Directly estimating epidemic curves from genomic data

Timothy G. Vaughan[1,*,‡], Gabriel E. Leventhal[2,3,*,†], David A. Rasmussen[4], Alexei J. Drummond[1], David Welch[1], Tanja Stadler[4,5]

(1) Centre for Computational Evolution, University of Auckland, Auckland, New Zealand
(2) Institute of Integrative Biology, ETH Zürich, Zurich, Switzerland
(3) Department of Civil and Environmental Engineering, Massachusetts Institute of Technology (MIT), Cambridge, MA, U.S.A.
(4) Department of Biosystems Science and Engineering, ETH Zürich, Basel, Switzerland
(5) Swiss Institute of Bioinformatics (SIB), Lausanne, Switzerland

*These authors contributed equally.
†ge@leventhal.ch; ‡tgvaughan@gmail.com

Keywords: phylodynamics, particle filter, epidemiology, Bayesian phylogenetics

Abstract

Modern phylodynamic methods interpret an inferred phylogenetic tree as a partial transmission chain providing information about the dynamic process of transmission and removal (where removal may be due to recovery, death or behaviour change). Birth-death and coalescent processes have been introduced to model the stochastic dynamics of epidemic spread under common epidemiological models such as the SIS and SIR models, and are successfully used to infer phylogenetic trees together with transmission (birth) and removal (death) rates. These methods integrate analytically over past incidence and prevalence to infer rate parameters, and thus cannot explicitly infer past incidence or prevalence. Here we introduce a particle filtering framework to explicitly infer prevalence and incidence trajectories along with phylogenies and epidemiological model parameters from genomic sequences under the birth-death model. After demonstrating the accuracy of this method on simulated data, we use it to assess the prevalence through time of the early 2014 Ebola outbreak in Sierra Leone.

1 Introduction

The main goal of epidemiological inference for infectious diseases is to estimate the number of current and new cases through time. When it is not possible to directly estimate such prevalence and incidence curves, it can still be possible to estimate the parameters of epidemiological compartmental models, such as SIR-like models [1]. These parametric compartmental models can then be used to calculate fundamental quantities like the basic reproductive number, R_0, or to simulate prevalence and incidence curves that recapitulate the true epidemic.

In recent years, several statistical methods have been developed for epidemiological inference from genomic data. These methods lie at the intersection of statistical phylogenetics and epidemiology, and exploit the rapid evolution of many pathogens that occurs on the same time-scale as their epidemiological spread. In these cases, pathogens are said to be "measurably evolving" [2] and the use of phylogenetics in this context is termed "phylodynamics" [3].

Early phylodynamic methods only indirectly infer incidence and prevalence or epidemiological parameters. One of the first phylodynamic investigations [4] used the mathematical relationship between the effective population size and the time between coalescent events in phylogenetic trees [5] to produce non-parametric estimates of Hepatitis C virus prevalence [6]. These non-parametric "skyline plots" were then fitted to a parametric epidemiological model to estimate the basic reproduction rate, R_0 [6]. A subsequent approach combined the estimation of the viral phylogeny and the effective viral population size through time into a joint Bayesian method [7], but still lacked an explicit model of the epidemiological process. Another variant of the skyline plot based on the birth-death process [8] allowed for piecewise-constant variation in the birth and death rates [9] from which R_0 could be derived.

An important limitation of all of these approaches is that they do not directly include the epidemiological model in the phylogenetic inference method. Rather, the epidemiological parameters are themselves estimated from estimates of the phylogeny and viral effective population size which are based on genomic

data. This multi-step approach obscures the propagation of the overall uncertainty from one method to the next.

There have recently been three distinct approaches to incorporate compartmental models into phylodynamic inference. First, Volz *et al.* [10, 11] showed how to derive prior probability distributions for viral gene trees in the coalescent limit from arbitrary birth-death processes. This method permits joint Bayesian inference of epidemic model parameters, prevalence curves and phylogenetic trees. The coalescent basis of this method requires epidemic curves to either be deterministic, or stochastic as long as the epidemic events are statistically independent from the events that make up the sampled epidemic transmission tree [12]. Either assumption is justified in the case of a large populations size (prevalence). But when prevalence is low, the coalescent method can lead to biases in the estimates of the phylogenetic tree and the epidemiological parameters [13].

Second, Kühnert *et al.* [14] used a parametric compartmental model — specifically, a stochastic SIR model — to produce the piecewise-constant rates of the birth-death skyline plot. Like the coalescent methods of Volz *et al.* [10, 11], this enables joint inference of epidemiological parameters, epidemic curves and phylogeny. The stochastic formulation of the epidemiological process does not rest on the assumption of large population sizes but, like the coalescent methods, the tree events and the epidemic events are assumed to be statistically independent.

Third, Leventhal *et al.* [15] presented the first inference approach to employ an approximation-free computation of the phylogenetic tree probability under a stochastic epidemiological model. The method involves a tailored numerical algorithm to integrate the master equations of stochastic epidemiological process that is conditioned on the phylogenetic tree. While this approach can be extended to full joint inference of epidemic model parameters and the phylogeny, the current implementation assumes a known phylogeny and integrates over all possible prevalence curves to infer epidemic model parameters.

