Dating Trees and Diversification rates

Emily Jane McTavish

Life and Environmental Sciences
University of California, Merced
ejmctavish@ucmerced.edu, twitter:snacktavish

(With thanks to Tracy Heath for slides!)

Many research questions require time-scaled phylogenies

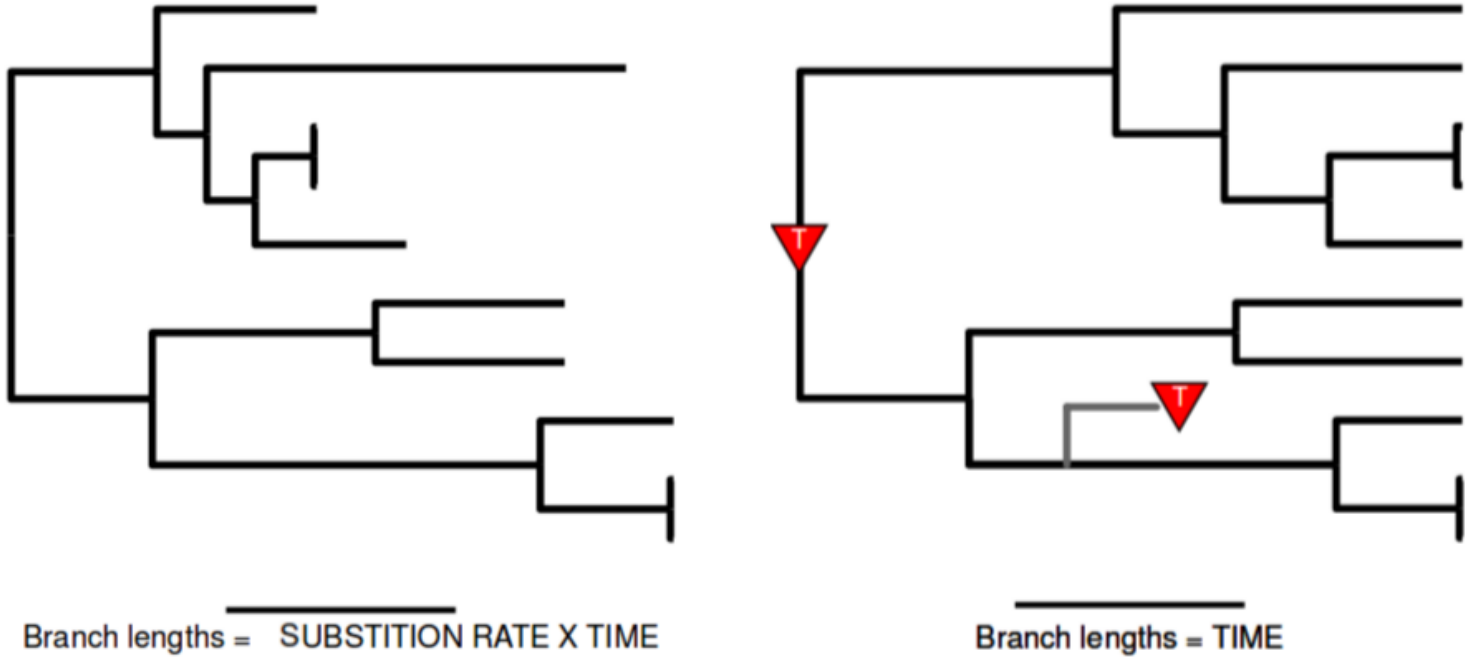
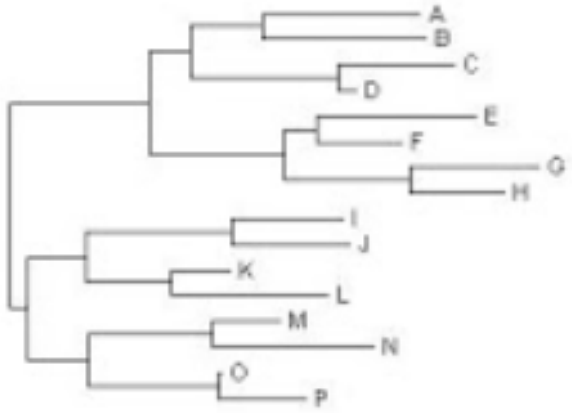
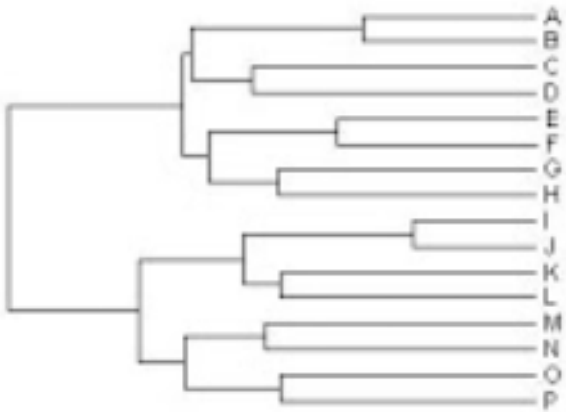


figure from Tracy Heath



NON-ULTRAMETRIC-TREE

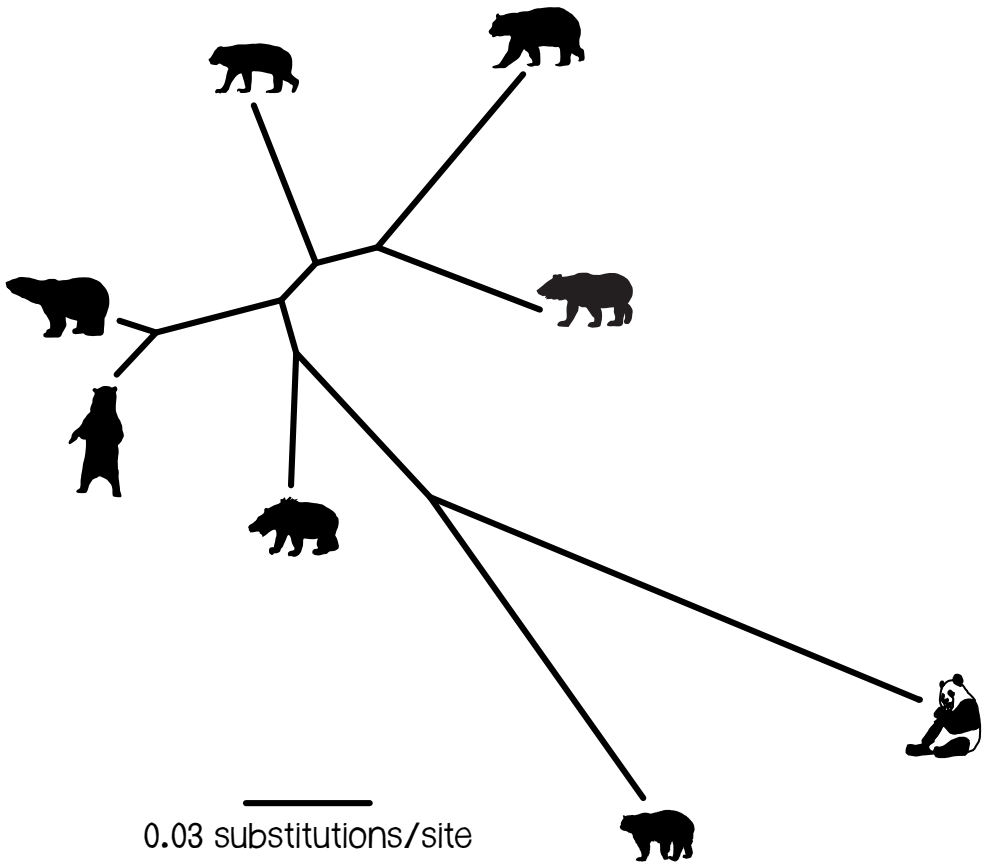


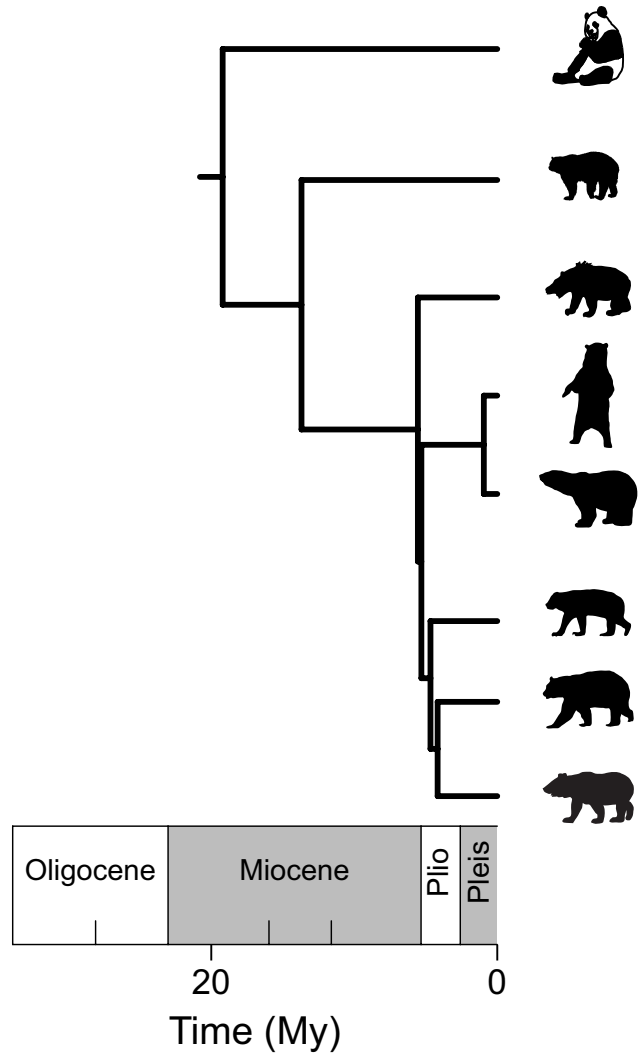
ULTRAMETRIC-TREE

figure from Indelible

A Time-Scale for Evolution

Phylogenies with branch lengths proportional to time provide more information about evolutionary history than unrooted trees with branch lengths in units of substitutions/site.

