Barcelona
Biomedical
Research
Park

CRG
Centre
de Regulació
Genòmica

Functional and Evolutonary Implications of Orthology

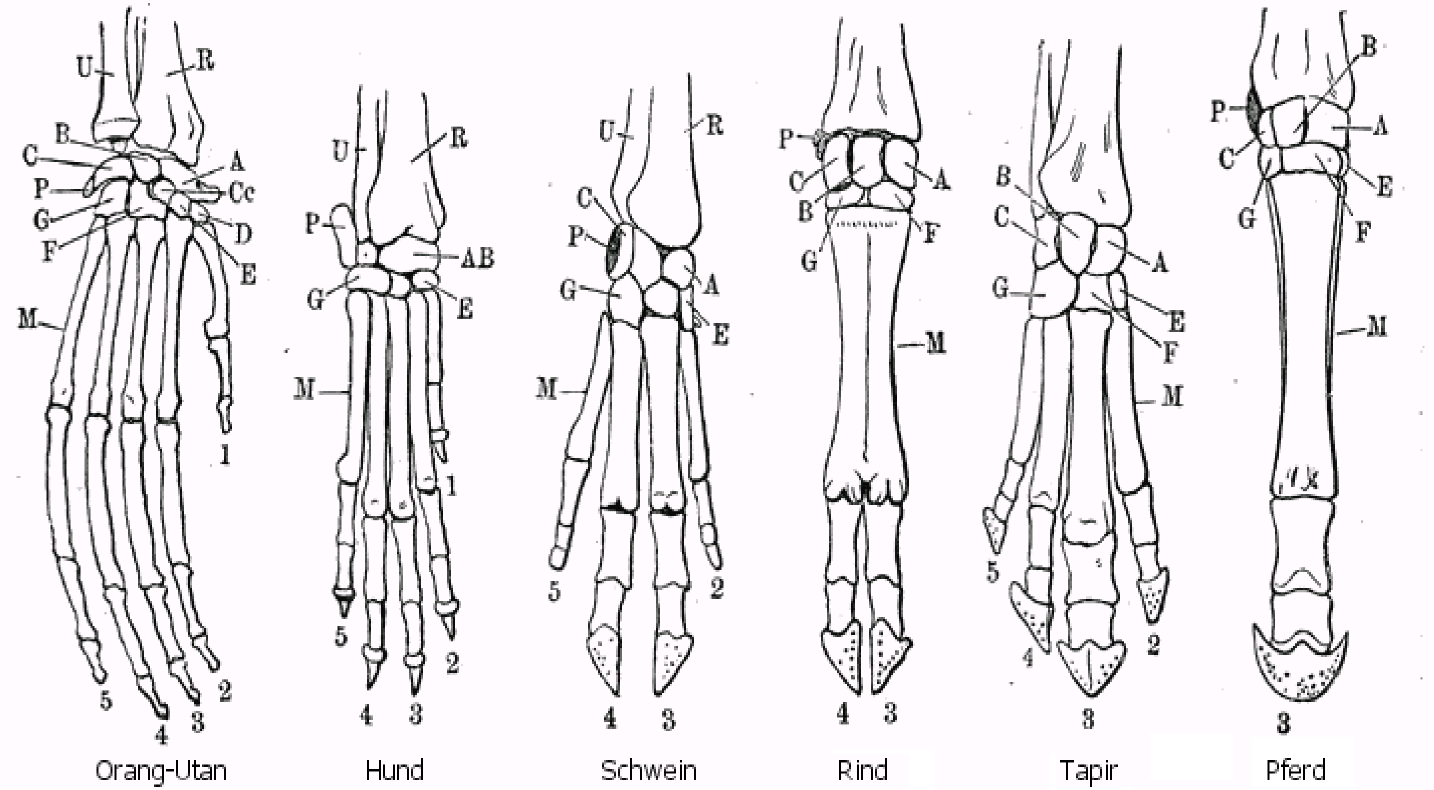
Toni Gabaldón

Centre for Genomic Regulation (CRG), Barcelona

(tgabaldon@crg.es)

http://gabaldonlab.crg.es

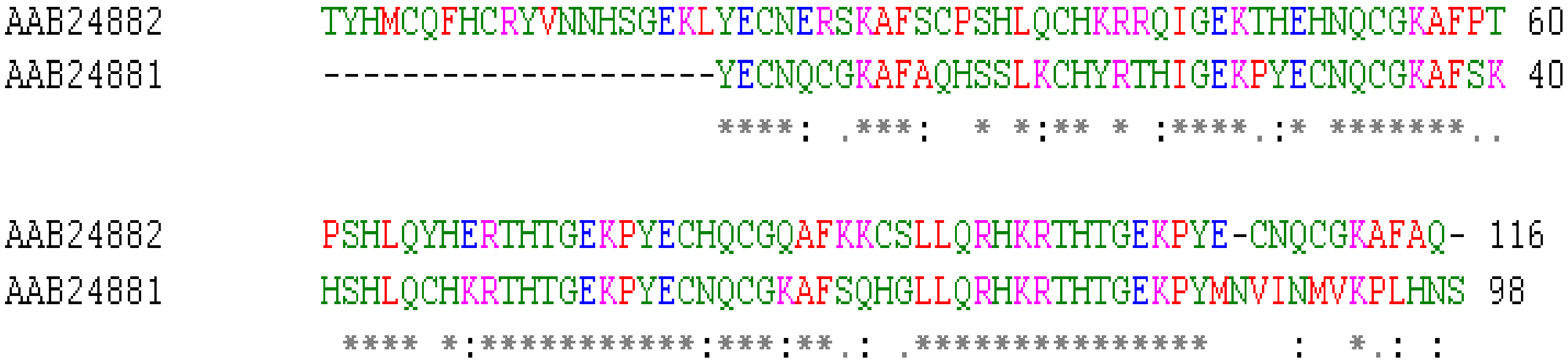
Finding which gene is which in a set of genomes (orthology)



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).

Homology: common ancestry



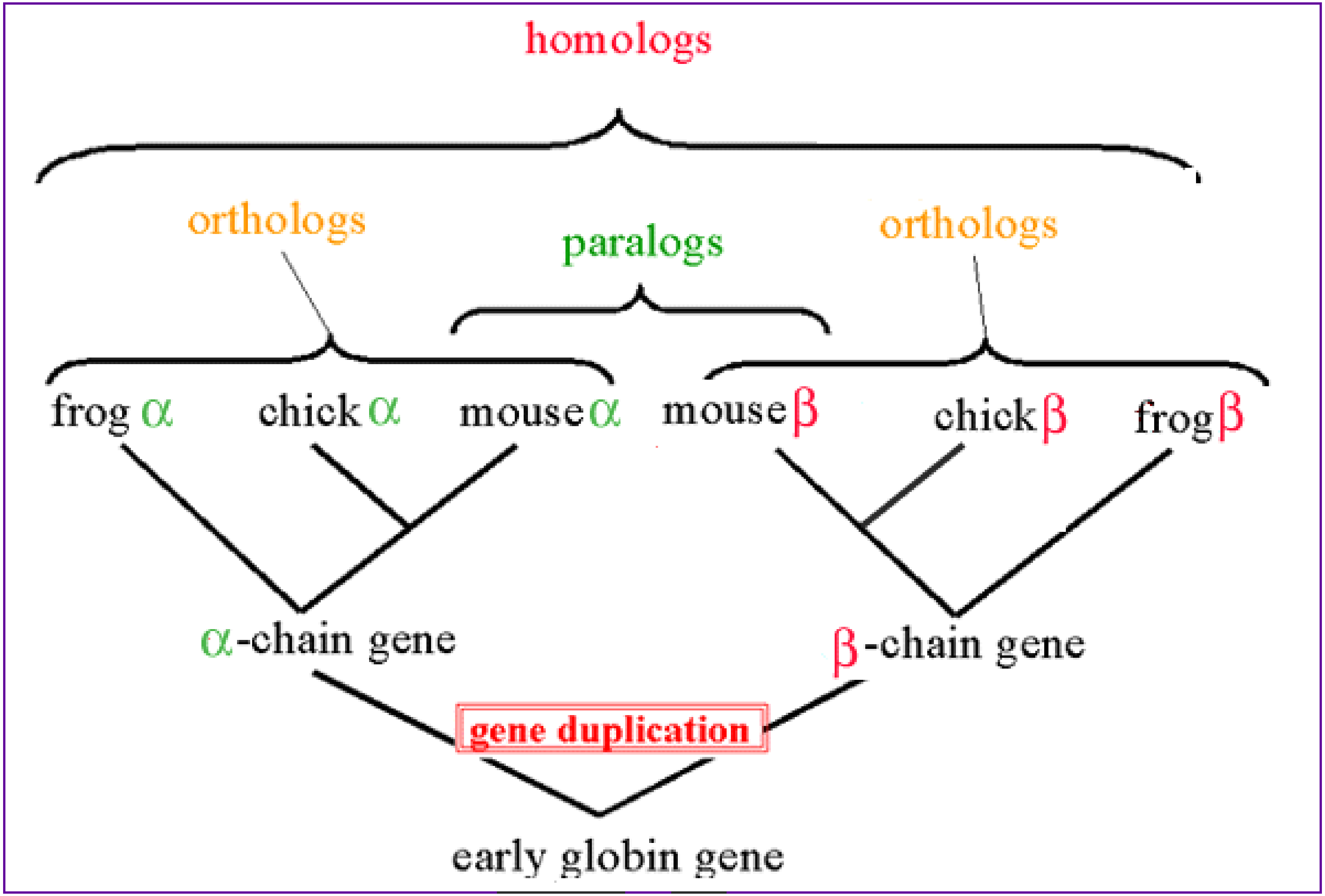
```
AAB24882        TYHMCQFHCRYVNNHSGEKLYECNERSKAFSCPSHLQCHKRRQIGEKTHEHNQCGKAFPT 60
AAB24881        --------------------YECNQCGKAFAQHSSLKCHYRTHIGEKPYECNQCGKAFSK 40
                                    ****: .***:  * *:** * :****.:* *******..

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

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…)
- 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)

Why predicting orthology is important?

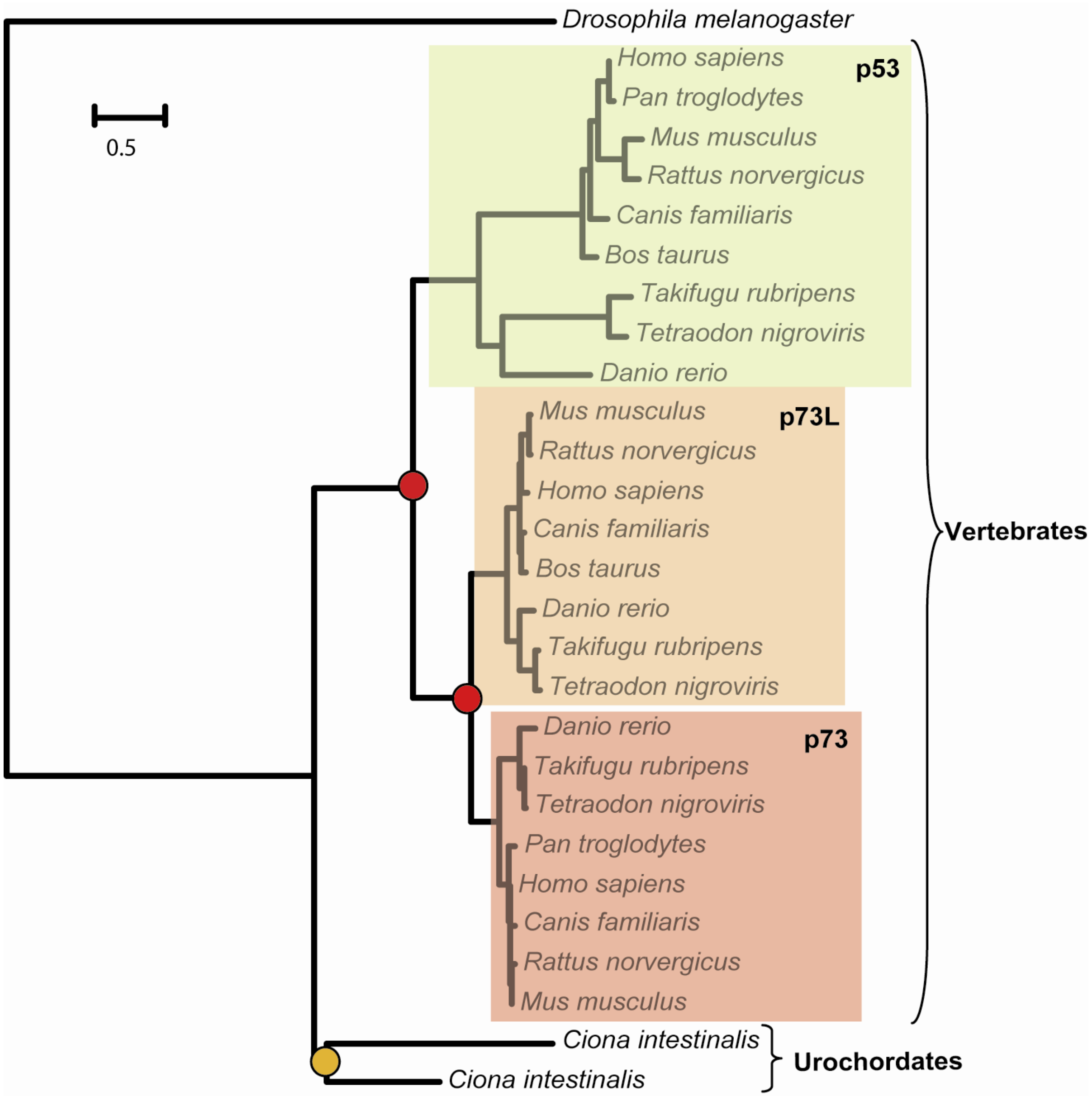
- **Important implications for phylogeny:** only sets of orthologous genes are expected to reflect the underlying species evolution (although there are many exceptions)