In this paper, we introduce a new method that uses the Particle Marginal Metropolis-Hastings algorithm [16] to jointly infer prevalence and incidence curves, phylogenetic trees, and epidemiological parameters under stochastic epidemiological models. Our approach addresses several of the short-comings of previous methods: (i) it accounts for the dependence of epidemic and tree events; (ii) it incorporates stochastic models of epidemic dynamics; (iii) it includes "sampled ancestors" to account for the possibility that some sampled viruses are directly "ancestral" to others on the transmission tree that relates them and (iv) it provides a natural route to the inclusion of additional (non-genetic) incidence data in full joint phylodynamic analyses. While particle filtering approaches have been previously applied to phylodynamic inference [12, 17, 18], they have only been used in the diffusion limit where the discrete nature of the compartment occupancies is ignored. In contrast, our particle filter is used to compute the exact probability of a transmission tree under the full stochastic compartmental model. This distinction is important near the start of epidemics where prevalence is low and diffusion or coalescent limits do not hold [13].

2 Methods

In this section we derive a flexible and exact inference method for unstructured stochastic compartmental models.

2.1 Stochastic compartmental epidemic models

Compartmental models are the centrepiece of epidemiological modeling. They partition individuals in a population into compartments according to their infection status, and allow for individuals to transition between some of the compartments. In an SIS model, individuals are either susceptible (S) or infectious (I). Susceptible individuals move to the infectious compartment upon infection, and infectious individuals move back to the susceptible compartment upon recovery. The SIR is similar to the SIS model, except that infectious individuals do not move back tot the susceptible compartment, but are modelled as removed (R), due to for example quarantine, recovery with immunity, or death. Let $S(t)$, $I(t)$ and $R(t)$ (or the relevant set for a given model) represent the number of individuals in the respective compartments at time t, and define $\mathcal{A}(t) = (S(t), I(t), R(t))$ to be the state of the epidemic at time time.

In this paper, we only consider unstructured compartmental models, i.e., models in which there is only one class of infected individual. This rules out (i) models that include an exposed compartment, often called E, where an individual can be infected but not yet infectious (such as SEIR and SEIS), and (ii) structuring via space, age or other factors of the infectious compartment. The reason for this restriction is that lineages of the transmission tree we discuss below are homogeneous and structuring requires labelling of the lineages to indicate the compartment each part of the lineage occupies.

The transitions of individuals between compartments can be described by a continuous-time Markov process on the state vector $\mathcal{A}$, as defined by the master equation

$$\frac{d}{dt}P(\mathcal{A}(t)|\mathcal{A}(0)) = \sum_{q\in Q}\left[\alpha_q(\mathcal{A}(t)-\boldsymbol{v}_q)P(\mathcal{A}(t)-\boldsymbol{v}_q|\mathcal{A}(0)) - \alpha_q(\mathcal{A}(t))P(\mathcal{A}(t)|\mathcal{A}(0))\right], \tag{1}$$

where $\alpha_q(\mathcal{A}(t))$ is the overall rate at which the epidemiological event of type q occurs, and $\boldsymbol{v}_q$ is the effect of event type q on the state: $\mathcal{A}\to\mathcal{A}+\boldsymbol{v}_q$.

The overall model is defined by the set of compartments combined with the set, $\mathcal{Q}$, of epidemic event types together with their corresponding rates. This formulation encompasses a broach range of epidemiological models. For instance, a linear birth-death model consists of just one compartment: $I(t)$, the number infected at time t. Possible events are infections and removals, so $\mathcal{Q}=\{\text{Infection},\text{Removal}\}$. The infection event produces a single new infection as described by $\boldsymbol{v}_{\text{Inf}}=+1$, and the overall infection rate is $\alpha_{\text{Inf}}(\mathcal{A}(t))=\beta I(t)$. Here, β is a constant describing the rate at which infectious individuals infect others. Similarly, the removal event removes an individual from the infectious compartment ($\boldsymbol{v}_{\text{Rem}}=-1$) at overall removal rate $\alpha_{\text{Rem}}(\mathcal{A}(t))=\delta$.

The SIR model has the same event type set as the linear birth-death process, $\mathcal{Q}=\{\text{Infection},\text{Removal}\}$, but different rate functions and event effects. An infection has effect vector $\boldsymbol{v}_{\text{Inf}}=(-1,1,0)$ and occurs at rate $\alpha_{\text{Inf}}[\mathcal{A}(t)]=\beta S(t)I(t)$, while a removal event has an effect vector $\boldsymbol{v}_{\text{Rem}}=(0,-1,+1)$ and occurs at rate $\alpha_{\text{Rem}}(\mathcal{A}(t))=\delta$. The SIS model is similar to the SIR model, only with effect vectors $\boldsymbol{v}_{\text{Inf}}=(-1,1)$ and $\boldsymbol{v}_{\text{Rem}}=(1,-1)$.

A specific model realization of an epidemic forward in time can be generated as follows: The epidemic starts at at time $t_0=0$ with a compartment occupancy distribution $\mathcal{A}(0)$. Typically, $I(0)=1$ for the infectious compartment, but other choices are possible. This initial distribution is changed by a series of events with types $e_1,e_2,\ldots e_s$ and times $t_1,t_2,\ldots t_s$, where s is a random variable indicating the number of events which occurred before some pre-determined stopping time T. The number of the individuals in each compartment after the i-th event has occurred at time t_i is denoted by $\mathcal{A}_i=\mathcal{A}(t_i)$. The population trajectory of the epidemic is then given by $(\mathcal{A}_0,\mathcal{A}_1,\ldots,\mathcal{A}_s)$. Figure 1a shows an example of the infectious compartment occupancy over time.

