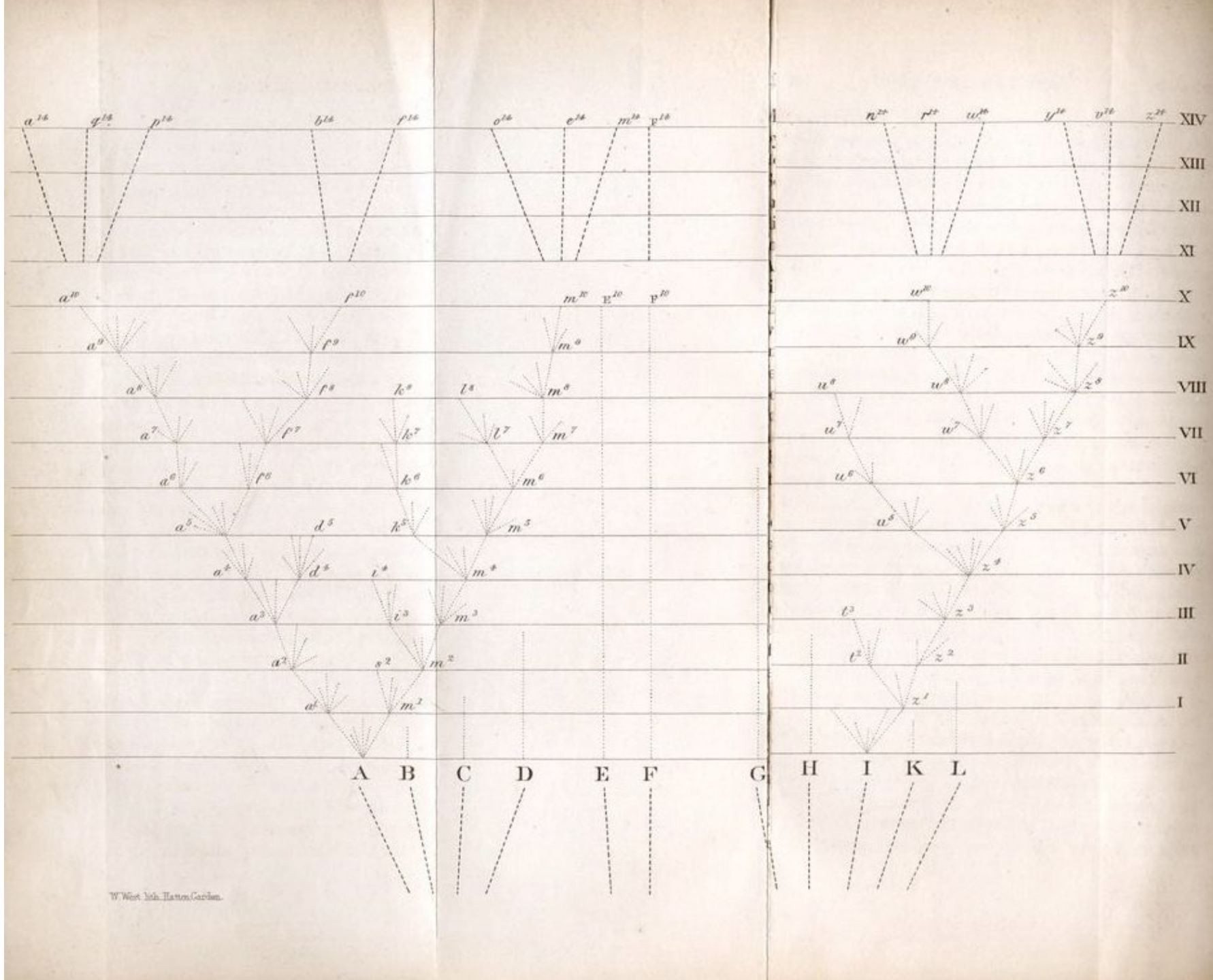
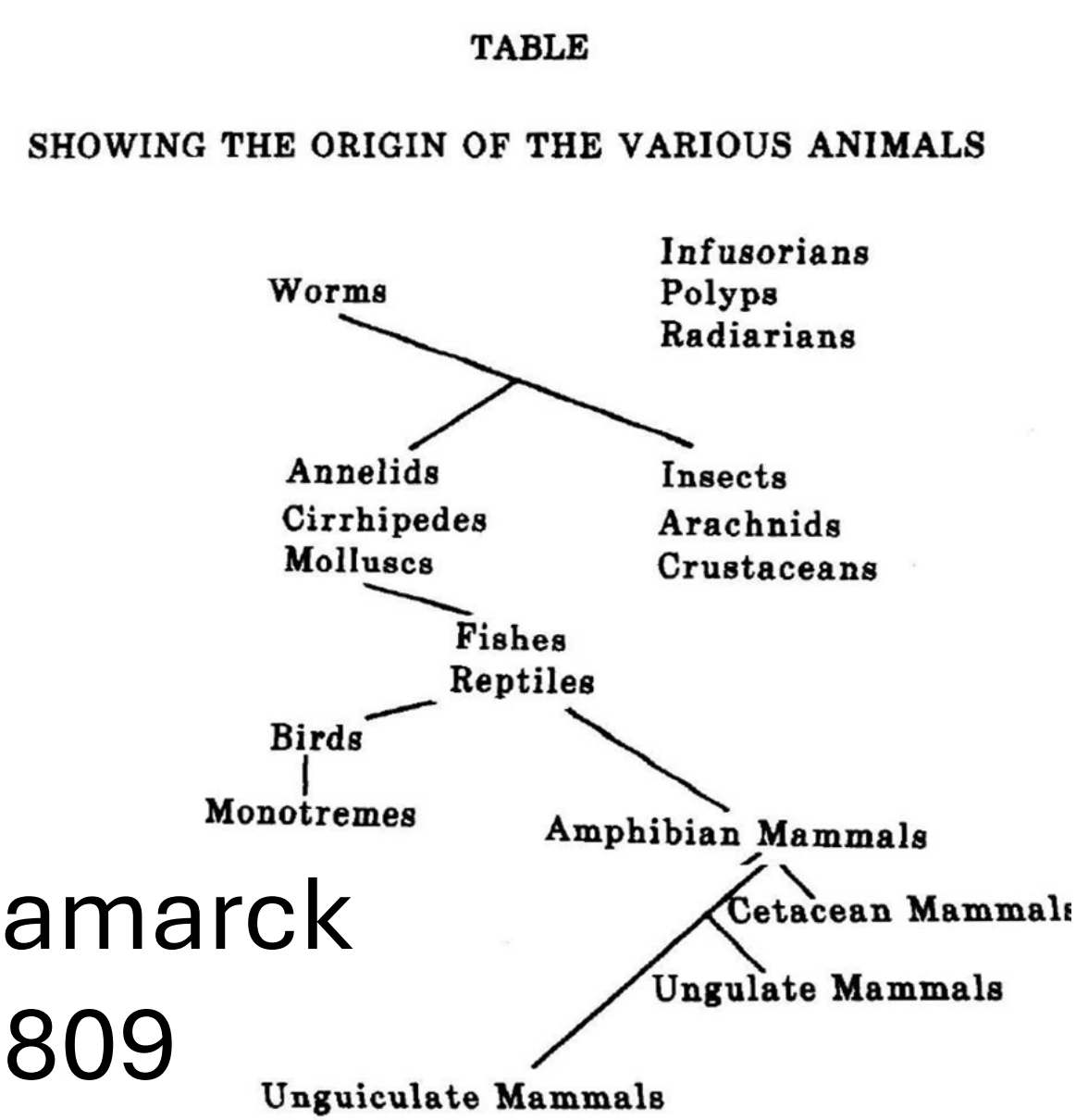
# Phylogenetic trees

## Methods and interpretation

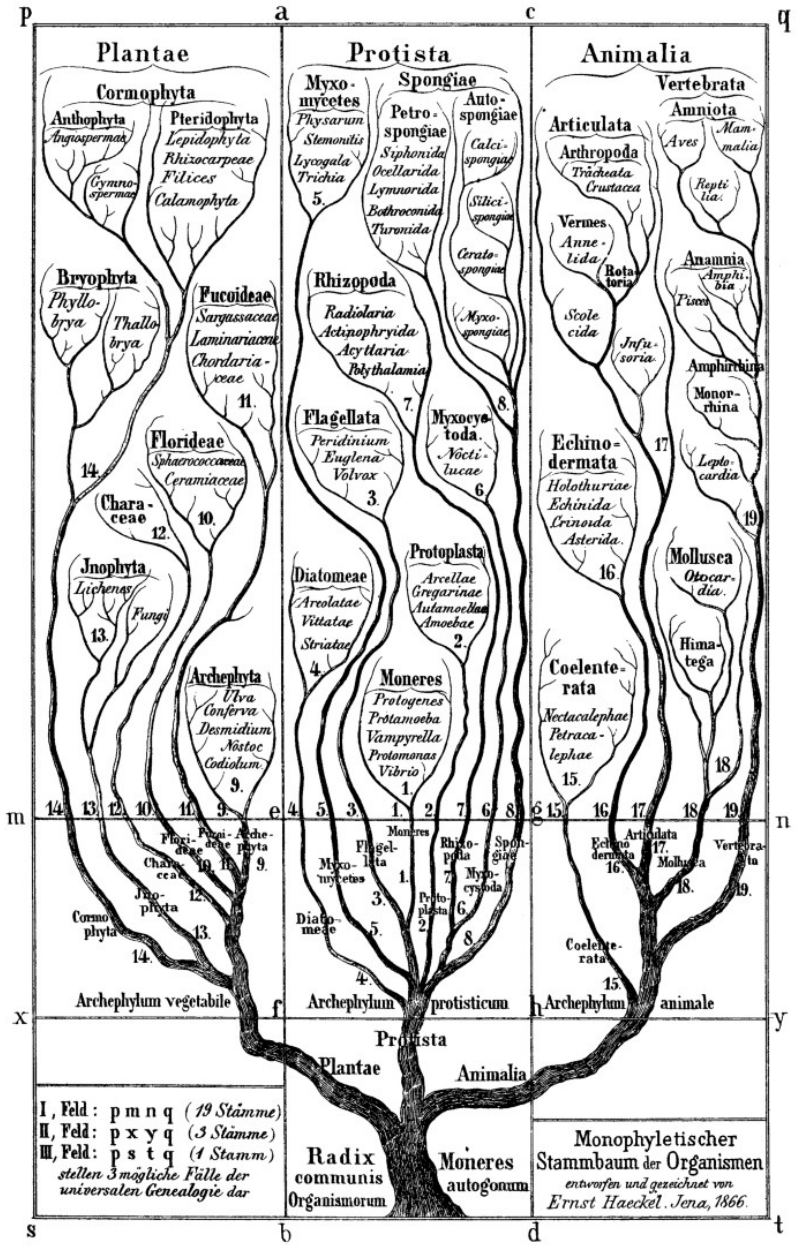
Darwin's one diagram, "The Tree of Life"



# Scientists were drawing trees before and after Darwin

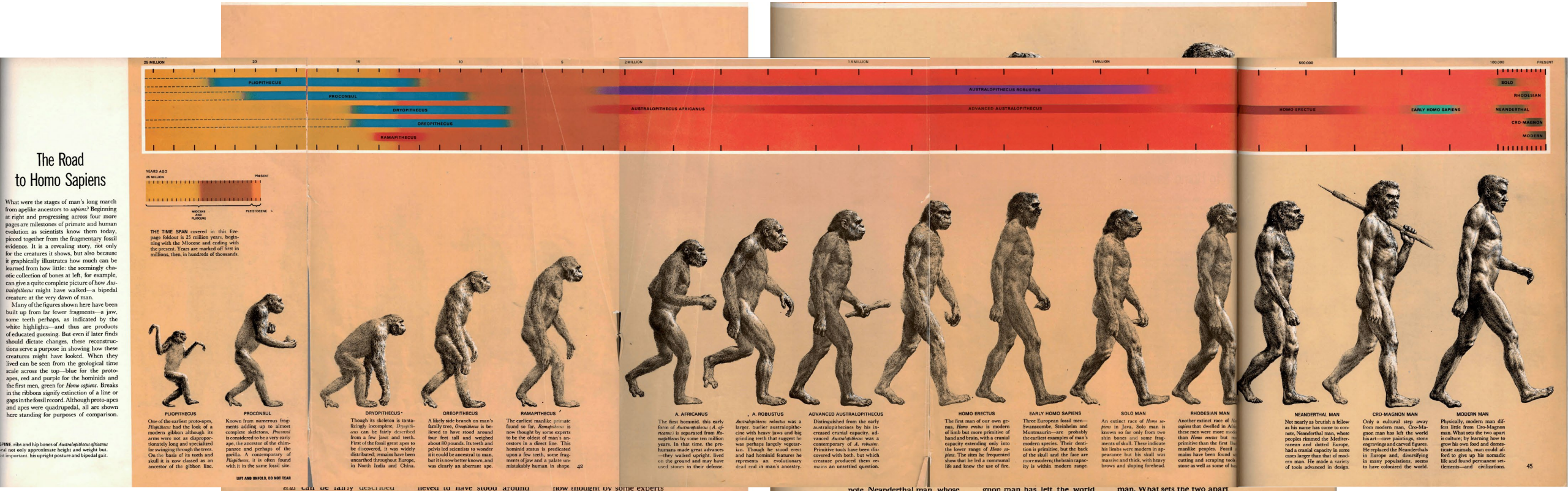



Lamarck
1809




Haeckel
1866

# However, many myths about “progress”




Actually, the author and the artist wanted to show fossil hominins in a folded-up time sequence, and some implications of progress were accidental!

Image by Rudolph Zallinger, for
Howell (1965) *Early Man*. Time-Life Books

# Meanwhile in flowering plants … (Stebbins 1974 system)


G. Ledyard Stebbins


Vernon H. Heywood

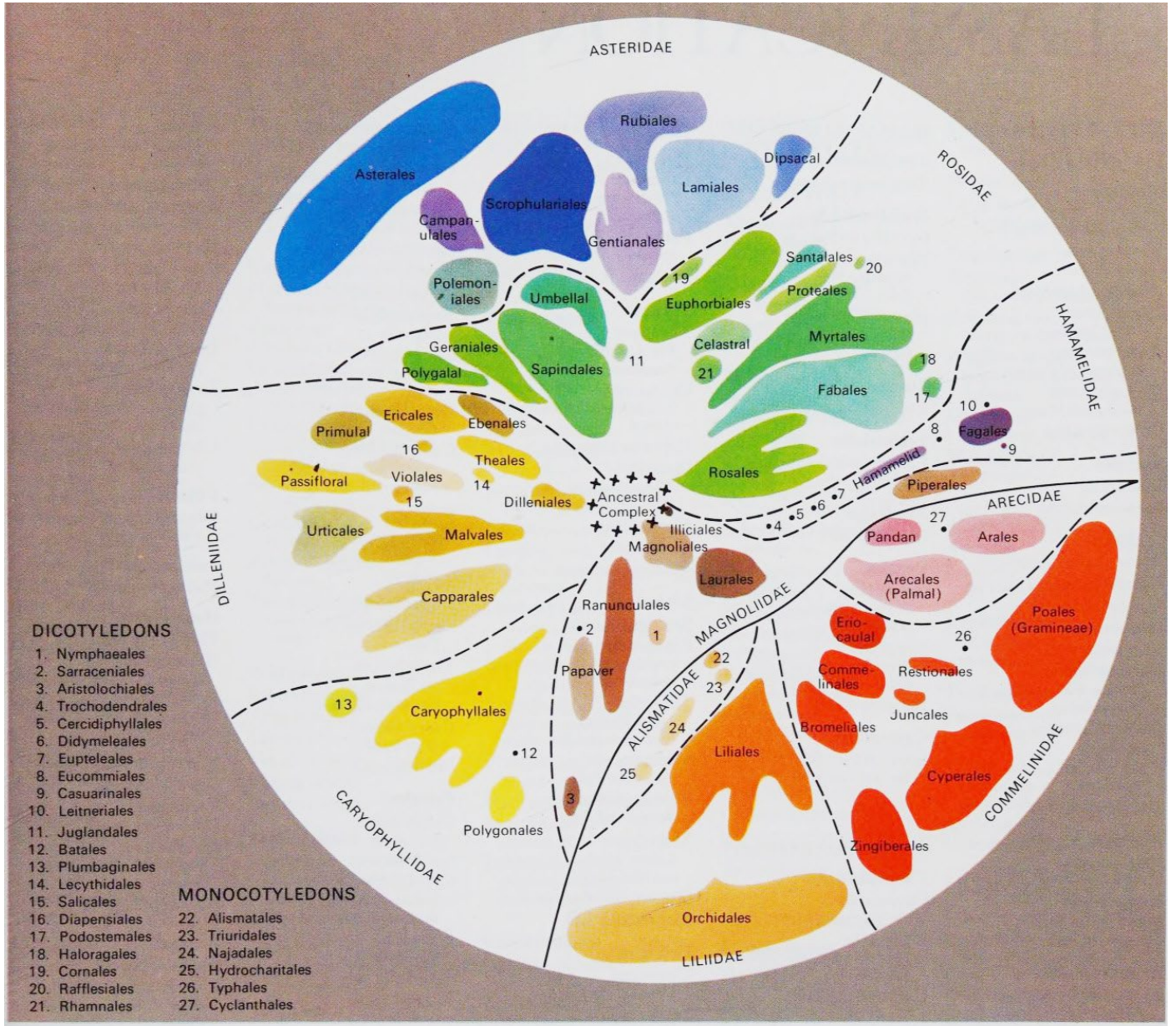


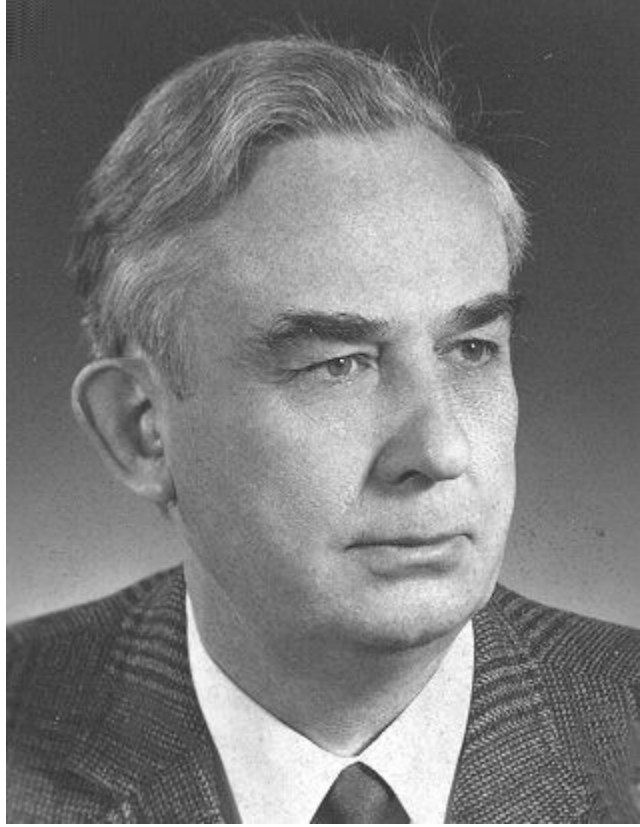
Plate III. Evolutionary diagram showing the relative degree of specialization of the orders of Angiosperms (after Stebbins, 1974).

A “cabbage” phylogeny!

from Heywood 1978

# The cladistics revolution

- 1950: Willi Hennig, a dipterist (taxonomist of flies) publishes the book *Grundzüge einer Theorie der phylogenetischen Systematik* in German
- Largely ignored by English-speaking biologists
- 1961: Warren H. Wagner promotes principles of phylogenetics based on optimal combinations of ancestral (0) and derived (1) character states (later called parsimony methods)



Willi Hennig

- 1966: A (bad) English translation of Hennig's book published in the USA: Phylogenetic Systematics
- 1960s-1980s: Conflict among scientists as to methods in systematics: phenetic similarity, “evolutionary” systematics, and cladistics

# 1960s-1980s cladistic revolution

- Phenetic classification: overall similarity
  Might lead to “paraphyly”
- Cladistic classification: use shared derived characters only
  Only monophyletic groups are considered real (following Hennig)

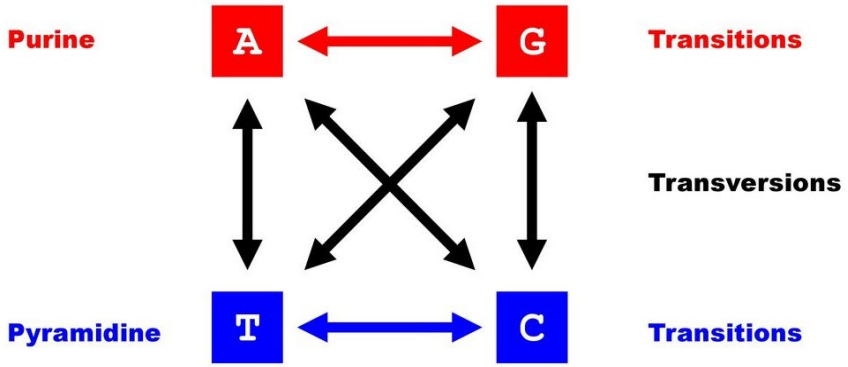
**Methods of tree construction**

- Use of overall similarity – distance based trees
- Parsimony – minimize numbers of character changes
  “Pattern cladists” prefer the most parsimonious tree, even if not true phylogeny!
- Likelihood/Bayesian – model evolution of characters (e.g. DNA)

