American Museum of Natural History

American Museum of Natural History
Richard Gilder Graduate School

RGGS Comparative Genomics 2 – Computational Methods (Session 14)

Jose Barba

Gerstner Scholar in Bioinformatics & Computational Biology

AMNH Arachnology Lab
Institute for Comparative Genomics
jbarba@amnh.org
josebarbamontoya.github.io

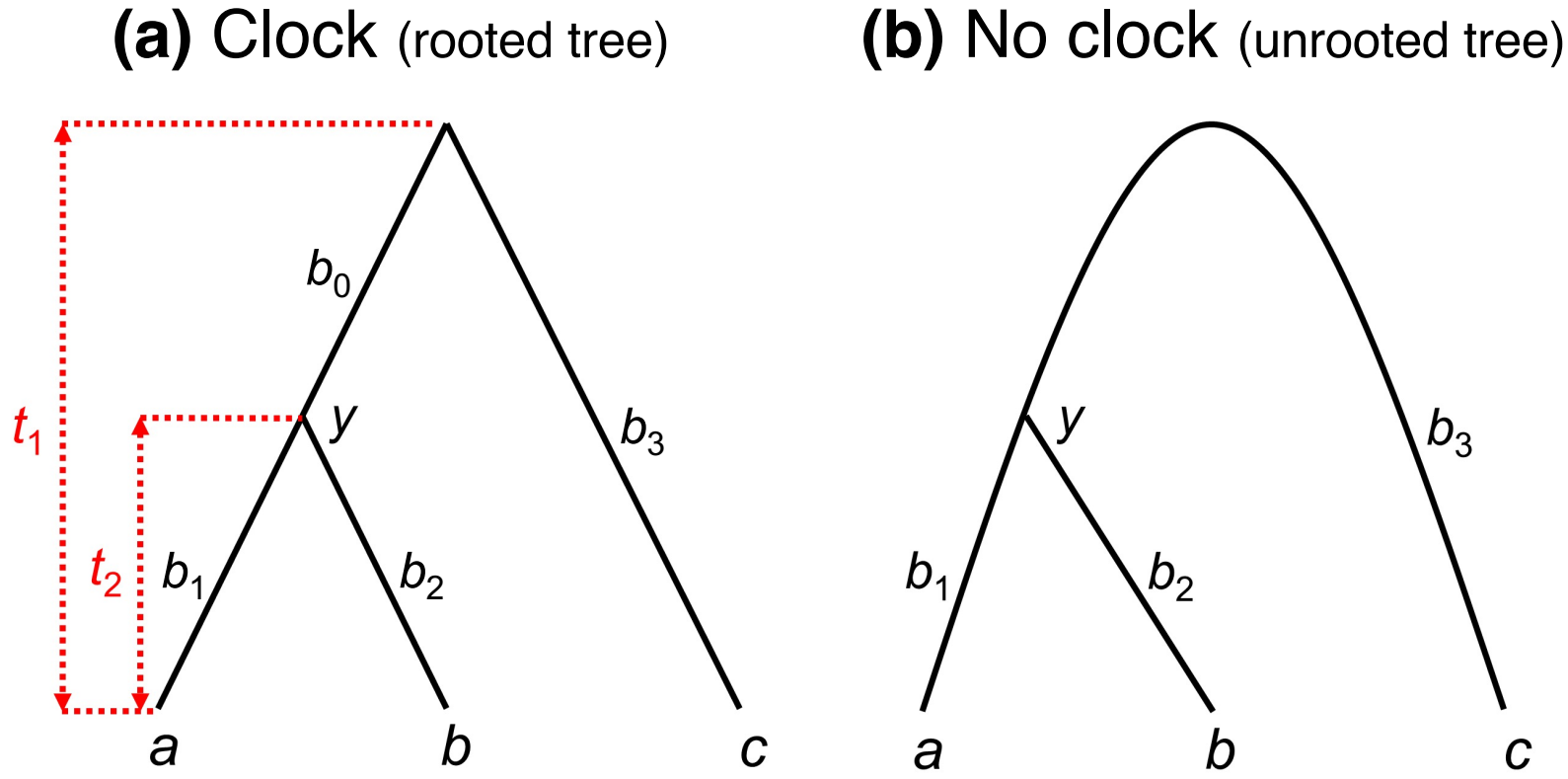
December 5, 2024
ICG conference room

Session 14 outline

- **ML and Bayesian phylogenetic inference (topics from Session 13 refer to that presentation)**
