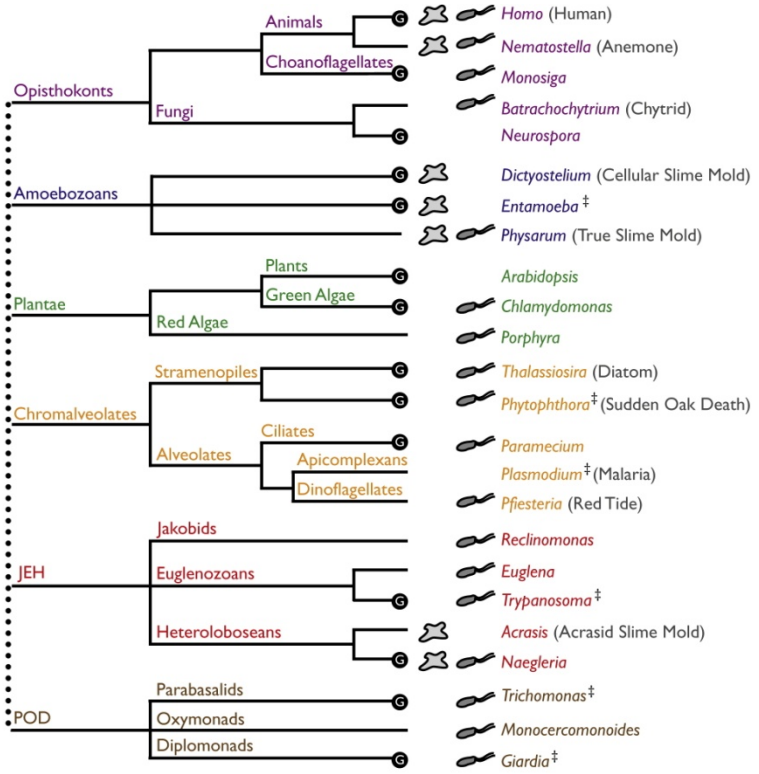
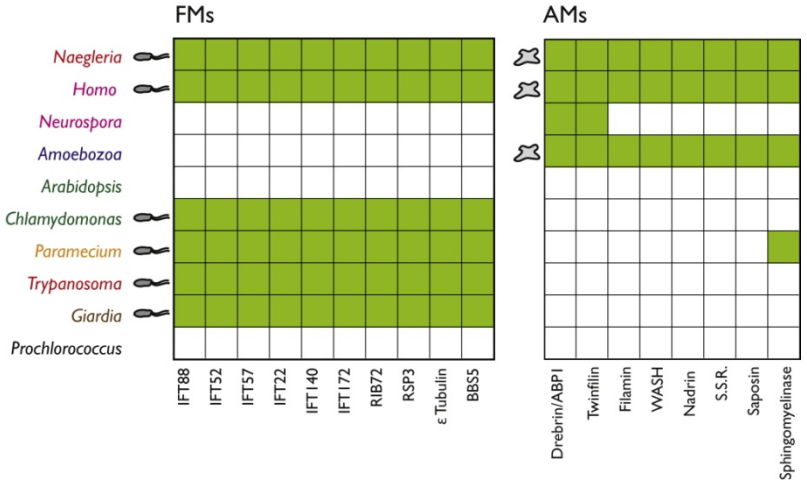
Phylogenetic trees & orthology



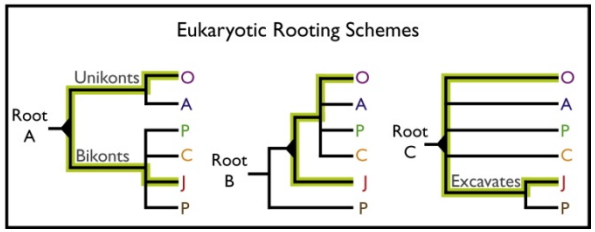


Amoeboid locomotion

Flagellar apparatus

Genome compared

‡ Parasitic



- Med11 vs kinases: orthology
- Trees are useful beyond that: HGT, timing of duplication, study of all kinds of evolutionary processes

Fritz-Laylin et al. cell 2010

Gene Trees, Gene Duplications, and Orthology

- How to make trees
- Bootstrap
- Interpreting trees
- duplications vs speciations vs loss, timing of duplications, HGT
- Orthology
- Duplications before LECA
- Endosymbiosis

Phylogenetic gene trees: how to make them

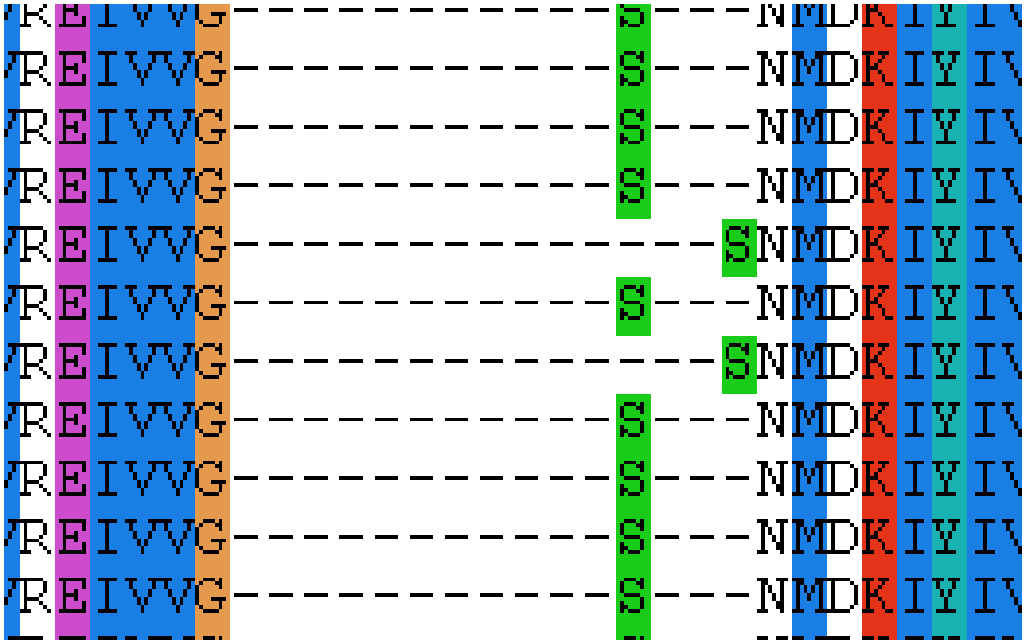
- Homology: *are* two pieces of sequence related; Trees: when did they diverge (*how* are they related)
- Start from a multiple sequence alignment
- All multiple sequence programs alignments make a global alignment, thus feed it regions that you know are homologous → Domains !
- MUSCLE / clustal / t_coffee /**MAFFT**
- Visual inspection of alignments (gaps, fragments/complete sequences, weird things e.g. A)

Put homologs in the alignment

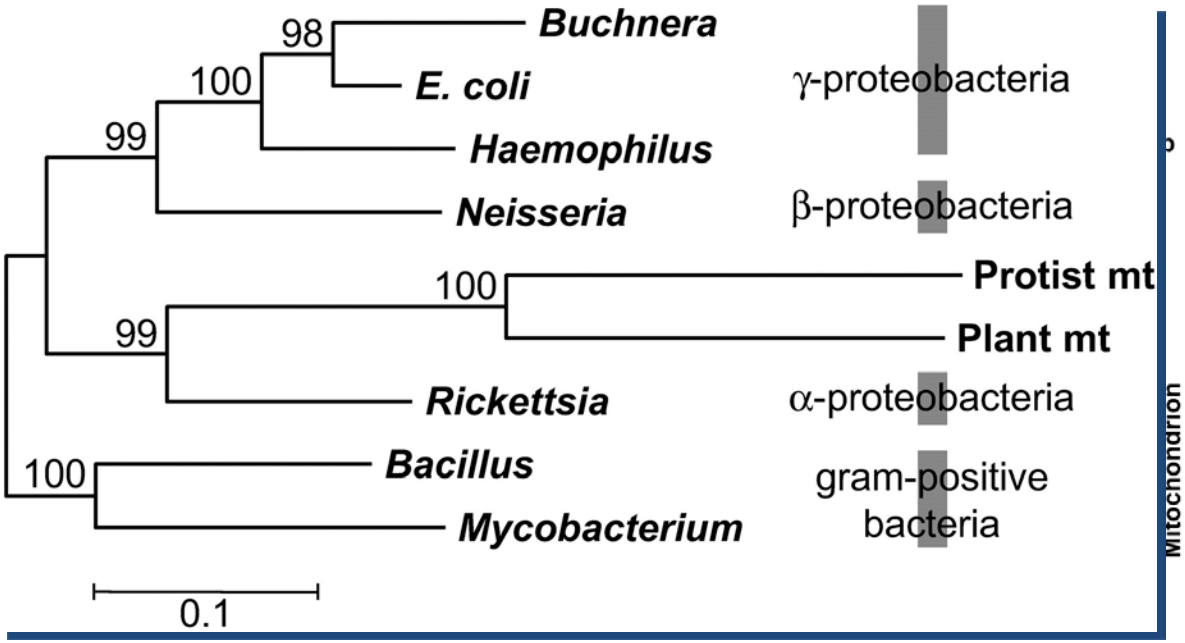
- Even if they are not homologous alignment programs will align them (muscle/clustal/mafft implicitly “assumes” that the sequences you feed it are homologous)

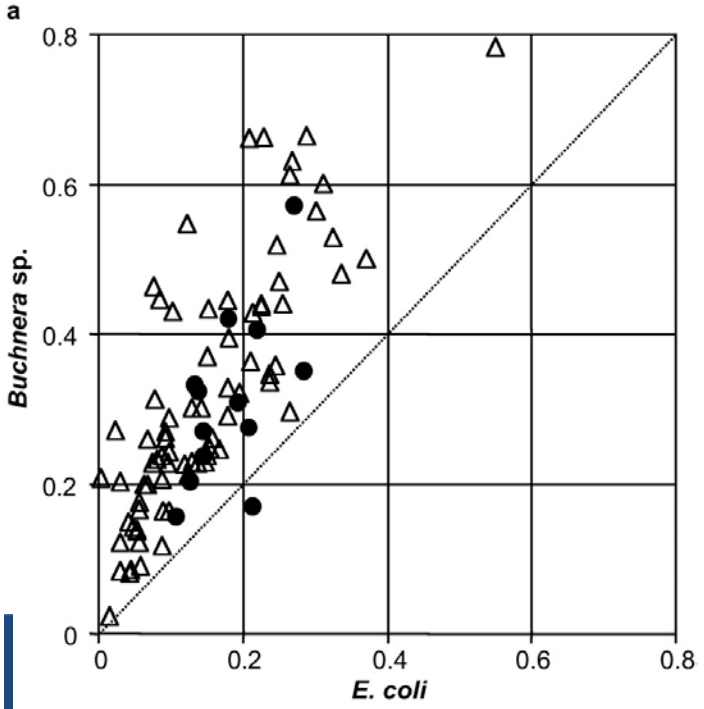
- And in a phylogeny program, non-homologous sequences ***will be*** clustered

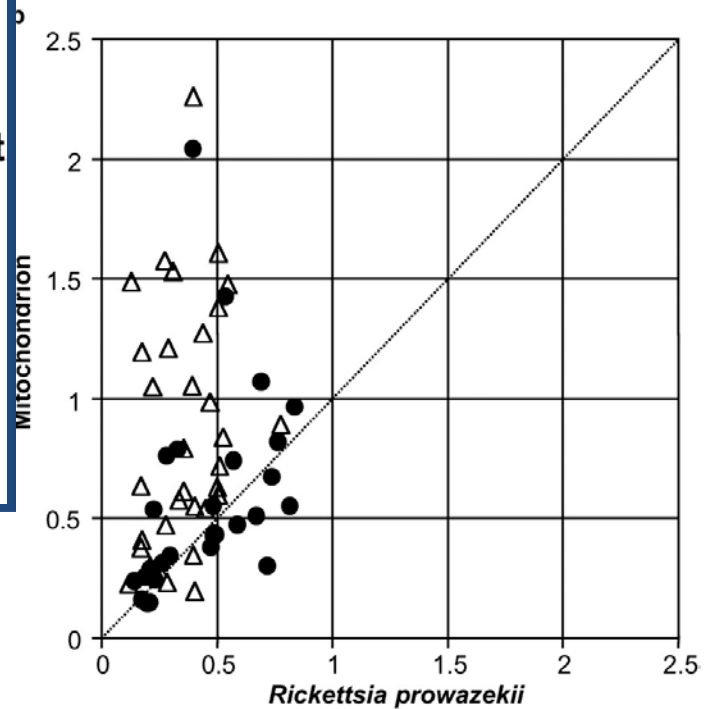
Visual inspection of alignments: ?!



Unequal rates between species are a very real phenomenon







Character based: parsimony and maximum likelihood

- Two way classification in phylogeny distance based vs character based
- character state method. Searches “directly” (i.e. without defining distances) for a tree that fits best to the data (the alignment)

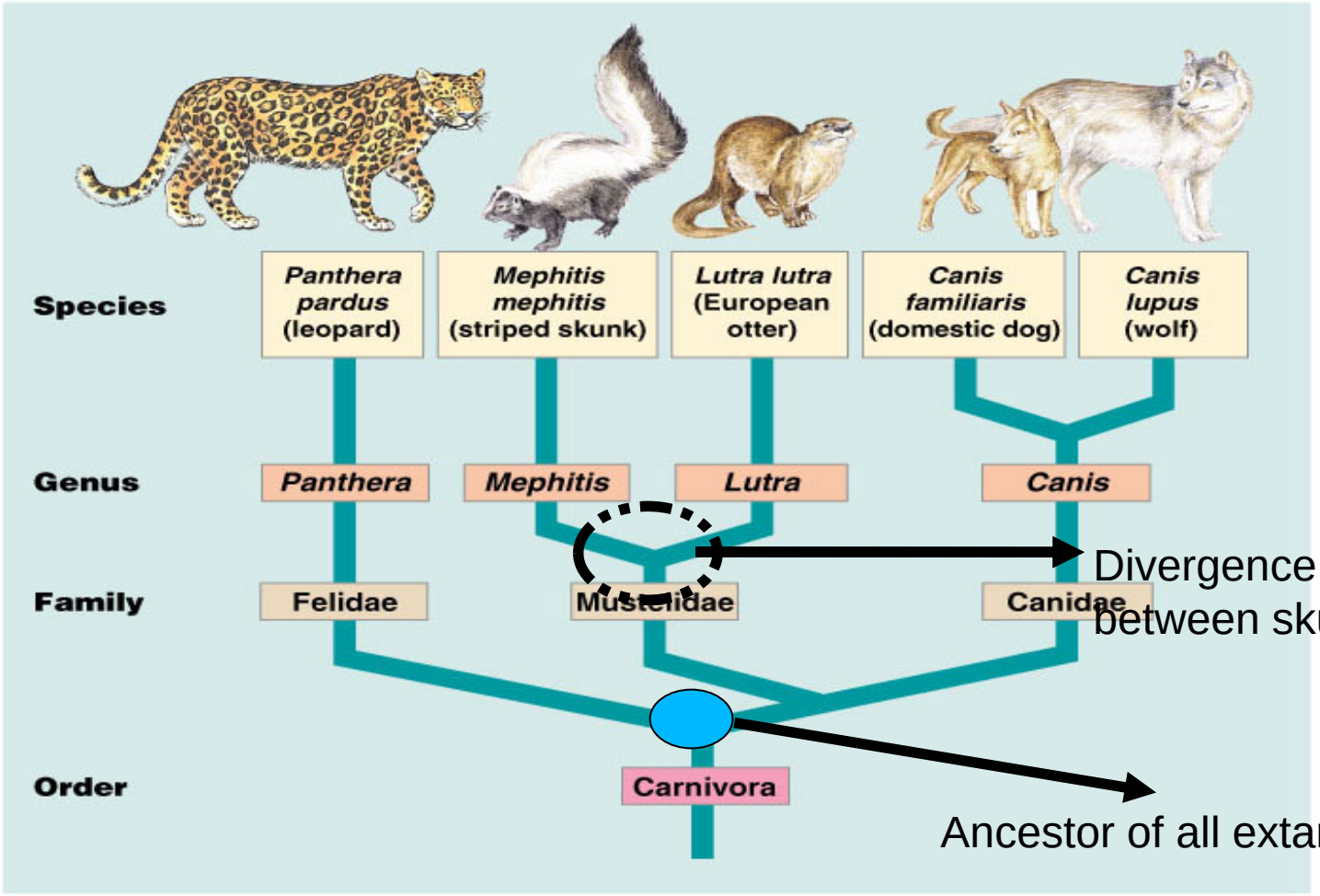
Maximum likelihood

- have to specify a model of sequence evolution
- likelihood for all sites is the product of the likelihoods for individual sites **assuming** all the sites evolve independently.
- maximum likelihood method computes the probabilities for all possible combinations of ancestral states!
- ML methods evaluate phylogenetic hypotheses in terms of the probability that a proposed **model** of the evolutionary process and the proposed unrooted tree (**hypothesis**) would give rise to the observed **data** (the alignment). The tree found to have the highest (log)ML value is considered to be the preferred tree.
- **Currently: RaxML and WAG**

