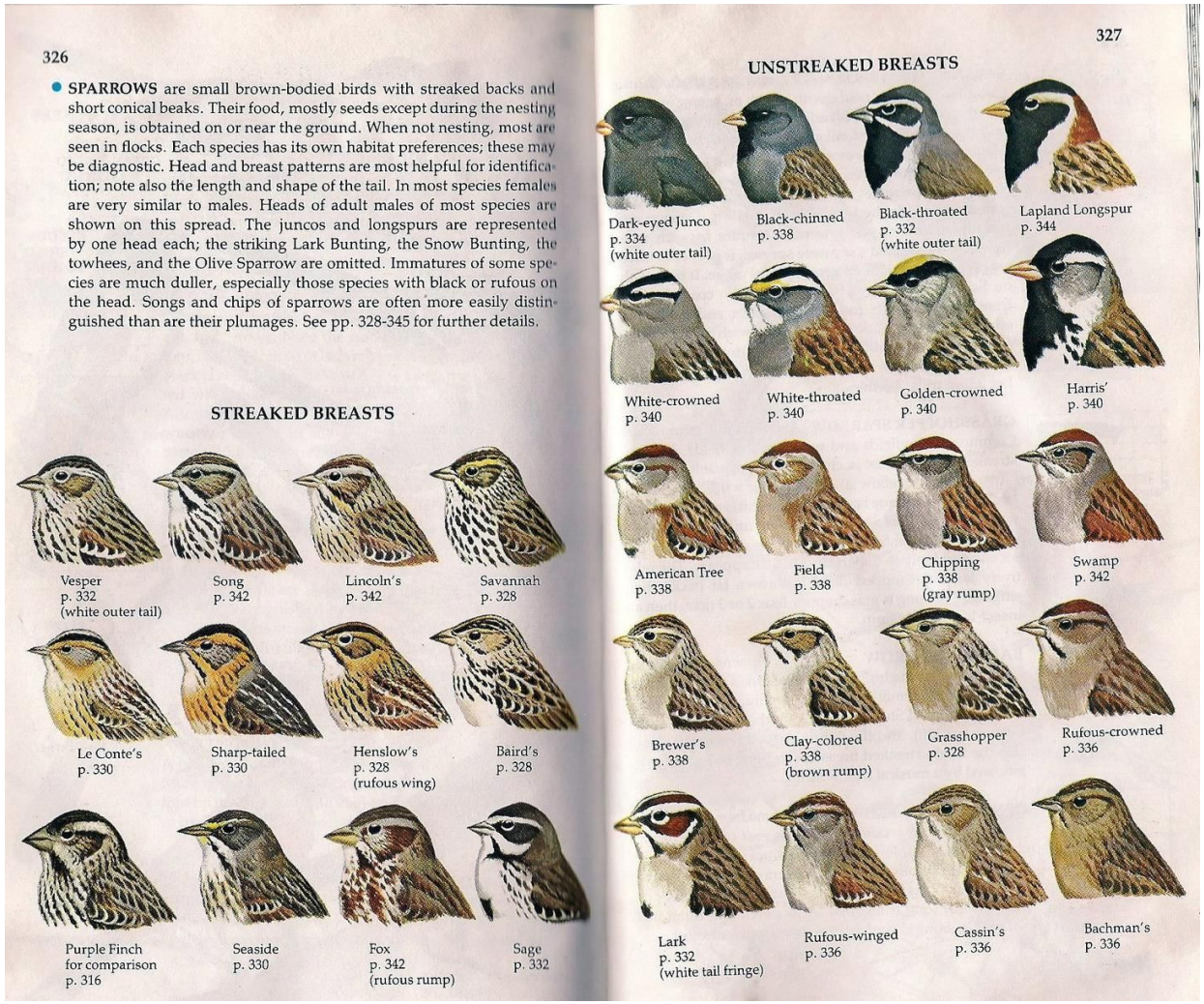
A field guide to the structured population models of BEAST 2

David Rasmussen
Dept. of Entomology and Plant Pathology
Bioinformatics Research Center
NC State University

Taming the BEAST Eh!
August 15th, 2019

Why are there so many little brown birds?



326

● **SPARROWS** are small brown-bodied birds with streaked backs and short conical beaks. Their food, mostly seeds except during the nesting season, is obtained on or near the ground. When not nesting, most are seen in flocks. Each species has its own habitat preferences; these may be diagnostic. Head and breast patterns are most helpful for identification; note also the length and shape of the tail. In most species females are very similar to males. Heads of adult males of most species are shown on this spread. The juncos and longspurs are represented by one head each; the striking Lark Bunting, the Snow Bunting, the towhees, and the Olive Sparrow are omitted. Immatures of some species are much duller, especially those species with black or rufous on the head. Songs and chips of sparrows are often more easily distinguished than are their plumages. See pp. 328-345 for further details.

STREAKED BREASTS

Vesper p. 332 (white outer tail)
Song p. 342
Lincoln's p. 342
Savannah p. 328
Le Conte's p. 330
Sharp-tailed p. 330
Henslow's p. 328 (rufous wing)
Baird's p. 328
Purple Finch for comparison p. 316
Seaside p. 330
Fox p. 342 (rufous rump)
Sage p. 332

327

UNSTREAKED BREASTS

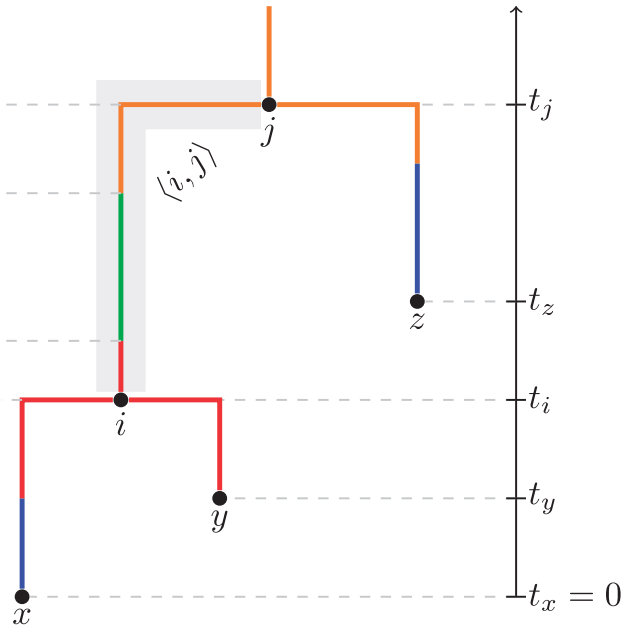
Dark-eyed Junco p. 334 (white outer tail)
Black-chinned p. 338
Black-throated p. 332 (white outer tail)
Lapland Longspur p. 344
White-crowned p. 340
White-throated p. 340
Golden-crowned p. 340
Harris' p. 340
American Tree p. 338
Field p. 338
Chipping p. 338 (gray rump)
Swamp p. 342
Brewer's p. 338
Clay-colored p. 338 (brown rump)
Grasshopper p. 328
Rufous-crowned p. 336
Lark p. 332 (white tail fringe)
Rufous-winged p. 336
Cassin's p. 336
Bachman's p. 336

Why are there so many structured models in BEAST 2?

- Discrete-trait models (Lemey *et al.*, 2009)
- MultiTypeTree (Vaughan, *et al.*, 2014)
- BASTA (De Maio et al., 2015)
- MASCOT (Müller *et al.*, 2018)
- PhyDyn (Volz & Siveroni, 2018)
- BDMM (Kühnert *et al.*, 2016)
- MSBD (Barido-Sottani et al., 2019)

What are structured population models?

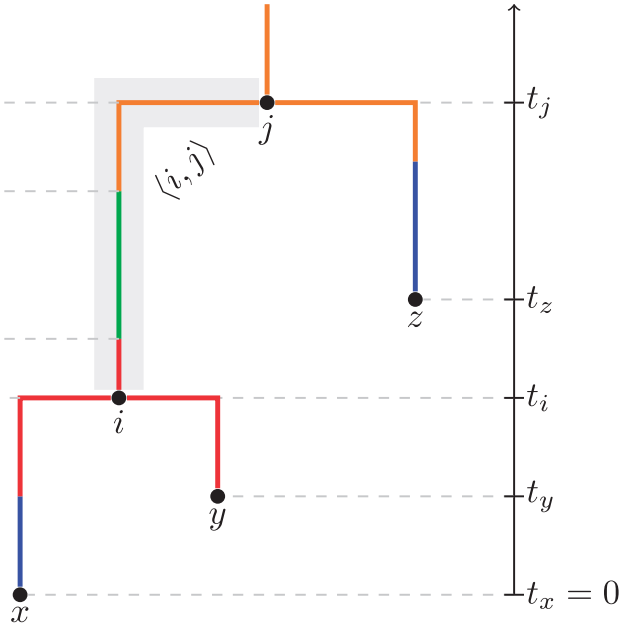
- Any model where lineages can reside in different populations or “type” states



Vaughan et al., (2014)

What are structured population models?

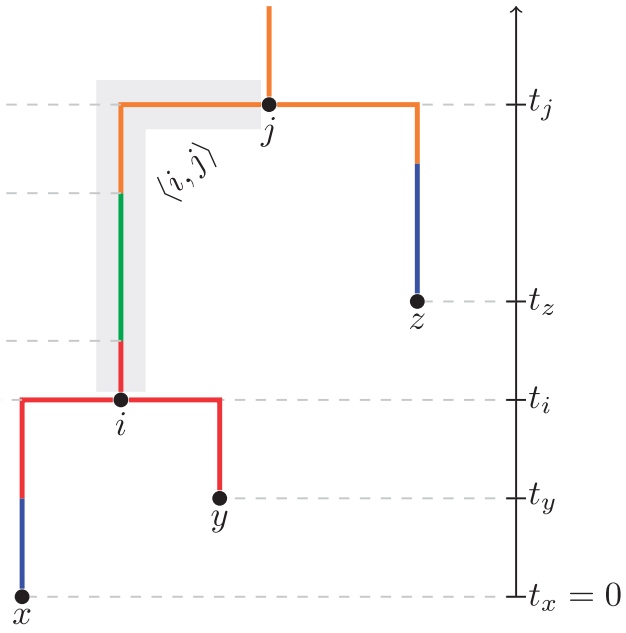
- Any model where lineages can reside in different populations or “type” states
- Types can represent different biological populations, geographic locations, infectious states, character traits, ect.



Vaughan et al., (2014)

What are structured population models?

- Any model where lineages can reside in different populations or “type” states
- Types can represent different biological populations, geographic locations, infectious states, character traits, ect.
- Key point is that not all lineages are equivalent or “exchangeable”



Vaughan et al., (2014)

Why are we interested in structured population models?

- Often we would like to be able to reconstruct ancestral states like the geographic location of a lineage.

- We may also be interested in estimating *migration* or transition rates between populations or states.

- Even if we are not directly interested in the movement of lineages, population structure can confound other demographic inferences (e.g. pop size estimates under coalescent models).

The most direct approach: DTA

- Discrete-trait or “mugation” models track the movement of lineages as a continuous time Markov process similar to the substitution models we use in molecular evolution (e.g. HKY or GTR).

The most direct approach: DTA

- Discrete-trait or "mugation" models track the movement of lineages as a continuous time Markov process similar to the substitution models we use in molecular evolution (e.g. HKY or GTR).

- We have an instantaneous transition rate matrix:

$$Q = \begin{bmatrix} -\sum_{i \neq 1}^{n} m_{1,i} & m_{1,2} & \cdots & m_{1,n} \\ m_{2,1} & -\sum_{i \neq 2}^{n} m_{2,i} & \cdots & m_{2,n} \\ \vdots & \vdots & \ddots & \vdots \\ m_{n,1} & m_{n,2} & \cdots & -\sum_{i \neq n}^{n} m_{n,i} \end{bmatrix}$$

The most direct approach: DTA

- Discrete-trait or "mugation" models track the movement of lineages as a continuous time Markov process similar to the substitution models we use in molecular evolution (e.g. HKY or GTR).

