Coalescent theory and the population genetics of molecular evolution

Molecular Epidemiology of Infectious Diseases

Lecture 4

February 7th, 2022
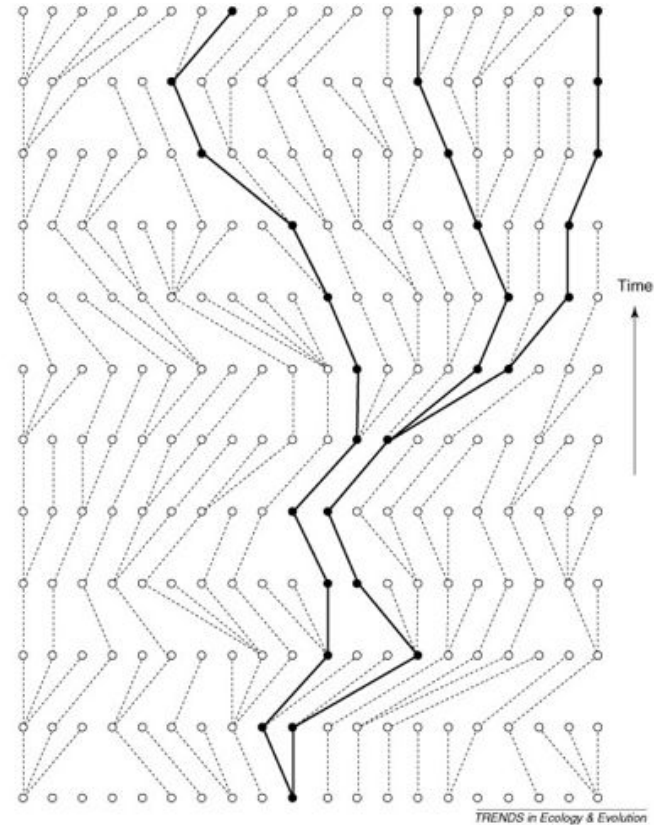
Coalescent models track the ancestry of sampled lineages backwards through time.

John Goddard, The Tree of Man's Life (1649)

Basic coalescent theory

Coalescent theory describes the ancestral relationships (i.e. genealogy) of individuals sampled from a larger population.

The sampled lineages are embedded within the full ancestral history of the population.
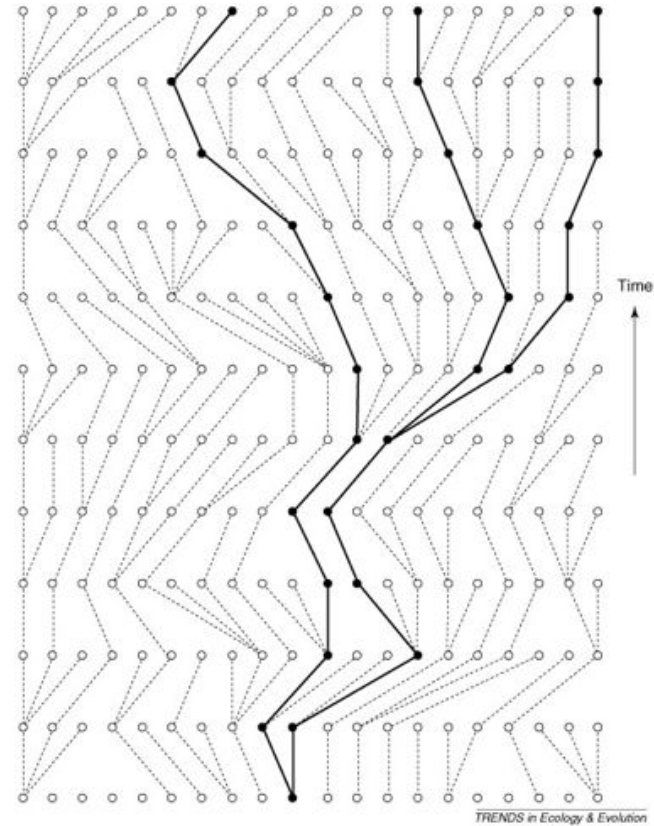

Kuhner *et al.* (2008)

Wright-Fisher assumptions

Coalescent theory is largely based on simple demographic models like the Wright-Fisher model

Constant population size of N haploid individuals

Discrete, non-overlapping generations

Reproduction is a stochastic process such that individuals leave a random number of offspring.

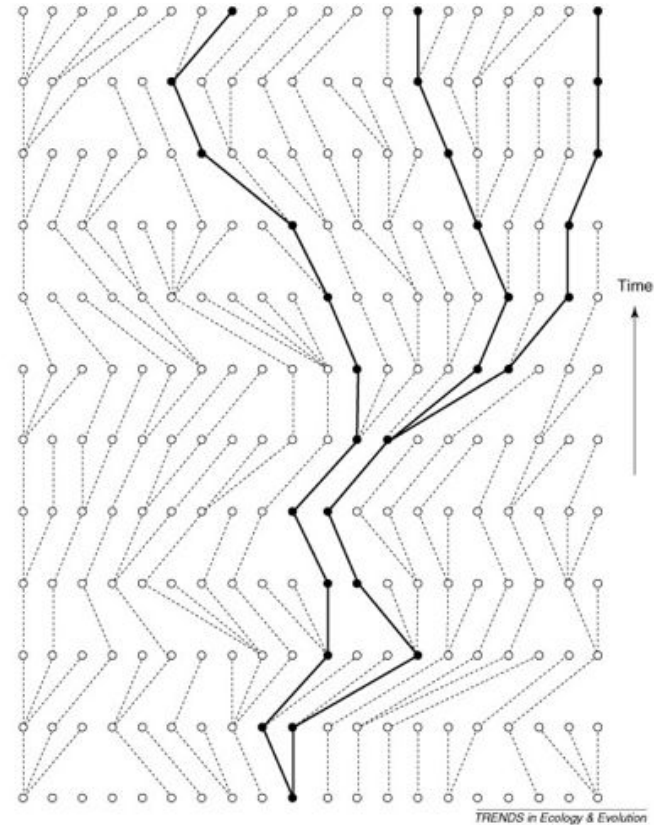Kuhner *et al.* (2008)

Basic coalescent theory

Because reproduction is random, the ancestral relationships among individuals is also viewed as a stochastic process that generates random coalescent trees.

