What is the probability that two individuals have the same parent?

The coalescent

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

The coalescent

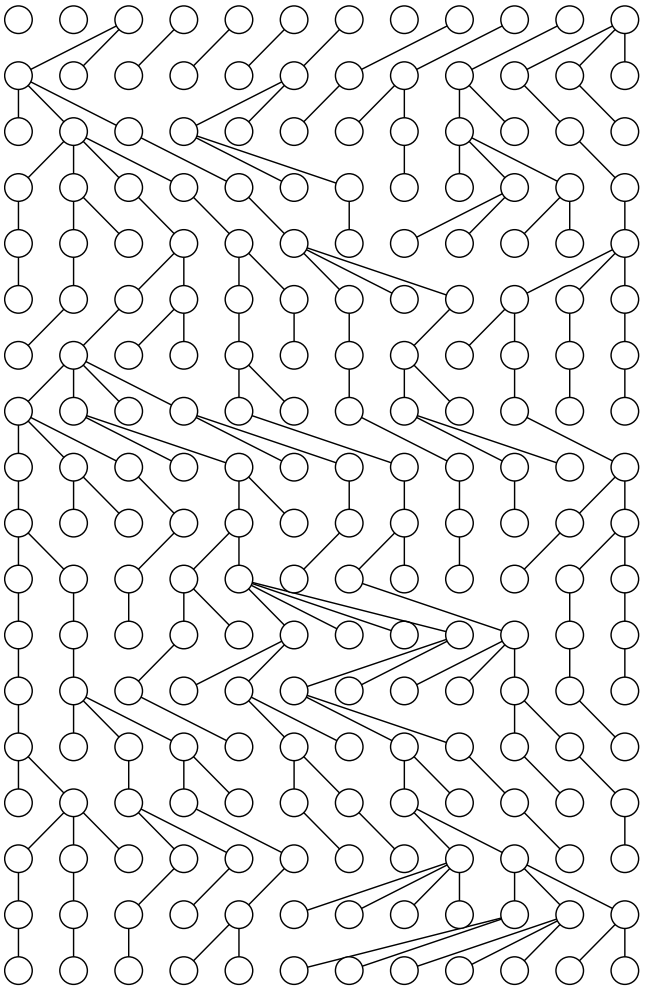
- is a model of the ancestral relationships of a sample of individuals taken from a larger population.
- describes a probability distribution on ancestral genealogies (trees) given a population history, $N(t)$.
  - Therefore the coalescent can convert information from ancestral genealogies into information 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

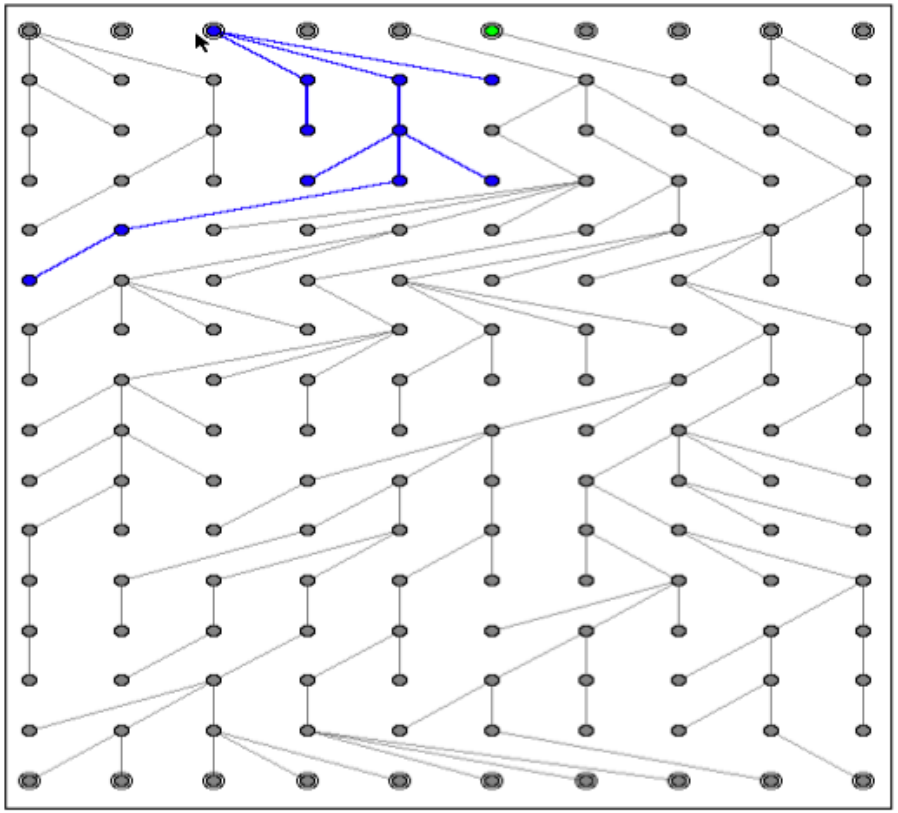
Most of theoretical population genetics is based on the idealized Wright-Fisher model of population which assumes

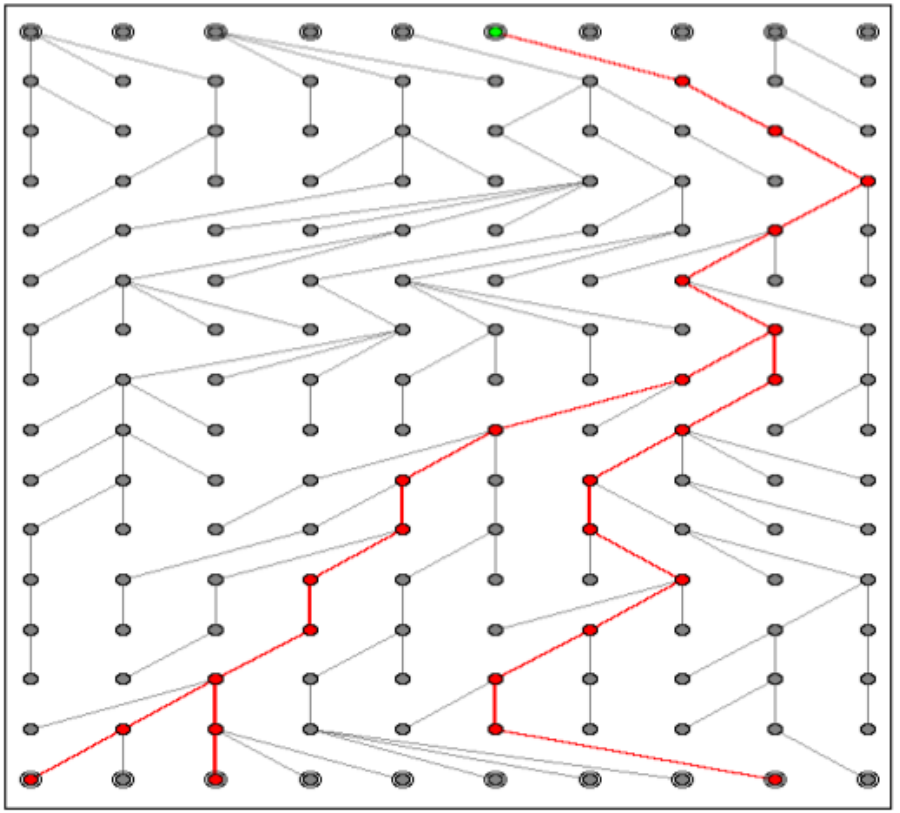
- Constant population size N
- Discrete generations
- Complete mixing

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



Genetic drift: extinction and ancestry



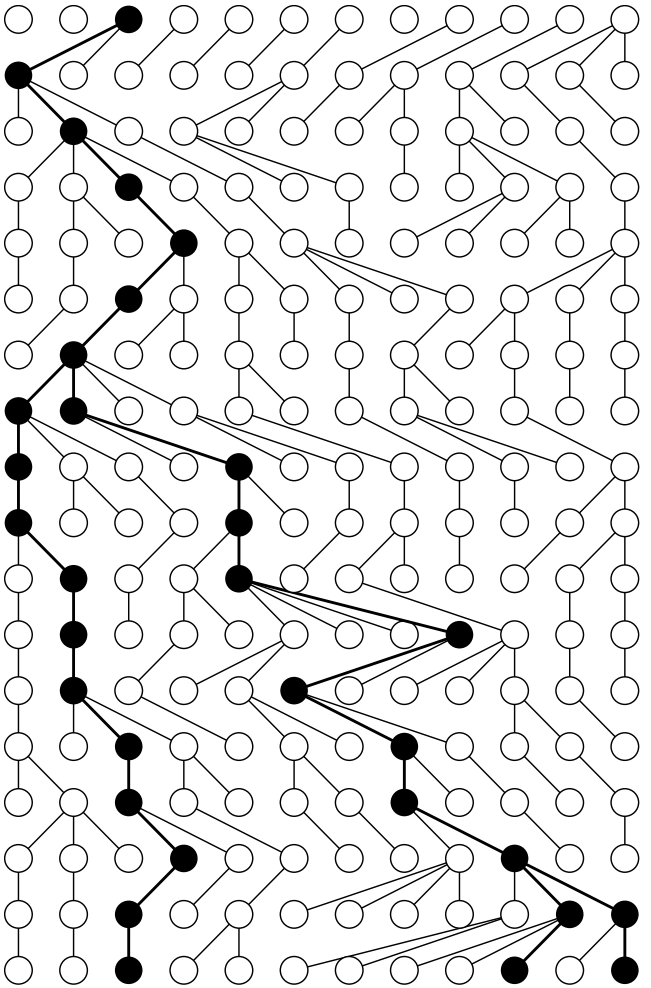


- If you trace the ancestry of a sample of individuals back in time you inevitably reach **a single most recent common ancestor**.
- If you pick a random individual and trace their descendents forward in time, all the descendents of that individual will **with high probability** eventually die out.

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. $N \gg k^2$.

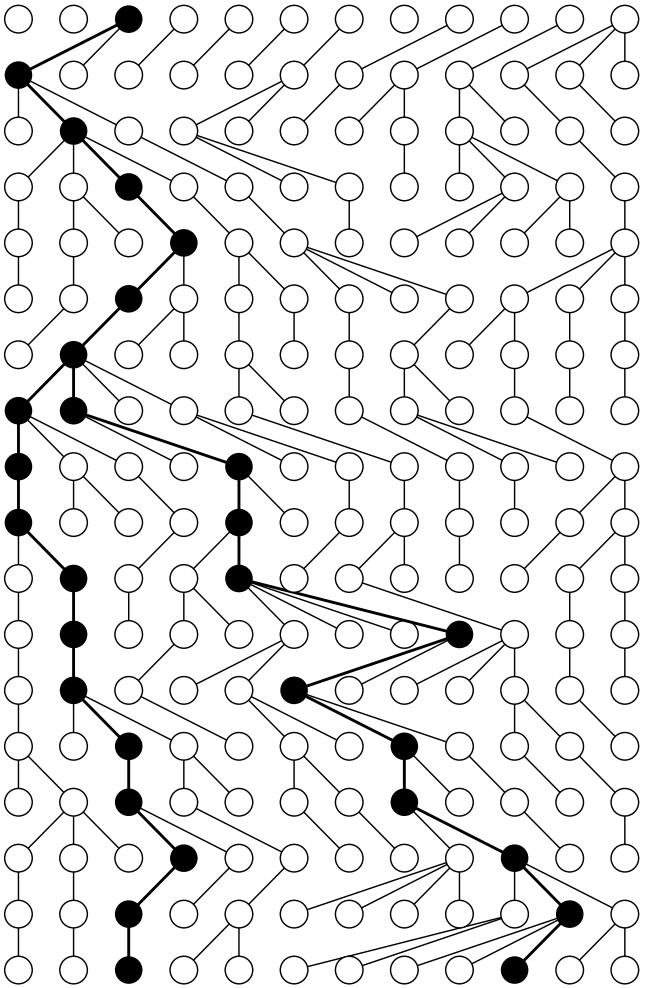


The coalescence of two ancestral lineages

- 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 $N^3/(N-1)$.

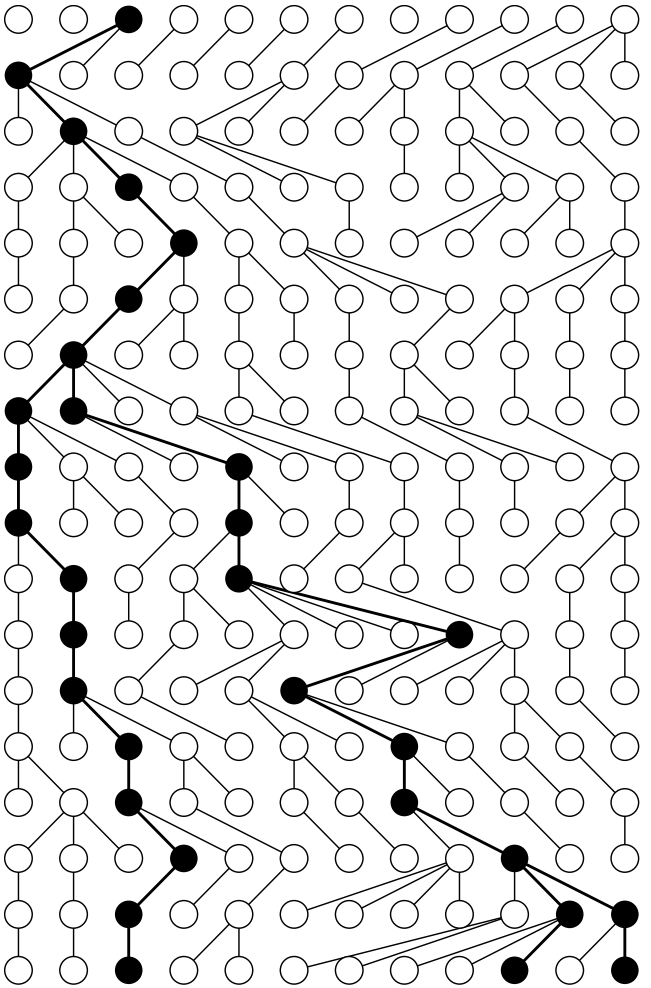


The coalescence of k lineages

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

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

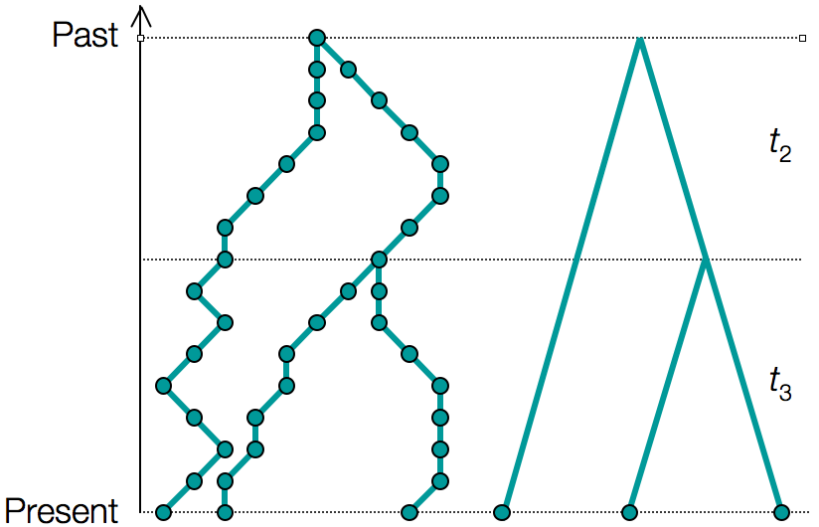
This implicitly assumes that N is much larger than $O(k^2)$, so that the probability of two coalescent events in the same generation is small.



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:

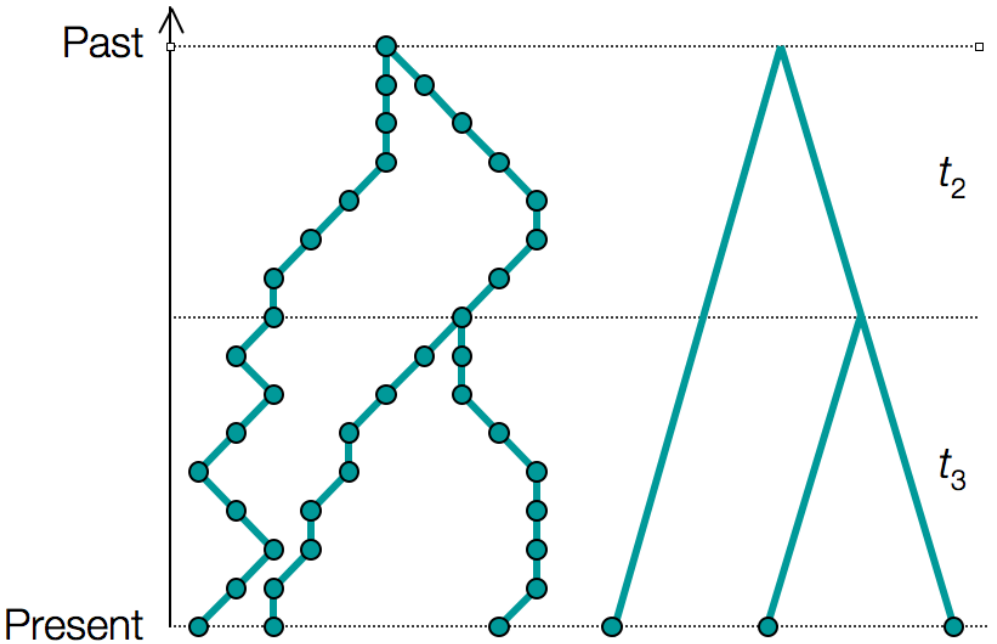


$$f(t_k) = \frac{\binom{k}{2}}{N} \exp\left(-\frac{\binom{k}{2} t_k}{N}\right) .$$

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:

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

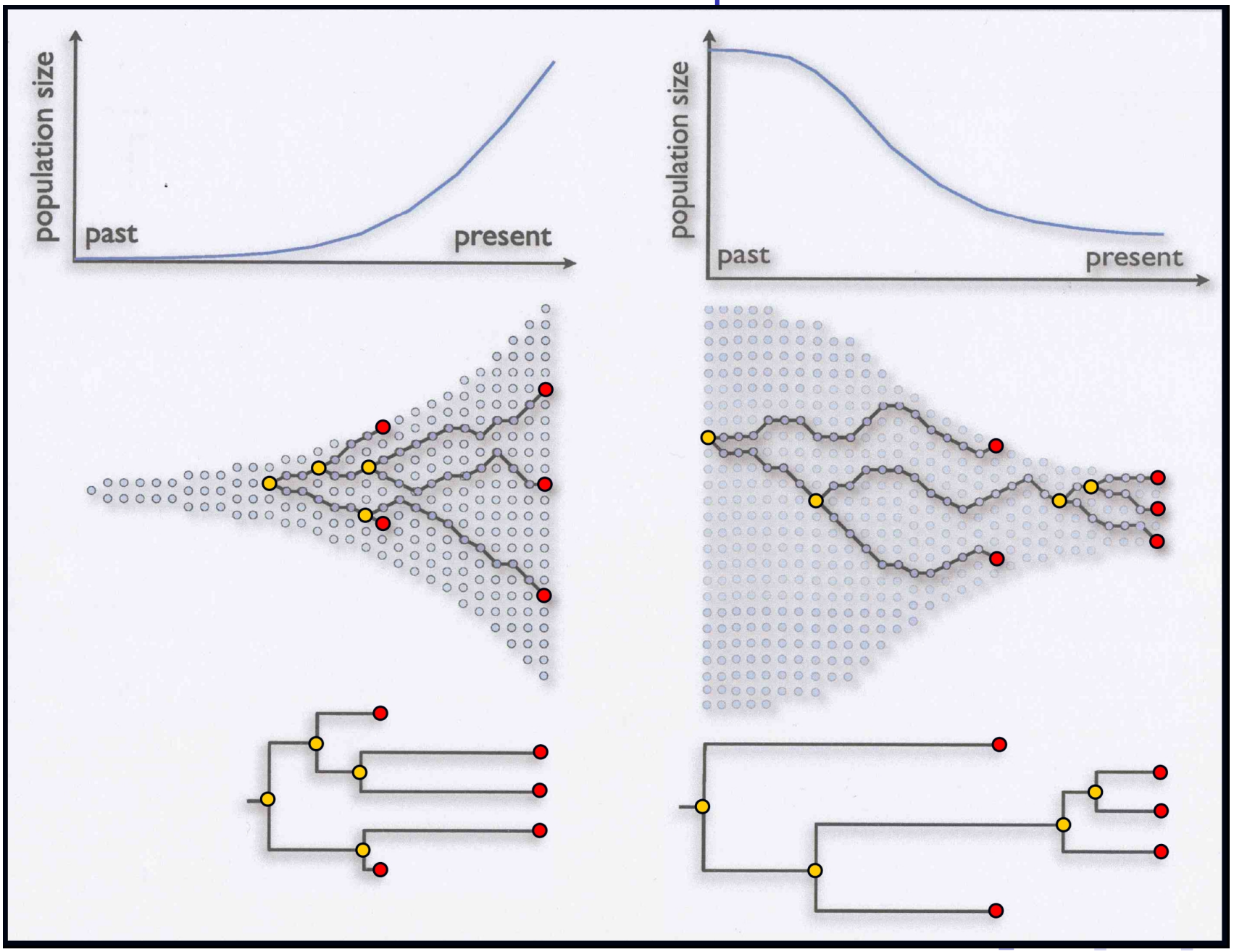


The coalescent density with varying population size

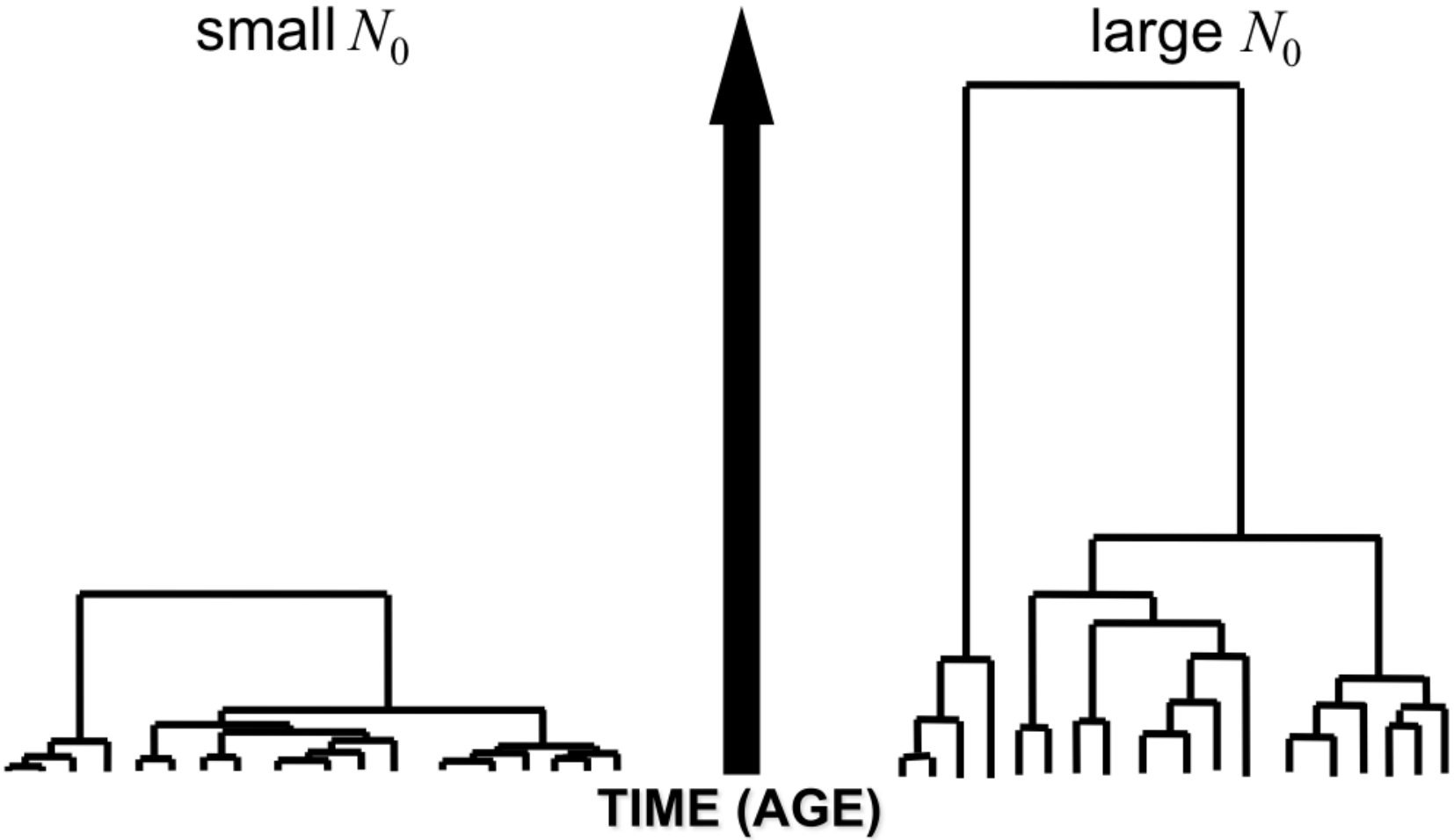
The generalization of the coalescent for the case where the population size changes over time, $N = N(t)$ is given by Griffiths and Tavare (1994). They showed that the coalescent density for the first coalescence event being at time t in the past given n lineages is:

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

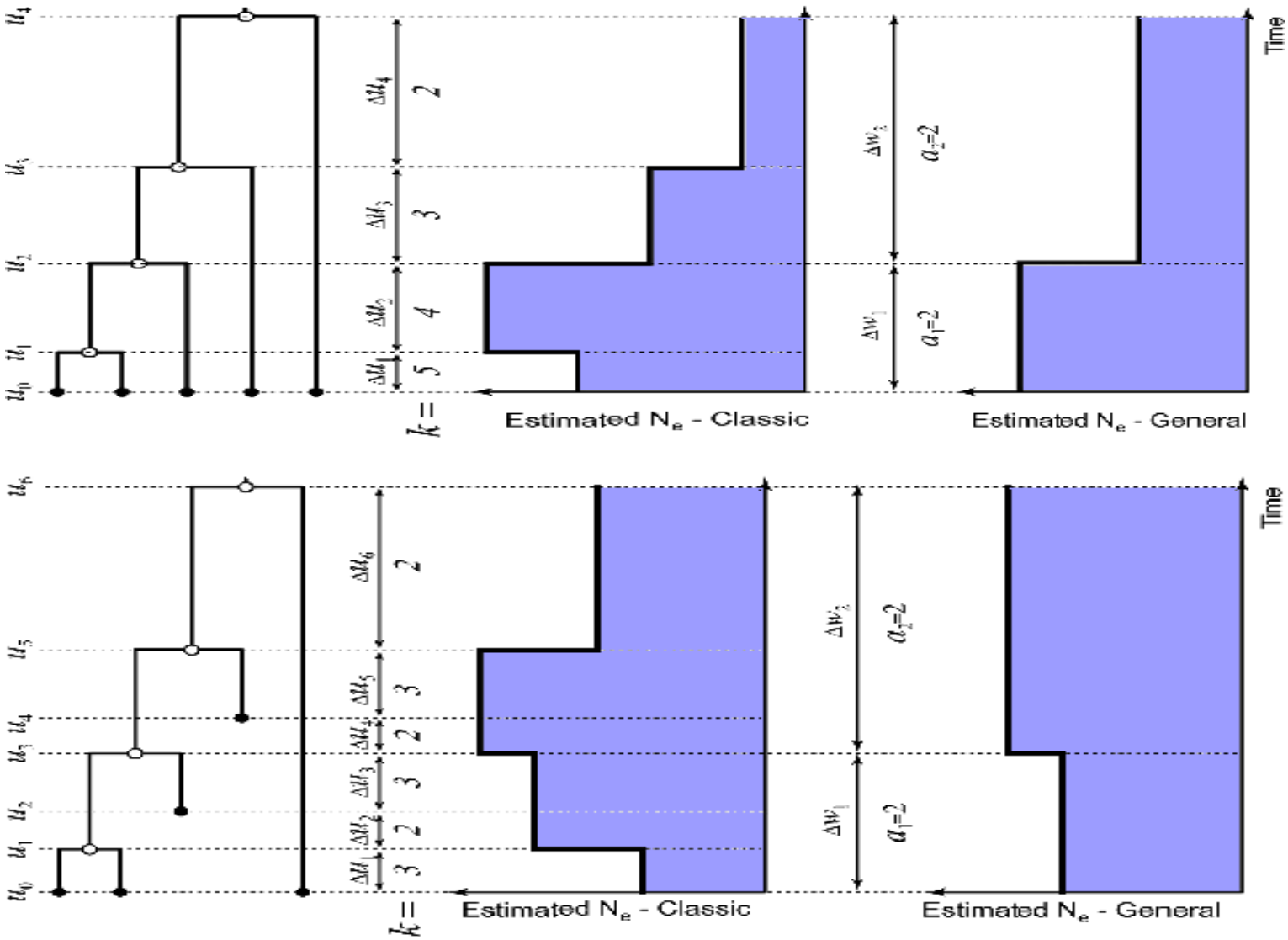
The coalescent with serial samples



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

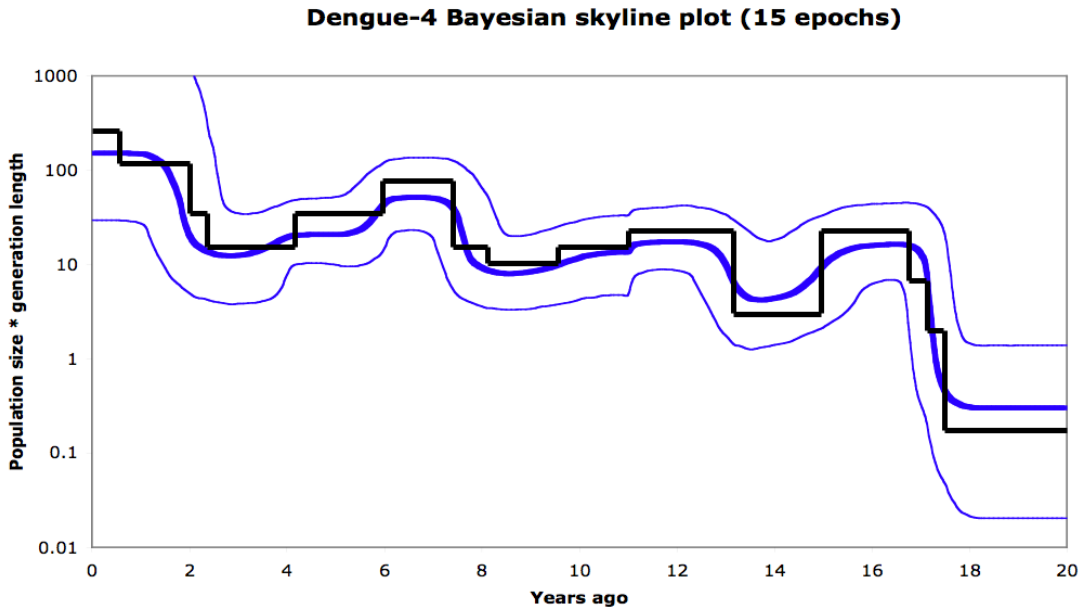


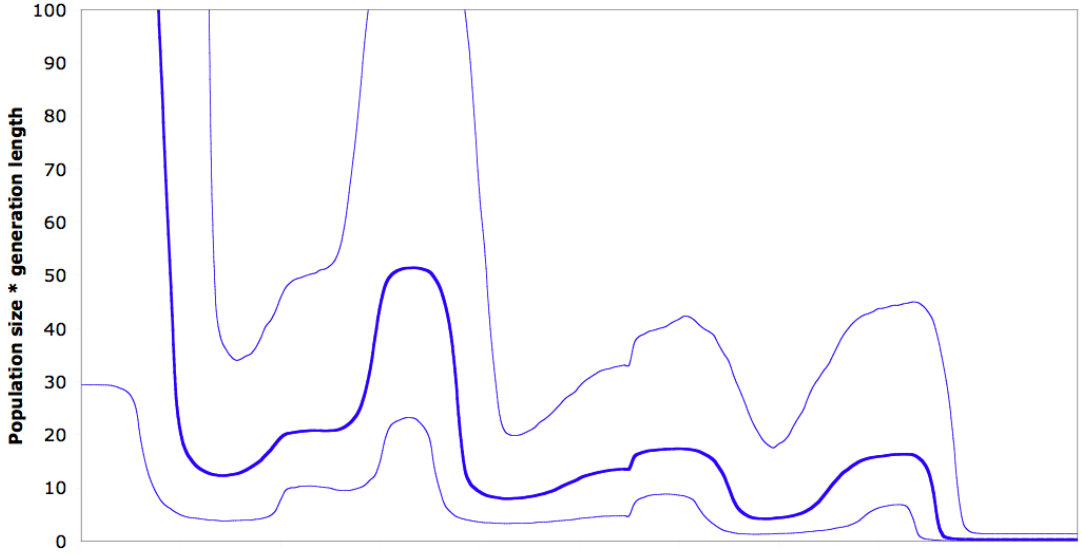
Bayesian skyline



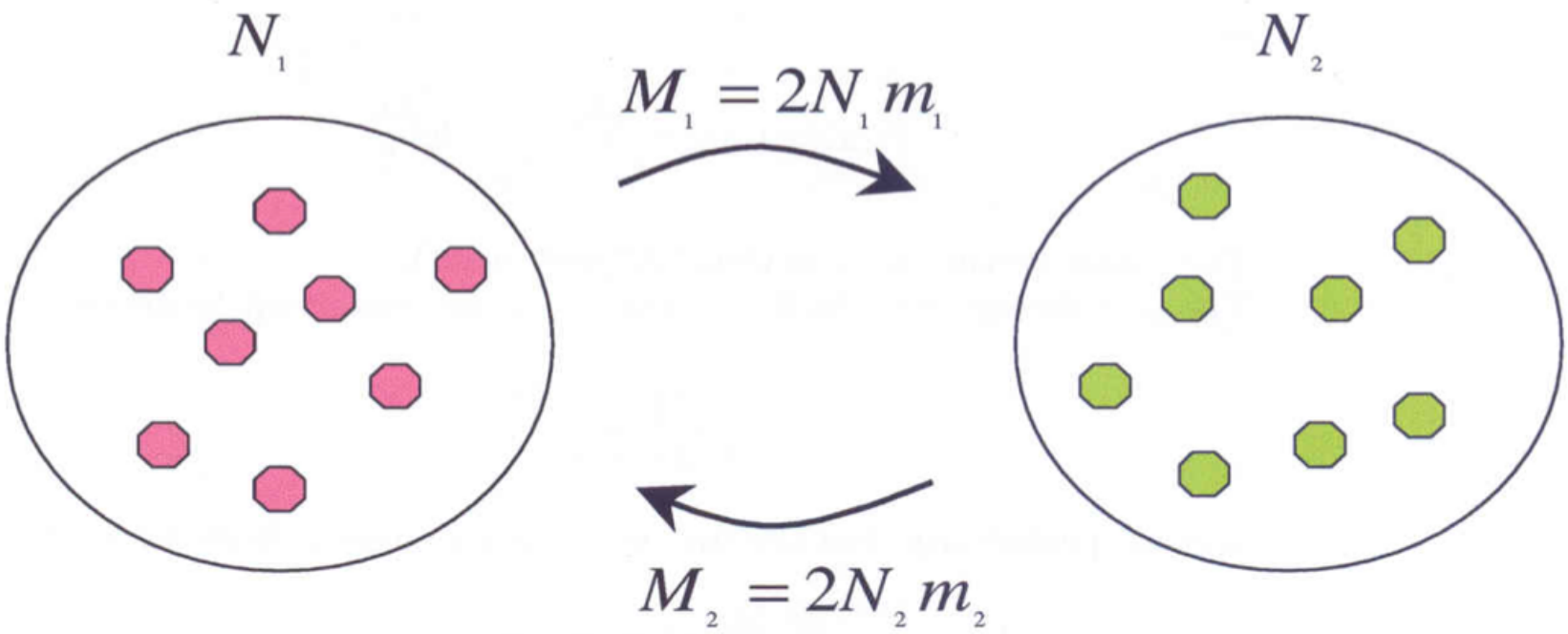
The Bayesian skyline plot (Drummond *et al*, 2005)

The Bayesian skyline plot estimates a demographic function that has a certain fixed number of steps (in this example 15) and then integrates over all possible positions of the break points, and population sizes within each epoch.

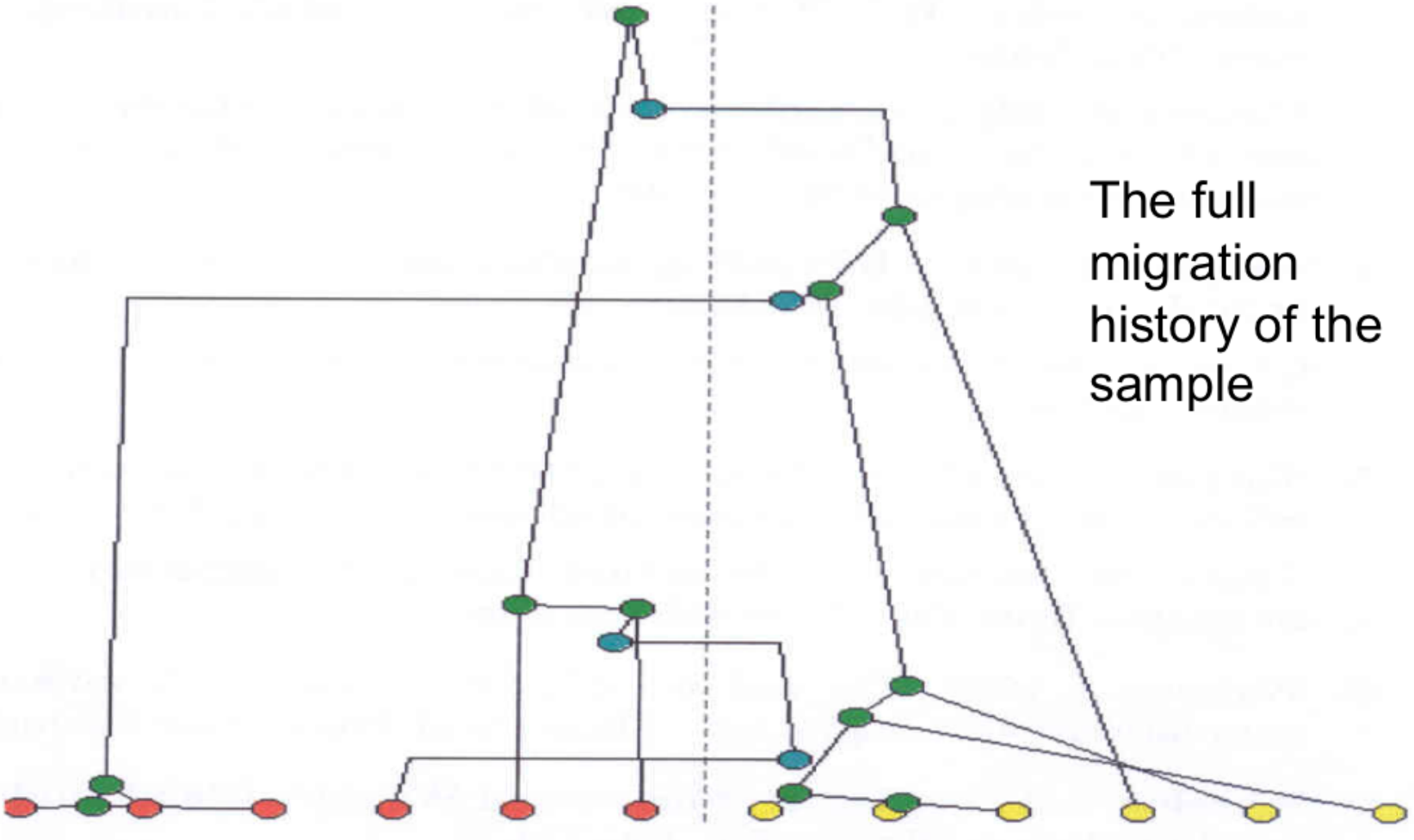




Coalescent with population structure

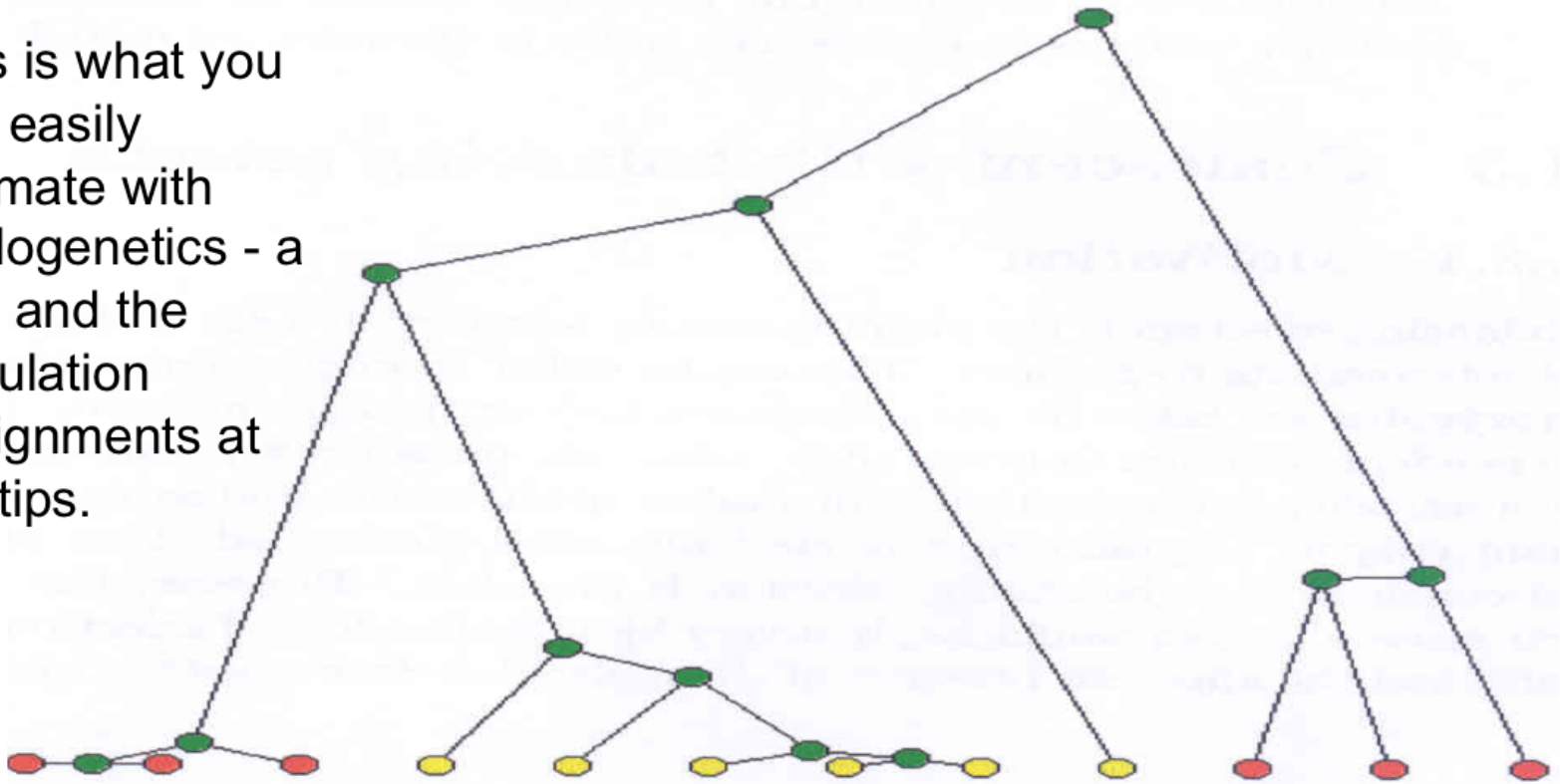


Population subdivision - two demes



Population subdivision - two demes

This is what you can easily estimate with phylogenetics - a tree and the population assignments at the tips.



