Introduction to Phylogeny

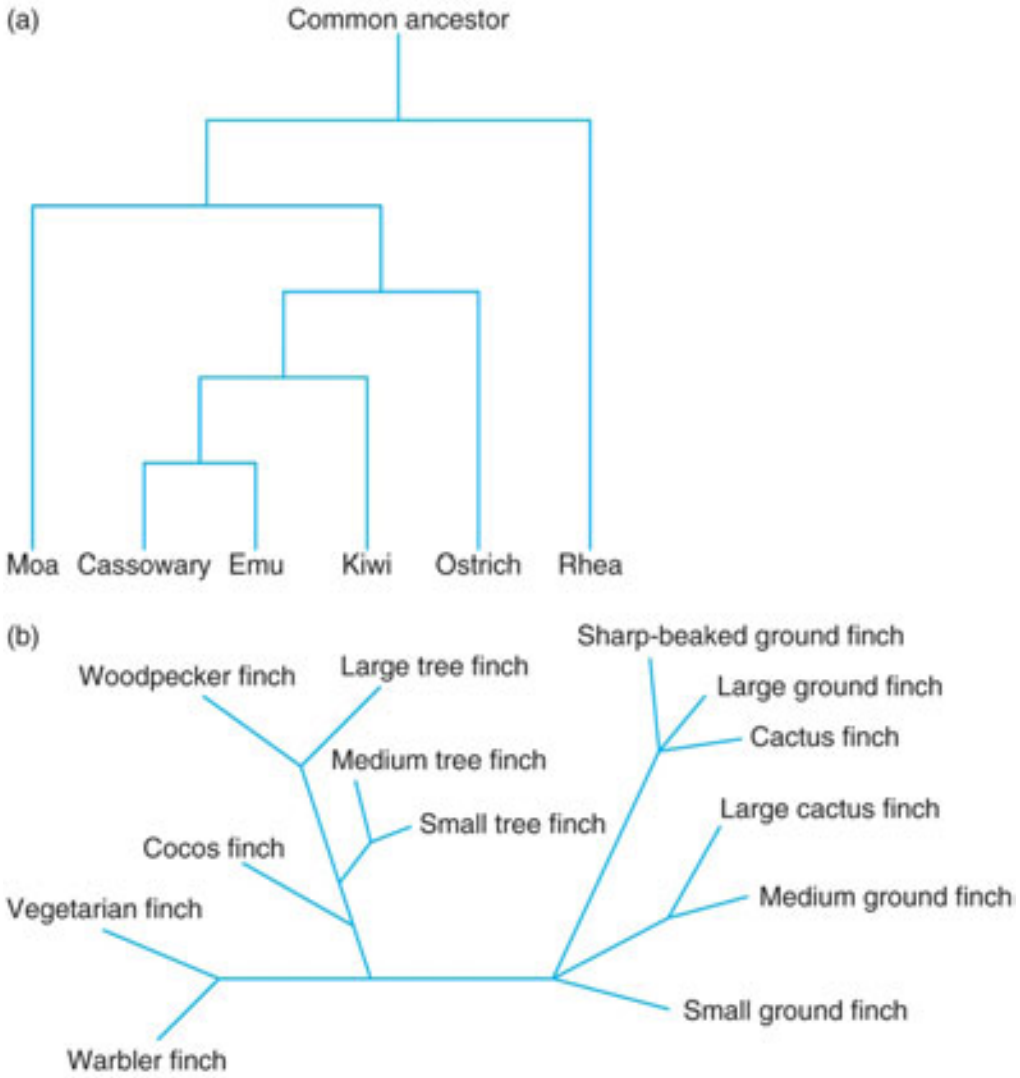
Arne Elofsson

Reading material:
“The Roots of Bioinformatics in Protein Evolution” Ross Doolittle
http://www.ploscollections.org/article/info%3Adoi%2F10.1371%2Fjournal.pcbi.1000875
http://en.wikipedia.org/wiki/Phylogeny
http://en.wikipedia.org/wiki/Neighbor_joining
https://en.wikipedia.org/wiki/Maximum_parsimony_(phylogenetics)

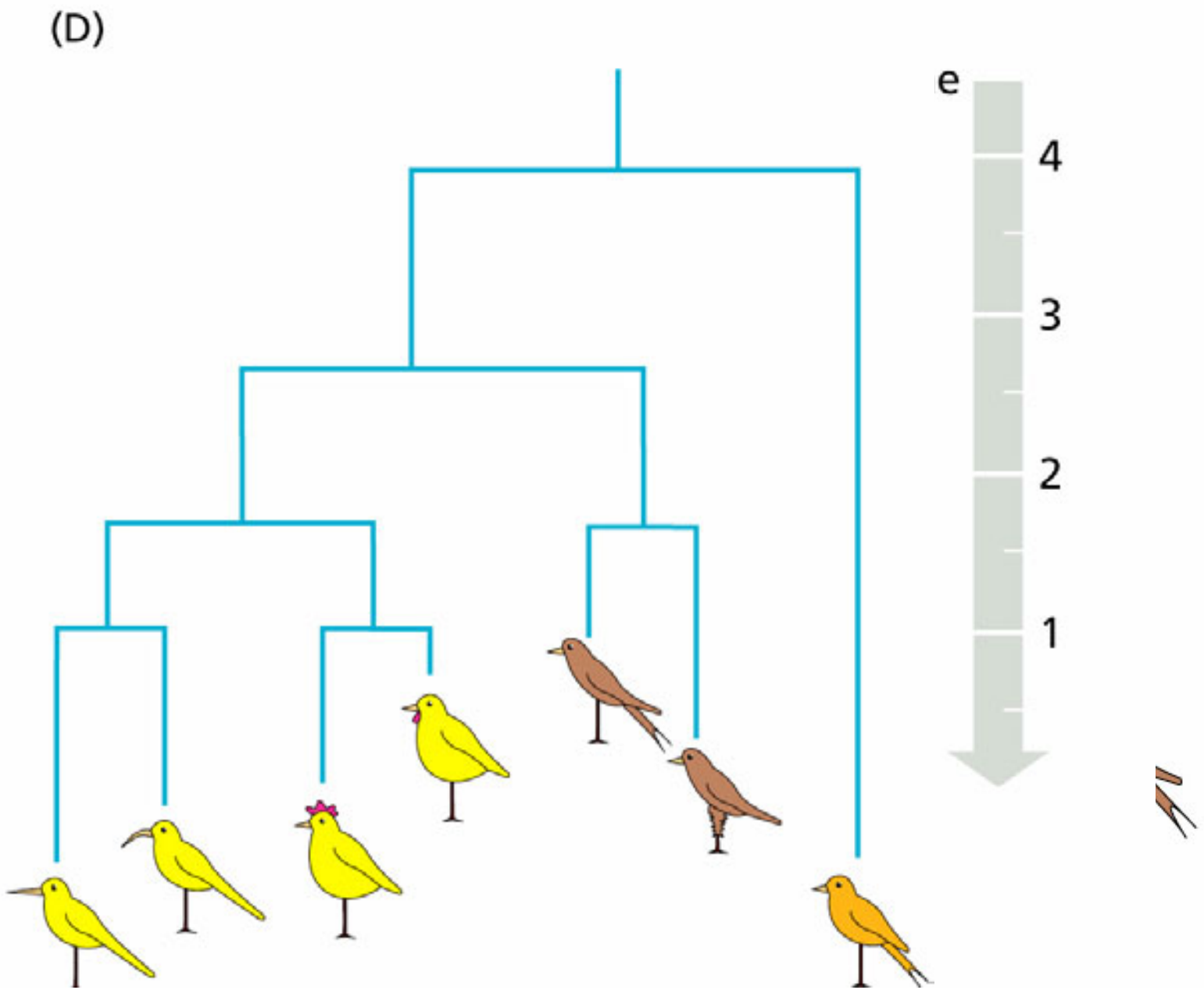
What is Phylogeny

- Evolutionary tree
- The (true) relationship of a group (greek: Phyla) through descent over time from their common ancestor

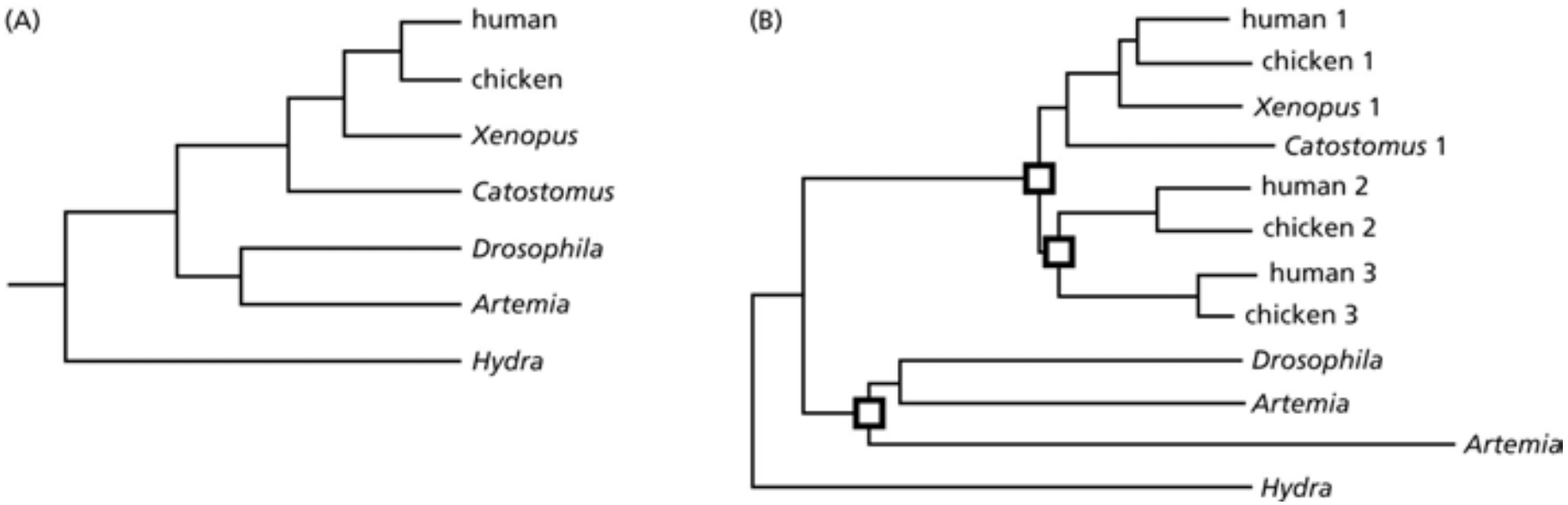
Trees


(a)
Common ancestor
Moa
Cassowary
Emu
Kiwi
Ostrich
Rhea
(b)
Woodpecker finch
Large tree finch
Medium tree finch
Small tree finch
Cocos finch
Vegetarian finch
Warbler finch
Sharp-beaked ground finch
Large ground finch
Cactus finch
Large cactus finch
Medium ground finch
Small ground finch

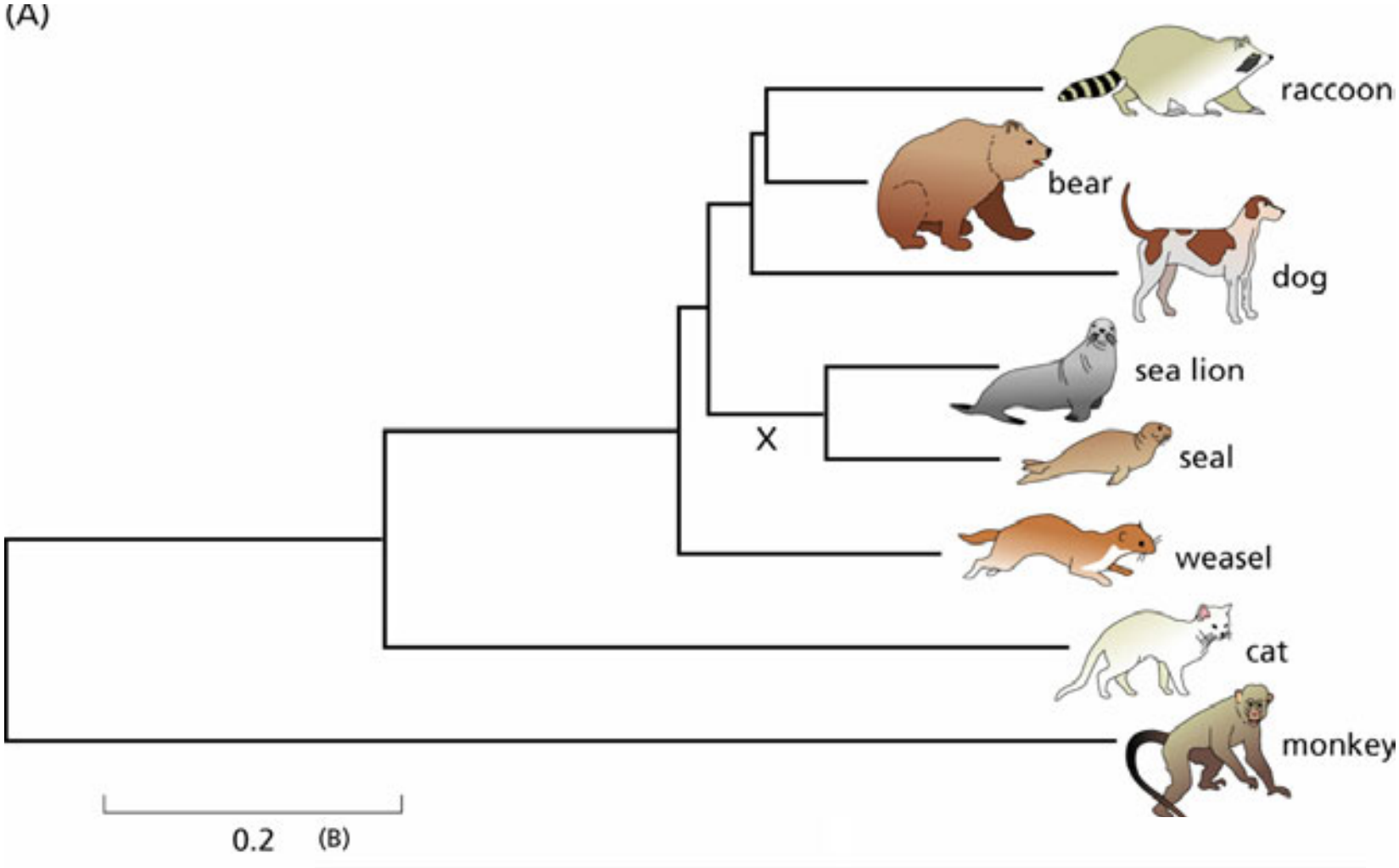
Different type of trees



Species trees and gene trees

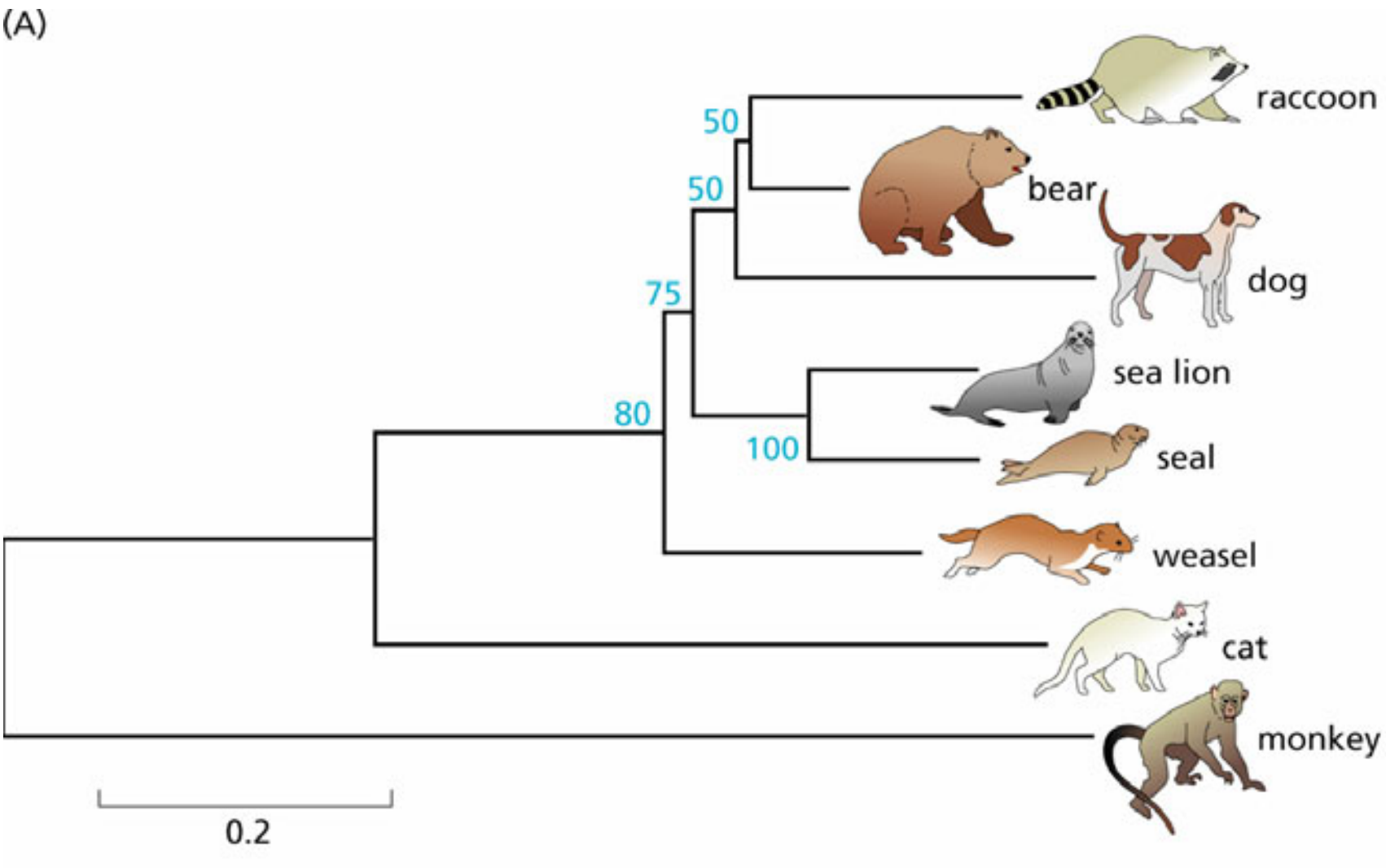


A tree can be represented as splits

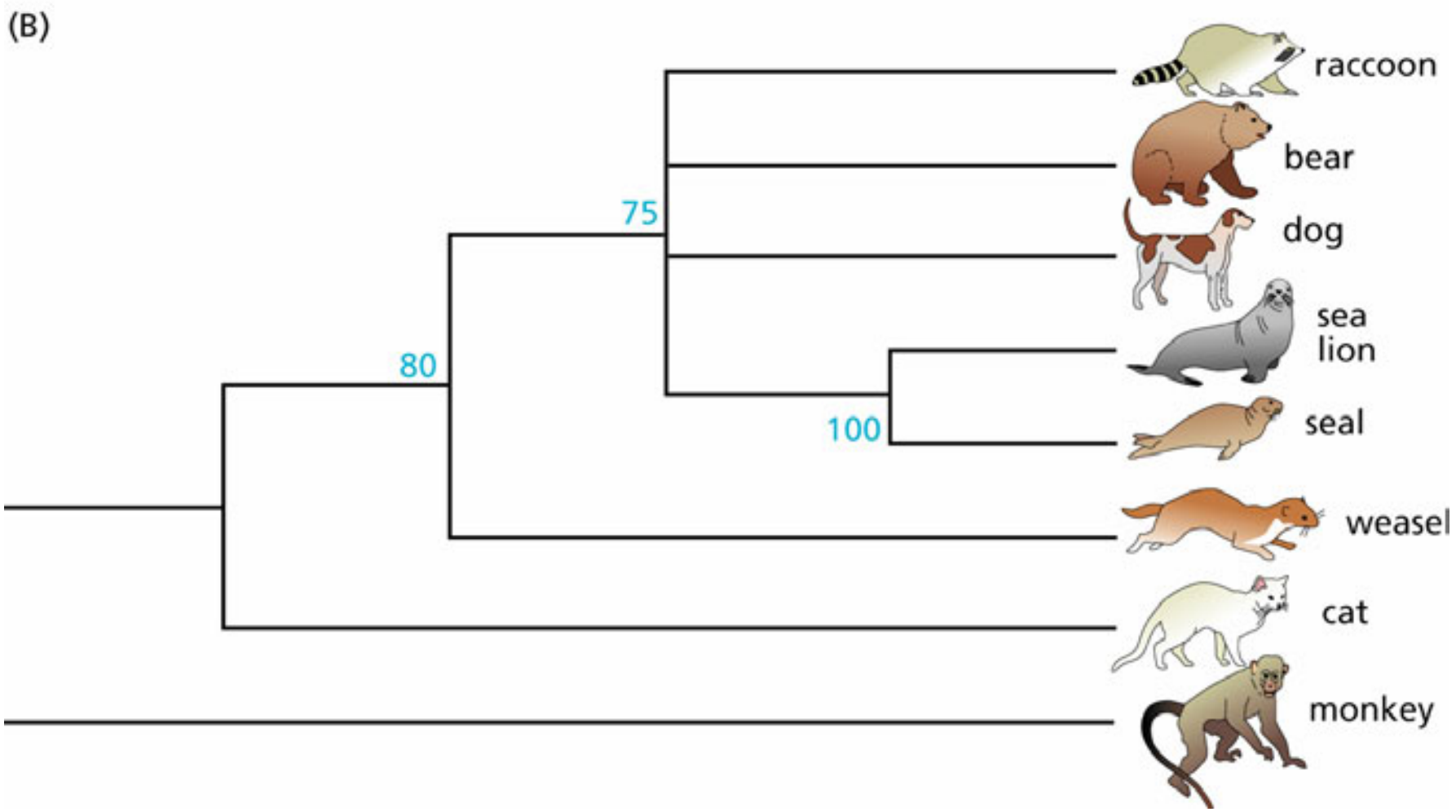


raccoon	bear	dog	sea lion	seal	weasel	cat	monkey
*	*						
*	*	*					
			*	*			
*	*	*	*	*			
*	*	*	*	*	*		

