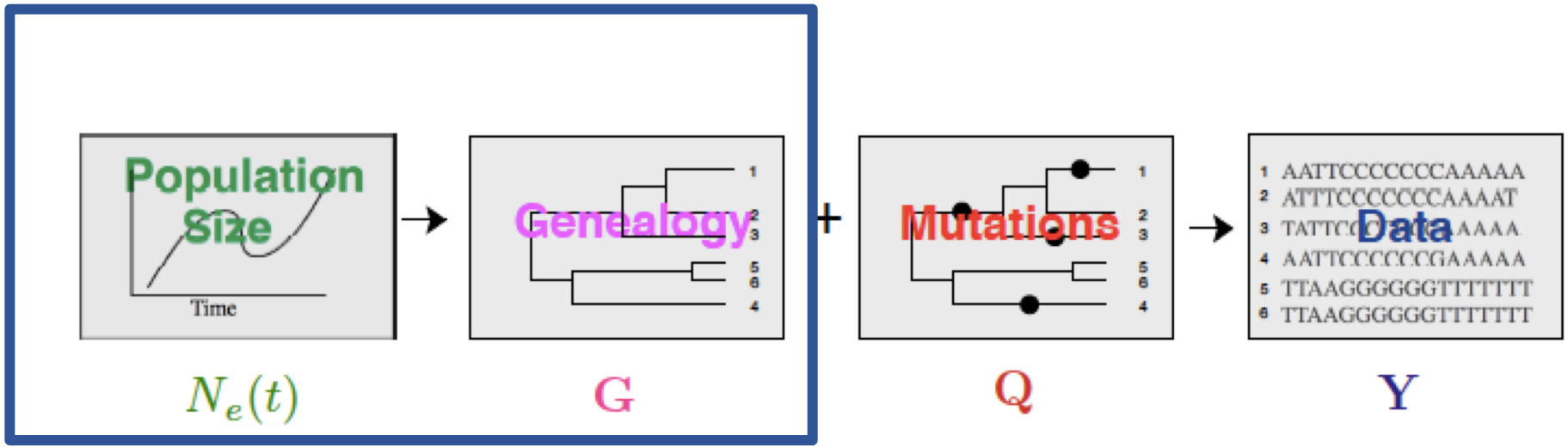
Tree Priors

Instructor: Julia A. Palacios

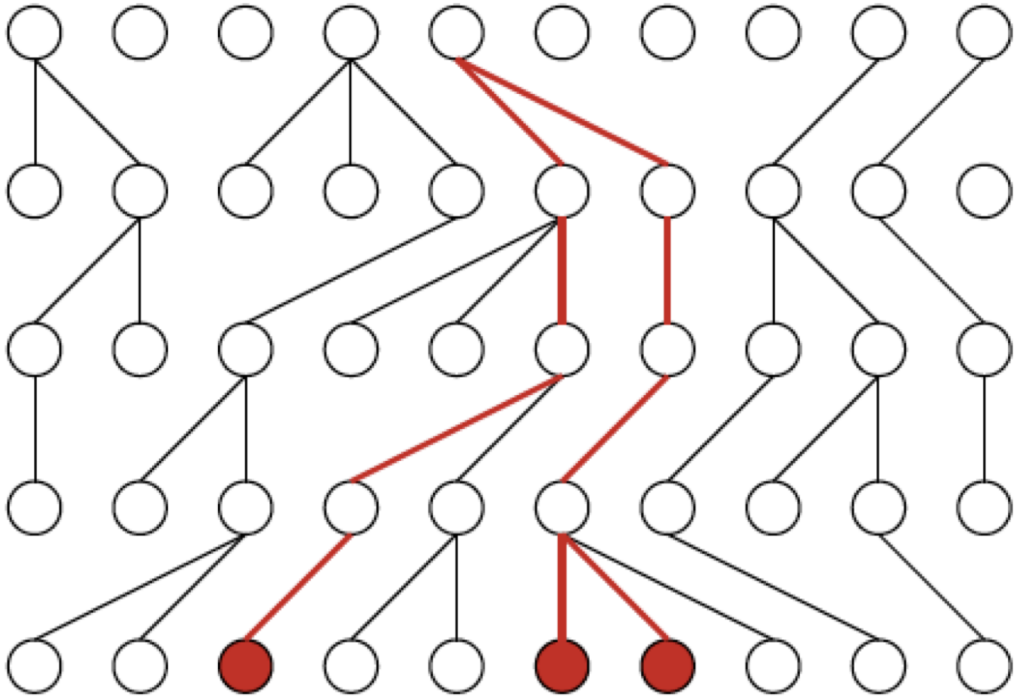
TA: Joëlle Barido-Sottani

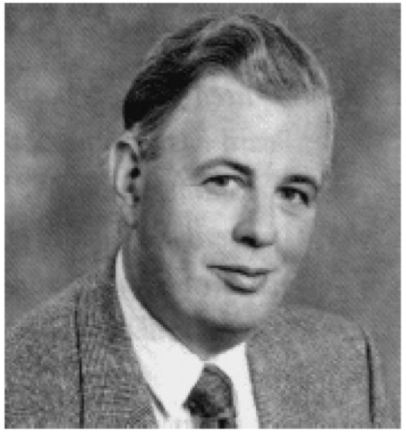
Tree priors



- **Coalescent Models**
- Birth-Death Models

Kingman's coalescent





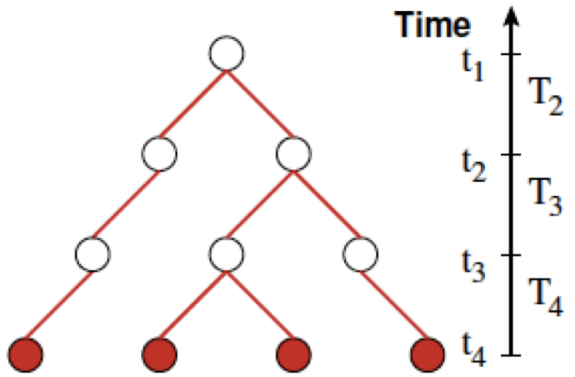
J.F.C. Kingman

- k genes coalesce after j generations w.p

$$\approx \left[1 - \binom{k}{2}\frac{1}{N}\right]^j \to e^{-\binom{k}{2}t}$$

when time is measured in N generation units

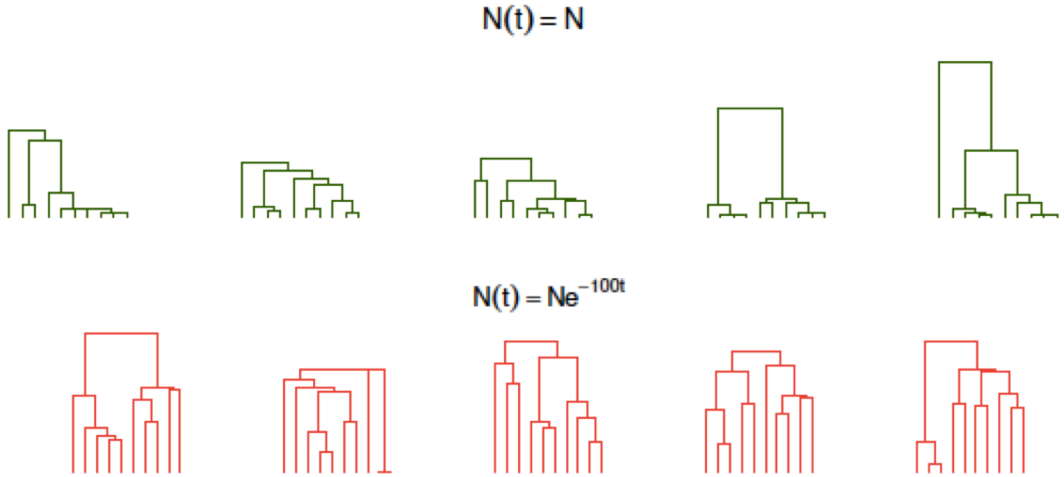
Coalescent with variable population size



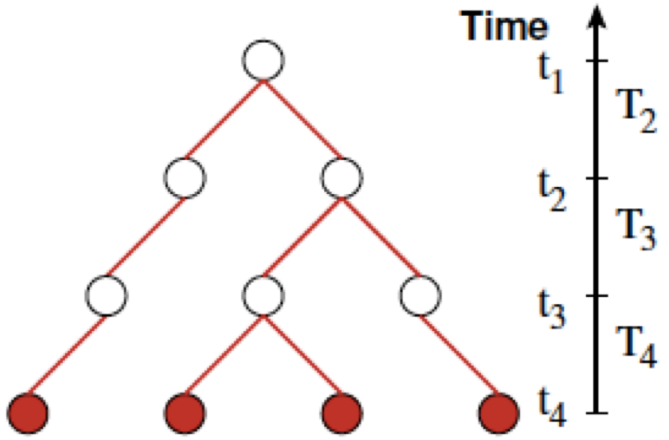
The coalescent with variable population size [Slatkin and Hudson, 1991]:

$$\Pr(T_k > t | t_{k+1}) = e^{-\binom{k}{2} \int_{t_{k+1}}^{t+t_{k+1}} \frac{1}{N_e(u)} du}$$

- Constant population size
- Exponential growth (variable pop. size)



Coalescent with variable population size



- It is a non-homogeneous CTMC.
- It can be viewed as a Markov point process on $[0, \infty)$.

