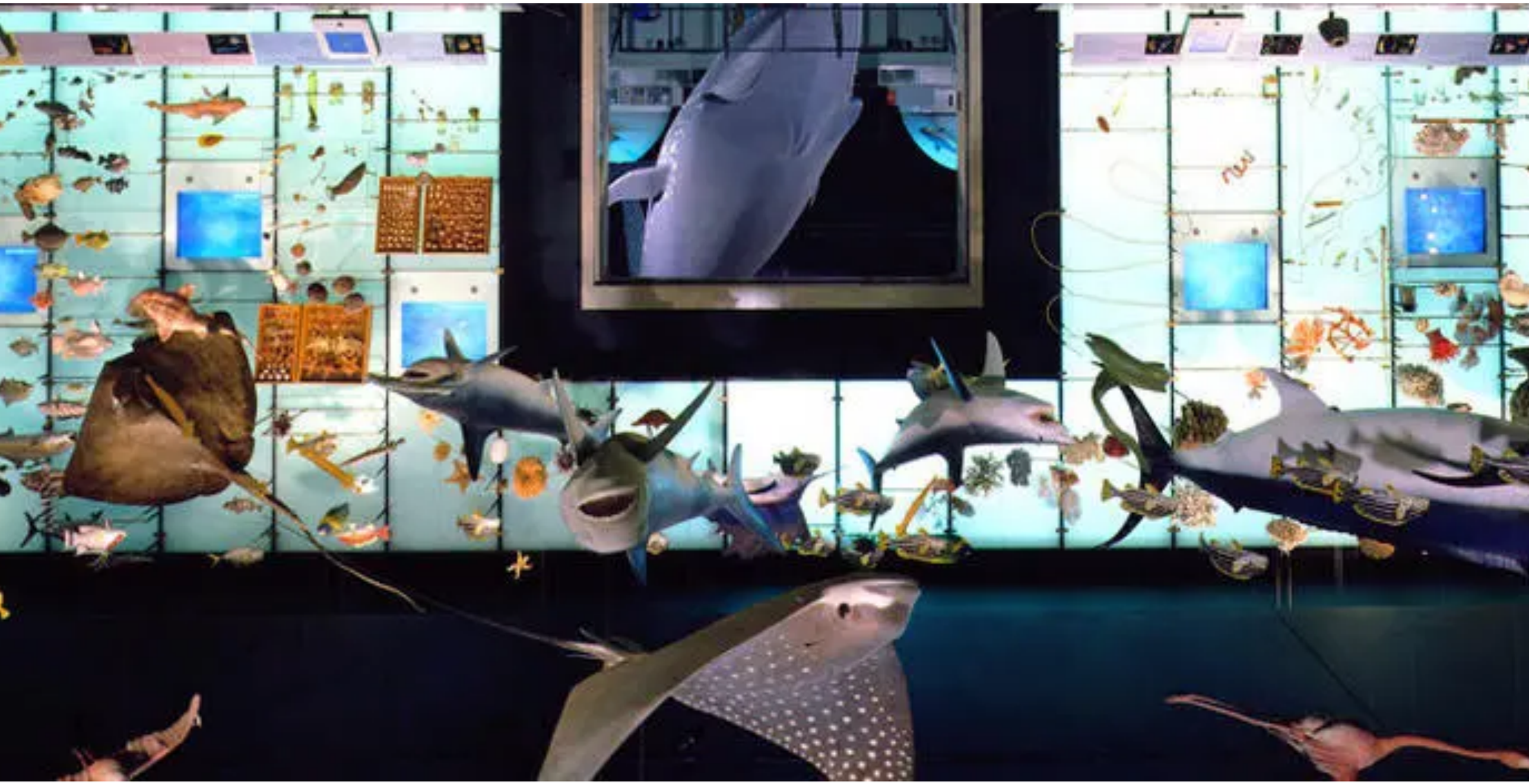
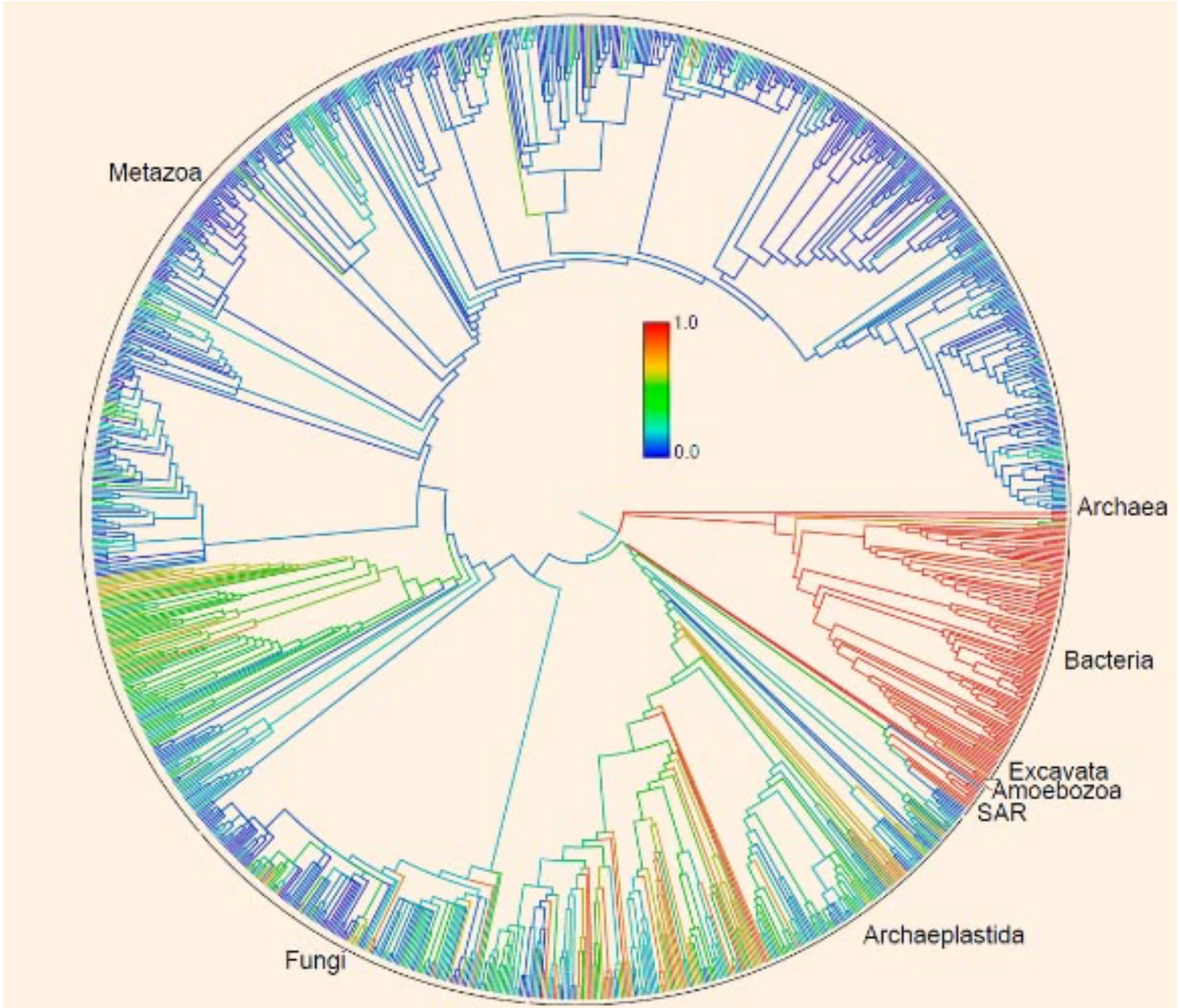
Molecular Ecology and Systematics

Lecture 6

JA Drew




Metazoa
1.0
0.0
Archaea
Bacteria
Excavata
Amoebozoa
SAR
Archaeplastida
Fungi

Systematics

- The goal of systematics is to have phylogeny (underlying evolutionary order) reflected in our classification
    - Requires taxonomy - the naming and ordering of species
    - Phylogenetics - understanding the evolutionary history of species

Species

Panthera pardus (leopard)

Mephitis mephitis (striped skunk)

Lutra lutra (European otter)

Canis familiaris (domestic dog)

Canis lupus (wolf)

Genus

Panthera

Mephitis

Lutra

Canis

Family

Felidae

Mustelidae

Canidae

Order

Carnivora

Systematics relies on taxonomy

- Taxonomy is a way of linking species together, however there can be many different ways of linking groups together, not all of which are related to evolutionary relationship


MORRISVILLE
EST. 1908
SUNY MORRISVILLE


HAWKS
1
New Paltz
STATE UNIVERSITY OF NEW YORK


ESF
State University of New York
College of Environmental Science and Forestry


UNIVERSITY AT ALBANY
State University of New York


STATE UNIVERSITY OF NEW YORK
PLATTSBURGH


SUNY
ADIRONDACK

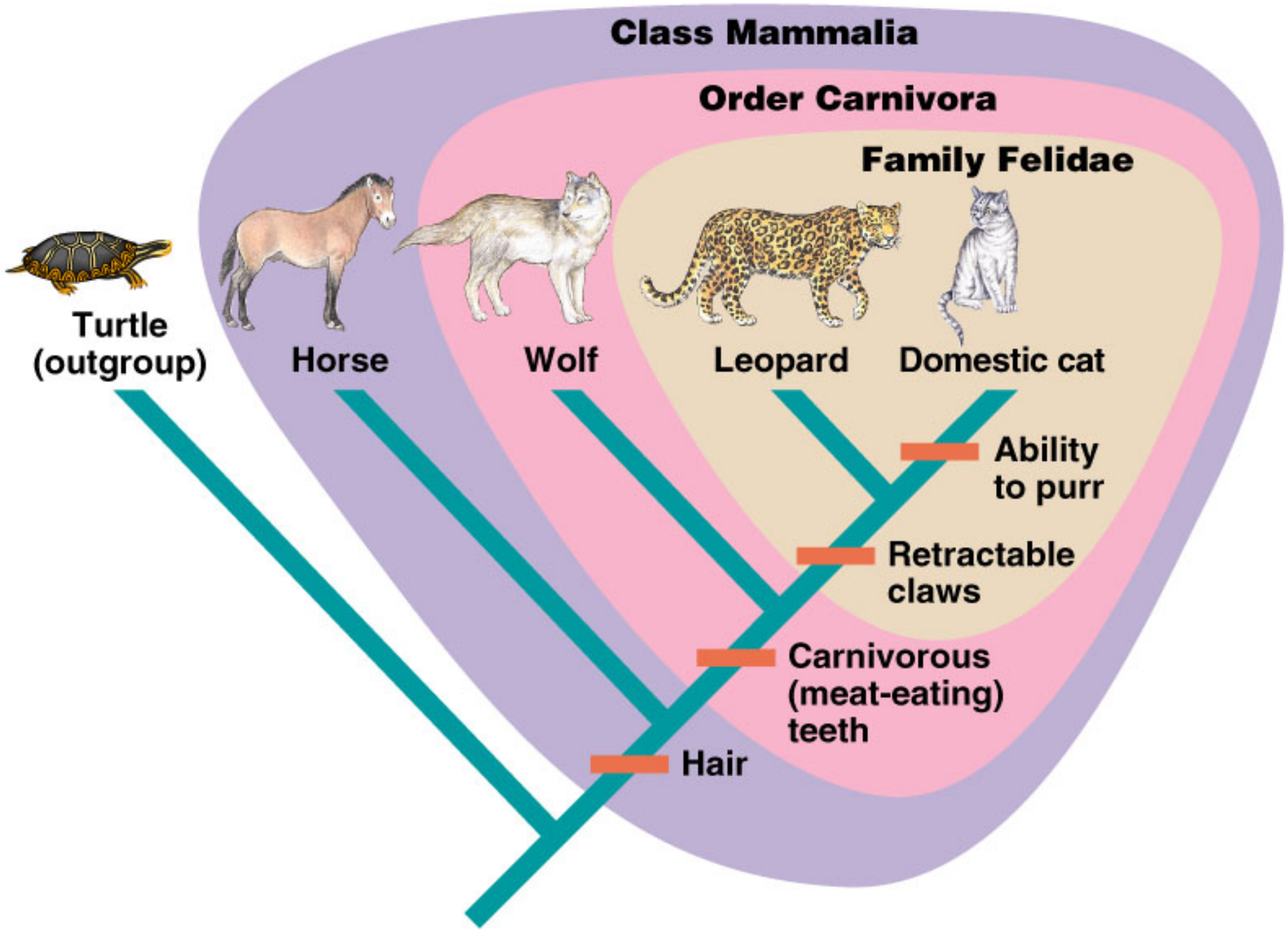

SUNY
Cobleskill


FIT
Fashion Institute
of Technology
State University of New York


SUFFOLK
Suffolk
COUNTY COMMUNITY COLLEGE

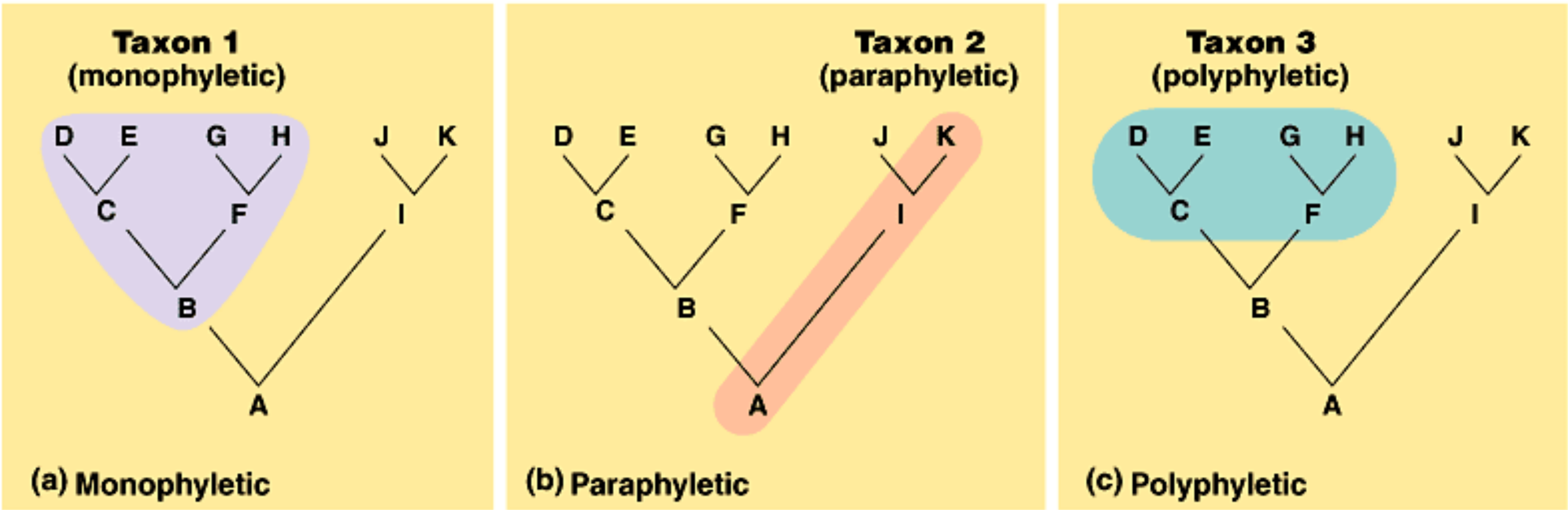
Phylogenetic systematics

- We use a philosophy called cladistics to help organize species into groups that represent a shared evolutionary history.
- This is often expressed as a tree like structure
- Cladistics classifies organisms according to the order in time that the branches arise along a phylogenetic tree without considering the degree of divergence (e.g., how different they are)