- 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

Classical approach: phylogenetic inference

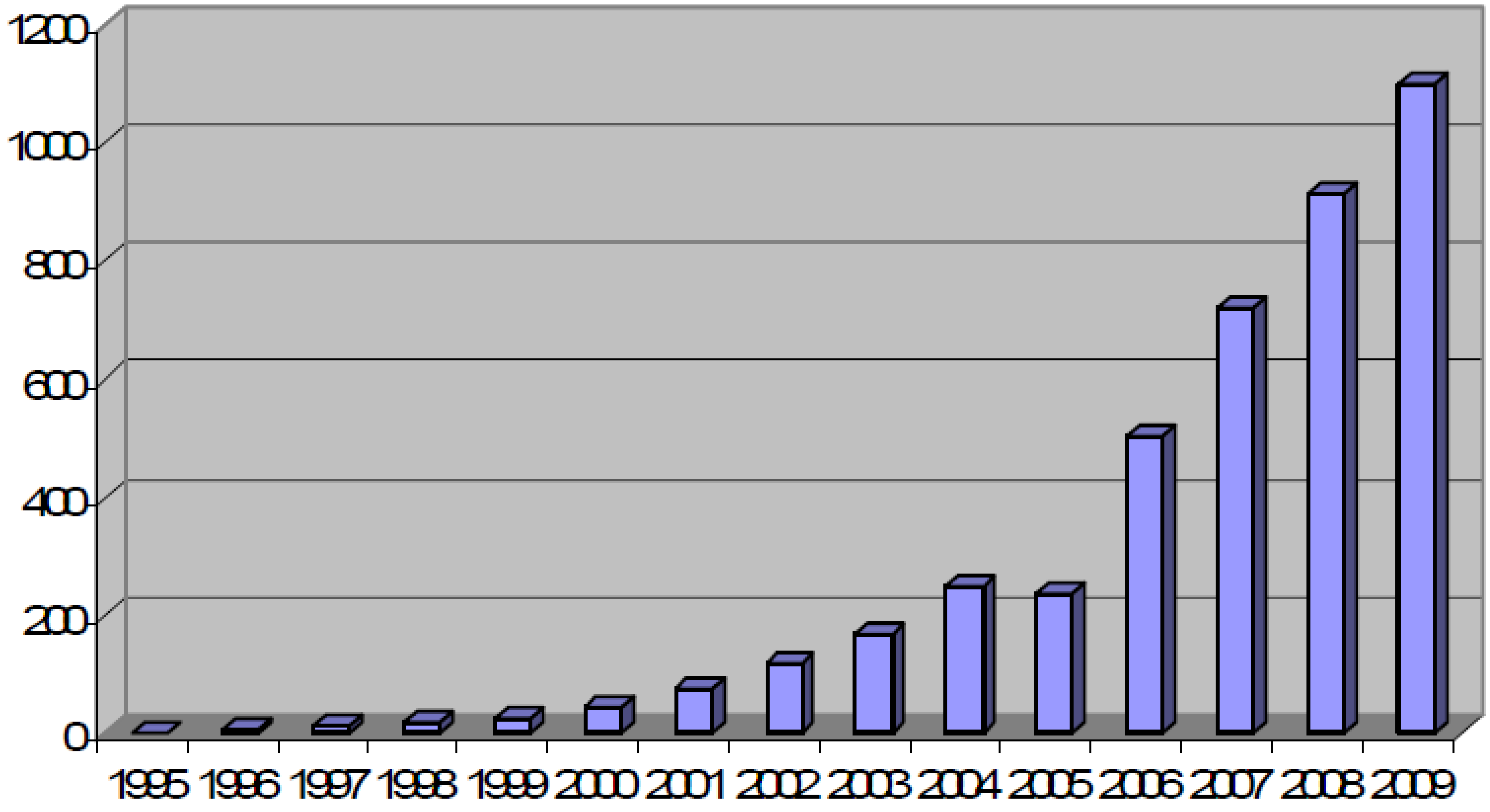
- Build a gene tree
- Compare to the species tree
- Infer duplications and speciation events
- Assign orthology and paralogy relationships accordingly

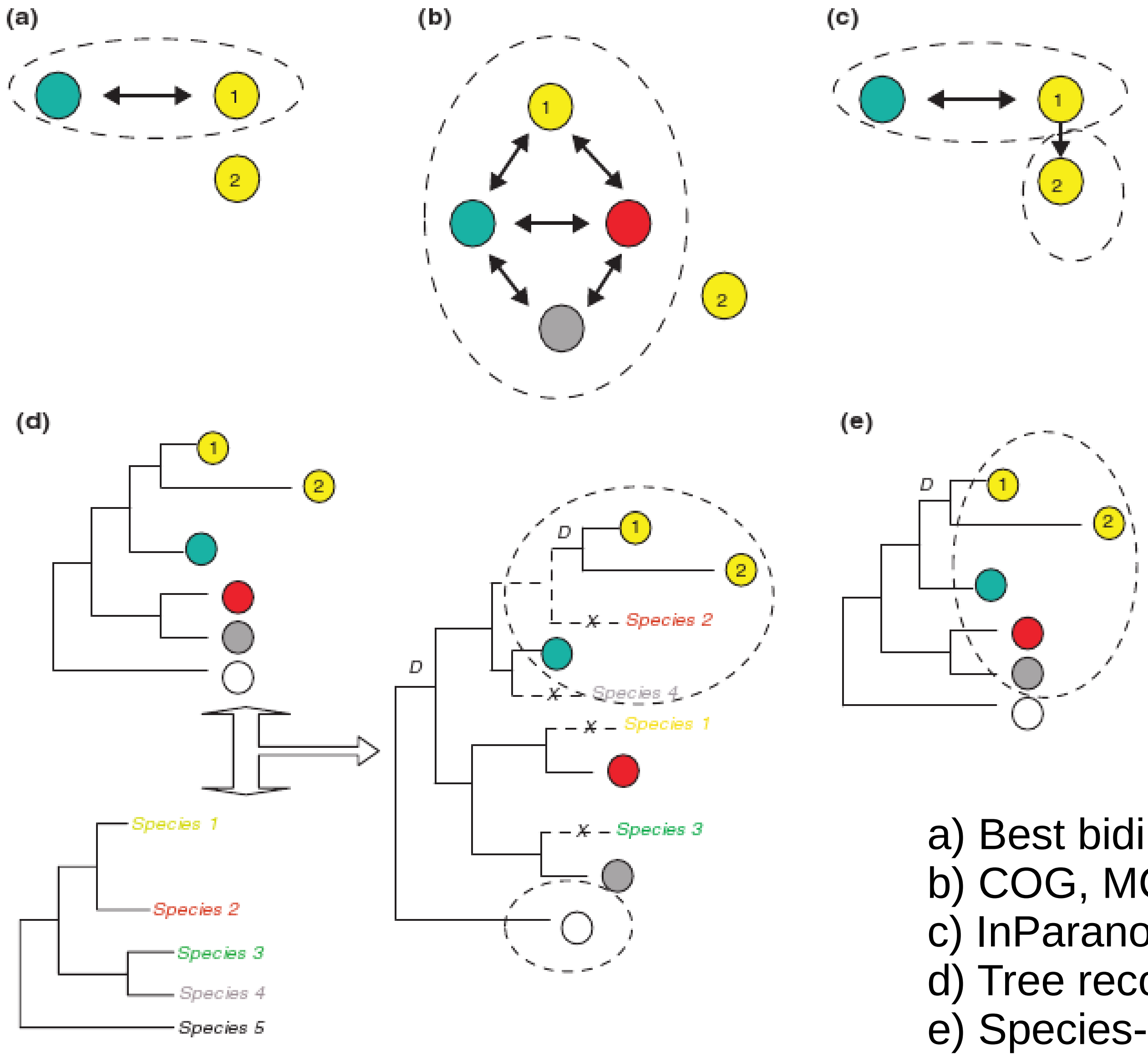

Drosophila melanogaster
0.5
p53
Homo sapiens
Pan troglodytes
Mus musculus
Rattus norvergicus
Canis familiaris
Bos taurus
Takifugu rubripens
Tetraodon nigroviris
Danio rerio
p73L
Mus musculus
Rattus norvergicus
Homo sapiens
Canis familiaris
Bos taurus
Danio rerio
Takifugu rubripens
Tetraodon nigroviris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





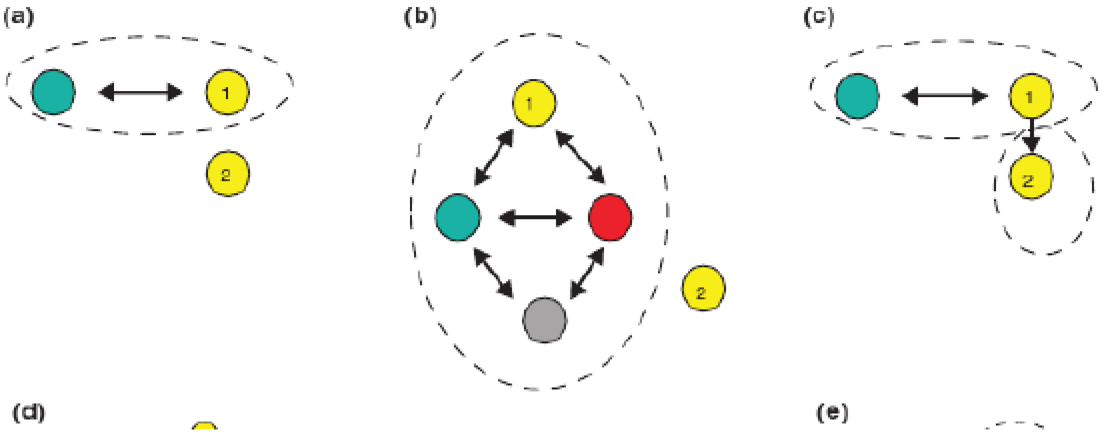
a) Best bidirectional hits
b) COG, MCL-clustering approach
c) InParanoid
d) Tree reconciliation
e) Species-overlap (PhylomeDB)

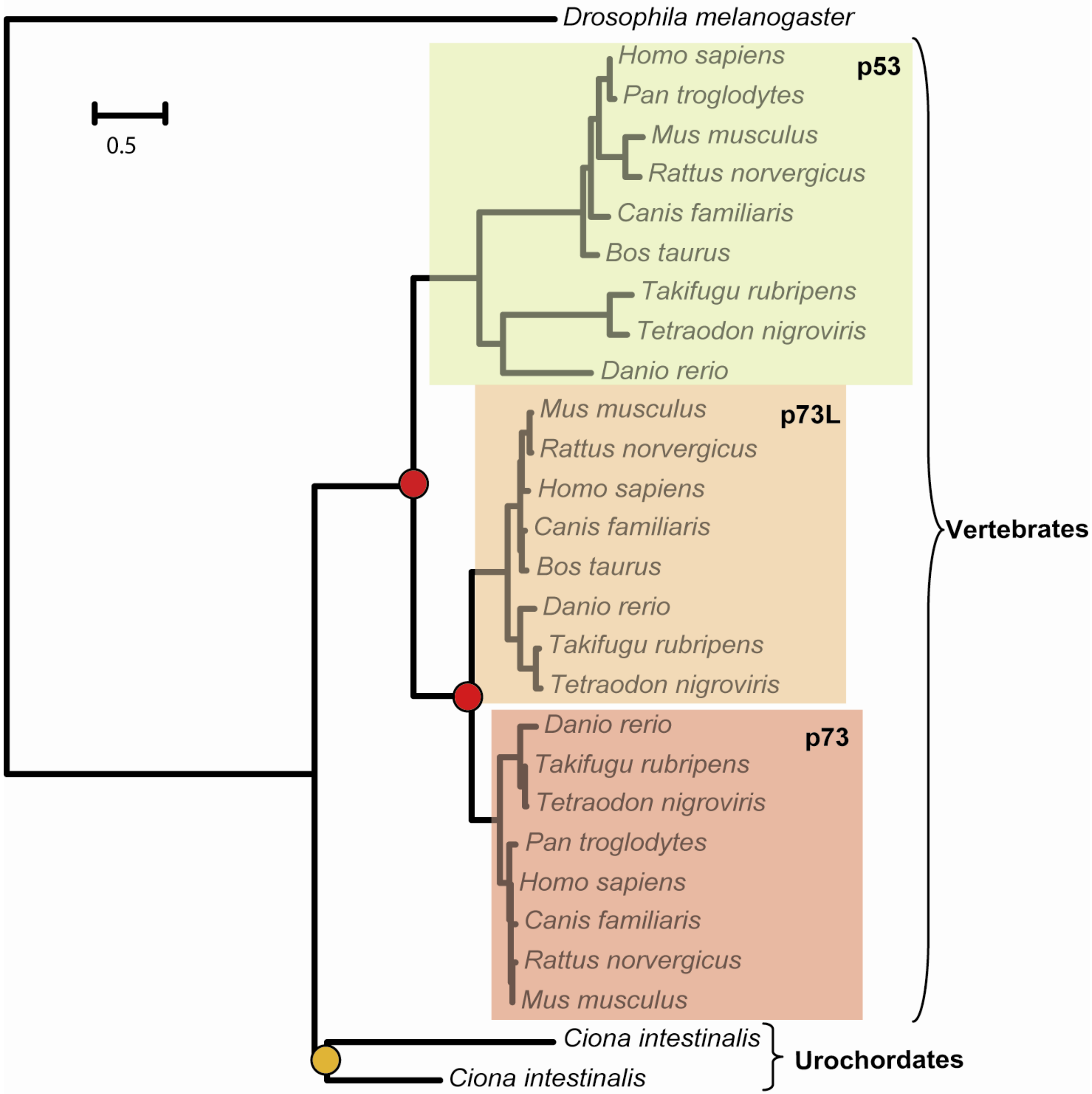
Gabaldón, T. *Genome Biology* (2008)

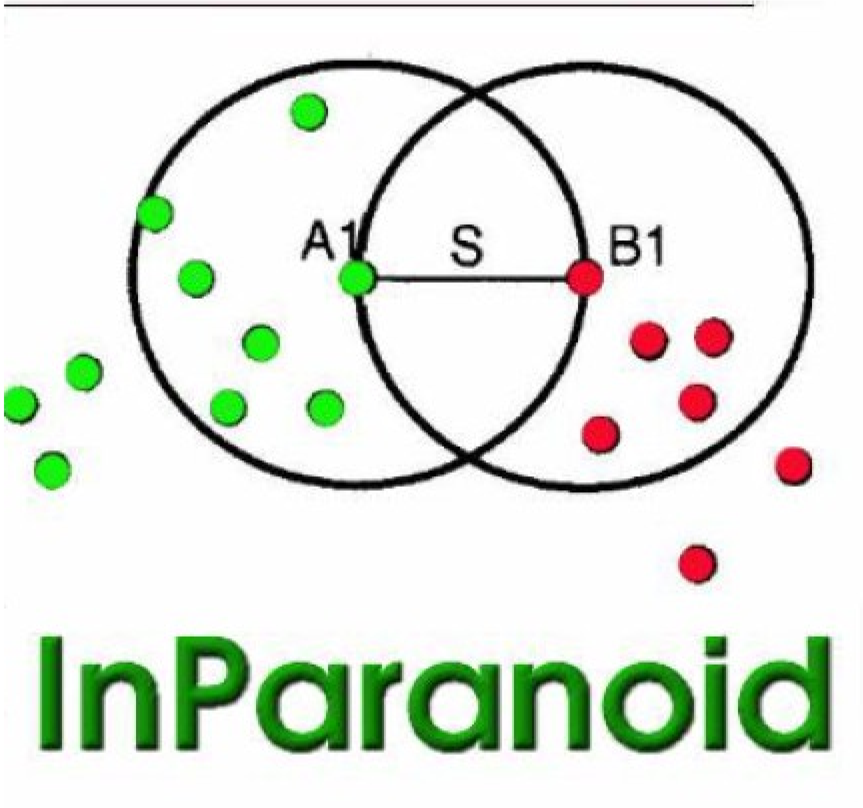
Similarity-based approaches (many more approaches):

- **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

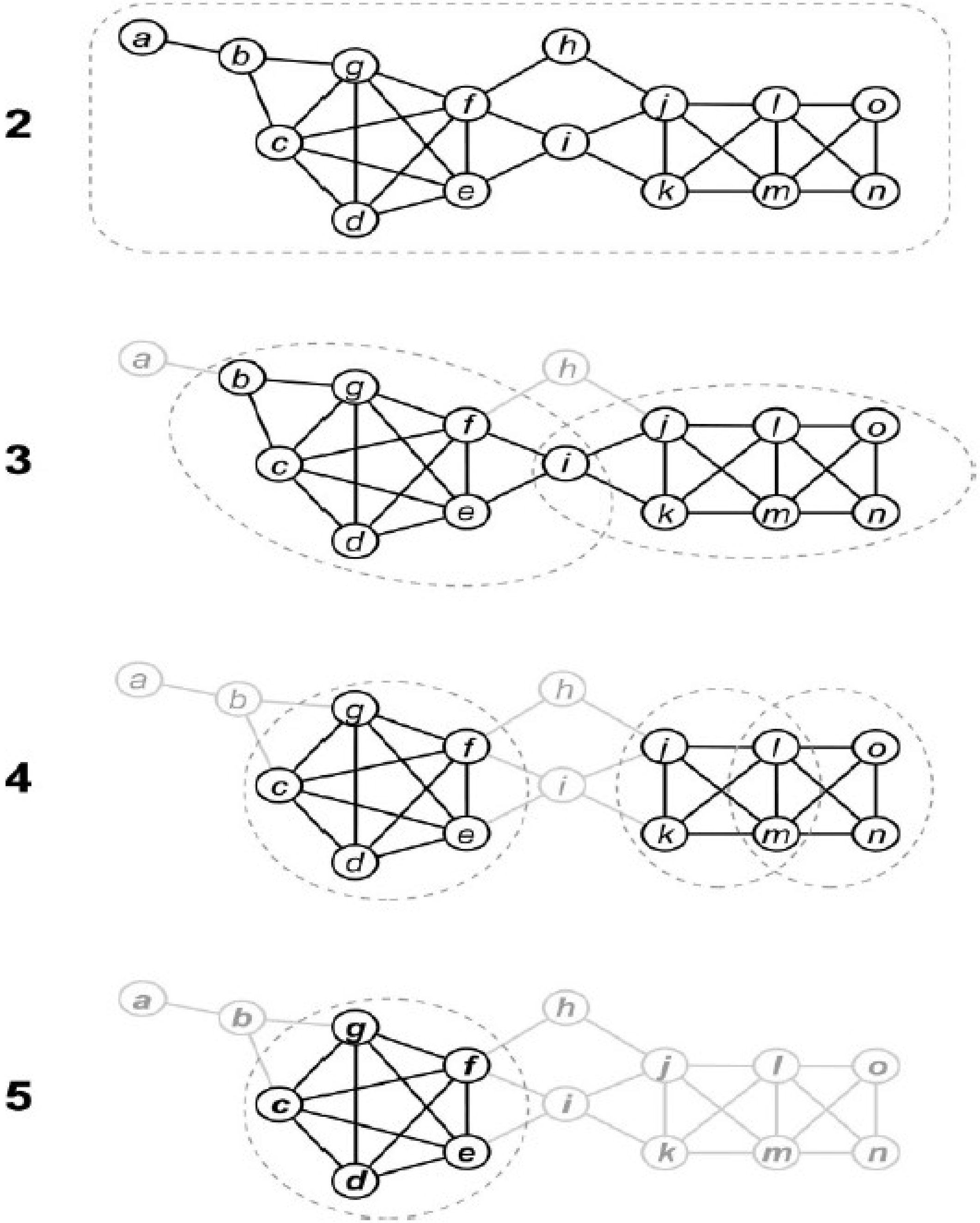


In-Paranoid.
