SAMPLING-AWARE **PHYLODYNAMICS** WITH THE **BAYESIAN EPOCH SAMPLING SKYLINE PLOT**

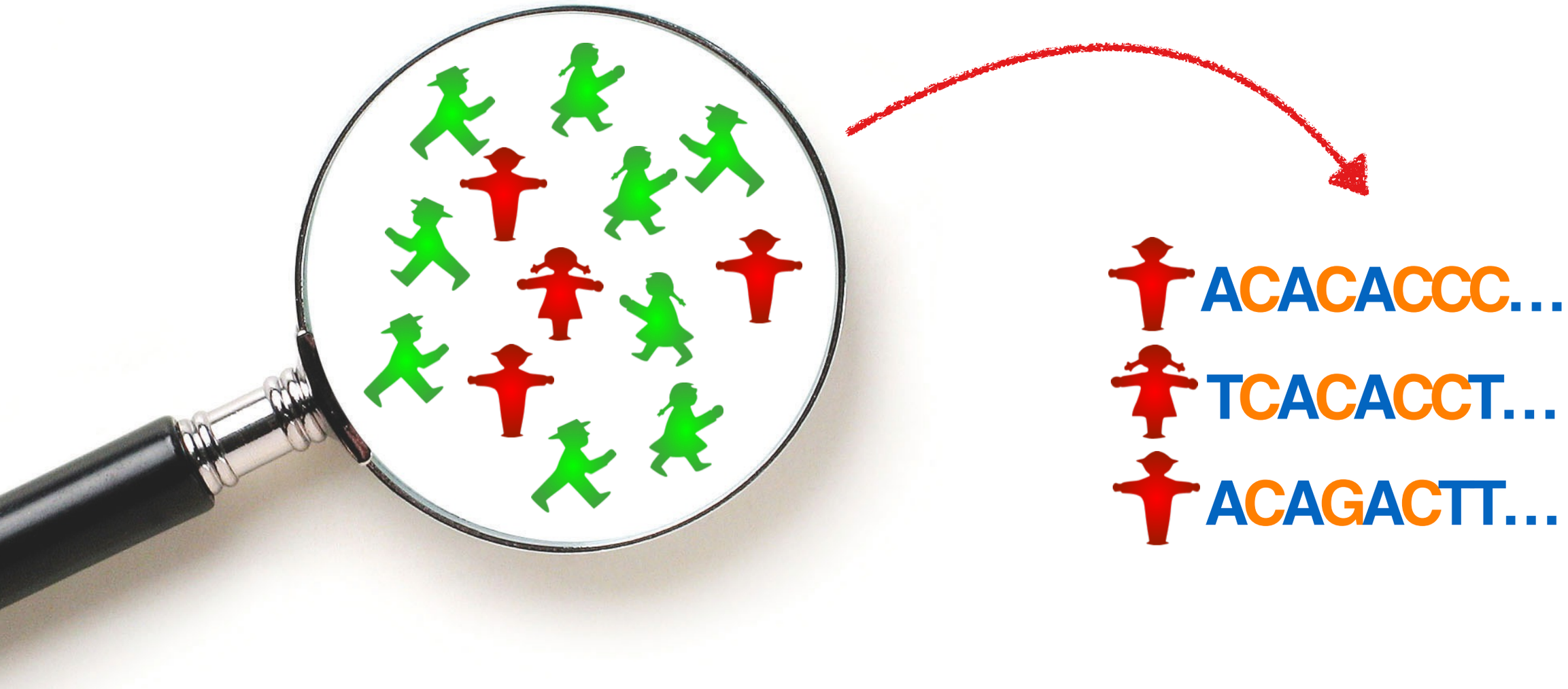
LOUIS DU PLESSIS, KRIS PARAG AND OLIVER PYBUS

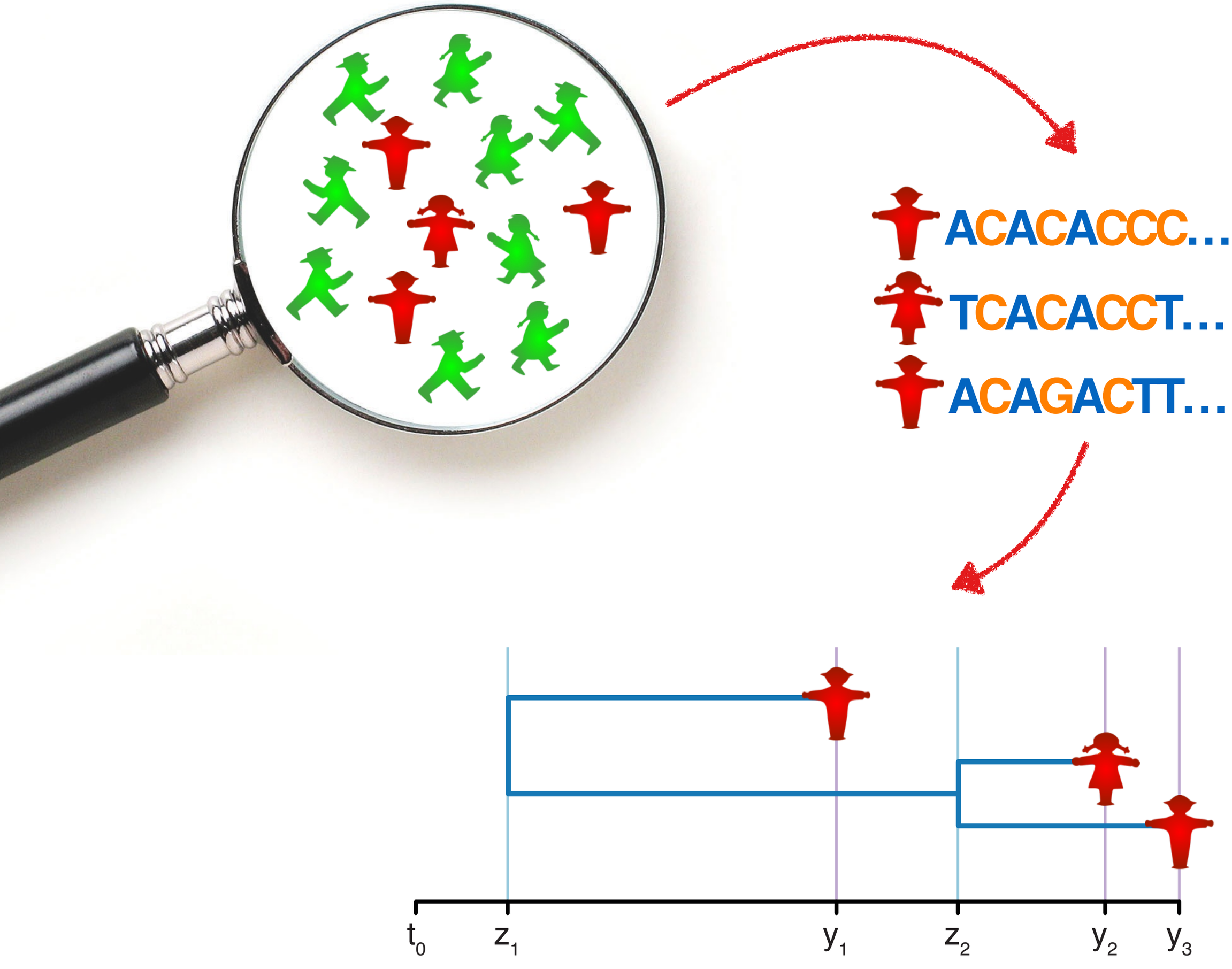
PHYLODYNAMIQUE

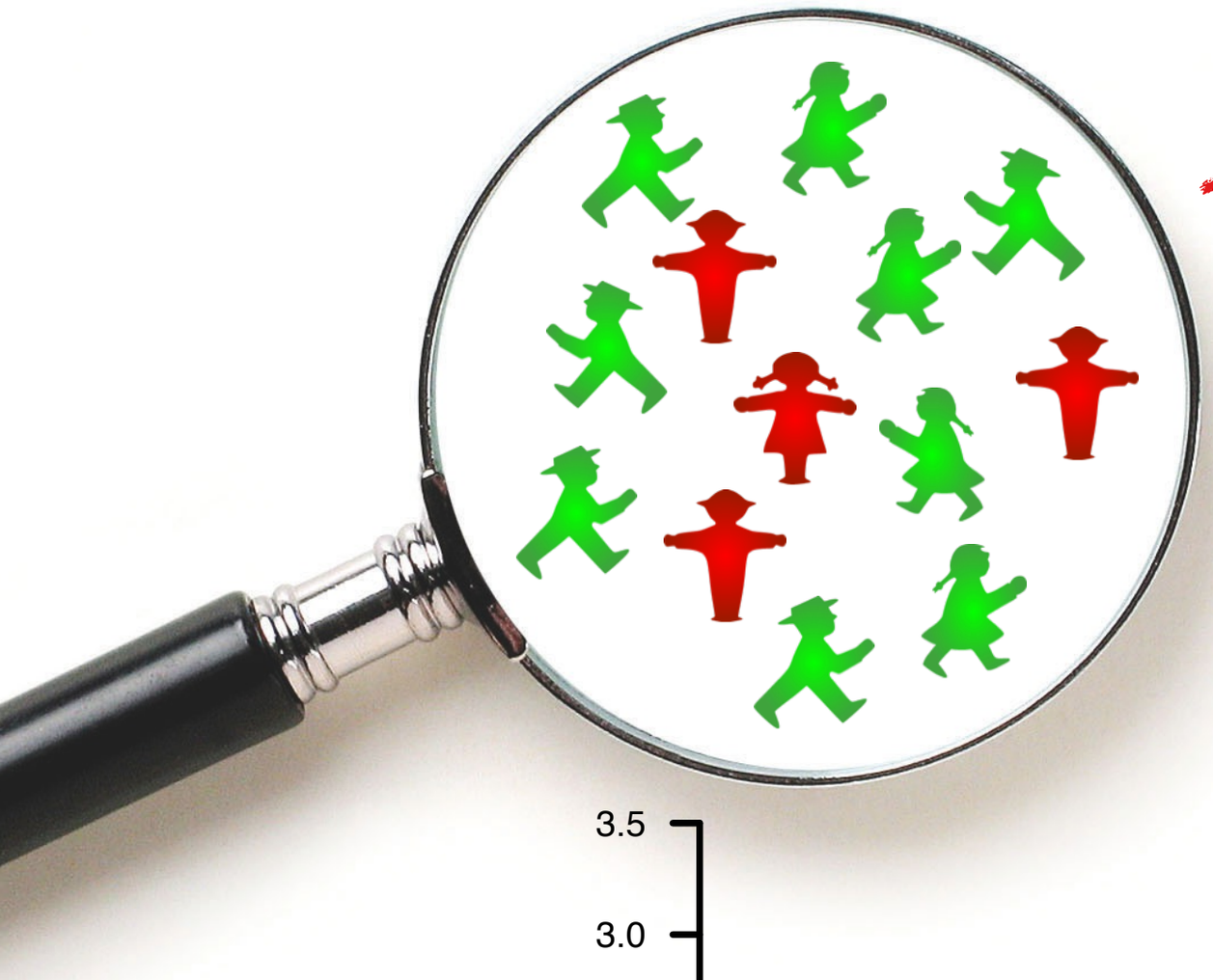
EN PRENANT EN COMPTE L'ÉCHANTILLONNAGE AVEC LE

BAYESIAN EPOCH SAMPLING SKYLINE PLOT

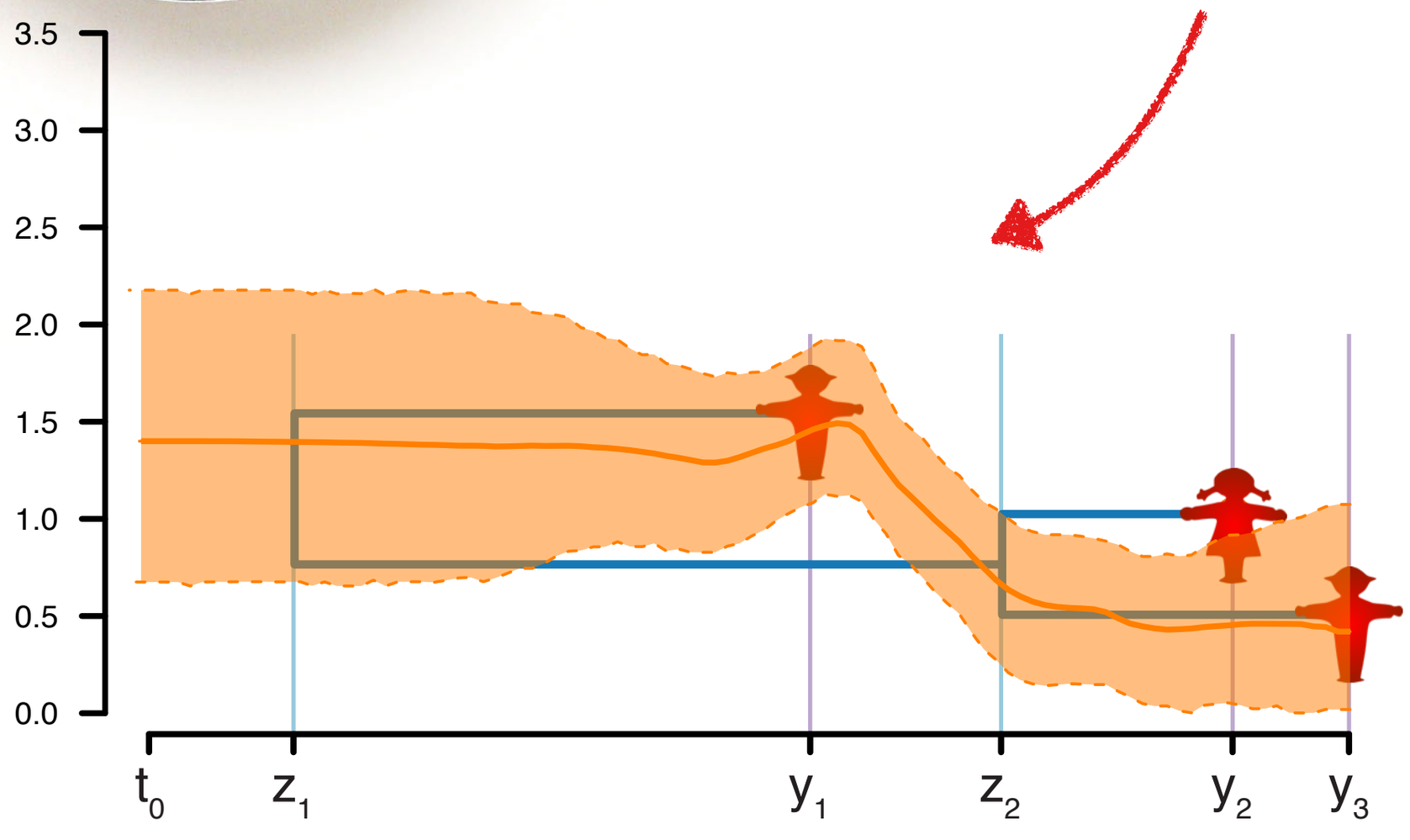
LOUIS DU PLESSIS, KRIS PARAG AND OLIVER PYBUS


ACACACCC…
TCACACCT…
ACAGACTT…


ACACACCC…
TCACACCT…
ACAGACTT…
t0
z1
y1
z2
y2
y3



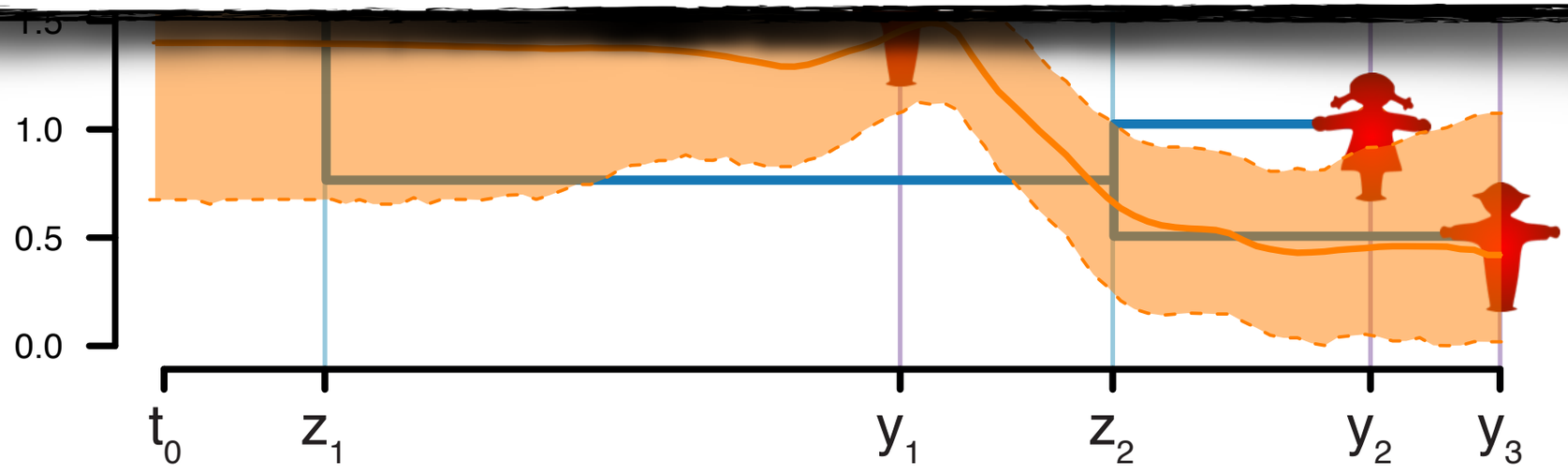

ACACACCC…
TCACACCT…
ACAGACTT…


3.5
3.0
2.5
2.0
1.5
1.0
0.5
0.0
t_0
z_1
y_1
z_2
y_2
y_3

Tree
Realisation of a stochastic process

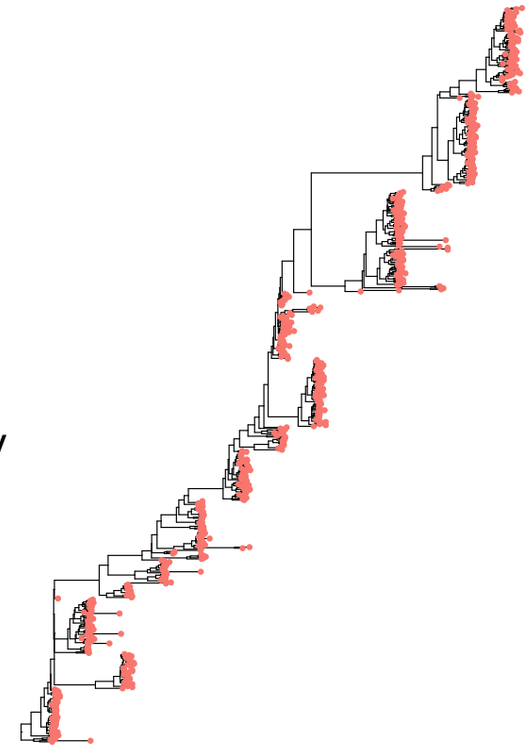
$P(\text{tree} \mid \text{demography})$

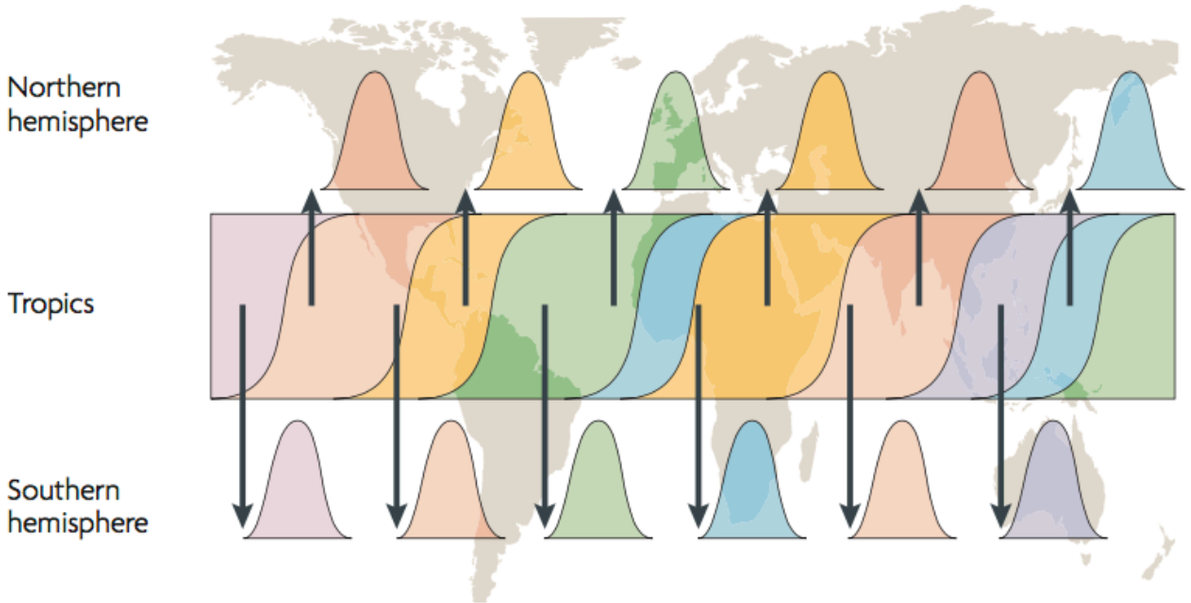
Demographic model
Describes the growth of the tree (spread of epidemic)

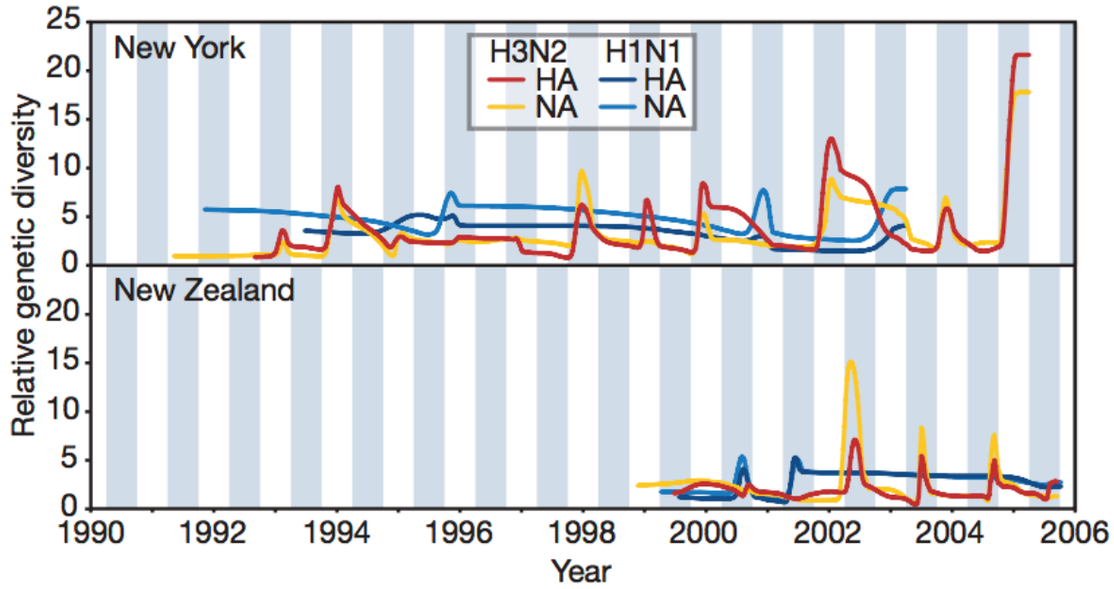


Seasonal influenza

- Yearly winter outbreaks in temperate regions
- Influenza **A/H3N2** is the dominant subtype
- Tropical source population seeds yearly epidemics in temperate sinks
- Strong directional selection maintains limited genetic diversity (most visible in HA gene)