Gene Trees, Gene Duplications, and Orthology

- How to make trees
- Bootstrap
- Interpreting trees
- duplications vs speciations vs loss, timing of duplications, HGT
- Orthology
- Duplications before LECA
- Endosymbiosis

Phylogenetic tree: historical pattern of relationships among organisms: interpretation of a tree

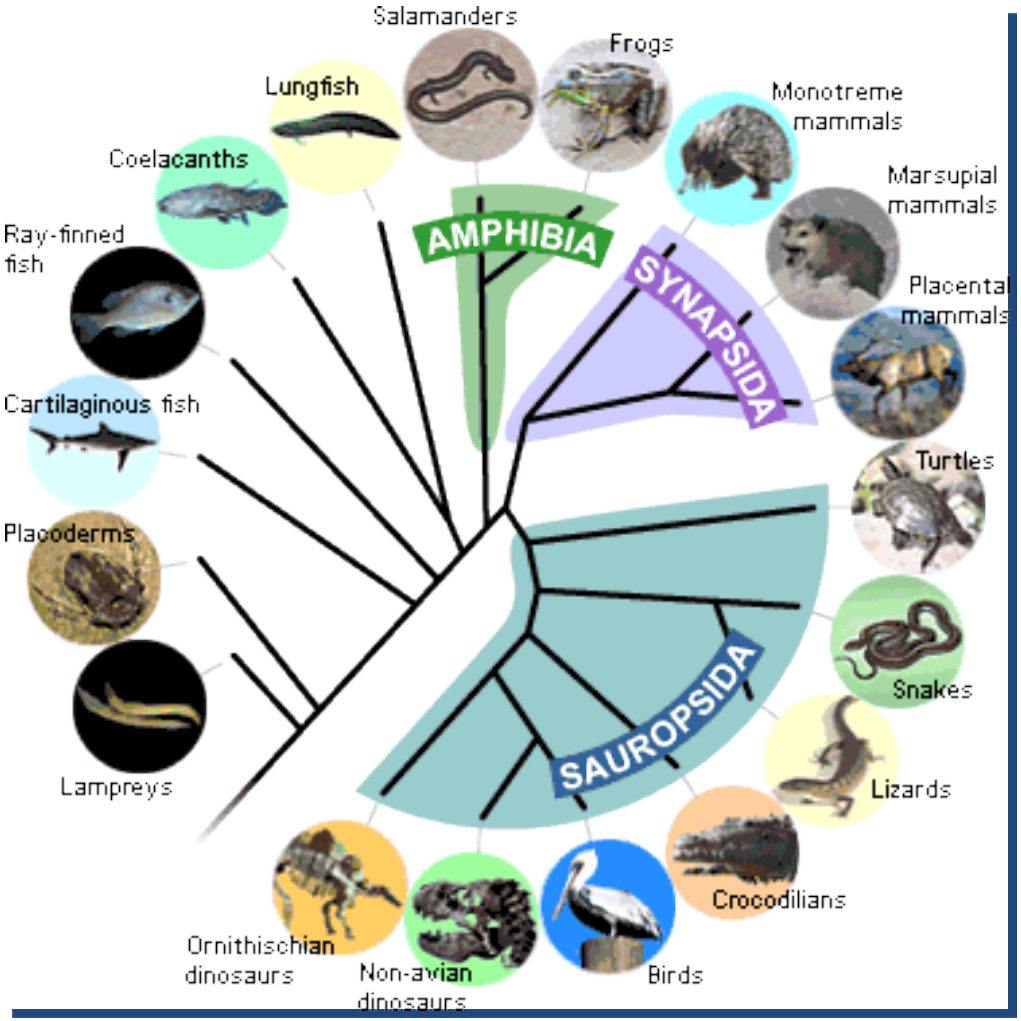


NB still no information in skunk left / otter right

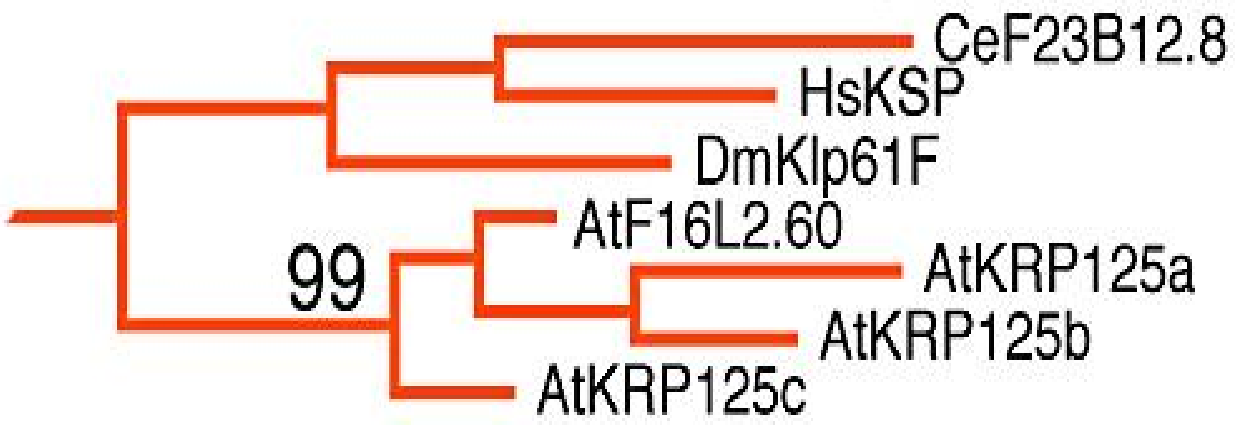
Interpreting the tree

- Taxonomic findings

- Paraphyly
- Monophyly



Simple example (kinase)



What are the nodes?

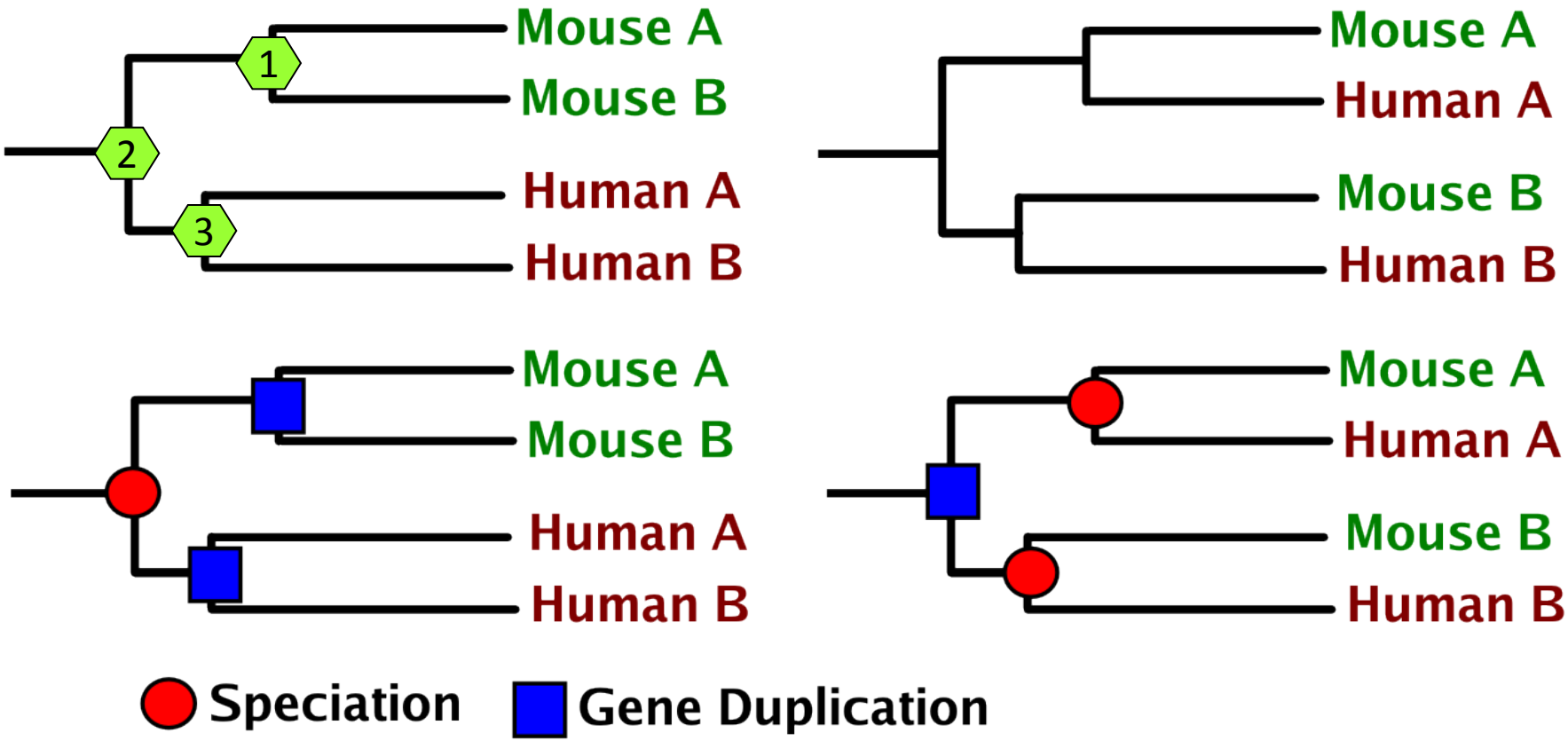
Gene Trees, Gene Duplications, and Orthology

- How to make trees
- Bootstrap
- Interpreting trees
- duplications vs speciations vs loss, timing of duplications, HGT
- Orthology
- Duplications before LECA
- Endosymbiosis

Two genes per species: how to differentiate between one ancient or two recent duplications?

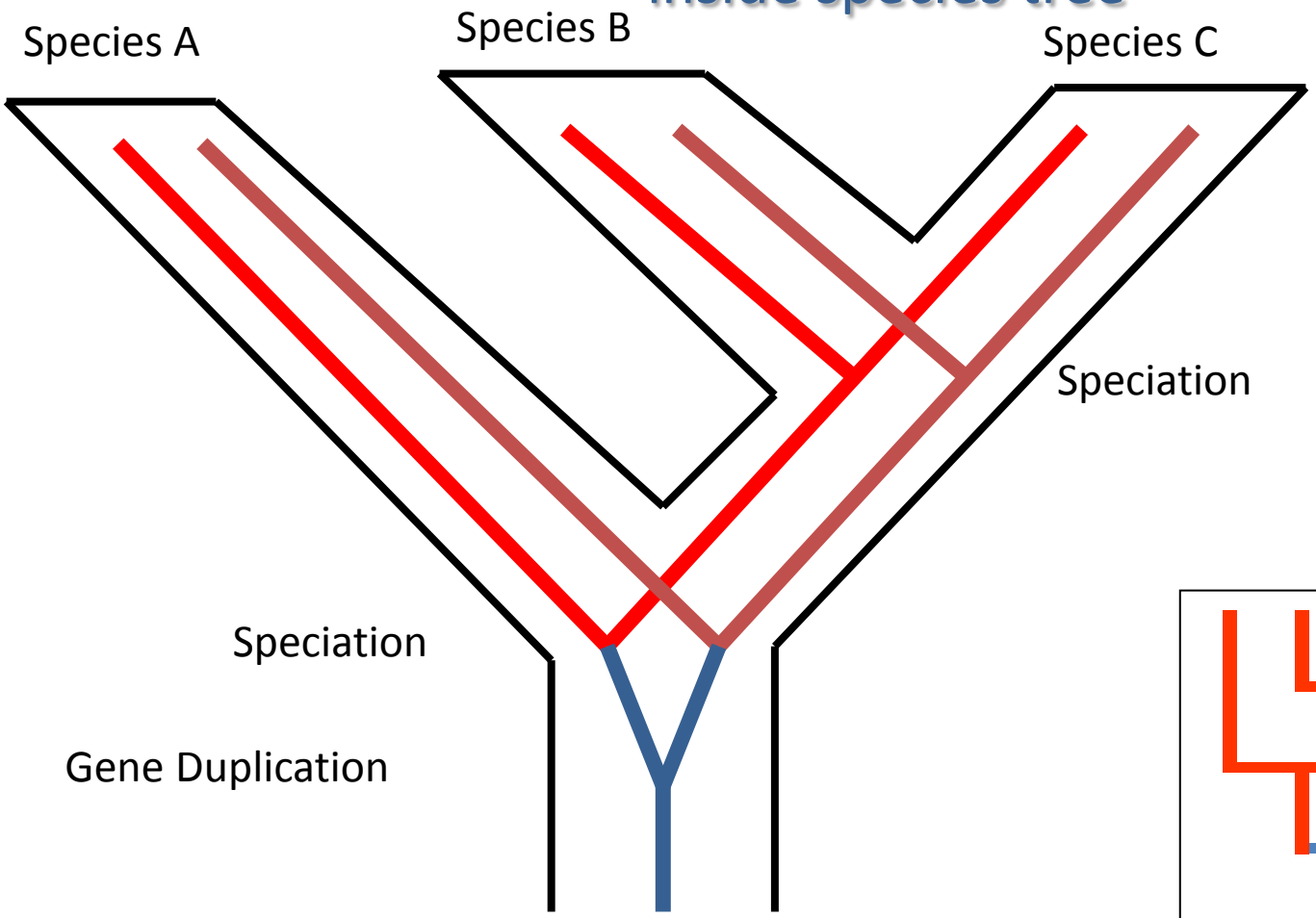
- Two genes in Human chromosomes (human A & Human B) & two genes in mouse chromosomes (Mouse A & Mouse B)

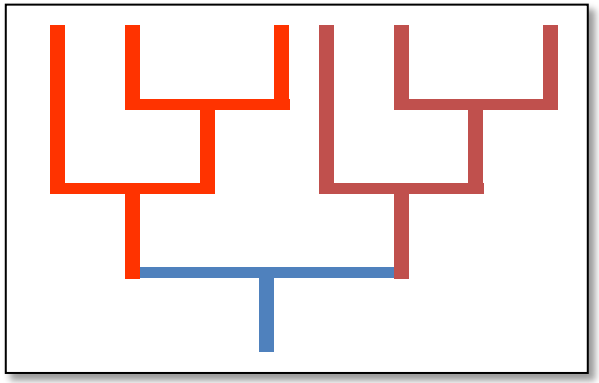
Duplications, Speciations



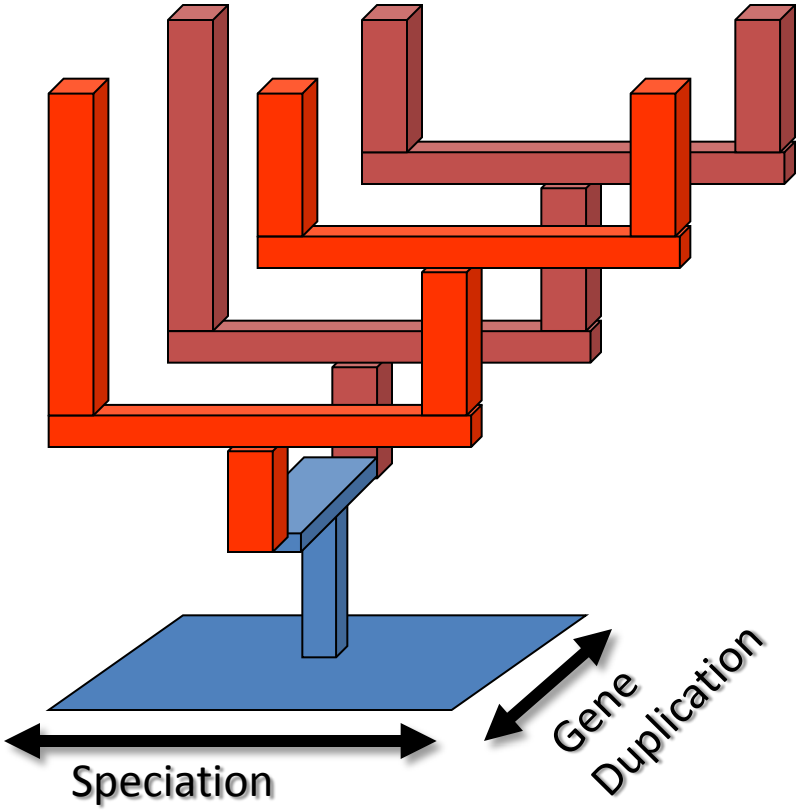
?