# Modelling base change: The HKY 85 model of DNA evolution

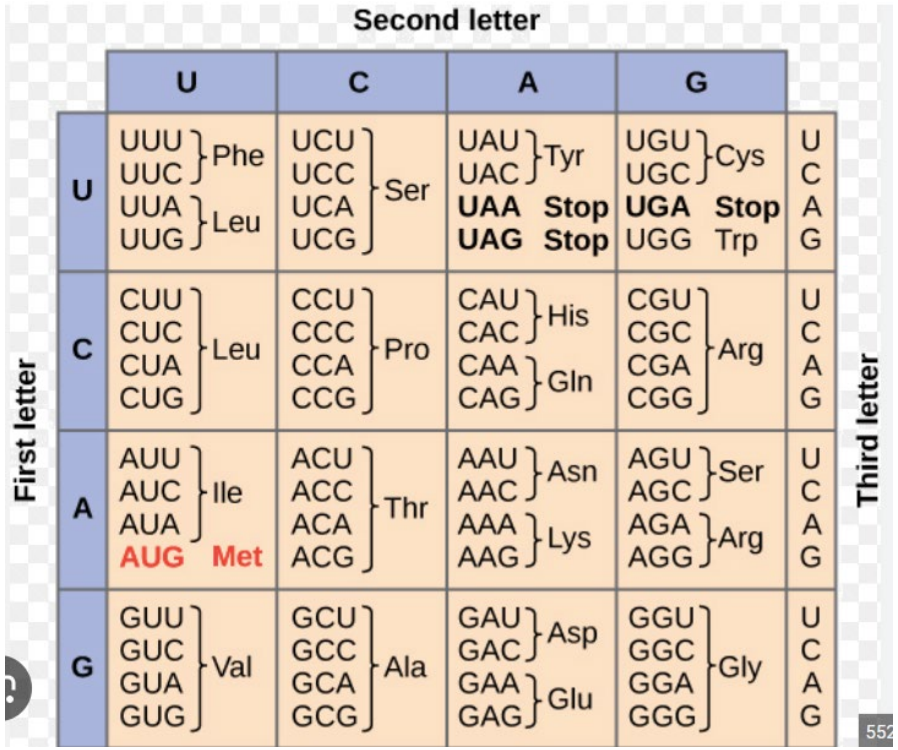
$$\text{Rate matrix } Q = \begin{array}{c} \\ A \\ G \\ C \\ T \end{array} \begin{array}{c} \begin{array}{cccc} A & G & C & T \end{array} \\ \begin{pmatrix} * & \kappa\pi_G & \pi_C & \pi_T \\ \kappa\pi_A & * & \pi_C & \pi_T \\ \pi_A & \pi_G & * & \kappa\pi_T \\ \pi_A & \pi_G & \kappa\pi_C & * \end{pmatrix} \end{array}$$

rates of
transitions > transversions

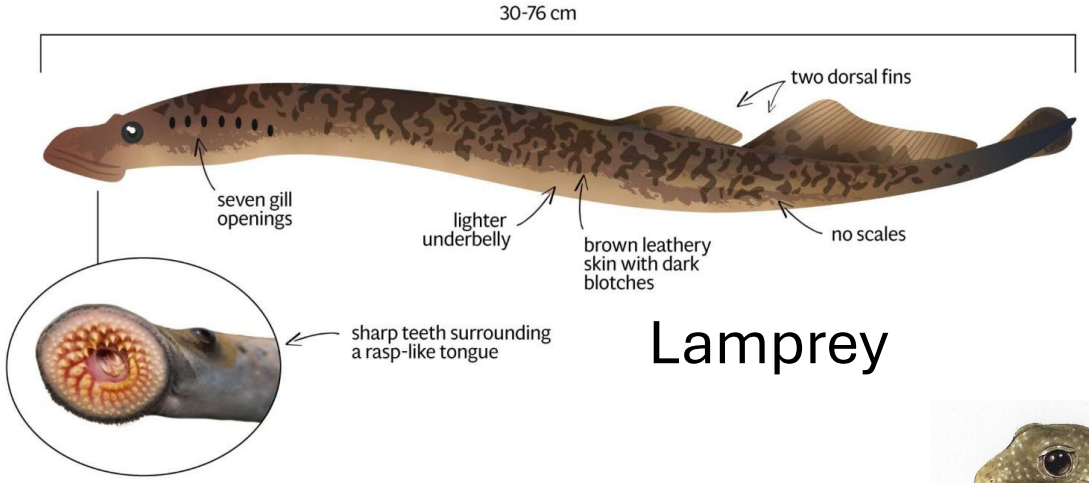



| | | | | | | | | | | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Ala | 4 | | | | | | | | | | | | | | | | | | | |
| Arg | -1 | 5 | | | | | | | | | | | | | | | | | | |
| Asn | -2 | 0 | 6 | | | | | | | | | | | | | | | | | |
| Asp | -2 | -2 | 1 | 6 | | | | | | | | | | | | | | | | |
| Cys | 0 | -3 | -3 | -3 | 9 | | | | | | | | | | | | | | | |
| Gln | -1 | 1 | 0 | 0 | -3 | 5 | | | | | | | | | | | | | | |
| Glu | -1 | 0 | 0 | 2 | -4 | 2 | 5 | | | | | | | | | | | | | |
| Gly | 0 | -2 | 0 | -1 | -3 | -2 | -2 | 6 | | | | | | | | | | | | |
| His | -2 | 0 | 1 | -1 | -3 | 0 | 0 | -2 | 8 | | | | | | | | | | | |
| Ile | -1 | -3 | -3 | -3 | -1 | -3 | -3 | -4 | -3 | 4 | | | | | | | | | | |
| Leu | -1 | -2 | -3 | -4 | -1 | -2 | -3 | -4 | -3 | 2 | 4 | | | | | | | | | |
| Lys | -1 | 2 | 0 | -1 | -3 | 1 | 1 | -2 | -1 | -3 | -2 | 5 | | | | | | | | |
| Met | -1 | -1 | -2 | -3 | -1 | 0 | -2 | -3 | -2 | 1 | 2 | -1 | 5 | | | | | | | |
| Phe | -2 | -3 | -3 | -3 | -2 | -3 | -3 | -3 | -1 | 0 | 0 | -3 | 0 | 6 | | | | | | |
| Pro | -1 | -2 | -2 | -1 | -3 | -1 | -1 | -2 | -2 | -3 | -3 | -1 | -2 | -4 | 7 | | | | | |
| Ser | 1 | -1 | 1 | 0 | -1 | 0 | 0 | 0 | -1 | -2 | -2 | 0 | -1 | -2 | -1 | 4 | | | | |
| Thr | 0 | -1 | 0 | -1 | -1 | -1 | -1 | -2 | -2 | -1 | -1 | -1 | -1 | -2 | -1 | 1 | 5 | | | |
| Trp | -3 | -3 | -4 | -4 | -2 | -2 | -3 | -2 | -2 | -3 | -2 | -3 | -1 | 1 | -4 | -3 | -2 | 11 | | |
| Tyr | -2 | -2 | -2 | -3 | -2 | -1 | -2 | -3 | 2 | -1 | -1 | -2 | -1 | 3 | -3 | -2 | -2 | 2 | 7 | |
| Val | 0 | -3 | -3 | -3 | -1 | -2 | -2 | -3 | -3 | 3 | 1 | -2 | 1 | -1 | -2 | -2 | 0 | -3 | -1 | 4 |
| | Ala | Arg | Asn | Asp | Cys | Gln | Glu | Gly | His | Ile | Leu | Lys | Met | Phe | Pro | Ser | Thr | Trp | Tyr | Val |

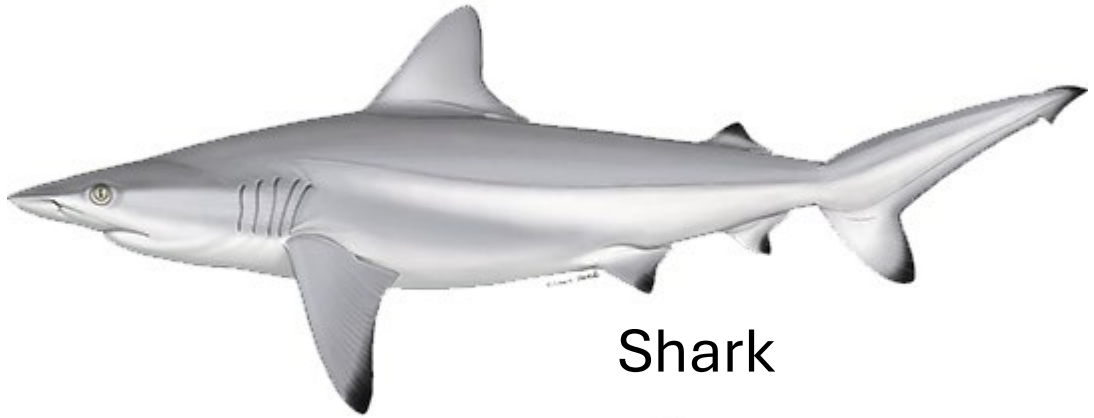
The BLOSUM62 model of amino acid evolution



# Example: classification of vertebrates




Lamprey



Shark

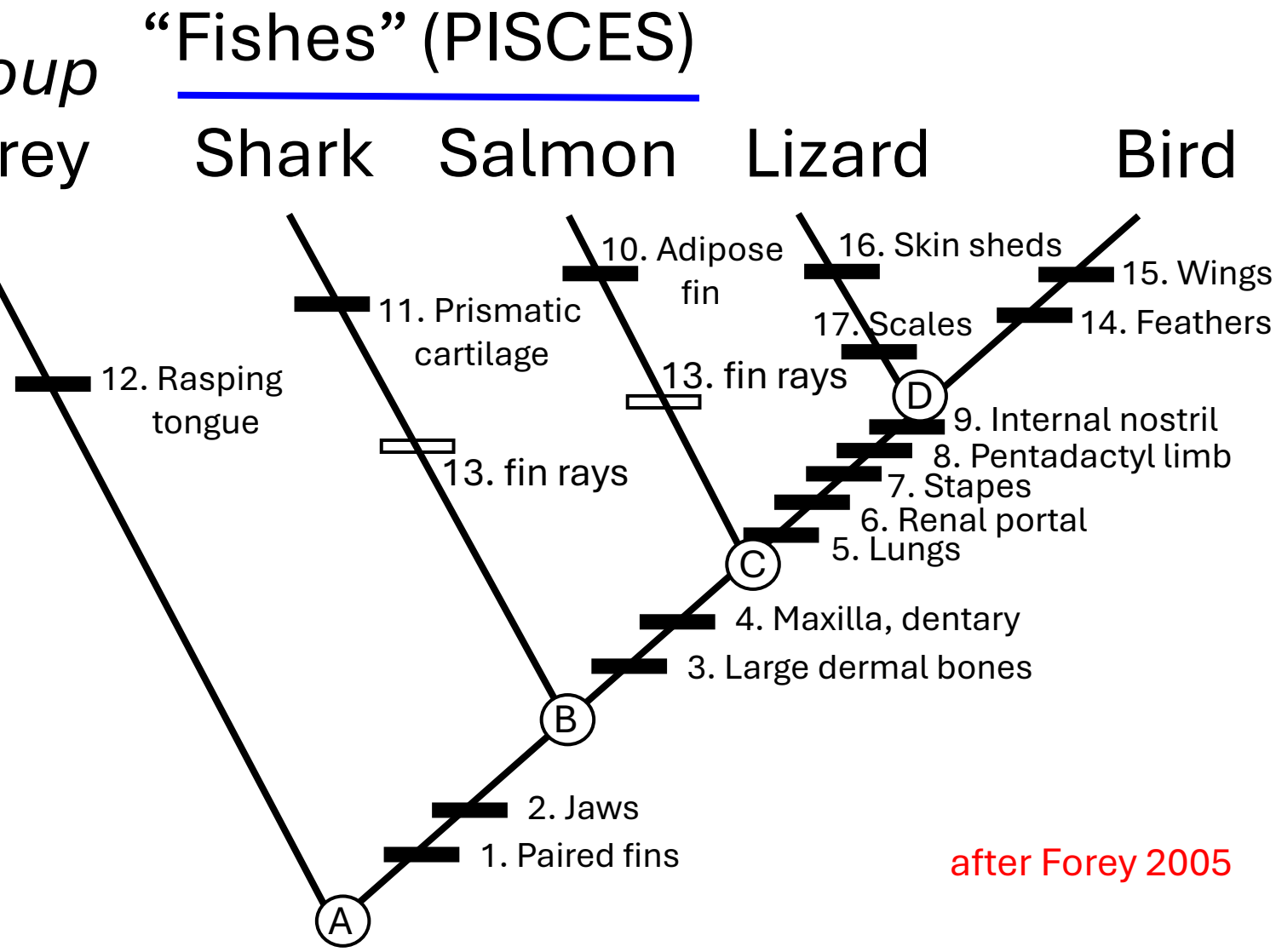

Salmon

Lizard

Bird

# Monophyly vs. paraphyly

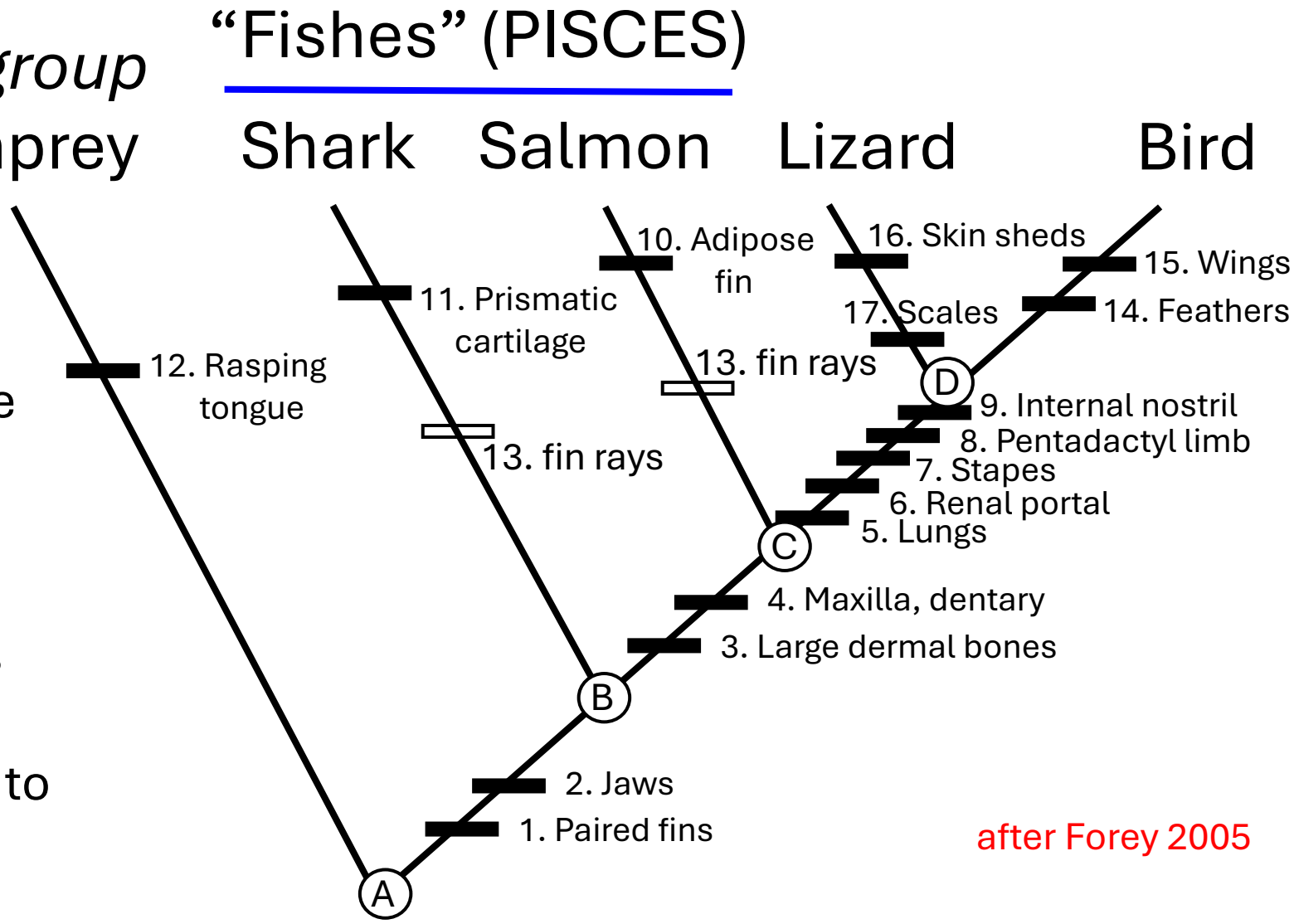
- A cladistic tree
- Jawed vertebrates are MONOphyletic
- “Fishes” are PARAphyletic




after Forey 2005

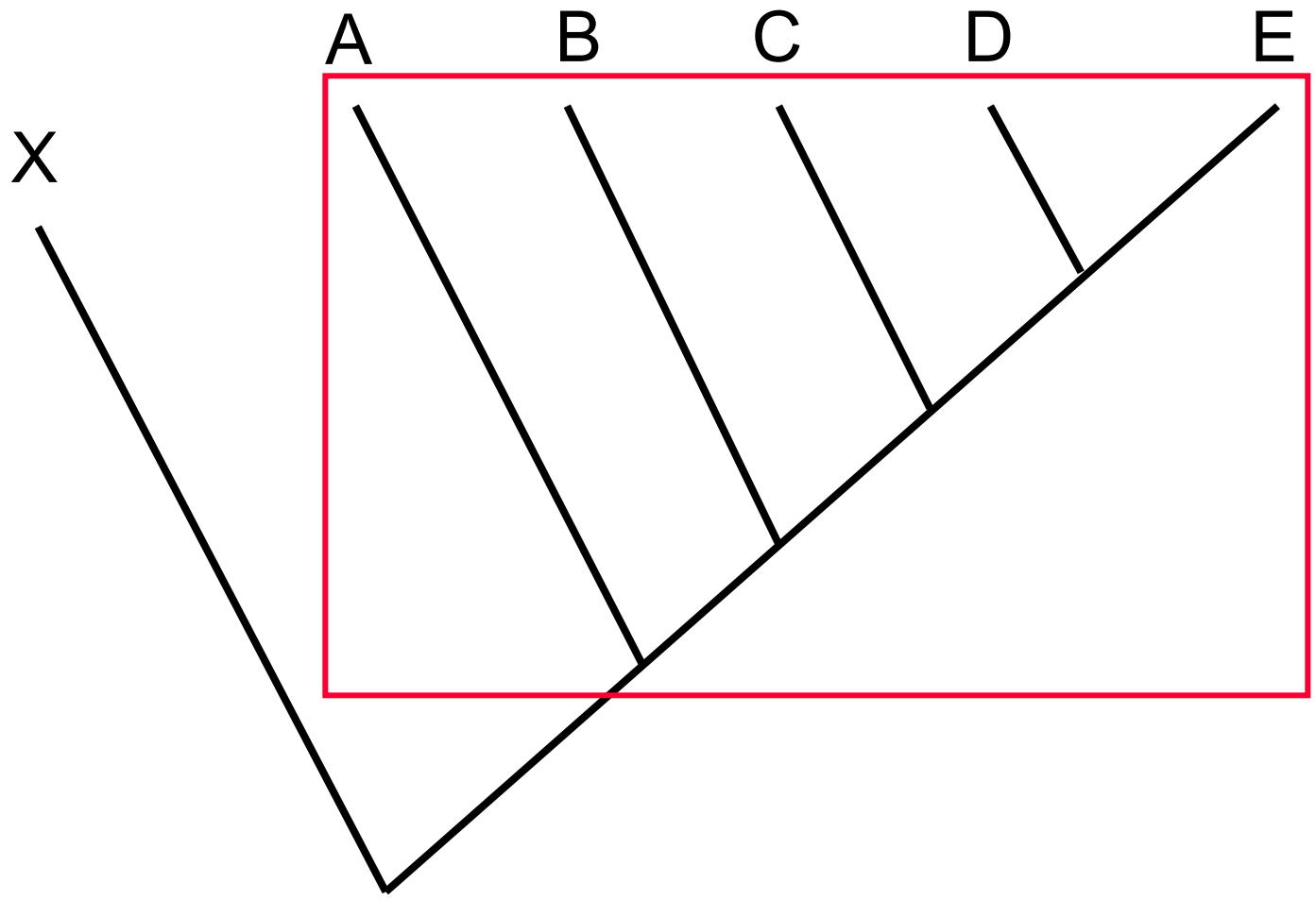
# Parsimony criterion for grouping

- Character 13 would group “fishes” together
- However this would require both characters 3 and 4 to evolve 2x
- The arrangement shown here is more parsimonious (fewer changes required; only character 13 required to change 2x)

