Barcelona Biomedical Research Park

CRG Centre de Regulació Genòmica

Orthology
Part I
concepts and implications

Toni Gabaldón

Centre for Genomic Regulation (CRG), Barcelona

Barcelona Biomedical Research Park

CRG Centre de Regulació Genòmica

Toni Gabaldón

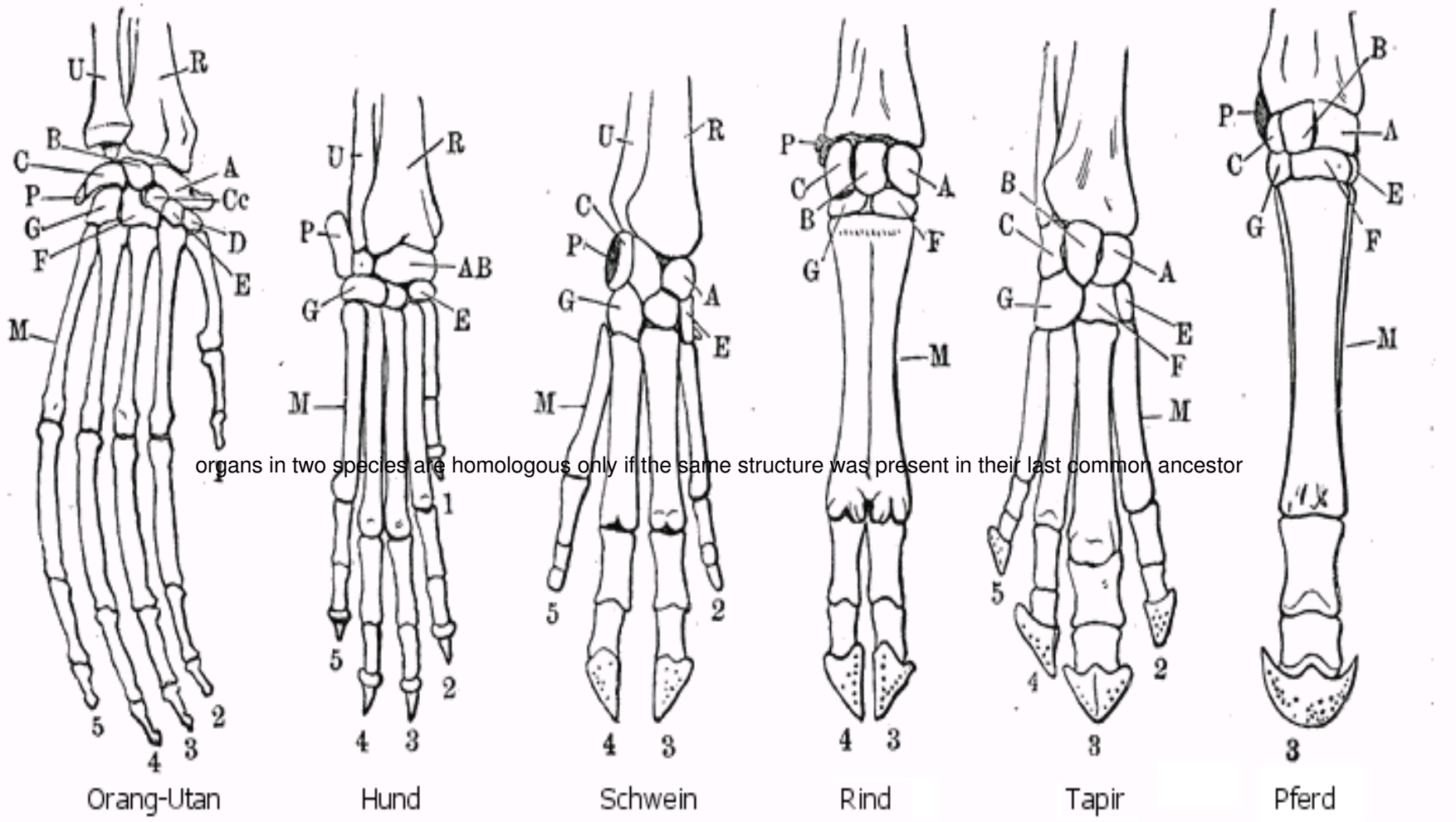
Contact: tgabaldon@crg.es
Group website: http://gabaldonlab.crg.es
Science blog: http://treevolution.blogspot.com
Twitter: @gabaldonlab, @Toni_Gabaldon

Barcelona Biomedical Research Park

CRG Centre de Regulació Genòmica

Orthology

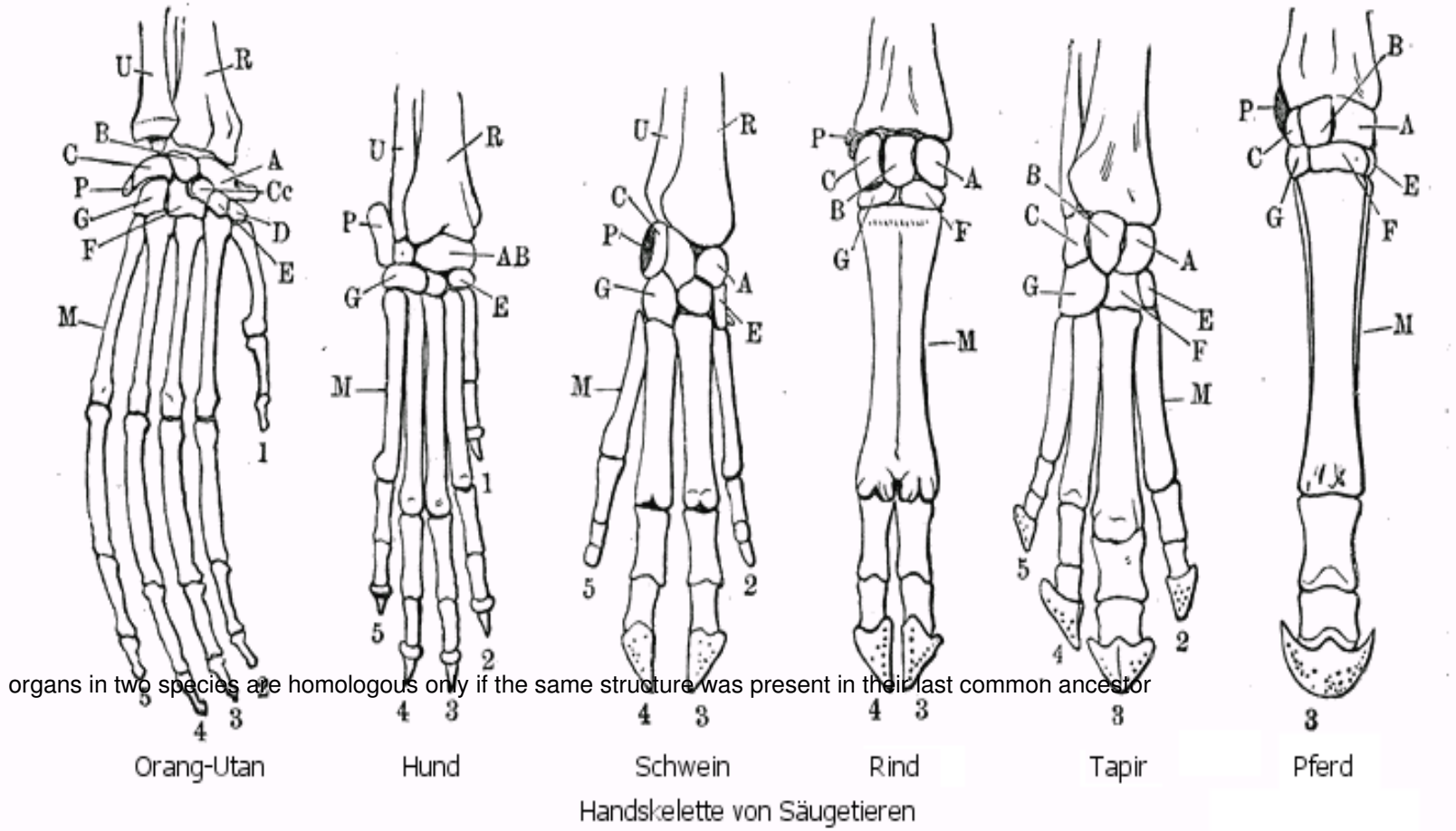
"concepts and implications"



organs in two species are homologous only if the same structure was present in their last common ancestor

Handskelette von Säugetieren

R Radius (Speiche), **U** Ulna (Elle), **A-G**, **Cc**, **P** Knochen des Carpus (Handwurzel): **A** Scaphoideum (Kahnbein), **B** Lunare (Mondbein), **C** Triquetrum (dreieckiges Bein), **D** Trapezium (großes vieleckiges Bein), **E** Trapezoides (kleines vieleckiges Bein), **F** Capitatum (Kopfbein), **G** Hamatum (Hafenbein), **P** Pisiforme (Erbsenbein), **Cc** Centrale Carpi, **M** Metacarpus (Mittelhand). Die Zahlen **1-5** bezeichnen die Finger (**1** Daumen, **5** kleiner Finger).



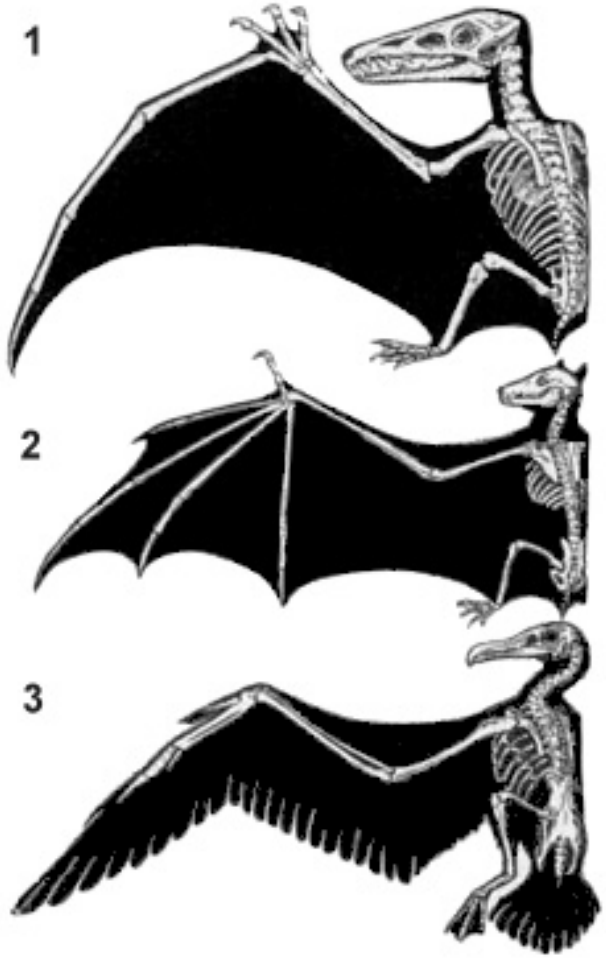
Handskelette von Säugetieren

R Radius (Speiche), **U** Ulna (Elle), **A-G**, **Cc**, **P** Knochen des Carpus (Handwurzel): **A** Scaphoideum (Kahnbein), **B** Lunare (Mondbein), **C** Triquetrum (dreieckiges Bein), **D** Trapezium (großes vieleckiges Bein), **E** Trapezoides (kleines vieleckiges Bein), **F** Capitatum (Kopfbein), **G** Hamatum (Hafenbein), **P** Pisiforme (Erbsenbein), **Cc** Centrale Carpi, **M** Metacarpus (Mittelhand). Die Zahlen **1-5** bezeichnen die Finger (**1** Daumen, **5** kleiner Finger).

“the same organ in different animals under every variety of form and function" **R. Owan**

→ organs in two species are **homologous** only if the same structure was present in their last common ancestor

Analogous structures:
Similar function but independet origin.



Homologous as forelimbs
But
Analogous as wings

Extension of the concept of homology to sequences:

Two sequences are homologous if they share common ancestry

```
AAB24882        TYHMCQFHCRYVNNHSGEKLYECNERSKAFSCPSHLQCHKRRQIGEKTHEHNQCGKAFPT 60
AAB24881        --------------------YECNQCGKAFAQHSSLKCHYRTHIGEKPYECNQCGKAFSK 40
                                    ****: .***:  * *:** * :****.:* *******..

AAB24882        PSHLQYHERTHTGEKPYECHQCGQAFKKCSLLQRHKRTHTGEKPYE-CNQCGKAFAQ- 116
AAB24881        HSHLQCHKRTHTGEKPYECNQCGKAFSQHGLLQRHKRTHTGEKPYMNVINMVKPLHNS 98
                 **** *:***********:***:**.: .***************    :  *.: :
```

Important: Similarity and Homology

Similarity and homology are often confused. e.g. “the sequences are 50% homologous”, “these two sequences are highly homologous”

Why is this incorrect? Where does the confusion comes from?

Detour

Sequence similarity, homology detection and blast database queries

```
AAB24882        TYHMCQFHCRYVNNHSGEKLYECNERSKAFSCPSHLQCHKRRQIGEKTHEHNQCGKAFPT 60
AAB24881        --------------------YECNQCGKAFAQHSSLKCHYRTHIGEKPYECNQCGKAFSK 40
                                    ****: .***:  * *:** * :****.:* *******..

AAB24882        PSHLQYHERTHTGEKPYECHQCGQAFKKCSLLQRHKRTHTGEKPYE-CNQCGKAFAQ- 116
AAB24881        HSHLQCHKRTHTGEKPYECNQCGKAFSQHGLLQRHKRTHTGEKPYMNVINMVKPLHNS 98
                 **** *:***********:***:**.: .***************    :  *.: :
```

Are this two sequences **significantly** similar?
(i.e how likely is that such an alignment is the result of chance)

```
> ref|NP_114344.1| NADH dehydrogenase subunit 5 [Macaca sylvanus]
Length=603

 GENE ID: 803075 ND5 | NADH dehydrogenase subunit 5 [Macaca sylvanus]
(10 or fewer PubMed links)

 Score =  796 bits (2056),  Expect = 0.0, Method: Compositional matrix adjust.
 Identities = 438/564 (77%), Positives = 478/564 (84%), Gaps = 0/564 (0%)

Query  24   VNPNKKNSYPHYVKSIVASTFIISLFPTTMFMCLDQEVIISNWHWATTQTTQLSLSFKLD  83
            +NPNKK+ YP+YVK+ V   FI SL  TT++M L+QE II +WHW  TQT  L+LSFKLD
Sbjct  24   INPNKKHLYPNYVKTAVMYAFITSLSSTTLYMFLNQETIIWSWHWMMTQTLSLTLSFKLD  83

Query  84   YFSMMFIPVALFVTWSIMEFSLWYMNSDPNINQFFKYLLIFLITMLILVTANNLFQLFIG  143
            YFSMMF P+AL  TWSIMEFSLWYM+SDPNI+QFFKYLLIFLITMLILVTANNLFQ FIG
Sbjct  84   YFSMMFTPIALLTTWSIMEFSLWYMSSDPNIDQFFKYLLIFLITMLILVTANNLFQFFIG  143

Query  144  WEGVGIMSFLLISWWYARADANTAAIQAVLYNRIGDIGFILALAWFILHSNSWDPQQMAL  203
            WEG+GIMSFLLISWW+AR DANTAAIQA+LYNRIGDIG IL + WF+LH NSWD QQM
Sbjct  144  WEGMGIMSFLLISWWHARTDANTAAIQAILYNRIGDIGLILTMTWFLLHYNSWDFQQMLA  203
```

Score of a High Scoring Pair (HSP)

Alignment scores are sums of residue-pairing scores according to a Scoring Matrix

BLOSUM62

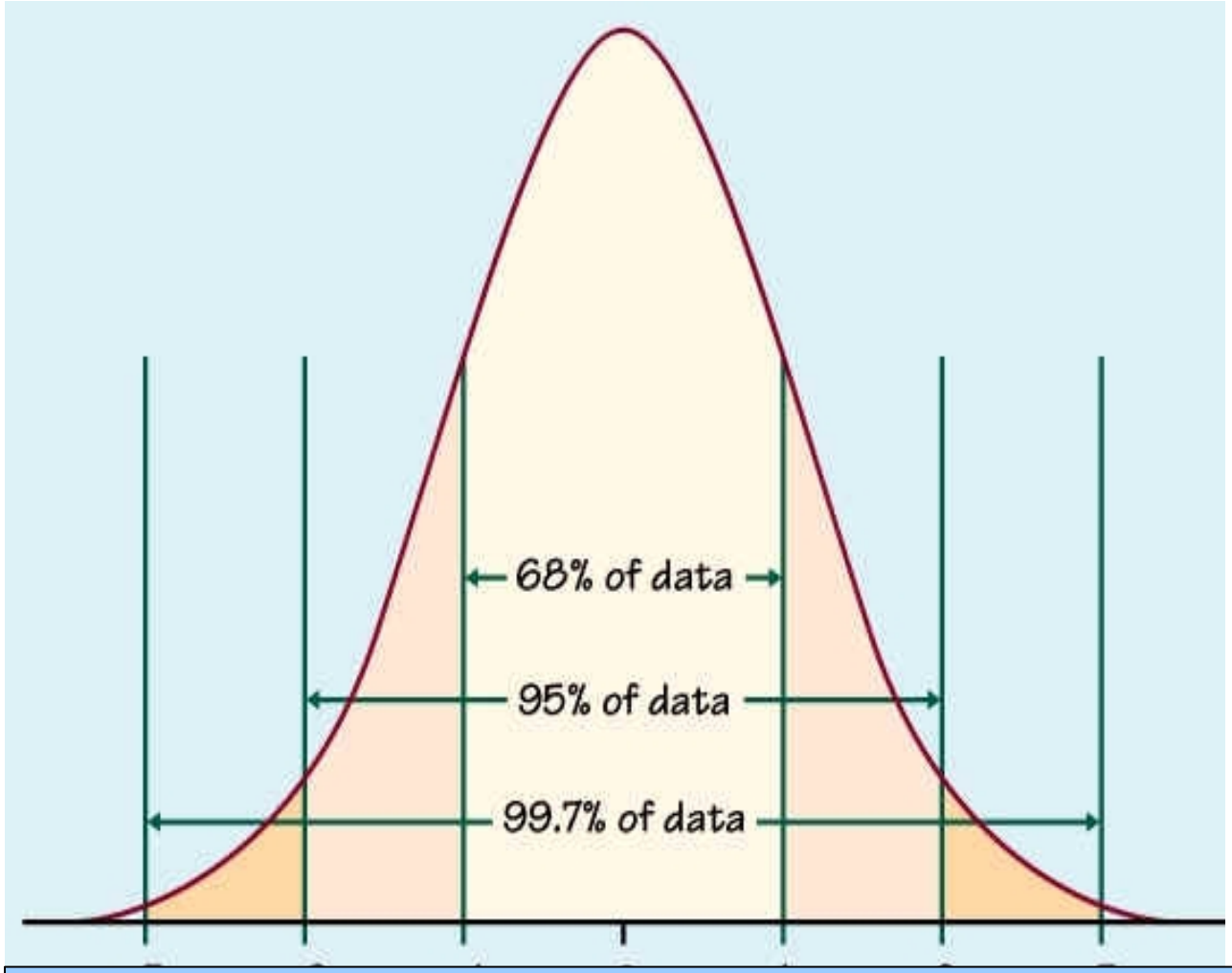
	A	R	N	D	C	Q	E	G	H	I	L	K	M	F	P	S	T	W	Y	V	X
A	4																				
R	-1	5																			
N	-2	0	6																		
D	-2	-2	1	6																	
C	0	-3	-3	-3	9																
Q	-1	1	0	0	-3	5															
E	-1	0	0	2	-4	2	5														
G	0	-2	0	-1	-3	-2	-2	6													
H	-2	0	1	-1	-3	0	0	-2	8												
I	-1	-3	-3	-3	-1	-3	-3	-4	-3	4											
L	-1	-2	-3	-4	-1	-2	-3	-4	-3	2	4										
K	-1	2	0	-1	-3	1	1	-2	-1	-3	-2	5									
M	-1	-1	-2	-3	-1	0	-2	-3	-2	1	2	-1	5								
F	-2	-3	-3	-3	-2	-3	-3	-3	-1	0	0	-3	0	6							
P	-1	-2	-2	-1	-3	-1	-1	-2	-2	-3	-3	-1	-2	-4	7						
S	1	-1	1	0	-1	0	0	0	-1	-2	-2	0	-1	-2	-1	4					
T	0	-1	0	-1	-1	-1	-1	-2	-2	-1	-1	-1	-1	-2	-1	1	5				
W	-3	-3	-4	-4	-2	-2	-3	-2	-2	-3	-2	-3	-1	1	-4	-3	-2	11			
Y	-2	-2	-2	-3	-2	-1	-2	-3	2	-1	-1	-2	-1	3	-3	-2	-2	2	7		
V	0	-3	-3	-3	-1	-2	-2	-3	-3	3	1	-2	1	-1	-2	-2	0	-3	-1	4	
X	0	-1	-1	-1	-2	-1	-1	-1	-1	-1	-1	-1	-1	-1	-2	0	0	-2	-1	-1	-1

