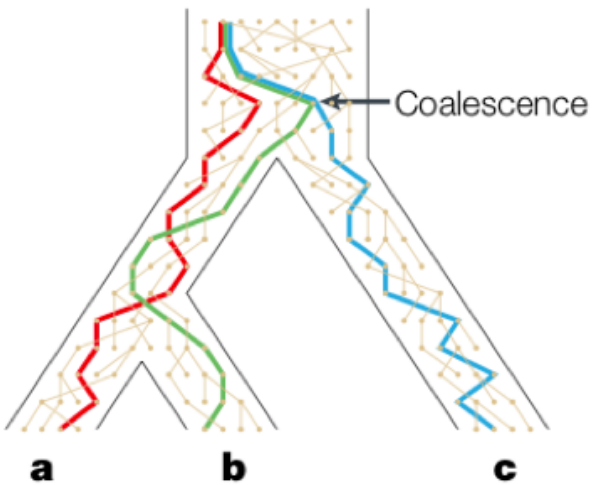
The multi-species coalescent

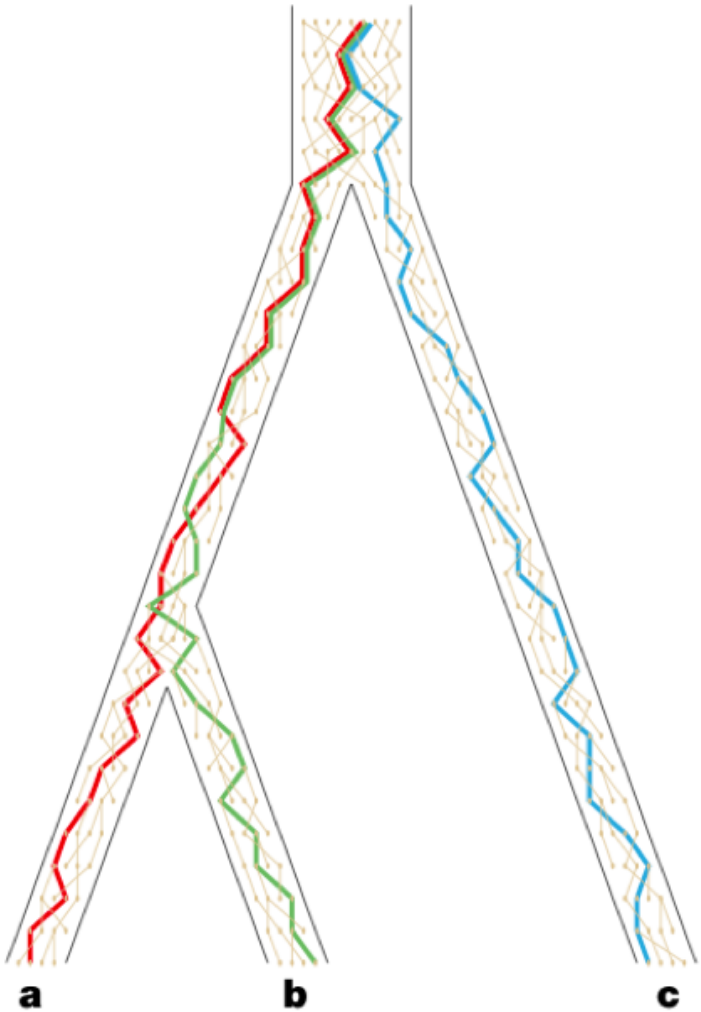
Alexei Drummond, alexei@cs.auckland.ac.nz

Center for Computational Evolution

March 29, 2017

Gene trees and species trees





The coalescent

Data: a small genetic sample from a large background population.

The coalescent

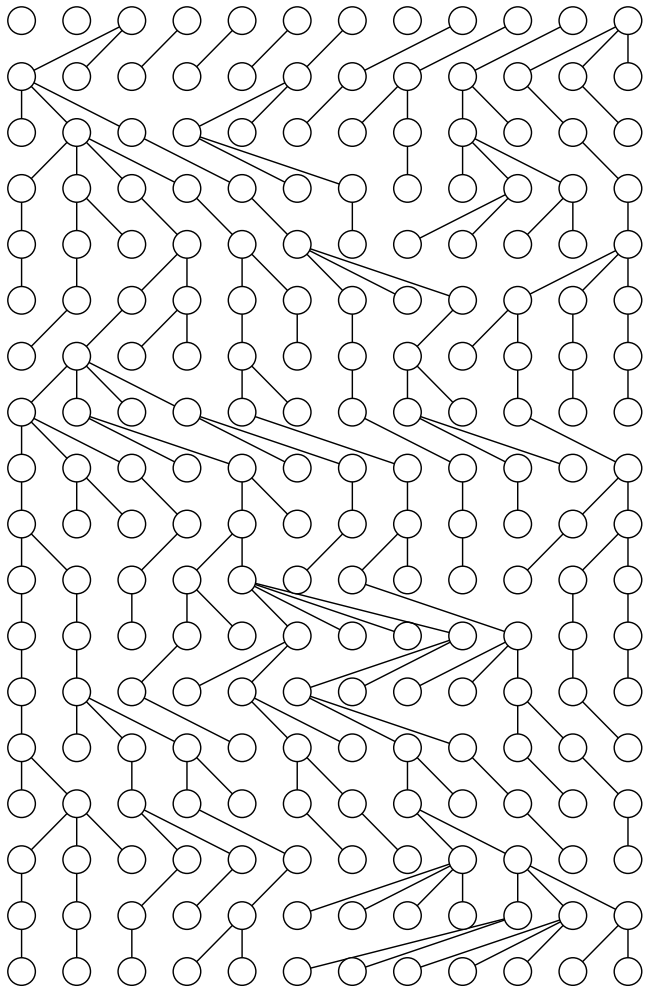
- is a model of the ancestral relationships of a sample of individuals taken from a larger population.
- describes a probability distribution on ancestral genealogies (trees) given a population history, $N(t)$.
  - Therefore the coalescent can convert information from ancestral genealogies into information about population history and vice versa.
- a model of ancestral genealogies, not sequences, and its simplest form assumes neutral evolution.
- can be thought of as a prior on the tree, in a Bayesian setting.

Theoretical population genetics

Most of theoretical population genetics is based on the idealized Wright-Fisher model of population which assumes

- Constant population size N
- Discrete generations
- Complete mixing

For the purposes of this presentation the population will be assumed to be haploid, as is the case for many pathogens.

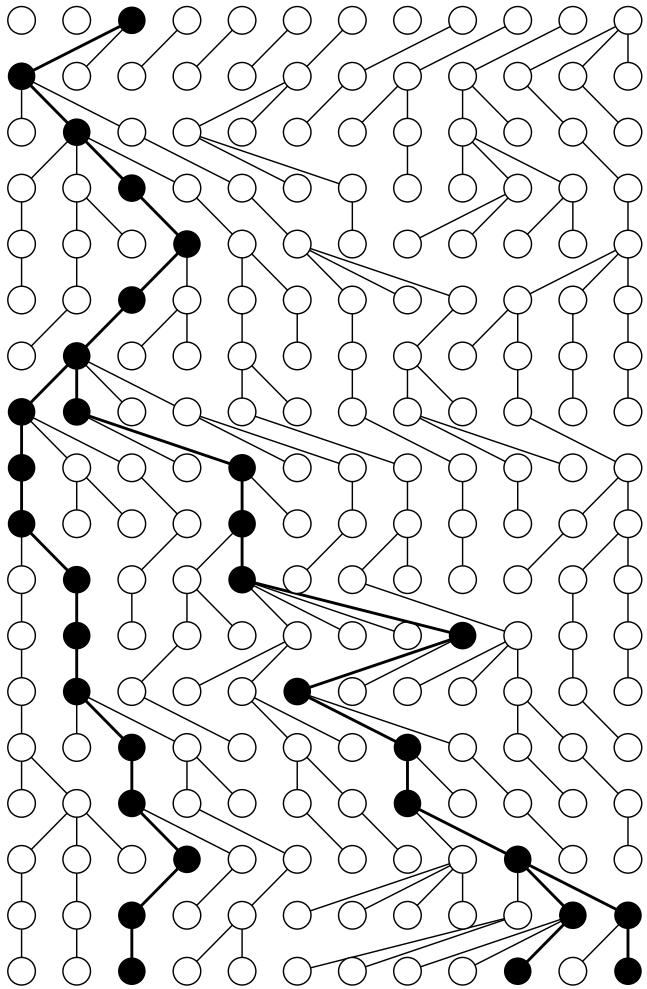


Kingman's n-coalescent

Consider tracing the ancestry of a sample of k individuals from the present, back into the past.

This process is a discrete-time Markov process that eventually *coalesces* to a single common ancestor (*concestor*) of the sample of individuals.

Kingman's n-coalescent is a *continuous-time* diffusion approximation of this process, in the limit of large N, i.e. $N \gg k^2$.

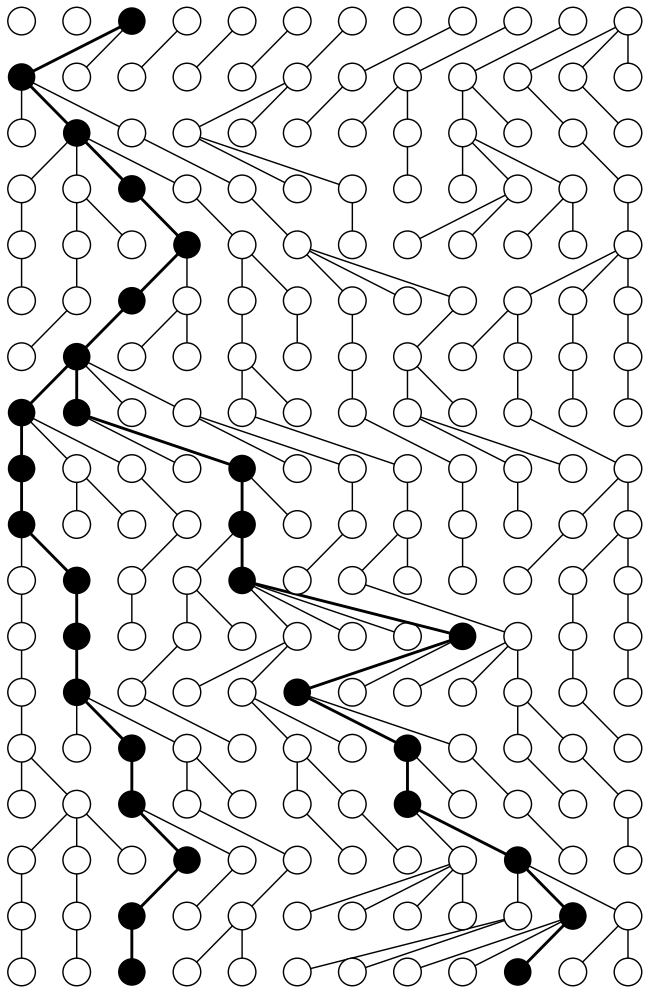


The coalescence of two ancestral lineages

- First, consider two random members from a population of fixed size N.
- By perfect mixing, the probability they share a *concestor* in the previous generation is $1/N$.
- The probability the concestor is t generations back is

$$Pr\{t\} = \frac{1}{N}(1 - \frac{1}{N})^{t-1}.$$

- It follows that $g = t - 1$, has a geometric distribution with a success rate of $\lambda = 1/N$, and so has mean N and variance of $N^3/(N-1)$.

