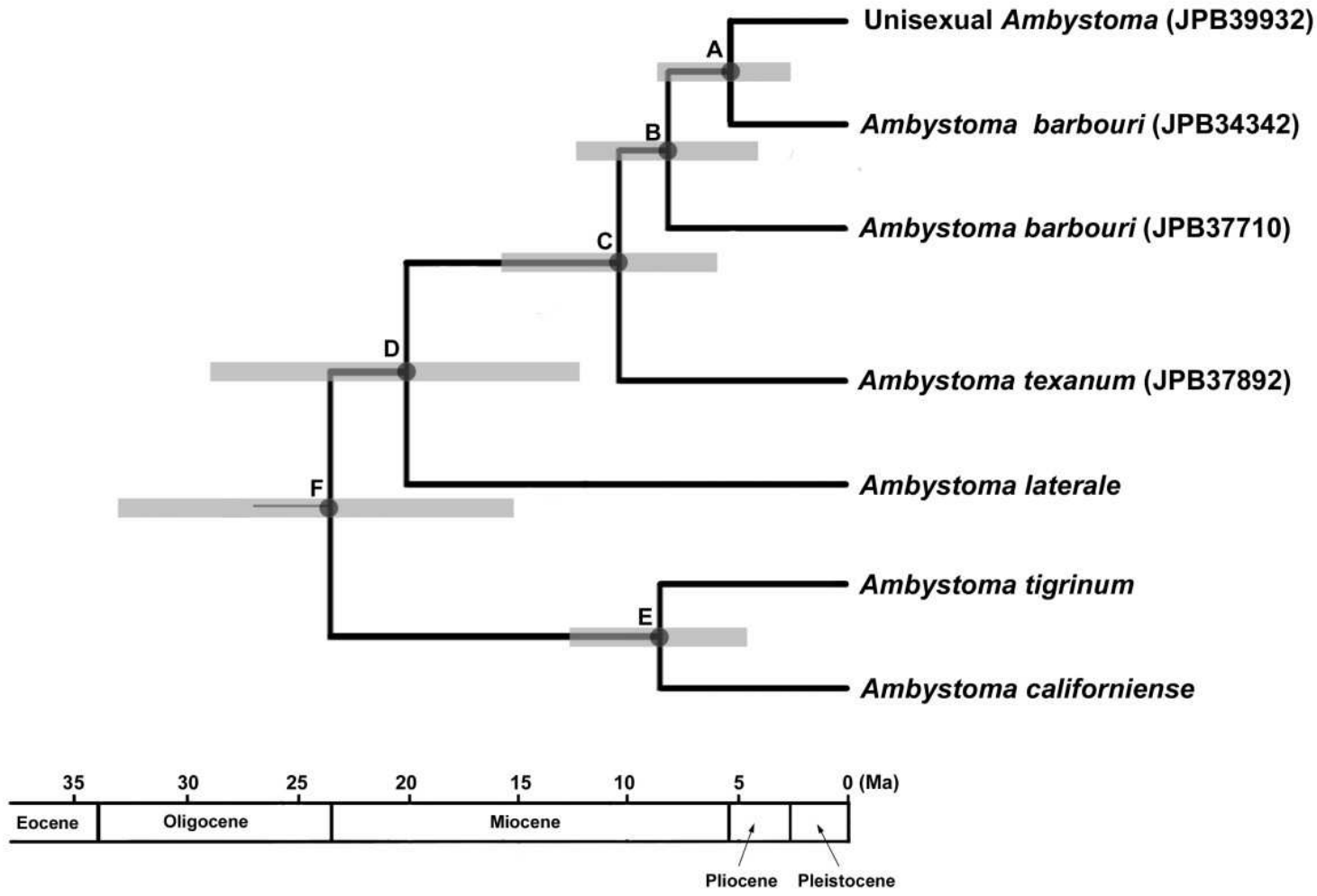
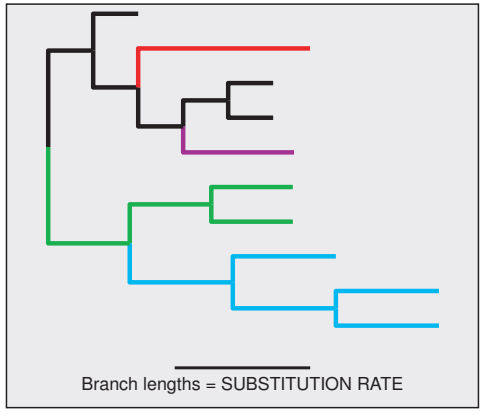
Divergence Time Estimation

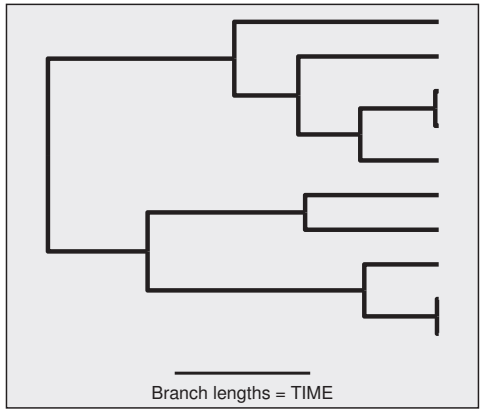


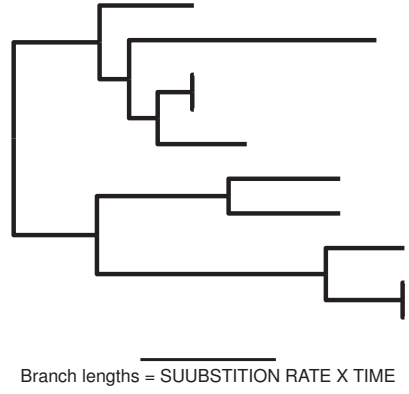
Bi & Bogart 2010. *BMC Evol. Biol.*

Divergence Time Estimation

Methods estimate the SUBSTITUTION RATE and TIME separately







Unconstrained methods provide branch length estimates that are the product of SUBSTITUTION RATE and TIME

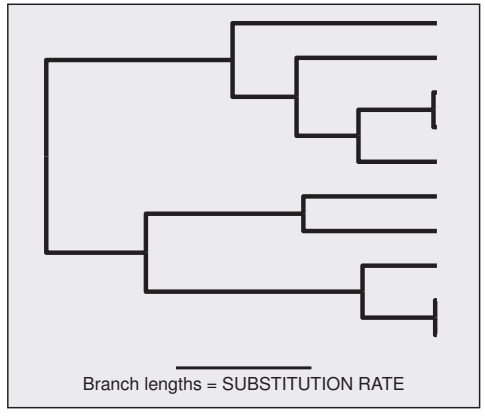
Models of Substitution-Rate Variation

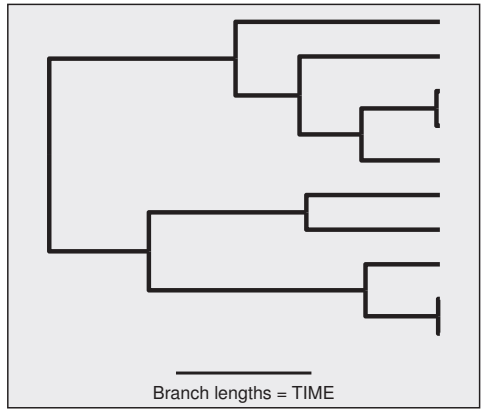
These models make assumptions about how substitution rate changes over the tree