The likelihood of observing the coalescent times:

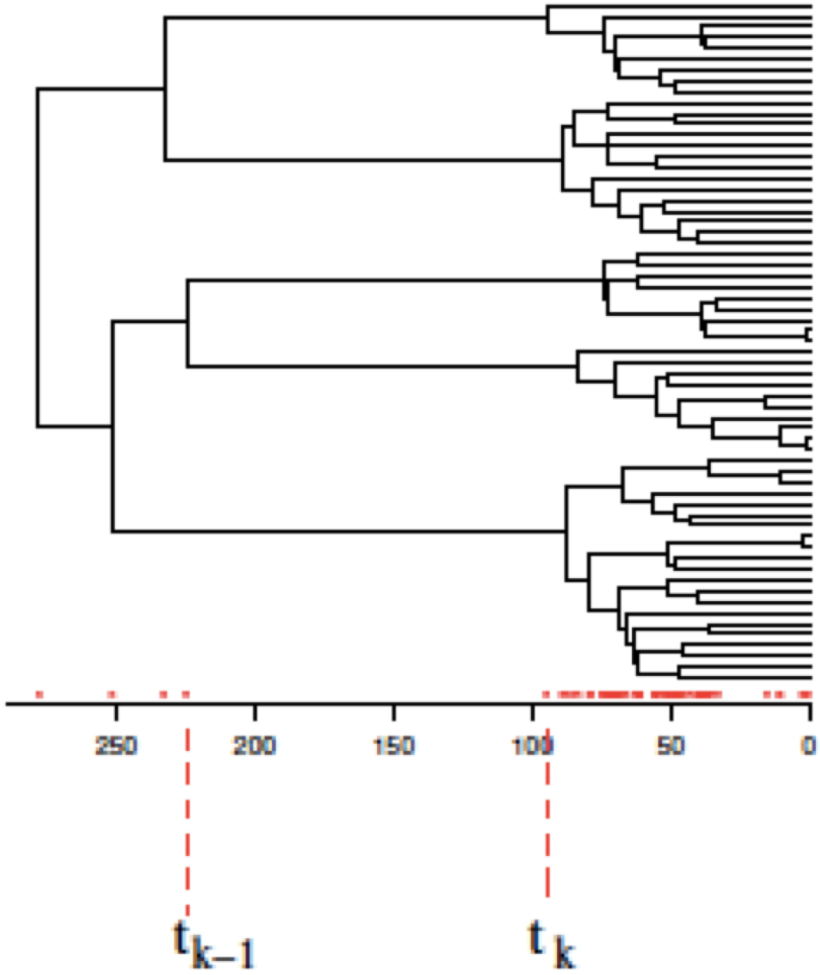
$$P(t_1, \ldots, t_{n-1}) = \prod_{k=1}^{n-1} P(t_{k-1} \mid t_k),$$

where

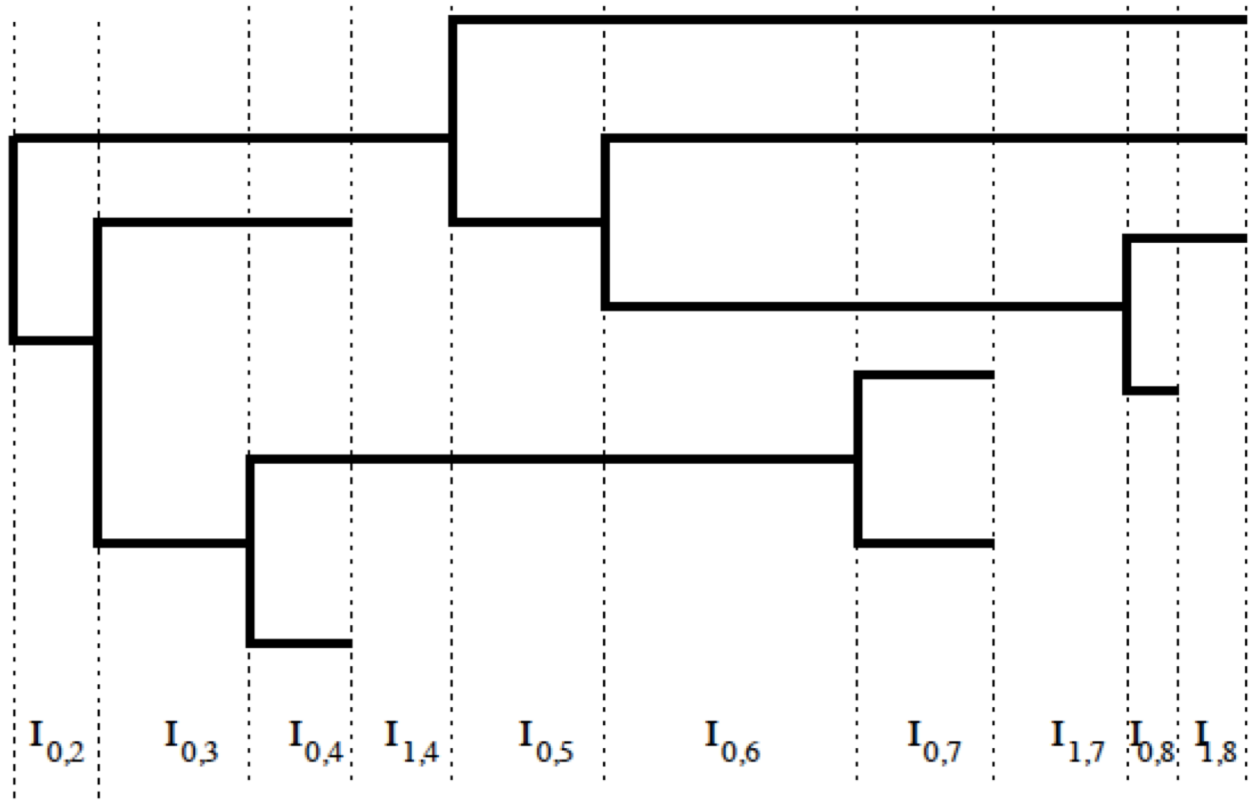
$$P(t_{k-1} \mid t_k, N_e(t)) = \frac{\binom{k}{2}}{N_e(t_{k-1})} \exp - \left\{ \int_{t_k}^{t_{k-1}} \frac{\binom{k}{2} dt}{N_e(t)} \right\}$$

Coalescent with variable population size

Isochronous coalescent:



Heterochronous coalescent



The coalescent as a Markov point process

Isochronous coalescent

$$P[t_{k-1}|t_k, N_e(t)] = \frac{C_k}{N_e(t_{k-1})} exp\left[-\int_{t_k}^{t_{k-1}} \frac{C_k dt}{N_e(t)}\right], \text{where } C_k = \binom{k}{2}.$$

As a point process with conditional intensity:

$$\lambda^*(t|t_k) = \binom{k}{2}\lambda(t)1_{\{t\in(t_k,t_{k-1}]\}}, \text{ for } k = 2, ..., n,$$

Heterochronous coalescent

(Felsenstein and Rodrigo, 1999)

$$P[t_{k-1}|t_k, \mathbf{s}, \mathbf{n}, N_e(t)] = \frac{C_{0,k}}{N_e(t_{k-1})} exp - \left\{\int_{I_{0,k}} \frac{C_{0,k}dt}{N_e(t)} + \sum_{i=1}^{m}\int_{I_{i,k}} \frac{C_{i,k}dt}{N_e(t)}\right\},$$