- We have an instantaneous transition rate matrix:

$$Q = \begin{bmatrix} -\sum_{i \neq 1}^{n} m_{1,i} & m_{1,2} & \cdots & m_{1,n} \\ m_{2,1} & -\sum_{i \neq 2}^{n} m_{2,i} & \cdots & m_{2,n} \\ \vdots & \vdots & \ddots & \vdots \\ m_{n,1} & m_{n,2} & \cdots & -\sum_{i \neq n}^{n} m_{n,i} \end{bmatrix}$$

- And we can use Q to compute the probability of a lineage being in any state at any time t in the past:

$$P(t) = e^{Qt} P(0)$$

DTA models in BEAST

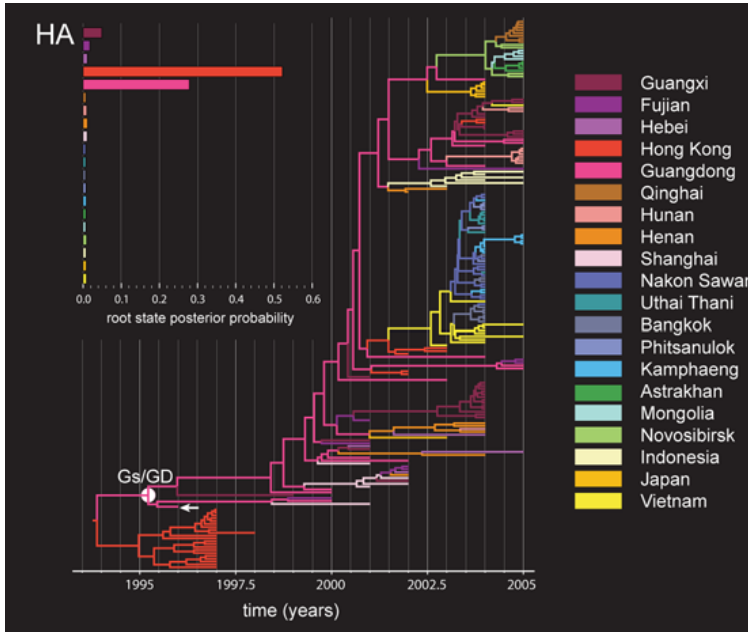
OPEN ACCESS Freely available online

PLoS COMPUTATIONAL BIOLOGY

Bayesian Phylogeography Finds Its Roots

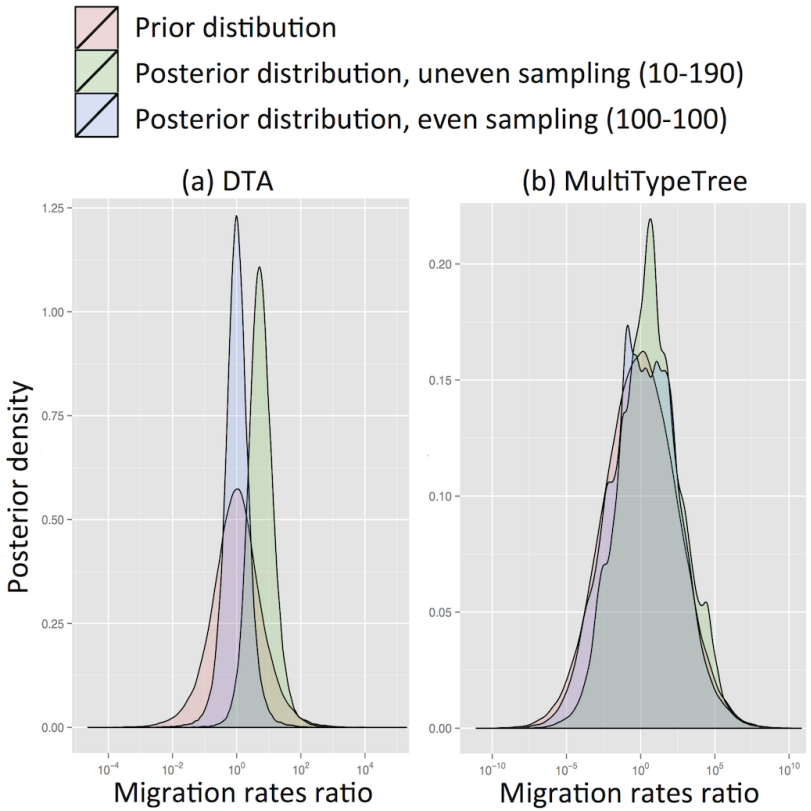
Philippe Lemey[1]*, Andrew Rambaut[2], Alexei J. Drummond[3], Marc A. Suchard[4,5]

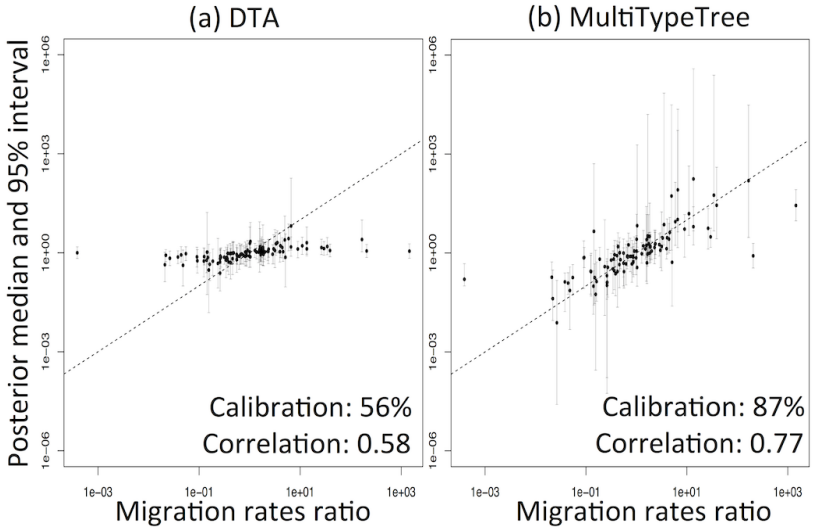
1 Department of Microbiology and Immunology, Katholieke Universiteit Leuven, Leuven, Belgium, **2** Institute of Evolutionary Biology, University of Edinburgh, Edinburgh, United Kingdom, **3** Department of Computer Science, University of Auckland, Auckland, New Zealand, **4** Departments of Biomathematics and Human Genetics, David Geffen School of Medicine, University of California, Los Angeles, California, United States of America, **5** Department of Biostatistics, School of Public Health, University of California, Los Angeles, California, United States of America



Uneven sampling strongly biases DTA

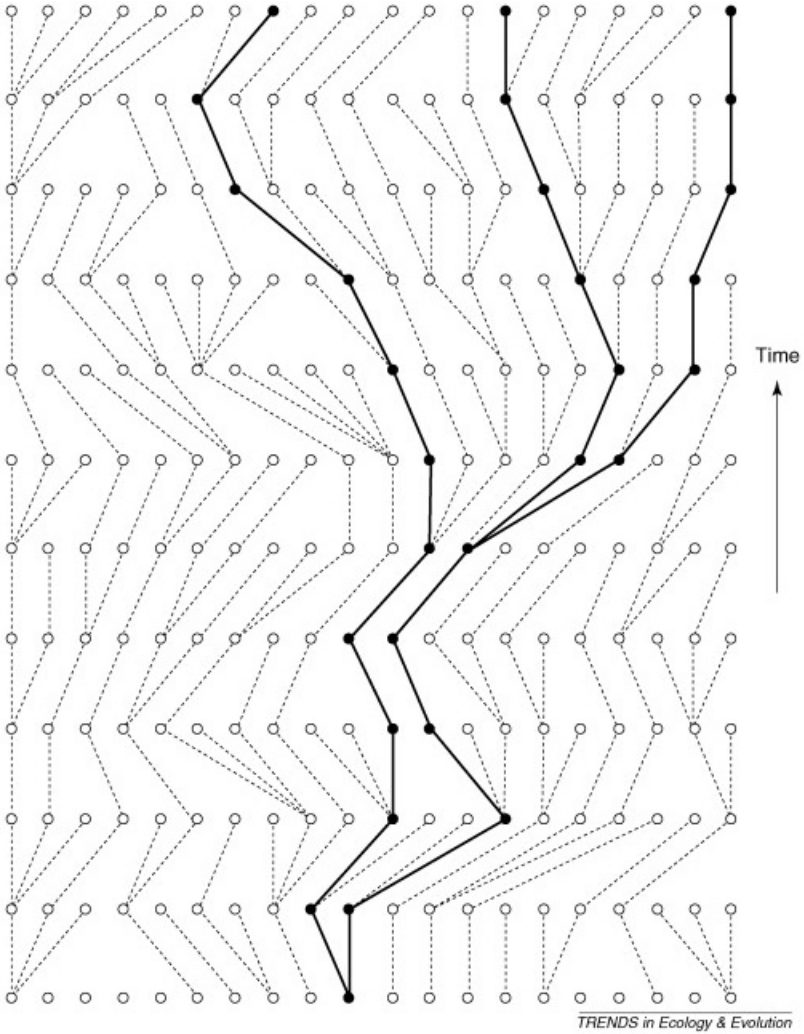
- DTA treats the sampling process as informative about the migration process.





De Maio et al., (2015)

Back to the coalescent!

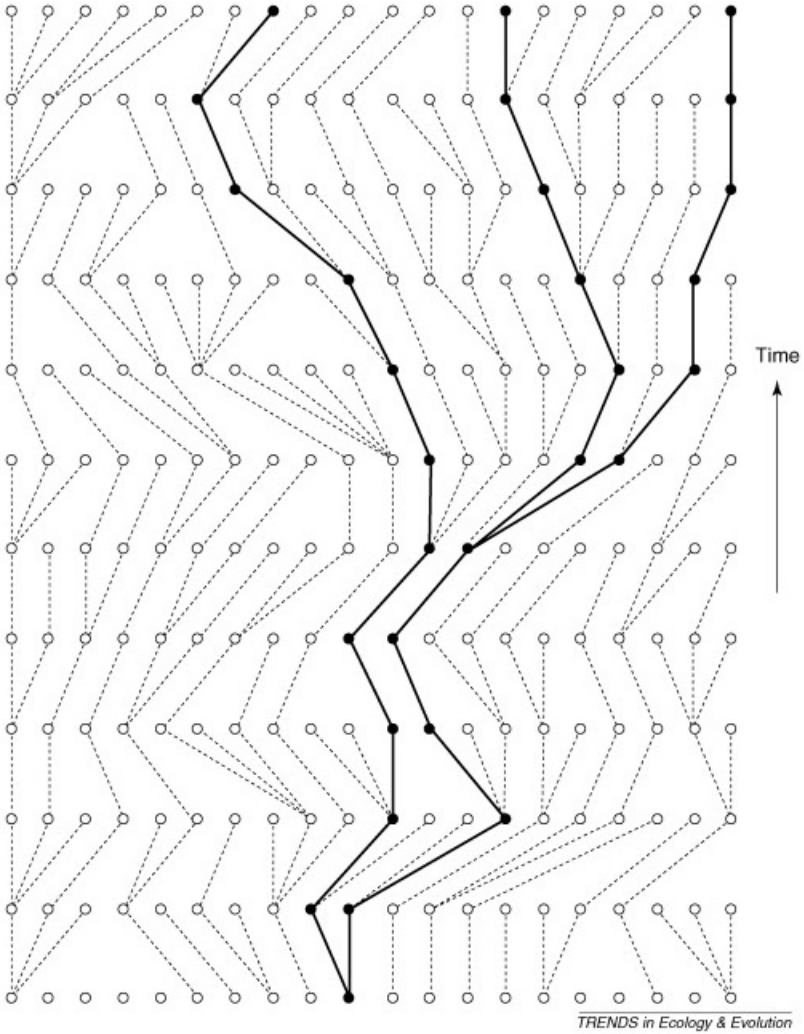


Kuhner (2008)

Probability of coalescing per generation:

$$p_{coal} = \frac{1}{N}$$

Back to the coalescent!