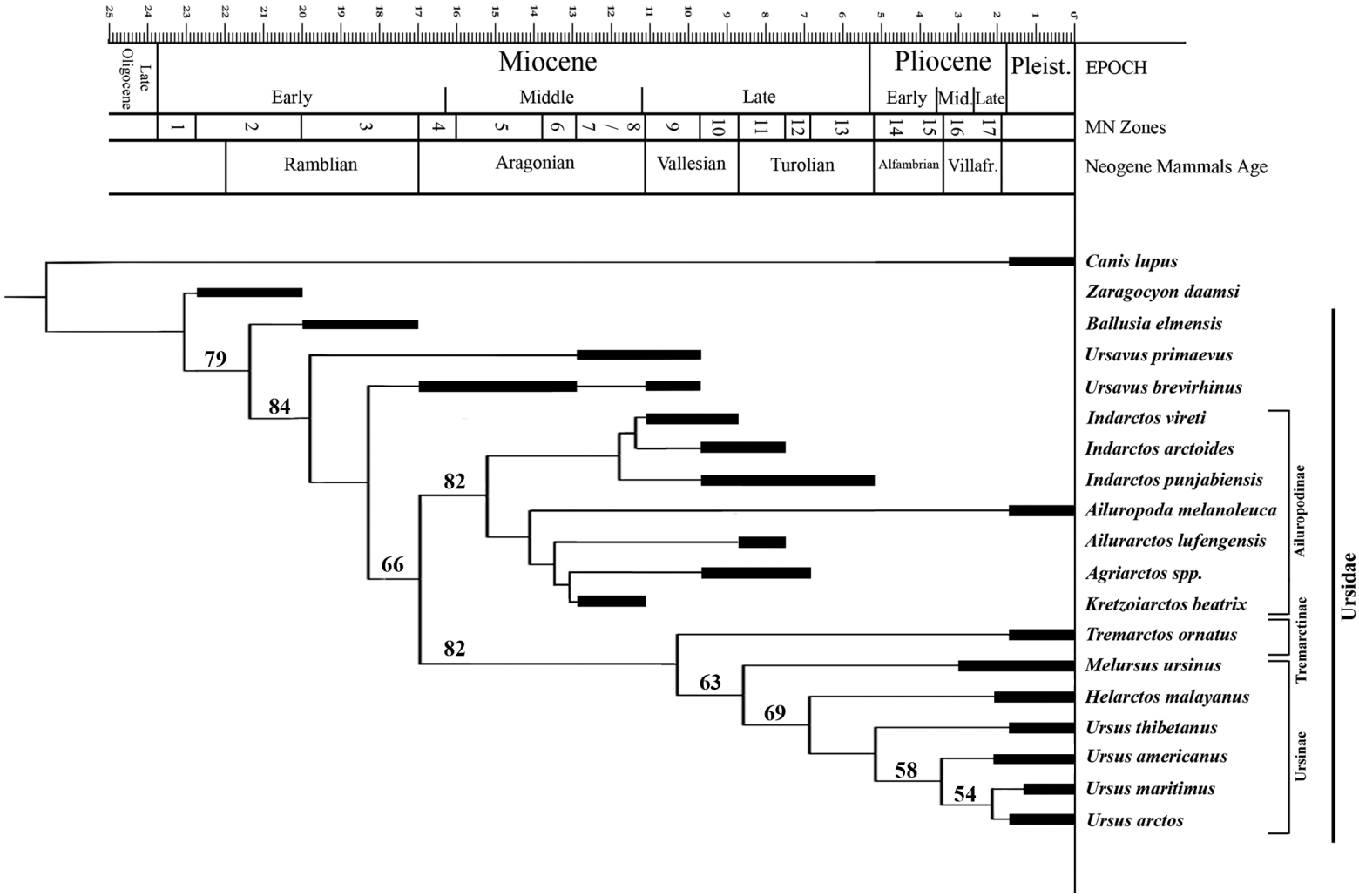




(silhouette images from `http://phylopic.org`)

A Time-Scale for Evolution

If fossil taxa are available, then these data can be incorporated into macroevolutionary studies

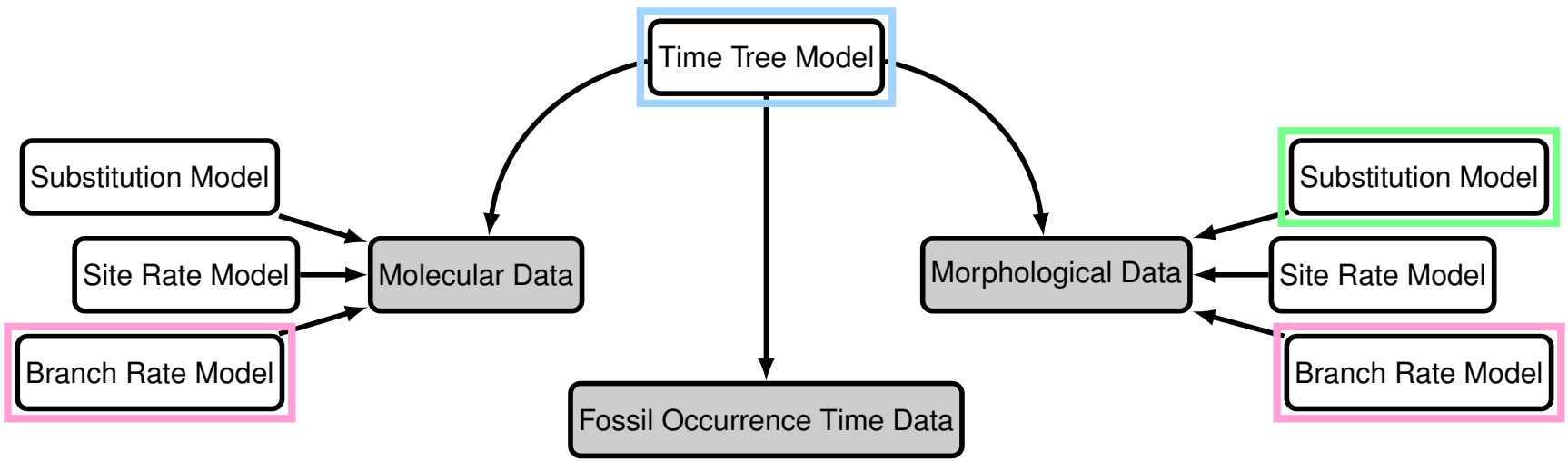


(Abella et al. PLoS One 2012)

Where to get dates? `https://nextstrain.org/ncov/global?l=clock`

Model Framework

Combine models for sequence evolution, morphological change, & fossil recovery to jointly estimate the tree topology, divergence times, & lineage diversification rates

