Phylogeographic Inference using the Structured Coalescent

Tim Vaughan

Taming the BEAST Online 2021

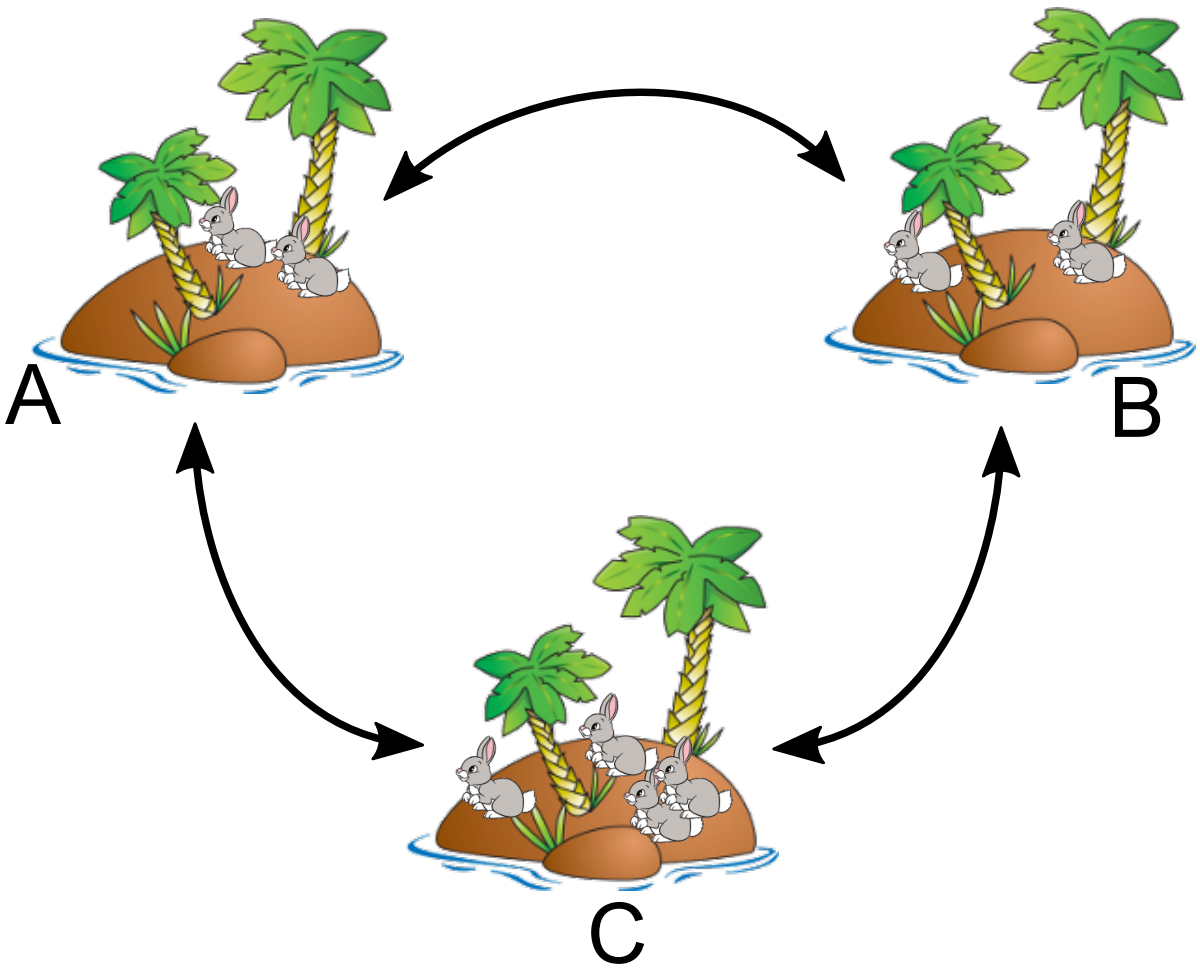
What is a structured population?

A structured population is able to be partitioned into groups (subpopulations) between which gene flow is limited.

- Population structure can dramatically influence the shape of the tree.
- Structure can be produced by
  - Geographic segregation with slow migration (cf. phylogeography),
  - Distinct phases of an infection which during which a pathogen is more or less contagious,
  - *et cetera!*

Generalized island models and demes

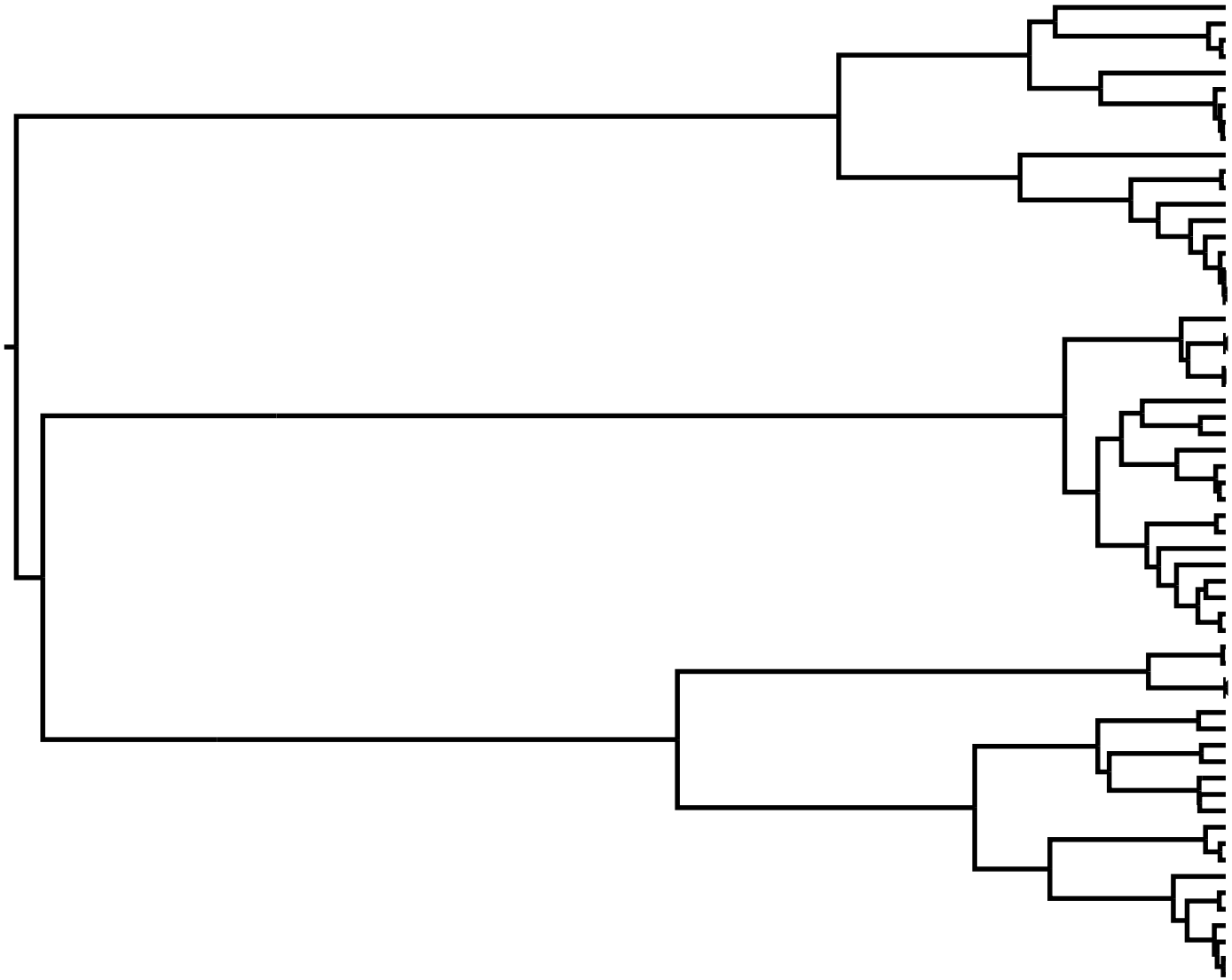
The island model is a common discrete model of spatial structure:



Locations are sometimes referred to as *demes*.

Effect of population structure on trees

Population structure can have a very strong effect on the shape of the trees sampled from that population:

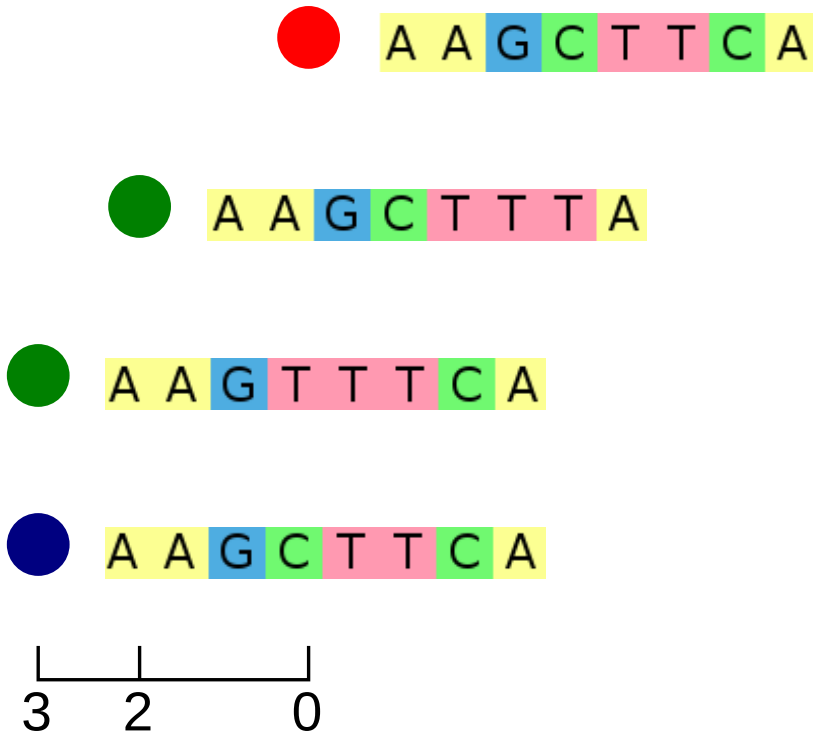


Phylogeographic inference data

Sample	Sequence	Location	Age/Time
1	A A G C T T C A	Place A	0
2	A A G C T T T A	Place B	2
3	A A G T T T C A	Place B	3
4	A A G C T T C A	Place C	3

