Viral epidemiology and the coalescent

Marc A. Suchard

msuchard@ucla.edu

UCLA

Key to population genetics

Key to population genetics

- Population size
- Structure
- Migration
- Selection

Tree-based population genetics

The coalescent . . .

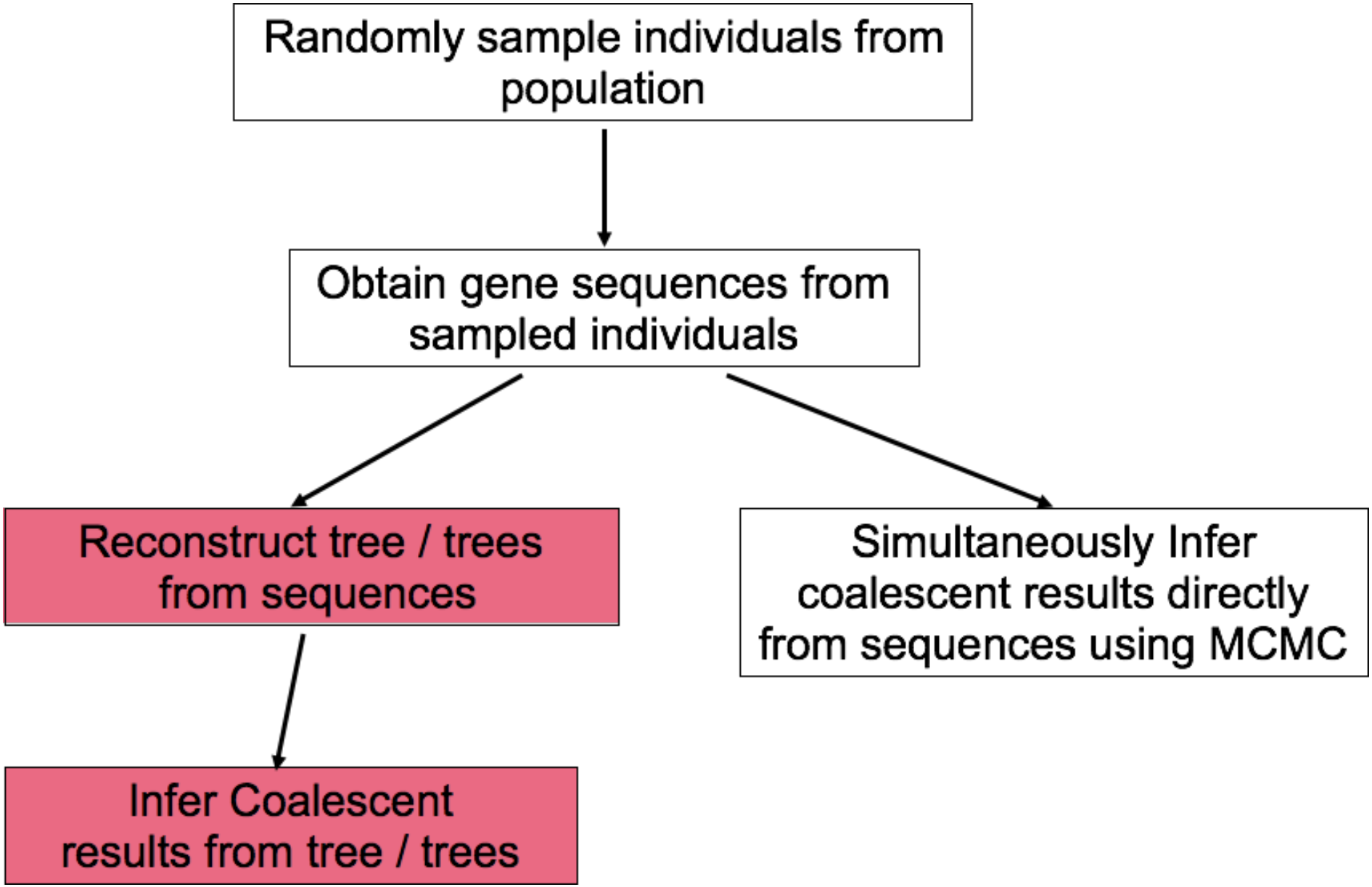
- Models the **ancestral relationships** of a random sample of individuals taken from a large background population.
- Describes a **probability distribution** on ancestral trees given a population history
- Covers ancestral trees, not sequences, and its simplest from **assumes neutral evolution**.

Population history inference

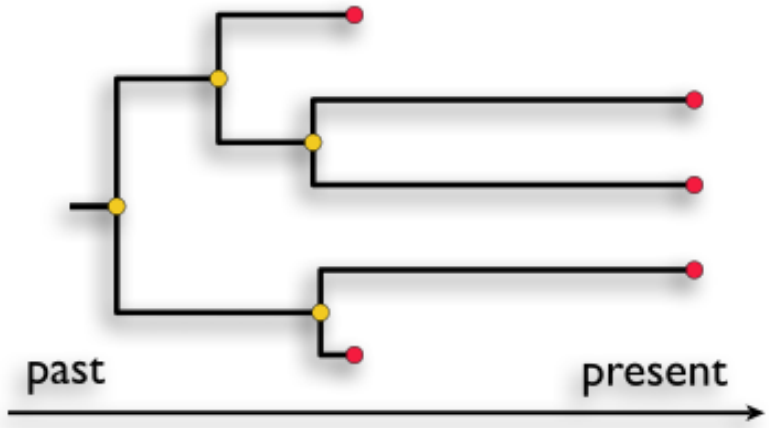
A population history → often called a demographic

- Inference: learn about changes in population size through time
- Applications include:
  - Reconstructing infection disease epidemics
  - Investigating viral dynamics within hosts
  - Using viral sequences as genetic markers for their hosts and host demographics
  - Identifying population bottlenecks caused by:
    - ★ Changes in climate/environment? aridifcation, ice ages
    - ★ Competition with other species? humans
    - ★ Transmission between hosts in viruses