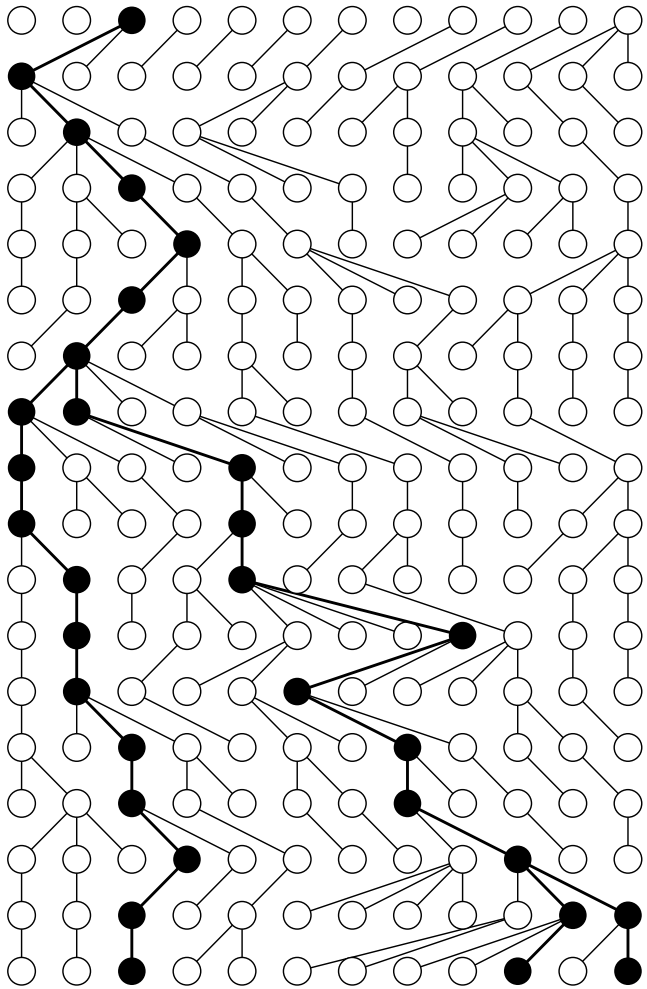


The coalescence of k lineages

With k lineages the time to the first coalescence is derived in the same way, only now there are $\binom{k}{2}$ possible pairs that may coalesce, resulting in a success rate of $\lambda = \binom{k}{2}/N$ and mean time to first coalescence (t_k) of

$$E[t_k] = \frac{N}{\binom{k}{2}}.$$

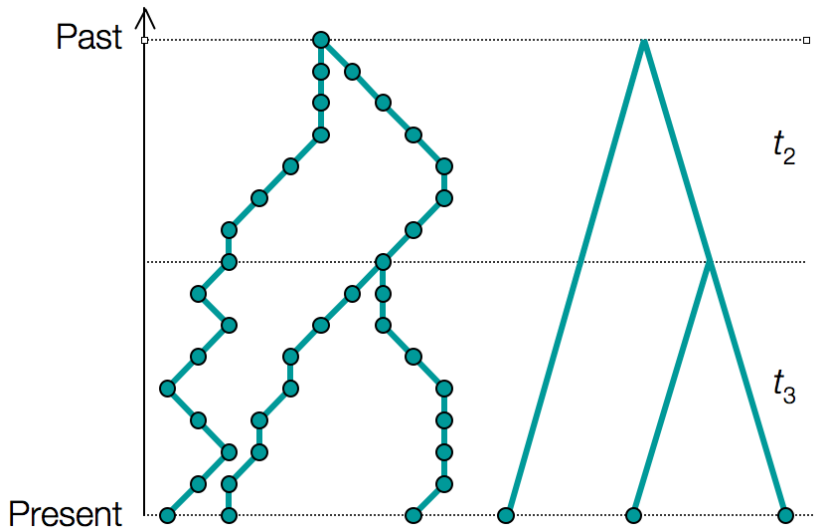
This implicitly assumes that N is much larger than $O(k^2)$, so that the probability of two coalescent events in the same generation is small.



The coalescent is a *diffusion approximation*

Kingman (1982) showed that as N grows the coalescent process converges to a continuous-time Markov chain.

$\lambda = \binom{k}{2}/N$ is the rate of coalescence, i.e. the probability of coalescing a pair from k lineages on a short time interval Δt is $O(\lambda \Delta t)$. Unsurprisingly the solution turns out to be the exponential distribution:

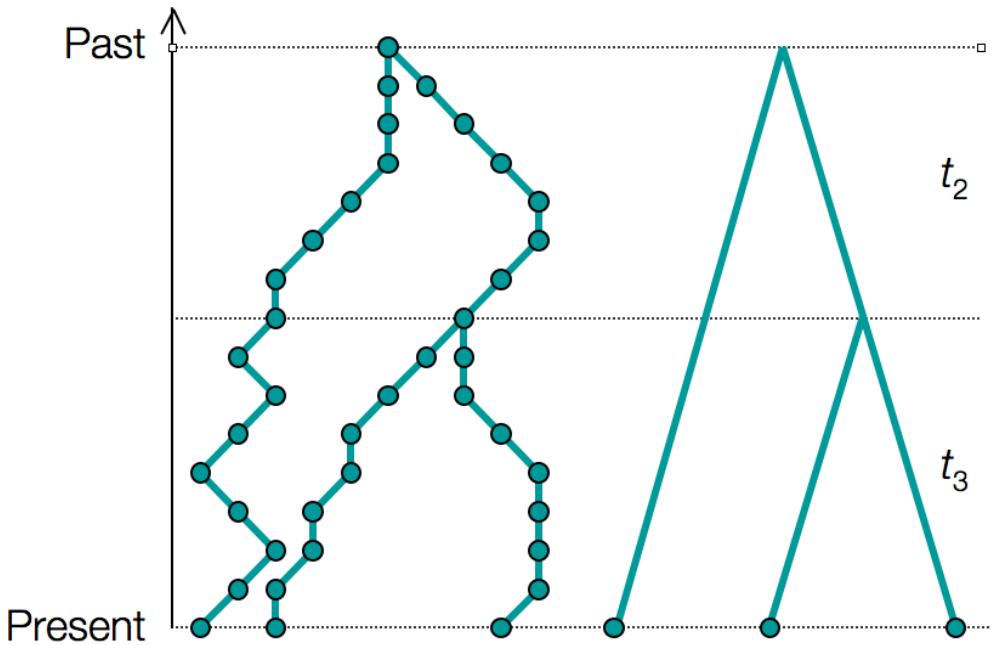


$$f(t_k) = \frac{\binom{k}{2}}{N} \exp\left(-\frac{\binom{k}{2} t_k}{N}\right) .$$

The coalescent density for a genealogy

For a genealogy with coalescent times $\mathbf{t} = \{t_2, t_3, ..., t_n\}$ we can write the probability density, given N:

$$f(\mathbf{t}|N) = \frac{1}{N^{n-1}} \prod_{k=2}^{n} \exp\left(-\frac{\binom{k}{2} t_k}{N}\right) .$$



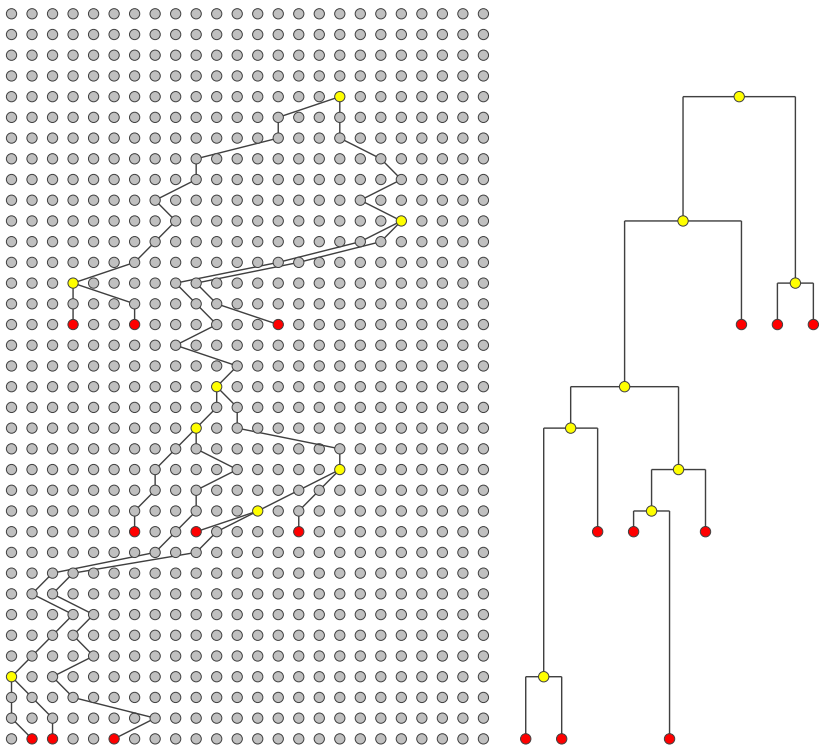
The coalescent density with varying population size

The generalization of the coalescent for the case where the population size changes over time, $N = N(t)$ is given by Griffiths and Tavare (1994). They showed that the coalescent density for the first coalescence event being at time t in the past given n lineages is:

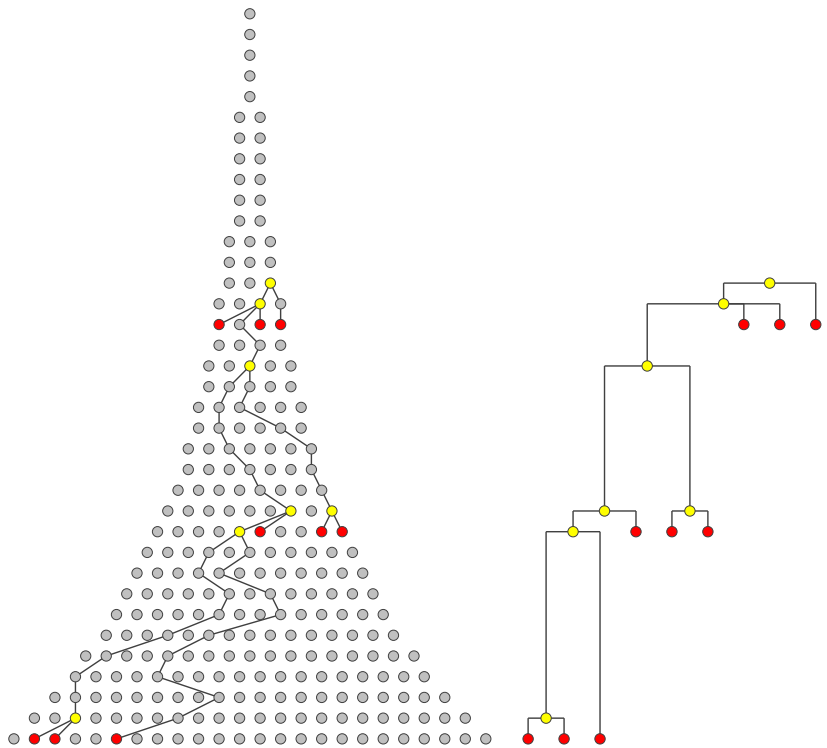
$$f(t) = \frac{1}{N(t)} \exp\left(-\int_0^t \frac{\binom{n}{2}}{N(x)} dx \right)$$

The coalescent with serial samples

Many epidemiological agents, like RNA viruses, evolve very rapidly, so that the effect of sampling the population at different times becomes important.