Summary

- Coalescent theory provides access to information about underlying population parameters from a small sample
- Big populations produce big sample trees
- Exponentially growing populations produce star-like sample trees
- For transmission chains, the coalescent can describe the relationship between the viral gene tree and the transmission tree.
- Selection, recombination, geography and population structure can all be theoretically incorporated into a coalescent analysis and progress is being made on all these fronts.

Aphelocoma (scrub-jays and relatives)

Aphelocoma (scrub-jays and relatives)

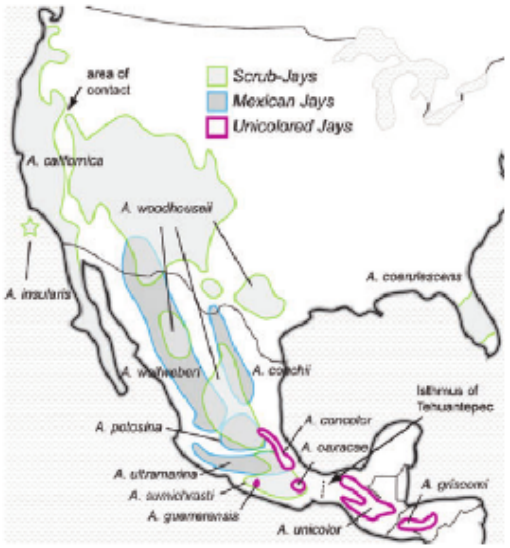
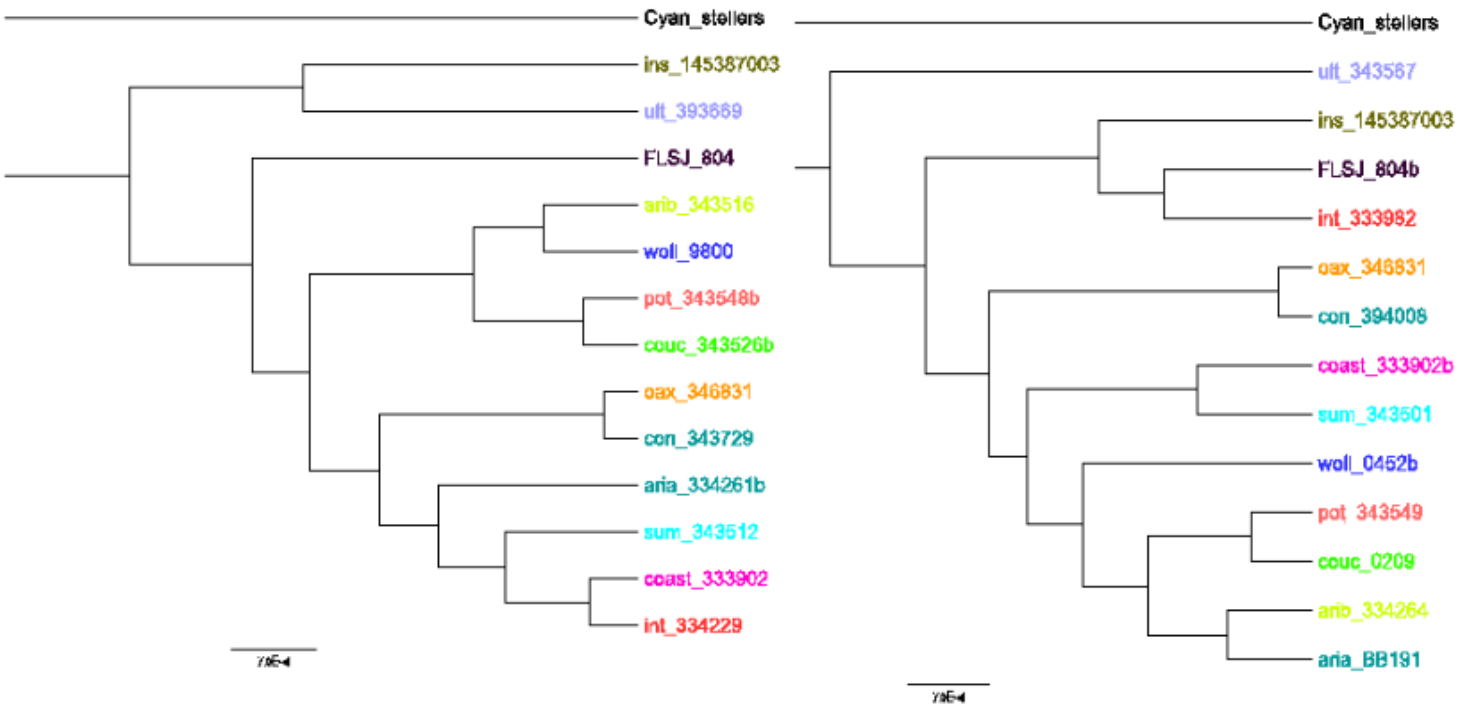


Figure 1. Geographic distributions of *Aphelocoma* jays. The three major species groups (i.e., Mexican Jays, Scrub-Jays, and Unicolored Jays) are indicated with colored outlines and all the species lineages discussed in text are labeled. See Methods for details about species designations and Appendix 1 for comparison with existing taxonomy.

Tree on 1 gene, rejected by expert

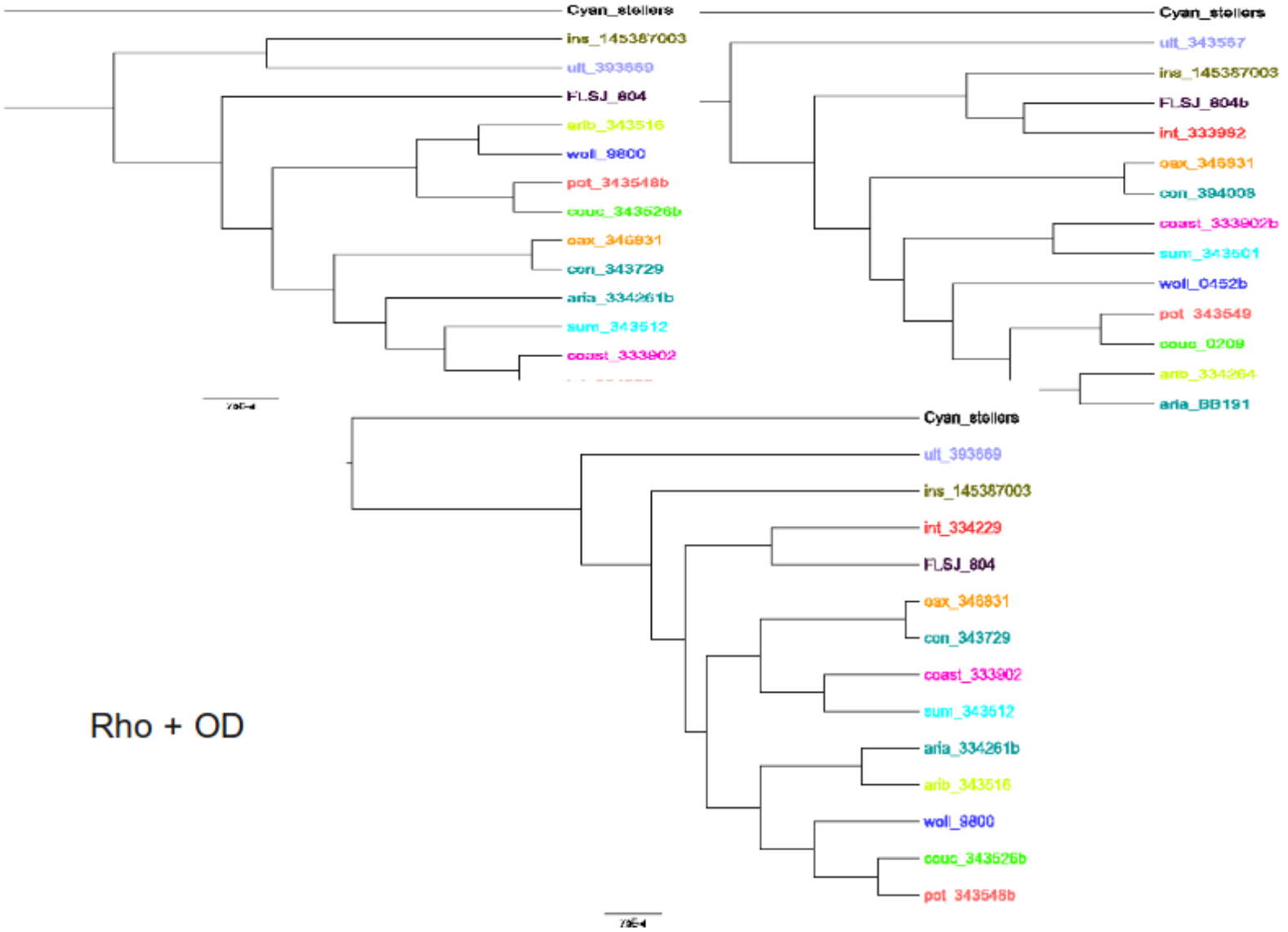
Aphelocoma (scrub-jays and relatives)



OD Rho

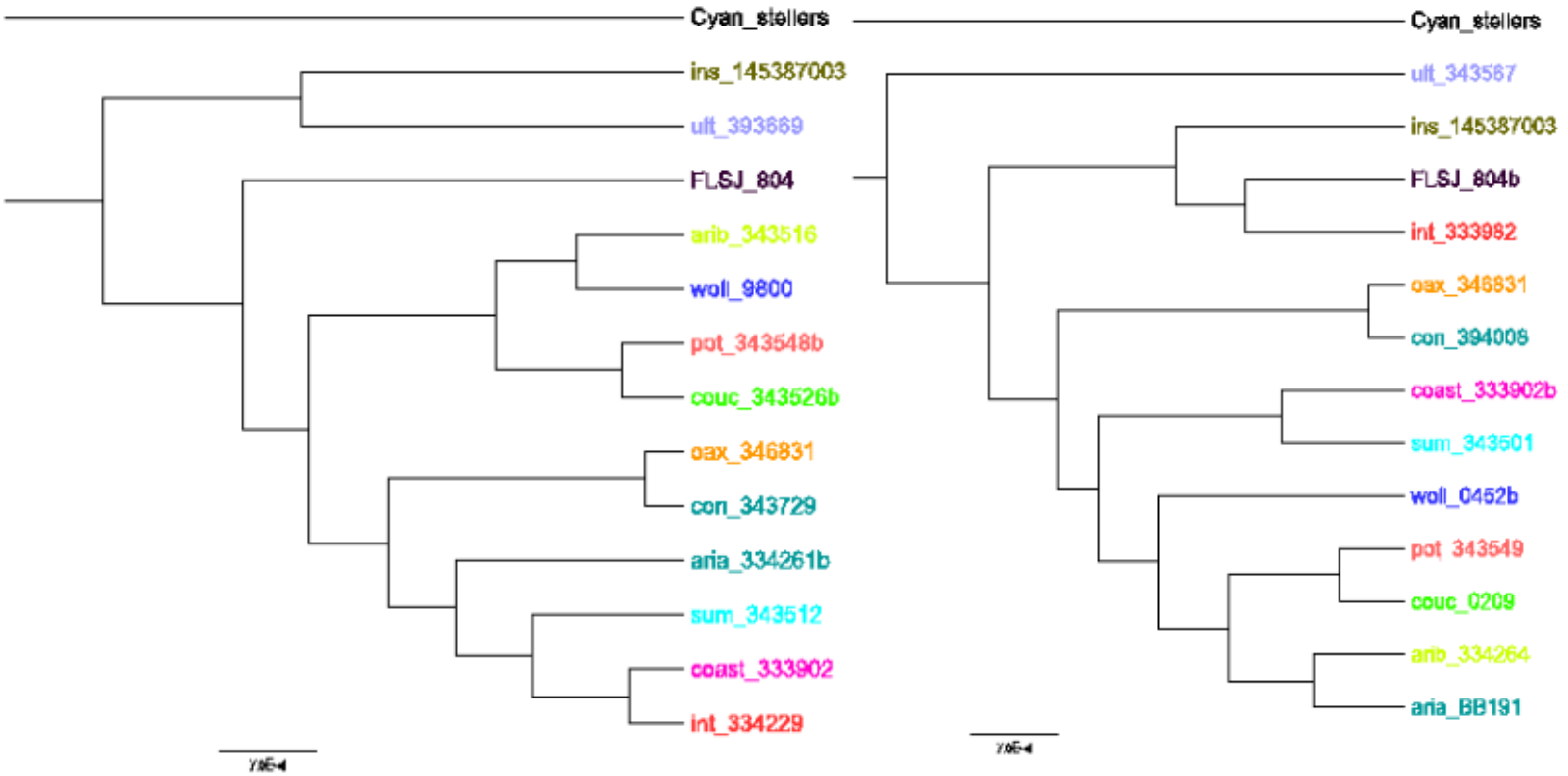
Different genes, different trees

Aphelocoma (scrub-jays and relatives)



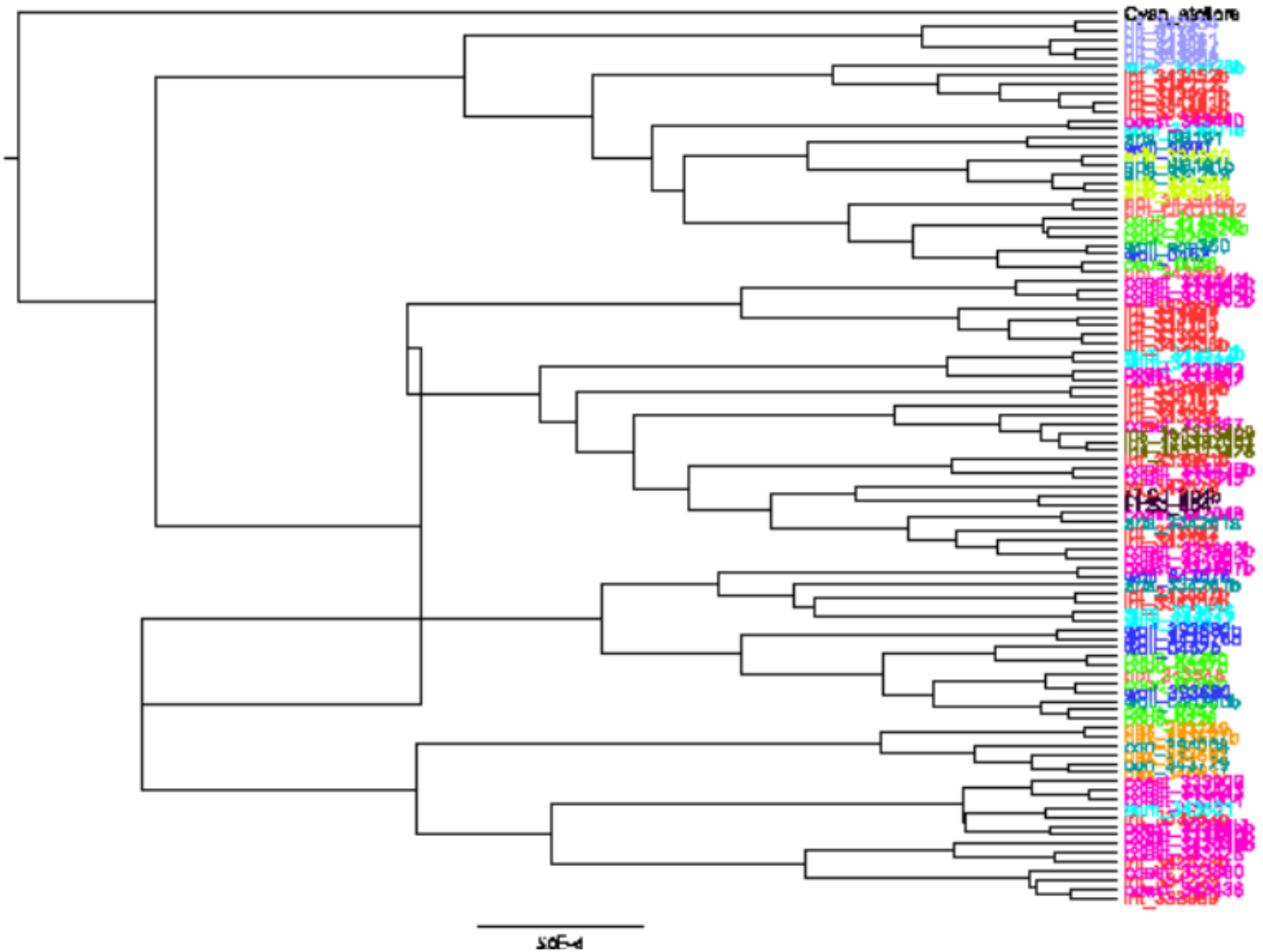
Appending genes, different trees

Aphelocoma (scrub-jays and relatives)



