Using the Structured Coalescent for Phylogeography in BEAST 2

Lecturer: Tim Vaughan

Centre for Computational Evolution
Department of Computer Science, University of Auckland

February 2017

Taming the BEAST in the South Pacific

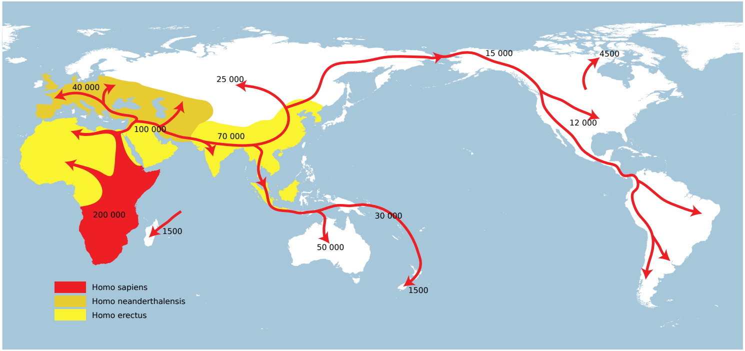
What is Phylogeography?

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:



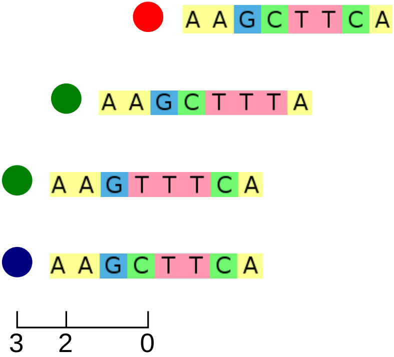
[Wikipedia]

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

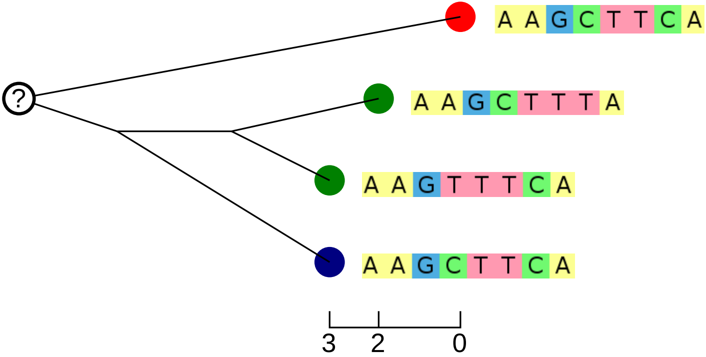
Phylogeographic inference questions

Common questions include:



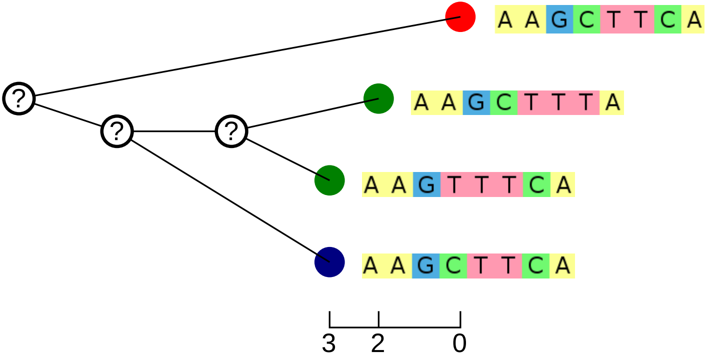
Phylogeographic inference questions

Common questions include:



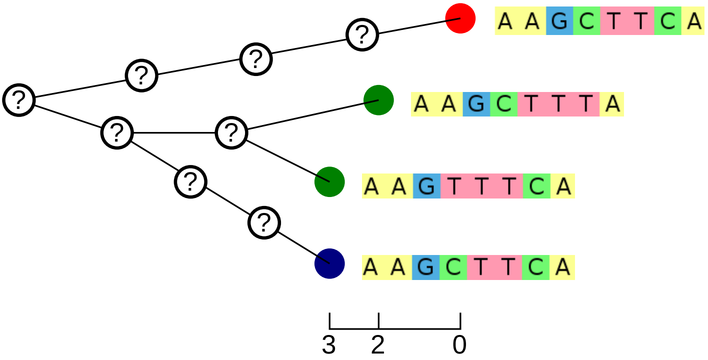
Phylogeographic inference questions

Common questions include:



