Bayesian coalescent inference of population size history

Alexei Drummond
Center for Computational Evolution
University of Auckland

Taming the BEAST, Engelberg, Switzerland, 2016

28th June 2016

Coalescent theory

The coalescent

Data: a small genetic sample from a large background population.

The coalescent

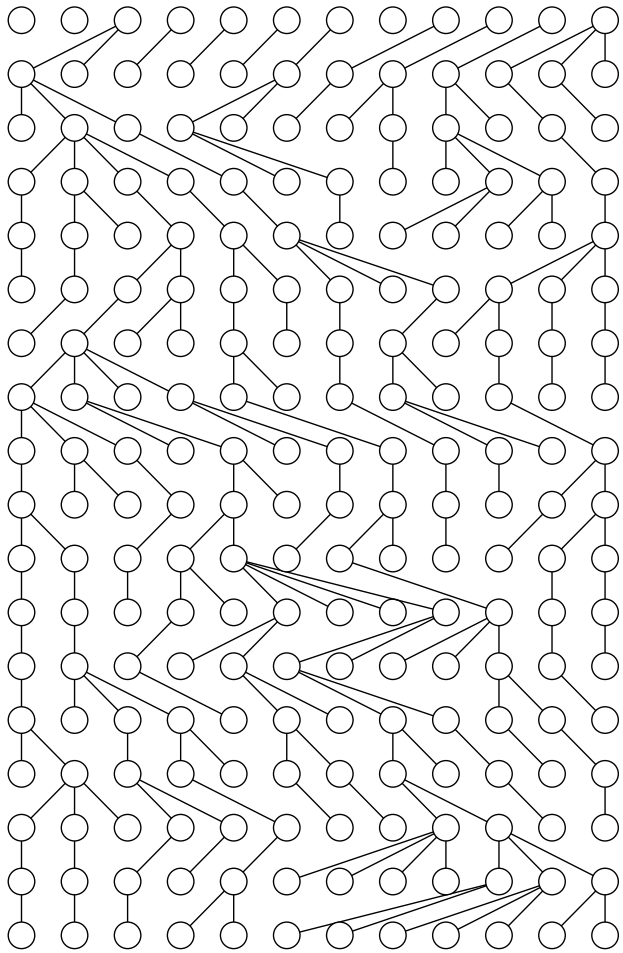
- is a model of the ancestral relationships of a sample of individuals taken from a larger population.
- describes a probability distribution on ancestral genealogies (trees) given a population history, $\mathbf{N(t)}$.
  - Therefore the coalescent can convert information from ancestral genealogies into information about population history and vice versa.
- a model of ancestral genealogies, not sequences, and its simplest form assumes neutral evolution.
- can be thought of as a prior on the tree, in a Bayesian setting.

Theoretical population genetics

Most of theoretical population genetics is based on the idealized Wright-Fisher model of population which assumes

- Constant population size N
- Discrete generations
- Complete mixing

For the purposes of this presentation the population will be assumed to be haploid, as is the case for many pathogens.

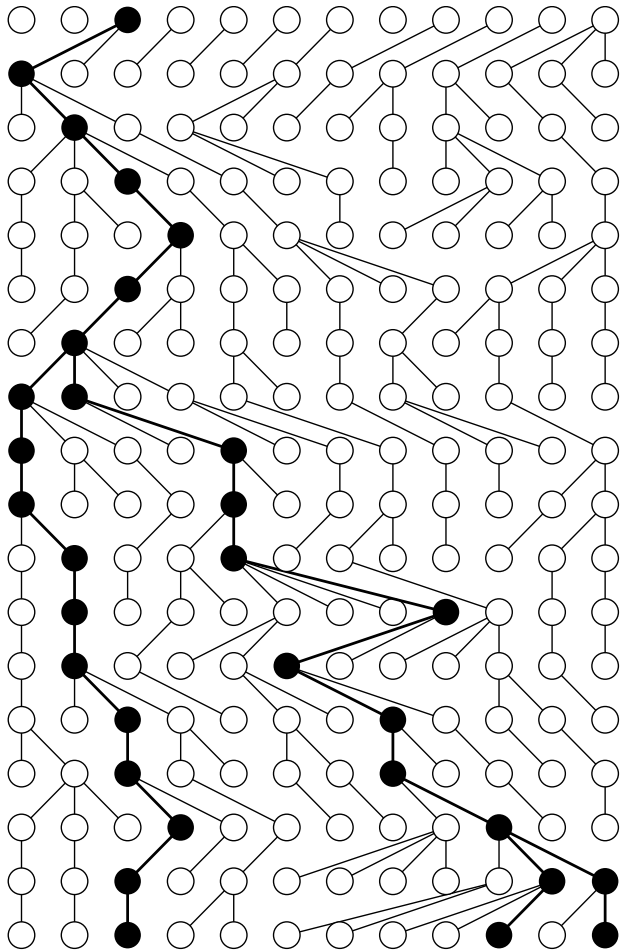


Kingman's n-coalescent

Consider tracing the ancestry of a sample of $\mathbf{k}$ individuals from the present, back into the past.

This process eventually *coalesces* to a single common ancestor (*concestor*) of the sample of individuals.

Kingman's n-coalescent describes the statistical properties of such an ancestry when $\mathbf{k}$ is small compared to the total population size $\mathbf{N}$.

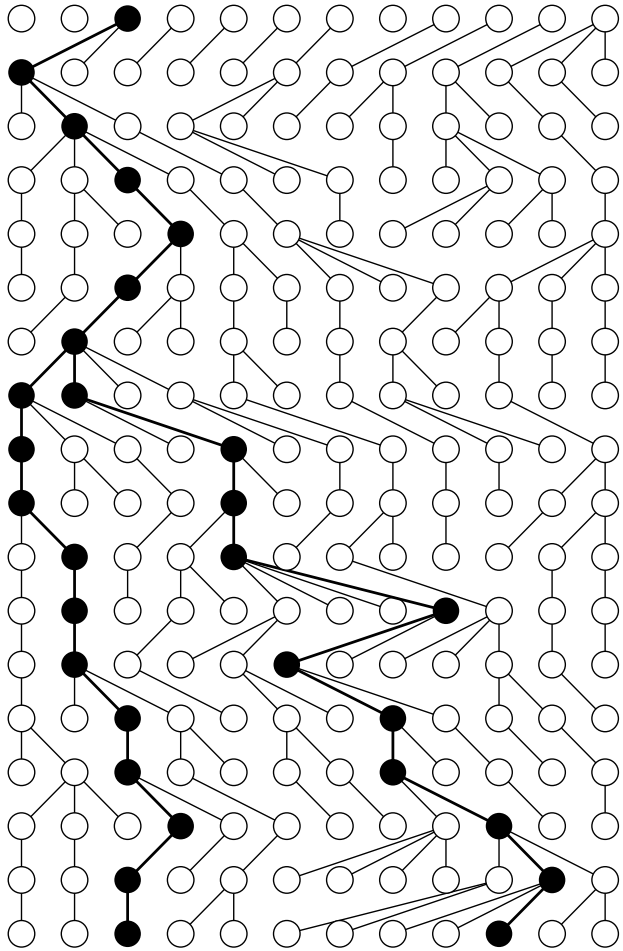


The coalescence of two ancestral lineages

- First, consider two random members from a population of fixed size N.
- By perfect mixing, the probability they share a *concestor* in the previous generation is $1/N$.
- The probability the concestor is t generations back is

$$Pr\{t\} = \frac{1}{N}(1 - \frac{1}{N})^{t-1}.$$

- It follows that $g = t - 1$, has a geometric distribution with a success rate of $\lambda = 1/N$, and so has mean N and variance of $N^3/(N-1)$.

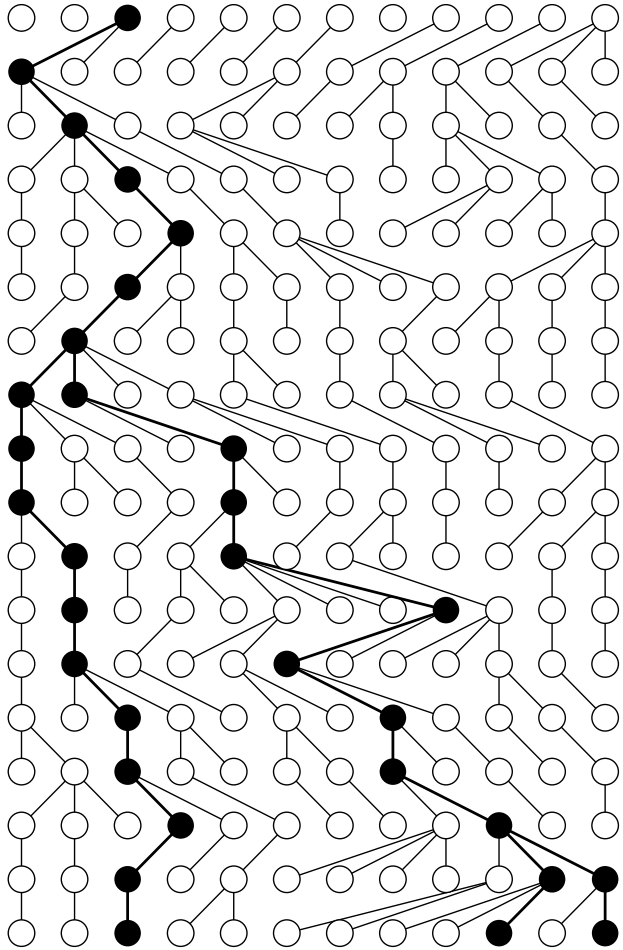


The coalescence of k lineages

With k lineages the time to the first coalescence is derived in the same way, only now there are $\binom{k}{2}$ possible pairs that may coalesce, resulting in a success rate of $\lambda = \binom{k}{2}/N$ and mean time to first coalescence (t_k) of

$$E[t_k] = \frac{N}{\binom{k}{2}}.$$

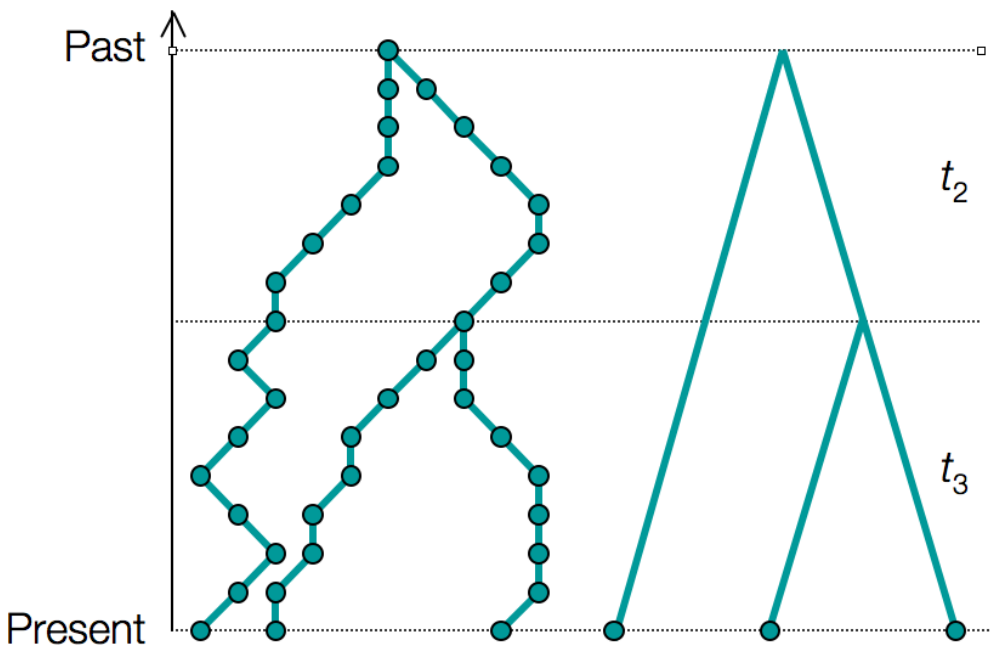
This implicitly assumes that N is much larger than $O(k^2)$, so that the probability of two coalescent events in the same generation is small.