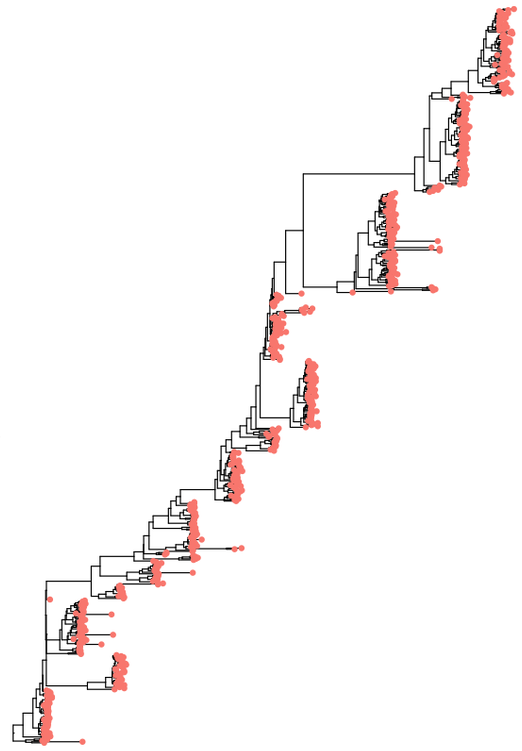




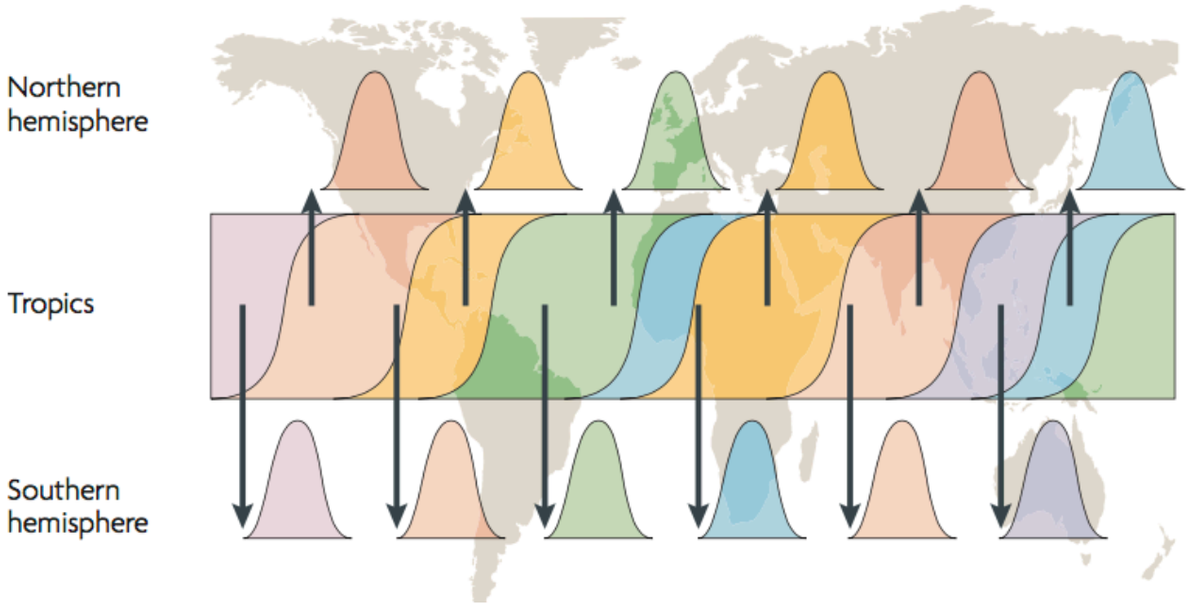
Rambaut *et al.* **Nature 2008**
Pybus and Rambaut **Nature Reviews Genetics 2009**

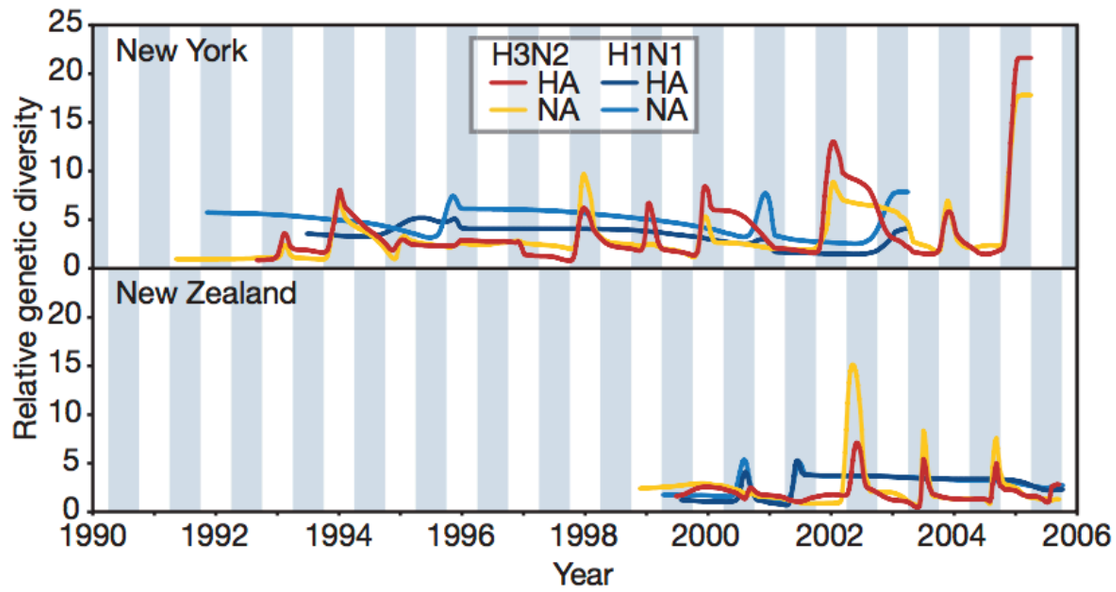
Seasonal influenza

- Yearly winter outbreaks in temperate regions
- Influenza **A/H3N2** is the dominant subtype
- Tropical source population seeds yearly epidemics in temperate sinks
- Strong directional selection maintains limited genetic diversity (most visible in HA gene)



Bayesian Skyline Plot

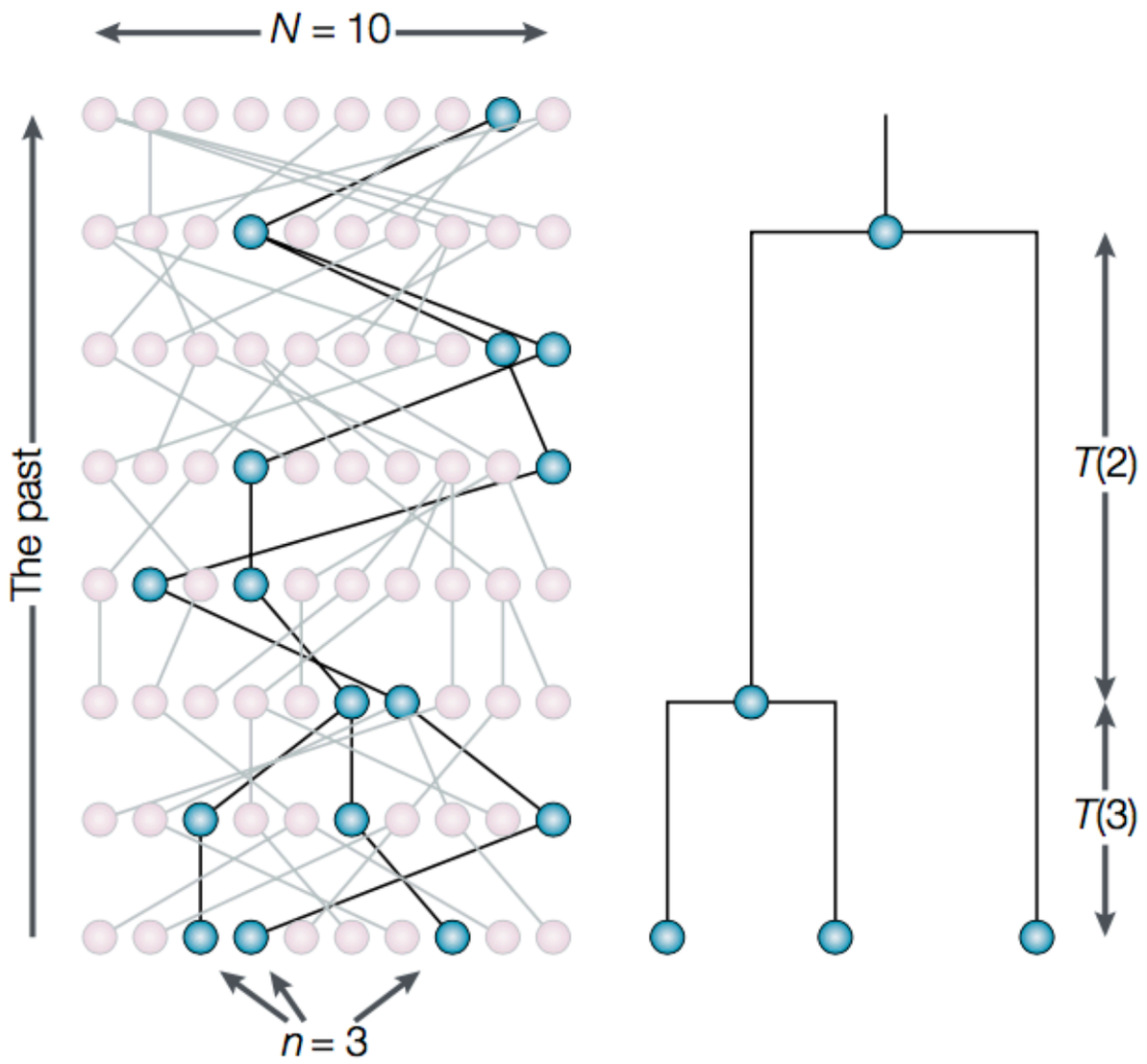




Rambaut *et al.* **Nature 2008**
Pybus and Rambaut **Nature Reviews Genetics 2009**

Kingman coalescent

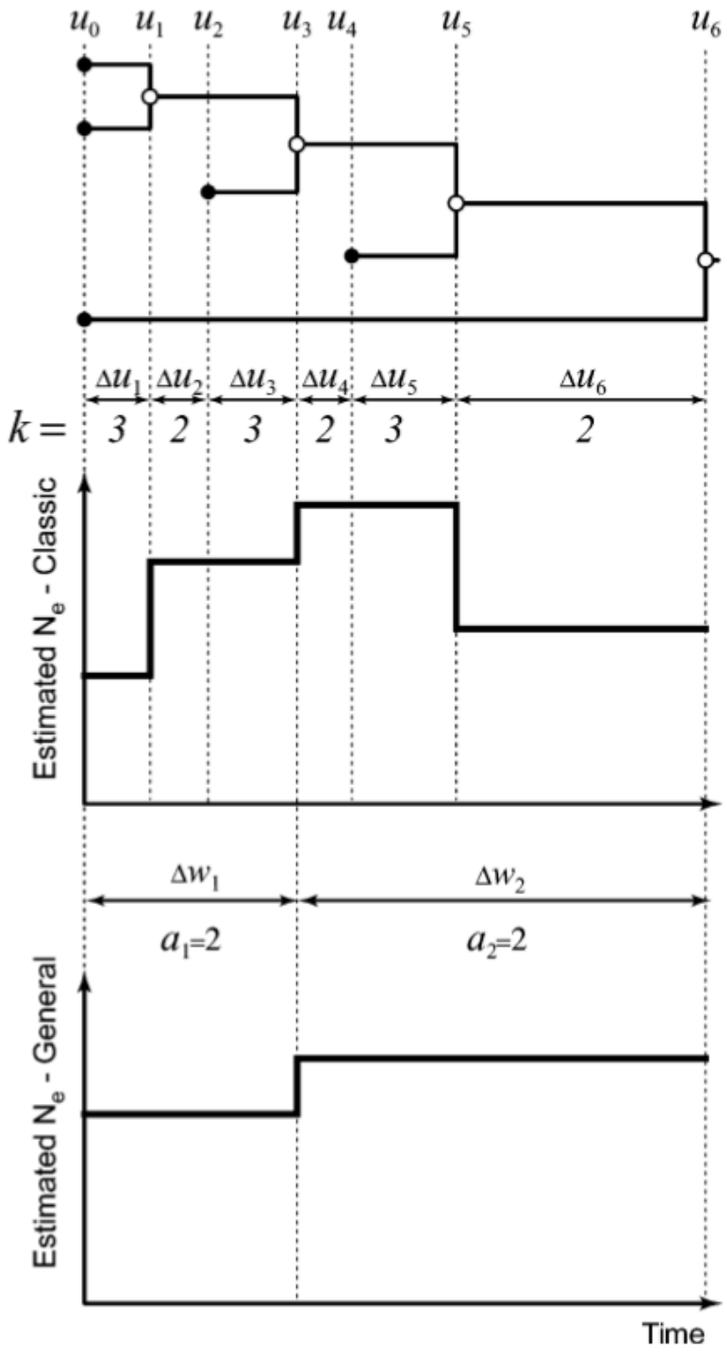
Full population **Sampled tree**



- Approximation to Wright-Fisher population dynamics (with large **N**)
- Trace ancestry of **n** samples in a population of size **N**
- Given **N** it is easy to calculate the probability for **2** nodes to coalesce in time **t**
- Calculate the probability of observing a given **tree** for a particular **N**
  → estimate **N**
  (**N_e** or relative genetic diversity in practice)
- Easy to extend to time changing **N(t)**

Rosenberg and Nordborg **Nature Reviews Genetics 2002**

Skyline plots



Classic skyline plot

- Piecewise constant N_e
- Change-points at coalescent times
- Noisy estimate

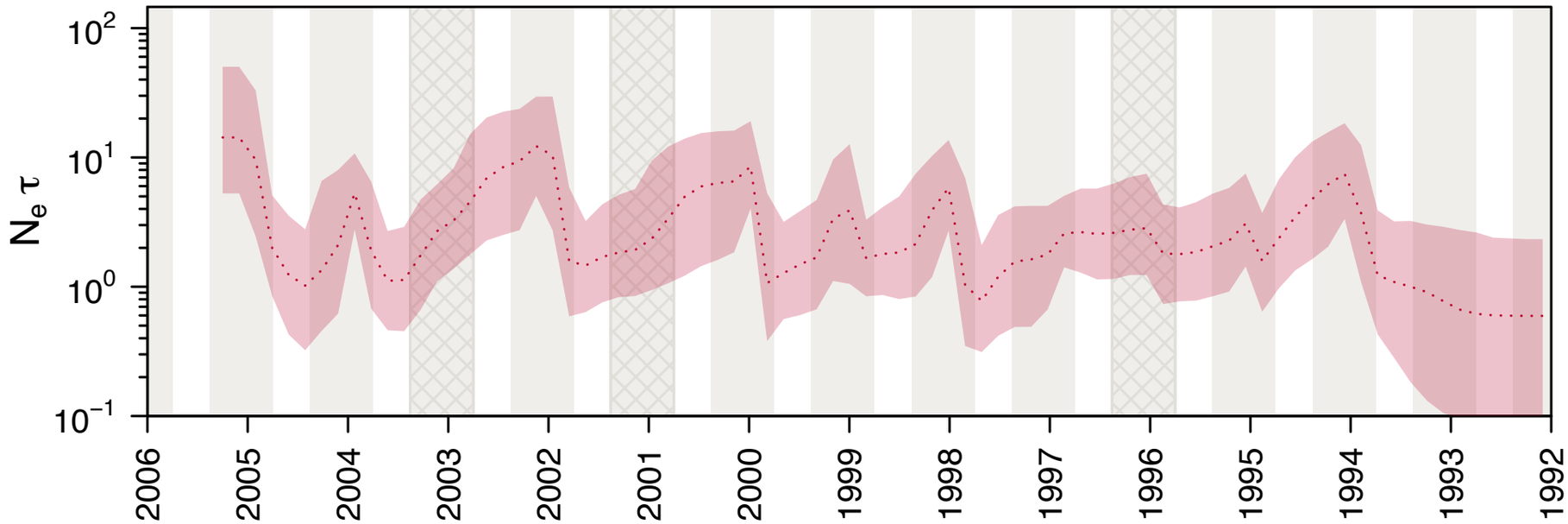
Generalised skyline plot

- Group neighbouring segments
- Smoother estimate

Drummond *et al.* **MBE 2005**

Bayesian Skyline Plot **(BSP)**

637 New York Influenza A/H3N2 HA sequences across 12 seasons



$$P(\text{genealogy}, \text{demographic}, \text{site}, \text{clock} \mid \text{ACAC}\ldots/\text{TCAC}\ldots/\text{ACAG}\ldots) = \frac{P(\text{ACAC}\ldots/\text{TCAC}\ldots/\text{ACAG}\ldots \mid \text{genealogy}, \text{site}, \text{clock})\, P(\text{genealogy} \mid \text{demographic})\, P(\text{demographic})\, P(\text{site})\, P(\text{clock})}{P(\text{ACAC}\ldots/\text{TCAC}\ldots/\text{ACAG}\ldots)}$$

ACAC…
TCAC…
ACAG…
genetic sequences

genealogy

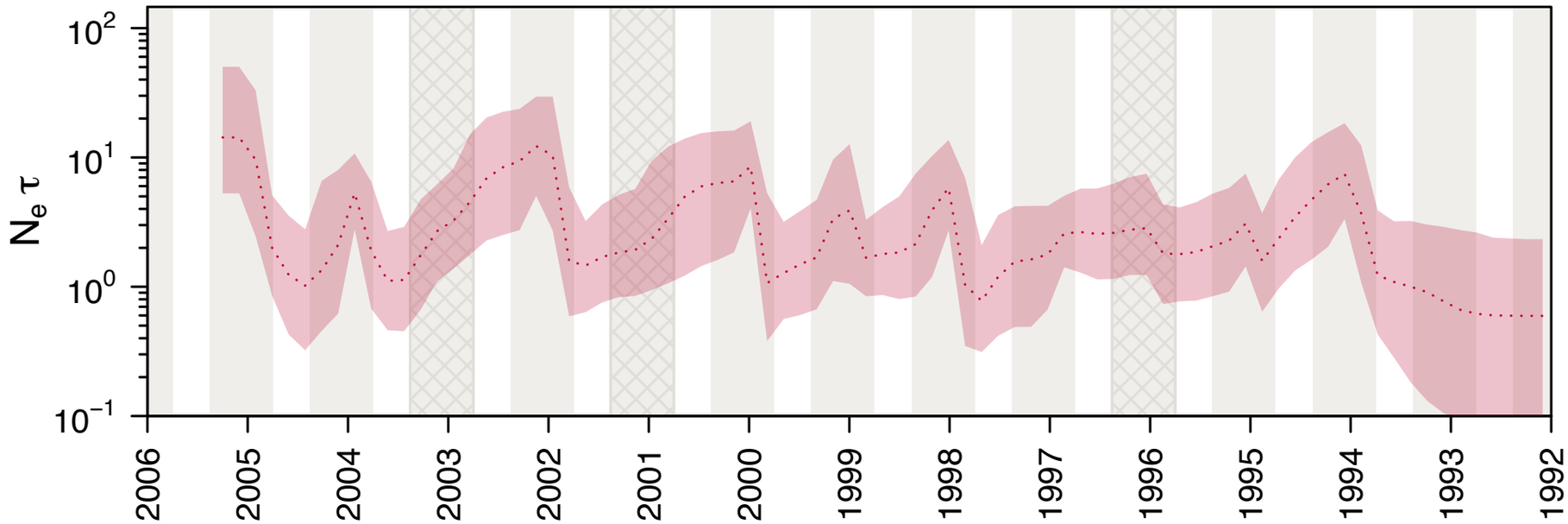
demographic model

site model

molecular clock model

Bayesian Skyline Plot **(BSP)**

637 New York Influenza A/H3N2 HA sequences across 12 seasons



Bayesian Skyline Plot
(Tree prior)

$$P(\text{tree}, \text{demo}, \text{site}, \text{clock} \mid \text{ACAC}\ldots/\text{TCAC}\ldots/\text{ACAG}\ldots) = \frac{P(\text{ACAC}\ldots/\text{TCAC}\ldots/\text{ACAG}\ldots \mid \text{tree}, \text{site}, \text{clock})\, P(\text{tree} \mid \text{demo})\, P(\text{demo})\, P(\text{site})\, P(\text{clock})}{P(\text{ACAC}\ldots/\text{TCAC}\ldots/\text{ACAG}\ldots)}$$

ACAC…
TCAC…
ACAG…
genetic sequences

genealogy

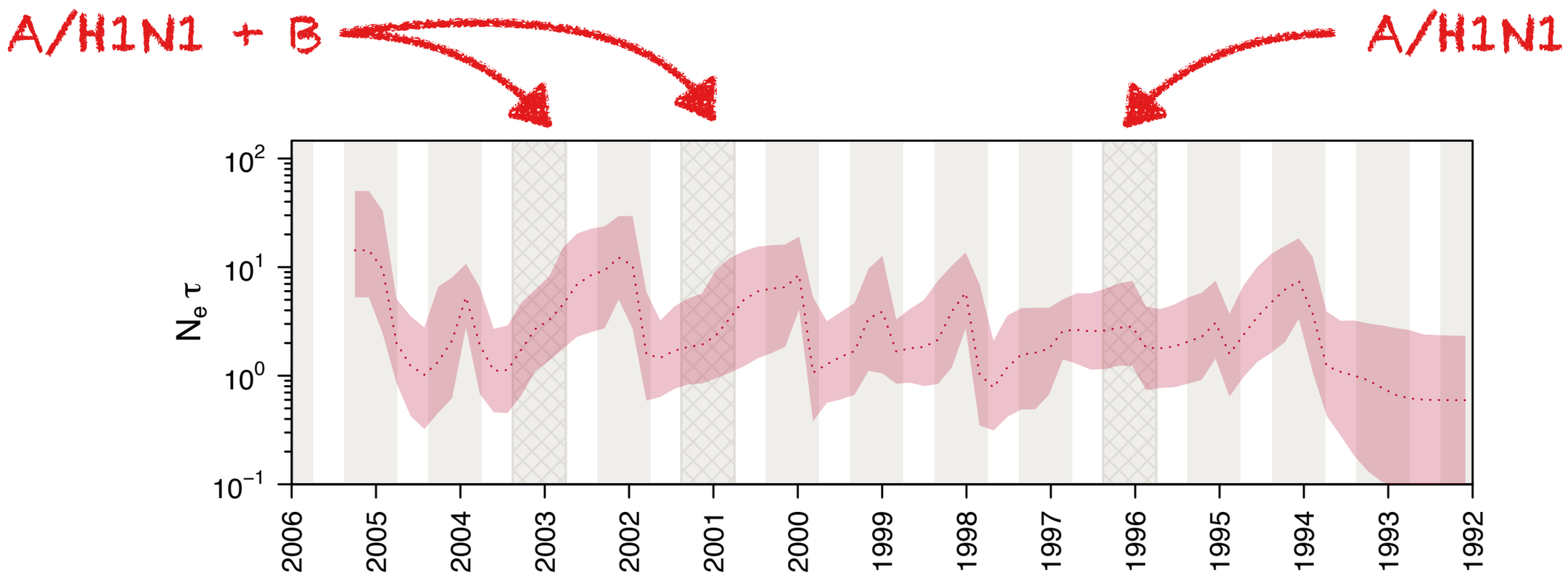
demographic model

site model

molecular clock model

Bayesian Skyline Plot **(BSP)**

637 New York Influenza A/H3N2 HA sequences across 12 seasons



Bayesian Skyline Plot (Tree prior)

$$P(\text{tree}, \text{demo}, \text{site}, \text{clock} \mid \text{ACAC}\ldots/\text{TCAC}\ldots/\text{ACAG}\ldots) = \frac{P(\text{ACAC}\ldots/\text{TCAC}\ldots/\text{ACAG}\ldots \mid \text{tree}, \text{site}, \text{clock})\,P(\text{tree} \mid \text{demo})\,P(\text{demo})\,P(\text{site})\,P(\text{clock})}{P(\text{ACAC}\ldots/\text{TCAC}\ldots/\text{ACAG}\ldots)}$$

