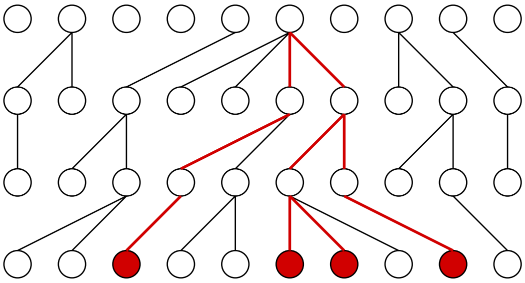
Non-Parametric Bayesian Population Dynamics Inference

Philippe Lemey and Marc A. Suchard

Department of Microbiology and Immunology
K.U. Leuven, Belgium, and
Departments of Biomathematics, Biostatistics and Human Genetics
University of California, Los Angeles

SISMID

Review: Continuous-Time Coalescent

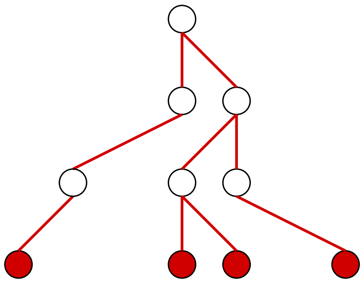


- Time measured in N generation units
- N = const $\rightarrow u_k \sim \text{Exp}\left[\binom{k}{2}\right]$
- $N = N(t) \rightarrow$

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

- Constant population size
- Exponential growth

Review: Continuous-Time Coalescent

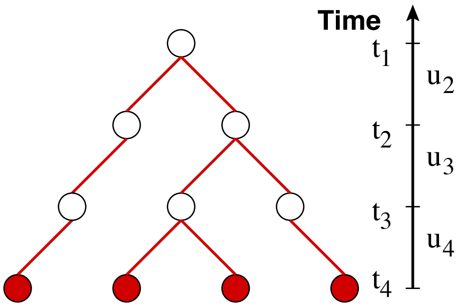


- Time measured in N generation units
- N = const → $u_k \sim \text{Exp}\left[\binom{k}{2}\right]$
- $N = N(t) \rightarrow$

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

- Constant population size
- Exponential growth

Review: Continuous-Time Coalescent

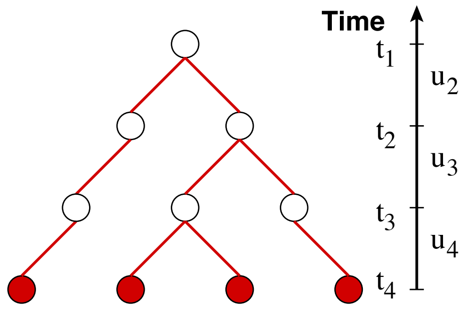


- Time measured in N generation units
- N = const → $u_k \sim \text{Exp}\left[\binom{k}{2}\right]$
- $N = N(t) \rightarrow$