Improved BRH to detect in-paralogs as well. Works well at the pairwise level. (multi-paranoid for multi-species comparisons

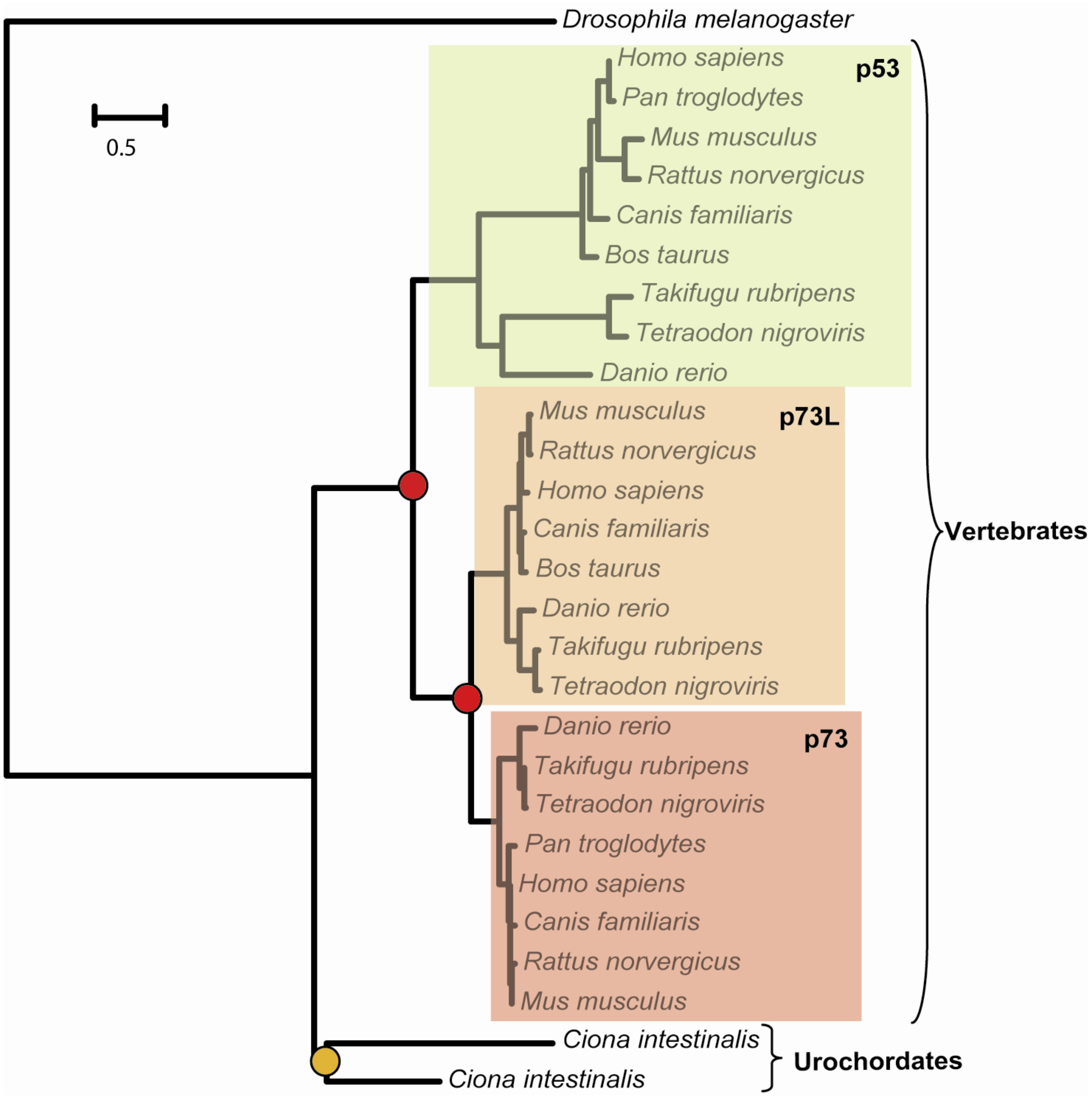
COG-like
(used by many DBs like STRING)

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



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



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)

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.

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)

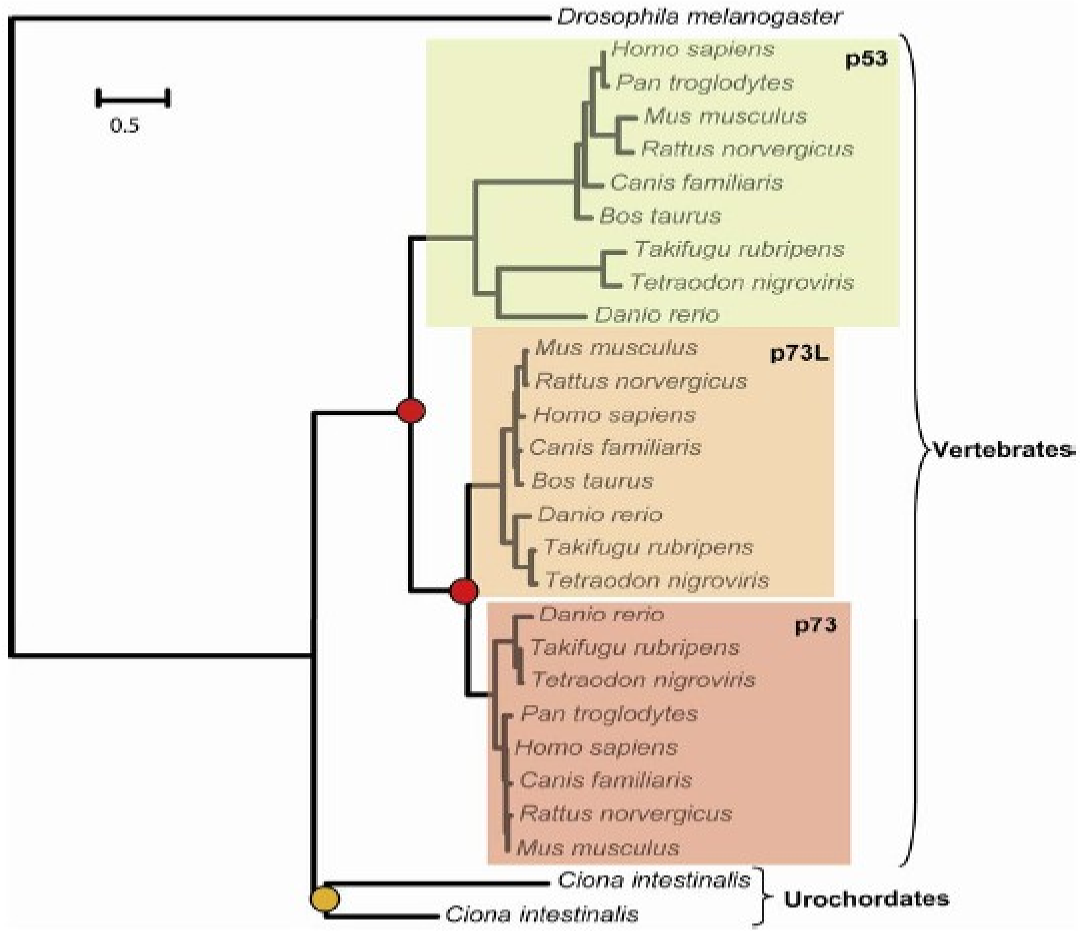
→ Species-overlap algorithms

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

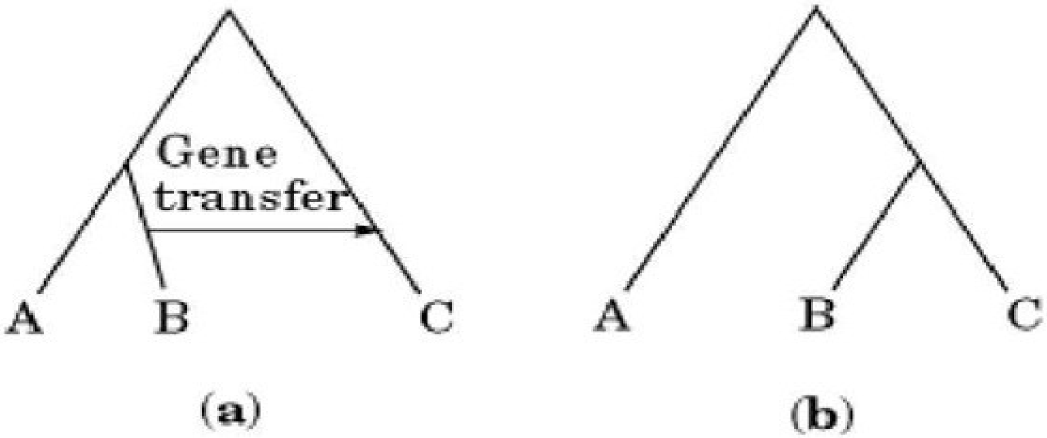
It works when these two assumptions are correct:

A) We know the true species tree

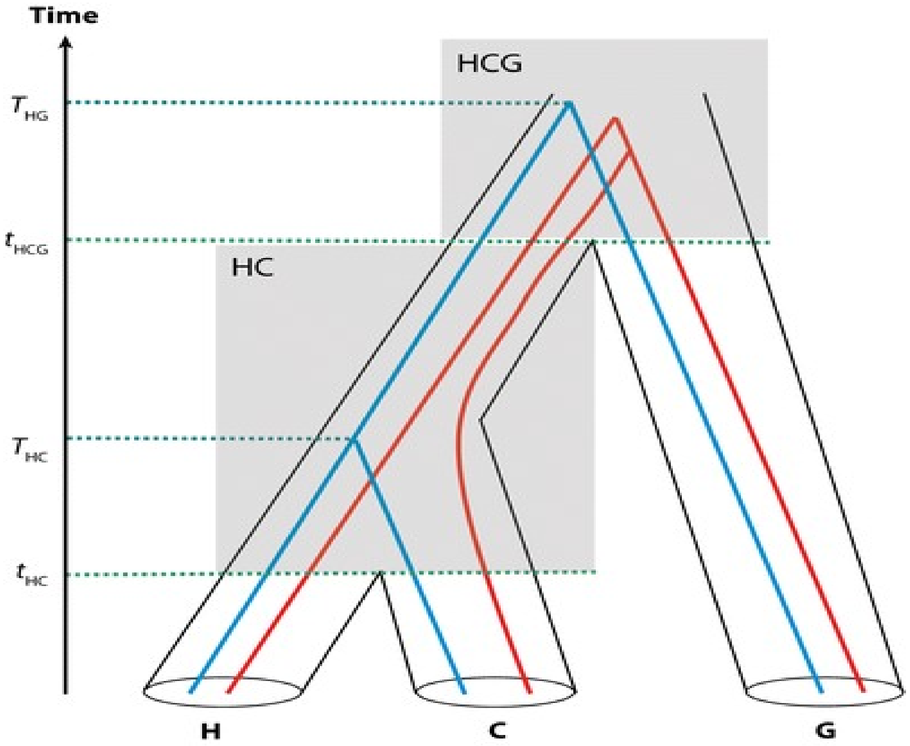
B) The gene tree is correct and reflects the species evolution



Horizontal gene transfer



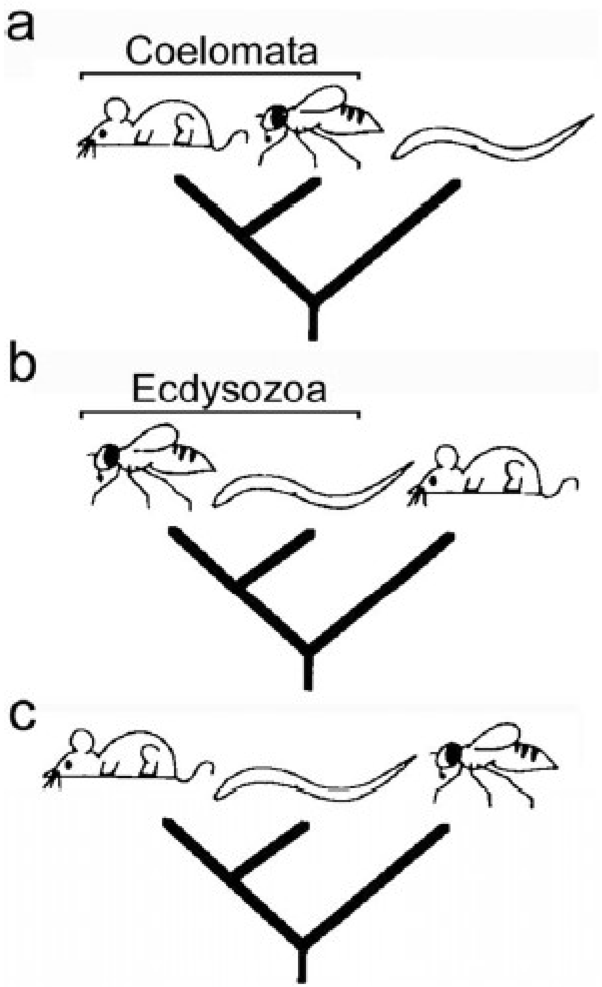
Incomplete lineage sorting

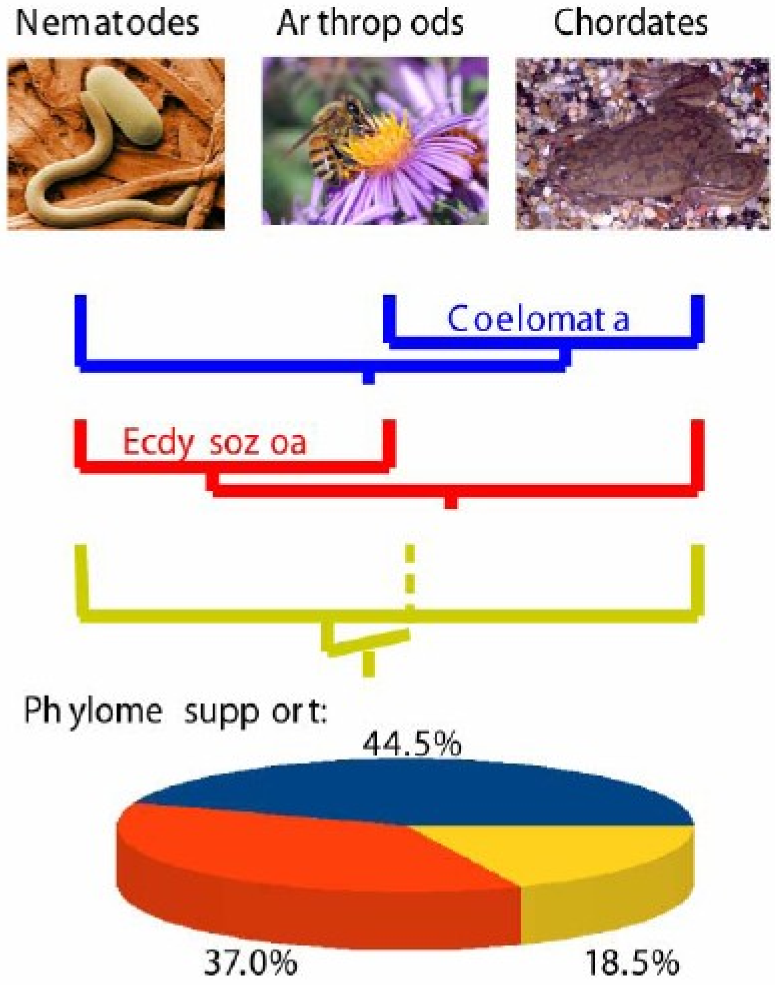


Rannala B, Yang Z. 2008.