We can then equivalently expand $\mathcal{A}_i$ as a sum of effect vectors:

$$\mathcal{A}_i=\mathcal{A}_0+\boldsymbol{v}_{e_1}+\cdots+\boldsymbol{v}_{e_i}=\mathcal{A}_0+\left(\sum_{k=1}^{i}\boldsymbol{v}_{e_k}\right).$$

An epidemiological trajectory $\mathcal{E}$ is thus well defined by the initial state, $\mathcal{A}_0$, the vector of transition events $\mathbf{e}=(e_1,e_2,\ldots,e_s)$, and the corresponding event times, $\boldsymbol{t}=(t_1,t_2,\ldots,t_s)$,

$$\mathcal{E}=\{\mathcal{A}_0,\mathbf{e},\boldsymbol{t}\}. \tag{2}$$

As for any time-homogeneous discrete state continuous time Markov process, the probability density of a particular realization (trajectory) is a product of exponential distributions for the waiting times between the s events together with factors representing the probability density of each of given event times. That is,

$$P(\mathcal{E}|\boldsymbol{\eta},T)=\prod_{i=1}^{s}\exp[-\bar{\alpha}_i(t_i-t_{i-1})]\alpha_{e_i}(A_i), \tag{3}$$

where $\bar{\alpha}_i=\sum_{q\in\mathcal{Q}}\alpha_q(\mathcal{A}_i)$ is the sum of the rates of all possible transitions in interval i.

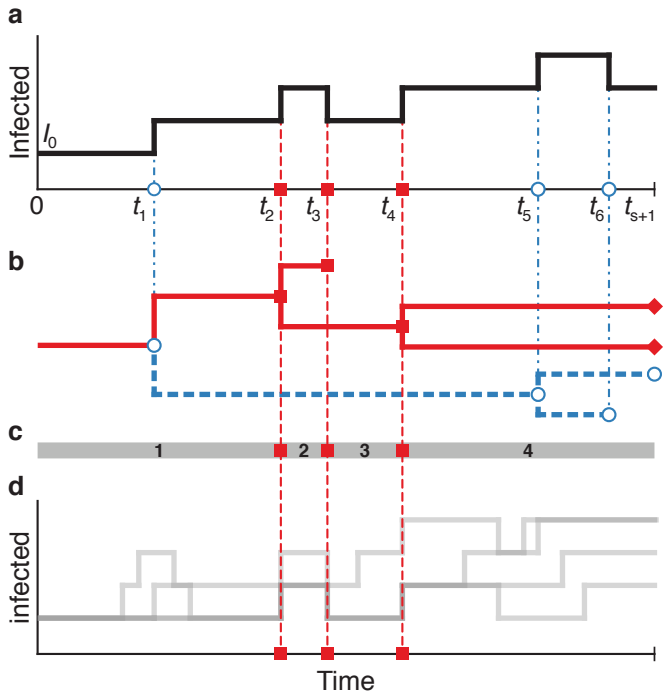

Figure 1: Relationships between (a) an epidemic trajectory, (b) the full and sampled transmission trees, (c) the sampled transmission tree coalescent intervals and (d) other epidemic trajectories compatible with the sampled tree. Note that the sampled transmission tree contains only a subset of the events represented by the full tree and trajectory, and each of these "observed" events must be present in any compatible trajectory.

In the case of an SIS model, new infections happen at a rate $\beta S_i I_i$ and infected individuals are removed at a rate γI_i. Let $\mathcal{I}_1 \subset \mathcal{I} = \{1, \ldots, s\}$ be the indices of infection events, and $\mathcal{I}_2 \subset \mathcal{I}$ the indices of the removal events. Then the probability density can be written,

$$P_{\mathrm{SIS}}(\mathcal{E}|\boldsymbol{\eta}, T) = \prod_{i\in\mathcal{I}} e^{-(\beta S_i I_i + \gamma I_i)(t_i - t_{i-1})} \prod_{i\in\mathcal{I}_1} \beta S_{i-1} I_{i-1} \prod_{i\in\mathcal{I}_2} \gamma I_{i-1}. \tag{4}$$

2.2 Modelling the sampling process

Sampling of individuals can be described by expanding $\mathcal{Q}$ to include two additional event types, sampling with and without removal. While the particular form of the effect vectors depend on the dimension of the compartmental model, their general result remains the same: $\boldsymbol{v}_{\mathrm{SampR}}$ removes an individual from the infectious class, while $\boldsymbol{v}_{\mathrm{SampNR}}$ leaves $\mathcal{A}(t)$ unchanged. We explicitly model the number and timing of sampling events by augmenting the stochastic process with sampling events included.

To model the timing of these events we use the following rate functions: $\alpha_{\mathrm{SampR}}(\mathcal{A}(t)) = r\psi I(t)$ and $\alpha_{\mathrm{SampNR}}(\mathcal{A}(t)) = (1-r)\psi I(t)$, where ψ is the per-individual sampling rate parameter and r is the probability of removal following sampling. Additionally, we allow for lineages extant at time T with probability ρ per lineage, yielding a binomial distribution of contemporaneous sampled leaf nodes. For convenience, we group these rate parameters together in the vector $\boldsymbol{\sigma} = (\psi, r, \rho)$.

We use $P(\mathcal{E}|\eta, \sigma, T)$ to represent the probability density of trajectories generated by this combined process.

2.3 From epidemiological trajectories to transmission trees

By tracking the identity of who infected whom in the stochastic epidemiological model, we obtain the transmission tree of the epidemic (full tree in Figure 1b). All transitions e_i in the trajectory correspond to either branching events or tips in the full phylogeny (circles and squares in Figure 1).