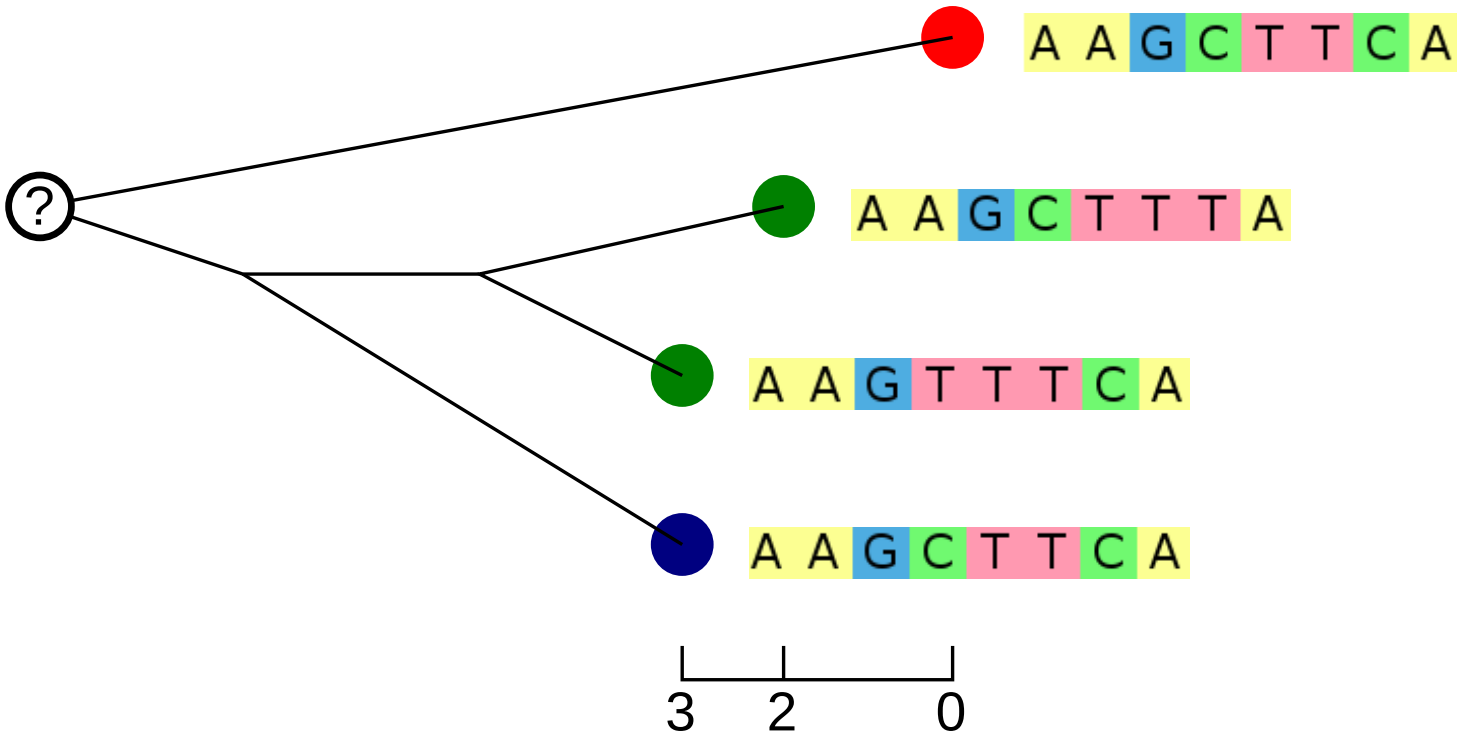
Phylogeographic inference questions

Common questions include:



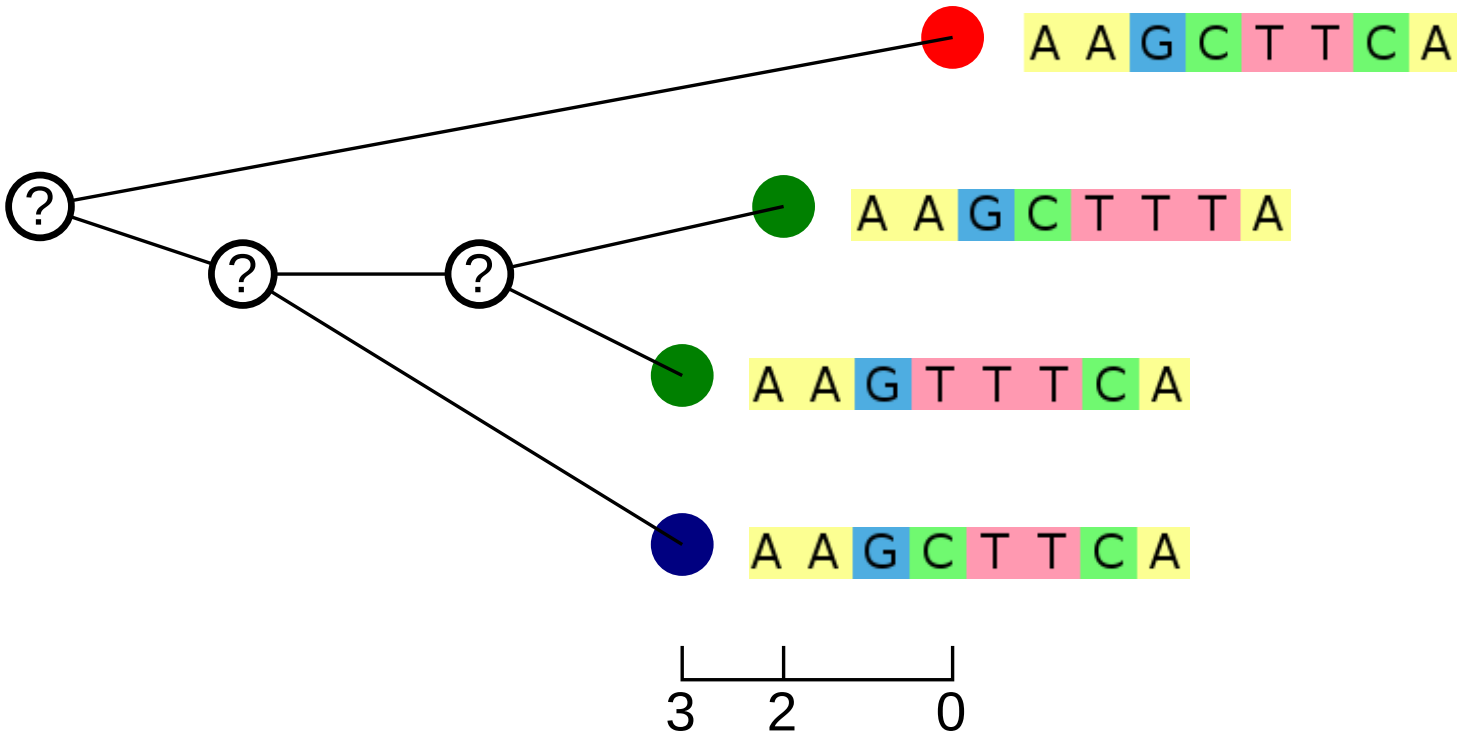
Phylogeographic inference questions

Common questions include:



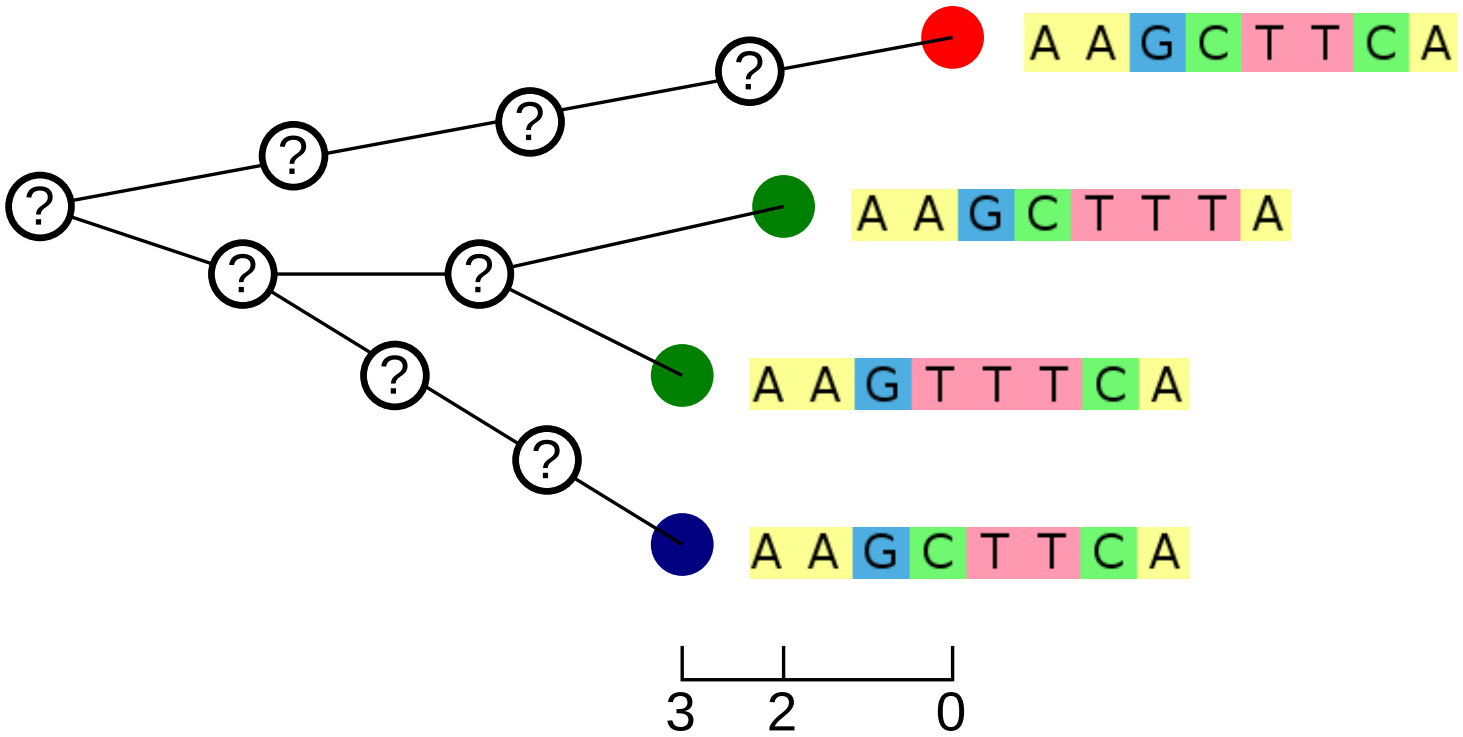
Phylogeographic inference questions

Common questions include:



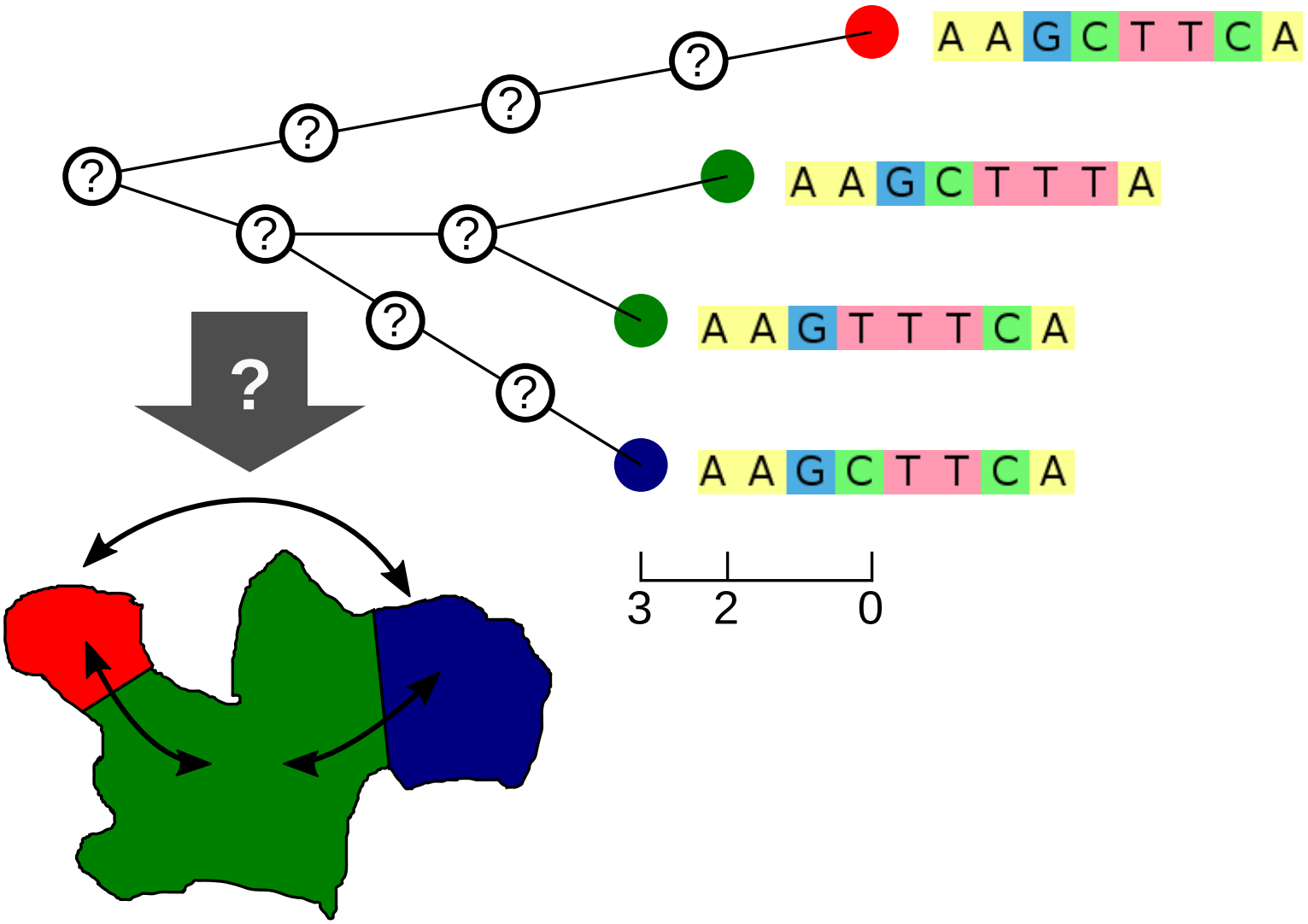
Phylogeographic inference questions

Common questions include:



Phylogeographic inference questions

Common questions include:



Models for Phylogeographic inference

Currently there are two main classes of structured models used in phylogenetic inference:

Models for Phylogeographic inference

Currently there are two main classes of structured models used in phylogenetic inference:

- **Mugration models (also Discrete Trait Analysis):**
  - Given tree and root location, what is the probability of sample locations?
  - Exist in continuous and discrete forms.
  - Developed by Phillipe Lemey et al. (Lemey et al., 2009, 2010).

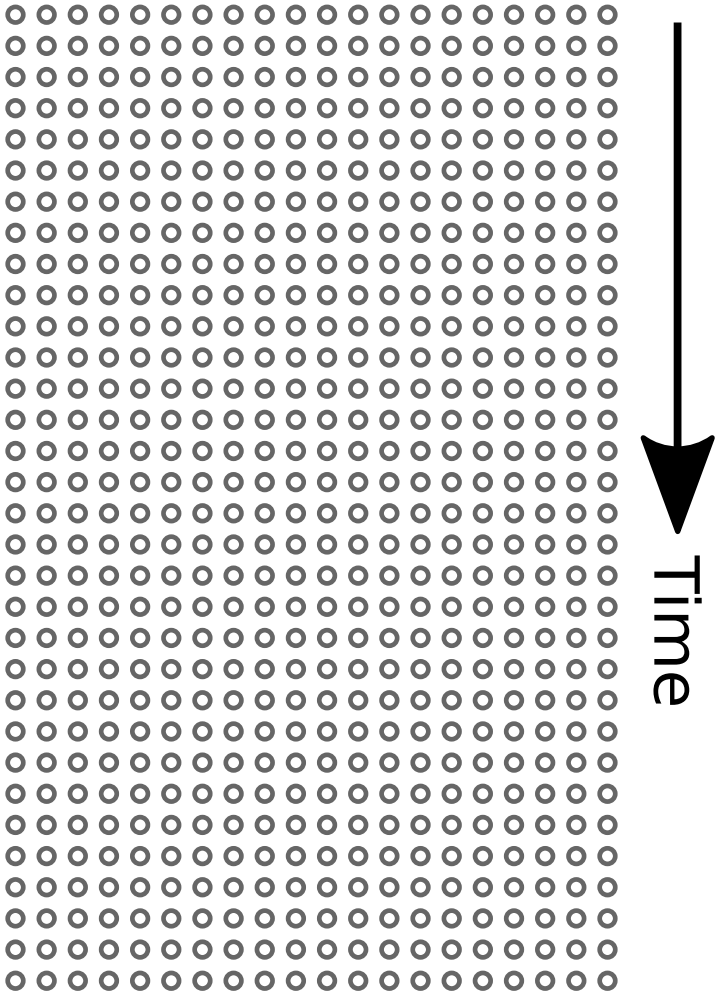
Models for Phylogeographic inference

Currently there are two main classes of structured models used in phylogenetic inference:

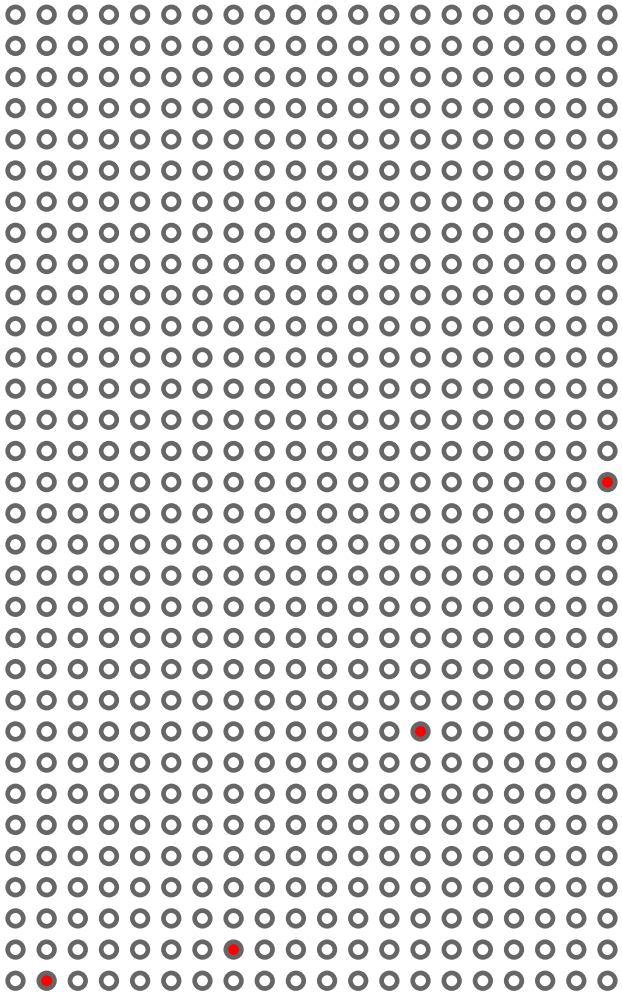
- **Mugration models (also Discrete Trait Analysis):**
  - Given tree and root location, what is the probability of sample locations?
  - Exist in continuous and discrete forms.
  - Developed by Phillipe Lemey et al. (Lemey et al., 2009, 2010).
- **Structured population models:**
  - Given sequences and locations, what is the probability of the tree?
  - Currently mostly discrete.
  - Many extend the structured coalescent framework of Hudson (1990) and Notohara (1990).
  - Others extend the birth-death-sampling framekwork of Stadler (2010).