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

- Constant population size
- Exponential growth

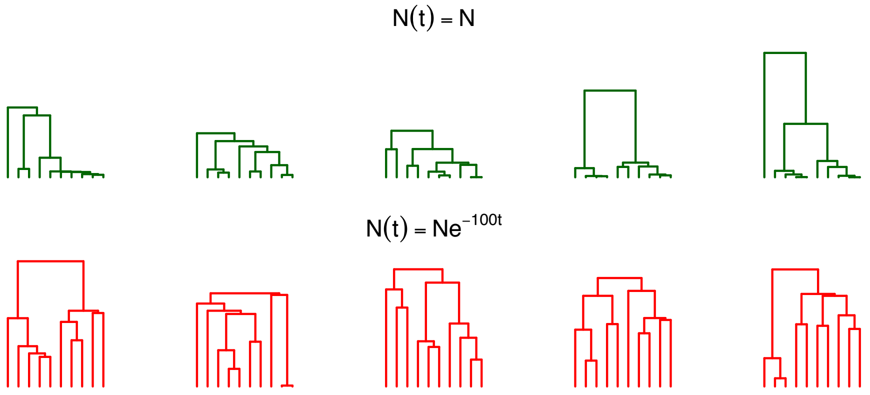
Review: Continuous-Time Coalescent



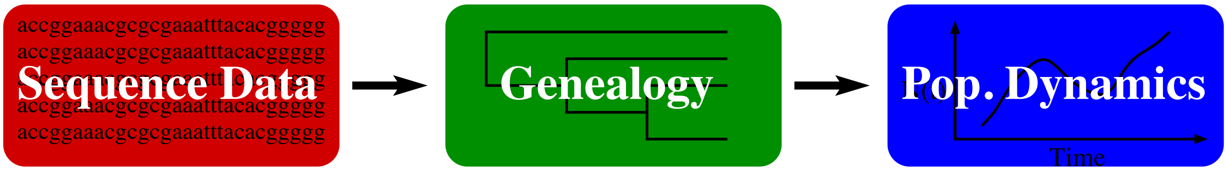
- Time measured in N generation units
- N = const → $u_k \sim \text{Exp}\left[\binom{k}{2}\right]$
- $N = N(t) \rightarrow$

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

- Constant population size
- Exponential growth



Sequence Data → Population Model Parameters

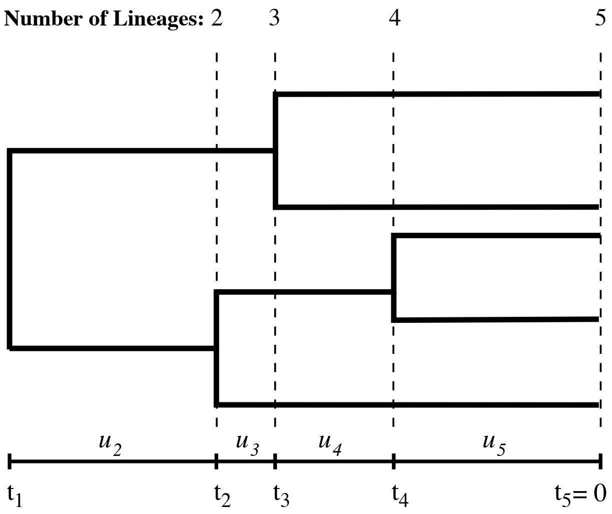

accggaaacgcgcgaaatttacacggggg
accggaaacgcgcgaaatttacacggggg
accggaaacgcgcgaaatttacacggggg
accggaaacgcgcgaaatttacacggggg
accggaaacgcgcgaaatttacacggggg
Time

Sequence Data → Genealogy → Pop. Dynamics

More Formally (Bayesian Approach):

- $\Pr(\mathbf{G}, \mathbf{Q}, \theta \mid \mathbf{D}) \propto \Pr(\mathbf{D} \mid \mathbf{G}, \mathbf{Q}) \Pr(\mathbf{Q}) \Pr(\mathbf{G} \mid \theta) \Pr(\theta)$
- **G** - genealogy with branch lengths
- **Q** - substitution matrix
- θ - population genetics parameters
- **D** - sequence data
- $\Pr(\mathbf{G} \mid \theta)$ - **Coalescent prior**

Piecewise Constant Demographic Model

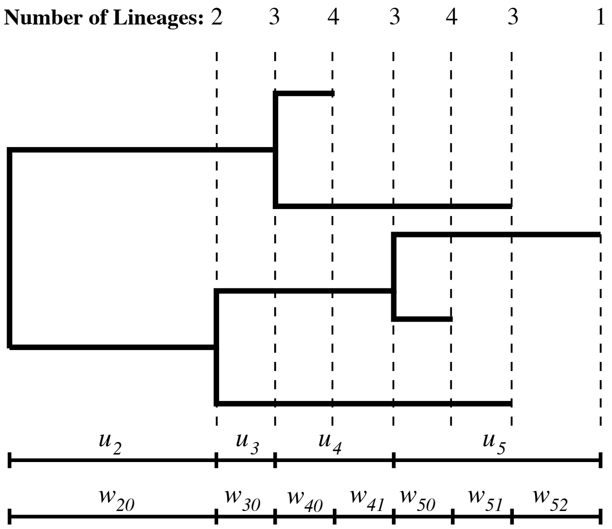


Isochronous Data

- $N_e(t) = \theta_k$ 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 θ!

