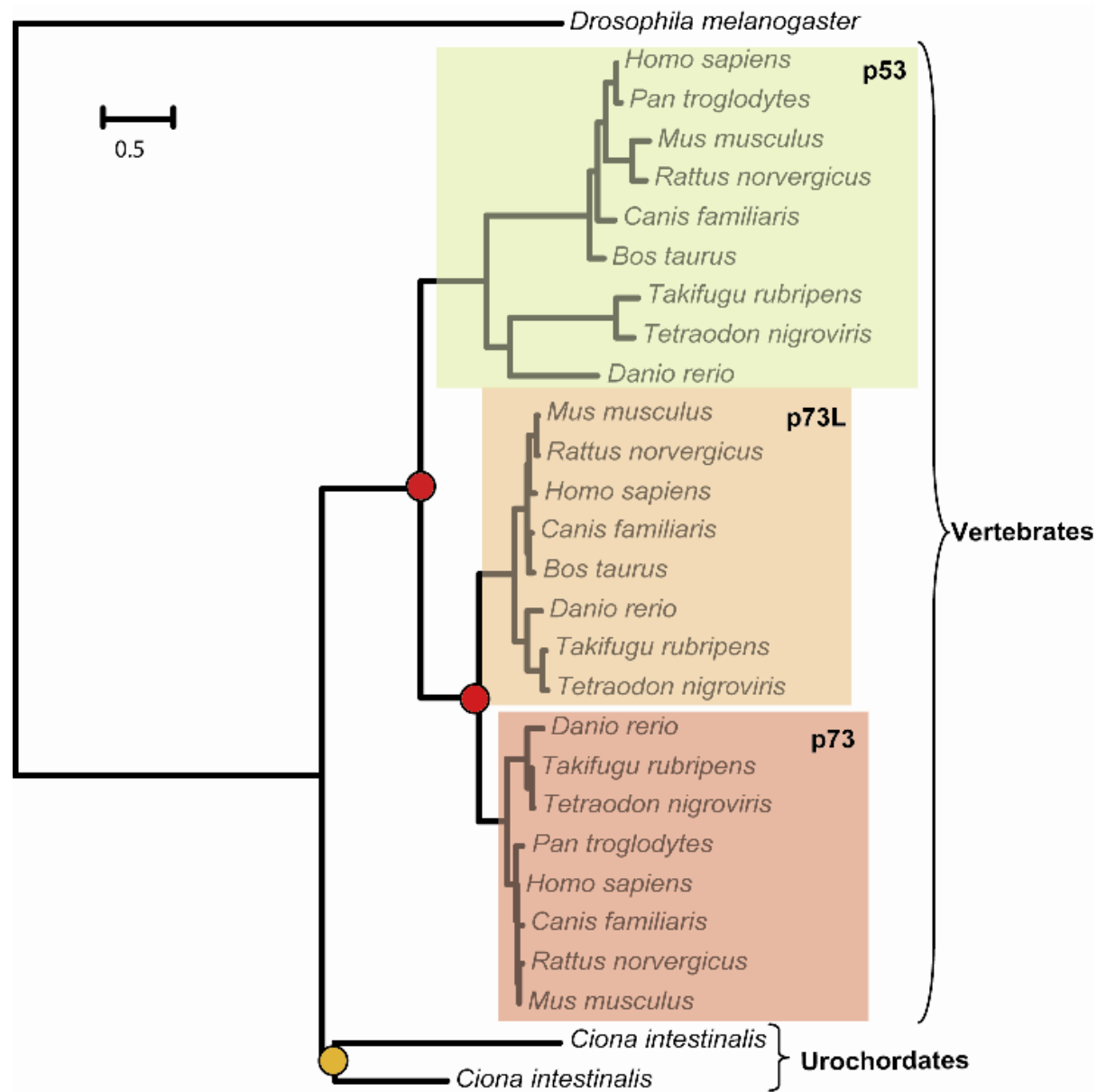
Part II

Orthology prediction methods

Toni Gabaldón
Centre for Genomic Regulation (CRG), Barcelona

Classical approach: phylogenetic inference

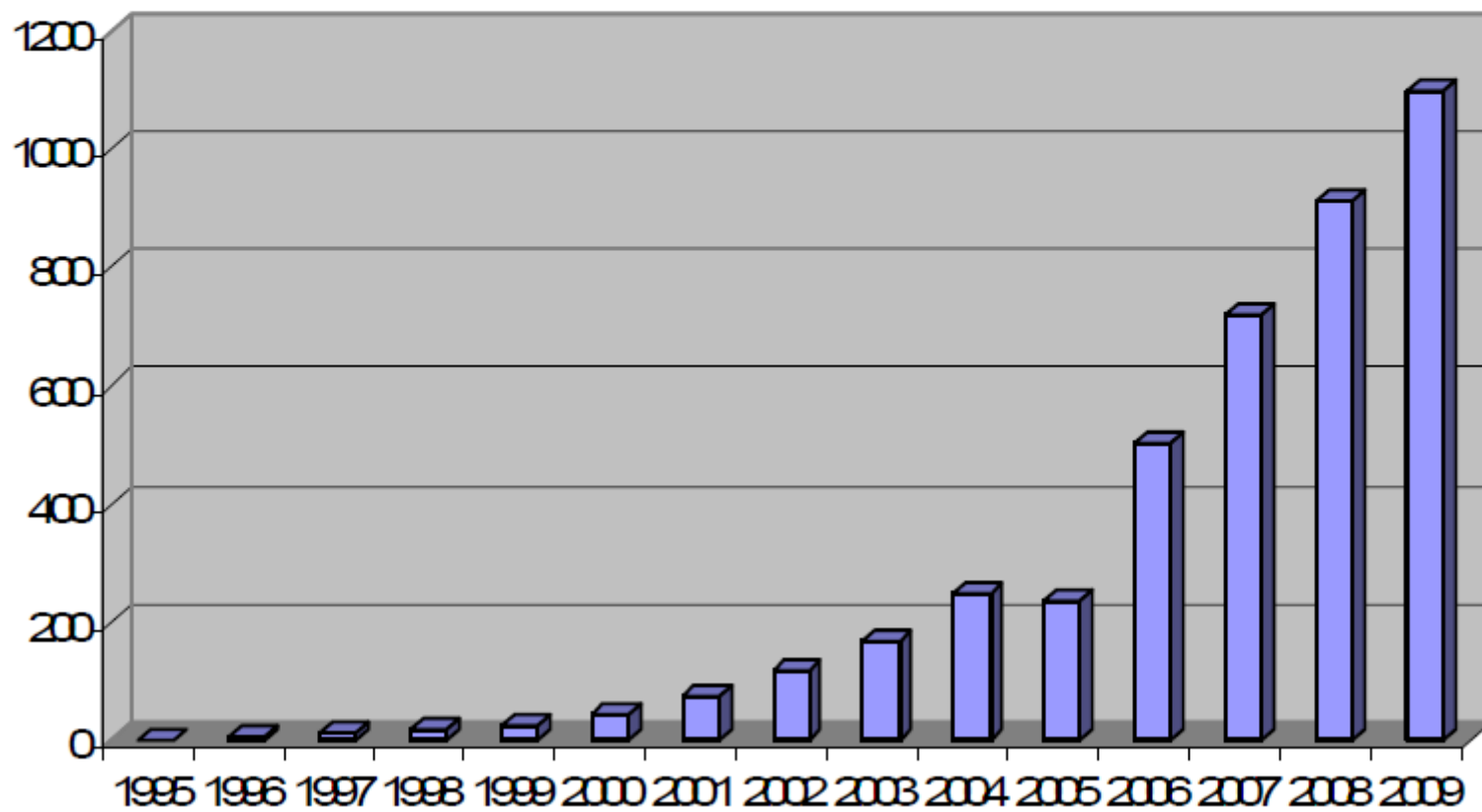
- Build a gene tree
- Compare to the species tree
- Infer duplications and speciation events
- Assign orthology and paralogy relationships accordingly

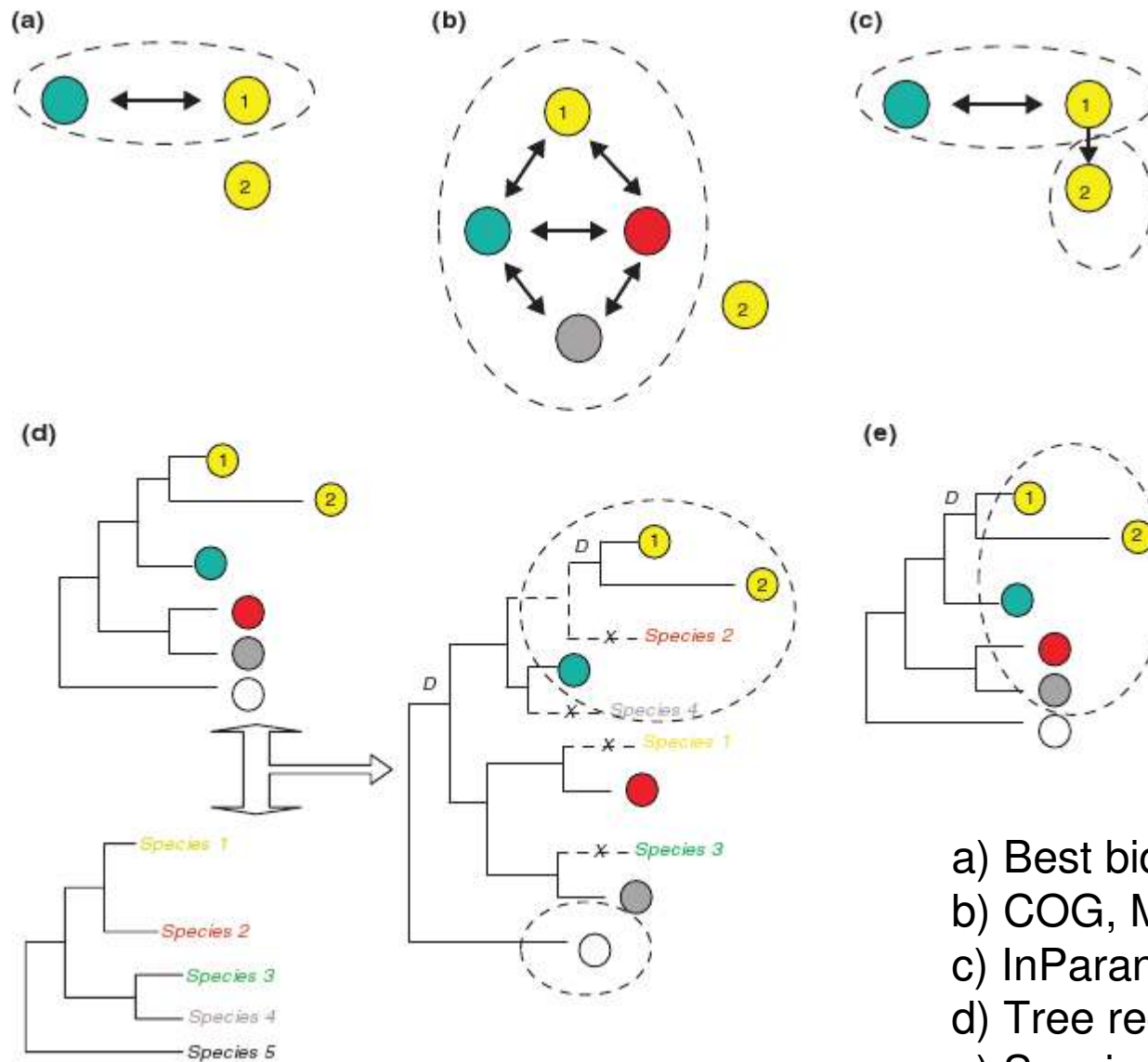


Going genome-wide scale:

Everything must be done automatic and “blind”

Completely sequenced genomes





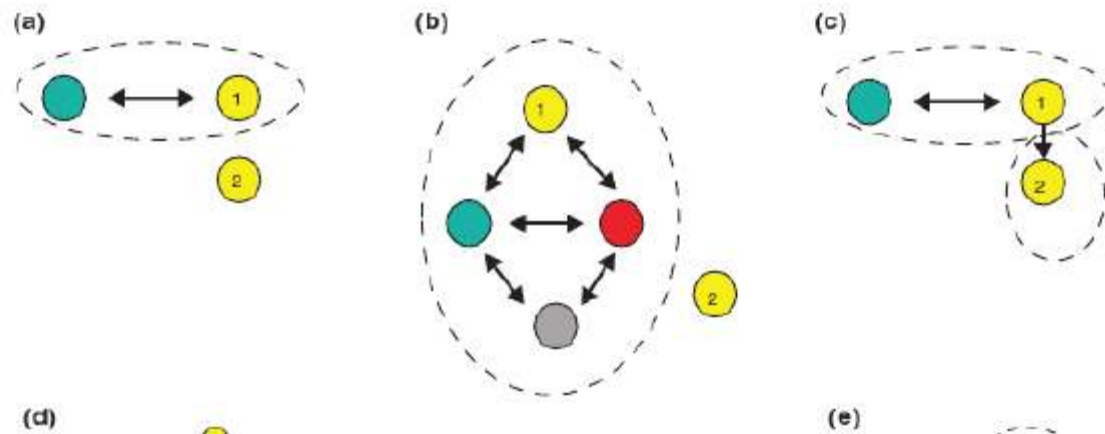
- a) Best bidirectional hits
- b) COG, MCL-clustering approach
- c) InParanoid
- d) Tree reconciliation
- e) Species-overlap (PhylomeDB)

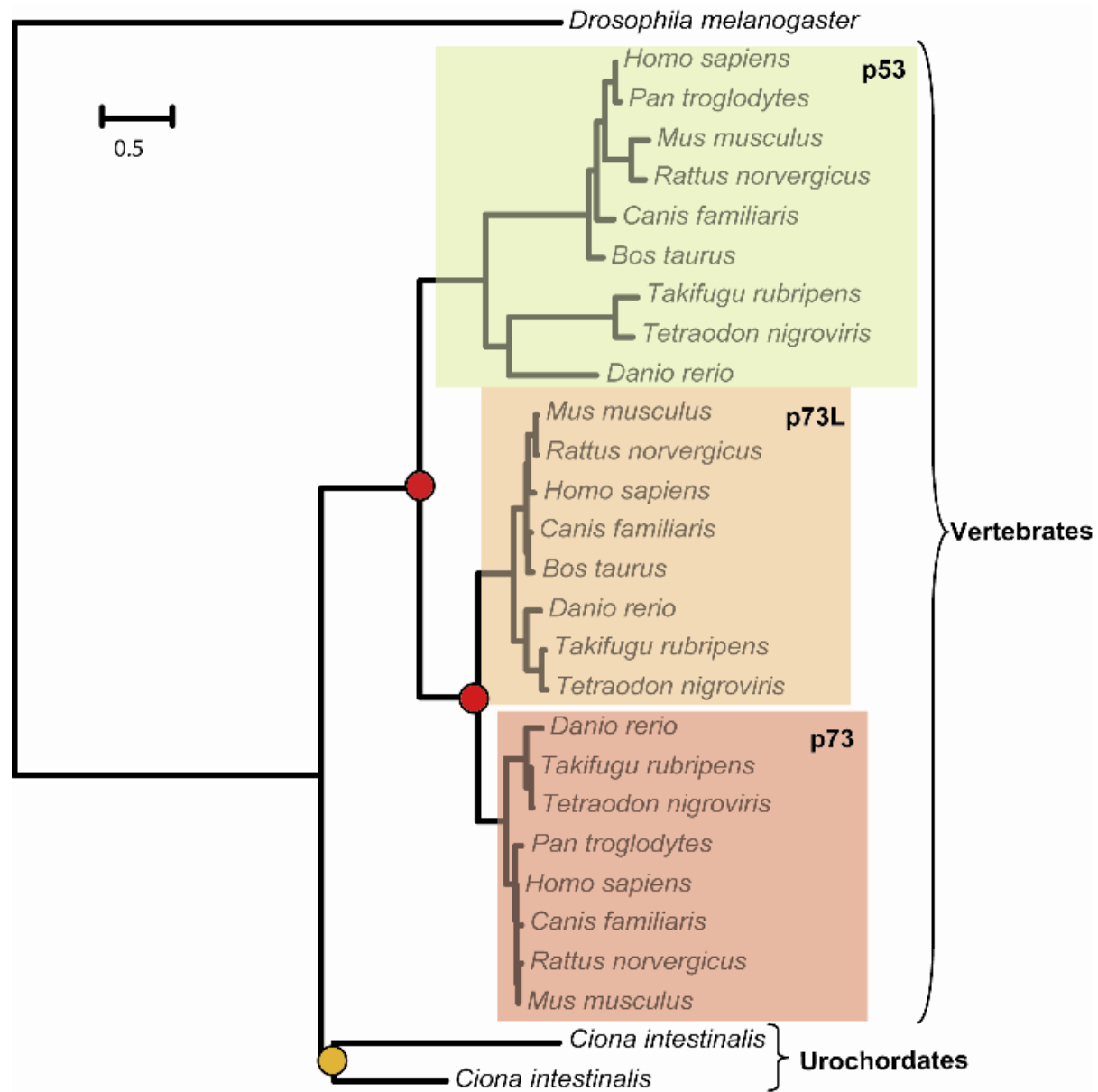
Similarity-based approaches (many more approaches):

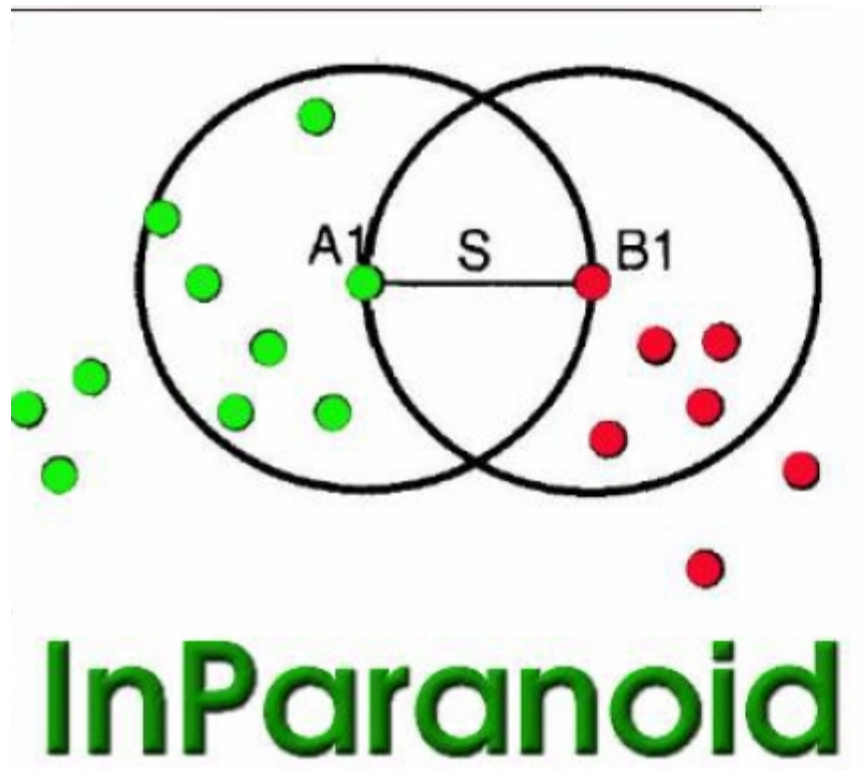
- **Best Reciprocal Hits**

- 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

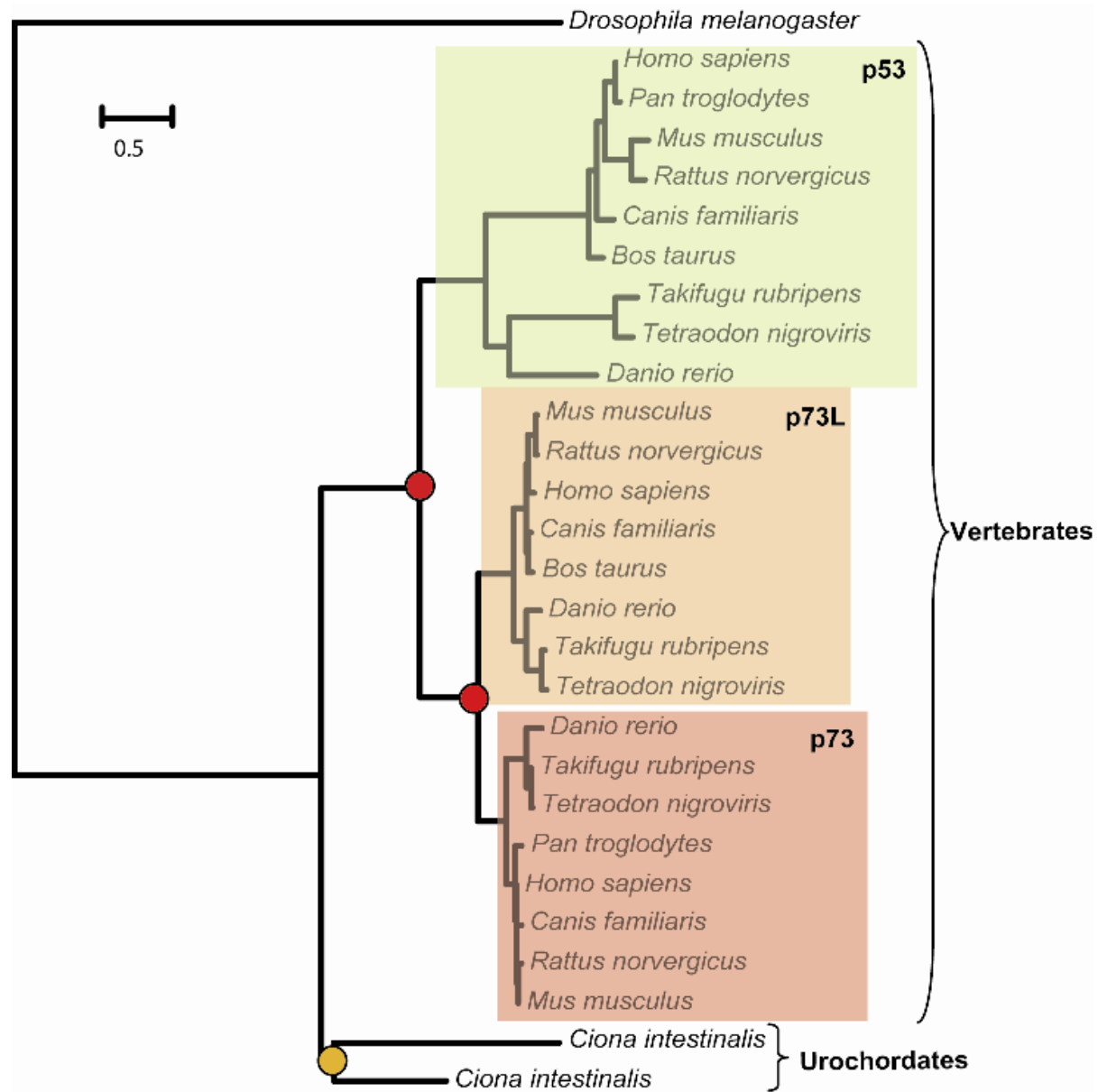






In-Paranoid.