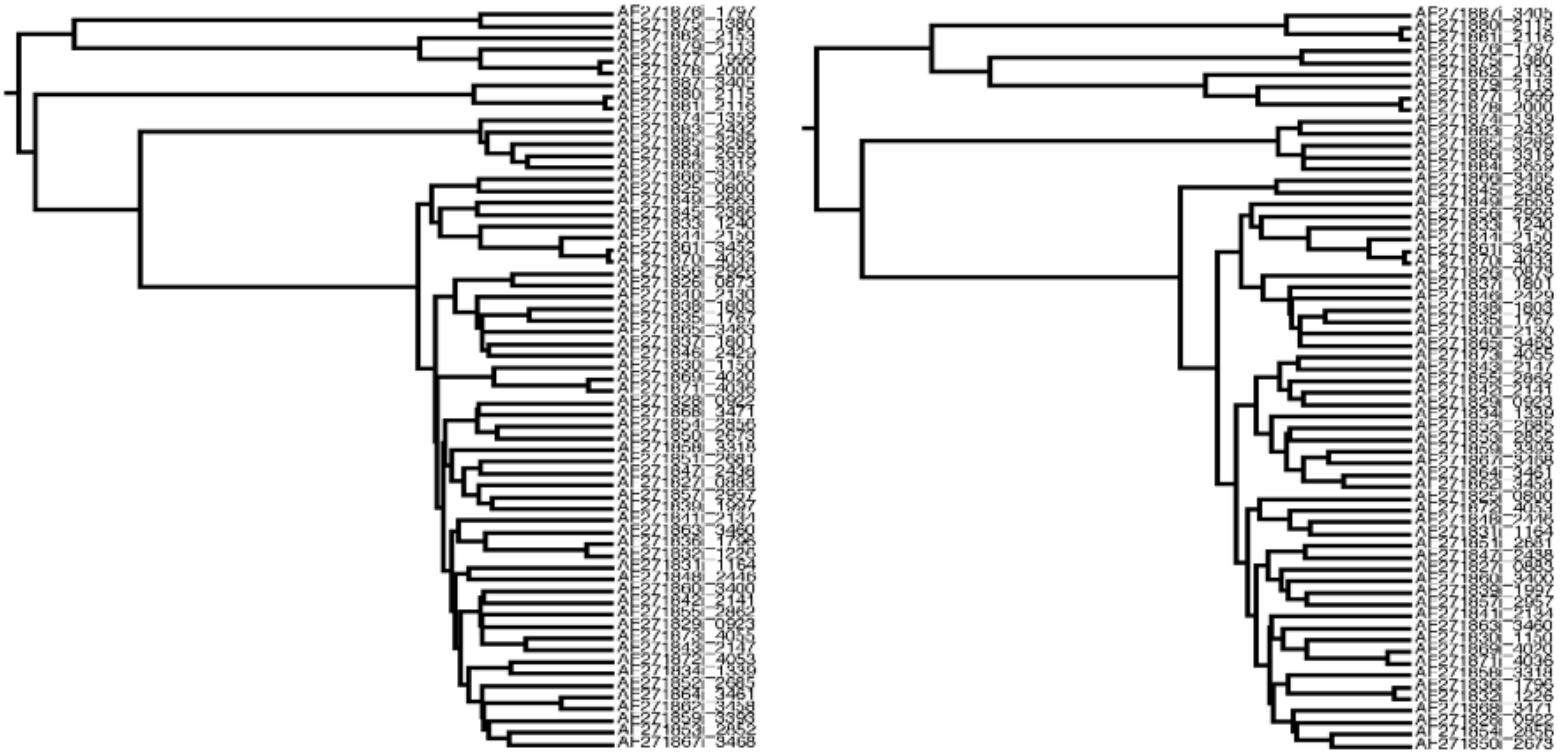


Constant size



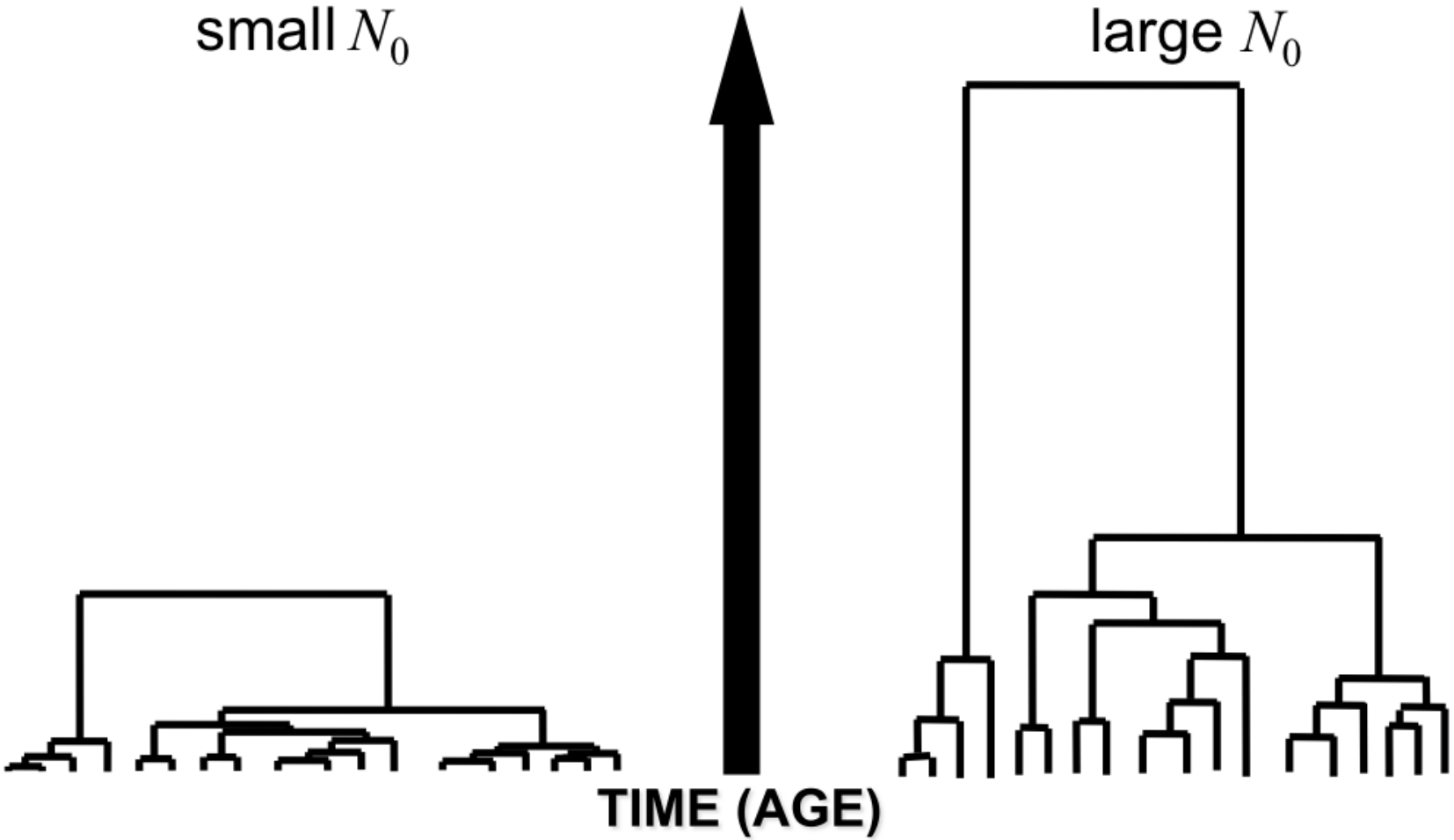
Exponential growth

Bayesian integration of uncertainty in genealogies



How similar are these two trees? Both of them are plausible given the data. We can use Bayesian Markov-chain Monte Carlo to average the coalescent over all plausible trees.

Constant population size: $N(t) = N_0$



The coalescent: shapes of genealogies

Constant size

$$N(t) = N_0$$

Exponential growth

$$N(t) = N_0 exp[-rt]$$

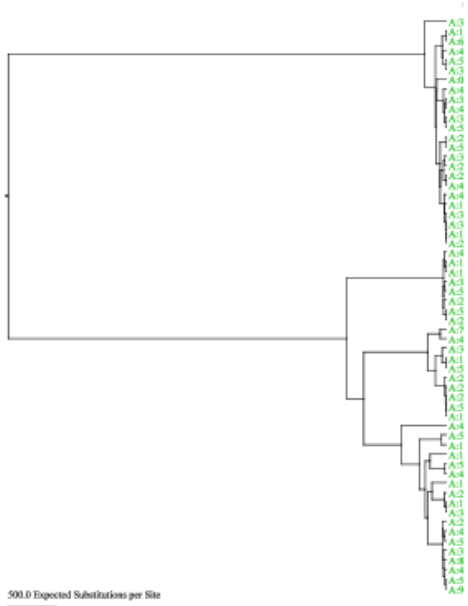
The coalescent can be used to convert coalescent times into knowledge about population size and its change though time.

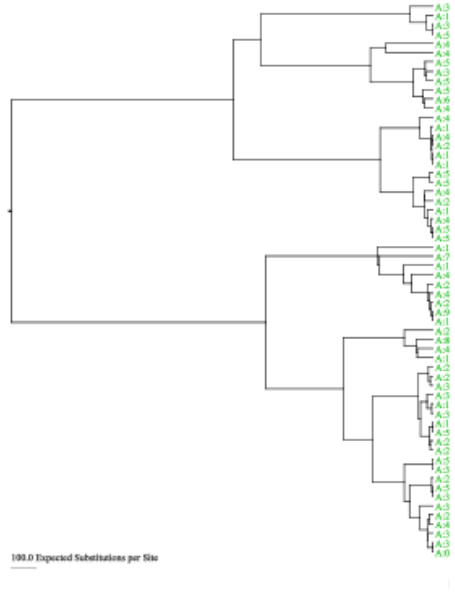
The murky boundary between population genetics and phylogenetics

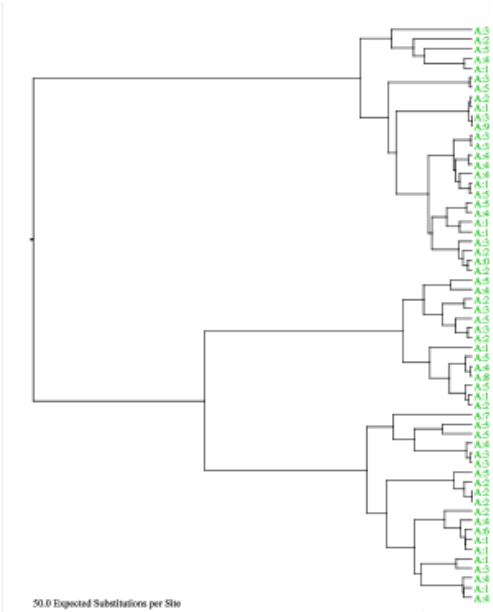
There has been increased interest in analyses of closely related species, where the effect of population genetic processes, such as the coalescent can't be ignored.

- Different gene trees can have different topologies due to incomplete lineage sorting
- Divergence times of species can be overestimated due to ancestral polymorphism
- Sometimes the exact species identities of individuals are not known
- Sometimes researchers identify species based on a split in a single gene tree.

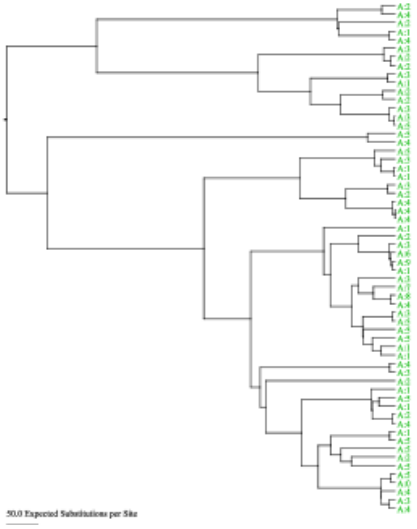
Enter the multi-species coalescent.