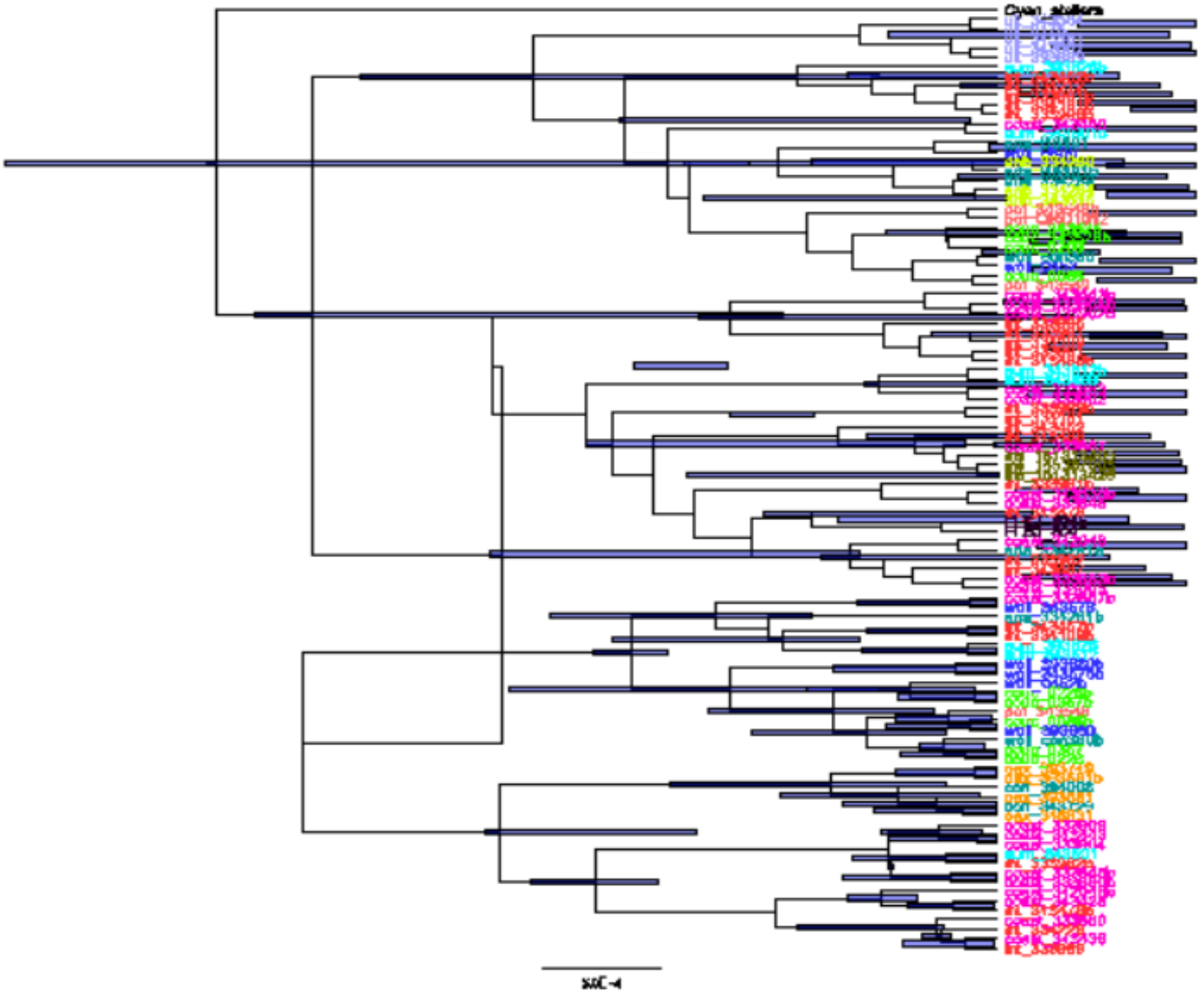
Different indiviudals, different trees

Aphelocoma (scrub-jays and relatives)



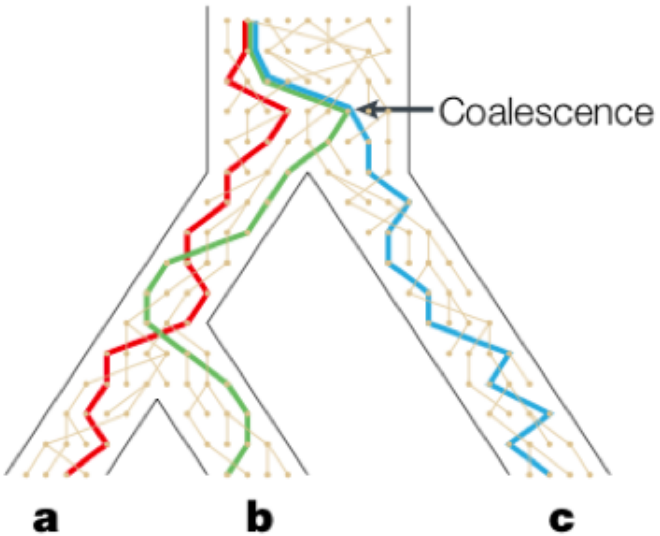
Combining all data: note negative tree lengths

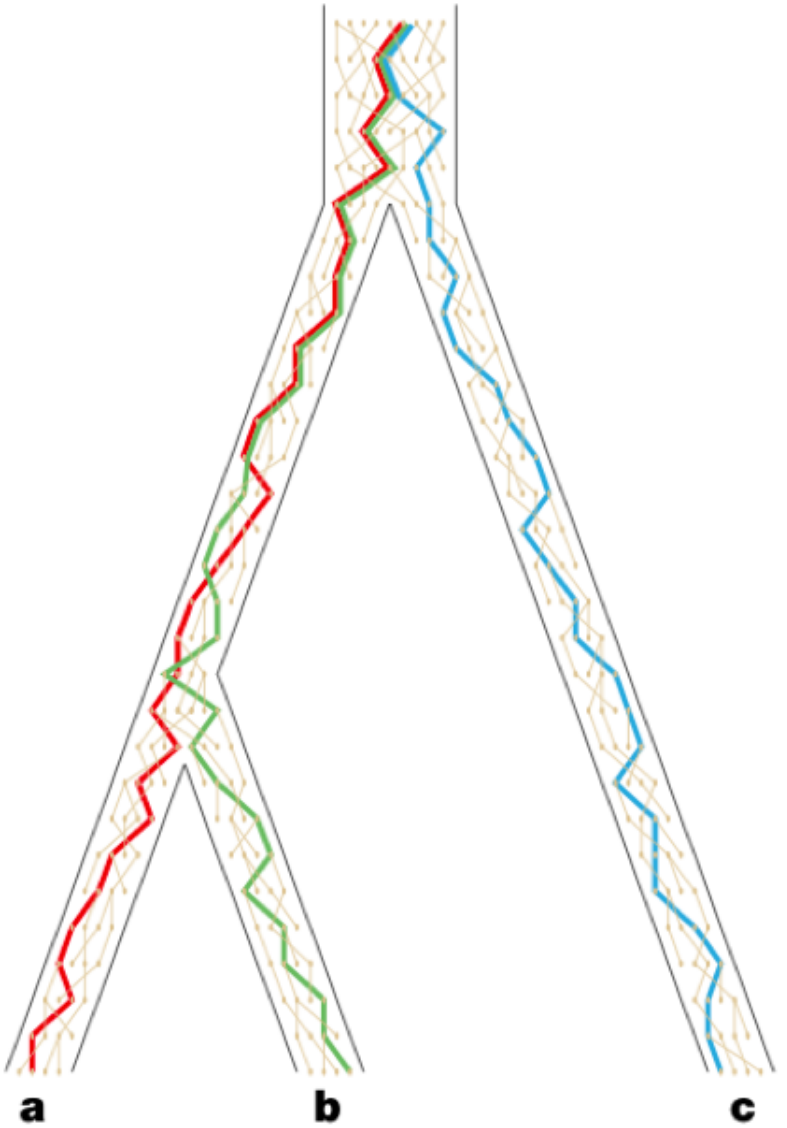
Aphelocoma (scrub-jays and relatives)



Combining all data: note massive uncertainty

Gene trees and species trees





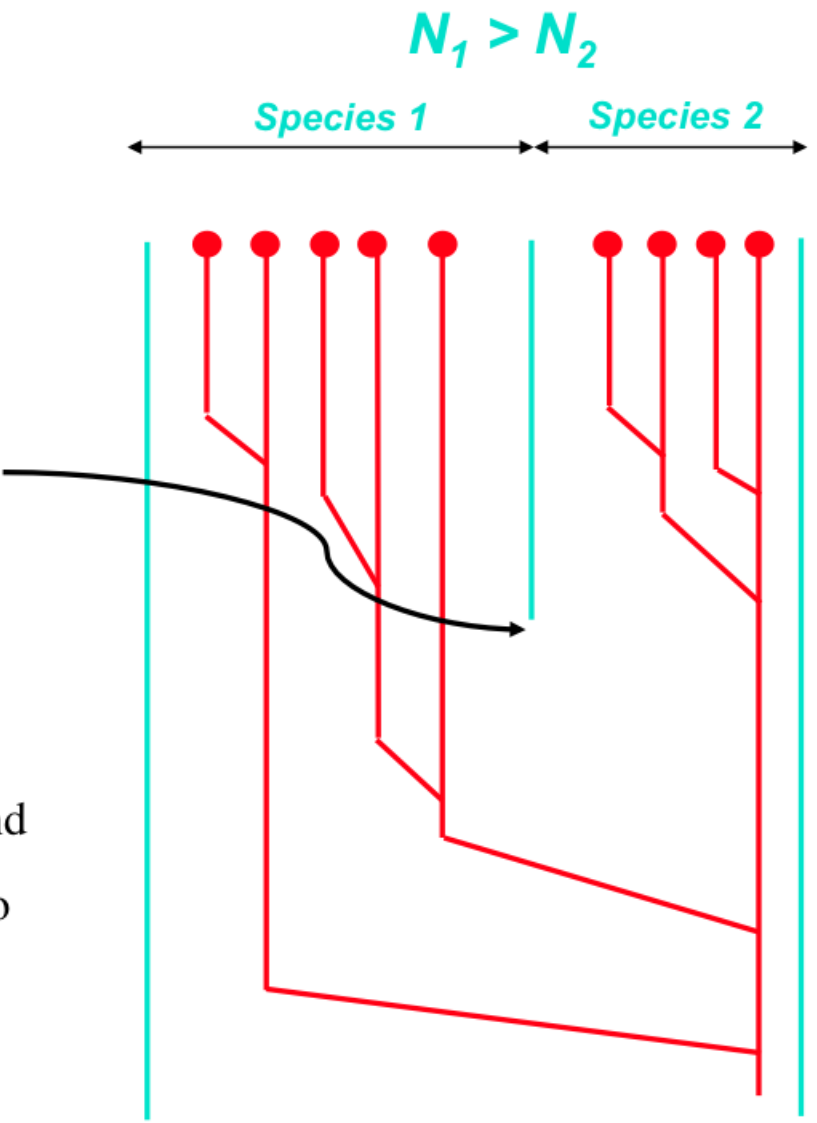
The murky boundary between population genetics and phylogenetics

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
- 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.

Incomplete Lineage Sorting

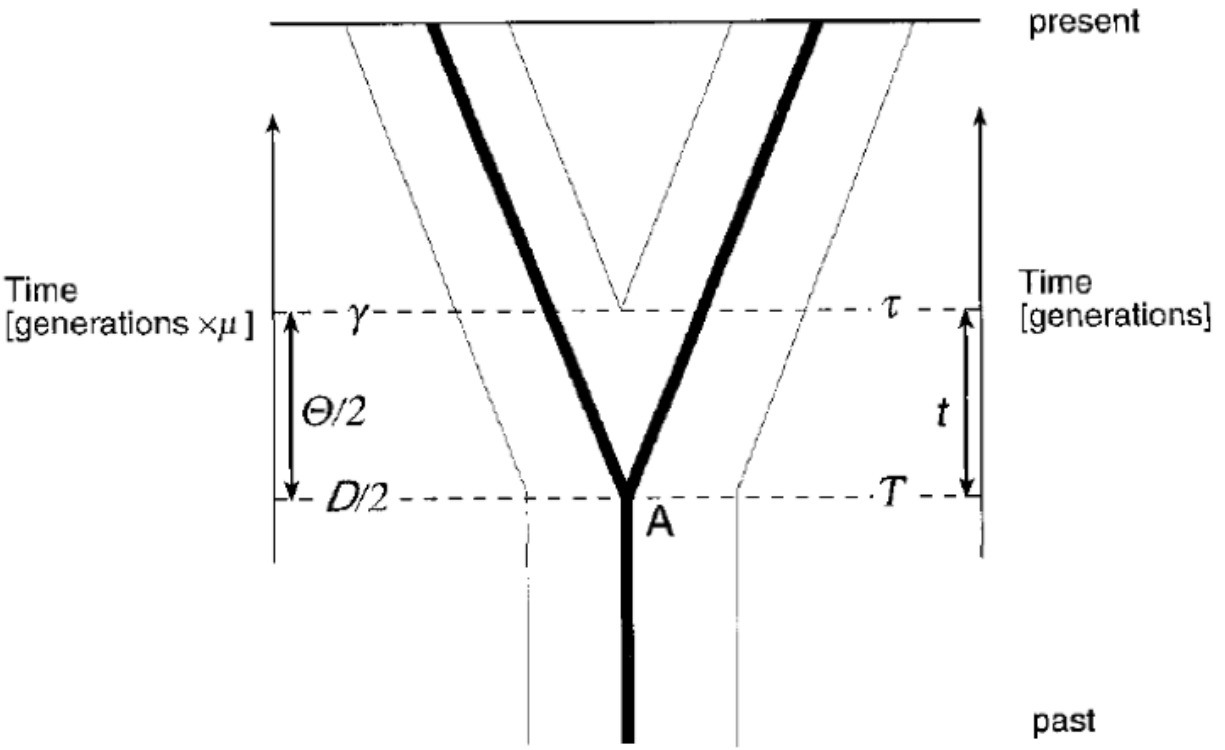


- The probabilities of seeing reciprocal monophyly, paraphyly or polyphyly, depend on:

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

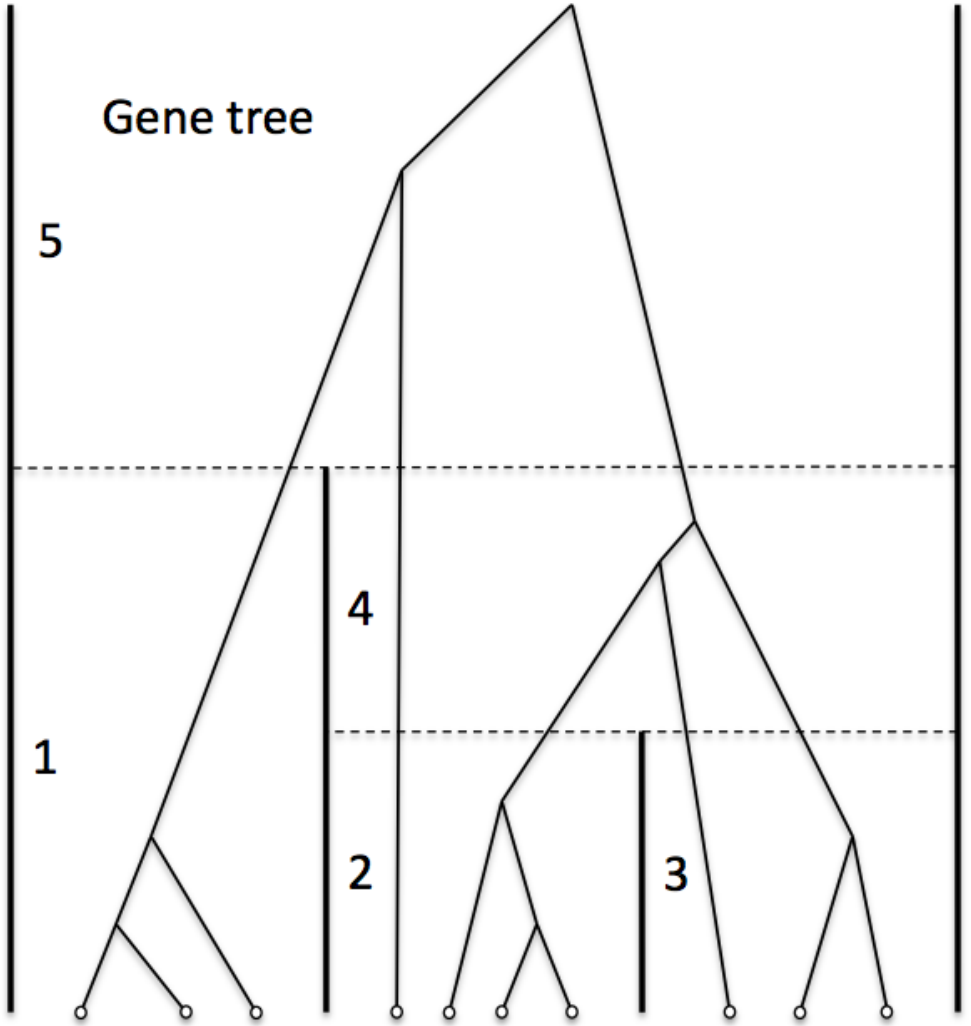
Problems with estimation of divergence times

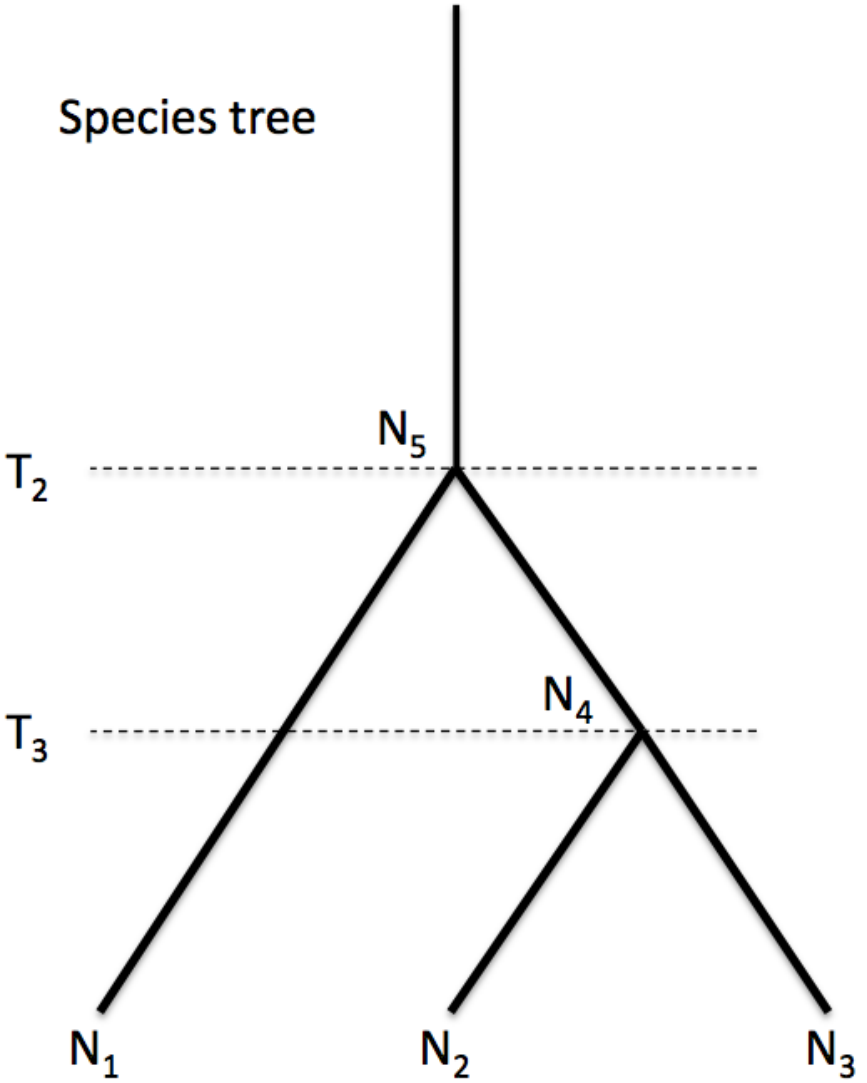
- Typically we use gene phylogenies to estimate species phylogenies
- But divergence time for genes will be longer than that of species.



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

The multispecies coalescent



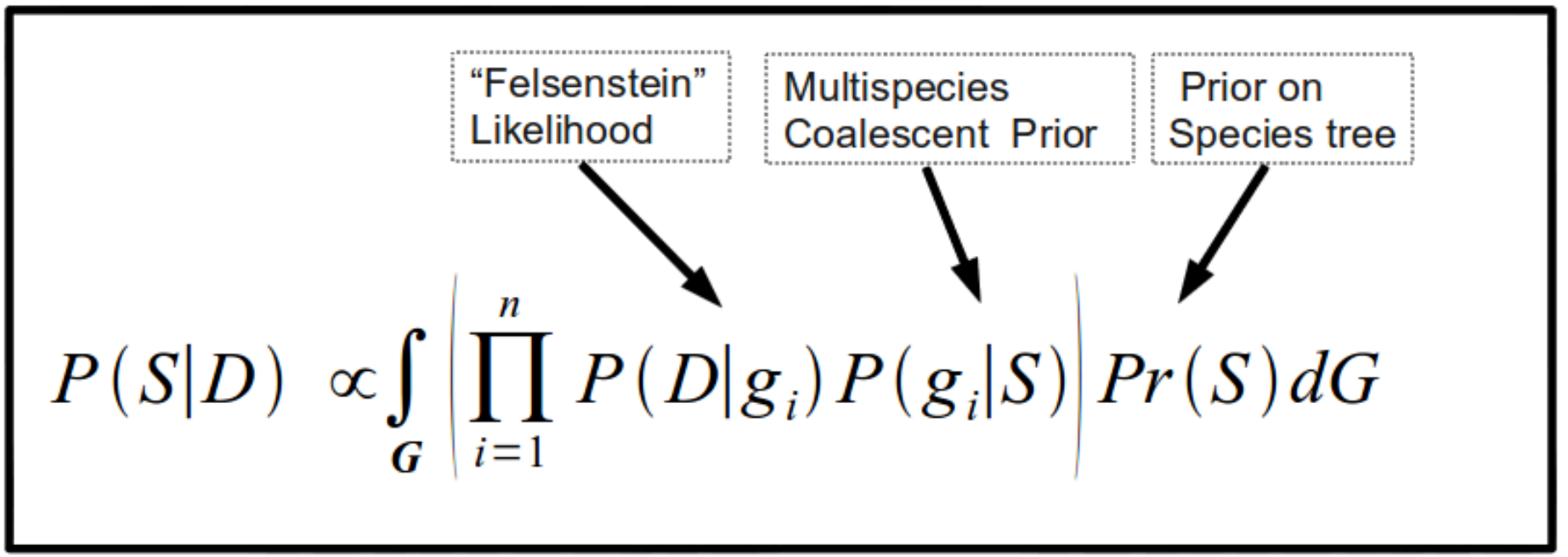


*BEAST

- 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

*BEAST is a Bayesian model – a probability is defined for each combination of gene trees, species tree and population sizes.