The coalescent likelihood for a genealogy

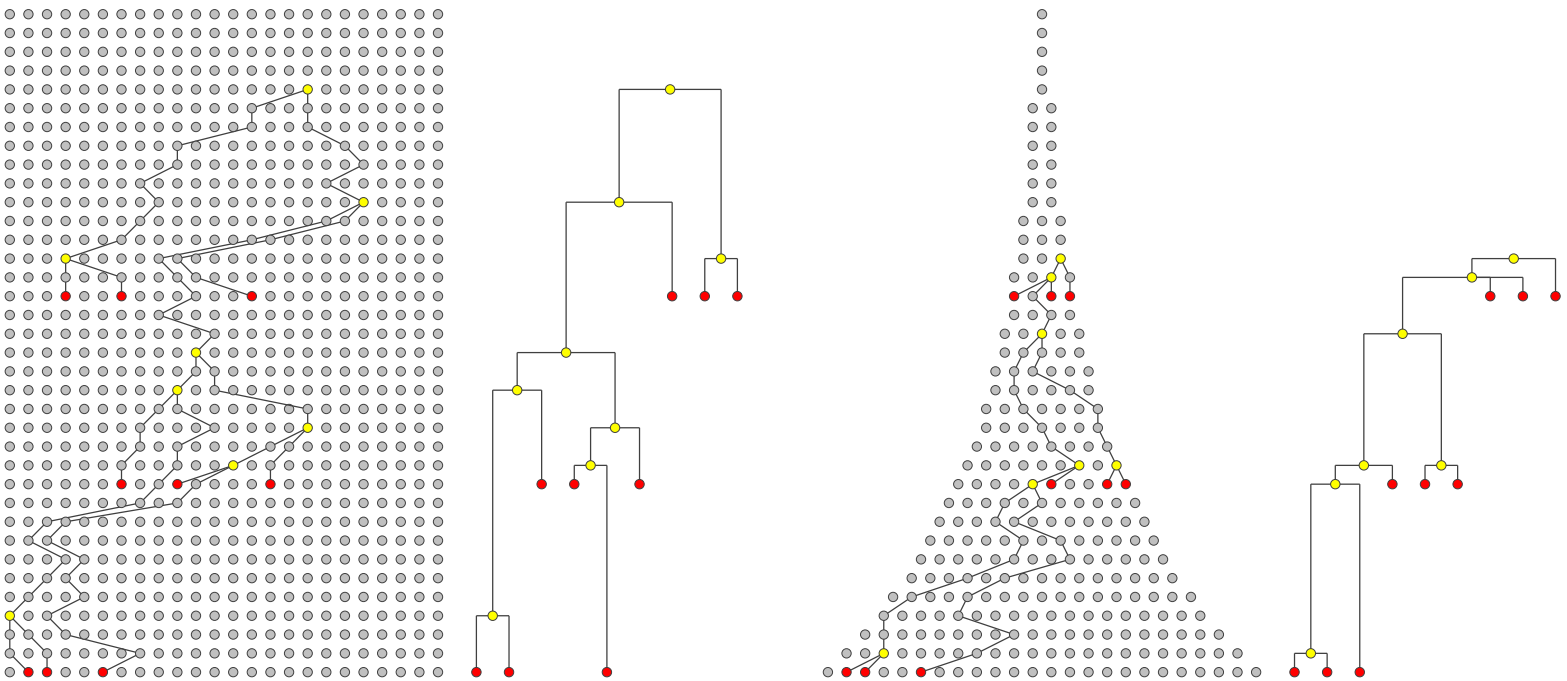
For a genealogy with known coalescent times $\mathbf{t} = \{t_2, t_3, ..., t_n\}$ we can write the likelihood:

$$f(\mathbf{t}|N) = \frac{1}{N^{n-1}} \prod_{k=2}^{n} \exp\left(-\frac{\binom{k}{2} t_k}{N}\right) .$$



The coalescent with serial samples

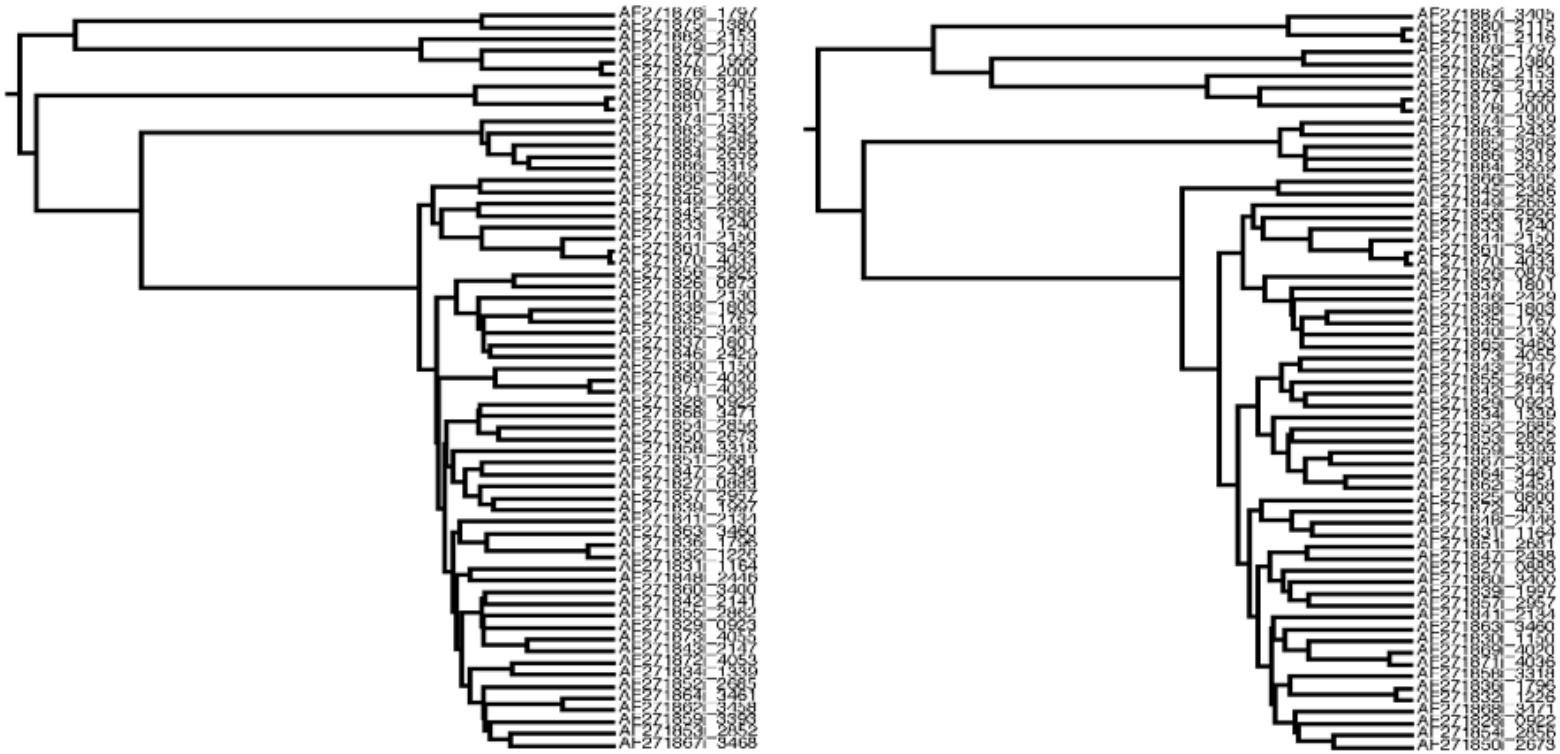
Many epidemiological agents, like RNA viruses, evolve very rapidly, so that the effect of sampling the population at different times becomes important. Ancient DNA also requires care handling of sampling times.



Constant size

Exponential growth

Bayesian integration of uncertainty in genealogies

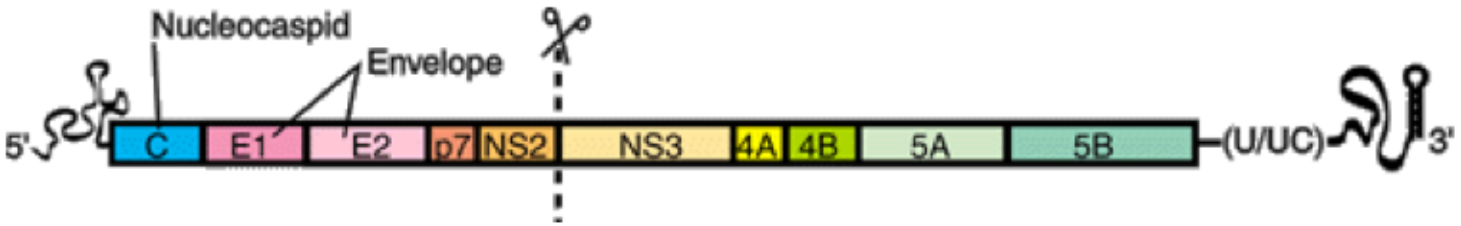


How similar are these two trees? Both of them are plausible given the data. We can use Bayesian Markov-chain Monte Carlo to average the coalescent over all plausible trees.