Piecewise Constant Demographic Model

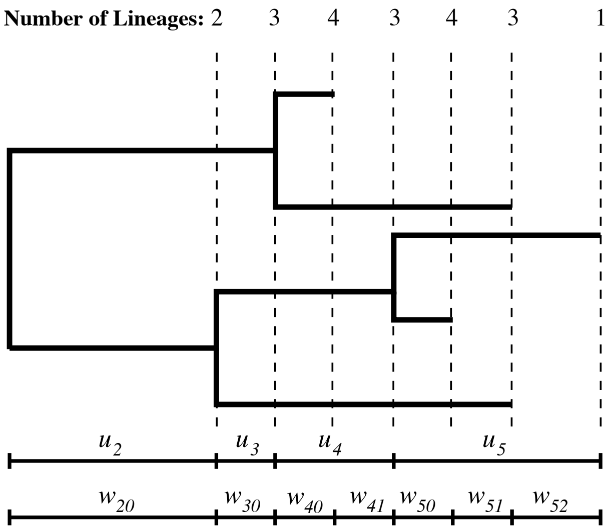


Heterochronous Data

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

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

Piecewise Constant Demographic Model



Heterochronous Data

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

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

Current Approaches

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

Current Approaches

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

Current Approaches

Strimmer and Pybus (2001)

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

Drummond et al. (2005)

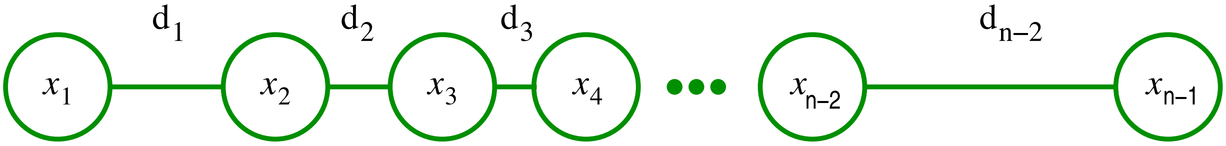
- Multiple change-point model with fixed number of change-points
- Change-points allowed only at Coalescent events
- Joint estimation of phylogenies and population dynamics

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 Prior (GMRF approach)

- 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 $\alpha = 0.01$, $\beta = 0.01$

