An overview of structured population models in BEAST 2

David Rasmussen

Department of Entomology and Plant Pathology

Bioinformatics Research Center

North Carolina State University

Taming the BEAST, Squamish

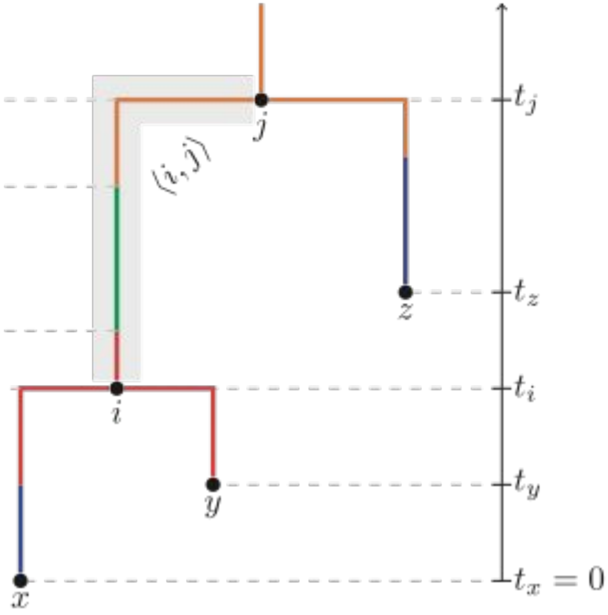
August 17th, 2023

Why are there so many structured models?

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

What is a structured model?

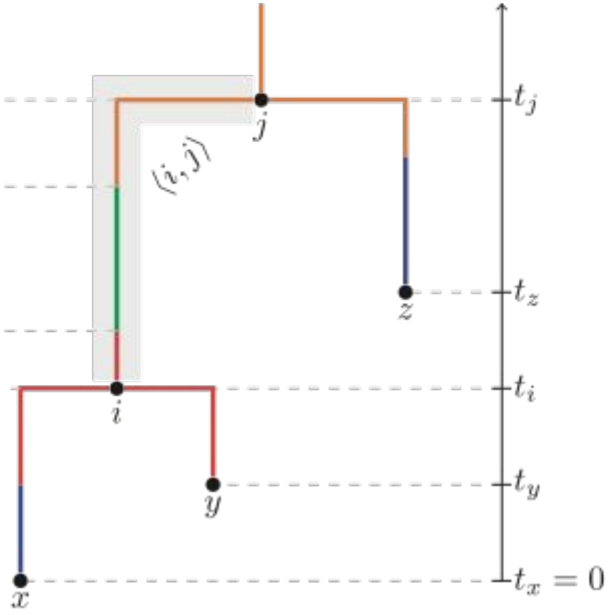
- Any model where lineages can reside in different populations or states



Vaughan *et al.* (2014)

What is a structured model?

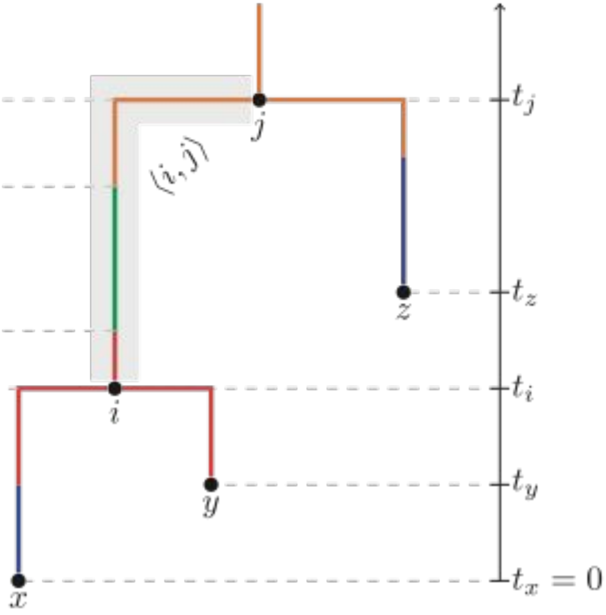
- Any model where lineages can reside in different populations or states
- States can represent different populations, geographic locations, infectious states, character traits, ect.



Vaughan *et al.* (2014)

What is a structured model?

- Any model where lineages can reside in different populations or states
- States can represent different 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 models?

- We may be interested in estimating transition rates (e.g. migration rates) between different populations or states
- We may be interested in reconstructing ancestral states like the geographic location of a lineage (e.g. origins of an epidemic)
- Even if we are not directly interested in structure for its own sake, structure can confound our inferences about other variables of interest (e.g. pop size estimates under coalescent models)

Three main types of structured models

1. Discrete-trait CTMC (DTA) models
2. Structured coalescent models
3. Multi-type birth-death models

Three main types of structured models

1. Discrete-trait CTMC (DTA) models
2. Structured coalescent models
3. Multi-type birth-death models

Discrete trait models (DTA)

Discrete trait methods model transitions between locations as a **continuous time Markov chain**, i.e. the same way we model sequence evolution at a single site.

Instead of a substitution rate matrix, we have a migration rate matrix *M*:

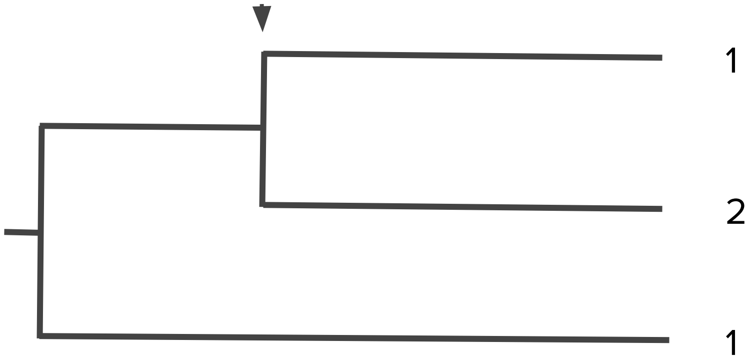
$$M = \begin{bmatrix} -\sum_{i\neq 1}^{n} m_{1,i} & m_{1,2} & \cdots & m_{1,n} \\ m_{2,1} & -\sum_{i\neq 2}^{n} m_{2,i} & \cdots & m_{2,n} \\ \vdots & \vdots & \ddots & \vdots \\ m_{n,1} & m_{n,2} & \cdots & -\sum_{i\neq n}^{n} m_{n,i} \end{bmatrix}$$

Because we model migration the same we model mutations, these models are sometimes referred to as “**mugation” models**.

