Molecular evolution

Joe Felsenstein

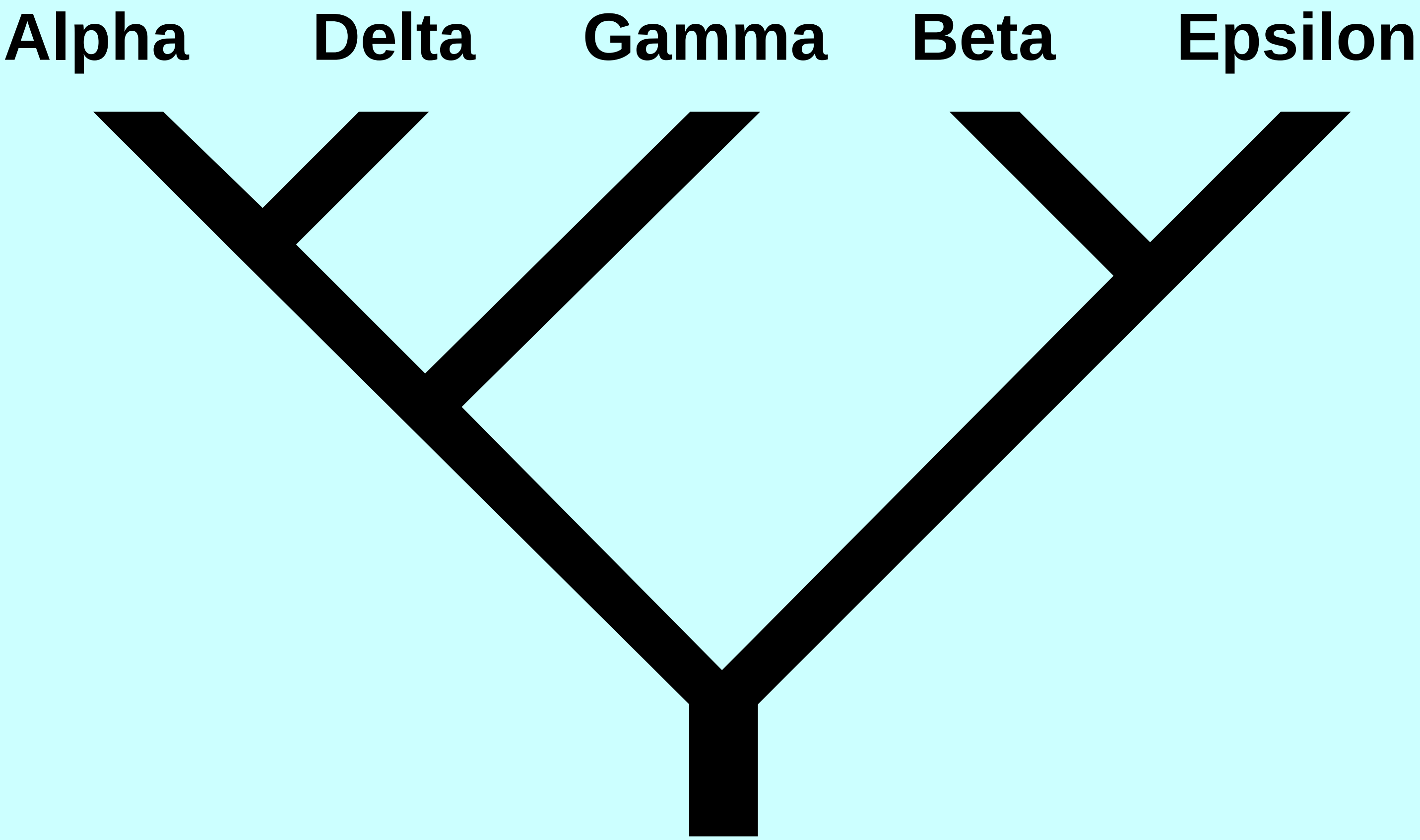
GENOME 453, Autumn 2013

A data example for phylogeny inference

Five DNA sequences, for some gene in an imaginary group of species whose names are Alpha, Beta, Gamma, Delta, and Epsilon:

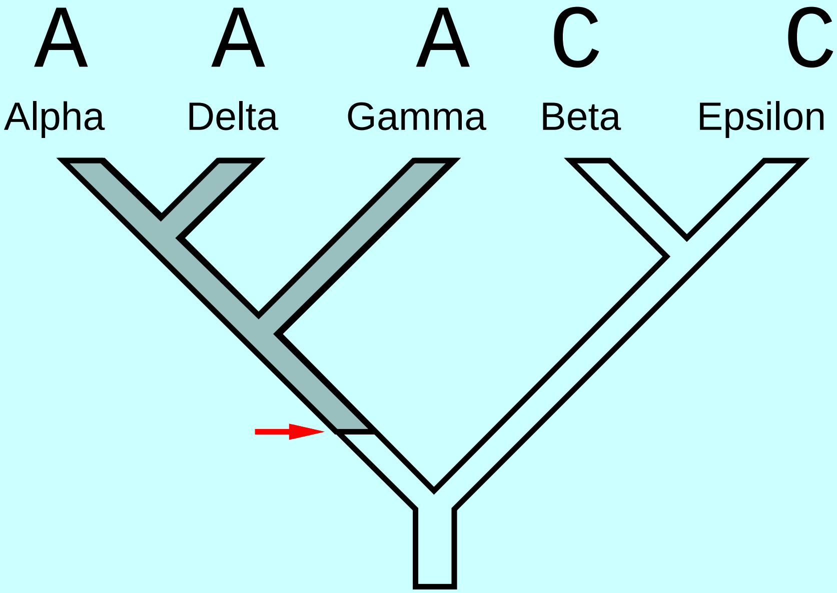
	site					
species	1	2	3	4	5	6
Alpha	A	T	G	A	G	C
Beta	C	T	C	T	A	C
Gamma	A	G	G	T	A	C
Delta	A	G	G	A	G	T
Epsilon	C	T	C	A	G	C

The tree we are evaluating

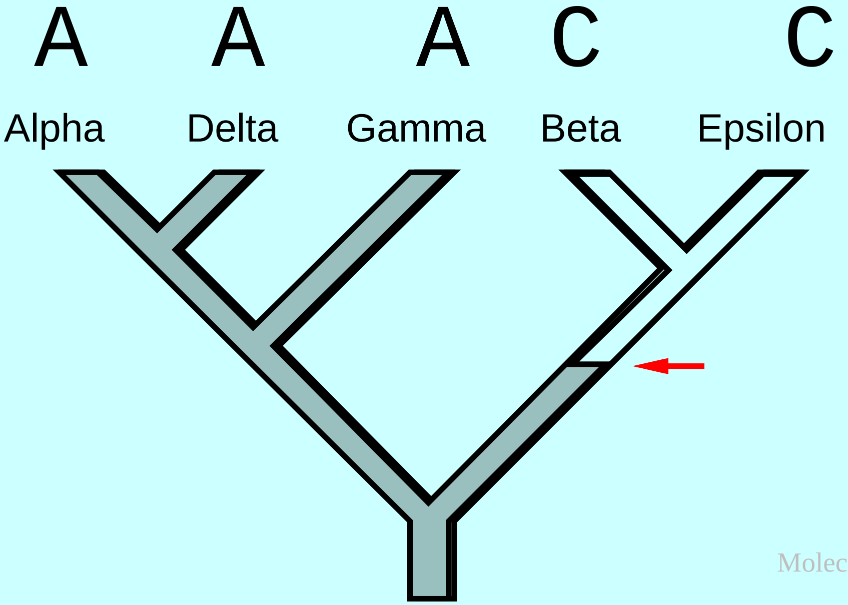


Steps in character 1

	1	2	3	4	5	6
Alpha	A	T	G	A	G	C
Beta	C	T	C	T	A	C
Gamma	A	G	G	T	A	C
Delta	A	G	G	A	G	T
Epsilon	C	T	C	A	G	C

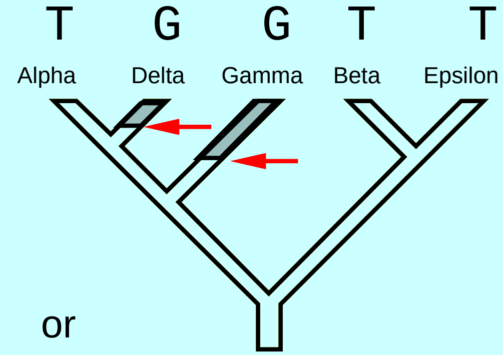


or

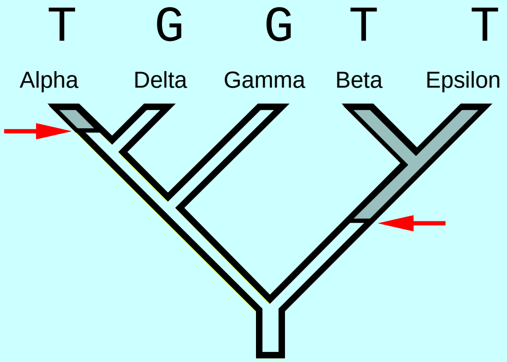


Steps in character 2

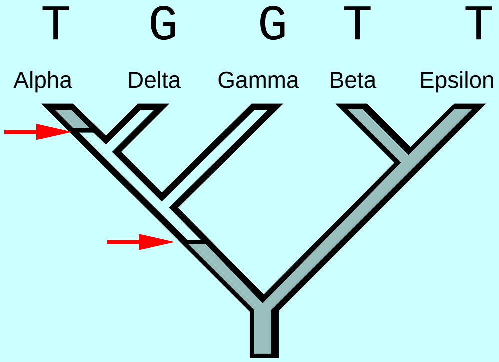
	1	2	3	4	5	6
Alpha	A	T	G	A	G	C
Beta	C	T	C	T	A	C
Gamma	A	G	G	T	A	C
Delta	A	G	G	A	G	T
Epsilon	C	T	C	A	G	C



or

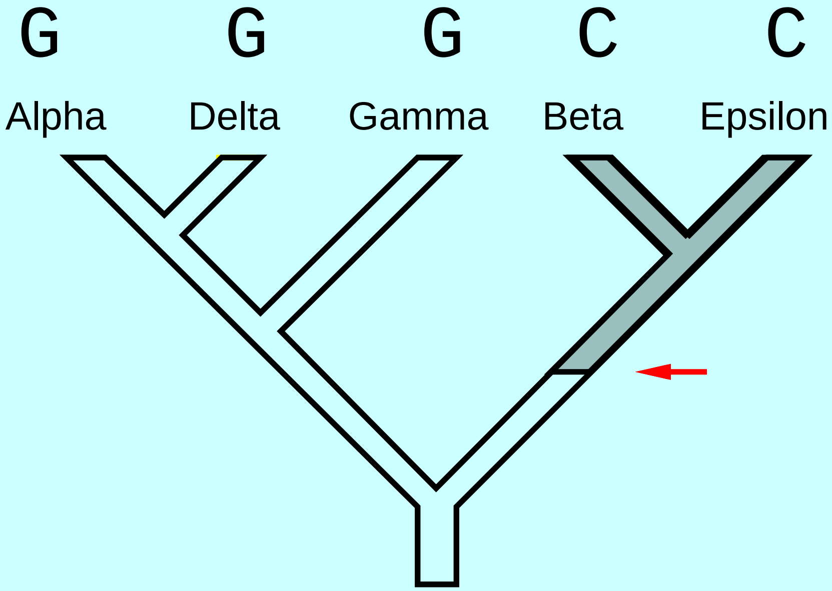


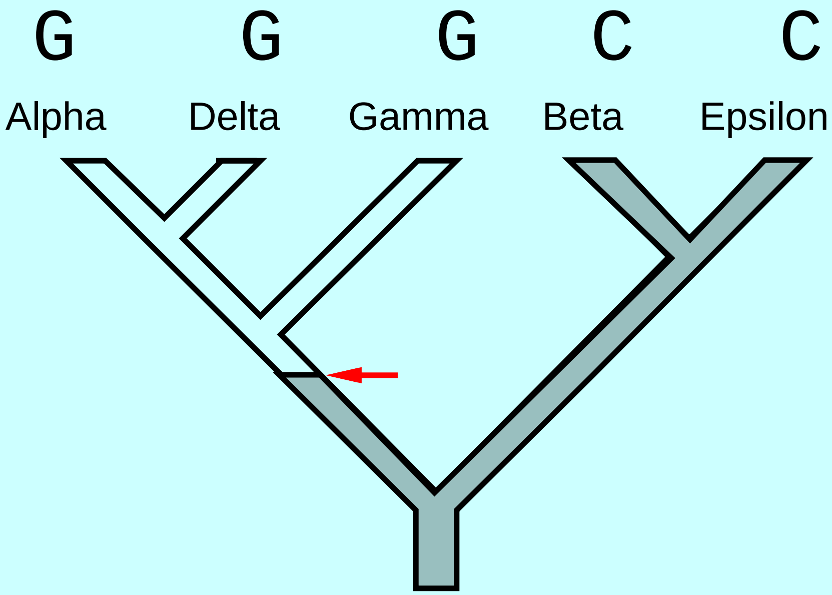
or



Steps in character 3

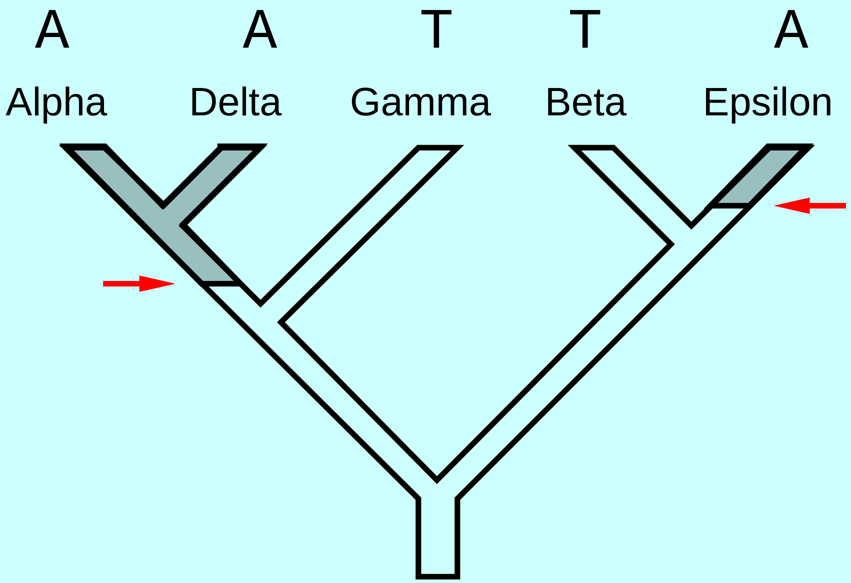
	1	2	3	4	5	6
Alpha	A	T	G	A	G	C
Beta	C	T	C	T	A	C
Gamma	A	G	G	T	A	C
Delta	A	G	G	A	G	T
Epsilon	C	T	C	A	G	C



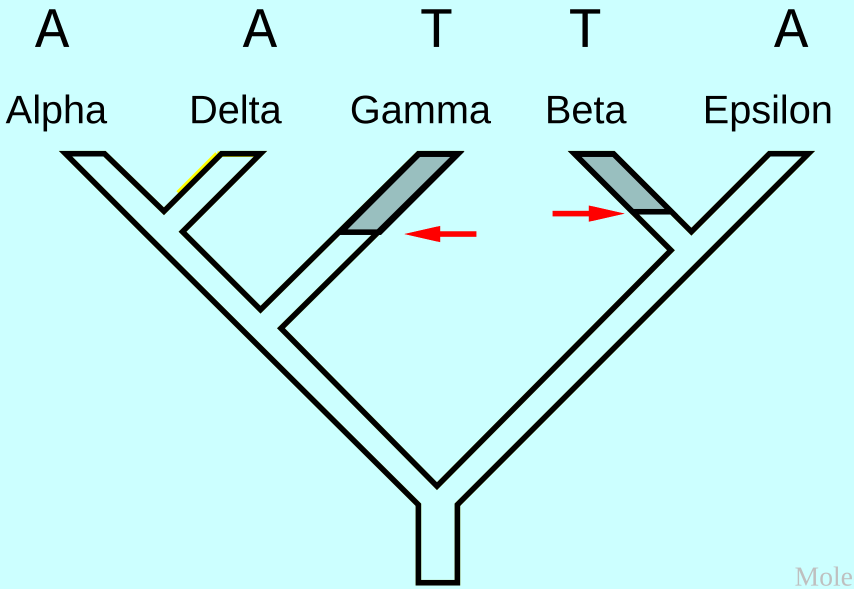


Steps in character 4

	1	2	3	4	5	6
Alpha	A	T	G	A	G	C
Beta	C	T	C	T	A	C
Gamma	A	G	G	T	A	C
Delta	A	G	G	A	G	T
Epsilon	C	T	C	A	G	C

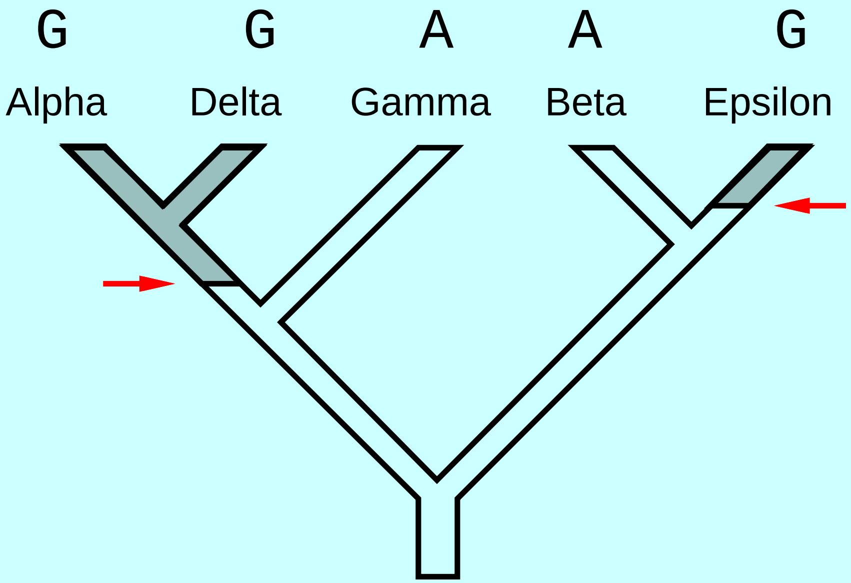


or

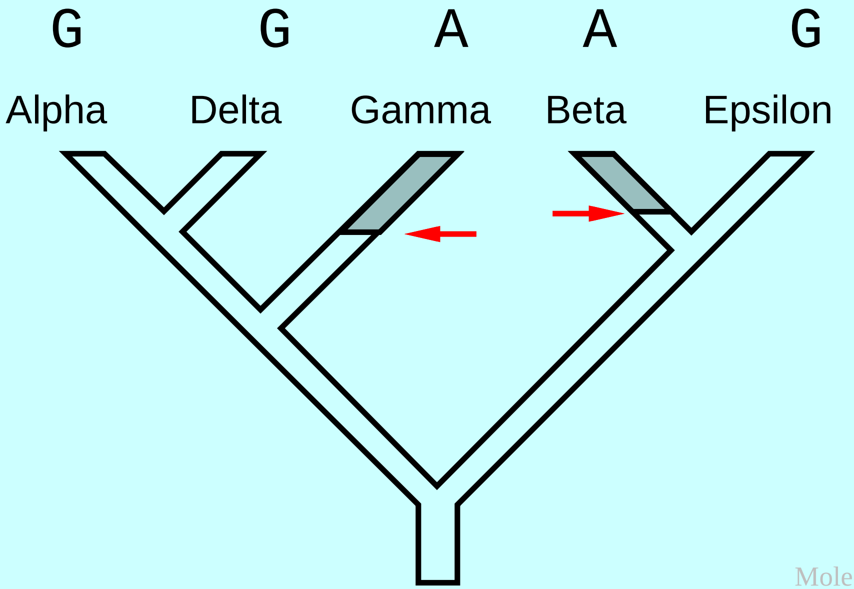


Steps in character 5

	1	2	3	4	5	6
Alpha	A	T	G	A	G	C
Beta	C	T	C	T	A	C
Gamma	A	G	G	T	A	C
Delta	A	G	G	A	G	T
Epsilon	C	T	C	A	G	C