MCMC Algorithm

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

Updating Population Size Trajectory

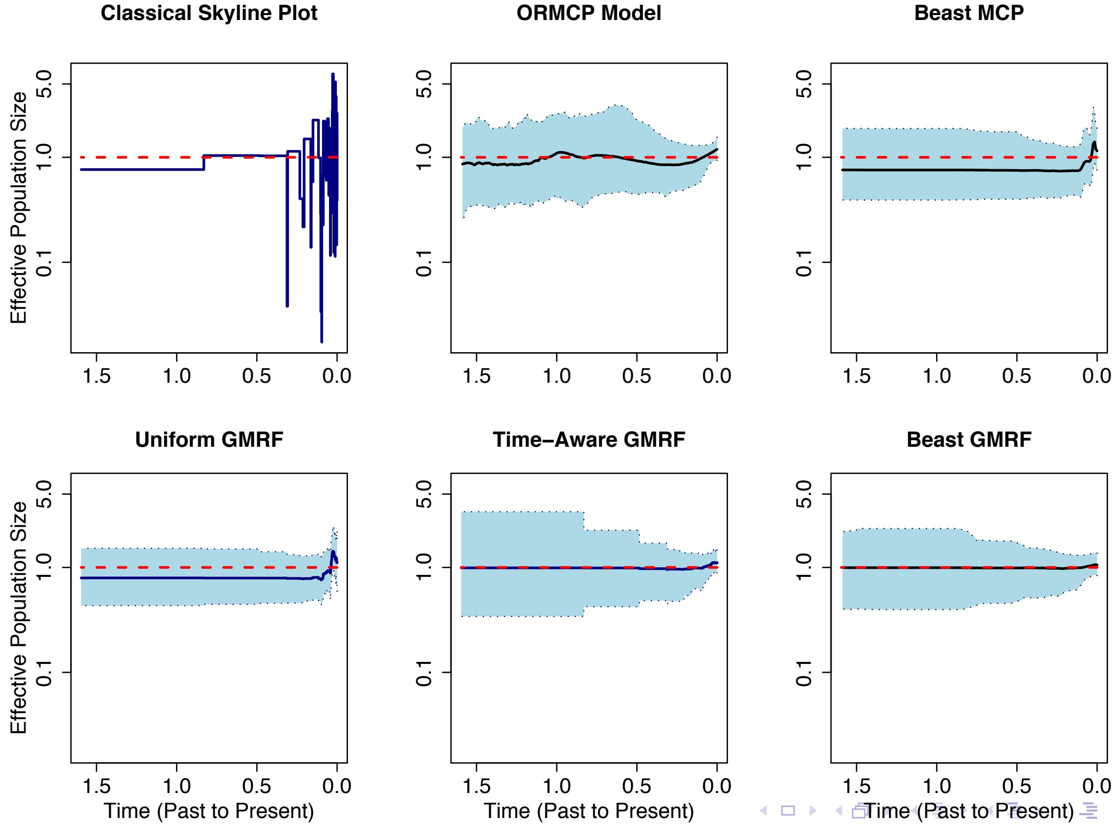
- Use fast GMRF sampling (Rue et al., 2001, 2004)
- Draw ω^* from an arbitrary univariate proposal distribution
- Use Gaussian approximation of $\Pr(\mathbf{x} \mid \omega^*, \mathbf{G})$ to propose $\mathbf{x}^*$
- Jointly accept/reject $(\omega^*, \mathbf{x}^*)$ in Metropolis-Hastings step

Object-Oriented Reality?