Positive for chemically similar substitution

Common amino acids have low weights

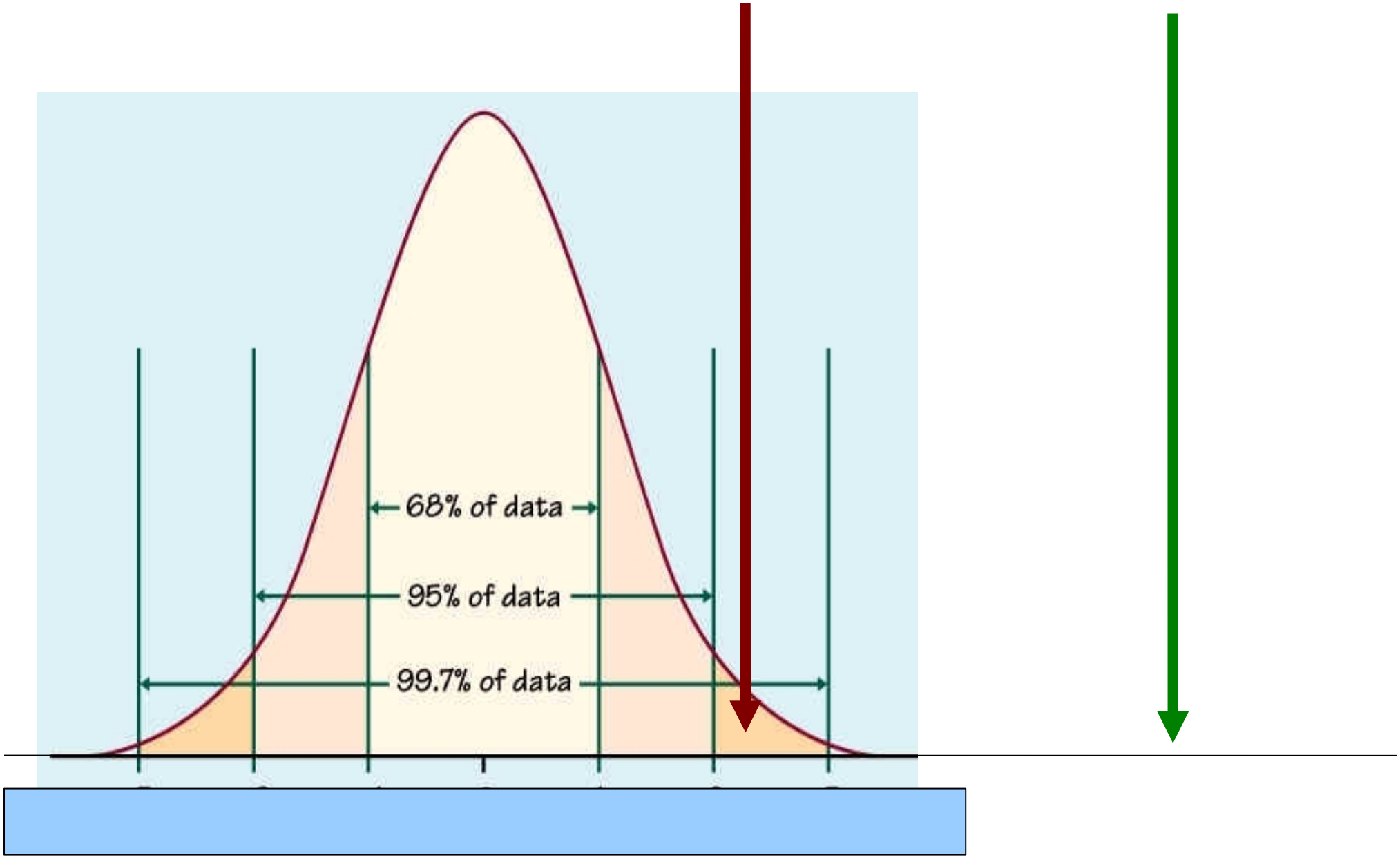
Rare amino acids have high weights

Distribution of scores in comparisons of **random***-sequences



* considering the representation of the different amino acids (nucleotides) in a DataBase

Your score



The significance of each alignment is computed as a P value or an E value

E value: Expectation value. The number of different alignents with scores equivalent to or better than S that are expected to occur in a database search by chance. The lower the E value, the more significant the score.

P value :The probability of an alignment occurring with the score in question or better. The p value is calculated by relating the observed alignment score, S, to the expected distribution of HSP scores from comparisons of random sequences of the same length and composition as the query to the database. The most highly significant P values will be those close to 0. P values and E values are different ways of representing the significance of the alignment.

Length of hit and query

Normalized score

$$E = m\,n\,2^{-S'}$$

E-value (Expectation value)= the number of sequences that would be expected to have that **score** (or higher) if the query sequence were compared against a **database** containing unrelated sequences

E-value= ranges from 0 to the number of sequences in the DB, **and depends on the Database!!!**

```
> ref|NP_114344.1|  NADH dehydrogenase subunit 5 [Macaca sylvanus]
Length=603

 GENE ID: 803075 ND5 | NADH dehydrogenase subunit 5 [Macaca sylvanus]
(10 or fewer PubMed links)

 Score =  796 bits (2056),  Expect = 0.0, Method: Compositional matrix adjust.
 Identities = 438/564 (77%), Positives = 478/564 (84%), Gaps = 0/564 (0%)

Query  24   VNPNKKNSYPHYVKSIVASTFIISLFPTTMFMCLDQEVIISNWHWATTQTTQLSLSFKLD  83
            +NPNKK+ YP+YVK+ V   FI SL  TT++M L+QE II +WHW  TQT  L+LSFKLD
Sbjct  24   INPNKKHLYPNYVKTAVMYAFITSLSSTTLYMFLNQETIIWSWHWMMTQTLSLTLSFKLD  83

Query  84   YFSMMFIPVALFVTWSIMEFSLWYMNSDPNINQFFKYLLIFLITMLILVTANNLFQLFIG  143
            YFSMMF P+AL  TWSIMEFSLWYM+SDPNI+QFFKYLLIFLITMLILVTANNLFQ FIG
Sbjct  84   YFSMMFTPIALLTTWSIMEFSLWYMSSDPNIDQFFKYLLIFLITMLILVTANNLFQFFIG  143

Query  144  WEGVGIMSFLLISWWYARADANTAAIQAVLYNRIGDIGFILALAWFILHSNSWDPQQMAL  203
            WEG+GIMSFLLISWW+AR DANTAAIQA+LYNRIGDIG IL + WF+LH NSWD QQM
Sbjct  144  WEGMGIMSFLLISWWHARTDANTAAIQAILYNRIGDIGLILTMTWFLLHYNSWDFQQMLA  203
```

E-value

Coverage over the query

Other aspects in Blast searches

- E-value depends on database (specially important when locally searching in small databases)
- Use of Low complexity filtering
- Why multiple HSPs in a hit
- PSI-Blast, HMMER searches

End of the detour

From homology to orthology

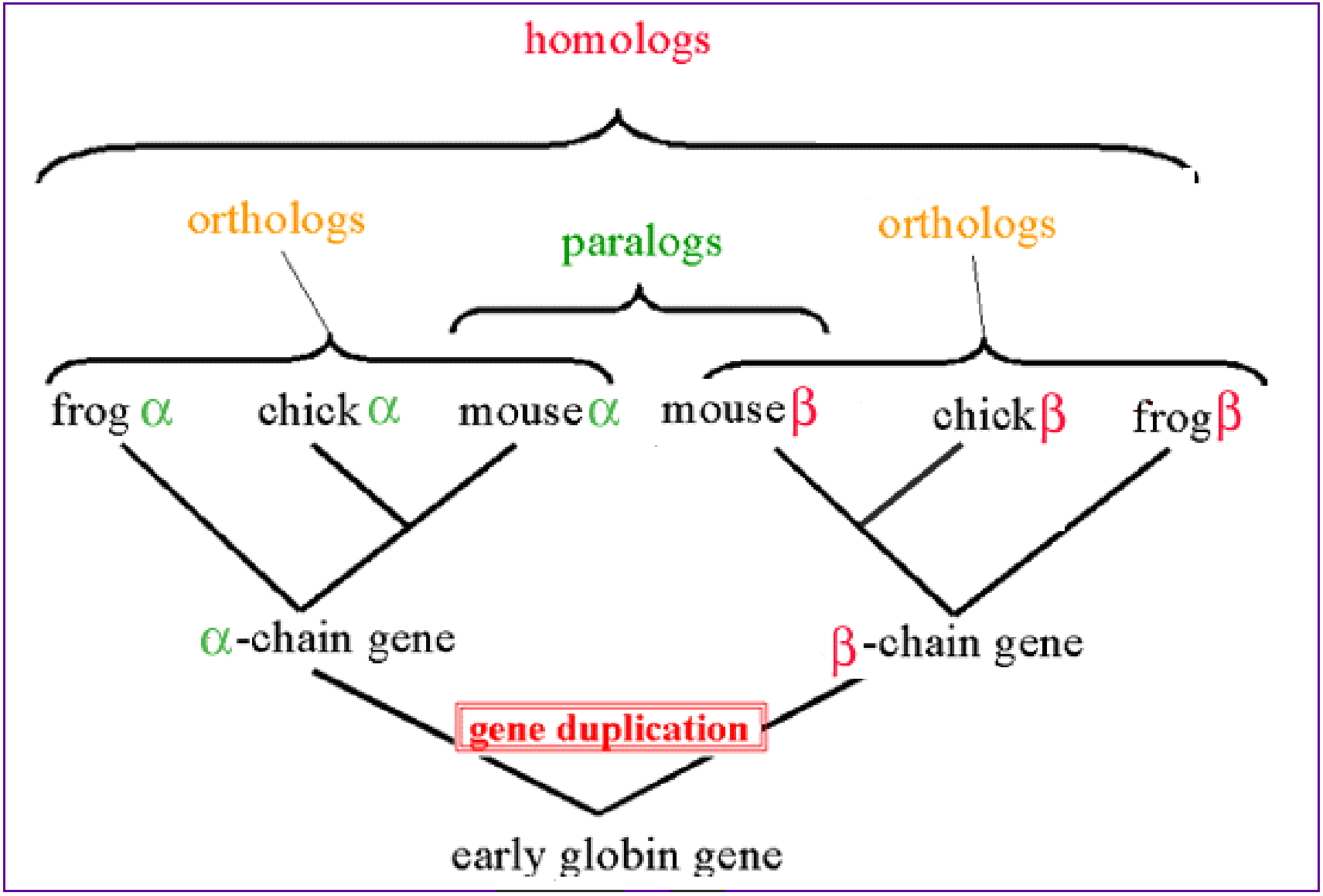
- Homologues are sequences derived from a common ancestor...

- What are then orthologues?.... and paralogues?

Original definition of orthology and paralogy by Walter Fitch (1970, Systematic Zoology 19:99-113):

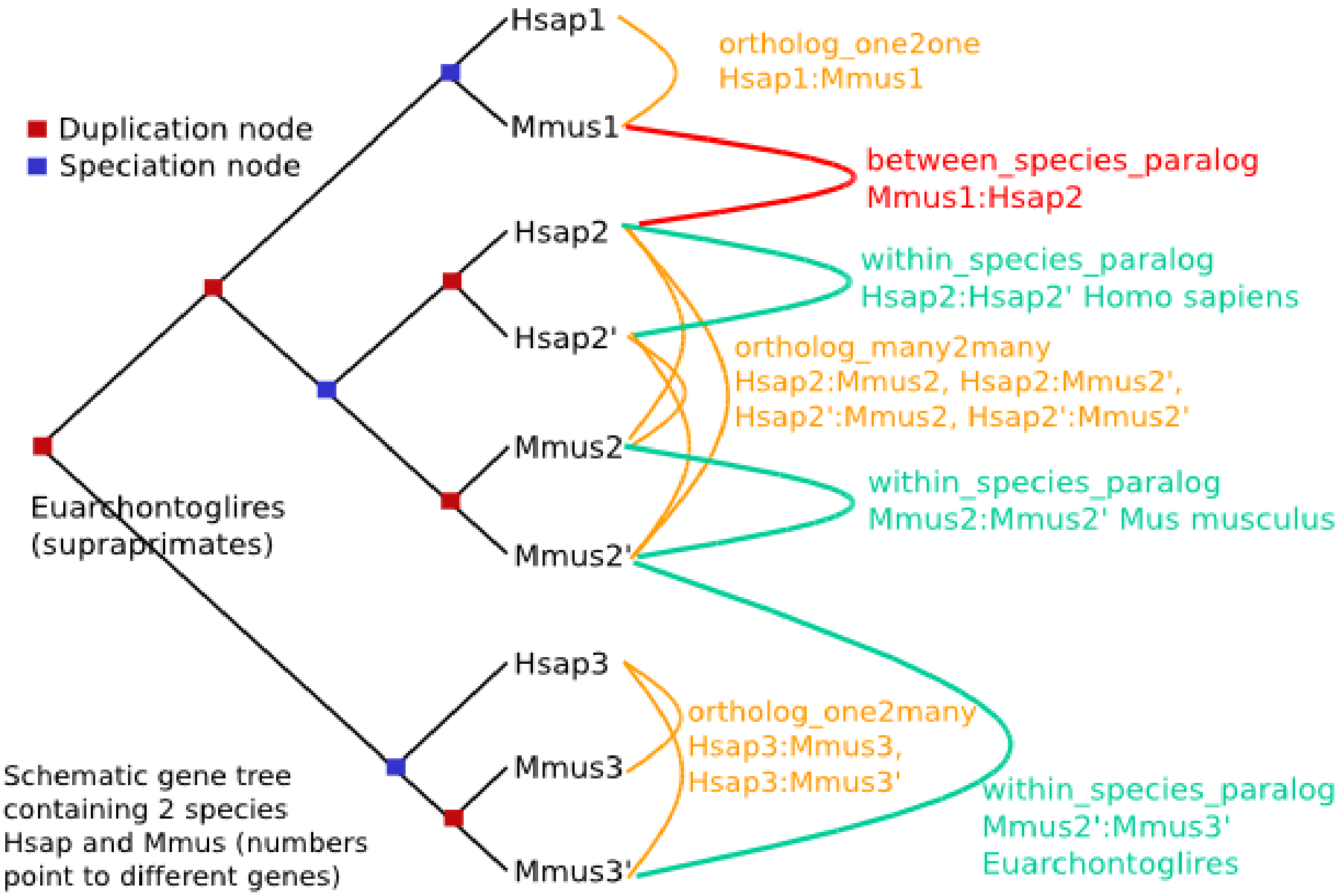
*"Where the homology is **the result of gene duplication** so that both copies have descended side by side during the history of an organism, (for example, alpha and beta hemoglobin) the genes should be called **paralogous** (para = in parallel).*

*Where the homology is **the result of speciation** so that the history of the gene reflects the history of the species (for example alpha hemoglobin in man and mouse) the genes should be called **orthologous** (ortho = exact)."*


homologs
orthologs
paralogs
orthologs
frog α
chick α
mouse α
mouse β
chick β
frog β
α-chain gene
β-chain gene
gene duplication
early globin gene

Corollary:

- Orthology definition is purely on evolutionary terms (not functional, not synteny…)
- Orthology/paralogy defines a pair-wise relationship between two genes
- There is no limit on the number of orthologs or paralogs that a given gene can have (when more than one ortholog exist, there is nothing such as "*the true ortholog*",)
- Many-to-Many orthology relationships do exist (co-orthology)
- No limit on how ancient/recent is the ancestral relationship of orthologs and paralogs
- Orthology is non-transitive (as opposed to homology)



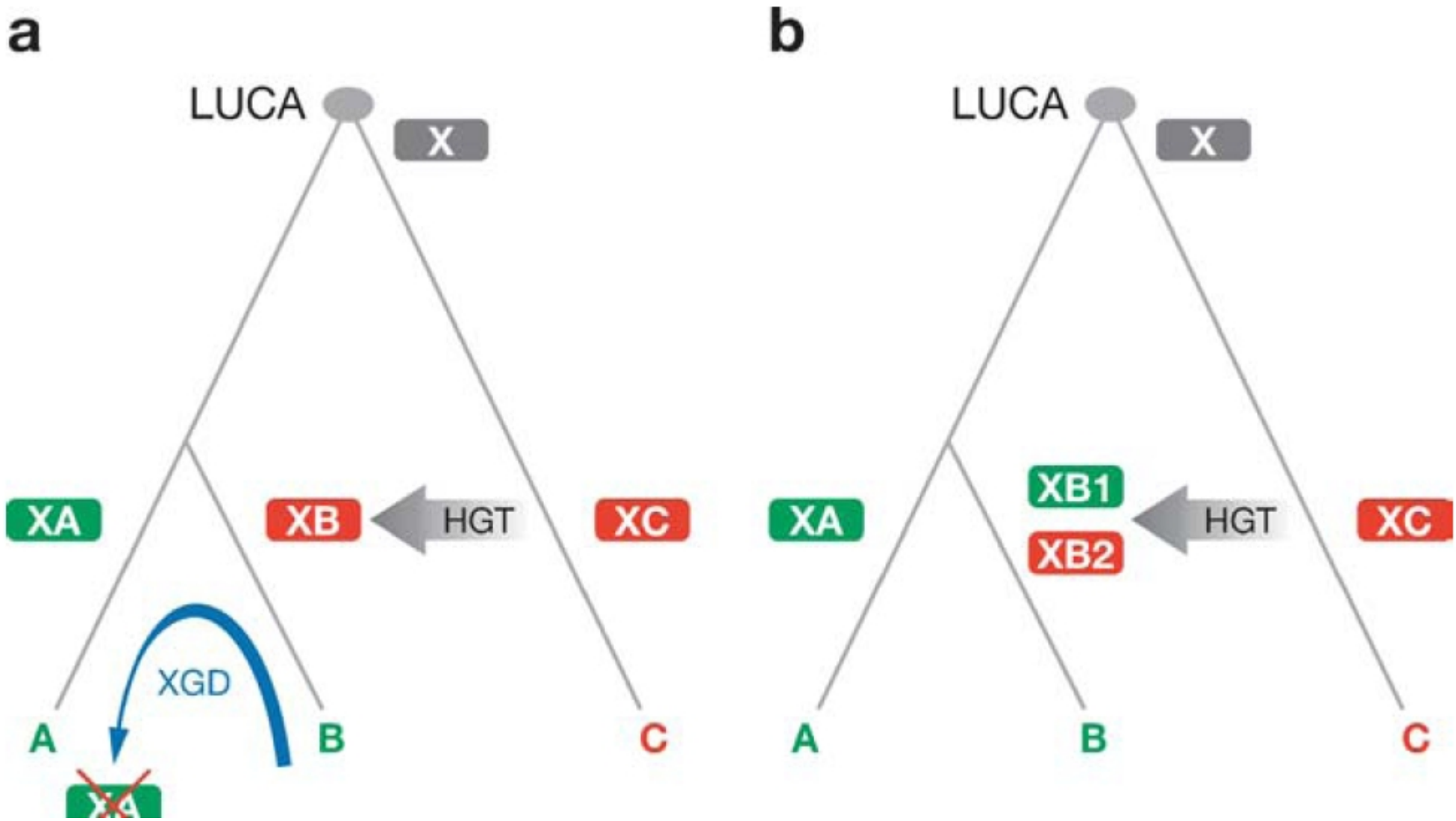
Schematic gene tree containing 2 species Hsap and Mmus (numbers point to different genes)