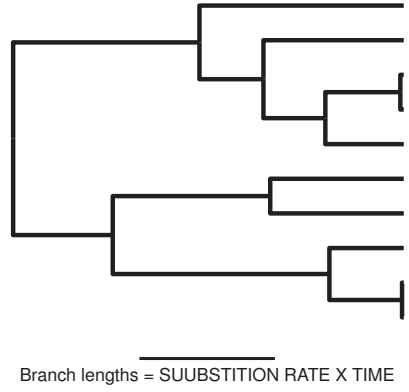
- Global molecular clock (Zuckerkandl & Pauling, 1962)
- Local molecular clocks (Kishino, 1990; Rambaut & Bromham 1998; Yang & Yoder, 2003)
- Compound Poisson process (Huelsenbeck et al., 2000)
- Autocorrelated rates – Log-normal distribution (Thorne et al., 1998; Kishino & Thorne, 2001; Thorne et al., 2002)
- Uncorrelated rates (Drummond et al., 2006, Rannala & Yang, 2007, Lepage et al., 2007)
- Non-parametric rate smoothing/Penalized likelihood (Sanderson, 1997, 2002)
- Ornstein-Uhlenbeck Process (Aris-Brosou & Yang, 2002)
- Cox–Ingersol–Ross process (Lepage et al., 2006)

Global Molecular Clock

Substitution rate is constant across all lineages





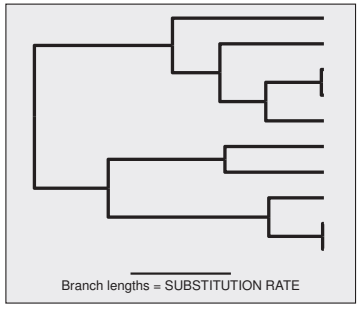


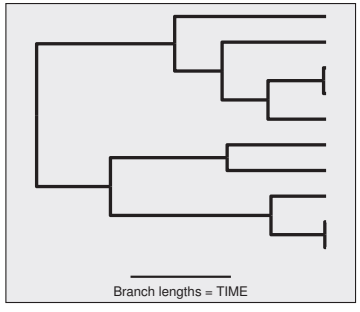
(Zuckerkandl & Pauling, 1962)

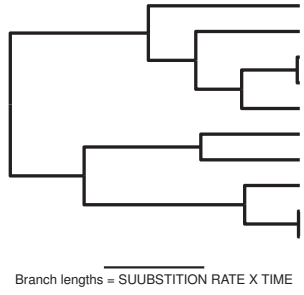
Rejecting the Global Molecular Clock

Incorrect models of sequence evolution lead to errors in the estimation of rates

- Almost any error in the model can lead to biases (or higher than needed variance) in detecting multiple hits
- Assumption of a Poisson clock can be wrong – even if we correctly count the number of changes, we don't account for over-dispersion (higher than Poisson-variance in the number of substitutions) (Cutler, 2000)



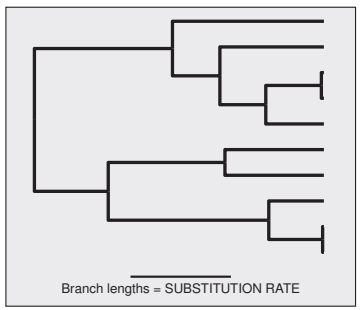


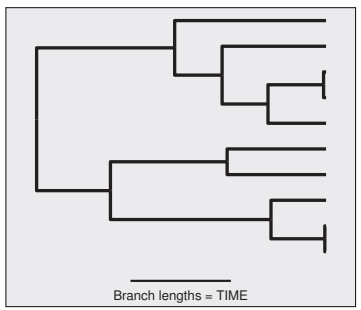


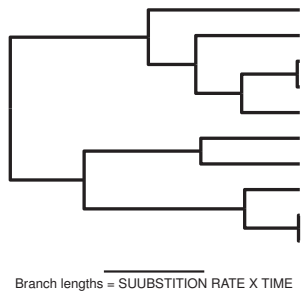
Rejecting the Global Molecular Clock

Rates of evolution can vary across lineages and over time

- mutation rates can vary (mutations per cell cycle, mutations per time, number of cell cycles per generation, generation time)
- strength and targets of selection can vary
- population sizes can vary