Species tree S, the data $D = d_1, d_2, \ldots, d_n$ is composed of n alignments

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.

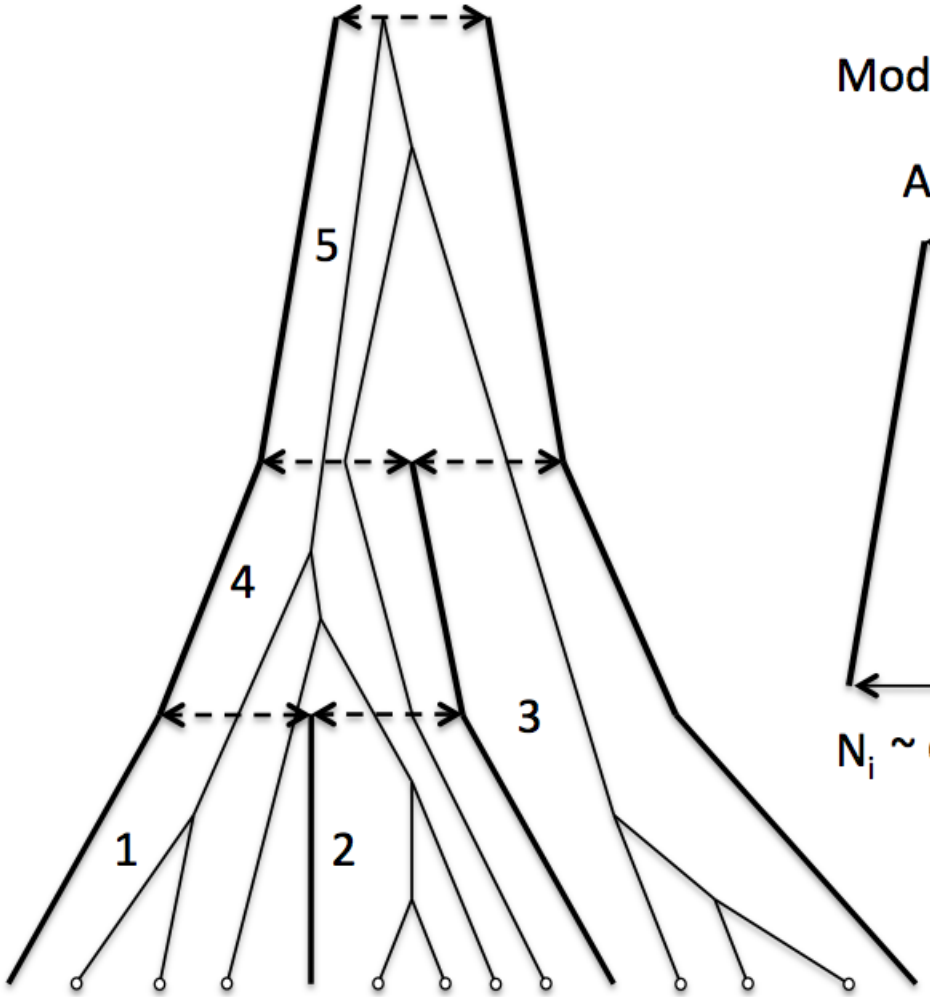
Define T to be a time tree with ranked topology and divergence times.

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

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

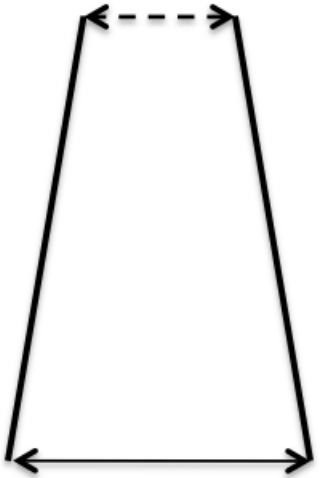
$P(S) = P(A, N|T)P(T)$

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



Modern species

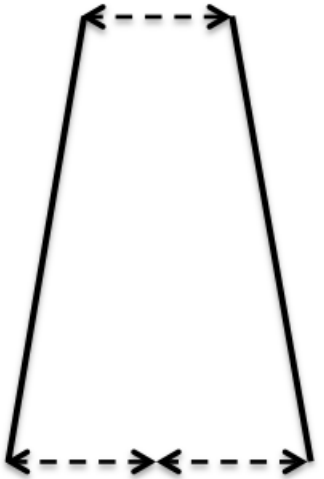
$A_i \sim Exp(\Theta)$



$N_i \sim Gamma(2,\Theta)$

Ancestral species

$A_i \sim Exp(\Theta)$



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

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

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:

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

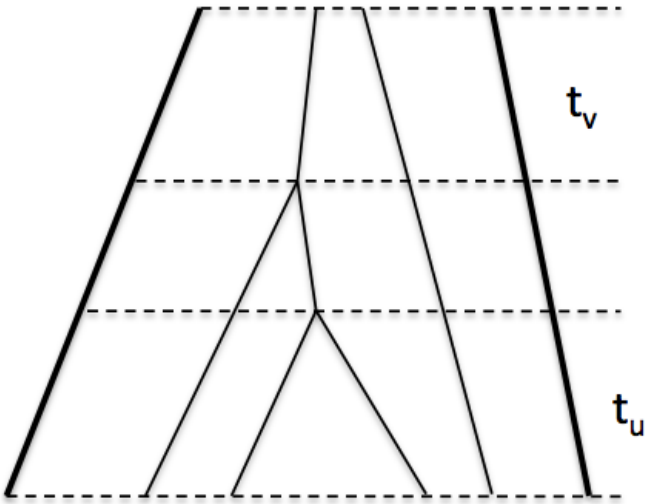
More complex species tree priors that admit species extinction (Birth-death prior; Gernhard, 2008) and incomplete sampling (Birth-death-sampling; Stadler, 2009) are also possible.

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

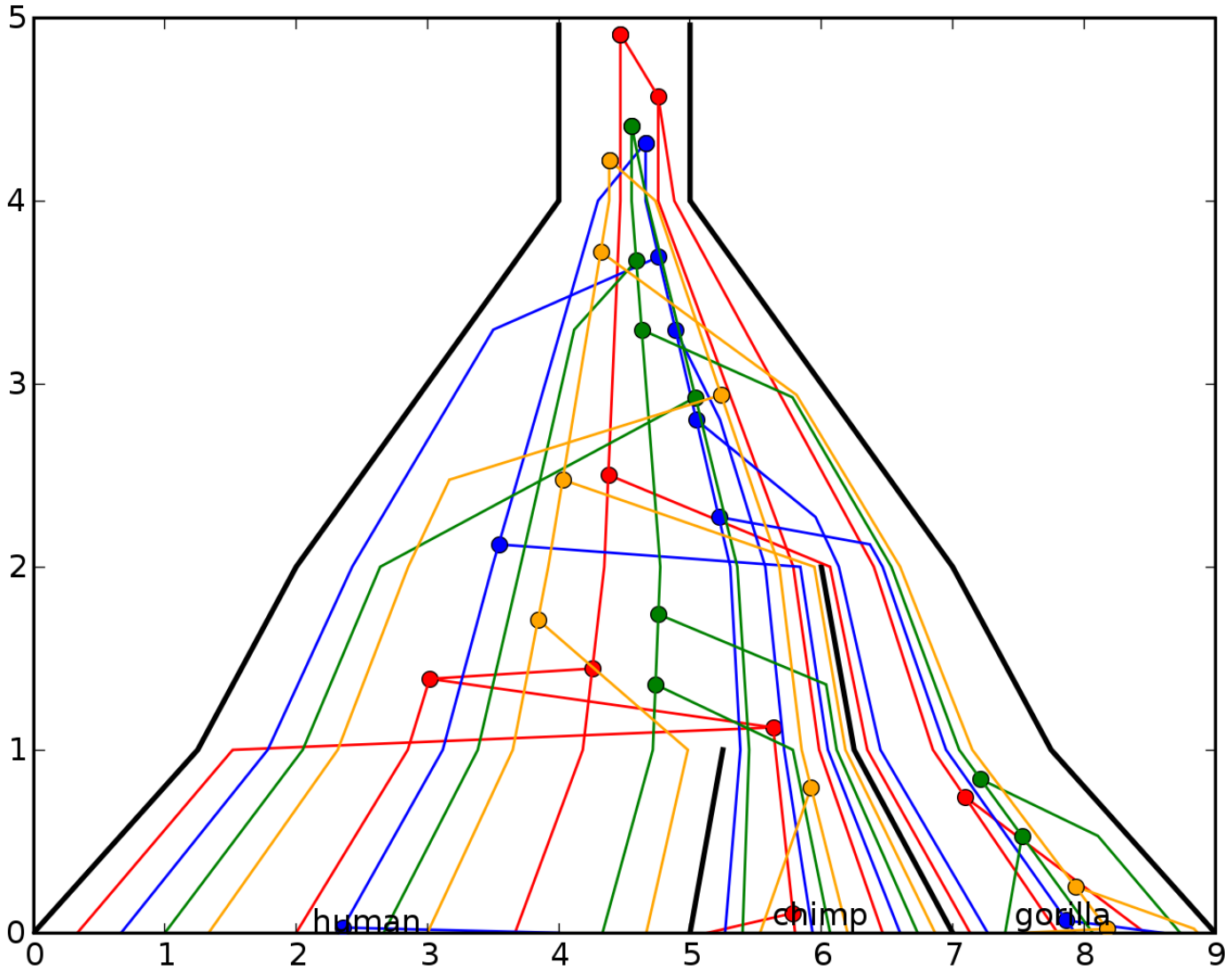
Coalescent prior for gene trees: $P(g_i|S)$

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:

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



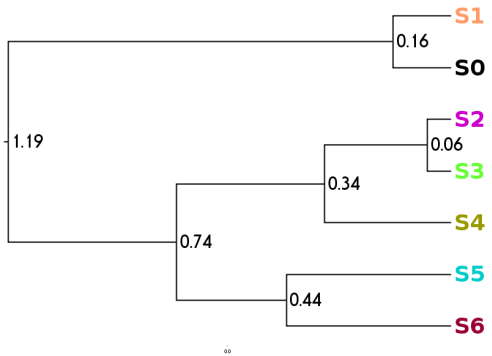
Four gene trees inside a 3-species tree



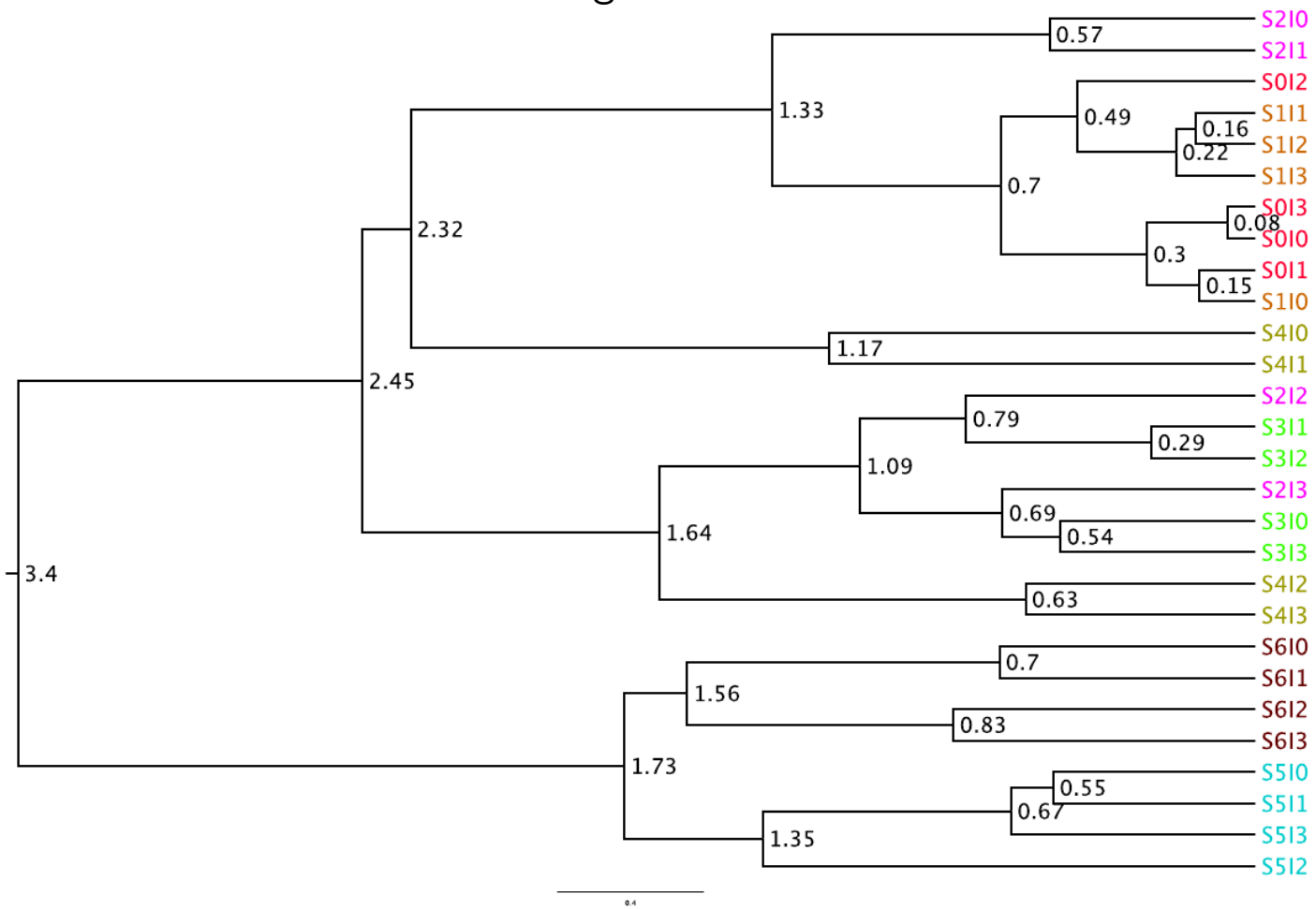
Supermatrix concatenation

a terrible idea for rapid radiations

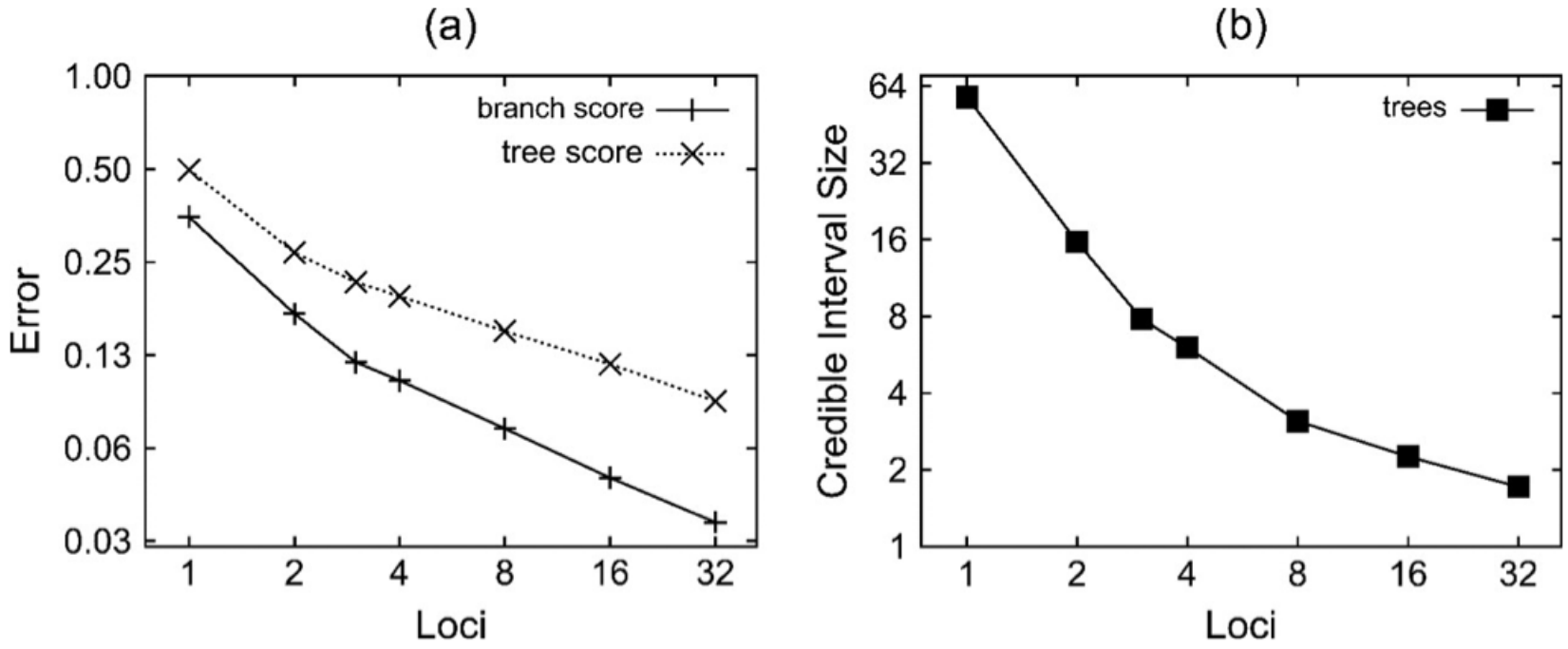
Species tree



"Gene" tree from concatenated genes



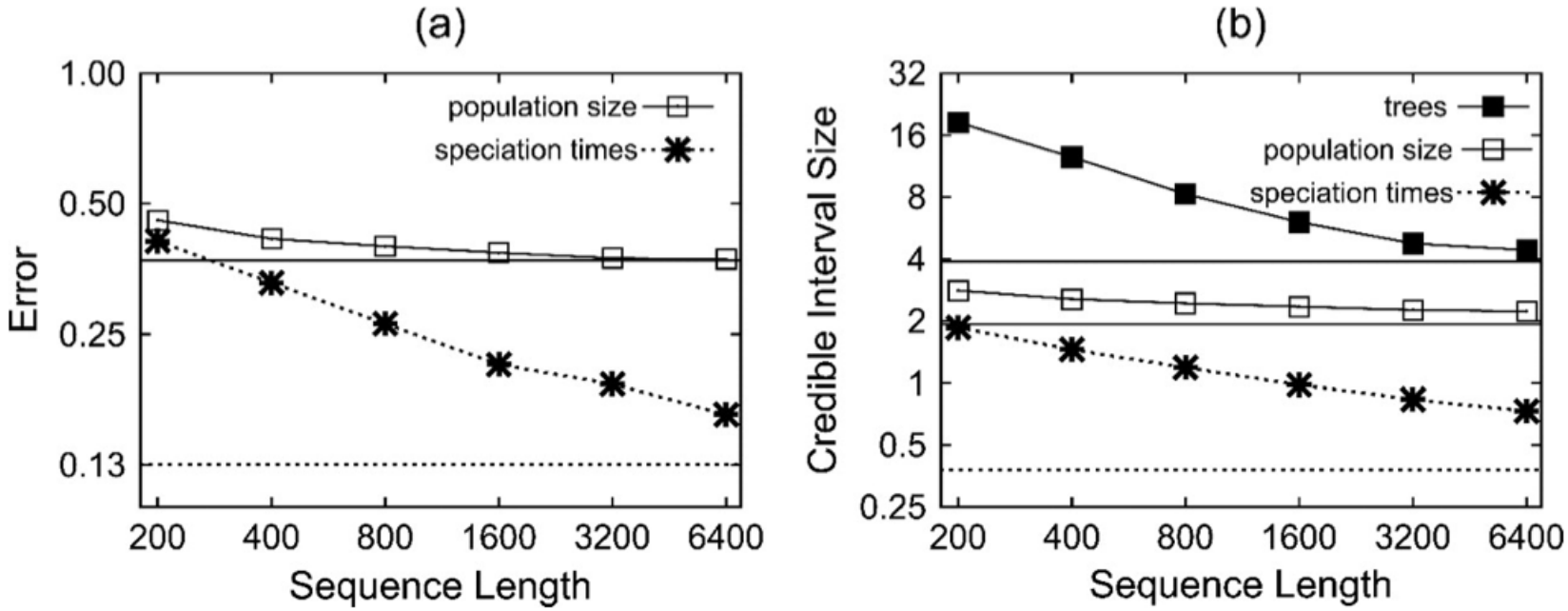
(a) Species tree estimation error and (b) 95% credible interval size as a function of the number of loci.



