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]*

1 Department of Statistics, University of Washington, Seattle, Washington, United States of America, **2** Department of Organismic and Evolutionary Biology, Harvard University, Cambridge, Massachusetts, United States of America, **3** Department of Ecology and Evolutionary Biology, Brown University, Providence, Rhode Island, United States of America, **4** Center for Computational Molecular Biology, Brown University, Providence, Rhode Island, United States of America, **5** Vaccine and Infectious Disease Division, Fred Hutchinson Cancer Research Center, Seattle, Washington, United States of America, **6** Department of Human Genetics, David Geffen School of Medicine at UCLA, University of California, Los Angeles, Los Angeles, CA, United States of America, **7** Department of Biomathematics, David Geffen School of Medicine at UCLA, University of California, Los Angeles, Los Angeles, CA, United States of America, **8** Department of Biostatistics, UCLA Fielding School of Public Health, University of California, Los Angeles, Los Angeles, California, United States of America, **9** Department of Biology, University of Washington, Seattle, Washington, United States of America

* vminin@uw.edu

Abstract

Phylodynamics seeks to estimate effective population size fluctuations from molecular sequences of individuals sampled from a population of interest. One way to accomplish this task formulates an observed sequence data likelihood exploiting a coalescent model for the sampled individuals' genealogy and then integrating over all possible genealogies via Monte Carlo or, less efficiently, by conditioning on one genealogy estimated from the sequence data. However, when analyzing sequences sampled serially through time, current methods implicitly assume either that sampling times are fixed deterministically by the data collection protocol or that their distribution does not depend on the size of the population. Through simulation, we first show that, when sampling times do probabilistically depend on effective population size, estimation methods may be systematically biased. To correct for this deficiency, we propose a new model that explicitly accounts for preferential sampling by modeling the sampling times as an inhomogeneous Poisson process dependent on effective population size. We demonstrate that in the presence of preferential sampling our new model not only reduces bias, but also improves estimation precision. Finally, we compare the performance of the currently used phylodynamic methods with our proposed model through clinically-relevant, seasonal human influenza examples.

Author Summary

Phylodynamics seeks to estimate changes in population size from genetic data sampled from individuals across a particular population. One approach to accomplish this task uses a model called the coalescent, which relates the shape of the individuals' shared ancestral tree to genetic diversity, which is in turn related to population size. However, when analyzing genetic data sampled at different times, current techniques assume that sampling times are fixed ahead of time or are distributed randomly without any relation to the size of the population. Through simulation, we show that when sampling times are related to population size, a situation referred to as preferential sampling, those estimation methods may be systematically biased. To fix this problem, we propose a new model that explicitly accounts for and models the preferential sampling. We show that in the presence of preferential sampling our new technique not only fixes the bias, but also has improved precision in its population size estimates. We also compare the performance of the old and new techniques on several real-world seasonal human influenza examples.

Introduction

Phylodynamics—a set of techniques for estimating population dynamics from genetic data—has proven useful in ecology and epidemiology [1, 2]. Phylodynamics is especially useful in cases where ascertaining population sizes via traditional sampling methods is infeasible; e.g., in infectious disease epidemiology it is impossible to obtain the total number of infected individuals in a large population. Estimating population dynamics from a limited sample of genetic data is possible because changes in population size leave evidence in the molecular sequences of the population. Recently, techniques employing a nonparametric approach to inferring population trajectories have improved upon earlier models in terms of flexibility, accuracy, and precision by, e.g., employing Gaussian Markov random fields [3, 4] and Gaussian processes [5]. However, none of these state-of-the-art methods currently account for randomness in sampling time data, potentially introducing bias in studies where sampling times have a relationship to population dynamics. Through a simulation study we characterize this bias in a demographic scenario with seasonally varying population size. We also extend the state-of-the-art by incorporating a sampling time model into phylodynamic inference, mitigating the bias and improving precision.

Phylodynamic methods use Kingman's coalescent model that, given a particular effective population size trajectory, defines the density of a genealogy relating the sampled individuals [6]. Effective population size measures genetic diversity present in the population and relates to census population size if certain assumptions are met [7]. Many early coalescent-based phylodynamic methods required strict parametric assumptions about the effective population size trajectory, such as constant through time [8] or exponential growth [9, 10]. A major alternative arose with the advent of nonparametric methods, one of the earliest and most influential being the piecewise constant classical skyline model [11]. This approach greatly increases the number of estimated parameters, leading to noisy effective population size trajectories. A number of algorithms seeking compromise between the relative stability of parametric approaches and the flexibility of nonparametric approaches have been implemented [3, 4, 12]. For a detailed comparison, see [13].

Many successful applications of phylodynamics methodology come from infectious disease epidemiology, where the effective population size is interpreted, albeit with caution, as the effective number of infections [14]. In these epidemiological applications, disease agent DNA

or RNA sequences are collected at multiple times. When analyzing such heterochronous data, researchers implicitly assume that sampling times are either fixed or follow a distribution that is functionally independent of the effective population size trajectory. However, it is conceivable that the infectious disease agent DNA samples are collected more frequently when the number of infections is high and less frequently during time periods with few infections. Therefore, the implicit assumption of no relationship between sampling times and population dynamics, made by all state-of-the-art phylodynamic methods, is troublesome, since unrecognized preferential sampling leads to systematic estimation bias, as explored by Diggle et al. [15] in the context of spatial statistics. Furthermore, preferential sampling could be present in the sequence databases, but it could also be introduced accidentally or intentionally by filtering during database queries or data mining.

To test the effect of preferential sampling on phylodynamic inference we first perform a simulation study. We simulate sampling times according to multiple distributions, contrasting distributions functionally dependent on effective population size with a functionally independent distribution. We then simulate genealogies based on the sampling times and perform state-of-the-art phylodynamic analyses, and we find that ignoring preferential sampling can bias effective population size estimation and that the size of the bias depends on the local properties of the effective population size trajectory.

In order to account for preferential sampling, we formulate a new phylodynamic model in which sampling times are generated from an inhomogeneous Poisson process with intensity functionally dependent on effective population size. Our model is similar to the augmented coalescent model of Volz and Frost, who work with a specific parametric model [16]. In contrast, we work within a nonparametric framework by incorporating our Poisson preferential sampling model into a Gaussian process-based Bayesian phylodynamic method [3–5]. Applying our new sampling-aware method to our simulations shows that modeling preferential sampling eliminates the aforementioned bias and can increase precision of the phylodynamic inference. In all of our developments, we assume that the genealogy of the sample is known without error. This assumption allows us to use an integrated nested Laplace approximation (INLA) to make our Bayesian inference computationally efficient [17, 18], which is important for executing our simulation studies.

Finally, we examine the performance of our algorithm on two real-world examples. Rambaut et al.[19] explore the seasonal variation of genetic diversity in the genes that code for several of the most important proteins in the two most common influenza subtypes, H3N2 and H1N1. For the sake of brevity we only analyze the hemagglutinin gene in H3N2. We find evidence of preferential sampling in the dataset, and our sampling-aware method produces a large improvement in precision over the conditional (sampling un-aware) method. Zinder et al.[20] specifically explore the patterns of seasonal migration of genetic diversity of H3N2 influenza across the regions of the world. We examine the regions separately and find differing strengths of preferential sampling, but in all regions our method performs better than the conditional model. In some regions, we see stronger relationships between sampling frequency and population size, most often in regions with the most seasonal variation in incidence.

Methods

State-of-the-art phylodynamics

Consider a sample of individuals from a well-mixed population. Some individuals will share a common ancestor more recently than others. One pair of individuals in particular will have the pairwise most recent common ancestor. Moving backwards in time, we can consider those two individuals to have *coalesced*, treating the two individuals as one. We can then repeat this

process of finding the pairwise most recent common ancestor and coalescing individuals until we reach the most recent common ancestor of the entire sample. If we keep track of the ancestral lineages and coalescences of the individuals, we see the data take the shape of a bifurcating tree, and we refer to this ancestry tree as a *genealogy* (illustrated in Fig 1).

We refer to the branching points of the genealogy tree as *coalescent events*. If the samples are all taken simultaneously, we refer to the genealogy as *isochronous*. Kingman's original coalescent provided a density for isochronous genealogies with a fixed effective population size [6]. Later extensions to the coalescent allowed for parametric and nonparametric specifications of effective population size trajectories along with *heterochronous* sampling times. Heterochronous sampling times (also called sampling events) can occur at any time up to the present.

We consider first the case of a fixed, heterochronous genealogy [21]. The coalescent likelihood has sufficient statistics $\mathbf{g} = \{t_i\}_{i=1}^{n}, 0 = t_n < t_{n-1} < \ldots < t_1$, representing the coalescent times, and $\mathbf{s} = \{s_i, n_i\}_{i=1}^{m}, 0 = s_m < s_{m-1} < \ldots < s_1, \sum_{j=1}^{m} n_j = n$, representing the sampling times along with the corresponding number of lineages sampled (see Fig 1). We define the number of *active lineages* at time t as the number of lineages sampled between t and the present, minus the number of coalescent events between t and the present. In Fig 1, this appears as the number of horizontal lines that a vertical line at time t will cross.