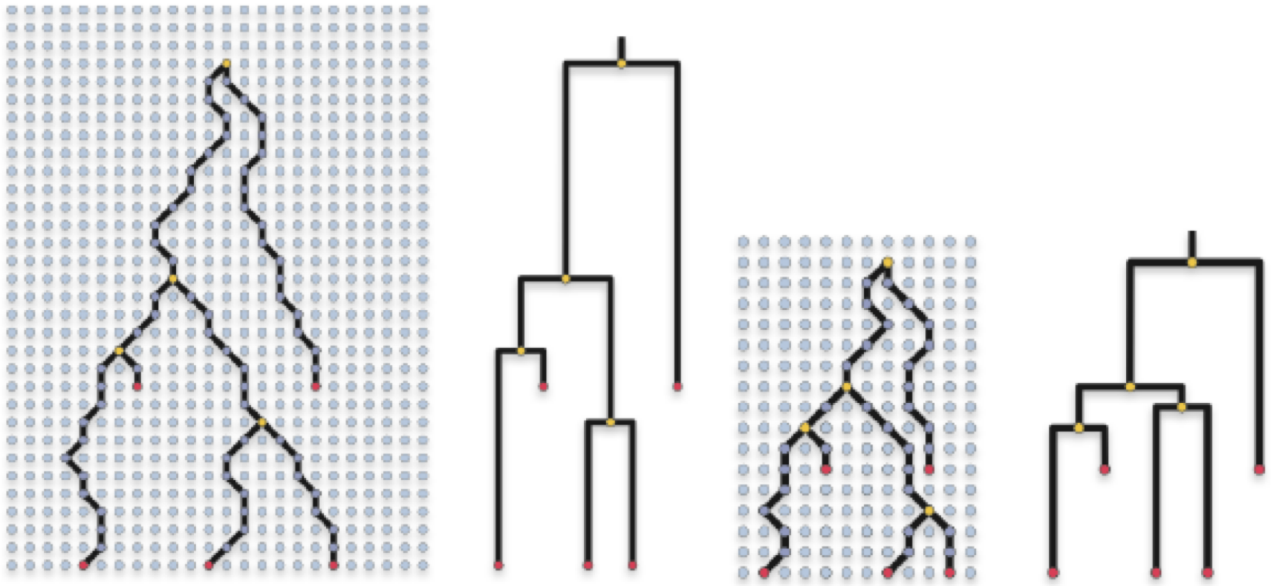
As a point process with conditional intensity:

$$\lambda^*(t|\mathbf{n}, \mathbf{s}, t_k) = \sum_{i=1}^{m}\binom{n_{i,k}}{2}\lambda(t)1_{\{t\in I_{i,k}\}}, \text{ for } k = 2, ...n.$$

$$\lambda(t) = 1/N_e(t).$$

Coalescent effective population size Ne(t)

- The coalescent allows us to calculate $\mathrm{P}(g \mid N_e(t))$
- Or, the inverse problem: learn $N_e(t)$ from g

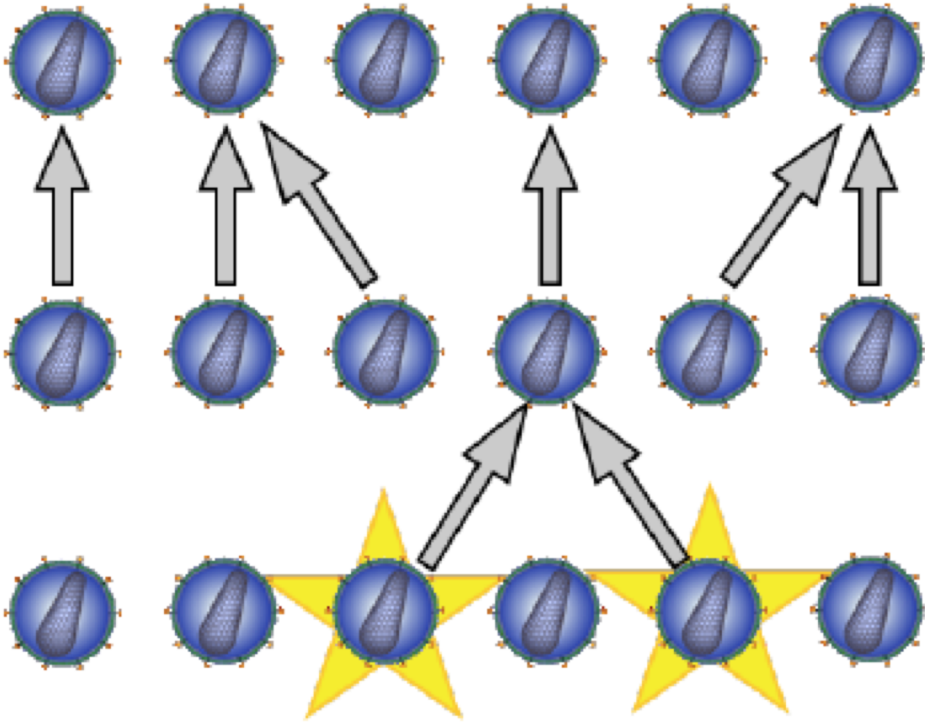


- Large and small population sizes

Coalescent effective population size Ne(t)

- The major weakness of the coalescent lies in its simplifying assumptions

- Neutral evolution?
- Reproductive variance?
- Panmitic population?



But, does this matter?

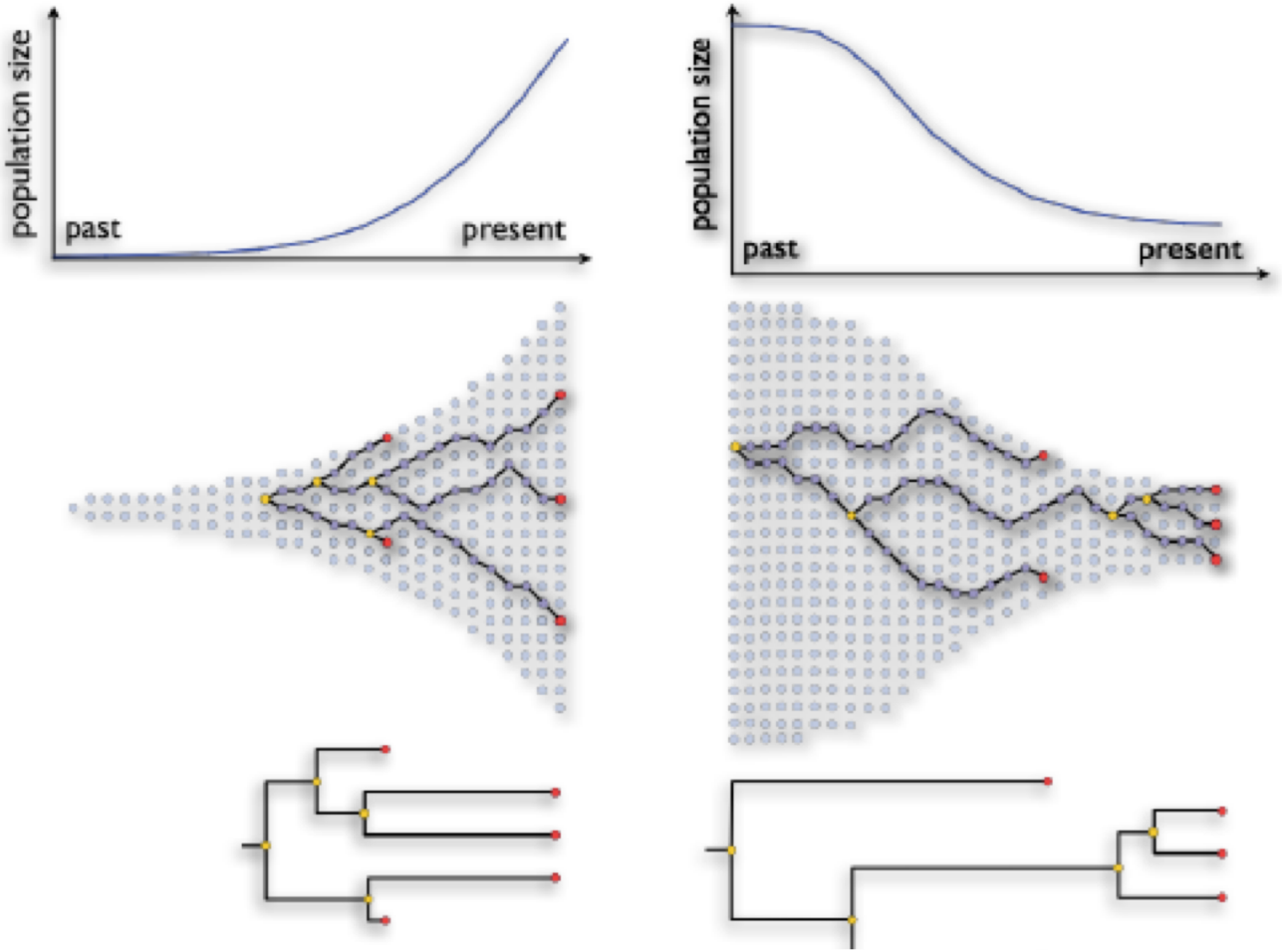
Coalescent effective population size Ne(t)

Effective Population Size Trajectory $Ne(t)$

$N_e(t)$ is a measure of relative genetic diversity over time.