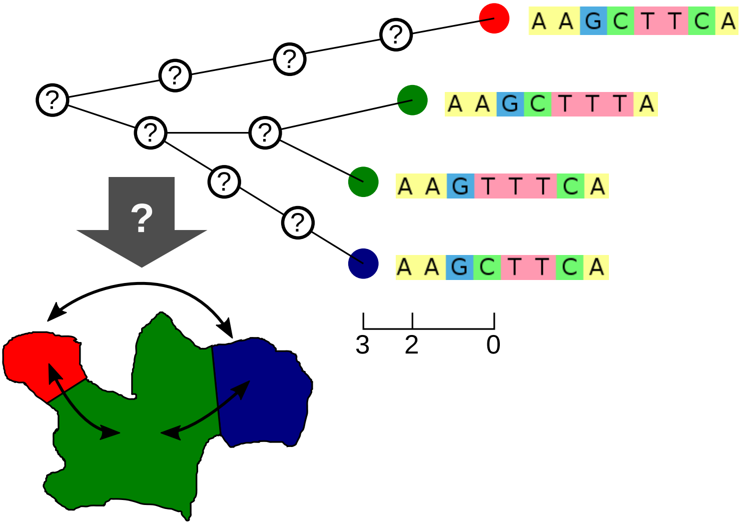
Phylogeographic inference questions

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?

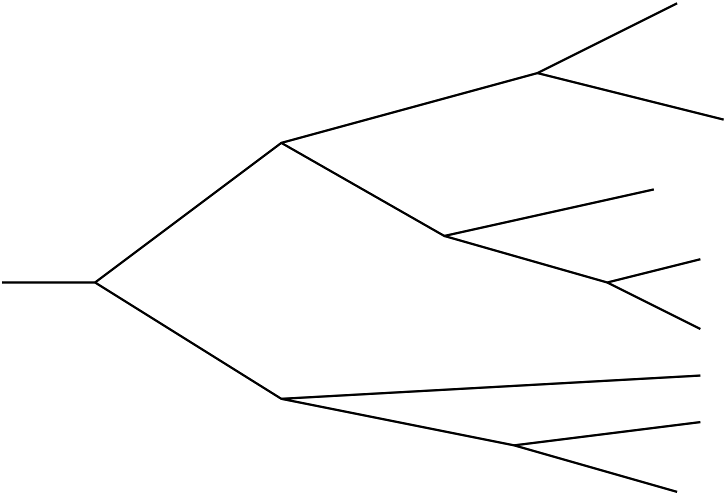
Models for Phylogeographic inference

Currently there are two main classes of models:

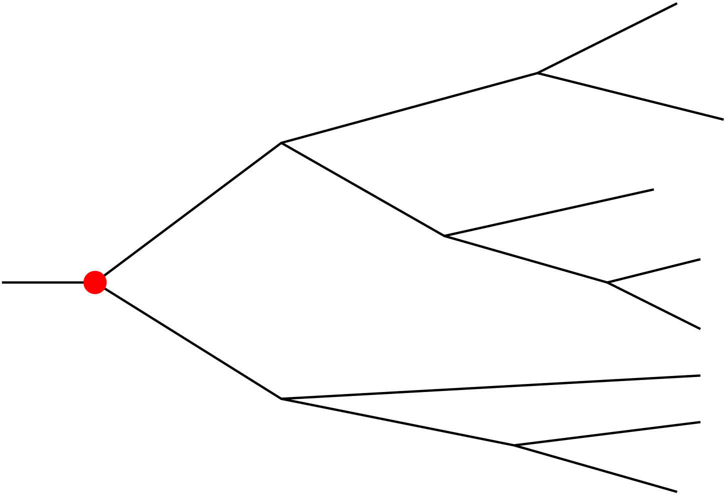
- **Mugration models:**
  - Given tree and root location, what is the probability of sample locations?
  - Exist in continuous and discrete forms.
  - Developed by Phillipe Lemey et al. [Lemey et al., 2009, Lemey et al., 2010].
- **Structured population models:**
  - Given sequences and locations, what is the probability of the location-coloured tree?
  - Currently mostly discrete.
  - Eariest examples by [Hudson, 1990] and [Notohara, 1990].