We define a partition of $(0, t_1)$ with intervals $I_{i,k}$ for $k = 1, \ldots, n$. We let $I_{0,k}$ represent the intervals ending with a coalescent event and let $I_{i,k}$ for $i = 1, \ldots, m_k$ represent the m_k intervals ending in a sampling event between the $(k - 1)$th and kth coalescent events (see Intervals in Fig 1). We let $C_{i,k} = \binom{n_{i,k}}{2}$, where $n_{i,k}$ is the number of active lineages in the interval $I_{i,k}$. Suppose $\mathbf{s}$ is fixed, then the coalescent likelihood is

$$\Pr\left[\mathbf{g} | N_e(t), \mathbf{s}\right] \propto \prod_{k=2}^{n} \frac{C_{0,k}}{N_e(t_{k-1})} \exp\left[-\sum_{i=0}^{m_k} \int_{I_{i,k}} \frac{C_{i,k}}{N_e(t)} dt\right].$$

In Bayesian phylodynamic inference, our aim is to explore the posterior distribution of the effective population size trajectory $N_e(t)$, so we employ a Gaussian process prior $\Pr[N_e(t)|\tau]$, where $N_e(t) = \exp[\gamma(t)]$, with $\gamma(t) \sim \mathcal{BM}(\tau)$ following a Brownian motion with precision parameter τ[18]. We assign a Gamma(0.01, 0.01) hyperprior to τ. This results in the posterior $\Pr[N_e(t), \tau|\mathbf{g}] \propto \Pr[\mathbf{g}|N_e(t)]\Pr[N_e(t)|\tau]\Pr(\tau)$.

The continuous case as written above involves an infinite-dimensional object—the function $N_e(t)$—which makes the problem as stated intractable. However, we can approximate the continuous function with a piecewise constant function. We construct a fine, regular grid $\mathbf{x} = \{x_j\}_{j=1}^{B}$ with grid width w over the interval that supports the genealogy and let $\gamma_j = \log[N_e(x_j)]$. We construct a piecewise constant approximation $N^\gamma(t) = \sum_{i=1}^{B} \exp(\gamma_i) 1_{t \in [x_i - w/2, x_i + w/2)}$. The discretized coalescent likelihood becomes

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

where $\gamma = (\gamma_1, \ldots, \gamma_B)$ and the integrals are simple to compute over the step function $N^\gamma(t)$. We discretize the Brownian process prior with an intrinsic random walk prior,

$$\Pr\left(\gamma \mid \tau\right) \propto \tau^{(n-1)/2} \exp\left[-\frac{\tau}{2} \sum_{k=1}^{B-1} (\gamma_{k+1} - \gamma_k)^2\right].$$

Finally, the discretized posterior becomes $\Pr(\gamma, \tau|\mathbf{g}) \propto \Pr(\mathbf{g}|\gamma)\Pr(\gamma|\tau)\Pr(\tau)$.
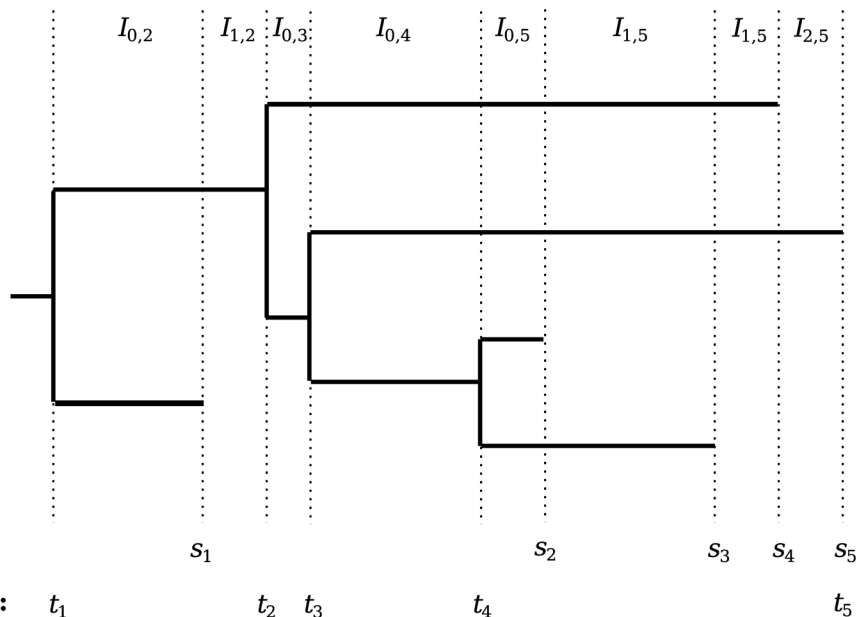
Fig 1. Illustration of an example heterochronous genealogy with $n = 5$ lineages. Sampling times $s_1, \ldots, s_5$ and coalescent times $t_1, \ldots, t_5$ are marked below the genealogy.

With the posterior known (up to a proportionality constant), we can proceed with numerical integration techniques such as Markov chain Monte Carlo (MCMC) or INLA—a deterministic algorithm for approximating posterior distributions. We select INLA and name the implementation Bayesian nonparametric phylodynamic reconstruction (BNPR).

Phylodynamics with preferential sampling

In the previous section we made the assumption that we could safely ignore any potential dependence of sampling times **s** on effective population size $N^\gamma(t)$ in our calculations. In this section, we relax this assumption. We model sampling times according to an inhomogeneous Poisson process in a fixed sampling window $[0, s_0]$, with intensity $\lambda(t) = \exp(\beta_0)[N^\gamma(t)]^{\beta_1}$, i.e. proportional to a power of the effective population size, where β_0 and β_1 are unknown parameters. The sampling log-likelihood is

$$\log[\Pr(\mathbf{s} \mid \boldsymbol{\gamma}, \beta_0, \beta_1)] = C + n\beta_0 + \sum_{i=1}^{n} \beta_1 \log[N^\gamma(s_i)] - \int_{s_m}^{s_0} \exp(\beta_0)[N^\gamma(r)]^{\beta_1} dr.$$

To illustrate our parameterization, sampling with $\beta_1 = 1$ would result in collecting genetic sequences with intensity directly proportional to effective population size, while higher β_1 values result in more clustered samples. Conversely, $\beta_1 = 0$ produces a uniform distribution of sampling times, with a Poisson distribution on the number of individuals sampled.

In many datasets, the sampling time data will have low enough resolution (for instance, only recording the date but not time of sampling) that some sampling times will appear to be coincident. Our sampling model is compatible with simultaneous sampling times because the model naturally bins the samples along our earlier discretization. The likelihood is proportional to a product of Poisson mass functions centered at the grid points **x**.

The genealogy depends on the sampling times, so we condition on **s** in the likelihood for **g**. We are treating **s** as random, so we insert the likelihood term for it as well as independent

Normal priors for parameters β_0 and β_1—specifically $\beta_i \sim N(\text{mean} = 0, \text{variance} = 1000)$ for $i = 0, 1$. We retain the same hyperprior for the precision parameter τ as above. This results in the posterior that accounts for preferential sampling,

$$\Pr(\boldsymbol{\gamma}, \tau, \boldsymbol{\beta} \mid \mathbf{g}, \mathbf{s}) \propto \Pr(\mathbf{g} \mid \mathbf{s}, \boldsymbol{\gamma}) \Pr(\mathbf{s} \mid \boldsymbol{\gamma}, \boldsymbol{\beta}) \Pr(\boldsymbol{\gamma} \mid \tau) \Pr(\tau) \Pr(\beta),$$

where $\Pr(\mathbf{g}|\mathbf{s}, \boldsymbol{\gamma})$ is defined by Eq (1), but now we add conditioning on $\mathbf{s}$ explicitly. In the case where the density of sampling times $\mathbf{s}$ is functionally independent of the vector of log effective population sizes $\boldsymbol{\gamma}$, the posterior for $\mathbf{g}$ simplifies to the form it had in the previous section, because the likelihood for $\mathbf{s}$ becomes a constant in $\boldsymbol{\gamma}$. We incorporate our sampling model into an INLA framework similar to BNPR and name the implementation Bayesian nonparametric phylodynamic reconstruction with preferential sampling (BNPR-PS).

INLA framework

Here we present a brief outline of the INLA methodology [17] in the context of our BNPR and BNPR-PS implementations. We first examine BNPR as the simpler model. In the end, we intend to estimate the marginal posteriors of the precision hyperparameter $\Pr(\tau \mid \mathbf{g})$ and the latent points $\Pr(\gamma_i \mid \mathbf{g})$, $i = 1, \ldots, B$, most often focusing on the posterior medians and the end points of the 95% Bayesian credible intervals. We approximate the marginal of τ with

$$\widehat{\Pr}(\tau \mid \mathbf{g}) \propto \left. \frac{\Pr(\boldsymbol{\gamma}, \tau, \mathbf{g})}{\widehat{\Pr}_{\mathbf{G}}(\boldsymbol{\gamma} \mid \tau, \mathbf{g})} \right|_{\gamma = \gamma^*(\tau)},$$

where $\widehat{\Pr}_{\mathbf{G}}(\boldsymbol{\gamma} \mid \tau, \mathbf{g})$ is the Gaussian approximation generated from a Taylor expansion around $\gamma^*(\tau)$, the mode of $\Pr(\boldsymbol{\gamma}|\tau, \mathbf{g})$ for a given τ. We can find $\gamma^*(\tau)$ using the Newton-Raphson method.

