Barcelona Biomedical Research Park

CRG Centre de Regulació Genòmica

Introduction to phylogeny, Orthology and Paralogy

Toni Gabaldón

Centre for Genomic Regulation (CRG), Barcelona

(tgabaldon@crg.es)

http://gabaldonlab.crg.es

Barcelona Supercomputing Center
Centro Nacional de Supercomputación

IRB BARCELONA
INSTITUTE FOR RESEARCH IN BIOMEDICINE

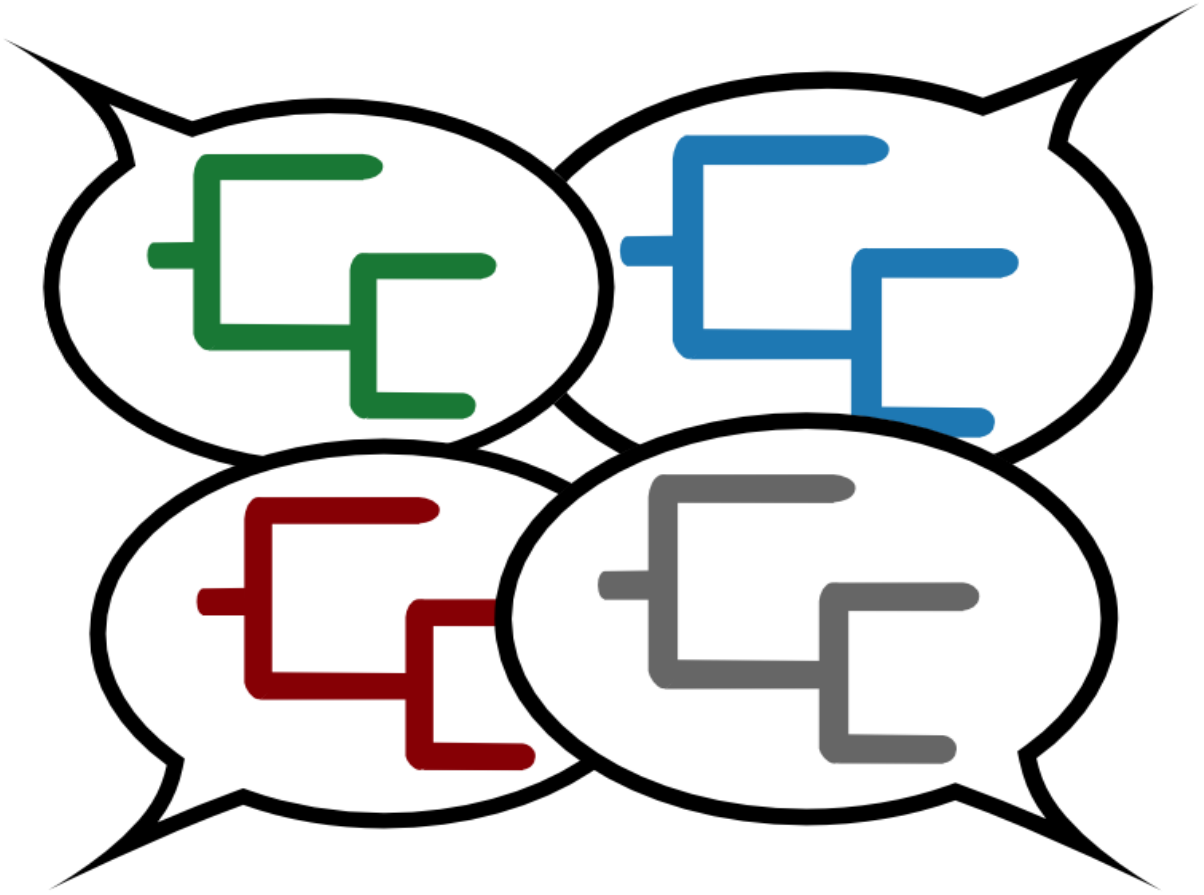
The ‘World’s Most Beautiful Data Center’ is a Supercomputer Housed in a Church

The MareNostrum 4 is only the world’s 25th most powerful supercomputer, but it definitely has the most style.

Image: Barcelona Supercomputing Center

From the outside, Torre Girona Chapel at the Polytechnic University of Catalonia in Barcelona looks like any one of the thousands of old churches that can be found throughout Spain, with a large cross mounted on the roof and a rose window perched above the entrance. Step through the chapel doors, however, and you won’t find any religious iconography or a congregation in prayer.

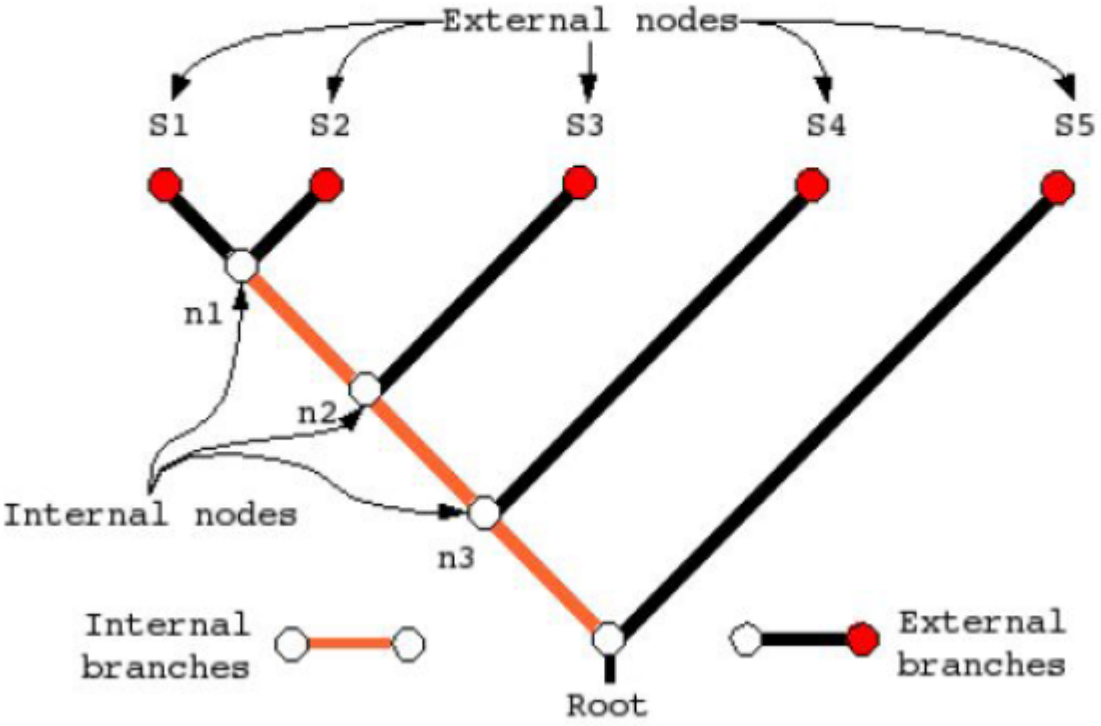
A brief primer on phylogenetics and tree reconstruction



A phylogenetic tree

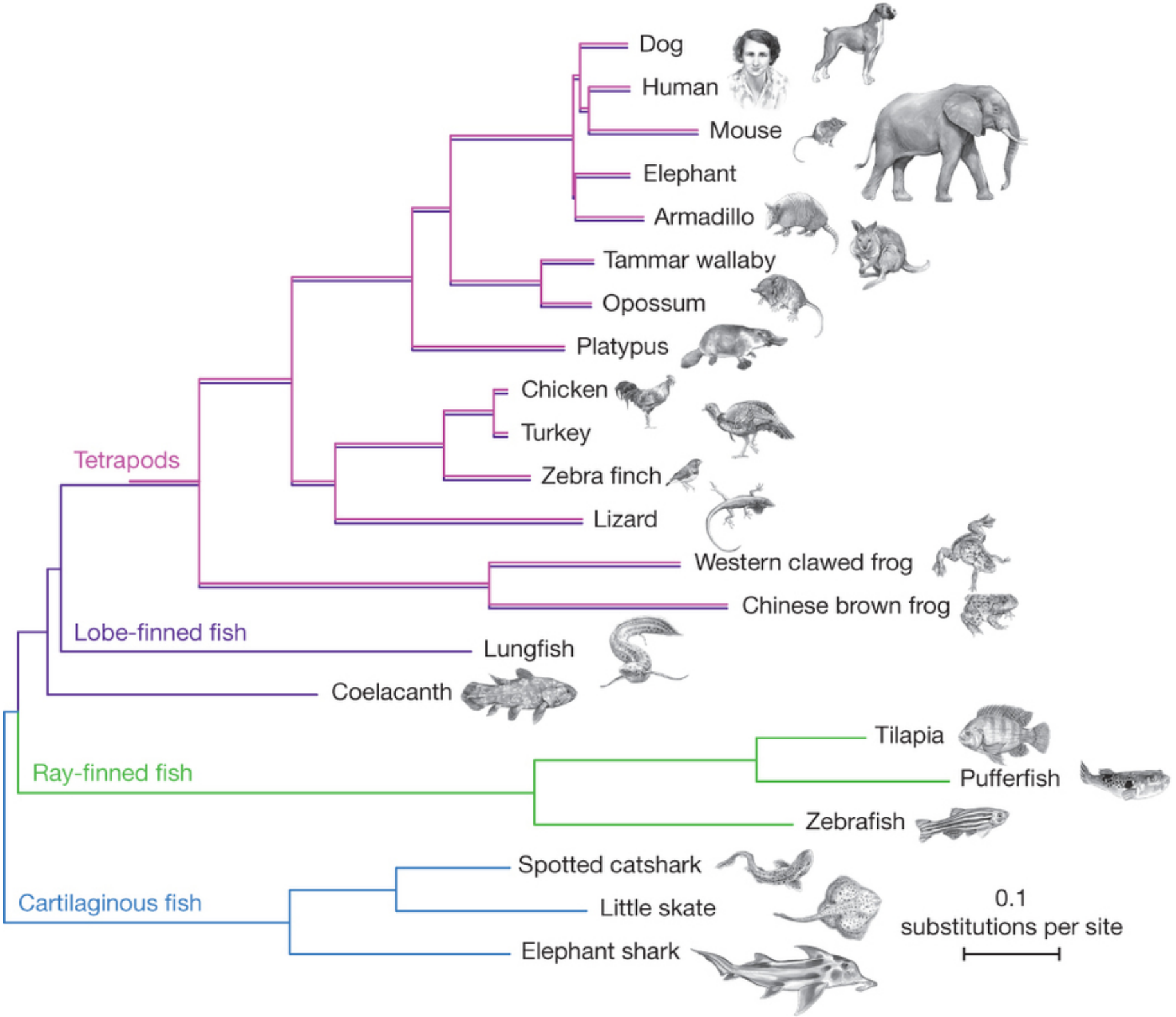
A branching diagram (**bipartite graph**) showing the inferred **evolutionary** relationships among various **biological species or other entities** (e.g sequences) **based on similarities and differences** in their physical and/or genetic characteristics.

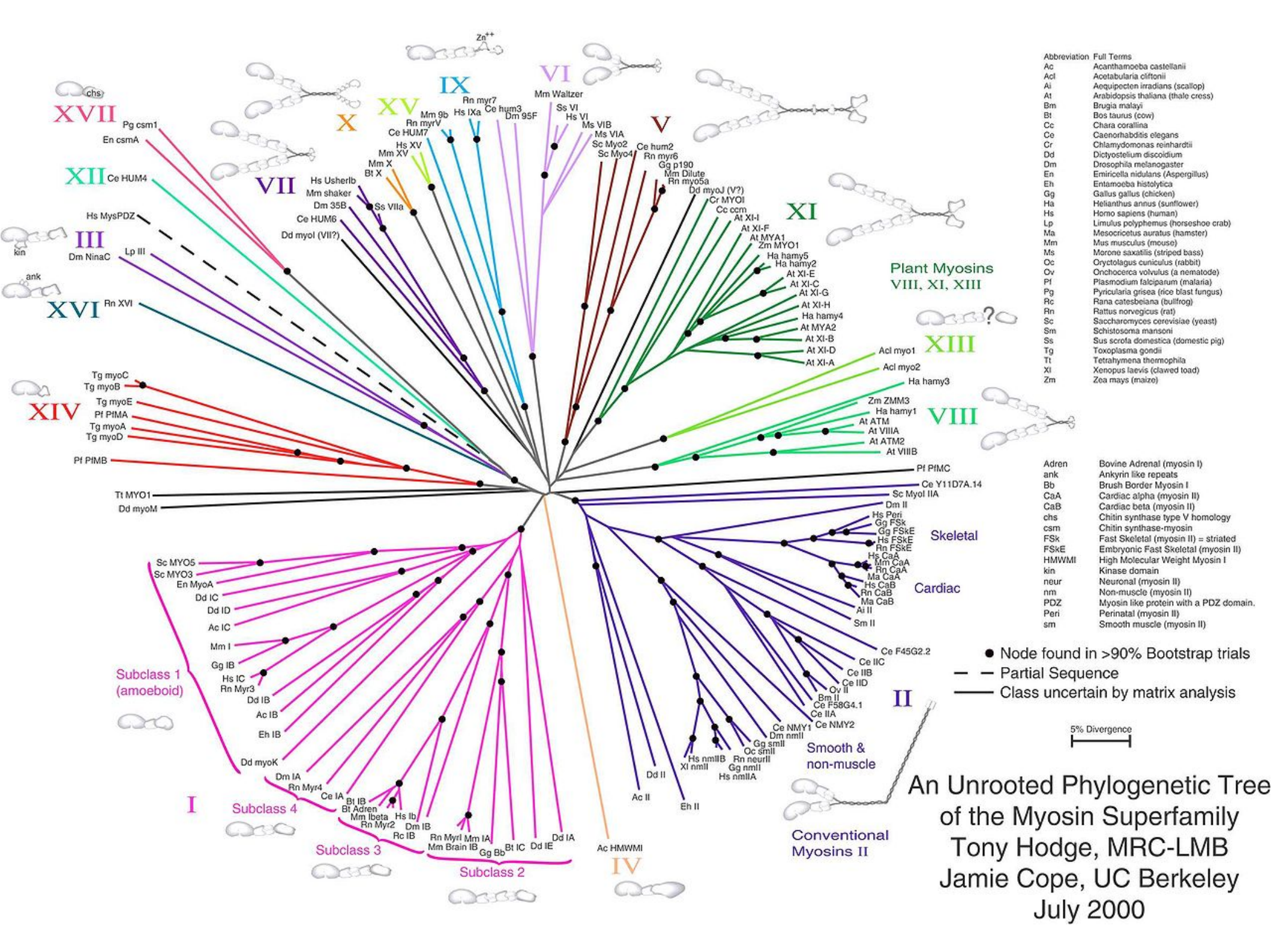
- **Nodes & branches.**Trees contain internal and external nodes and branches. In molecular phylogenetics, **external nodes** are sequences representing **genes, populations or species!**. Sometimes, **internal nodes** contain the ancestral information of the clustered species. A **branch** defines the relationship between sequences in terms of descent and ancestry.



A phylogenetic tree is a **hypothesis** of how things* are related through evolution

* species, genes, ….


Dog
Human
Mouse
Elephant
Armadillo
Tammar wallaby
Opossum
Platypus
Chicken
Turkey
Zebra finch
Lizard
Western clawed frog
Chinese brown frog
Tetrapods
Lobe-finned fish
Lungfish
Coelacanth
Tilapia
Pufferfish
Zebrafish
Ray-finned fish
Spotted catshark
Little skate
Elephant shark
Cartilaginous fish
0.1
substitutions per site



Abbreviation	Full Terms
Ac	Acanthamoeba castellanii
Acl	Acetabularia cliftonii
Ai	Aequipecten irradians (scallop)
At	Arabidopsis thaliana (thale cress)
Bm	Brugia malayi
Bt	Bos taurus (cow)
Cc	Chara corallina
Ce	Caenorhabditis elegans
Cr	Chlamydomonas reinhardtii
Dd	Dictyostelium discoidium
Dm	Drosophila melanogaster
En	Emiricella nidulans (Aspergillus)
Eh	Entamoeba histolytica
Gg	Gallus gallus (chicken)
Ha	Helianthus annus (sunflower)
Hs	Homo sapiens (human)
Lp	Limulus polyphemus (horseshoe crab)
Ma	Mesocricetus auratus (hamster)
Mm	Mus musculus (mouse)
Ms	Morone saxatilis (striped bass)
Oc	Oryctolagus cuniculus (rabbit)
Ov	Onchocerca volvulus (a nematode)
Pf	Plasmodium falciparum (malaria)
Pg	Pyricularia grisea (rice blast fungus)
Rc	Rana catesbeiana (bullfrog)
Rn	Rattus norvegicus (rat)
Sc	Saccharomyces cerevisiae (yeast)
Sm	Schistosoma mansoni
Ss	Sus scrofa domestica (domestic pig)
Tg	Toxoplasma gondii
Tt	Tetrahymena thermophila
Xl	Xenopus laevis (clawed toad)
Zm	Zea mays (maize)

Adren	Bovine Adrenal (myosin I)
ank	Ankyrin like repeats
Bb	Brush Border Myosin I
CaA	Cardiac alpha (myosin II)
CaB	Cardiac beta (myosin II)
chs	Chitin synthase type V homology
csm	Chitin synthase-myosin
FSk	Fast Skeletal (myosin II) = striated
FSkE	Embryonic Fast Skeletal (myosin II)
HMWMI	High Molecular Weight Myosin I
kin	Kinase domain
neur	Neuronal (myosin II)
nm	Non-muscle (myosin II)
PDZ	Myosin like protein with a PDZ domain.
Peri	Perinatal (myosin II)
sm	Smooth muscle (myosin II)

● Node found in >90% Bootstrap trials

– – Partial Sequence

—— Class uncertain by matrix analysis

5% Divergence

An Unrooted Phylogenetic Tree of the Myosin Superfamily

Tony Hodge, MRC-LMB
Jamie Cope, UC Berkeley

July 2000

A phylogenetic tree is a **hypothesis** of how things* are related through evolution

How we find our best hypothesis:

- data (i.e. sequences)
- a “model” of how this type of data evolve
- a way to assess how good our hypothesis is as compared to other possible hypotheses

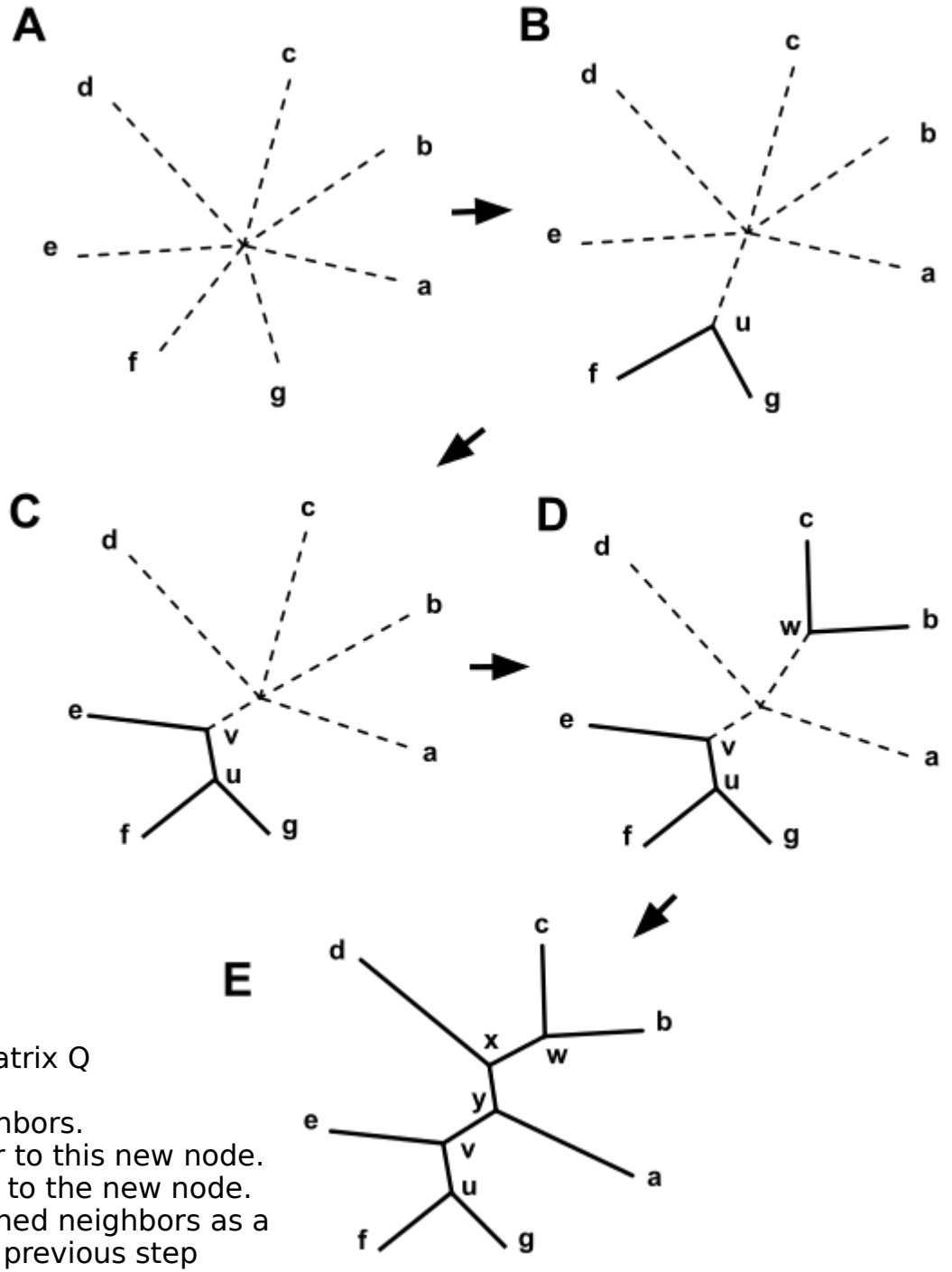
Phylogenetic approaches:

Distance methods (NJ, UPGMA)

Maximum Parsimony

Probabilistic Methods (Maximum Likelihood, Bayesian Inference, including coalescence methods)

Neighbor Joining



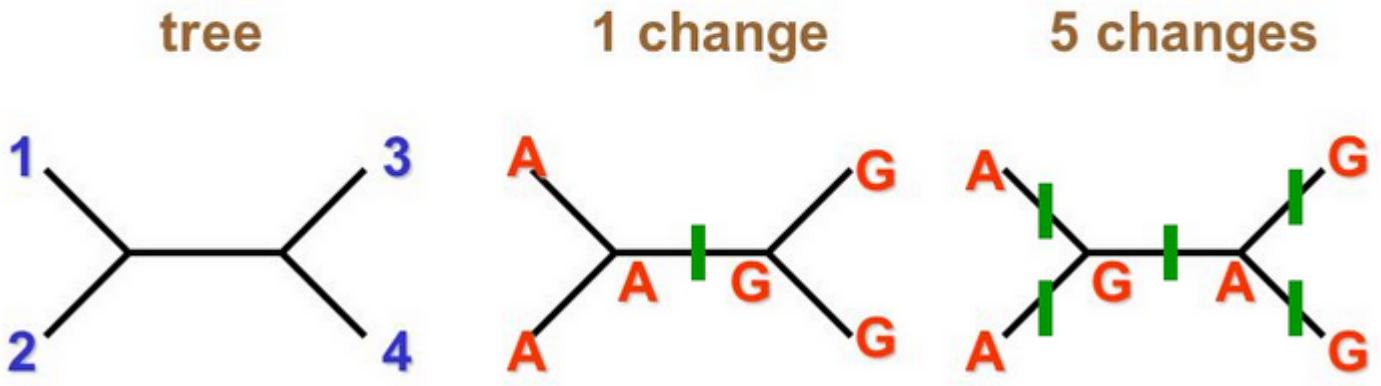
Based on the current distances matrix calculate the matrix Q
2. Find the pair of taxa in Q with the lowest value.
Create a node on the tree that joins these closest neighbors.
3. Calculate the distance of each of the taxa in the pair to this new node.
4. Calculate the distance of all taxa outside of this pair to the new node.
5. Start the algorithm again, considering the pair of joined neighbors as a single taxon and using the distances calculated in the previous step

Maximum Parsimony

Finding the tree with that implies the minimal number of changes along its branches.

Taxon-1	ATATT
Taxon-2	ATCGT
Taxon-3	GCAGT
Taxon-4	GCCGT

For each site, the goal is to reconstruct the evolution of that site on a tree subject to the constraint of invoking the fewest possible evolutionary changes.



So, how to find the best tree?

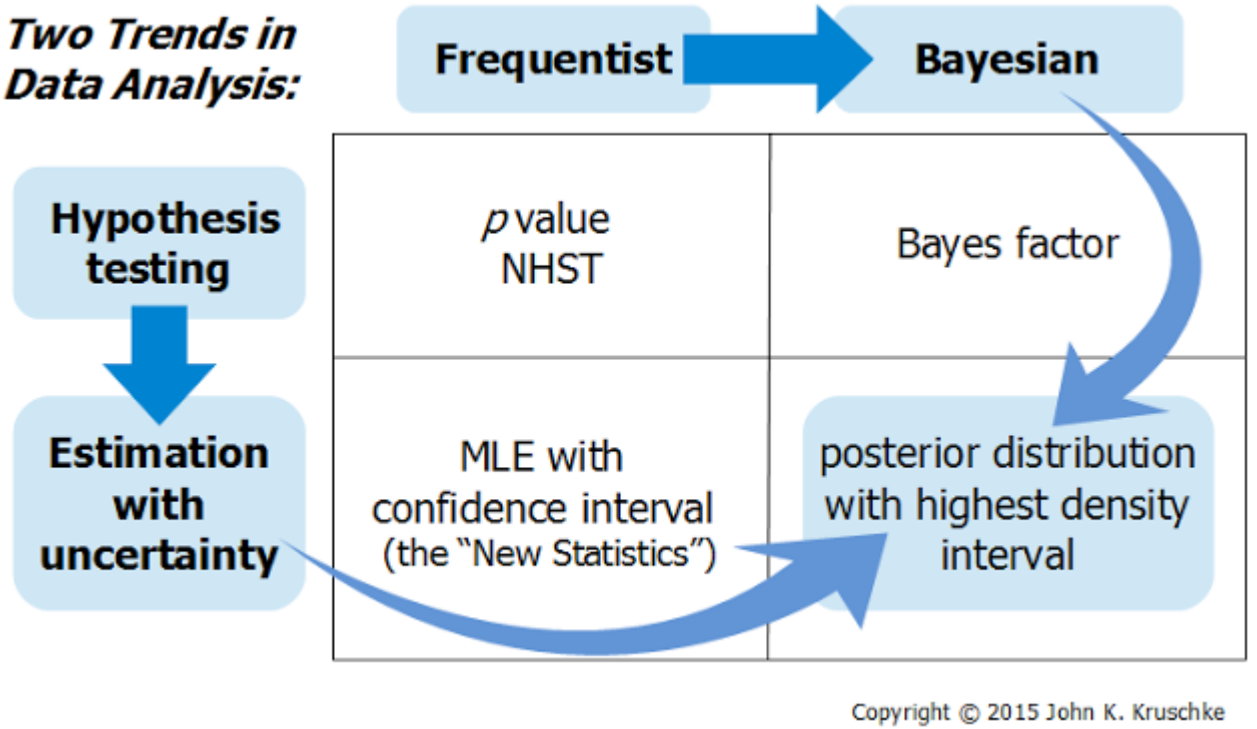
Exhaustive search: make ALL trees first, and then see which one best fits the data (you need an optimality criterion)

Heuritisc search: Try to find a way to find an optimal tree (hopefully the best) without testing them all. You also need an optimality criterion and you are not guaranteed to find the best, but you save time.

Different software differ in searching heuristics

Number of taxa *T*	Number of unrooted bifurcating trees *B(T)*
3	1
4	3
5	15
6	105
7	945
8	10,395
9	135,135
10	2,027,025
22	3×10^{23}
50	3×10^{74}

Probabilistic methods render themselves for testing



Likelihood

Given some data **(D)** a decission must be made about an adequate explanation (**H**, hypothesis)

D: alignment

H: Model of evolution, tree topology, branch lengths, parameters of the model

--> Each H will have a certain probability of producing the data P(D/H)

The best **H** is that of the greatest **P**

Important remark!!

The likelihood function **is not** the probability of a hypothesis being correct!!

The likelihood function is defined in terms of probability of producing the observed events not of the unknown parameters

Thus: the probability of observing the data has nothing to do with the probability that the underlying model is correct.

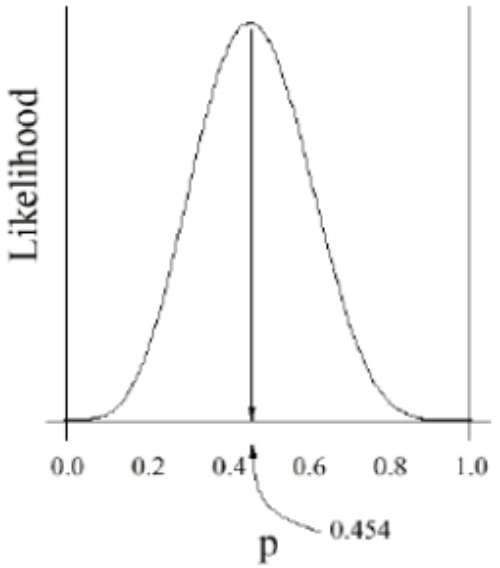
If tosses are all **independent**, and all have the same **unknown heads probability** p, then the observing sequence of tosses:

HHTTHTHHTTT

we can calculate the ML of these data as:

$$L = Prob(D/p) = pp(1-p)(1-p)p(1-p)pp(1-p)(1-p)(1-p) = p^5(1-p)^6$$

Ploting L against p, we observe the probabilities of the same data (D) for different values of p.



Thus the ML or the maximum probability to observe the above sequence of events is at $p = 0.4545$,

Suppose we have:

- **Data**:

Sequence 1 **C C A T**

Sequence 2 **C C G T**

- **Model**:[29]

$$\pi = [0.1, 0.4, 0.2, 0.3]$$

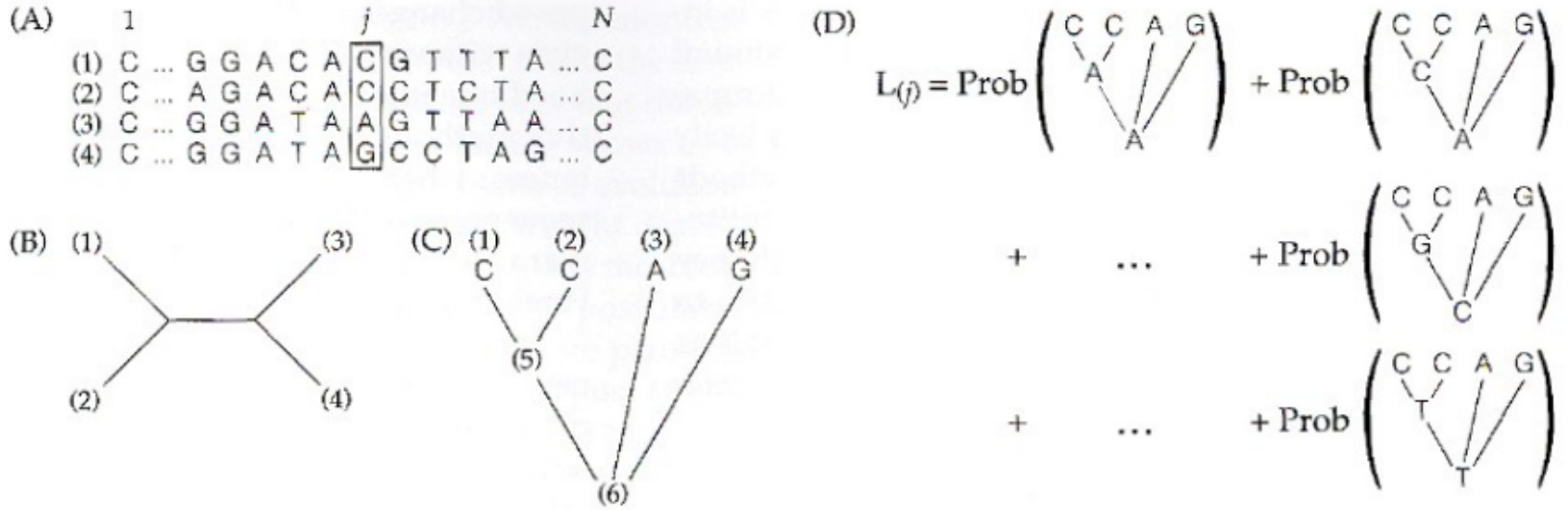
$$\boldsymbol{P} = \begin{bmatrix} 0.976 & 0.01 & 0.007 & 0.007 \\ 0.002 & 0.983 & 0.005 & 0.01 \\ 0.003 & 0.01 & 0.979 & 0.007 \\ 0.002 & 0.013 & 0.005 & 0.979 \end{bmatrix}$$

$$L_{(Seq.1 \to Seq.2)} = \pi_C P_{C \to C} \pi_C P_{C \to C} \pi_A P_{A \to G} \pi_T P_{T \to T}$$
$$0.4\text{x}0.983\text{x}0.4\text{x}0.983\text{x}0.1\text{x}0.007\text{x}0.3\text{x}0.979$$
$$= 0.0000300$$

$$lnL_{tree:Seq_1 \to Seq_2} = -10.414$$

[29] Note that the base composition sum one, but indeed the the rows of substitution matrix sum one. Why?

computation in a real problem



- Tree after rooting in an arbitrary node (reversible model).
- The likelihood for a particular site is the sum of the probabilities of every possible reconstruction of ancestral states given some model of base substitution.
- The likelihood of the tree is the product of the likelihood at each site.

$$L = L_{(1)} \cdot L_{(2)} \cdot \ldots \cdot L_{(N)} = \prod_{j=1}^{N} L_{(j)}$$

- The likelihood is reported as the sum of the log likelihhod of the full tree.

$$lnL = lnL_{(1)} + lnL_{(2)} + \ldots + lnL_{(N)} = \sum_{j=1}^{N} lnL_{(j)}$$

Models can be very complex to capture different processes, increasing the number of parameters

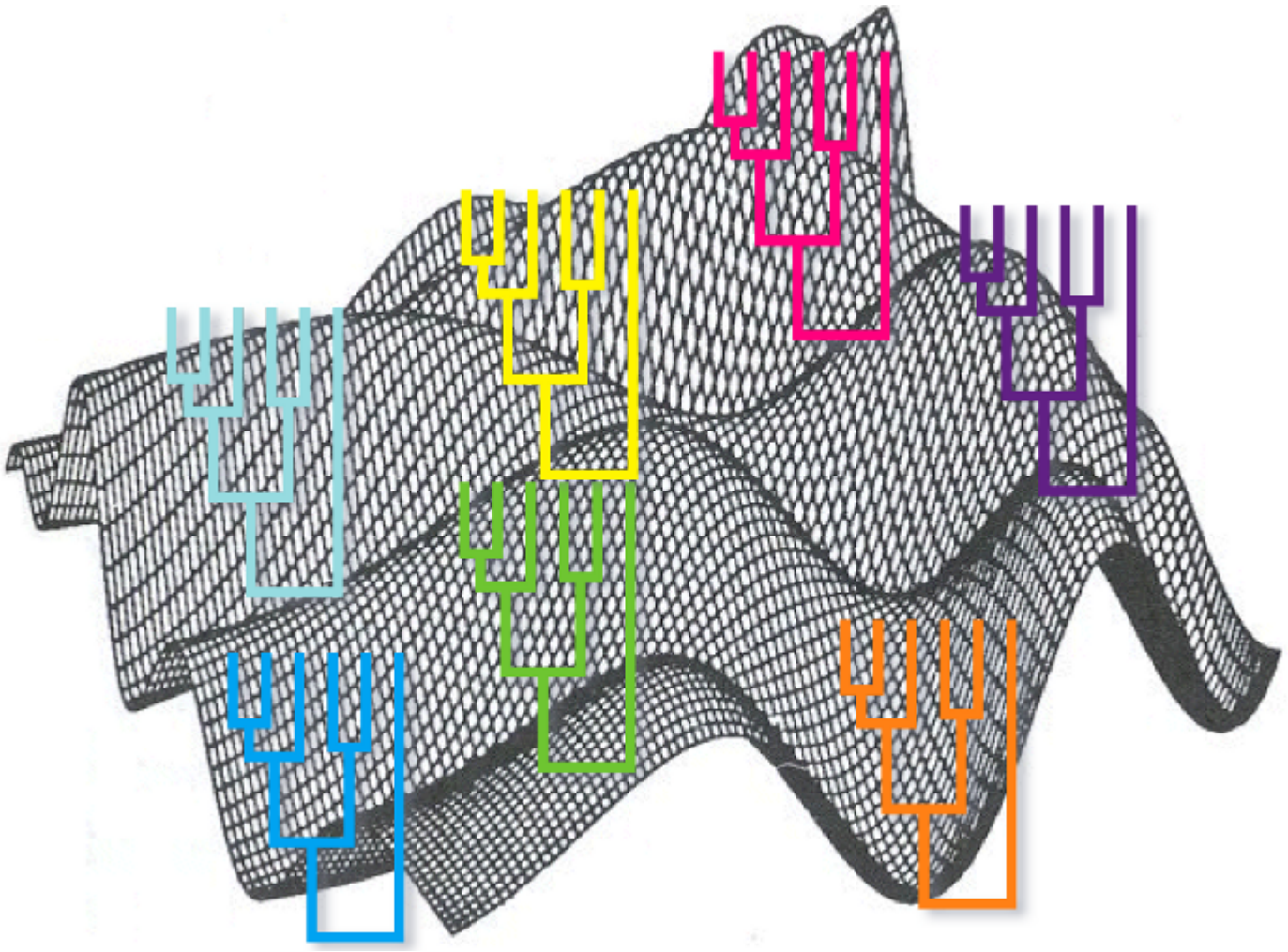
- i.e. fast and slow evolving sites

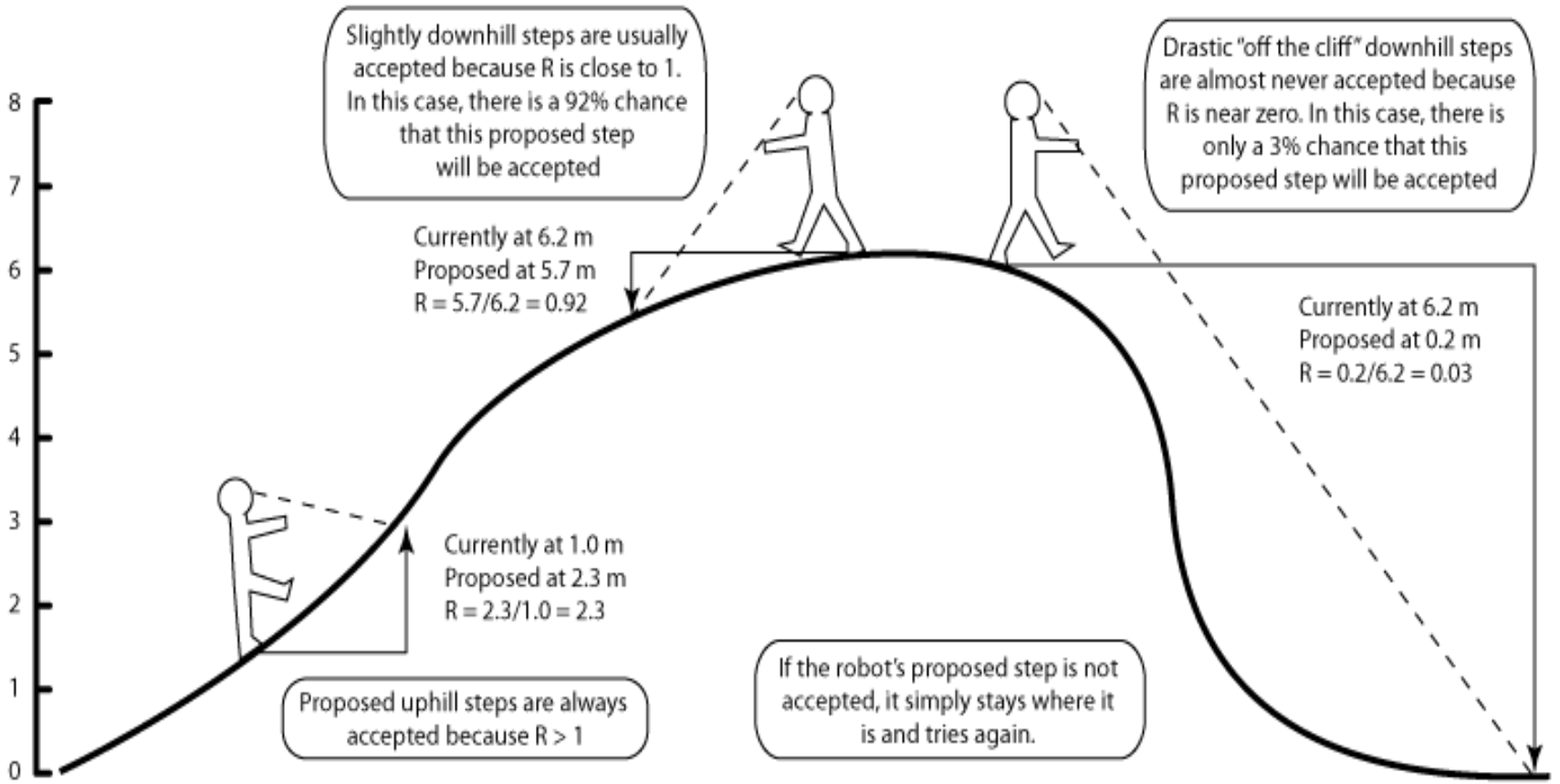
Models can be specificly made for specific groups of sequences

- i.e. for mitochondrially encoded proteins

Model choice can influence results.