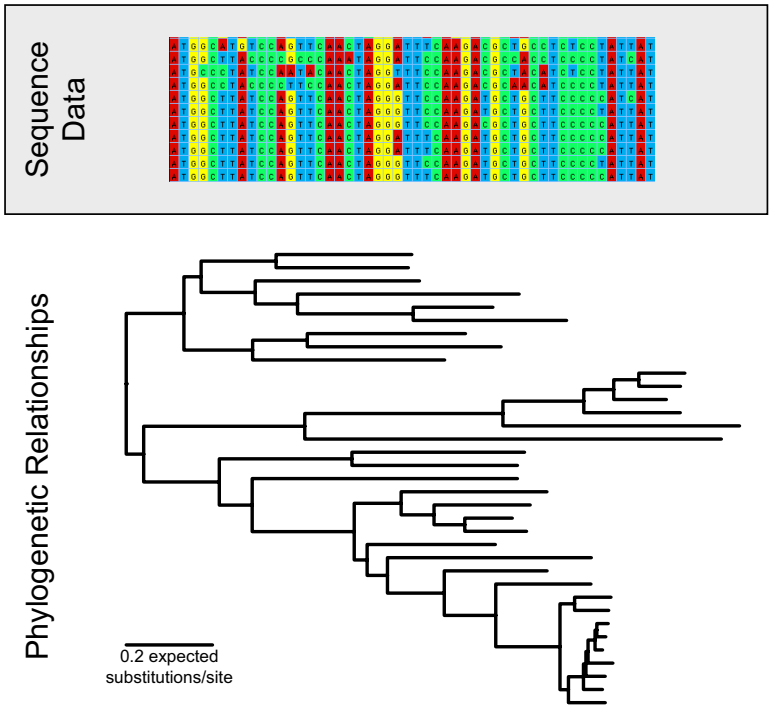


Estimating Rate & Time

Sequence data provide information about **branch lengths**

In units of **the expected # of substitutions per site**

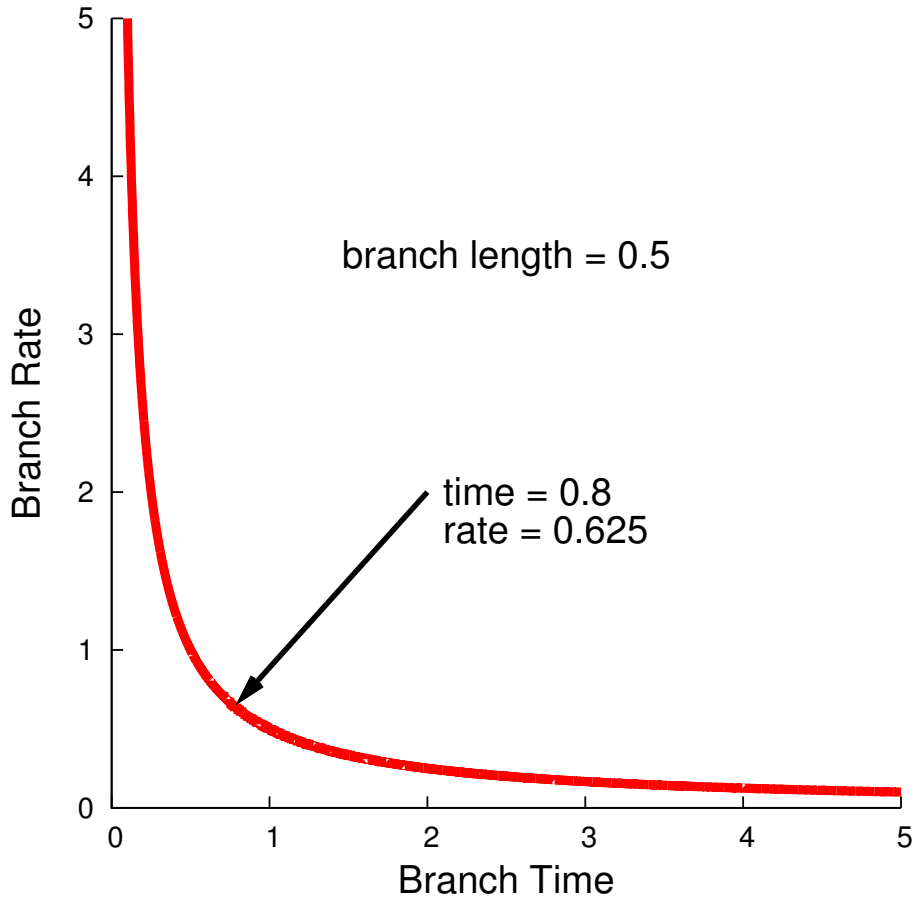
branch length = rate × time



Estimating Rate & Time

The sequence data provide information about branch length

for any possible rate, there's a time that fits the branch length perfectly

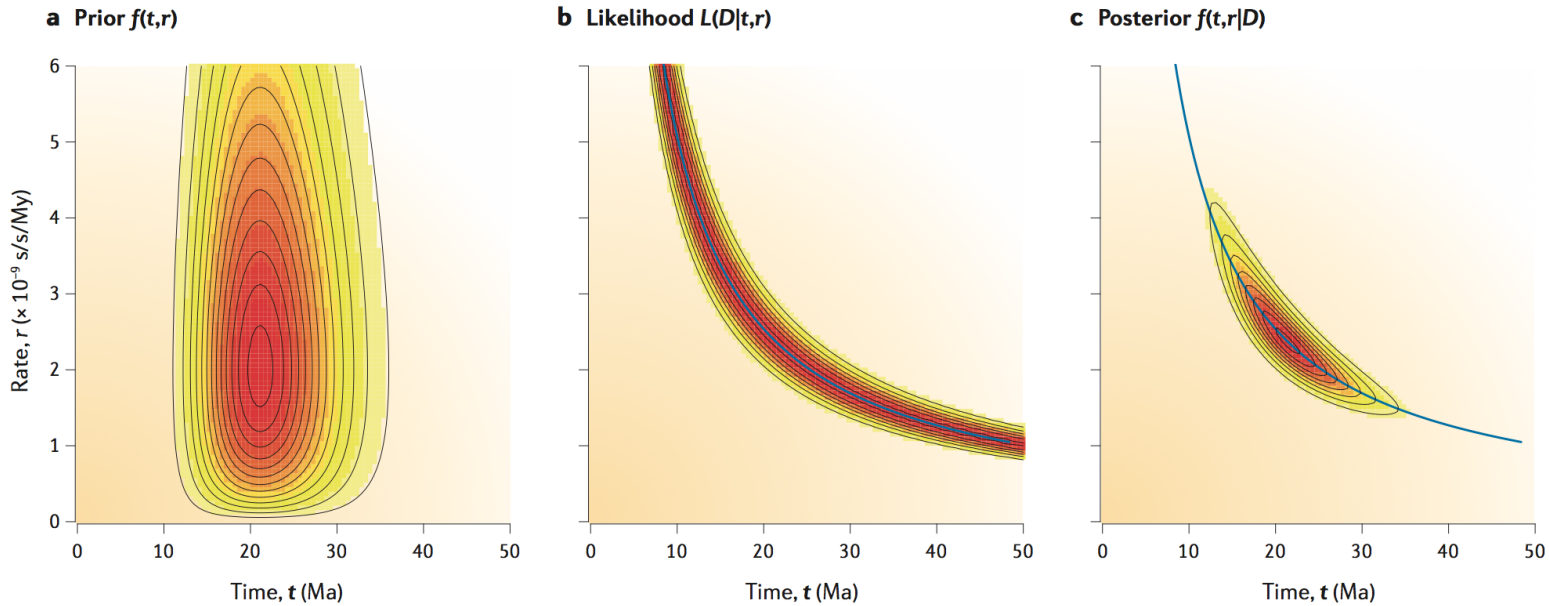


(figure based on Thorne & Kishino, 2005)

Methods for dating species divergences estimate the substitution rate and time separately

Estimating Rate & Time

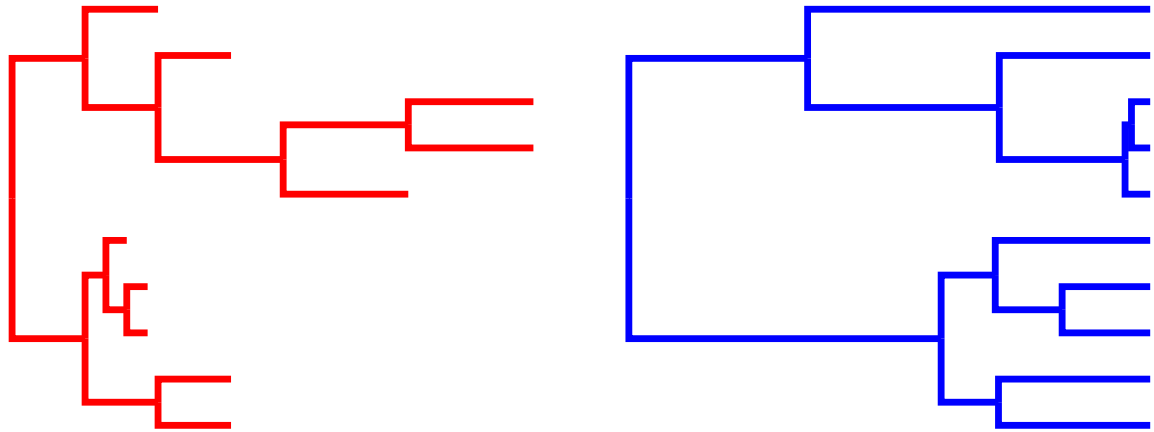
Methods for dating species divergences estimate the substitution rate and time separately



(dos Reis et al. *Nature Reviews Genetics*, 2016)

Tree-time priors for molecular phylogenies are only informative on a **relative** time scale