- **Molecular clock dating using phylogenomic data**
- **A tutorial for conducting a phylogenomic analysis using simulated data will be completed**
  - A comparison of phylogenetic methods will be conducted
  - A Bayesian timetree will be inferred

Relative-rate test

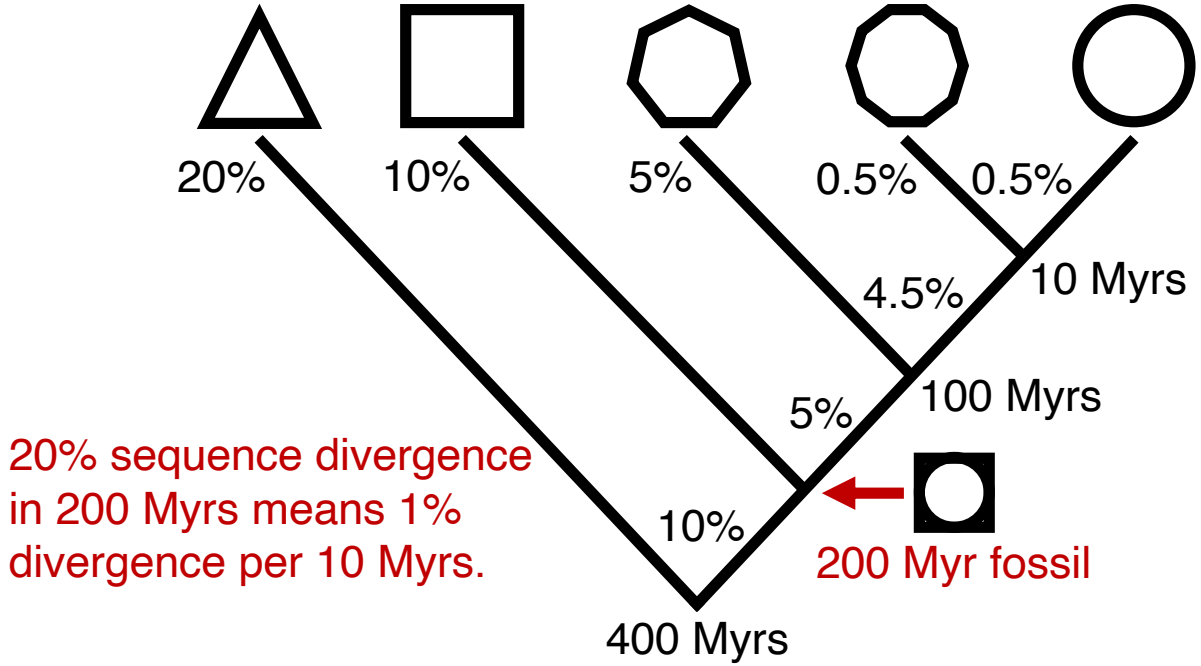
- **If substitution rate is constant over time or among lineages, molecular clock holds. For distantly related species this hypothesis should not be assumed**



Sarich and Wilson (*Science* 1973)