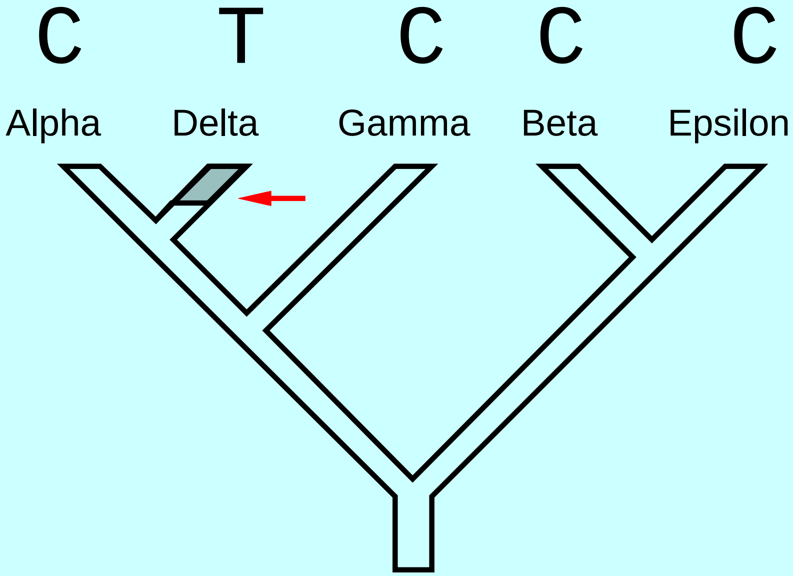


or



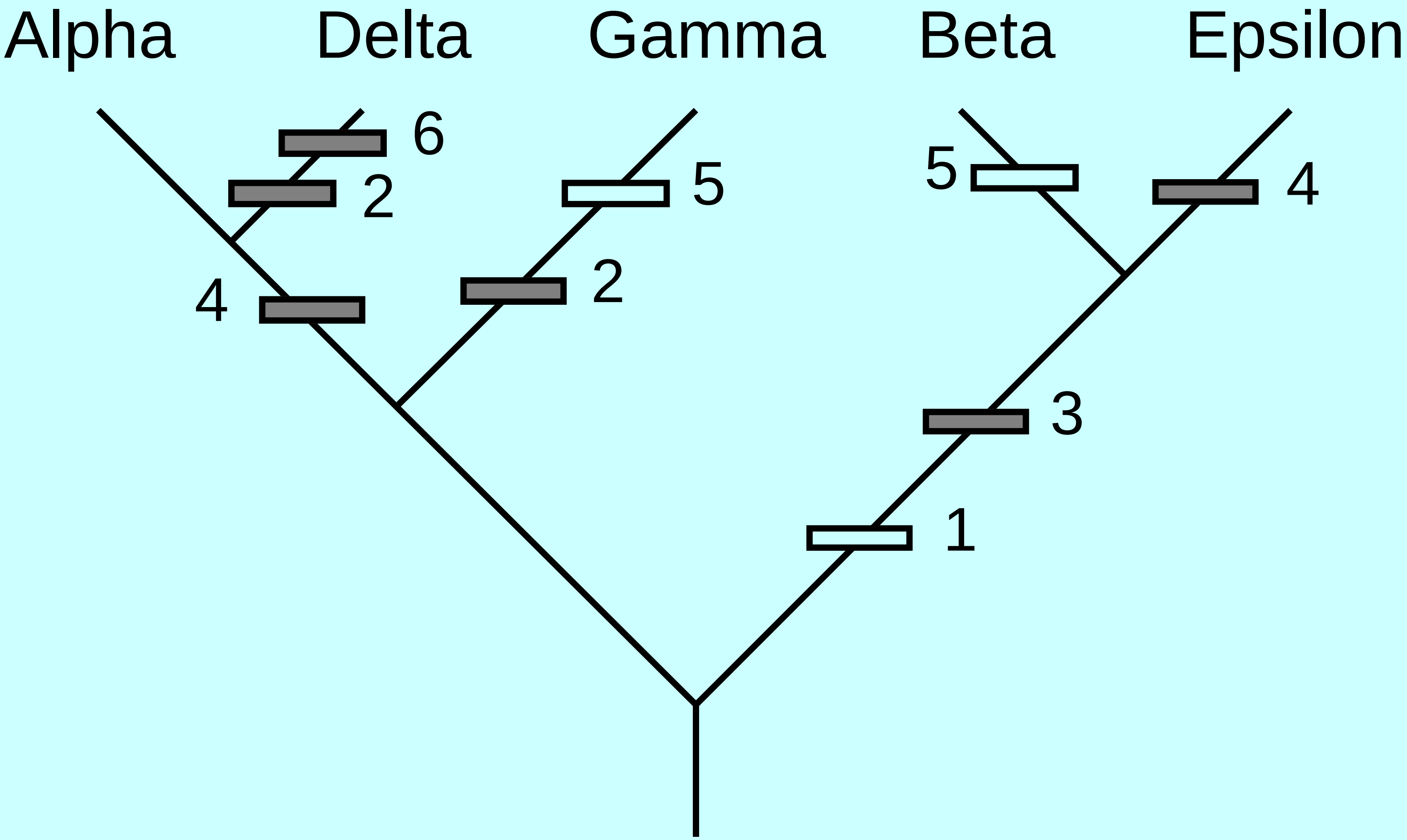
Steps in character 6

	1	2	3	4	5	6
Alpha	A	T	G	A	G	C
Beta	C	T	C	T	A	C
Gamma	A	G	G	T	A	C
Delta	A	G	G	A	G	T
Epsilon	C	T	C	A	G	C



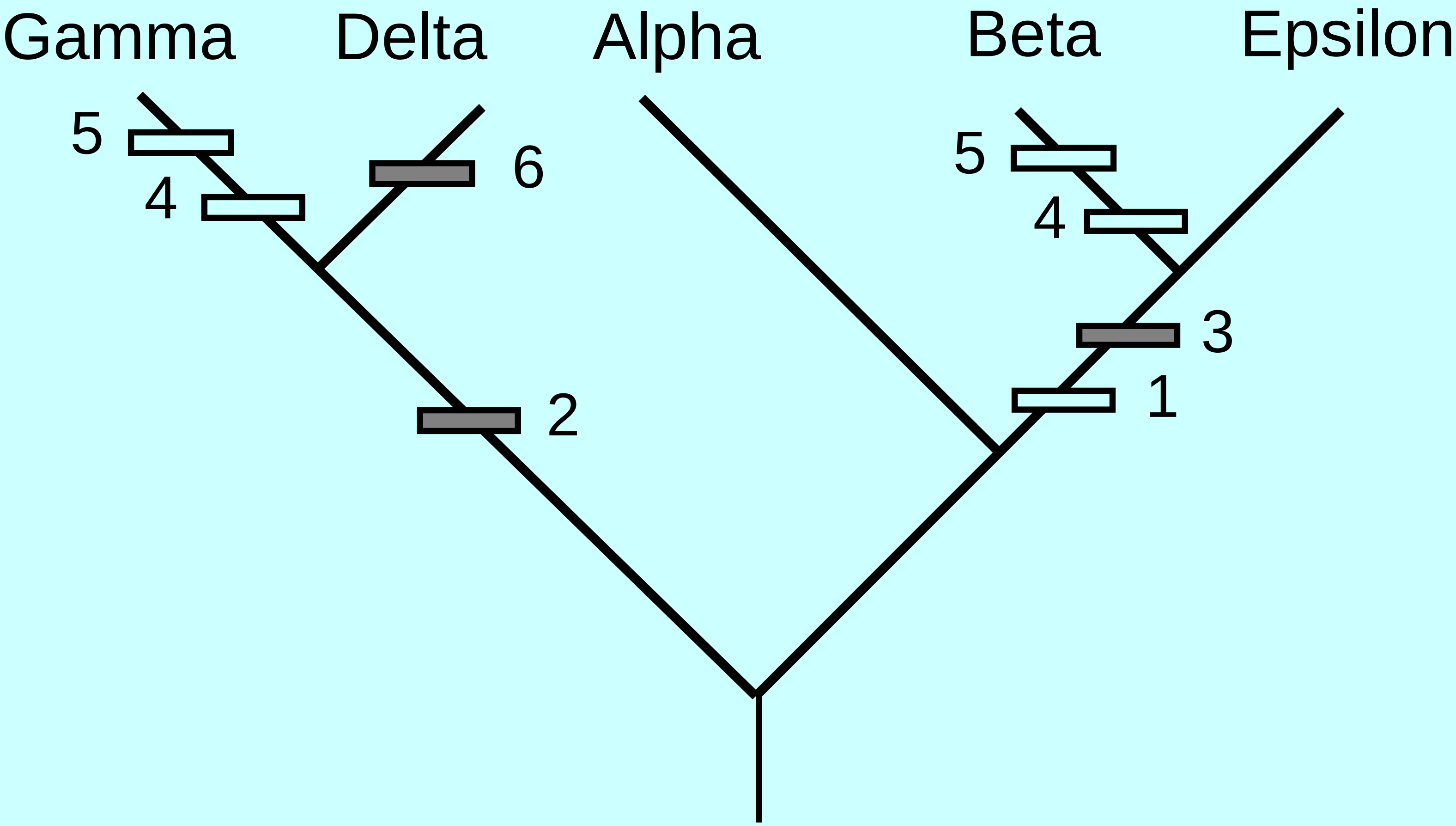
Steps in all characters

showing one of their possible placements



The most parsimonious tree

with one possible placement of the changes

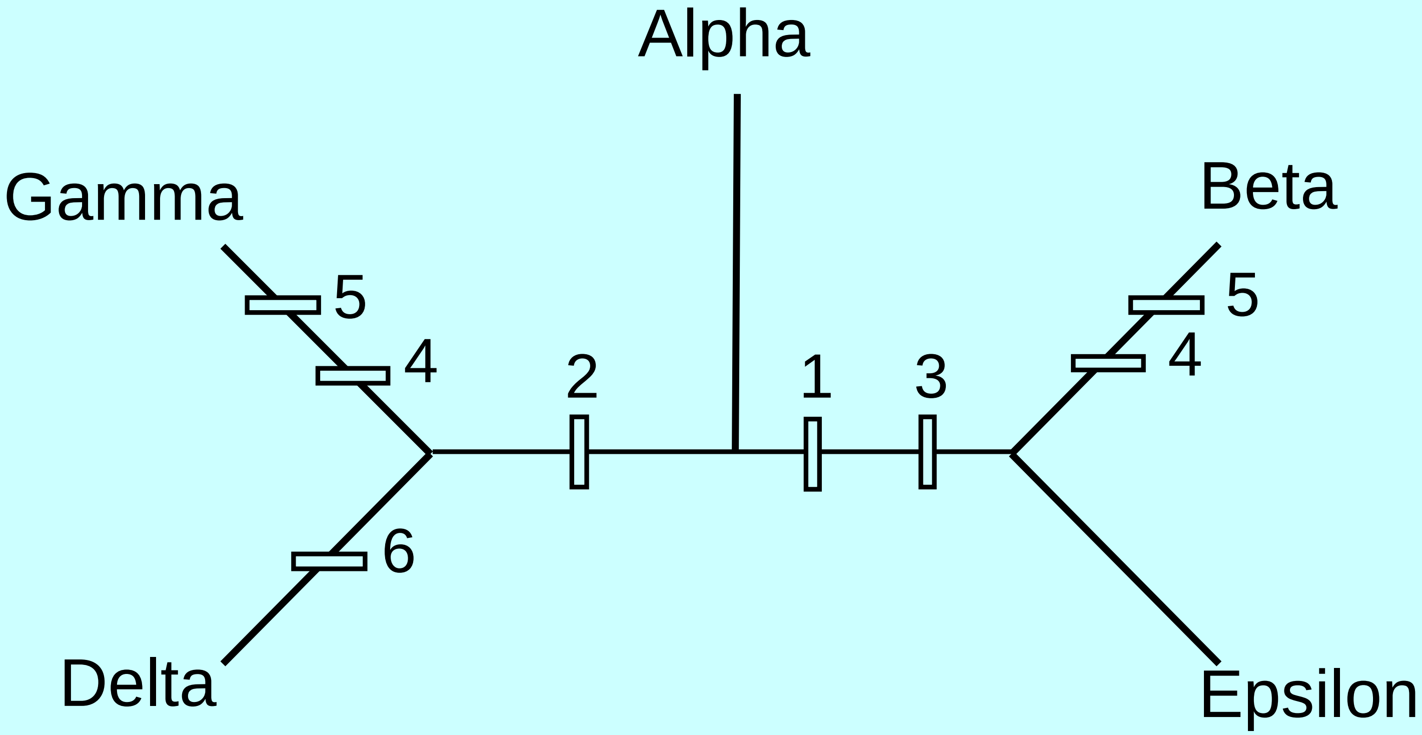


The same tree as an unrooted tree

shown as an unrooted tree

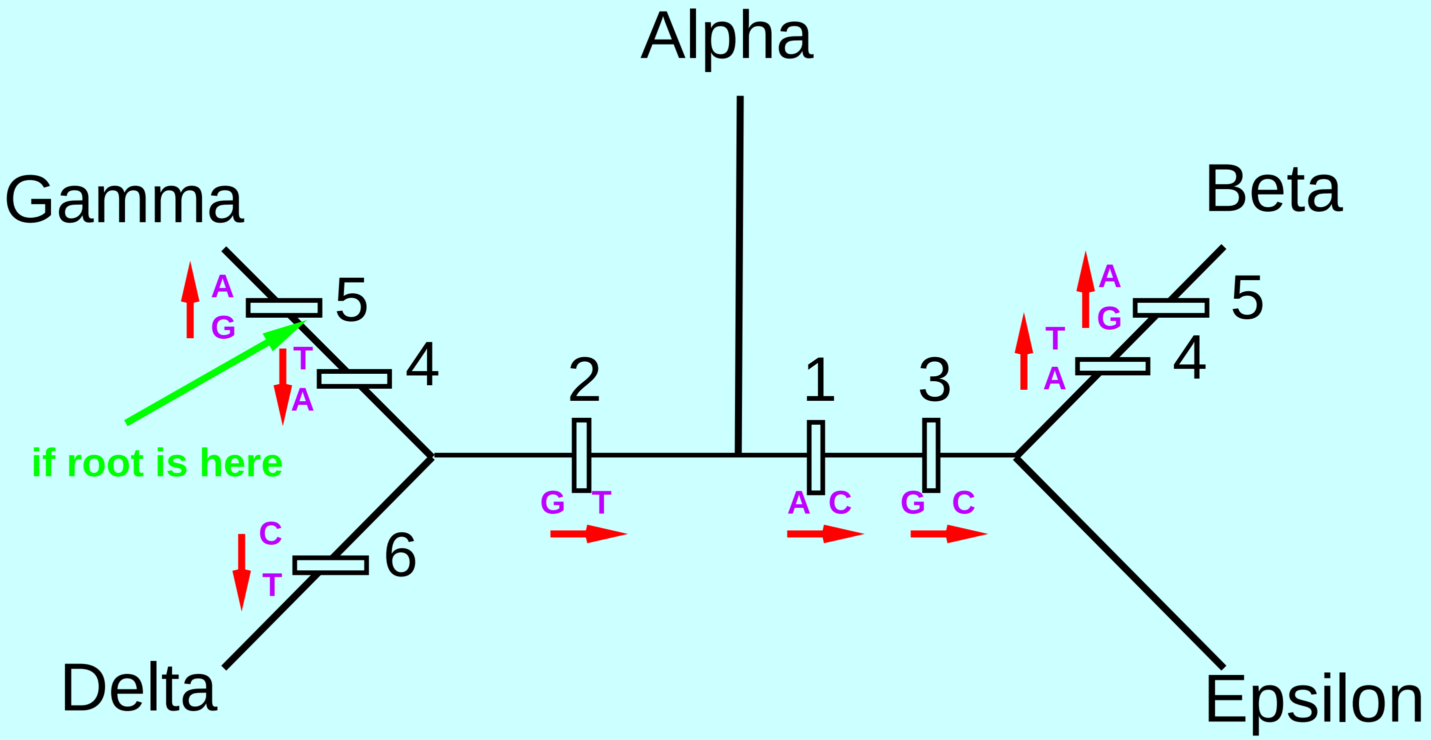
root can be anywhere

changes can occur in either direction



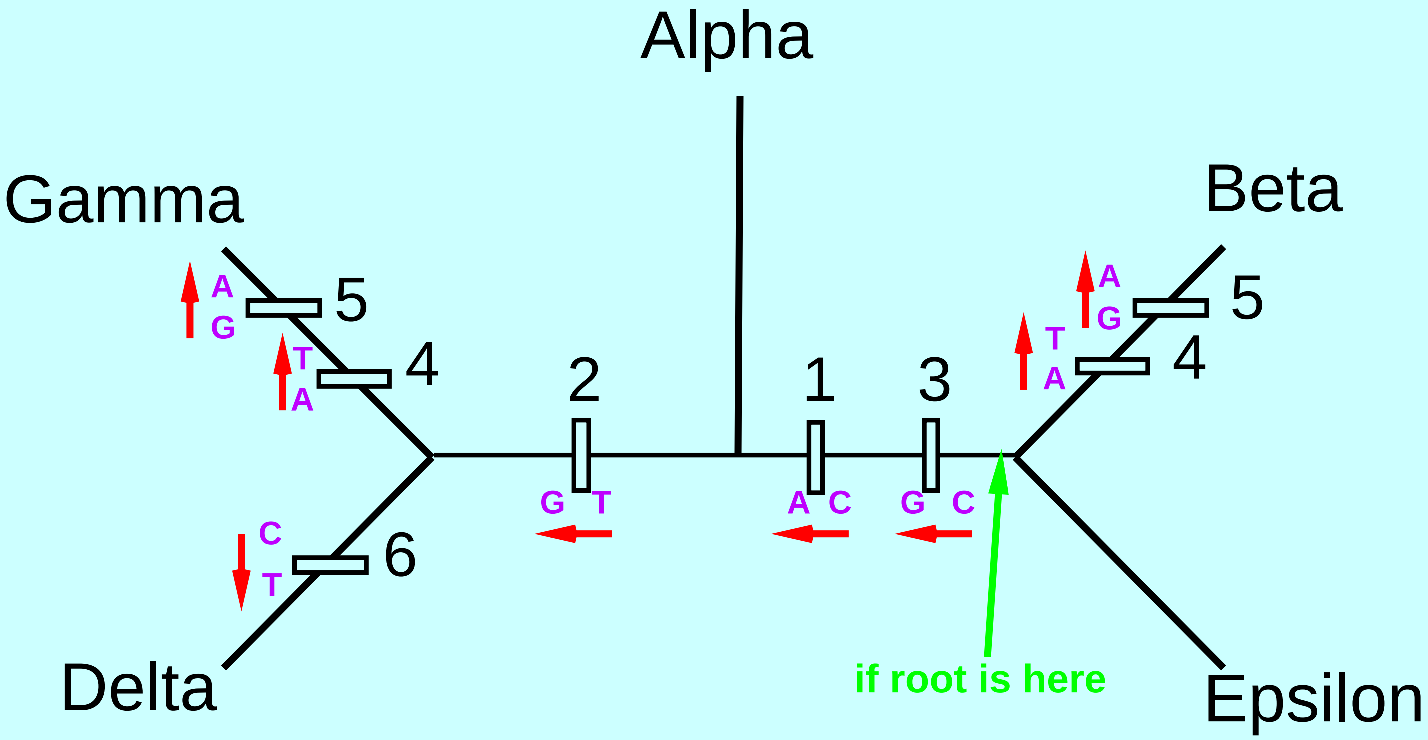
Direction of change depends on where root is

Placement of the root
affects which way bases change
but not how many changes there are