Computing ancestral state probabilities

$$p(t) = e^{Mt}p(0)$$

$$p(t) = \begin{bmatrix} ? \\ ? \end{bmatrix}$$



$$p(0) = \begin{bmatrix} 1 \\ 0 \end{bmatrix}$$

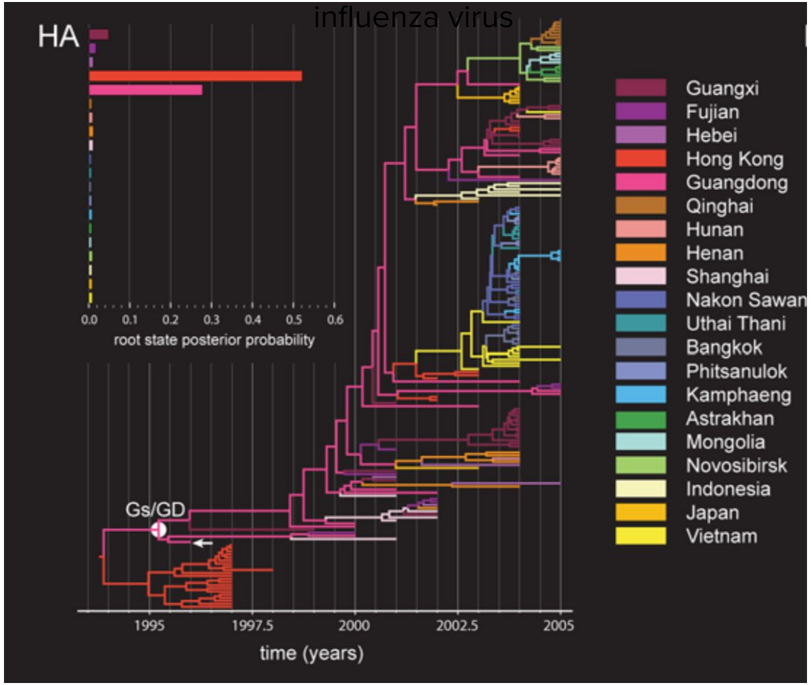
We can easily compute ancestral state probabilities under a CTMC given our migration rate matrix ***M*** and the time elapsed along a branch ***t***.

Bayesian phylogeography with DTA

Bayesian methods based on CTMC efficiently reconstruct migration histories even for relatively large trees and many different sampling locations.

Posterior probabilities directly quantify our uncertainty about ancestral locations.

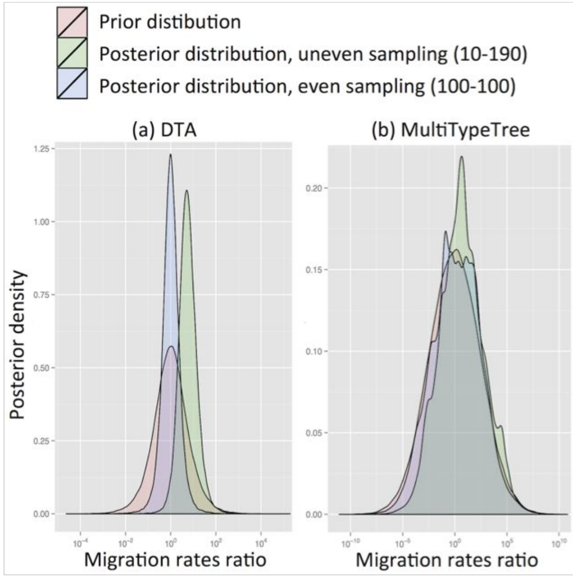
MCC tree for avian H5N1 influenza virus

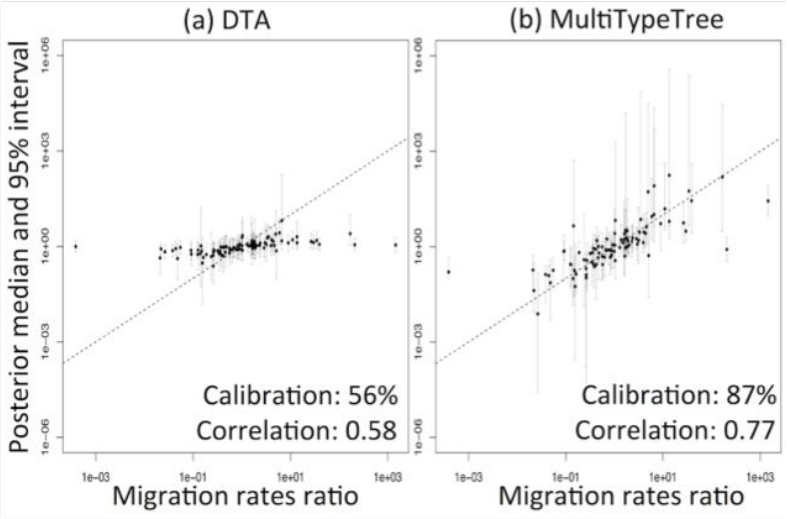


Lemey *et al.* (2009)

Uneven sampling strongly biases DTA

Uneven sampling strongly biases DTA because it treats sampling as informative about the migration process.





De Maio *et al.* (PLoS Genetics, 2015)

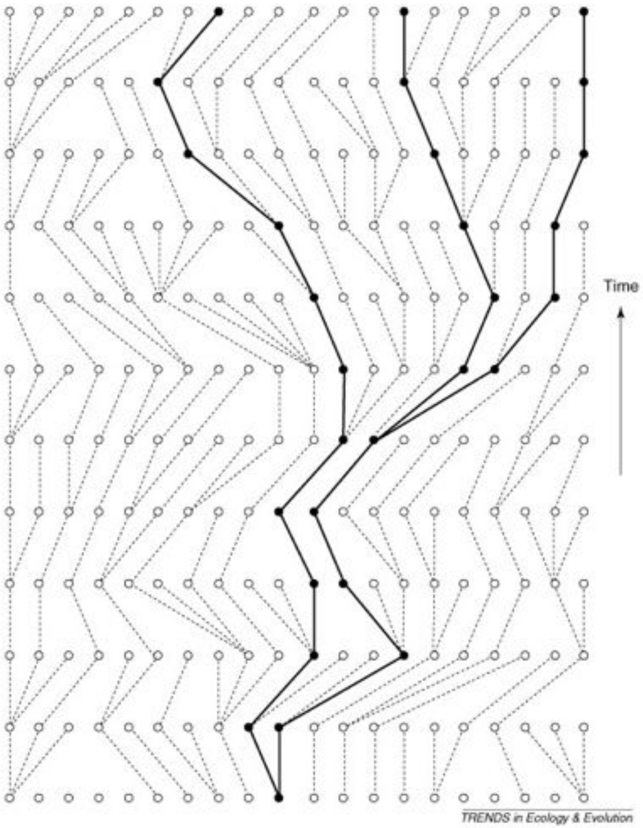
Three main types of structured models

1. Discrete-trait CTMC (DTA) models
2. Structured coalescent models
3. Multi-type birth-death models

Back to the coalescent!

