The Annals of Applied Statistics
2023, Vol. 17, No. 1, 1–22
https://doi.org/10.1214/21-AOAS1583

FITTING STOCHASTIC EPIDEMIC MODELS TO GENE GENEALOGIES USING LINEAR NOISE APPROXIMATION

BY MINGWEI TANG[1,a], GYTIS DUDAS[2,3,b], TREVOR BEDFORD[3,c] AND VLADIMIR N. MININ[4,d]

[1]*Department of Statistics, University of Washington, Seattle,* [a]*tangmw1991@gmail.com*
[2]*Gothenburg Global Biodiversity Centre (GGBC),* [b]*gytisdudas@gmail.com*
[3]*Vaccine and Infectious Disease Division, Fred Hutchinson Cancer Research Center,* [c]*trevor@bedford.io*
[4]*Department of Statistics, University of California, Irvine,* [d]*vminin@uci.edu*

Phylodynamics is a set of population genetics tools that aim at reconstructing demographic history of a population based on molecular sequences of individuals sampled from the population of interest. One important task in phylodynamics is to estimate changes in (effective) population size. When applied to infectious disease sequences, such estimation of population size trajectories can provide information about changes in the number of infections. To model changes in the number of infected individuals, current phylodynamic methods use nonparametric approaches (e.g., Bayesian curve-fitting based on change-point models or Gaussian process priors), parametric approaches (e.g., based on differential equations), and stochastic modeling in conjunction with likelihood-free Bayesian methods. The first class of methods yields results that are hard to interpret epidemiologically. The second class of methods provides estimates of important epidemiological parameters, such as infection and removal/recovery rates, but ignores variation in the dynamics of infectious disease spread. The third class of methods is the most advantageous statistically but relies on computationally intensive particle filtering techniques that limits its applications. We propose a Bayesian model that combines phylodynamic inference and stochastic epidemic models and achieves computational tractability by using a linear noise approximation (LNA)—a technique that allows us to approximate probability densities of stochastic epidemic model trajectories. LNA opens the door for using modern Markov chain Monte Carlo tools to approximate the joint posterior distribution of the disease transmission parameters and of high dimensional vectors describing unobserved changes in the stochastic epidemic model compartment sizes (e.g., numbers of infectious and susceptible individuals). In a simulation study we show that our method can successfully recover parameters of stochastic epidemic models. We apply our estimation technique to Ebola genealogies estimated using viral genetic data from the 2014 epidemic in Sierra Leone and Liberia.

1. Introduction.

Phylodynamics is an area at the intersection of phylogenetics and population genetics that studies how epidemiological, immunological, and evolutionary processes affect viral genealogies/phylogenies that were constructed based on molecular sequences sampled from the population of interest (Grenfell et al. (2004), Volz, Koelle and Bedford (2013b)). Phylodynamics is especially useful in infectious disease modeling because genetic data provide a source of information that is complementary to the traditional disease case count data. Here, we are interested in inferring parameters governing infectious disease dynamics from the genealogy/phylogeny estimated from infectious disease agent molecular sequences collected during the disease outbreak. Working in a Bayesian framework, we develop an efficient Markov chain Monte Carlo (MCMC) algorithm that allows us to work with

Received February 2019; revised October 2021.

Key words and phrases. Coalescent, susceptible-infectious-recovered model, state-space model, phylodynamics, Ebola virus.

stochastic models of infectious disease dynamics, properly accounting for stochastic nature of the dynamics.

Infectious disease phylodynamics methods handle densely and sparsely sampled outbreaks differently (but see Smith, Ionides and King (2017), Vaughan et al. (2019) for potentially universal methods). In a densely sampled outbreak scenario, it is possible to simultaneously infer infectious disease dynamics parameters and a transmission network (Ypma, van Ballegooijen and Wallinga (2013), Jombart et al. (2014), Klinkenberg et al. (2017)). When an outbreak is sampled sparsely, a setting we are interested in this paper, it is impossible to determine who infected whom, so additional modeling is needed to connect sampled hosts to the unobserved population dynamics. Currently, learning about population-level infectious disease dynamics from a sparse sample of molecular sequences can be accomplished using three general strategies. The first strategy relies on the coalescent theory—a set of population genetics tools that specify probability models for genealogies relating individuals randomly sampled from the population of interest (Kingman (1982), Griffiths and Tavaré (1994), Donnelly and Tavare (1995)). Using a subset of these models (Griffiths and Tavaré (1994)), it is possible to estimate changes in effective population size—the number of breeding individuals in an idealized population that evolves according to a Wright–Fisher model (Wright (1931)). Such reconstruction can be done assuming parametric (Kuhner, Yamato and Felsenstein (1998), Drummond et al. (2002)) or nonparametric (Drummond et al. (2002), Drummond et al. (2005), Minin, Bloomquist and Suchard (2008), Palacios and Minin (2013), Gill et al. (2013)) functional forms of the effective population size trajectory. In the context of infectious disease phylodynamics, nonparametric inference is the norm, and the estimated effective population size is often interpreted as the effective number of infections or the effective number of infectious individuals. However, reconstructed effective population size trajectories are not easy to interpret, and estimation of parameters of disease dynamics is difficult to accomplish if one wishes to maintain statistical rigor (Pybus et al. (2001), Frost and Volz (2010)).

Another way to learn about infectious disease dynamics from molecular sequences is to model explicitly events that occur during the infectious disease spread and to link these events to the genealogy/phylogeny of sampled individuals using birth-death processes. For example, a susceptible-infectious-removed (SIR) model includes two possible events, infections and removals (e.g., recoveries and deaths), represented by births and deaths in the corresponding birth-death model (Kühnert et al. (2014), Stadler et al. (2013)). Other SIR-like models (e.g., SI and SIS models) differ by the number and types of the events that are needed to accurately describe natural history of the infectious disease (Leventhal et al. (2013)).

Structured coalescent models provide the third strategy of inferring parameters governing spread of an infectious disease (Volz (2012), Volz et al. (2009), Dearlove and Wilson (2013)). These models assume infectious disease agent genetic data have been obtained from a random sample of infected individuals, allowing for serial sampling over time. Although similar to the birth-death modeling framework, the structured coalescent models have two advantages. First, one does not have to keep track, analytically or computationally of extinct and not sampled genetic lineages. Second, the density of the genealogy can be obtained given the population level information about status of individuals: for example, in the SIR model it is sufficient to know the numbers of susceptible $(S(t))$, infectious $(I(t))$, and recovered, $(R(t))$ individuals at each time point t. The second advantage comes with two caveats: (1) such densities can be obtained only approximately, and (2) evaluating densities of genealogies is not straightforward and involves numerical solutions of differential equations. Even in cases when these caveats are manageable, the density of the assumed stochastic epidemic model population trajectory remains computationally intractable. One way around this intractability assumes a deterministic model of infectious disease dynamics (Volz (2012), Volz and Pond (2014), Volz et al. (2009)) which potentially leads to overconfidence in estimation of model parameters. Particle filter MCMC offers another solution (Rasmussen, Ratmann and Koelle (2011), Rasmussen, Volz and Koelle (2014)).