Bootstrapping provides confidence



Condensed tree



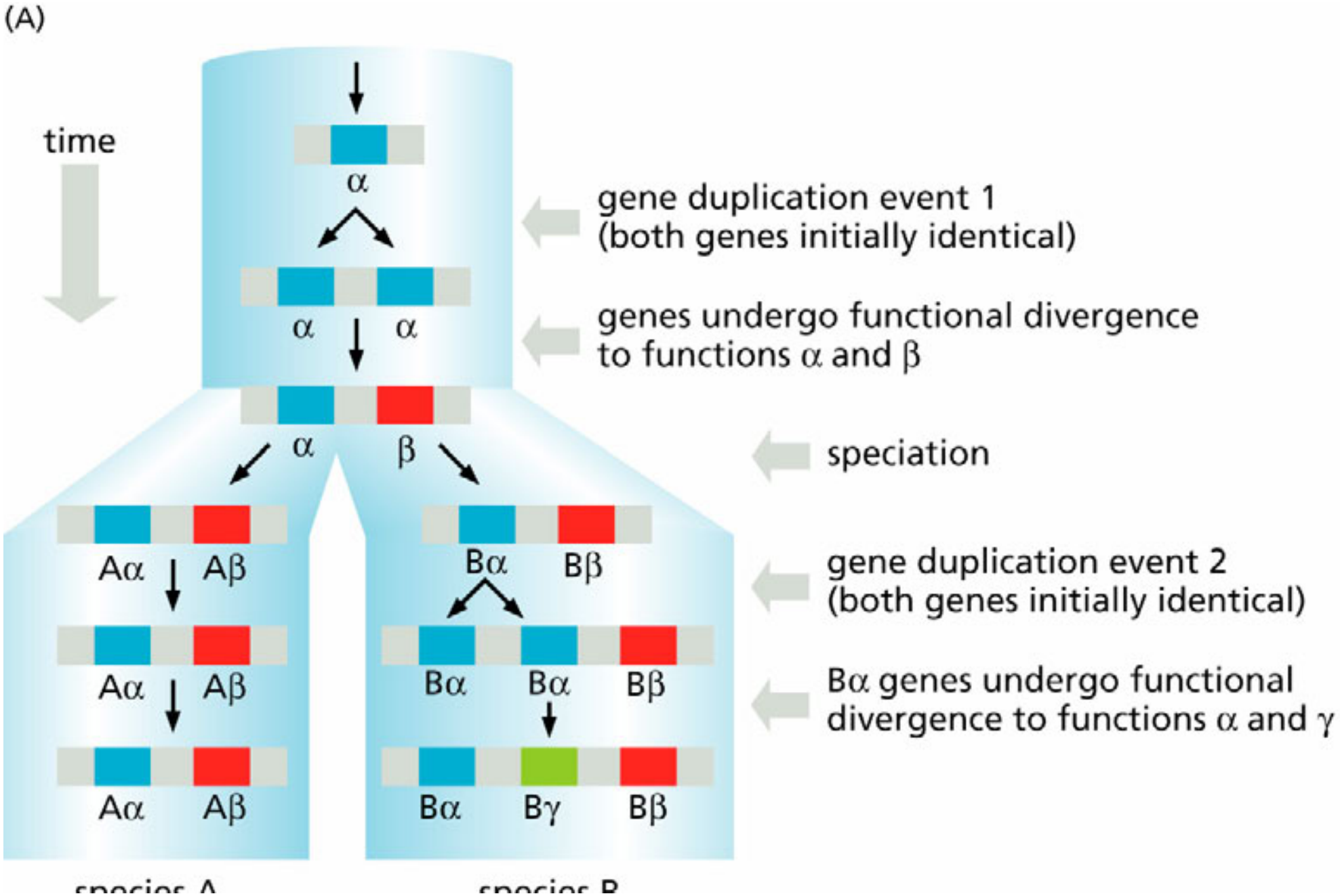
Orthologs vs. Paralogs

- Two types of homologs
- Orthologs - genes separated by a speciation event
  - More similar function
- Paralogs - genes separated by a gene duplication events
  - Duplication allows new function to evolve
  - Sub-functionalisation

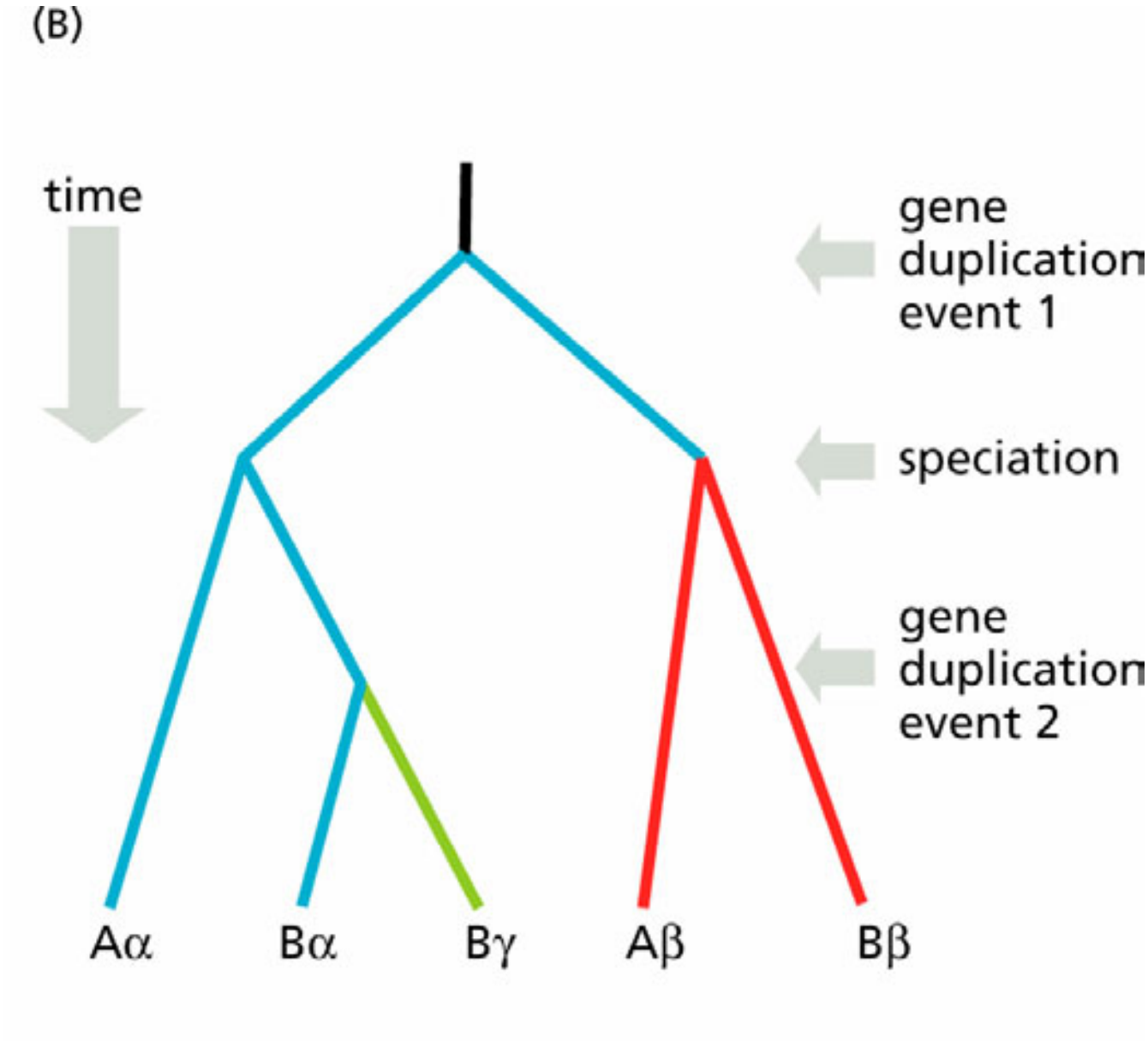
Biological definitions for related sequences

- **Homologues** are similar sequences in two different organisms that have been derived from a common ancestor sequence. Homologues can be described as either orthologues or paralogues.
- **Orthologues** are similar sequences in two different organisms that have arisen due to a speciation event. Orthologs typically retain identical or similar functionality throughout evolution.
- **Paralogues** are similar sequences within a single organism that have arisen due to a gene duplication event.
- **Xenologues** are similar sequences that do not share the same evolutionary origin, but rather have arisen out of horizontal transfer events through symbiosis, viruses, etc.

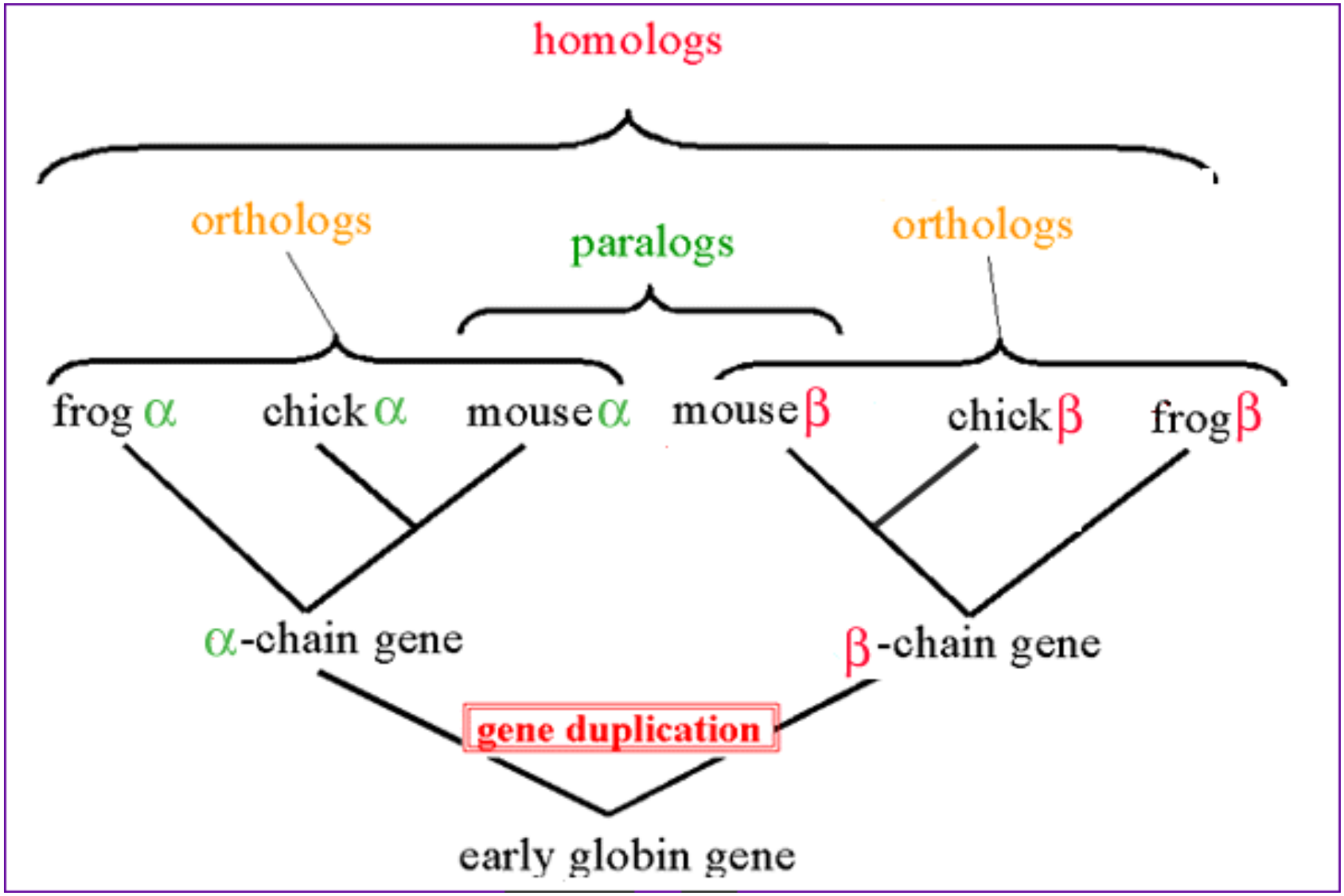
Gene evolution



The phylogenetic tree

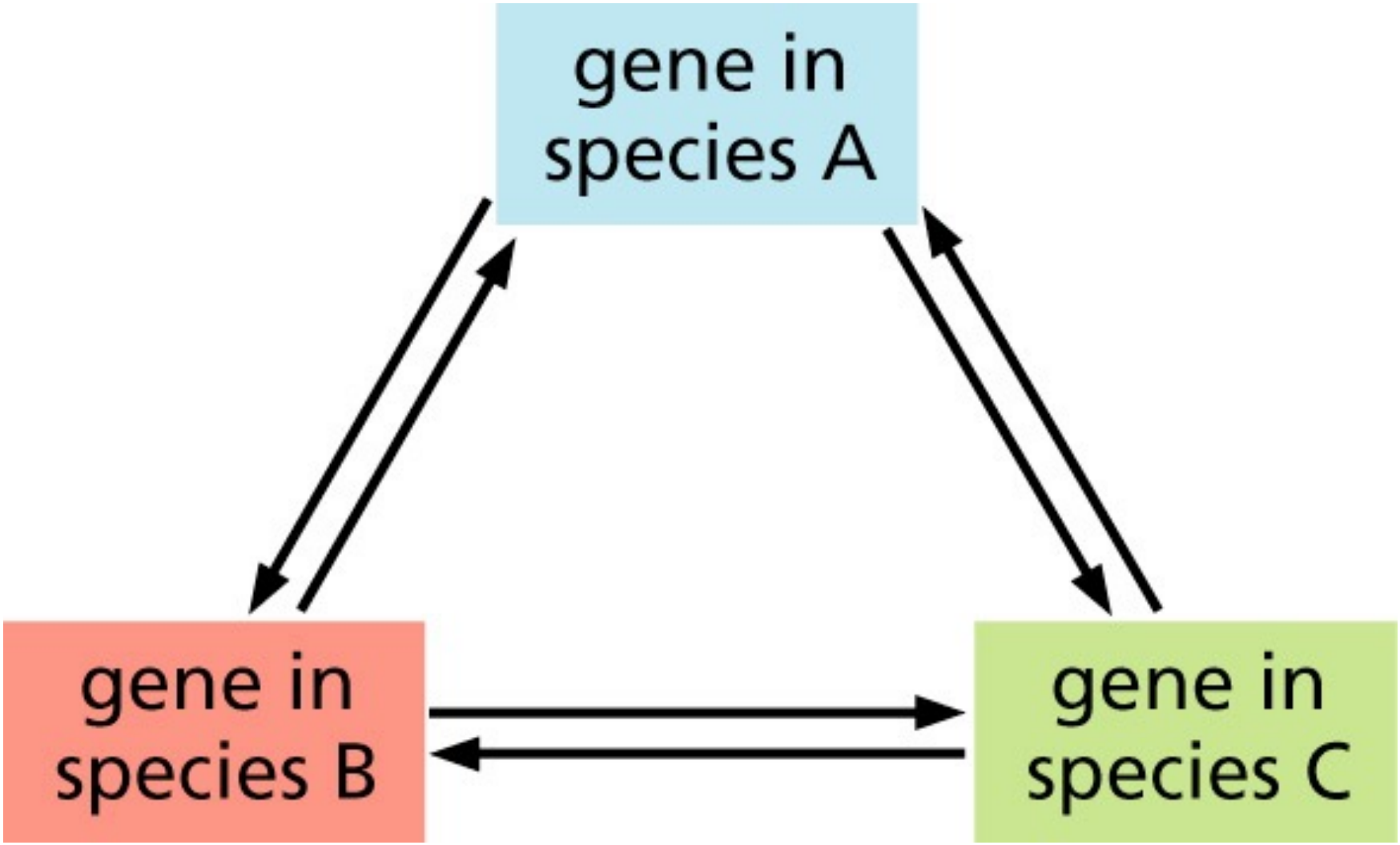


So this means …

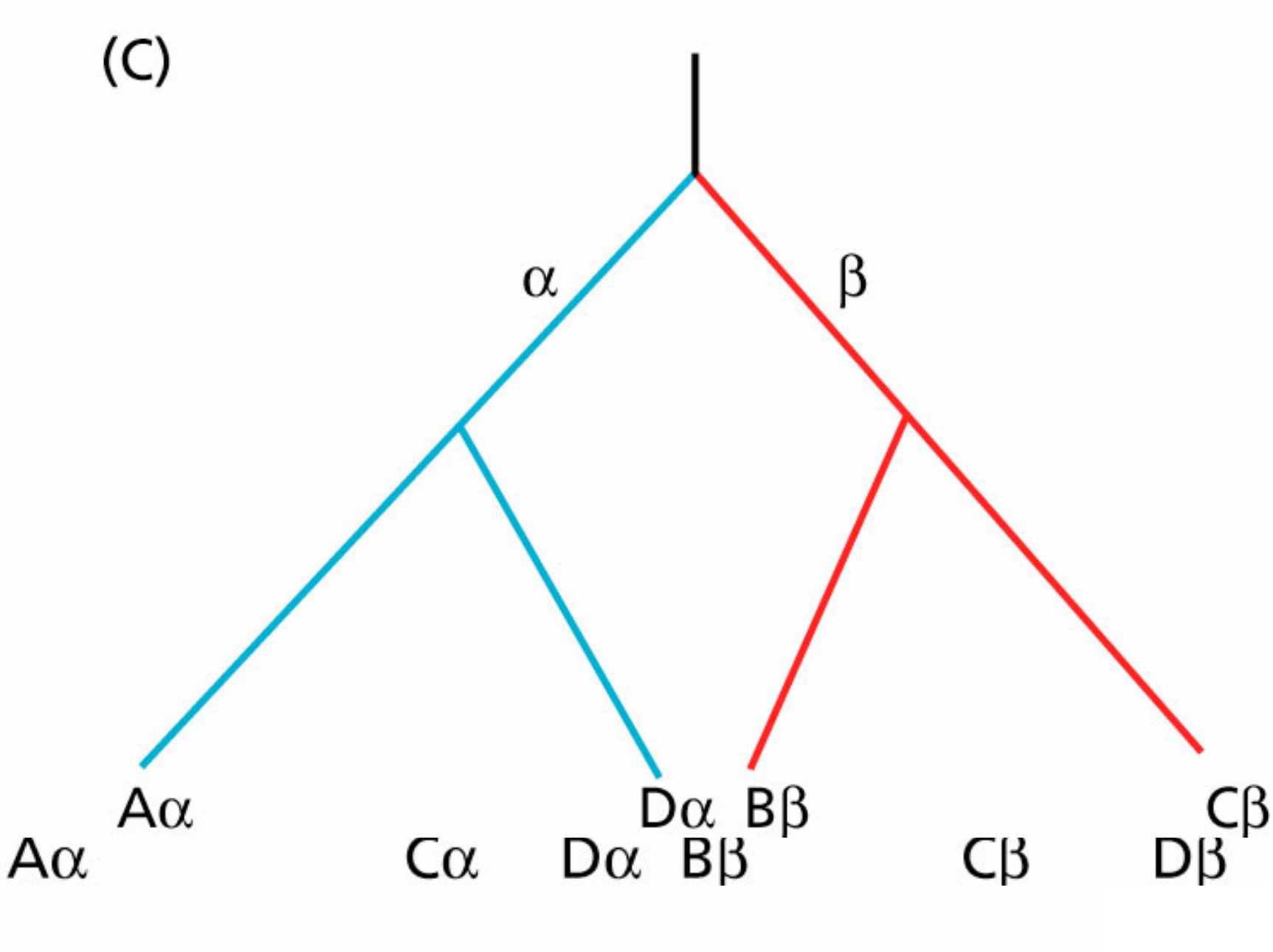


Source: http://www.ncbi.nlm.nih.gov/Education/BLASTinfo/Orthology.html

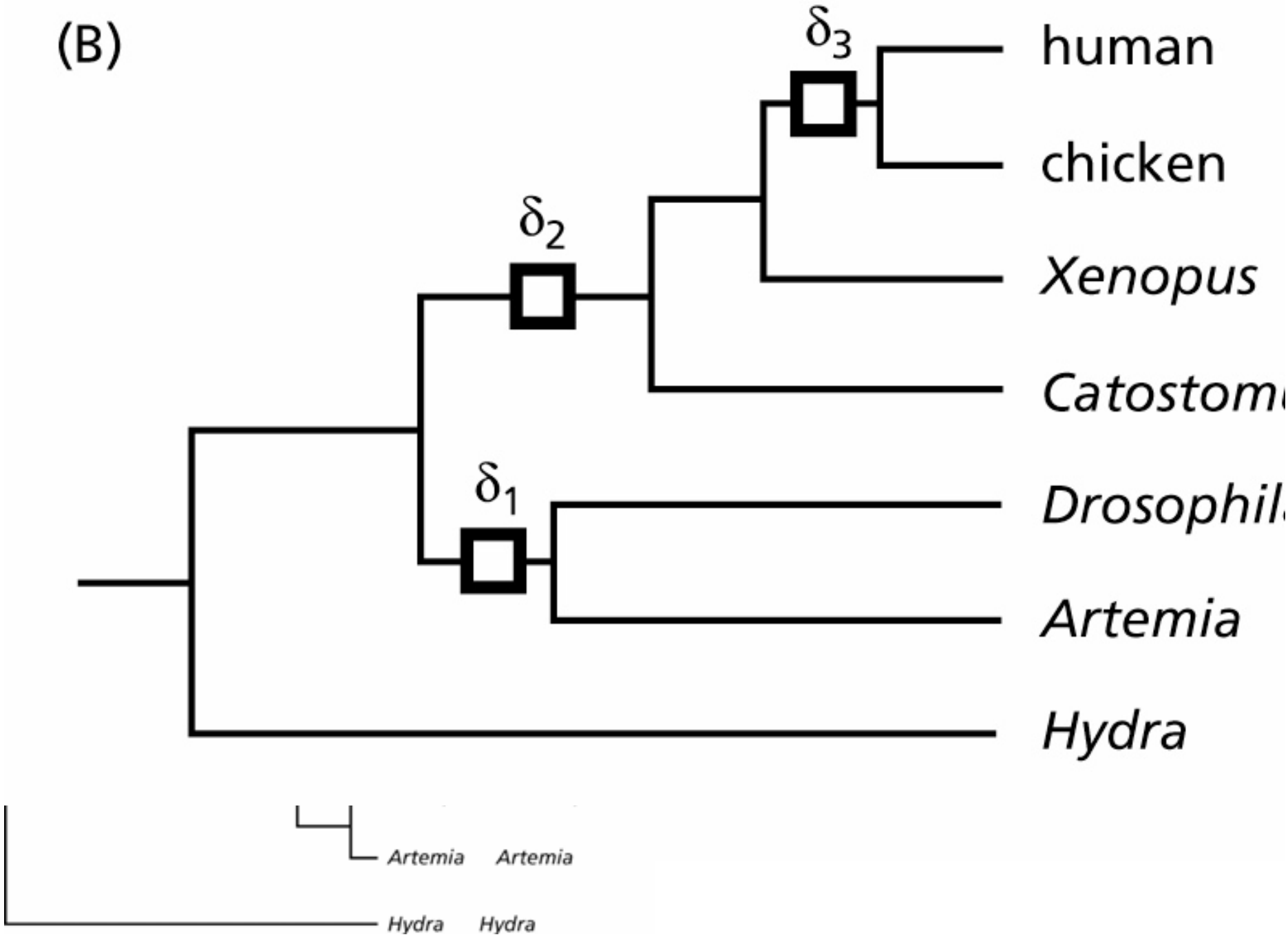
To detect orthhologs (COGs)