Hepatitis C in Egypt

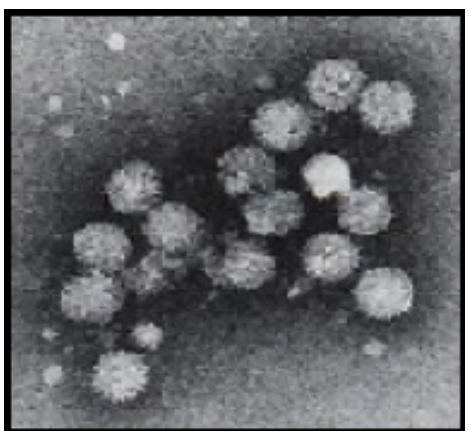
Hepatitis C in Egypt

A case study of the coalescent approach to molecular epidemiology



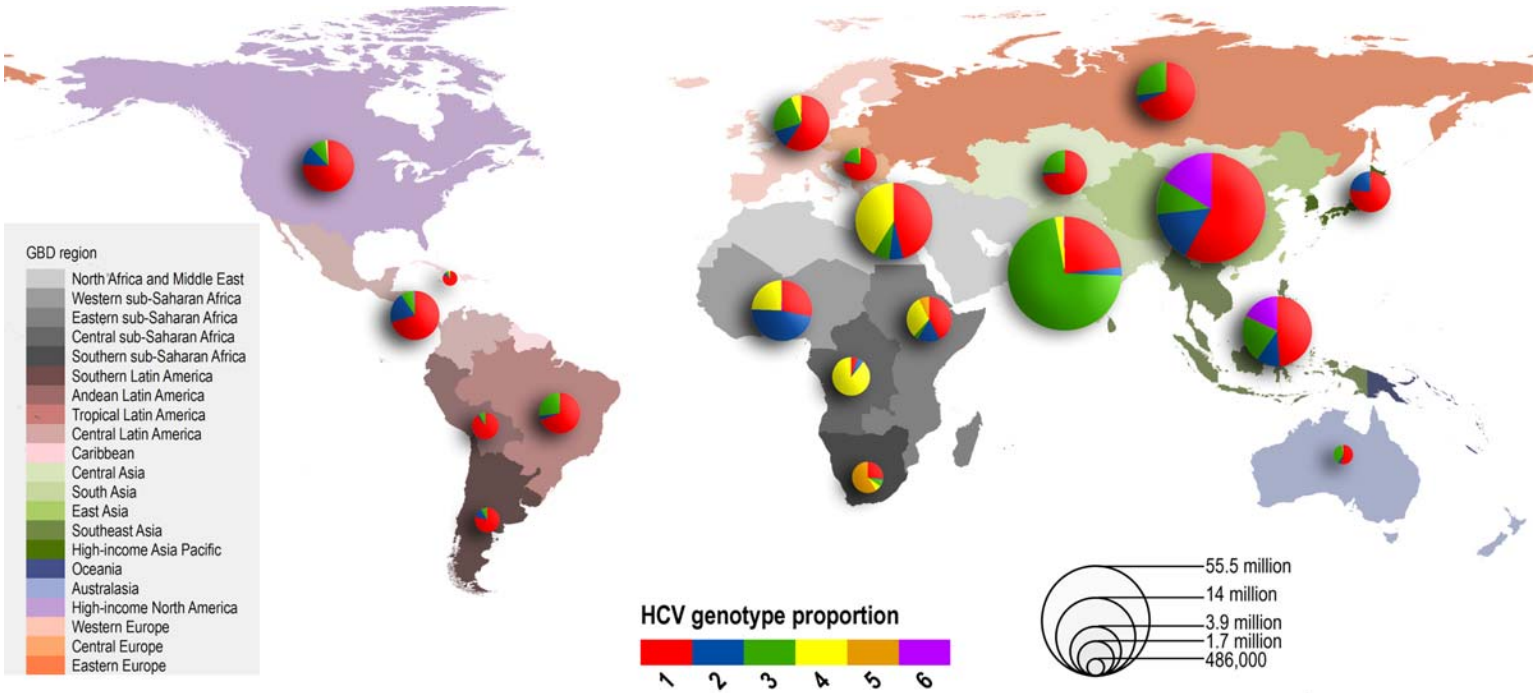
Hepatitis C (HCV)

- Identified in 1989
- 9.6kb single-stranded RNA genome
- Polyprotein cleaved by proteases
- Tissue culture system only recently developed



How important is Hepatitis C?

$>$ 185 million people infected worldwide



- ~80% infections are chronic
- Liver cirrhosis and cancer risk
- 10,000 deaths per year in USA
- No protective immunity?

HCV Transmission

By percutaneous exposure to infected blood

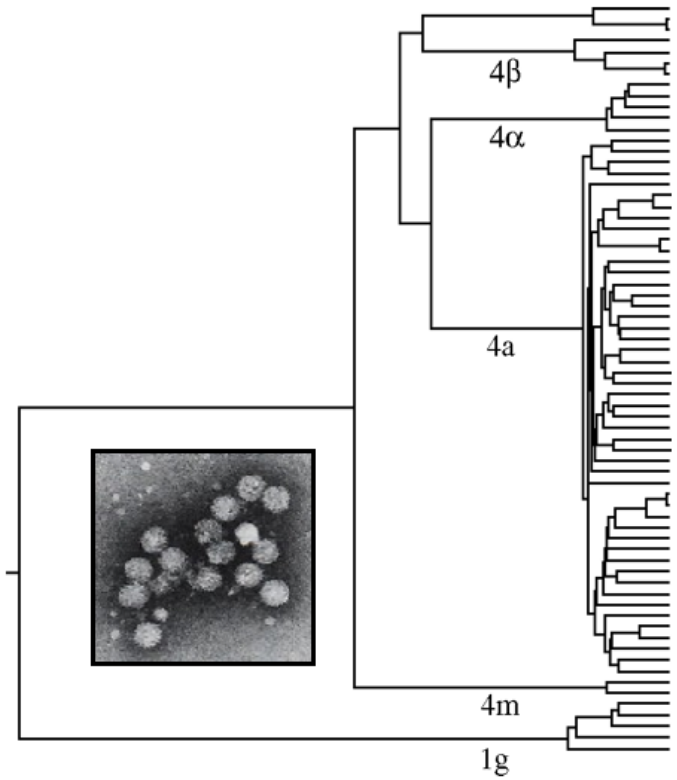
- Blood transfusion / blood products
- Injecting and nasal drug use
- Sexual and vertical transmission
- Unsafe injections
- Unidentified routes

Coalescent population inference of HCV

Pybus *et al* (2003) *Molecular Biology and Evolution*

Egyptian HCV gene sequences
n=61
E1 gene, 411bp

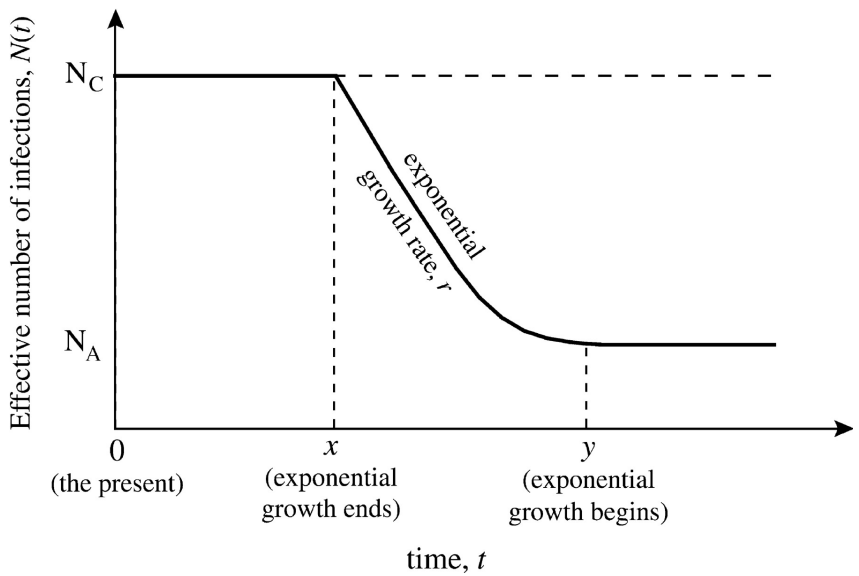
- All sequence contemporaneous
- Egypt has highest prevalence of HCV worldwide (10-20%)
- But low prevalence in neighbouring states
- Why is Egypt so seriously affected?
- Parenteral antischistosomal therapy (PAT)?



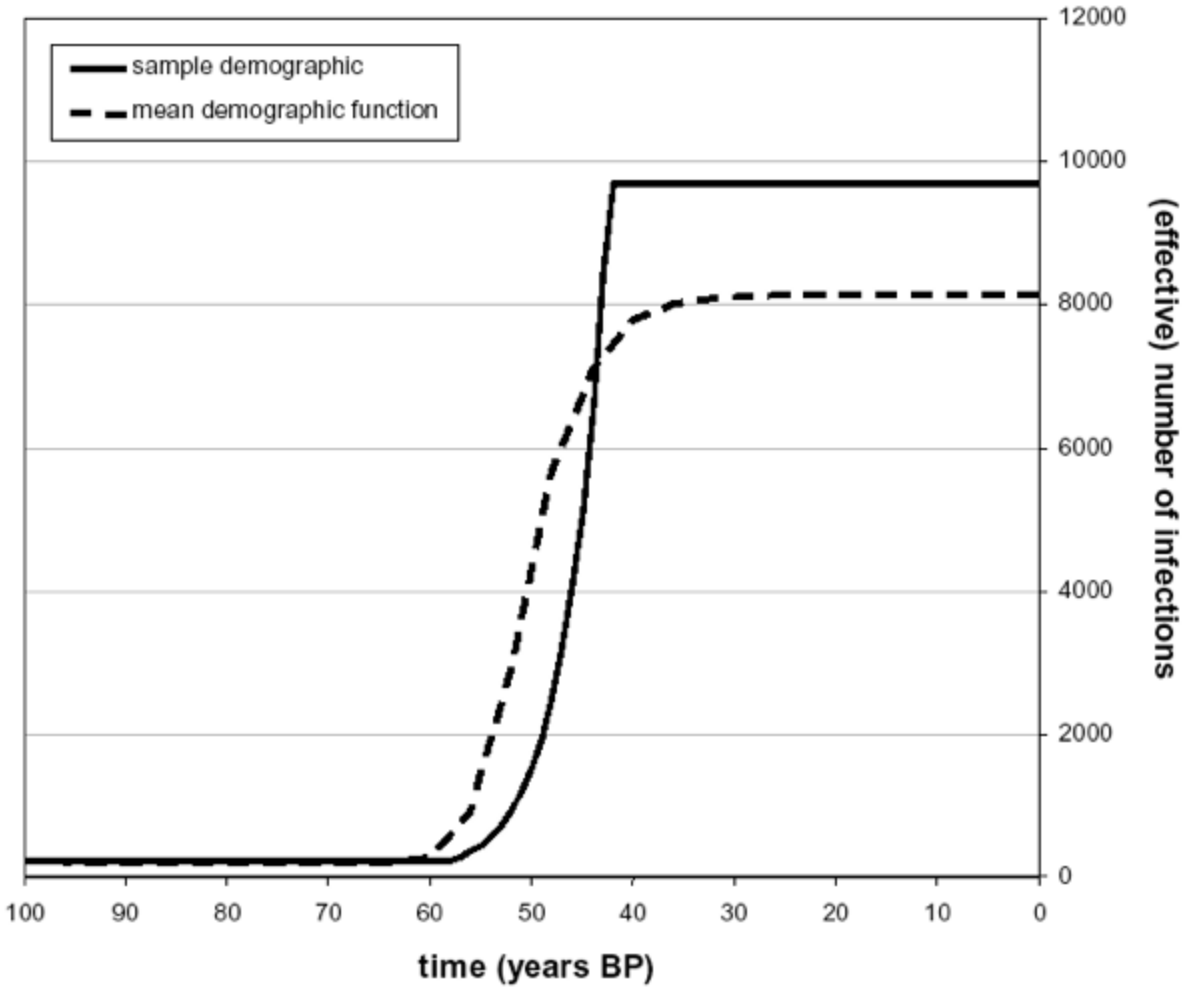
Demographic model for Hepatitis C in Egypt

- The coalescent can be extended to model any integrable function of varying population size.
- The model we used was a const-exp-const model.
- A Bayesian MCMC method was developed to sample the gene genealogy, the substitution model and demographic function simultaneously.

$$N(x) = \begin{cases} N_c & \text{if } t \leqslant x \\ N_c e^{-r(t-x)} & \text{if } x < t < y \\ N_A & \text{if } t \geqslant y \end{cases}$$