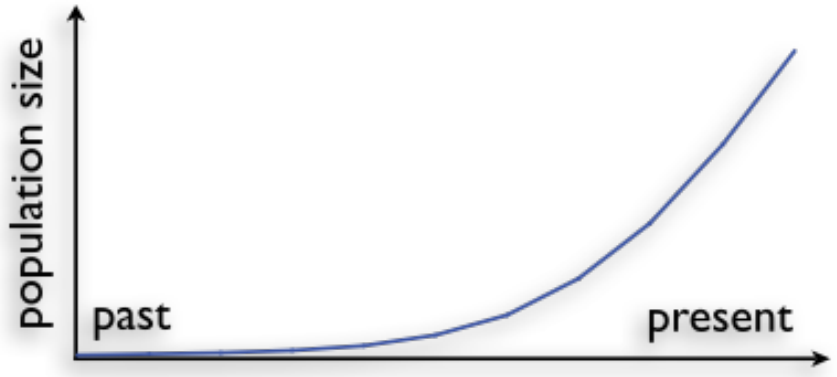
Information pipe-line



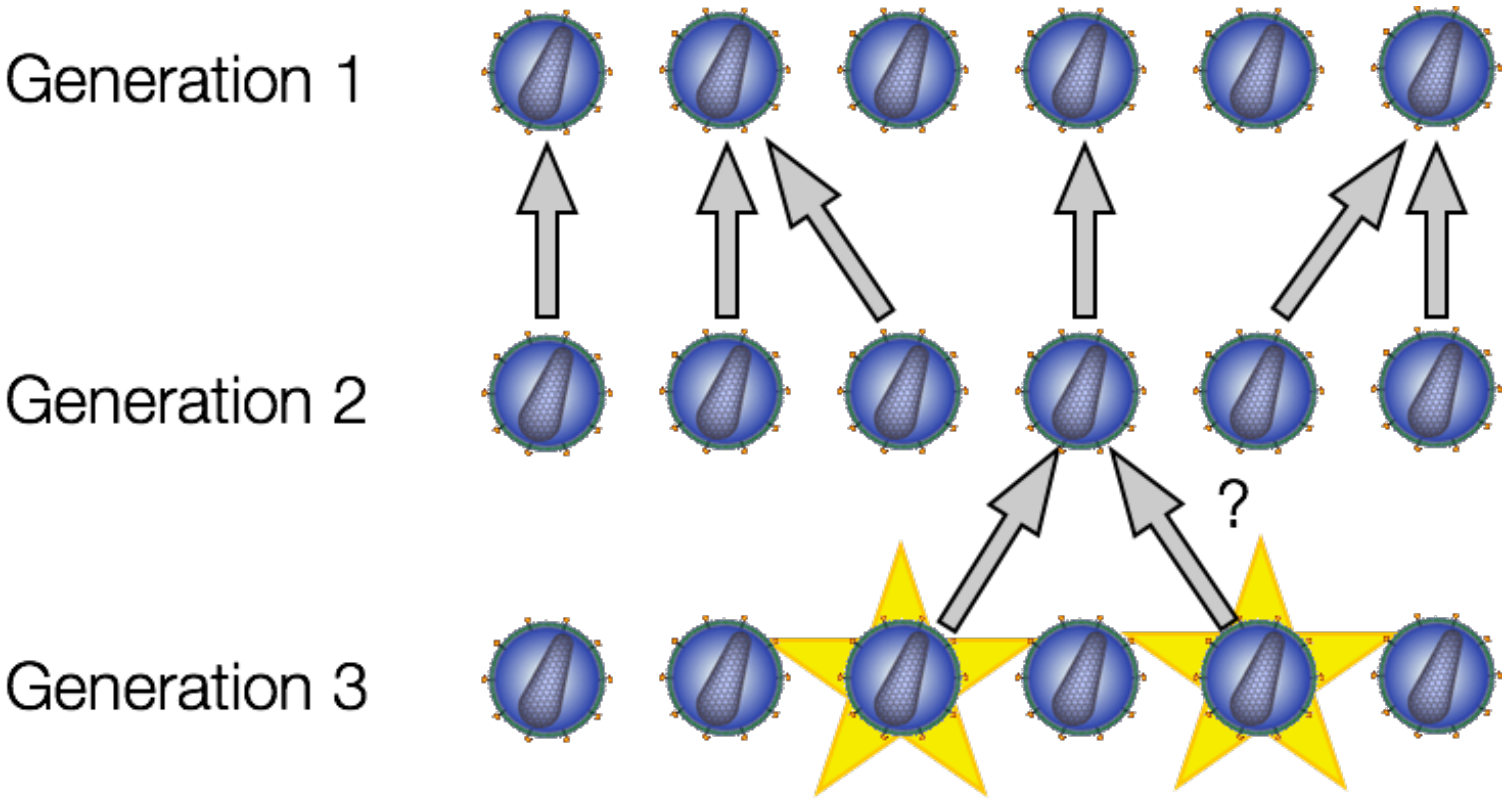
Coalescent inference



Coalescent theory

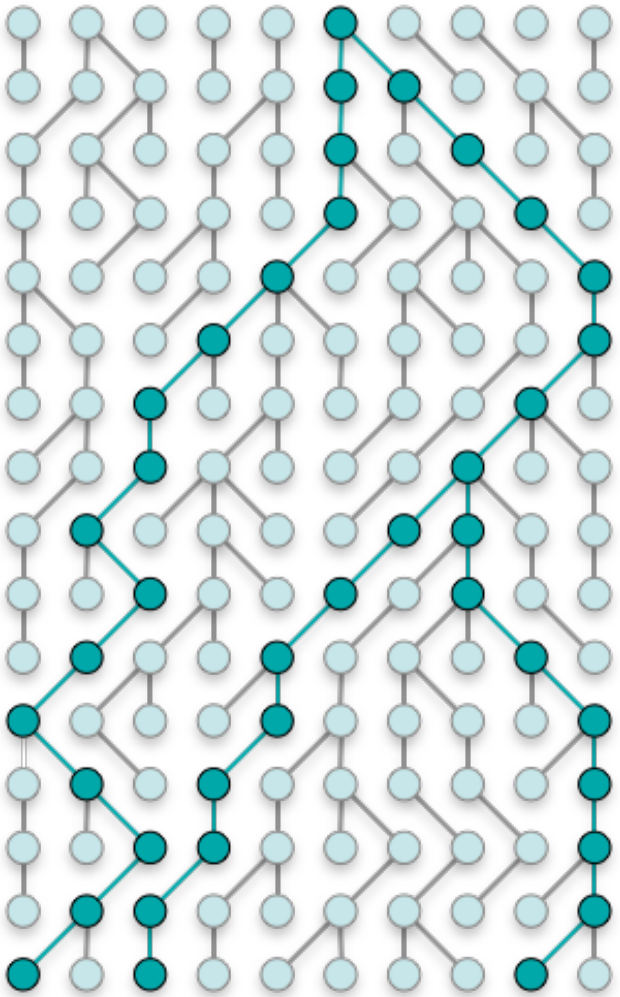


Simple model of reproduction

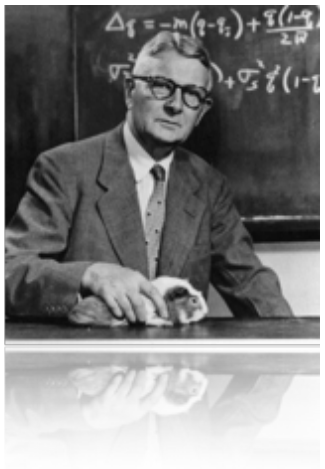


For a randomly chosen pair of individuals, they share a common ancestor (coalesce) in the previous generation with probability $1/N$.

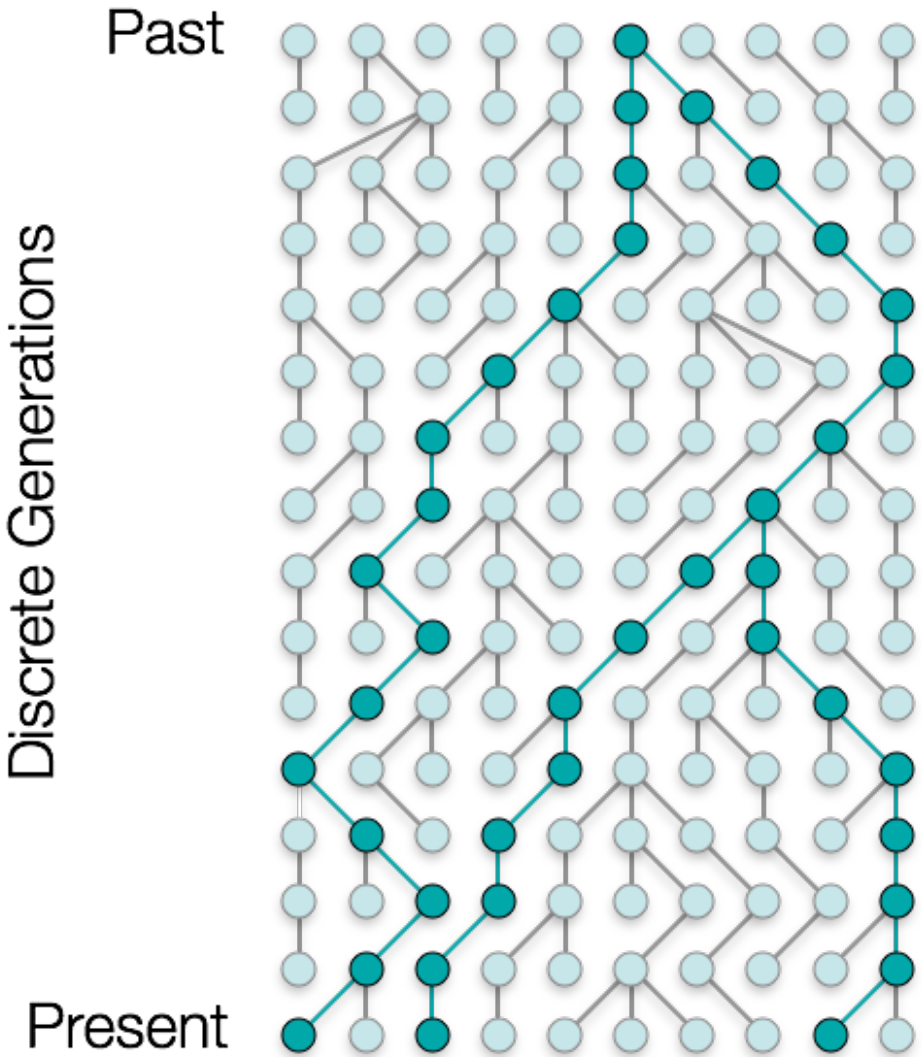
Wright-Fisher reproduction model



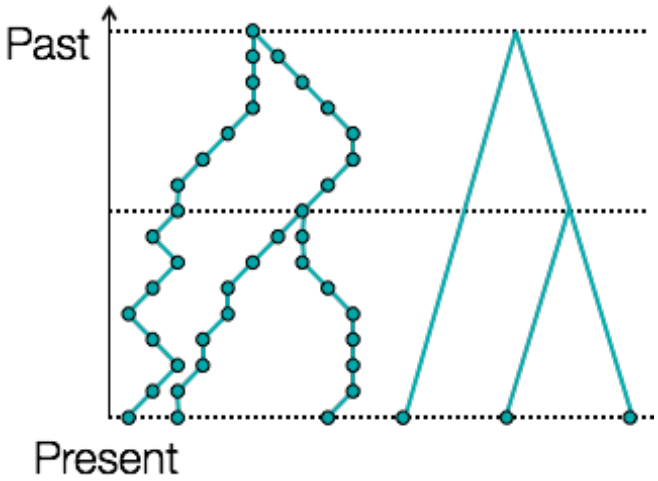
- A constant population size of N individuals (usually $2N$)
- Each new (non-overlapping) generation “chooses" its parents from the previous generation at random with replacement
- No geographic/social structure, no recombination, no selection



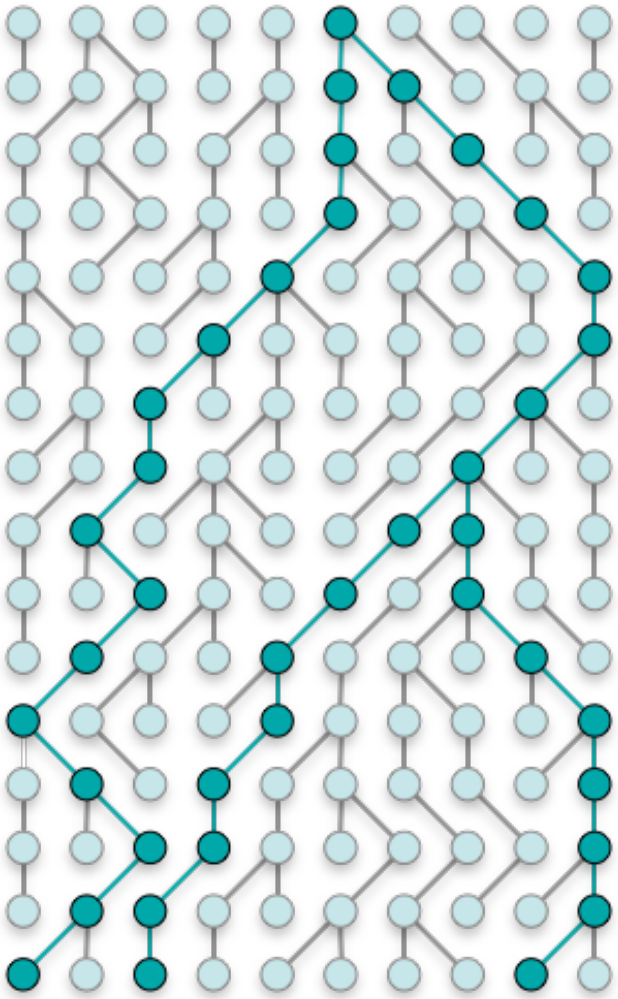
Sample tree from a Wright-Fisher population



- A sample tree of 3 sequences from a population of $N = 10$



Kingman discrete-time coalescent



- 2 individuals coalesce in 1 generation w.p. $\frac{1}{N}$
- 2 individuals coalesce in j generations w.p.

$$\frac{1}{N}\left(1-\frac{1}{N}\right)^{j-1}$$

- k individuals coalesce in j generations w.p.

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

Kingman continuous-time coalescent

Kingman (1982) *J Appl Prob* 19, 27-42

Kingman (1982) *Stoch Proc Appl* 13, 235-48

- Let $t \sim j$ define a rescaled time in the past, and
- Assume a sample of n individuals with $n \ll N$
- Then, the waiting time for k individuals to have $k-1$ ancestors

$$\mathbb{P}(u_k \leq t) = 1 - e^{-\binom{k}{2}\frac{t}{N}}$$

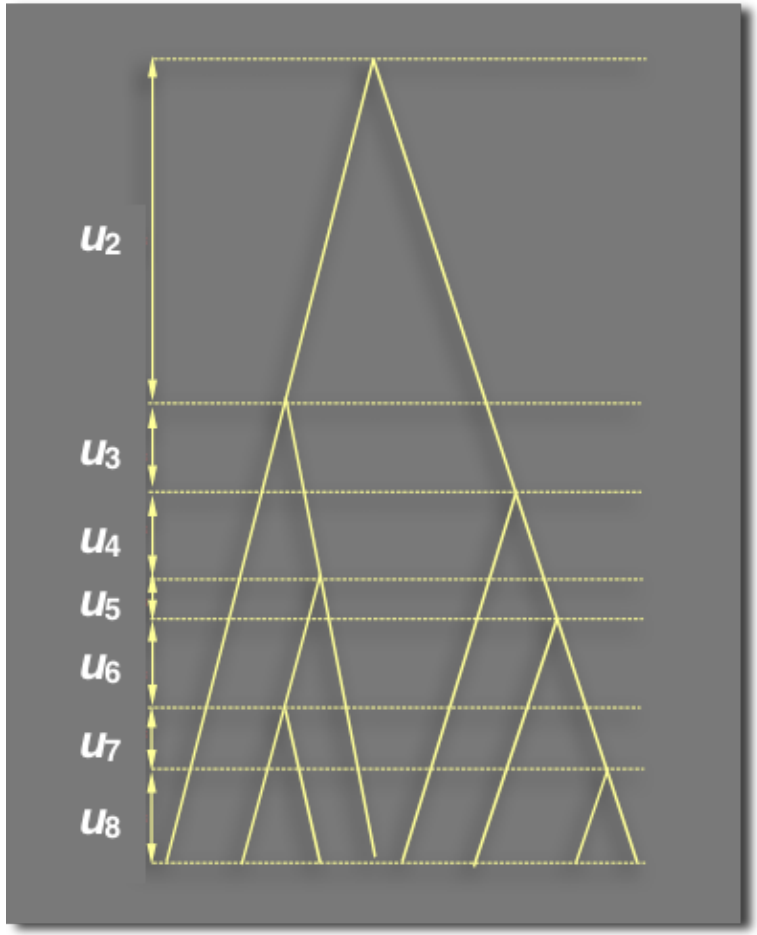
- Exponential (memoryless) $\rightarrow$ defines a continuous-time Markov chain

$$\mathbb{E}(u_k) = \frac{2N}{k(k-1)}$$

Kingman coalescent: CTMC

- The number of ancestral lineages decreases by one at each coalescence
- The process continues until the most recent common ancestor (MRCA) is reached
- What is the expected time to MRCA?

