# Phylogeography and structured populations

Lecturer: Tim Vaughan

Department of Biosystems Science and Enginerring
ETH Zürich, Switzerland

Taming the BEAST, London, July 2017

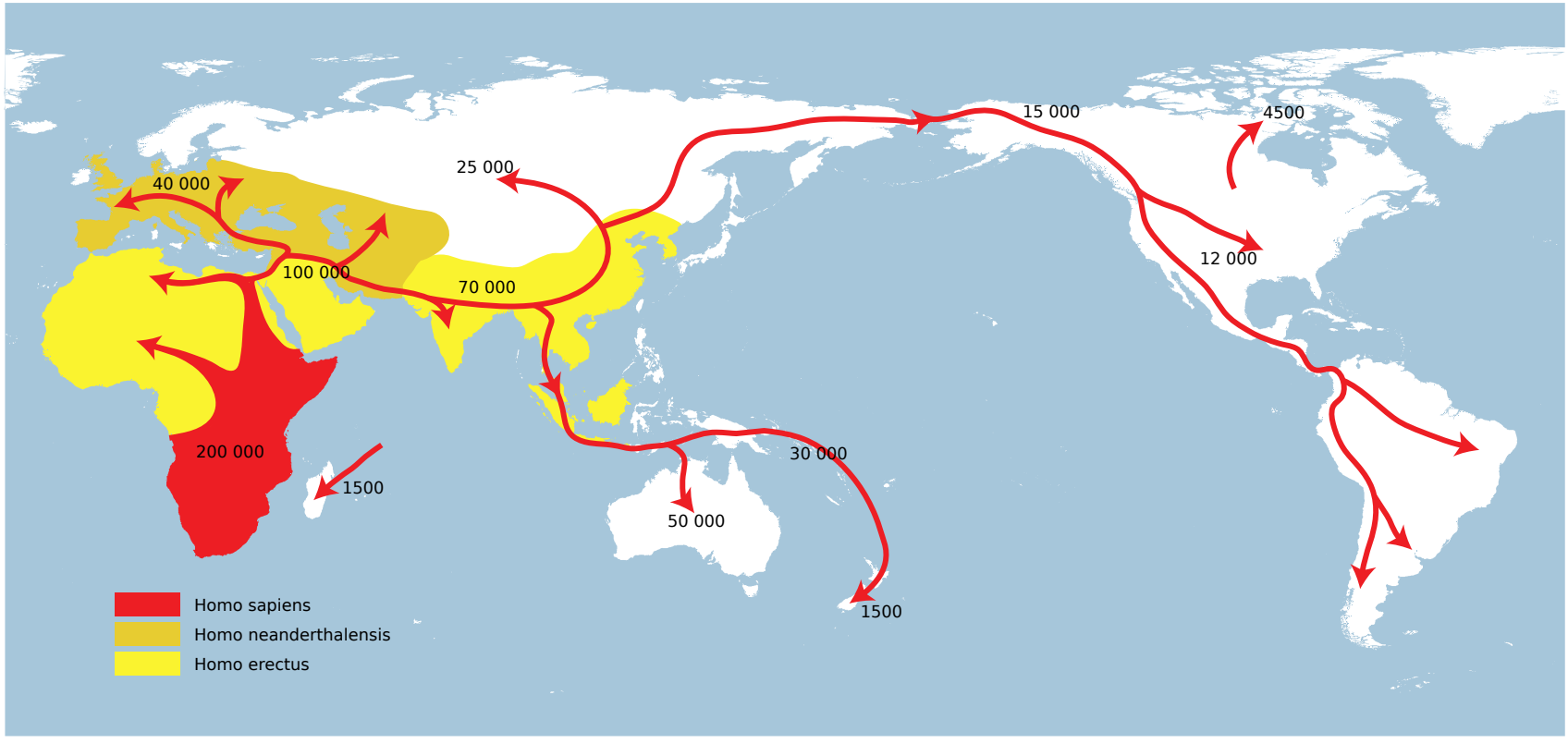
# What is Phylogeography?

Phylogeography is a field of study concerned with the principles and processes governing the geographic distributions of genealogical lineages, especially those within and among closely related species.

Avise (2000)

# What is Phylogeography?

Early human migrations:




Wikipedia

# 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.

# 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),

# 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,

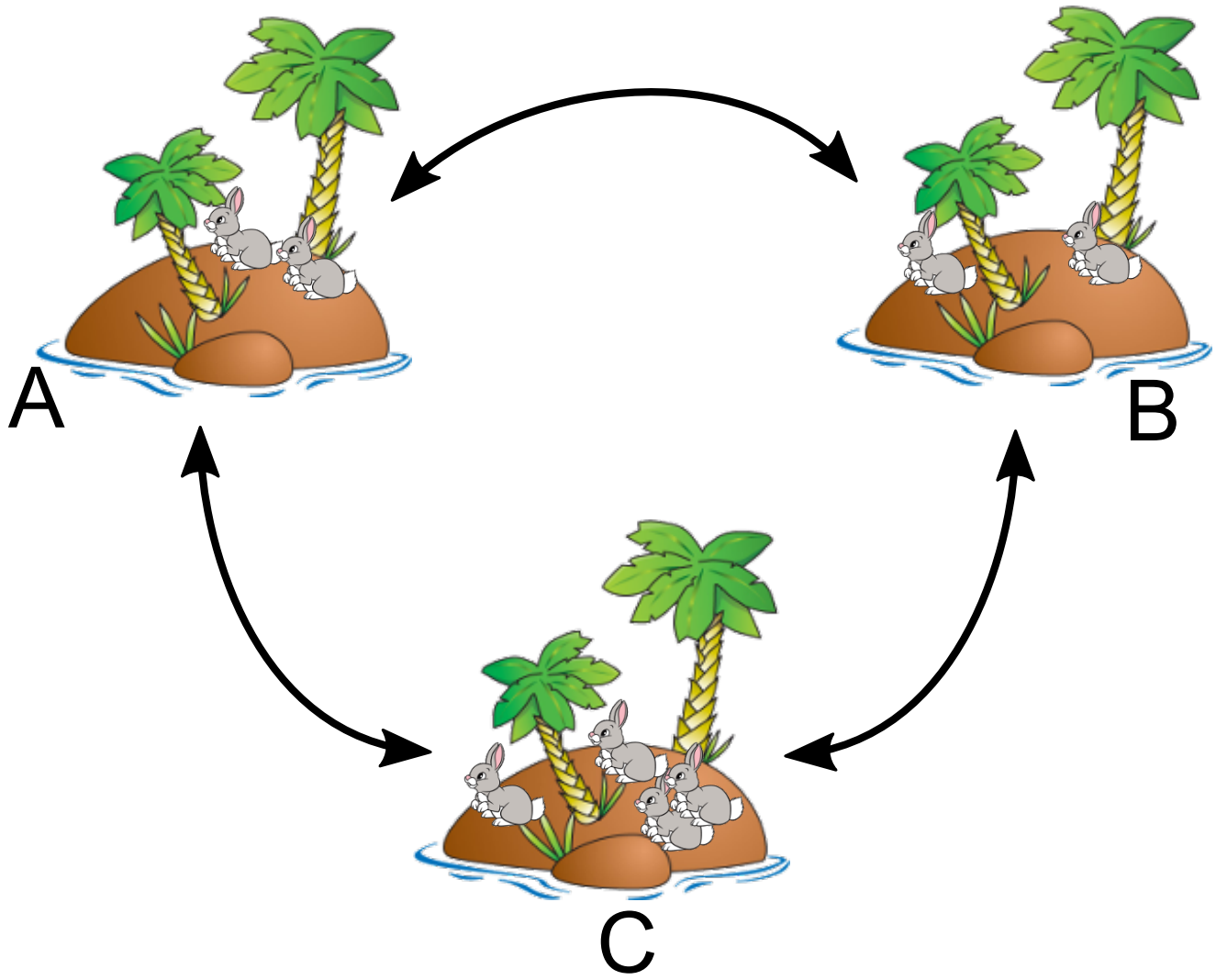
# What is a structured population?

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

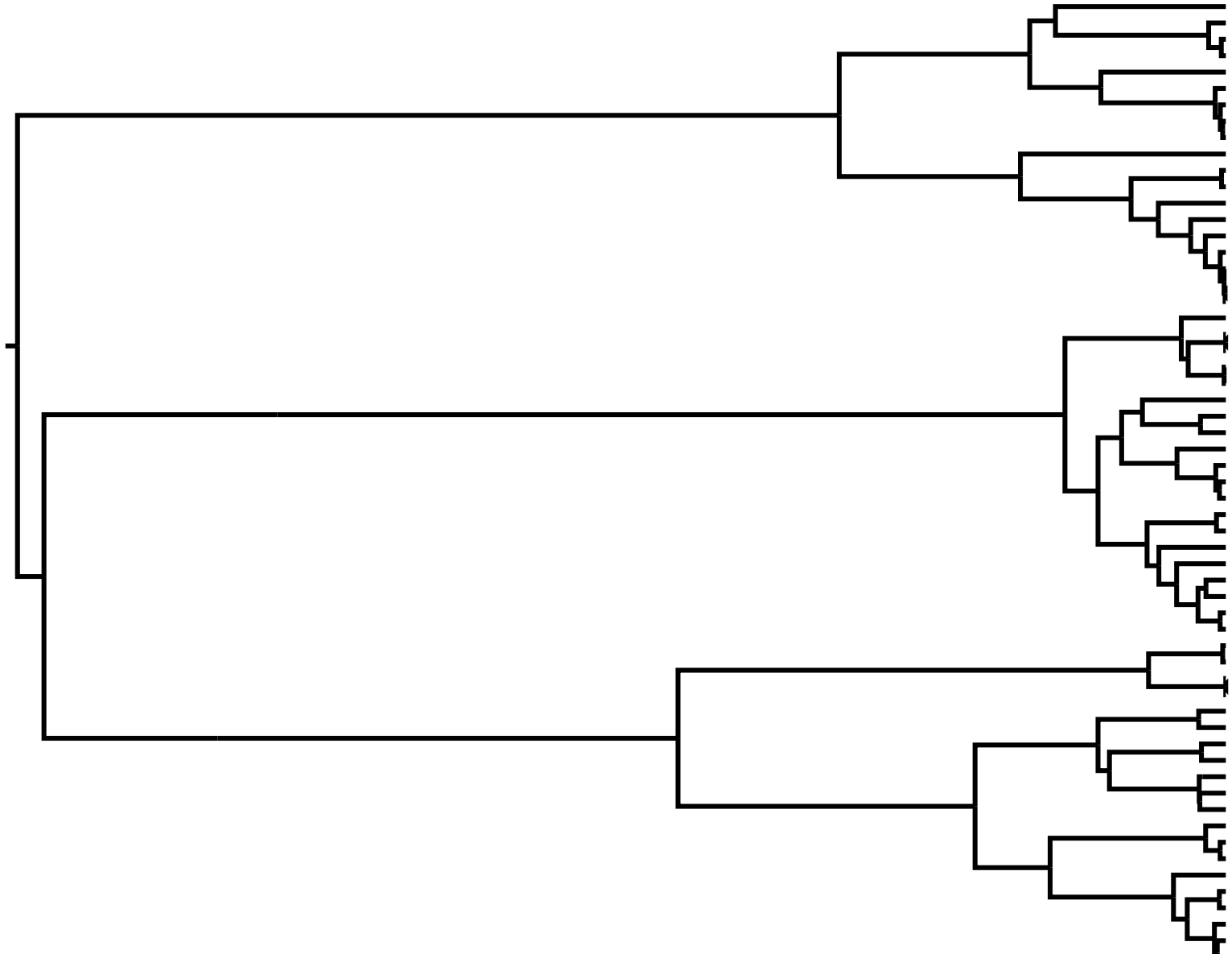
The island model is a common discrete model of spatial structure:




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:

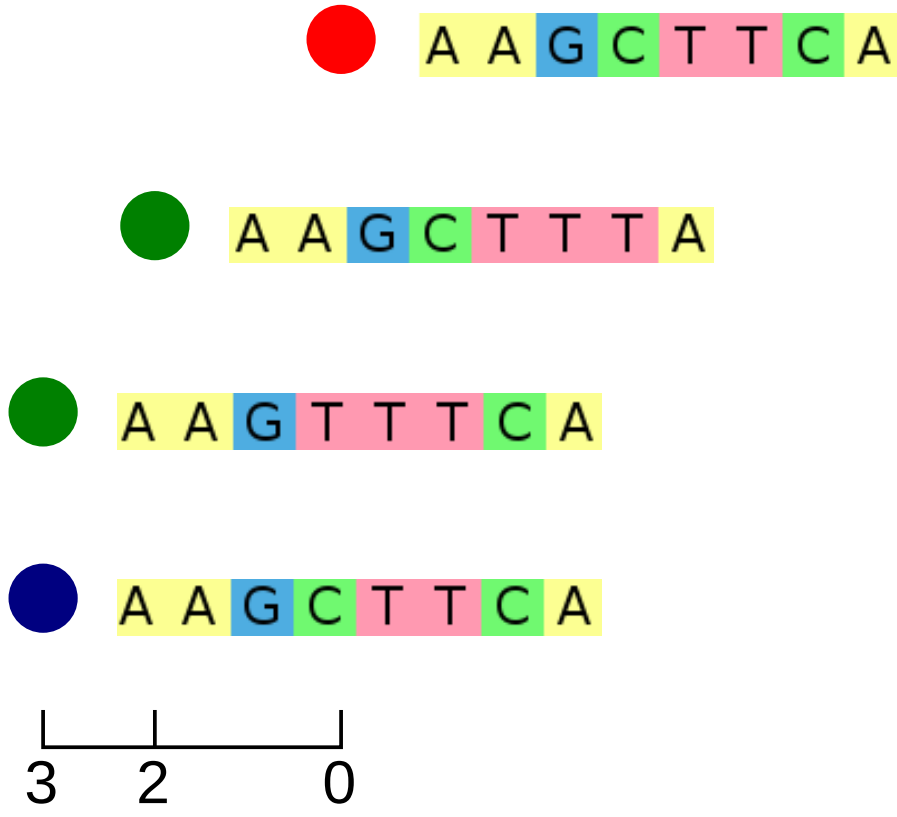


# Phylogeographic inference data

| 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 |

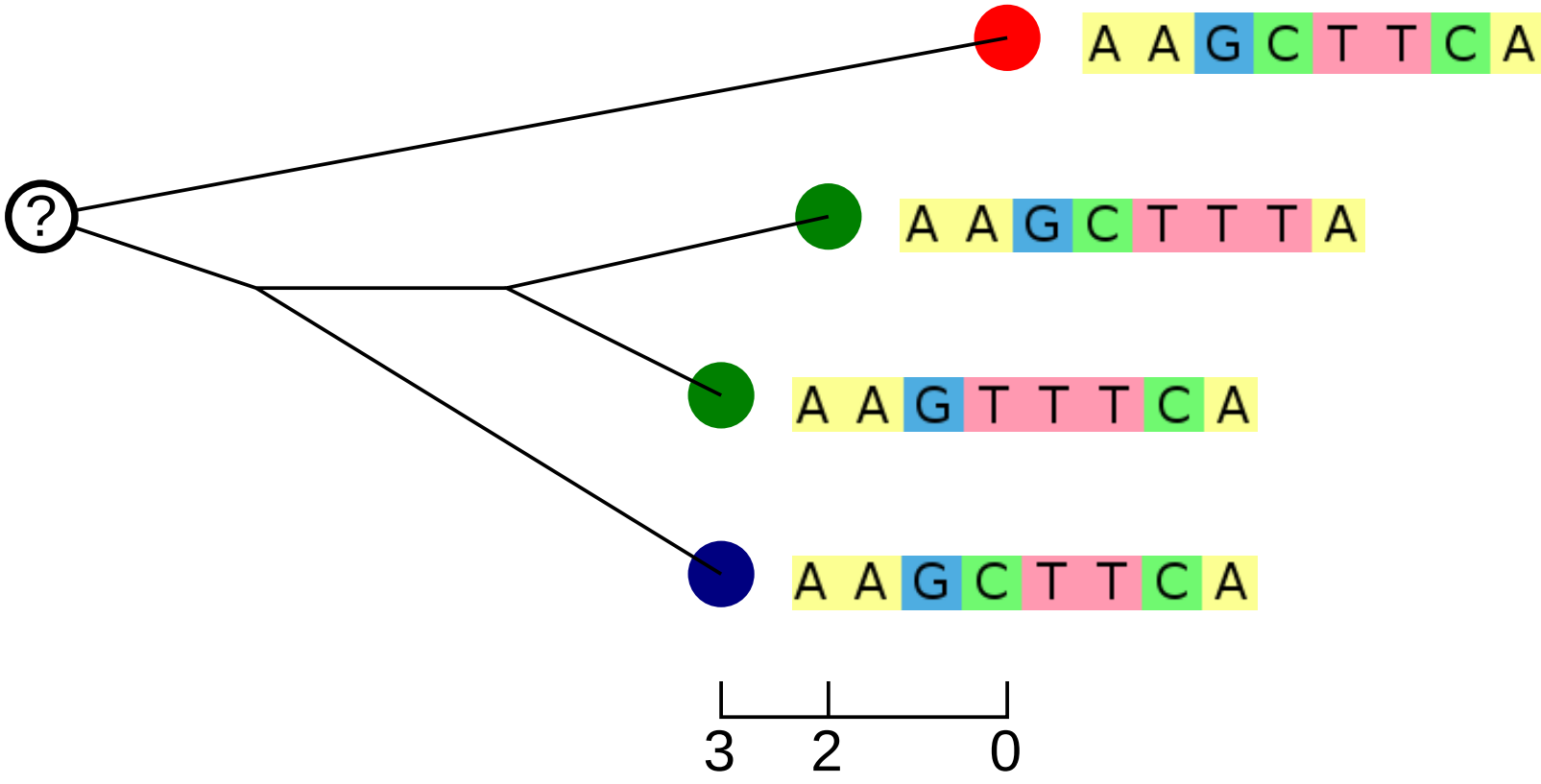
# Phylogeographic inference questions

Common questions include:



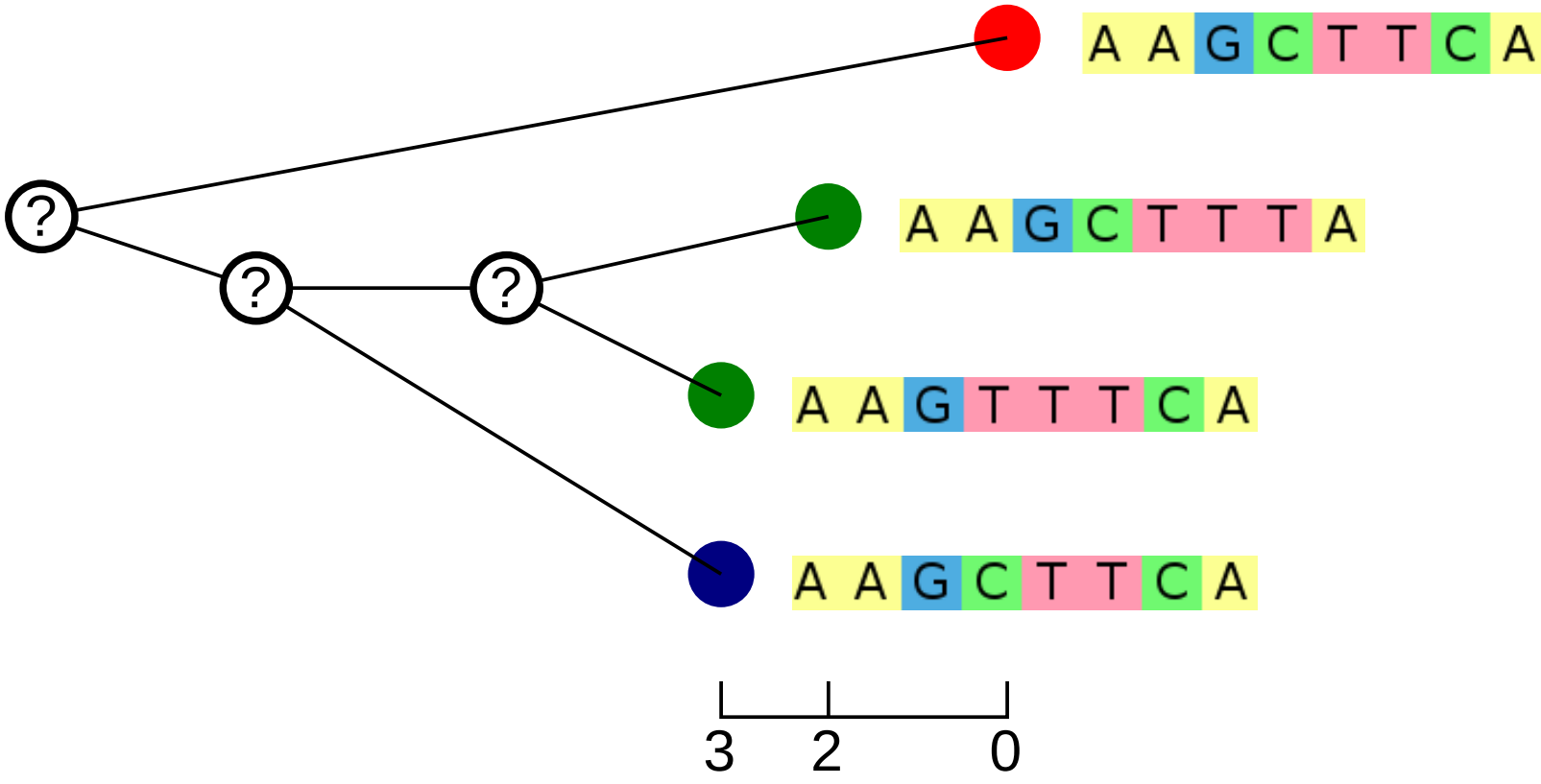
# Phylogeographic inference questions

Common questions include:



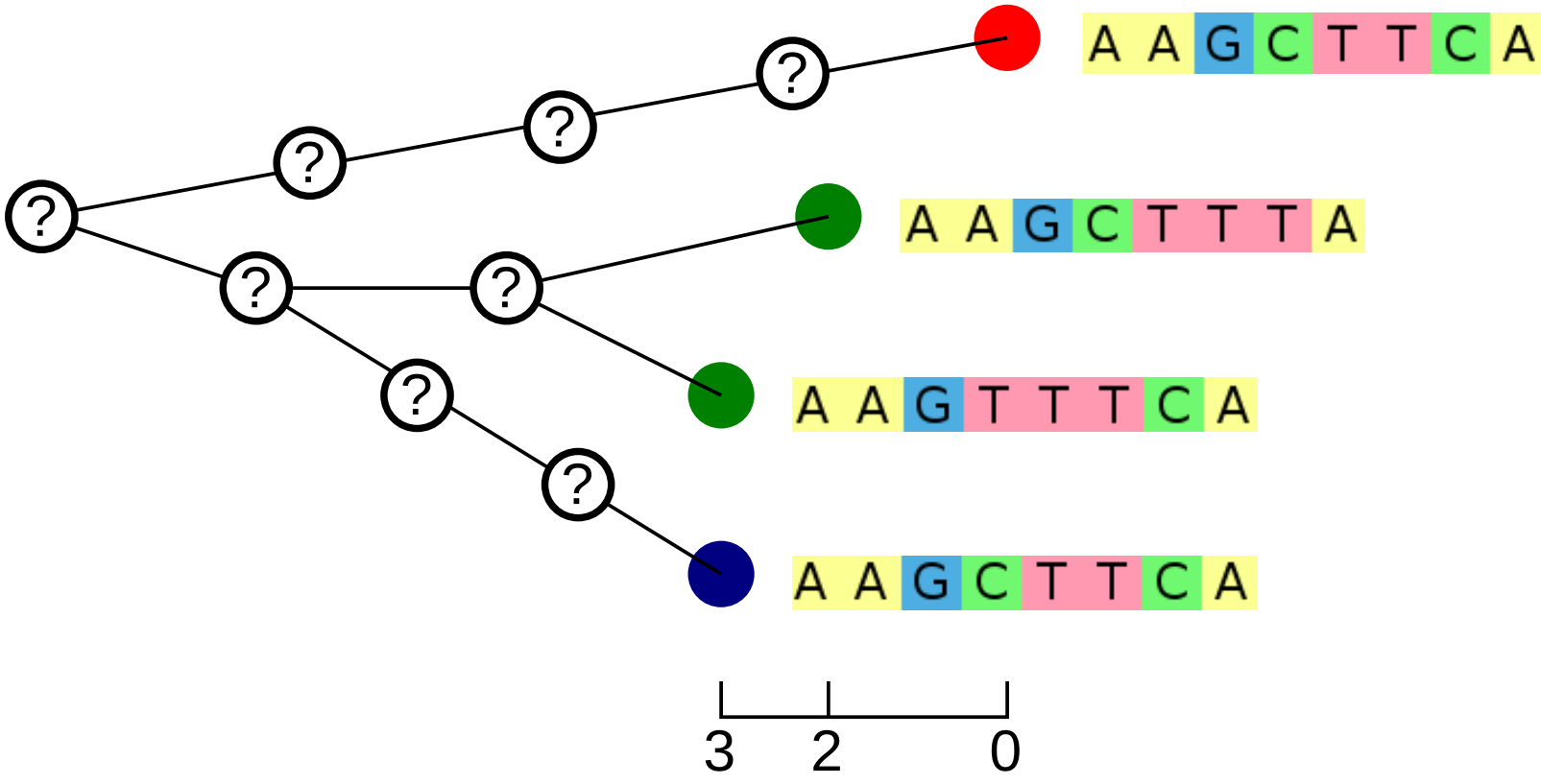
# Phylogeographic inference questions

Common questions include:



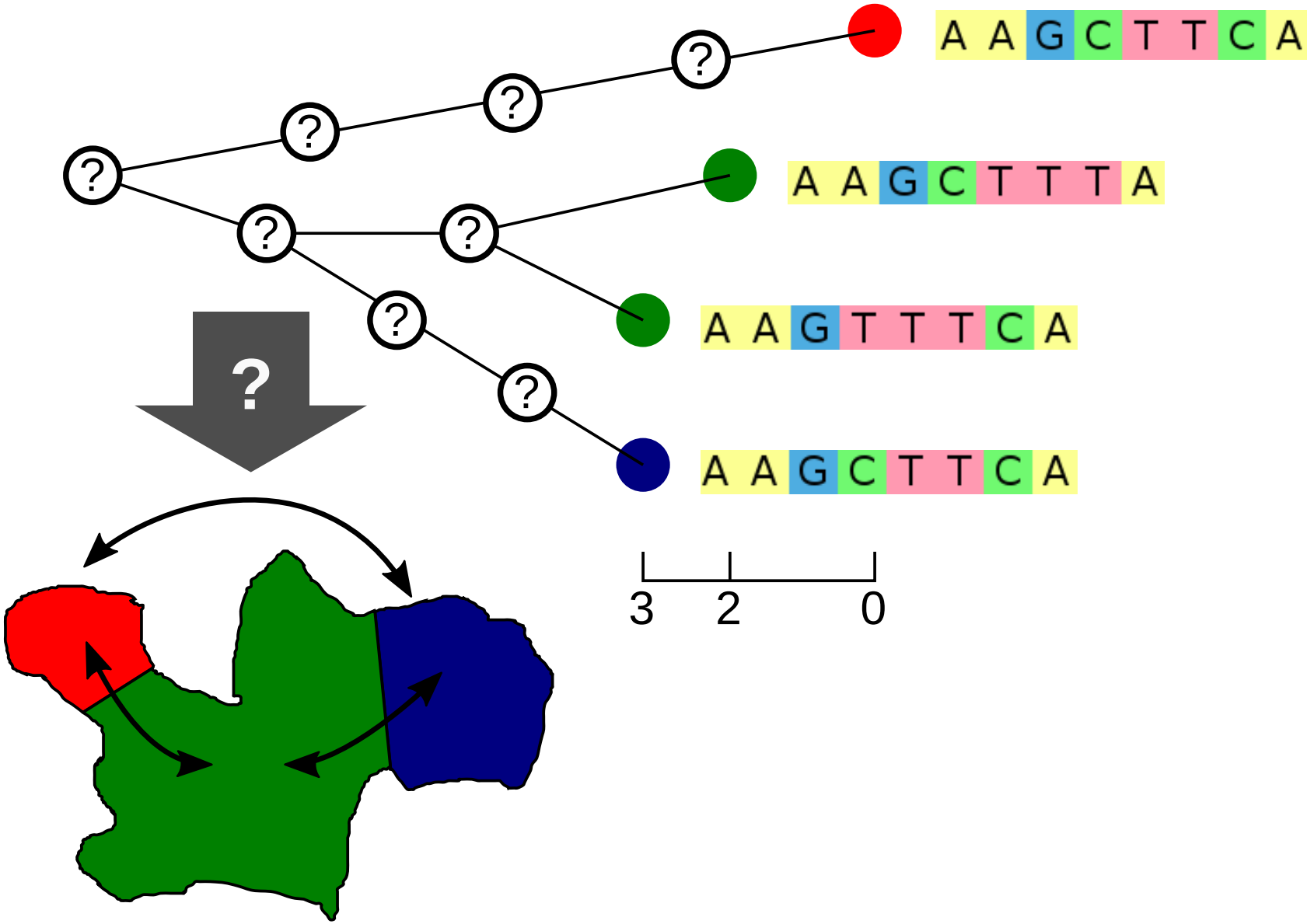
# Phylogeographic inference questions

Common questions include:



# Phylogeographic inference questions

Common questions include:



# Bayesian Phylogeographic Inference?

The usual phylogenetic posterior is:

$$P(T, \mu, \theta|A) = \frac{1}{P(A)} P(A|T, \mu) P(T|\theta) P(\mu) P(\theta)$$

where

$P(A|T, \mu)$ is a the *tree likelihood*,

$P(T|\theta)$ is the *tree prior*, and

$P(\mu)$ and $P(\theta)$ are the *parameter priors*.

# Bayesian Phylogeographic Inference?

The usual phylogenetic posterior is:

$$P(T, \mu, \theta|A) = \frac{1}{P(A)} P(A|T, \mu) P(T|\theta) P(\mu) P(\theta)$$

where

$P(A|T, \mu)$ is a the *tree likelihood*,

$P(T|\theta)$ is the *tree prior*, and

$P(\mu)$ and $P(\theta)$ are the *parameter priors*.

Where does geography fit in?

# 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:

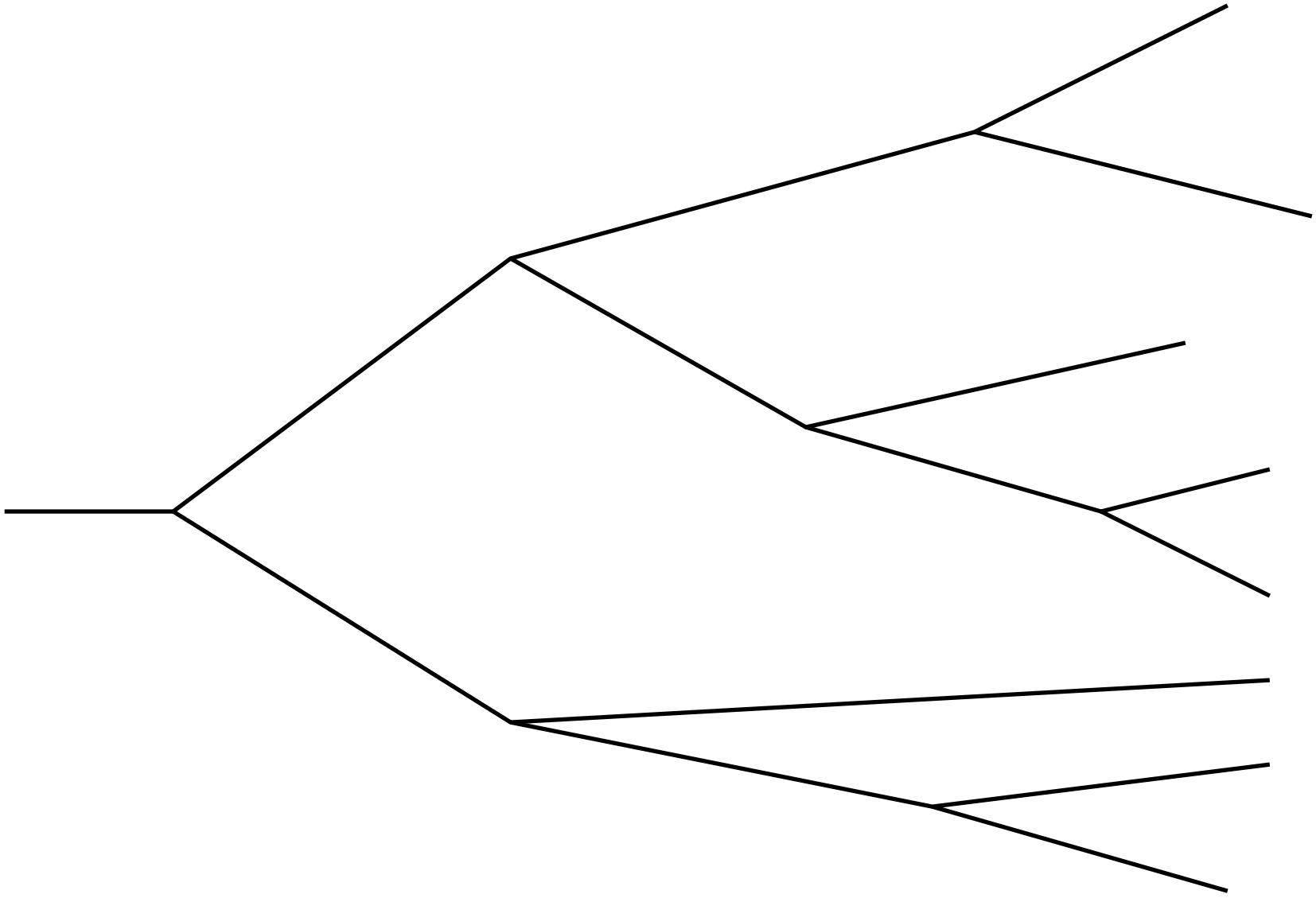
- **Mugration models (also Discrete Trait Analysis):**
  - Given tree and root location, what is the probability of sample locations?
  - Exist in continuous and discrete forms.
  - Developed by Phillipe Lemey et al. (Lemey et al., 2009, 2010).
- **Structured population models:**
  - Given sequences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