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

MBE MOLECULAR BIOLOGY AND EVOLUTION

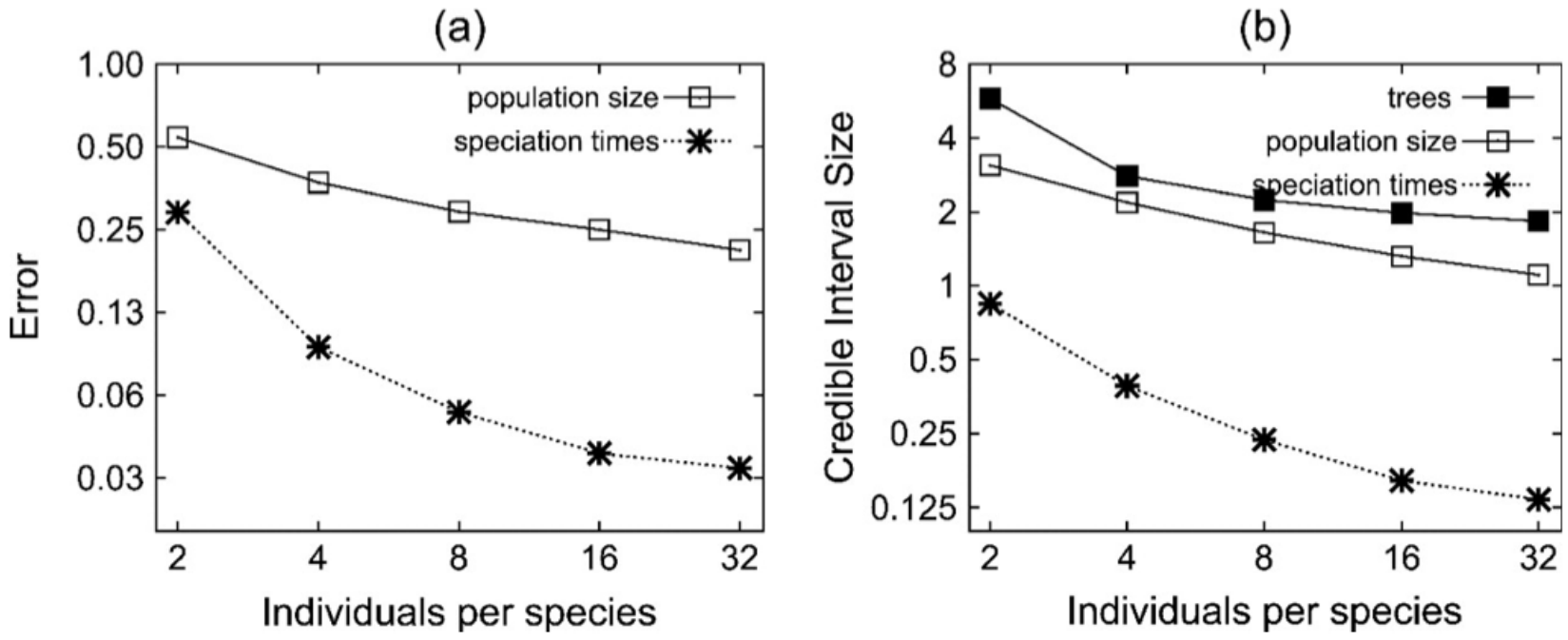
(a) Relative error and (b) credible interval sizes as a function of sequence length.



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

Integrate out population sizes on branches

Problem: with few samples per species pop sizes are hard to estimate

Math trick: If pop-size prior is inverse gamma it can be integrated out.

- useful if demographic history is just a nuisance parameter
- much faster convergence
- especially useful when poor pop size info is available (few gene trees, few samples per species)
- `STACEY/starbeast2` packages

Better tree proposals

Issue: Species tree moves slow in tree space, due to gene tree restrictions

- NNI or SPR move on species tree
- SPR moves on gene trees
- much better convergence
- `STACEY/starbeast2` packages

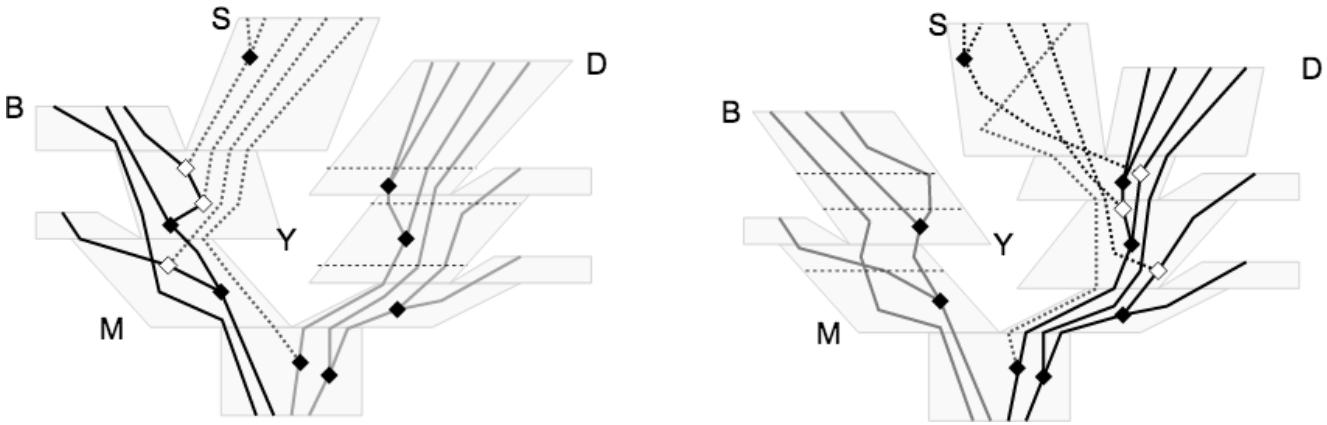
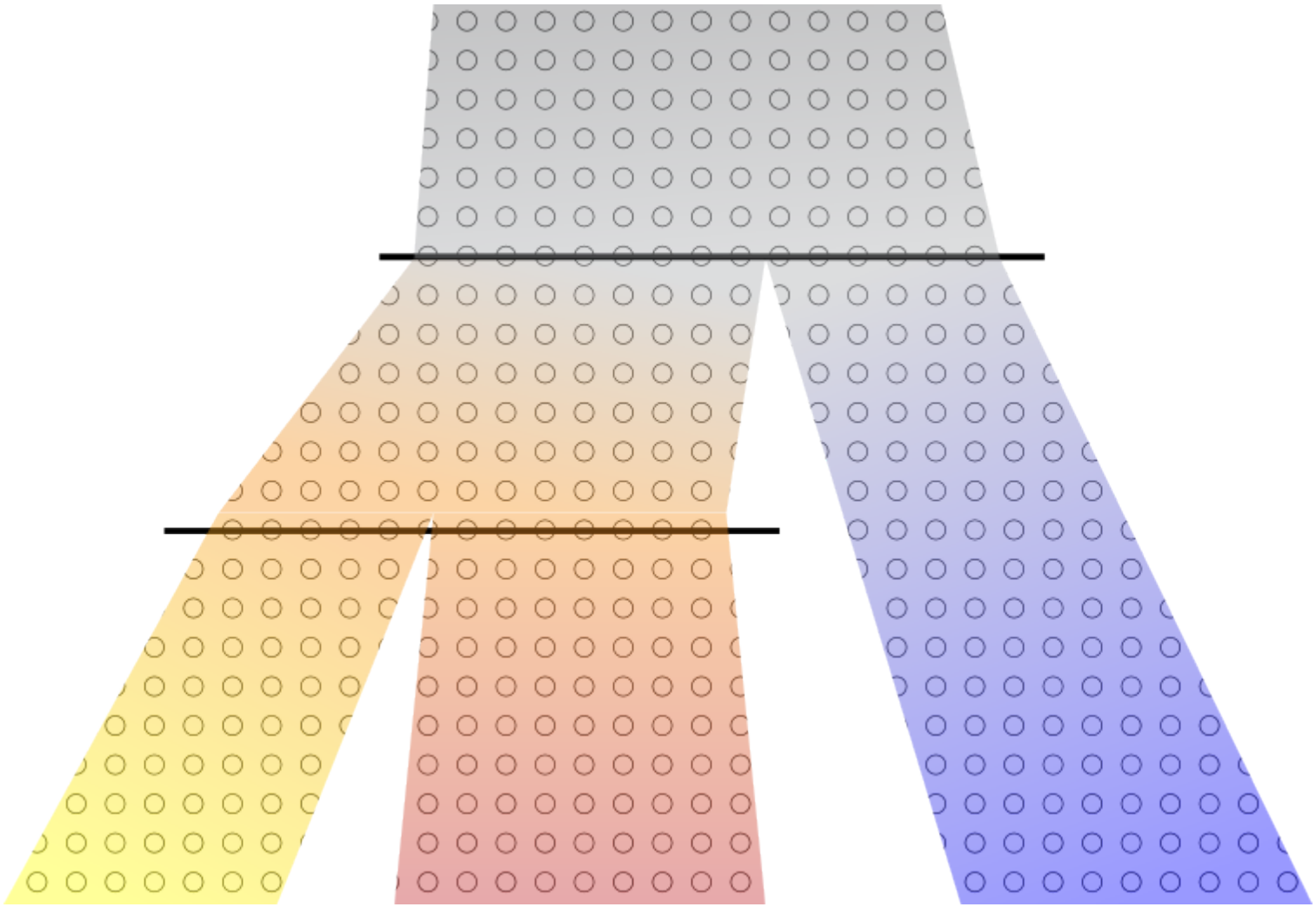
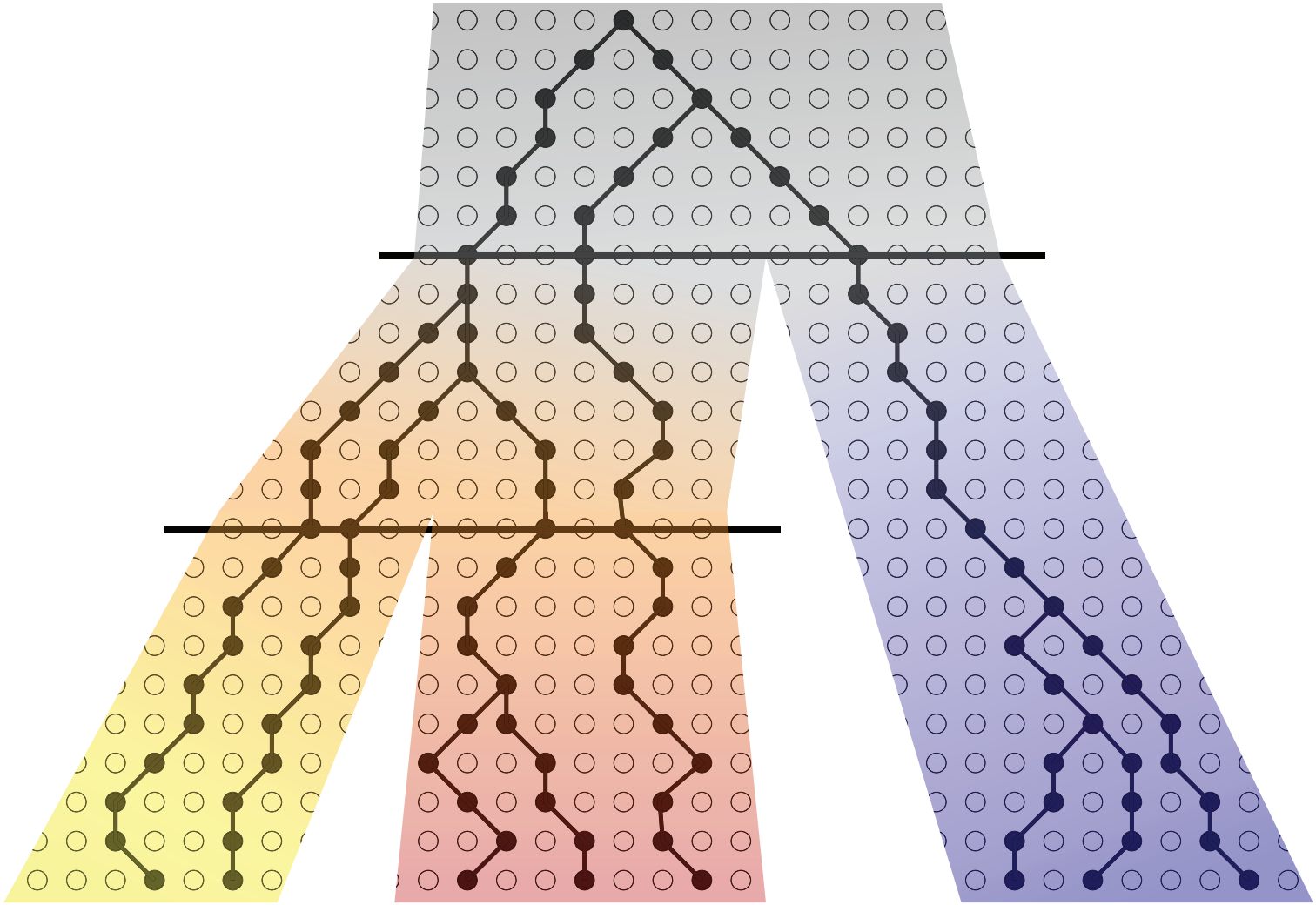


image from G.Jones, 2015, bioRxiv

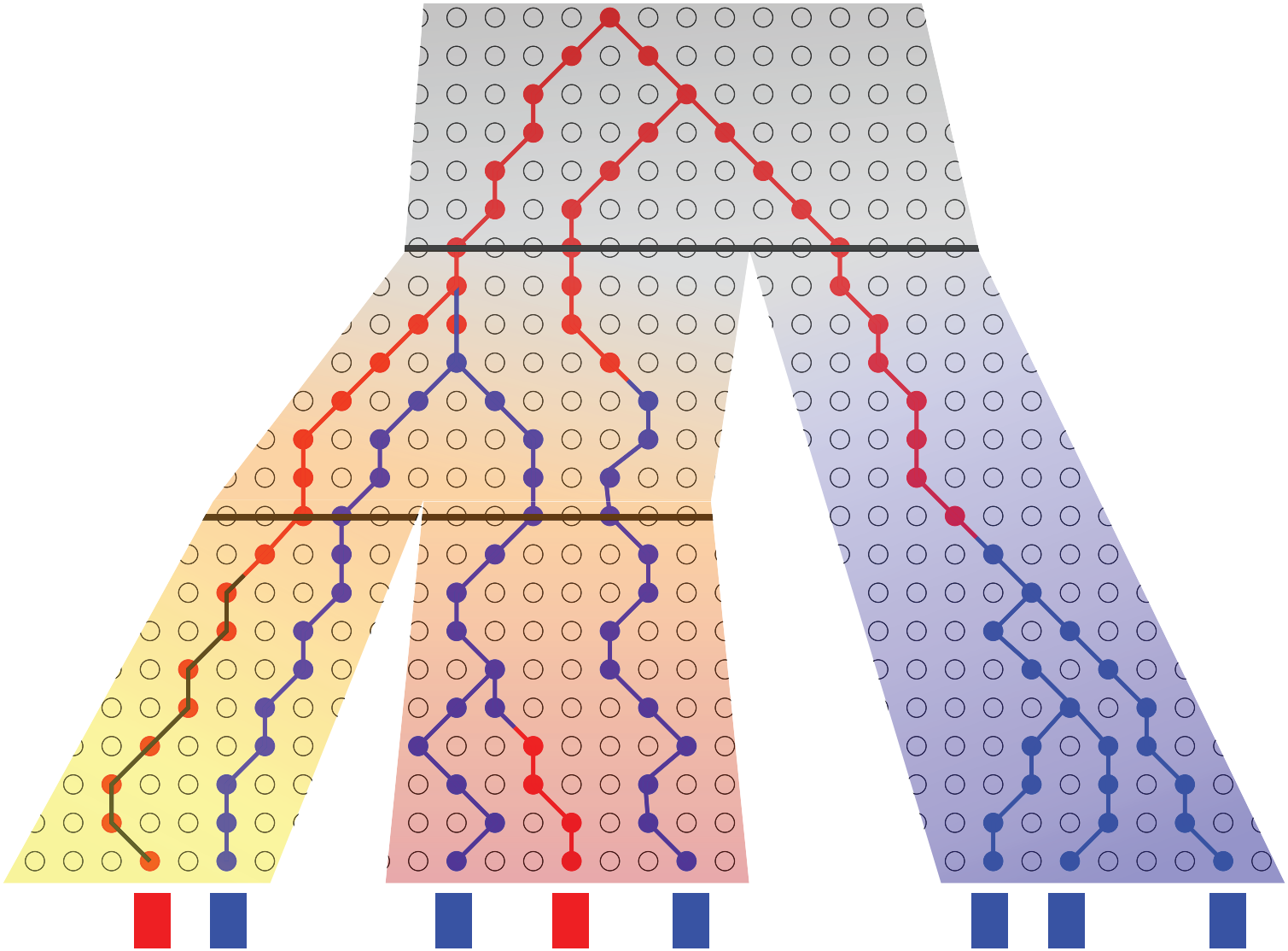
Gene trees and species trees



Gene trees and species trees

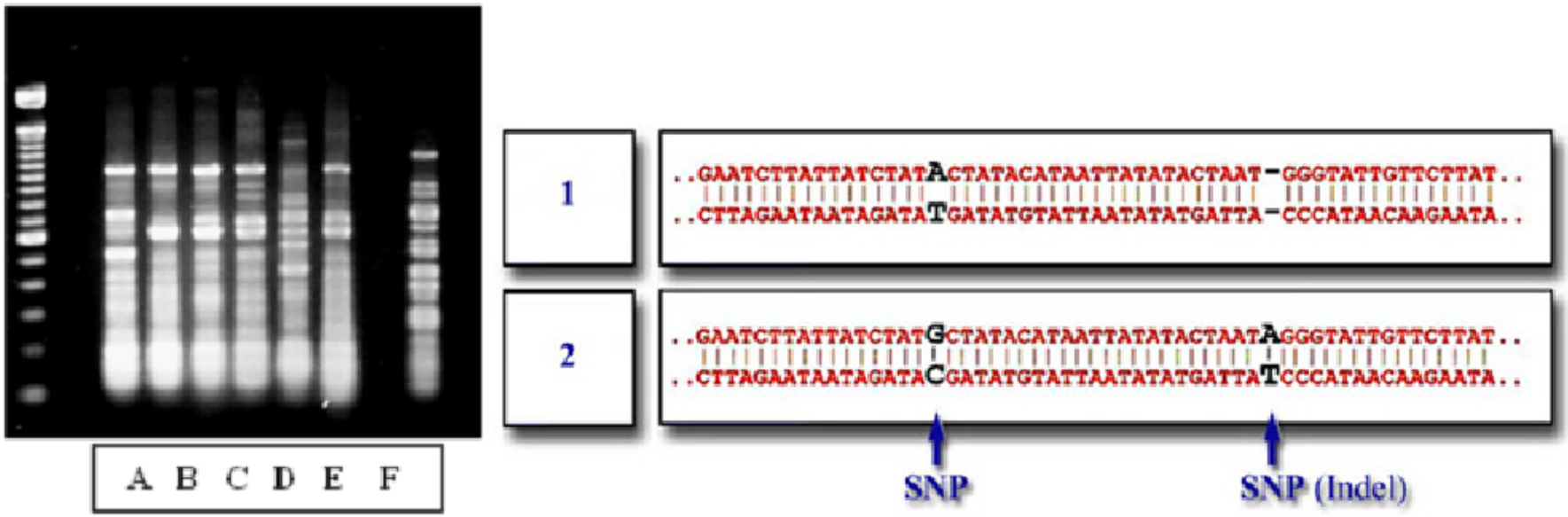


Gene trees and species trees



Bi-allelic markers (SNPs and AFLPs)

SNAPP = SNP and AFLP Package for Phylogenetic Analysis
= multi species coalescent without those pesky gene trees.