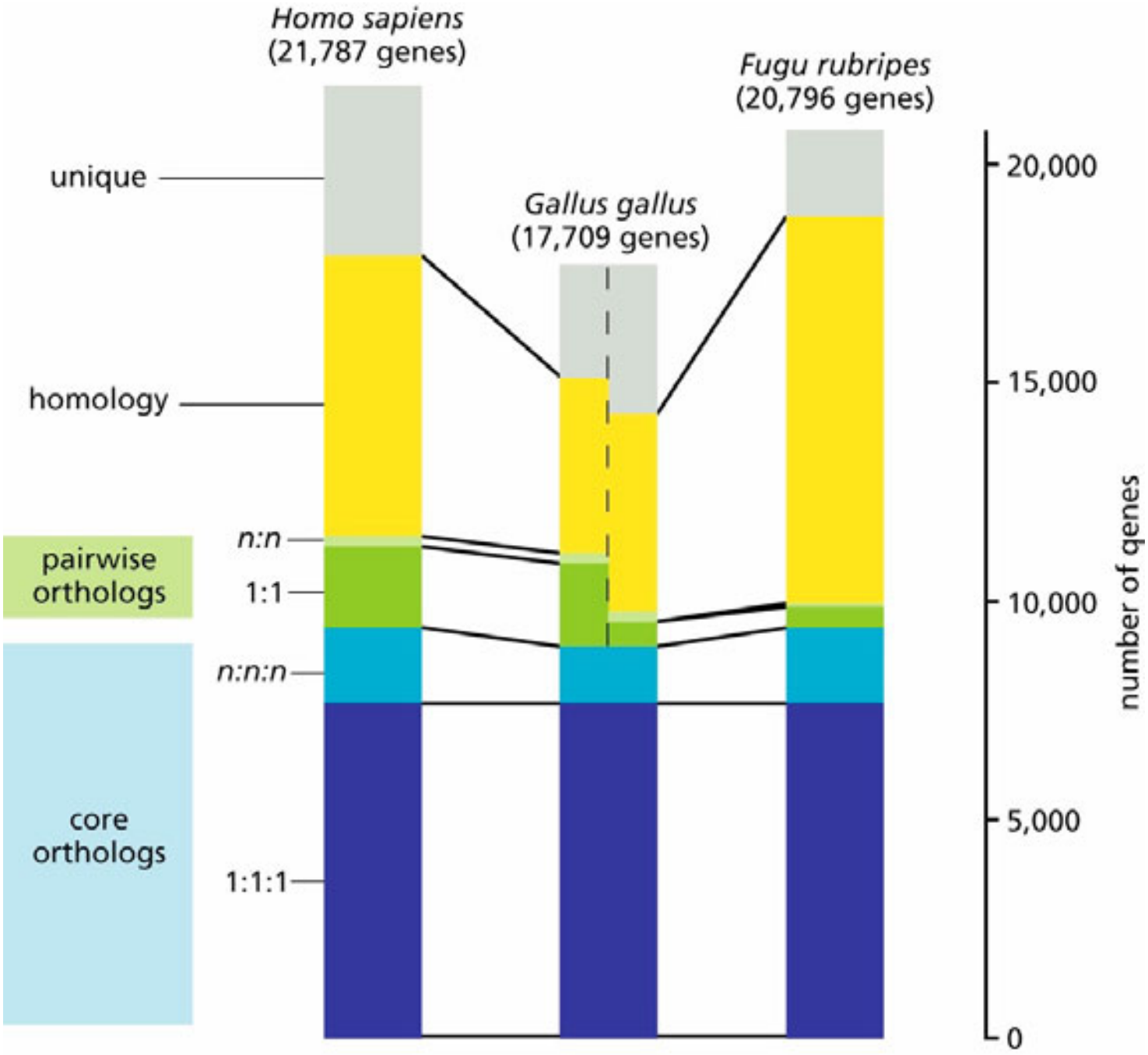
Why gene loss might influence species trees



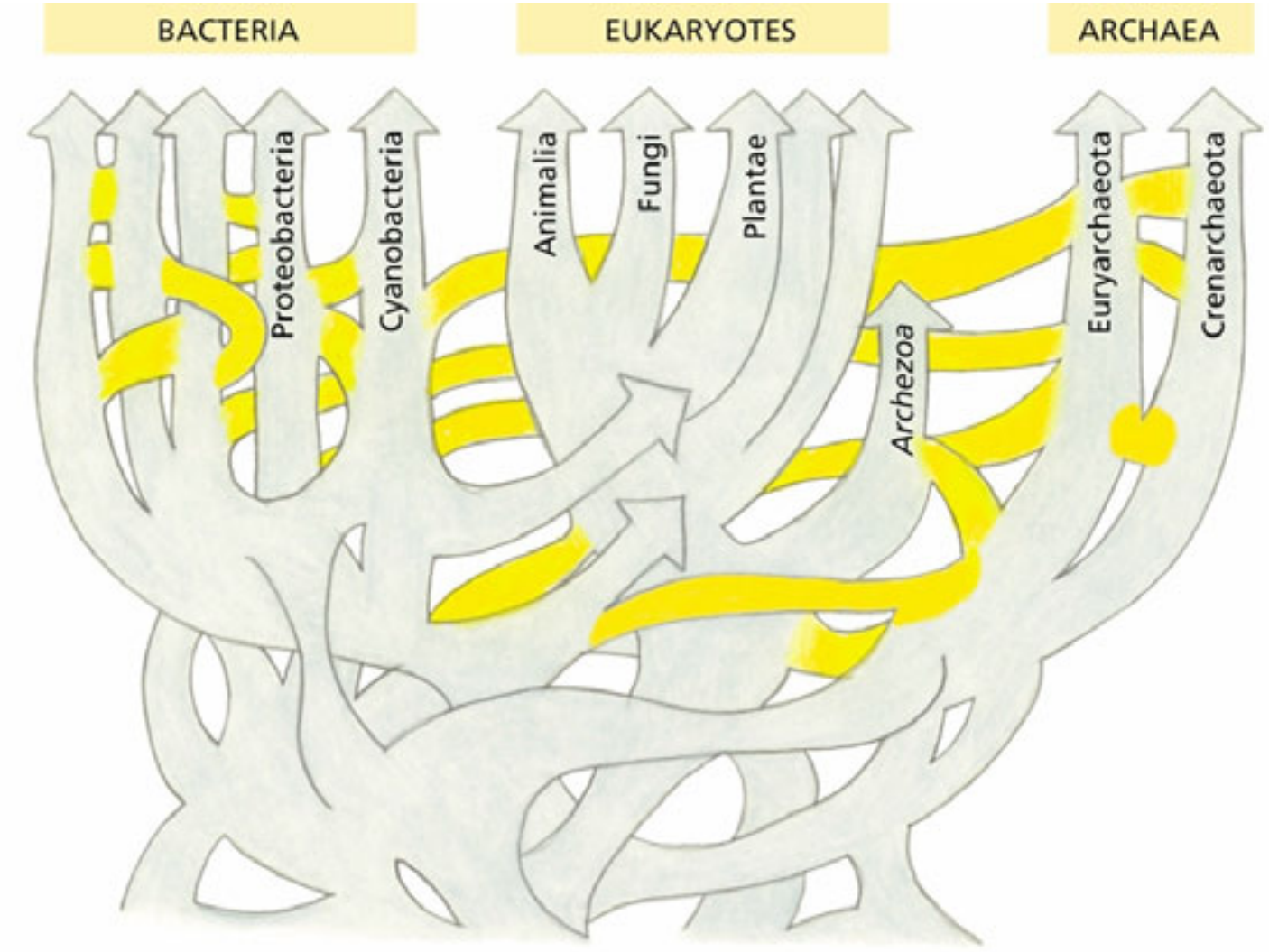
Reconciled gene/species trees



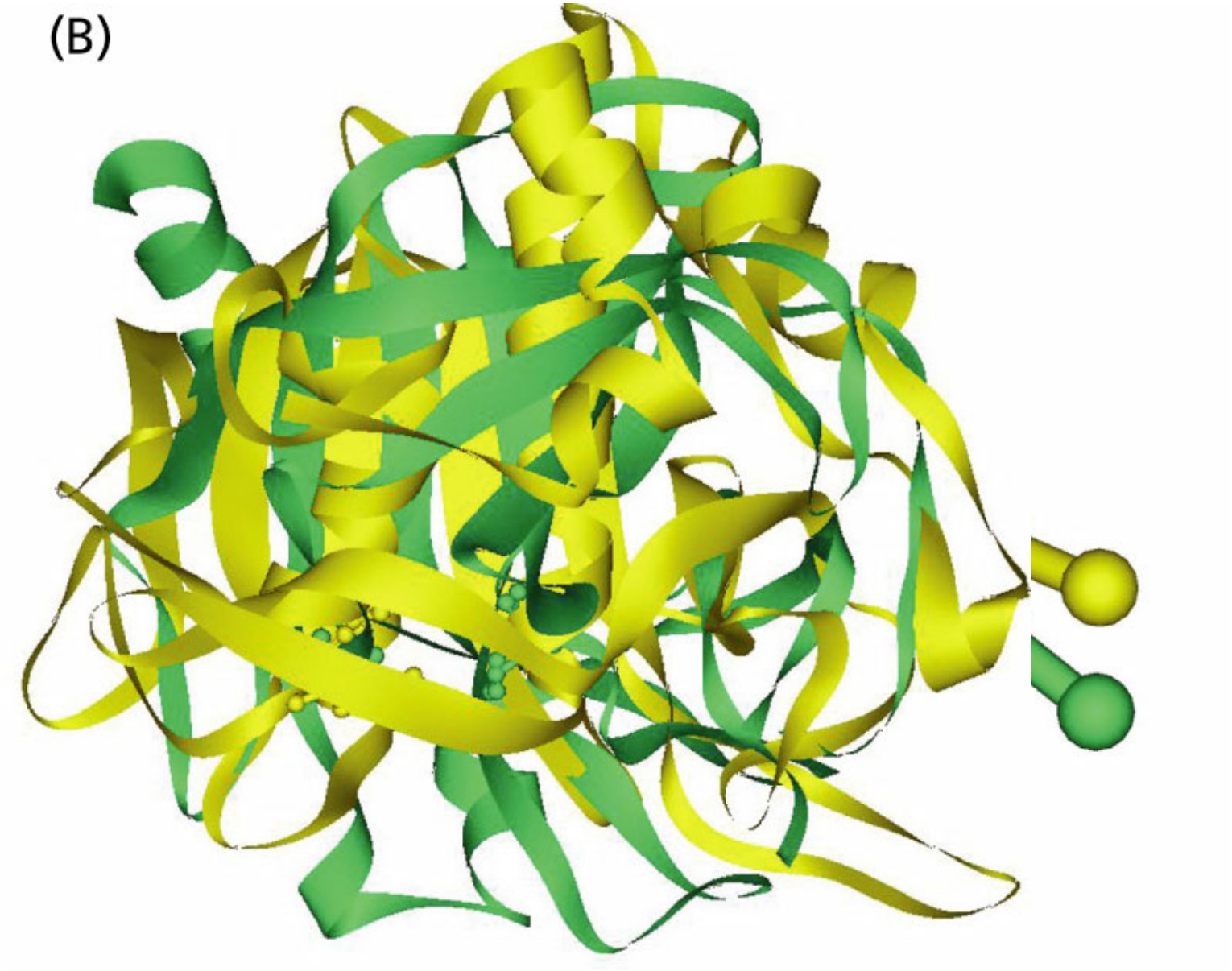
Number of orthologs



Horizontal gene transfer

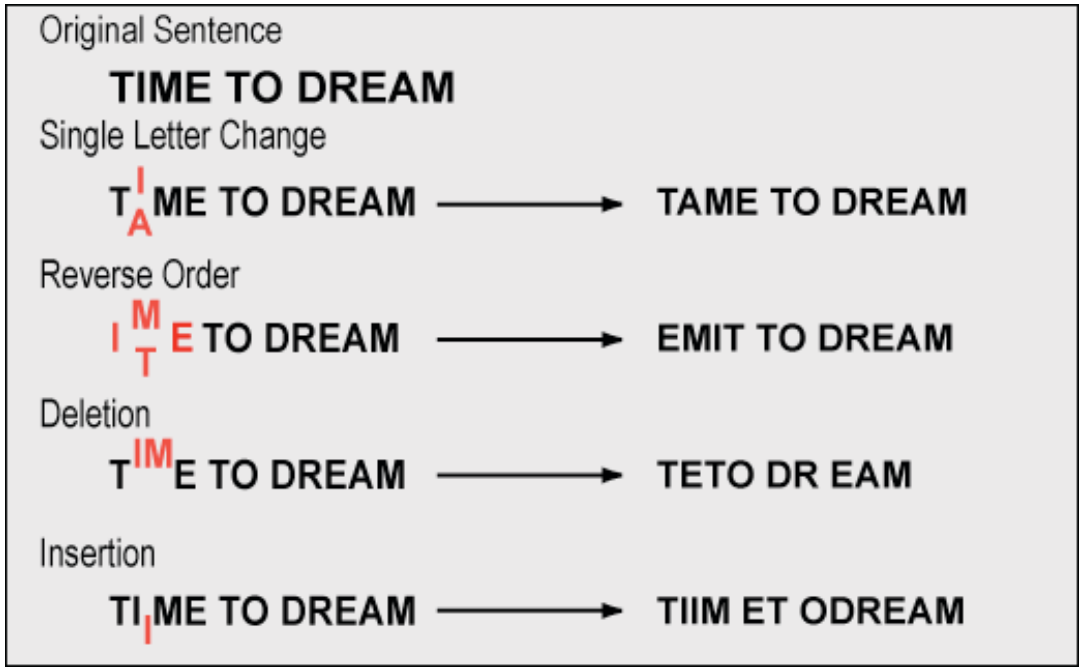


All similarities are not homologous



Basic evolutionary steps

- Mutations
  - Non-synonymous
  - synonymous
  - Gaps
- Gene duplication
- Gene fusion
- Horizontal transfer



Evolutionary steps II

- Selection
  - Positive
  - Negative
  - Neutral
- Selectionists vs. neutralists
- Speciation

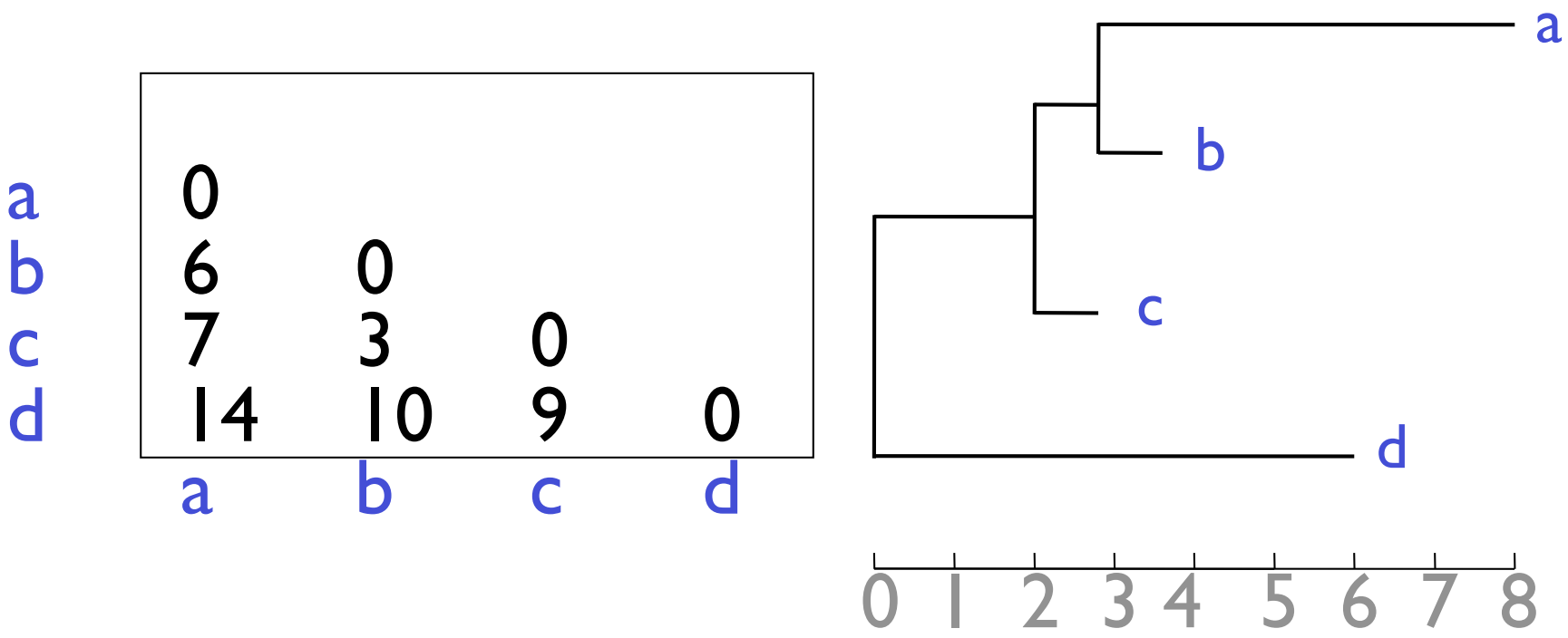
Adaptive evolution

- Selection
  - No selection – ”neutral evolution”
  - The ’null frequency’ with which amino acid-mutations are expected to happen by chance alone
- Negative selection – ”conservation”
  - AA-mutations happens less often than the null frequency
  - Positive selection – ”functional change”
  - AA-mutations happens more often than null frequency

How to make a phylogenetic tree

- Single phylogenetic method
  - UPMGA, Fitch, Neighbor joining
- Multiple trees
  - Parsimony
  - Maximum likelihood

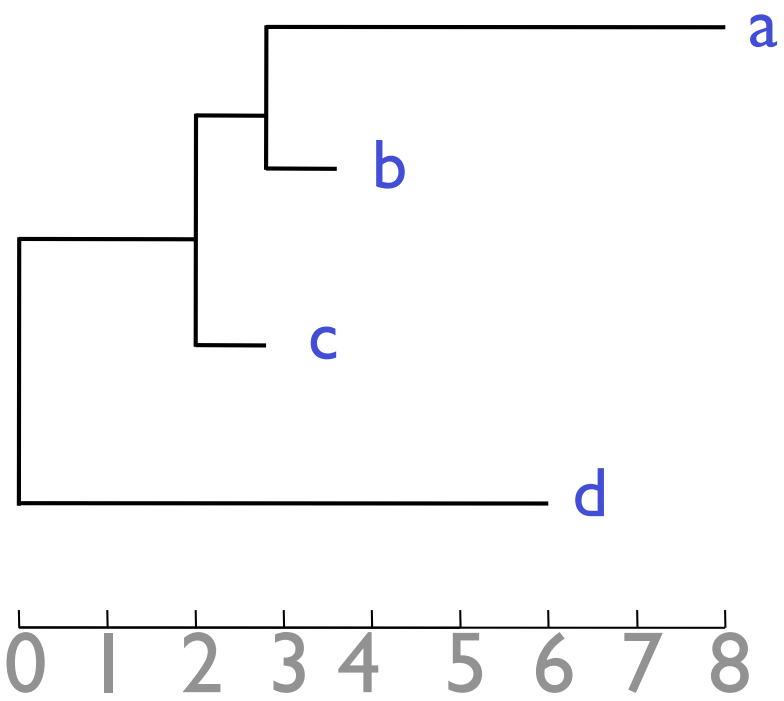
Distance Matrices



Distance matrix is ***additive*** if there is a tree that fits it exactly

Distance Matrices

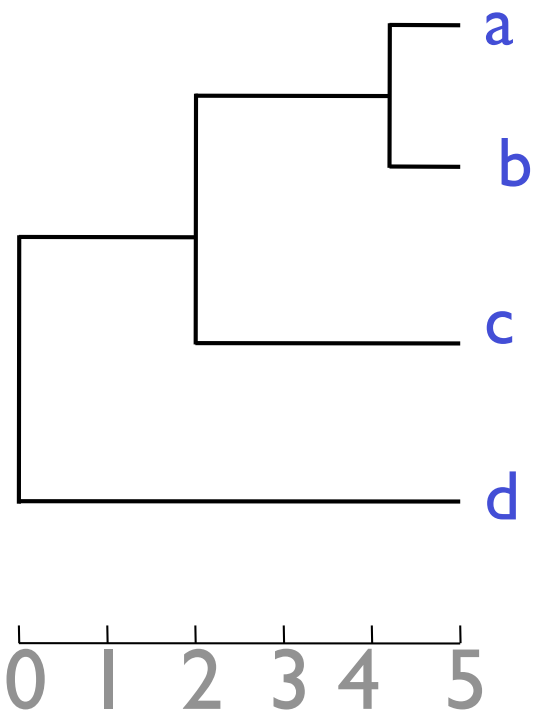
	a	b	c	d
a	0			
b	6	0		
c	7	3	0	
d	14	10	9	0