Wright-Fisher model

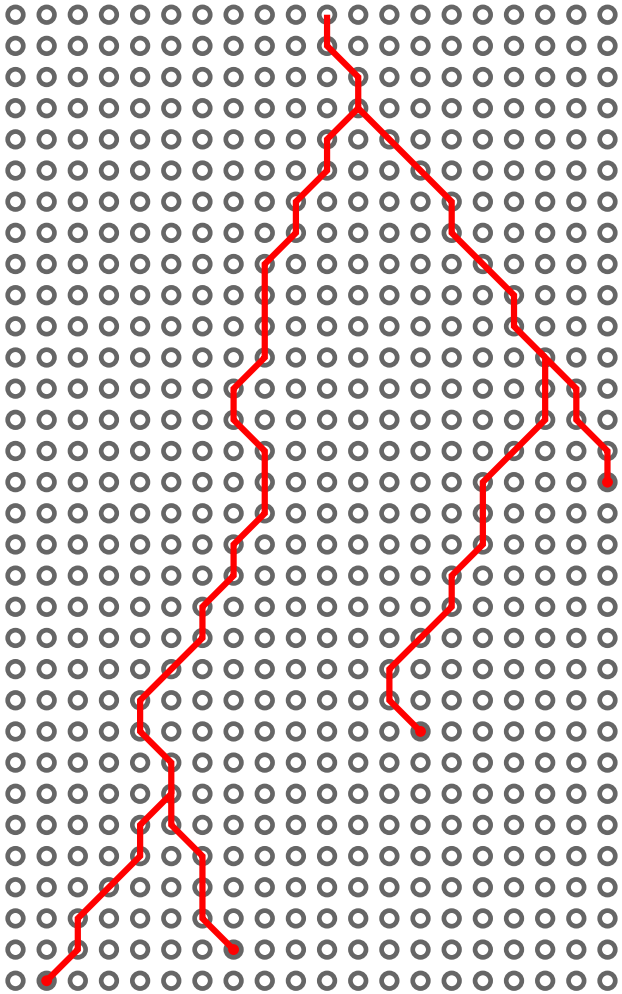
Wright-Fisher model



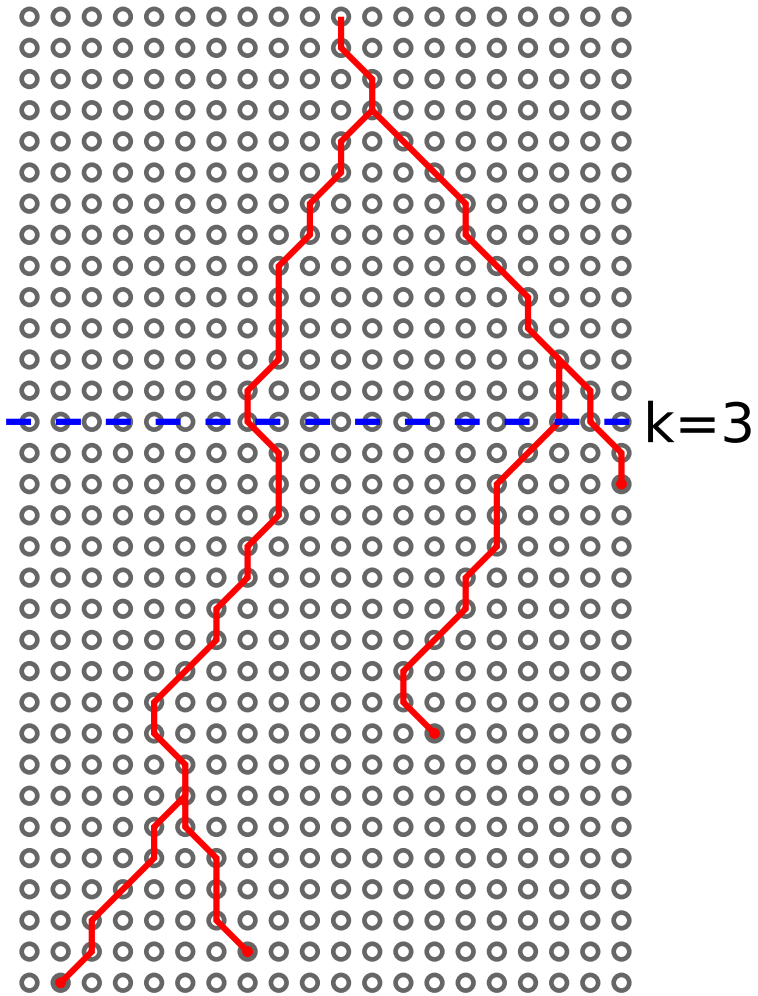
Wright-Fisher model



Wright-Fisher model



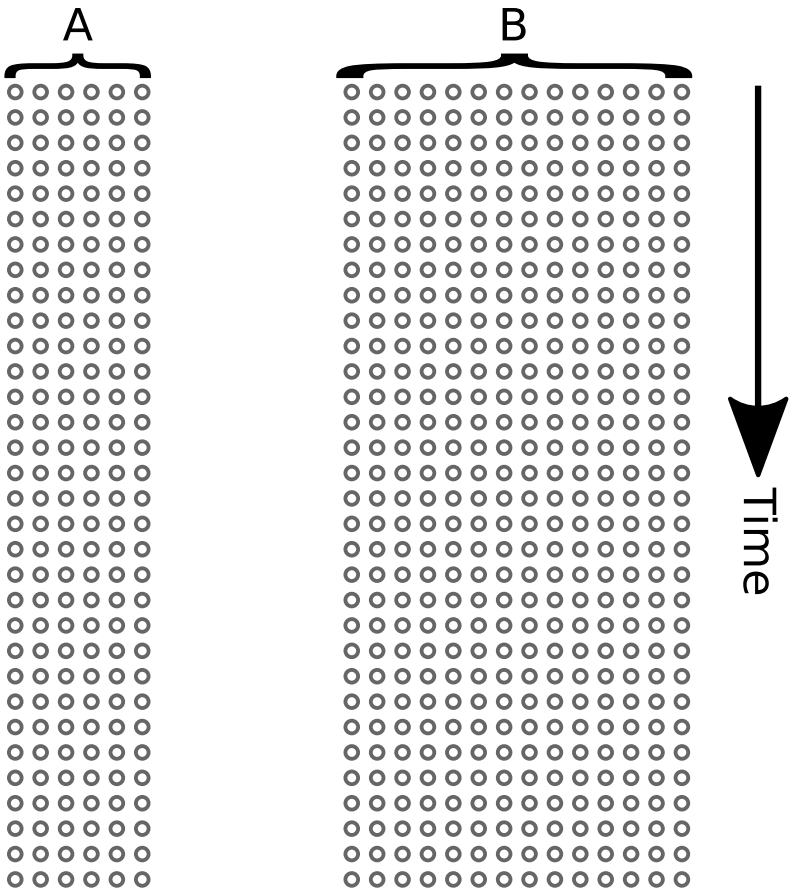
Wright-Fisher model



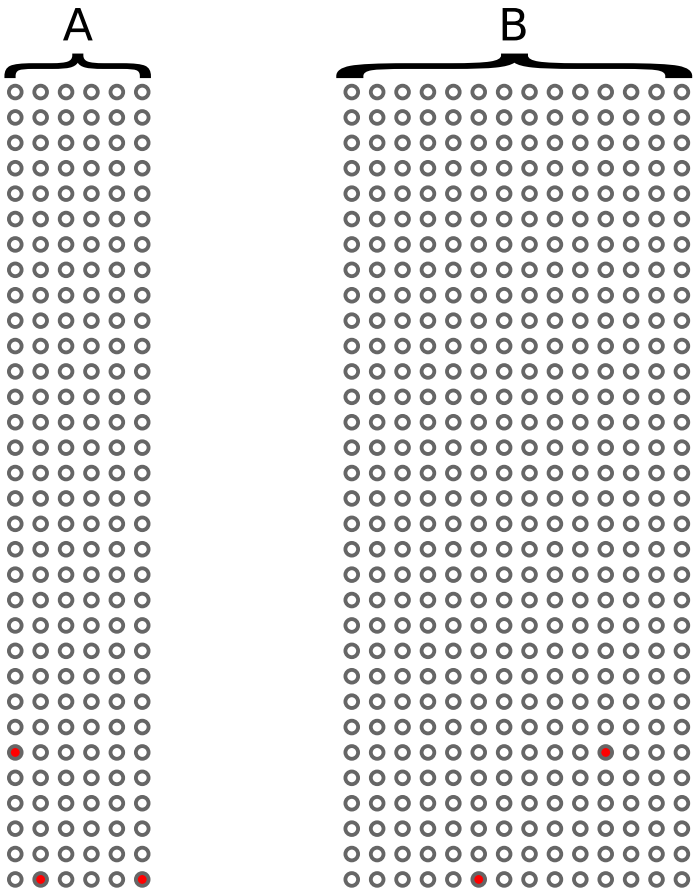
Probability of coalescence per generation:

$$\sim \binom{k}{2} \frac{1}{N}$$

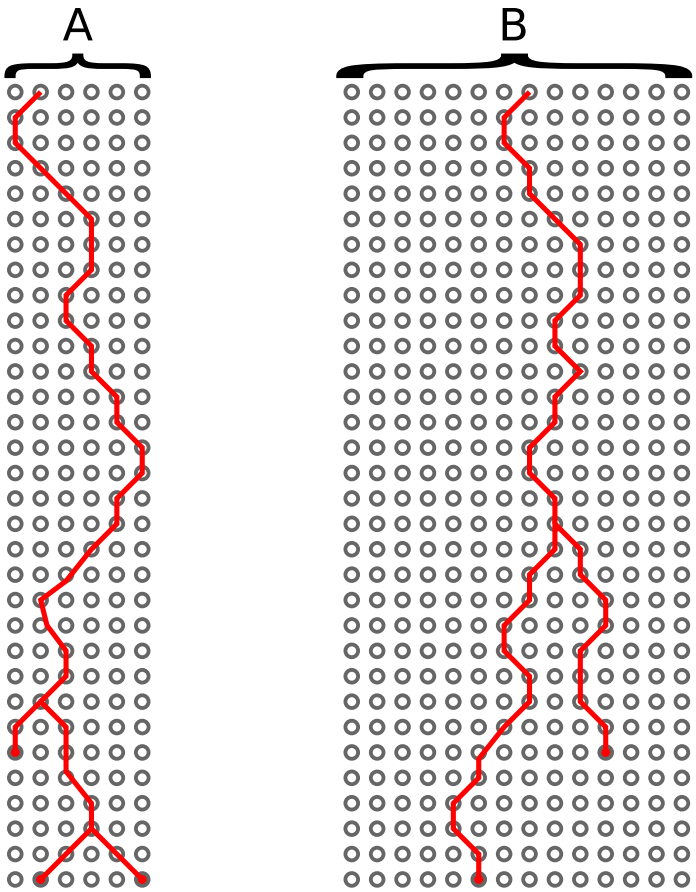
Partitioned Wright-Fisher model



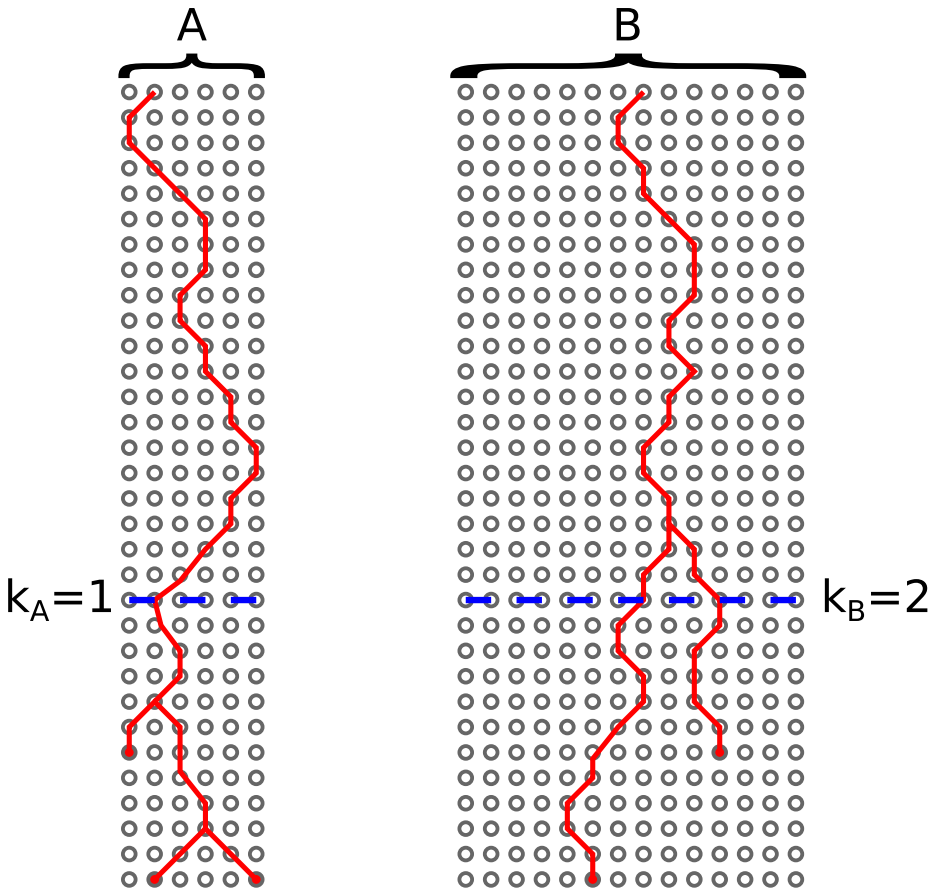
Partitioned Wright-Fisher model



Partitioned Wright-Fisher model



Partitioned Wright-Fisher model



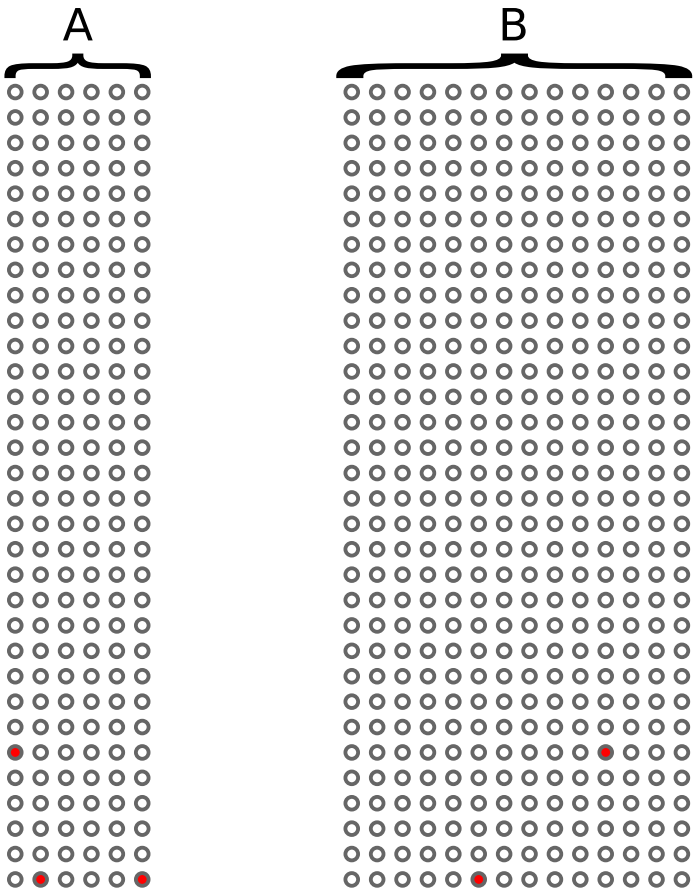
Probability of coalescence per generation in A:

$$\binom{k_A}{2}\frac{1}{N_A}$$

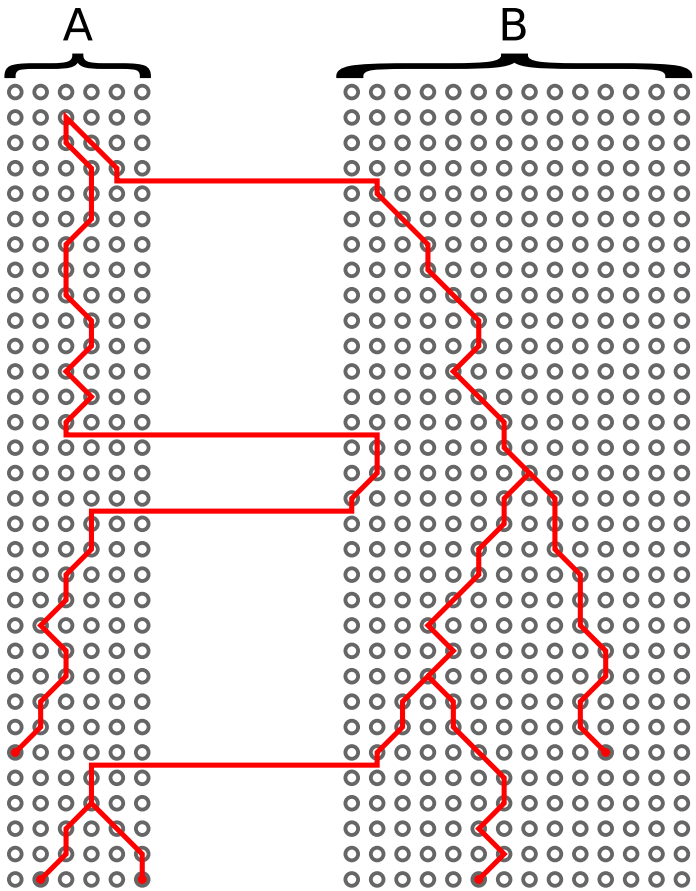
Probability of coalescence per generation in B:

$$\binom{k_B}{2}\frac{1}{N_B}$$

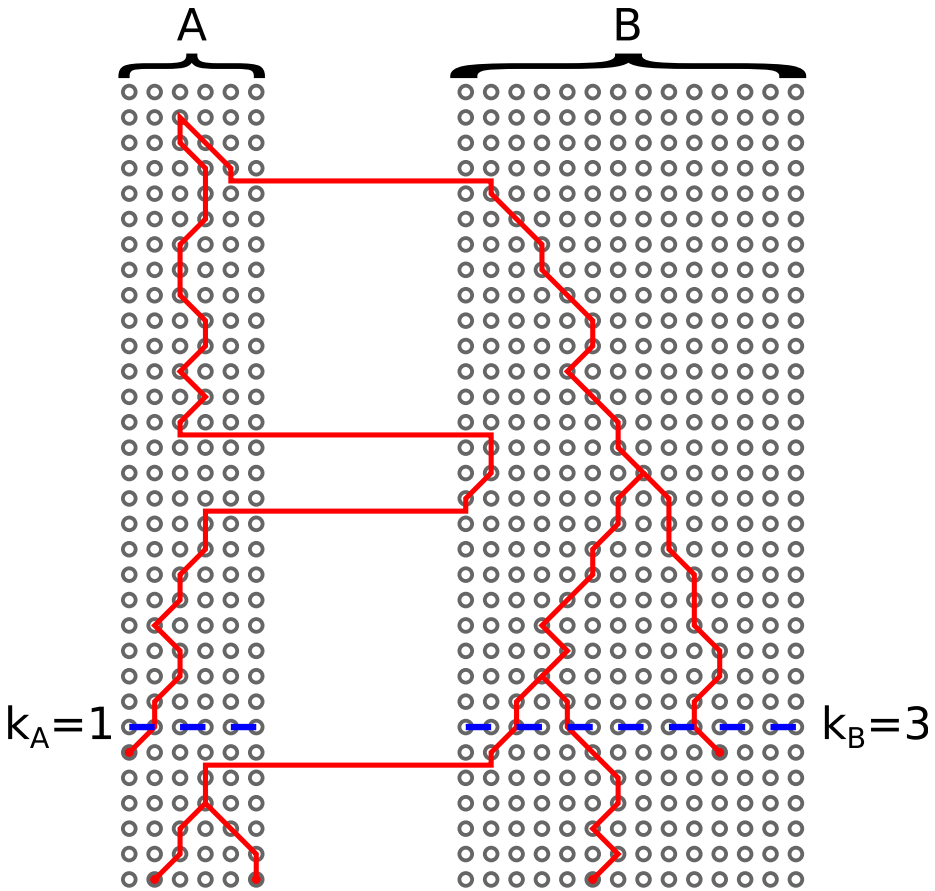
Structured Wright-Fisher model



Structured Wright-Fisher model



Structured Wright-Fisher model



Probability of migration from $A \rightarrow B$ per individual in A:

$$q_{AB}$$

Probability of single lineage migration from $B \rightarrow A$ (**backward time**):

$$m_{BA} = q_{AB} \frac{N_A}{N_B}$$

Structured Coalescent

Backwards-in-time Markov process that generates both the sampled tree and ancestral locations.
(Hudson, 1990; Notohara, 1990)

Structured Coalescent

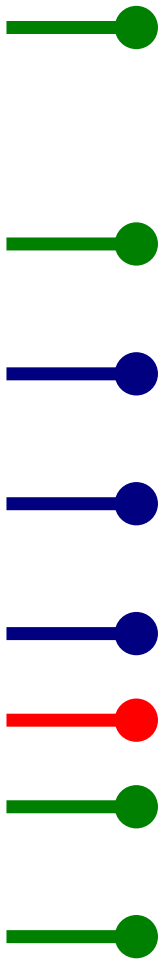
Backwards-in-time Markov process that generates both the sampled tree and ancestral locations.

(Hudson, 1990; Notohara, 1990)

Structured Coalescent

Backwards-in-time Markov process that generates both the sampled tree and ancestral locations.

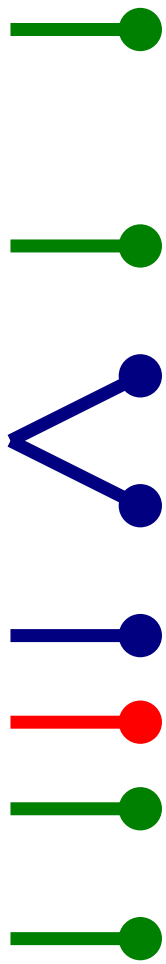
(Hudson, 1990; Notohara, 1990)



Structured Coalescent

Backwards-in-time Markov process that generates both the sampled tree and ancestral locations.

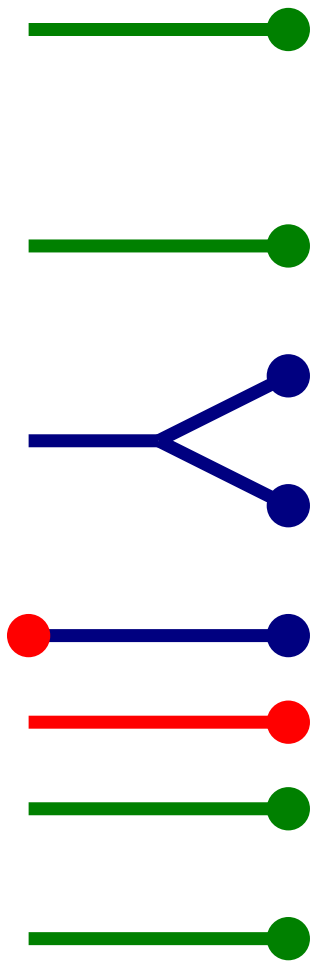
(Hudson, 1990; Notohara, 1990)



Structured Coalescent

Backwards-in-time Markov process that generates both the sampled tree and ancestral locations.

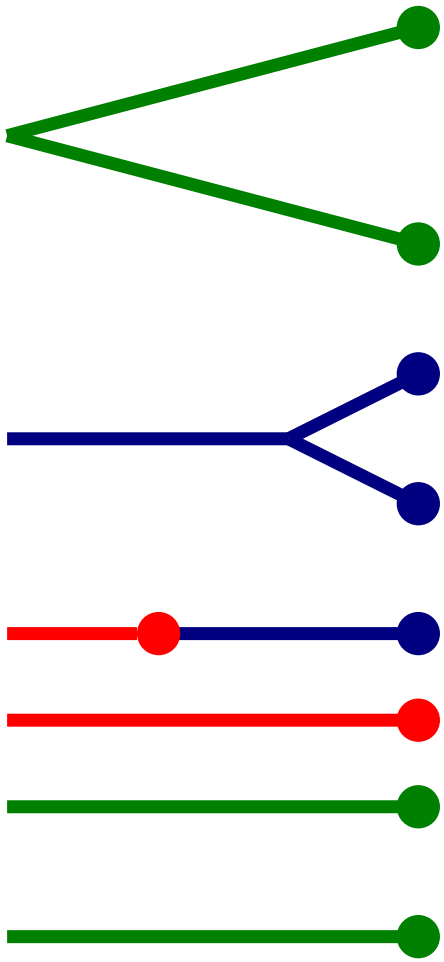
(Hudson, 1990; Notohara, 1990)



Structured Coalescent

Backwards-in-time Markov process that generates both the sampled tree and ancestral locations.