- Assumptions: independent sites, only coalescent and mutation (no selection, migration, gene flow, ...)
- one gene tree per site
- integrates out gene trees
- tree in units of substitutions

(Bryant et al, MBE 2012)

Ascertainment correction

SNP data is selected under various conditions

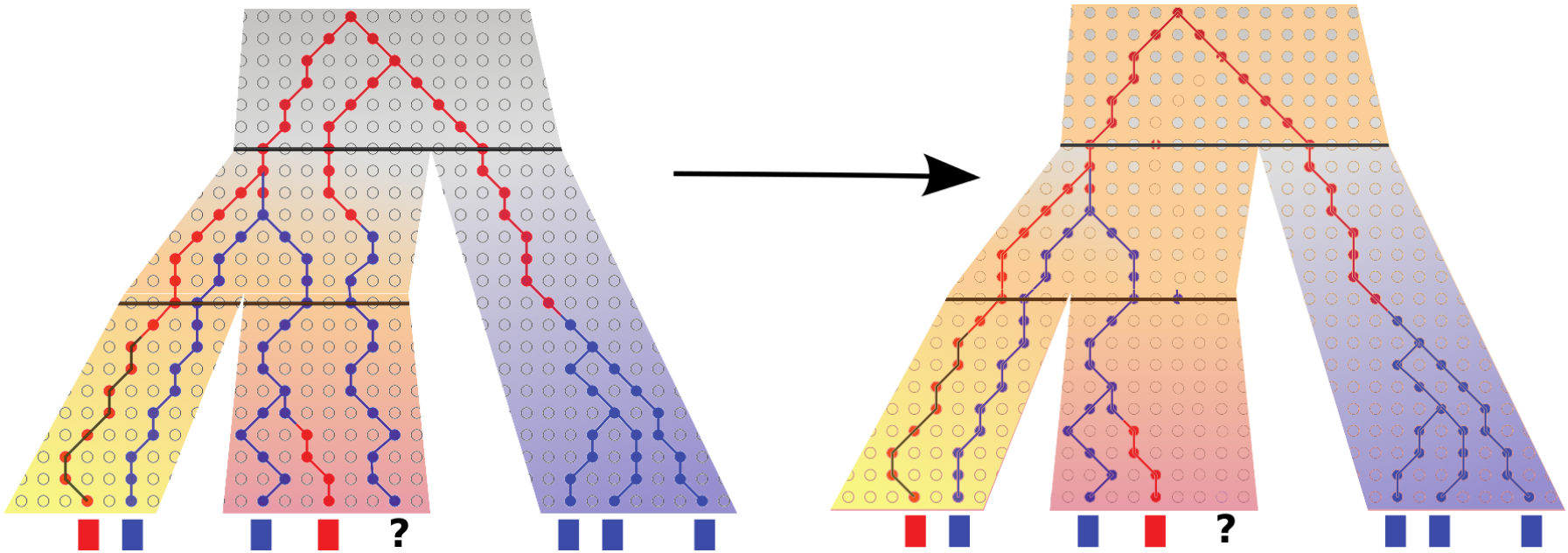
- Non constant sites only
- At least N different sites only, e.g. no constant and no singletons
- Panels – different ascertainment within species.
- Others...

This has considerable impact on pop sizes/tree height estimates

Active field of research/work in progress...

Missing data

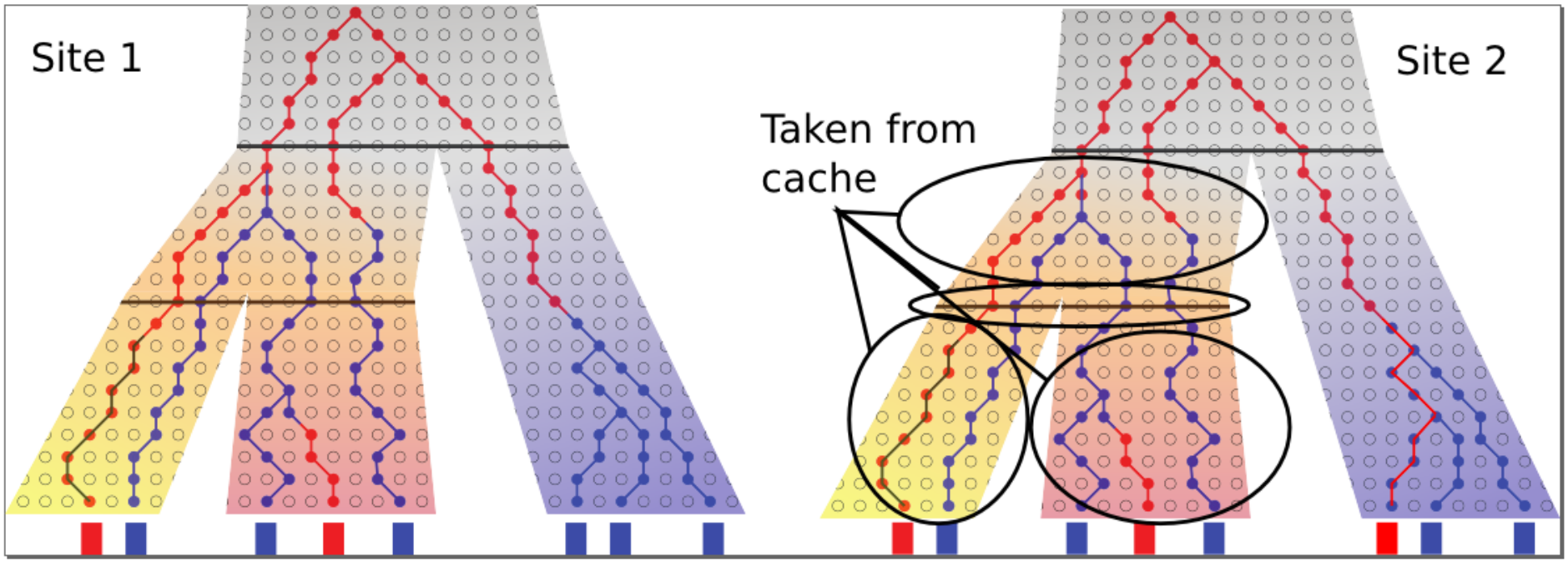
SNAPP handles a site with missing data as if the lineage does not exist



Best not used without non-polymorphic sites

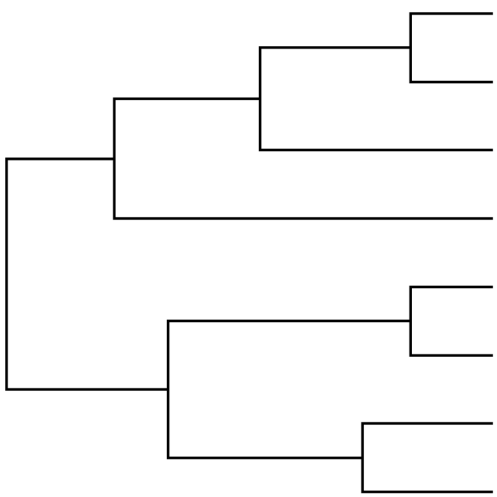
Performance tuning

= tree likelihood calculation optimisation

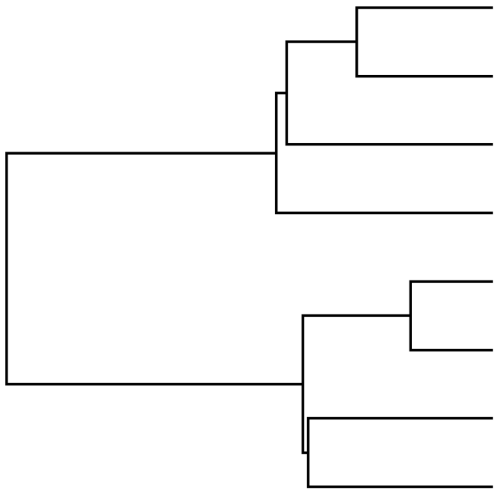


- Use threads – need to find optimal number for your data
- MCMCMC – from BEASTLabs package
- Start with small nr of lineages, increase till running out of patience
- Subsample lineages, not sites
- (No BEAGLE support, none expected)

Species tree topology simulations



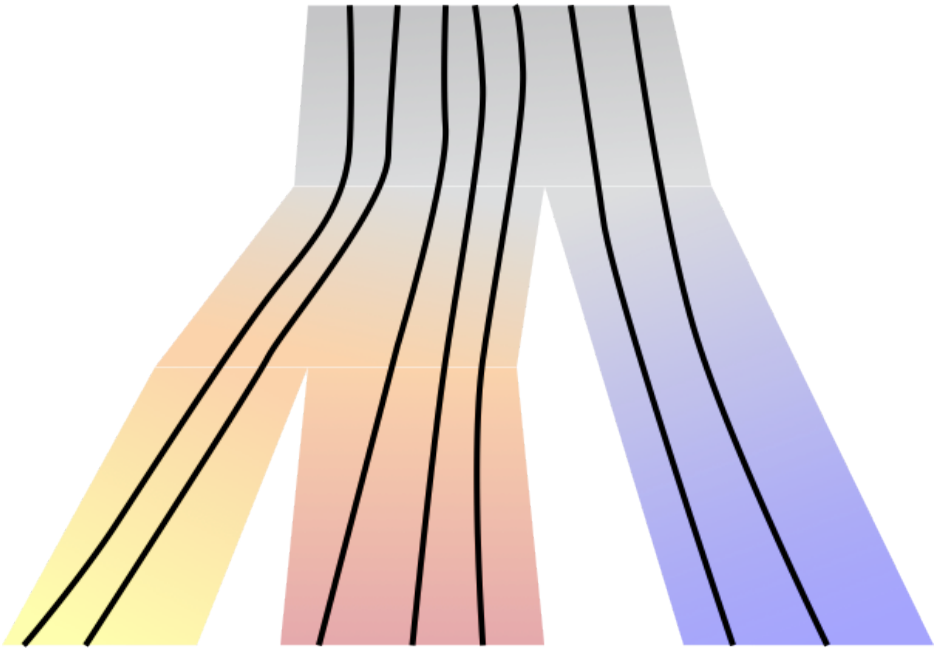
'Easy tree'
3 trees in credibility set for $\leq$ 400 loci
1 tree in credibility set for $\geq$ 500 loci.
True tree always in credibility set



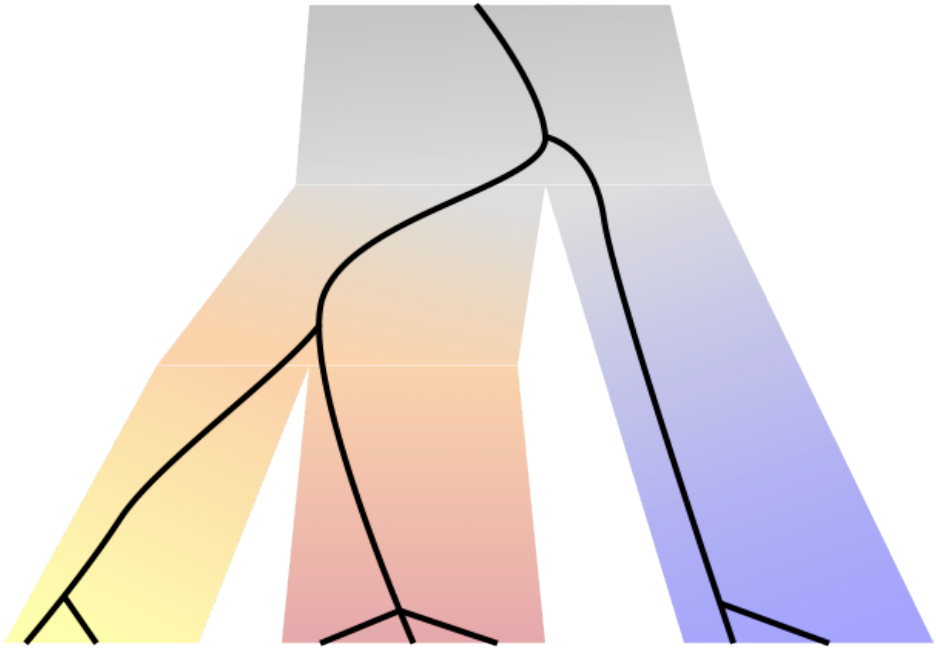
'Hard tree'
2-3 trees in credibility set for $\leq$ 10000 loci
1 tree in credibility set for $\geq$ 100000 loci.
True tree always in credibility set

Species tree topology is recovered well

Finding the lineage sorting sweet spot



No knowledge about topology
'Infinite' population size estimates
Quite a bit of knowledge about ancestral population



Some knowledge about topology (like a standard phylogenetic analysis)
Poor population size estimates
Some knowledge about present-day populations sizes

Powers and limitations of SNAPP

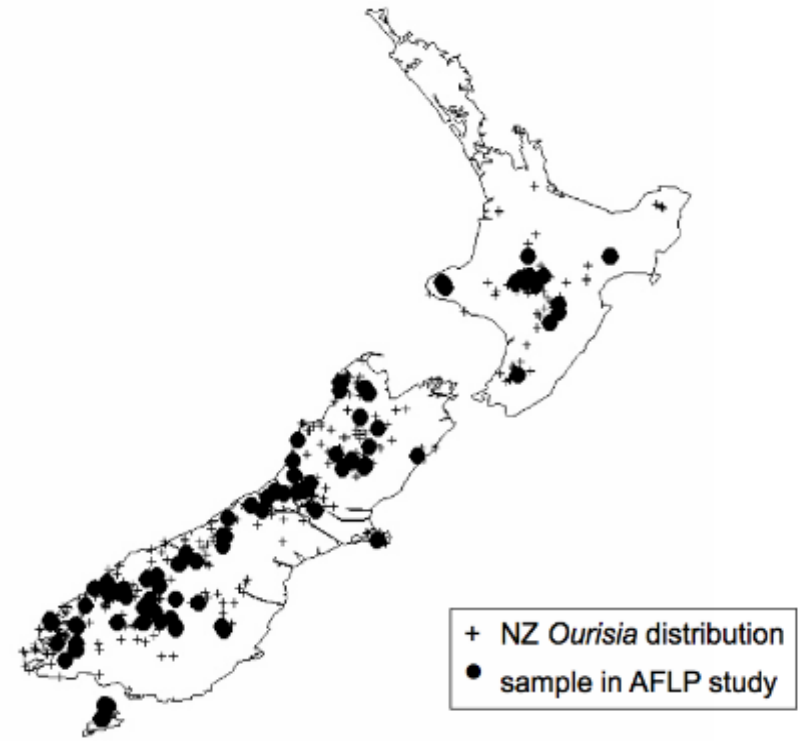
SNAPP recovers

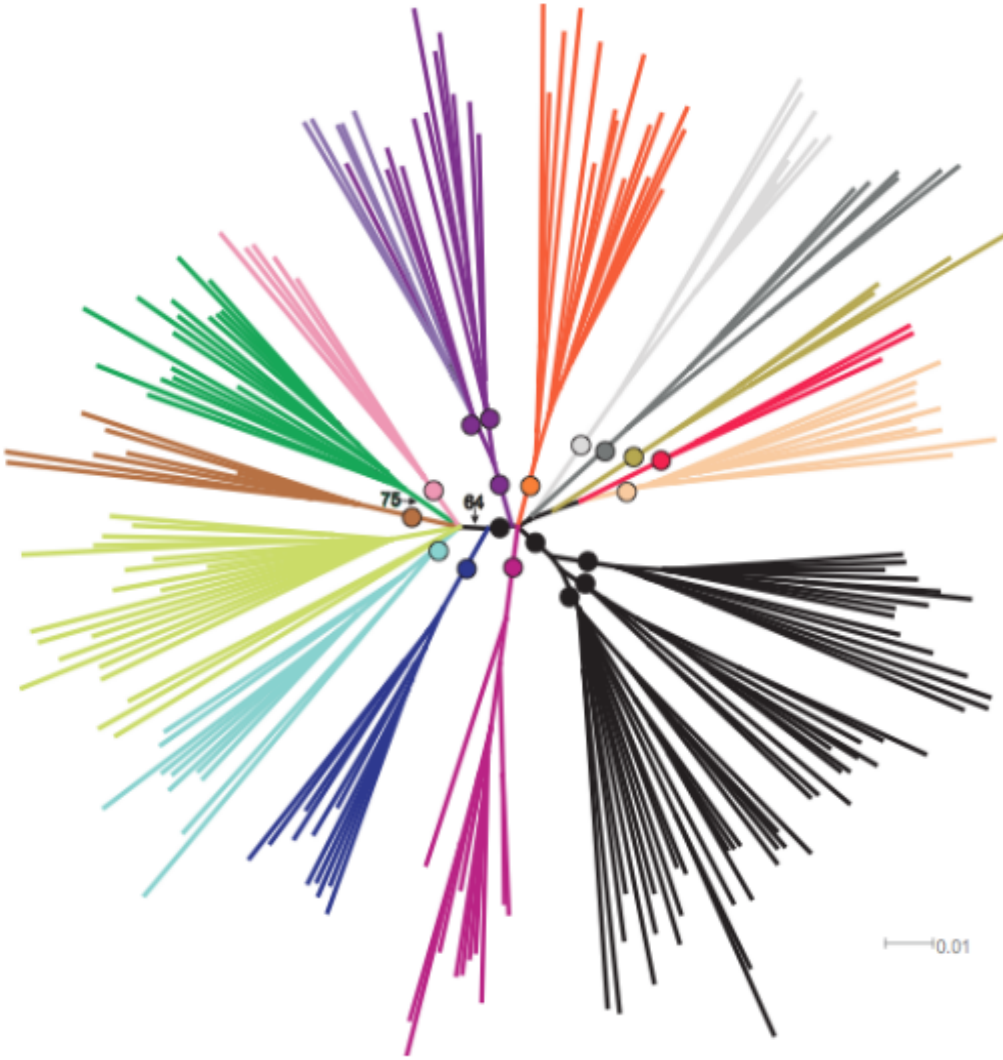
- Topology
- Coalescent times
- Population sizes per branch

in order of decreasing accuracy

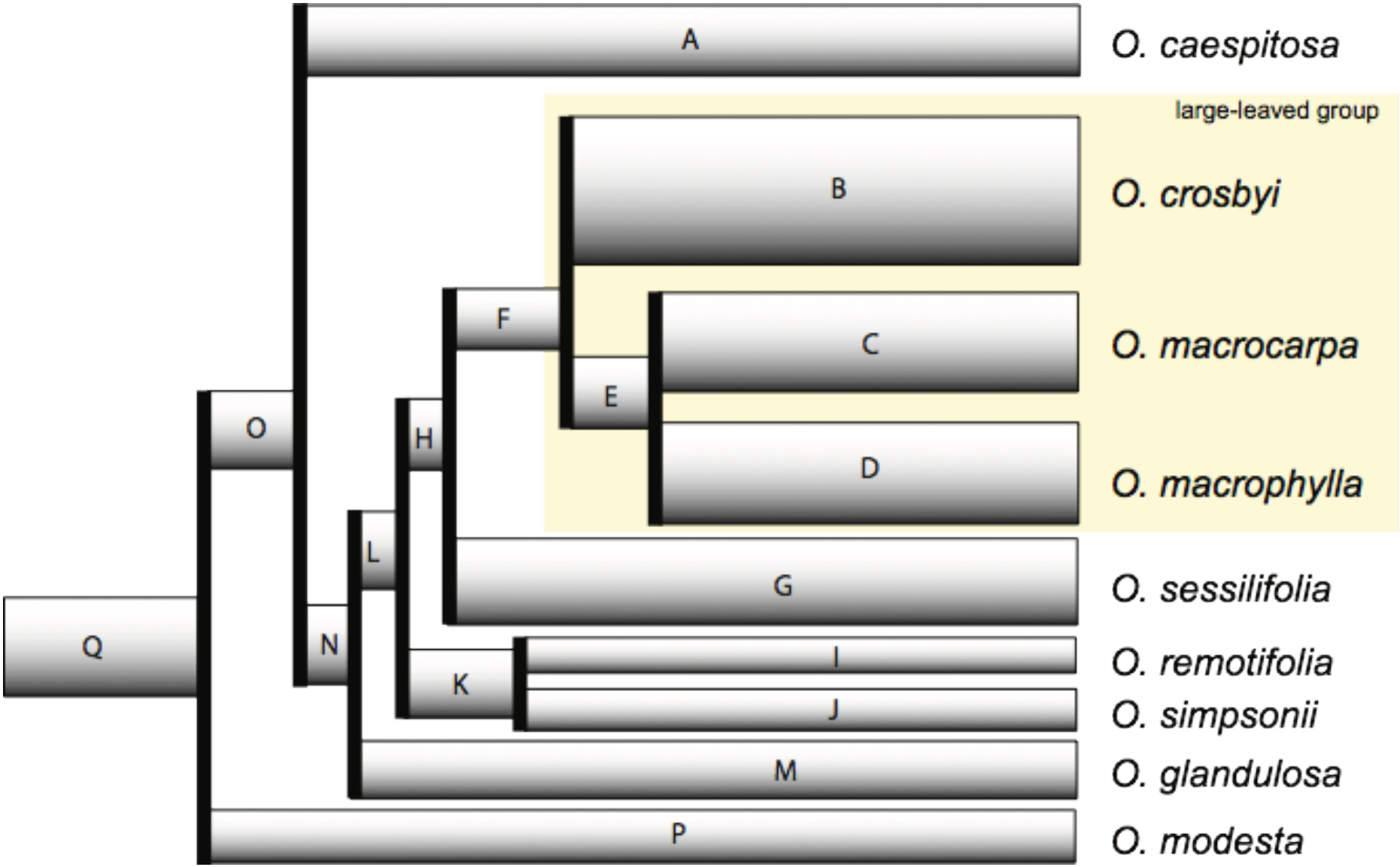
- Not enough lineages $\Rightarrow$ pop size samples from prior (happens often near root)
- θ and coalescent time estimates are often more accurate at the bottom than the top of tree
  (higher uncertainty when going back in time)