Each model would indiuce a different likelihood landscape




Slightly downhill steps are usually accepted because R is close to 1. In this case, there is a 92% chance that this proposed step will be accepted
Drastic "off the cliff" downhill steps are almost never accepted because R is near zero. In this case, there is only a 3% chance that this proposed step will be accepted
Currently at 6.2 m
Proposed at 5.7 m
R = 5.7/6.2 = 0.92
Currently at 6.2 m
Proposed at 0.2 m
R = 0.2/6.2 = 0.03
Currently at 1.0 m
Proposed at 2.3 m
R = 2.3/1.0 = 2.3
Proposed uphill steps are always accepted because R > 1
If the robot's proposed step is not accepted, it simply stays where it is and tries again.
0
1
2
3
4
5
6
7
8

Bayesian Inference

♣ **Maximum Likelihood** will find the tree that is most likely to have produced the observed sequences, or formally $P(D/H)$ (the probability of seeing the data given the hypothesis).

♠ **A Bayesian approach** will give you the tree (or set of trees) that is most likely to be explained by the sequences, or formally $P(H/D)$ (the probability of the hypothesis being correct given the data).

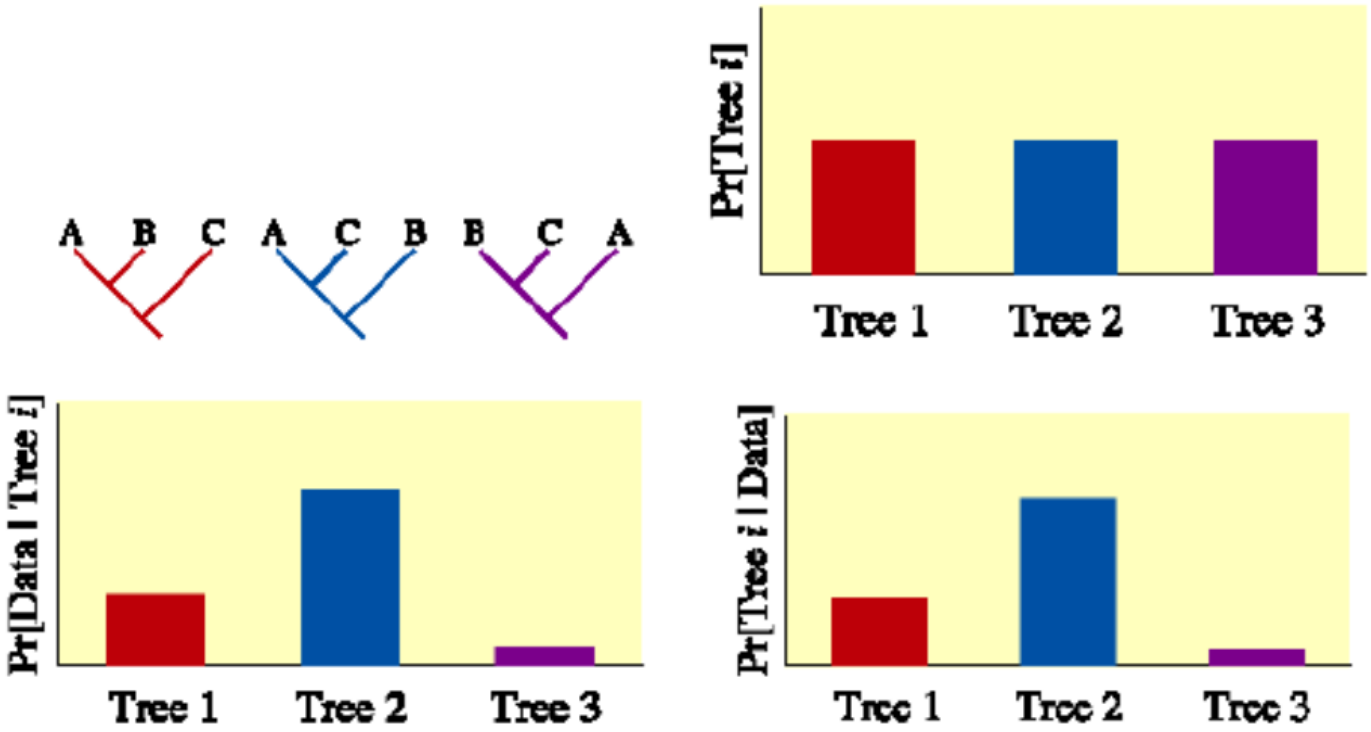
♢ **Bayes Theorem** provides a way to calculate the probability of a model (*tree topology and evolutionary model*) from the results it produces (*the aligned sequences we have*), what we call a **posterior probability**[31].

Thomas Bayes (1702-1761)

$$P(\theta/D) = \frac{P(\theta) \cdot P(D/\theta)}{P(D)}$$

The main components of Bayes analysis

- $P(\theta)$ The **prior probability** of a tree represents the probability of the tree before the observations have been made. Typically, all trees are considered equally probable.

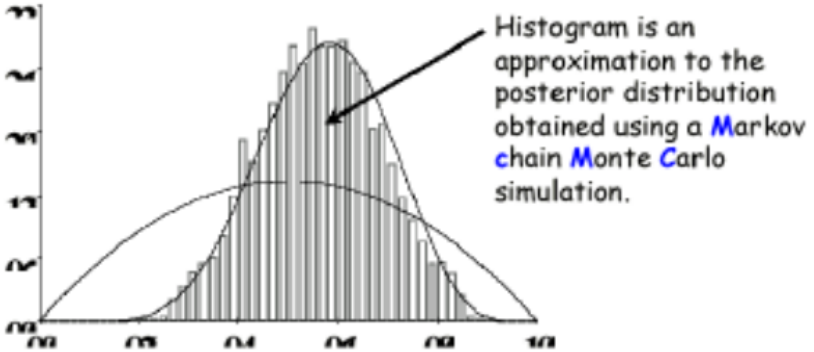


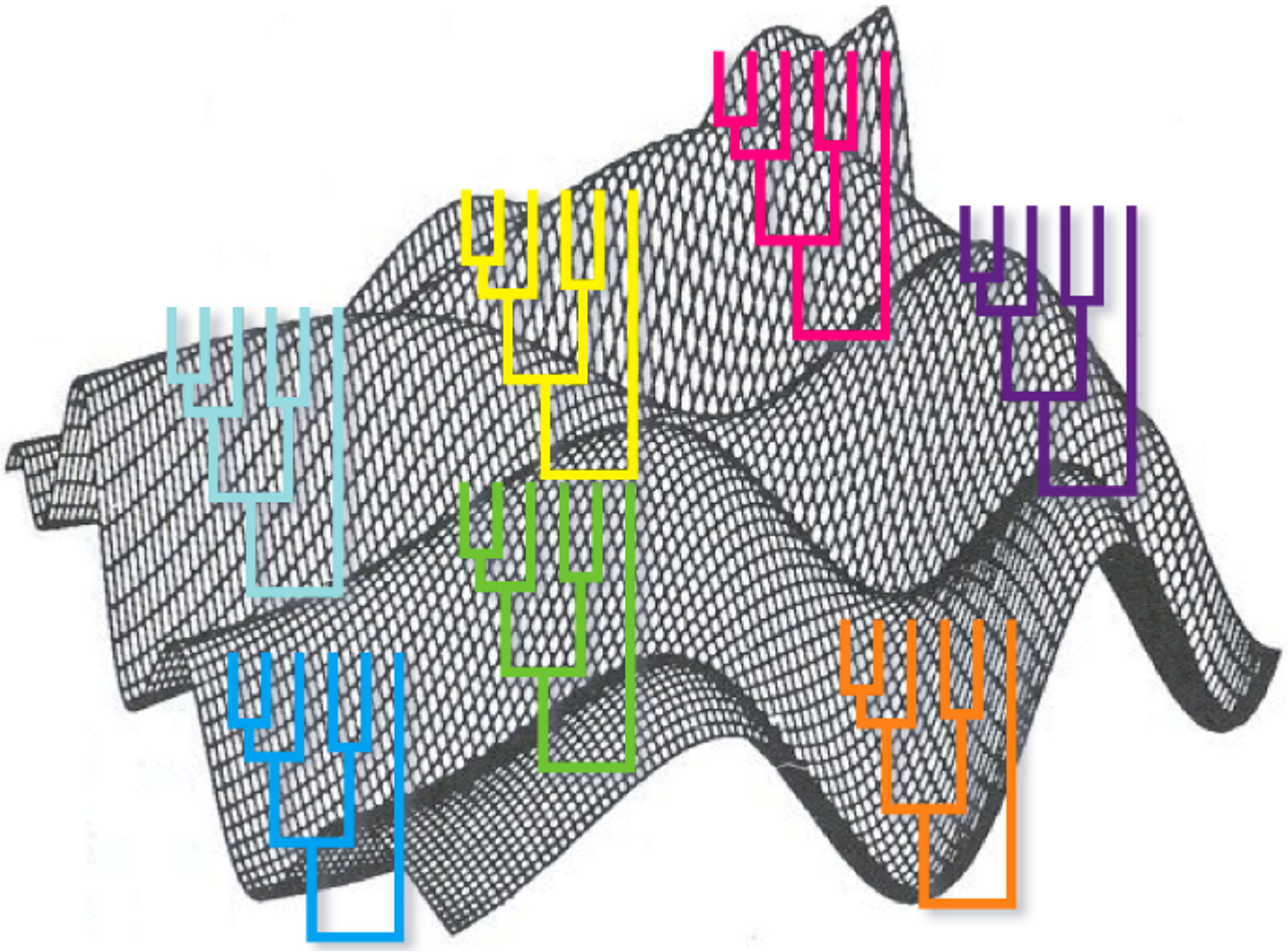
- $P(D/\theta)$ The **likelihood** is proportional to the probability of the observations (data sets) conditional on the tree.

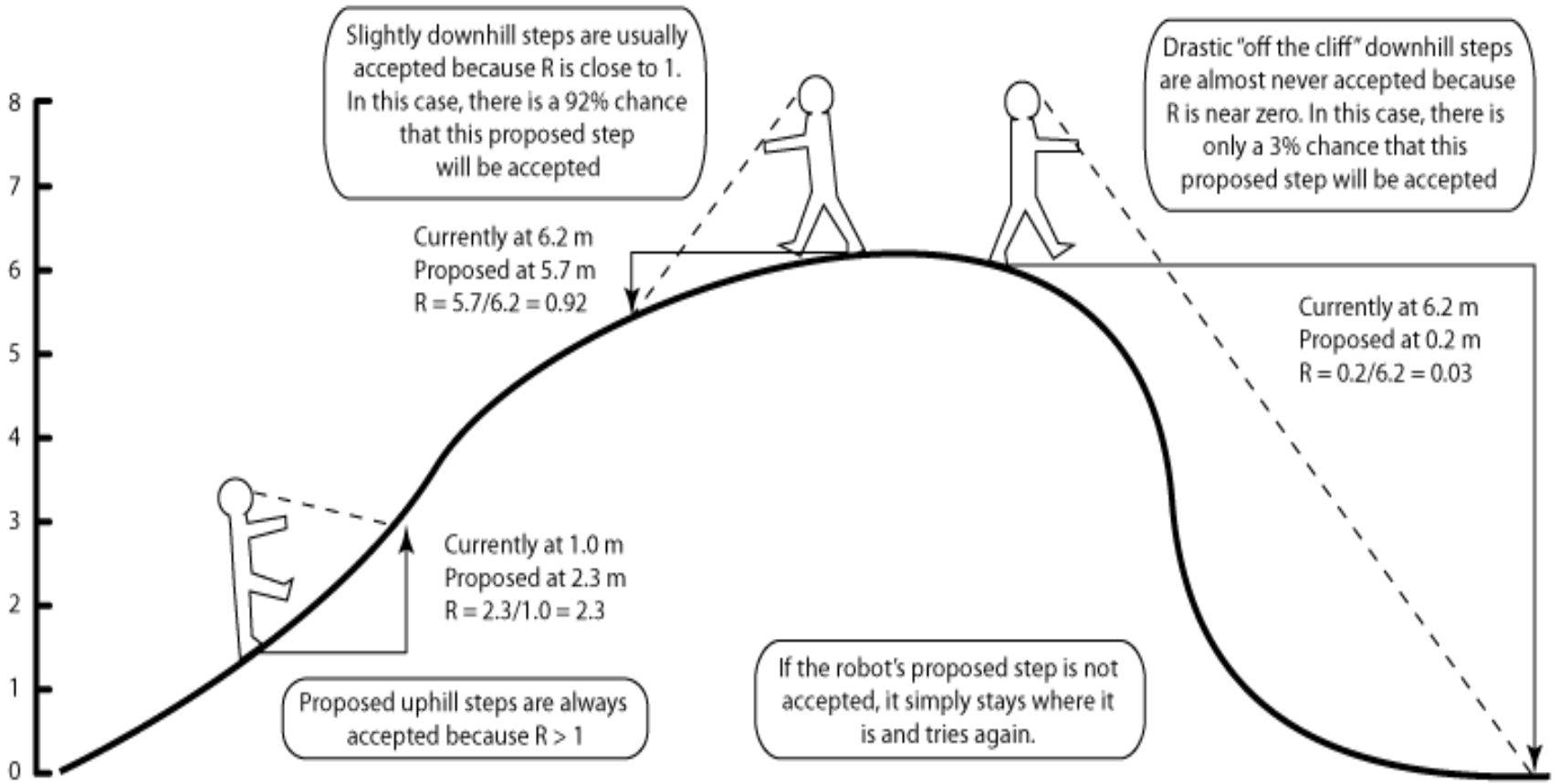
How to find the solution

There's no analytical solution for a Bayesian system. However, giving:

- **Data**: Sequence data,
- **Model**: The evolutionary model, base frequencies, among site rate variation parameters, a tree topology, branch lengths
- **Priors** distribution on the model parameters, and
- **A method** for calculating posterior distribution from prior distribution and data: **MCMC** technique[32]






Slightly downhill steps are usually accepted because R is close to 1. In this case, there is a 92% chance that this proposed step will be accepted
Drastic "off the cliff" downhill steps are almost never accepted because R is near zero. In this case, there is only a 3% chance that this proposed step will be accepted
Currently at 6.2 m
Proposed at 5.7 m
R = 5.7/6.2 = 0.92
Currently at 6.2 m
Proposed at 0.2 m
R = 0.2/6.2 = 0.03
Currently at 1.0 m
Proposed at 2.3 m
R = 2.3/1.0 = 2.3
Proposed uphill steps are always accepted because R > 1
If the robot's proposed step is not accepted, it simply stays where it is and tries again.
0
1
2
3
4
5
6
7
8

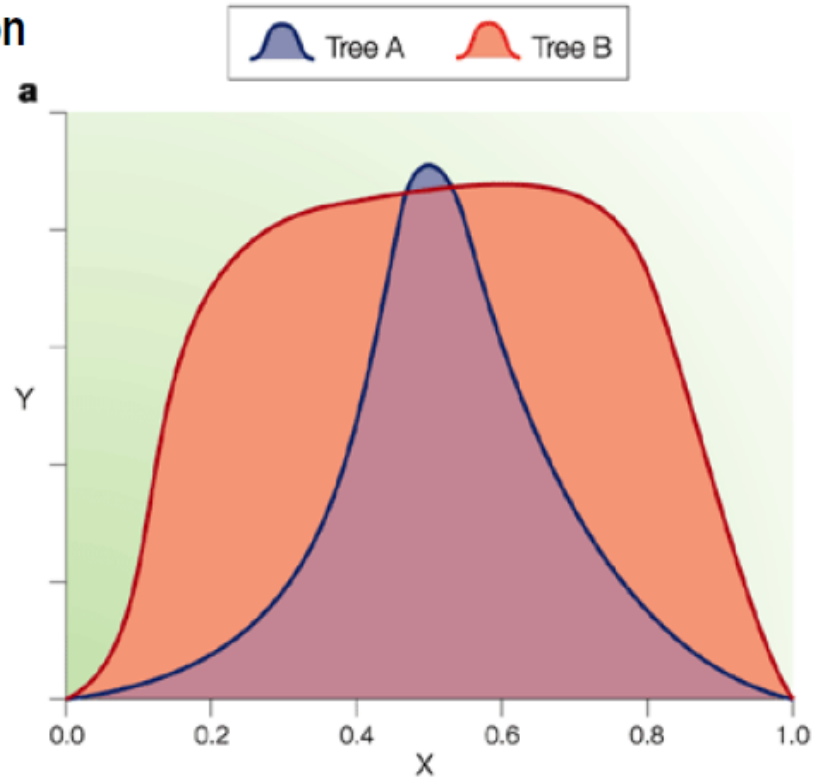
A fundamental difference:

ML commonly uses **Joint Estimation**: finding the highest point in the parameter landscape

Bayesian analyses measure the volume under the posterior probability surface, the parameters are integrated (marginalized) to obtain the marginal posterior probability of a tree (**Marginal estimation**)

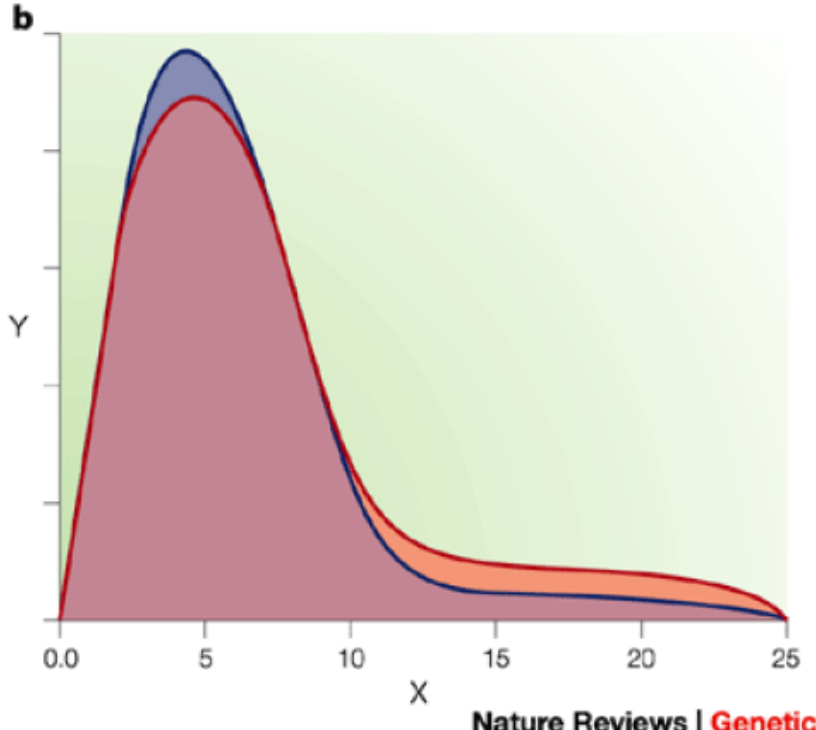
Joint versus Marginal estimation

Likelihood



ML favours A
Bayesian favours B

Every part of the surface affects the results, so the prior distributions may be seriously considered.



ML favours A
Bayesian favours B

More to come in the following days

Xiaofan Zhou

Alexey Kozlov

Stephen Crotty

Laura Kubatko

Mario Dos Reis

Tracy Heath

Barcelona Biomedical Research Park

CRG Centre de Regulació Genòmica

Orthology Part I: concepts and implications

Toni Gabaldón

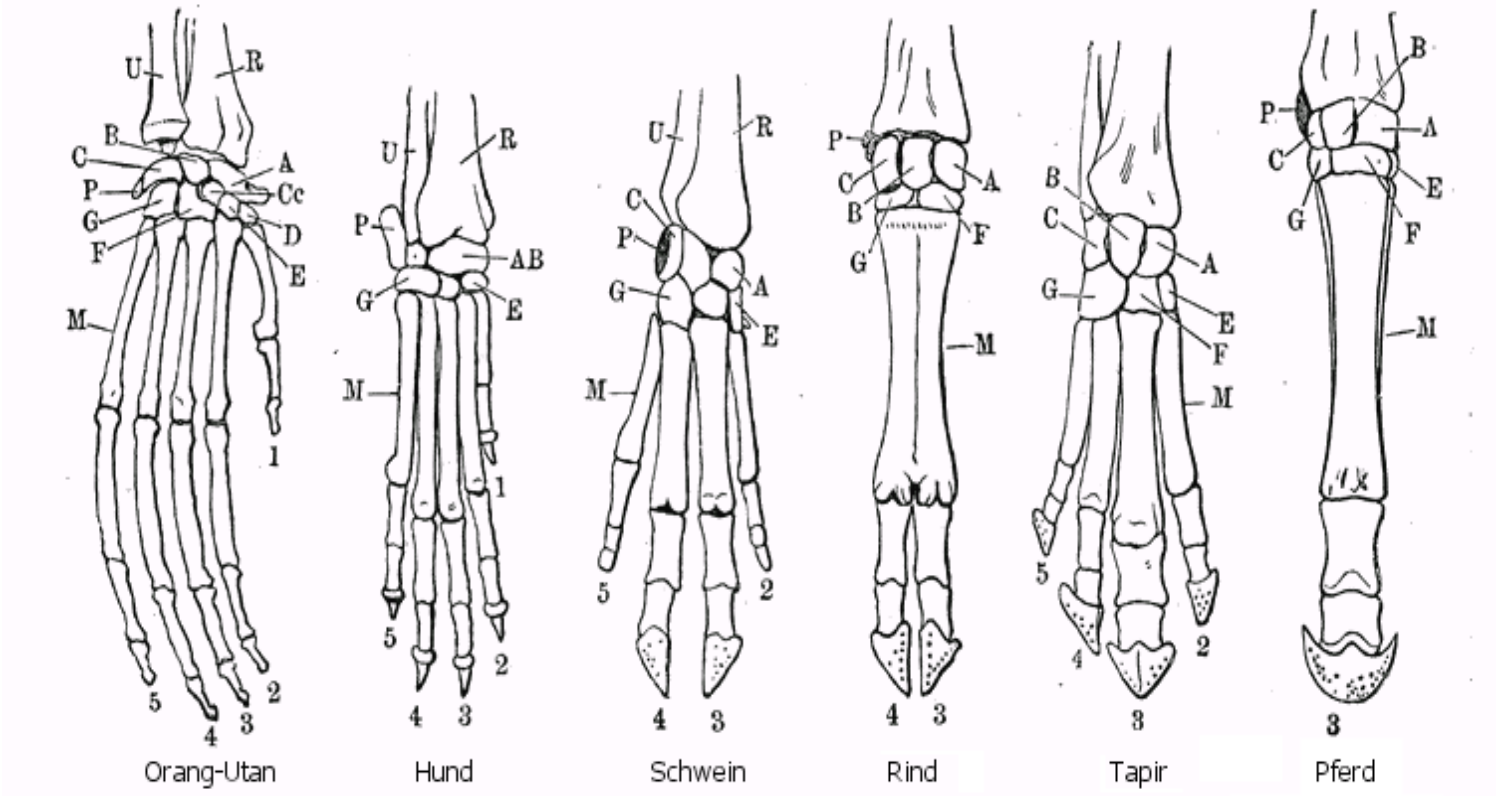
Centre for Genomic Regulation (CRG), Barcelona

(tgabaldon@crg.es)

http://gabaldonlab.crg.es

Richard Owen.

Homology

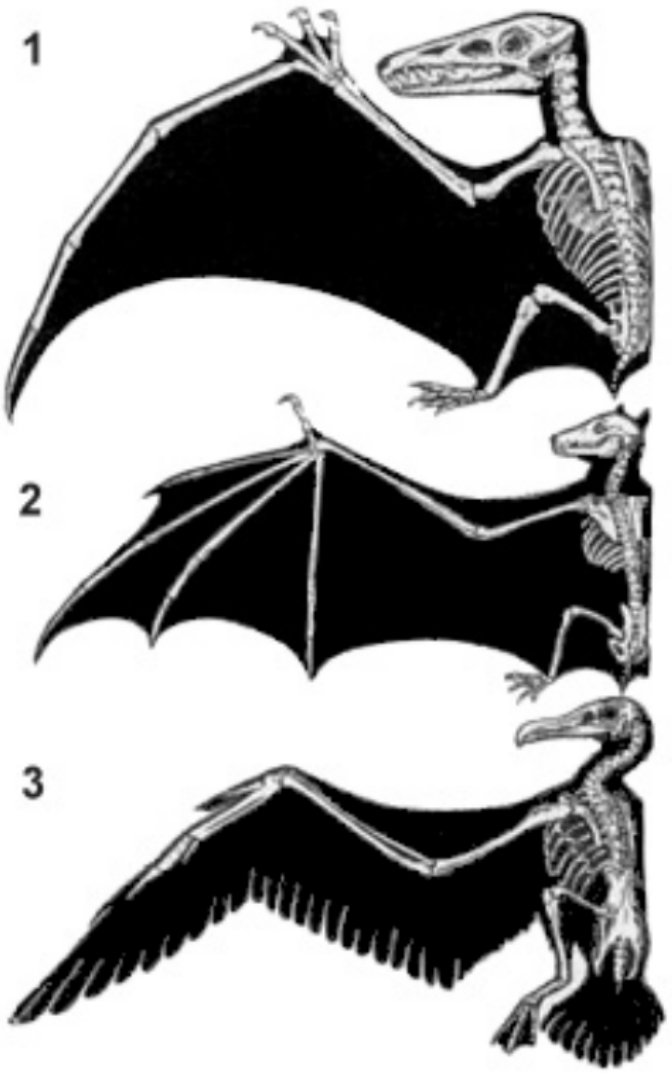


Handskelette von Säugetieren

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

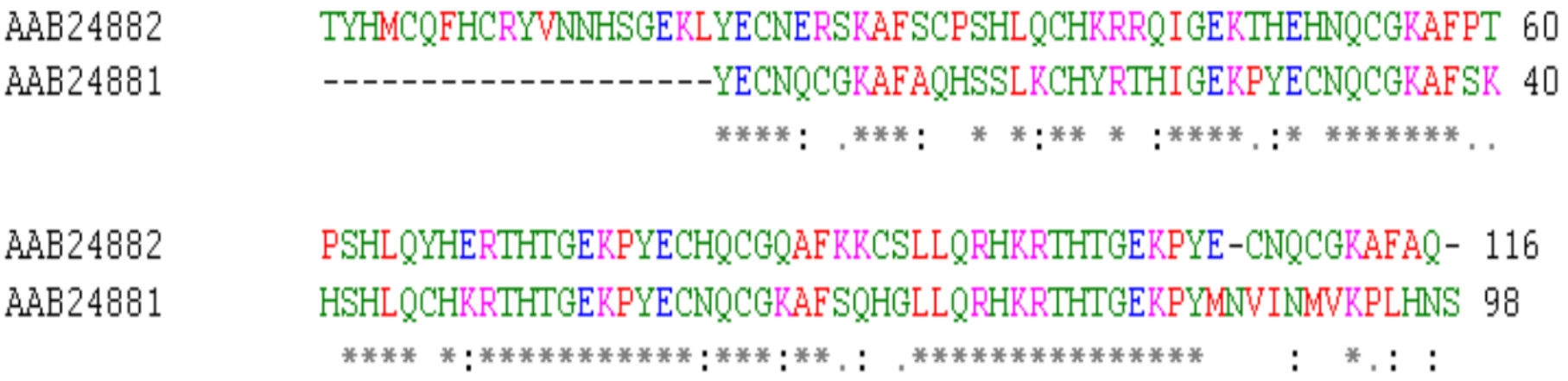
Analogous structures:
Similar function but independet origin.