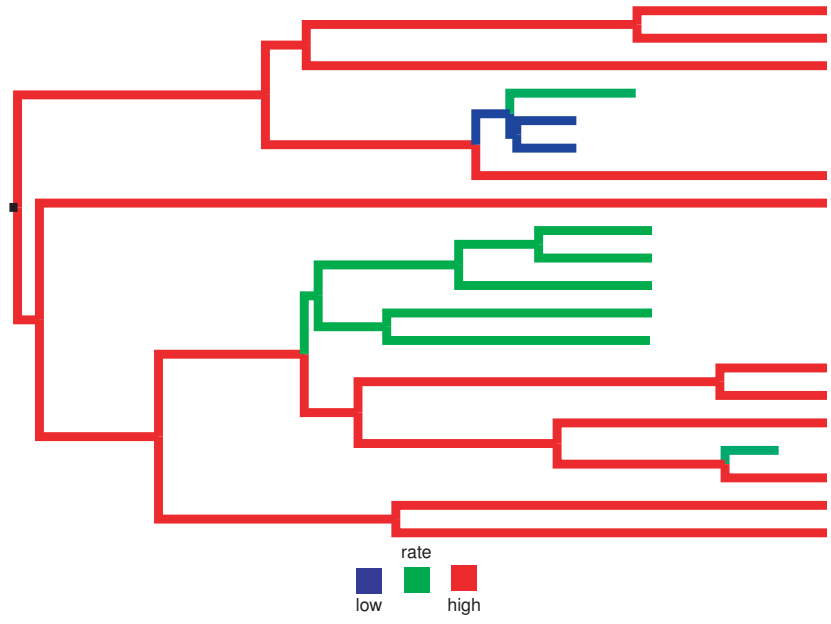






Local Molecular Clocks

Closely related lineages share the same rate — rates are clustered by subclades

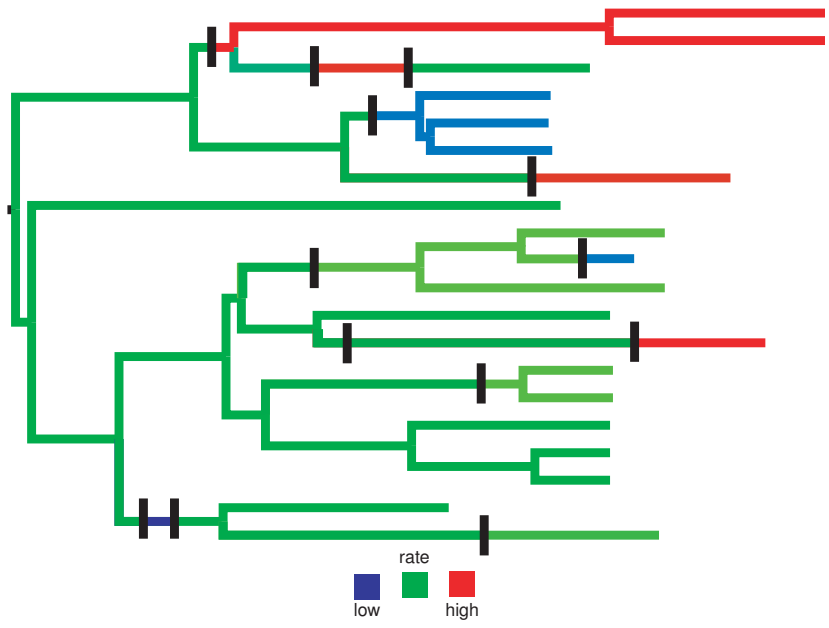


(Kishino, 1990; Rambaut & Bromham 1998; Yang & Yoder, 2003)

Compound Poisson Process

Rate changes occur along lineages according to a point process

At rate-change events, the new rate is a product of the old rate and a Γ - distributed multiplier

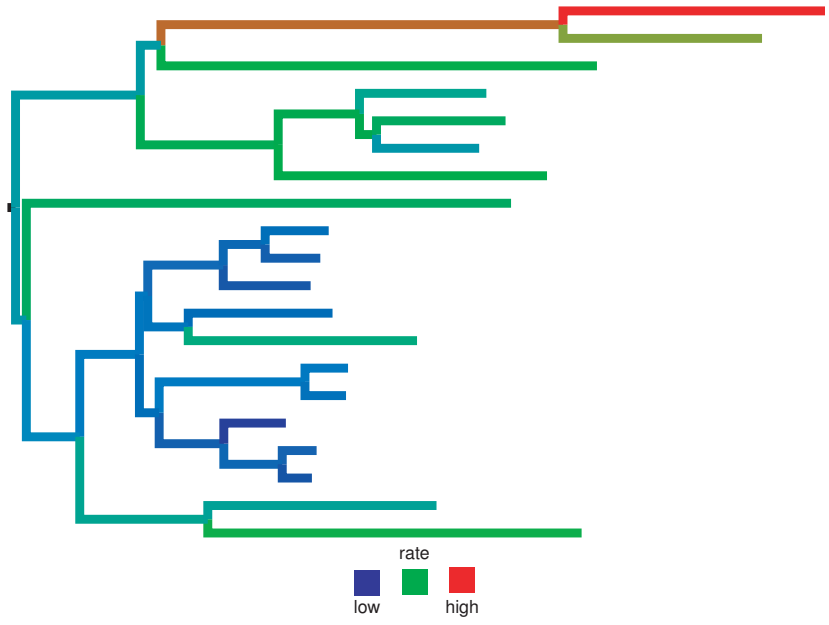


(Huelsenbeck et al., 2000)

Autocorrelated Rates (Log-normal)

Substitution rates evolve gradually over the tree - closely related lineages have similar rates

The rate at a node is drawn from a lognormal distribution with a mean equal to the parent rate

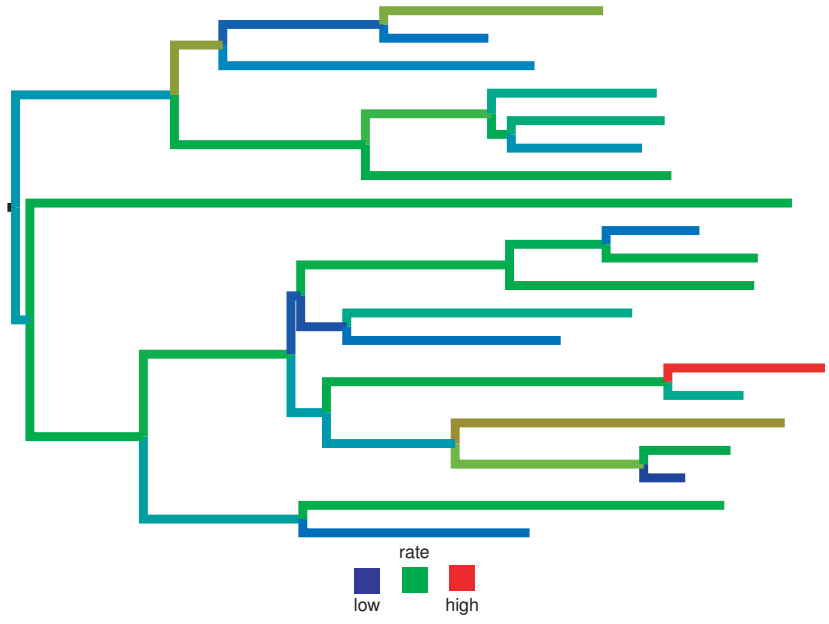


(Thorne et al., 1998; Kishino & Thorne, 2001; Thorne et al., 2002)

Uncorrelated Rates (Log-normal or Gamma)

The rates associated with each lineage are drawn, independently from a log-normal or gamma distribution

