Tree Priors (Tree Models)

Fred Ronquist
Swedish Museum of Natural History,
Stockholm, Sweden

Workshop in Advanced Bayesian Phylogenetics
Adelaide, Nov. 17-21, 2014

Topics

- Some model terminology
- Models for branch-length trees
- Models for time trees

1. Some Model Terminology

Some Terminology

Prior probability	Probability of all unclamped stochastic non-sink nodes in the model
Likelihood	Probability of all clamped stochastic sink nodes in the model
Posterior prob.	Probability of all stochastic nodes in the model (clamped and unclamped)
Hyperprior	Second-level prior in a hierarchical model. Hyper-hyperprior is third level, etc.

These distinctions are partly arbitrary. Better to just talk about the probability of the model = the probability of all stochastic nodes (unclamped or clamped) in the model.

More Terminology

Model	The structure of the entire model or any part of the model. Could also implicitly include the probability distributions.
Prior	Either the structure of upstream parts of the model and the probability distributions associated with them OR just the probability distributions associated with the upstream parts of the model.

In other words, model and prior can sometimes mean the same thing. For instance, tree prior and tree model would typically refer to the same thing.

Even More Terminology

Topology	The topological structure of a tree.
Tree	The topology and the branch lengths. Sometimes used to refer only to the topology.
Branch	A taxon bipartition in the tree. Also known as a clade, an edge or a split in the tree.
Node	A node in the tree. Also known as a vertex, a divergence event, a dichotomy, or a bifurcation in the tree.

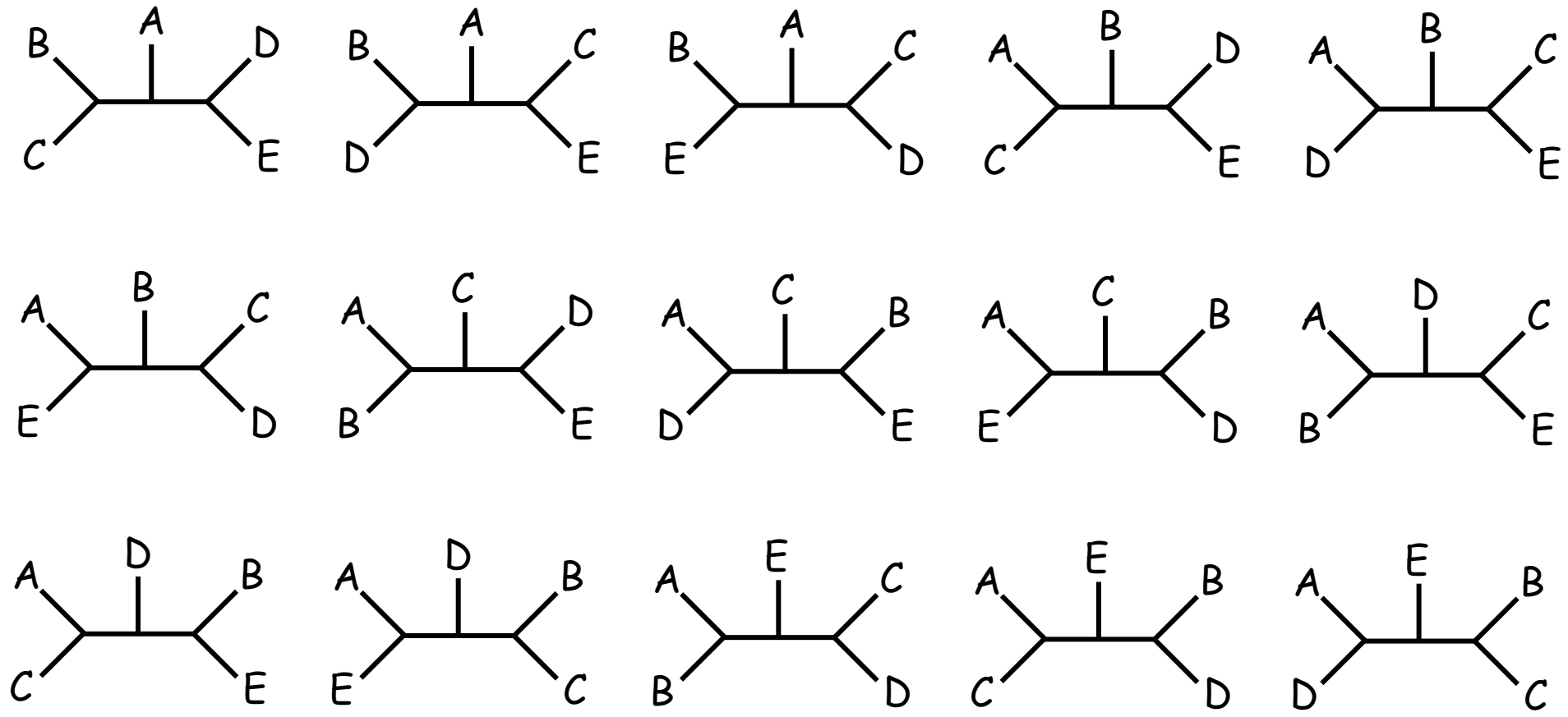
Note that tree and topology sometimes mean the same thing, and sometimes not.

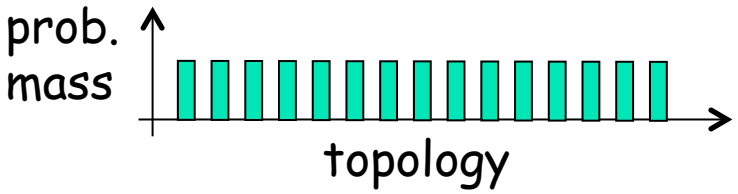
Yet More Terminology

Branch-length tree	A tree where branches are measured in units of evolutionary change.
Time tree	A tree where branches are measured in units of relative or absolute time.
Clock model	A model for converting time trees into branch-length trees

2. Models for Branch-Length Trees

Uniform prior on topologies





Combine with choice of branch length prior

Exponential prior

Exponential prior with rate drawn from a hyperprior

Terminal and internal branches have different priors

Dirichlet prior for branch length proportions combined with separate prior for total tree length

Separate branch length for every character ("parsimony model")

