# Introduction to Bayesian Phylogenetics (part 2)

Vladimir N. Minin

UCI Department of Statistics
Donald Bren School of Information & Computer Sciences

Taming the BEAST Workshop
June 2018

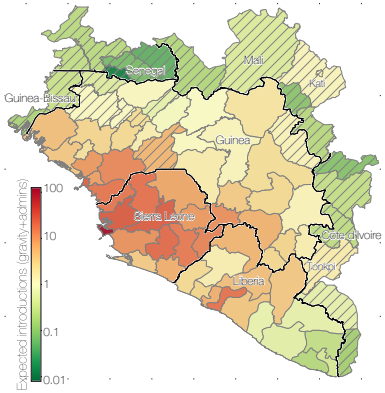
**featuring joint work with** Trevor Bedford (Fred Hutch), Gytis Dudas (Fred Hutch), Jim Faulkner (UW, NOAA), Michael Karcher (UW), Julia Palacios (Stanford), Marc Suchard (UCLA), Mingwei Tang (UW)

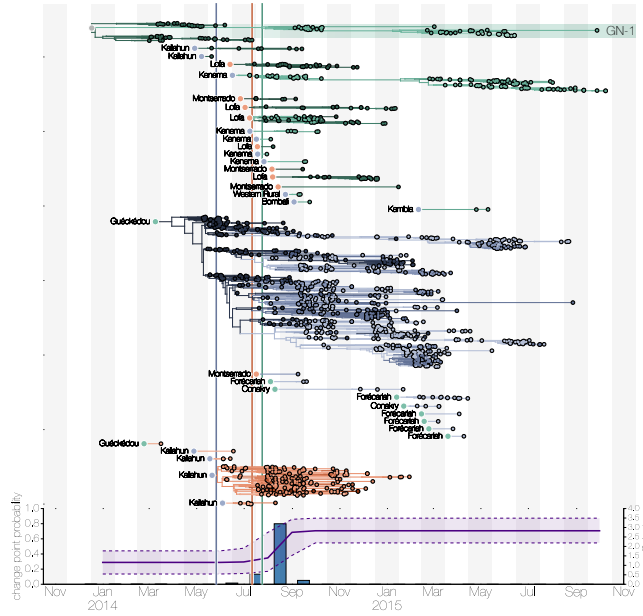
**funding:** NIH R01-AI107034 and NIH U54 GM111274

# Motivating phylodynamics by example

- We start with sequence data and sampling times
- In some cases the tree and/or dates of branching events are of interest
- Often phylogeny is a nuisance parameter
- Common objective 1: to estimate changes in population size
- Common objective 2: impute unobserved phenotype (“spatial location can be a “phenotype”) at internal nodes

Ebola in West Africa




From Dudas et al. (Nature, 2017)

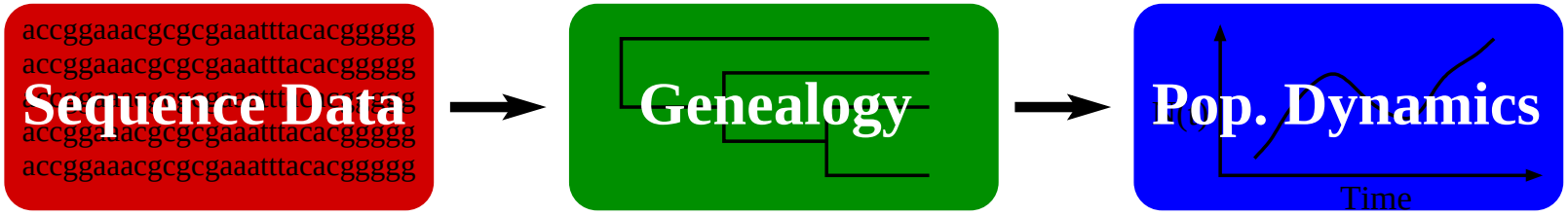
# When a Bayesian hat fits better

- You have an estimation "pipeline:"
  data $\rightarrow$ estimates/pseudo-data $\rightarrow$ more estimates

  The joint model that does everything simultaneously has better statistical properties

- You have latent variables, missing data, nuisance parameters
- You need to compare non-nested models
- You want to use multiple models to make predictions
- You want to build a model with parametric and nonparametric components
- You have prior information about some parameters of your model

# Big picture

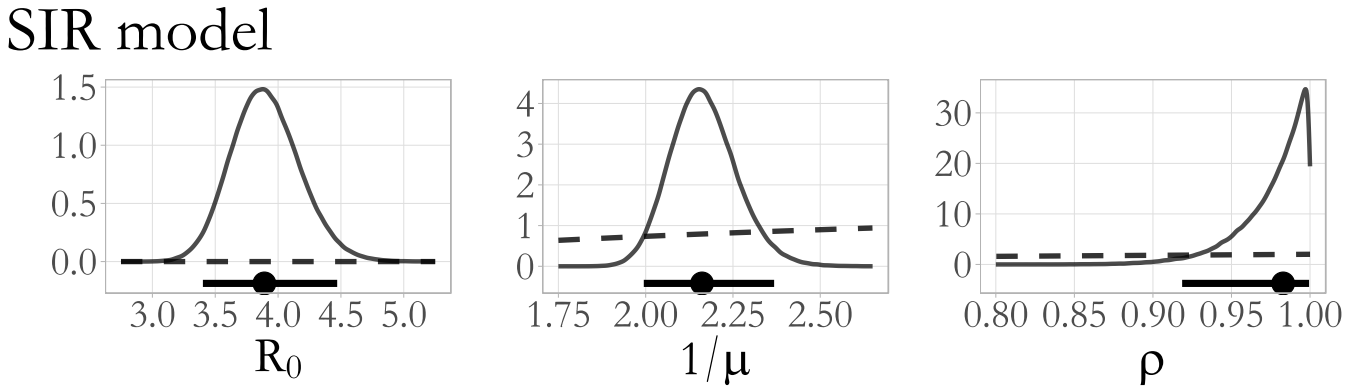



$$P(\mathbf{G}, \mathbf{Q}, N_e(t), \theta \mid \mathbf{D}) \propto$$
$$P(\mathbf{D} \mid \mathbf{G}, \mathbf{Q})\, P(\mathbf{Q})\, P(\mathbf{G} \mid N_e(t))\, P(N_e(t) \mid \theta)\, P(\theta)$$

- **G** - genealogy with branch lengths
- $N_e(t)$ - effective population size trajectory
- $P(\mathbf{G} \mid N_e(t))$ - **coalescent prior**
- $P(N_e(t) \mid \theta)$ - model for population size changes
- **Q** - substitution matrix
- **D** - sequence data
- $P(\mathbf{D} \mid \mathbf{G}, \mathbf{Q})$ - phylogenetic likelihood
- $\theta$ - hyper-parameters

# General thoughts about priors

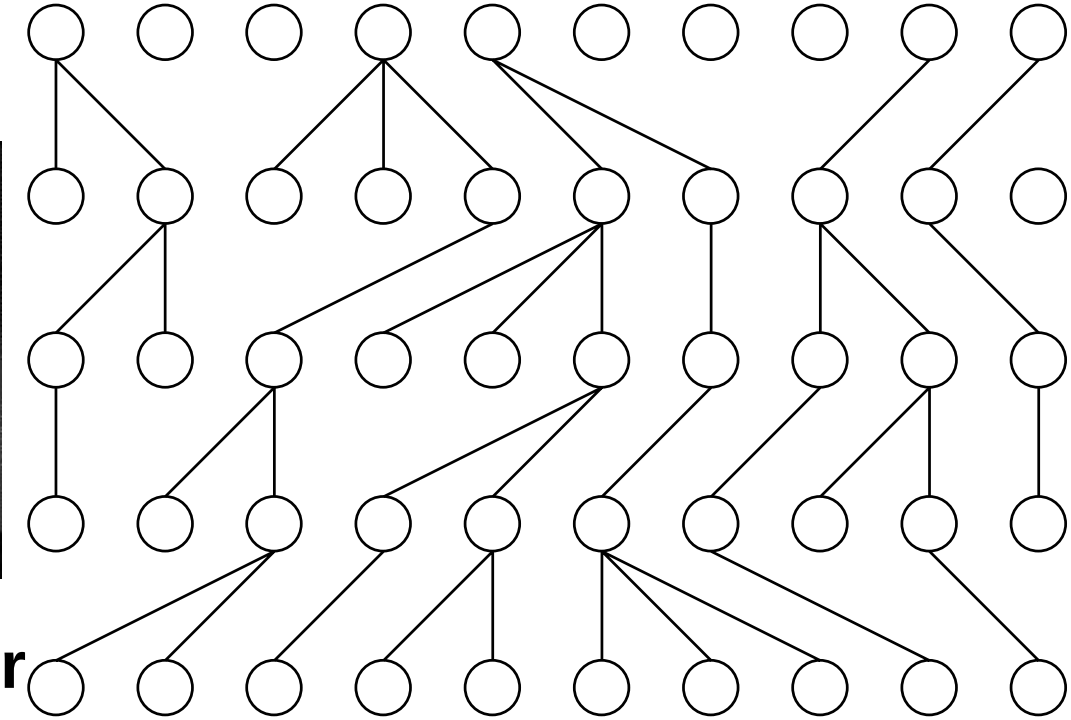
- Think about plausible ranges of your parameters
- Simulate from priors when they become complex and multivariate
- For key parameters of interest, plot posterior and prior distributions together

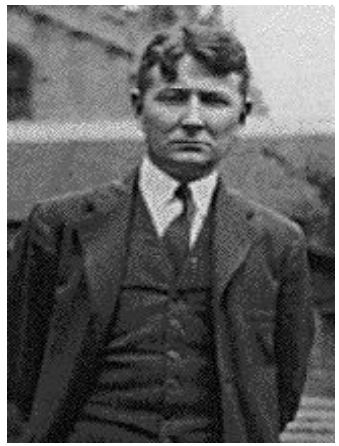



- Mess your priors and see how sensitive your posterior is to these purturbations