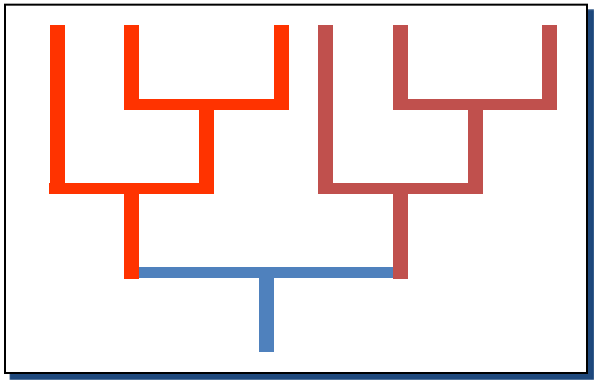
Interpreting the tree: duplications vs speciations, gene tree inside species tree



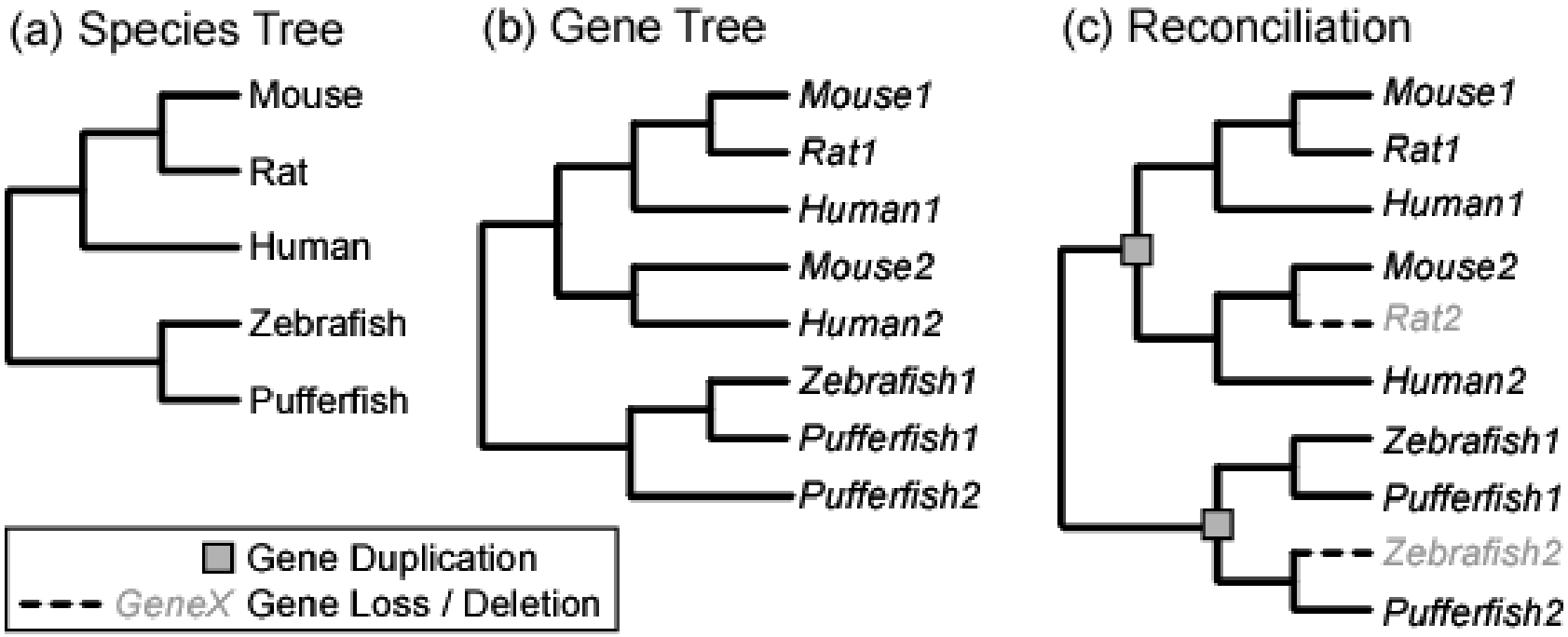


Interpreting the tree: duplications vs speciations, going pseudo 3D

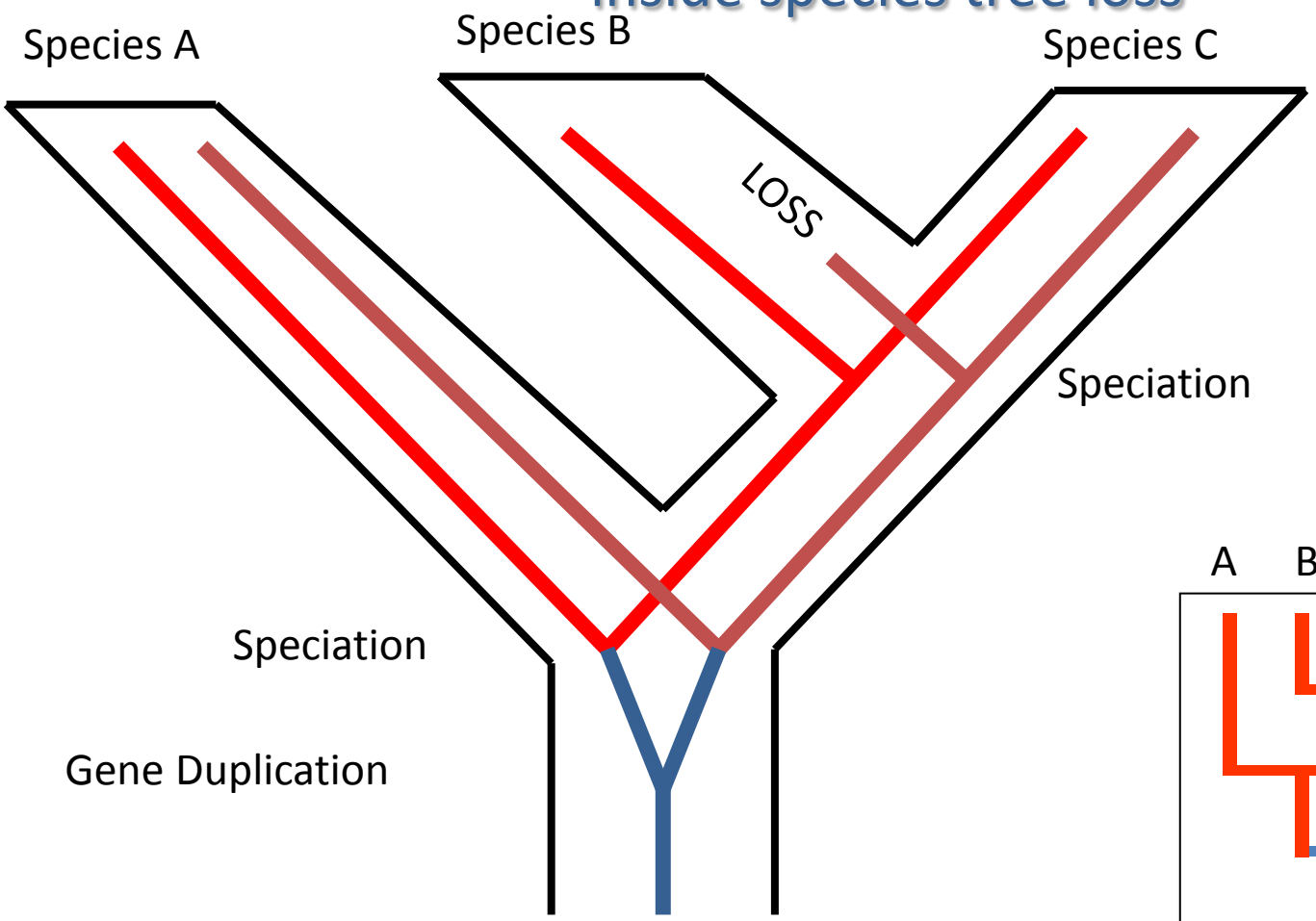


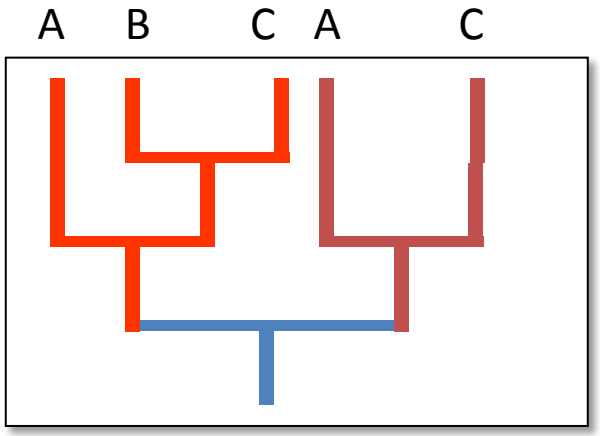


Interpreting the tree: gene trees vs species trees

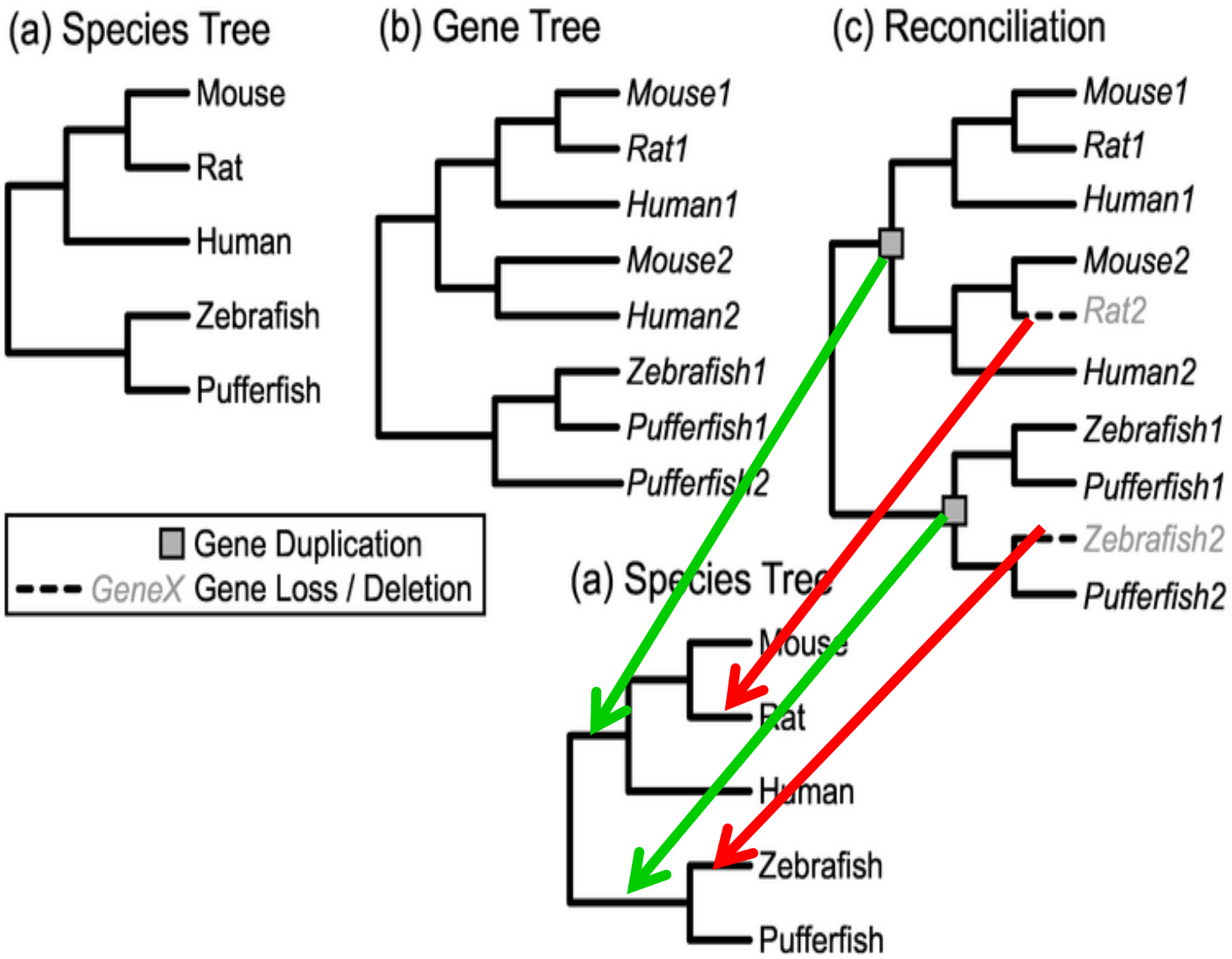


Interpreting the tree: duplications vs speciations, gene tree inside species tree loss



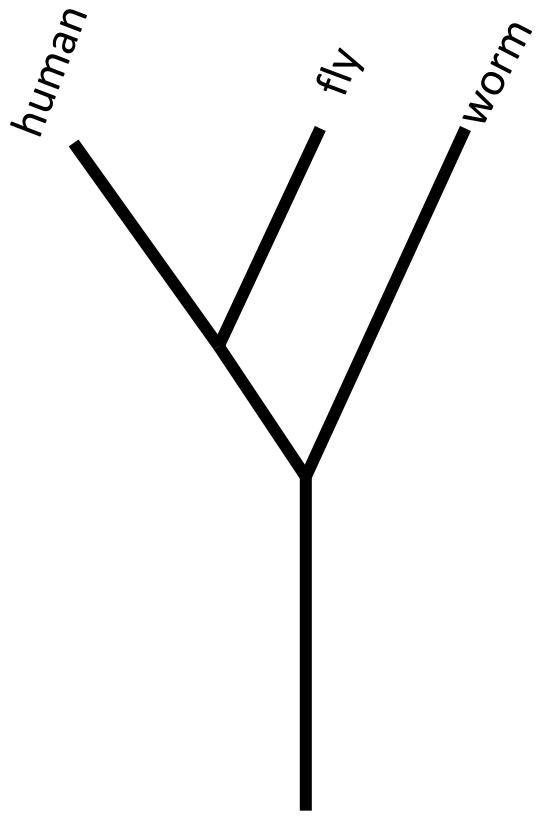


Tree reconciliation gives gene family dynamics on a species tree

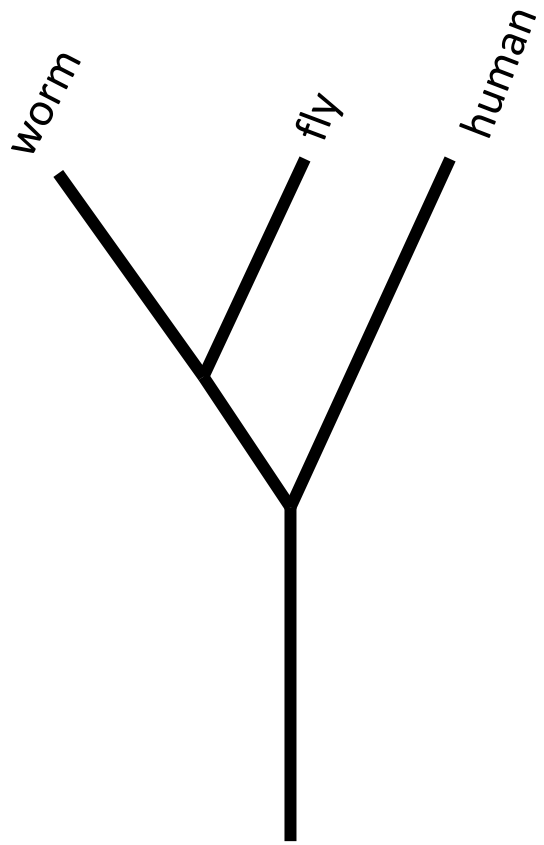


Annotating a gene tree in terms of duplications and losses, assuming vs not assuming a species tree AND/OR assuming the gene tree is wrong

Gene tree

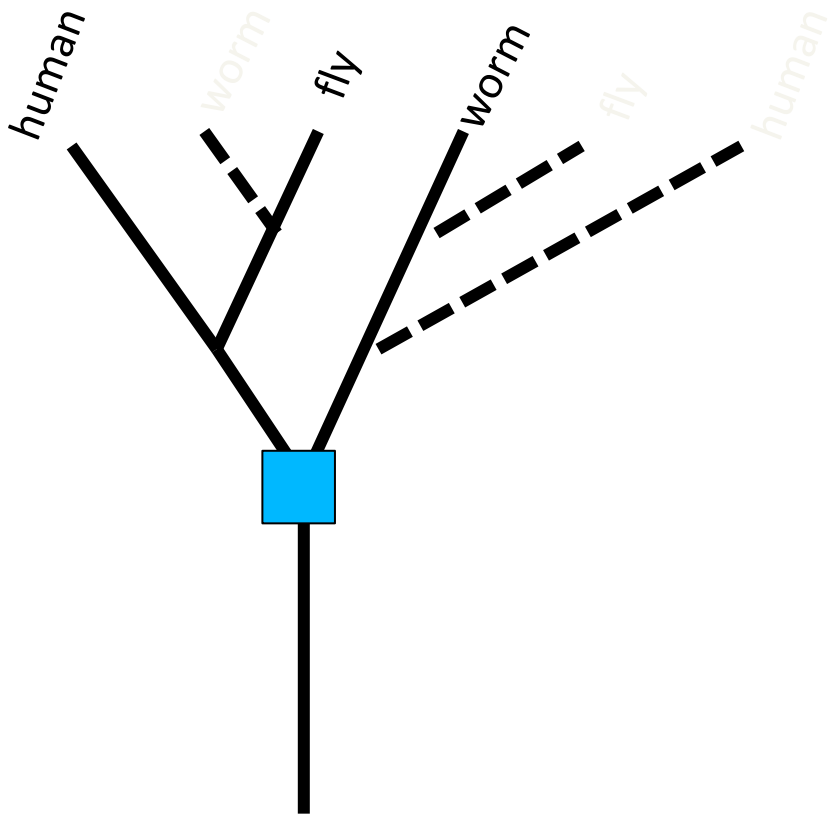


Species tree

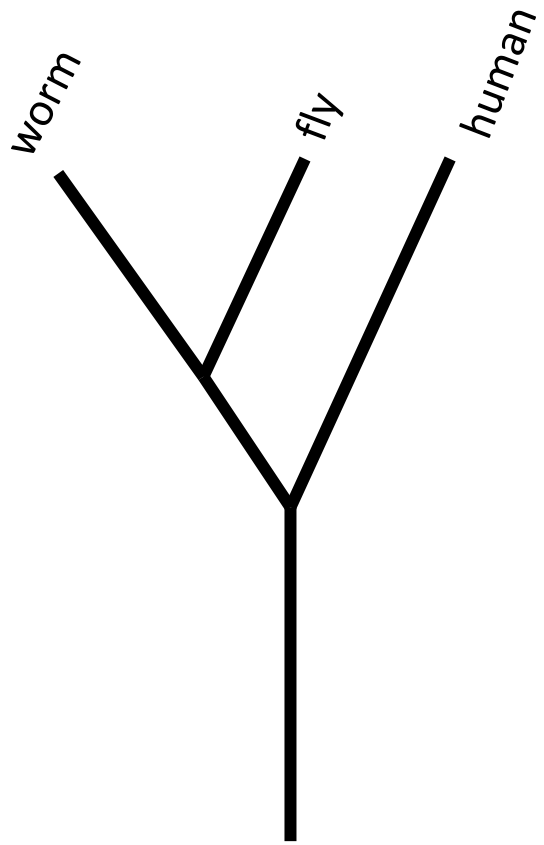


Strict reconciliation?

