Functional and Comparative Genomics

ESCI upf.
School of International Studies

Comparative and Functional Genomics

Session 2
Comparative sequence analysis

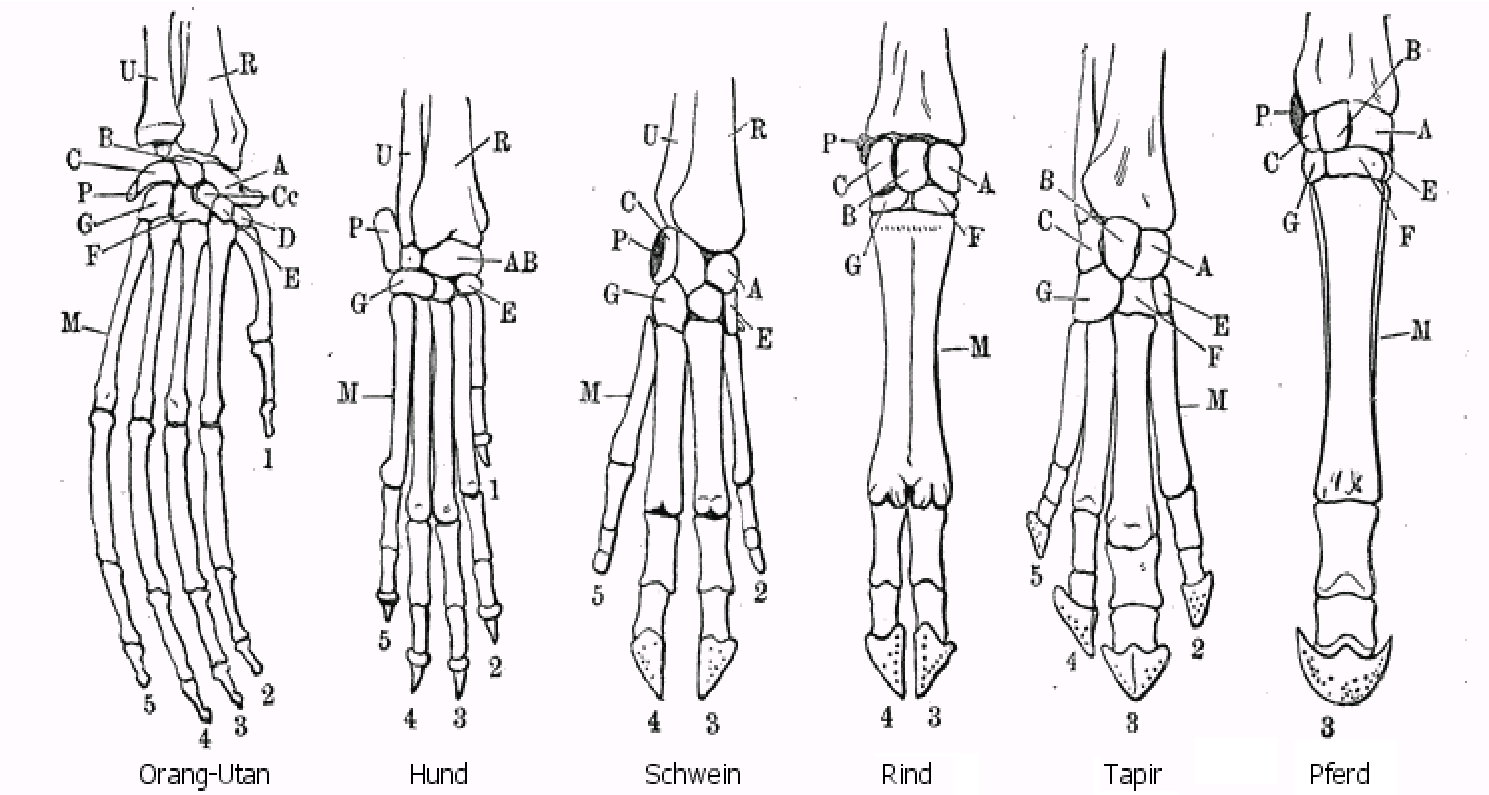
ESCI upf.
School of International Studies

Comparative sequence analysis

- Homology, Paralogy and Orthology.
- Methods for predicting orthology and paralogy clustering-based and phylogeny-based.
- Gene families.
- Gene duplication, neo- and sub-functionalization.
- Gene family expansions and contractions.
- Adaptation and genome evolution.

- **Homology, Paralogy and Orthology**

Homology

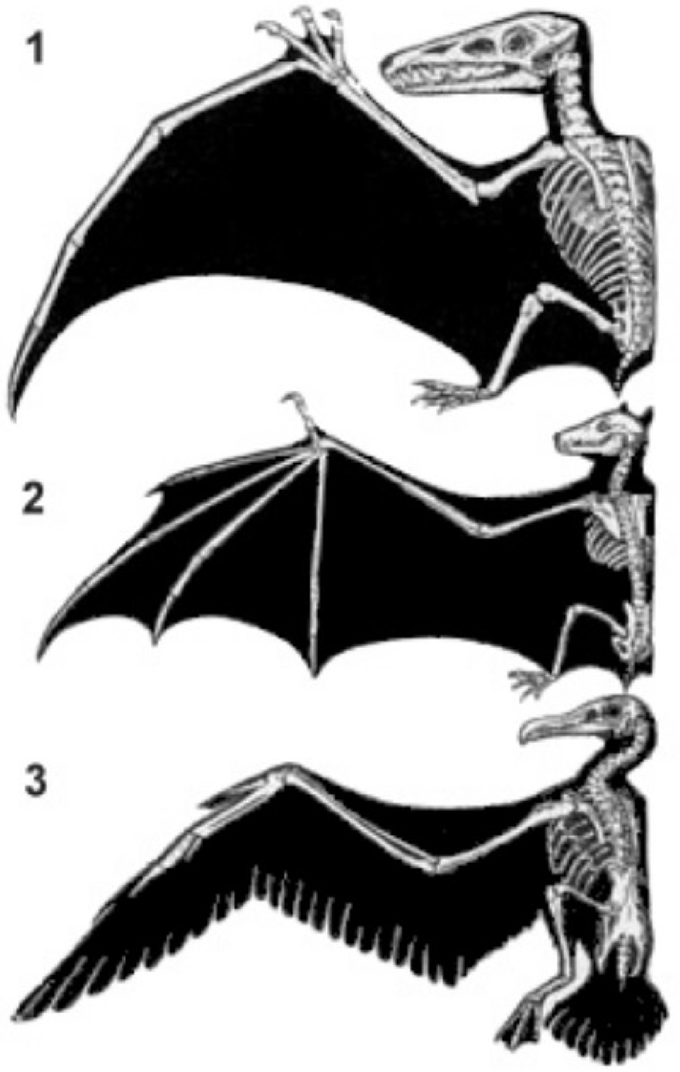


Handskelette von Säugetieren

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. Owen → organs in two species are homologous only if the same structure was present in their last common ancestor. Homology → common ancestry

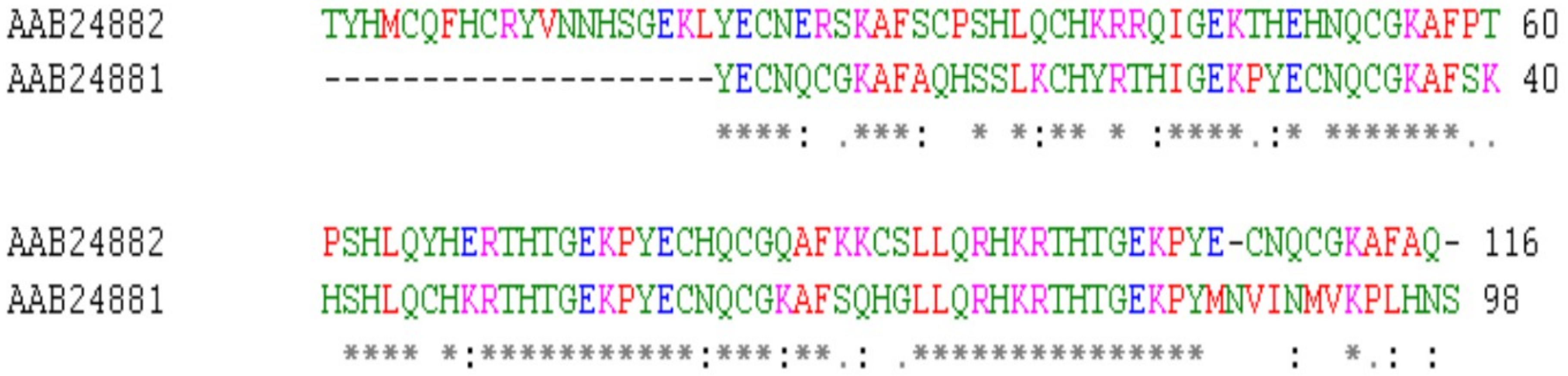
Analogous structures:
Similar function but independet origin.



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.

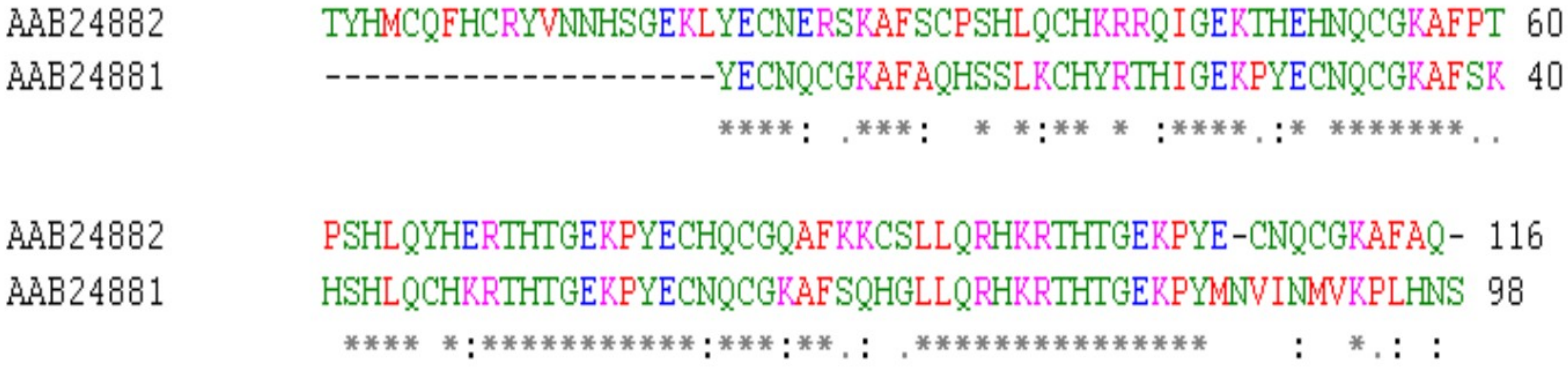
“the sequences are 50% homologous”, “these two sequences are highly homologous”

Why is this incorrect?

Where does the confusion comes from?

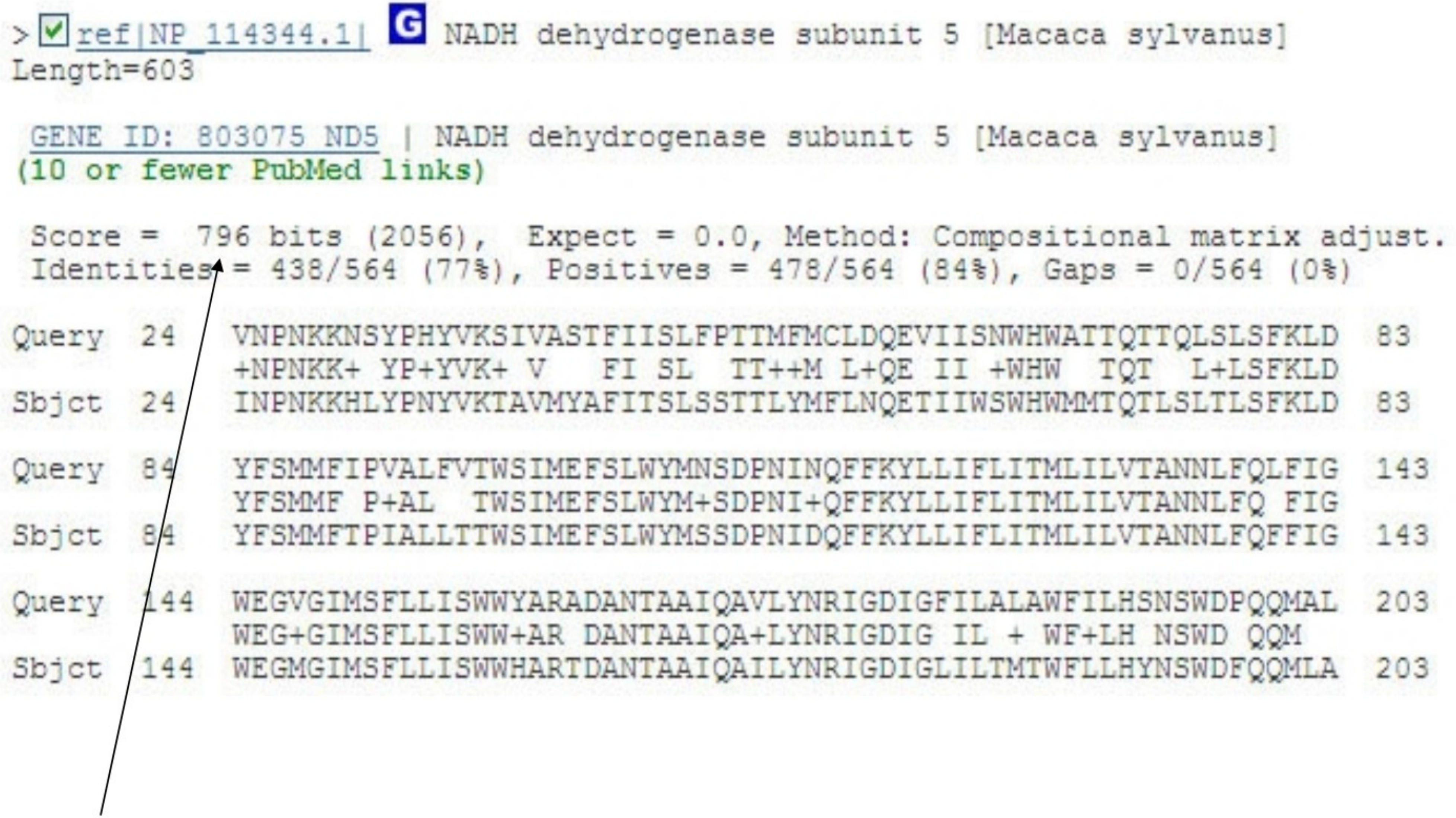
Detour

Sequence similarity, homology detection and blast database queries



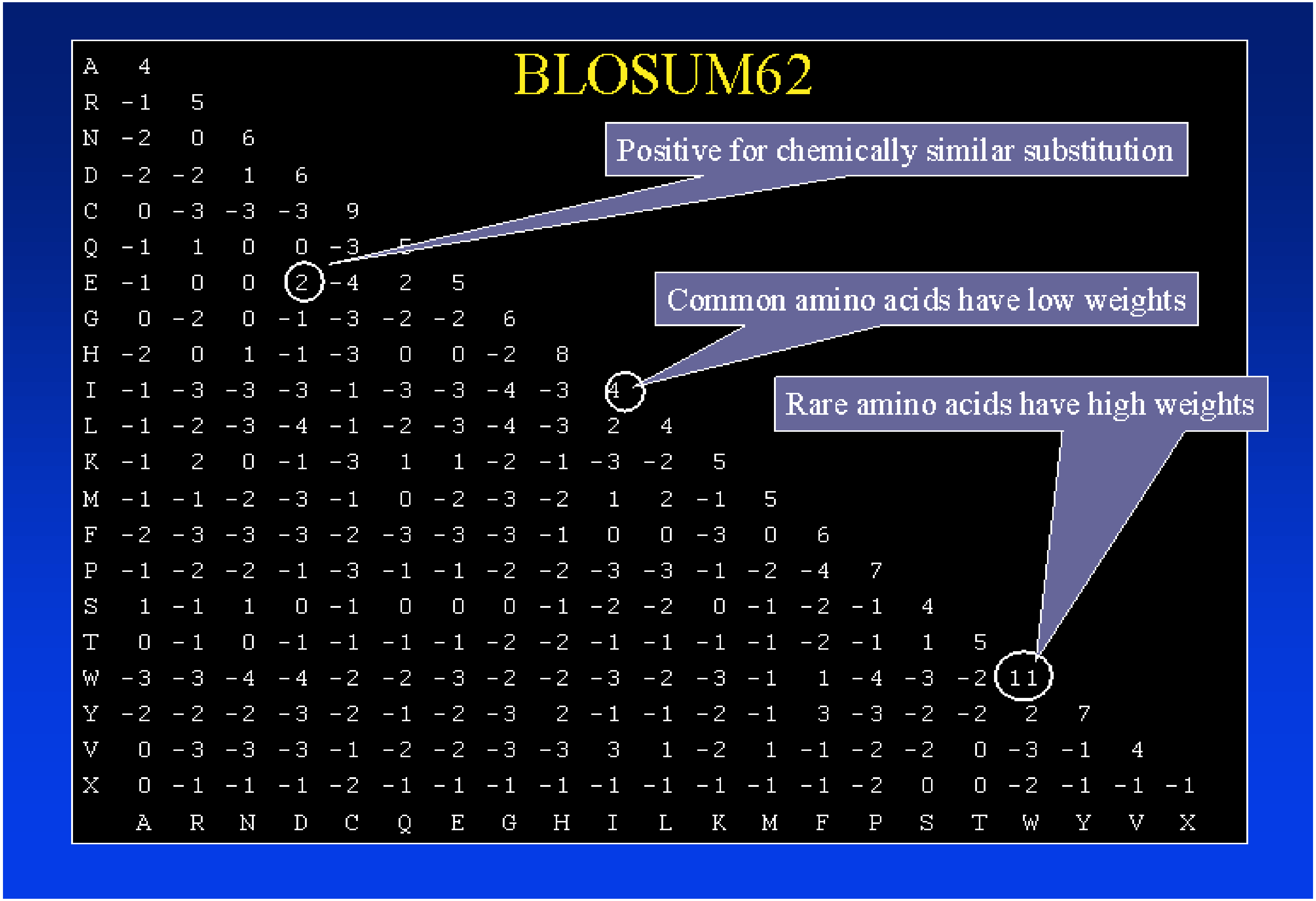
```
AAB24882        TYHMCQFHCRYVNNHSGEKLYECNERSKAFSCPSHLQCHKRRQIGEKTHEHNQCGKAFPT 60
AAB24881        --------------------YECNQCGKAFAQHSSLKCHYRTHIGEKPYECNQCGKAFSK 40
                                    ****: .***:  * *:** * :****.:* *******..

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| G 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
```

Alignment scores are sums of residue pairing scores according to a scoring Matrix



BLOSUM62

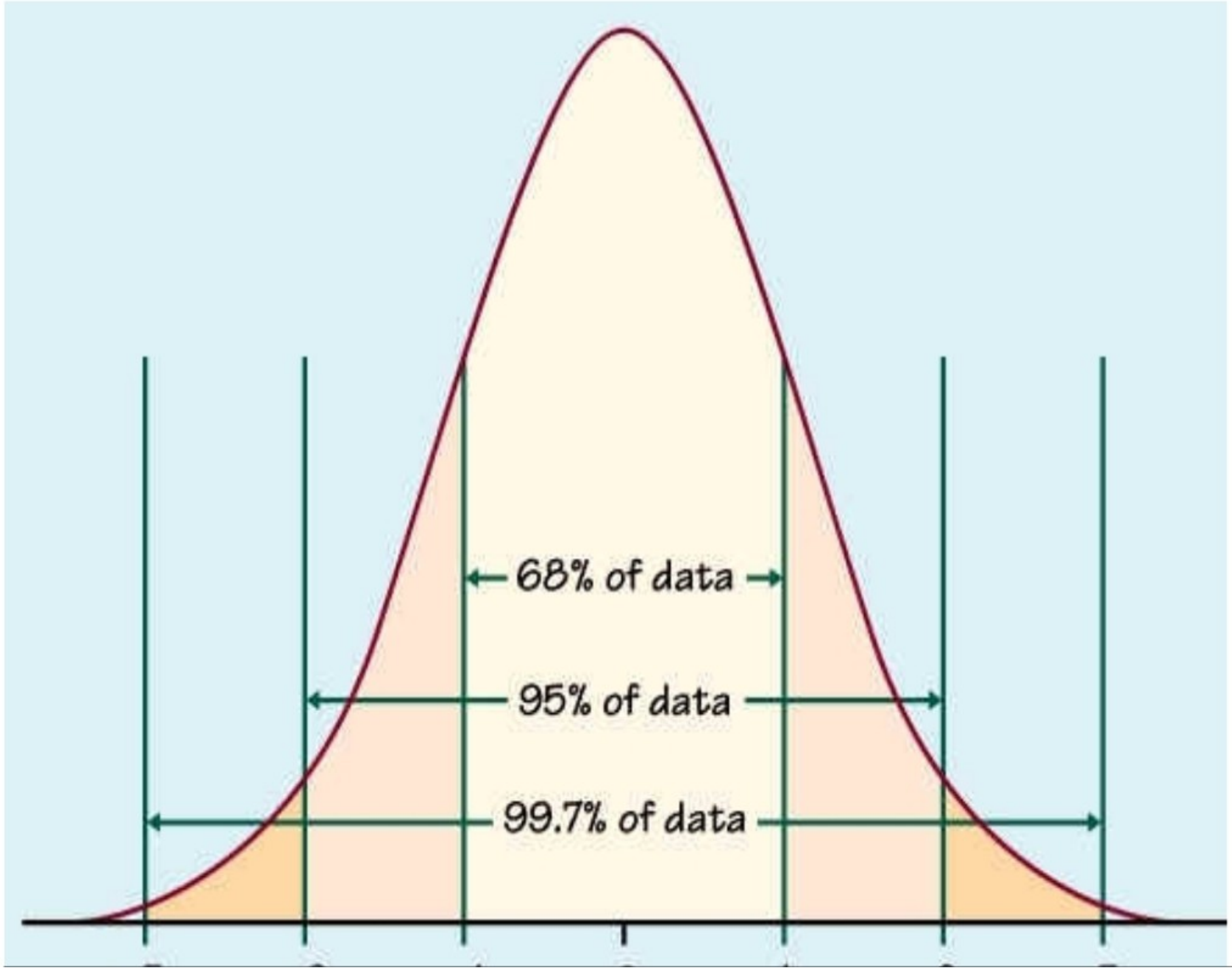
	A	R	N	D	C	Q	E	G	H	I	L	K	M	F	P	S	T	W	Y	V	X
A	4																				
R	-1	5																			
N	-2	0	6																		
D	-2	-2	1	6																	
C	0	-3	-3	-3	9																
Q	-1	1	0	0	-3	5															
E	-1	0	0	2	-4	2	5														
G	0	-2	0	-1	-3	-2	-2	6													
H	-2	0	1	-1	-3	0	0	-2	8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