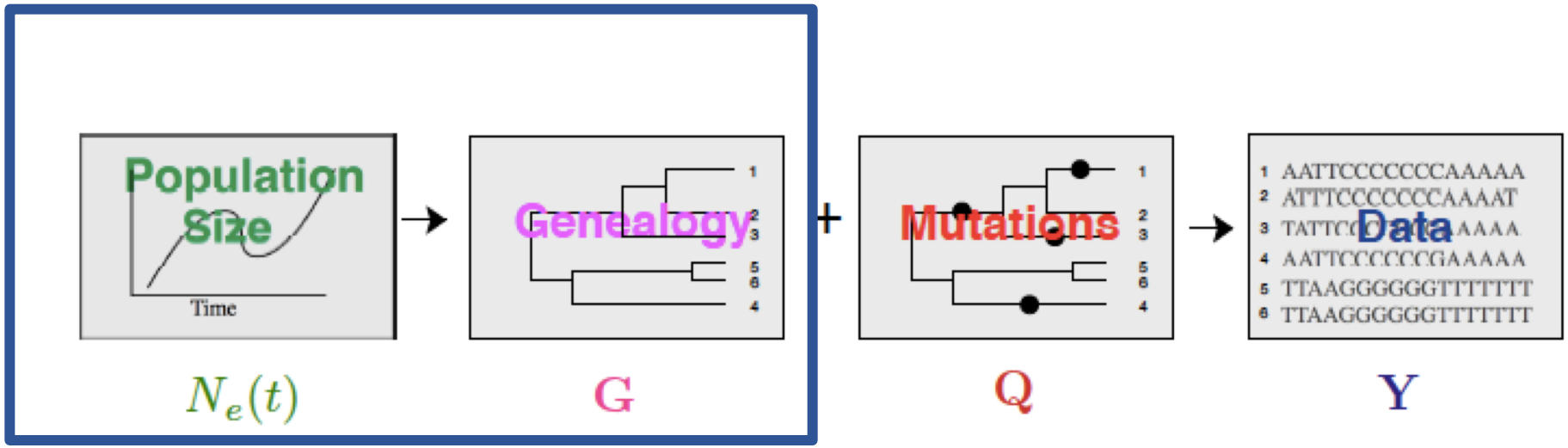
- The N_e of a real biological population is the size of an idealized Wright-Fisher population that loses or gains genetic diversity at the same rate
- N_e is generally smaller than the census population
- The coalescent N_e provides the time-to-ancestry distribution for a sample tree from a real population

Coalescent effective population size Ne(t)



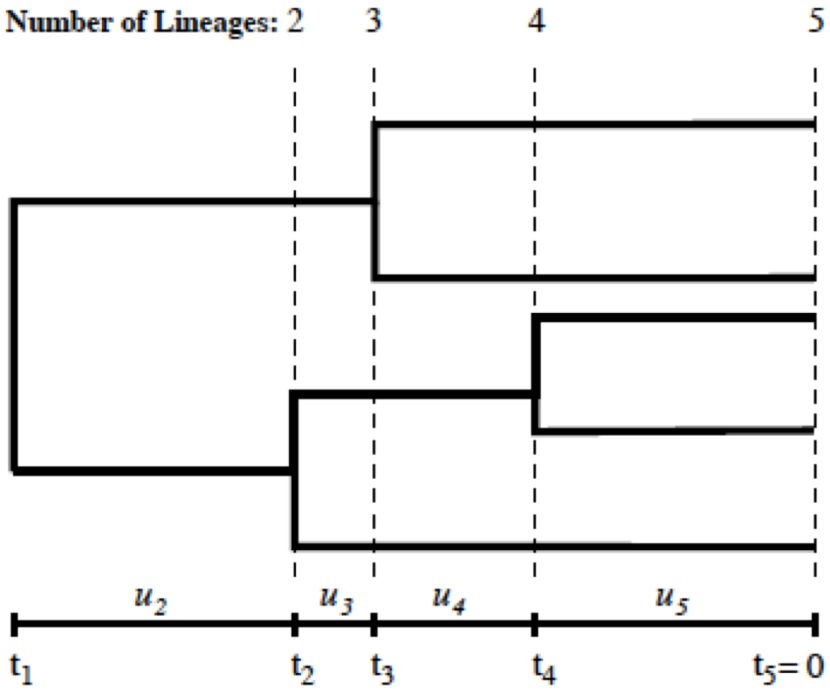
- Changes in N_e reflect changes in the census population size

Tree priors



- Coalescent Models
  - **Priors on Ne(t)**

Ne(t) as a piecewise constant function

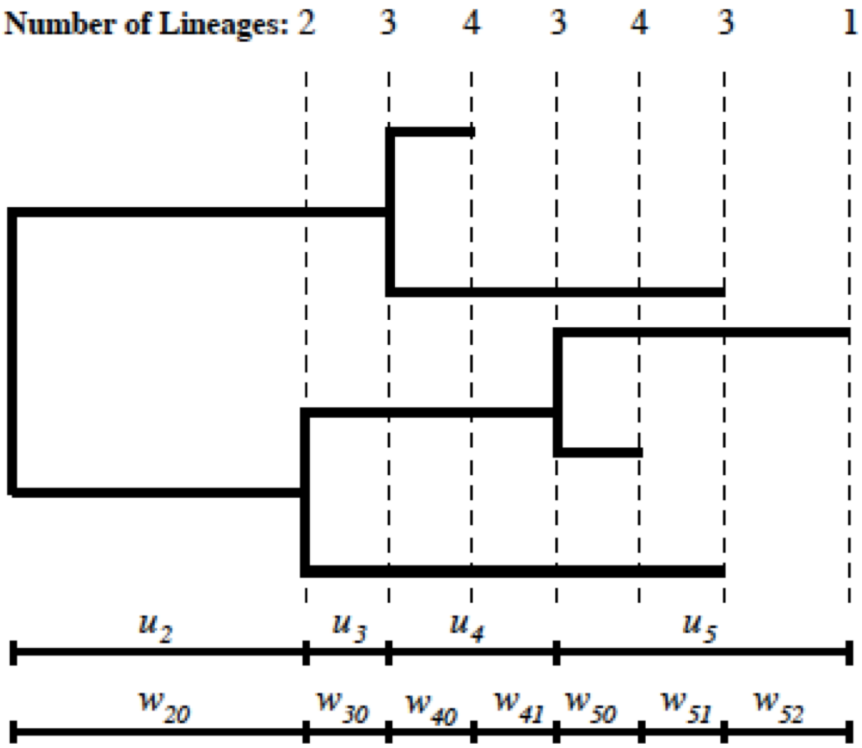


Isochronous Data

- $N_e(t) = \theta_k$ for $t_k < t \leq t_{k-1}$.
- $u_2, \ldots, u_n$ are independent
- $\Pr(u_k \mid \theta_k) = \frac{k(k-1)}{2\theta_k} e^{-\frac{k(k-1)u_k}{2\theta_k}}$
- $\Pr(\mathbf{F} \mid \theta) \propto \prod_{k=2}^{n} \Pr(u_k \mid \theta_k)$

- Equivalent to estimating exponential mean from one observation.
- Need further restrictions to estimate all effective pop sizes θ!