Homologous as forelimbs
But
Analogous as wings

Extension of the concept of homology to sequences:

Two sequences are homologous if they share common ancestry



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

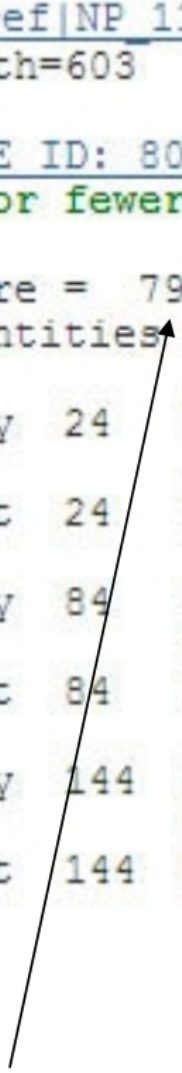
GENE ID: 803075 ND5 | NADH dehydrogenase subunit 5 [Macaca sylvanus]
(10 or fewer PubMed links)

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

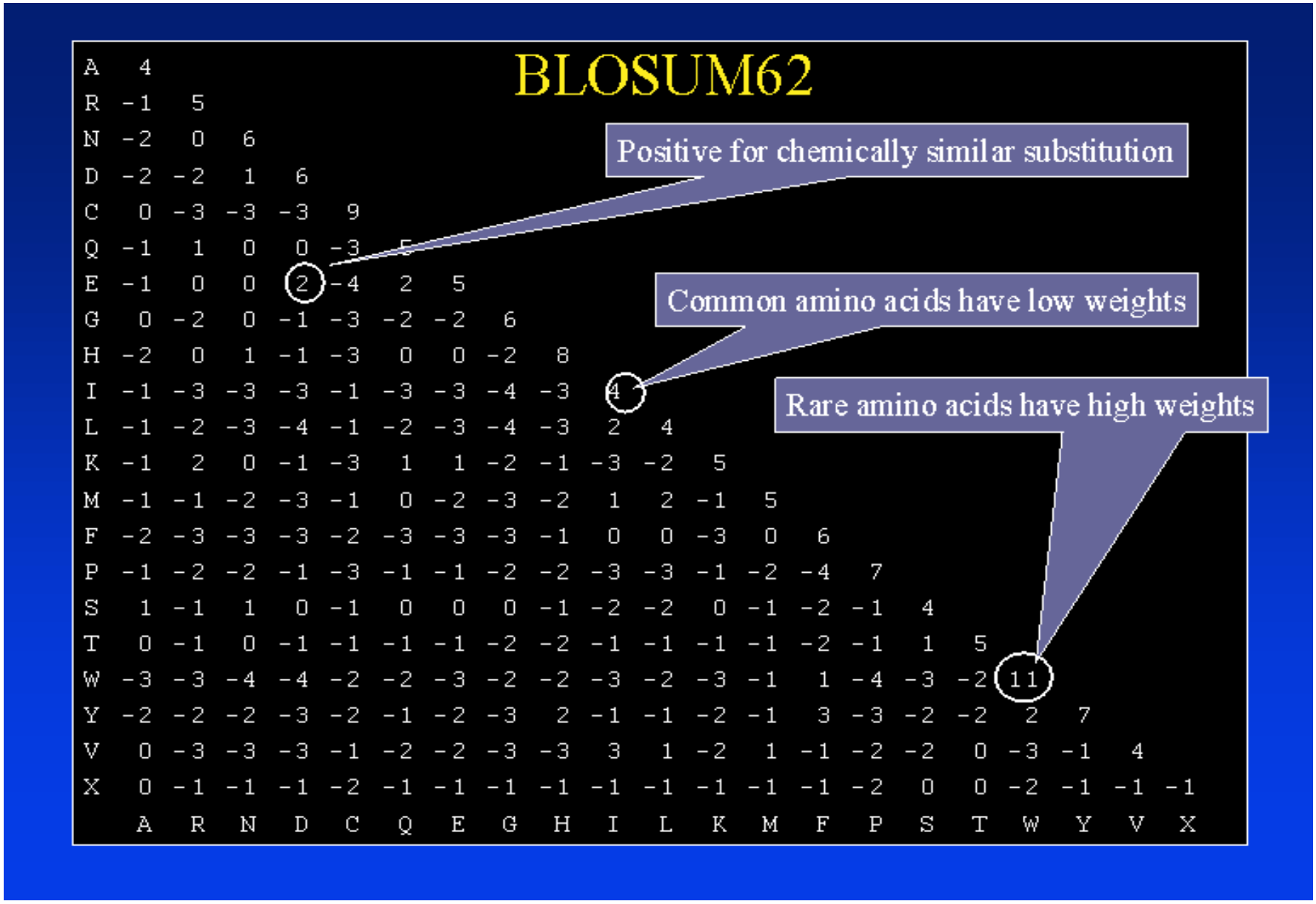
```
Query  24   VNPNKKNSYPHYVKSIVASTFIISLFPTTMFMCLDQEVIISNWHWATTQTTQLSLSFKLD  83
            +NPNKK+ YP+YVK+ V   FI SL  TT++M L+QE II +WHW  TQT  L+LSFKLD
Sbjct  24   INPNKKHLYPNYVKTAVMYAFITSLSSTTLYMFLNQETIIWSWHWMMTQTLSLTLSFKLD  83

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
```

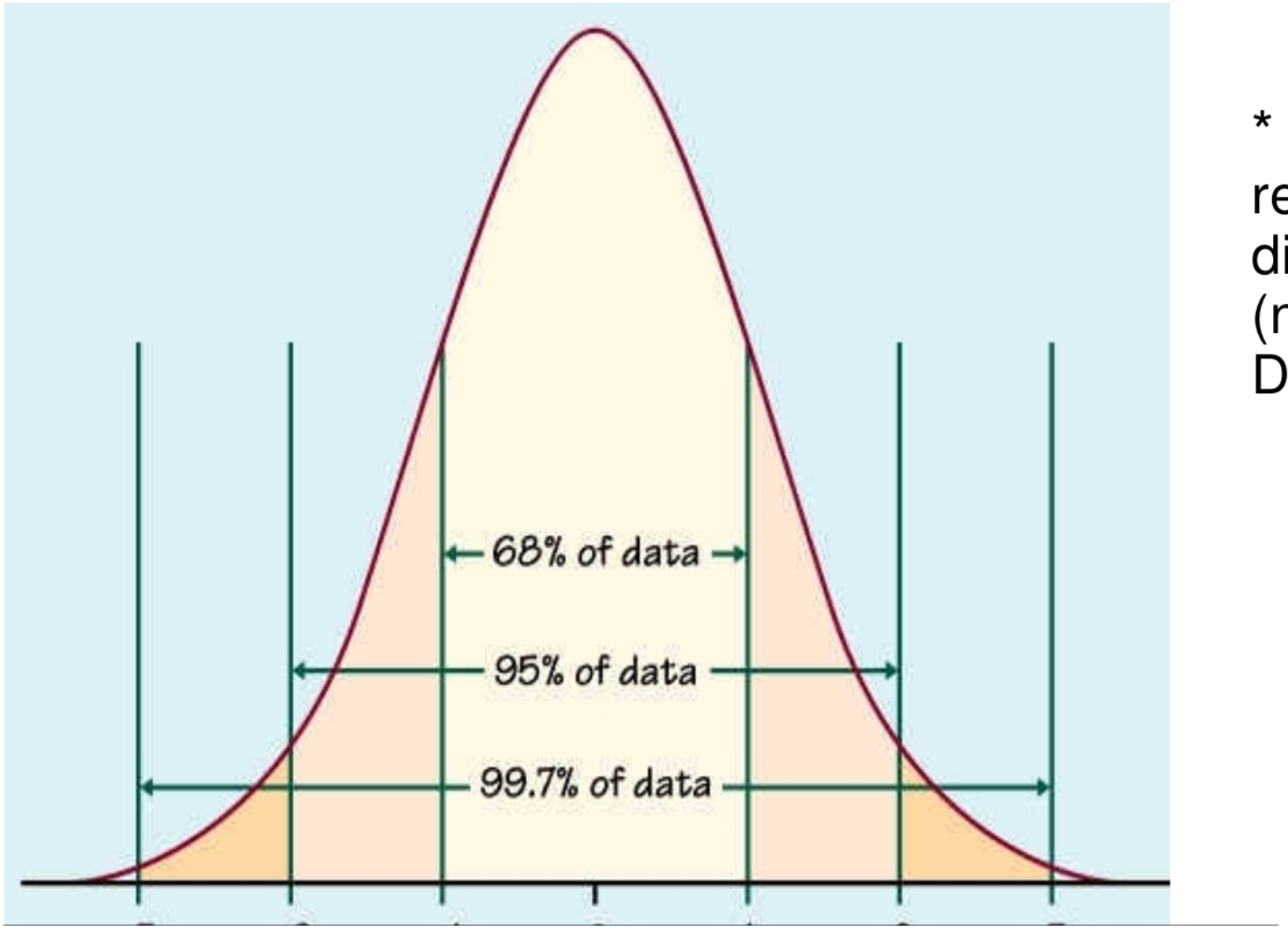


Alignment scores are sums of residue pairing scores according to a scoring Matrix

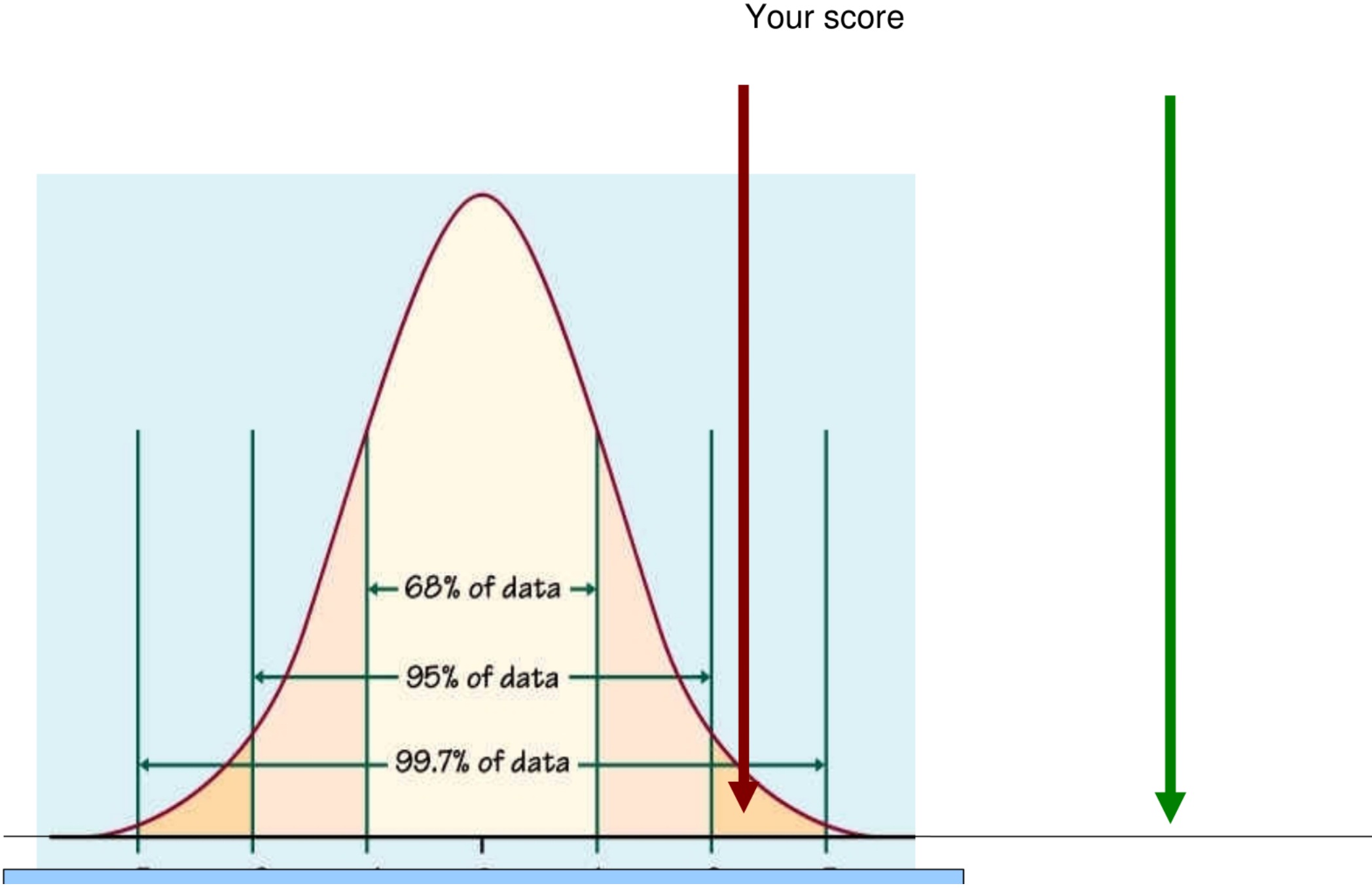


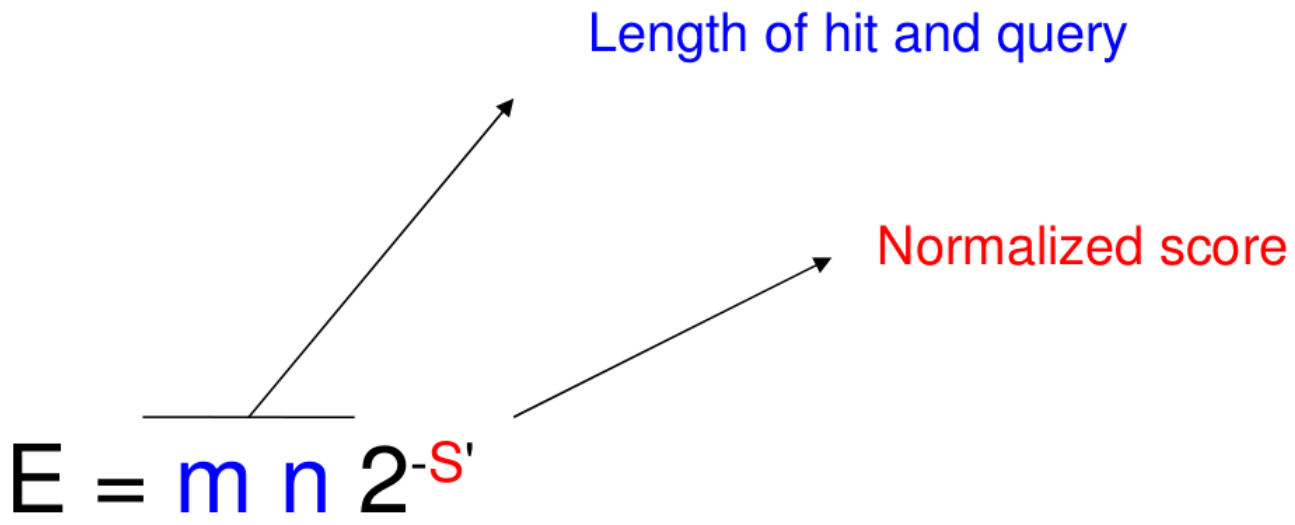
	A	R	N	D	C	Q	E	G	H	I	L	K	M	F	P	S	T	W	Y	V	X
A	4																				
R	-1	5																			
N	-2	0	6																		
D	-2	-2	1	6																	
C	0	-3	-3	-3	9																
Q	-1	1	0	0	-3	5															
E	-1	0	0	2	-4	2	5														
G	0	-2	0	-1	-3	-2	-2	6													
H	-2	0	1	-1	-3	0	0	-2	8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

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



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


Your score
68% of data
95% of data
99.7% of data



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| 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

From homology to orthology

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

- What are then orthologs?.... and paralogs?

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.

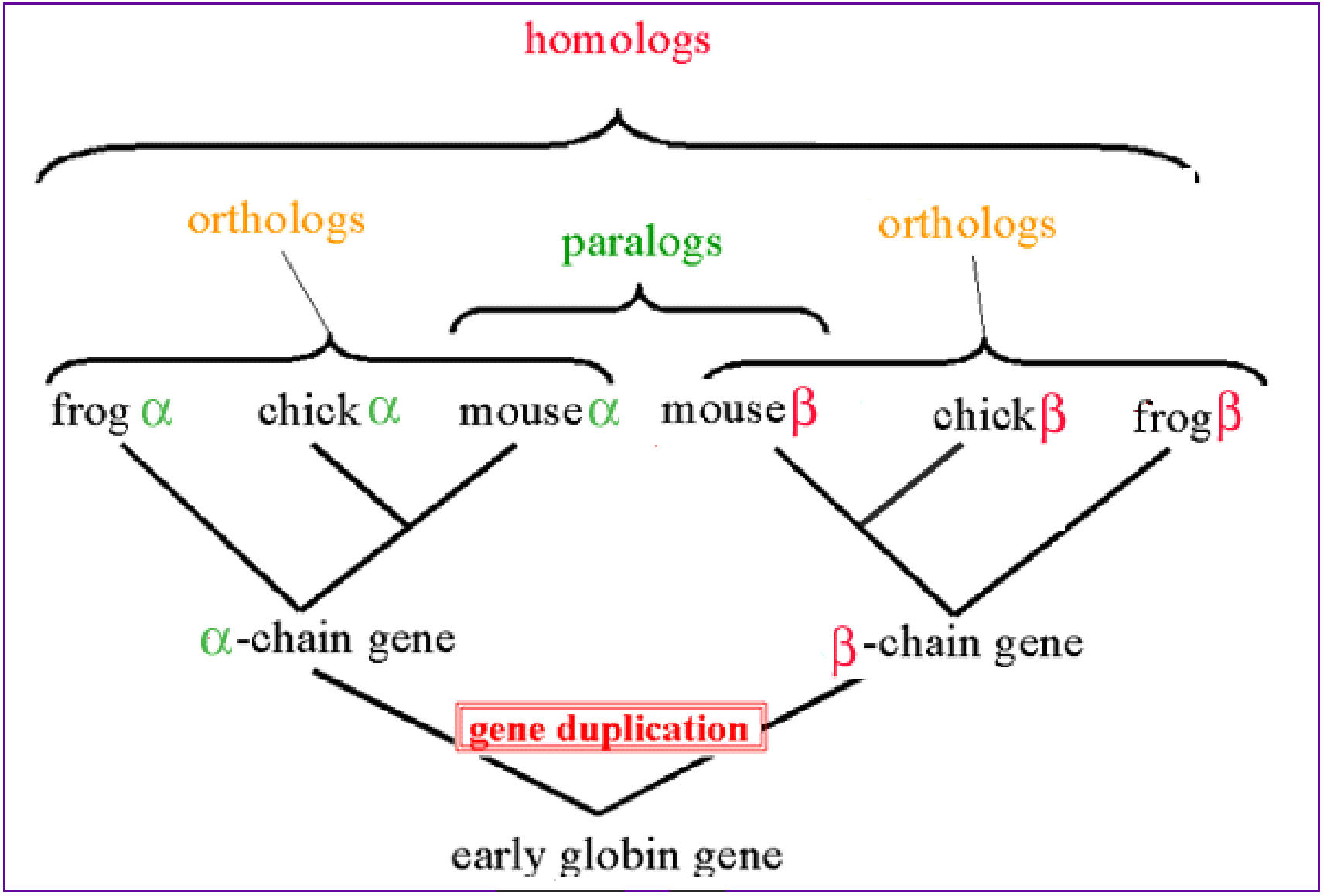
Distinguishing homologous from analogous proteins.

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).*

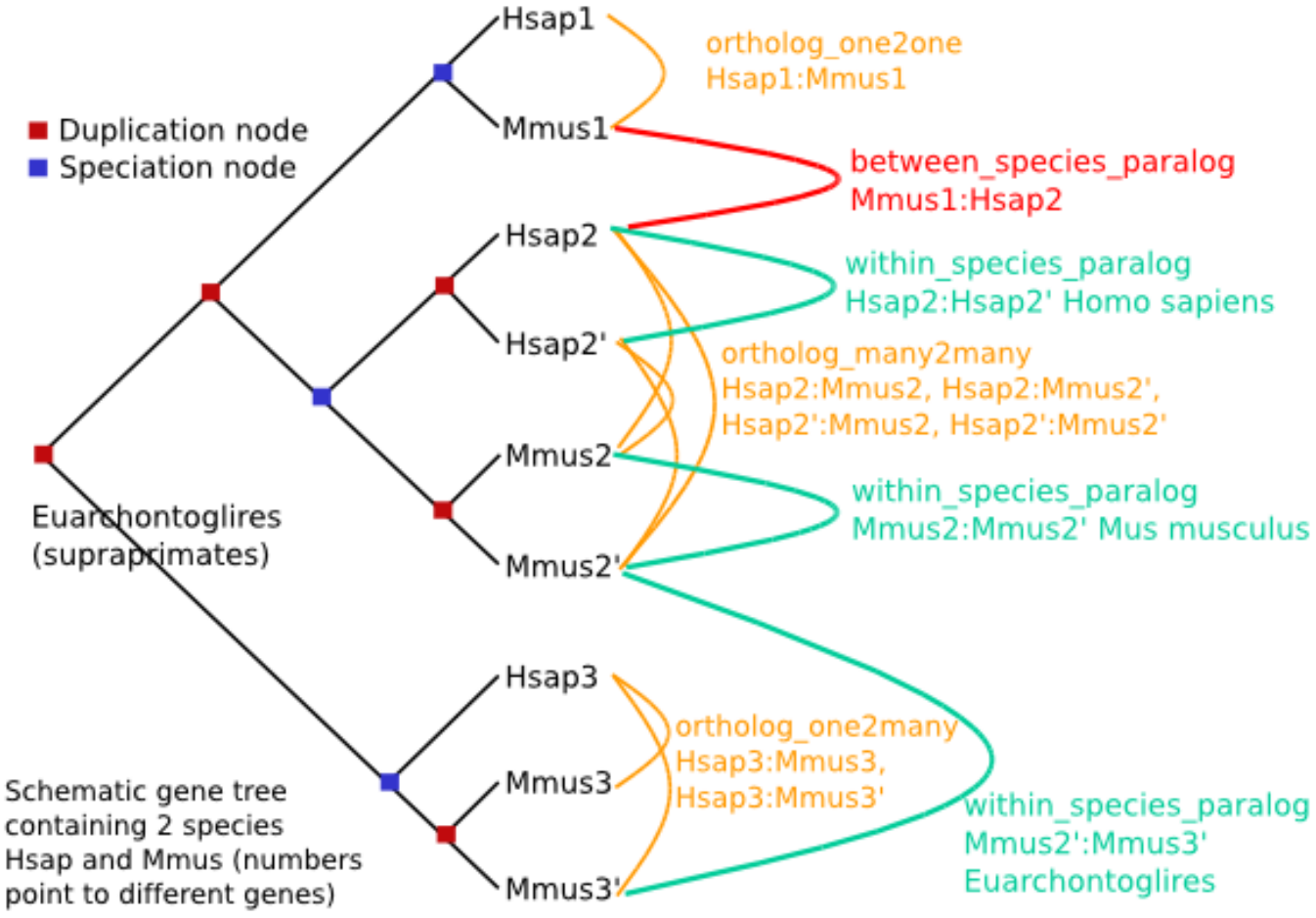
*Where the homology is **the result of speciation** so that the history of the gene reflects the history of the species (for example alpha hemoglobin in man and mouse) the genes should be called **orthologous** (ortho = exact)."*


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)

Orthology relationships can be complex, and intricate



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)

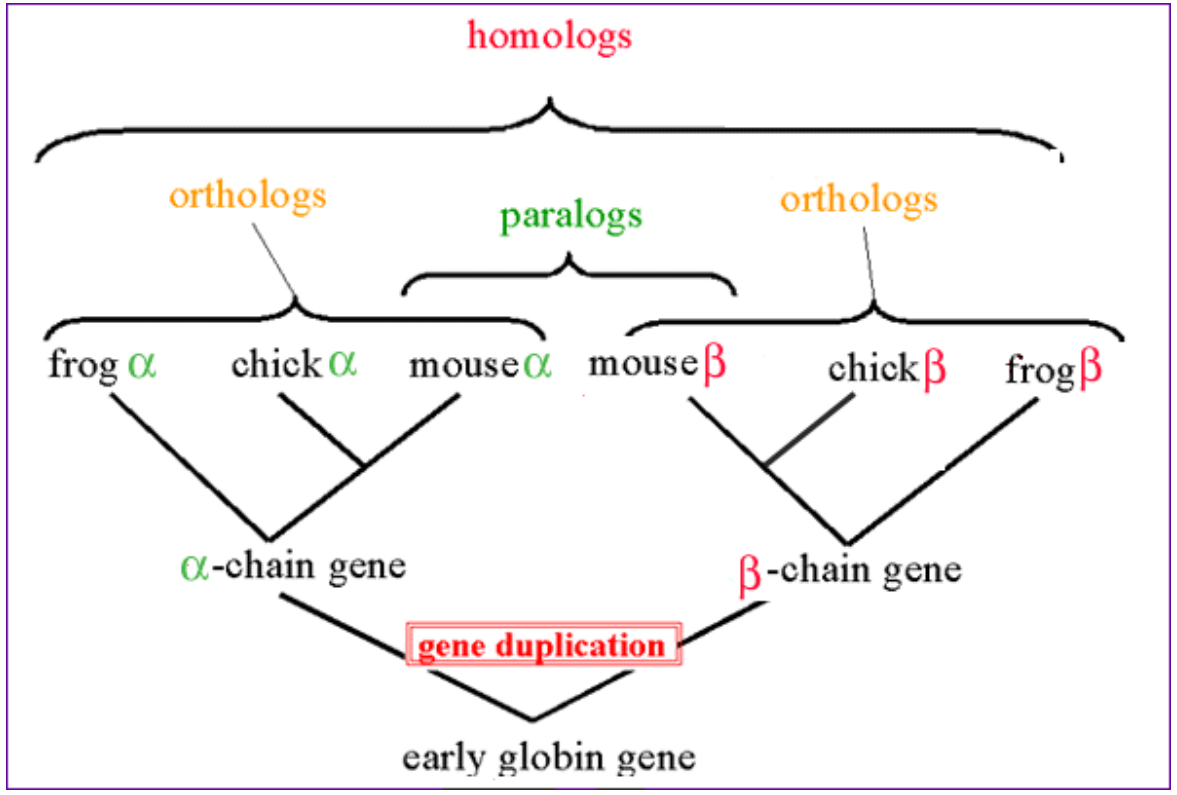
- The most exact way of **comparing two (or more) genomes** in terms of their gene content. Necessary to uncover how genomes evolve.

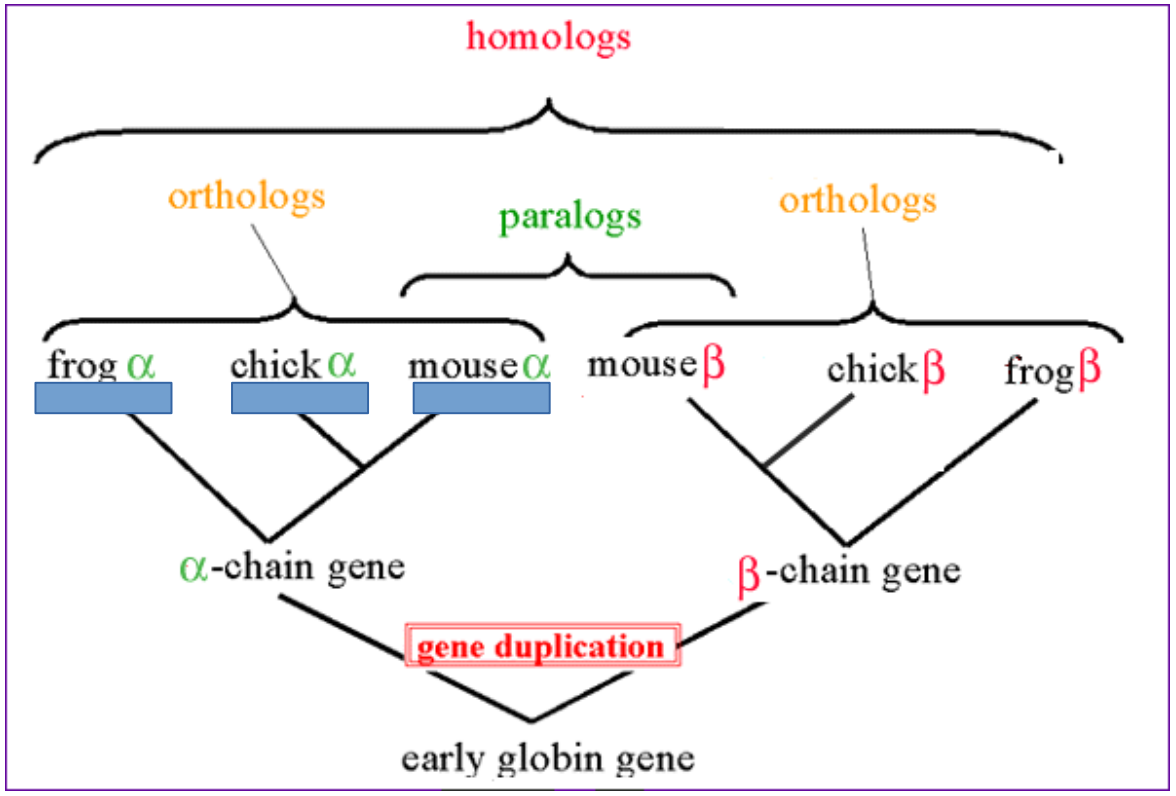
- Implications for **functional inference**: orthologs, as compared to paralogs, are more likely to share the same function

Why predicting orthology is important?

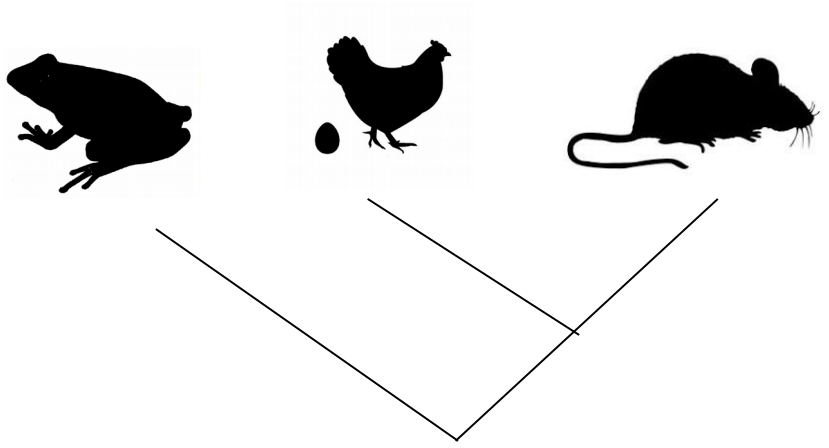
- **Important implications for phylogeny:** only sets of orthologous genes are expected to reflect the underlying species evolution (although there are many exceptions)

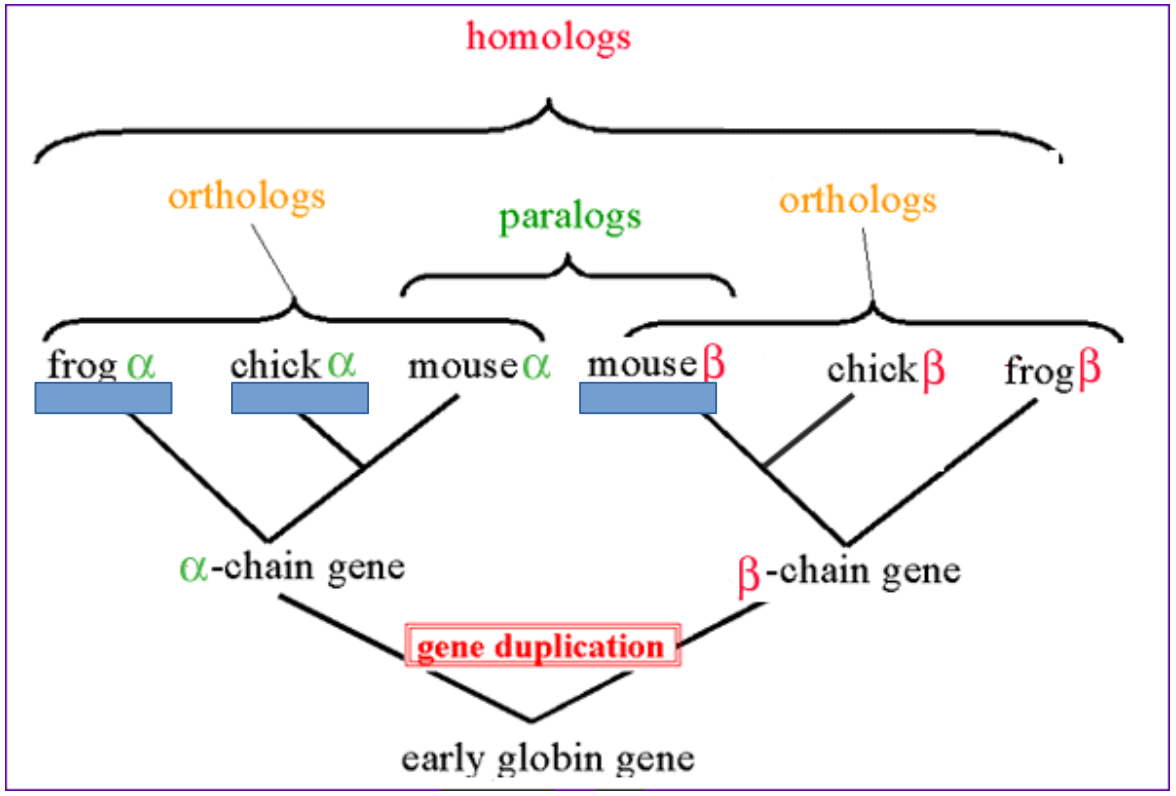
*Where the homology is **the result of speciation** so that the history of the gene reflects the history of the species (for example alpha hemoglobin in man and mouse) the genes should be called **orthologous** (ortho = exact).*


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

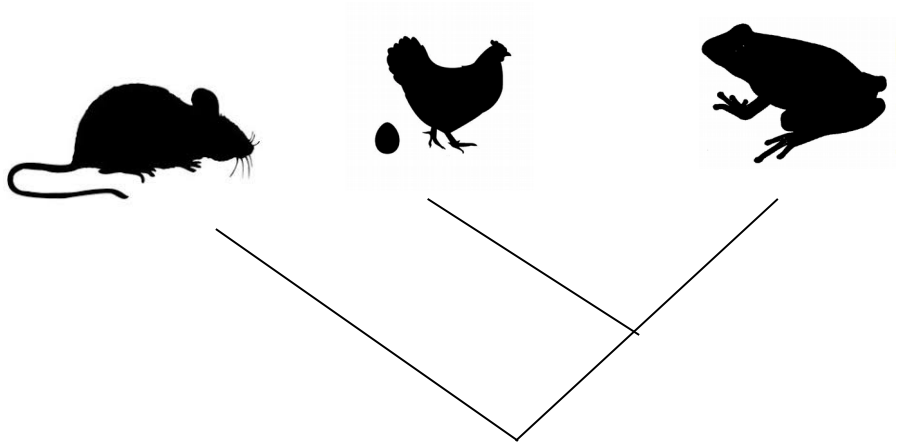


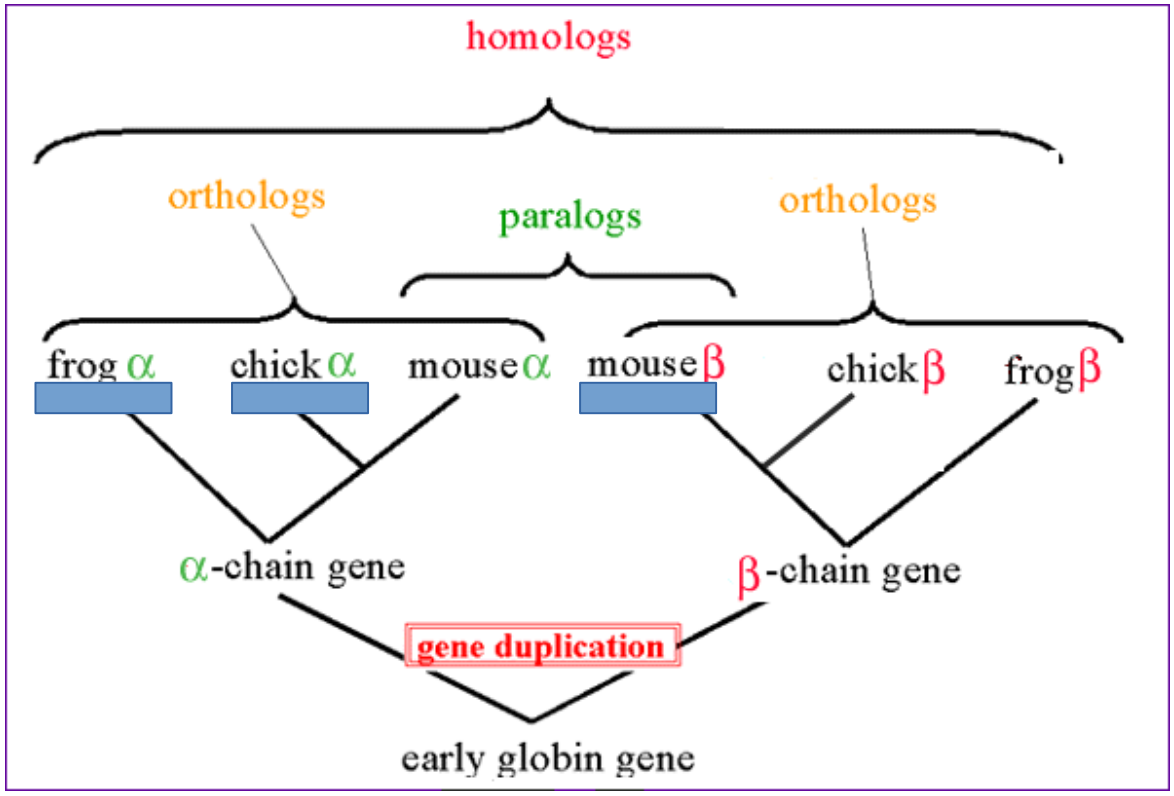
Tree from orthologous dataset:



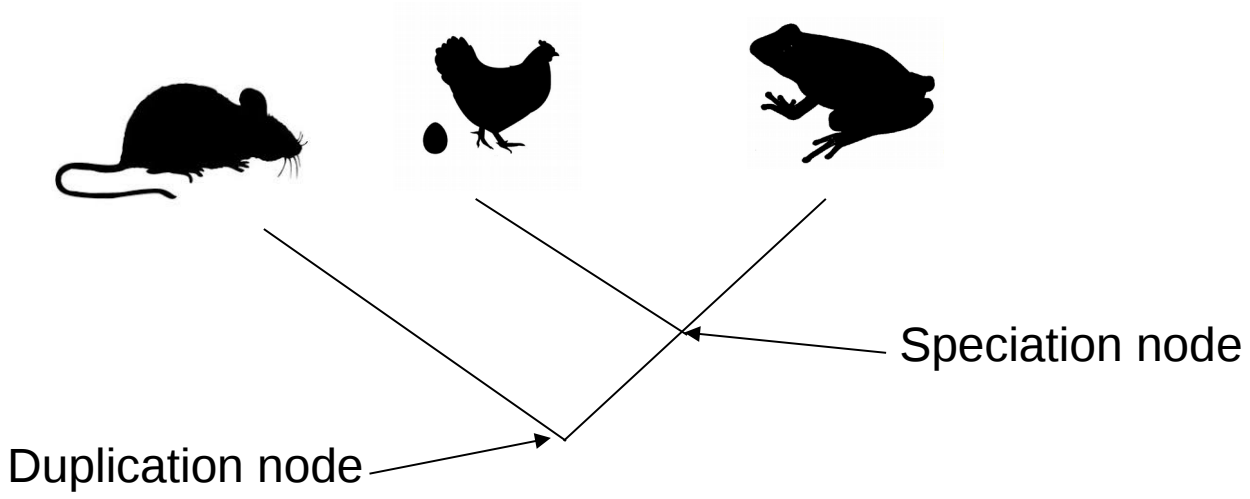


Tree from non-orthologous dataset:





Tree from non-orthologous dataset: NOT a SPECIES TREE



Seems easy, but it’s not*

*at least not always.