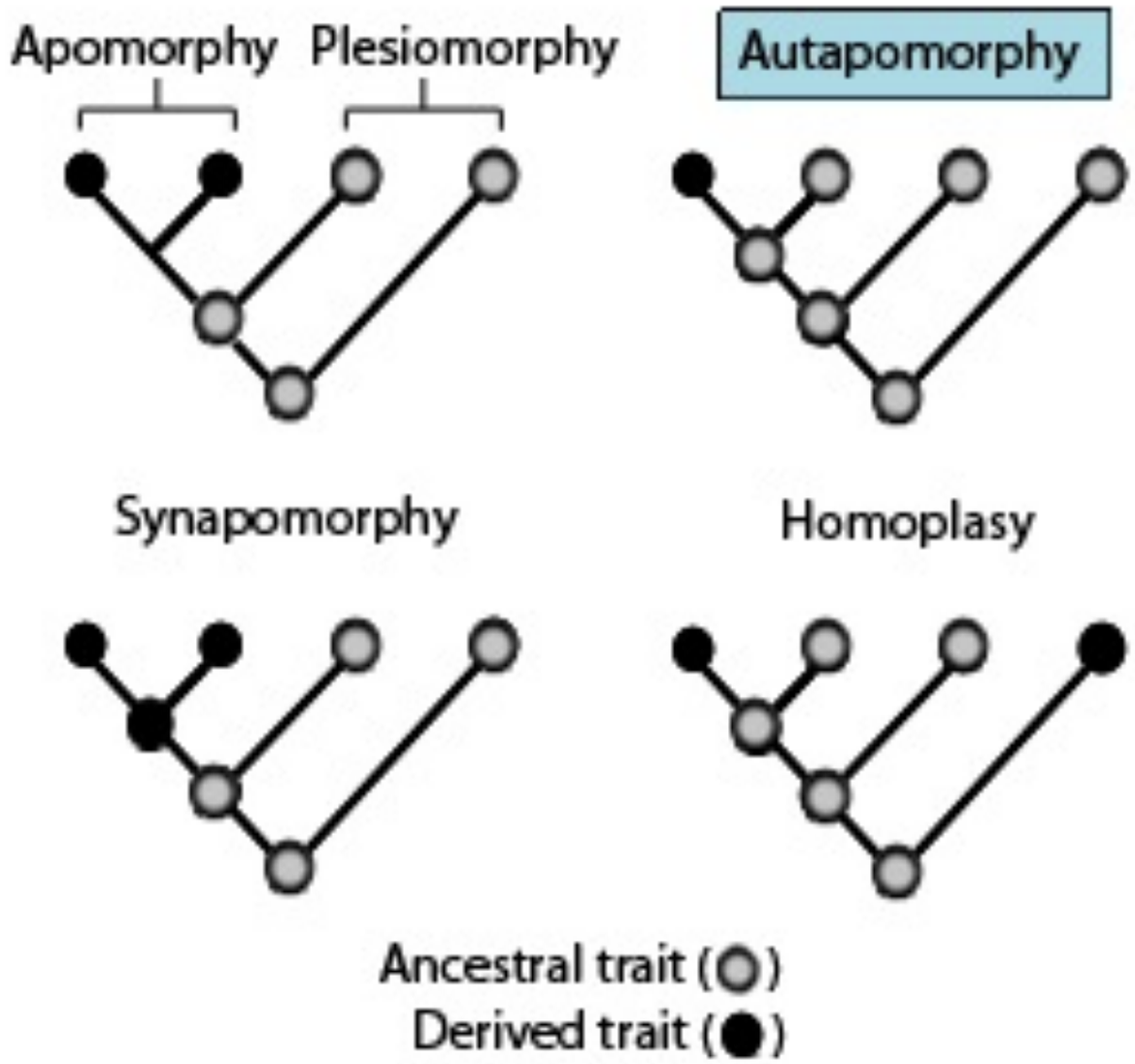
How do we build these trees

- Groups are organized into clades and these clades are united by shared characters
- A fully resolved tree has only monophyletic taxa
- If a group has multiple individuals with different evolutionary histories it is called polyphyletic
- If a group has does not contain all of the lineages of a shared ancestry it is paraphyletic



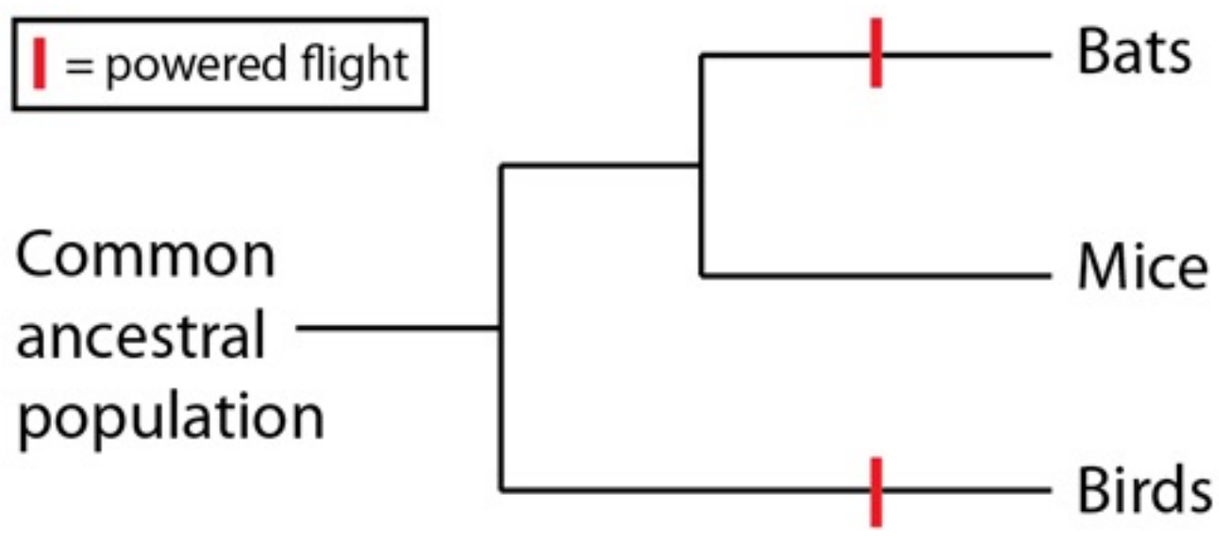
Synapomorphies etc.

- To define the groups we use characters which exist in multiple states
- An apomorphy is a shared derived trait that evolved from an ancestral or plesiomorphic trait
- An autapomorphy is a trait that is found only in one taxa and is phylogenetically uninformative
- Synplesiomorphy a shared ancestral trait


Apomorphy
Plesiomorphy
Autapomorphy
Synapomorphy
Homoplasy
Ancestral trait ( )
Derived trait ( )

The problem with homoplasy

- A character is homoplastic when it is shared for reasons other than common ancestry. Commonly called “convergent evolution”



Molecular systematics

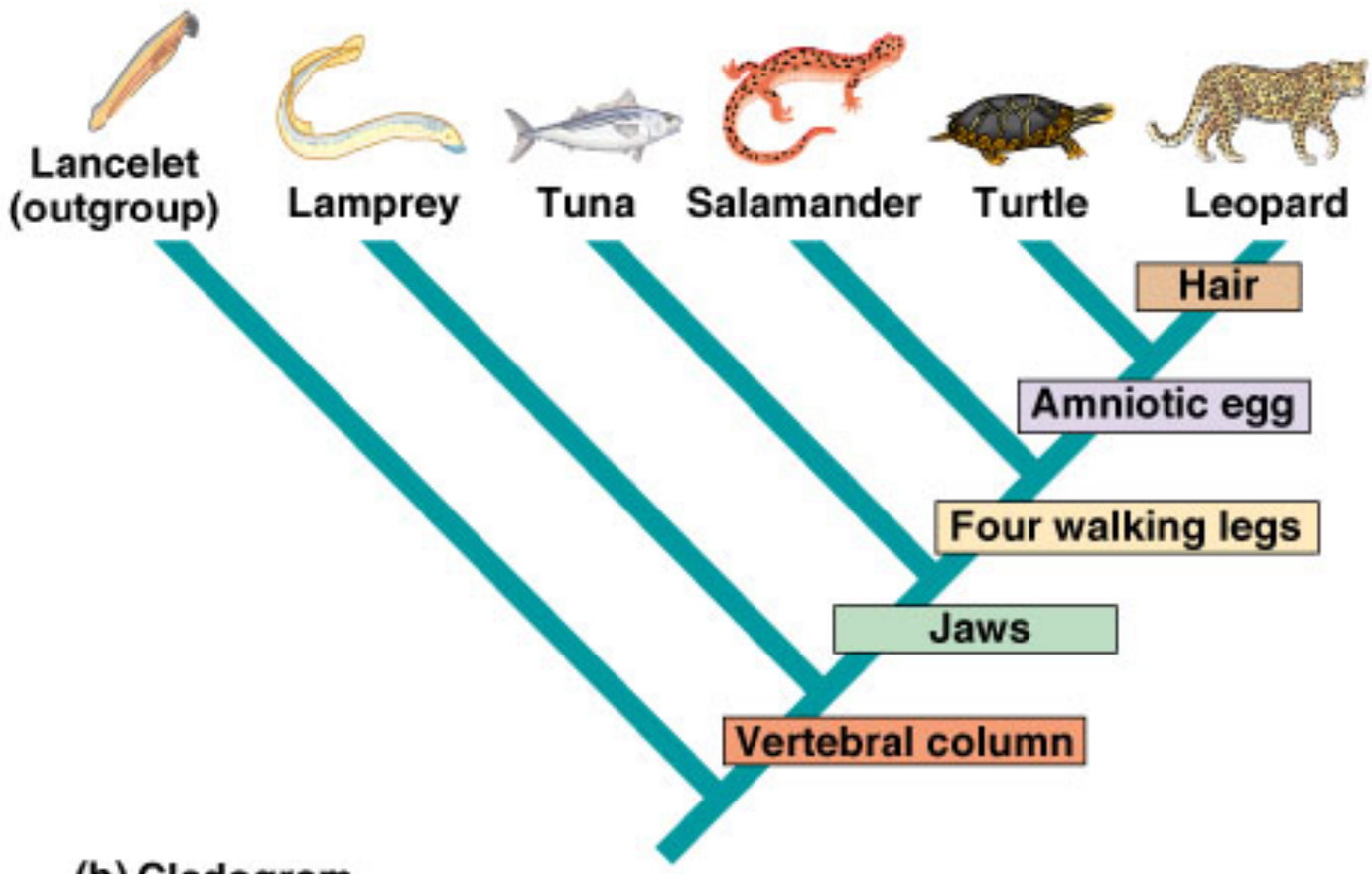
- Phylogenetic trees can be built using any kinds of categories
- (My upcoming DOL lecture on Dinosaurs for a lot more on this)
- However in the context of here we are going to focus on trees that are built by using DNA sequences