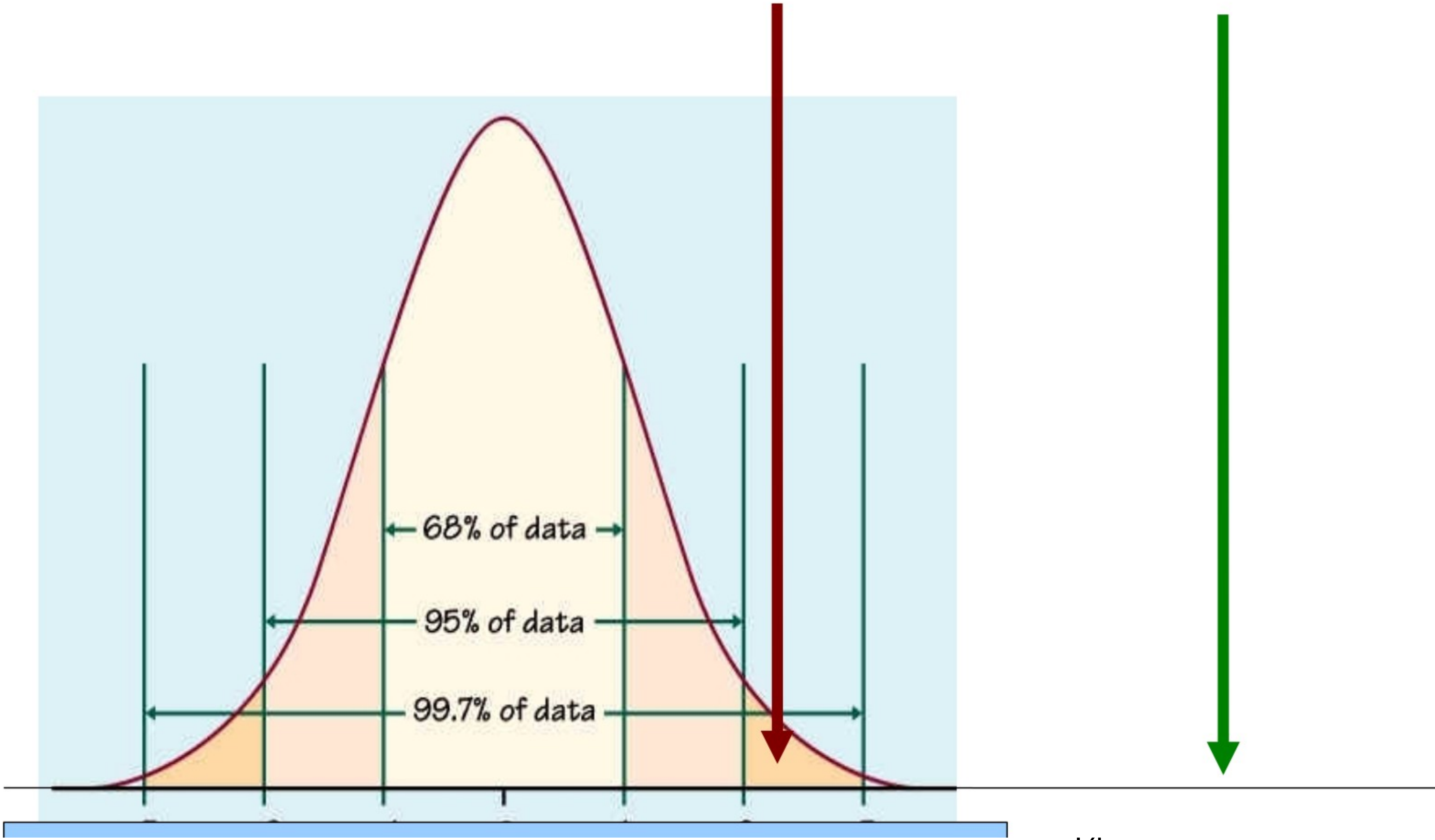
Rare amino acids have high weights

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



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

Your score

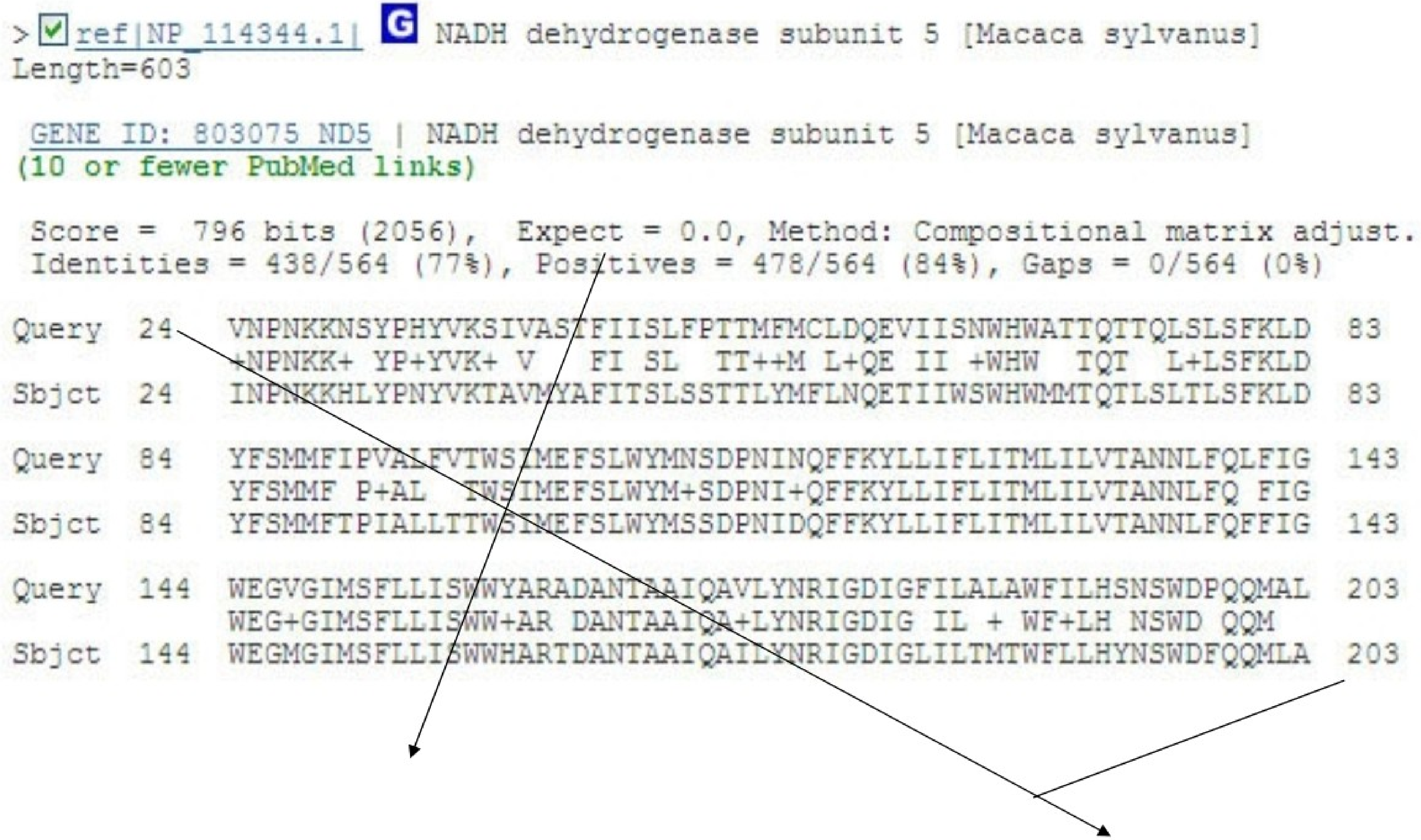


Length of hit and query

Normalized score

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

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



```
> ☑ ref|NP_114344.1|  G  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 (important when locally searching in small databases)
- Low complexity filtering
- Why multiple HSPs in a hit
- PSI-Blast, HMMER searches
- Issues with multi-domain protein

From homology to orthology

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

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

Are these sentences correct?

- Orthologs are homologous genes that have the same function
- Orthologs are homologous genes in different species, while paralogs are homologous genes in the same species
- The ortholog is the most similar sequence among the homologs in another species
- If gene A is orthologous to gene B, and gene B is orthologous to gene C, then A and C are orthologous to each other.
- Orthologs are genes that do not duplicate and, when they exist, they are always present in single copy
- After a duplication, the orthologous copy is the one that keeps the function of the ancestral gene

Fitch W.M.

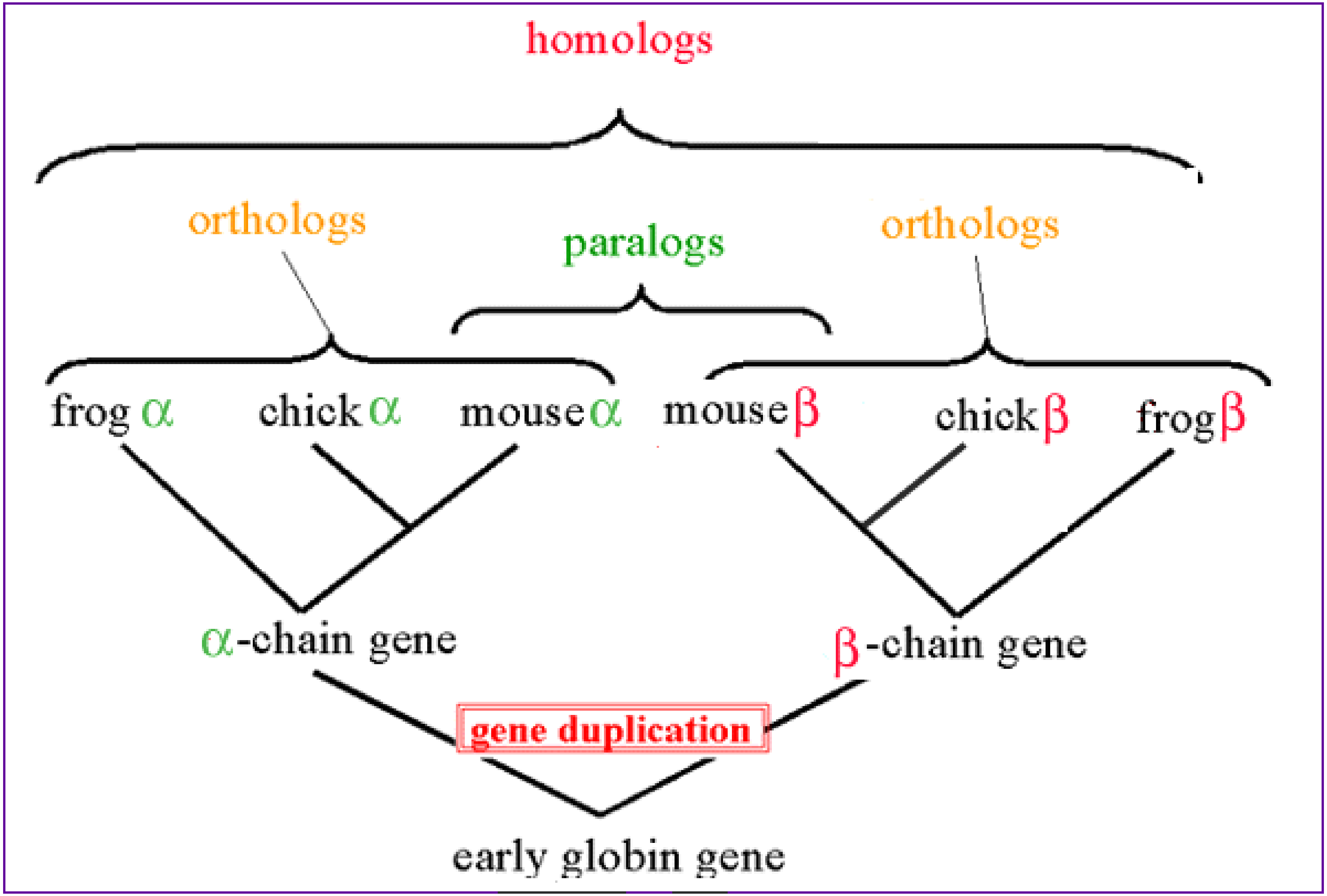
Distinguishing homologous from analogous proteins.

Syst. Zool. 1970; 19: 99-113

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

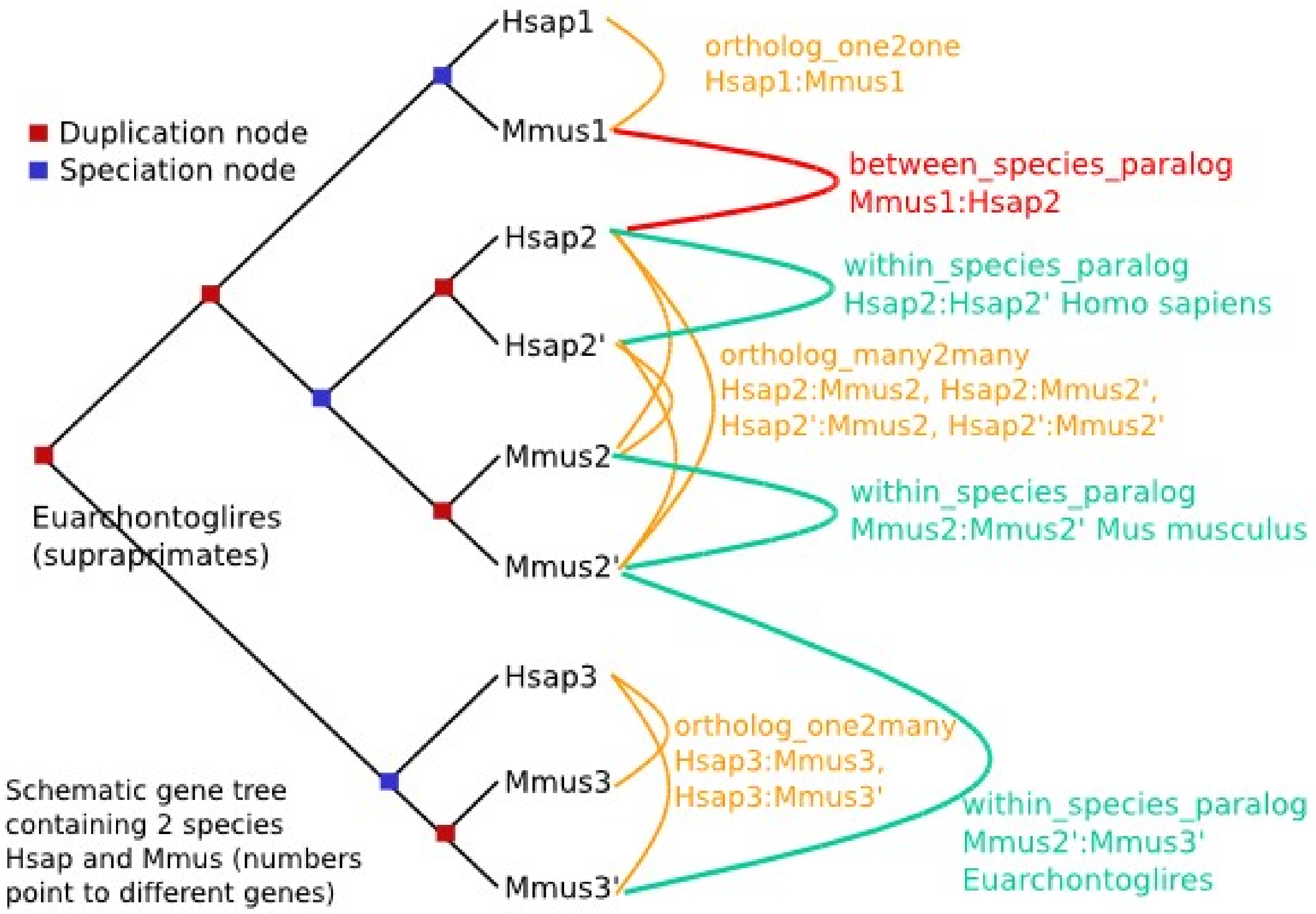
"Where the homology is ***the result of gene duplication*** *so that both copies have descended side by side during the history of an organism, (for example, alpha and beta hemoglobin) the genes should be called* ***paralogous*** *(para = in parallel).*

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)


Hsap1
Mmus1
Hsap2
Hsap2'
Mmus2
Mmus2'
Hsap3
Mmus3
Mmus3'
🟥 Duplication node
🟦 Speciation node
Euarchontoglires
(supraprimates)
Schematic gene tree
containing 2 species
Hsap and Mmus (numbers
point to different genes)
ortholog_one2one
Hsap1:Mmus1
between_species_paralog
Mmus1:Hsap2
within_species_paralog
Hsap2:Hsap2' Homo sapiens
ortholog_many2many
Hsap2:Mmus2, Hsap2:Mmus2',
Hsap2':Mmus2, Hsap2':Mmus2'
within_species_paralog
Mmus2:Mmus2' Mus musculus
ortholog_one2many
Hsap3:Mmus3,
Hsap3:Mmus3'
within_species_paralog
Mmus2':Mmus3'