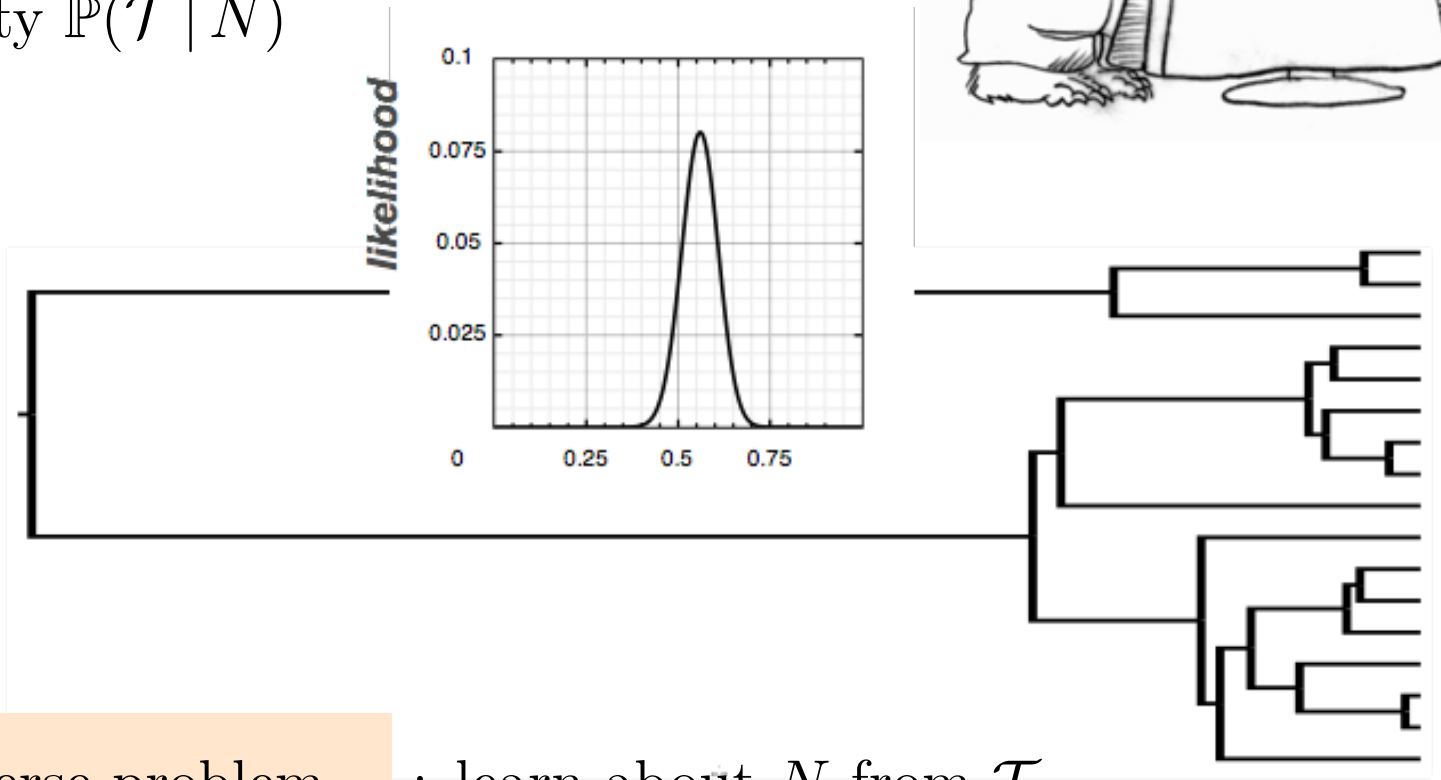
$$\begin{aligned}\mathbb{E}\left(\sum_{k=2}^{n} u_k\right) &= \sum_{k=2}^{n} \mathbb{E}(u_k) \\ &= \sum_{k=2}^{n} \frac{2N}{k(k-1)} \\ &= 2N\left(1 - \frac{1}{n}\right)\end{aligned}$$



- Note: $\text{tMRCA}/2 < \mathbb{E}(u_2)$

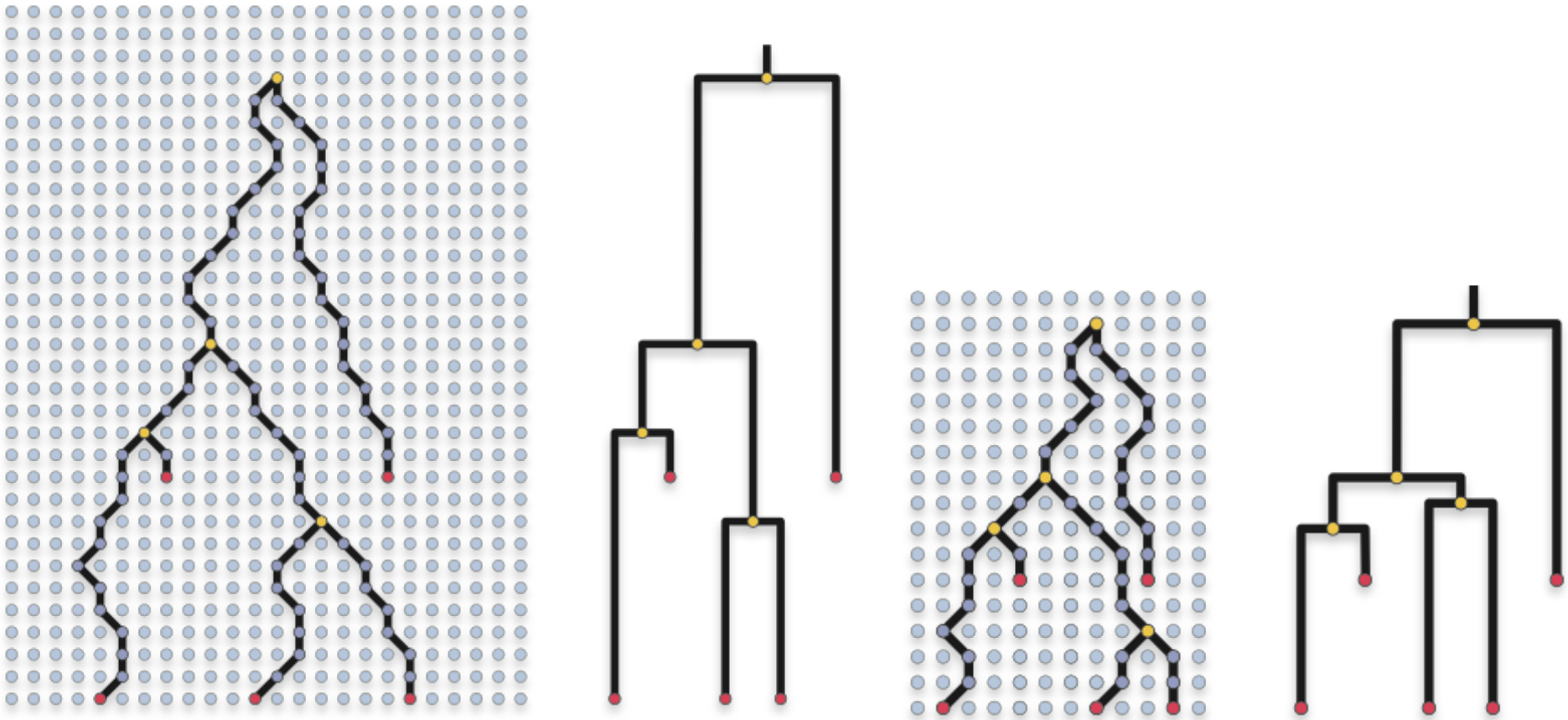
Kingman coalescent: probability distribution

- Given a known tree $\mathcal{T}$ from a sample of individuals from a population
- The coalescent allows us to calculation the probability $\mathbb{P}(\mathcal{T} \mid N)$



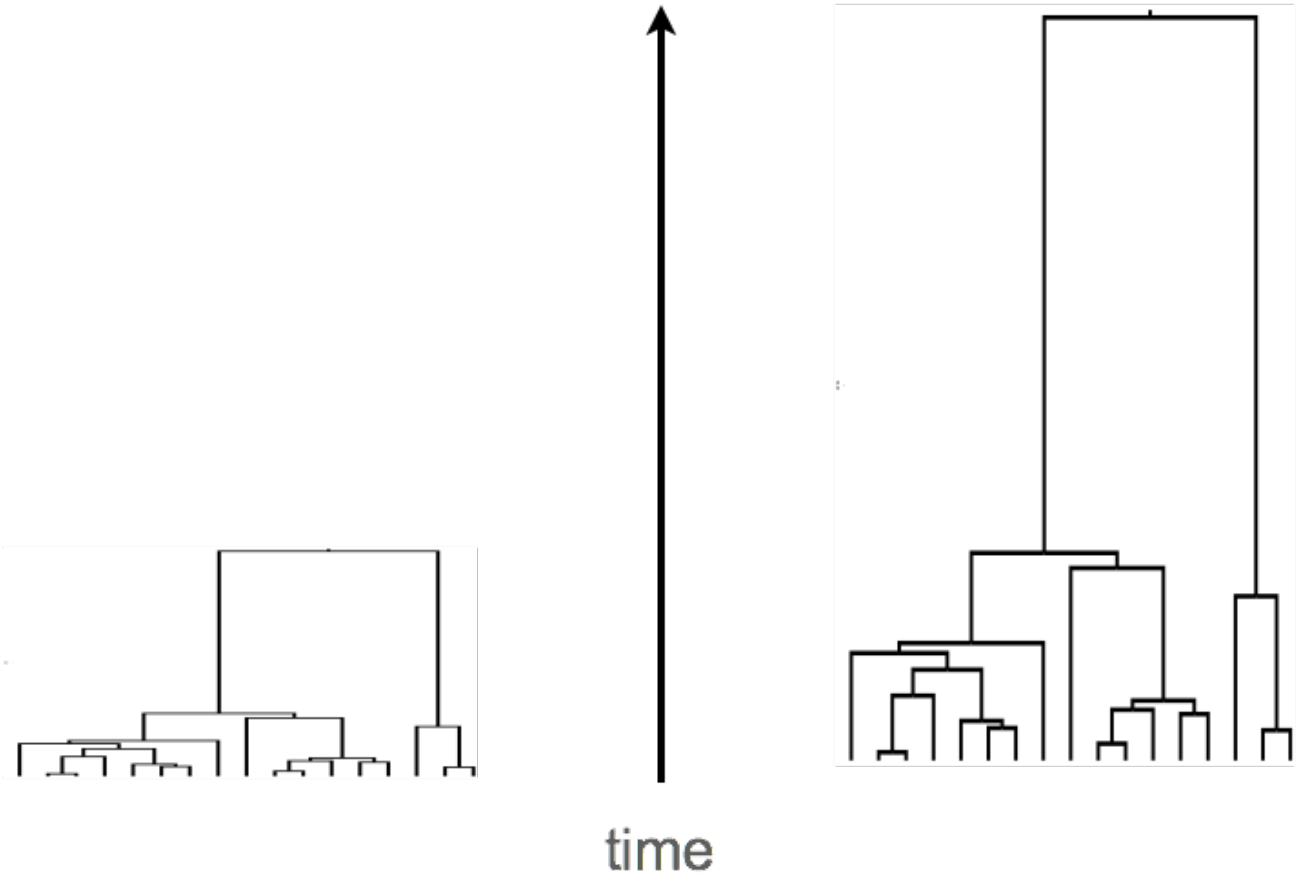
Or, the **inverse problem** : learn about N from $\mathcal{T}$