Rosenberg & Nordborg (2002)

Basic coalescent theory

The probability of two lineages coalescing per generation is:

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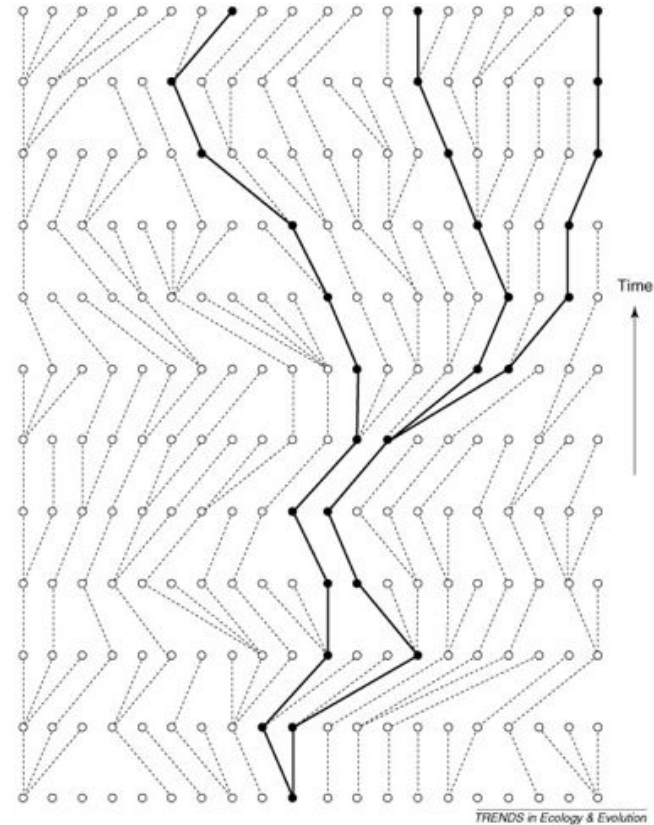

Kuhner *et al.* (2008)

Basic coalescent theory

The probability of two lineages coalescing per generation is:

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

The probability of coalescing after n generations is:

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

Kuhner *et al.* (2008)

Basic coalescent theory

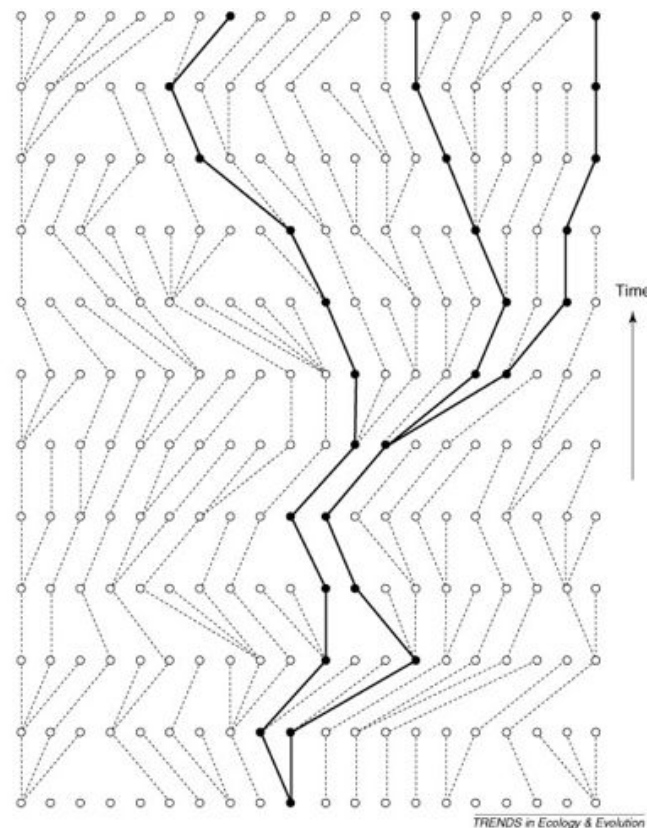The probability of two lineages coalescing per generation is:

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

The probability of coalescing after *n* generations is:

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

In continuous time:

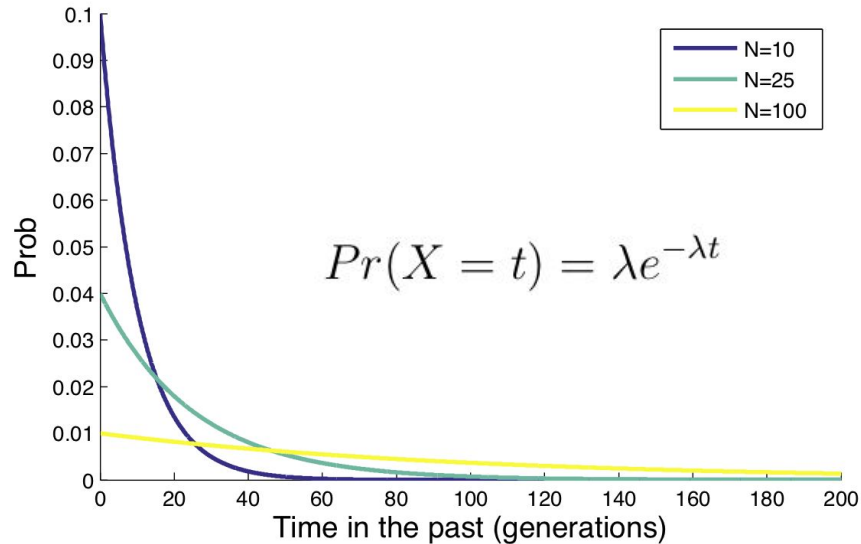$$Pr(X = t) = \lambda e^{-\lambda t} \qquad \lambda = p_{coal} = \frac{1}{N}$$

Time

TRENDS in Ecology & Evolution

Kuhner *et al.* (2008)

Basic coalescent theory

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

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

A slightly more general coalescent model

The coalescent rate increases with the the amount of reproductive variance σ² in the population:
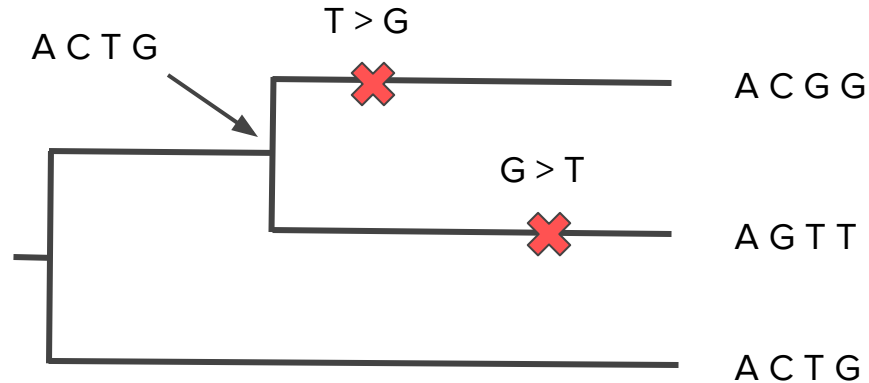
$$\lambda_{coal} = \frac{\sigma^2}{N}$$