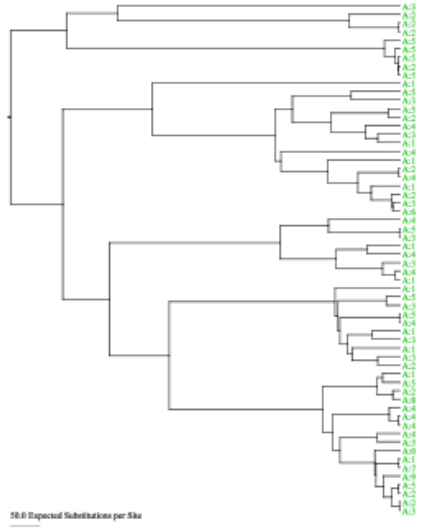




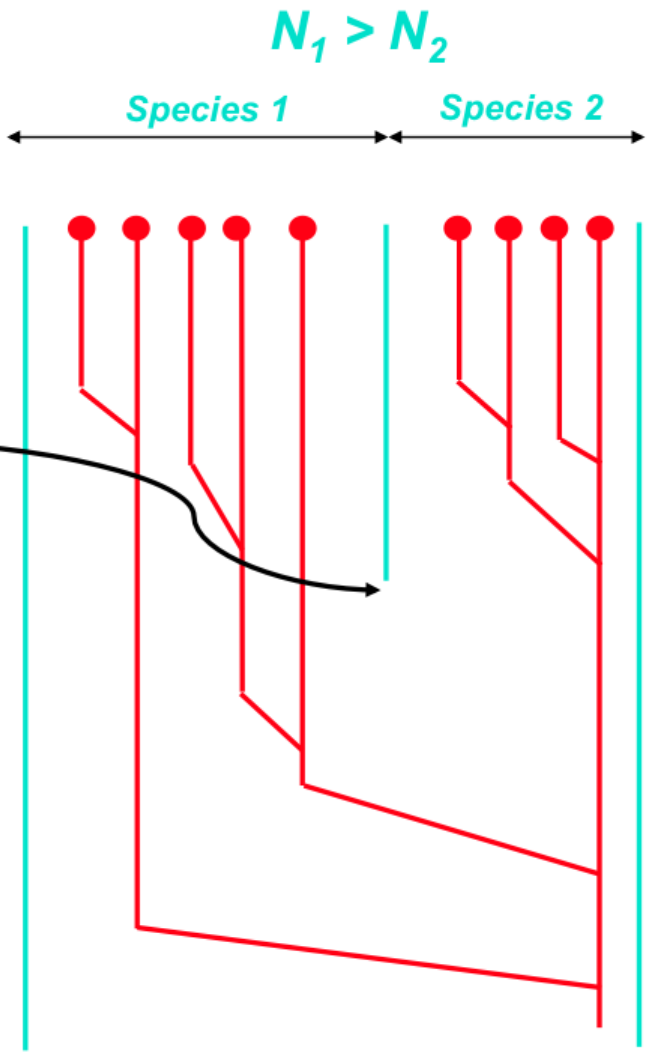


All these patterns are the consequence of lineage drift within a single population





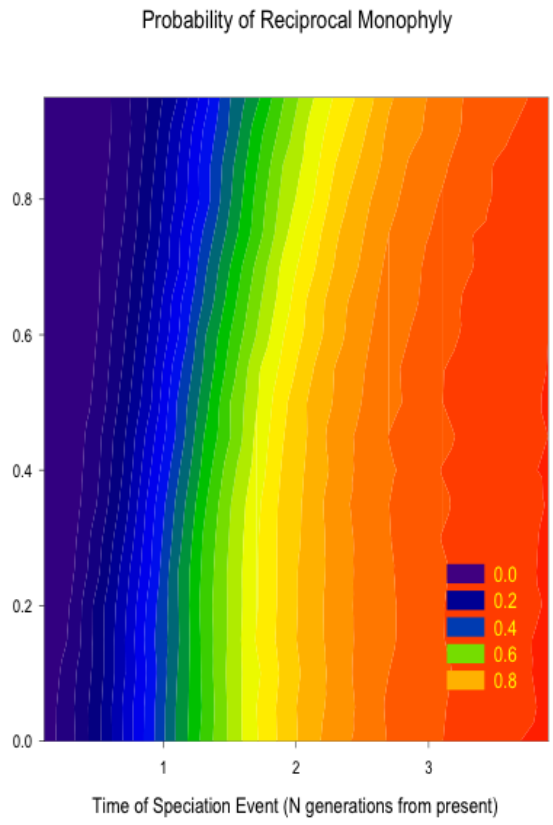
Incomplete Lineage Sorting

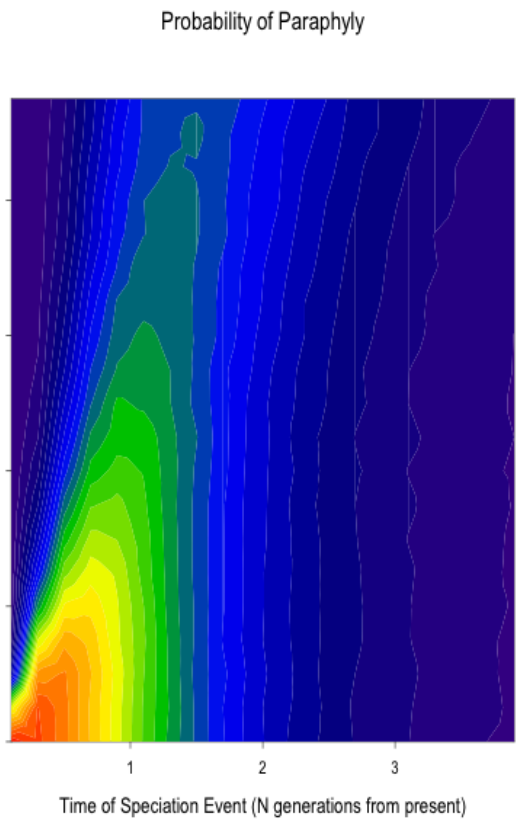


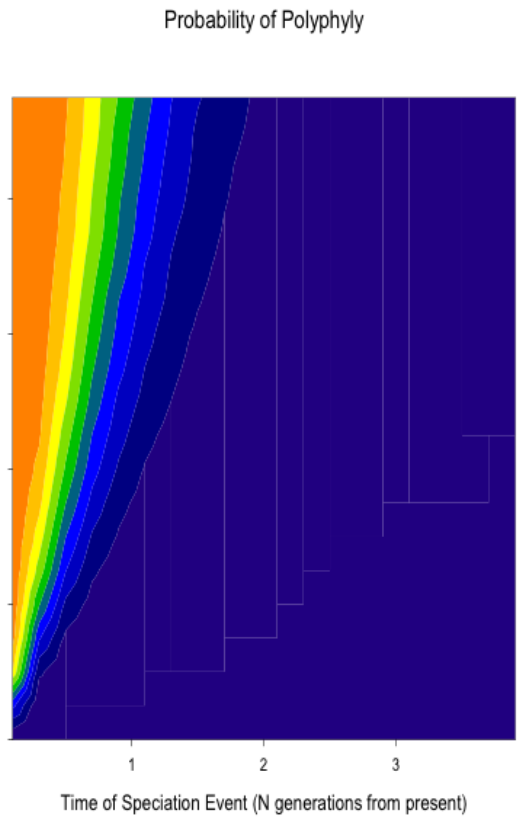
- The probabilities of seeing reciprocal monophyly, paraphyly or polyphyly, depend on:

1. The time since speciation, and
2. The effective sizes of the two species.

Probabilities of Monophyly, Paraphyly and Polyphyly

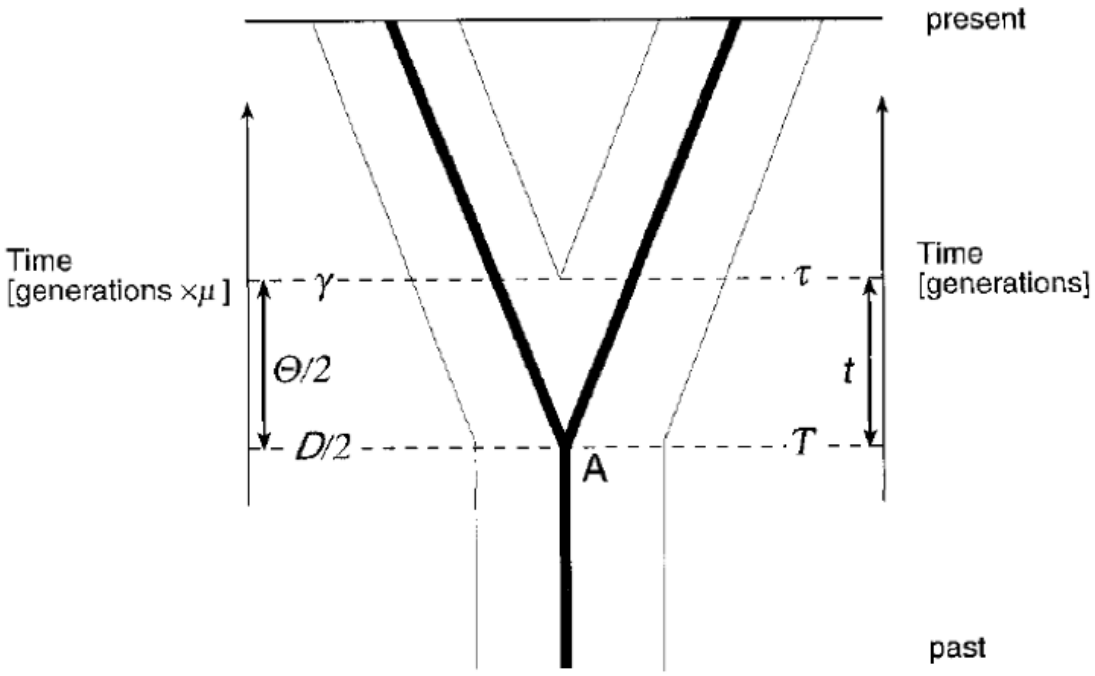






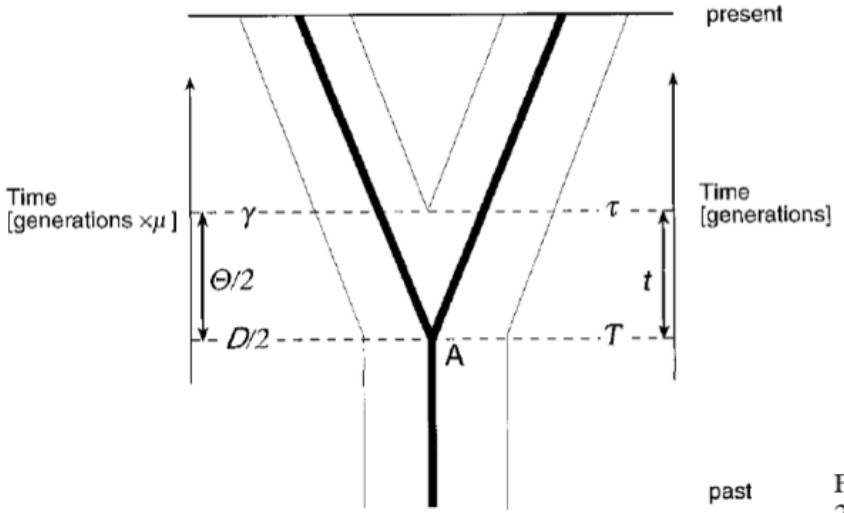
Problems with estimation of divergence times

- Typically we use gene phylogenies to estimate species phylogenies
- But divergence time for genes will be longer than that of species.