Kuhner (2008)

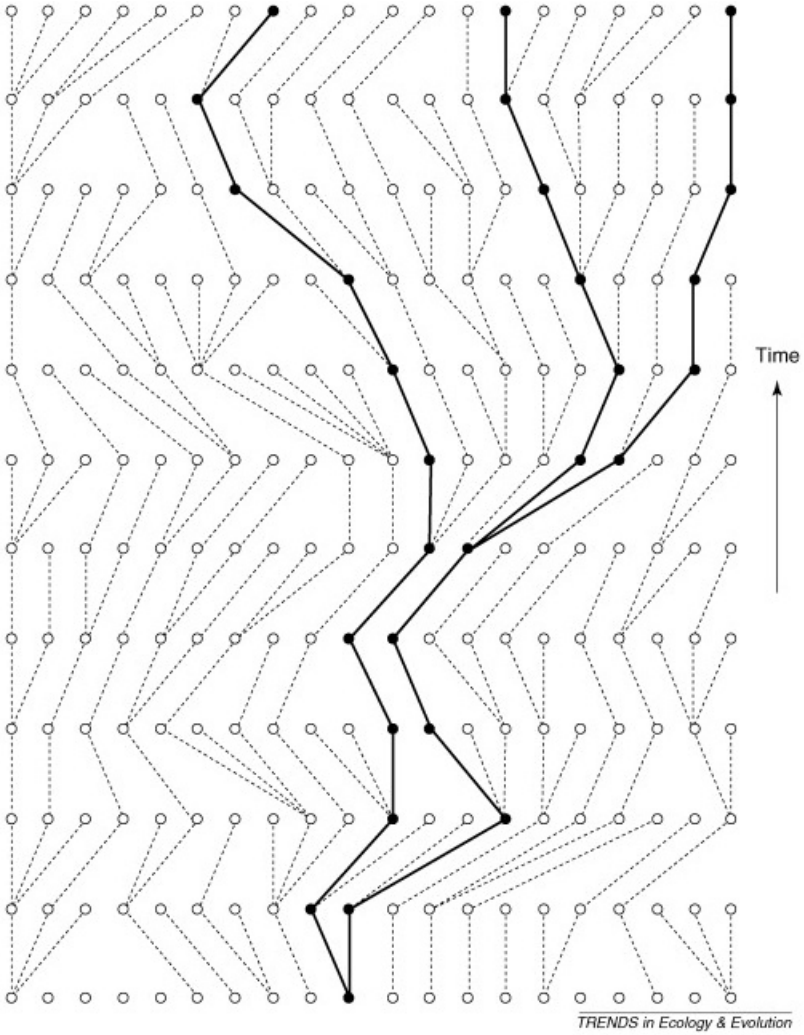
Probability of coalescing per generation:

$$p_{coal} = \frac{1}{N}$$

Probability of coalescing after *n* generations:

$$Pr(X = n) = (1 - p_{coal})^{n-1} p_{coal}$$

Back to the coalescent!



Kuhner (2008)

Probability of coalescing per generation:

$$p_{coal} = \frac{1}{N}$$

Probability of coalescing after *n* generations:

$$Pr(X = n) = (1 - p_{coal})^{n-1} p_{coal}$$

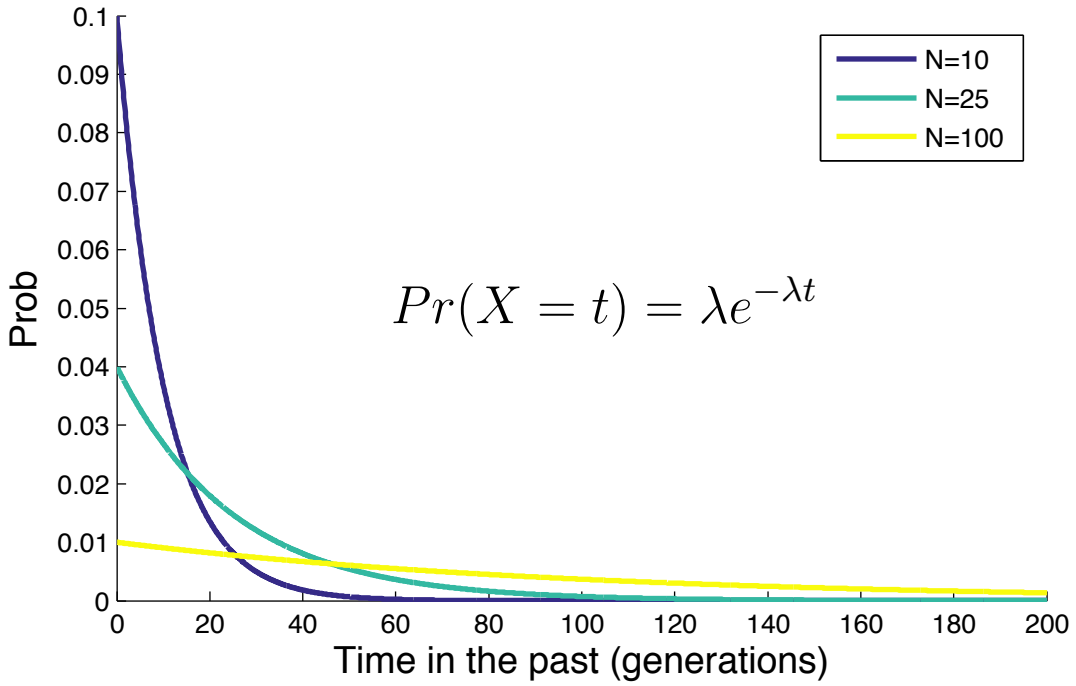
In continuous time:

$$Pr(X = t) = \lambda e^{-\lambda t}$$

$$\lambda = p_{coal} = \frac{1}{N}$$

The distribution of coalescent times

- The waiting time for a pair of lineages to coalesce follows an exponential distribution:



More than two lineages

- For more than two lineages, the rate at which any pair of lineages coalesces becomes:

$$\lambda_{coal} = \frac{\binom{k}{2}}{N_e}$$

- The binomial coefficient gives the total number of pairs we can form from *k* lineages:

$$\binom{k}{2} = \frac{k(k-1)}{2}$$

- For a tree with *n* samples and *n-1* known coalescent times, we can compute the coalescent likelihood as:

$$L(T|N_e) = \frac{1}{N_e^{(n-1)}} \prod_{k=2}^{n} \exp\left(- \frac{\binom{k}{2}}{N_e} t_k \right)$$

The problem with population structure

- Standard coalescent models assume that all lineages are exchangeable.

- Exchangeability here means that any lineage is equally likely to coalesce with any other lineage.

- Many forms of population structure violate this assumption.

The structured coalescent

- Relaxes the exchangeability assumption by letting lineages reside and move between different populations.

- Each pair of lineages is allowed to coalesce at a different rate λ_{ij} based on the states of lineages i and j.

$$L(T|\theta) = \prod_{k=2}^{n} \lambda_{ij} \exp\left[-\sum_{i}^{k} \sum_{j>i}^{k} \lambda_{ij} t_k \right].$$

- But inference is much more difficult because we must infer the location of each lineage in the tree over time.

The Migrate-n model

- A structured population model with migration between *n* subpopulations or demes.

- Model is parameterized in terms of a migration rate matrix *M* and a vector of effective population sizes Θ

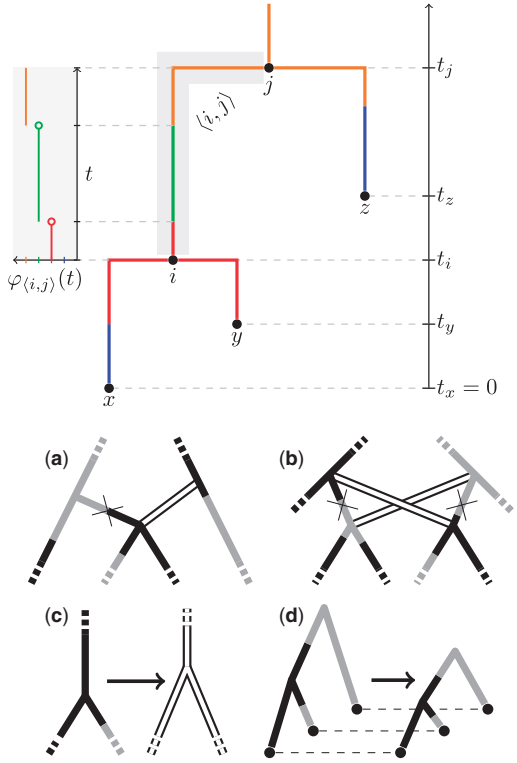
$$M = \begin{bmatrix} 0 & m_{1,2} & \cdots & m_{1,n} \\ m_{2,1} & 0 & \cdots & m_{2,n} \\ \vdots & \vdots & \ddots & \vdots \\ m_{n,1} & m_{n,2} & \cdots & 0 \end{bmatrix} \qquad \Theta = \begin{bmatrix} N_e^1 \\ N_e^2 \\ \vdots \\ N_e^n \end{bmatrix}$$

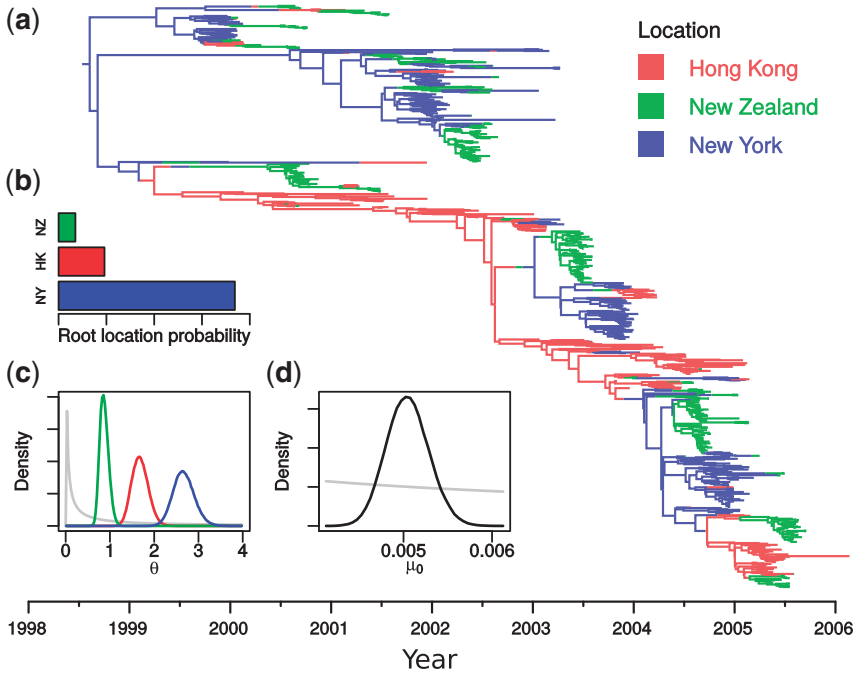
- Model allows for likelihood-based inference of *M* and Θ BUT we need to use MCMC to sample full migration histories along each lineage of the tree.

Beerli & Felsenstein (PNAS, 2001)

MultiTypeTree

- MultiTypeTree (Vaughan *et al.*, 2014) is an implementation of the Migrate-n model in BEAST 2 with a more efficient MCMC algorithm for sampling migration histories.





Sampling migration histories

- MTT offers dramatic gains in efficiency over Migrate, but it's still fundamentally limited by the need to sample migration histories using MCMC.

- Does not allow for very efficient MCMC sampling from the posterior due to strong correlations between migration histories and model parameters. Limited to about 5 or 6 states and trees < 1000 tips.

- But what if we can "integrate out" migration histories?

The Volz (2012) Structured Coalescent

- Rather than explicitly sampling migration histories, we probabilistically track the movement of each lineage.

- We can then write pairwise coalescent rates in terms of lineage state probabilities p_{ik}

$$\lambda_{ij} = \sum_{k}^{m} \sum_{l}^{m} \frac{f_{kl}}{y_k y_l} \left(p_{ik} p_{jl} + p_{il} p_{jk} \right)$$

- We track how the lineage state probabilities evolve backwards in time using a system of master equations. This assumes each lineage evolves independently of other lineages.

BASTA

- BASTA (De Maio et al., 2015) implements a specific form of the Volz (2012) structured coalescent in BEAST 2

- Assumes constant migration rates and population sizes through time and no birth/transmission events between populations.