ACAC…
TCAC…
ACAG…
genetic sequences

genealogy

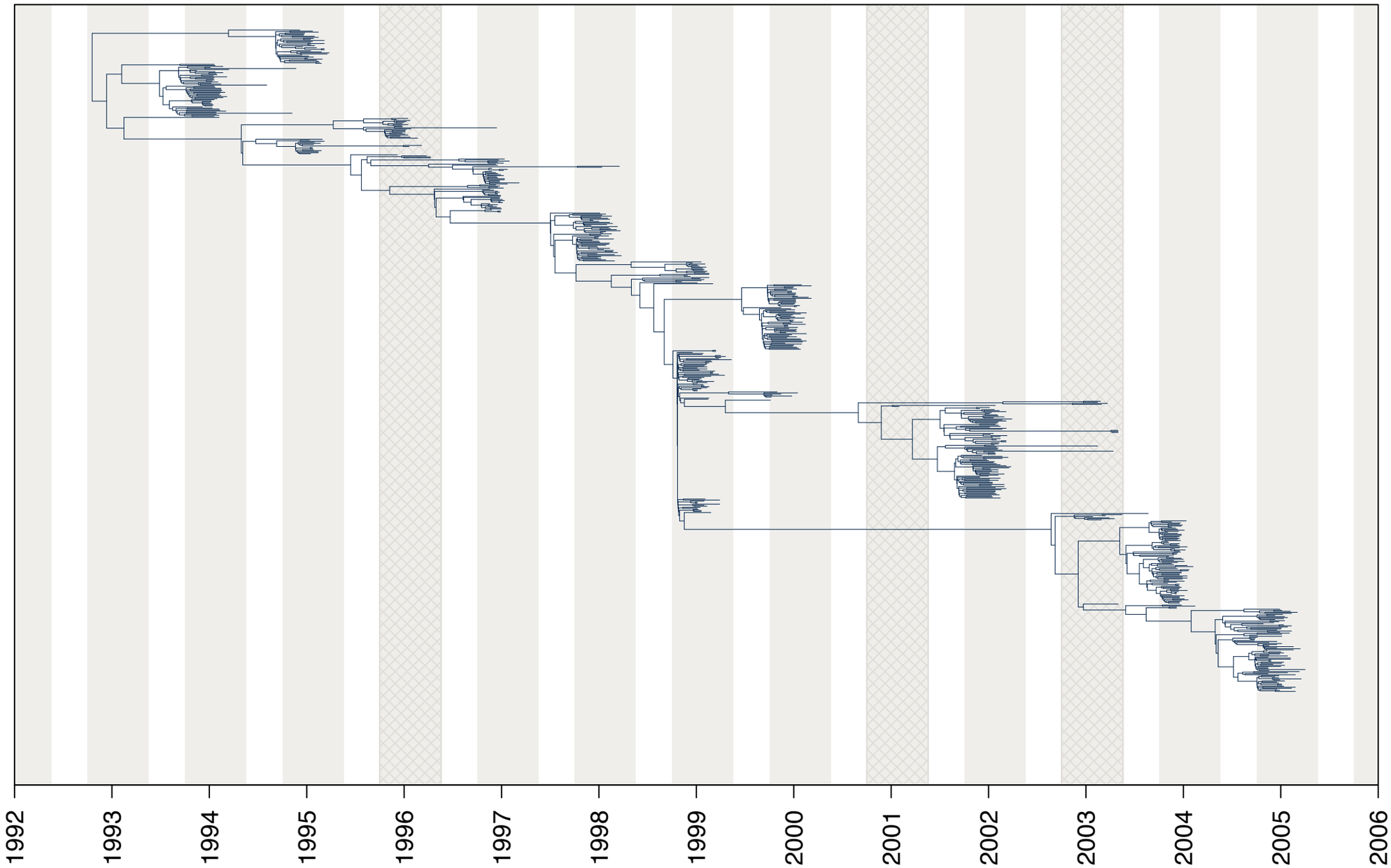
demographic model

site model

molecular clock model

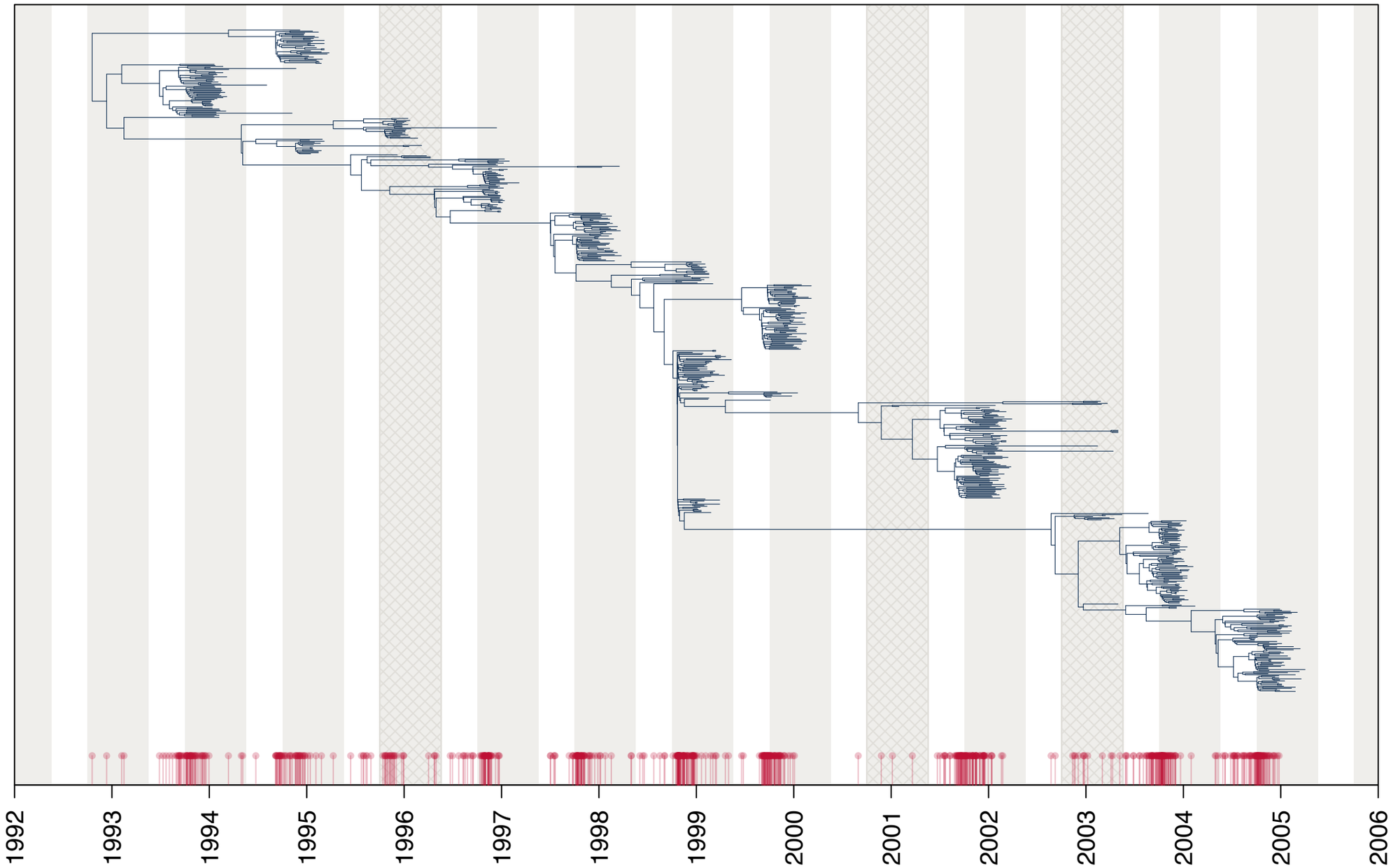
Bayesian Skyline Plot **(BSP)**

637 New York Influenza A/H3N2 HA sequences across 12 seasons



Bayesian Skyline Plot **(BSP)**

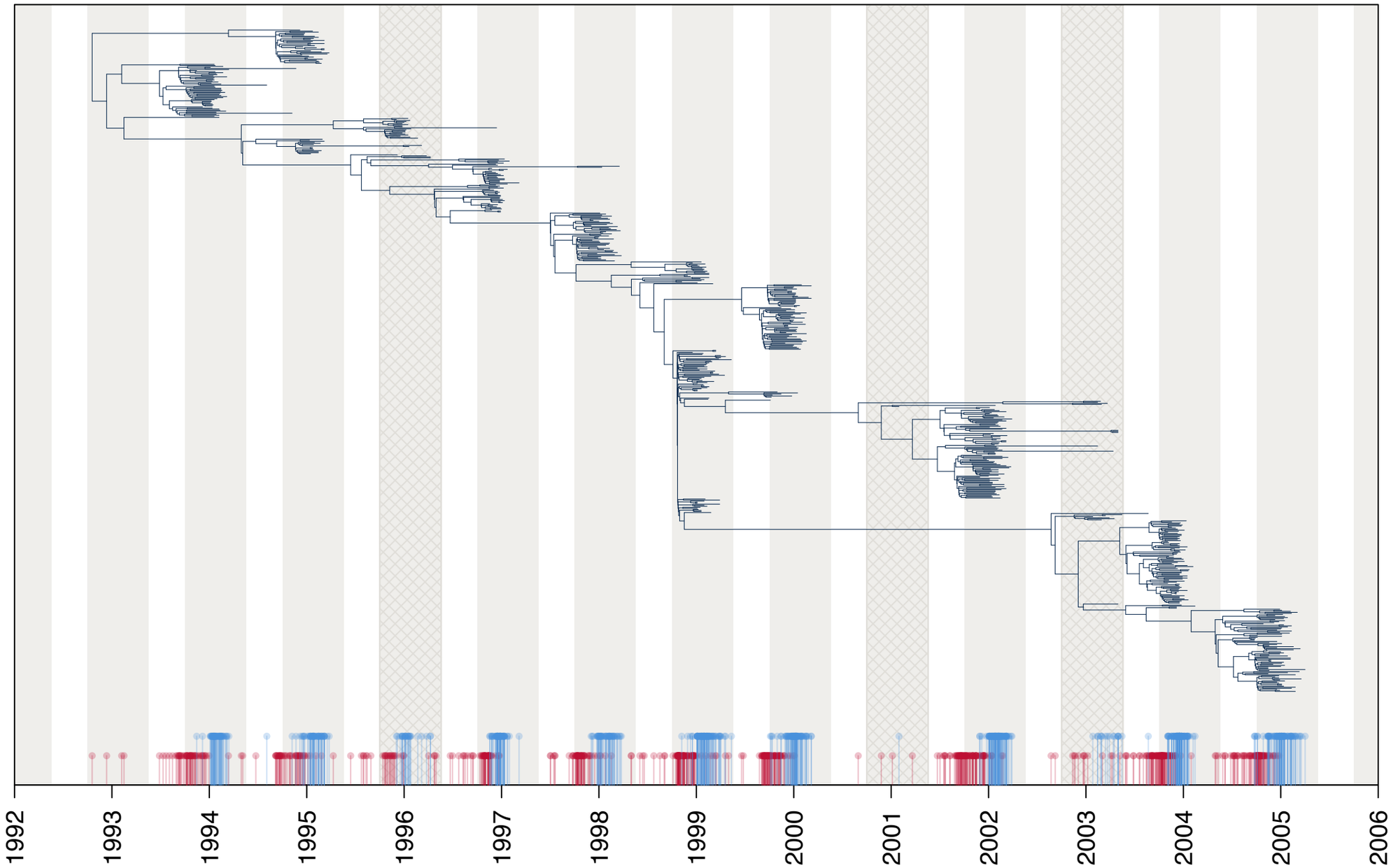
637 New York Influenza A/H3N2 HA sequences across 12 seasons



Only coalescent times are informative to the Bayesian Skyline Plot!

Bayesian Skyline Plot **(BSP)**

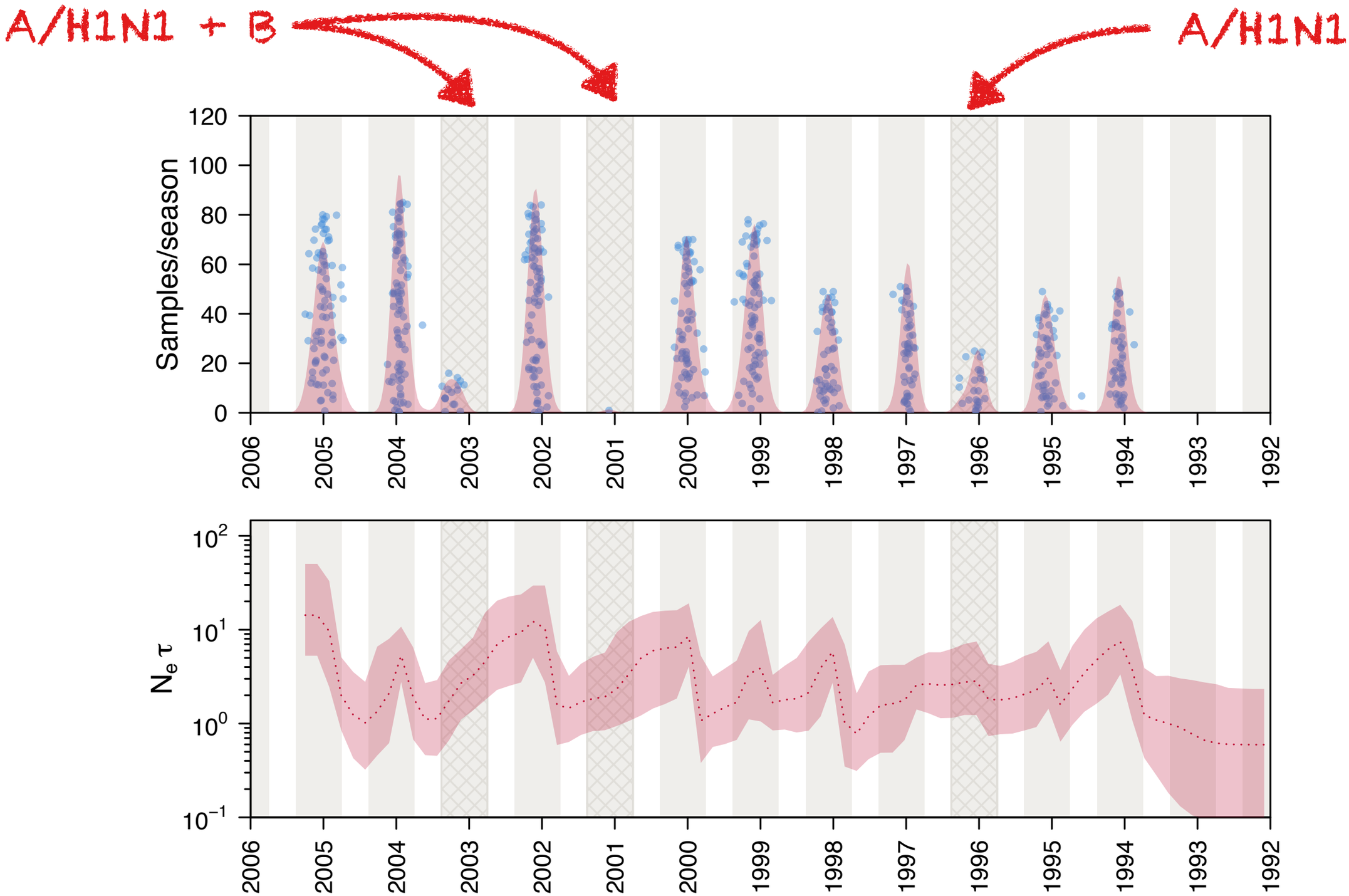
637 New York Influenza A/H3N2 HA sequences across 12 seasons



What about the sampling times?

Bayesian Skyline Plot **(BSP)**

637 New York Influenza A/H3N2 HA sequences across 12 seasons



Coalescent with sampling

- Sampling events follow a Poisson process
- Sampling is proportional to population size with intensity β

Sampling rate: $\psi = \beta \cdot N(t)$

JOURNAL OF THE ROYAL SOCIETY Interface

rsif.royalsocietypublishing.org

Sampling through time and phylodynamic inference with coalescent and birth–death models

Erik M. Volz[1] and Simon D. W. Frost[2]

- Added to exponential growth coalescent

PLOS | COMPUTATIONAL BIOLOGY

RESEARCH ARTICLE

Quantifying and Mitigating the Effect of Preferential Sampling on Phylodynamic Inference

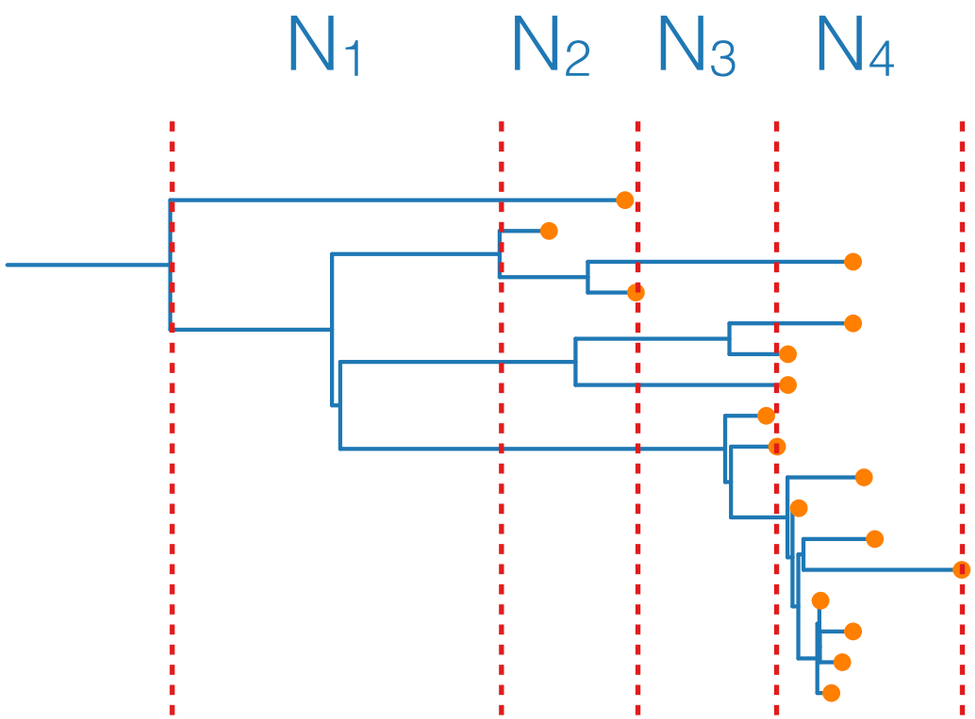
Michael D. Karcher[1], Julia A. Palacios[2,3,4], Trevor Bedford[5], Marc A. Suchard[6,7,8], Vladimir N. Minin[1,9]*

- Added to a Skygrid model
- Non-linear dependence on population size (preferential sampling)

Bayesian Epoch Sampling Skyline plot **(BESP)**

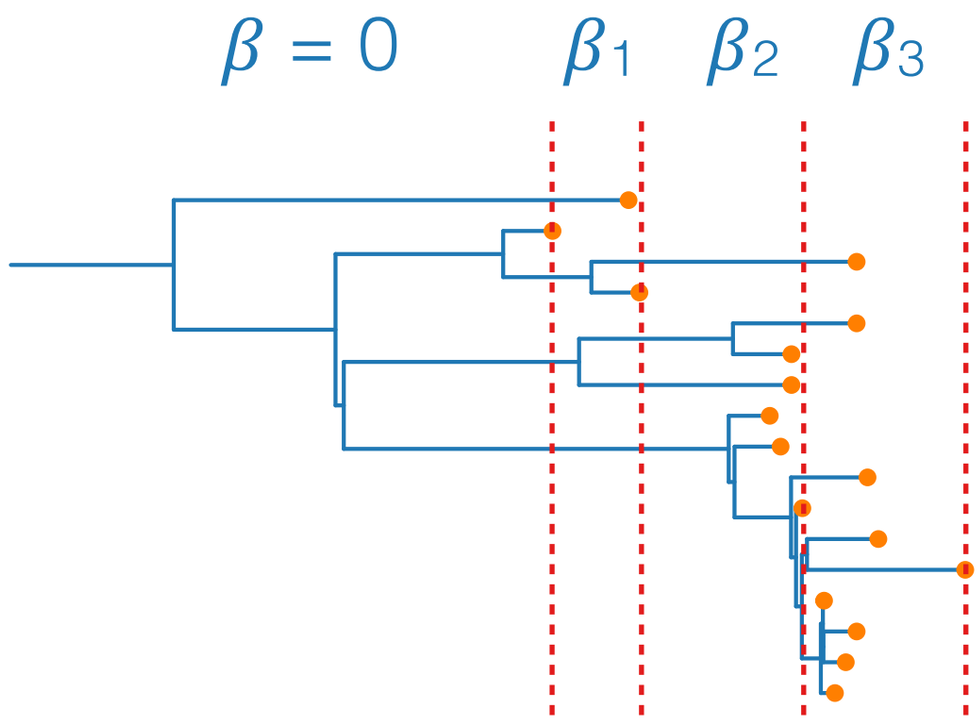
Allow for flexible nonparametric sampling model while ensuring identifiability

Population size



- Group into segments between TMRCA and present
- Change-point times can fall on coalescent or sampling events
- ≥ 2 events / segment

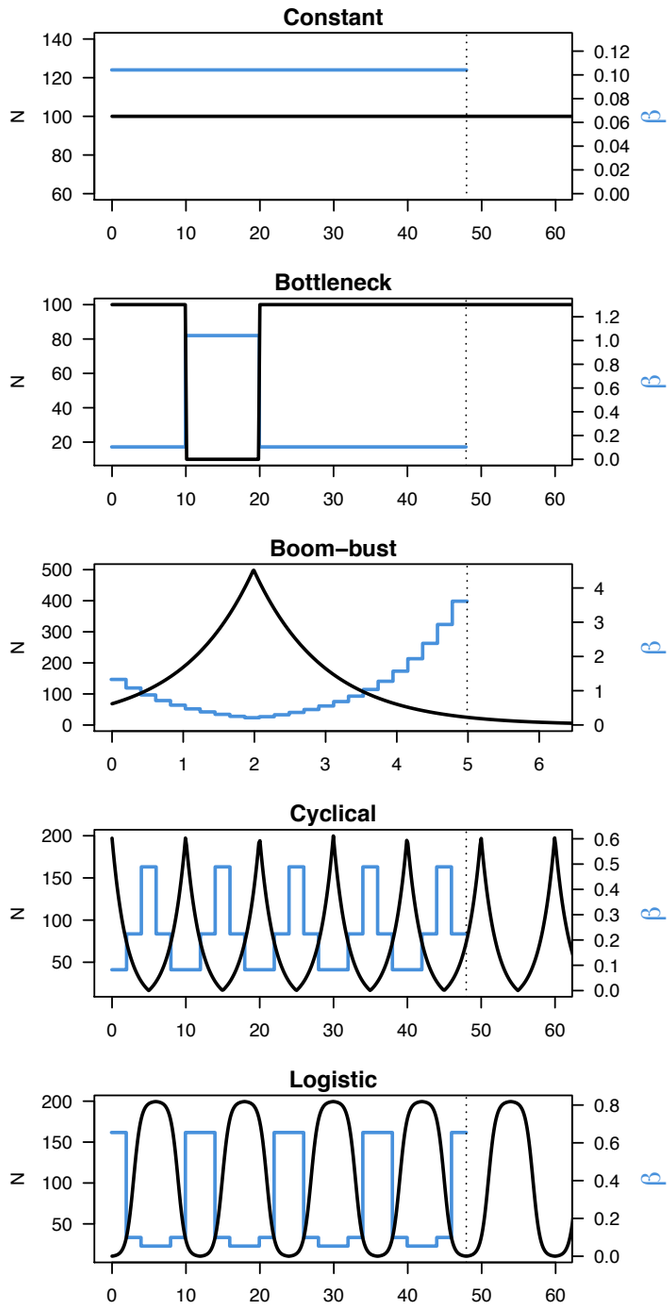
Sampling intensity



- Group into epochs between oldest and most recent samples
- Change-point times fall on sampling events
- ≥ 2 events / epoch

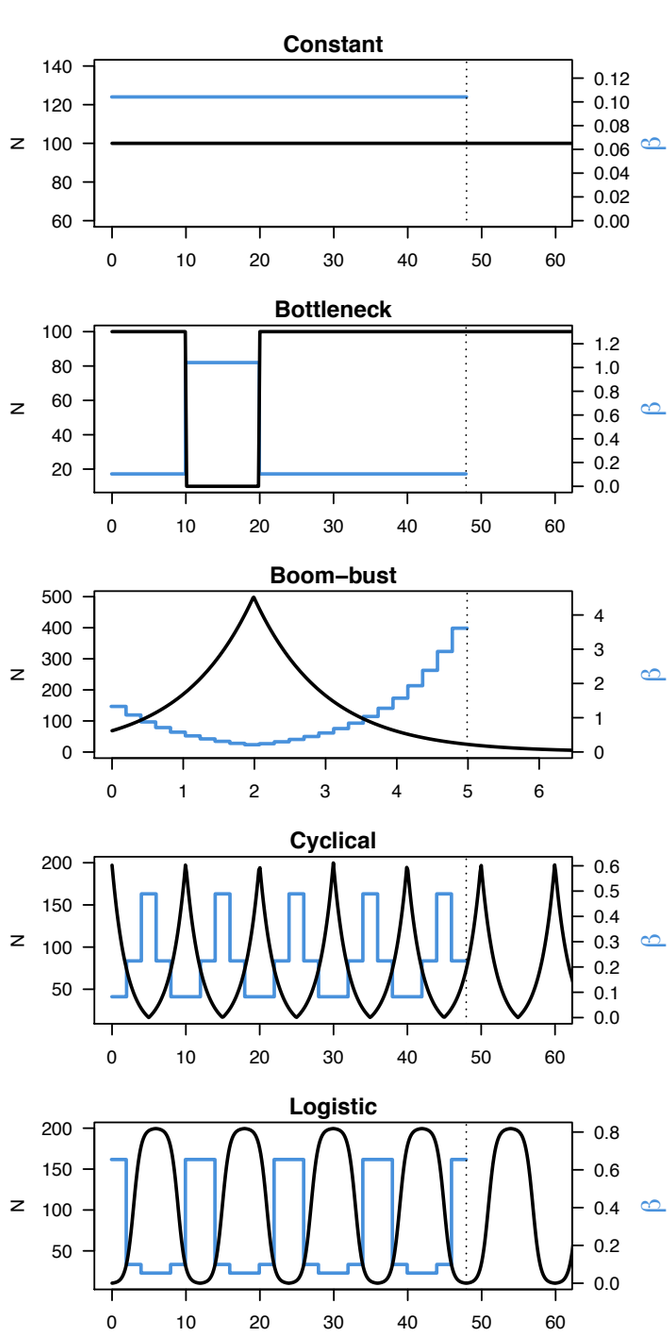
Simulation study **(BEAST2 implementation)**