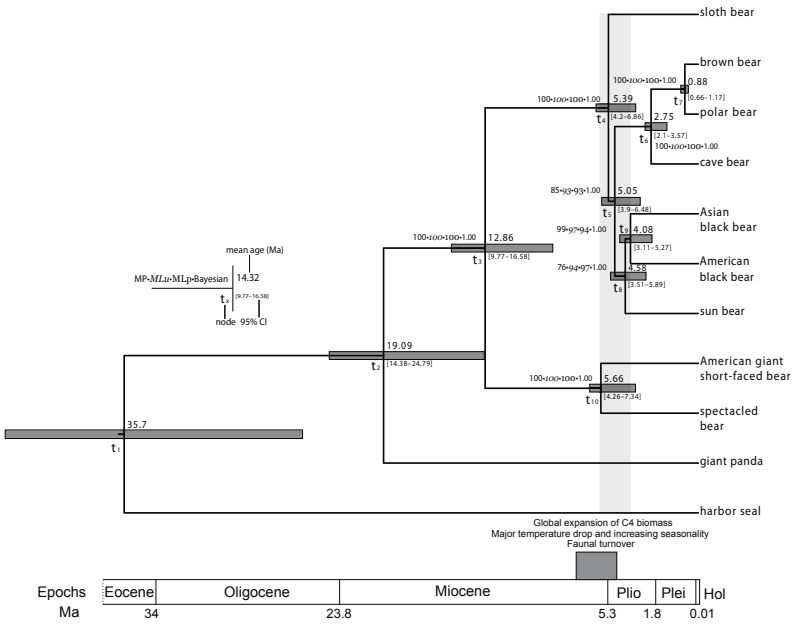
Common models used in BEAST



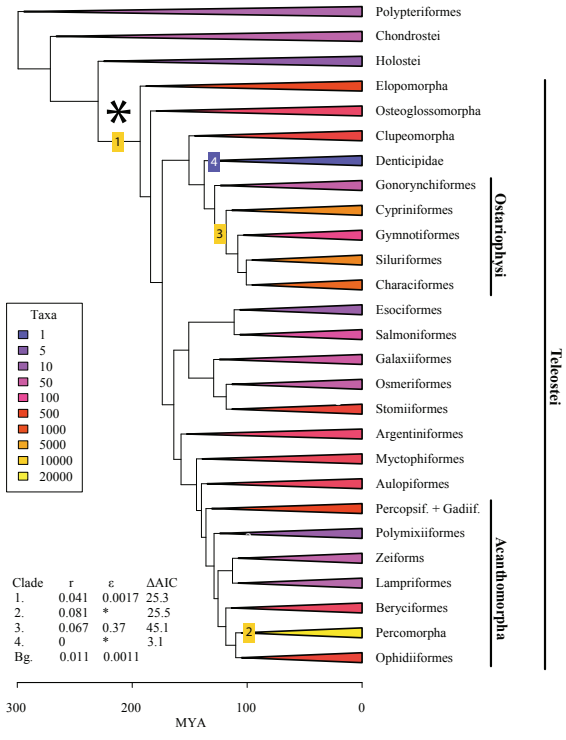
(Drummond et al., 2006)

Models of Substitution-rate Variation

Are our models appropriate across all data sets?



Krause et al., 2008. Mitochondrial genomes reveal an explosive radiation of extinct and extant bears near the Miocene-Pliocene boundary. BMC Evol. Biol. 8.



Santini et al., 2009. Did genome duplication drive the origin of teleosts? A comparative study of diversification in ray-finned fishes. BMC Evol. Biol. 9.

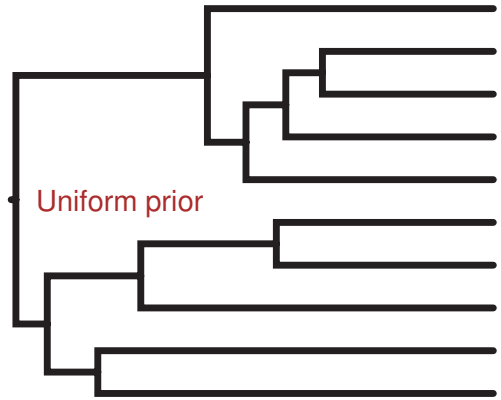
Divergence Time Estimation Software

Program	**Models/Method**
r8s	Strict clock, local clocks, NPRS, PL
ape (R)	NPRS, PL
multidivtime	log-n autocorrelated (plus some others)
PhyBayes	OU, log-n autocorrelated (plus some others)
PhyloBayes	CIR, white noise (uncorrelated) (plus some others)
BEAST	Uncorrelated (log-n & gamma), local clocks
TreeTime	Dirichlet model, CPP, uncorrelated
RevBayes	CPP, strict clock, DPP, autocorrelated

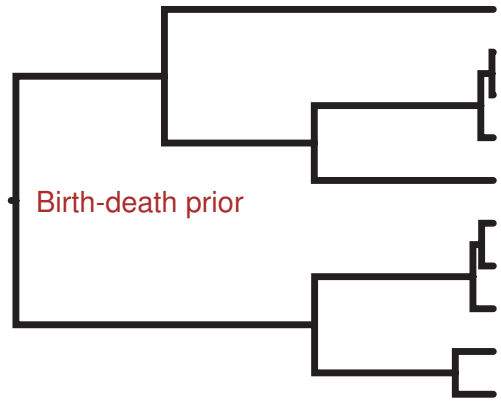
Priors on Node Times

Relaxed clock Bayesian analyses require a prior distribution on node times

Uniform prior: the time at a given node has equal probability across the interval between the time of the parent node and the time of the oldest daughter node

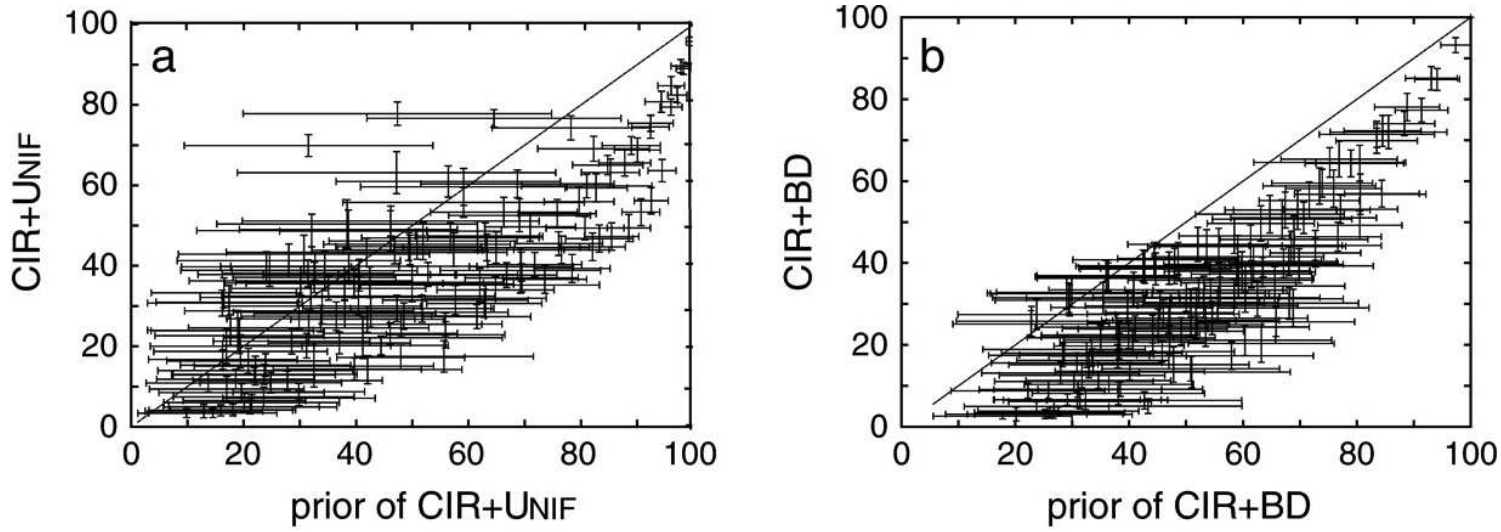


Birth-death prior: node times are sampled from a stochastic process with parameters for speciation and extinction (and in some cases taxon sampling)