Bayesian Divergence Time Estimation

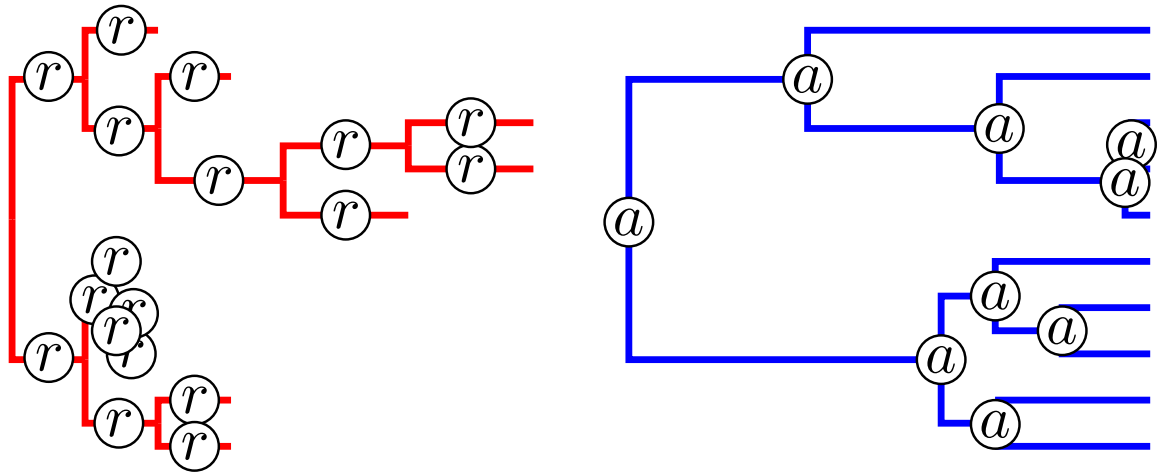


$$\boldsymbol{\mathcal{R}} = (r_1, r_2, r_3, \ldots, r_{2N-2})$$

$$\boldsymbol{\mathcal{A}} = (a_1, a_2, a_3, \ldots, a_{N-1})$$

$$N = \text{number of tips}$$

Bayesian Divergence Time Estimation



$$\boldsymbol{\mathcal{R}} = (r_1, r_2, r_3, \ldots, r_{2N-2})$$

$$\boldsymbol{\mathcal{A}} = (a_1, a_2, a_3, \ldots, a_{N-1})$$

$$N = \text{number of tips}$$

Bayesian Divergence Time Estimation

Posterior probability

$$f(\boldsymbol{\mathcal{R}}, \boldsymbol{\mathcal{A}}, \theta_{\mathcal{R}}, \theta_{\mathcal{A}}, \theta_s, \mathcal{T} \mid D)$$

$\boldsymbol{\mathcal{R}}$	Vector of rates on branches
$\boldsymbol{\mathcal{A}}$	Vector of internal node ages
$\theta_{\mathcal{R}}, \theta_{\mathcal{A}}, \theta_s$	Model parameters
D	Molecular or morphology data
$\mathcal{T}$	Tree topology

Bayesian Divergence Time Estimation

$$f(\boldsymbol{\mathcal{R}}, \boldsymbol{\mathcal{A}}, \theta_{\mathcal{R}}, \theta_{\mathcal{A}}, \theta_s \mid D) = \frac{f(D \mid \boldsymbol{\mathcal{R}}, \boldsymbol{\mathcal{A}}, \theta_s)\, f(\boldsymbol{\mathcal{R}} \mid \theta_{\mathcal{R}})\, f(\boldsymbol{\mathcal{A}} \mid \theta_{\mathcal{A}})\, f(\theta_s)}{f(D)}$$

$f(D \mid \boldsymbol{\mathcal{R}}, \boldsymbol{\mathcal{A}}, \theta_{\mathcal{R}}, \theta_{\mathcal{A}}, \theta_s)$	Likelihood
$f(\boldsymbol{\mathcal{R}} \mid \theta_{\mathcal{R}})$	Prior on rates
$f(\boldsymbol{\mathcal{A}} \mid \theta_{\mathcal{A}})$	Prior on node ages
$f(\theta_s)$	Prior on substitution parameters
$f(D)$	Marginal probability of the data

Modeling Rate Variation

Some models describing lineage-specific substitution rate variation:

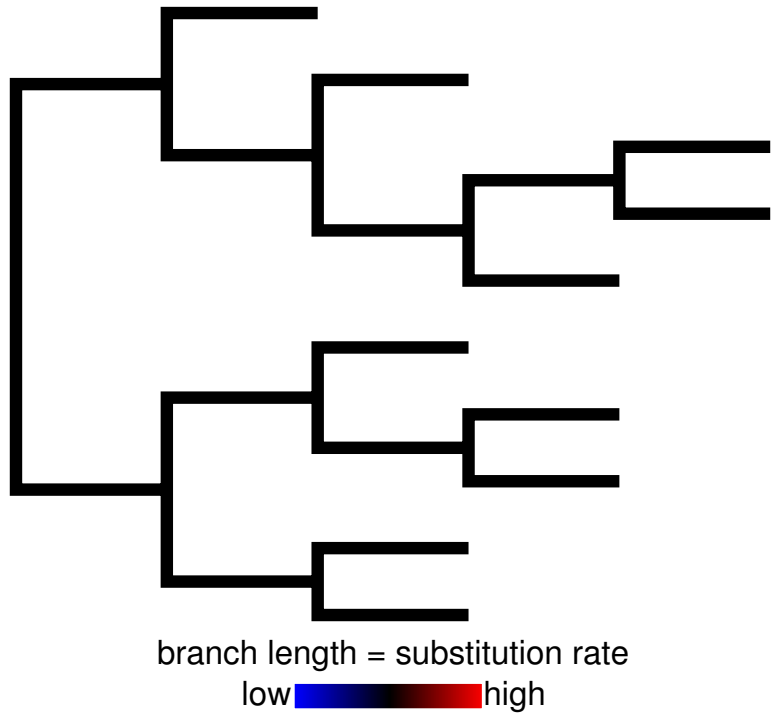
- **Global clock** (Zuckerkandl & Pauling, 1962)
- **Local clocks** (Hasegawa, Kishino & Yano 1989; Kishino & Hasegawa 1990; Yoder & Yang 2000; Yang & Yoder 2003, Drummond and Suchard 2010)
- **Punctuated rate change model** (Huelsenbeck, Larget and Swofford 2000)
- **Log-normally distributed autocorrelated rates** (Thorne, Kishino & Painter 1998; Kishino, Thorne & Bruno 2001; Thorne & Kishino 2002)
- **Uncorrelated/independent rates models** (Drummond et al. 2006; Rannala & Yang 2007; Lepage et al. 2007)
- **Mixture models on branch rates** (Heath, Holder, Huelsenbeck 2012)

Global Clock

The substitution rate is constant over time

All lineages share the same rate

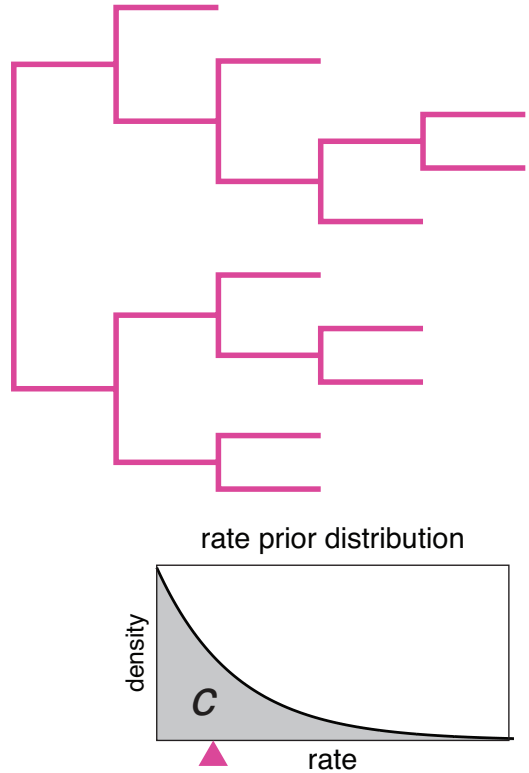
(Zuckerkandl & Pauling, 1962)



..