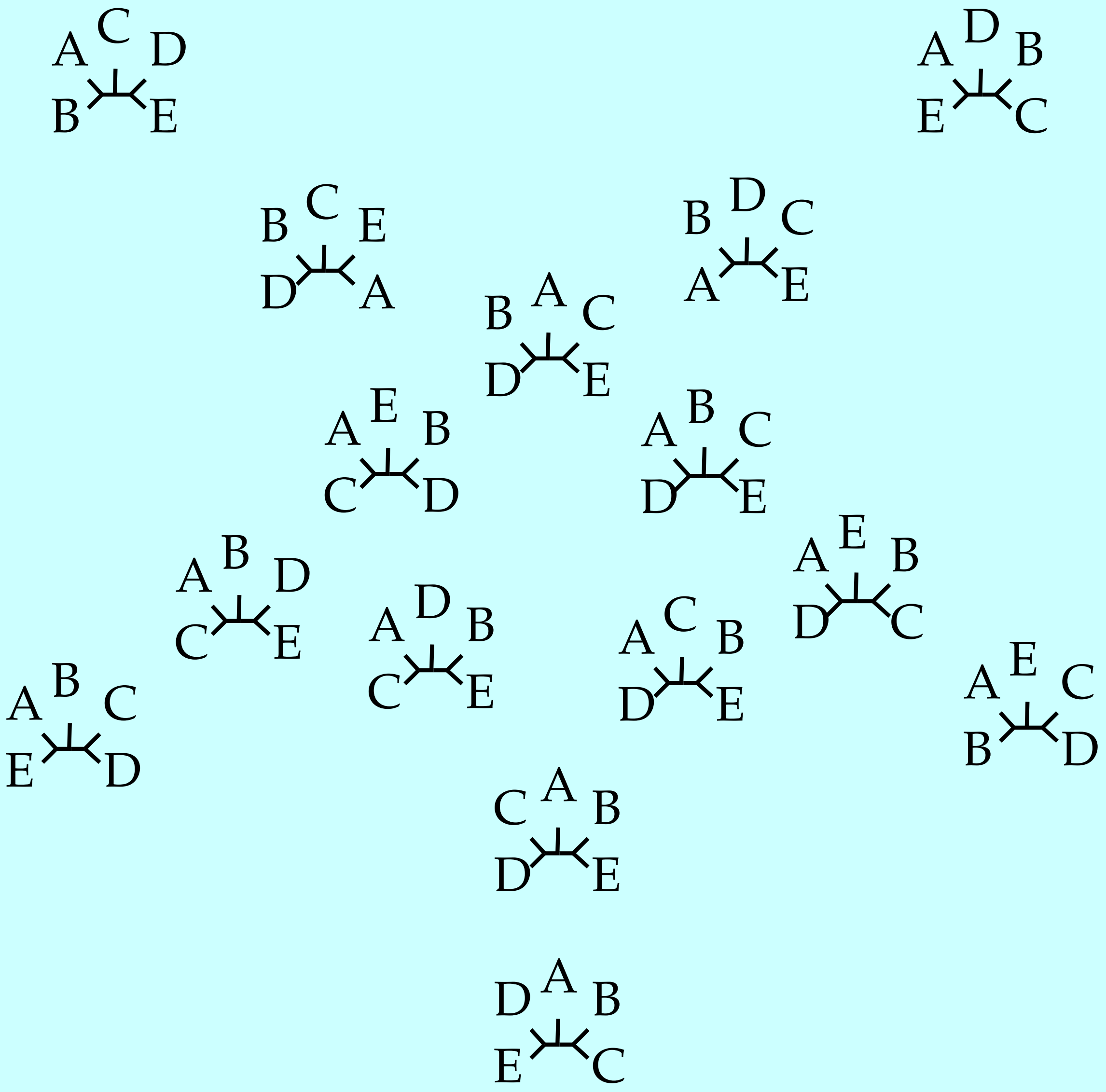


Changing the root changes the direction

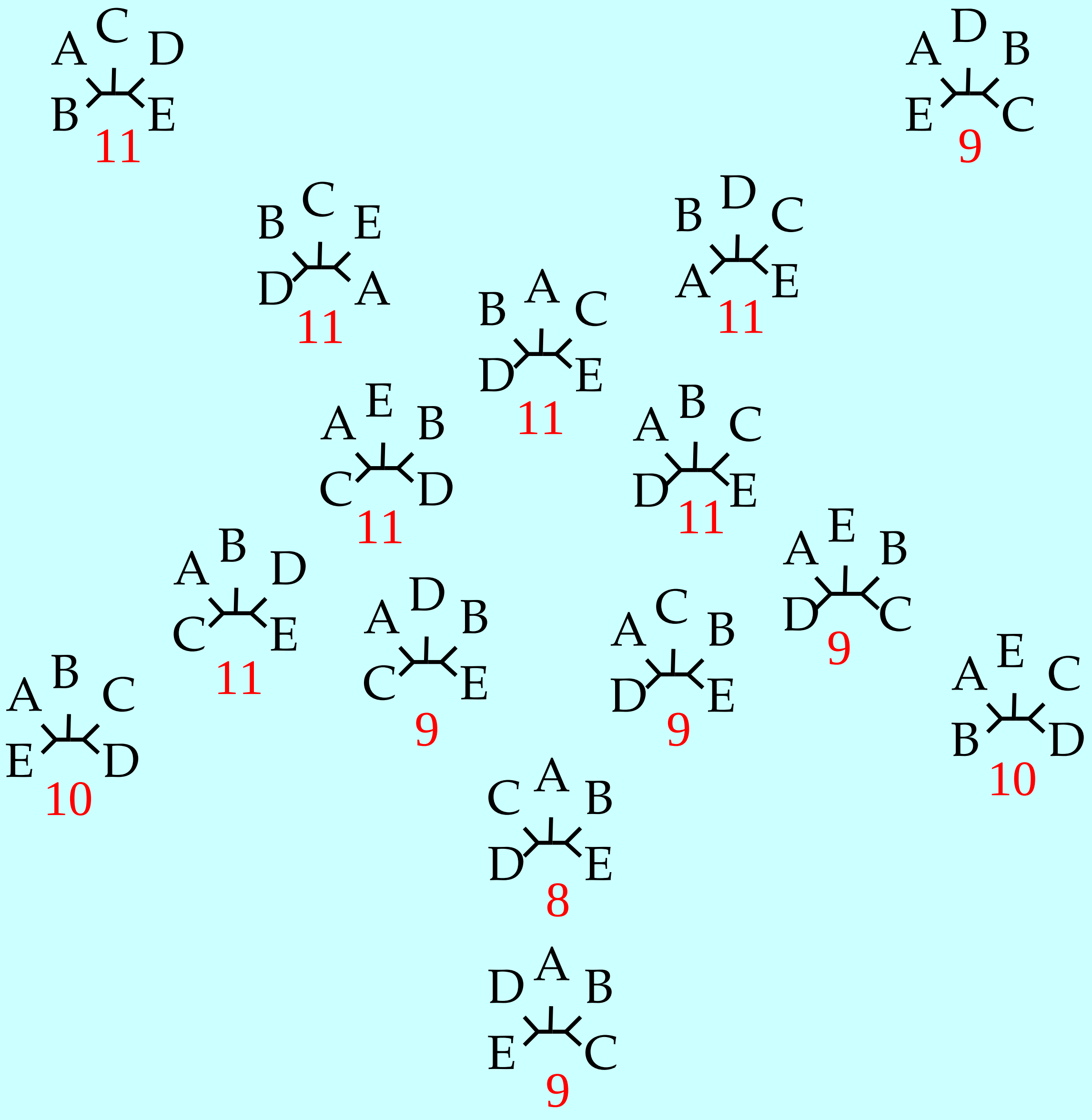
Placement of the root
affects which way bases change
but not how many changes there are



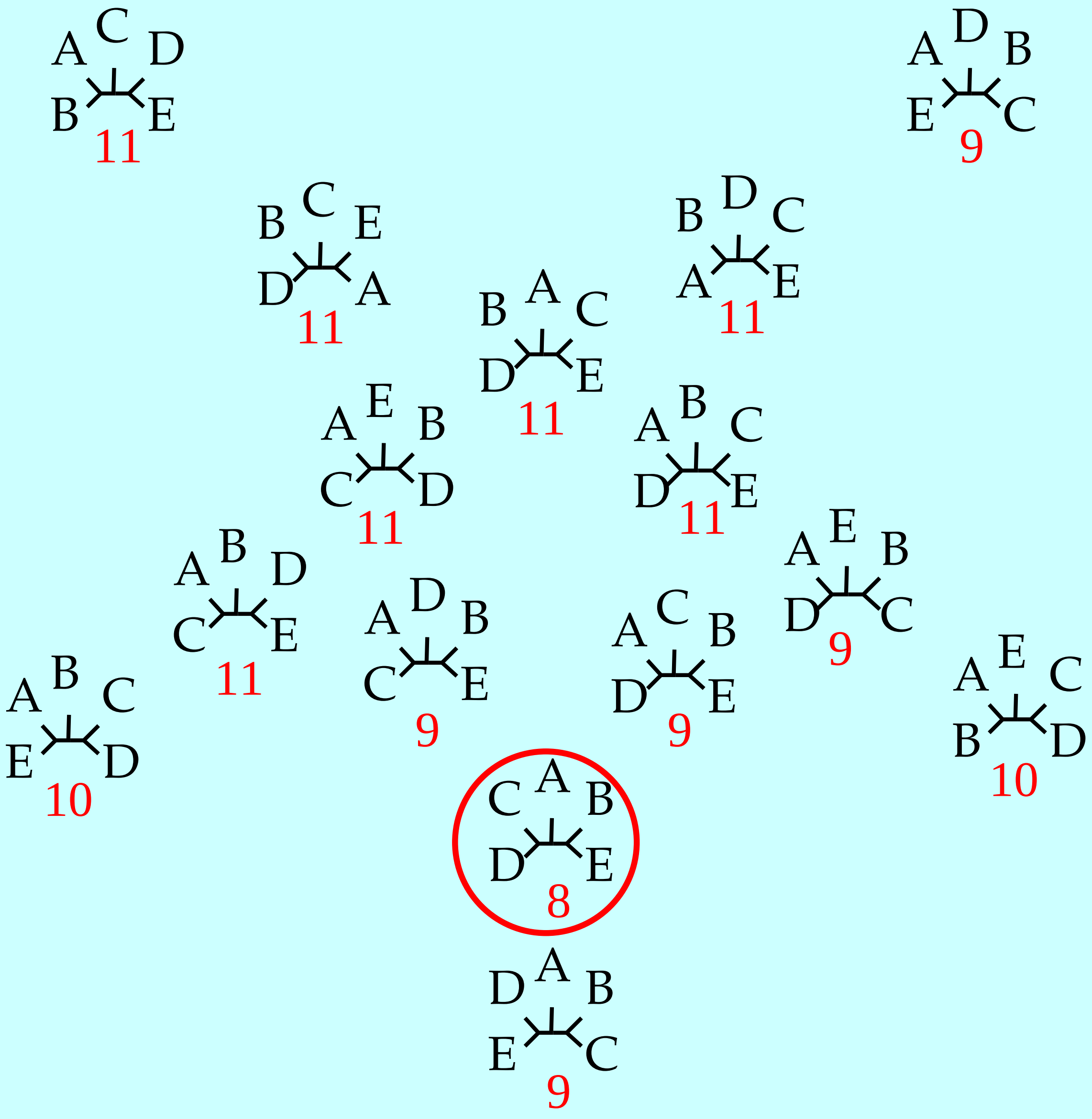
All possible trees (15 in all)



Their best numbers of nucleotide substitutions

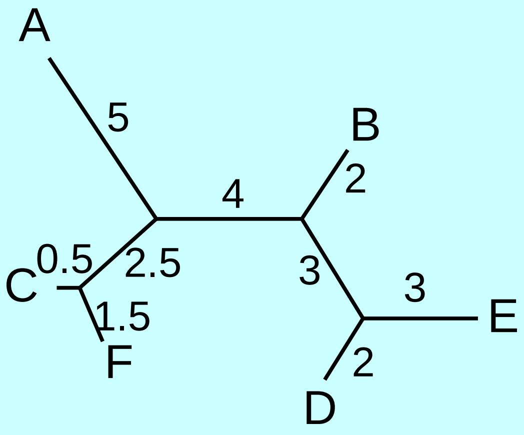


The most parsimonious tree



Distance matrix methods

Each possible tree (with branch lengths) predict pairwise distances



	A	B	C	D	E	F
A	0	11	8	14	15	9
B	11	0	9	7	8	10
C	8	9	0	13	14	2
D	14	7	13	0	5	13
E	15	8	14	5	0	14
F	9	10	2	13	14	0

Find the tree which comes closest to predicting

the observed pairwise distances

compare

observed distances
calculated from
the data

	A	B	C	D	E	F
A	0	10	9	12	16	9
B	10	0	10	6	9	9
C	9	10	0	10	15	2
D	12	6	10	0	6	13
E	16	9	15	6	0	15
F	9	9	2	13	15	0

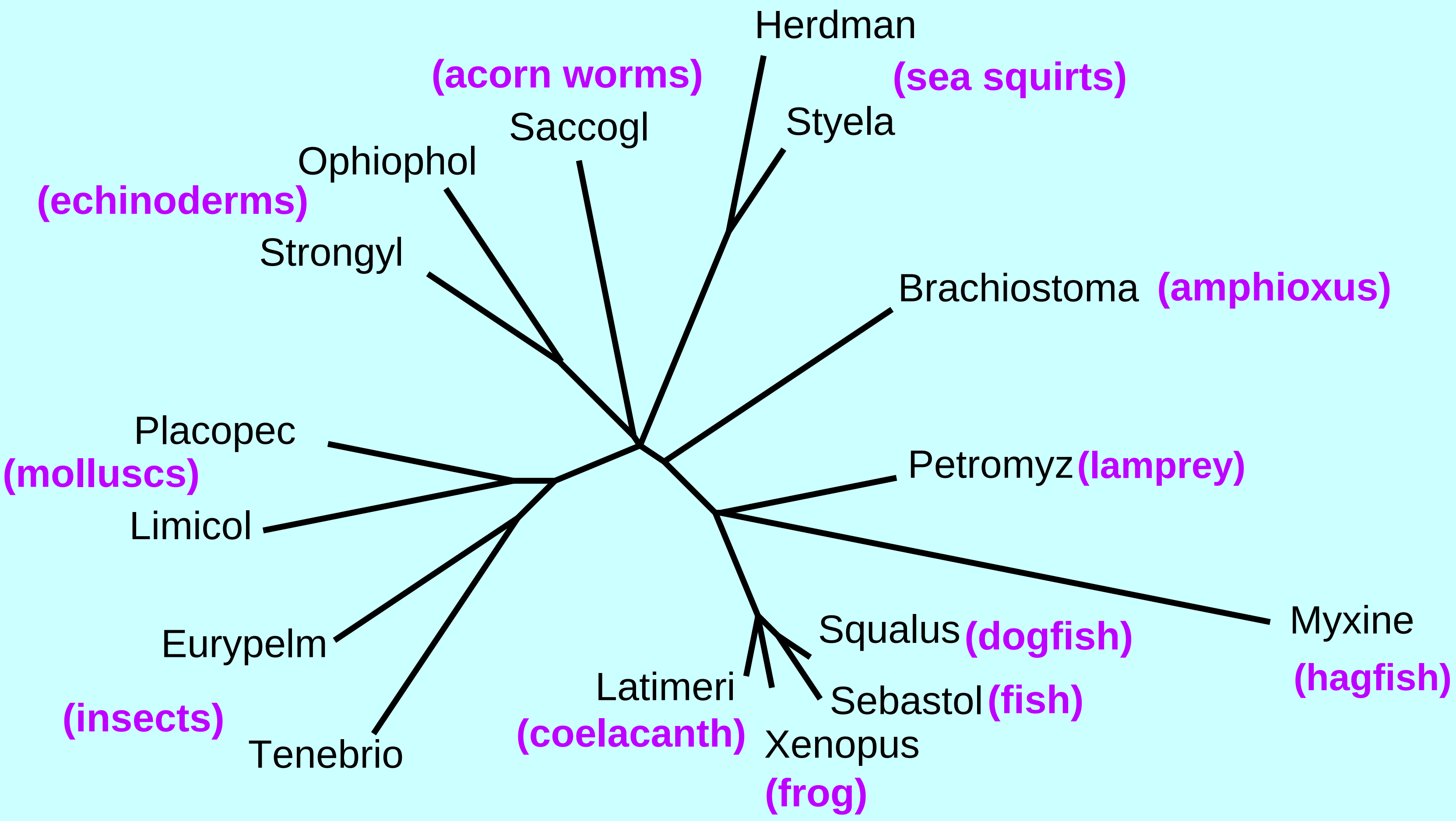
An example

Turbeville. J. McC., Schulz, J .R. and R. A. Raff. 1994. Deuterostome phylogeny and the sister group of the chordates: evidence from molecules and morphology. *Molecular Biology and Evolution* **11:** 648-655.

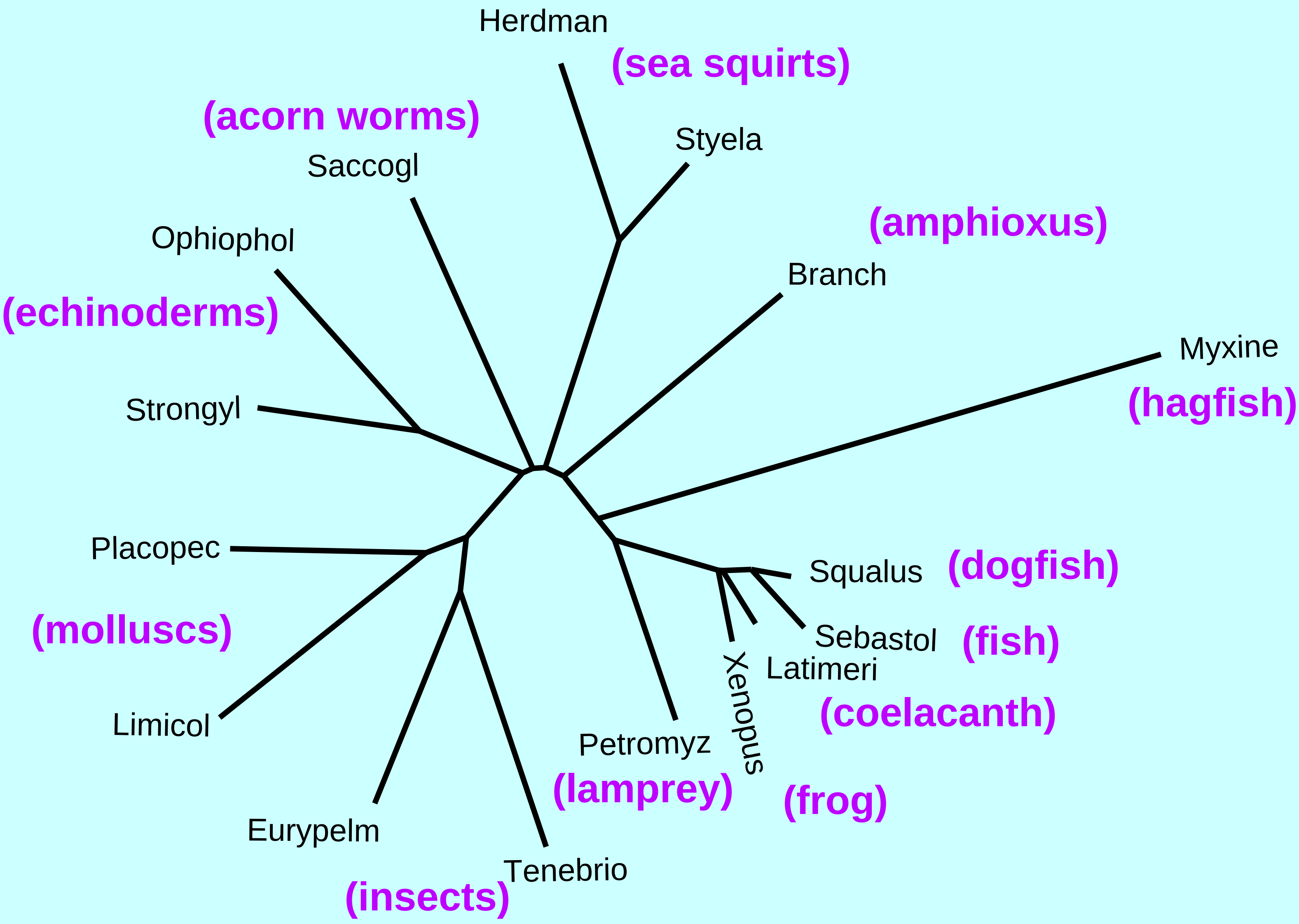
```
Xenopus      ?TACCTGGTTGATCCTGCCAGTAG-CATATGCTTGTCTCAAAGATTAAGCCATGCACG
Sebastol     ??????????????????????AG-CATATGCTTGTCTCAAAGATTAAGCCATGCAAG
Latimeri     ?TACCTGGTTGATCCTGCCAGTAG-CATATGCTTGTCTCAAAGATTAAGCCATGCATG
Squalus      ??????????????????????AG-CATATGCTTGTCTCAAAGATTAAGCCATGCATG
Myxine       ??CCCTGGTTGATCCTGCCAGCCG-CATATGCTTGTCTCAAAGACTAAGCCATGCATG
Petromyz     ???CCTGGTTGATCCTGCCAGTAG-CATATGCTTGTCTCAAAGATTAAGCCATGCATG
Branch       ???CCTGGTTGATCCTGCCAGTAGTCATATGCTTGTCTCAAAGATTAAGCCATGCACG
Styela       ??ATCTGGTTGATCCTGCCAGTAGTGATATGCTTGTCTCAAAGATTAAGCCATGCAGG
Herdman      ?TATCTGGTTGATCCTGCCAGTAGTGATATGCTTGTCTCAA-GATTAAGCCATGCAGG
Saccogl      ??ACCTGGTTGATCCTGCCAGTAGTCATATGCTTGTCTCAAAGATTAAGCCATGCATG
Ophiophol    ??ACCTGGTTGATCCTGCCAGTAGTCATATGCTTGTCTCAAAGATTAAGCCATGCATG
Strongyl     ??ACCTGGTTGATCCTGCCAGTAGTCATATGCTTGTCTCAAAGATTAAGCCATGCATG
Placopec     CAACCTGGTTGATCCTGCCAGTAGTCATATGCTTGTCTCAAAGATTAAGCCATGCATG
Limicol      ?TATCTGGTTGATCCTGCCAGTAGTCATATGCTTGTCTCAAAGATTAAGCCATGCATG
Eurypelm     ?TACCTGGTTGATCCTGCCAGTAGTCATATGCTTGTCTCAAAGATTAAGCCATGCATG
Tenebrio     ?TCCCTGGTTGATCCTGCCAGTAGTCATATGCTTGTCTCAAAGATTAAGCCATGCATG
```

```
(and so on for 33 more pages)
```