Priors on Node Times

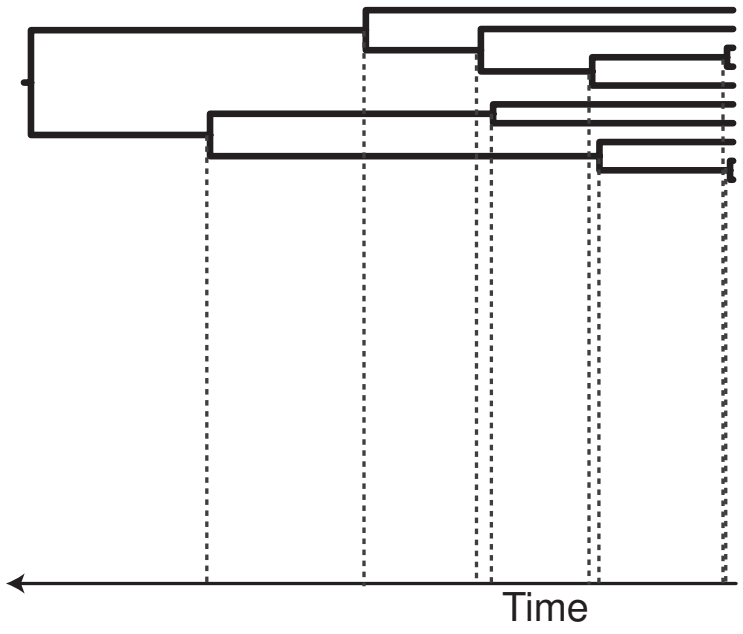
A comparison of the prior and posterior estimates of relative node ages using an autocorrelated rates model and the uniform or birth-death priors on node times



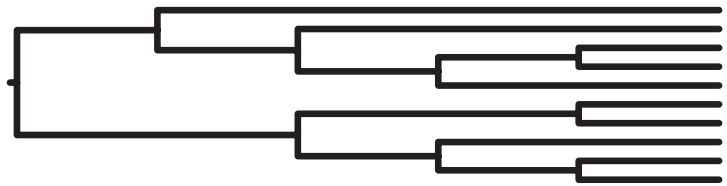
Lepage et al., 2007. A General Comparison of Relaxed Molecular Clock Models. *MBE* 24:2669–2680.

Calibrating the Tree

Goal: branch lengths in absolute time



Calibrating the Tree

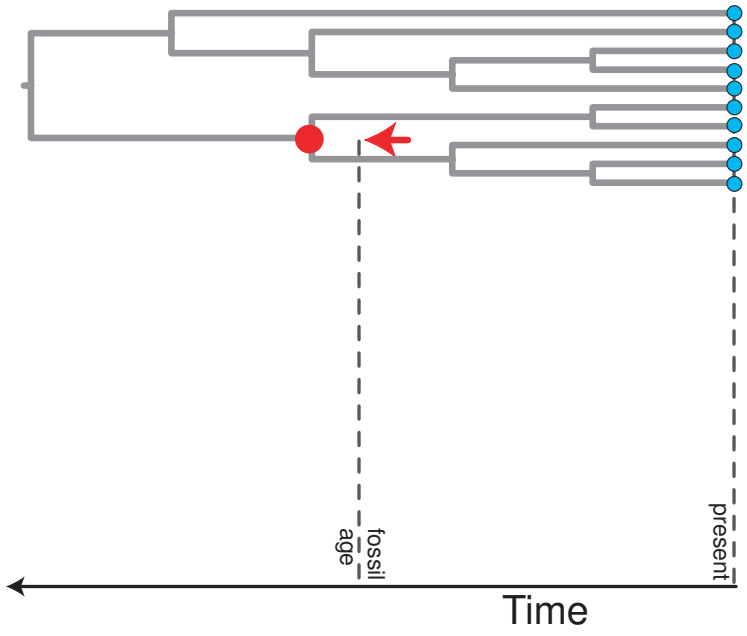


We have an estimate of the tree topology

Calibrating the Tree

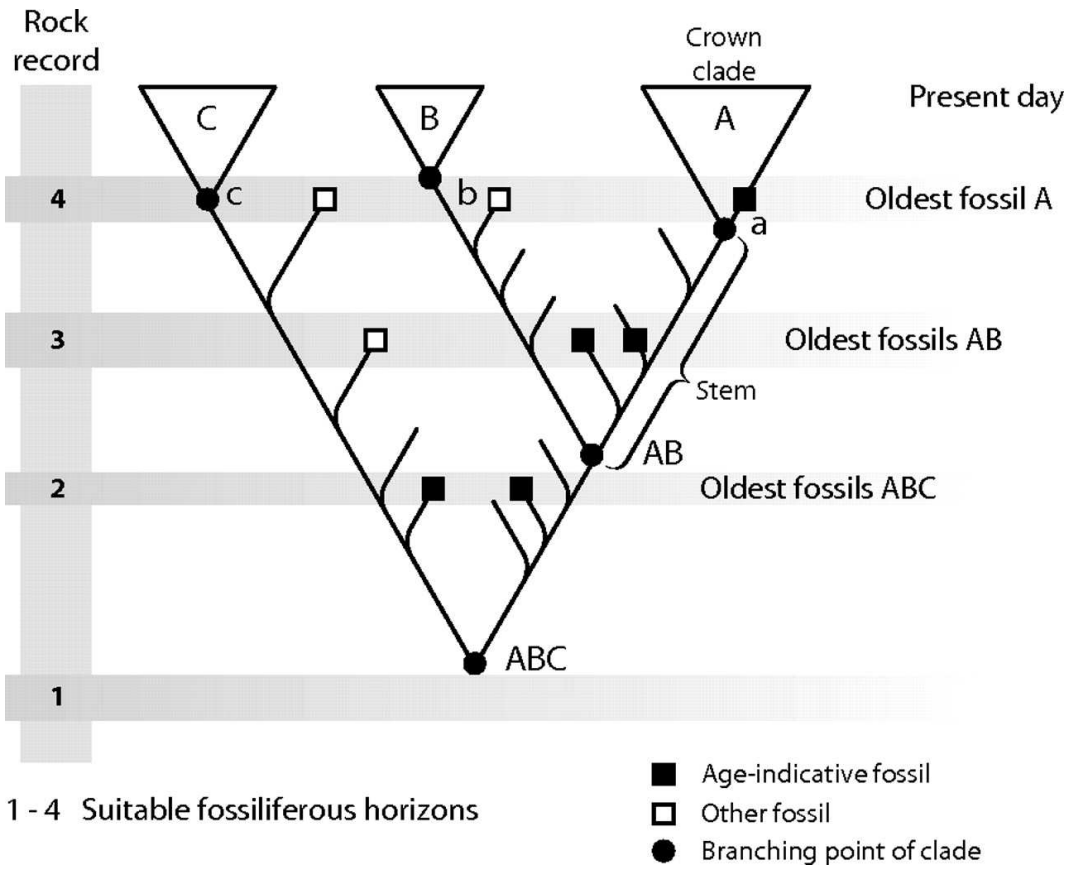
Known ages for sampled extant taxa

Estimates of minimum ages (from fossils or biogeographical data) that can be applied to nodes on the tree