The Volz (2012) Structured Coalescent

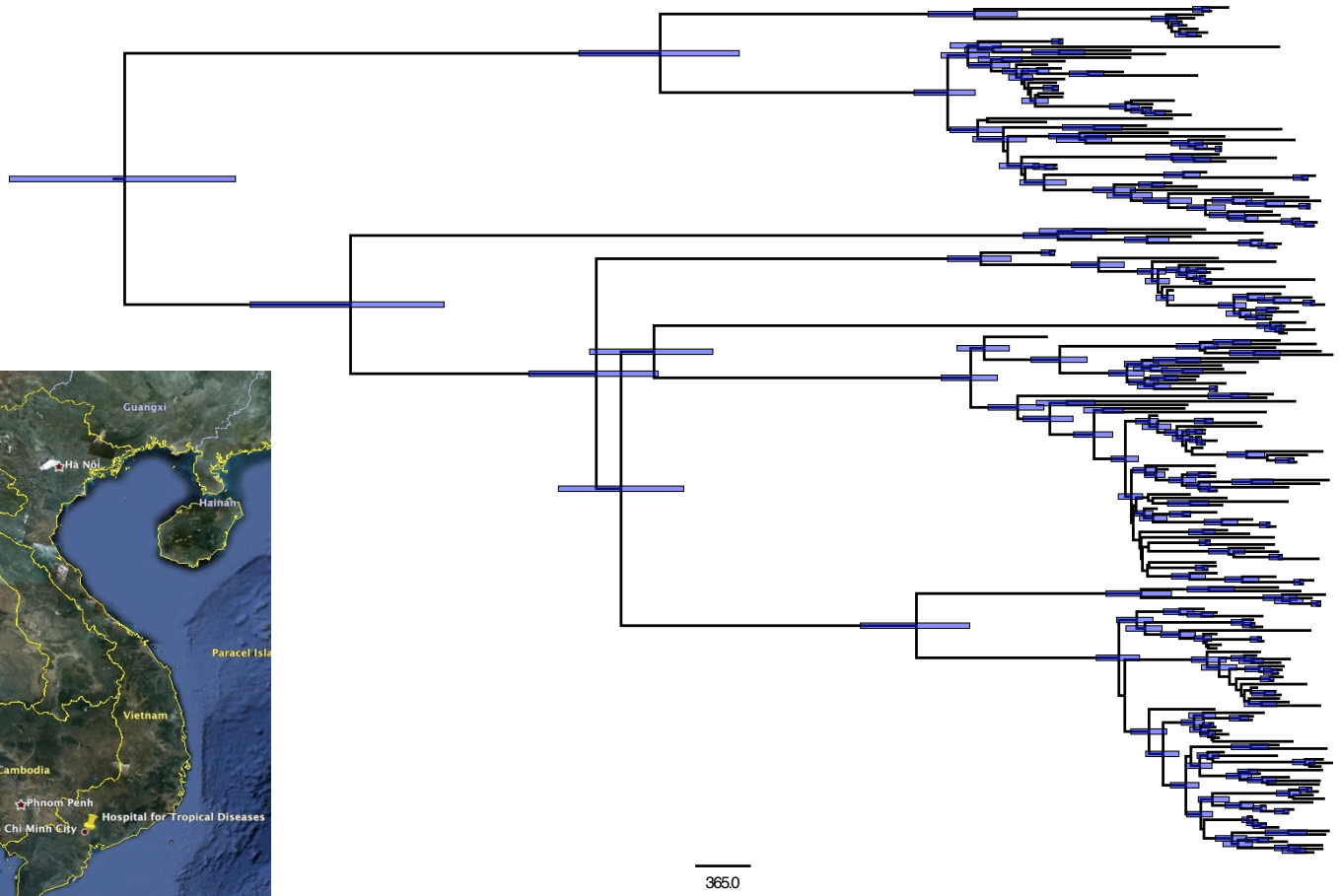
- Rather than explicitly sampling migration histories, we probabilistically track the movement of each lineage.

- We can then write pairwise coalescent rates in terms of lineage state probabilities p_{ik}

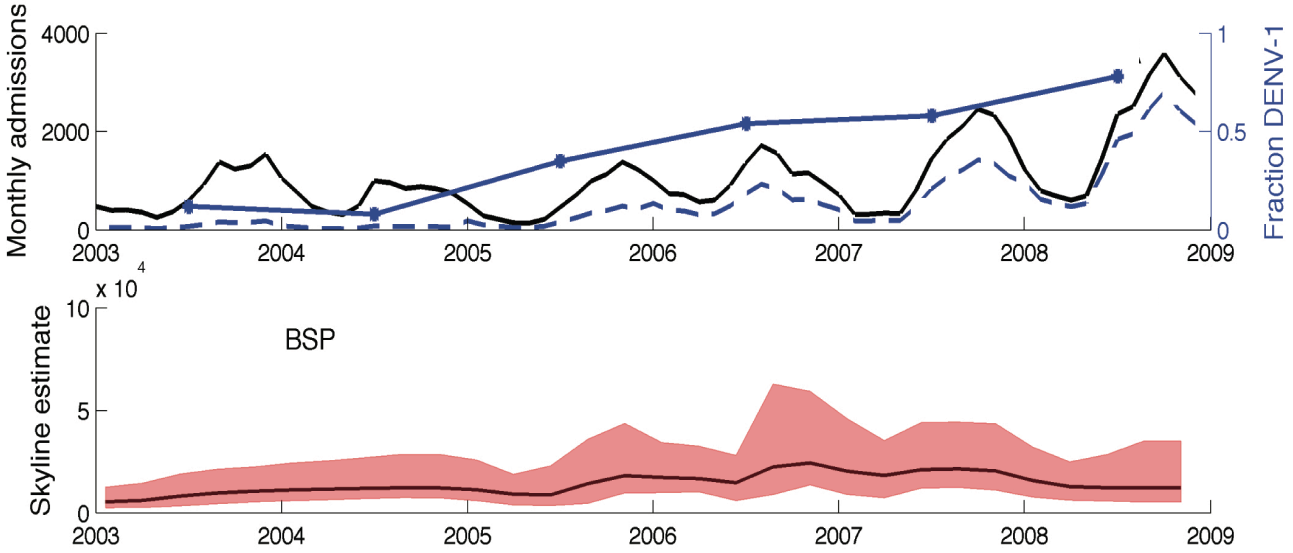
$$\lambda_{ij} = \sum_{k}^{m} \sum_{l}^{m} \frac{f_{kl}}{y_k y_l} \left(p_{ik} p_{jl} + p_{il} p_{jk} \right)$$

- We track how the lineage state probabilities evolve backwards in time using a system of master equations. This assumes each lineage evolves independently of other lineages.

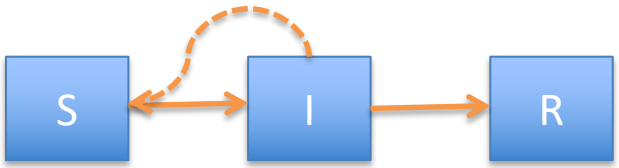
Dengue in southern Vietnam



Standard model estimates



SIR with seasonality



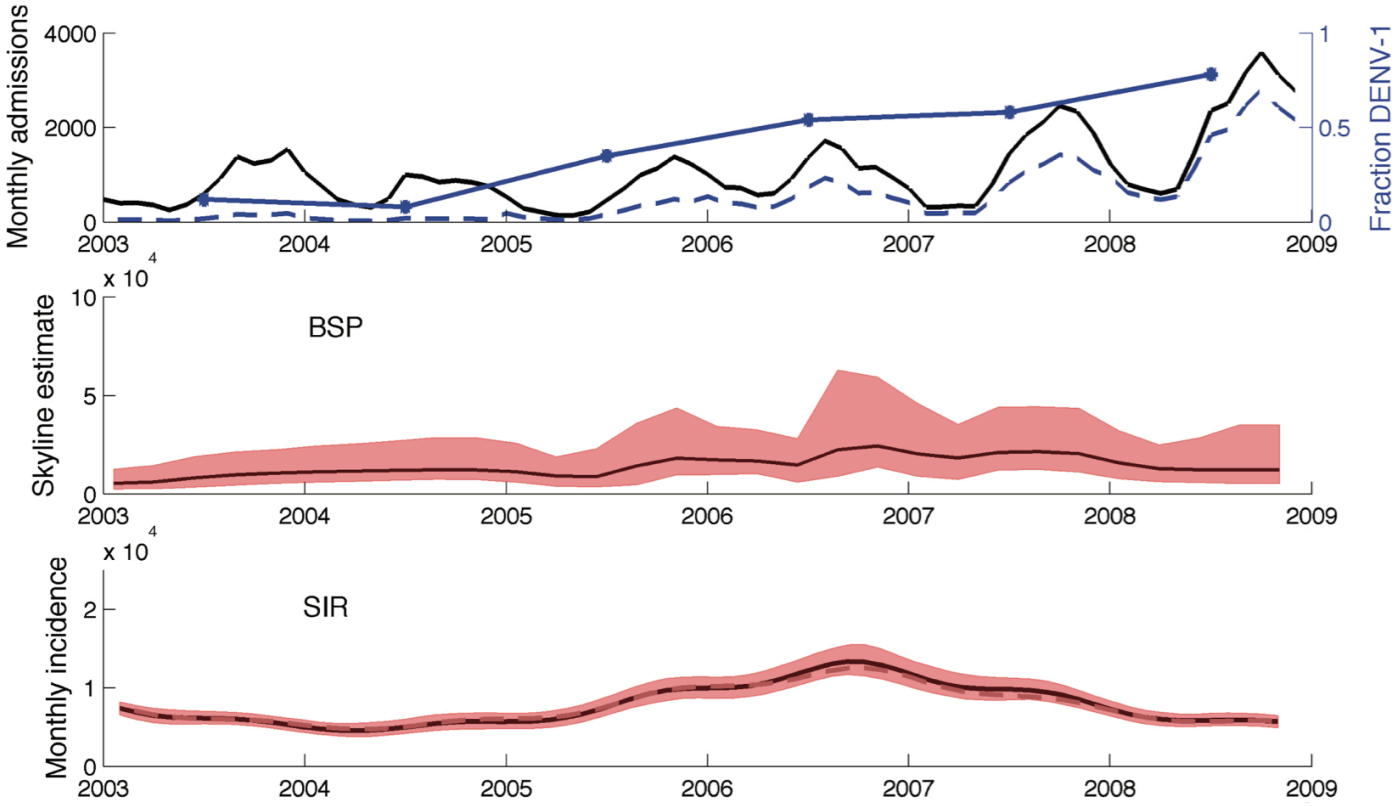
$$\frac{dS}{dt} = \mu N - \beta(t)\frac{S}{N}I - \mu S$$
$$\frac{dI}{dt} = \beta(t)\frac{S}{N}I - I(\mu + \nu)$$
$$\frac{dR}{dt} = \nu I - \mu R$$

$$\beta(t) = \bar{\beta}\left(1 + \alpha\cos\left(\frac{t+\delta}{2\pi}\right)\right)$$

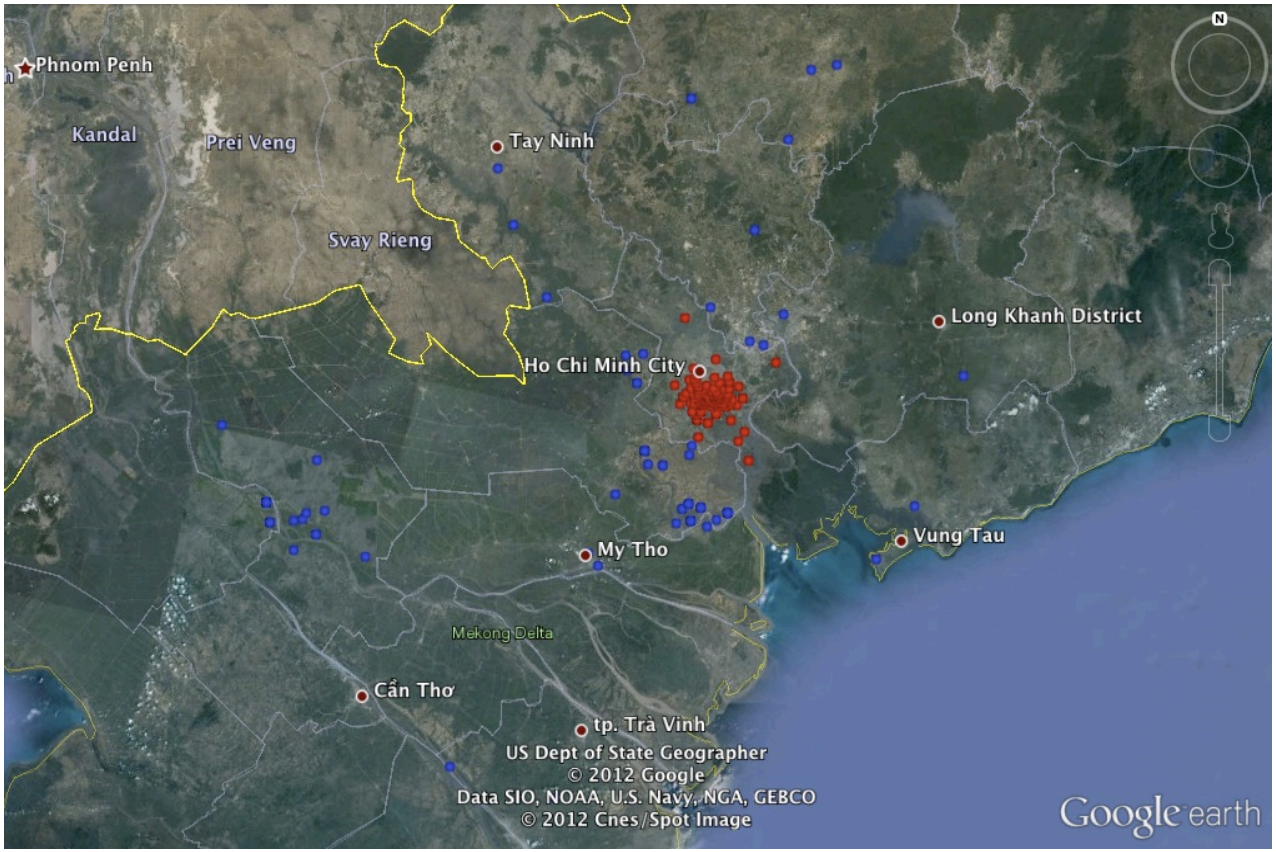
Coalescent model:

$$\lambda = \frac{2\beta(t)\frac{S}{N}}{I}$$

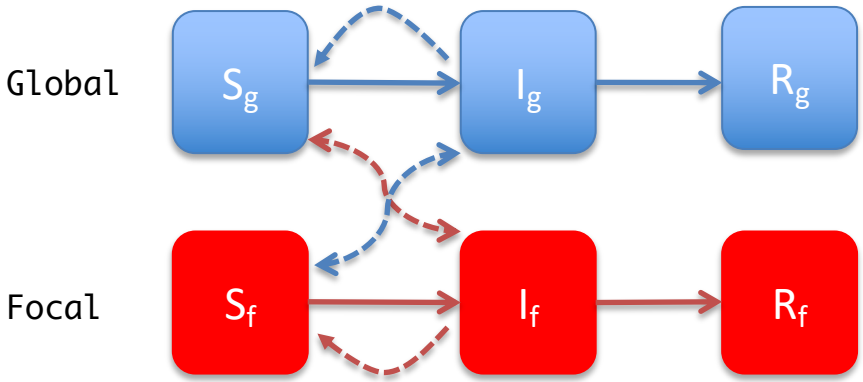
SIR model with seasonality

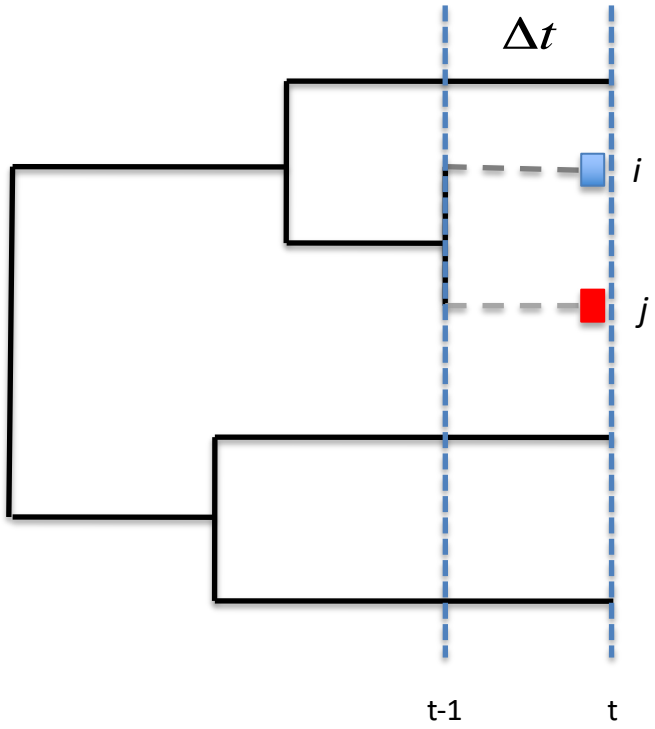


Spatial structure



Spatial SIR model

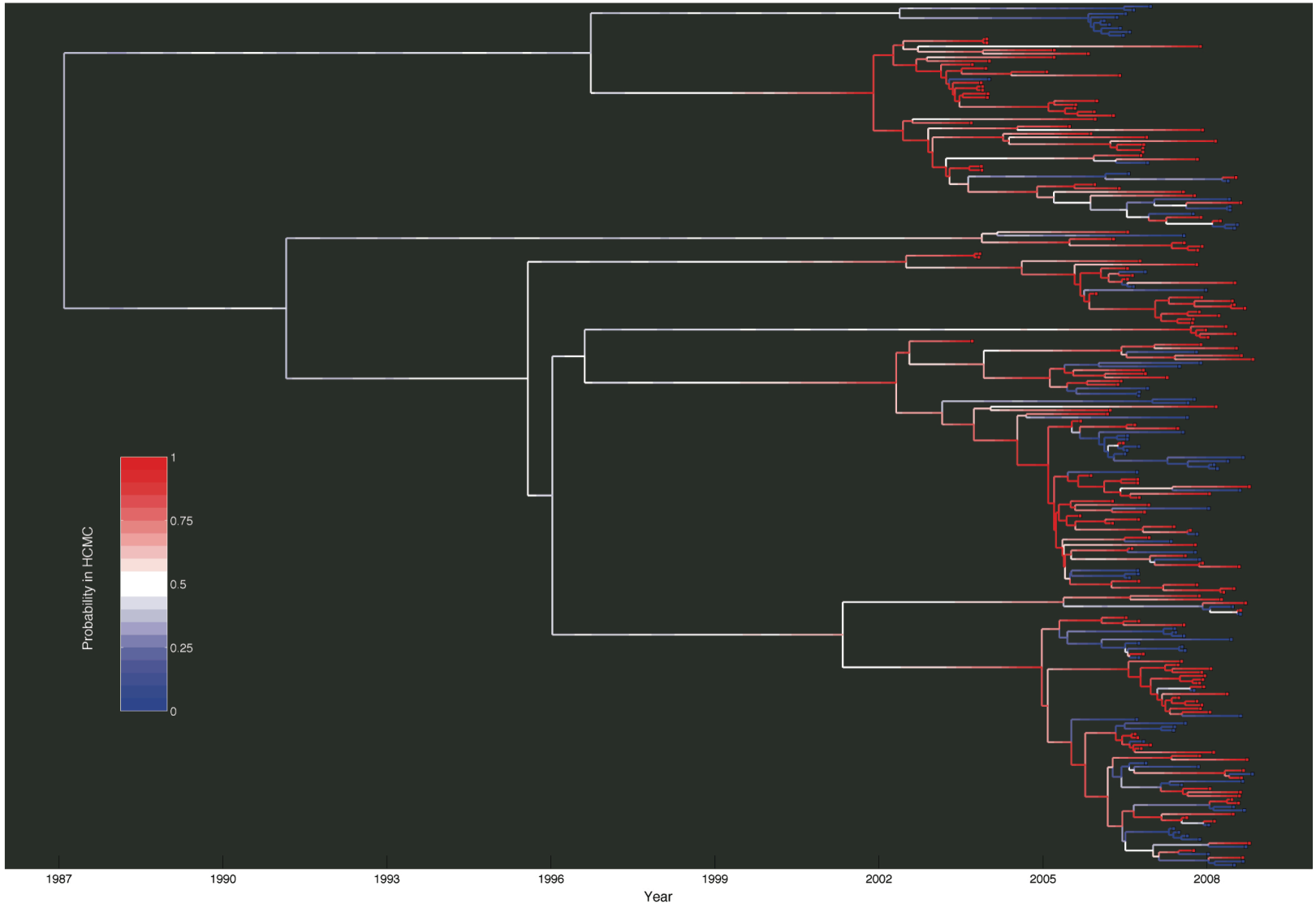




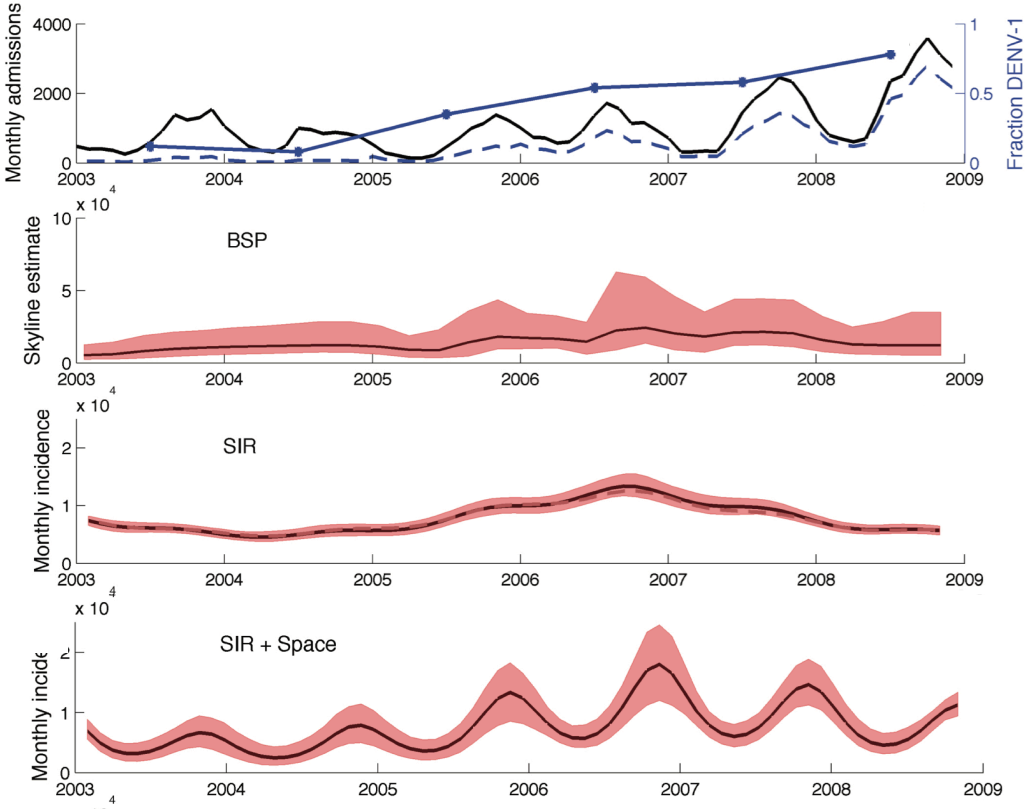
Structured coalescent model:

$$\lambda_{ij} = \sum_{k}^{m} \sum_{l}^{m} \frac{\beta_{kl} \frac{S_l}{N_l} I_k}{I_k I_l} \left(p_{ik} p_{jl} + p_{il} p_{jk} \right)$$

Spatial SIR estimates



Spatial SIR estimates



Rasmussen et al., MBE, 2014

PhyDyn: Epi modeling in BEAST

RESEARCH ARTICLE

Bayesian phylodynamic inference with complex models

Erik M. Volz*, Igor Siveroni

Department of Infectious Disease Epidemiology and the MRC Centre for Global Infectious Disease Analysis, Imperial College London, London, United Kingdom