The probability of two lineages coalescing per generation is:

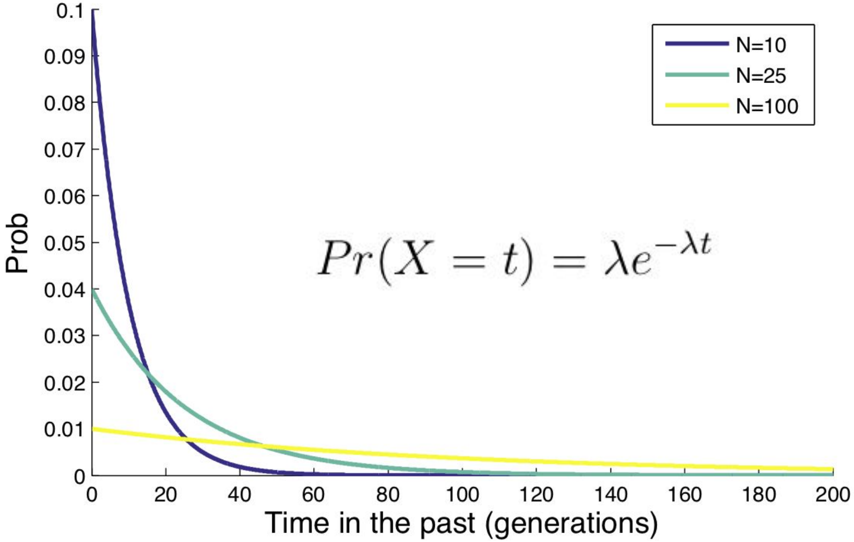
$$p_{coal} = \frac{1}{N}$$



Kuhner *et al.* (2008)

Basic coalescent theory

The waiting time for a pair of lineages to coalesce is exponentially distributed.



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

Now with more than two lineages

With *k* lineages present, the coalescent rate becomes:

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

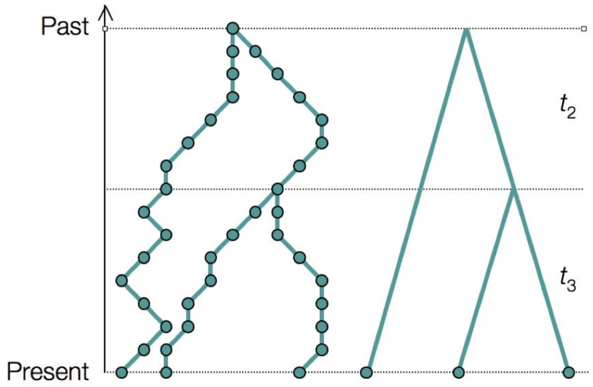
The binomial coefficient gives the total number of lineage pairs that could have coalesced:

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

The coalescent likelihood

For a tree with *n* samples and *n-1* coalescent events we can compute the likelihood of the tree as:

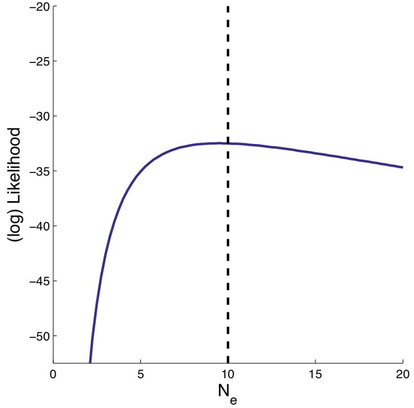
$$L(T|N_e) = \frac{1}{N_e^{(n-1)}} \prod_{k=2}^{n} \exp\Big(-\frac{\binom{k}{2}}{N_e} t_k\Big)$$

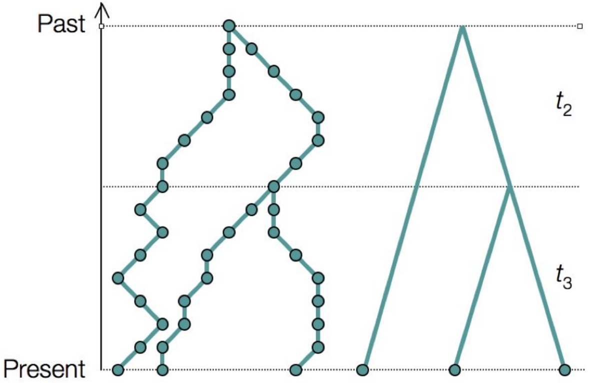


Coalescent-based inference

We can therefore infer demographic parameters like N_e from a known phylogeny.

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





The problem with population structure

Standard coalescent models assume that all lineages in the tree are **exchangeable**.

Exchangeability here means that any lineage is equally likely to coalesce with any other lineage in the tree.

Many forms of population structure violate this key assumption.

The structured coalescent

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

Each lineage pair is allowed to coalesce at a different rate λ_{ij} based on the locations of lineages *i* and *j*:

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

The Migrate-n model

A structured coalescent model with migration between *n* subpopulations or demes

Models 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*.