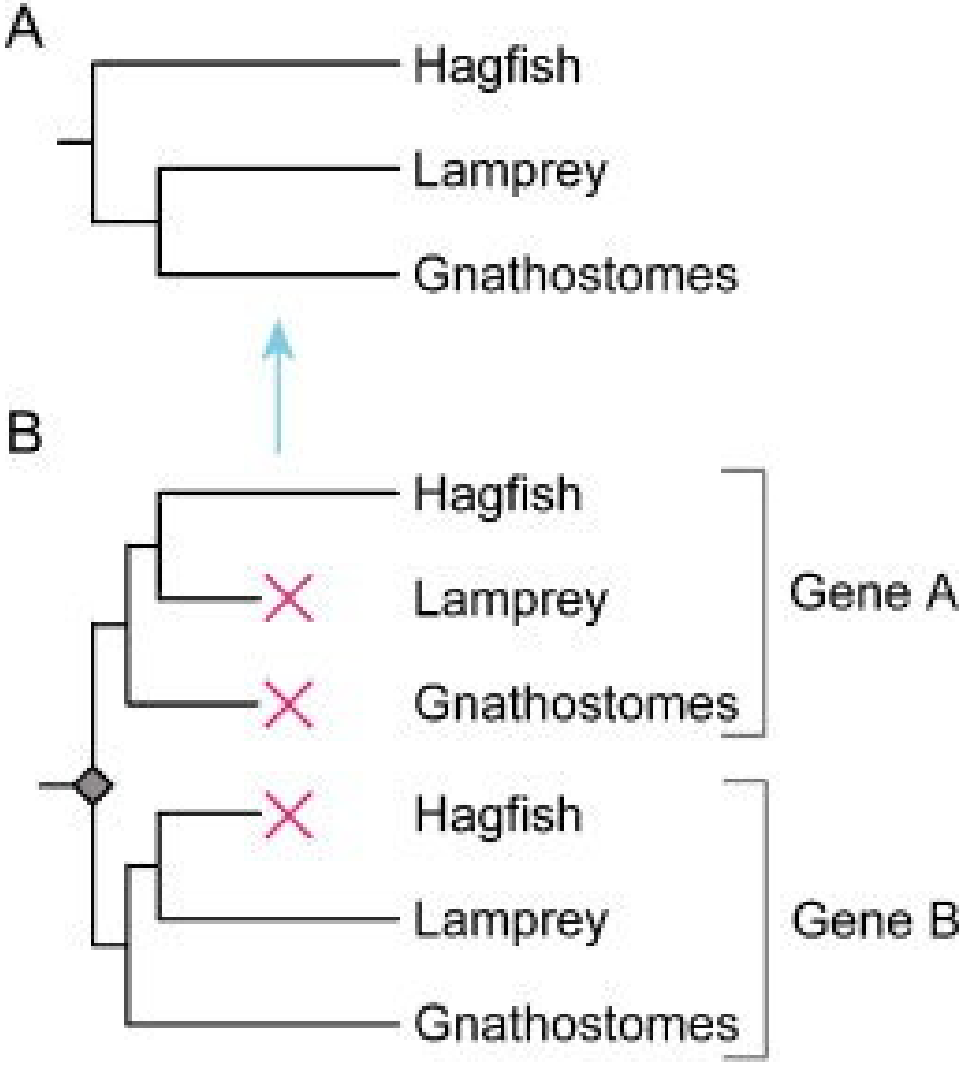
The hidden paralogy problem

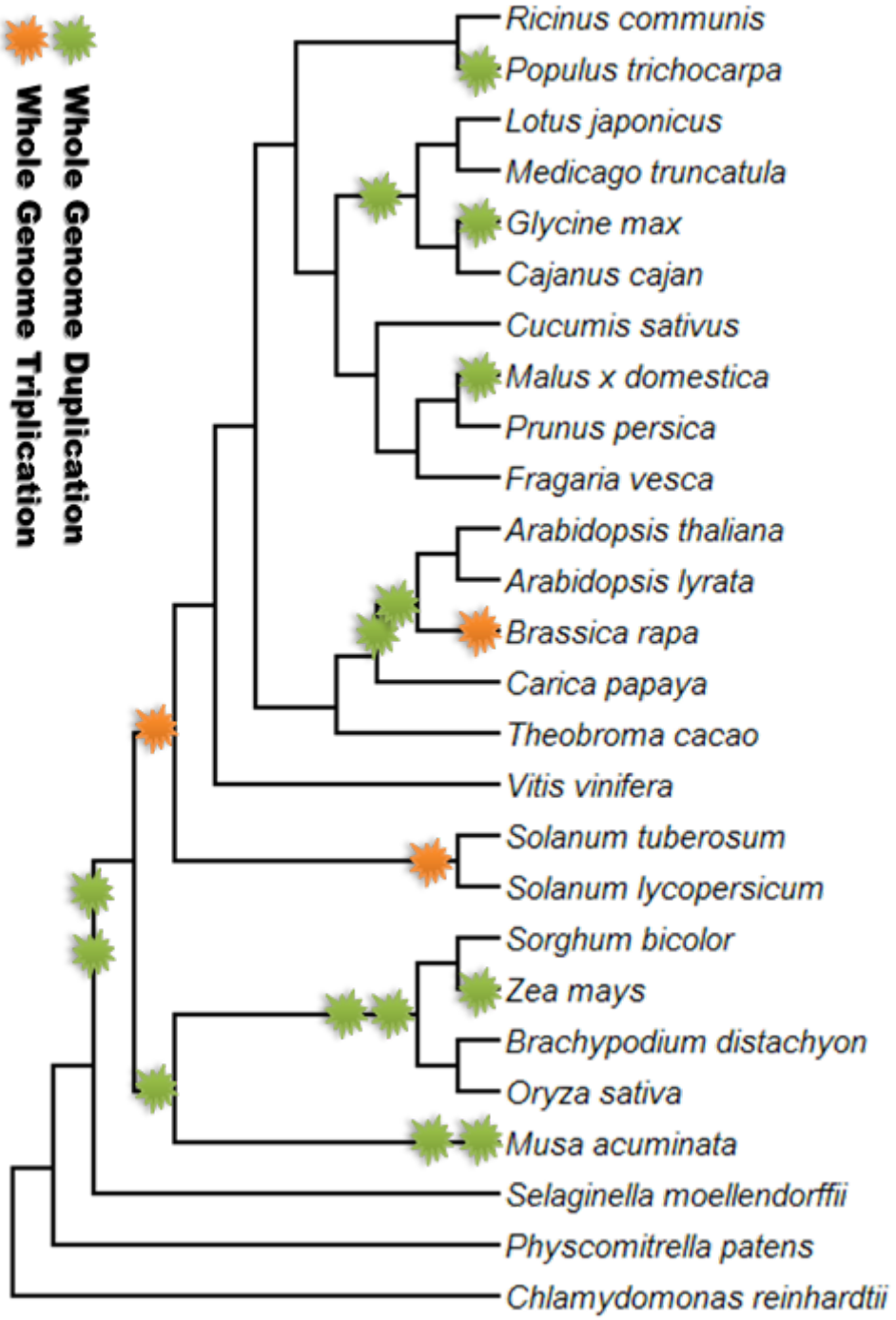
Complex duplications and loss patterns can result in paralogous genes being recovered as putative orthologs by most methods, resulting in faulty phylogenetic relationships.

This problem is exacerbated following whole genome duplication events, usually followed by massive differential gene loss.

Annotation problems and fragmented genomes could result in similar patterns.

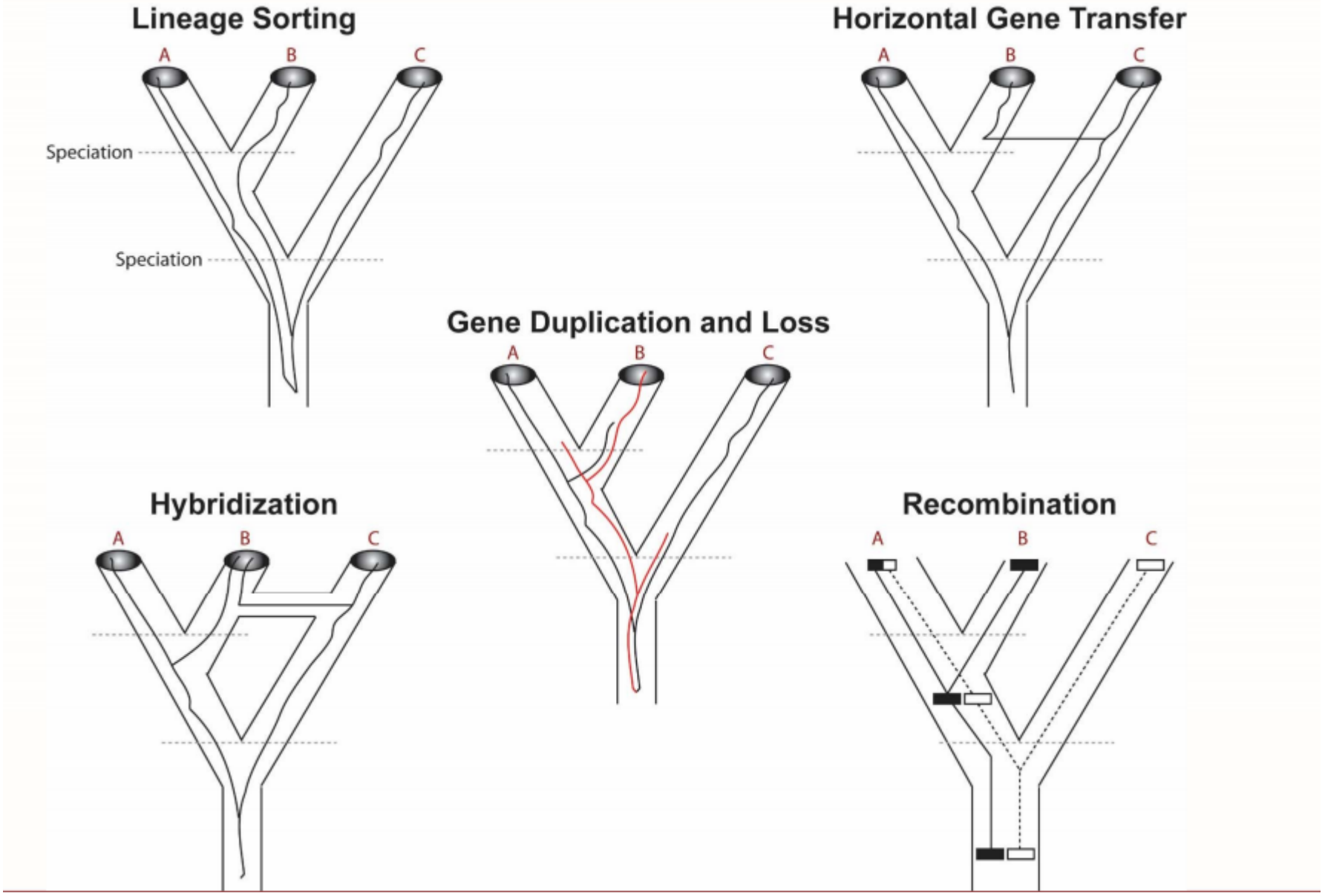
Increasing taxonomic coverage is one approach to aleviate this problem .



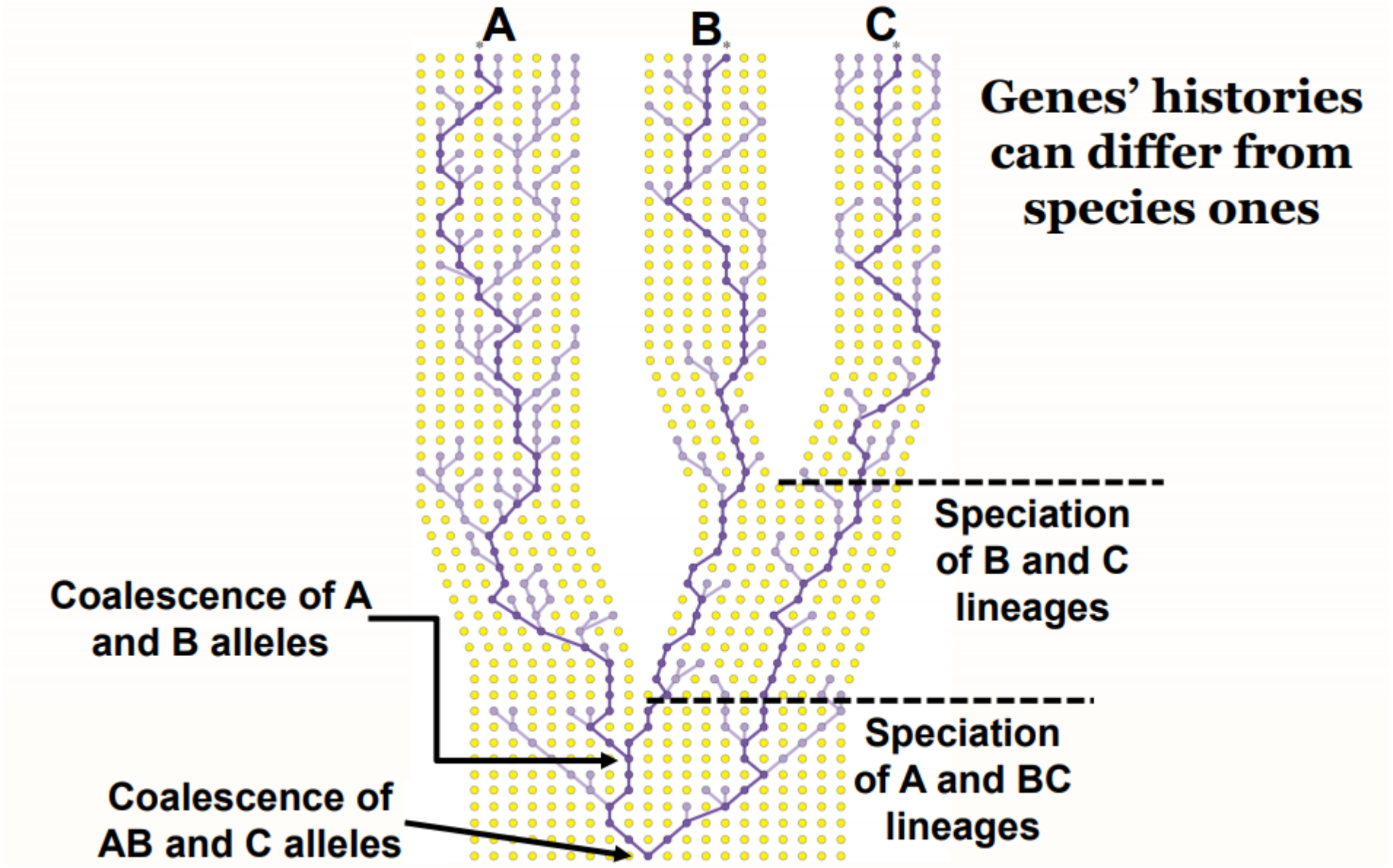

Whole Genome Duplication
Whole Genome Triplication
Ricinus communis
Populus trichocarpa
Lotus japonicus
Medicago truncatula
Glycine max
Cajanus cajan
Cucumis sativus
Malus x domestica
Prunus persica
Fragaria vesca
Arabidopsis thaliana
Arabidopsis lyrata
Brassica rapa
Carica papaya
Theobroma cacao
Vitis vinifera
Solanum tuberosum
Solanum lycopersicum
Sorghum bicolor
Zea mays
Brachypodium distachyon
Oryza sativa
Musa acuminata
Selaginella moellendorffii
Physcomitrella patens
Chlamydomonas reinhardtii

Even if you get a truly orthologous dataset.

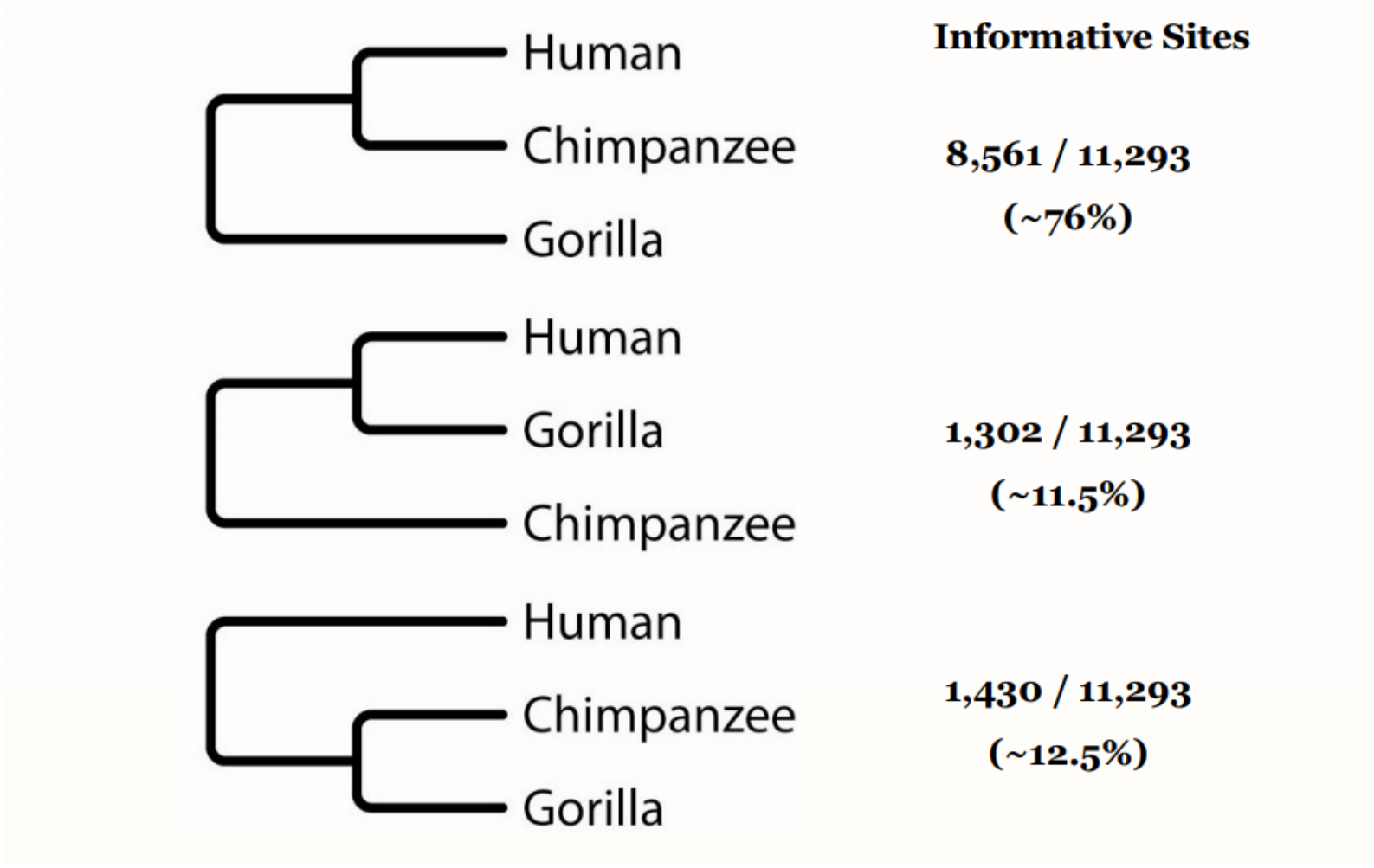
Orthologus genes are not guaranteed to reflect the species tree!


Lineage Sorting
A
B
C
Speciation
Speciation
Horizontal Gene Transfer
A
B
C
Gene Duplication and Loss
A
B
C
Hybridization
A
B
C
Recombination
A
B
C

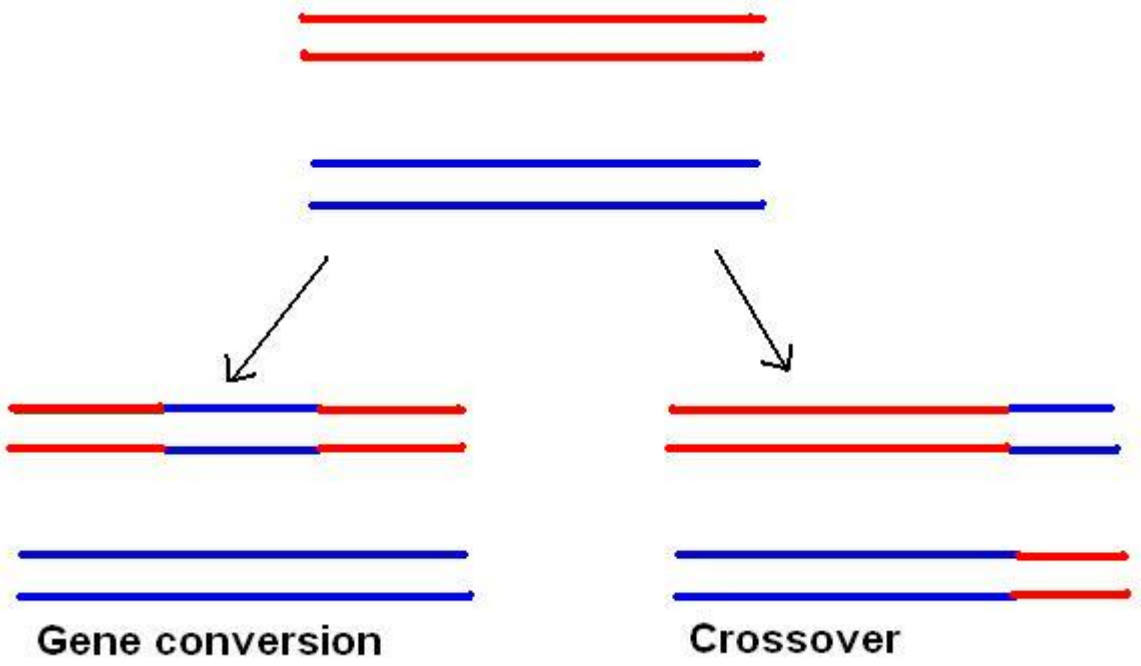
Incomplete lineage sorting

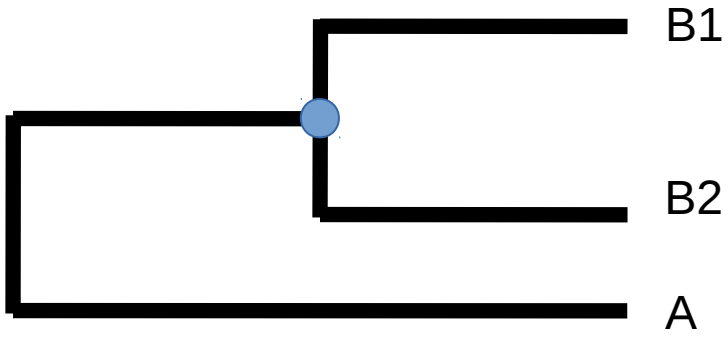


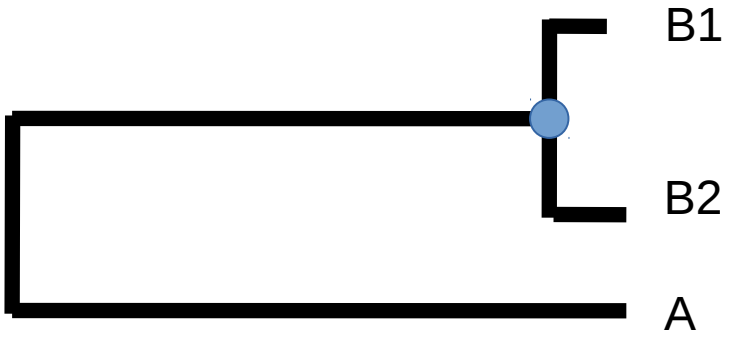
Incomplete lineage sorting in the primate lineage:



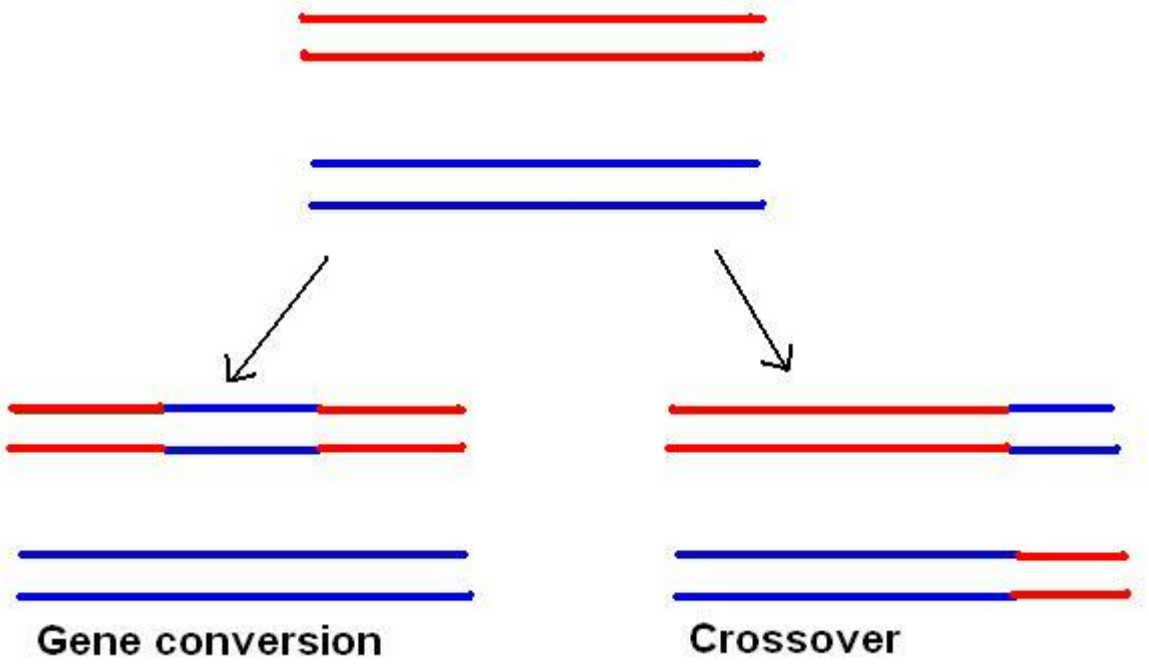
Gene conversion

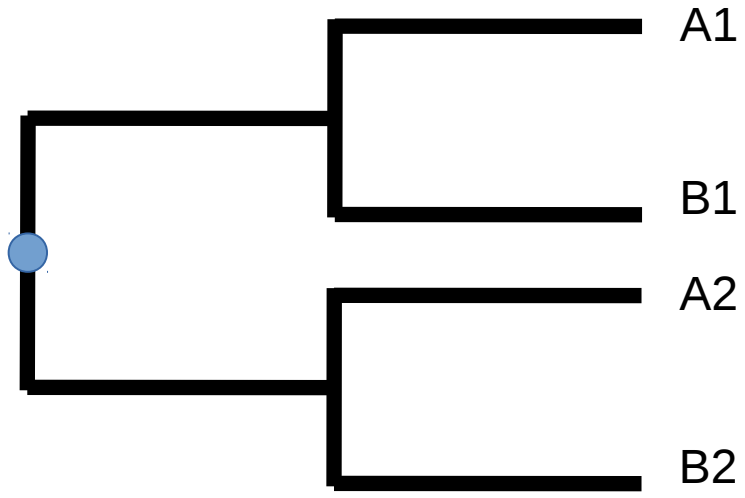


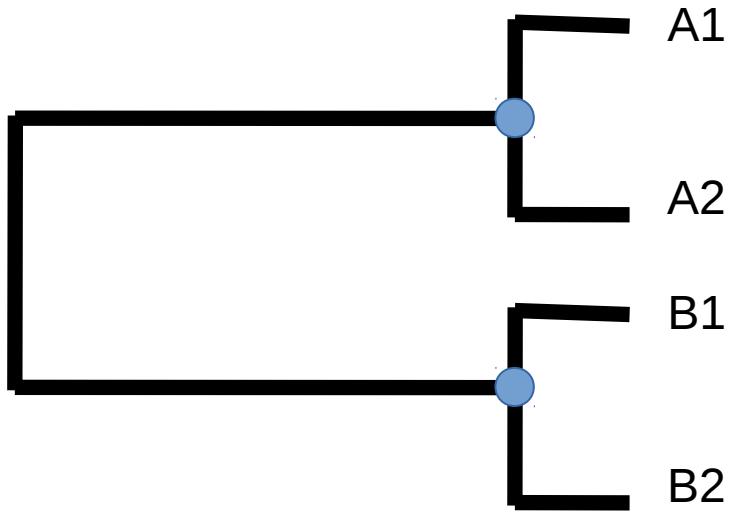




Gene conversion





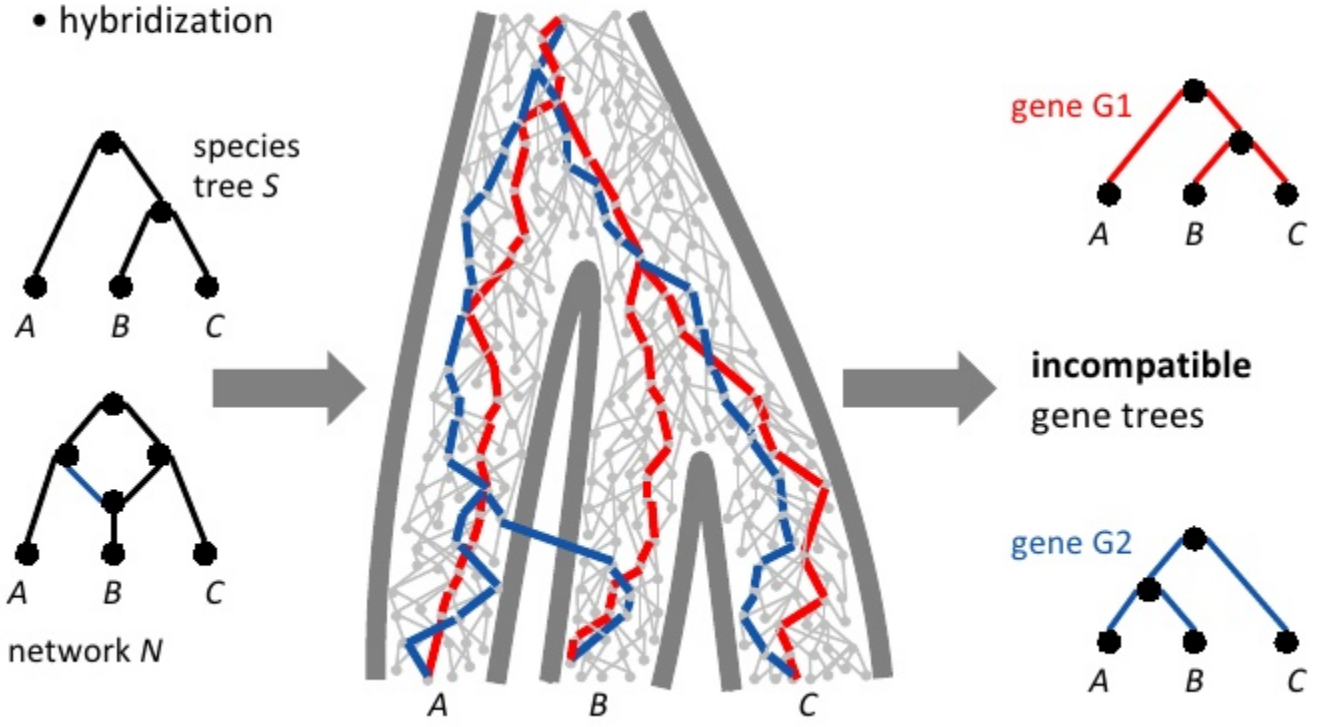


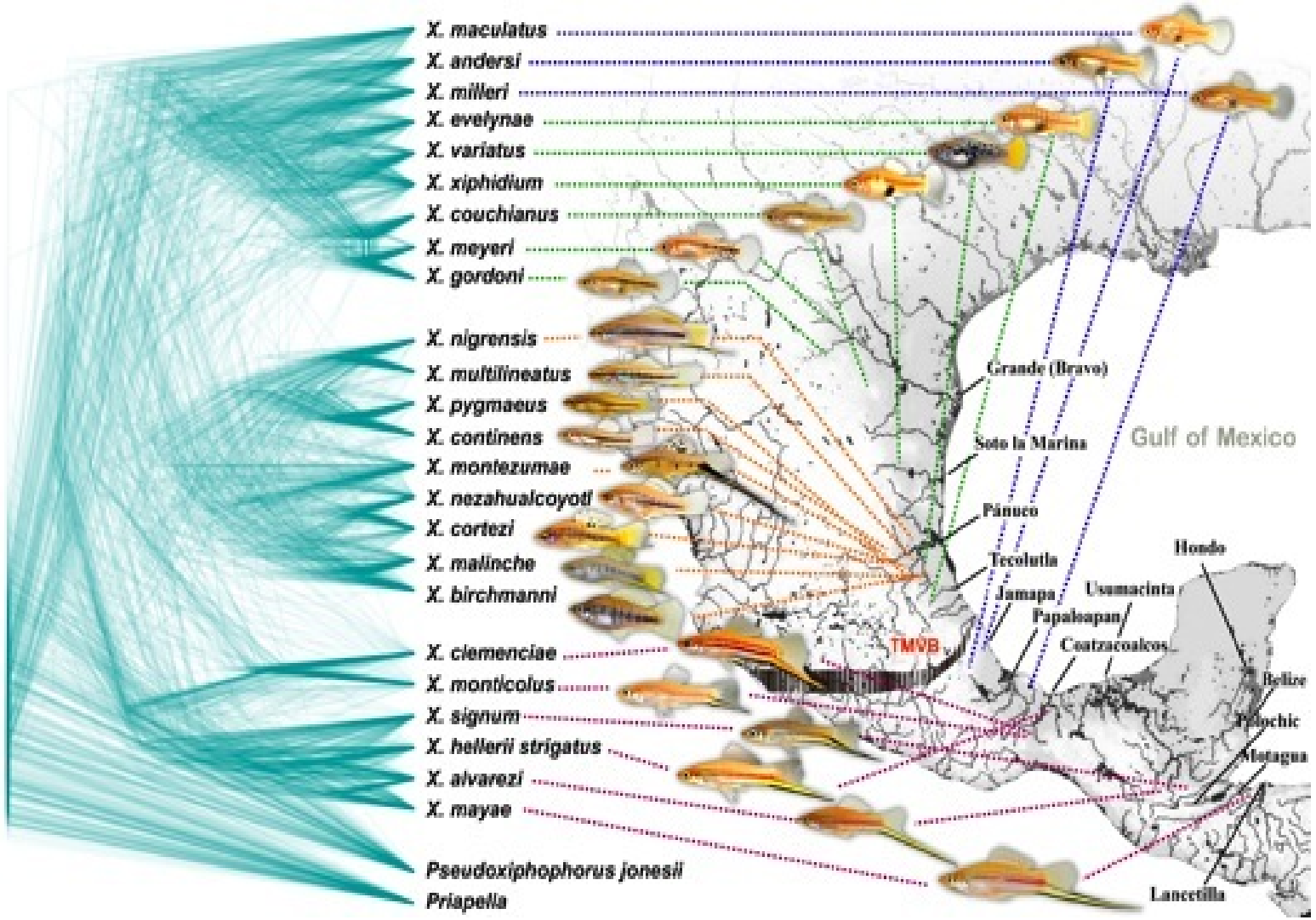
Introgression / Hybrization

Genetic material transfer

Genetic material transfers between coexisting species:

- horizontal gene transfer
- hybridization




X. maculatus
X. andersi
X. milleri
X. evelynae
X. variatus
X. xiphidium
X. couchianus
X. meyeri
X. gordoni
X. nigrensis
X. multilineatus
X. pygmaeus
X. continens
X. montezumae
X. nezahualcoyotl
X. cortezi
X. malinche
X. birchmanni
X. clemenciae
X. monticolus
X. signum
X. hellerii strigatus
X. alvarezi
X. mayae
Pseudoxiphophorus jonesii
Priapella
Grande (Bravo)
Soto la Marina
Gulf of Mexico
Pánuco
Tecolutla
Jamapa
Papaloapan
TMVB
Coatzacoalcos
Usumacinta
Hondo
Belize
Motagua
Lancetilla

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)

Yes, orthologs are useful to retrieve species trees.