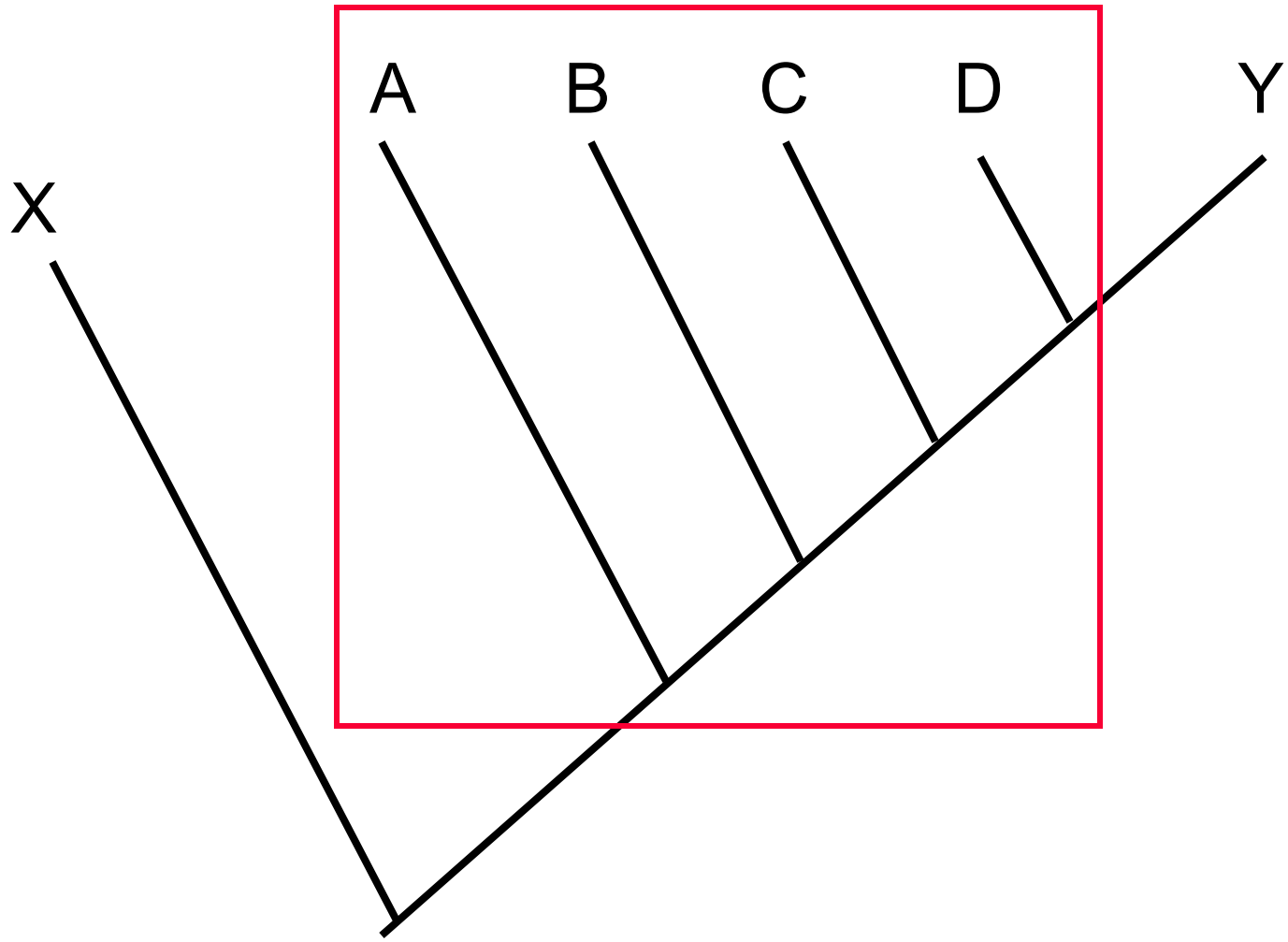



after Forey 2005

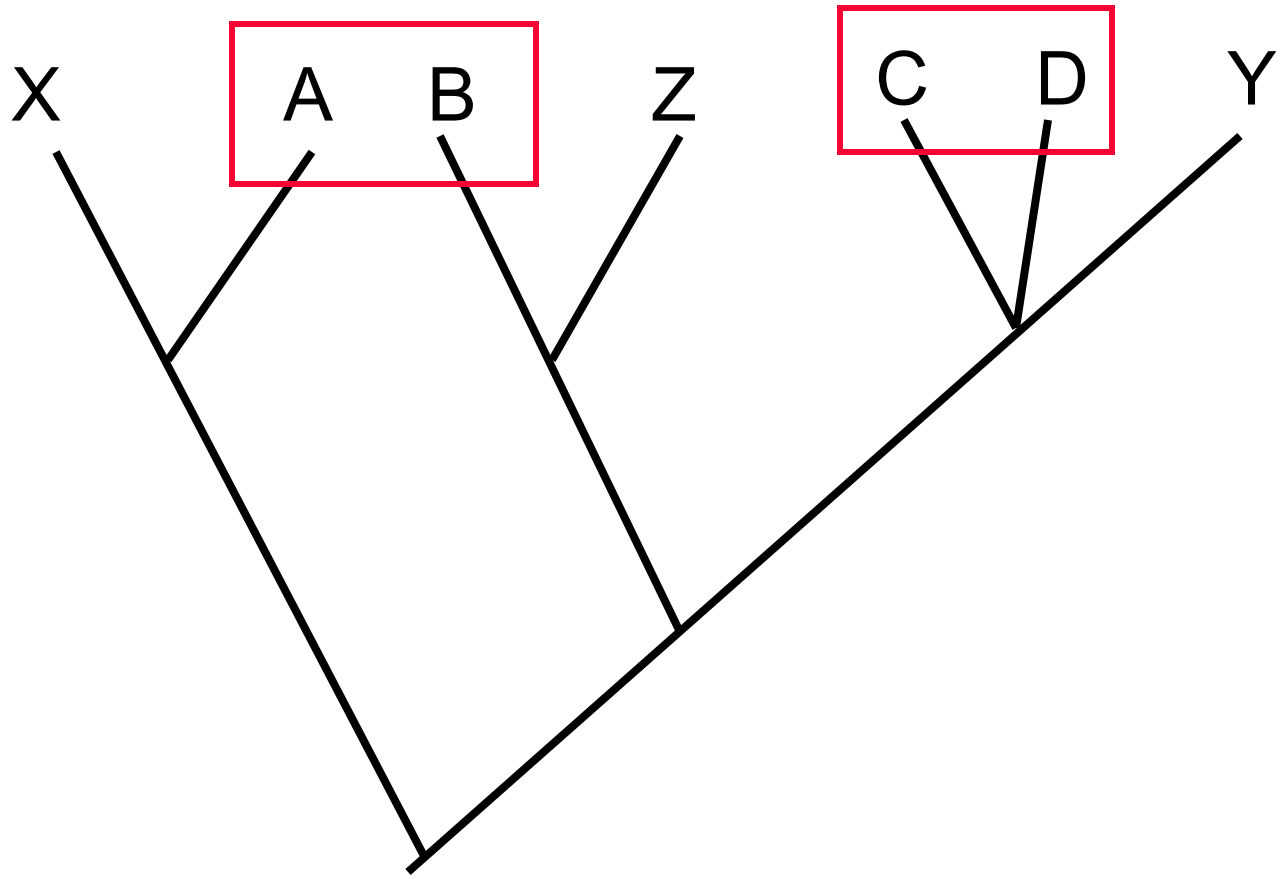
# A monophyletic clade



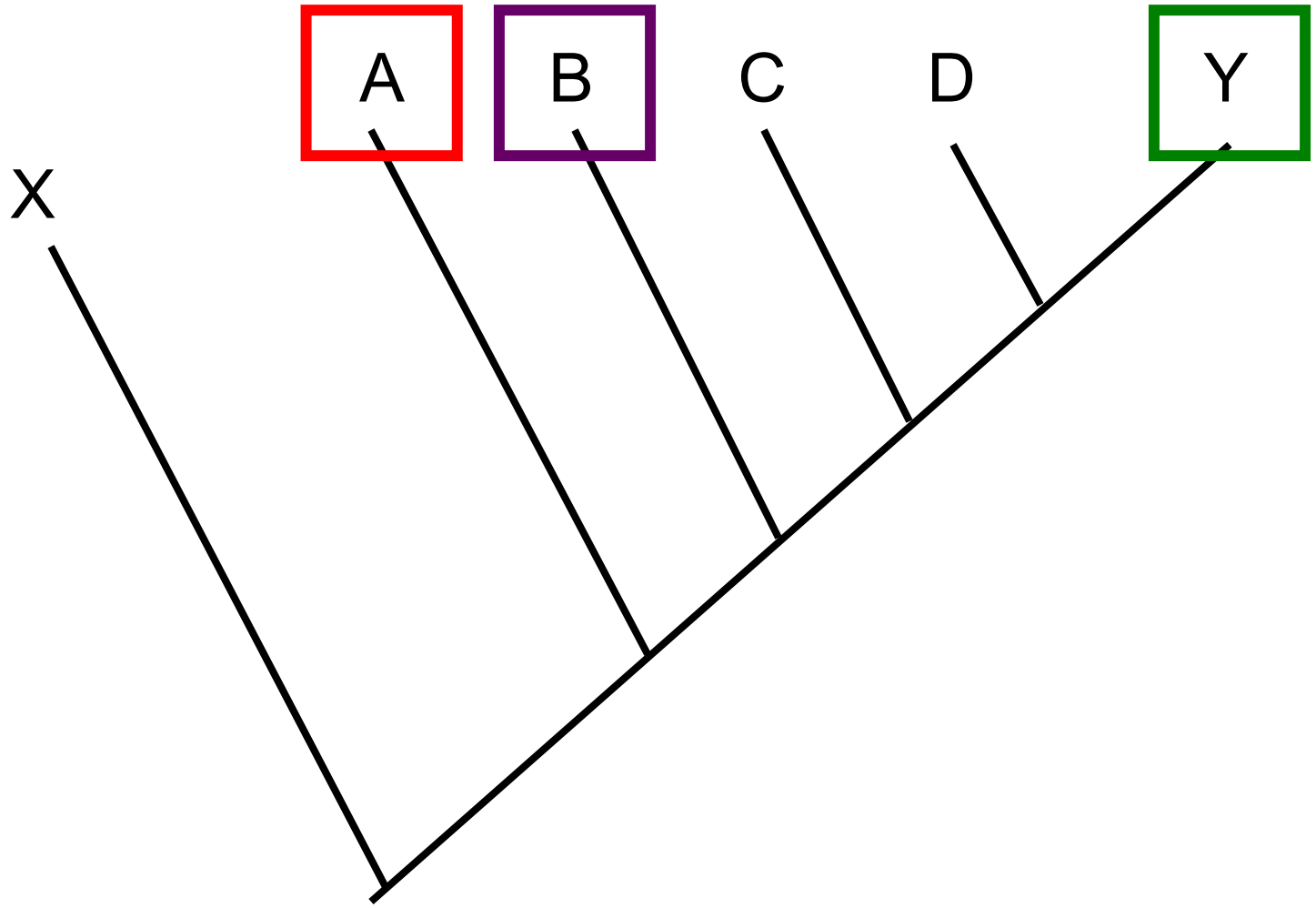
# A paraphyletic group



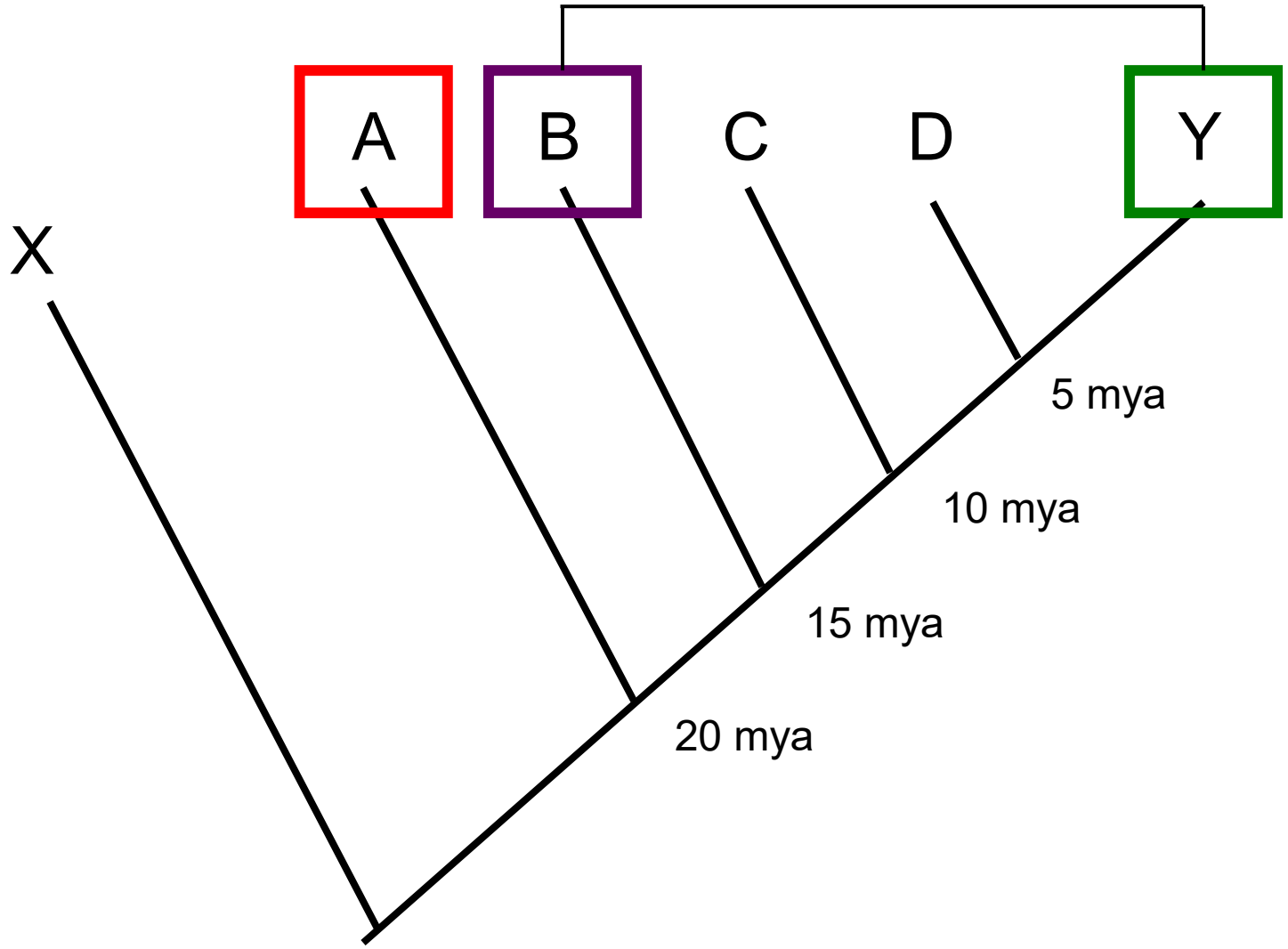
# A polyphyletic group



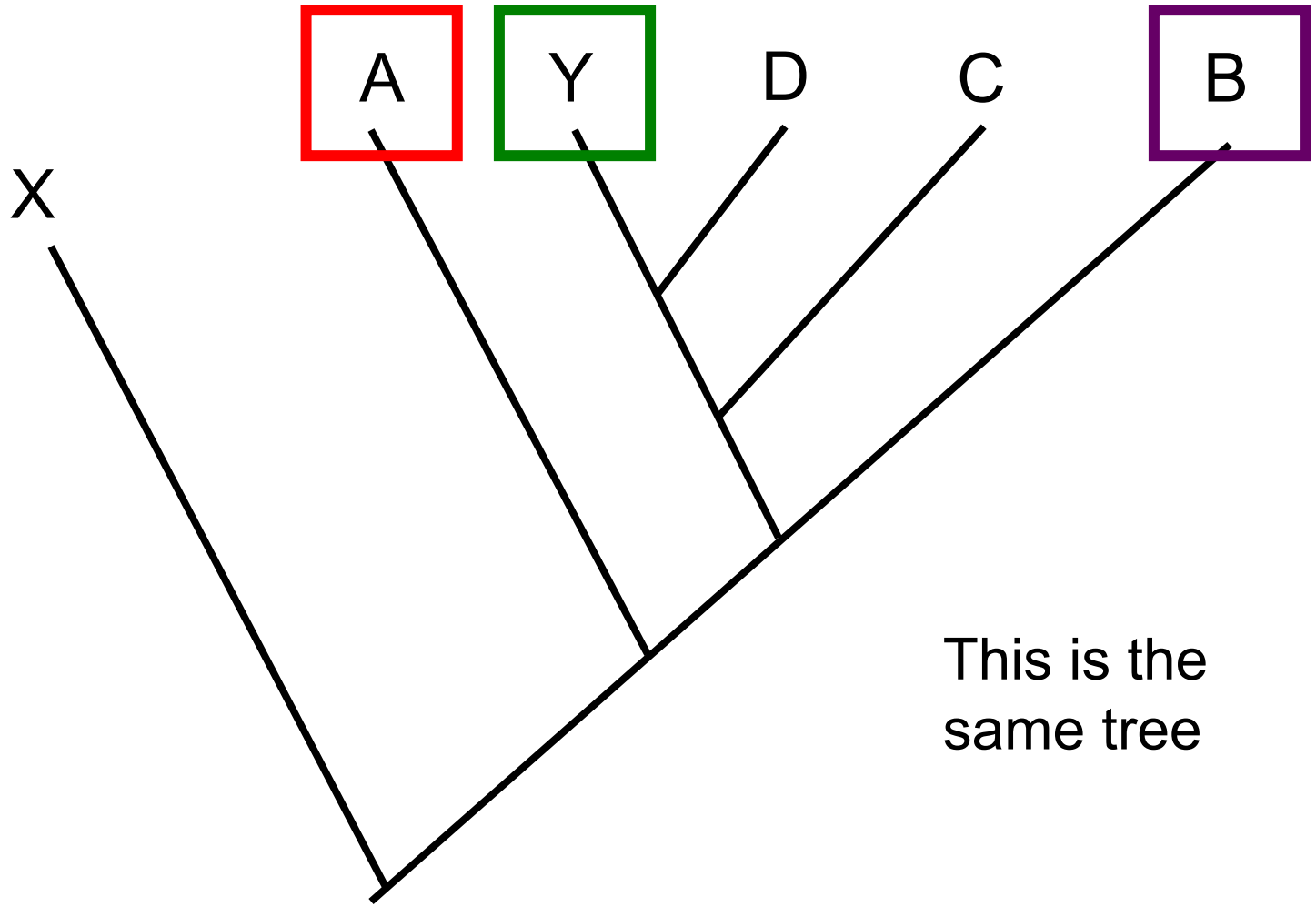
# Quick Quiz: Is B more closely related to A or to Y?



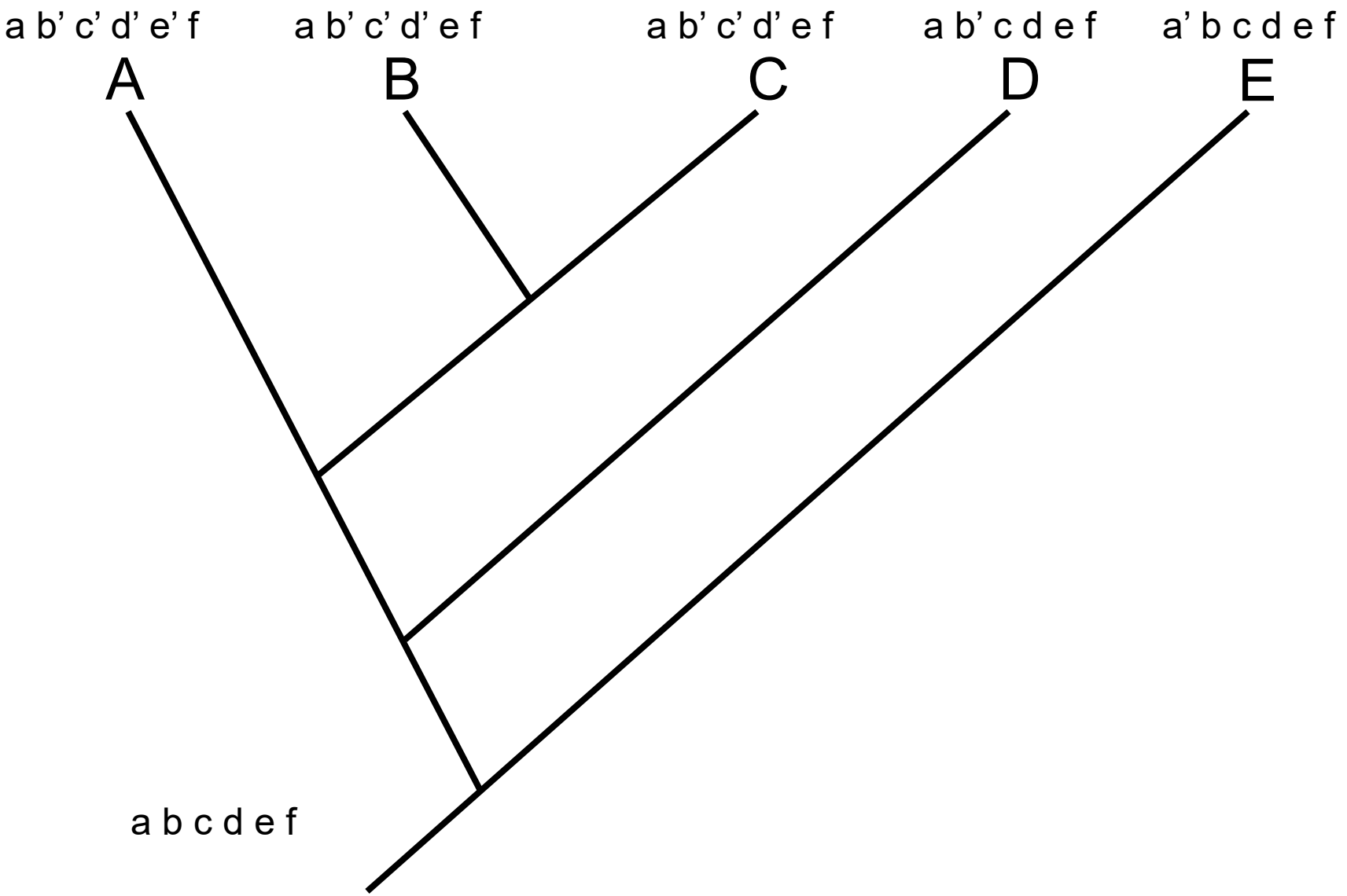
B more closely related to Y because they share a more recent common ancestor.

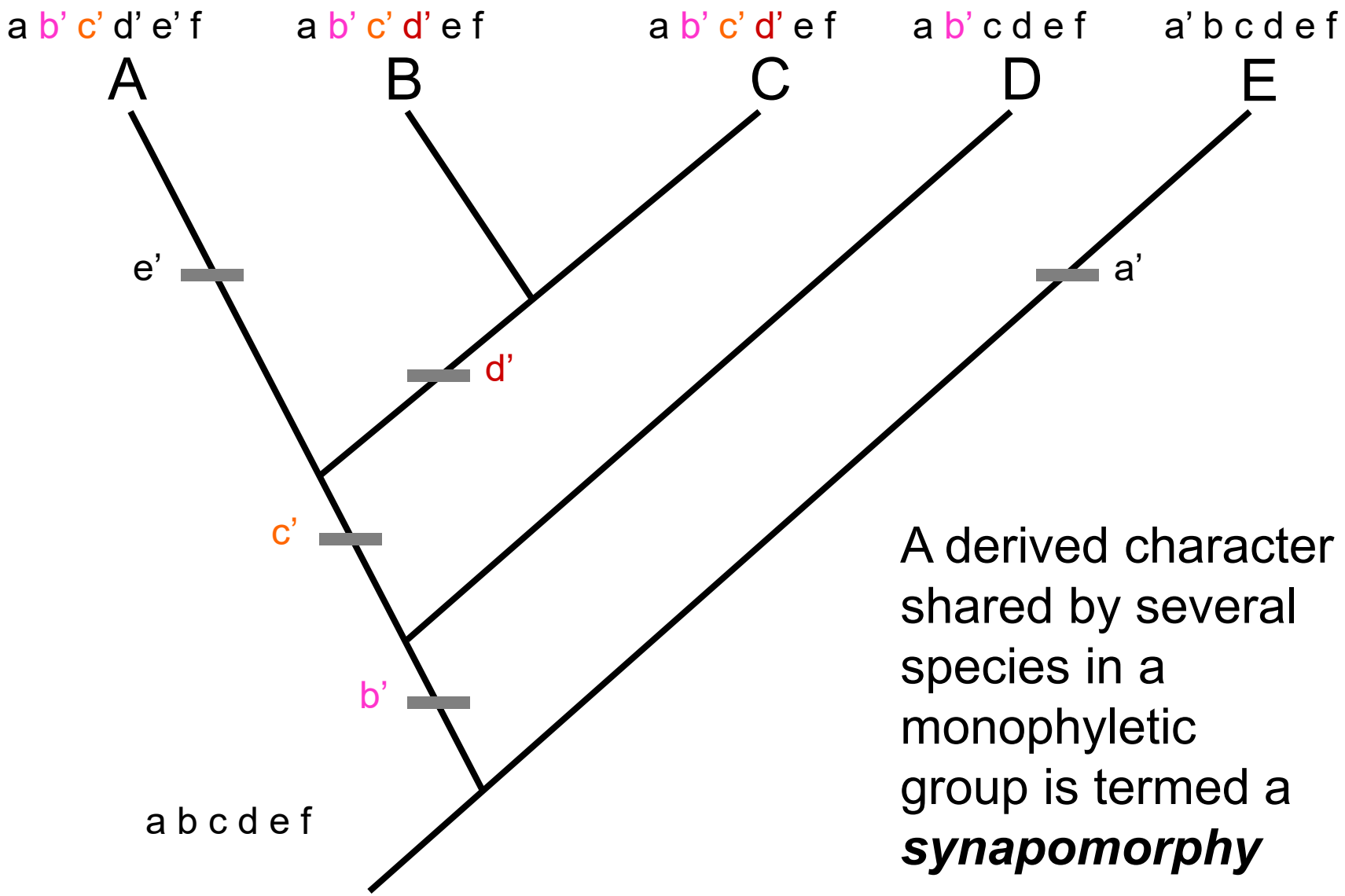


B more closely related to Y because they share a more recent common ancestor.

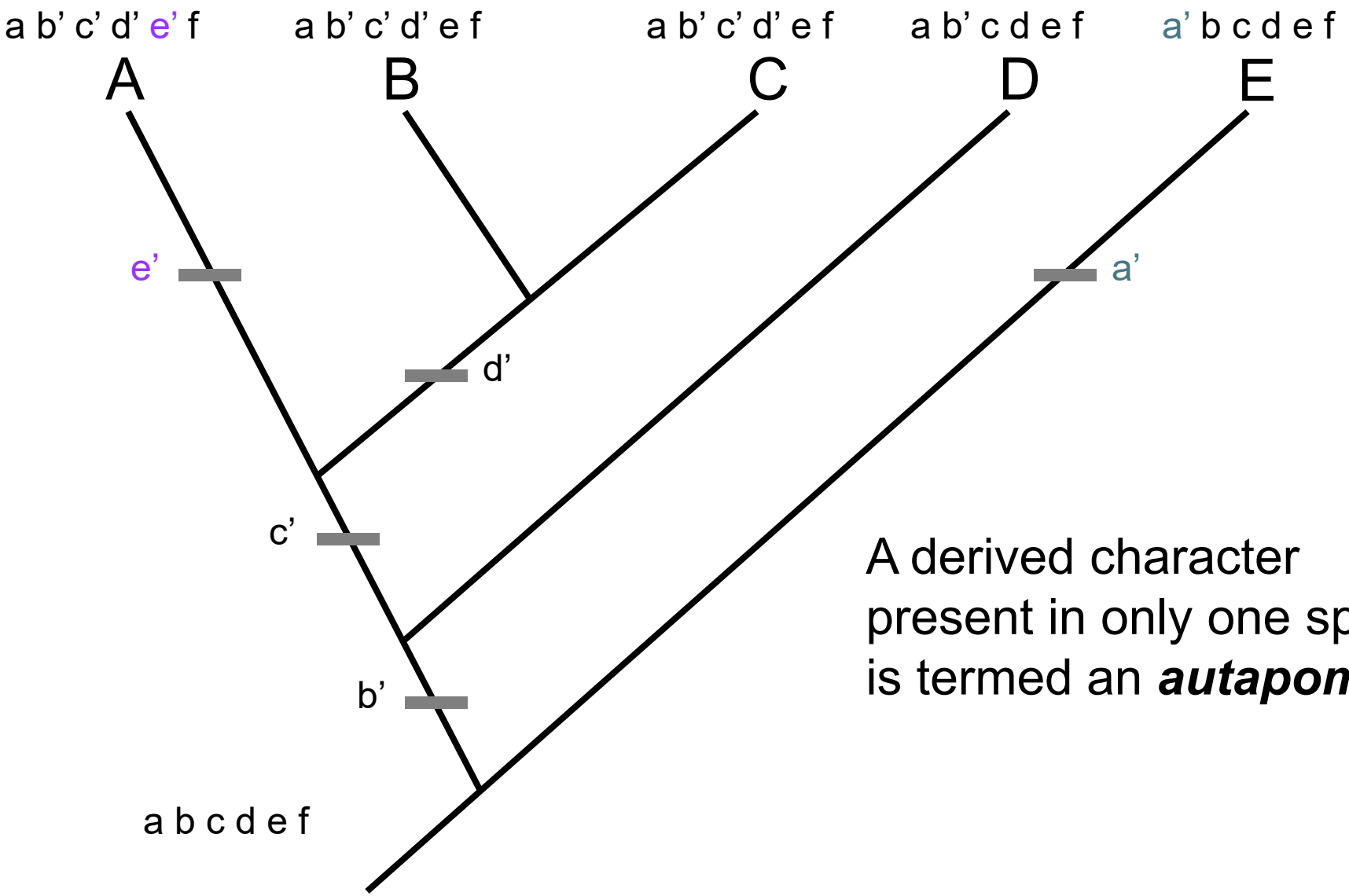



This is the same tree

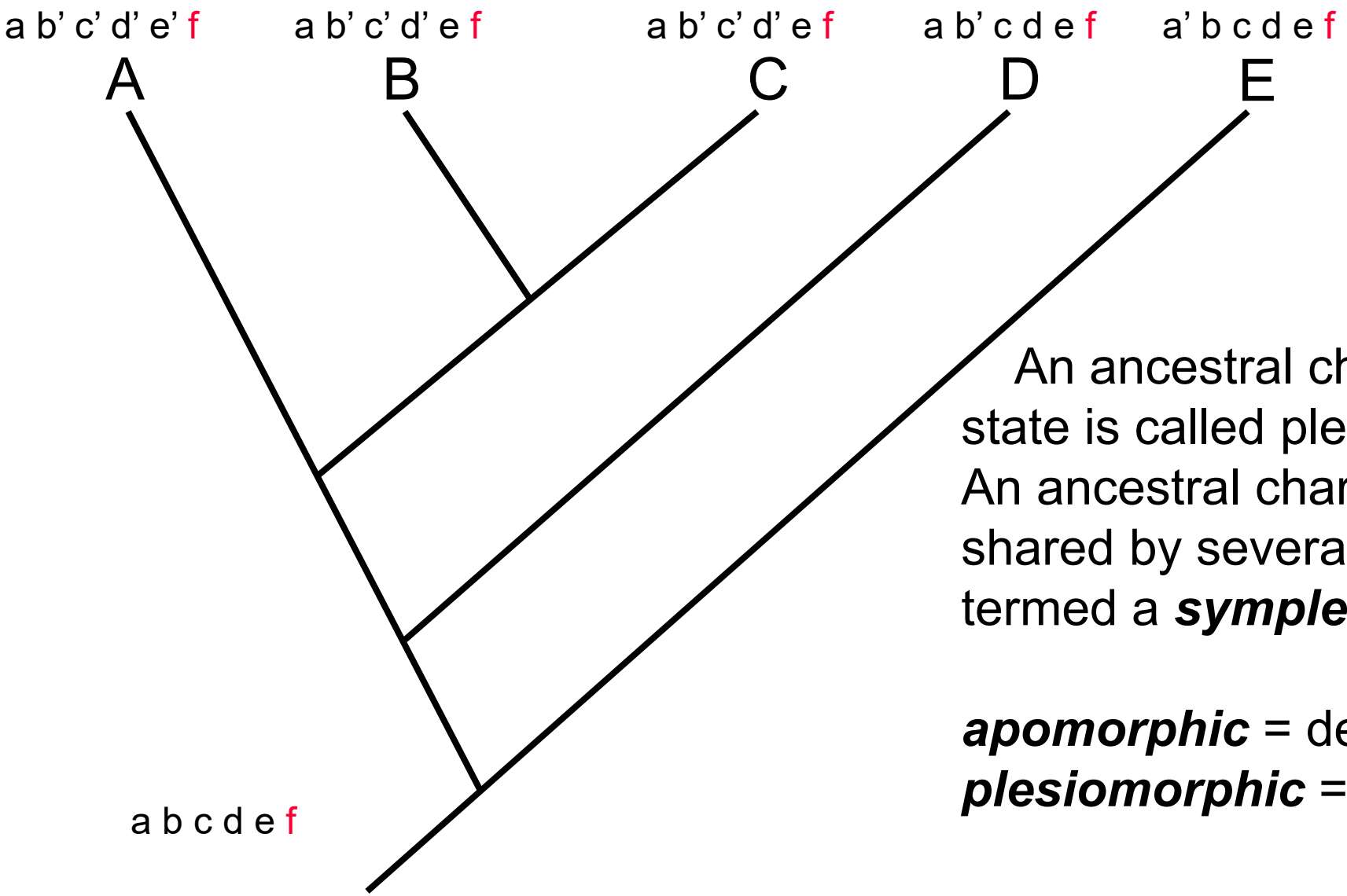

a b’ c’ d’ e’ f
a b’ c’ d’ e f
a b’ c’ d’ e f
a b’ c d e f
a’ b c d e f
A
B
C
D
E
a b c d e f




A derived character shared by several species in a monophyletic group is termed a ***synapomorphy***




A derived character present in only one species is termed an ***autapomorphy***




An ancestral character state is called plesiomorphic. An ancestral character shared by several species is termed a ***symplesiomorphy***

***apomorphic*** = derived
***plesiomorphic*** = ancestral

# Today…

- “Tree thinking”
- Most people try to model evolution to estimate trees (see later…)
- In taxonomy we generally use cladistic arrangements of taxa. We attempt to employ monophyletic groups only
- It’s wrong to think of long branches as “primitive” and short branches as “advanced”. e.g. a mouse or an amoeba is not more *primitive* than a human!
- Humans are not advanced monkeys, or fish, or bacteria
- We did not evolve from monkeys, or fish, or bacteria – they aren’t our ancestors!
- We did share ***common ancestors*** with all of these, however!