Beerli and Felsenstein (2001)

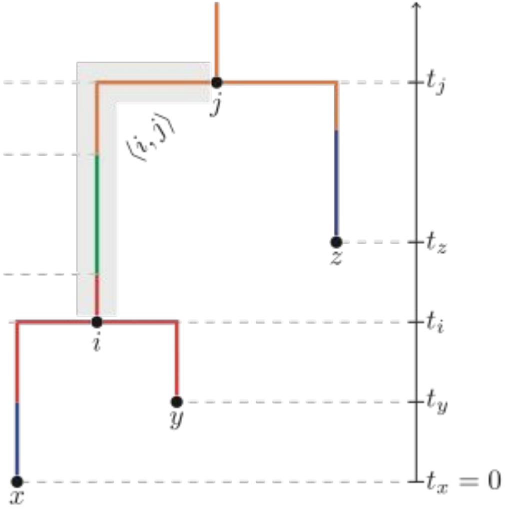
The Migrate-n model

The pairwise rate of coalescent between two lineages *i* and *j* will depend on their population states *k* and *l*:

$$\lambda_{ij} = \begin{cases} \frac{1}{N_e^k} & \text{if } k = l \\ 0 & \text{if } k \neq l. \end{cases}$$

Lineages sampled in different populations therefore need to migrate back to the same population before they can coalesce.

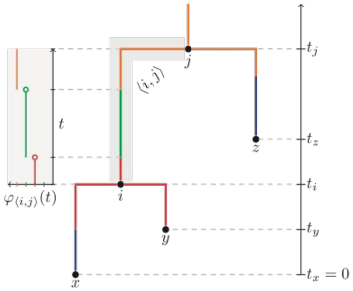
But we now need to infer **migration histories** to know the location of each lineage at any point in the past.

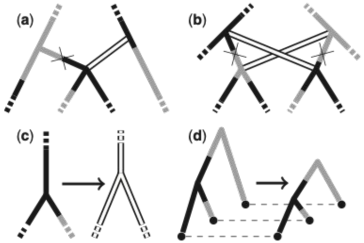


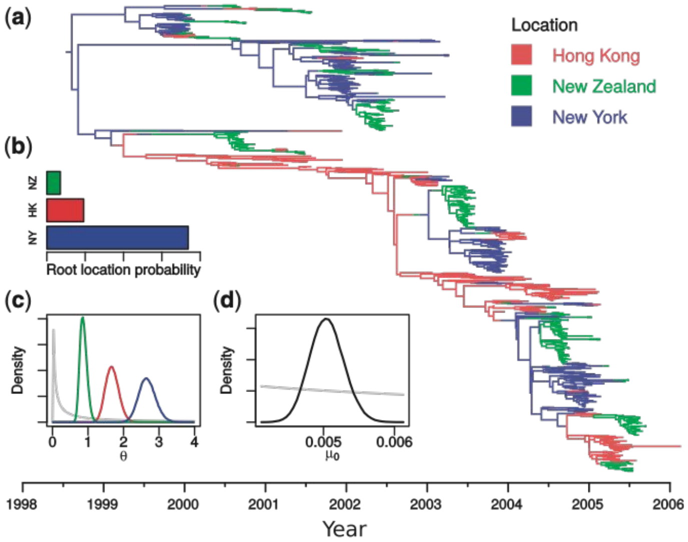
Beerli and Felsenstein (2001)

Migrate-N and MultiTypeTree

MCMC implementations of the structured coalescent like MIGRATE and MultiTypeTree (Vaughan *et al.*, 2014) sample migration histories on trees







Vaughan *et al.* (2014)

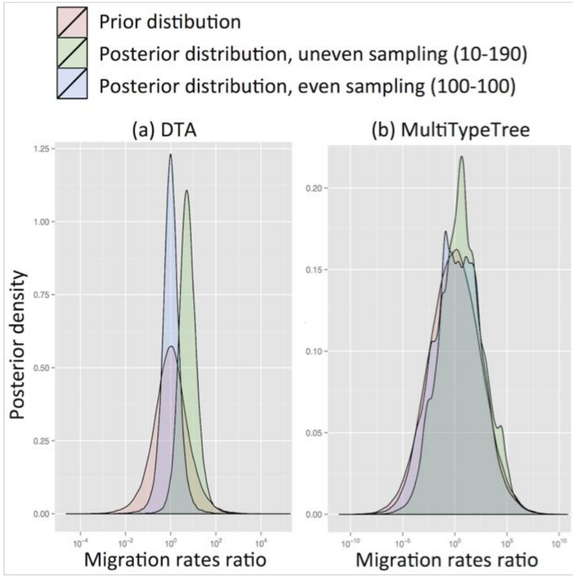
DTA versus the structured coalescent

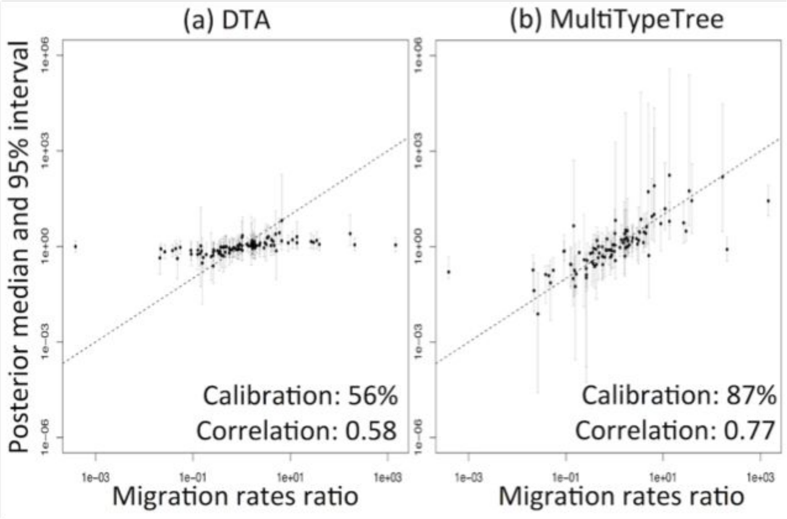
The structured coalescent has become an attractive alternative to discrete-trait analysis (DTA) for phylogeography.

Because the structured coalescent conditions on sampling locations instead of treating them as informative about migration like DTA, the SC is much more robust to uneven sampling.

DTA versus the structured coalescent

Uneven sampling strongly biases DTA but not the structured coalescent.





De Maio *et al.* (PLoS Genetics, 2015)

DTA versus the structured coalescent

Structured coalescent models improve statistical performance but are fundamentally limited by the need to sample migration histories on trees.

This does not allow for very efficient MCMC sampling due to strong correlations between the migration histories and model parameters. Generally limited to about 5 or 6 states and trees with < 1000 tips.