# Part I

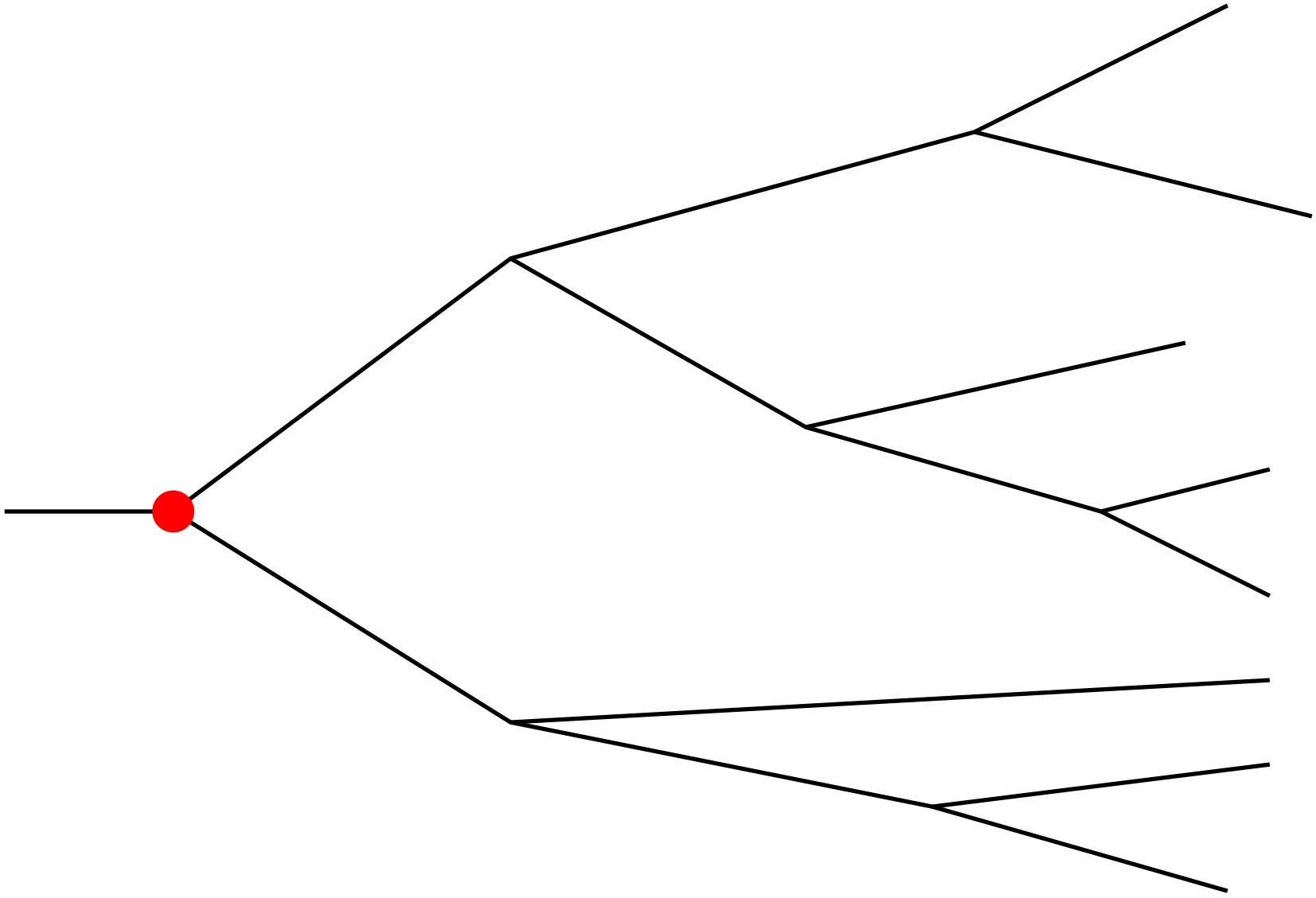
## Mugration models

# Discrete mugration model

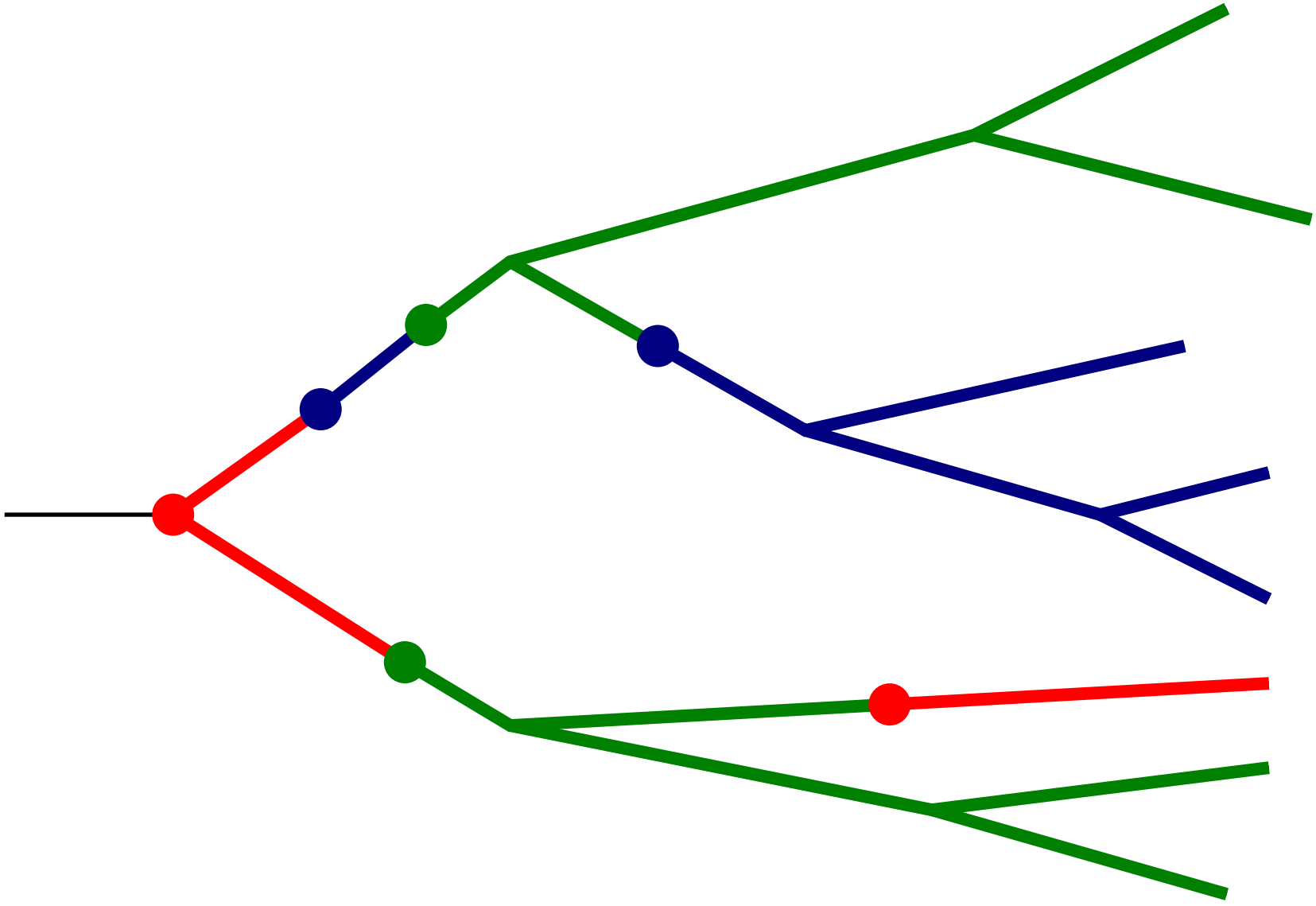
# Discrete mugration model



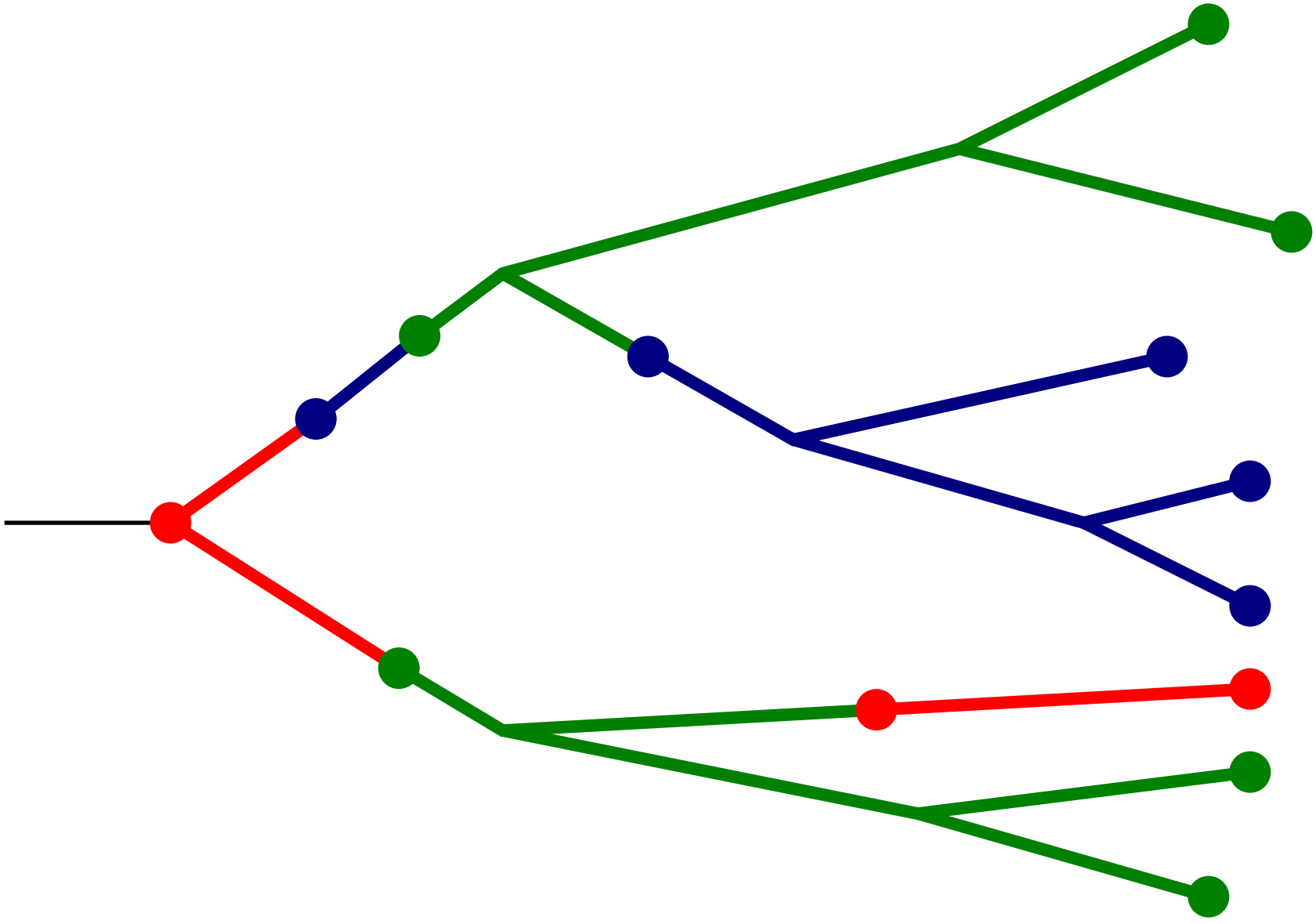
# Discrete mugration model



# Discrete mugration model



# Discrete mugration model



# Mugration Inference: Modified tree likelihood

The standard phylogenetic posterior is modified:

$$P(T, \mu, \theta|A, L) = \frac{1}{P(A, L)} P(A|T, \mu) P(L|T, M) \times P(T|\theta) P(\mu) P(M) P(\theta)$$

where

- $L$ are the sampled locations, and
- $M$ is a matrix specifying the random walk.

# Mugration Inference: Modified tree likelihood

The standard phylogenetic posterior is modified:

$$P(T, \mu, \theta|A, L) = \frac{1}{P(A, L)} P(A|T, \mu) P(L|T, M) \times P(T|\theta) P(\mu) P(M) P(\theta)$$

where

- $L$ are the sampled locations, and
- $M$ is a matrix specifying the random walk.

Notice the similarity between the two likelihood terms.

# Mugration Inference: Modified tree likelihood

The standard phylogenetic posterior is modified:

$$P(T, \mu, \theta | A, L) = \frac{1}{P(A, L)} P(A|T, \mu) P(L|T, M) \times P(T|\theta) P(\mu) P(M) P(\theta)$$

where

- $L$ are the sampled locations, and
- $M$ is a matrix specifying the random walk.

Notice the similarity between the two likelihood terms.

**Mugration models treat location as just another trait/character.**

# Mugration transition matrix

The matrix $\mathrm{M}$ is often assumed to be symmetric and thus describe a reversible process. In exact analogy to the GTR model of nucleotide evolution, we can expand:

$$\mathrm{M} = \mathfrak{m}\mathrm{S}\Pi \tag{1}$$

where

- $\mathfrak{m}$ is the average overall transition rate,
- $\mathrm{S}$ is the normalized transition rate matrix, and
- $\Pi$ is a diagonal matrix containing the equilibrium probabilities for each location in the long term.

# Sampling assumption

The following very important assumption is made by the mugration model posterior:

# Sampling assumption

The following very important assumption is made by the mugration model posterior:

**Samples are to be collected in a manner that is blind to their location.**

# Sampling assumption

The following very important assumption is made by the mugration model posterior:

**Samples are to be collected in a manner that is blind to their location.**

- Mugration models use sample location as data.

# Sampling assumption

The following very important assumption is made by the mugration model posterior:

**Samples are to be collected in a manner that is blind to their location.**

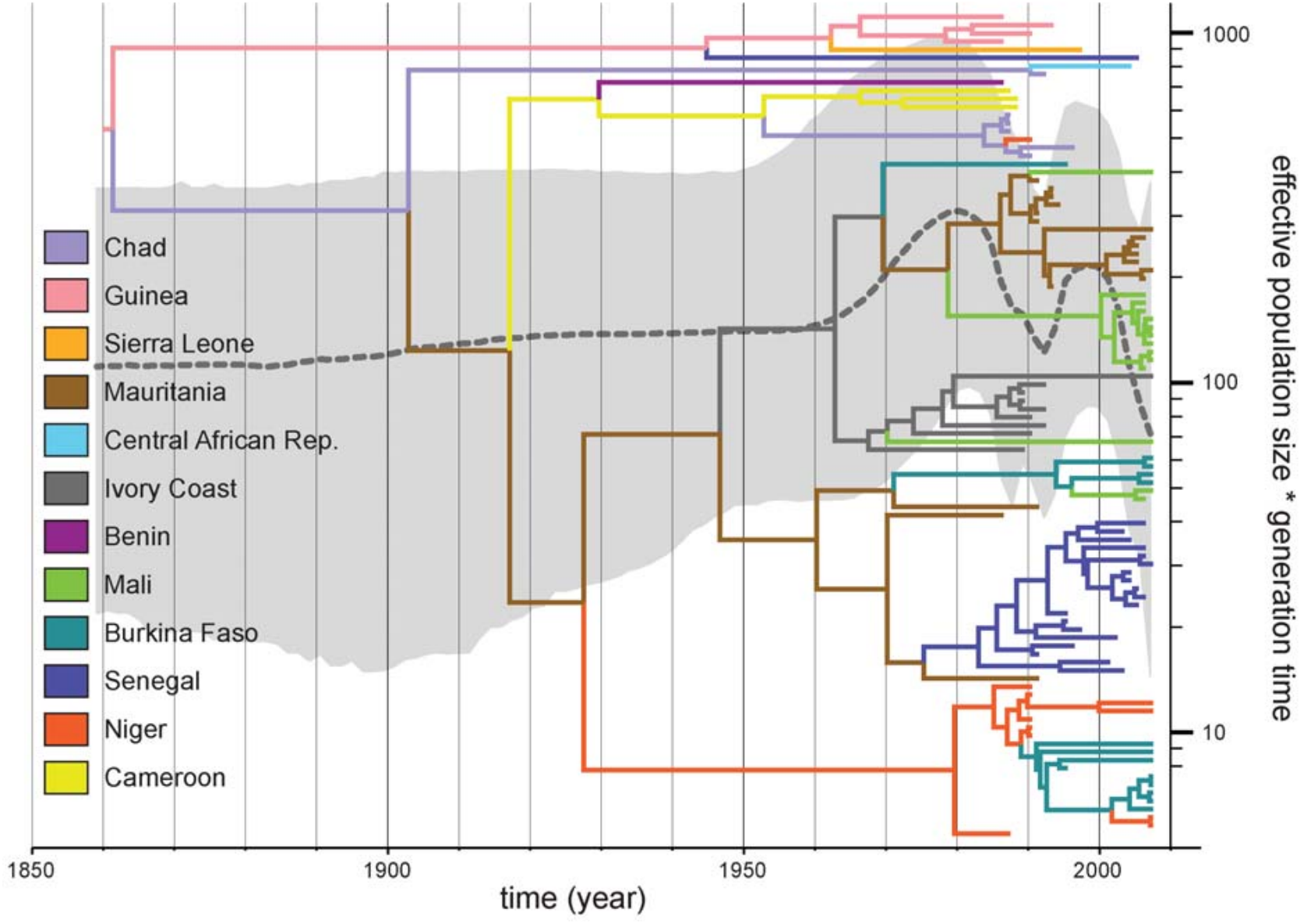
- Mugration models use sample location as data.
- Just as for genetic data, non-random sampling procedures will **bias results**.