Euarchontoglires

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)

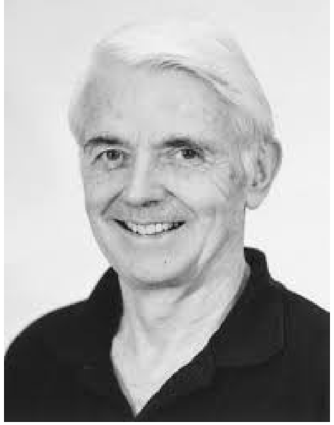
- The most exact way of **comparing two (or more) genomes** in terms of their gene content. Necessary to uncover how genomes evolve.

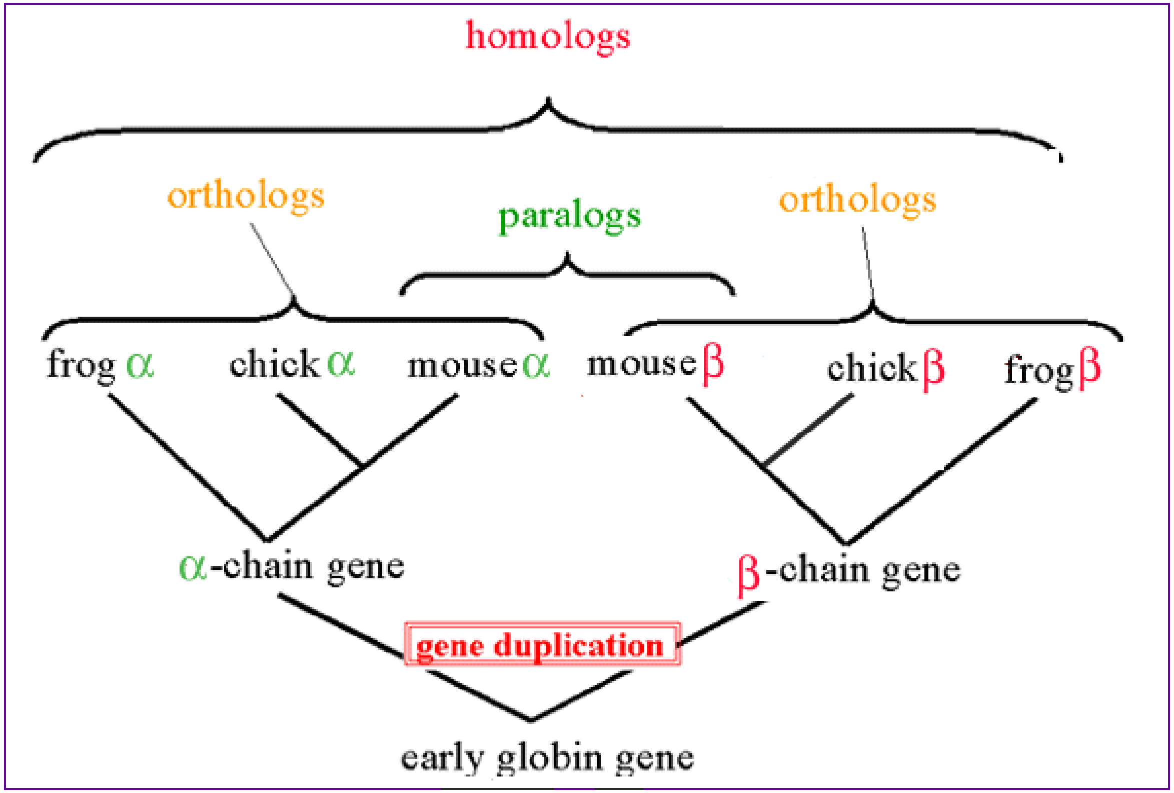
- Implications for **functional inference**: orthologs, as compared to paralogs, are more likely to share the same function

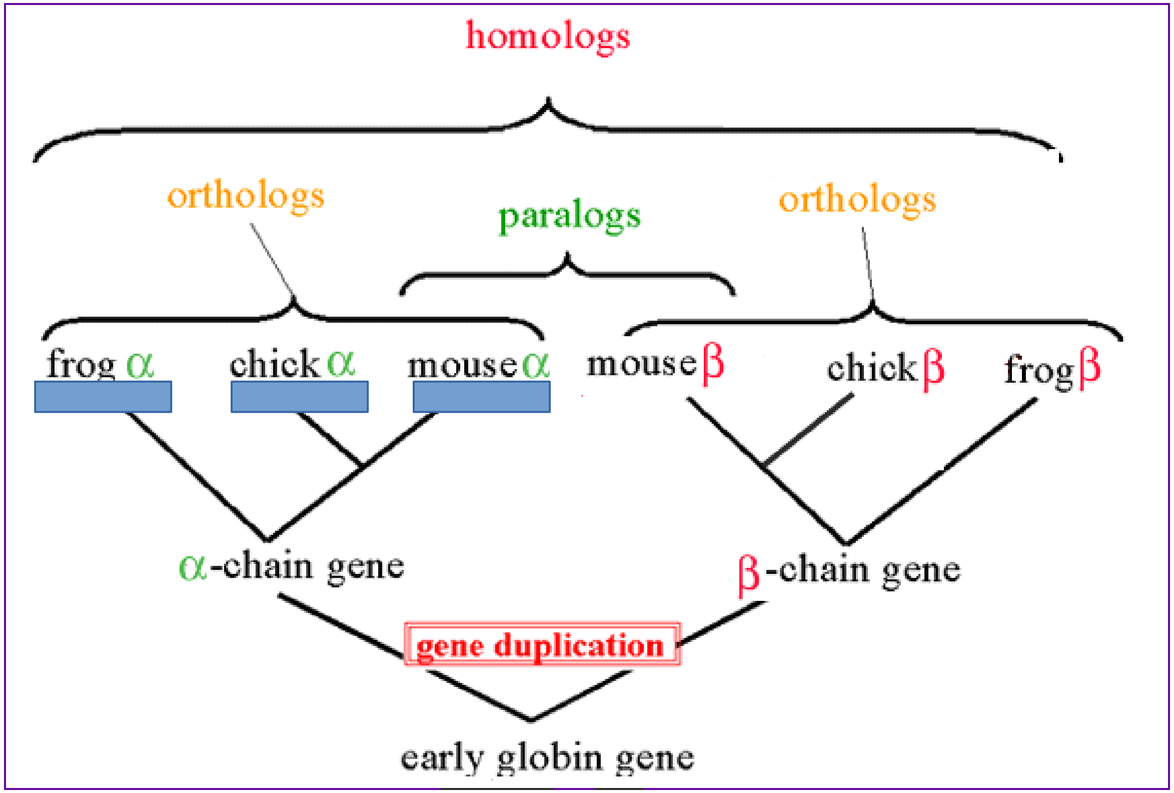
Why predicting orthology is important?

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

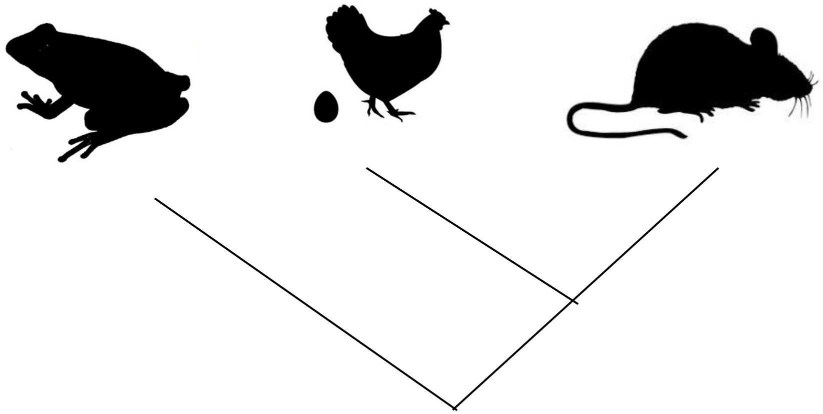
*Where the homology is **the result of speciation** so that the history of the gene reflects the history of the species (for example alpha hemoglobin in man and mouse) the genes should be called **orthologous** (ortho = exact).*

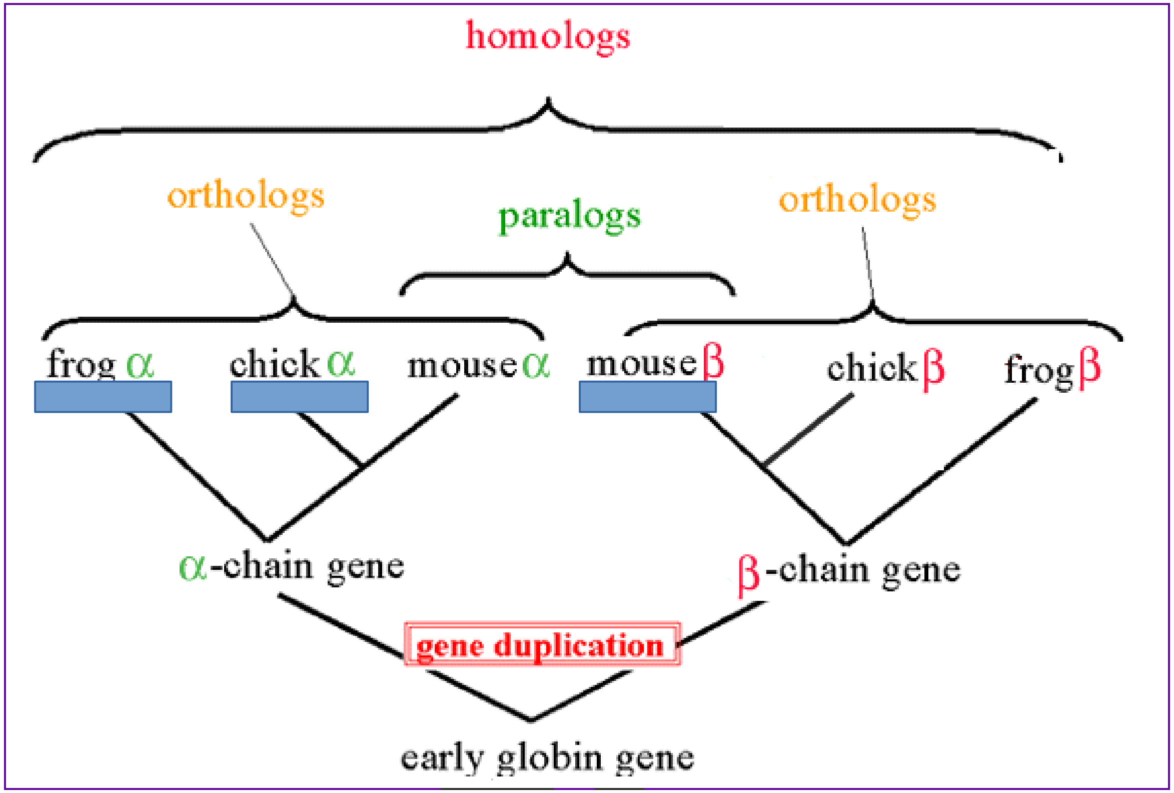



homologs
orthologs
paralogs
orthologs
frog α
chick α
mouse α
mouse β
chick β
frog β
α-chain gene
β-chain gene
gene duplication
early globin gene

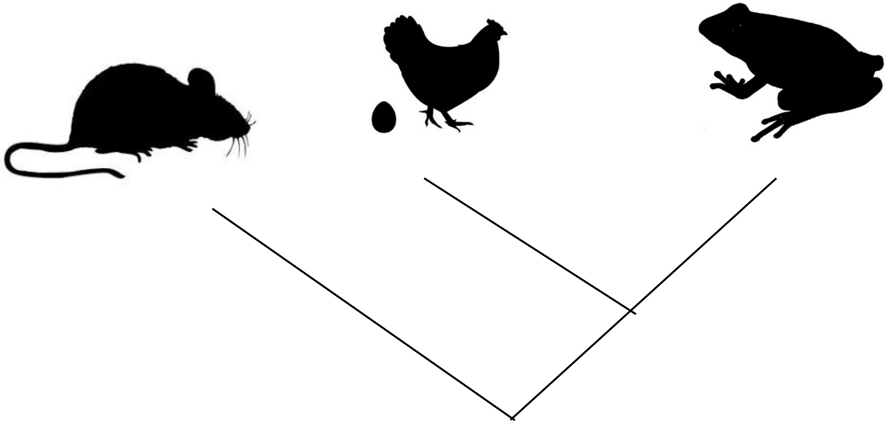


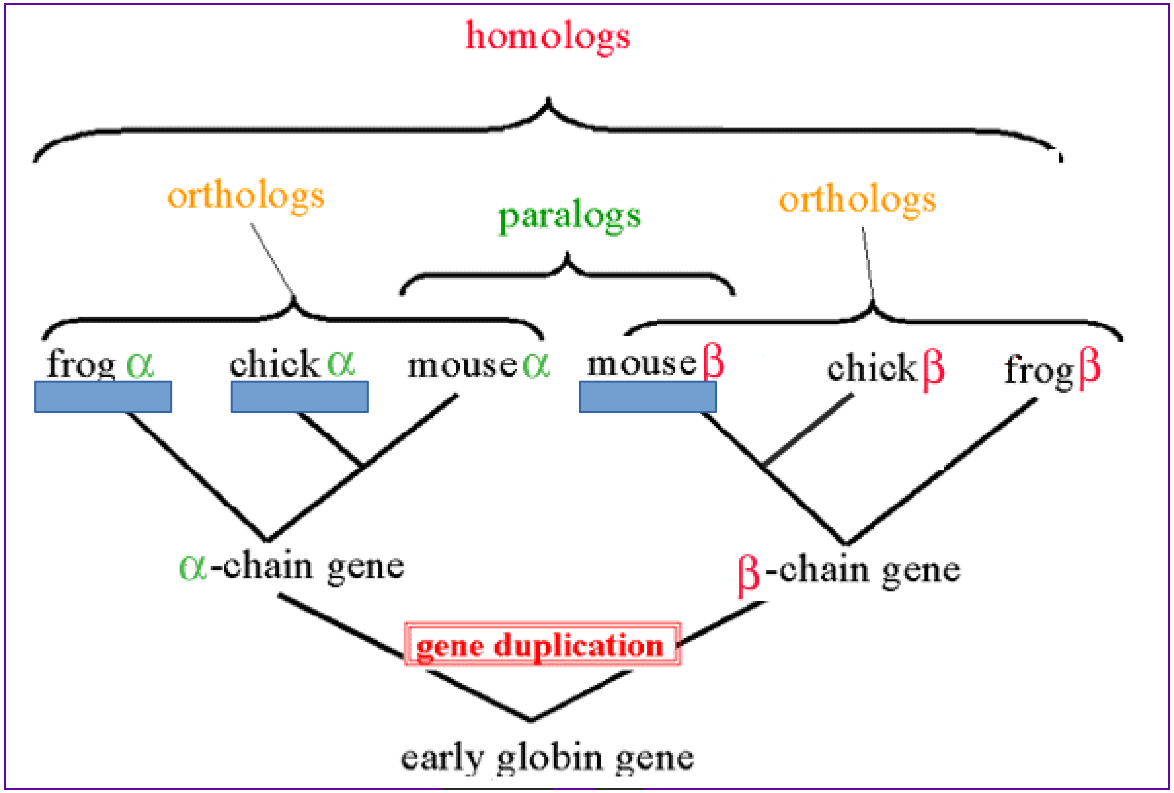
Tree from orthologous dataset:



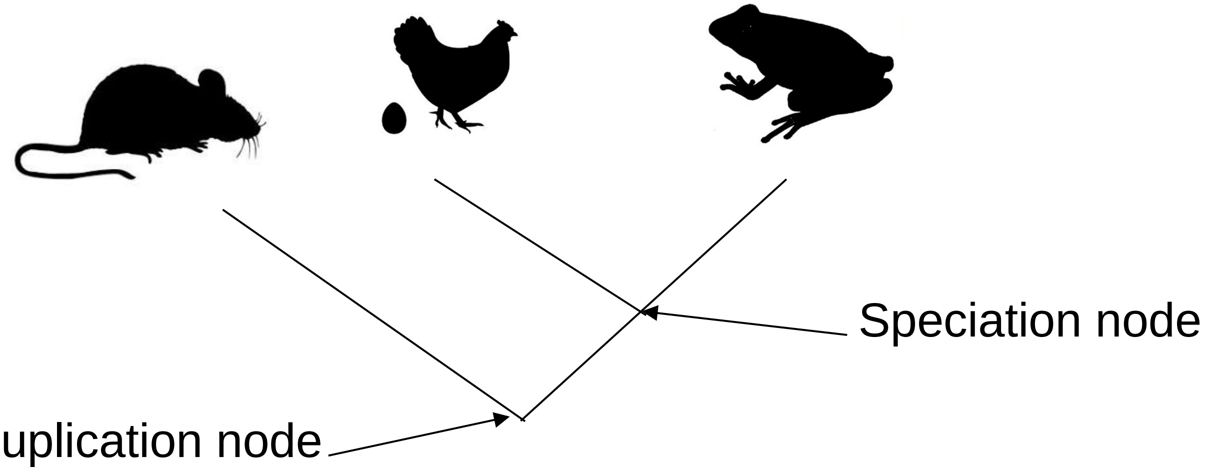


Tree from non-orthologous dataset:





Tree from non-orthologous dataset: NOT a SPECIES TREE



Speciation node

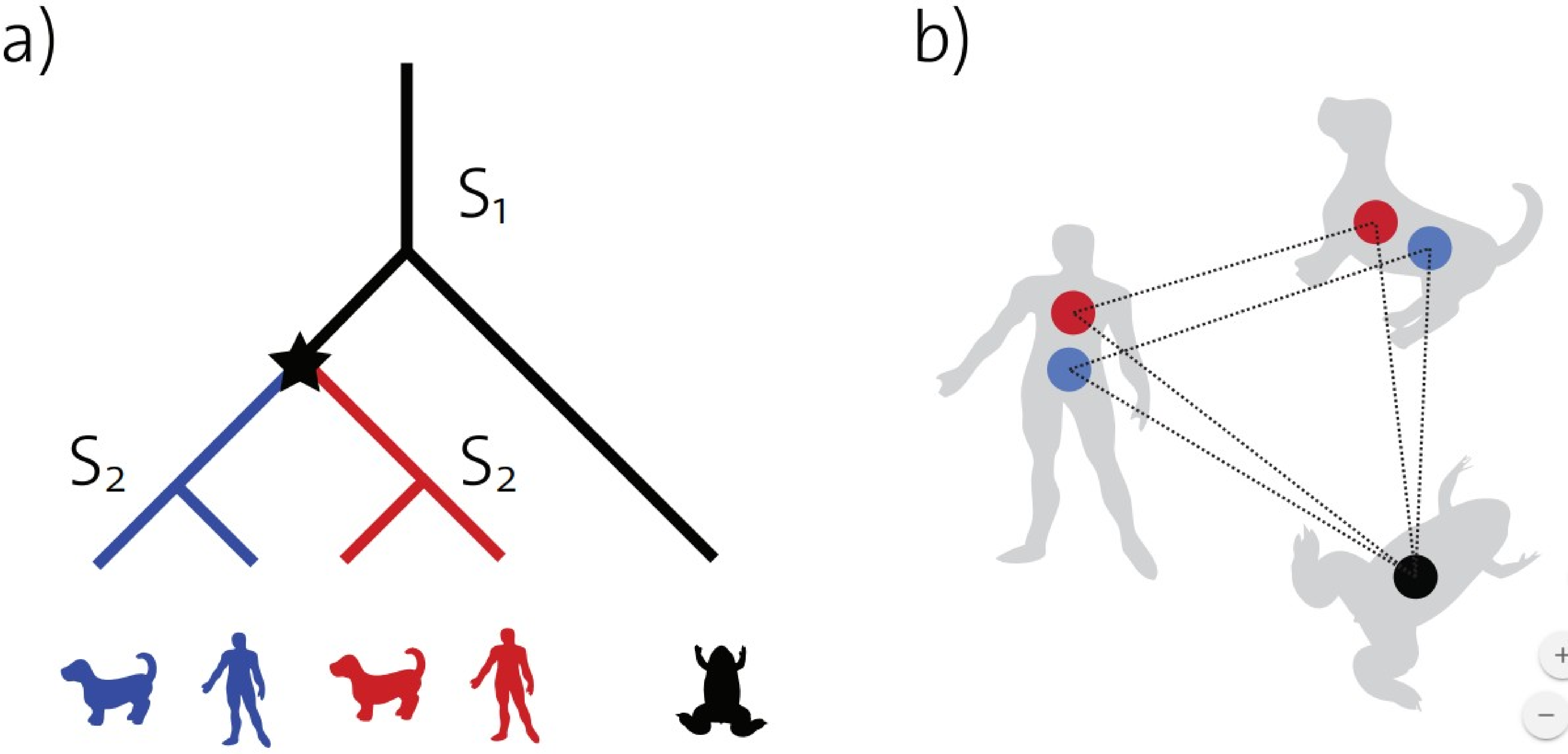
Duplication node

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


a)
S_1
S_2
S_2
b)
+
−

Why predicting orthology is important?

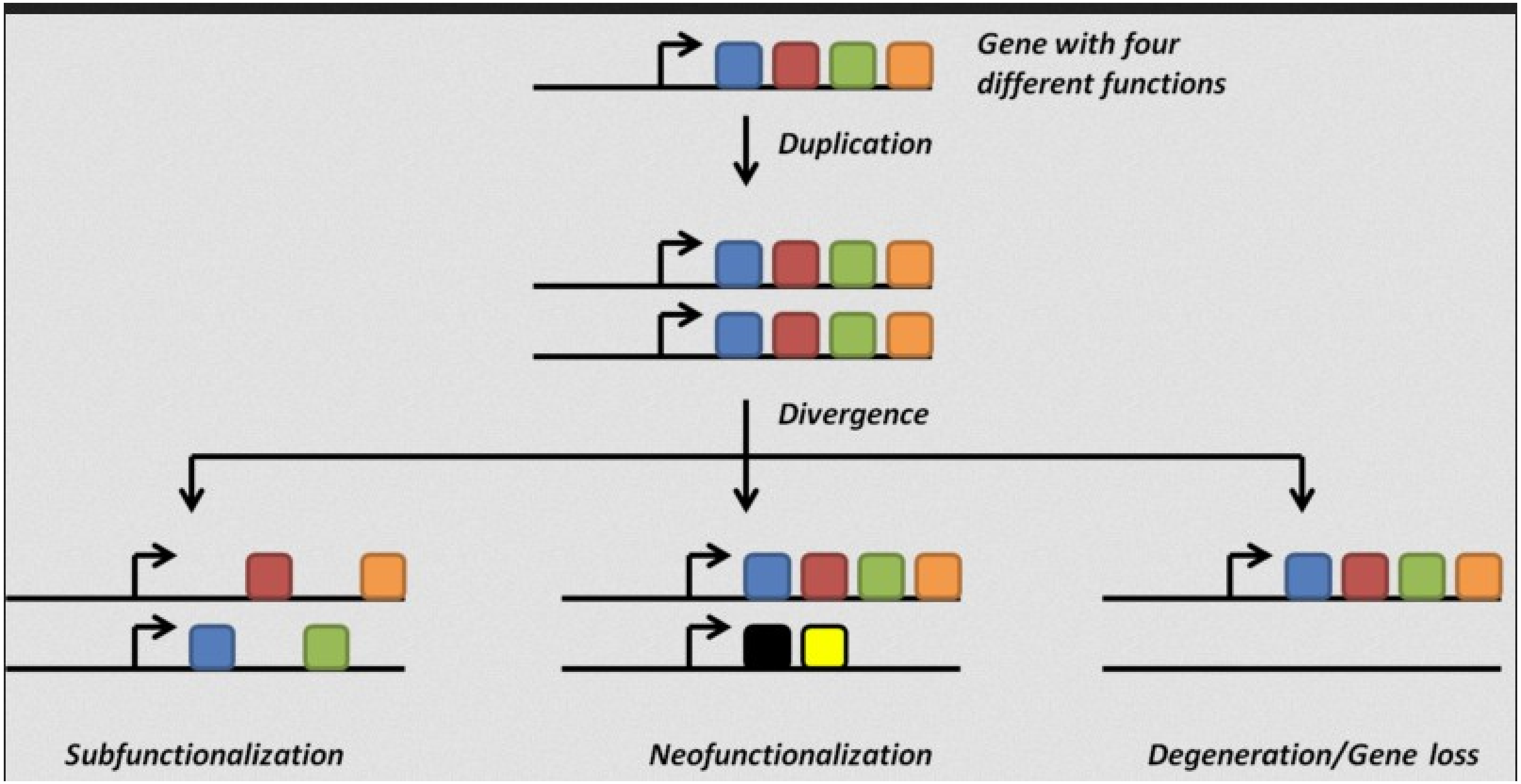
- **Important implications for phylogeny:** only sets of orthologous genes are expected to reflect the underlying species evolution (although there are many exceptions)

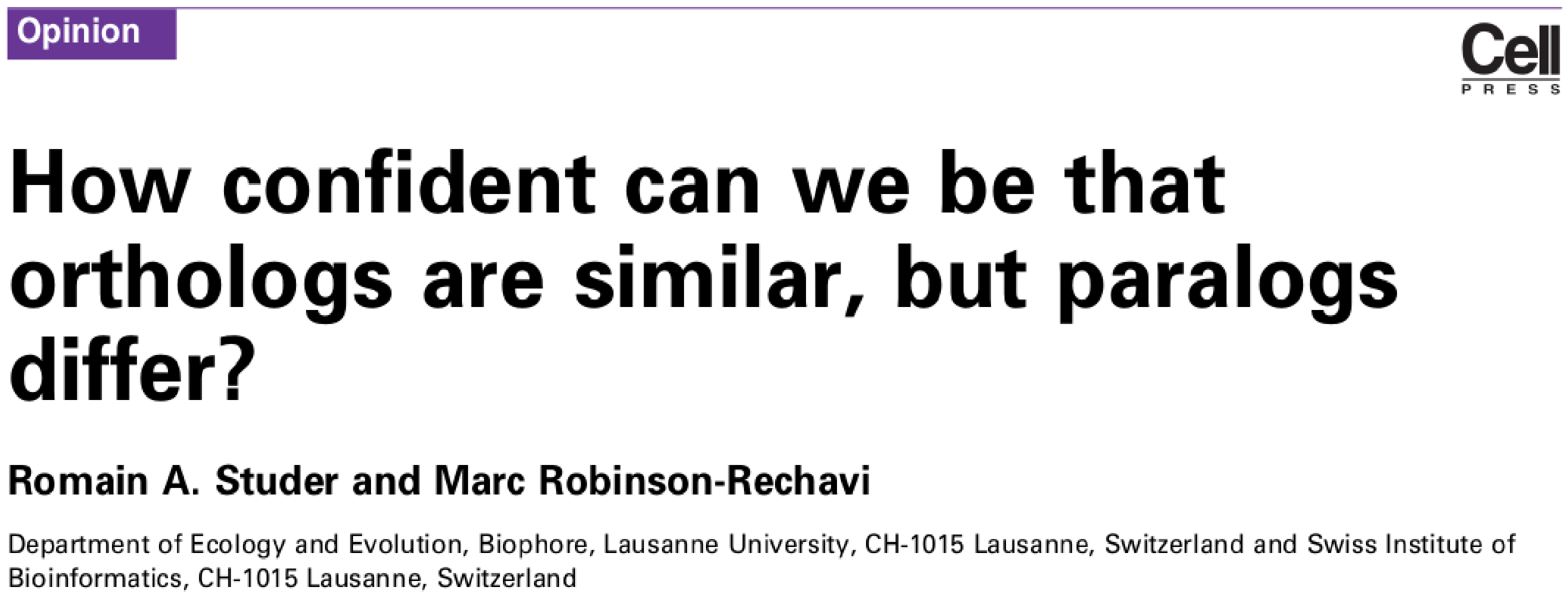
- The most exact way of **comparing two (or more) genomes** in terms of their gene content. Necessary to uncover how genomes evolve.

- 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?

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





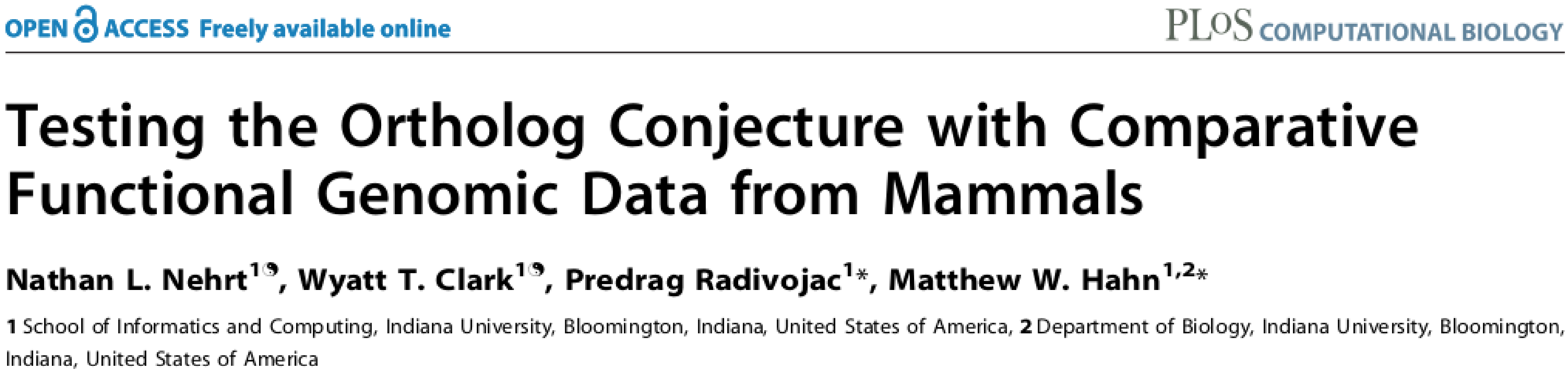
Opinion

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

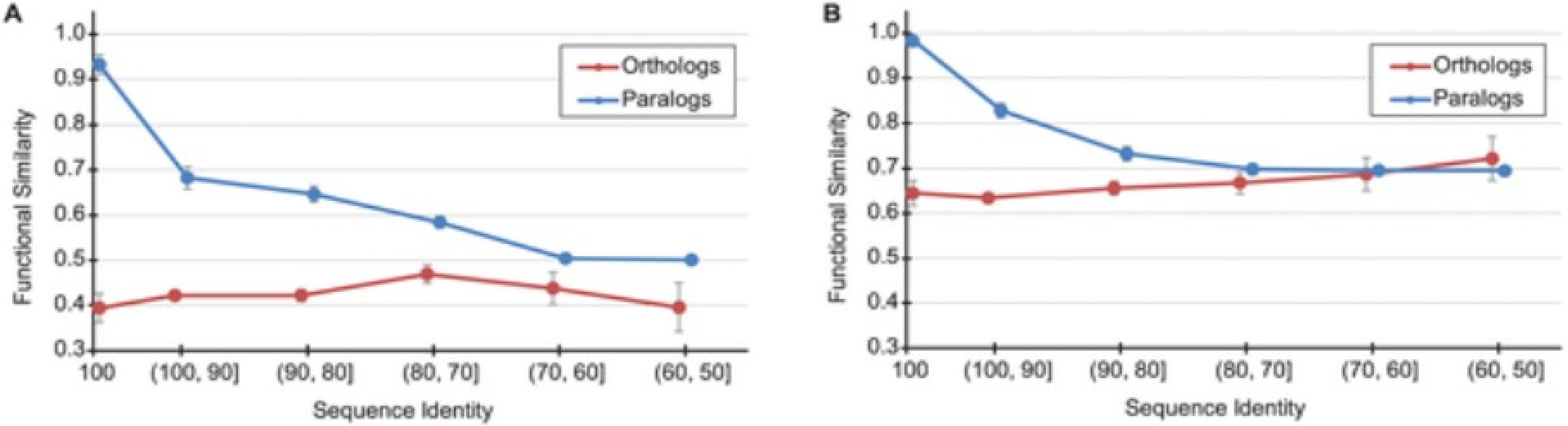
PLoS COMPUTATIONAL BIOLOGY

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]*

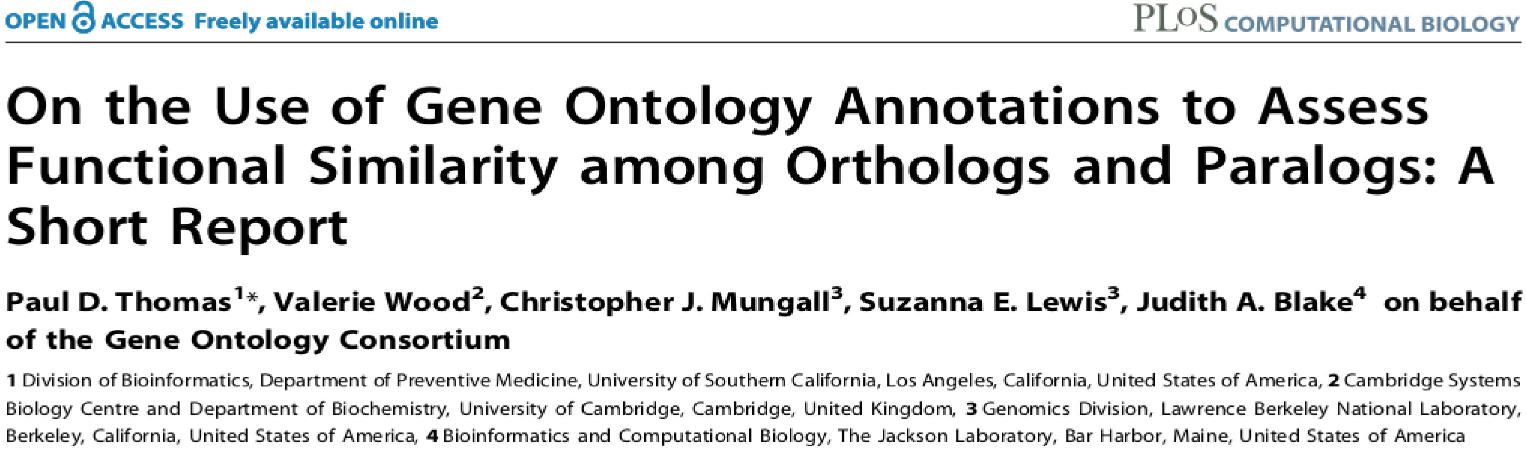
[1] School of Informatics and Computing, Indiana University, Bloomington, Indiana, United States of America, [2] Department of Biology, Indiana University, Bloomington, Indiana, United States of America

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