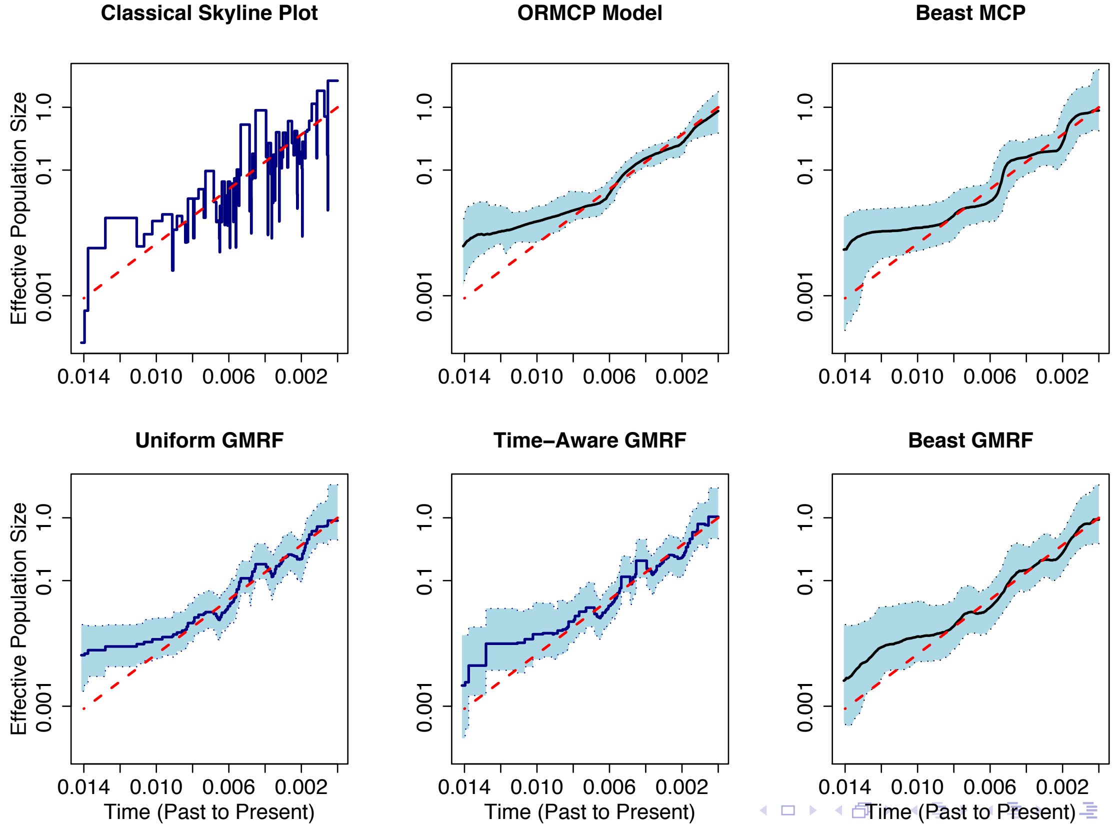
BEAST = **B**ayesian **E**volutionary **A**nalysis **S**ampling **T**rees