We can define an **effective population size N_e**:

$$N_e = \frac{N}{\sigma^2}$$

More generally then, the coalescent rate is:
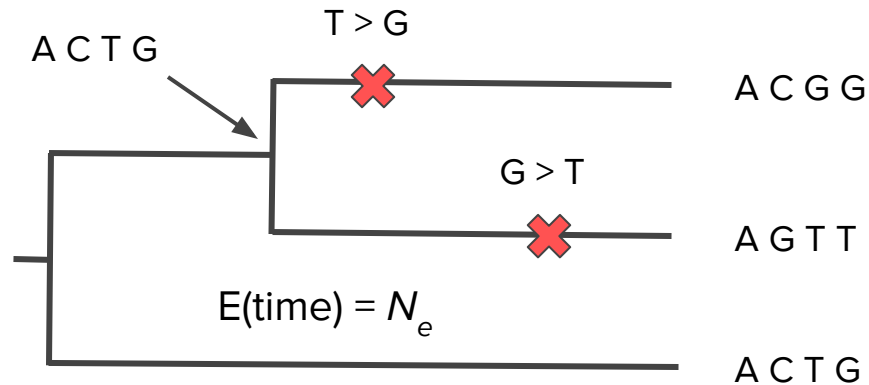
$$\lambda_{coal} = \frac{1}{N_e}$$

Coalescent trees with mutations

Most coalescent models assume mutations are neutral such that mutations occur independently of the coalescent process.

Coalescent trees and genetic diversity

A C T G

T > G

A C G G

G > T

A G T T

E(time) = N_e

A C T G

Genetic diversity depends directly on both the mutation rate μ and the coalescent rate. The expected average pairwise diversity is: $\theta = 2N_e\mu$

Now with more than two lineages

With k lineages present, the coalescent rate becomes:

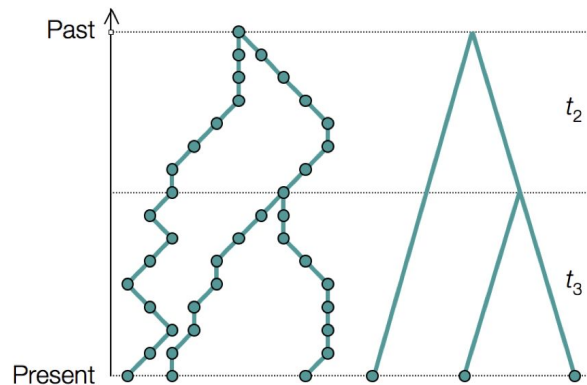$$\lambda_{coal} = \frac{\binom{k}{2}}{N_e}$$

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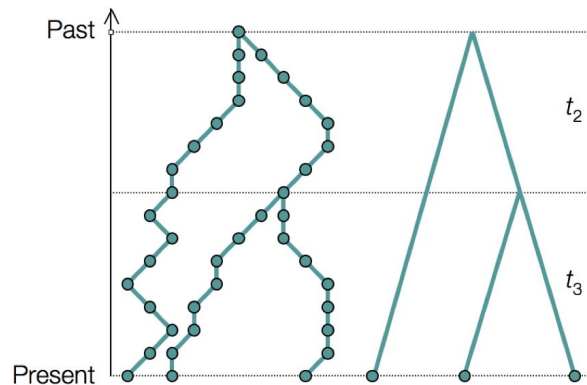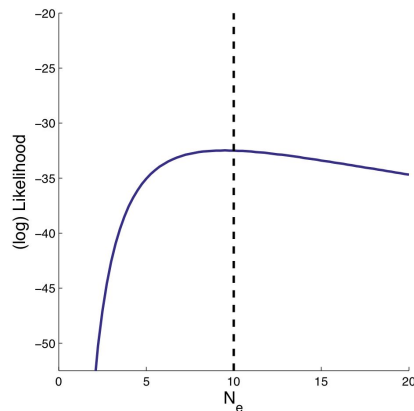
$$L(T|N_e) = \frac{1}{N_e^{(n-1)}} \prod_{k=2}^{n} \exp\left(-\frac{\binom{k}{2}}{N_e} t_k\right)$$
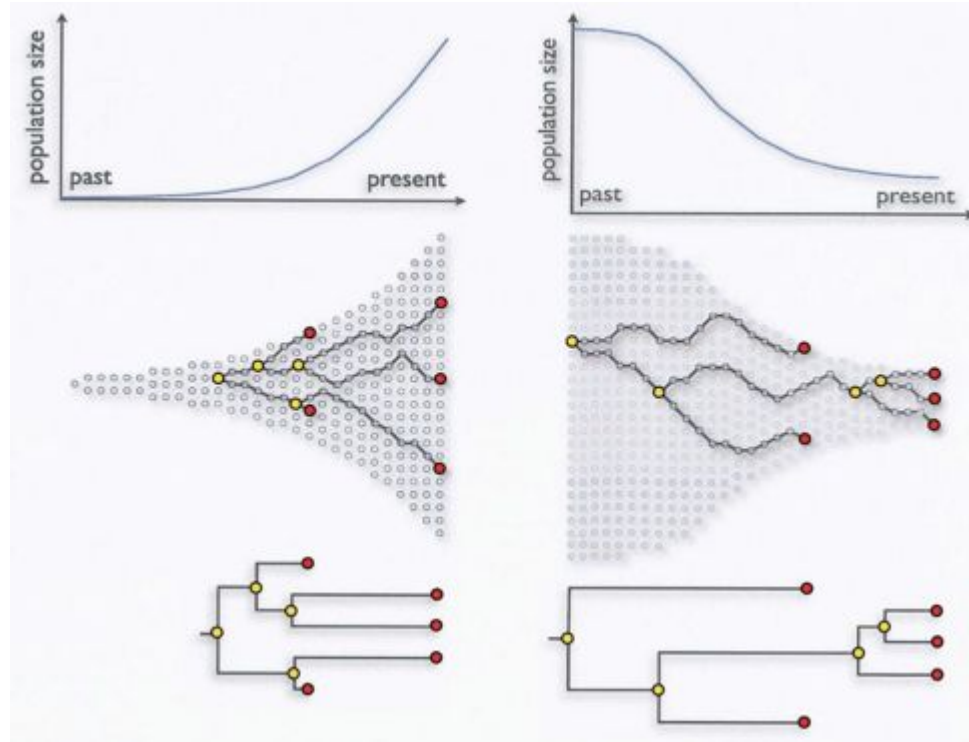
Coalescent-based inference

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

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

We can therefore use coalescent models to infer the demographic history of a population.