Additional useful definitions

- **In-paralogs and out-paralogs** (Sohnhammer and koonin): It is defined relative to a given speciation event. In-paralogs are derived from duplications occurred subsequent to the speciation event and are therefore specific of one lineage. Out-paralogs are paralogs emerged from duplications occurred before the speciation. (Important: if you change the speciation events these relationships change)

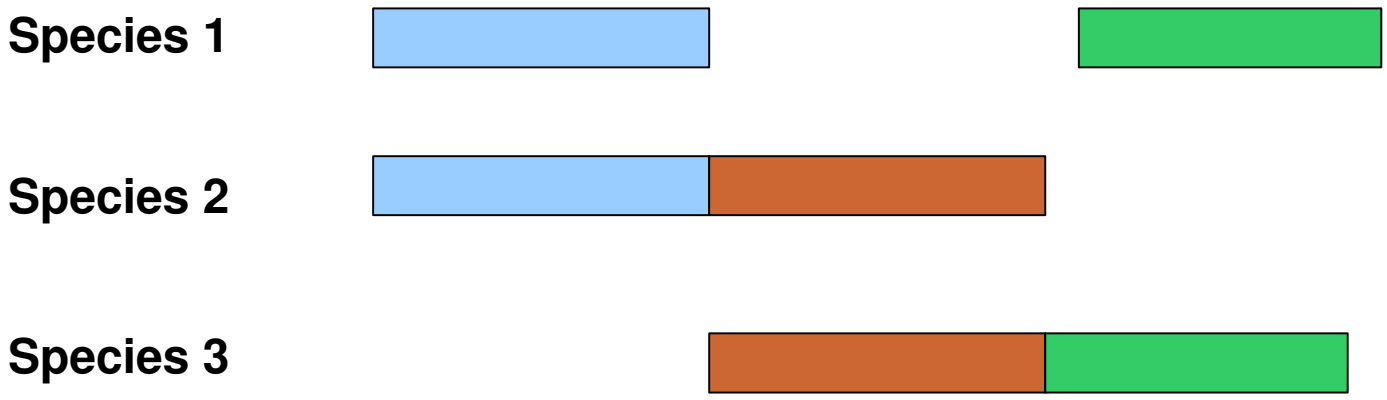
- **Orthologous group (~Orthogroup):** Also defined relative to a speciation event. It is the complete set of genes in one of the lineages formed by a speciation event. (it includes orthologs and in-paralogs, so not all the genes in an orthologous group are orthologs to each other)

The effect of HGT: Xenology and pseudoparalogy



Orthology and multi-domain proteins

- Orthology was defined at the level of genes, but this is not always the smallest level of evolution: domains do constitute smaller units of evolution, due to gene fusion/fission and recombination.



Why predicting orthology is important?

- **Important implications for phylogeny:** only sets of orthologous genes are expected to reflect the underlying species evolution (although there are many exceptions)

- The most exact way of **comparing two (or more) genomes** in terms of their gene content. Necessary to uncover how genomes evolve.

- Implications for **functional inference**: orthologs, as compared to paralogs, are more likely to share the same function

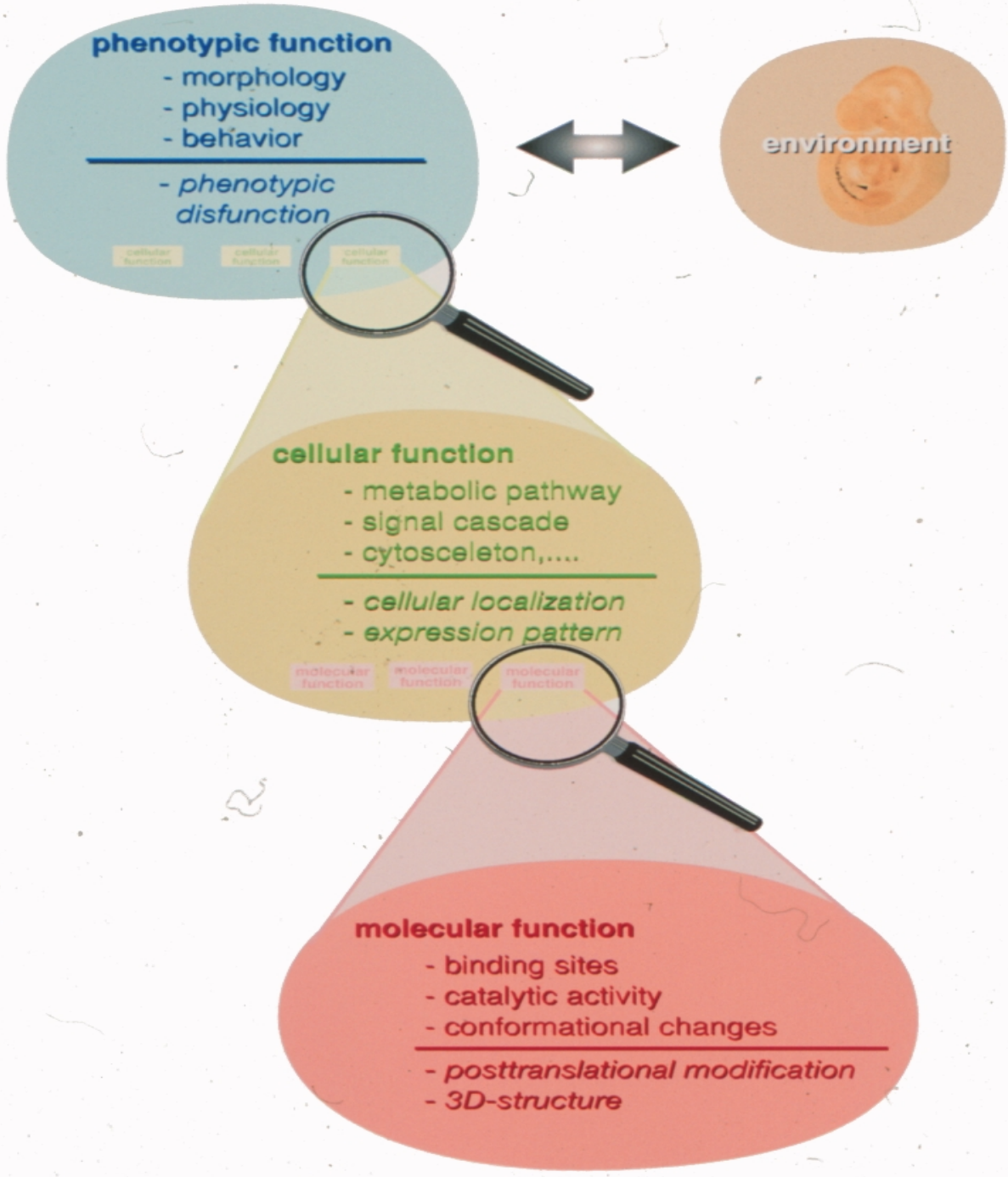
Why predicting orthology is important?

- **Important implications for phylogeny:** only sets of orthologous genes are expected to reflect the underlying species evolution (although there are many exceptions)

- The most exact way of **comparing two (or more) genomes** in terms of their gene content. Necessary to uncover how genomes evolve.

- Implications for **functional inference**: orthologs, as compared to paralogs, are more likely to share the same function

REALLY???, IS THIS TRUE IF SO, WHY IS THAT?


phenotypic function
- morphology
- physiology
- behavior
- phenotypic disfunction
environment
cellular function
- metabolic pathway
- signal cascade
- cytosceleton,....
- cellular localization
- expression pattern
molecular function
- binding sites
- catalytic activity
- conformational changes
- posttranslational modification
- 3D-structure

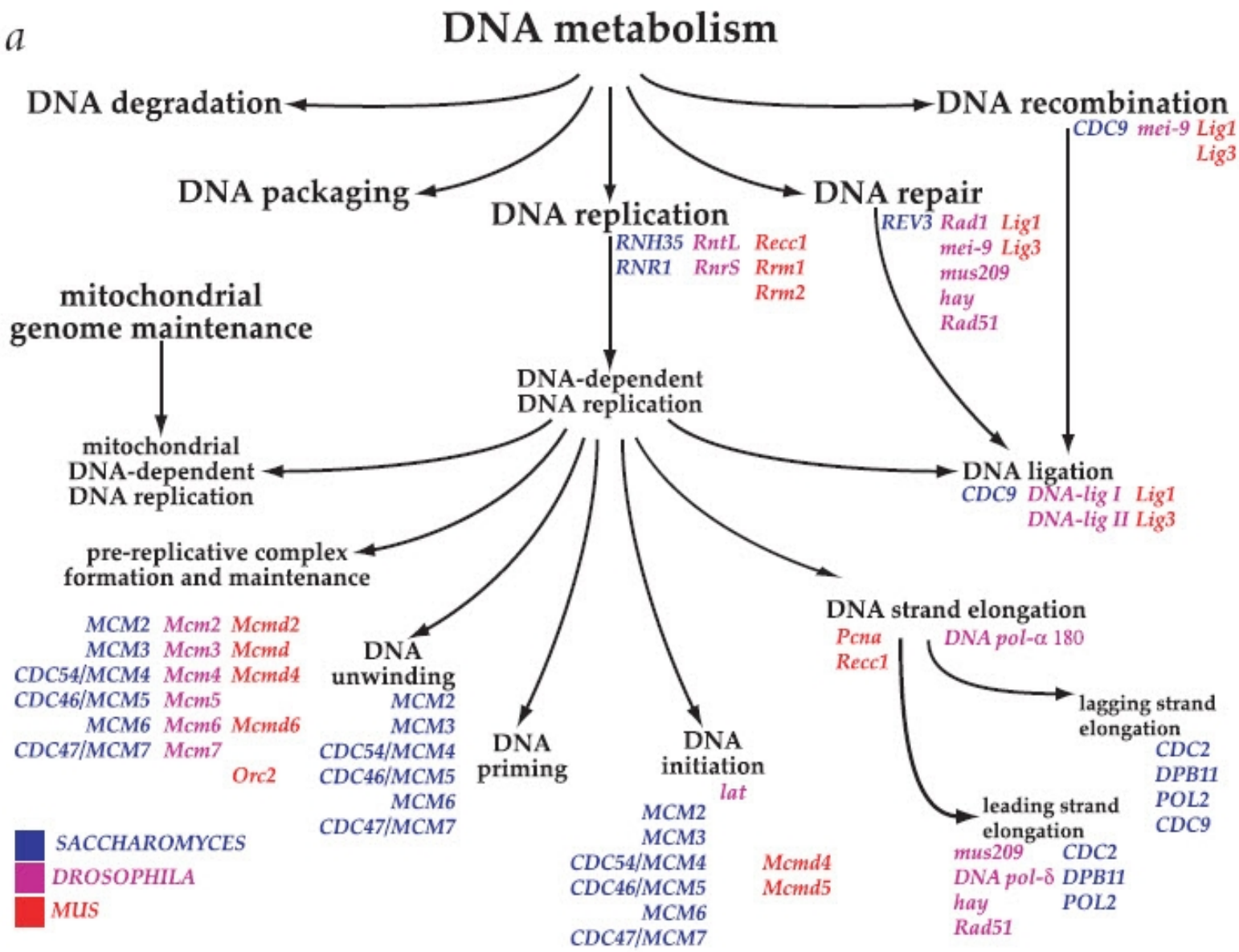
the Gene Ontology

Downloads

Accession, Term		Ontology	Qualifier	Evidence
GO:0006915 : apoptotic process	10407 gene products view in tree	biological process		ISS With UniProtKB:P04637
GO:0002326 : B cell lineage commitment	34 gene products view in tree	biological process		IEA With Ensembl:ENSMUSP00000
GO:0007569 : cell aging	878 gene products view in tree	biological process		ISS With UniProtKB:P04637
GO:0035690 : cellular response to drug	1521 gene products view in tree	biological process		IEA With Ensembl:ENSP00000

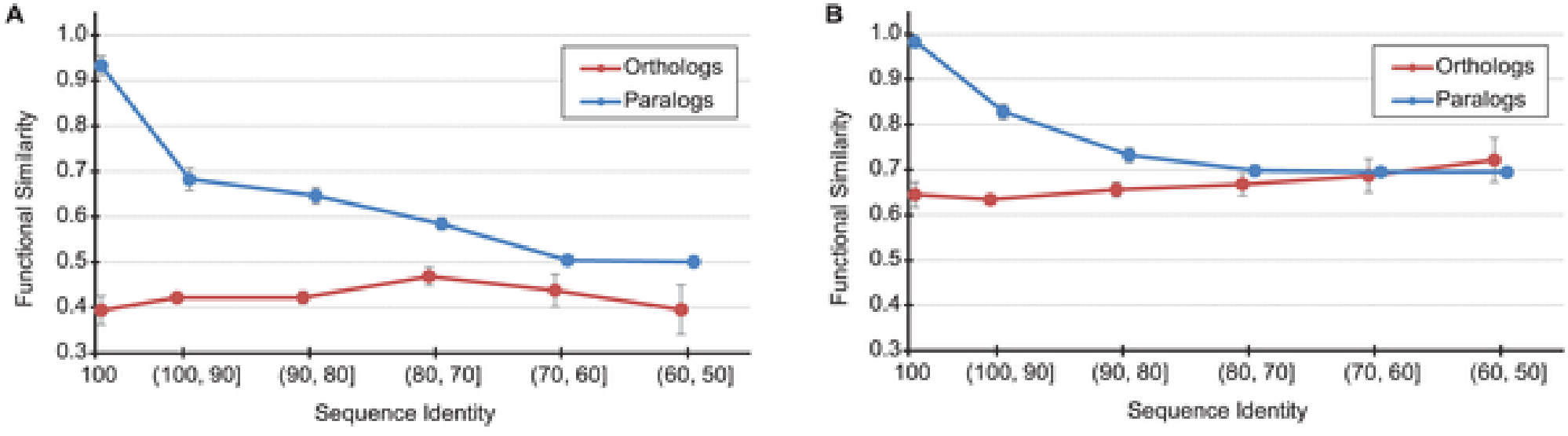
GO:0006915 (Apoptotic process)

A programmed cell death process which begins when a cell receives an internal (e.g. DNA damage) or external signal (e.g. an extracellular death ligand), and proceeds through a series of biochemical events (signaling pathways) which typically lead to rounding-up of the cell, retraction of pseudopodes, reduction of cellular volume (pyknosis), chromatin condensation, nuclear fragmentation (karyorrhexis), plasma membrane blebbing and fragmentation of the cell into apoptotic bodies. The process ends when the cell has died. The process is divided into a signaling pathway phase, and an execution phase, which is triggered by the former.


a
DNA metabolism
DNA degradation
DNA packaging
DNA replication
RNH35 RntL Recc1
RNR1 RnrS Rrm1
Rrm2
DNA repair
REV3 Rad1 Lig1
mei-9 Lig3
mus209
hay
Rad51
DNA recombination
CDC9 mei-9 Lig1
Lig3
mitochondrial genome maintenance
DNA-dependent DNA replication
mitochondrial DNA-dependent DNA replication
pre-replicative complex formation and maintenance
MCM2 Mcm2 Mcmd2
MCM3 Mcm3 Mcmd
CDC54/MCM4 Mcm4 Mcmd4
CDC46/MCM5 Mcm5
MCM6 Mcm6 Mcmd6
CDC47/MCM7 Mcm7
Orc2
DNA unwinding
MCM2
MCM3
CDC54/MCM4
CDC46/MCM5
MCM6
CDC47/MCM7
DNA priming
DNA initiation
lat
MCM2
MCM3
CDC54/MCM4
CDC46/MCM5
MCM6
CDC47/MCM7
Mcmd4
Mcmd5
DNA ligation
CDC9 DNA-lig I Lig1
DNA-lig II Lig3
DNA strand elongation
Pcna
Recc1
DNA pol-α 180
lagging strand elongation
CDC2
DPB11
POL2
CDC9
leading strand elongation
mus209 CDC2
DNA pol-δ DPB11
hay POL2
Rad51
SACCHAROMYCES
DROSOPHILA
MUS

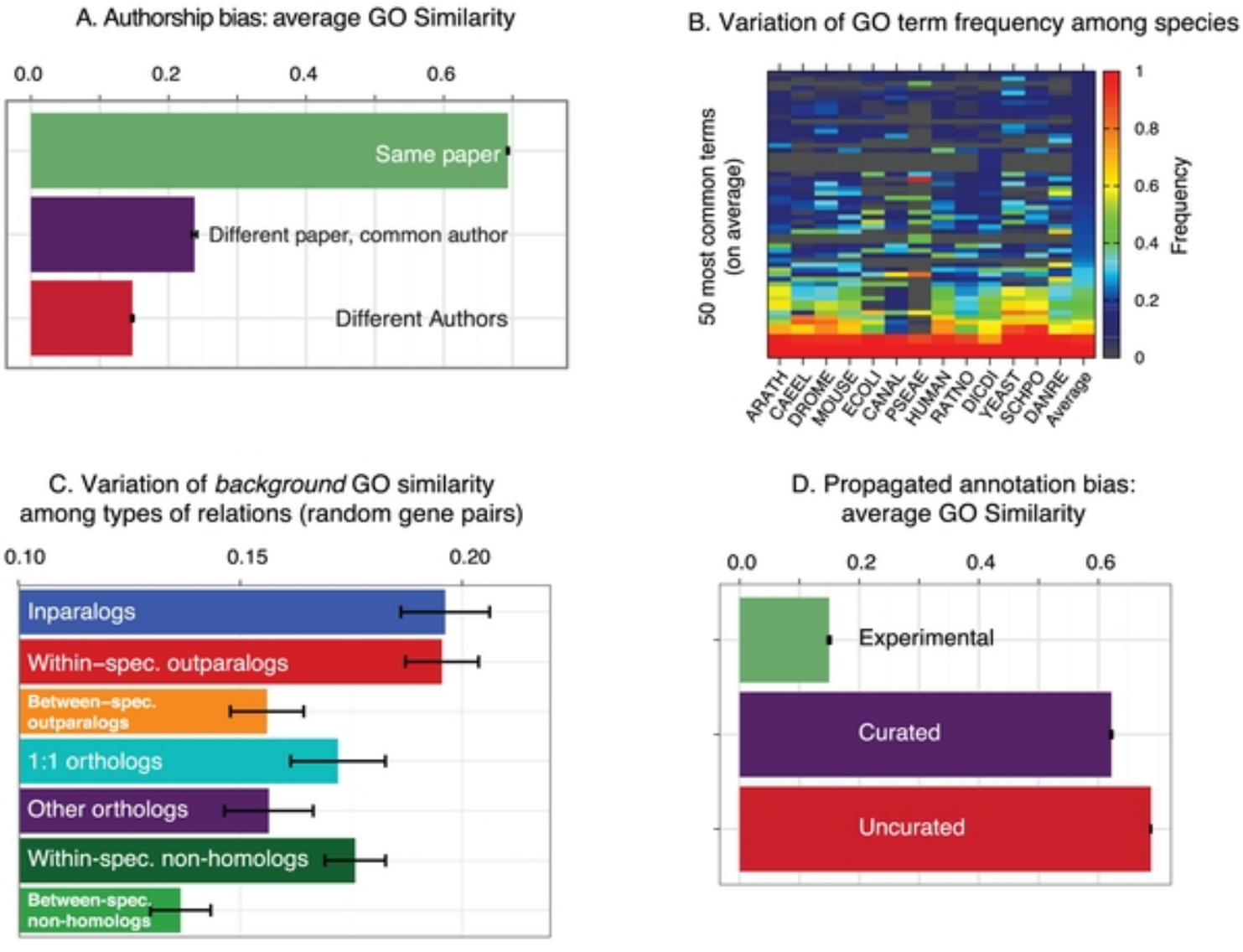
Do orthologs have more similar GO terms than paralogs?