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

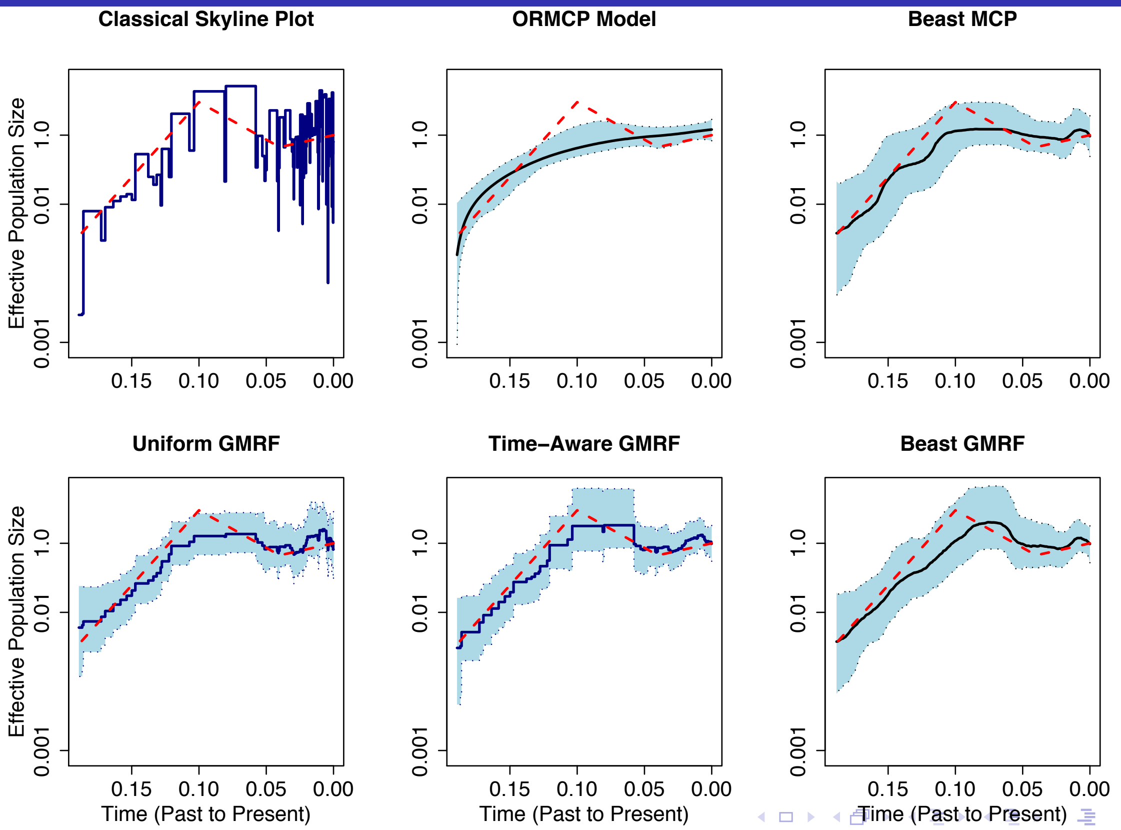
Simulation: Constant Population Size



Simulation: Exponential Growth



Simulation: Exponential Growth with Bottleneck



Accuracy in Simulations

$$\text{Percent Error} = \int_0^{\text{TMRCA}} \frac{|\widehat{N}_e(t) - N_e(t)|}{N_e(t)} dt \times 100, \qquad (1)$$

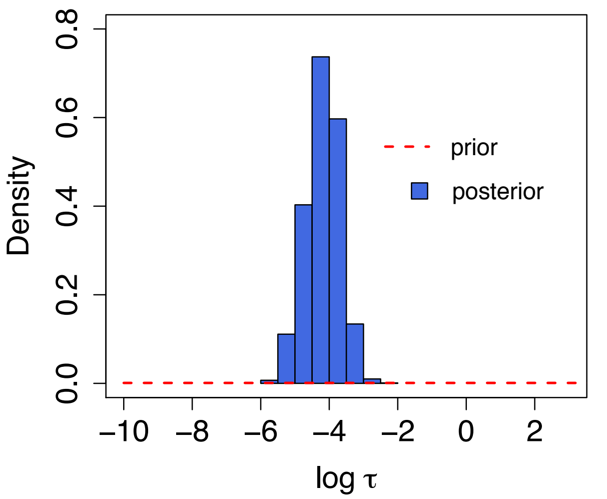
Table: Percent error in simulations. We compare percent errors, defined in equation (1), for the Opgen-Rhein multiple change-point (ORMCP), uniform and fixed-tree time-aware Gaussian Markov random field (GMRF) smoothing, BEAST multiple change-point (MCP) model, and BEAST GMRF smoothing.