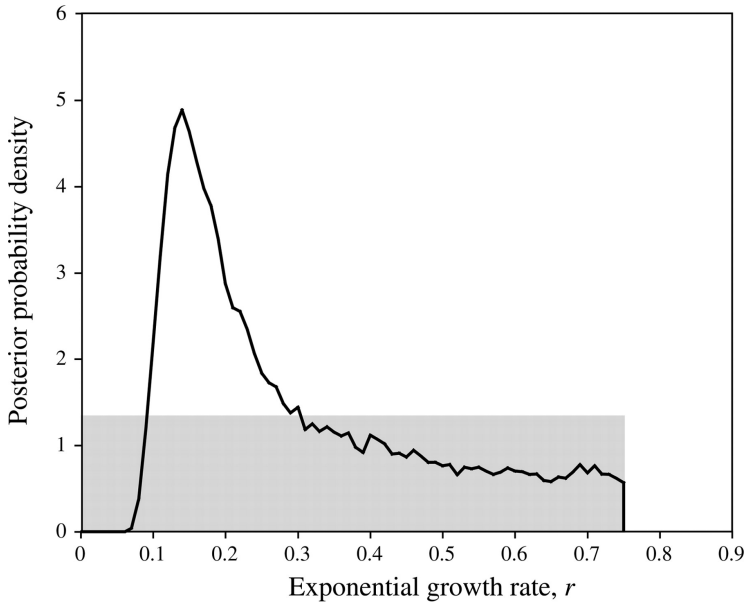


Estimated population history of HCV in Egypt



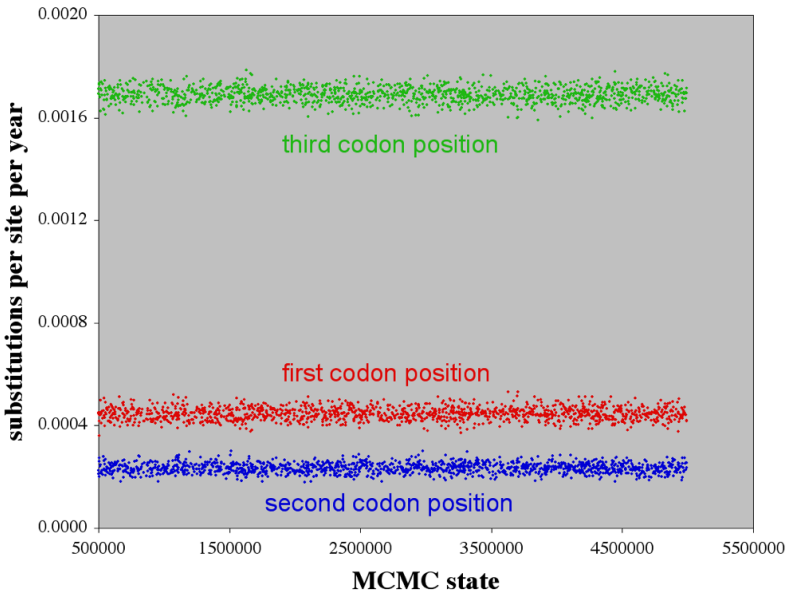
Uncertainty in parameter estimates

Demographic parameters



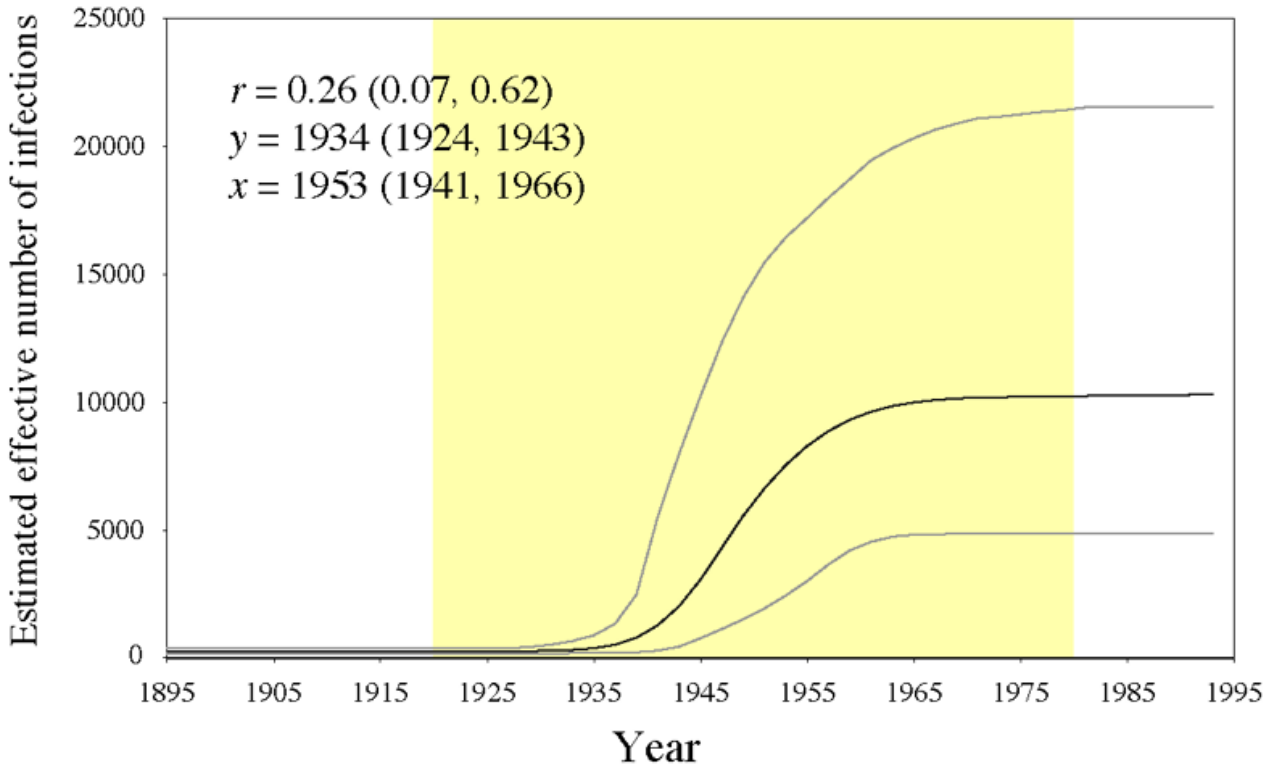
Growth rate of the growth phase
Grey box is the prior

Mutational parameters



Rates at different codon positions,
All significantly different

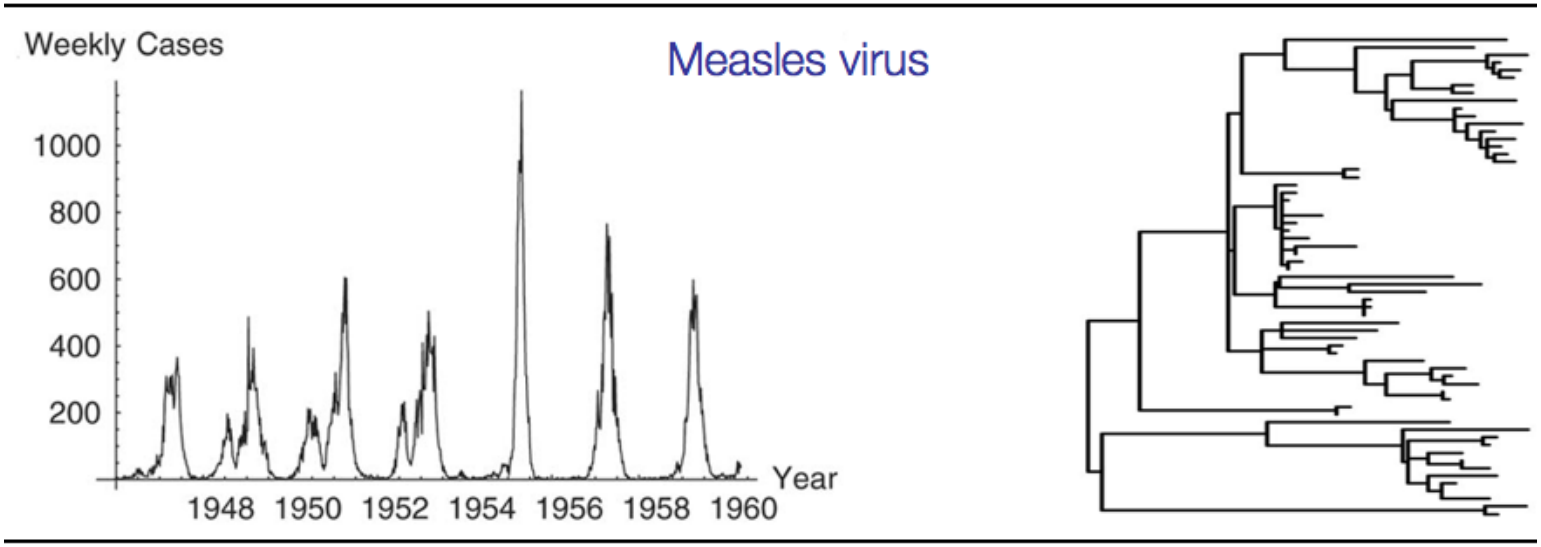
Full Bayesian Estimation

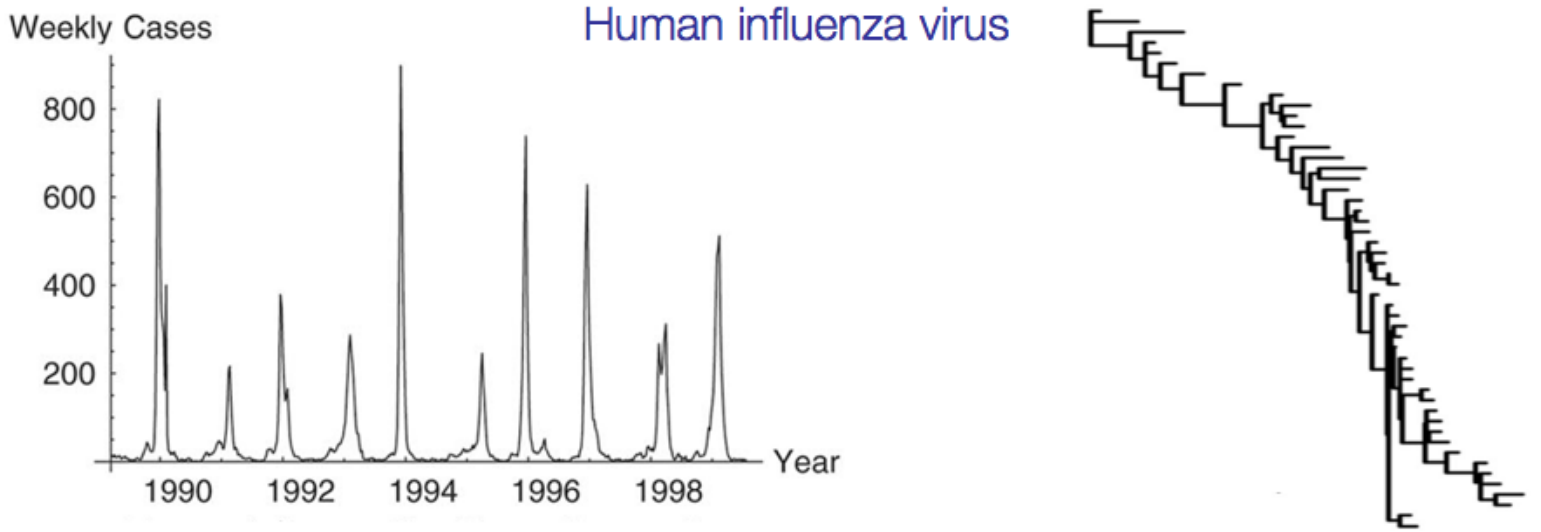


- Marginalized over uncertainty in genealogy and mutational processes
- Yellow band represents time over which PAT was employed in Egypt