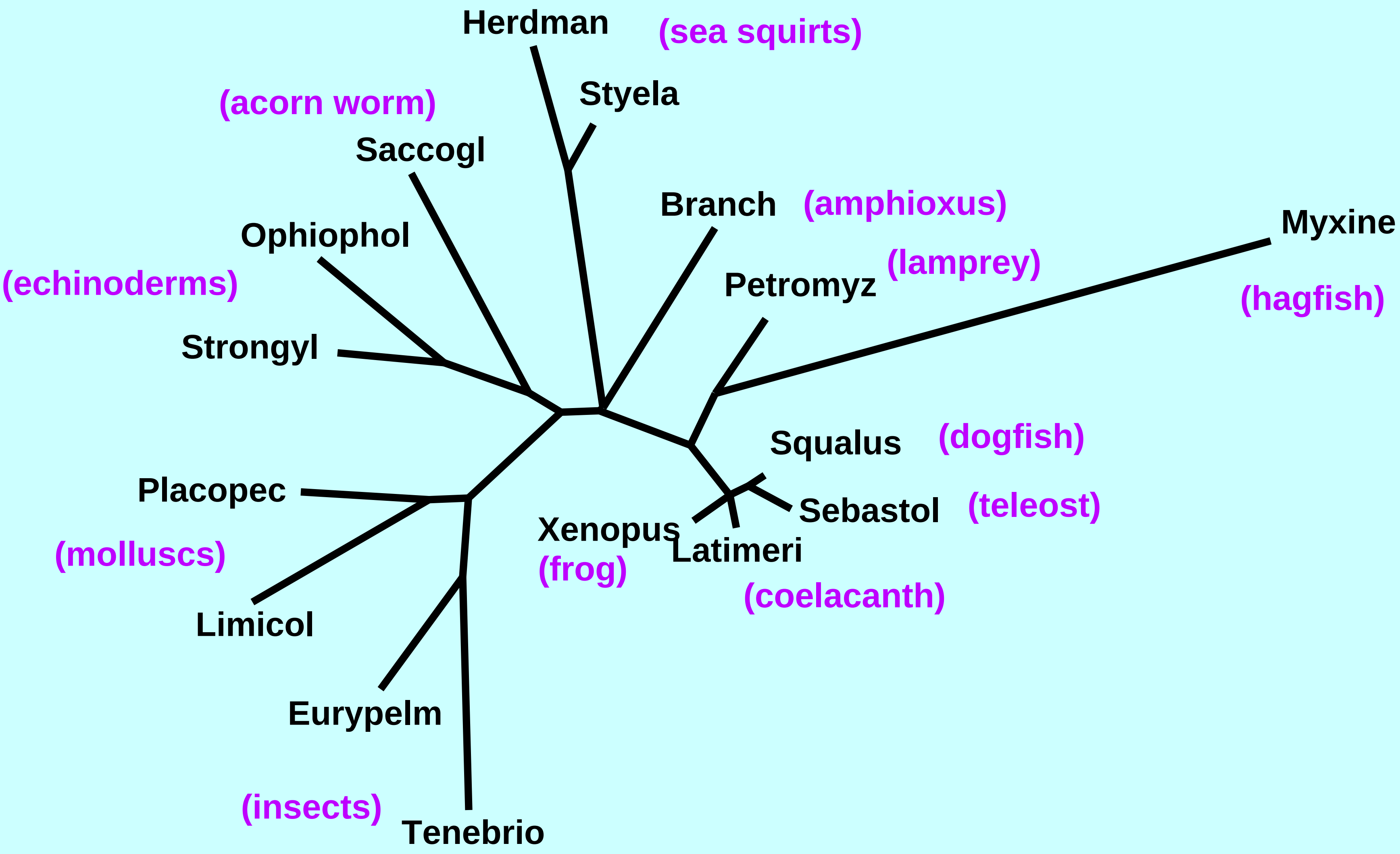
Tree using parsimony



Tree using a distance method



Tree using a maximum likelihood method

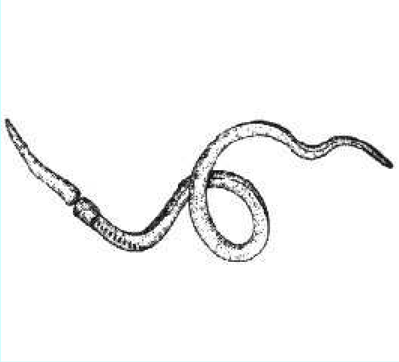


The tree with images of the animals

Placopecten
(scallop)

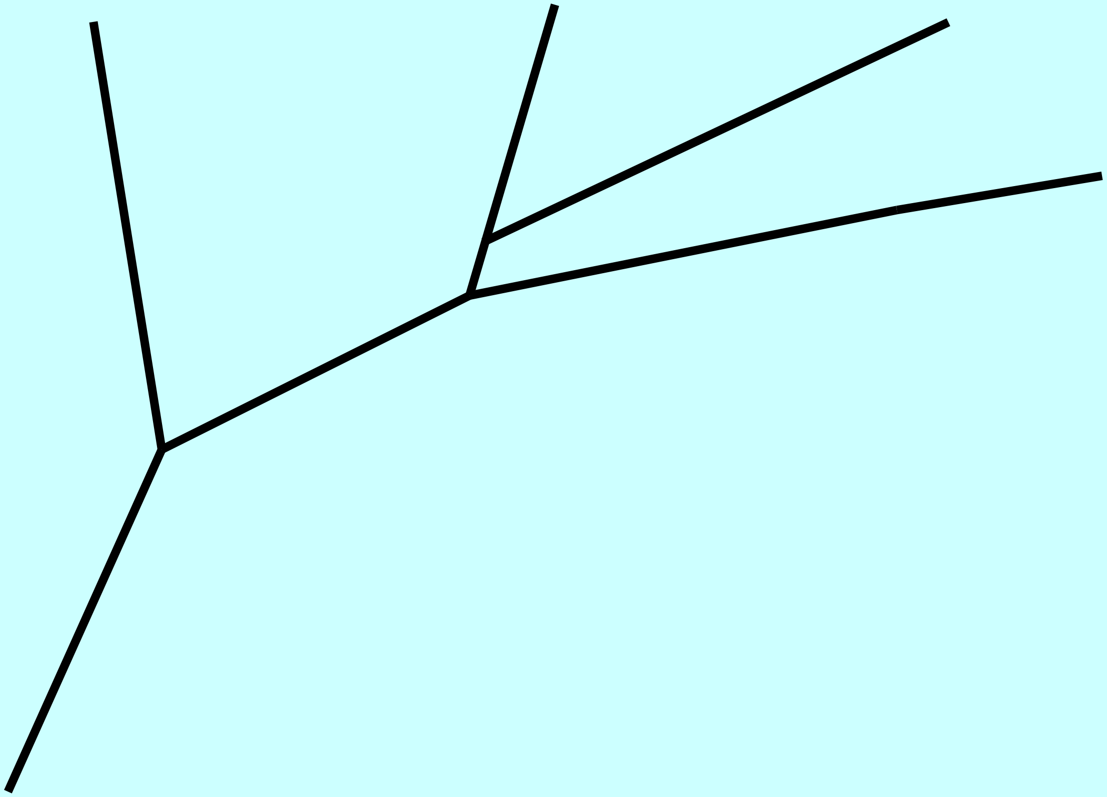
Strongylocentrotus

(sea urchin)



Saccoglossus

(acorn worm)



The tree with images of the animals

Branchiostoma

(amphioxus)

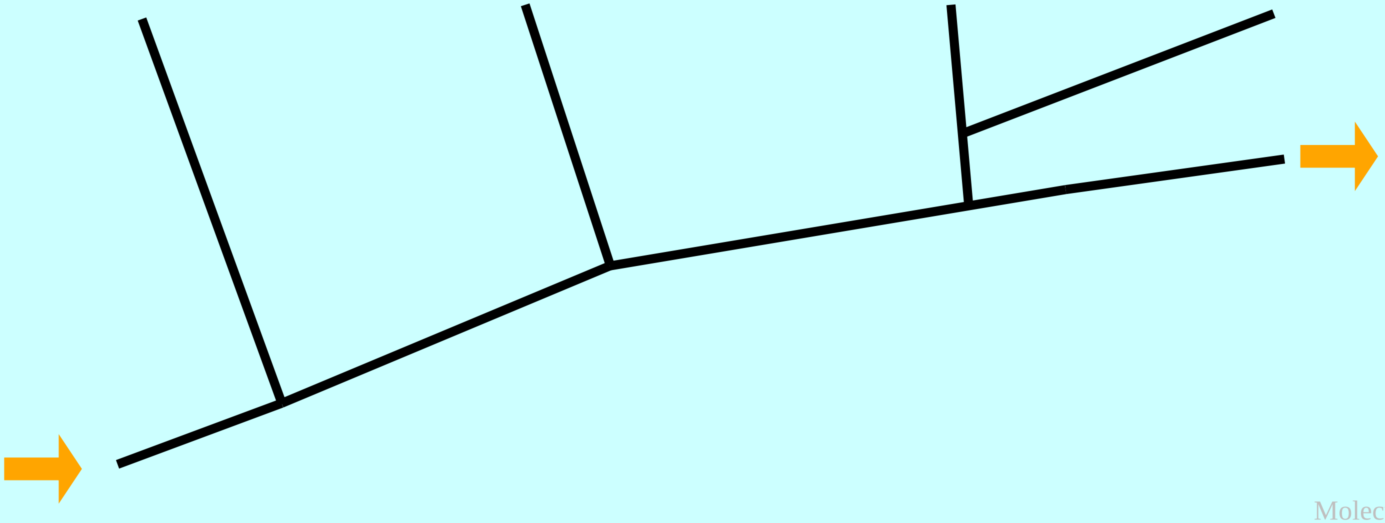
tunicate

Myxine

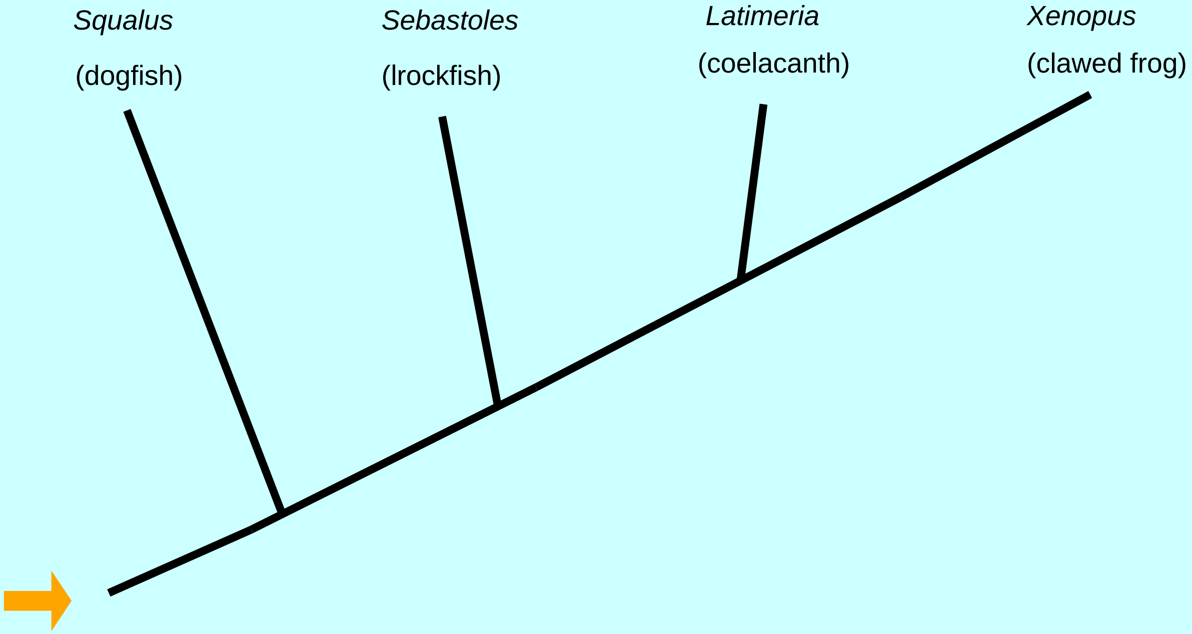
(hagfish)

Petromyzon

(lamprey)



The tree with images of the animals



Squalus

(dogfish)

Sebastoles

(lrockfish)

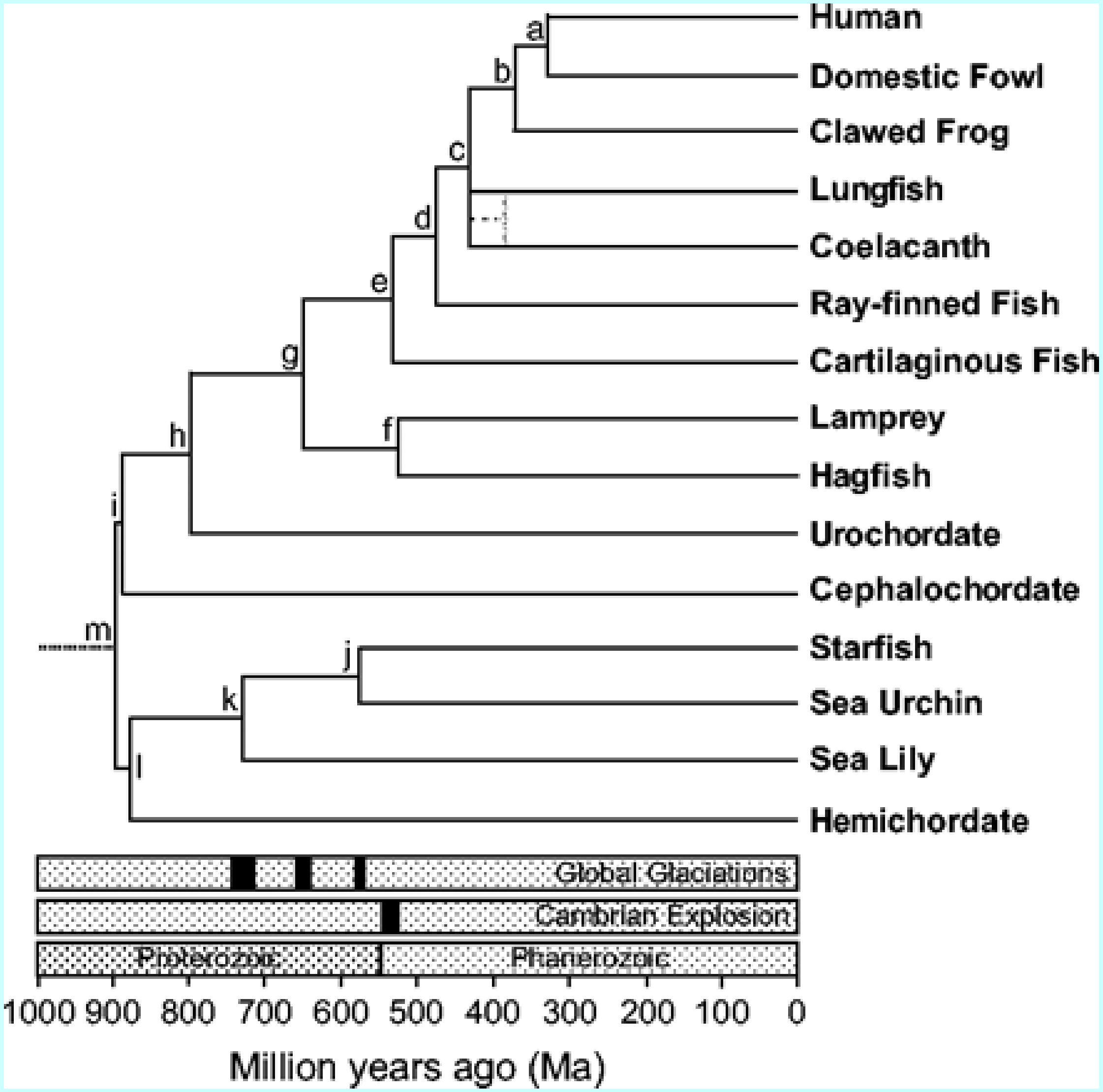
Latimeria

(coelacanth)

Xenopus

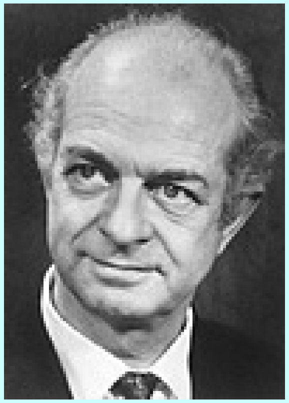
(clawed frog)

Blair and Hedges’ alternative tree



(in *Molecular Biology and Evolution*, 2005.)

Molecular evolution (1963 on)



Linus Pauling in 1963

Emile Zuckerkandl, more recently

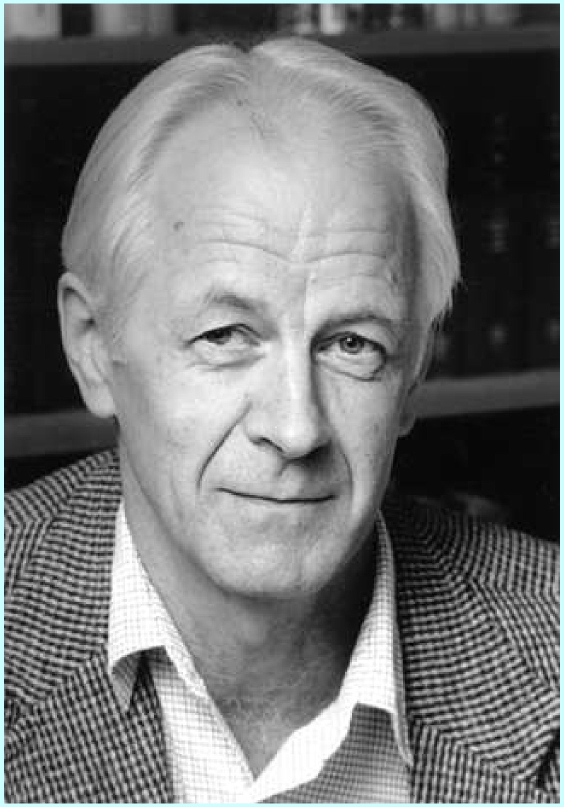
The late Margaret Dayhoff

Responsible for the first computer-produced molecular phylogeny (1966), the start of protein sequence databases (1965), the recognition of gene families (1960s-1970s)

Morris Goodman

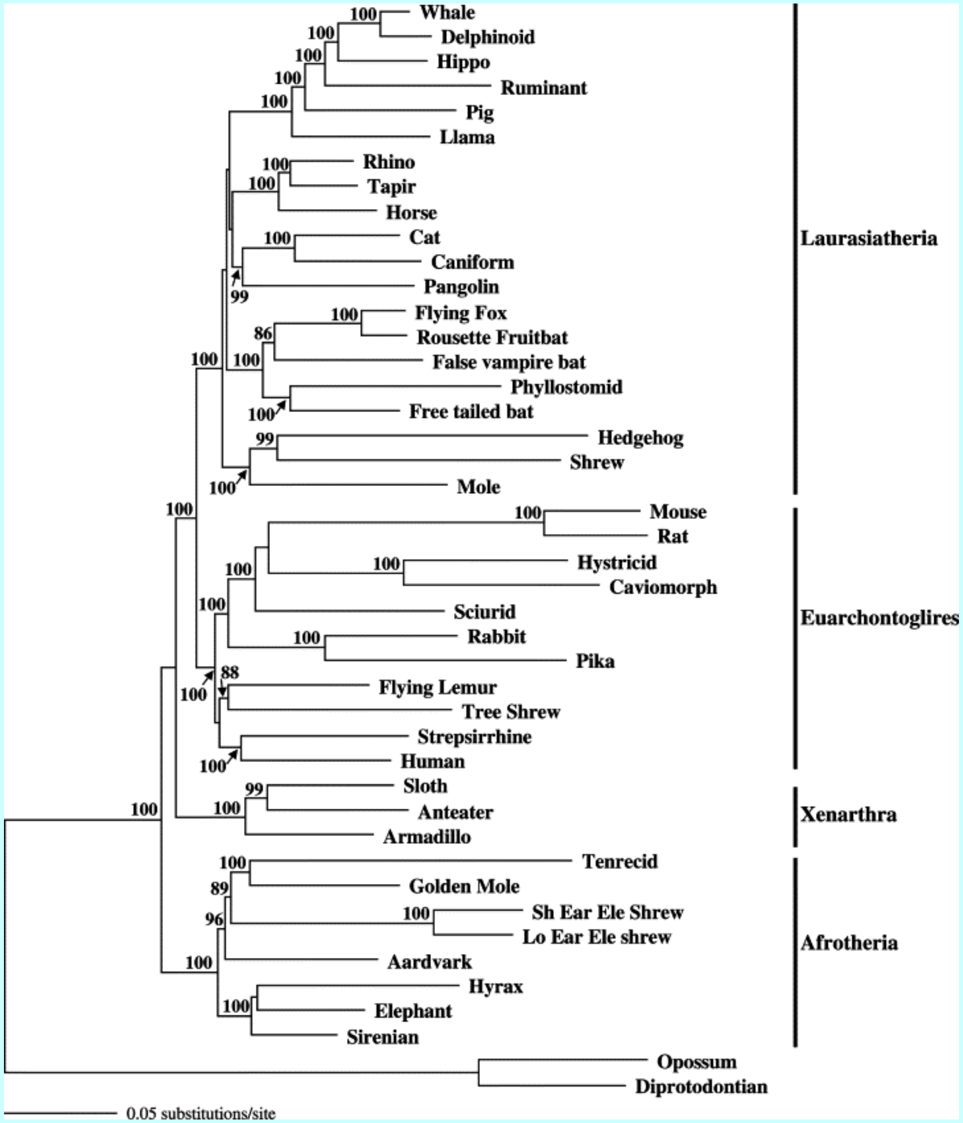
Using immunological methods with proteins, defined the human-chimp-gorilla clade (1962), later pioneered work on evolution of gene families, especially the globin family, and on the phylogeny of mammals.