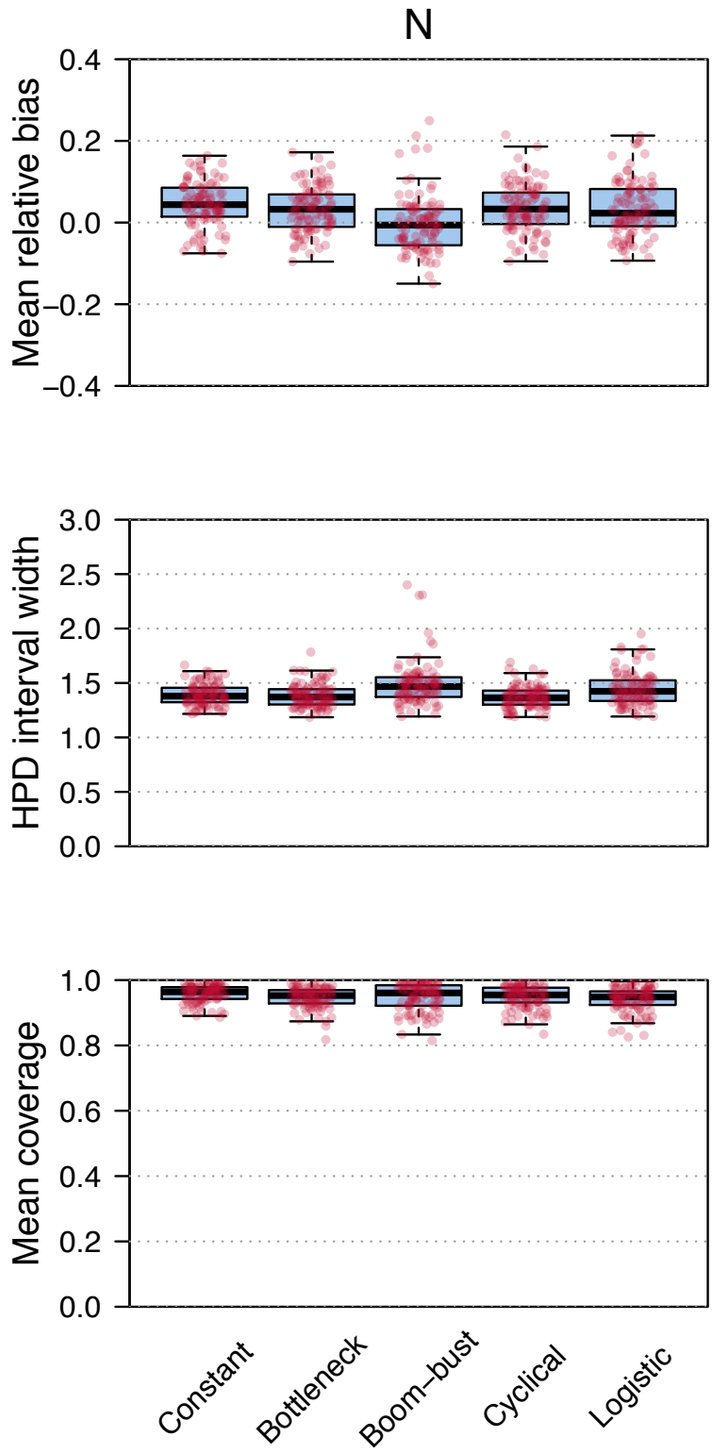
100 replicates / population trajectory simulated with **24 sampling epochs**



Simulation study **(BEAST2 implementation)**

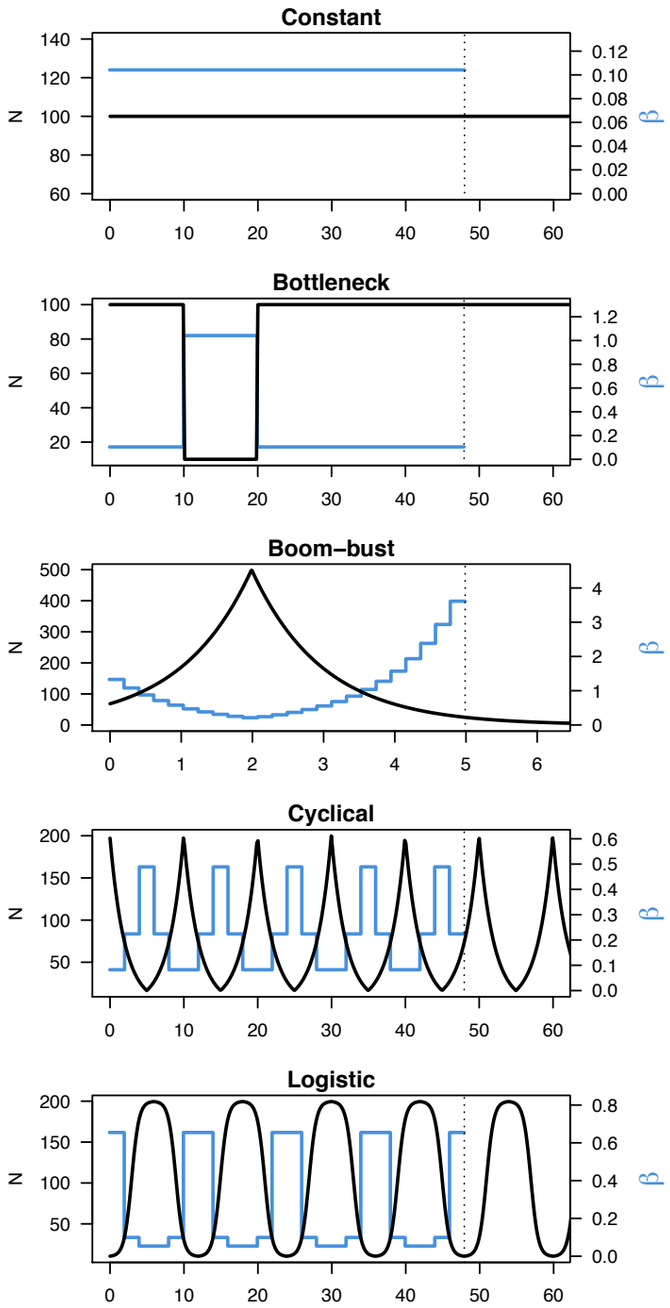
100 replicates / population trajectory simulated with **24 sampling epochs**

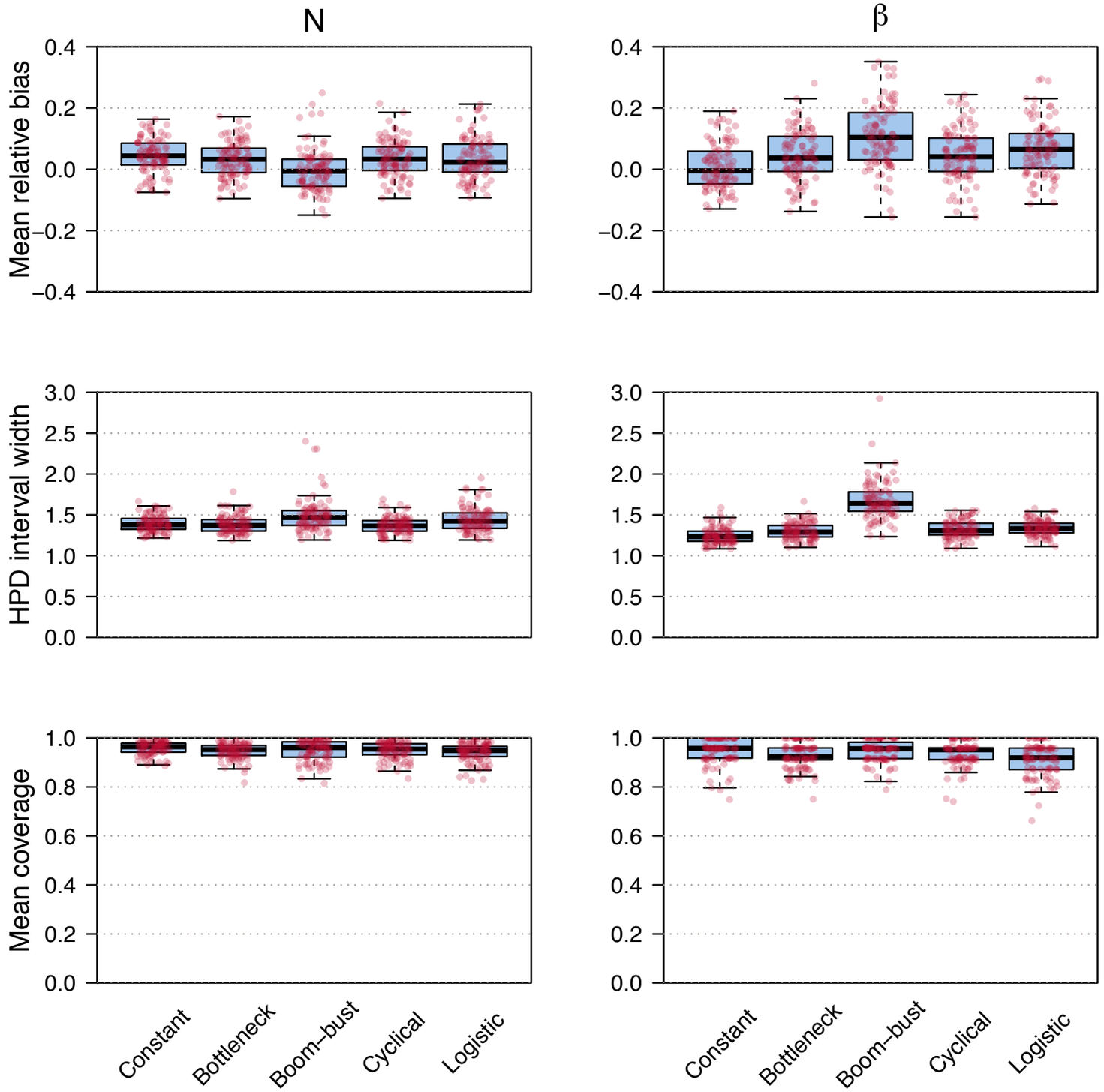




Simulation study **(BEAST2 implementation)**

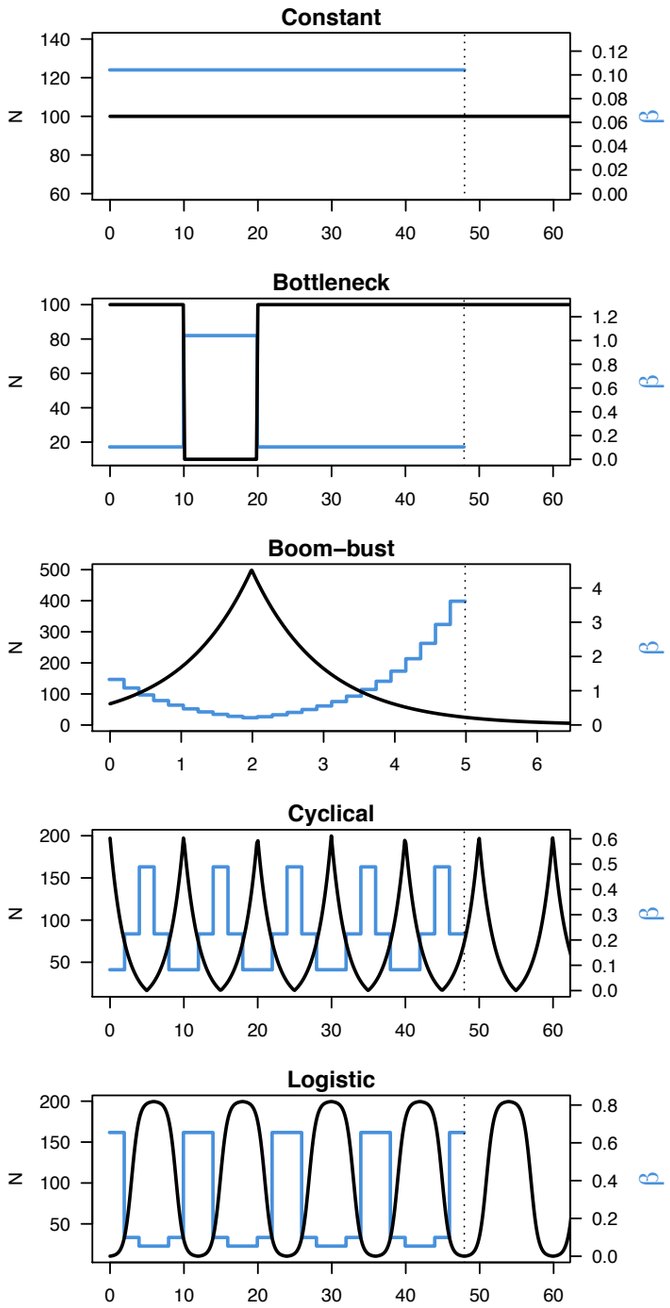
100 replicates / population trajectory simulated with **24 sampling epochs**

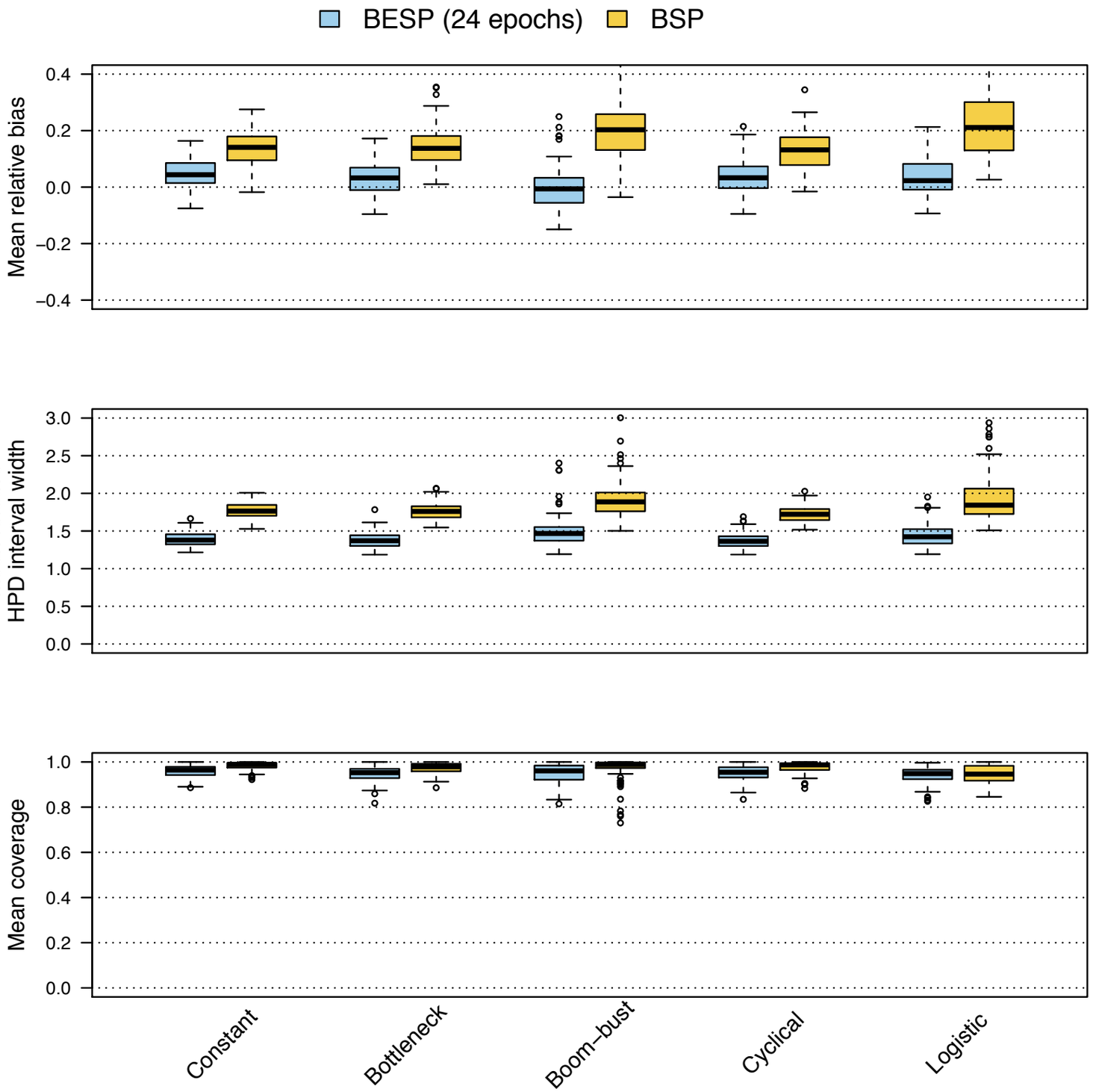




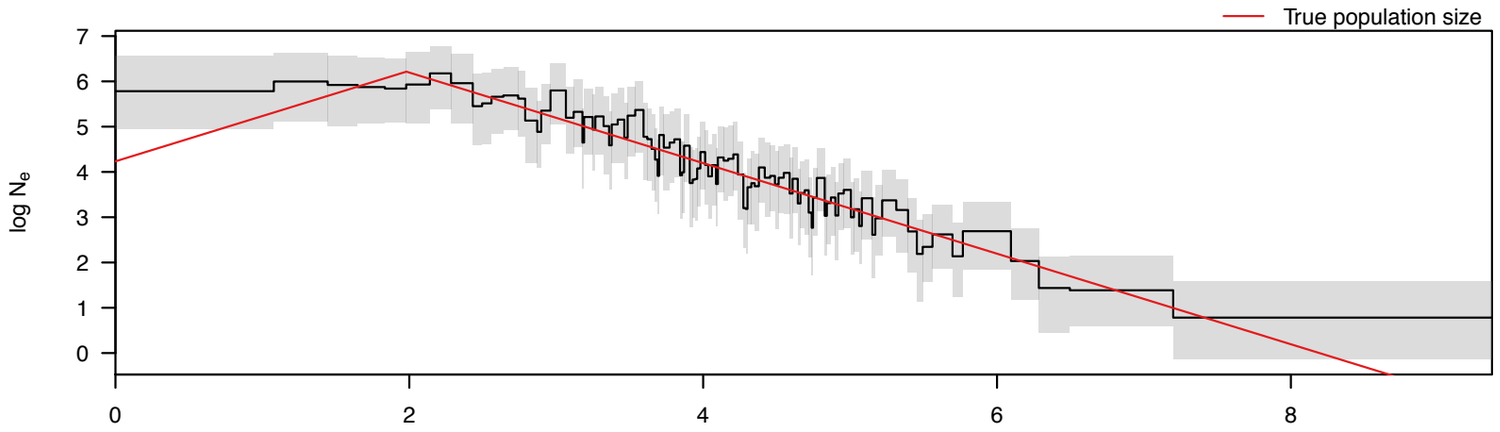
Simulation study **(population size — N)**