PLOS | COMPUTATIONAL BIOLOGY

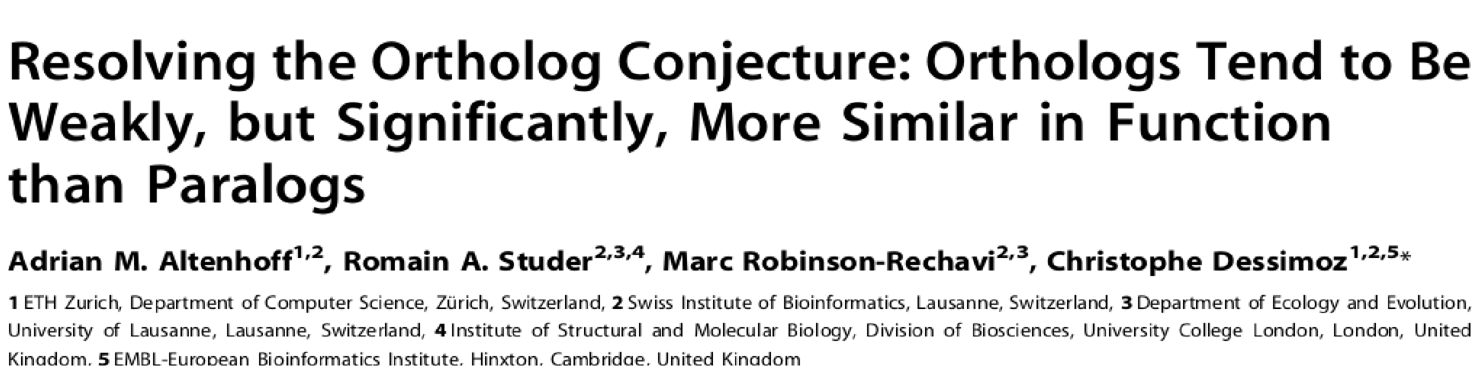


OPEN ACCESS Freely available online PLOS COMPUTATIONAL BIOLOGY

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

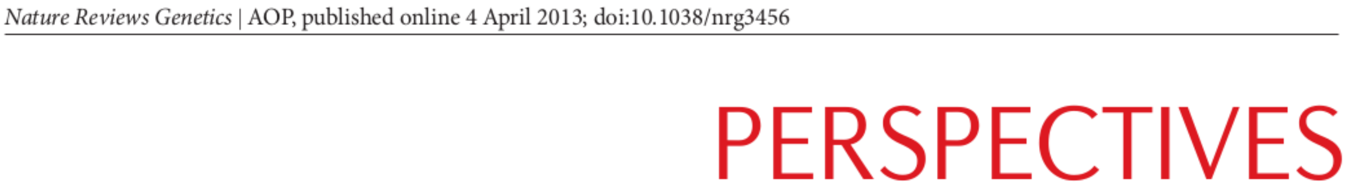


OPEN ACCESS Freely available onlin

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

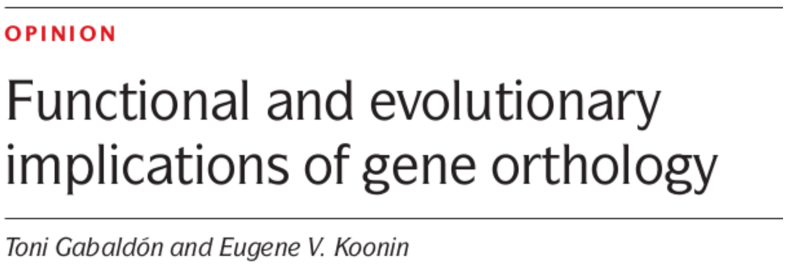
Adrian M. Altenhoff[1,2], Romain A. Studer[2,3,4], Marc Robinson-Rechavi[2,3], Christophe Dessimoz[1,2,5]*

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

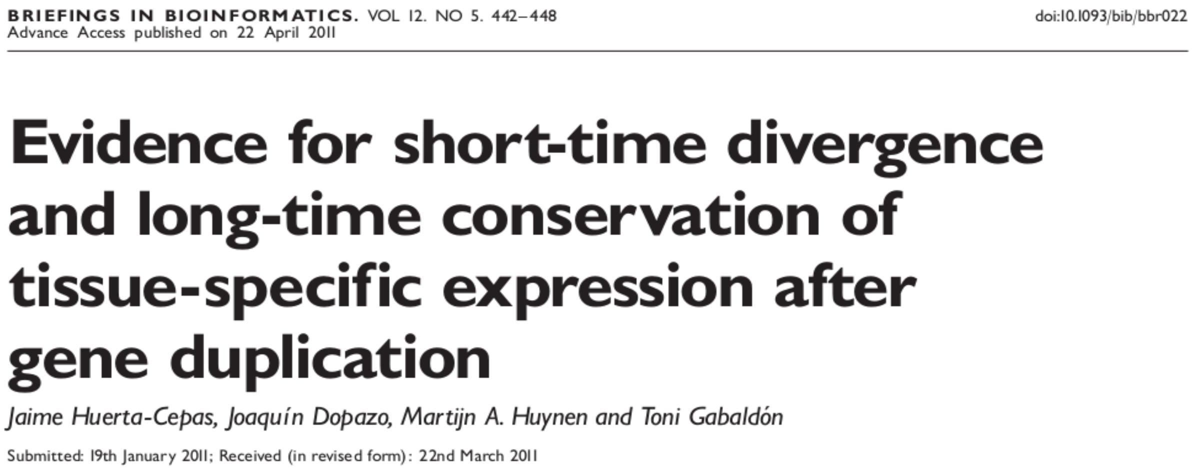
PERSPECTIVES



OPINION

Functional and evolutionary implications of gene orthology

Toni Gabaldón and Eugene V. Koonin



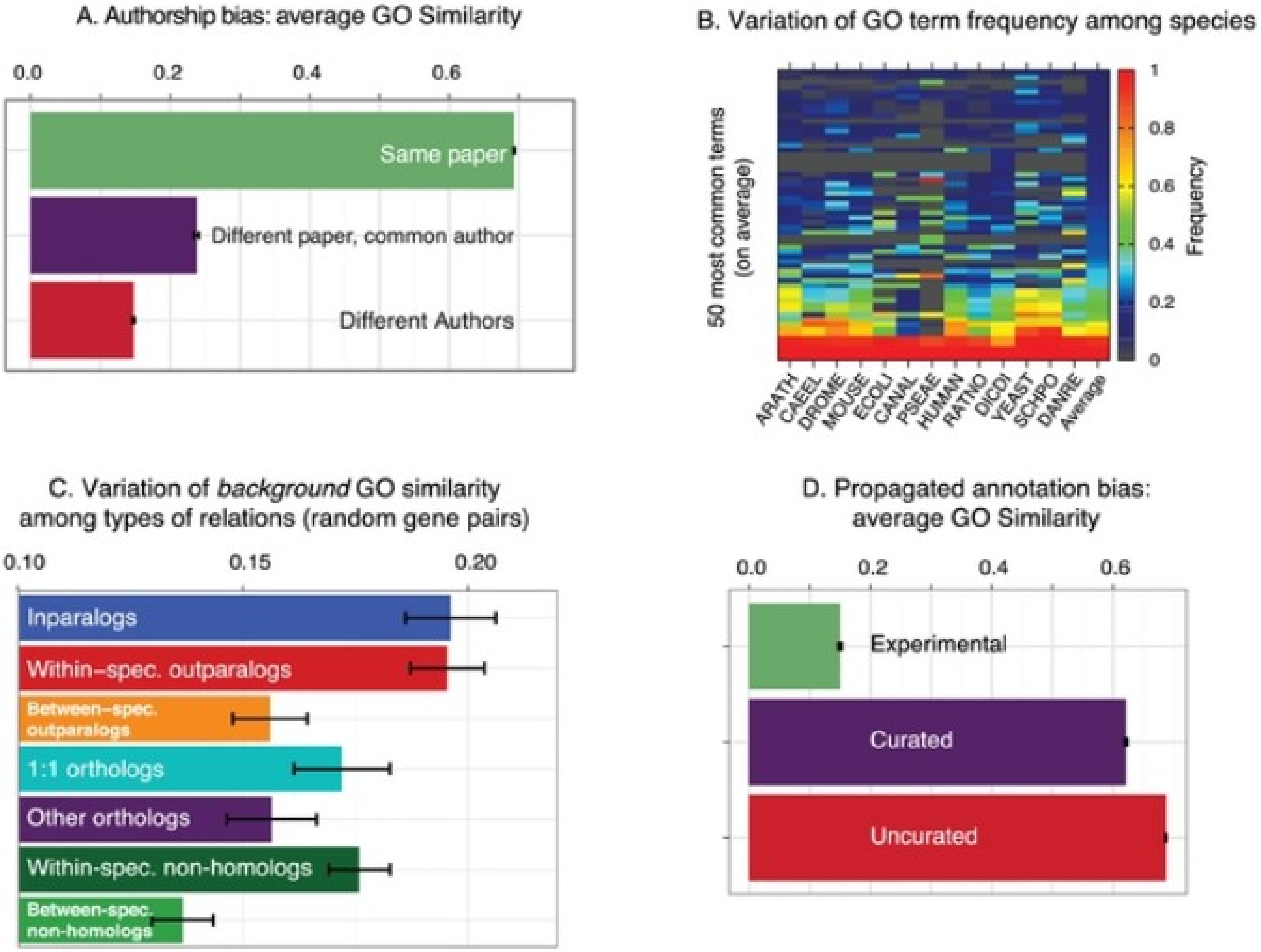
BRIEFINGS IN BIOINFORMATICS. VOL 12. NO 5. 442–448 doi:10.1093/bib/bbr022
Advance Access published on 22 April 2011

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

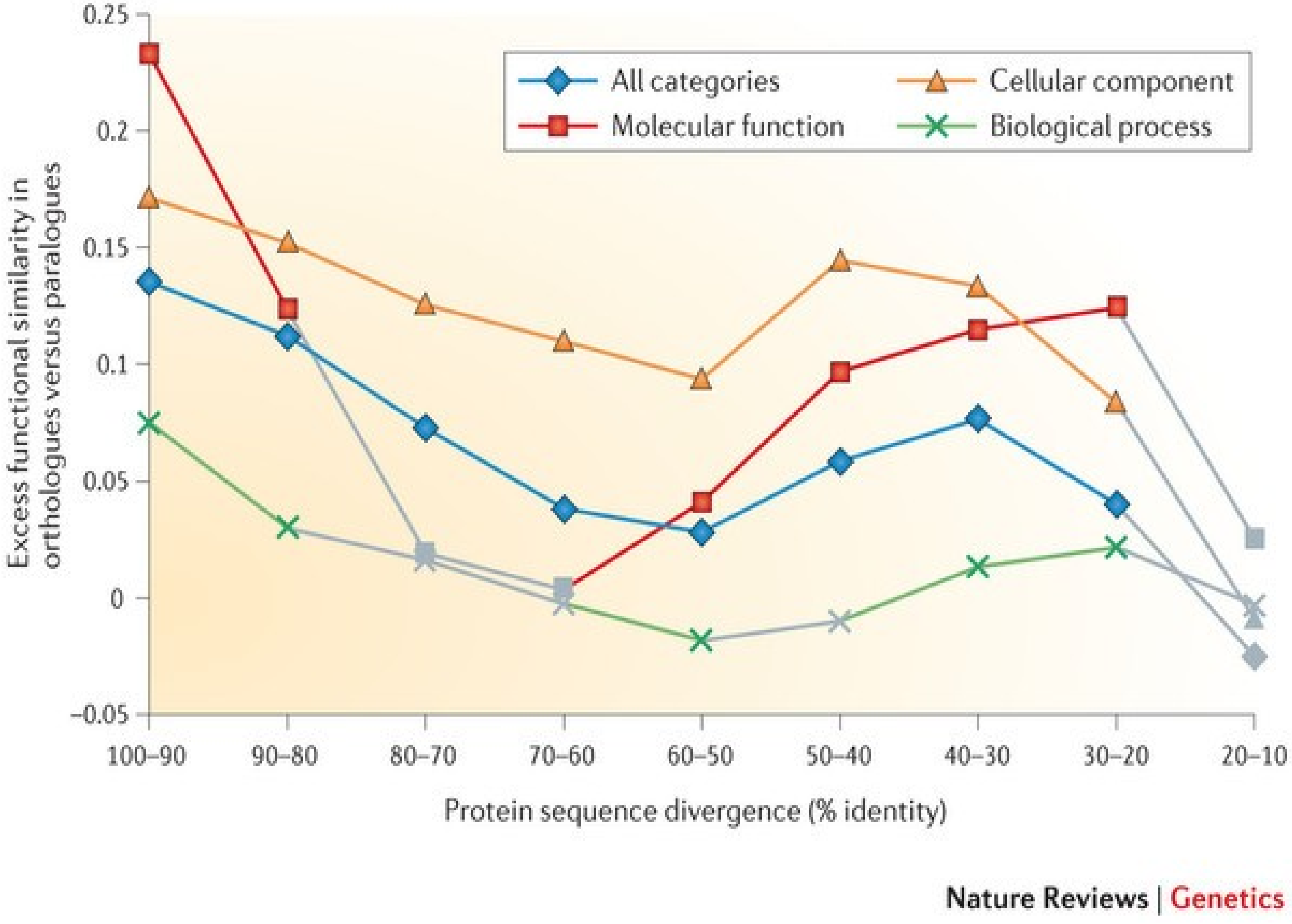
Submitted: 19th January 2011; Received (in revised form): 22nd March 2011

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

PLOS | COMPUTATIONAL BIOLOGY



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

Gene families

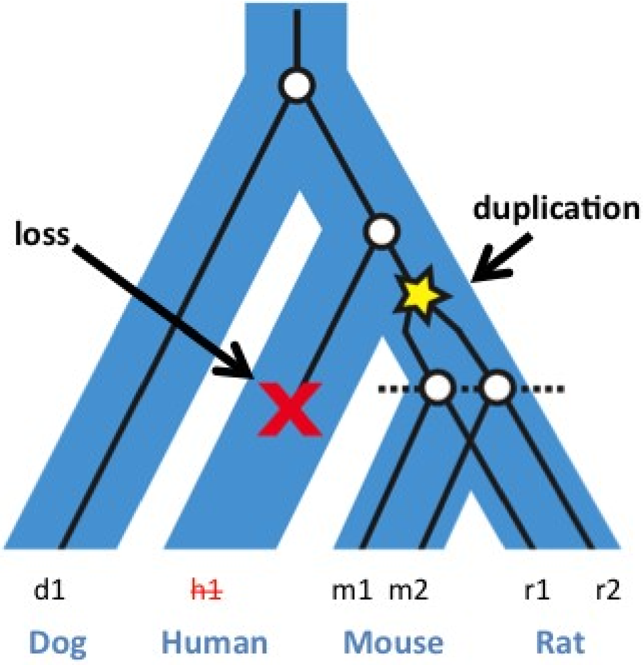
A group of genes that share a common ancestry (they are homologs)

Gene families have hierarchical evolutionary relationship (best represented by a tree)

Members of a gene family can be orthologs or paralogs between them

An orthologous group is a (or part of) a gene family

Gene families evolve by duplication and loss (birth and death)

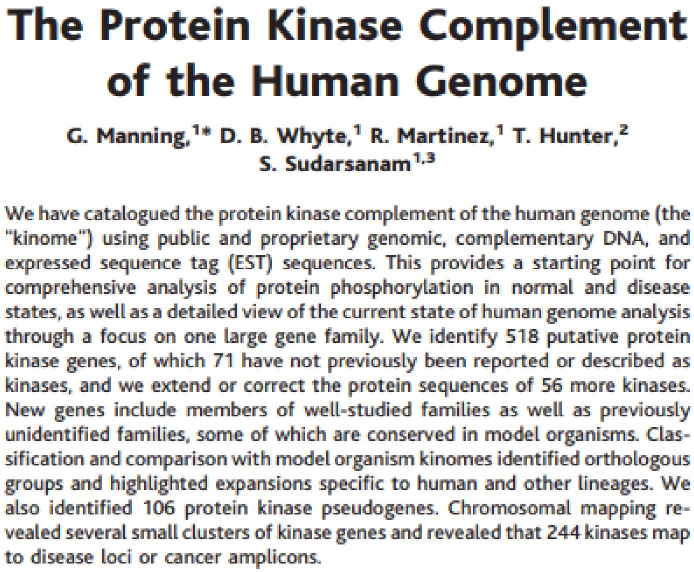


Gene families

Because of loss/duplictain dynamics gene families will vary in size and phylogenetic distribution.

Single copy families
Multigene families

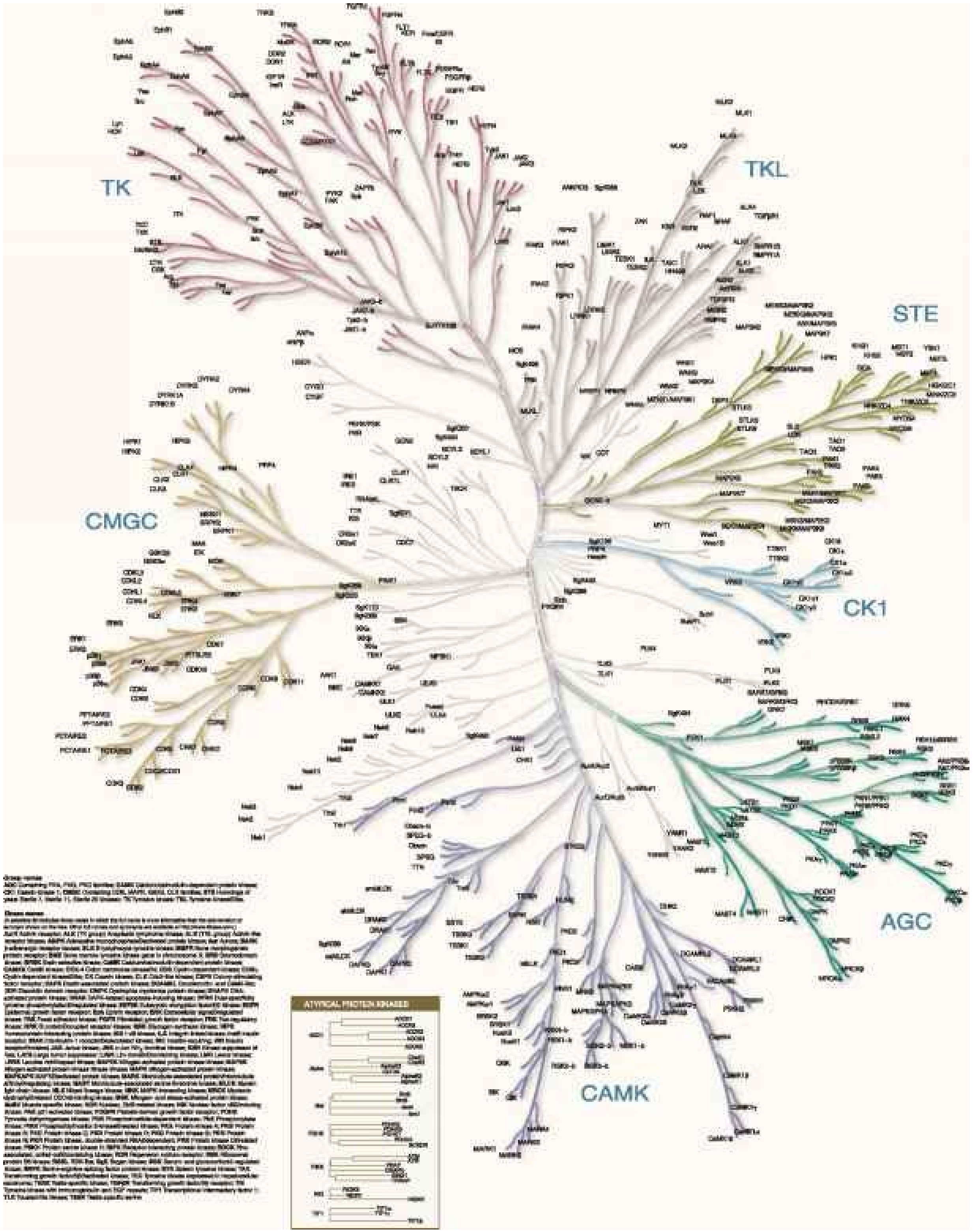
E.g. more that 518 protein kinases only in human



The Protein Kinase Complement of the Human Genome

G. Manning,[1]* D. B. Whyte,[1] R. Martinez,[1] T. Hunter,[2] S. Sudarsanam[1,3]

We have catalogued the protein kinase complement of the human genome (the "kinome") using public and proprietary genomic, complementary DNA, and expressed sequence tag (EST) sequences. This provides a starting point for comprehensive analysis of protein phosphorylation in normal and disease states, as well as a detailed view of the current state of human genome analysis through a focus on one large gene family. We identify 518 putative protein kinase genes, of which 71 