But what if there was a way to efficiently “integrate over” migration histories and therefore average over all possible paths a lineage could have taken?

The Volz (2012) Structured Coalescent

Rather than sampling migration histories, we can probabilistically track the movement of each lineage back through time.

We can then write pairwise coalescent rates in terms of lineage state probabilities p_{ik}. Assuming lineage pairs can only coalesce if they're in the same population:

$$\lambda_{ij} = \sum_{k}^{m} \frac{p_{ik} p_{jk}}{N_k}$$

The Volz (2012) Structured Coalescent

Rather than sampling migration histories, we can probabilistically track the movement of each lineage back through time.

We can then write pairwise coalescent rates in terms of lineage state probabilities p_{ik}. Assuming lineage pairs can only coalesce if they're in the same population:

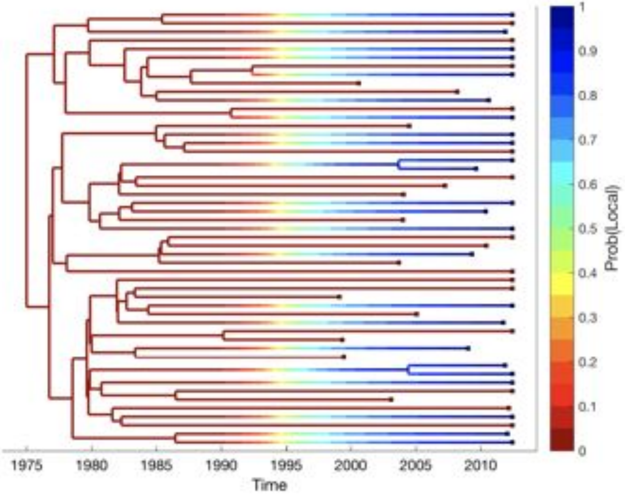
$$\lambda_{ij} = \sum_{k}^{m} \frac{p_{ik} p_{jk}}{N_k}$$

The theory in Volz (2012) is actually more general and allows birth/coalescent events to occur between populations at rate f_{kl}:

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

The Volz (2012) Structured Coalescent

Lineage state probabilities p_{ik} can then be tracked backwards in time using a system of master equations (ODEs) based on the transition rates g_{kl}:

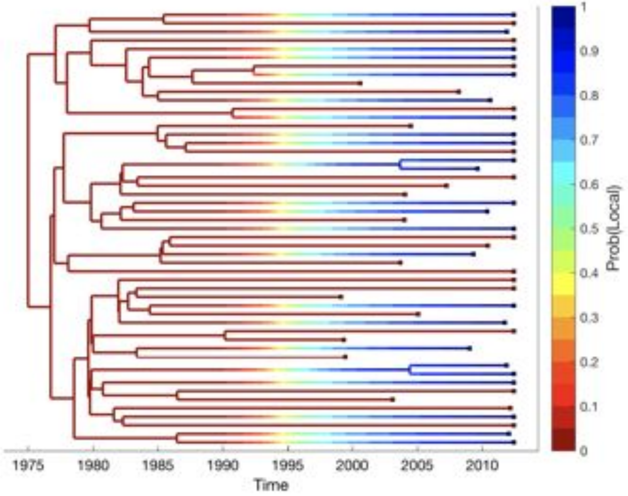


Rasmussen et al., (Virus Evo., 2018)

$$\frac{d}{dt}p_{ik} = \sum_{k}^{m} \left(p_{il}g_{kl} - p_{ik}g_{lk}\right)$$

The Volz (2012) Structured Coalescent

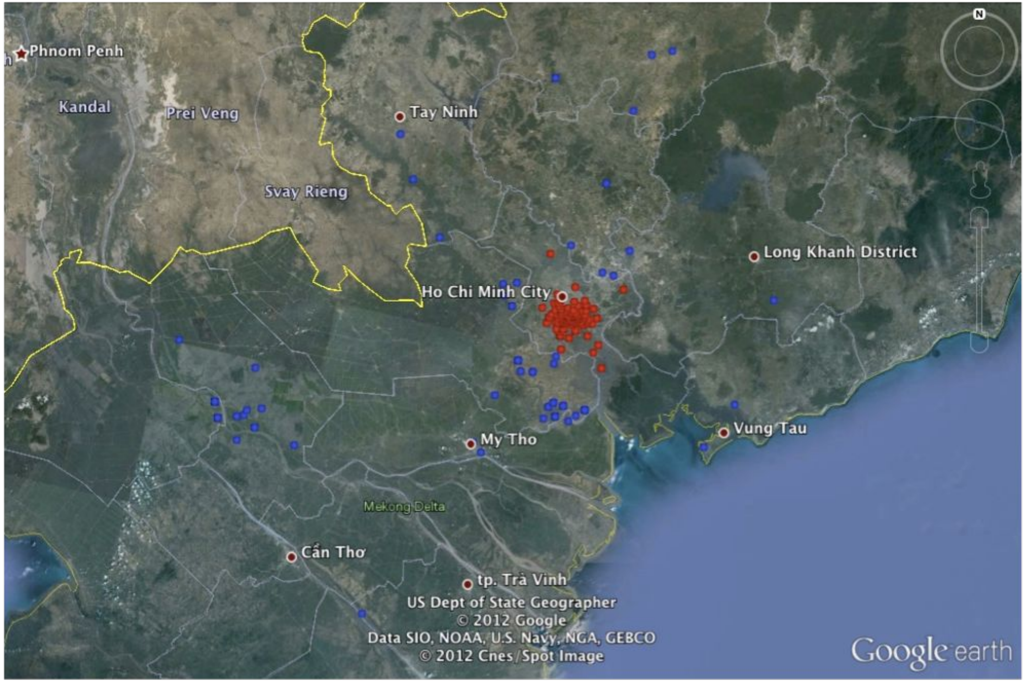
Lineage state probabilities p_{ik} can then be tracked backwards in time using a system of master equations (ODEs) based on the transition rates g_{kl}:



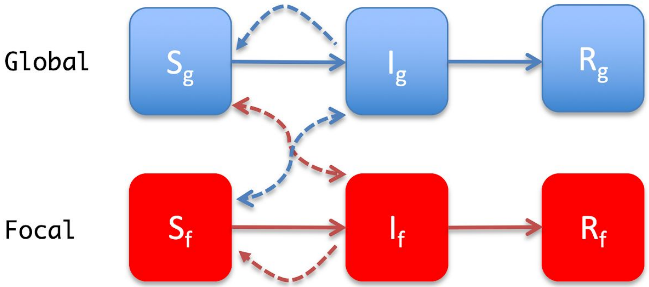
Rasmussen et al., (Virus Evo., 2018)

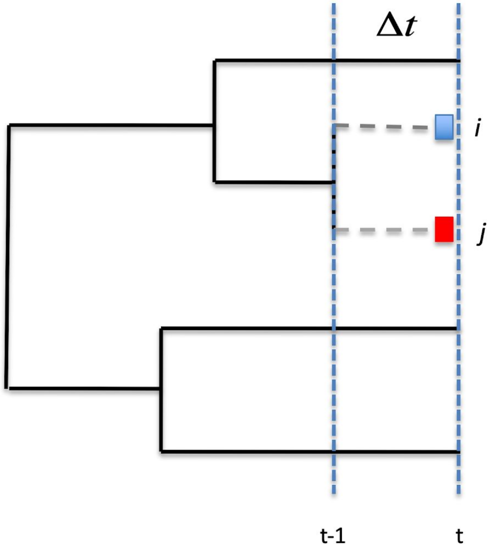
$$\frac{d}{dt}p_{ik} = \sum_{k}^{m} \left(\underbrace{p_{il}g_{kl}}_{\text{Flow in from other pops}} - \underbrace{p_{ik}g_{lk}}_{\text{Flow out to other pops}} \right)$$

Spatial epidemiological dynamics



Spatial SIR model



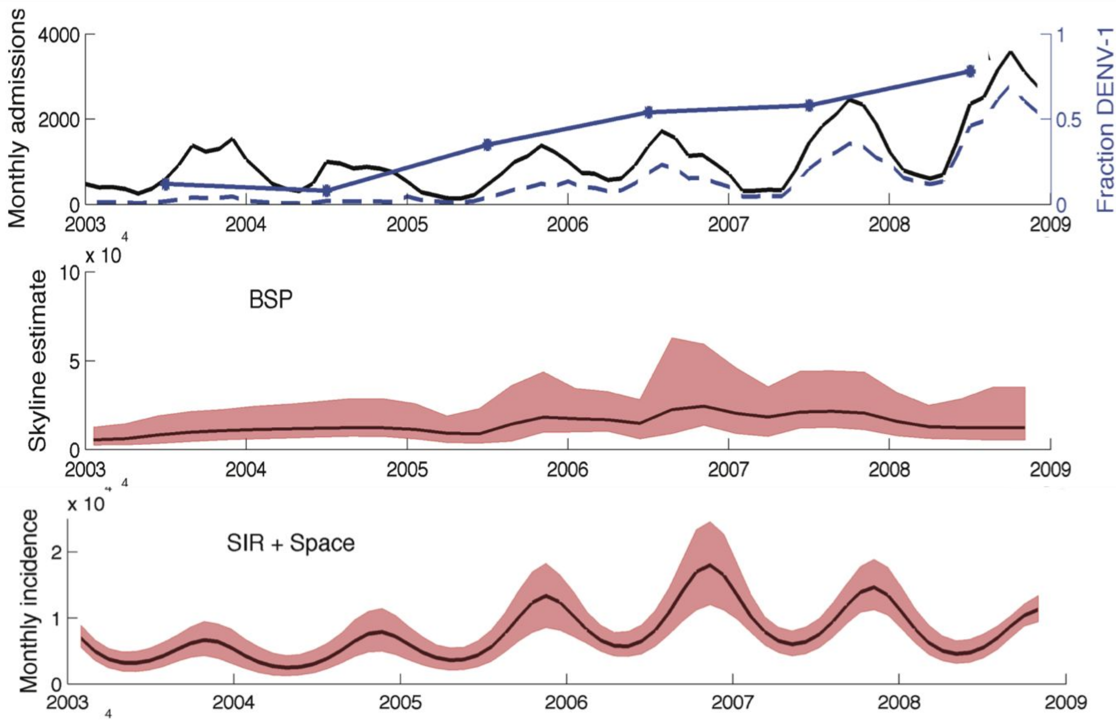


Structured coalescent model:

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

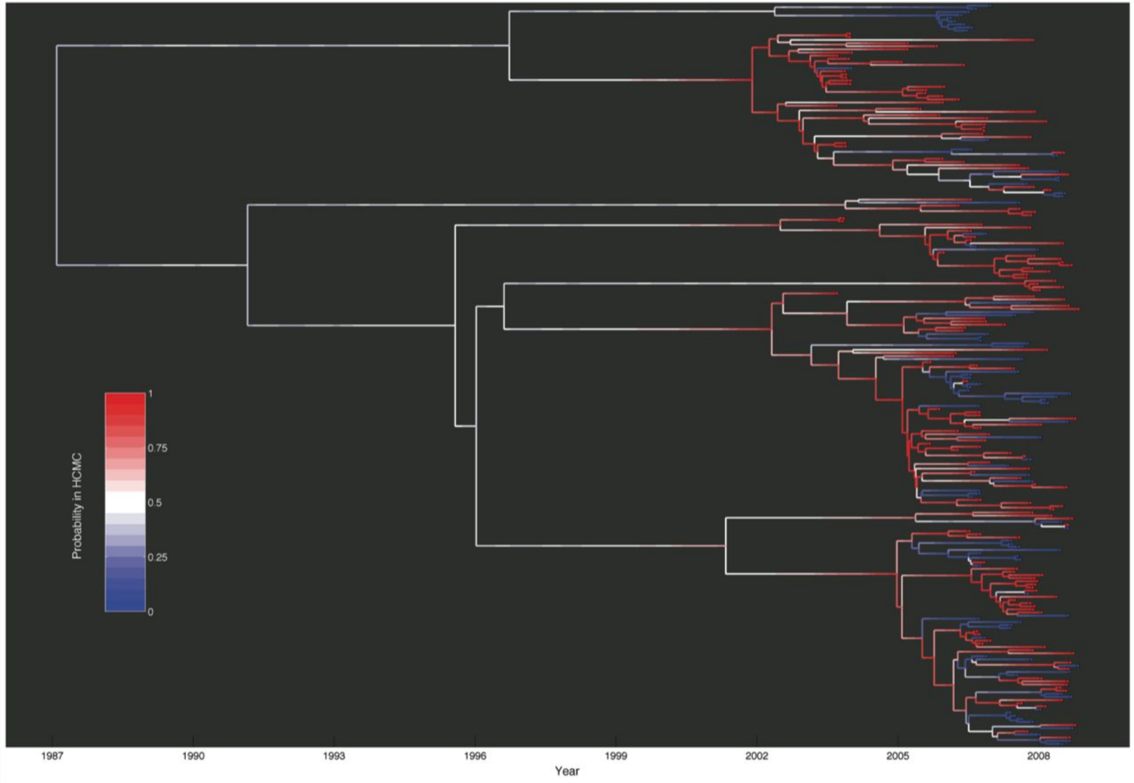
Rasmussen *et al.*, (MBE, 2014)