Data Matrix with characters

Cladogram built based from characters

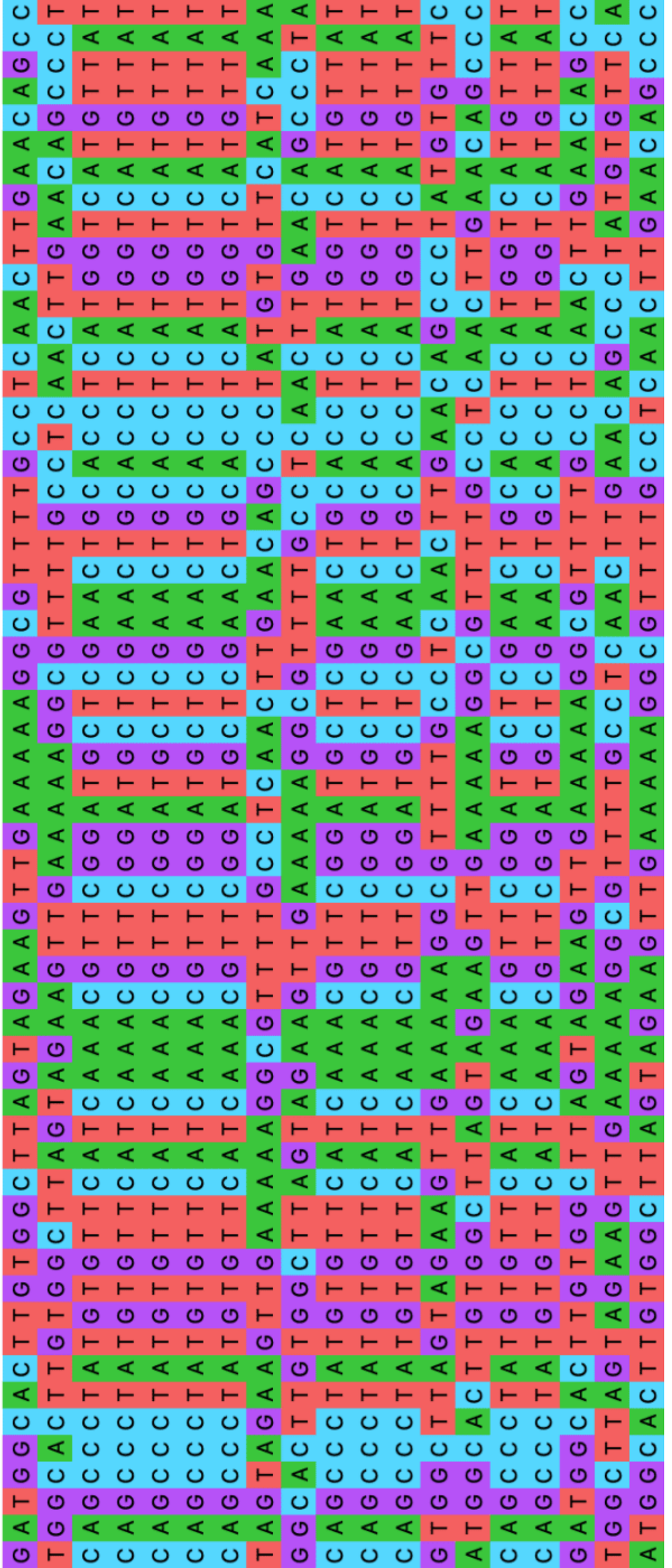
TAXA

CHARACTERS	Lancelet (outgroup)	Lamprey	Tuna	Salamander	Turtle	Leopard
Hair	0	0	0	0	0	1
Amniotic (shelled) egg	0	0	0	0	1	1
Four walking legs	0	0	0	1	1	1
Jaws	0	0	1	1	1	1
Vertebral column (backbone)	0	1	1	1	1	1



(a) Character table

(b) Cladogram



Models of Evolution

- These are used to help understand the observed differences. E.g., is an A->T always the same across the genome? Can we weigh some changes more than others?
- Factors include base pair frequency (human genome is AT rich)
- Transitions and Transversion frequency
- Gamma shaped parameter

Jukes Cantor (JC69)

- Assumes equal base pair frequency

$$\left(\pi_A = \pi_G = \pi_C = \pi_T = \frac{1}{4}\right)$$

- Assumes p(transitions) = p(transversions)

$$Q = \begin{pmatrix} * & \frac{\mu}{4} & \frac{\mu}{4} & \frac{\mu}{4} \\ \frac{\mu}{4} & * & \frac{\mu}{4} & \frac{\mu}{4} \\ \frac{\mu}{4} & \frac{\mu}{4} & * & \frac{\mu}{4} \\ \frac{\mu}{4} & \frac{\mu}{4} & \frac{\mu}{4} & * \end{pmatrix}$$

- Thus only parameter is overall substitution rate

Kimura 1980 (K2P)

- Assumes equal basepair frequency

$$\left(\pi_A = \pi_G = \pi_C = \pi_T = \frac{1}{4}\right)$$

- However allows for transitions and transversions to vary

Rate matrix $Q = \begin{pmatrix} * & \kappa & 1 & 1 \\ \kappa & * & 1 & 1 \\ 1 & 1 & * & \kappa \\ 1 & 1 & \kappa & * \end{pmatrix}$ with columns corresponding to A, G, C, and T, respectively.

Kimura 81 (K3P)

- Assumes equal basepair frequencies

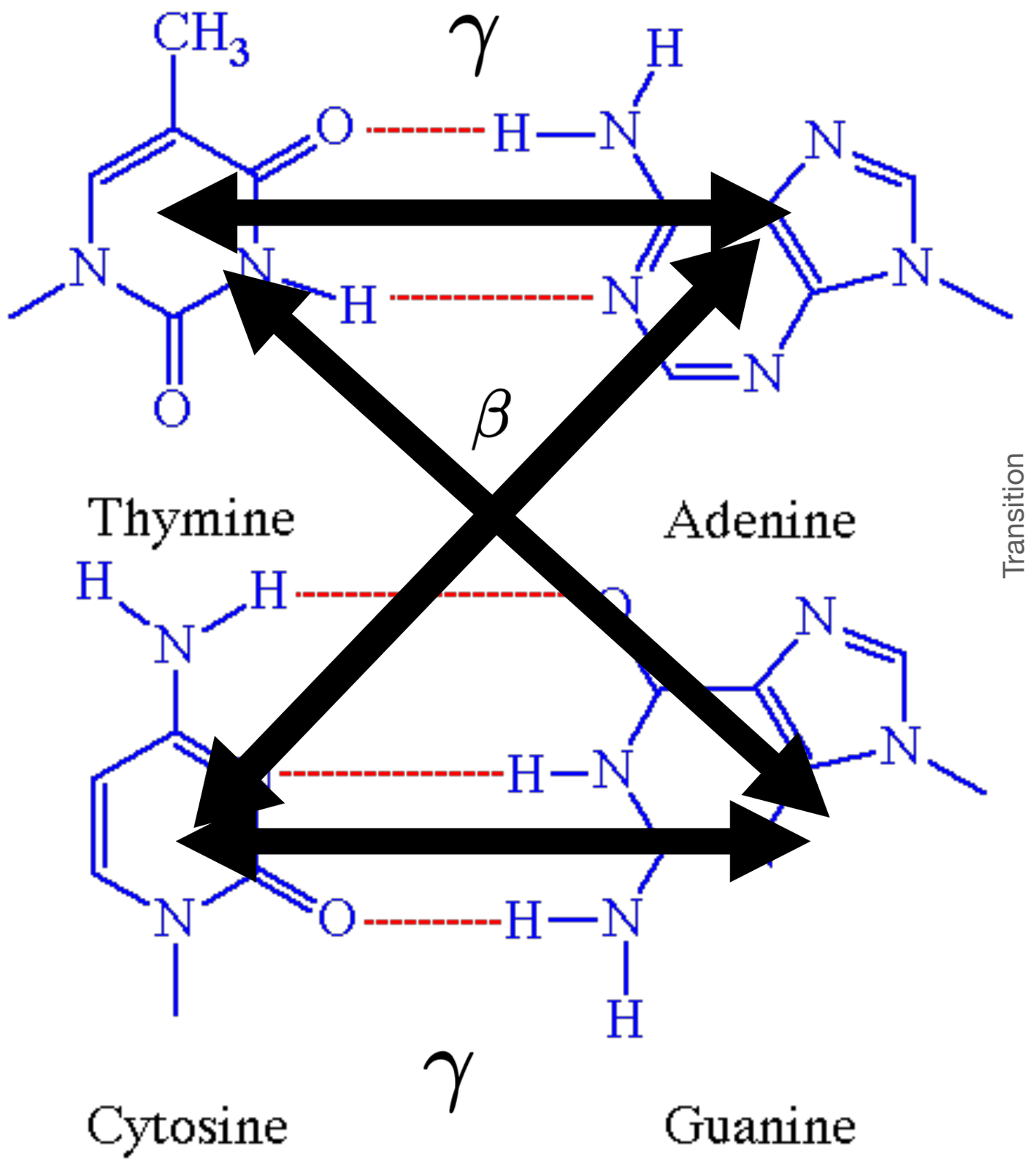
$$\left(\pi_A = \pi_G = \pi_C = \pi_T = \frac{1}{4}\right)$$

- Has different rates for transitions and transversions

Rate matrix $Q = \begin{pmatrix} * & \alpha & \beta & \gamma \\ \alpha & * & \gamma & \beta \\ \beta & \gamma & * & \alpha \\ \gamma & \beta & \alpha & * \end{pmatrix}$ with columns corresponding to A, G, C, and T, respectively.

Pyrimidines

Purines



Transition

Felsenstein 81 (F81)

- An extension of the JC69, but now allows basepair frequency to vary

$$(\pi_A \neq \pi_G \neq \pi_C \neq \pi_T)$$

- But holds still the p(transition) and p(transversion)

$$Q = \begin{pmatrix} * & \pi_G & \pi_C & \pi_T \\ \pi_A & * & \pi_C & \pi_T \\ \pi_A & \pi_G & * & \pi_T \\ \pi_A & \pi_G & \pi_C & * \end{pmatrix}$$

HKY85

- A combination of the K2P and F85 where we allow both basepair frequency AND transition and transversions to vary

$$(\pi_A \neq \pi_G \neq \pi_C \neq \pi_T)$$

$$\text{Rate matrix } Q = \begin{pmatrix} * & \kappa\pi_G & \pi_C & \pi_T \\ \kappa\pi_A & * & \pi_C & \pi_T \\ \pi_A & \pi_G & * & \kappa\pi_T \\ \pi_A & \pi_G & \kappa\pi_C & * \end{pmatrix}$$

General Time Reversible (GTR)

- Every possible combination has its own parameter

$$Q = \begin{pmatrix} -(\alpha\pi_G + \beta\pi_C + \gamma\pi_T) & \alpha\pi_G & \beta\pi_C & \gamma\pi_T \\ \alpha\pi_A & -(\alpha\pi_A + \delta\pi_C + \epsilon\pi_T) & \delta\pi_C & \epsilon\pi_T \\ \beta\pi_A & \delta\pi_G & -(\beta\pi_A + \delta\pi_G + \eta\pi_T) & \eta\pi_T \\ \gamma\pi_A & \epsilon\pi_G & \eta\pi_C & -(\gamma\pi_A + \epsilon\pi_G + \eta\pi_C) \end{pmatrix}$$

Where

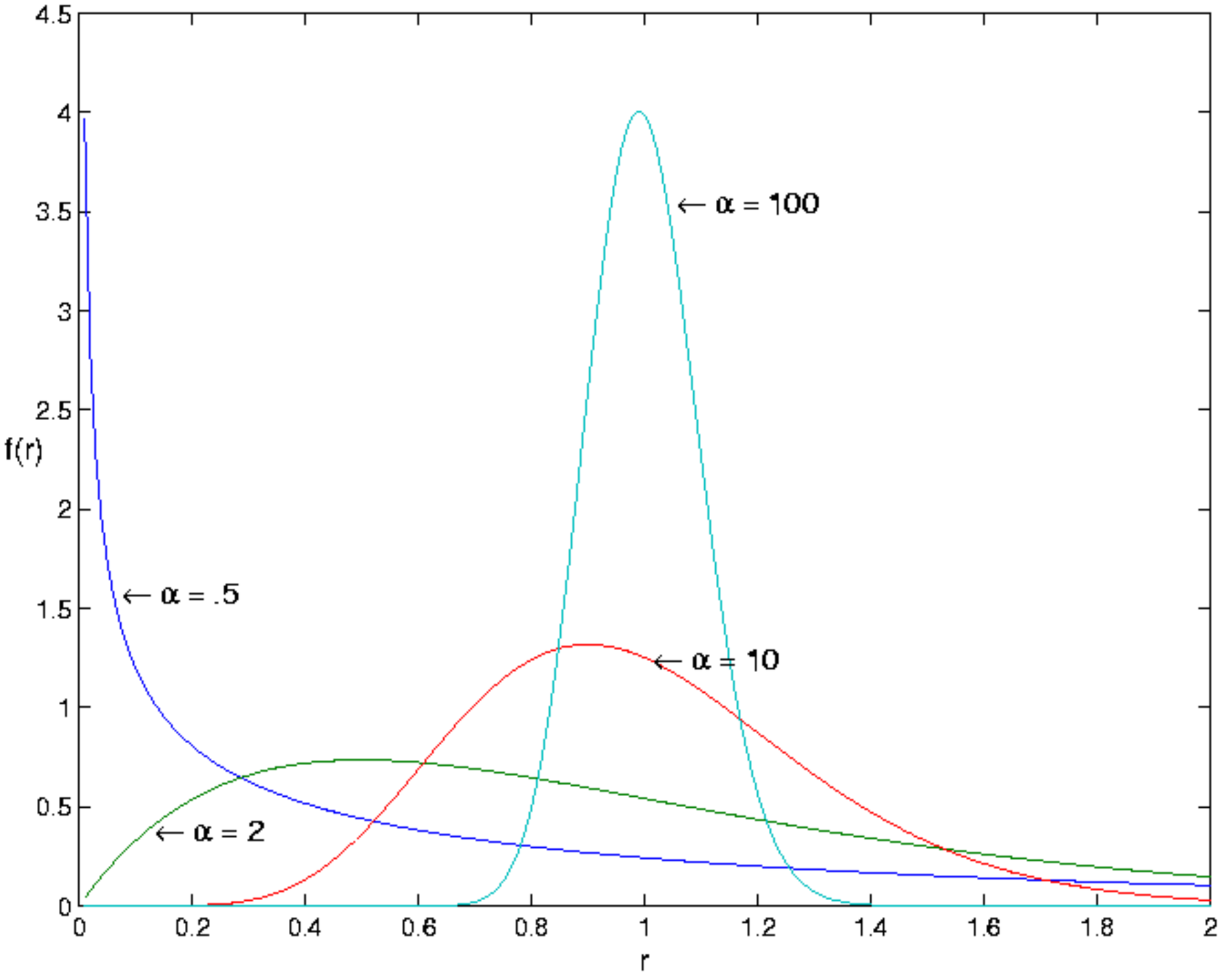
$$\alpha = r(A \to G) = r(G \to A)$$
$$\beta = r(A \to C) = r(C \to A)$$
$$\gamma = r(A \to T) = r(T \to A)$$
$$\delta = r(G \to C) = r(C \to G)$$
$$\epsilon = r(G \to T) = r(T \to G)$$
$$\eta = r(C \to T) = r(T \to C)$$

Gamma shape parameter

- This will allow us to vary how often we see transitions and transversions across the sequence. I.e. some areas may be more active than others (remember 3rd vs 1st codon position)