N governs the rate of coalescence



- Large and small population sizes

Quiz

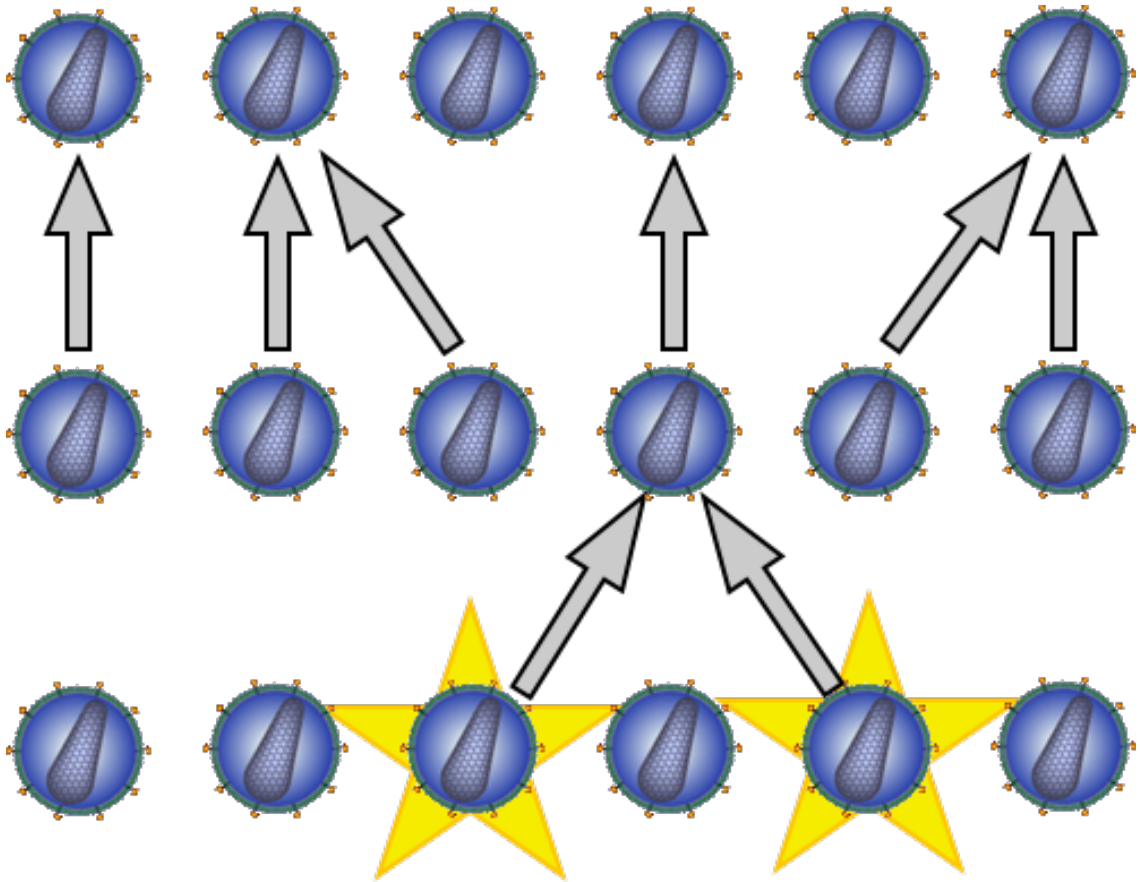


- Which population is large?

Coalescent assumptions

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

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

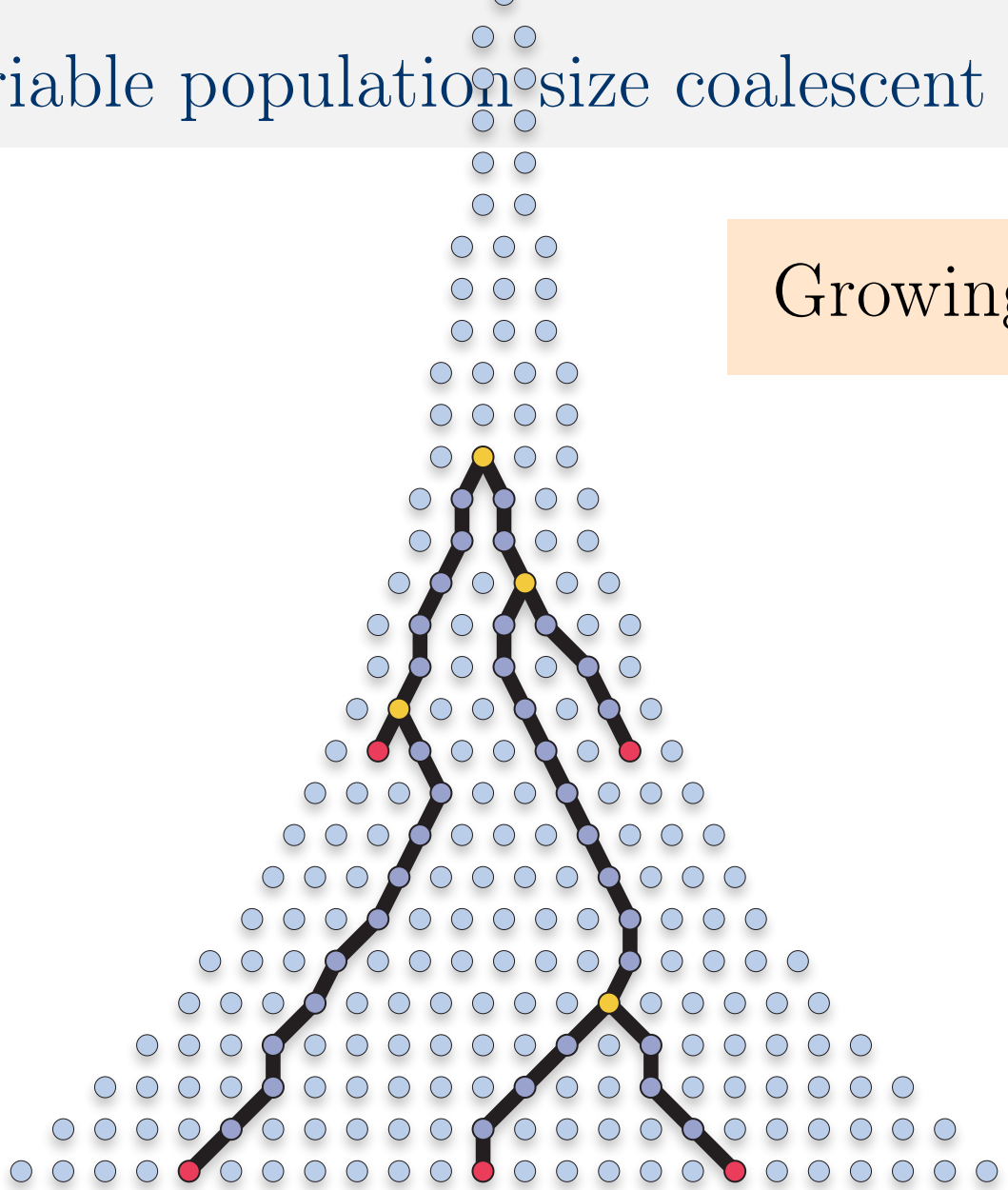


But, does this matter?

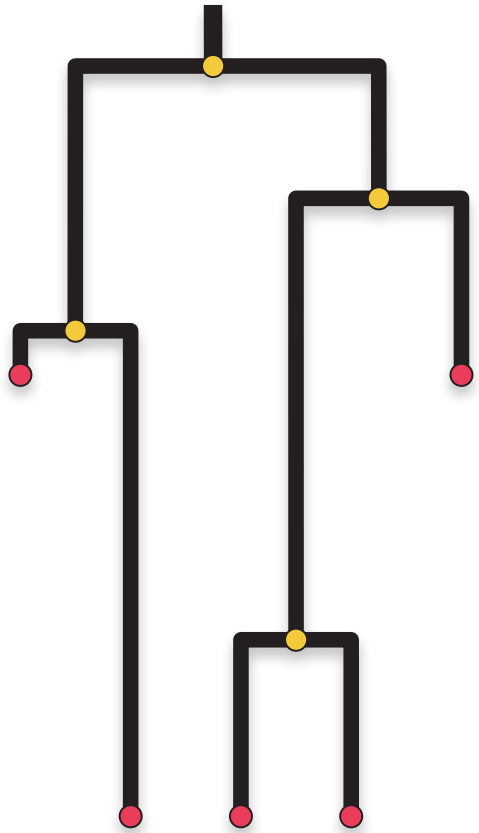
Solution: *Effective* population size

- Consider an abstract parameter, the effective population size N_e
- 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