The signal of population size change
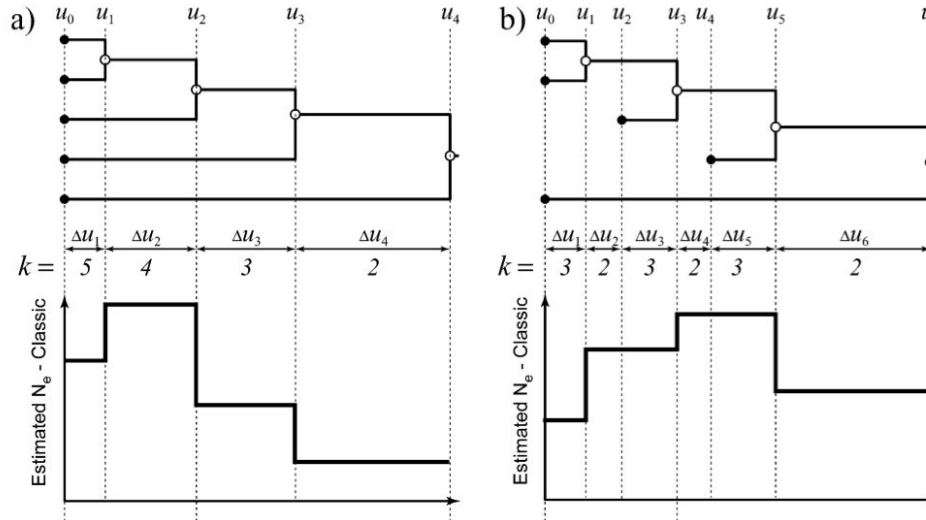
Modeling population size changes

Parametric models assume population sizes change according to some population dynamic model (e.g. exponential growth)

Nonparametric methods allow population sizes to change over time in an unconstrained manner.

- Bayesian Skyline (Drummond *et al.*, 2005)
- Bayesian Skygrid (Minin *et al.*, 2008)

Nonparametric approaches

Generally assume population sizes change over time in a piecewise-constant manner.

Drummond *et al.* (MBE, 2005)

Bayesian skyline

Assumes that the population size can only change at a fixed number of change points.

Uses Bayesian MCMC to integrate (average) over all possible change point positions in addition to N_e within each interval/epoch.

Intervals can contain multiple coalescent events allowing for better estimates of N_e

Produces a smoothed estimate of N_e through time with credible intervals.
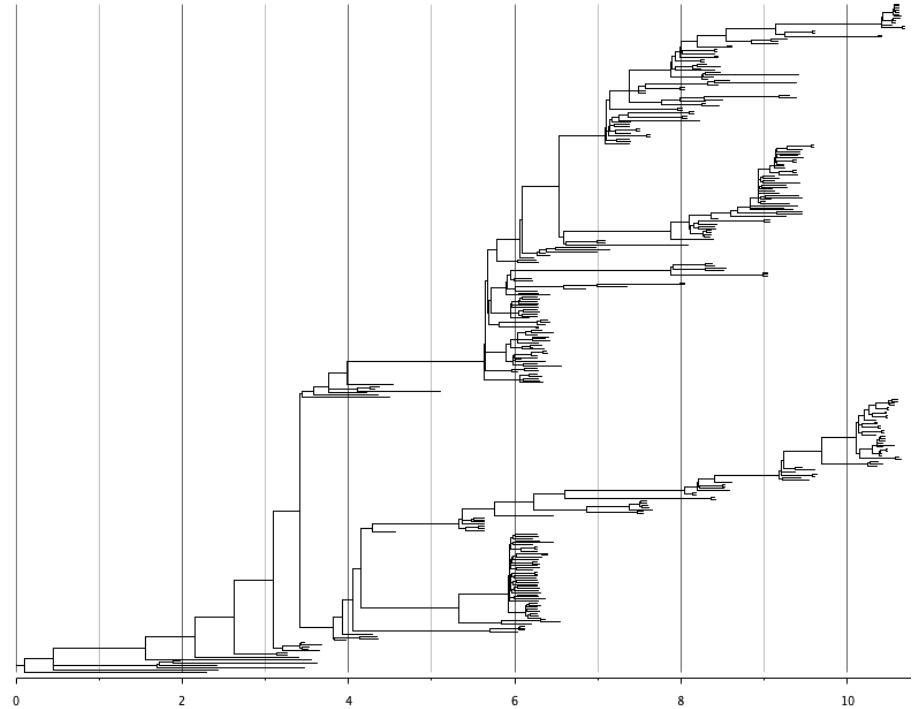
Drummond *et al.* (MBE, 2005)

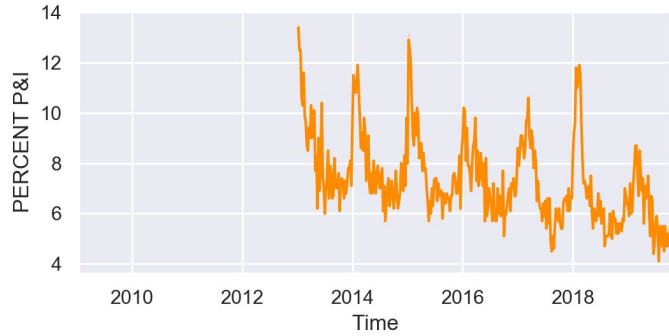# Bayesian skyline plots

Bayesian skyline estimates of for Hepatitis C virus in Egypt
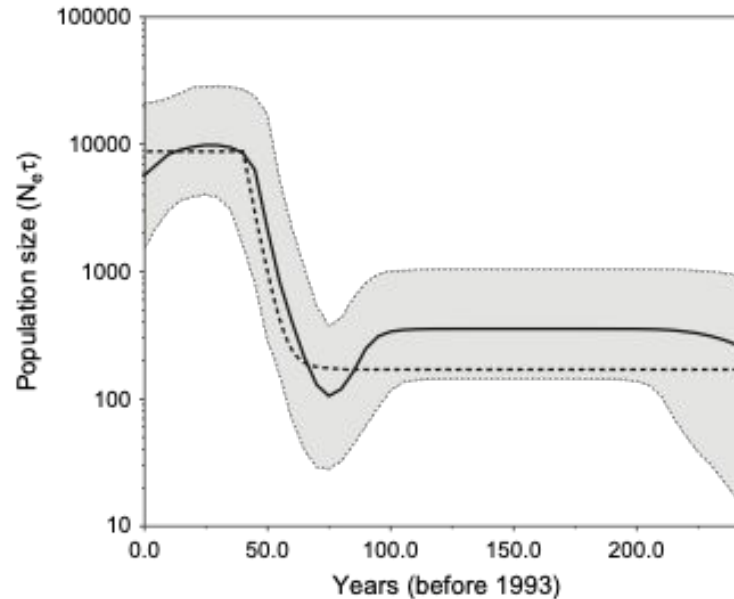
Drummond *et al.* (2005)