If we consider only the subtree ancestral to sampling events and pruning all other lineages from the transmission chain, we obtain the *sampled phylogeny* $\mathcal{T}$ (represented by the red subtree in figure 1b). Only a subset of the transition events are included in the sampled phylogeny, $\mathcal{K} \subset \mathcal{I}$ (red squares in the figure). The probability density $P(\mathcal{T}|\mathcal{E})$ of the sampled phylogeny given the full epidemic trajectory is independent of the specific epidemiological model, and is a straight-forward product of combinatorial factors which depend on whether an epidemic event occurs on the sampled transmission tree or not.

For example, using the notation from Eq. (4), the probability density for unstructured models such as the constant rate birth-death process, the SIR or SIS model, the probability of a labeled tree conditioned on the trajectory is,

$$P(\mathcal{T}|\mathcal{E}) = \prod_{i \in \mathcal{I}_1 \setminus \mathcal{K}} \frac{\binom{I_i}{2} - \binom{k_i}{2}}{\binom{I_i}{2}} \prod_{i \in \mathcal{I}_1 \cap \mathcal{K}} \frac{1}{\binom{I_i}{2}}. \tag{5}$$

2.4 Bayesian inference

Our primary goal is to perform asymptotically exact Bayesian inference of the epidemiological parameters using a set of aligned pathogen genetic sequences sampled throughout an epidemic. Our secondary goal is to simultaneously infer the prevalence and incidence trajectory under the same model.

To this end, for a given pathogen sequence alignment D, we use Bayes' rule to express the joint posterior distribution for the model parameters and the epidemic trajectories in terms of the conditional distributions composing the full model:

$$P(\mathcal{E}, \boldsymbol{\eta}, \boldsymbol{\sigma}, T|D) = \frac{1}{Z(D)} \int_{\mathcal{T}, \boldsymbol{\mu}} P(D|\mathcal{T}, \boldsymbol{\mu}) P(\mathcal{T}|\mathcal{E}) P(\mathcal{E}|\boldsymbol{\eta}, \boldsymbol{\sigma}, T) P(\boldsymbol{\mu}, \boldsymbol{\eta}, \boldsymbol{\sigma}, T). \tag{6}$$

Here $Z(D)$ is a normalization constant, and $P(D|\mathcal{T}, \boldsymbol{\mu})$ is the probability of D evolving down the sampled transmission tree $\mathcal{T}$ under a substitution model parameterized by $\boldsymbol{\mu}$, also known as the *tree likelihood*. $P(\boldsymbol{\mu}, \boldsymbol{\eta}, \boldsymbol{\sigma}, T)$ represents the joint prior probability distribution for the model parameters.

Several approaches to characterizing this posterior for particular models already exist in the literature, all of which involve using Markov chain Monte Carlo (MCMC) to sample (or maximum likelihood to optimize) a marginalized and/or approximate form of Eq. (6). For instance, Stadler [8] and Stadler & Bonhoeffer [19] analytically marginalize over the trajectory sub-space in the case of the linear birth-death model and use MCMC to sample from $(\mathcal{T}, \boldsymbol{\eta}, \boldsymbol{\sigma}, T)$. Similarly, Leventhal et al. [15] express the marginalization of Eq. (6) for the nonlinear stochastic SIS model as the solution to a master equation which is then integrated numerically; with parameter inferences being drawn by applying MCMC or ML.

Kühnert et al. [14] provide a approximation to the posterior for discretized trajectories $\mathcal{E}^*$ under the SIR model and use MCMC to sample $(\mathcal{E}^*, \mathcal{T}, \boldsymbol{\eta}, \boldsymbol{\sigma}, T)$. Finally, Volz et al. [10] and Volz [11] present an approximation to this posterior under the assumption that the relative amplitude of the demographic noise in $\mathcal{E}$ is negligible and that $P(\mathcal{E}|\boldsymbol{\eta}, \boldsymbol{\sigma}, T)$ therefore collapses to a point mass centred on the approximate deterministic solution of the model.

In contrast to these methods, we use the particle marginal Metropolis-Hastings (PMMH) algorithm [16]. This has previously been applied in a phylodynamic context by Rasmussen, Ratmann & Koelle [12] and Rasmussen, Volz & Koelle [17] using a diffusion-limit approximation to the stochastic epidemiological dynamics, but not to sample directly from the full phylodynamic posterior as we do in the algorithm described below.

2.5 Particle filtering algorithm

The PMMH algorithm is closely related to the exact pseudo-marginal algorithm [20]. In the form presented here, it involves using a bootstrap particle filter to simulate trajectories $\mathcal{E}$ conditional on a sampled transmission tree $\mathcal{T}$.

We divide the tree into intervals bounding each of its nodes, which are indexed as shown in Fig. 1c. The first interval begins at time T prior to the most recent sample. We refer to the portion of the tree within interval i as the *partial tree* $\mathcal{T}_i$, which is understood to include both the k_i lineages within the interval

and the tree node immediately to the right of the interval. Similarly, we divide the full trajectory $\mathcal{E}$ into corresponding partial trajectories $\mathcal{E}_i$.

The algorithm involves simulating an ensemble of m trajectories or "particles" in each of the intervals between t_0 and $t_M = T$. The initial condition for each particle is sampled from the ensemble of finishing states of particles simulated in the previous interval, weighted according to the probability of the tree event that divides the intervals. Each interval are sampled from the simulated end states of the previous interval.