… and topology constraints

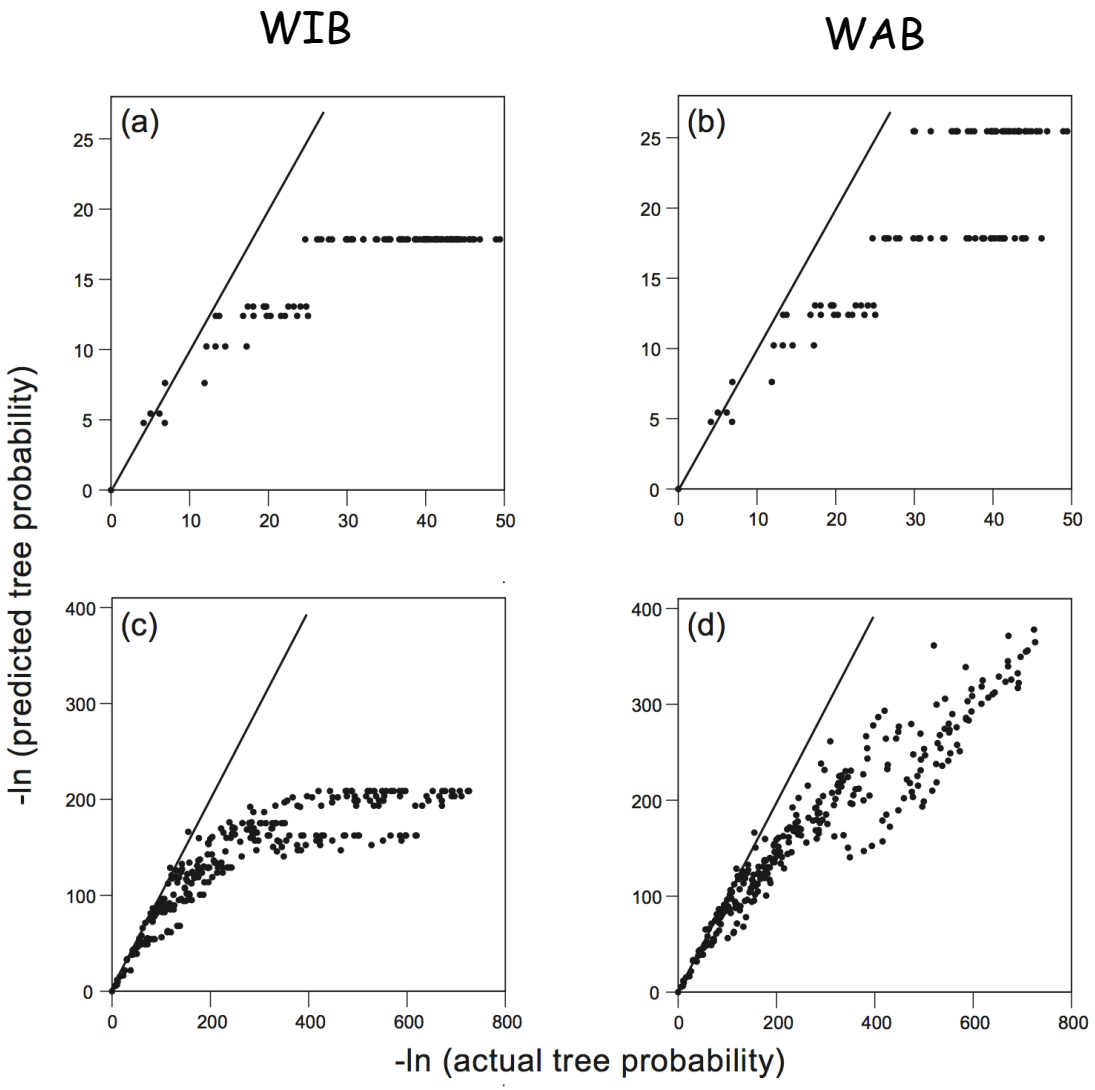
Hard constraint: Only put probability on trees satisfying certain conditions, e.g., clade A being monophyletic.

Soft constraint: Weight probability with factors depending on how well they satisfy constraints, e.g., clade A being monophyletic.

Backbone constraint: Only include certain taxa in the constraint, have other taxa float around in the tree.

Negative constraint: Only put probability on trees **not** satisfying the constraint. For instance, put all probability on trees **not having** clade A.

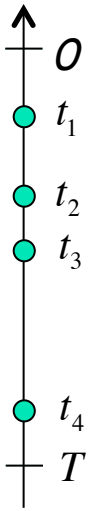
Soft constraints to represent tree space priors



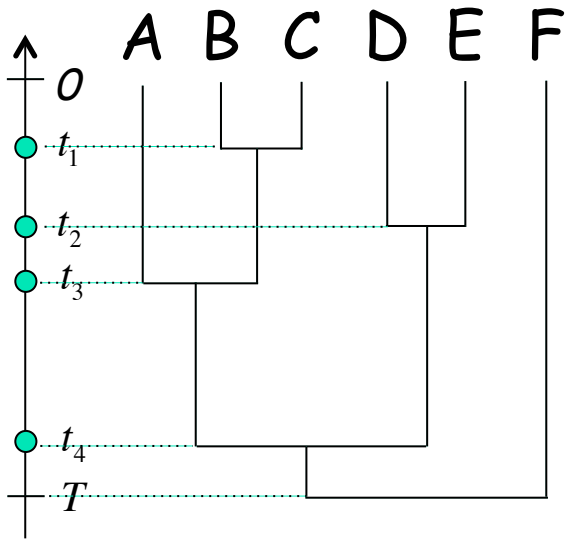
Ronquist et al. (2004). In Bininda-Emonds, Supertrees

3. Models for Time Trees

Uniform time tree prior

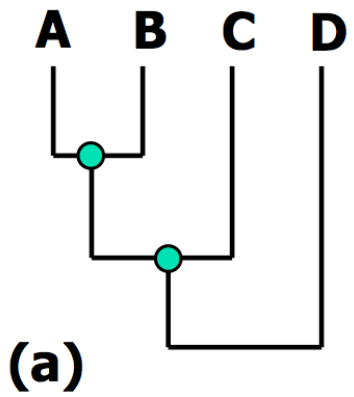


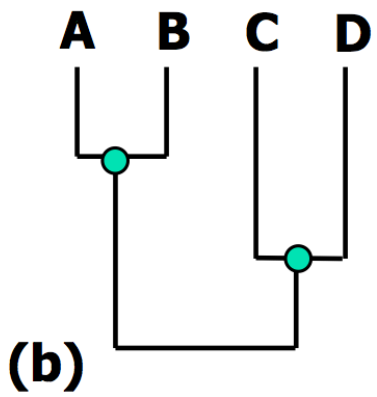
1. Speciation times are drawn uniformly at random in the interval (0,T), and then ordered.