Distance matrix is ***additive*** if there is a tree that fits it exactly

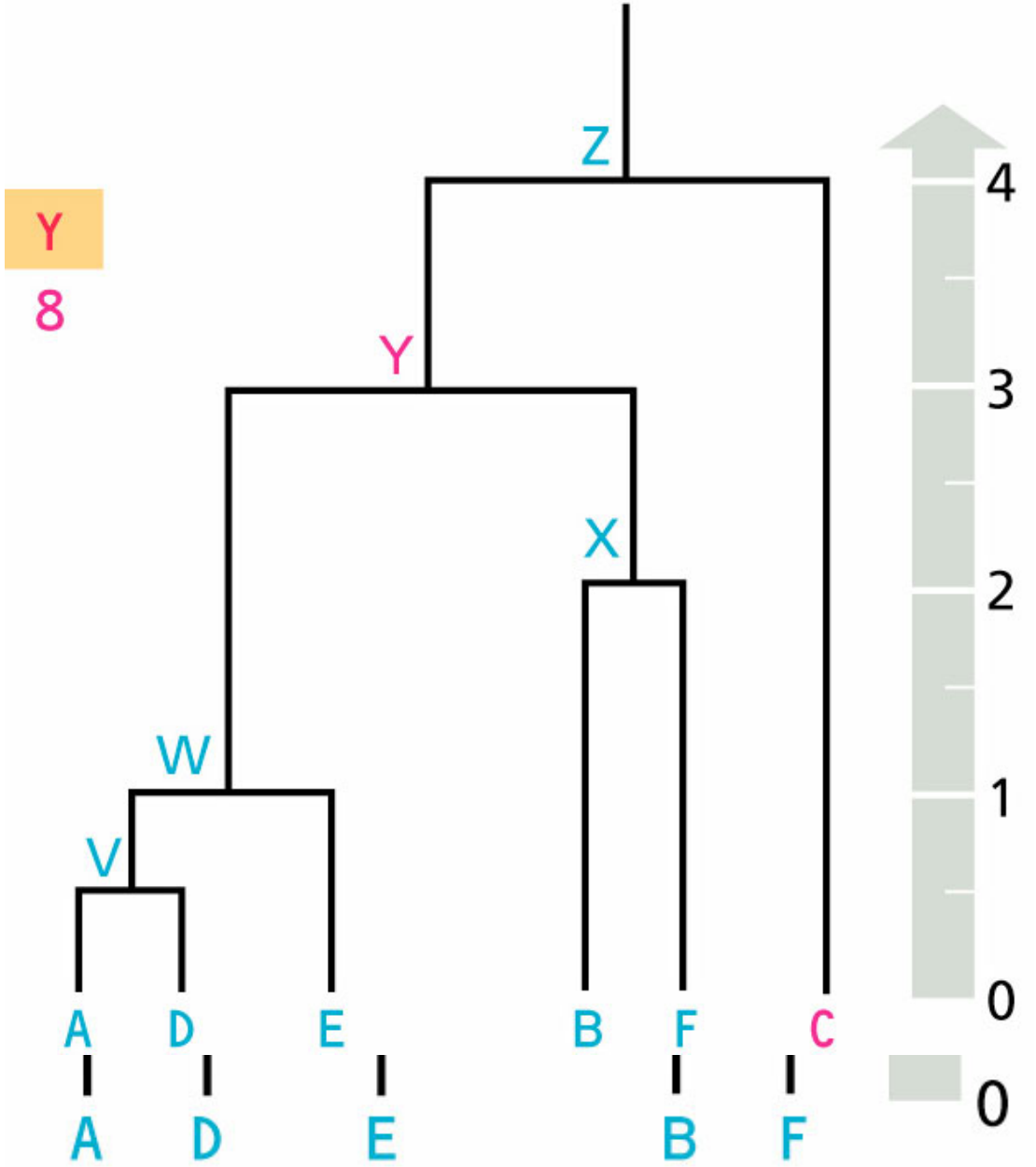
Ultrametric Matrices

	a	b	c	d
a	0			
b	2	0		
c	6	6	0	
d	10	10	10	0



Additive + molecular clock assumption

UPMGA



d_{ij}	A	B	C	D	E	F
A	–	6	8	1	2	6
B		–	8	6	6	4
C			–	8	8	8
D				–	2	6
E					–	6

Neighbor joining

(C) STEP 3 ($N = 3$)

d_{ij}

	B	X
A	5	3
B		4
X		

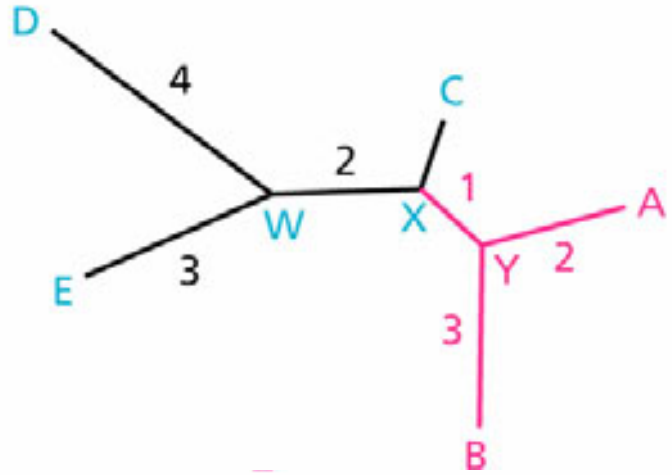
U_i

8
9
7

δ_{ij}

B	X	
−12	−12	A
	−12	B
		X

Three alternatives (of which here we choose one of the two with an internal node):
A and X are neighbors through internal node Y with $d_{AY} = 2$ and $d_{XY} = 1$ or
B and X are neighbors through internal node Y with $d_{BY} = 3$ and $d_{XY} = 1$.
Whichever is chosen, the remaining distance d_{AY} or d_{BY} will be found in the next d_{ij} matrix.



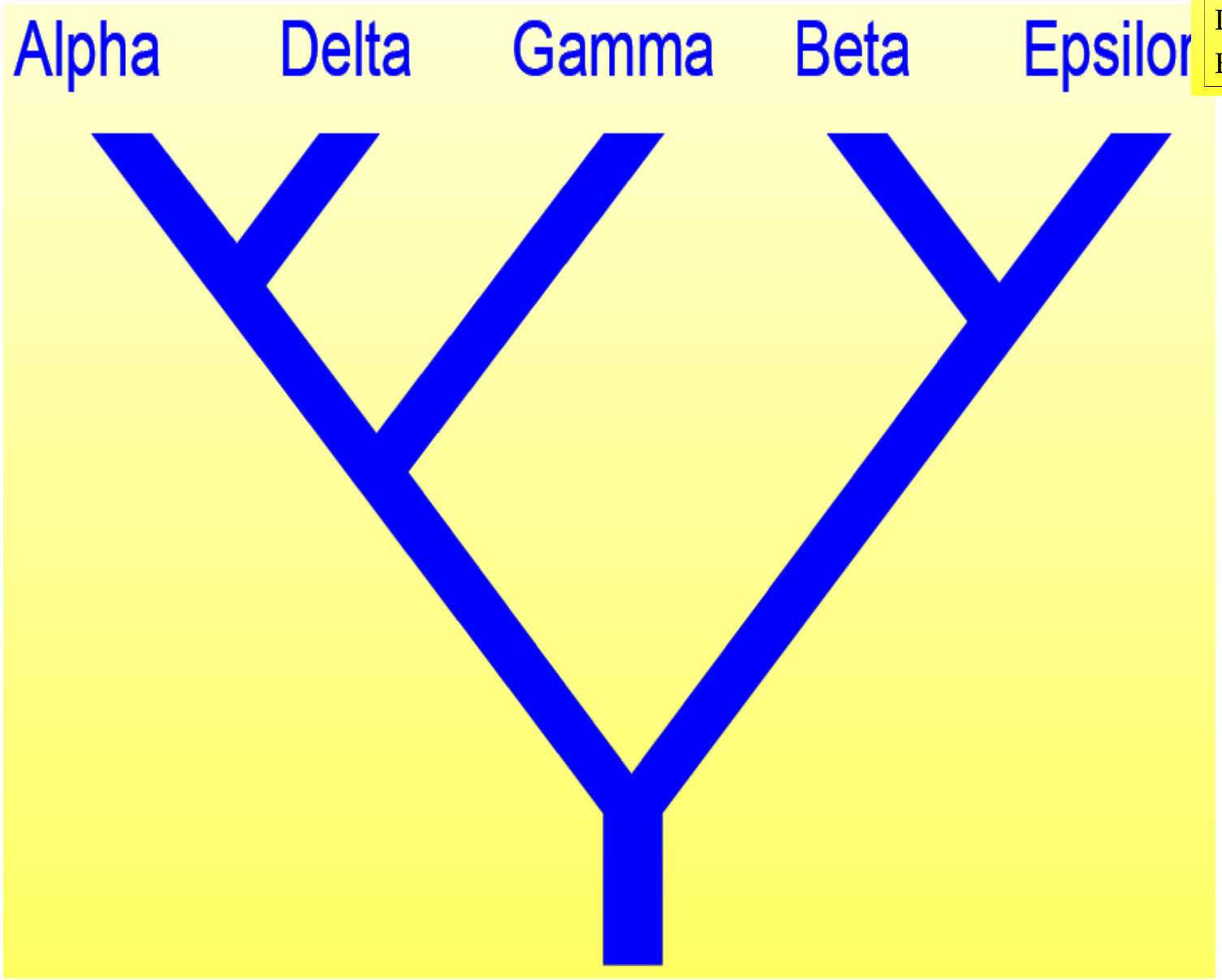
Parsimony example: A DNA data set

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

[F05]

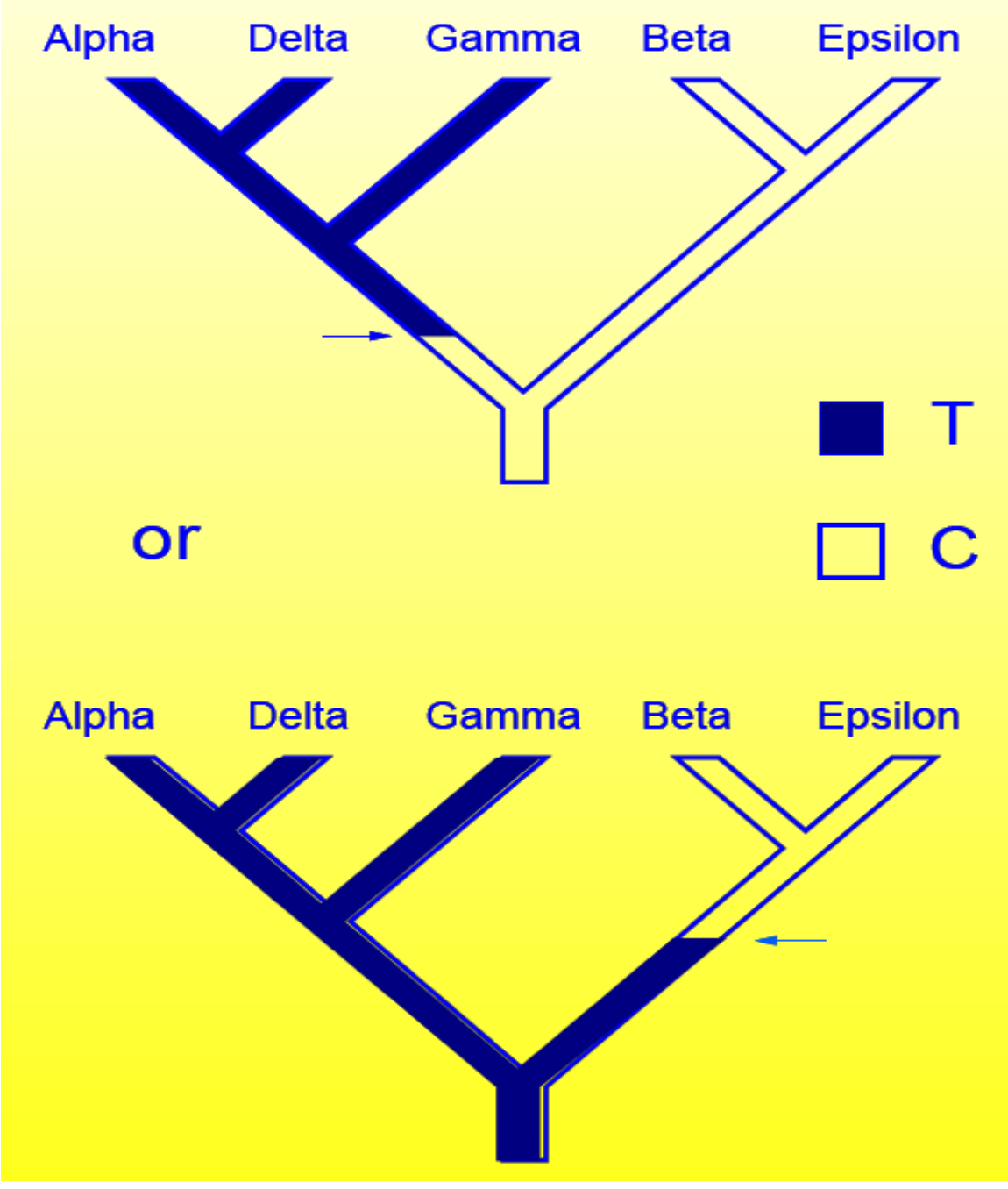
An example tree

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



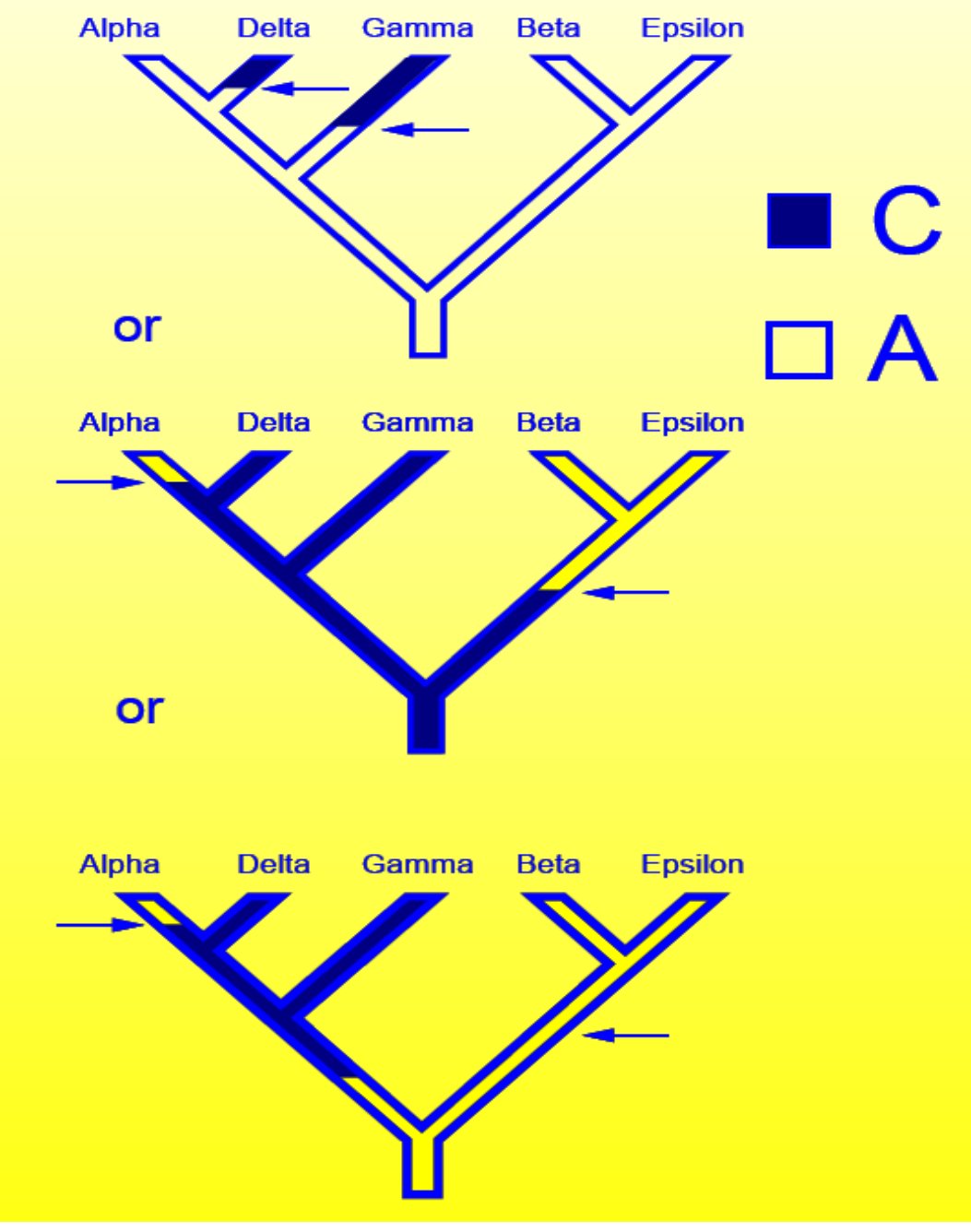
[F05]

Most parsimonious states for site 1



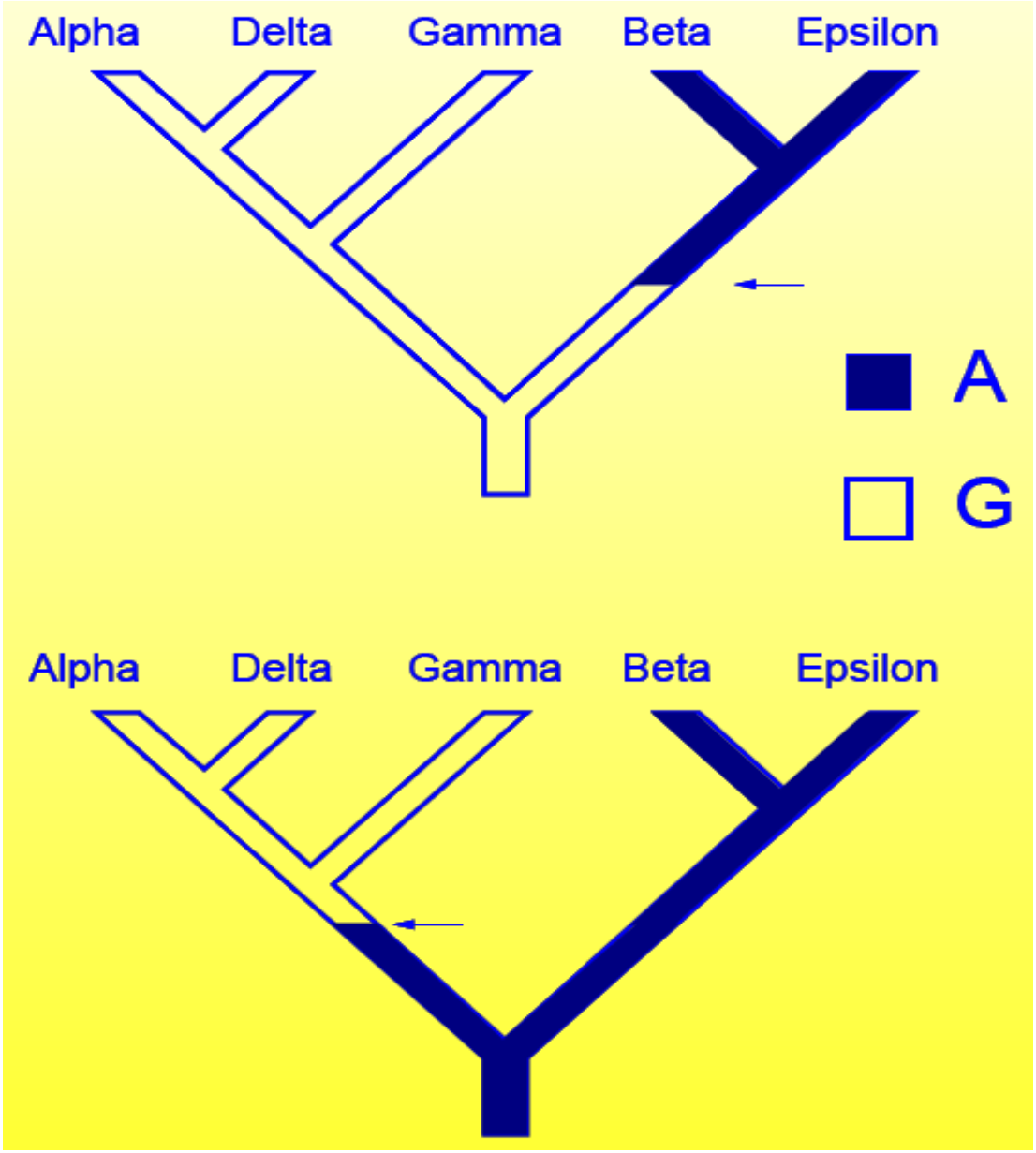
	Characters					
Species	1	2	3	4	5	6
Alpha	T	A	G	C	A	T
Beta	C	A	A	G	C	T
Gamma	T	C	G	G	C	T
Delta	T	C	G	C	A	A
Epsilon	C	A	A	C	A	T

Most parsimonious states for site 2



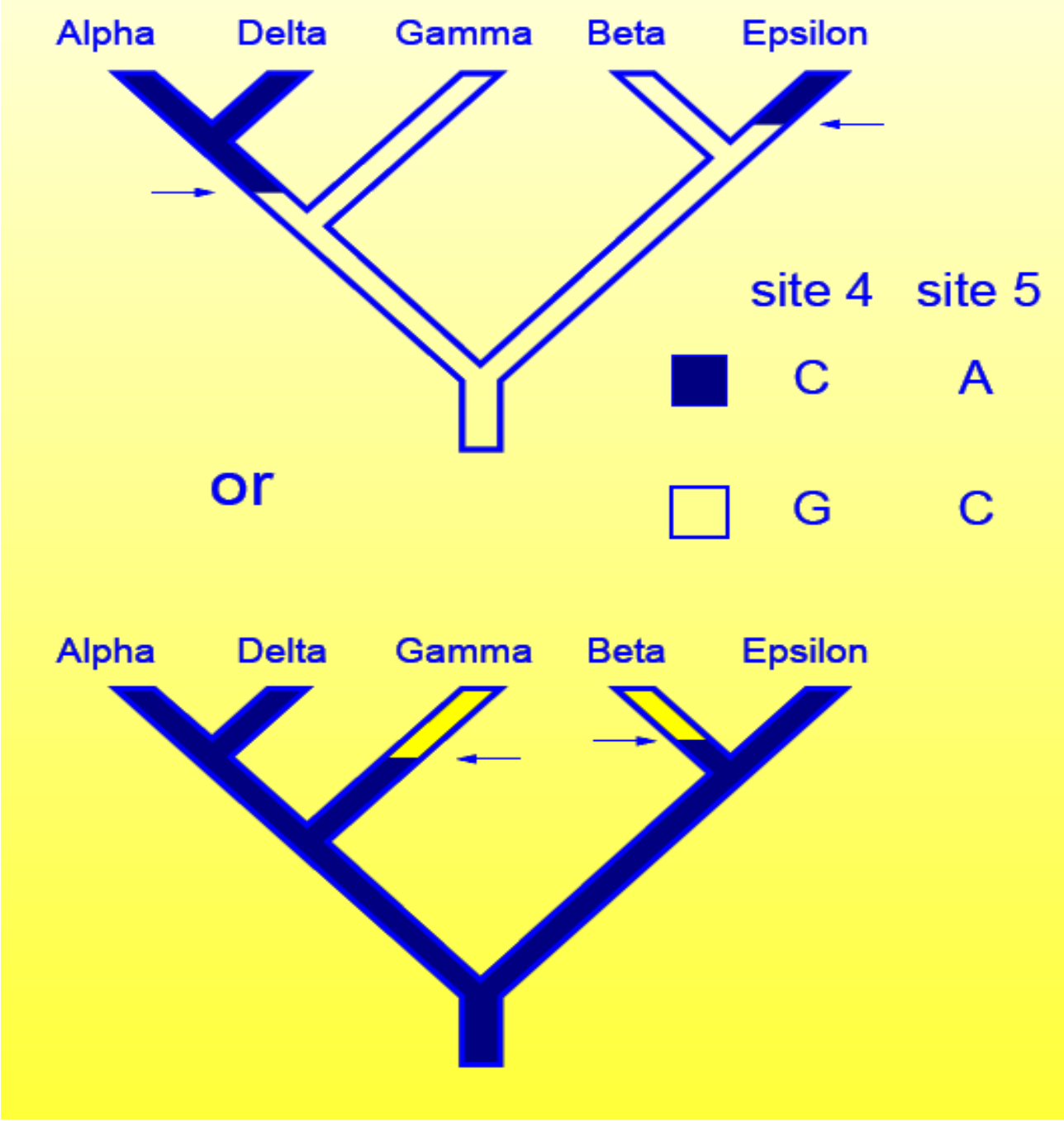
Species	Characters					
	1	2	3	4	5	6
Alpha	T	A	G	C	A	T
Beta	C	A	A	G	C	T
Gamma	T	C	G	G	C	T
Delta	T	C	G	C	A	A
Epsilon	C	A	A	C	A	T