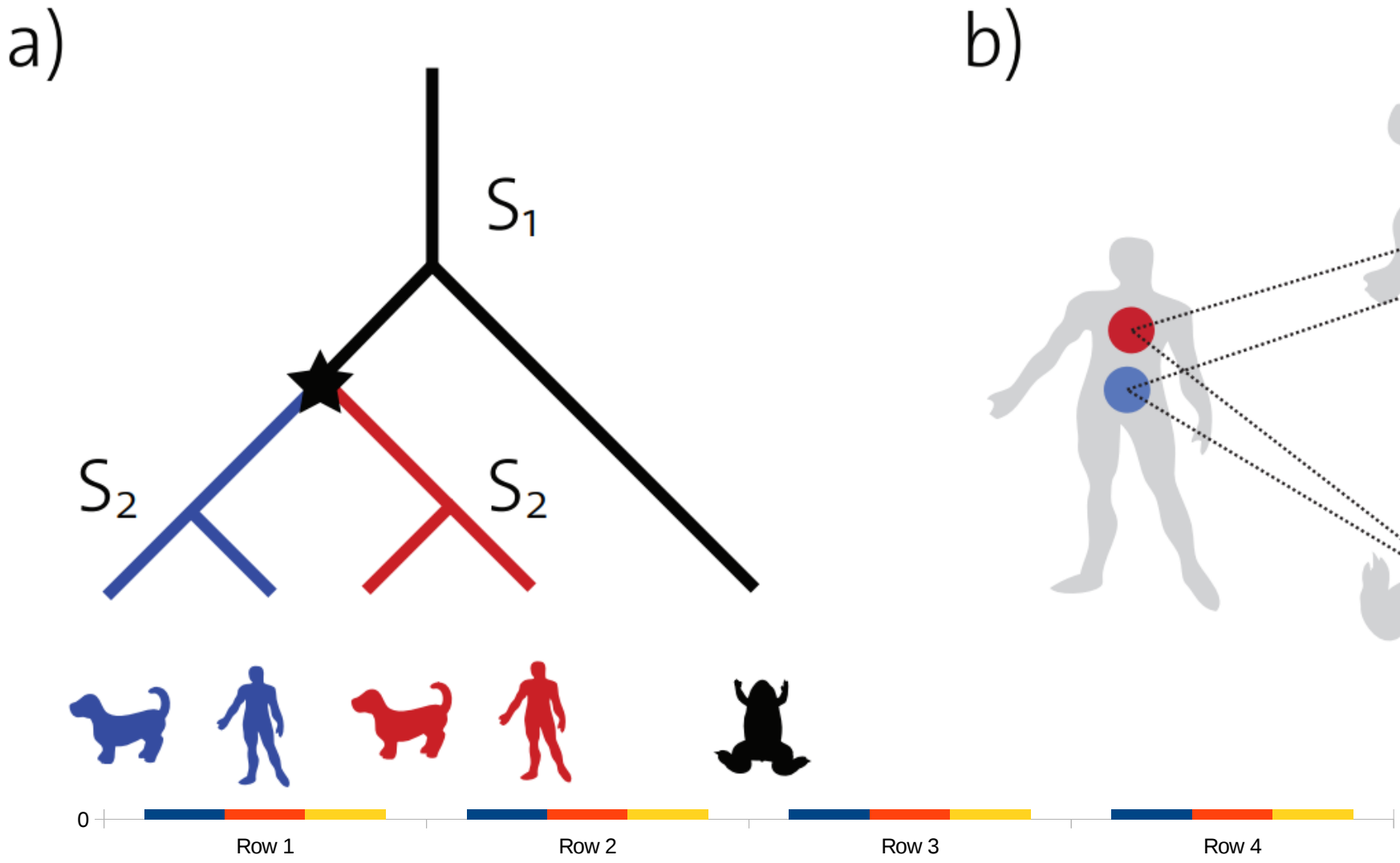
However, even if you have the orthologs, getting the species tree is not straightforward.

Why predicting orthology is important?

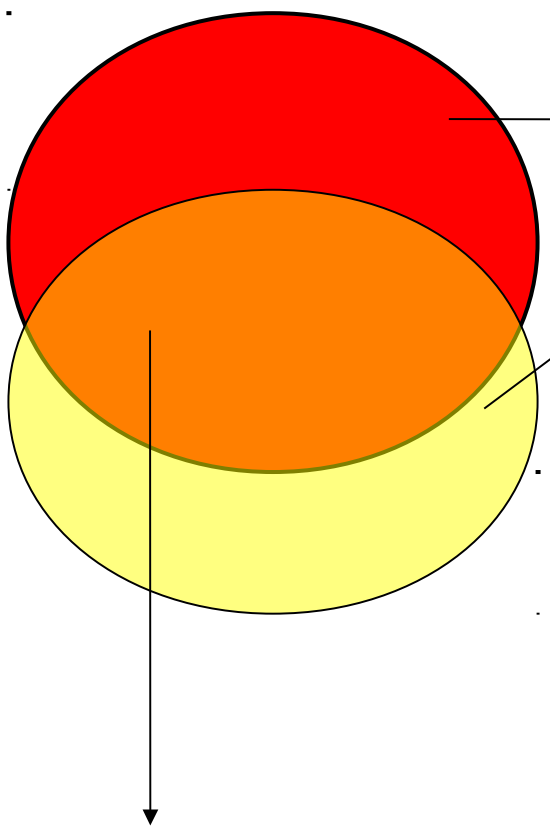
- **Important implications for phylogeny:** only sets of orthologous genes are expected to reflect the underlying species evolution (although there are many exceptions)

- 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

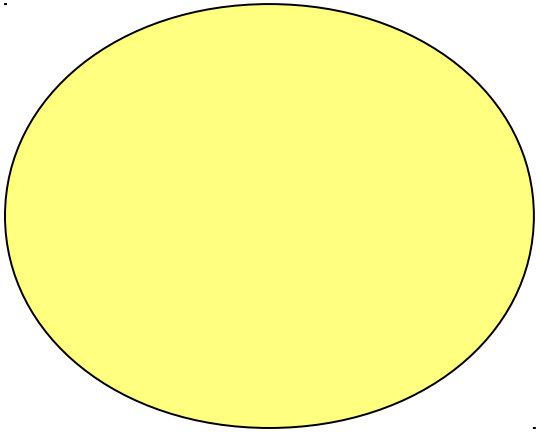

a)
b)
S_1
S_2
S_2
0
Row 1
Row 2
Row 3
Row 4

Gene content → co-evolution. (The easy case, few genomes.)

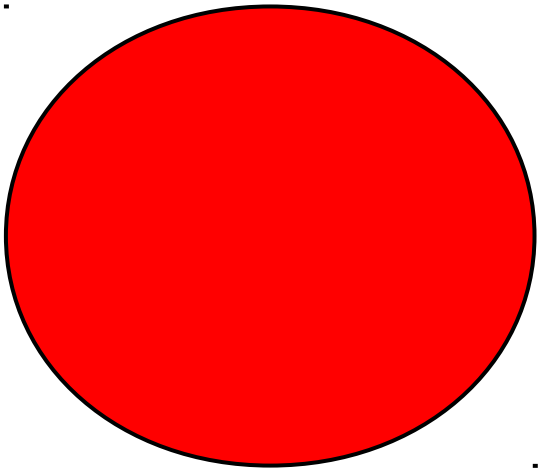


Differences between gene Content reflect differences in Phenotypic potentialities

Genomes share genes for phenotypes they have in common

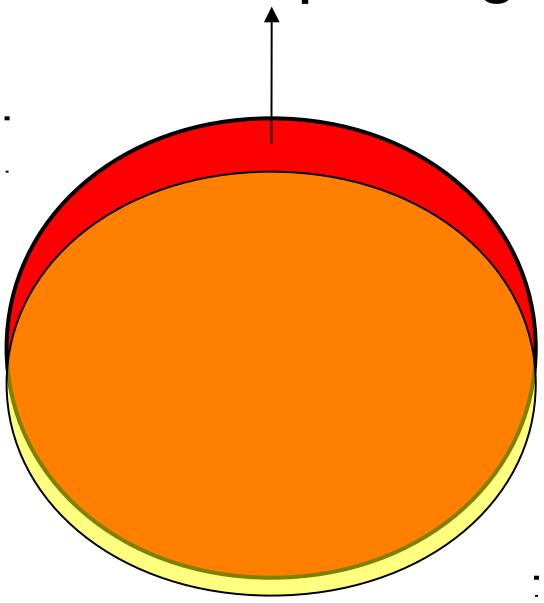


L. innocua (non-pathogen)



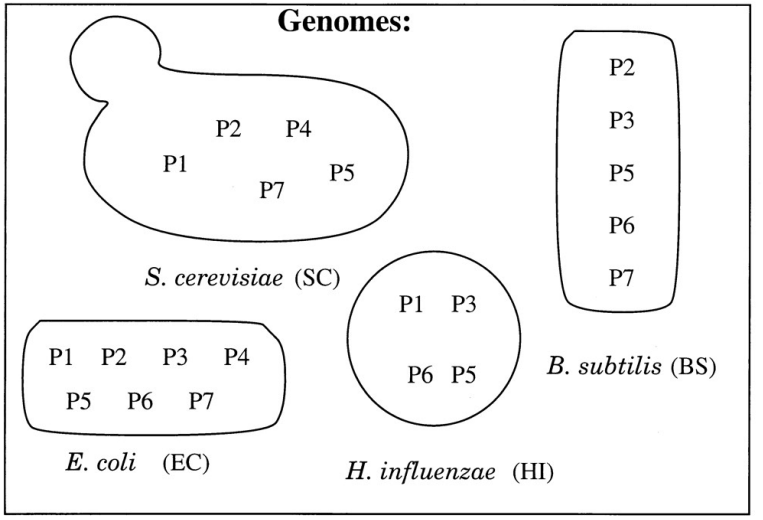
L. monocytogenes (pathogen)

Genes involved in pathogenecity



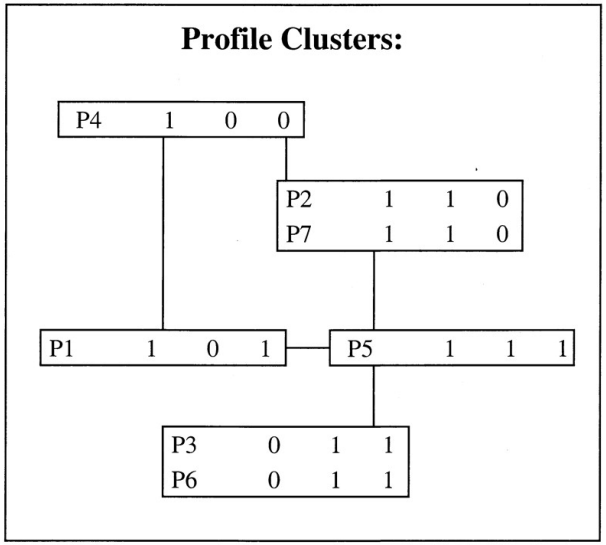
L. innocua (non-pathogenic) L. monocytogenes (pathogenic)

More than two genomes



Phylogenetic Profile:

EC	SC	BS	HI
P1	1	0	1
P2	1	1	0
P3	0	1	1
P4	1	0	0
P5	1	1	1
P6	0	1	1
P7	1	1	0



Conclusion: P2 and P7 are functionally linked , P3 and P6 are functionally linked

Assigning protein functions by comparative genome analysis: Protein phylogenetic profiles

Matteo Pellegrini, Edward M. Marcotte, Michael J. Thompson, David Eisenberg, and Todd O. Yeates

PNAS April 13, 1999. 96 (8) 4285-4288; https://doi.org/10.1073/pnas.96.8.4285

Contributed by David S. Eisenberg

Why predicting orthology is important?

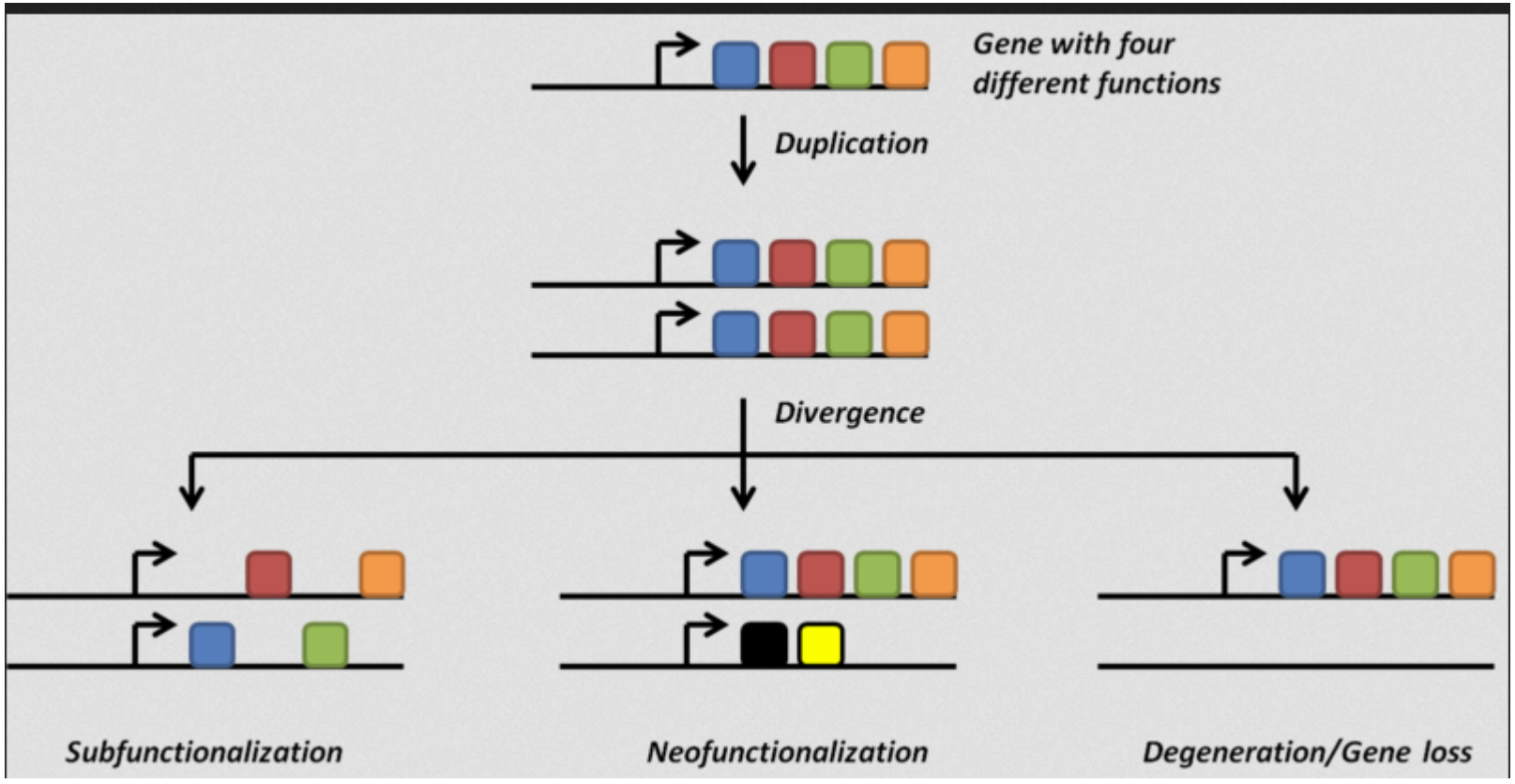
- **Important implications for phylogeny:** only sets of orthologous genes are expected to reflect the underlying species evolution (although there are many exceptions)

- 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

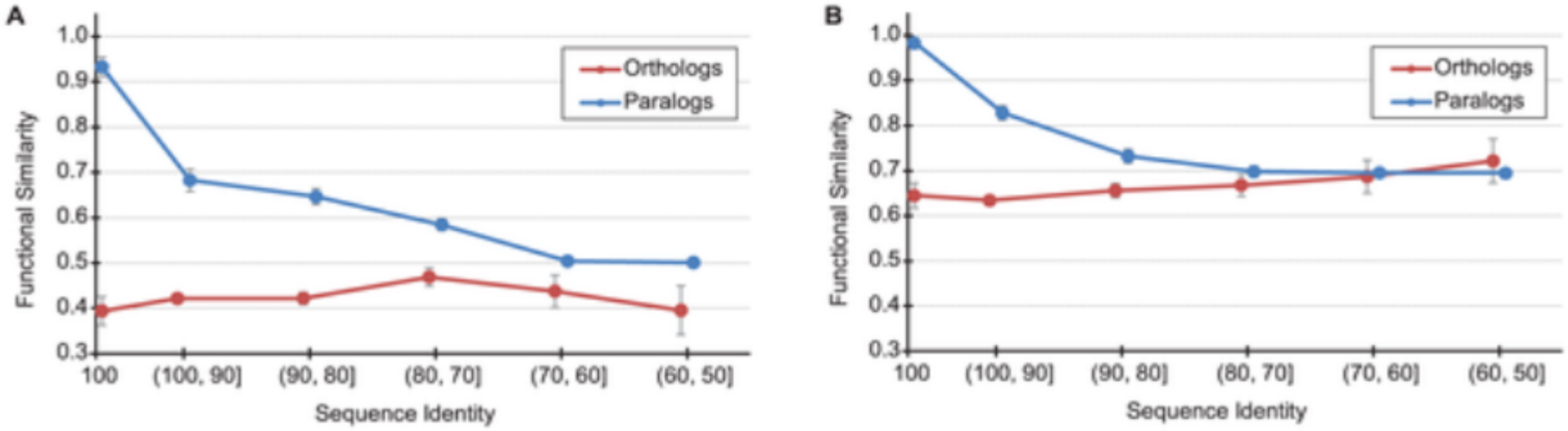
PLoS COMPUTATIONAL BIOLOGY

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

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

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

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

OPINION

Functional and evolutionary implications of gene orthology

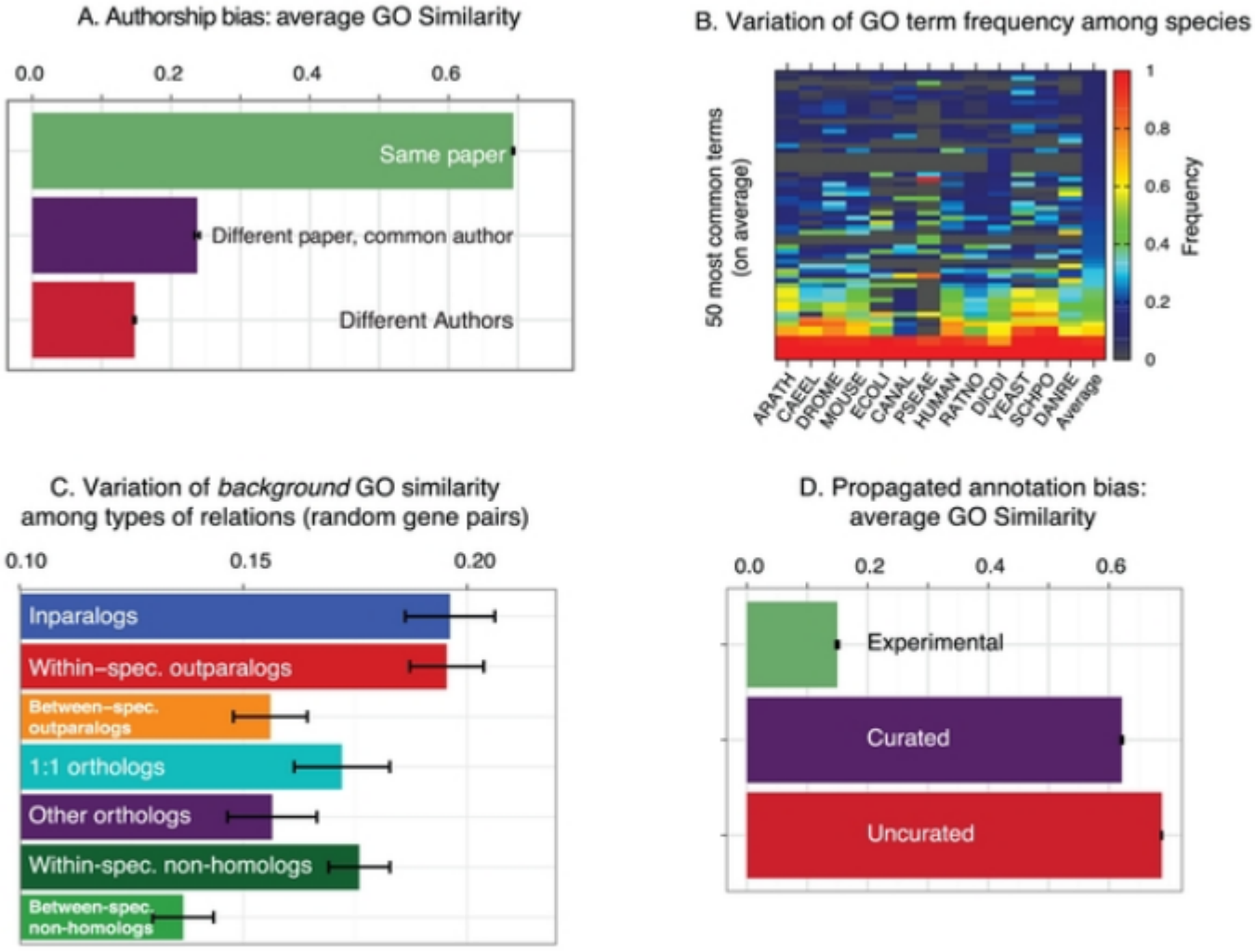
Toni Gabaldón and Eugene V. Koonin

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

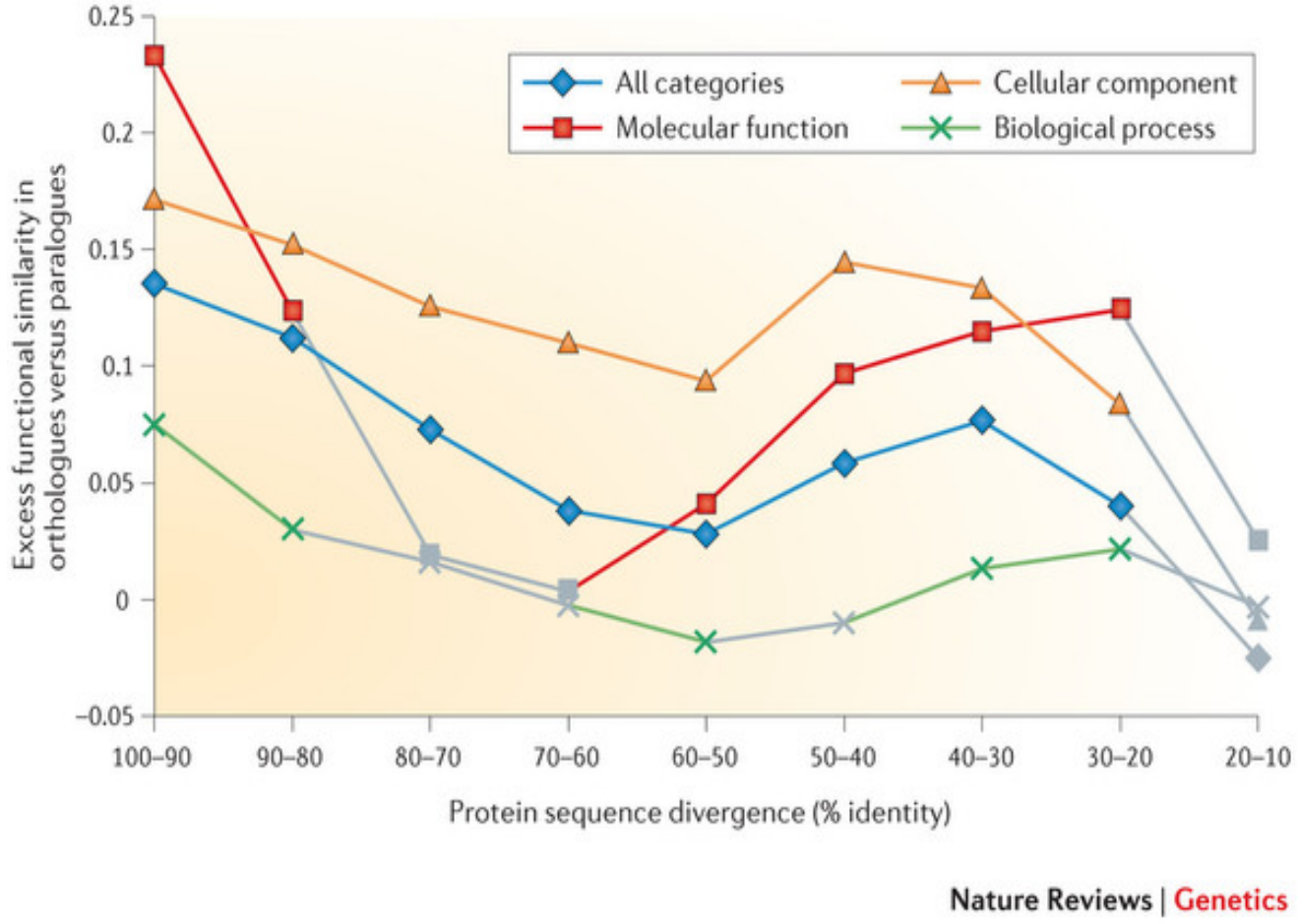
Jaime Huerta-Cepas, Joaquín Dopazo, Martijn A. Huynen and Toni Gabaldón

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



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

Questions about this part?

Barcelona Biomedical Research Park

CRG Centre de Regulació Genòmica

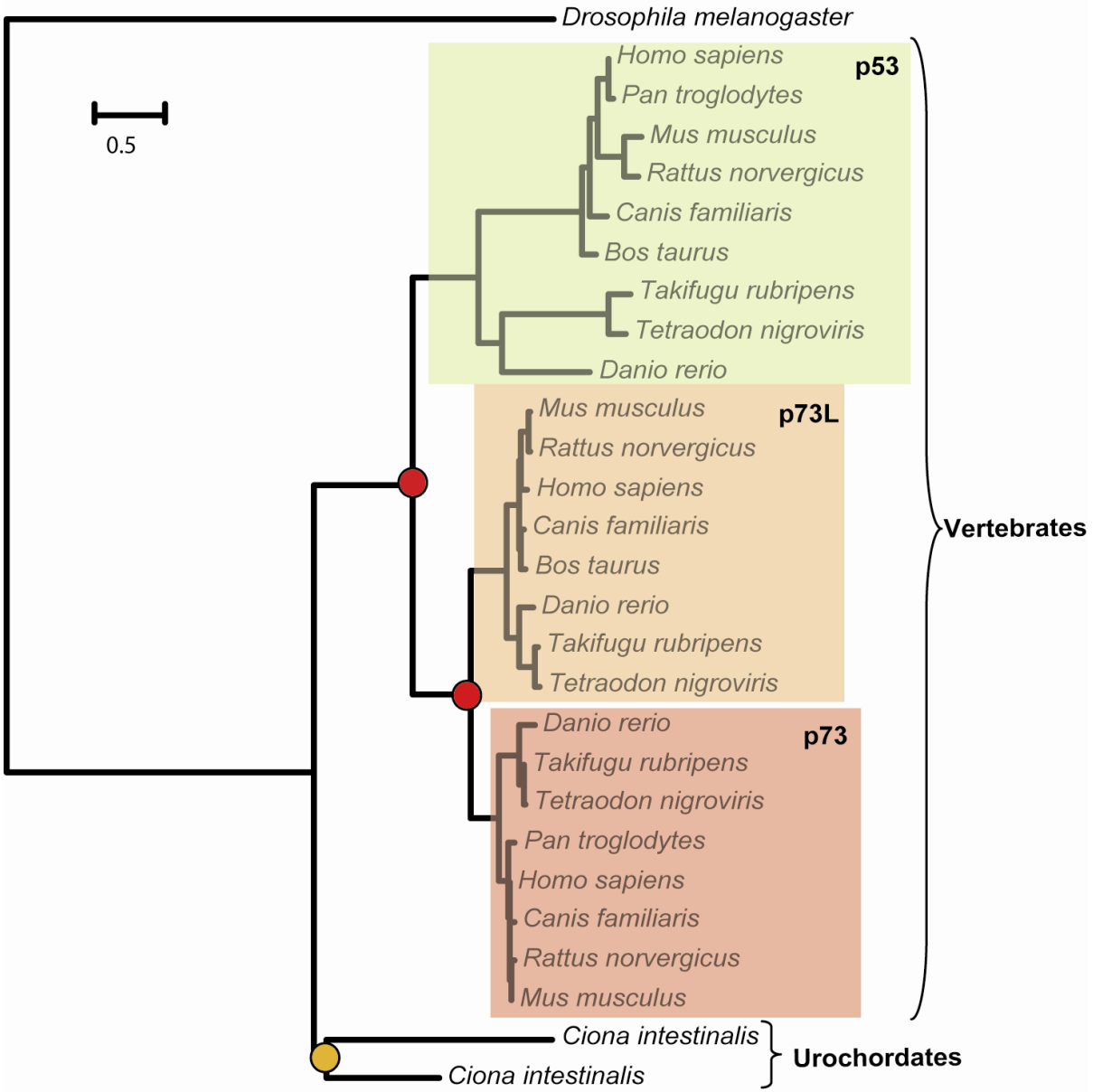
Orthology
Part II

Orthology prediction methods

Toni Gabaldón
Centre for Genomic Regulation (CRG), Barcelona

Classical approach: phylogenetic inference

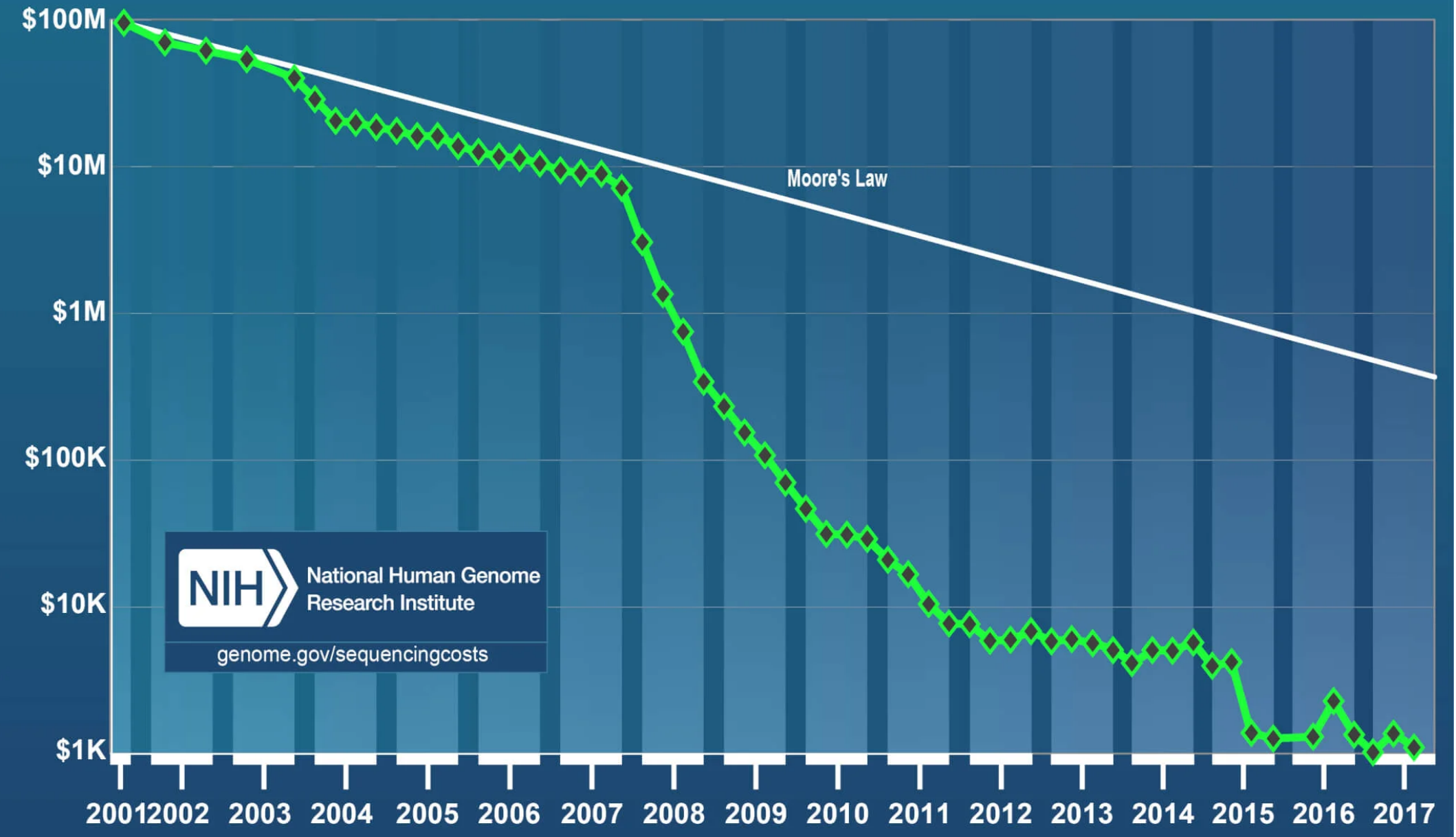
- Build a gene tree
- Compare to the species tree
- Infer duplications and speciation events
- Assign orthology and paralogy relationships accordingly

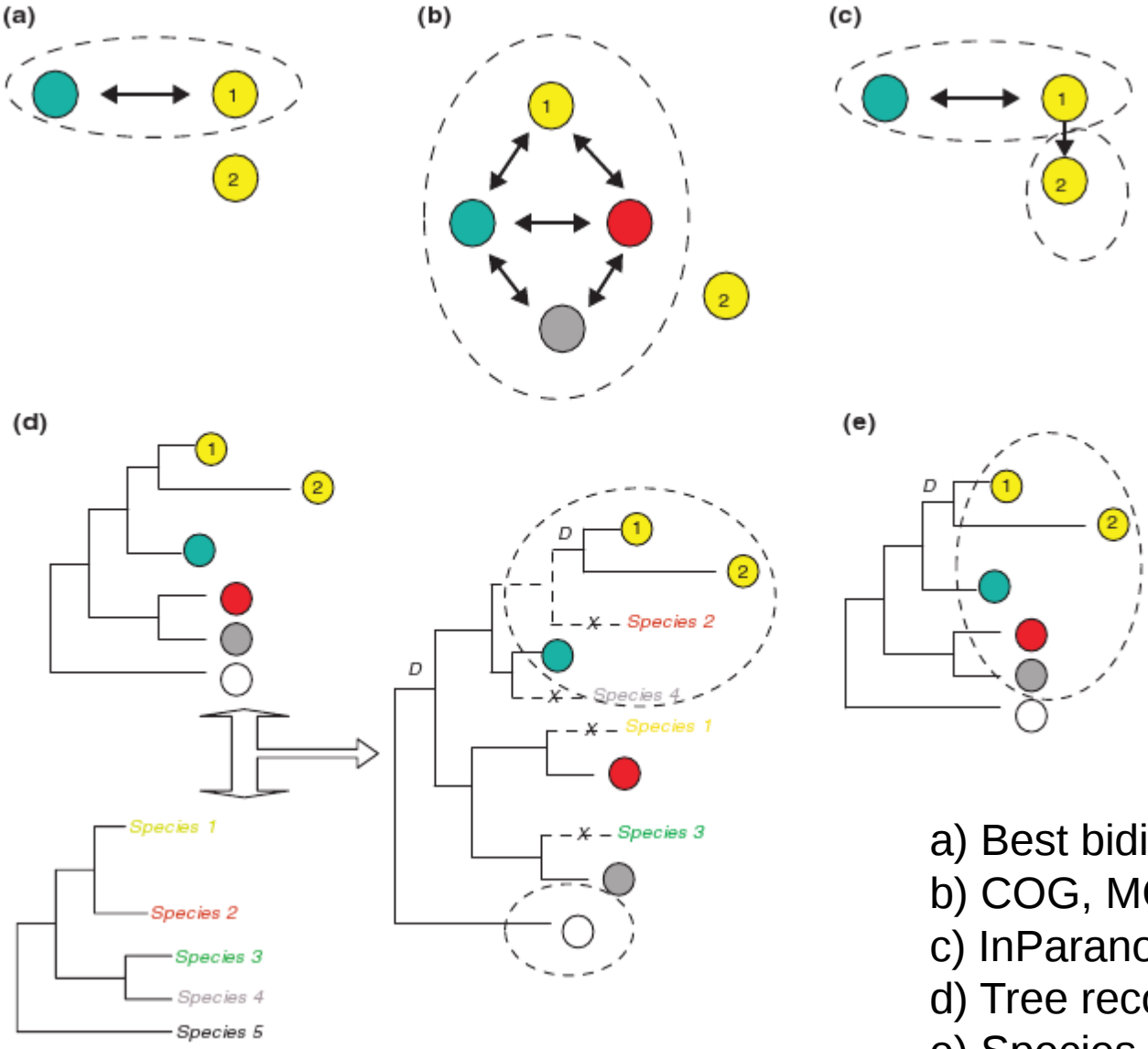

Drosophila melanogaster
0.5
Homo sapiens
Pan troglodytes
Mus musculus
Rattus norvergicus
Canis familiaris
Bos taurus
Takifugu rubripens
Tetraodon nigroviris
Danio rerio
p53
Mus musculus
Rattus norvergicus
Homo sapiens
Canis familiaris
Bos taurus
Danio rerio
Takifugu rubripens
Tetraodon nigroviris
p73L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

Going genome-wide scale:

Everything must be done automatic and “blind”

Cost per Genome





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

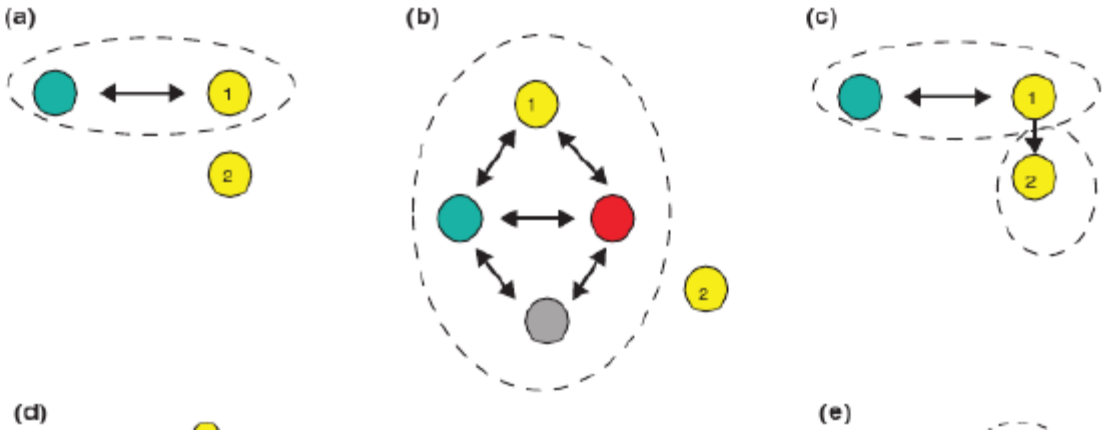
Gabaldón, T. *Genome Biology* (2008)

Similarity-based approaches (many more approaches):

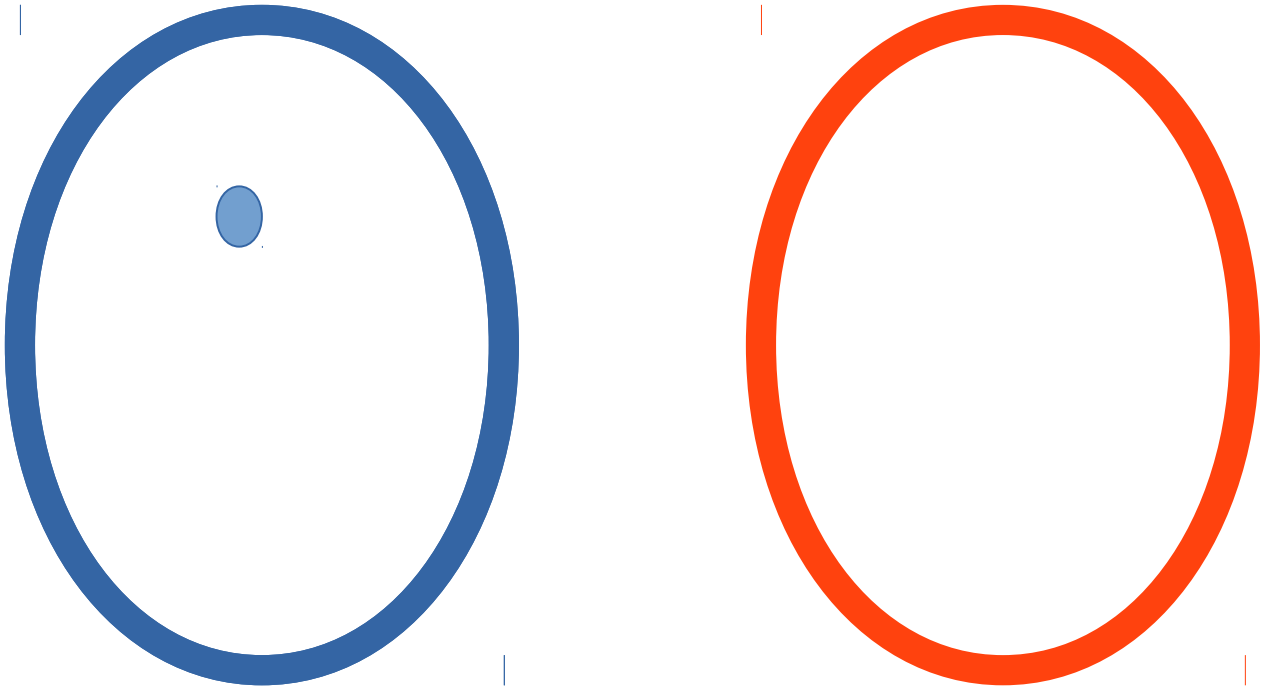
- **Best Reciprocal Hits**
- Detects all orthologies as one-to one. Highly affected by paralogy. Low rate of false positives but high rates of false negatives.