Molecular clock hypothesis

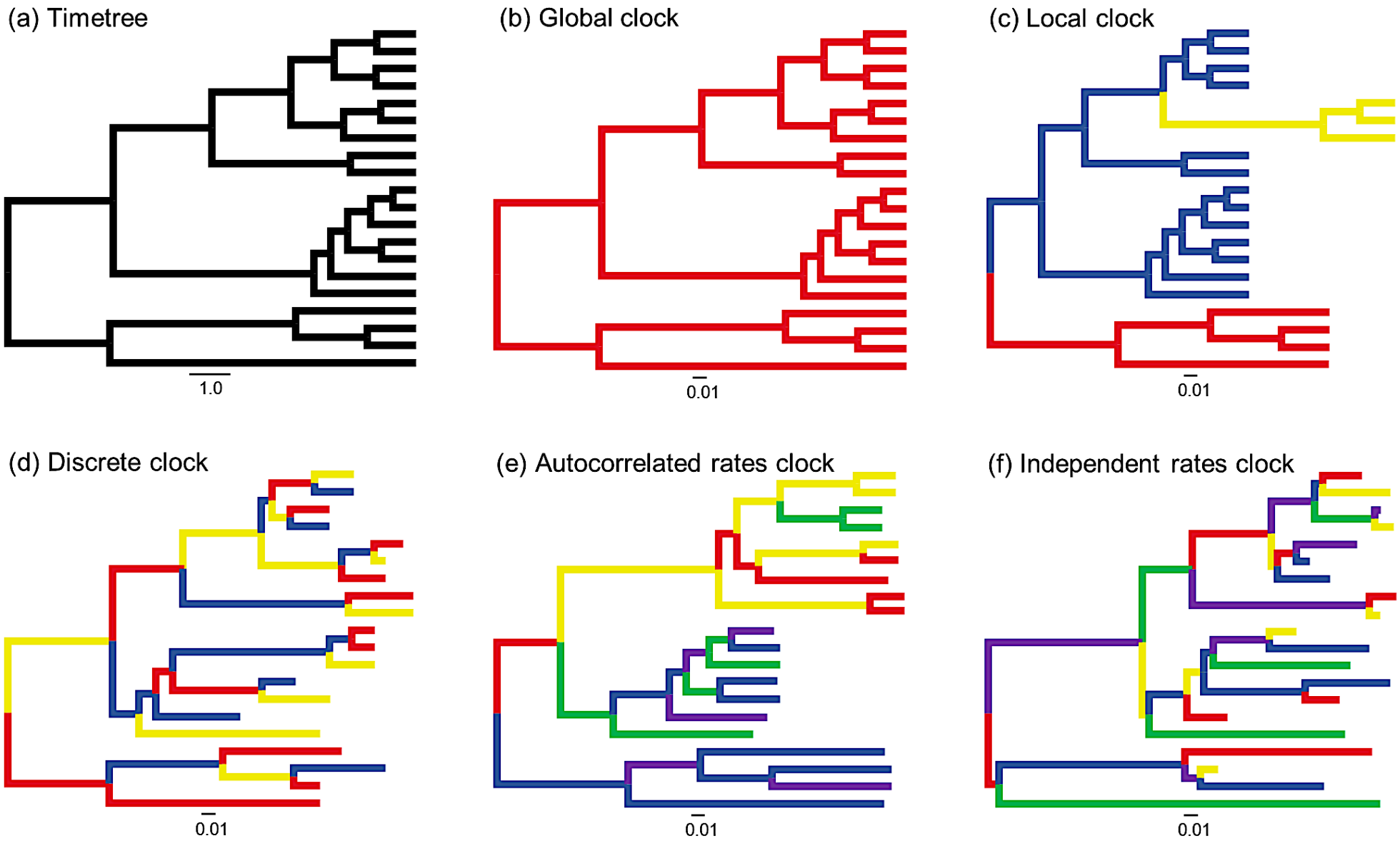
- **Simple but powerful approach measuring timescale of evolutionary divergences. Expected distance between sequences grows linearly with time of divergence**



- **Ages from fossil record or geological events, can be used to translate distances between sequences or tree branch lengths into absolute geological times**

Zuckerkandl and Pauling (1962)

Molecular clock model



Phylogenetic uncertainty in molecular dating

- **Types of molecular clock methods**
  - Generate reliable phylogeny before estimating divergence times
  - Jointly infer phylogeny and divergence times