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



Note:

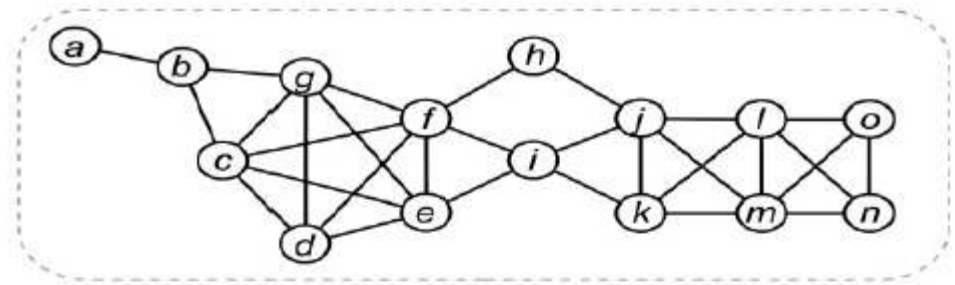
Definition of **in-** and **out-paralogues** require the specification of a given **speciation-node** of reference

COG-like
(used by many DBs like **STRING**)

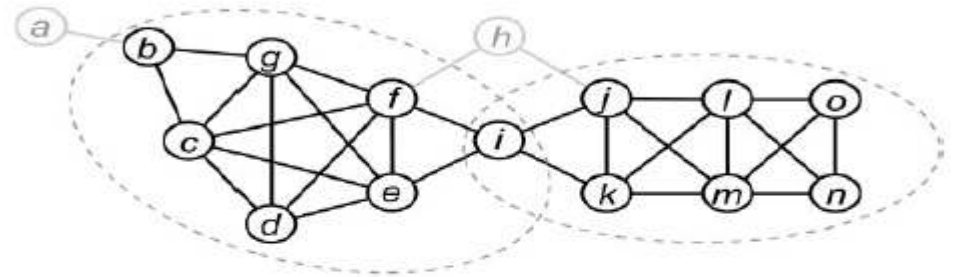
Exploits multi-species information.
Predicts clusters of orthologous
groups (in-paralogs) not all pairs in
a cluster are paralogs.

Can be used at different stringent
levels

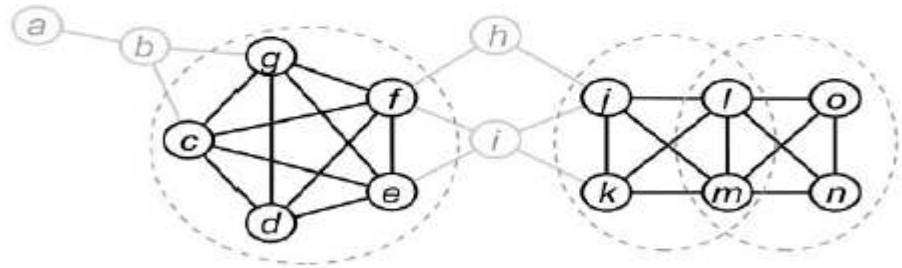
2



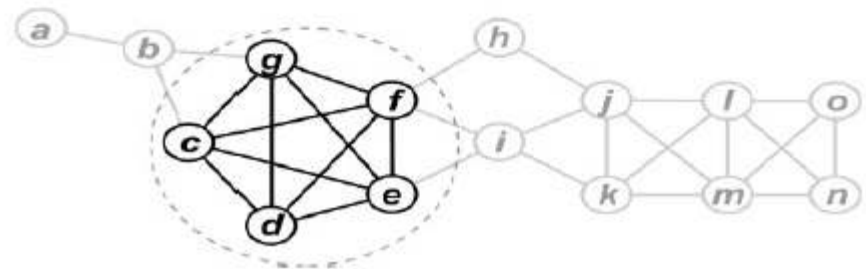
3



4



5



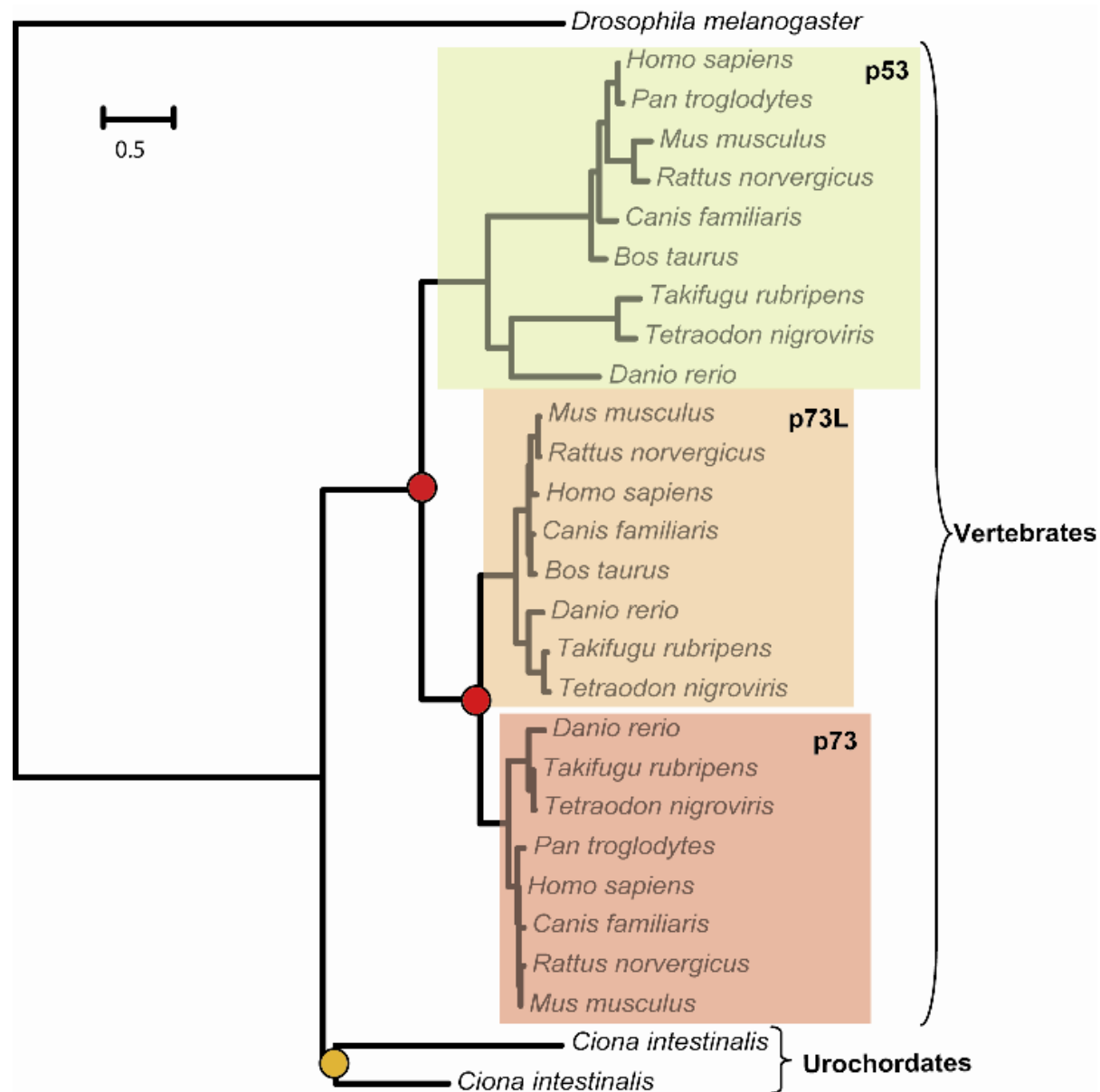
Clustering methods produce: orthologous groups