Calibrating the Tree

Assigning fossils to clades

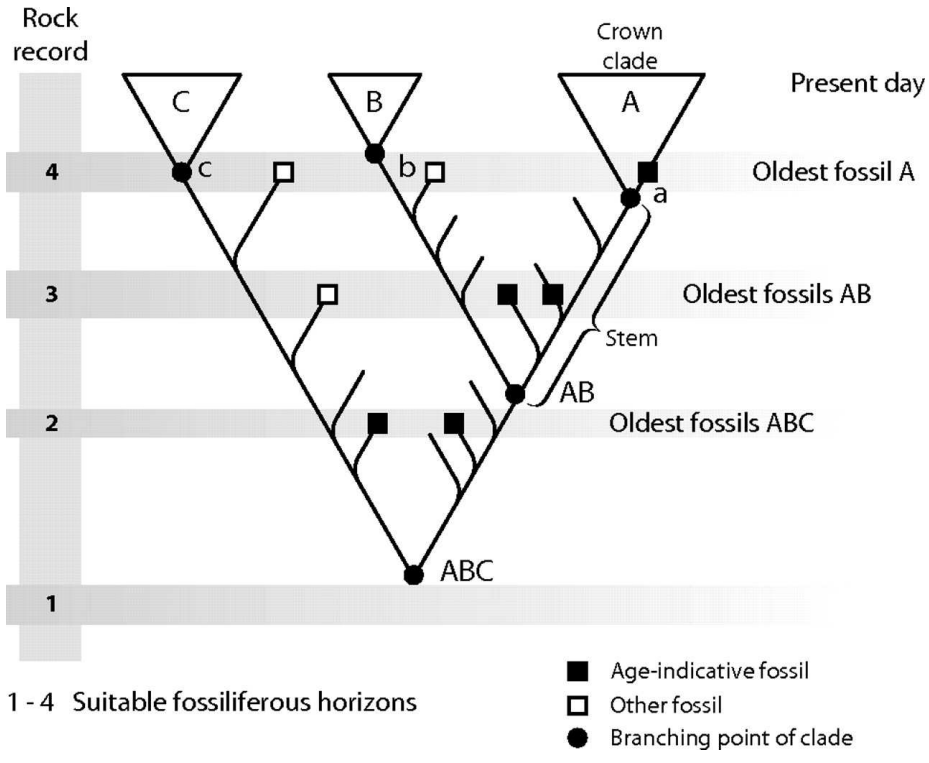


(Benton & Donoghue 2007 *Mol. Biol. Evol.* 24(1):26–53)

Calibrating the Tree

Assigning fossils to clades

Crown clade: all living species and their most-recent common ancestor (MRCA)

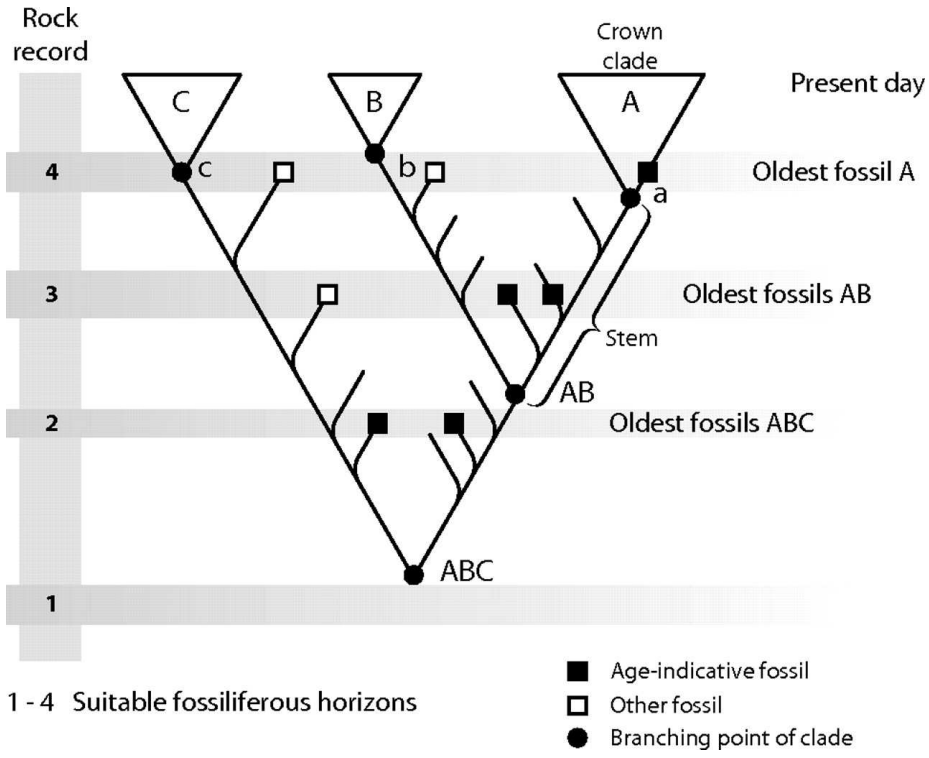


(Benton & Donoghue 2007 *Mol. Biol. Evol.* 24(1):26–53)

Calibrating the Tree

Assigning fossils to clades

Stem lineages: purely fossil forms that are closer to their descendant crown clade than any other crown clade