have not previously been reported or described as kinases, and we extend or correct the protein sequences of 56 more kinases. New genes include members of well-studied families as well as previously unidentified families, some of which are conserved in model organisms. Classification and comparison with model organism kinomes identified orthologous groups and highlighted expansions specific to human and other lineages. We also identified 106 protein kinase pseudogenes. Chromosomal mapping revealed several small clusters of kinase genes and revealed that 244 kinases map to disease loci or cancer amplicons.

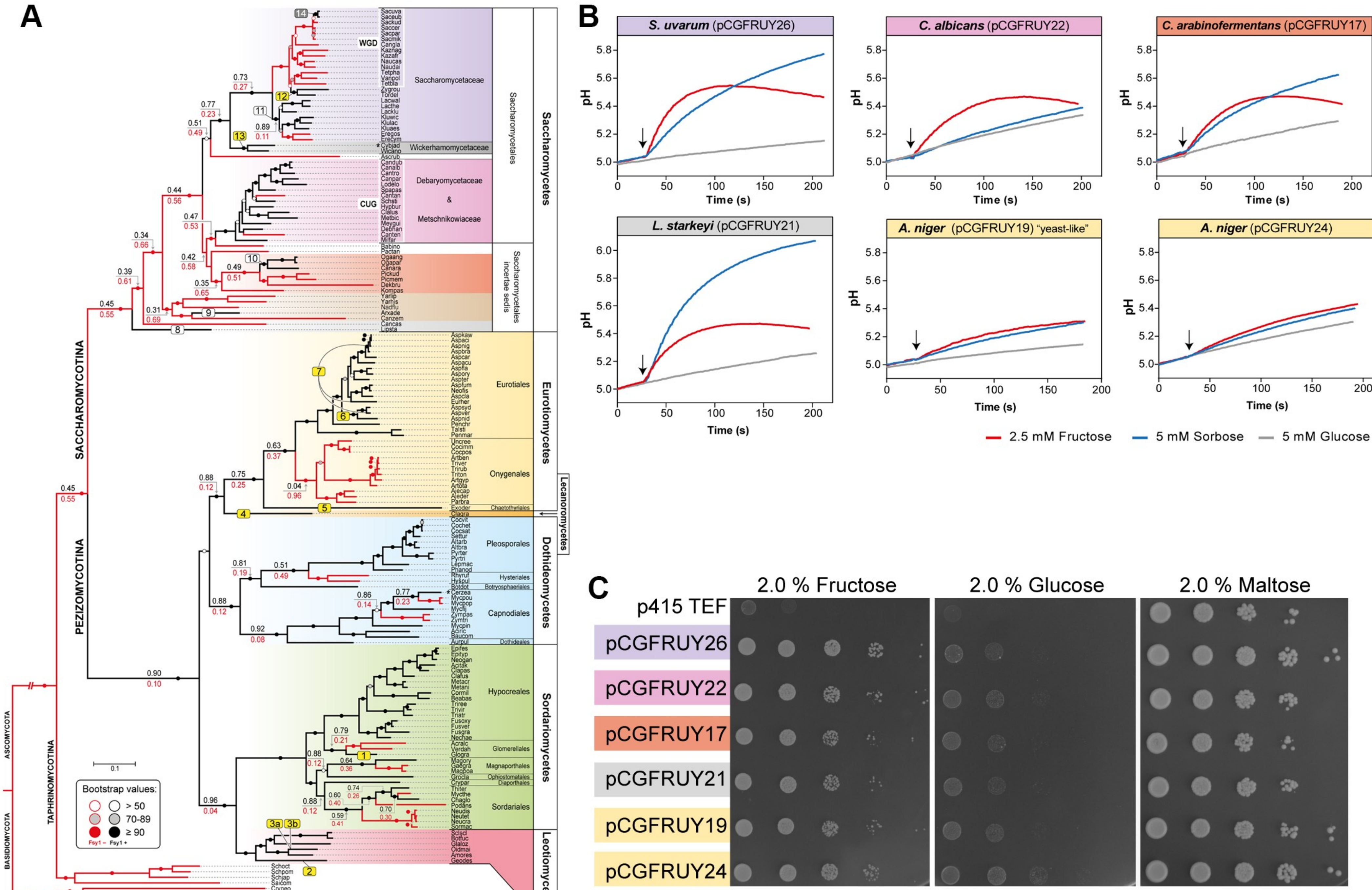


Human kinome

Proteins within the same gene family tend to have **related f**unctions
(remember homology based function prediction)

But functions can evolve through time.

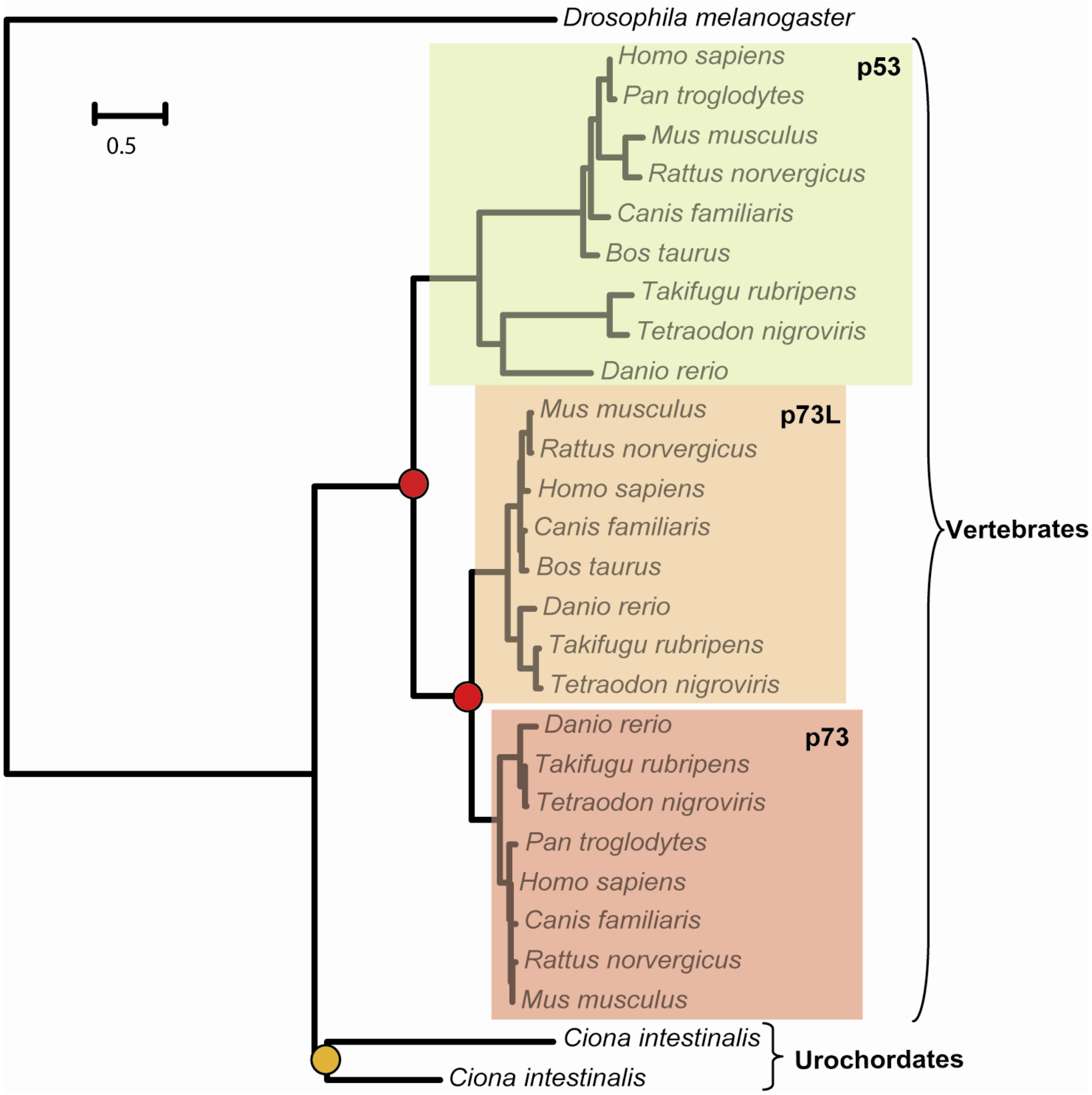
Functional evolution through species diversification



Orthology prediction methods

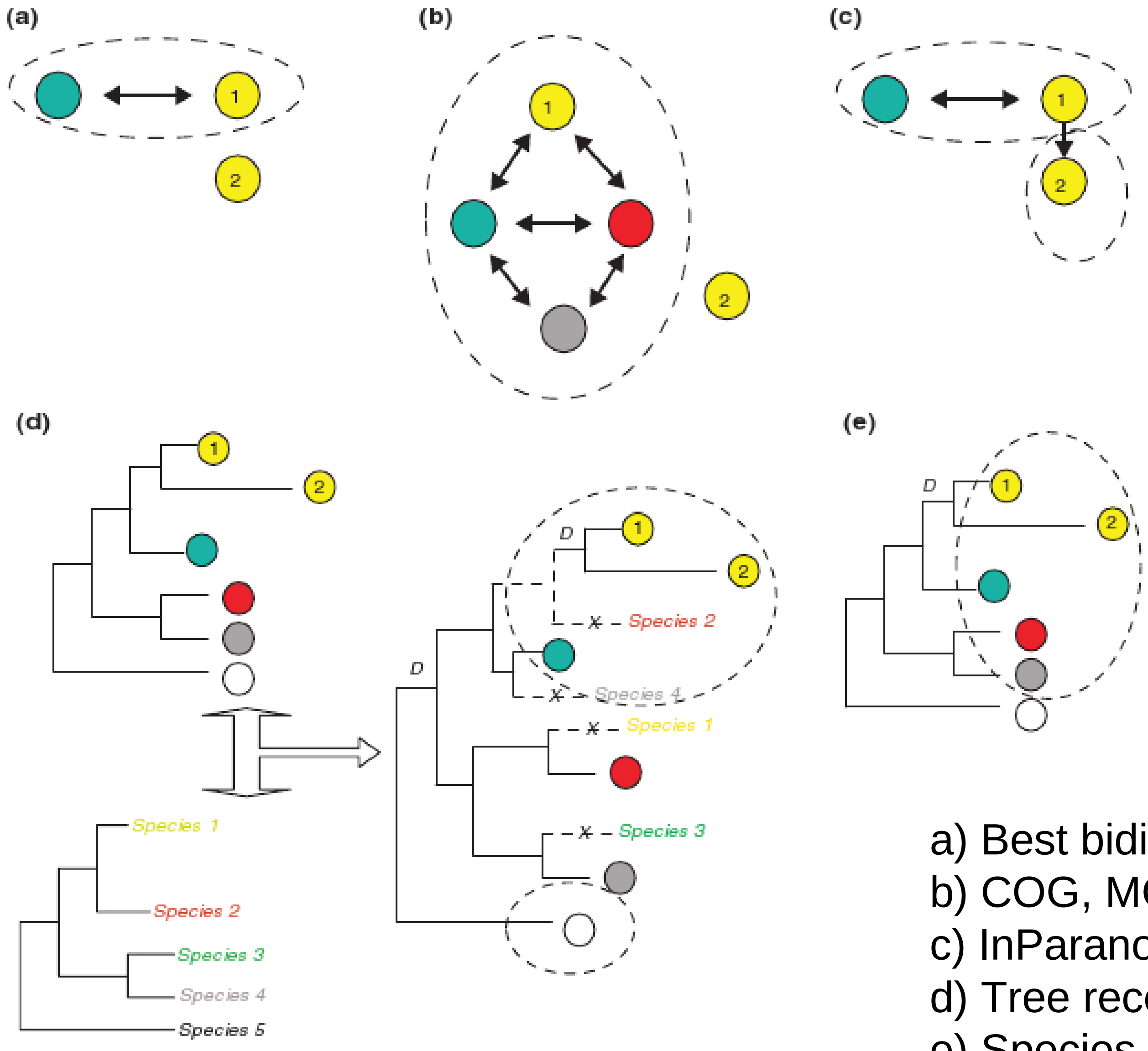
Classical approach: phylogenetic inference

- 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”



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

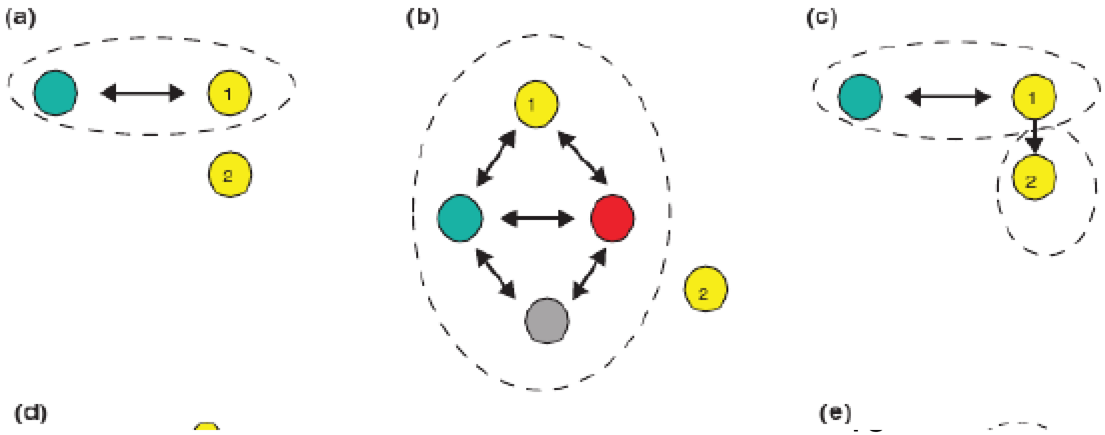
Gabaldón, T. *Genome Biology* (2008)

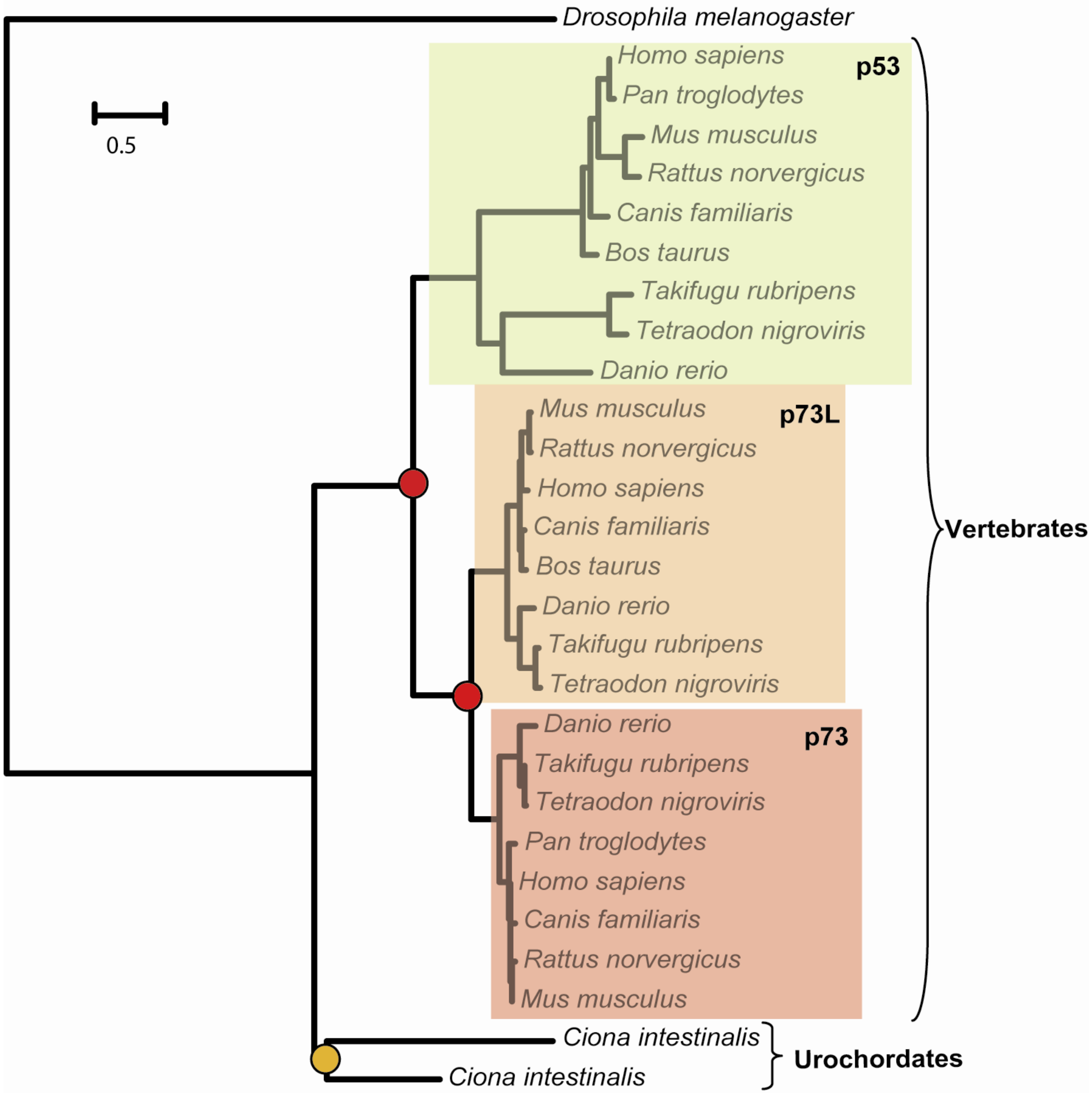
Similarity-based approaches (many more approaches):

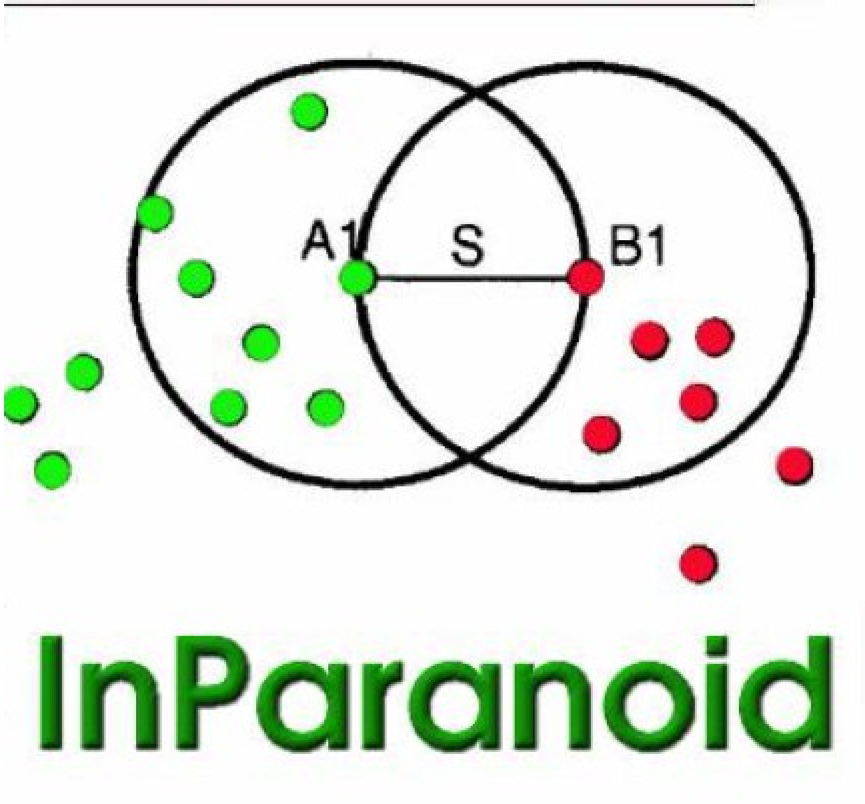
- **Best Reciprocal Hits**
- Detects all orthologies as one-to one. Highly affected by paralogy. Low rate of false positives but high rates of false negatives.

- The simplest and fastest method, still widely used

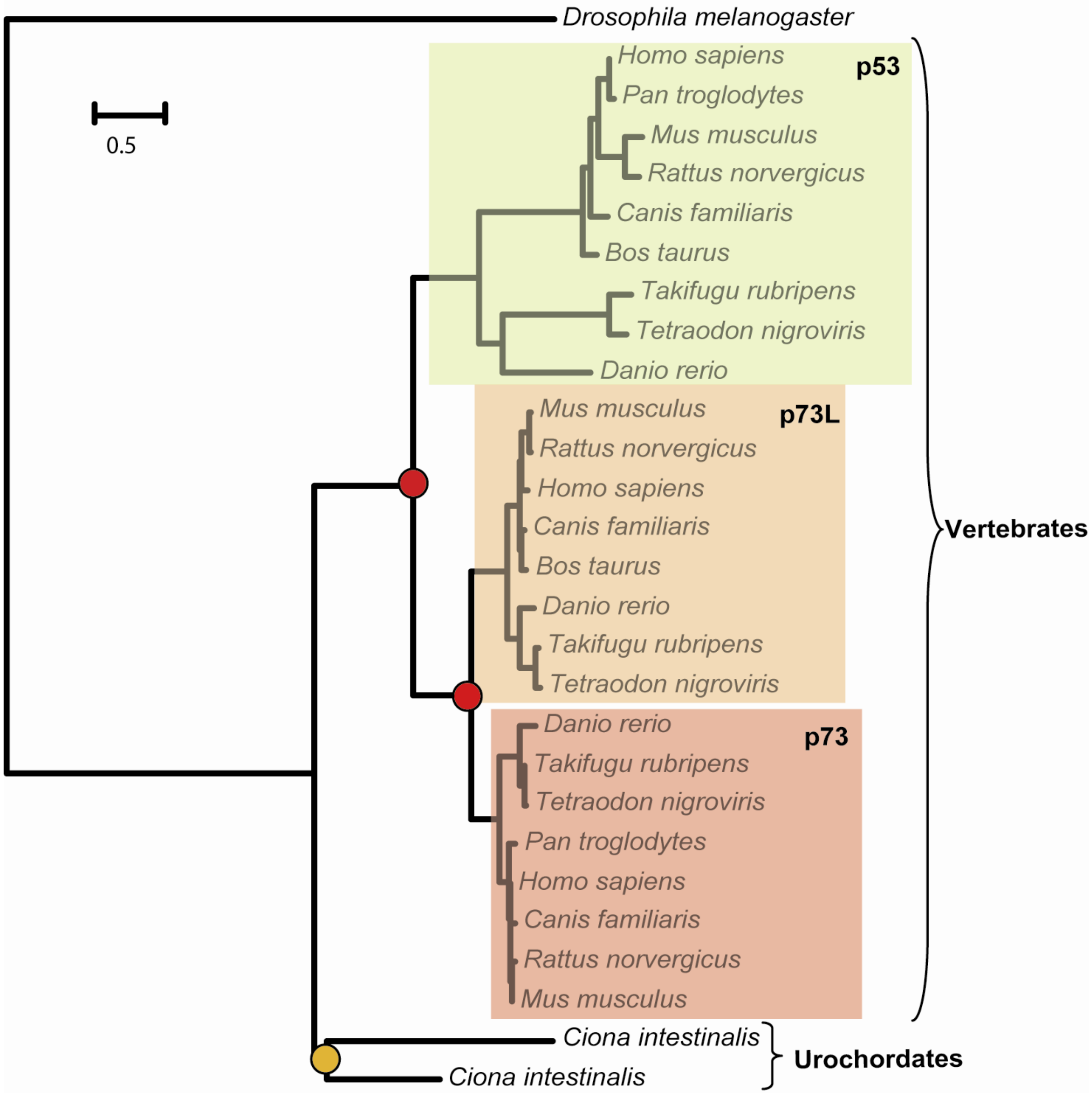
-




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
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:

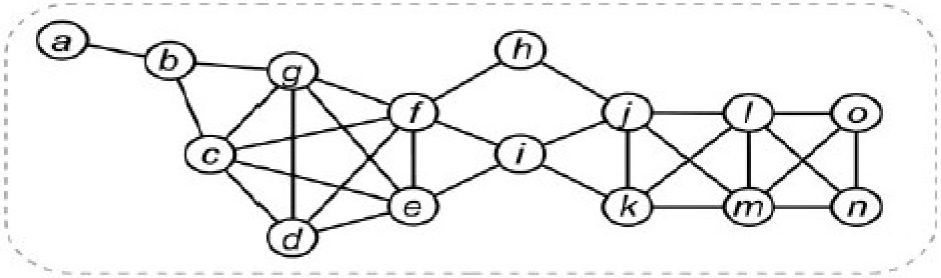
Definition of **in-** and **out-paralogues** require the specification of a given **speciation-node** of reference

COG-like
(used by many DBs like STRING)

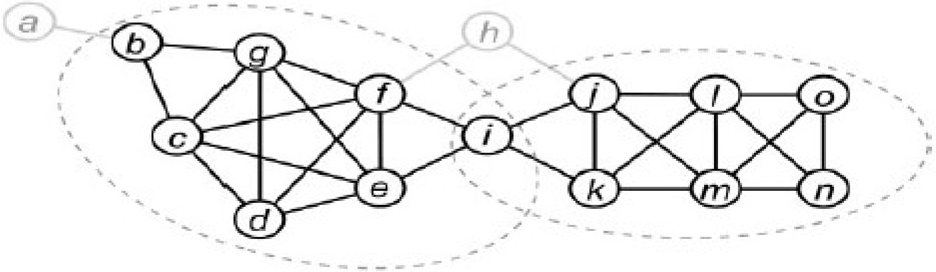
Exploits multi-species information. Predicts clusters of orthologous groups (in-paralogs) not all pairs in a cluster are paralogs.

Can be used at different stringent levels

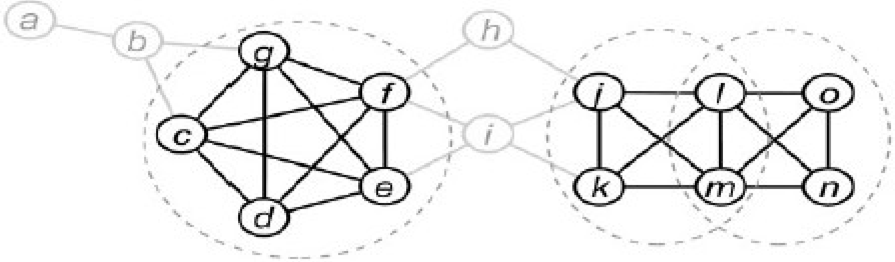
2



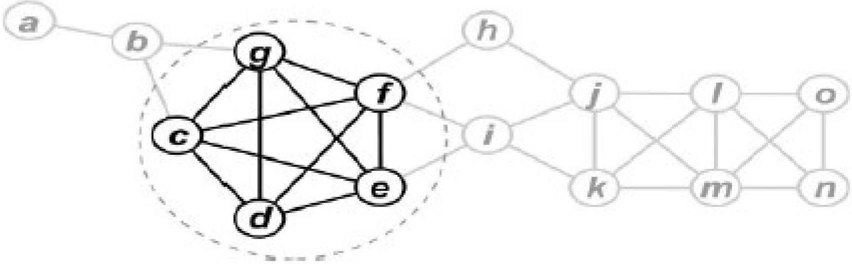
3



4



5



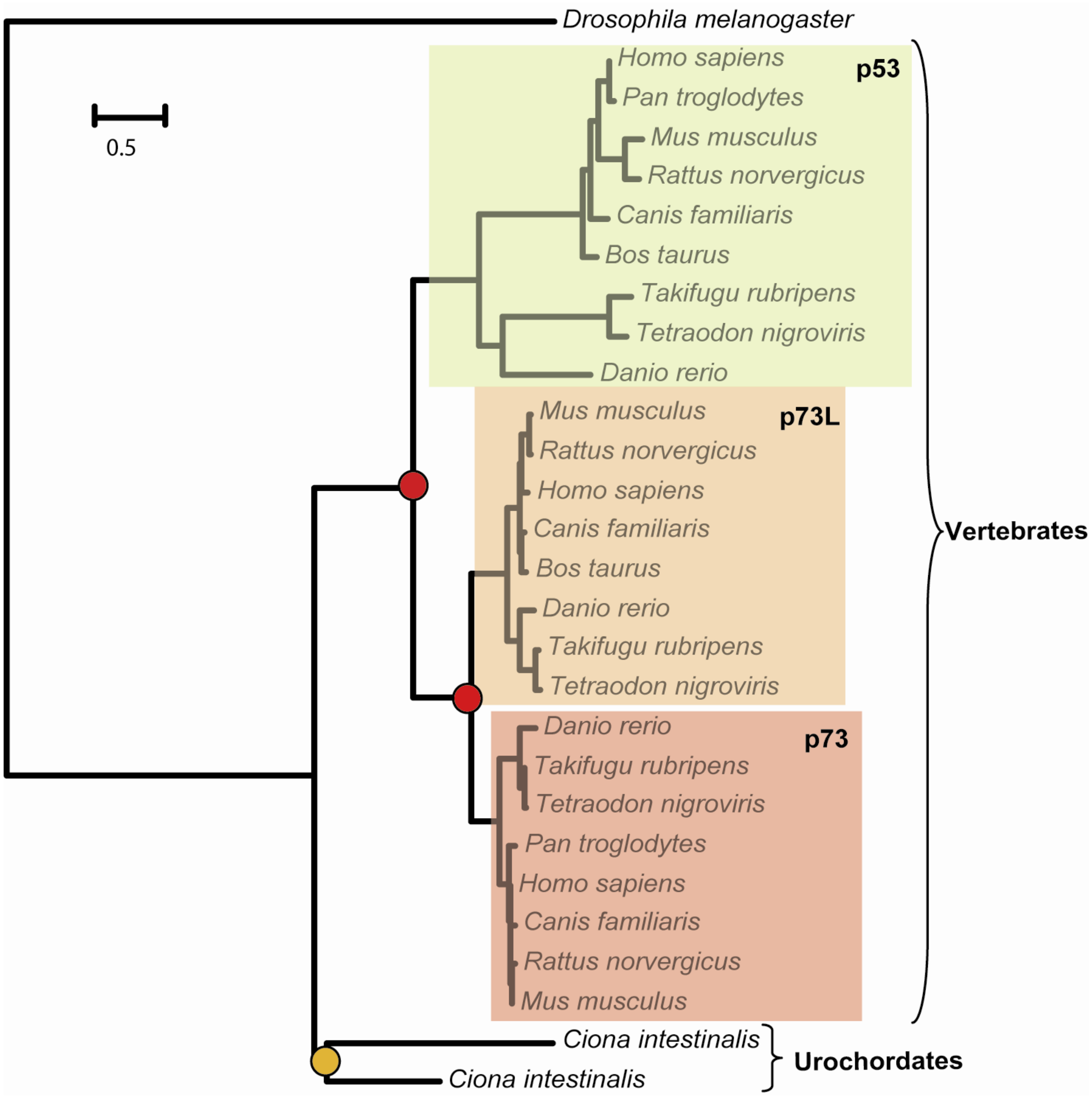
Clustering methods produce: orthologous groups

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



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)

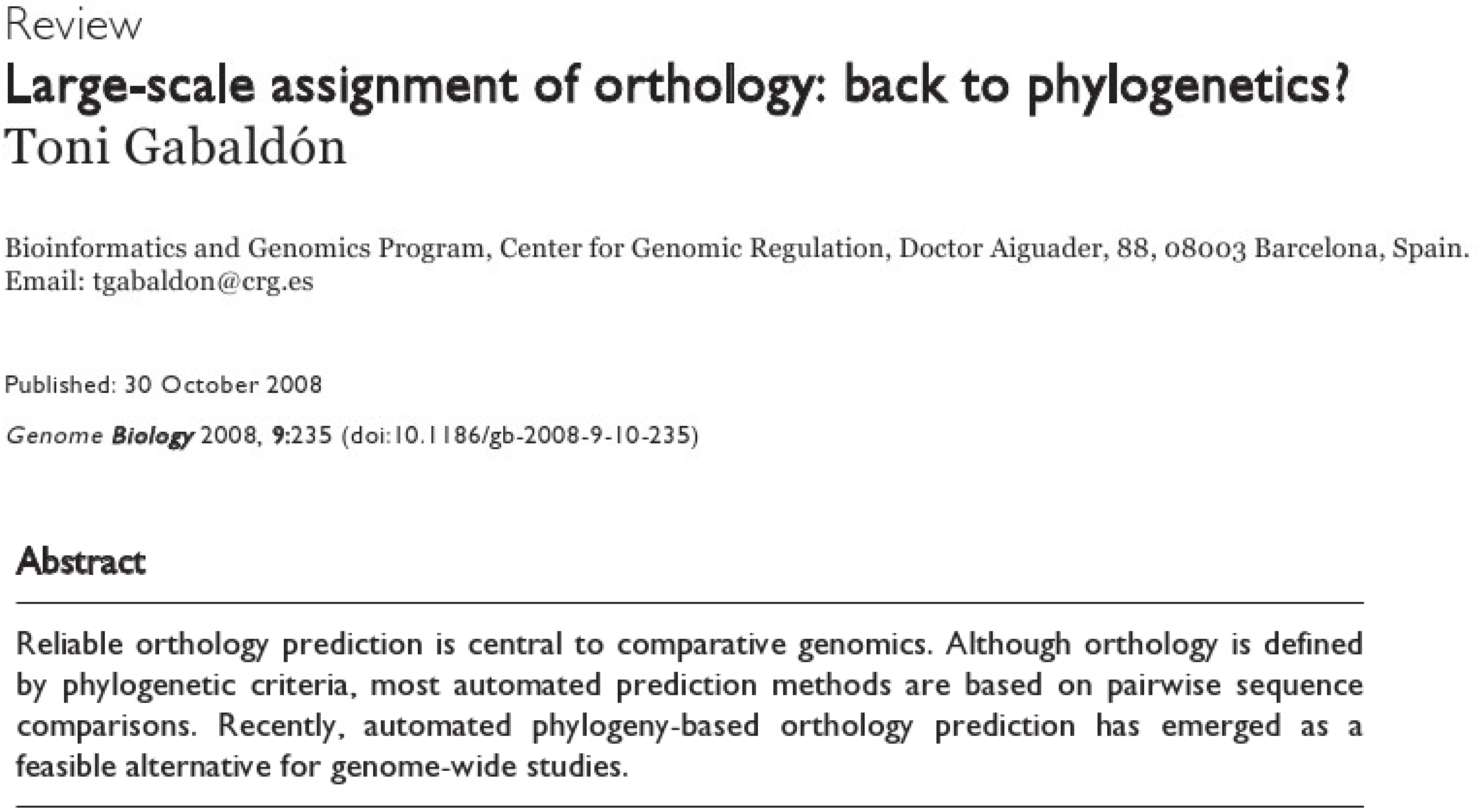