# Sampling assumption

The following very important assumption is made by the mugration model posterior:

**Samples are to be collected in a manner that is blind to their location.**

- Mugration models use sample location as data.
- Just as for genetic data, non-random sampling procedures will **bias results**.
- This isn't a property of the generative model, just the way it is conditioned. (Conditioning on sampling location would prevent the use of regular unstructured tree priors.)



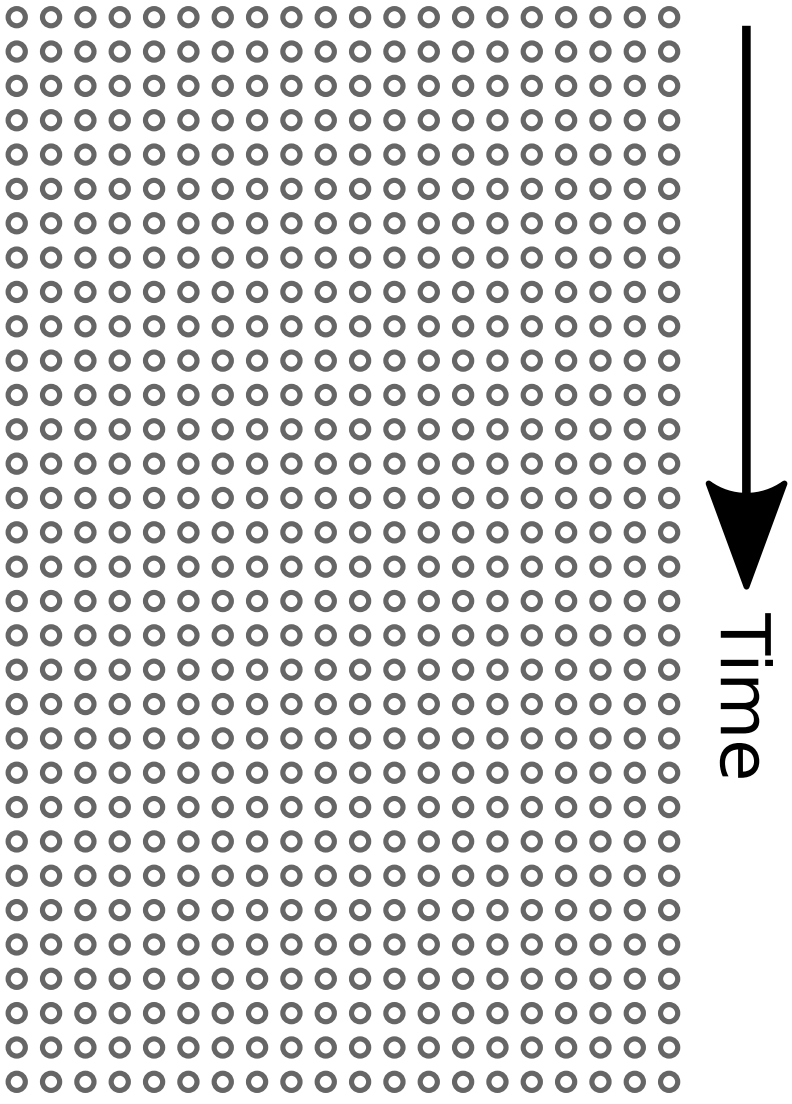

Analysis of "Africa 2" rabies virus, Lemey et al. (2009)

# Part II

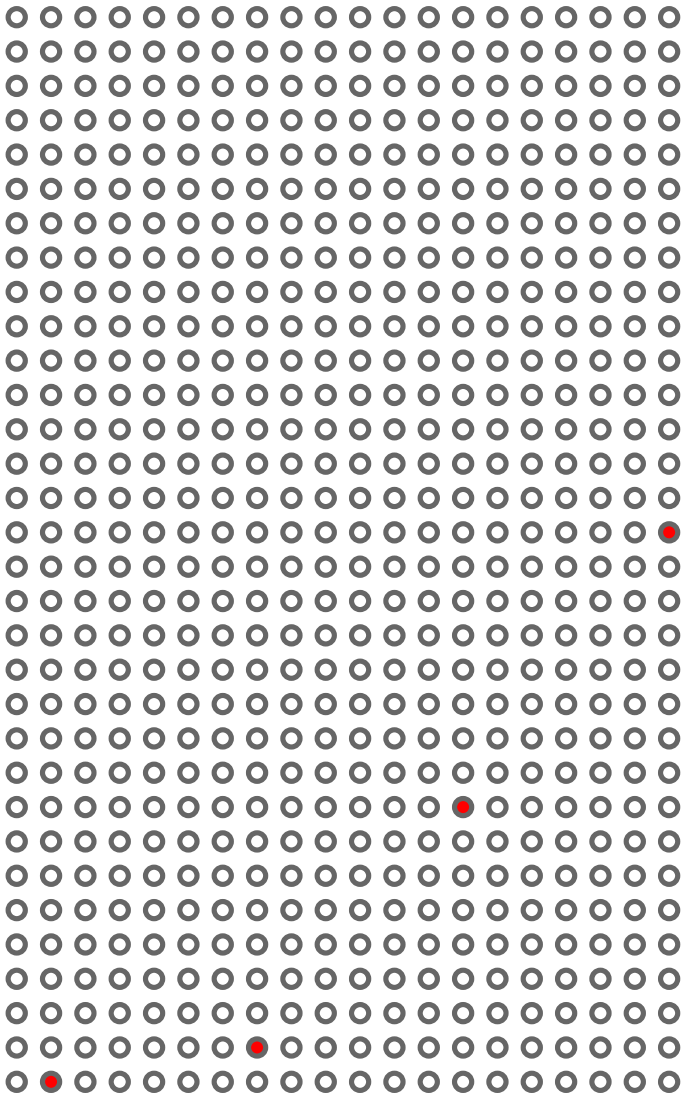
## Structured coalescent models

# Wright-Fisher model

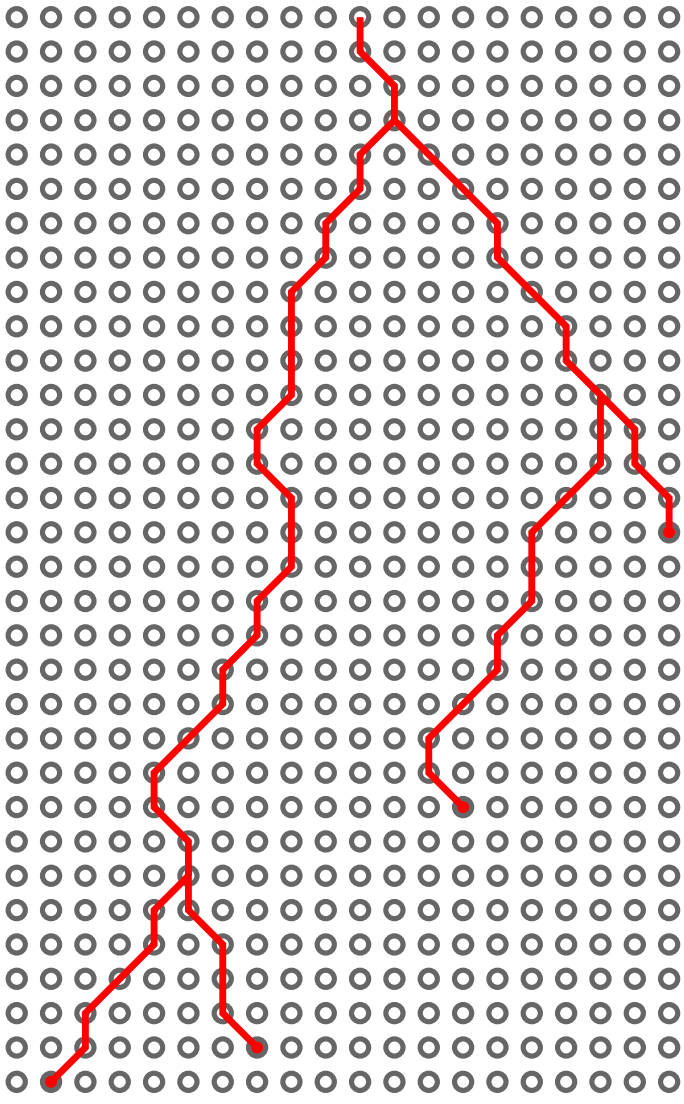
# Wright-Fisher model



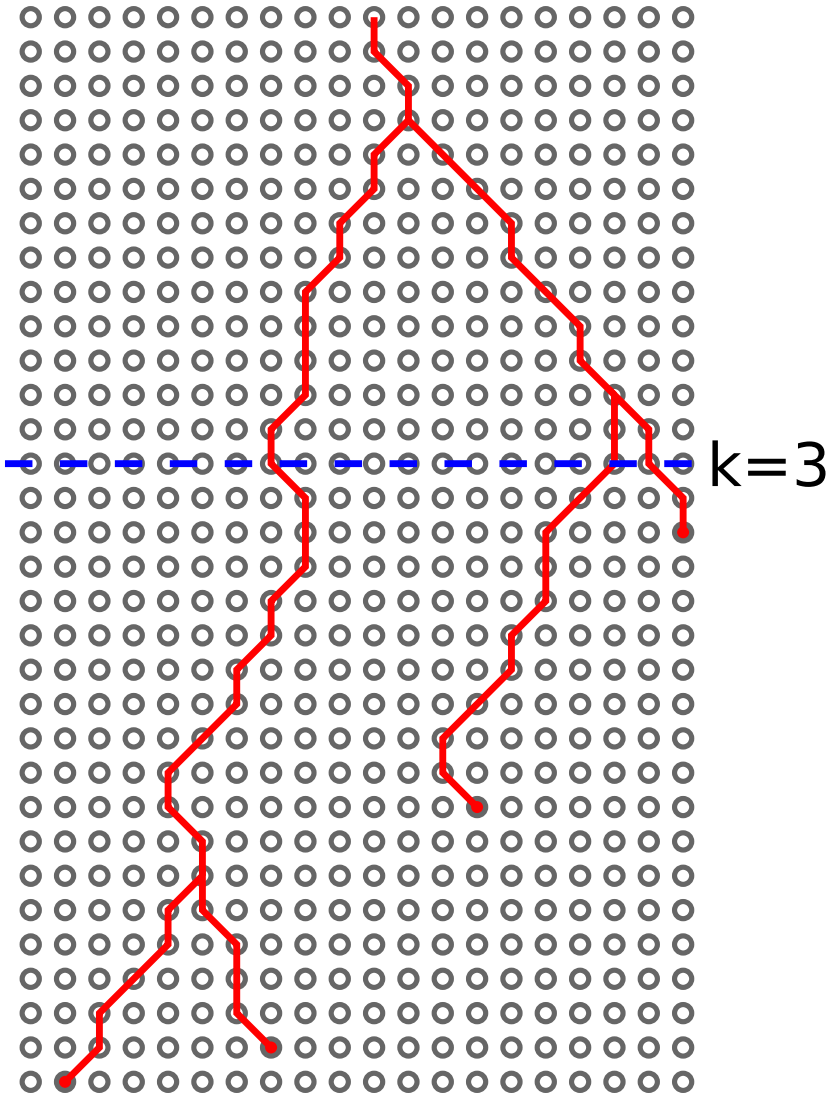
# Wright-Fisher model



# Wright-Fisher model



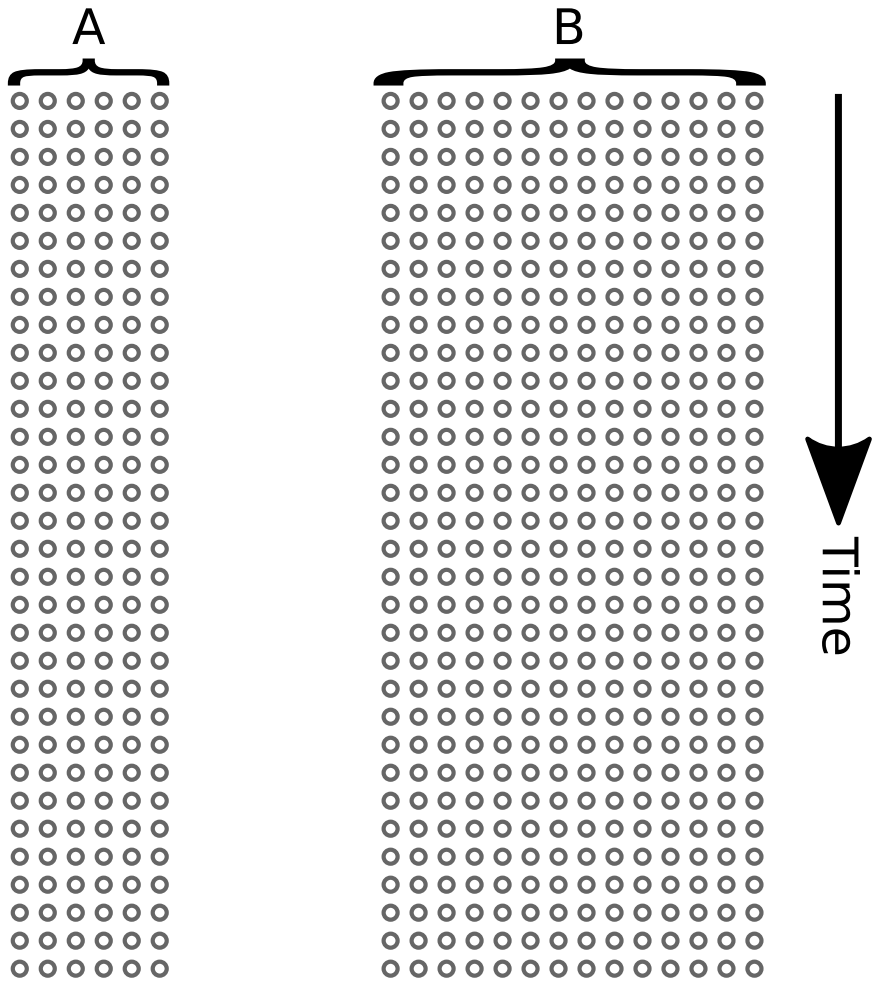
# Wright-Fisher model



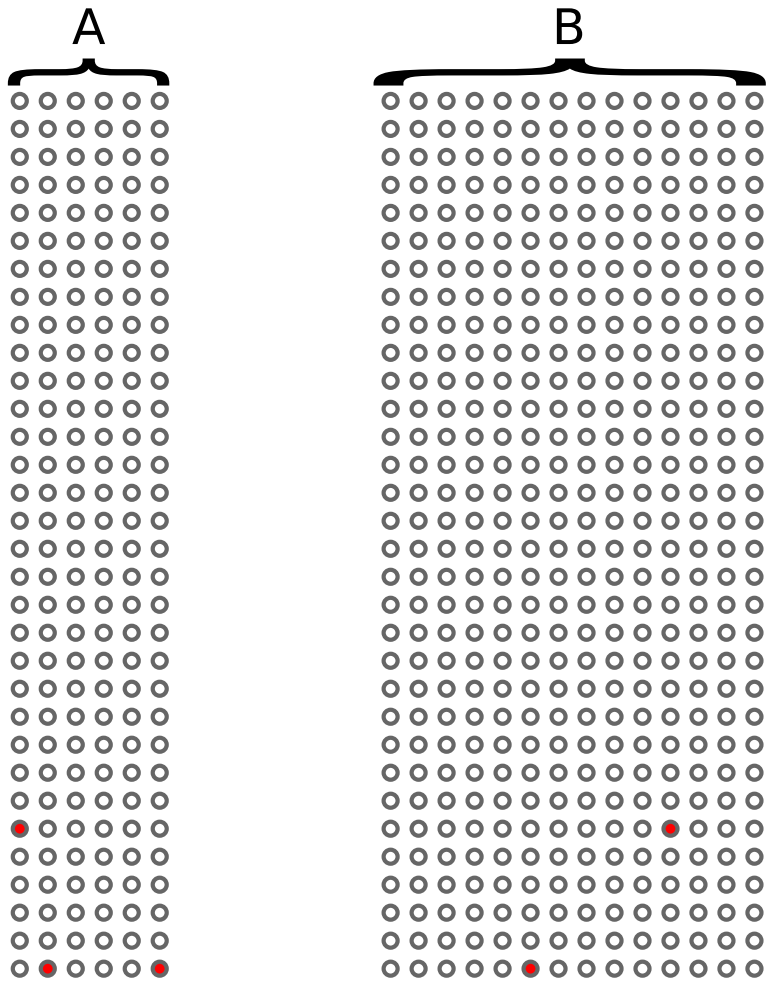

Probability of coalescence per generation:

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

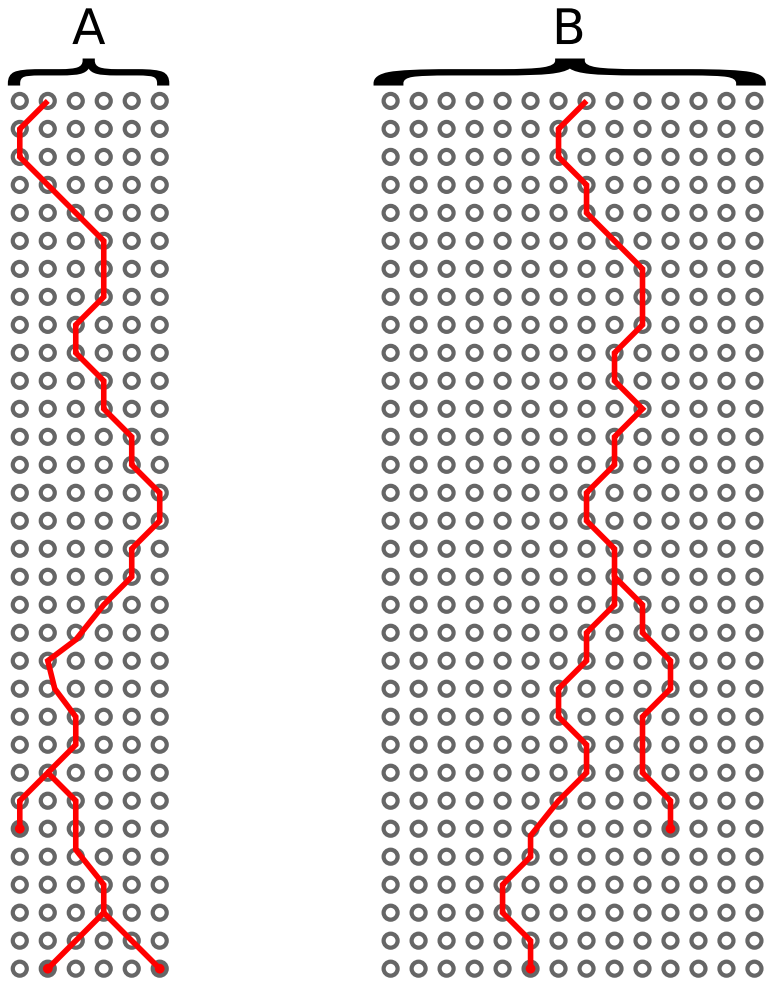
# Partitioned Wright-Fisher model



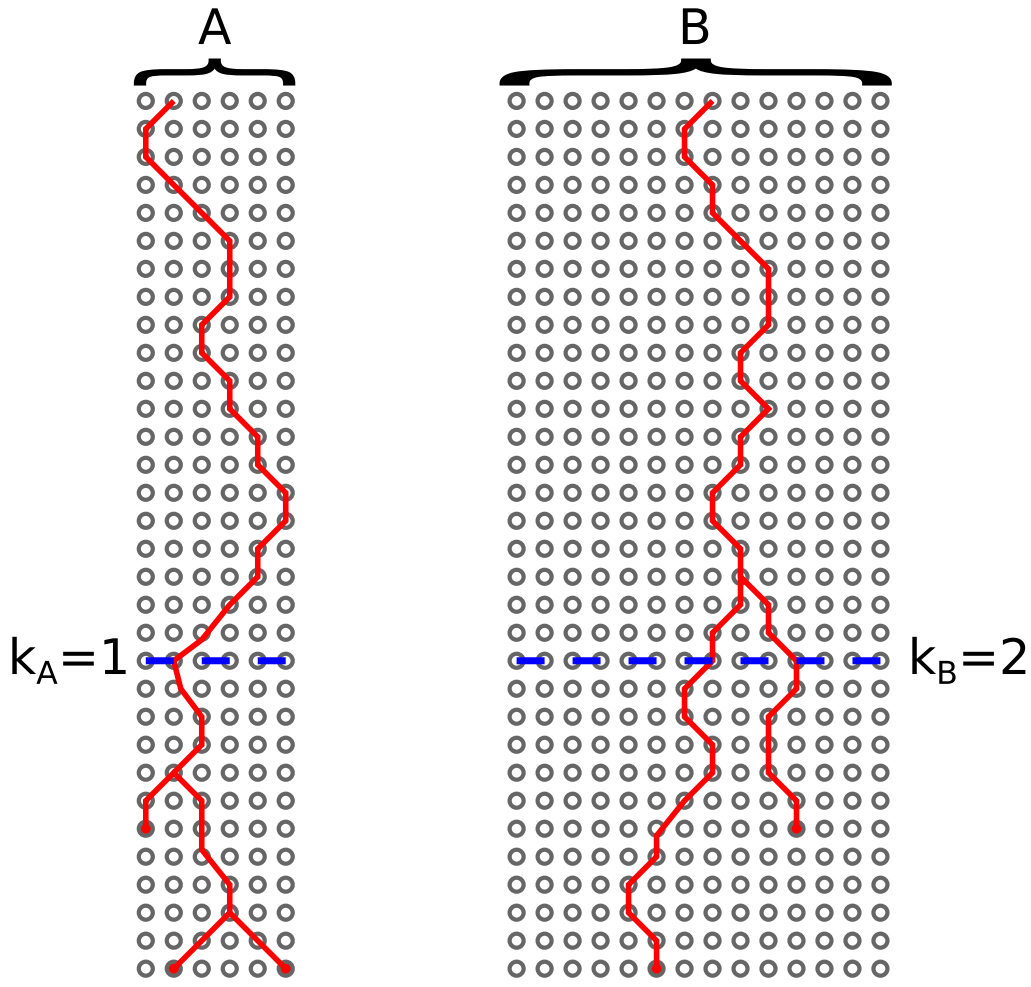
# Partitioned Wright-Fisher model



# Partitioned Wright-Fisher model



# Partitioned Wright-Fisher model



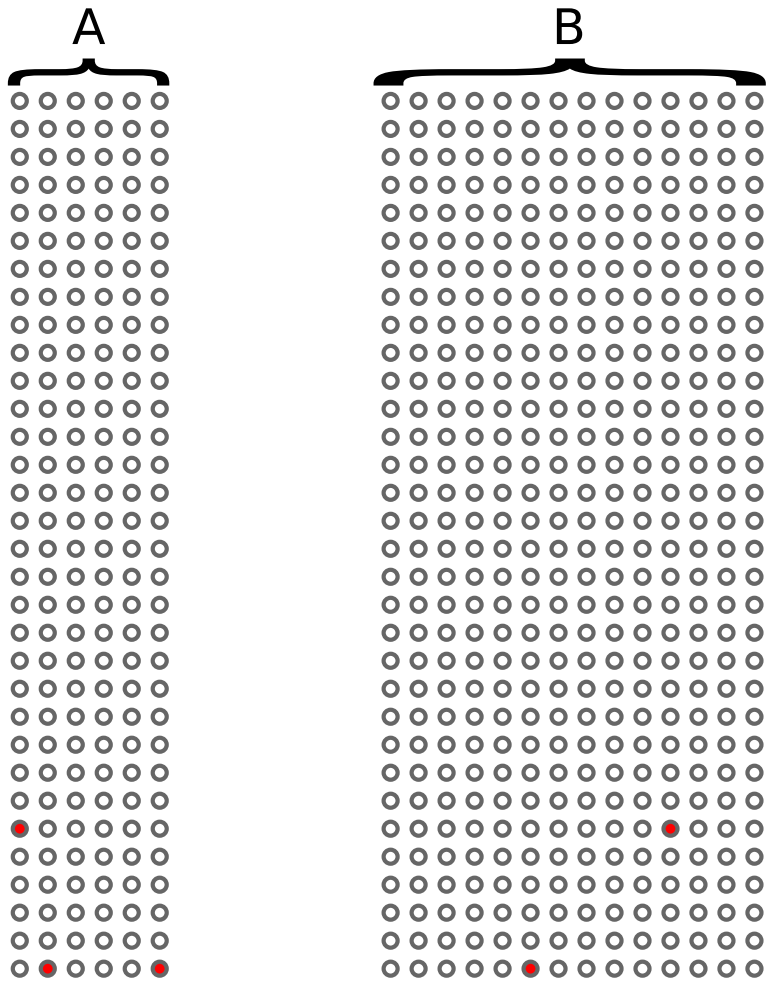

Probability of coalescence per generation in A:

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

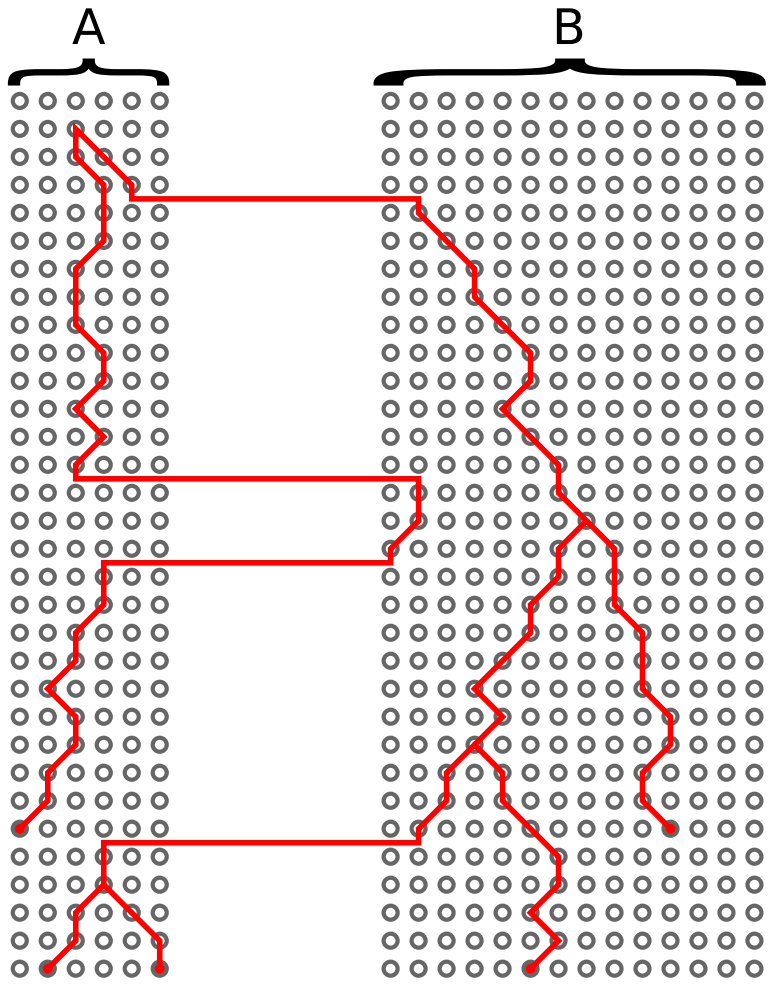
Probability of coalescence per generation in B:

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

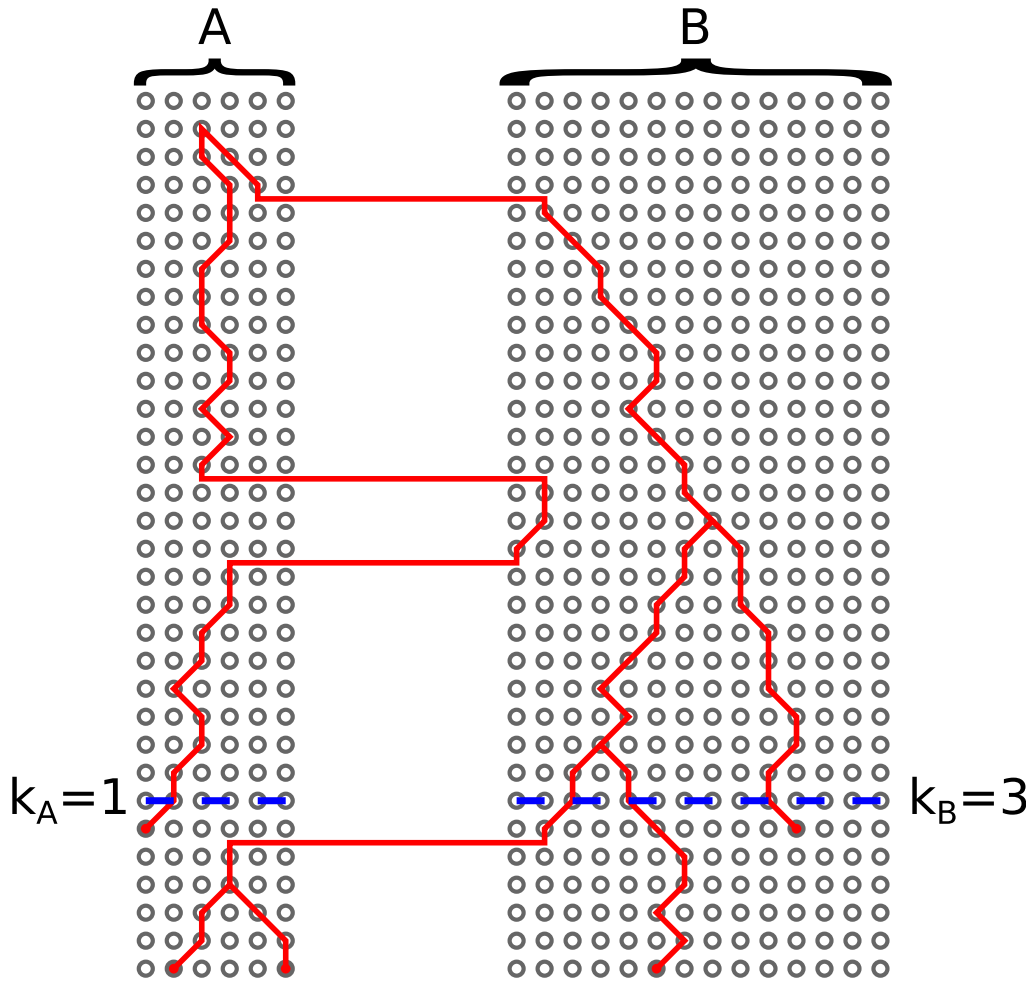
# Structured Wright-Fisher model



# 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)

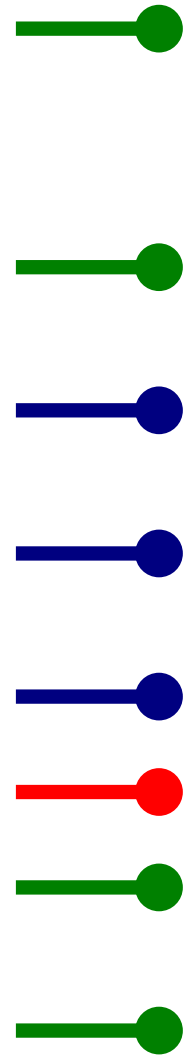
# Structured Coalescent

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

# Structured Coalescent

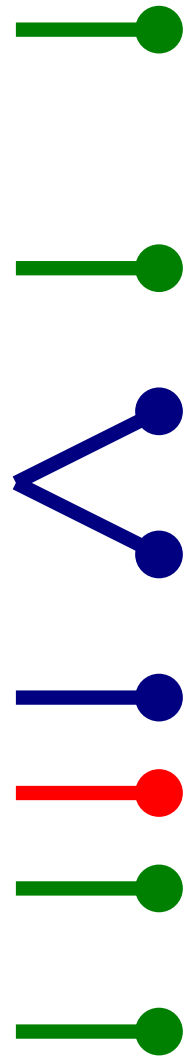
Backwards-in-time Markov process that generates both the sampled tree and ancestral locations.