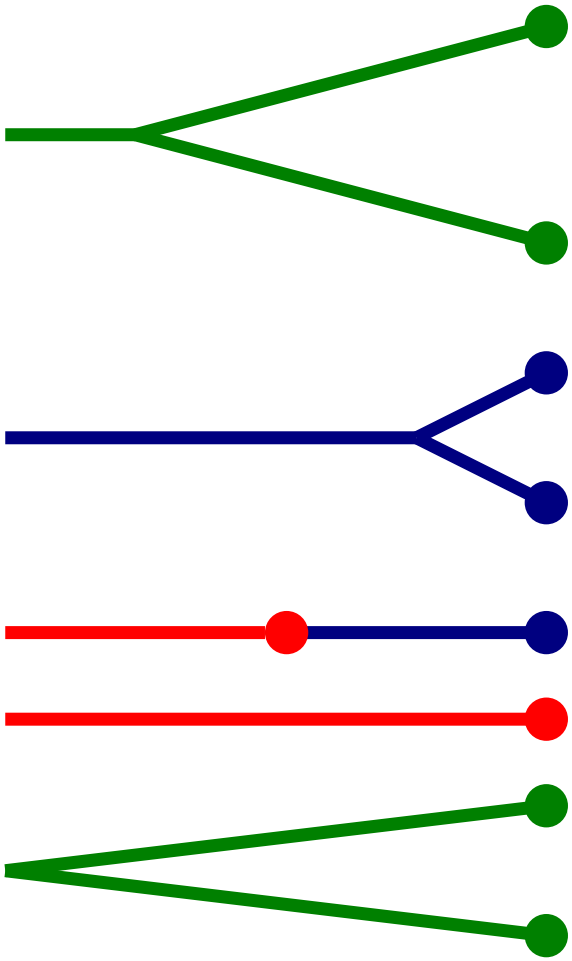
(Hudson, 1990; Notohara, 1990)



Structured Coalescent

Backwards-in-time Markov process that generates both the sampled tree and ancestral locations.

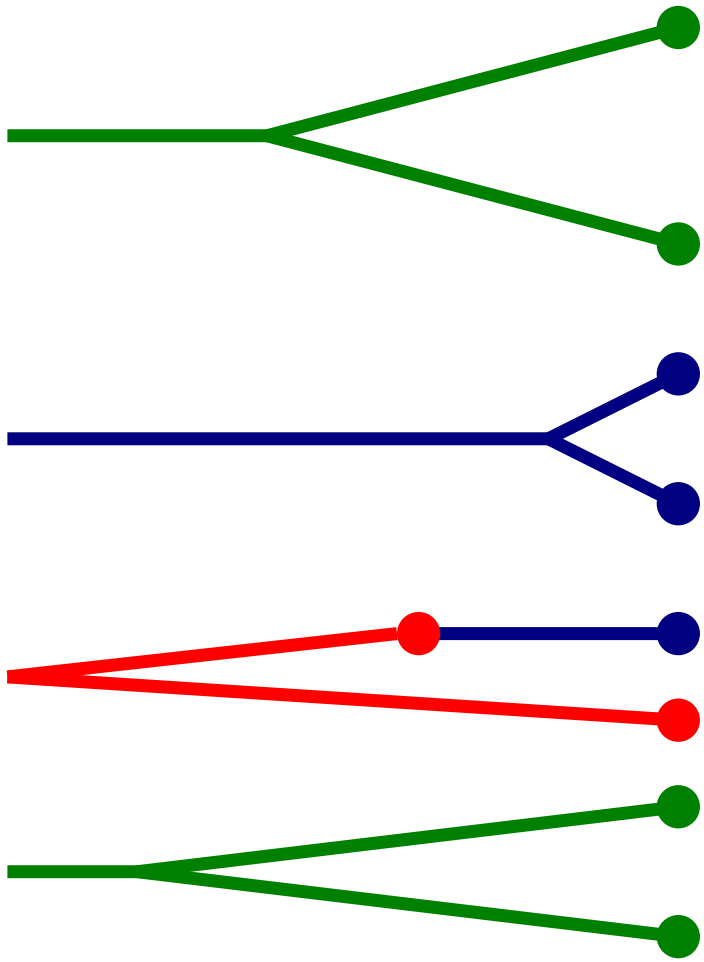
(Hudson, 1990; Notohara, 1990)



Structured Coalescent

Backwards-in-time Markov process that generates both the sampled tree and ancestral locations.

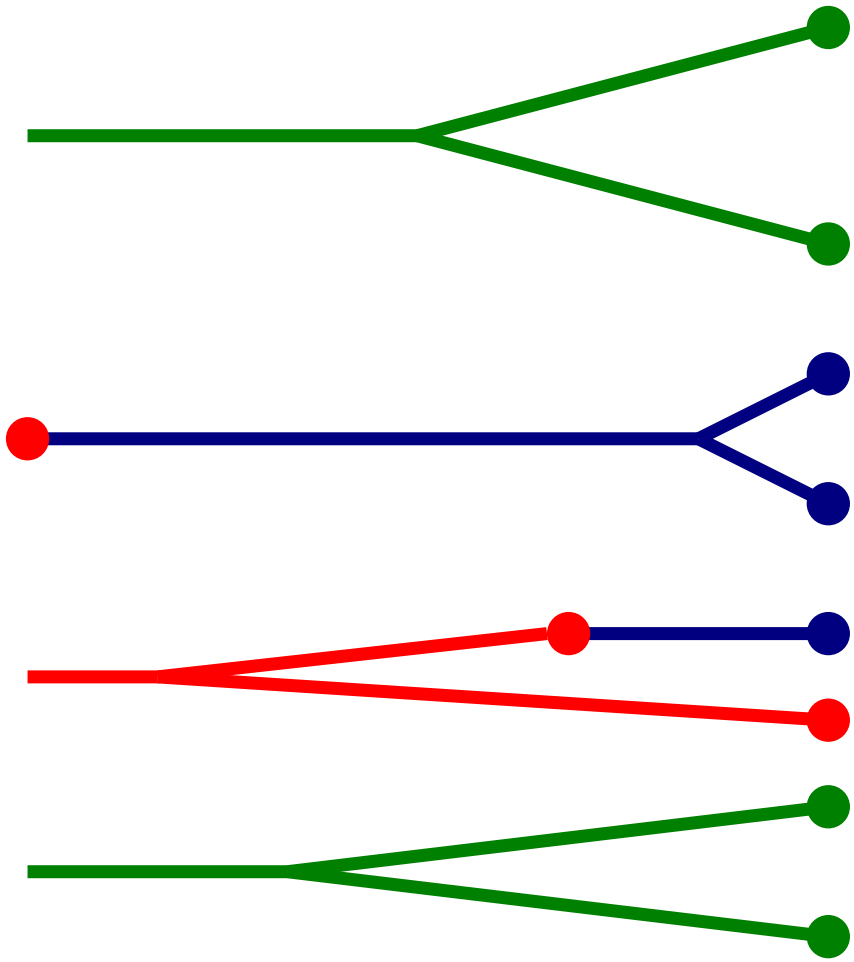
(Hudson, 1990; Notohara, 1990)



Structured Coalescent

Backwards-in-time Markov process that generates both the sampled tree and ancestral locations.

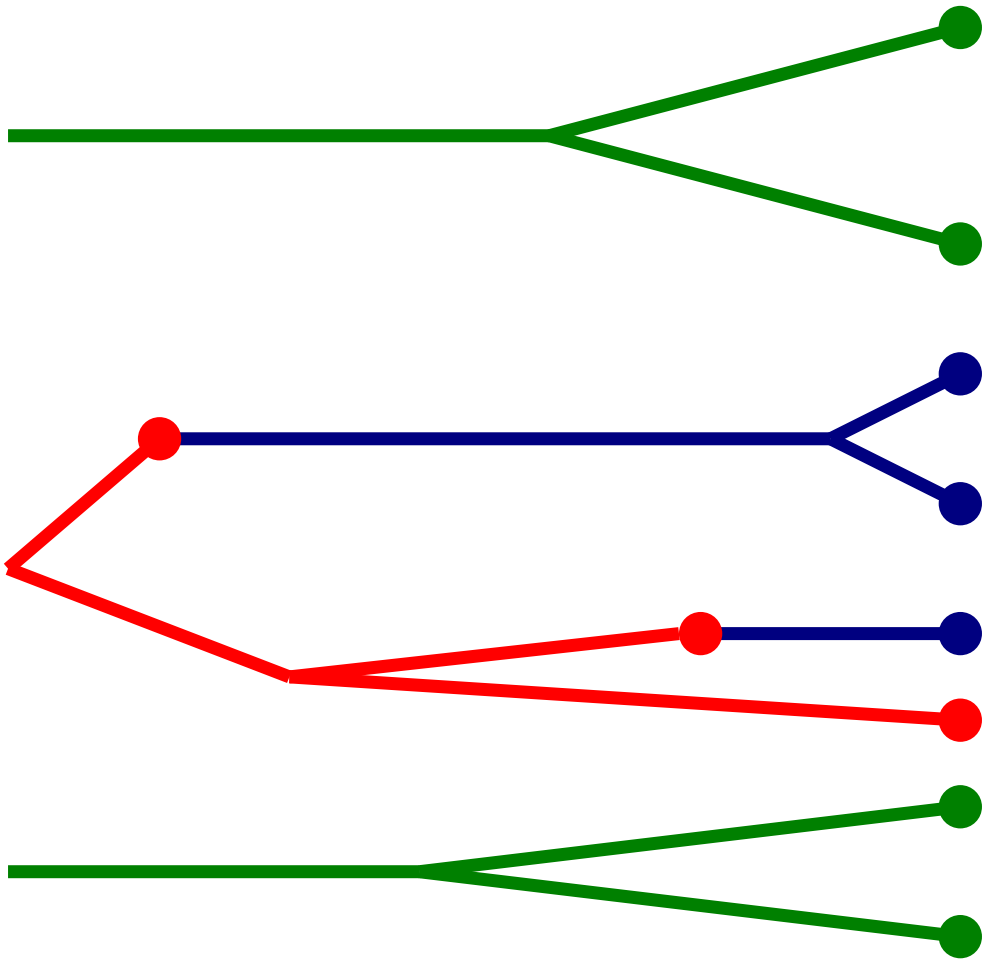
(Hudson, 1990; Notohara, 1990)



Structured Coalescent

Backwards-in-time Markov process that generates both the sampled tree and ancestral locations.