Figure 1. The relationship between functional similarity and sequence identity for human-mouse orthologs (red) and all paralogs (blue).



Nehrt NL, Clark WT, Radivojac P, Hahn MW (2011) Testing the Ortholog Conjecture with Comparative Functional Genomic Data from Mammals. PLoS Comput Biol 7(6): e1002073. doi:10.1371/journal.pcbi.1002073
http://www.ploscompbiol.org/article/info:doi/10.1371/journal.pcbi.1002073

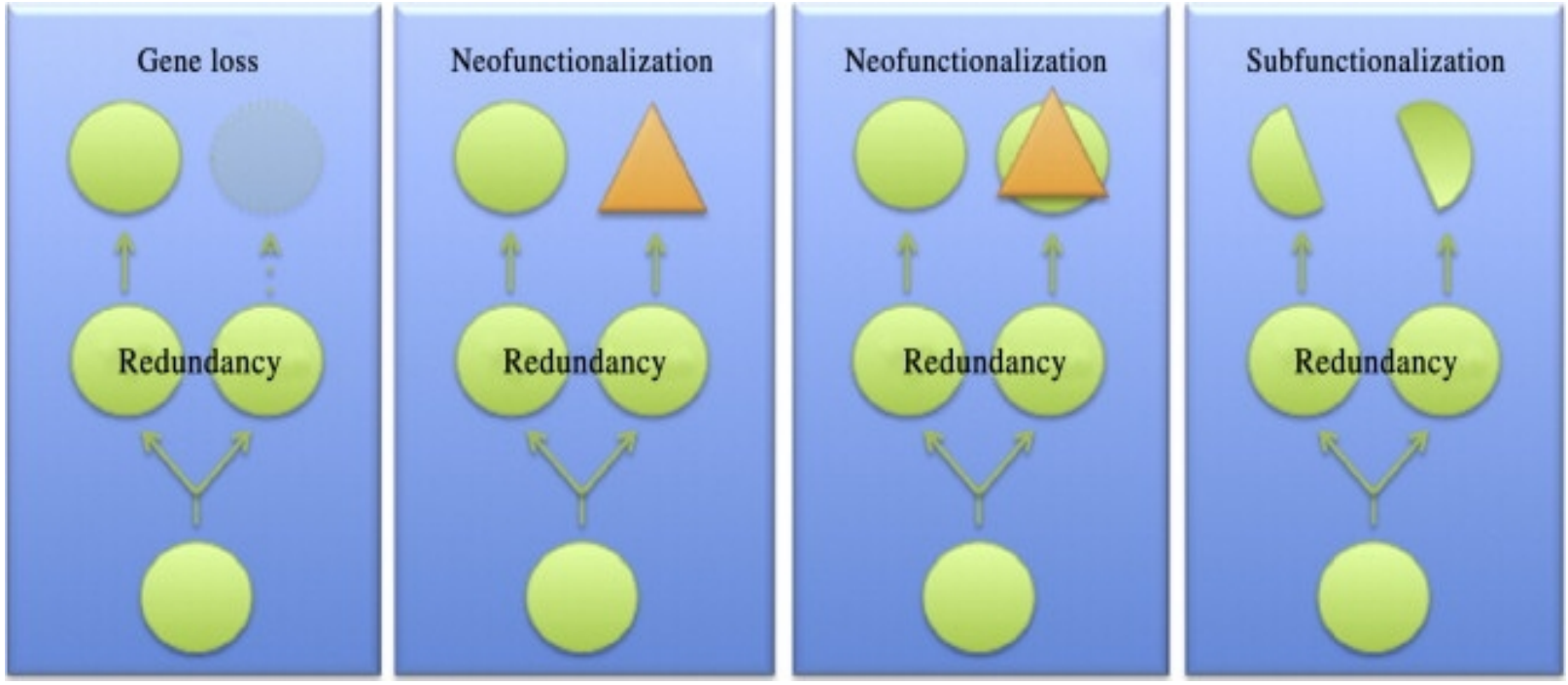
Figure 1. Potential confounding factors in GO analyses.

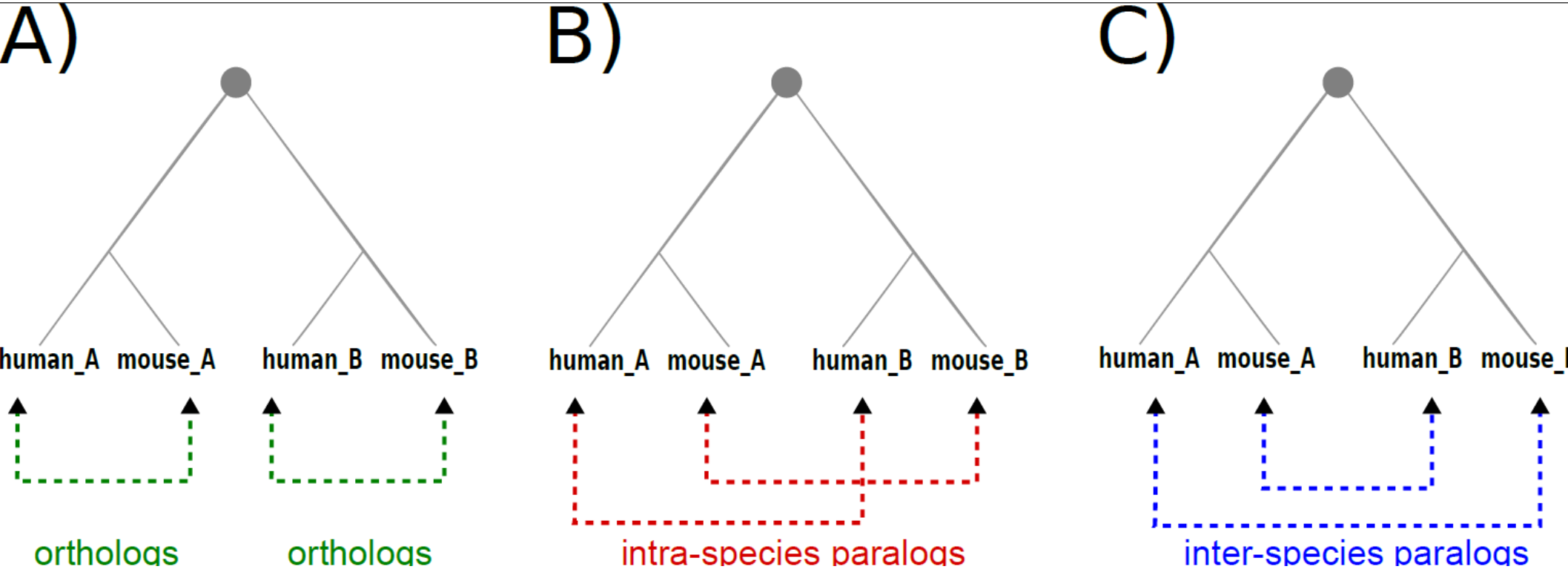


Altenhoff AM, Studer RA, Robinson-Rechavi M, Dessimoz C (2012) Resolving the Ortholog Conjecture: Orthologs Tend to Be Weakly, but Significantly, More Similar in Function than Paralogs. PLoS Comput Biol 8(5): e1002514. doi:10.1371/journal.pcbi.1002514
http://www.ploscompbiol.org/article/info:doi/10.1371/journal.pcbi.1002514

Orthologs do **tend** to have a more similar function because duplications promote functional divergence.

However, orthologs do also may vary their functions with time.



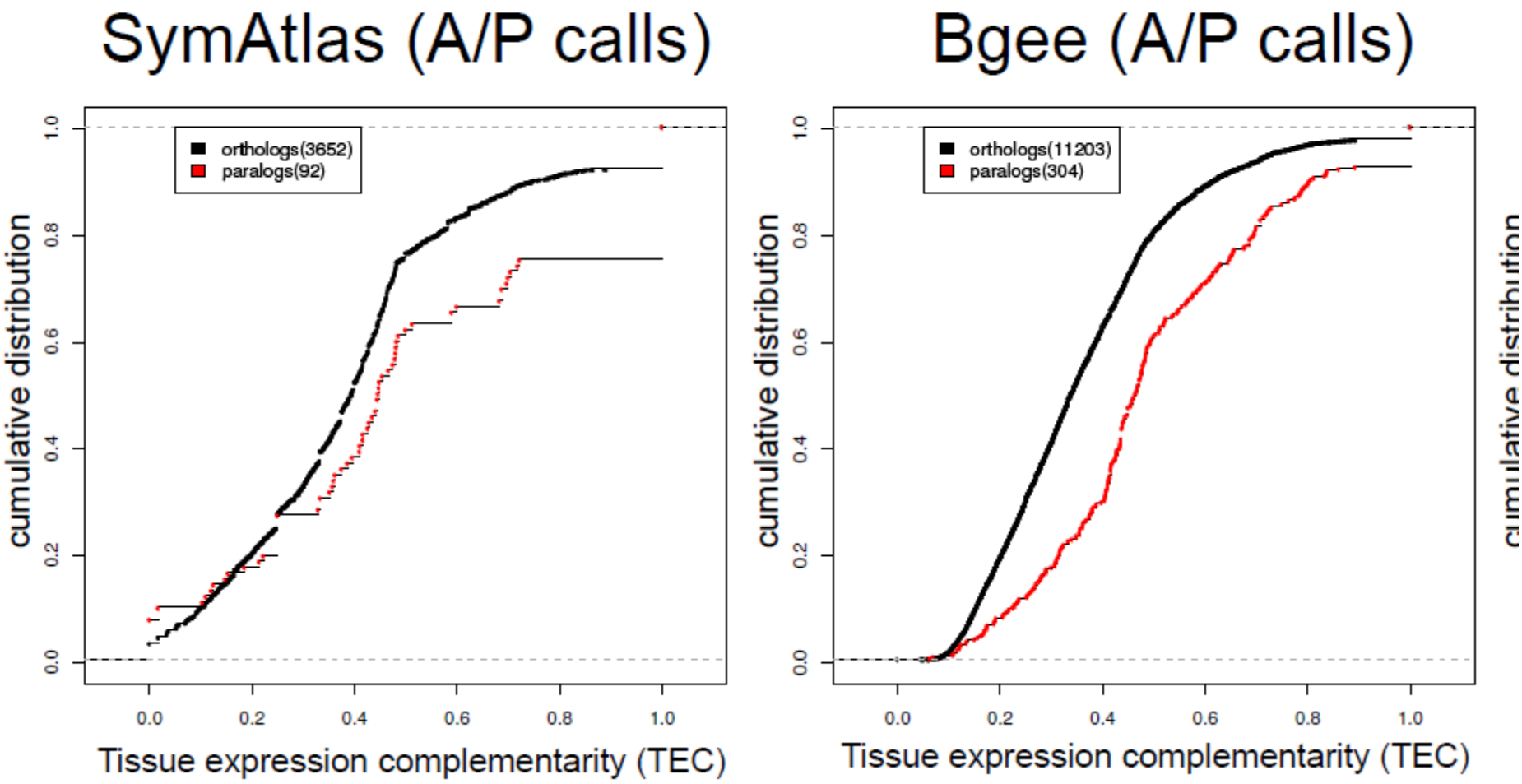


Comparison of differences in tissue-specific patterns of expression across orthologs and paralogs.

Evidence for short-time divergence and long-time conservation of tissue-specific expression after gene duplication.

Huerta-Cepas J, Dopazo J, Huynen MA, Gabaldón T.

Brief Bioinform. 2011 Sep;12(5):442-8. doi: 10.1093/bib/bbr022


SymAtlas (A/P calls)
orthologs(3652)
paralogs(92)
cumulative distribution
Tissue expression complementarity (TEC)
Bgee (A/P calls)
orthologs(11203)
paralogs(304)
cumulative distribution
Tissue expression complementarity (TEC)

Barcelona Biomedical Research Park

CRG Centre de Regulació Genòmica

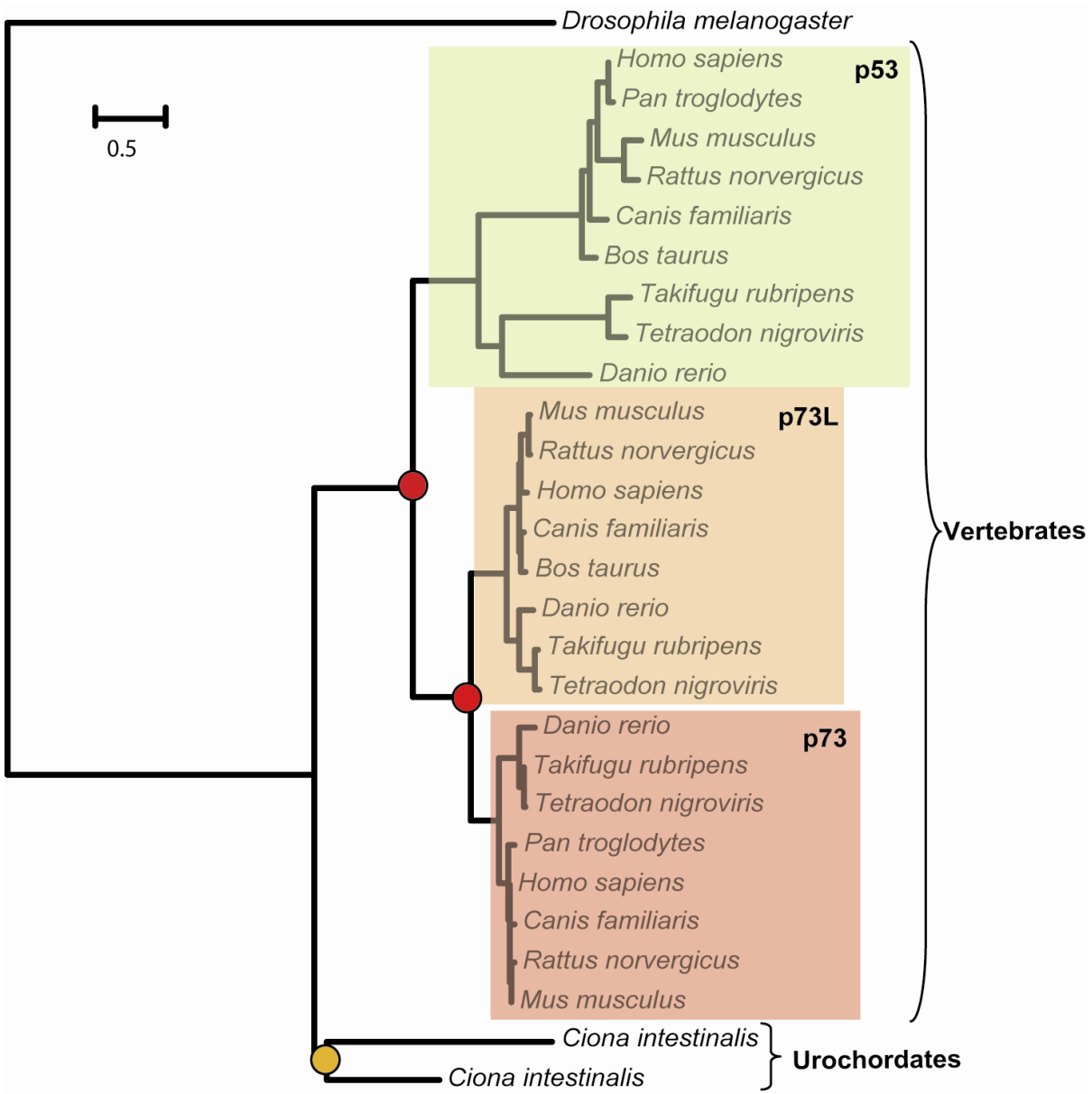
Orthology
Part II

Orthology prediction methods

Toni Gabaldón
Centre for Genomic Regulation (CRG), Barcelona

Classical approach: phylogenetic inference

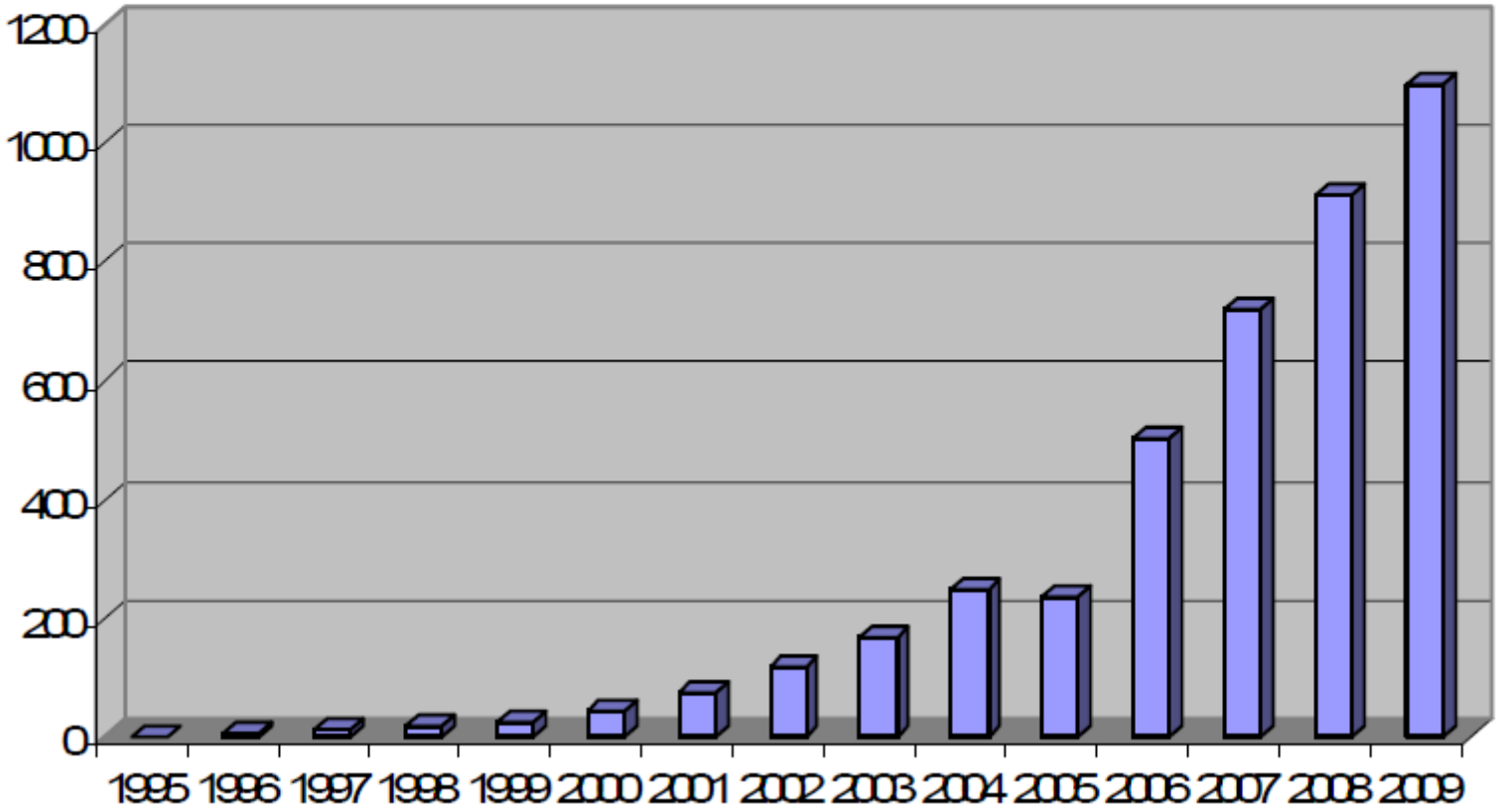
- Build a gene tree
- Compare to the species tree
- Infer duplications and speciation events
- Assign orthology and paralogy relationships accordingly