Ne(t) as a piecewise constant function



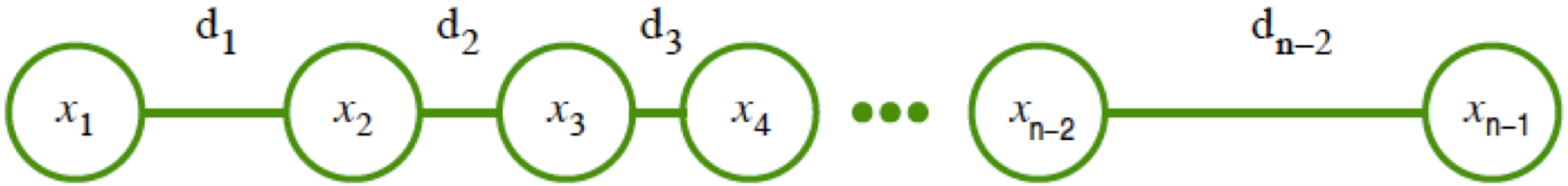
Heterochronous Data

- $w_{20}, \ldots, w_{nj_n}$ are independent
- $\Pr(w_{k0} \mid \theta_k) = \frac{n_{k0}(n_{k0}-1)}{2\theta_k} e^{-\frac{n_{k0}(n_{k0}-1)w_{k0}}{2\theta_k}}$
- $\Pr(w_{kj} \mid \theta_k) = e^{-\frac{n_{kj}(n_{kj}-1)w_{kj}}{2\theta_k}}, j>0$
- $\Pr(\mathbf{F} \mid \boldsymbol{\theta}) \propto \prod_{k=2}^{n} \prod_{j=0}^{j_k} \Pr(w_{kj} \mid \theta_k)$

- Equivalent to estimating exponential mean from one observation.
- Need further restrictions to estimate θ!

Smooth GMRF prior on log Ne(t)

- Go to the log scale $x_k = \log \theta_k$
- $\Pr(\mathbf{x} \mid \omega) \propto \omega^{(n-2)/2} \exp\left[-\frac{\omega}{2}\sum_{k=1}^{n-2}\frac{1}{d_k}(x_{k+1}-x_k)^2\right]$



Weighting Schemes

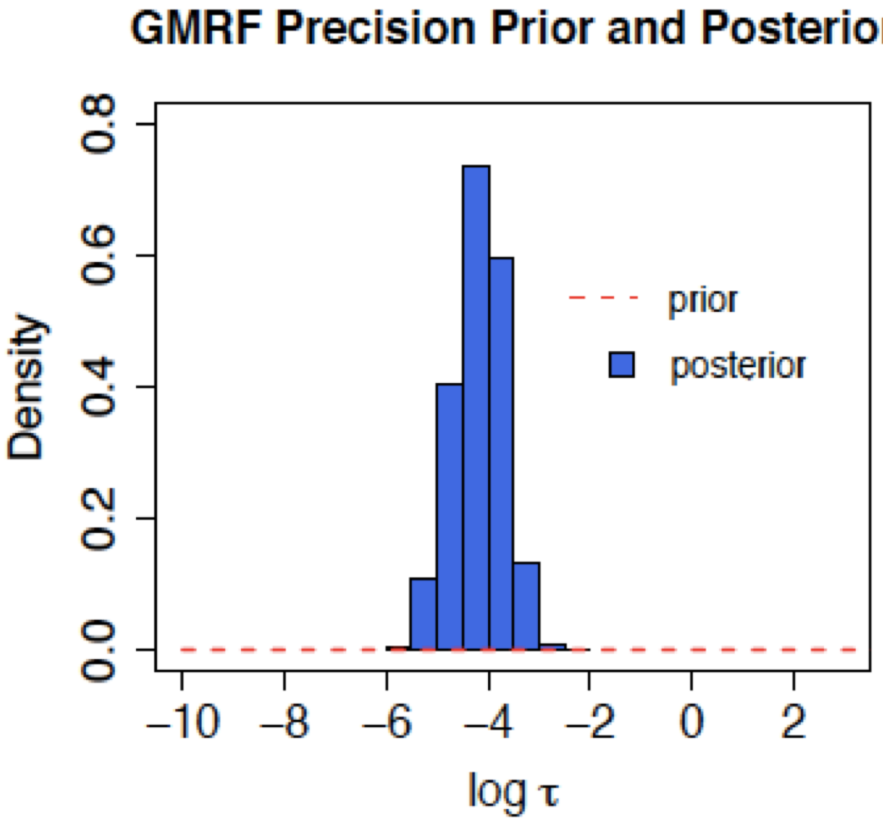
1. Uniform: $d_k = 1$
2. Time-Aware: $d_k = \frac{u_{k+1}+u_k}{2}$

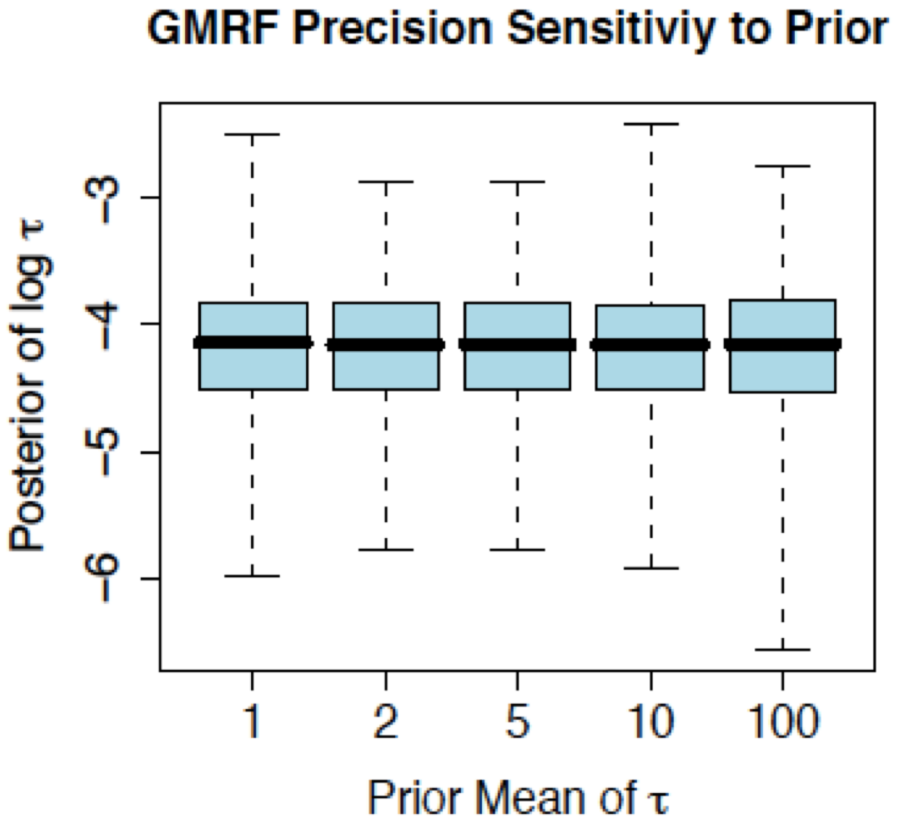
- $\Pr(\mathbf{x}, \omega) = \Pr(\mathbf{x} \mid \omega)\Pr(\omega)$
- $\Pr(\omega) \propto \omega^{\alpha-1}e^{-\beta\omega}$, diffuse prior with $\alpha = 0.01$, $\beta = 0.01$

[Phylodyn: An R package for phylodynamic simulation and inference, Karcher et al., MER 2017]

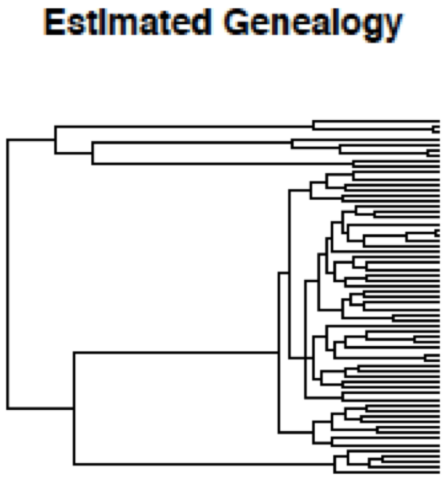
Precision parameter of GMRF

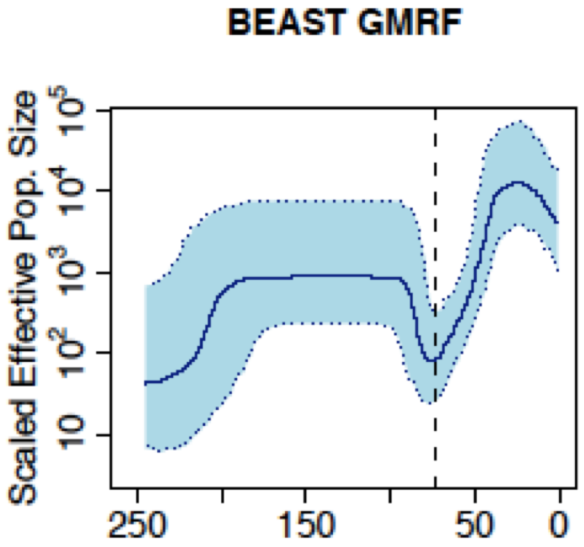
- ω - GMRF precision, controls smoothness
- Usually $\Pr(\omega \mid \mathbf{D})$ is sensitive to perturbations of $\Pr(\omega)$
- Not in our Coalescent model!