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

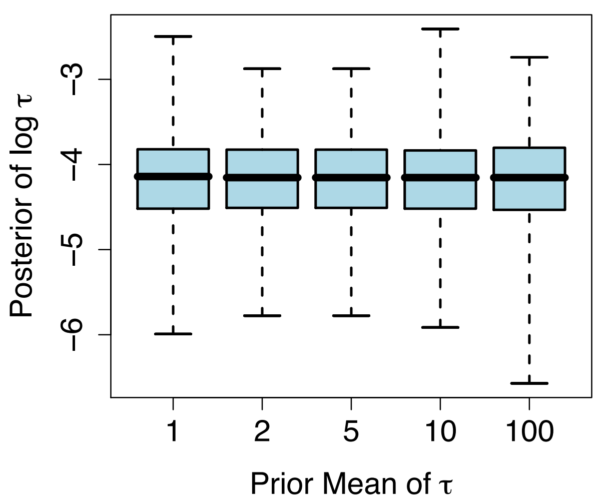
GMRF Precision Prior Sensitivity

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

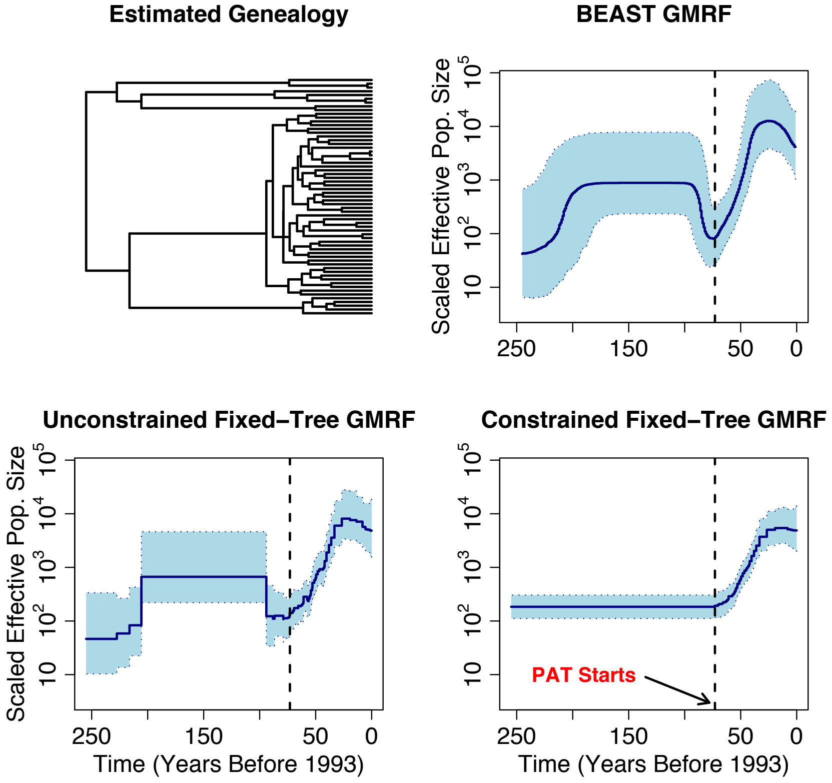
GMRF Precision Prior and Posterior



GMRF Precision Sensitiviy to Prior

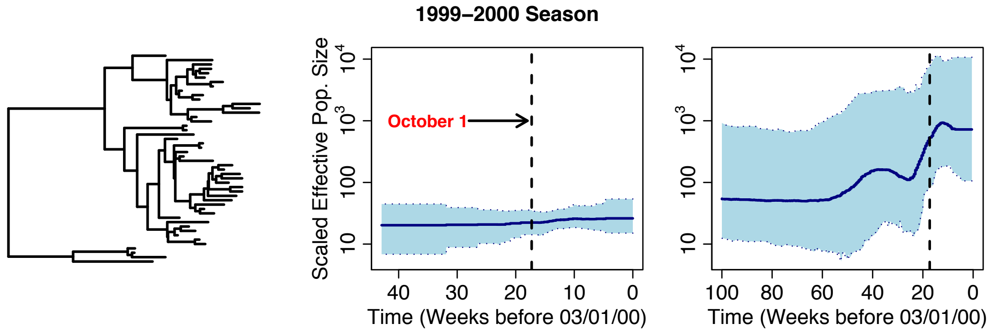


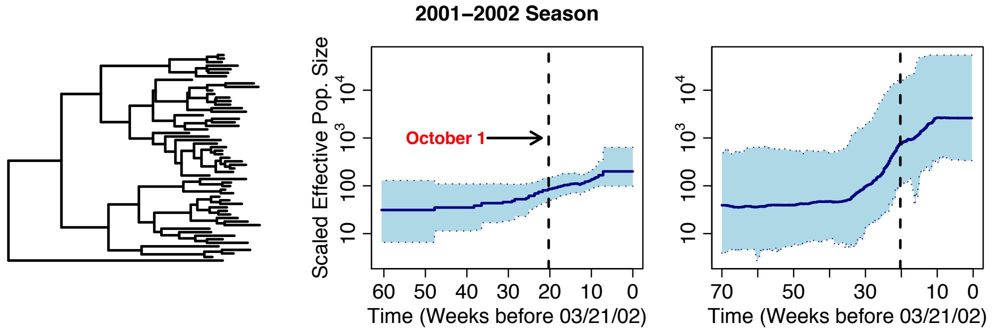
HCV Epidemics in Egypt



- 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

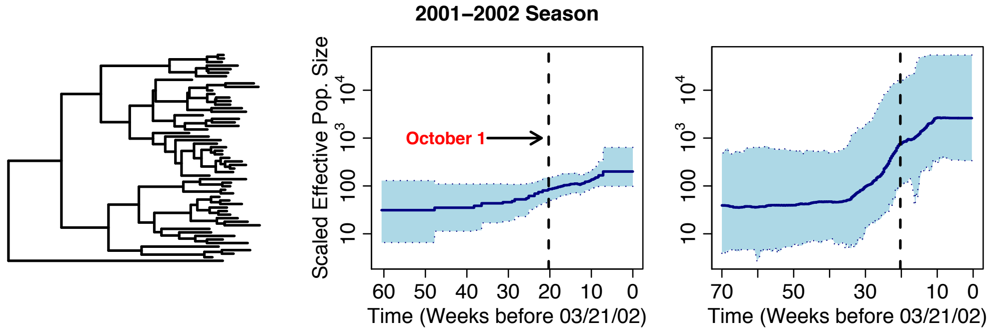
Influenza Intra-Season Population Dynamics

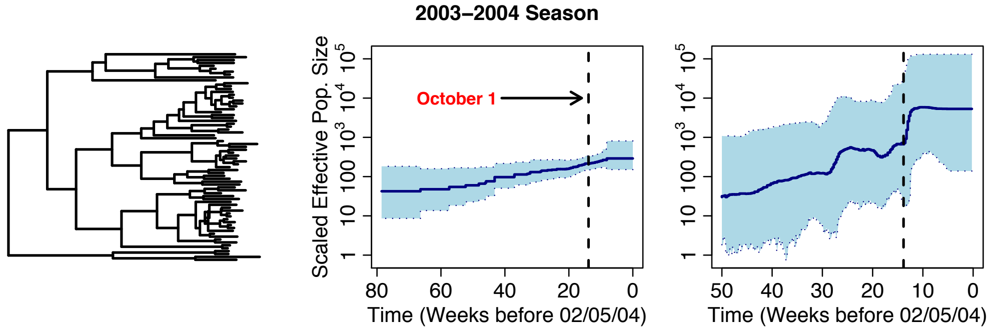




New York state hemagglutinin sequences serially sampled (Ghedin et al., 2005)