The algorithm is as follows:

1. Set the interval $i \leftarrow 1$.
2. For each $j \in [1 \dots m]$ we use Gillespie's stochastic simulation algorithm [21, 22] or its asymptotically exact equivalent [23] to sample a partial trajectory $\mathcal{E}_i^{(j)}$ from $P(\mathcal{E}_i|\boldsymbol{\eta}, \boldsymbol{\sigma}, T)$ if $i = 1$ or $P(\mathcal{E}_i|\boldsymbol{\eta}, \boldsymbol{\sigma}, T, \tilde{\mathcal{E}_{i-1}^{(j)}})$ otherwise.
3. Each sampled partial trajectory $\mathcal{E}_i^{(j)}$ is then assigned a weight $\omega_i^j = P(\mathcal{T}_i|\mathcal{E}_i^{(j)}, \boldsymbol{\eta}, \boldsymbol{\sigma})$. The sample average of these weights $\Omega_i = (\sum_{j=1}^m \omega_i^j)/m$ is then recorded, and a new set of m unweighted partial trajectories $\tilde{\mathcal{E}}_i$ is sampled with replacement from the weighted distribution.
4. If $i < M$, set $i \leftarrow i + 1$ and go to step 2.
5. Compute the product $\hat{P}(\mathcal{T}|\boldsymbol{\eta}, \boldsymbol{\sigma}, T) \equiv \prod_{i=1}^M \Omega_i$. Also, sample a single final partial trajectory $\tilde{\mathcal{E}}_i$ from the final distribution of weights and follow it back to $t = 0$, yielding a single sampled trajectory $\tilde{\mathcal{E}}$.

This algorithm can be intuitively understood by considering that, in the limit $m \to \infty$, each resampled partial trajectory $\tilde{\mathcal{E}}_v^{(j)}$ is drawn from the distribution $P(\mathcal{E}_i|\mathcal{E}_{i-1}, \dots, \mathcal{E}_1, \mathcal{T}_i, \dots, \mathcal{T}_1, \boldsymbol{\eta}, \boldsymbol{\sigma}, T)$, meaning that the average of each set of weights Ω_i becomes

$$\begin{aligned}\lim_{M\to\infty} \Omega_i &= \sum_{i=1}^{\mathcal{M}} P(\mathcal{T}_i, \mathcal{E}_i = \tilde{\mathcal{E}}_i^{(j)}, \dots, \mathcal{E}_1 = \tilde{\mathcal{E}}_1^{(j)}|\mathcal{T}_{i-1}, \dots \mathcal{T}_1, \boldsymbol{\eta}, \boldsymbol{\sigma}, T) \\ &= P(\mathcal{T}_i|\mathcal{T}_{i-1}, \dots, \mathcal{T}_1, \boldsymbol{\eta}, \boldsymbol{\sigma}, T). \end{aligned} \tag{7}$$

In this same limit, the product of these averages $\hat{P}(\mathcal{T}, \boldsymbol{\eta}, \boldsymbol{\sigma}, T)$ approaches the marginal tree density $P(\mathcal{T}|\boldsymbol{\eta}, \boldsymbol{\sigma}, T)$.

For finite m, the sequences of resampled partial trajectories $\tilde{\mathcal{E}}_M^{(j)}, \dots \tilde{\mathcal{E}}_1^{(j)}$ are not independent. Despite this fact, $\hat{P}(\mathcal{T}|\boldsymbol{\eta}, \boldsymbol{\sigma}, T)$ provided by the bootstrap particle filter is an unbiased estimate of the true marginal probability of the transmission tree $P(\mathcal{T}|\boldsymbol{\eta}, \boldsymbol{\sigma}, T)$.

We use this estimate of the marginal tree density in place of the terms $P(\mathcal{T}|\mathcal{E})P(\mathcal{E}|\boldsymbol{\eta}, \boldsymbol{\sigma}, T)$ in Eq. (6). Samples from the posterior Eq. (6) marginalized over the epidemic trajectory distribution are then obtained using MCMC. By keeping track of the sampled trajectories, $\tilde{\mathcal{E}}$, the algorithm generates samples from the full joint posterior.

3 Results

3.1 Implementation

We have implemented the algorithm described above as a BEAST 2 [24] package. This allows the algorithm to be used in conjunction with standard phylogenetic models such as those describing the nucleotide substitution process as well as existing algorithms for performing the MCMC sampling of the phylogenetic tree space. The package is released under the GNU General Public License and instructions for installing and using it can be found, along with source code, at `http://tgvaughan.github.io/EpiInf`.