Variable population size coalescent

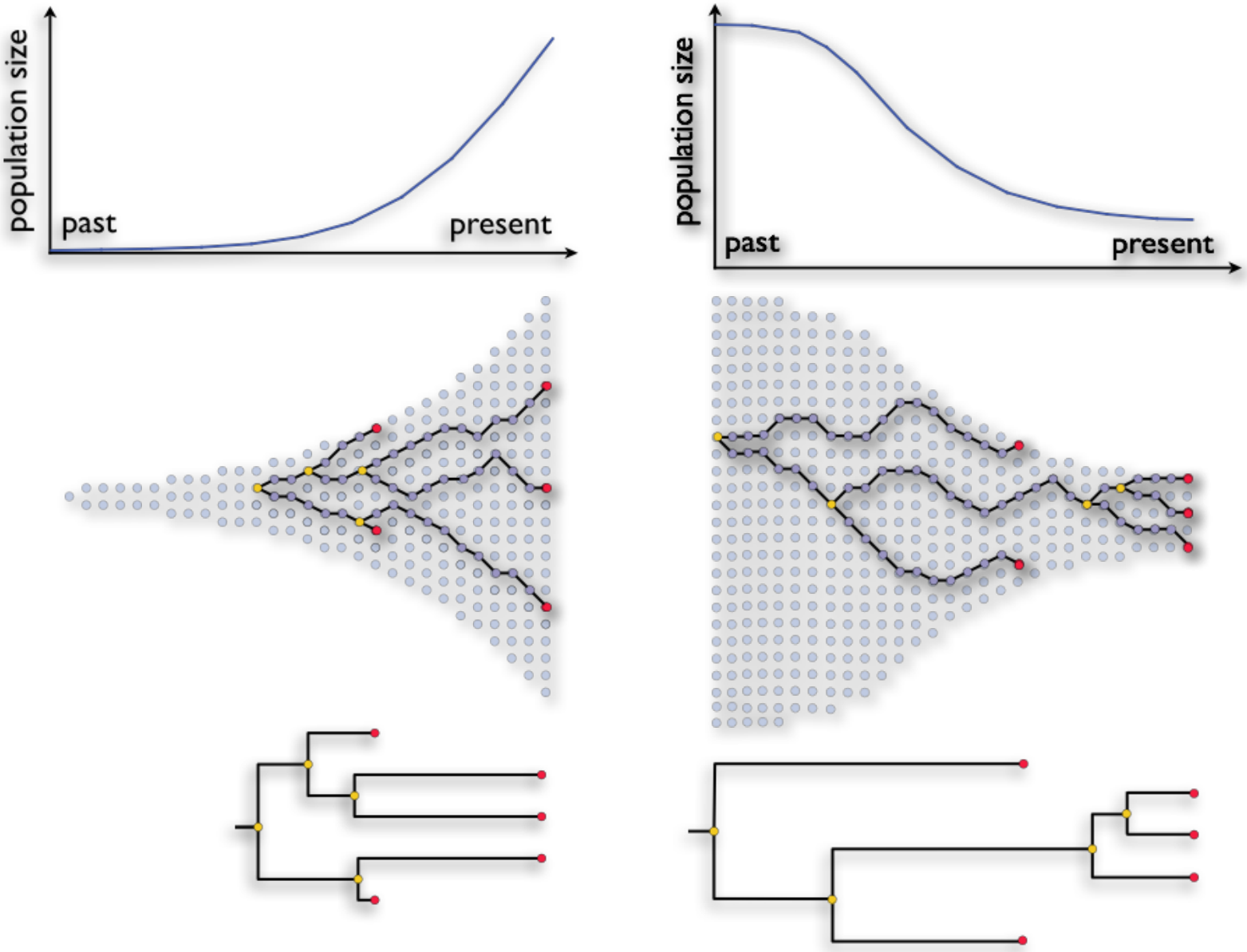


Growing population



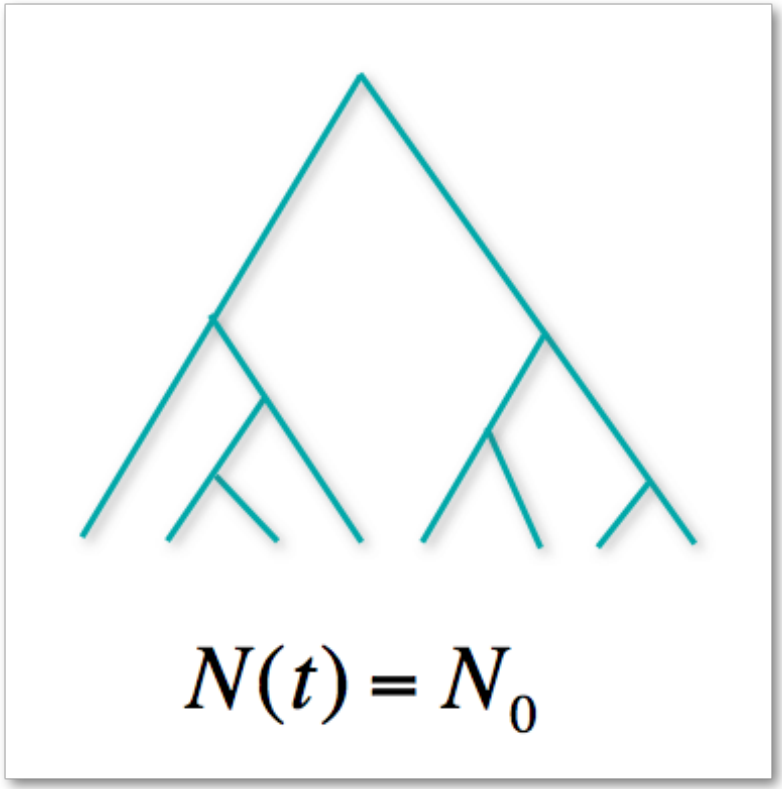
- Changes in N_e reflect changes in the census population size

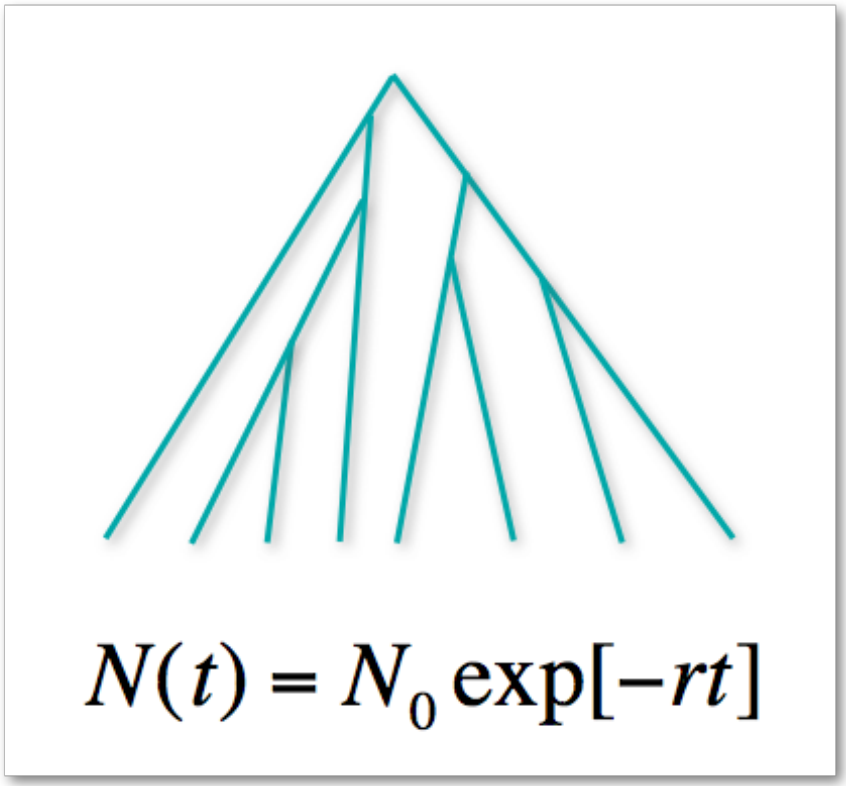
Variable population size coalescent



Parametric models of $N(t)$ through time

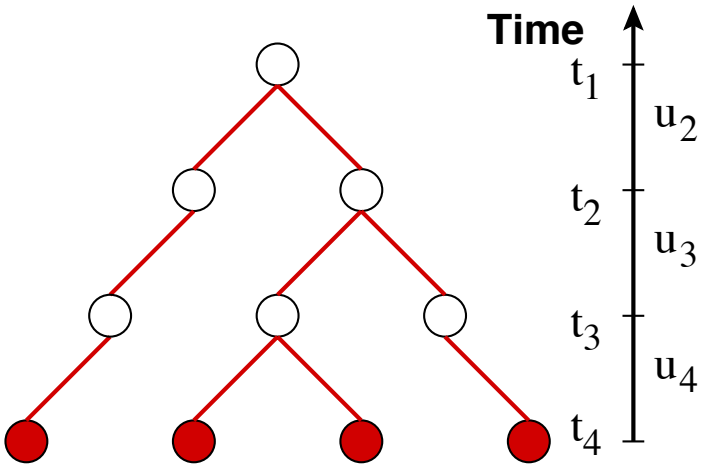
- The standard coalescent can be extended to accommodate various scenarios of demographic change through time





- However, few parametric forms (constant, exponential, logistic) are available. Can use piece-wise combinations

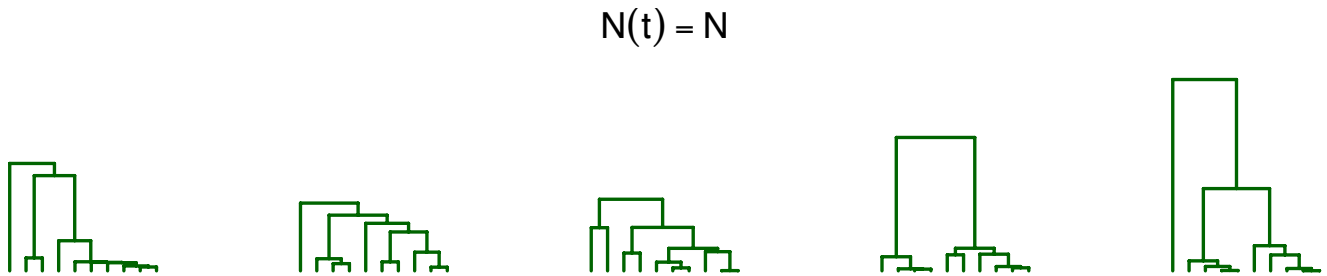
Review: continuous-time coalescent



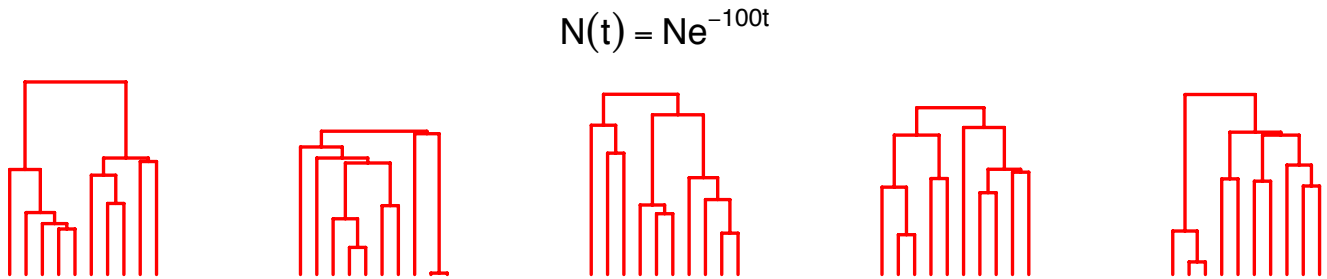
- Time measured in generation units
- $N = \text{const} \rightarrow u_k \sim \text{Exp}\left[\binom{k}{2} N\right]$
- $N = N(t) \rightarrow$

  $\Pr(u_k > t | t_{k+1}) = e^{-\binom{k}{2} \int_{t_{k+1}}^{t+t_{k+1}} \frac{N}{N(u)} du}$
- u_k are not independent any more

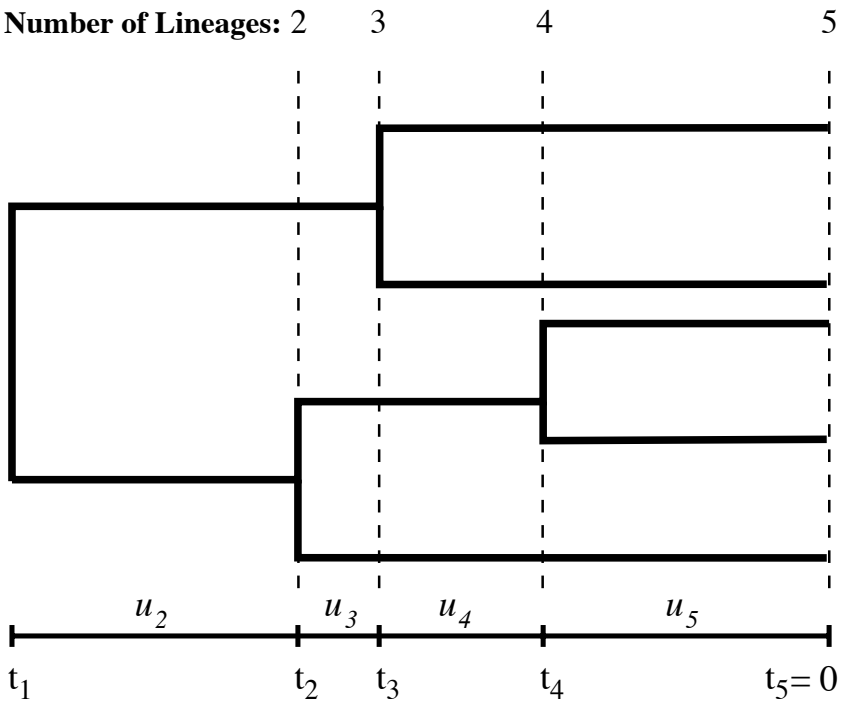
- Constant population size



- Exponential growth



Piecewise constant demographic model

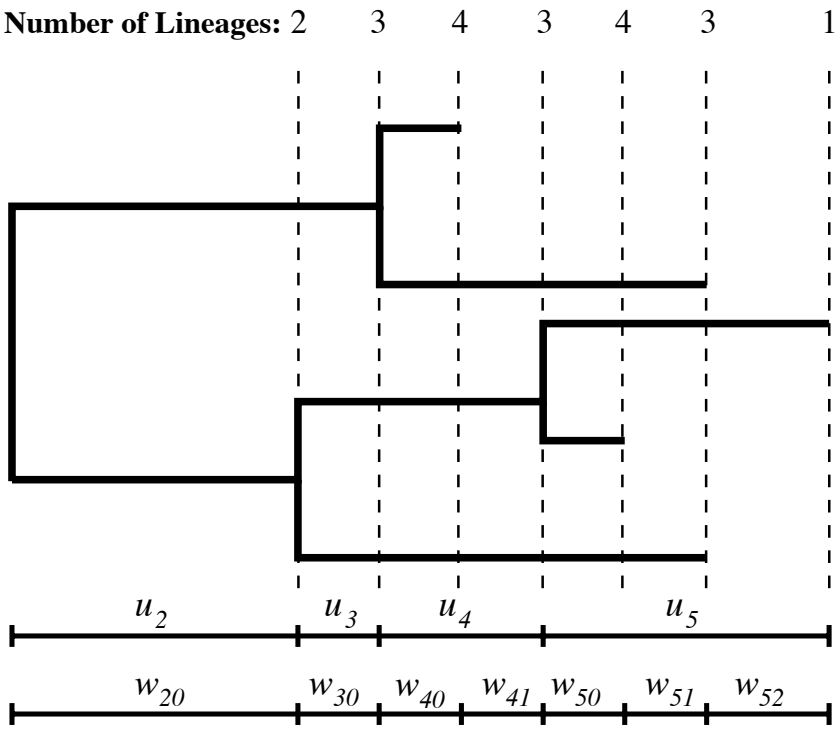


Isochronous Data

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

- Equivalent to estimating exponential mean from one observation.
- Need further restrictions to estimate all effective pop sizes $\boldsymbol{\theta}$!