Sequential Analysis

- MCMCTree
- MultiDivTime
- PhyloBayes
- RelTime
- TreePL

Joint analysis

- BEAST
- BEAST2
- MrBayes
- RevBayes
- TREETIME

Bayesian divergence time estimation

- **Using Bayesian method, can integrate fossil, molecule and clock uncertainty**

$$f(\mathbf{t}, \mathbf{r}, \theta | D) = f(\theta) f(\mathbf{t}) f(\mathbf{r} | \mathbf{t}, \theta) L(D | \mathbf{t}, \mathbf{r}, \theta) / k$$

time and rate estimates

fossil uncertainty

rate uncertainty

branch length uncertainty

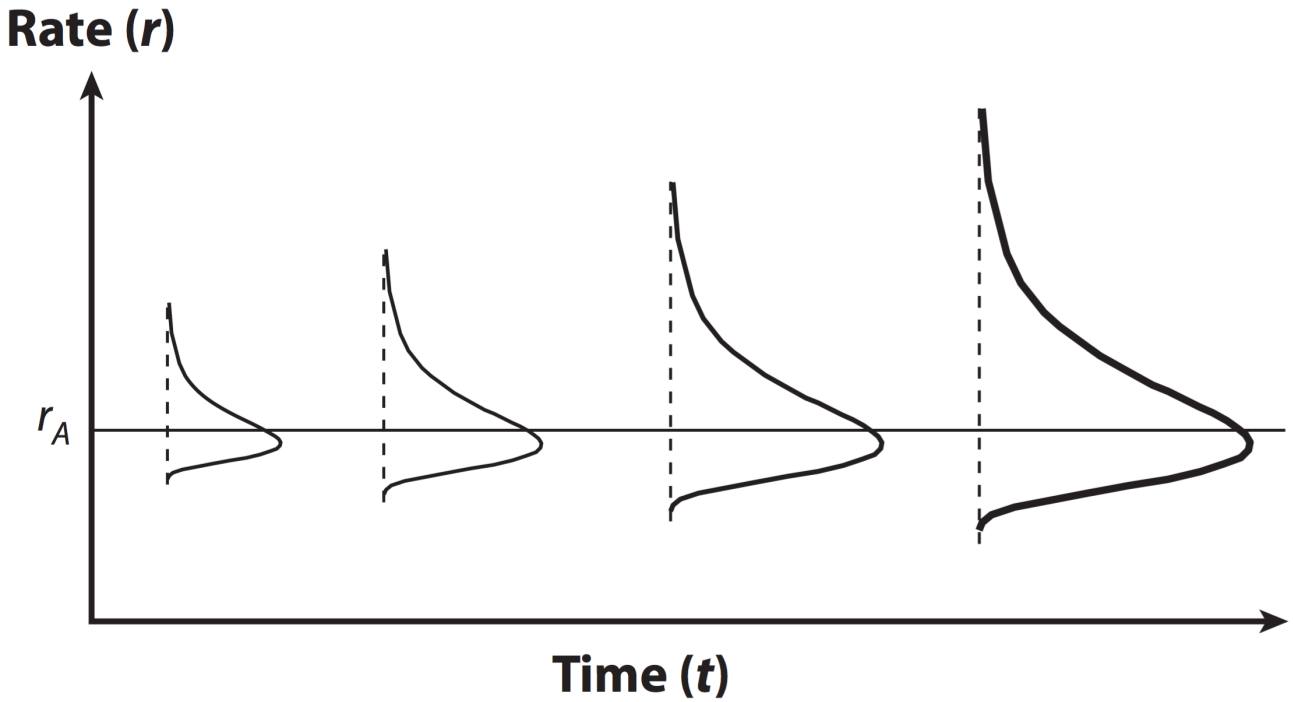
- **Calculating the posterior involves a complex integral (*k*), so MCMC algorithms are used to generate samples from joint posterior dist.**

Yang and Rannala (*MBE* 2006)

Relaxed clocks and prior model of rate drift

- **In AR model, rate at each node specified by conditioning on rate at ancestral node**
  - Given the rate r_A at ancestral node, rate r at current node has a lognormal dist.