Most parsimonious states for site 3



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

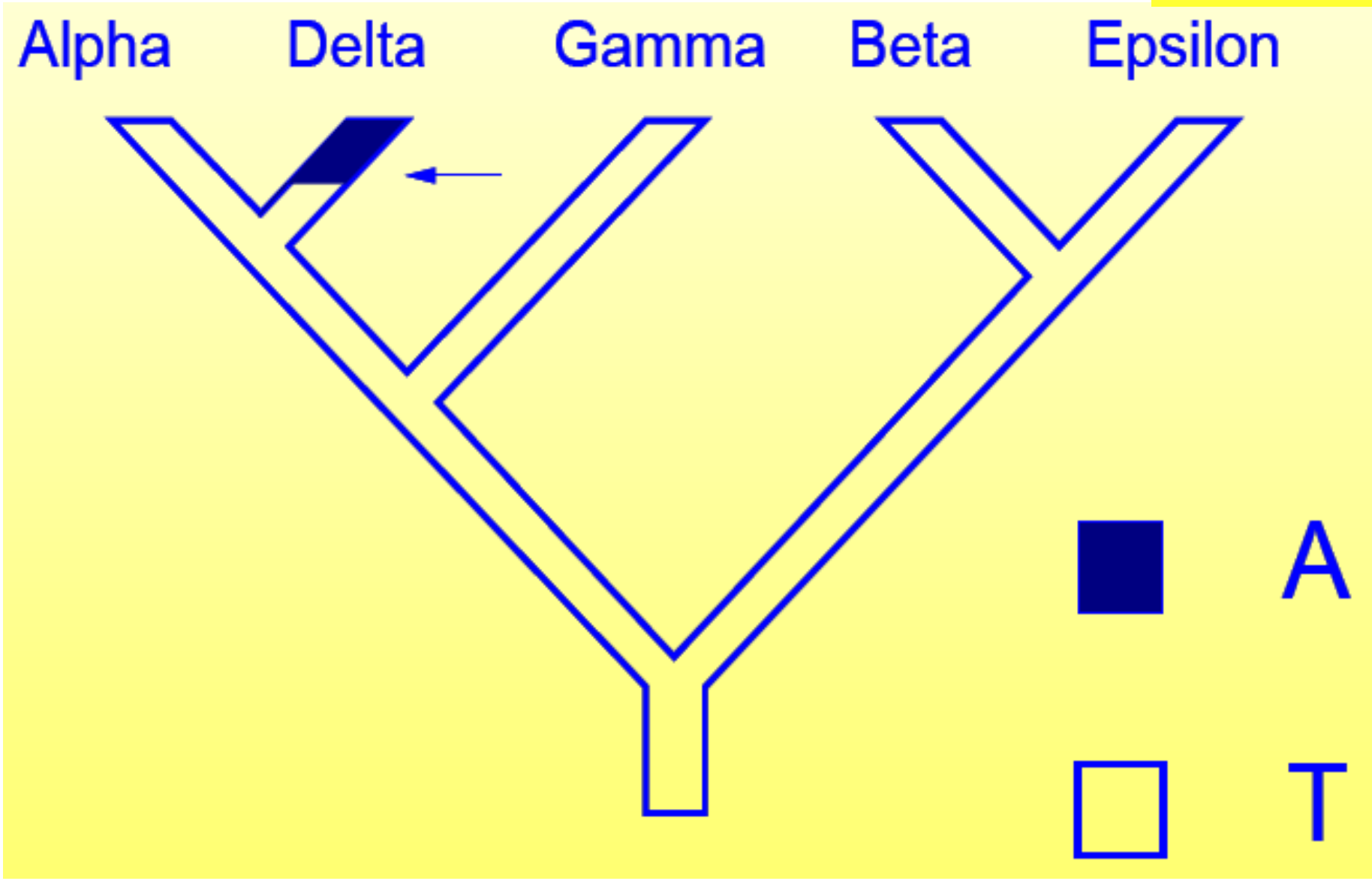
Most parsimonious states for sites 4 and 5



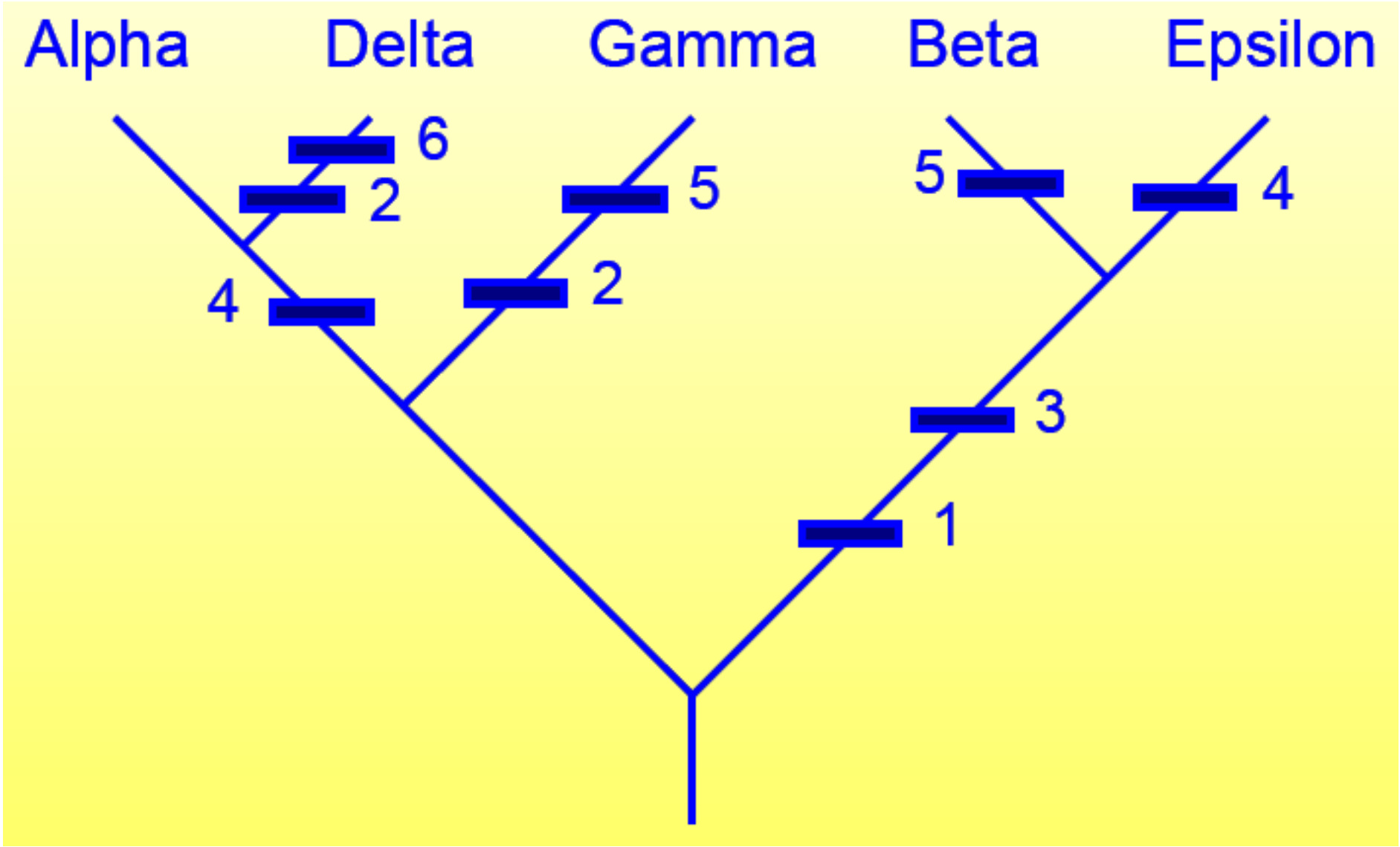
Species	Characters 1	2	3	4	5	6
Alpha	T	A	G	C	A	T
Beta	C	A	A	G	C	T
Gamma	T	C	G	G	C	T
Delta	T	C	G	C	A	A
Epsilon	C	A	A	C	A	T

Most parsimonious states for site 6

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



Evolutionary steps on tree



Only one choice of reconstruction at each site is shown
9 steps in all

Some interesting facts !!

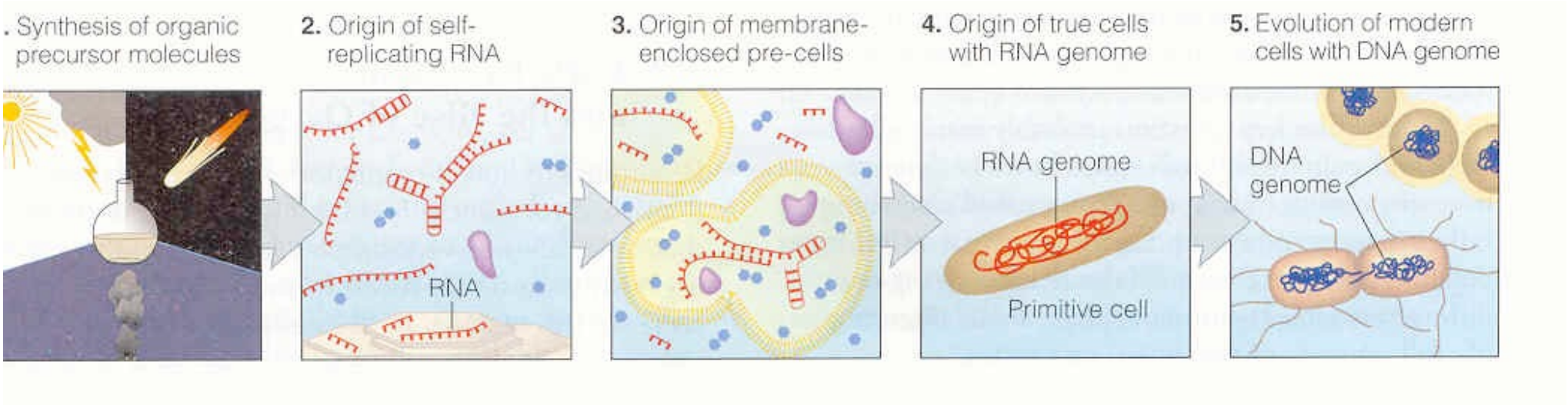
Evolution of life on earth

- Million Years Ago
- 4000 M Early life started early – HOW ?
- 3500 M Photosynthesis
- 2600 M LUCA (Last Universal Common Ancestor)
  - Three kingdoms of life
- 545 M Cambrian explosion
  - Animals
- 200 M Rise of O_2 – current levels

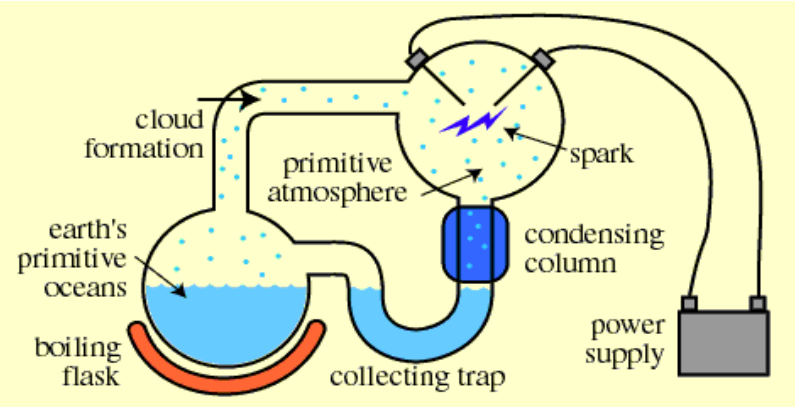
Arne Elofsson (arne@bioinfo.se)

How did life start

- Life started early in the history of life
- Exact time is hard to tell
- Microfossils dating to 3.5 billion years ago
- Difficult to distinguish from mineral structures
- Evidence in metamorphic rocks that life existed 3.85 billions years ago

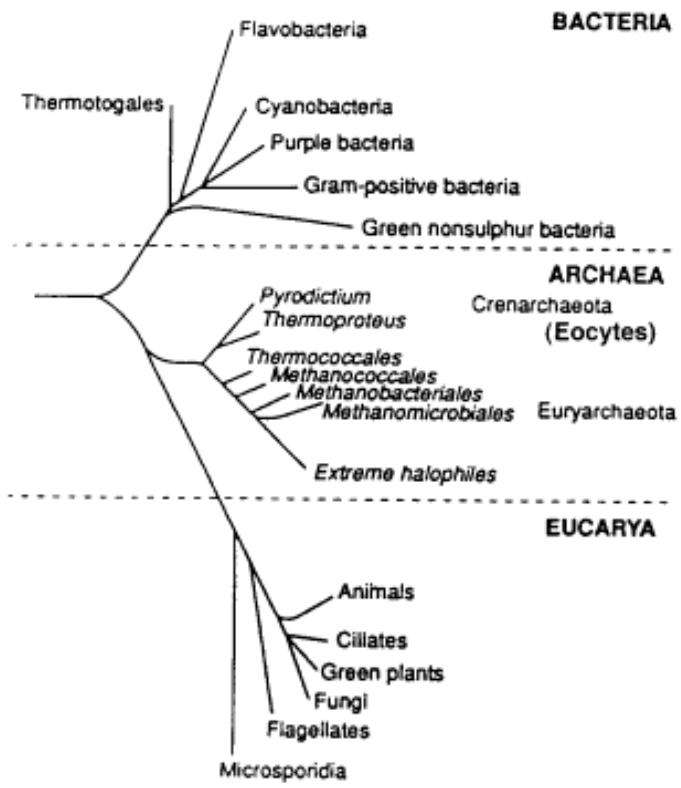


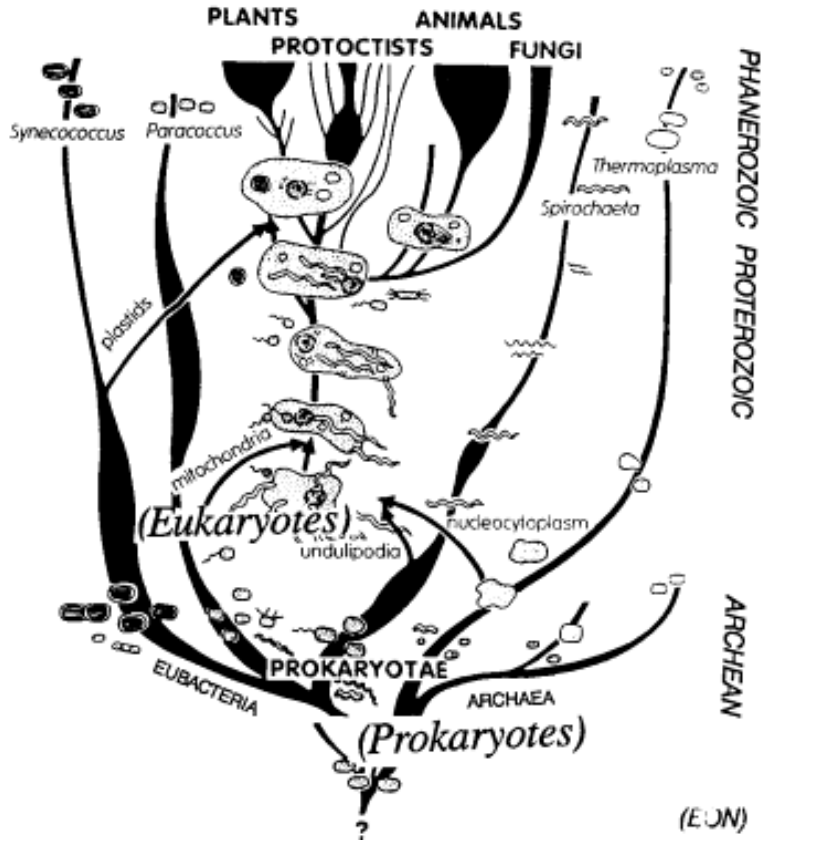
Miller-Urey Experiment



- First flask partially filled with water and heated to produce water vapor (sea)
- Water vapor was moved to a second flask where methane and ammonia vapor was added (atmosphere)
- Electric sparks (lightening) in second flask was energy source for chemical reactions
- Below second flask, water vapor cooled (rain) and recycled to first flask (sea)
- Result: turned brown with amino acids and other complex organic molecules

Tree of life

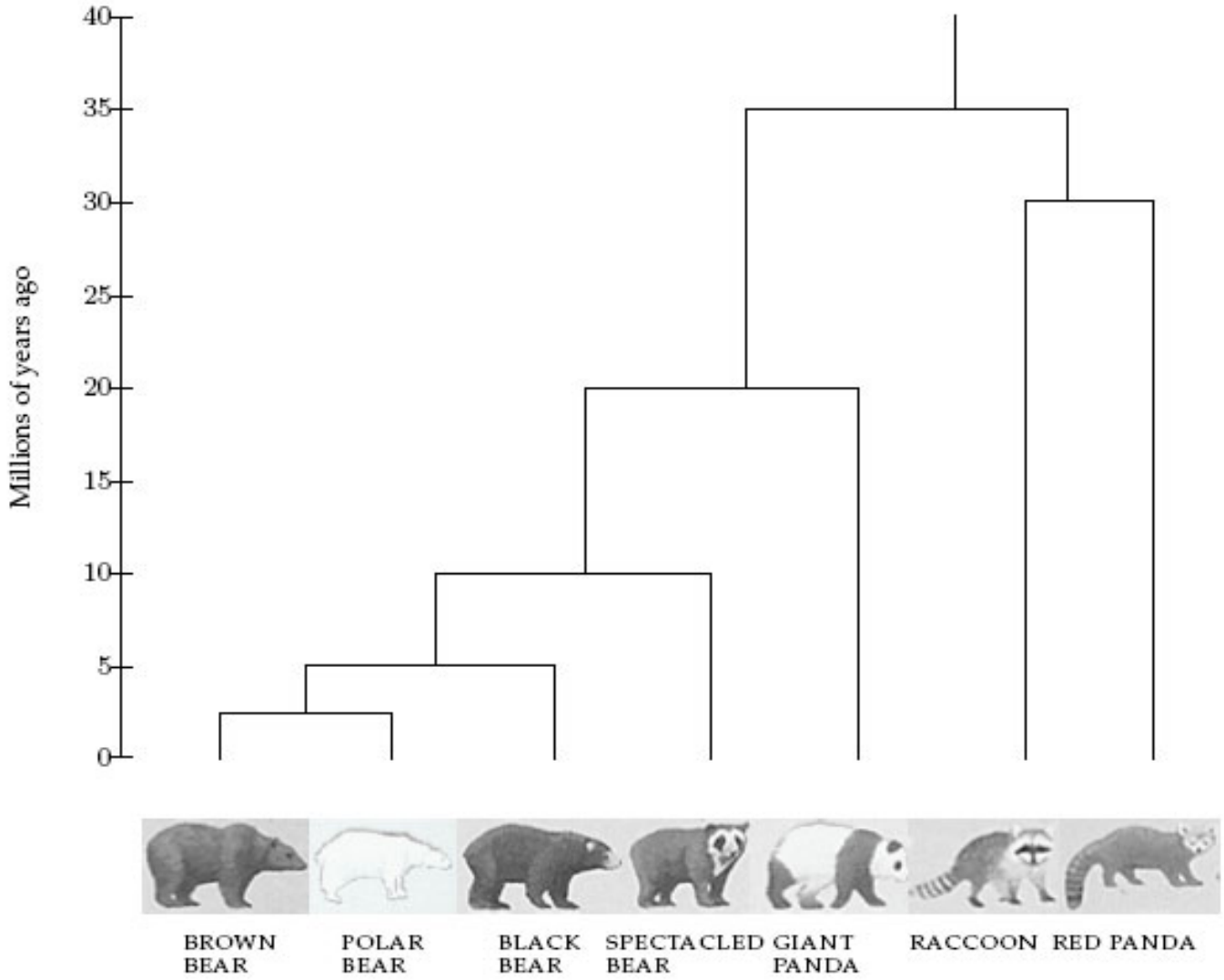

BACTERIA
Flavobacteria
Thermotogales
Cyanobacteria
Purple bacteria
Gram-positive bacteria
Green nonsulphur bacteria
ARCHAEA
Pyrodictium
Thermoproteus
Crenarchaeota
(Eocytes)
Thermococcales
Methanococcales
Methanobacteriales
Methanomicrobiales
Euryarchaeota
Extreme halophiles
EUCARYA
Animals
Ciliates
Green plants
Fungi
Flagellates
Microsporidia


PLANTS
ANIMALS
PROTOCTISTS
FUNGI
Synecococcus
Paracoccus
Thermoplasma
Spirochaeta
plastids
mitochondria
(Eukaryotes)
undulipodia
nucleocytoplasm
PROKARYOTAE
EUBACTERIA
ARCHAEA
(Prokaryotes)
?
PHANEROZOIC
PROTEROZOIC
ARCHEAN
(EON)