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

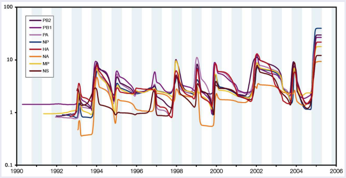
Software: The Skyride

- Implemented as a Coalescent prior in BEAST
- Exploits approximate Gibbs sampling
- Faster convergence? Better mixing?

Reference: Minin, Bloomquist and Suchard (2008) *Molecular Biology & Evolution*, 25, 1459–1471.

Active Ideas: GMRFs are Highly Generalizable

Hierarchical Modeling

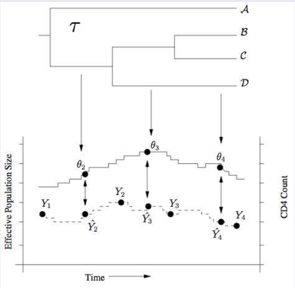


Flu genes display similar (not equal) dynamics

- 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)


$\mathcal{T}$
$\mathcal{A}$
$\mathcal{B}$
$\mathcal{C}$
$\mathcal{D}$
θ_2
θ_3
θ_4
Y_1
Y_2
Y_3
Y_4
$\hat{Y}_2$
$\hat{Y}_3$
$\hat{Y}_4$
Effective Population Size
CD4 Count
Time