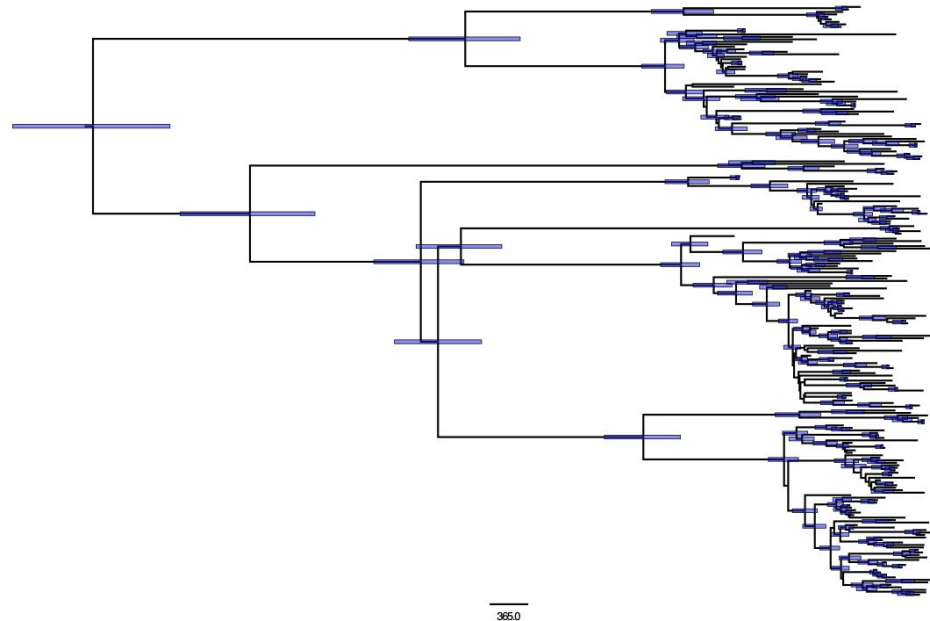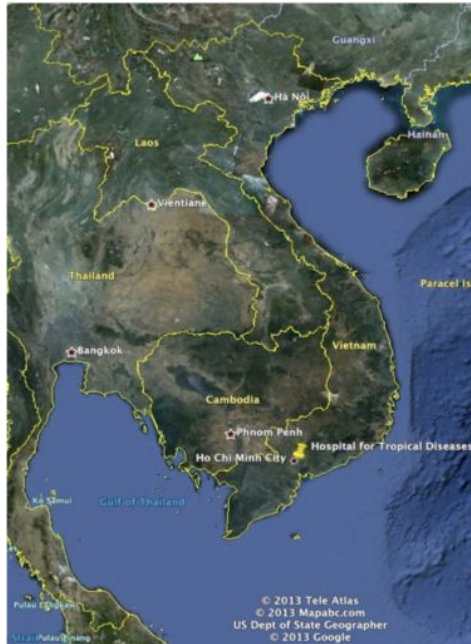# Bayesian skyline plots for the flu in NC

Limitations of simple coalescent models

Other forces beyond population dynamics influence coalescent rates and can strongly shape phylogenies:
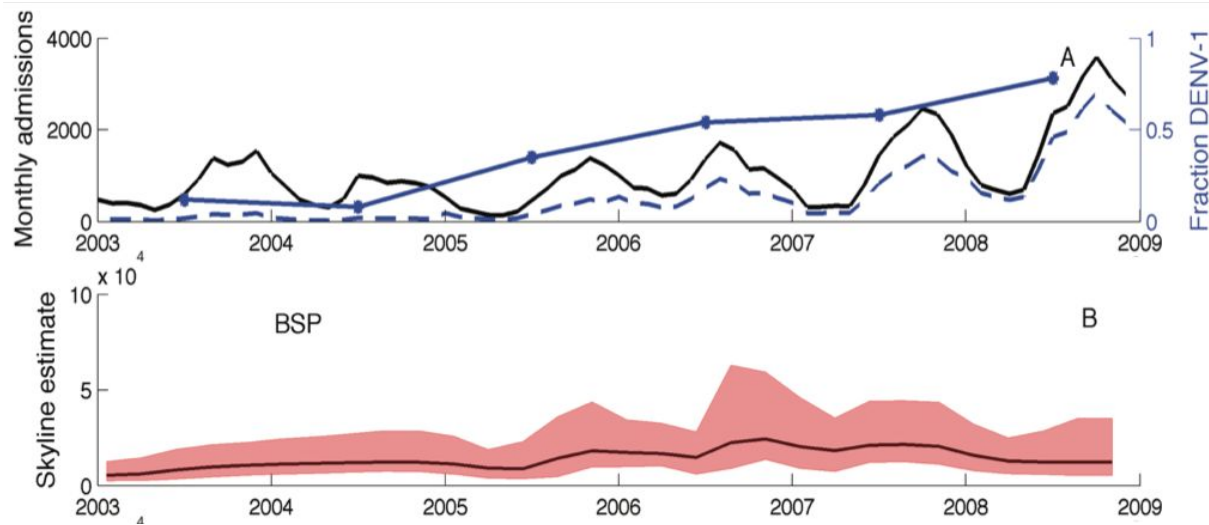
- Reproductive variance

- Population/spatial structure

- Selection

How do we account for population structure?
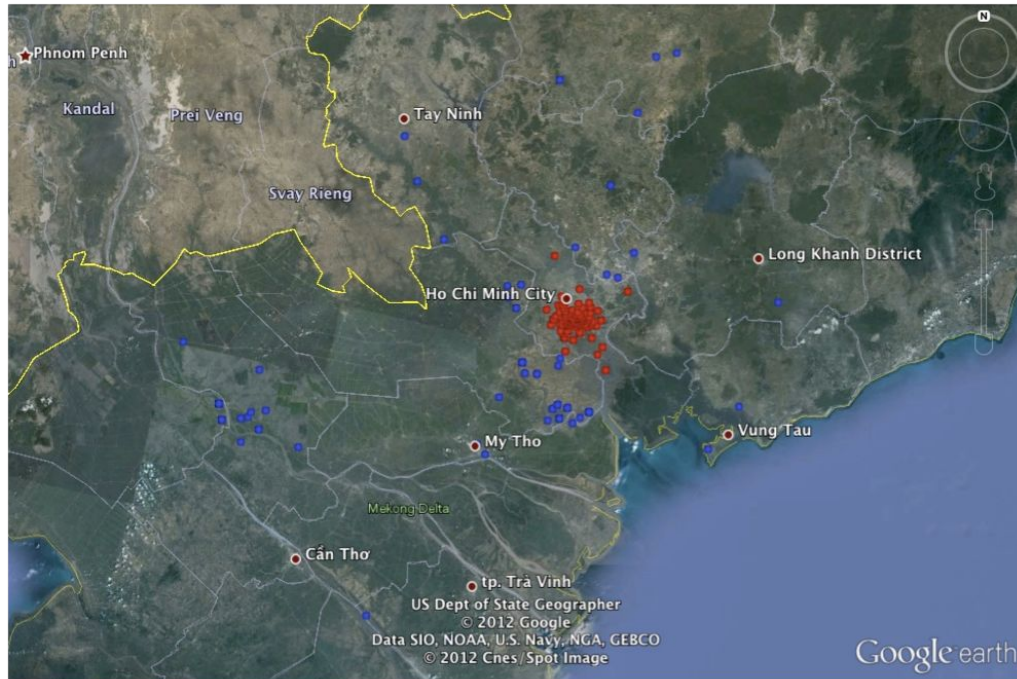
Dengue in southern Vietnam

Bayesian skyline estimates

Spatial structure

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*:

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

However, inference is much more difficult because we must now also infer the location of each lineage through time.