Evolution and DNA Analysis: the Giant Panda Riddle

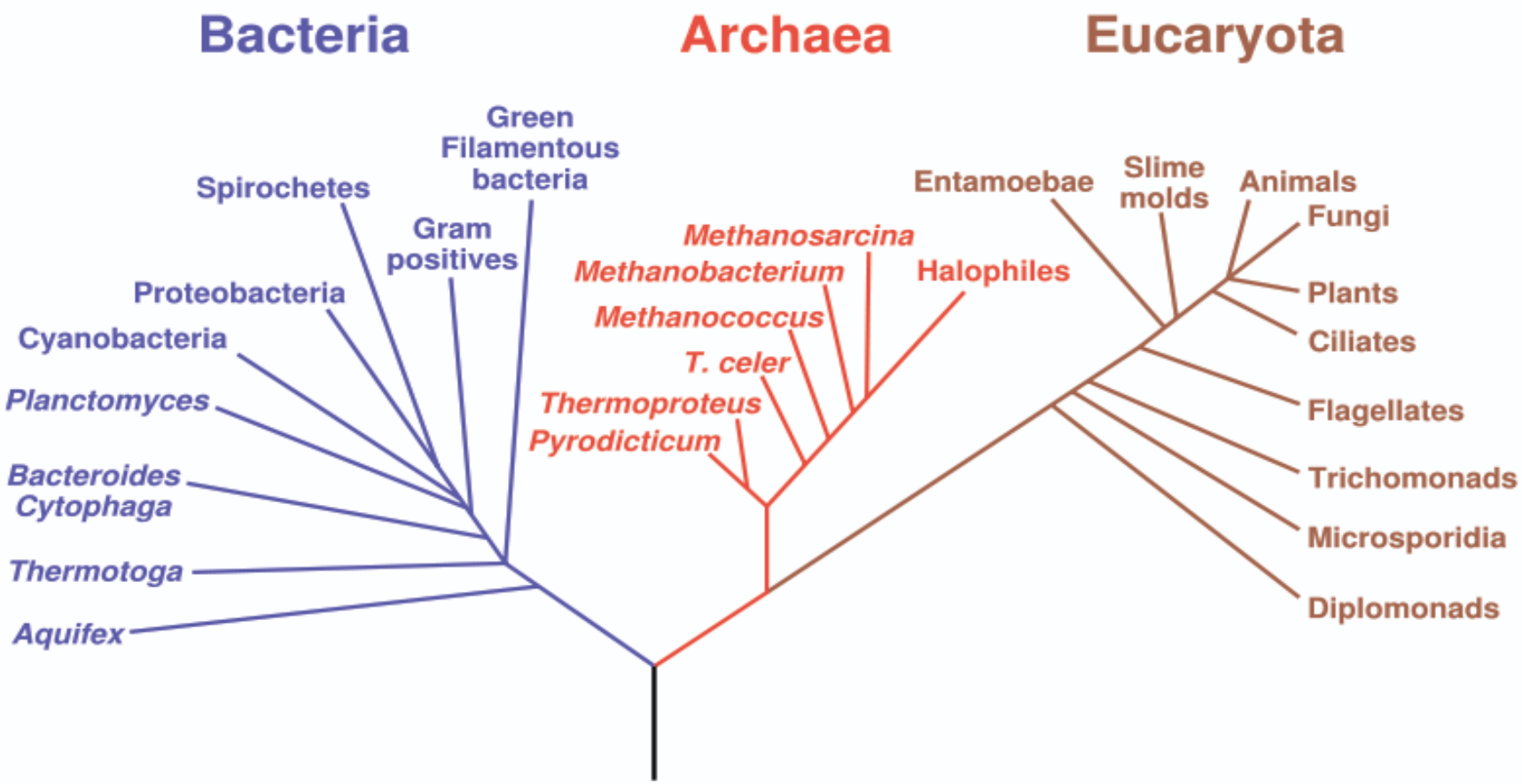
- For roughly 100 years scientists were unable to figure out which family the giant panda belongs to
- Giant pandas look like bears but have features that are unusual for bears and typical for raccoons, e.g., they do not hibernate
- In 1985, Steven O'Brien and colleagues solved the giant panda classification problem using DNA sequences and algorithms

Evolutionary Tree of Bears and Raccoons



Arne Elofsson (arne@bioinfo.se)

Phylogenetic Tree of Life



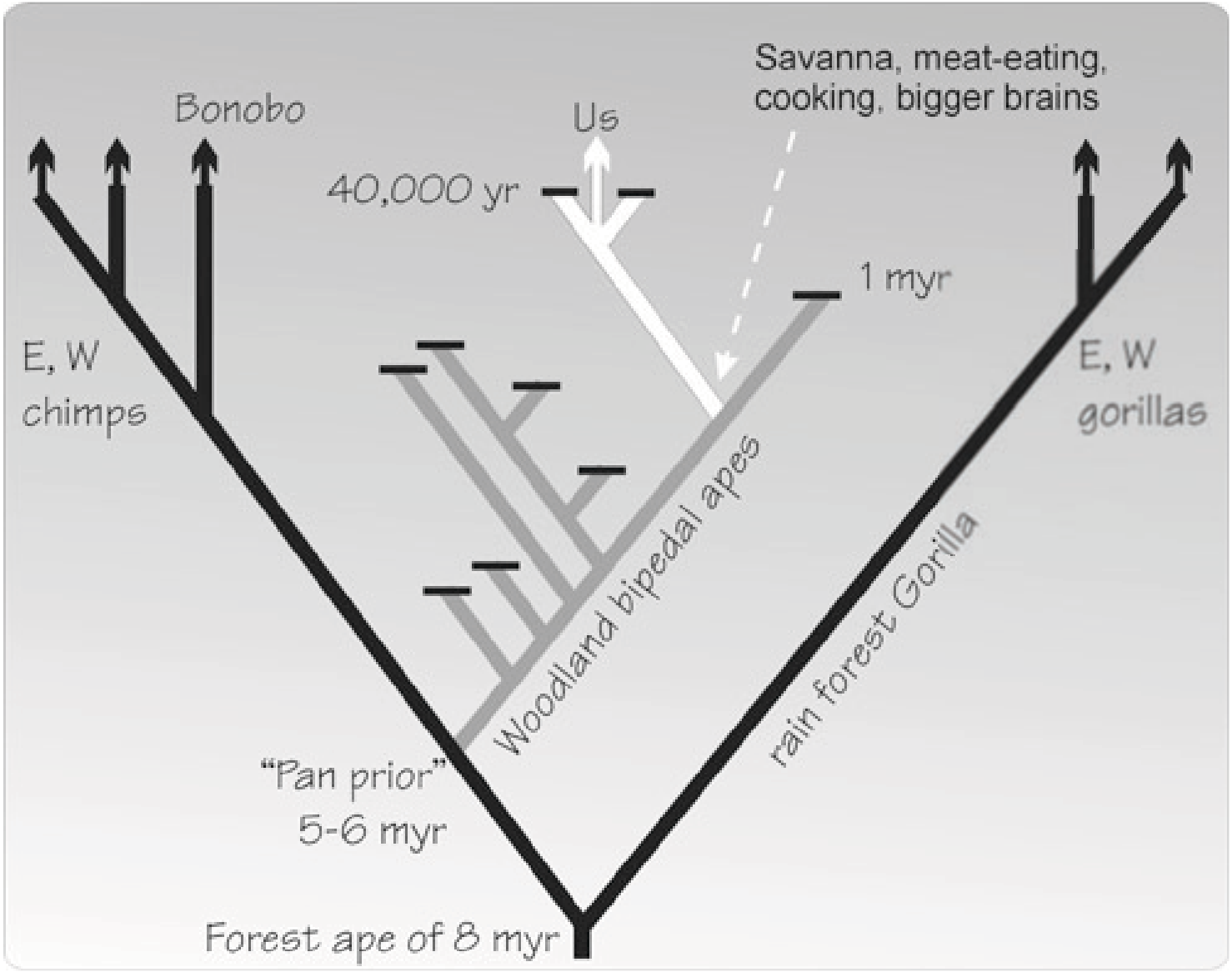
Archaea

- Archaea were identified in 1977 by Carl Woese and George E. Fox as being a separate branch based on their separation from other prokaryotes on 16S rRNA phylogenetic trees
- Archaea are similar to other prokaryotes in most aspects of cell structure and metabolism. However, their genetic transcription and translation — the two central processes in molecular biology — do not show many typical bacterial features, and are in many aspects similar to those of eukaryotes.'
- Often found in "extreme" environments.

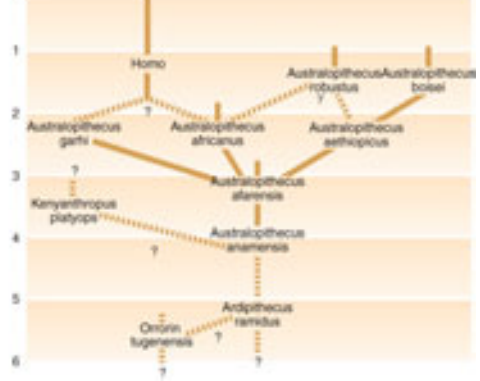
Human origin - who are we ?

- How did we evolve ?
- Can genomics help us with this questions ?
- What is our closest relative ?
- How are Neanderthals related to us.
- Are we sons/daughters of Monkeys ?
- Where did we origin ?
- How closely related are humans ?

Human, chimps and Gorilla



Evolutionary trends from hominid to human



What are some of the characteristics that evolved to make us uniquely human?

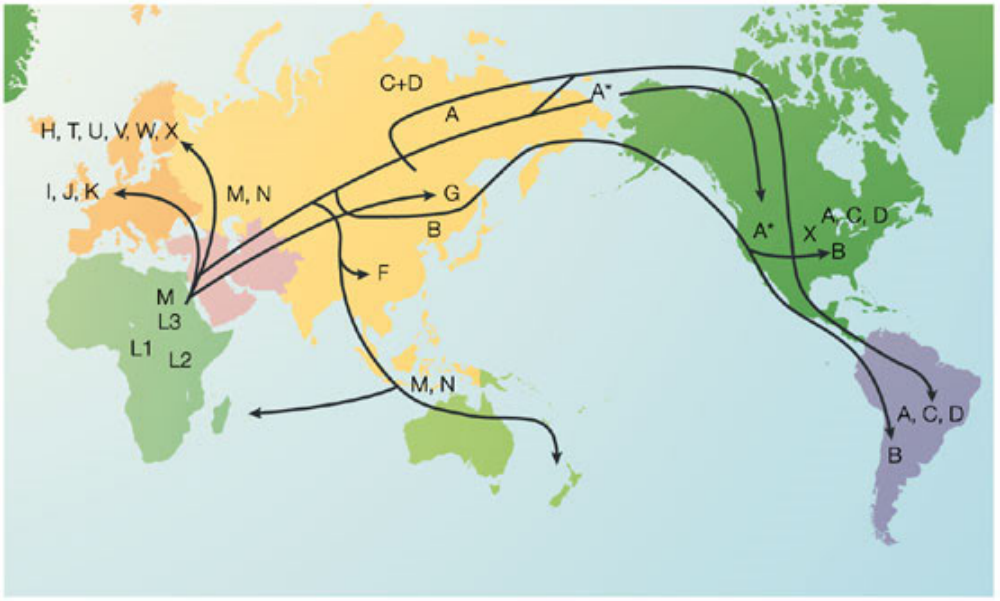
- Bipedalism - ~4 mill years ago
  - *Ardipithecus ramidus* (oldest hominid) and *Australopithecus anamensis*
- Smaller teeth (change in diet) - ~3 mill years ago
  - *Australopithecus afarensis* (Lucy)
- Reduction in robustness - ~2.5 – 3 mill years ago
  - *Australopithecus africanus* or *A. garhi*
- ↑ in brain size - ~2 – 2.5 mill years ago
  - *H. erectus*
- Art (symbolic expression) – 40,000 years ago
  - *H. sapiens*

http://www.youtube.com/watch?v=kU0ei9ApmsY

Arne Elofsson
(arne@bioinfo.se)

How do we interpret genetic variants to ask anthropological questions?

- Look at the patterns of genetic variation
- What has created the patterns of genetic variation?
  - Evolutionary history of humans

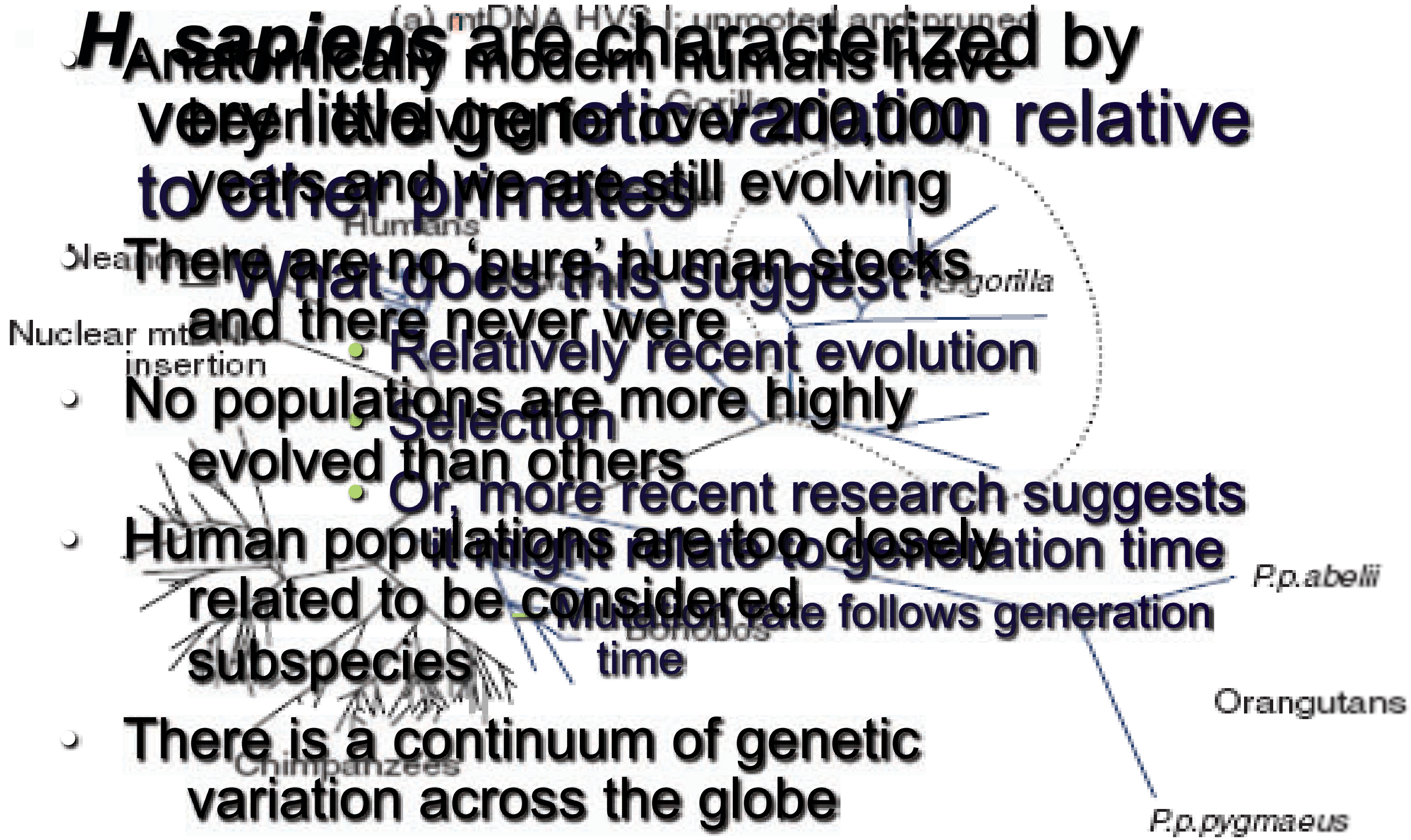


Shriver and Kittles, 2004

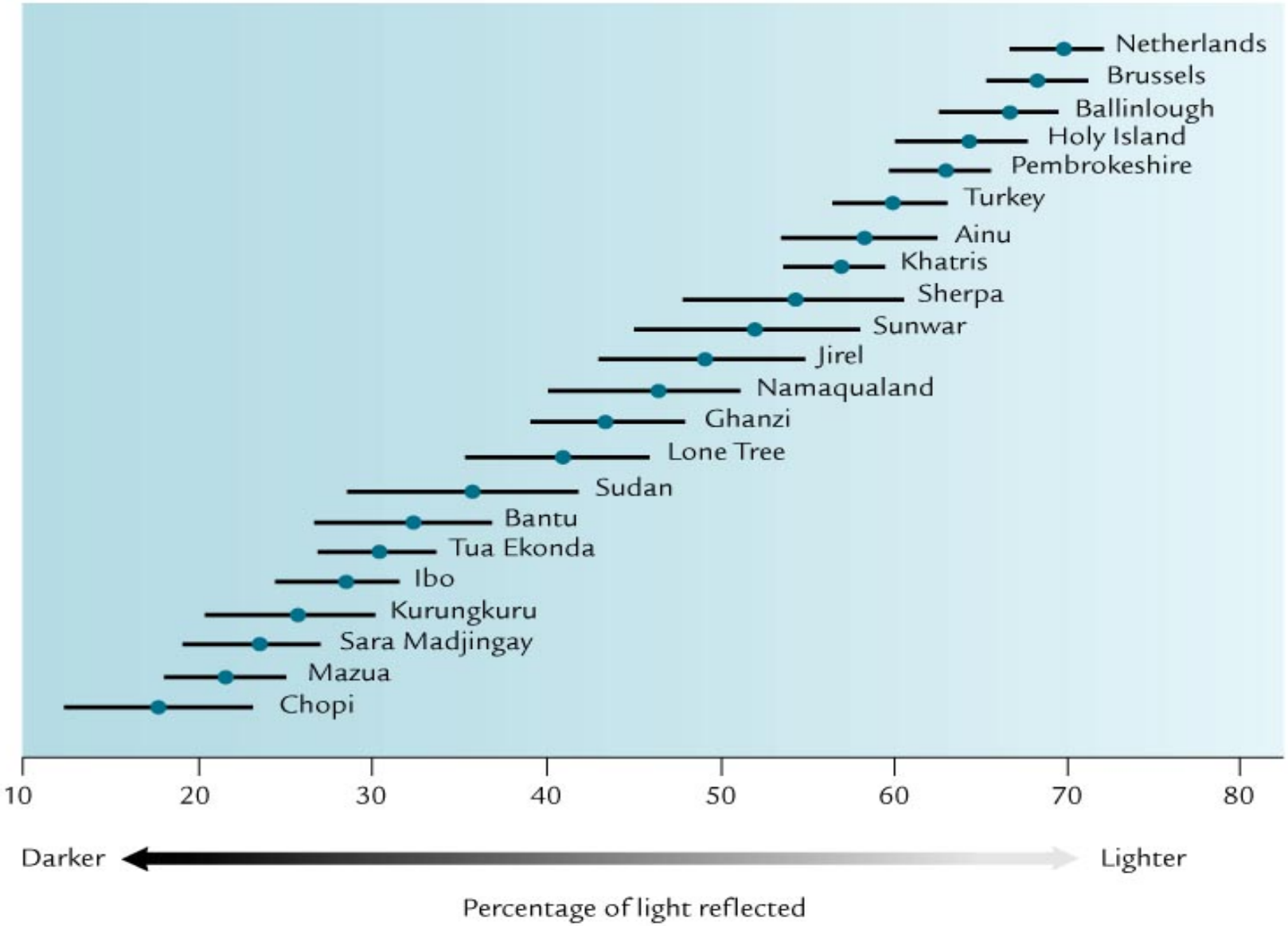
Human genetic diversity is low. What does that mean ?