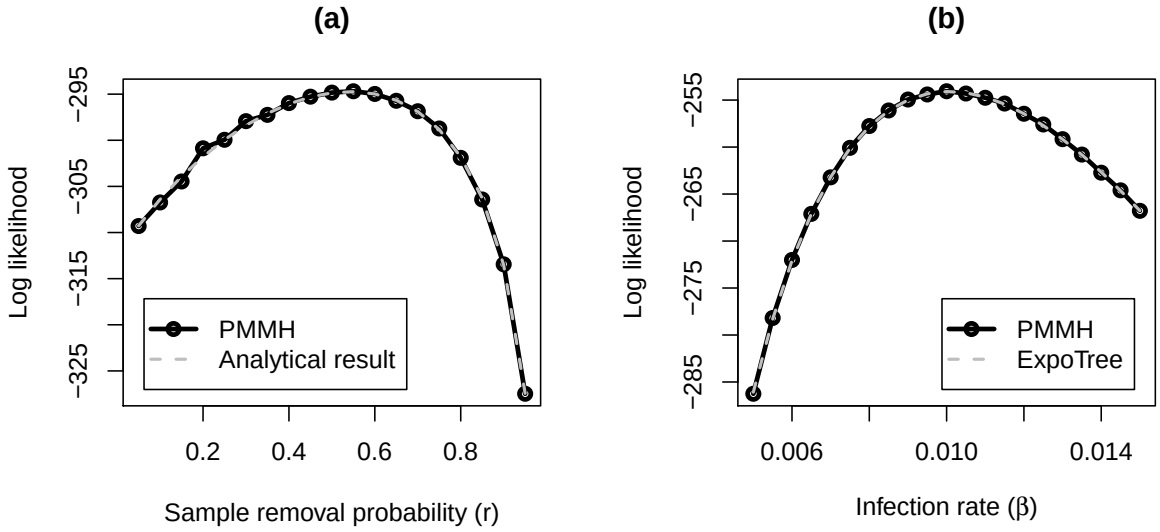

Figure 2: Comparison between values of the marginal epidemiological model parameter likelihoods computed using the PMMH algorithm with those calculated using other approaches: (a) marginal likelihood of r under linear birth-death model compared with the analytical result [8] and (b) marginal likelihood of β under stochastic SIS model compared with a numerical result [15] (ExpoTree).

3.2 Verification of methods

We validated our algorithm and its implementation by comparing the likelihoods generated by the particle filter with those computed (a) analytically under the linear birth-death model [8] and (b) numerically under the nonlinear stochastic SIS model [15]. These comparisons were performed for a variety of parameter combinations and in all cases yielded perfect agreement, as shown in Figure 2.

In order to assess the capability of the sampler to recover prevalence trajectories, we simulated trajectories under each of the three models supported by our implementation: linear birth-death, stochastic SIS and stochastic SIR. Sampled transmission trees were then simulated from each of these these trajectories, which were in turn used to simulate genetic sequence alignments. For each alignment, we used our algorithm to sample from the joint posterior for the transmission tree, epidemic trajectory and model parameters.

Figure 3 illustrates the agreement between the marginal posterior prevalence distributions obtained from each of these analyses (red lines) and the true prevalence curves (black lines). Also shown is the distribution of prevalence curves generated directly from the marginal posterior of the model parameters (blue lines). These trajectories are not explicitly conditioned on the corresponding sampled transmission trees, hence the greater variance in their distribution.

3.3 Inference of Ebola prevalence in Sierra Leone

We analyzed 101 full EBOV genomes collected from Kailahun in eastern Sierra Leone during the 2014 epidemic [25, 26, 27, 28], as collected and aligned by Dudas et al. [29]. We assumed a HKY+Γ substitution model with a strict clock rate whose value was jointly estimated using an informative prior derived from a recent meta-analysis [30]. For the epidemiology, we assumed a stochastic SIR model in which the data are accrued via a fixed rate linear sampling process between the times of the most recent and earliest samples. The initial occupancy of the susceptible compartment was fixed at 3×10^4, which is comparable to a recent estimate [31] of the urban population size of the Luawa chiefdom, which includes Kailahun town. The removal probability r was likewise fixed to 0, meaning that sampling was not assumed to affect infectious potential. The complete list of distributions used for the compartmental model parameters is presented in the second column of table 1.

A total of five independent MCMC chains were run for 5×10^7 steps each compared to assess convergence.

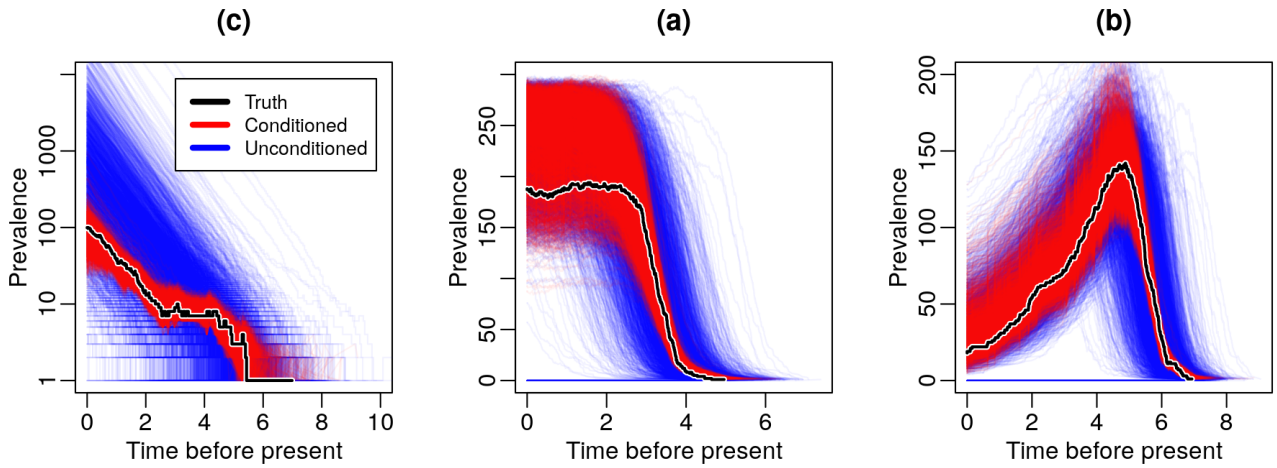

Figure 3: Inference of prevalence dynamics from sequence data simulated under (a) linear birth-death, (b) stochastic SIS and (c) stochastic SIR model. Samples from the marginal posterior of the prevalence trajectory are shown in red, while the black line represents the truth. The blue lines are prevalence trajectories simulated from the posterior samples of the compartmental model parameters.