The Migrate-n model

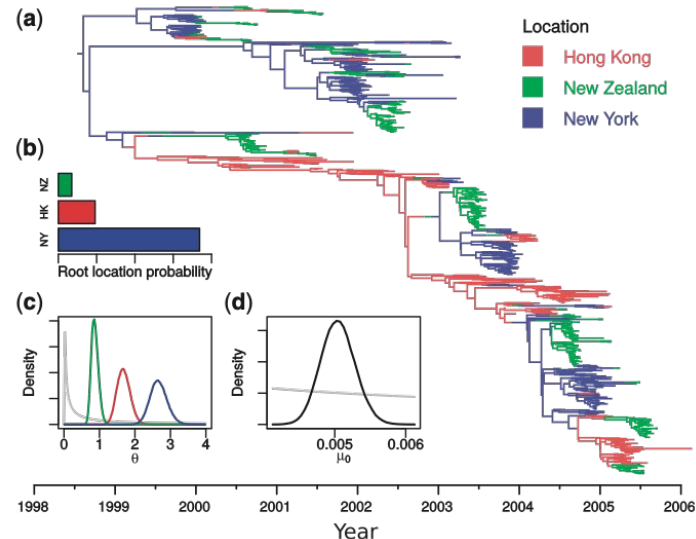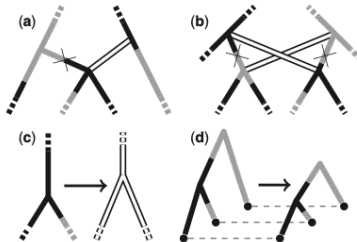A structured coalescent model with migration between *n* subpopulations or demes

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 BUT we need to use MCMC to sample full **migration histories** along each lineage

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
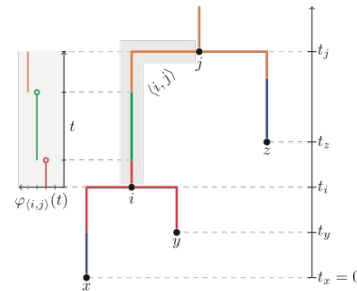
Phylogeography with the SC

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

DTA treats sampling locations as informative about the migration process whereas the SC conditions on sampling locations.

DTA can therefore be highly biased by disproportionate sampling while the SC is more robust to uneven sampling.

DTA vs. the SC

Uneven sampling strongly biases DTA but not the structured coalescent.

De Maio *et al.* (PLoS Genetics, 2015)
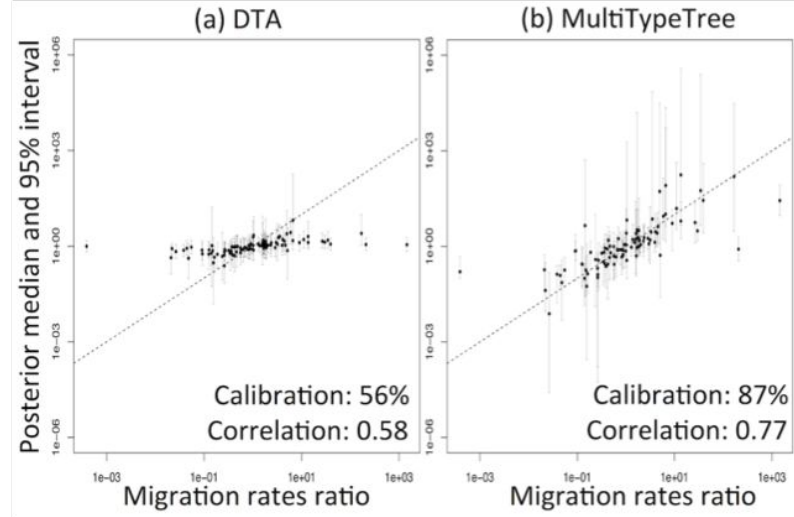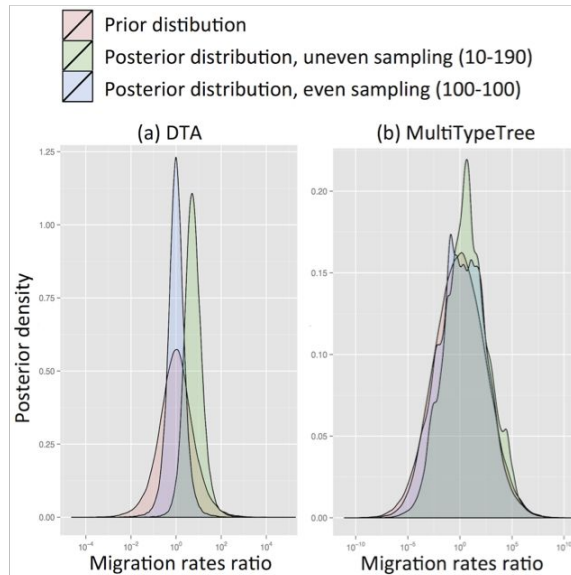
DTA vs. the SC

DTA is also over confident in assigning ancestral state probabilities.