From Edwards and Beerli, 2000, Evolution **54**: 1839-1854

Problems with estimation of divergence times



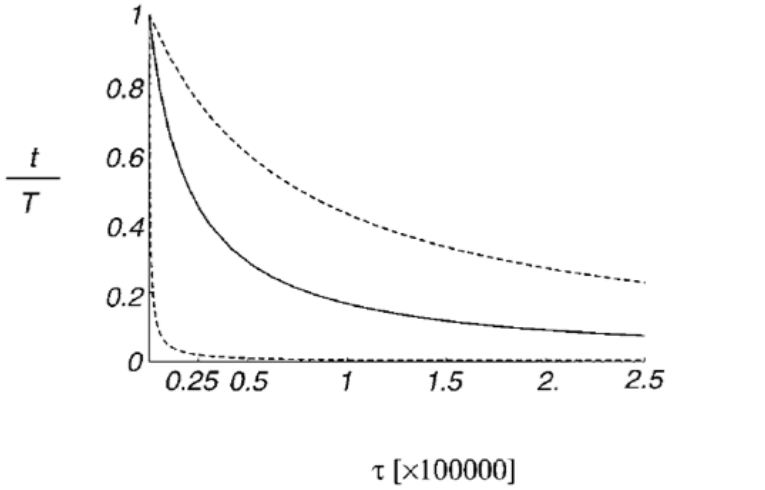
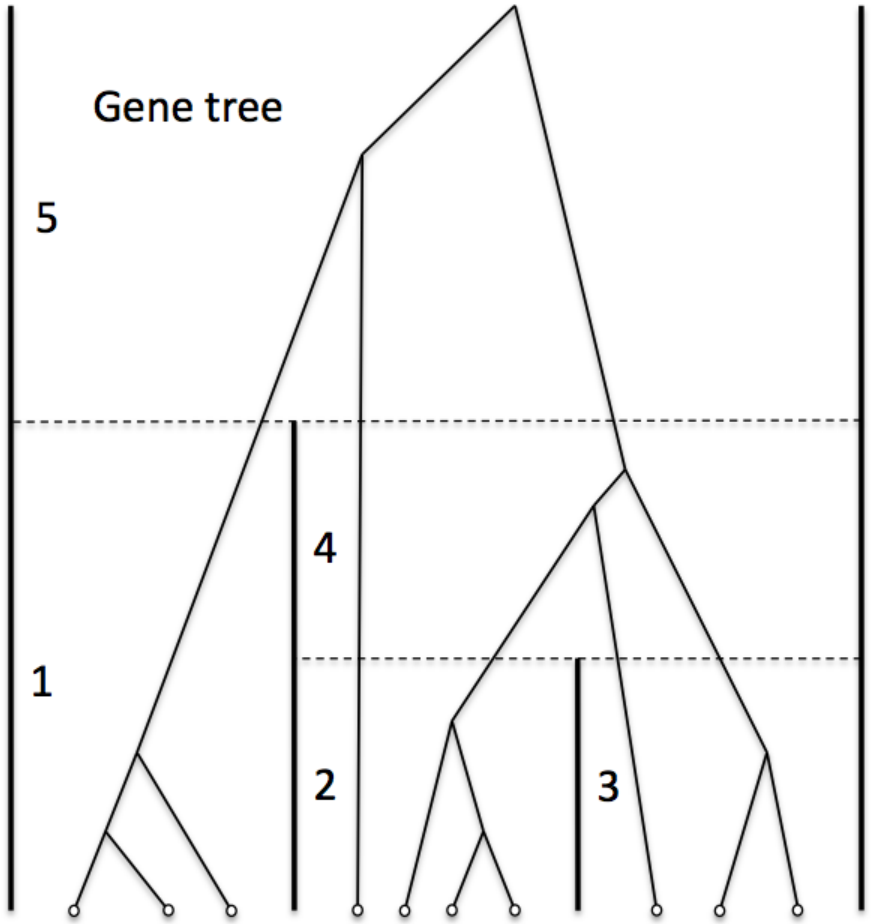
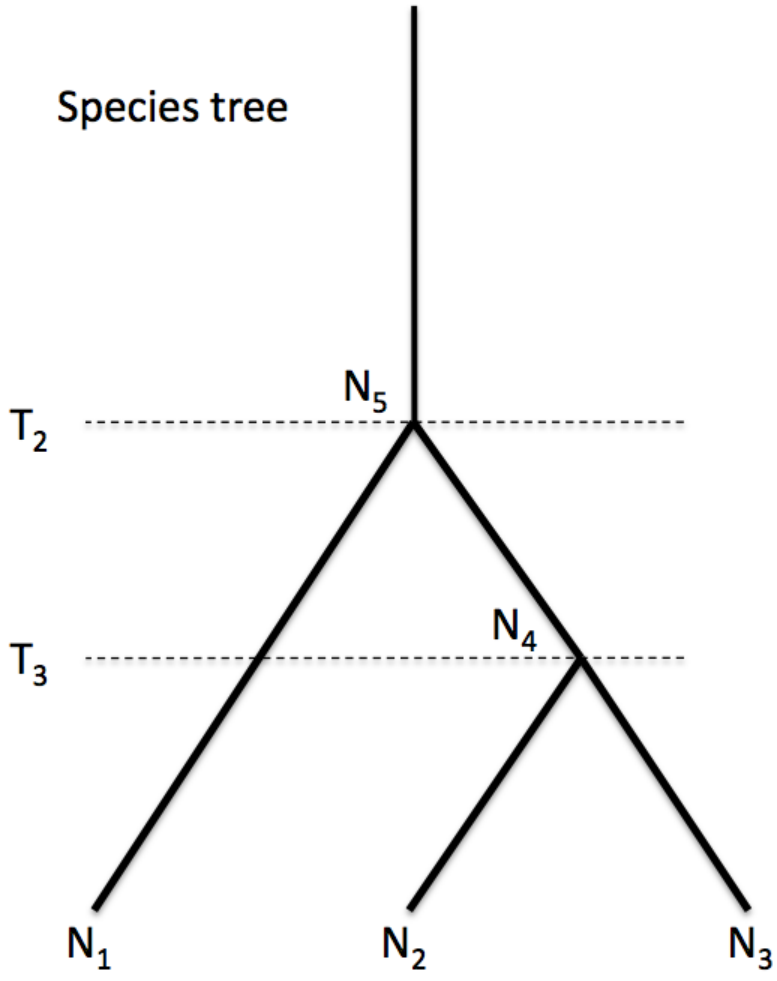


FIG. 2. The fraction of the total interspecific coalescence $T = t + 2N_e$ generations taken up by ancestral divergence coalescence t. An effective population size of 100,000 was assumed. Dashed lines indicate upper and lower 95% confidence intervals.

From Edwards and Beerli, 2000, Evolution **54**: 1839-1854

The multispecies coalescent





*BEAST

- Coestimate Species tree, Gene trees, Population sizes and all other parameters.
- Any combination of number of individual and genes from each individual.
- Gene trees estimated using any BEAST models, any type of linkage between parameters.
- For example, all mutations rates may be equal or separate.

*BEAST

Given data D we define the posterior distribution of the species tree S as follows,

$$P(S|D) \propto \int_G \left(\prod_{i=1}^{n} P(d_i|g_i) P(g_i|S) \right) P(S) dG \qquad (1)$$

The data $D = d_1, d_2, \ldots, d_n$ is composed of n alignments, one per locus. $G = (G_1 \times G_2 \times \cdots \times G_n)$ is the space of all gene trees over the respective alignments where $g_i \in G_i$ is one specific gene tree on the i^{th} alignment.

The term $P(d_i|g_i)$ is the "Felsenstein" likelihood of the data given a gene tree, $P(g_i|S)$ is the multispecies coalescent and $P(S)$ is a prior distribution on the space of species trees. Like most coalescent models this assumes no recombination within loci and free recombination between loci.

The Species Tree: S

Define N_i to be the effective population size at the present for the species $i \in \{1, 2, \ldots, n\}$, and A_i the ancestral effective population of species i at the time of the species origin.

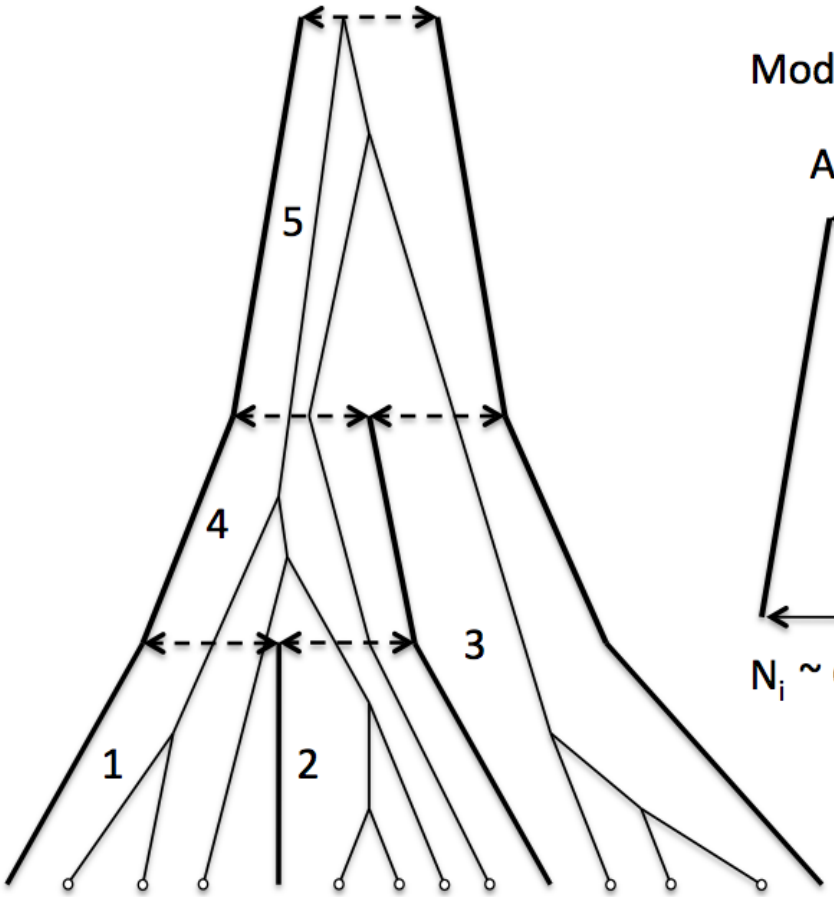
Define T to be a time tree with ranked topology and divergence times.

Then a species tree S is a time tree T with population sizes associated with both internal and external nodes:

$S = \{A, N, T\}$

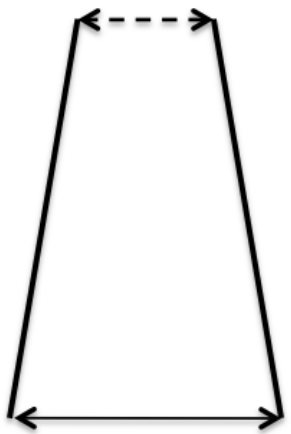
$P(S) = P(A, N|T)P(T)$

Population sizes prior: $P(A, N|T)$



Modern species

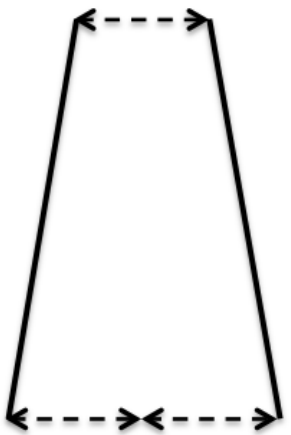
$A_i \sim \text{Exp}(\Theta)$



$N_i \sim \text{Gamma}(2,\Theta)$

Ancestral species

$A_i \sim \text{Exp}(\Theta)$



$A_{left(i)} \sim \text{Exp}(\Theta)$ $A_{right(i)} \sim \text{Exp}(\Theta)$

i.e. $A_{left(i)} + A_{right(i)} \sim \text{Gamma}(2,\Theta)$

Species divergence times prior: $P(T|\lambda)$

For a species tree of n species, define T_i to be the time at which the species tree goes from having i to $i-1$ species, back in time. Additionally define $\tau_i = T_i - T_{i+1}, i \in \{2, \ldots, n-1\}$ and $\tau_n = T_n$.

The Yule speciation prior supposes a uniform rate species birth (λ) on all lineages, implying a prior of:

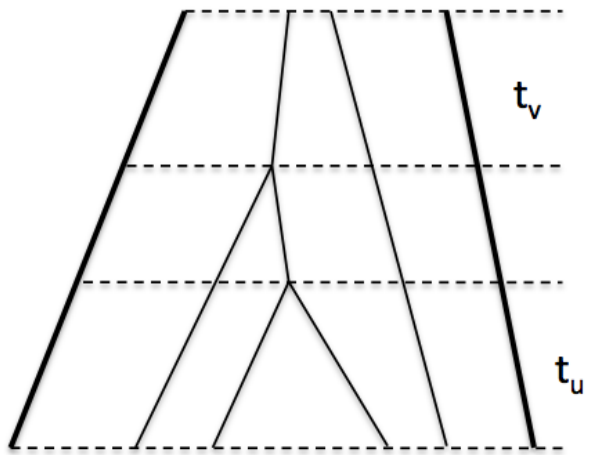
$$\tau_i \sim \text{Exp}(1/i\lambda)$$

More complex species tree priors that admit species extinction (Birth-death prior; Gernhard, 2008) and incomplete sampling (Birth-death-sampling; Stadler, 2009) are also possible.

All of these species tree priors imply a uniform prior on labelled histories.