# The numbers of possible trees (unrooted, here) explodes with the number of taxa studied!

| Number of Organisms | Number of alternative Trees |
|---|---|
| 3 | 1 |
| 4 | 3 |
| 5 | 15 |
| 6 | 105 |
| 7 | 945 |
| 10 | 2.027.025 |
| 15 | 7.905.853.580.625 |
| 20 | $2.21 * 10^{20}$ |
| 50 | $2.84 * 10^{76}$ |

Table 2.1: Number of possible trees for phylogenies with 3–50 organisms

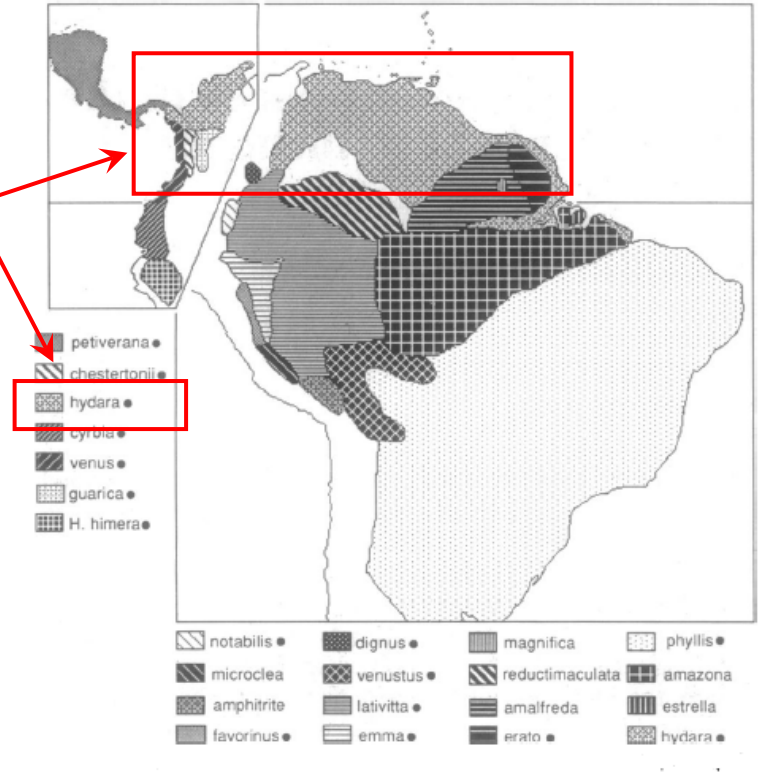
# Methods of estimating phylogenetic trees

- Most people these days **use DNA (or amino acid) sequence data** to construct phylogenetic trees by some sort of numerical algorithm
- Evolution, especially of DNA, can be **unparsimonious**, so parsimony not often used any more (but *Cladistics* journal mandates it!)
- **Distance-based** (phenetic) tree construction, e.g. “neighbor-joining” algorithm, works reasonably well for molecular characters, because of approximate neutrality and molecular clock. Computationally easier.
- However, most these days use “model-based” approaches, involving maximizing the **likelihood***, or the **Bayesian probability***, of the tree based on modelling DNA or amino-acid evolutionary rates. Computationally much more difficult.

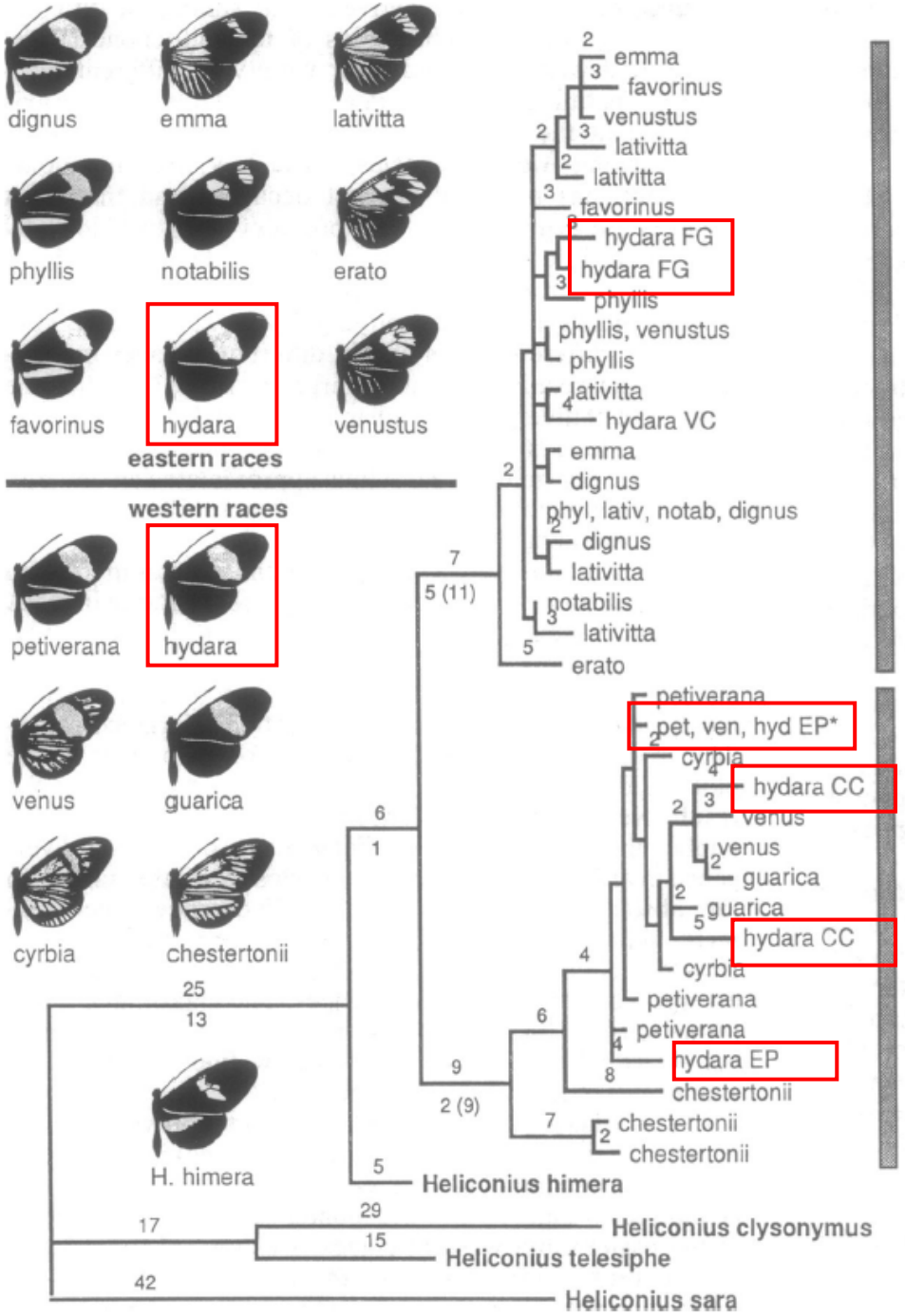
* see slides at end (may not get to them in lecture!)

# Parsimony-based mtDNA tree for *Heliconius erato* subspecies

*Heliconius erato hydara* distribution in Northern South America




Brower 1994




Shown here is one of the 2094 most parsimonius trees for 47 sequenced individuals, with 272 steps. Branch lengths = steps (# steps above branches).

Andy Brower carried out ***THE*** first DNA-based study of *Heliconius*. He scored mitochondrial DNA bases by hand using autoradiographs.

I thought it odd that Brower argued that the hydara colour pattern had evolved twice, based on mtDNA. I was asked to write a news article, so I expressed doubts.

Brower never forgave me!

# Evolution: **Mimicry meets the mitochondrion**

James Mallet, Chris D. Jiggins and W. Owen McMillan

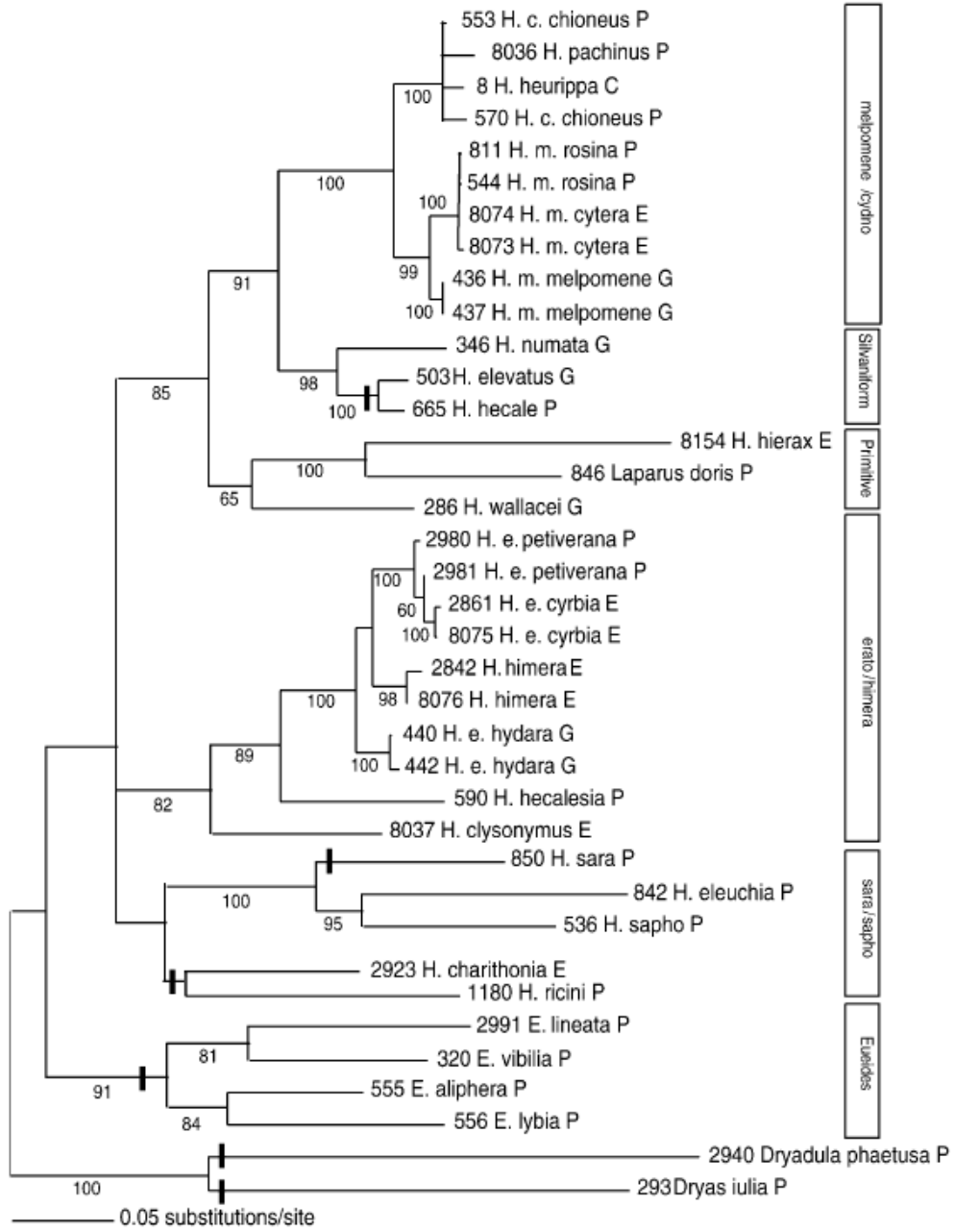
**A recent molecular study of the evolution of mimicry in tropical butterflies of the genus *Heliconius* proves that the mimics adapted to previously diverged 'model' species, but does not clearly distinguish between opposing views of how the model species diverged.**

Address: Galton Laboratory, Department of Biology, 4 Stephenson Way, London, NW1 2HE, UK.

**Current Biology** 1996, Vol 6 No 8:937–940

# Likelihood-based mtDNA tree: *Heliconius* species

- Branch lengths represent numbers of substitutions.
- Numbers below branches represent % bootstrap support.
- Vertical bars show insertions/deletions
- We imagined that with more genes, we would be able to discover the true tree of life for all species!
- (Let's see how that worked out!....).



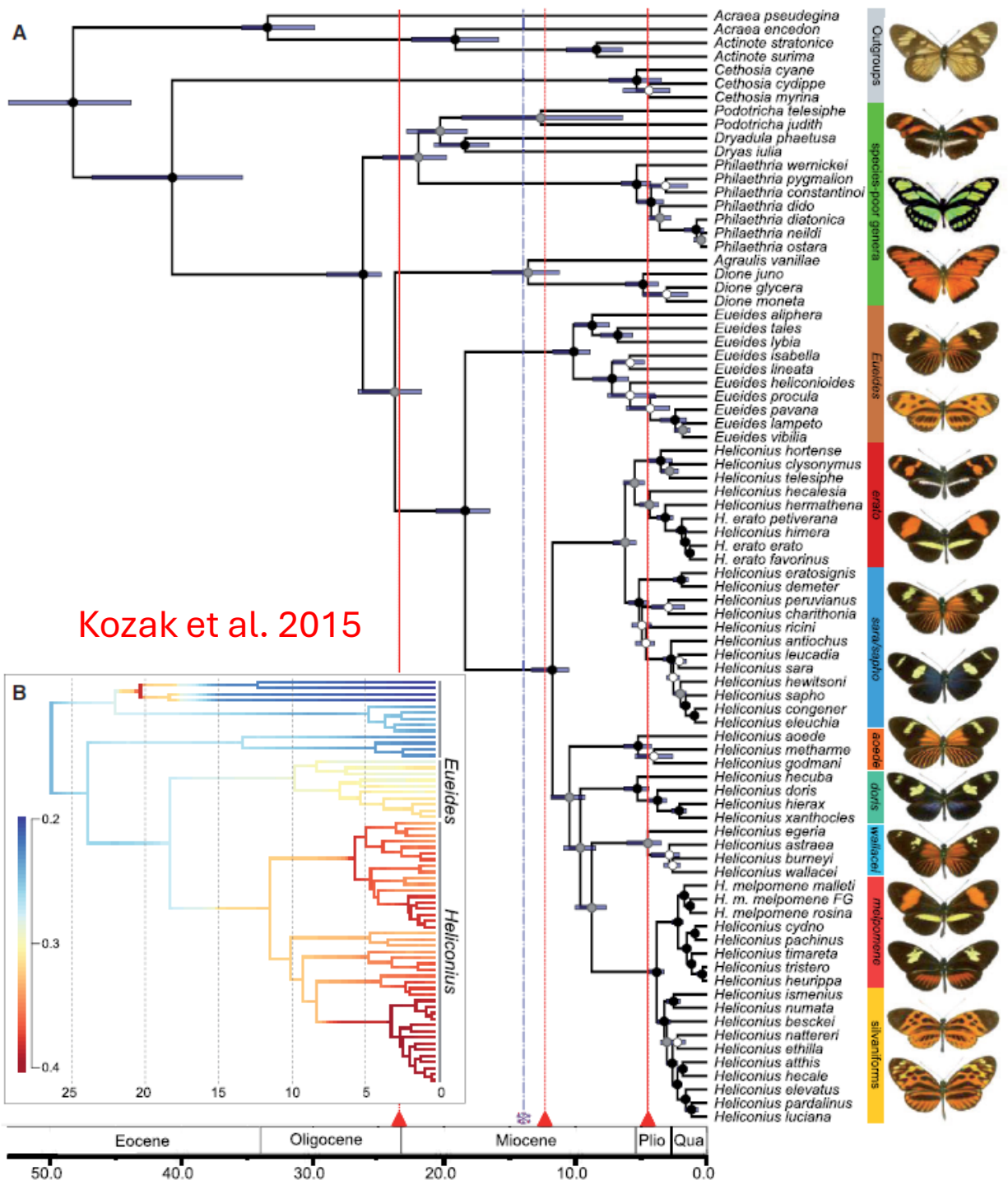

Beltran et al. 2002

# Bayesian “time tree” (21 genes) *Heliconius*

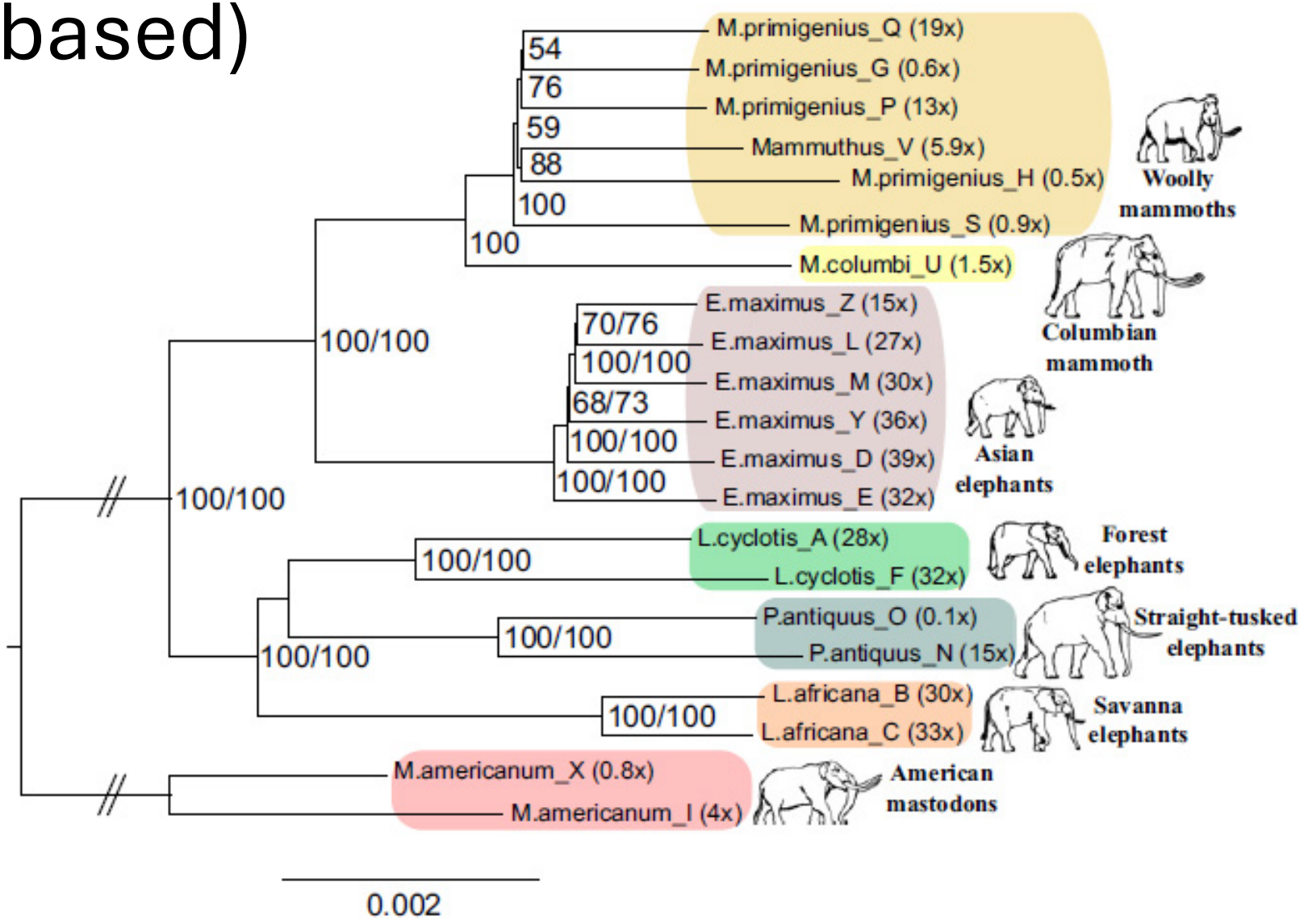
- Branch lengths represent time
- Support for each node shown as horizontal bars.
- Clearly, the genus *Heliconius* is an (adaptive) radiation, compared with other related taxa

Kozak et al. 2015

Scale of rapidity of diversi-fication

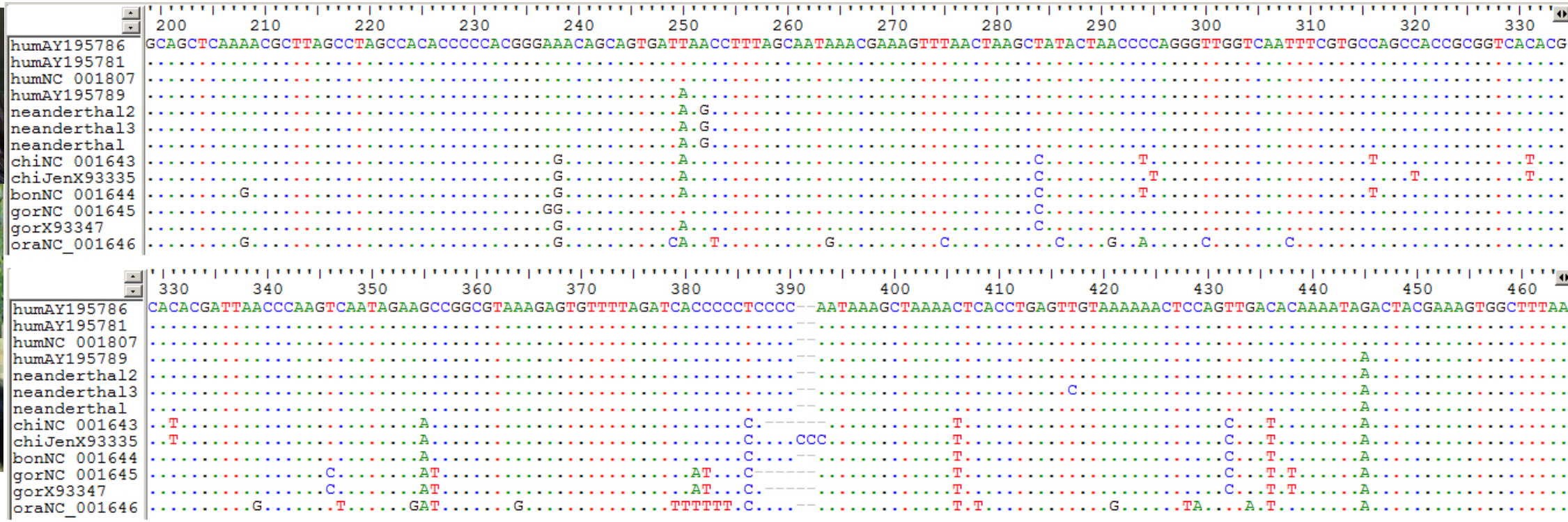


# Mammoths – Neighbor-joining tree – whole genome (distance-based)




Palkopolou et al. 2018

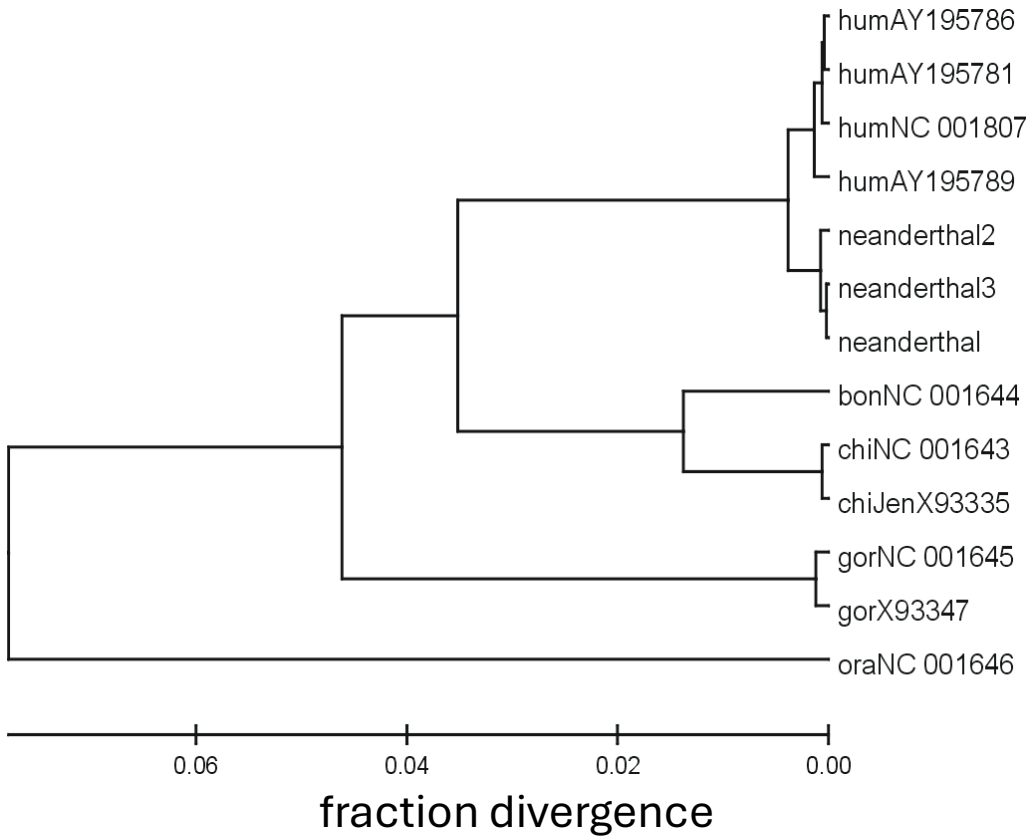
# DNA sequence variation




- 260 base pairs (bp) of 17,000 bp of mtDNA of apes
- Nuclear genome: 3,200,000,000 bp (3.2Gb)
- Average nuclear genome divergence human-chimp genome: 1% of base pairs

whole mitochondrial genome sequences downloaded from GENBANK
https://www.ncbi.nlm.nih.gov/genbank/