Tomato yellow leaf curl virus

De Maio *et al.* (PLoS Genetics, 2015)
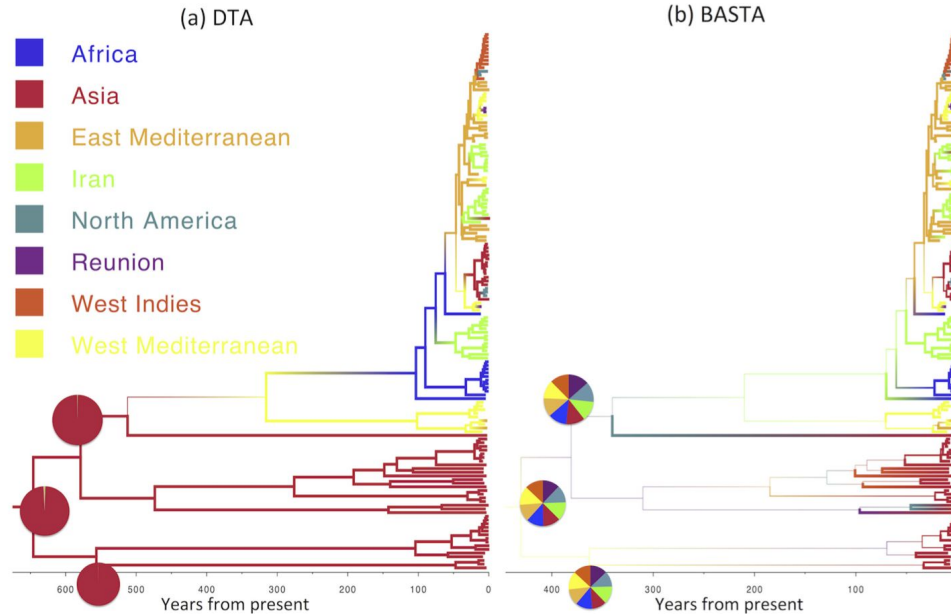
DTA vs. the SC

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 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}. But we'll return to this later.

$$\lambda_{ij} = \sum_k^m \sum_l^m \frac{f_{kl}}{y_k y_l} \left(p_{ik}p_{jl} + p_{il}p_{jk} \right)$$
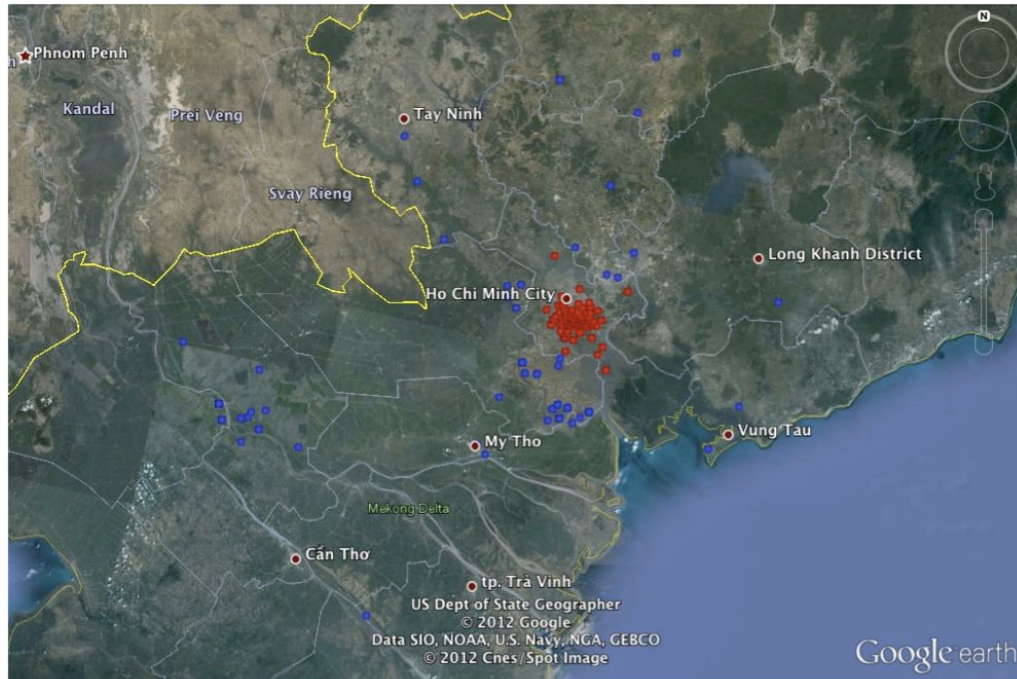
The Volz (2012) Structured Coalescent

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}:
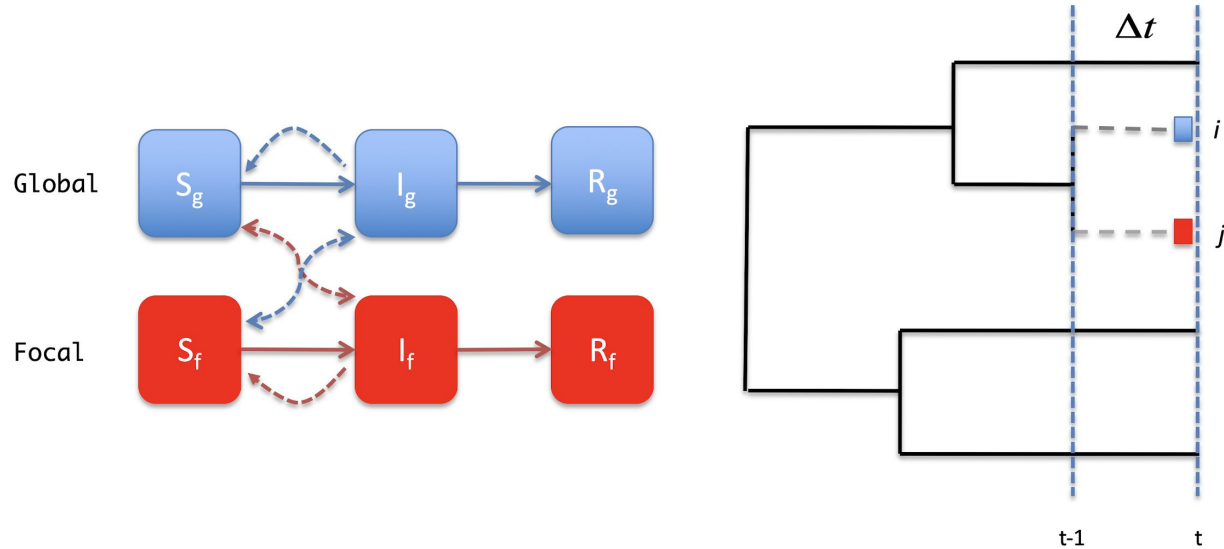