Annu. Rev. Genomics Hum. Genet. 9:217–31

Gene conversion
Hybridization
Introgression

Uncertainty in species trees and topological variability in gene trees



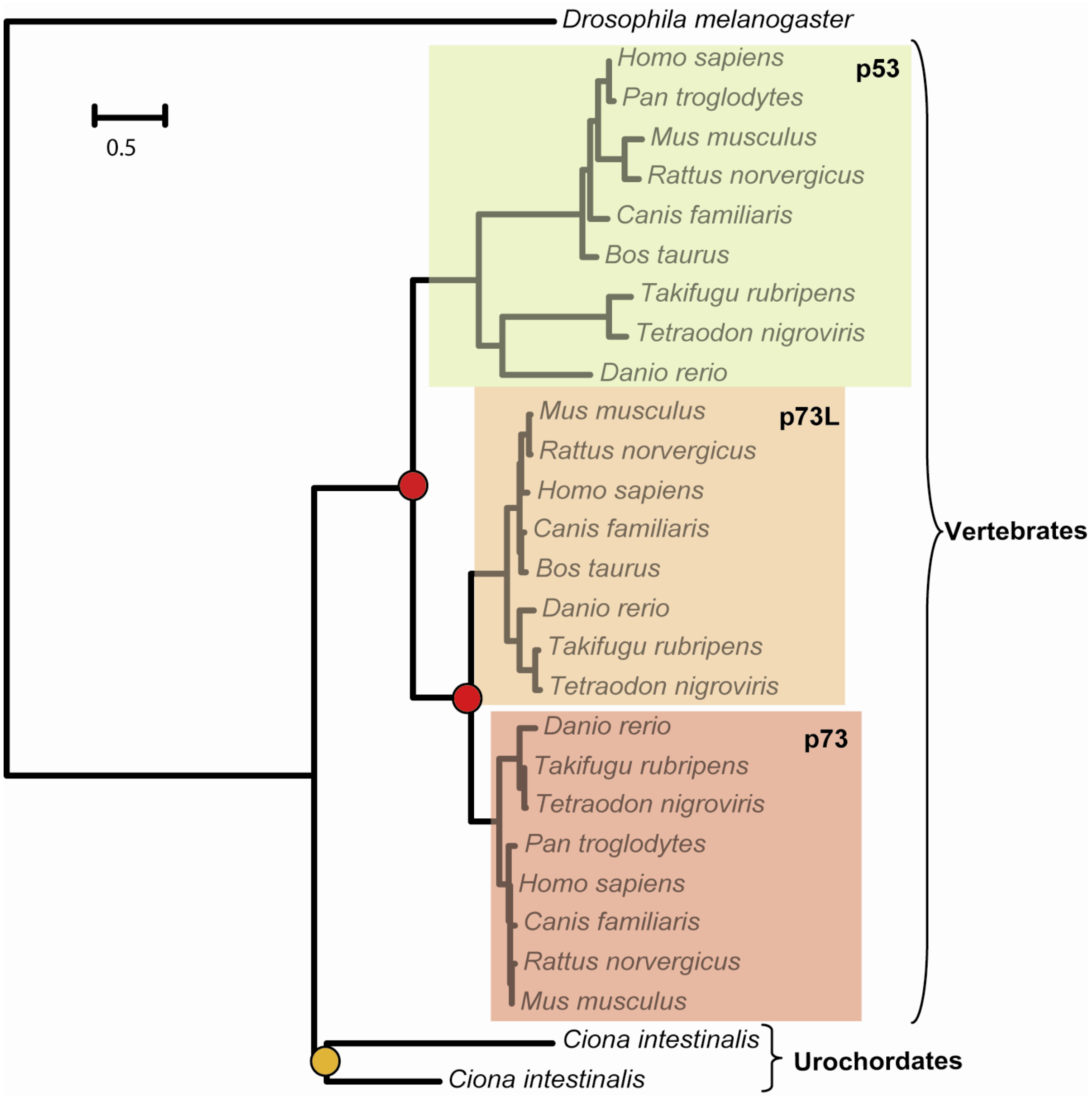


What percentage of gene trees from the human phylome support each topology?

Similar results for

Primates
Rodents
laurasatheria

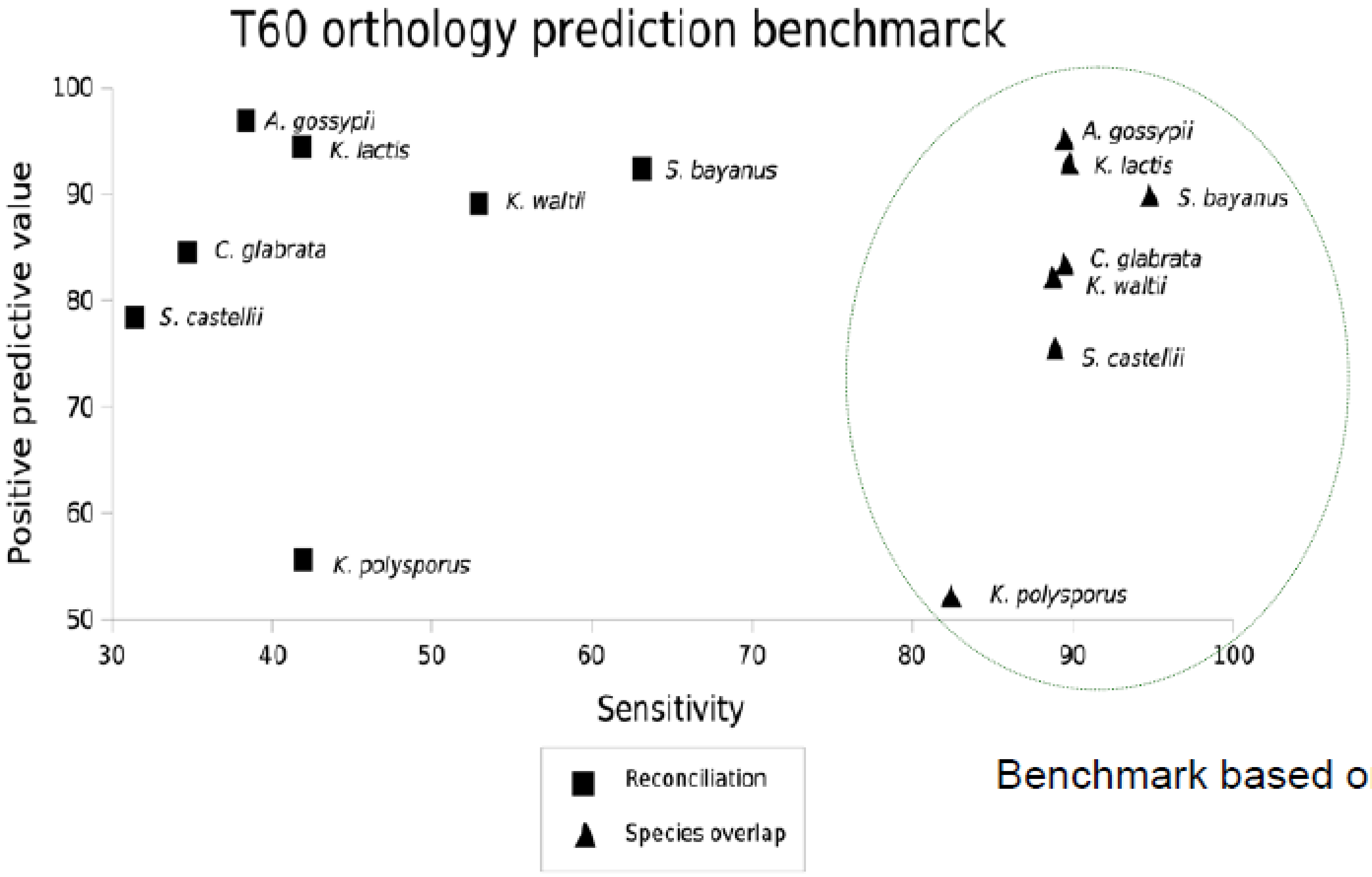
Species overlap to detect duplications and speciations



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

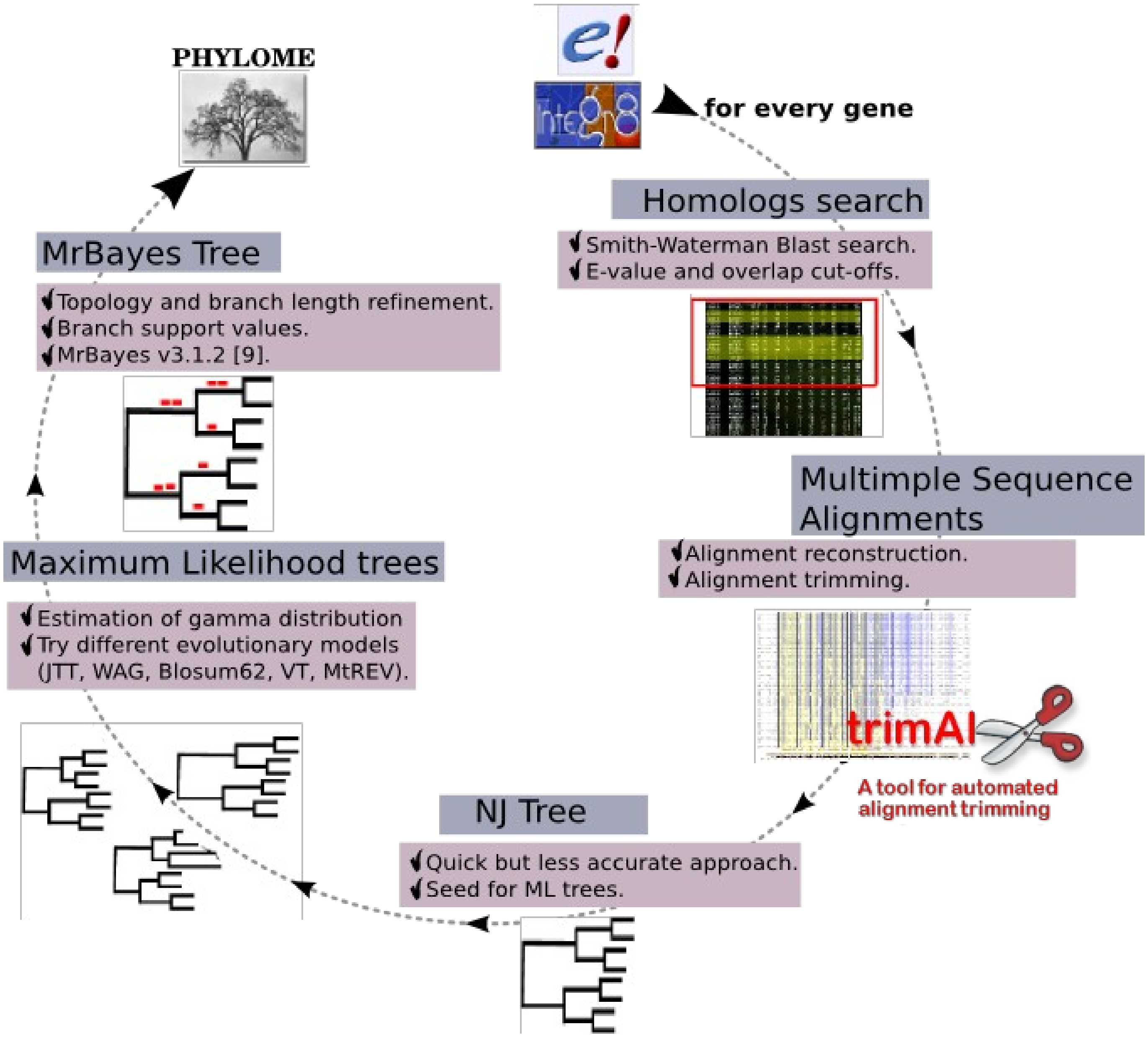
Tree reconciliation / species overlap

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



for every gene

Homologs search

✓Smith-Waterman Blast search.