Next, we need to approximate the distribution of γ_i conditional on τ. The simplest method of using the Gaussian approximations above can produce errors [17], so we briefly describe the use of nested Laplace approximations. The full implementation details can be found in [17]. We define

$$\widehat{\Pr}_{LA}(\gamma_i \mid \tau, \mathbf{g}) \propto \left. \frac{\Pr(\boldsymbol{\gamma}, \tau, \mathbf{g})}{\widehat{\Pr}_{\mathbf{GG}}(\boldsymbol{\gamma}_{-i} \mid \gamma_i, \tau, \mathbf{g})} \right|_{\gamma_{-i} = \gamma^*_{-i}},$$

where $\widehat{\Pr}_{\mathbf{GG}}(\boldsymbol{\gamma}_{-i} \mid \gamma_i, \tau, \mathbf{g})$ is a Gaussian approximation of $\Pr(\boldsymbol{\gamma}_{-i}|\gamma_i, \tau, \mathbf{g})$ obtained by a Taylor expansion around $\gamma^*_{-i} = \mathrm{E}_{\mathbf{G}}(\boldsymbol{\gamma}_{-i} \mid \gamma_i, \tau, \mathbf{g})$, which is computed using $\widehat{\Pr}_{\mathbf{G}}(\boldsymbol{\gamma} \mid \tau, \mathbf{g})$. Finally, we normalize and combine the two approximations, then use numerical integration to calculate

$$\widehat{\Pr}(\gamma_i \mid \mathbf{g}) = \int \widehat{\Pr}(\gamma_i \mid \tau, \mathbf{g}) \widehat{\Pr}(\tau \mid \mathbf{g}) d\tau.$$

The outline for BNPR-PS is very similar. The approximate marginal of the hyperparameters is

$$\widehat{\Pr}(\tau, \boldsymbol{\beta} \mid \mathbf{g}, \mathbf{s}) \propto \left. \frac{\Pr(\boldsymbol{\gamma}, \tau, \boldsymbol{\beta}, \mathbf{g}, \mathbf{s})}{\widehat{\Pr}_{\mathbf{G}}(\boldsymbol{\gamma} \mid \tau, \boldsymbol{\beta}, \mathbf{g}, \mathbf{s})} \right|_{\gamma = \gamma^*(\tau, \boldsymbol{\beta})},$$

for similarly defined factors. We take advantage of an INLA extension by Martins et al. [22] that allows for multiple likelihoods. The approximate distribution of γ_i conditional on $\tau, \boldsymbol{\beta}$

becomes

$$\widehat{\Pr}_{LA}(\gamma_i \mid \tau, \boldsymbol{\beta}, \mathbf{g}, \mathbf{s}) \propto \left. \frac{\Pr(\boldsymbol{\gamma}, \tau, \boldsymbol{\beta}, \mathbf{g}, \mathbf{s})}{\widehat{\Pr}_{\mathbf{GG}}(\boldsymbol{\gamma}_{-i} \mid \gamma_i, \tau, \boldsymbol{\beta}, \mathbf{g}, \mathbf{s})} \right|_{\gamma_{-i} = \gamma_{-i}^*},$$

and the final numerical integration is analogously more complex but still tractable, since we integrate over both τ and $\boldsymbol{\beta}$.

We use the R-INLA package [17, 22] to perform the above calculations. We make INLA approximations of BNPR and BNPR-PS posteriors available, along with other phylodynamic tools, in the R package `phylodyn` which can be found at https://github.com/mdkarcher/phylodyn.

Results

Simulation study

We investigate estimating effective population size in the presence of preferential sampling via simulated data. First, we seek to show where and how the model misspecification resulting from ignoring preferential sampling manifests itself in terms of posterior median and Bayesian credible interval width estimation. Our second goal is to show what we gain by properly modeling preferential sampling.

Our primary set of simulation results use the family of seasonally-varying effective population size functions characterized by

$$N_{e,a,o}(t) = \begin{cases} 10 + 90/(1 + \exp\{a[3 - (t + o \pmod{12})]\}), & \text{if } t + o \pmod{12} \leq 6, \\ 10 + 90/(1 + \exp\{a[3 + (t + o \pmod{12}) - 12]\}), & \text{if } t + o \pmod{12} > 6. \end{cases} \quad (2)$$

For all of our experiments, the smoothness parameter $a = 2$ will be used. This family emulates a cyclical population time series with t in nominal months. The shape is loosely modeled after flu seasons, with o controlling which part of the year $t = 0$ represents ($o = 0, 3, 6$ emulates summer, spring, and winter, respectively). We simulate genealogies with varying tip sampling times using two sampling schedules. The uniform schedule distributes n sampling times uniformly throughout a given sampling interval. The proportional schedule distributes sampling times in the sampling interval according to an inhomogeneous Poisson process with intensity proportional to effective population size. The proportionality constant here is tuned to have an expected number of sampling times equal to n.

We explore the properties of our two methods using a Monte Carlo approach. To create a Monte Carlo iteration, we generate our sampling times according to their sampling schedules, then simulate our genealogies using coalescent theory via the rejection sampling method of [5]. Given the genealogy and the samples, we infer the effective population time series, using BNPR and BNPR-PS to approximate grids of marginal posteriors. For each iteration, this gives us approximate estimates of the posterior median and quantiles at each point in the effective population size time series. In Fig 2, we see outputs from BNPR and BNPR-PS on the same example iteration.

Our first set of experiments is aimed at determining the extent of the bias introduced by unaccounted preferential sampling. With r Monte Carlo iterations, we take two approaches to locating model misspecification error—time interval analysis and pointwise analysis. For time interval analyses, we calculate summary statistics for a pre-specified time interval (a, b) and average them over the set of r simulation iterations. For pointwise analyses however, we
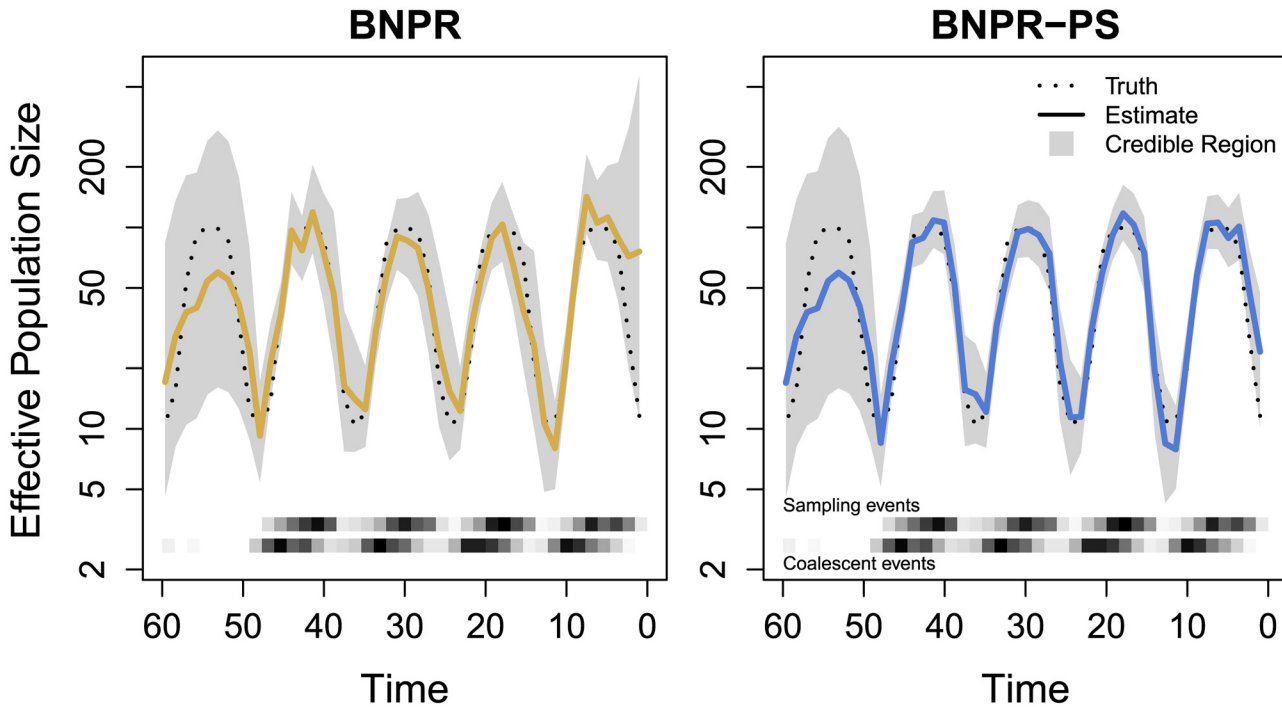
Fig 2. Graphical representation of the output of a single genealogy simulation and integrated nested Laplace approximation (INLA) estimation.
The dotted black lines represent the true population trajectory. The solid colored lines represent the posterior median estimates, while the shaded regions represent the 95% credible regions. At bottom, the upper and lower heatmaps represent frequencies of sampling events and coalescent events, respectively. For this figure, we sampled individuals according to an inhomogeneous Poisson process with intensity proportional to effective population size $N_e(t)$. The plot on the left is generated by Bayesian nonparametric phylodynamic reconstruction (BNPR) and does not account for preferential sampling, while the plot on the right is generated by Bayesian nonparametric phylodynamic reconstruction with preferential sampling (BNPR-PS) and incorporates our sampling time model. Time is in months.

consider the time series of point estimates from each iteration, and then on a pointwise basis we calculate aggregate point estimates and confidence intervals.