(Hudson, 1990; Notohara, 1990)



# Structured Coalescent

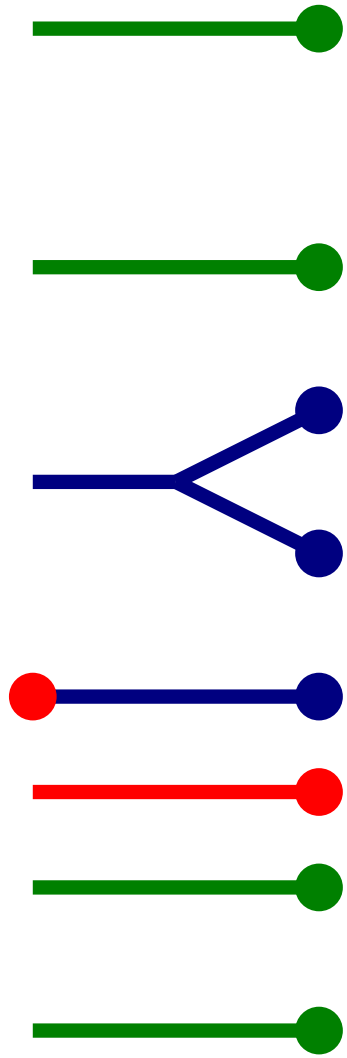
Backwards-in-time Markov process that generates both the sampled tree and ancestral locations.
(Hudson, 1990; Notohara, 1990)



# Structured Coalescent

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

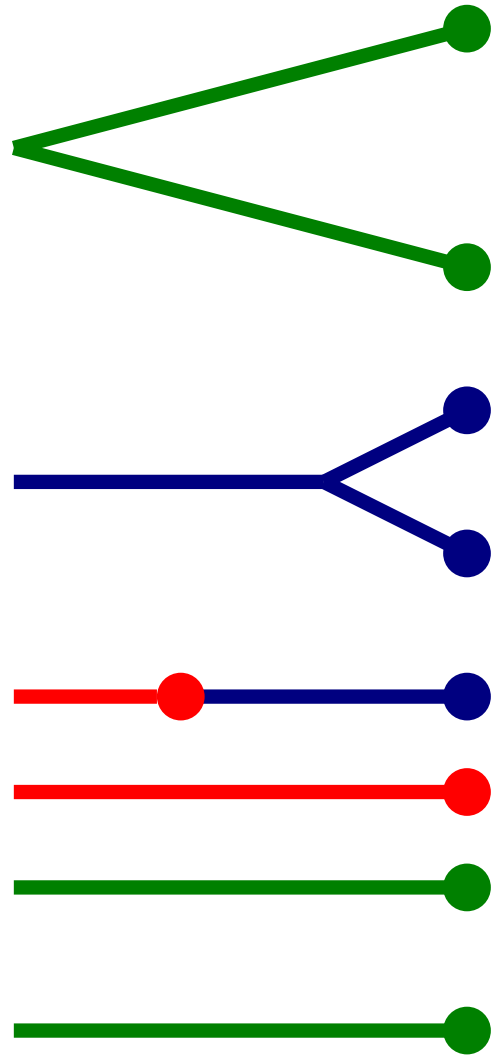
(Hudson, 1990; Notohara, 1990)



# Structured Coalescent

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

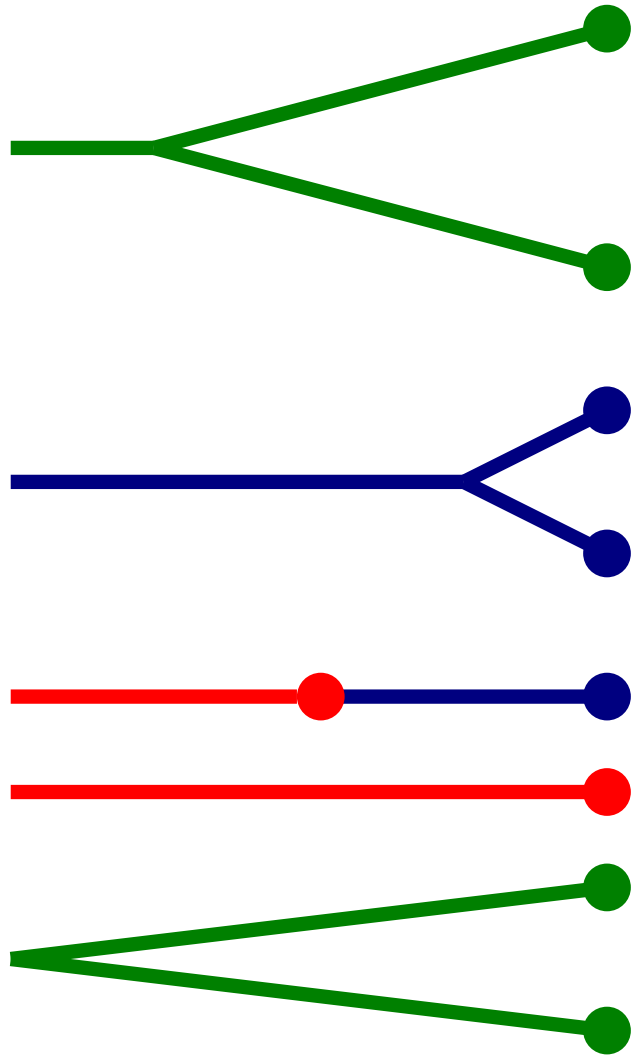
(Hudson, 1990; Notohara, 1990)



# Structured Coalescent

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

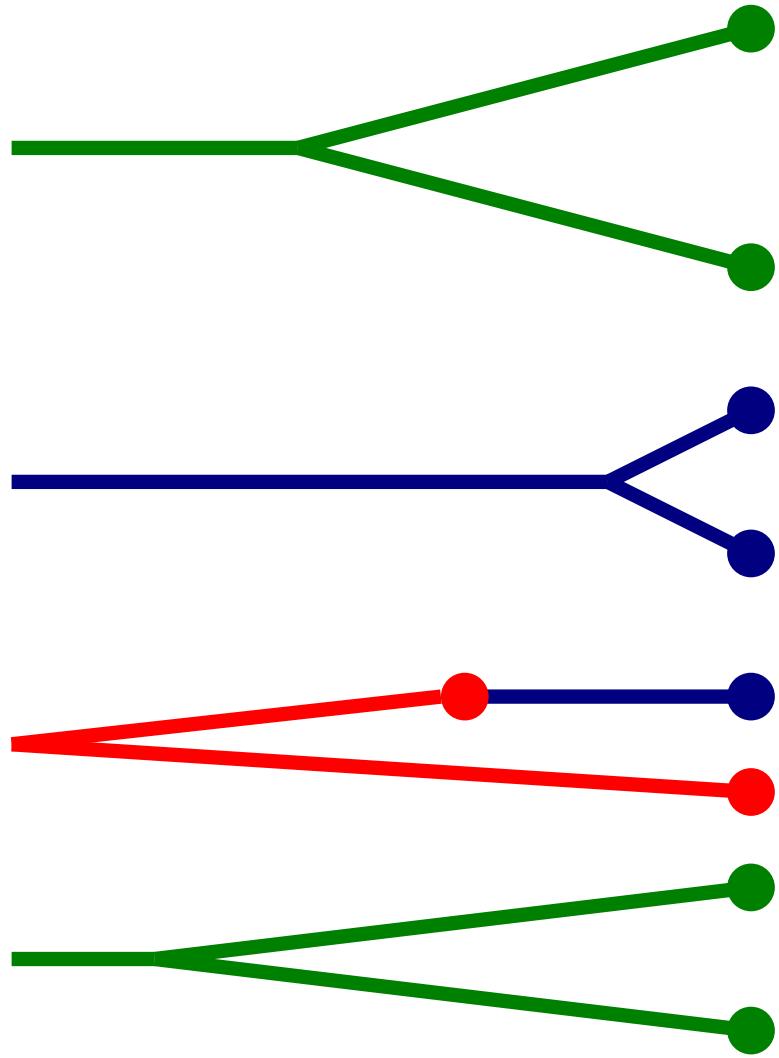
(Hudson, 1990; Notohara, 1990)



# Structured Coalescent

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

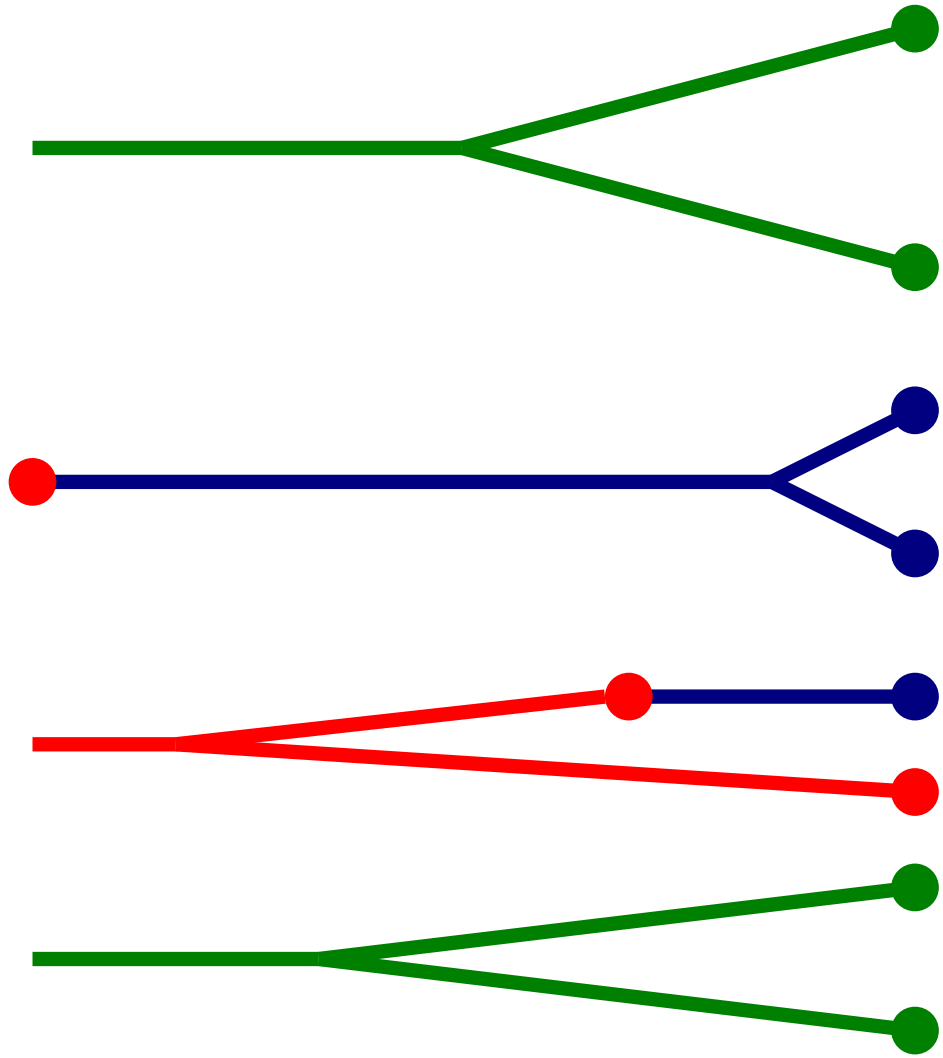
(Hudson, 1990; Notohara, 1990)



# Structured Coalescent

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

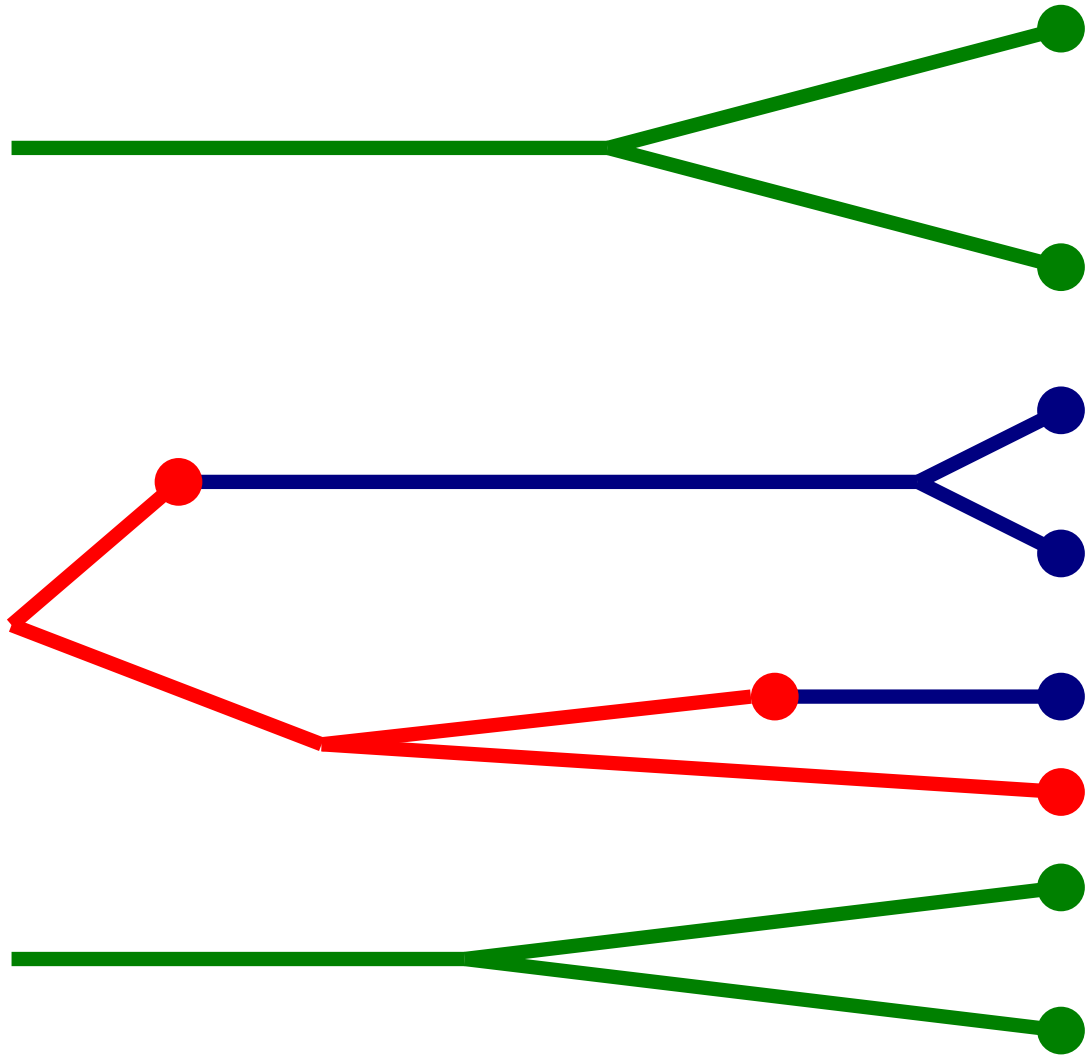
(Hudson, 1990; Notohara, 1990)



# Structured Coalescent

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

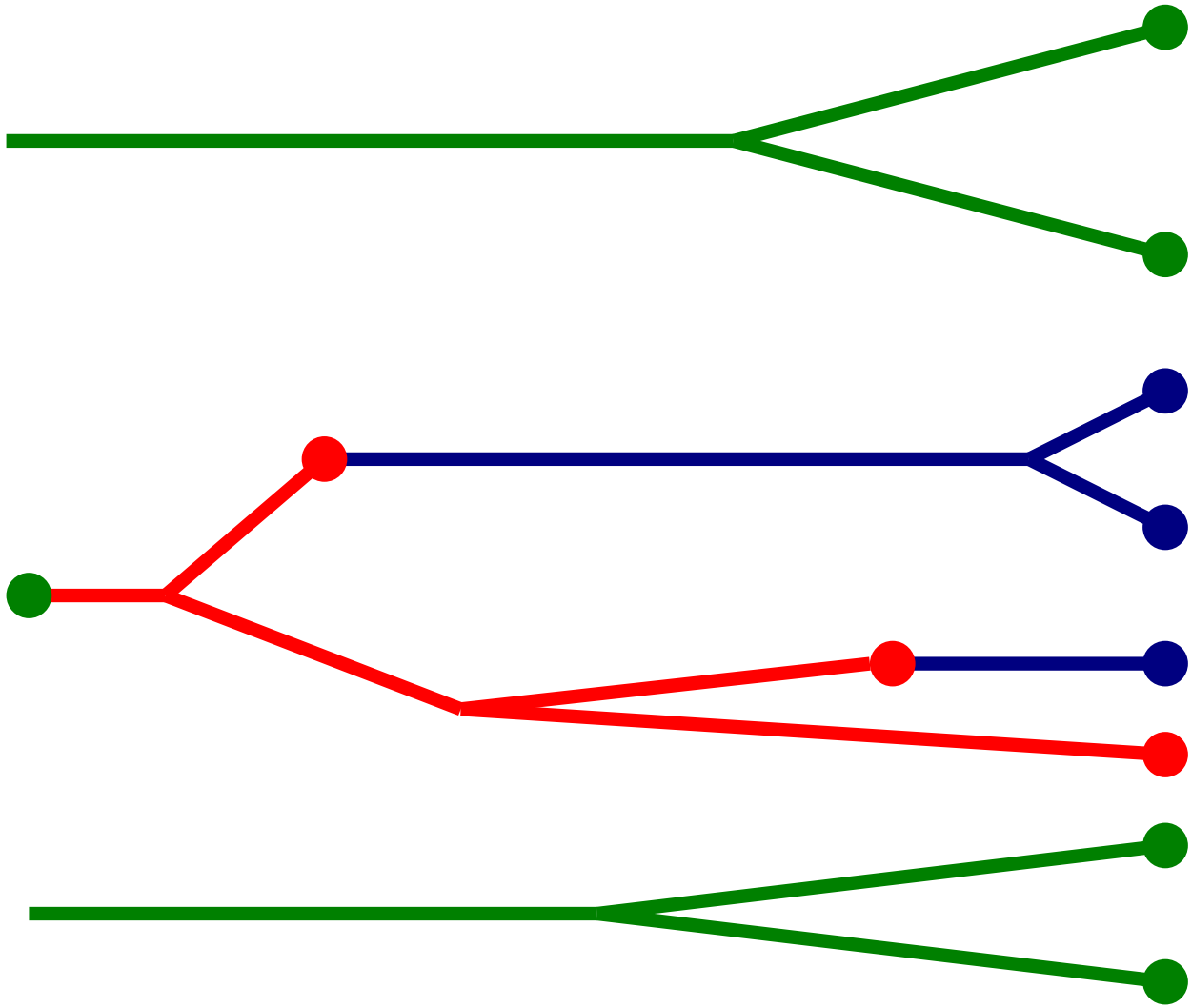
(Hudson, 1990; Notohara, 1990)