- The simplest and fastest method, still widely used

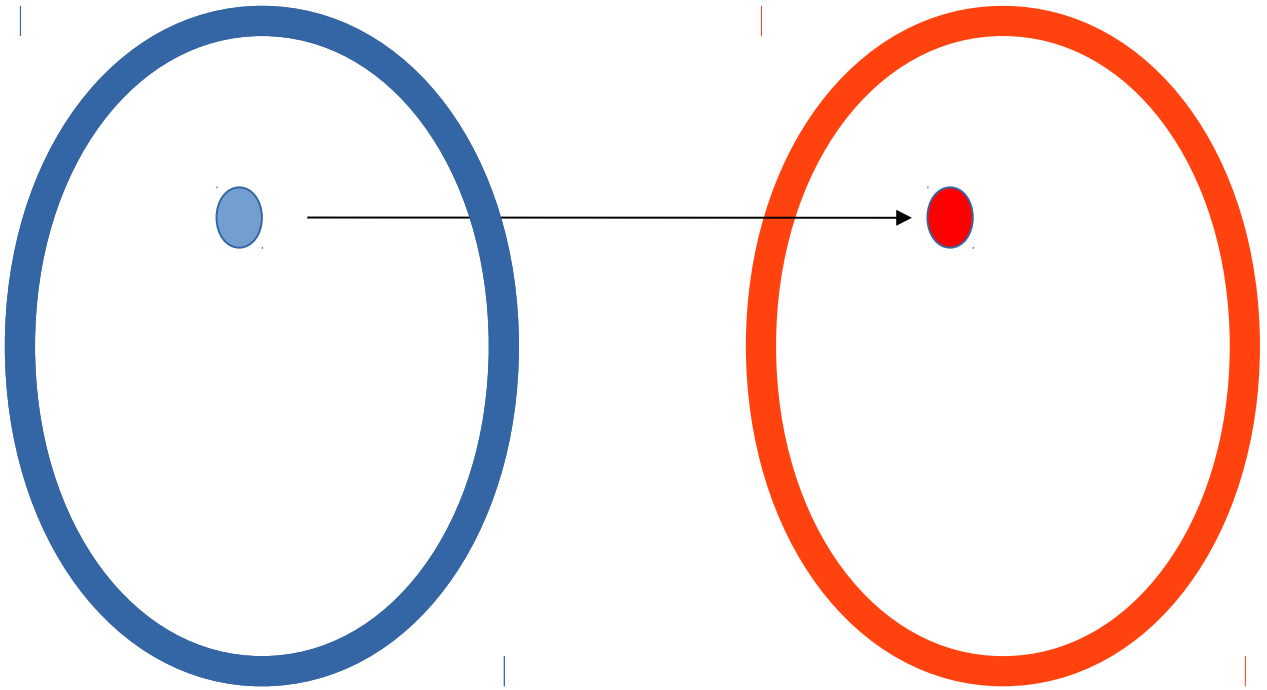
-



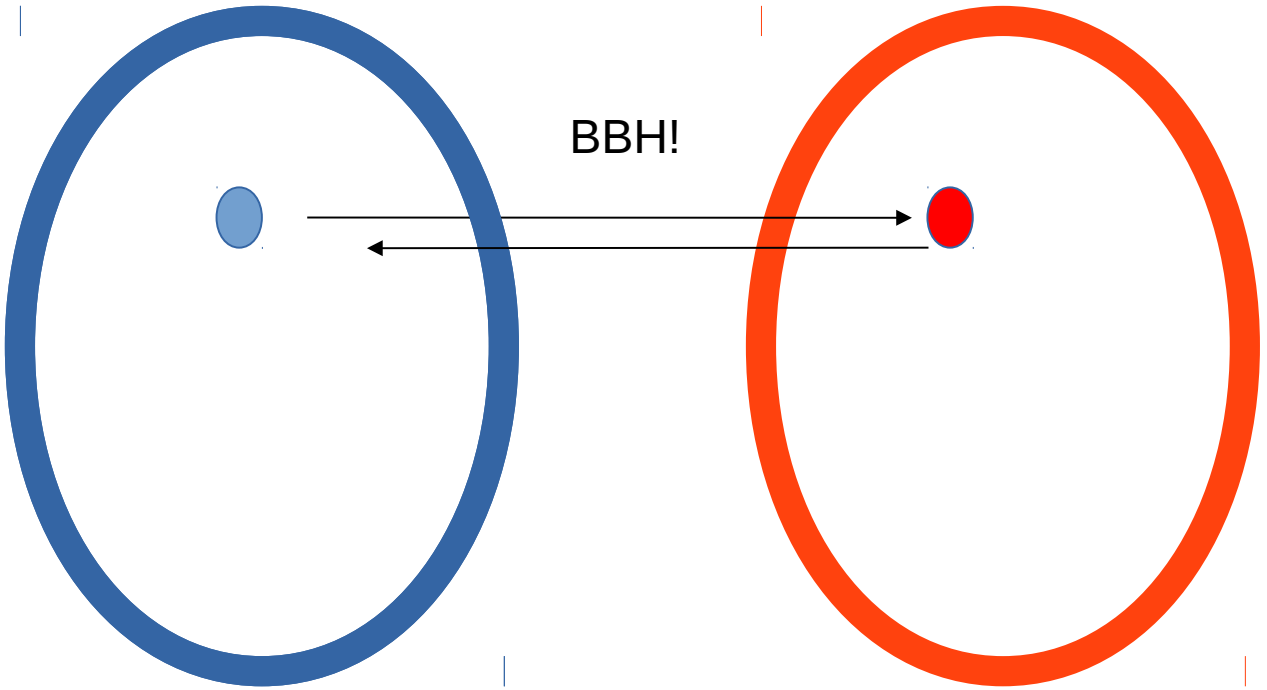
Best bidirectional hit (BBH), Best reciprocal hits (BRH)

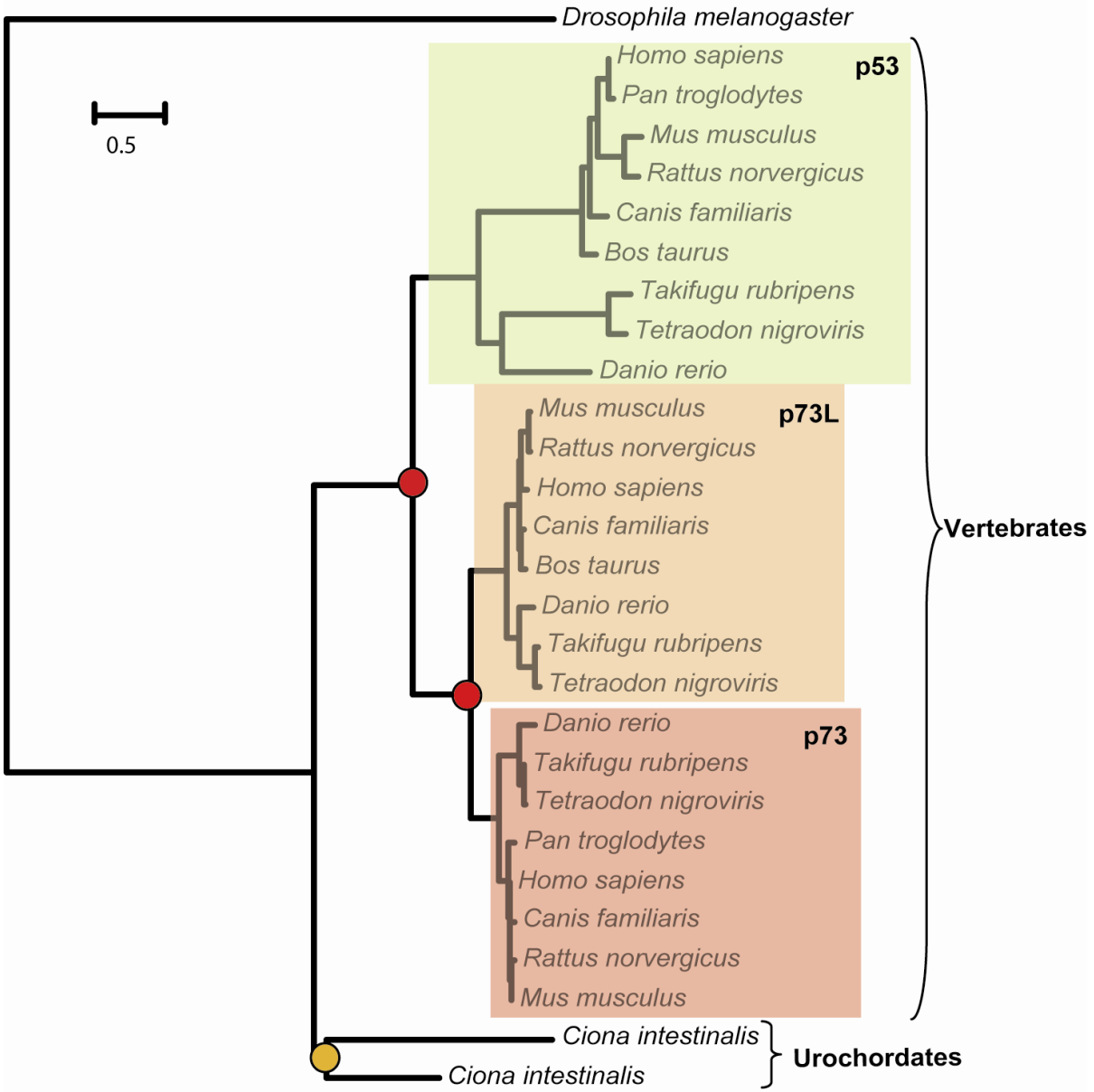


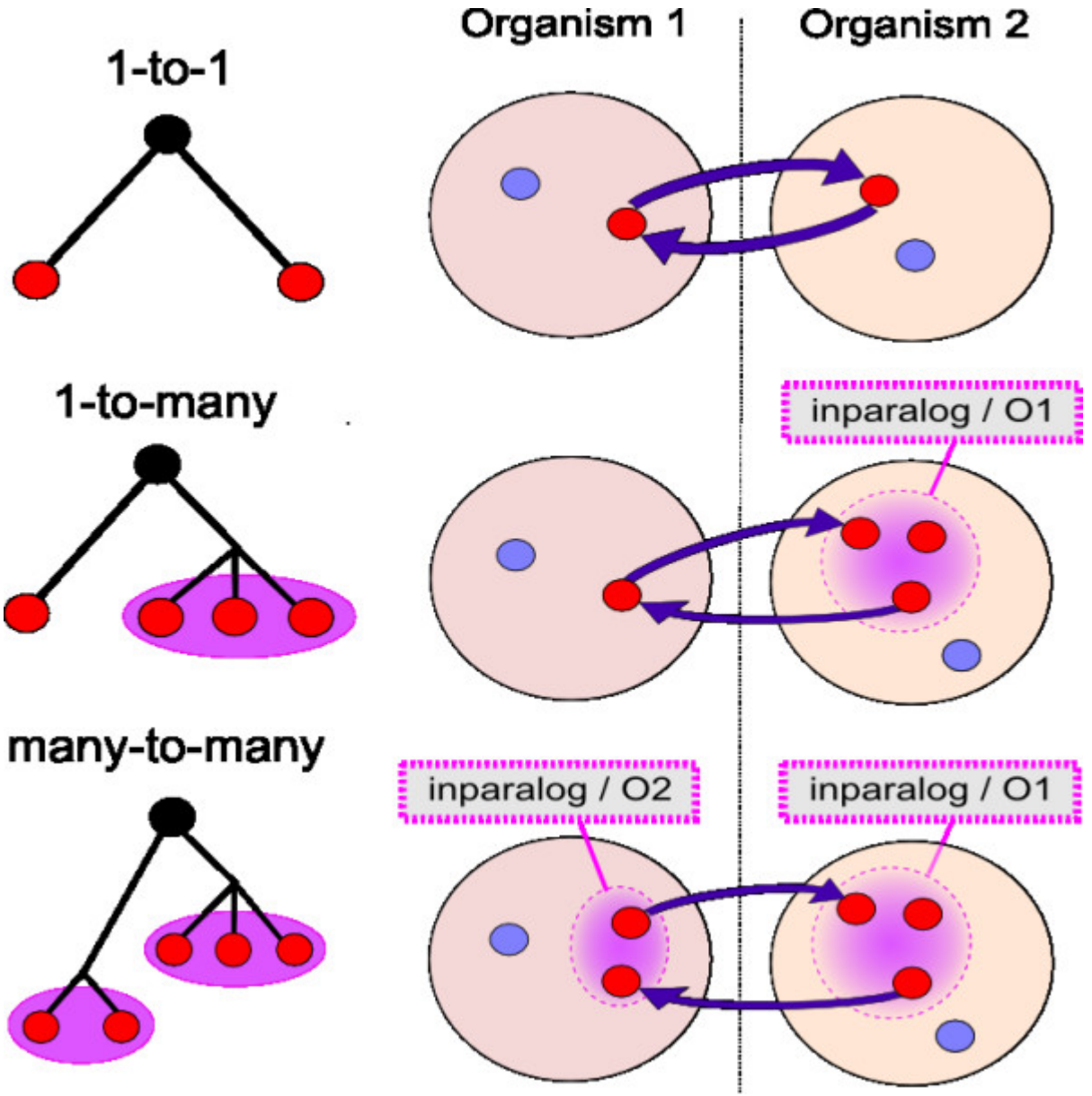
Best bidirectional hit (BBH), Best reciprocal hits (BRH)

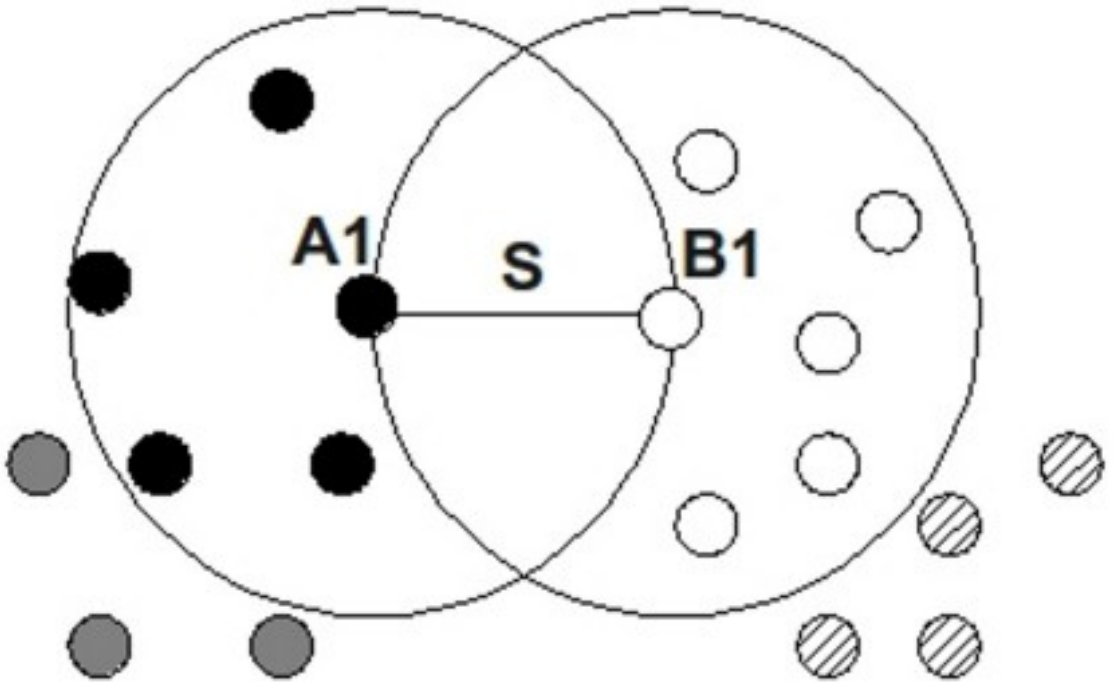


Best bidirectional hit (BBH), Best reciprocal hits (BRH)




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


Organism 1
Organism 2
1-to-1
1-to-many
inparalog / O1
many-to-many
inparalog / O2
inparalog / O1



A1 Seed ortholog species A
B1 Seed ortholog species B
S Score seed orthologs

Inparalogs species A
Inparalogs species B
Outparalogs species A
Outparalogs species B

In-paranoid: improved BRH to detect in-paralogs as well. Works well at the pairwise level.

Note:

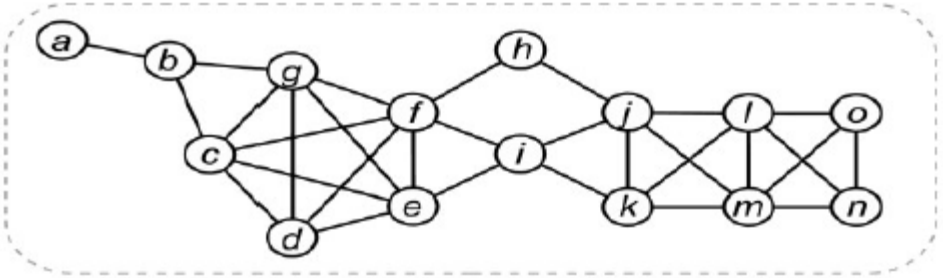
Definition of **in-** and **out-paralogues** require the specification of a given **speciation-node** of reference

COG-like
(used by many DBs like STRING)

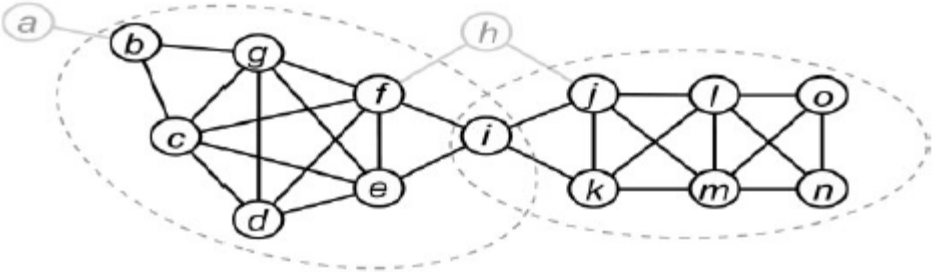
Exploits multi-species information.
Predicts clusters of orthologous groups (in-paralogs) not all pairs in a cluster are paralogs.

Can be used at different stringent levels

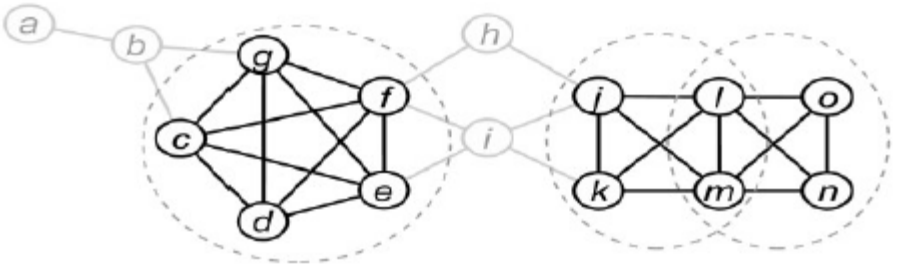
2



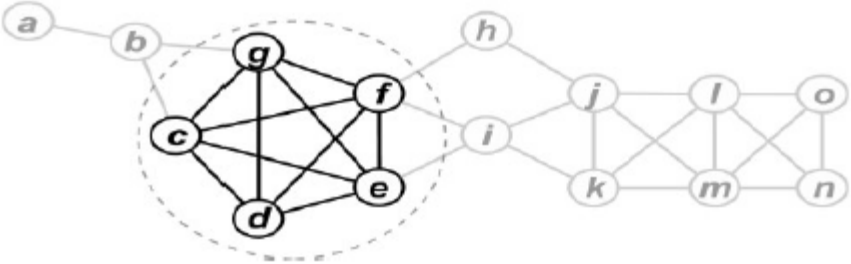
3



4



5



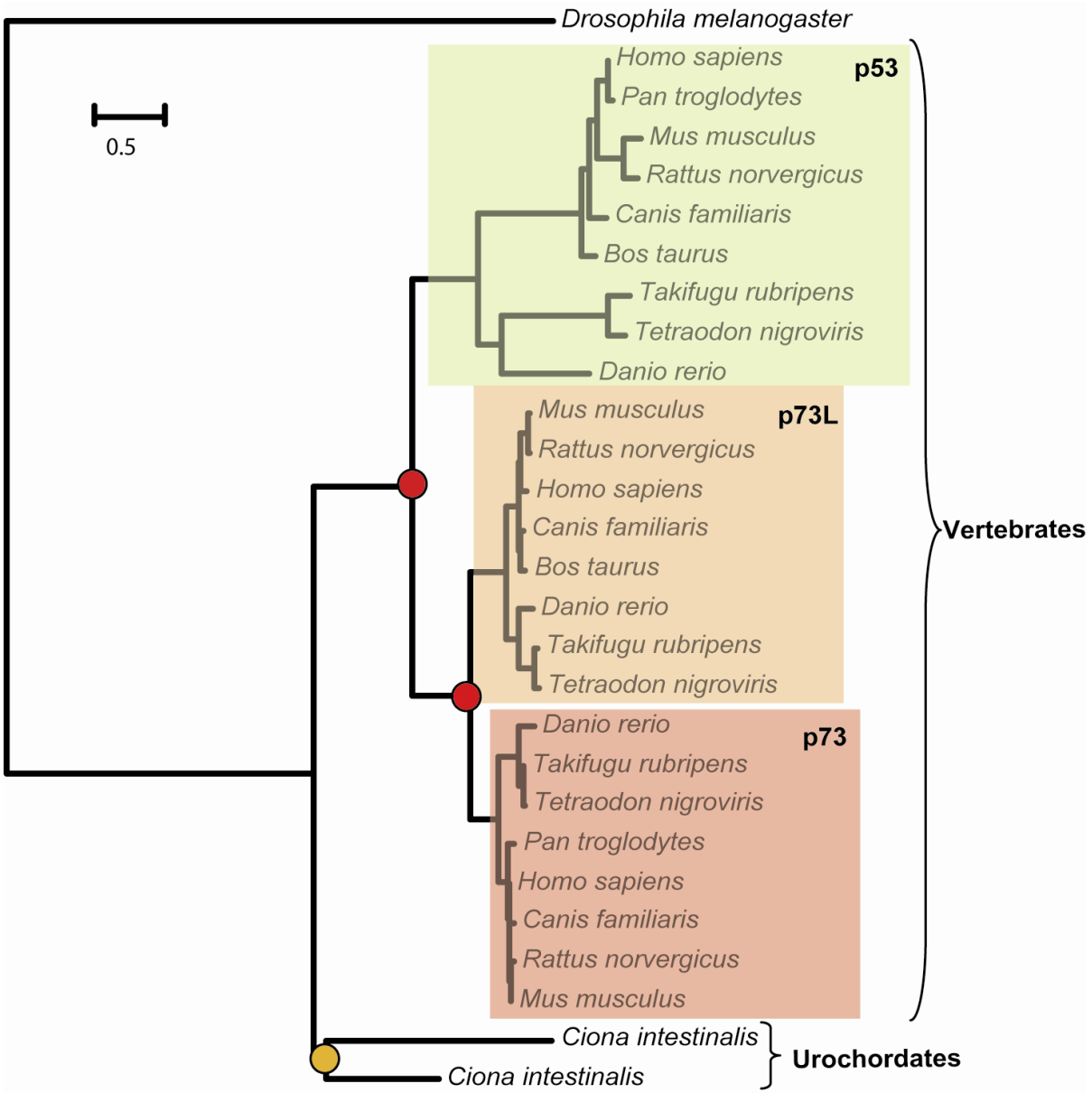
Clustering methods produce: orthologous groups

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



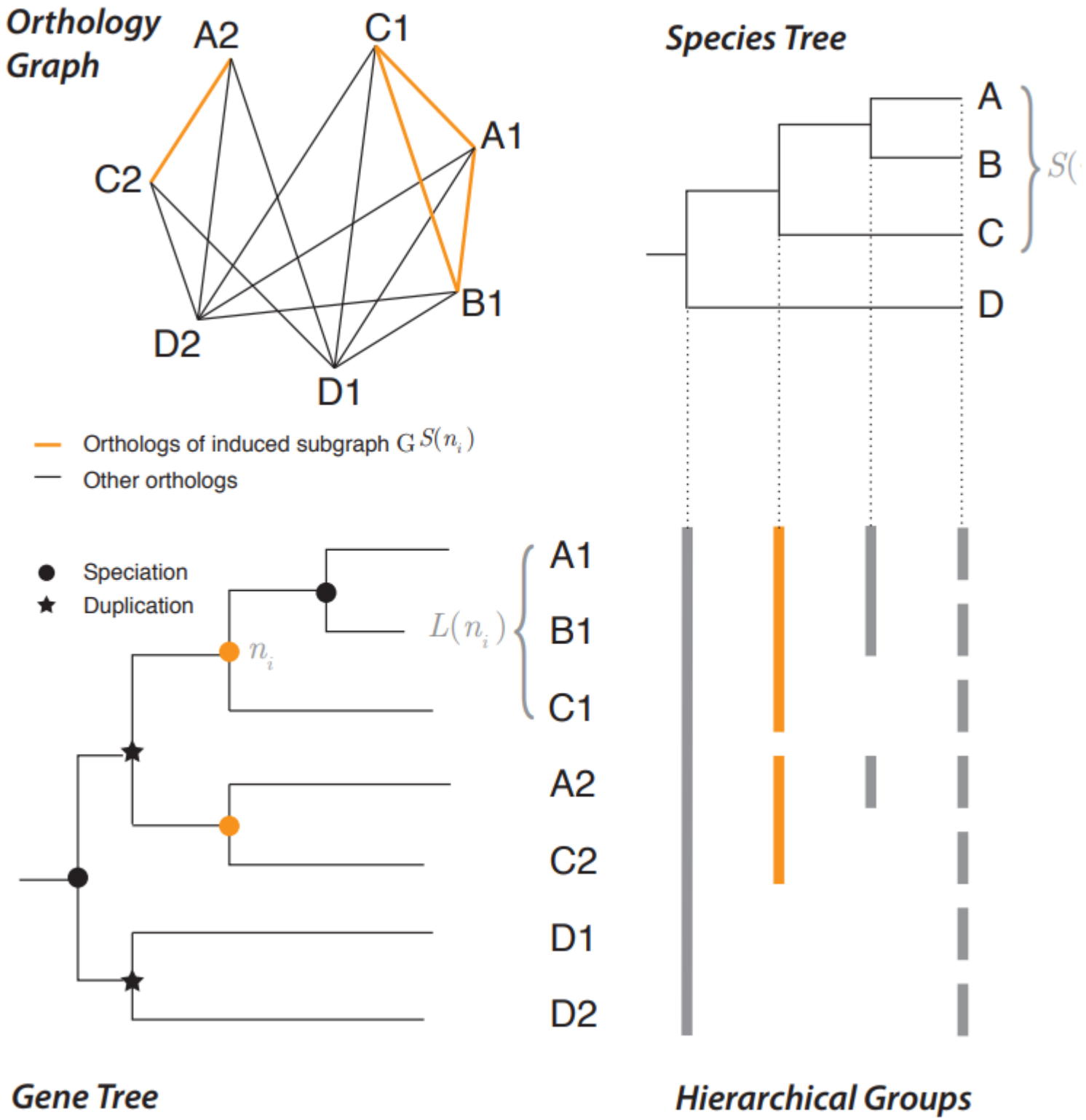
Additional useful definitions

- **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)


Orthology Graph
A2
C1
A1
C2
B1
D2
D1
Orthologs of induced subgraph G S(n_i)
Other orthologs
Species Tree
A
B
C
D
S(
Speciation
Duplication
n_i
L(n_i)
A1
B1
C1
A2
C2
D1
D2
Gene Tree
Hierarchical Groups

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:

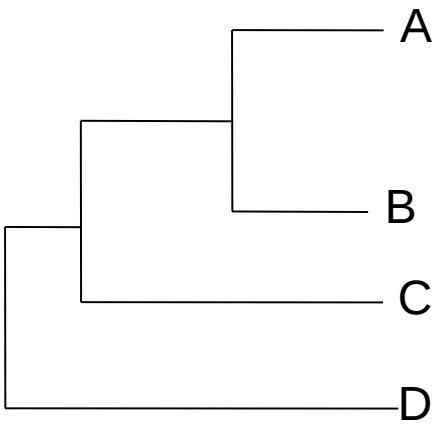
→ Species tree reconciliation (RIO, Ensembl)

→ Species-overlap algorithms

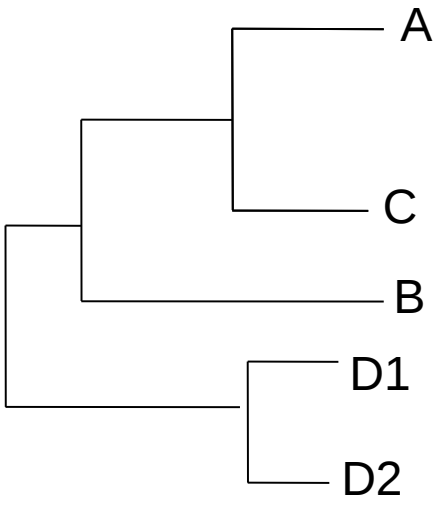
Reconciliation algorithm.