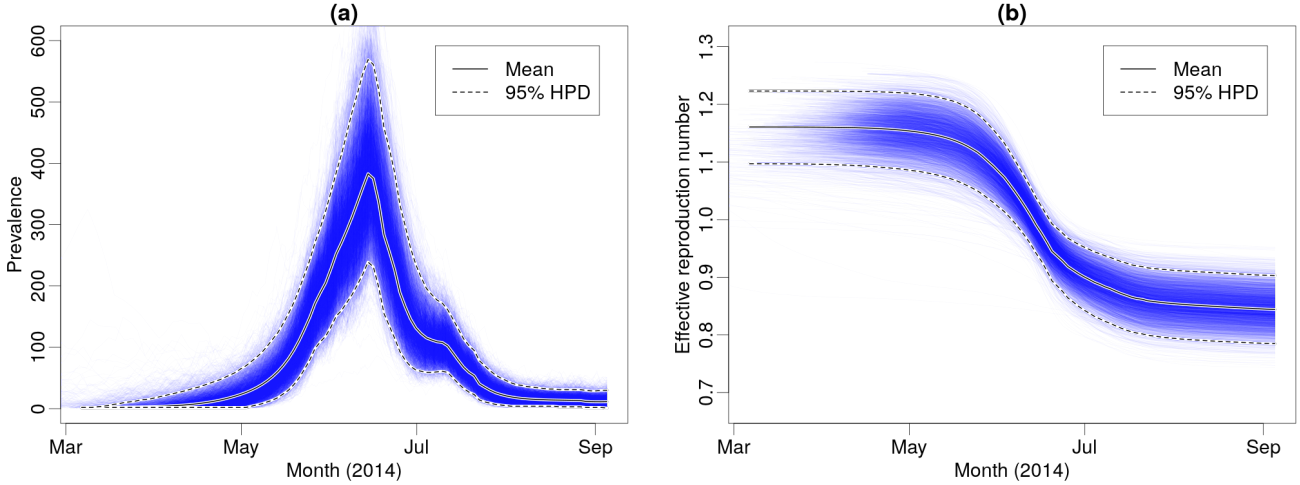

Figure 4: Marginal posterior distributions for (a) prevalence and (b) effective reproduction number during the Kailahun EVD outbreak.

The initial 10% of each chain was removed to account for burn-in and the remainder combined into a single chain from which the final results were derived.

The 95% highest posterior density (HPD) intervals for each of the estimated compartmental model parameters are presented in the right-most column of table 1. The corresponding marginal posterior distributions for the prevalence, $I(t)$, and effective reproduction number, $R_e(t) = \beta S(t)/\mu$, trajectories are shown in Figure 4.

4 Discussion

The primary strength of the PMMH algorithm presented here is its versatility. With only minor modifications, the algorithm can in principle be used to jointly sample the epidemic trajectory and transmission tree under any unstructured stochastic compartmental model whose dynamics may be described by Eq. (1). It owes this versatility to its use of simulation to compute Monte Carlo estimates of the probability density of the transmission tree under the model.

This versatility comes with a cost. Generating the transmission tree probability estimates using sim-

Parameter	Unit	Prior distribution	Posterior 95% HPD	
			Lower	Upper
β	year^{-1}	$\log\mathcal{N}(-5.5, 1.25)$	6.3×10^{-3}	9.9×10^{-3}
μ	year^{-1}	$\log\mathcal{N}(4.6, 1.25)$	160	250
$\psi/(\psi+\mu)$	—	$\text{Unif}(0, 1)$	9.7×10^{-3}	2.1×10^{-2}
t_{or}	year	$1/t_{or}$	0.33	0.47

Table 1: Parameter priors distributions used in and 95% highest posterior density intervals derived from our analysis of EBOV genomes sampled from the 2014 EVD outbreak in Kailahun.

ulation is computationally demanding, particularly when these estimates must be computed hundreds of millions of times as part of a larger MCMC analysis, as it must be in our algorithm. And while PMMH allows one to sacrifice the accuracy of these estimates without affecting the exact nature of the MCMC sampling by reducing the number of stochastic simulations used in the estimate, the corresponding increase in the variance of the estimates can dramatically reduce the rate at which the overall MCMC algorithm "mixes" (i.e. produces effectively independent draws from the posterior).

Furthermore, the inability of the present algorithm to handle structure in the infected population is a significant impediment to realistic epidemic modeling, as this structure can not only originate from spatial segmentation of an infected host population but is also necessary for describing distinct phases of infections with varying degrees of transmissibility or a non-infectious period. To an extent this has already been addressed through the work of Rasmussen, Volz & Koelle [17], although in an approximate way.

The question of whether or not it is possible to construct a phylodynamic sampling algorithm which is simultaneously (a) completely general, (b) mathematically exact and (c) practically useful is therefore still very much open.

5 Acknowledgements

The authors thank Louis du Plessis for helpful suggestions. We also thank the New Zealand eScience Infrastructure for access to high-performance computing facilities (`http://www.nesi.org.nz`). This work was supported by Marsden grant UOA1324 from the Royal Society of New Zealand.

References

1. Kermack WO, McKendrick AG (1927) A Contribution to the Mathematical Theory of Epidemics. Proc. R. Soc. Lond. A 115: 700. DOI: 10.1098/rspa.1927.0118.
2. Drummond AJ et al. (2003) Measurably evolving populations. Trends in Ecology & Evolution 18: 481–488. DOI: 10.1016/S0169-5347(03)00216-7.
3. Grenfell BT et al. (2004) Unifying the epidemiological and evolutionary dynamics of pathogens. Science 303: 327–32. DOI: 10.1126/science.1090727.
4. Pybus OG et al. (2001) The epidemic behavior of the hepatitis C virus. Science 292: 2323–5. DOI: 10.1126/science.1058321.
5. Kingman J (1982) The Coalescent. Stochastic Processes and their Applications 13: 235–248. ISSN: 0304-4149. DOI: 10.1016/0304-4149(82)90011-4. URL: http://www.sciencedirect.com/science/article/pii/0304414982900114.
6. Pybus OG, Rambaut A, Harvey PH (2000) An integrated framework for the inference of viral population history from reconstructed genealogies. eng. Genetics 155: 1429–1437. URL: http://www.genetics.org/content/155/3/1429.short.
7. Drummond AJ et al. (2005) Bayesian coalescent inference of past population dynamics from molecular sequences. Mol Biol Evol 22: 1185–92. DOI: 10.1093/molbev/msi103.
8. Stadler T (2010) Sampling-through-time in birth-death trees. J Theor Biol 267: 396–404. DOI: 10.1016/j.jtbi.2010.09.010.