100 replicates / population trajectory simulated with **24 sampling epochs**

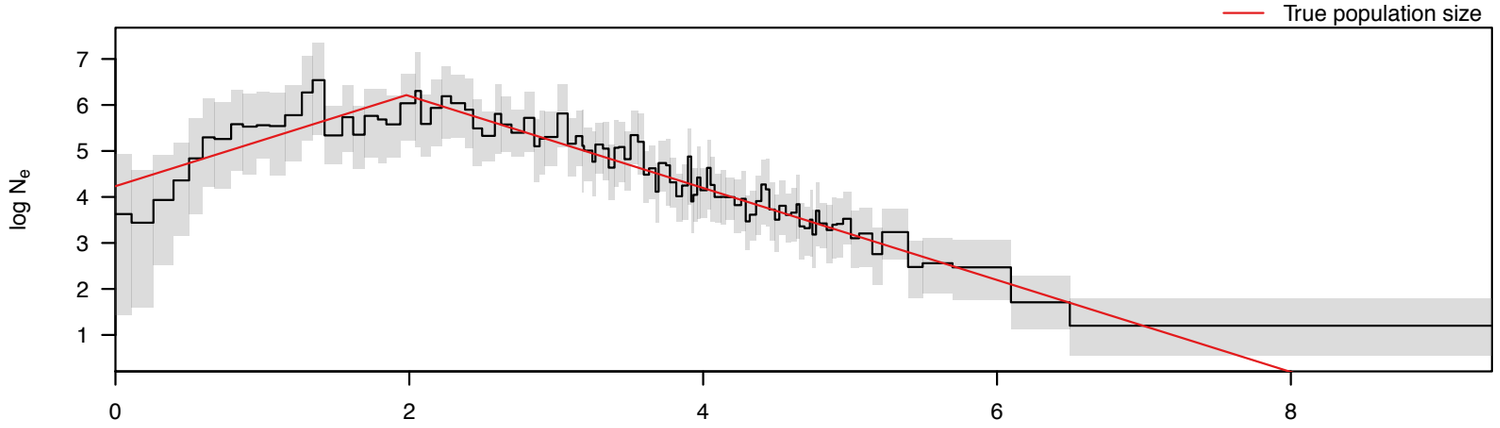




Boom-bust example

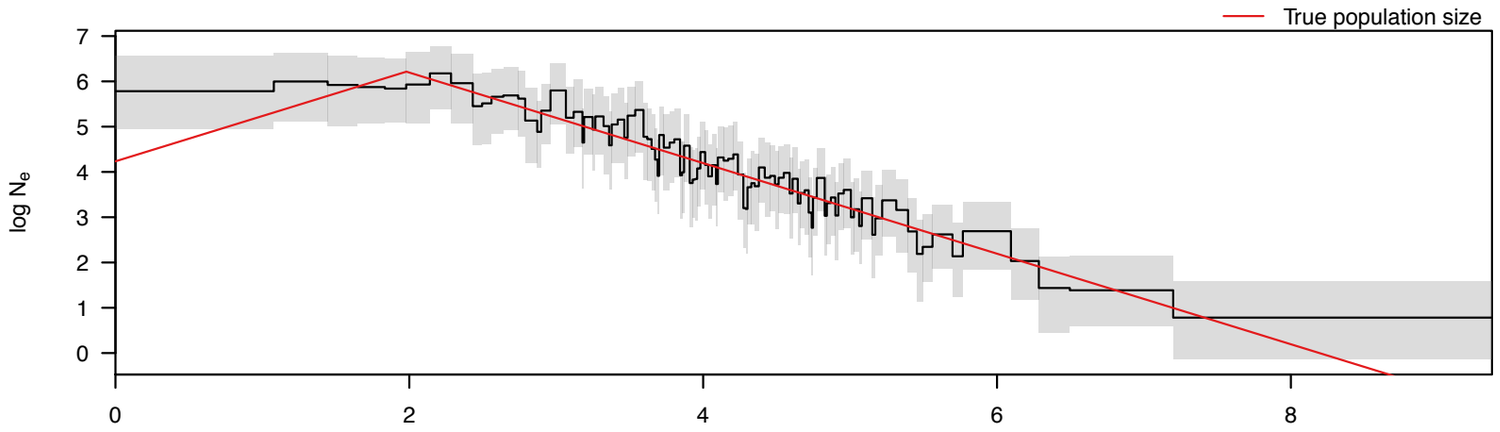


Bayesian Skyline plot

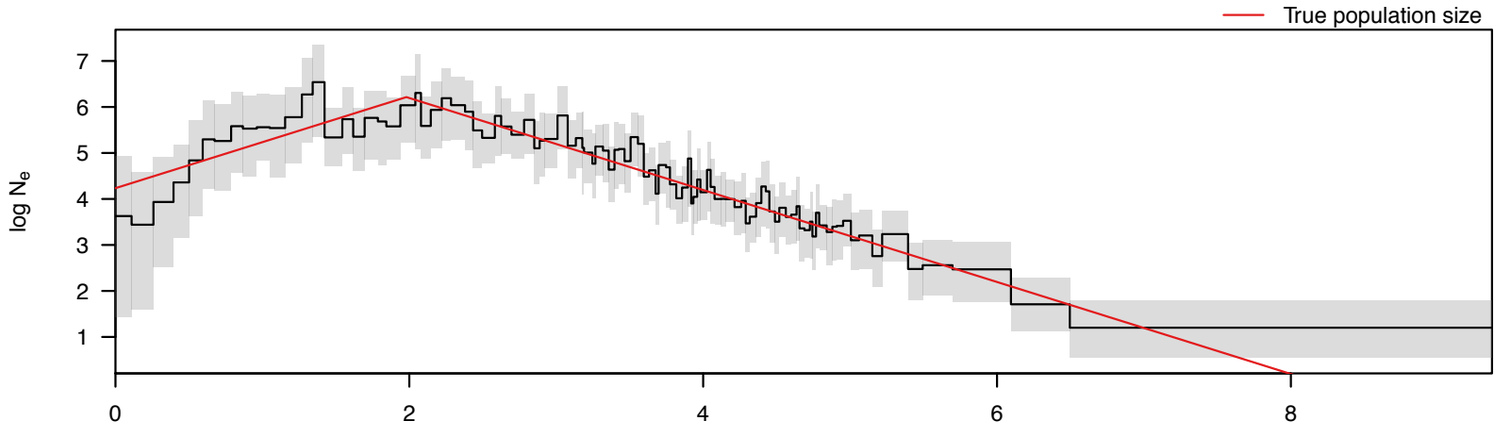


Bayesian Epoch Sampling Skyline plot

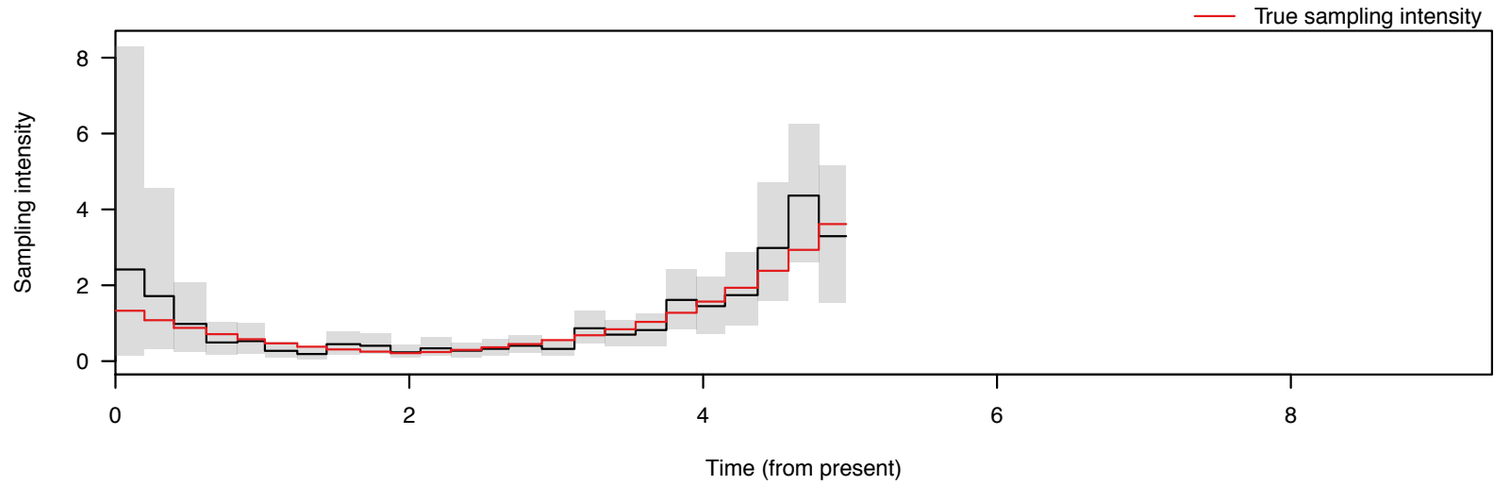
Boom-bust example

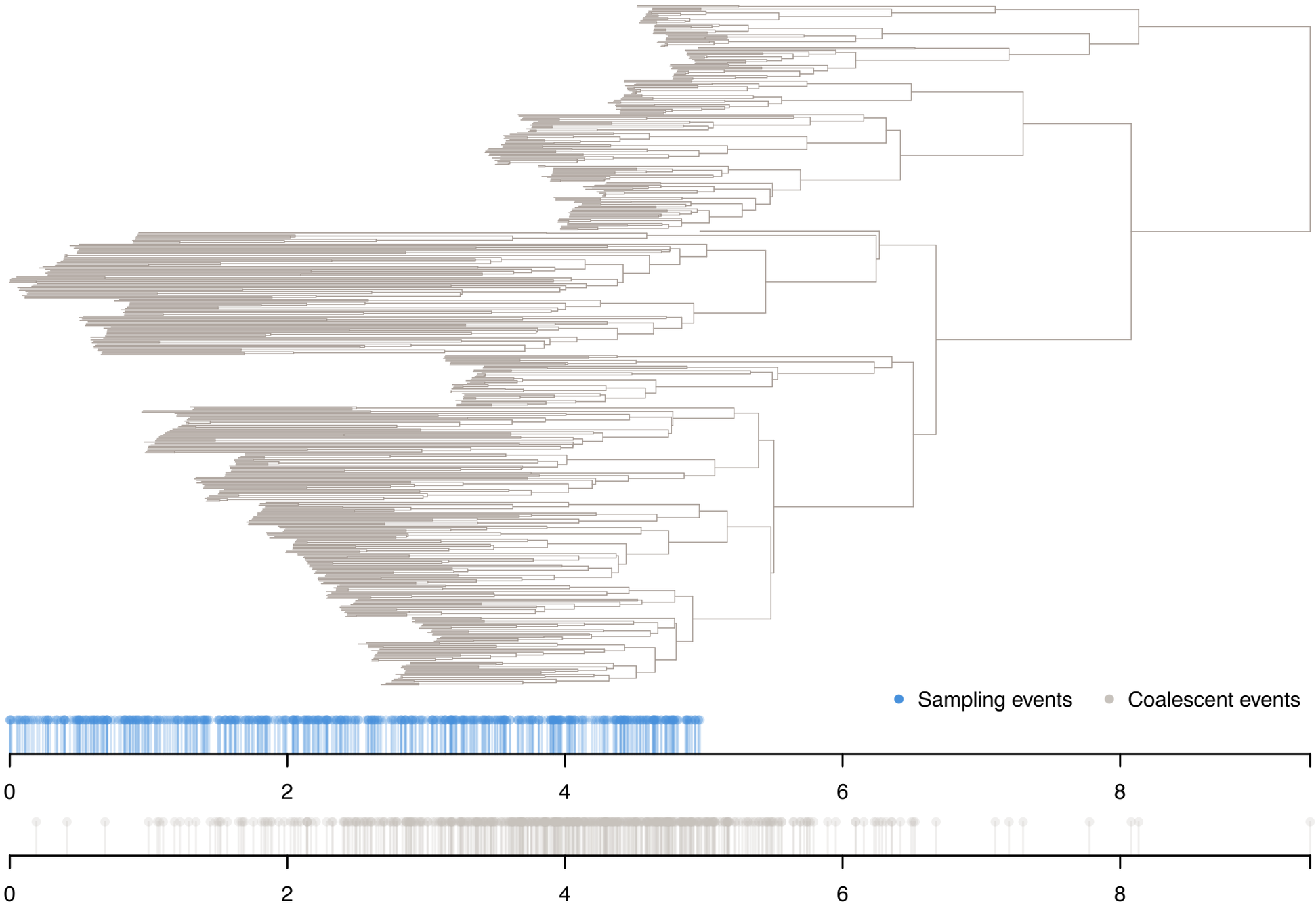


Bayesian Skyline plot



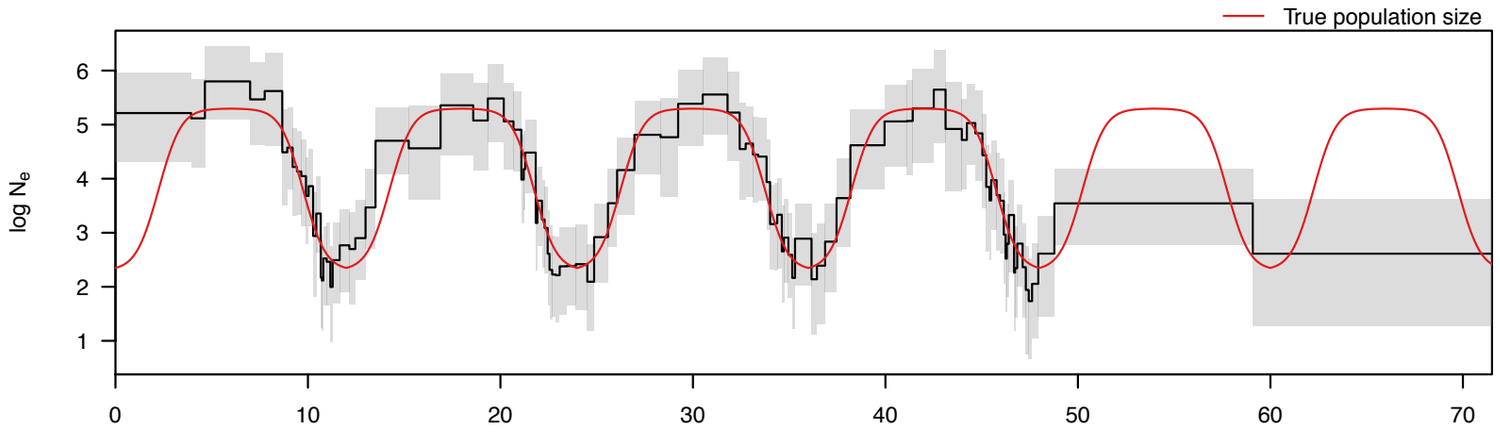
Bayesian Epoch Sampling Skyline plot



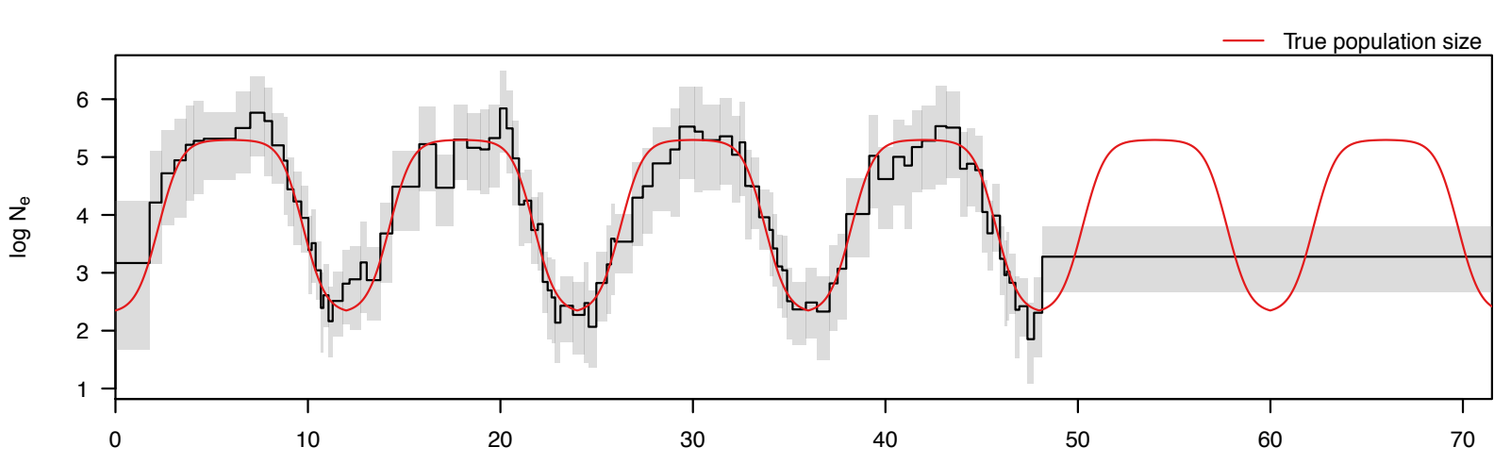


- Not many coalescences during population decline…
- Using sampling time information **doubles** the number of informative data points

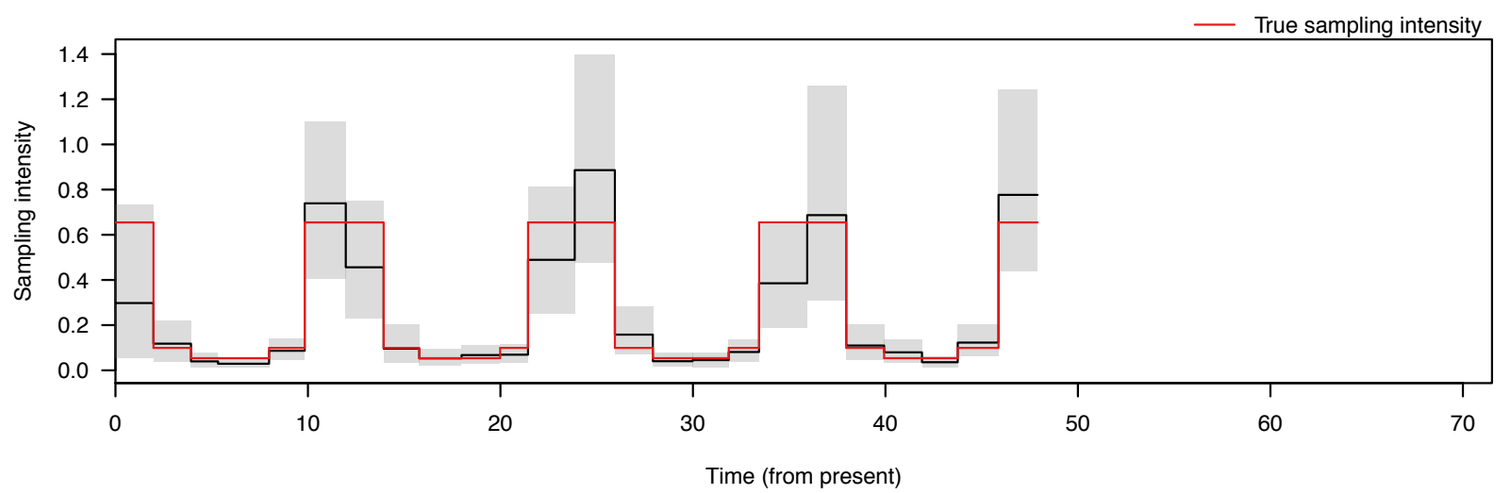
Logistic growth and decline example

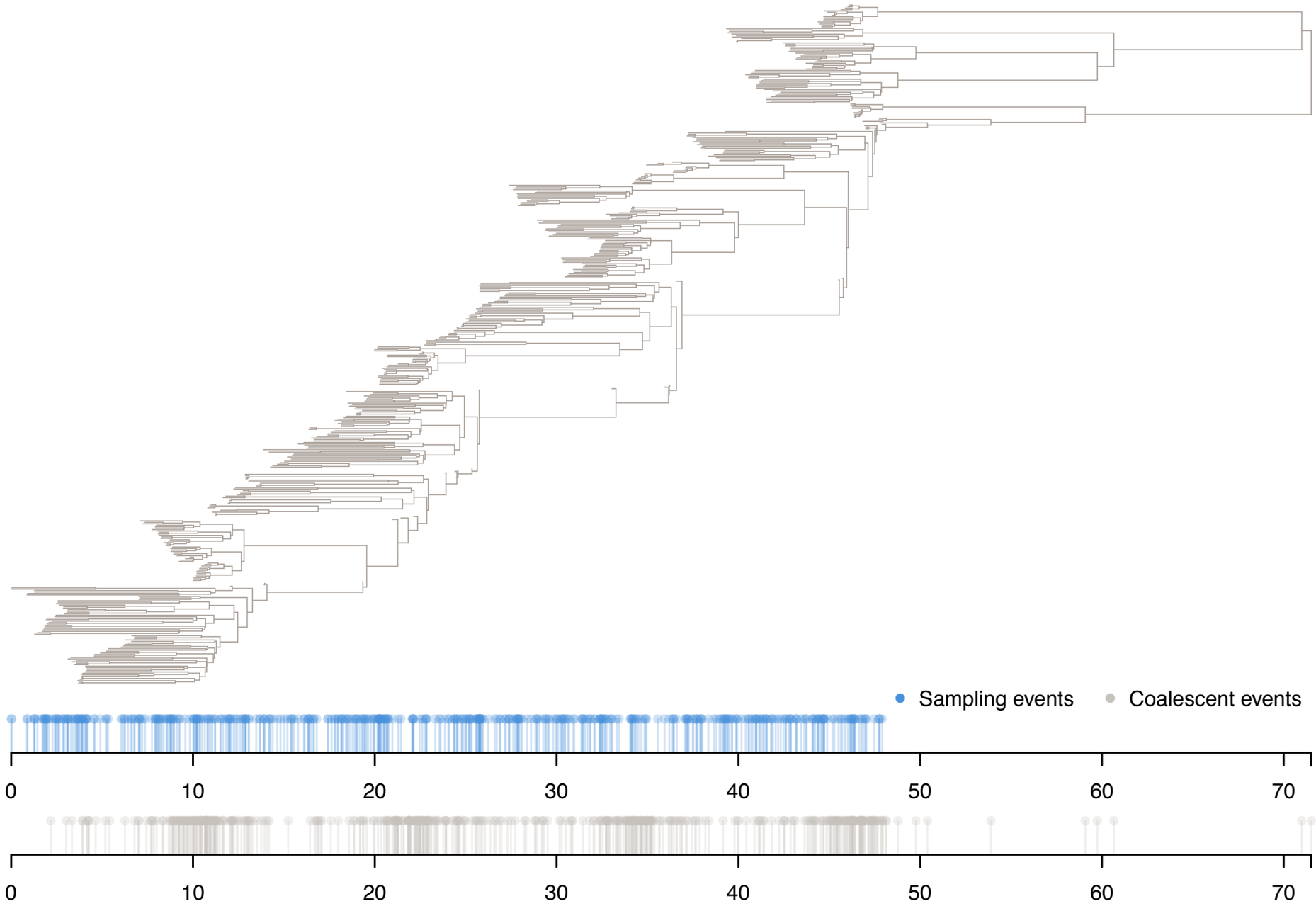


Bayesian Skyline plot



Bayesian Epoch Sampling Skyline plot

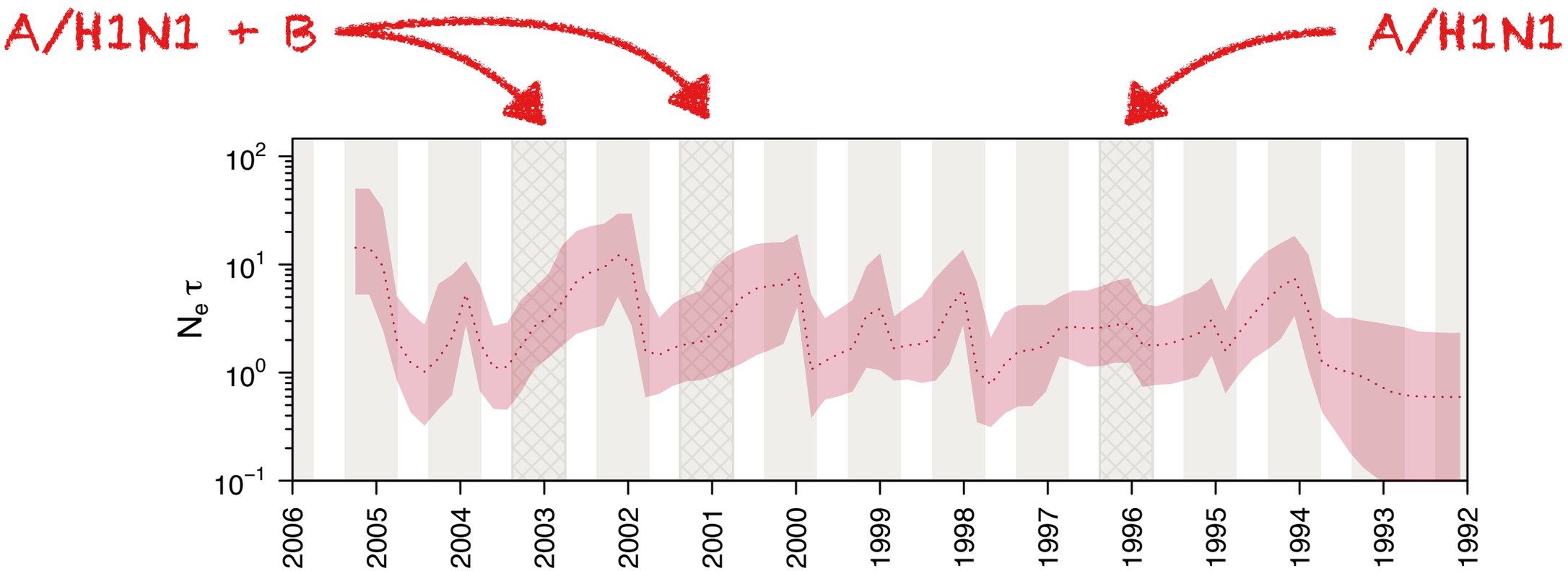




- Not many coalescences during population decline…
- Using sampling time information **doubles** the number of informative data points

Bayesian Epoch Sampling Skyline Plot

637 New York Influenza A/H3N2 HA sequences across 12 seasons

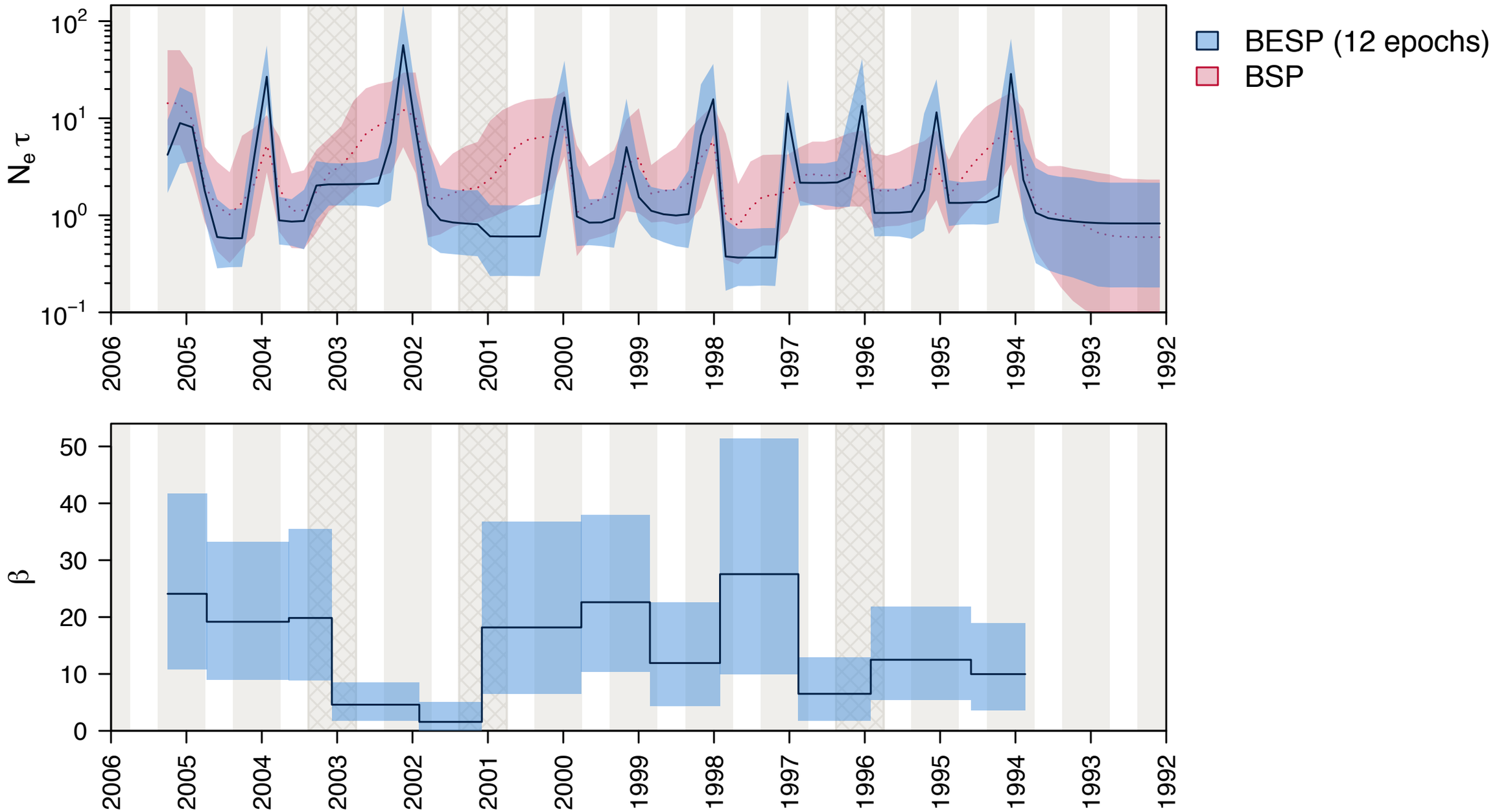


Bayesian Epoch Sampling Skyline Plot

637 New York Influenza A/H3N2 HA sequences across 12 seasons

A/H1N1 + B

A/H1N1



New Results

Jointly inferring the dynamics of population size and sampling intensity from molecular sequences