Piecewise constant demographic model



Heterochronous Data

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

- Equivalent to estimating exponential mean from one observation.
- Need further restrictions to estimate all effective pop sizes $\boldsymbol{\theta}$!

Previous priors to restrict $\boldsymbol{\theta}$

Strimmer and Pybus (2001)

- Make $N_e(t)$ constant across some inter-coalescent times
- Group inter-coalescent intervals with AIC

Drummond et al. (2005)

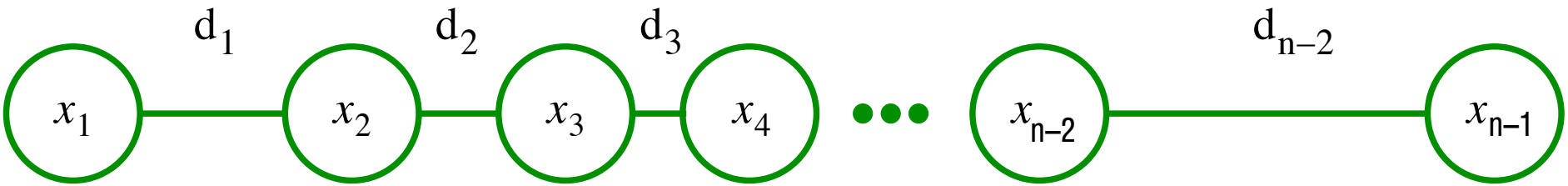
- Multiple change-point model with fixed number of change-points
- Change-points allowed only at coalescent events
- Joint estimation of phylogenies and population dynamics

Opgen-Rhein et al. (2005)

- Multiple change-point model with random number of change-points
- Change-points allowed anywhere in interval $(0, t_1]$
- Posterior is approximated with rjMCMC

Smoothing priors - Gaussian Markov random fields

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



Weighting Schemes

1. Skyride : weights d_k determined by tree (in relative time)
2. Skygrid : d_k on a regular grid in absolute time + multi-locus

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

MCMC algorithm

$$\Pr(\mathbf{G}, \mathbf{Q}, \mathbf{x} \,|\, \mathbf{D}) \propto \Pr(\mathbf{D} \,|\, \mathbf{G}, \mathbf{Q}) \Pr(\mathbf{Q}) \Pr(\mathbf{G} \,|\, \mathbf{x}) \Pr(\mathbf{x})$$

Updating Population Size Trajectory

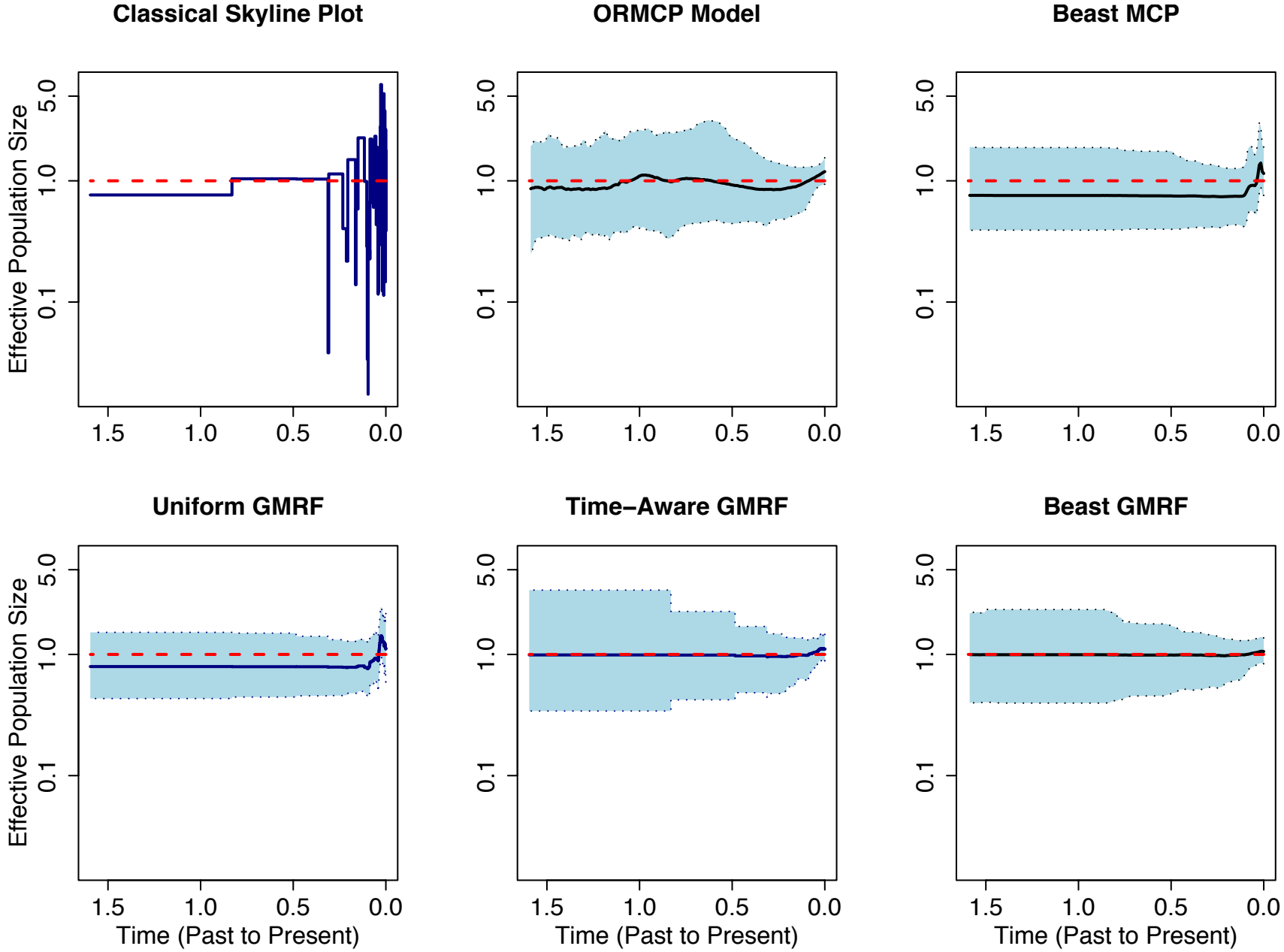
- Use fast GMRF sampling (Rue et al., 2001, 2004)
- Draw ω^* from an arbitrary univariate proposal distribution
- Use Gaussian approximation of $\Pr(\mathbf{x} \,|\, \omega^*, \mathbf{G})$ to propose $\mathbf{x}^*$
- Jointly accept/reject $(\omega^*, \mathbf{x}^*)$ in Metropolis-Hastings step

Object-Oriented Reality?

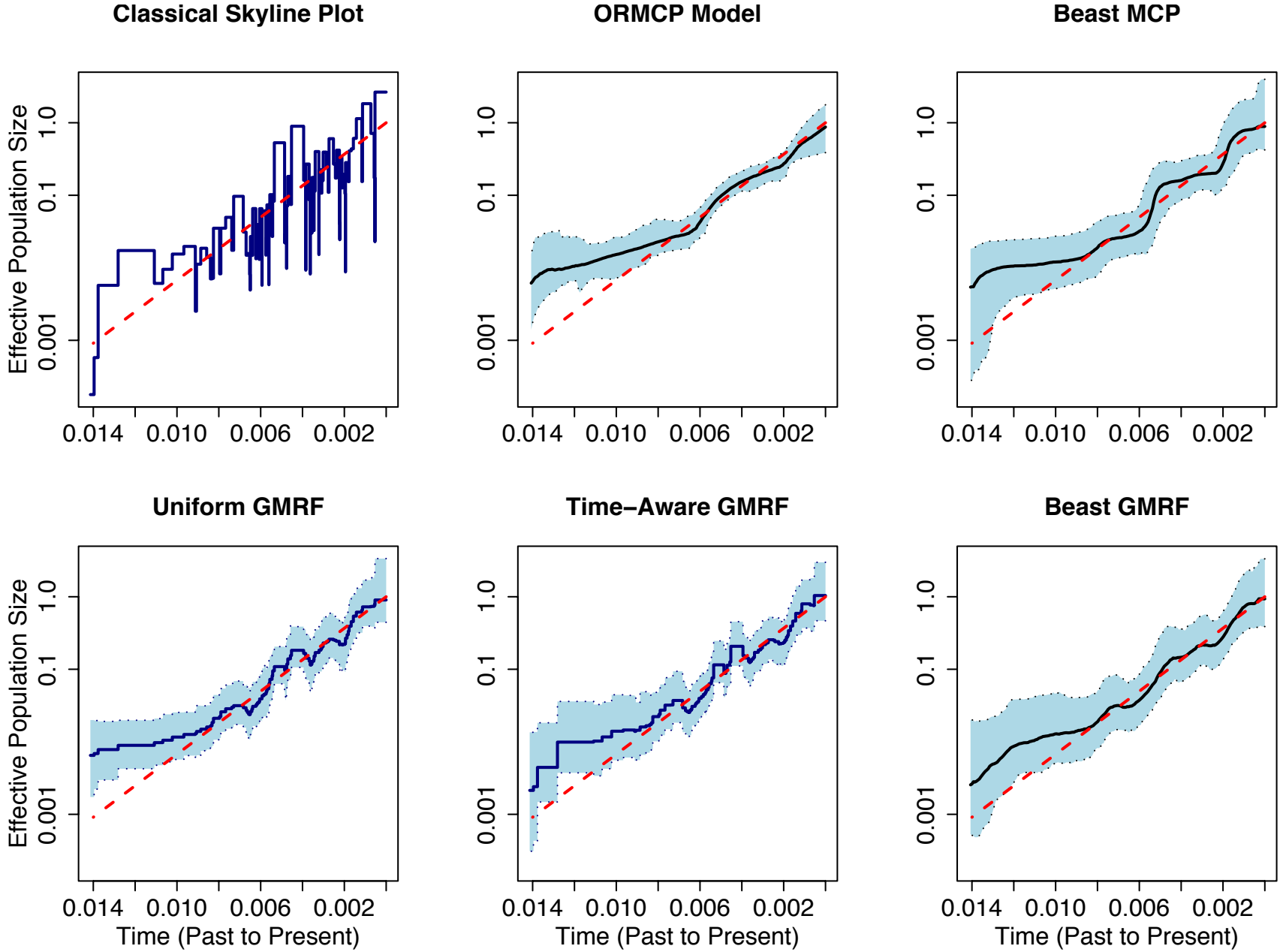
BEAST = **B**ayesian **E**volutionary **A**nalysis **S**ampling **T**rees