Reconciled
Gene tree

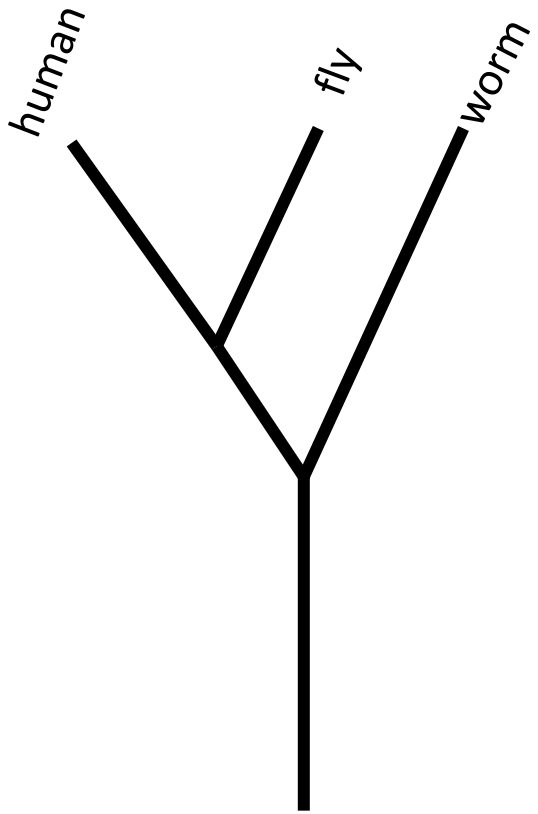


Species tree

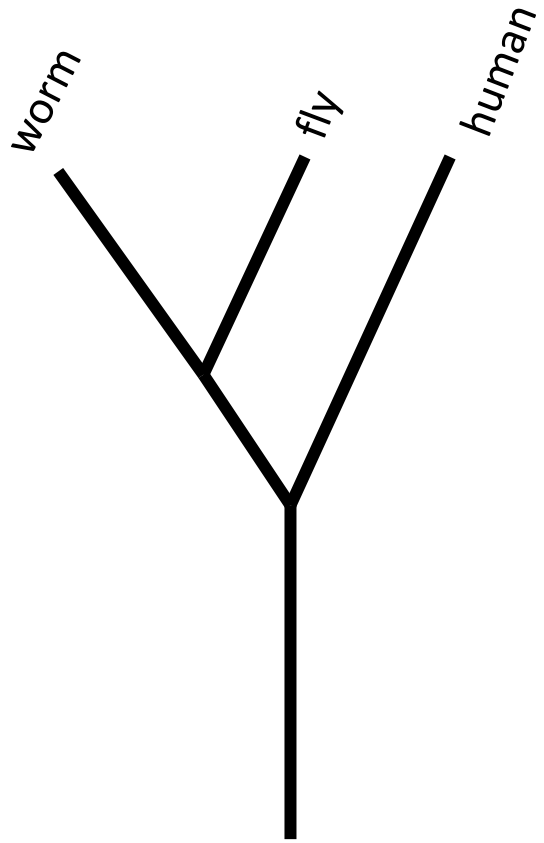


Two similar solutions: species tree guided tree reconstruction

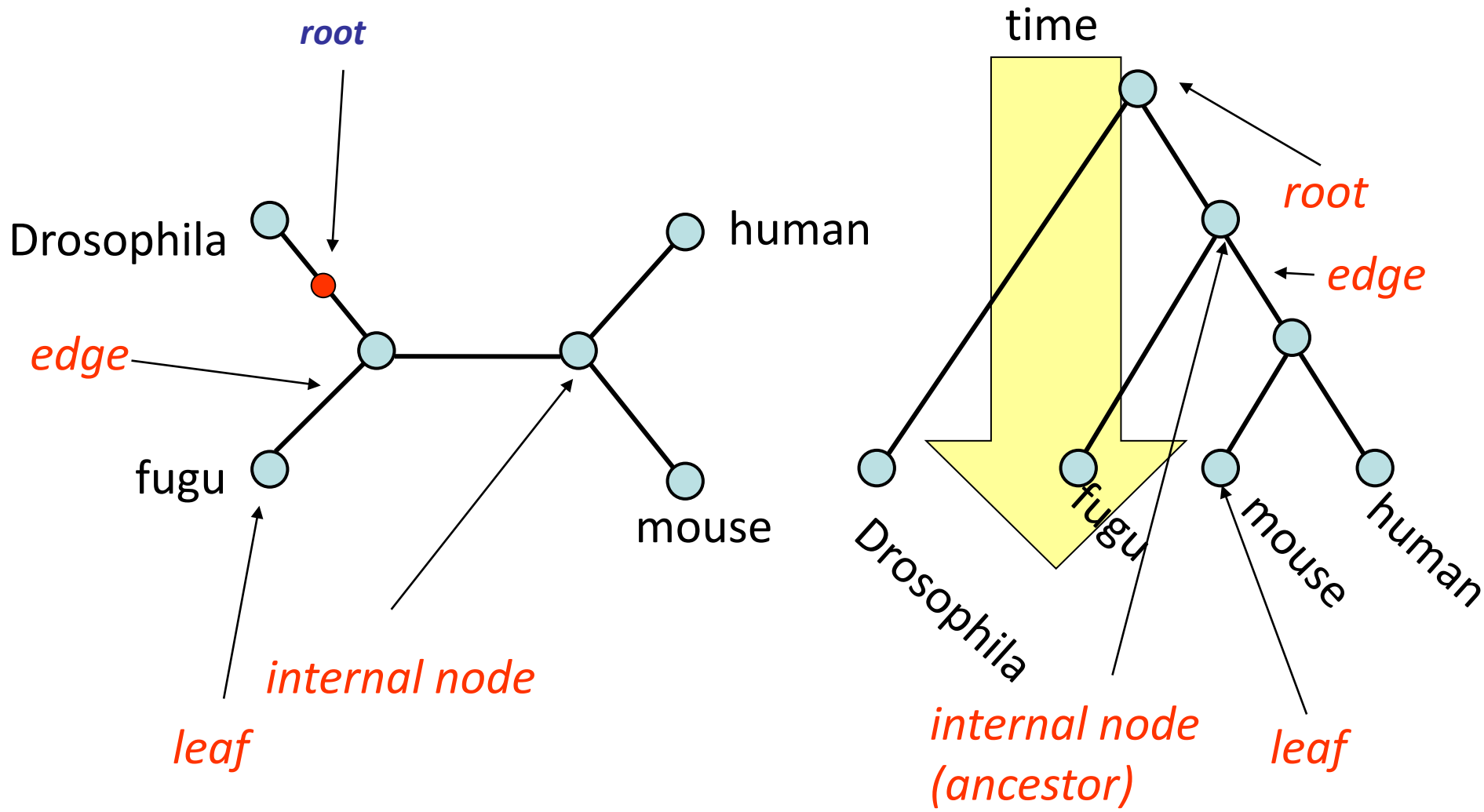
Gene tree



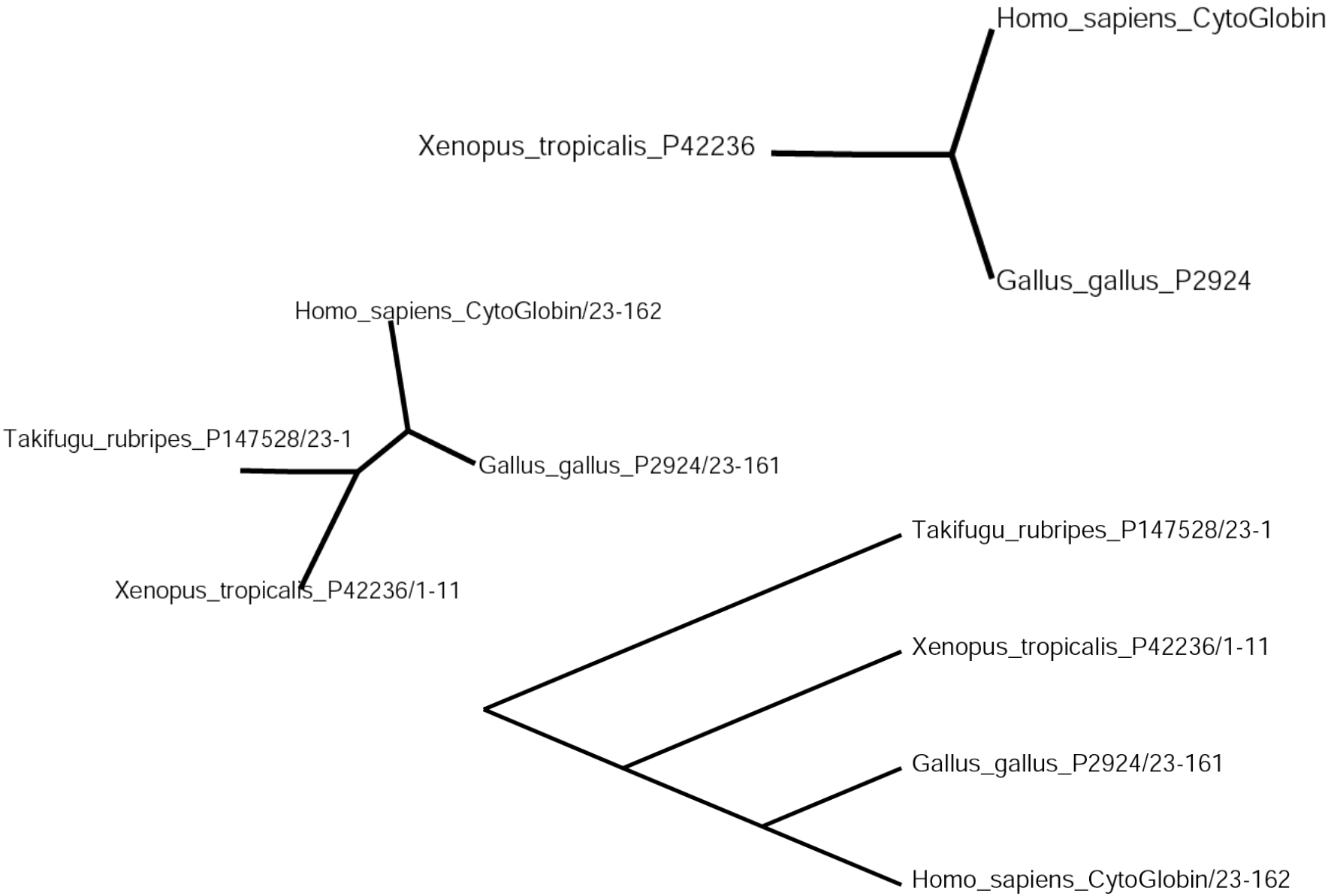
Species tree



Introduce a root



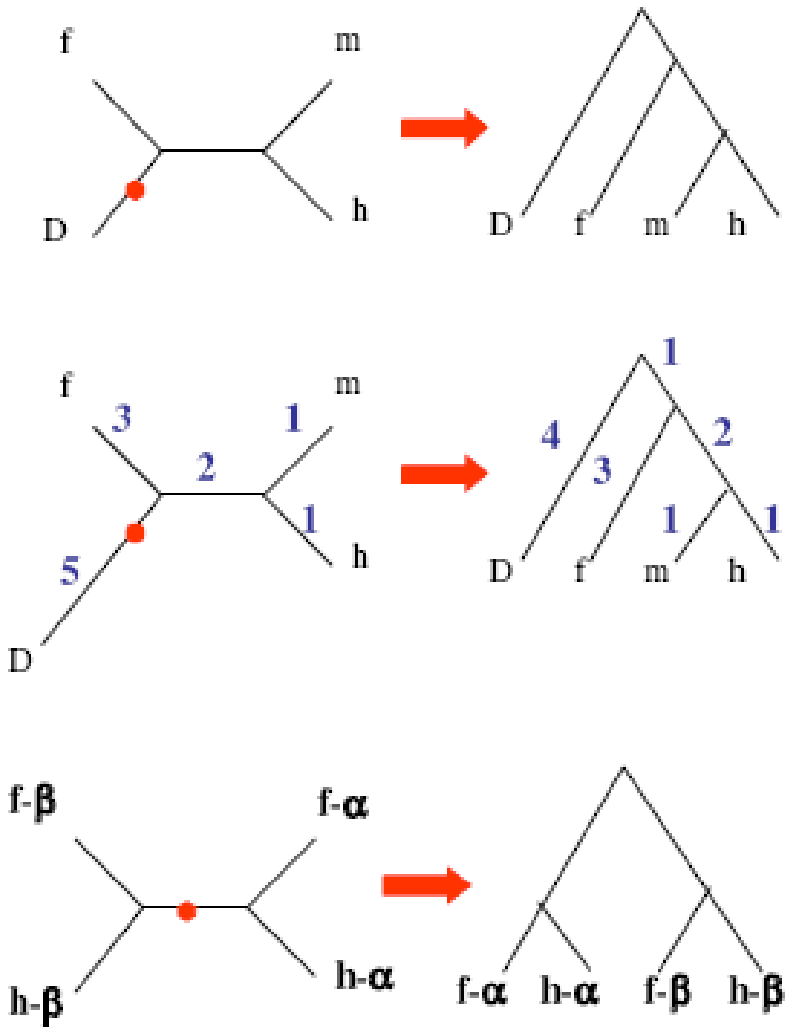
How to root a tree: outgroup



Interpreting the tree

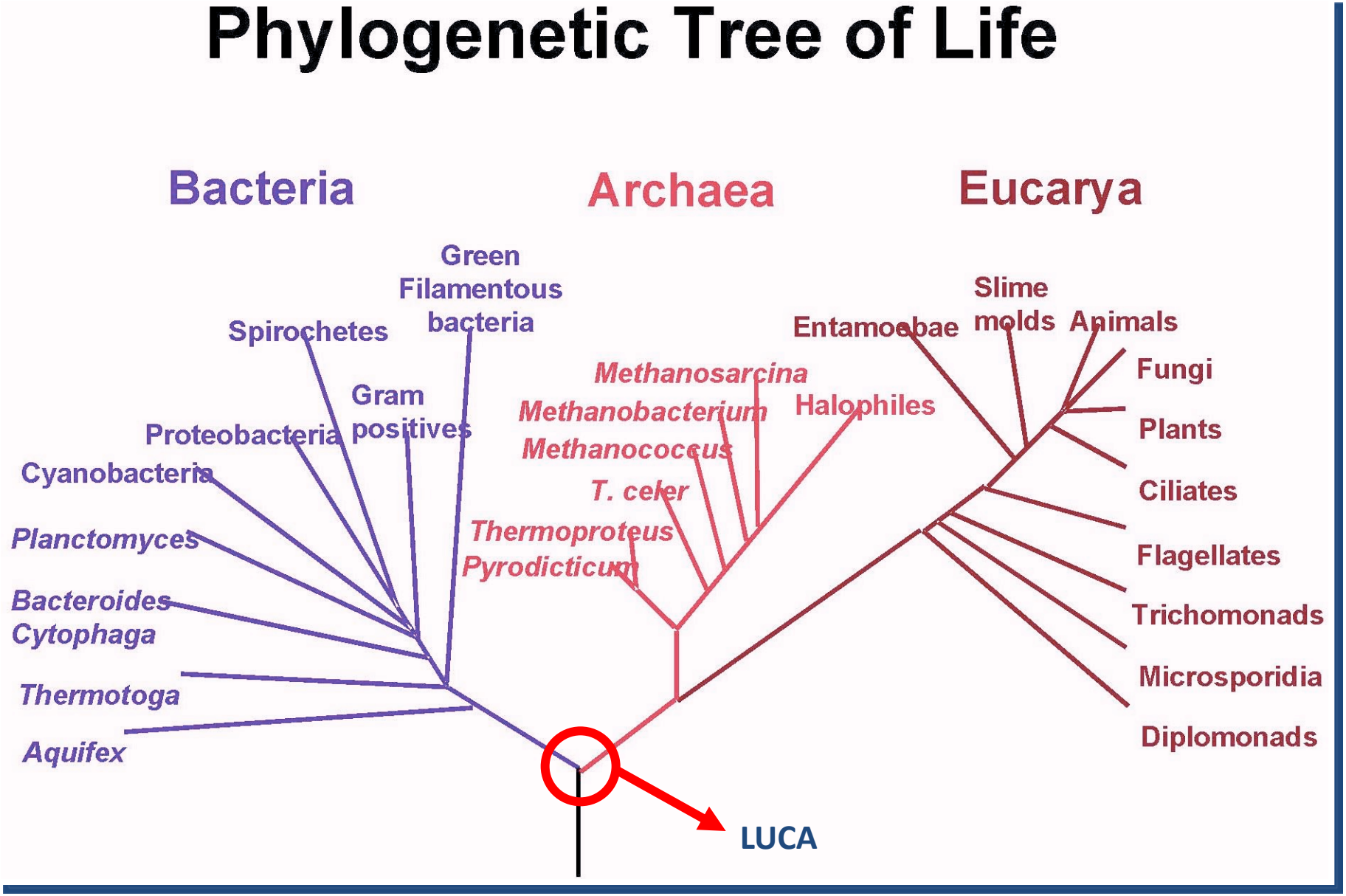
- **Outgroup**. place root between distant homologouss sequence and rest group (b)
- **Midpoint.** place root at midpoint of longest path (sum of branches between any two leafs) NB njplot
- **Gene duplication**. Place root between paralogous gene copies (b)

- NB all affected by rates !