# Structured Coalescent

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

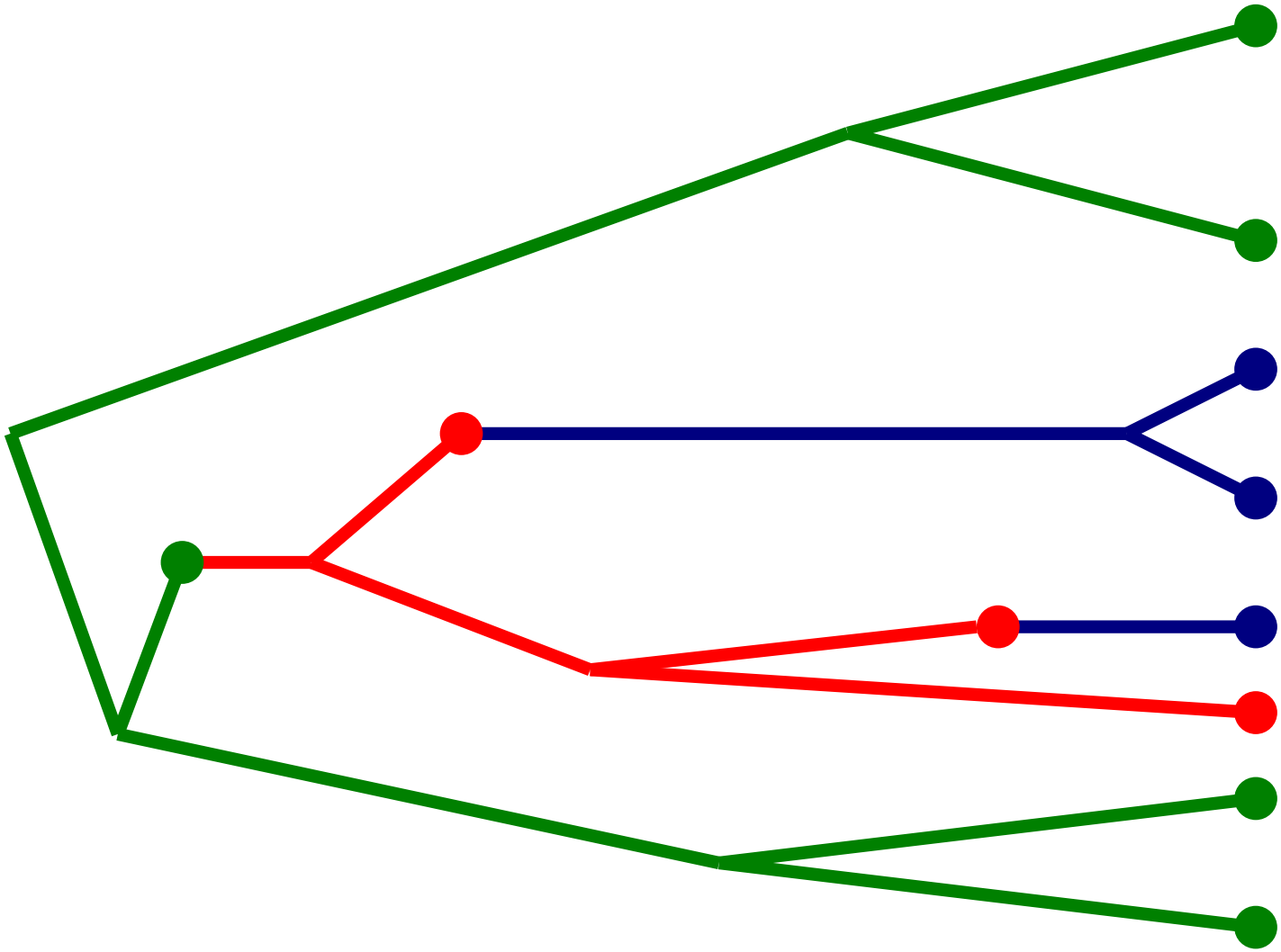
(Hudson, 1990; Notohara, 1990)



# 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

Again, 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,
- $\bar{M}$ is the **backward-time** migration rate matrix, and
- $C$ are the ancestral locations on the tree.

# SC Inference: Modified tree prior

Again, 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,
- $\bar{M}$ 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**.

# SC Inference: Modified tree prior

Again, 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,
- $\bar{M}$ 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 will *not* bias results!

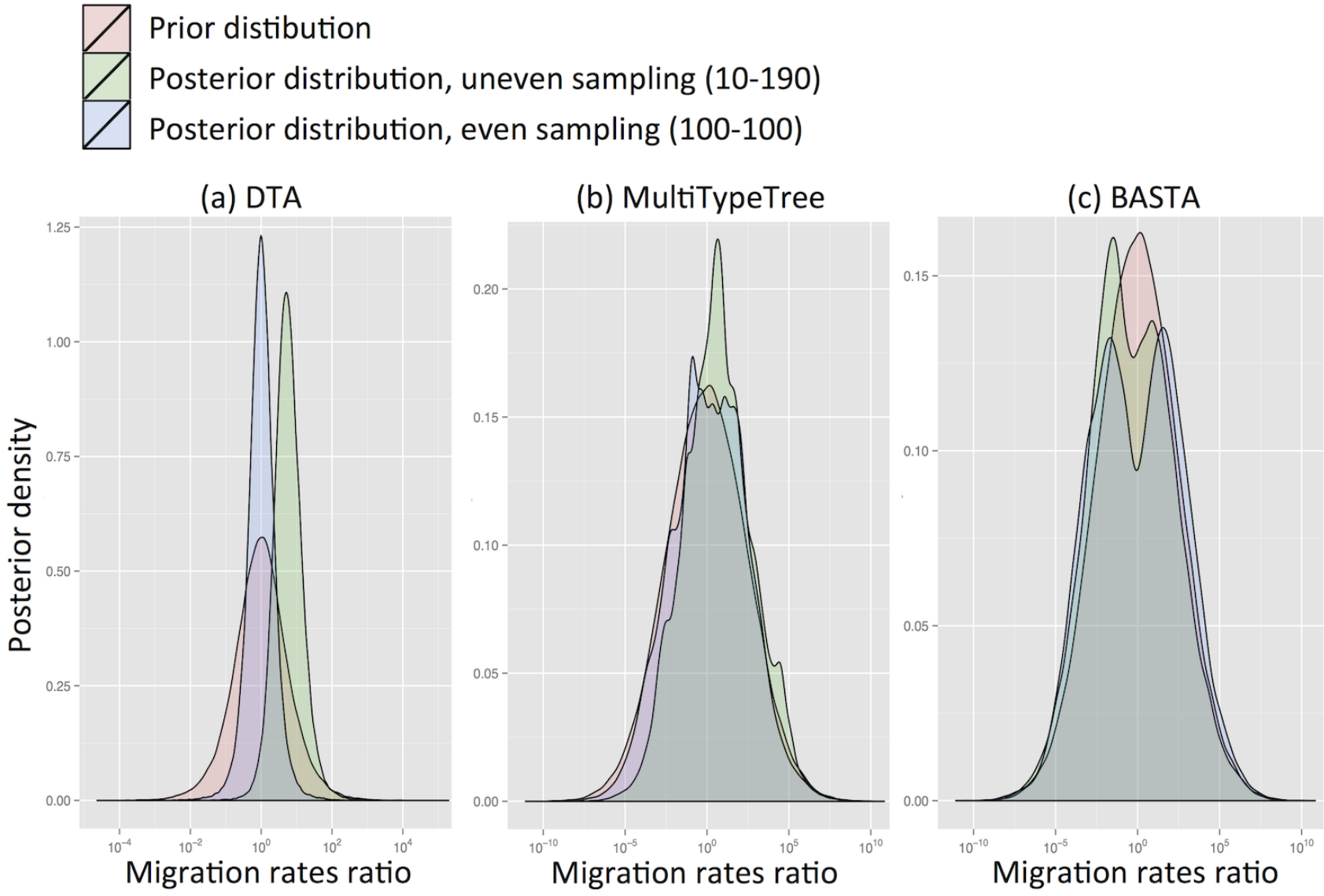
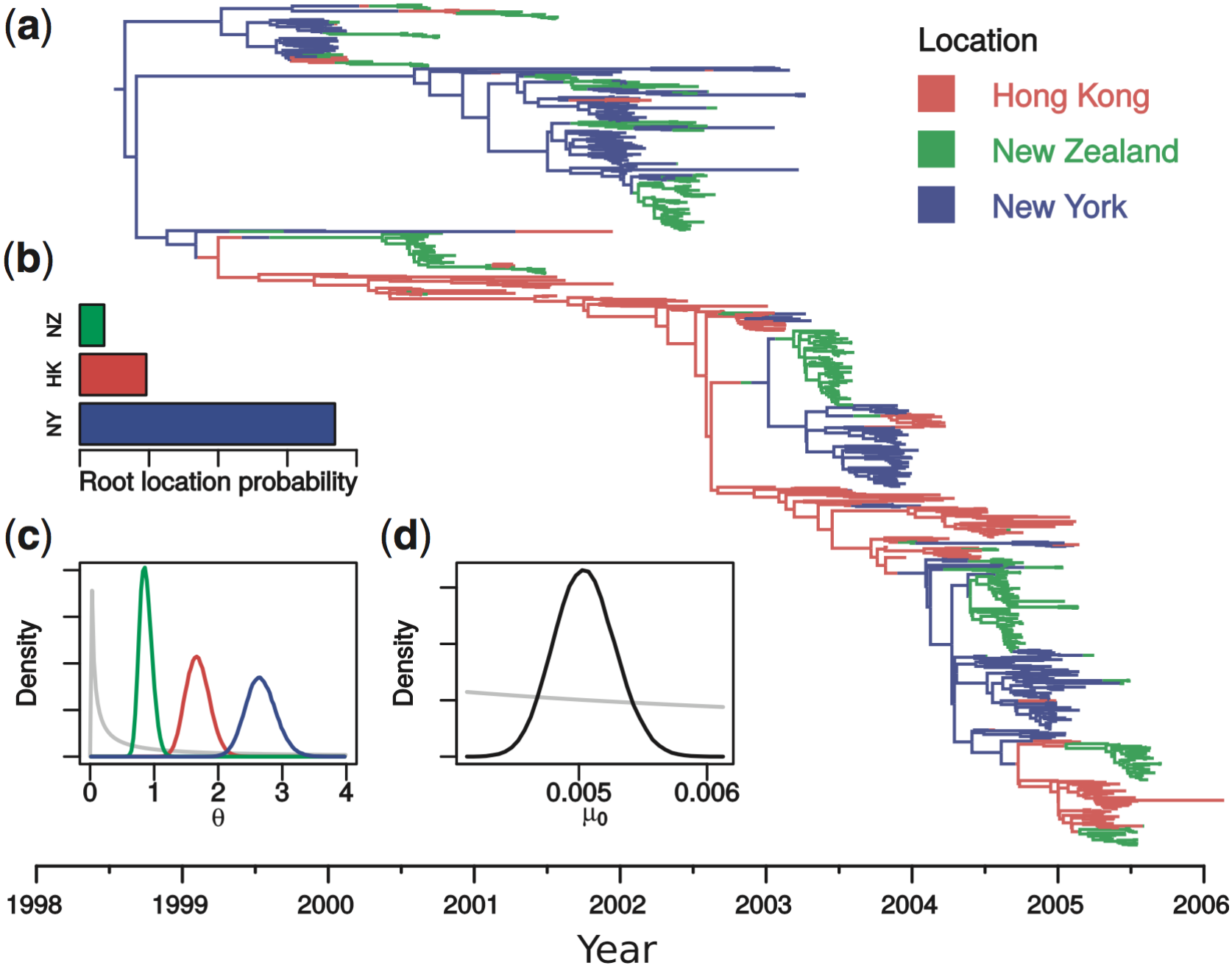



Figure 2, De Maio et al. (2015)




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

# Part III

## Structured birth-death-sampling models

# Structured birth-death-sampling models

- Introduced by Kühnert et al. (2016).
- A birth-death model of population dynamics in which individuals are permitted to change location due to discrete migration events.
- Sampling process is explicitly modelled.
- Birth and death rates may be location-dependent: not "neutral"! (Tree shape affected by structure.)
- Inference is performed using modified tree prior.

# Examples: two-type structured birth-death models

# Examples: two-type structured birth-death models

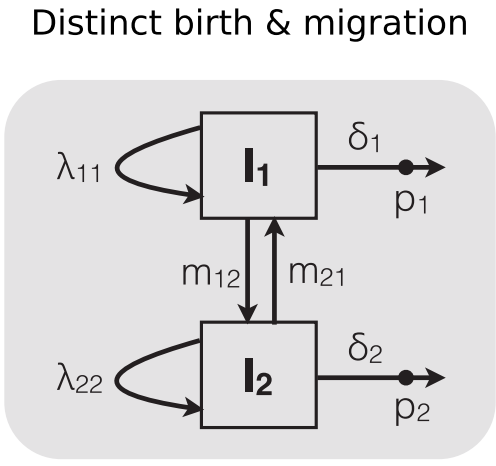
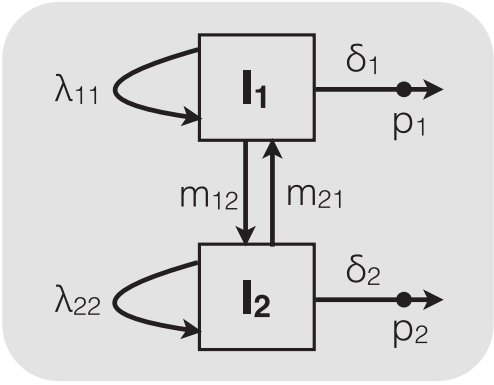



Figure 2, Kühnert et al. (2016)

# Examples: two-type structured birth-death models

Distinct birth & migration




Simultaneous birth & migration

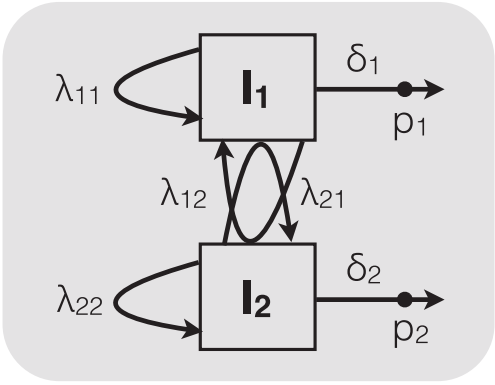
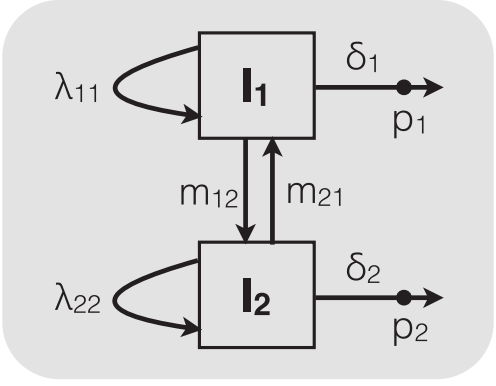


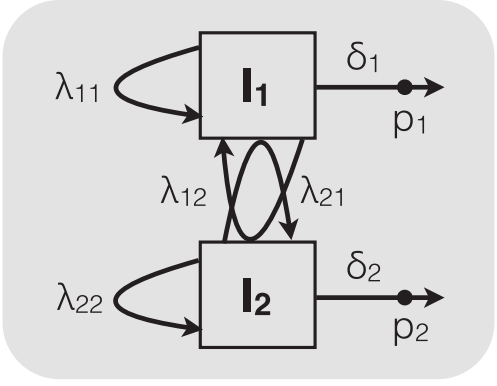

Figure 2, Kühnert et al. (2016)