# Wright-Fisher Reproduction Model

**R.A. Fisher**

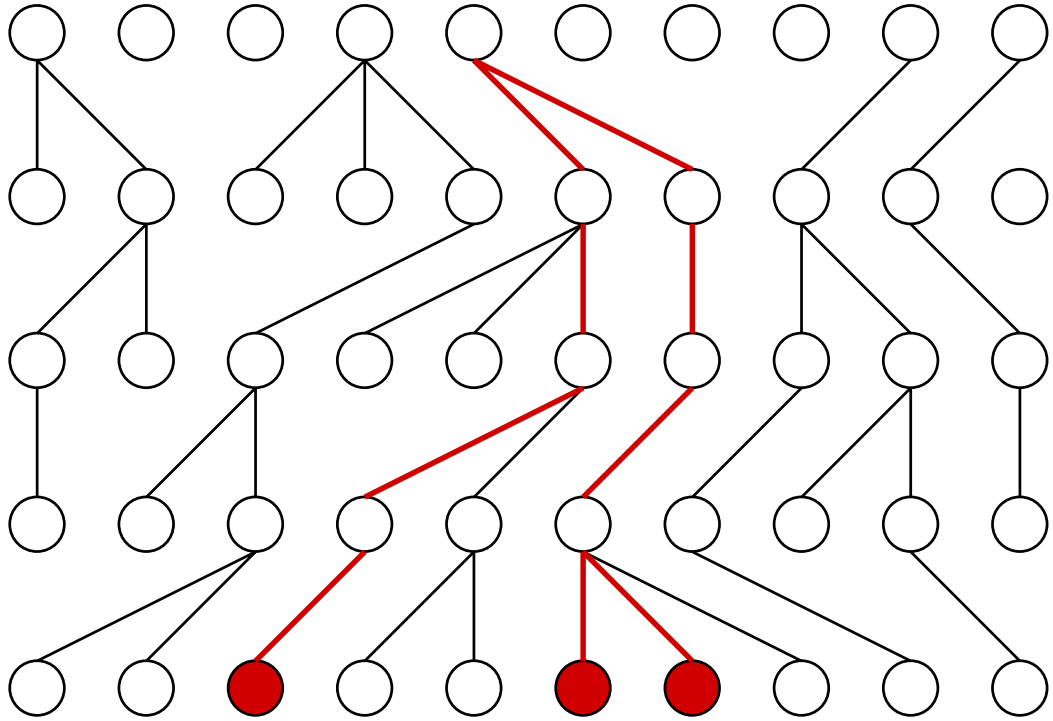


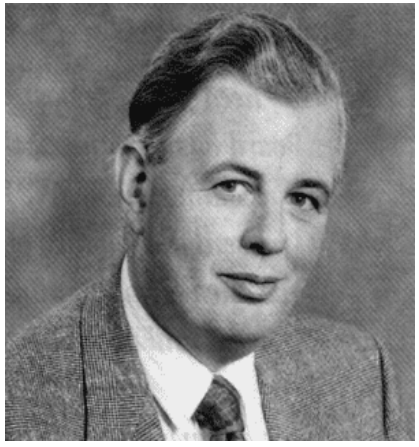


**S. Wright**

- Constant population size = 2N,
- Non-overlapping generations
- “Genes” are sampled randomly with replacement
- Population has no geographical or social structure
- No recombination, no selection

# Kingman's Discrete-Time Coalescent



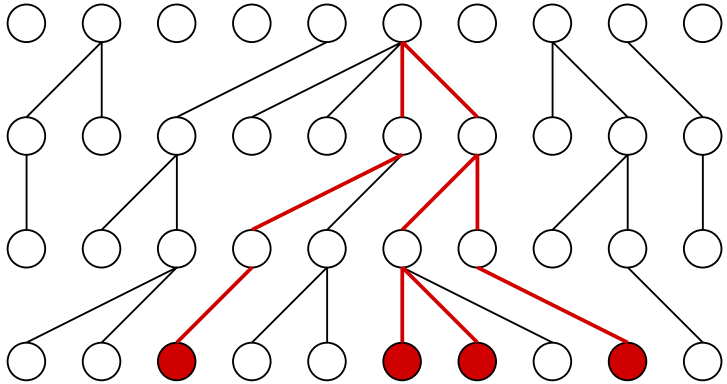


**J.F.C. Kingman**

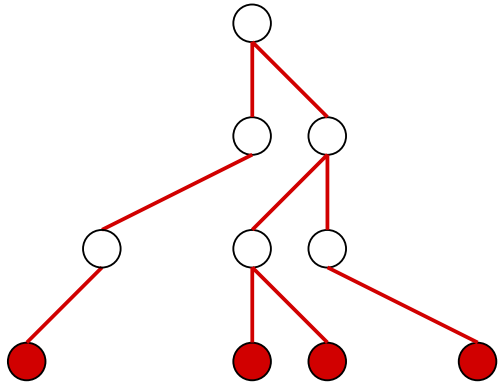
- 2 genes coalesce in one generation w/ prob. $\frac{1}{2N}$
- 2 genes coalesce in $j$ generations w/ prob. $\left(1-\frac{1}{2N}\right)^{j-1}\frac{1}{2N}$
- $k$ genes coalesce in $j$ generations w/ prob.

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

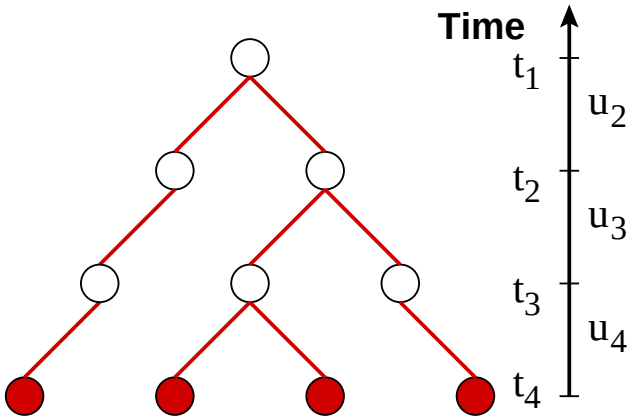
# Continuous-Time Coalescent



# Continuous-Time Coalescent

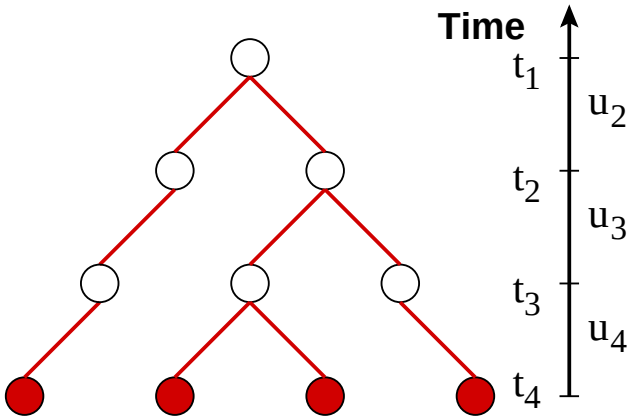


# Continuous-Time Coalescent




- Time measured in *N* generation units
- N = const $\rightarrow u_k \sim \text{Exp}\left[\binom{k}{2}\right]$ + indep.

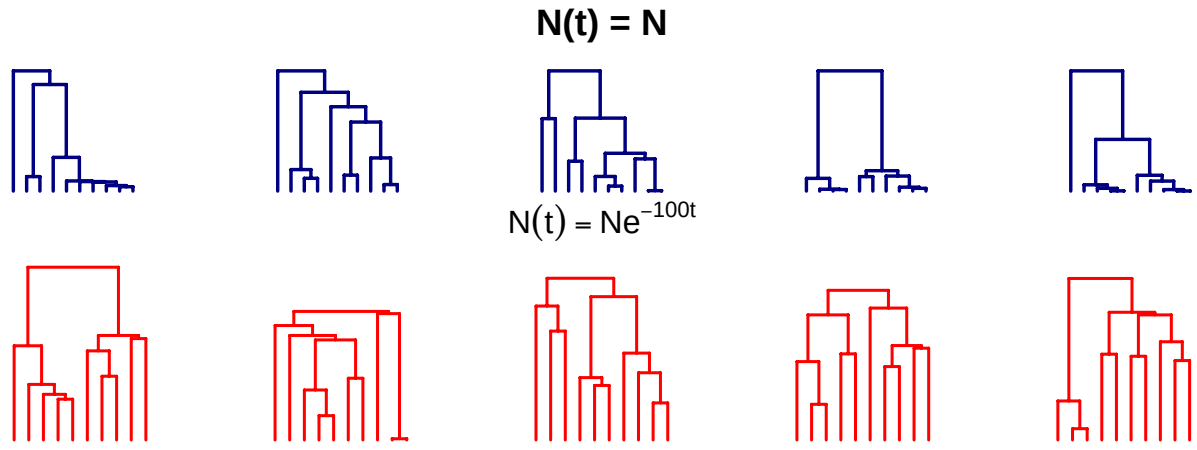
# Continuous-Time Coalescent




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

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

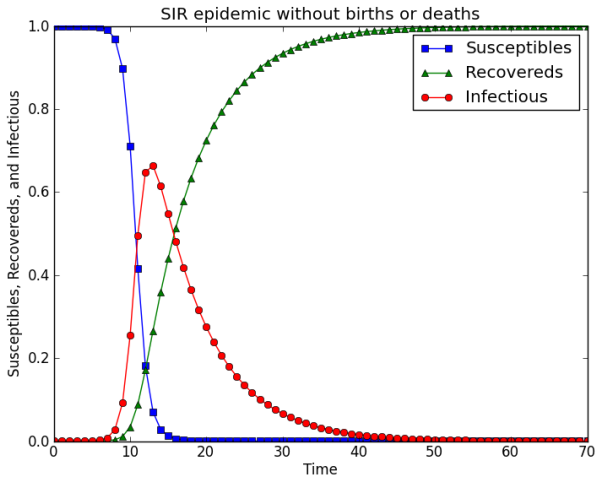
- Constant population size
- Exponential growth



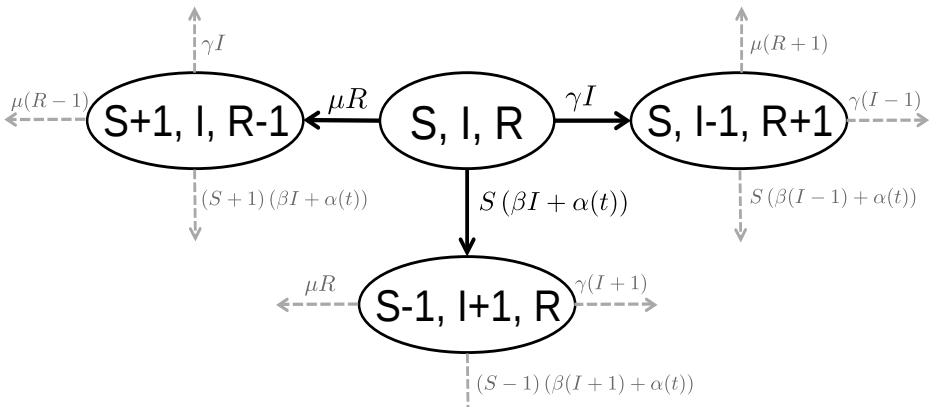
# Examples of parametric demographic models