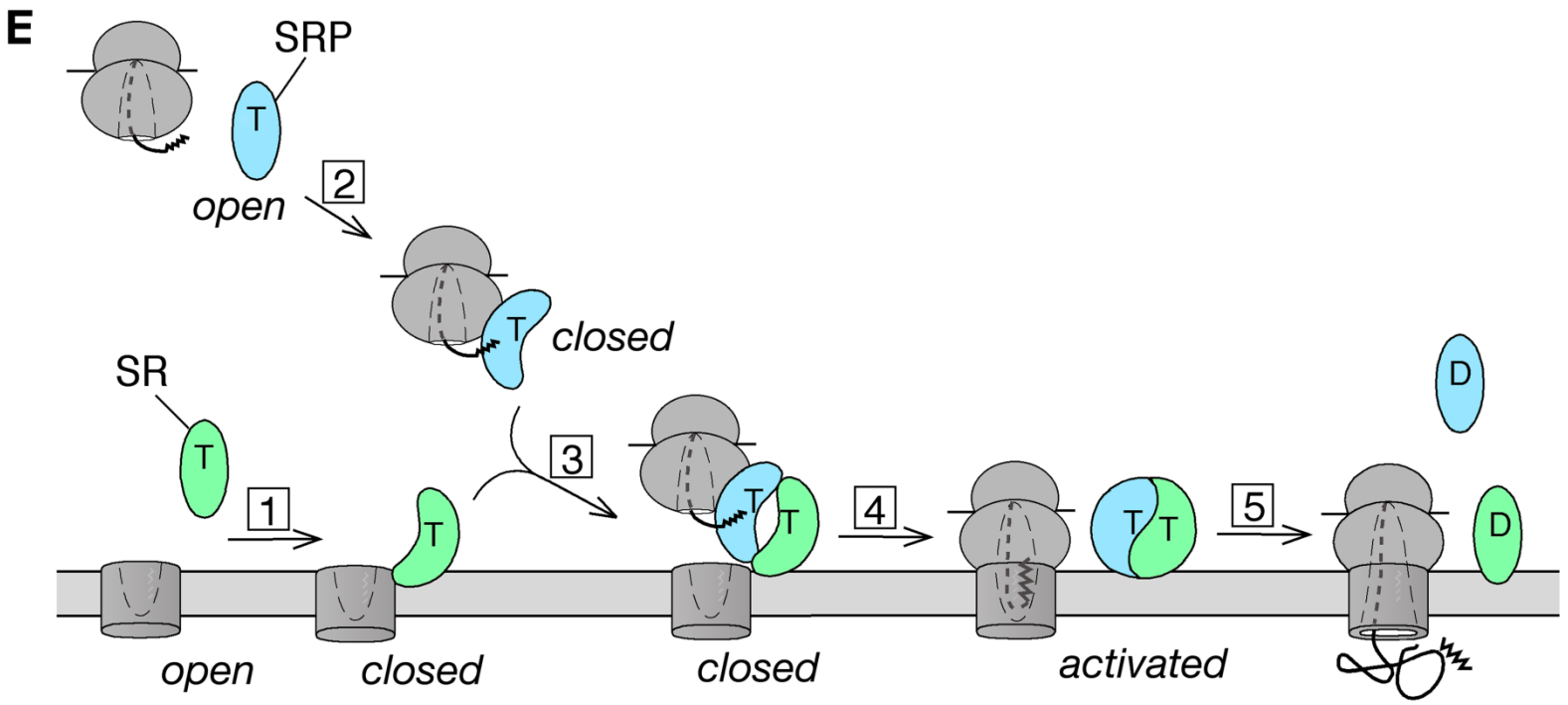
b

Phylogenetic Tree of Life



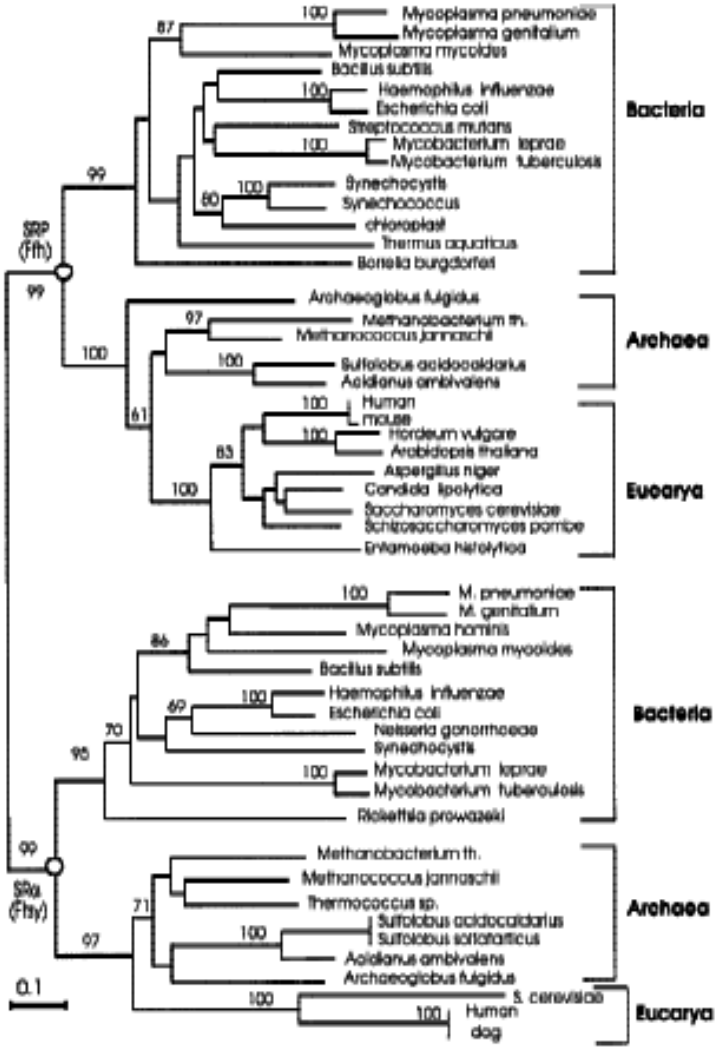
"three kingdoms"

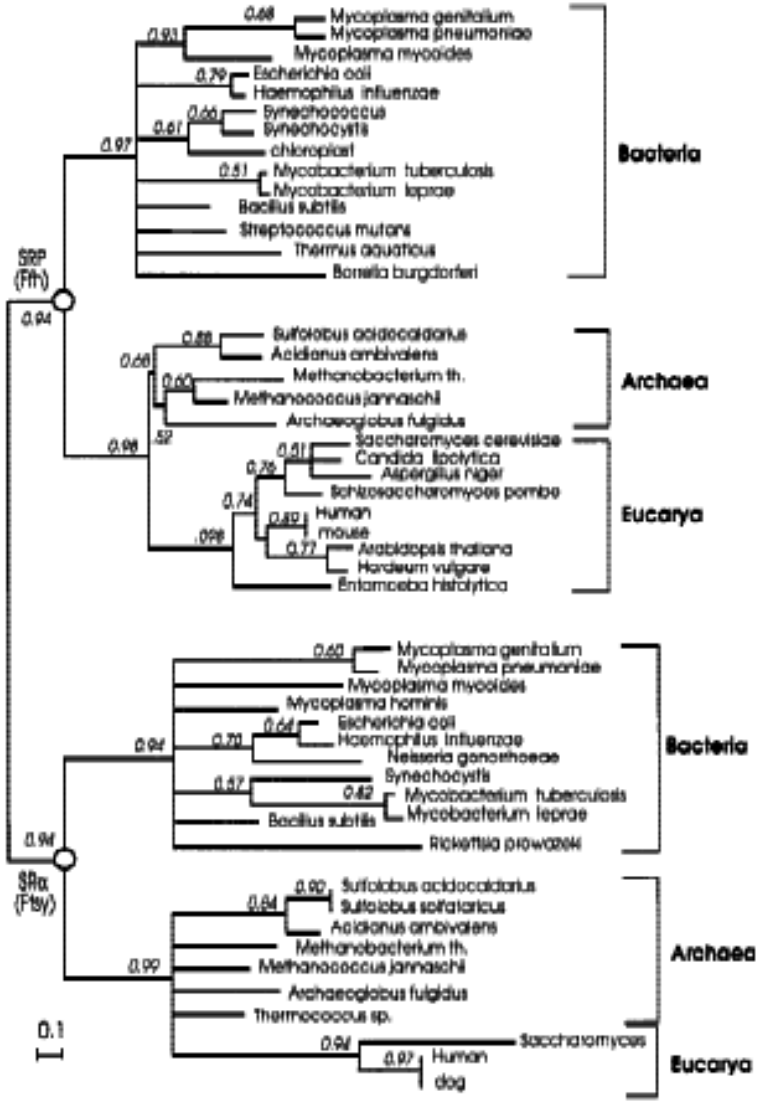
How to root the tree of life?
1: Find paralogs that duplicated before the LUCA



6 found so far

How to root the tree of life? 2: Make a tree of paralogs that duplicated before the LUCA

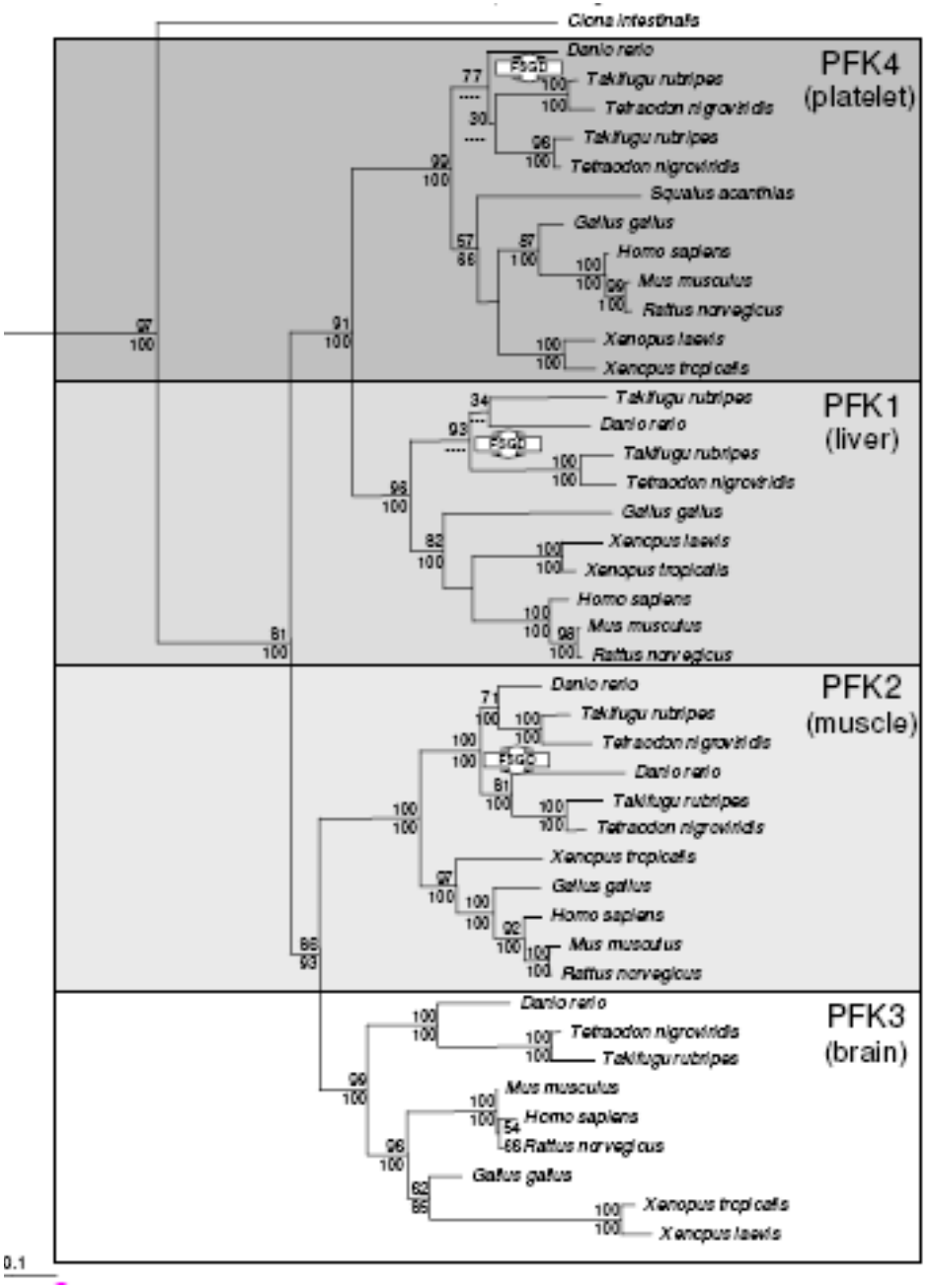




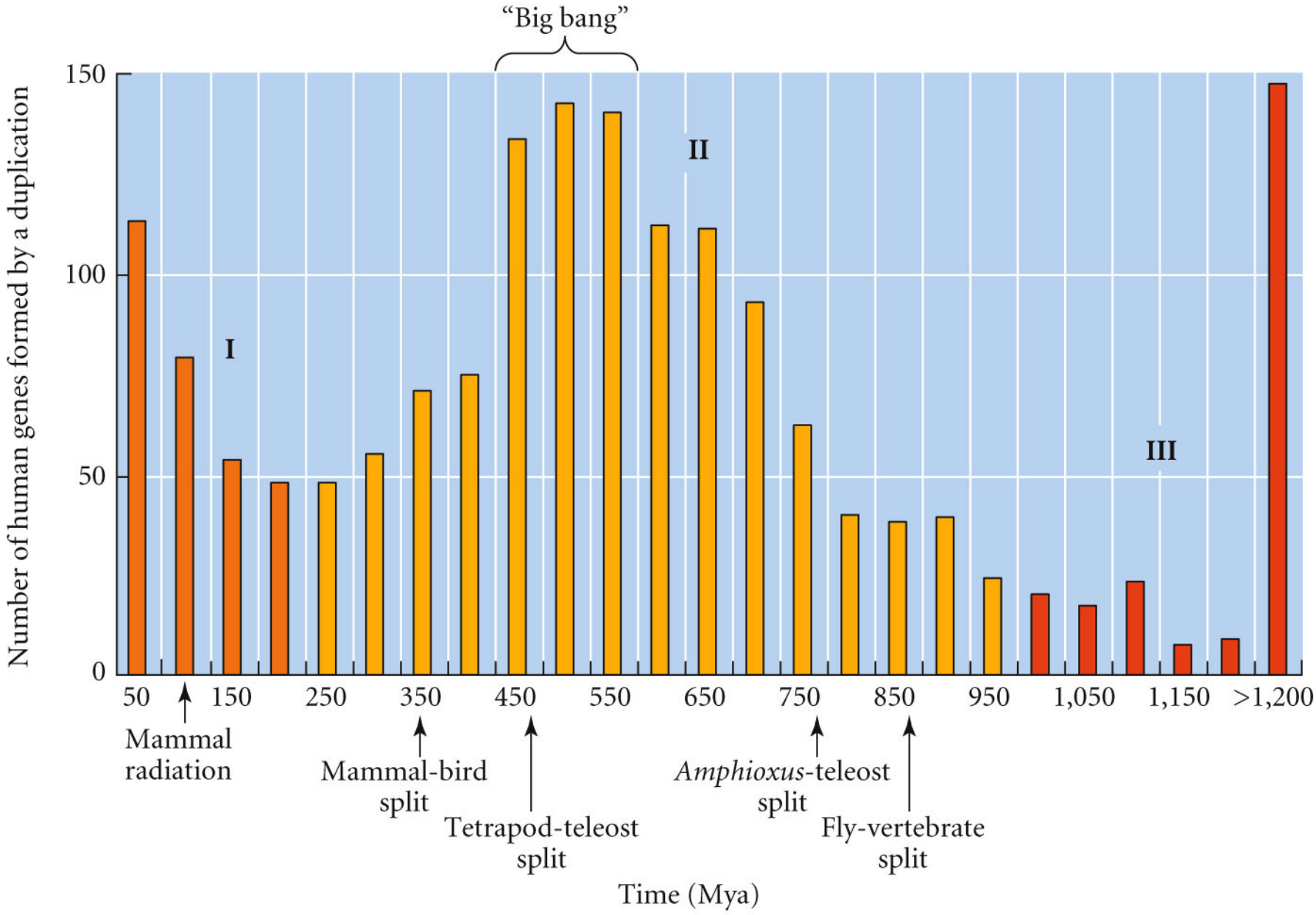
Griblado 1998 J Mol Evol

Interpreting the tree Example: vertebrate duplications

- Tetraploidy?