Discrete mugration model

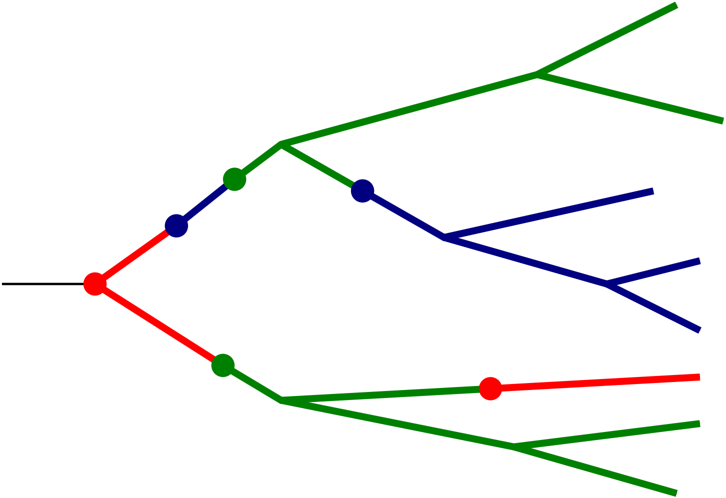
Discrete mugration model



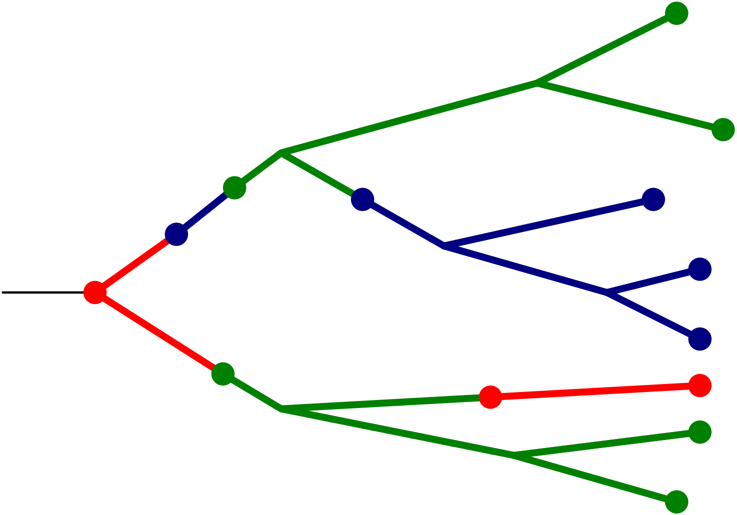
Discrete mugration model



Discrete mugration model



Discrete mugration model



Recap: Bayesian Phylogenetic 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

A is a sequence alignment,

T is the tree.

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:

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

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.

Sampling assumption

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

Sampling assumption

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

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

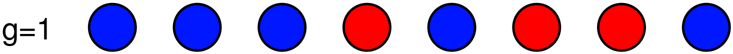
- Mugration models use sample location as data.
- Just as for genetic data, non-random sampling procedures will **bias results**.

Equivalent population genetic model

A helpful way to visualise the mugration model is to imagine its effect on the population as a whole:

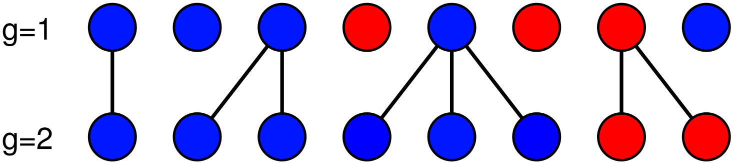
Equivalent population genetic model

A helpful way to visualise the mugration model is to imagine its effect on the population as a whole:



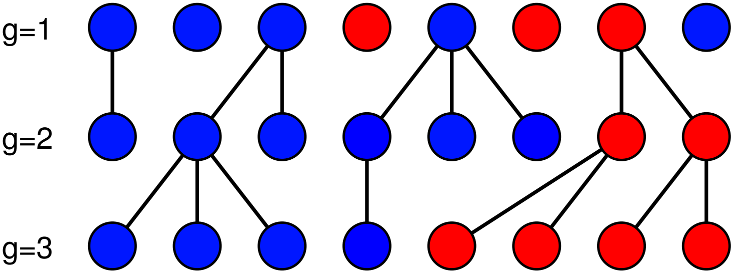
Equivalent population genetic model

A helpful way to visualise the mugration model is to imagine its effect on the population as a whole:



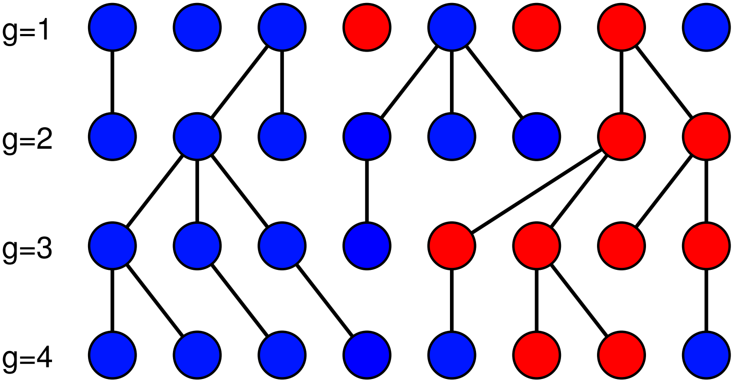
Equivalent population genetic model

A helpful way to visualise the mugration model is to imagine its effect on the population as a whole:



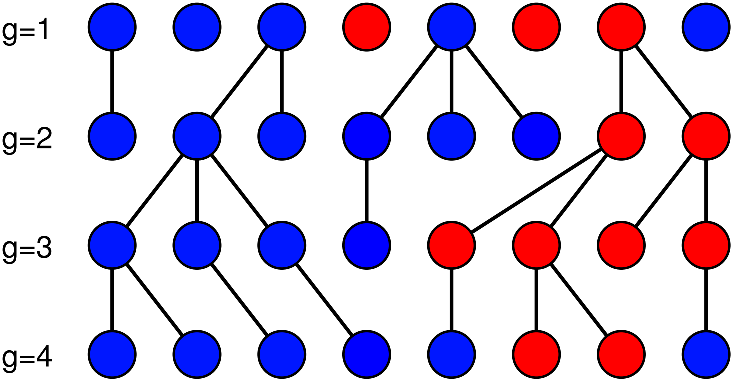
Equivalent population genetic model

A helpful way to visualise the mugration model is to imagine its effect on the population as a whole:



Equivalent population genetic model

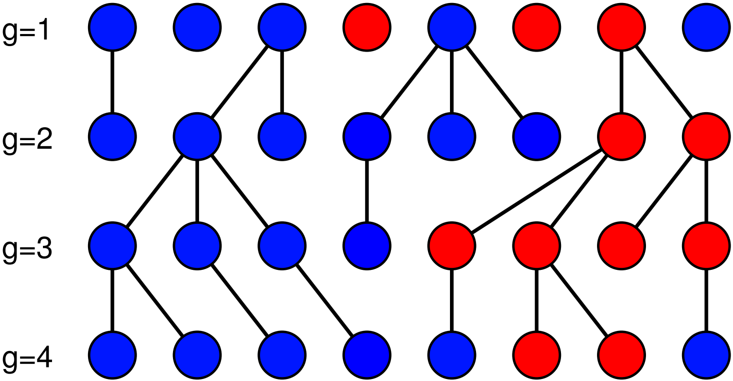
A helpful way to visualise the mugration model is to imagine its effect on the population as a whole:



- Mugration $\Longrightarrow$ stochastically varying population sizes.

Equivalent population genetic model

A helpful way to visualise the mugration model is to imagine its effect on the population as a whole:



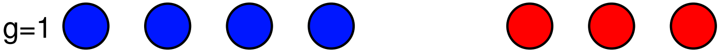
- Mugration $\implies$ stochastically varying population sizes.
- A "neutral" model.

Structured Wright-Fisher model

Imagine two sub-populations connected by weak migration:

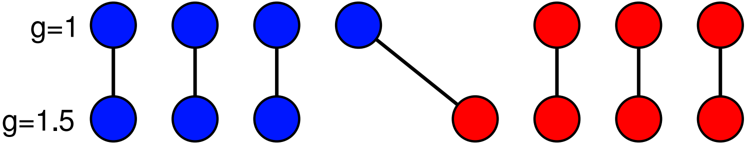
Structured Wright-Fisher model

Imagine two sub-populations connected by weak migration:



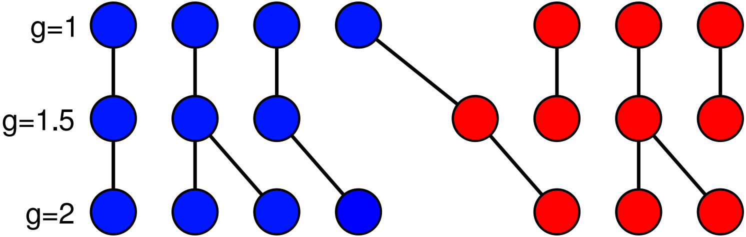
Structured Wright-Fisher model

Imagine two sub-populations connected by weak migration:



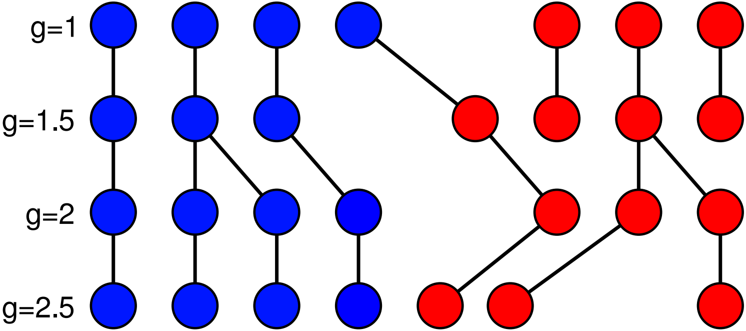
Structured Wright-Fisher model

Imagine two sub-populations connected by weak migration:



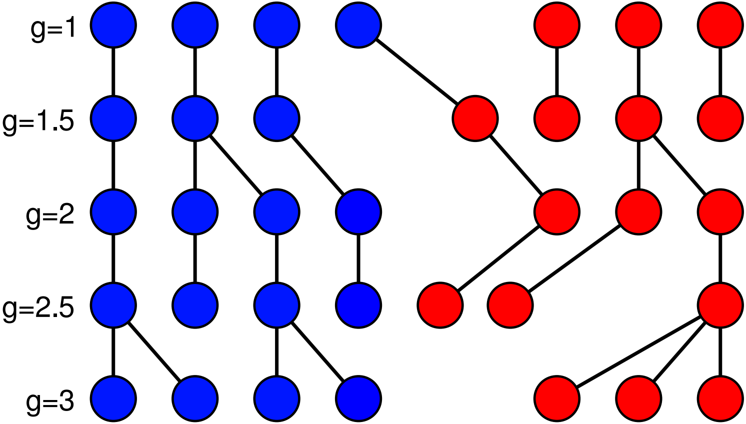
Structured Wright-Fisher model

Imagine two sub-populations connected by weak migration:



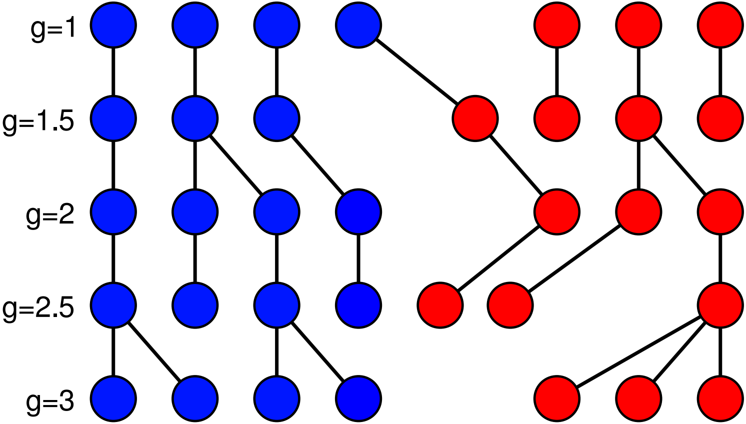
Structured Wright-Fisher model

Imagine two sub-populations connected by weak migration:



Structured Wright-Fisher model

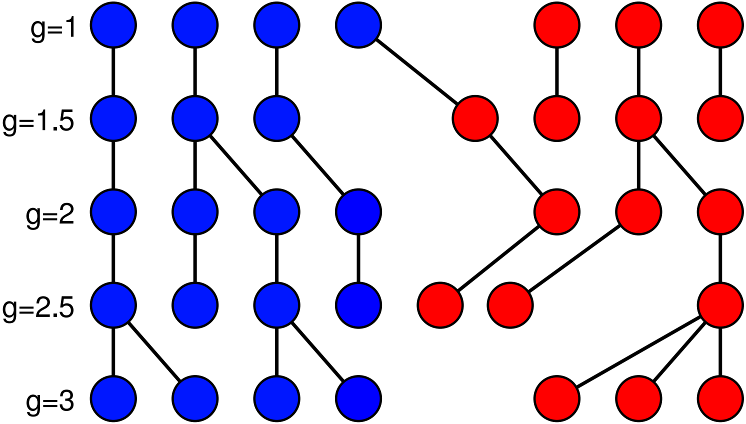
Imagine two sub-populations connected by weak migration:



- Model described by [Notohara, 1990].

Structured Wright-Fisher model

Imagine two sub-populations connected by weak migration:



- Model described by [Notohara, 1990].
- Island populations are held constant by respective carrying capacities.

Structured Coalescent

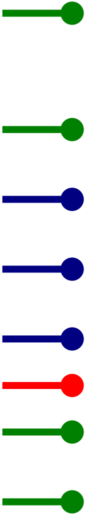
Backwards-in-time process that generates both the tree and ancestral locations.