* e.volz@imperial.ac.uk

$$\frac{d}{dt}I_1 = \frac{S}{N}(\beta_1 I_1 + \beta_2 I_2) - \gamma_1 I_1$$

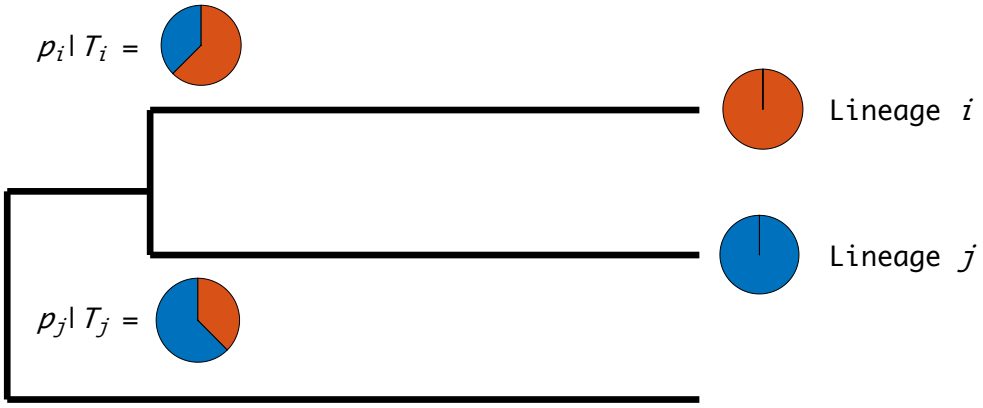
$$\frac{d}{dt}I_2 = \gamma_1 I_1 - \gamma_2 I_2$$

```
<model spec='PopModelODE' id='twodeme'  popParams='@initialValues'
          modelParams='@rates' evaluator="compiled">
  <definition spec="Definition"> N = S + I1 + I2 </definition>
  <matrixeq type="birth" origin="I1" destination="I1"> beta1*I1*S/N  </matrixeq>
  <matrixeq type="birth" origin="I2" destination="I1"> beta2*I2*S/N </matrixeq>
 <matrixeq type="migration" origin="I1" destination="I2"> gamma1*I1</matrixeq>
 <matrixeq type="death" origin="I1"> gamma1*I1 </matrixeq>
 <matrixeq type="death" origin="I2"> gamma2*I2 </matrixeq>
 <matrixeq type="nondeme" origin="S">
           b*S - (beta1*I1+ beta2*I2)*S/N
 </matrixeq>
</model>
```

Based on template by Igor Siveroni

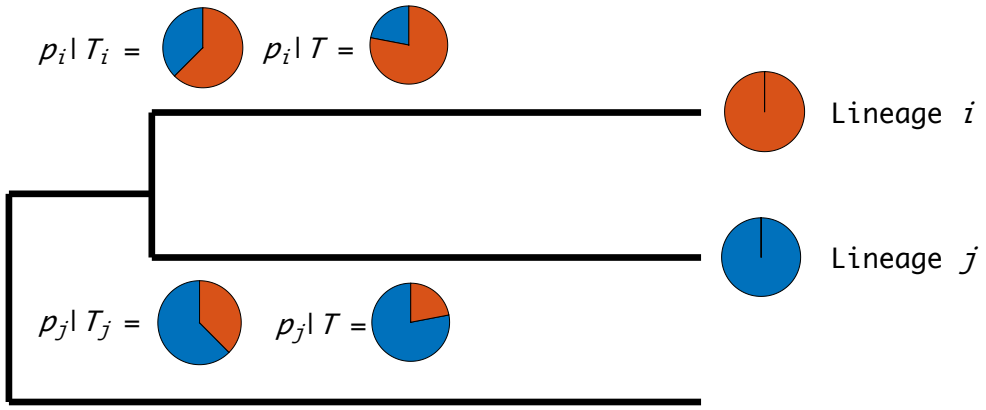
Tracking lineage state probabilities

- Lineage state probabilities can differ between methods due to assumptions of independence among lineages.



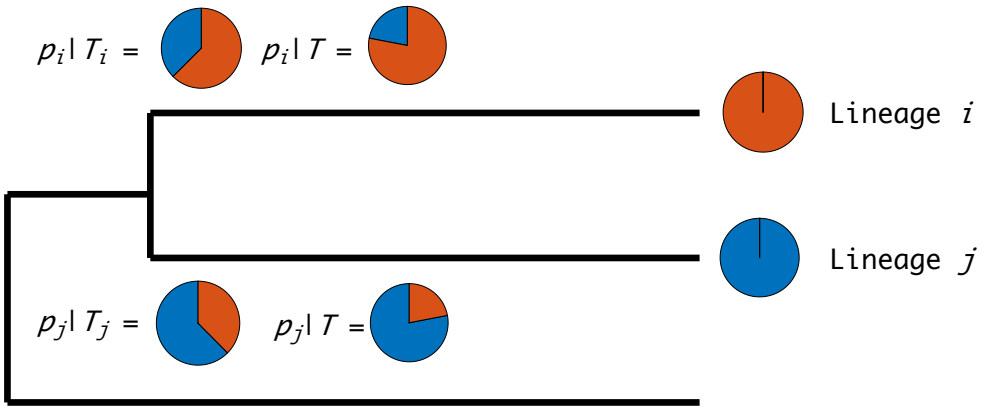
Tracking lineage state probabilities

- Lineage state probabilities can differ between methods due to assumptions of independence among lineages.



Tracking lineage state probabilities

- Lineage state probabilities can differ between methods due to assumptions of independence among lineages.



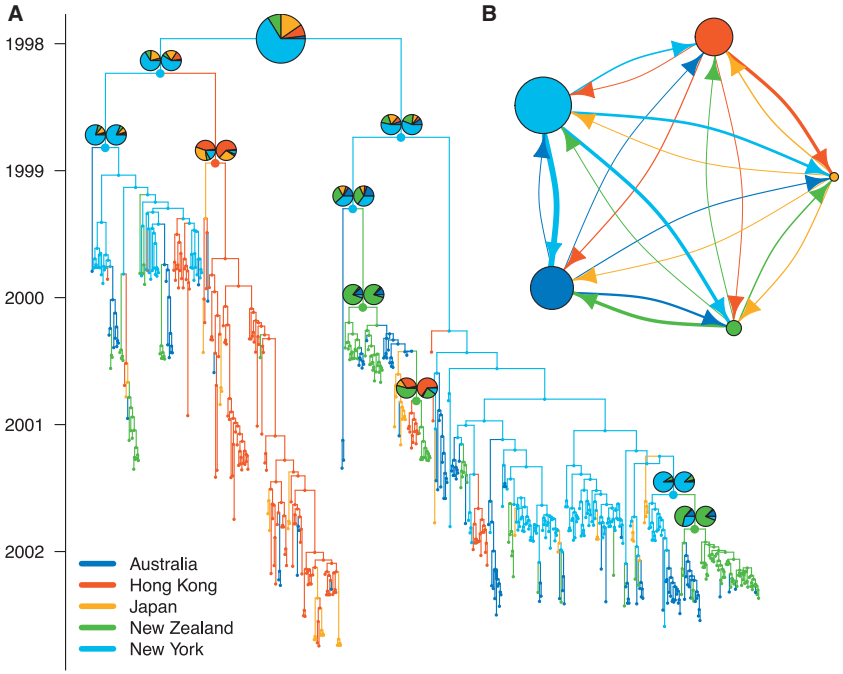
- Assuming independence can bias inferences of ancestral states and migration rates when migration is slow relative to coalescent rates and when sampling is highly asymmetric (Müller *et al.*, MBE, 2018).

MASCOT

- MASCOT uses an improved approximation to the structured coalescent model that does not assume lineage independence to track lineage states probabilistically rather than sampling migration histories.

MASCOT: parameter and state inference under the marginal structured coalescent approximation

Nicola F. Müller[1,2,*], David Rasmussen[1,2,3,4] and Tanja Stadler[1,2,*]

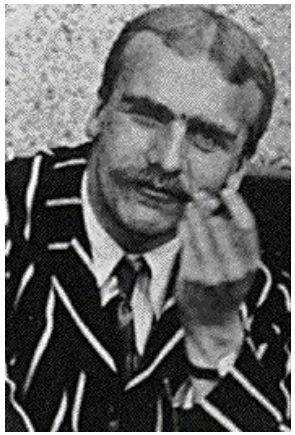


Why are there so many structured models in BEAST 2?

- Discrete-trait models (Lemey *et al.*, 2009)
- MultiTypeTree (Vaughan, *et al.*, 2014)
- BASTA (De Maio et al., 2015)
- MASCOT (Müller *et al.*, 2018)
- PhyDyn (Volz & Siveroni, 2018)
- BDMM (Kühnert *et al.*, 2016)
- MSBD (Barido-Sottani et al., 2019)

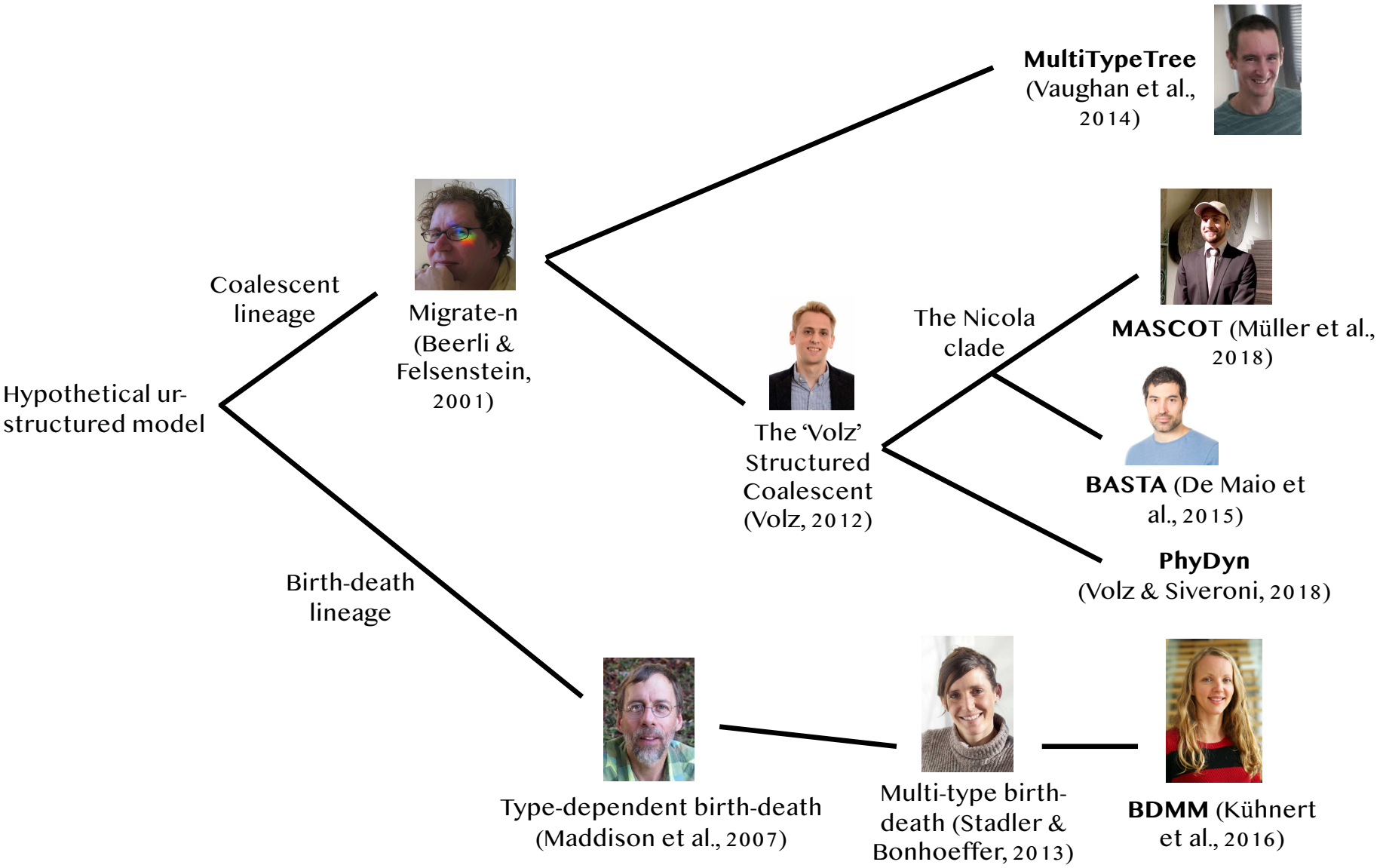
Structured population models

Asked by theologians what he had concluded about The Developers of BEAST, J.B.S. Haldane (1892-1964) remarked:



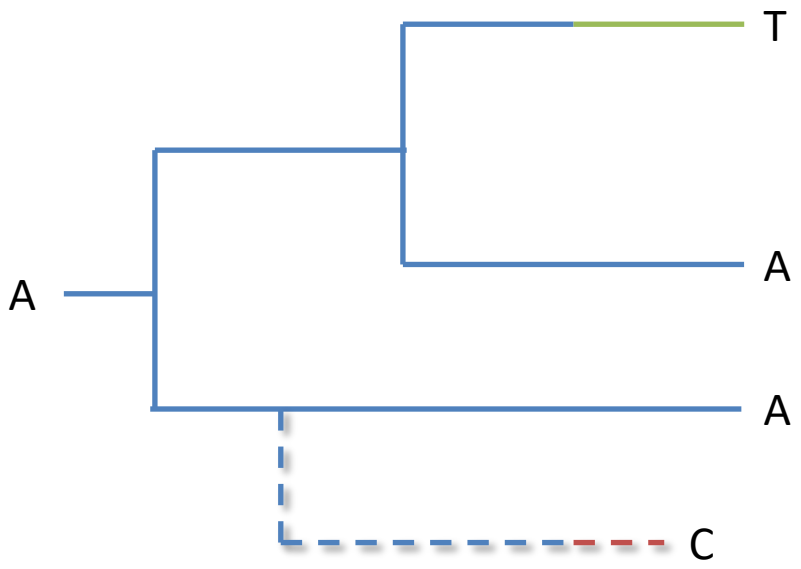
“An inordinate fondness for structured population models”

Decent with modification: A phylogeny of structured population models in BEAST 2

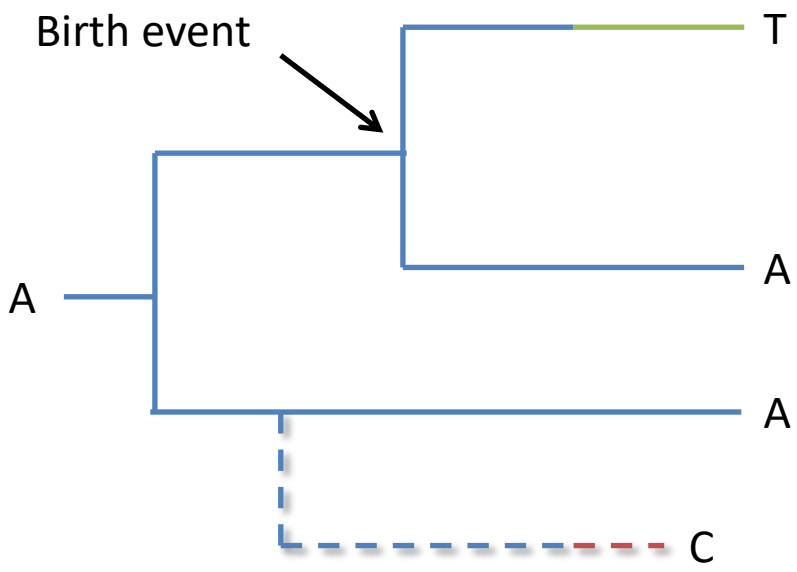


The multi-type birth-death model

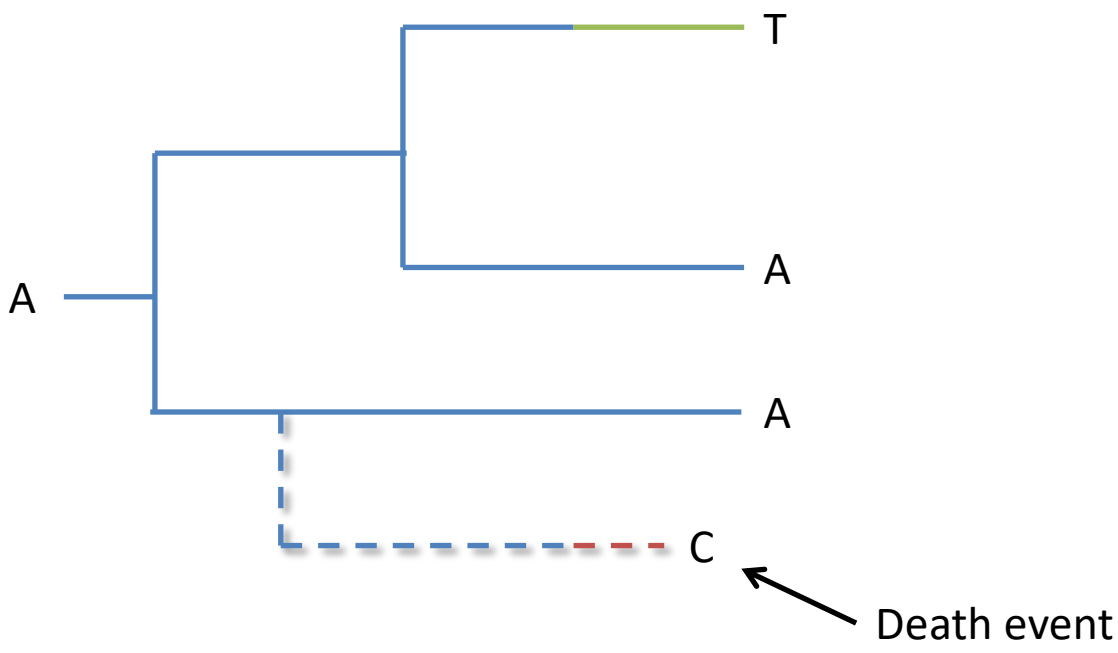
- Allows for type transitions along lineages under a stochastic birth-death model with incomplete sampling (Stadler & Bonhoeffer, 2013).



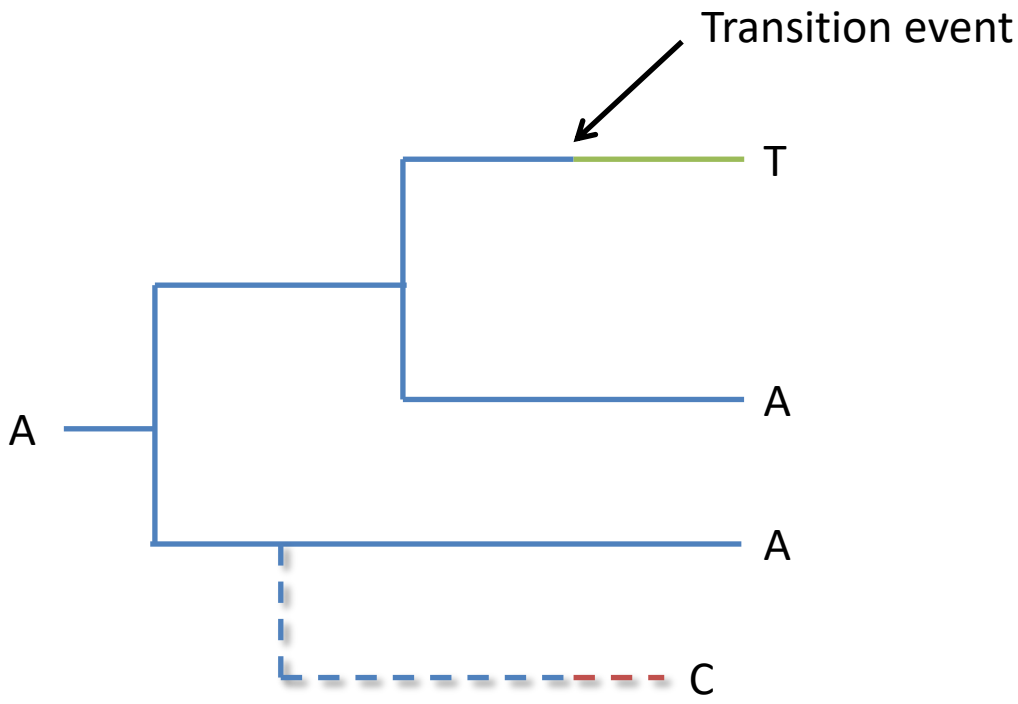
The multi-type birth-death model



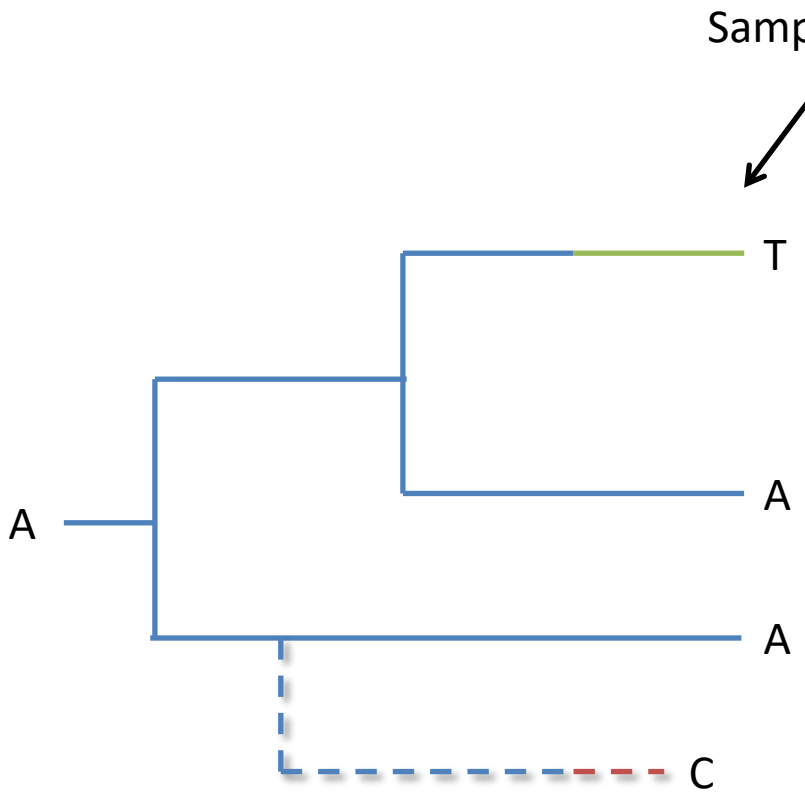
The multi-type birth-death model



The multi-type birth-death model

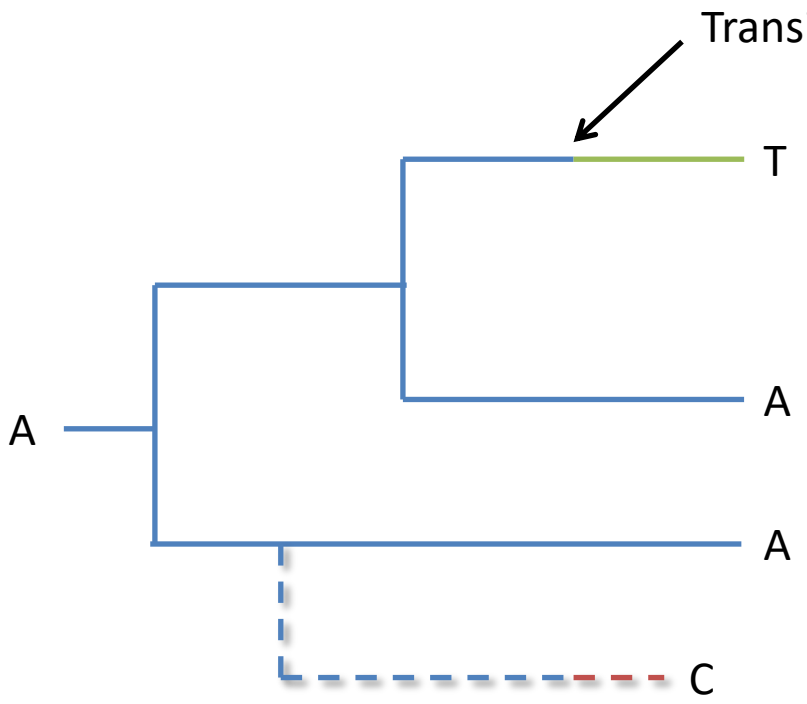


The multi-type birth-death model



Sampling rates and proportions can differ between populations

The multi-type birth-death model



$$M = \begin{bmatrix} 0 & m_{1,2} & \cdots & m_{1,n} \\ m_{2,1} & 0 & \cdots & m_{2,n} \\ \vdots & \vdots & \ddots & \vdots \\ m_{n,1} & m_{n,2} & \cdots & 0 \end{bmatrix}$$

BDMM (Kühnert *et al.*, 2016)

- Allows for inference of type-dependent birth-death rates and transition rates under a MTBD model.