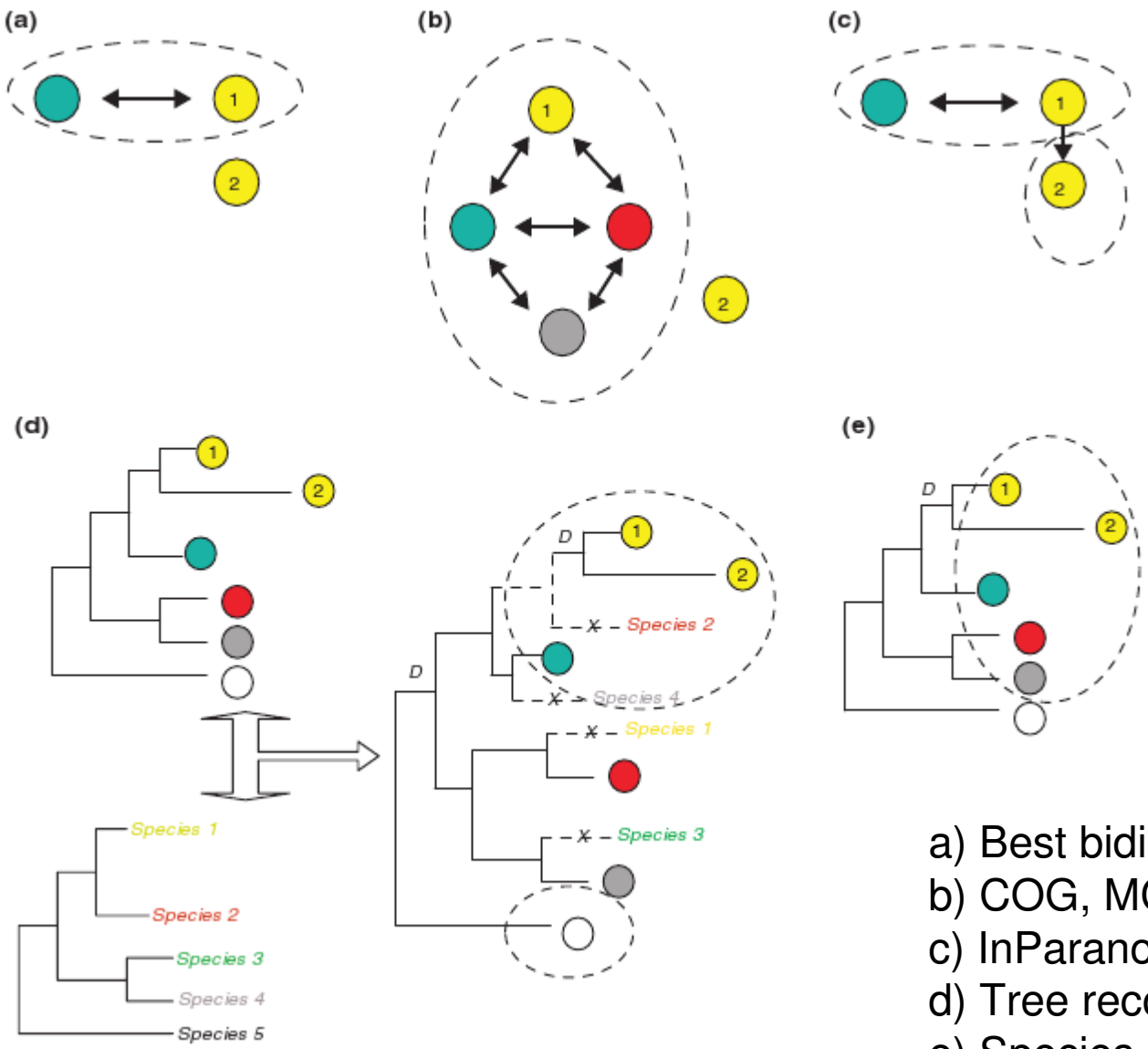

Drosophila melanogaster
0.5
p53
Homo sapiens
Pan troglodytes
Mus musculus
Rattus norvergicus
Canis familiaris
Bos taurus
Takifugu rubripens
Tetraodon nigroviris
Danio rerio
p73L
Mus musculus
Rattus norvergicus
Homo sapiens
Canis familiaris
Bos taurus
Danio rerio
Takifugu rubripens
Tetraodon nigroviris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

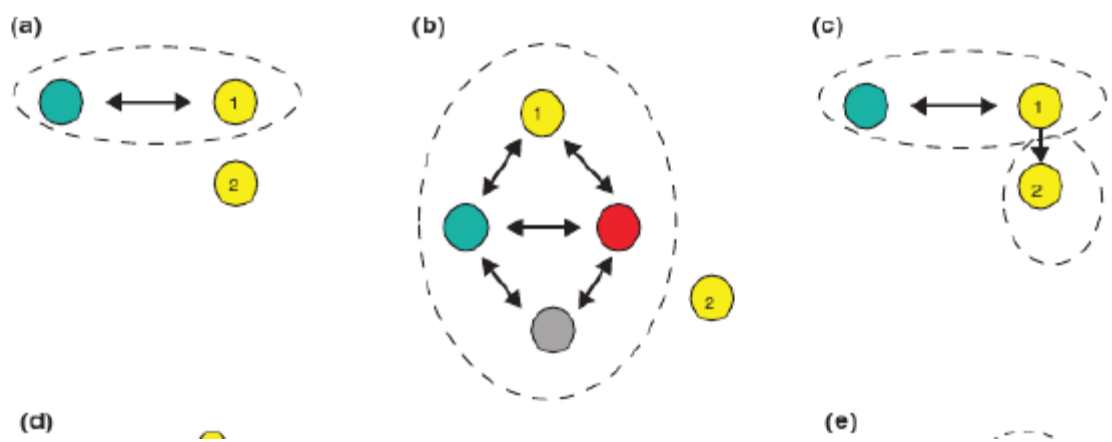
Gabaldón, T. *Genome Biology* (2008)

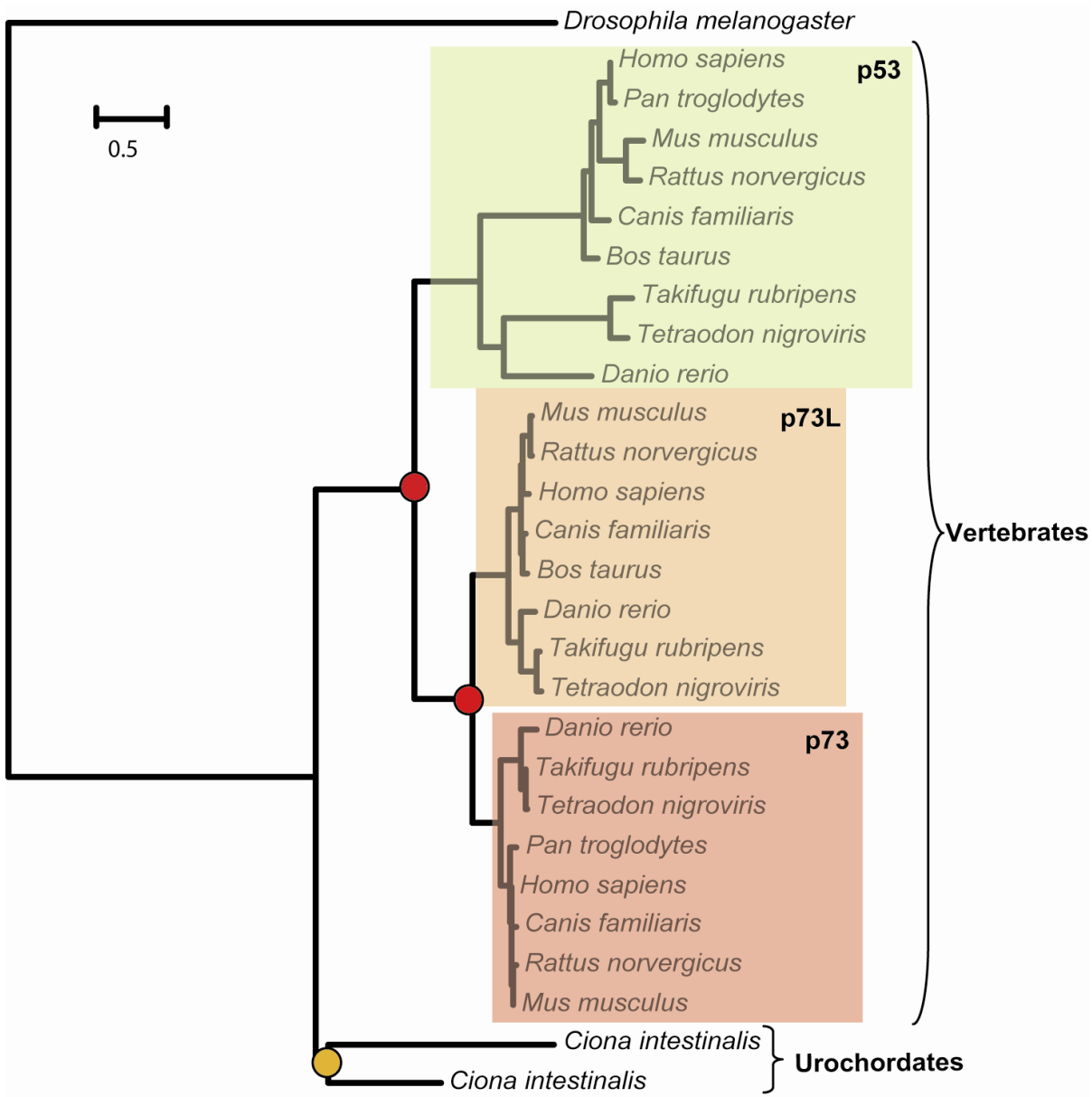
Similarity-based approaches (many more approaches):

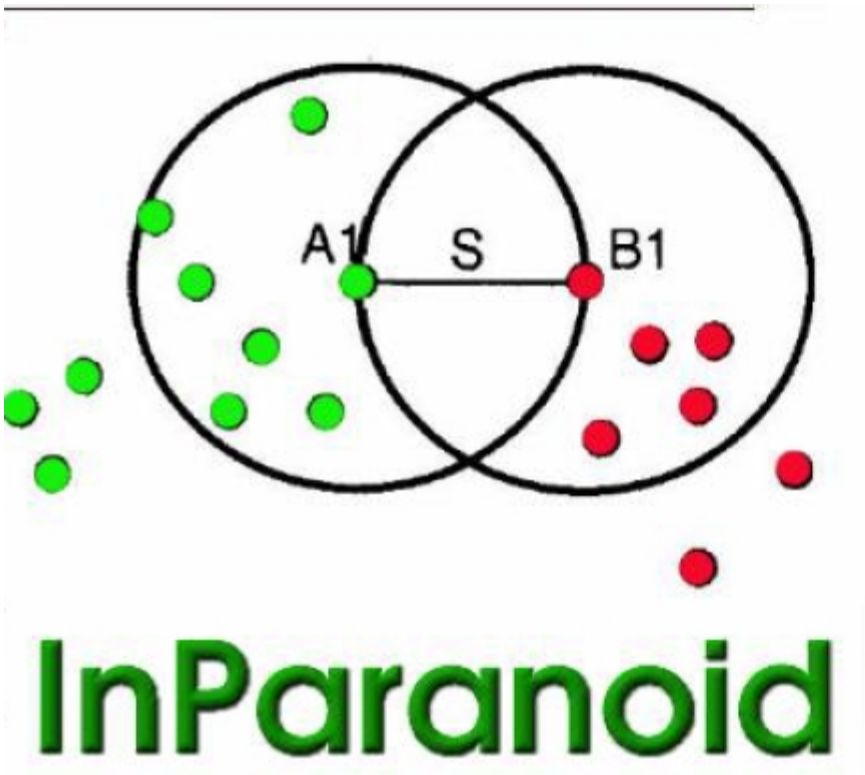
- **Best Reciprocal Hits**
- Detects all orthologies as one-to one. Highly affected by paralogy. Low rate of false positives but high rates of false negatives.

- The simplest and fastest method, still widely used

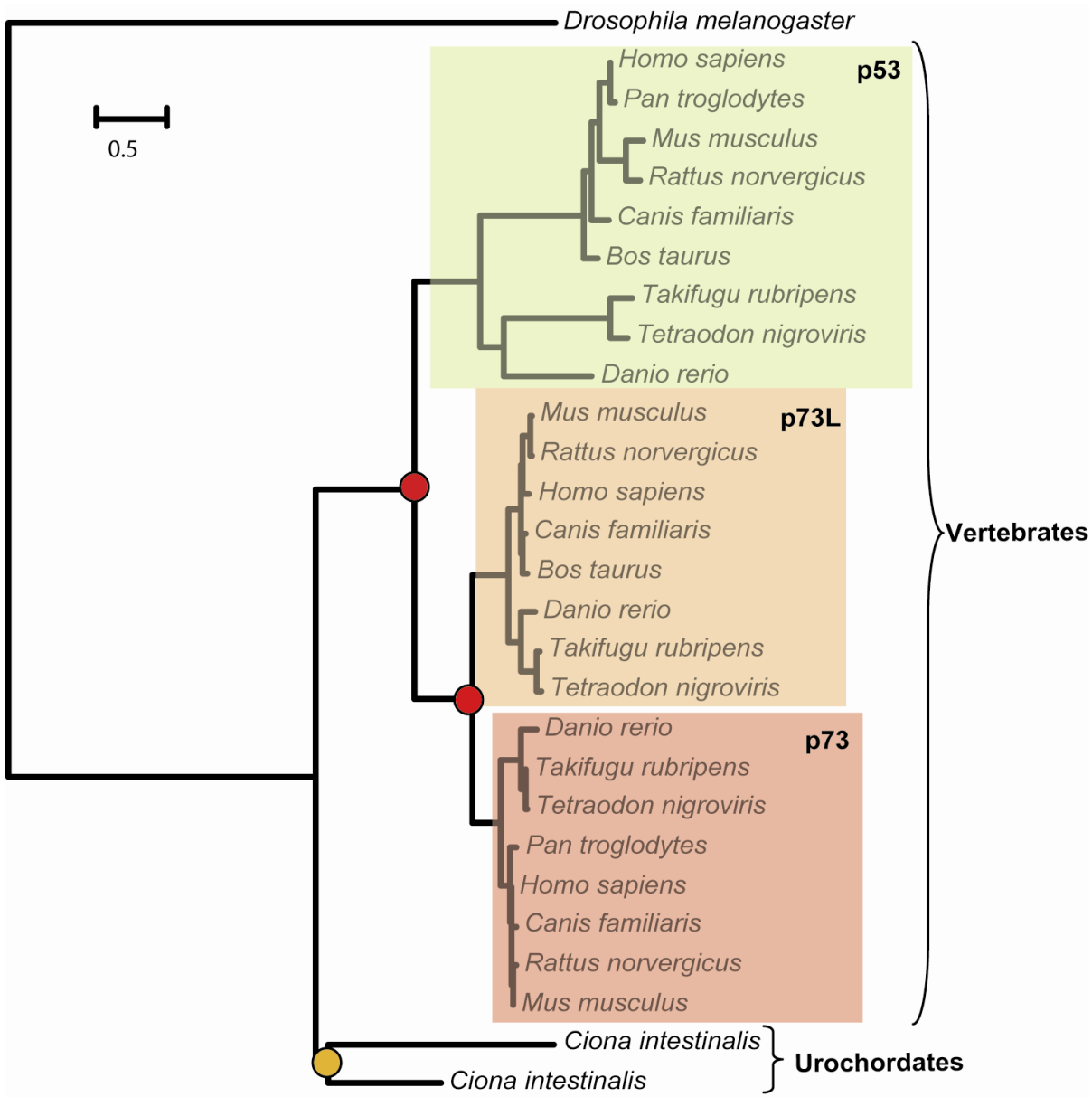
-




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



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

Note:

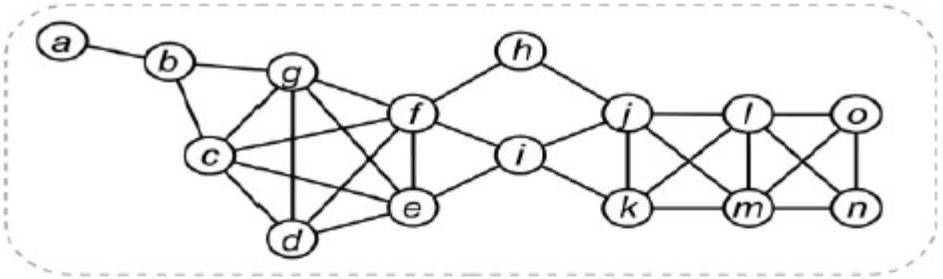
Definition of **in-** and **out-paralogues** require the specification of a given **speciation-node** of reference

COG-like
(used by many DBs like STRING)

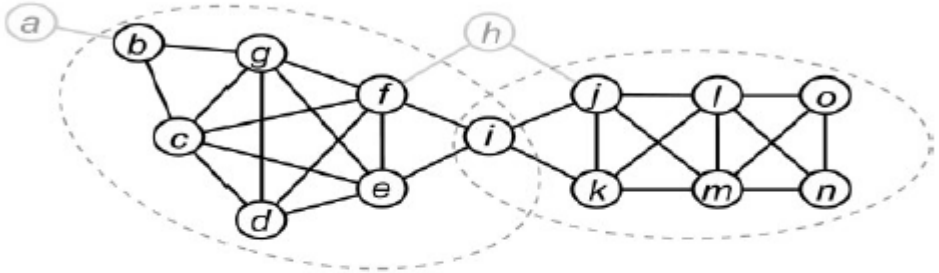
Exploits multi-species information. Predicts clusters of orthologous groups (in-paralogs) not all pairs in a cluster are paralogs.