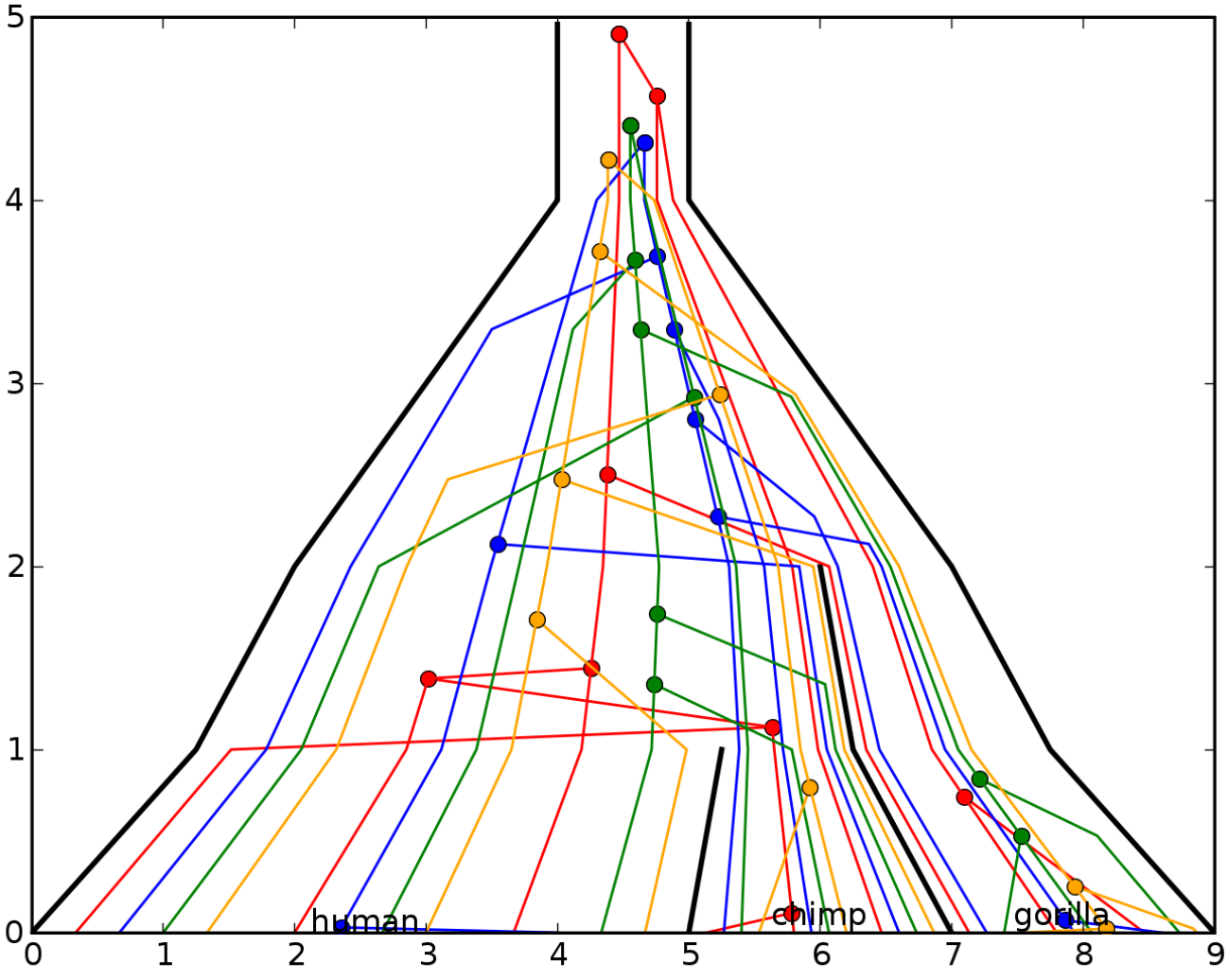
Coalescent prior for gene trees: $P(g_i|S)$

Consider a single species in the species tree, spanned by $k = u - v$ coalescent intervals (and a final interval without a coalescent event). t_k is the time during which there are k lineages. Define $N(s)$ as the population size of this species at time s. Define $s_i = \sum_{k=u}^{i} t_k$. The prior density for each interval ending in a coalescent is:



$$f(t_k) = \frac{1}{N(s_k)} \exp\left(- \int_{s_{k-1}}^{s_k} \frac{\binom{n}{2}}{N(x)} dx \right)$$

Four gene trees inside a 3-species tree

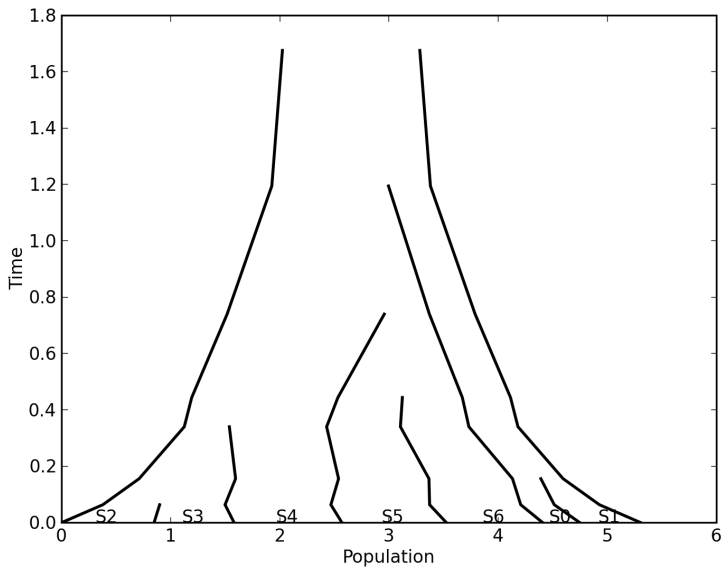


Simulating a rapid radiation of 7 species

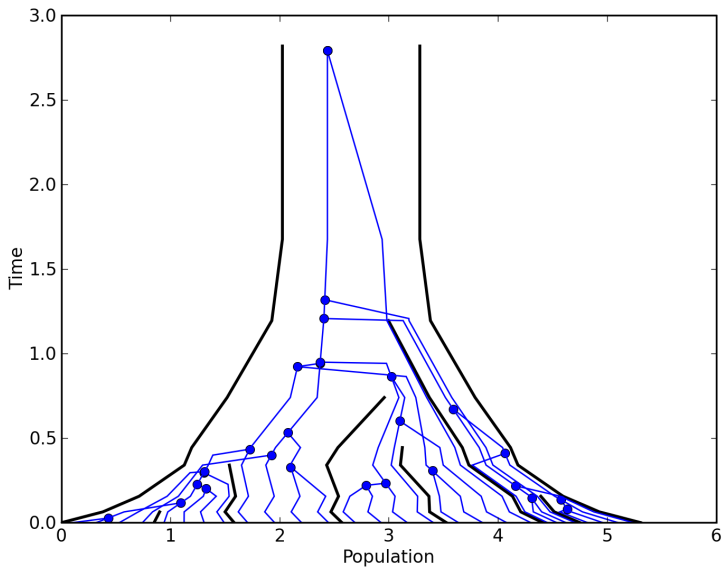
Repeat 100 times:

1. Simulate a random species tree of 7 extant species from Birth-death process
2. Simulate gene trees in species tree $(1, 2, 3, \ldots$ genes)
3. Simulate sequences down each gene tree $(400, 800, 1600$ nucleotides)