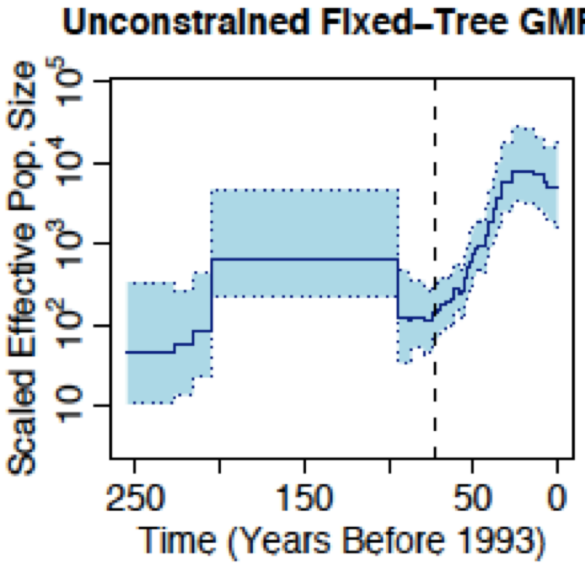


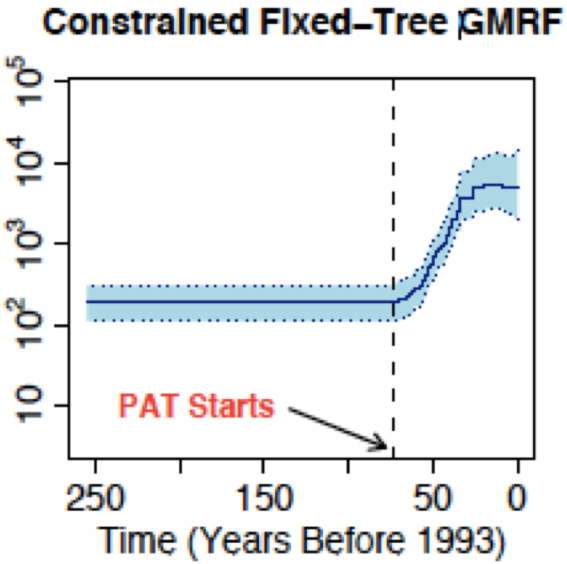


Hepatitis C virus in Egypt



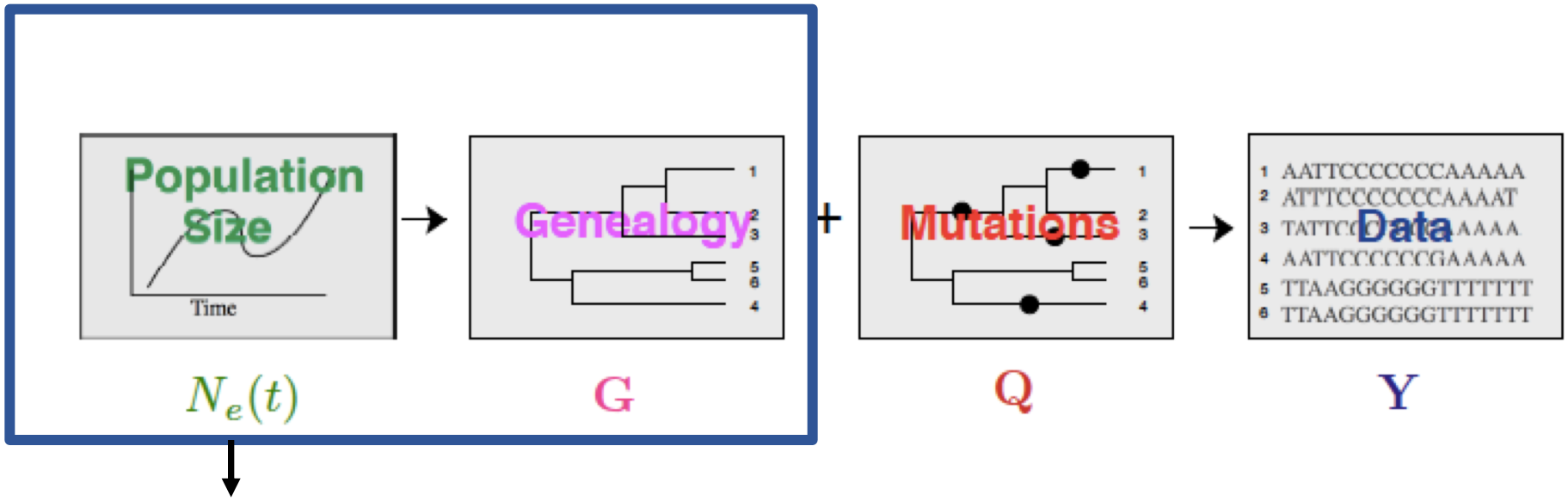






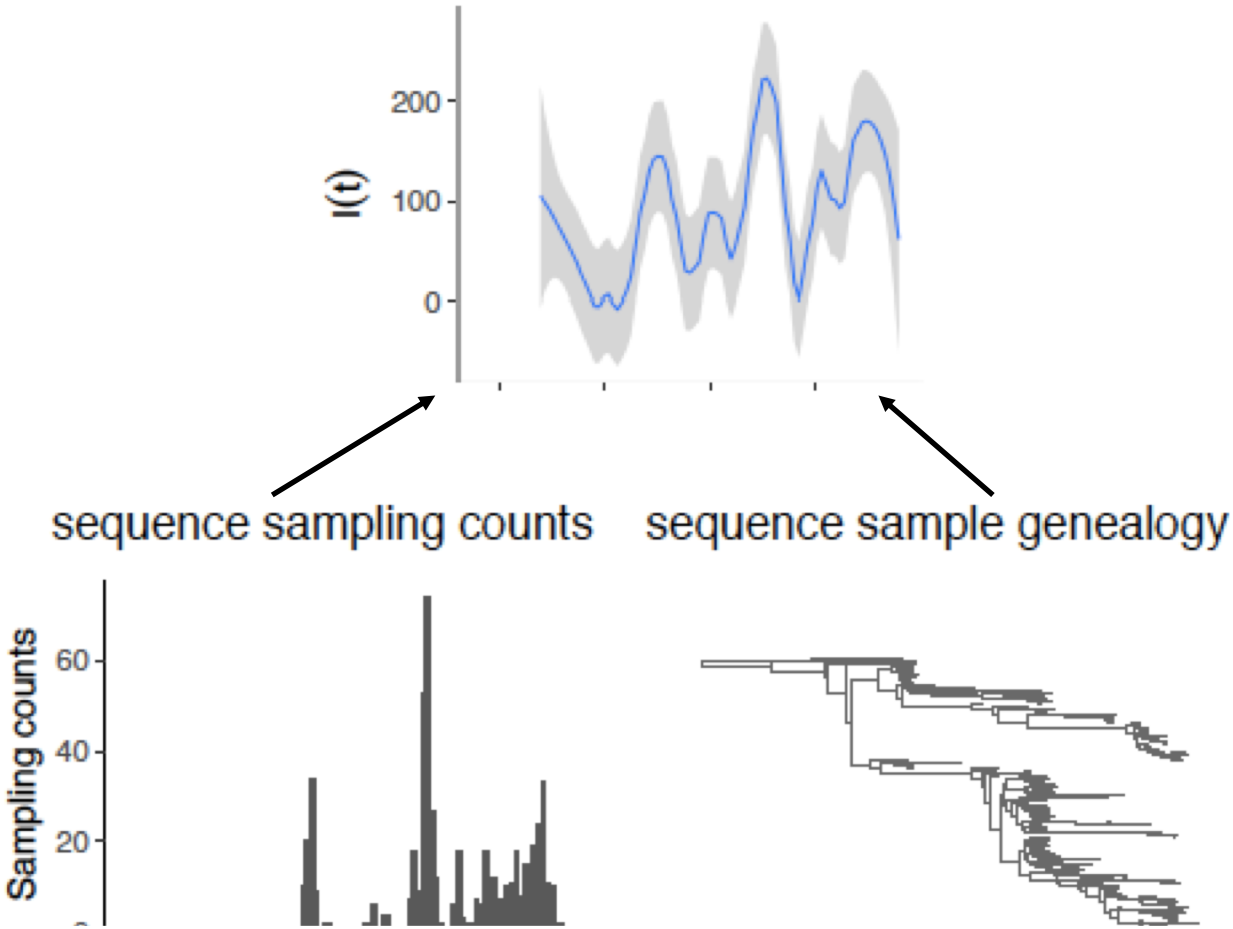
- Random population sample
- No sign of population sub-structure
- Parenteral antischistosomal therapy (PAT) was practiced from 1920s to 1980s
- Bayes Factor 12,880 in favor of constant population size prior to 1920