# Human/Neanderthal/chimp/gorilla NJ tree




Anatomically modern humans

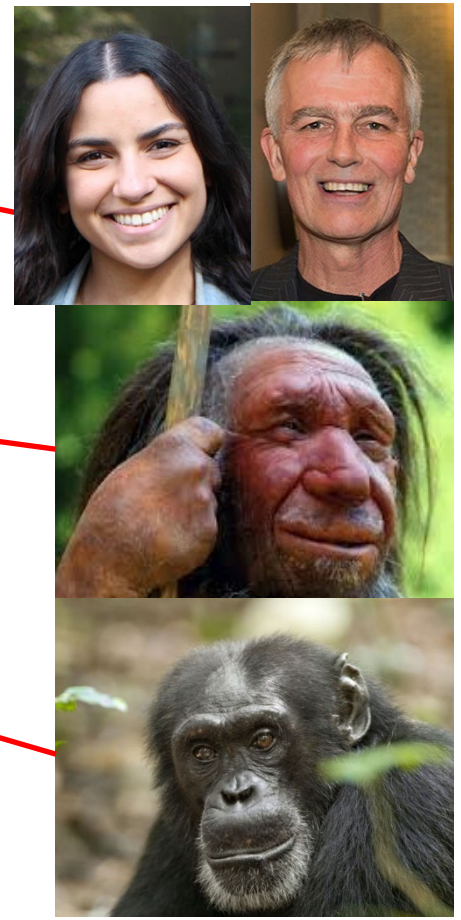
Neanderthal

Bonobo

Chimp

Gorilla

Orang



Based on comparison of 16Kb mtDNA

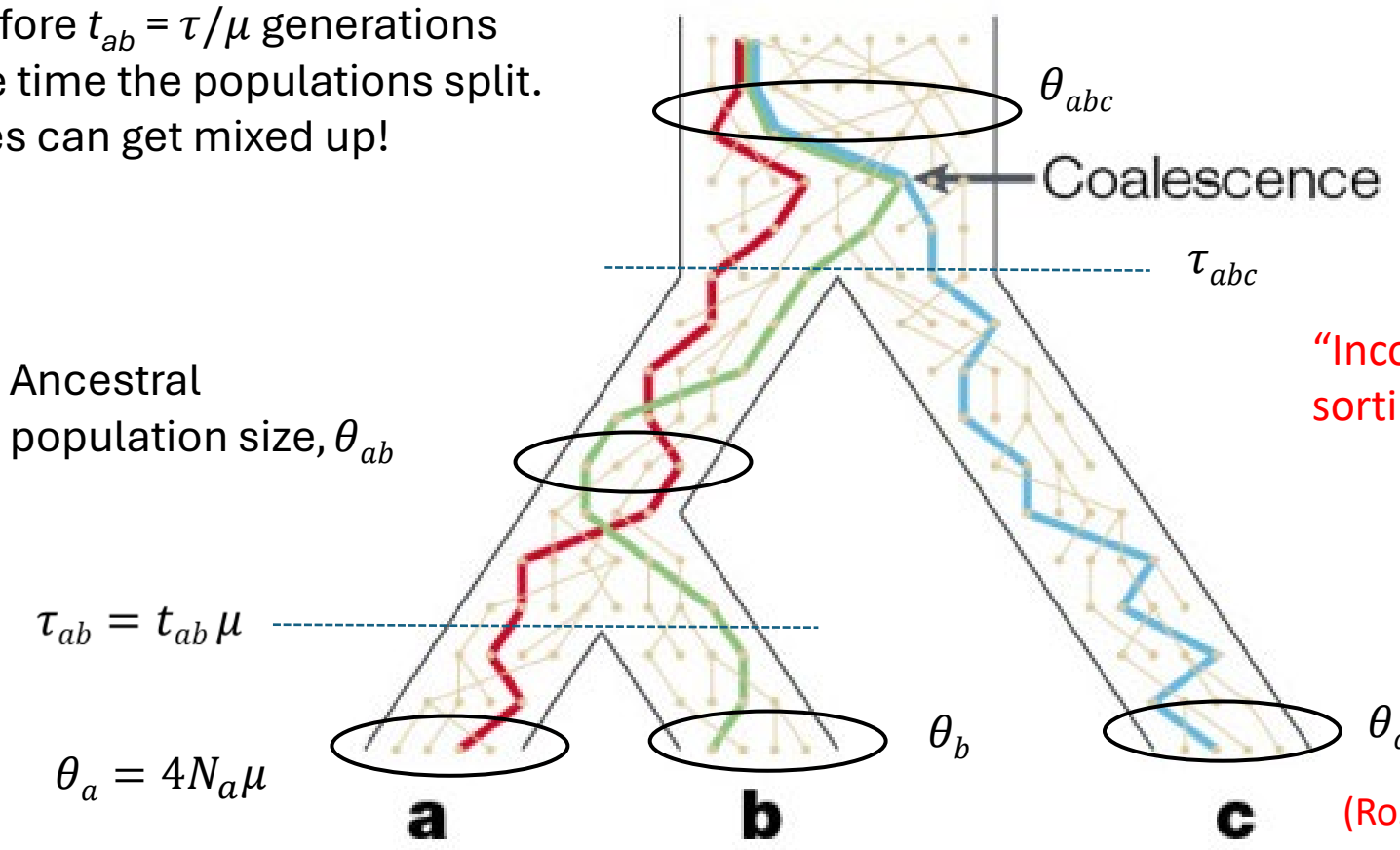
I made this distance-based tree in about ½ hr by (1) downloading genbank whole mt genome sequences, (2) aligning them and (3) running them through tree-building software

# So trees are complicated, but wait …!

- So far, we have been assuming you have haploid sequence data, with no sex or recombination
- What we have been looking at so far are “gene trees.” For multi-gene, mtDNA, or whole genome data, we have been looking at “concatenated sequence data”, i.e. assuming a single haplotype.
- Trees would be easy if all species were asexual haploids; but they aren’t (but methods so far mentioned have assumed this!).
- Real “species” consist of more than one haploid genome! Species are in *populations* of recombining genomes.

# The neutral coalescent between species

Coalescence of a "gene" between two species a, b, will occur some time before $t_{ab} = \tau/\mu$ generations ago, the time the populations split. Lineages can get mixed up!

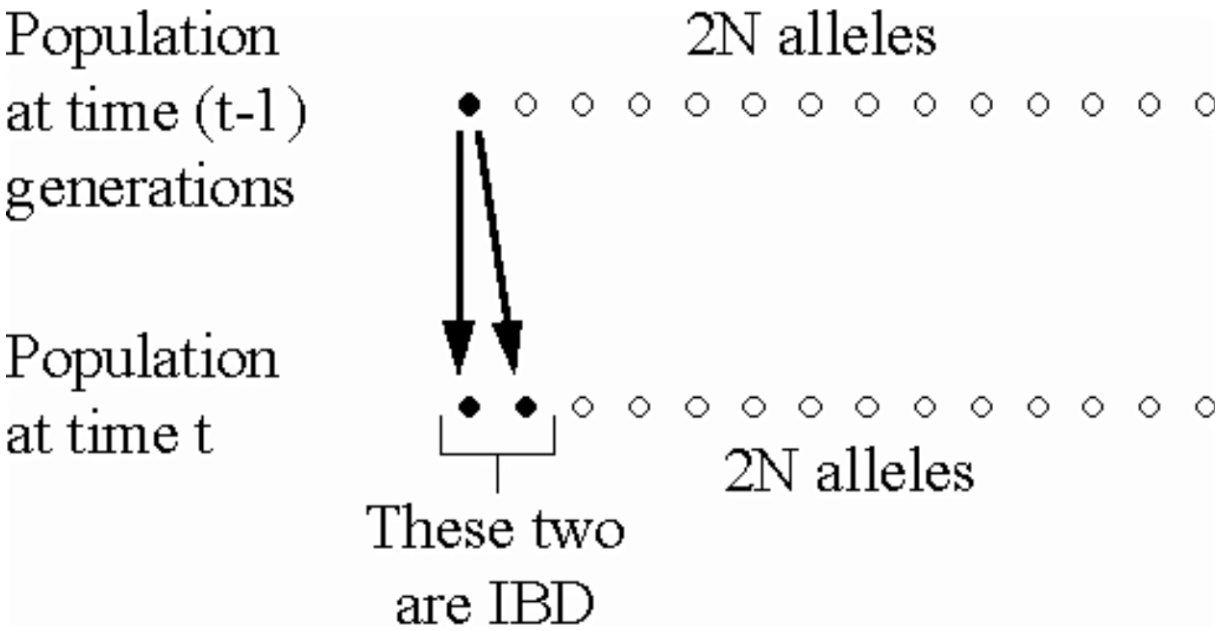



"Incomplete lineage sorting" will be common!

(Rosenberg & Nordborg)

Note: to simplify, **we suppose we can find genes that do not recombine.**
Note: the species tree is often different from the gene tree!

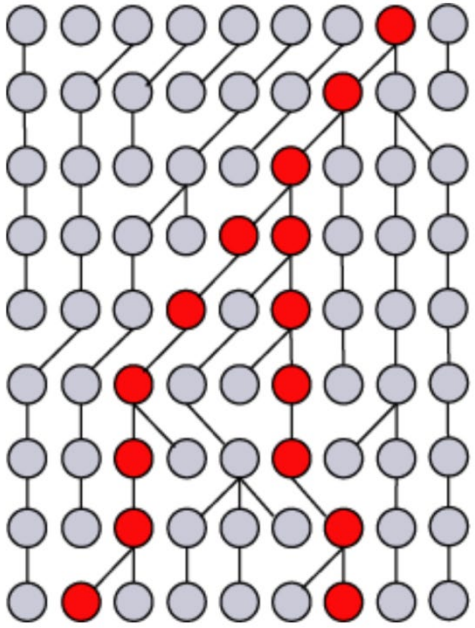
# Assume neutral drift in finite N populations: *forwards*

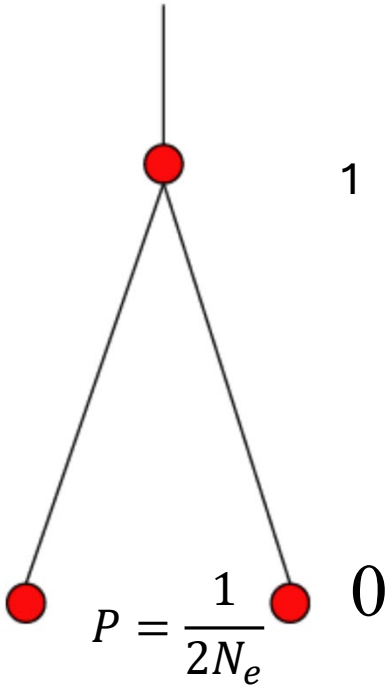



In a population of size $N$, the extra probability that two alleles picked at random in generation $t$ are "identical by descent" due to copying from generation $t$ - 1

is (on average): $\frac{1}{2N}$

# Probability of coalescence over time: *backwards*






$P$

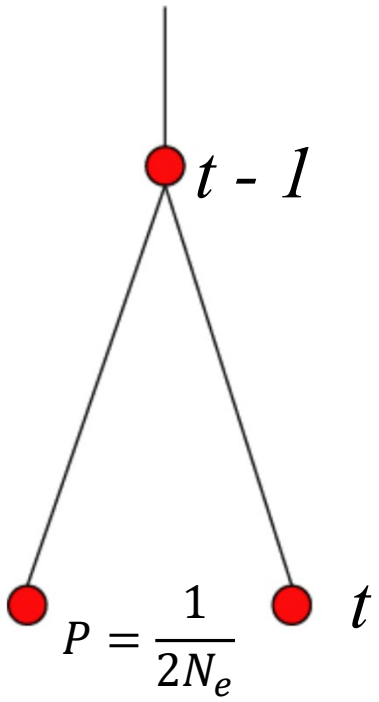
1-$P$

| | |
|---|---|
| P(Coalesces 1 generation ago) | 1/2N |
| P(Coalesces 2 generations ago) | (1-1/2N) * 1/2N |
| P(Coalesces 3 generations ago) | $(1-1/2N)^2 * 1/2N$ |
| P(Coalesces 4 generations ago) | $(1-1/2N)^3 * 1/2N$ |
| P(Coalesced *t* generations ago) | $(1-1/2N)^{t-1} * 1/2N$ |

Coalescence of **two gene copies** follows a *geometric distribution* with mean **2N**

(from Elliot & Mooers)

# Time back to coalescence of a pair of alleles

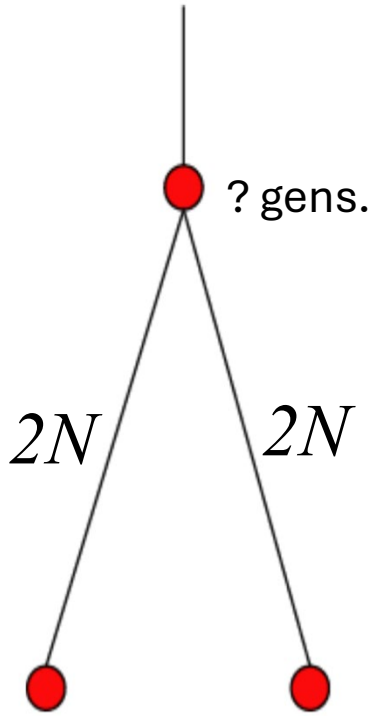



In a single generation, the probability of coalescence for any particular pair of alleles is 1/(2*N*). This is the average number of coalescences per allele pair per generation.

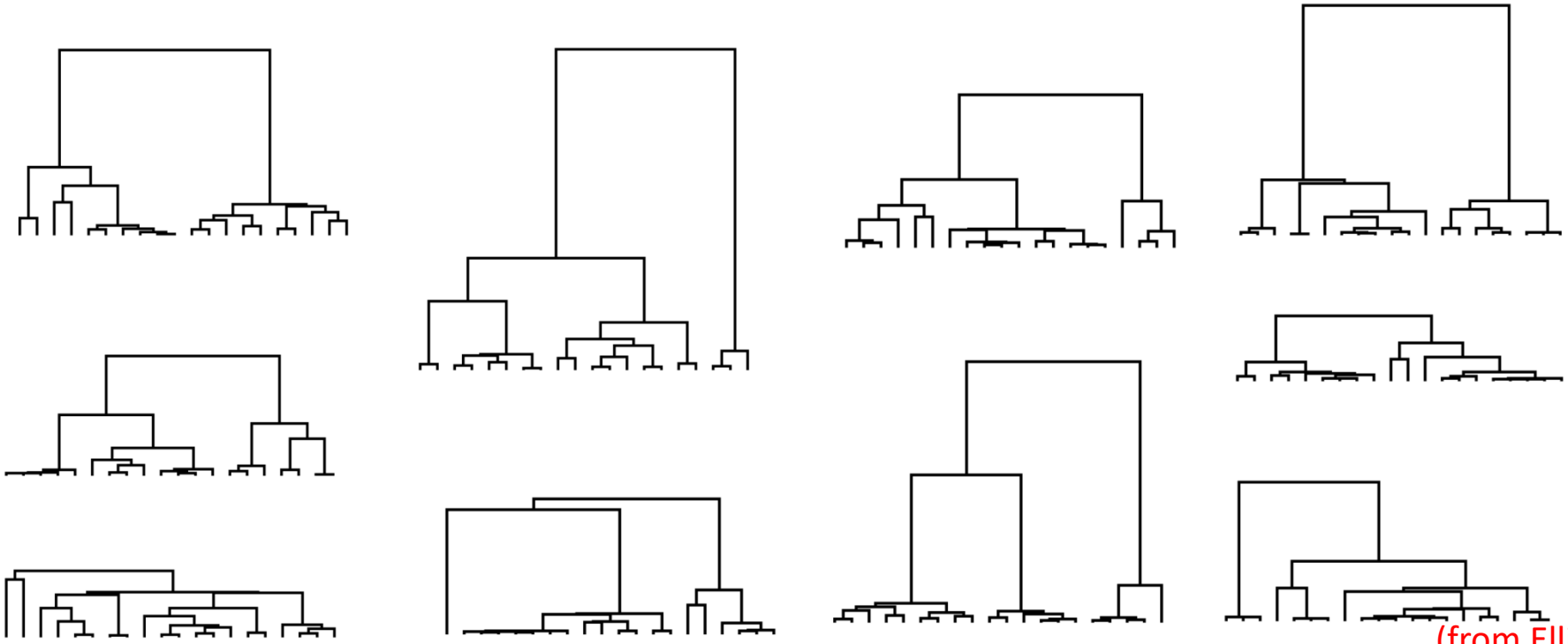
So the mean numbers of generations per coalescence is the reciprocal: ~2*N*

It takes about 2*N* generations for a randomly selected pair of alleles to coalesce.

So we expect heterozygosity, $\pi \approx \theta = 4N\mu$



# Simulations of the coalescent (sample size, $k$ = 20, for a given $2N_e$)



(from Elliot & Mooers)

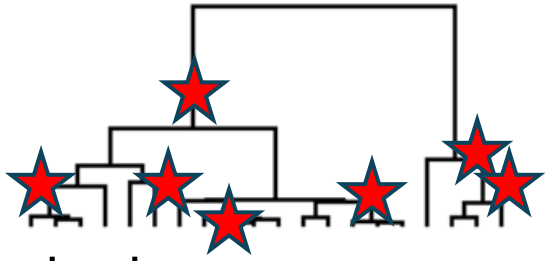
Very variable! A single gene does not give a good idea of average behavior

e.g. $N_e$ = 1000, mean = $4N_e$ ~ 4000 gens., SD = 7430 gens
(using Tajima 1983 theory).

# Practical uses of the neutral coalescent

Typical uses in understanding evolution within species:

i) simulate a genealogy with given $2N_e$
(backwards in time from sample, of course)



ii) Pepper mutations randomly onto genealogy at rate $\mu$
iii) Repeat steps i and ii many times to get distributions based on different parameter values
iv) Compare with data, e.g. expect heterozygosity, or % difference between two randomly selected haplotypes:

$$\pi \approx \theta = 4N\mu$$

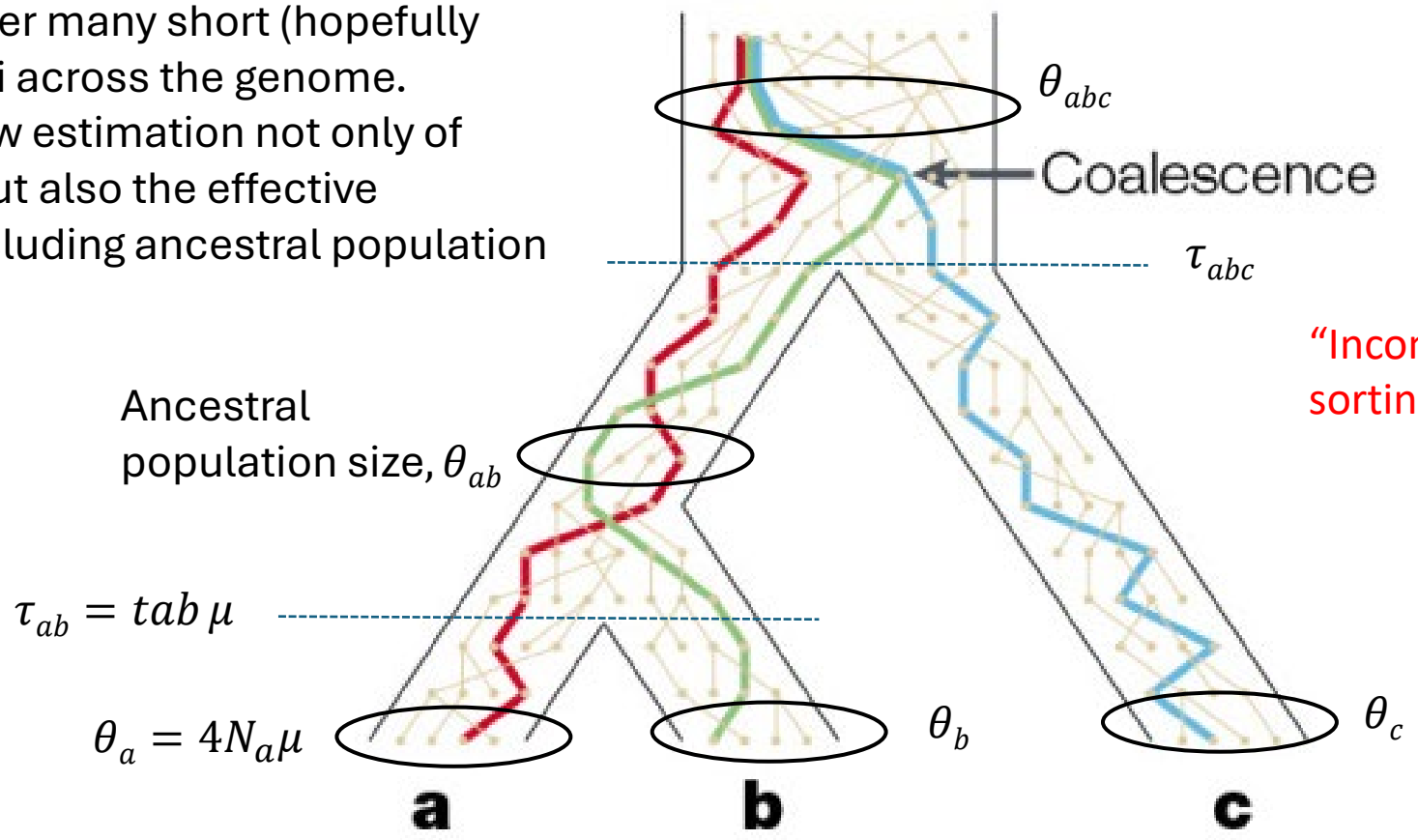
v) Do some sort of stats
vi) Very useful in computation because you don't have to model all the ancestral populations back in time. You just need to follow back the sample haplotypes you have taken now.

(partly from Robert Berwick at MIT)

Human polymorphism level, $\pi$ = ??