- Exponential growth: $N_e(t) = ae^{bt}$, $a$ and $b$ are hyper-parameters
- Logistic growth: $f(x) = \frac{a}{1+e^{b(x-x_0)}}$, $a$, $b$, and $x_0$ are hyper-parameters
- ODE-based SIR model

$$\frac{dS}{dt} = -\beta \frac{SI}{N},$$
$$\frac{dI}{dt} = \beta \frac{SI}{N} - \gamma I,$$
$$\frac{dR}{dt} = \gamma I.$$



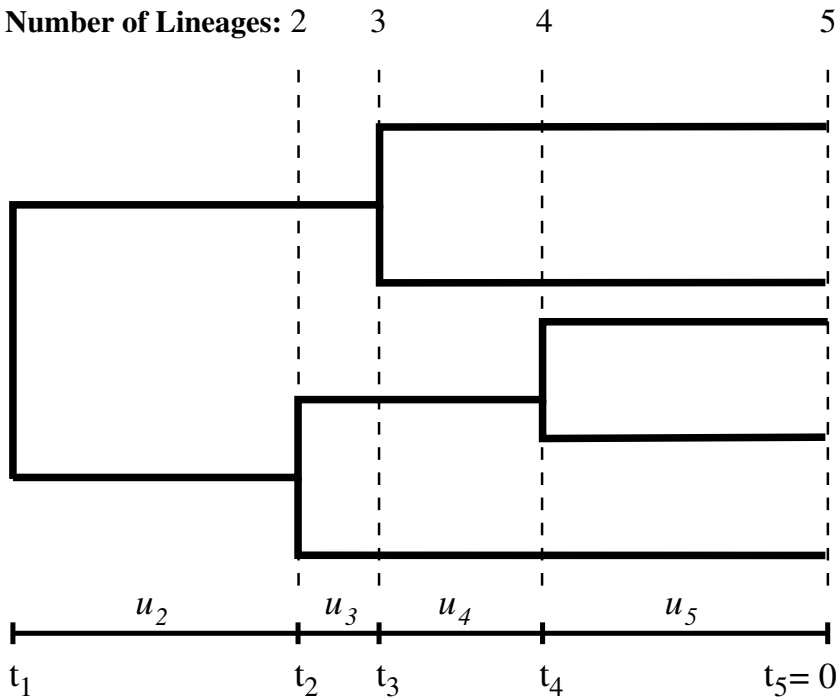

- Stochastic SIRS model:



# Nonparametric demographic models

- Parametric models are attractive, because they have a small number of interpretable parameters (e.g., rate of exponential growth)
- But parametric models are not flexible enough to capture surprising data features (e.g., you will never see a population decline in an exponential growth model)
- Nonparametric models offer an alternative
- In Statistics, nonparametric usually means highly parametric. Often, the number of parameters grows as one adds more data points
- Let's look at a couple of ways one can approach nonparametric effective population size estimation

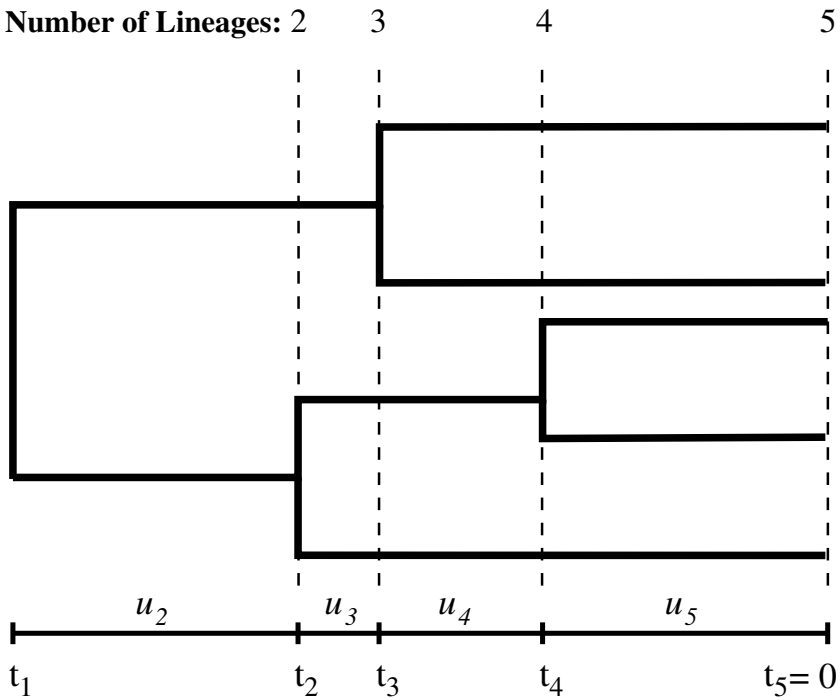
# Piecewise constant demographic models




## Isochronous data

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

# Piecewise constant demographic models



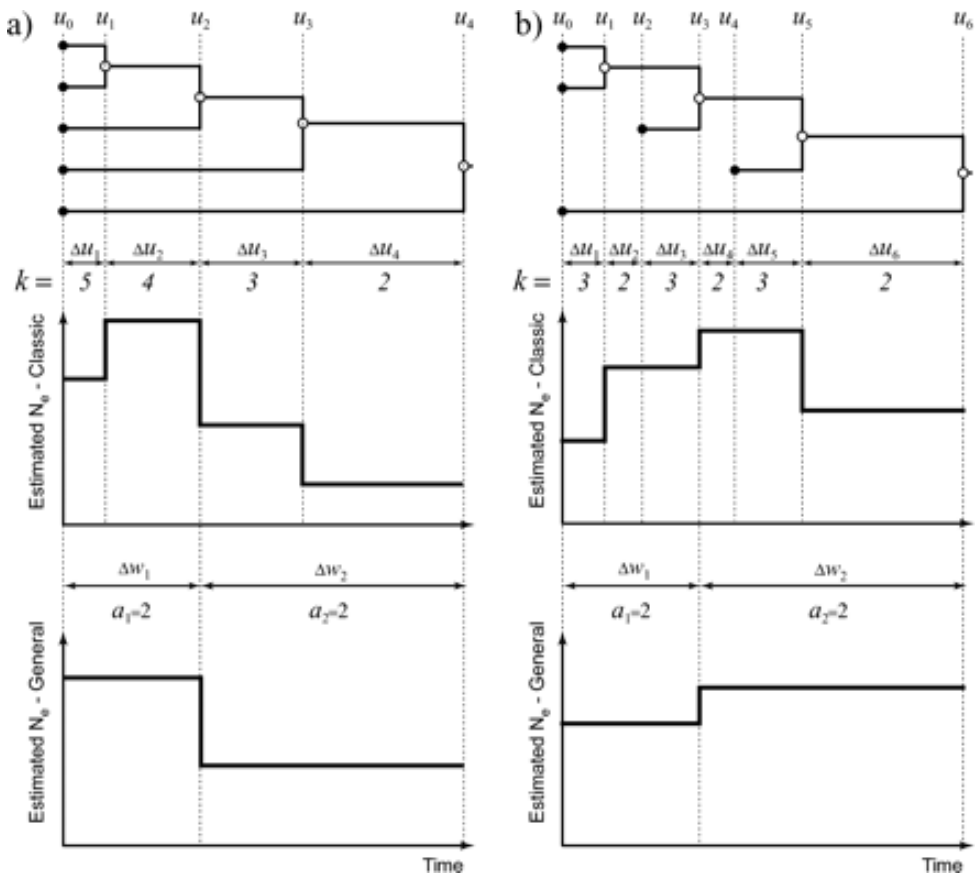

## Isochronous data

- $N_e(t) = \theta_k$ 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 — this is called a classical Skyline plot.
- Need further restrictions to estimate $\theta$!
- For example, Strimmer and Pybus (2001) made $N_e(t)$ constant across some consecutive inter-coalescent times using AIC.

# Bayesian Skyline

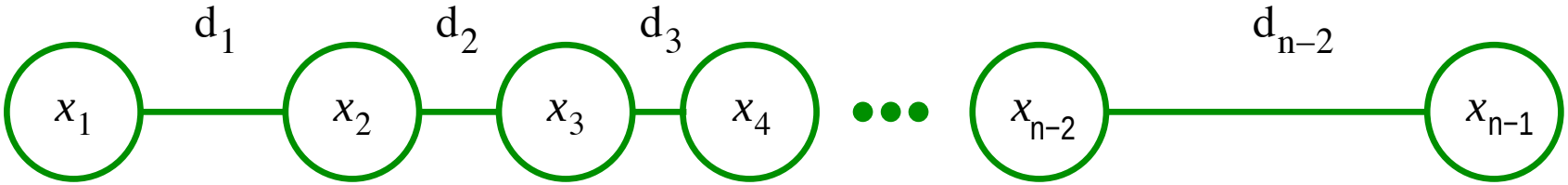
- Drummond et al. (2005) introduced a Bayesian multiple change-point model with a fixed number of change-points
- Change-points allowed only at coalescent events
- The model jointly estimates phylogeny and population dynamics




- Extended Bayesian Skyline makes the number of change-points an unknown parameter and estimates it together with everything else.