- $\Pr(\mathbf{G} \,|\, \mathbf{x}, \mathbf{D}, \mathbf{Q})$ - sampled by BEAST
- $\Pr(\mathbf{Q} \,|\, \mathbf{G}, \mathbf{D})$ - sampled by BEAST

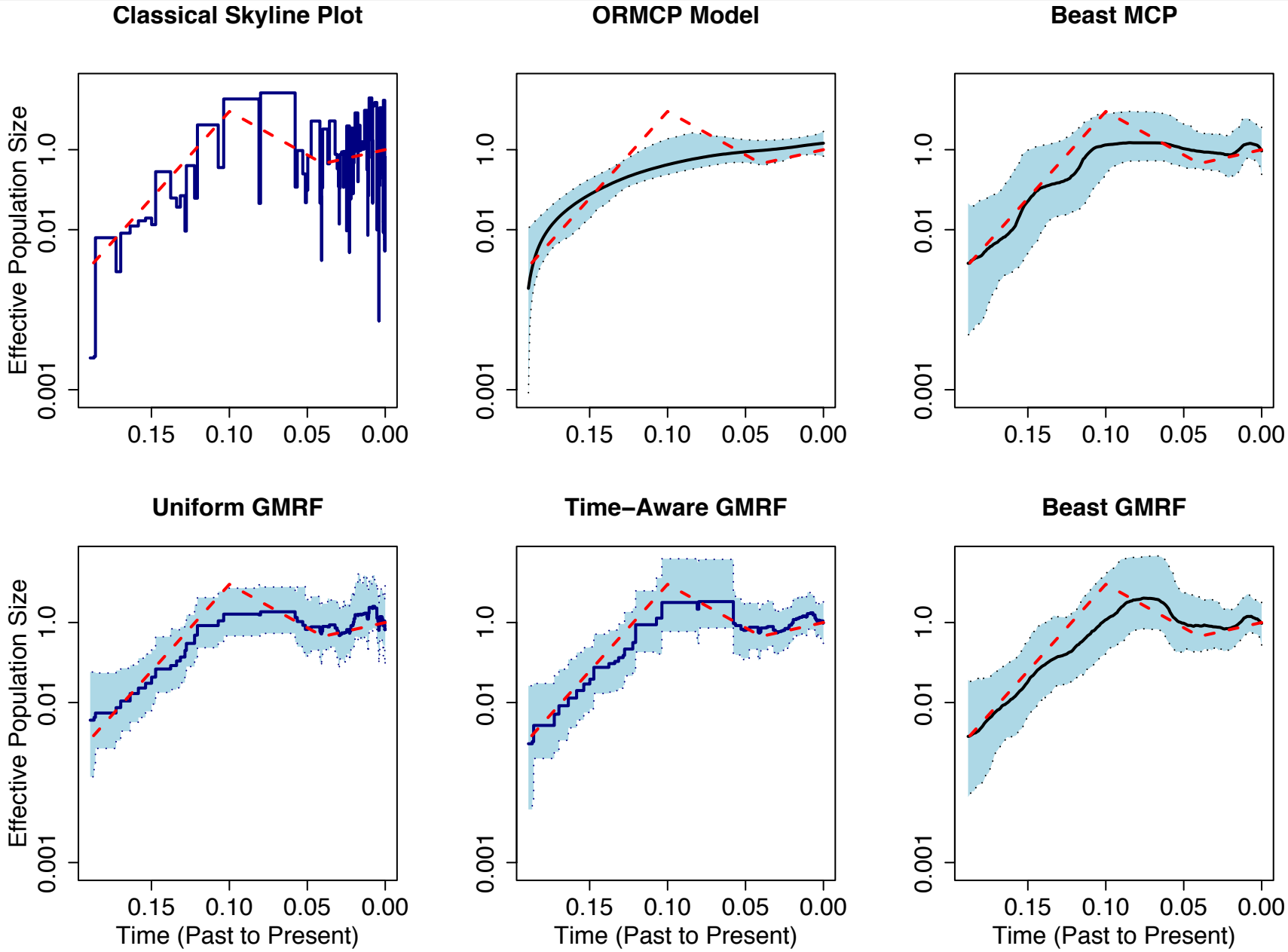
Simulation: constant population size



Simulation: exponential growth



Simulation: exponential growth with bottleneck



Accuracy in simulations

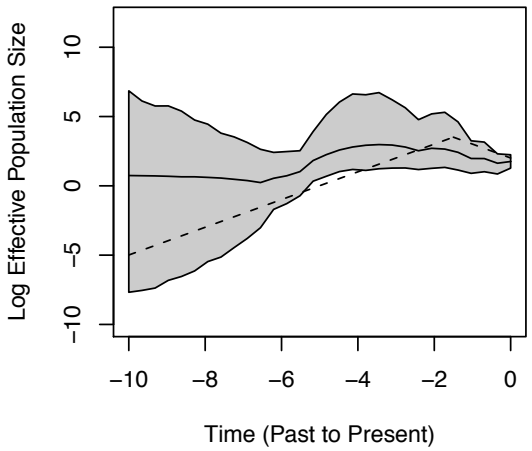
$$\text{Percent error} = \int_0^{\text{TMRCA}} \frac{|\widehat{N}_e(t) - N_e(t)|}{N_e(t)} dt \times 100,$$

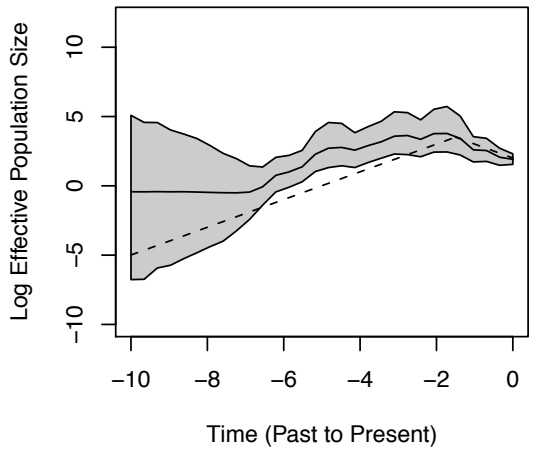
Table: Percent error in simulations. We compare percent errors, defined in equation (1), for the Opgen-Rhein multiple change-point (ORMCP), uniform and fixed-tree time-aware Gaussian Markov random field (GMRF) smoothing, BEAST multiple change-point (MCP) model, and BEAST GMRF smoothing.

Model	Constant	Exponential	Bottleneck
ORMCP	14.0	1.7	7.4
Uniform GMRF	32.8	1.5	5.9
Time-Aware GMRF	2.8	1.2	4.8
BEAST MCP	38.2	1.6	5.2
BEAST GMRF	1.7	1.0	5.4