The late Allan Wilson



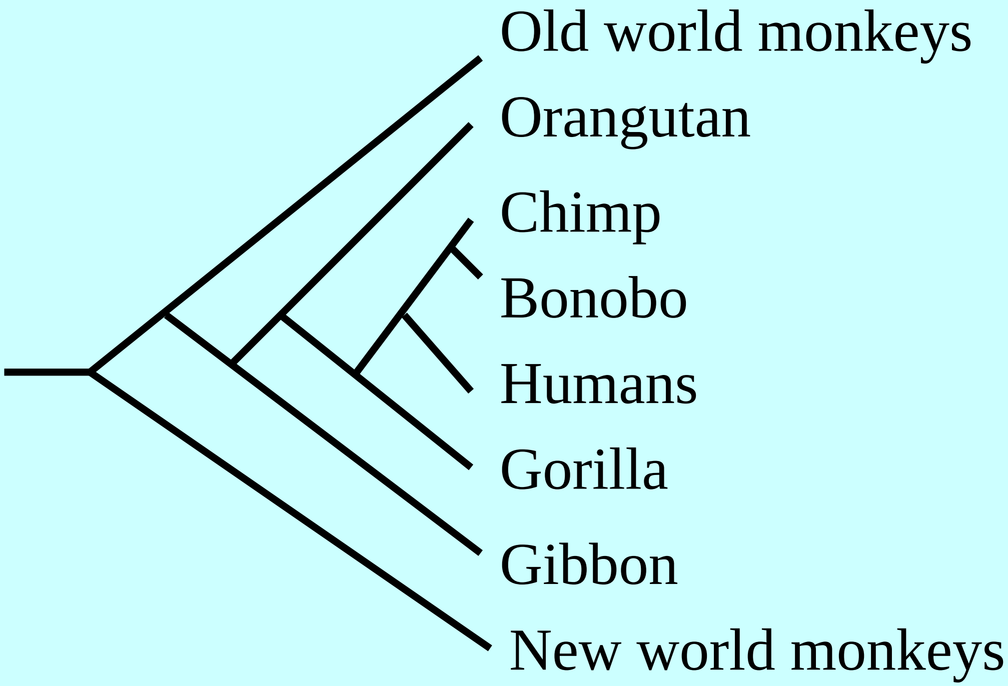
With Vincent Sarich, supported the human-chimp-gorilla clade. Advocated use of the “molecular clock” by which they dated the divergence of these species to 5 million years ago. Later (as we shall see) found the tree of human mitochondria, whose ancestor was “mitochondrial Eve”.

An example: who is most closely related to whales?

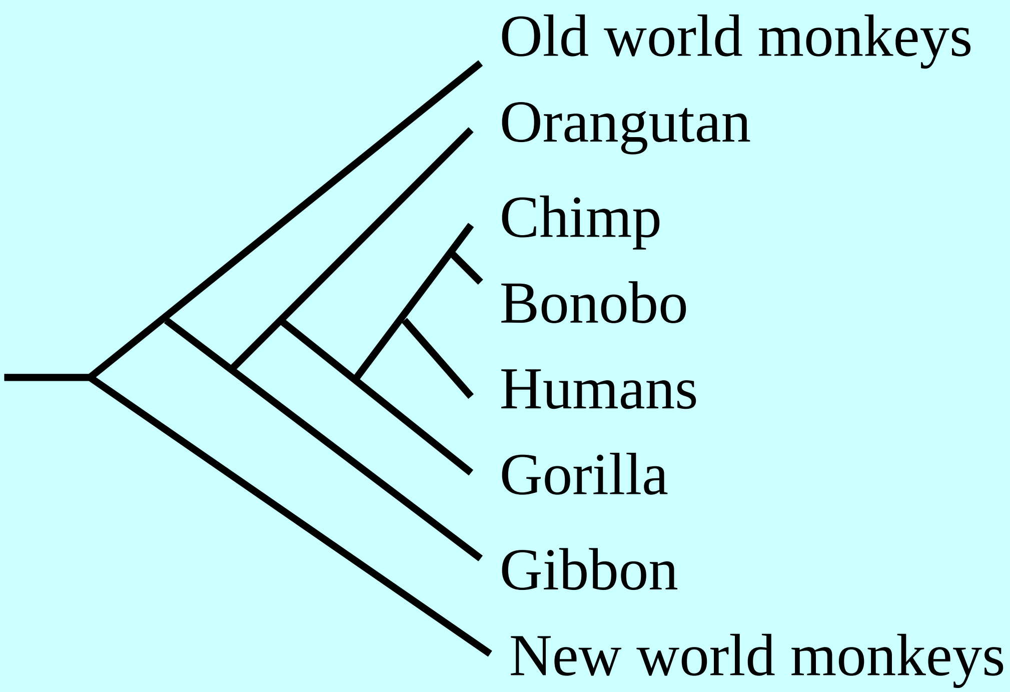


from Amrine-Madsen, H. et al., 2003, *Molecular Phylogenetics and Evolution*

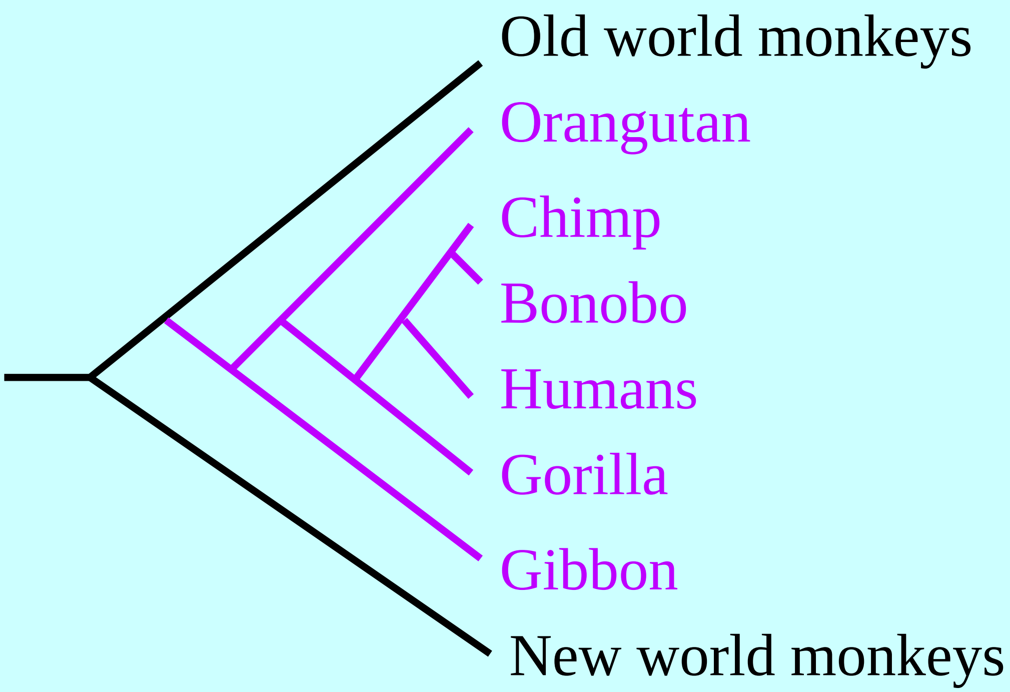
The tree of human ancestry



Just who are you calling an ape?

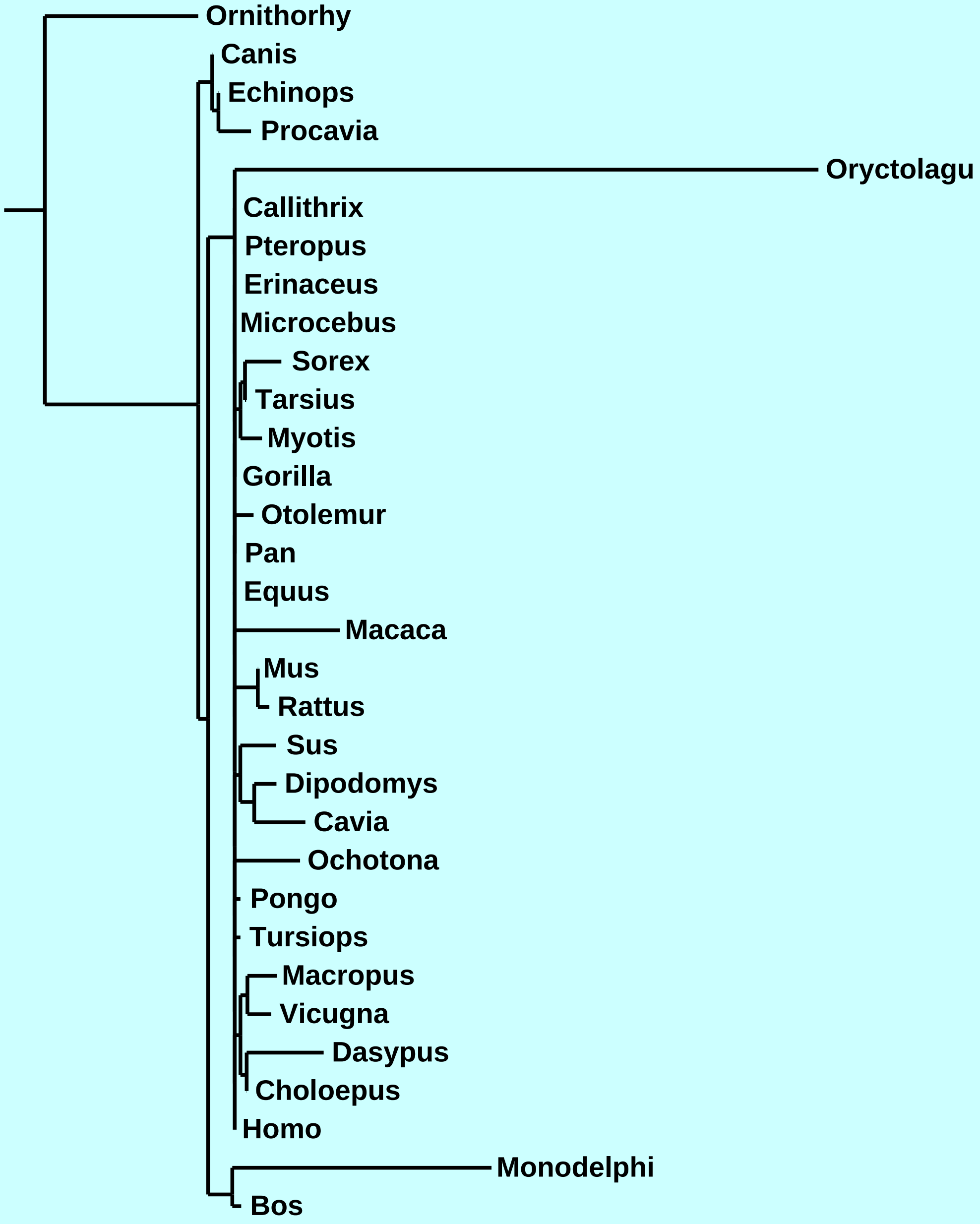


Just who are you calling an ape?

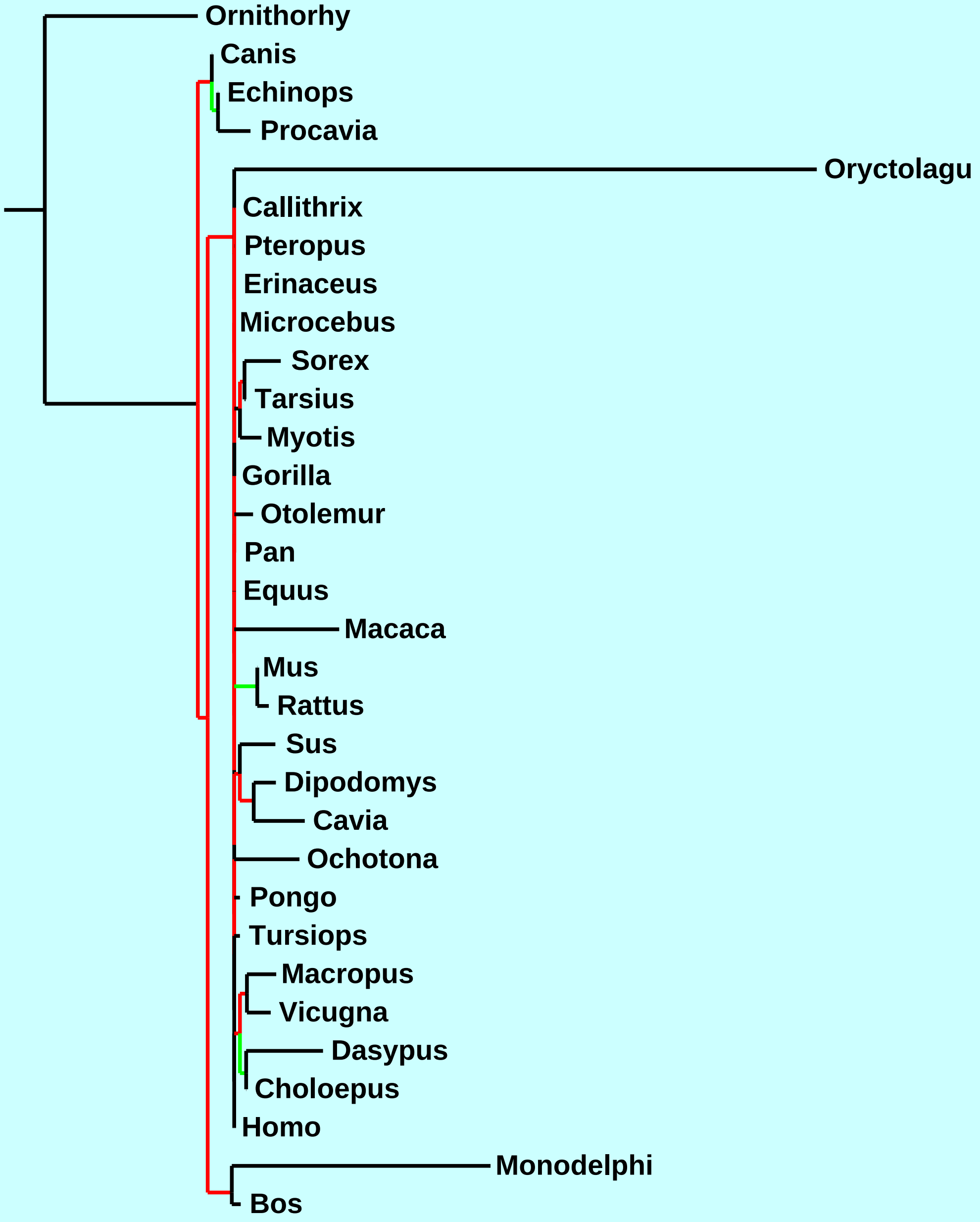


All of us, actually.

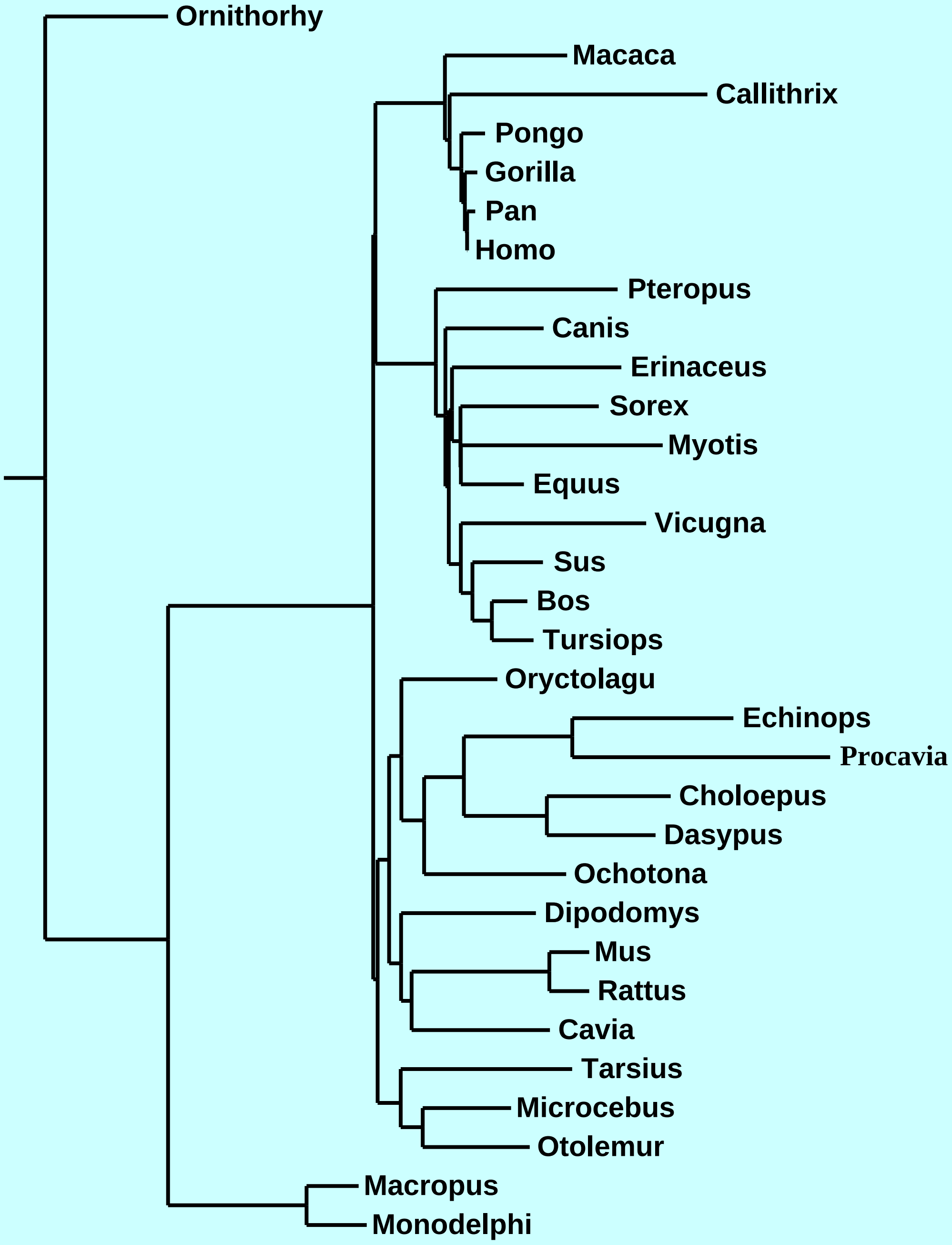
32 mammals from Homeobox-containing protein 1



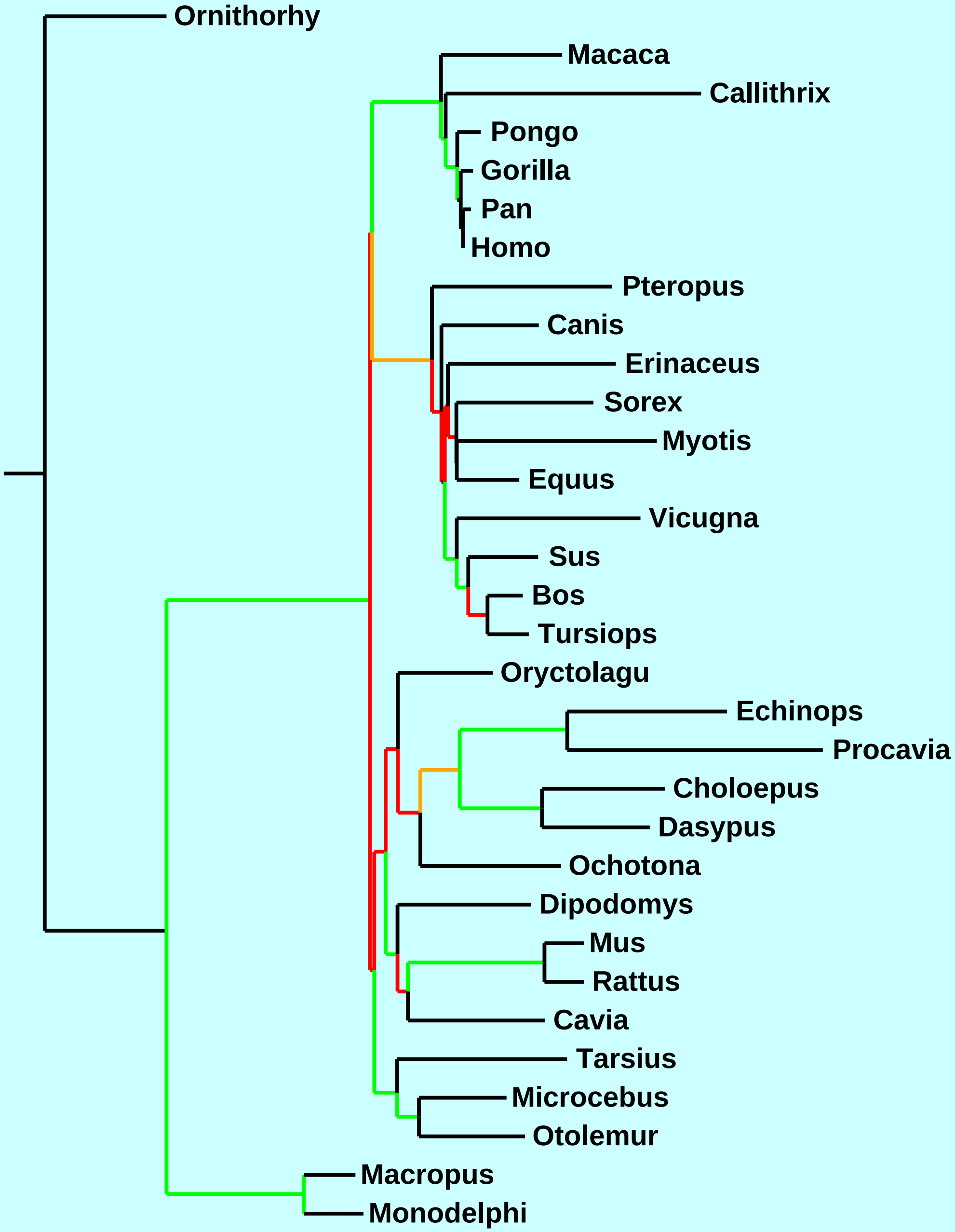
... coloring in branches that are or aren’t true