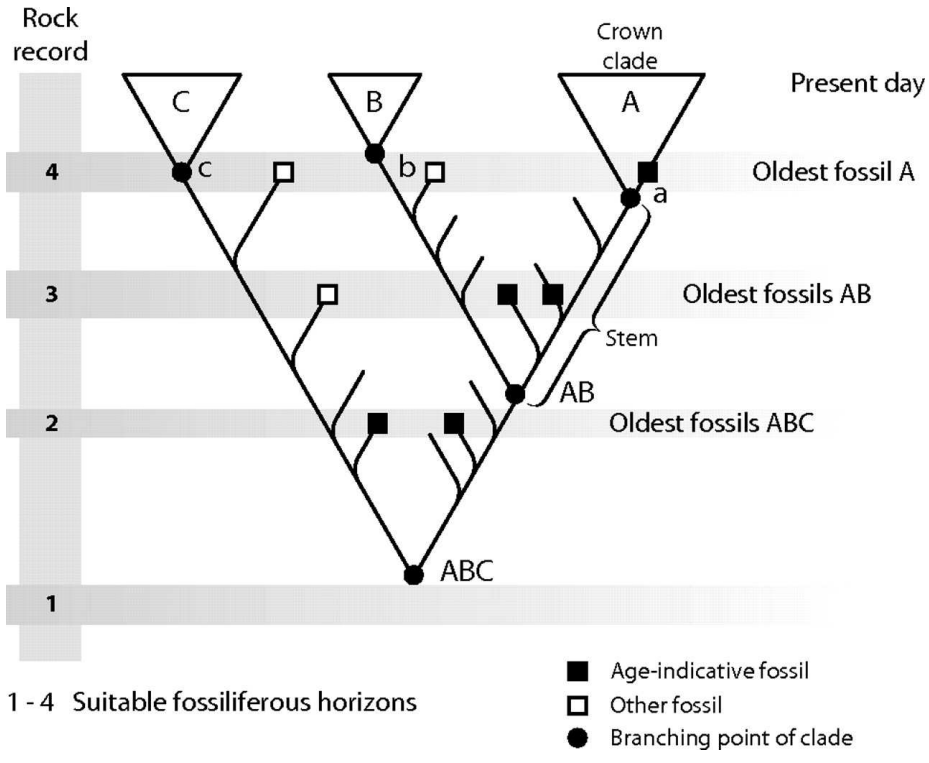


(Benton & Donoghue 2007 *Mol. Biol. Evol.* 24(1):26–53)

Calibrating the Tree

Assigning fossils to clades

Fossiliferous horizons: the sources in the rock record for relevant fossils

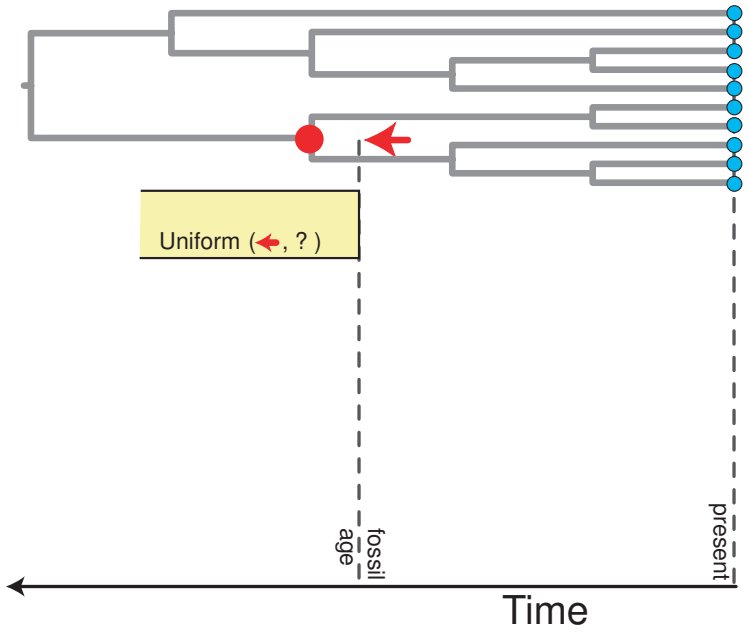


(Benton & Donoghue 2007 *Mol. Biol. Evol.* 24(1):26–53)

Priors on Calibrated Nodes

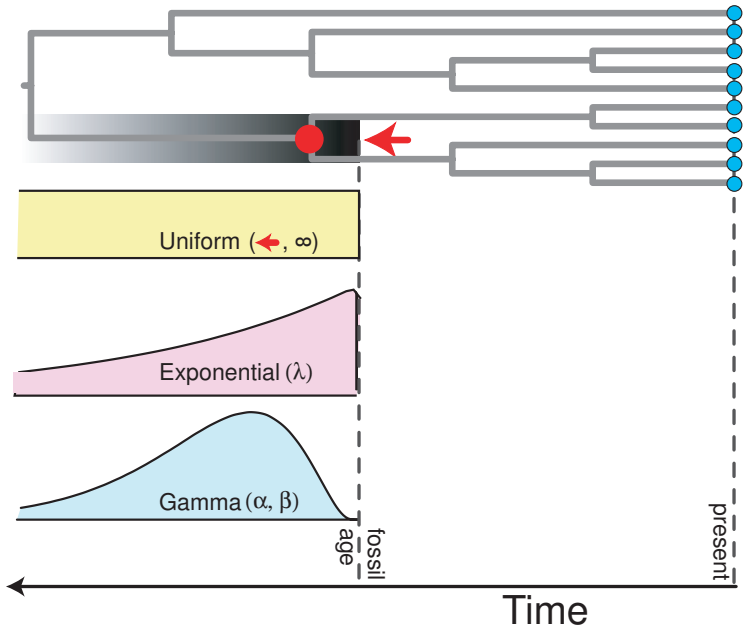
Fossils typically provide MINIMUM bounds for calibrating nodes

Reliable MAXIMUM bounds are difficult to obtain

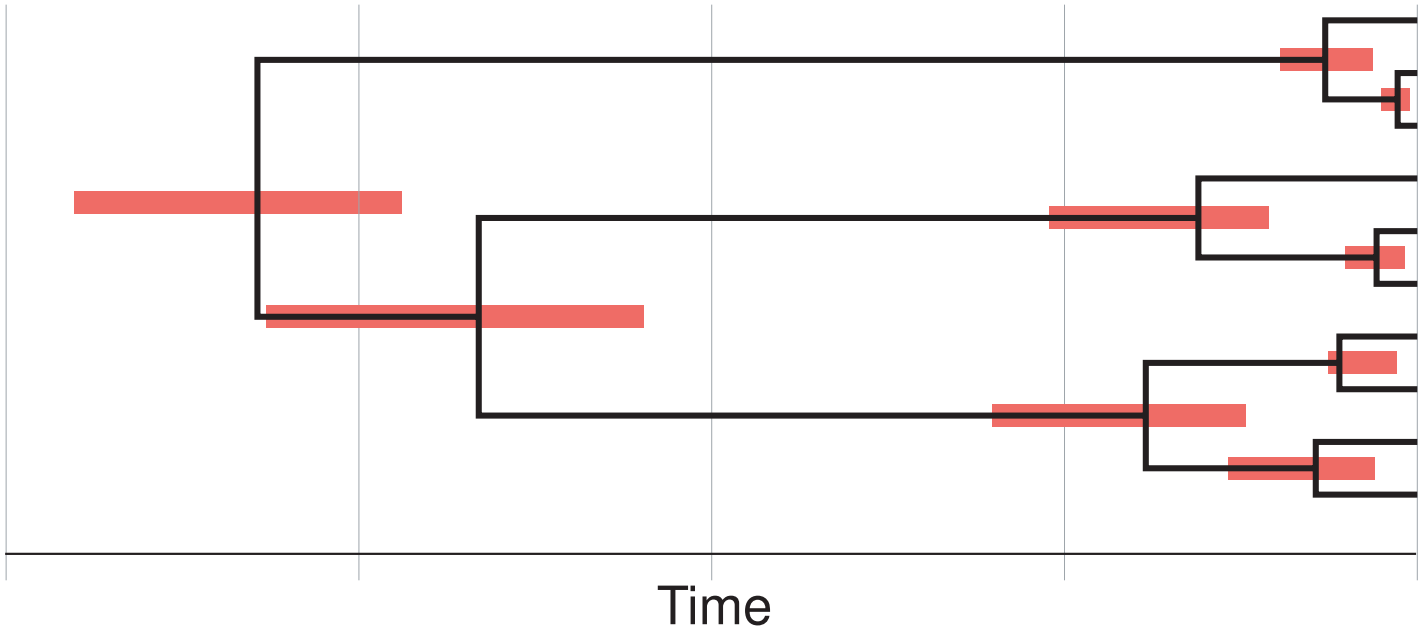


Priors on Calibrated Nodes

Different types of distributions that do not require maximum bounds can be applied to calibrated nodes