Our time interval summary statistics are *mean relative deviation*,

$$\text{MRD} = \frac{1}{r}\sum_{i=1}^{r}\left[\frac{1}{b-a}\int_{a}^{b}\frac{|\hat{N}_i^{\gamma}(t) - N^{\gamma}(t)|}{N^{\gamma}(t)}\,dt\right],$$

mean relative width of the 95% Bayesian credible intervals,

$$\text{MRW} = \frac{1}{r}\sum_{i=1}^{r}\left[\frac{1}{b-a}\int_{a}^{b}\frac{\hat{N}_{i,0.975}^{\gamma}(t) - \hat{N}_{i,0.025}^{\gamma}(t)}{N^{\gamma}(t)}\,dt\right],$$

where $N^{\gamma}(t)$ is the discretized true effective population size trajectory, $\hat{N}_i^{\gamma}(t)$ is the estimated posterior median of effective population sizes for iteration i, and $\hat{N}_{i,q}^{\gamma}(t)$ is the estimated qth posterior quantile for iteration i. We also look at *mean envelope*, ME, the proportion of grid points where the credible interval contains the true trajectory, averaged over all grid points contained in $[a, b]$ across all Monte Carlo iterations.

For a given grid of time points $\{t_j\}_{j=0}^k$, pointwise analysis computes the means of pointwise posterior medians,

$$\mathrm{mpmedian}(t_j) = \frac{1}{r} \sum_{i=1}^r \hat{N}_{i,0.5}^\gamma(t_j), \ \text{ for } j = 0, \ldots, k,$$

pointwise mean relative errors,

$$\mathrm{mre}(t_j) = \frac{1}{r} \sum_{i=1}^r \frac{\hat{N}_{i,0.5}^\gamma(t_j) - N^\gamma(t_j)}{N^\gamma(t_j)}, \ \text{ for } j = 0, \ldots, k,$$

and a sequence of mean relative widths of the pointwise Bayesian credible intervals,

$$\mathrm{mrw}(t_j) = \frac{1}{r} \sum_{i=1}^r \frac{\hat{N}_{i,0.975}^\gamma(t_j) - \hat{N}_{i,0.025}^\gamma(t_j)}{N^\gamma(t_j)}, \ \text{ for } j = 0, \ldots, k.$$

We choose grid size $k = 100$, number of simulation iterations $r = 512$, and expected number of lineages per genealogy $n = 500$. We choose the sampling interval $[0, 48]$ for all simulations.

Ignoring preferential sampling. Table 1 shows the averaged time interval summary statistics for simulated genealogies under uniform and proportional schedules for the time intervals $(0, 6)$ and $(6, 48)$. Genealogies were simulated assuming effective population size function $N_{e,2,0}(t)$ defined in Eq 2. We show the time interval summary statistics for inferred effective population sizes both ignoring and considering preferential sampling. Ignoring preferential sampling (Table 1 under BNPR), we note a 17% increase in mean relative deviation from uniform to proportional schedules, as well as a 20% increase in mean relative width of Bayesian credible intervals for $(6, 48)$. For $(0, 6)$ the increase is more stark. We see a 407% increase in mean relative deviation from uniform to proportional, and a 799% increase in mean relative width of Bayesian credible intervals. Under proportional sampling, we see a notable increase in mean envelope, ME, on the $(0, 6)$ interval. All other cases show BNPR and BNPR-PS having ME within Monte Carlo error. These results confirm that ignoring preferential sampling affects both bias and variance of Bayesian nonparametric estimators of the effective populations size.

Fig 3 (solid lines) compares the average pointwise statistics for the uniform and proportional sampling schedules. Note the marked increase in mean relative error in several locations. We also see much larger mean relative widths in the same locations. Fig 4 compares the time interval statistics for the uniform and proportional sampling schedules, and we see increases in mean relative deviation and mean relative width. We conjecture that these features are representative of the model misspecification error that we would expect while sampling sequences/lineages preferentially in time but not accounting for it in the model.

Accounting for preferential sampling. Table 1 under BNPR-PS shows the time interval statistics for the sampling-aware model. For interval $(6, 48)$, mean relative deviation decreases by 23% versus BNPR under proportional sampling, while mean relative width of Bayesian credible intervals decreases by a larger margin of 33%. For interval $(0, 6)$ mean relative deviation and mean relative width decrease by 80% and 91%, respectively. Under uniform sampling, BNPR-PS performs almost identically to BNPR for both intervals.

Fig 3 (dashed lines) compares the average pointwise statistics for the uniform and proportional sampling schedules under BNPR-PS. We see that BNPR-PS does not experience the increase in relative error that BNPR experiences under preferential sampling. The plots also show an improvement in mean relative width of Bayesian credible intervals under preferential sampling due to the additional information available. Fig 4 compares the time interval statistics

Table 1. Averaged time interval summary statistics for BNPR and BNPR-PS.

	Uniform—(6, 48)		Proportional—(6, 48)		Uniform—(0, 6)		Proportional—(0, 6)	
	BNPR	BNPR-PS	BNPR	BNPR-PS	BNPR	BNPR-PS	BNPR	BNPR-PS
MRD	0.205	0.205	0.239	0.183	0.430	0.436	2.181	0.432
MRW	1.255	1.255	1.500	1.008	2.816	2.816	19.681	1.682
ME	0.965	0.964	0.962	0.957	0.950	0.948	0.833	0.898

We compare the performance of the models under two different sampling distributions. Uniform distributes sampling times according to a uniform distribution on the interval (0, 48), while proportional distributes sampling times according to a inhomogeneous Poisson process with intensity proportional to effective population size $N_e(t)$ on the same interval. We examine the statistics mean relative deviation (MRD), mean relative width of the 95% Bayesian credible interval (MRW), and mean envelope (ME). We average over statistics over the interval (6, 48) where both methods perform well and over the most recent interval (0, 6) where BNPR-PS performs considerably better.

doi:10.1371/journal.pcbi.1004789.t001

for the uniform and proportional sampling schedules under BNPR-PS, and shows improvements in mean relative deviation and mean relative width.

Negative control simulations.   In the previous sections, we find a pattern of increased mean relative deviation and mean relative width while using a conditional model in a scenario involving preferential sampling. However, it is possible that this behavior of the conditional, state-of-the-art coalescent model can be seen under other simulation scenarios that cluster sampling times, even when such clustering has no relationship to the effective population size fluctuations. To test this assertion, we design a pair of negative control simulation studies to have random clusters of sampling times, but no preferential sampling.

First, we apply BNPR to genealogies generated from randomly constructed piecewise constant sampling intensity functions, independent of effective population size; see Appendix. We see some examples of increased mean relative error, but nothing as consistent nor prevalent as in the preferential sampling case above (see Figs [2] and [3]). Similarly, we see increased mean relative width in several locations, but decreased widths in others. Second, we apply BNPR to genealogies generated from Gaussian process evaluations (subsampled for relatively similar shapes and number of peaks and troughs to the true population trajectory); see Appendix. This model has shape characteristics closer to the population trajectory since we are sampling from a Gaussian process. Despite the similar shapes, we see fewer increases in mean relative width and smaller increases in mean relative width. We conclude that unaccounted preferential sampling produces markedly more error more consistently than the negative control cases.

We also apply BNPR-PS to the same scenarios as above. BNPR-PS's performance suffers significantly due to both scenarios violating its fundamental assumption of a fixed relationship between effective population size and sampling intensity. We see BNPR-PS performs worst locally when there is a nearly fixed relationship which is suddenly reversed in a small time interval.