In this paper we develop methods that allow us to bypass particle filter MCMC with the help of a linear noise approximation (LNA). LNA is a low order correction of the deterministic ordinary differential equation describing the asymptotic mean trajectories of compartmental models of population dynamics defined as Markov jump processes (e.g., chemical reaction models and SIR-like models of infectious disease dynamics) (Kurtz (1970), Kurtz (1971), Van Kampen and Reinhardt (1981)). LNA can also be viewed as a first-order Taylor approximation of Markov population dynamics models represented by stochastic differential equations (Giagos (2010), Wallace (2010)). A key feature of the LNA method is that it approximates the transition density of a stochastic population model with a Gaussian density (Komorowski et al. (2009)).

Inspired by recent applications of LNA to analysis of Google flu trends data (Fearnhead, Giagos and Sherlock (2014)) and disease case counts (Buckingham-Jeffery, Isham and House (2018)), we develop a Bayesian framework that combines LNA for stochastic models of infectious disease dynamics with structured coalescent models for genealogies of infectious disease agent genetic samples. Our approach yields a latent Gaussian Markov model that closely resembles a Gaussian state-space model. We use this resemblance to develop an efficient MCMC algorithm that combines high-dimensional elliptical slice sampler updates (Murray, Adams and MacKay (2010)) with low-dimensional Metropolis–Hastings (MH) moves. Using simulations, we demonstrate that this algorithm can handle reasonably complex models, including an SIR model with a time-varying infection rate. We apply this SIR model to a recent Ebola outbreak in West Africa. Our analysis of data from Liberia and Sierra Leone illuminates significant changes in the Ebola infection rate over time, likely caused by the public health response measures and increased awareness of the outbreak in the population.

2. Methodology.

2.1. *Genealogy as data.* We start with n infectious disease agent molecular sequences obtained from infected individuals sampled uniformly at random from the total infected population. Further, we assume that a phylogenetic tree or genealogy, $\mathbf{g}$, relating these sequences has been estimated in such a way that the tree branch lengths respect the known sequence sampling times. Such estimation can be performed with, for example, BEAST—a software package for Bayesian phylogenetic inference (Suchard et al. (2018)). The genealogy is represented by a tree structure with its nodes containing two sources of temporal information: coalescent and sampling times. The coalescent times correspond to the internal nodes of the tree, which are defined as the times at which two lineages in the tree are merged into a common ancestor. The sampling times, corresponding to the tips of the tree, are the times at which molecular sequences were sampled. Note that sampling times are observed directly, while coalescent times are estimated from molecular sequences during phylogenetic reconstruction.

To perform inference about infectious disease dynamics using the above genealogy, we need a probability model that relates the genealogy and infectious disease dynamics model parameters. We assume that the infectious disease is spreading through the population according to the SIR model—a canonical compartmental model that at each time point t tracks the number of susceptible individuals $S(t)$, number of infected/infectious individuals $I(t)$, and number of removed individuals $R(t)$ (Bailey (1975), Anderson and May (1992)). We assume that the population is closed so $S(t) + I(t) + R(t) = N$ for all times t, where N is the population size that we assume to be known. This constraint implies that vector $\mathbf{X}(t) = (S(t), I(t))$ is sufficient to keep track of the population state at time t. We follow common practice and

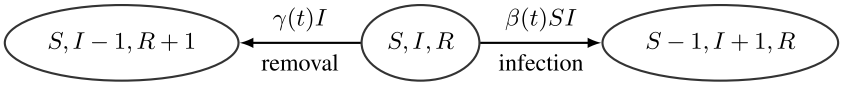

FIG. 1. *SIR Markov jump process. From the current state with the counts S, I, R, the population can transition to state $S-1$, $I+1$, R (an infection event) with rate $\beta(t)SI$ or to state S, $I-1$, $R+1$ (a removal event) with rate $\gamma(t)I$. No other instantaneous transitions are allowed.*

model $\mathbf{X}(t)$ as a Markov jump process (MJP) with allowable instantaneous jumps, shown in Figure 1 (O'Neill and Roberts (1999)). Because we allow the infection rate $\beta(t)$ and removal rate $\gamma(t)$ to be time-varying, the assumed MJP process $\mathbf{X}(t)$ is inhomogeneous.

The structured coalescent models assume that only coalescent times $c_1 < c_2 < \cdots < c_{n-1}$ provide information about the population dynamics. These times are modeled as jumps of an inhomogeneous pure death process with rate $\lambda(t)$, where each "death" event corresponds to coalescence of two lineages and $\lambda(t)$ is called a coalescent

rate. Then, the density of the genealogy, which serves as a likelihood in our work, is written as

$$\Pr(\mathbf{g}) \propto \prod_{k=2}^{n} \lambda(c_{k-1}) \exp\biggl(-\int_{c_{k-1}}^{c_k} \lambda(\tau)\,\mathrm{d}\tau\biggr),$$

where c_n denotes the most recent sequence sampling time. The dependence of coalescent rate on the assumed population dynamics can be complicated and mathematically intractable, but luckily, approximations exist for some specific cases. For the SIR model the approximate coalescent rate can be obtained via the following formula:

$$\lambda(t) = \lambda\bigl(l(t), \beta(t), \mathbf{X}(t)\bigr) = \binom{l(t)}{2} \frac{2\beta(t)S(t)}{I(t)}, \tag{1}$$

where $l(t)$ is the number of lineages present at time t (Rasmussen, Ratmann and Koelle (2011), Volz, Koelle and Bedford (2013b)). The coalescent rate in the SIR model can be interpreted as the rate of infection events between sampled lineages present at time t: $\lambda(t) \approx \binom{l(t)}{2}/\binom{I(t)}{2} \cdot \beta(t)S(t)I(t)$, where $\beta(t)S(t)I(t)$ is the total infection rate in the population and $\binom{l(t)}{2}/\binom{I(t)}{2}$ corresponds to the probability that the infection occurs between lineages present at time t. Note that, when the number of susceptibles is not changing significantly relative to the total population size (i.e., $S(t) \approx N$) and infection rate is constant (i.e., $\beta(t) = \beta$), the structured coalescent reduces to the classical Kingman's coalescent, where we interpret $I(t)/(2\beta N)$ as the effective population size trajectory (Kingman (1982)). It is possible to find approximate coalescence rate for general compartmental models, but closed form expressions exist only for a few models with a low number of compartments (e.g., SI, SIR) (Volz (2012), Volz et al. (2009), Dearlove and Wilson (2013)).

Since we allow sequences to be sampled at different times $s_1 < s_2 < \cdots < s_m = c_n$, some intercoalescent times are censored. To deal with this censoring algebraically, each intercoalesecent interval $[c_{k-1}, c_k)$ is partitioned by the sampling events into i_k subintervals: $\mathcal{I}_{0,k}, \ldots, \mathcal{I}_{i_k-1,k}$. The intervals that start with a coalescent event are defined as $\mathcal{I}_{0,k} = [c_{k-1}, \min\{c_k, s_j\})$, for $s_j > c_{k-1}$ and $k = 2, \ldots, n$. Let the number of lineages in each interval $\mathcal{I}_{i,k}$ be $l_{i,k}$. Then, the number of lineages at each time point t can be written as $l(t) = \sum_{k=2}^{n} \sum_{i=0}^{i_{k-1}} 1_{\{t \in I_{i,k}\}} l_{i,k}$. If the interval $\mathcal{I}_{i,k}$ ends with a coalescent time, the number of lineages in the next interval will be decreased by 1. If the interval ends with a sampling event s_i, then the number of lineages in the next interval is increased by n_i—the number of sequences sampled at time s_i. Figure 2 shows an example of a genealogy with labeled coalescent times, sampling times, number of lineages, and the corresponding intervals.