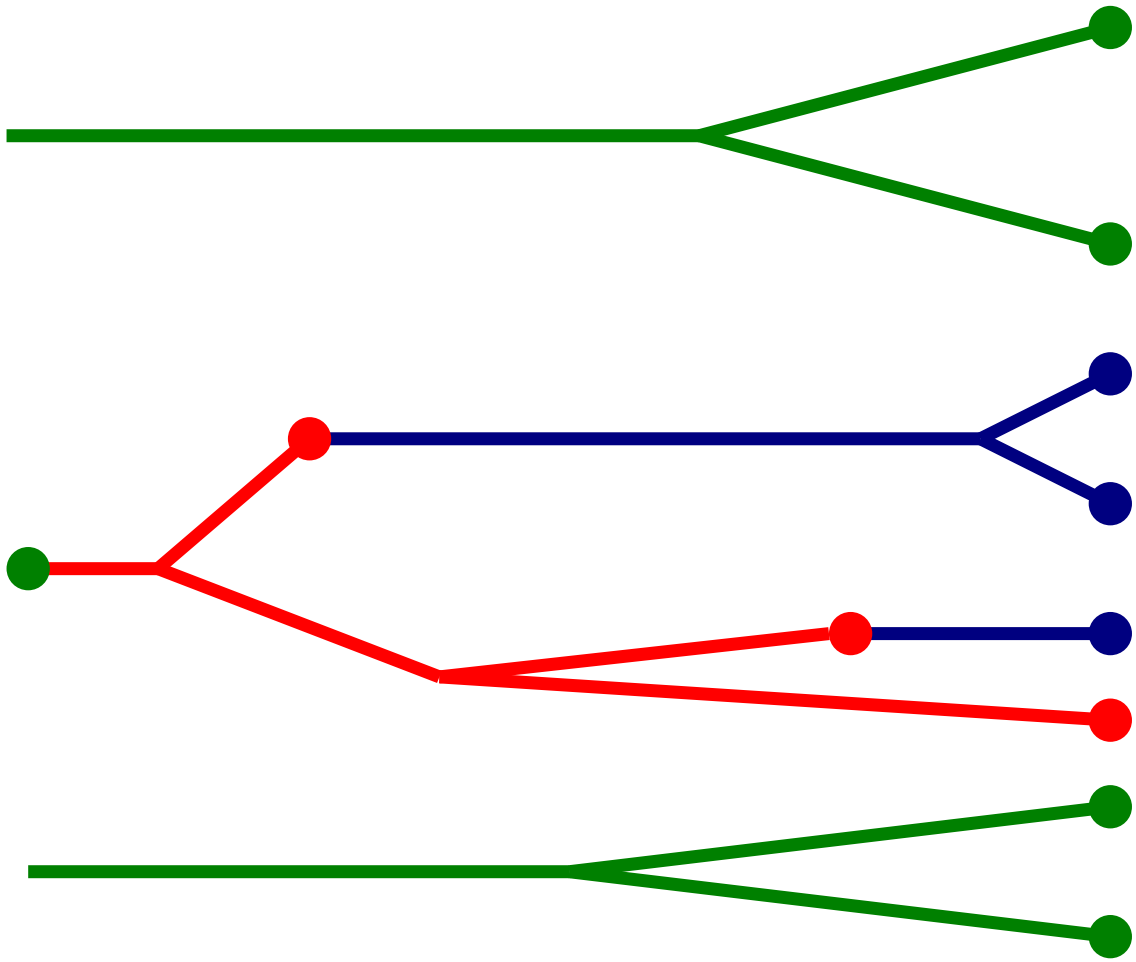
(Hudson, 1990; Notohara, 1990)



Structured Coalescent

Backwards-in-time Markov process that generates both the sampled tree and ancestral locations.

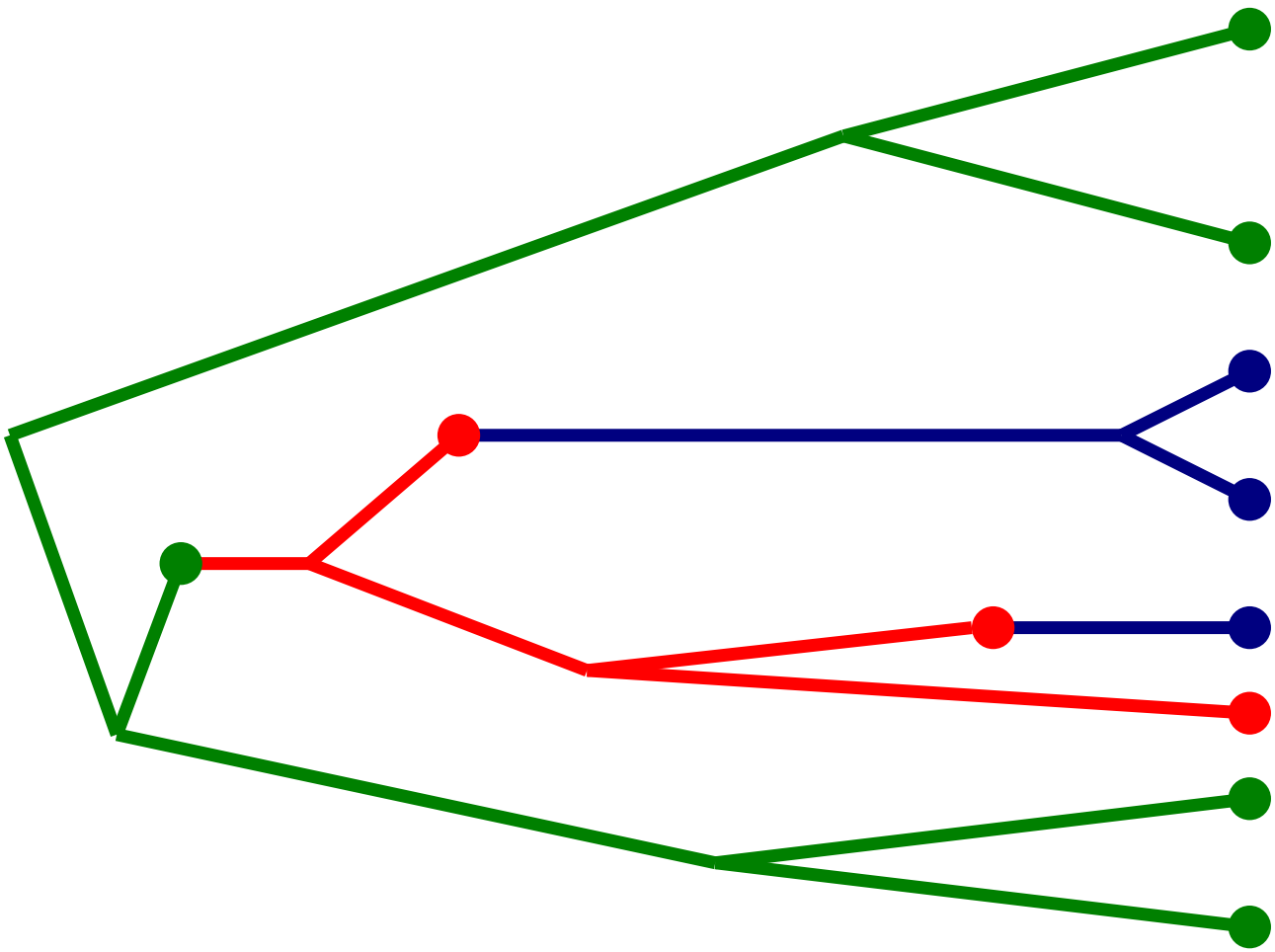
(Hudson, 1990; Notohara, 1990)



Structured Coalescent

Backwards-in-time Markov process that generates both the sampled tree and ancestral locations.

(Hudson, 1990; Notohara, 1990)



SC Inference: Modified tree prior

The standard phylogenetic posterior is modified:

$$\begin{aligned} P(T, C, \mu, \theta, \bar{M}, \vec{N}|A, L) = & \frac{1}{P(A|L)} P(A|T, \mu) \\ & \times P(T, C|\vec{N}, \bar{M}, L) \\ & \times P(\mu)P(\theta)P(\bar{M})P(\vec{N}) \end{aligned}$$

where

- L are the sampled locations,
- $\vec{N}$ are the deme-specific population sizes,
- $\bar{M}$ is the **backward-time** migration rate matrix, and
- C are the ancestral locations on the tree.

The sample locations and SC model affect the **tree prior**.

The *shape* of the tree is affected by structure.

Sampling assumption

- The coalescent tree prior is explicitly conditioned on the sample times.
- Similarly, the structured coalescent tree prior is conditioned on sample locations.

Sampling assumption

- The coalescent tree prior is explicitly conditioned on the sample times.
- Similarly, the structured coalescent tree prior is conditioned on sample locations.

The strucured coalescent makes no assumption about the manner in which samples are collected with respect to location.

Sampling assumption

- The coalescent tree prior is explicitly conditioned on the sample times.
- Similarly, the structured coalescent tree prior is conditioned on sample locations.

The strucured coalescent makes no assumption about the manner in which samples are collected with respect to location.

- Sample distribution not used as data.
- Uneven sampling can reduce inference power, but won't bias results!