Parametric simulations.   Finally, we also explored model misspecification in a correctly-specified parametric context. We simulated 100,000 genealogies from the coalescent with an exponential effective population size trajectory $N_e(t) = \exp(a+bt)$, under uniform and proportional sampling schedules. We applied an exponential growth/decline parametric maximum likelihood method and summarized the results in the Appendix. In both uniform and preferential sampling we see small, but comparable biases in estimates of parameter a. However, estimates of the exponential growth rate parameter b have no detectable bias under uniform sampling, but have small but significant bias under preferential sampling. This verifies that ignoring preferential sampling causes systematic bias, perhaps of small magnitude, in
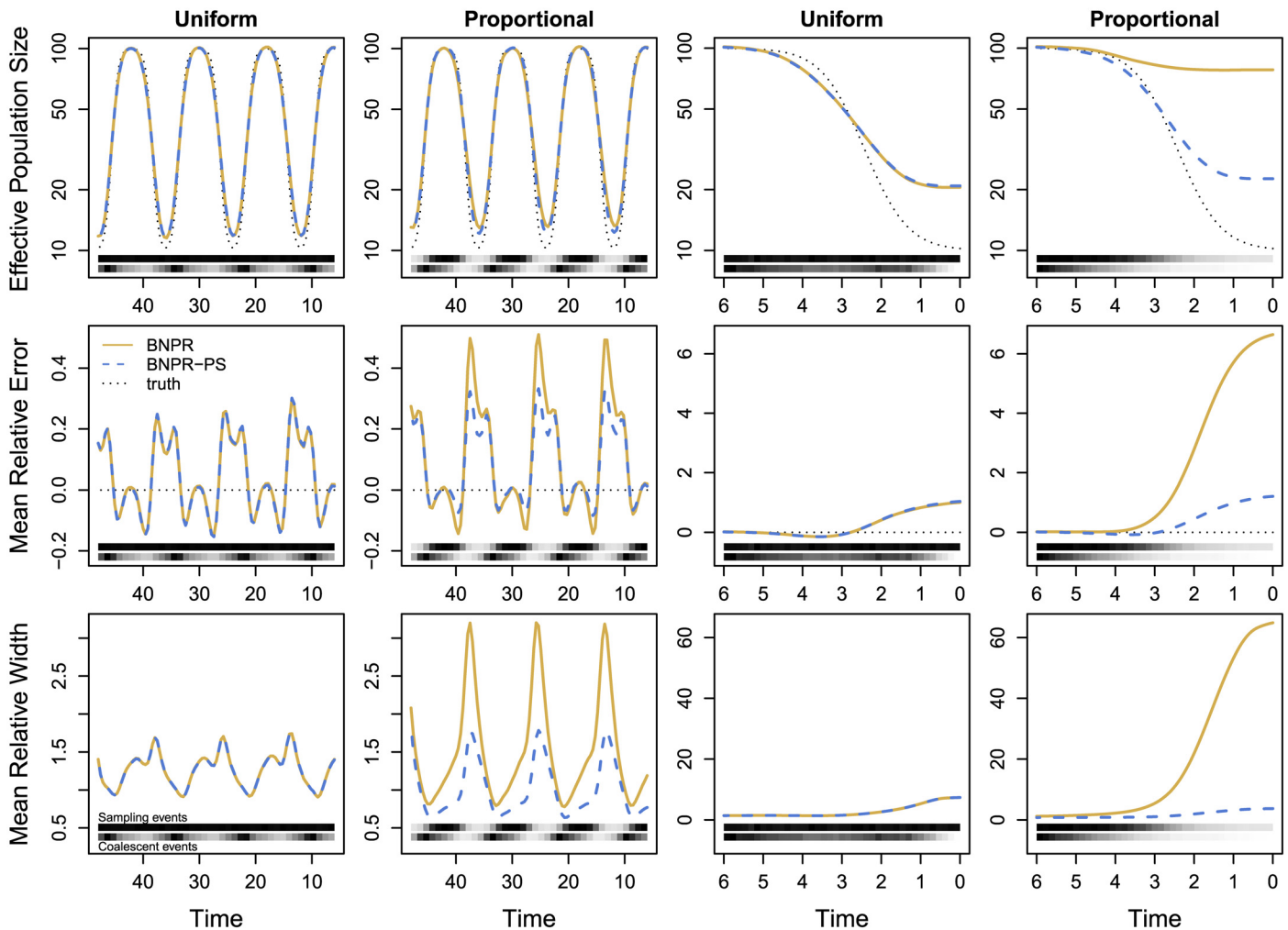
Fig 3. Comparison of pointwise statistics. Dotted black lines represent the truth, where applicable. Solid yellow lines represent the conditional method BNPR (ignoring preferential sampling), while dashed blue lines represent the sampling-aware method BNPR-PS (accounting for preferential sampling). The first row shows true and estimated effective population sizes, the second shows mean relative error, while the third shows mean relative width of the 95% Bayesian credible interval. The left two columns show the interval (6, 48) where both models perform at their best. The right two columns show (0, 6), where BNPR-PS performs significantly better. At the bottom of each plot, the distribution of sampling events (above) and coalescent events (below) are shown as heat maps. Time is in months.

doi:10.1371/journal.pcbi.1004789.g003

maximum likelihood phylodynamic estimation even under a simple low-dimensional parametric model.

Case studies

New York influenza. We base our first case study on a subset of the data from [19], also analyzed by Palacios and Minin [5]. We focus on the 709 hemagglutinin gene sequences of H3N2 human influenza type A obtained from the National Center for Biotechnology Information (NCBI) Influenza Virus Sequence Database for years 1992 through 2005 from New York State. We align the sequences using the software MUSCLE [23], and infer a maximum clade credibility genealogy using the software BEAST [24]. We infer the genealogy branch lengths in units of years using a strict molecular clock, a constant effective population size prior, and an HKY substitution model with the first two nucleotides of a codon sharing the same estimated
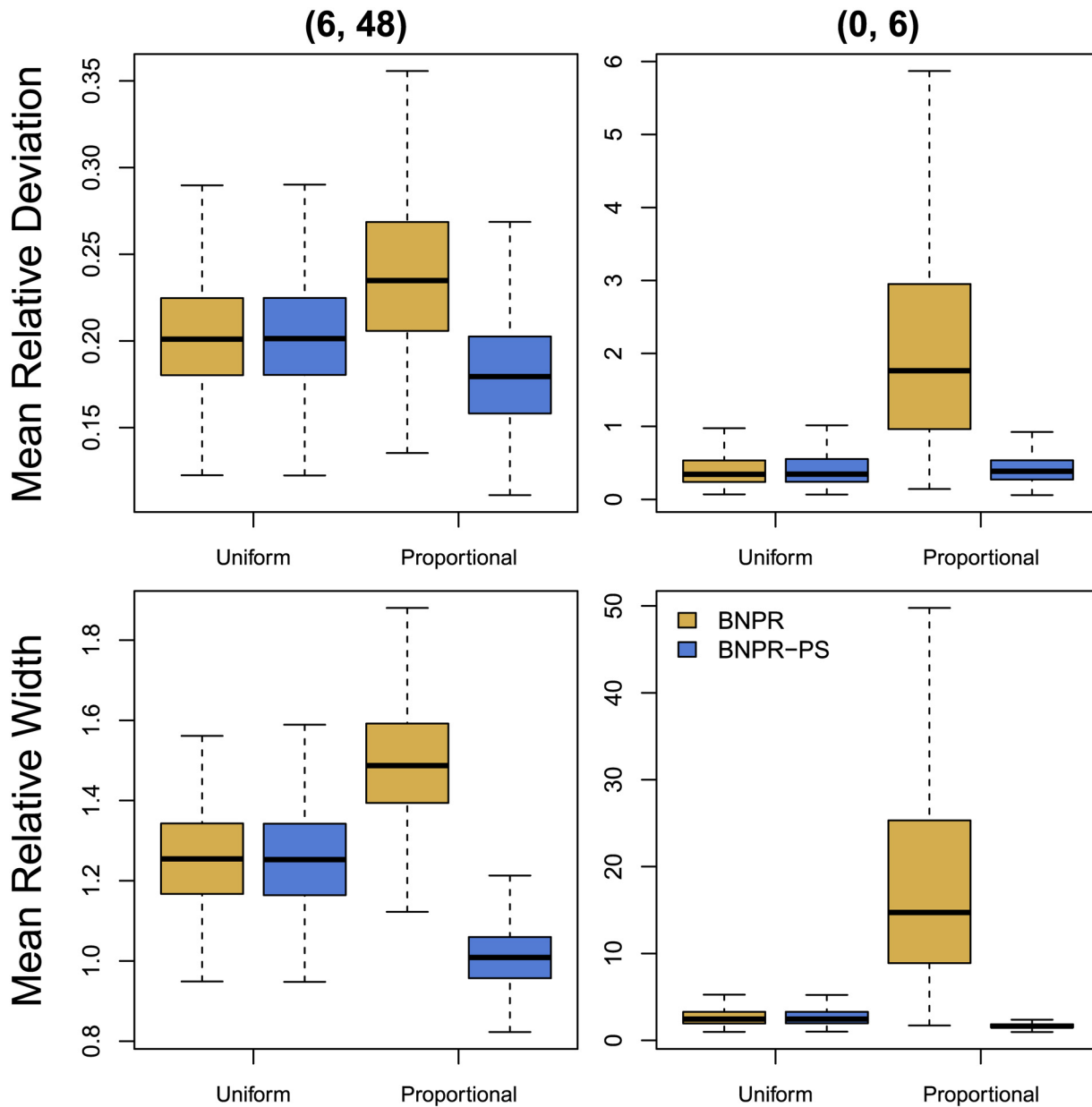
Fig 4. Comparison of time interval statistics. Within each plot, we apply BNPR and BNPR-PS to sampling times generated according to a Uniform distribution on the left and proportionally to effective population size on the right. In the left column of plots, we examine the interval (6, 48) where the performances of both models are comparable. In the right column, we show (0, 6), and note that BNPR-PS performs well, while BNPR performs considerably worse.

transition matrix, while the third nucleotide's transition matrix is estimated separately. We then apply our two algorithms to the estimated genealogy.