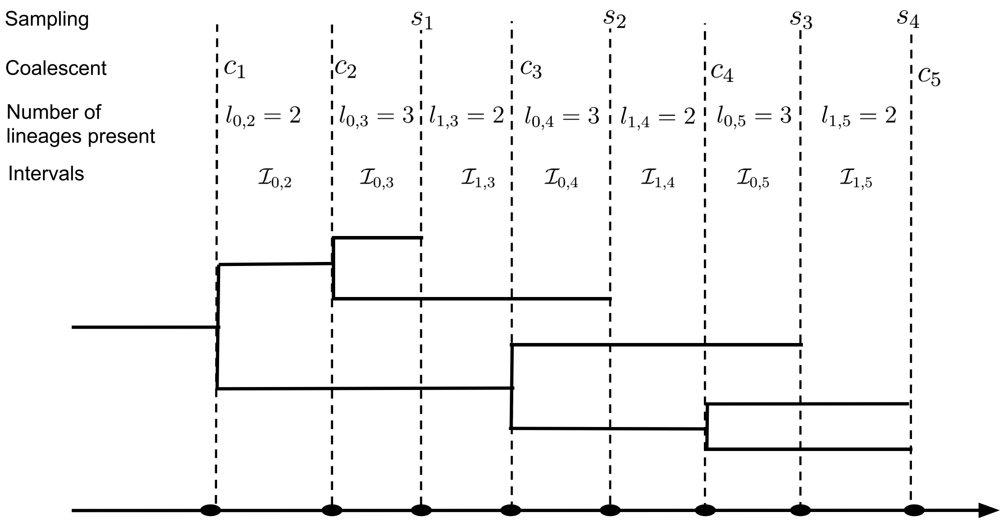

FIG. 2. *Example of a genealogy. Black solid lines show the genealogy structure. The colescent times $c_1, \dots, c_4$ and sampling times $s_1, \dots, s_4$ are labeled with vertical dashed lines. The number of lineages $l_{i,k}$ is given in each intervals $\mathcal{I}_{i,k}$.*

We are now ready to connect the SIR model and a genealogy with serially sampled tips with the help of a structured coalescent density/likelihood. First, we discretize the time interval between the time to the most recent common ancestor c_1 (time corresponding to the root of the tree) and the most recent sampling time s_m using a regular grid $t_0 < t_1 < \cdots < t_T$ ($t_0 < c_1$ and $t_T > s_m$). Using this grid, we discretize the latent epidemic trajectory by assuming that $\mathbf{X}(t) = \sum_{j=1}^{T} \mathbf{X}_{j-1} \mathbf{1}_{[t_{j-1}, t_j)}(t)$, where $\mathbf{X}_j = (S_j, I_j)$ is a column vector. Similarly, we discretize the infectious disease dynamics parameter vector trajectory $\boldsymbol{\theta}(t) = (\beta(t), \gamma(t))$ so that $\boldsymbol{\theta}(t) = \sum_{j=1}^{T} \boldsymbol{\theta}_{j-1} \mathbf{1}_{[t_{j-1}, t_j)}(t)$, where $\boldsymbol{\theta}_j = (\beta_j, \gamma_j)$ is also a column vector. We collect latent variables $\mathbf{X}_j$s and parameters $\boldsymbol{\theta}_j$s into matrices $\mathbf{X}_{0:T}$ and $\boldsymbol{\theta}_{0:T}$ respectively. The SIR structured coalescent density/likelihood then becomes

$$
\begin{aligned}
\Pr(\mathbf{g} \mid \mathbf{X}_{0:T}, \boldsymbol{\theta}_{0:T}) \propto{} & \prod_{k=2}^{n} \binom{l(c_{k-1})}{2} \frac{2\beta(c_{k-1}) S(c_{k-1})}{I(c_{k-1})} \\
& \times \exp\left(- \sum_{i=0}^{i_k - 1} \int_{\mathcal{I}_{i,k}} \binom{l_{i,k}}{2} \frac{2\beta(\tau) S(\tau)}{I(\tau)} \, \mathrm{d}\tau \right).
\end{aligned}
\tag{2}
$$

Since $S(t)$, $I(t)$, and $\beta(t)$ are piecewise constant functions, the integrals in the above formula are readily available in closed form and are fast to compute.

2.2. *Bayesian data augmentation.*

2.2.1. *Posterior distribution.* Given genealogy $\mathbf{g}$, our goal is to infer the latent SIR population dynamic $\mathbf{X}_{0:T}$ and rate parameters $\boldsymbol{\theta}_{0:T}$ over time grid $t_0 < t_1 < \cdots < t_T$. Let $\Pr(\mathbf{X}_0)$ and $\Pr(\boldsymbol{\theta}_{0:T})$ denote the prior densities for the initial compartment states and the SIR parameters, respectively. The posterior distribution for the population trajectory $\mathbf{X}_{0:T}$ and parameters $\boldsymbol{\theta}_{0:T}$, given observed genealogy $\mathbf{g}$, is

$$
\begin{aligned}
\Pr(\mathbf{X}_{0:T}, \boldsymbol{\theta}_{0:T} \mid \mathbf{g}) \propto{} & \Pr(\mathbf{g} \mid \mathbf{X}_{0:T}, \boldsymbol{\theta}_{0:T}) \Pr(\mathbf{X}_{1:T} \mid \mathbf{X}_0, \boldsymbol{\theta}_{0:T}) \\
& \times \Pr(\boldsymbol{\theta}_{0:T}) \Pr(\mathbf{X}_0),
\end{aligned}
\tag{3}
$$

where $\Pr(\mathbf{g} \mid \mathbf{X}_{0:T}, \boldsymbol{\theta}_{0:T})$ is the structured coalescent likelihood introduced in Section 2.1 and $\Pr(\mathbf{X}_{1:T} \mid \mathbf{X}_0, \boldsymbol{\theta}_{0:T})$ is the likelihood function for discrete observations of trajectory $\mathbf{X}_{1:T}$, given the initial value $\mathbf{X}_0$,

$$\Pr(\mathbf{X}_{1:T} \mid \mathbf{X}_0, \boldsymbol{\theta}_{0:T}) = \prod_{i=1}^{T} \Pr(\mathbf{X}_i \mid \mathbf{X}_{i-1}, \boldsymbol{\theta}_{i-1}), \tag{4}$$

where the factorization comes from the assumed Markov property of the disease dynamics. However, the SIR transition density $\Pr(\mathbf{X}_i \mid \mathbf{X}_{i-1}, \boldsymbol{\theta}_{i-1})$ becomes intractable, as population size N grows large, making it difficult to perform likelihood-based inference for outbreaks in large populations.

2.2.2. *Linear noise approximation.* To furnish a feasible computation strategy for large populations, we use a linear noise approximation (LNA) method in which the computationally intractable transition probability $\Pr(\mathbf{X}_i \mid \mathbf{X}_{i-1}, \boldsymbol{\theta}_{i-1})$ is approximated using a closed form Gaussian transition density (Kurtz (1970, 1971), Komorowski et al. (2009)).