# Examples: two-type structured birth-death models

Distinct birth & migration




Simultaneous birth & migration




Hybrid model

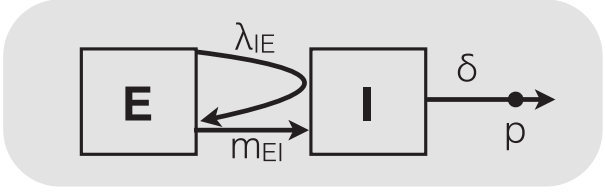
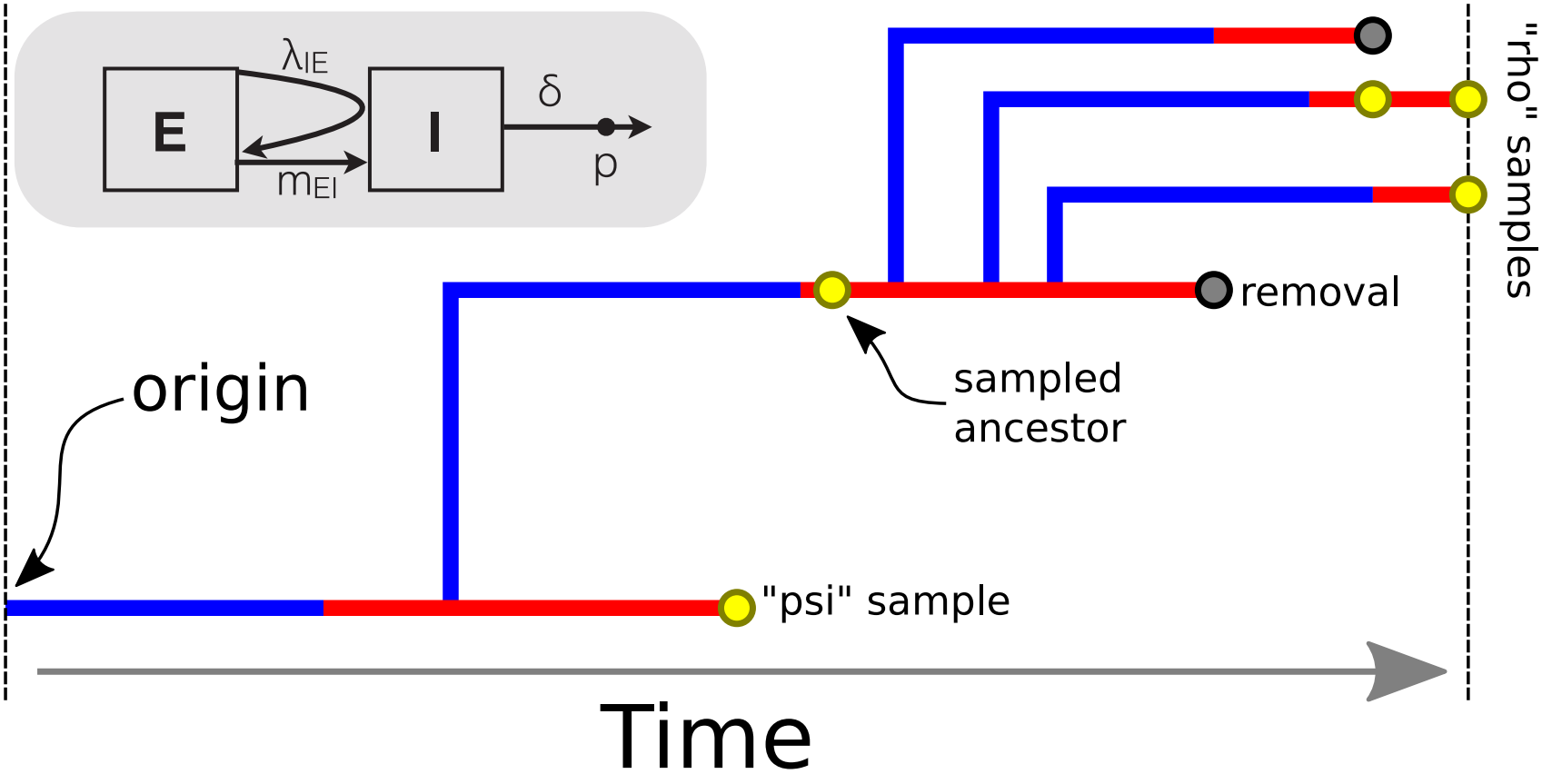



Figure 2, Kühnert et al. (2016)

# Examples: two-type structured birth-death models



# BDMM Inference: Modified tree prior

The standard phylogenetic posterior is once more modified:

$$\begin{aligned} P(T, [C], \mu, \vec{\lambda}, M, \vec{\gamma}, \vec{\psi}, \vec{\rho}, r, t_0 | A, L) = & \frac{1}{P(A|L)} P(A|T, \mu) \\ & \times P(T, [C] | \vec{\lambda}, M, \vec{\gamma}, \vec{\psi}, \vec{\rho}, r, t_0) \\ & \times P(\mu) P(\vec{\lambda}, M, \vec{\gamma}, \vec{\psi}, \vec{\rho}, r, t_0) \end{aligned}$$

where

| | |
|---|---|
| $L$ | are the sampled locations, |
| $\vec{\lambda}, \vec{\gamma}$ | are the type-specific birth and death rates, |
| $M$ | is the type-change transition rate matrix, |
| $\vec{\psi}, \vec{\rho}$ | are the type-specific $\psi$ and $\rho$-sampling rates, |
| $r$ | is the removal probability on sampling, and |
| $t_0$ | is the age of the start of the forward process. |

# BDMM Inference: Modified tree prior

The standard phylogenetic posterior is once more modified:

$$P(T, [C], \mu, \vec{\lambda}, M, \vec{\gamma}, \vec{\psi}, \vec{\rho}, r, t_0 | A, L) = \frac{1}{P(A|L)} P(A|T, \mu) \times P(T, [C]|\vec{\lambda}, M, \vec{\gamma}, \vec{\psi}, \vec{\rho}, r, t_0) \times P(\mu) P(\vec{\lambda}, M, \vec{\gamma}, \vec{\psi}, \vec{\rho}, r, t_0)$$

where

- $L$ are the sampled locations,
- $\vec{\lambda}, \vec{\gamma}$ are the type-specific birth and death rates,
- $M$ is the type-change transition rate matrix,
- $\vec{\psi}, \vec{\rho}$ are the type-specific $\psi$ and $\rho$-sampling rates,
- $r$ is the removal probability on sampling, and
- $t_0$ is the age of the start of the forward process.

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

# Sampling assumption

- As with the birth-death-sampling process (Stadler, 2010), these models explicitly model the sampling process.
- Both temporal and spatial aspects of the sampling process are modeled.

# Sampling assumption

- As with the birth-death-sampling process (Stadler, 2010), these models explicitly model the sampling process.
- Both temporal and spatial aspects of the sampling process are modeled.

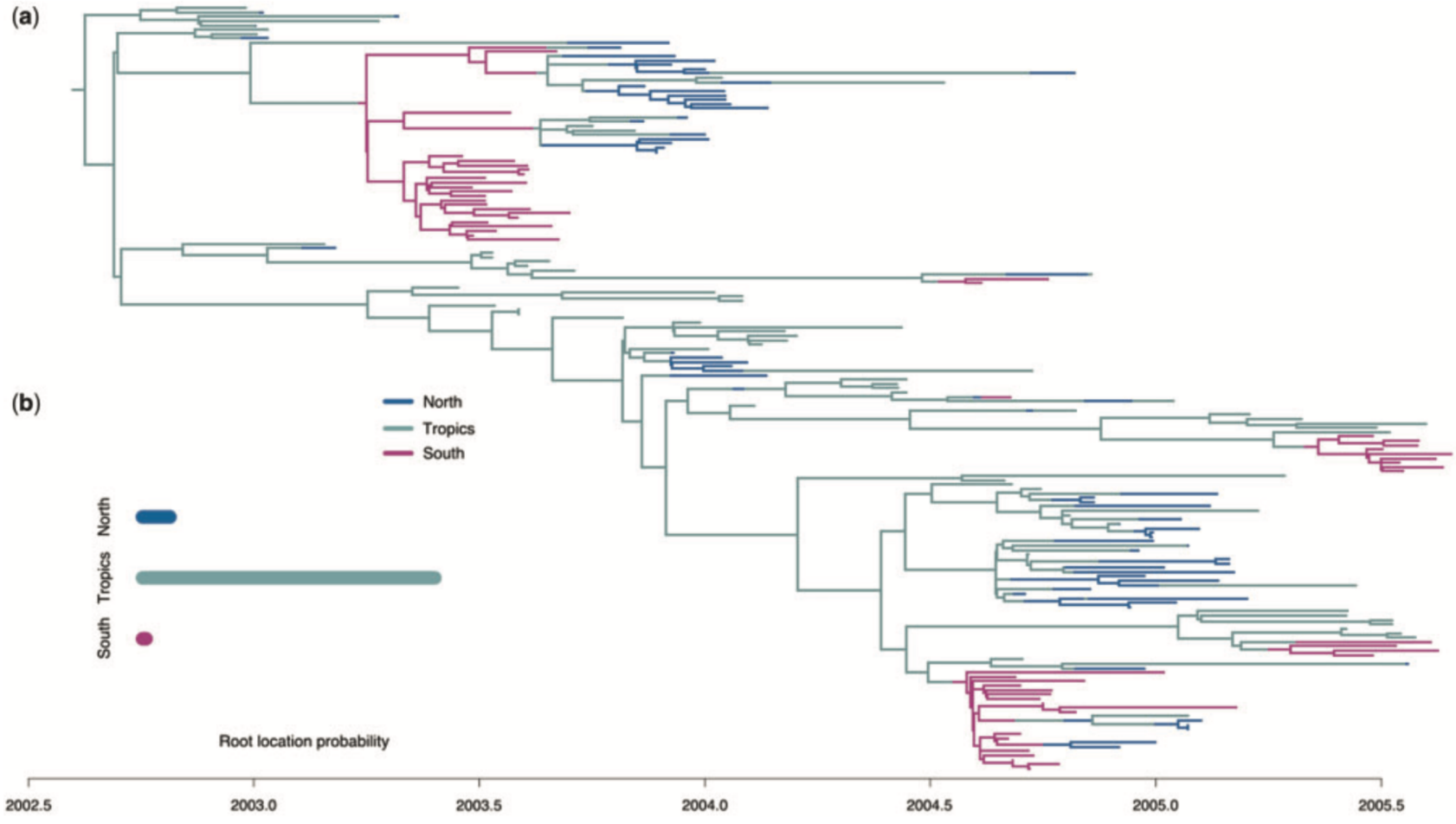
The acquisition of samples from the metapopulation may be non-uniform, assuming the type-specific sampling rates are correspondingly non-uniform.

# Sampling assumption

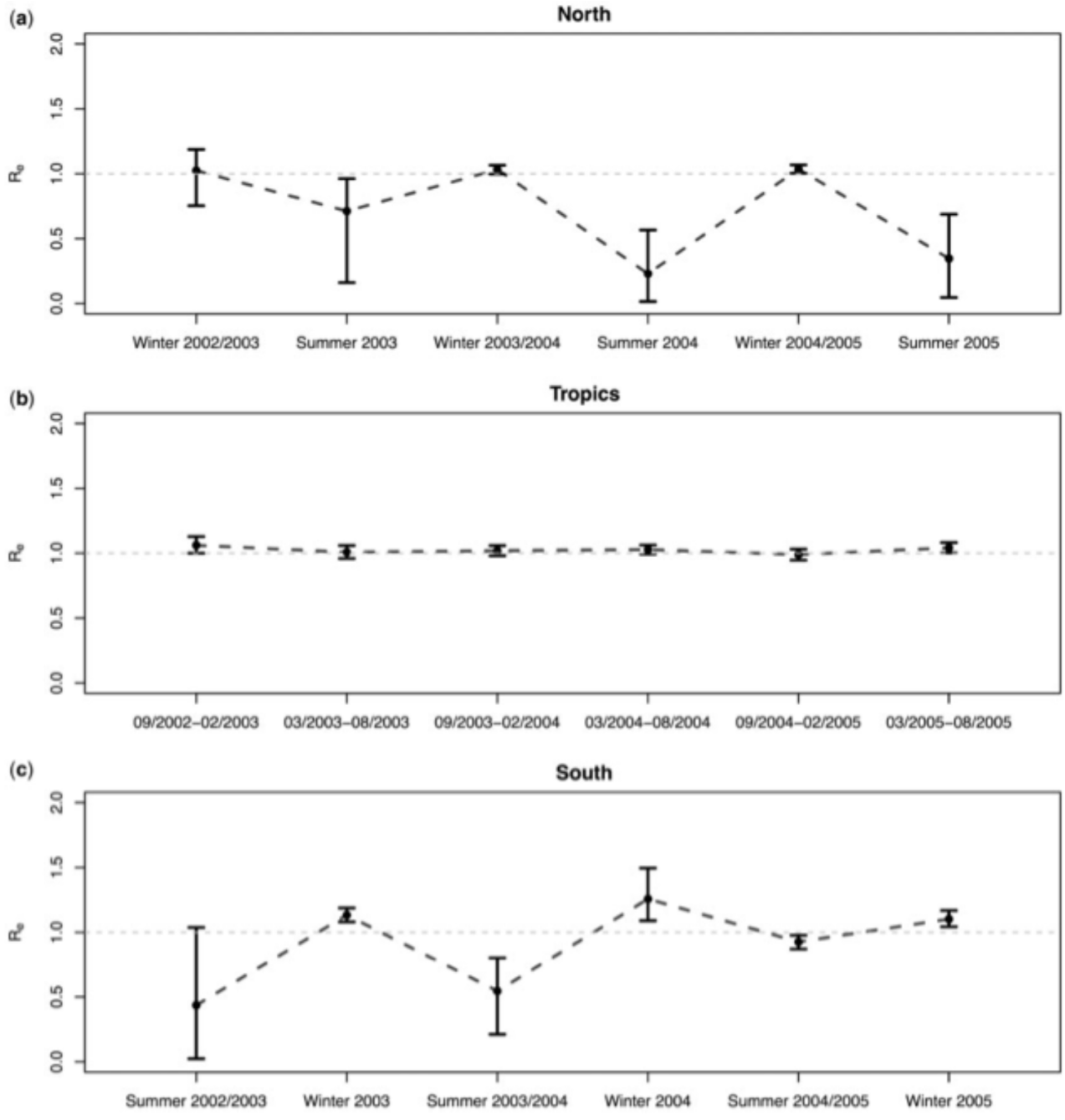
- As with the birth-death-sampling process (Stadler, 2010), these models explicitly model the sampling process.
- Both temporal and spatial aspects of the sampling process are modeled.

The acquisition of samples from the metapopulation may be non-uniform, assuming the type-specific sampling rates are correspondingly non-uniform.

- Jointly estimating the spatial distribution of sampling rates allows non-uniform sampling to be appropriately handled.
- In contrast with the vanilla SC model, modelling the sampling process can provide additional inference power.




Analysis of H3N2 under BDMM, Kühnert et al. (2016).




Analysis of H3N2 BDMM, Kühnert et al. (2016).

# Part IV

# Conclusions

- Spatially labelled sequence data can be used to draw inferences about
  - the specific geographical ancestry of the samples,
  - the spatial dynamics of the underlying population (when this is a component of the model.

- Spatially labelled sequence data can be used to draw inferences about
  - the specific geographical ancestry of the samples,
  - the spatial dynamics of the underlying population (when this is a component of the model.
- Population structure due to weak gene flow between different locations (or other compartments) can directly affect the shape of the tree.

- Spatially labelled sequence data can be used to draw inferences about
  - the specific geographical ancestry of the samples,
  - the spatial dynamics of the underlying population (when this is a component of the model.
- Population structure due to weak gene flow between different locations (or other compartments) can directly affect the shape of the tree.
- Models such as the structured coalescent and the birth-death-sampling with migration (BDMM) models explicitly capture this effect.

- Spatially labelled sequence data can be used to draw inferences about
  - the specific geographical ancestry of the samples,
  - the spatial dynamics of the underlying population (when this is a component of the model.
- Population structure due to weak gene flow between different locations (or other compartments) can directly affect the shape of the tree.
- Models such as the structured coalescent and the birth-death-sampling with migration (BDMM) models explicitly capture this effect.
- All models discussed make specific assumptions about the sampling process. Be careful that your data satisfy the assumptions of the model you use.

# References


Avise, J. C. (2000). *Phylogeography: the history and formation of species*. Harvard university press.

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.

Kühnert, D., Stadler, T., Vaughan, T. G., and Drummond, A. J. (2016). Phylodynamics with migration: A computational framework to quantify population structure from genomic data. *Mol Biol Evol*, 33:2102–2116.

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.

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.