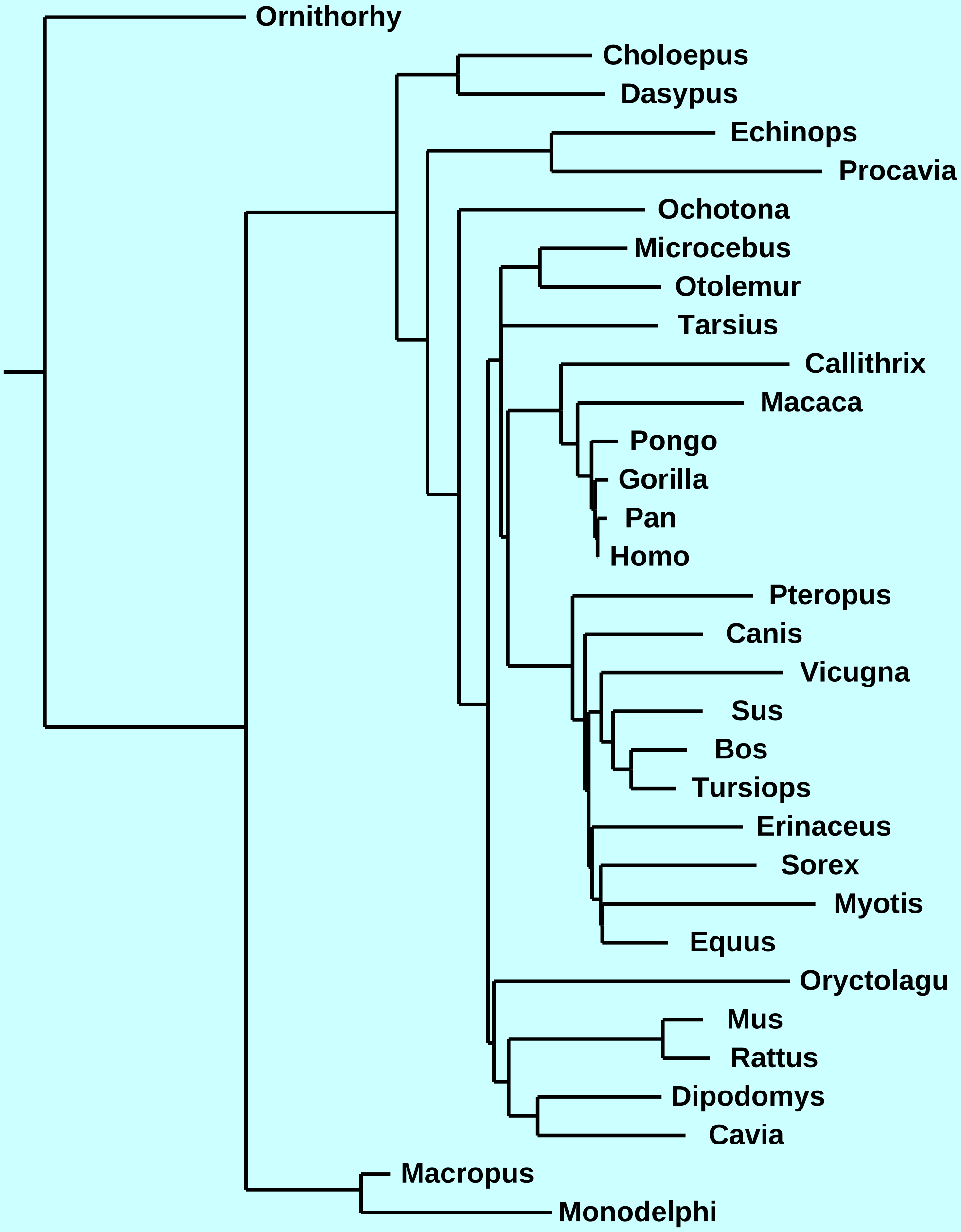
The same 32 using E3 ubiquitin-protein ligase



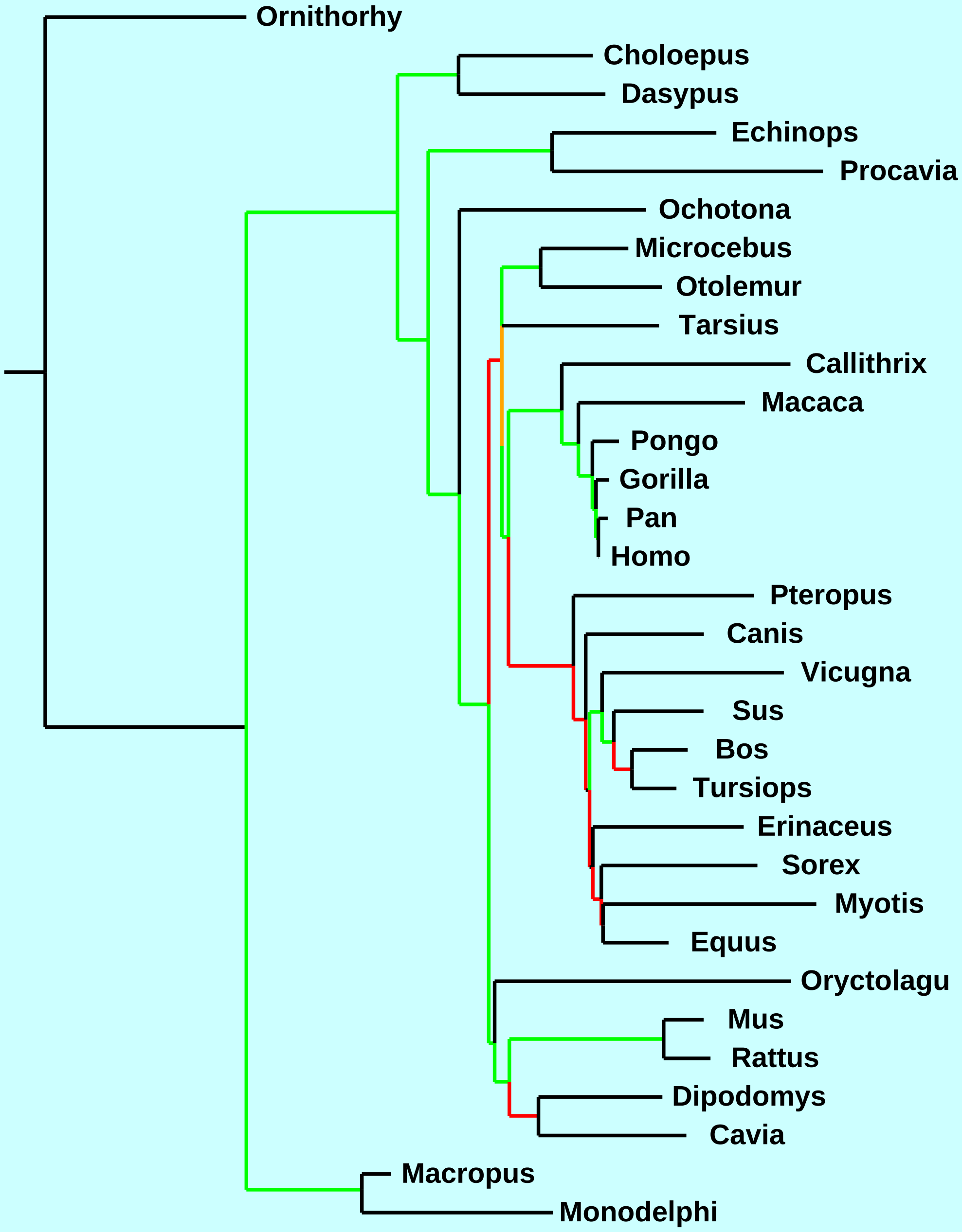
... does considerably better



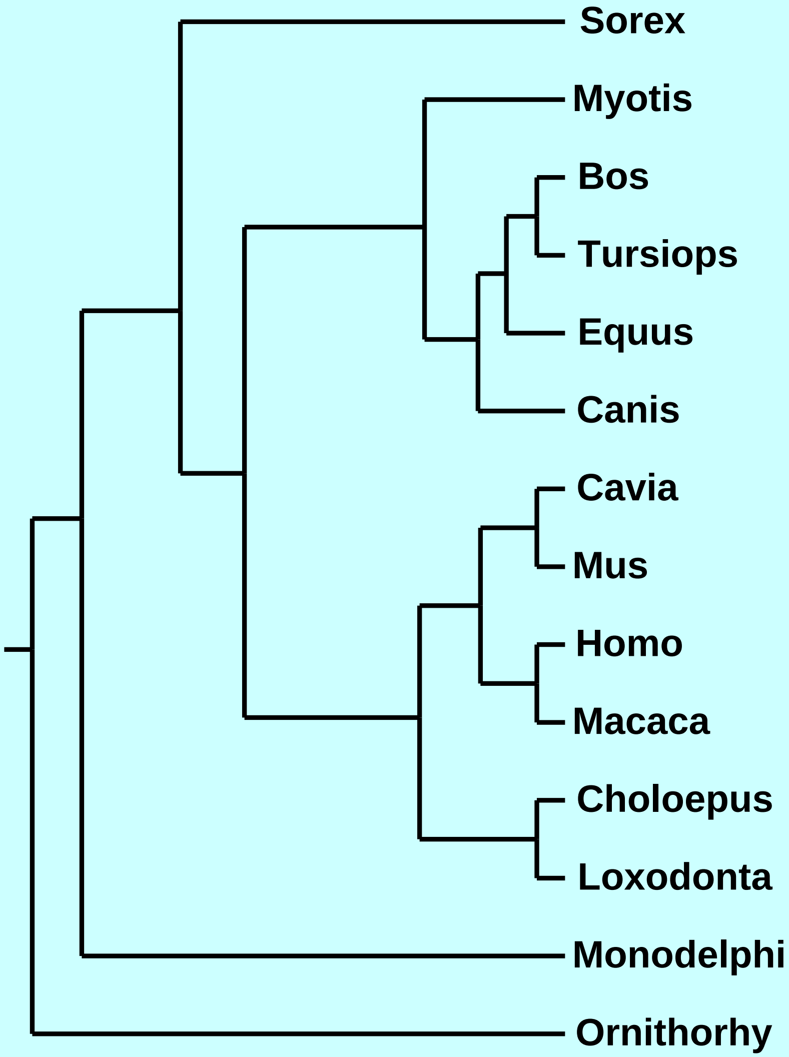
But using both of these loci ...



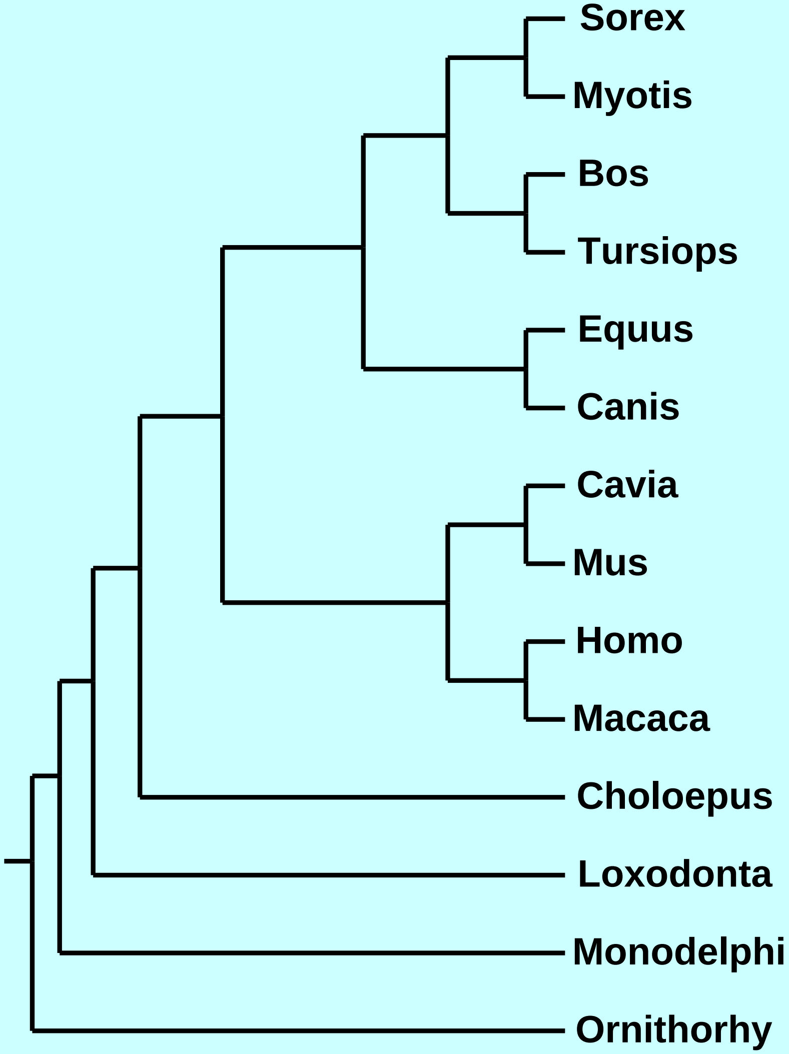
... is not bad at all!



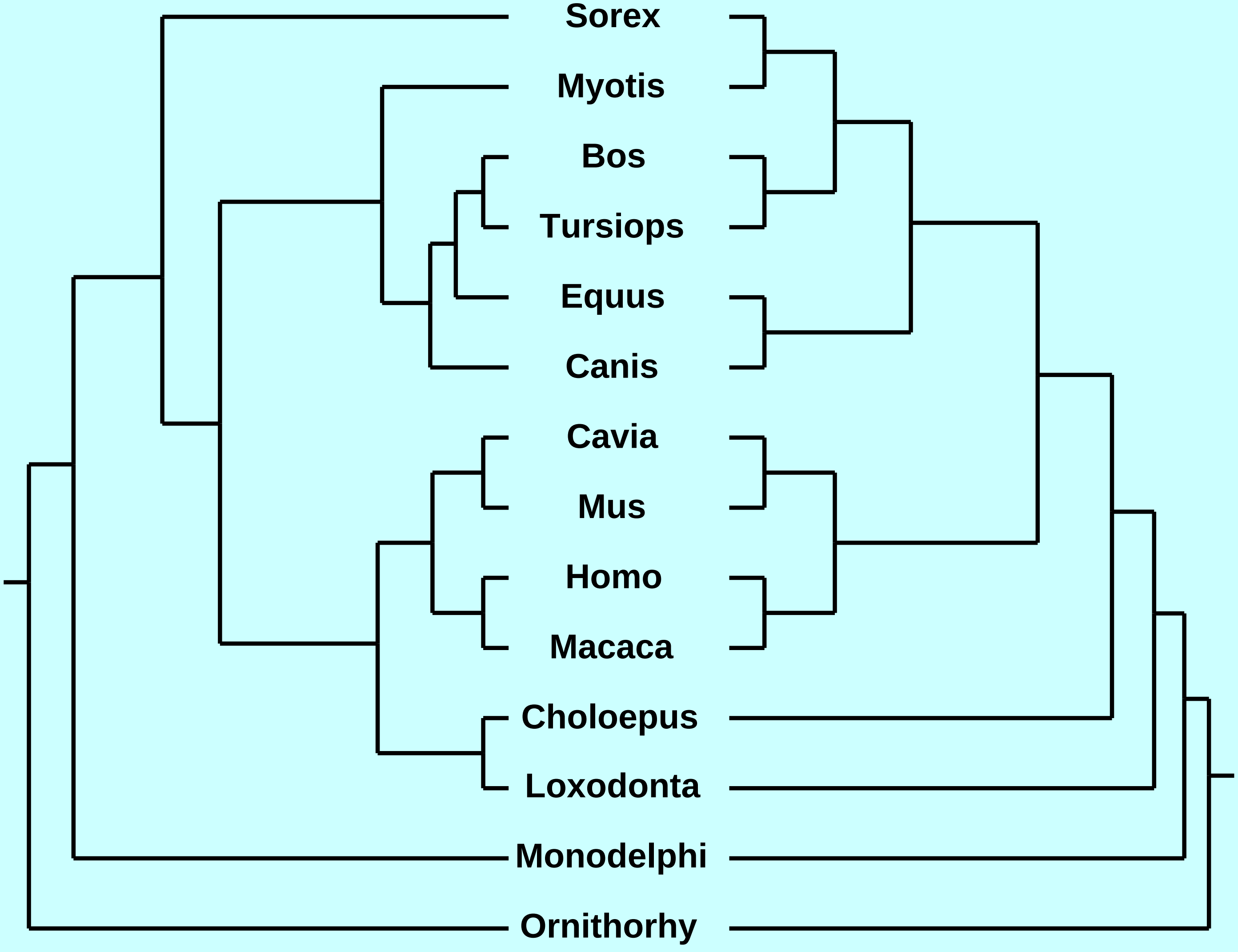
Using 19 loci, the consensus tree is



Using 19 loci concatenated, the tree is



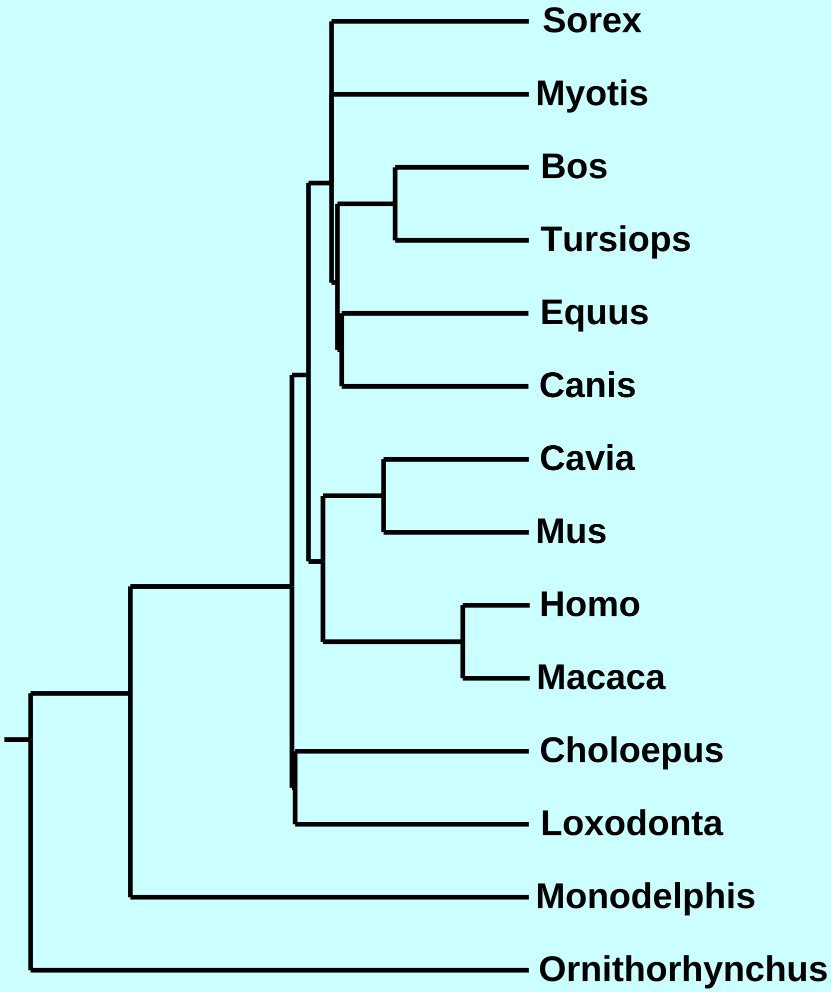
Do these trees agree with each other? Well, ...