The definition of a reference ancestral species is just an approximation to the inherently hierarchical nature of gene family evolution: and is thus incomplete.

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.

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:

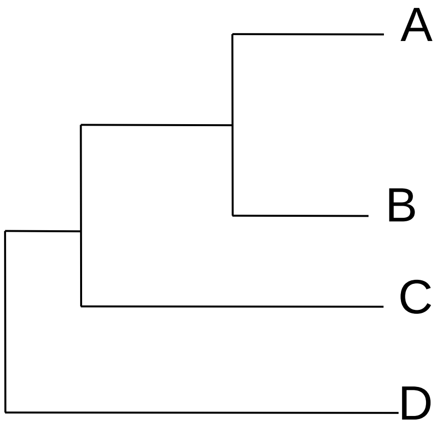
→ Species tree reconciliation (RIO, Ensembl)

→ Species-overlap algorithms

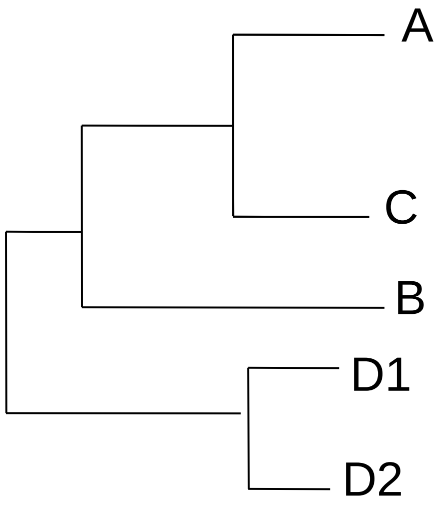
Reconciliation algorithm.

(Hard reconciliation) Resolve any incogurence between gene tree and species tree by introducing the minimal number of gene duplicatios and losses.

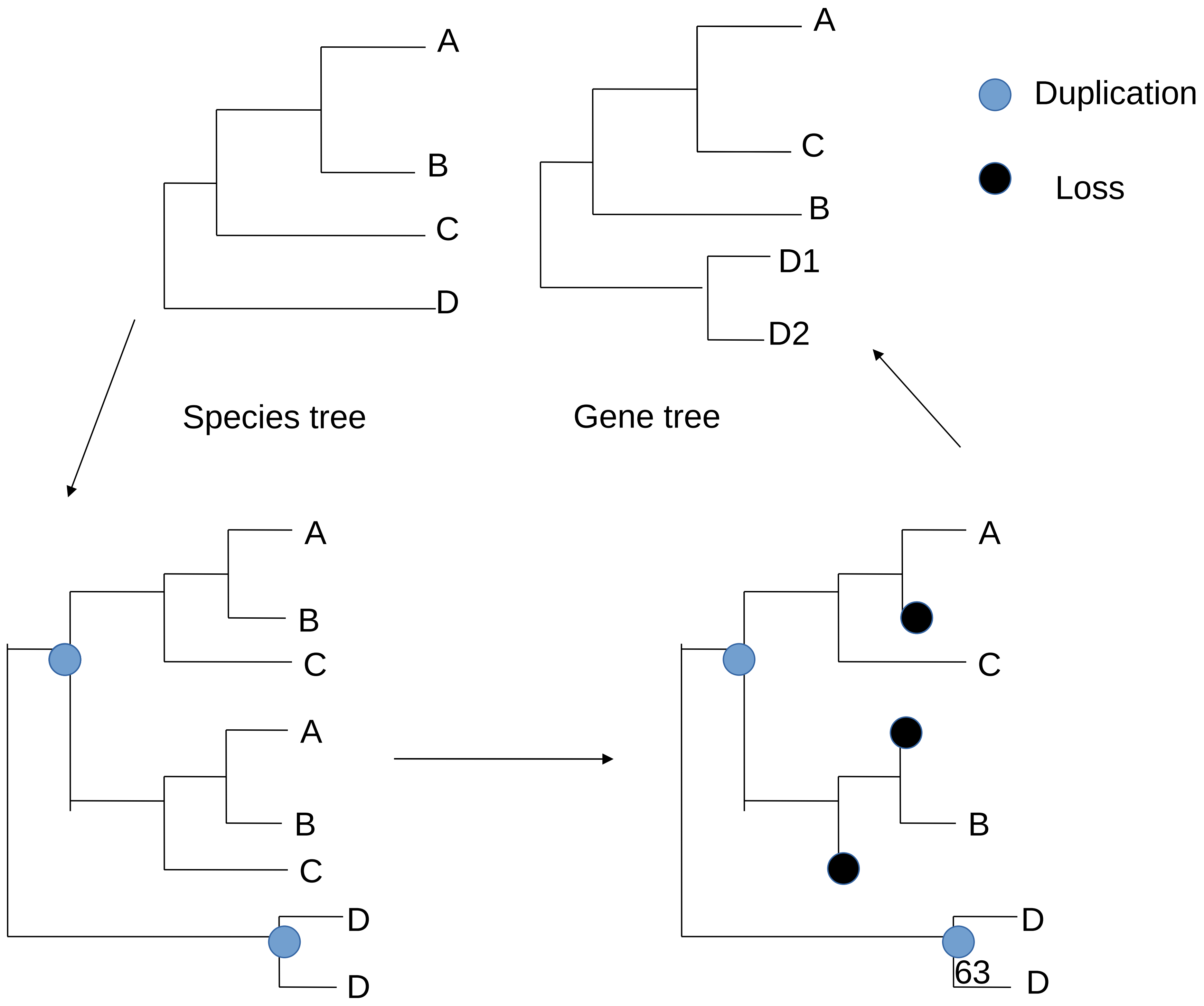
(Soft reconciliation) Allow incongruences below a given support value



Species tree



Gene tree



Species tree

Gene tree

● Duplication

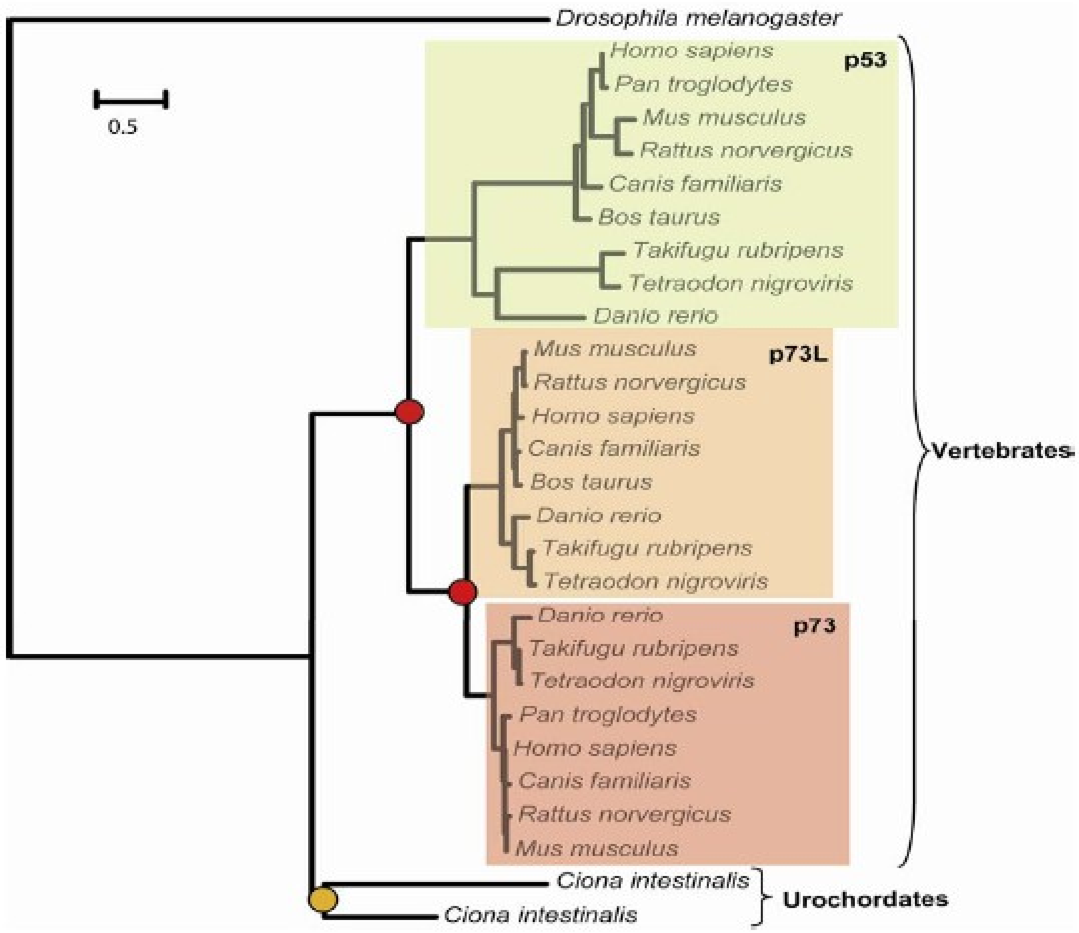
● Loss

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

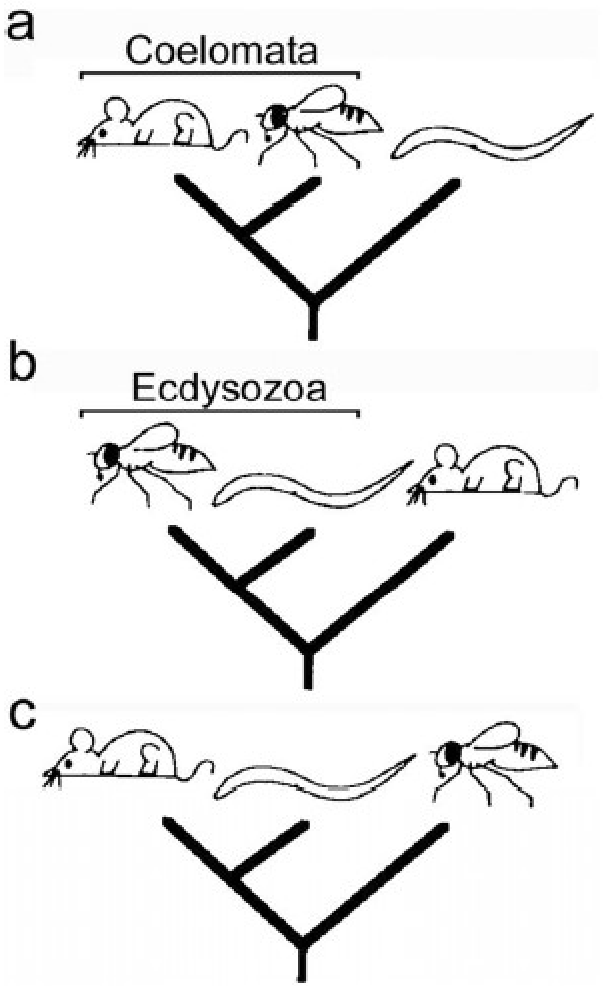
It works when these two assumptions are correct:

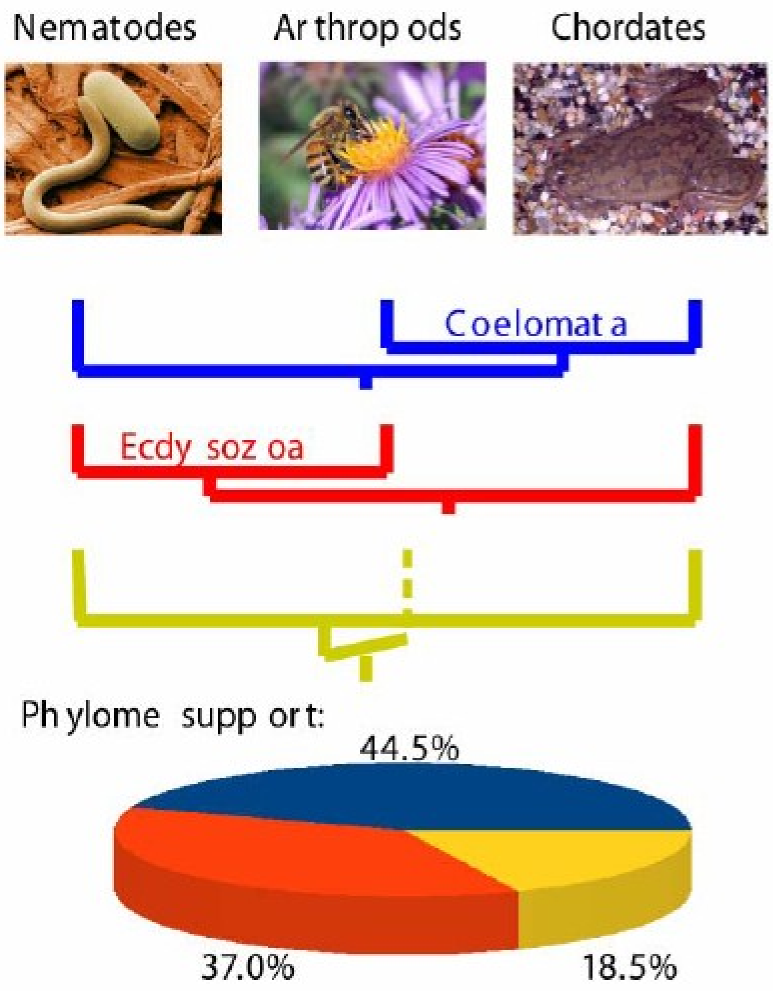
A) We know the true species tree

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

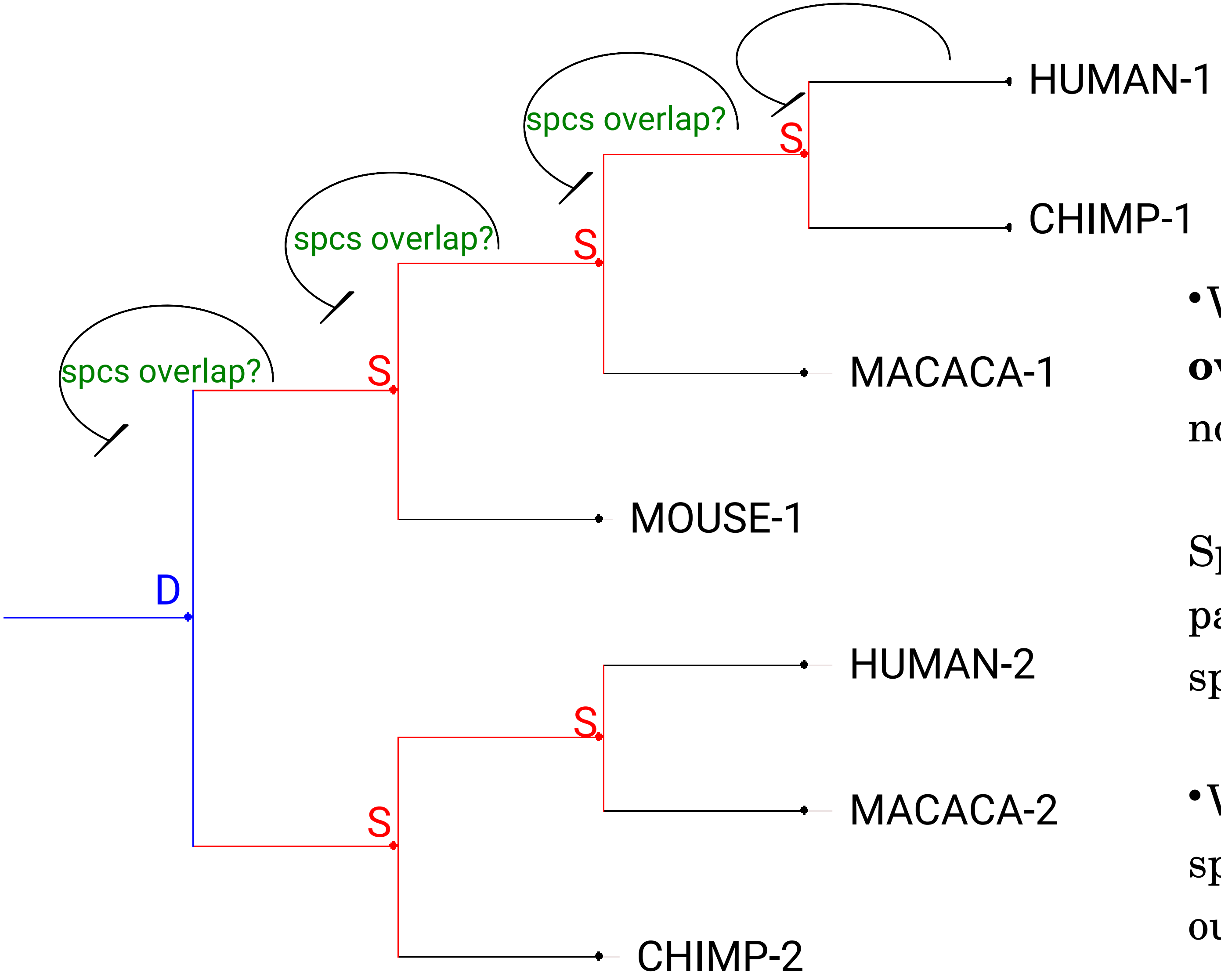
Primates
Rodents
laurasatheria

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.

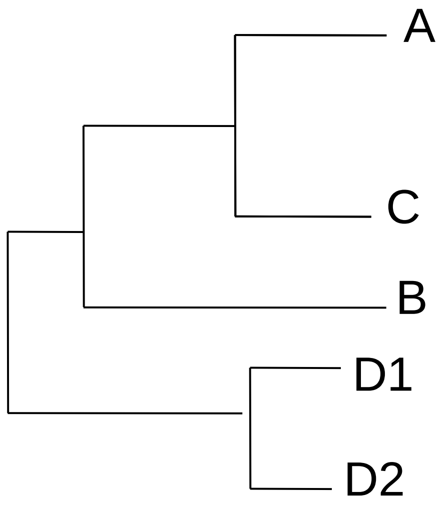
Species common to both partitions / sum of the species in both partitions

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

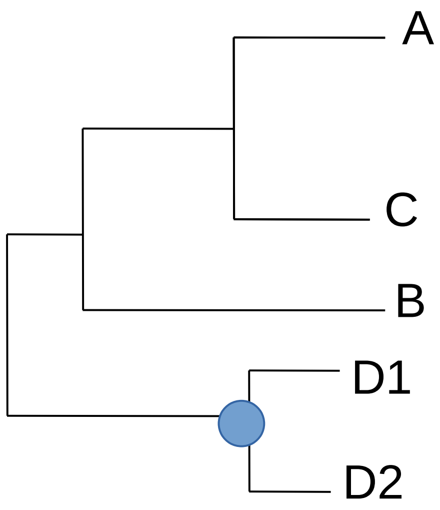
Species ovelap algorithm.

It does not require a species-tree but needs to know the species to which The genes belong
In essence can be seen as a reconciliation with an unresolved species tree

For every node in the gene tree evaluate whether the daughter partitions share any species. If the overlap (number of species shared over total number of species) is higher than the given threshold. Inpute a duplication at that node.



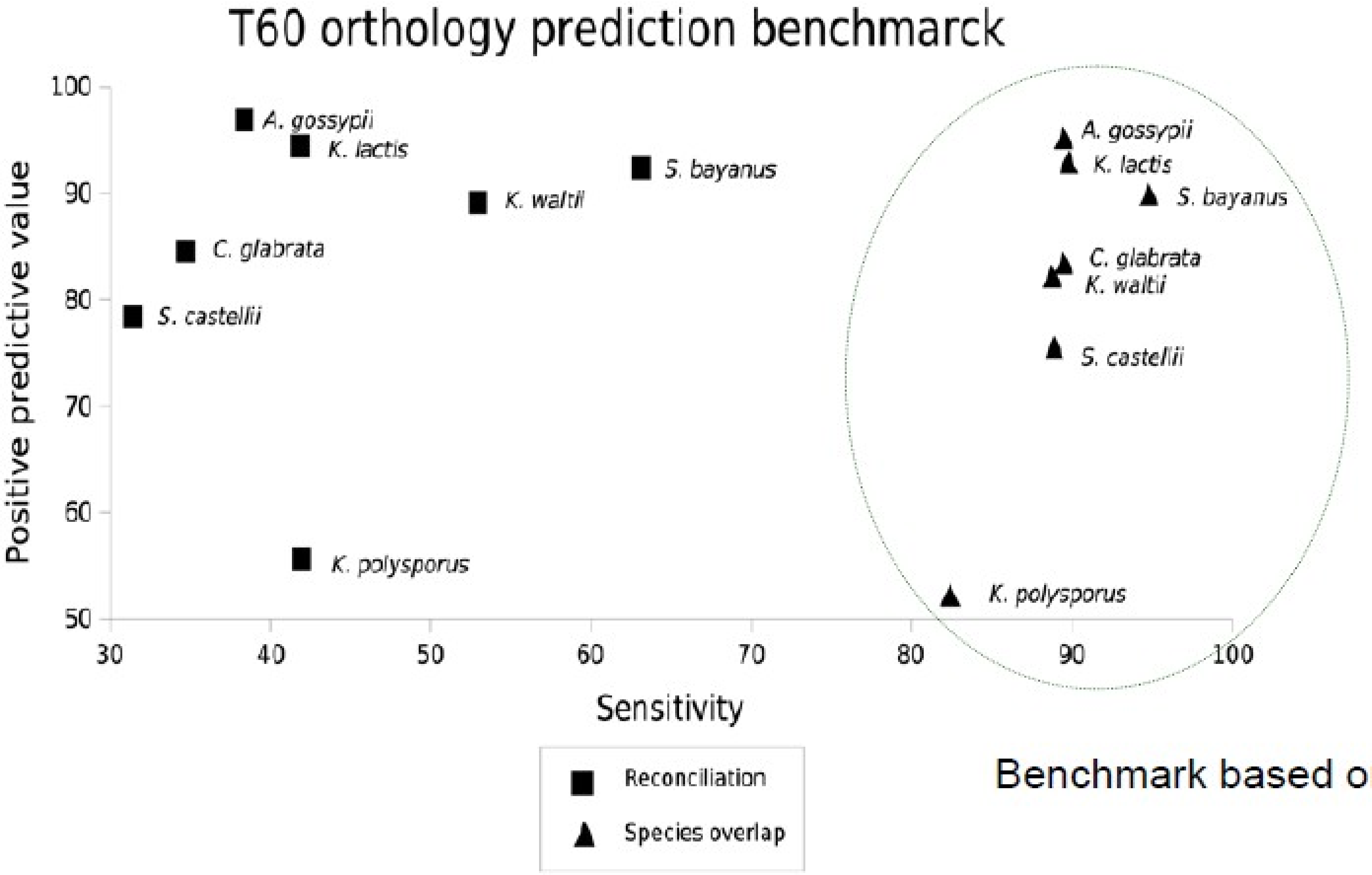
Gene tree



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)

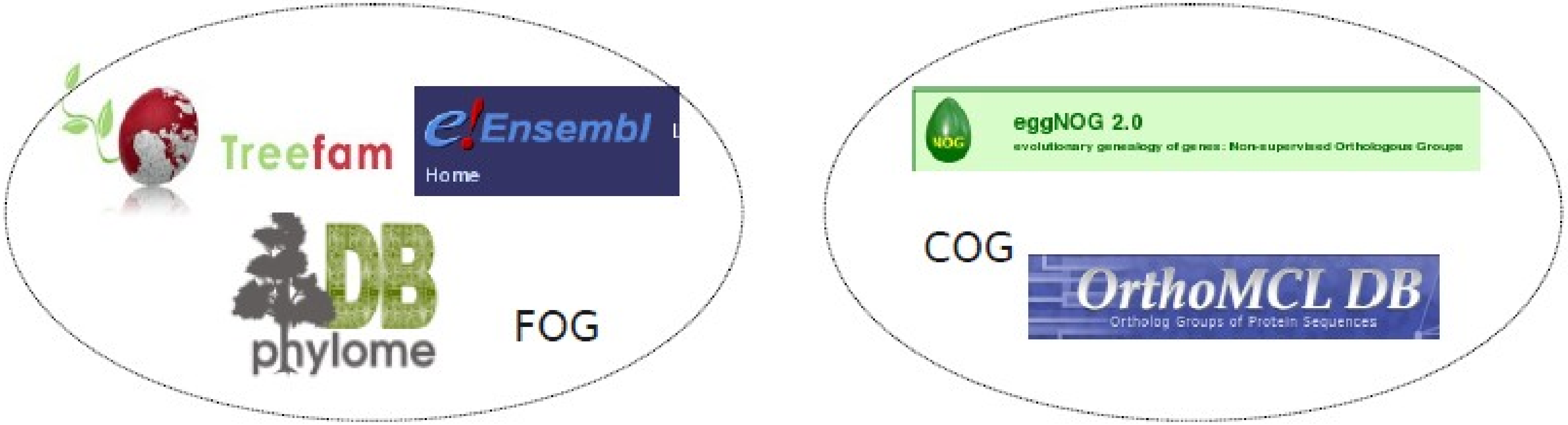


Benchmark based on YGOB (synteny)

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

MetaPhOrs

(Meta-Phylogeny-Based-Orthologs)


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

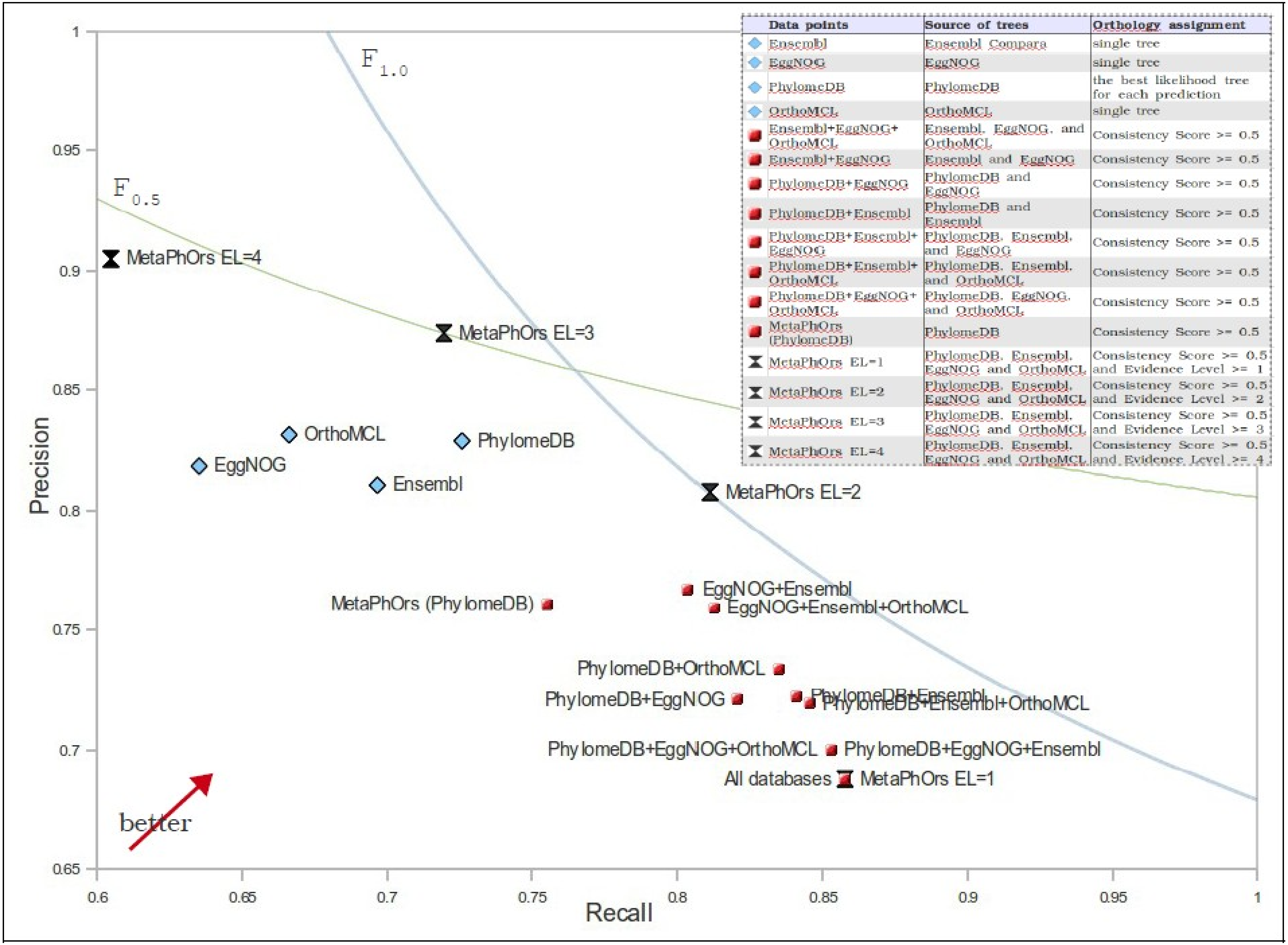
Use existing tree repositories

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

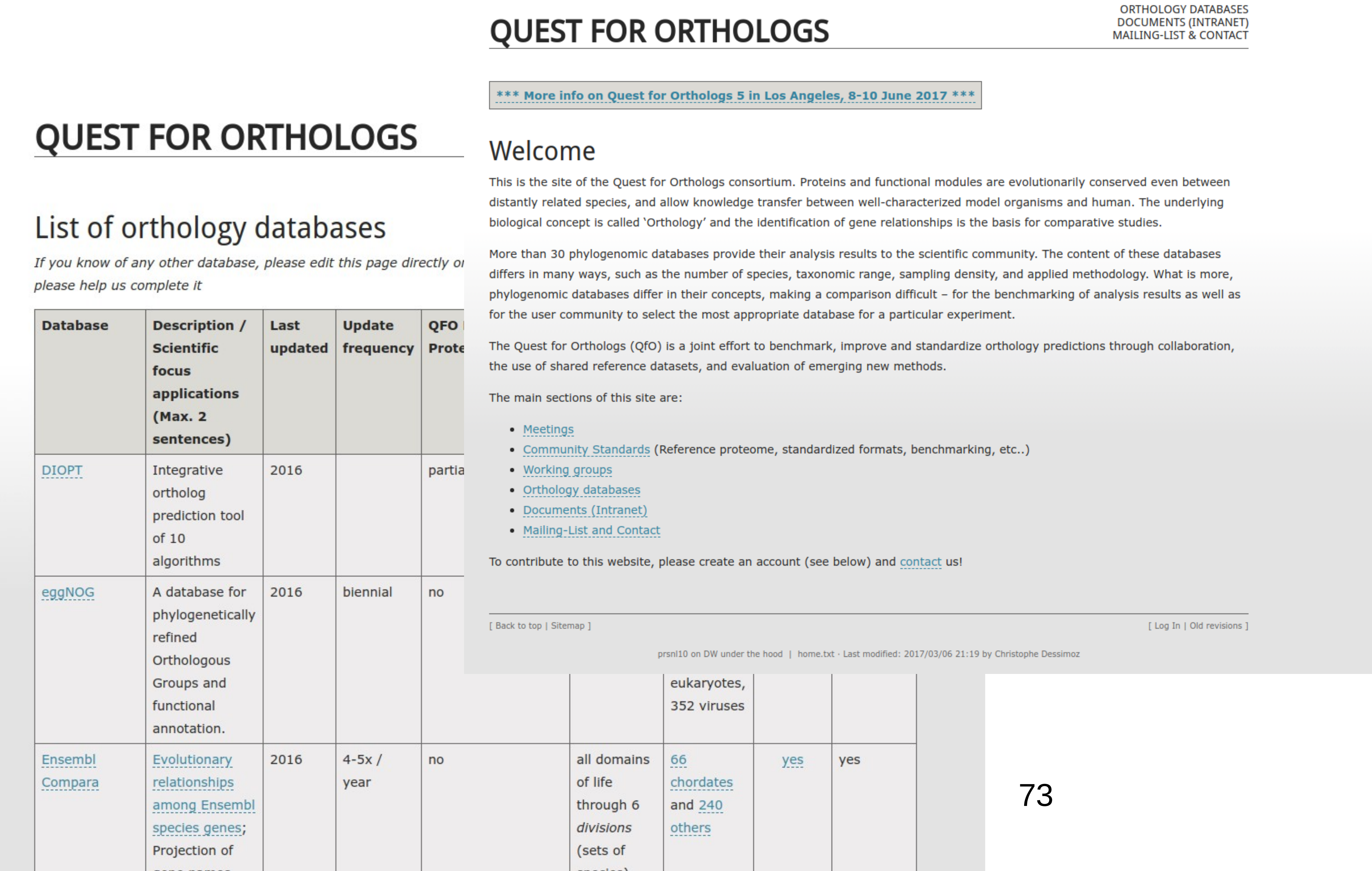
Pryszcz et. al. (NAR, 2011)



	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