Can be used at different stringent levels

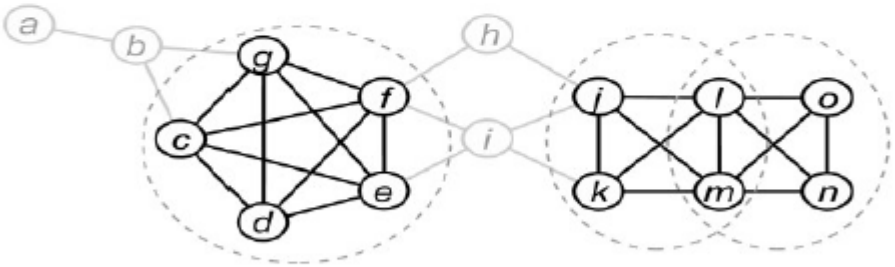
2



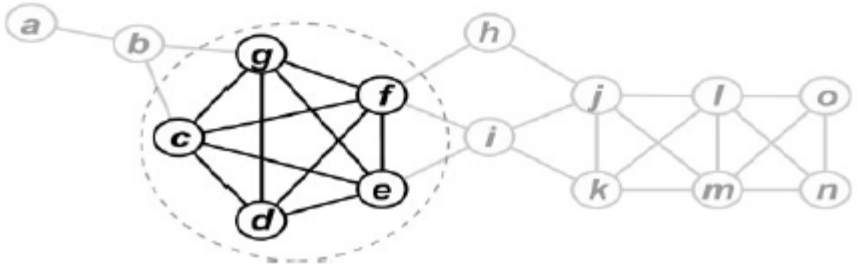
3



4



5



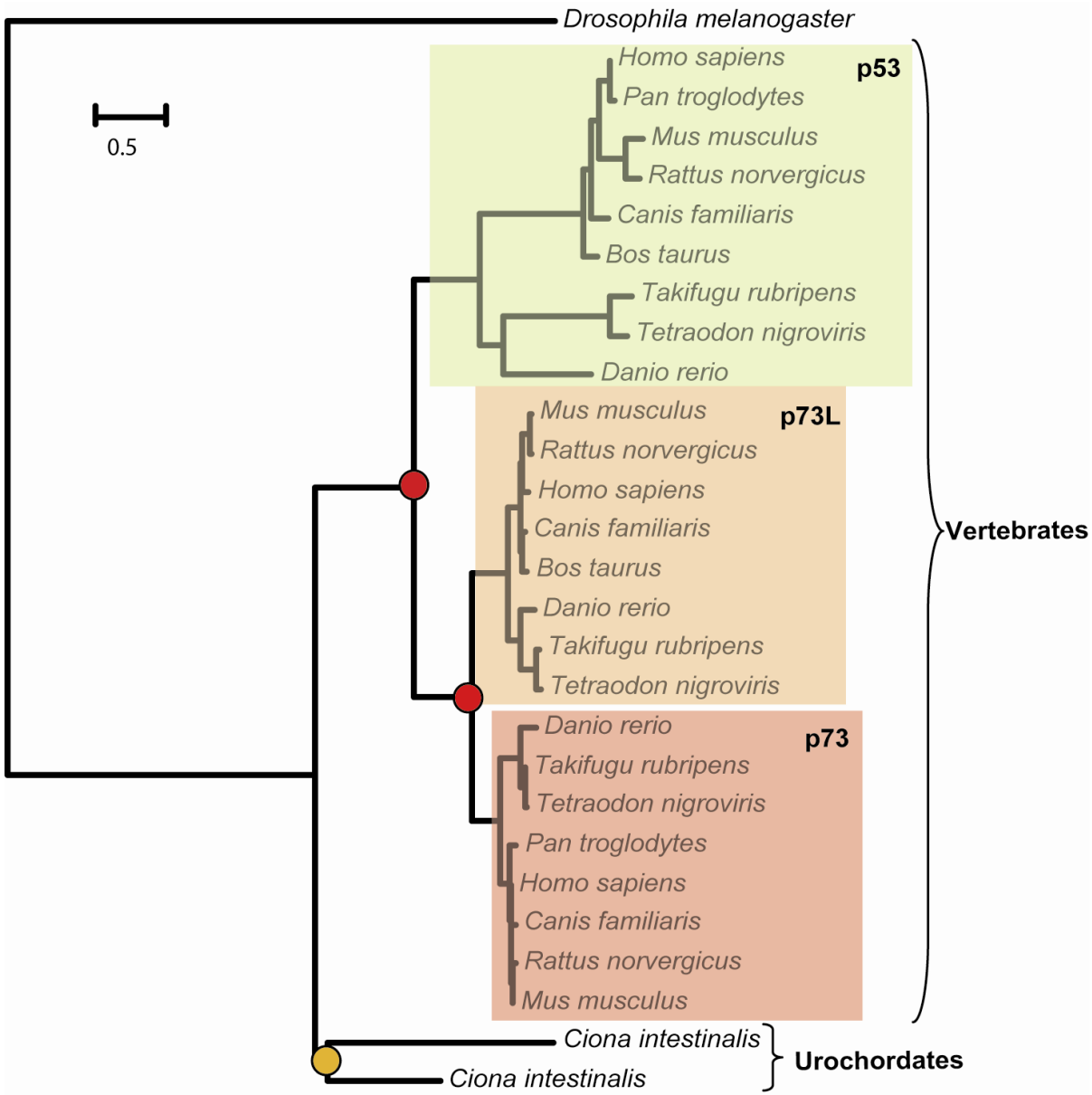
Clustering methods produce: orthologous groups

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)

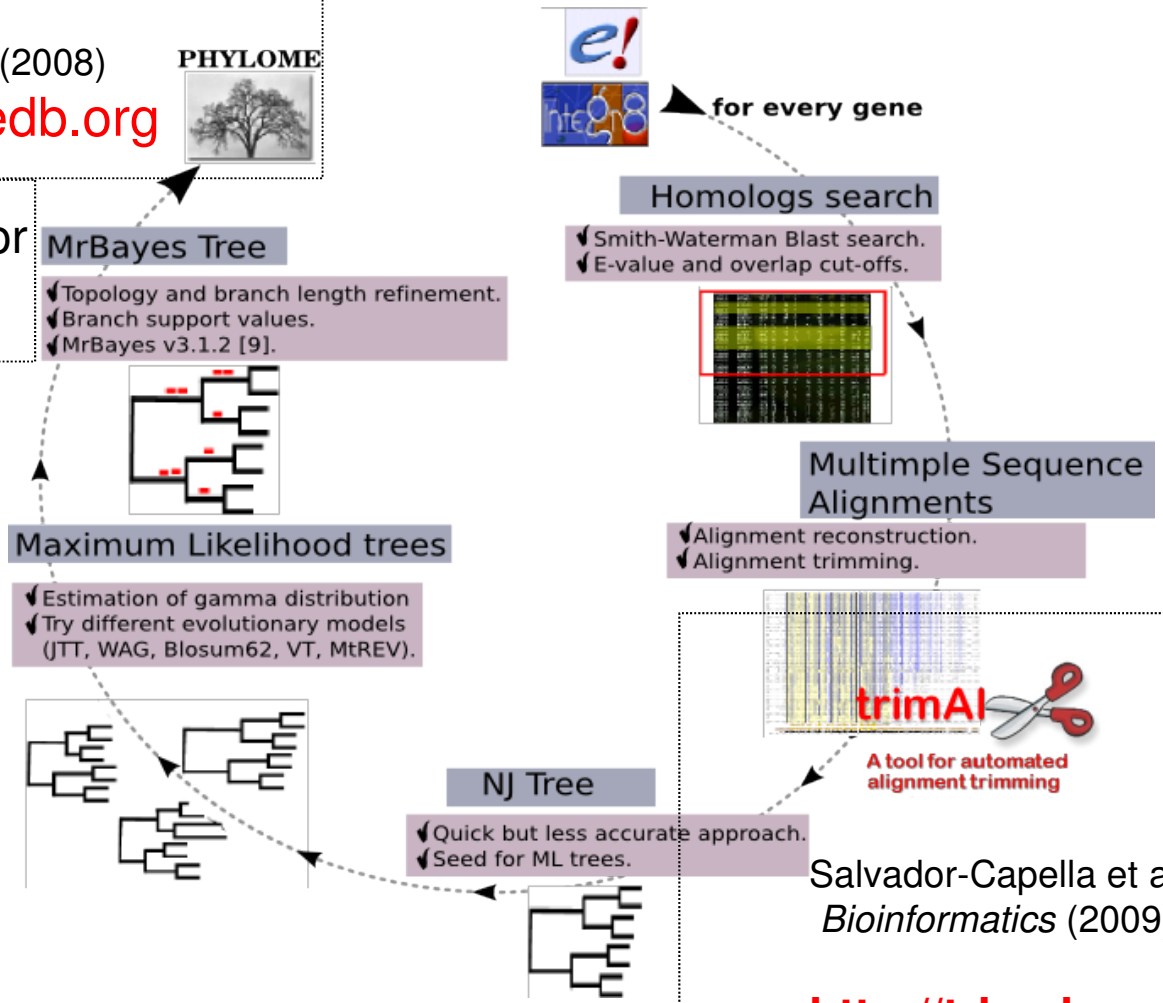
Abstract

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)

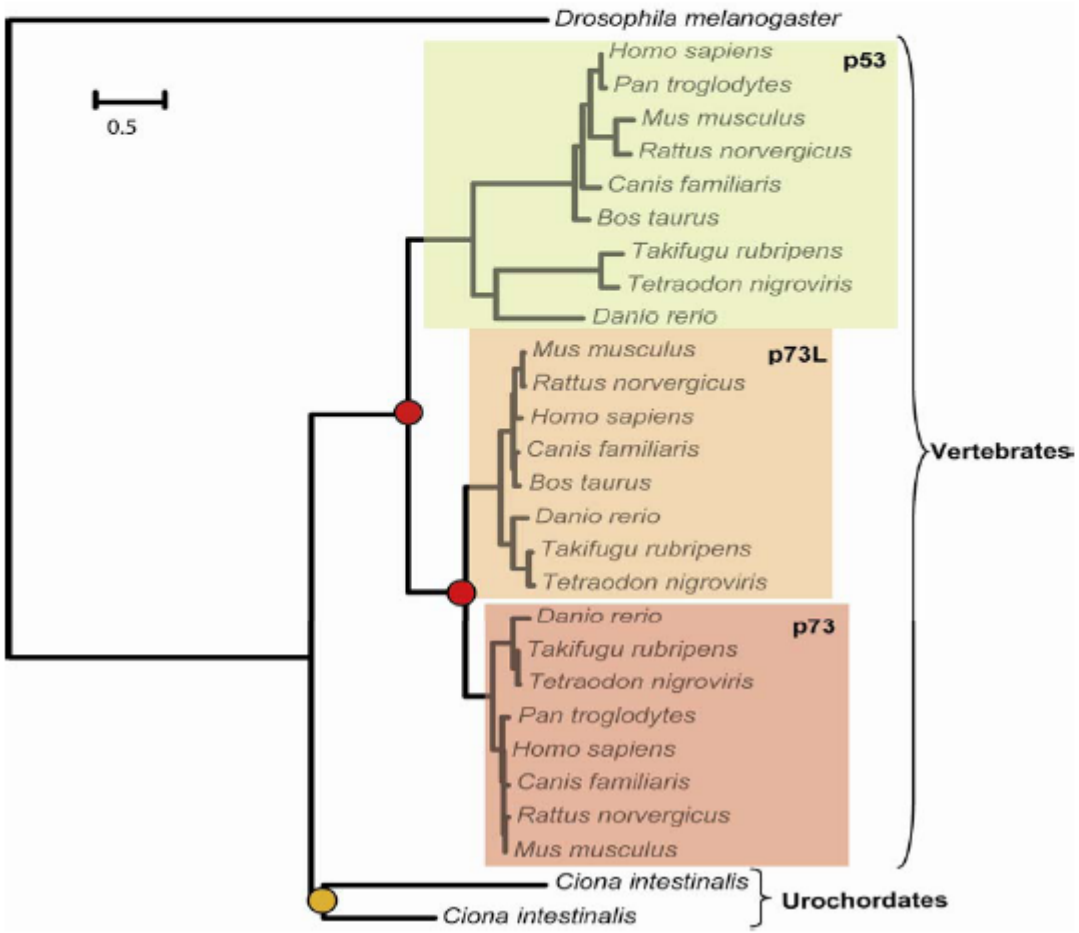
→ Species-overlap algorithms

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

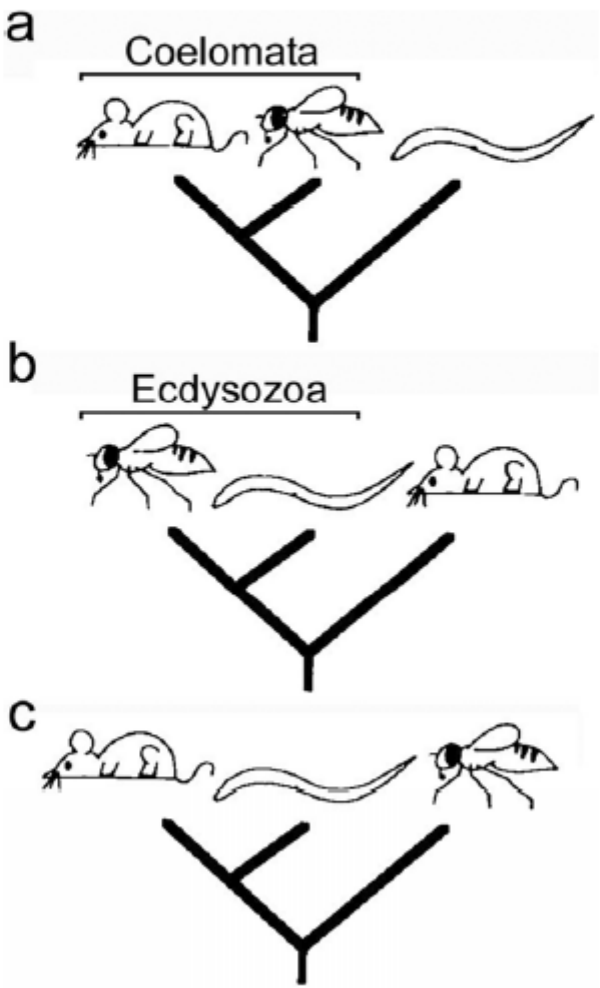
It works when these two assumptions are correct:

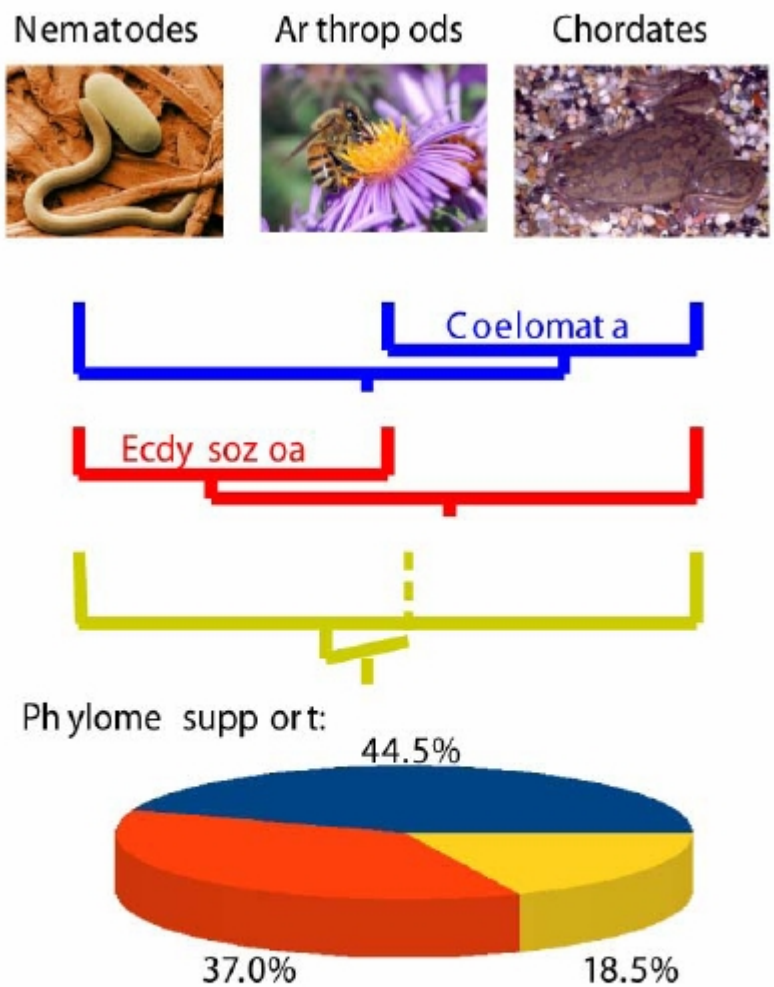
A) We know the true species tree

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?