✓E-value and overlap cut-offs.

Multimple Sequence Alignments

✓Alignment reconstruction.
✓Alignment trimming.

NJ Tree

✓Quick but less accurate approach.
✓Seed for ML trees.

Maximum Likelihood trees

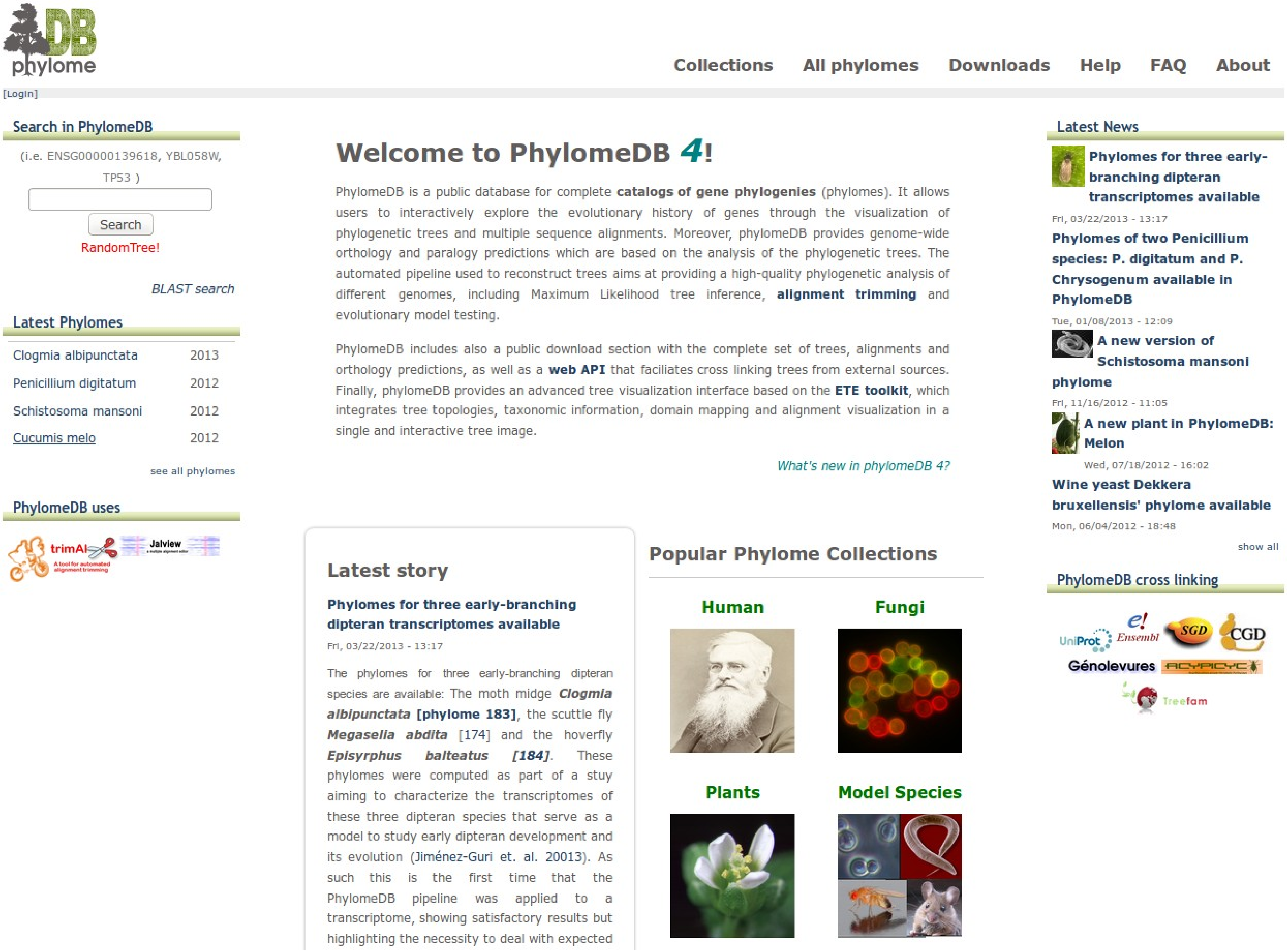
✓Estimation of gamma distribution
✓Try different evolutionary models (JTT, WAG, Blosum62, VT, MtREV).

MrBayes Tree

✓Topology and branch length refinement.
✓Branch support values.
✓MrBayes v3.1.2 [9].

PHYLOME

www.phylomedb.org



Search in PhylomeDB

(i.e. ENSG00000139618, YBL058W, TP53)

Latest Phylomes

Clogmia albipunctata	2013
Penicillium digitatum	2012
Schistosoma mansoni	2012
Cucumis melo	2012

see all phylomes

PhylomeDB uses

Welcome to PhylomeDB *4*!

PhylomeDB is a public database for complete **catalogs 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 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, as well as a **web API** that faciliates cross linking trees from external sources. Finally, phylomeDB provides an advanced tree visualization interface based on the **ETE toolkit**, which integrates tree topologies, taxonomic information, domain mapping and alignment visualization in a single and interactive tree image.

What's new in phylomeDB 4?

Latest story

Phylomes for three early-branching dipteran transcriptomes available

Fri, 03/22/2013 - 13:17

The phylomes for three early-branching dipteran species are available: The moth midge ***Clogmia albipunctata*** **[phylome 183]**, the scuttle fly ***Megaselia abdita*** [174] and the hoverfly ***Episyrphus balteatus*** ***[184]***. These phylomes were computed as part of a stuy aiming to characterize the transcriptomes of these three dipteran species that serve as a model to study early dipteran development and its evolution (Jiménez-Guri et. al. 20013). As such this is the first time that the PhylomeDB pipeline was applied to a transcriptome, showing satisfactory results but highlighting the necessity to deal with expected

Popular Phylome Collections

Latest News

Phylomes for three early-branching dipteran transcriptomes available

Fri, 03/22/2013 - 13:17

Phylomes of two Penicillium species: P. digitatum and P. Chrysogenum available in PhylomeDB

Tue, 01/08/2013 - 12:09