Structured Coalescent

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

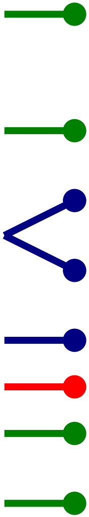
Structured Coalescent

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



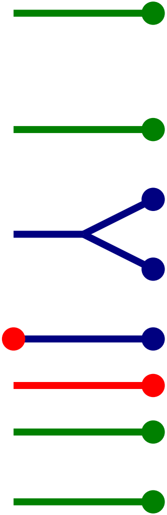
Structured Coalescent

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



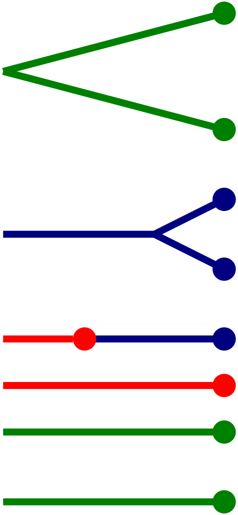
Structured Coalescent

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



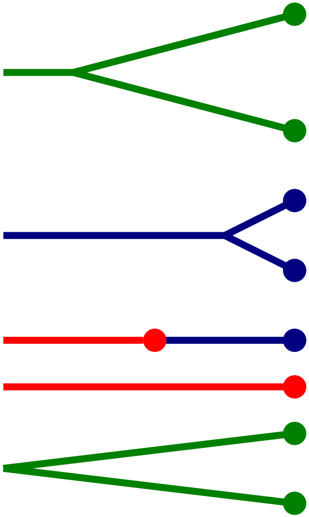
Structured Coalescent

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



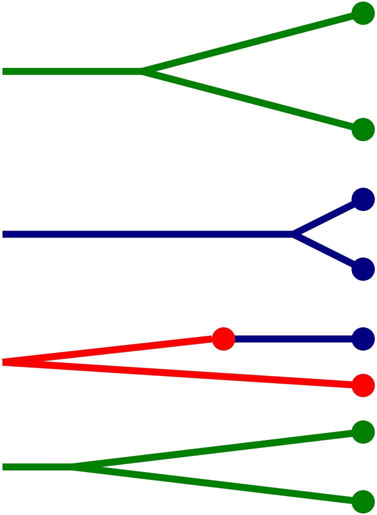
Structured Coalescent

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



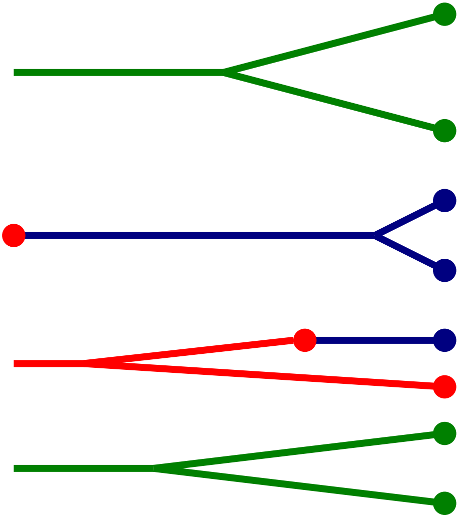
Structured Coalescent

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



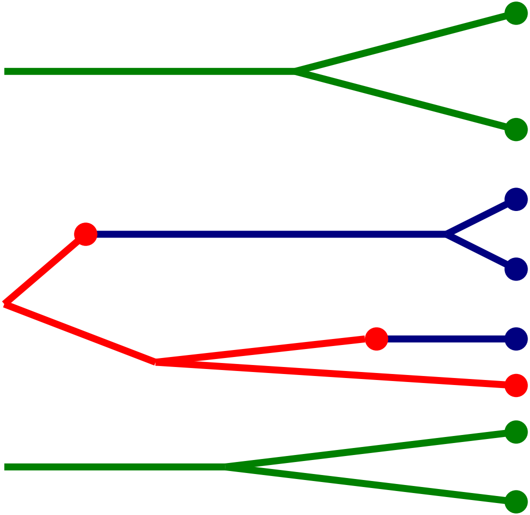
Structured Coalescent

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



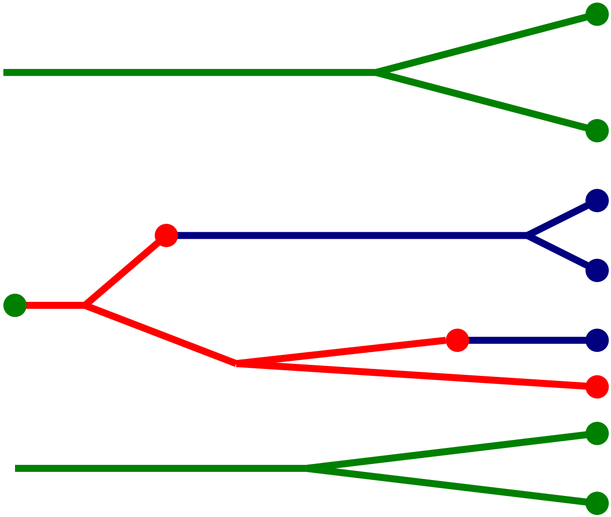
Structured Coalescent

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



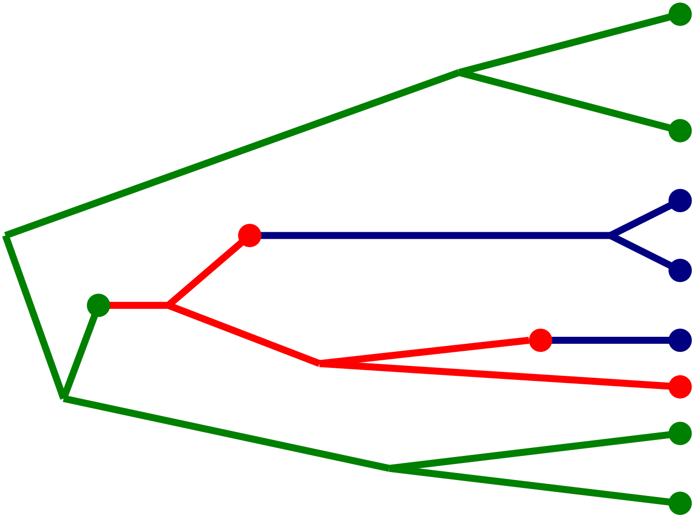
Structured Coalescent

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



Structured Coalescent

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



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 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.