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

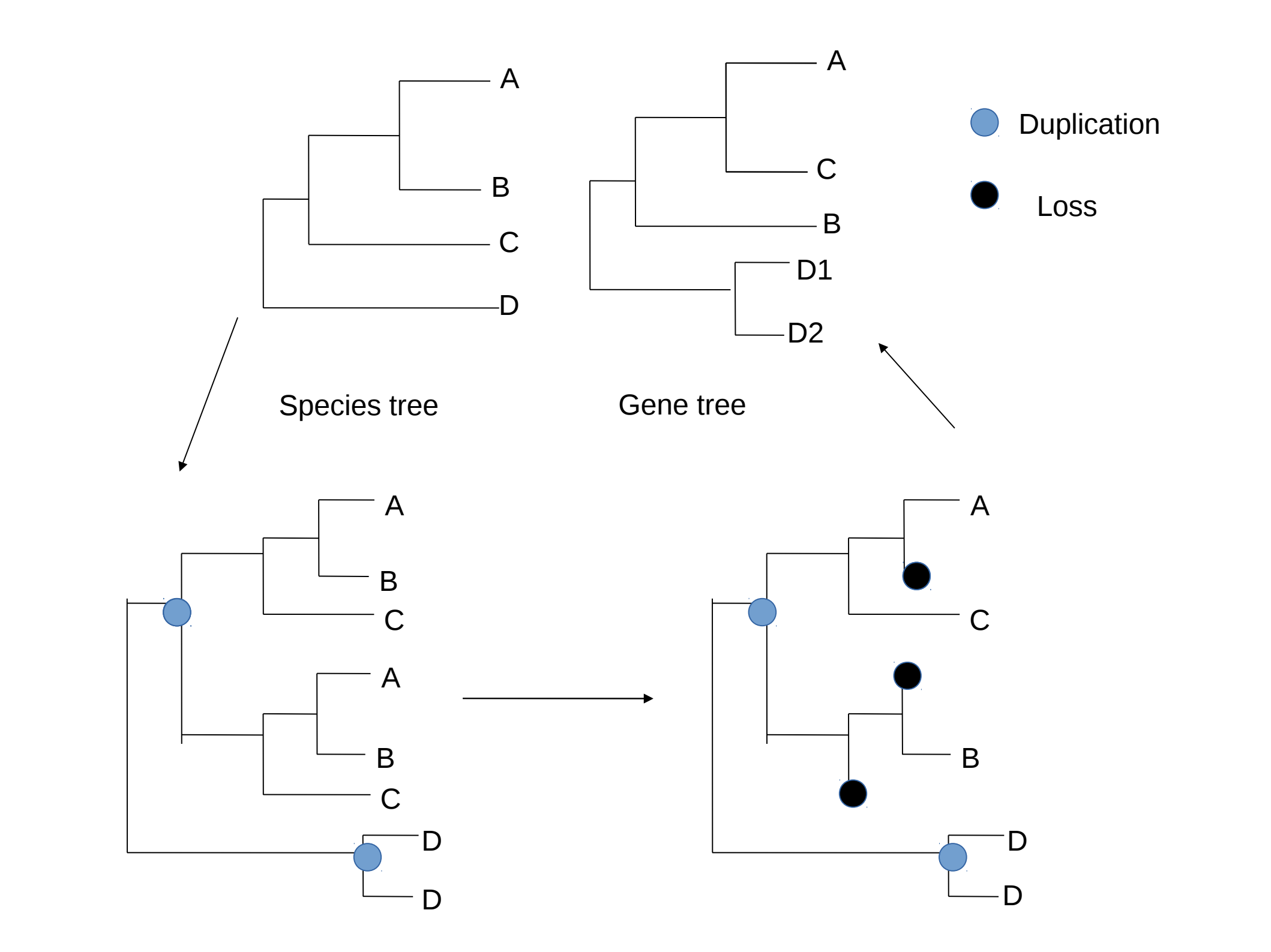
(Soft reconciliation) Allow incongruences below a given support value



Species tree



Gene tree

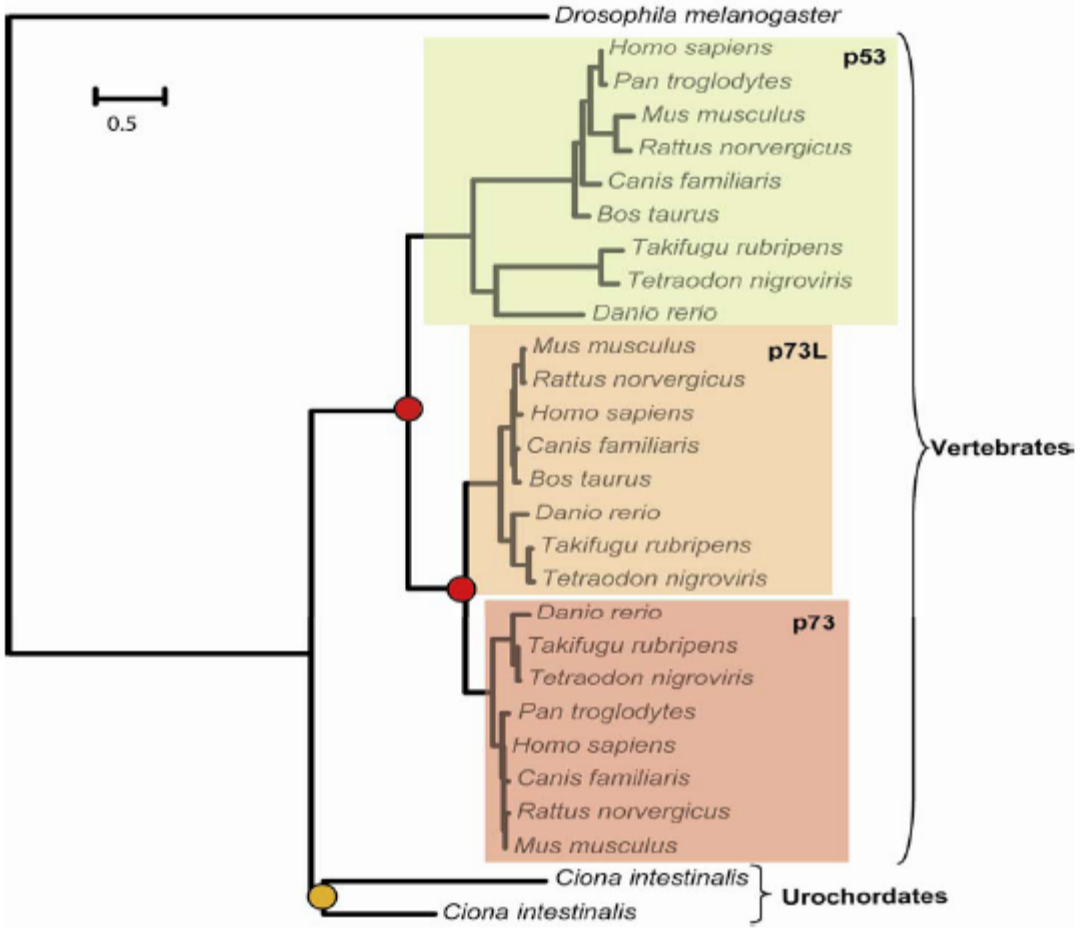

A
B
C
D
Species tree
A
C
B
D1
D2
Gene tree
Duplication
Loss
A
B
C
A
B
C
D
D
A
C
B
D
D

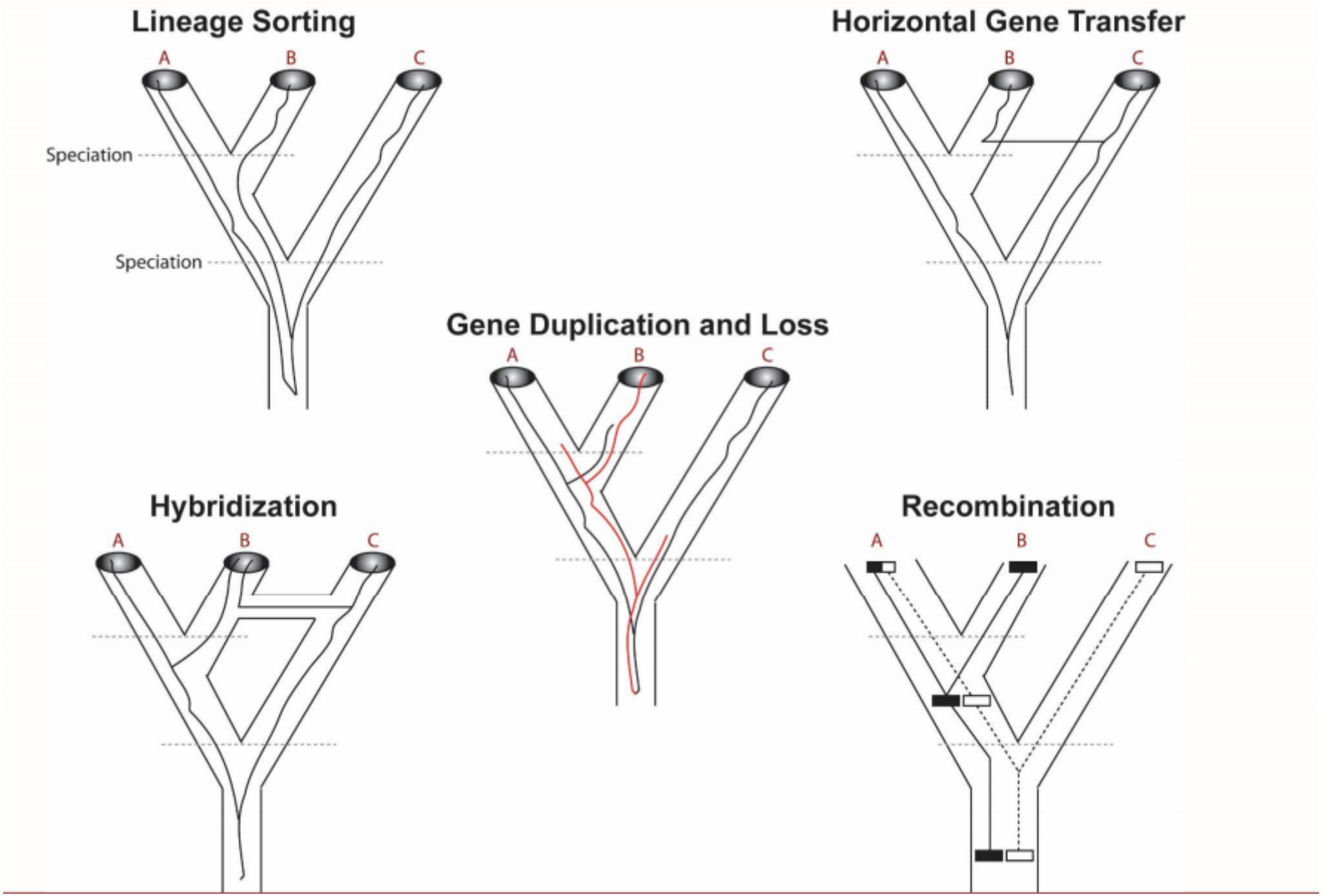
Reconciliation with the species tree readily provides you information on speciation and duplication nodes in a tree

It works when these two assumptions are correct:

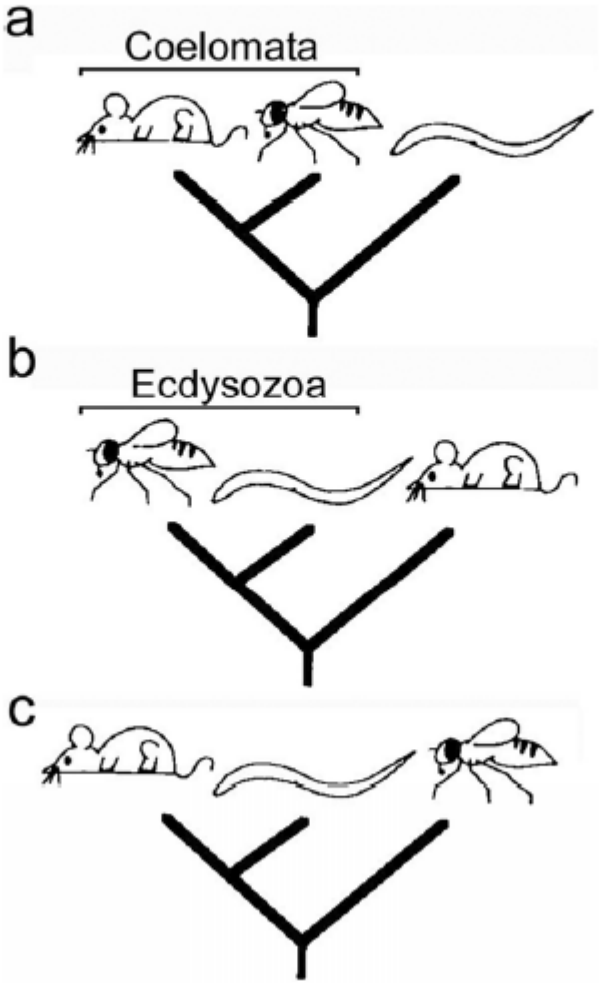
A) We know the true species tree

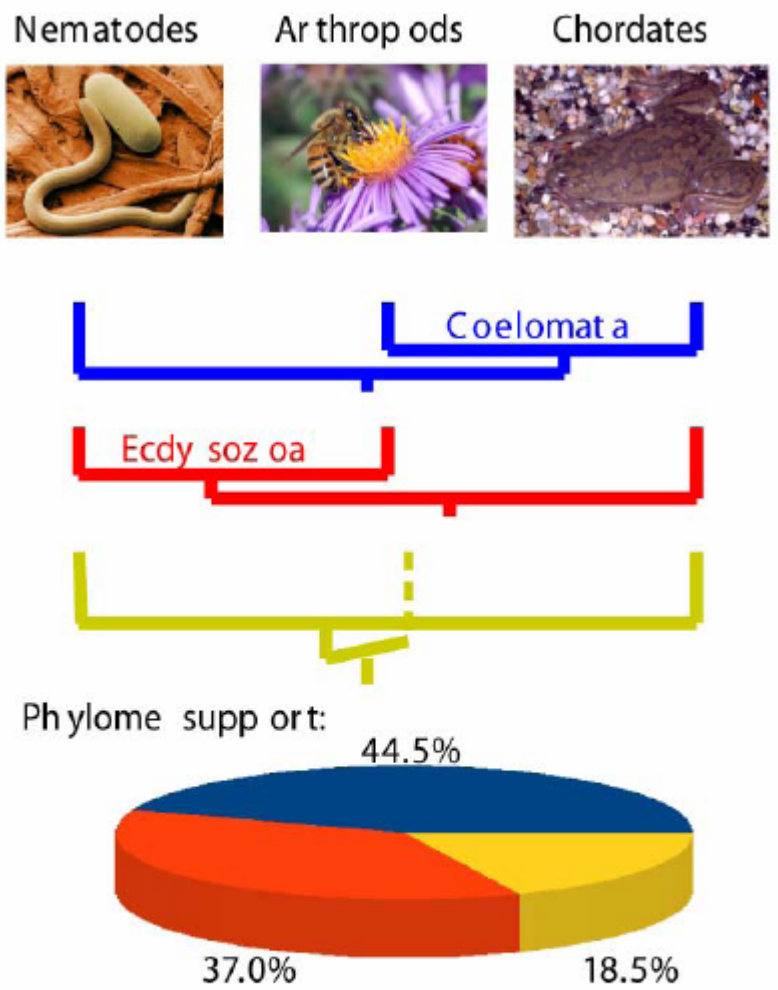
B) The gene tree is correct and reflects the species evolution




Lineage Sorting
A
B
C
Speciation
Speciation
Horizontal Gene Transfer
A
B
C
Gene Duplication and Loss
A
B
C
Hybridization
A
B
C
Recombination
A
B
C

Uncertainty in species trees and topological variability in gene trees





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

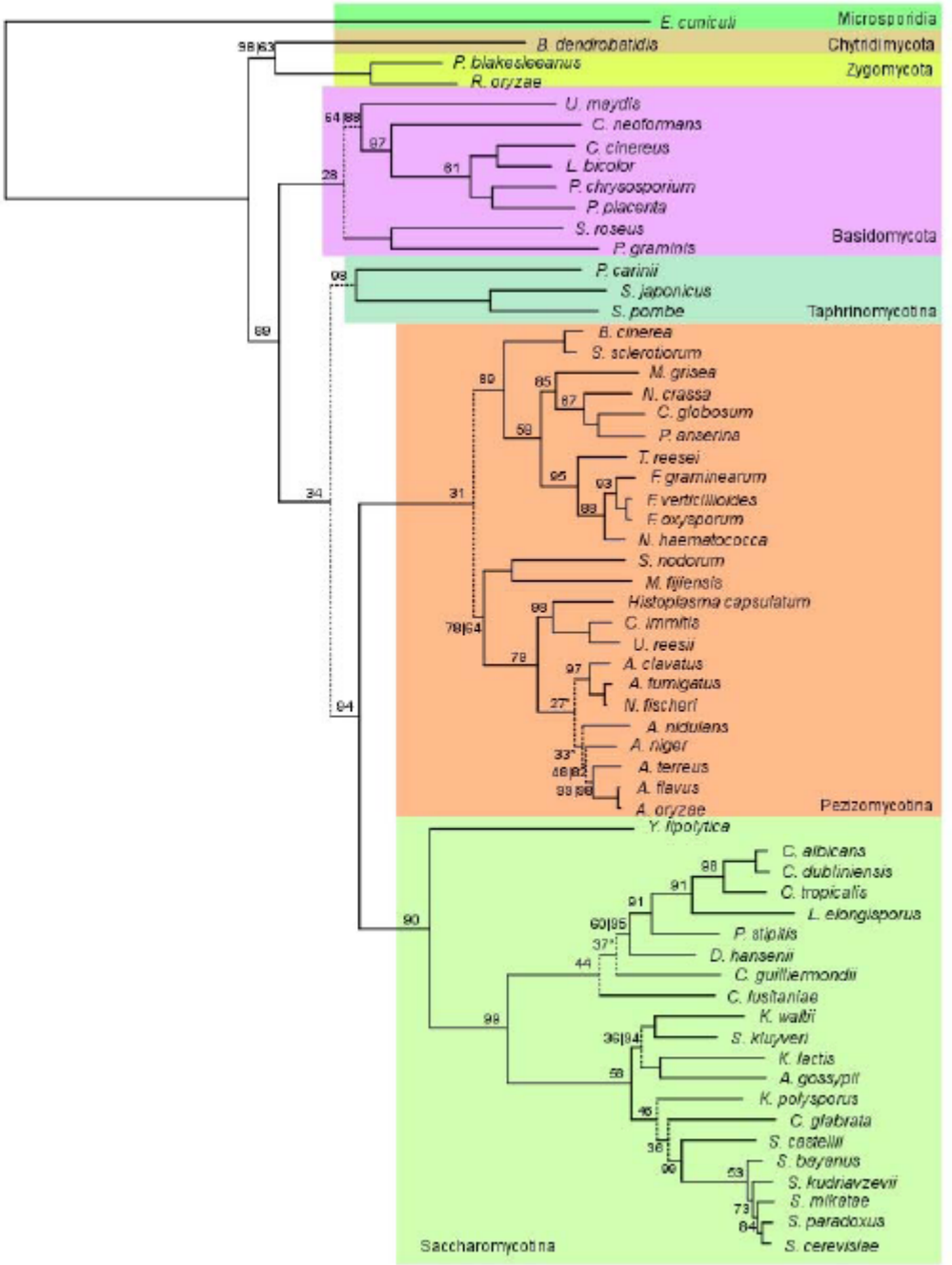
Similar results for

Primates
Rodents
laurasatheria

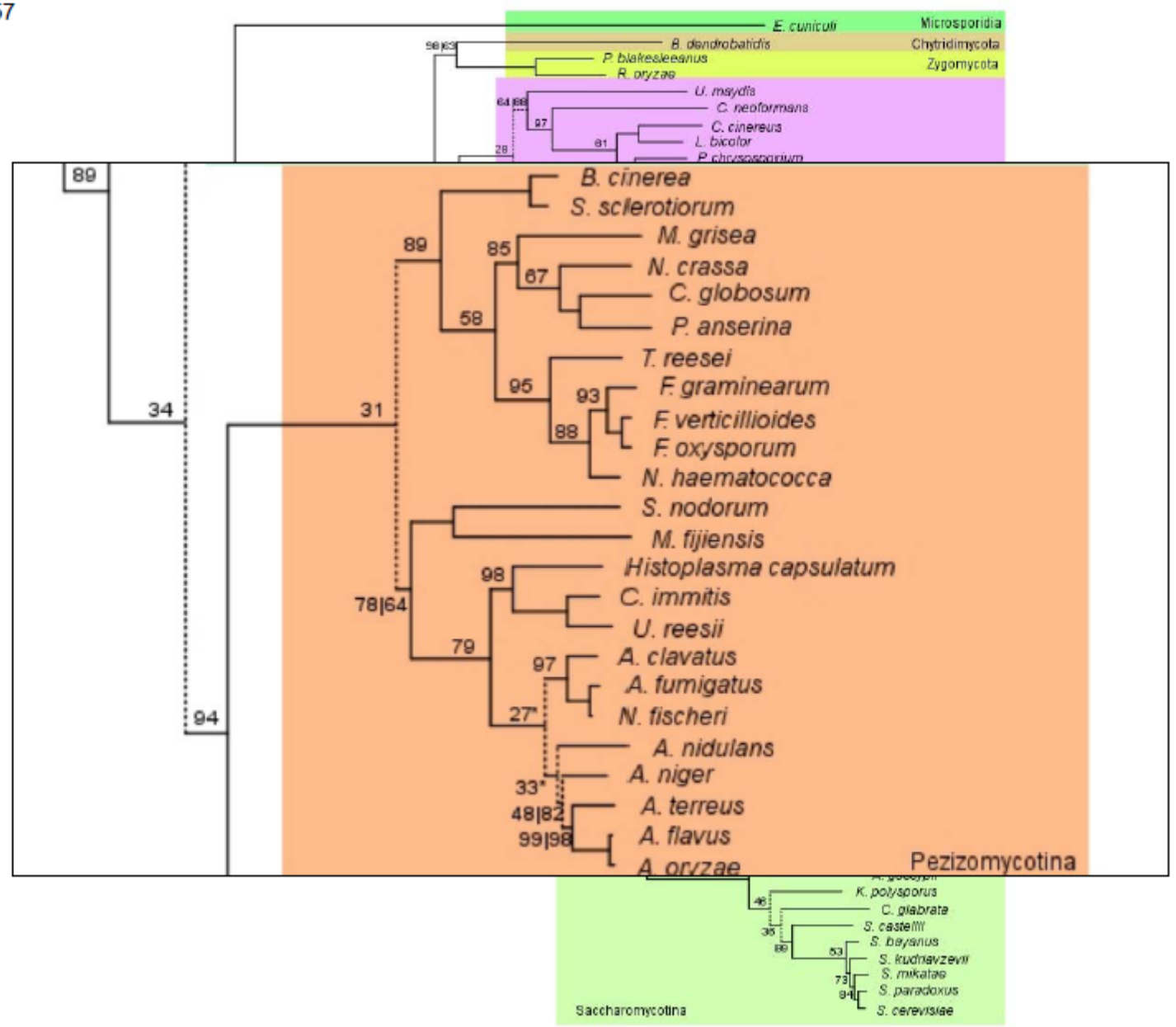
The tree vs the forest:

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



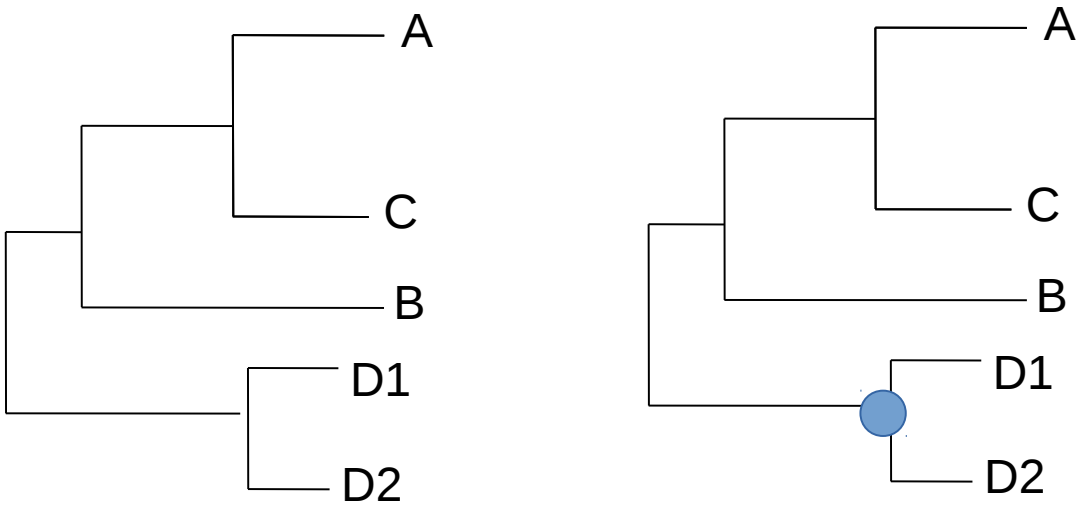
This large-degree of topological variability might be in part due to phylogenetic artifacts, insuficient phylogenetic signal, etc. But also to real evolutionary processes that render a gene tree different from a species tree: lineage sorting, gene conversion, etc

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

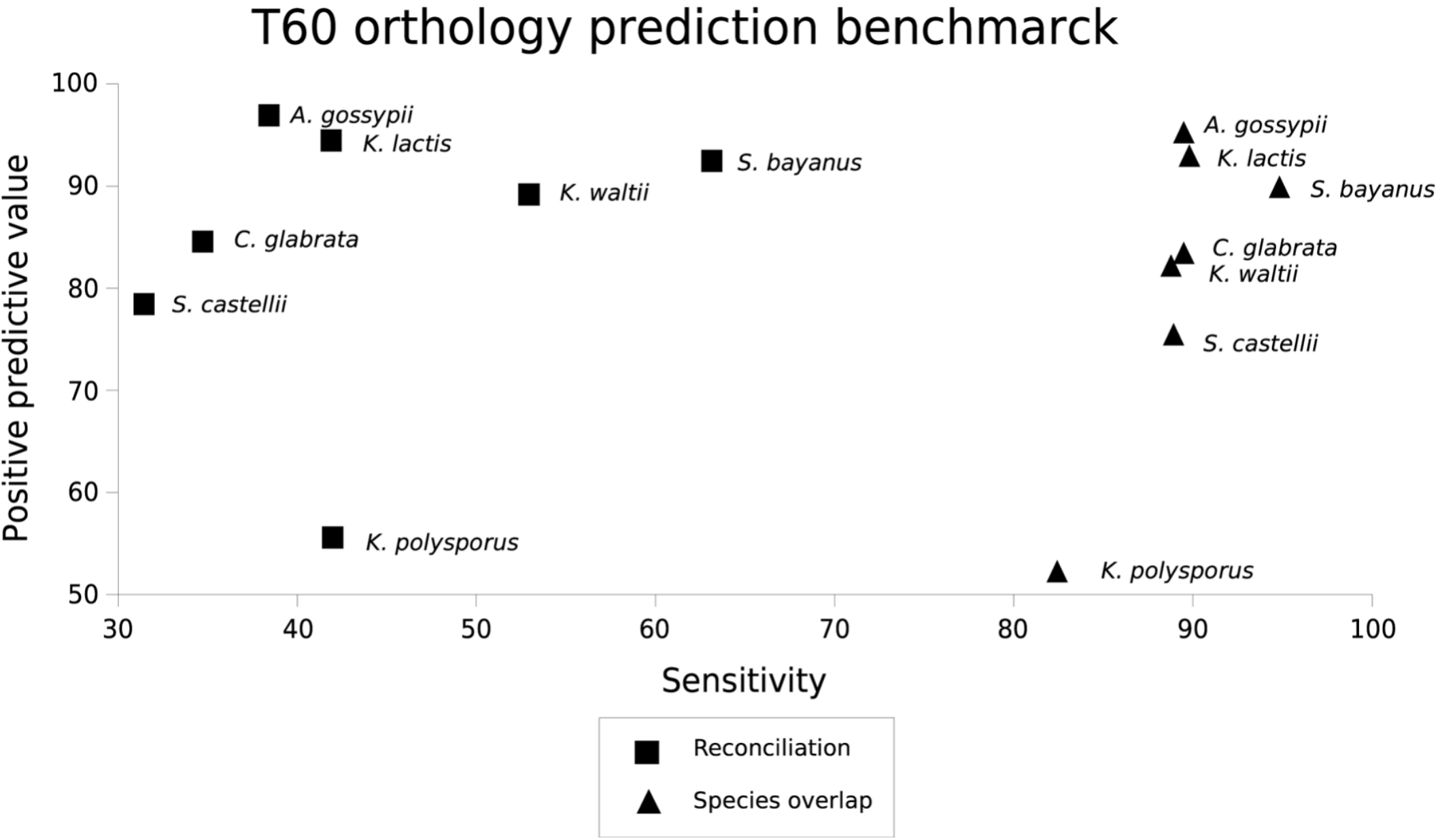
Species ovelap algorithm.

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



Benchmark on YGOB (Genome alignments and synteny information)

Our pipeline:

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

PHYLOME

ETE: Environment for Tree Exploration
ete.cgenomics.org

for every gene

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

Multimple Sequence Alignments
- Alignment reconstruction.
- Alignment trimming.

trimAl
A tool for automated alignment trimming

Salvador-Capella et al
Bioinformatics (2009).

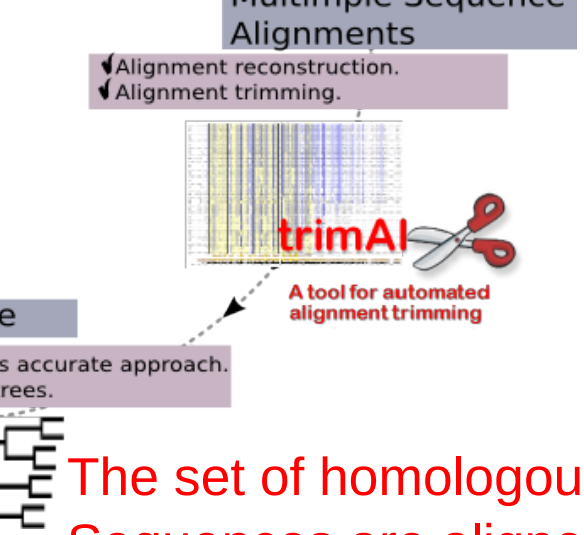
http://trimal.cgenomics.org

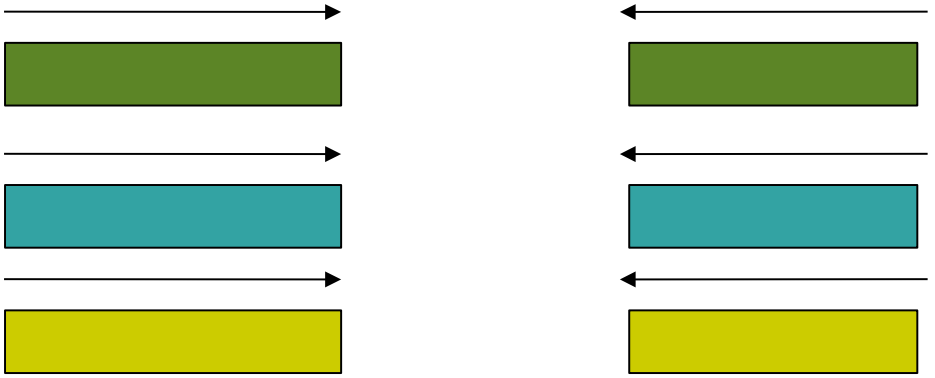
NJ Tree
- Quick but less accurate approach.
- Seed for ML trees.

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].

Pipeline described in Huerta-Cepas et al NAR (2011)

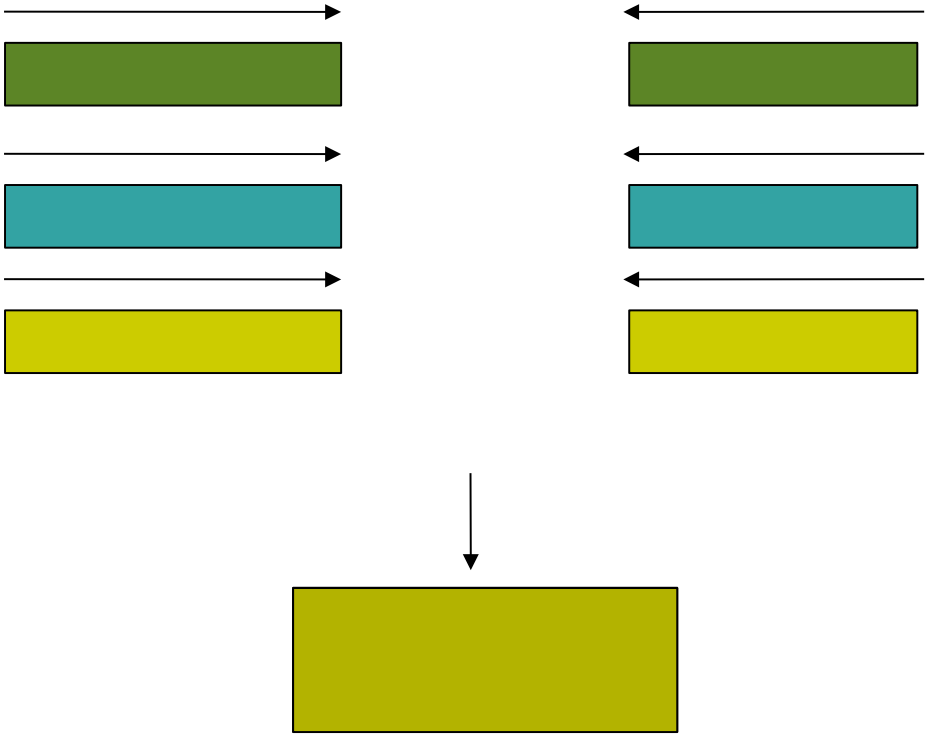




The set of homologous Sequences are aligned by 3 different aligners in forward and reverse modes (Head or Tails approach)

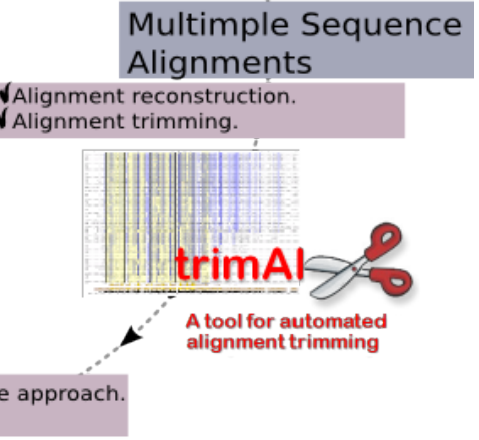
The set of homologous
Sequences are aligned by 3 different aligners in forward and reverse modes (Head or Tails approach)

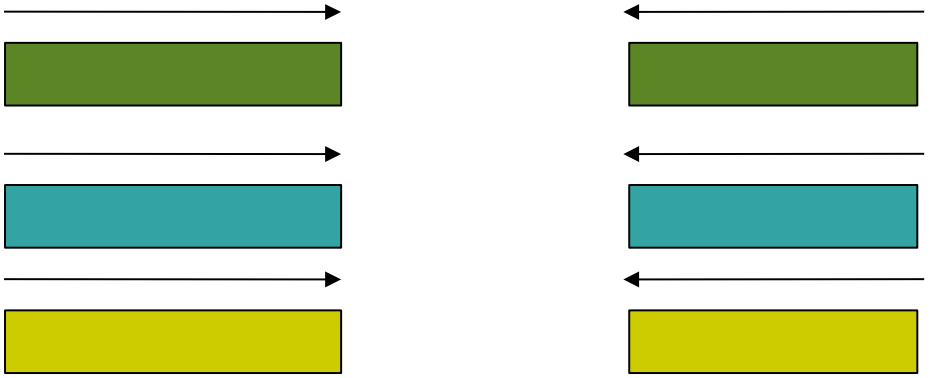
A consensus is buit



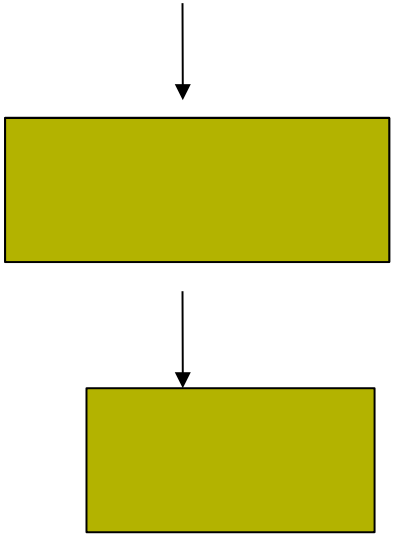
```
sw_DSBA_PSESM/1   ---MRNLIISAALVAASLFGMSAQAAEPIESGKQYV-ELTSAVPV
sw_DSBA_SALTY/1   ---MKKIWLA---LAGMVLAFSASAAQISD-GKQYI-TLDKP--V
sw_DSBA_ENTAM/3   AKWINSIFKSVVLTAALALPFTAS--AFTE-GTDYM-VLEKP---
sw_DSBA_LEGPN/1   ----------------LMPMTALATQFIE-GKDYQTVASAQ-LS
cons                                : ::*   :   *.:*

sw_DSBA_PSESM/1   AVPGK-IEVIELFWYGCPHCYAFEPTI---NPWVEKLPSDVNFVR
sw_DSBA_SALTY/1   --AGE-PQVLEFFSFYCPHCYQFEEVLHVSDNVKKKLPEGTKMTK
sw_DSBA_ENTAM/3   -IPDADKTLIKVFSYACPFCYKYDKAVT--GPVADKVADLVTFVP
sw_DSBA_LEGPN/1   TNKDKTPLITEFFSYGCPWCYKIDAPLN--D-WATRMGKGAHLER
cons                .    : :.* : ** **  :  :   .   :: . . :
```





The set of homologous Sequences are aligned by 3 different aligners in forward and reverse modes (Head or Tails approach)

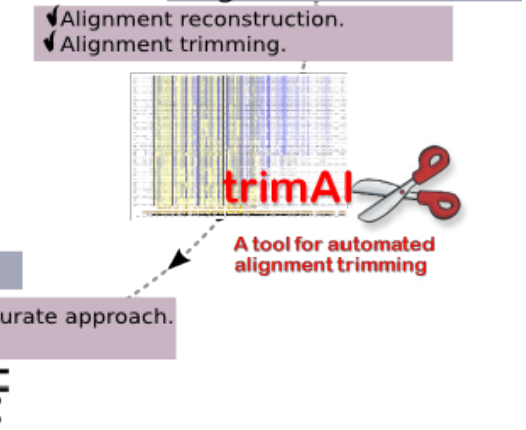


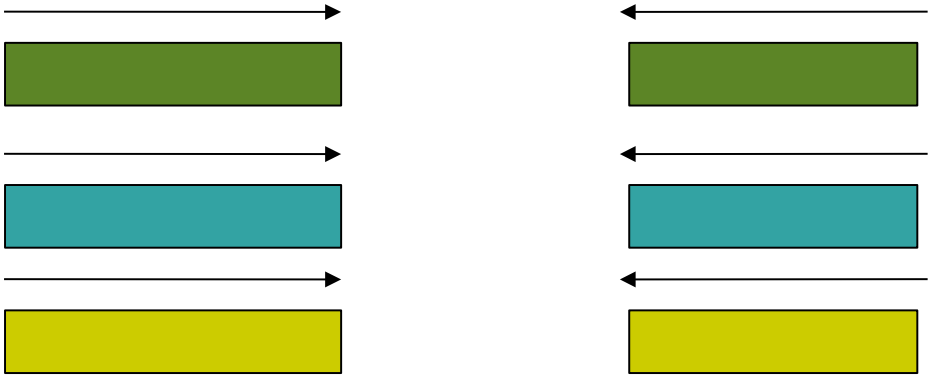
A consensus is buit

The consensus is trimmed (trimAl) based on:

- consistency across the 6 alignments
- gap content

Http://trimal.cgenomics.com



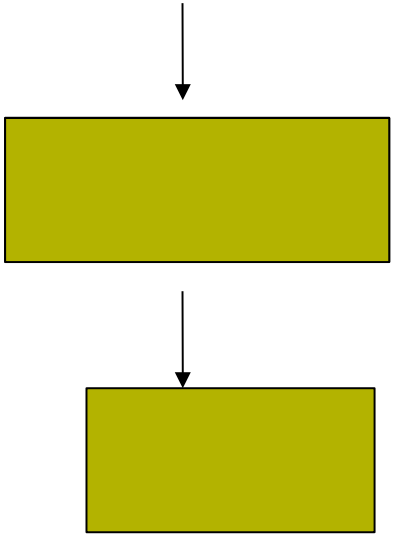


The set of homologous Sequences are aligned by 3 different aligners in forward and reverse modes (Head or Tails approach)

A consensus is buit

The consensus is trimmed (trimAl) based on:

- consistency across the 6 alignments
- gap content



Http://trimal.cgenomics.com

Our pipeline:

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

PHYLOME

ETE: Environment for Tree Exploration
ete.cgenomics.org

for every gene

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

Multimple Sequence Alignments
- Alignment reconstruction.
- Alignment trimming.

trimAl
A tool for automated alignment trimming

Salvador-Capella et al
Bioinformatics (2009).

http://trimal.cgenomics.org

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].