- ***H. sapiens*** **are characterized by very little genetic variation relative to other primates**
  - What does this suggest?
    - Relatively recent evolution
    - Selection
    - Or, more recent research suggests it might relate to generation time
  - Mutation rate follows generation time

- Anatomically modern humans have been evolving for over 200,000 years and we are still evolving
- There are no 'pure' human stocks and there never were
- No populations are more highly evolved than others
- Human populations are too closely related to be considered subspecies
- There is a continuum of genetic variation across the globe



Variation in skin color in 22 populations

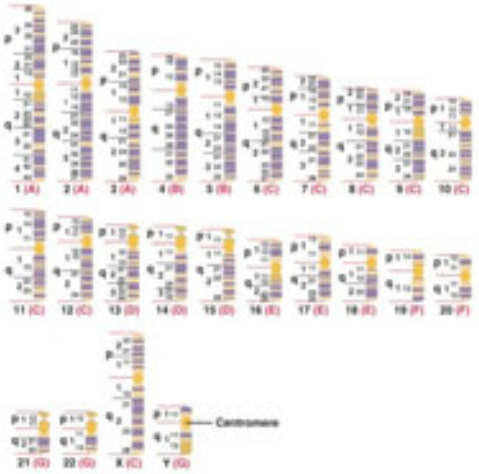


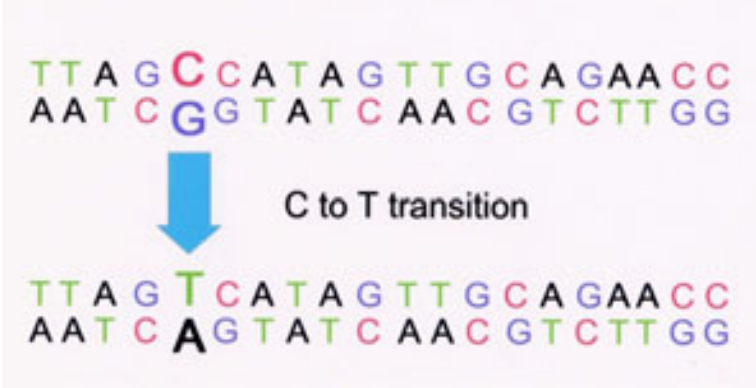
Genetic evidence for the origin of modern humans

- Greatest genetic diversity is i
  populations
  - Consistent with out-of-Africa theory of human origin
- Most genetic variation in humans is within populations, not between populations
  - ~85% within populations
  - ~5% between populations on same continent
  - ~10% between populations on different continents, i.e. races

But there are different frequencies of alleles in different populations

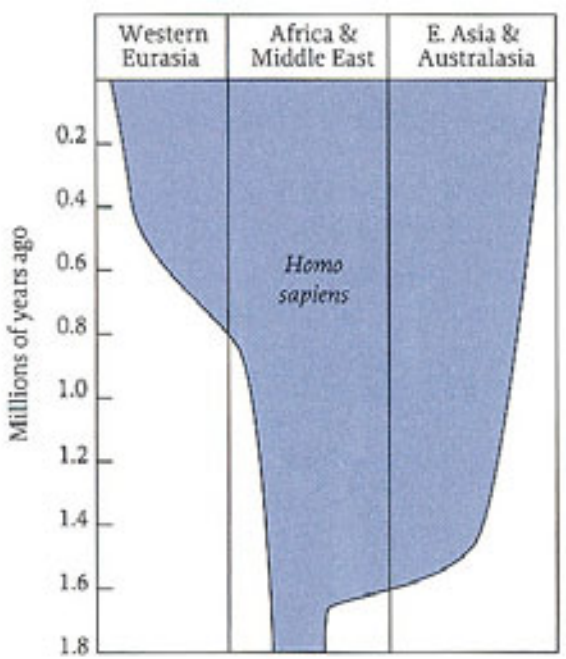
- Alleles differ in frequency between people and populations, genes don't differ in frequency
  - Gene – DNA sequence that encodes a protein
  - Allele – one of several alternative forms of a DNA sequence (can be coding or non-coding)



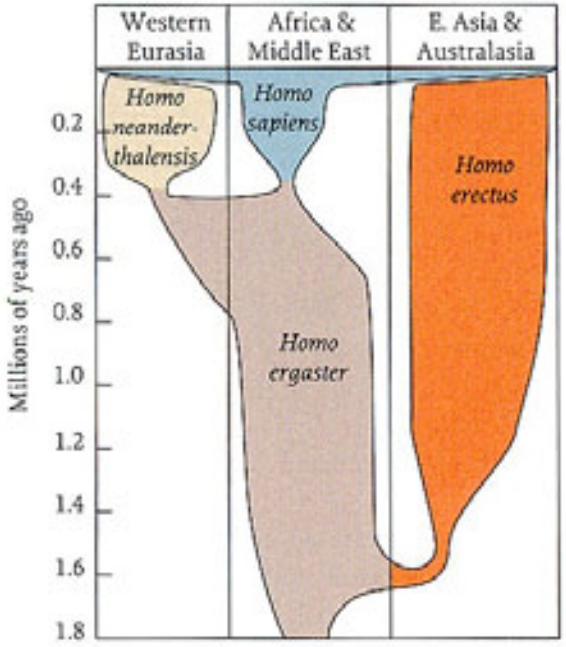


Human paths over the world

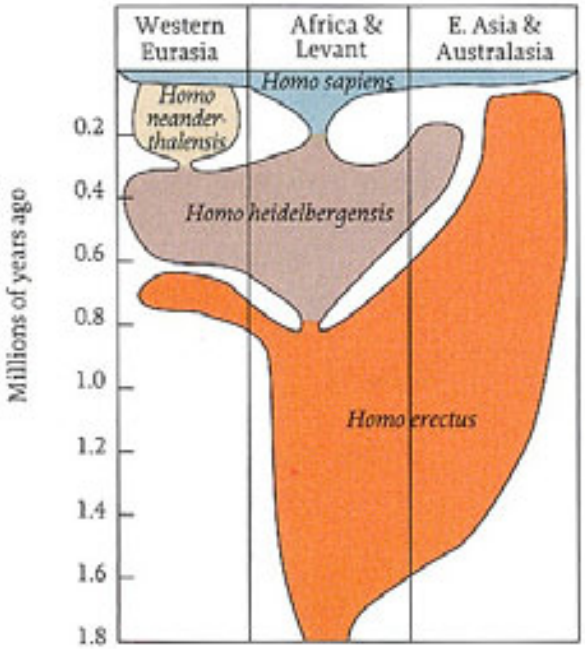
Hominid phylogenies



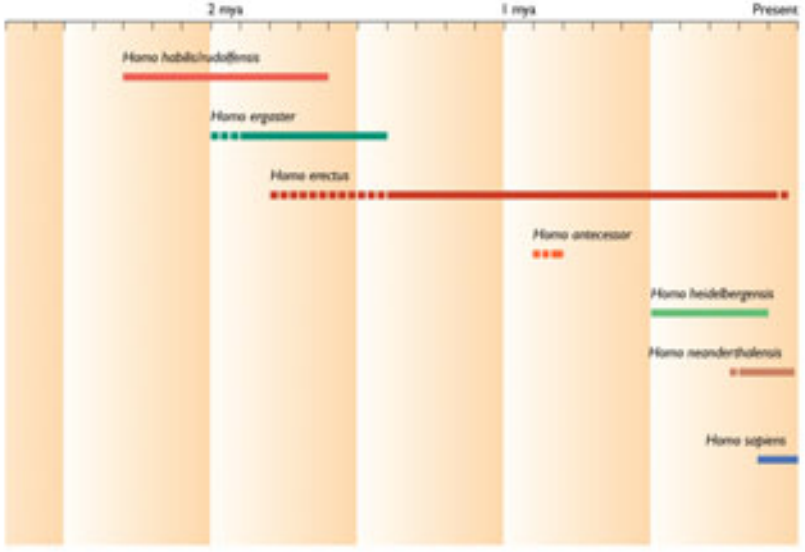
Milford Wolpoff



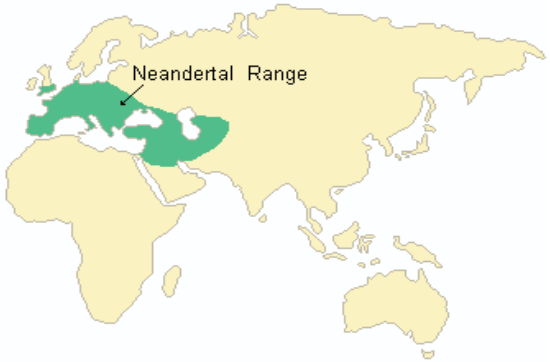
G. Phillip Rightmire



Richard Klein

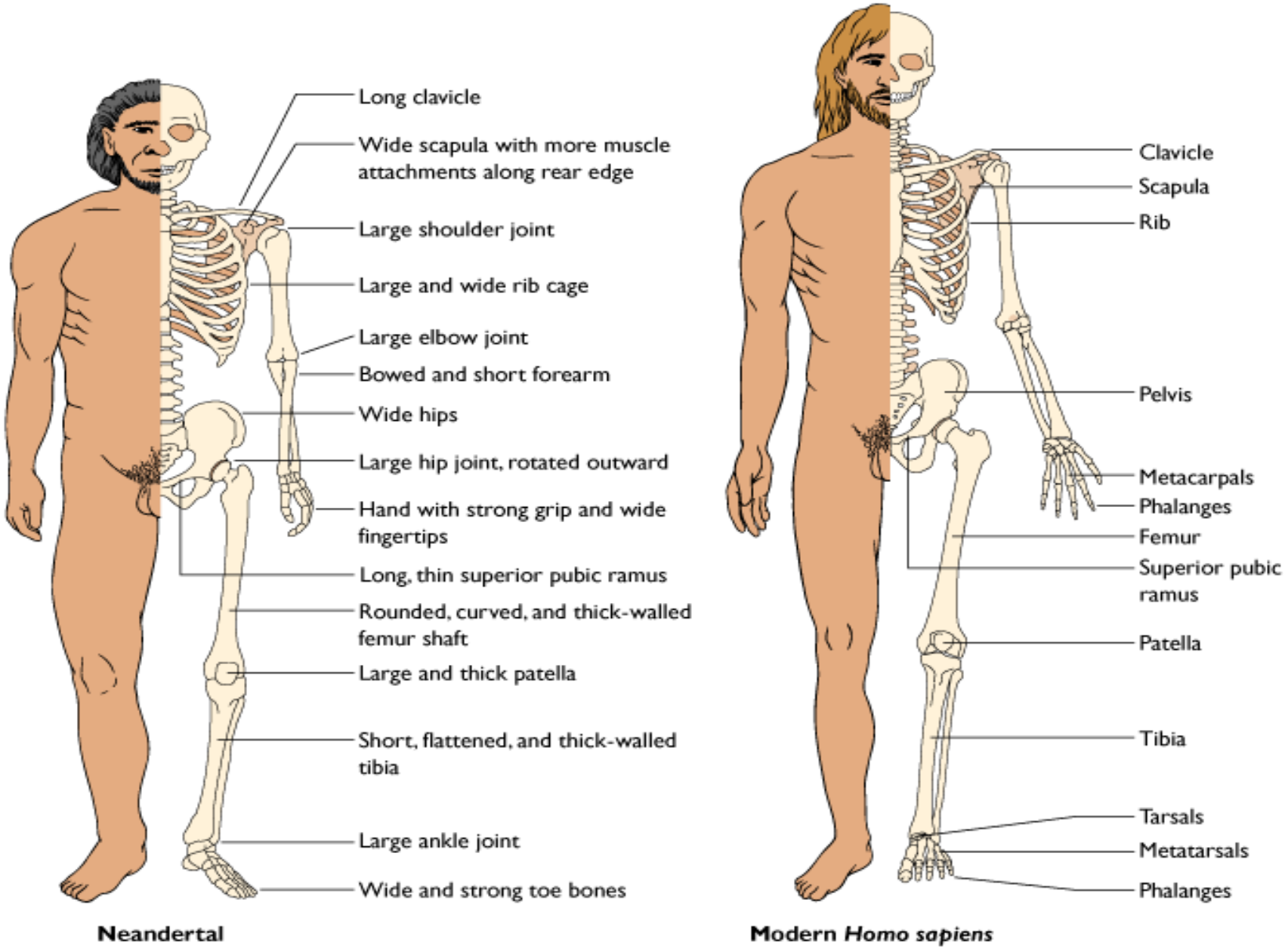


Neanderthals



- Most closely related hominid group to modern humans
- First recognized 150 years ago in Germany
- Existed 500,000 years ago
- Lived in Europe and West Asia
- Evolved away from humans
- Disappeared 30,000 years ago

http://www.msnbc.msn.com/id/13154583/



Arne Elofsson
(arne@bioinfo.se)

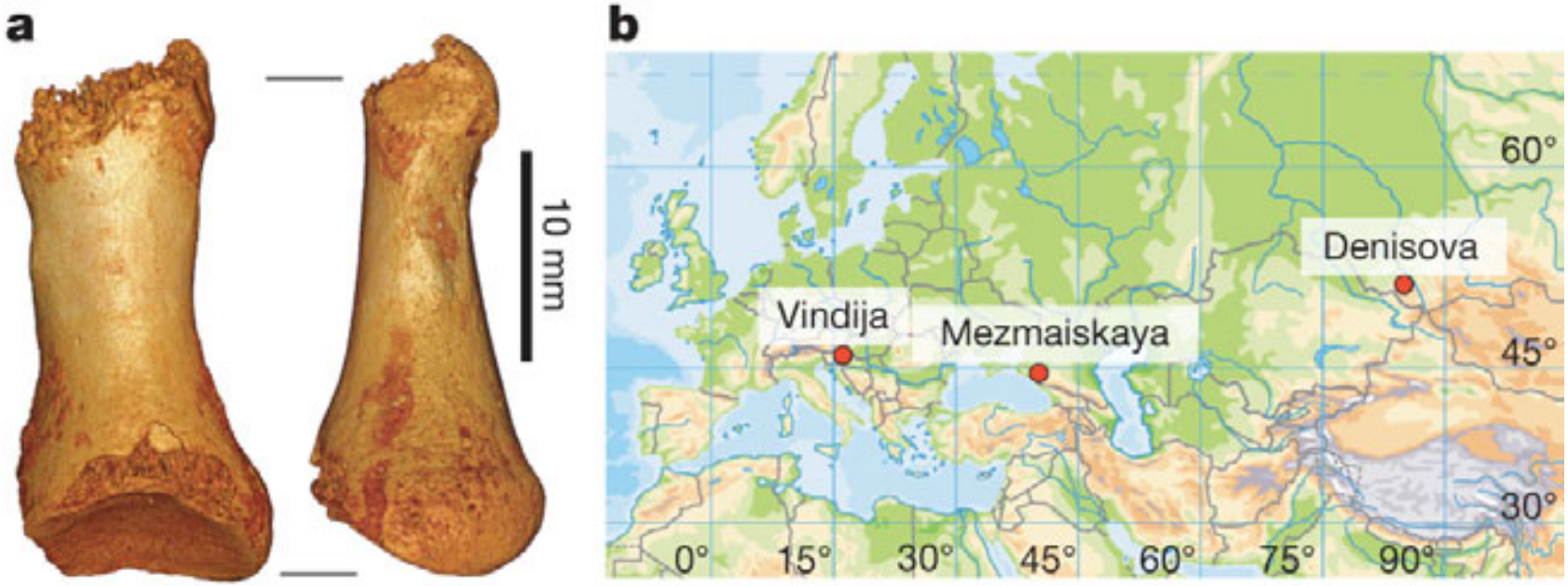
Neanderthal vs. Cro Magnon

- Are Europeans descended purely from Cro Magnons? Pure Neanderthals? Or mixed?

Neanderthal

Cro Magnon

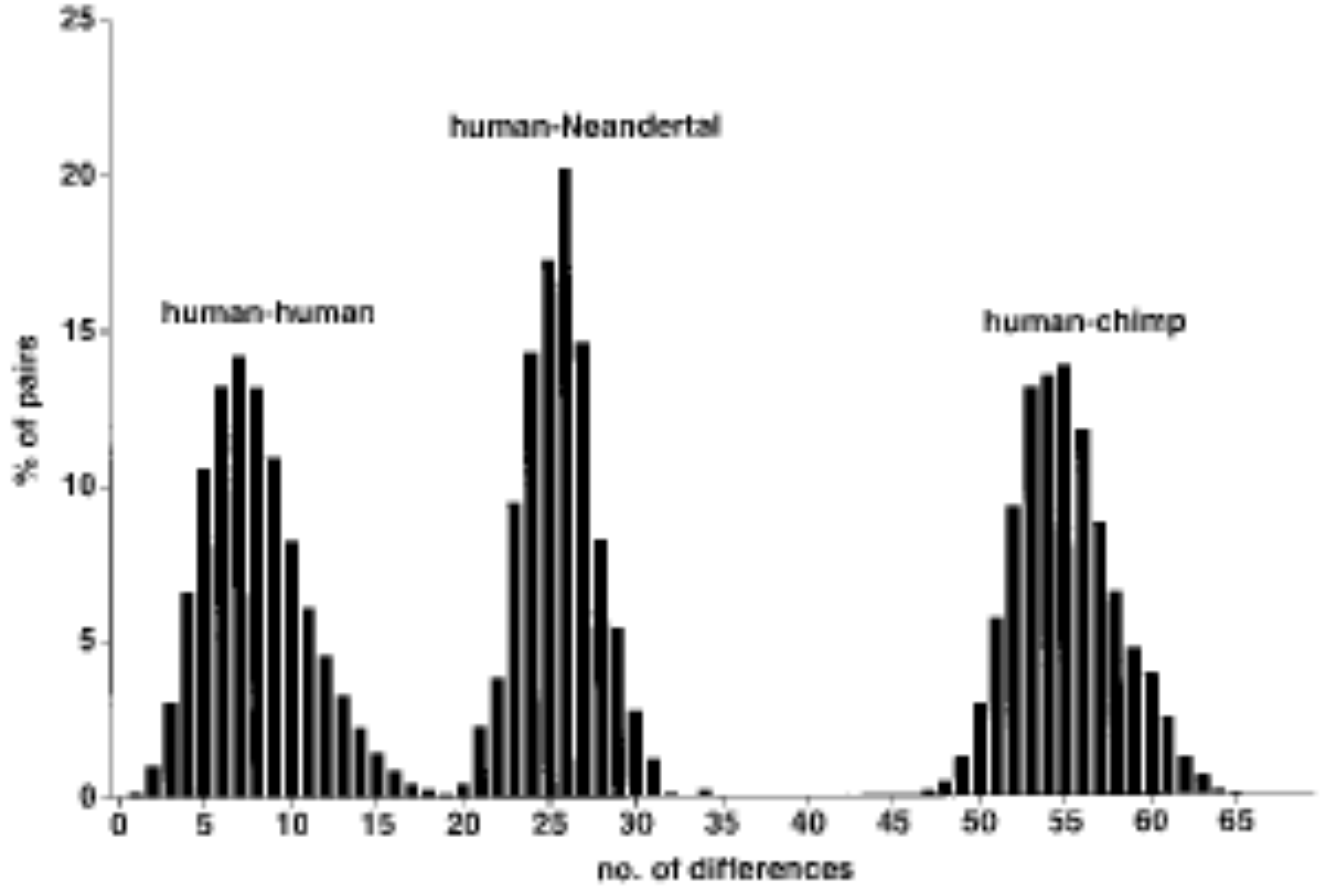
Toe phalanx and location of Neanderthal samples for which genome-wide data are available.



K Prüfer *et al. Nature* **000**, 1-7 (2013) doi:10.1038/nature12886

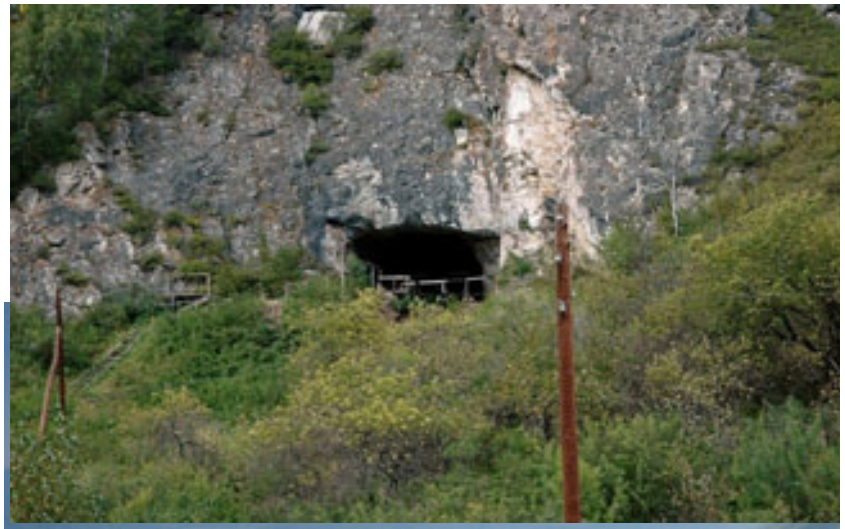
nature

Modern human mtDNA is distinct from Neanderthal mtDNA

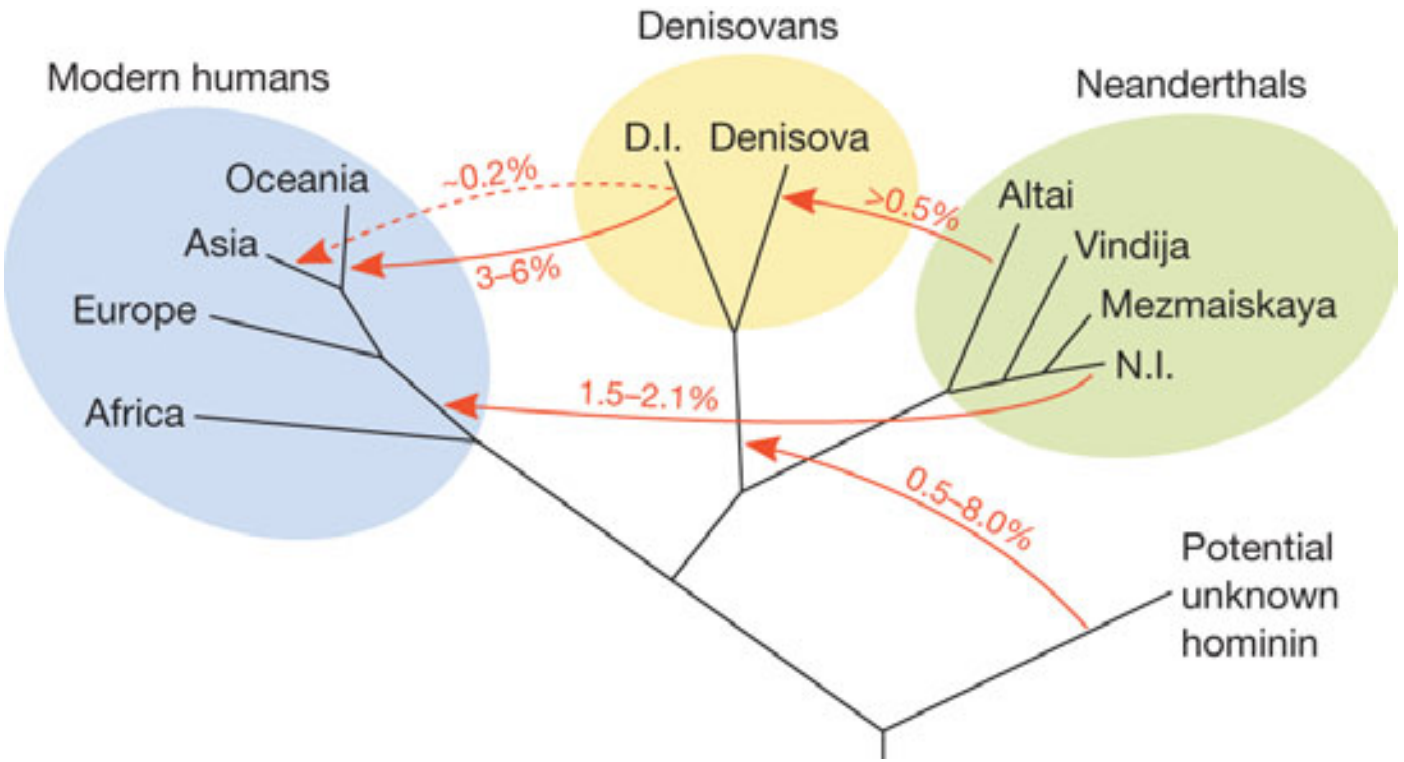


Denisova

- In 2008, a hominin bone fragment was discovered in the Denisova cave in southern Siberia's Altai mountains
  - Child's finger bone scattered among stone tools and bone implements in layer dated to 48-30 kya
- It was believed that modern humans and Neanderthals were the only hominins present there at the time



A possible model of gene flow events in the Late Pleistocene.