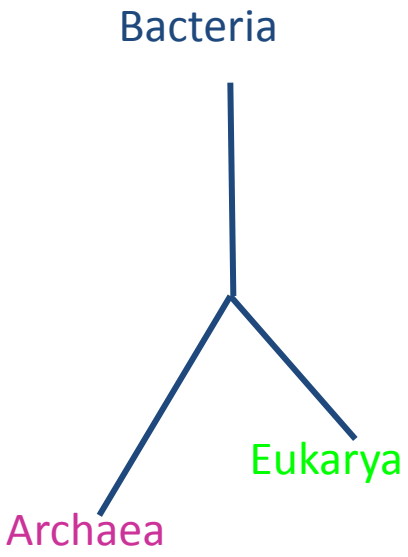
[Three rounds (1R/2R/3R) of genome duplications and the evolution of the glycolytic pathway in vertebrates.]
Steinke D, **Hoegg** S, **Brinkmann** H, Meyer A.
BMC Biol. 2006 Jun 6;4:16.

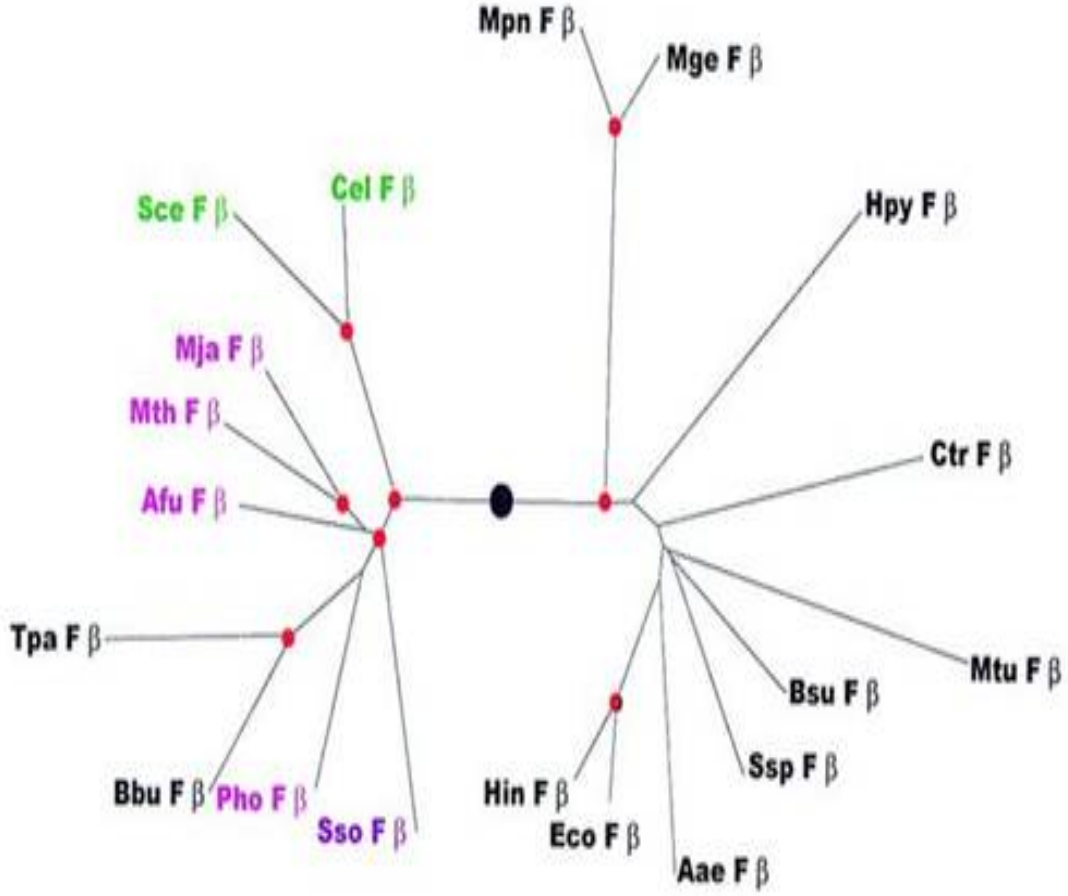


B

***EVOLUTION 2e*, Figure 20.19**

Interpreting the tree: Horizontal Gene Transfer (HGT)





Also endosymbiosis …

- Separate lecture …

So annotating gene tree can give

- Timing of duplications & gene loss
- Horizontal gene transfer
- History of endosymbiosis
- A root to a tree in the absence of an outgroup sequence / species

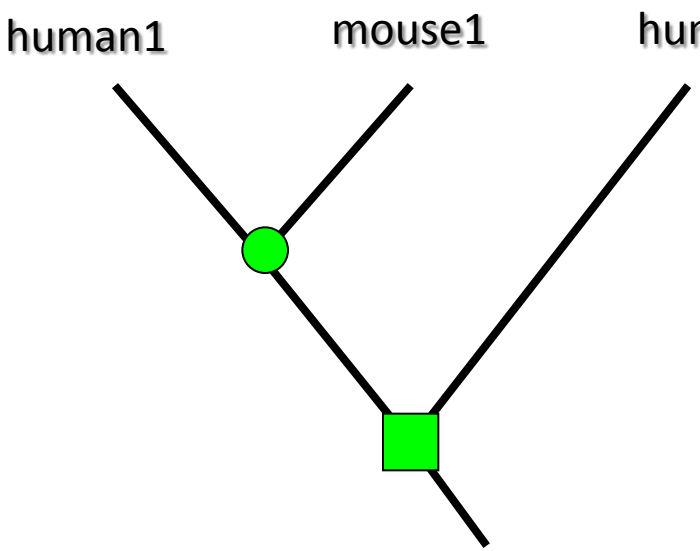
So annotating gene tree can give all kinds of incredibly cool things but in reality gene trees are very noisy …

- For ToL -> gigantic concatenated alignments ~phylogenomics
- If strict reconciliation / annotation gene trees would give e.g. many spurious duplications, B

Gene Trees, Gene Duplications, and Orthology

- How to make trees
- Bootstrap
- Interpreting trees
- duplications vs speciations vs loss, timing of duplications, HGT
- Orthology

Jargon for interpretation: Orthology (and paralogy) as a specification of homology when discussing two species



Fitch 1970
Two genes in two species are orthologous if they derive from a single gene in their *last* common ancestor

"the corresponding gene"

"last common node is a speciation node"

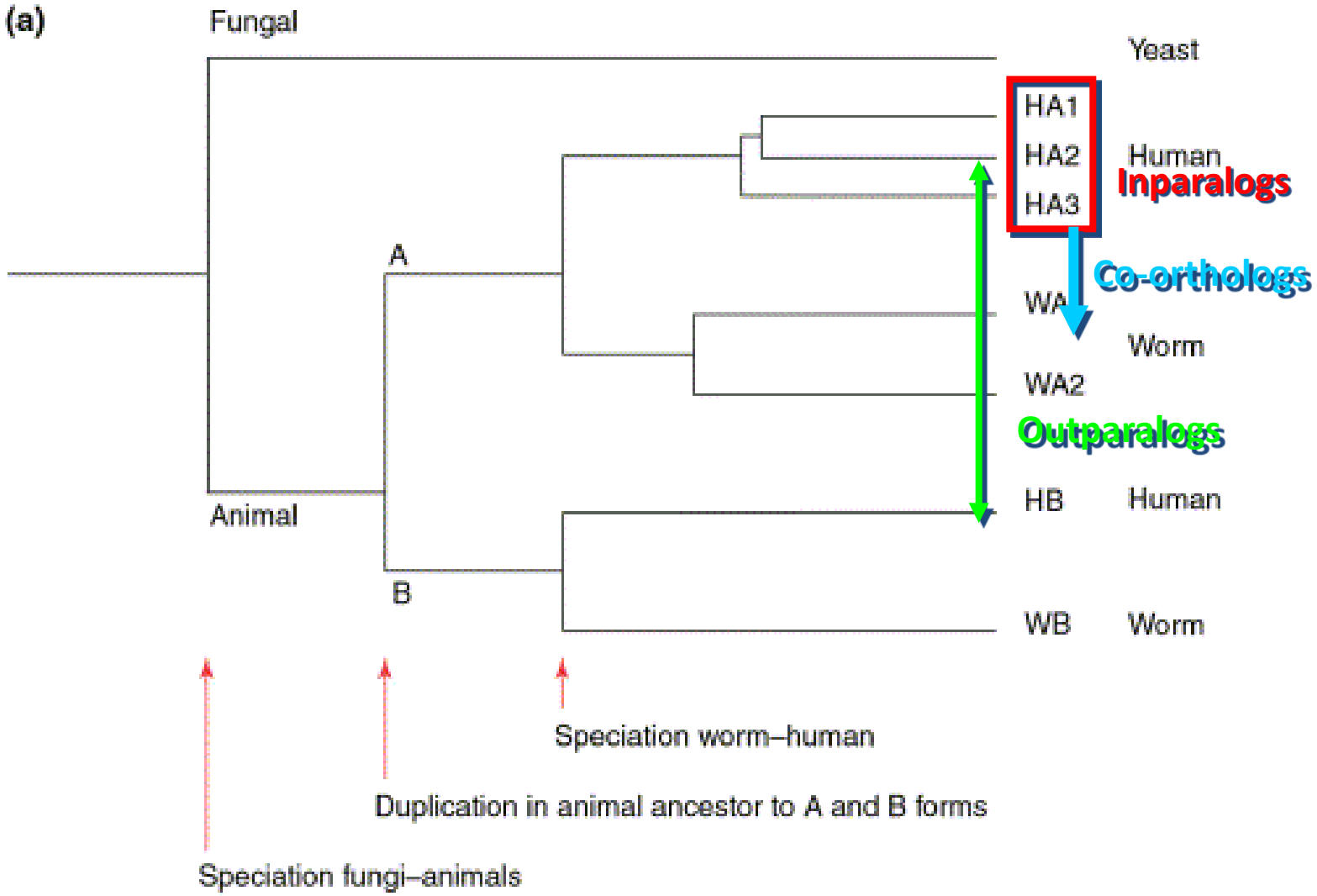
implied to have the same function

Orthology ~ annotating internal nodes as duplications or speciations

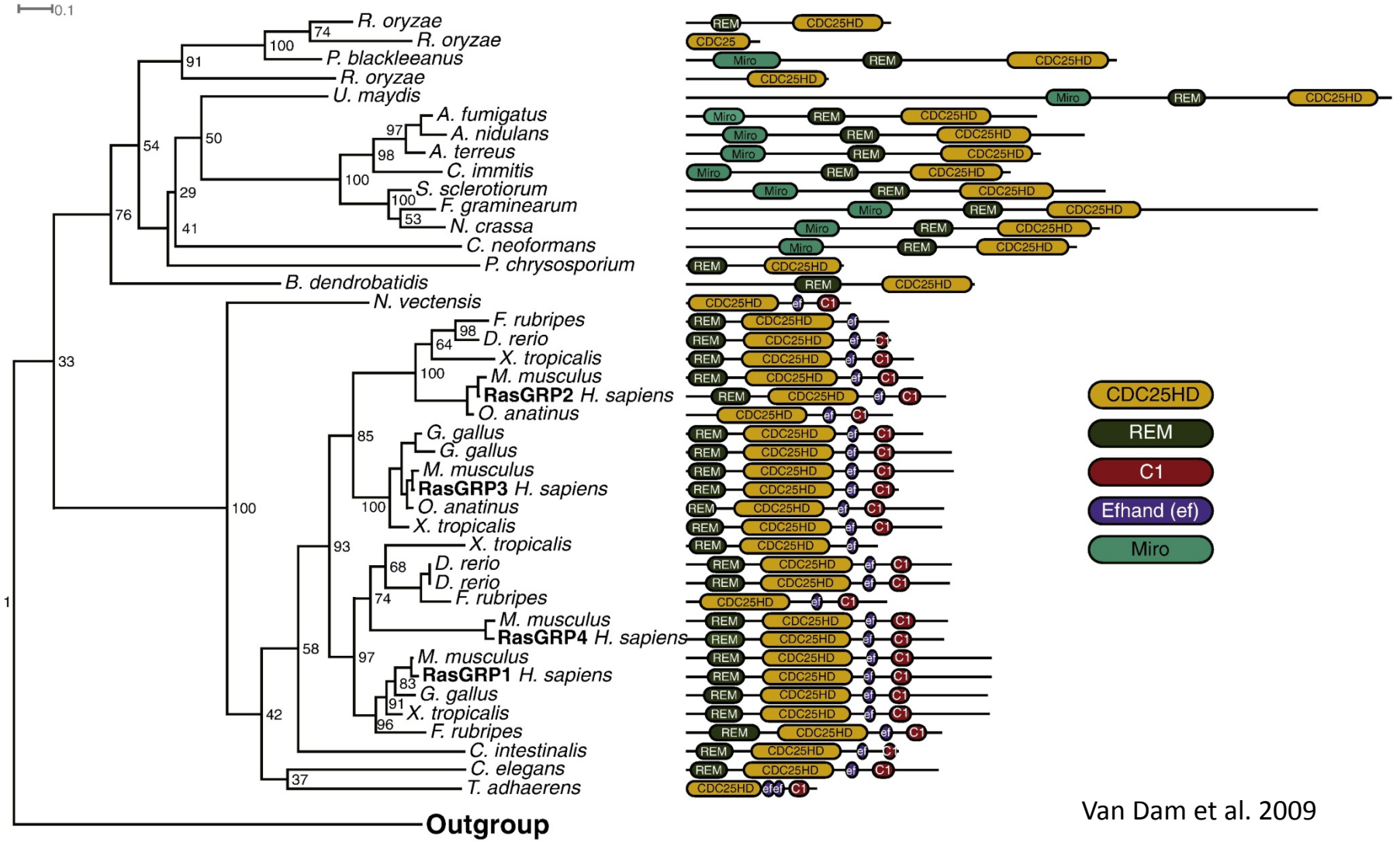
Genes can diverge by

- Speciation, or
- Duplication

Terminology: inparalogs, outparalogs, co-orthologs

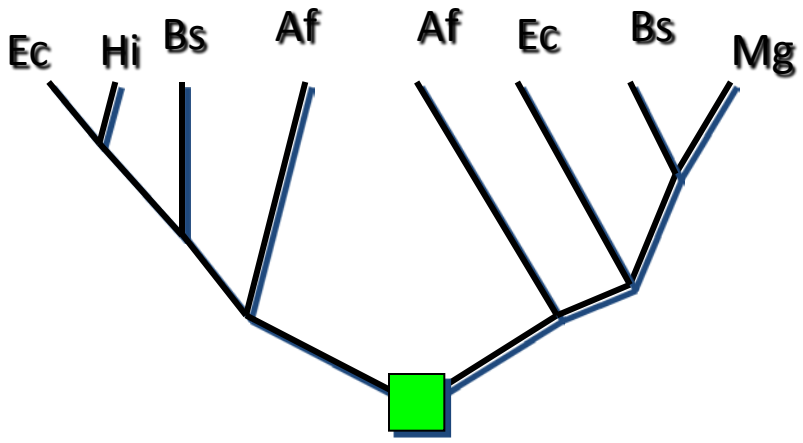


Orthologs can have different domain composition: likely changed function; only the domain is orthologous



Van Dam et al. 2009

Importance of orthology for comparative genomics: more resolution



Gene family present in
Ec Hi Bs Mg Af
Orthologs 1 present in
Ec Hi Bs Af
Orthologs 2 present in Ec Bs Mg Af

Phenotype ~ gene correlation
Func prediction if Hi is only biochem characterized enzyme
Func prediction by co-oc
Evolution of gene content: loss vs dupl

Re-cap 1

- Sometimes sequence similarity is the bottle neck for finding orthologs e.g. med11, apc15???, spindly
  - Fulfill separated by speciation and bi-directional best hit criterion
  - are occasionally found via experiments rather than sequence

Recap 2: inparalogs.