# Bayesian Skyride

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



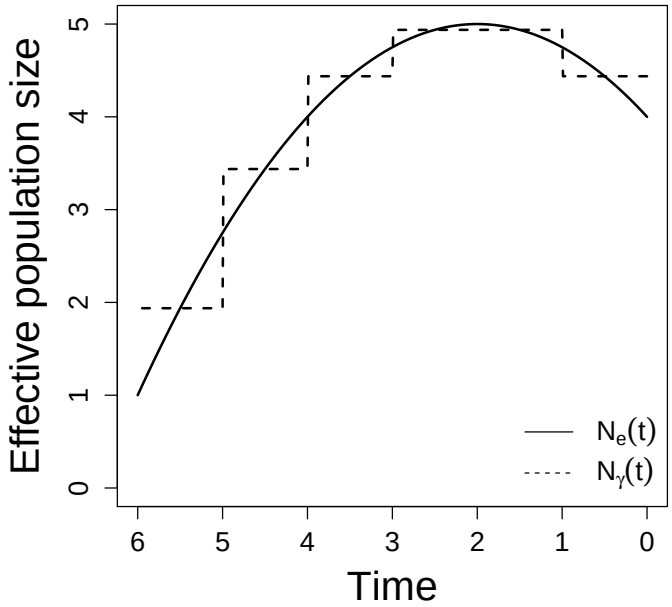

## Weighting Schemes

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

# Bayesian Skygrid

- Palacios and Minin (2012) and Gill et al. (2013) proposed a regular grid approach
- Construct fine, regular grid $\mathbf{x} = \{x_j\}_{j=1}^{k}$ with grid width $w$
- Let $\gamma_j = N_e(x_j)$




- Construct piecewise constant approximation

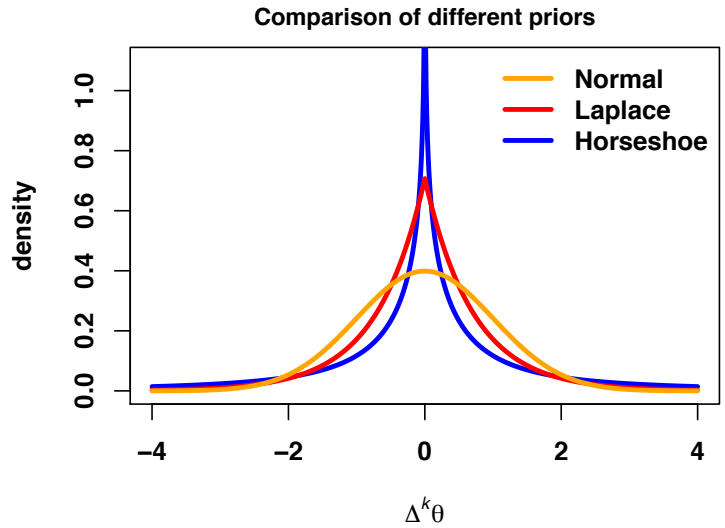
$$N^{\gamma}(t) = \sum_{i=1}^{k} \gamma_i 1_{t \in [x_i - w/2, x_i + w/2)}$$

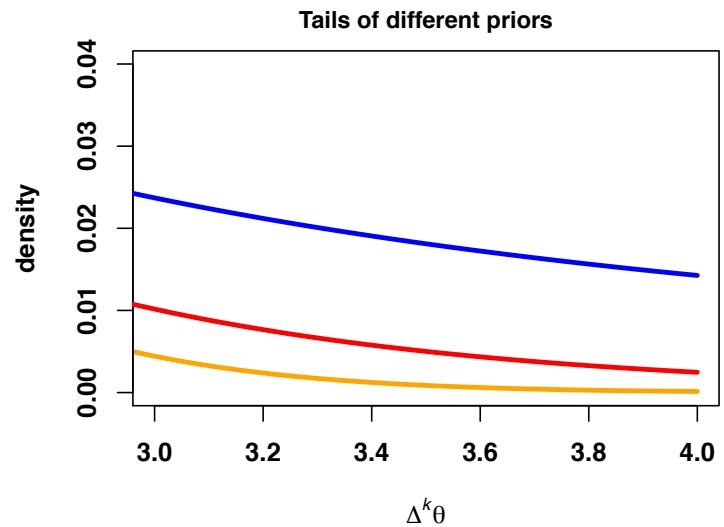
- Use the same Gaussian Markov random field prior as Skyride
- Skygrid allows to model $N_e(t)$ as a log-linear function of external covariates

# Preview of ongoing work: adaptive Skygrid

Our default temporal smoothing technique is a Brownian motion prior, which on a grid translates to a random walk of order 1. Sometimes we use higher order random walks:

- $\gamma_i = \log[N_e(t_i)]$, $\Delta^k \gamma_i \sim \mathcal{N}(0, \tau^2)$, where $\Delta^k$ is a $k$th-order forward difference operator.
- Adaptive version of the random walk prior:
- $\gamma_i = \log[N_e(t_i)]$, $\Delta^k \gamma_i \sim \text{Horseshoe}(0, \tau)$.



# Summary of Sky[line,ride,grid] methods

- Bayesian skyline
  - Statistical pros: good at detecting discontinuous jumps
  - Statistical cons: coalescent grid can produce strange artifacts
  - Computational pros: easy to implement, no need for fancy pants MCMC
  - Computational cons: slow convergence/mixing when number of change-points in high
  - Extended Bayesian skyline can handle multiple loci
- Bayesian skyride
  - Statistical pros: good at modeling gradual changes
  - Statistical cons: coalescent grid can produce strange artifacts, too smooth
  - Computational pros: efficient MCMC based on Gaussian tricks
  - Computational cons: implementation is not straightforward
- Bayesian skygrid
  - Statistical pros: good at modeling gradual changes
  - Statistical cons: setting up a grid could be tricky, too smooth
  - Computational pros: efficient MCMC based on Gaussian tricks
  - Computational cons: implementation is not straightforward
  - Skygrid can handle multiple loci
  - Effective population size can be a function of external time-varying covariates

# Case study I: Egyptian HCV

World Health Organization

عربي | 中文 | English | Français | Русский | Español

Search

All WHO | This site only

Home
About WHO
Countries
Health topics
Publications
Data and statistics
Programmes and projects
GAR Home
Alert & Response Operations
Diseases
Global Outbreak Alert & Response Network
Biorisk Reduction

Global Alert and Response (GAR)

Country activities | Outbreak news | Resources | Media centre

WHO > Programmes and projects > Global Alert and Response (GAR) > Diseases covered by EPR > Hepatitis

printable version

**The role of parenteral antischistosomal therapy in the spread of hepatitis C virus in Egpypt**

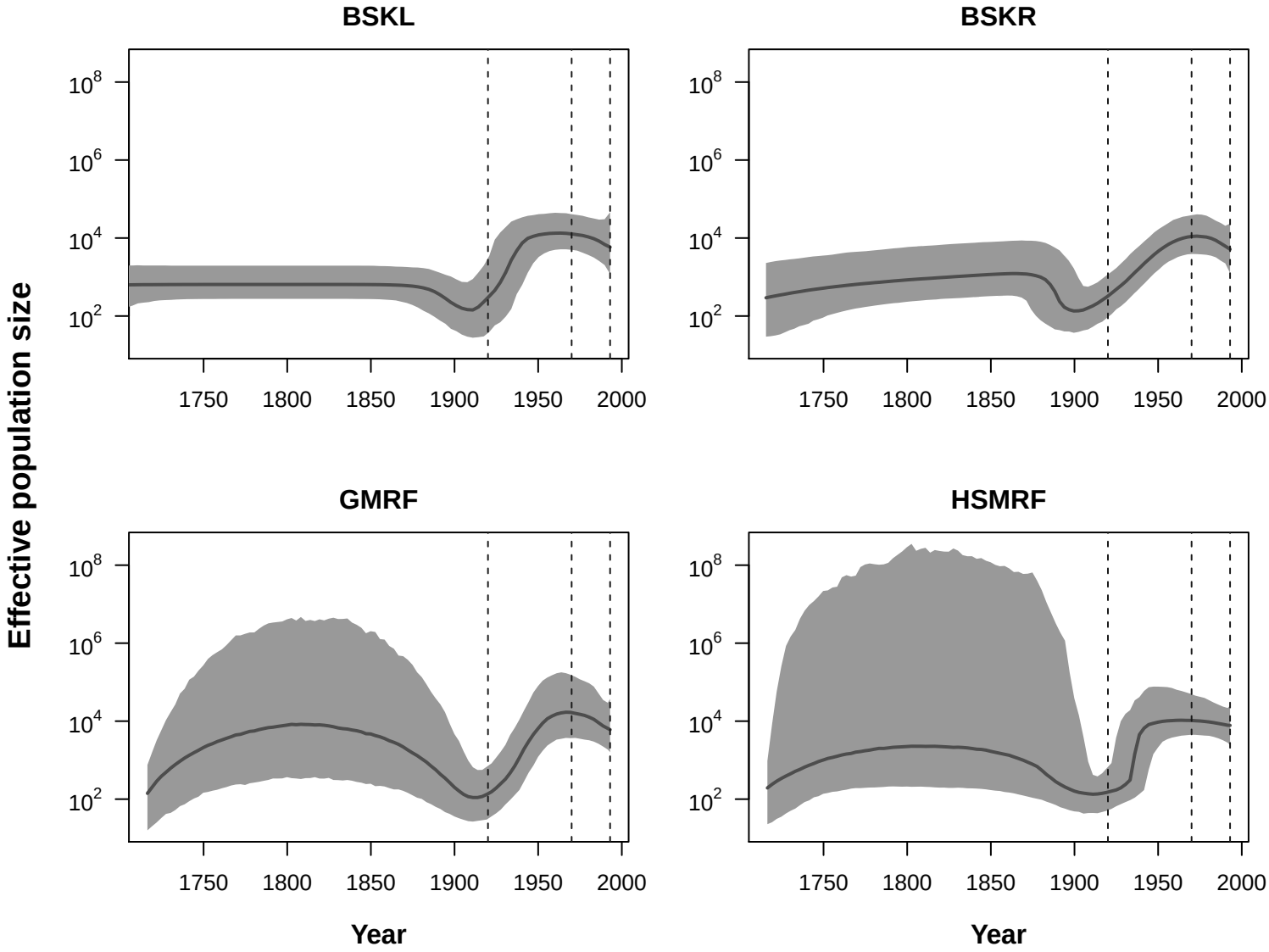
*Christina Frank, Mustafa K Mohamed, G Thomas Strickland, Daniel Lavanchy, Ray R Arthur, Laurence S Magder, Taha El Khoby, Yehia Abdel-Wahab, El Said Aly Ohn, Wagida Anwar, Ismail Sallam*

**Summary**

**Background** The population of Egypt has a heavy burden of liver disease, mostly due to chronic infection with hepatitis C virus (HCV). Overall prevalence of antibody to HCV in the general population is around 15-20%. The risk factor for HCV transmission that specifically sets Egypt apart from other countries is a personal history of parenteral antischistosomal therapy (PAT). A review of the Egyptian PAT mass-treatment campaigns, discontinued only in the 1980s, show a very high potential for transmission of blood-borne pathogens. We examine the relative importance of PAT in the spread of HCV in the Egypt.