9. Stadler T et al. (2013) Birth-death skyline plot reveals temporal changes of epidemic spread in HIV and hepatitis C virus (HCV). Proc Natl Acad Sci U S A 110: 228–33. DOI: 10.1073/pnas.1207965110.
10. Volz EM et al. (2009) Phylodynamics of infectious disease epidemics. Genetics 183: 1421–30. DOI: 10.1534/genetics.109.106021.
11. Volz EM (2012) Complex Population Dynamics and the Coalescent Under Neutrality. Genetics 190: 187–201.
12. Rasmussen DA, Ratmann O, Koelle K (2011) Inference for Nonlinear Epidemiological Models Using Genealogies and Time Series. PLoS Comput Biol 7: e1002136.
13. Stadler T et al. (2015) How well can the exponential-growth coalescent approximate constant-rate birth-death population dynamics? eng. Proc Biol Sci 282: 20150420. DOI: 10.1098/rspb.2015.0420. URL: http://dx.doi.org/10.1098/rspb.2015.0420.
14. Kühnert D et al. (2014) Simultaneous reconstruction of evolutionary history and epidemiological dynamics from viral sequences with the birth-death SIR model. J R Soc Interface 11: 20131106. DOI: 10.1098/rsif.2013.1106.
15. Leventhal GE et al. (2014) Using an epidemiological model for phylogenetic inference reveals density dependence in HIV transmission. Mol Biol Evol 31: 6–17. DOI: 10.1093/molbev/mst172.
16. Andrieu C, Doucet A, Holenstein R (2010) Particle Markov chain Monte Carlo methods. J Roy Stat Soc B 72: 269–342. ISSN: 1467-9868. DOI: 10.1111/j.1467-9868.2009.00736.x.
17. Rasmussen DA, Volz EM, Koelle K (2014) Phylodynamic inference for structured epidemiological models. PLoS Comput Biol 10: e1003570. DOI: 10.1371/journal.pcbi.1003570.
18. Smith RA, Ionides EL, King AA (2016) Infectious Disease Dynamics Inferred from Genetic Data via Sequential Monte Carlo. bioRxiv: 096396.
19. Stadler T, Bonhoeffer S (2013) Uncovering epidemiological dynamics in heterogeneous host populations using phylogenetic methods. Philos Trans R Soc Lond B Biol Sci 368: 20120198. DOI: 10.1098/rstb.2012.0198.
20. Andrieu C, Roberts GO (2009) The pseudo-marginal approach for efficient Monte Carlo computations. The Annals of Statistics 37: 697–725. URL: http://projecteuclid.org/euclid.aos/1236693147.
21. Gillespie DT (1976) A general method for numerically simulating the stochastic time evolution of coupled chemical reactions. J Comp Phys 22: 403.
22. Gillespie DT (1977) Stochastic simulation of coupled chemical reactions. J Phys Chem 81: 2340.
23. Gillespie DT (2001) Approximate accelerated stochastic simulation of chemically reacting systems. J. Chem. Phys. 115: 1716. DOI: 10.1063/1.1378322.
24. Bouckaert R et al. (2014) BEAST 2: a software platform for Bayesian evolutionary analysis. eng. PLoS Comput Biol 10: e1003537. DOI: 10.1371/journal.pcbi.1003537. URL: http://dx.doi.org/10.1371/journal.pcbi.1003537.
25. Gire SK et al. (2014) Genomic surveillance elucidates Ebola virus origin and transmission during the 2014 outbreak. Science 345: 1369–1372. DOI: 10.1126/science.1259657.
26. Park D et al. (2015) Ebola Virus Epidemiology, Transmission, and Evolution during Seven Months in Sierra Leone. Cell 161: 1516–1526. DOI: 10.1016/j.cell.2015.06.007.
27. Carroll MW et al. (2015) Temporal and spatial analysis of the 2014–2015 Ebola virus outbreak in West Africa. Nature 524: 97–101. DOI: 10.1038/nature14594.
28. Bell A et al. (2015) Genome sequence analysis of Ebola virus in clinical samples from three British healthcare workers, August 2014 to March 2015. Eurosurveillance 20: 21131. DOI: 10.2807/1560-7917.es2015.20.20.21131.
29. Dudas G et al. (2016) Virus genomes reveal the factors that spread and sustained the West African Ebola epidemic. bioRxiv. DOI: 10.1101/071779. eprint: http://biorxiv.org/content/early/2016/09/16/071779.full.pdf. URL: http://biorxiv.org/content/early/2016/09/16/071779.
30. Holmes EC et al. (2016) The evolution of Ebola virus: Insights from the 2013–2016 epidemic. Nature 538: 193–200. DOI: 10.1038/nature19790.
31. 2015 Popoulation and Housing Census, Summmary of Final Results, Statistics Sierra Leone (2015). URL: http://www.statistics.sl/.