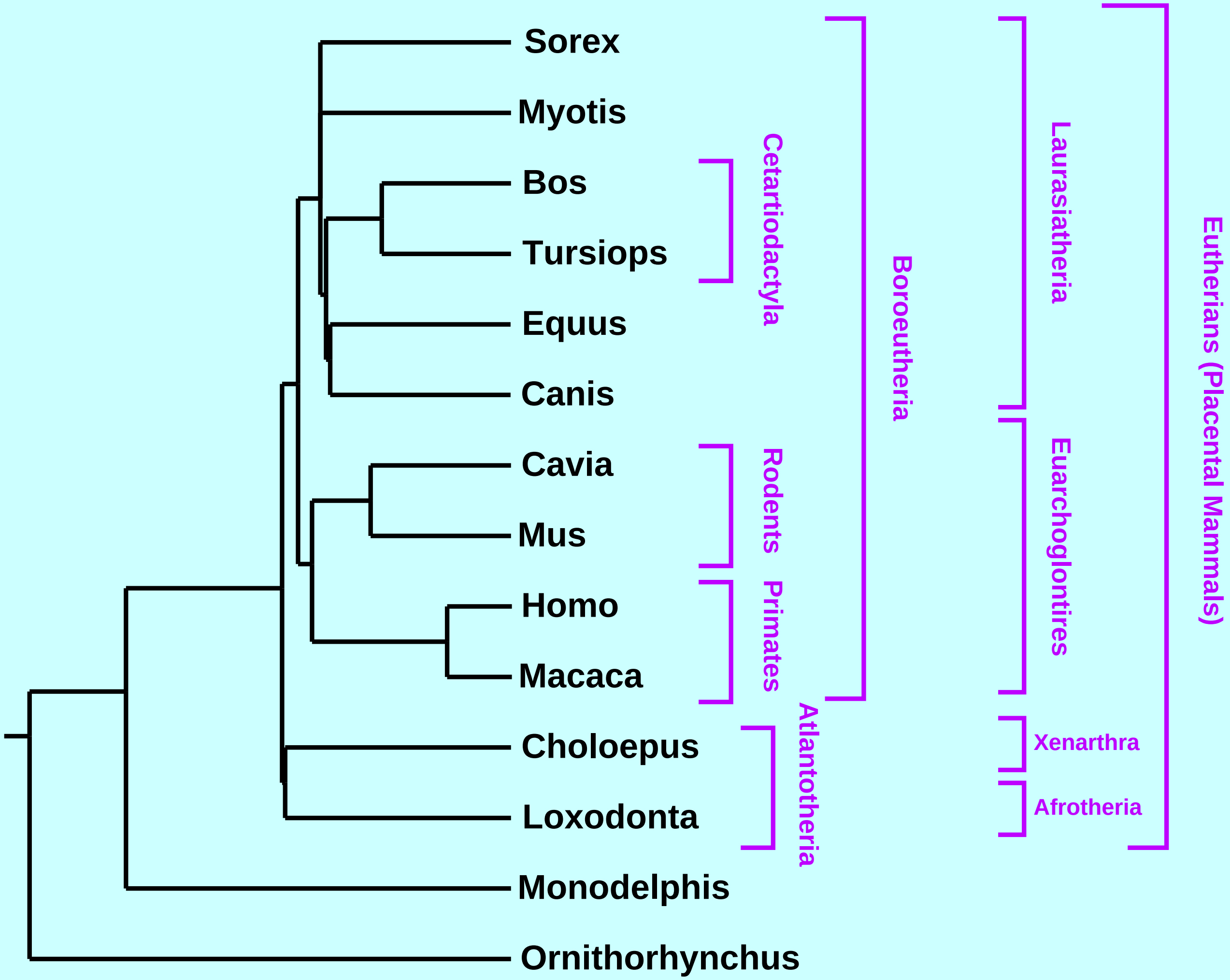
Consensus tree

Concatated tree

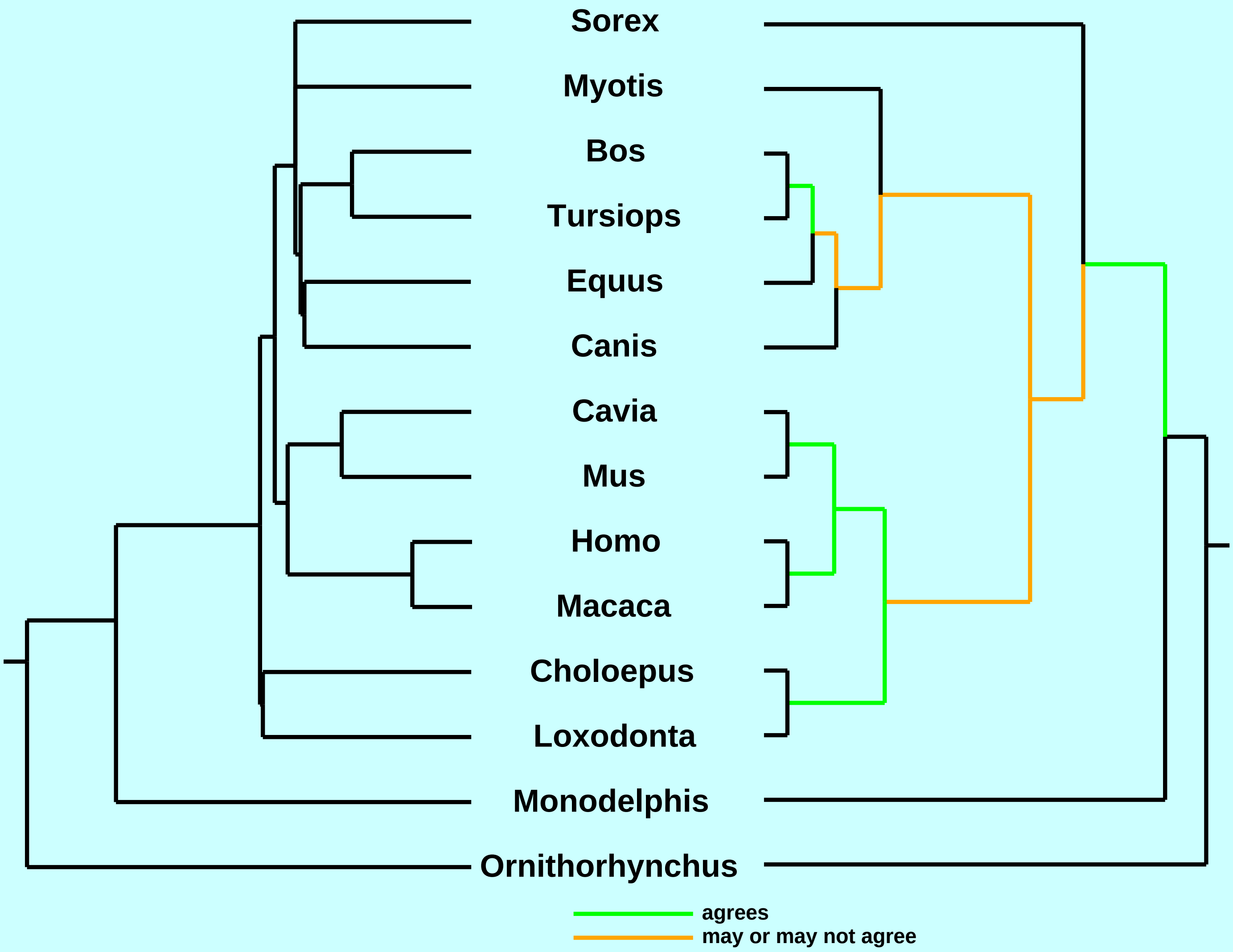
The “expert tree” from Timetree.org



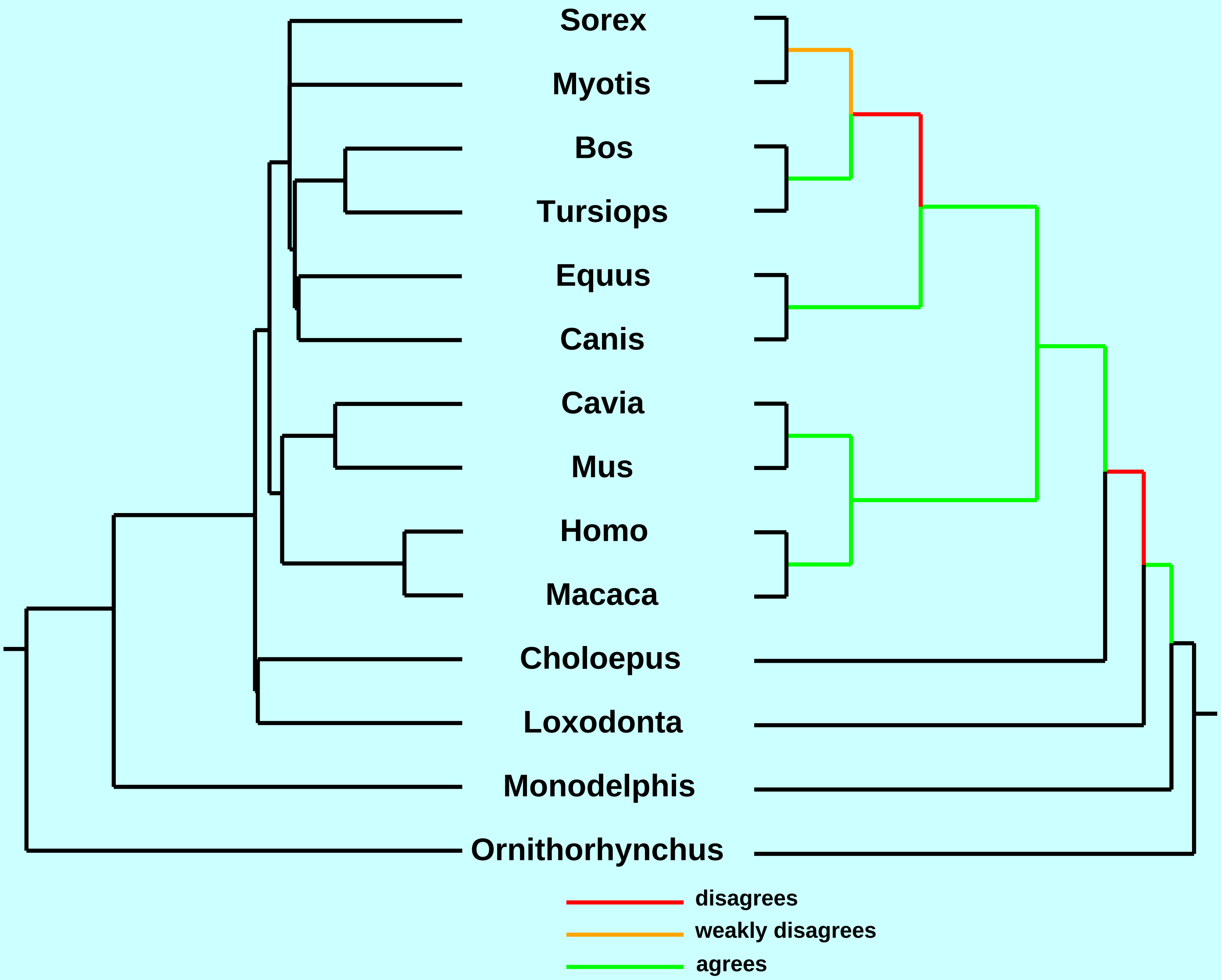
Some named groups widely agreed upon



How does the consensus tree agree?



How does the concatenated tree agree?



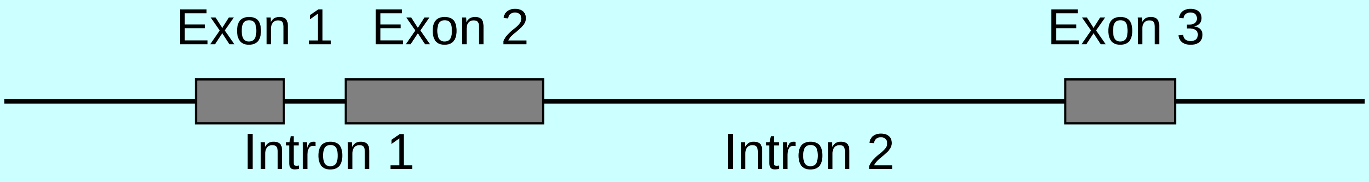
Molecular phylogenies

Some examples of other important conclusions from molecular phylogenies:

- Using immunological distances, Morris Goodman (1962 on) and later Wilson and Sarich (1966) show that humans, gorilla, and chimps were a clade.
- Wilson and Sarich (in that work, 1967) date the divergence of humans to 5 million years.
- Charles Sibley and Jon Ahlquist (1984) use DNA hybridization to argue for the clade humans-chimps.
- Carl Woese (1978) uses rRNA trees to introduce evolution into microbiology, argue for the domain Archaea.
- Much progress on early radiation of angiosperms
- Protostome-deuterostome tree of metazoans (more or less) replaced by deuterostome-lophotrichozoa-ecdysozoa tree.
- Fungi closer to animals than either is to plants.
- Symbiotic origin of mitochondria and of chloroplasts verified.
- Amphioxus diverged before split of tunicates from craniate chordates.
- Lots of horizontal gene transfer in prokaryotes, almost not a tree.

Different parts of hemoglobin genes

Alignment of hemoglobin ε loci of Human, Tarsier



region	bases	differences	% different
upstream	**100**	**12**	**12.0**
exon 1	**92**	**9**	**9.8**
intron 1	**126**	**26**	**20.6**
exon2	**223**	**26**	**11.7**
intron 2	**820**	**239**	**29.1**
exon 3	**129**	**13**	**10.1**
downstream	**100**	**13**	**13.0**

Differences in exons

position 1	**10**
position 2	**5**
position 3	**33**

Higher Rates of Substitution for ...

1. ... some proteins than others
2. ... some sites within proteins than others (less in active sites, interior sites)
3. ... some amino acid replacements than others (less changes to chemically more similar amino acids)
4. ... silent changes than nonsilent ones
5. ...“in-between" DNA than introns, introns than coding sequences
6. ... transitions than transversions

Morris Goodman tabulation for β hemoglobin

where	sites	change/100myr
Heme contacts	21	0.02
Nonheme contacts	10	0.02
Salt bridges β-β	4	0.00
2,3-DPG binding	4	0.10
Nonsalt bridge α,β contacts	16	0.16
Remaining interior	21	0.09
Remaining exterior	70	0.20
All	146	0.13

PhyloHMM analysis of multiple genomes

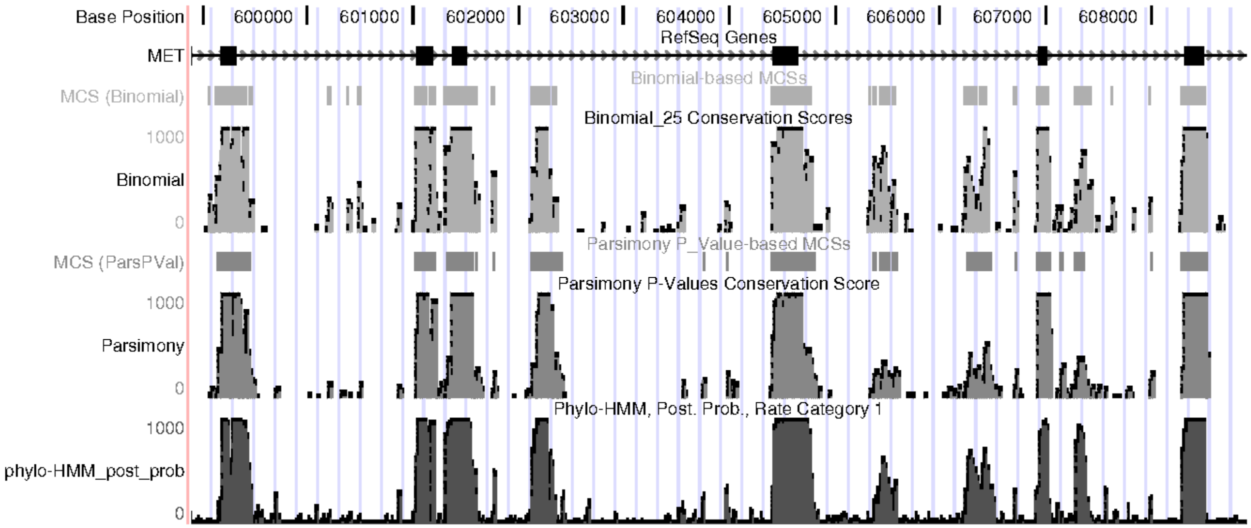


Fig. 5. A screen shot from the UCSC Genome Browser [24] showing a selected region of the data set of example 2, including several exons of the *MET* gene (black boxes at top). The binomial-based (light gray) and parsimony-based (medium gray) conservation scores of Margulies et al. [30] are shown as tracks in the browser, as are the posterior probabilities (×1000) of state s_1 in the phylo-HMM (dark gray). Plots similar to this one, showing phylo-HMM-based conservation scores across the whole human genome, can be viewed online at http://genome.ucsc.edu.

From a paper by Siepel and Haussler (*Journal of Computational Biology*, 2004) describing the machinery for finding conserved regions of multiple genomes.

Rates of change from neutral and selective mechanisms

Neutral mutations

A fraction μ of all copies of a gene mutate. Of these $\frac{1}{2\mathrm{N}}$ (equal to the initial frequency of the mutant) succeed in drifting to fixation for the mutant.

There are in all $2\mathrm{N}$ copies of the gene available to mutate.

The resulting rate of substitution is

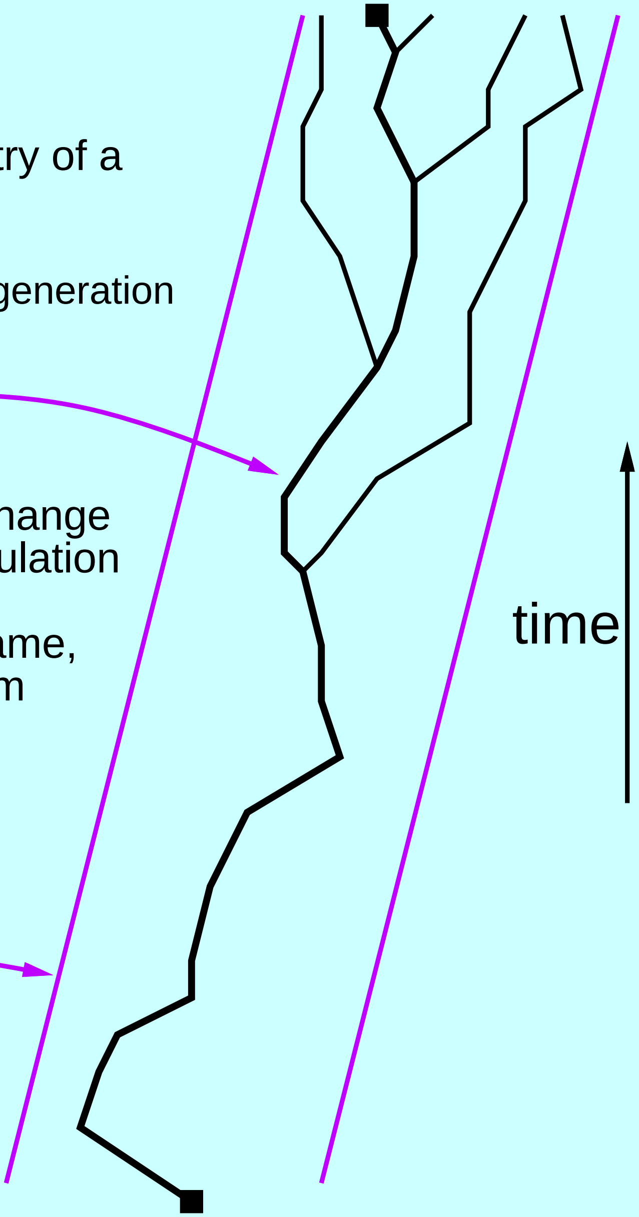
$$\mu \times \frac{1}{2\mathrm{N}} \times 2\mathrm{N} = \mu$$

So the rate of substitution of neutral mutations is equal to the mutation rate (the mutation rate of neutral mutants, not the total mutation rate).

Change by neutral mutation

end of a lineage which is the ancestry of a single copy of the gene:

change is expected to be μ per generation



gene copy ancestry

Interestingly, the expected rate of change is the same no matter what the population size is (small populations make all copies more likely to become the same, but all are still expected to differ from their ancestors by this amount)

species boundary

Rates from neutral and selective mechanisms

Selectively advantageous mutations

A fraction μ of copies of the gene mutate. There are in all $2N$ copies available. A fraction $2s$ succeed in fixing.

The resulting rate of substitution is

$$\mu \times 2N \times 2s = 4Ns\mu$$

Note that this is $4Ns$ times as high as for neutral mutants, if the mutation rate in both categories were equal (which it isn't).