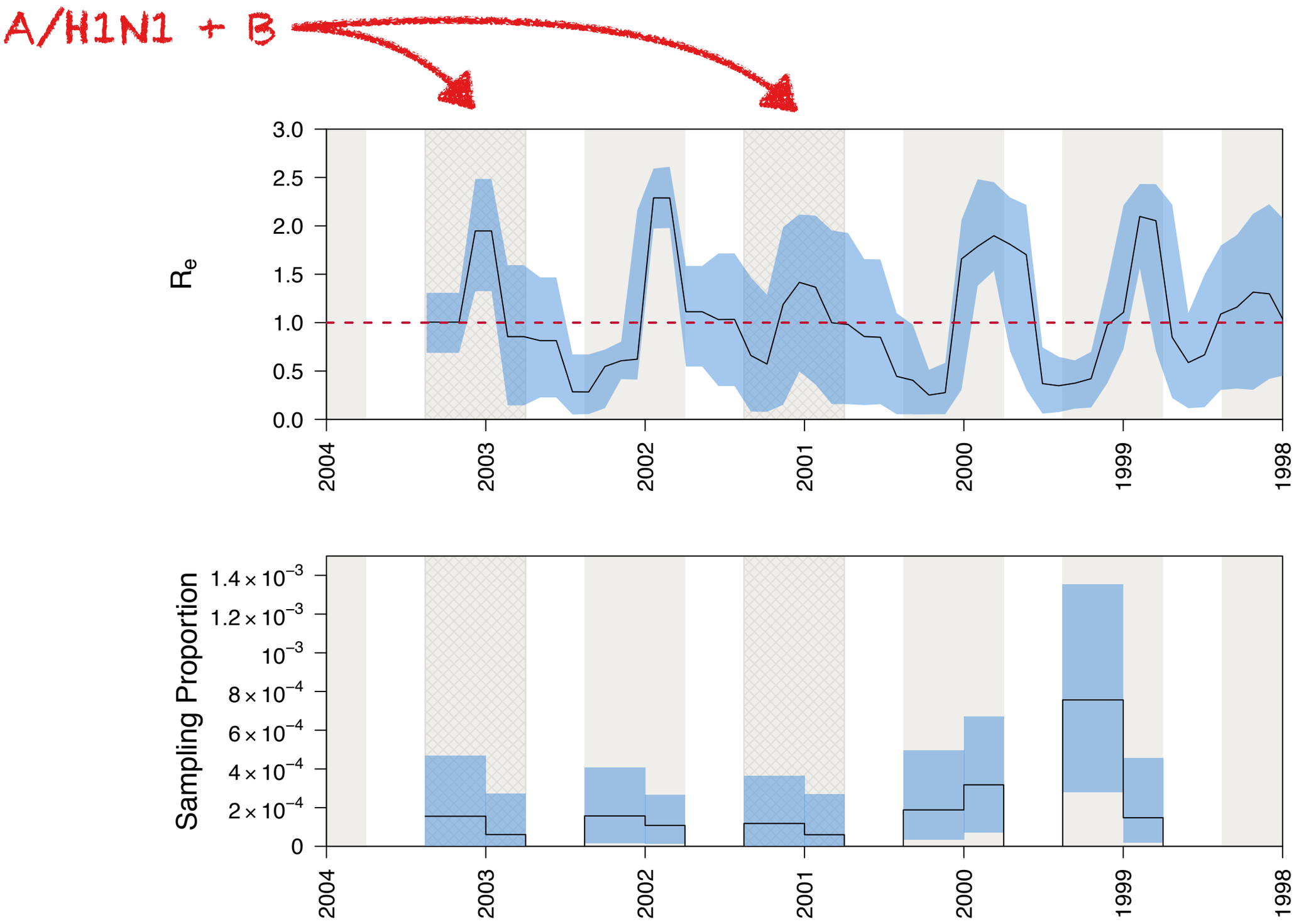
KV Parag, L du Plessis, OG Pybus

doi: https://doi.org/10.1101/686378

BEAST2 package:
github.com/laduplessis/besp/

Birth-death skyline digression

249 New York Influenza A/H3N2 HA sequences across 5 seasons



Birth-death skyline digression

249 New York Influenza A/H3N2 HA sequences across 5 seasons

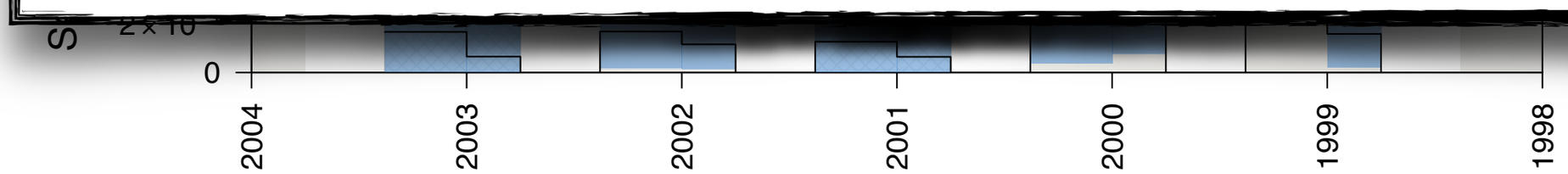
A/H1N

Caveats:

- Parameter-rich models
- To ensure identifiability need to fix parameters *a priori* or use very strong priors

Personal experience:

- When the fraction of sampled cases is **small** the sampling proportion is **not identifiable** (Zika, Influenza, Dengue, Chikungunya etc.)
- Only relative changes are identifiable



Summary

- **Sampling times are informative about population dynamics** (if they are non-random)
- **BESP** is a **flexible nonparametric** method for the **demographic history** and the **sampling process**
- **BEAST2** package

Things I left out

- Fast **maximum-likelihood solution**
- Accounting for sampling at least **doubles the precision**
- **Model is identifiable** and not overly reliant on strong priors

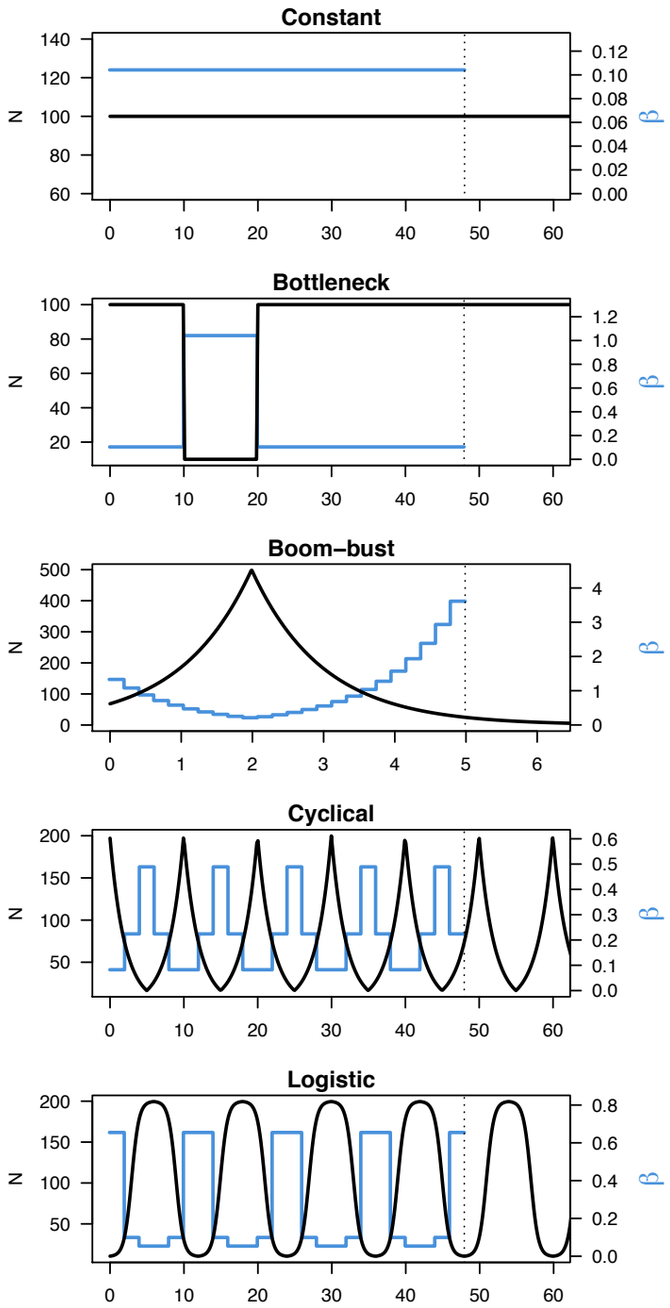
Thanks to:

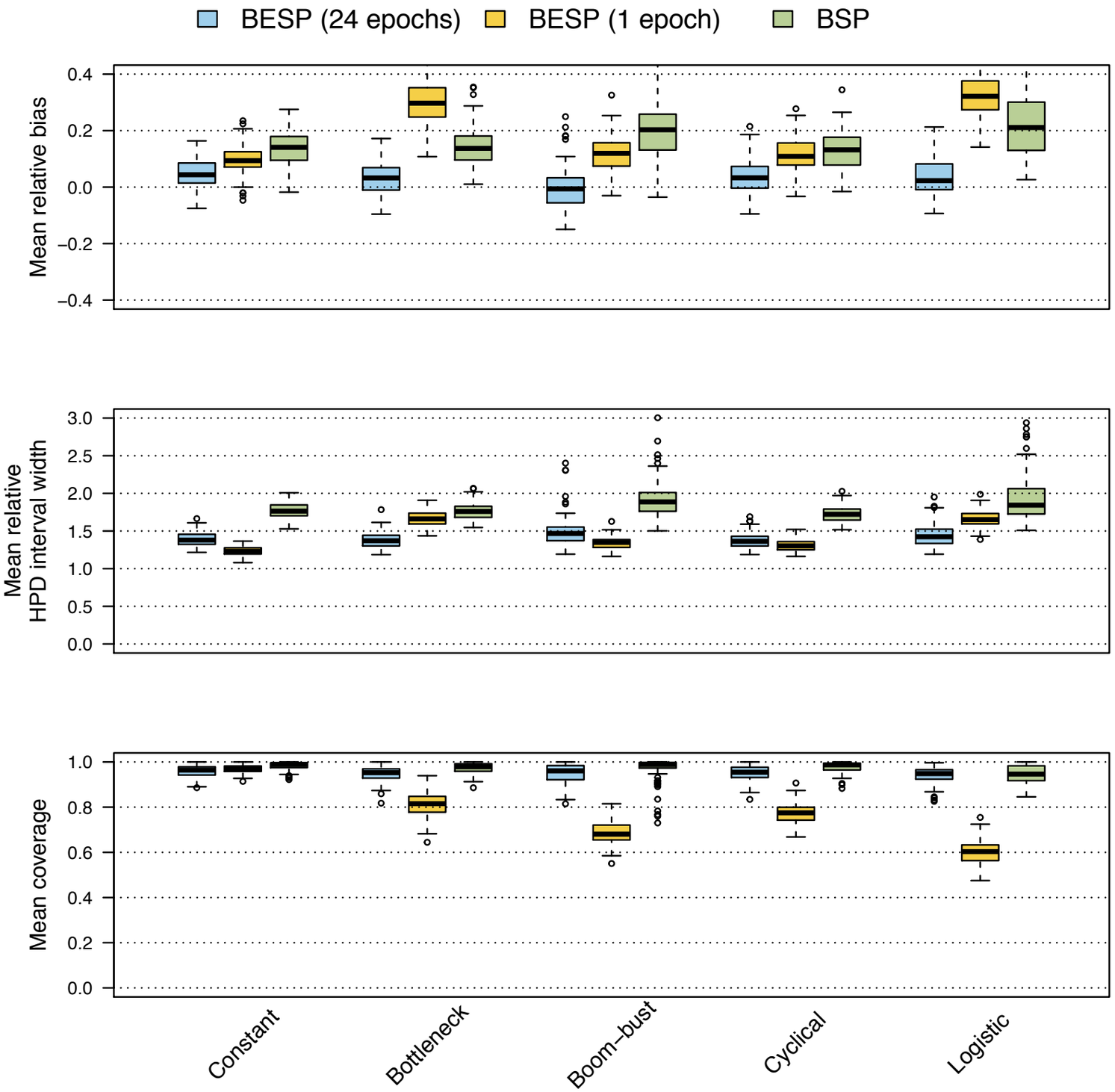
Kris Parag and Oliver Pybus

Edward Holmes, Beth Shapiro, Cécile Viboud, Amanda Perofsky

Simulation study **(population size — N)**

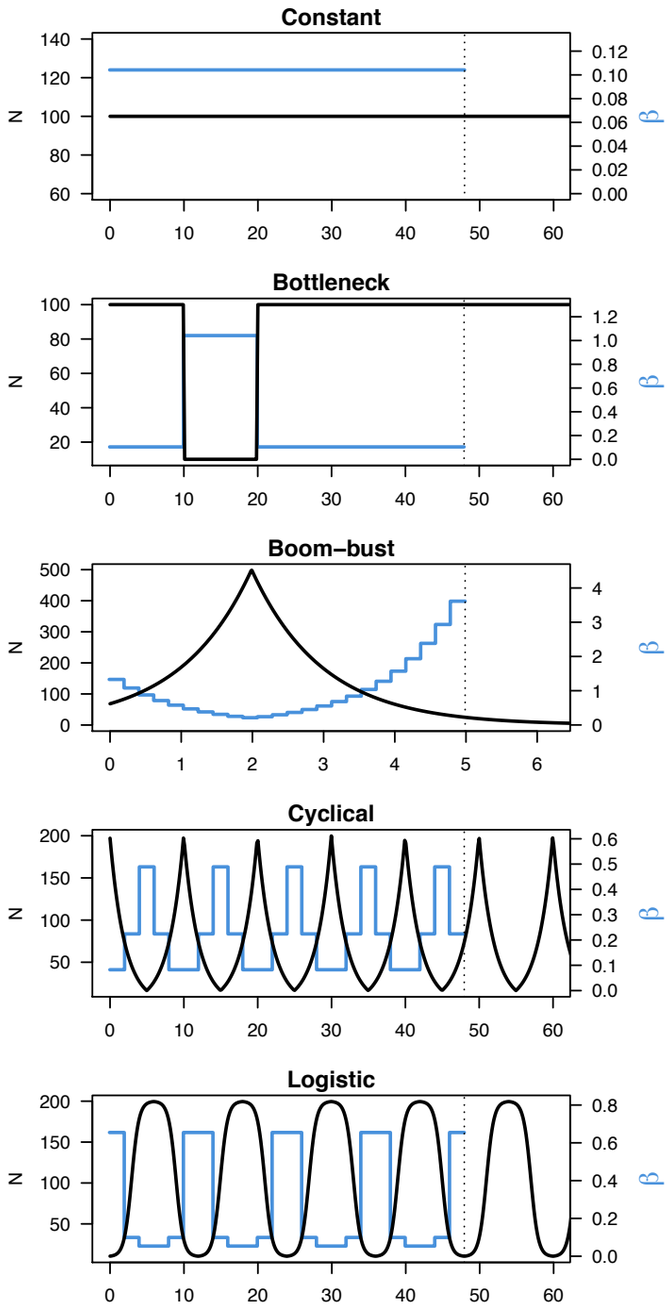
100 replicates / population trajectory using BEAST2 implementation

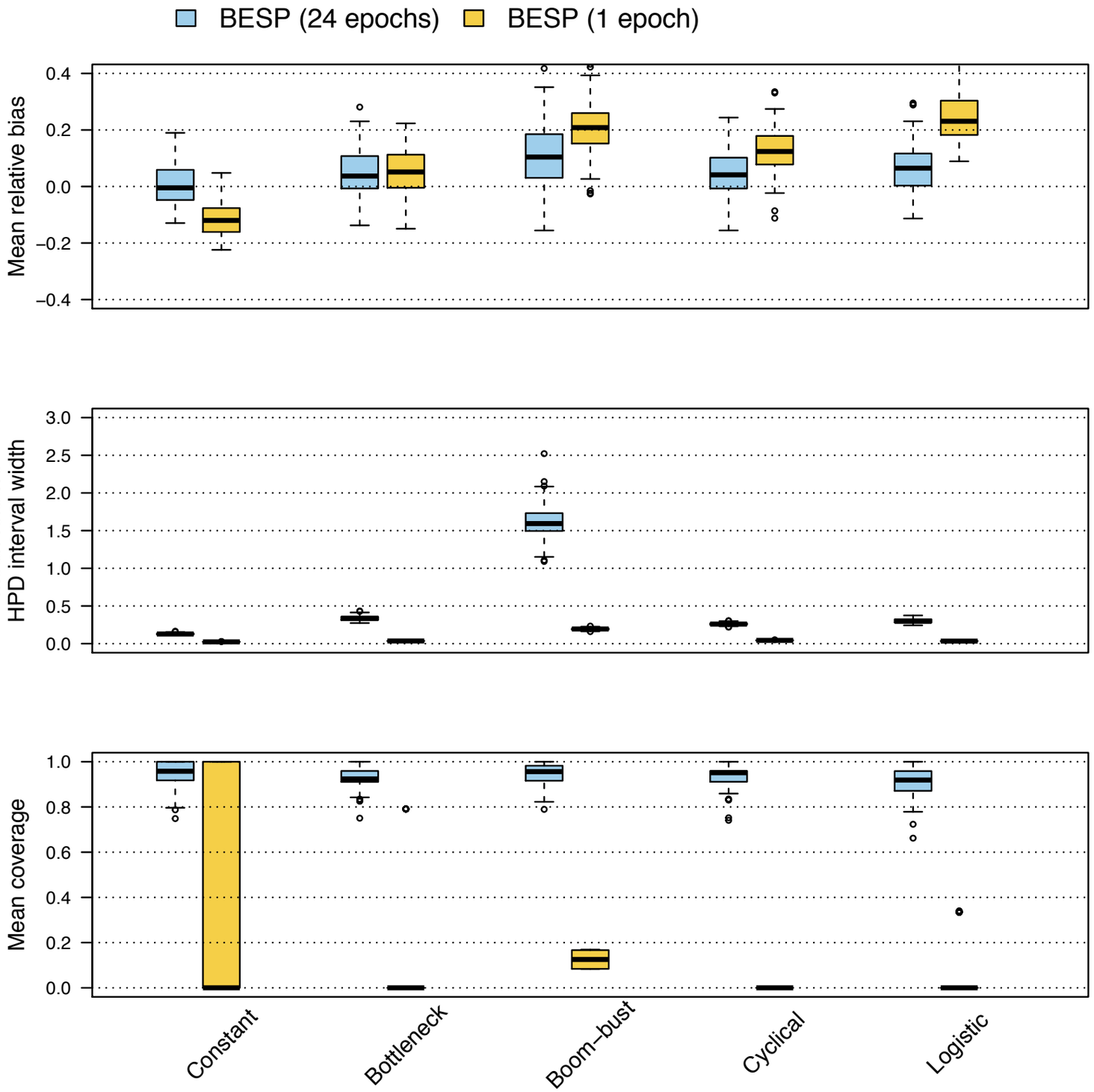




Simulation study **(sampling intensity — β)**

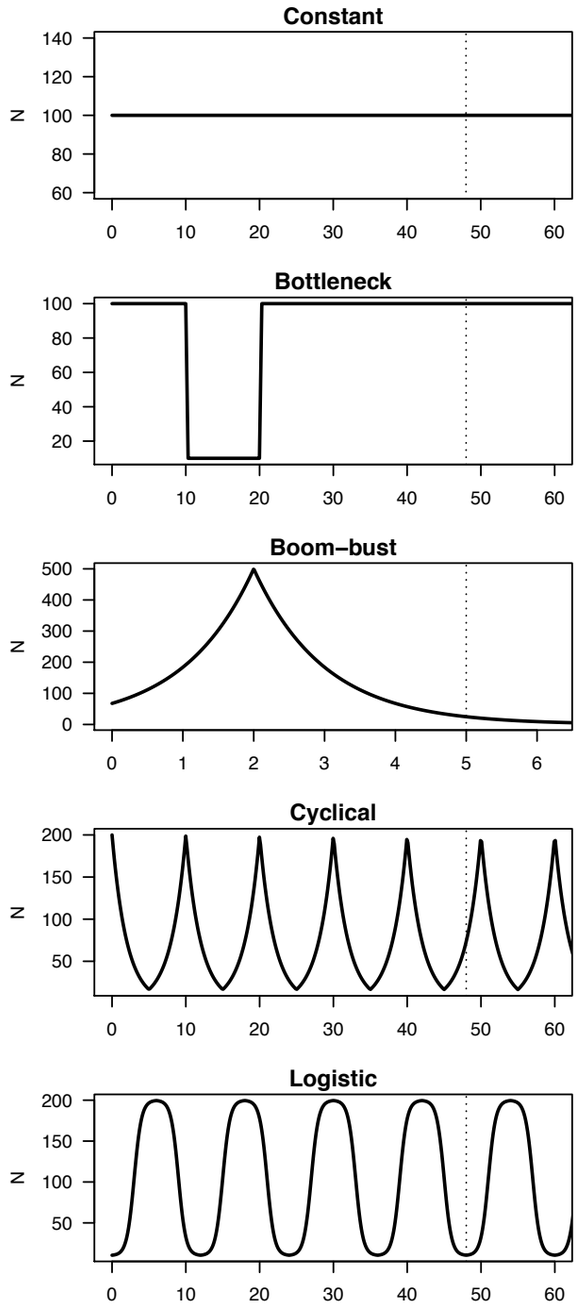
100 replicates / population trajectory using BEAST2 implementation