2. Species are clustered randomly, using the speciation times.

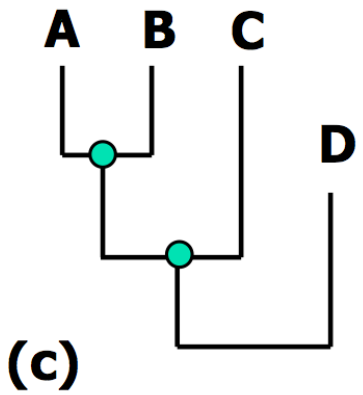
Properties of the uniform tree prior

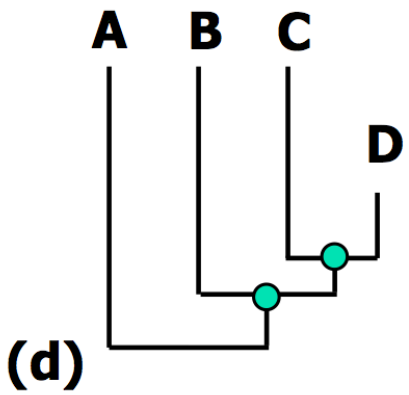




Probability is proportional to labeled histories and not to topologies: order of speciation events counts.

Probability of topology (b) is twice the probability of topology (a)

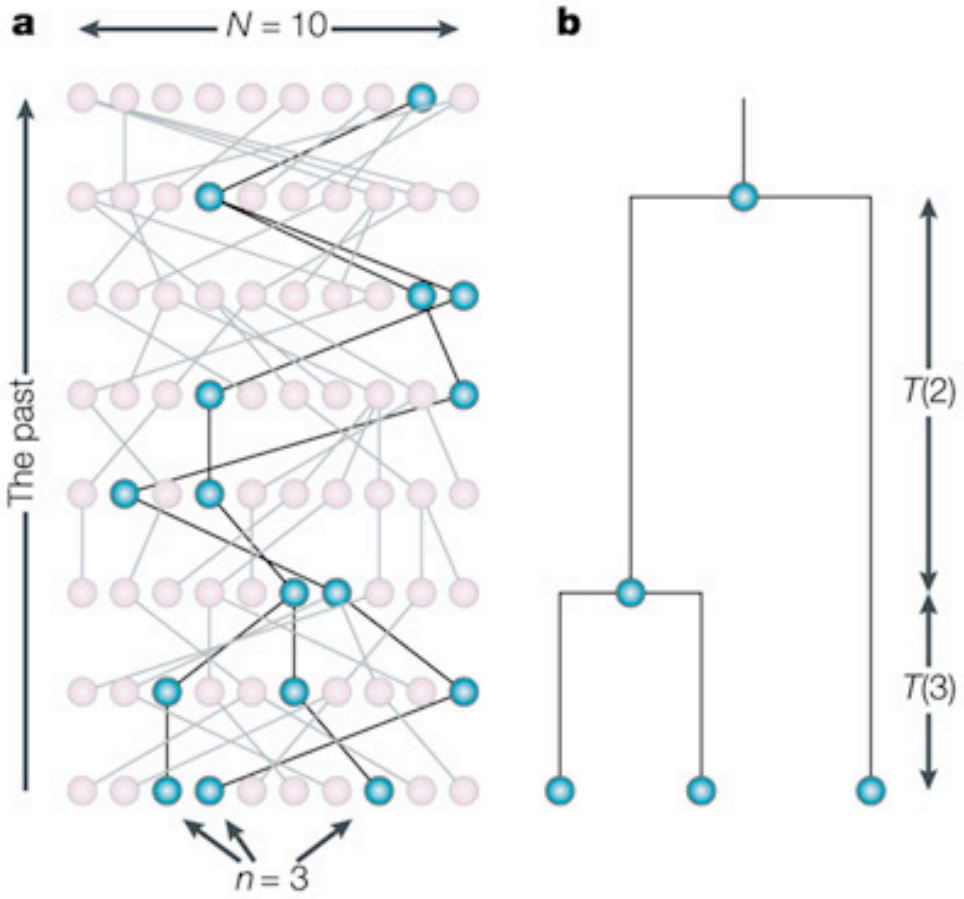




Probability is proportional to flexibility in assigning node dates.

Probability of topology (c) is much higher than probability of topology (d).

The coalescent model



The coalescent model

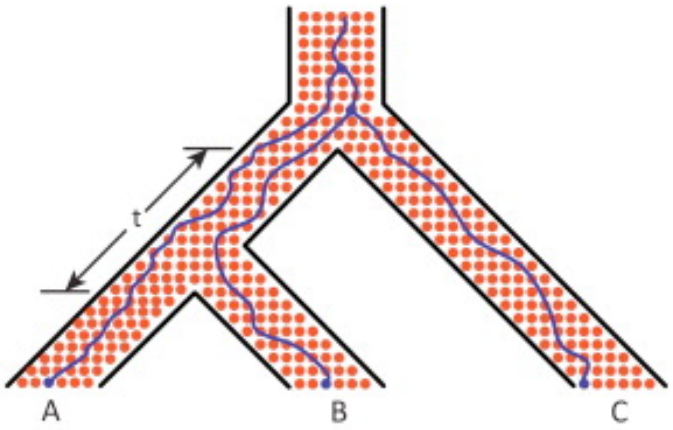
At each successive preceding generation, the probability of coalescence is geometrically distributed — that is, it is the probability of *non*coalescence at the $t-1$ preceding generations multiplied by the probability of coalescence at the generation of interest:

$$P_c(t) = \left(1 - \frac{1}{2N_e}\right)^{t-1} \left(\frac{1}{2N_e}\right).$$

For sufficiently large values of N_e, this distribution is well approximated by the continuously defined exponential distribution