Estimates accounting for spatial structure



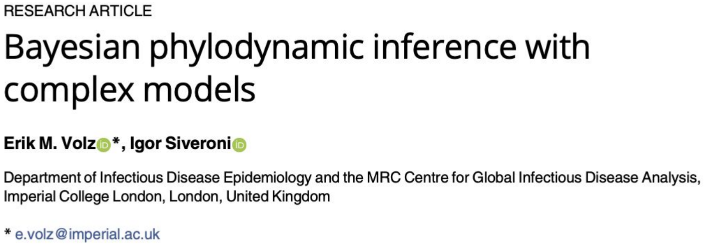
Rasmussen *et al.*, (MBE, 2014)

Movement of DENV lineages



PhyDyn: Epi models in BEAST 2

A BEAST 2 package for fitting generic SIR-type models to pathogen phylogenies using an XML markup schema to specify differential equations.



RESEARCH ARTICLE

Bayesian phylodynamic inference with complex models

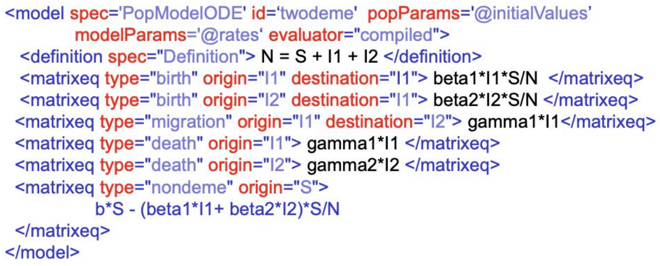
Erik M. Volz*, Igor Siveroni

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

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

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

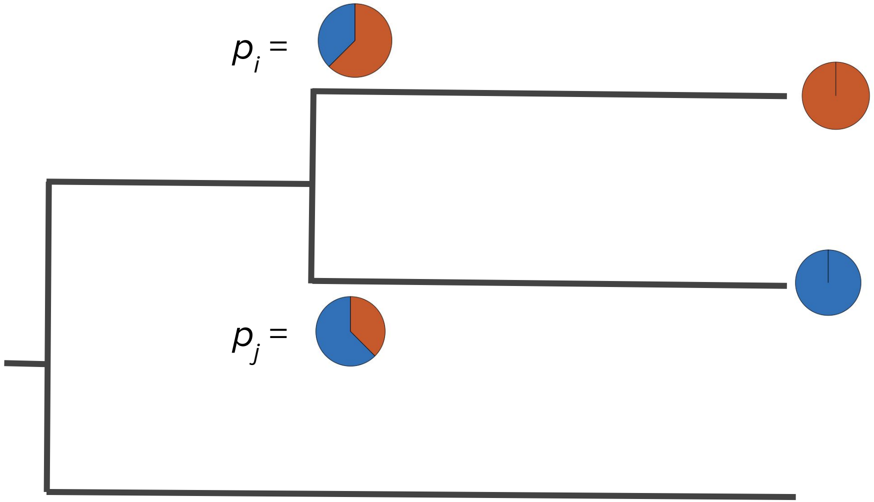


```
<model spec='PopModelODE' id=‘twodeme’  popParams='@initialValues’
         modelParams='@rates‘ evaluator="compiled">
  <definition spec="Definition"> N = S + I1 + I2 </definition>
  <matrixeq type="birth" origin="I1" destination="I1"> beta1*I1*S/N  </matrixeq>
  <matrixeq type="birth" origin="I2" 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>
```

Tutorial available at: https://davidrasm.github.io/MolEpi/tutorials/phydyn-week12/

Tracking lineage state probabilities

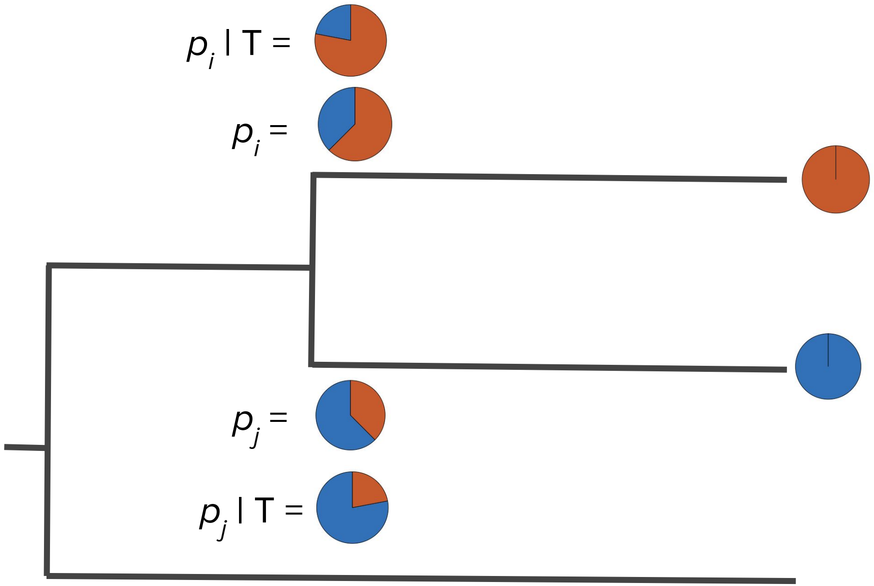
Lineage state probabilities can differ between methods depending on our assumptions about how the migration process relates to the tree.



So is the probability of a lineage being in a given population independent of the locations of other lineages?