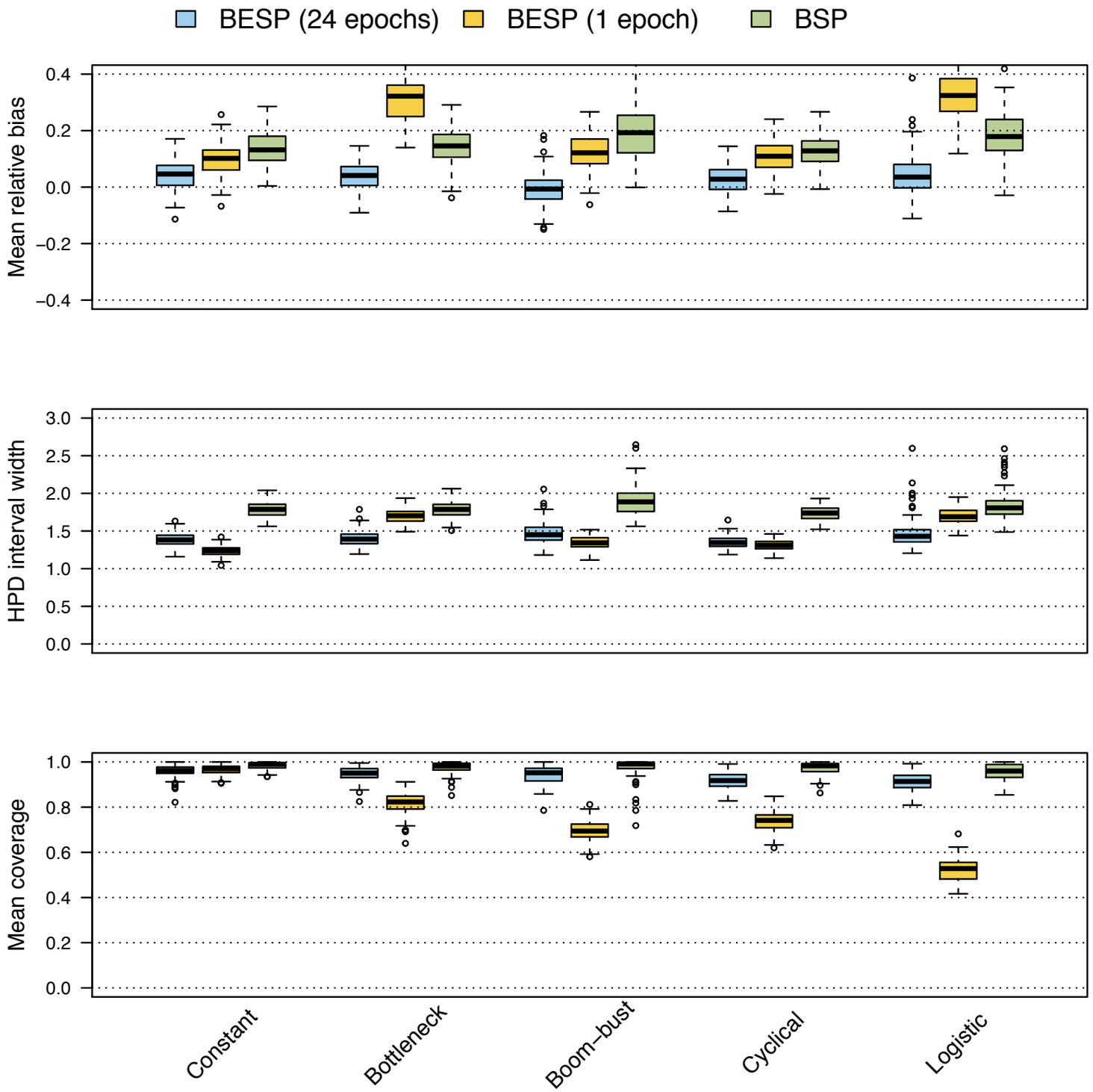




Simulation study **(population size — N)**

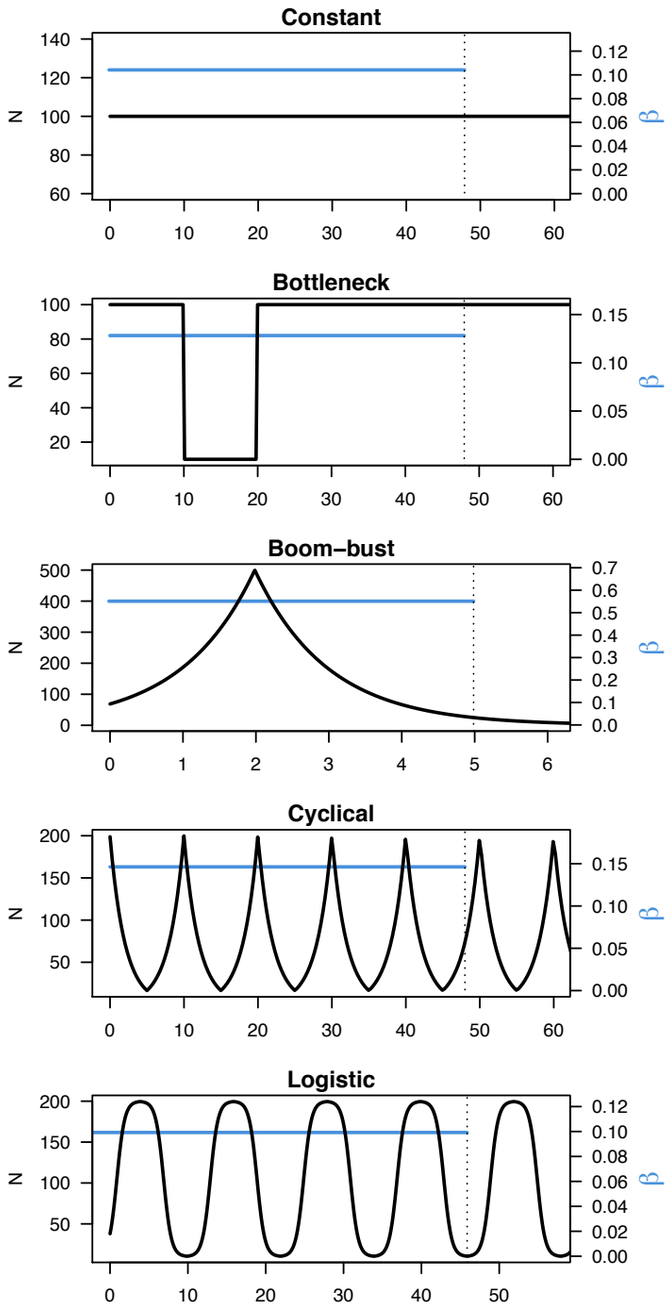
Simulated with **uninformative** sampling times

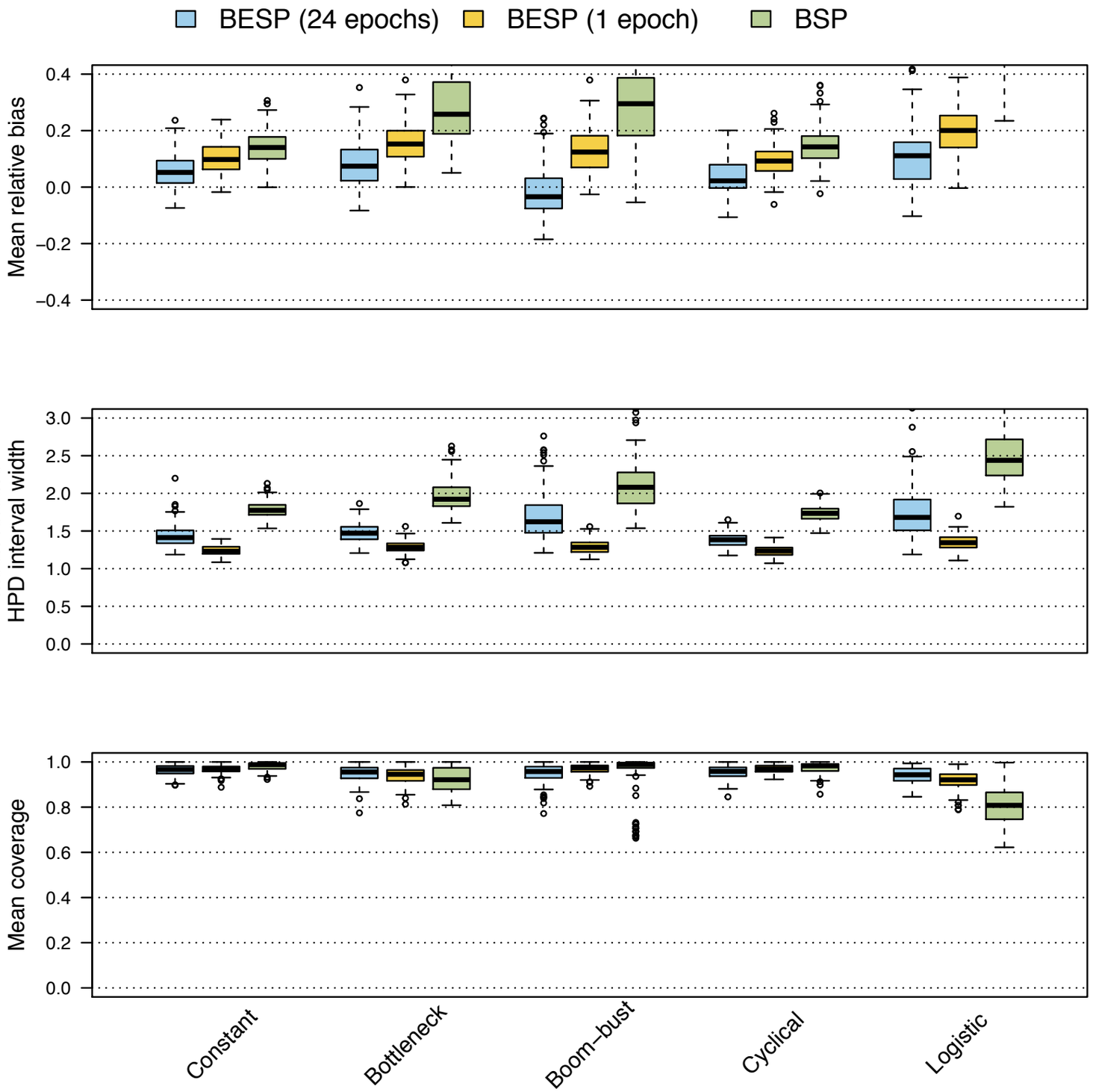




Simulation study **(population size — N)**

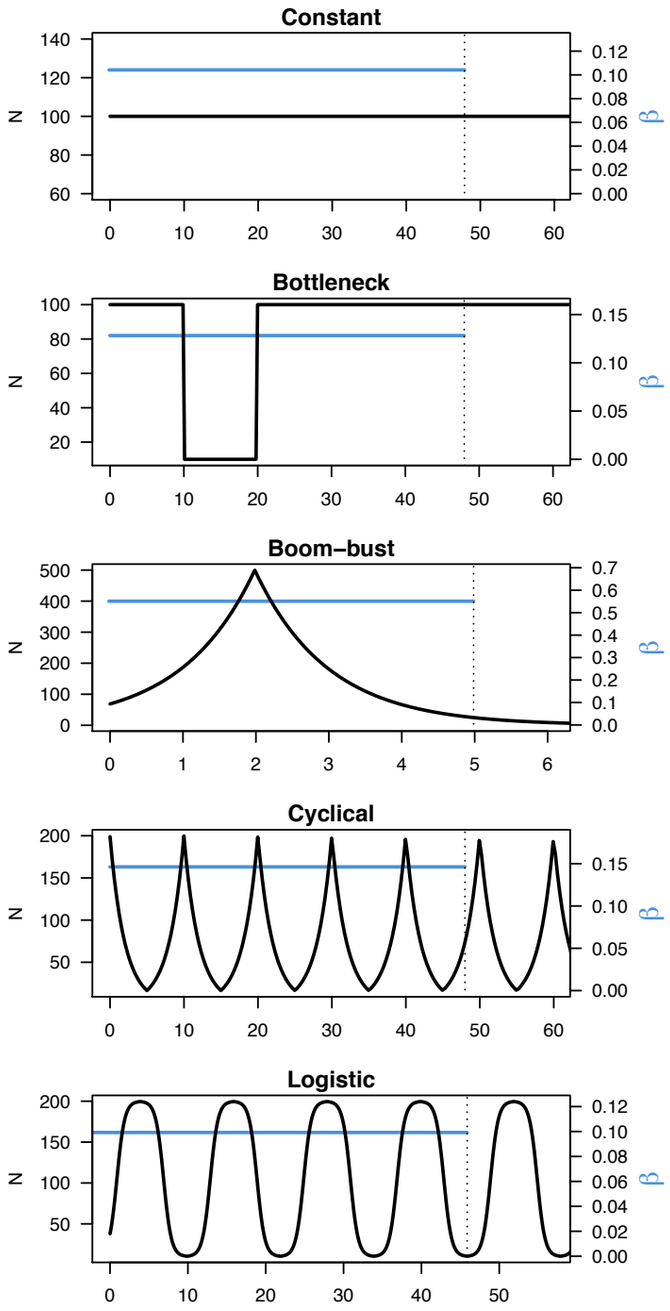
Simulated with **constant** sampling intensity

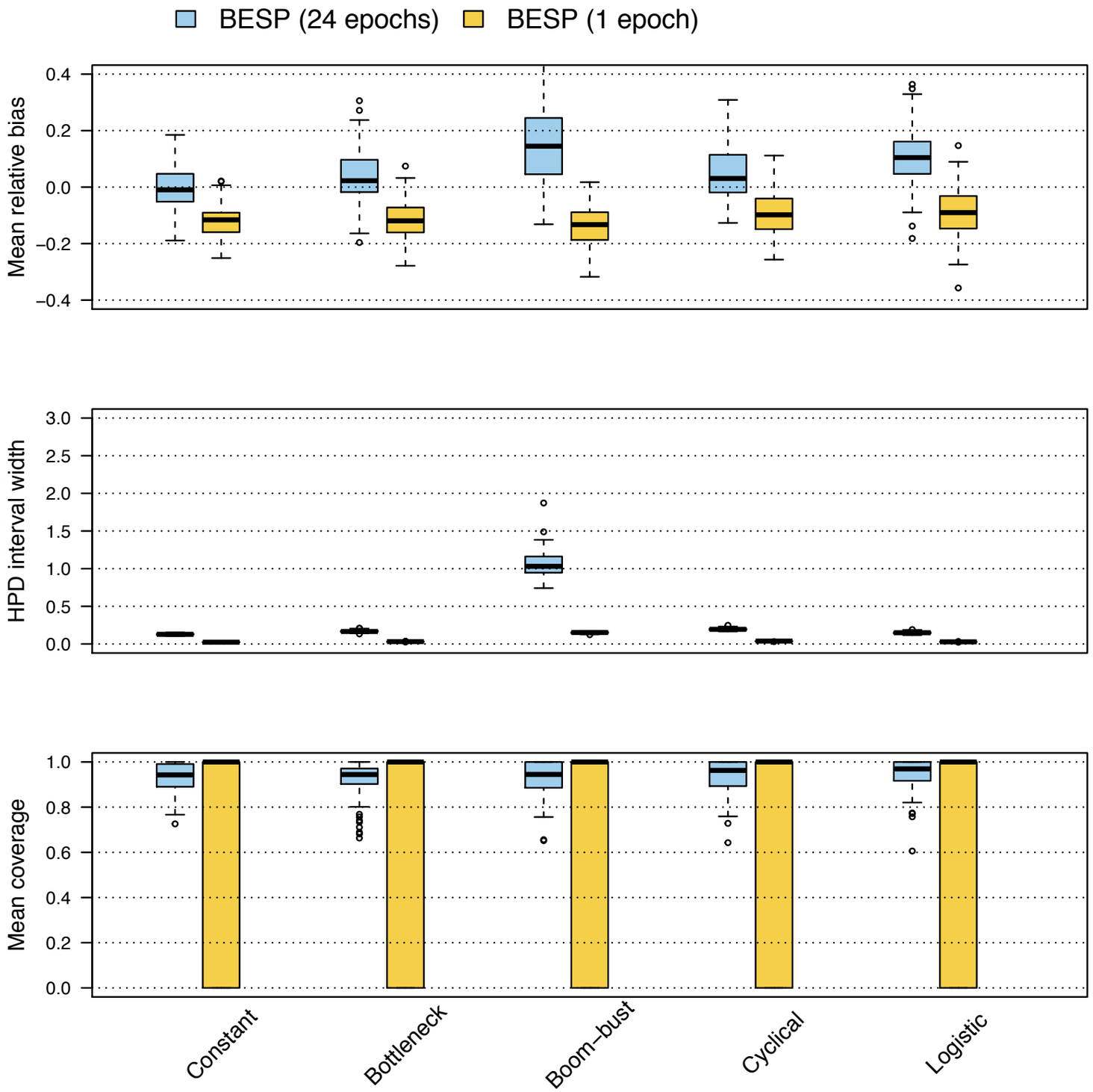




Simulation study **(sampling intensity — β)**

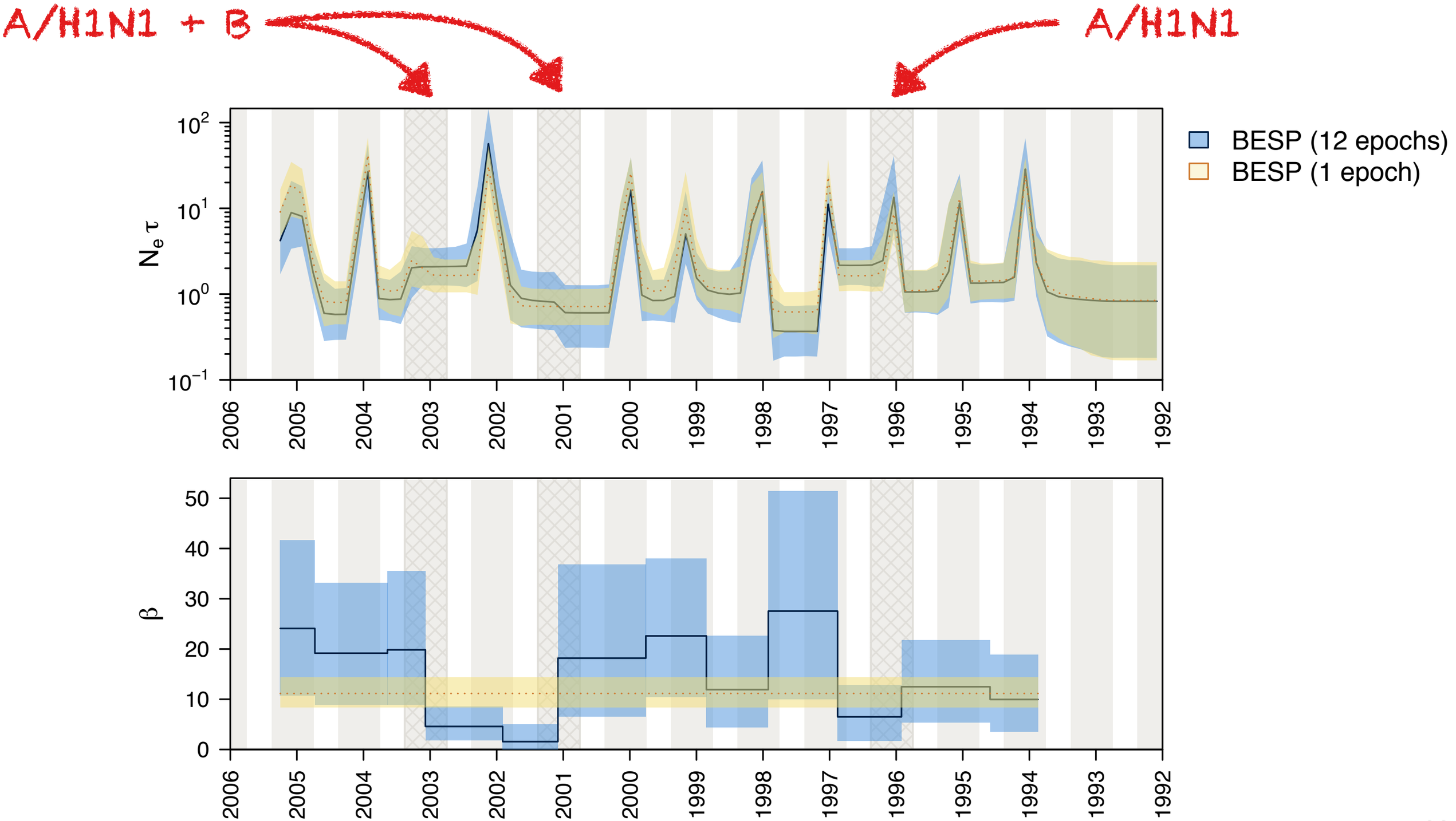
Simulated with **constant** sampling intensity