The LNA method replaces the MJP discrete state space with a continuous state space of $\mathbf{X}(t)$ to approximate the counts of at time t, under the following constraints: $S(t) > 0$, $I(t) > 0$ and $S(t) + I(t) \leq N$. To briefly explain how this approximation is obtained, we will need additional notation.

The SIR MJP instantaneous transitions, depicted in Figure 1, are encoded in an effect matrix

$$\mathbf{A} = \begin{array}{cc} \text{susceptible} & \text{infected} \\ \left(\begin{array}{c} -1 \\ 0 \end{array}\right. & \left.\begin{array}{c} 1 \\ -1 \end{array}\right) \end{array} \begin{array}{l} \\ \text{infection} \\ \text{removal.} \end{array} \tag{5}$$

Each row in matrix (5) represents a type of transition event, and each column corresponds to a change in the susceptible and infected populations. Next, we define a rate vector $\mathbf{h}$ and a rate matrix $\mathbf{H}$,

$$\mathbf{h}\big(\mathbf{X}(t), \boldsymbol{\theta}(t)\big) = \begin{pmatrix} \beta(t)S(t)I(t) \\ \gamma(t)I(t) \end{pmatrix}, \qquad \mathbf{H} = \begin{pmatrix} \beta(t)S(t)I(t) & 0 \\ 0 & \gamma(t)I(t) \end{pmatrix}. \tag{6}$$

The above notation as well as subsequent developments based on it can be generalized to other epidemic models and, more generally, to a large class of density dependent stochastic processes, such as chemical reaction and gene regulation models (Wilkinson (2011)); see Section A-1 in the Appendix (Tang et al. (2023)) for more details on this generalization.

Consider a transition from $\mathbf{X}_{i-1}$ at time t_{i-1} to $\mathbf{X}_i$ at t_i. Recall that we assume that the SIR rates $\boldsymbol{\theta}(t)$ take constant values $\boldsymbol{\theta}_{i-1}$ in $[t_{i-1}, t_i)$. The LNA represents the value of the next state $\mathbf{X}_i$, as $\mathbf{X}_i = \boldsymbol{\eta}(t_i) + \mathbf{M}(t_i)$, where $\boldsymbol{\eta}(t_i)$ is a deterministic component and $\mathbf{M}(t_i)$ is a stochastic component. The deterministic component $\boldsymbol{\eta}(t_i)$ can be obtained by solving the standard SIR ODE that in our notation can be written as

$$\mathrm{d}\boldsymbol{\eta}(t) = \mathbf{A}^T \mathbf{h}\big(\boldsymbol{\eta}(t), \boldsymbol{\theta}_{i-1}\big)\,\mathrm{d}t, \quad t \in [t_{i-1}, t_i]. \tag{7}$$

The stochastic part $\mathbf{M}(t_i)$ corresponds to the solution of the following SDE at time t_i:

$$\mathrm{d}\mathbf{M}(t) = \mathbf{F}\big(\boldsymbol{\eta}(t), \boldsymbol{\theta}_{i-1}\big)\mathbf{M}(t)\,\mathrm{d}t + \sqrt{\mathbf{A}^T \mathbf{H}\big(\boldsymbol{\eta}(t), \boldsymbol{\theta}_{i-1}\big)\mathbf{A}}\,\mathrm{d}\mathbf{W}_t, \quad t \in [t_{i-1}, t_i], \tag{8}$$

where $\mathbf{F}(\boldsymbol{\eta}(t), \boldsymbol{\theta}_{i-1}) := \frac{\partial \mathbf{A}^T \mathbf{h}(\mathbf{X}(t), \boldsymbol{\theta}_{i-1})}{\partial \mathbf{X}}|_{\mathbf{X}=\boldsymbol{\eta}(t)}$ is the Jacobian matrix of the deterministic part $\mathbf{A}^T \mathbf{h}(\mathbf{X}(t), \boldsymbol{\theta}_{i-1})$ in (7) evaluated at $\boldsymbol{\eta}(t)$. The solution of SDE (8), $\mathbf{M}(t)$, is a Gaussian process and can be recovered by solving two ordinary differential equations governing the mean function $\mathbf{m}(t) := \mathbf{E}[\mathbf{M}(t)]$ and covariance function $\boldsymbol{\Phi}(t) := \mathbf{Var}(\mathbf{M}(t))$,

$$\mathrm{d}\mathbf{m}(t) = \mathbf{F}(\boldsymbol{\eta}(t), \boldsymbol{\theta}_{i-1})\mathbf{m}(t)\,\mathrm{d}t, \tag{9}$$

$$\begin{aligned}\mathrm{d}\boldsymbol{\Phi}(t) = \big(&\mathbf{F}(\boldsymbol{\eta}(t), \boldsymbol{\theta}_{i-1})\boldsymbol{\Phi}(t) + \boldsymbol{\Phi}(t)\mathbf{F}^T(\boldsymbol{\eta}(t), \boldsymbol{\theta}_{i-1}) \\ &+ \mathbf{A}^T\mathbf{H}(\boldsymbol{\eta}(t), \boldsymbol{\theta}_{i-1})\mathbf{A}\big)\,\mathrm{d}t,\end{aligned} \tag{10}$$

for $t \in [t_{i-1}, t_i]$. A heuristic derivation of LNA, based on (Wallace (2010)), is given in Section A-2 of the Appendix. Let $\boldsymbol{\eta}_{t_{i-1}}$, $\mathbf{m}_{t_{i-1}}$, $\boldsymbol{\Phi}_{t_{i-1}}$ denote the initial values of $\boldsymbol{\eta}(t)$, $\mathbf{m}(t)$, $\boldsymbol{\Phi}(t)$ at time t_{i-1} in differential equations (7), (9), and (10), respectively. There are two options for choosing these initial conditions: the nonrestarting LNA of Komorowski et al. (2009) and the restarting LNA of Fearnhead, Giagos and Sherlock (2014). In this paper we will use the nonrestarting LNA by Komorowski et al. (2009) since it allows us to isolate the effect of adding stochasticity to the ODE method, as the mean population trajectory of the non-restarting LNA is the trajectory from the ODE method. The nonrestarting LNA has the following choice of initial conditions:

1. $\boldsymbol{\eta}_{t_{i-1}} = \boldsymbol{\eta}(t_{i-1})$, where $\boldsymbol{\eta}(t_{i-1})$ was obtained by solving the ODE (7) using parameter vector $\boldsymbol{\theta}_{i-2}$ over the interval $[t_{i-2}, t_{i-1}]$,
2. $\mathbf{m}_{t_{i-1}} = \mathbf{X}_{i-1} - \boldsymbol{\eta}(t_{i-1})$,
3. $\boldsymbol{\Phi}_{t_{i-1}} = \mathbf{0}$.