A plethora of methods for ortholog prediction

http://questfororthologs.org



QUEST FOR ORTHOLOGS

List of orthology databases

If you know of any other database, please edit this page directly or
please help us complete it

Database	Description / Scientific focus applications (Max. 2 sentences)	Last updated	Update frequency	QFO Prote				
DIOPT	Integrative ortholog prediction tool of 10 algorithms	2016		partia				
eggNOG	A database for phylogenetically refined Orthologous Groups and functional annotation.	2016	biennial	no		eukaryotes, 352 viruses		
Ensembl Compara	Evolutionary relationships among Ensembl species genes; Projection of	2016	4-5x / year	no	all domains of life through 6 *divisions* (sets of	66 chordates and 240 others	yes	yes

QUEST FOR ORTHOLOGS

ORTHOLOGY DATABASES
DOCUMENTS (INTRANET)
MAILING-LIST & CONTACT

*** More info on Quest for Orthologs 5 in Los Angeles, 8-10 June 2017 ***

Welcome

This is the site of the Quest for Orthologs consortium. Proteins and functional modules are evolutionarily conserved even between distantly related species, and allow knowledge transfer between well-characterized model organisms and human. The underlying biological concept is called 'Orthology' and the identification of gene relationships is the basis for comparative studies.

More than 30 phylogenomic databases provide their analysis results to the scientific community. The content of these databases differs in many ways, such as the number of species, taxonomic range, sampling density, and applied methodology. What is more, phylogenomic databases differ in their concepts, making a comparison difficult – for the benchmarking of analysis results as well as for the user community to select the most appropriate database for a particular experiment.

The Quest for Orthologs (QfO) is a joint effort to benchmark, improve and standardize orthology predictions through collaboration, the use of shared reference datasets, and evaluation of emerging new methods.

The main sections of this site are:

- Meetings
- Community Standards (Reference proteome, standardized formats, benchmarking, etc..)
- Working groups
- Orthology databases
- Documents (Intranet)
- Mailing-List and Contact

To contribute to this website, please create an account (see below) and contact us!

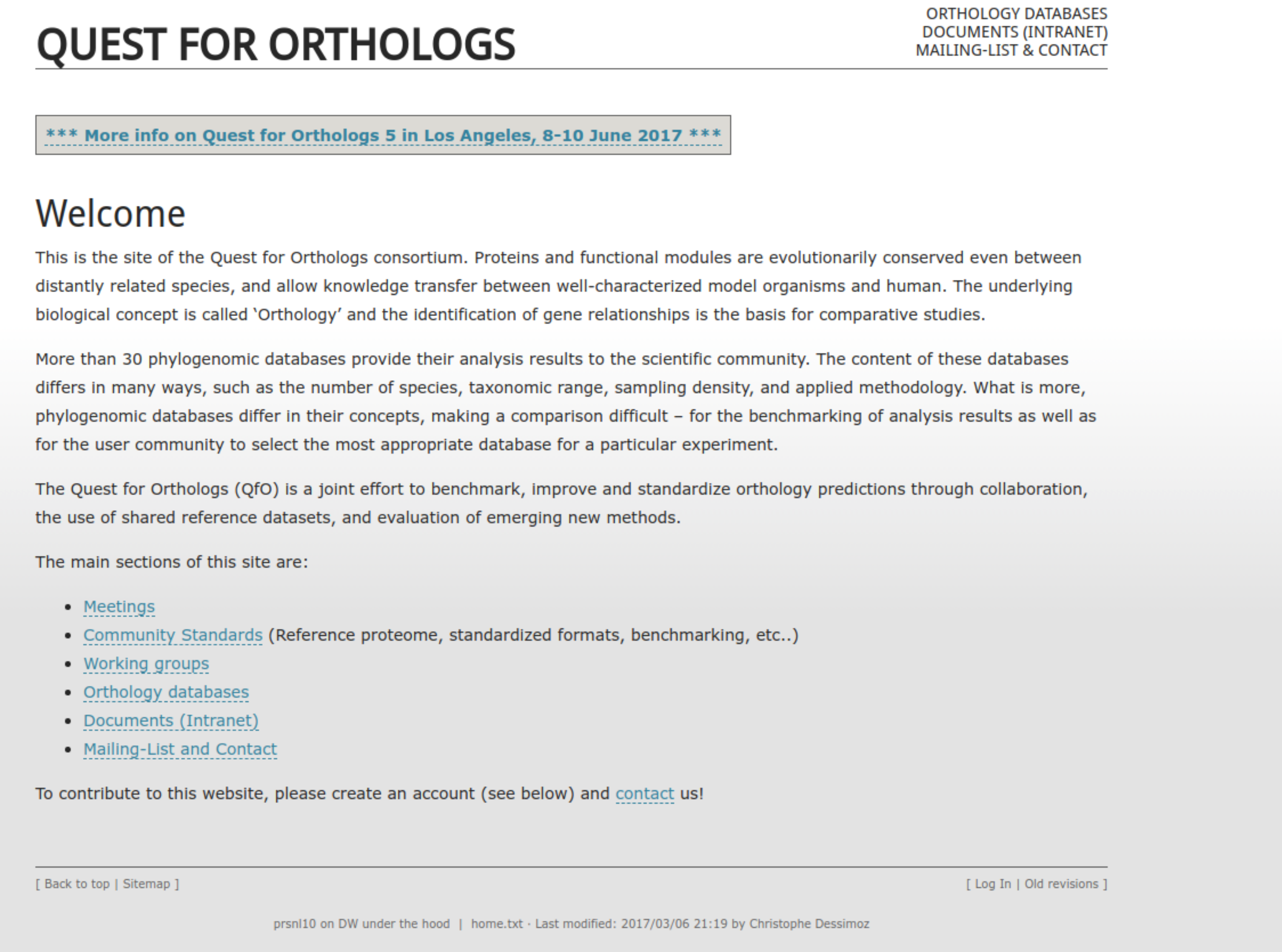
[Back to top | Sitemap]　　[Log In | Old revisions]

prsnl10 on DW under the hood | home.txt · Last modified: 2017/03/06 21:19 by Christophe Dessimoz