Example: alpine plant species radiation

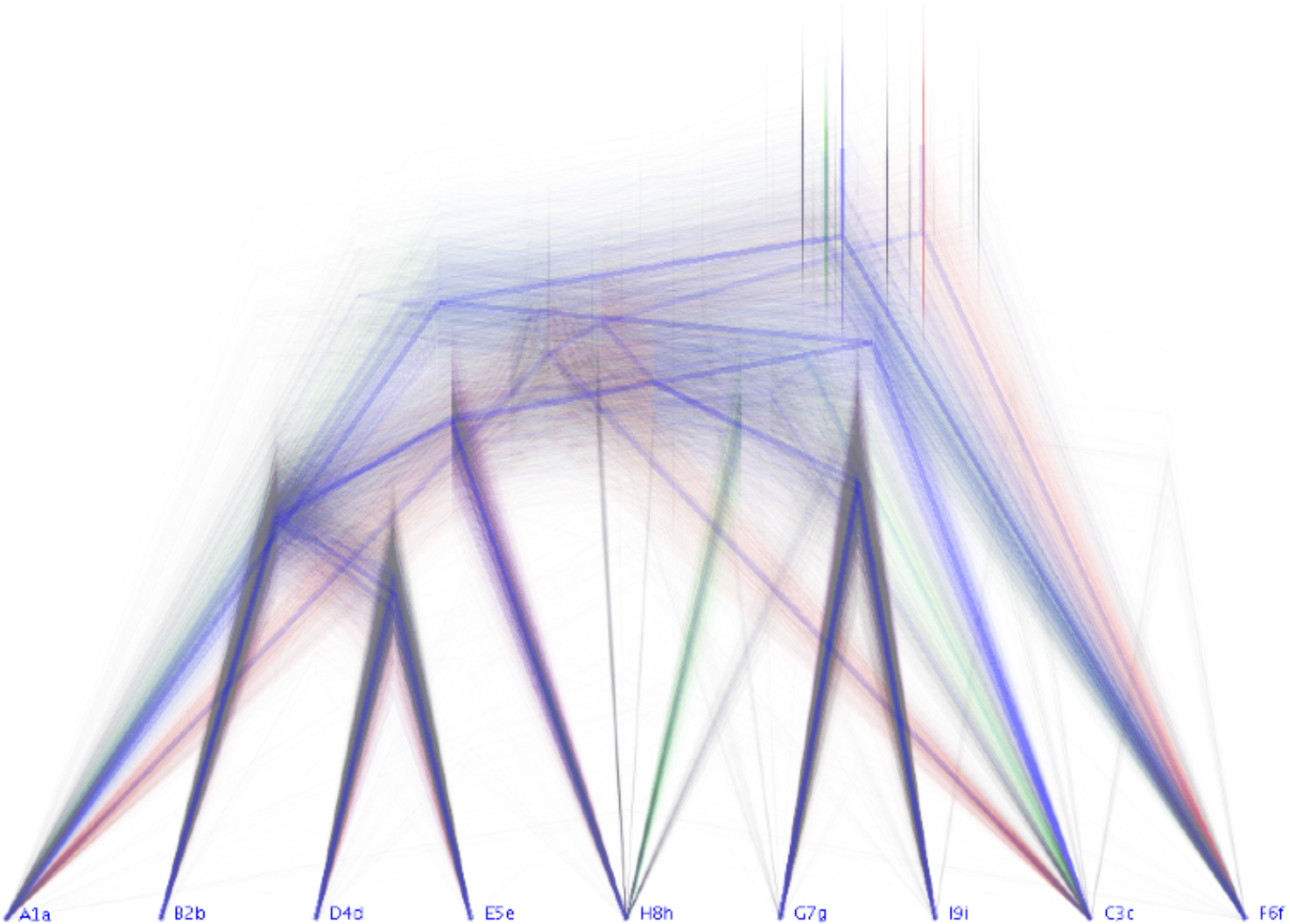




Pipe tree from SNAPP



DensiTree from SNAPP



Final analysis

Species delimitation

What is a species?

Home » News & events » Species delimitation in the age of genomics

Species delimitation in the age of genomics

The 2015 CBA conference aims to explore the impact of genomics on species discovery and delimitation in the context of taxonomic practice and applications.

Over three days, local and international speakers will expose to Australian systematists to developing analytical techniques and software for genomic species delimitation (separate or simultaneous with phylogenetic inference), in the context of integrative taxonomy.

Species delimitation in the age of genomics

- Submissions
- Registration
- Attendee information
- Program
- Organisation
- Financial support

Date

28–30 April 2015

Location

Australian National Botanic Gardens, Canberra

NATIONAL CENTER for Science Education

NCSE

DEFENDING THE TEACHING OF EVOLUTION & CLIMATE SCIENCE

NCSE provides information and advice as the premier institution dedicated to keeping evolution and climate change in the science classroom and to keeping creationism and climate change denial out. LEARN MORE >>

ABOUT | NEWS | TAKING ACTION | CREATIONISM | EVOLUTION | CLIMATE | PUBLICATIONS | BLOG

Evolution
Polls
Science & Religion

QUICK LINKS

- "Intelligent Design" on Trial
- *Reports of the NCSE*
- Events
- Library/Archives
- Polls
- Home

Home » Reports of the NCSE » Volume 26 (2006) » RNCSE 26 (4) » Species, Kinds, and Evolution

Species Concepts in Modern Literature

September 17th, 2008

Please note: This text is part of Species, Kinds, and Evolution, by John Wilkins, *Reports of NCSE* **26** (4), 2006.

Summary of 26 species concepts

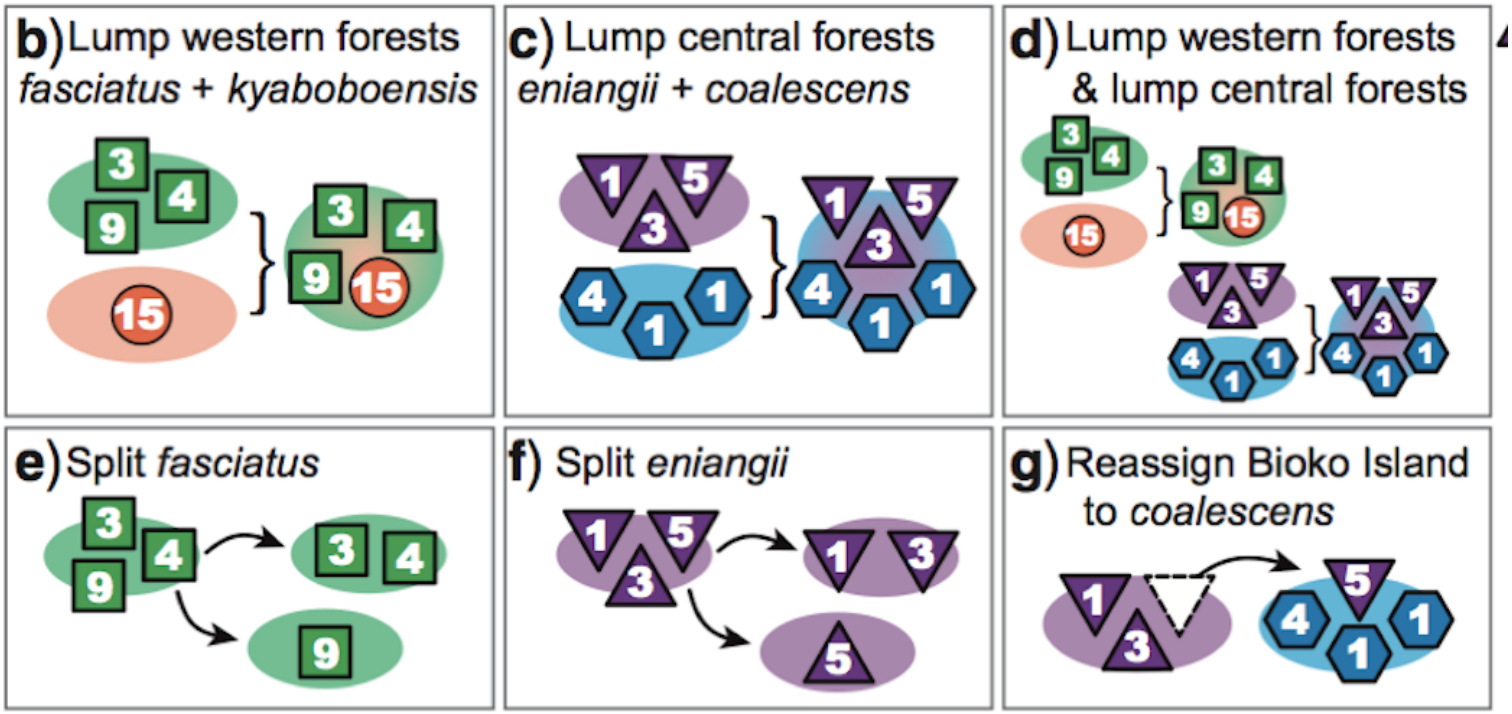
There are numerous species "concepts" at the research and practical level in the

coywolf/woyote

Species delimitation – BFD*

Variant of Bayesian model comparison

- Suppose there are two species assignments A and B



- Estimate log marginal likelihood (ML_A) using stepping stone analysis of taxon sets A
- Estimate log marginal likelihood (ML_B) of taxon sets B

The difference $\Delta = ML_A - ML_B$ is the log Bayes factor.

Leaché et al 2014 SysBio

Species delimitation

The difference $\Delta = ML_A - ML_B$ is the log Bayes factor.

	Δ	< 0	support for B
$0 <$	Δ	< 1.1	barely worth mentioning
$1.1 <$	Δ	< 3	substantial support for A
$3 <$	Δ	< 5	strong support for A
$5 <$	Δ		decisive

Stepping stone/path sampling analysis

Requires model-selection package in BEAST
Setting up and running an analysis:

- Start BEAST AppStore , select PathSampler
- or hack XML (see wiki or tutorial)

How many steps?

- start with small number of steps, say 8
- increase nr of steps till ML estimate does not decrease any more

How long

- no hard and fast rule
- ESS $>$ 200 for each step is over kill
- multiple (say 4) runs giving same estimate

Stepping stone/path sampling analysis

Requires model-selection package in BEAST
Setting up and running an analysis:

- Start BEAST AppStore , select PathSampler
- or hack XML (see wiki or tutorial)

How many steps?

- start with small number of steps, say 8
- increase nr of steps till ML estimate does not decrease any more

How long

- no hard and fast rule
- ESS $>$ 200 for each step is over kill
- multiple (say 4) runs giving same estimate

Stepping stone/path sampling analysis

Requires model-selection package in BEAST
Setting up and running an analysis:

- Start BEAST AppStore , select PathSampler
- or hack XML (see wiki or tutorial)

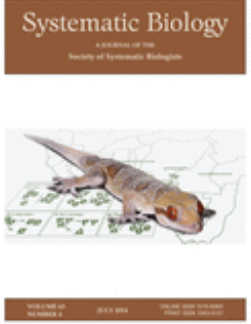
How many steps?

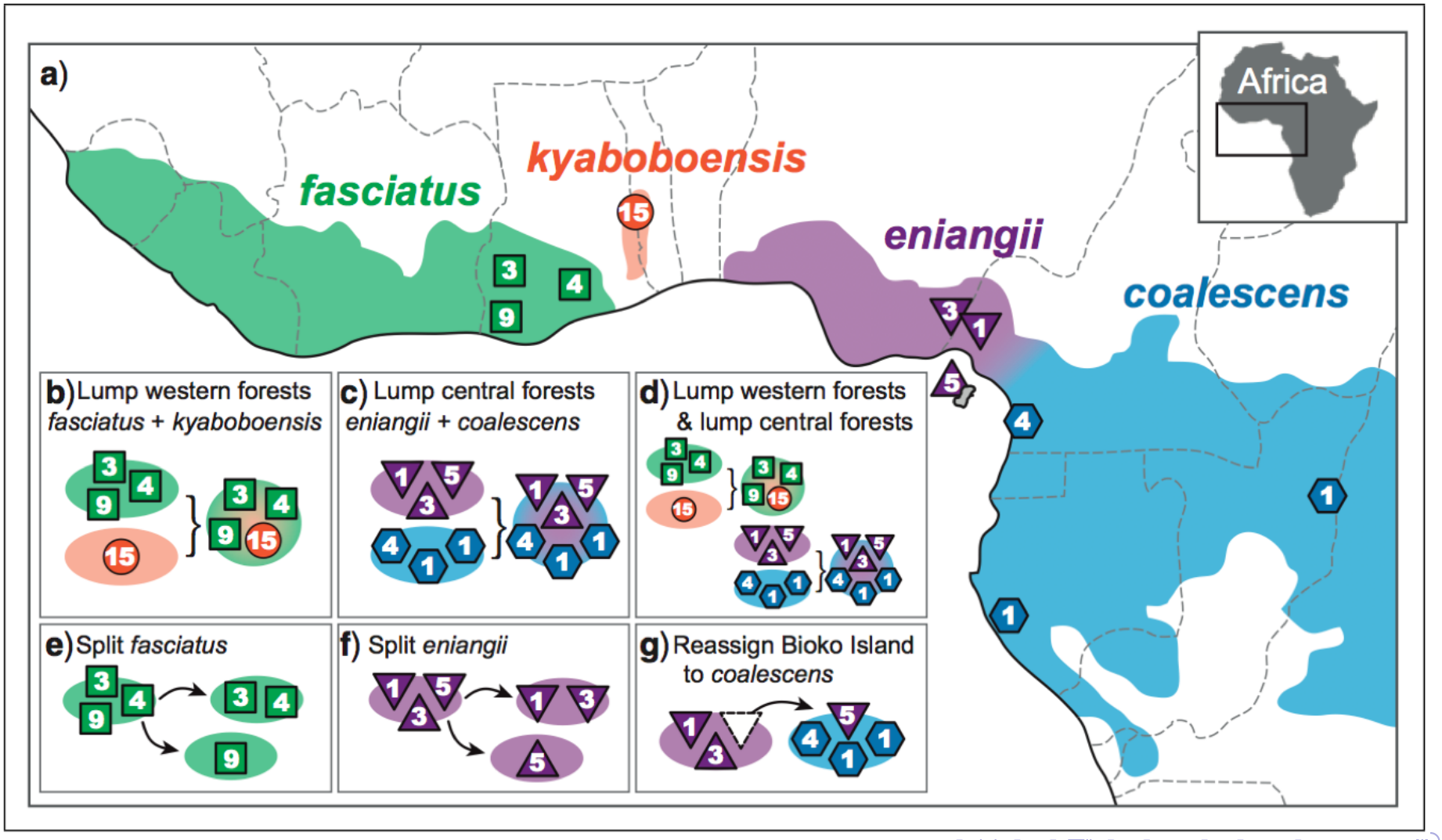
- start with small number of steps, say 8
- increase nr of steps till ML estimate does not decrease any more

How long

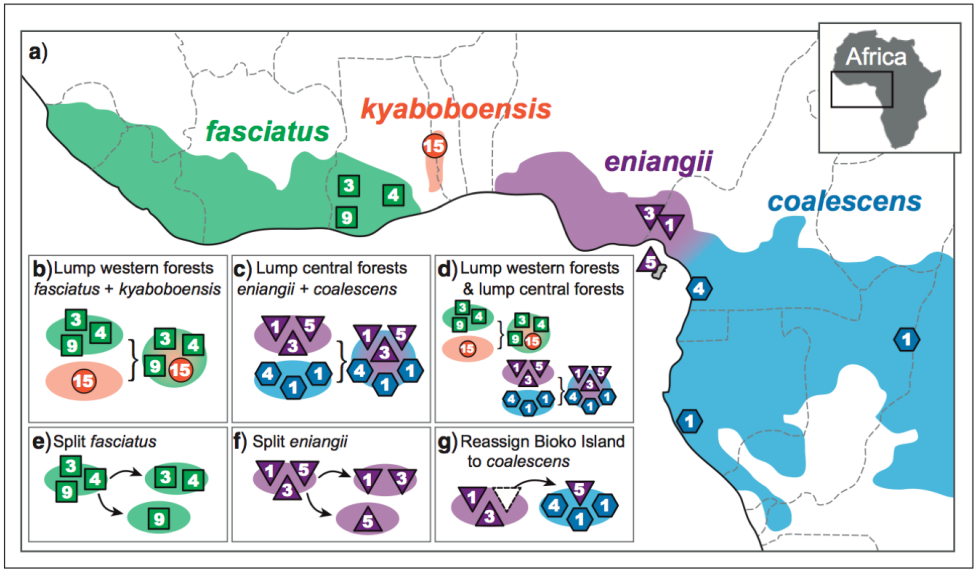
- no hard and fast rule
- ESS $>$ 200 for each step is over kill
- multiple (say 4) runs giving same estimate

BFD* example


Systematic Biology


a)
fasciatus
kyaboboensis
eniangii
coalescens
Africa
b) Lump western forests
fasciatus + kyaboboensis
c) Lump central forests
eniangii + coalescens
d) Lump western forests
& lump central forests
e) Split fasciatus
f) Split eniangii
g) Reassign Bioko Island
to coalescens

BFD* example



Model	Species	MLE	Rank	BF
runA, current taxonomy	4	-1673.4	2	–
runB, lump western forests	3	-1724.2	5	+101.5
runC, lump central forests	3	-1788.0	6	+229.2
runD, lump western & central forests	2	-1842.9	7	+339.0
runE, split *fasciatus*	5	-1713.2	4	+79.7
runF, split *eniangii*	5	-1625.9	1	-95.1
runG, reassign Bioko Island	4	-1712.6	3	+78.4

Detecting Anomalies

Consider two species

- What happens when they are grouped as one species? Pop size increases
- Where to look for cryptic species? Large pop size branches
- Where to look for mislabeled lineages? Large pop size branches

Species delimitation with *BEAST (variants)

- Bayesian Factor Delimitation aka BFD
  - Fix species tree under a number of hypotheses
  - Use Bayesian model comparison (stepping stone)
  - Use Bayes factors to select hypothesi/es with highest support
- Threshold methods: `DISSECT/STACEY` packages
  - Sample species tree topology
  - Spike in prior on near-zero leaf branch lengths
  - Any coalescent with time $< \delta$ is considered to be within species

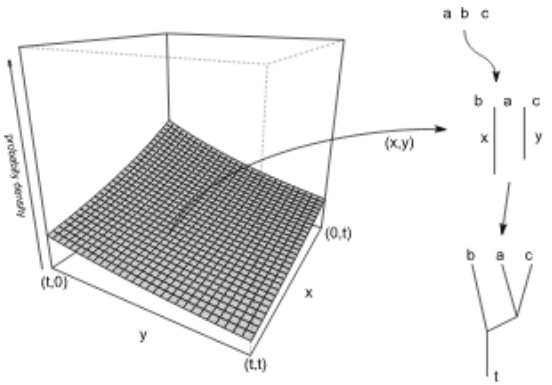


Figure 1: Sampling trees from the usual birth-death density

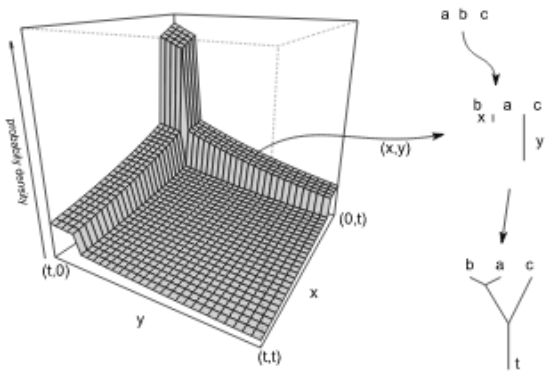


Figure 2: Sampling trees from the mixturedensity

image from Jones & Oxelman, 2014, www.indriid.com