Global Clock

The sampled rate is applied to every branch in the tree



Relaxed-Clock Models

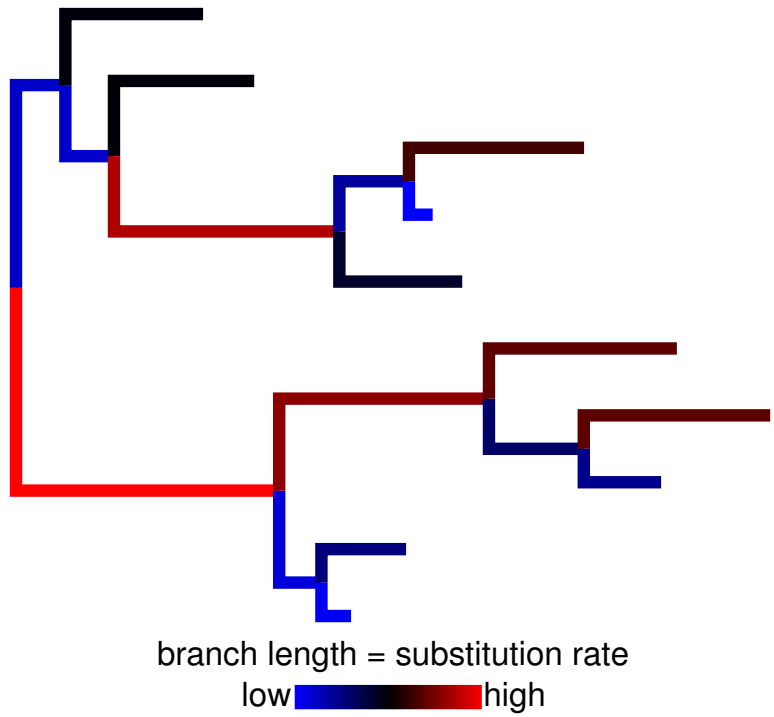
To accommodate variation in substitution rates 'relaxed-clock' models estimate lineage-specific substitution rates

- **Local clocks**
- **Punctuated rate change model**
- **Log-normally distributed autocorrelated rates**
- **Uncorrelated/independent rates models**
- **Mixture models on branch rates**

Independent/Uncorrelated Rates

Lineage-specific rates are uncorrelated when the rate assigned to each branch is independently drawn from an underlying distribution

(Drummond et al. 2006; Rannala & Yang 2007; Lepage et al. 2007)



..

Modeling Rate Variation

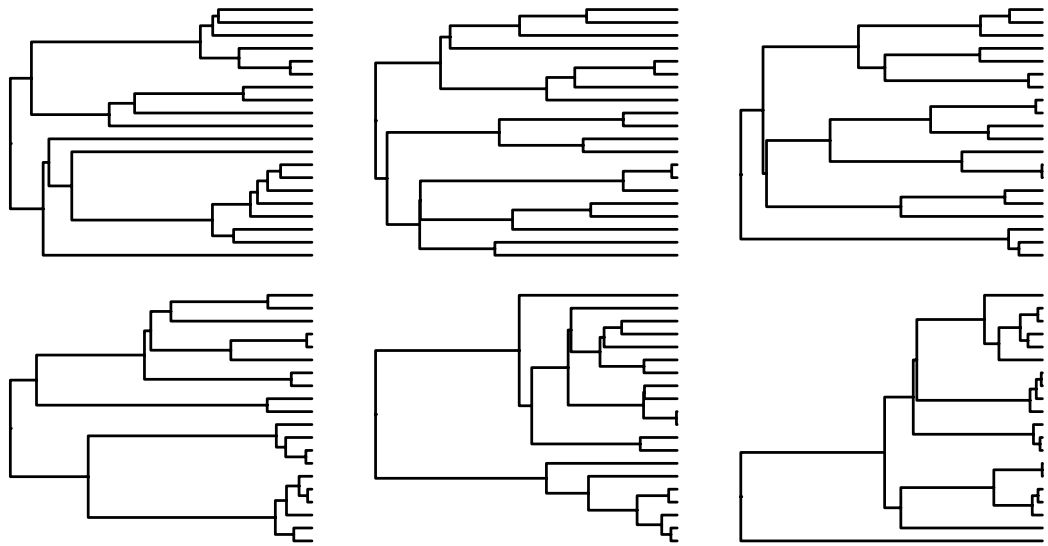
These are only a subset of the available models for branch-rate variation

- **Global clock**
- **Local clocks**
- **Punctuated rate change model**
- **Log-normally distributed autocorrelated rates**
- **Uncorrelated/independent rates models**
- **Dirchlet process prior**

Considering model selection, uncertainty, & plausibility is **very** important for Bayesian divergence time analysis

Priors on the Tree and Node Ages

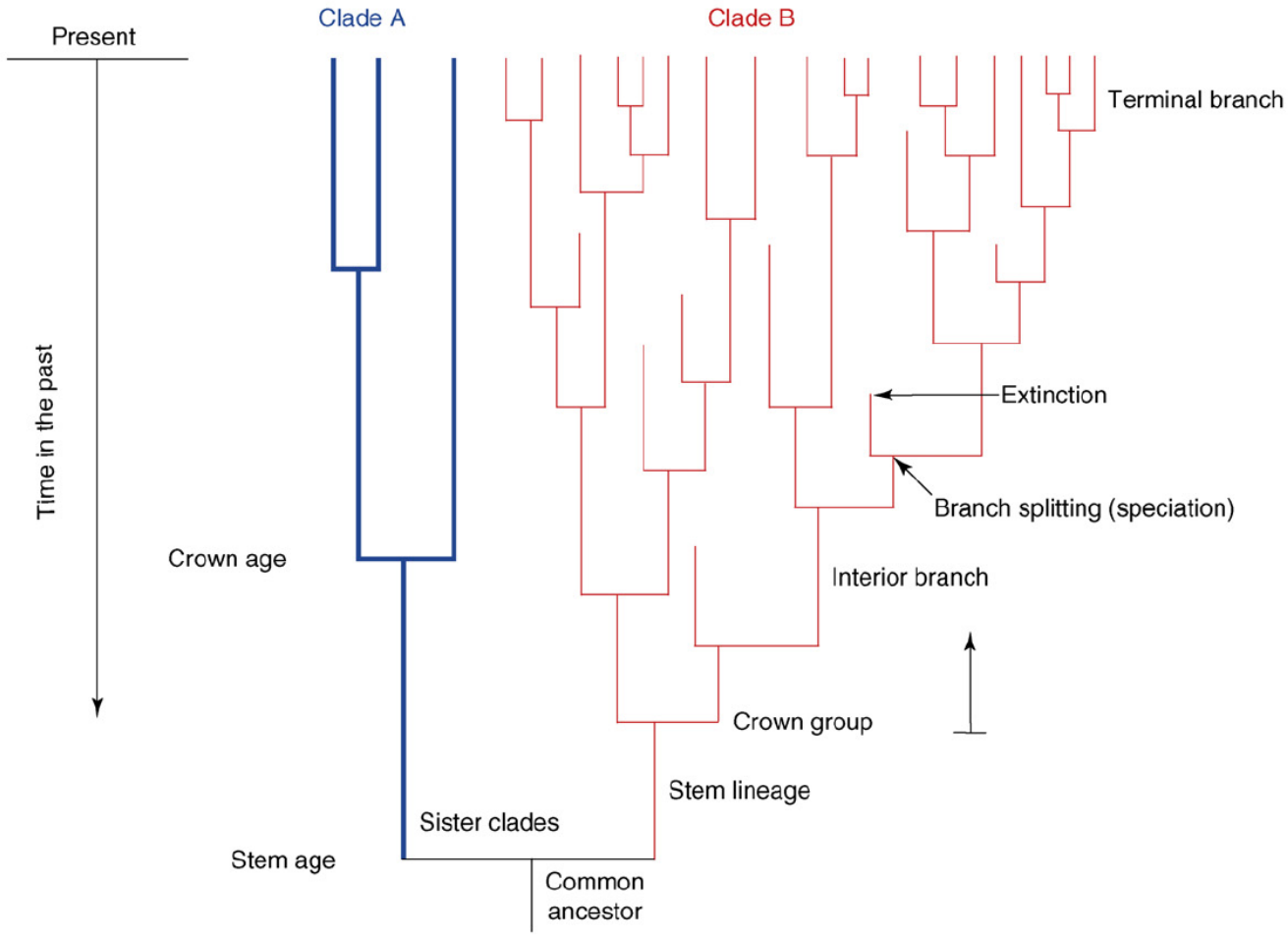
Relaxed clock Bayesian analyses require a prior distribution on time trees



Different node-age priors make different assumptions about the timing of divergence events

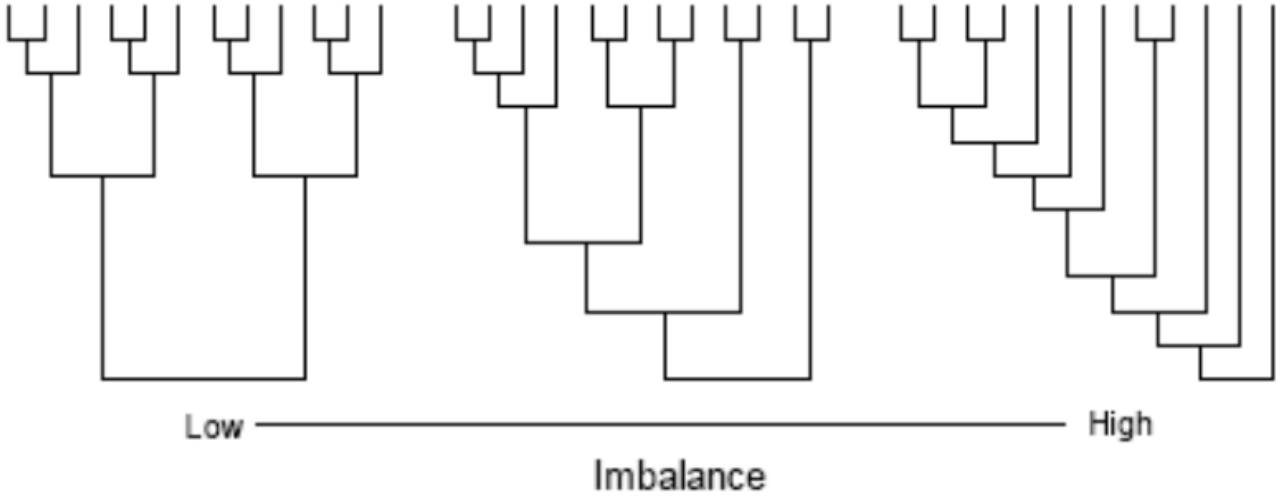
Models for tree shapes:

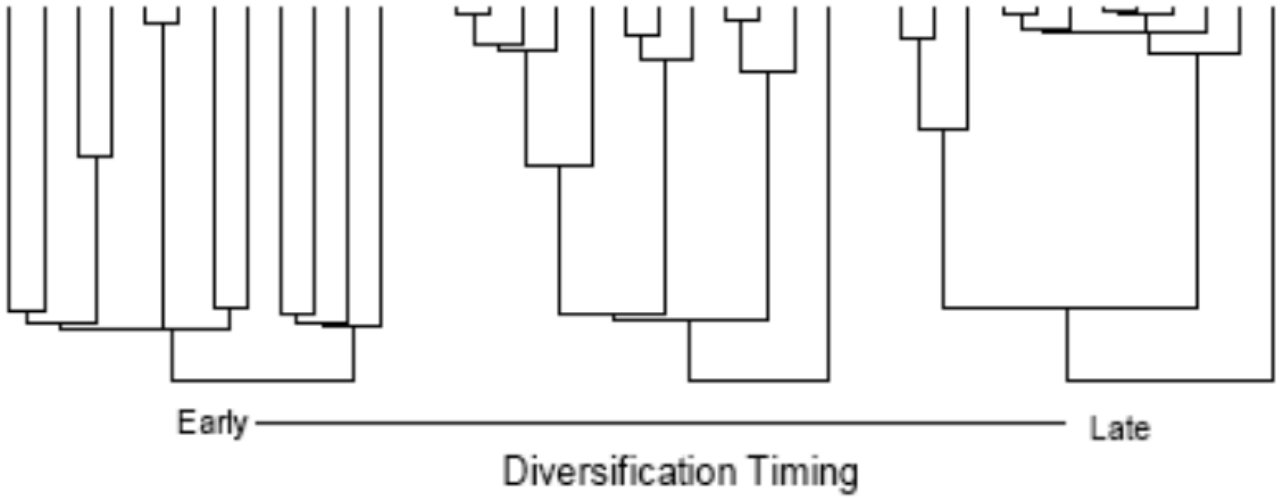
- Yule or 'Pure Birth' model: lineages have a constant probability of speciation.λ.
- Birth-Death: Lineages have a constant probabilities λ of speciation, or μ extinction.
- Diversification rate: ρ, equals the $\lambda - \mu$
- Relative extinction rate or 'turnover rate': μ/λ



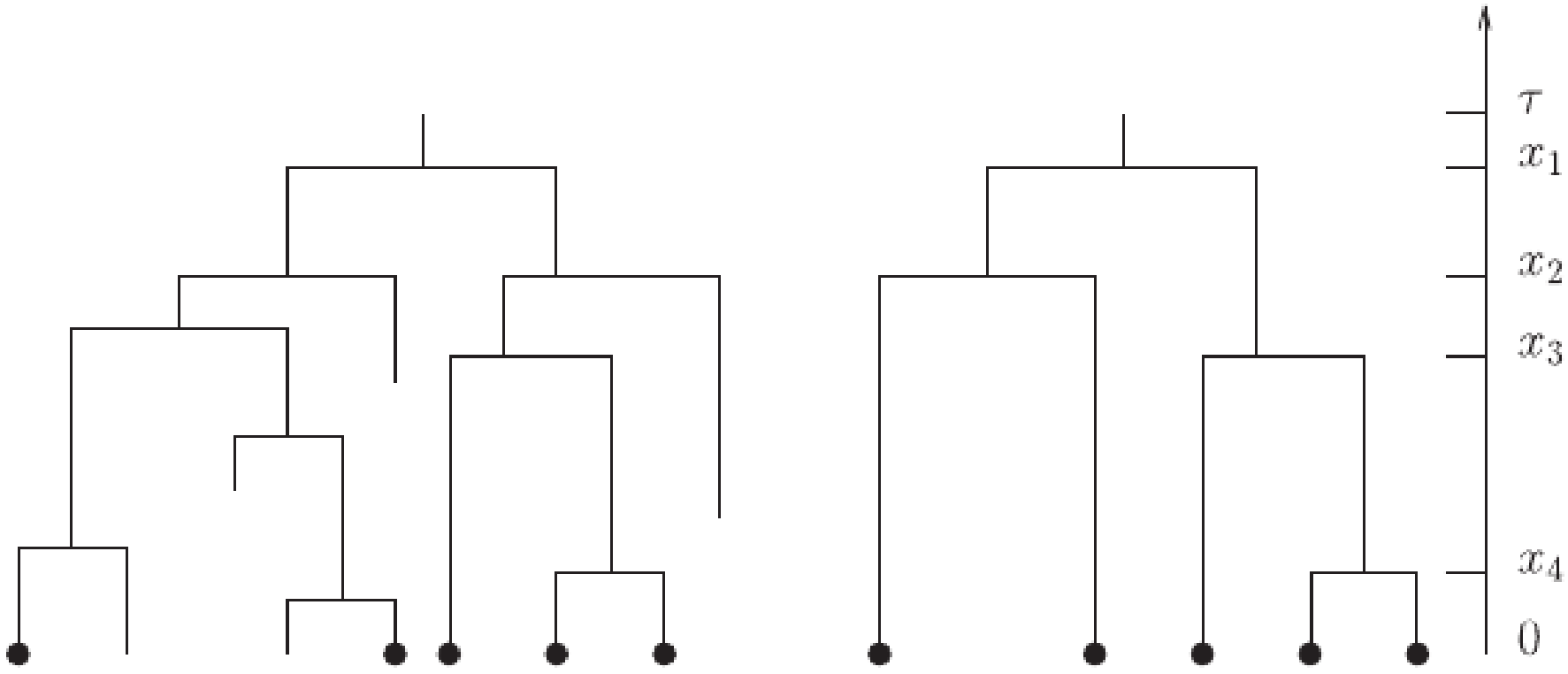
(Ricklefs, TREE, 2007)

Tree shapes are the same for Yule and Birth Death processes, but waiting times are not.