Solving the system of ODEs (7), (9), (10), we obtain $\boldsymbol{\eta}(t_i)$, $\mathbf{m}(t_i)$, and $\boldsymbol{\Phi}(t_i)$. The solution $\mathbf{m}(t_i)$ will be a function of the initial value $\mathbf{X}_{i-1} - \boldsymbol{\eta}(t_{i-1})$, the interval length $\Delta t_i := t_i - t_{i-1}$, and the SIR rates $\boldsymbol{\theta}_{i-1}$. To make this dependence explicit, we write $\mathbf{m}(t_i) := \boldsymbol{\mu}(\mathbf{X}_{i-1} - \boldsymbol{\eta}(t_{i-1}), \Delta t_i, \boldsymbol{\theta}_{i-1})$. Since (9) is a first-order homogeneous linear ODE, the solution $\boldsymbol{\mu}(\mathbf{X}_{i-1} - \boldsymbol{\eta}(t_{i-1}), \Delta t_i, \boldsymbol{\theta}_{i-1})$ is a linear function of $\mathbf{X}_{i-1} - \boldsymbol{\eta}(t_{i-1})$. Hence, the transition from $\mathbf{X}_{i-1}$ to $\mathbf{X}_i$ follows the following Gaussian distribution:

$$\mathbf{X}_i \mid \mathbf{X}_{i-1}, \boldsymbol{\theta}_{i-1} \sim \mathcal{N}\big(\boldsymbol{\eta}(t_i) + \boldsymbol{\mu}(\mathbf{X}_{i-1} - \boldsymbol{\eta}(t_{i-1}), \Delta t_i, \boldsymbol{\theta}_{i-1}), \boldsymbol{\Phi}(t_i)\big). \tag{11}$$

To summarize, the derived conditional Gaussian densities $\Pr(\mathbf{X}_i \mid \mathbf{X}_{i-1}, \boldsymbol{\theta}_{i-1})$ allow us to compute the density of the latent SIR trajectory (4). As a result, our augmented posterior distribution of $\mathbf{X}_{0:T}$ and $\boldsymbol{\theta}_{0:T}$, shown in equation (3), can be computed up to proportionality constant and approximated via “standard” (not particle filter) MCMC approaches.

2.3. *Reparameterization, priors, and MCMC algorithm.*

2.3.1. *Reparameterizing SIR rates.* We have experimented with multiple parameterizations of our inhomogeneous SIR model and found that the following parameterization works best with our proposed MCMC algorithm for approximating the posterior distribution (3). First, recall that we allow SIR rates to vary with time. Since it is much more likely for the infection rate to be time variable, we are going to assume a constant removal/recovery rate γ. This leaves us with the following parameters: infection rates on a grid $\boldsymbol{\beta}$, removal rate γ, and initial SIR state $\mathbf{X}_0 = (S_0, I_0)$. Since we are interested in modeling an emerging infectious disease outbreak, we set the initial counts of susceptibles to $S_0 = N - I_0$. Initial counts of infected individuals, I_0, is assumed to be low and treated as an unknown parameter with a lognormal prior distribution. Instead of the time-varying infection rate $\beta(t)$, we parameterize our SIR model with a time-varying basic reproduction number $R_0(t) = [\beta(t)N]/\gamma$. The reproduction number is interpreted as the average number of cases that one case generates over its infectious period in a completely susceptible population. Since our infection rate changes in a piecewise manner, the basic reproduction number varies over time in a piecewise manner too,

$$R_0(t) = \sum_{i=1}^{T} R_{0_{i-1}} \mathbf{1}_{[t_{i-1}, t_i)}(t), \tag{12}$$

where $R_{0_i} = [\beta_i N]/\gamma$ is the reproduction number corresponding to the time interval $[t_{i-1}, t_i)$. Let $R_0 = R_{0_0}$ be the initial basic reproductive number and $\delta_i = \log(R_{0_i}/R_{0_{i-1}})/\sigma$ be a normalized log ratio of $R_0(t)$ over two successive time intervals. Then, interval-specific basic reproduction numbers can be written as

$$R_{0_i} = R_0(t, \boldsymbol{\delta}_{1:T}, \sigma) = R_0 \exp\left(\sum_{k=1}^{i} \sigma \delta_k\right), \quad \text{for } i = 1, \dots, T, \tag{13}$$

where we assume a priori that δ_is are independent standard normal random variables.

This construction implies that log-transformed piecewise constant reproduction numbers, $\log(R_{0_i})$s, a priori follow a first-order Gaussian Markov random field (GMRF) with standard deviation σ that controls the a priori smoothness of $R_0(t)$ trajectory (Rue (2001), Rue and Held (2005)). In addition to speeding MCMC convergence, working with $R_0(t)$ is convenient, because this trajectory is dimensionless and retains its interpretation when one changes the population size N. The initial R_0 is assigned a lognormal(a_1, b_1) prior. We use a lognormal(a_2, b_2) prior for the inverse of standard deviation $1/\sigma$.

2.3.2. *Grid size and prior for GMRF standard deviation.* The number of grid intervals T can be thought of as a tuning parameter in our model. Increasing T linearly increases complexity of the coalescent likelihood and $R_0(t)$ prior density calculations, suggesting that keeping T small is prudent from a computational point of view. However, if the chosen T is too small, we may miss large changes of the latent numbers of susceptible and infectious individuals and changes of the basic reproduction number. We recommend choosing T large enough to capture these changes, possibly experimenting with multiple grid sizes. We recommend setting the prior distribution for σ in conjunction with T, for example, by controlling the probability that $R_0(t)$ a priori stays within a reasonable range.

2.3.3. *Reparameterizing SIR latent trajectories.* We reparameterize the latent SIR trajectory $\mathbf{X}_{1:T}$ with a sequence of independent Gaussian random variables $\boldsymbol{\xi}_{1:T}$, following a noncentered parameterization framework of Papaspiliopoulos, Roberts and Sköld (2007). According to formula (11), conditional on $\mathbf{X}_{i-1}$, $\mathbf{X}_i$ can be written as

$$\mathbf{X}_i = \boldsymbol{\eta}(t_i) + \boldsymbol{\mu}\big(\mathbf{X}_{i-1} - \boldsymbol{\eta}(t_{i-1}), \Delta t_i, \boldsymbol{\theta}_{i-1}\big) + \boldsymbol{\Phi}_i^{1/2} \boldsymbol{\xi}_i, \tag{14}$$

where $\boldsymbol{\xi}_i \overset{\text{iid}}{\sim} \mathcal{N}(0, \mathbf{I})$ for $i = 1, \dots, T$ and $\mathbf{I}$ is a 2×2 identity matrix. In our parameterization we will treat $\boldsymbol{\xi}_{1:T}$ as random latent variables and the SIR latent trajectory $\mathbf{X}_{1:T}$ as a deterministic transformation of $\boldsymbol{\xi}_{1:T}$. More details about our noncentered parameterization of $\mathbf{X}_{1:T}$ can be found in Section A-3 of the Appendix.