We find that BNPR produces results in line with previous analyses of this dataset, showing a characteristic uncertainty around the flu seasons of 2000–2001 and 2002–2003 (see Fig 5). In contrast, BNPR-PS shows a marked improvement in the regularity of the reconstructed flu seasons, as well as thinner Bayesian credible intervals across the the whole observation interval.
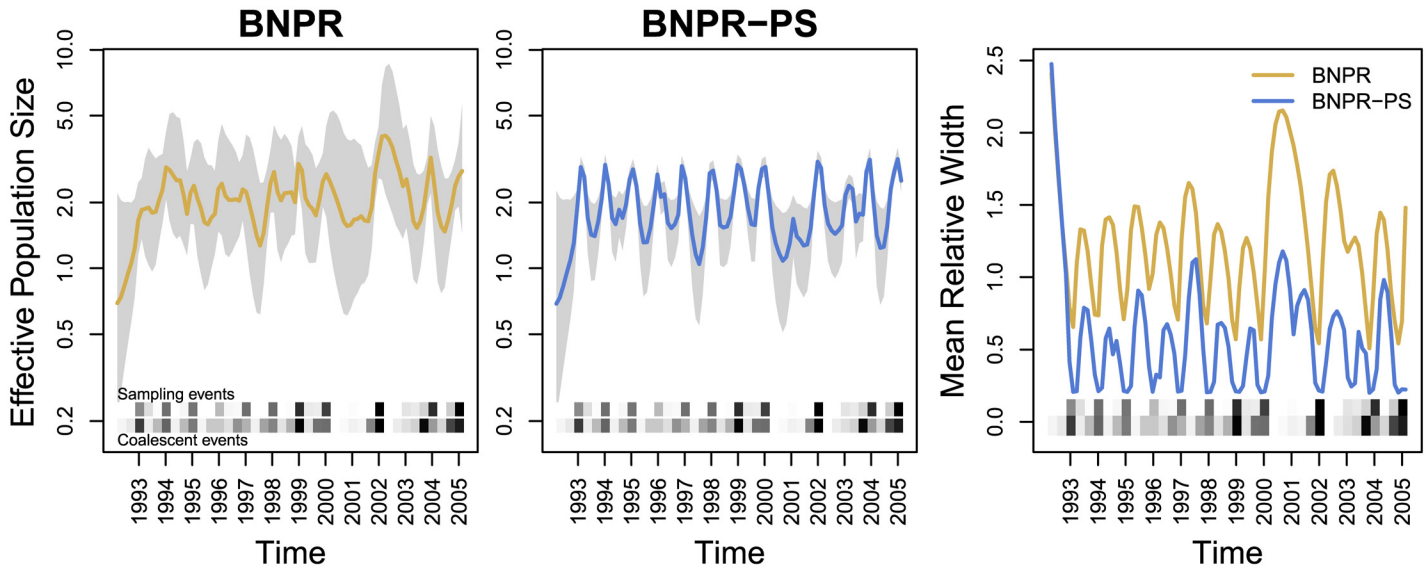
Fig 5. BNPR and BNPR-PS models applied to the genealogy inferred from the New York influenza data [19]. Years mark January of the corresponding year. Note the correlation of higher effective population size $N^\gamma(t)$ with more intense sampling frequencies (darker regions in the Sampling events heatmap), suggesting preferential sampling. We see a marked improvement in discerning the seasonal influenza patterns and significantly thinner credible regions under BNPR-PS.

doi:10.1371/journal.pcbi.1004789.g005

Estimations also improved during the unusual flu seasons of 2000–2001 and 2002–2003, consistent with these seasons being H1N1 dominant seasons instead of H3N2 dominant [25].

To compare performance of the BNPR and BNPR-PS models, we introduce an empirical measure of performance because we cannot know the true population size trajectory. We calculate the time interval and pointwise *empirical mean relative width* (EMRW) of the 95% Bayesian credible intervals,

$$\text{EMRW} = \frac{1}{r}\sum_{i=1}^{r}\left[\frac{1}{b-a}\int_a^b \frac{\hat{N}_{i,0.975}^\gamma(t) - \hat{N}_{i,0.025}^\gamma(t)}{\widehat{N}_i^\gamma(t)}\, dt\right], \text{ for } j = 0, \dots, k,$$

and

$$emrw(t_j) = \frac{1}{r}\sum_{i=1}^{r}\frac{\hat{N}_{i,0.975}^\gamma(t_j) - \hat{N}_{i,0.025}^\gamma(t_j)}{\widehat{N}_i^\gamma(t_j)}, \text{ for } j = 0, \dots, k.$$

Table 2 shows a very high value of β_1 for this dataset, suggesting a strong pattern of preferential sampling, and accordingly we see a marked improvement of BNPR-PS model over its BNPR counterpart in estimation precision as measured by the Bayesian credible interval widths (EMRW).

Regional influenza. Zinder et al. examine world-wide seasonal patterns of migration of H3N2 influenza across the regions of the world [20]. They also examine different seasonal incidence patterns, with tropical regions having a relatively flat incident rate throughout the year, while temperate regions show larger seasonal variation with higher incidence in winter months. In order to explore the effects of seasonality on preferential sampling, we examine the regions separately. We align the sequences using the software MUSCLE [23], and infer a maximum clade credibility genealogy using the software BEAST [24]. We infer the genealogy branch lengths in units of years using a strict molecular clock, a constant effective population size prior, and an HKY substitution model with the first two nucleotides of a codon sharing the

Table 2. Case studies' empirical mean relative widths and Bayesian credible intervals of β_0 and β_1.

	n	EMRW		95% credible interval of β_0	95% credible interval of β_1
		BNPR	BNPR-PS		
New York influenza	709	1.23	0.58	(-47.4, -30.3)	(5.88, 10.23)
Regional influenza					
USA & Canada	520	1.83	1.11	(-3.02, -0.79)	(2.52, 4.05)
South America	191	0.86	0.91	(-4.21, -0.42)	(3.27, 7.52)
Europe	361	1.73	0.96	(-6.61, -2.44)	(3.68, 6.88)
India	233	1.79	1.30	(-2.18, 0.50)	(2.34, 4.78)
Japan & Korea	444	1.82	1.09	(-2.23, -0.25)	(2.35, 3.76)
North China	384	1.80	1.09	(-2.63, -0.27)	(2.22, 3.89)
South China	528	1.27	0.78	(-1.05, 1.00)	(1.68, 3.23)
Southeast Asia	494	0.99	0.54	(-7.93, -2.55)	(4.39, 8.86)
Oceania	461	1.53	0.88	(-1.51, 0.43)	(2.71, 4.52)

The regional influenza dataset is broken down into world regions. In all but one region, we see improvements, or at worst near-equality, in empirical mean relative width (EMRW) using BNPR-PS over BNPR.

doi:10.1371/journal.pcbi.1004789.t002

same estimated transition matrix, while the third nucleotide's transition matrix is estimated separately. We then apply our two algorithms to the estimated genealogy.

We find that none of the regions contain 0 in their β_1 Bayesian credible interval (see Table 2), suggesting a relationship between effective population size and sampling frequency. Across all regions except South America, we see improvements of the BNPR-PS model over the BNPR model in estimation precision (EMRW). We examine three of the regions more closely in Figs 6 and 7 and the remaining six regions in the appendix. We see noticeable improvements in the relative widths of the Bayesian credible intervals. We also see more pronounced seasonality in the estimated effective population size trajectories produced by BNPR-PS. The USA/Canada region shows the expected seasonal peak in January-February, while the Oceania region shows the same in July-September. South China shows less seasonality overall, but BNPR-PS shows a more pronounced August peak despite the region being in the northern hemisphere. This is, however, in line with previous findings, most likely due to southern China's more tropical climate [26].

Discussion

Researchers who study measurably evolving populations [27], such as viruses, can inadvertently or purposefully preferentially select sequences in accordance to the changes in size of the population of interest. Failing to account for such an ascertainment bias can compromise the statistical properties of phylodynamic inference. Our simulation study shows that the effect of preferential sampling is particularly severe when the effective population size is decreasing. We propose an extension to the state-of-the-art in Gaussian process-based Bayesian phylodynamic methods, in which we assume that sampling times *a priori* follow an inhomogeneous Poisson process with intensity proportional to a power of the effective population size. This model extension eliminates the systematic estimation bias resulting from having unrecognized preferential sampling, and also gives us better population size estimates by incorporating sampling times as an additional source of information.