Tracking lineage state probabilities

Assume N_e is small so lineages should coalesce rapidly. Conditional on the observation that two lineages have not yet coalesced, the probability that both lineages are in the same population could be drastically lower.

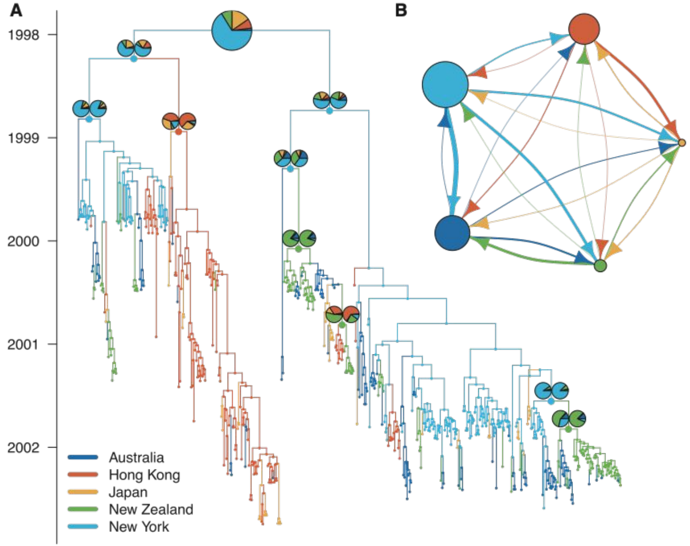


MASCOT

Assuming lineage independence can bias ancestral state inference when migration is slow relative and when coalescent rates are highly asymmetric (Müller *et al.*, MBE, 2018)

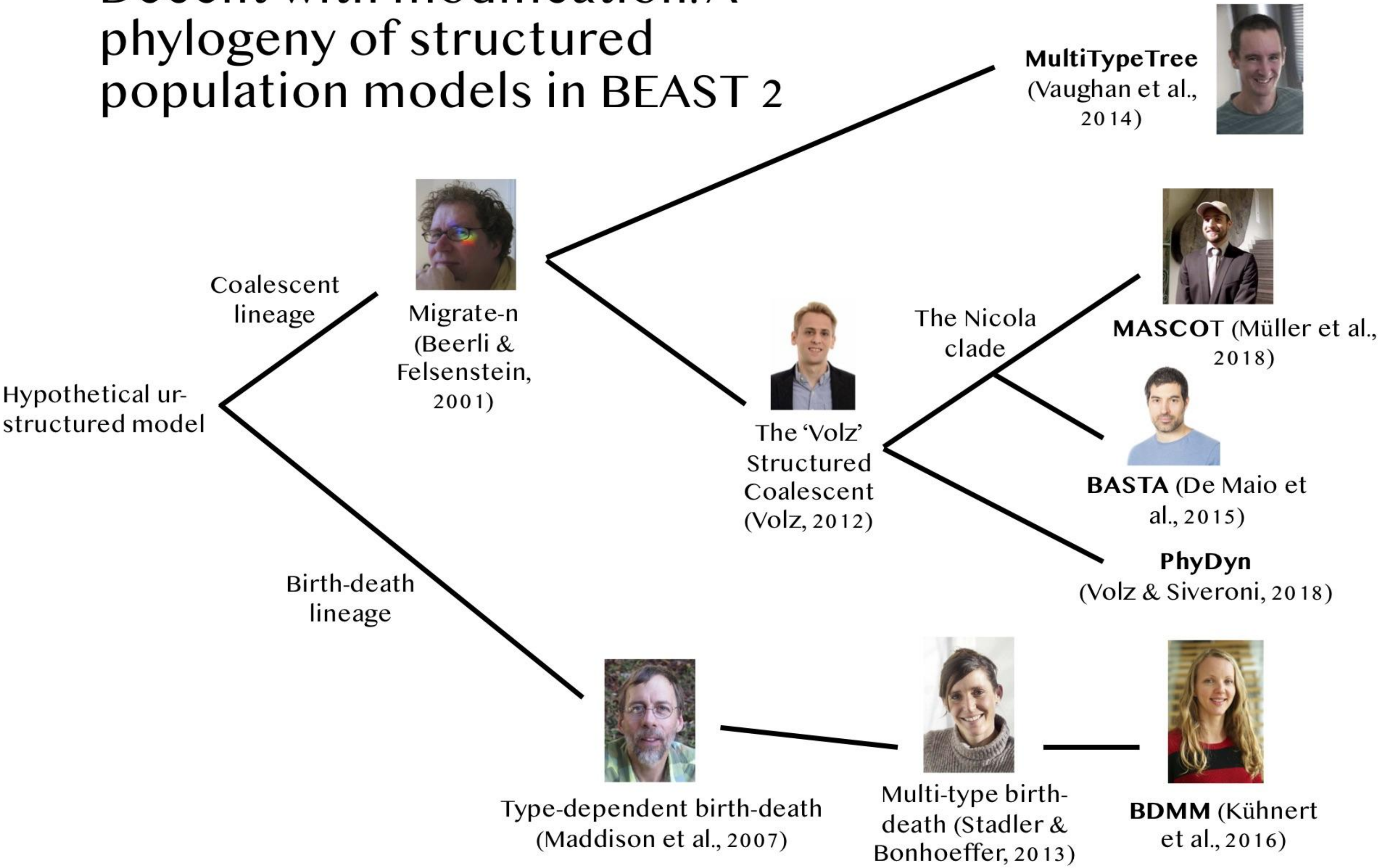
MASCOT uses and improved approximation to the structured coalescent by tracking configurations of lineage states rather than treating lineages independently.

Also allows for inference of time-varying pop sizes and migration rates



Müller *et al.* (2018)

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



Hypothetical ur-structured model

Coalescent lineage

Migrate-n (Beerli & Felsenstein, 2001)

MultiTypeTree (Vaughan et al., 2014)

The 'Volz' Structured Coalescent (Volz, 2012)

The Nicola clade

MASCOT (Müller et al., 2018)

BASTA (De Maio et al., 2015)

PhyDyn (Volz & Siveroni, 2018)

Birth-death lineage

Type-dependent birth-death (Maddison et al., 2007)

Multi-type birth-death (Stadler & Bonhoeffer, 2013)

BDMM (Kühnert et al., 2016)

Joëlle will now tell you about recent diversification along the MTBD lineage