$$P_c(t) = \frac{1}{2N_e} e^{-\frac{t-1}{2N_e}}.$$

The multispecies coalescent



Coalescent process within lineages, combined with speciation

Birth-death model in phylogenetics

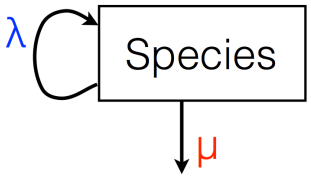
Parameters

λ Speciation rate

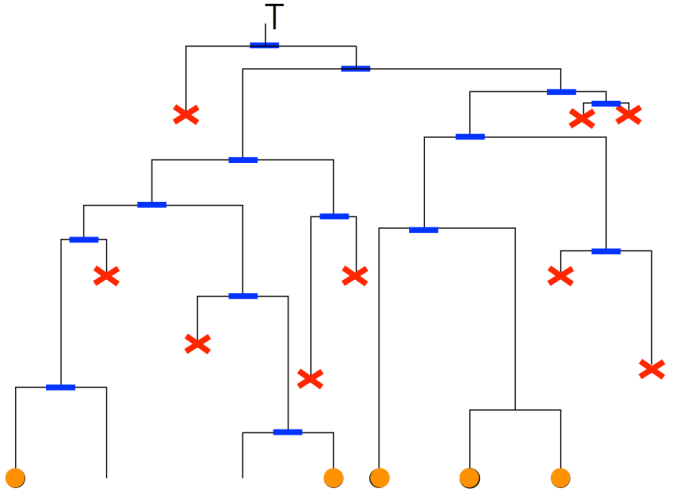
μ Extinction rate

ρ Sampling probability

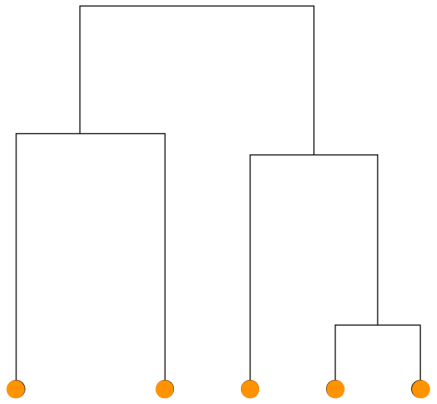
T Time of origin



State machine representation