Tree priors



- Coalescent Models
  - Priors on Ne(t)
  - **Prior on Ne(t) with Preferential sampling**

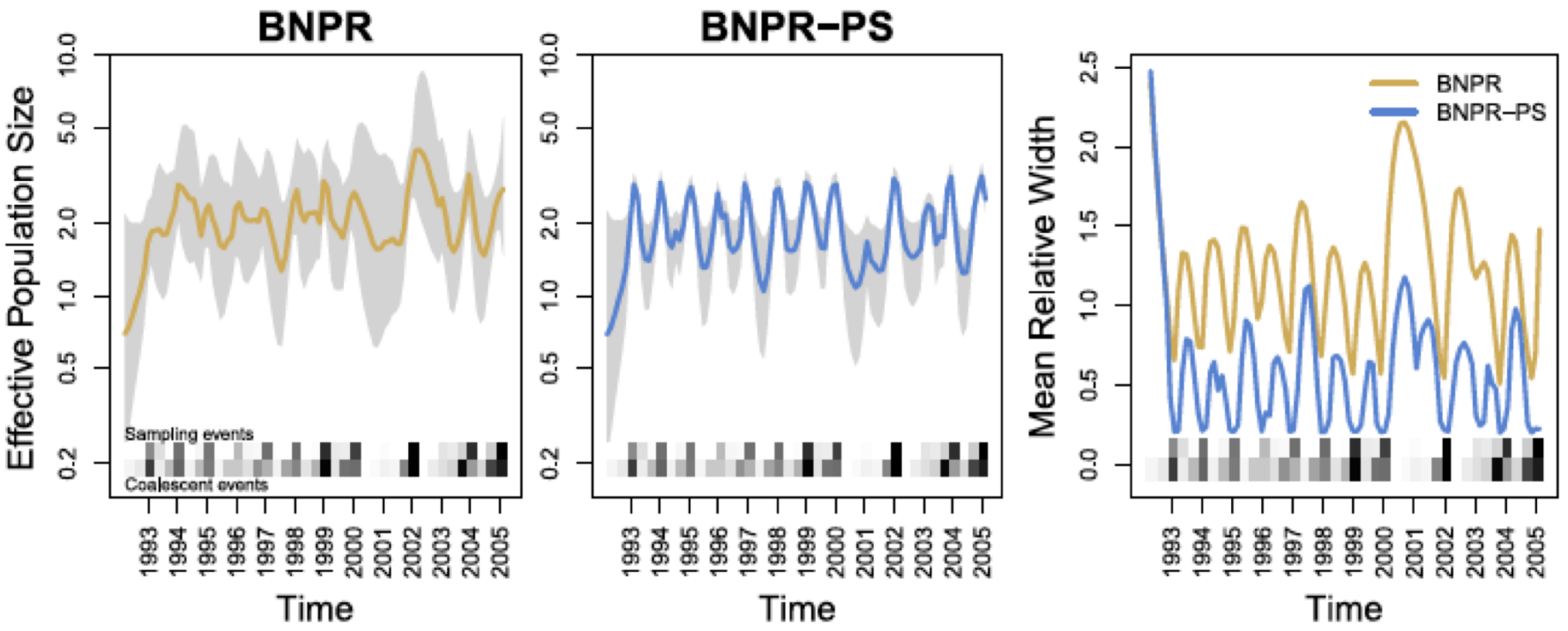
Preferential Sampling



Sampling log-likelihood:

$$\log[\Pr(\mathbf{s} \mid N_e(t), \beta_0, \beta_1] = C + n\beta_0 + \sum_{i=1}^{n} \beta_1 \log[N(s_i)] - \int_{s_0}^{s_m} \exp(\beta_0)[N_e(t)]^{\beta_1} dt \quad (1)$$

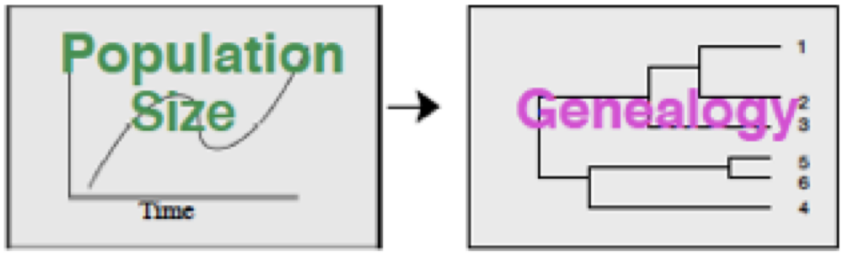
Preferential Sampling



$n = 709$ hemagglutinin gene sequences of H3N2 human influenza type A

Karcher et al., *PLOS Computational Biology* 2016

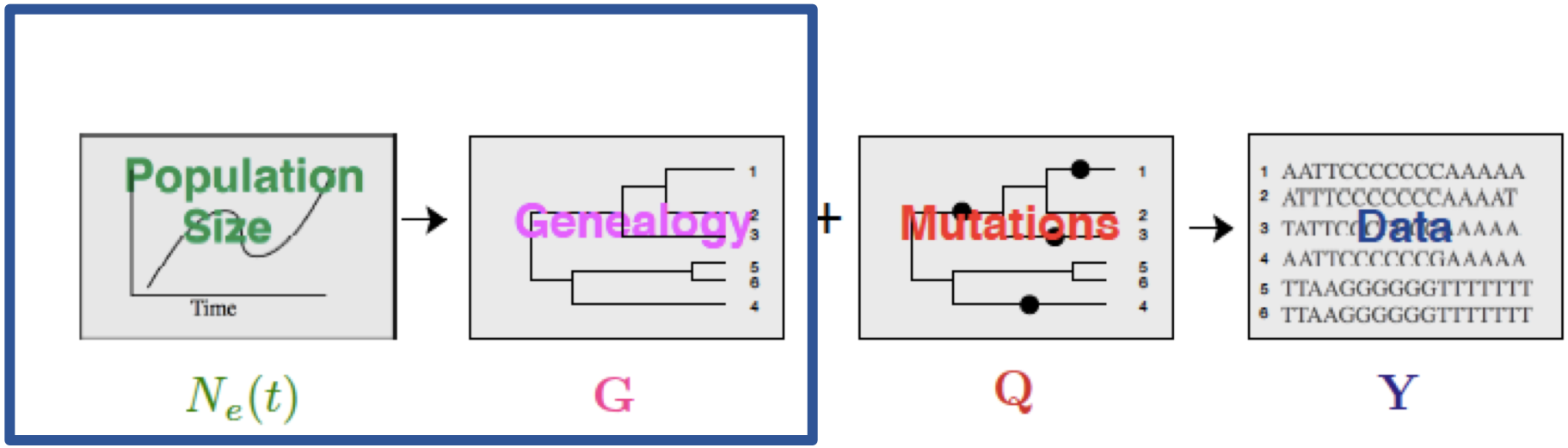
Genealogy branch lengths depend on population dynamics (neutral theory)



Assumptions:

- Evolution is neutral – tree topology and branch lengths are independent.
- We have a random sample from a large population.
- No population structure