Species tree



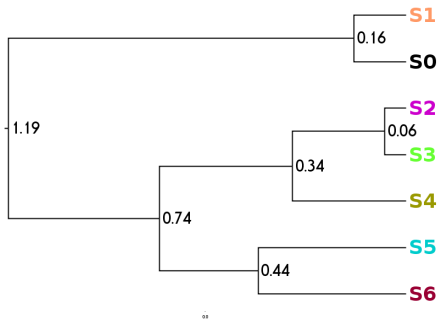
Species tree and one gene tree



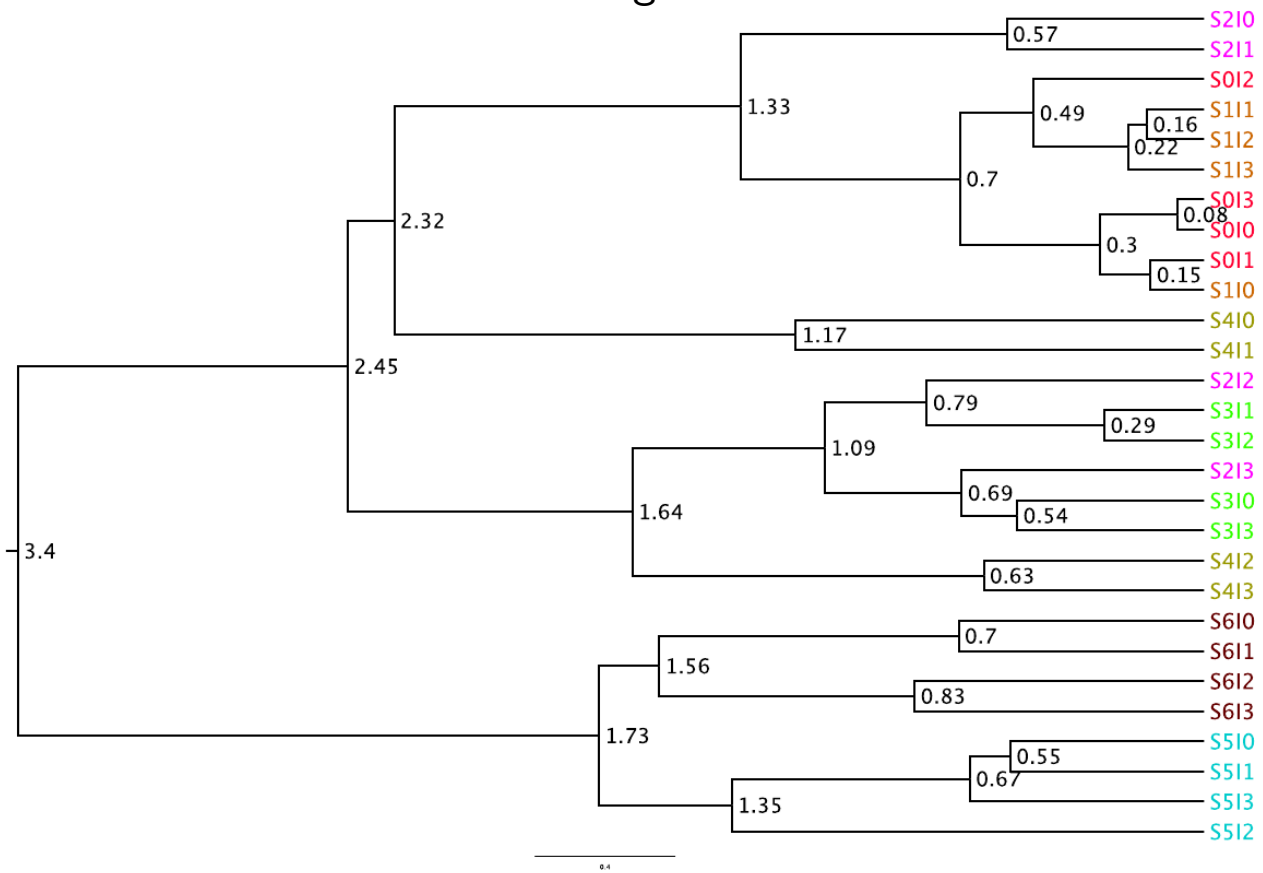
Supermatrix concatenation

a terrible idea for rapid radiations

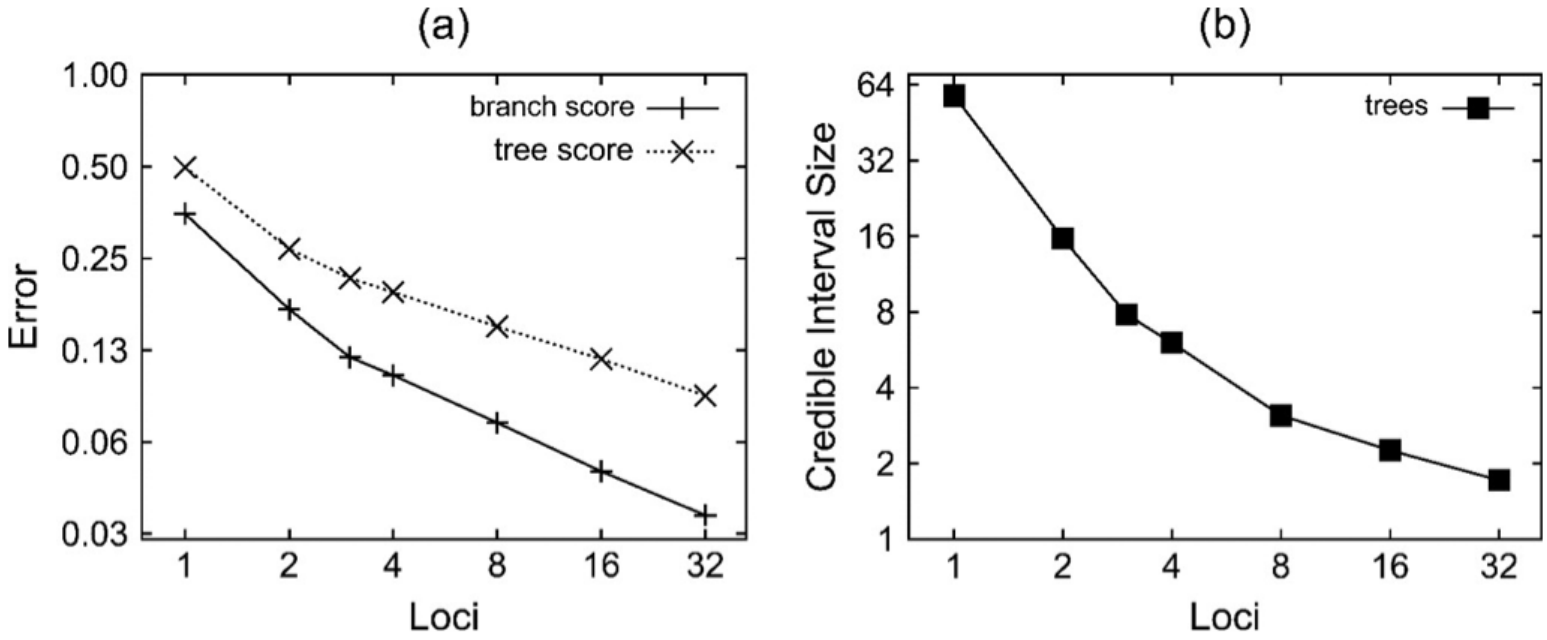
Species tree



"Gene" tree from concatenated genes



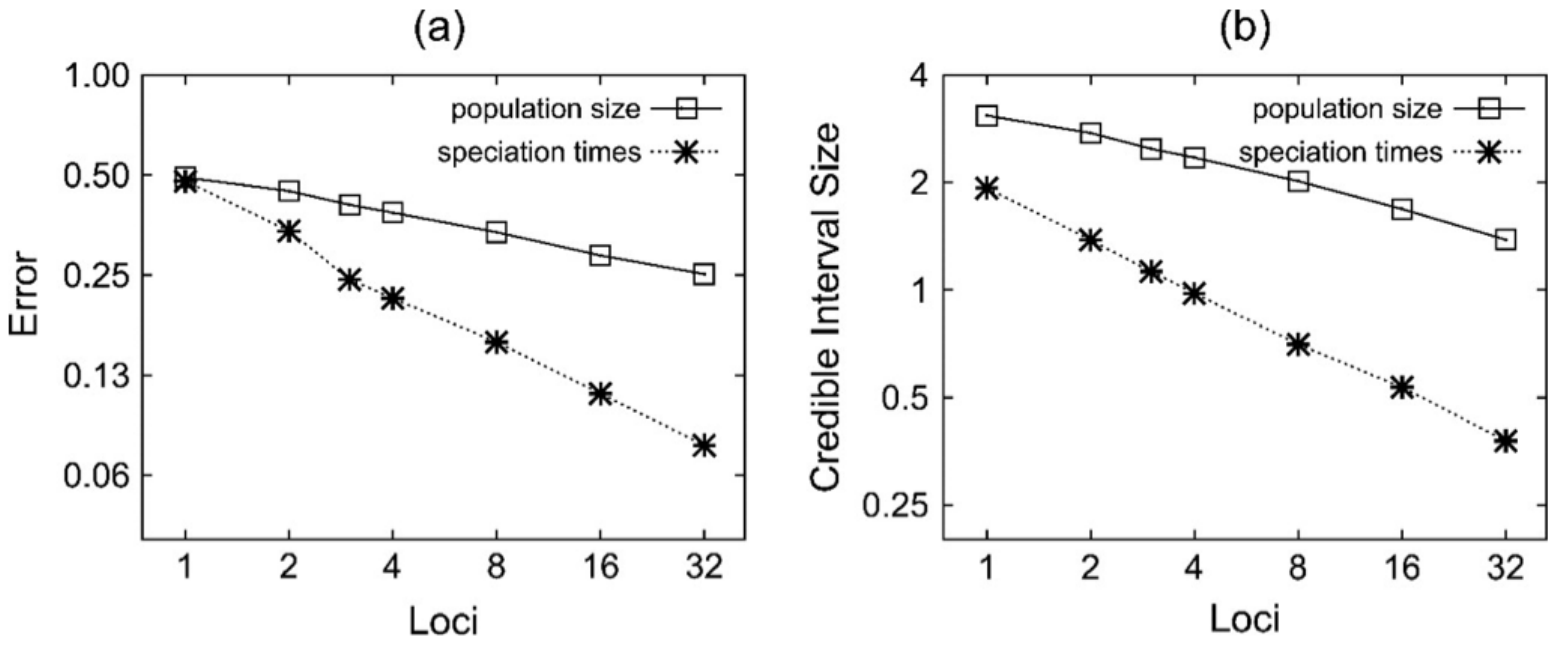
(a) Species tree estimation error and (b) 95% credible interval size as a function of the number of loci.



Heled J , Drummond A J Mol Biol Evol 2010;27:570-580

MBE MOLECULAR BIOLOGY AND EVOLUTION

(a) Relative error and (b) credible interval size for both population size and speciation time point estimates.



Heled J , Drummond A J Mol Biol Evol 2010;27:570-580

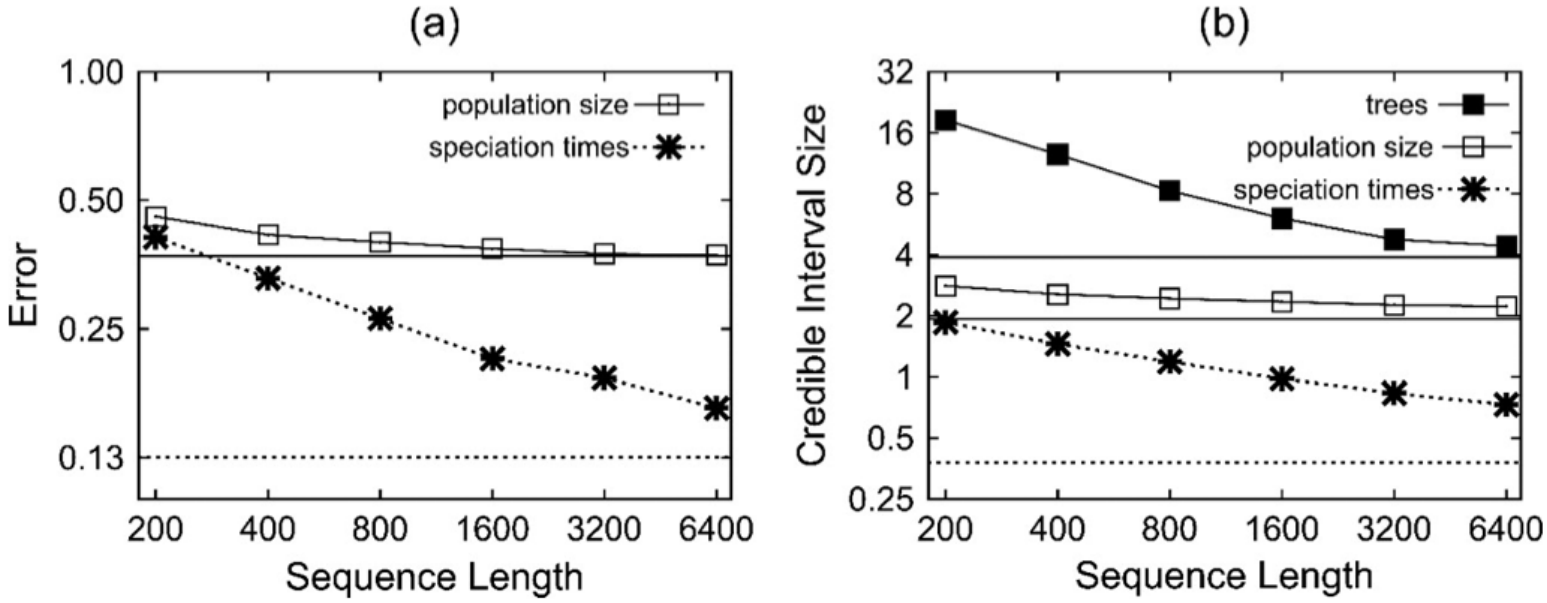
MBE MOLECULAR BIOLOGY AND EVOLUTION

Table 1.
Summary of Seven Taxa, Four Loci Species Tree Estimation Where Genes Evolve at Different Rates. The Final Row Represents the Case Where the Model is Misspecified: The Truth is That Each Gene Has a Different Rate, But the Method Assumes That All Genes Have the Same Substitution Rate

Data/model	Topology inside 95%	Mean 95% size	Normalized branch score	Normalized tree score	Speciation time inside 95%	Speciation time error/CI size	Population size error/CI size
Equal rates/equal rates	94	7.86	0.10	0.19	93	1.36/10.0	2.2/162
Rates vary/rates vary	95	8.08	0.12	0.19	93	1.43/12.0	2.2/189
Rates vary/equal rates	92	10.24	0.16	0.20	67	1.57/12.6	2.5/189

CI, credible interval.

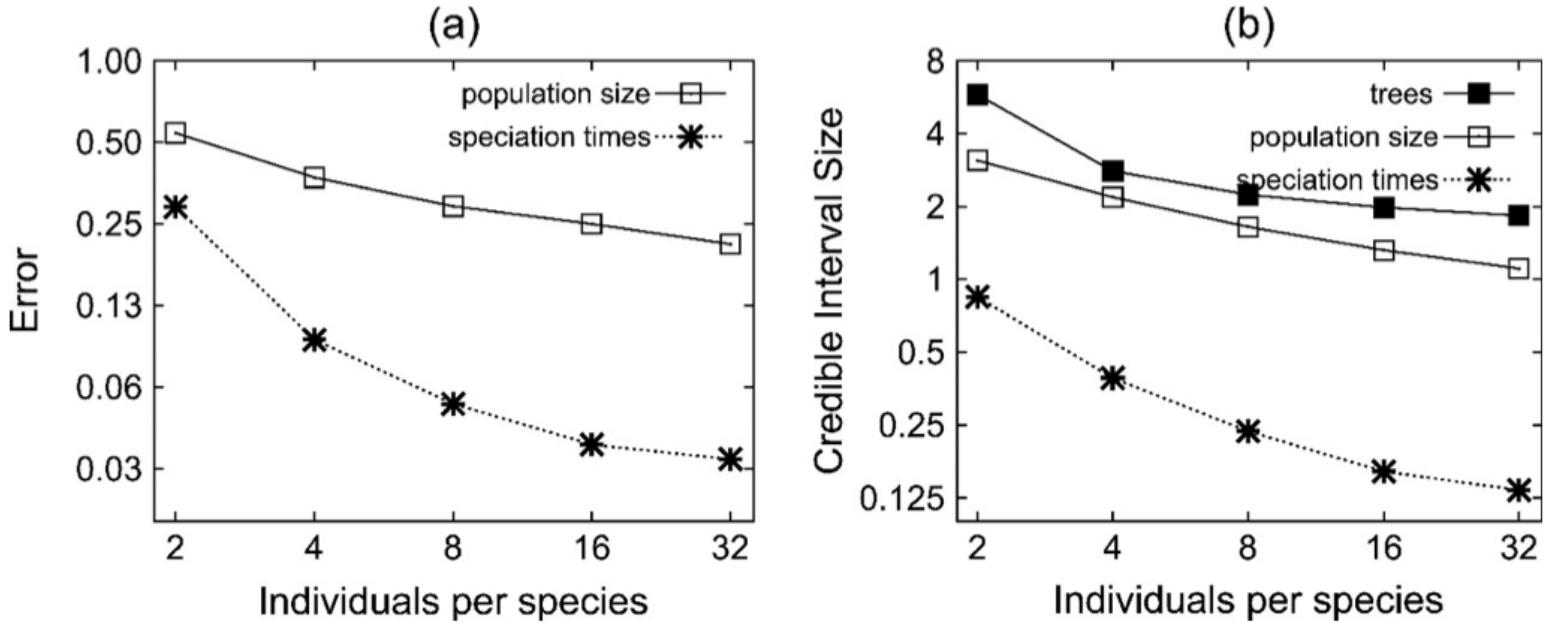
(a) Relative error and (b) credible interval sizes as a function of sequence length.