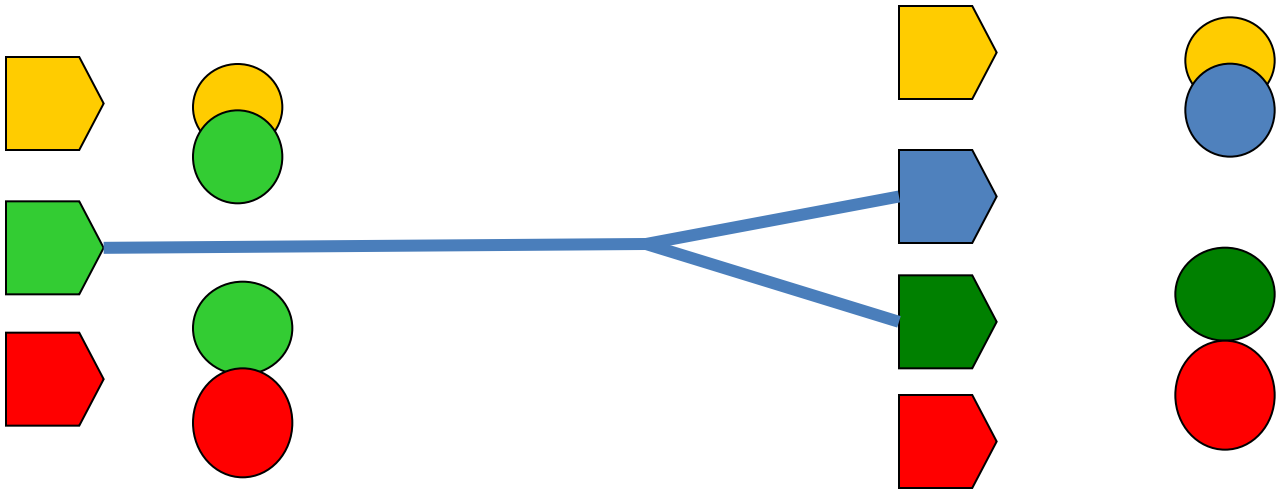
- When comparing plant or plasmodium proteins to human or yeast proteins, plenty of time for duplications to make genes that are still co-orthologs. Such duplications are thus very frequent, also at shorter time scales (i.e. vertebrates vs invertebrates, flowering plants vs green algae). What do we think of their evolution, function and how do we deal with them?

Fate after gene duplication

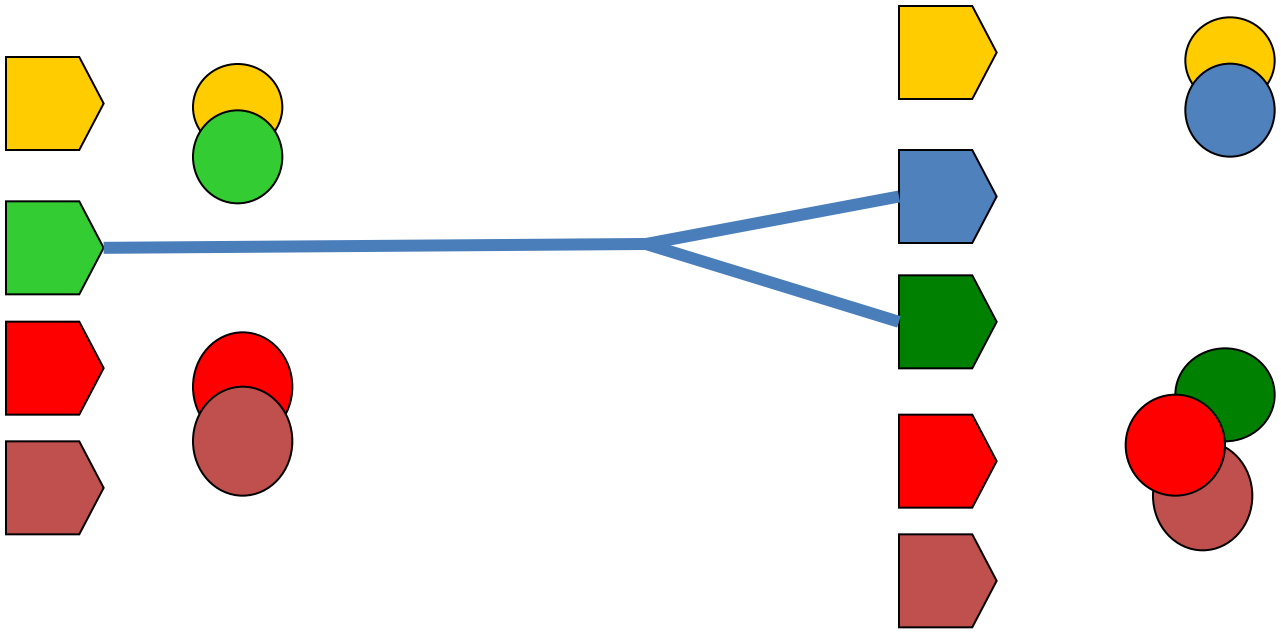
- Most duplications are thought to be deleterious (like most mutations). Hence theories (A) on why they stay (are neutral/selected) on short time scale vs (B) how they evolve and are “used” on longer time scale. We focus on B
- Dosage
- Redundancy
- Subfunctionalization
- Neofunctionalization
- (pseudogenization)

The evolution of gene duplications: classifying and distinguishing between models.
Innan H, Kondrashov F.
Nat Rev Genet. **2010** Feb;11(2):97-108

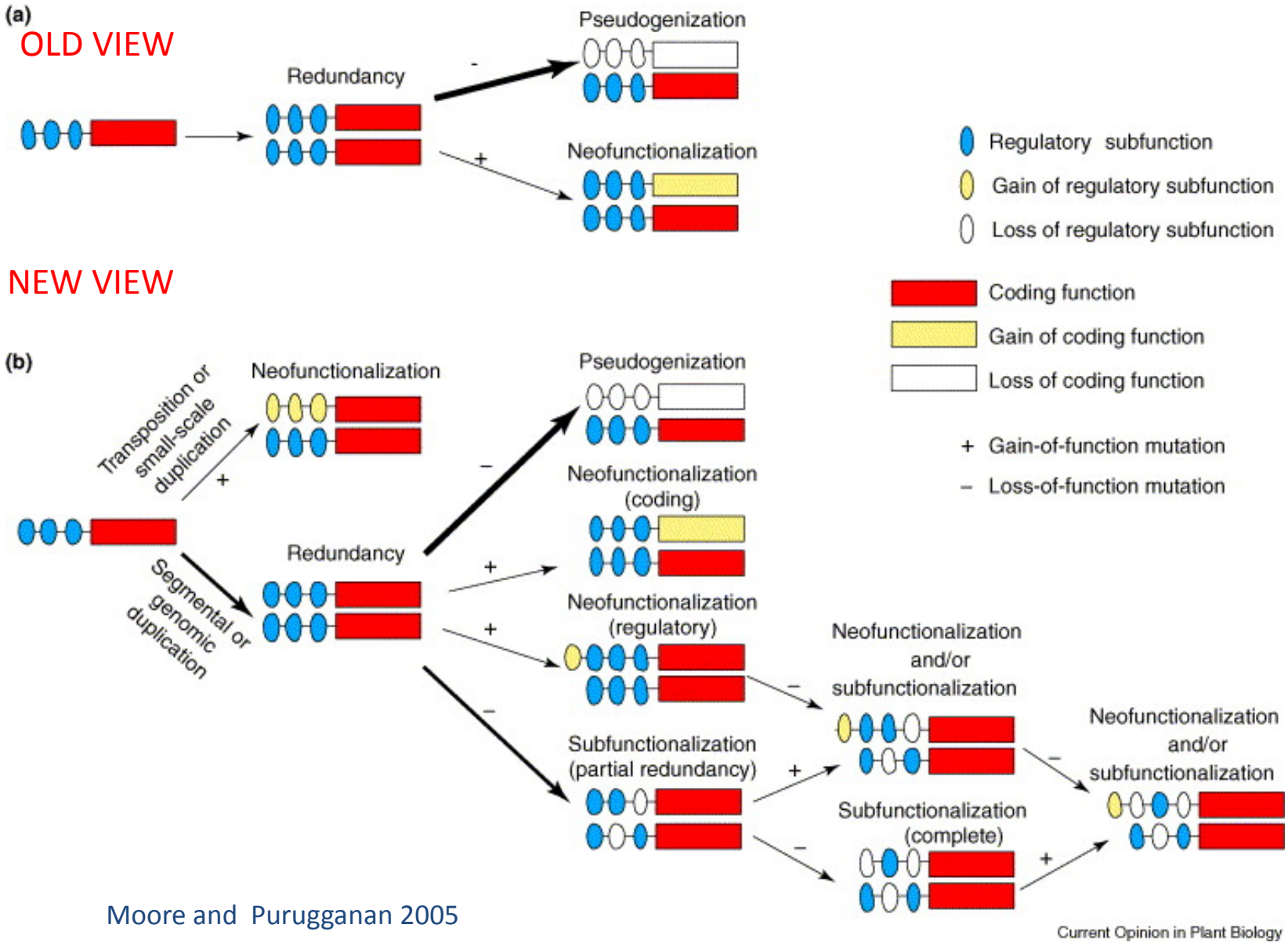
subfunctionalization: example in terms of protein complexes (=GO cellular component)



neofunctionalization: example in terms of protein complexes (=GO cellular component)

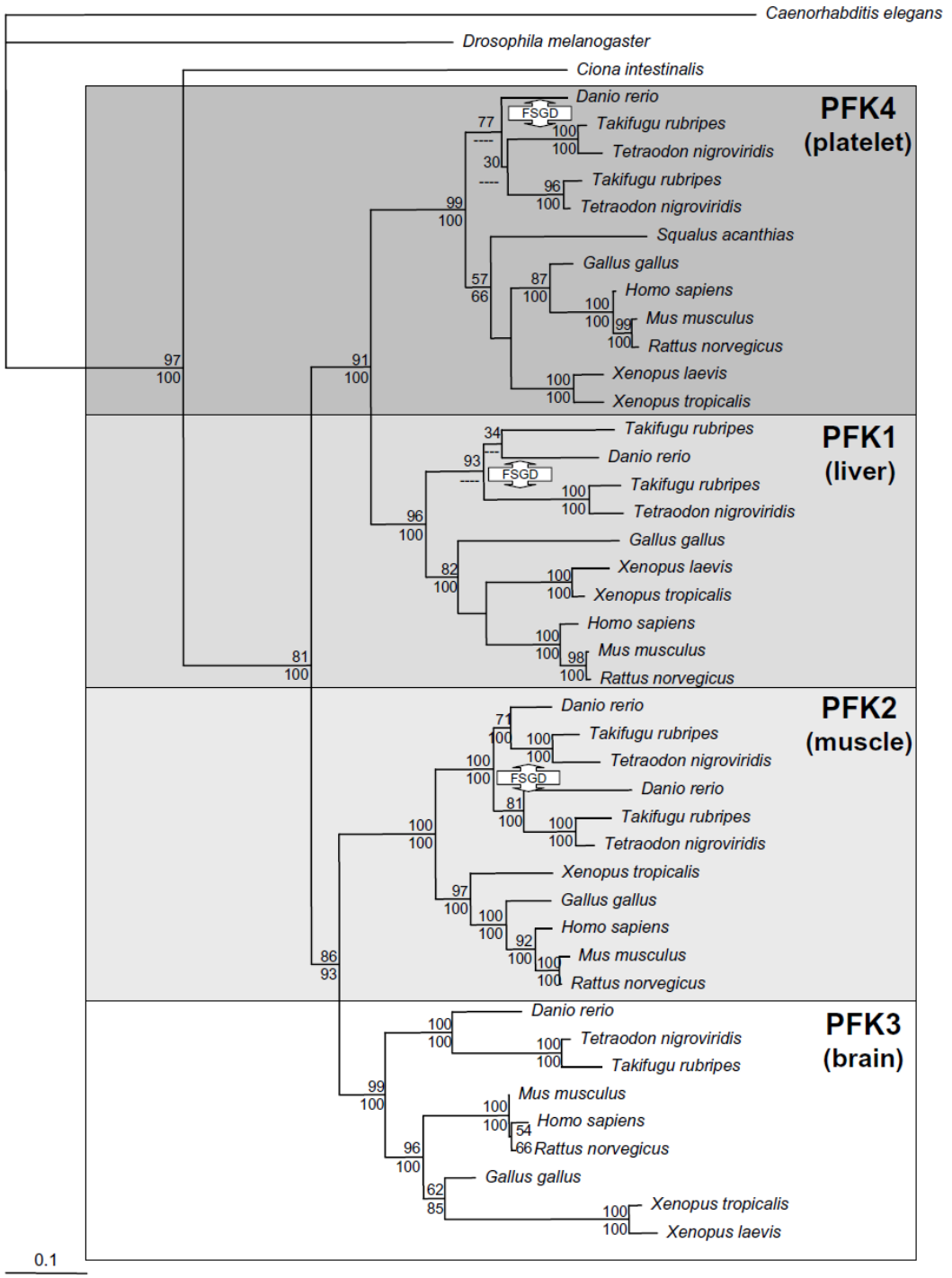


Sub vs neo in regulatory context



Moore and Purugganan 2005

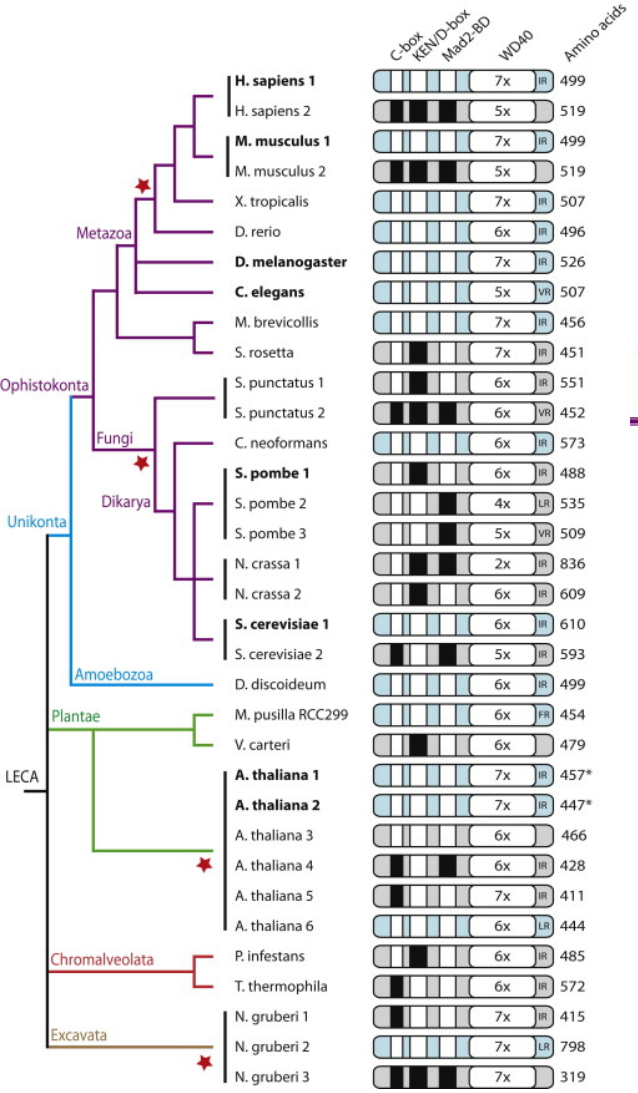
b

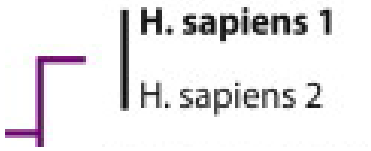


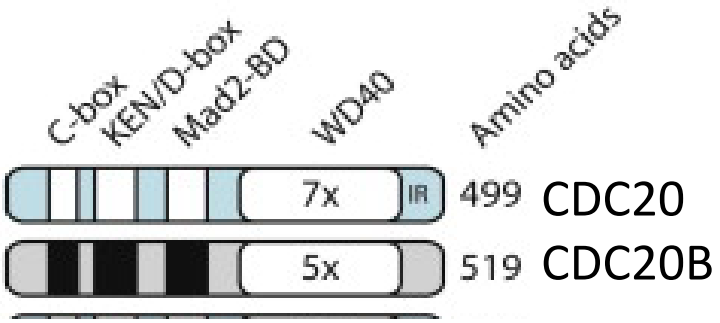
[Three rounds (1R/2R/3R) of genome duplications evolution of the glycolytic pathway in vertebrates]

Steinke D, **Hoegg** S, **Brinkmann** H, Meyer A.
BMC Biol. 2006 Jun 6;4:16.