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!

Birth-death migration model

- 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.

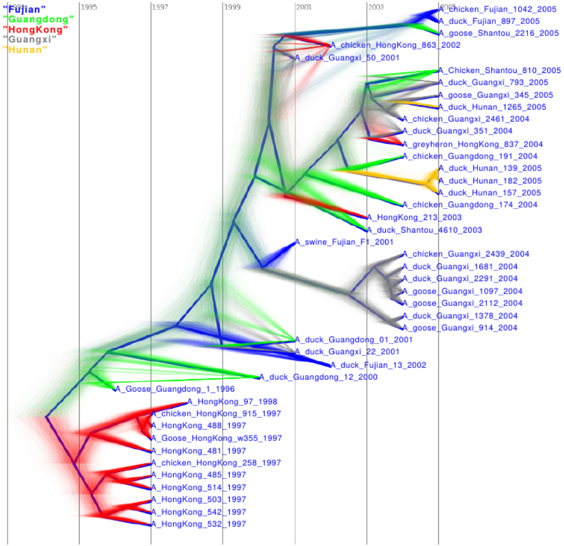
Discrete Phylogeography (Mugration)

Required packages:

- BEAST_CLASSIC

- Very well supported, BEAUti analysis setup.
- Tutorial on beast2.org/tutorials.
- Very fast, allows inference of which migrations are necessary to describe data.
- Prone to sampling biases.

Discrete Phylogeography (Mugration)



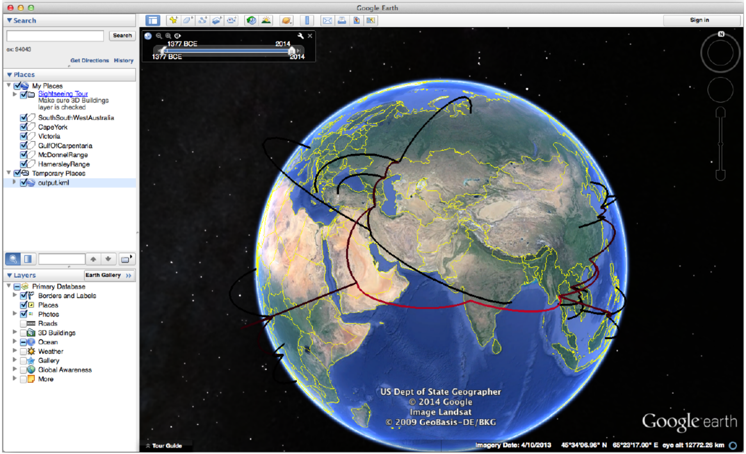
Continuous Phylogeography (Mugration)

Required packages:

- GEO_SPHERE

- Also well supported, BEAUti analysis setup.
- Tutorial on beast2.org/tutorials.
- Output can be summarized using Spread and visualized using Google Earth.
- Prone to sampling biases.

Continuous Phylogeography (Mugration)



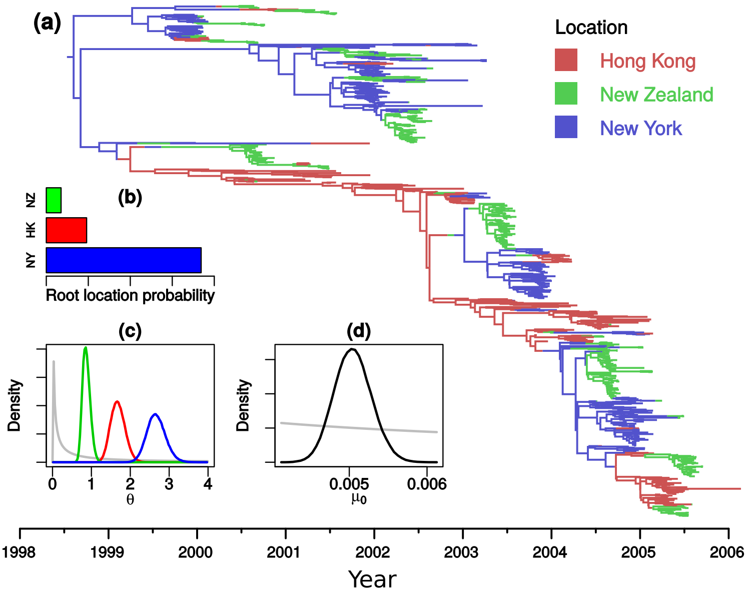
Structured Coalescent (full model)

Required packages:

- MultiTypeTree

- Newer analysis option, BEAUti setup.
- Tutorial at beast2.org/tutorials.
- No built-in assumptions regarding sampling procedure.
- More computationally demanding than mugration models, only smaller numbers of demes are feasible.

