

# Phylodynamic Inference for Structured Epidemiological Models



David A. Rasmussen<sup>1\*</sup>, Erik M. Volz<sup>2</sup>, Katia Koelle<sup>1,3</sup>

**1** Biology Department, Duke University, Durham, North Carolina, United States of America, **2** Department of Infectious Disease Epidemiology, Imperial College London, London, United Kingdom, **3** Fogarty International Center, National Institutes of Health, Bethesda, Maryland, United States of America

## Abstract

Coalescent theory is routinely used to estimate past population dynamics and demographic parameters from genealogies. While early work in coalescent theory only considered simple demographic models, advances in theory have allowed for increasingly complex demographic scenarios to be considered. The success of this approach has led to coalescent-based inference methods being applied to populations with rapidly changing population dynamics, including pathogens like RNA viruses. However, fitting epidemiological models to genealogies via coalescent models remains a challenging task, because pathogen populations often exhibit complex, nonlinear dynamics and are structured by multiple factors. Moreover, it often becomes necessary to consider stochastic variation in population dynamics when fitting such complex models to real data. Using recently developed structured coalescent models that accommodate complex population dynamics and population structure, we develop a statistical framework for fitting stochastic epidemiological models to genealogies. By combining particle filtering methods with Bayesian Markov chain Monte Carlo methods, we are able to fit a wide class of stochastic, nonlinear epidemiological models with different forms of population structure to genealogies. We demonstrate our framework using two structured epidemiological models: a model with disease progression between multiple stages of infection and a two-population model reflecting spatial structure. We apply the multi-stage model to HIV genealogies and show that the proposed method can be used to estimate the stage-specific transmission rates and prevalence of HIV. Finally, using the two-population model we explore how much information about population structure is contained in genealogies and what sample sizes are necessary to reliably infer parameters like migration rates.

**Citation:** Rasmussen DA, Volz EM, Koelle K (2014) Phylodynamic Inference for Structured Epidemiological Models. *PLoS Comput Biol* 10(4): e1003570. doi:10.1371/journal.pcbi.1003570

**Editor:** Sergei L. Kosakovsky Pond, University of California San Diego, United States of America

**Received:** November 1, 2013; **Accepted:** February 28, 2014; **Published:** April 17, 2014

**Copyright:** © 2014 Rasmussen et al. This is an open-access article distributed under the terms of the Creative Commons Attribution License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original author and source are credited.

**Funding:** DAR was funded by U.S. National Science Foundation Graduate Research Fellowship. EMV was supported by an NIAID K01 Career Development Award (AI91440). KK was supported by a NSF grant (NSF-EF-08-27416) and by a Complex Systems grant from the James S. McDonnell Foundation to KK. The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

**Competing Interests:** The authors have declared that no competing interests exist.

\* E-mail: david.rasmussen@duke.edu

## Introduction

Genealogies can provide valuable information about the demographic history of a population because the demography of a population can dramatically shape the structure of a genealogy [1,2]. For example, fluctuations in population size will shift the distribution of branching events, or coalescent times, over a genealogy relative to what would be expected for a population with a constant size [3]. Other aspects of a population's demographic history can also leave behind distinctive genealogical patterns. For example, the structuring of a population into different subpopulations can influence the topology of genealogies, which is often seen as clustering among individuals sampled from the same subpopulation [4]. These observations have led to great interest in statistical methods for inferring demographic trends and parameters from genealogies and given rise to the new field of phylodynamic inference [2,5–8].

Most statistical methods for reconstructing the demographic history of a population from genealogies have been motivated by coalescent theory, which provides a probabilistic framework for relating the demographic history of a population to a genealogy of individuals sampled from that population [9,10]. Critically, coalescent models provide a way to compute the probability of a

given genealogy under a given demographic model. It is therefore possible to estimate parameters of a demographic model, such as population size, from a genealogy using likelihood-based inference methods. Extensions of this basic idea have been used to estimate changes in population size over time, for example by the Bayesian skyline methods available in the BEAST phylogenetic software package [11,12]. Coalescent theory has also been extended to consider different forms of population structure, giving rise to structured coalescent models [13,14]. Statistical methods that allow fitting of structured coalescent models to genealogies have the ability to estimate parameters relating to population structure, including migration rates between populations [7,15].

Recent developments in phylodynamics have focused on developing models and statistical methods for more complex demographic scenarios, which have been largely motivated by the application of coalescent methods to pathogens like RNA viruses with rapidly changing population sizes. For example, coalescent models have been developed for populations where birth (i.e. transmission) rates vary over time [16,17]. Importantly, the framework of Volz *et al.* [16] also considers the coalescent process in populations where transmission rates change over time in a nonlinear manner, as is often the case for epidemiological models

## Author Summary

Mathematical models play an important role in our understanding of what processes drive the complex population dynamics of infectious pathogens. Yet developing statistical methods for fitting models to epidemiological data is difficult. Epidemiological data is often noisy, incomplete, aggregated across different scales and generally provides only a partial picture of the underlying disease dynamics. Using nontraditional sources of data, like molecular sequences of pathogens, can provide additional information about epidemiological dynamics. But current “phylogenetic” inference methods for fitting models to genealogies reconstructed from sequence data have a number of major limitations. We present a statistical framework that builds upon earlier work to address two of these limitations: population structure and stochasticity. By incorporating population structure, our framework can be applied in cases where the host population is divided into different subpopulations, such as by spatial isolation. Our framework also takes into consideration stochastic noise and can therefore capture the inherent variability of epidemiological dynamics. These advances allow for a much wider class of epidemiological models to be fit to genealogies in order to estimate key epidemiological parameters and to reconstruct past disease dynamics.

like the well-known Susceptible-Infected-Recovered (SIR) model [18]. Coalescent models have also been developed for common epidemiological scenarios with population structure that alters the rate of coalescence in the population [19], but these models are limited to populations at equilibrium. Finally, Volz [20] presented a framework that brings together both complex population dynamics and population structure. This approach has great appeal as it generalizes coalescent models to allow both birth and migration rates to change over time as a function of the underlying population dynamics, which may be nonlinear and far from equilibrium.

Although recent advances with structured coalescent models have enabled the analysis of more complex epidemiological models, the statistical challenge remains of efficiently fitting stochastic population dynamic models to genealogies. These models can be extremely high-dimensional due to a large number of latent state variables for which we have no direct observations. In Rasmussen *et al.* [21], a particle filtering approach was used to marginalize out these latent variables by forward simulating population dynamic trajectories from the epidemiological model and then averaging over these trajectories to compute a marginal likelihood. For unstructured models, adapting particle filtering methods to coalescent-based inference is relatively straightforward as the likelihood of a genealogy is simply a function of the simulated population dynamic trajectories. However, for structured models the likelihood also depends on the internal states of lineages in the genealogy, which may change over time as lineages move between populations [20]. The probable state of a lineage can only be calculated retrospectively conditional on the population’s demographic history and the state of the lineage at the time of sampling. As we show below, these backward-time dependencies prevent the direct application of forward-time particle filtering methods to structured models.

We therefore present a new statistical approach for fitting stochastic population dynamics models to genealogies using the structured coalescent approach presented in Volz [20] using a

modified particle filtering algorithm. This modified algorithm allows for efficient particle filtering under structured coalescent models where the probability that a lineage is in a certain population may depend on both the past dynamics of the population as well as future sampling of lineages. Using this algorithm, we can fit stochastic, nonlinear epidemiological models with essentially any form of population structure to genealogies as long as the model is Markovian. Because population structure arises naturally in many epidemiological models, we define population structure in a very broad sense and consider any model where the population of infected hosts is structured into different nonequivalent states and therefore lineages in different infected hosts do not necessarily have an equal probability of coalescing. This includes models with spatial structure, multiple stages of infection and models of vector-borne and other multi-host pathogens.

The paper has the following structure. First, we present the forward-time epidemiological models that we use as examples throughout the paper. Next, we review the framework first developed in Volz [20] for how coalescent models can be derived for a corresponding forward-time population dynamic model. We then describe how we can fit structured epidemiological models to genealogies given the corresponding structured coalescent model. The statistical method we describe combines MCMC methods with our particle filtering algorithm, and is a variation of the particle MCMC algorithm of Andrieu *et al.* [22]. Using simulated genealogies, we show that this algorithm can accurately reconstruct population dynamics in structured populations and obtain reliable estimates of epidemiological parameters such as transmission rates. We then apply our approach to the HIV epidemic in Detroit, Michigan in order to estimate stage-specific transmission rates and infer how prevalence and incidence have changed over the course of the epidemic. Finally, we explore under what conditions parameters relating to population structure can be inferred from genealogies and how factors such as sample size affect uncertainty in our estimates.

## Methods

### Epidemiological models

In this paper, we use epidemiological models to demonstrate how mechanistic population dynamic models can be fit to genealogies. More specifically, we will consider the type of Susceptible-Infected-Recovered (SIR) models widely used to study the transmission dynamics of infectious diseases [18,23]. In SIR-type models, the host population is divided into different compartments depending on the host’s state (e.g. susceptible or infected). For generality, we let  $x_t$  be the vector that holds the number of hosts in each compartment at time  $t$ , for example  $x_t = \{S_t, I_t, R_t\}$  for the standard SIR model. For stochastic models, the state variables in  $x_t$  are treated as random variables. We consider an epidemiological model to be structured if there is more than one class of infected host.

**Applications.** We use two simple structured epidemiological models throughout the paper as illustrative examples. The first is a SIR model with three stages of infection, which illustrates how our approach can be applied to models where infected hosts progress through different stages of infection. In the Results section, we apply this model to HIV data so we assume that these three stages correspond to the early, chronic and AIDS stages of HIV infection. The deterministic skeleton of the three-stage SIR model can be written as the following system of ordinary differential equations:

$$\begin{aligned}
\frac{dS}{dt} &= \mu N - \Lambda(t)S - \mu S \\
\frac{dI_E}{dt} &= \Lambda(t)S - \gamma_E I_E - \mu I_E \\
\frac{dI_C}{dt} &= \gamma_E I_E - \gamma_C I_C - \mu I_C \\
\frac{dI_A}{dt} &= \gamma_C I_C - \gamma_A I_A.
\end{aligned} \tag{1}$$

Infections progress from one stage to the next according to the rates  $\gamma_E$  and  $\gamma_C$ . We assume there is no recovery and that individuals with AIDS infection die at rate  $\gamma_A$  instead of the normal host mortality rate  $\mu$ , where generally  $\gamma_A > \mu$ .

The force of infection  $\Lambda(t)$  is given by

$$\Lambda(t) = e^{-\alpha(I_E + I_C + I_A)} \frac{(\beta_E I_E + \beta_C I_C + \beta_A I_A)}{N}, \tag{2}$$

where  $N$  is the host population size ( $N = S + I_E + I_C + I_A$ ). The exponential term in (2) allows incidence to scale nonlinearly with the prevalence of HIV in the population and has been frequently used in HIV models [24–26]. This nonlinear scaling may reflect heterogeneity in sexual contact rates or behavioral changes as awareness or diagnosis of the disease grows.

The second model we consider is a simple two-population SIR model where transmission can occur both within and between the two populations due to infectious individuals coming into contact with susceptible individuals in either population. While we do not explicitly define the factor that structures the population, population structure could be due to spatial structure or other factors like age that affect the probability of different hosts contacting one another. The deterministic skeleton of this model can be written as follows:

$$\begin{aligned}
\frac{dS_1(t)}{dt} &= \mu N_1(t) - \beta_W(t) \frac{S_1(t)}{N_1(t)} I_1(t) - \beta_B(t) \frac{S_1(t)}{N_1(t)} I_2(t) - \mu S_1(t) \\
\frac{dI_1(t)}{dt} &= \beta_W(t) \frac{S_1(t)}{N_1(t)} I_1(t) + \beta_B(t) \frac{S_1(t)}{N_1(t)} I_2(t) - \nu I_1(t) - \mu I_1(t) \\
\frac{dS_2(t)}{dt} &= \mu N_2(t) - \beta_W(t) \frac{S_2(t)}{N_2(t)} I_2(t) - \beta_B(t) \frac{S_2(t)}{N_2(t)} I_1(t) - \mu S_2(t) \\
\frac{dI_2(t)}{dt} &= \beta_W(t) \frac{S_2(t)}{N_2(t)} I_2(t) + \beta_B(t) \frac{S_2(t)}{N_2(t)} I_1(t) - \nu I_2(t) - \mu I_2(t).
\end{aligned} \tag{3}$$

The parameter  $\mu$  is the host birth/death rate and  $\nu$  is the rate at which infected hosts recover.  $N_1$  and  $N_2$  are the host population sizes, respectively.  $\beta_W$  is the within-population transmission rate and  $\beta_B$  is the between-population transmission rate. We write the transmission rates as  $\beta_W(t)$  and  $\beta_B(t)$  to allow the transmission rate to vary seasonally. Both  $\beta_W$  and  $\beta_B$  are scaled relative to a base transmission rate  $\beta$  such that  $\beta_W = \beta\rho$  and  $\beta_B = \beta(1 - \rho)$ , so that the parameter  $\rho$  controls the extent of mixing or coupling between the two populations, as in Keeling and Rohani [27]. Under this parameterization, the basic reproductive number  $R_0 = \beta/(\mu + \nu)$  and is therefore invariant to changes in  $\rho$  so that we can vary the degree of mixing between populations while not significantly altering the overall epidemiological dynamics.