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

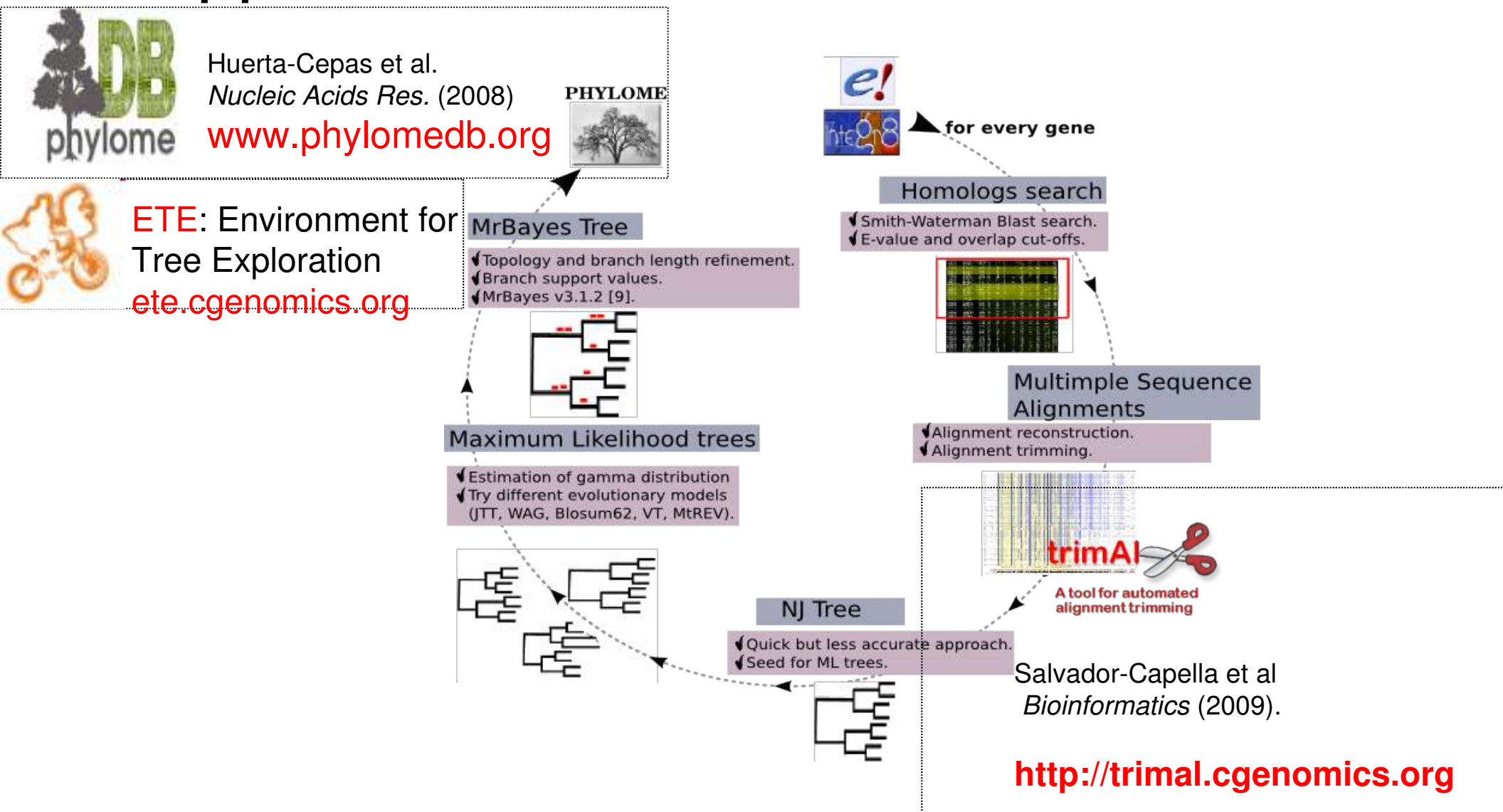
Published: 30 October 2008

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:



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

Phylogeny-based methods

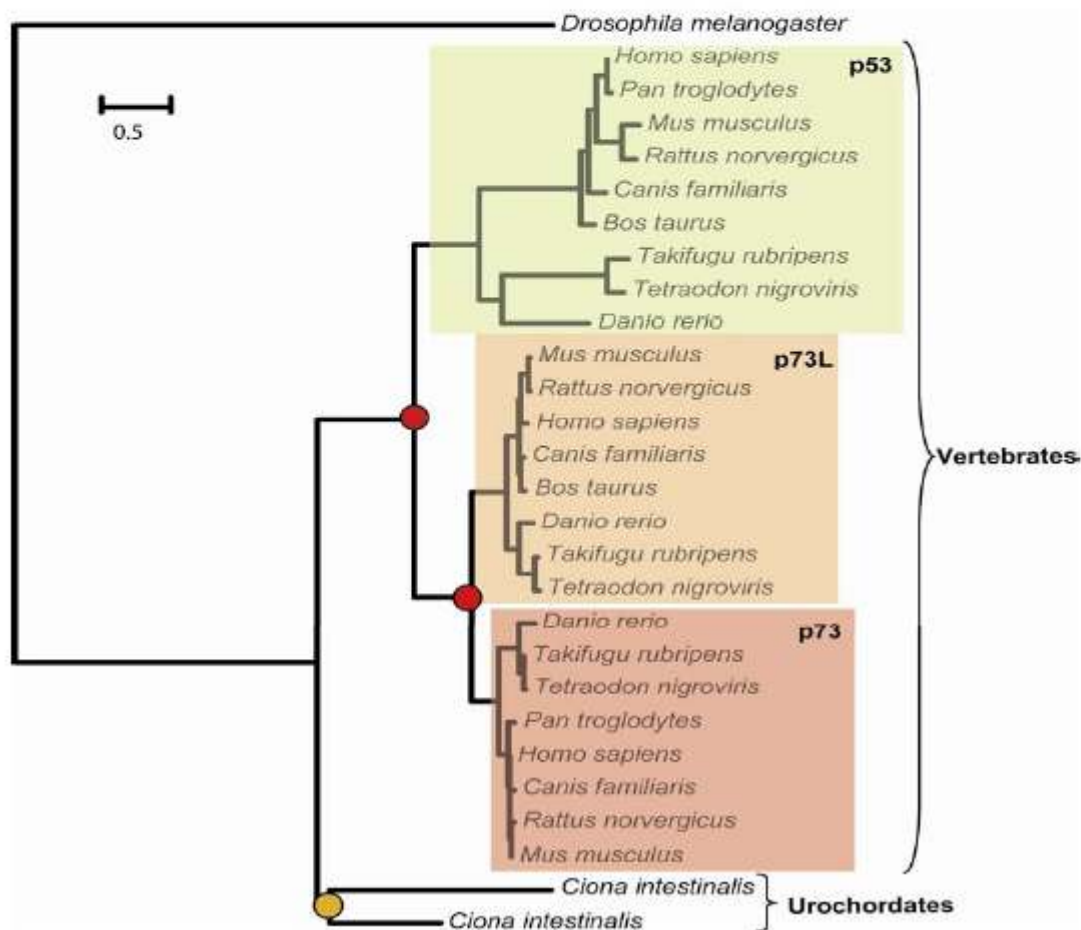
- General procedure: reconstruct the evolution of a gene family (phylogenetics), detect duplication and speciation nodes and predict orthology and paralogy accordingly.
- Two main methods for predicting duplication and speciation nodes from a tree:
 - Species tree reconciliation (RIO, Ensembl)
 - Species-overlap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

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

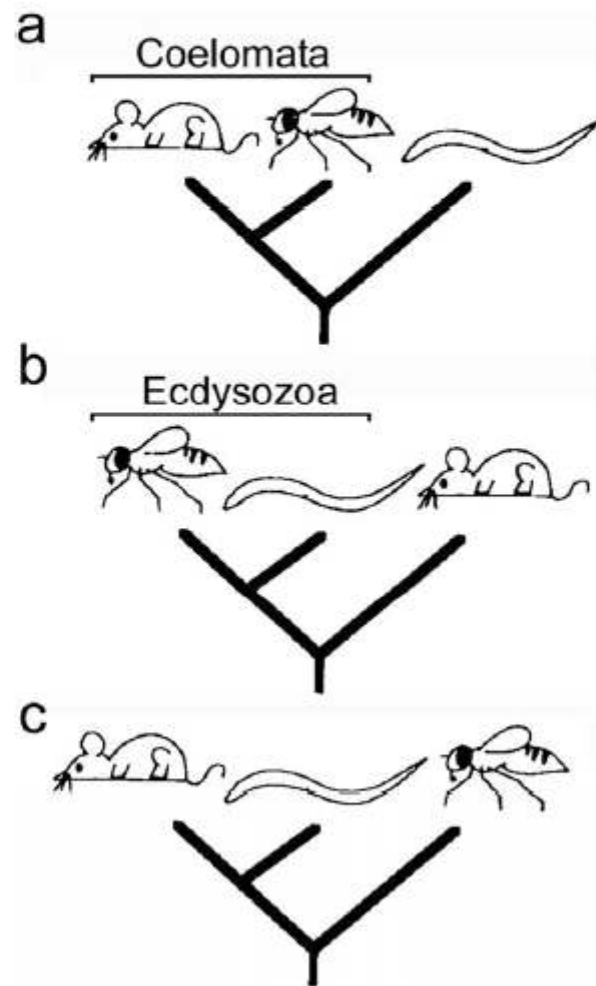
It works when these two assumptions are correct:

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



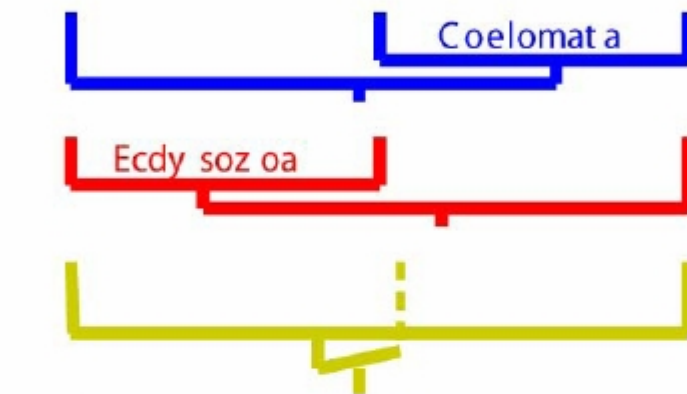
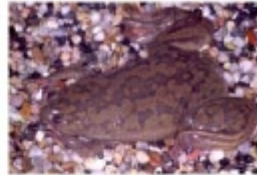
Nematodes



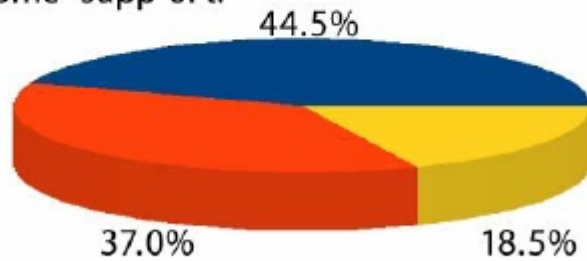
Arthropods



Chordates



Phylogenetic support:



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

Similar results for

Primates

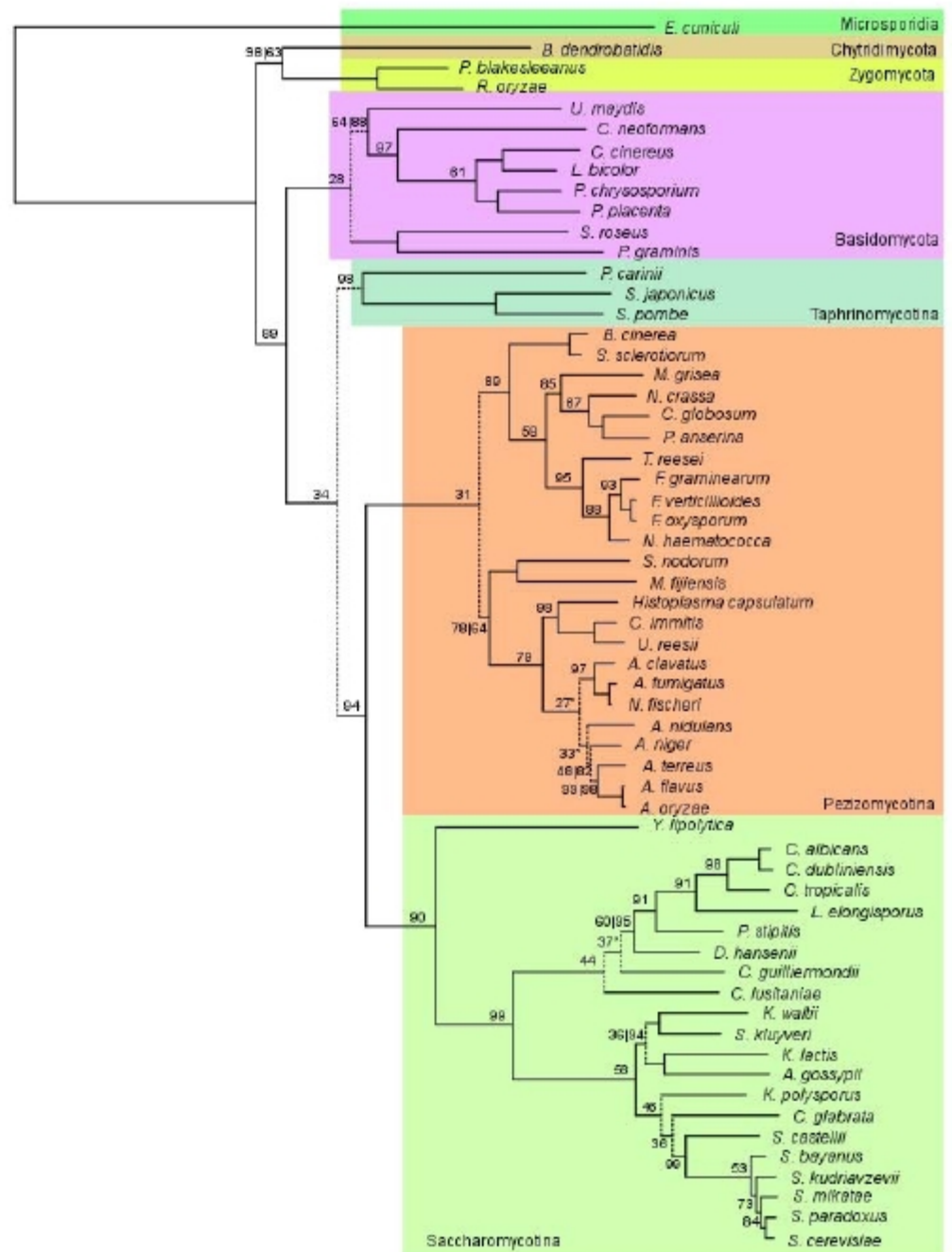
Rodents

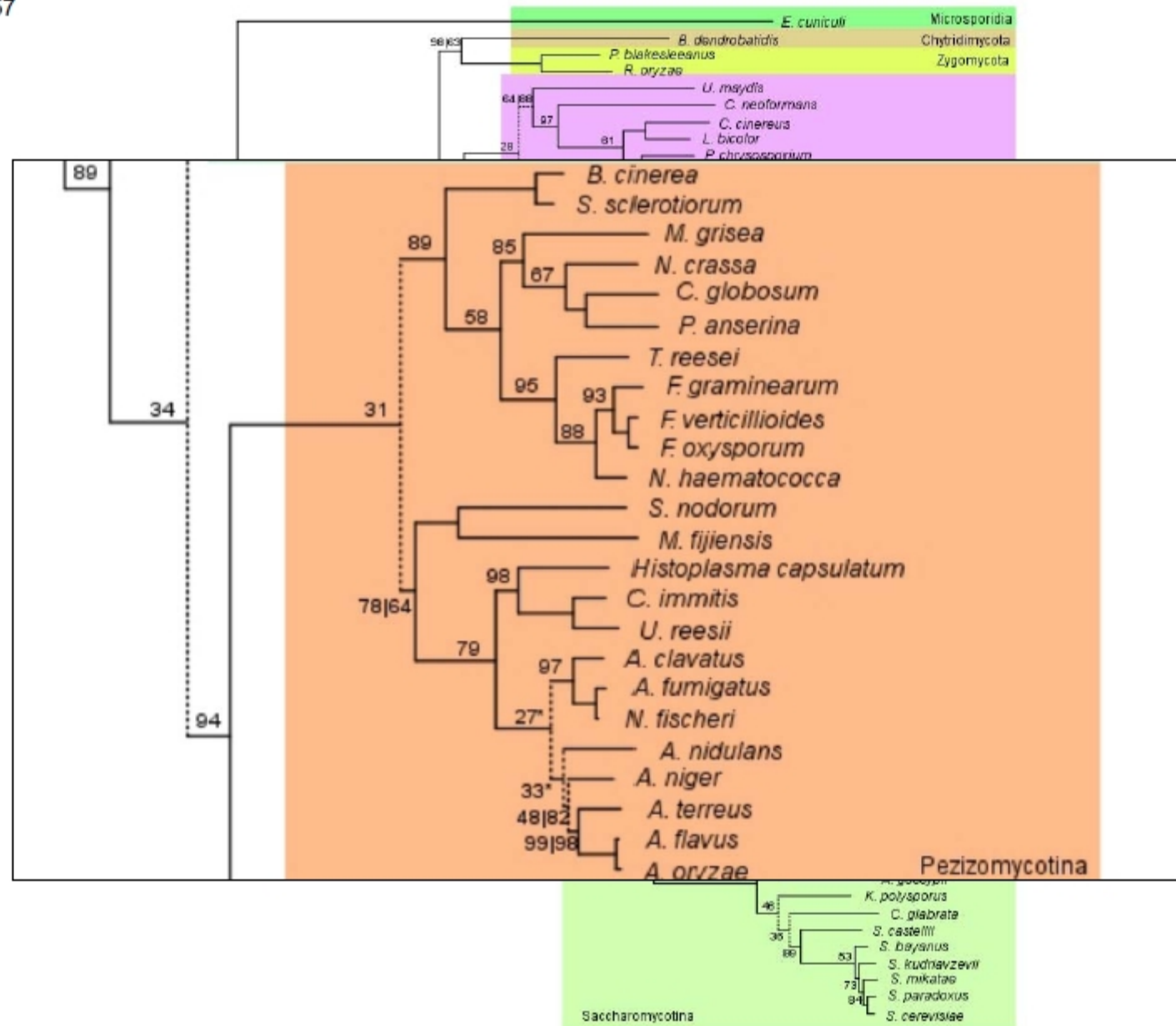
Mammalia

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

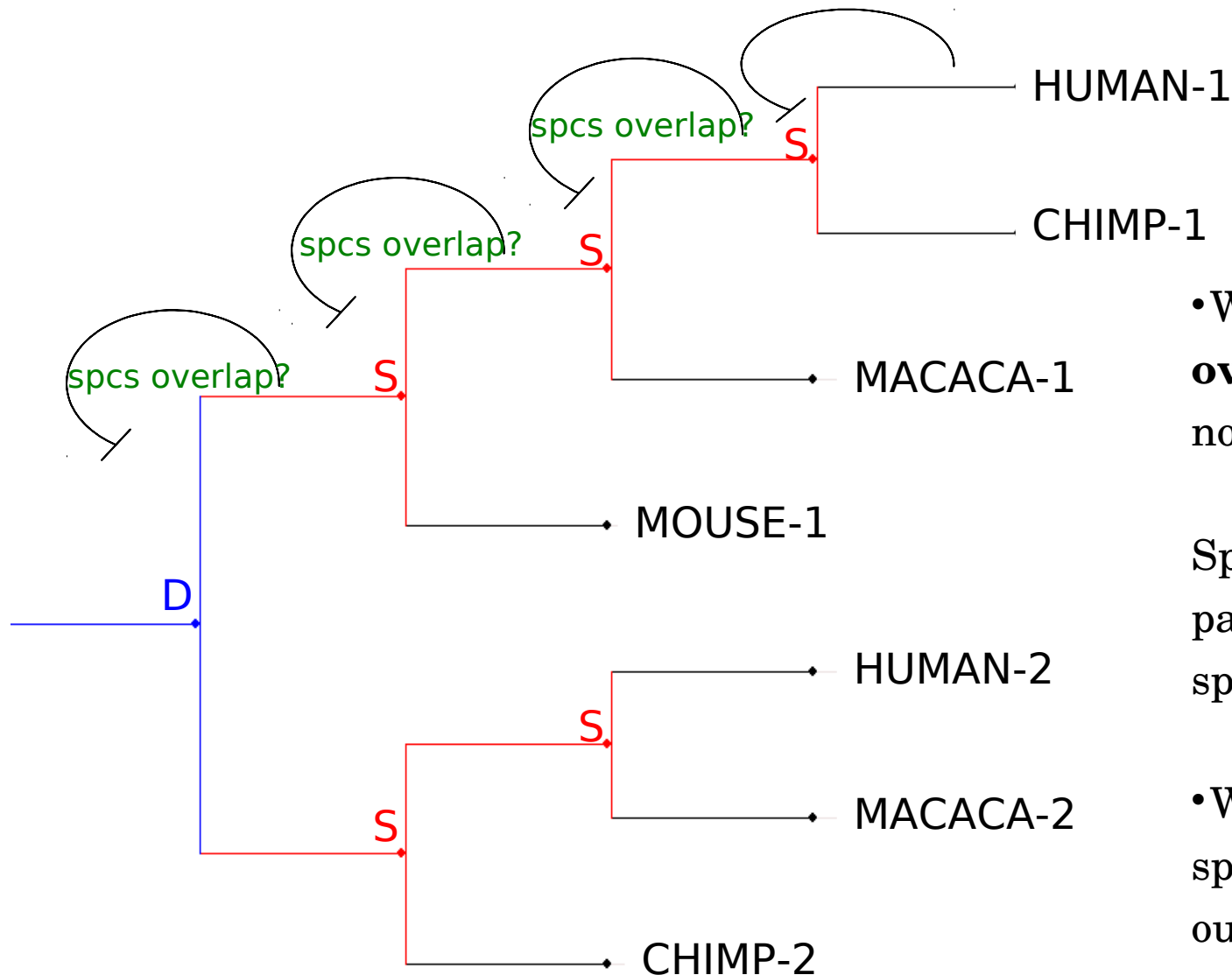




This large-degree of topological variability might be in part due to phylogenetic artifacts, insufficient phylogenetic signal, etc. But also to real evolutionary processes that render a gene tree different from a species tree: lineage sorting, gene conversion, etc

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)