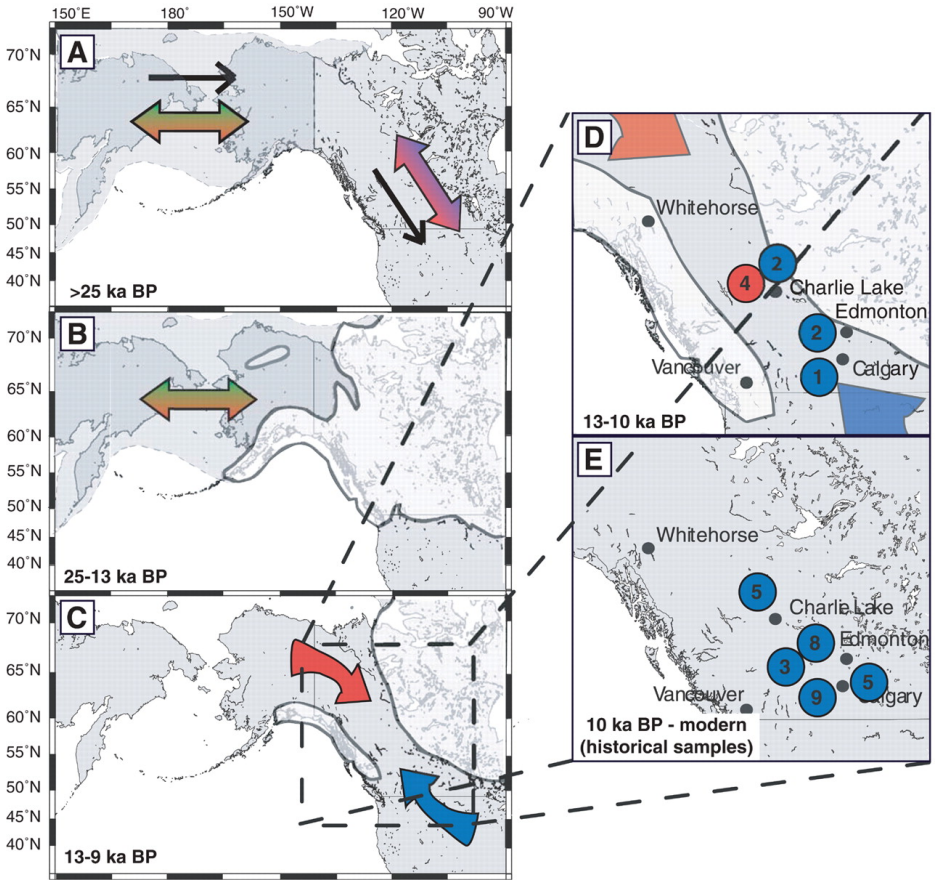
- 63 molecular sequences sampled at random from the population
- Parenteral antischistosomal therapy (PAT) was practiced from 1920s to 1980s
- In the 1970s started a transition from the intravenous to the oral administration of the PAT

# Egyptian HCV results



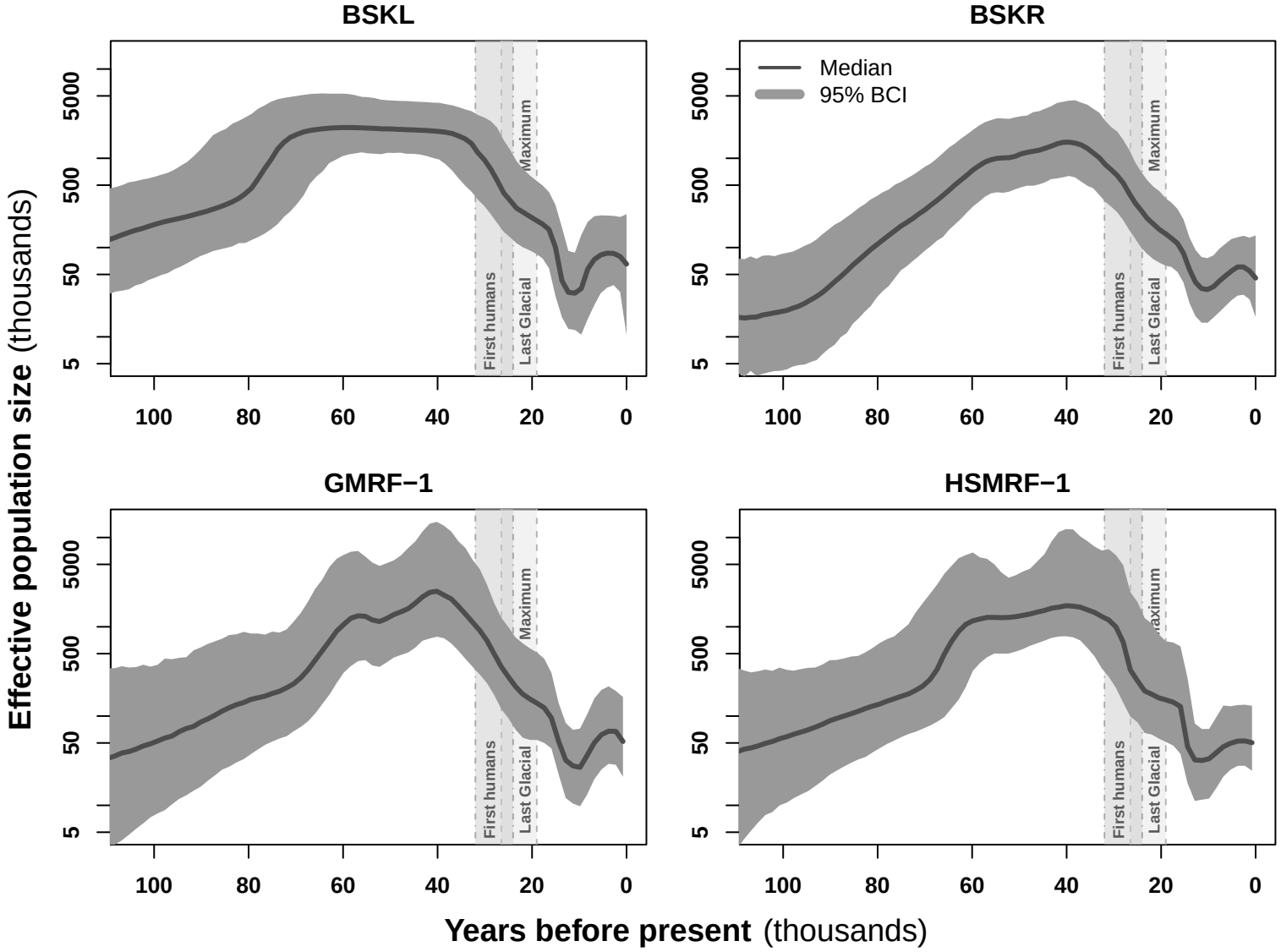

BSKL = Skyline, BSKR = Skyride, GMRF = Skygrid,
HSMRF = Adaptive Skygrid

# Case study II: Beringian steppe bison




- The data are 152 sequences (135 ancient and 17 modern) of mitrochondrial DNA with 602 base pairs
- DNA was extracted from bison fossils from Alaska (68), Canada (46), Siberia (13), the lower 48 United States (6), and China (2)

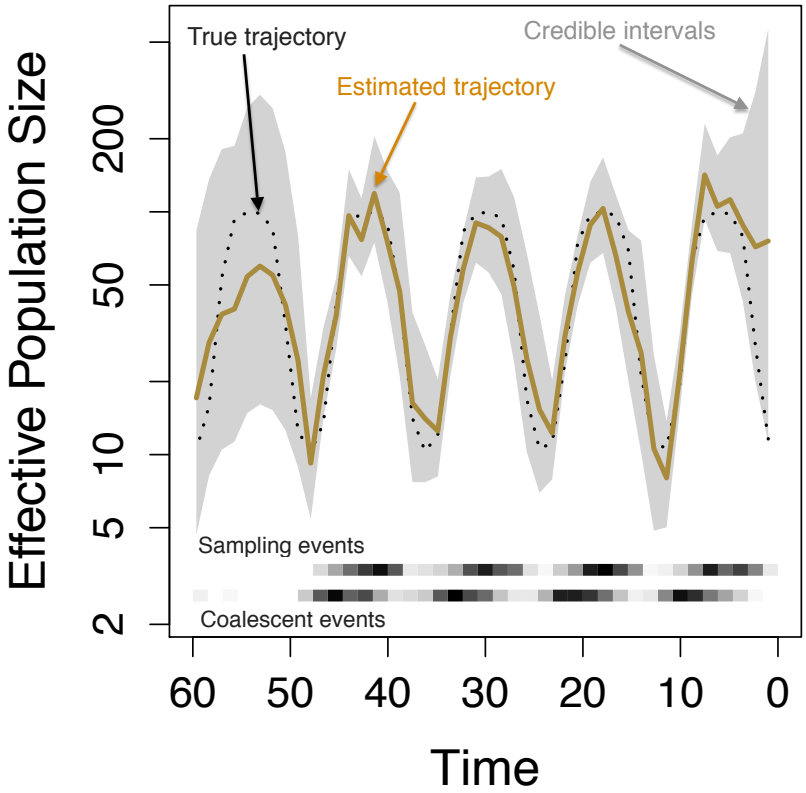
# Beringian steppe bison results




BSKL = Skyline, BSKR = Skyride, GMRF = Skygrid,
HSMRF = Adaptive Skygrid

# Preferential sampling

What if we sample more when incidence is high? This is called **preferential sampling** in spatial statistics (Diggle et al., 2010).




- **G** — genealogy with branch lengths
- $\gamma$ — heights of a piecewise constant effective population size trajectory
- $\tau$ — hyper-parameter controlling smoothness of the population size trajectory

**Preferential sampling model**:

$$P(\mathbf{G}, \mathbf{Q}, N_e(t), \theta \mid \mathbf{D}, \mathbf{s}) \propto$$
$$P(\mathbf{D} \mid \mathbf{G}, \mathbf{Q})\, P(\mathbf{s} \mid N_e(t), \beta)\, P(\mathbf{Q})\, P(\mathbf{G} \mid N_e(t))\, P(N_e(t) \mid \theta, \mathbf{s})\, P(\theta)\, P(\beta)$$