Geometric Brownian motion model of rate drift



Rannala and Yang (*Systematic Biology* 2007)

Relaxed clocks and prior model of rate drift

- **In IR model, rate for branch is a random variable drawn from a common probability distribution**
- **Rates effectively evolve independently on each lineage, but extent of rate variation has evolutionary constraint (imposed by prior distribution on rates)**

$$f(r \mid \mu, \sigma^2) = \frac{1}{r\sqrt{2\pi\sigma^2}} \exp\{-\tfrac{1}{2\sigma^2}(\log(r/\mu) + \tfrac{1}{2}\sigma^2)^2\}, \quad 0 < r < \infty$$

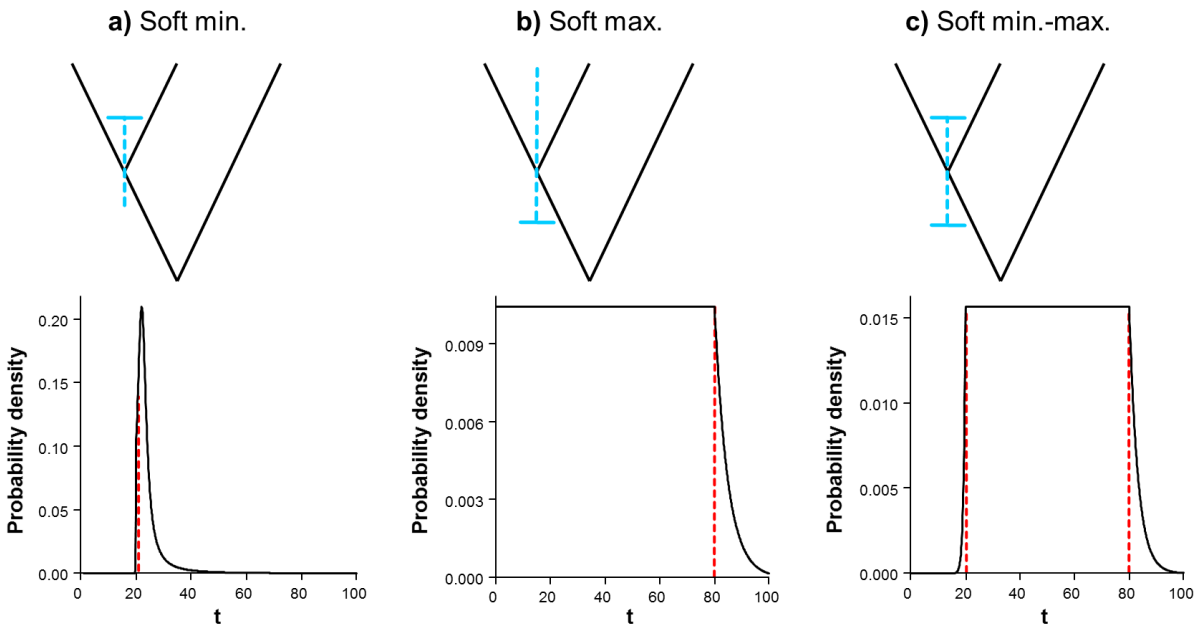
mean rate across loci

variance of the logarithm of the rate

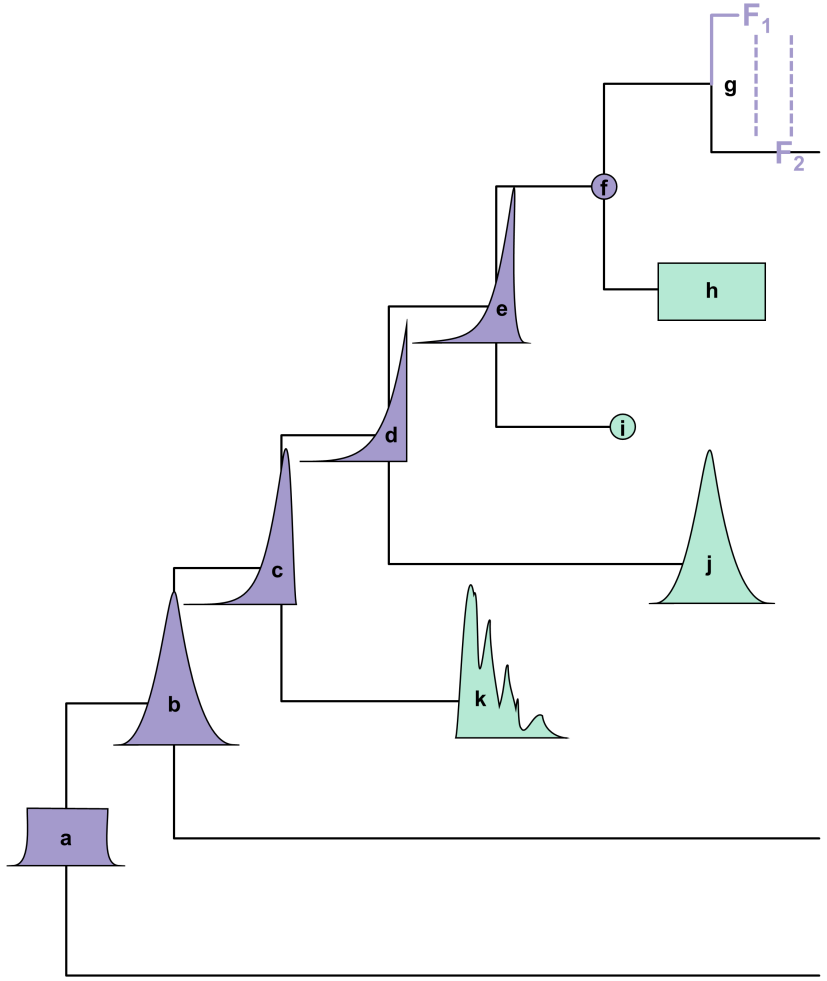
Rannala and Yang (*Systematic Biology* 2007)

Fossil calibration

Probability densities for describing uncertainty in fossil calibrations

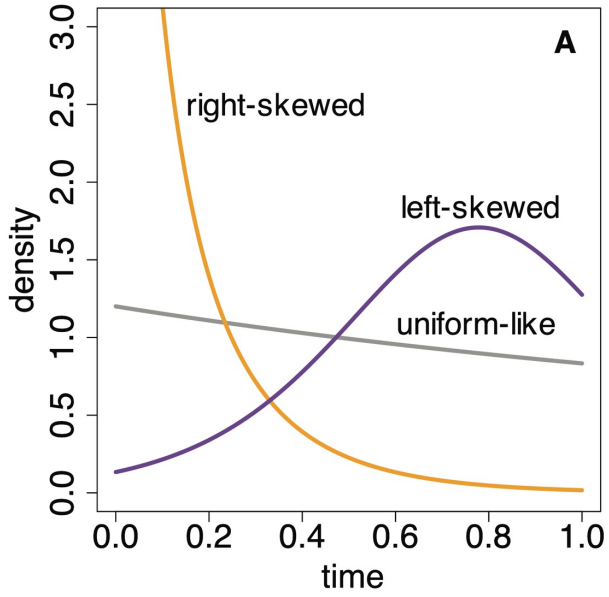


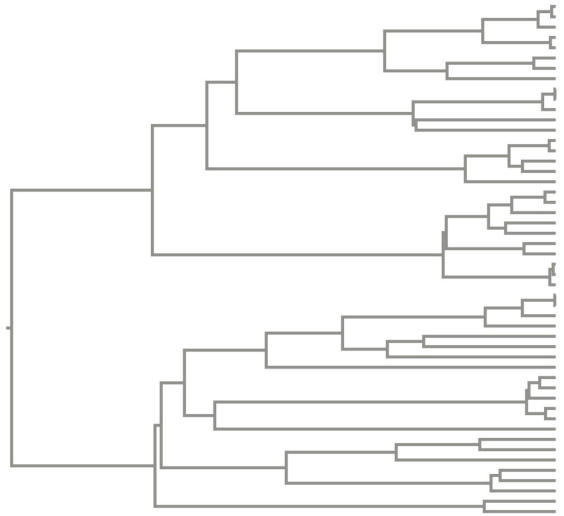
Approaches to represent uncertainty of calibrations in phylogenetic tree