¿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?)

A plethora of methods for ortholog prediction



QUEST FOR ORTHOLOGS

*** More info on Quest for Orthologs 5 in Los Angeles, 8-10 June 2017 ***

Welcome

This is the site of the Quest for Orthologs consortium. Proteins and functional modules are evolutionarily conserved even between distantly related species, and allow knowledge transfer between well-characterized model organisms and human. The underlying biological concept is called 'Orthology' and the identification of gene relationships is the basis for comparative studies.

More than 30 phylogenomic databases provide their analysis results to the scientific community. The content of these databases differs in many ways, such as the number of species, taxonomic range, sampling density, and applied methodology. What is more, phylogenomic databases differ in their concepts, making a comparison difficult – for the benchmarking of analysis results as well as for the user community to select the most appropriate database for a particular experiment.

The Quest for Orthologs (QfO) is a joint effort to benchmark, improve and standardize orthology predictions through collaboration, the use of shared reference datasets, and evaluation of emerging new methods.

The main sections of this site are:

- Meetings
- Community Standards (Reference proteome, standardized formats, benchmarking, etc..)
- Working groups
- Orthology databases
- Documents (Intranet)
- Mailing-List and Contact

To contribute to this website, please create an account (see below) and contact us!

[Back to top | Sitemap]
[Log In | Old revisions]
prsnl10 on DW under the hood | home.txt · Last modified: 2017/03/06 21:19 by Christophe Dessimoz

THINGS TO CONSIDER:

Working with incomplete genomes (Transcriptome data, etc):

Check number of family members in related species with complete genomes

Compare relative distances with other genes

Use complete genomes as an anchor in blasts and phylogenies

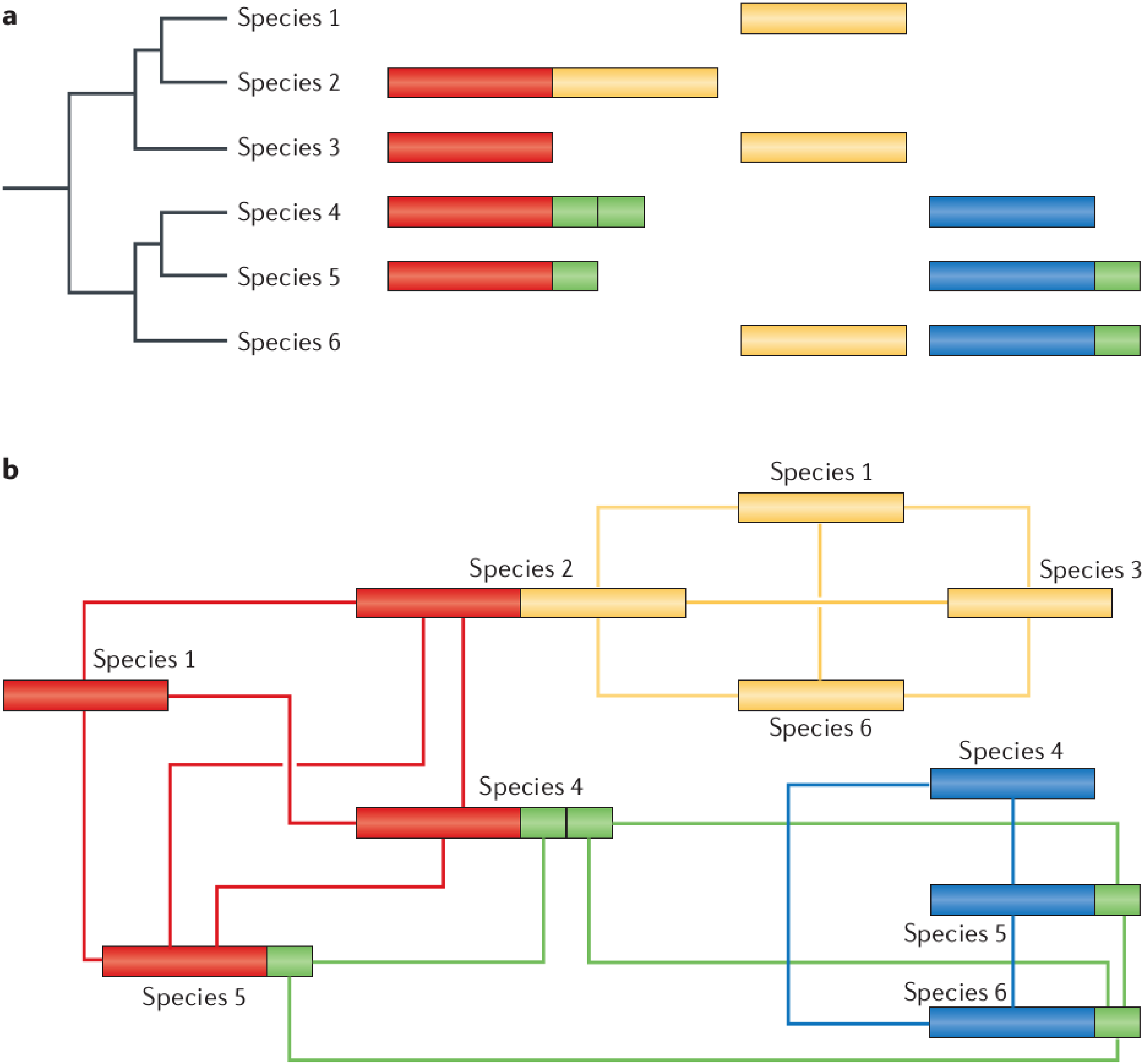
Be aware of artificial duplications caused by split gene models

Non-vertical modes of inheritance

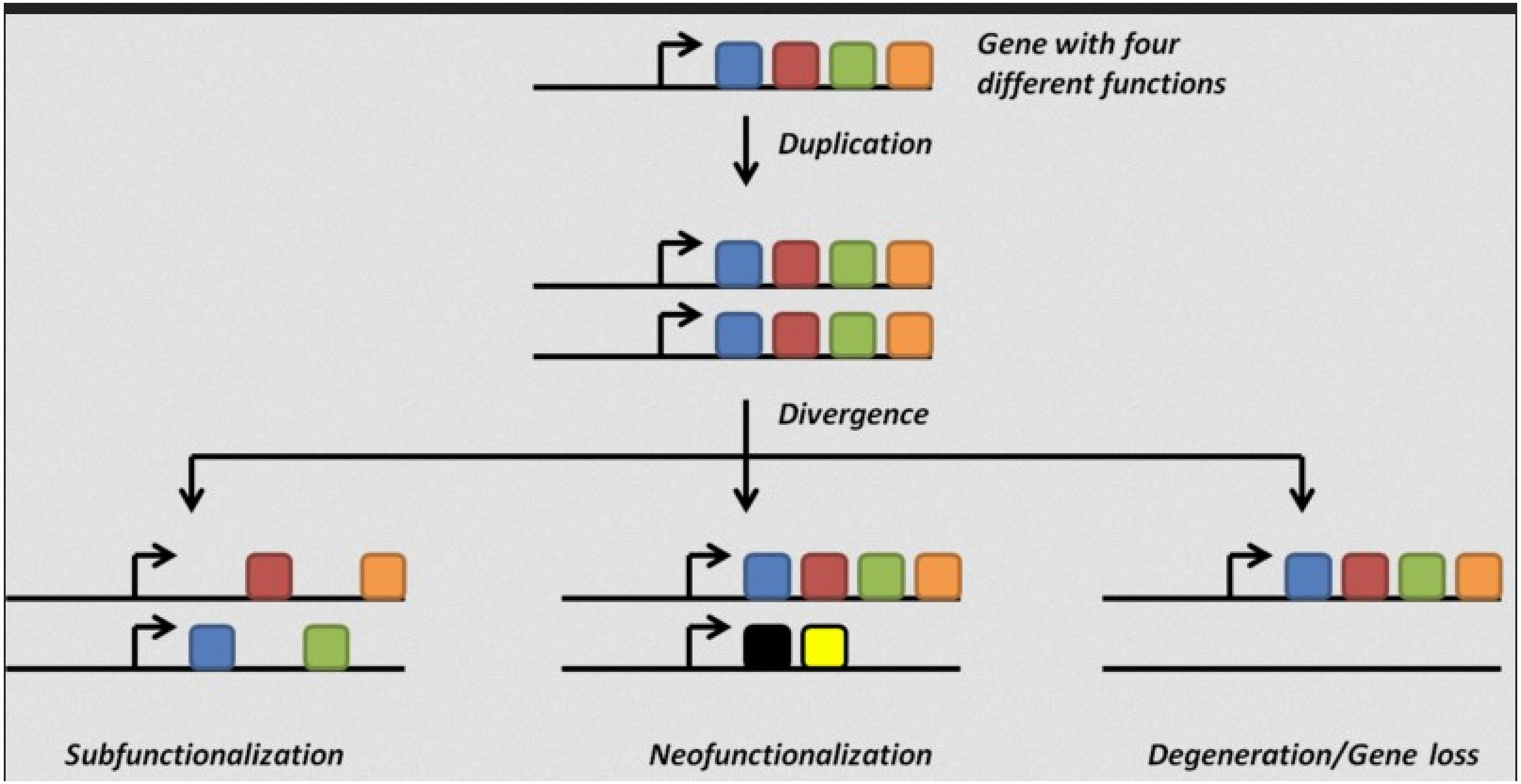
Multidomain proteins

Functional inference from orthology

Box 2 | Units of orthology



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


Gene with four different functions
Duplication
Divergence
Subfunctionalization
Neofunctionalization
Degeneration/Gene loss