Bayesian Divergence time estimation



Criticism of Relaxed Clock Methods

- Dependent on and sensitive to fossil calibrations — fossil age estimates and node assignment are not without error
- Models are not biologically realistic
- Different methods/models can produce very different estimates of the same divergence times
- Priors are too informative
- Studies comparing methods have produced conflicting and unclear results

HIV Transmission in Libya

An outbreak of HIV (and HepC) among patients in a Libyan hospital resulted in the 8-yr imprisonment of 6 foreign medical workers

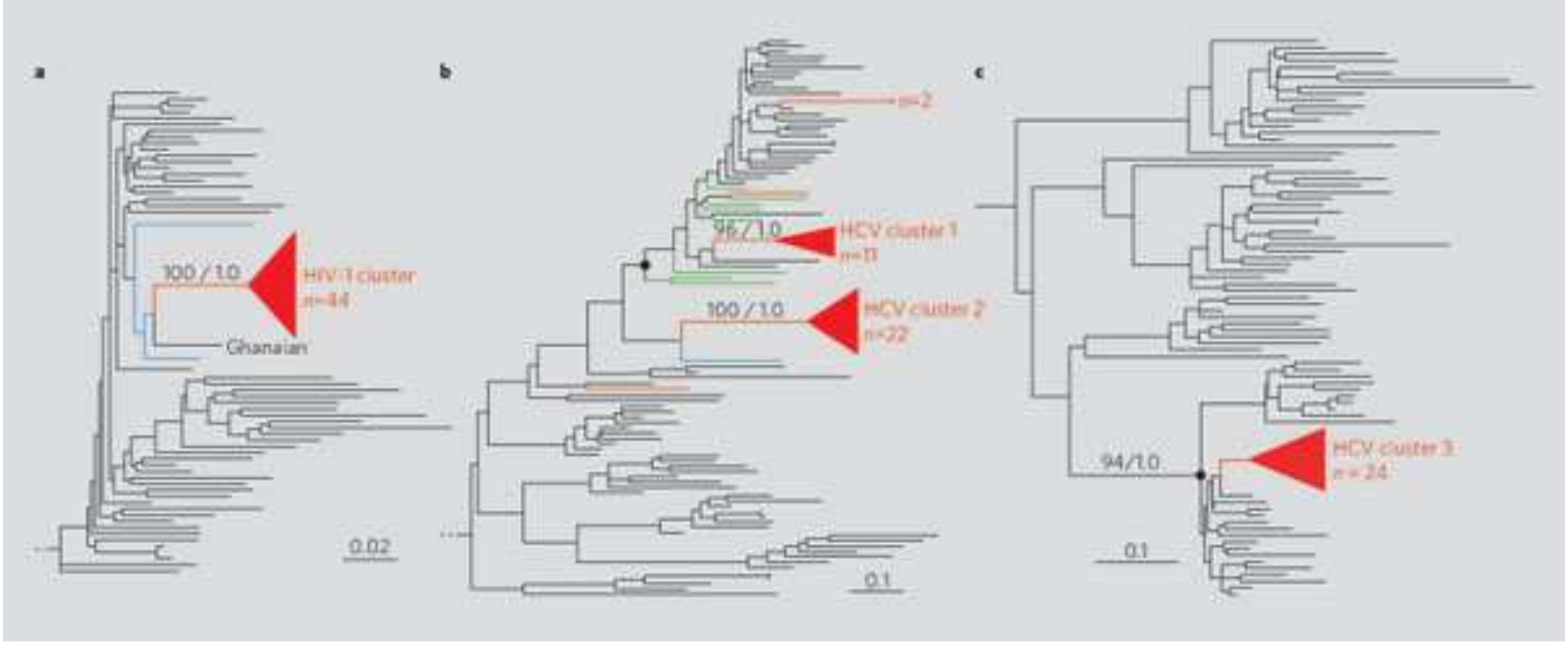
The defendants were accused of deliberately infecting over 400 children and sentenced to death

(Butler, 2007 *Nature*)

Prosecutors claimed the medical workers used the kids as test subjects in an illicit clinical trial

HIV Transmission in Libya

A study by HIV experts – Were the viral strains present in Libya before or after the arrival of the foreign medics (March 1998)?



(de Oliveria et al., 2006 *Nature*)

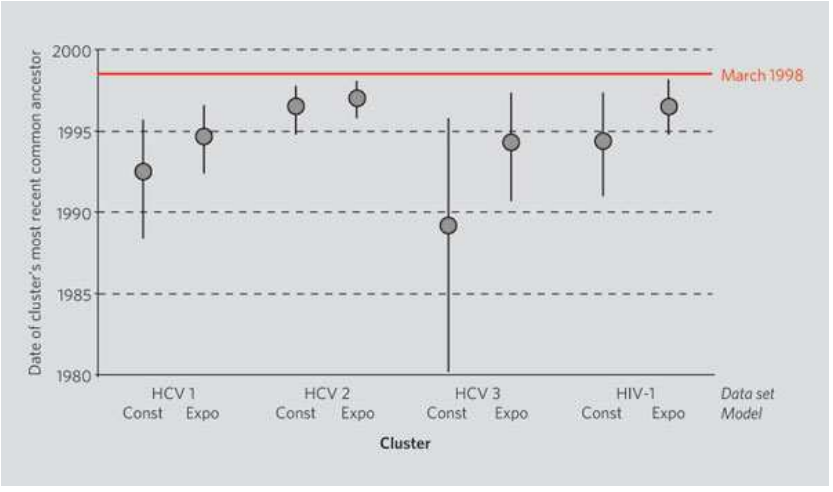
HIV Transmission in Libya

Strict and relaxed clock methods were used to estimate the date of the MRCA for each cluster

Every MRCA predated March 1998

Many different models were used to obtain robust results and build a solid case against deliberate infection

(de Oliveria et al., 2006 *Nature*)



This study supported the findings of a previous epidemiological study: the outbreak was the result of poor and unsanitary medical practices

HIV Transmission in Libya

The Libyan court denied the validity of the findings and held that the outbreak was deliberate

International outrage and the phylogenetic analyses resulted in the eventual release of the defendants