Complete tree

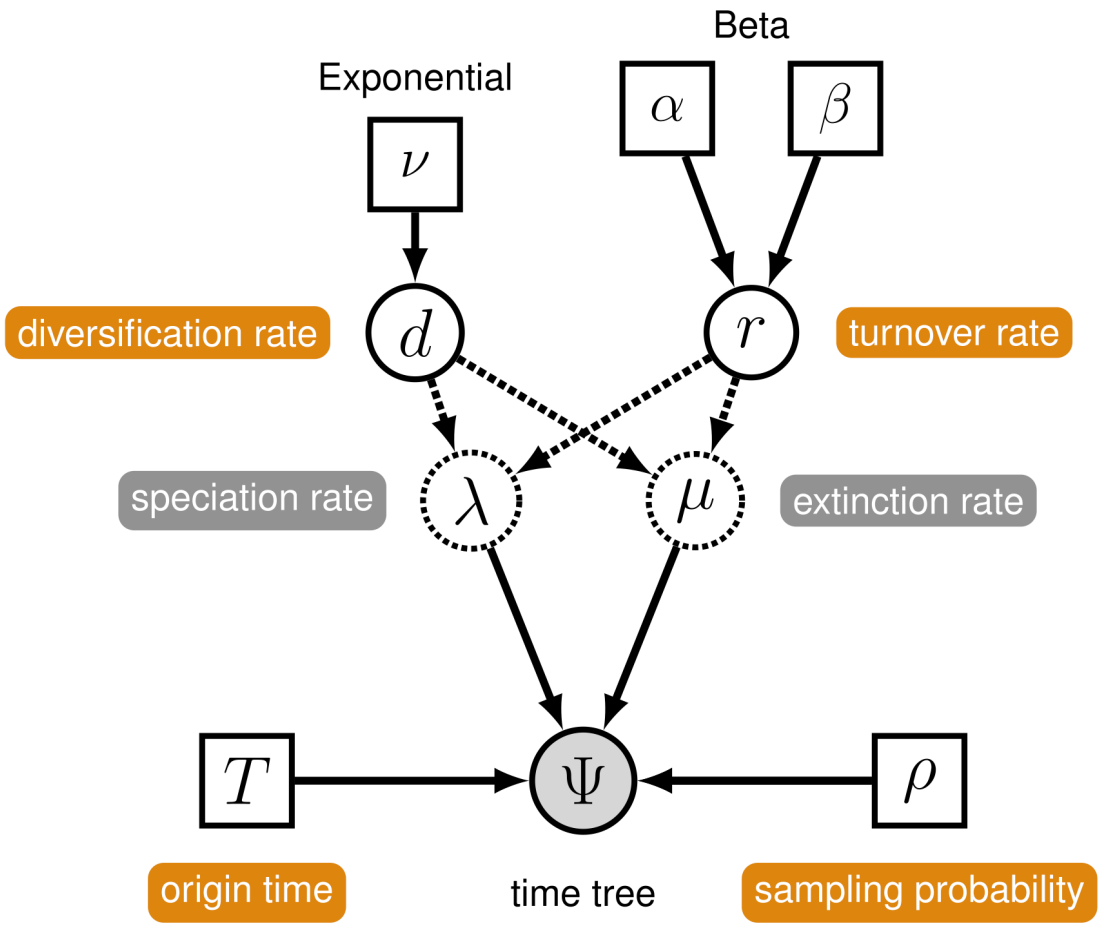


Sampled tree

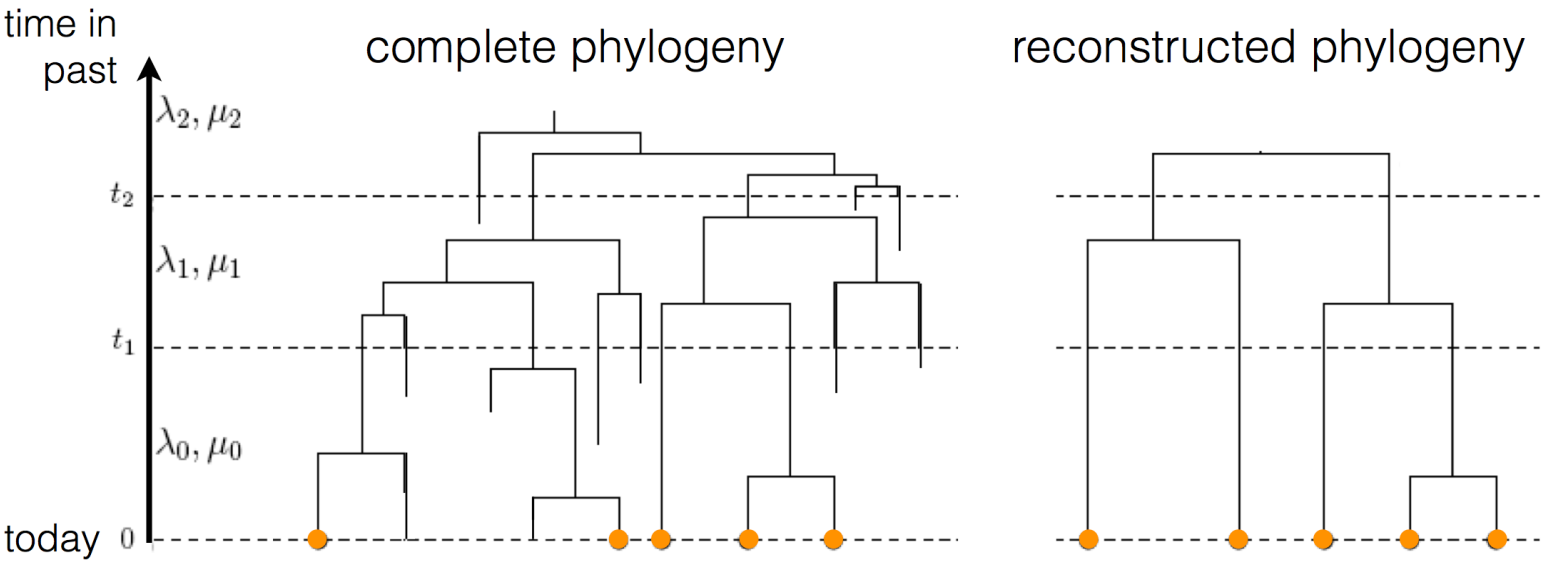
Generalized birth-death model graph

$$d = \lambda - \mu$$

$$r = \mu / \lambda$$



The piece-wise constant birth-death model



Probability of the reconstructed tree is an integral over all complete trees. It can be calculated efficiently using recursion and by solving differential equations.