Heled J , Drummond A J Mol Biol Evol 2010;27:570-580

MBE MOLECULAR BIOLOGY AND EVOLUTION

(a) Relative error and (b) credible interval sizes, as a function of number of individuals sample from each species.



Heled J , Drummond A J Mol Biol Evol 2010;27:570-580

MBE MOLECULAR BIOLOGY AND EVOLUTION

Comparing *BEAST and BEST and Supermatrix methods

100 simulated species trees (each with 7 species) with four individuals per species and four loci.

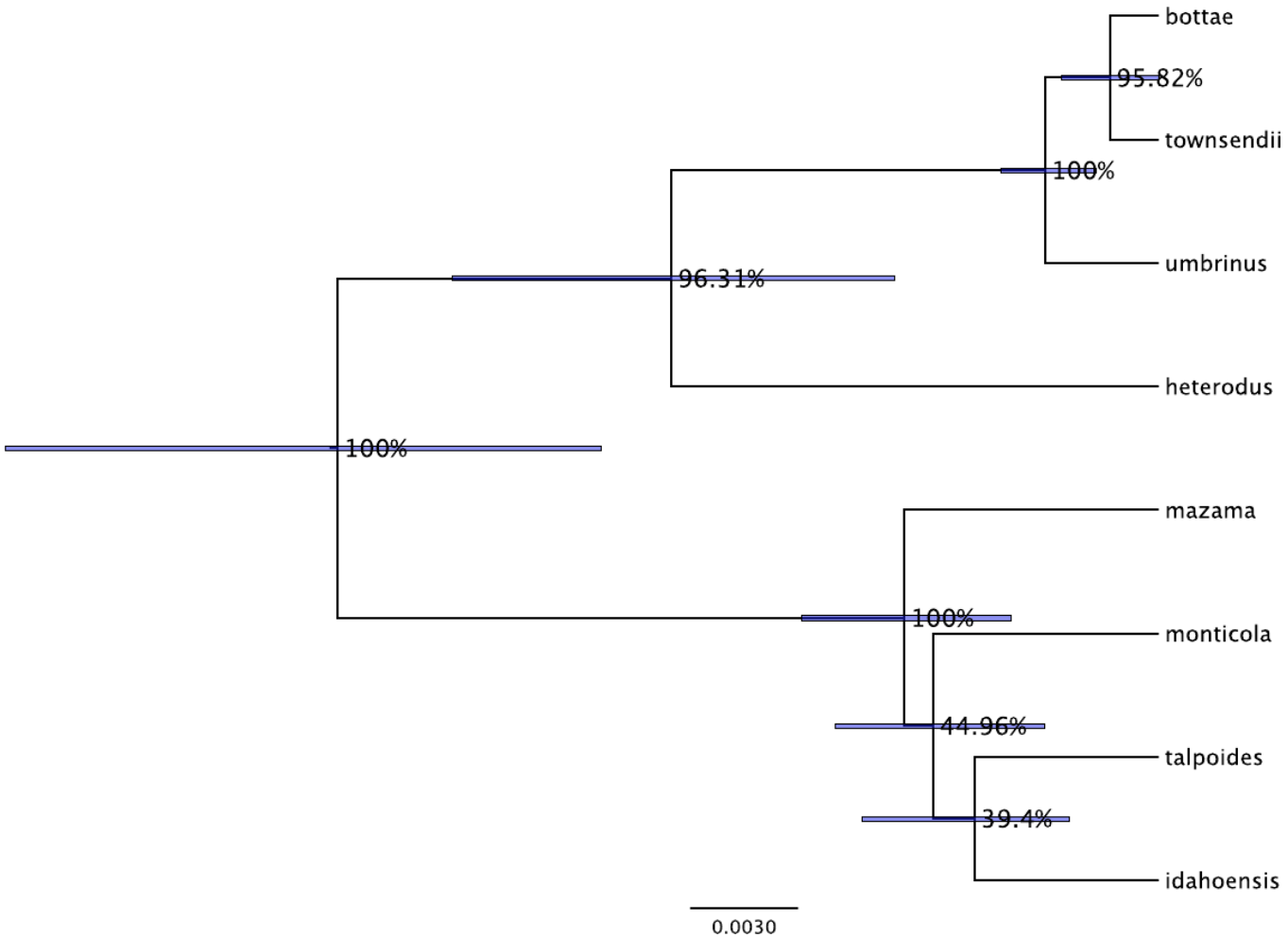
	topology inside 95%	mean 95% size	normalized branch score
*BEAST	97	11.78	0.10
BEST	88	12.88	0.58
Supermatrix	9	1.4	0.77

	tree score	speciation time inside 95%	speciation time error/CI size
*BEAST	0.20	96%	0.41/1.49
BEST	0.64	56%	1.32/2.06
Supermatrix	N/A	0.6%	21.12/5.28

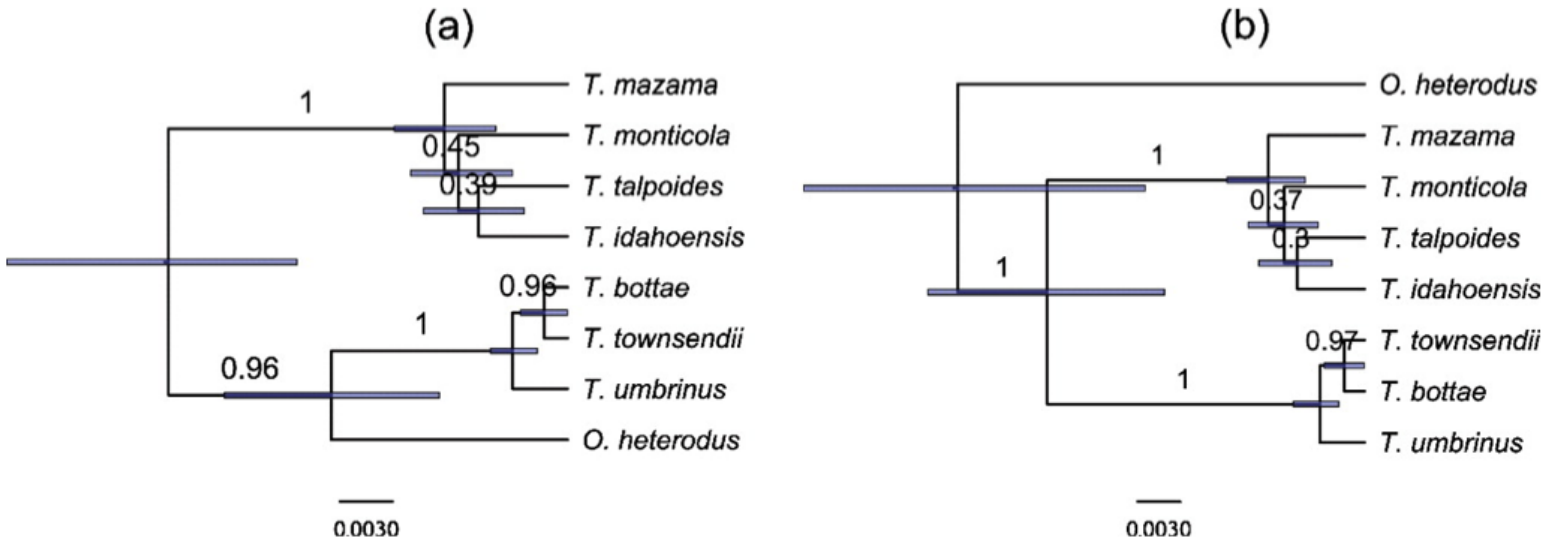
	population size inside 95%	population size error/CI size
*BEAST	98%	0.33/1.11
BEST	56%	0.59/2.15
Supermatrix	N/A	N/A

Pocket Gophers
Data from (Belfiore, 2008)

27 individuals, 7 loci (12 from *T. bottae*, 23 from others, 1 from outgroup)



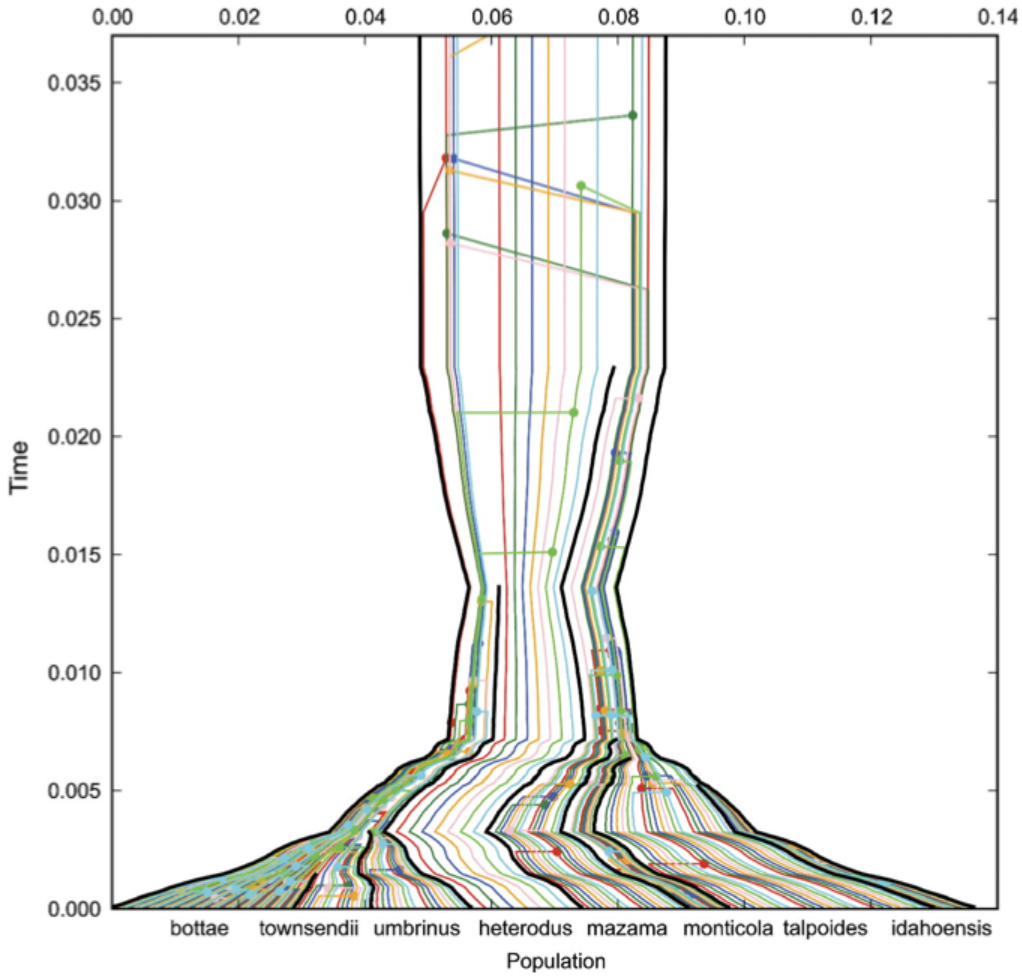
Phylogeny for seven groups of western pocket gophers (Geomyidae, Thomomys).



Heled J , Drummond A J Mol Biol Evol 2010;27:570-580

MBE MOLECULAR BIOLOGY AND EVOLUTION

Western pocket gophers (Geomyidae, Thomomys) species tree with embedded gene trees, each in a different color.



Heled J , Drummond A J Mol Biol Evol 2010;27:570-580

MBE MOLECULAR BIOLOGY AND EVOLUTION

Open questions

- Are there better priors for the population sizes?
- What are efficient proposal distributions for MCMC on the multispecies coalescent?
- What about uncertain species identification?
  - How do we characterize the prior distribution on species associations? (geography, morphology...)
- What about uncertain numbers of species?!
  - How do we characterize the hypothesis space over a species trees with a *random* number of species, and gene tree tips with an uncertain species identity?
  - Reversible-jump, yes, but what proprosal distributions? ;-) What about Bayesian stochastic variable selection?