## Coalescent models

In this section, we consider formulating structured coalescent models for the type of structured epidemiological models just presented. As shown in Volz [20], thinking about population dynamic models as simple birth-death processes can be useful when deriving coalescent models that correspond to a given forward-time model. If we randomly sample individuals from a population and trace their ancestry back in time, then coalescent events in the genealogy will correspond to birth events in the population when both the parent and child lineages are ancestral to sampled individuals. While deaths may affect the overall population size, deaths can be ignored along lineages ancestral to sampled individuals because we know that a lineage could not have died out at an earlier time if it persisted to be sampled at some later time. For a structured population, we also must consider individuals transitioning between different subpopulations through migration events that occur independently of birth events, although for the type of models we will consider here a lineage can also transition between populations by being born into a different population than its parent.

The same birth-death-migration framework can be applied to pathogens if we assume that each infected host corresponds to a single individual in the pathogen population. In this case, births in the pathogen population occur at transmission events between hosts. Deaths in the population will correspond to recovery or mortality of infected hosts. If each infected host is represented by a single pathogen lineage, coalescent events in the genealogy will correspond to transmission events if both the infected host and the infector are sampled or give rise to descendent infections that are sampled. For structured epidemiological models, we also must consider a pathogen lineage transitioning among populations, or compartments in SIR-type models, independent of transmission events. For example, in the three-stage model, pathogen lineages can transition between different stages of infection. Here, we will refer to all transitions between states that occur independently of transmission as migration for generality. This allows many epidemiological models with some form of population structure to be thought of as a birth-death-migration process.

To formalize the birth process, we adopt the notation of Volz [20] and let  $F(t)$  be a matrix that specifies the birth rate of new lineages in the population at time  $t$ , where  $F(t) = F(\theta, x_t)$ , meaning that  $F(t)$  can be a function of the epidemiological parameters  $\theta$  and the population state variables  $x_t$ . Lineages may be in any one of  $m$  states. The rate at which lineages currently in state  $k$  give birth to lineages in state  $l$  is given by the element  $f_{kl}$ . The rate at which migration, or transitions between states independent of birth events, occurs is given by another matrix  $G(t) = G(\theta, x_t)$ . The rate at which lineages currently in state  $k$  migrate to state  $l$  is given by the element  $g_{kl}$ . We treat birth and migration as distinct processes because, as we will see, they affect the coalescent process in different ways since coalescent events can only occur at birth events but migration events can affect the probability of a particular lineage coalescing with another lineage. The total number of lineages in each state is given by a vector  $Y(t)$ , such that  $y_k(t)$  gives the total number of individuals in the population in state  $k$  at time  $t$ . From here in, we drop the time indices and just refer to the matrices  $F$  and  $G$  or the vector  $Y$ , but emphasize that the rates in  $F$  and  $G$  and the population sizes in  $Y$  can be time-dependent.

We illustrate the  $F$  and  $G$  matrix notation by decomposing the three-stage and two-population SIR models presented above into their component birth and migration processes. For the three-stage model, we have

$$F = \begin{pmatrix} \beta_E \frac{S}{N} I_E & 0 & 0 \\ \beta_C \frac{S}{N} I_C & 0 & 0 \\ \beta_A \frac{S}{N} I_A & 0 & 0 \end{pmatrix}, \quad (4)$$

$$G = \begin{pmatrix} 0 & \gamma_E I_E & 0 \\ 0 & 0 & \gamma_C I_C \\ 0 & 0 & 0 \end{pmatrix}. \quad (5)$$

In the  $F$  matrix, births occur through transmission of the pathogen from any of the three stages of infection to susceptible individuals. Because all new infections begin in the early stage, only the leftmost column of the  $F$  matrix has nonzero elements. The nonzero elements in the  $G$  matrix correspond to migration between stages through disease progression from early to chronic and from chronic to AIDS.

For the two-population model, we have

$$F = \begin{pmatrix} \beta_W \frac{S_1}{N_1} I_1 & \beta_B \frac{S_2}{N_2} I_1 \\ \beta_B \frac{S_1}{N_1} I_2 & \beta_W \frac{S_2}{N_2} I_2 \end{pmatrix}, \quad (6)$$

$$G = \begin{pmatrix} 0 & 0 \\ 0 & 0 \end{pmatrix}. \quad (7)$$

Because transmission events can move the pathogen within and between the two populations in either direction, all entries in the  $F$  matrix are nonzero. The  $G$  matrix has all zero entries because there is no migration between populations independent of transmission.

Before moving on, we note that for an infectious pathogen our coalescent models make the implicit assumption that coalescent events in the genealogy correspond to transmission events between hosts. In essence then, we are ignoring the within-host coalescent process and assuming that all infected hosts are represented by a single lineage. This implies that lineages immediately coalesce once in the same infected host, which may not be true for certain pathogens where multiple lineages can persist within a host for long periods of time. Nevertheless, in general our assumption that each infected host is represented by a single pathogen lineage will be valid as long as super-infection is rare and there is a strong bottleneck in the pathogen population at transmission events so that it is unlikely that more than one lineage is transmitted between hosts.

### Coalescent likelihoods

To fit a structured coalescent model to a genealogy, we need to compute the likelihood of the coalescent model given the genealogy. To compute this likelihood, we can partition the genealogy into any number of discrete time intervals. We label the time partitioned genealogy  $\mathcal{G}_{1:T}$ , where  $t=1$  is the time of the first event in the genealogy and  $t=T$  is the final event time going forwards in time (usually the terminal-most sampling event). Time points are chosen to correspond to the times at which events in the genealogy occur such as coalescent and sampling events. We can then further subdivide the genealogy into smaller intervals that correspond to the  $\Delta t$  time steps used to simulate from the epidemiological model so that at any time point  $t$  we have the state

variables  $x_t$  corresponding to that time. With the time partitioned genealogy  $\mathcal{G}_{1:T}$ , we can compute the likelihood over each interval in the genealogy,  $\mathcal{G}_{t-1:t}$ , and then take the product over all intervals to compute the total likelihood of the model given  $\mathcal{G}_{1:T}$ .

Computing the likelihood over a time interval  $\mathcal{G}_{t-1:t}$  requires us to first compute the probabilities that the lineages present in the genealogy did or did not coalesce within that time interval. The probability of a coalescent event in turn depends on the expected rate of coalescence under the model. This expected rate can be computed for a coalescent model with any arbitrary population structure using the formalism summarized above for the rates of birth in  $F$ . As shown in Volz [20], the rate of coalescence  $\lambda_{ij}$  for two lineages  $i$  and  $j$  is

$$\lambda_{ij} = \sum_k^m \sum_l^m \frac{f_{kl}}{y_k y_l} (p_{ik} p_{jl} + p_{il} p_{jk}) \quad (8)$$

where, for example,  $p_{ik}$  is the probability that lineage  $i$  is in state  $k$ . How these lineage state probabilities are computed is explained below. We can make intuitive sense of the coalescent rate in (8) by noting that  $f_{kl}$  is the total rate at which lineages in state  $k$  give birth to lineages in state  $l$  in the population and that  $\frac{1}{y_k y_l}$  is the probability that lineages  $i$  and  $j$  are the two lineages involved in a particular birth event. However, since we do not know the true states of  $i$  and  $j$  we must sum over all possible combinations of states for these two lineages.

The total rate of coalescence  $\lambda_{\mathcal{A}}$  for all lineages  $\mathcal{A}$  present in the genealogy over an interval of time is then

$$\lambda_{\mathcal{A}} = \sum_{i \in \mathcal{A}} \sum_{j \in \mathcal{A}, j < i} \lambda_{ij}. \quad (9)$$

Given the rates of coalescence, we can then compute the likelihood over a time interval  $\mathcal{G}_{t-1:t}$  under the coalescent model. If the time interval does not end in a coalescent event, we have

$$L(\mathcal{G}_{t-1:t}) = e^{-\lambda_{\mathcal{A}} \Delta t}. \quad (10)$$

Alternatively, if the interval does end in a coalescent event between two lineages  $i$  and  $j$ , we have

$$L(\mathcal{G}_{t-1:t}) = \lambda_{ij} e^{-\lambda_{\mathcal{A}} \Delta t}. \quad (11)$$

### Lineage state probabilities

As alluded to above, computing the coalescent rates requires us to compute the probability of each lineage in the genealogy being in each possible state. At the time of sampling, we may know the state of a lineage from information gathered from the infected host from which the sample was obtained. Alternatively, if we do not know the state of the host at the time of sampling exactly, we can assign prior probabilities to the lineage being in each state under a multinomial distribution. Either way, given the initial state or state probabilities at the time of sampling, we need to be able to compute the probability of the lineage being in each state at any point in the past.

Going backwards in time, the lineages transition between states at the rates given in the  $F$  and  $G$  matrices, which in turn depends on the population states  $x_{1:T}$  and the parameters  $\theta$ . Given these transition rates, we have a continuous time Markov process on a

discrete state space along each branch. We can therefore use master equations to track how the lineage state probabilities change going backwards through time. In other words, we can write down differential equations for how the probability mass assigned to each state flows between states as we move into the past. As shown in Volz [20], the general form that these master equations take for any lineage  $i$  and state  $k$  is

$$\frac{d}{dt}p_{ik} = \sum_l \left( p_{il} \frac{g_{kl}}{y_l} - p_{ik} \frac{g_{lk}}{y_k} + p_{il} \frac{f_{kl}y_k - A_k}{y_l y_k} - p_{ik} \frac{f_{lk}y_l - A_l}{y_k y_l} \right), \quad (12)$$

where  $A_k = \sum_{i \in A} p_{ik}$ ; that is  $A_k$  is the expected number of lineages in state  $k$  in the genealogy at a given point in time. Further details on how the lineage state probabilities are computed and get updated at coalescent events are given in Text S1. For convenience, we introduce the notation  $\mathcal{P}_t$  to denote the lineage state probabilities for all lineages in the genealogy at time  $t$  and  $\mathcal{P}_{1:T}$  to denote the complete mapping of lineage state probabilities onto the genealogy over the entire time partitioned genealogy  $\mathcal{G}_{1:T}$ .

### Statistical inference

The goal of phylodynamic inference for the type of models presented above will generally be to infer the parameters of interest from the genealogy along with the latent population state variables, such as the number of infected or susceptible hosts over time. In a Bayesian context then, we would like to infer the joint posterior density of the model parameters  $\theta$  and the latent state variables  $x_{1:T}$ . Up to a normalizing constant, this posterior density is given by

$$p(\theta, x_{1:T} | \mathcal{G}_{1:T}) \propto p(\mathcal{G}_{1:T} | \theta, x_{1:T}) p(x_{1:T} | \theta) p(\theta). \quad (13)$$

From (13), we see that this joint density can be factored into three parts: the coalescent likelihood  $p(\mathcal{G}_{1:T} | \theta, x_{1:T})$  which we outlined how to compute above; the prior density on the population state variables  $p(x_{1:T} | \theta)$  as defined by the epidemiological process model; and the prior density on the parameters  $p(\theta)$ . Although we may be able to compute each component individually and thereby the posterior probability of a given set of parameters  $\theta$  and population states  $x_{1:T}$ , the posterior density is not analytically tractable in general and we must resort to sampling from the posterior using MCMC methods.