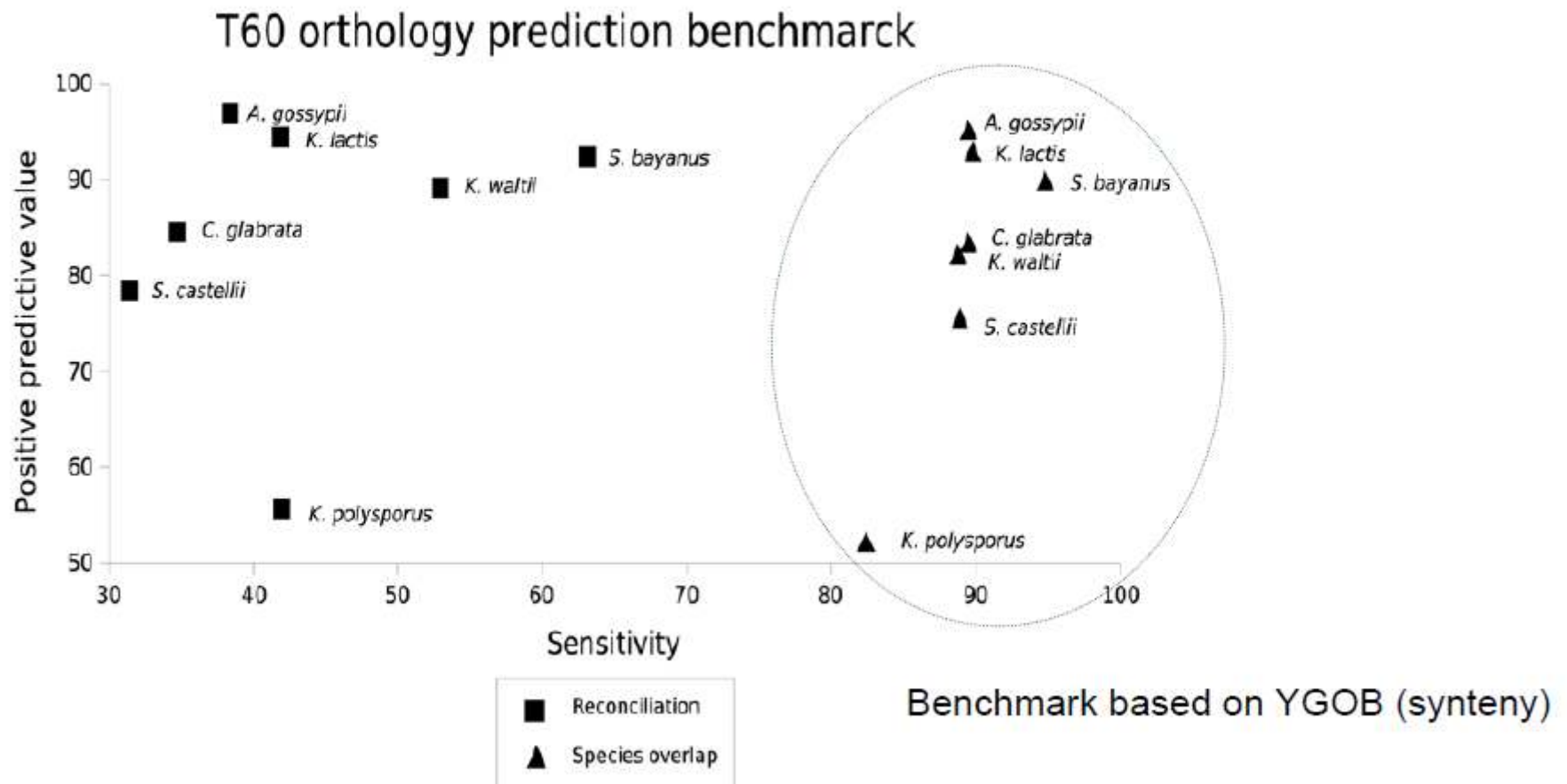
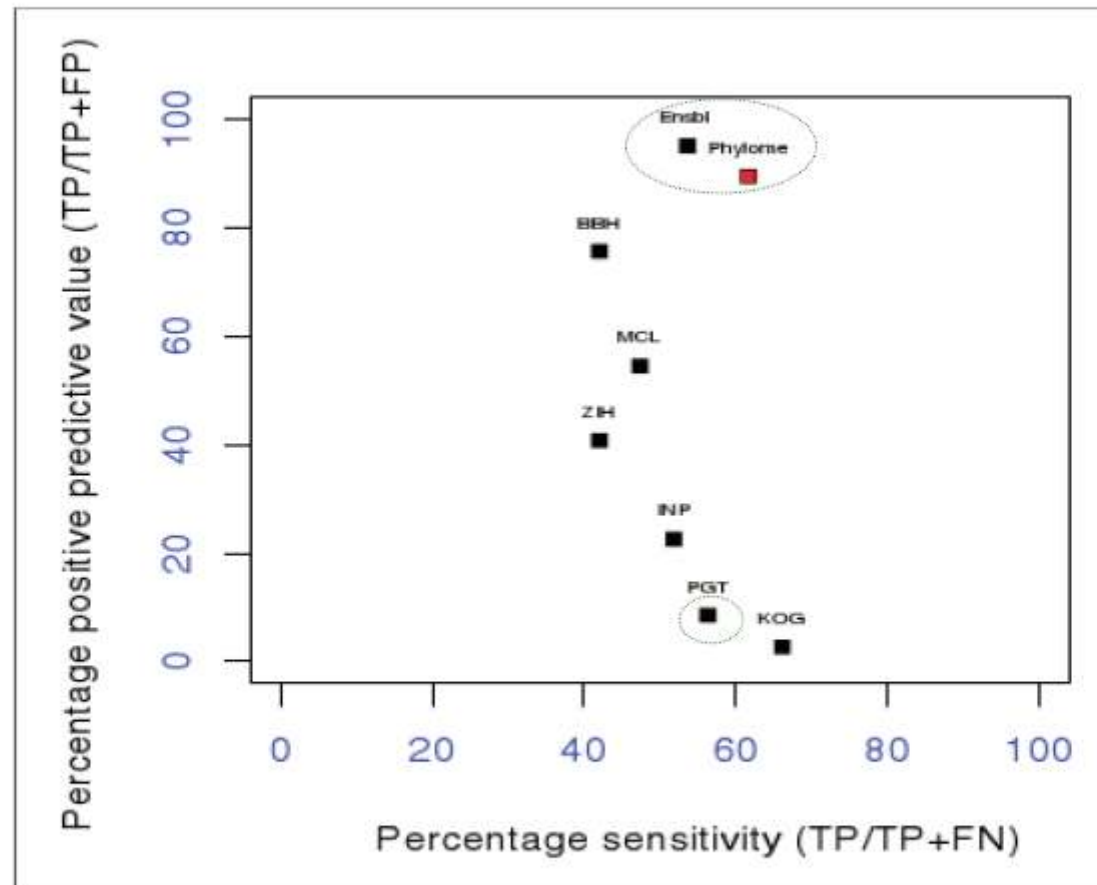


Figure 2. Comparison of different orthology inference algorithms. The synteny based and manually curated orthology predictions available at YGOB database [18] is taken as a golden set to compute the number of true positives (TP), false positives (FP) and false negatives (FN) yielded by each method. For 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:

[HOME](#)

[BROWSE PHYLOMES](#)

[DOWNLOADS](#)

[FAQ](#)

[HELP](#)

[ABOUT](#)

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**.



user:
pass:
[login](#)

[HOME](#) [BROWSE PHYLOMES](#) [DOWNLOADS](#) [FAQ](#) [HELP](#) [ABOUT](#)

YBL058W

[Search!](#)

[Blast Search!](#)

[Select phylomes](#)

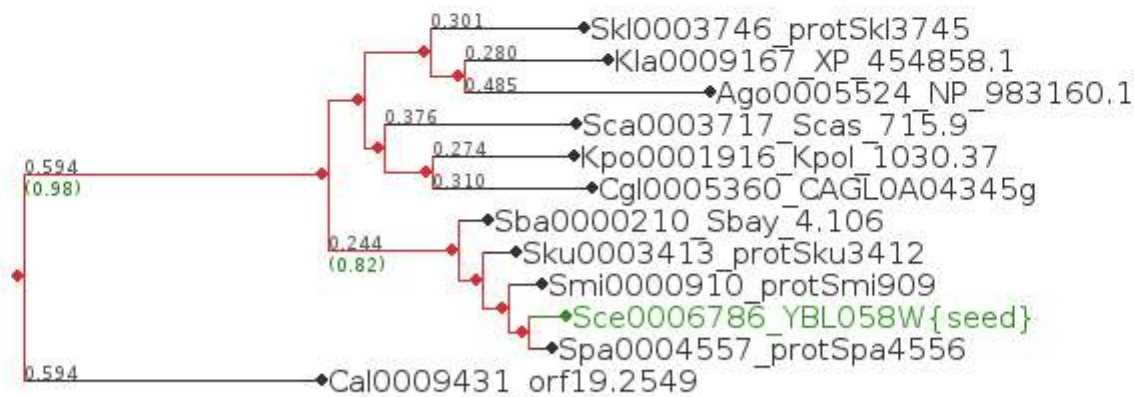
- YBL058W
 - Info
 - Orthologs
 - Seed Trees (4)
 - SceP12a
 - SceP21
 - SceP12b
 - SceP60
 - Collateral Trees (4)
 - Hsa0028724
 - Hsa0028192
 - Hsa0018651
 - Hsa0016629