2.3.4. *MCMC algorithm.* Using our new parameterization, we are now interested in the posterior distribution of the initial number of infected individuals, I_0, removal rate, γ, the initial basic reproduction number, R_0, standardized vectors, $\boldsymbol{\delta}_{1:T}$ and $\boldsymbol{\xi}_{1:T}$, and GMRF standard deviation, σ,

$$\begin{aligned}
\Pr(I_0, R_0, \gamma, \boldsymbol{\delta}_{1:T}, \boldsymbol{\xi}_{1:T}, \sigma | \mathbf{g}) &\propto \Pr(\mathbf{g} | I_0, R_0, \gamma, \boldsymbol{\delta}_{1:T}, \boldsymbol{\xi}_{1:T}, \sigma) \Pr(I_0) \\
&\quad \times \Pr(R_0) \Pr(\gamma) \Pr(\boldsymbol{\delta}_{1:T}) \Pr(\boldsymbol{\xi}_{1:T}) \Pr(\sigma) \\
&\propto \Pr(\mathbf{g} | \mathbf{X}_{0:T}, \boldsymbol{\theta}_{0:T}) \Pr(I_0) \Pr(R_0) \Pr(\gamma) \\
&\quad \times \Pr(\boldsymbol{\delta}_{1:T}) \Pr(\boldsymbol{\xi}_{1:T}) \Pr(\sigma).
\end{aligned}$$

The latent variables $\mathbf{X}_{0:T}$ and parameter vector $\boldsymbol{\theta}_{0:T}$ are deterministic functions of new parameters I_0, γ, R_0, $\boldsymbol{\delta}_{1:T}$, $\boldsymbol{\xi}_{1:T}$, and σ. We use the following MCMC with block updates to approximate this posterior distribution. We update high-dimensional vector $\mathbf{U} = (\log(R_0), \boldsymbol{\delta}_{1:T}, \log(\sigma))$ using the efficient elliptical slice sampler (Murray, Adams and MacKay (2010)). Vector $\boldsymbol{\xi}_{1:T}$ is updated the same way in a separate step. Initial number of invected individuals I_0 and removal rate γ are updated using univariate Metropolis steps. The full procedure is described in Algorithm 2 which, together with details of the elliptical slice sampler, can be found in Section A-4.1 of the Appendix. After MCMC is done, we report posterior summaries using natural parameterization. For example, we report posterior medians and 95% Bayesian credible intervals (BCIs) of the piecewise latent reproduction number trajectory, R_{0_i} for $i = 0, \dots, T$, and latent trajectory $\mathbf{X}_{0:T}$.

2.3.5. *Implementation.* Our R package called `LNAPhylodyn` provides an implementation of our MCMC algorithm. The package code is publicly available at https://github.com/MingweiWilliamTang/LNAphyloDyn. This repository also contains scripts that should allow one to reproduce key numerical results in this manuscript. The PhyDyn simulation example is also included in https://github.com/MingweiWilliamTang/LNAphyloDyn/blob/master/inst/SIR_phydyn_example.xml.

3. Simulation experiments.

3.1. *Simulations based on single genealogy realizations.* In this section we use simulated genealogies to assess performance of our LNA-based method and to compare it with an ODE-based method, where we replace equation (14) with its simplified version, $\mathbf{X}_i = \boldsymbol{\eta}(t_i)$. Under our assumption of a fixed and known genealogy and constant R_0, our ODE-based method closely resembles previously developed methods by Volz et al. (2009) and Volz and Siveroni (2018). To compare ODE-based and LNA-based models in a Bayesian nonparametric setting, we equip the ODE model with the GMRF prior for time-varying $R_0(t)$, described in Section 2.3.1. We use the same MCMC algorithm for both LNA-based and ODE-based models, except we do not have a separate step to update latent vector $\boldsymbol{\xi}_{1:T}$ (equivalently, $\mathbf{X}_{0:T}$) in the ODE-based inference; see Algorithm 3 in the Appendix for a more detailed description of the ODE-based MCMC.

The simulation protocol consists of two steps. First, given the population size N and prespecified parameters γ, I_0, and $R_0(t)$, we simulate one realization of the SIR population trajectory based on the MJP using the Gillespie algorithm (Gillespie (1977)). Next, we generate realistic lineage sampling times and simulate coalescent times from the distribution specified by density (2) using a thinning algorithm by Palacios and Minin (2013). We specified several sampling times spanning the time of the epidemic. The number of sampled sequences at each sampling time in each scenario is set to be approximately proportional to the true prevalence. More details are given in Appendix Section A-5.1.

We test LNA-based and ODE-based methods under three "true" $R_0(t)$ trajectories over the time interval $[0, 90]$:

1. Constant (CONST) $R_0(t)$. $R_0(t) = 2.2$ for $t \in [0, 90]$. Recovery rate $\gamma = 0.2$. Initial counts of infected individuals $I_0 = 1$. Total population size is $N = 100{,}000$. The total number of sampled sequences is 1022.
2. Stepwise decreasing (SD) $R_0(t)$. $R_0(t) = 2$, $t \in [0, 30)$, $R_0(t) = 1$, $t \in [30, 60)$, and $R_0(t) = 0.6$, $t \in [60, 90]$. Recovery rate $\gamma = 0.2$. Initial counts of infected individuals $I_0 = 1$. Total population size $N = 1{,}000{,}000$. The total number of sampled sequences is 342.

3. Nonmonotonic (NM) $R_0(t)$. $R_0(t) = 1.4 \times 1.015^{0.5t}$, $t \in [0, 30]$, $R_0(t) = 1.750 \times 0.975^{t-30}$, $t \in [30, 80]$, and $R_0(t) = 0.4583$, $t \in [80, 90]$. Recovery rate $\gamma = 0.3$. Initial counts of infected individuals $I_0 = 3$. Total population size $N = 1{,}000{,}000$. The total number of sampled sequences is 442.

For all simulations we use lognormal$(1, 1)$ prior for I_0. The parameters of the lognormal priors for the initial R_0 and inverse standard deviation $1/\sigma$ are set to $a_1 = 0.7$, $b_1 = 0.5$, and $a_2 = 3$, $b_2 = 0.2$, respectively, in such a way that a priori $R_0(t)$ trajectory stays within a reasonable range of [0, 5] with 0.9 probability. We assign an informative prior for γ in each simulation scenario, assuming that prior information about this parameter is available: (1) CONST: $\gamma \sim$ lognormal$(-1.7, 0.1)$, (2) SD: $\gamma \sim$ lognormal$(-1.7, 0.1)$, (3) NM: $\gamma \sim$ lognormal$(-1.2, 0.1)$. We set the grid size to $T = 36$, with $t_i - t_{i-1} = 2.5$ for $i = 1, \dots, 36$. As a result, each scenario has 72 latent variables that keep track of latent numbers of infectious and removed individuals, $\mathbf{X}_{1:36}$, and 36 parameters that describe changes in the basic reproduction number, $\boldsymbol{\delta}_{1:36}$, plus parameters R_0, I_0, γ, and σ. For both LNA-based and ODE-based methods, we use 1,000,000 MCMC iterations. All MCMC chains appeared to converge (trace plots are shown in Section A-5.4.1 of the Appendix). The effective sample sizes of all unknown quantities were above 400 (see Table A-1 for more details).

The first row of Figure 3 shows pointwise posterior medians and 95% BCIs for the basic reproduction number trajectory, $R_0(t)$. Our LNA-based method performs well in capturing the continuous dynamics of $R_0(t)$. Though our approach may not perfectly catch the discontinuous changes in R_0 in the SD scenario, the method provides BCIs that are able to capture most of the $R_0(t)$ trajectory. The ODE-based method yields similar results in the CONST case and the SD case but underestimates the magnitude of the decrease in $R_0(t)$ toward the end of the epidemic.

The second row in Figure 3 shows posterior summaries of removal rate γ. Both LNA-based and ODE-based methods provide good estimates in the CONST scenario with posterior modes centered at the true value and higher posterior densities at truth when compared with the prior. In the SD and NM scenarios with the time varying $R_0(t)$, the posterior estimates from the LNA-based method and ODE-based method, though still centered at the truth, do not differ much from the prior distribution.