However, it may be difficult or impossible to sample from complex, high-dimensional densities such as  $p(\theta, x_{1:T} | \mathcal{G}_{1:T})$  using standard MCMC methods. We could, for example, use a Gibbs sampler to iteratively sample from the conditional posterior densities of  $\theta$  and any component of  $x_{1:T}$ , but this strategy can be extremely inefficient owing to strong correlations among the parameters and the state variables, leading to slow MCMC mixing [28]. In Rasmussen *et al.* [21], a particle MCMC approach known as the particle marginal Metropolis-Hastings (PMMH) algorithm was therefore used to sample from the joint posterior density of  $\theta$  and  $x_{1:T}$ . The main motivation behind using the PMMH algorithm is that we can jointly update  $\theta$  and  $x_{1:T}$  together [22]. Each MCMC iteration, we first propose new parameter values  $\theta^*$  and then run a particle filtering algorithm to get a numerical approximation of the posterior density of the latent state variables  $p(x_{1:T} | \mathcal{G}_{1:T}, \theta^*)$ , which we refer to as  $\hat{p}(x_{1:T} | \mathcal{G}_{1:T}, \theta^*)$ . Particle filtering, also known as sequential Monte Carlo, provides a computational means of approximating high dimensional densities by providing samples (i.e. the particles) distributed according to the

desired density, and are often used in the context of nonlinear and non-Gaussian state space models [29–31]. We review how particle filters can be used to fit epidemiological models to genealogies in Text S1.

After running the particle filtering step in the PMMH algorithm, we can then sample a particle from  $\hat{p}(x_{1:T} | \mathcal{G}_{1:T}, \theta^*)$  to get a proposal  $x_{1:T}^*$  for the latent state variables that is adapted to the parameters in  $\theta^*$ . We can also use the particle filter to compute the marginal likelihood of  $\theta^*$  by marginalizing out the state variables. Because we jointly accept  $\theta^*$  and  $x_{1:T}^*$  based on the marginal likelihood, we do not have to independently update  $x_{1:T}$ , leading to a much more efficient MCMC sampler. Despite marginalizing out the latent state variables, the remarkable feature of the PMMH algorithm is it provides an exact (i.e. unbiased) approximation to the density of interest,  $p(\theta, x_{1:T} | \mathcal{G}_{1:T})$ . The PMMH algorithm is summarized in pseudo-code below.

**Algorithm 1.** The PMMH sampler targeting  $p(\theta, x_{1:T} | \mathcal{G}_{1:T})$

At each MCMC iteration, with current parameter values  $\theta$ :

1. Sample  $\theta^*$  from a proposal density  $q(\theta^* | \theta)$ .
2. Run particle filter to sample  $x_{1:T}^*$  from  $\hat{p}(x_{1:T} | \mathcal{G}_{1:T}, \theta^*)$  and obtain the marginal likelihood estimate  $\hat{p}(\mathcal{G}_{1:T} | \theta^*)$ .
3. Accept  $\theta^*$  and  $x_{1:T}^*$  with probability

$$\min \left( \frac{\hat{p}(\mathcal{G}_{1:T} | \theta^*) p(\theta^*) q(\theta | \theta^*)}{\hat{p}(\mathcal{G}_{1:T} | \theta) p(\theta) q(\theta^* | \theta)}, 1 \right). \quad (14)$$

While the PMMH algorithm described above works for unstructured epidemiological models where all infected hosts are assumed to be in the same population, we encounter an additional problem for structured epidemiological models. In this case, the inference task at hand becomes more difficult because we need to take into account the unknown lineage states. This is done by conditioning on the lineage state probabilities  $\mathcal{P}_{1:T}$  when computing the coalescent likelihood. We can make this dependence on the lineage state probabilities clear by rewriting the likelihood as  $p(\mathcal{G}_{1:T} | \theta, x_{1:T}, \{\mathcal{P}_{1:T}\})$ . We use the notation  $\{\mathcal{P}_{1:T}\}$  to indicate that while the lineage state probabilities are required to compute the coalescent likelihood, we are not treating the lineage states as random variables but rather as probabilities that are completely determined by the master equations shown in (12) given  $\theta$  and a population state trajectory  $x_{1:T}$ .

Recall that the probability of a lineage being in a certain state in the past depends conditionally on the state of the lineage at the time of sampling. This creates a backwards-time dependence structure that cannot easily be accommodated by the forwards in time particle filtering methods used in the PMMH approach. This is because the computational efficiency of the particle filter largely relies on the ability to resample—replacing particles with low weights with particles with high weights. In order to resample, we need to be able to compute the particle weights at any time  $t$ , which in turn requires the ability to compute the likelihood  $p(\mathcal{G}_{t-1:t} | \theta, x_t, \{\mathcal{P}_t\})$  over any time interval. Computing this likelihood therefore entails being able to compute the lineage state probabilities  $\mathcal{P}_t$ , which will depend on the future states of the system  $x_{t+1:T}$  for any structured model. The backward-time dependency of the lineage state probabilities therefore prohibits resampling and thus compromises the efficiency of the particle filter. We therefore use a modified particle filtering scheme that allows us to resample by computing expected lineage state probabilities before running the particle filter and then applying a correction step to counteract any bias introduced by using the

expected rather than the true lineage state probabilities while filtering.

In more detail, the algorithm proceeds as follows. We initially simulate a deterministic trajectory from the epidemiological model for the state variables in  $x_{1:T}$ , which we refer to as  $\bar{x}_{1:T}$ . We can then compute the expected lineage state probabilities  $\bar{\mathcal{P}}_{1:T}$  going backwards in time conditional on  $\bar{x}_{1:T}$ . We then run the particle filter forward in time to approximate the density  $p(x_{1:T}|\mathcal{G}_{1:T}, \theta, \{\bar{\mathcal{P}}_{1:T}\})$ . Although  $p(x_{1:T}|\mathcal{G}_{1:T}, \theta, \{\bar{\mathcal{P}}_{1:T}\})$  is not ultimately the target density we are interested in, it serves a useful intermediate purpose. Once we have sampled particles representing trajectories from  $p(x_{1:T}|\mathcal{G}_{1:T}, \theta, \{\bar{\mathcal{P}}_{1:T}\})$ , we can re-weight these particles and use an additional round of importance sampling to get particles representing samples from  $p(x_{1:T}|\mathcal{G}_{1:T}, \theta, \{\mathcal{P}_{1:T}\})$ , our density of interest.

To do this, for each particle  $j$  sampled using the particle filter we compute the true lineage state probabilities  $\mathcal{P}_{1:T}^j$  conditional on the actual population state trajectory of the particle  $x_{1:T}^j$ . We can therefore compute the corrected weights  $v_T^j = p(x_{1:T}^j|\mathcal{G}_{1:T}, \theta, \{\mathcal{P}_{1:T}^j\})$  using the true lineage state probabilities. We also store the expected weights  $u_T^j = p(x_{1:T}^j|\mathcal{G}_{1:T}, \theta, \{\bar{\mathcal{P}}_{1:T}\})$  computed using the expected lineage state probabilities. With both the expected weights  $u_T^j$  and corrected weights  $v_T^j$  we can assign final importance weights  $w_T^j = \frac{v_T^j}{u_T^j}$  and resample particles again according to the final weights in  $w_T$ . This final round of importance sampling corrects for any bias we may have introduced by resampling particles using the expected lineage state probabilities while filtering and thereby gives us particles approximately distributed according to the target density  $p(x_{1:T}|\mathcal{G}_{1:T}, \theta, \{\mathcal{P}_{1:T}\})$ .

**Algorithm 2.** The particle filter/importance sampler targeting  $p(x_{1:T}|\mathcal{G}_{1:T}, \theta, \{\mathcal{P}_{1:T}\})$

1. Run deterministic simulation to obtain  $\bar{x}_{1:T}$  and compute the expected lineage state probabilities  $\bar{\mathcal{P}}_{1:T}$  conditional on  $\bar{x}_{1:T}$ .
2. Initialize the particle filter at time  $t=1$  with  $N$  particles.
  - (a) Set  $x_1^j$  to initial values for all particles.
  - (b) Assign normalized weights,  $U_1^j = \frac{1}{N}$ .
3. Run filter from  $t=2$  to  $t=T$ .
  - (a) Propagate particles forward by simulating from the process model  $p(x_t^j|x_{t-1}^j, \theta)$ .
  - (b) Set  $x_{1:t}^j = (x_{1:t-1}^j, x_t^j)$  for all particles.
  - (c) Compute unnormalized weights conditional on  $\bar{\mathcal{P}}_{1:t}$ ,
$$u_t^j = (u_{t-1}^j) p(\mathcal{G}_{t-1:t}|\theta, x_{1:t}^j, \{\bar{\mathcal{P}}_{1:t}\}). \quad (15)$$
  - (d) Normalize weights, so that  $U_t^j = \frac{u_t^j}{\sum_{j=1}^N u_t^j}$ .
  - (e) If resampling at  $t$ , resample according to  $U_t^j$ .
4. At time  $T$ , resample particles again according to  $U_T^j$  to get particles distributed according to  $p(x_{1:T}|\mathcal{G}_{1:T}, \theta, \{\bar{\mathcal{P}}_{1:T}\})$ .
5. Compute corrected lineage state probabilities  $\mathcal{P}_{1:T}^j$  for each particle conditional on  $x_{1:T}^j$ .
6. Compute corrected weights  $v_T^j = p(x_{1:T}^j|\mathcal{G}_{1:T}, \theta, \{\mathcal{P}_{1:T}^j\})$ .

7. Compute final importance weights  $w_T^j = \frac{v_T^j}{u_T^j}$  and normalize to get  $W_T^j$ .
8. Sample  $x_{1:T}^*$  from  $\hat{p}(x_{1:T}|\mathcal{G}_{1:T}, \theta, \{\mathcal{P}_{1:T}\})$  according to  $W_T^j$ .
9. Compute marginal likelihood estimate

$$\hat{p}(\mathcal{G}_{1:T}|\theta) = \prod_{t=1}^T \sum_{j=1}^N W_T^j v_t^j. \quad (16)$$

The particle filter/importance sampler therefore provides us with a proposal for the population state variables  $x_{1:T}^*$  and an estimate of the marginal likelihood. We can therefore plug the particle filter/importance sampler into step 2 of Algorithm 1 to obtain a PMMH algorithm for sampling from the joint posterior density of  $\theta$  and  $x_{1:T}$  under structured epidemiological models. Moreover, because we marginalize over the population state variables  $x_{1:T}$  using the particle filter and then marginalize over the lineage states by summing over all possible lineage states when computing the likelihood, we can efficiently sample from the posterior density using the PMMH algorithm without having to design proposal updates for the population states or lineage states.

Before moving on, we make a few notes about the potential limitations and efficiency of the particle filter/importance sampler. As a basic requirement of importance sampling, the support of the importance density  $p(x_{1:T}|\mathcal{G}_{1:T}, \theta, \{\bar{\mathcal{P}}_{1:T}\})$  must span the support of the target density  $p(x_{1:T}|\mathcal{G}_{1:T}, \theta, \{\mathcal{P}_{1:T}\})$ , so that wherever  $p(x_{1:T}|\mathcal{G}_{1:T}, \theta, \{\mathcal{P}_{1:T}\}) > 0$  so must  $p(x_{1:T}|\mathcal{G}_{1:T}, \theta, \{\bar{\mathcal{P}}_{1:T}\}) > 0$ . However, in order for the particle filter/importance sampler to be efficient, the density  $p(x_{1:T}|\mathcal{G}_{1:T}, \theta, \{\mathcal{P}_{1:T}\})$  should also be similar to  $p(x_{1:T}|\mathcal{G}_{1:T}, \theta, \{\bar{\mathcal{P}}_{1:T}\})$ . Of course, this might not always be the case. If the stochastic particle trajectories can diverge largely from  $\bar{x}_{1:T}$  or the lineage state probabilities are highly correlated with  $x_{1:T}$  (such that small changes in the population states lead to large jumps in the lineage state probabilities), then these two densities may be quite different, causing the importance sampler to become very inefficient and requiring us to sample many particle trajectories in order to obtain a reasonable particle approximation to  $p(x_{1:T}|\mathcal{G}_{1:T}, \theta, \{\mathcal{P}_{1:T}\})$ . In such cases, it may be unwise to resample particles according to their expected weights  $u_T$  during the particle filtering stage because these weights will not be predictive of the corrected weights  $v_T$ , meaning we may be discarding particles with high posterior probability under the desired density  $p(x_{1:T}|\mathcal{G}_{1:T}, \theta, \{\mathcal{P}_{1:T}\})$ . In practice, this can easily be checked by making sure that there is a strong positive correlation between the expected and corrected weights. We found this to be true for all cases considered in this paper and found that resampling according to the expected weights during the filtering stage measurably improved the performance of the algorithm by reducing the variance in the marginal likelihood estimates, which tends to improve MCMC mixing overall.