A new version of Schistosoma mansoni phylome

Fri, 11/16/2012 - 11:05

A new plant in PhylomeDB: Melon

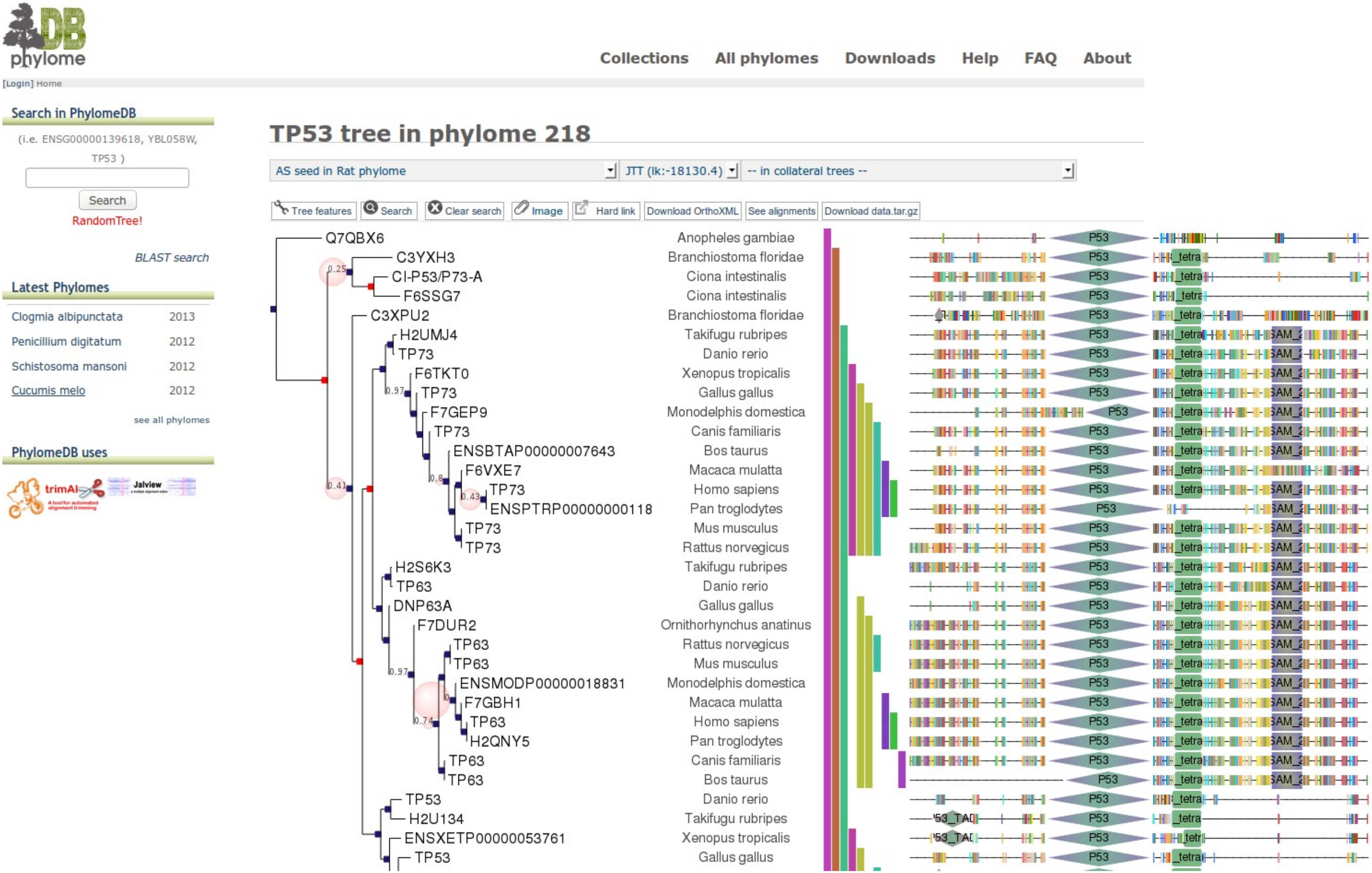
Wed, 07/18/2012 - 16:02

Wine yeast Dekkera bruxellensis' phylome available

Mon, 06/04/2012 - 18:48

show all

PhylomeDB cross linking

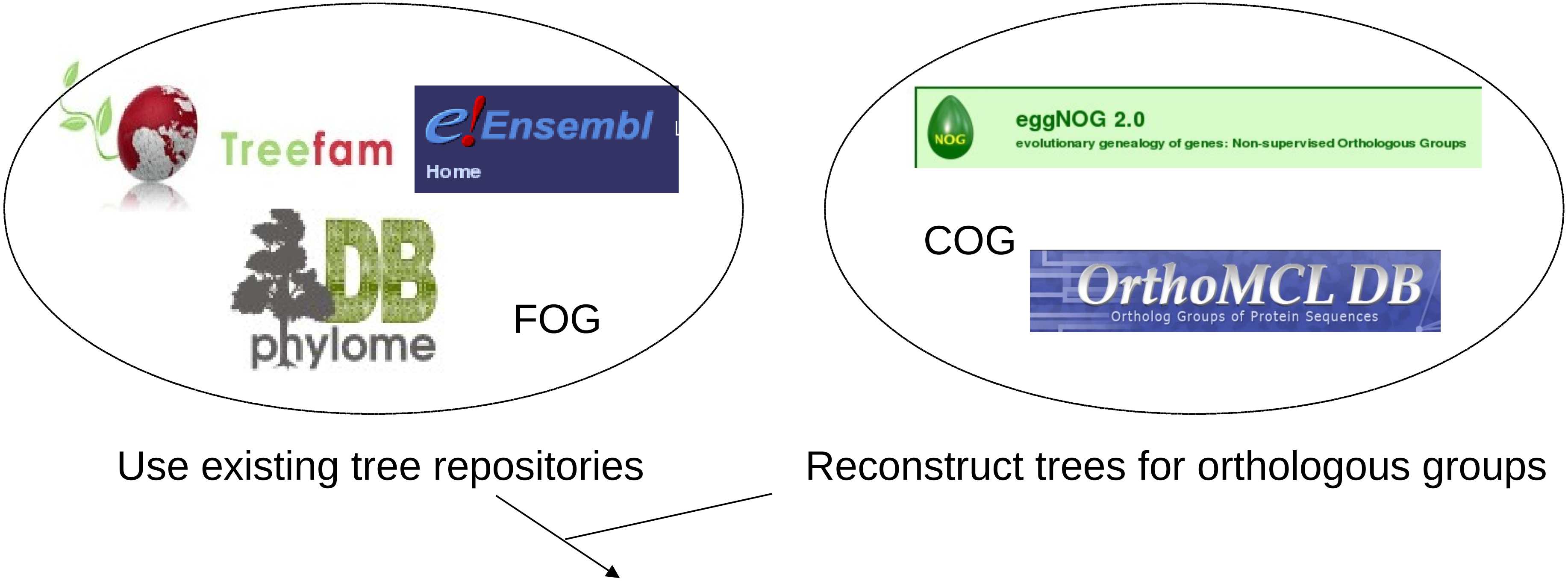


TP53 tree in phylome 218

Q7QBX6	Anopheles gambiae
C3YXH3	Branchiostoma floridae
CI-P53/P73-A	Ciona intestinalis
F6SSG7	Ciona intestinalis
C3XPU2	Branchiostoma floridae
H2UMJ4	Takifugu rubripes
TP73	Danio rerio
F6TKT0	Xenopus tropicalis
TP73	Gallus gallus
F7GEP9	Monodelphis domestica
TP73	Canis familiaris
ENSBTAP00000007643	Bos taurus
F6VXE7	Macaca mulatta
TP73	Homo sapiens
ENSPTRP00000000118	Pan troglodytes
TP73	Mus musculus
TP73	Rattus norvegicus
H2S6K3	Takifugu rubripes
TP63	Danio rerio
DNP63A	Gallus gallus
F7DUR2	Ornithorhynchus anatinus
TP63	Rattus norvegicus
TP63	Mus musculus
ENSMODP00000018831	Monodelphis domestica
F7GBH1	Macaca mulatta
TP63	Homo sapiens
H2QNY5	Pan troglodytes
TP63	Canis familiaris
TP63	Bos taurus
TP53	Danio rerio
H2U134	Takifugu rubripes
ENSXETP00000053761	Xenopus tropicalis
TP53	Gallus gallus

MetaPhOrs

(Meta-Phylogeny-Based-Orthologs)


Treefam
Ensembl
Home
phylome DB
eggNOG 2.0
evolutionary genealogy of genes: Non-supervised Orthologous Groups
NOG
OrthoMCL DB
Ortholog Groups of Protein Sequences

FOG

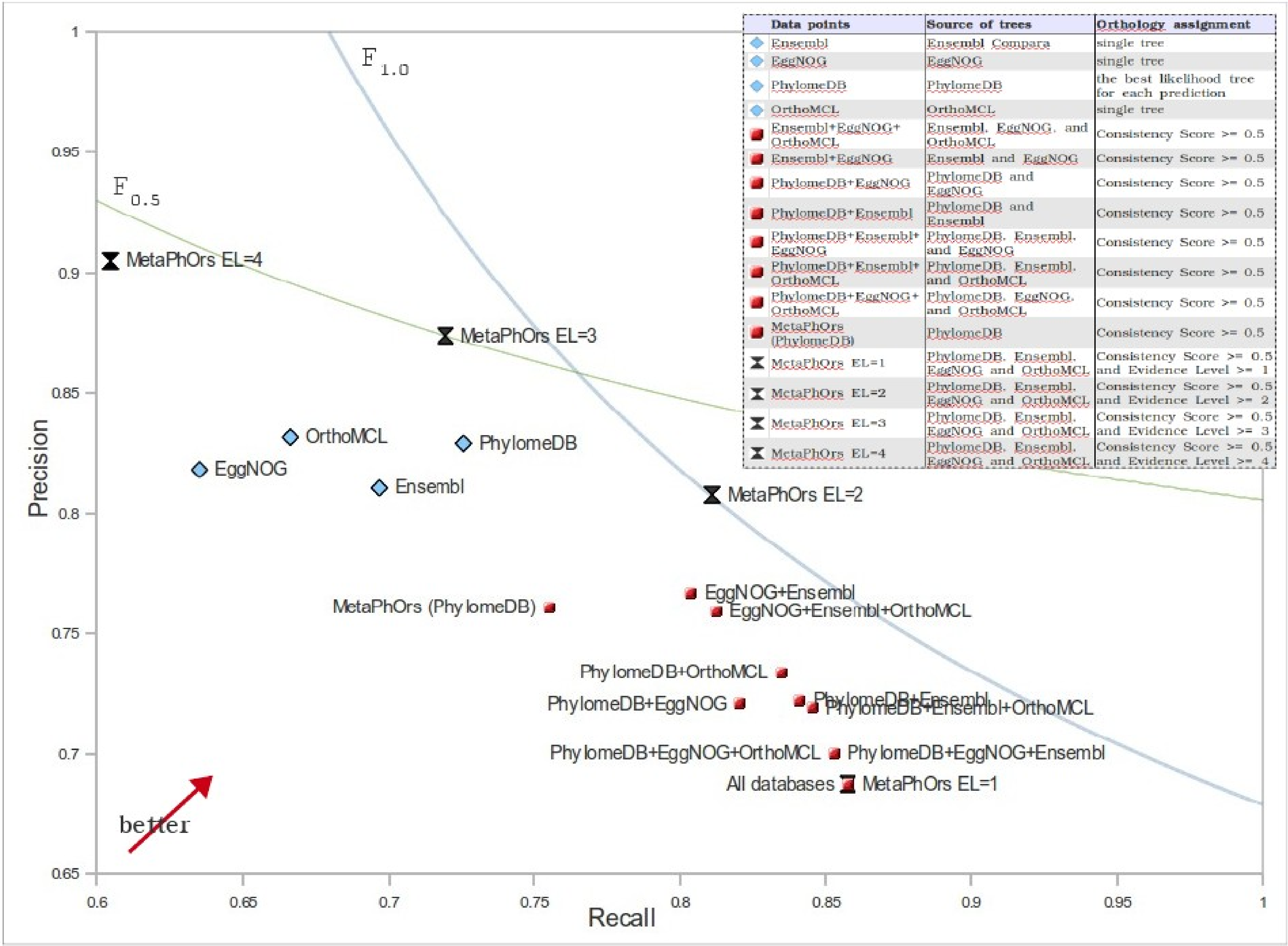
COG

Use existing tree repositories

Reconstruct trees for orthologous groups

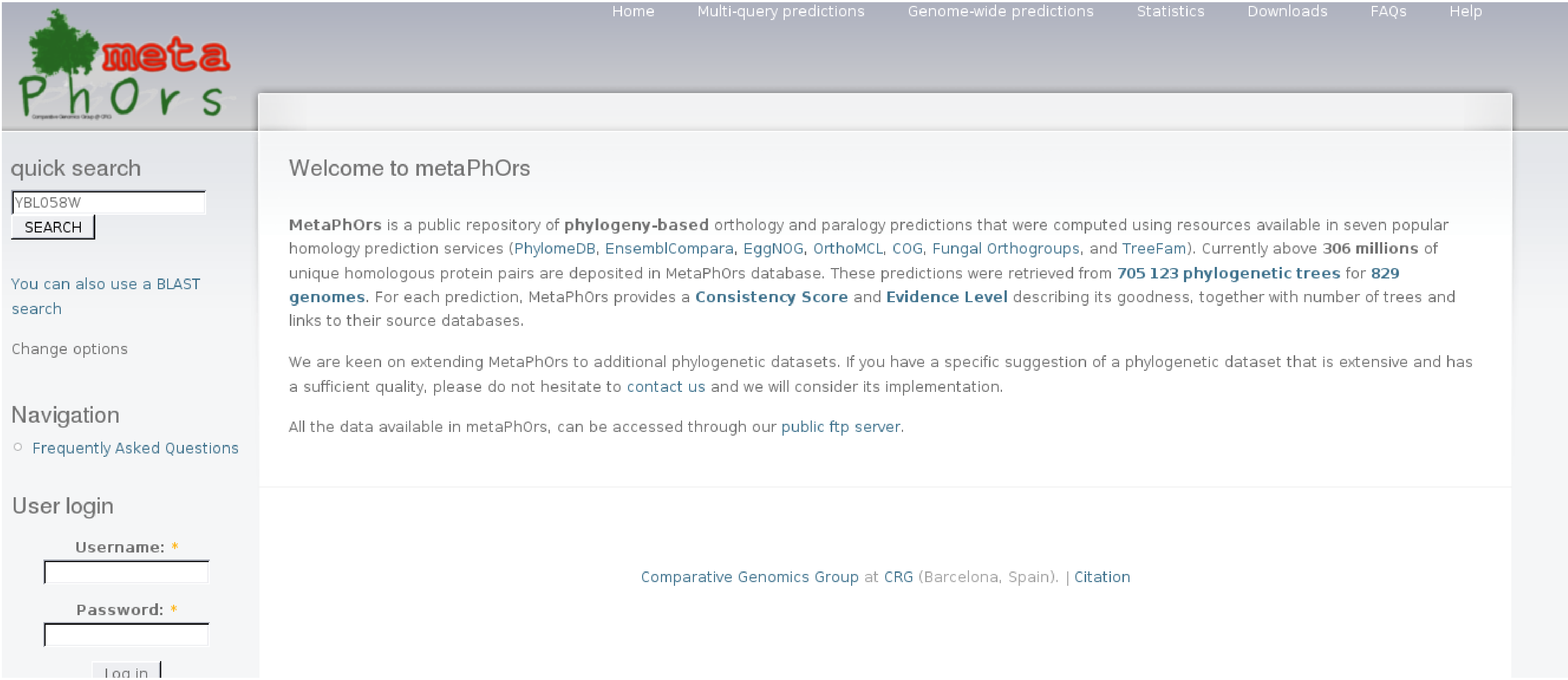
Integrate and use consistency across datasets as a proxy of reliability

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



	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



You can also use a BLAST search

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.

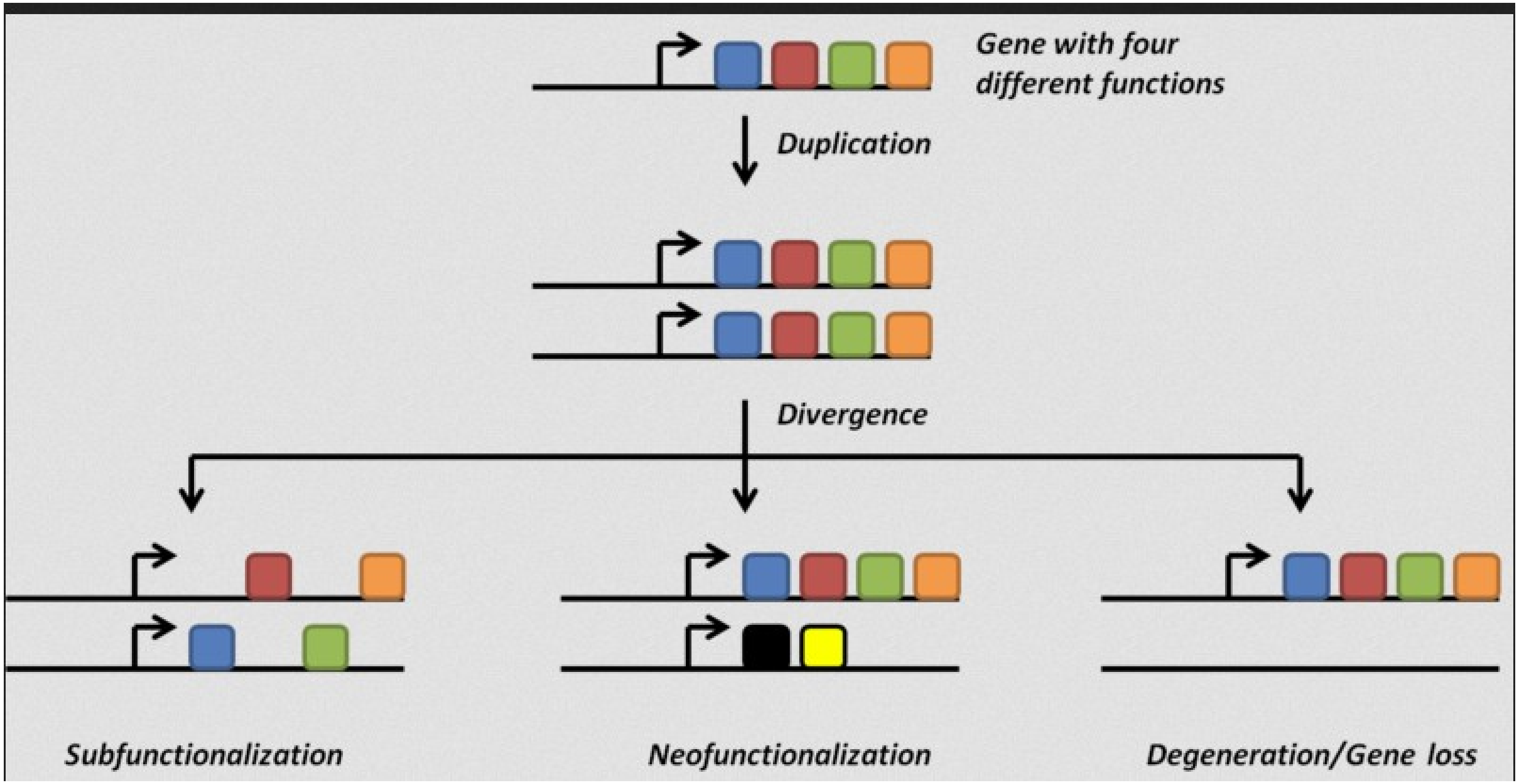
We are keen on extending MetaPhOrs to additional phylogenetic datasets. If you have a specific suggestion of a phylogenetic dataset that is extensive and has a sufficient quality, please do not hesitate to contact us and we will consider its implementation.

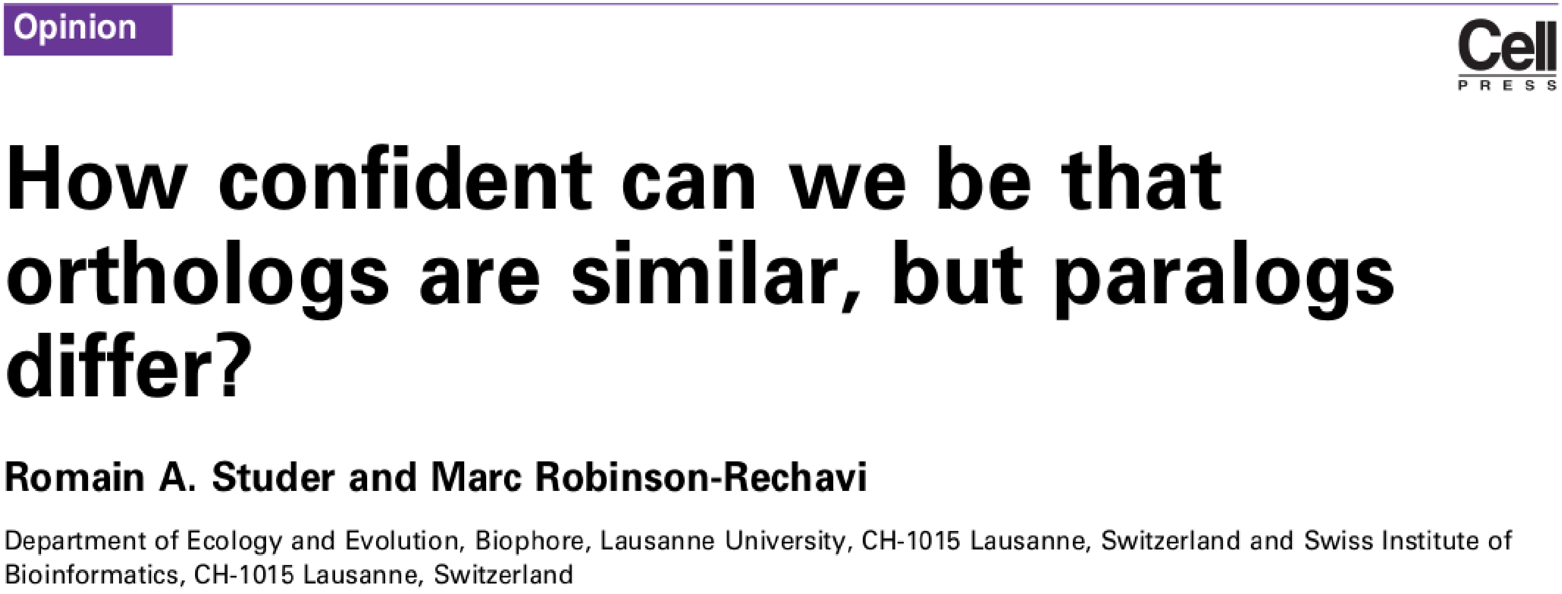
All the data available in metaPhOrs, can be accessed through our public ftp server.