- When gamma is low there is strong among site variation, as gamma increases the rate of variation decreases. When small most evolve very slowly and a few are hyper variable, as it increases, the rate decreases to most evolving at the same rate

- Sometimes we also allow for a fixed number of invariant sites (+I)


4.5
4
3.5
3
2.5
2
1.5
1
0.5
0
f(r)
← α = 100
← α = .5
← α = 10
← α = 2
0
0.2
0.4
0.6
0.8
1
1.2
1.4
1.6
1.8
2
r

Models of Evolution

- Depending on how you build trees can strongly to very strongly influence the outcome

- Therefore choosing the best model is important

- In general you want the model with the least parameters that can best explain the data

- We use model testing program such as AIC to determine how to chose among

```
library(AICcmodavg)

#define list of models
models <- list(model1, model2, model3)

#specify model names
mod.names <- c('disp.hp.wt.qsec', 'disp.qsec', 'disp.wt')

#calculate AIC of each model
aictab(cand.set = models, modnames = mod.names)

Model selection based on AICc:

                K   AICc Delta_AICc AICcWt Cum.Wt     LL
disp.hp.wt.qsec 6 162.43       0.00   0.83   0.83 -73.53
disp.wt         4 165.65       3.22   0.17   1.00 -78.08
disp.qsec       4 173.32      10.89   0.00   1.00 -81.92
```

In general you want the model with the lowest AIC

Types of construction methods

- Easiest Neighbor joining
- Uses a distance matrix (in this case usually differences in substitutions) to aggregate tips into a tree
- Advantages: very fast
- Disadvantages: Will only give one tree, can create ‘negative’ branch lengths, dependent on the model of evolution

Maximum Liklihood

- Requires a data set and an explicit model of evolution
- Will look to maximize likelihood (a statistical process that looks to match statistical distributions with observed data) over a tree
- Not only supports different rates of evolution across different branches of the tree
- Will produce multiple equally likely trees which can then be used to bootstrap
- Robust to violations of underlying assumptions (to a point)
- Much slower

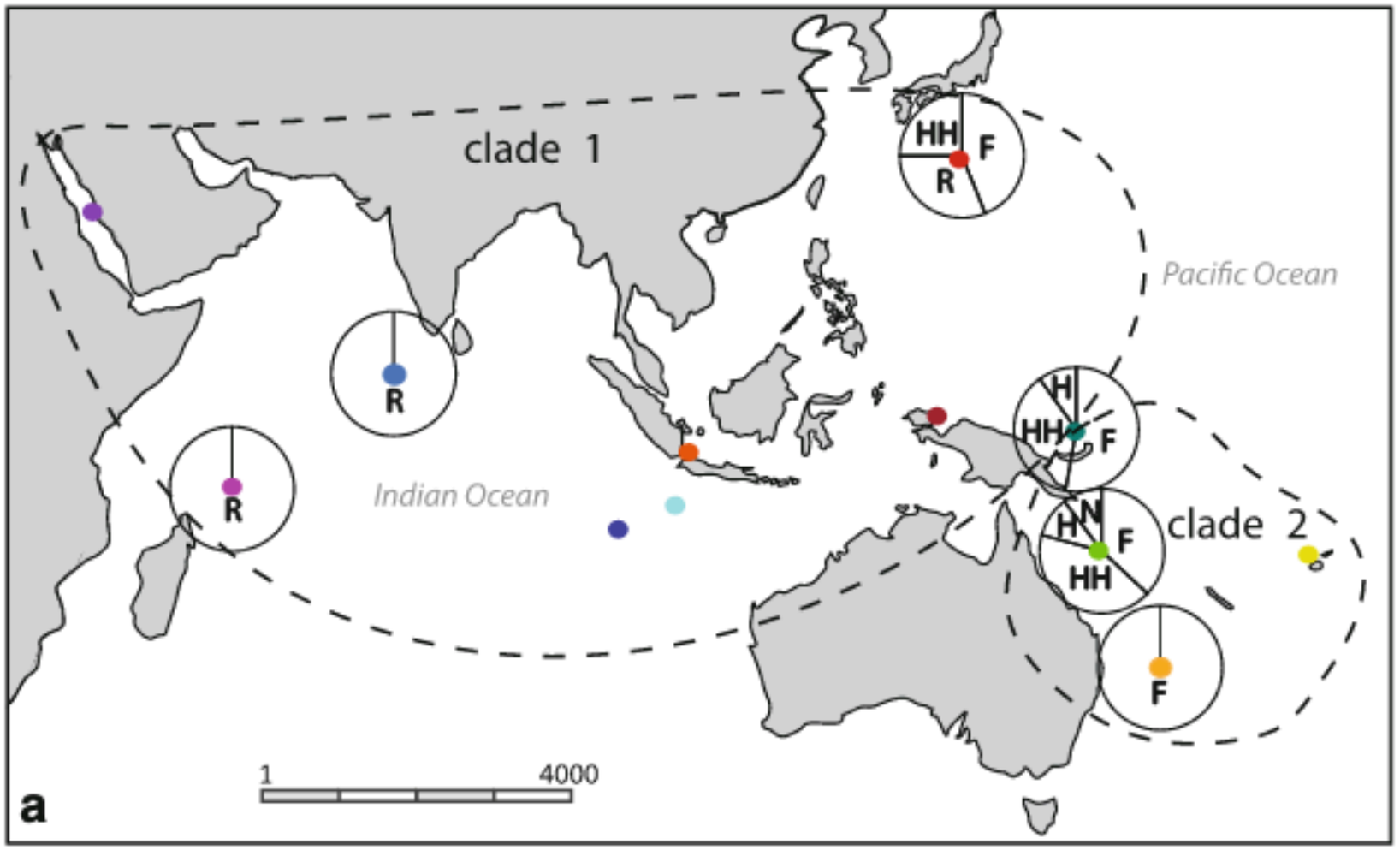
Bayesian approaches

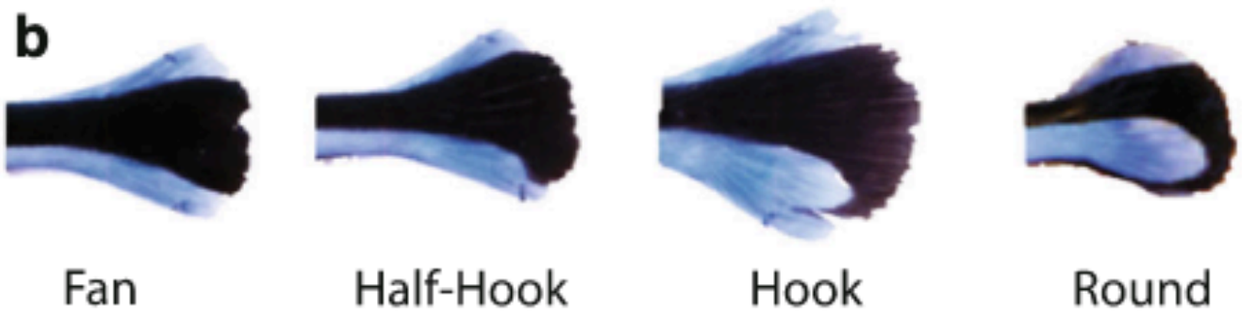
- Will generate a tree based on Bayes' theorem - a statistical technique that relies on incorporating previous data to inform future model predictions

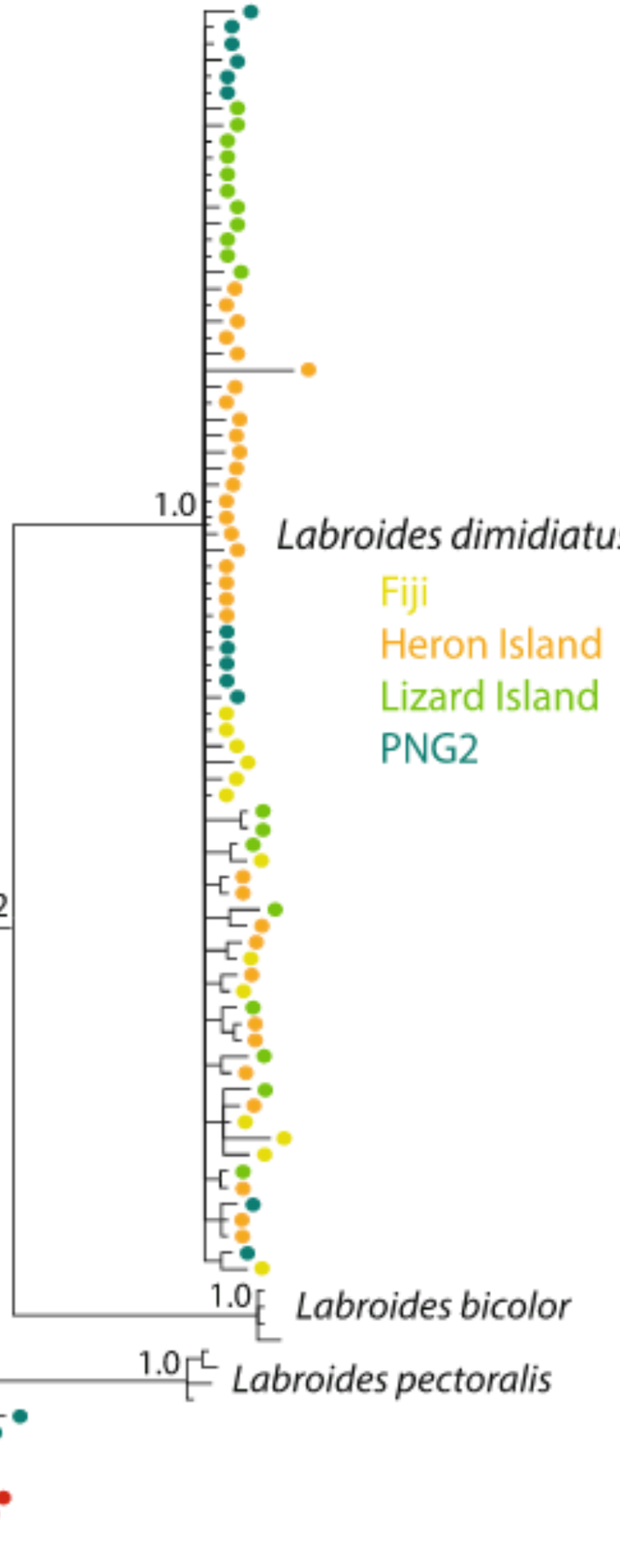
- Requires a prior probability distribution of trees to inform where new ones might develop

- Essentially each new iteration uses previous generations to create a slightly better version of a tree

- Will ultimately produce multiple equally probable trees which we can use to the posterior probabilities (functionally equivalent to bootstrapping)


clade 1
clade 2
Pacific Ocean
Indian Ocean
HH
F
R
H
HH
F
H
N
F
HH
F
R
R
1
4000
a


b
Fan
Half-Hook
Hook
Round

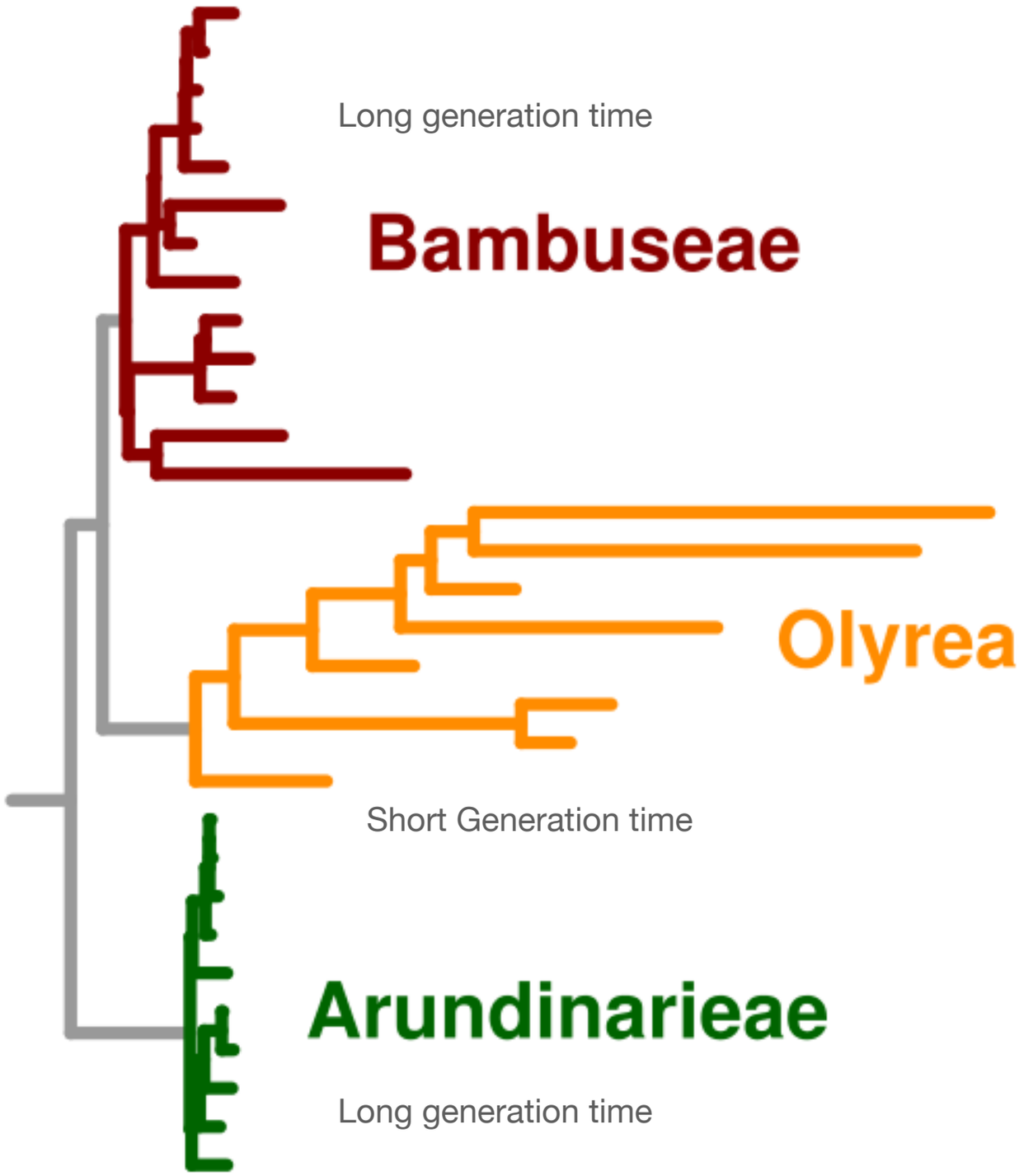

1.0
Labroides dimidiatus cl
Fiji
Heron Island
Lizard Island
PNG2
0.92
1.0
1.0
Labroides bicolor
1.0
Labroides pectoralis

A molecular clock

- Remember we assumed that there was a steady rate of substitutions
- Therefore we we know that rate and we know how many substitutions there are we can algebraically deduce how long ago those two sequences diverged
- This is known as a molecular clock.