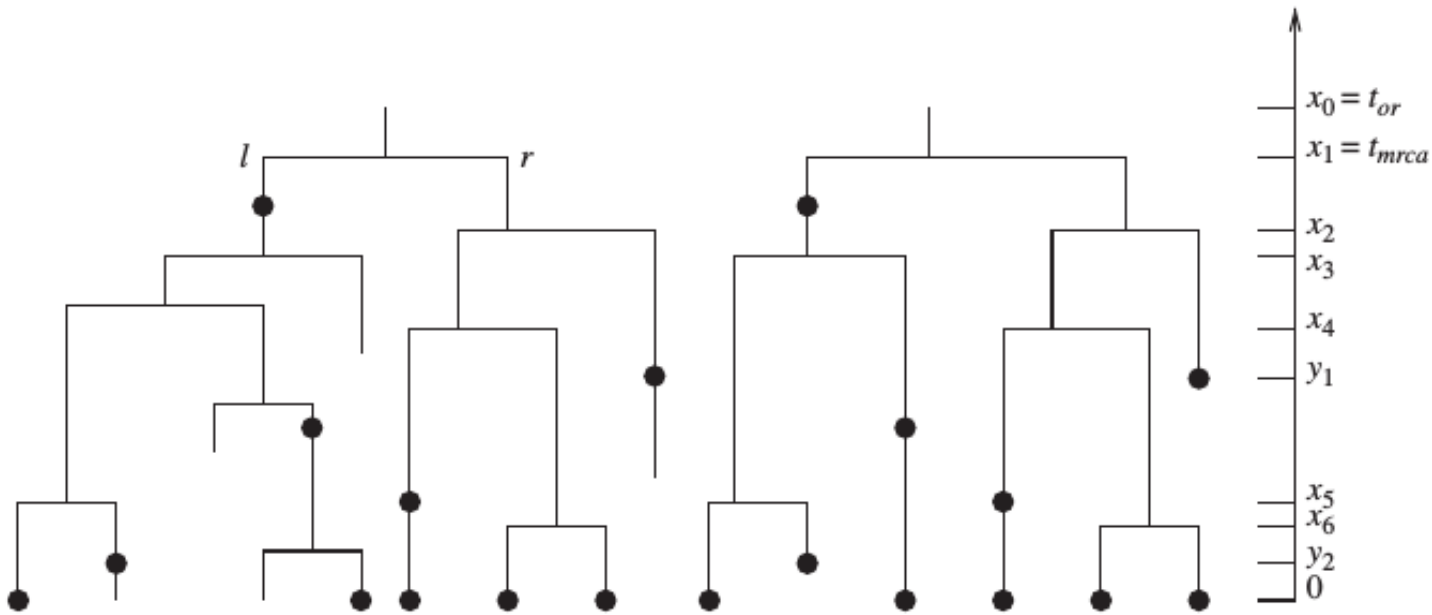


Sampled birth-death process tree prior



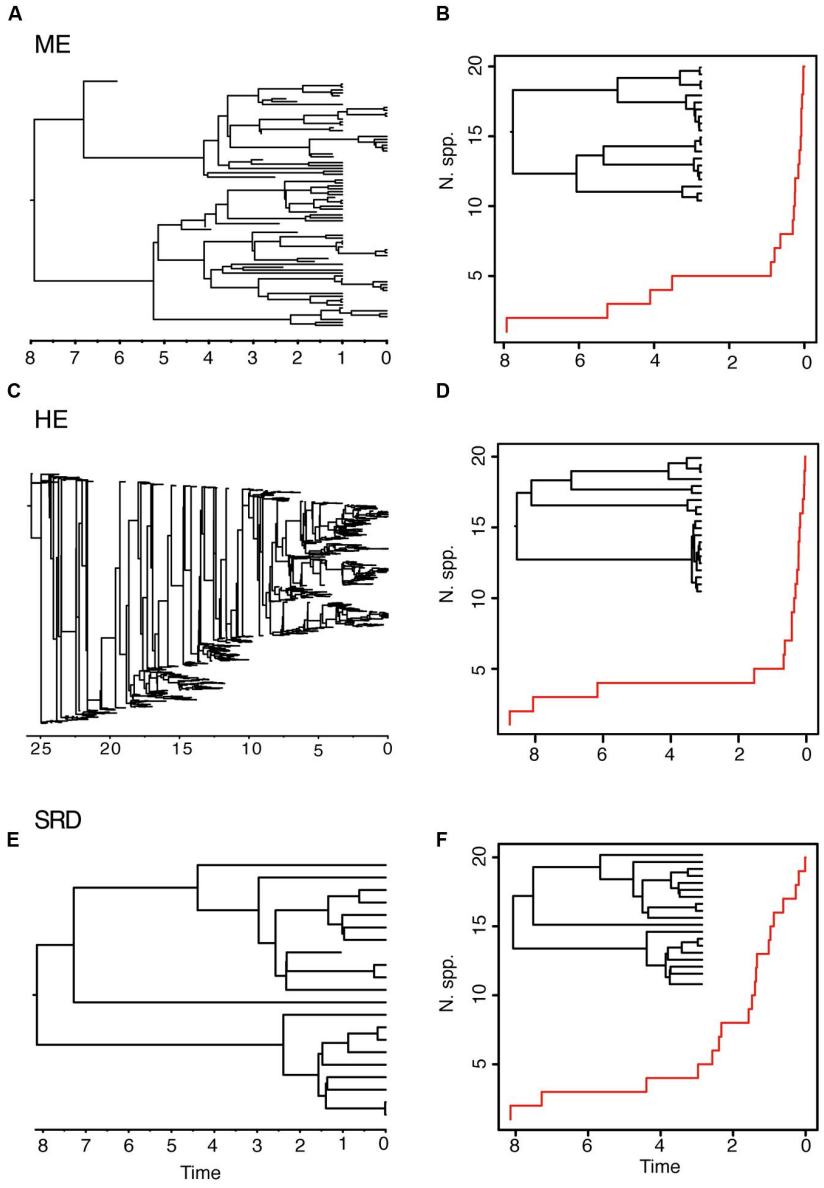
Stadler, J. Theoretical Biology. 2009

Sampling through time in birth-death trees



Stadler (2010) J. Theoretical Biology

Different birth–death models can generate reconstructed phylogenies with similar realized tree shapes.



(Sanmartin and Meseguer, Frontiers in Genetics, 2016)

‘Pull of the present’: diversification appears to increase close to the present because lineages that arose more recently are less likely to have gone extinct. (Nee et al. PNAS, 1993)

Diversification rates often appear to decrease (Etitenne and Rosindel, Syst Bio, 2012)

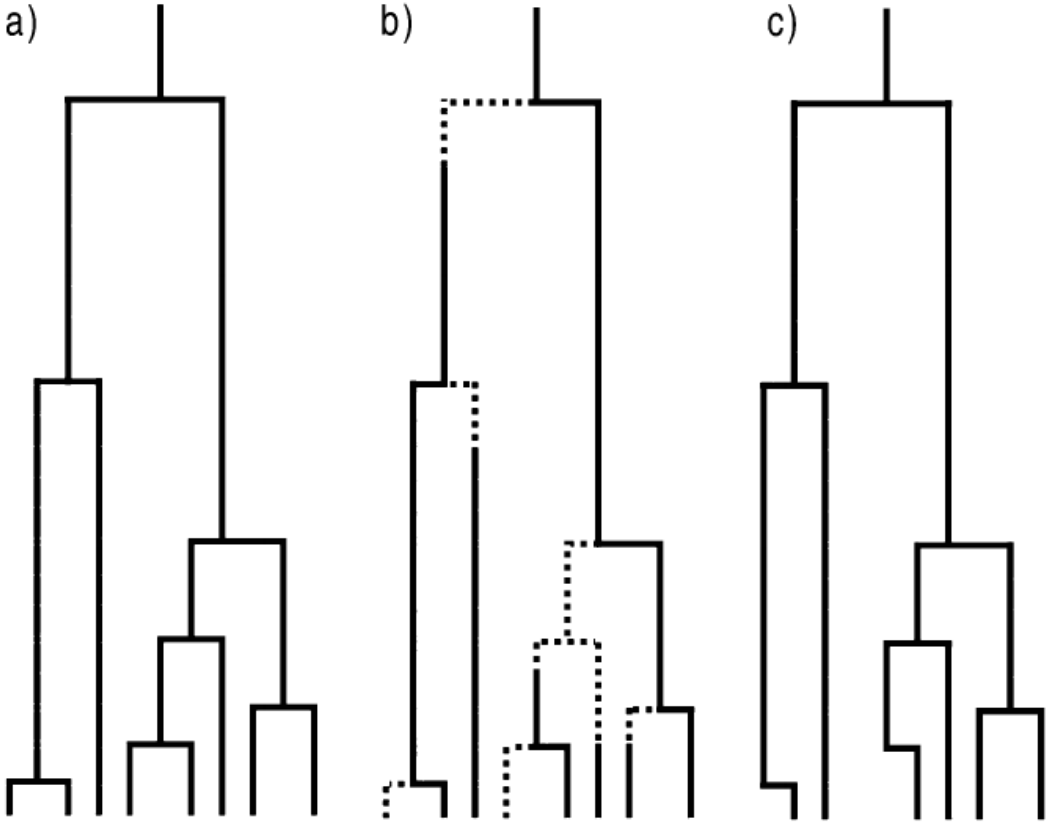
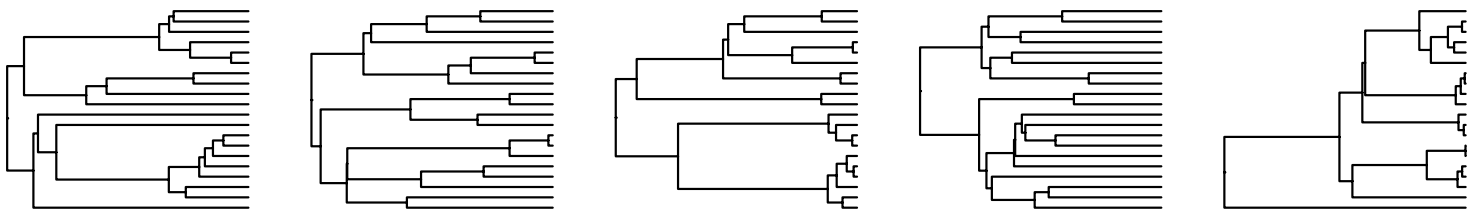


FIGURE 1. The pure birth model a) with and b) without protracted speciation. Dotted lines indicate an incipient species and solid lines are good species. c) Phylogeny of the protracted pure birth process of panel b: only those lineages that have completed speciation before the present will show up in the phylogeny. Note that the branching points are at the times that the incipient species are produced, not at the times that they become good species.