## Simulations

We simulated mock genealogies under each model to test the performance of the PMMH algorithm before applying the method to real data. Mock genealogies were obtained by first forward simulating from the population dynamic model while tracking all infected hosts in the population and the parent-offspring relationships at transmission events. From the forward simulations, we could then trace the lineages of infected individuals backwards through time to obtain the true genealogy for a fraction of sampled lineages. All population dynamic simulations were performed

using the tau-leaping algorithm so that the epidemiological dynamics included demographic noise [32].

The three-stage model was parameterized to reflect the natural history of HIV because we planned to apply our method to real HIV genealogies (see Table 1). We set the disease progression and AIDS death rate to values that give an average time between infection and death of about 10 years, consistent with observed patterns. The incidence scaling parameter  $\alpha$  was set to zero so that in the simulations there was a linear scaling between incidence and prevalence. The epidemic simulations were seeded with one early-stage infection at time zero and run for 37 years to reflect the timespan of the HIV epidemic in the U.S. To obtain mock genealogies from the complete infection trees, we sampled 200 individuals in the last six years of the epidemic to reflect the fact that most HIV sequences have been sampled in the recent past. For all parameters, we chose to use uniform priors over a wide range of biologically plausible values so that the choice of prior would have minimal influence on our estimates.

For the two-population model, we added seasonality to the model by seasonally forcing the base transmission rate  $\beta(t)$  using a sinusoidal forcing function, where

$$\beta(t) = \beta \left( 1 + \alpha \cos \left( \frac{(t + \delta)}{2\pi} \right) \right). \quad (17)$$

The strength of seasonality  $\alpha$  was the same in both populations but we allowed  $\delta$  to differ between the two populations to get asynchronous dynamics between populations. The values of all fixed parameters in the model are also shown in Table 1. For the genealogies, 120 infected hosts were randomly sampled over time with sampling effort proportional to disease prevalence in each population. For the two-population model, we fixed the initial conditions for the number of susceptible and infected hosts in each population.

For the simulation experiments, we wished to compare estimates obtained by fitting stochastic models using the PMMH algorithm against estimates obtained by fitting deterministic models. To fit deterministic models, we used a Metropolis-Hastings sampler where, whenever new parameters are proposed, the likelihood of the genealogy under the new parameters is computed by conditioning on a deterministic trajectory of the state variables  $x_{1:T}$  simulated from the model using these new parameters.

## HIV data

We applied our method to a set of HIV-1 partial *pol* sequences collected from men who have sex with men (MSM) in the metropolitan area of Detroit, Michigan. The dataset contained 437 HIV-1 subtype B sequences which were originally collected for drug resistance testing between 2004 and 2011. More information about this dataset can be found in Volz *et al.* [26]. Data were anonymized by staff at the Michigan Department of Community Health before being provided to investigators. Because this research falls under the original mandate for HIV surveillance and was de-identified, it was classified as human subjects research but was exempt from further IRB review.

We reconstructed time-scaled genealogies from the HIV sequences in BEAST using a relaxed molecular clock [33]. All sequences identified as likely recombinants were removed from the alignment prior to the analysis. Tips in the genealogy corresponding to sampled infected individuals were assigned prior probabilities of being in each infection stage based on the time since infection estimated from CD4 cell counts and genetic diversity within the host [34].

From the HIV genealogies, we estimated the transmission rates  $\beta_E$ ,  $\beta_C$  and  $\beta_A$  as well as the incidence scaling parameter  $\alpha$ . All other parameters were fixed at the values given in Table 1. Rather than estimate initial conditions, the time of the initial introduction of HIV into Detroit was estimated, at which point the epidemic was seeded with one early-stage infection in a completely susceptible population. All priors on the parameters were uniform. For the time of initial introduction the prior was truncated at 1973 as a lower bound and the root time of each tree as an upper bound. To ensure our phylodynamic estimates of HIV incidence were reasonable, we compared our estimates against incidence back-calculated from Michigan Department of Community Health surveillance data using the method of Yan *et al.* [35].

## Implementation

For all results shown in this paper, the PMMH algorithm was run for at least 100,000 iterations or until the MCMC fully converged. For the Metropolis-Hastings step, we chose a multivariate normal proposal density for  $q(\theta|\theta^*)$ , which can take into account the correlations among different parameters by optimizing the covariance parameters that specify the density.

For the particle filter, we found that using a small number of particles ( $N = 10$ ) was sufficient. Running the particle filter with a small number of particles tends to increase the error, or variance, in the marginal likelihood estimates. However, this error will not affect inference as long as the marginal likelihood estimates are not systematically biased because the error in the estimates will get averaged out in the encompassing MCMC algorithm. Nevertheless, with too few particles we run the risk of the MCMC getting stuck at erroneously high values of the likelihood. Our choice of  $N = 10$  was therefore a compromise between minimizing the error in the marginal likelihood estimates and the time taken to run the particle filter. Resampling within the particle filter was done by multinomial sampling with replacement. Resampling times were chosen to minimize the variance in the marginal likelihood estimates and were usually placed around coalescent events, as most of the variation in particle weights arises at coalescent times.

The PMMH algorithm was implemented in the software package PHYLter and Java source code is freely available at <http://code.google.com/p/phylyter/>. Running the PMMH algorithm for 100,000 iterations using the simulated HIV genealogies took approximately 10 hours (0.36 s per iteration) on a 3.4 GHz Intel i7 processor without any parallelization across cores. The most computationally intensive component of the algorithm is computing the lineage state probabilities, which involves numerically solving the master equations for each lineage in the genealogy and has a time complexity of  $\mathcal{O}(m^2)$ , where  $m$  is the number of possible lineage states. On the other hand, run times scale linearly with the number of particles and lineages in the genealogy. Thus, the efficiency of the algorithm is mainly limited by the number of states in the model.

## Results

### Testing the algorithm

Before applying the PMMH algorithm to genealogies reconstructed from real data, we ran extensive simulations to ensure that we could accurately recover epidemiological parameters and population dynamics from mock genealogies. We simulated 100 stochastic realizations of an epidemic from the three-stage model, keeping track of the underlying infection tree so that we could obtain the true genealogy for a fraction of sampled lineages. From the simulated epidemic dynamics, we can see that demographic stochasticity generates considerable variation in when the



**Table 1.** Fixed parameters in the epidemiological models.

Three-stage model		Two-population model	
Initial population size	$N = 10,000$	Initial population sizes	$N_1 = N_2 = 2 \text{ million}$
Birth/death rate	$\mu = \frac{1}{40.28} \text{ yr}^{-1}$	Birth/death rate	$\mu = \frac{1}{70} \text{ yr}^{-1}$
Progression rate	$\gamma_E = 1 \text{ yr}^{-1}$	Recovery rate	$\nu = \frac{1}{7} \text{ day}^{-1}$
Progression rate	$\gamma_C = \frac{1}{6.31} \text{ yr}^{-1}$	Seasonal amplitude	$\alpha = 0.08$
AIDS death rate	$\gamma_A = \frac{1}{2.55} \text{ yr}^{-1}$	Seasonal phase	$\delta_1 = 0.0 \text{ yrs}$ $\delta_2 = 0.5 \text{ yrs}$

doi:10.1371/journal.pcbi.1003570.t001

epidemic begins and peaks (Figure S1). Even with this variability, we accurately inferred stage-specific prevalence and transmission rates from the mock genealogies using the PMMH algorithm (Figure 1). The 95% credible intervals generally contained the true prevalence for all three stages of infection (Figure 1A). We were also able to estimate the stage-specific transmission rates associated with each stage of infection (Figure 1B–D), even though there were strong correlations among the different transmission rates as seen in the pairwise joint posterior densities (Figure 1E–G). Overall, out of all 100 simulations, the 95% credible intervals contained all three transmission rates 94 times, while the posterior coverage was greater than 95% for each parameter individually. In contrast, when we fit deterministic models to the same set of genealogies, the credible intervals contained the true parameters only 79% of the time. The PMMH algorithm therefore appears to give reliable estimates of parameters and epidemiological dynamics and outperforms deterministic methods when stochasticity plays a role in the epidemic dynamics.

### HIV in Detroit

Given that we were able to reliably estimate transmission parameters and prevalence in our simulation study, we next applied the method to HIV genealogies reconstructed from sequences collected in Detroit, Michigan. A critical question in HIV epidemiology is to what extent transmission during the early stages of infection contributes to overall HIV incidence. Transmission during early infection may influence the effectiveness of interventions based on antiretroviral treatment in limiting the epidemic [36,37]. If most new cases of HIV result from recently infected individuals, then prevention strategies that rely on treating diagnosed individuals, who are likely in later stages of infection, will directly prevent few transmissions. Thus, the transmission rate from early HIV infections (EHI) is a key parameter of great interest, although difficult to measure directly from traditional surveillance data. Phylogenetic studies of HIV have used the high degree of clustering and short branch times within these clusters to argue for a high EHI transmission rate [4,38]. However, clustering alone cannot be taken as definitive evidence for high EHI transmission as similar patterns can arise simply from epidemic transmission dynamics [26]. In this section, we demonstrate that our inference framework can be used to estimate the EHI transmission rate and the number of new HIV infections attributable to EHI from HIV genealogies using models that explicitly consider HIV's transmission dynamics, as well as the stochastic nature of the epidemic dynamics.

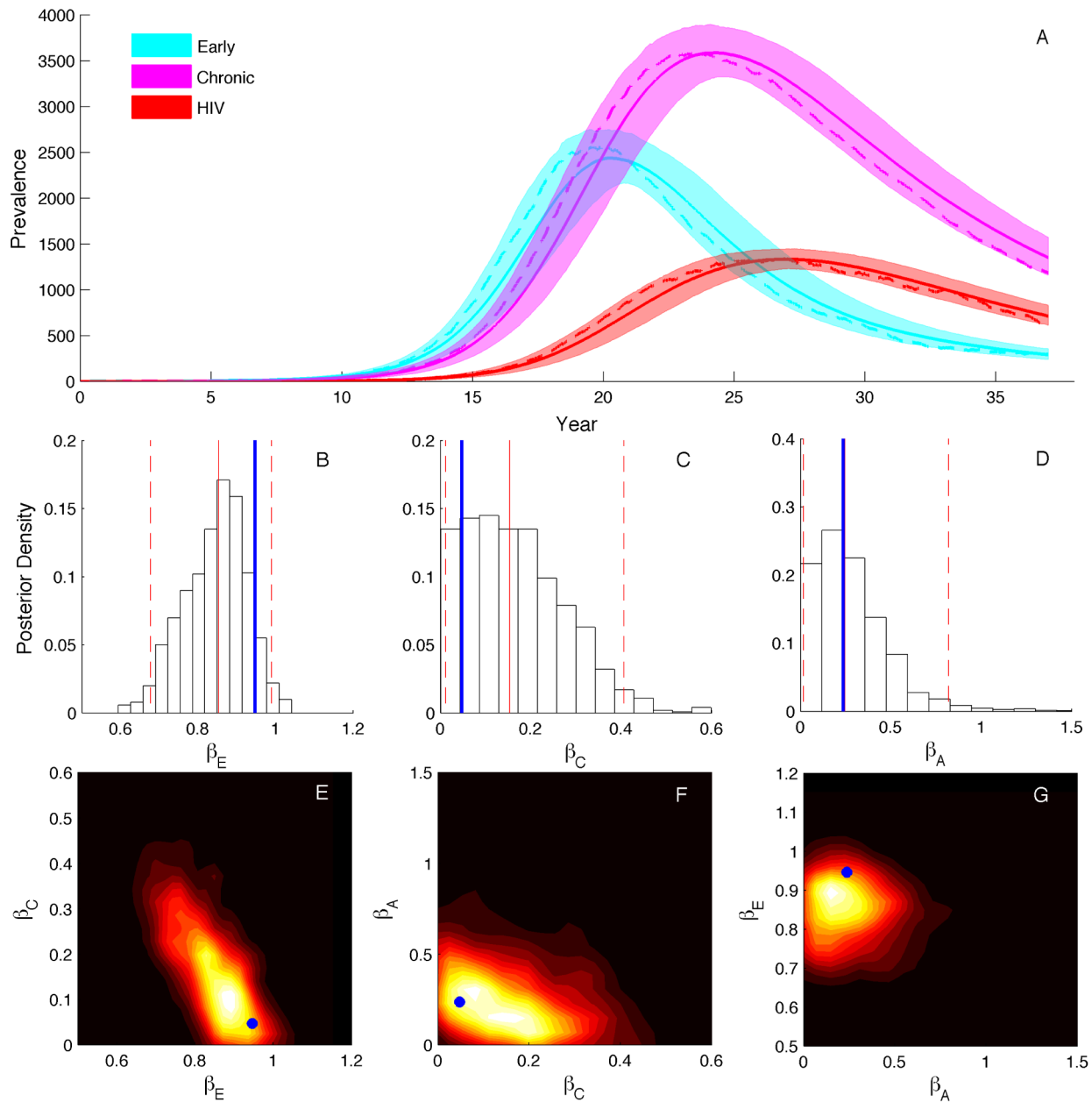
Time-scaled genealogies were reconstructed using BEAST from HIV-1 partial *pol* sequences isolated from men who have sex with