# The neutral coalescent between species

- The species tree can be estimated using a model-based Bayesian approach that assumes the neutral coalescent over many short (hopefully non-recombining) loci across the genome.
- The methods will allow estimation not only of divergence times $\tau$, but also the effective population sizes $\theta$ including ancestral population sizes

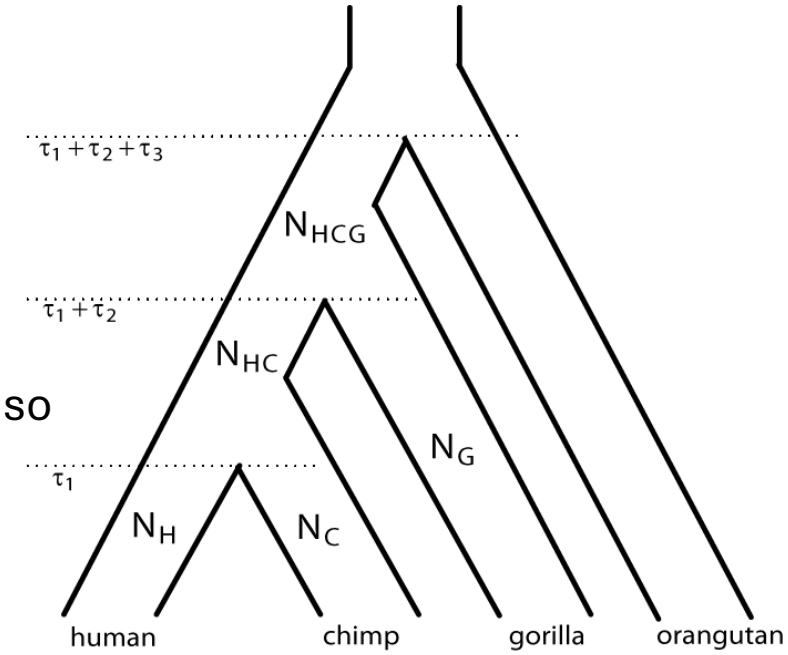



"Incomplete lineage sorting" will be common!

(Rosenberg & Nordborg)

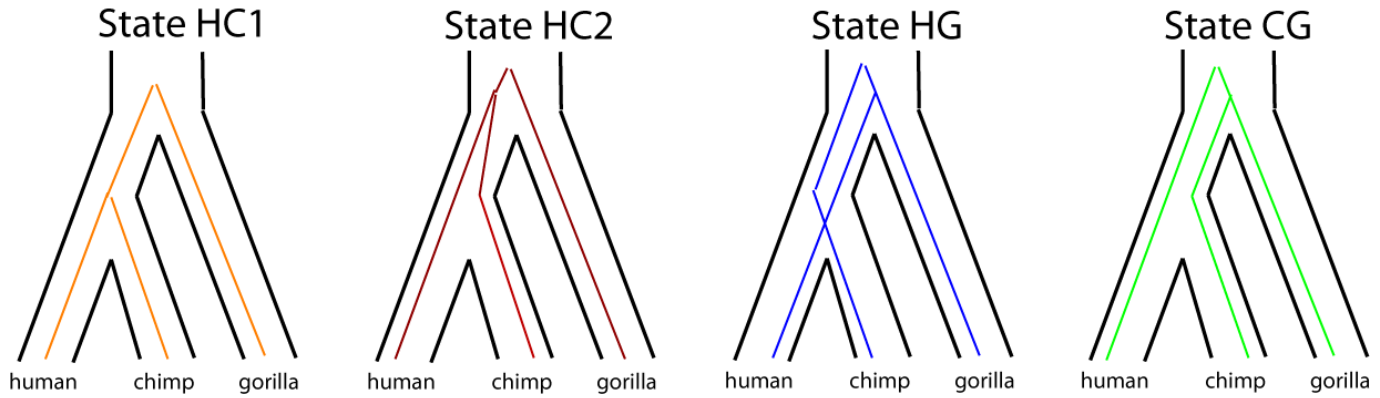
# Gene trees vs. species trees

Gorilla-chimp branch is short so expect …




Incomplete lineage sorting (ILS)

Coalescence in ancestors can lead to "*Incomplete Lineage Sorting*" of the gene trees!




~70% of the genome

~15% of the genome

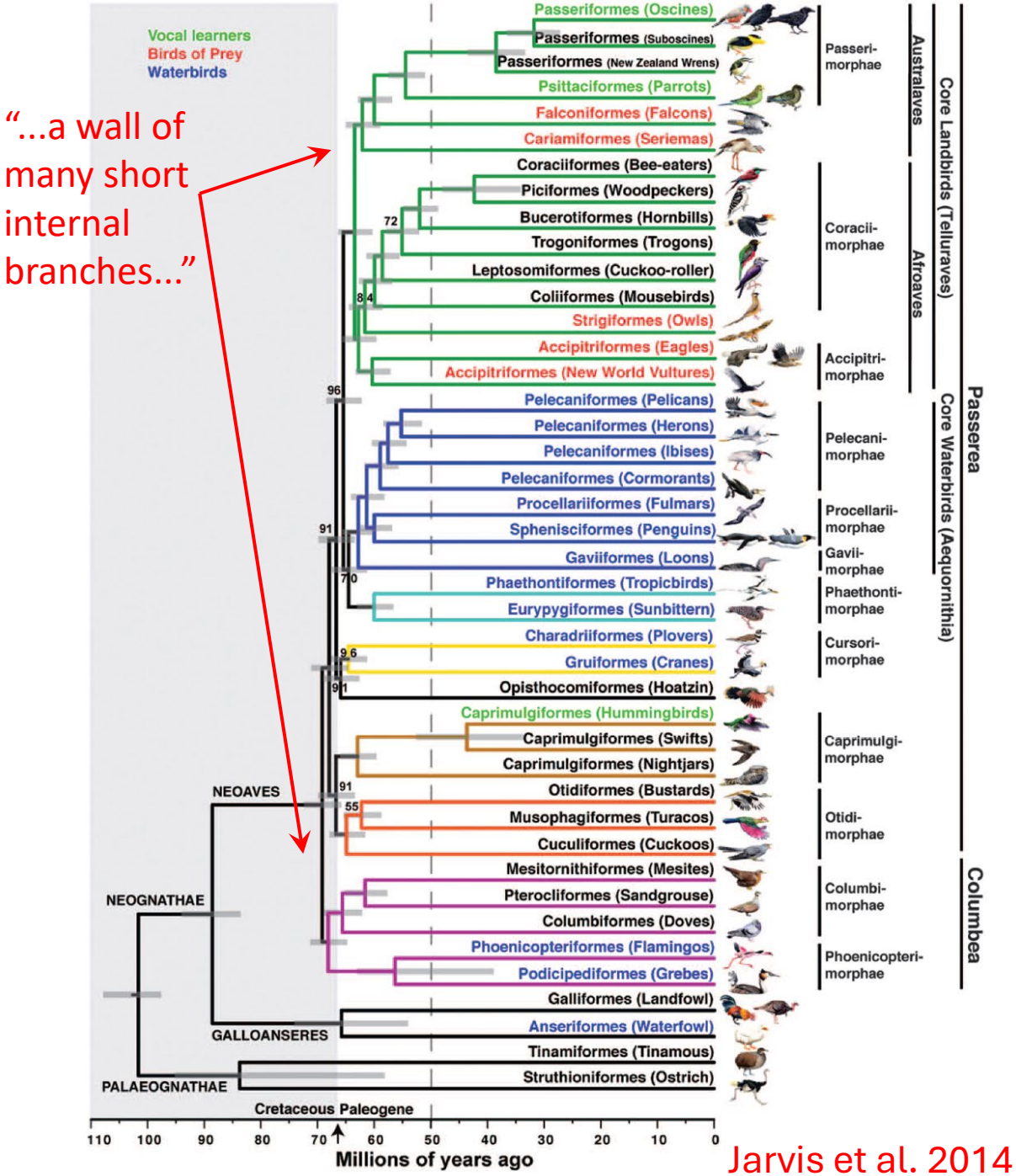
~15% of the genome

What about gene flow?

Holboth et a. 2007; Burgess & Yang 2008

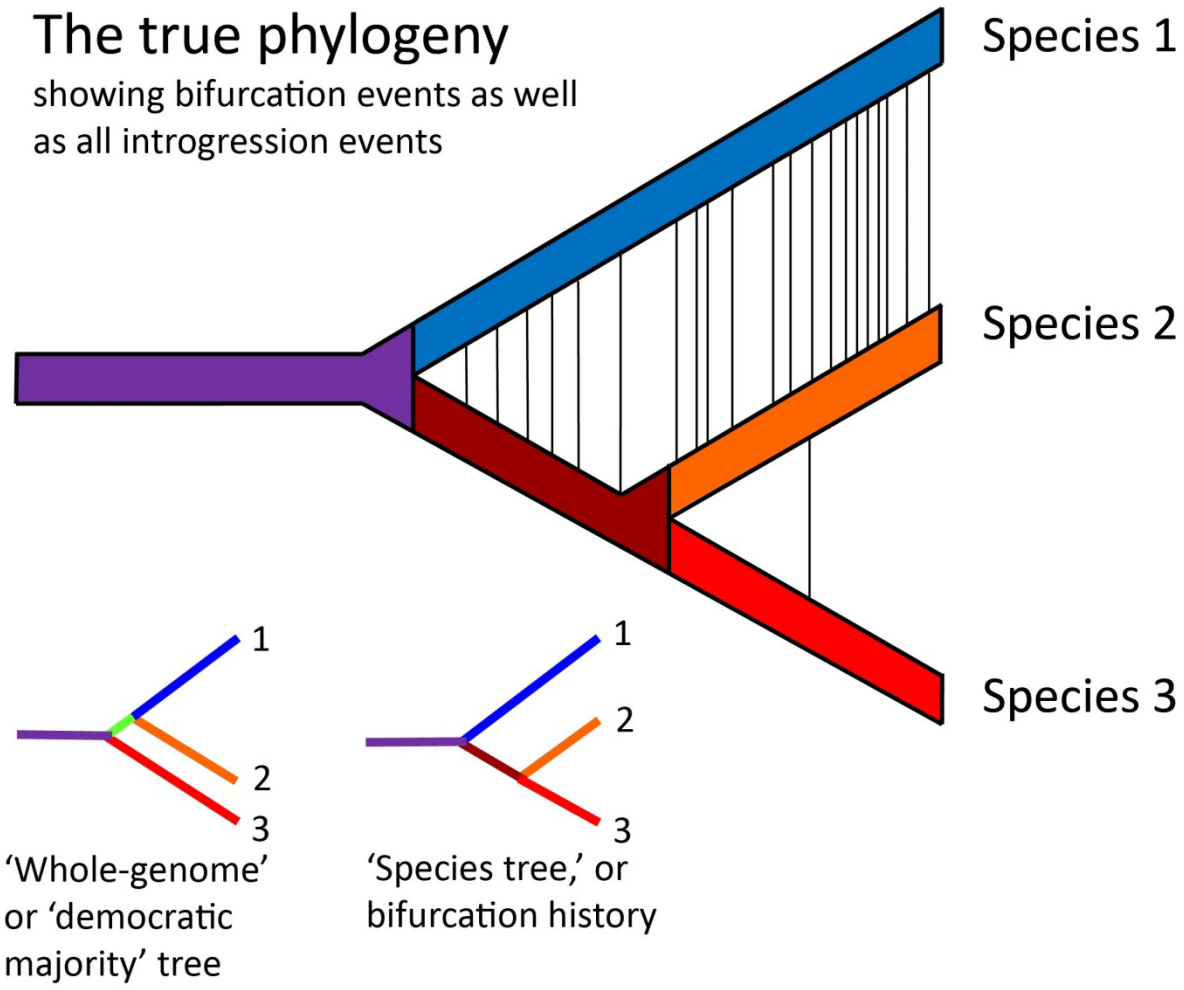
# Bird Time Tree

- There are ~11,000 spp. of birds
- This is a maximum likelihood tree, a Total Evidence Nucleotide Tree (TENT) of the major lineages.
- 33/38 major lineages of the Neoaves arose in 15My near K/Pg boundary, 66 MyA.
- Among 8251 high confidence protein-coding genes (41.8 Mbp, ~3.5% of genome), no single gene tree agreed with the TENT, or with coalescent-based species trees
- Phylogenetic conflict due to ILS?




Jarvis et al. 2014

What about introgression? This will alter the gene trees; In this case, gene flow has caused average gene trees to differ from the species tree!

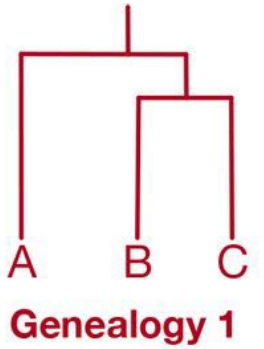


# Previously, estimated species trees ignored introgression!

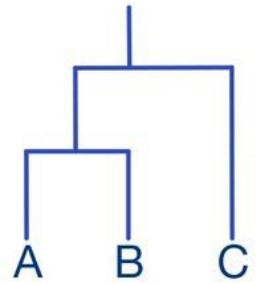
- We need new methods to study gene flow in species trees!
- Very few methods do this properly.  e.g. BPP a program that estimates species trees and introgression jointly, while integrating over possible gene trees

# Introgression? Or incomplete lineage sorting?

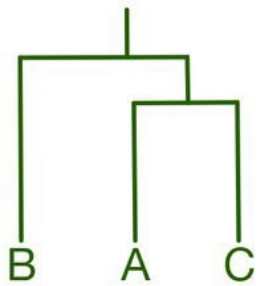
Both incomplete lineage sorting and gene flow will produce gene trees like genealogy 2. But, with enough loci, enough substitutions per branch, one should be able to detect gene flow.