Nuances

- Essentially a molecular clock assumes neutral evolution but we know that the rate of fixation varies due to several factors including population size and generation length
- Different parts of the genome may have different clocks
- For very long branches there may be saturation as multiple intermediary mutations will be invisible.



Woody

Herbaceous

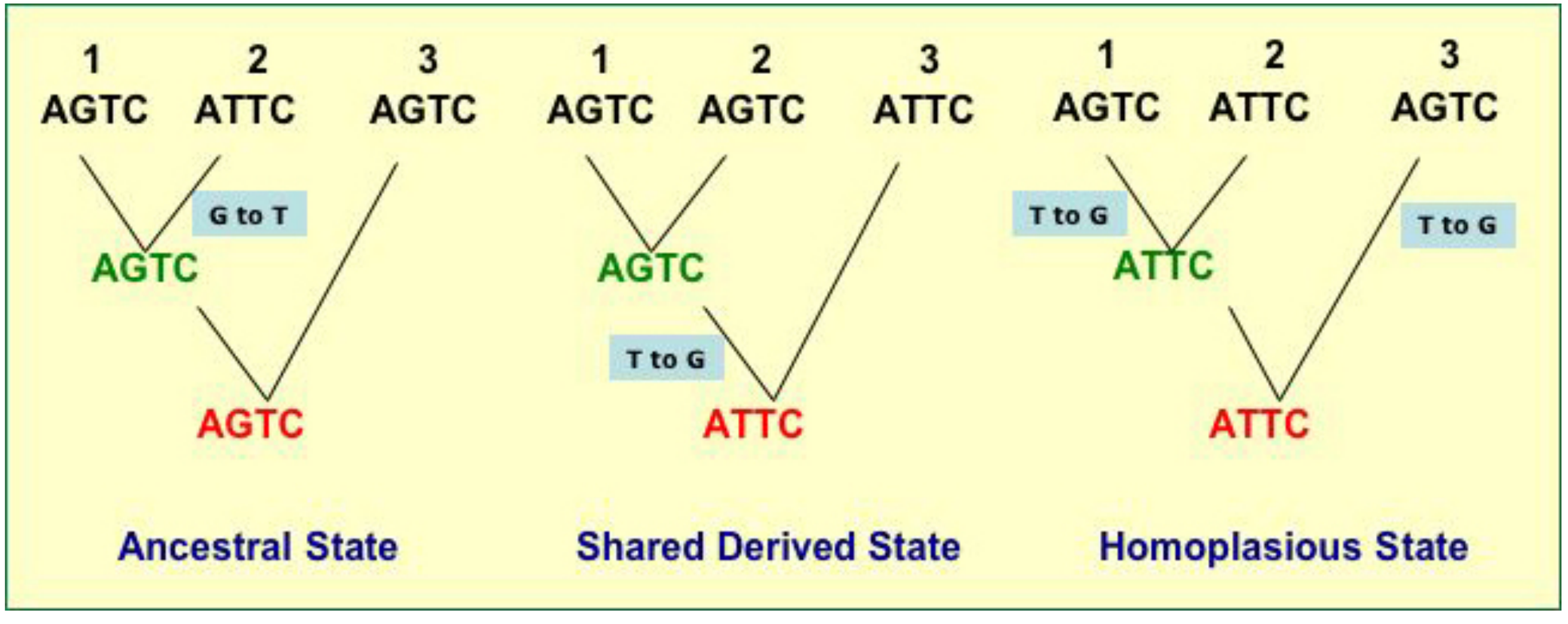
Woody

Phylogenies are hypotheses

- One problem inherent to all phylogenies is that they are hypotheses. You cannot know the true evolutionary relationship, you can only make a hypothesis based in induction from available data.
- If those data change then your hypothesis may become less well suited and therefore you may need to change it

Problems with molecular systematics

- Back mutations can lead to homoplasies: E.g. characters which are shared for reasons other than common ancestry



Incomplete Lineage Sorting

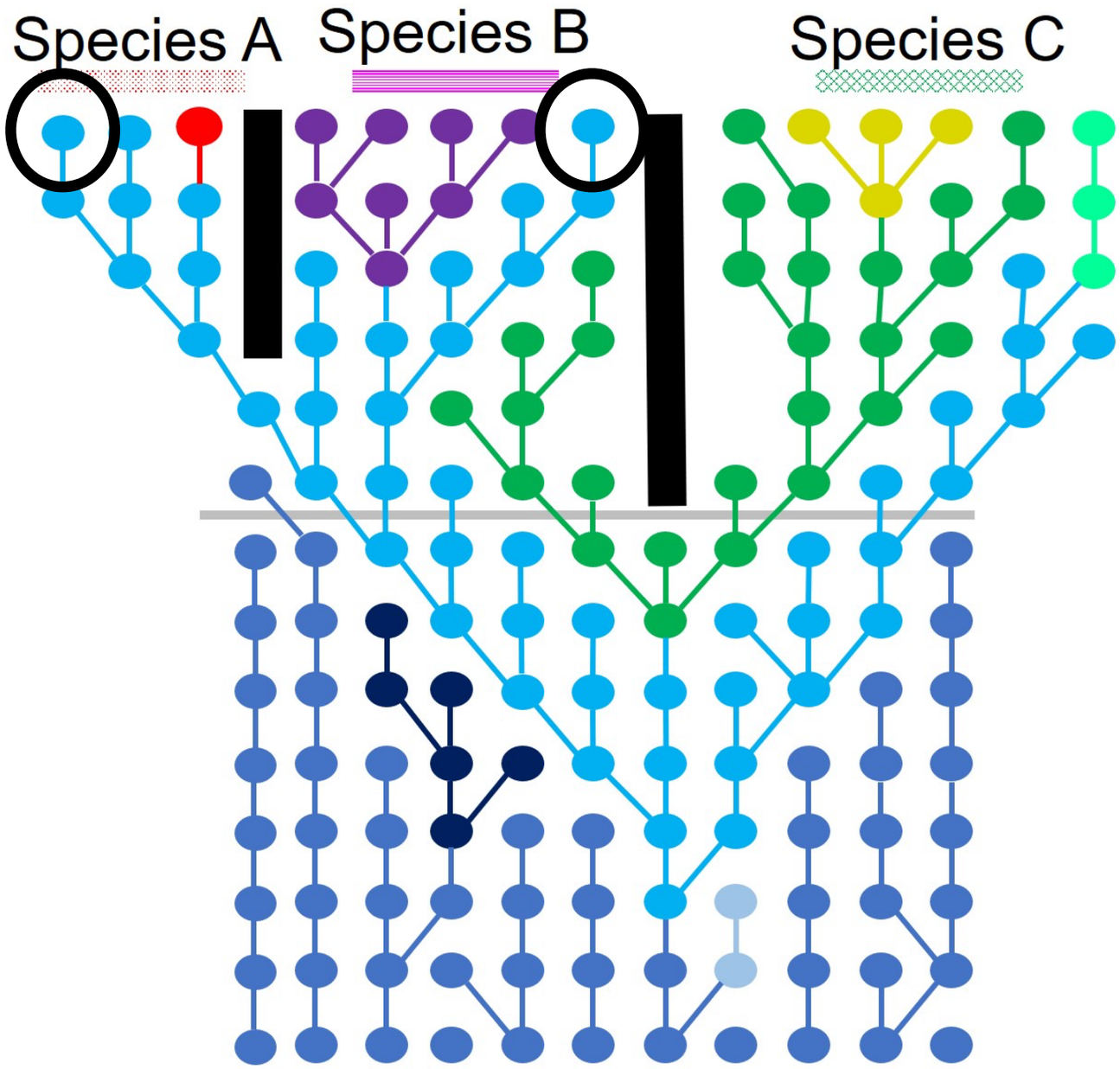
- When genes evolve at different rates, they may give different results when building trees based on them
- Often found when comparing faster evolving mitochondria DNA to slower evolving nuclear DNA
- This can also include ancestral polymorphism

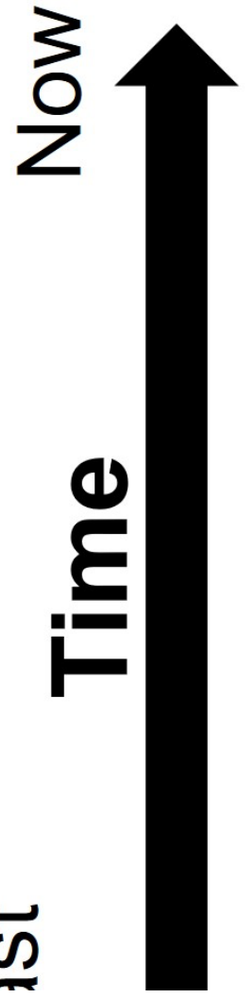
polymorphism

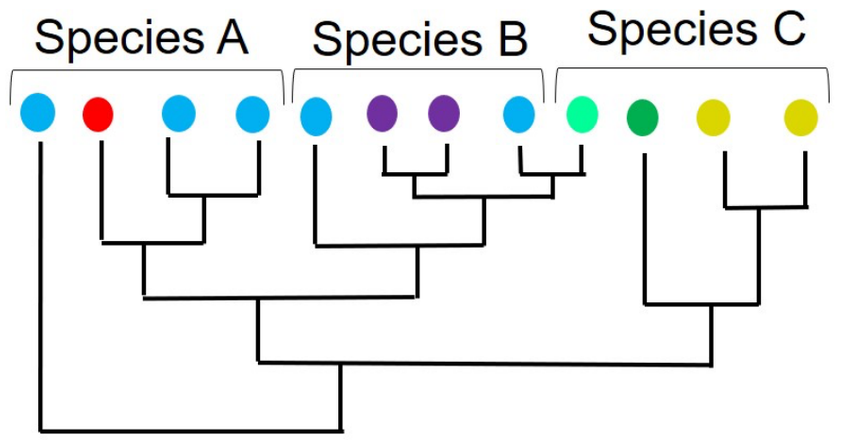
A

B

C







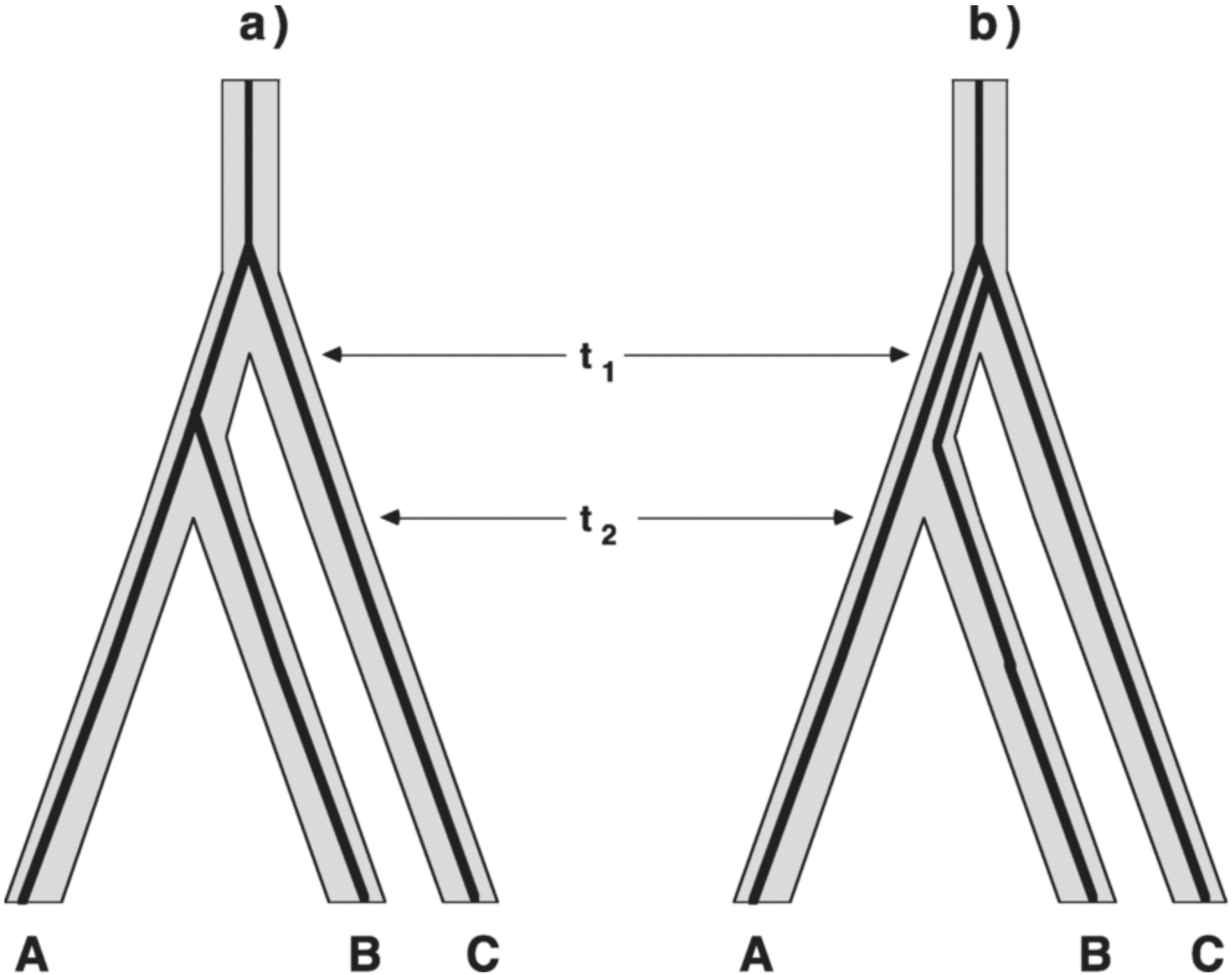
Hemiplasy - When incomplete lineage sorting yields a signal from one gene that differs from the consensus evolutionary