Priors on Node Times

Sequence data are only informative on *relative* rates & times

Node-time priors cannot give precise estimates of *absolute* node ages



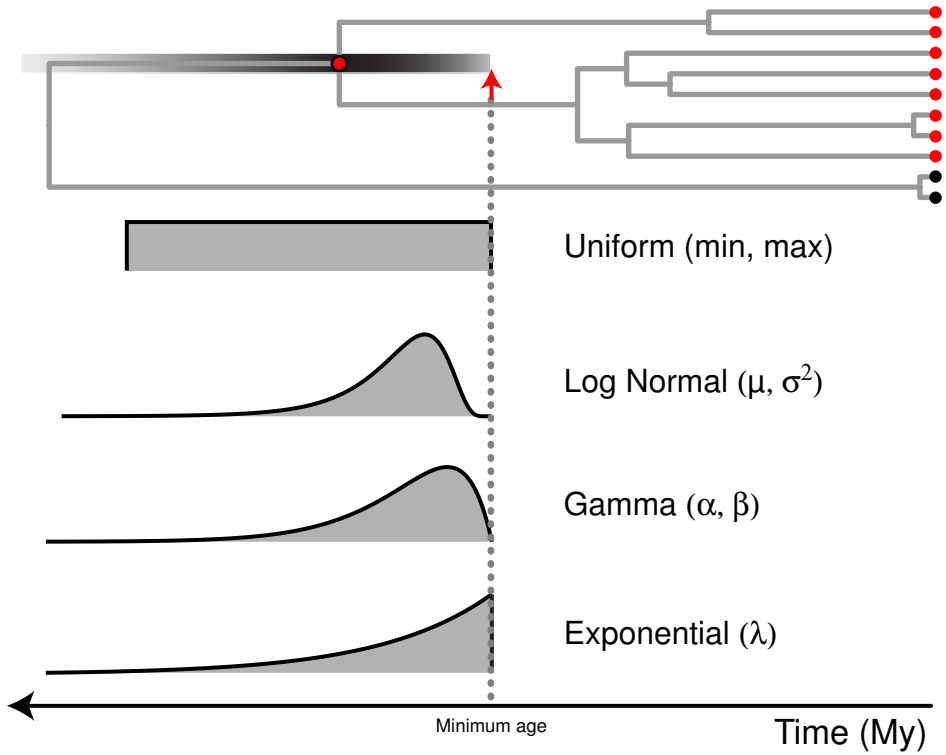
We need additional data (like fossils) to provide absolute time scale

Prior Densities on Calibrated Nodes

Common practice in Bayesian divergence-time estimation:

Parametric distributions are typically off-set by the age of the oldest fossil assigned to a clade

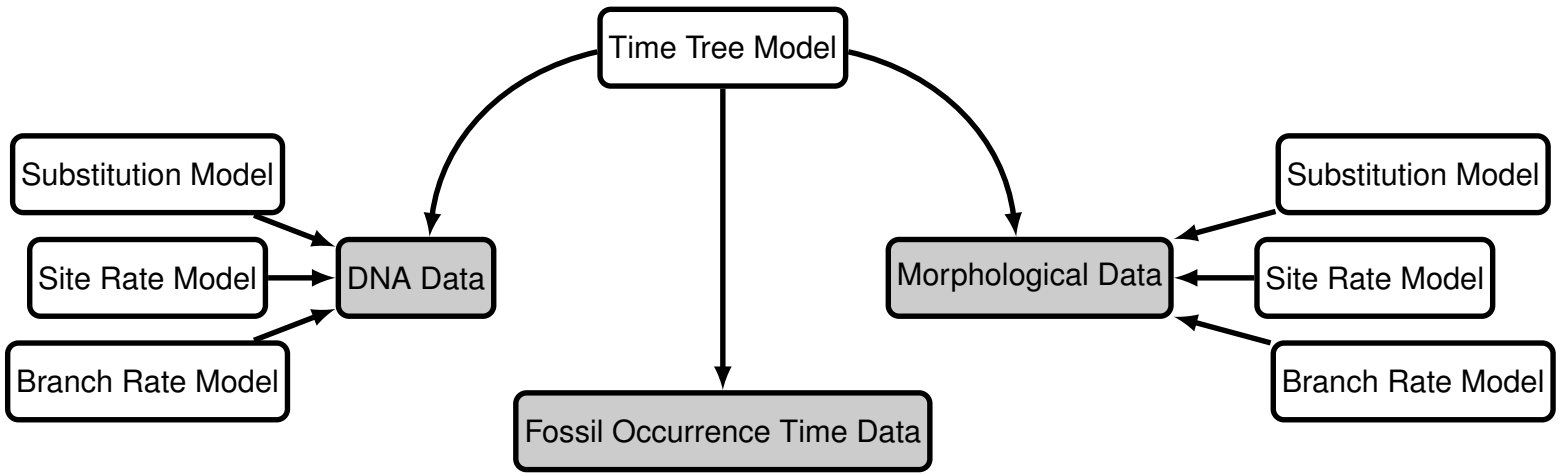
These prior densities do not (necessarily) require specification of maximum bounds



For a more nuanced approach integrating uncertainty, see optional tutorial on Fossilized-Birth-Death process
`https://taming-the-beast.org/tutorials/FBD-tutorial/`

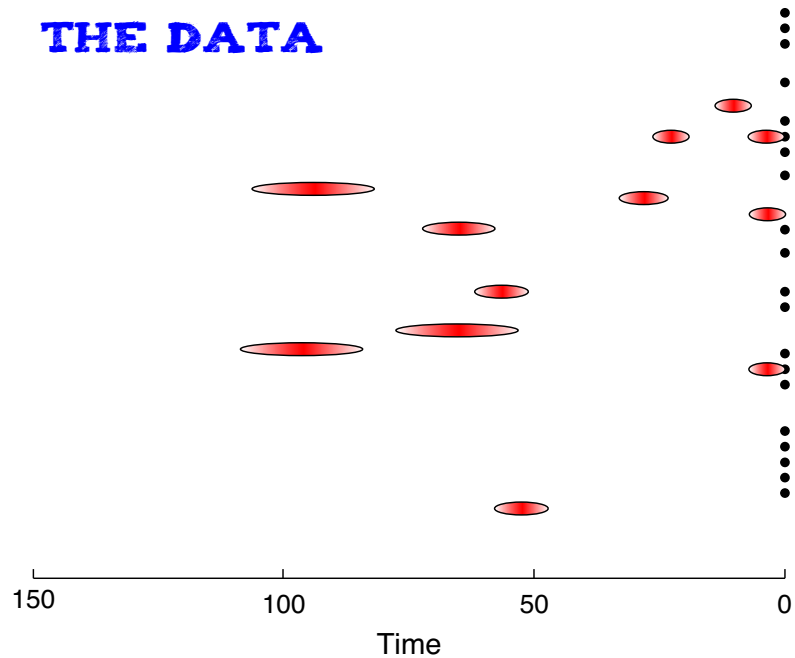
"Total-Evidence" Analysis

Integrating models of molecular and morphological evolution with improved tree priors enables joint inference of the tree topology (extant & extinct) and divergence times



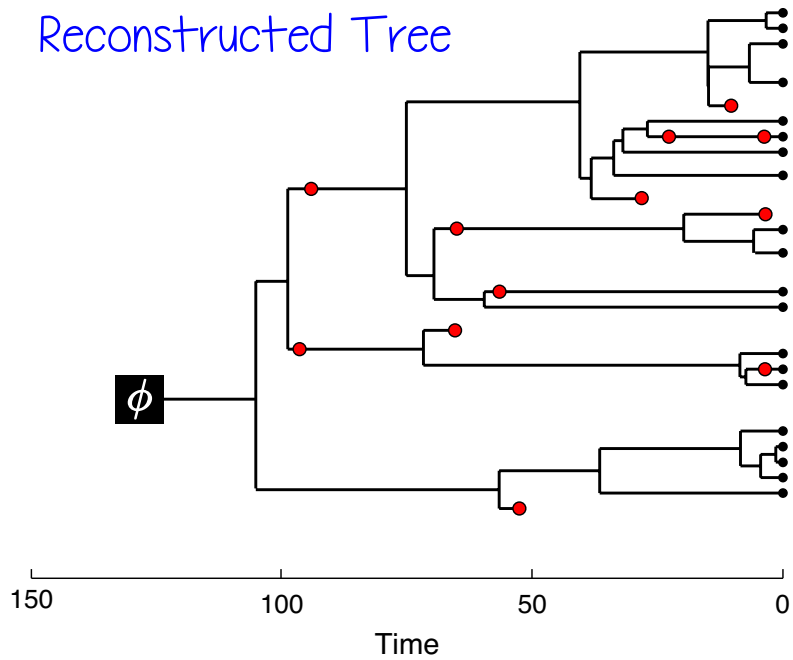
"Total-Evidence" Analysis

Allows us to estimate the reconstructed tree of our sampled fossils and extant taxa along with the diversification dynamics and rates of molecular and morphological evolution



"Total-Evidence" Analysis

Allows us to estimate the reconstructed tree of our sampled fossils and extant taxa along with the diversification dynamics and rates of molecular and morphological evolution



Conclusions:

- Few trees fit an exact molecular clock
- Time estimates on trees require external dates, either fossils or samples through time
- Uncertainty in estimates is often high.
- Fossil calibration choices can have major effects on inferences.