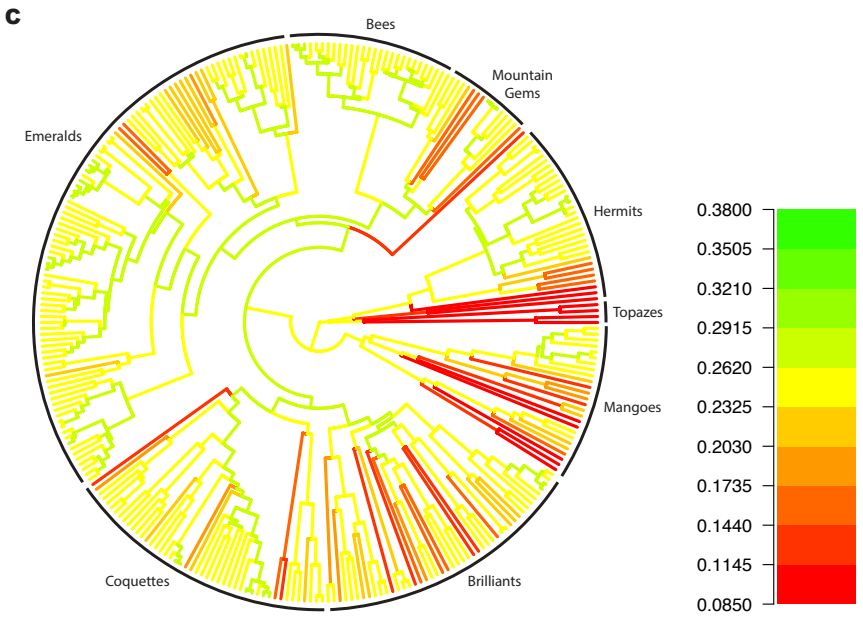
- Accounts for demographic stochasticity under a linear (exponential growth) BD model

- Allows for time-varying parameters under a piece-wise constant ‘Skyline’ model.

MSBD (Barido-Sottani et al., 2019)

- Estimates lineage-specific birth and death rates from a phylogeny **without** given tip states under a MTBD model.

- MSBD estimates the number of different types and their placement on the tree.



Comparison of structured models

Model	Family	Biased by uneven sampling	Time-varying parameters	Stochastic pop dynamics	GLM integration
DTA	Substitution	Very biased	No	No	Yes
MultiTypeTree	Struct Coal	Low bias	No	No	No
BASTA	Struct Coal	Some bias	No	No	No
MASCOT	Struct Coal	Low bias	No	No	Yes
PhyDyn	Struct Coal	Low bias	Yes	No	No
BDMM	Multi-type BD	Low bias	Yes	Yes	No
MSBD	Multi-type BD	Low bias	No	Yes	No

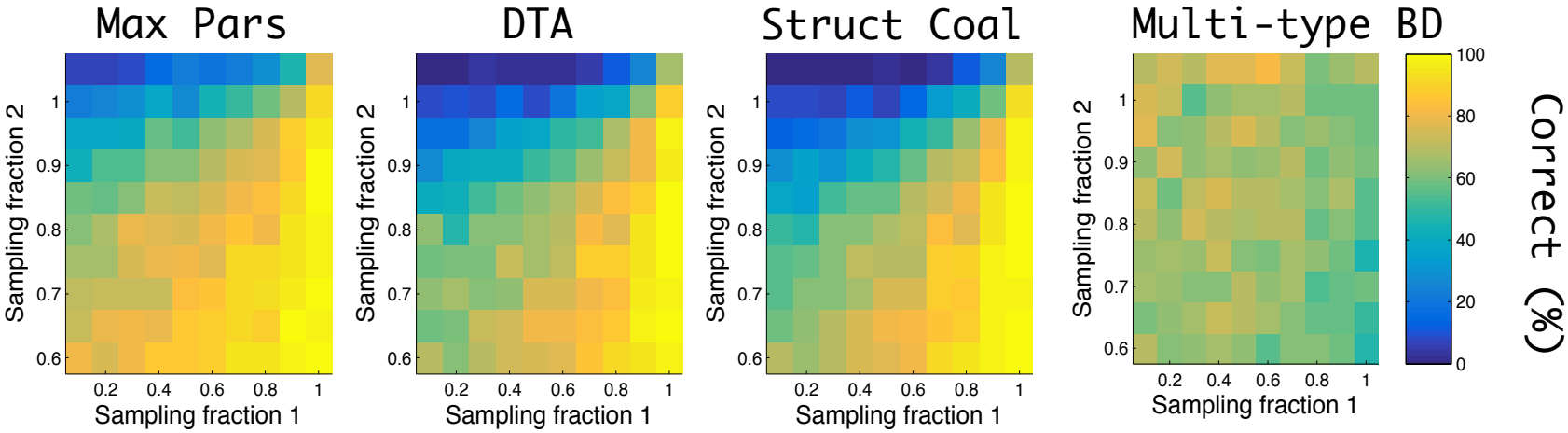
Strategies for biased sampling

- Subsampling from oversampled populations in order to make sampling proportional across populations is common.

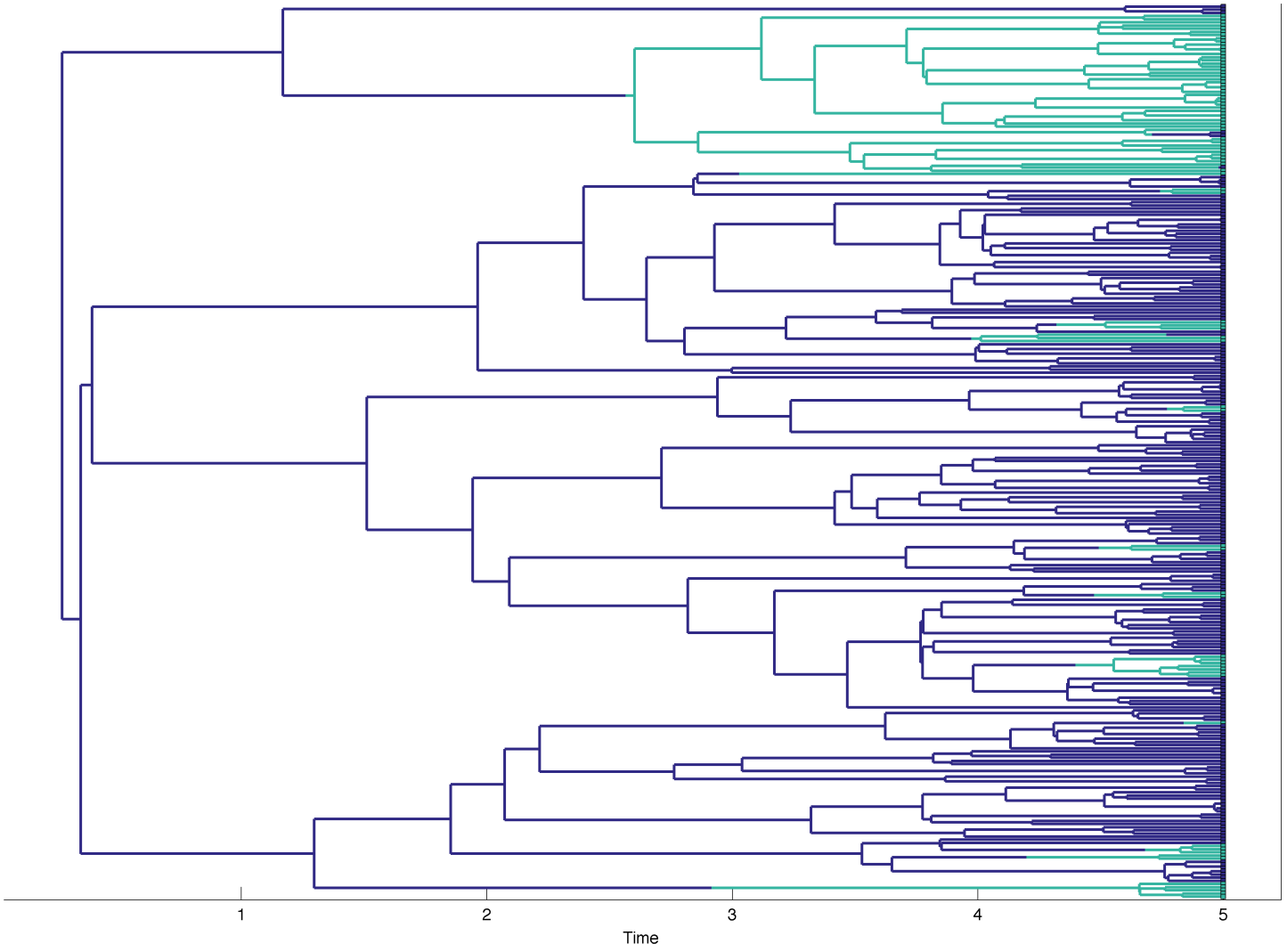
- Arguably, there's no reason to do this under structured coalescent and birth-death models because sampling can be uneven between populations as long as sampling is random within populations.

- BUT all methods can be biased by extremely uneven sampling as the tree will no longer be "representative" of the general population's history.

Root state inference with sampling bias



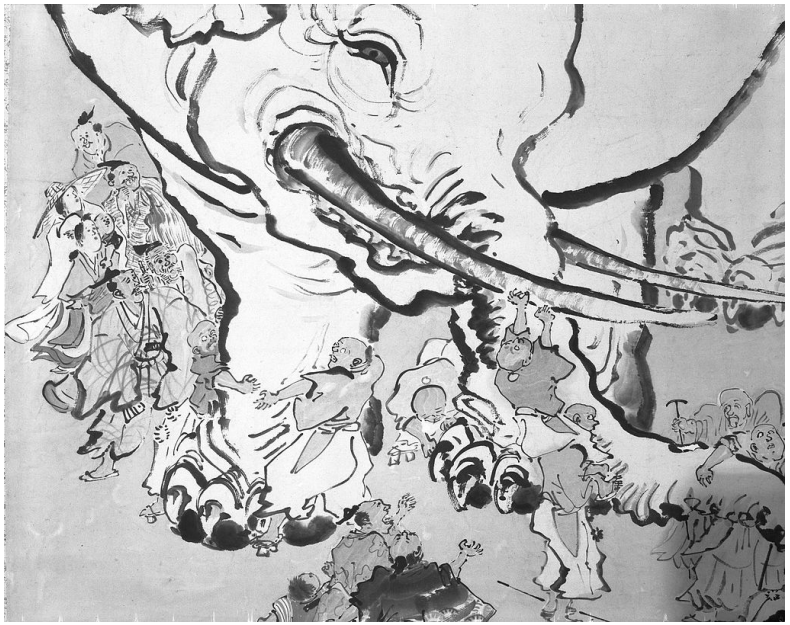
Slow migration tree



Sampling bias

When the blind men had each felt a part of the elephant, the king went to each of them and said to each: ‘Well, blind man, have you seen the elephant? Tell me, what sort of thing is an elephant?’

—Buddhist proverb



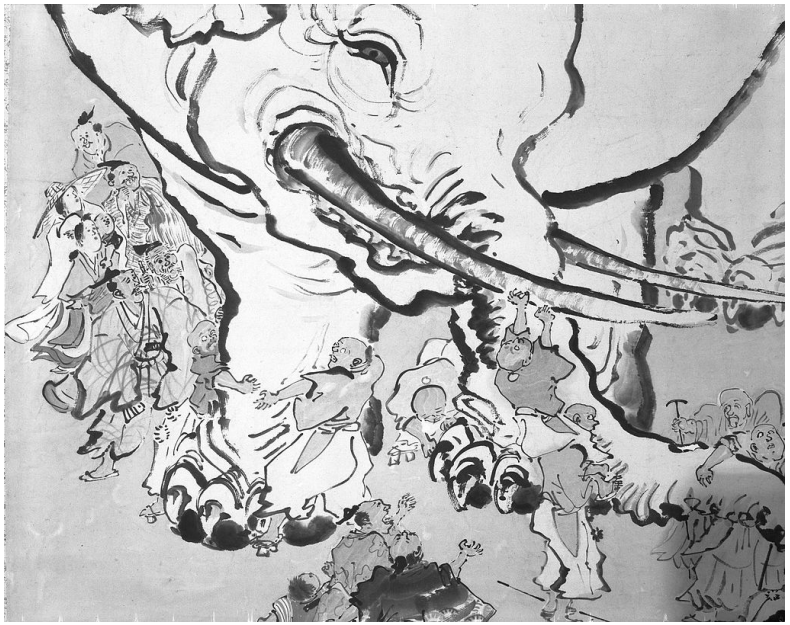
Ohara Donshu

Sampling bias

When the blind men had each felt a part of the elephant, the king went to each of them and said to each: ‘Well, blind man, have you seen the elephant? Tell me, what sort of thing is an elephant?’

—Buddhist proverb

How do we know if a sampled tree is “representative” of a population’s history?



Ohara Donshu