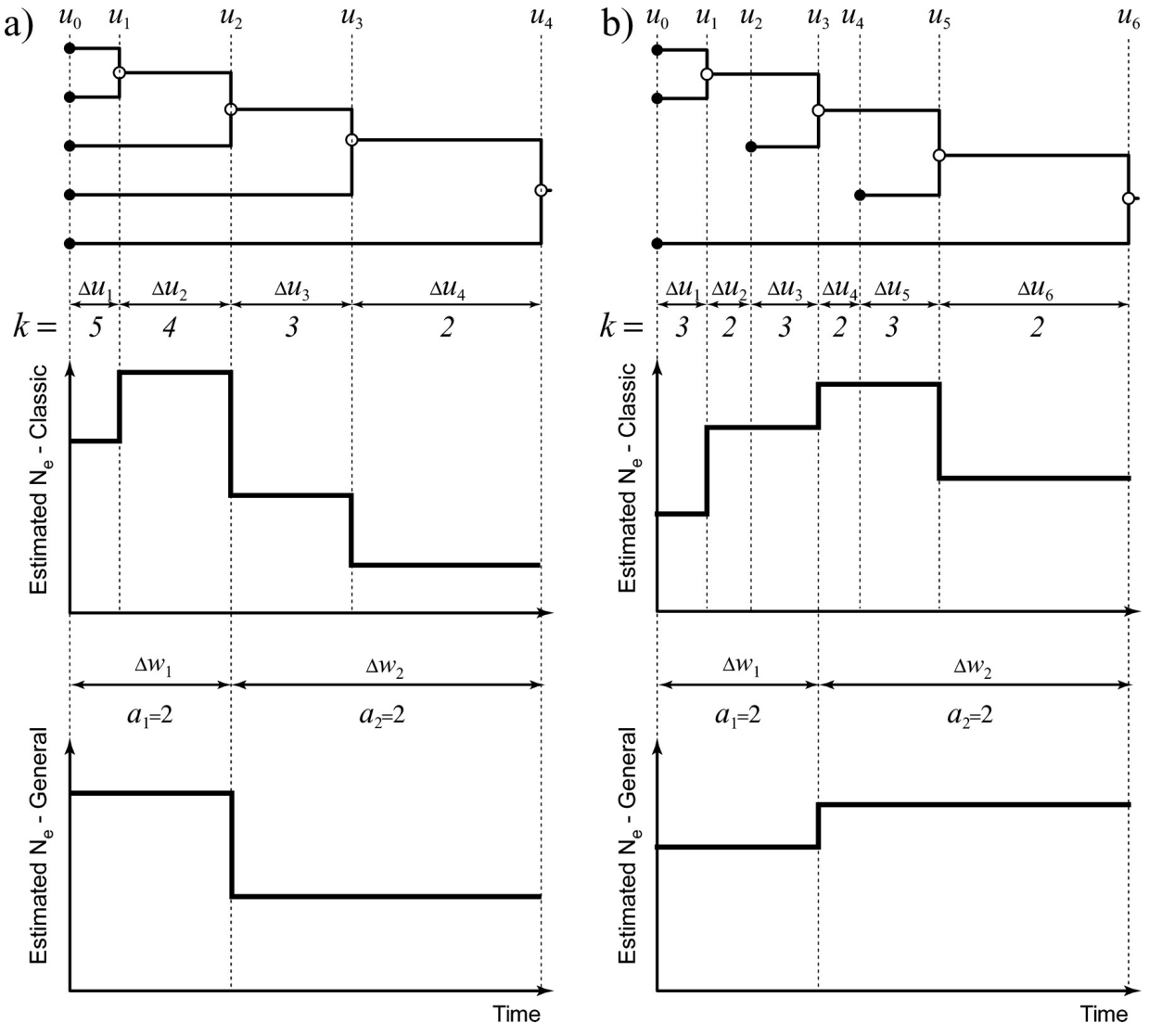
Bayesian skyline plot

Virus Phylodynamics



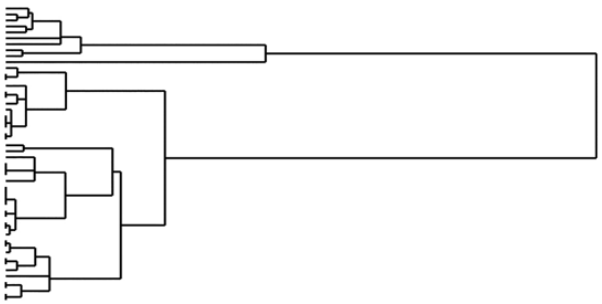


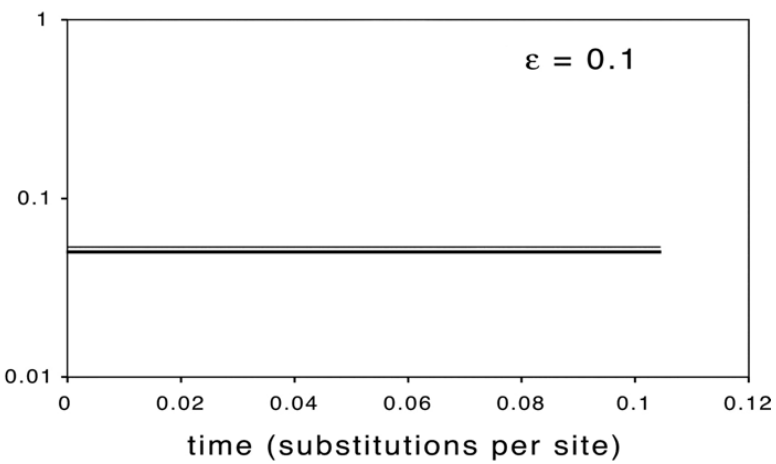
"Skyline" coalescent model



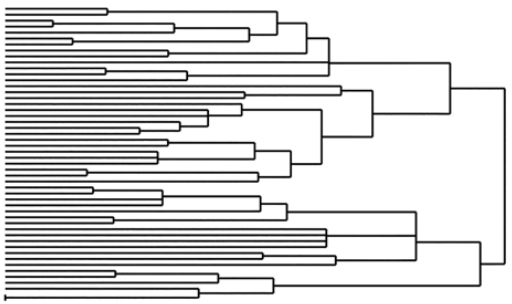
The generalized skyline plot - simulated data

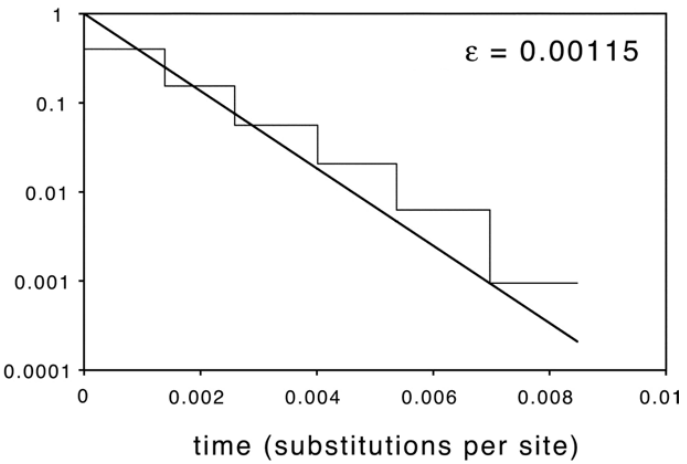
Constant population size,
$N(t) = N_0$





Exponential growth,
$N(t) = N_0 e^{-rt}$



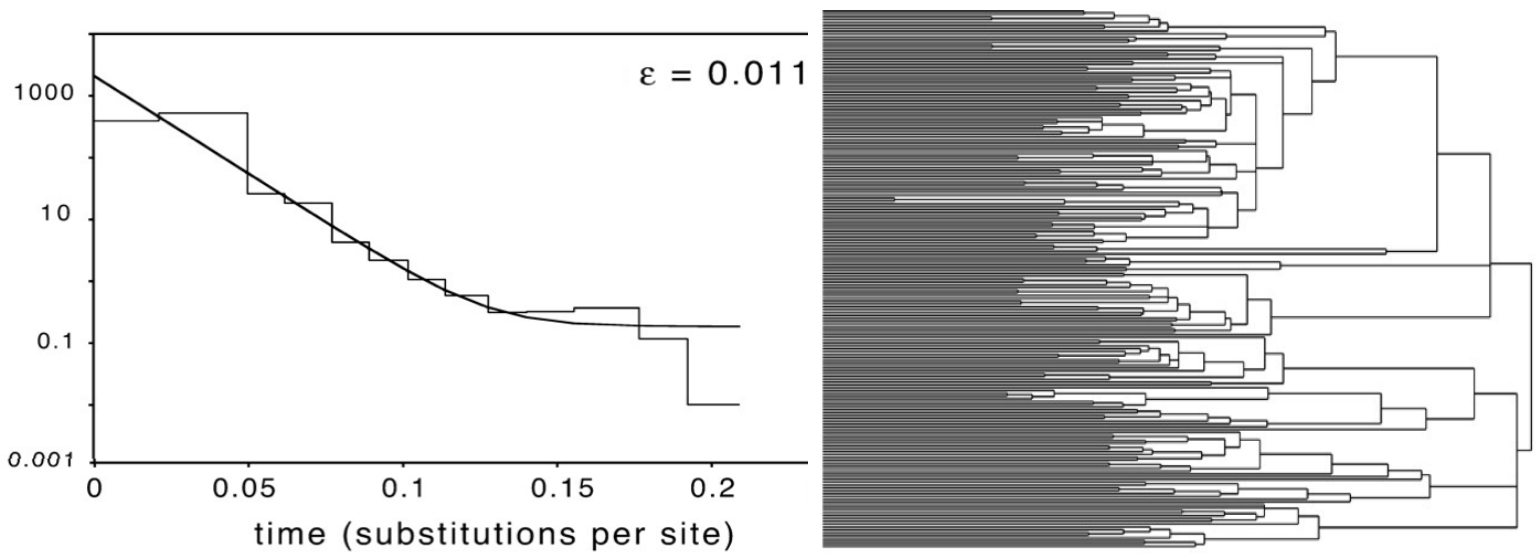


The generalized skyline plot - HIV-1 group M

The tree used here was estimated in Yusim *et al* (2001) *Phil. Trans. Roy. Soc. Lond. B* **356**:855-866.

The black curve is a parametric coalescent estimate obtained from the same data under the expansion model,