Stadler 2011, PNAS

The fossilized birth-death (FBD) model

Parameters

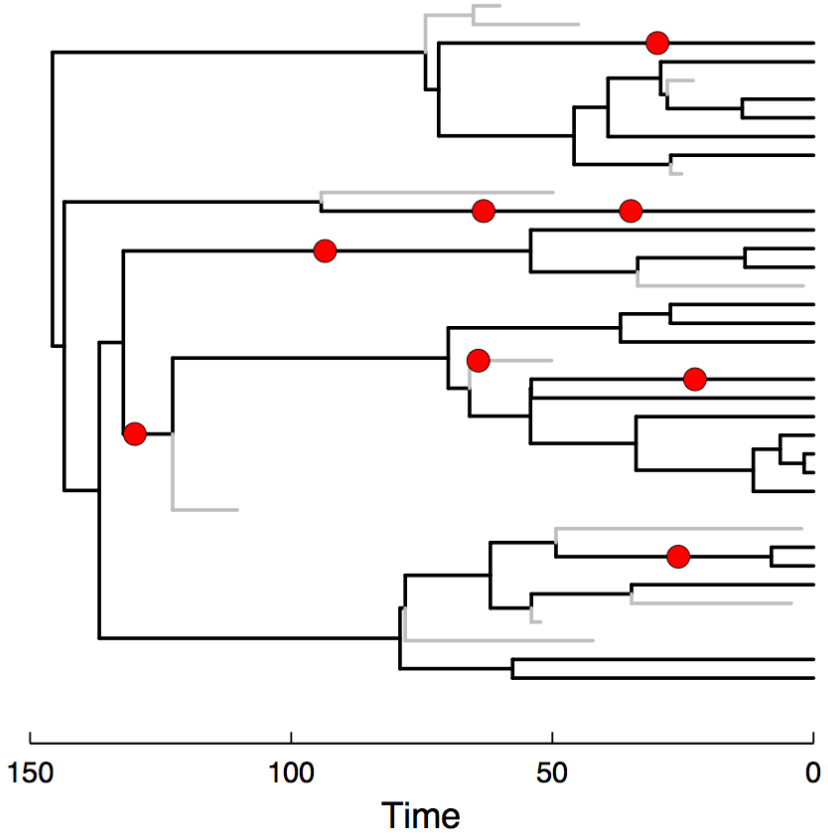
λ Speciation rate

μ Extinction rate

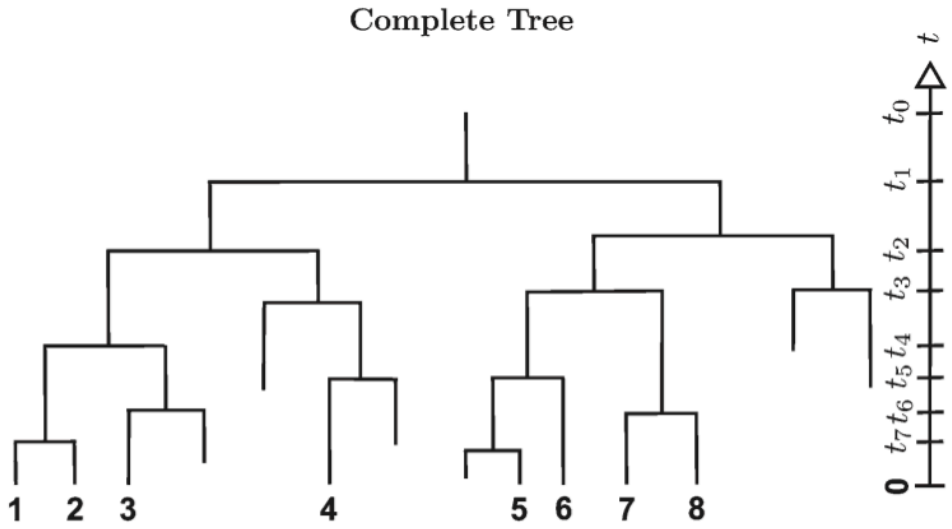
ψ Fossilization rate

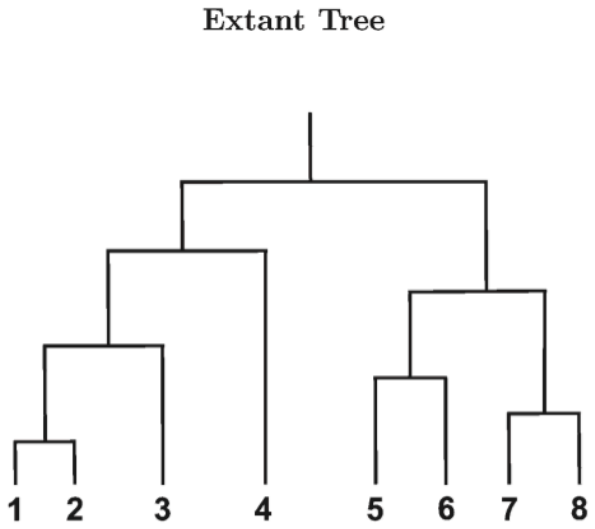
ρ Sampling probability

T Time of origin