men (MSM) in the metropolitan area of Detroit. A representative genealogy randomly sampled from the BEAST posterior is shown in Figure S2. We then fit our three-stage SIR model to 10 genealogies sampled from the BEAST posterior to take into account uncertainty in the genealogy. From these genealogies, we estimated the transmission rate for each stage, including the EHI transmission rate, along with the stage-specific dynamics of prevalence and the incidence (i.e. number of new cases) attributable to each stage over the course of the epidemic.

Parameters estimated from the representative HIV genealogy are shown in Figure 2 and estimates from all 10 genealogies are given in Table 2. We estimated that transmission rates are higher during the early and AIDS stages than during the chronic stage, as expected from previous studies [39–41]. The transmission rate from EHI is about 20 times higher than during the chronic stage and about five times higher than during the AIDS stage (Figure 2A–C). We also found evidence for a nonlinear dependence of incidence on prevalence, quantified through the incidence scaling parameter  $\alpha$ . Although estimated values of  $\alpha$  are small, the posterior density is clearly centered away from zero, indicating that incidence scales nonlinearly with prevalence (Figure 2D). Overall, parameter estimates were largely consistent across genealogies, although there was considerable variation in the time of initial introduction of HIV into Detroit estimated from different trees. This is likely attributable to the large amount of variation in the root times inferred for different trees, as we inferred earlier times of introduction from trees with earlier root times.

Stage-specific HIV prevalence inferred from the genealogies shows a predictable transition from most infections being in the early stage at the beginning of the epidemic to most infections being in the chronic or AIDS stages later in the epidemic (Figure 3A). This is expected given the longer duration of the chronic and AIDS stages. In general, our phylodynamic estimates of the epidemic dynamics closely track HIV incidence imputed from surveillance data from the beginning of the epidemic through the peak (Figure 3B). While our phylodynamic estimates do not capture the fluctuations in incidence that occur after 1990, there was nothing in our model that would allow us to reproduce this pattern, which likely results from complex changes in HIV treatment and behavioral changes [34]. Although there was also considerable variability in the population dynamics inferred from different genealogies, this variation occurs primarily during the early stages of the epidemic (Figure 3C). Again, this appears to be associated with uncertainty in the root times of trees; dynamics inferred from trees with earlier root times show an earlier rise and peak in incidence. After the epidemic peaks, the incidence estimated from different trees seems to converge on similar values.



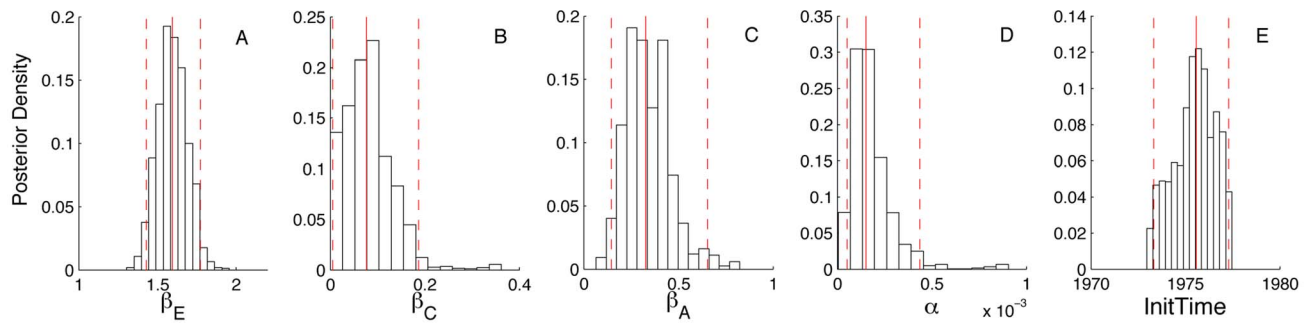


**Figure 1. Prevalence and transmission rates estimated from a representative genealogy simulated under the three-stage SIR model.** (A) Stage-specific prevalence estimates with the 95% credible intervals shaded and the posterior medians shown as solid lines. Dashed lines show the true prevalence. (B–D) The marginal posterior densities of the stage-specific transmission rates. (E–G) The corresponding pairwise joint densities of the transmission rates, which were constructed from the MCMC samples using nonparametric kernel density estimation. doi:10.1371/journal.pcbi.1003570.g001

Estimates of incidence attributable to each stage show that EHI contributed to most new infections at the beginning of epidemic when EHI prevalence was high (Figure 3B). After the epidemic peak, infections arising from EHI remains high proportional to EHI prevalence, consistent with the higher transmission rate we estimated for EHI. In the late 2000's, we estimated that between 40 to 50% of all new infections arise from EHI, indicating that early stage infections still play a major role in driving HIV transmission. These large estimates for number of new infections arising from EHI are consistent with the phylodynamic estimates of Volz *et al.* [34], who fit a more complex but deterministic epidemiological model to the same set of HIV sequences.

### Inferring population structure

While our results for the three-stage model suggest that the PMMH algorithm works effectively and can be used to estimate key epidemiological parameters like HIV transmission rates, we were also interested in how much information genealogies contain about the structure of populations in general. To explore this question, we used the two-population model presented in (3), for which we can tune the strength of population structure by altering the mixing rate  $\rho$  between populations. Mock genealogies were simulated under three values of  $\rho$ : low (0.01), medium (0.05) and high (0.2). At  $\rho=0.01$ , for example, about one in every one hundred transmission events occurs between populations. For all



**Figure 2. Posterior densities of parameters inferred from one HIV genealogy.** Solid red lines mark the median values and dashed lines indicate the 95% credible intervals. The estimated parameters are the early stage transmission rate  $\beta_E$ , the chronic stage transmission rate  $\beta_C$ , the AIDS stage transmission rate  $\beta_A$ , the incidence scaler  $\alpha$  and the initial introduction time of HIV into Detroit. doi:10.1371/journal.pcbi.1003570.g002

three values of  $\rho$ , we were able to accurately infer the epidemiological parameters of interest and the population dynamics from the simulated genealogies (Figure 4 & Figure S3). While we can easily estimate  $\beta$  under all three demographic scenarios, the posterior densities become skewed towards increasingly high values of  $\rho$  as mixing increases between the populations (Figure 4A–C). This indicates that it may be very difficult to obtain precise estimates of  $\rho$  or other parameters pertaining to population structure when populations are only weakly structured.

We can visually explore how much information a genealogy contains about population structure and pathogen movement by comparing the true lineage states to the computed lineage state probabilities. In Figure 5A–C, the true state of each lineage over time is mapped onto the genealogies. For ease of viewing, we only display a representative subtree of each genealogy. As expected, under low mixing lineages change states very slowly leading to a high degree of clustering among lineages sampled from the same population, whereas under high mixing lineages move rapidly between states and there is little clustering. We can then compare the true lineage states with the state probabilities computed under the median posterior values of the estimated parameters (Figure 5D–F). When  $\rho$  is low, the state of the lineages at the time of sampling is highly informative about the state of the lineage going into the past. However, when we increase  $\rho$  to 0.05, the state of the sampled lineages is less informative about the past states and we can see that the lineage state probabilities fluctuate seasonally according to the asynchronous dynamics between populations.

When  $\rho$  is high, the lineages move between states so rapidly that there is high uncertainty in the lineage states over the entire tree. This loss of information regarding the lineage states is readily observed by considering how the entropy, or uncertainty, in the lineage states changes going backwards in time (Figure 5G–H).

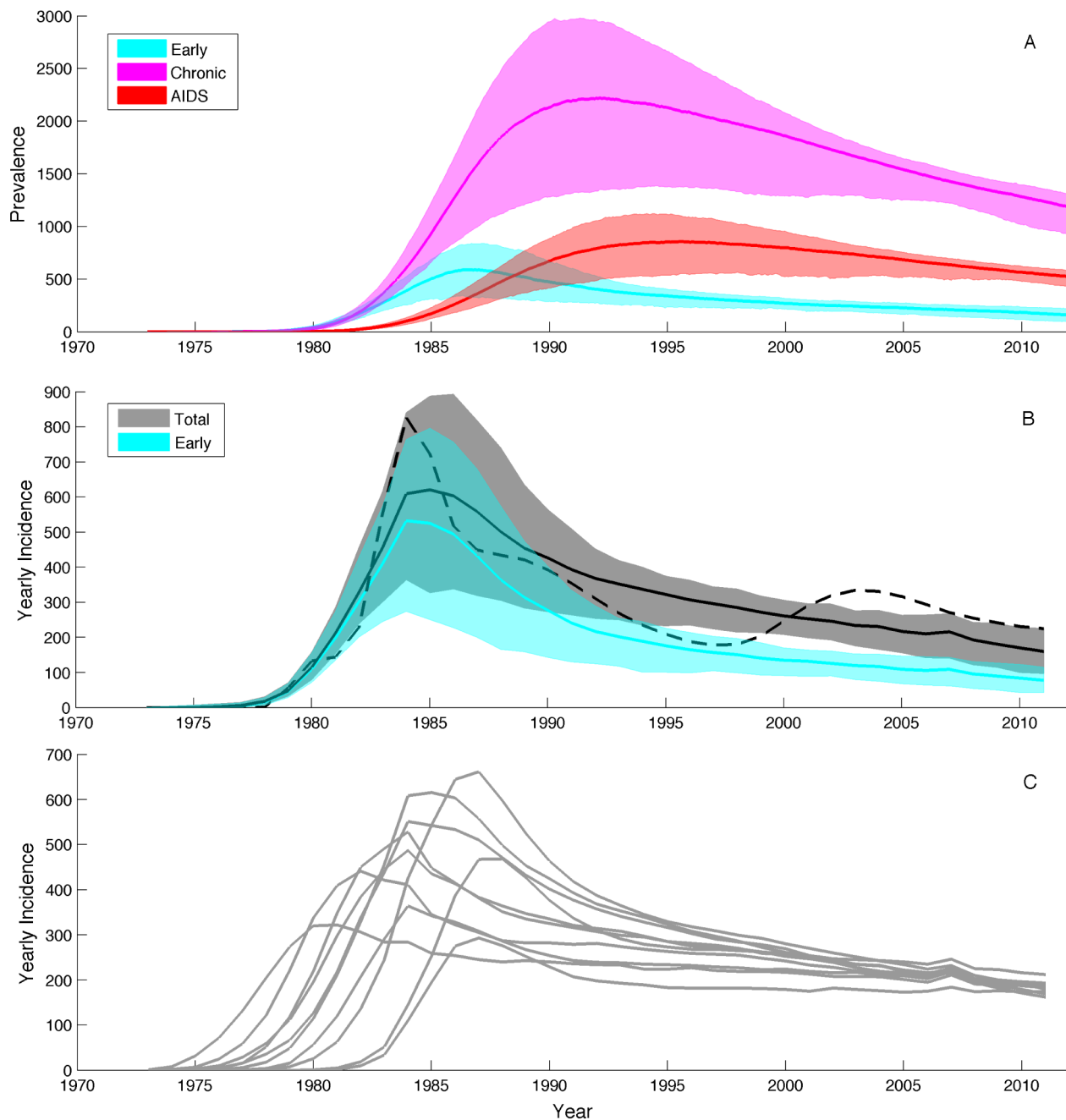
Visualizing the flow of information along the lineages in the trees shows how uncertainty in parameters like  $\rho$  depends on how rapidly information about the lineage states decays. When  $\rho$  is low, lineages remain in the same state long enough that once a coalescent event is reached, information about the probable state of the coalescing lineages provides additional information about the transmission event with respect to whether the transmission event occurred within or between populations. By combining information from coalescent events across the entire tree, we can then estimate the rates at which transmission occurs within and between populations. However, if all information about the past lineage states is lost before lineages coalesce, the observed coalescent events will no longer be informative about whether transmission occurred within or between populations and therefore parameters like  $\rho$  will be difficult to precisely estimate.