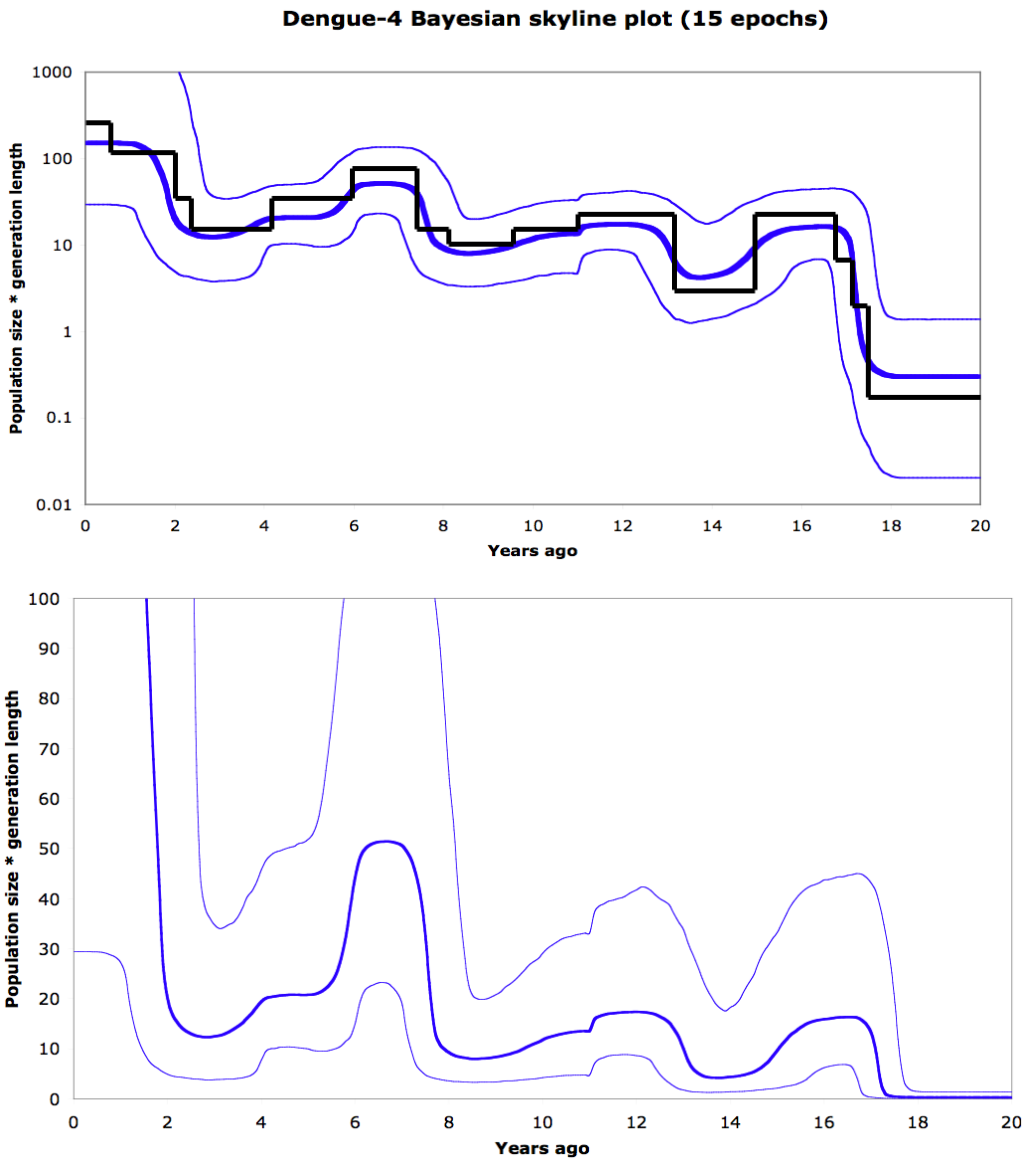
$N(t) = (N_0 - N_A)e^{-rt} + N_A$



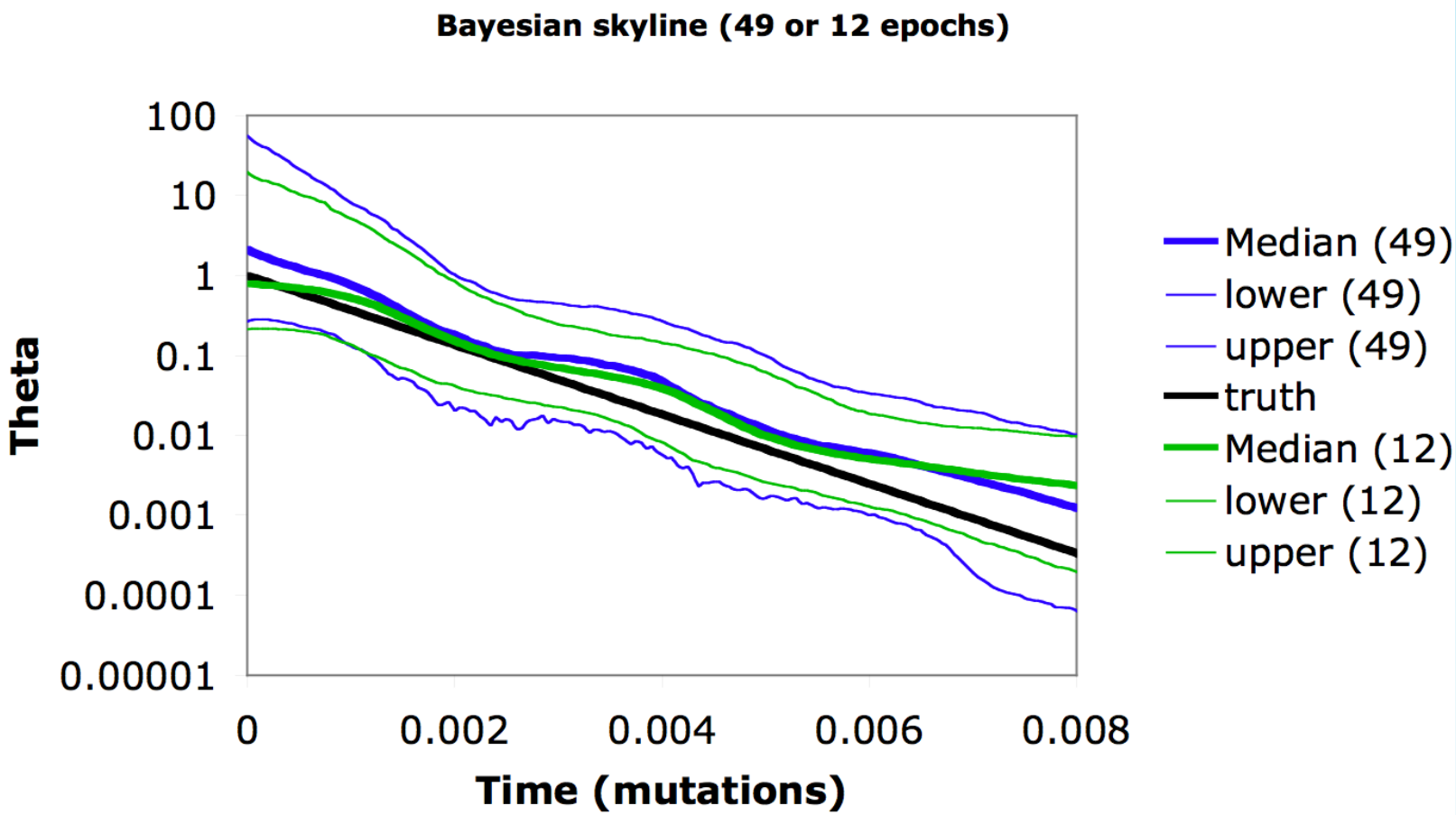
The Bayesian skyline plot

Drummond *et al* (2005) *Molecular Biology and Evolution*

The Bayesian skyline plot estimates a demographic function that has a certain fixed number of steps (in this example 15) and then integrates over all possible positions of the break points, and population sizes within each epoch.

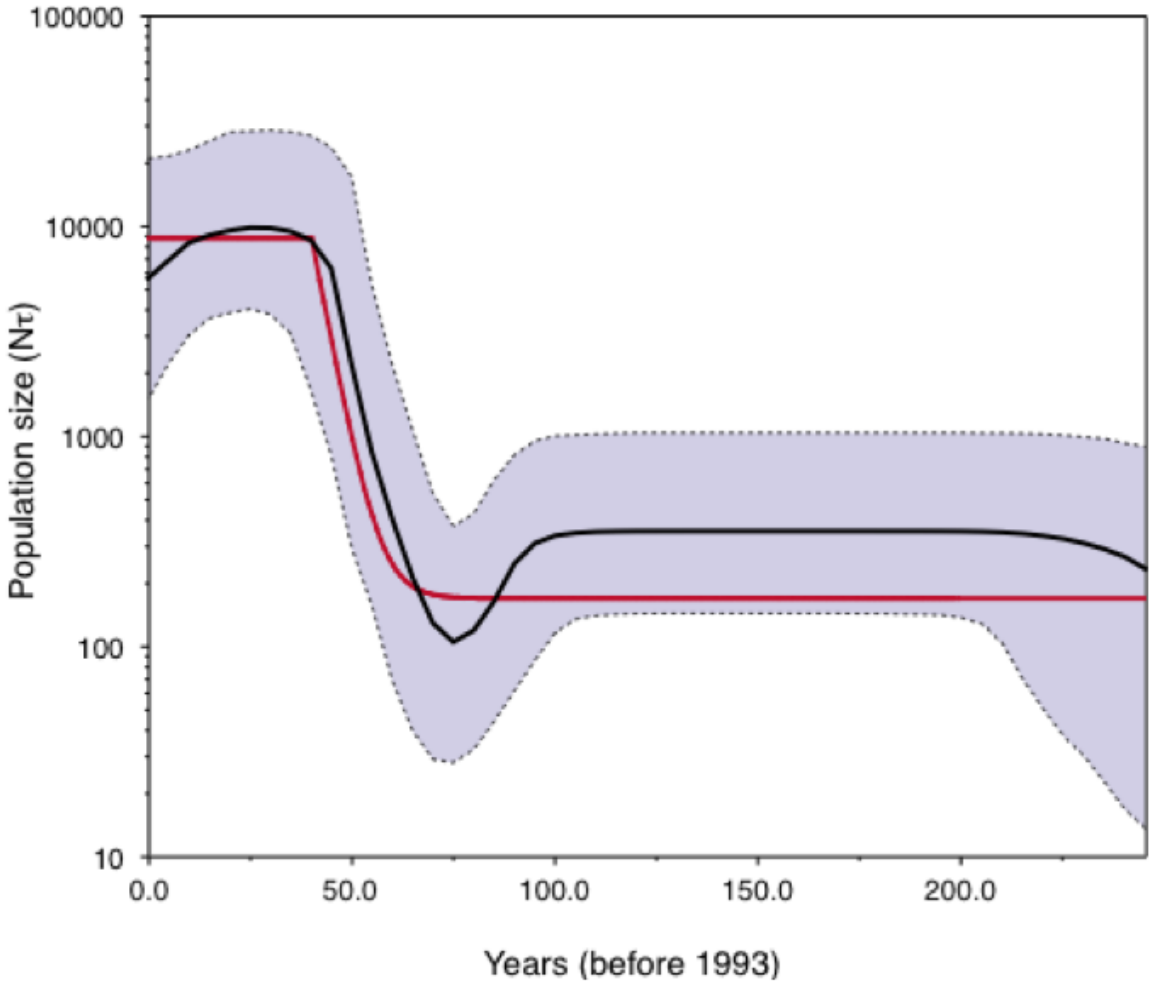


Validating the Bayesian skyline plot



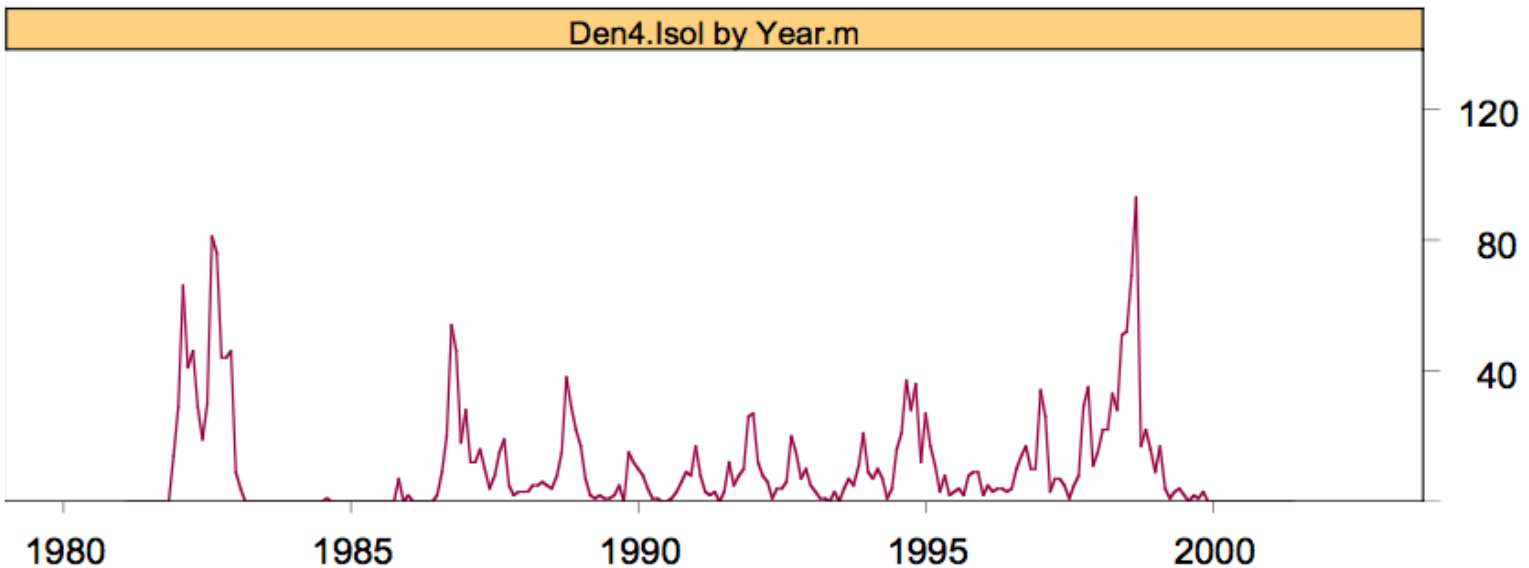
Comparison of BSP to parametric coalescent