$$\frac{d}{dt}p_{ik} = \sum_{k}^{m}\left(p_{il}g_{kl} - p_{ik}g_{lk}\right)$$

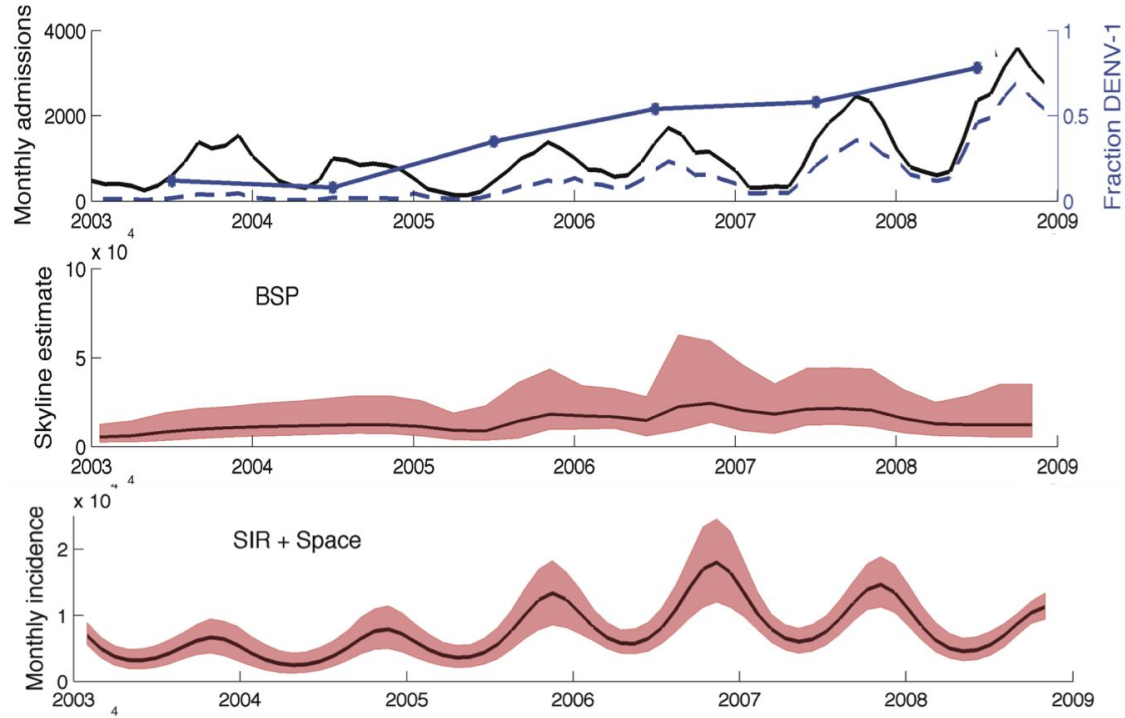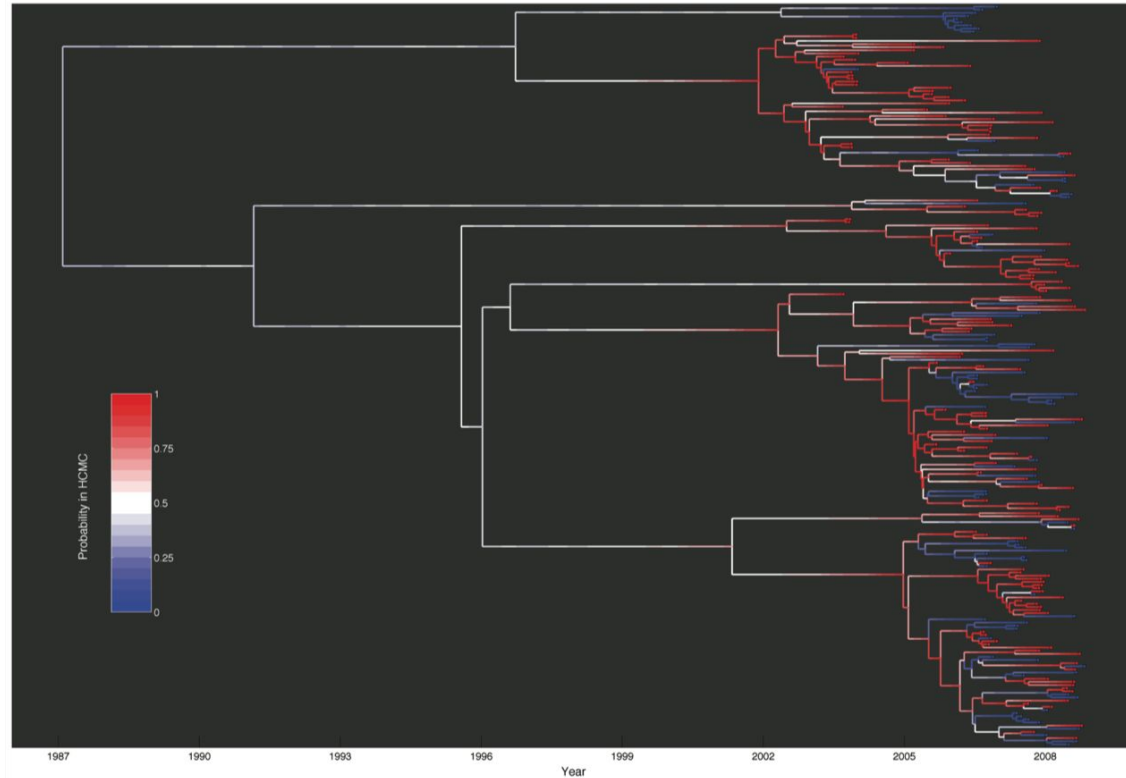Volz (Genetics, 2012)
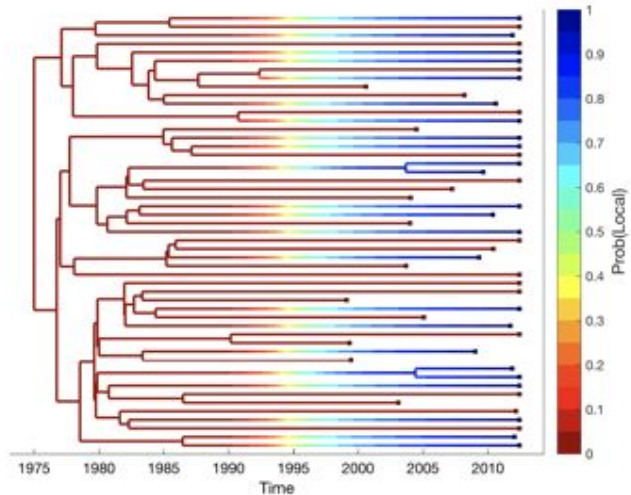
Spatial structure

Spatial SIR model

Structured coalescent model:

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

Estimates accounting for spatial structure

Movement of lineages

Conclusions

Coalescent models relate phylogenies of sampled lineages back to the larger demographic history of a population.

Coalescent methods can be used to reconstruct past population dynamics but other forces, especially population structure, also shape trees.

Structured coalescent models generalize coalescent models and are incredibly useful for modeling different forms of population structure.

SC models improve upon earlier discrete-trait phylogeographic methods but are less computationally efficient. Newer approaches like MASCOT that approximate lineage state probabilities offer a promising alternative.

Bonus lab: MASCOT

The MASCOT package for BEAST 2 implements a structured coalescent model that tracks lineage states probabilistically as in Volz (2012).

Uses an improved approximation to track lineage state probabilities

Allows for inference of pop sizes, migration rates and ancestral states

Can also use GLMs to predict migration rates based on explanatory variables

Müller et al. (Bioinformatics, 2017)