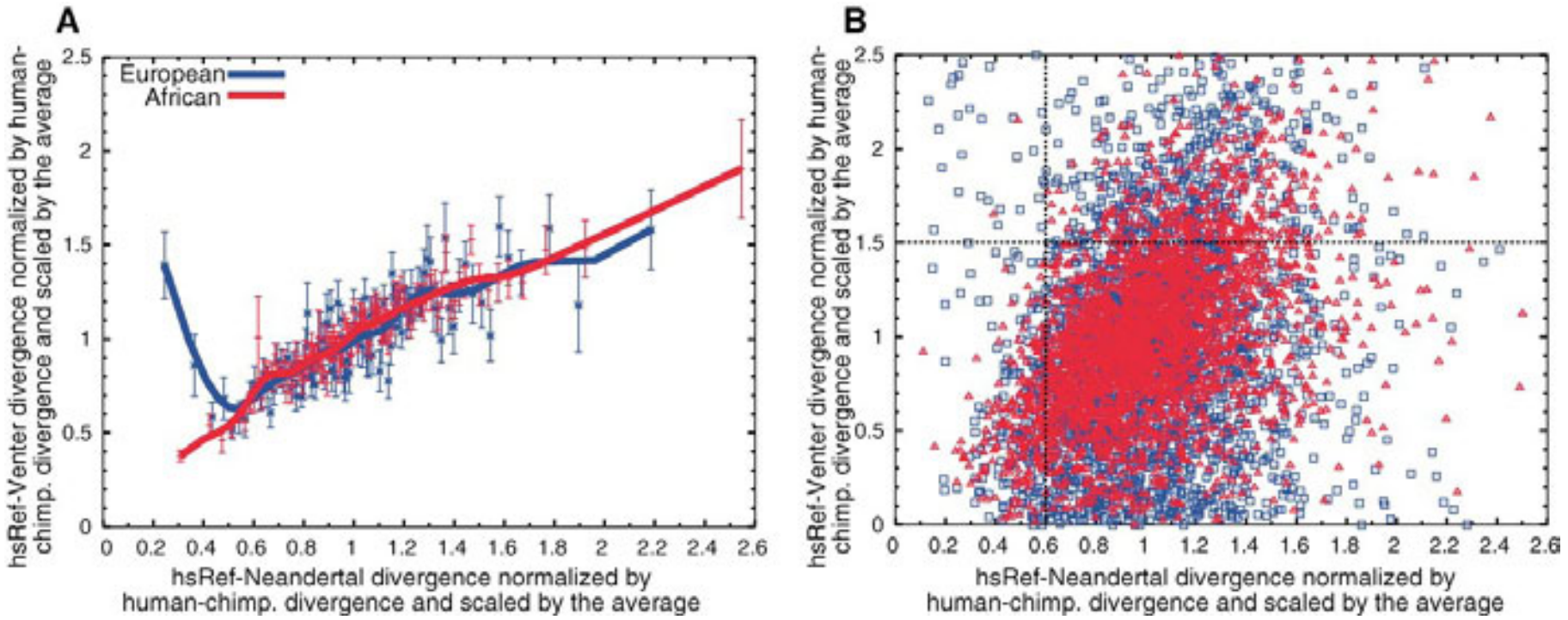


K Prüfer *et al. Nature* **000**, 1-7 (2013) doi:10.1038/nature12886

What about nuclear DNA?
Represents 3 billion bases, not 16 thousand

- Green et al. (2010) published draft nuclear genome sequences of 3 Neanderthals
  - Posits 1-4% Neanderthal admixture in Europeans and Asians
  - Since it occurred in both Europeans and Asians, likely to have occurred before those groups split, i.e. ~50-80 kya
  - No Neanderthal DNA in Africans
  - Expected difference between mitochondrial and nuclear DNA

Segments of Neanderthal ancestry in the human reference genome.

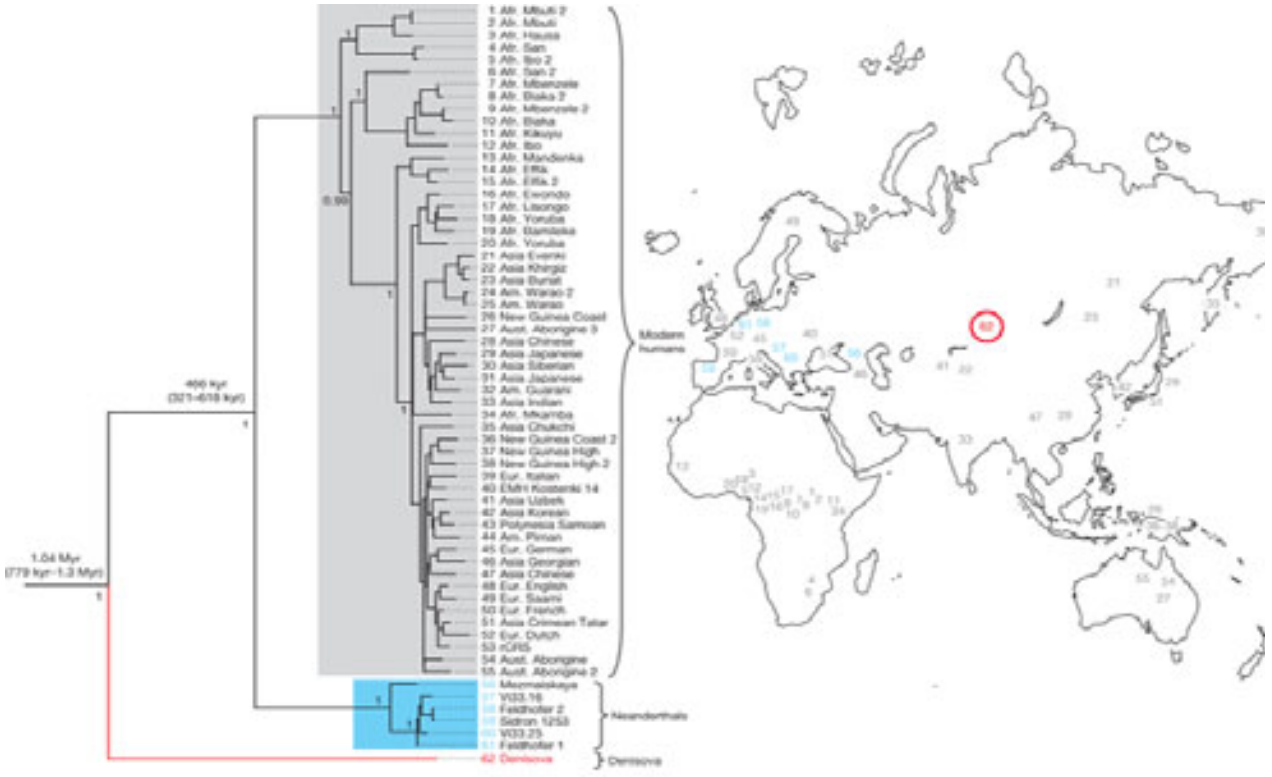


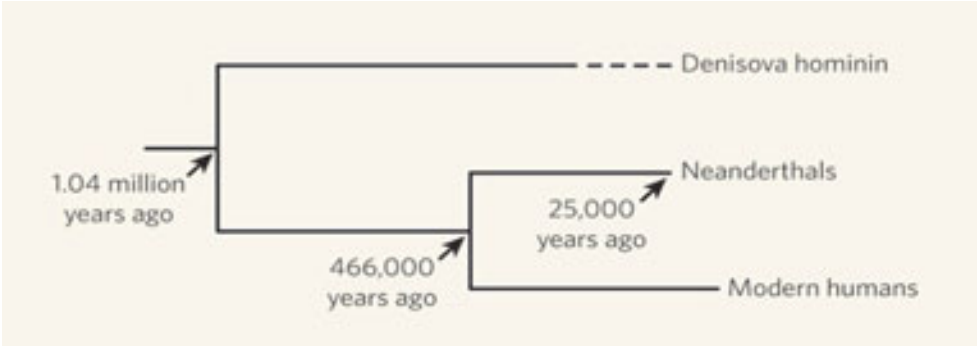
R E Green et al. Science 2010;328:710-722

Published by AAAS

Phylogenetic analysis

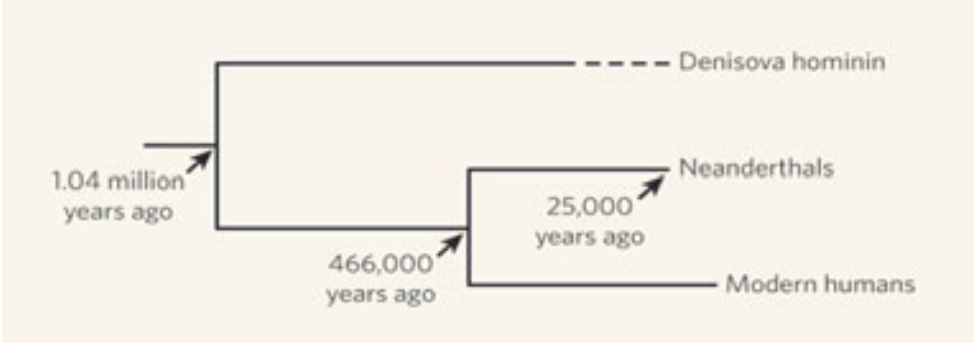
- Denisova mtDNA lineage branches much earlier than human and Neanderthal lineages
  - Most recent common ancestor (MRCA) between humans and Denisovans is ~1mya
  - MRCA is twice as old as MRCA of humans and Neanderthals

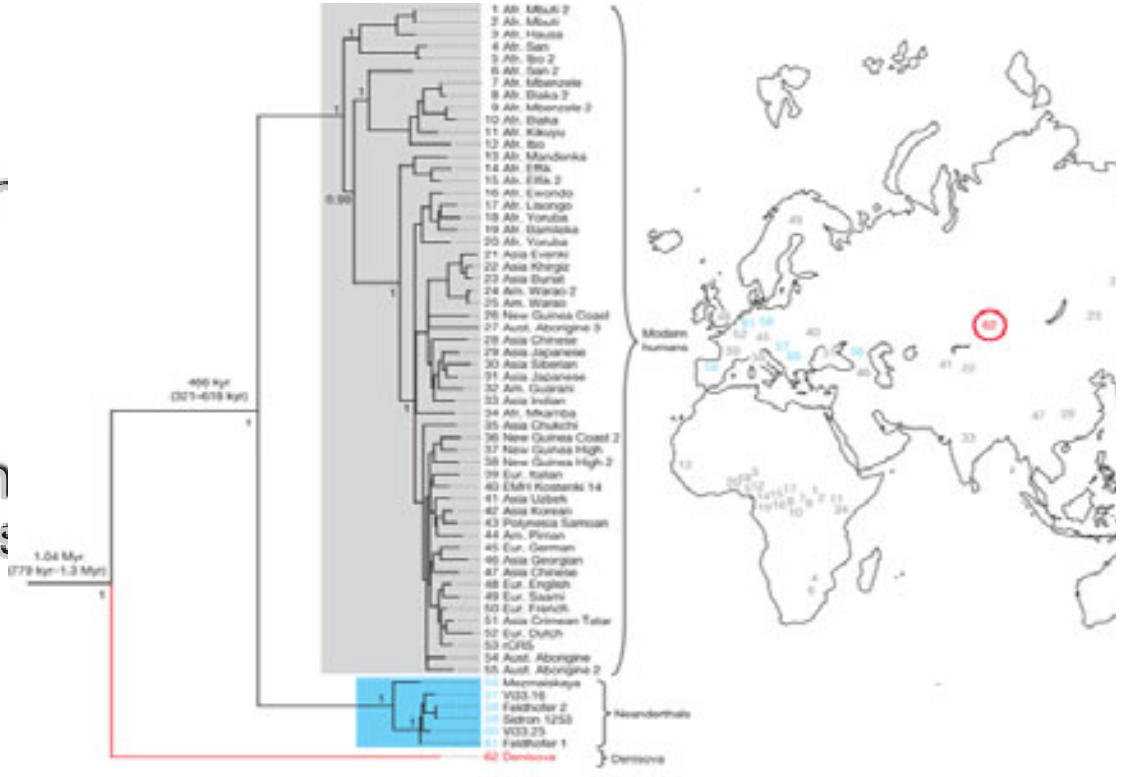




Phylogenetic analysis

- Denisova mtDNA lineage branches much earlier than human and Neanderthal lineages
  - Most recent common ancestor (MRCA) between humans and Denisovans is ~1mya
  - MRCA is twice as old as MRCA of humans and Neanderthals



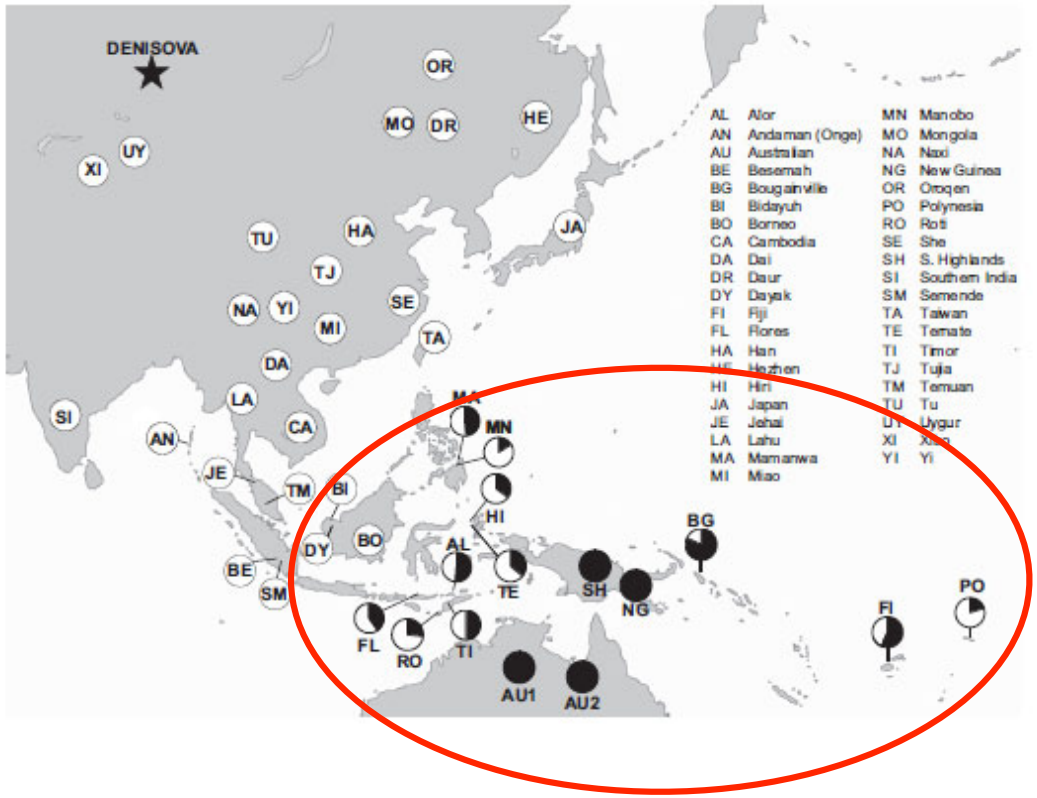


Denisova can't be *H erectus* b/c *H erectus* wasn't in mainland Asia ~40 kya and *H erectus* left Africa ~2 mya

- Denisova must have been in Africa ~2 mya to share a common ancestor with modern humans and Neanderthal

What about nuclear DNA?

- Changes phylogeny – Denisovans closer to Neanderthal than modern humans
- 4-6% admixture in Southeast Asians



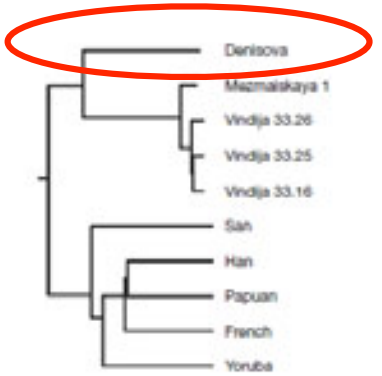
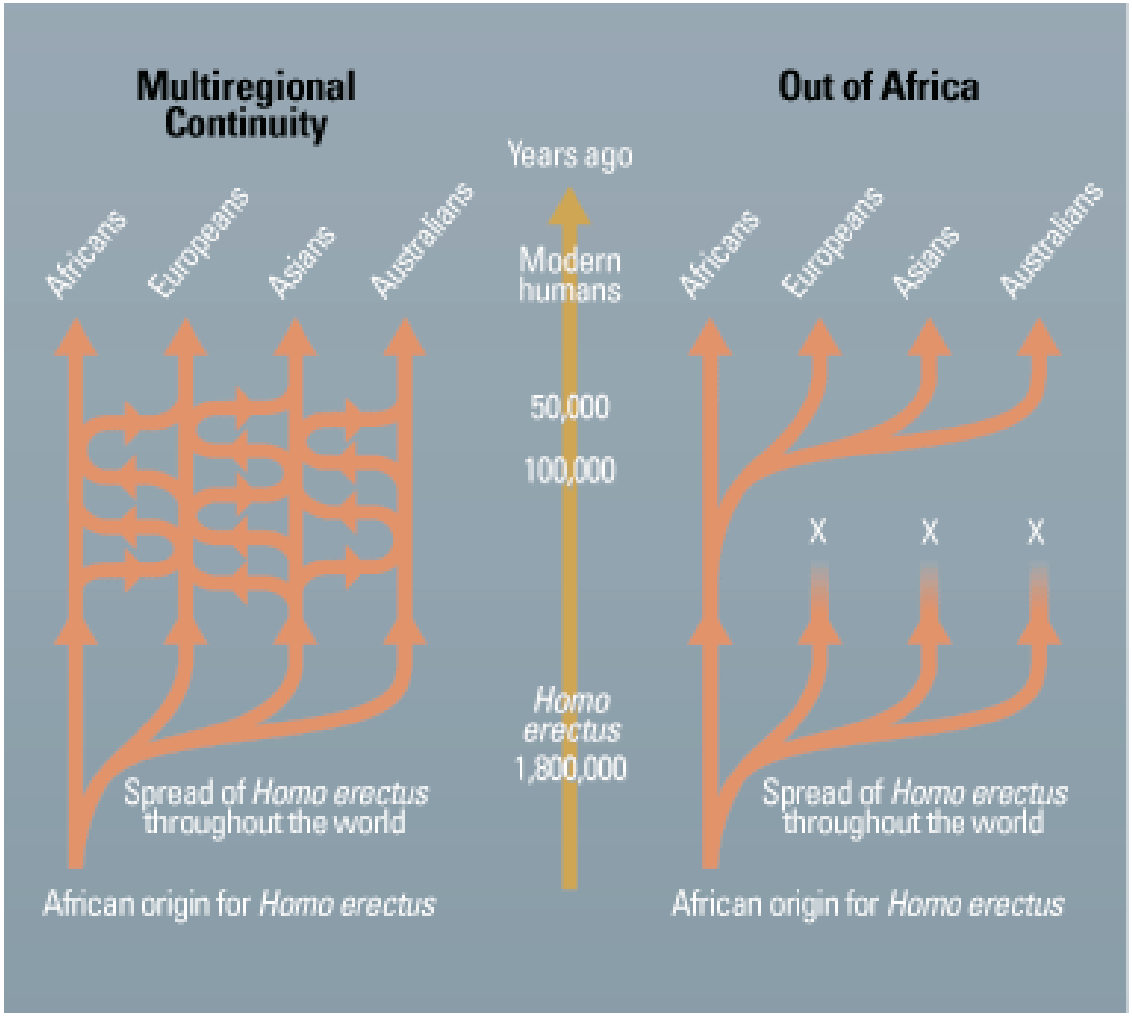
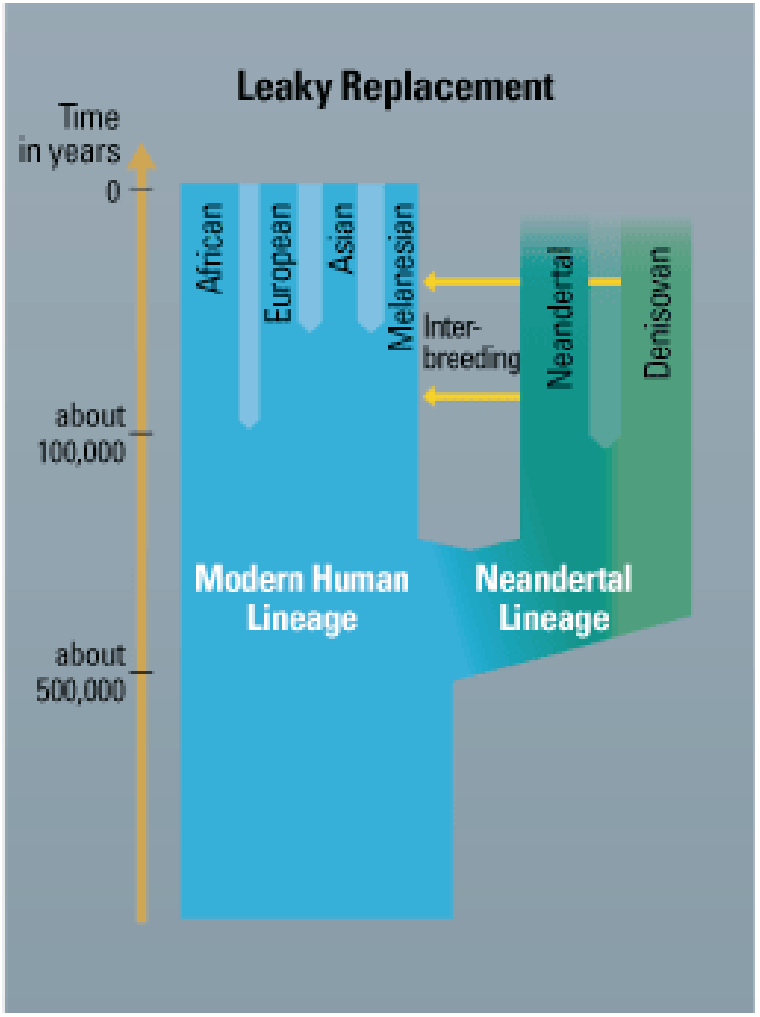


Figure 1 | A neighbour-joining tree based on pairwise autosomal DNA sequence divergences for five ancient and five present-day hominins. Vindija 33.16, Vindija 33.25 and Vindija 33.26 refer to the catalogue numbers of the Neanderthal bones.





Arne Elofsson
(arne@bioinfo.se)

Major paradigm shift

We do have Neanderthal DNA in us!

And Denisovan, another archaic hominin

Recent papers propose admixture from possibly two more, unidentified archaic hominins

Allele in a gene in our immune system has recently been identified as coming from Neanderthals and conferring a selective advantage (Mendez et al. AJHG, 2012)