The preceding observations about uncertainty in lineage states suggest that it may be possible to estimate  $\rho$  more precisely if we increase the number of sampled lineages. Increasing the sampling fraction will also increase the coalescent rate among lineages, thereby increasing the probability of lineages coalescing before all information about their probable state is lost. To test this idea, we

**Table 2. HIV parameter estimates.**

Tree	$\beta_E$	$\beta_C$	$\beta_A$	$\alpha \times 10^{-4}$	Initial Time
1	1.59 (1.43–1.77)	0.077 (0.006–0.187)	0.328 (0.146–0.653)	1.52 (0.51–4.35)	1975 (1973–1977)
2	1.69 (1.51–1.9)	0.079 (0.007–0.184)	0.318 (0.126–0.59)	1.72 (0.62–4.02)	1976 (1973–1979)
3	1.69 (1.39–1.99)	0.184 (0.02–0.539)	0.593 (0.151–1.26)	5.78 (1.43–11.8)	1976 (1973–1978)
4	1.51 (1.32–1.73)	0.079 (0.007–0.191)	0.316 (0.114–0.648)	1.55 (0.48–4.44)	1974 (1973–1976)
5	1.6 (1.29–1.92)	0.146 (0.016–0.395)	0.786 (0.235–1.64)	5.99 (1.91–11.1)	1973 (1973–1974)
6	1.69 (1.43–1.99)	0.114 (0.008–0.28)	0.543 (0.193–1.0)	3.75 (1.23–8.0)	1974 (1973–1975)
7	1.67 (1.45–1.91)	0.1 (0.012–0.224)	0.431 (0.155–0.823)	2.93 (1.21–5.57)	1974 (1973–1977)
8	1.87 (1.62–2.14)	0.095 (0.006–0.239)	0.521 (0.214–0.964)	4.06 (1.22–8.3)	1979 (1974–1981)
9	1.9 (1.51–2.31)	0.295 (0.025–0.636)	0.823 (0.247–1.74)	10.2 (2.76–17.5)	1980 (1974–1982)
10	1.52 (1.32–1.73)	0.08 (0.009–0.248)	0.433 (0.184–1.29)	2.5 (0.73–10.5)	1974 (1973–1976)

Median posterior and 95% credible intervals for parameters estimated from 10 HIV genealogies. doi:10.1371/journal.pcbi.1003570.t002



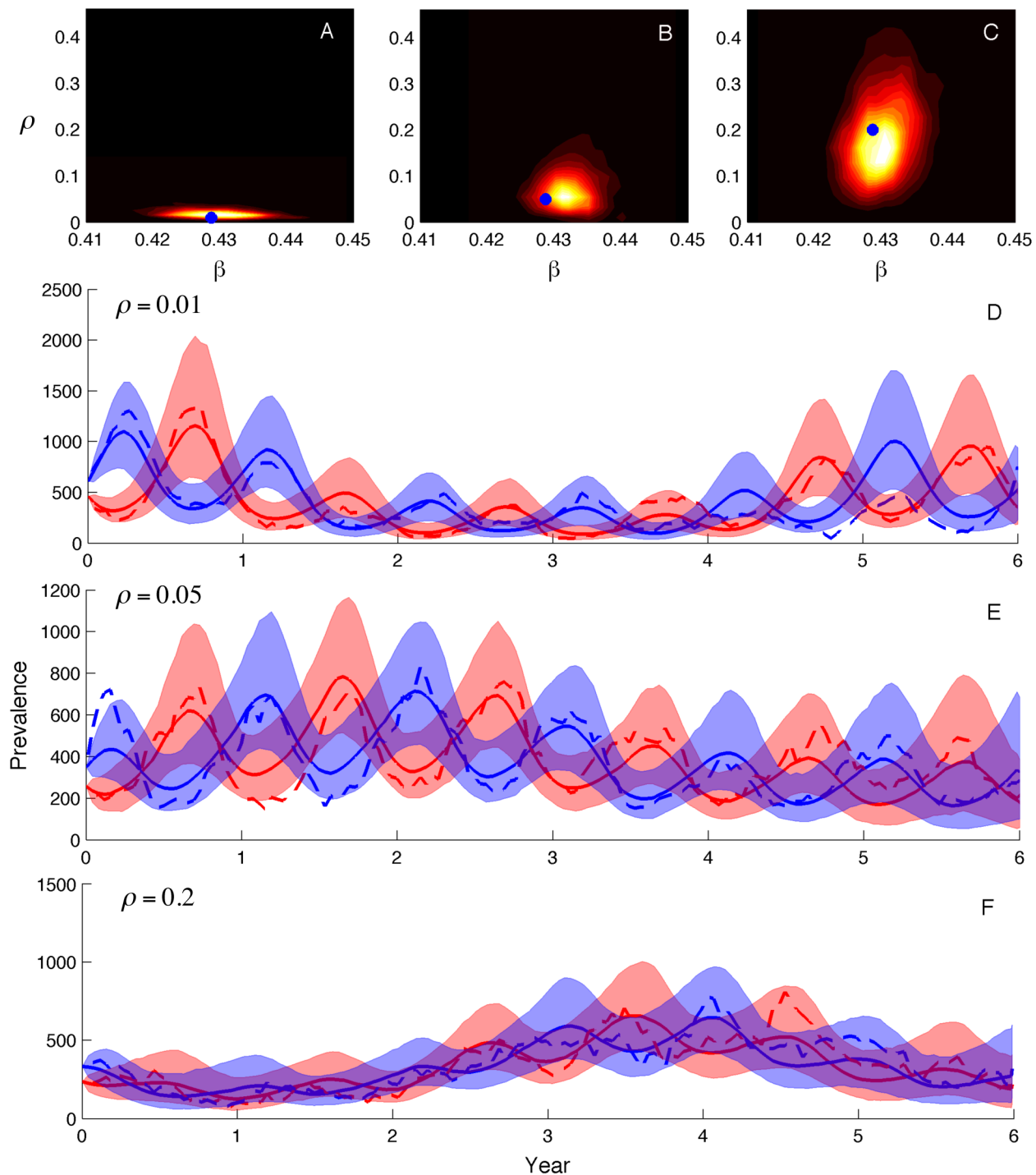
**Figure 3. Population dynamics inferred from the Detroit HIV genealogies.** (A) Stage-specific prevalence estimates from one genealogy with shaded regions showing the 95% credible intervals and lines the median of the posterior densities. (B) Estimated total yearly incidence and the estimated incidence attributable to early stage infections. The dashed black line shows the incidence back-calculated from Michigan Department of Community Health surveillance data. (C) Total incidence estimated from 10 randomly sampled HIV genealogies. doi:10.1371/journal.pcbi.1003570.g003

simulated genealogies under the same three values of  $\rho$  but varied the sample size. With a sample size of 100, the same as used above, we see that the likelihood is peaked around the true value of  $\rho$  when mixing is low but the likelihood profile is fairly flat when mixing is high (Figure 6A–C). Increasing the sample size to 500 resulted in more curved likelihood profiles but the likelihood remains relatively flat with high mixing (Figure 6D–F). Doubling the sample size again to 1,000, the likelihood profiles show significant curvature for all values of  $\rho$  (Figure 6G–I). This suggests that while the sample size does play a significant role in

determining whether parameters like  $\rho$  can be precisely estimated from genealogies, extremely large sample sizes may be required to estimate parameters pertaining to population structure when the population is only weakly structured.

## Discussion

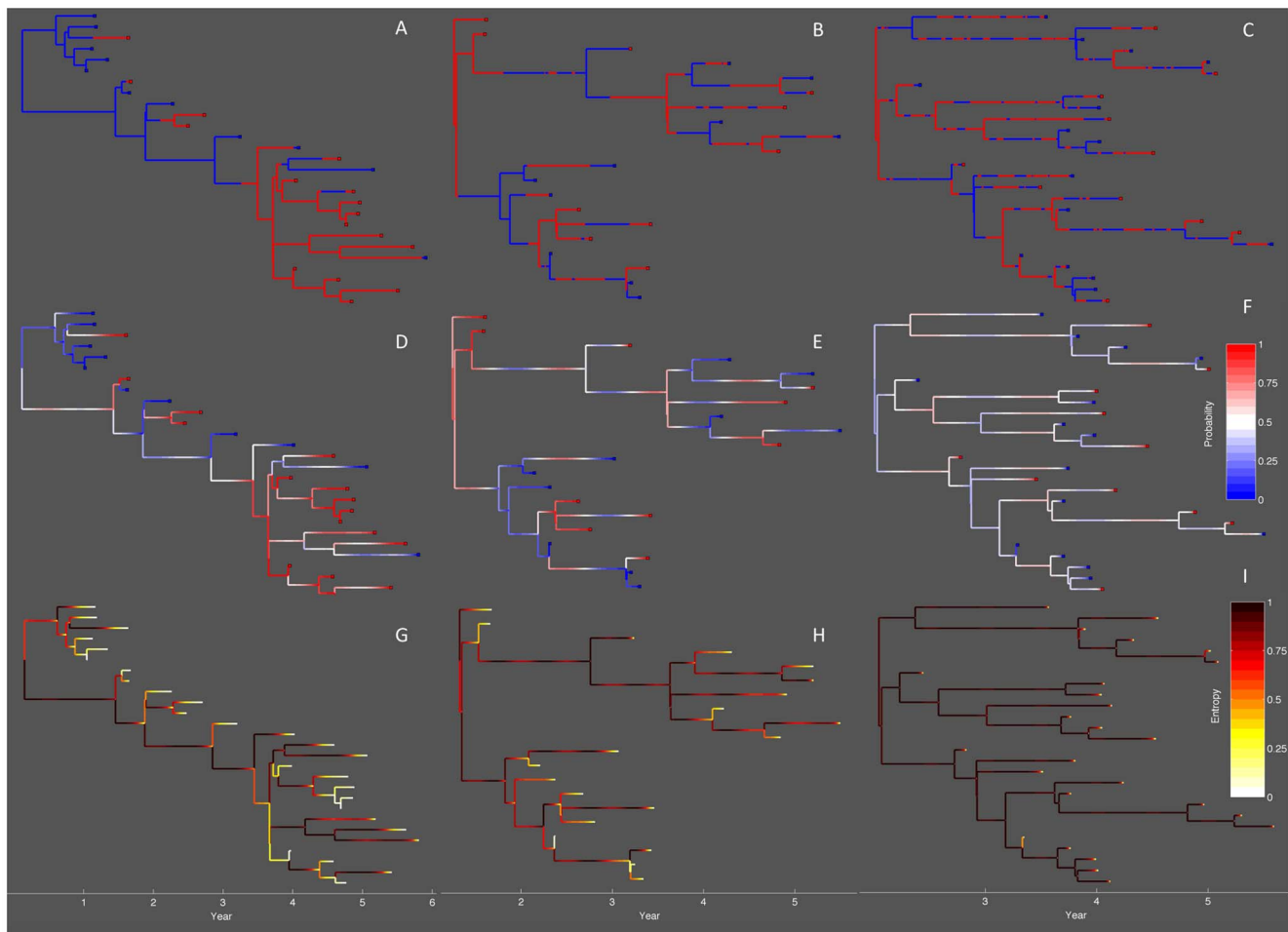
The approach outlined in this paper allows for structured, stochastic epidemiological and other population dynamic models to be fit to genealogies in order to jointly infer past population



**Figure 4. Parameter and prevalence estimates for the two-population model.** Mixing rates between the two populations were varied from low ( $\rho=0.01$ ), medium ( $\rho=0.05$ ) to high ( $\rho=0.2$ ). (A–C) Joint posterior densities for the transmission rate  $\beta$  and the mixing parameter  $\rho$ . (D–F) Prevalence estimates for the two populations with the 95% credible intervals shaded and the posterior medians shown as solid lines. Dashed lines show the true prevalence. Initial conditions for the number of susceptible and infected individuals in each population were fixed at their true values for these simulations.  
doi:10.1371/journal.pcbi.1003570.g004

dynamics and model parameters. We believe this to be an important step forward in the field of phylodynamics because many populations are structured in ways that could bias estimates of demographic parameters when using coalescent-based methods if population structure is not properly taken into account.