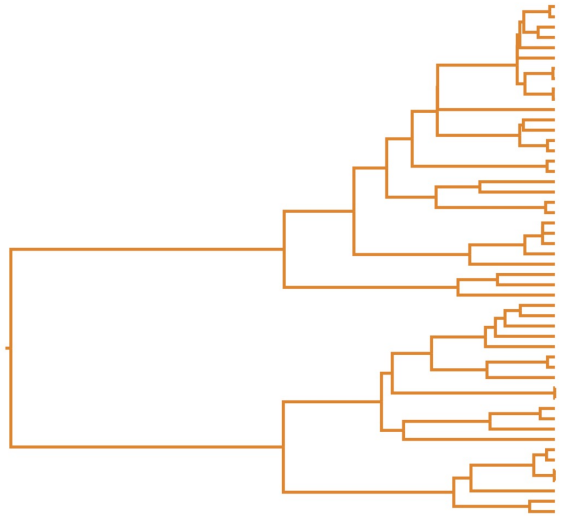
Cladogenesis model

- **BD process with species sampling:**
  - Specified by a per-lineage birth rate λ, a per-lineage death rate μ and a sampling fraction ρ

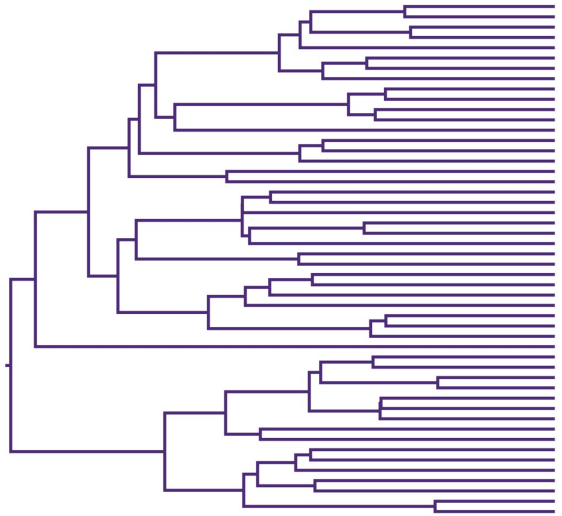




(B) uniform-like



(C) right-skewed

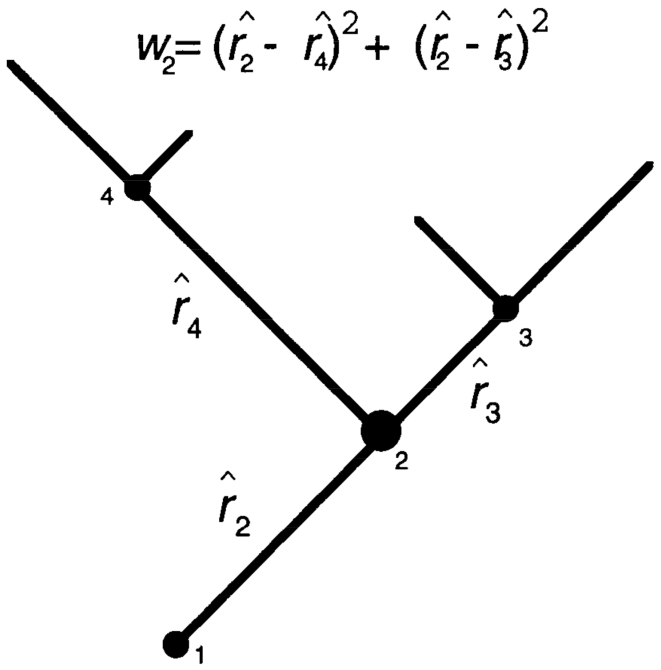


(D) left-skewed

Yang and Rannala (*MBE* 1997)