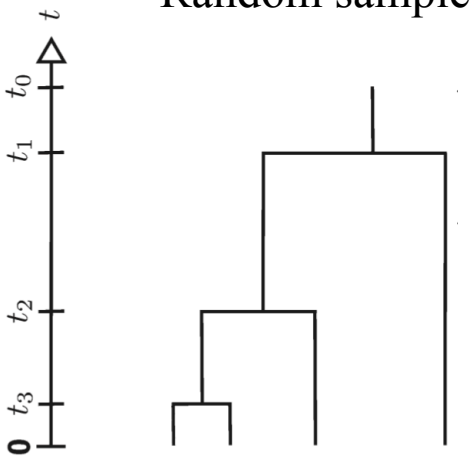


Sampling of extant taxa

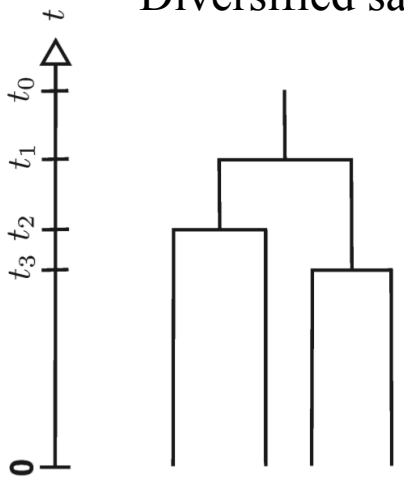




Random sample

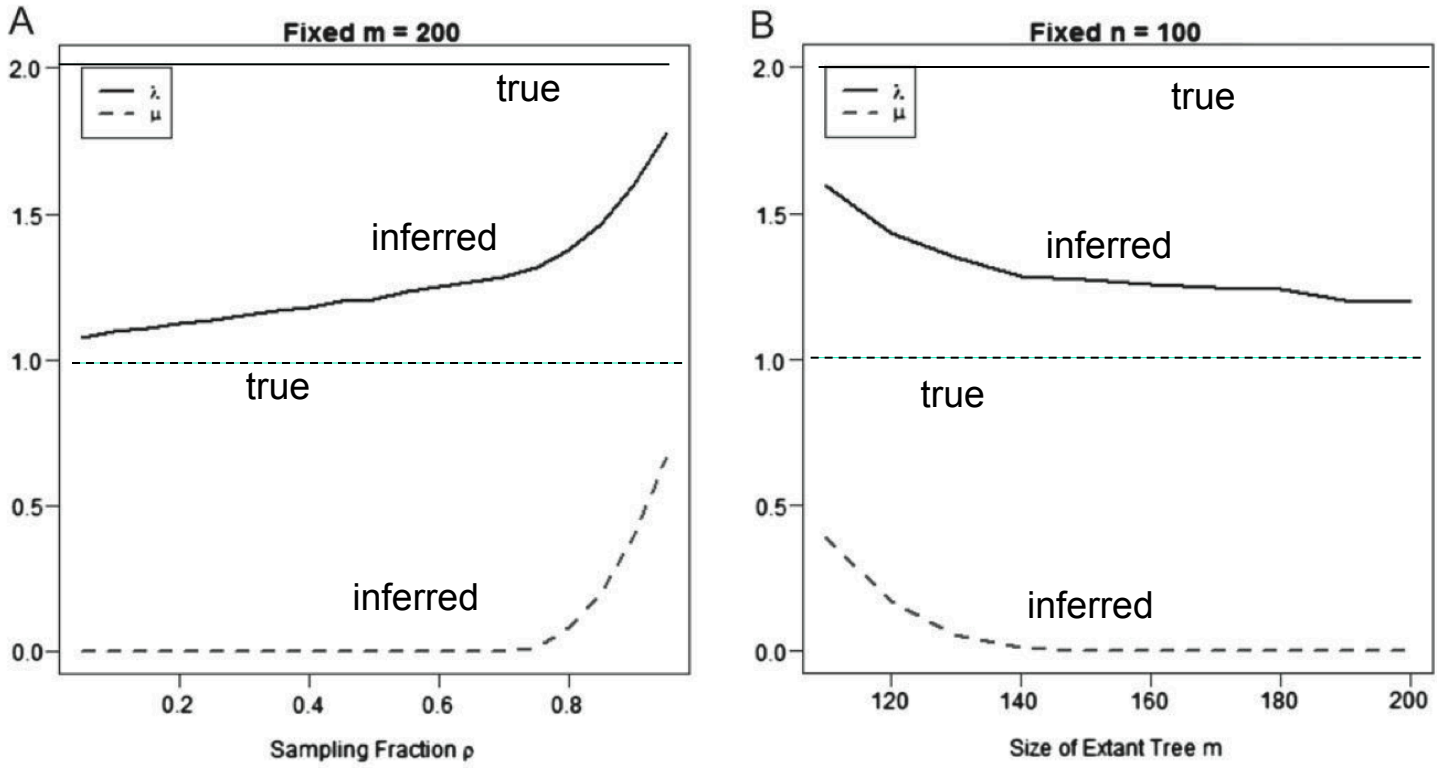


Diversified sample



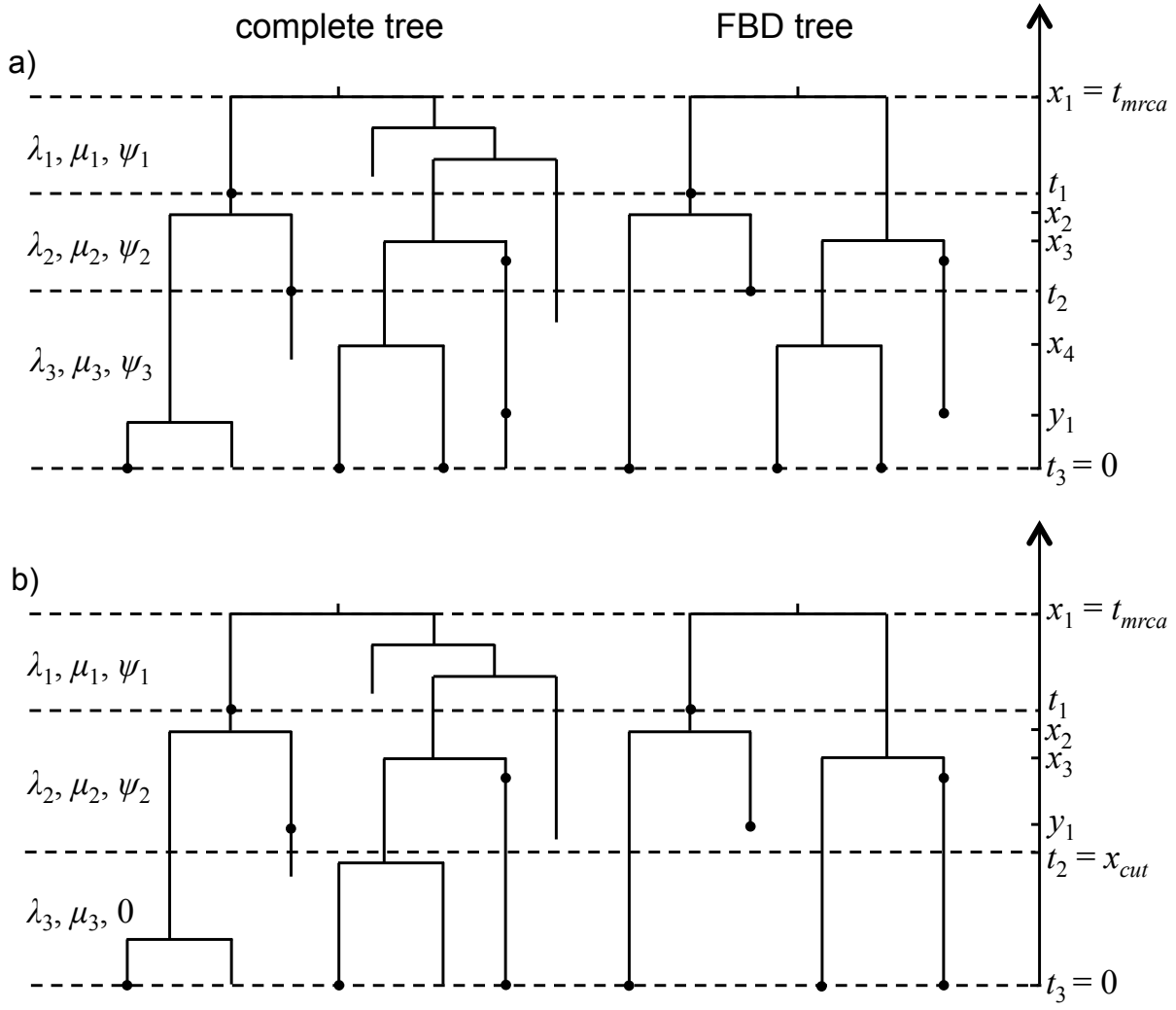
Sampling affects inferred speciation and extinction rates