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

Functional Implications of orthology

After duplication: diversify or die (neofunctionalization or subfunctionalization models)





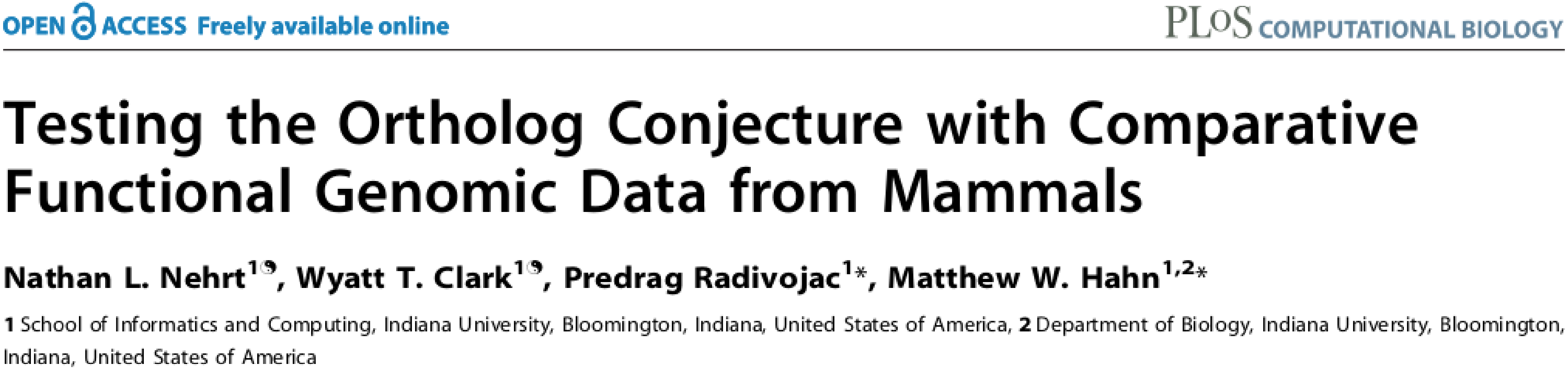
Opinion

Cell PRESS

How confident can we be that orthologs are similar, but paralogs differ?

Romain A. Studer and Marc Robinson-Rechavi

Department of Ecology and Evolution, Biophore, Lausanne University, CH-1015 Lausanne, Switzerland and Swiss Institute of Bioinformatics, CH-1015 Lausanne, Switzerland



OPEN ACCESS Freely available online

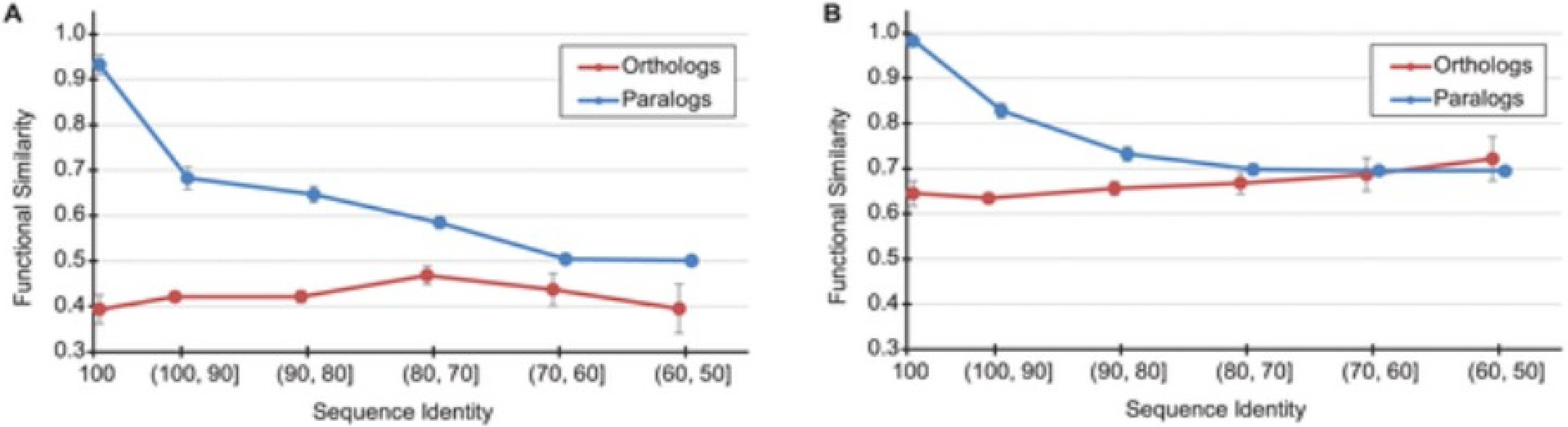
PLoS COMPUTATIONAL BIOLOGY

Testing the Ortholog Conjecture with Comparative Functional Genomic Data from Mammals

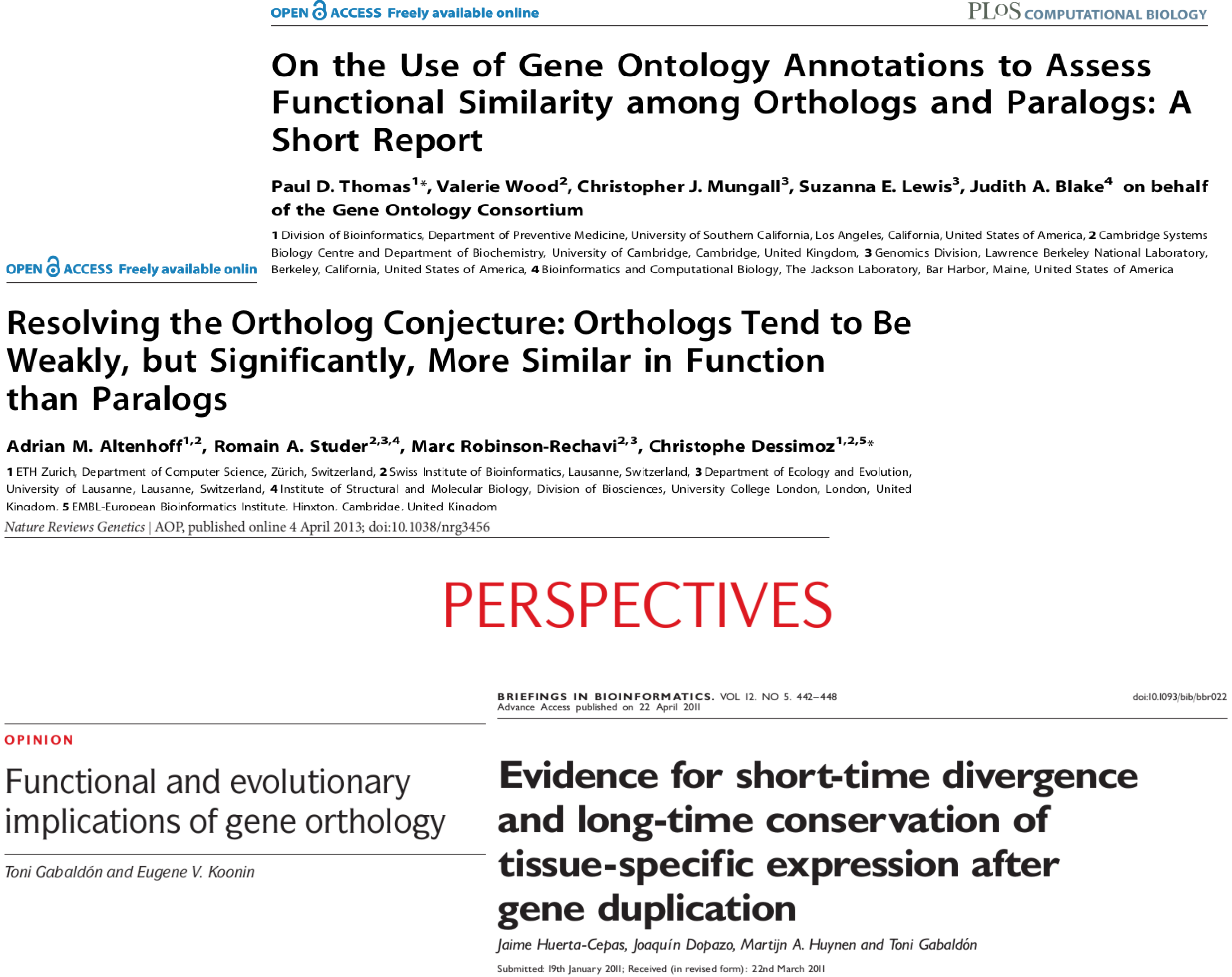
Nathan L. Nehrt[1¶], Wyatt T. Clark[1¶], Predrag Radivojac[1]*, Matthew W. Hahn[1,2]*

1 School of Informatics and Computing, Indiana University, Bloomington, Indiana, United States of America, **2** Department of Biology, Indiana University, Bloomington, Indiana, United States of America

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



OPEN ACCESS Freely available online PLoS COMPUTATIONAL BIOLOGY

On the Use of Gene Ontology Annotations to Assess Functional Similarity among Orthologs and Paralogs: A Short Report

Paul D. Thomas[1]*, Valerie Wood[2], Christopher J. Mungall[3], Suzanna E. Lewis[3], Judith A. Blake[4] on behalf of the Gene Ontology Consortium

1 Division of Bioinformatics, Department of Preventive Medicine, University of Southern California, Los Angeles, California, United States of America, 2 Cambridge Systems Biology Centre and Department of Biochemistry, University of Cambridge, Cambridge, United Kingdom, 3 Genomics Division, Lawrence Berkeley National Laboratory, Berkeley, California, United States of America, 4 Bioinformatics and Computational Biology, The Jackson Laboratory, Bar Harbor, Maine, United States of America

OPEN ACCESS Freely available onlin

Resolving the Ortholog Conjecture: Orthologs Tend to Be Weakly, but Significantly, More Similar in Function than Paralogs

Adrian M. Altenhoff[1,2], Romain A. Studer[2,3,4], Marc Robinson-Rechavi[2,3], Christophe Dessimoz[1,2,5]*

1 ETH Zurich, Department of Computer Science, Zürich, Switzerland, 2 Swiss Institute of Bioinformatics, Lausanne, Switzerland, 3 Department of Ecology and Evolution, University of Lausanne, Lausanne, Switzerland, 4 Institute of Structural and Molecular Biology, Division of Biosciences, University College London, London, United Kingdom, 5 EMBL-European Bioinformatics Institute, Hinxton, Cambridge, United Kingdom

Nature Reviews Genetics | AOP, published online 4 April 2013; doi:10.1038/nrg3456

PERSPECTIVES

OPINION

Functional and evolutionary implications of gene orthology

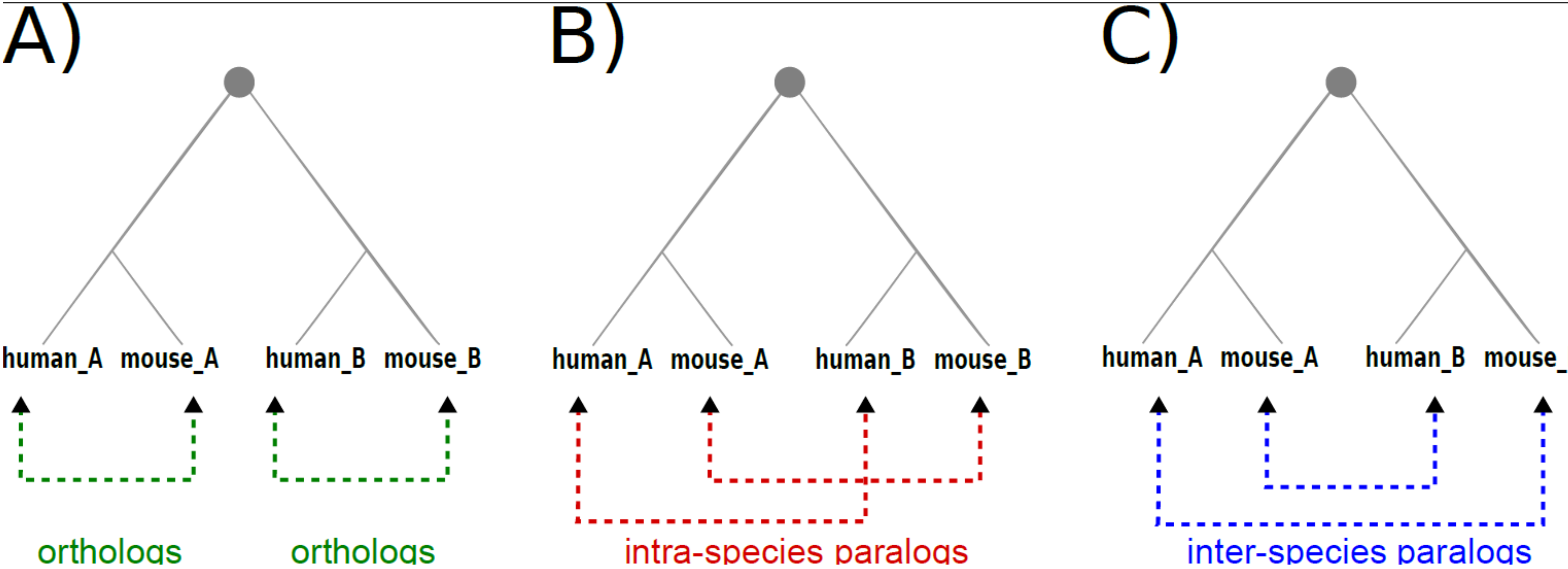
Toni Gabaldón and Eugene V. Koonin

BRIEFINGS IN BIOINFORMATICS. VOL 12. NO 5. 442–448
Advance Access published on 22 April 2011
doi:10.1093/bib/bbr022

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

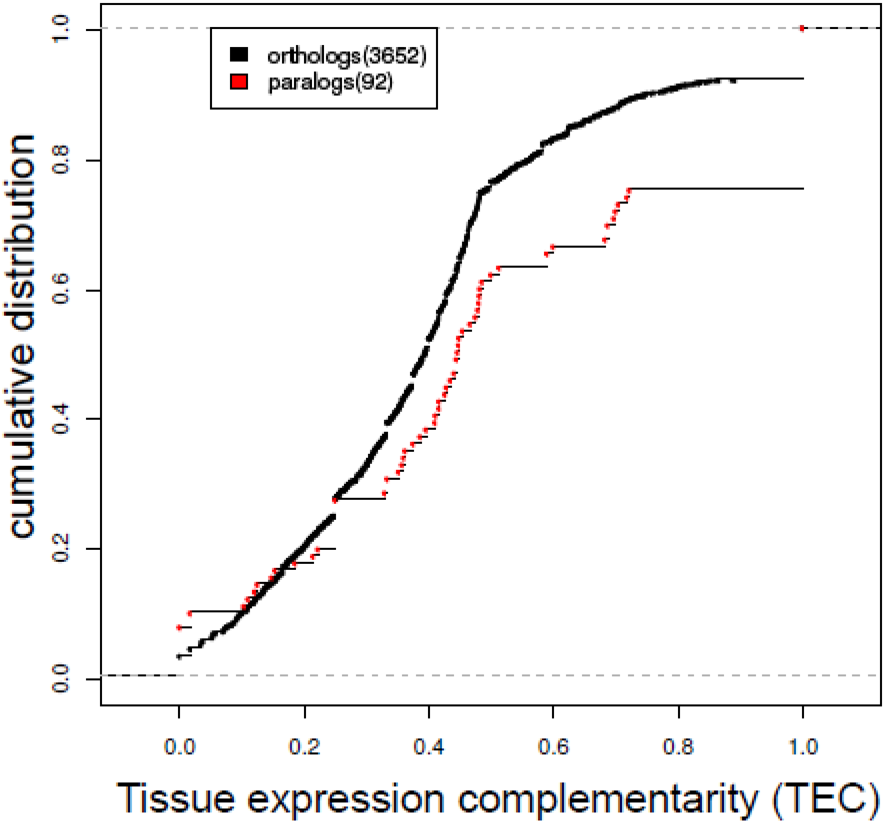
Jaime Huerta-Cepas, Joaquín Dopazo, Martijn A. Huynen and Toni Gabaldón

Submitted: 19th January 2011; Received (in revised form): 22nd March 2011

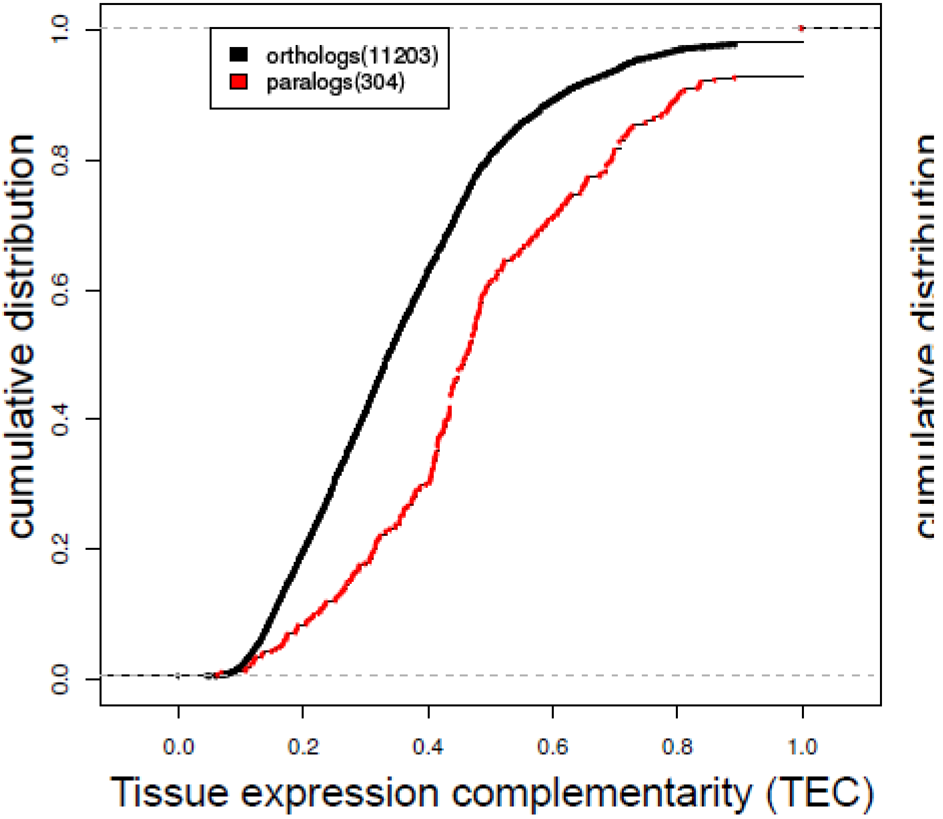


Comparison of differences in tissue-specific patterns of expression across orthologs and paralogs.