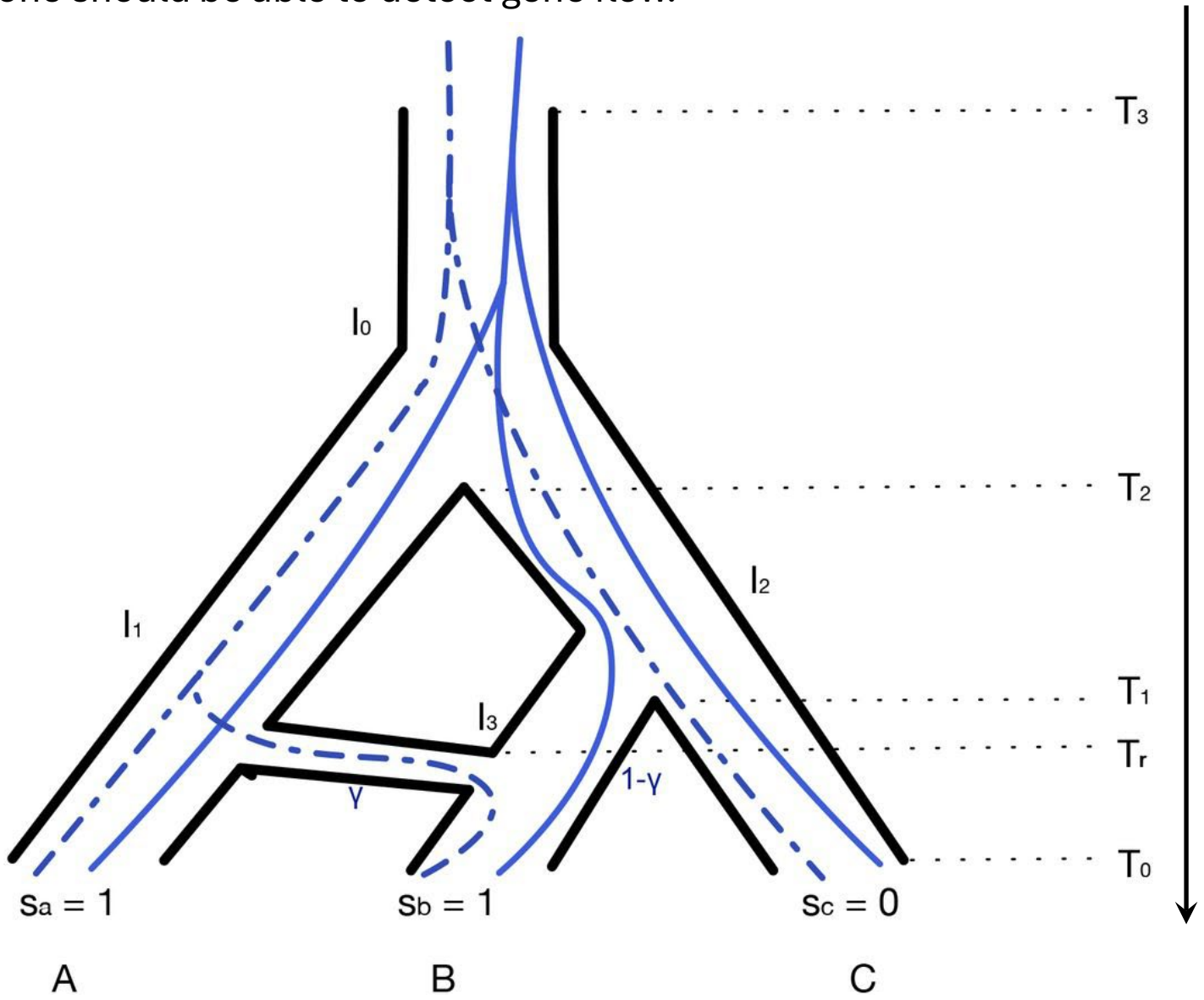

Genealogy 1



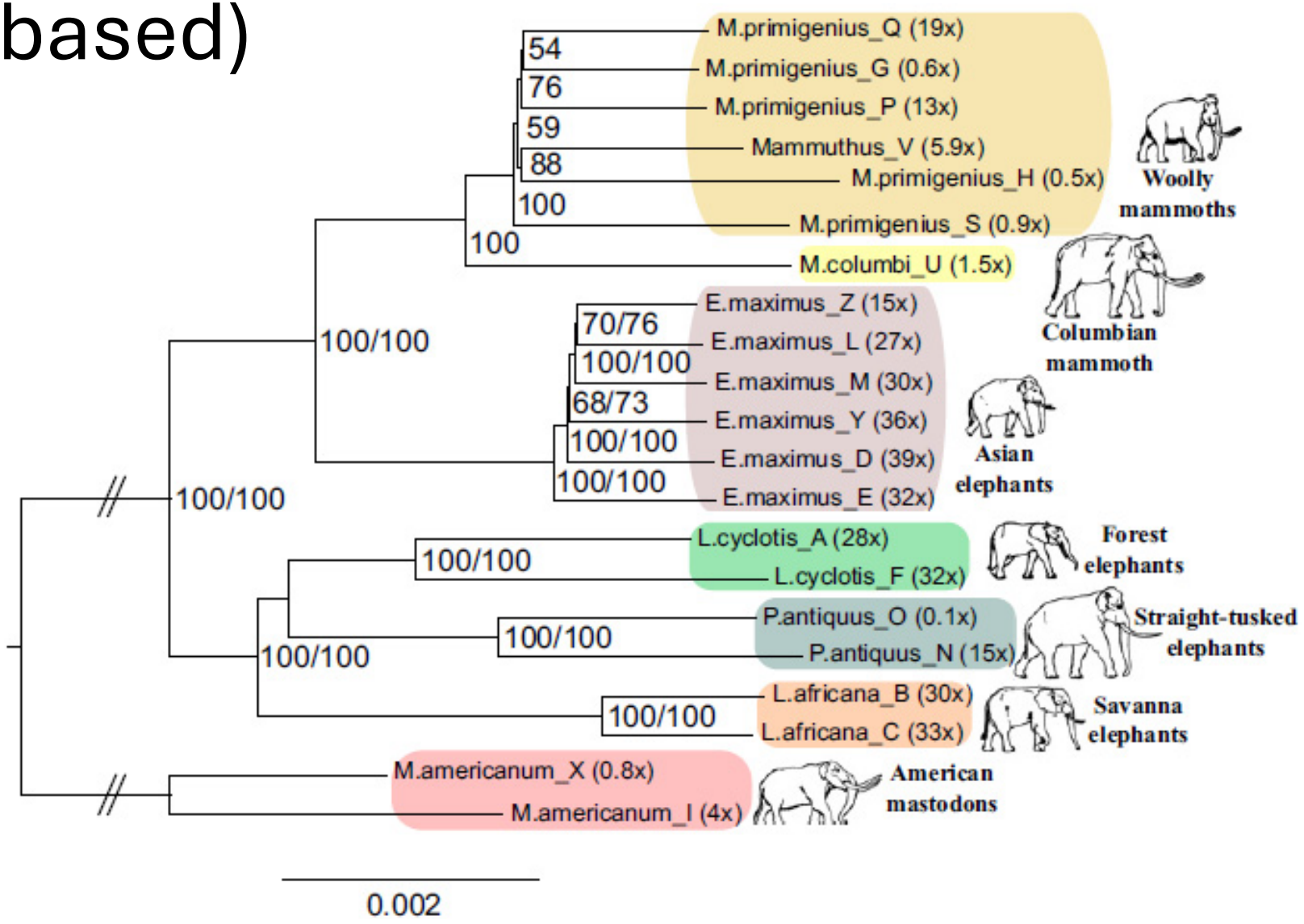

Genealogy 2




Genealogy 3

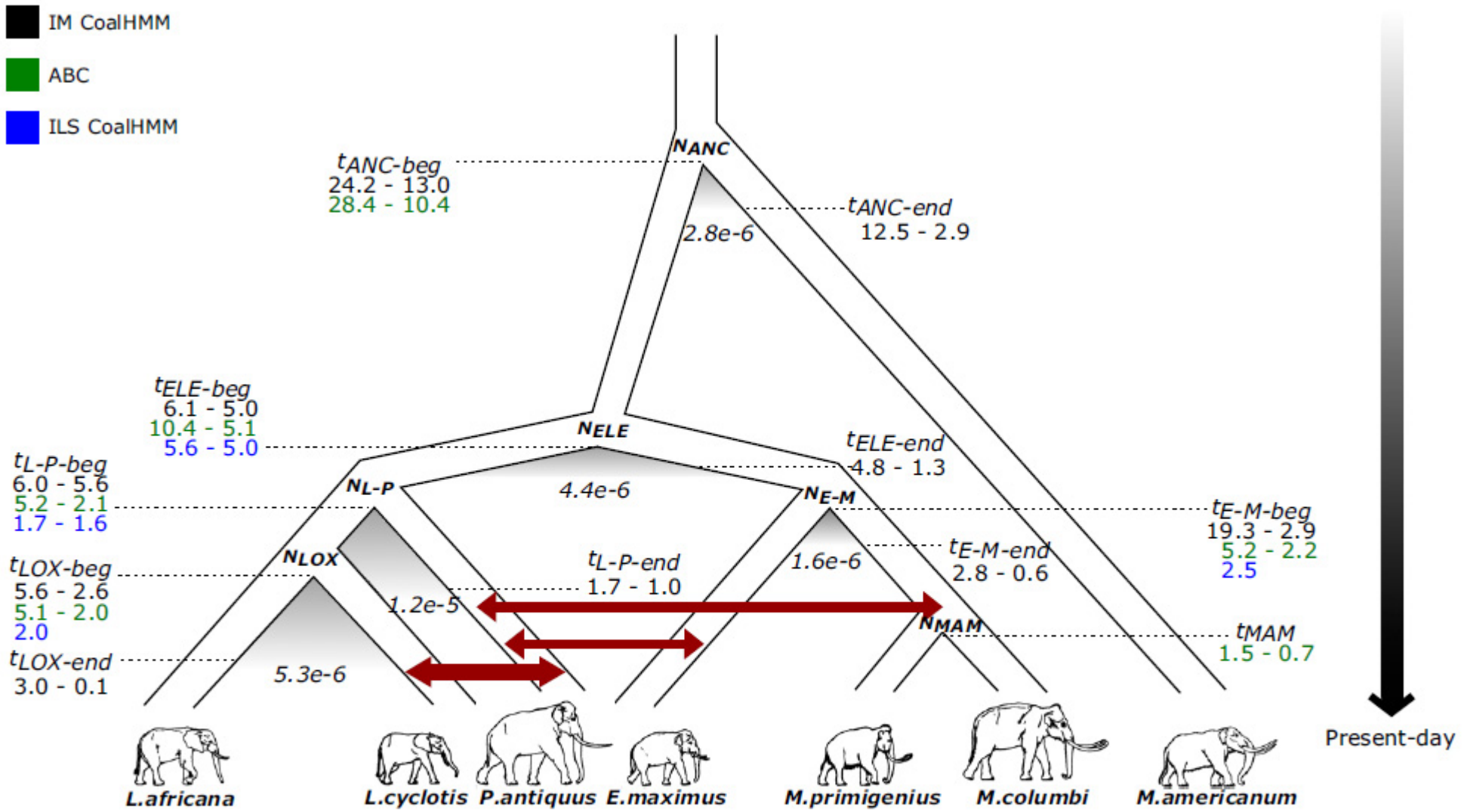


# Mammoths – Neighbor-joining tree – whole genome (distance-based)

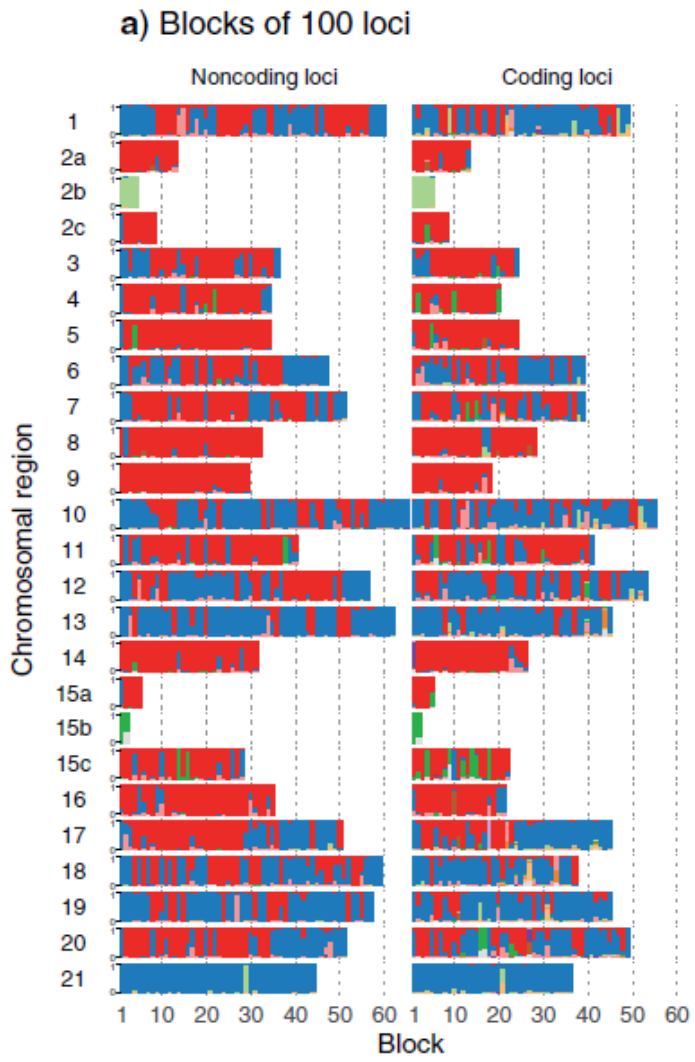



Palkopolou et al. 2018

# Mammoths – species tree with gene flow




Palkopolou et al. 2018




*Heliconius* species

***era**to*

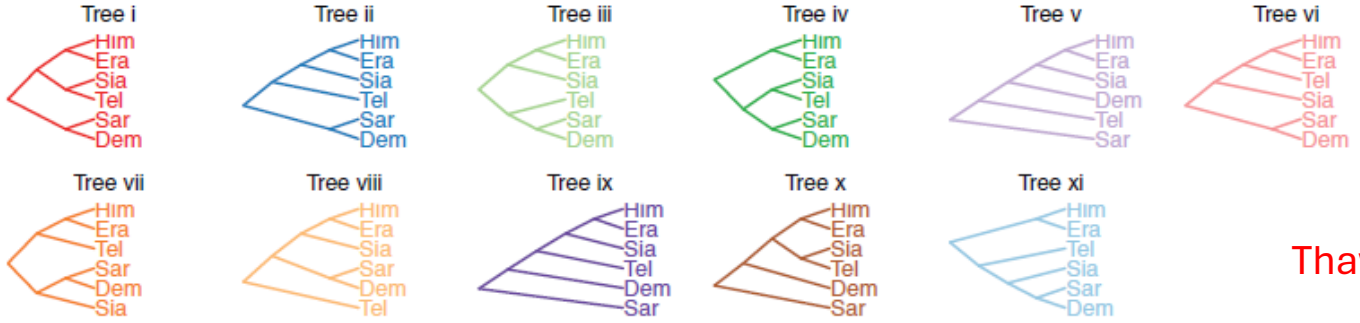
***him**era*

*hecale**sia***

***tel**esiphe*

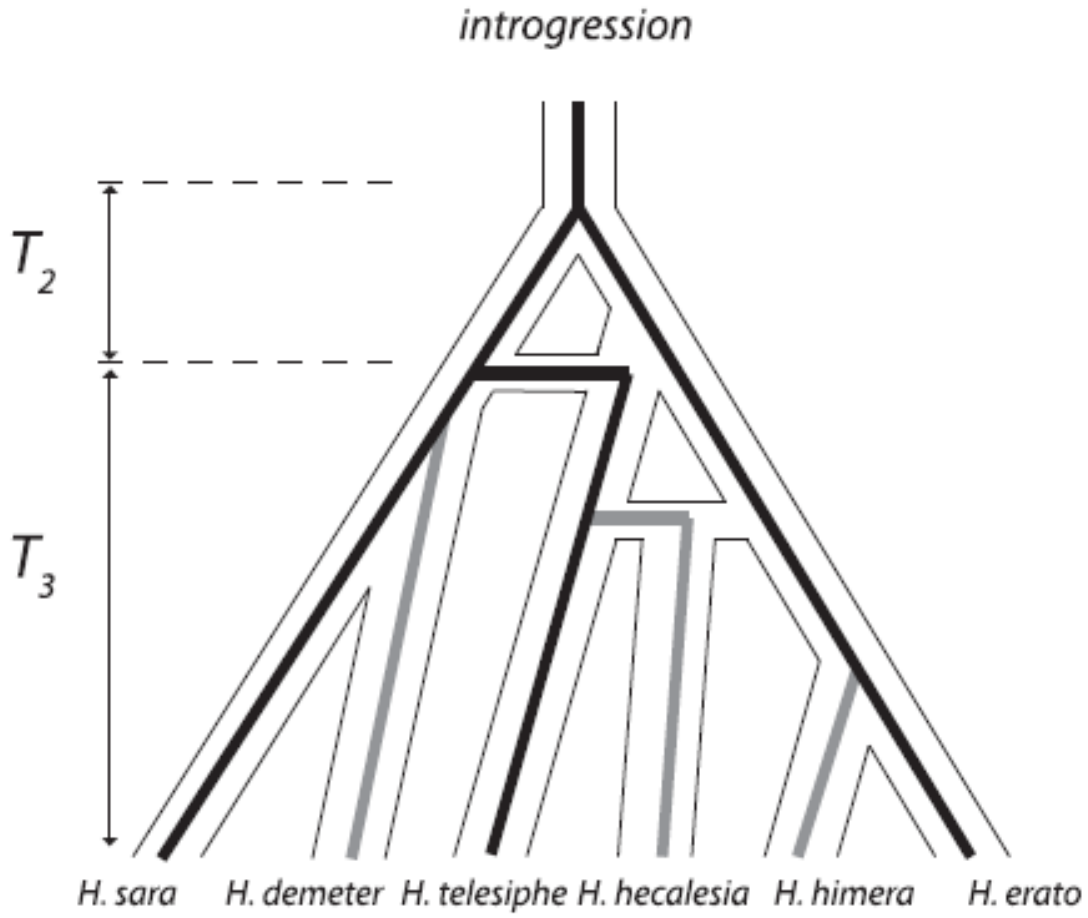
***dem**eter*

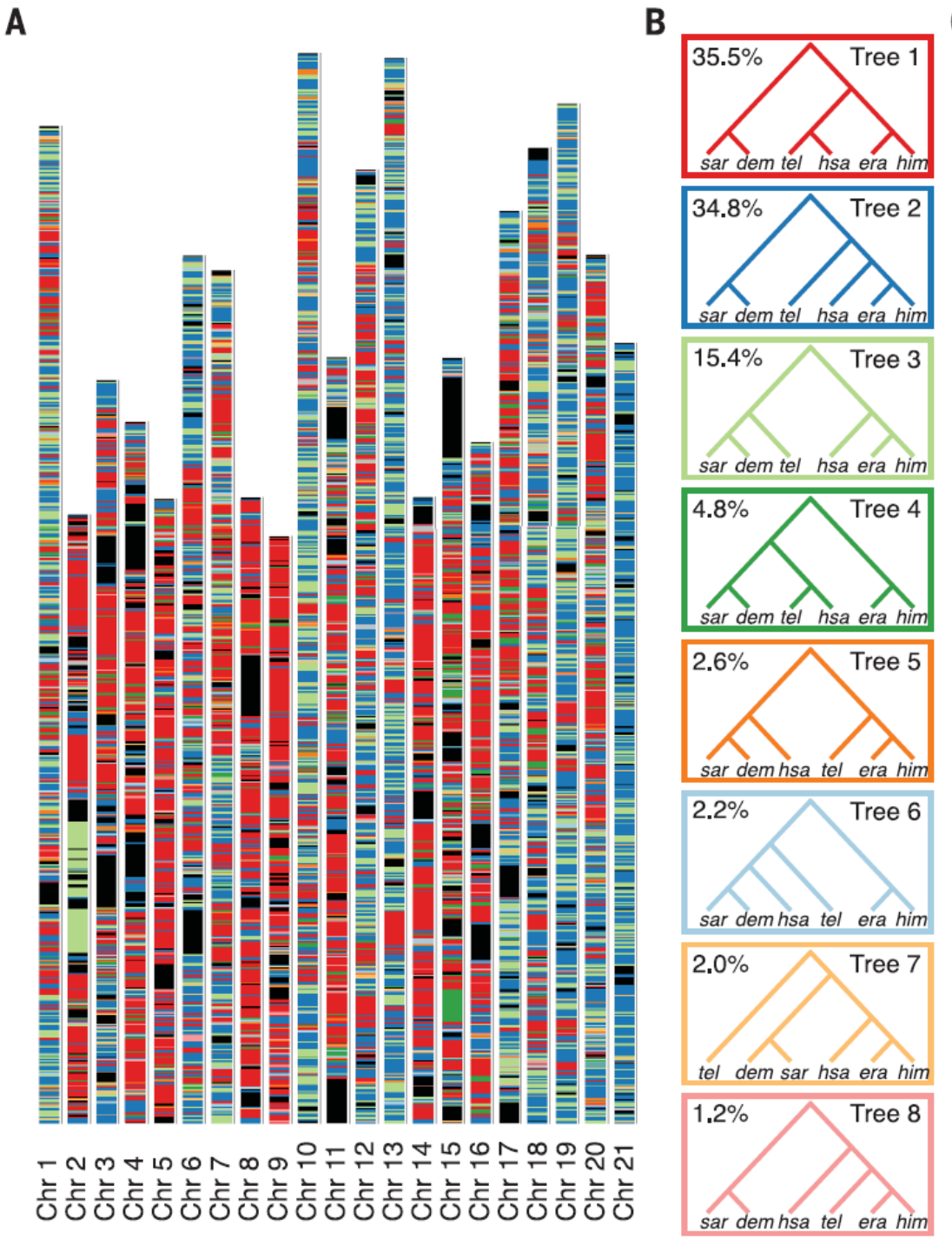
***sar**a*



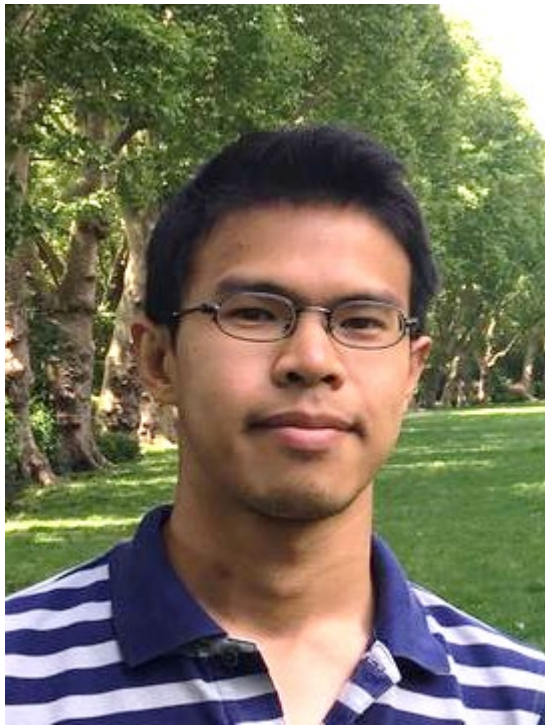

Thawornwattana et al. 2022

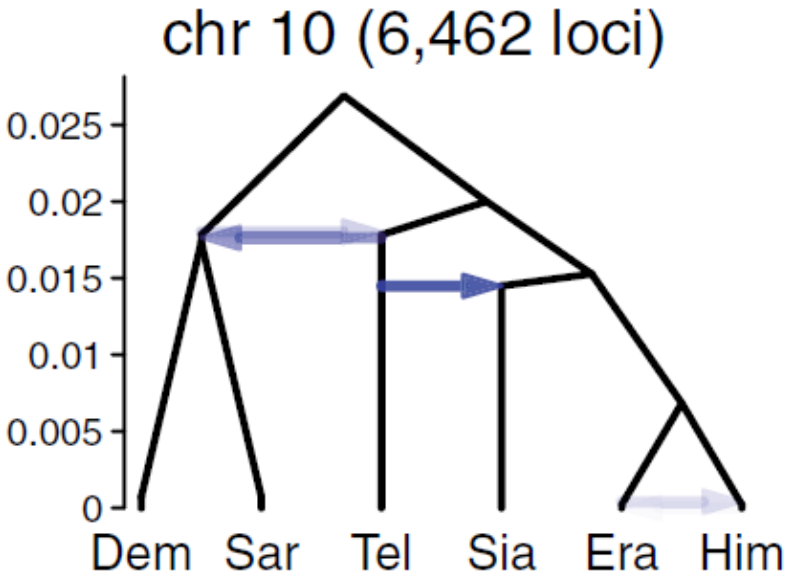
# Heliconius whole genome trees – with gene flow

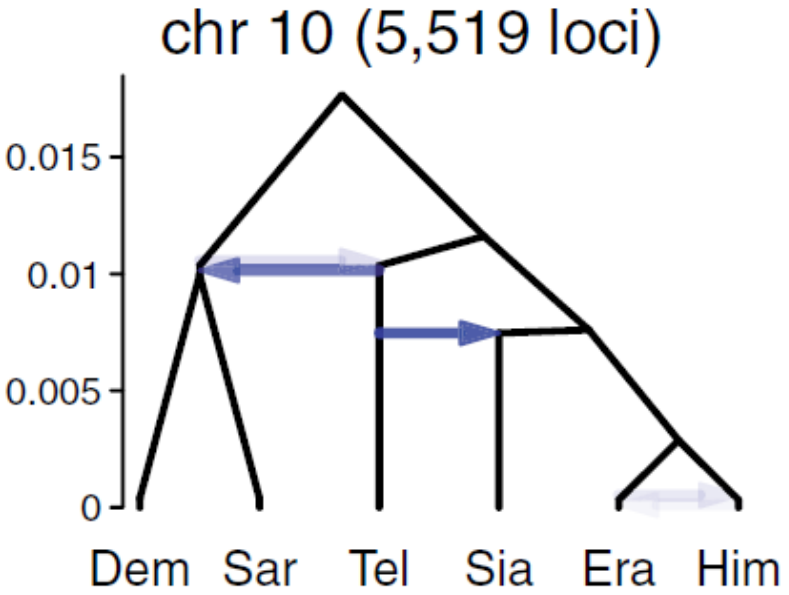


# *Heliconius* – species tree with gene flow






Thawornwattana et al. 2022

# Phylogenetics and tree thinking

- The “cladistics revolution” – monophyly
- The DNA revolution – gene trees
- Methods: parsimony, distance-based, likelihood, Bayesian
- Species trees and gene trees (problems with ILS)
- The genomics revolution
- Bayesian coalescent-based methods for species trees
- Not a tree, instead a network. We need to model introgression as well!
- Newer methods for whole genomes…

# The nature of scientific inference

“I’m sure this is true”
“This has to be wrong”
“It is likely that...”
“This seems most probable to me”

All of inference about the world is likely to be based on probability; it’s statistical

(Except divine revelation!)

# Models and hypotheses in statistical inference

*Models* consist of a set of interrelated parameters of interest. They are assumed to be true for the purposes of the particular test or problem

e.g. let's assume height in humans to be normally distributed with true mean, $\mu$, and true variance, $\sigma^2$.

*Hypotheses* are particular sets of "parameter values" that are the focus of interest in estimates or tests

e.g. estimates of the mean, $m$, and variance, $s^2$

The problem of inference: choose the best hypotheses based on the data

# Data is typically discrete

... Counts of things

... Measurements to nearest mm, 0.1$^{o}$C

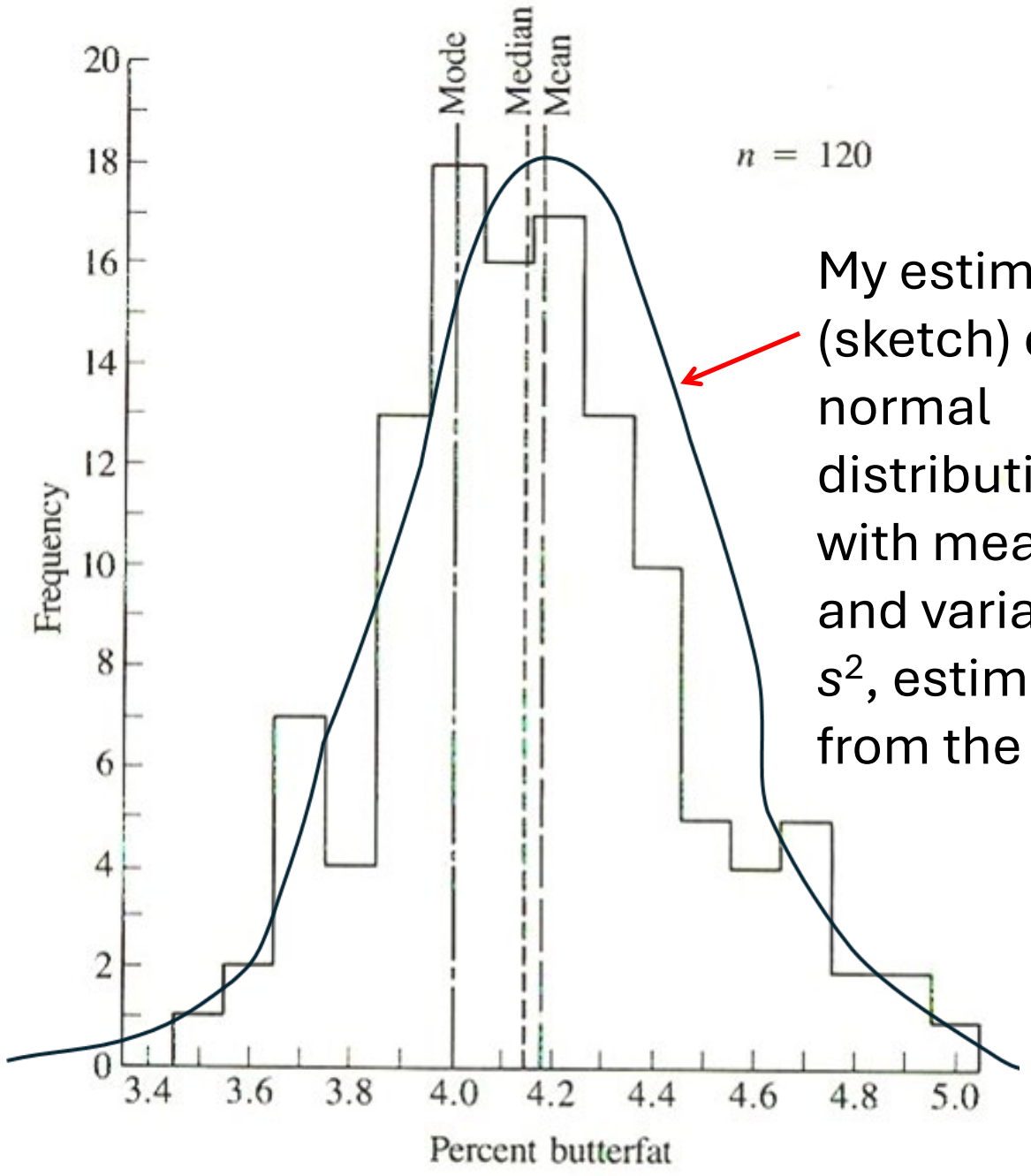
Data consists of finite # of observations

Models, hypotheses can be discrete too, or continuous. Models and hypotheses may be finite, or infinite in scope.

*A good method of inference should take the discreteness of data into account* when we analyse the data, even if the model is not.

*Many analyses, particularly frequentist, don't!*

# For example, milk fat in cow milk




My estimate (sketch) of a normal distribution with mean, *m*, and variance, $s^2$, estimated from the data

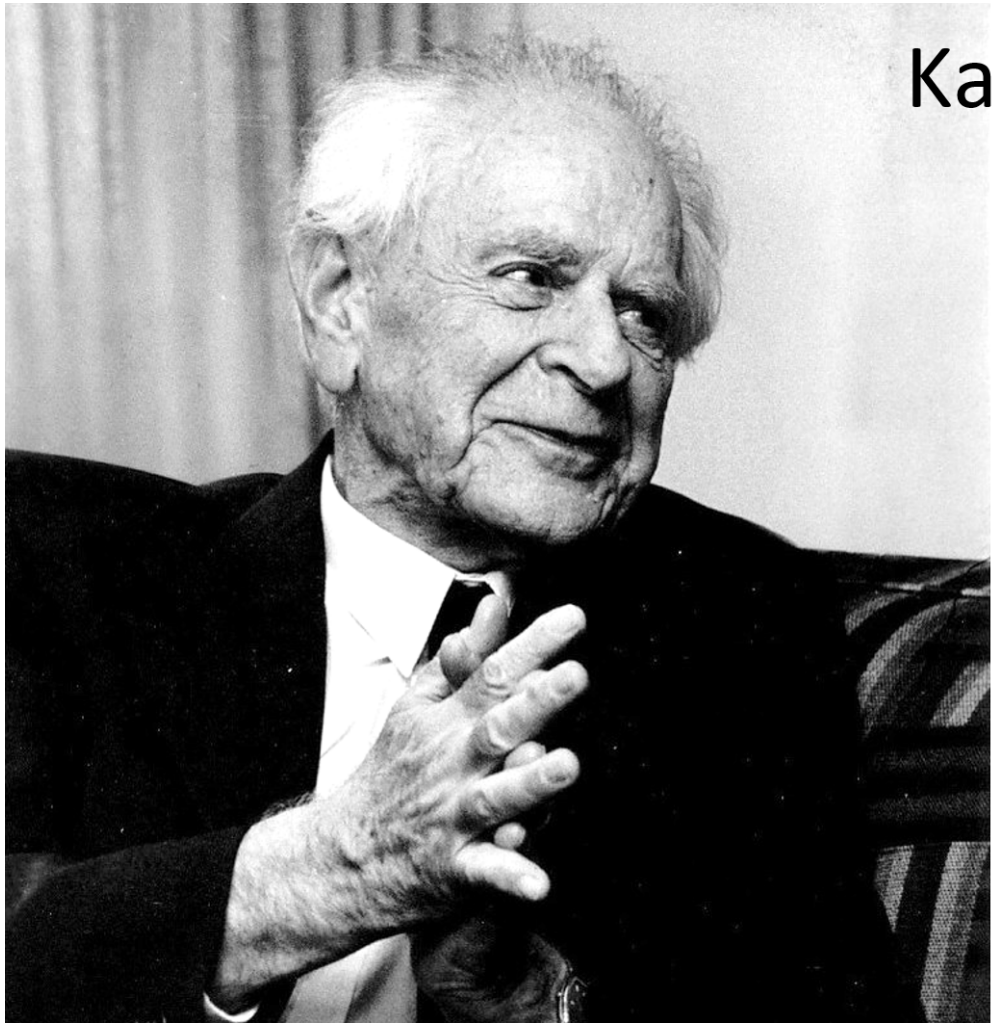
From Sokal & Rohlf 1981, Biometry, p. 47

# Null hypotheses in biological statistics

We are often taught in biology a simplistic kind of "Popperian" approach to science, to falsify simple hypotheses.

We try to test the "null hypothesis"

Physics-envy?



Karl Popper

1934 – 1994: "The criterion of the scientific status of a theory is its falsifiability, .. refutability, or testability"

1972: *Objective Knowledge: An Evolutionary Approach*

# Estimation is primary

Many statisticians today, since the 1970s (e.g. Anthony Edwards) instead argues that we should turn this argument on its head

Estimation of the parameters of a model can lead to testing of an infinitude of hypotheses, *including* the null hypothesis

It seems obvious that we should use some sort of probability measure when making scientific inferences, or estimation. But what sort of probability?

# The three philosophies

- What is scientific inference?
- Three philosophies of statistical inference:
  1. Frequentist (probability in the long run)
  2. Likelihood (probability of data given a hypothesis)
  3. Bayesian (posterior probability of a hypothesis)
- Common ground: Opposing philosophies agree (approximately), in results many problems.

Historical order was actually:

Bayesian, frequentist, likelihood

I start with frequentism, because often taught…

## Probability and its symbols

All scientific inference depends on some kind of probability. What is it? Philosophical problems…

With coin-tossing, dice, or cards, we have an intuitive idea of what we mean. Symbols used:

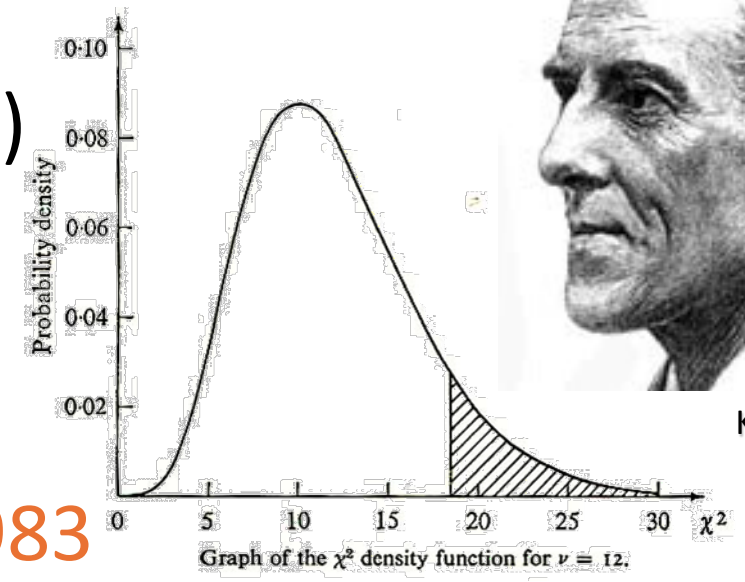
$P(A)$ = prob. of event $A$. $P(B)$ prob. of event $B$
$P(A|B)$ = probability of event $A$ **given** $B$

If $A$ and $B$ independent: $P(A \text{ or } B, \text{ or both})$? $P(A \text{ and } B)$?