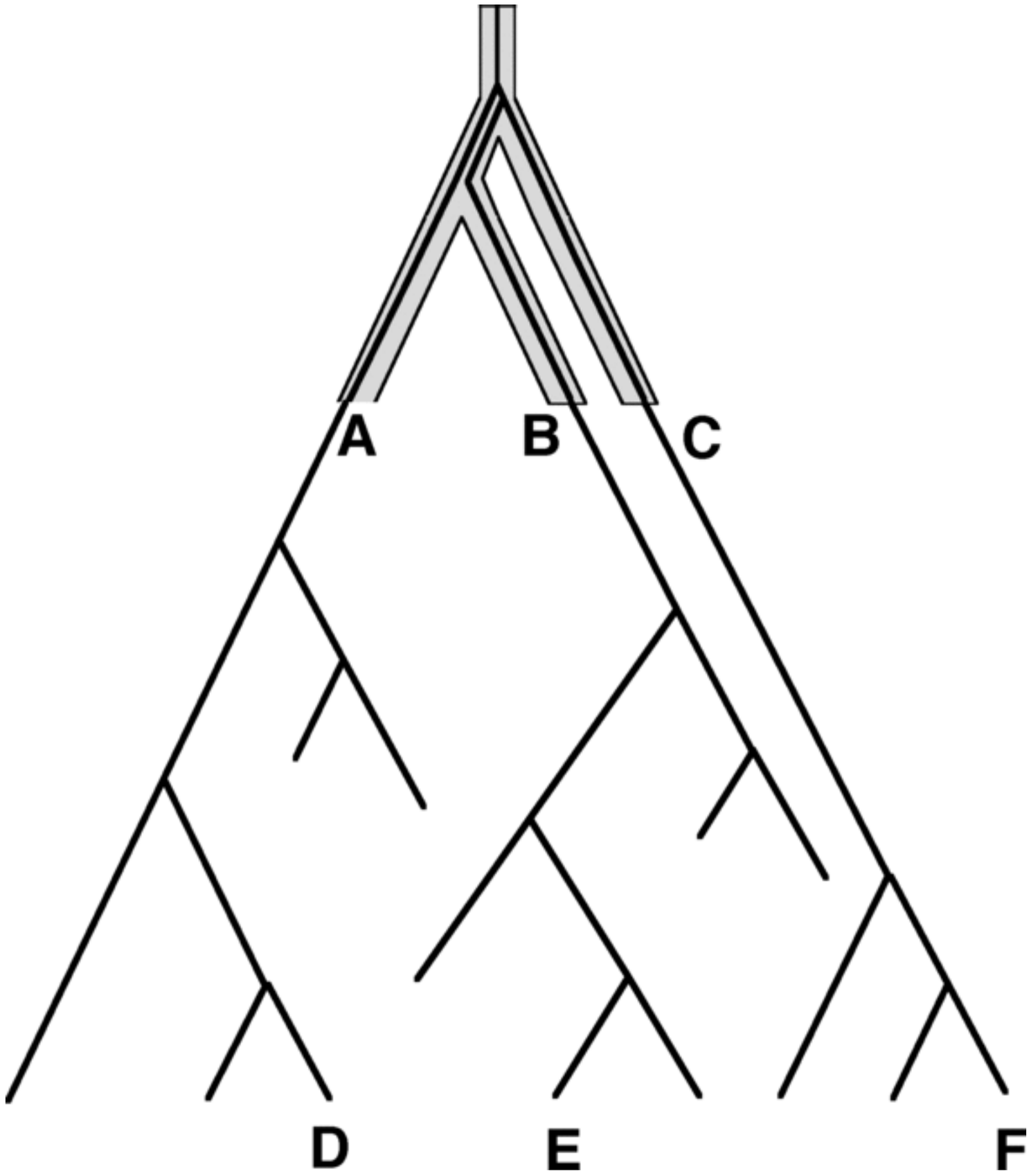
Hemiplasy: A New Term in the Lexicon o Phylogenetics FREE

John C. Avise, Terence J. Robinson

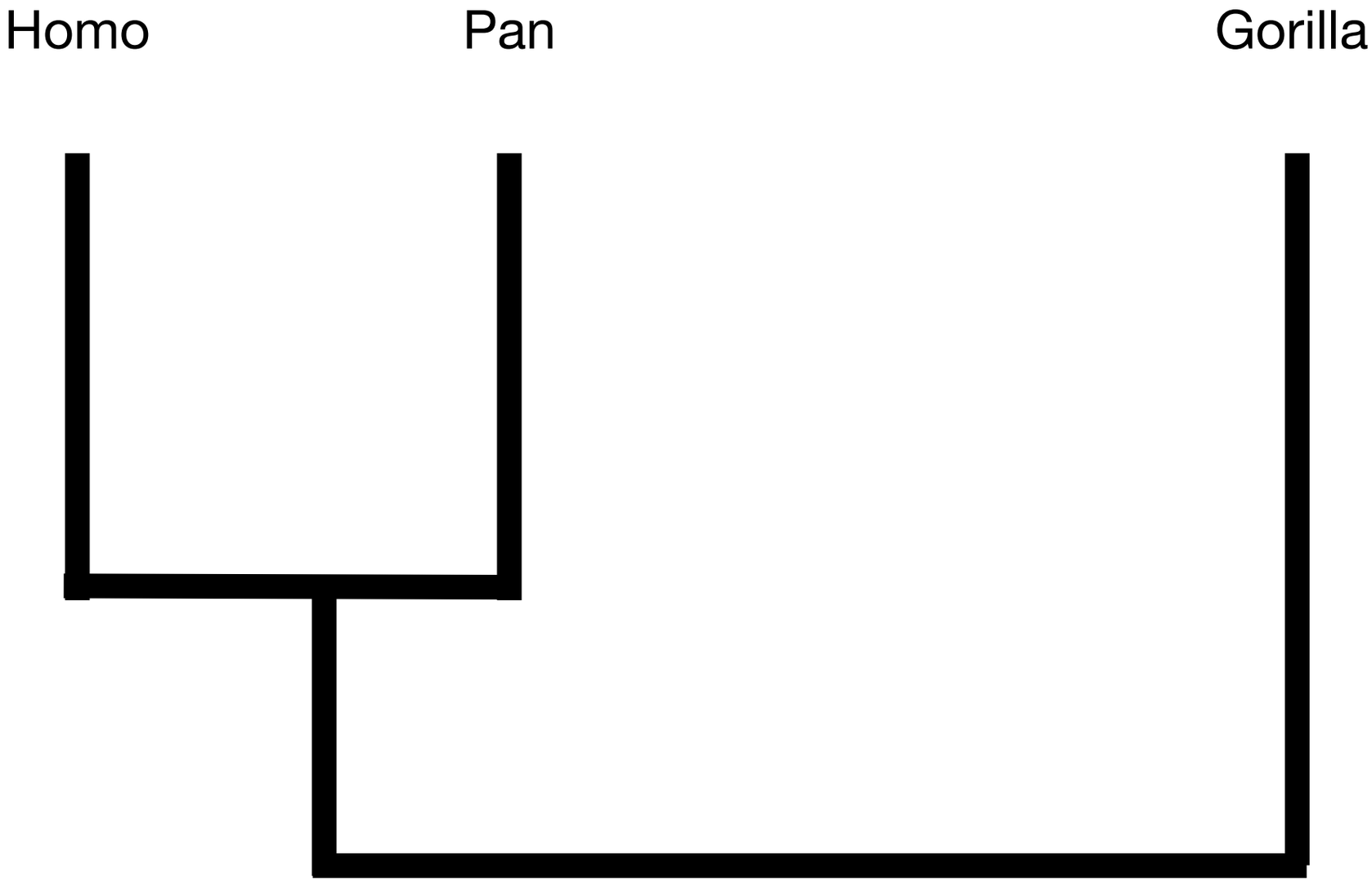
Systematic Biology, Volume 57, Issue 3, June 2008, Pages 503–507,
https://doi.org/10.1080/10635150802164587
Published: 01 June 2008 **Article history ▾**

Gene trees vs species trees




A
B
C
D
E
F

31/43 gene trees

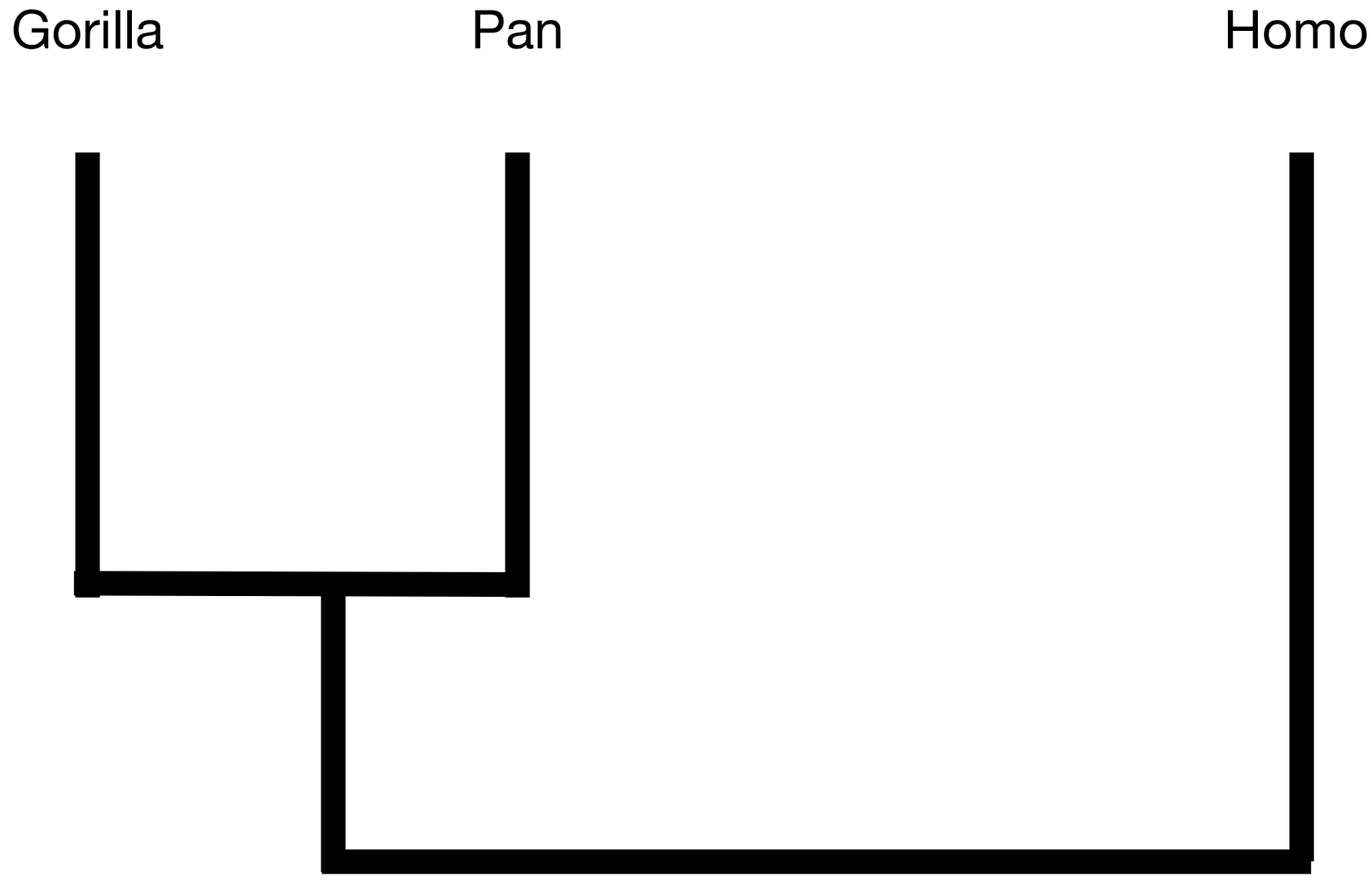


AJHG
Volume 68, Issue 2, February 2001, Pages 444-456

Genomic Divergences between Humans and Other Hominoids and the Effective Population Size of the Common Ancestor of Humans and Chimpanzees

Feng-Chi Chen [1, *], Wen-Hsiung Li [2]