Looking at diversified sample, assuming random sample

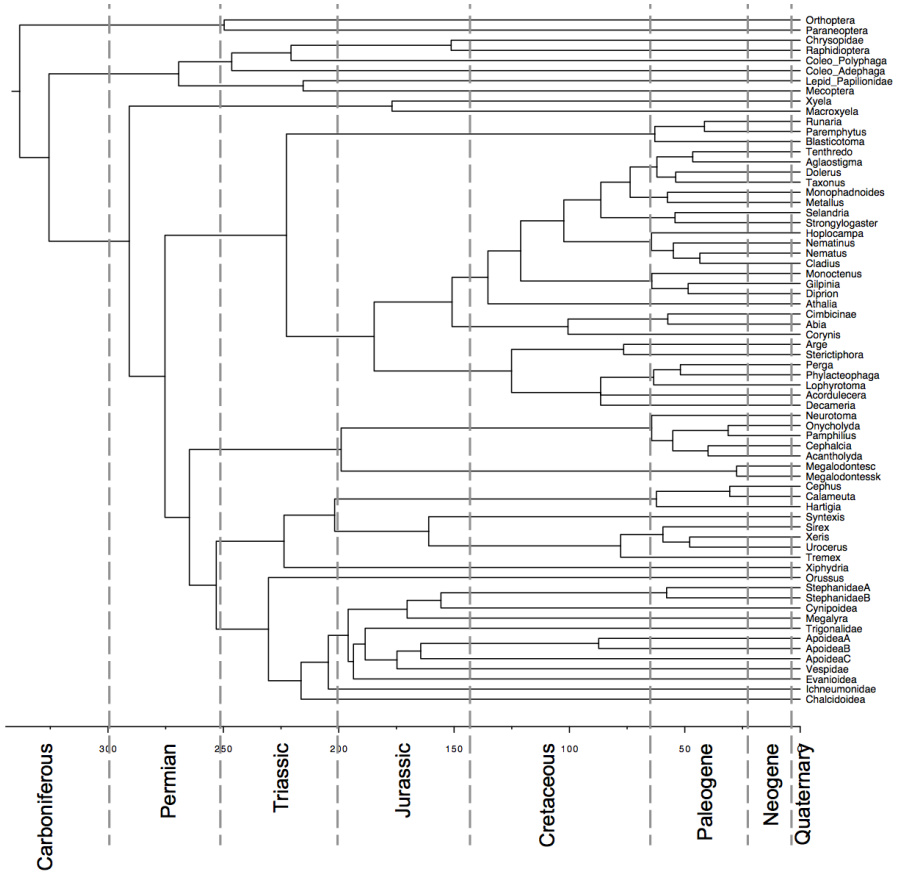


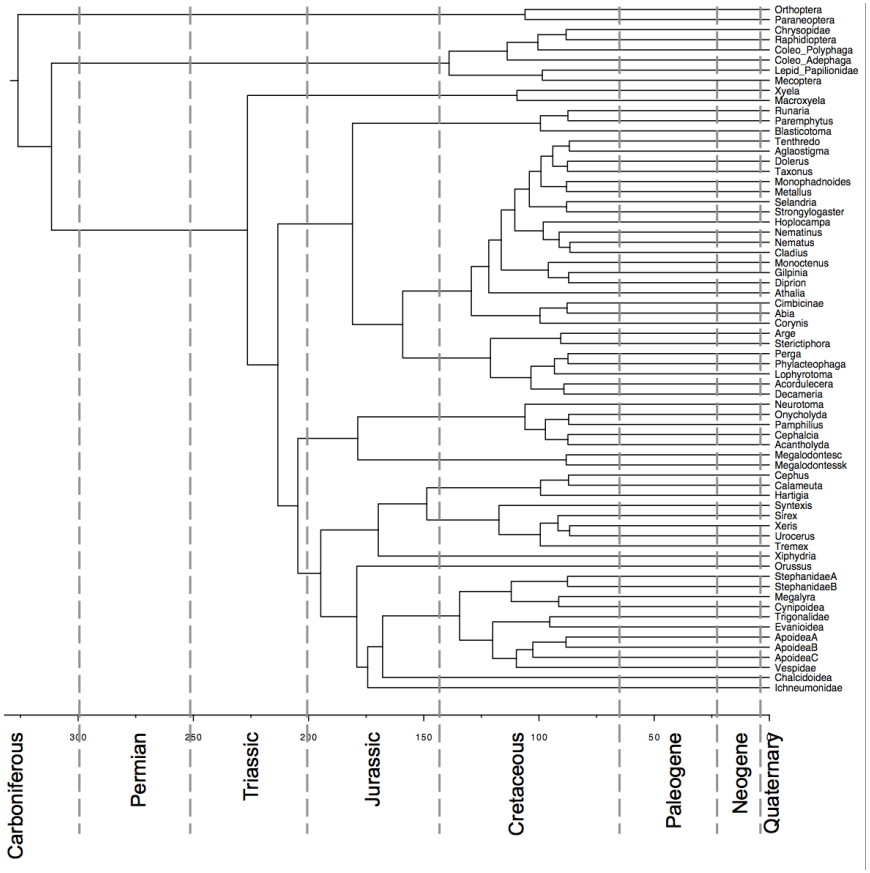
Summary: Both speciation and extinction rates are underestimated, especially extinction rates if we assume random sampling when sampling is diversified.

Piece-wise constant FBD model with slice sampling



FBD total-evidence dating assuming random sampling (left) or diversified sampling (right)

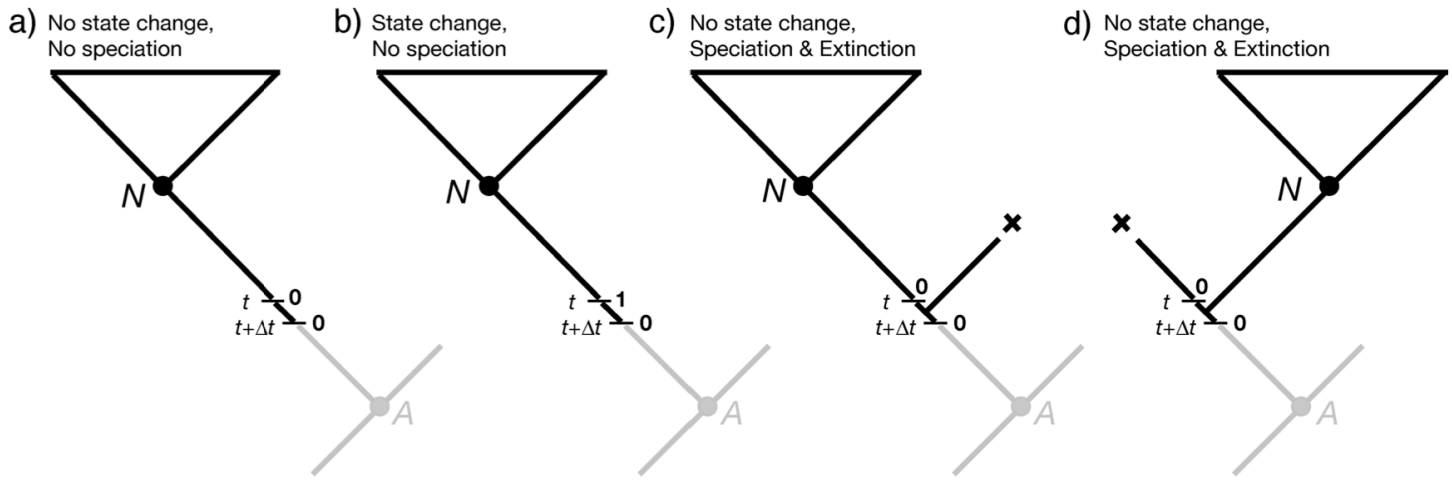




Binary State Speciation and Extinction (BiSSE)

Speciation and extinction rates are affected by the state of a binary character.

Probabilities can no longer be calculated analytically, they have to be estimated numerically.



Maddison et al. 2007 Syst Biol