Rapid relaxed clock methods

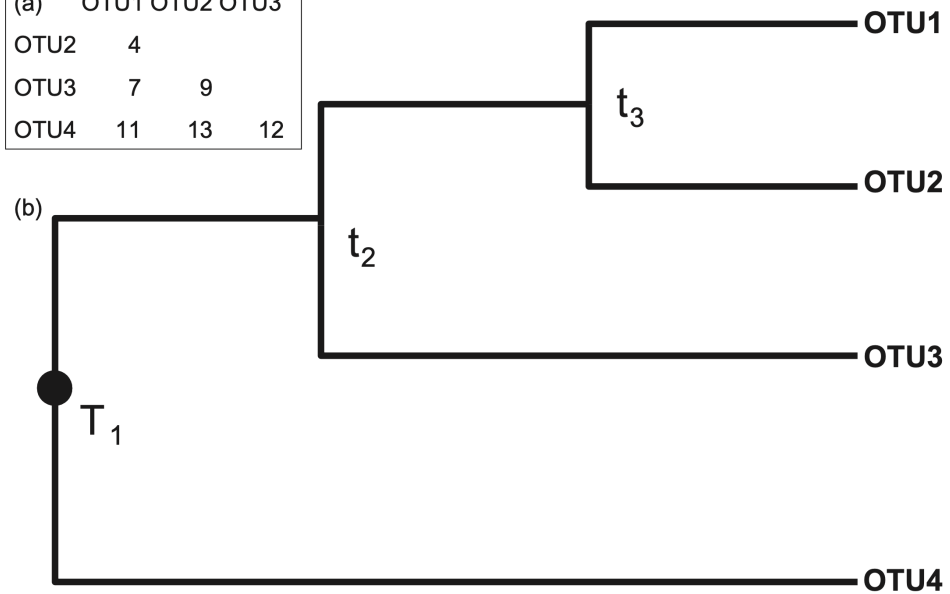
Penalized likelihood



Sanderson (1997, 2002)

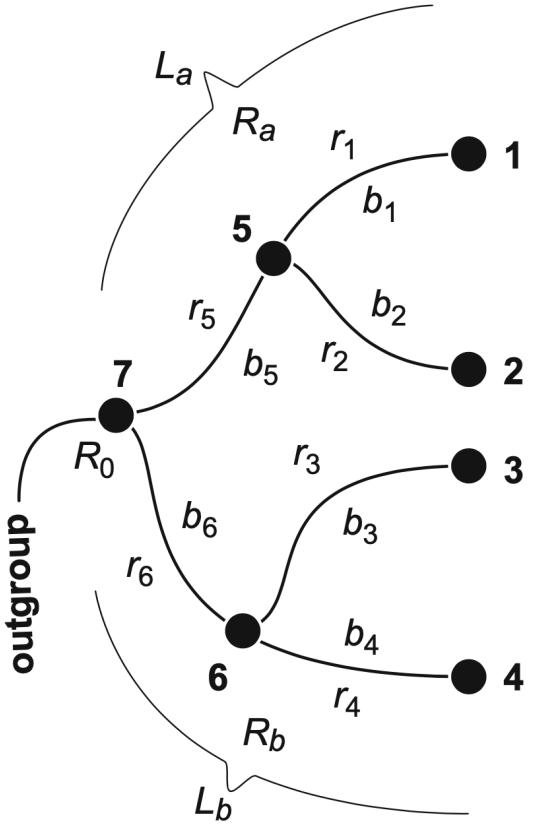
Least-squares

(a)	OTU1	OTU2	OTU3
OTU2	4		
OTU3	7	9	
OTU4	11	13	12



Xia and Yang (2011); To *et al.* (2016)

Relative rate framework



Tamura *et al.* (2012, 2018)