12/43 gene trees



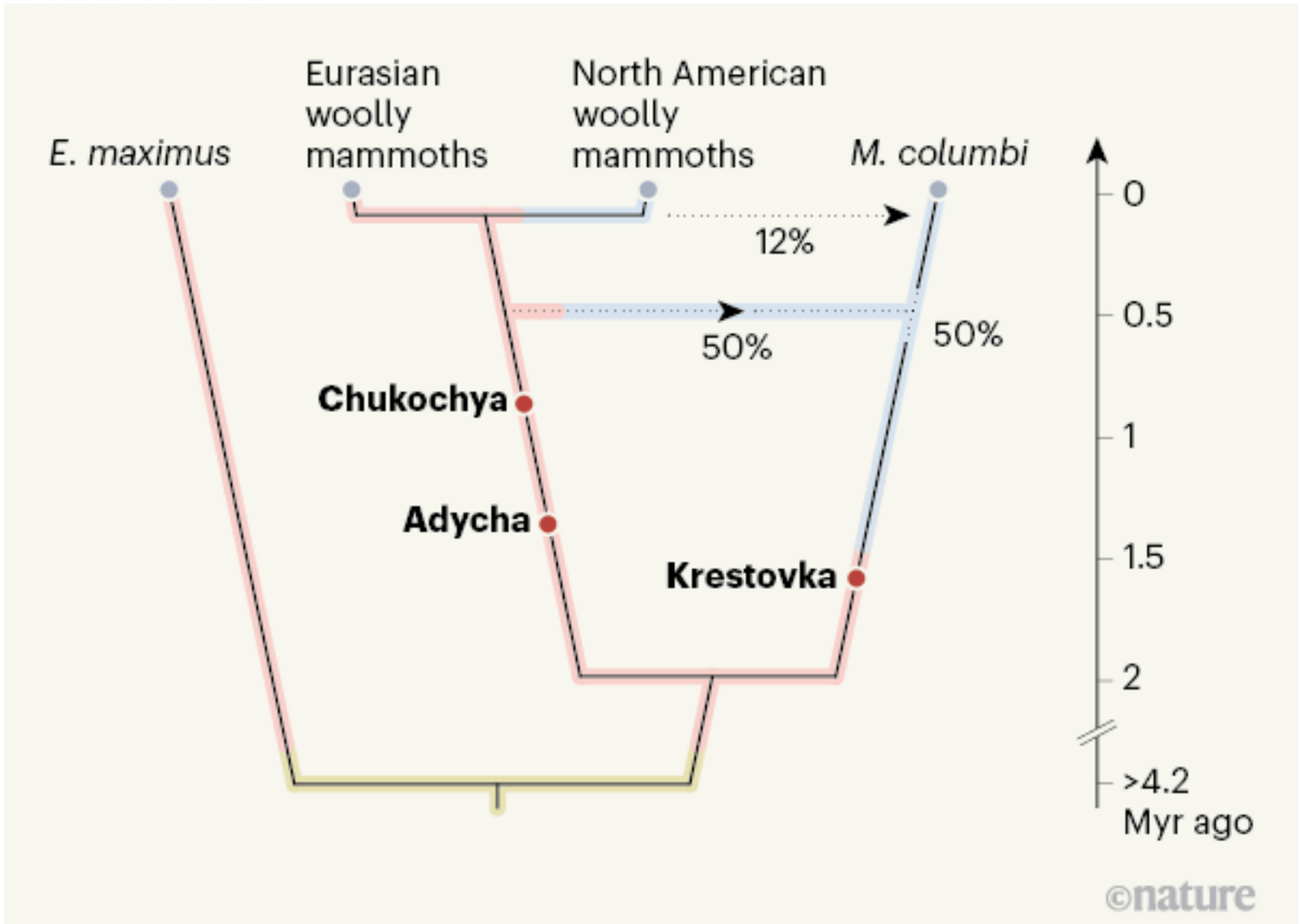
AJHG
Volume 68, Issue 2, February 2001, Pages 444-456

Genomic Divergences between Humans and Other Hominoids and the Effective Population Size of the Common Ancestor of Humans and Chimpanzees

Feng-Chi Chen [1, *], Wen-Hsiung Li [2]

Taxon sampling bias

- We can only get data from species we have in hand
- In some cases we might be able to dig deeper with what we have and increases in sequencing technology have made that possible
- In some cases we can even sequence extinct species

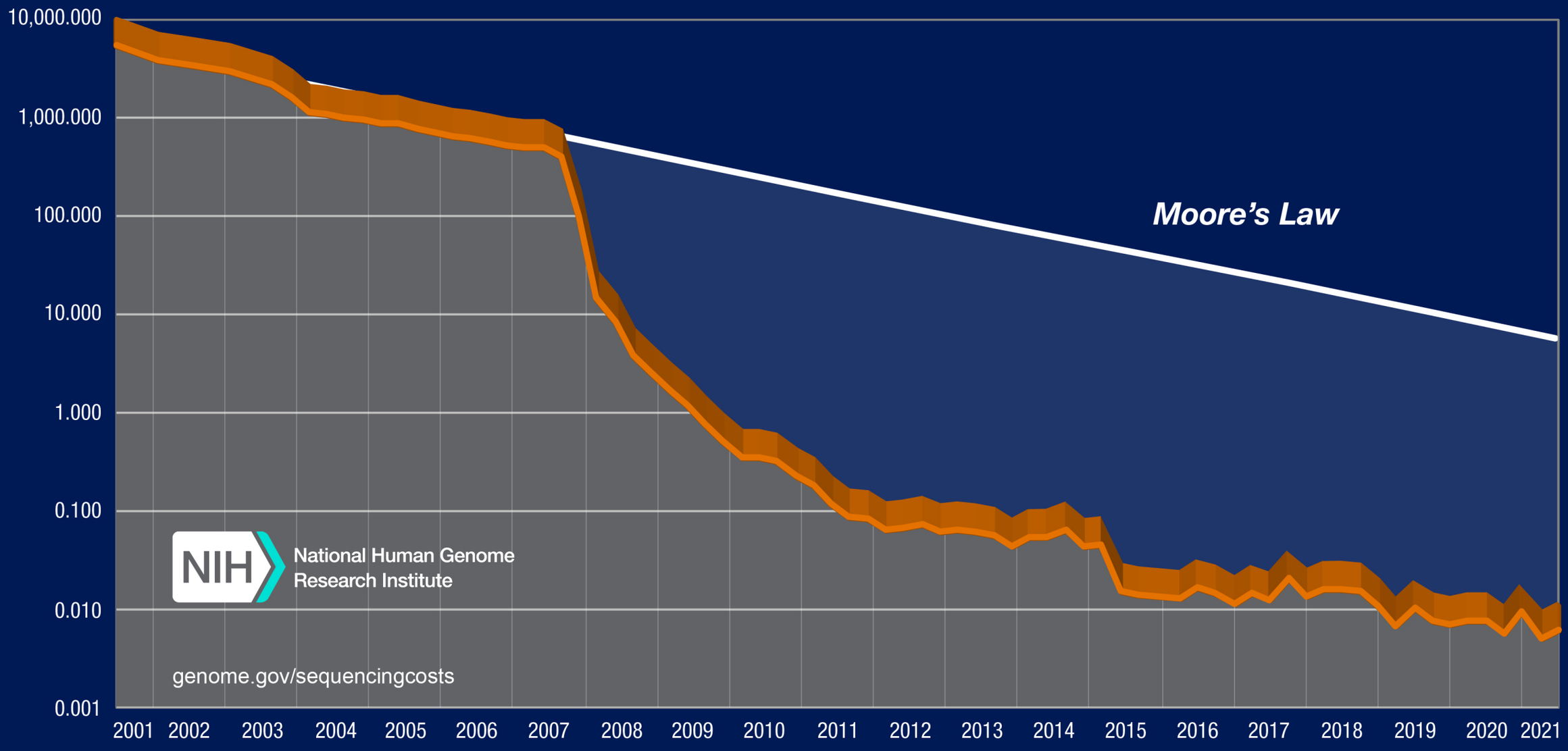


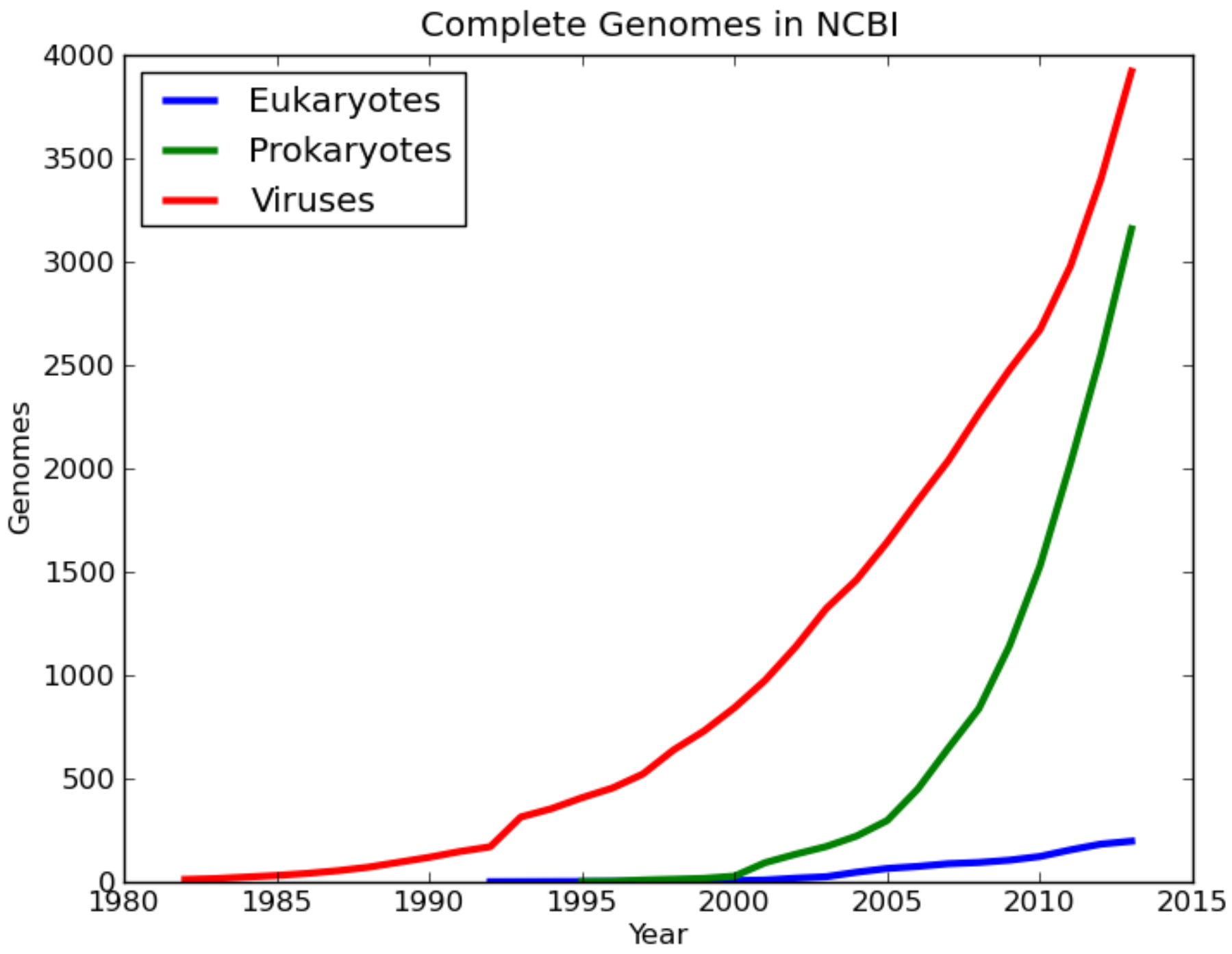
Article | Published: 17 February 2021

Million-year-old DNA sheds light on the genomic history of mammoths

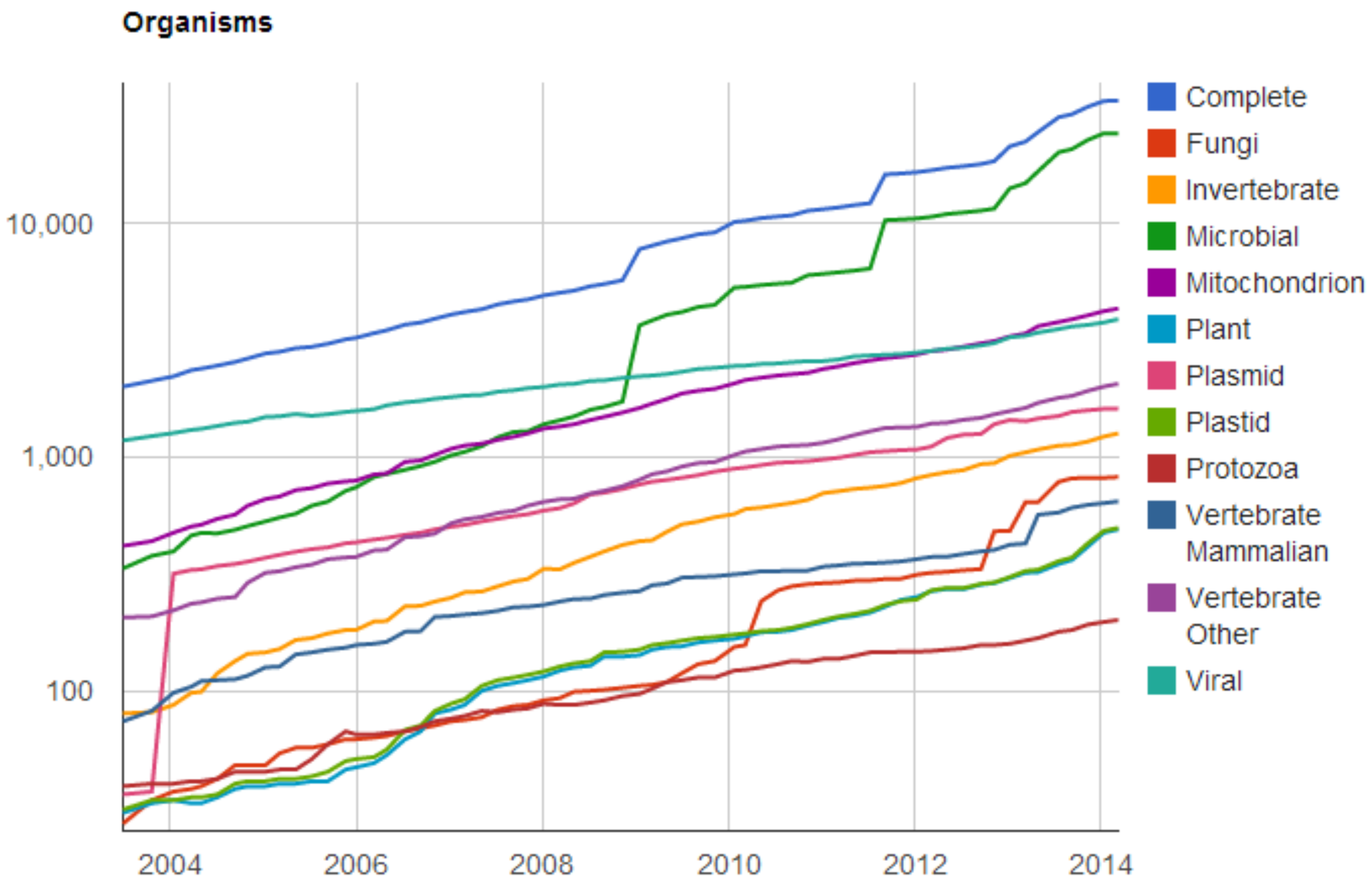
Tom van der Valk, Patrícia Pečnerová, David Díez-del-Molino, Anders Bergström, Jonas Oppenheimer, Stefanie Hartmann, Georgios Xenikoudakis, Jessica A. Thomas, Marianne Dehasque, Ekin Sağlıcan, Fatma Rabia Fidan, Ian Barnes, Shanlin Liu, Mehmet Somel, Peter D. Heintzman, Pavel Nikolskiy, Beth Shapiro, Pontus Skoglund, Michael Hofreiter, Adrian M. Lister, Anders Götherström & Love Dalén