Hepatitis C in Egypt



Modeling complex demographic history

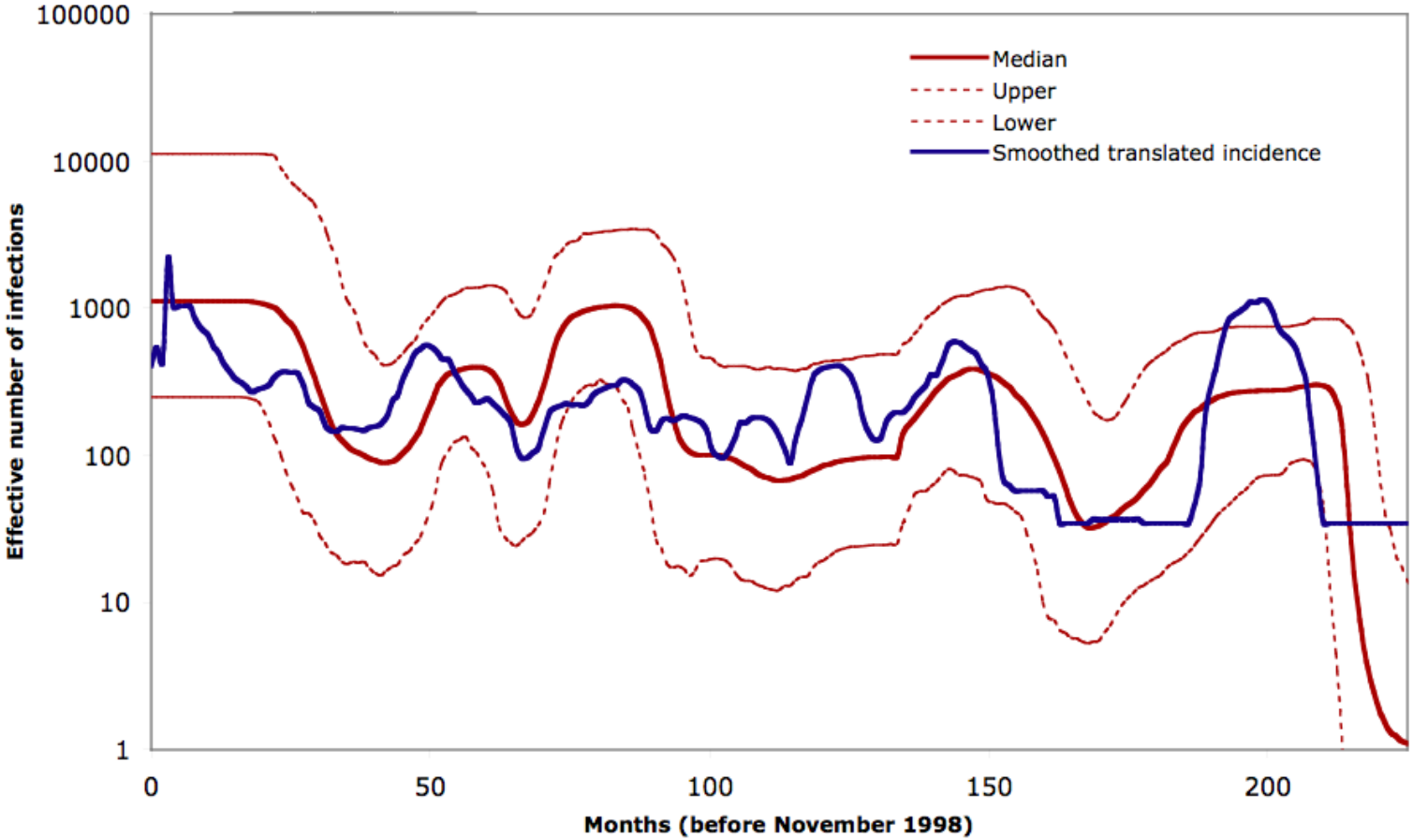
Dengue 4 in Puerto Rico



- $N(t) = N_0 \exp(-rt)$
  - log marginal likelihood = -10566.421
- $N(t)$ = scaled-translated case data
  - log marginal likelihood = -10478.572

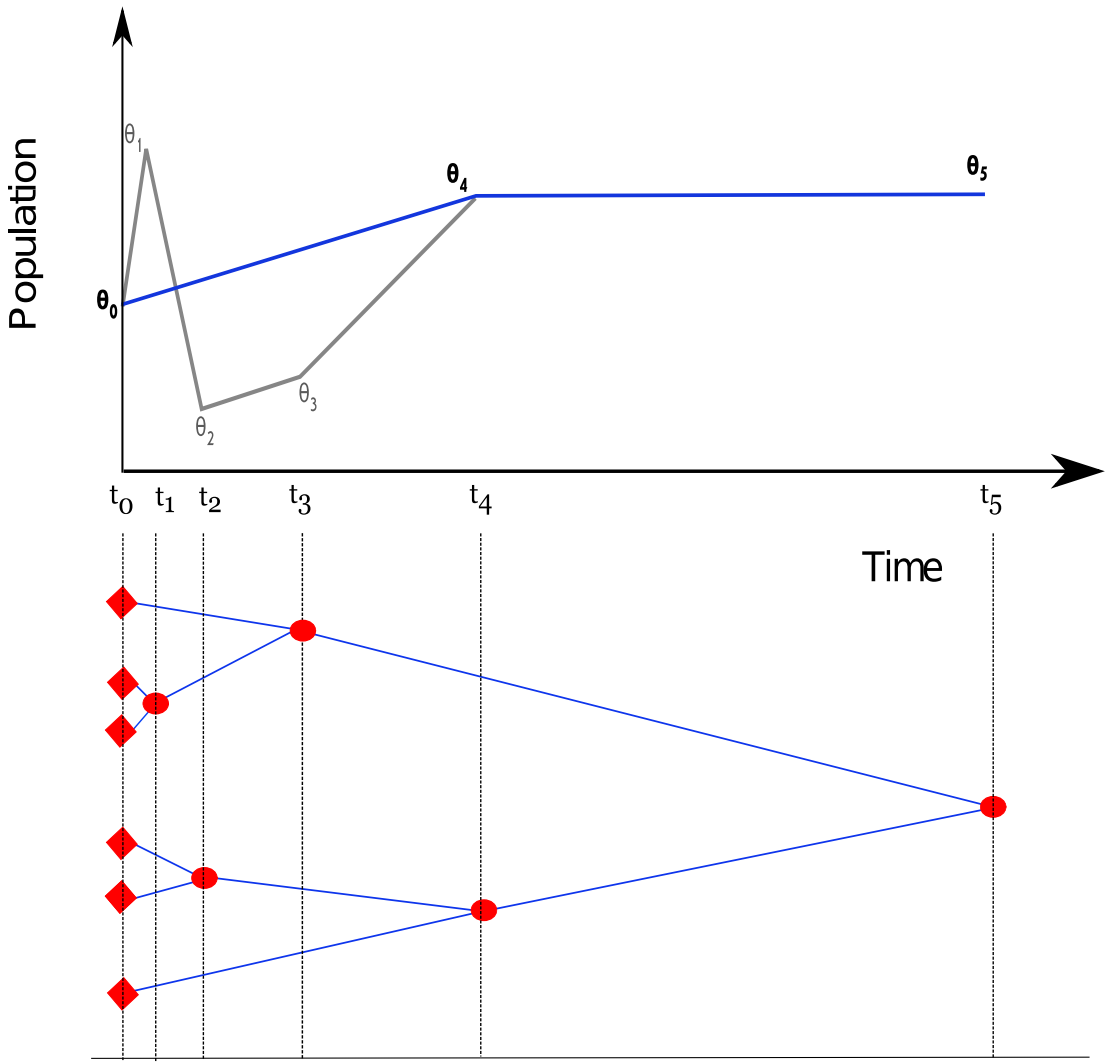
Comparing BSP to incidence data

Dengue 4 in Puerto Rico

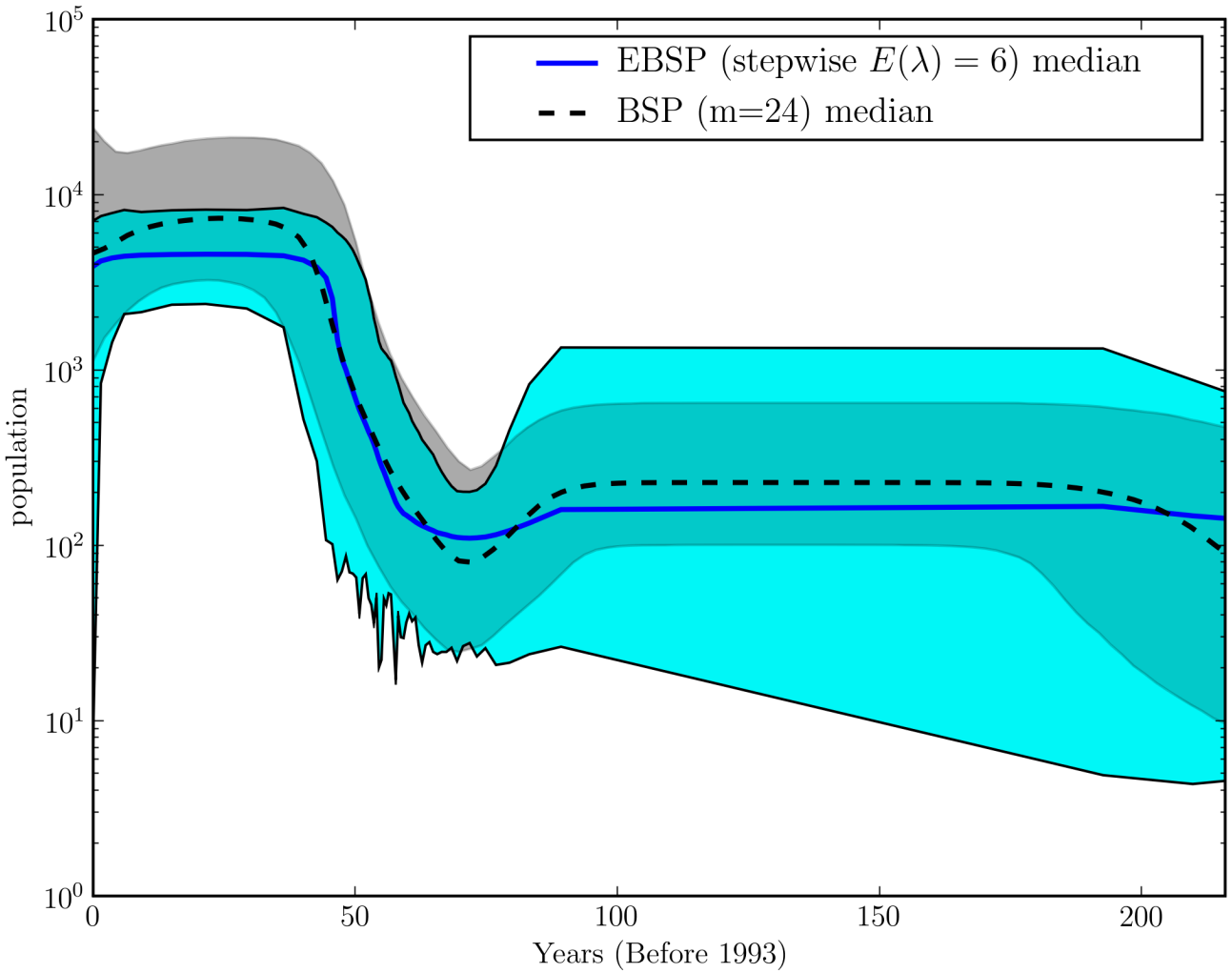


Extended BSP: Stochastic Variable Selection

Heled and Drummond (2008)