Pseudogenization vs neofunctionalization vs “marginalization”, eg. CDC20B

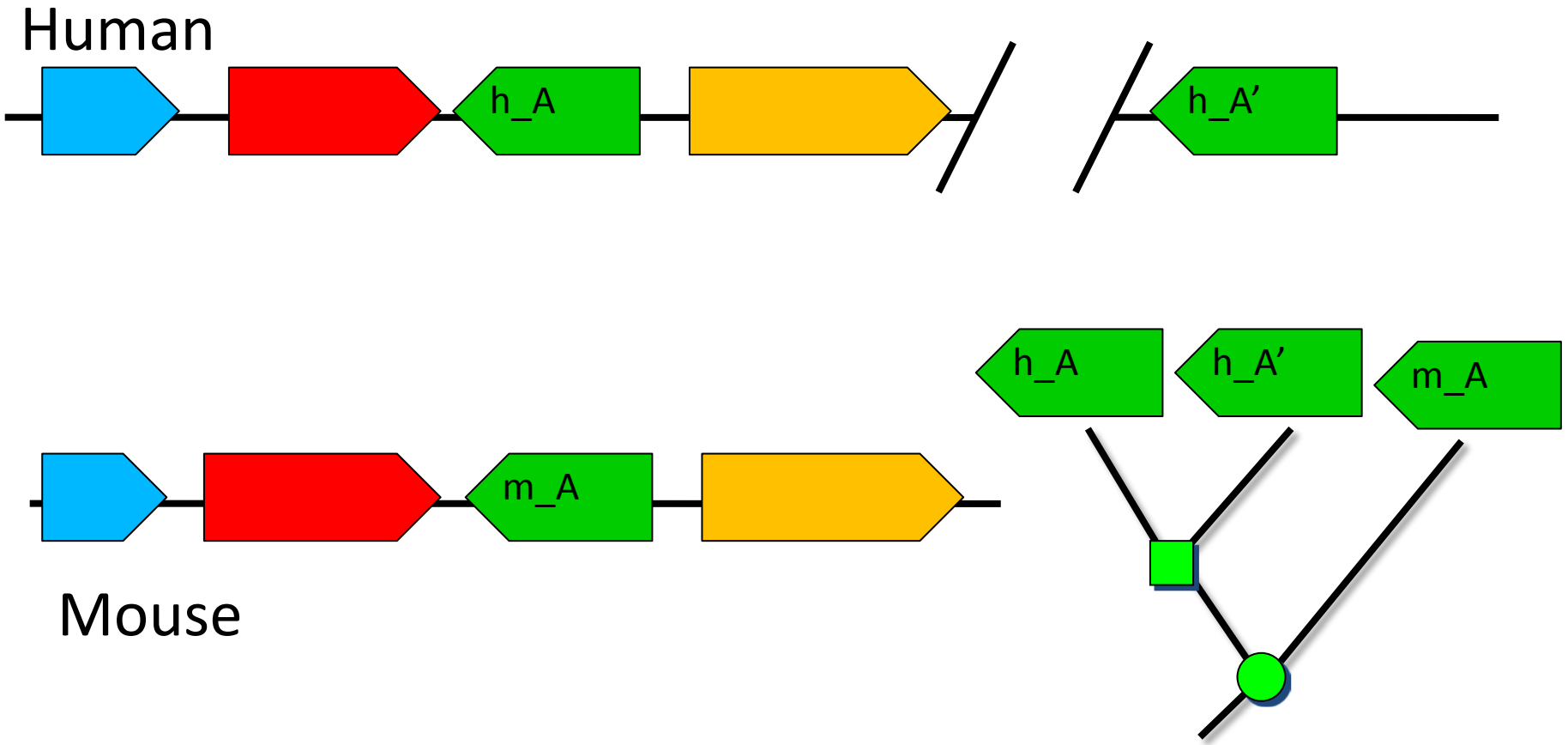




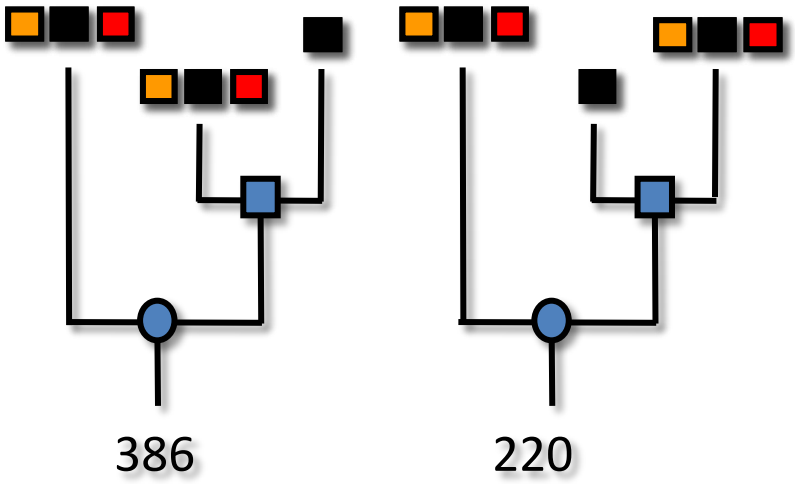


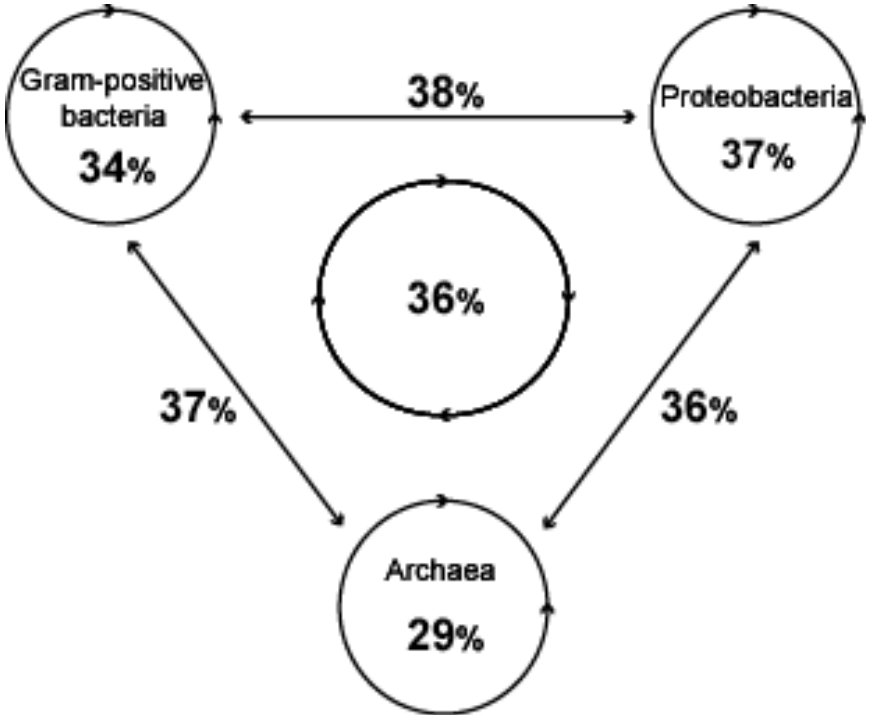
Genome alignments and orthology:

- to asses which is the “ancestral” gene
- a more detailed reconstruction of the past
- benchmarking orthology methods

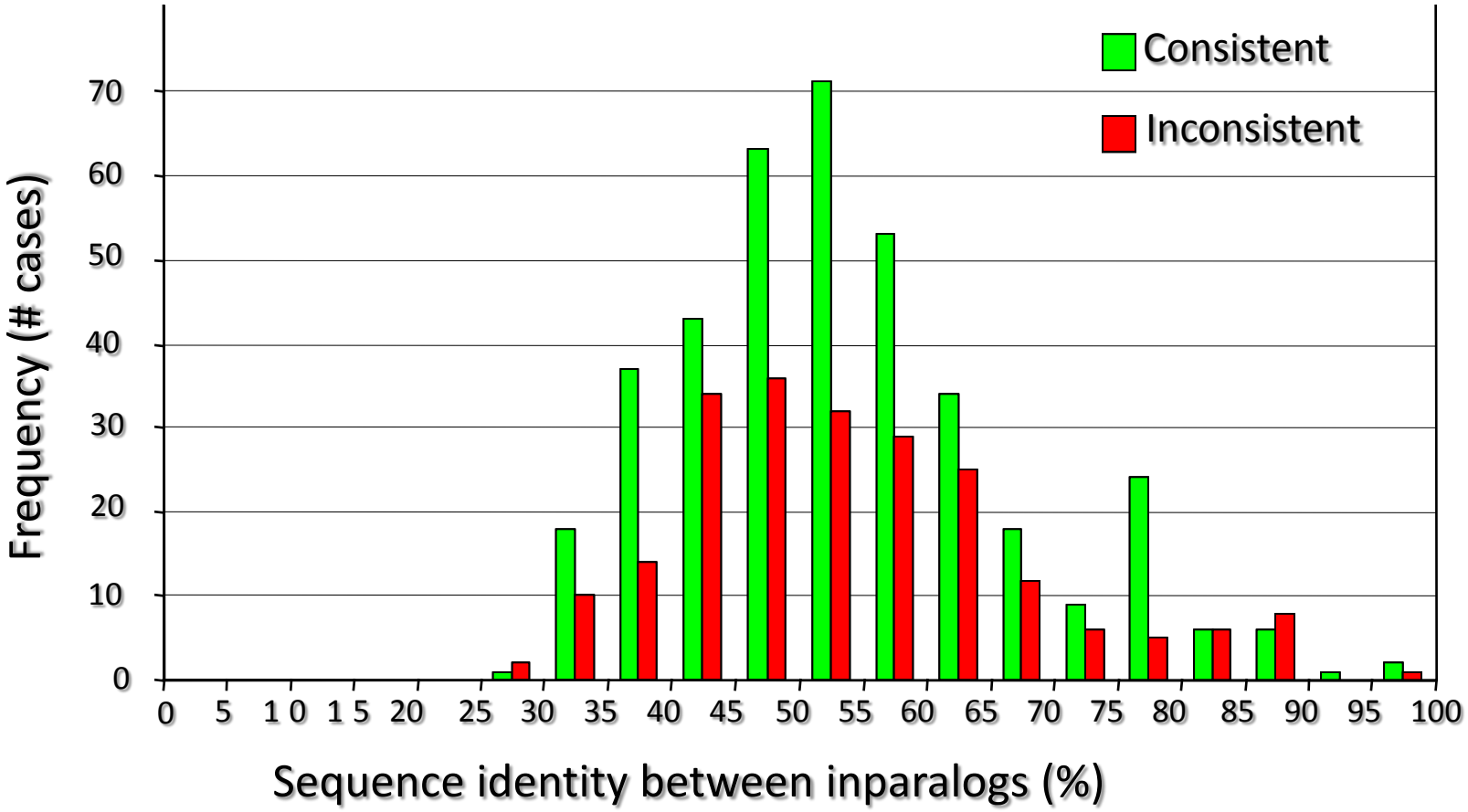


Does retaining the ancestral “role” correlate with speed of sequence evolution: yes but a substantial minority is inconsistent



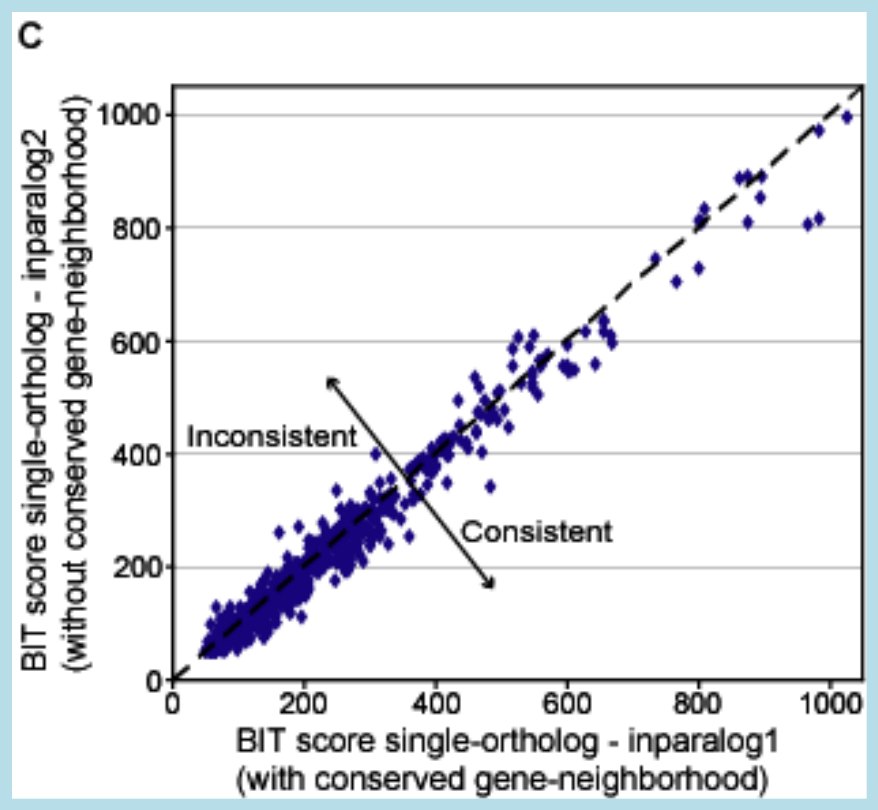


Why inconsistencies?



Not because of chance due to lack of divergence time

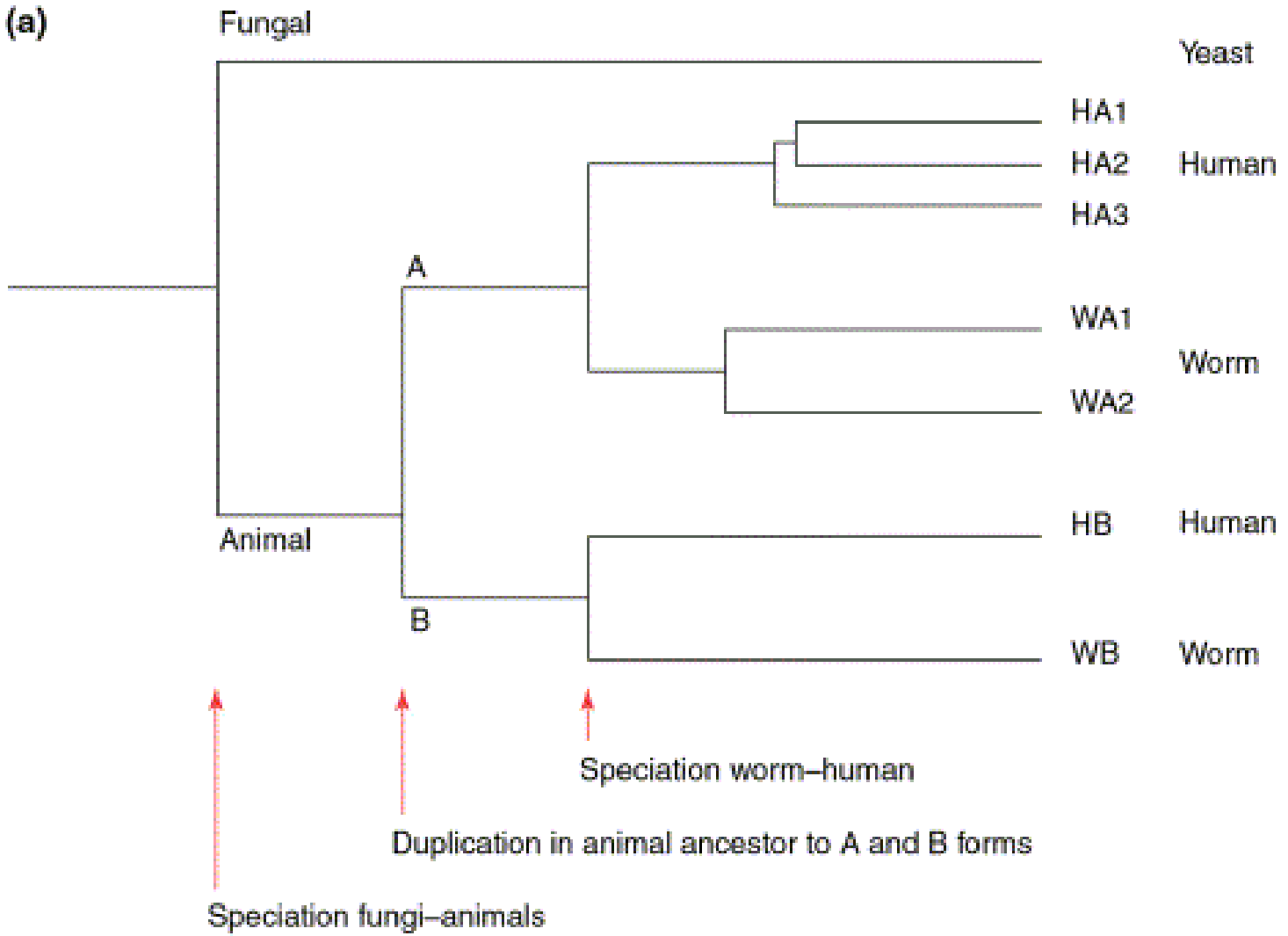
Why do observe inconsistencies?



Similar sequence divergence of inparalogs relative to their single-ortholog, molecular function similar?

Any inconsistencies are then a chance outcome: both duplicates have diverged, but at (roughly) the same evolutionary speed (most amino acids substitutions are only been subject to purifying selection and not to adaptive selection)

Orthology is officially defined between pairs of species, but for many questions you think about a set of species, i.e. what to put in the excel-sheet



Orthologous groups

- Conceptually: all proteins that are directly descended from one protein in the last common ancestor of all species in the set are considered orthologous to each other (i.e. includes inparalogs relative to this potentially quite ancient speciation)

b