Furthermore, unlike earlier methods for fitting structured coalescent models to genealogies (e.g. [7,15]), our framework can accommodate non-equilibrium and nonlinear population dynamics and allows birth and migration rates to vary over time. We can also include stochasticity in our models when fitting them to data



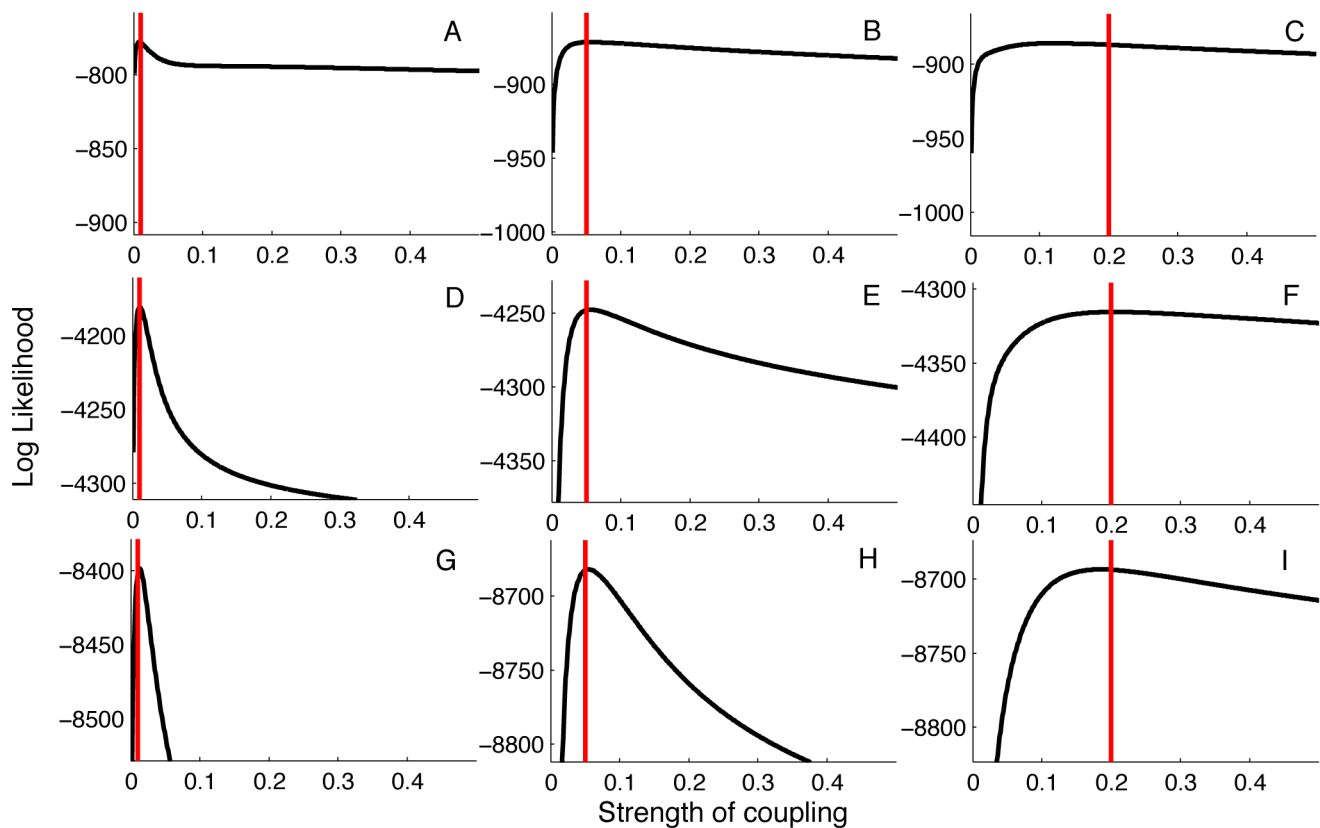
**Figure 5. Genealogies simulated under different mixing rates.** Mixing rates between the red and blue population were varied from low ( $\rho = 0.01$ ), medium ( $\rho = 0.05$ ) to high ( $\rho = 0.2$ ). (A–C) The true lineage states mapped onto the genealogy. (D–F) Lineage state probabilities given with respect to the probability that the lineage is in the red state. (G–I) Entropy in the lineage states, which shows how much uncertainty there is in the lineage states. For each lineage  $i$ , the entropy  $H_i = -\sum_k^m p_{ik} \log_2 \left( \frac{1}{p_{ik}} \right)$ .  
doi:10.1371/journal.pcbi.1003570.g005

obtained from real populations, which may behave very differently than what would be expected under deterministic models. We can therefore fit the type of mechanistic population dynamic models typically used by epidemiologists and ecologists, which often include population structure, to genealogies.

As we have shown, fitting stochastic population dynamic models to genealogies through a structured coalescent model poses some challenges to statistical inference not normally dealt with in the statistical literature on fitting generic state space models to observational data. Under our structured coalescent models, the probability of a genealogy depends conditionally on both the population state variables as well as the states of individual lineages over time. However, going backwards in time, the probability that a lineage is in a certain state can strongly depend on the state that the lineage was sampled in at some future point in time. Particle filtering methods, which are widely used to fit state space models to other sources of data, can perform very poorly under these circumstances because the state of the system, in this case the lineage states, can depend strongly on the future states of the system. One strategy we initially tried was therefore to use a Gibbs sampling approach to iteratively sample from the conditional posterior densities of the population state and lineage state

variables in independent steps to avoid the problem of having both forward and backward time dependencies in the model. Unfortunately, we found that such a Gibbs sampling strategy can be very inefficient and suffer from extremely poor MCMC mixing when there are strong correlations among the parameters and the lineage states. For example, in our two-population model, the mixing parameter  $\rho$  controls how rapidly lineages move between states and is thus highly correlated with the lineage states. If we update  $\rho$  conditional on our current lineage states, the proposed value of  $\rho$  will need to be very close to the current value in order for the proposal to have high enough probability to be accepted conditional on the current lineage states. We therefore explore a potentially very large parameter space taking only small steps at a time.

Given these issues, we decided to use a modified version of the PMMH algorithm originally proposed by Andrieu *et al.* [22]. In this approach, we simply propose new parameter values each MCMC iteration and then run the particle filter to numerically integrate over the population state variables. To make the particle filtering algorithm as efficient as possible within each MCMC step, we allow for resampling by first weighting the particles according to the expected lineage state probabilities. Once we have run the



**Figure 6. Likelihood profiles for the strength of coupling  $\rho$  obtained from genealogies simulated under different values of  $\rho$ .** Red lines correspond to the true value of  $\rho$ . The likelihoods were computed from genealogies with 100 samples in (A–C), 500 samples in (D–F) and 1000 samples in (G–I). These sample sizes correspond, respectively, to approximately 0.2%, 1.0% and 2% of all infected individuals being sampled. doi:10.1371/journal.pcbi.1003570.g006

particle filter forwards in time, we can then compute the true lineage state probabilities backwards in time and apply an additional round of importance sampling to correct for any bias introduced by using the expected lineage state probabilities. With the true lineage state probabilities of each particle, we can compute the coalescent likelihood of the genealogy while summing over all possible lineage states. We can therefore integrate over both the unobserved population state variables and the lineage state variables when computing the marginal likelihood of the parameter proposal. We thus have an efficient MCMC algorithm for sampling from the posterior density of the parameters without having to design independent proposals for the population states or the lineage states. The PMMH sampler therefore has a major practical advantage over other MCMC approaches that can be easily quantified. For the models considered in this paper, the PMMH algorithm typically converged in less than 100,000 iterations whereas for the Gibbs sampler we could run millions of MCMC iterations and still not converge. The efficiency of this approach will hopefully make it possible to also consider phylogenetic uncertainty in the future by sampling genealogies in addition to epidemiological parameters in the MCMC algorithm.

Whether or not the type of coalescent models considered here are appropriate for a particular pathogen is another important issue. The coalescent models assume that each infected host corresponds to a single pathogen lineage. If this were indeed always the case then coalescent events in the genealogy would always correspond to transmission events in the population. In reality, coalescent events will not occur instantaneously at

transmission events but at some time before the actual transmission event because there will be a waiting time between when a lineage is transmitted and when it coalesces with another sampled lineage in the host. How closely the actual transmission event corresponds in time with the coalescent event will likely depend on the within-host dynamics of the pathogen [42]. For chronic viral infections like HIV where multiple lineages can persist within a given host for months or years, this may result in a large discrepancy in the timing of transmission and coalescent events. Nevertheless, a simulation study using a realistic distribution of within-host coalescent times for HIV found that the difference in timing between coalescent and transmission events was not sufficient to bias estimates of epidemiological parameters [34]. This may be due to the fact that a large fraction of HIV transmissions are due to recently infected individuals, in which case the within-host coalescent event cannot have occurred very long before the actual transmission event. A more principled approach to pursue in the future may be to impute the actual times of transmission conditional on the time of the coalescent events using information about within-host population dynamics. For example, additional information about pathogen population sizes over the course of a typical infection could provide an informative prior on waiting times between transmission events and coalescent events within hosts.

Another possible violation of the coalescent model occurs if sampled individuals have descendants that are themselves sampled, which can occur when samples are collected serially over time. The coalescent model implicitly assumes that when a



new lineage is sampled, that lineage is sampled from a different host than any other lineage already in the genealogy. However, if a lineage is sampled from a host that has other sampled descendant lineages in the genealogy, then this results in a coalescent event in the tree that does not correspond to a transmission event in the population. A similar problem would arise if we unwittingly sampled more than one lineage from a single infected host. However this is likely to occur only if sampling is dense relative to prevalence over time. For example, if sampling is dense at the beginning and the end of an epidemic, then with a high probability hosts sampled at the beginning of the epidemic will likely have sampled descendants at the end of the epidemic. We acknowledge that the coalescent models used in this paper cannot adequately handle these types of situations, although for the HIV analysis it is unlikely that this is a serious problem seeing as all sequences were sampled in the recent past when prevalence was high. In cases where this is likely to be a serious problem, it may be worth developing metapopulation coalescent models, such as those introduced by Dearlove and Wilson [43], that allow hosts to be infected by more than a single lineage.

As our application to HIV showed, the PMMH algorithm allowed us to infer key epidemiological parameters like stage-specific transmission rates directly from genealogies. However, in the case of HIV, individuals stay in the same stage of infection for long periods of time relative to the timescale of the epidemic. The stage of infection of sampled individuals is therefore highly informative about the state of the lineage going into the past. Our experience with HIV may therefore not be representative of our general ability to infer parameters pertaining to pathogen transmission or movement in structured populations. In fact, our simple two-population SIR model revealed certain conditions under which it may be inherently difficult to estimate parameters relating to population structure. When lineages move between states rapidly due to transmission or migration any particular lineage is likely to have changed states multiple times before a coalescent event is reached, leading to high uncertainty about the state of lineage over the majority of the genealogy. This is somewhat analogous to the problem of site saturation in phylogenetic inference, where multiple transitions at a particular site along branches can render that site phylogenetically uninformative [44]. In the case of rapid transition rates among population states, observing the state of lineages at the time of sampling offers little or no information about the structure of the population because all information about the state of the lineage is quickly lost. Under these circumstances, it will be difficult to precisely estimate migration rates or other parameters relating to population structure from genealogies as we saw from the likelihood profiles of the mixing parameter in the two-population model, although it may be possible with many samples or a large sample fraction. This echoes earlier work on inference with structured coalescent models, where researchers have found it difficult to estimate migration rates from genealogies even without the complication of complex population dynamics [7,45].