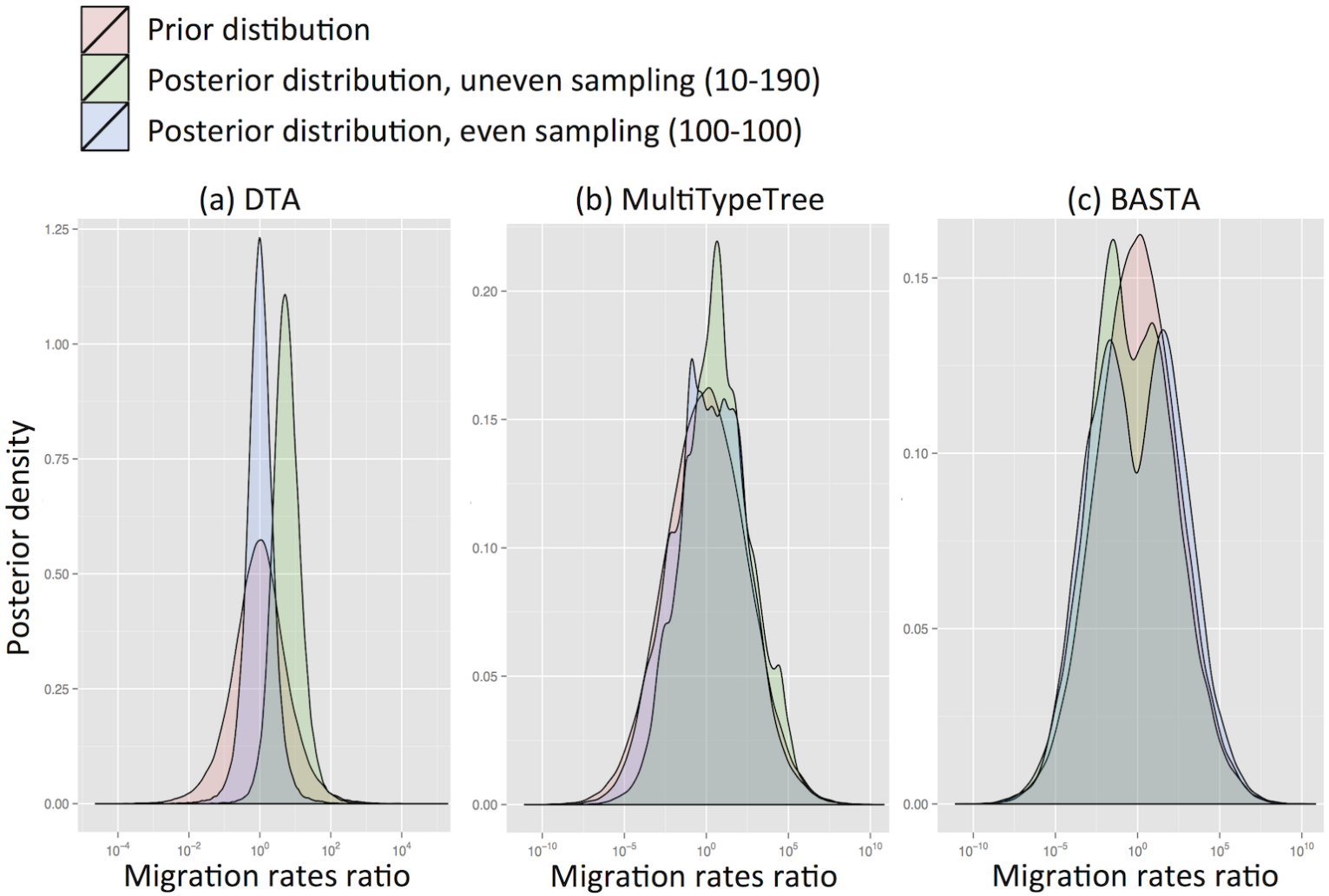
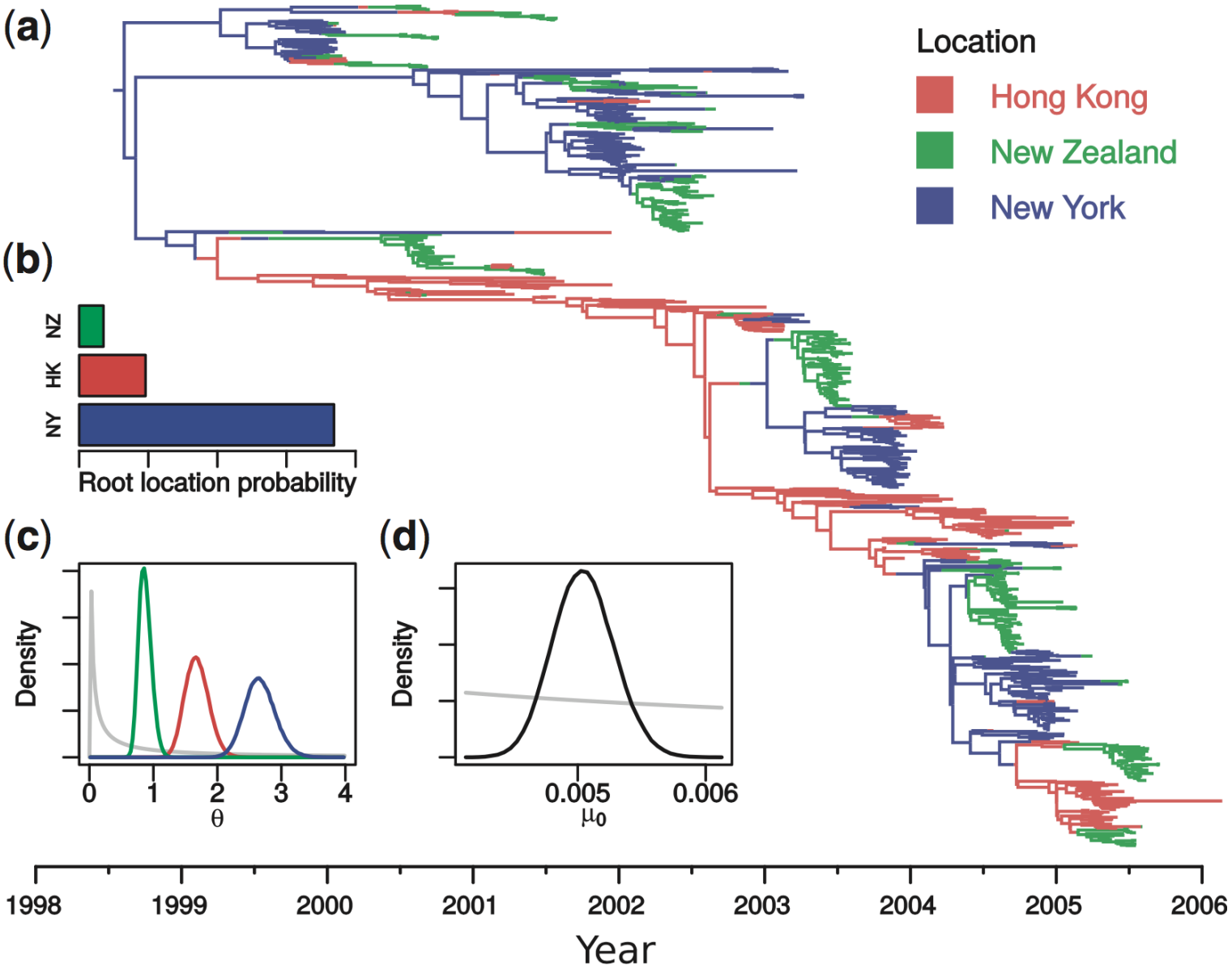
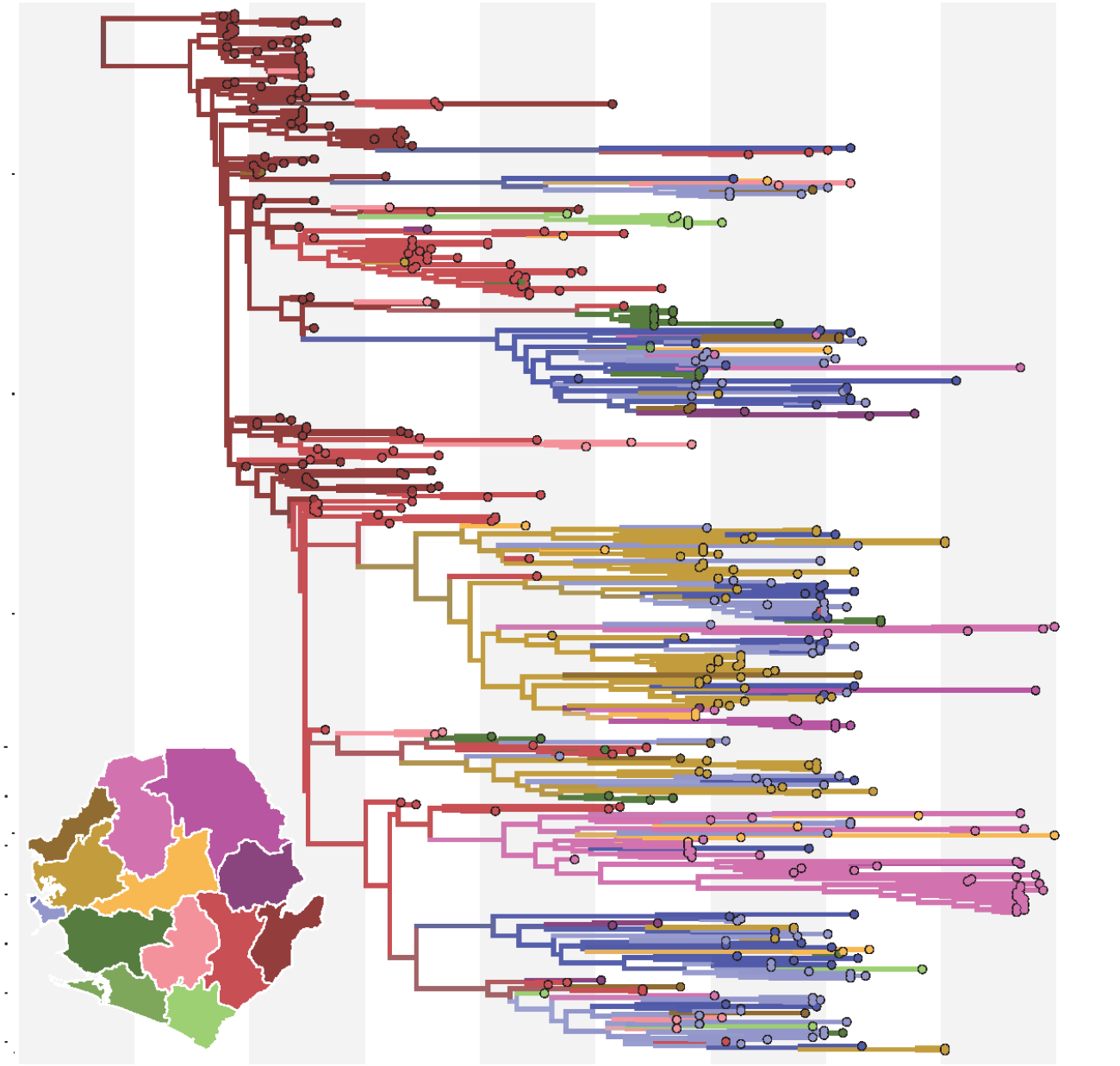


Figure 2, De Maio et al. (2015)



Inference of H3N2 movement using SC, Vaughan et al. (2014)



Inference of geographical spread of Ebola virus during 2014-2015 West-African epidemic, Müller et al. (2019)

The Structured Coalescent in BEAST 2

MultiTypeTree Exact inference under the model. (Vaughan et al., 2014)

- Pro: exact, entire history is sampled.
- Con: restricted to $\leqslant 4$ demes.

BASTA Faster approach which approximately includes all migration histories in each MCMC step. (De Maio et al., 2015)

- Pro: more efficient (handles more demes) than MTT.
- Con: no BEAUti integration, awkward to set up analyses.

MASCOT A more recent approximation more accurate than BASTA. (Müller et al., 2017, 2018)

- Pro: more efficient than MTT, modern GLM approaches supported, BEAUti interface.
- Con: full histories not yet accessible.

Tutorial

Structured coalescent

Population structure using MultiTypeTree

by Nicola F. Müller and Tim Vaughan

Tutorial location: `https://taming-the-beast.org/tutorials/Structured-coalescent/`

Tutorial Slack channel: #t-struct-coal

Wrap-up time: 15:35

References

De Maio, N., Wu, C.-H., O'Reilly, K. M., and Wilson, D. (2015). New routes to phylogeography: A bayesian structured coalescent approximation. *PLoS Genet*, 11(8):e1005421.

Hudson, R. R. (1990). Gene genealogies and the coalescent process. *Oxford Surveys in Evolutionary Biology*, 7:1.

Lemey, P., Rambaut, A., Drummond, A. J., and Suchard, M. A. (2009). Bayesian phylogeography finds its roots. *PLoS Comput Biol*, 5(9):e1000520.

Lemey, P., Rambaut, A., Welch, J. J., and Suchard, M. A. (2010). Phylogeography takes a relaxed random walk in continuous space and time. *Mol Biol Evol*, 27:1877–1885.

Müller, N. F., Dudas, G., and Stadler, T. (2019). Inferring time-dependent migration and coalescence patterns from genetic sequence and predictor data in structured populations. *Virus Evolution*, 5(2).

Müller, N. F., Rasmussen, D., and Stadler, T. (2018). MASCOT: parameter and state inference under the marginal structured coalescent approximation. *Bioinformatics*, 34(22):3843–3848.

Müller, N. F., Rasmussen, D. A., and Stadler, T. (2017). The structured coalescent and its approximations. *Molecular biology and evolution*.

Notohara, M. (1990). The coalescent and the genealogical process in geographically structured population. *J Math Biol*, 29(1):59–75.

Stadler, T. (2010). Sampling-through-time in birth-death trees. *J Theor Biol*, 267(3):396–404.

Vaughan, T. G., Kühnert, D., Popinga, A., Welch, D., and Drummond, A. J. (2014). Efficient bayesian inference under the structured coalescent. *Bioinformatics*, 30(16):2272–2279.