Posterior summaries of $S(t)$ and $I(t)$ are depicted in the third and fourth rows of Figure 3. The two methods produce similar results in the CONST and SD scenario, as both of them have narrow BCIs covering the true trajectories. However, in the NM case, while the LNA-based method manages to recover the latent SIR trajectory trend, the BCIs from the ODE-based method fail to cover the true prevalence trajectory in the middle and at the end of the epidemic. Somewhat counterintuitively, LNA-based method produces BCIs for the latent trajectories, $S(t)$ and $I(t)$, that are narrower than its ODE counterparts. We suspect this is a result of the ODE-based method poor estimation of the basic reproduction number trajectory at the end of the epidemic.

3.2. *Frequentist properties of posterior summaries.* In this section we design a simulation study based on repeatedly simulating SIR trajectories using MJP with prespecified parameters. We report simulations based on the nonmonotonic $R_0(t)$ trajectory scenario in Section 3.1 with the same parameter setup, except the parameters of the lognormal prior for the initial R_0 are set to $a_1 = 0.7$, $b_1 = 0.3$. Results of repeatedly simulating SIR trajectories with constant and monotonic $R_0(t)$ trajectories are reported in Appendix Section A-5.3. Simulating SIR dynamics under low initial number of infected individuals I_0 can end up with low prevalence trajectories that end at the beginning of the epidemic, or trajectories having unrealistically high prevalence, which are less likely to be observed during real infectious disease outbreaks. Therefore, while simulating SIR trajectories, we reject such “unreasonable”

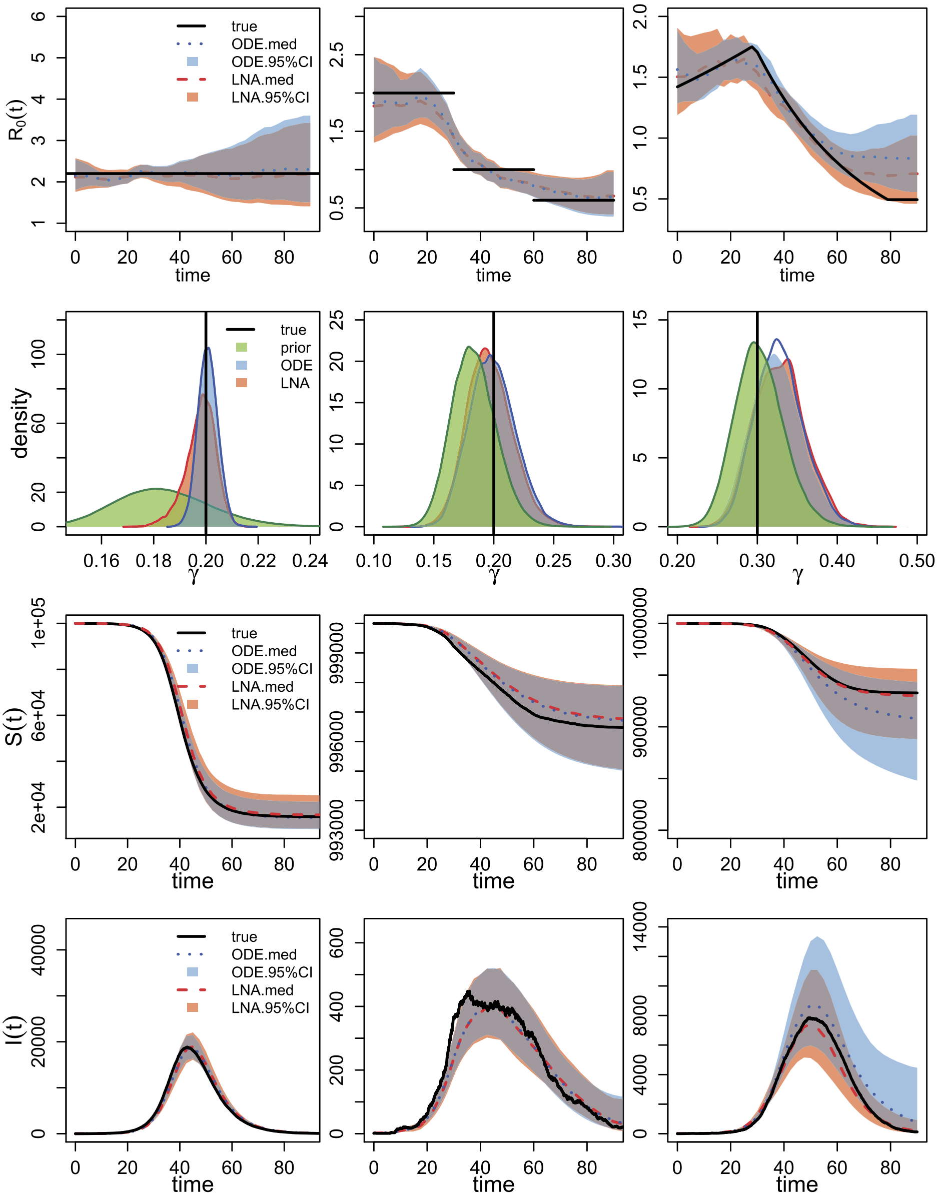

FIG. 3. *Analysis of three simulation scenarios. Columns correspond to CONST, SD, and NM simulated $R_0(t)$ trajectories. The first row shows the estimated $R_0(t)$ trajectories for the three scenarios, with the black solid lines representing the truth, the red dashed lines depicting the posterior median, and the red-shaded area showing the 95% BCIs for the LNA-based method. For the ODE-based method, the posterior median is plotted in blue dotted lines, with blue shading showing the 95% BCIs. The second row corresponds to the estimation for the removal rate γ. Posterior density curves from the LNA are shown in red lines, and the posterior density for ODE is plotted in blue lines, compared with prior density curve in green lines. The bottom two figures shows the estimated trajectory of $S(t)$ and $I(t)$, respectively.*

realizations to arrive at 100 simulated trajectories. The details of the rejection criteria are given in Section A-5.2 of the Appendix. For each simulated SIR trajectory, a realization of a genealogy is generated using the structured coalescent process. We use both LNA-based and ODE-based models to approximate the posterior distribution of model parameters and latent variables for each genealogy. In addition to the informative prior for removal rate γ, used in

Section 3.1, we use a weaker prior $\gamma \sim \text{lognormal}(-1.2, 0.25)$ to probe prior sensitivity of both LNA-based and ODE-based methods.