Applied to the real-world examples, our method produces improvements over the state-of-the-art. We see significantly improved precision, as well as more realistic estimation of seasonal
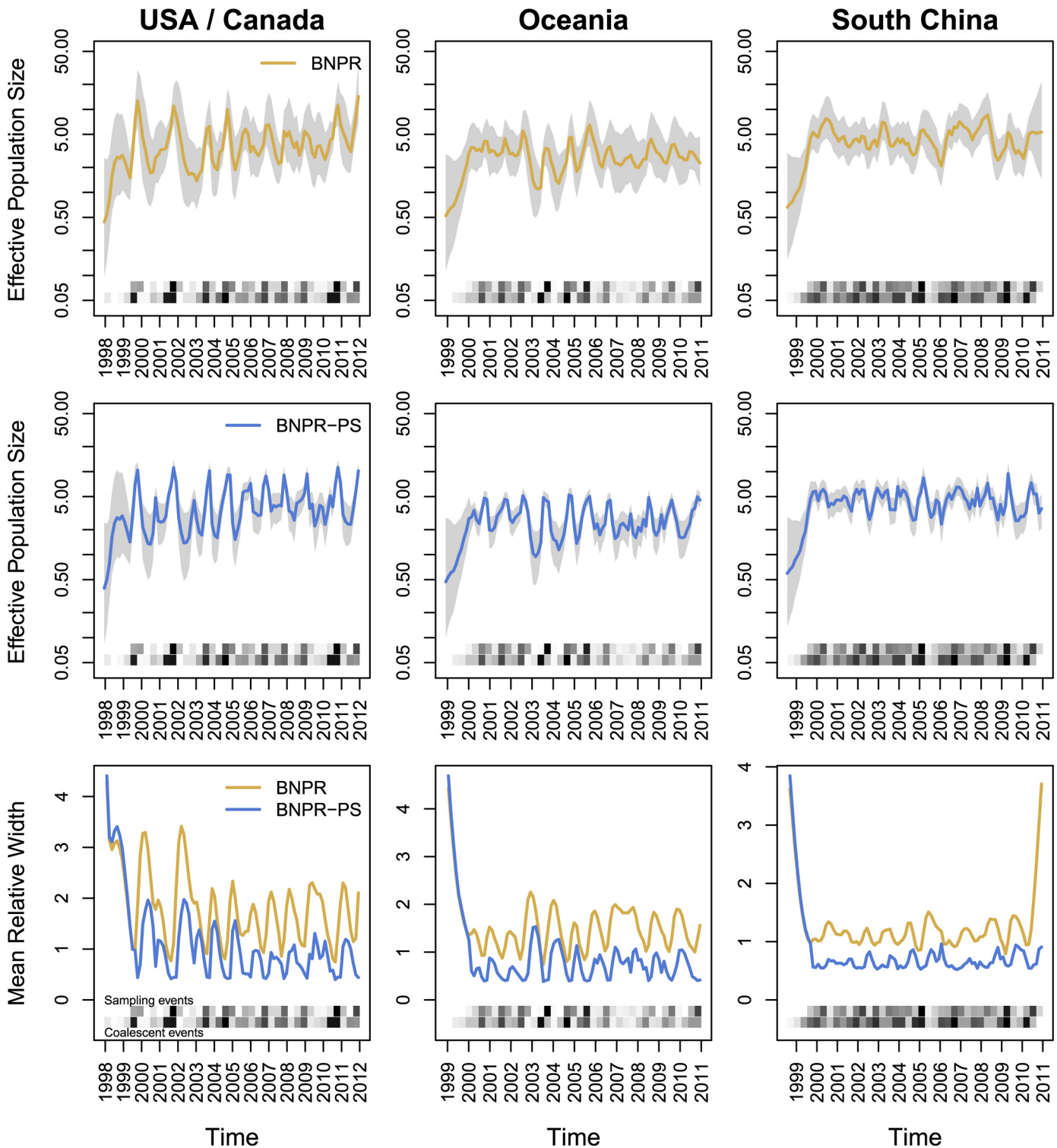
Fig 6. BNPR and BNPR-PS models applied to the genealogies inferred from the regional influenza example [20]. We see moderate correlation between effective population size $N^\gamma(t)$ and sampling frequencies in the data (Table 2). We see improvements in Bayesian credible interval widths, and BNPR-PS performs as well or better than BNPR everywhere in these examples.

doi:10.1371/journal.pcbi.1004789.g006
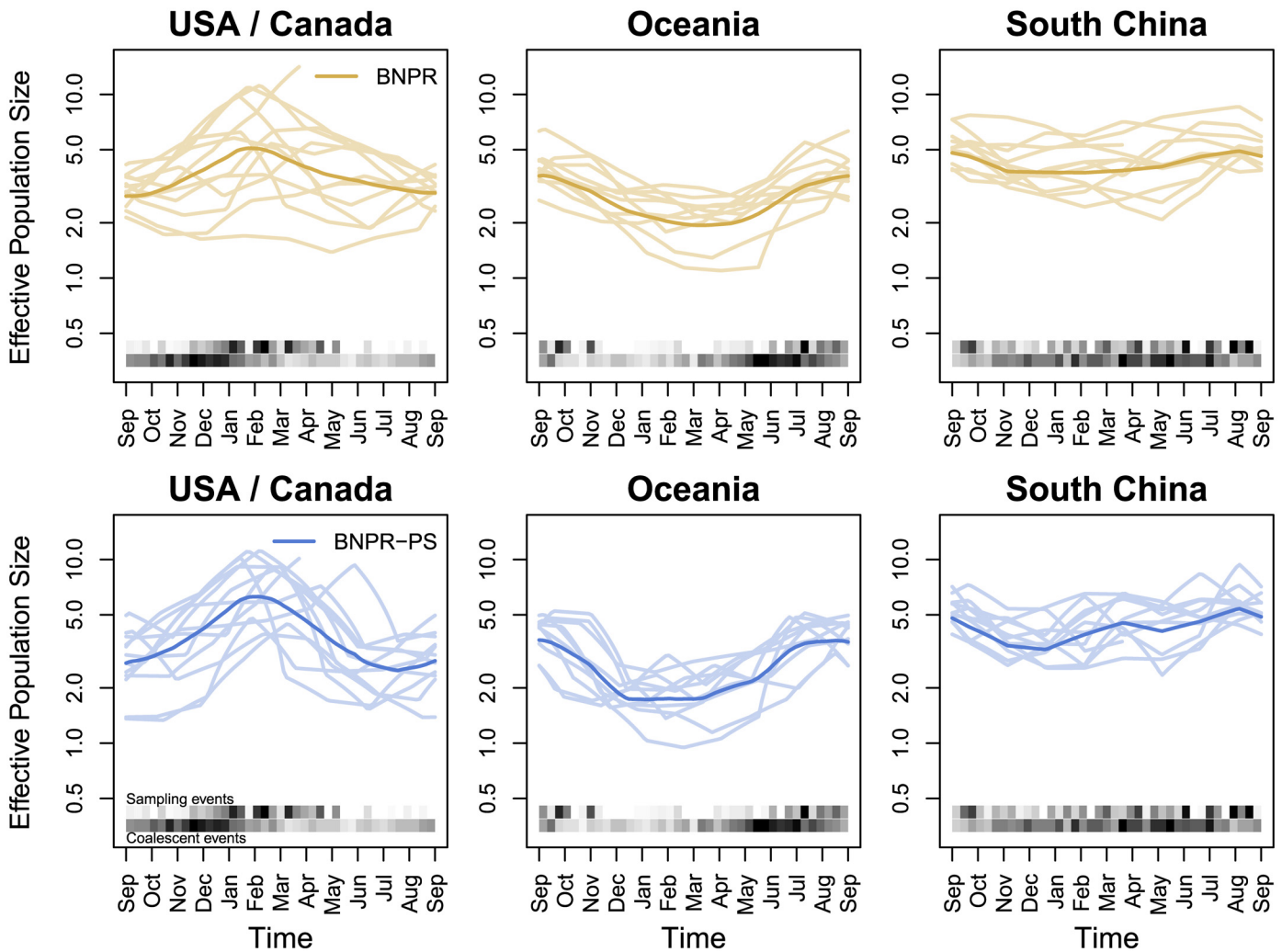
Fig 7. Seasonality in regional influenza. BNPR and BNPR-PS models applied to the genealogies inferred from the regional influenza example with years overlaid. We see more pronounced seasonality in the BNPR-PS plots.

variation of influenza diversity. In the presence of weaker preferential sampling, as in some of the regional influenza examples, we note that our method still performs better than the current state-of-the-art, with no loss of performance aside from a slightly longer computation time. In addition, by estimating β_1, the effect of population size on the log-intensity of sampling times, we gain the ability to quantify the strength of the preferential sampling relationship in the different regions. Such quantification is scientifically useful in infectious disease phylodynamics, because researchers may want to know whether frequency of sampling times can be used as a proxy for incidence.

One avenue of future exploration is to intentionally guarantee preferential sampling during the sequence data collection phase. For example, if an epidemiological study contains noisy incidence data, we can subsample sequences with intensity proportional to incidence and apply our sampling-aware BNPR-PS model to the resulting sequence data. Such a procedure will indirectly combine sequence and incidence data to estimate the effective number of infections —a nontrivial task for the current methods [28]. We contrast this to the approach of [29], which examined the effect of sampling infectious disease agent sequences in batches at different

points in an epidemic's life-cycle compared to uniform and preferential sampling. They found that during epidemic declines their estimates had the largest mean squared error and benefited most in terms of this metric when samples were collected more frequently during the declines. This is consistent with our results, as we see the most error and widest credible intervals during effective population size declines. However, they did not consider the effect of the relationship between their proposed sampling intensity and population size trajectories on estimation of population dynamics—the primary goal of our work.

Our current implementation of the BNPR-PS model assumes a fixed, known genealogy. However, in practice, genealogies are inferred with inherent uncertainty from sequence data. We have found that point estimates produced by our method on the Regional influenza data are robust to genealogical uncertainty (see Regional influenza section in the Appendix), but a method that jointly estimates both genealogy and effective population is still necessary to properly assign uncertainty to population size estimates. One limitation of our method is that the INLA framework cannot be extended to include inference of genealogies. However, it should be straightforward to incorporate the core of our approach —the sampling times model— into an MCMC sampler that targets the joint posterior distribution of population size trajectory, genealogy of sampled sequences, and other parameters. We intend to implement such an MCMC approach in the software BEAST [24].