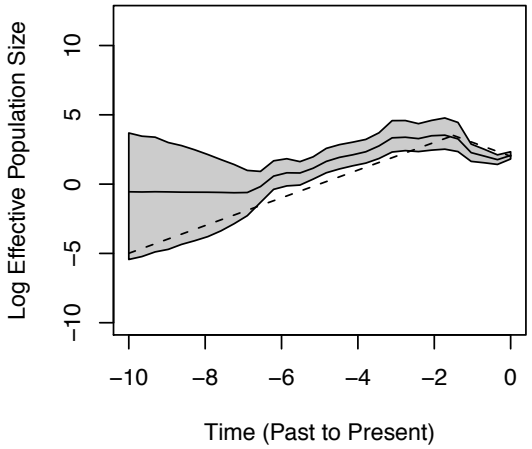
Multi-locus performance

- Adding loci improves dynamics recovery under the Skygrid

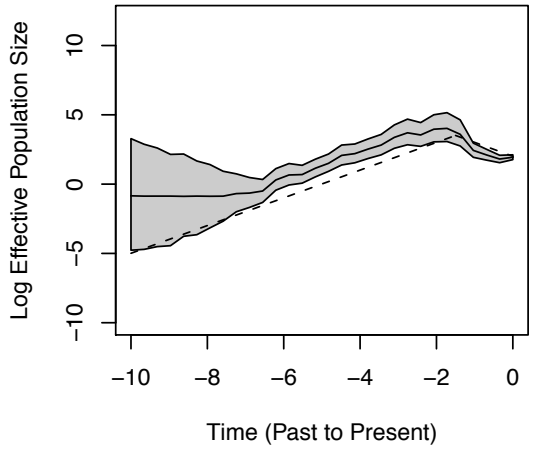




Five Loci



Ten Loci

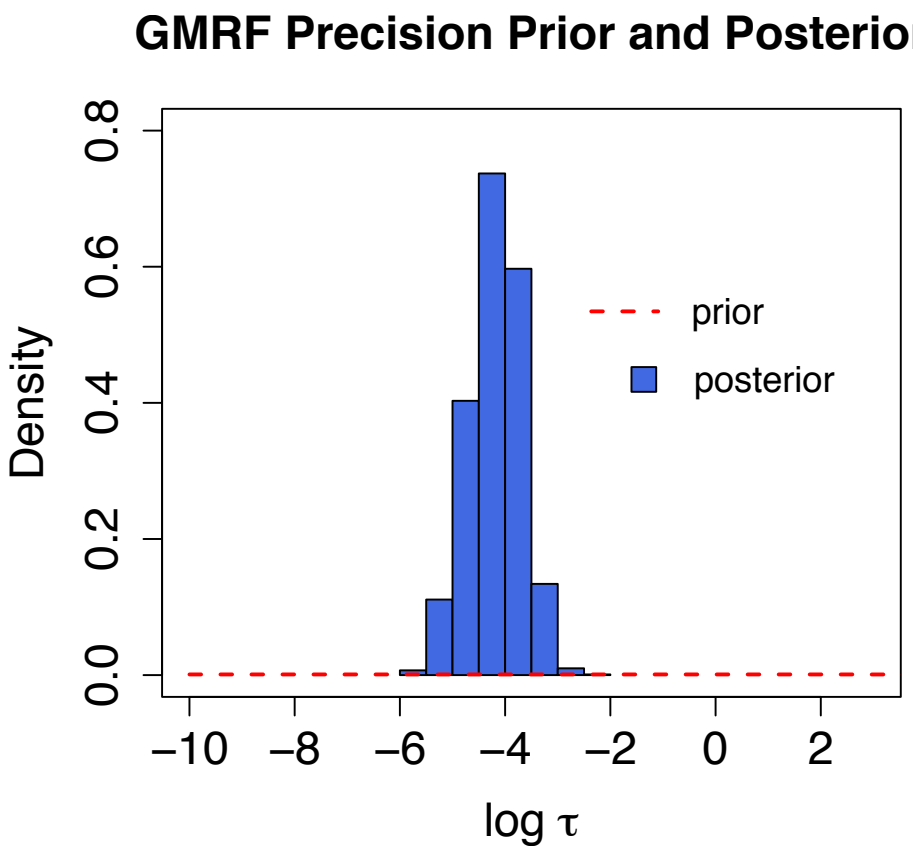


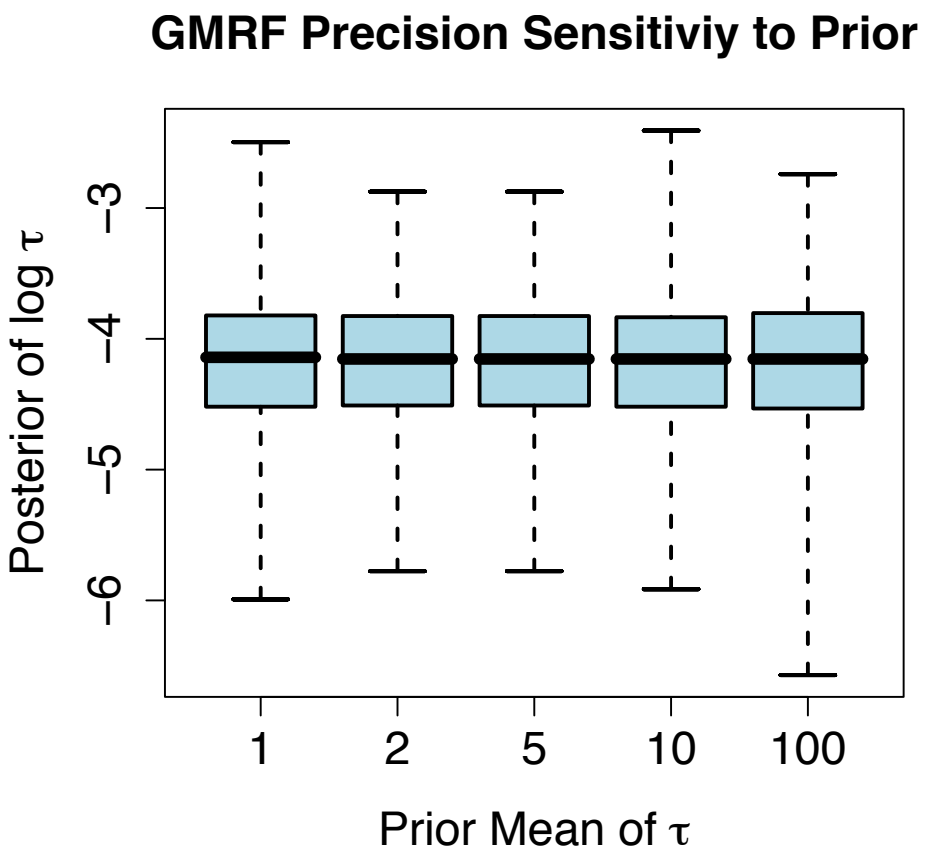
loci	Relative % error
1	1.23
2	1.66
5	1.90
10	2.41

Table: Relative error of extended Bayesian skyline plot (EBSP) to skygrid under exponential growth

GMRF Precision Prior Sensitivity

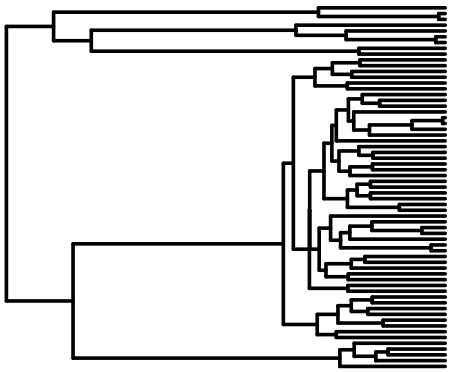
- ω - GMRF precision, controls smoothness
- Usually $\Pr(\omega \mid \mathbf{D})$ is sensitive to perturbations of $\Pr(\omega)$
- Not in our coalescent model!



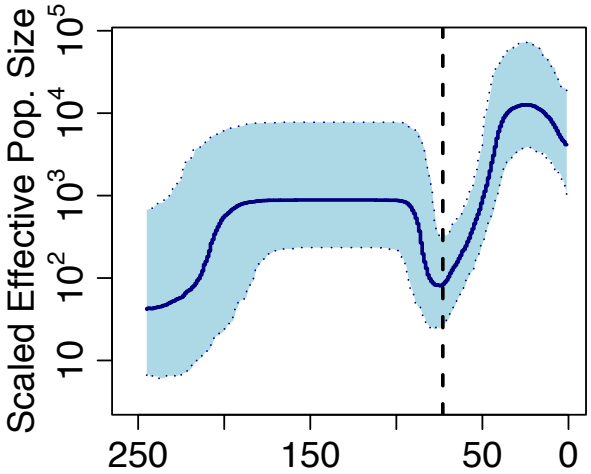


HCV Epidemics in Egypt

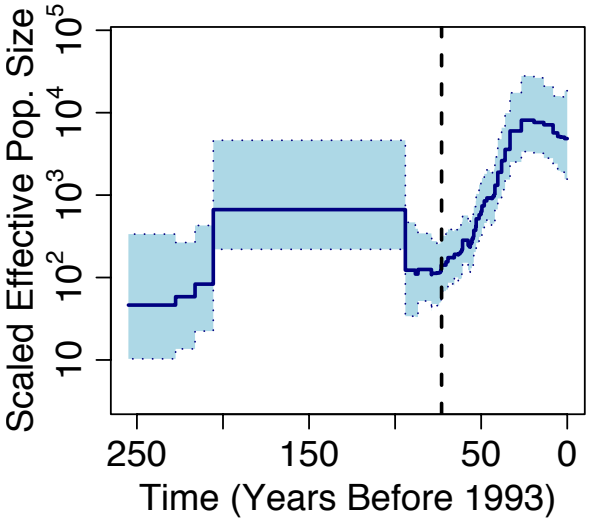
Estimated Genealogy



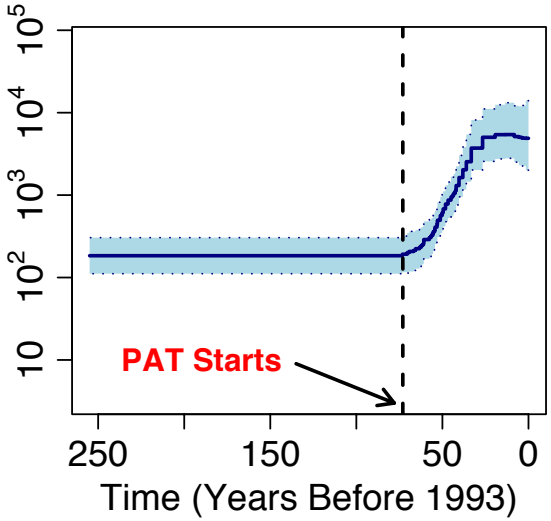
BEAST GMRF



Unconstrained Fixed–Tree GMRF

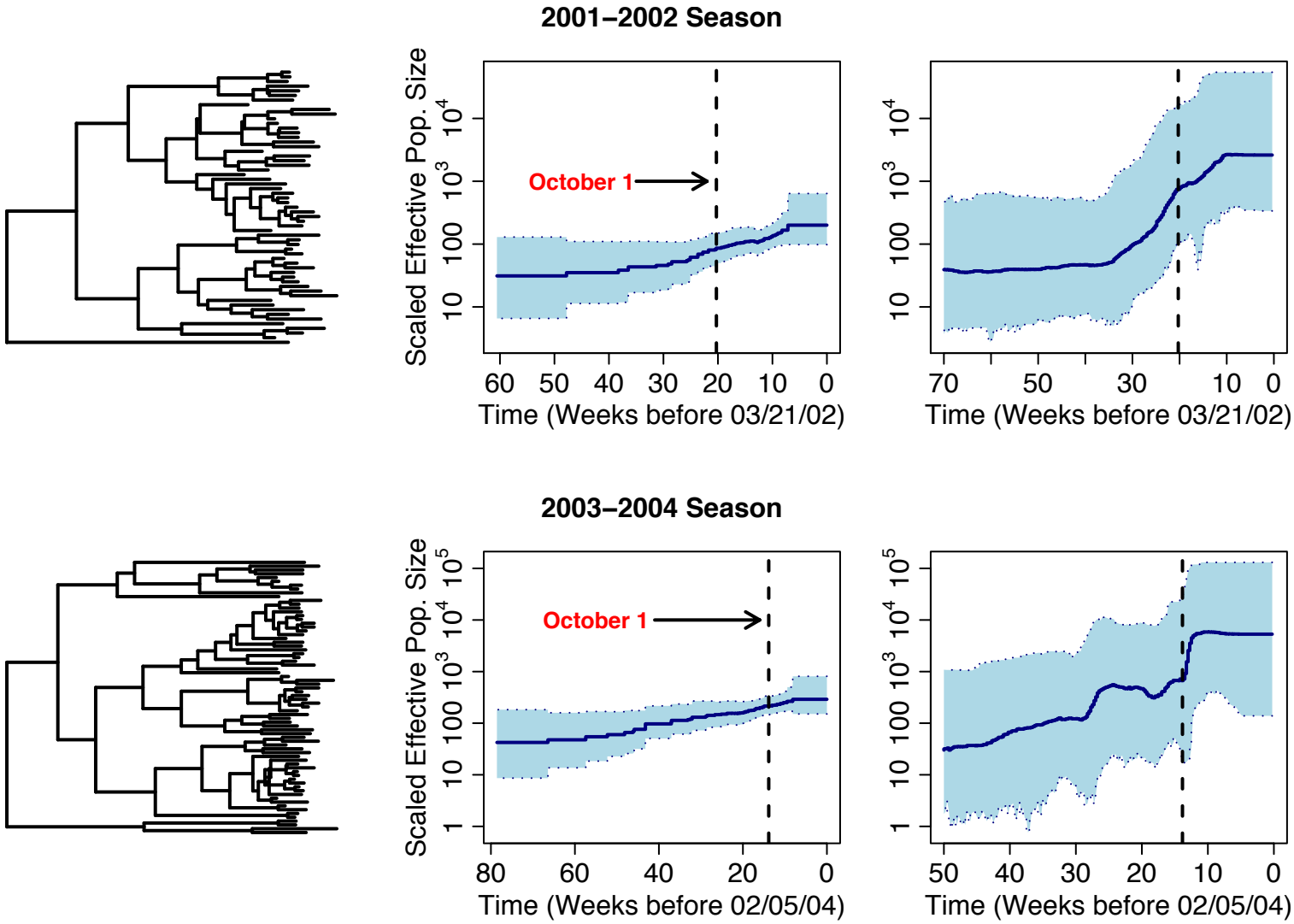


Constrained Fixed–Tree GMRF



- 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

Influenza Intra-Season Population Dynamics



New York state hemagglutinin sequences serially sampled (Ghedin et al., 2005)

Summary

- Genealogies inform us about population size trajectories
- Prior restrictions are necessary for non(semi)-parametric estimation of $N_e(t)$
- Smoothing can be imposed by GMRF priors

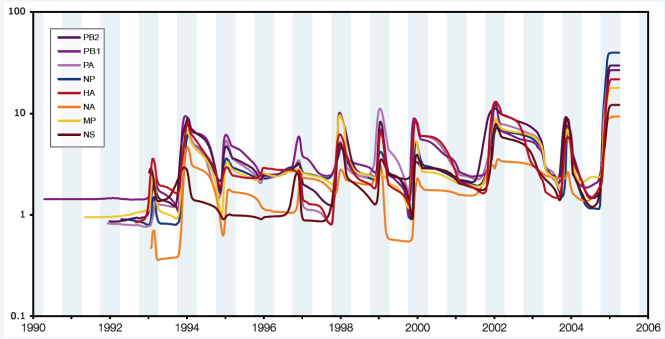
Software: skyride and skygrid

- Implemented as coalescent priors in BEAST
- Exploit approximate Gibbs sampling
- Faster convergence? Better mixing?

References:
Minin et al. (2008) *Molecular Biology & Evolution*, 25, 1459–1471.
Gill et al. (2013) *Molecular Biology & Evolution*, 30, 713–724.

Active ideas: GMRFs are highly generalizable

Hierarchical Modeling



Flu genes display similar (not equal) dynamics

- Incorporate multiple loci simultaneously
- Pool information for statistical power
- No need for strict equality

Introducing Covariates

- Augment field at fixed observation times
- Formal statistical testing for:
  - External factors (environment, drug tx)
  - Population dynamics (bottle-necks, growth)

Gill et al. (in press) *Systematic Biology*.