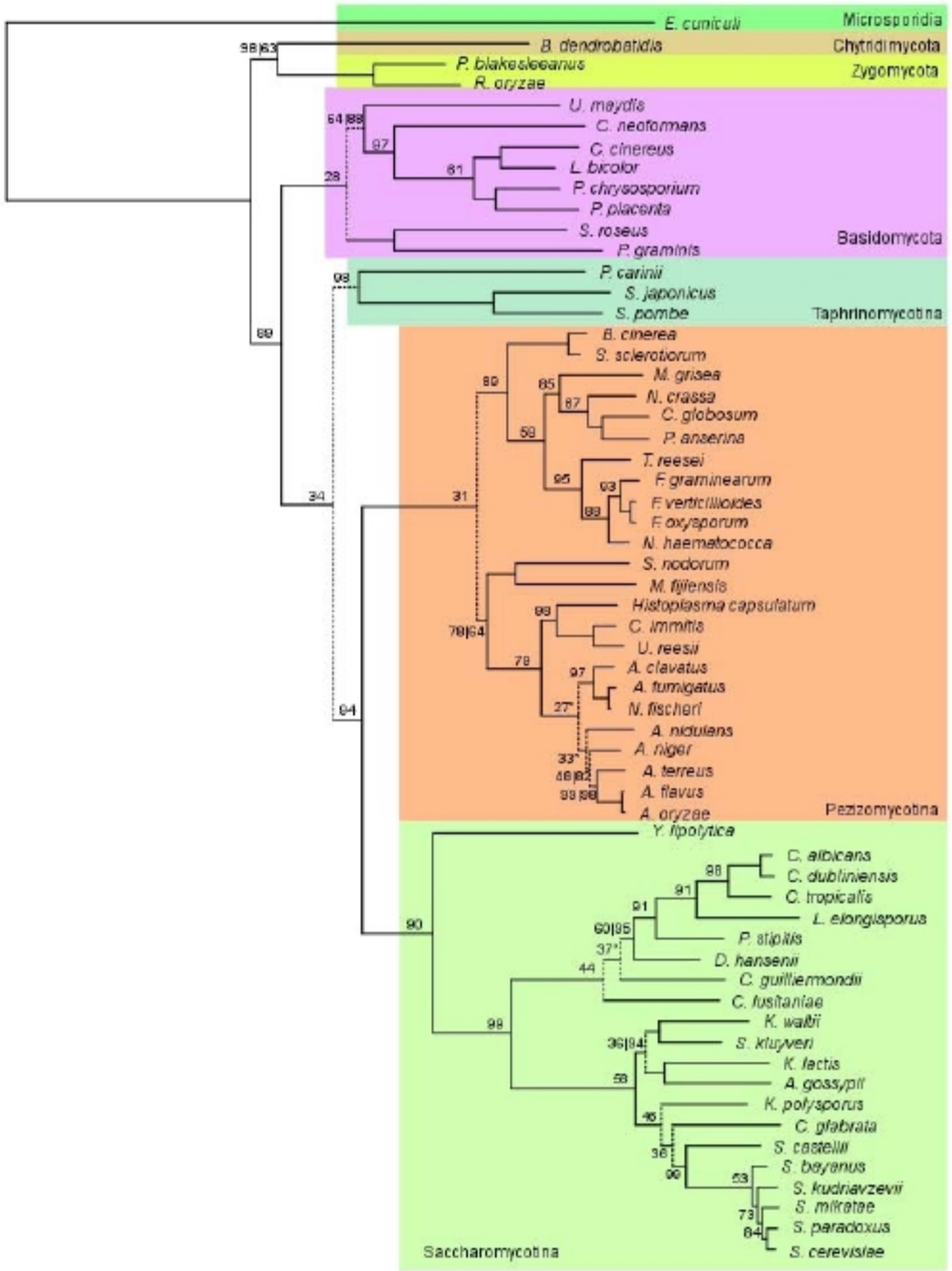
Similar results for

Primates
Rodents
laurasatheria

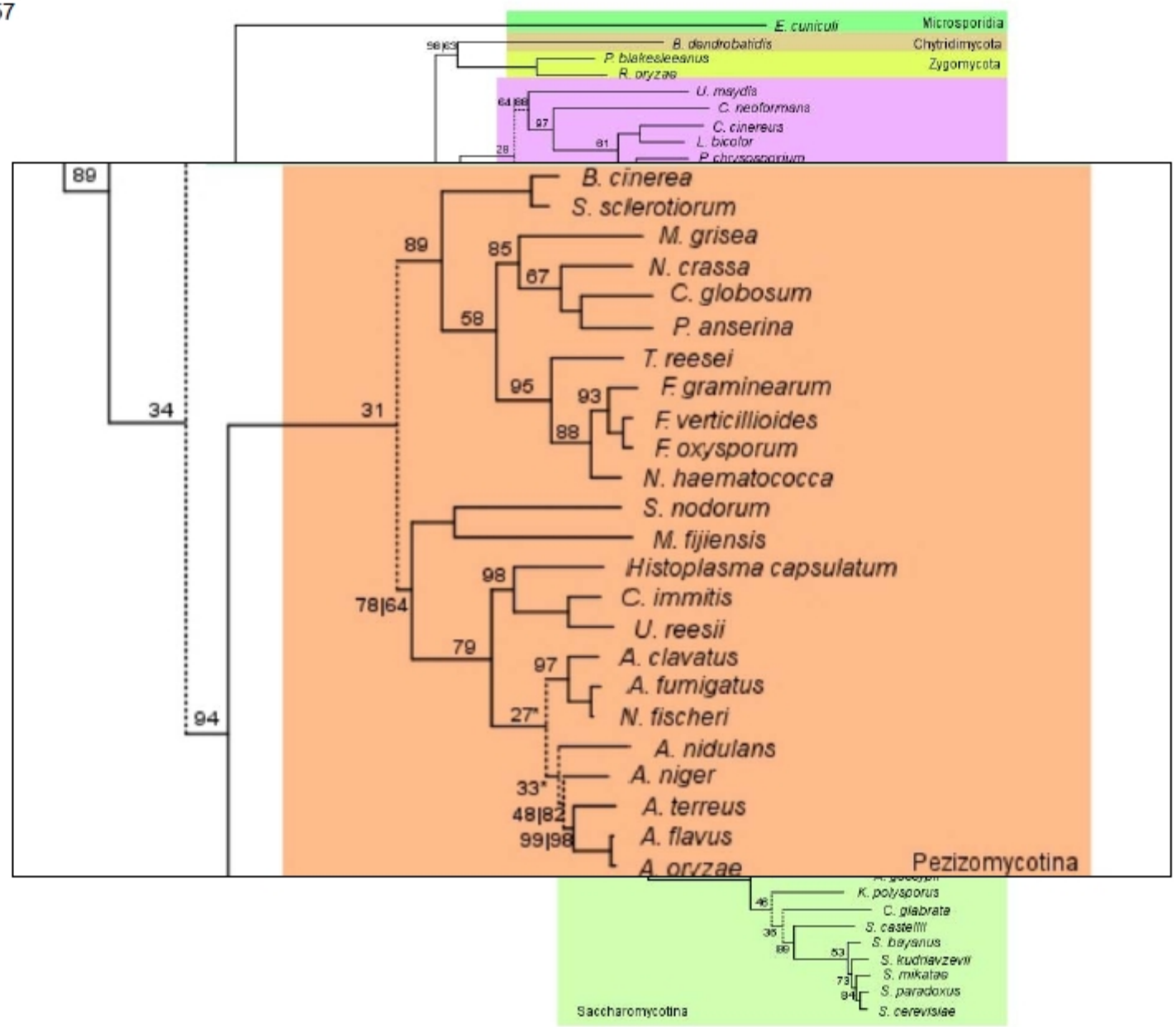
The tree vs the forest:

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

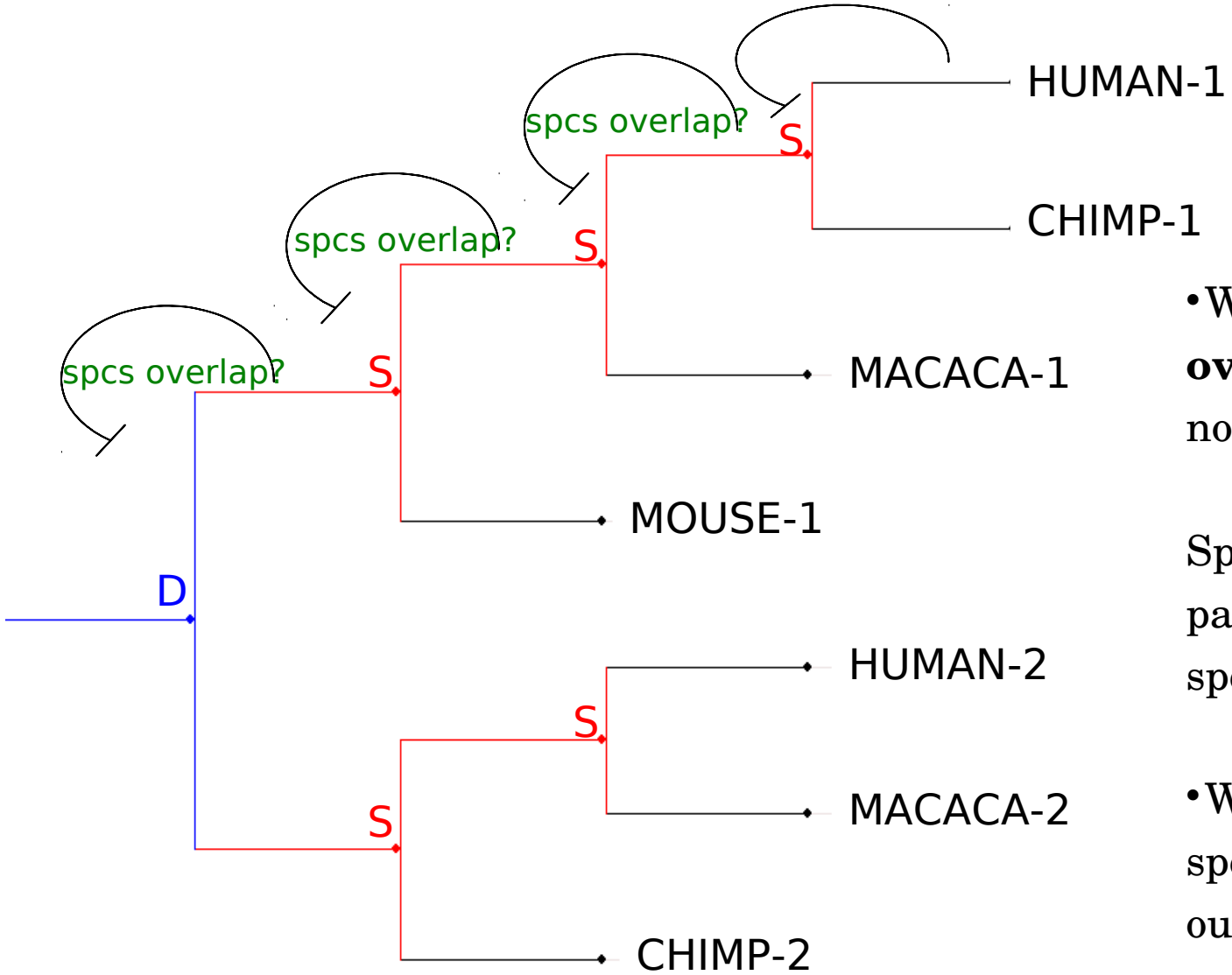


This large-degree of topological variability might be in part due to phylogenetic artifacts, insuficient phylogenetic signal, etc. But also to real evolutionary processes that render a gene tree different from a species tree: lineage sorting, gene conversion, etc

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

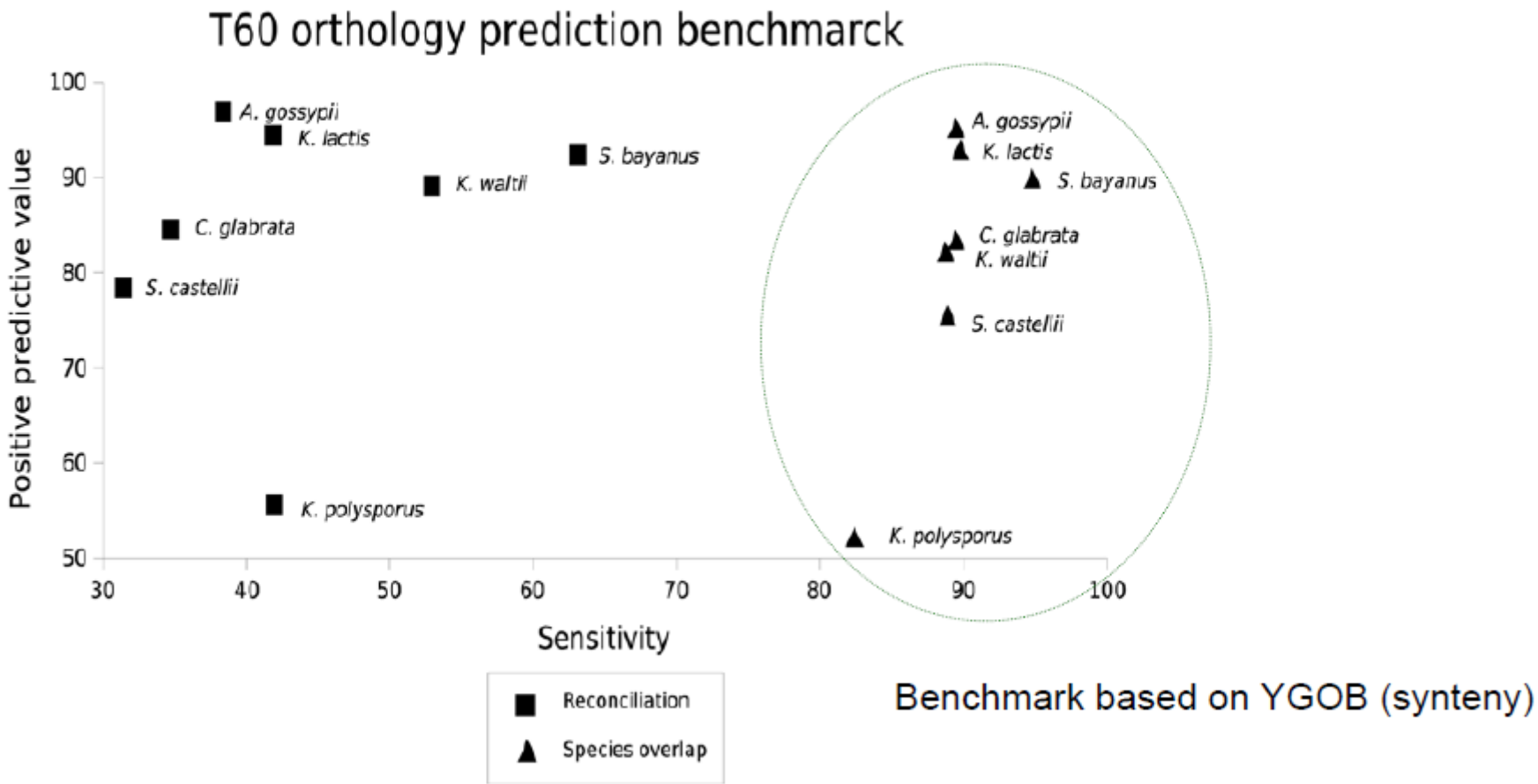
- We calculate a **species overlap score** for every node.

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

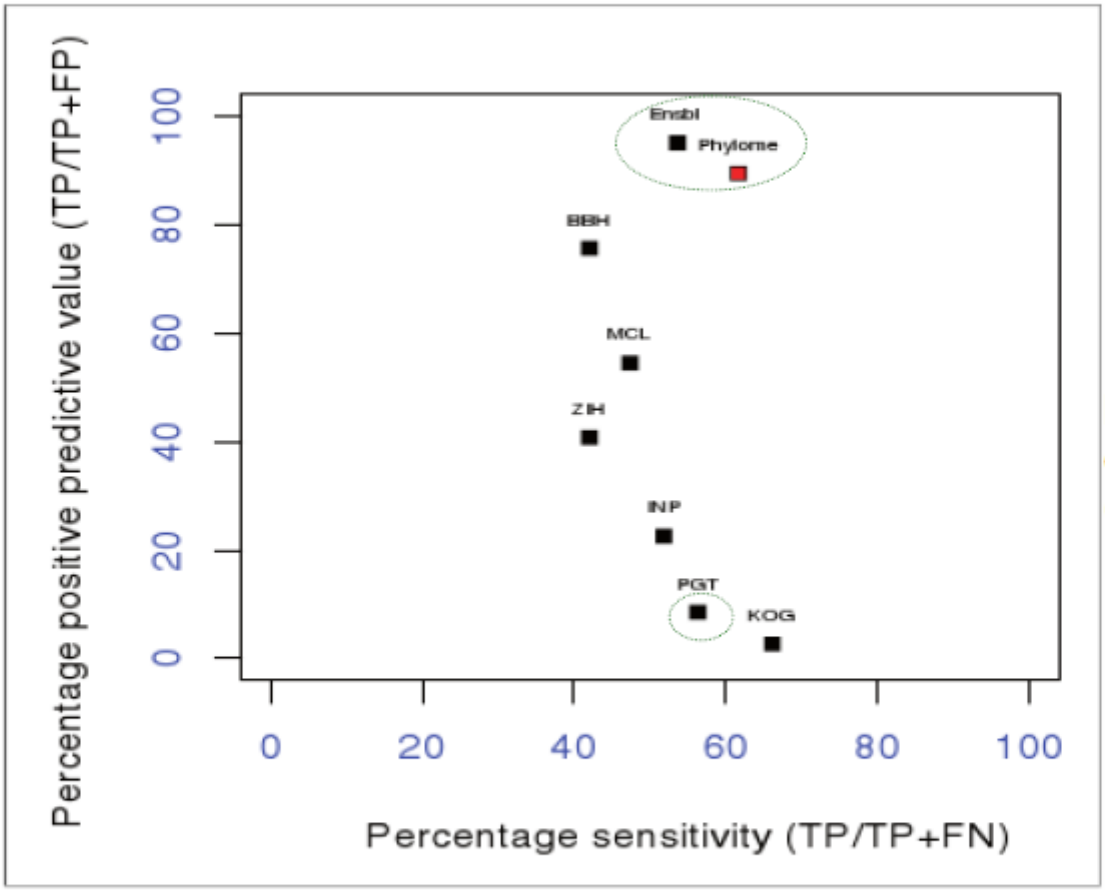
Tree reconciliation / species overlap
Marcet-Houben and Gabaldón. *PLoS ONE* (2009)



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.**

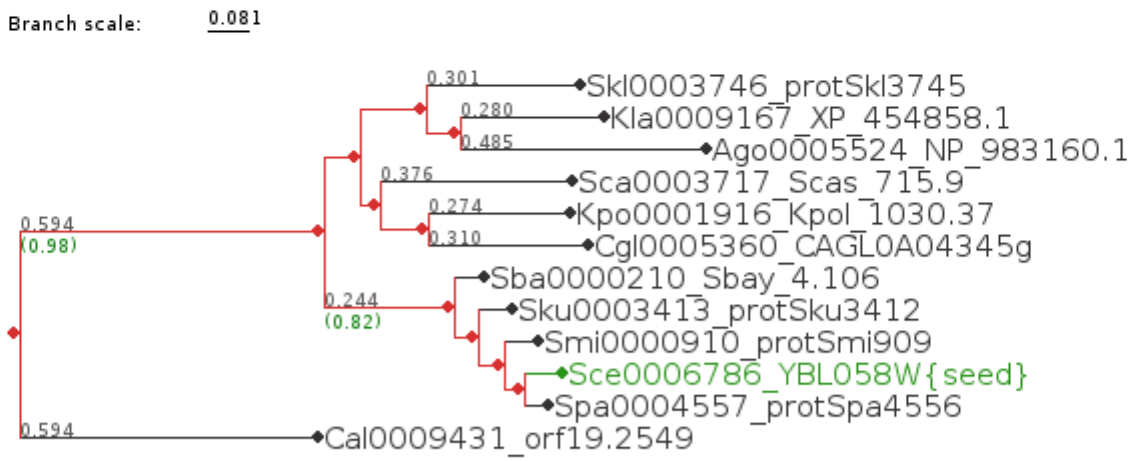
Comparative Genomics Group at CRG (Barcelona, Spain). | Citation

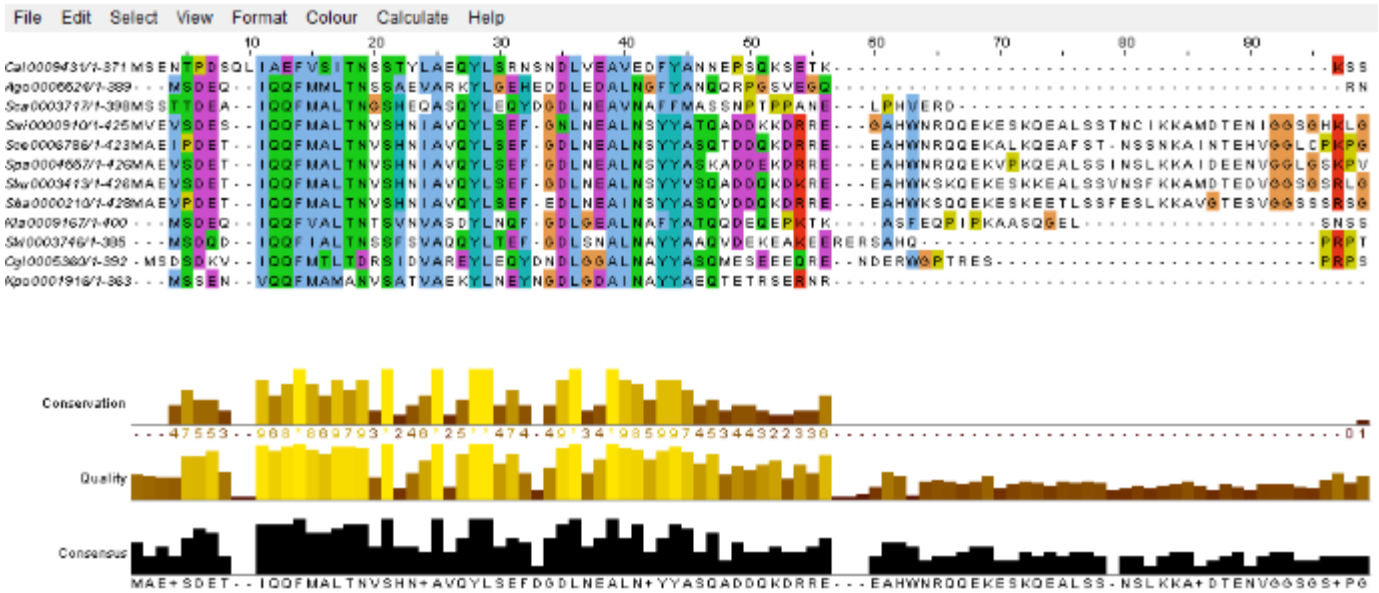
YBL058W
- Info
- Orthologs
- Seed Trees (4)
  - ScеP12a
  - ScеP21
  - ScеP12b
  - ScеP60
- Collateral Trees (4)
  - Hsa0028724
  - Hsa0028192
  - Hsa0018651
  - Hsa0016629