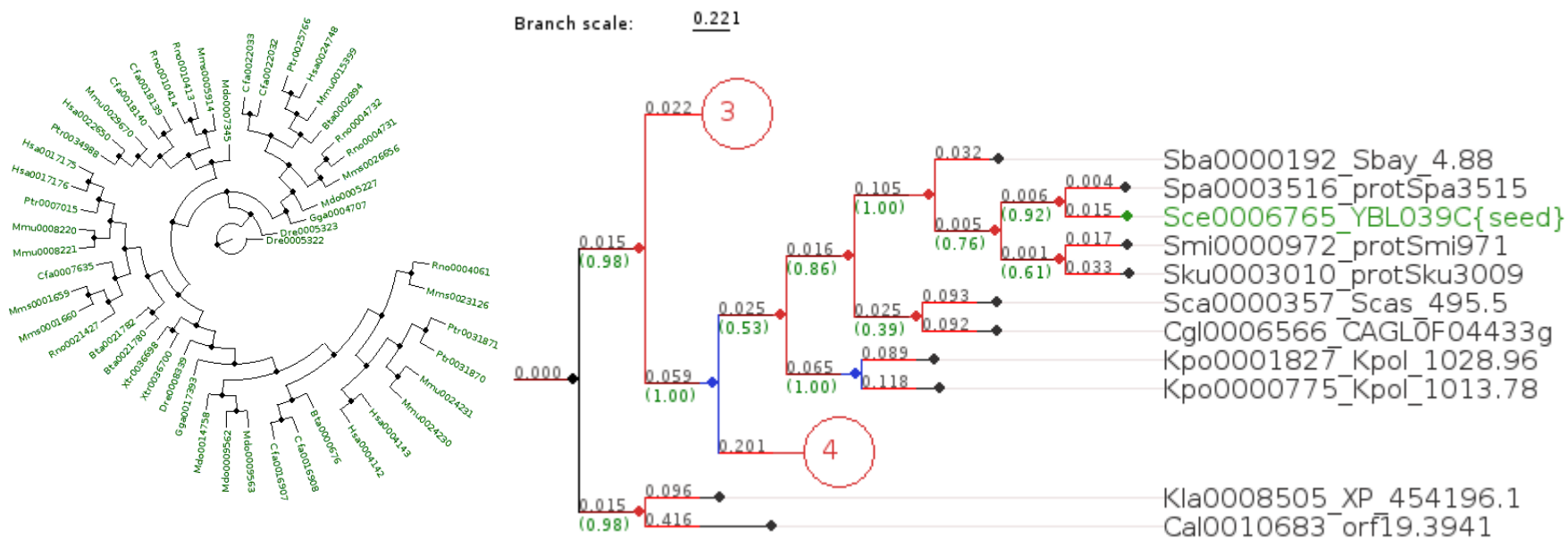
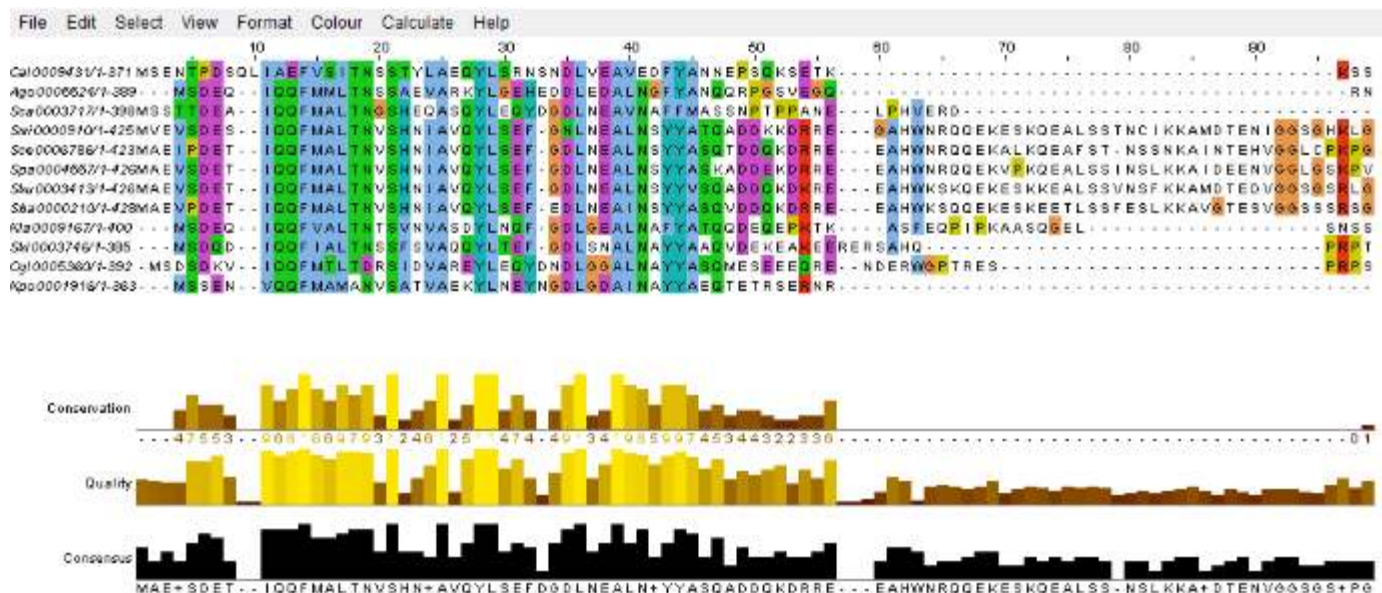
Tree: Sce0006786 (in phylome SceP12a)

Tree model:

Tree Tools and Actions:

Branch scale: 0.081







MetaPhOrs

(Meta-Phylogeny-Based-Orthologs)

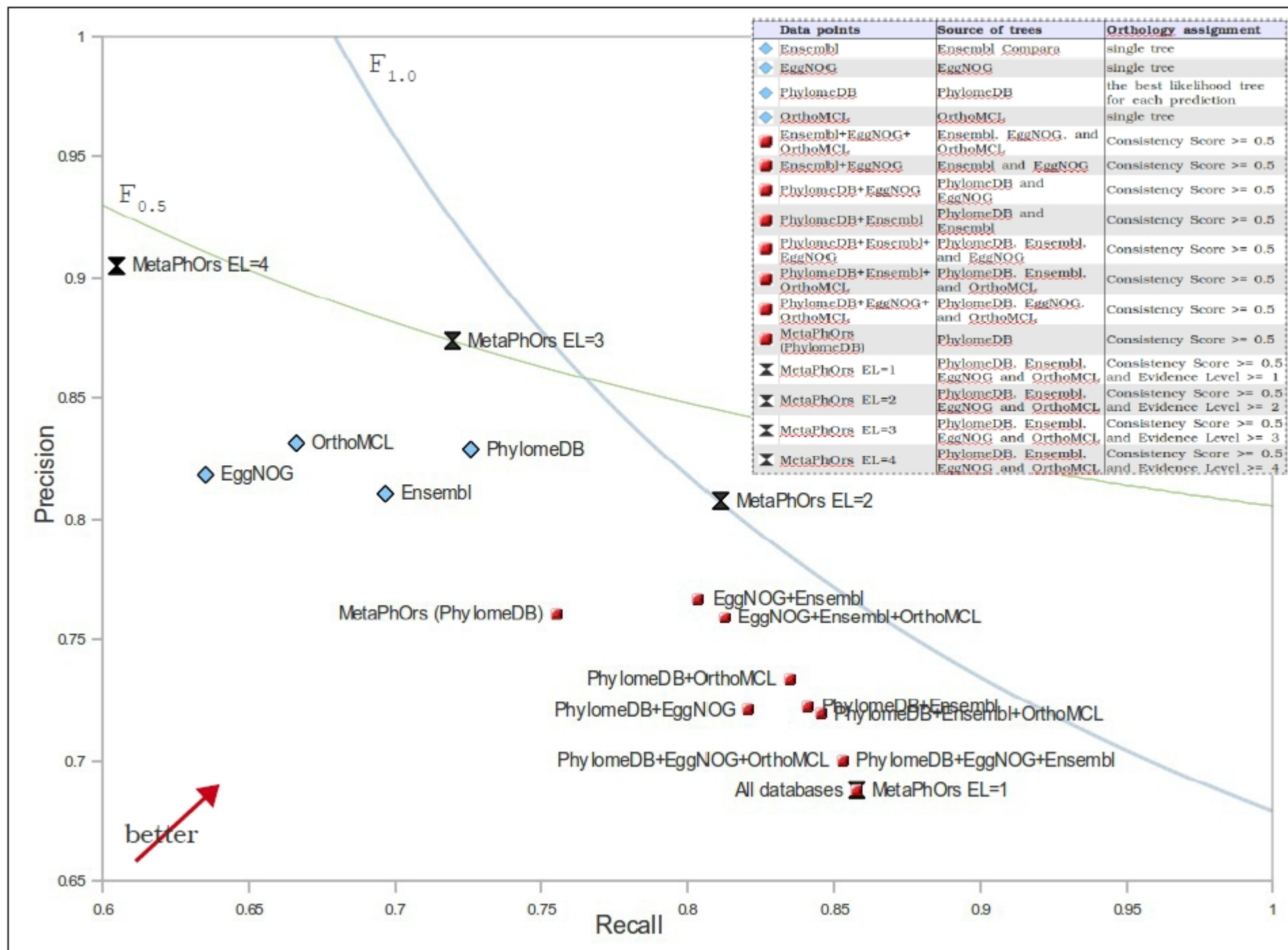


Use existing tree repositories

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



<http://orthology.phylomedb.org>

Give me all orthologs for a list of IDs

The screenshot shows the metaPhOrs website interface. At the top is a navigation bar with links: Home, Multi-query predictions, Genome-wide predictions, Statistics, Downloads, FAQs, and Help. On the left is a sidebar with a 'quick search' section containing a text input field with 'YBL058W' and a 'SEARCH' button. Below this are links for 'You can also use a BLAST search' and 'Change options'. A 'Navigation' section includes a link for 'Frequently Asked Questions'. The main content area has a 'Welcome to metaPhOrs' heading, followed by a paragraph describing the database as a public repository of phylogeny-based orthology and paralogy predictions, mentioning 306 million unique homologous protein pairs from 705,123 phylogenetic trees for 829 genomes. It also provides a 'Consistency Score' and 'Evidence Level' for each prediction. Below this is a paragraph about extending the database and a link to a public ftp server. Three arrows point from external text to specific features: one from 'Give me all orthologs for a list of IDs' to the 'Multi-query predictions' link; one from 'Give me all orthologs to TP53' to the 'SEARCH' button; and one from 'Blast my sequence and give me its orthologs' to the 'You can also use a BLAST search' link. A fourth arrow points from 'Give me all orthologs between Human and Mouse' to the 'Genome-wide predictions' link.

metaPhOrs

quick search

YBL058W

SEARCH

You can also use a BLAST search

Change options

Navigation

○ Frequently Asked Questions

Welcome to metaPhOrs

MetaPhOrs is a public repository of **phylogeny-based** orthology and paralogy predictions that were computed using resources available in seven popular homology prediction services (PhylomeDB, EnsemblCompara, EggNOG, OrthoMCL, COG, Fungal Orthogroups, and TreeFam). Currently above **306 millions** of unique homologous protein pairs are deposited in MetaPhOrs database. These predictions were retrieved from **705 123 phylogenetic trees** for **829 genomes**. For each prediction, MetaPhOrs provides a **Consistency Score** and **Evidence Level** describing its goodness, together with number of trees and links to their source databases.

We are keen on extending MetaPhOrs to additional phylogenetic datasets. If you have a specific suggestion of a phylogenetic dataset that is extensive and has a sufficient quality, please do not hesitate to contact us and we will consider its implementation.

All the data available in metaPhOrs, can be accessed through our [public ftp server](#).

Home Multi-query predictions Genome-wide predictions Statistics Downloads FAQs Help

Give me all orthologs for a list of IDs

Give me all orthologs to TP53

Blast my sequence and give me its orthologs

Give me all orthologs between Human and Mouse

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

quick search

You can also use a BLAST search

[Change options](#)

Navigation

[Frequently Asked Questions](#)

User login

Username: *

Password: *

[Request new password](#)

Orthology predictions for P04637

Phy000865J_HUMAN (Homo sapiens) mapped as: P04637

Target species	H. sapiens co-orthologs (CS)	Target orthologs	CS	Evidence level	Trees	PhylomeDB CS / EL	Ens	Egg	Ort	COG	FO	TF
Acyrtosiphon pisum	4 co-orthologs	Phy000YFHA	0.833	3	6	0.833 / 3						
	4 co-orthologs	Phy000YLR7	0.833	3	6	0.833 / 3						
	4 co-orthologs	C4WXY0	0.833	3	6	0.833 / 3						
Aedes aegypti	4 co-orthologs	Q171M5	1.000	2	3	1.000 / 1						
	4 co-orthologs	Q171M1	0.800	3	5	0.667 / 1						
Anopheles gambiae	4 co-orthologs	Q7QAB9	0.833	4	6	0.800 / 3						
	4 co-orthologs	Q7QBX6	0.875	5	8	0.833 / 3						
Apis mellifera	3 co-orthologs	Phy000ZPXS	0.667	1	3	0.667 / 1						
Bombyx mori	3 co-orthologs	Phy000VIB2	1.000	2	3	1.000 / 2						
Bos taurus	Phy000865J_ (1.00)	P67939	1.000	4	7	1.000 / 1						
Branchiostoma floridae	3 co-orthologs	C3XPU2	1.000	1	1	-						
	3 co-orthologs	C3YXH3	1.000	1	1	-						
	3 co-orthologs	C3ZIW1	1.000	1	1	-						

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?