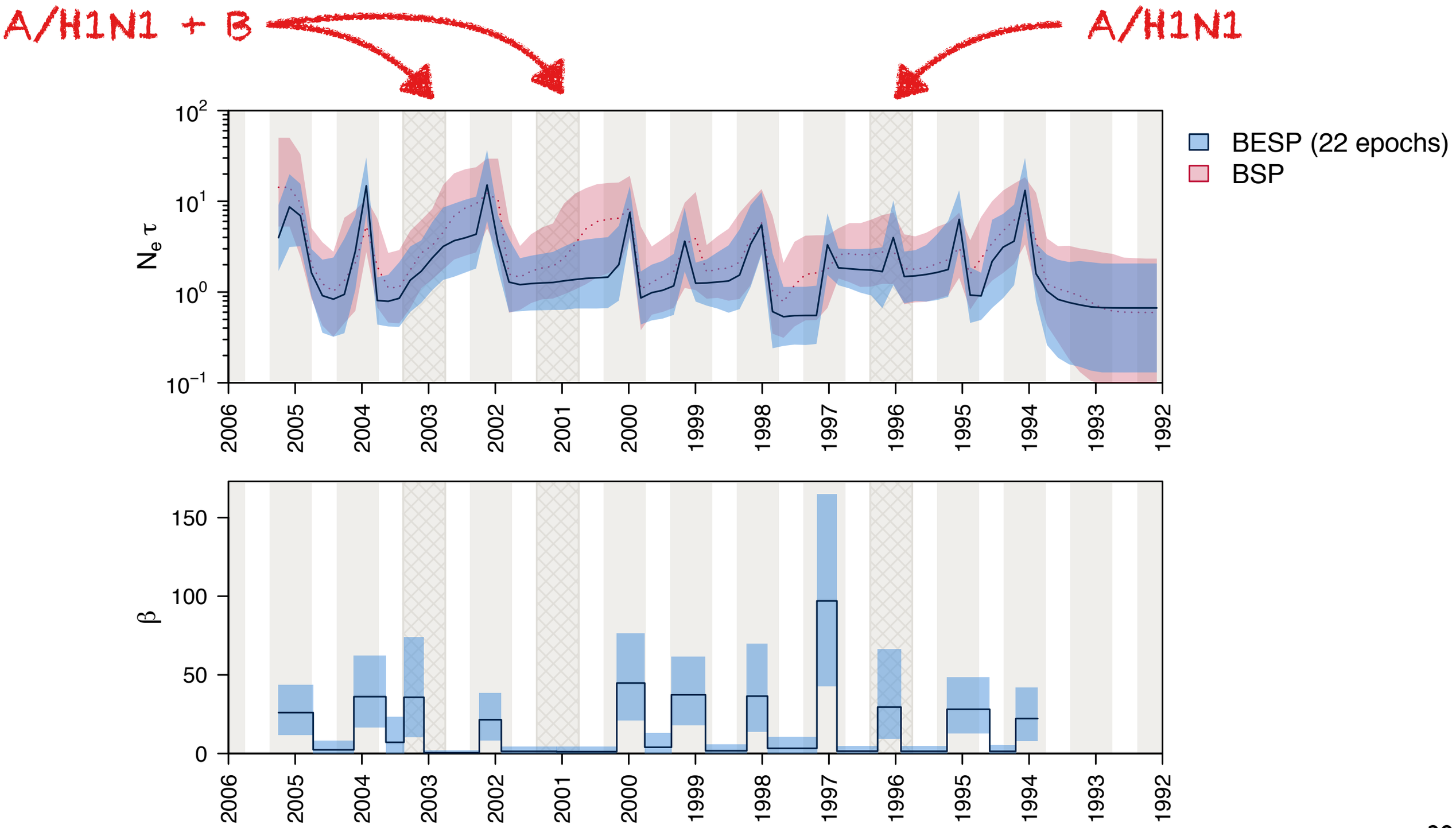
Bayesian Epoch Sampling Skyline Plot

637 New York Influenza A/H3N2 HA sequences across 12 seasons



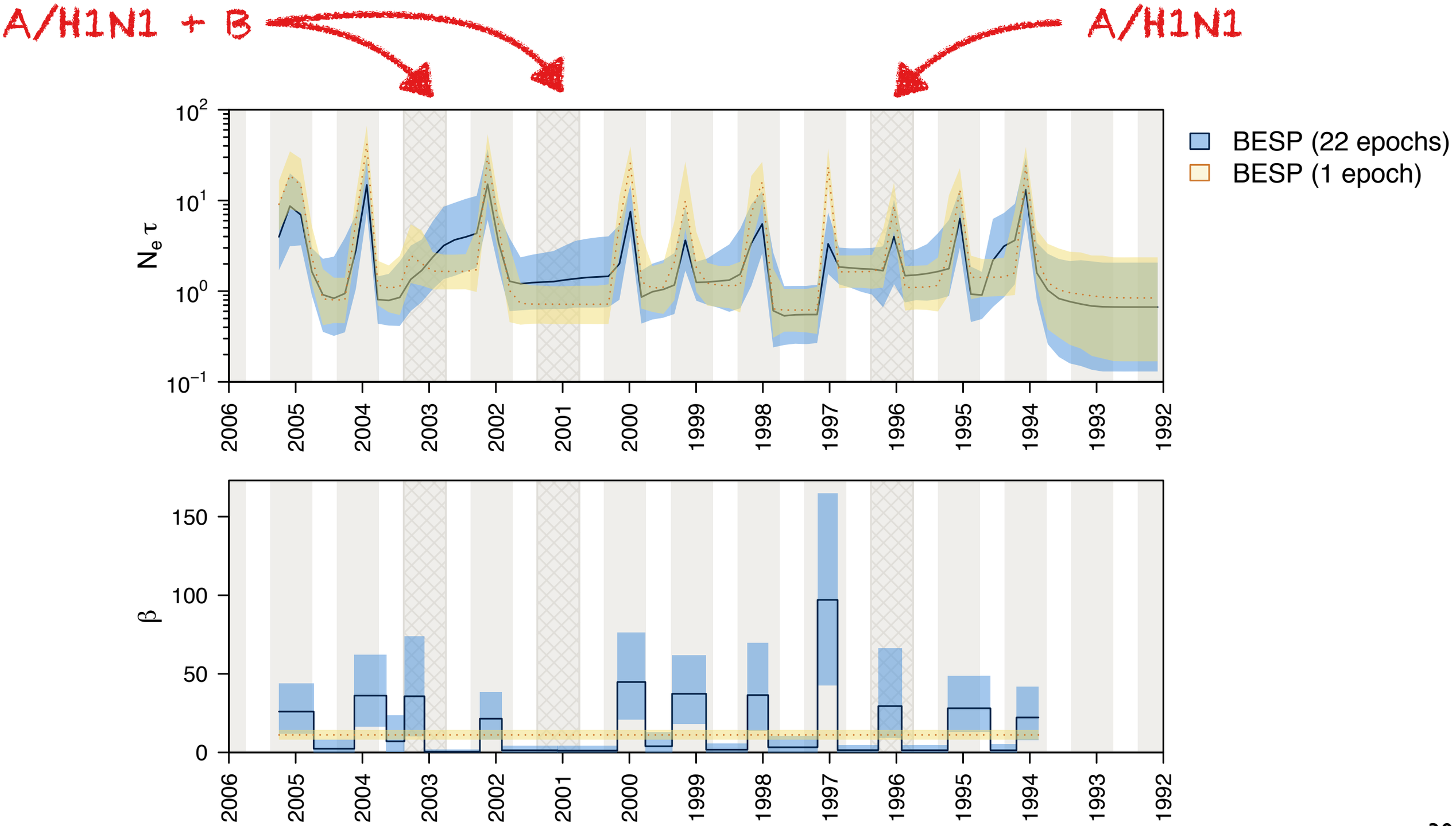
Bayesian Epoch Sampling Skyline Plot

637 New York Influenza A/H3N2 HA sequences across 12 seasons

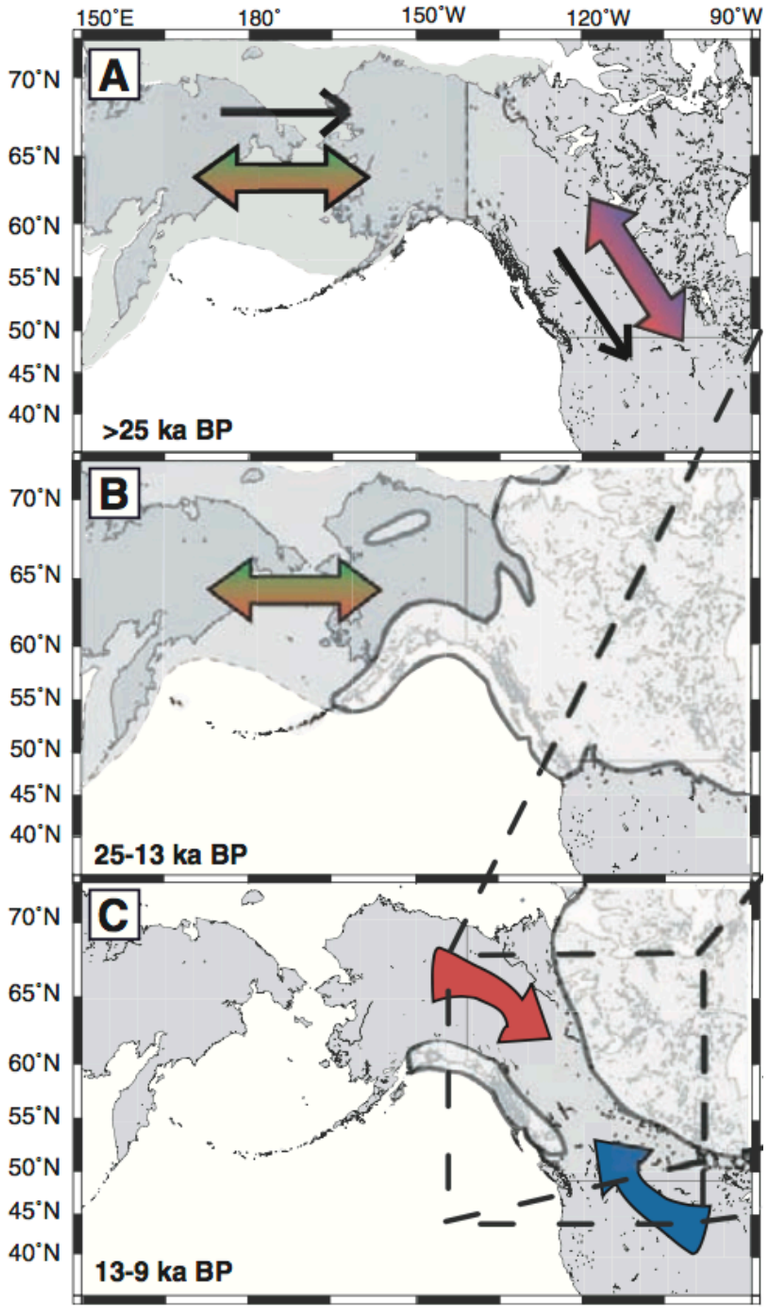


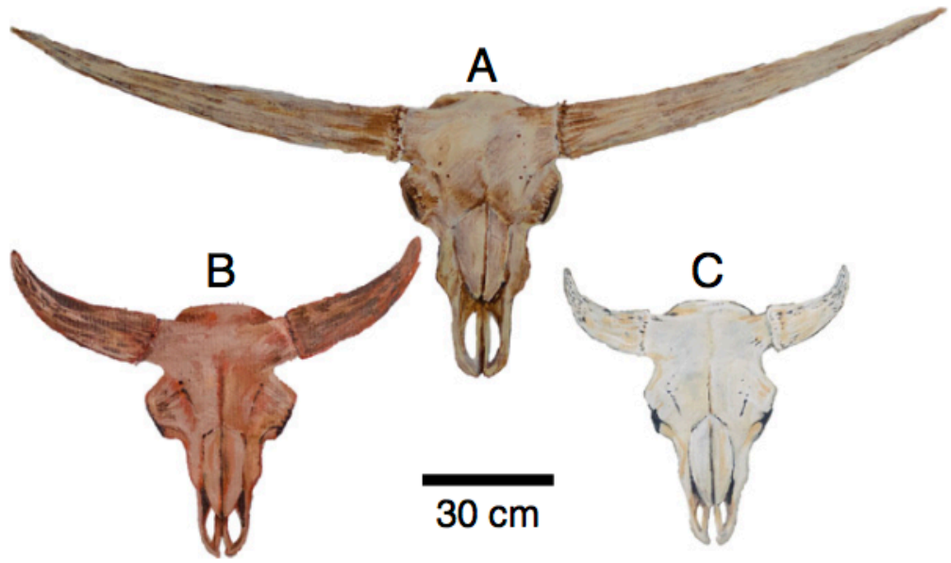
Bayesian Epoch Sampling Skyline Plot

637 New York Influenza A/H3N2 HA sequences across 12 seasons



Bison in Beringia



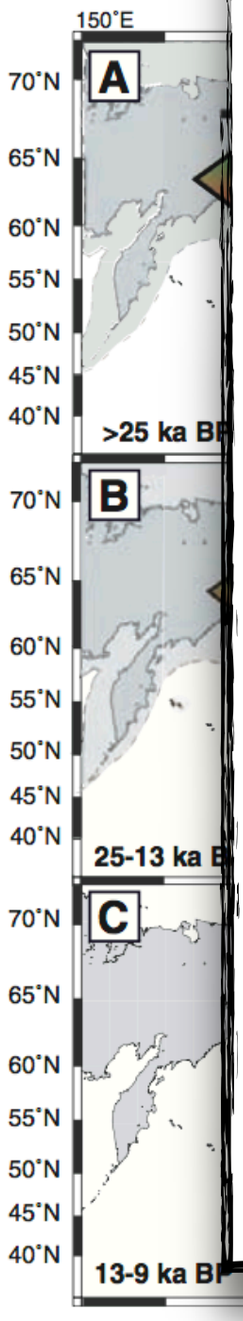


- Some of the most abundant megafauna in Beringia
- Colonised North America from Eurasia
- Big population bottleneck after the Late Quaternary Extinction event
- Extinct in Eurasia but survived in North America

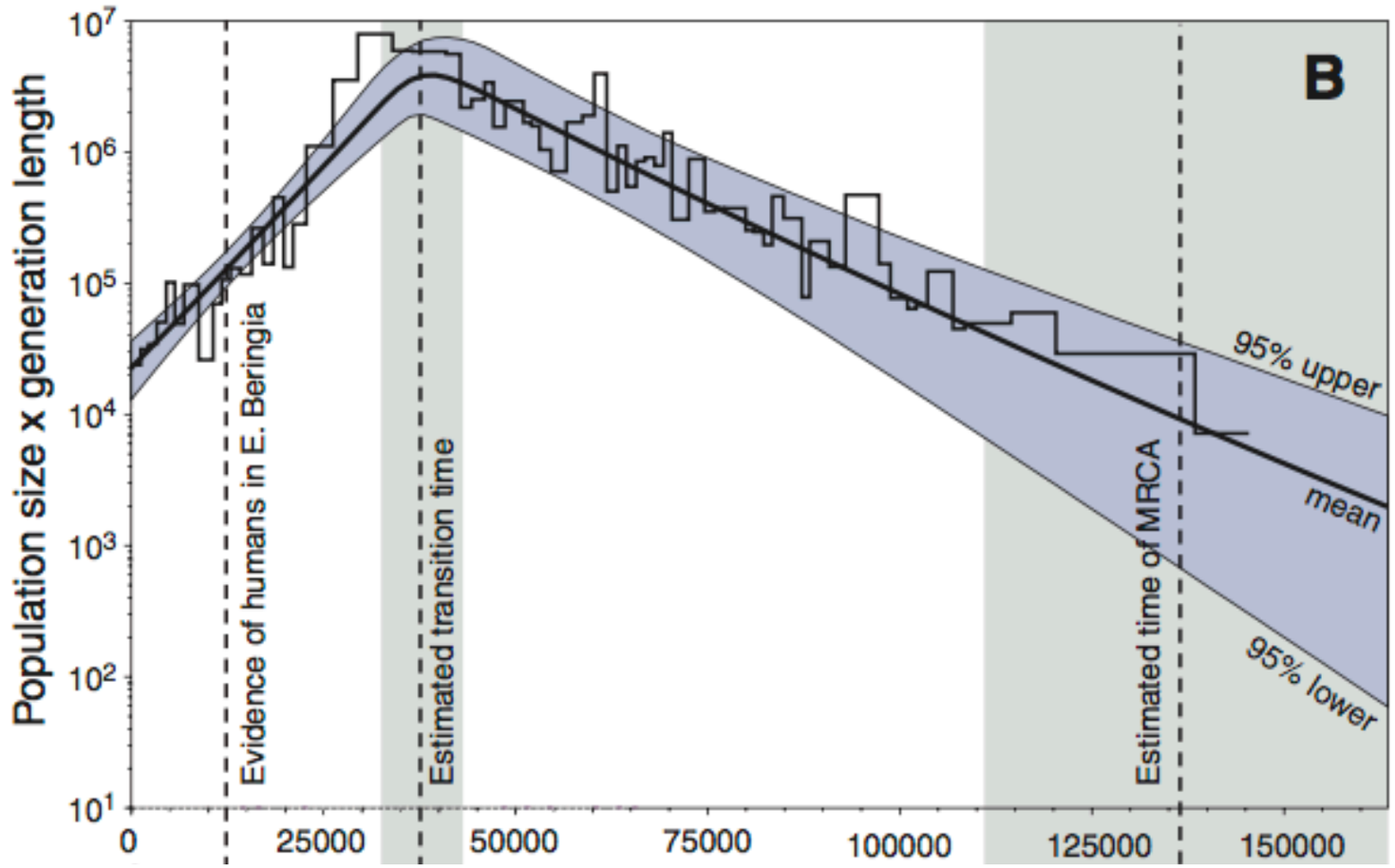
Shapiro *et al.* **Science 2004**

Froese *et al.* **PNAS 2017**

Bison in Beringia



Demographic history



Parametric boom-bust coalescent model

urasia

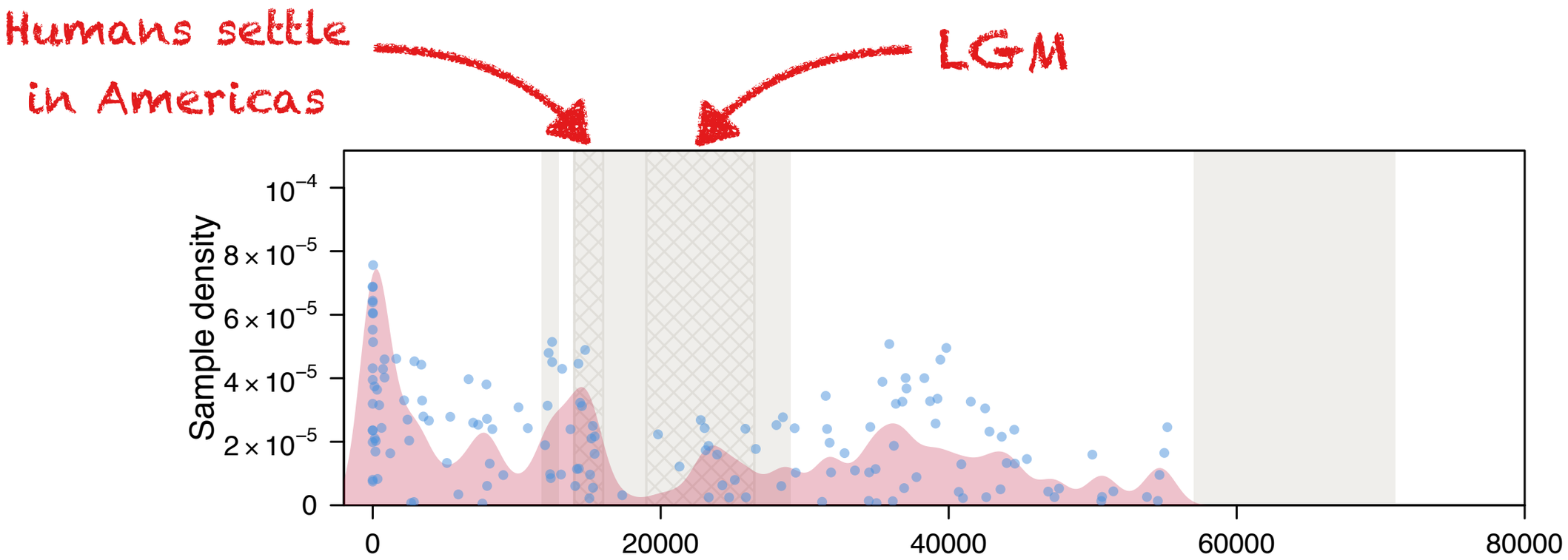
he

Shapiro *et al.* **Science 2004**

Froese *et al.* **PNAS 2017**

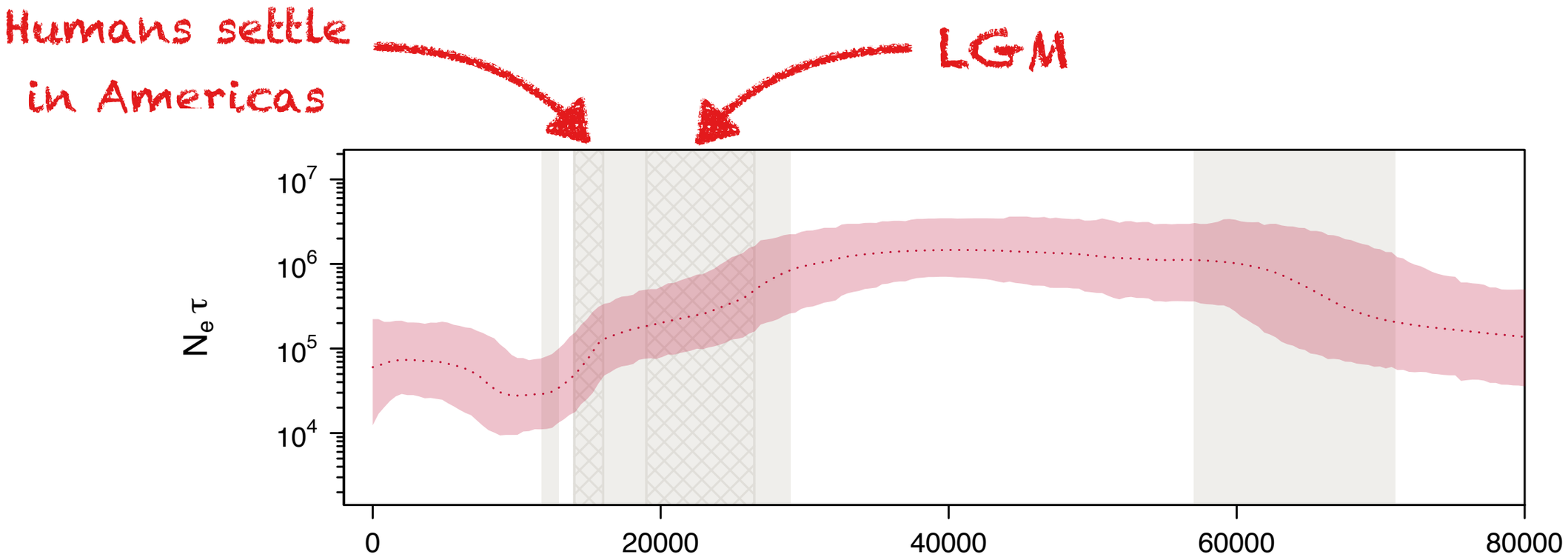
Bison in Beringia

152 mtDNA sequences across 55,000 years



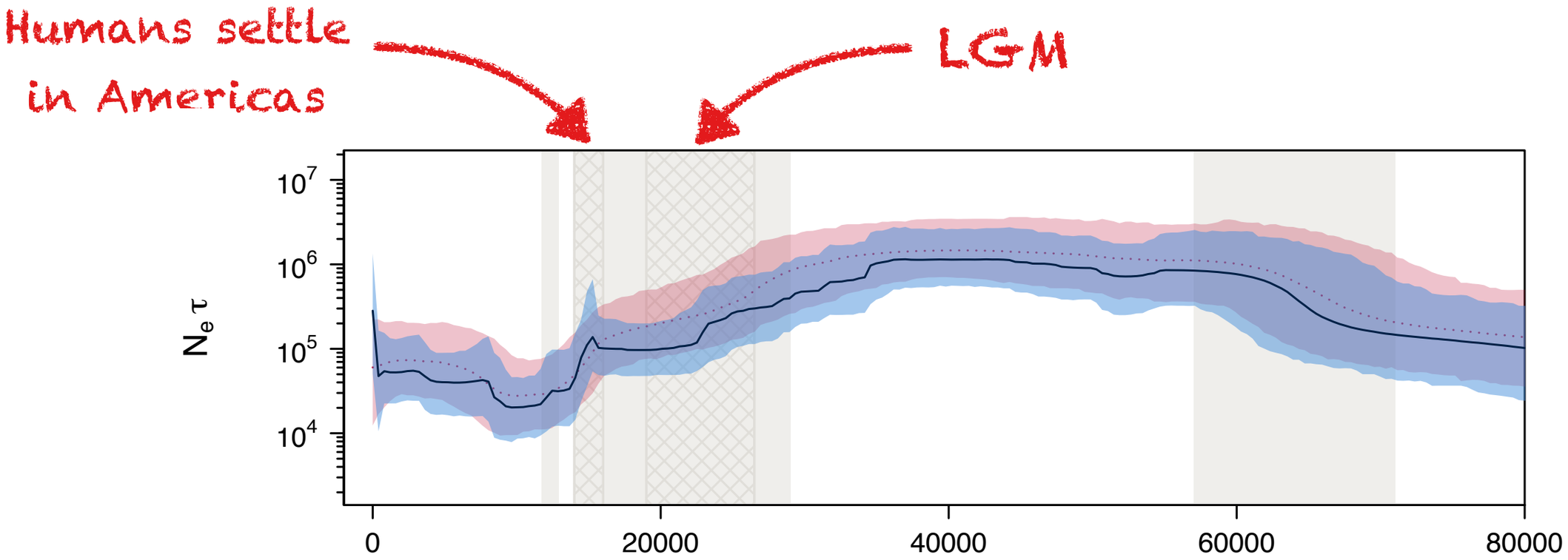
Bison in Beringia

152 mtDNA sequences across 55,000 years



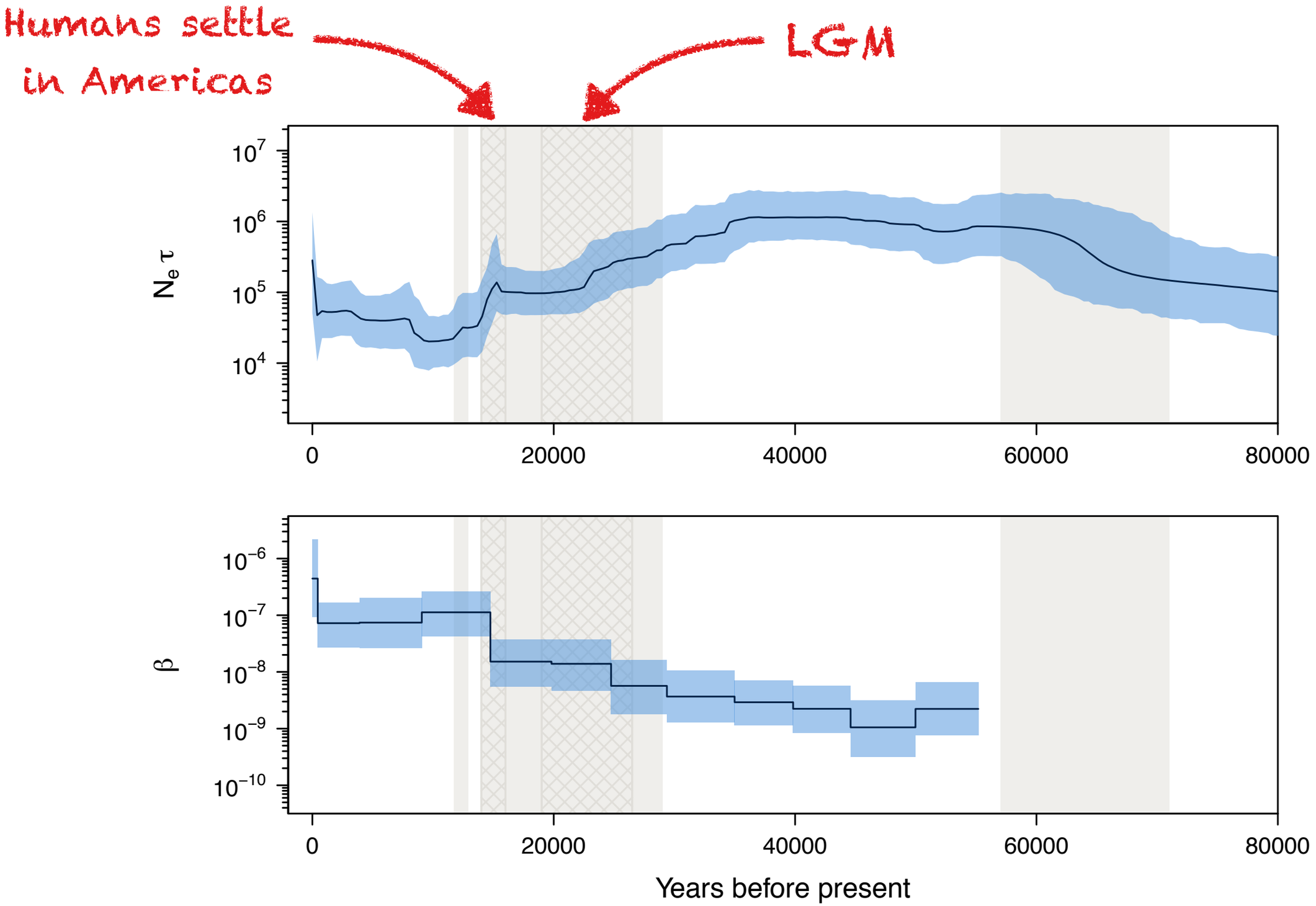
Bison in Beringia

152 mtDNA sequences across 55,000 years



Bison in Beringia

152 mtDNA sequences across 55,000 years



Not accounting for changes in sampling intensity gives biased results!