Huerta-Cepas. et. al. (Brief. In Bioinf. Special issue on orthology)



SymAtlas (A/P calls)



Bgee (A/P calls)

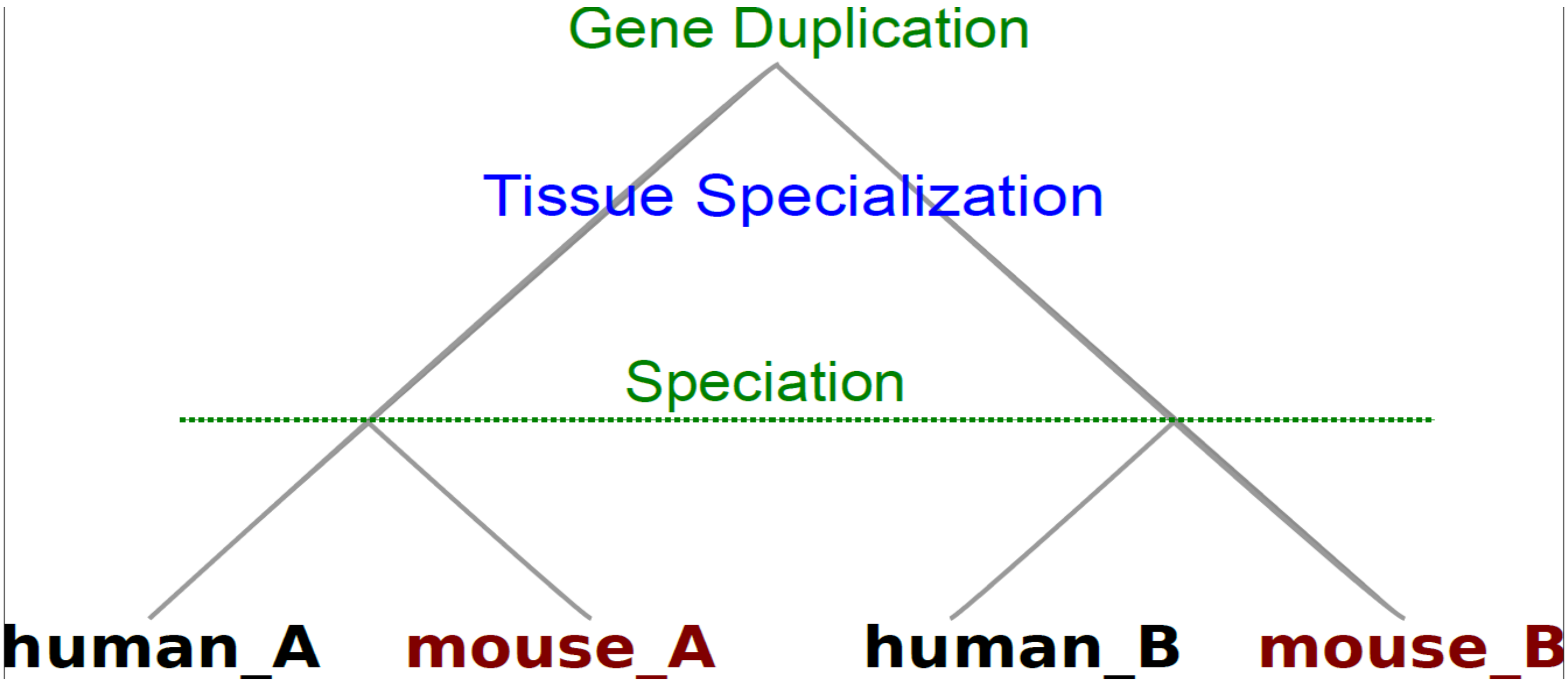
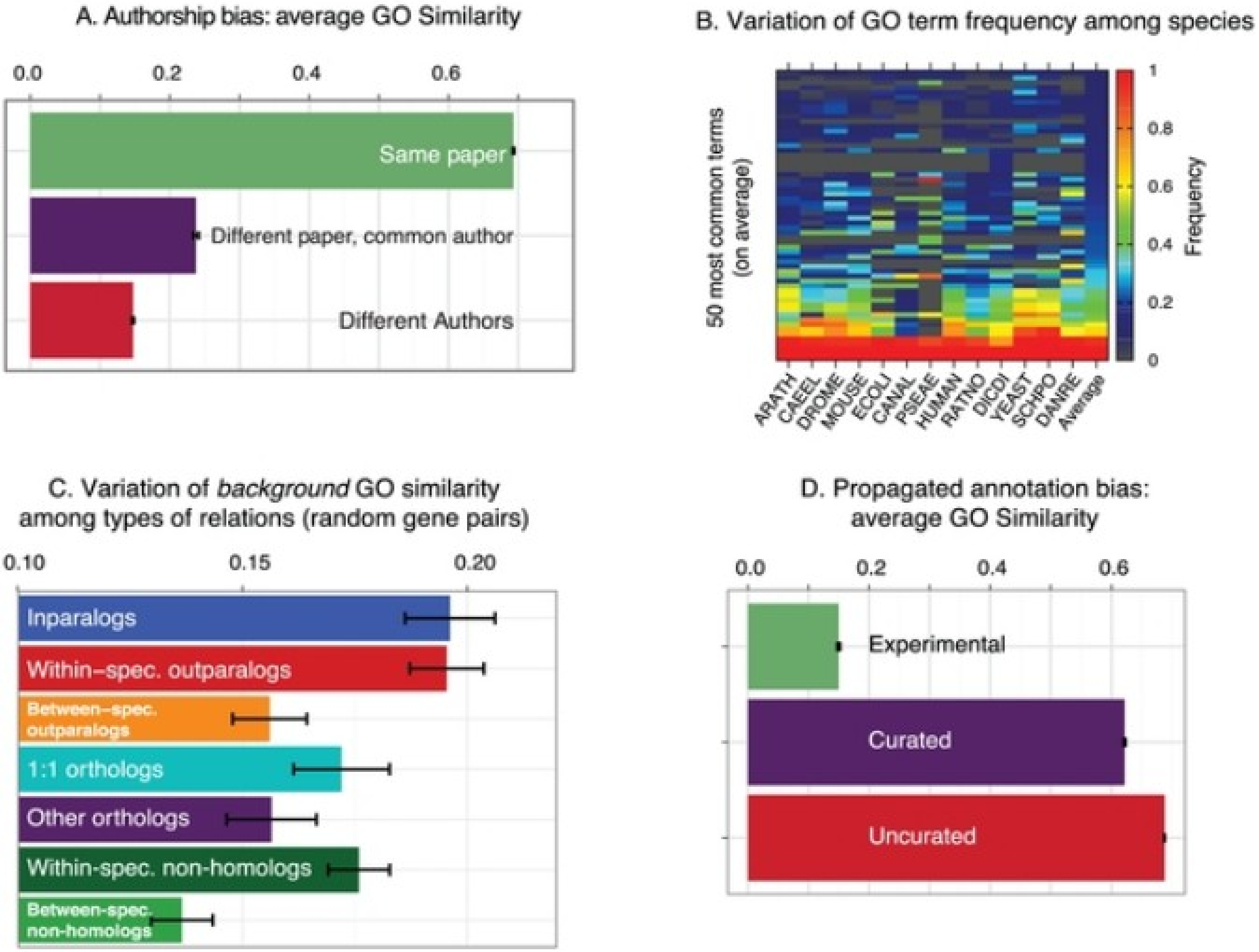
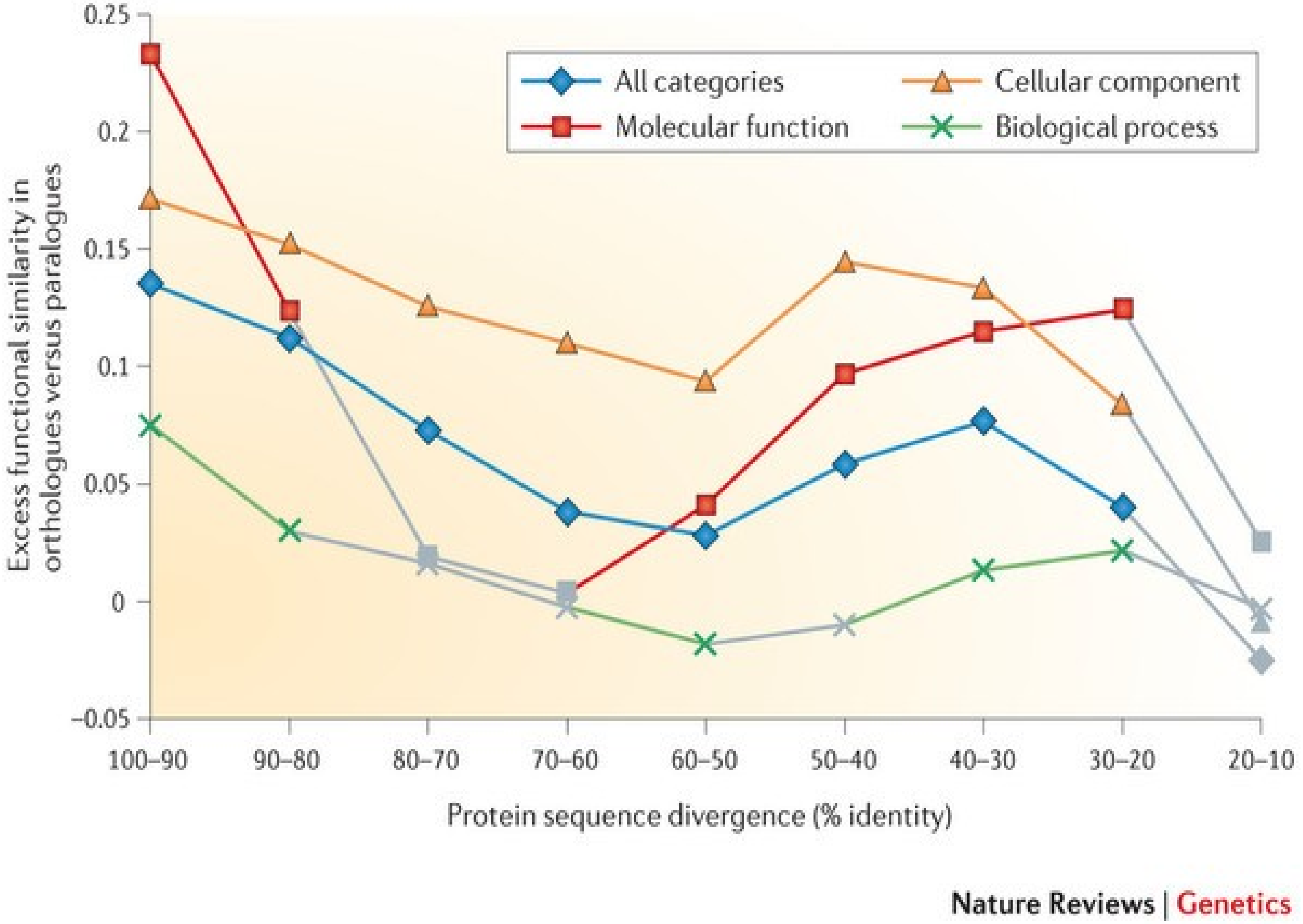

Gene Duplication
Tissue Specialization
Speciation
human_A
mouse_A
human_B
mouse_B

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



Gabaldón and Koonin (2013) Nat. Rev. Gen.

Thanks