Cost per Raw Megabase of DNA Sequence

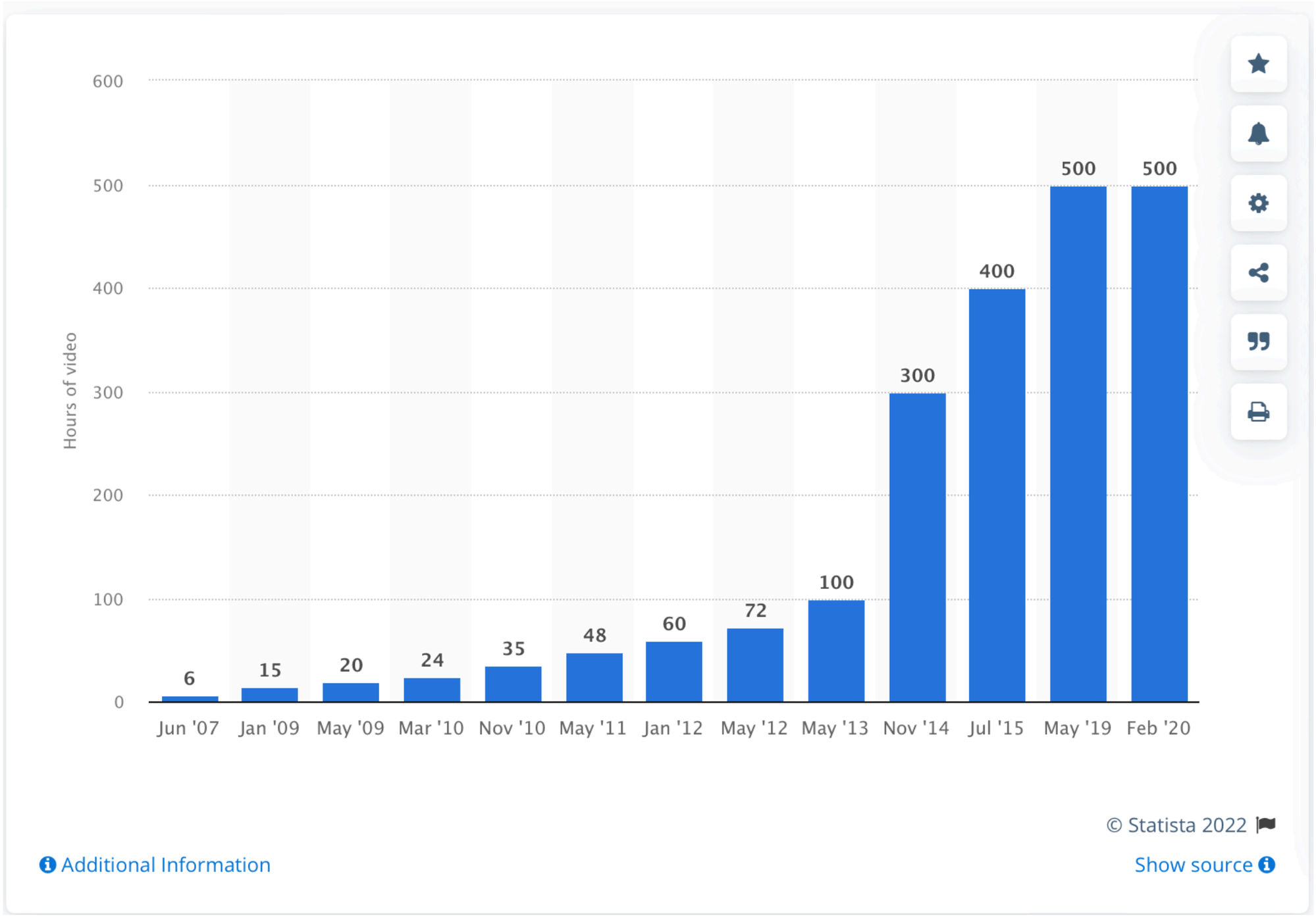




http://gregoryzynda.com/ncbi/genome/python/2014/03/31/ncbi-genome.html


Organisms
10,000
1,000
100
2004
2006
2008
2010
2012
2014
Complete
Fungi
Invertebrate
Microbial
Mitochondrion
Plant
Plasmid
Plastid
Protozoa
Vertebrate Mammalian
Vertebrate Other
Viral

Hours of YouTube Video per *minute*

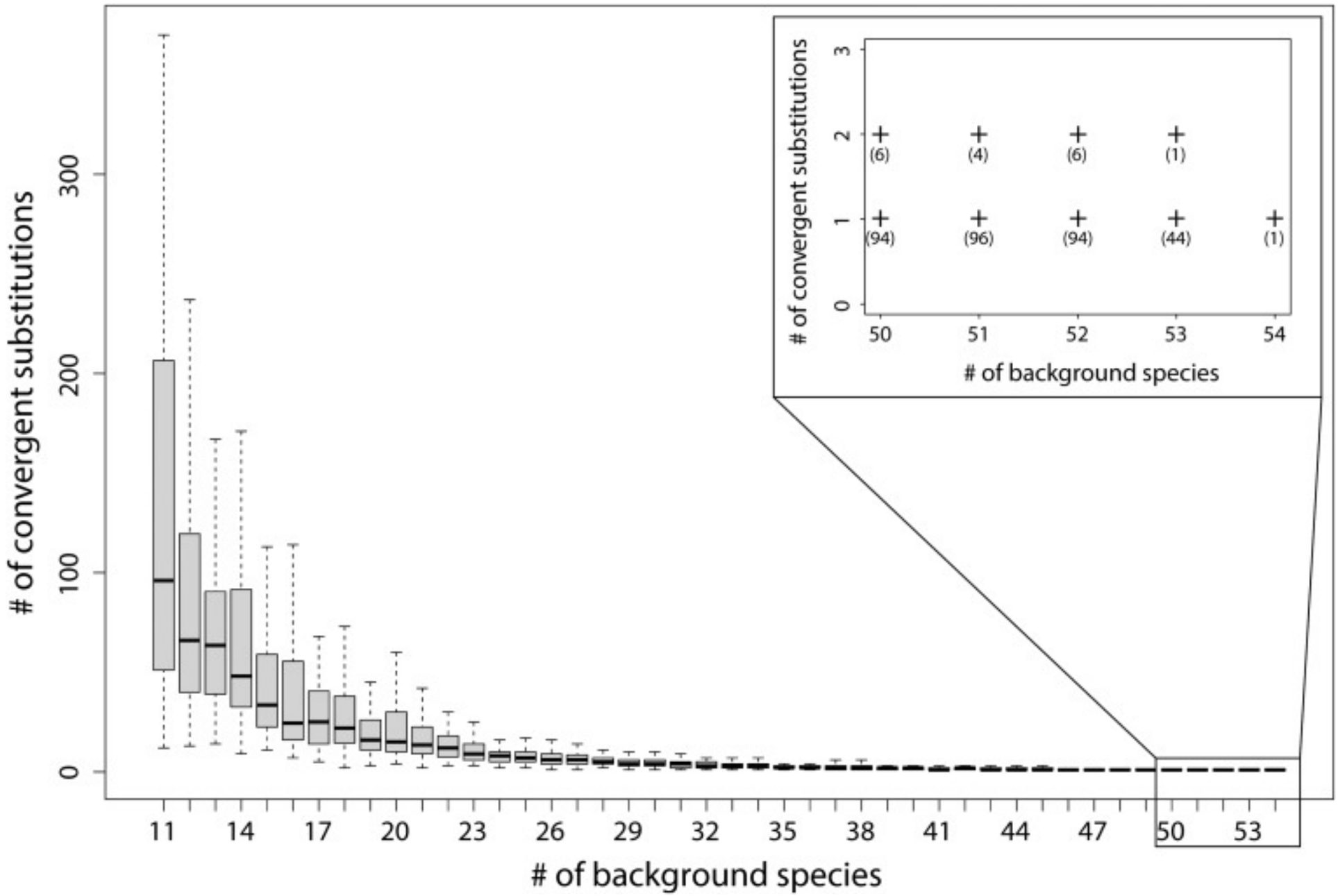


https://www.statista.com/statistics/259477/hours-of-video-uploaded-to-youtube-every-minute/

However more of the same thing isn’t always better

- While there is some debate, the majority of simulations suggest that adding additional species is going to give a better phylogenetic signal than adding additional genes from the samples you already have in hand

Marine Mammals

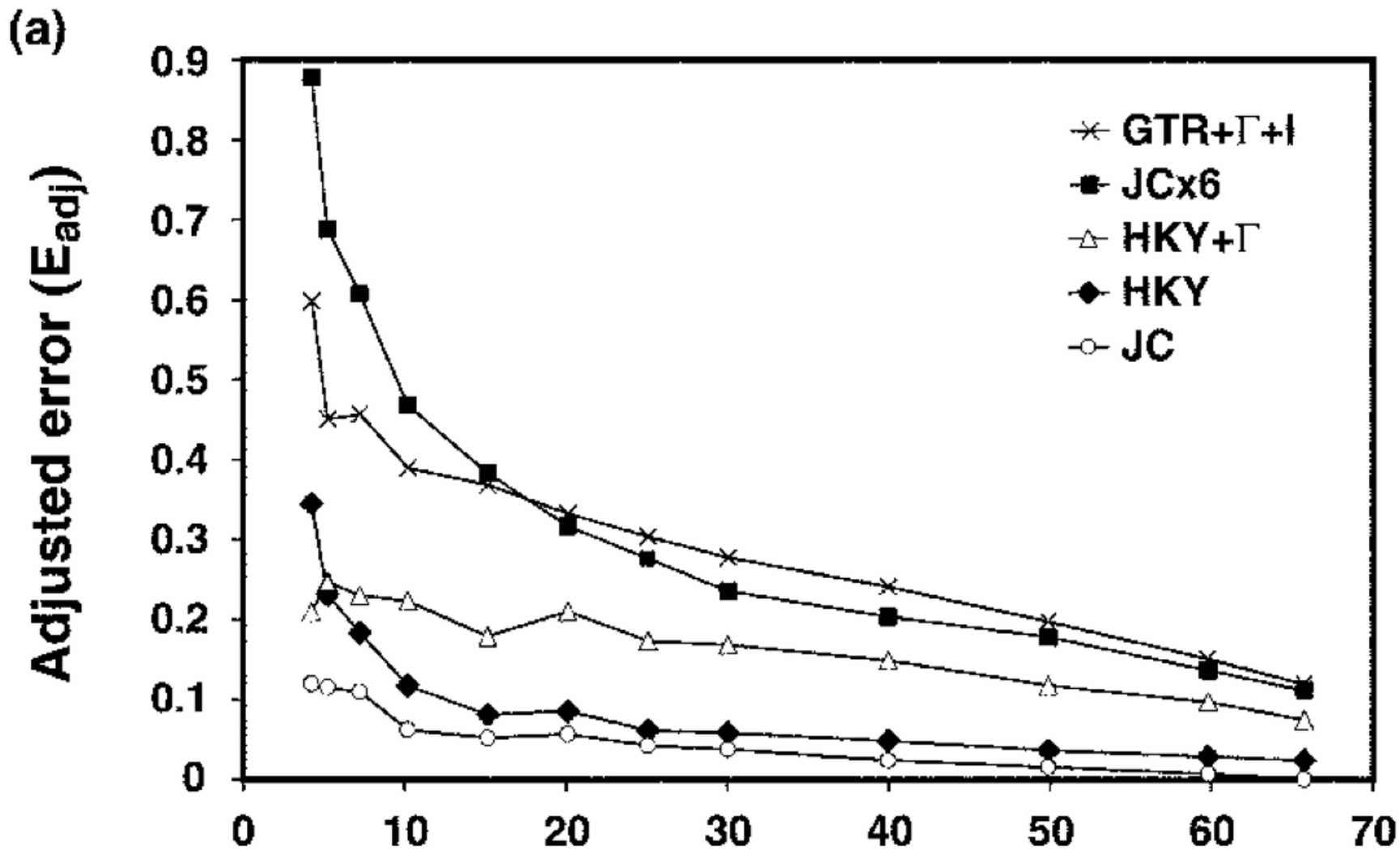


Genome Biol Evol. 2017 Jan; 9(1): 213–221.
Published online 2017 Feb 1. doi: 10.1093/gbe/evw306
PMCID: PMC5381636
PMID: 28057728

The Effects of Increasing the Number of Taxa on Inferences of Molecular Convergence

Gregg W. C. Thomas,[✉1] Matthew W. Hahn,[1] and Yoonsoo Hahn[2]

Author information · Article notes · Copyright and License information Disclaimer



Syst. Biol. 51(4):588–598, 2002
DOI: 10.1080/10635150290102339

Increased Taxon Sampling Greatly Reduces Phylogenetic Error

DERRICK J. ZWICKL AND DAVID M. HILLIS

Section of Integrative Biology and Center for Computational Biology and Bioinformatics, University of Texas, Austin, Texas 78712, USA; E-mail: zwickl@mail.utexas.edu and dhillis@mail.utexas.edu