Tree priors



- Coalescent Models
  - **Extension with SIR**
- **Birth-Death Models**

Structured coalescent SIR

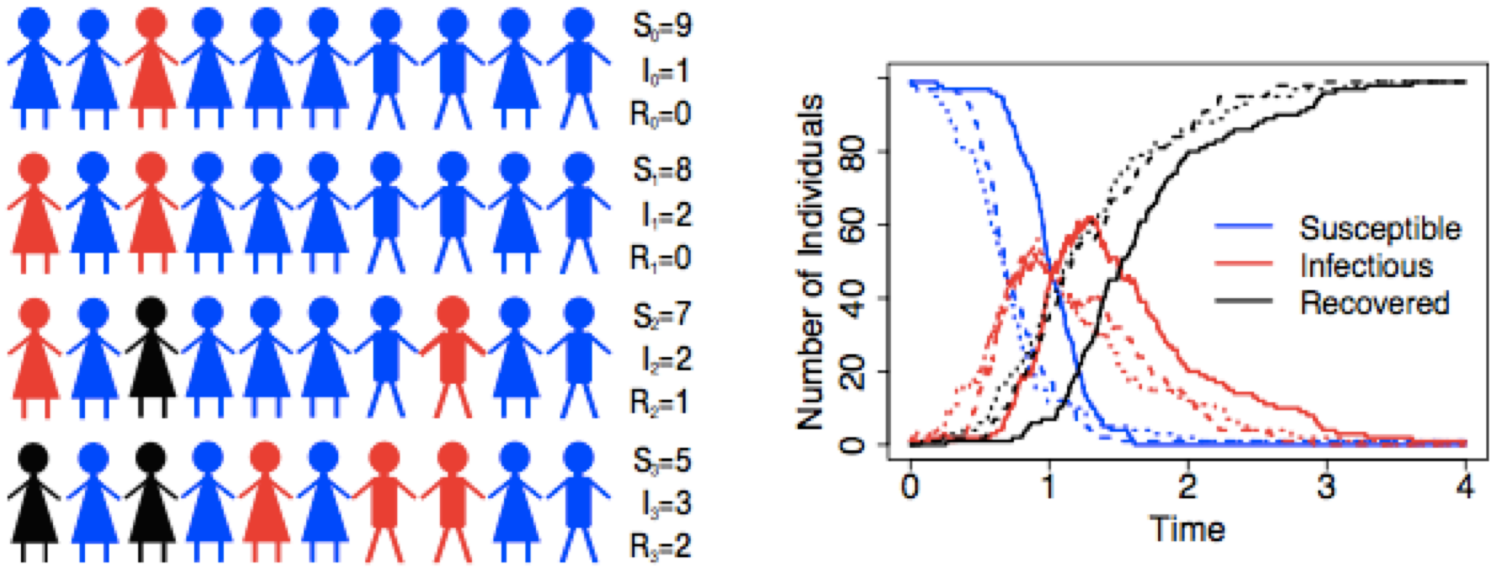
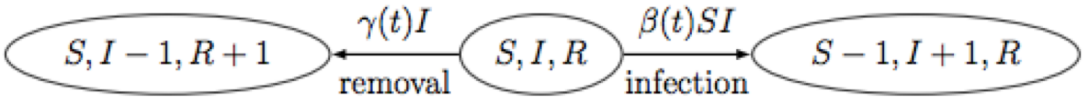


FIGURE 1.2: Susceptible-infectious-recovered (SIR) stochastic epidemic model. The left plot shows 4 times points during an epidemic in the population with 10 individuals. Susceptible individuals are shown in blue, infectious individuals are shown in red, and recovered individuals are shown in black. The right plot shows three realizations of the SIR epidemic model.

Figure credit: Vladimir Minin

Structured coalescent SIR

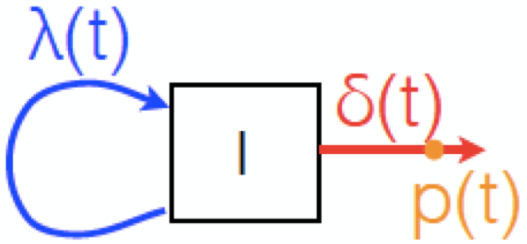


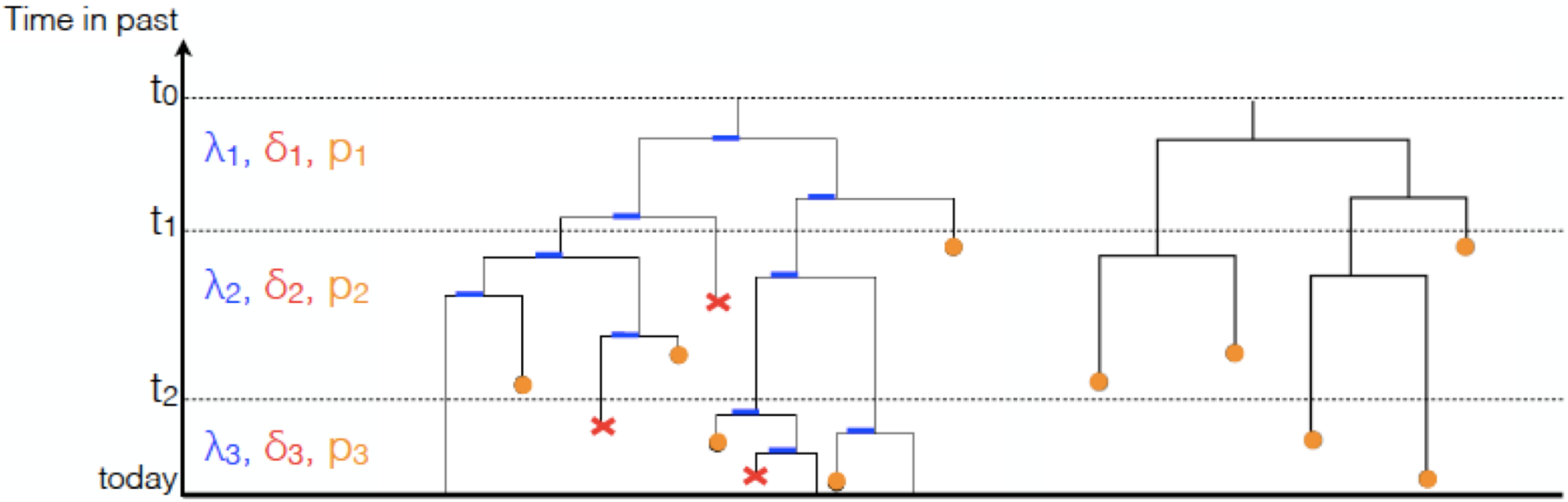
$$N_e(t) = \frac{I(t)}{2\beta(t)S(t)}$$

Volz et al., *Genetics* 2009

Birth-death stochastic model

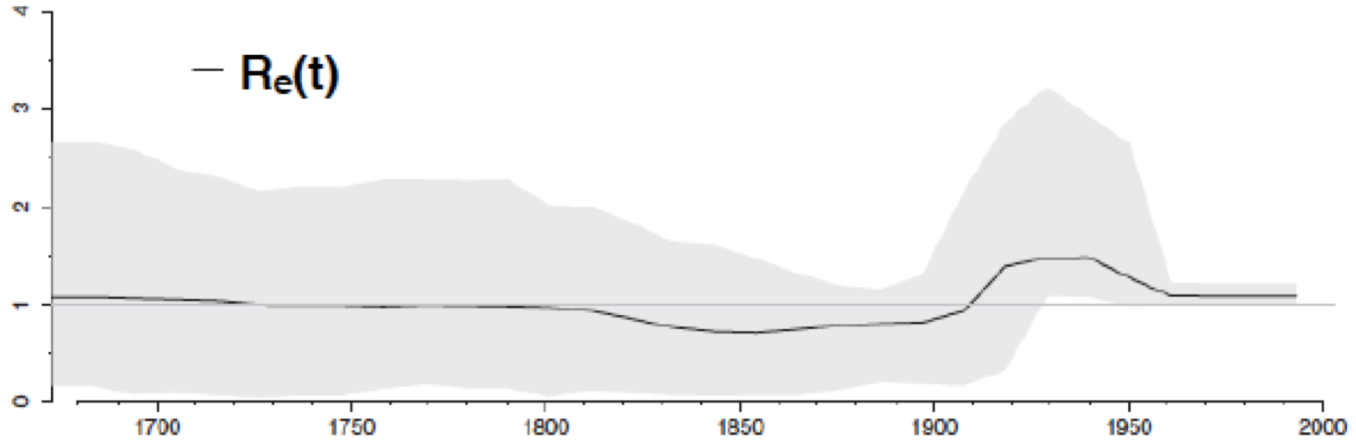
Epidemiological rates may change through time:





Hepatitis C virus in Egypt

Birth-death skyline plot: effective reproductive number



Coalescent skyline plot: effective population size