# Two point processes controlled by one rate

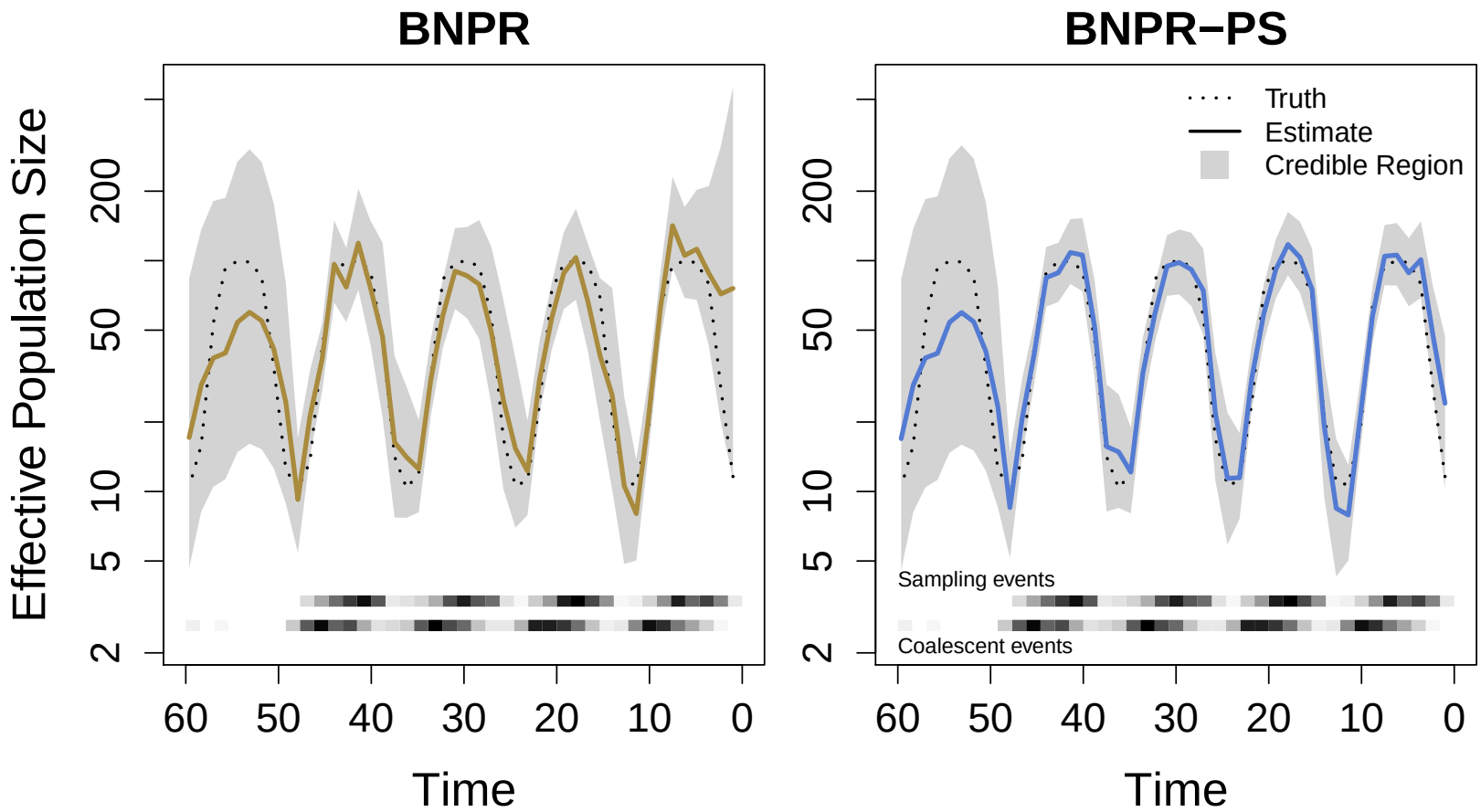
- The coalescent is a non-homogeneous continuous time Markov chain and can be viewed as a Markov point process on $[0, \infty)$.

$$\Pr(\mathbf{G} \mid \gamma) \propto \prod_{k=2}^{n} \frac{C_{0,k}}{N^{\gamma}(t_{k-1})} \exp\left[ -\sum_{i=0}^{m_k} \int_{I_{i,k}} \frac{C_{i,k}}{N^{\gamma}(t)} dt \right], \text{ where } C_{i,k} = \binom{n_{i,k}}{2}.$$

- Let's model sampling times as an inhomogeneous Poisson process with rate proportional to a power $(\beta_1)$ of $N^{\gamma}(t)$.

$$\Pr(\mathbf{s} \mid \gamma, \beta) \propto \prod_{i=1}^{n} [\beta_0 N^{\gamma}(s_i)]^{\beta_1} \exp\left[ -\int_{s_0}^{s_n} \beta_0 N^{\gamma}(t)^{\beta_1} dr \right]$$

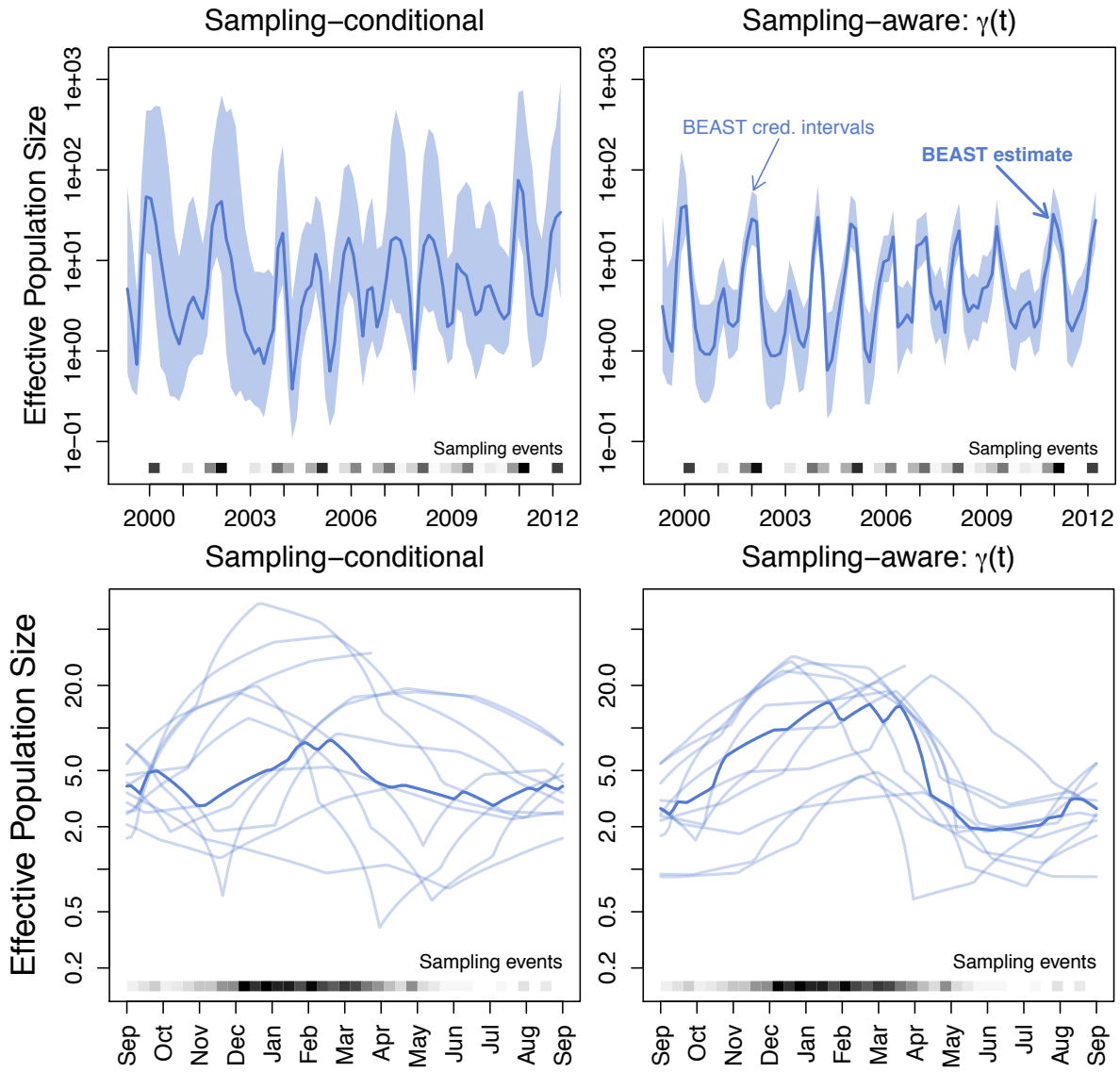
# Example of incorporating preferential sampling




**Summary of simulation study (not shown)**:

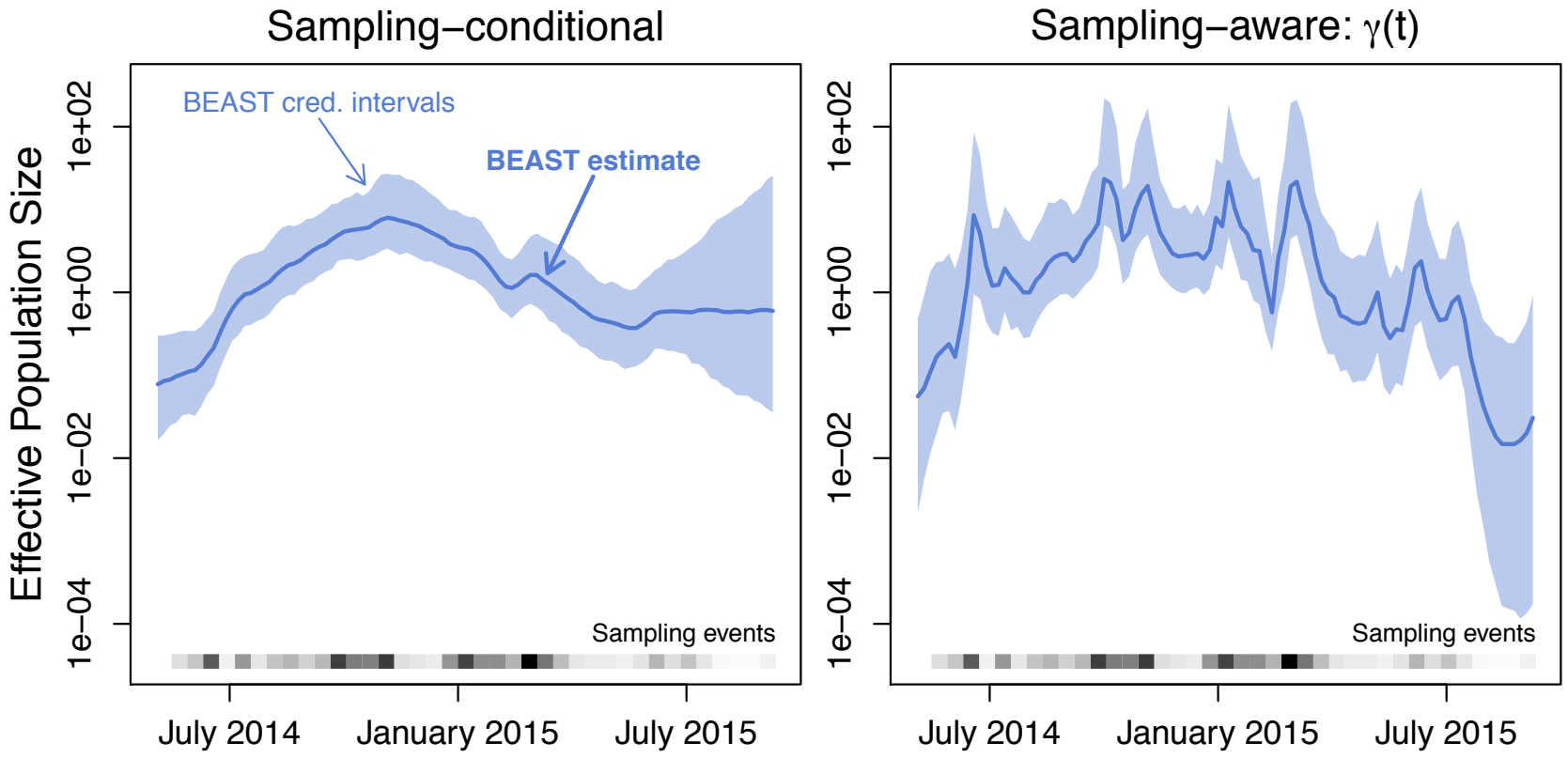
- Unrecognized preferential sampling can introduce systematic estimation bias, especially during the decline of population size
- In the presence of preferential sampling, modeling sampling times gives better accuracy and precision.