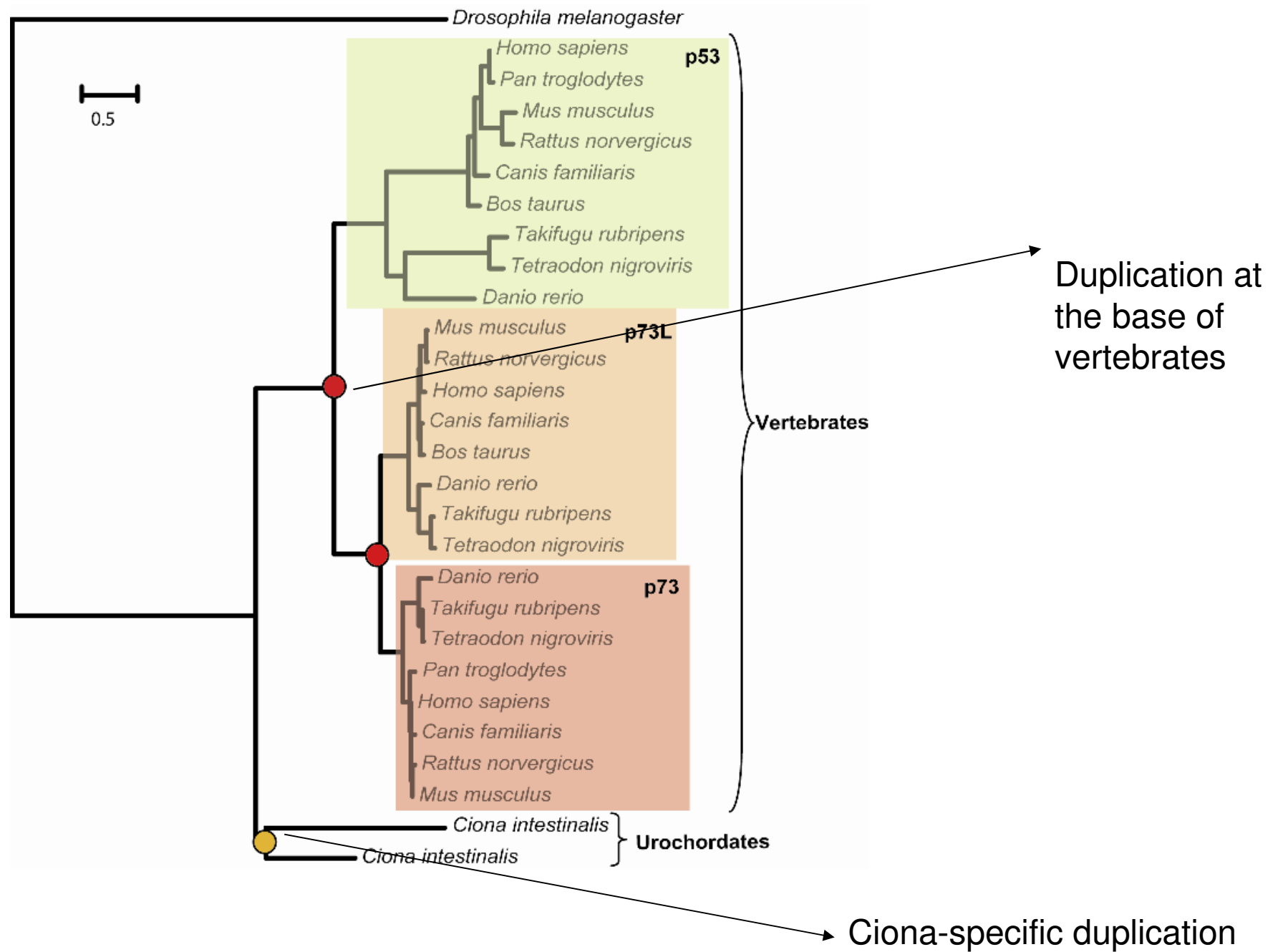
- Different taxonomic focuses
- Different methodologies
- Different outputs (pairwise relationships, groups, etc)
- Different interfaces
- Different accuracies (**how to benchmark this?**)

What about paralogy?

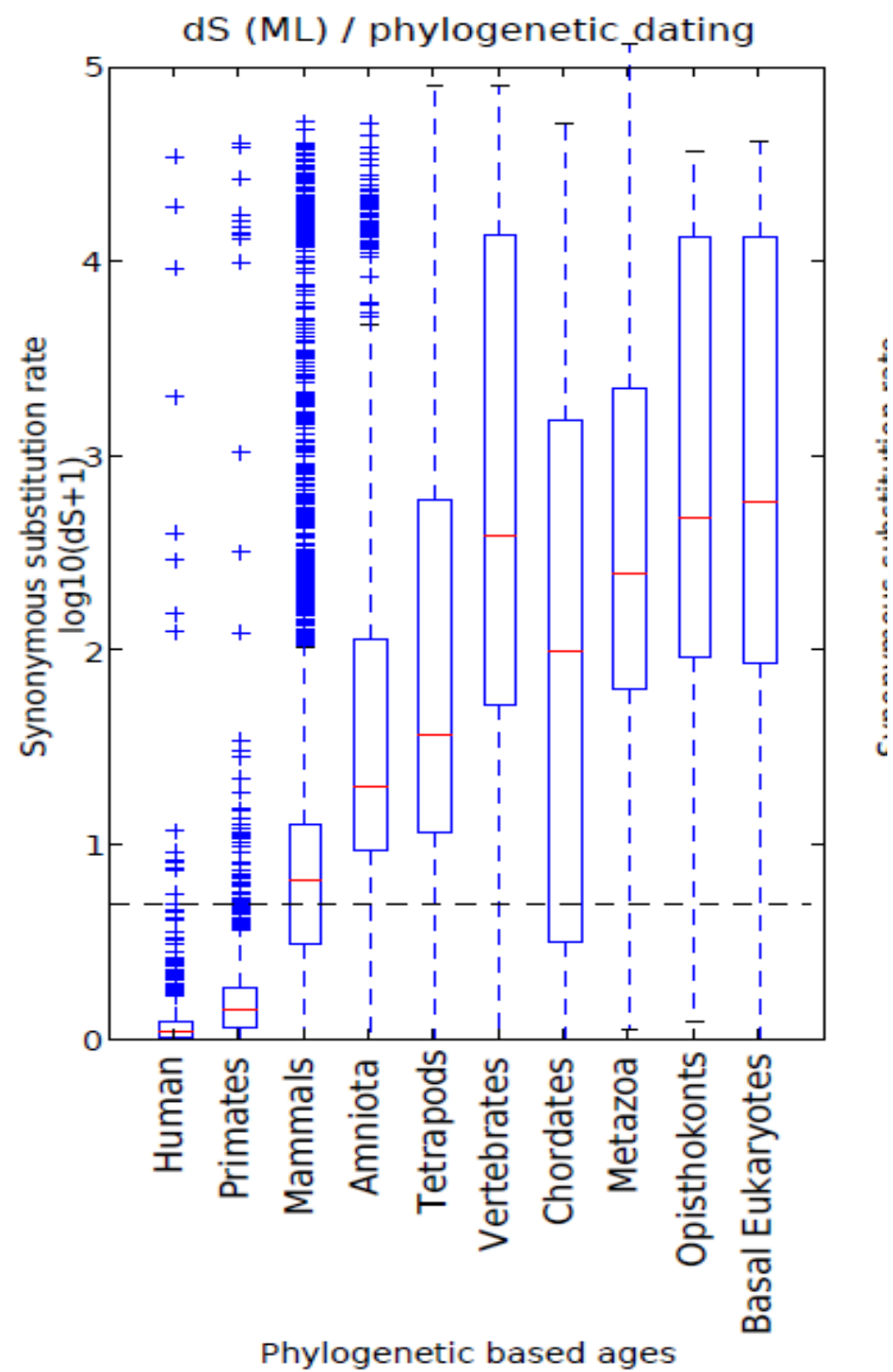
Most pairwise methods focus on orthologs, only in-paralogs are taken into account sometimes.

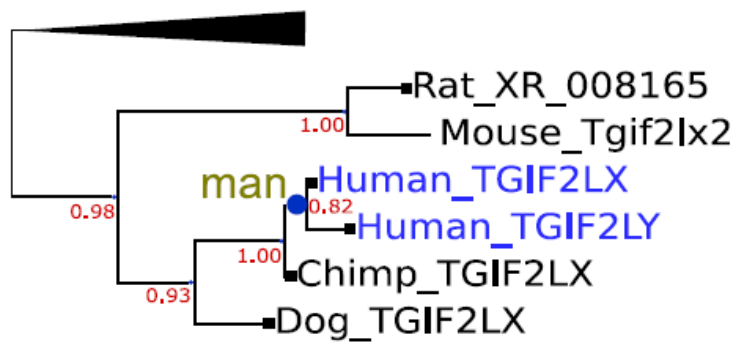
Phylogeny-based methods readily inform both on orthology and paralogy.

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



Comparison of topological dating vs dS

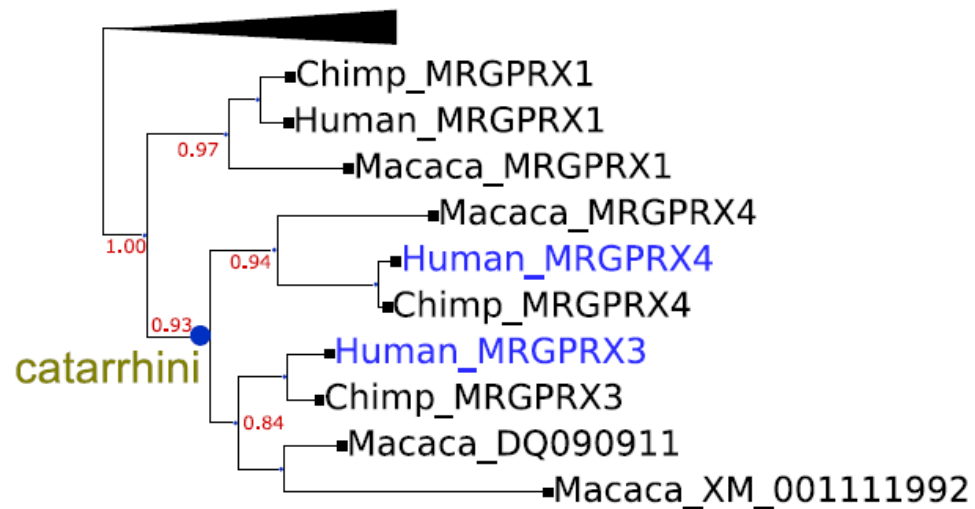




0.66

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

C)



0.10

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

D)