Structured Coalescent (full model)



[Vaughan et al., 2014]

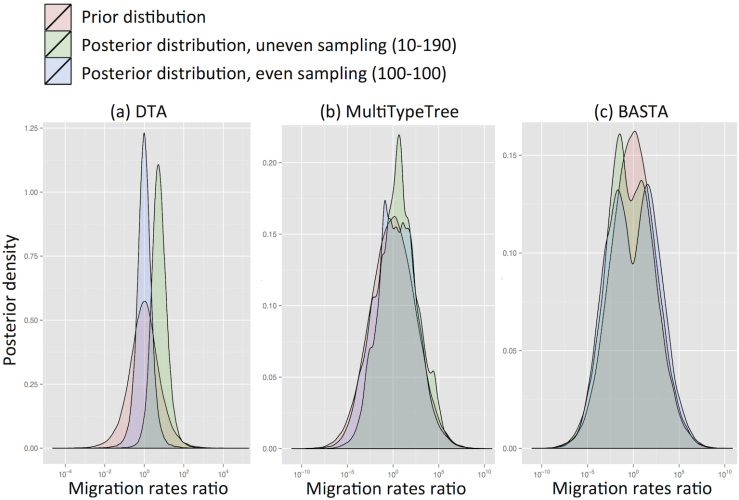
Structured Coalescent (approximation)

Required packages:

- BASTA

- Very new analysis option, **no BEAUti setup**.
- Temporary tutorial at github.com/tgvaughan/MultiTypeTree/wiki
- Approximation cuts down on the computational demands of the SC model, allowing many more locations to be considered.
- Produces similar results to MultiTypeTree.

Structured Coalescent (approximation)



[De Maio et al., 2015]

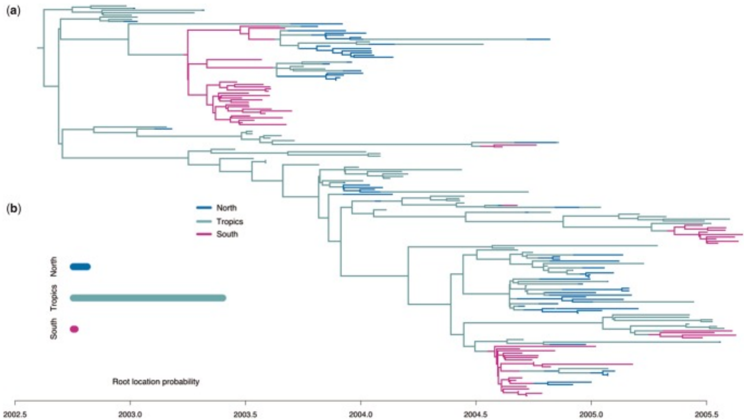
Birth-death Migration Model

Required packages:

- bdmm
- MultiTypeTree
- SA
- MASTER

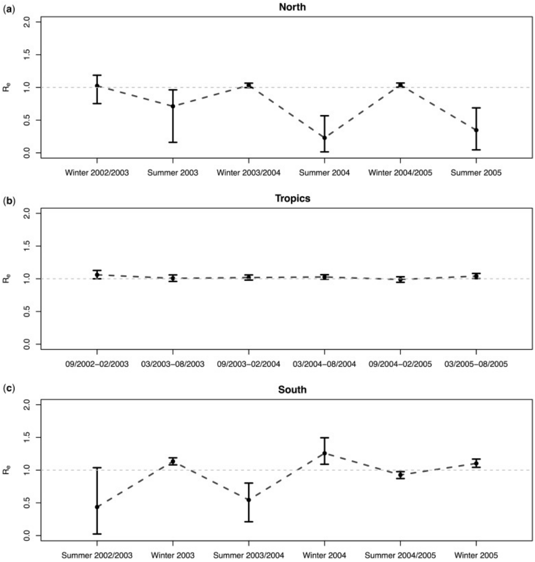
- Very new analysis option, BEAUti setup possible.
- Information can be found on the GitHub repository at github.com/denisekuehnert/bdmm.

Birth-death Migration Model



[Kühnert et al., 2016]

Birth-death Migration Model



[Kühnert et al., 2016]

Summary

- Bayesian phylogeographic methods provide a systematic way of combining geographic and genetic data.
- BEAST 2 provides two main routes:
  - Mugration models
  - Structured population models
- Mugration models tend to allow computationally efficient inference, but have questionable foundation and are subject to sampling biases.
- Structured population models may be more closely tied to the biology and don't necessarily depend on the samplling process.

MultiTypeTree Tutorial

1. Open MultiTypeTree tutorial at taming-the-beast.github.io/tutorials/Structured-coalescent
2. Start the tutorial!

References I

- 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.
- 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.