Substitution of neutral and advantageous mutations

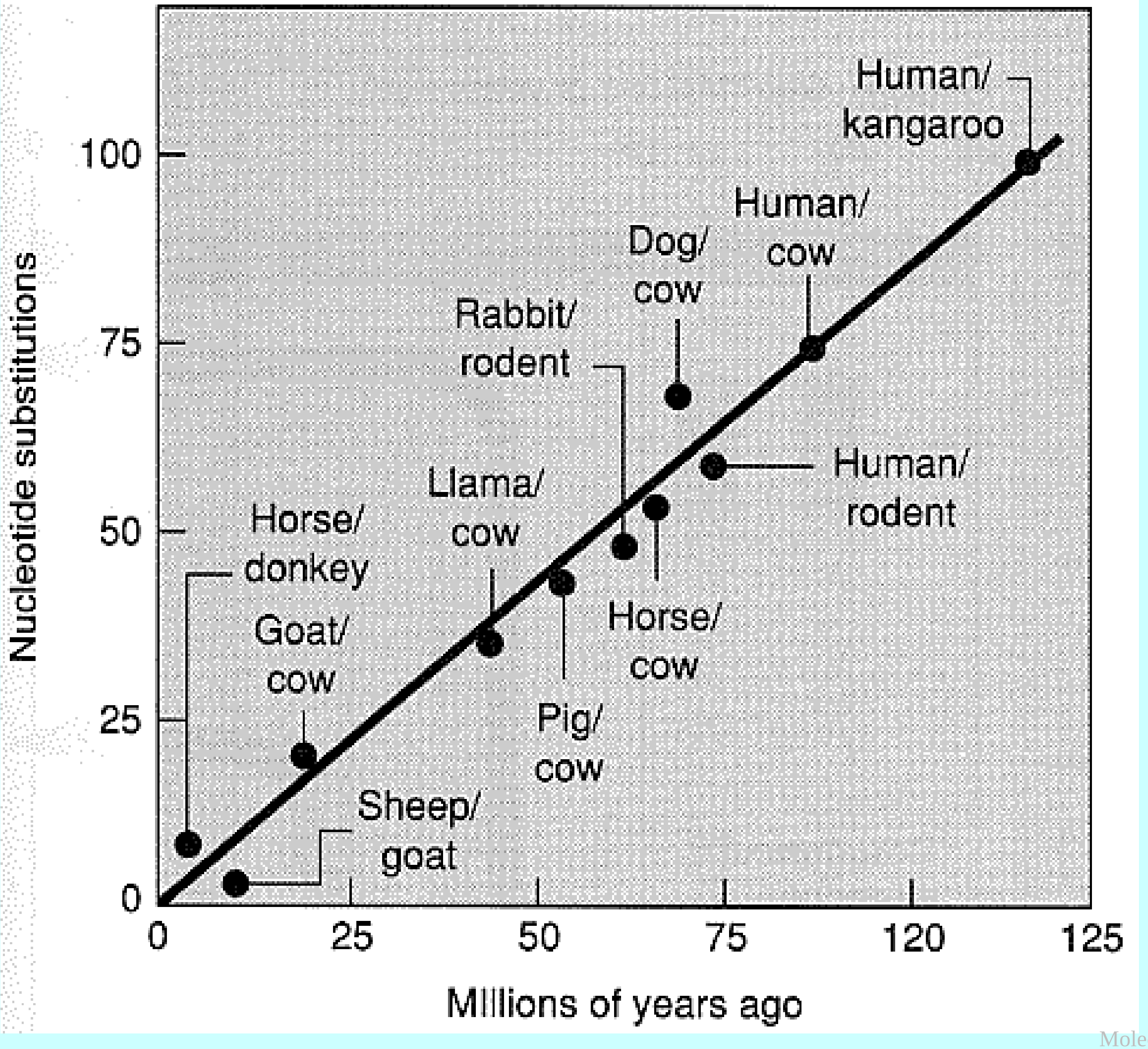
Suppose that the population size is $N = 1,000,000$, and mutation rates are:

$$\begin{array}{lrcl} \text{Advantageous mutations} & u_a & = & 10^{-7} \\ \text{Neutral mutations} & u_n & = & 10^{-6} \end{array}$$

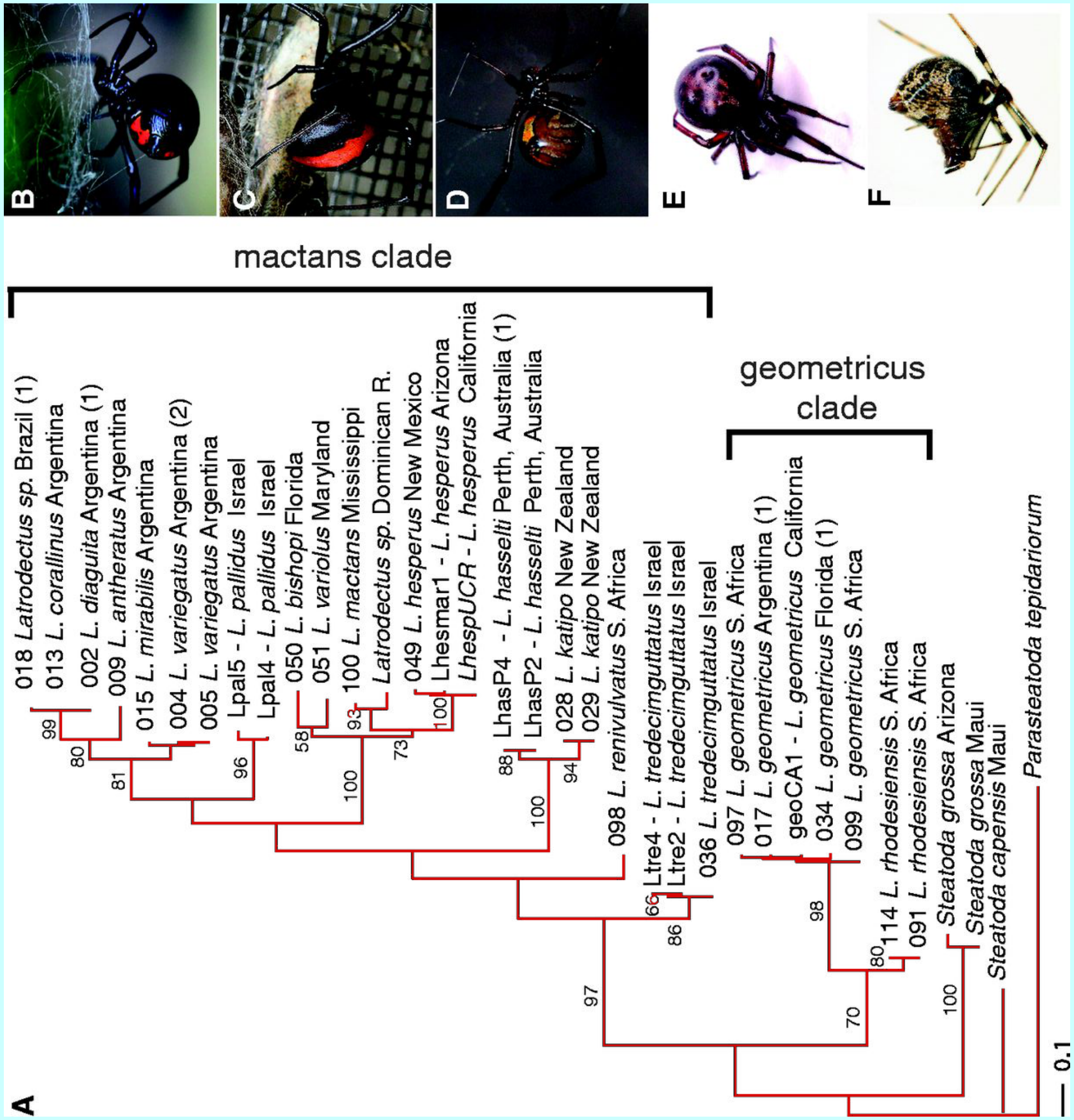
If the selection coefficient in favor of advantageous mutations is $s \;=\; 0.0001$, the rates of substitution expected are:

$$\begin{array}{lrcl} \text{Advantageous mutations} & (4Ns)u_a & = & 4 \times 10^{-5} \\ \text{Neutral mutations} & u_n & = & 10^{-6} \end{array}$$

The molecular clock (from Wilson, 1976)



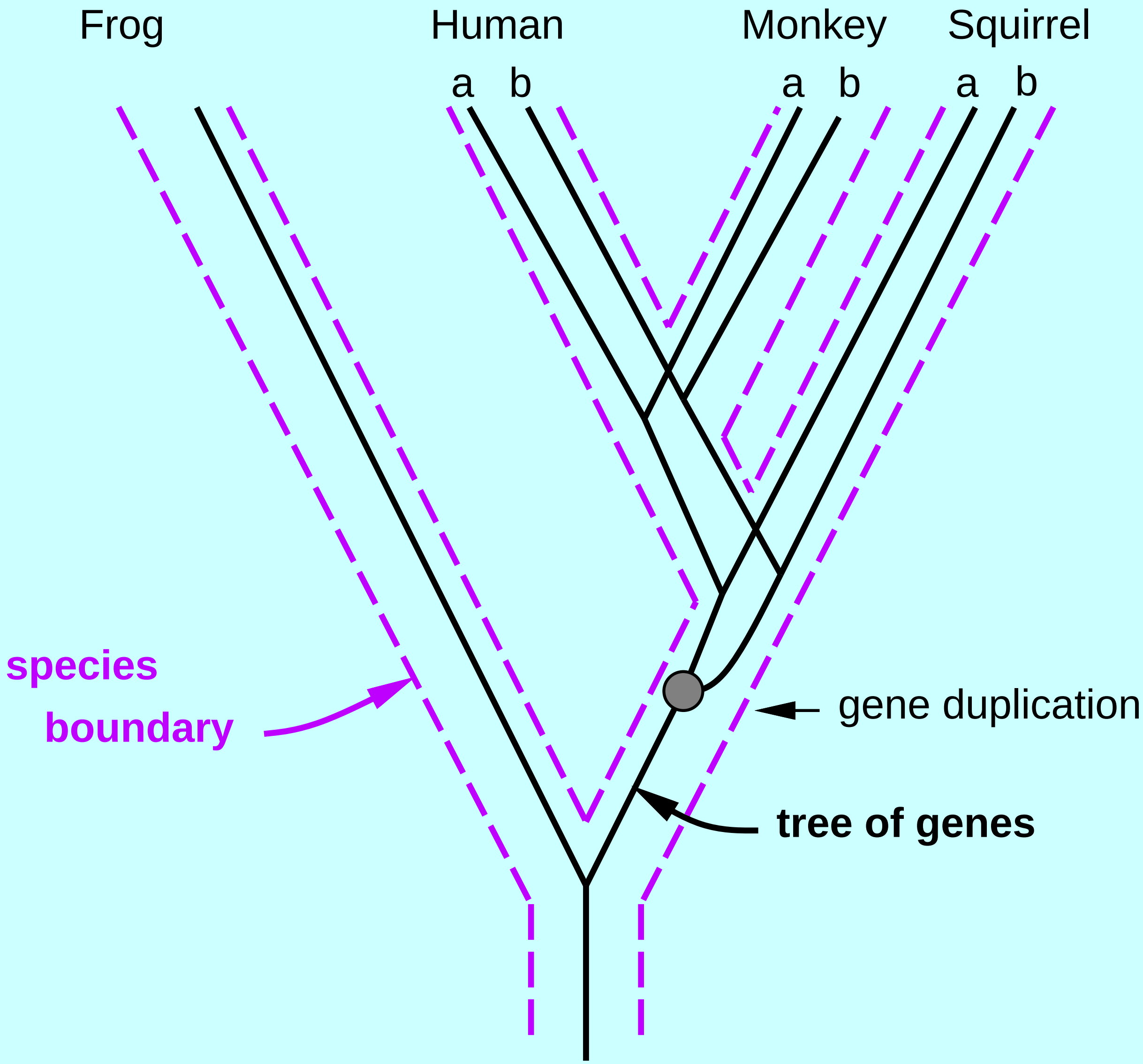
A not-quite-clocklike tree of black widow spiders


B
C
D
E
F
mactans clade
geometricus clade
018 *Latrodectus sp.* Brazil (1)
013 *L. corallinus* Argentina
002 *L. diaguita* Argentina (1)
009 *L. antheratus* Argentina
015 *L. mirabilis* Argentina
004 *L. variegatus* Argentina (2)
005 *L. variegatus* Argentina
Lpal5 - *L. pallidus* Israel
Lpal4 - *L. pallidus* Israel
050 *L. bishopi* Florida
051 *L. variolus* Maryland
100 *L. mactans* Mississippi
Latrodectus sp. Dominican R.
049 *L. hesperus* New Mexico
Lhesmar1 - *L. hesperus* Arizona
LhespUCR - *L. hesperus* California
LhasP4 - *L. hasselti* Perth, Australia (1)
LhasP2 - *L. hasselti* Perth, Australia
028 *L. katipo* New Zealand
029 *L. katipo* New Zealand
098 *L. renivulvatus* S. Africa
Ltre4 - *L. tredecimguttatus* Israel
Ltre2 - *L. tredecimguttatus* Israel
036 *L. tredecimguttatus* Israel
097 *L. geometricus* S. Africa
017 *L. geometricus* Argentina (1)
geoCA1 - *L. geometricus* California
034 *L. geometricus* Florida (1)
099 *L. geometricus* S. Africa
114 *L. rhodesiensis* S. Africa
091 *L. rhodesiensis* S. Africa
Steatoda grossa Arizona
Steatoda grossa Maui
Steatoda capensis Maui
Parasteatoda tepidariorum
99
80
81
96
58
93
100
73
100
88
94
100
66
86
97
98
80
70
100
A
0.1

Tree for α-latrotoxin protein from J. E. Garb and C. Y. Hayashi, 2013, in *Molecular Biology and Evolution*, vol. 30, issue 5, pp. 999-1004.

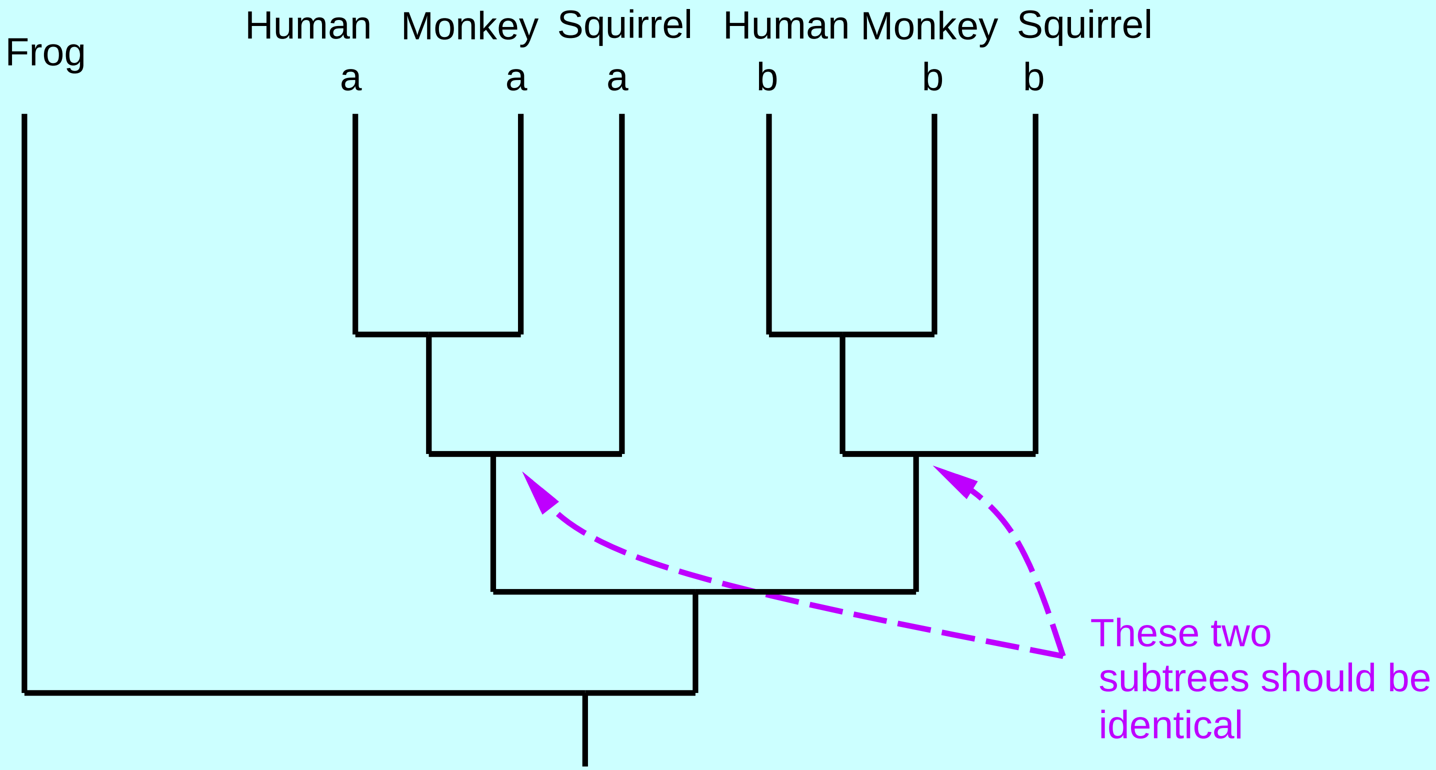
Phylogeny and gene tree with a gene duplication

A phylogeny with a gene duplication event:



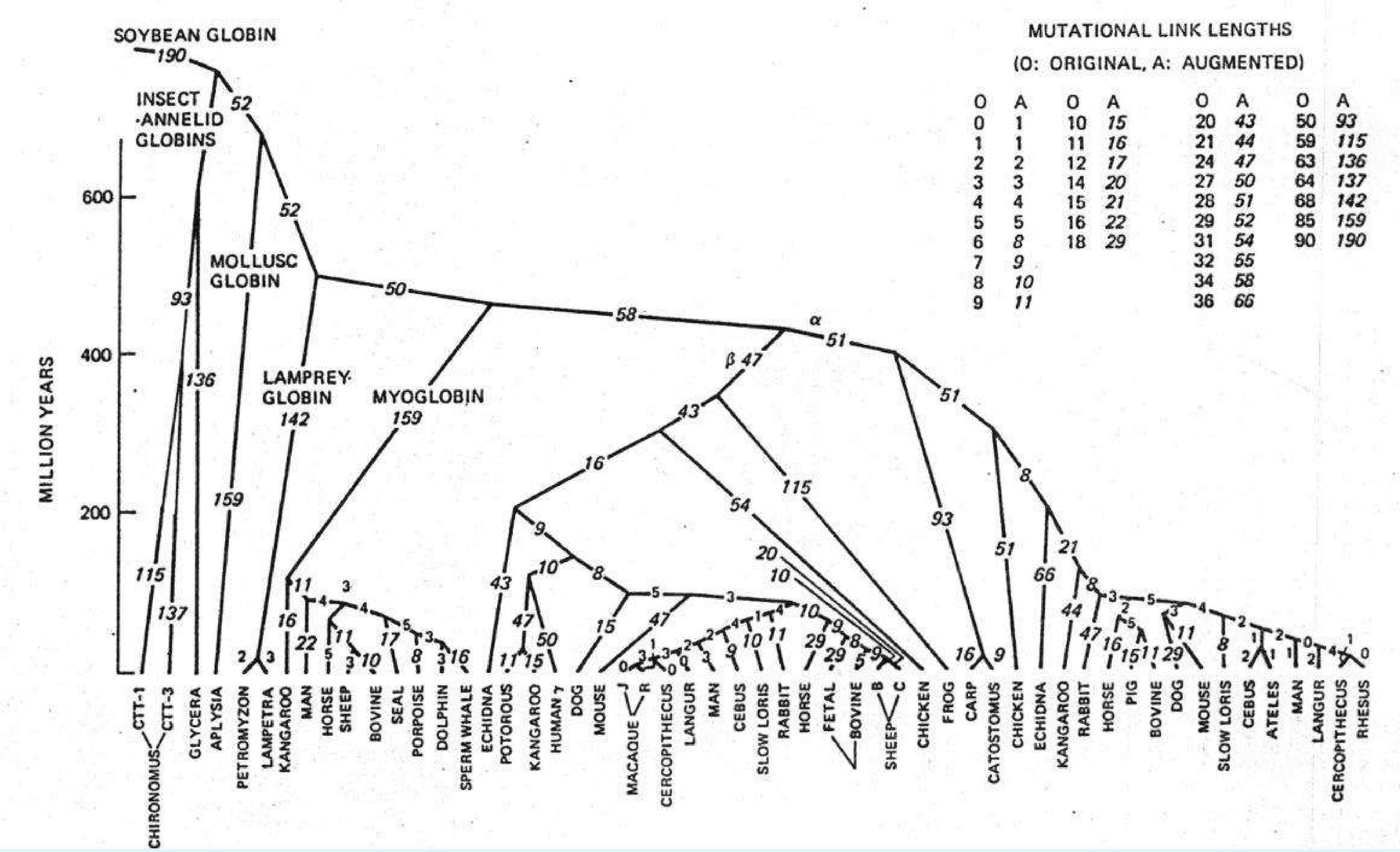
Phylogenies, gene trees, and gene duplication

So when genes are all aligned with each other, their "gene tree" is:



These two subtrees should be identical

A parsimony tree for the globin gene family



MUTATIONAL LINK LENGTHS

(O: ORIGINAL, A: AUGMENTED)

O	A	O	A	O	A	O	A
0	1	10	*15*	20	*43*	50	*93*
1	1	11	*16*	21	*44*	59	*115*
2	2	12	*17*	24	*47*	63	*136*
3	3	14	*20*	27	*50*	64	*137*
4	4	15	*21*	28	*51*	68	*142*
5	5	16	*22*	29	*52*	85	*159*
6	*8*	18	*29*	31	*54*	90	*190*
7	*9*			32	*55*		
8	*10*			34	*58*		
9	*11*			36	*66*		