Pipeline described in Huerta-Cepas et al NAR (2011)

Search in PhylomeDB

(i.e. ENSG00000139618, YBL058W, TP53)

Search

RandomTree!

BLAST search

Latest Phylomes

Arxula adeninivorans	2014
Beta vulgaris	2013
Clogmia albipunctata	2013
Penicillium digitatum	2012
Schistosoma mansoni	2012

see all phylomes

PhylomeDB uses

trimAl A tool for automated alignment trimming · Jalview a multiple alignment editor

PhylomeDB cross linking

UniProt · Ensembl · SGD · CGD · Génolevures · ACYPICYC · Treefam

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

New Zygomycete phylome: the human pathogen Lichtheimia corymbifera

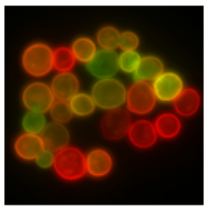
Mon, 09/15/2014 - 21:09

PhylomeDB extends its repertoire of fungal phylomes with that of a genome of a poorly sample clade, that of the basal group zygomycetes. In this case the phylome (245) of the human pathogenic mucorales Lichtemia corymbifera has served to reveal extensive past gene duplications in this group. Lichtheimia species are the second most important cause of mucormycosis in Europe. The sequencing of its genome and the comparison with other Zygomycete species, particularly of Rhizopus delemar, the main

Popular Phylome Collections

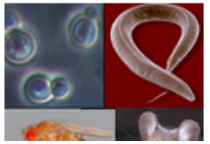
Human

Fungi



Plants

Model Species



Latest News

New Zygomycete phylome: the human pathogen Lichtheimia corymbifera

Mon, 09/15/2014 - 21:09

New hemyascomycete phylome: Blastobotrys (Arxula) adeninivorans, a yeast of biotechnological interest.

Mon, 05/19/2014 - 11:01

Help us to improve PhylomeDB: complete our survey.

Thu, 02/20/2014 - 16:22

show all

PhylomeDB Twitter

Tweets Follow

phylomedb @phylomedb 23 Oct

New birds, crocs, and fungal phylomes to come soon at phylomeDB. stay tuned!

Expand

phylomedb @phylomedb 15 Sep

New Zygomycete phylome: the human fungal pathogen Lichteimia corymbifera phylomedb.org/?q=node/537

Expand

phylomedb @phylomedb 26 Aug

NOTICE: PhylomeDB will be down due to MAINTENANCE

Search in PhylomeDB

(i.e. ENSG00000139618, YBL058W, TP53)

Search

RandomTree!

BLAST search

Latest Phylomes

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

see all phylomes

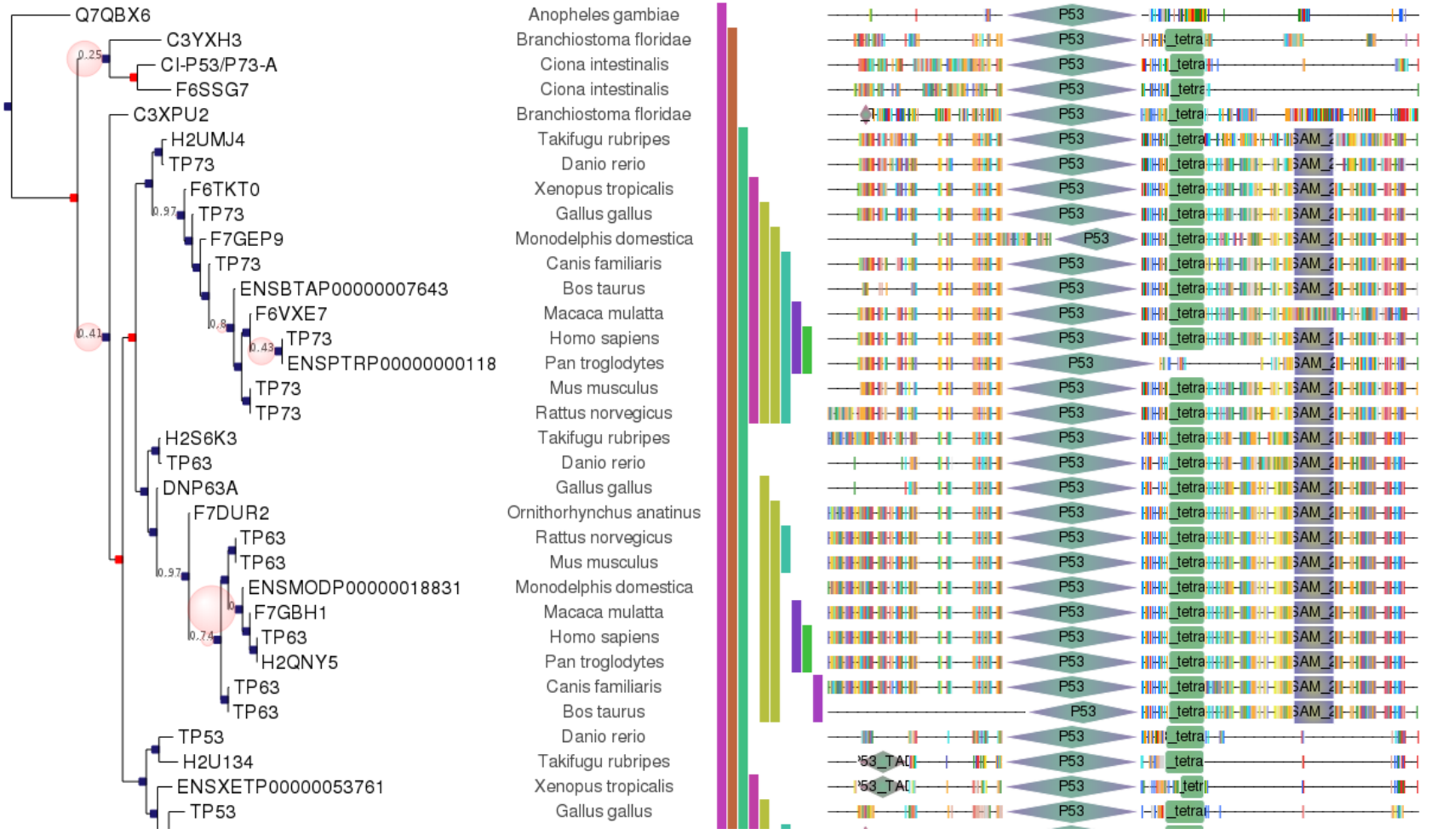
PhylomeDB uses

trimAl Jalview

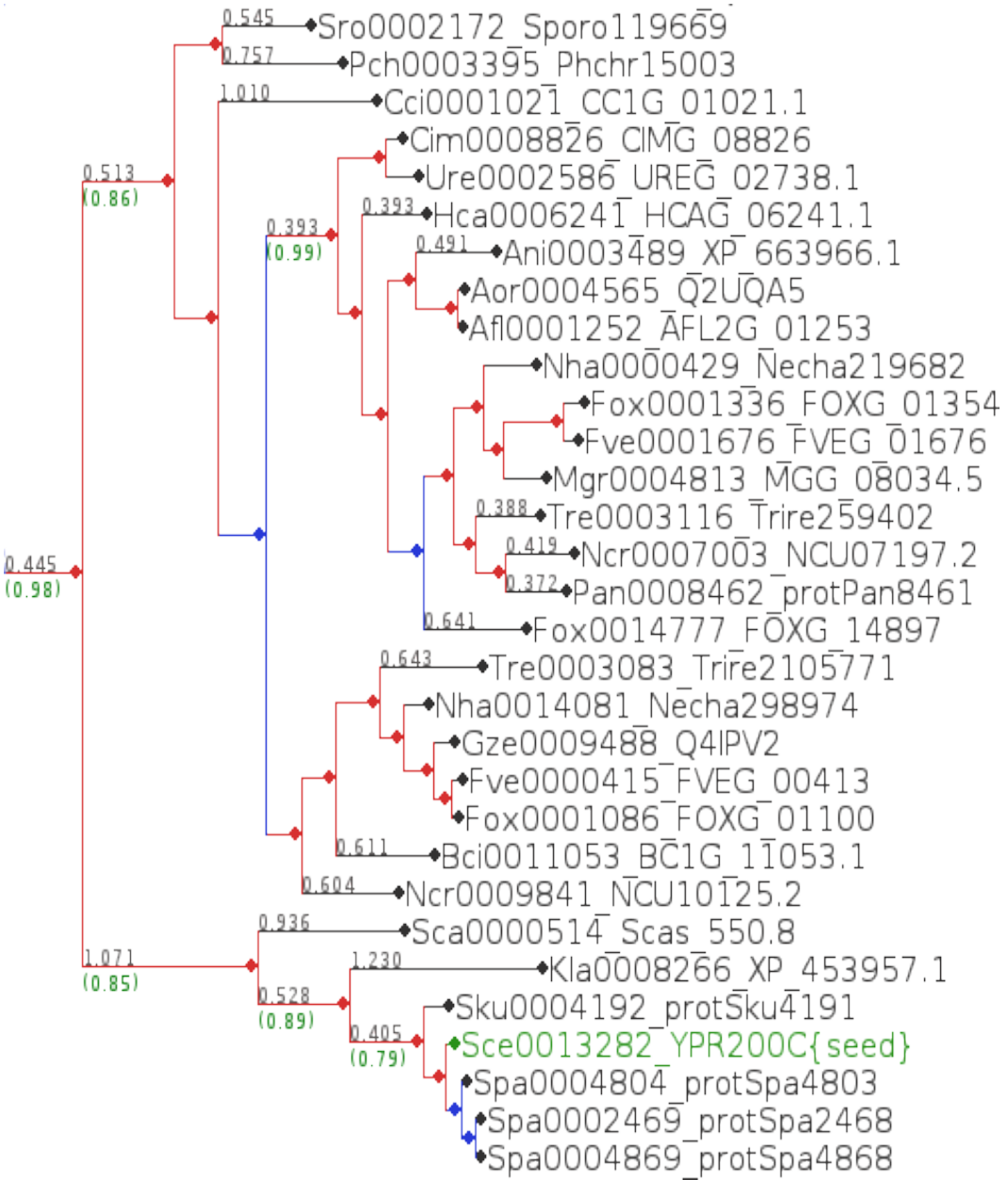
TP53 tree in phylome 218

AS seed in Rat phylome | JTT (lk:-18130.4) | -- in collateral trees --

Tree features | Search | Clear search | Image | Hard link | Download OrthoXML | See alignments | Download data.tar.gz



These phylomes can now be interrogated in many ways



- Families that show a particular topology
-Detect and date duplication events

- Genes that have accelerated evolutionary rates at a particular lineage (positive/relaxed selection)

- Families expanded at particular lineages

- Footprints of horizontal gene transfer, lineage sorting, gene conversion and other evolutionary processes

- Search for co-evolving genes

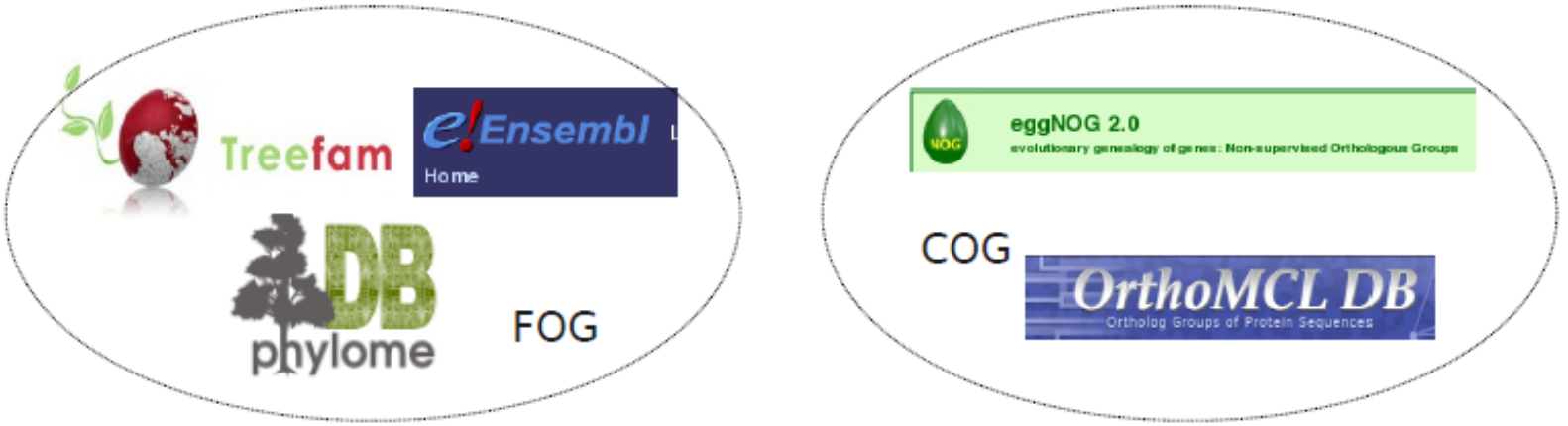
- predict functional properties

- across-species prediction of orthology and paralogy

-

MetaPhOrs

(Meta-Phylogeny-Based-Orthologs)



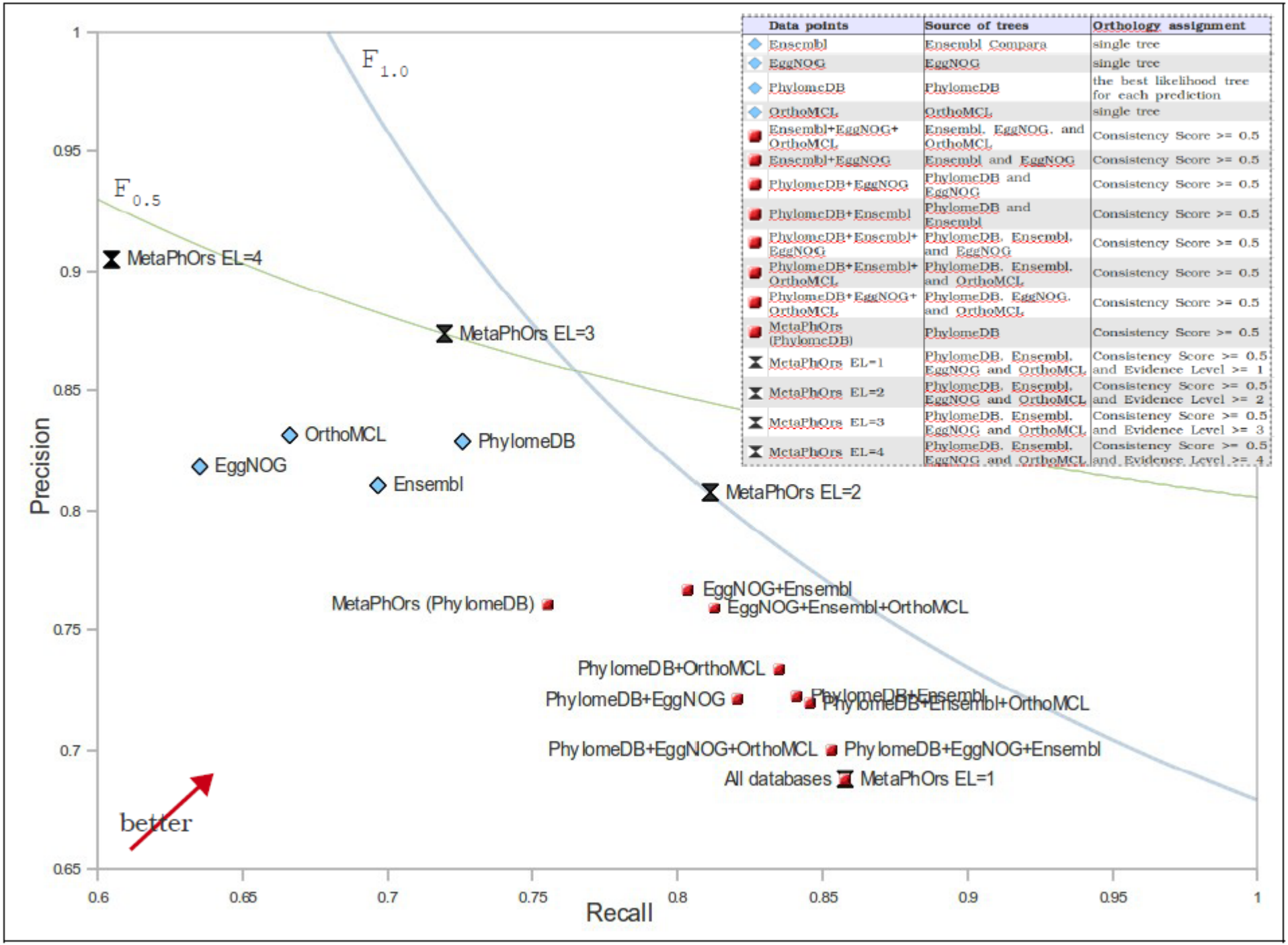
Use existing tree repositories

Reconstruct trees for orthologous groups

Integrate and use consistency across datasets as a proxy of reliability

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

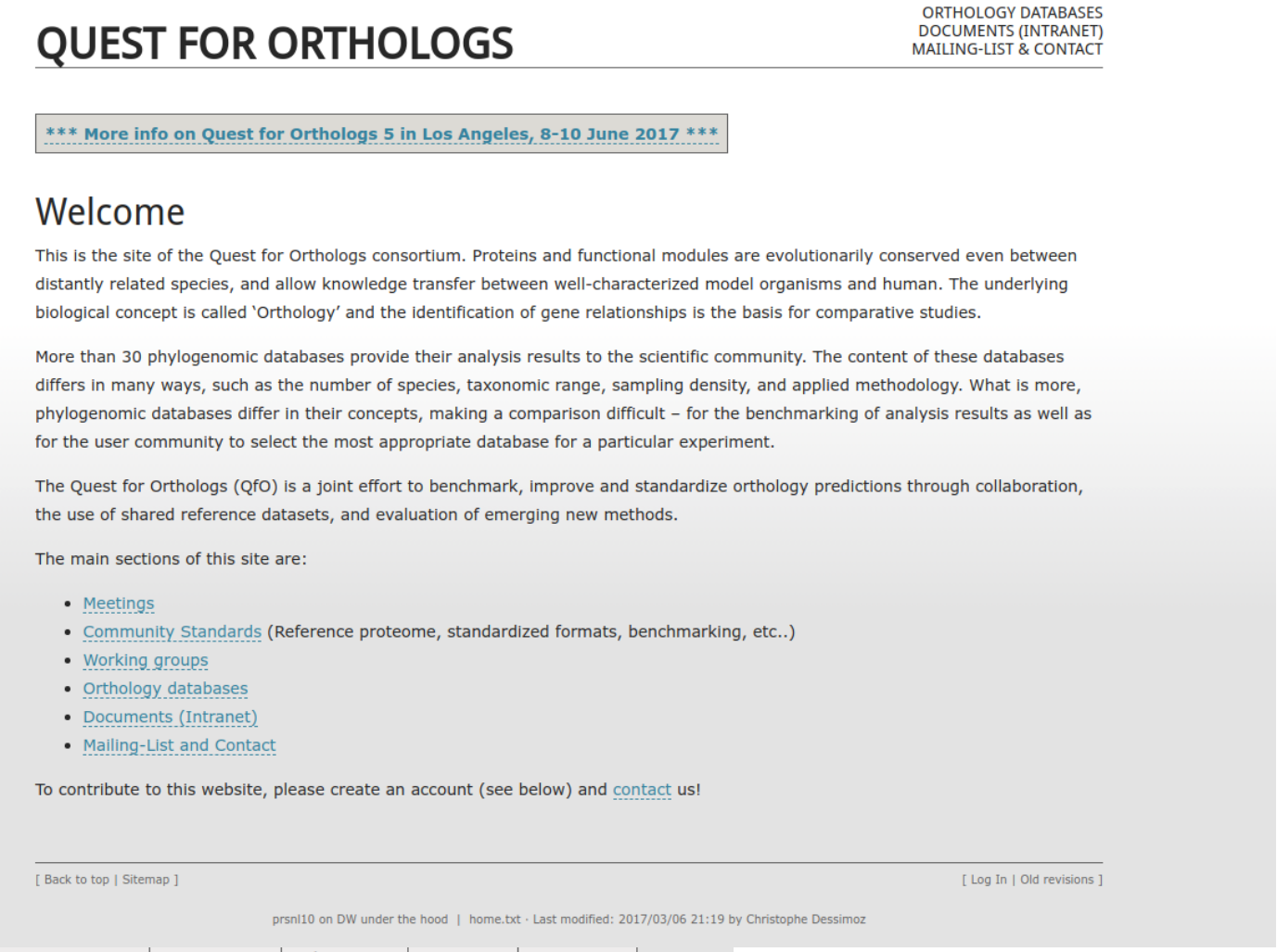
Pryszcz et. al. (NAR, 2011)



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

A plethora of methods for ortholog prediction

http://questfororthologs.org



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

QUEST FOR ORTHOLOGS

List of orthology databases

If you know of any other database, please edit this page directly o
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

¿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?)

Final warnings:

Most methods assume the complete, fully (and correctly) annotated genome for each of the compared species is available.

Deserve special considerations:

- Working with highly fragmented/incomplete genomes or transcriptomes

- Working with bacteria (pangenome concept, rampant HGT)

The “boundaries” of the orthology concept

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)."

The “boundaries” of the orthology concept

*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).*

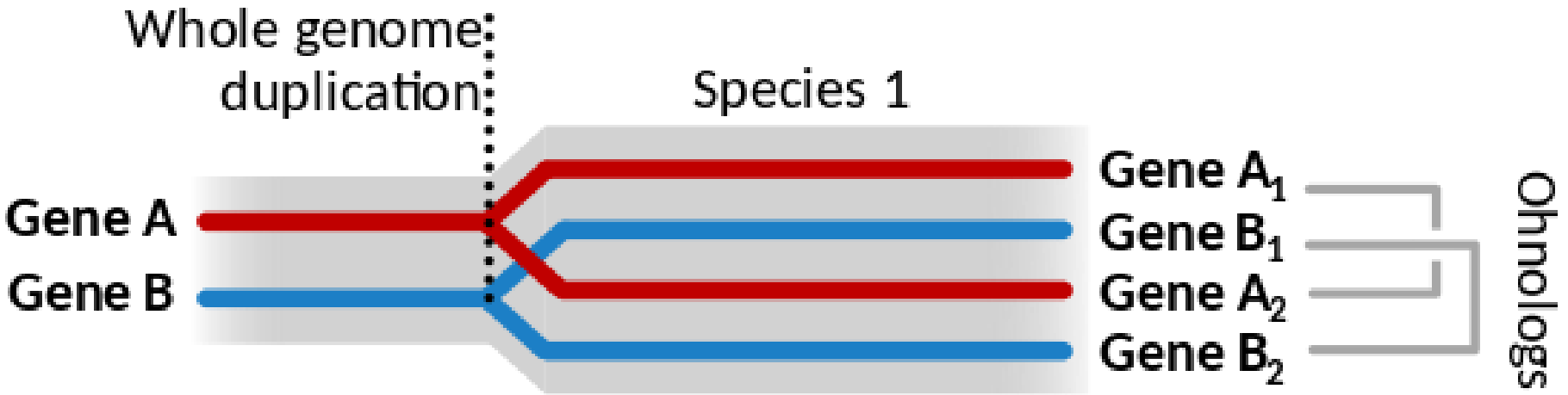
*Where the homology is the result of **speciation** so that the history of the **gene** reflects the history of the **species** (for example alpha hemoglobin in man and mouse) the genes should be called orthologous (ortho = exact)."*

GENE, SPECIATION, DUPLICATION

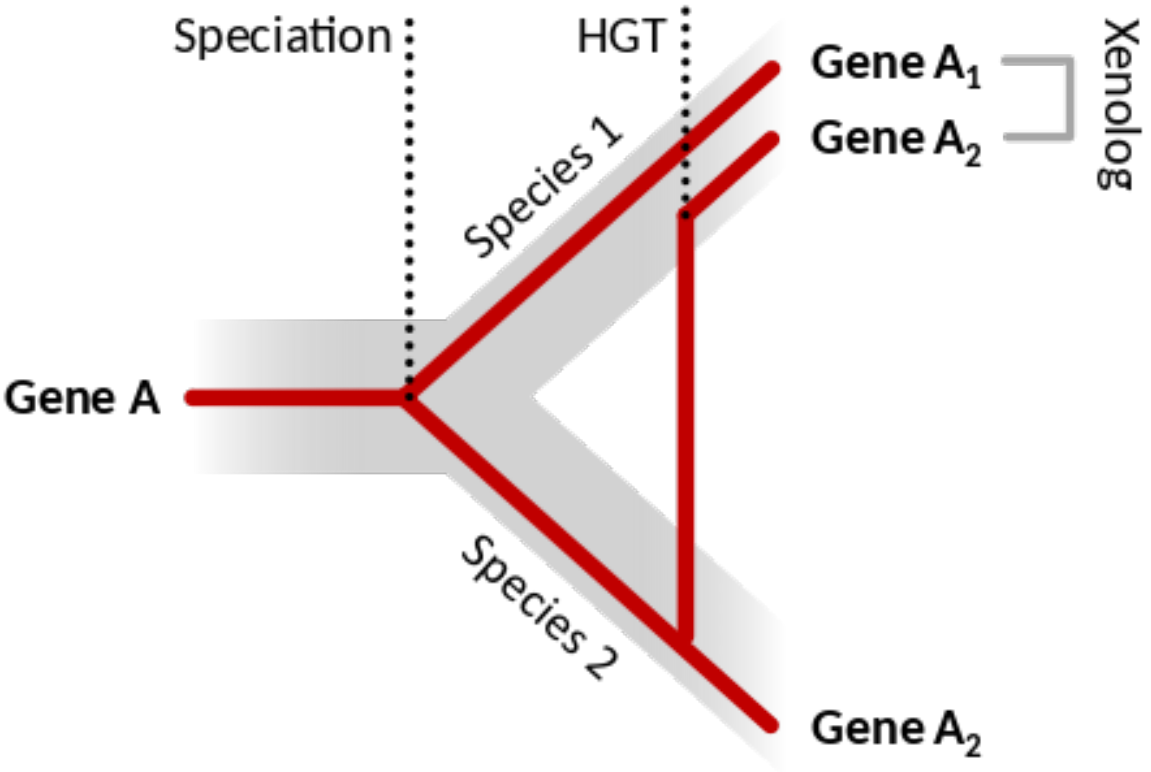
Can we accomodate orthology to evolutionary processes other than speciation and duplication?

Ohnologs, Xenologs, Homeologs

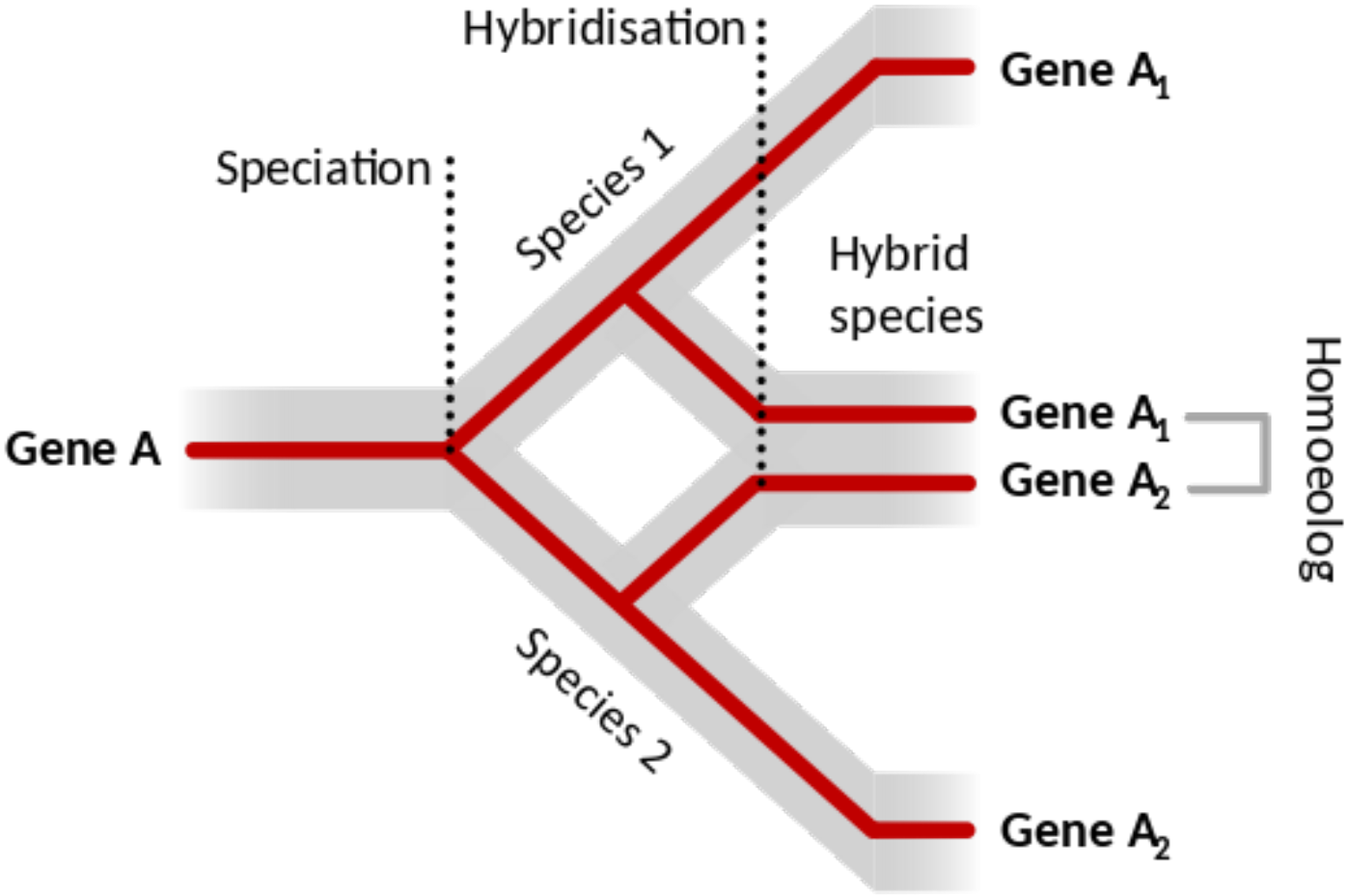
Ohnologs, Xenologs, and Homeologs



Ohnologs, Xenologs, and Homeologs



Ohnologs, Xenologs, and Homeologs

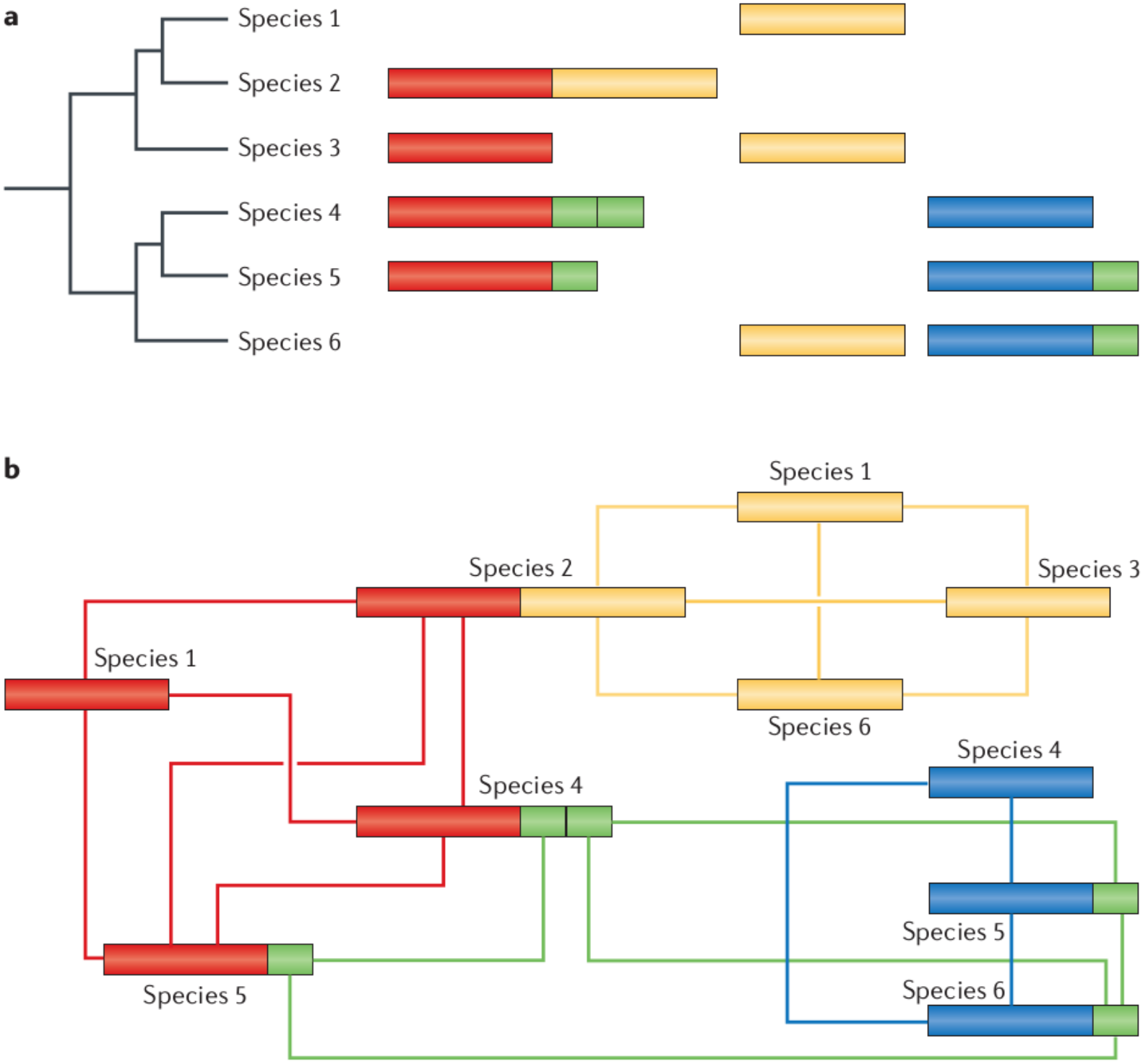


Can orthology be defined beyond genes?

In principle, the concept can simply be extrapolated to any loci that involves through duplication and speciation, but where to set the level of resolution?

Domains?
Single nucleotides?

Box 2 | **Units of orthology**



And what about the species boundary?

Two alleles of the same gene segregating in a population are diverging from each other (they are clearly homologs) but there is no speciation event separating them, they are still “the same gene in the same species”.. however, they can diverge and even change their chromosomal location

This can be particularly problematic in microbial organisms (pan/core genome, rampant gene flow, etc).

Questions?