The main goal of this manuscript is to point out the danger of ignoring preferential sampling in phylodynamics. Providing a solution to this problem, in the form of BNPR-PS model, remains our secondary goal, but we emphasize that much work is still needed to refine our proposed approach. The main weakness of our new model lies in its rigid parametric form of dependence between effective population size $N_e(t)$ and sequence sampling intensity $\lambda(t)$. In our negative control simulations we see that BNPR-PS performance suffers, possibly greatly, when this assumption of a fixed relationship between effective population size $N_e(t)$ and sampling intensity $\lambda(t)$ is violated. Similar results under model misspecification are observed by Volz and Frost in the context of birth-death-sampling models for phylodynamic inference [16, 30].

Sampling times model misspecification is most likely to occur if other variables besides effective population size $N_e(t)$ effect changes in the sampling intensity $\lambda(t)$. For instance, not accounting for a lag between $N_e(t)$ and $\lambda(t)$ may cause a severe model misspecification. Similarly, not accounting for increases in sampling intensity on longer time scales due to decreases in the cost of sequencing will bias our BNPR-PS estimation. We plan to address these issues by modeling our sampling intensity $\lambda(t)$ as a log-linear combination of effective sample size and other covariates:

$$\log[\lambda(t)] = \beta^T \mathbf{c}(t),$$

where $\mathbf{c}(t)^T = (1, N_e(t), c_1(t), \ldots, c_p(t))$ and $c_i(t)$, $i = 1, \ldots, p$ are covariates of interest. For example, the cost of genome sequencing over time and lagged population size $N_e(t−l)$ are among prime candidates for covariates to be included into our BNPR-PS model. Another example of a promising time-varying sampling covariate is an indicator of 'outbreak' status, allowing for changes in sampling intensity during times of increased epidemiological oversight. We hope to explore these model extensions in our future research.

Supporting Information

S1 Appendix. Additional materials.
(PDF)

Author Contributions

Conceived and designed the experiments: MDK JAP TB MAS VNM. Performed the experiments: MDK JAP. Analyzed the data: MDK JAP. Contributed reagents/materials/analysis tools: MDK JAP TB MAS VNM. Wrote the paper: MDK JAP TB MAS VNM.

References

1. Grenfell BT, Pybus OG, Gog JR, Wood JLN, Daly JM, Mumford JA, et al. Unifying the epidemiological and evolutionary dynamics of pathogens. Science. 2004; 303(5656):327–332. doi: 10.1126/science.1090727 PMID: 14726583

2. Holmes EC, Grenfell BT. Discovering the phylodynamics of RNA viruses. PLoS Computational Biology. 2009; 5(10):e1000505. doi: 10.1371/journal.pcbi.1000505 PMID: 19855824

3. Minin VN, Bloomquist EW, Suchard MA. Smooth skyride through a rough skyline: Bayesian coalescent-based inference of population dynamics. Molecular Biology and Evolution. 2008; 25(7):1459–1471. doi: 10.1093/molbev/msn090 PMID: 18408232

4. 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. 2013; 30(3):713–724. doi: 10.1093/molbev/mss265 PMID: 23180580

5. Palacios JA, Minin VN. Gaussian Process-Based Bayesian Nonparametric Inference of Population Size Trajectories from Gene Genealogies. Biometrics. 2013; 69(1):8–18. doi: 10.1111/biom.12003 PMID: 23409705

6. Kingman JFC. The coalescent. Stochastic processes and their applications. 1982; 13(3):235–248. doi: 10.1016/0304-4149(82)90011-4

7. Wakeley J, Sargsyan O. Extensions of the coalescent effective population size. Genetics. 2009; 181(1):341–345. doi: 10.1534/genetics.108.092460 PMID: 19001293

8. Griffiths RC, Tavaré S. Sampling theory for neutral alleles in a varying environment. Philosophical Transactions of the Royal Society of London Series B: Biological Sciences. 1994; 344(1310):403–410. doi: 10.1098/rstb.1994.0079 PMID: 7800710

9. Drummond AJ, Nicholls GK, Rodrigo AG, Solomon W. Estimating mutation parameters, population history and genealogy simultaneously from temporally spaced sequence data. Genetics. 2002; 161(3):1307–1320. PMID: 12136032

10. Kuhner MK, Yamato J, Felsenstein J. Maximum likelihood estimation of population growth rates based on the coalescent. Genetics. 1998; 149(1):429–434. PMID: 9584114

11. Pybus OG, Rambaut A, Harvey PH. An integrated framework for the inference of viral population history from reconstructed genealogies. Genetics. 2000; 155(3):1429–1437. PMID: 10880500

12. Drummond AJ, Rambaut A, Shapiro B, Pybus OG. Bayesian coalescent inference of past population dynamics from molecular sequences. Molecular Biology and Evolution. 2005; 22(5):1185–1192. PMID: 15703244

13. Ho SYW, Shapiro B. Skyline-plot methods for estimating demographic history from nucleotide sequences. Molecular Ecology Resources. 2011; 11(3):423–434. doi: 10.1111/j.1755-0998.2011.02988.x PMID: 21481200

14. Frost SDW, Volz EM. Viral phylodynamics and the search for an 'effective number of infections'. Philosophical Transactions of the Royal Society B: Biological Sciences. 2010; 365(1548):1879–1890. doi: 10.1098/rstb.2010.0060

15. Diggle PJ, Menezes R, Su T. Geostatistical inference under preferential sampling. Journal of the Royal Statistical Society: Series C (Applied Statistics). 2010; 59(2):191–232. doi: 10.1111/j.1467-9876.2009.00701.x

16. Volz EM, Frost SDW. Sampling through time and phylodynamic inference with coalescent and birth–death models. Journal of The Royal Society Interface. 2014; 11(101):20140945. doi: 10.1098/rsif.2014.0945

17. Rue H, Martino S, Chopin N. Approximate Bayesian inference for latent Gaussian models by using integrated nested Laplace approximations. Journal of the Royal Statistical Society: Series B. 2009; 71(2):319–392. doi: 10.1111/j.1467-9868.2008.00700.x

18. Palacios JA, Minin VN. Integrated Nested Laplace Approximation for Bayesian Nonparametric Phylodynamics. In: Proceedings of the Twenty-Eighth International Conference on Uncertainty in Artificial Intelligence; 2012. p. 726–735.

19. Rambaut A, Pybus OG, Nelson MI, Viboud C, Taubenberger JK, Holmes EC. The genomic and epidemiological dynamics of human influenza A virus. Nature. 2008; 453(7195):615–619. doi: 10.1038/nature06945 PMID: 18418375

20. Zinder D, Bedford T, Baskerville EB, Woods RJ, Roy M, Pascual M. Seasonality in the migration and establishment of H3N2 Influenza lineages with epidemic growth and decline. BMC Evolutionary Biology. 2014; 14(1):272. doi: 10.1186/s12862-014-0272-2 PMID: 25539729

21. Felsenstein J, Rodrigo AG. Coalescent Approaches to HIV Population Genetics. In: The Evolution of HIV. Johns Hopkins University Press; 1999. p. 233–272.

22. Martins TG, Simpson D, Lindgren F, Rue H. Bayesian computing with INLA: new features. Computational Statistics & Data Analysis. 2013; 67:68–83. doi: 10.1016/j.csda.2013.04.014

23. Edgar RC. MUSCLE: Multiple sequence alignment with high accuracy and high throughput. Nucleic Acids Research. 2004; 32:1792–1797. PMID: 15034147

24. Drummond AJ, Suchard MA, Xie D, Rambaut A. Bayesian phylogenetics with BEAUti and the BEAST 1.7. Molecular Biology and Evolution. 2012; 29:1969–1973. doi: 10.1093/molbev/mss075 PMID: 22367748

25. Goldstein E, Cobey S, Takahashi S, Miller JC, Lipsitch M. Predicting the epidemic sizes of influenza A/ H1N1, A/H3N2, and B: a statistical method. PLoS medicine. 2011; 8(7):952. doi: 10.1371/journal. pmed.1001051

26. Shu YL, Fang LQ, de Vlas SJ, Gao Y, Richardus JH, Cao WC. Dual seasonal patterns for influenza, China. Emerging Infectious Diseases. 2010; 16(4):725. doi: 10.3201/eid1604.091578 PMID: 20350403

27. Drummond AJ, Pybus OG, Rambaut A, Forsberg R, Rodrigo AG. Measurably evolving populations. Trends in Ecology & Evolution. 2003; 18(9):481–488. doi: 10.1016/S0169-5347(03)00216-7

28. Rasmussen DA, Ratmann O, Koelle K. Inference for nonlinear epidemiological models using genealogies and time series. PLoS Computational Biology. 2011; 7:e1002136. doi: 10.1371/journal.pcbi. 1002136 PMID: 21901082

29. Stack JC, Welch JD, Ferrari MJ, Shapiro BU, Grenfell BT. Protocols for sampling viral sequences to study epidemic dynamics. Journal of The Royal Society Interface. 2010; 7(48):1119–1127. doi: 10. 1098/rsif.2009.0530

30. Stadler T. Sampling-through-time in birth–death trees. Journal of Theoretical Biology. 2010; 267 (3):396–404. doi: 10.1016/j.jtbi.2010.09.010 PMID: 20851708