$P(A \cup B)$ $P(A \cap B)$

# 1. Frequentism, significance testing, *P*-values

Perfected 1900-1930
(Neyman, Pearson, Fisher)
*e.g.* $\chi^2$ test, or *t*-test
$\chi^2 = 5.28$, d.f. = 1;
or $t = 3.92$, d.f. = 10
We find $P<0.05$; $P = 0.00983$



Graph of the $\chi^2$ density function for $\nu = 12$.


K. Pearson

This is "tail probability" or "probability in the long run" of getting results at least as extreme as the data under the *null hypothesis*. "*P* value."

# Philosophical problems with frequentist approach

We only have one set of data. But *P* values seem to imagine the experiment done a very large number of times. (Randomization tests,such as bootstrapping or jackknifing similar)

We often assume the data come from a continuous distribution;

e.g. $\chi^2$ tests on count data, $\sum(O\text{-}E)^2/E$

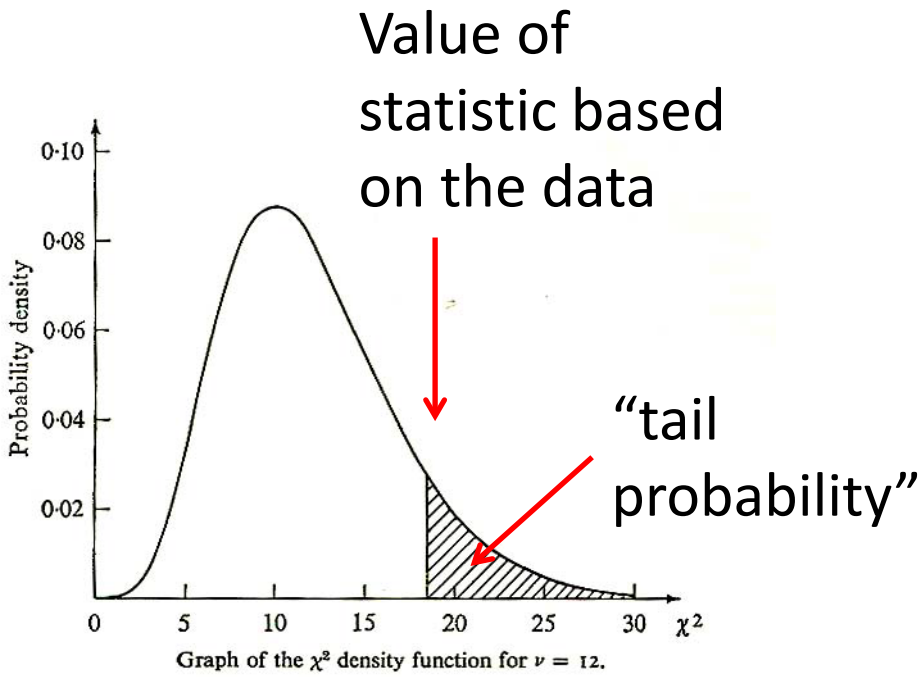
Encourages testing of null hypothesis

# *P* - values

$P > 0.05$ or

$P = 0.064215$

The *P*-value is the probability of obtaining the observed sample results, or more extreme results, when the null hypothesis is true




For example:

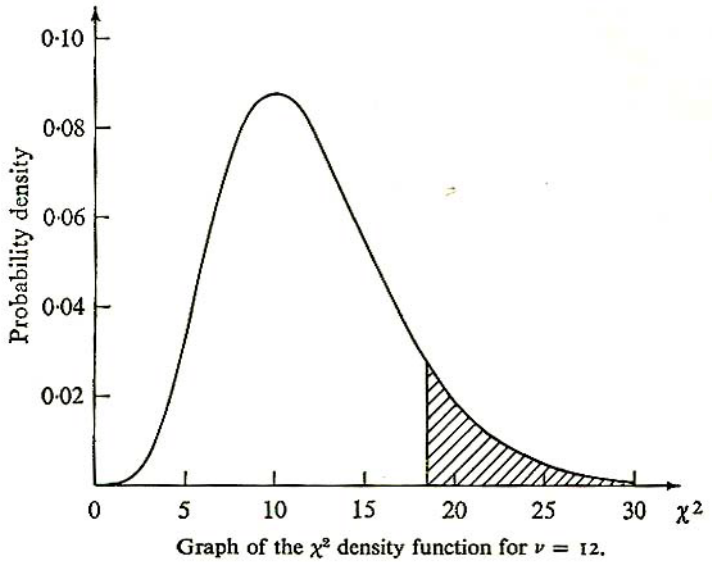
"No effect of heredity on IQ"

# *P* - values

"Null hypotheses"

*P*-values are "tail probabilities"

"What the use of *P* implies, therefore, is that a hypothesis that may be true may be rejected because it has not predicted observable results that have not occurred" Jeffreys 1961



Graph of the $\chi^2$ density function for $\nu = 12$.

e.g. No effect of heredity on IQ …

… a null hypothesis that is almost certainly **not** true

# Problems of frequentism

1) We only have a single set of data; yet frequentism seems to imagine the experiment done a very large number of times under a null hypothesis

2) Assuming that data come from a continuous distribution; we know it doesn't: e.g. "$\chi^2$ tests" on count data, $X^2 = \Sigma(O\text{-}E)^2/E \approx \chi^2$

3) Estimation usually more useful than test of null hypothesis, anyway

4) The null hypothesis, if not rejected, is often unbelievable

5) Sequential tests are powerless: need a way of integrating all the tests together

# Alternatives to frequentism

- Frequentism: P-values, “Probability in the long run”

- Two alternative probability measures of scientific inference:
  - Likelihood (RA Fisher 1920s, Edwards 1972)
    “The probability of the data given a hypothesis”
    (can be viewed as a simplified form of Bayesian probability – see below)

  - Bayesian Probability (Thomas Bayes 1763, Marquis de Laplace 1820)
    “The probability of a hypothesis given the data”
    “The posterior probability”

# 2. Likelihood

The *likelihood* of a hypothesis (*H*) after doing an experiment or gathering data (*D*) is proportional to the *probability of the data given the hypothesis*

$$L(H|D) = kP(D|H)$$

Probabilities add to 1 for each hypothesis (by definition), but do not add to 1 across *different* hypotheses – "Likelihood" is therefore *not* a kind of probability, although closely related to probability, of course!
*k* is an arbitrary constant

The hypotheses are variable; the data remain the same!

# The Law/Axiom of Likelihood

"Within the framework of a statistical model, a particular set of data supports one statistical hypothesis better than another if the likelihood of the first hypothesis on the data exceeds the likelihood of the second hypothesis. All the information which the data provide is contained in the likelihood ratio"

$$Likelihood\ Ratio = \frac{P(D \mid H_1)}{P(D \mid H_2)} > 1$$

# Method of support

*Support* for one hypothesis against another is defined as the natural logarithm of the likelihood ratio

$$Support = \log_e \frac{P(D \mid H_1)}{P(D \mid H_2)}$$

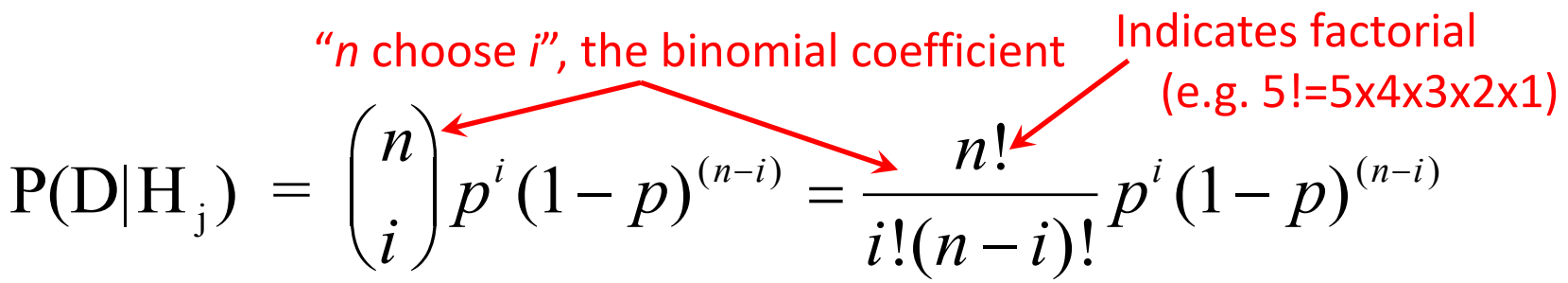
$$= \log_e P(D \mid H_1) - \log_e P(D \mid H_2)$$

# Example: binomial distribution

Supposing we are interested in estimating the frequency of a SNP in a sample of human genomes:

| **A** | **G** | **Total alleles** |
|---|---|---|
| 2 | 8 | 10 |
| *i* | *(n-i)* | *n* |

This is a problem that fits the binomial theorem:



# Likelihood approach

To get the support for hypotheses, we need to calculate the "log likelihood ratio" (LR):

$$Support, \ln L = \log_e \frac{P(D \mid H_1)}{P(D \mid H_2)}$$

$$\mathrm{P(D|H_j)} = \binom{n}{i} p^i (1-p)^{(n-i)} = \frac{n!}{i!(n-i)!} p^i (1-p)^{(n-i)}$$

Note! The binomial coefficient depends only on the data (D: *n,i*), not on the hypothesis (H: *p*). Thus Binomial coeffs. $\binom{n}{i}$ cancel! No need to calculate the tedious constants! Just need the $p^i(1-p)^{(n-i)}$ terms

# Likelihood and the binomial

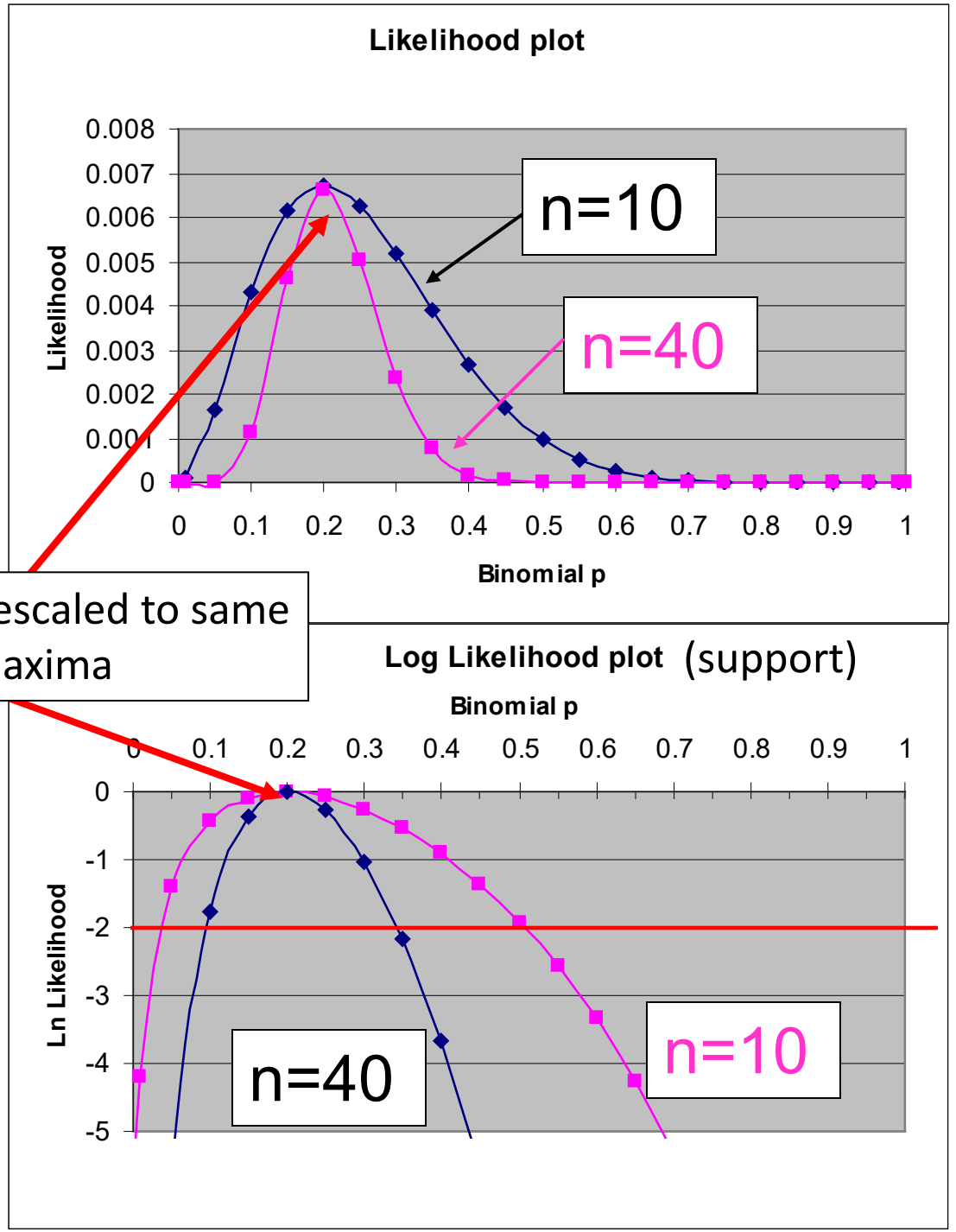
| **Binomial probability using likelihood** | | | sample size n= 10 | | "successes" i= 2 | |
|---|---|---|---|---|---|---|
| | Likelihood/B | | ln likelihood | | ln likelihood ratio (relative to $p$ = 0.2) | |
| Hj = p | p^i(1-p)^(n-i) | | | | | |
| 0 | 0 | | #NUM! | | #NUM! | |
| 0.001 | 1.002E-06 | | -13.81351 | | -8.36546 | |
| 0.01 | 9.22745E-05 | | -9.290743 | | -4.19635 | |
| 0.05 | 0.001658551 | | -6.401811 | | -1.39779 | |
| 0.1 | 0.004304672 | | -5.448054 | | -0.44403 | |
| 0.15 | 0.006131037 | | -5.094391 | | -0.09037 | |
| ➔ 0.2 | 0.006710886 | | -5.004024 | *=max (i=2)! | 0 | |
| 0.25 | 0.006257057 | | -5.074045 | | -0.07002 | |
| 0.3 | 0.005188321 | | -5.261345 | | -0.25732 | |
| 0.35 | 0.003903399 | | -5.545908 | | -0.54188 | |
| 0.4 | 0.002687386 | | -5.919186 | | -0.91516 | |
| 0.45 | 0.001695612 | | -6.379711 | | -1.37569 | |
| 0.5 | 0.000976563 | | -6.931472 | | -1.92745 | |
| 0.55 | 0.000508658 | | -7.583736 | | -2.57971 | |
| 0.6 | 0.00023593 | | -8.351977 | | -3.34795 | |
| 0.65 | 9.51417E-05 | | -9.260143 | | -4.25612 | |
| 0.7 | 3.21489E-05 | | -10.34513 | | -5.34111 | |

# Likelihood and the binomial

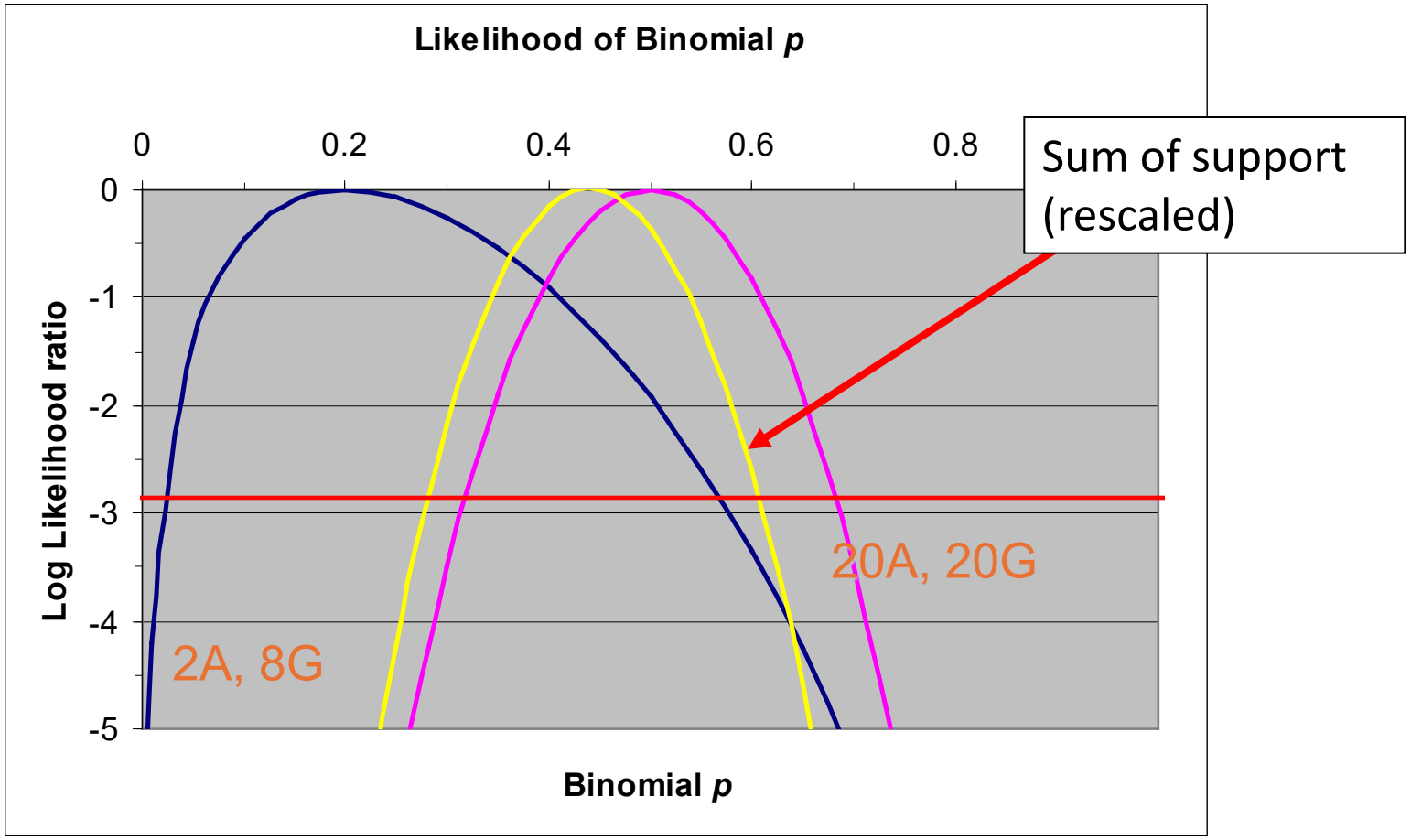
The support curves (rescaled here), with data = 2/10 or data = 8/40, give measures of belief in the continuously variable hypotheses

Edwards: 2 lnL units below max. lnL viewed as "support limits" (equivalent to approx. 2 standard errors, or ~95% probability density in the frequentist approach)

$\log_e LR=2$ implies $LR=e^2$, the best is 7.4x as good.



# Sum of support from different experiments

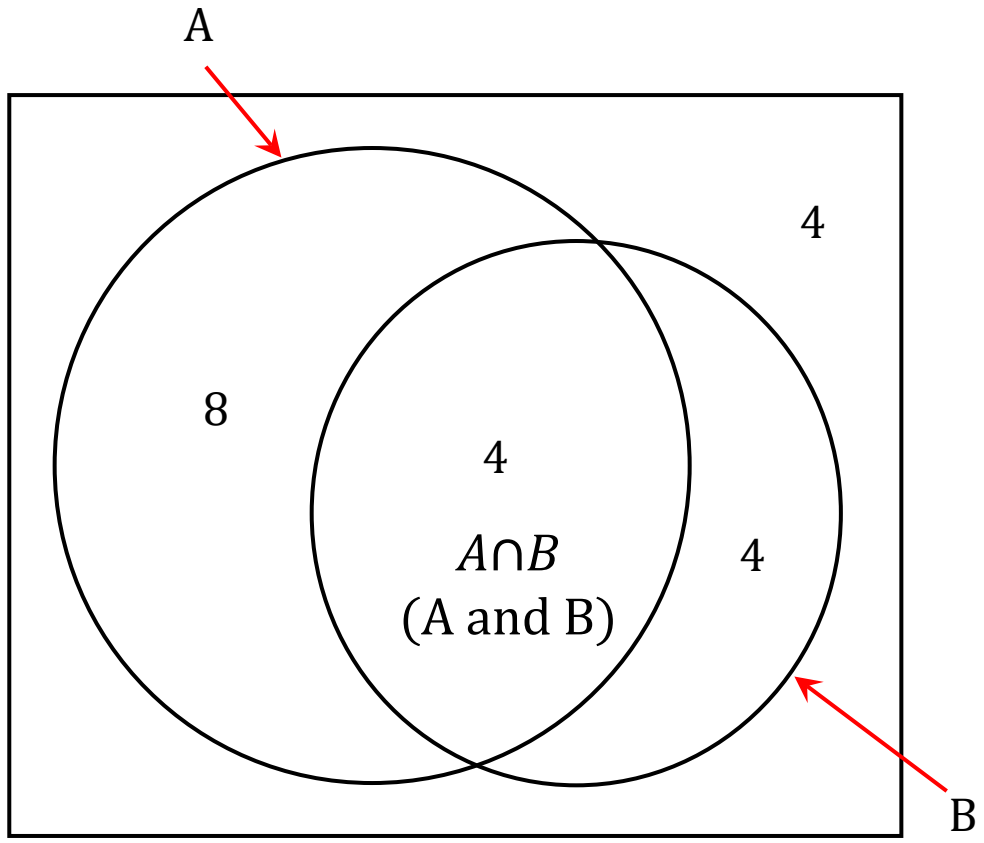



Support provides a way to adjudicate between data from different experiments. Example shows SNP frequencies.

# 3. Bayes' Theorem: Bayesianism

$$P(A \mid B) = \frac{P(B \mid A)P(A)}{P(B)}$$

Named after its inventor, Reverend Thomas Bayes in 18th Century England. Led by Bayes and Laplace, Bayes' theorem and "Posterior Probability" has come to be used in a system of inference …




*Product rule*:

If A and B are independent, then $P(A|B) = P(A)$ and $P(B|A) = P(B)$, and

$$P(A \cap B) = P(A)P(B) = \frac{4.8}{20}$$

(total = 20)

Here, A and B are not independent ...

*Product rule*: $P(A \cap B) = P(A|B)P(B) = P(B|A)P(A)$

*Example*: $\frac{4}{20} = \frac{4}{8} \times \frac{8}{20} = \frac{4}{12} \times \frac{12}{20}$

*Bayes Theorem*: $P(A|B) = \frac{P(B|A)P(A)}{P(B)} \Rightarrow P(H|D) = \frac{P(D|H)P(H)}{P(D)=k}$

Posterior probability

Likelihood

"Prior" probability

*N.B.* [$P(A|B)$ *means* $P(A$ *given* $B)$]

# Inference from Bayes' Theorem

$$P(A \mid B) = \frac{P(B \mid A)P(A)}{P(B)}$$

Rev. Thomas Bayes this theorem could be used for inference. (Anthony Edwards argues that this did not include many types of inference now attempted using "Bayesian methods")

$$P(H \mid D) = k.P(D \mid H)P(H)$$

Posterior Probability

Likelihood

"Prior probability" of the hypothesis

*H*: hypothesis; *D*: data, *k*: a "normalizing constant" equivalent to "the probability of the data")

# *Bayes' Theorem*

as a means of inference

$$\frac{P(H_1 \mid D)}{P(H_2 \mid D)} = \frac{k.P(D \mid H_1)P(H_1)}{k.P(D \mid H_2)P(H_2)}$$

If the prior is "uniform", $P(H_1)=P(H_2)$

$$\frac{P(H_1 \mid D)}{P(H_2 \mid D)} = \frac{P(D \mid H_1)}{P(D \mid H_2)}$$

The ratio of posterior probabilities collapses to
... a likelihood ratio!

# In practice

In well-behaved applications, all three approaches tend to support (or reject) similar hypotheses.

Significance tests are justifiable by appealing to likelihood ratios – tail probability low when likelihood ratio (itself often proportional to relative Bayesian probability) is high. And vice-versa.

In complex estimation problems (e.g. GLM ), where we test for "significance" of $\nu$ extra parameters, we use likelihood with a the chi-square approximation:

$$-2\log_e LR = \text{"deviance"} \approx \chi_1^2$$

This interpretation employs a frequentist approach.