# Case Study III — human influenza



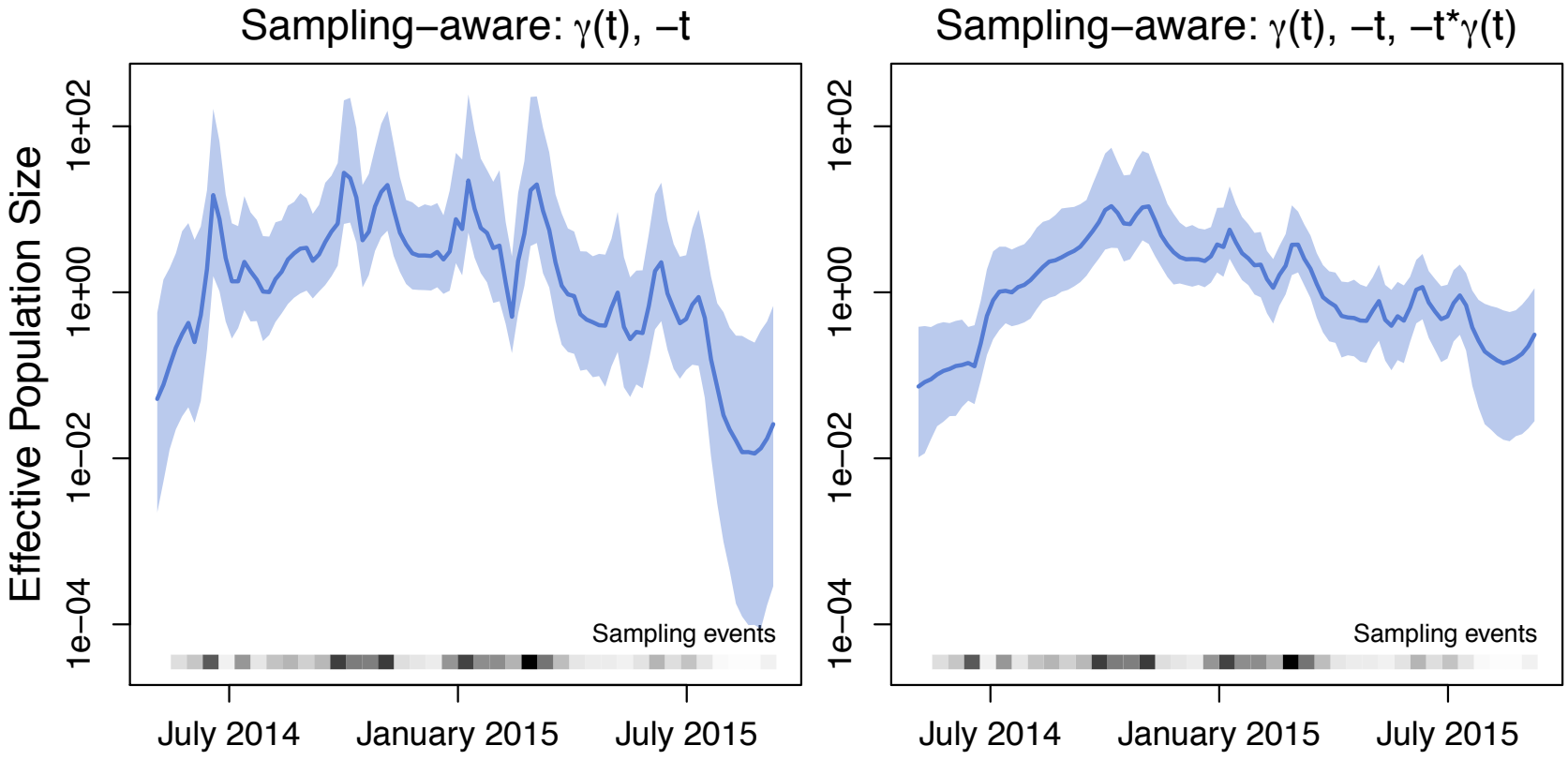

Subsample of the data assembled by Zinder et al. (2014).

# Case Study IV — Ebola in Sierra Lione



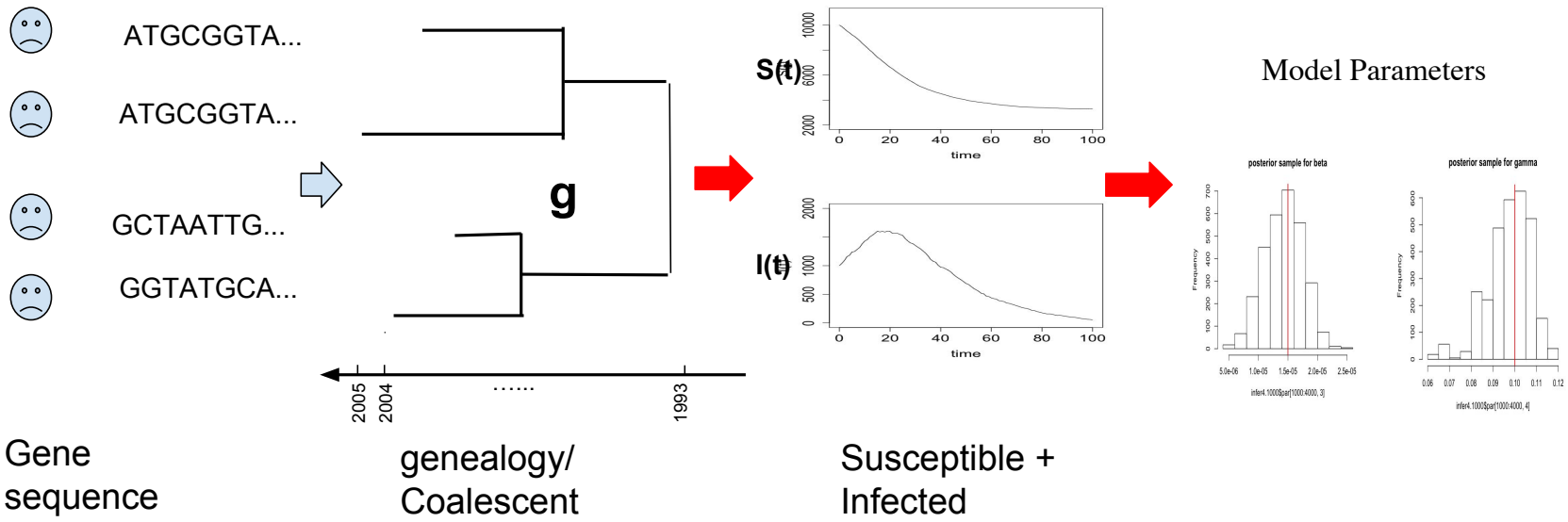

- Subsampled 200 sequence from Sierra Lione 2014-2015 outbreak
- This is an example of preferential sampling model being too simple
- Including time varying covariates can fix this

# Case Study IV — Ebola in Sierra Lione




- Subsampled 200 sequence from Sierra Lione 2014-2015 outbreak
- This is an example of preferential sampling model being too simple
- Including time varying covariates can fix this

# Fitting SIR




Gene sequence

genealogy/ Coalescent

Susceptible + Infected

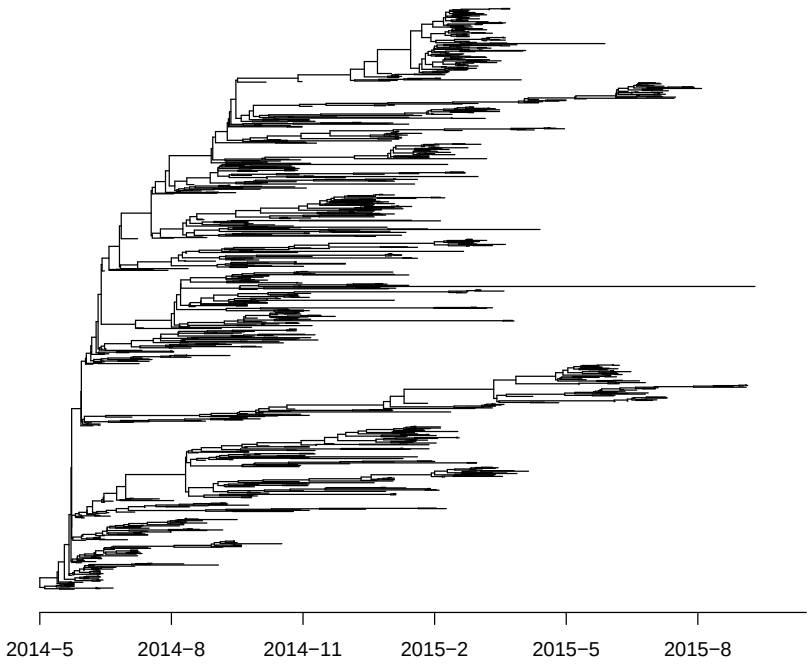
- Kingman's coalescent likelihood allows us to reconstruct the effective population size trajectory by connecting coalescent rates and the population size:

$$\lambda(t) = \frac{\binom{n_t}{2}}{N_e(t)}.$$

- Volz (2009) formulated a structured coalescent likelihood for compartmental models (e.g., SIR) by connecting coalescent rates and compartment sizes:

$$\lambda(t) = \frac{\binom{n_t}{2}}{\binom{I(t)}{2}} \cdot \beta S(t) I(t) \approx \binom{n_t}{2} \cdot \frac{2\beta S(t)}{I(t)}.$$