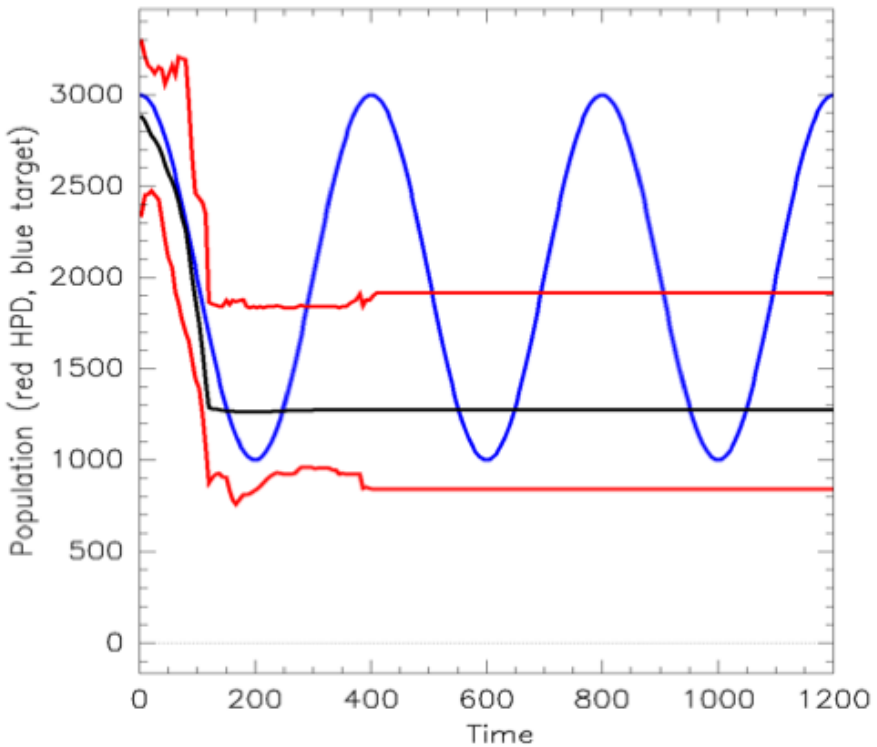


Comparison of EBSP with BSP on Egypt HCV



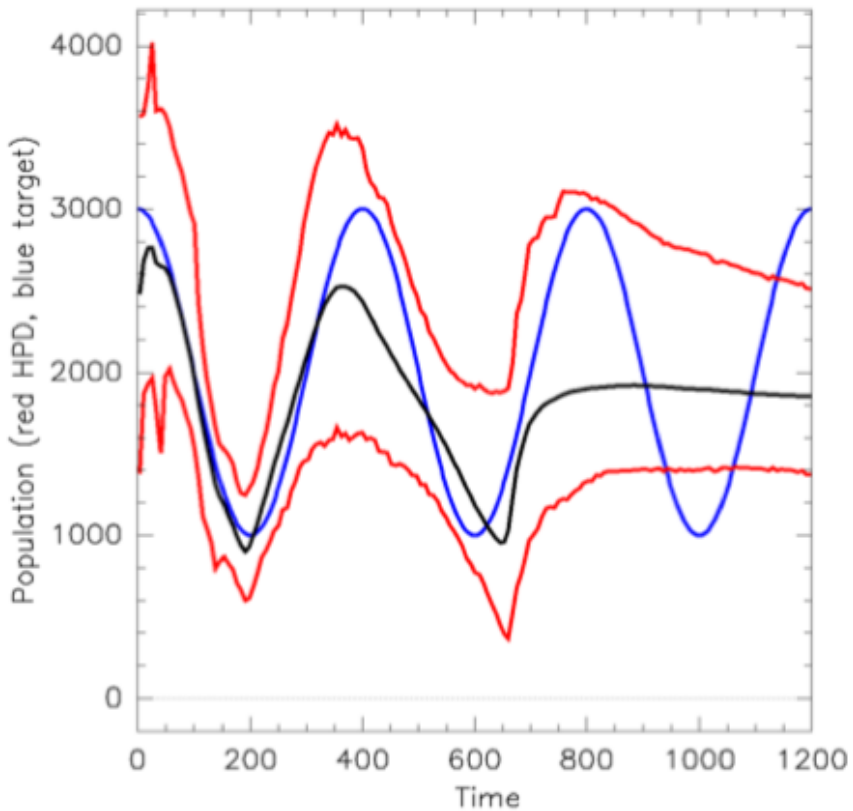
Detecting evolutionary bottlenecks using EBSP

480 contemporaneous samples from a single locus



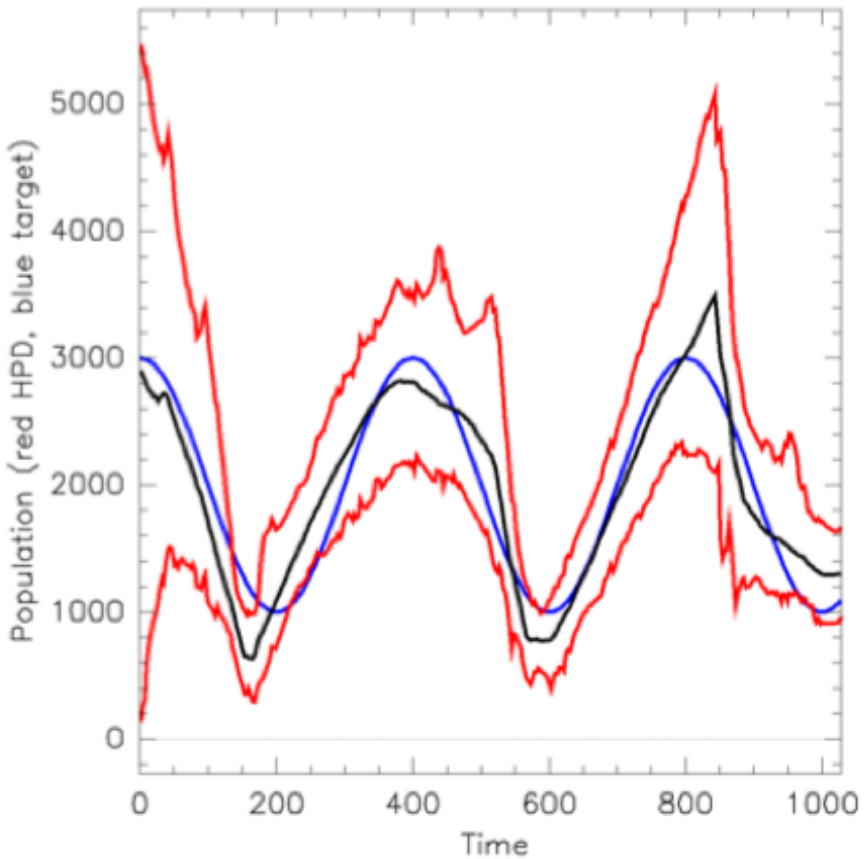
Detecting evolutionary bottlenecks using EBSP

16 contemporaneous samples from each of 32 loci



Detecting evolutionary bottlenecks using EBSP

480 samples sampled through time from a single locus



The population genetic dynamics of Influenza A

Rambaut *et al* (2008) *Nature* 453:615-620

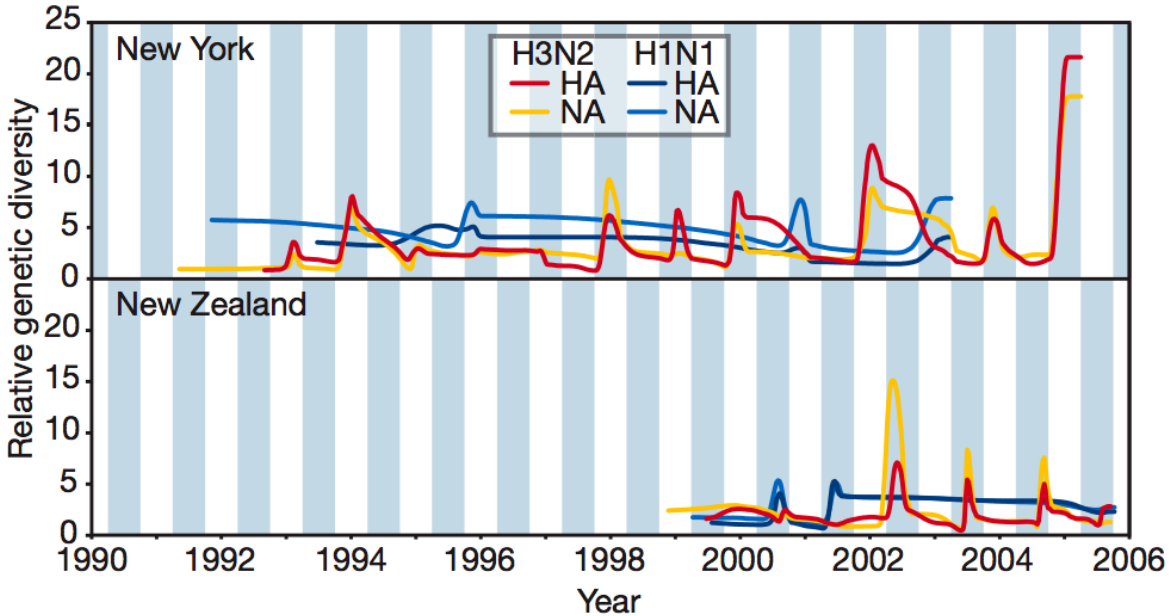


Figure 1 | Population dynamics of genetic diversity in influenza A virus. Bayesian skyline plots of the HA and NA segments for the A/H3N2 and A/H1N1 subtypes in New York state (top) and New Zealand (bottom). The horizontal shaded blocks represent the winter seasons. The *y*-axes represent a measure of relative genetic diversity (see Methods for details). The shorter timescale of New Zealand skyline plot is due to the shorter sampling period.

Conclusions

- Coalescent theory provides a mathematical framework for accurately modelling small genetic samples and the stochastic outcomes of genetic drift due to random mating.
- Coalescent theory provides backward probabilities for population parameters.
- In most implementations the coalescent does not directly model small populations or stochastically fluctuating populations.
- Coalescent theory can be extended to:
  - non-parametric fluctuating populations
  - structured populations
  - gene conversion

BEAST book

theory / practice / programming