Although it may not always be possible to precisely estimate parameters relating to population structure from genealogies, we can imagine several cases in which the ability to fit mechanistic epidemiological models to genealogies that include population structure may be extremely useful. For example, our methods could be used to fit spatially structured models to genealogies of samples collected in different locations and could potentially complement recently developed phylogeographic methods that consider spatial structure but do not generally take into account local population dynamics at any particular location [46,47]. For instance, incorporating both spatial and temporal dynamics could

be important when the structure of a population is not static but changes over time due to changes in migration rates, which themselves may vary due to non-stationary population dynamics across locations. Our approach can also be applied in cases where sampling effort is distributed unevenly among populations so that the assumption of random sampling in unstructured coalescent models has obviously been violated. In this case, structured coalescent models can be used to control for non-random sampling as long as sampling is random within the subpopulations defined in the coalescent model. Finally, our methods can be applied to multi-host or vectored pathogens where lineages can move among different host or vector species. As shown in Rasmussen *et al.* [48] for the case of dengue, including the dynamics of both the host and vector populations in coalescent models may be necessary in order for population dynamics inferred from genealogies of vector-borne pathogens to be accurate.

We end by noting that the methods presented here can be used to fit epidemiological models to genealogies as well as other sources of data simultaneously. For example, we previously showed how unstructured epidemiological models can be fit to a genealogy and a time series of case reports simultaneously and it would be straightforward to extend the methods presented here to include time series or other observational data [21]. This could be especially helpful when certain parameters or aspects of the dynamics are difficult to infer from one data source but for which an alternative data source could be highly informative. For example, case report data may be aggregated over different subpopulations obscuring some of the heterogeneity present in the population but could be revealed by also considering information present in a genealogy. Consolidating data sources in this way will likely play an important role in epidemiological modeling in the future, especially as molecular sequence data become increasingly available and phylogenetic methods become integrated into modern epidemiology.

## Supporting Information

**Figure S1 Simulated epidemic dynamics for 100 stochastic realizations of the three-stage SIR model.** Total prevalence includes all three stages of infection.  
(TIF)

**Figure S2 Representative time-scaled HIV genealogy from Detroit, Michigan.**  
(TIF)

**Figure S3 Marginal posterior densities of the parameters in the two-population model with low mixing ( $\rho=0.01$ ) between populations.** Blue lines show the true values. The estimated parameters are: (A) the transmission rate  $\beta$ ; (B) the mixing rate  $\rho$ ; (C) the amplitude of seasonal forcing  $\alpha$ ; (D–E) the seasonal phases for the two populations  $\delta_1$  and  $\delta_2$ . Similar estimates of the seasonality parameters were obtained with medium and high mixing between populations.  
(TIF)

**Text S1 Summary of the particle filtering algorithm for genealogies and details on how lineage state probabilities are computed.**  
(PDF)

## Author Contributions

Conceived and designed the experiments: DAR EMV KK. Performed the experiments: DAR. Analyzed the data: DAR. Contributed reagents/materials/analysis tools: DAR EMV. Wrote the paper: DAR EMV KK.



## References

1. Nee S, Holmes EC, Rambaut A, Harvey PH (1995) Inferring population history from molecular phylogenies. *Philos Trans R Soc Lond B Biol Sci* 349: 25–31.
2. Grenfell BT, Pybus OG, Gog JR, Wood JLN, Daly JM, et al. (2004) Unifying the epidemiological and evolutionary dynamics of pathogens. *Science* 303: 327–32.
3. Donnelly P, Tavaré S (1995) Coalescents and genealogical structure under neutrality. *Annu Rev Genet* 29: 401–21.
4. Lewis F, Hughes GJ, Rambaut A, Pozniak A, Leigh Brown AJ (2008) Episodic sexual transmission of HIV revealed by molecular phylodynamics. *PLoS Medicine* 5: e50
5. Kuhner M, Yamato J, Felsenstein J (1998) Maximum likelihood estimation of population growth rates based on the coalescent. *Genetics* 149: 429–34.
6. Pybus O, Rambaut A, Harvey P (2000) An integrated framework for the inference of viral population history from reconstructed genealogies. *Genetics* 155: 1429–37.
7. Beerli P, Felsenstein J (2001) Maximum likelihood estimation of a migration matrix and effective population sizes in *n* subpopulations by using a coalescent approach. *Proc Natl Acad Sci USA* 98: 4563–8.
8. Stadler T, Kouyos R, von Wyl V, Yerly S, Böni J, et al. (2012) Estimating the basic reproductive number from viral sequence data. *Mol Biol Evol* 29: 347–357.
9. Kingman J (1982) On the genealogy of large populations. *Journal of Applied Probability* 19: 27–43.
10. Wakeley J (2009) *Coalescent Theory: An Introduction*. Greenwood Village, Colo.: Roberts and Company Publishers.
11. Strimmer K, Pybus OG (2001) Exploring the demographic history of dna sequences using the generalized skyline plot. *Mol Biol Evol* 18: 2298–305.
12. Drummond AJ, Rambaut A, Shapiro B, Pybus OG (2005) Bayesian coalescent inference of past population dynamics from molecular sequences. *Mol Biol Evol* 22: 1185–92.
13. Notohara M (1990) The coalescent and the genealogical process in geographically structured population. *J Math Biol* 29: 59–75.
14. Takahata N, Slatkin M (1990) Genealogy of neutral genes in two partially isolated populations. *Theor Popul Biol* 38: 331–50.
15. Kuhner M (2006) LAMARC 2.0: Maximum likelihood and Bayesian estimation of population parameters. *Bioinformatics* 22: 768–70.
16. Volz EM, Kosakovsky Pond SL, Ward MJ, Leigh Brown AJ, Frost SDW (2009) Phylodynamics of infectious disease epidemics. *Genetics* 183: 1421–30.
17. Frost S, Volz E (2010) Viral phylodynamics and the search for an ‘effective number of infections’. *Philos Trans R Soc Lond B Biol Sci* 365: 1879–90.
18. Anderson RM, May RM (1991) *Infectious Diseases of Humans: Dynamics and Control*. Oxford, U.K.: Oxford University Press.
19. Koelle K, Rasmussen D (2012) Rates of coalescence for common epidemiological models at equilibrium. *J R Soc Interface* 9: 997–1007.
20. Volz E (2012) Complex population dynamics and the coalescent under neutrality. *Genetics* 190: 187–201.
21. Rasmussen D, Ratmann O, Koelle K (2011) Inference for nonlinear epidemiological models using genealogies and time series. *PLoS Comput Biol* 7: e1002136
22. Andrieu C, Doucet A, Holenstein R (2010) Particle Markov chain Monte Carlo methods. *Journal of the Royal Statistical Society: Series B* 72: 269–342.
23. Keeling M, Rohani P (2008) *Modeling infectious diseases in humans and animals*. Princeton: Princeton University Press.
24. Williams BG, Lloyd-Smith JO, Gouws E, Hankins C, Getz WM, et al. (2006) The potential impact of male circumcision on HIV in sub-Saharan Africa. *PLoS Medicine* 3: e262.
25. Granich RM, Gilks CF, Dye C, De Cock KM, Williams BG (2009) Universal voluntary HIV testing with immediate antiretroviral therapy as a strategy for elimination of HIV transmission: a mathematical model. *The Lancet* 373: 48–57.
26. Volz EM, Koopman JS, Ward MJ, Brown AL, Frost SD (2012) Simple epidemiological dynamics explain phylogenetic clustering of HIV from patients with recent infection. *PLoS Comput Biol* 8: e1002552.
27. Keeling M, Rohani P (2002) Estimating spatial coupling in epidemiological systems: a mechanistic approach. *Ecology Letters* 5: 20–29.
28. Andrieu C, Roberts GO (2009) The pseudo-marginal approach for efficient Monte Carlo computations. *The Annals of Statistics* 37: 697–725.
29. Doucet A, De Freitas N, Gordon N, et al. (2001) *Sequential Monte Carlo methods in practice*, volume 1. Springer New York.
30. Cappe O, Godsill S, Moulines E (2007) An overview of existing methods and recent advances in sequential Monte Carlo. *Proc Inst Electr Elect* 95: 899–924.
31. Doucet A, Johansen A (2009) A tutorial on particle filtering and smoothing: Fifteen years later. *Handbook of Nonlinear Filtering*: 656–704.
32. Gillespie DT (2007) Stochastic simulation of chemical kinetics. *Annu Rev Phys Chem* 58: 35–55.
33. Drummond AJ, Rambaut A (2007) BEAST: Bayesian evolutionary analysis by sampling trees. *BMC Evolutionary Biology* 7: 214.
34. Volz EM, Ionides E, Romero-Severson EO, Brandt MG, Mokotoff E, et al. (2013) HIV-1 Transmission during Early Infection in Men Who Have Sex with Men: A Phylodynamic Analysis. *PLoS Medicine* 10: e1001568.
35. Yan P, Zhang F, Wand H (2011) Using HIV diagnostic data to estimate HIV incidence: method and simulation. *Statistical Communications in Infectious Diseases* 3: 1–28
36. Cohen MS, Dye C, Fraser C, Miller WC, Powers KA, et al. (2012) HIV treatment as prevention: debate and commentary—will early infection compromise treatment-as-prevention strategies? *PLoS medicine* 9: e1001232.
37. Kretzschmar ME, van der Loeff MFS, Birrell PJ, De Angelis D, Coutinho RA (2013) Prospects of elimination of HIV with test-and-treat strategy. *Proceedings of the National Academy of Sciences* 110: 15538–15543.
38. Brown AJL, Lycett SJ, Weinert L, Hughes GJ, Fearnhill E, et al. (2011) Transmission network parameters estimated from HIV sequences for a nationwide epidemic. *Journal of Infectious Diseases* 204: 1463–1469.
39. Pilcher CD, Tien HC, Eron JJ, Vernazza PL, Leu SY, et al. (2004) Brief but efficient: acute HIV infection and the sexual transmission of HIV. *Journal of Infectious Diseases* 189: 1785–1792.
40. Hollingsworth TD, Anderson RM, Fraser C (2008) HIV-1 transmission, by stage of infection. *Journal of Infectious Diseases* 198: 687–693.
41. Powers KA, Ghani AC, Miller WC, Hoffman IF, Pettifor AE, et al. (2011) The role of acute and early HIV infection in the spread of HIV and implications for transmission prevention strategies in Lilongwe, Malawi: a modelling study. *The Lancet* 378: 256–268.
42. Ypma RJ, van Ballegoijen WM, Wallinga J (2013) Relating phylogenetic trees to transmission trees of infectious disease outbreaks. *Genetics* 195: 1055–1062.
43. Dearlove B, Wilson DJ (2013) Coalescent inference for infectious disease: meta-analysis of hepatitis C. *Philos Trans R Soc Lond B Biol Sci* 368: 1614
44. Yang Z (1998) On the best evolutionary rate for phylogenetic analysis. *Syst Biol* 47: 125–33.
45. Beerli P, Felsenstein J (1999) Maximum-likelihood estimation of migration rates and effective population numbers in two populations using a coalescent approach. *Genetics* 152: 763–73.
46. Lemey P, Rambaut A, Drummond AJ, Suchard MA (2009) Bayesian phylogeography finds its roots. *PLoS Comput Biol* 5: e1000520.
47. Pybus OG, Suchard MA, Lemey P, Bernardin FJ, Rambaut A, et al. (2012) Unifying the spatial epidemiology and molecular evolution of emerging epidemics. *Proc Natl Acad Sci USA* 109: 15066–71.
48. Rasmussen D, Boni M, Koelle K (2014) Reconciling epidemiology with phylogeny: The case of dengue virus in southern Vietnam. *Mol Biol Evol* 31: 258–271.