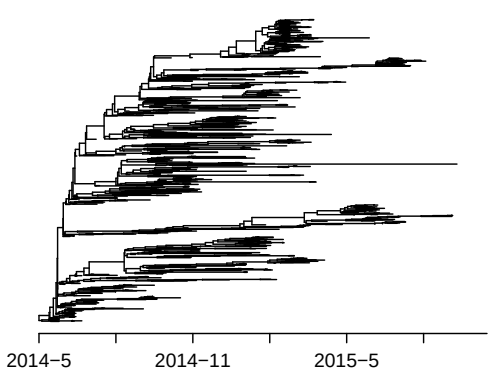
# Ebola in Sierra Lione set up

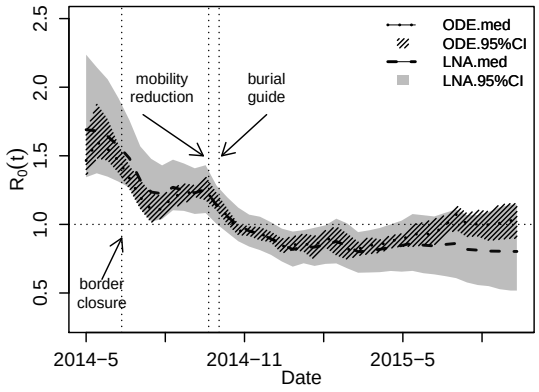


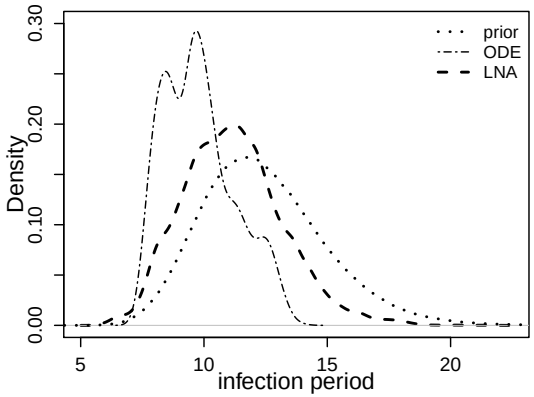

- The tree suggests a single introduction of Ebola.
- We analyzed 982 sequences from Sierra Lione (8% of all known cases).
- Time stamped tree was inferred using BEAST.

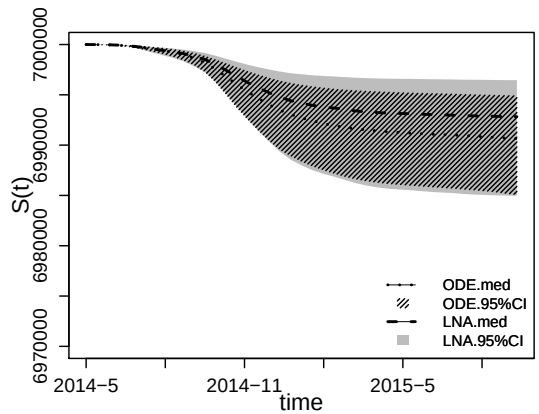
- We fit time inhomogeneous SIR model to the genealogy above. Time inhomogeneity comes from variable infection rate (or equivalently, variable basic reproduction number $R_0$).
- Objective: $\Pr(\mathbf{S}, \mathbf{I}, \mathbf{R}, \beta, \gamma | \mathbf{g})$, where $\beta$ is a set of infection rates, $\gamma$ is an infection rate, and $\mathbf{g}$ is a genealogy.

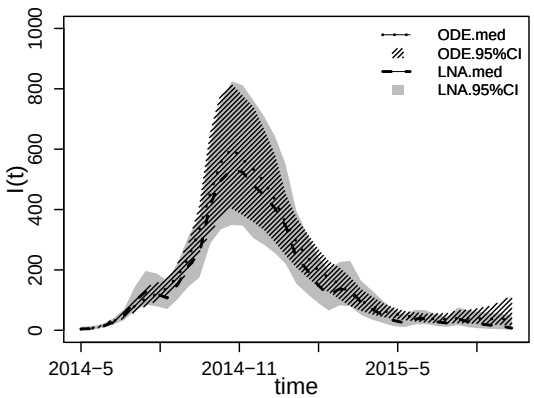
# Ebola in Sierra Lione results

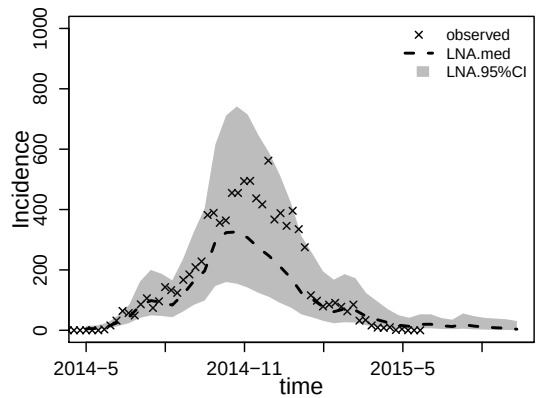


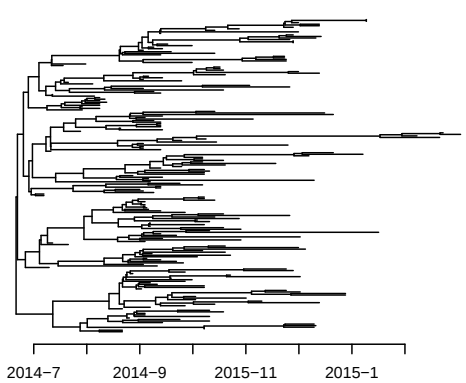



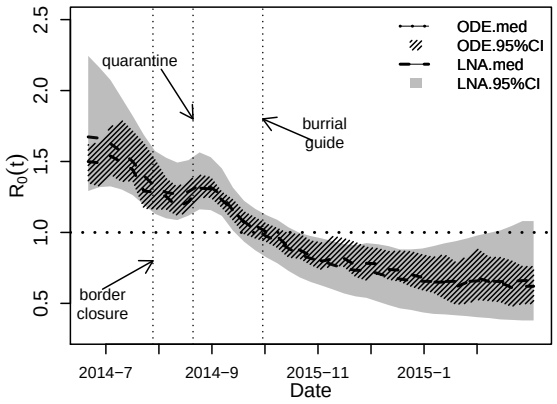



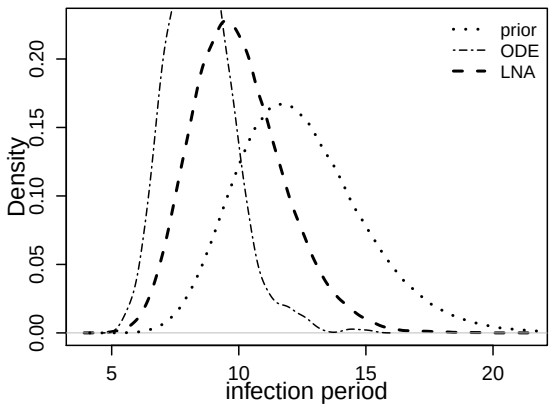



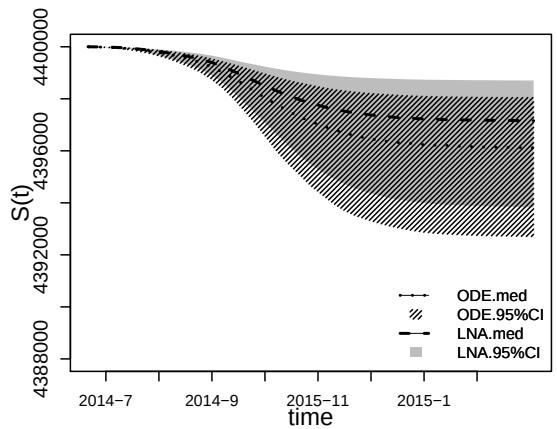

- The size of the epidemic (total number of infections $R_T$) is matching total number of known Sierra Lione cases.
- Number of cases started to decline when $R_0$ dropped below 1.
- SEIR model is more appropriate for Ebola, but Volz' coalescent likelihood is numerically unstable.

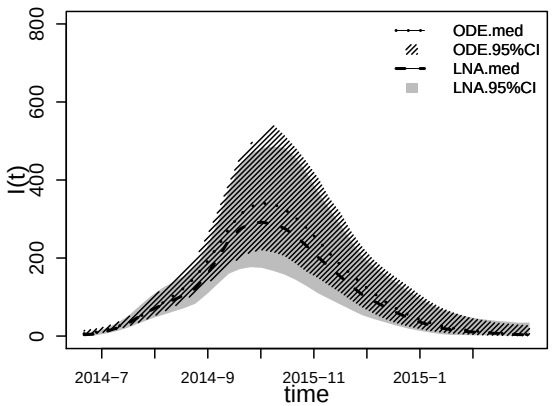
# Ebola in Liberia results

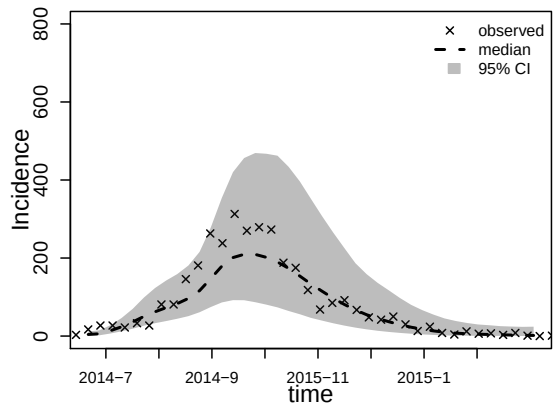



- The same analysis as above, but with $\sim$ 200 sequences from Liberia

# References


- Drummond AJ, Rambaut A, Shapiro B, Pybus OG. Bayesian coalescent inference of past population dynamics from molecular sequences. *Molecular Biology and Evolution* 22(5):1185–1192, 2005

- Minin VN, Bloomquist EW, Suchard MA. Smooth skyride through a rough skyline: Bayesian coales- cent-based inference of population dynamics. *Molecular Biology and Evolution*, 25(7):1459–1471, 2008.

- Gill MS, Lemey P, Faria NR, Rambaut A, Shapiro B, Suchard MA. Improving Bayesian population dynamics inference: a coalescent-based model for multiple loci. *Molecular Biology and Evolution*, 30(3):713–724, 2013.

- Karcher MD, Palacios JA, Bedford T, Suchard MA, and Minin VN. Quantifying and mitigating the effect of preferential sampling on phylodynamic inference. *PLoS Computational Biology*, 12:e1004789, 2016.

- Volz E, Siveroni I. Bayesian phylodynamic inference with complex models. *bioRxiv*, 2018.