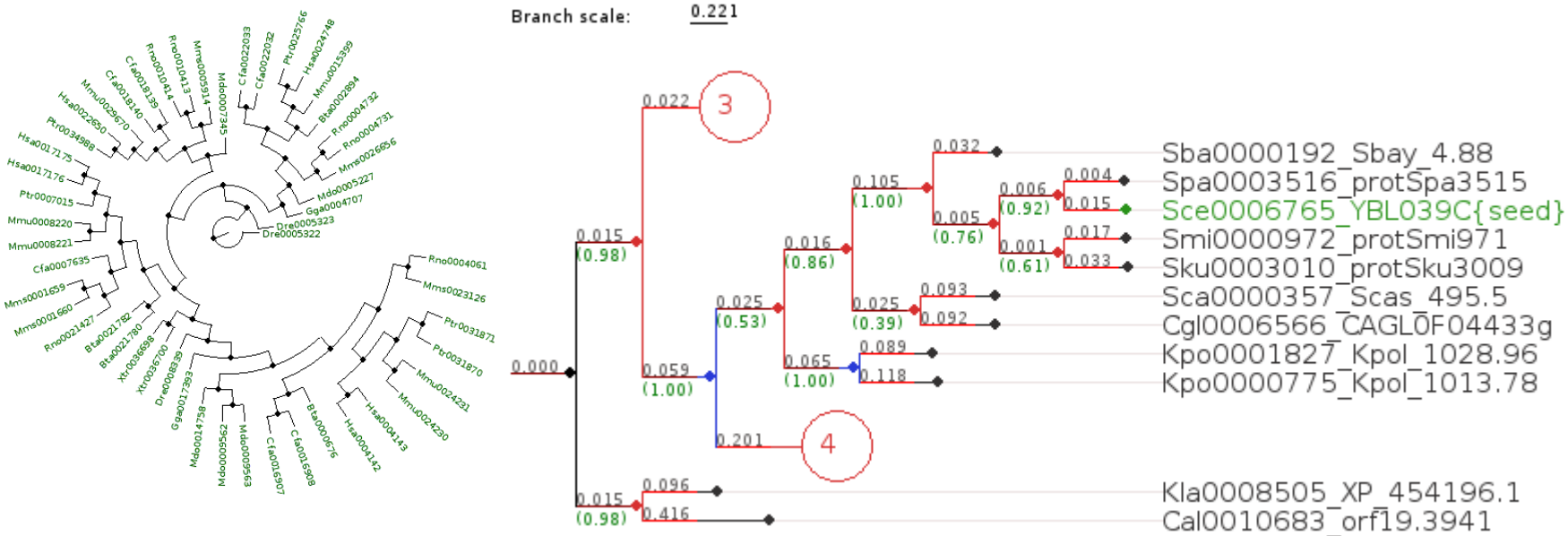
Tree: Sce0006786 (in phylome ScеP12a)

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

Tree Tools and Actions: Colapse tool Search in tree Clear

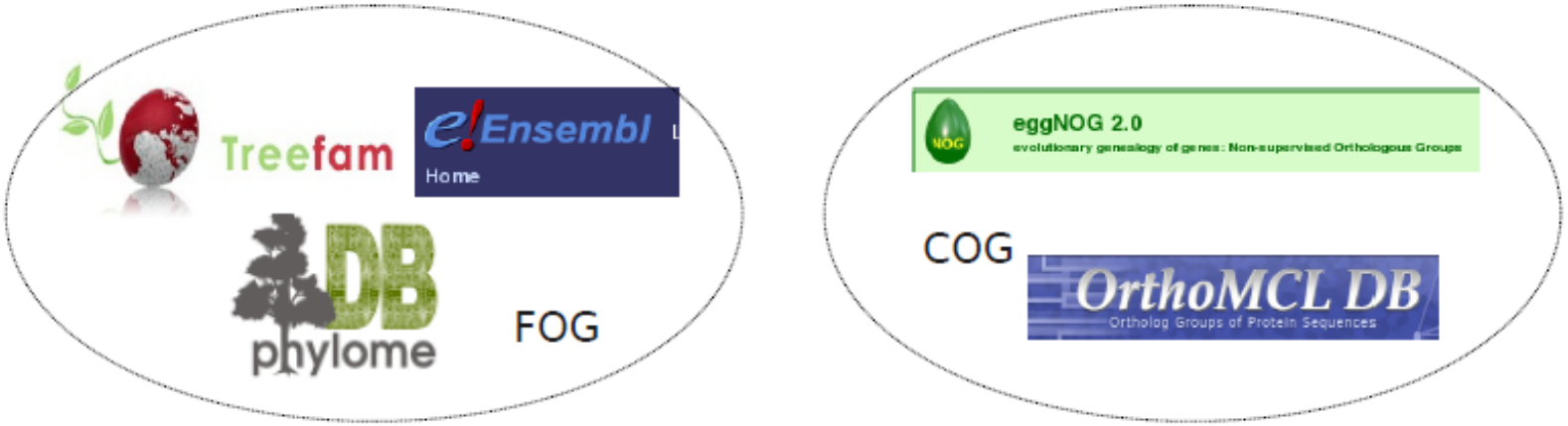



File Edit Select View Format Colour Calculate Help
Conservation
Quality
Consensus


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)



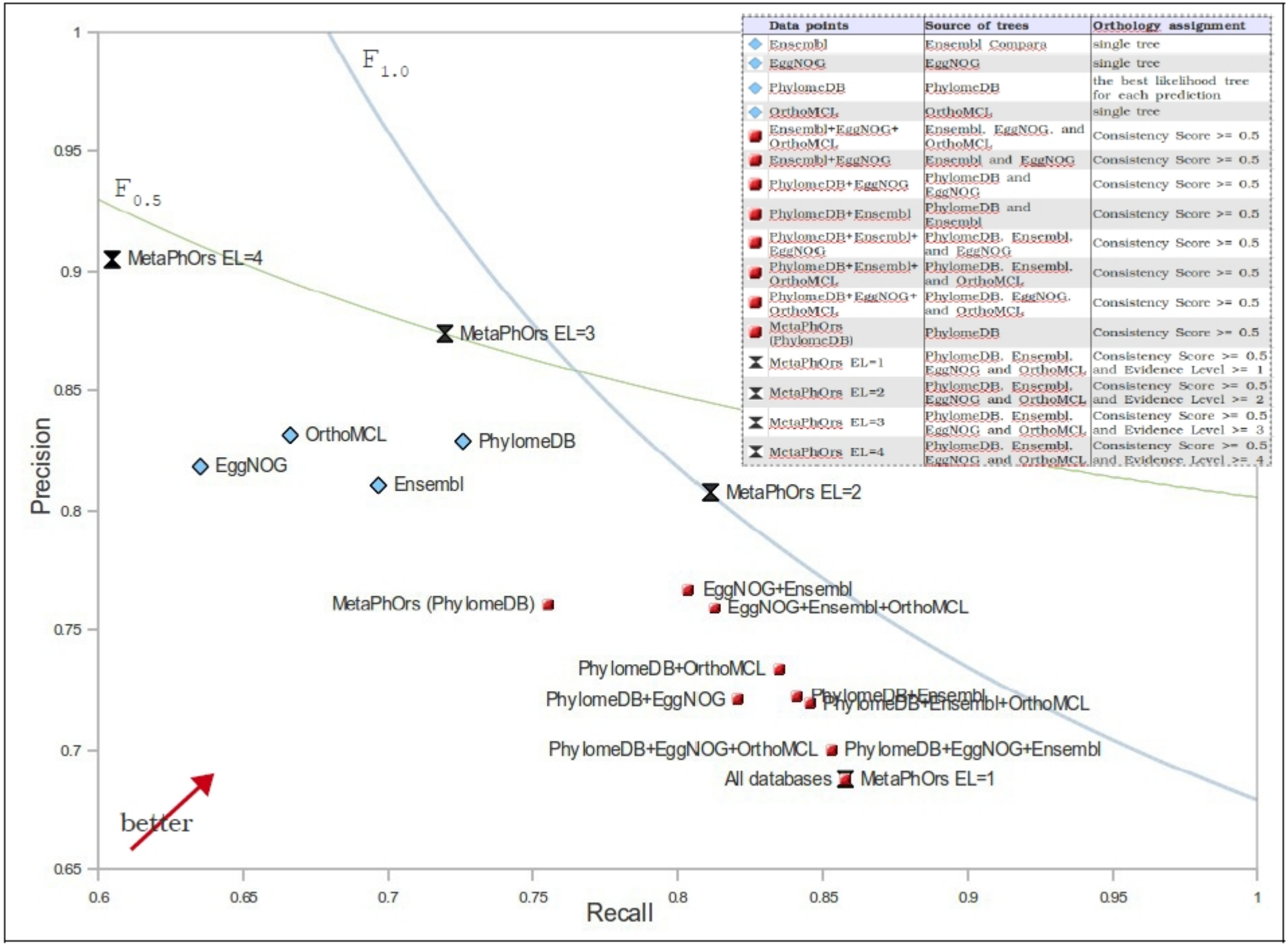
Use existing tree repositories

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)


Precision
Recall
1
0.95
0.9
0.85
0.8
0.75
0.7
0.65
0.6
0.65
0.7
0.75
0.8
0.85
0.9
0.95
1
F 1.0
F 0.5
MetaPhOrs EL=4
MetaPhOrs EL=3
MetaPhOrs EL=2
MetaPhOrs EL=1
OrthoMCL
PhylomeDB
EggNOG
Ensembl
MetaPhOrs (PhylomeDB)
EggNOG+Ensembl
EggNOG+Ensembl+OrthoMCL
PhylomeDB+OrthoMCL
PhylomeDB+EggNOG
PhylomeDB+Ensembl
PhylomeDB+Ensembl+OrthoMCL
PhylomeDB+EggNOG+OrthoMCL
PhylomeDB+EggNOG+Ensembl
All databases
better
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

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

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.

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!

Home Get all sequences

quick search

P04637

SEARCH

You can also use a BLAST search

Change options

Navigation

- Frequently Asked Questions

User login

Username: *

Password: *

Log in

- Request new password

Orthology predictions for P04637

Phy00086SJ_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
Acyrthosiphon 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	Phy00086SJ_(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?

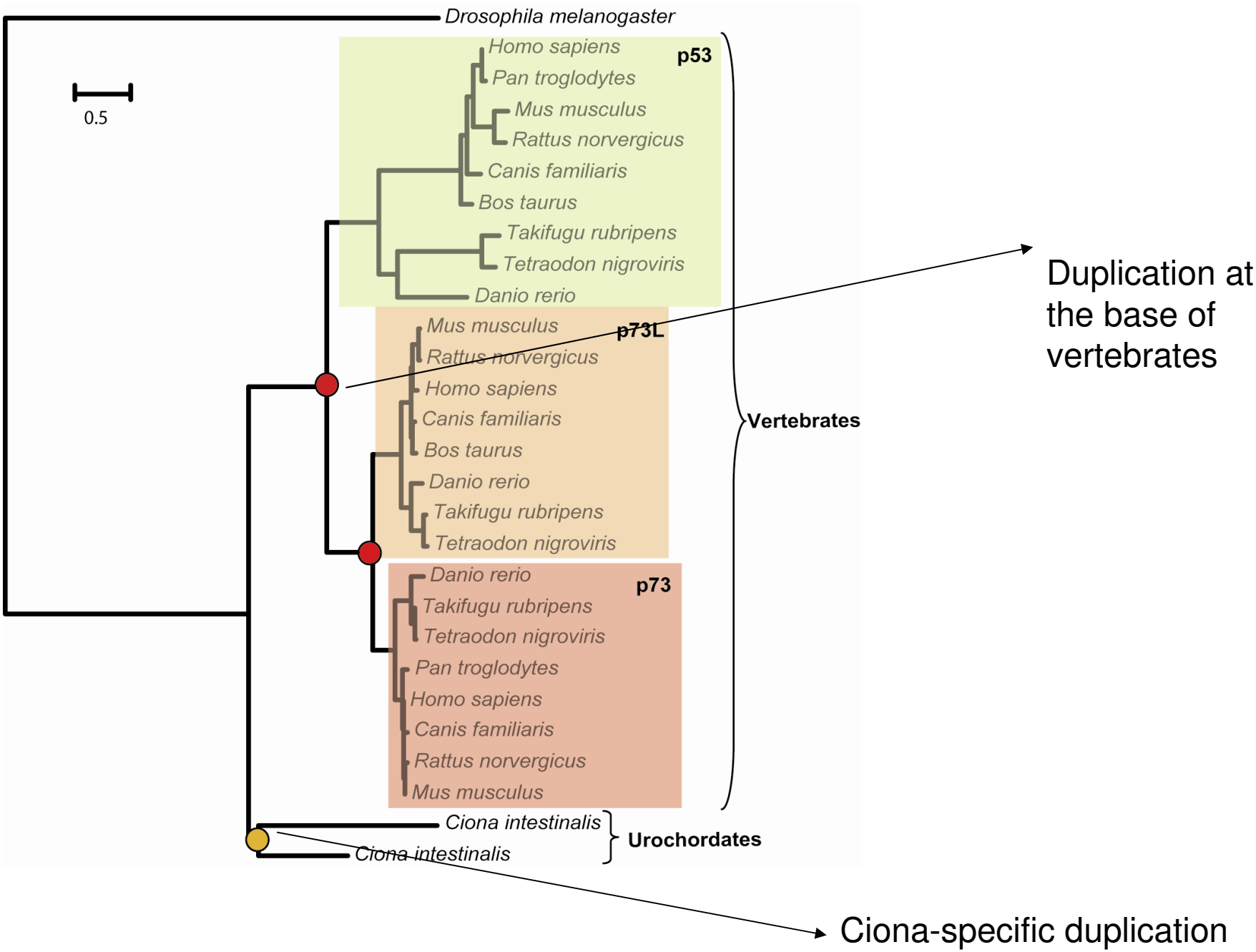
- Different taxonomic focuses
- Different methodologies
- Different outputs (pairwise relationships, groups, etc)
- Different interfaces
- Different accuracies (how to benchmark this?)

What about paralogy?

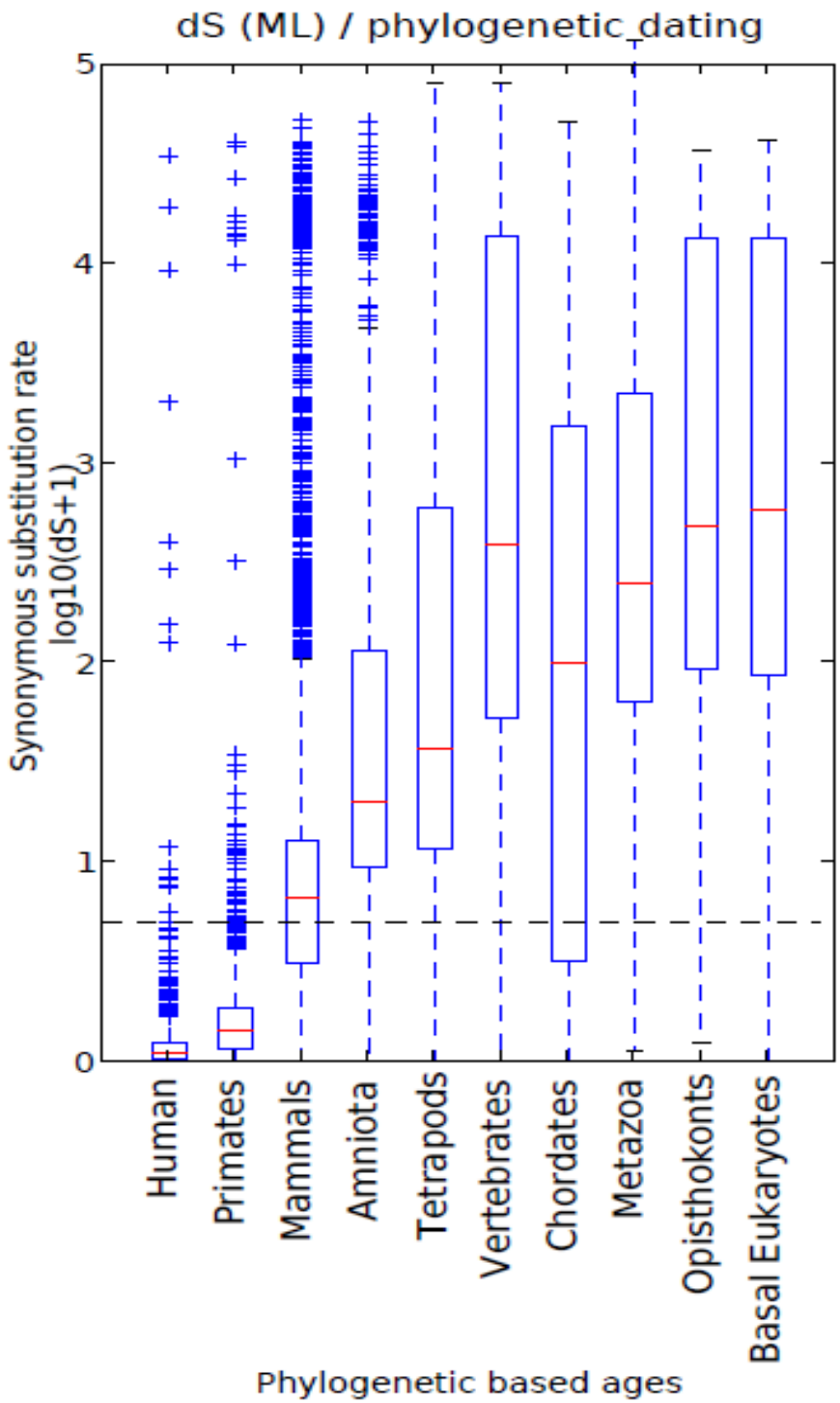
Most pairwise methods focus on orthologs, only in-paralogs are taken into account sometimes.

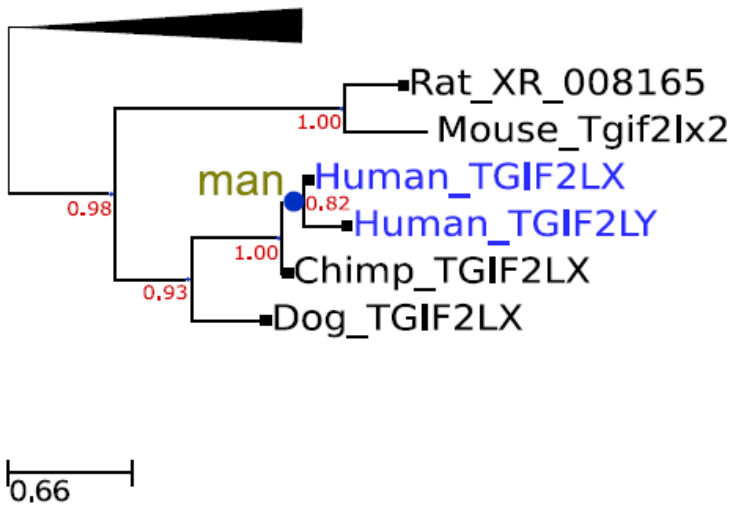
Phylogeny-based methods readily inform both on orthology and paralogy.

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


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

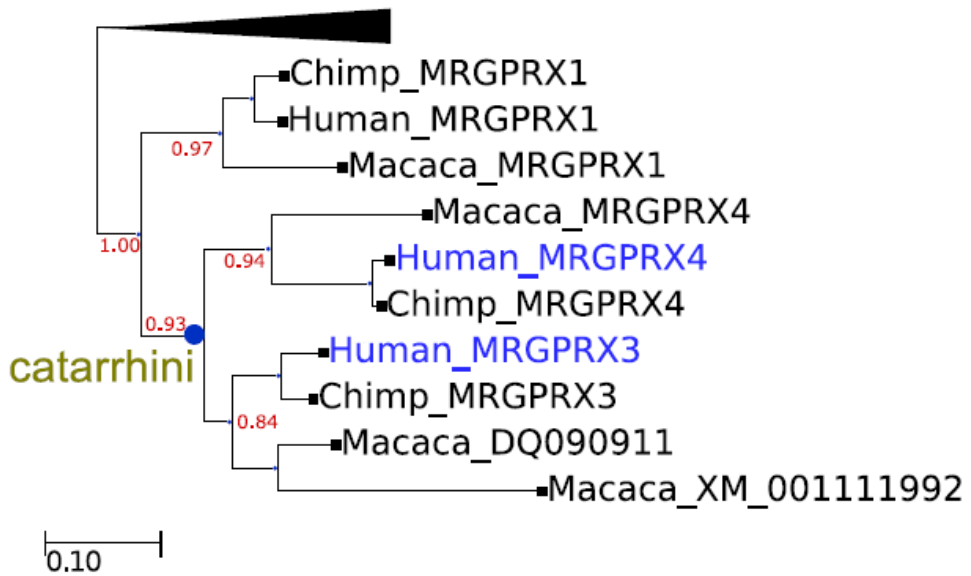
Comparison of topological dating vs dS





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)