We use three metrics to evaluate models based on their estimates of $R_0(t)$ and $I(t)$: average error of point estimates (posterior medians), width of credible intervals, and frequentist coverage of credible intervals. Since the value of $R_0(t)$ is greater than 0 and usually upper-bounded by 20 (i.e, it stays within the same order of magnitude), we will measure accuracy using an unnormalized mean absolute error (MAE),

$$\text{MAE} = \frac{1}{T+1}\sum_{i=0}^{T} \big|\hat{R}_{0_i} - R_0(t_i)\big|, \tag{15}$$

where $\hat{R}_{0_i}$ is the posterior median of $R_0(t_i)$. In contrast, $I(t)$ varies from one at the beginning of the epidemic to thousands at the peak, so to evaluate accuracy of prevalence estimation, we use the mean relative absolute error (MRAE),

$$\text{MRAE} = \frac{1}{T+1}\sum_{i=0}^{T} \frac{|\hat{I}_i - I(t_i)|}{I(t_i)+1}, \tag{16}$$

where $\hat{I}_i$ is the posterior median of I(t_i). We assess precision of $R_0(t)$ estimation based on the mean credible interval width (MCIW),

$$\text{MCIW} = \frac{1}{T+1}\sum_{i=0}^{T} \big[\hat{R}_{0_i}^{0.975} - \hat{R}_{0_i}^{0.025}\big], \tag{17}$$

where $\hat{R}_{0_i}^{0.025}$ and $\hat{R}_{0_i}^{0.975}$ denote the lower and upper bounds of the 95% BCI for R_{0_i}. Similar as our measure of accuracy, precision of $I(t)$ estimation is quantified via mean relative credible interval width (MRCIW),

$$\text{MRCIW} = \frac{1}{T+1}\sum_{i=0}^{T} \frac{\hat{I}_i^{0.975} - \hat{I}_i^{0.025}}{I(t_i)+1}, \tag{18}$$

where $\hat{I}_i^{0.025}$ and $\hat{I}_i^{0.975}$ specify the lower and upper bounds of the 95% BCI of $I(t_i)$. In addition, we compute the "envelope" (ENV)—a measure of coverage of BCIs the true trajectory—for $R_0(t)$ and $I(t)$ as follows:

$$\text{ENV-R}_0 = \frac{1}{T+1}\sum_{i=0}^{T} \mathbb{1}\big(\hat{R}_{0_i}^{0.025} \le R_0(t_i) \le \hat{R}_{0_i}^{0.975}\big),$$

$$\text{ENV-I} = \frac{1}{T+1}\sum_{i=0}^{T} \mathbb{1}\big(\hat{I}_i^{0.025} \le I(t_i) \le \hat{I}_i^{0.975}\big).$$

Sampling distribution boxplots of $R_0(t)$, posterior summaries are depicted in the left three plots of Figure 4. The LNA-based method yields lower MAE than the ODE-based method under both informative and weakly informative priors for the removal rate γ. As a trade-off, the MCIWs produced by the LNA-method are generally higher, as expected, since the LNA-based method incorporates the stochasticity in the population dynamics. With less bias and wider BCIs, the LNA-based method BCIs result in better $R_0(t)$ coverage than ODE-based BCIs, as shown by the envelope boxplots. Informative prior for the removal rate γ helps both LNA-based and ODE-based methods to estimate $R_0(t)$.

Sampling distribution boxplots of $I(t)$ posterior summaries, shown in Figure 4, are similar to the $R_0(t)$ results, with the LNA-based method generally having lower MRAEs, higher MRCIWs and a better coverage/envelope than the ODE-based method. Again, somewhat counter-intuitively, the MRCIWs for the LNA-based method are smaller than the ODE counterparts.

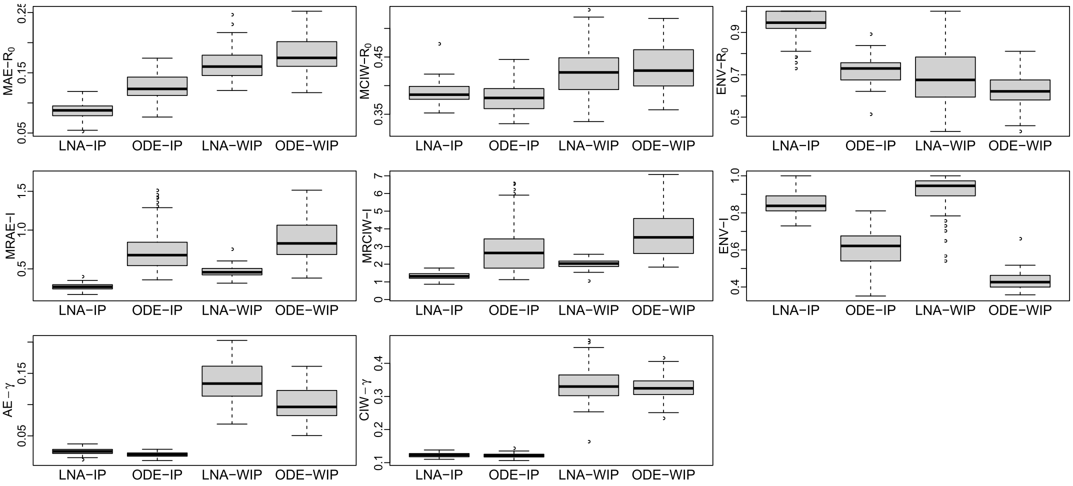

FIG. 4. *Boxplots comparing performance of LNA-based and ODE-based methods using* 100 *simulated genealogies under informative prior* (*IP*) *and weakly informative prior* (*WIP*) *for removal rate* γ. *The first row shows mean absolute error* (*MAE*), *mean credible interval width* (*MCIW*), *and envelope* (*ENV-*R_0) *for* $R_0(t)$ *trajectory. The second row depicts mean relative absolute error* (*MRAE*), *mean relative credible interval width* (*MRCIW*), *and envelope* (*ENV-I*) *for* $I(t)$ (*prevalence*) *trajectory* (*ENV-I*). *The last two plots show the absolute error* (*AE*) *and Bayesian credible interval* (*BCI*) *width for* γ.

This is likely caused by significant bias in $R_0(t)$ estimation by the ODE-based method. The contrast between results of informative and weakly informative prior is a little different from $R_0(t)$ estimation results, because the LNA-based method is estimating $I(t)$ better than $R_0(t)$ under the weakly informative prior.

We also report the absolute error (AE) and 95% BCI widths for removal rate γ in Figure 4. The LNA-based method yields slightly higher AEs than the ODE method. Under the informative prior, both LNA-based and ODE-based methods have coverage of 95% BCIs equal to 1.0. However, coverage of LNA-based method drops to 0.65 under the weakly informative prior, while the ODE-based method's 95% BCI coverage becomes 0.99.

In conclusion, the ODE-based method tends to be biased and overconfident when estimating basic reproduction number $R_0(t)$ and prevalence $I(t)$. By modeling stochasticity of the population trajectory dynamics, our LNA-based method produces more accurate and less precise estimators of $R_0(t)$ and $I(t)$ that enjoy good frequentist properties. However, the ODE-based method does better in estimating the recovery rate γ which is only weakly identifiable.

3.3. *Additional simulations and validation.* We perform the same repeated simulations for the constant and stepwise decreasing $R_0(t)$ scenarios under the same parameter setup as in Section 3.1 and report the corresponding frequentist properties of the posterior summaries in Appendix Figures A-5 and A-6. Both LNA-based and ODE-based methods results are similar to the results from the nonmonotonic $R_0(t)$ simulation scenario, but the differences between LNA-based and ODE-based methods are less pronounced than in the nonmonotonic $R_0(t)$ scenario.

Theoretically, both structured coalescent models and LNA are designed to work for epidemics in large populations. We test performance of LNA-based and ODE-based methods in a relatively small population with the size of $N = 1000$. For simplicity, we use a constant $R_0(t)$ simulation scenario. Assuming that R_0 is constant also allows us to compare our method to the BEAST 2 `PhyDyn` module that implements the ODE-based approach. `PhyDyn` can handle a wide range of different compartmental models of infectious disease dynamics, but we use only a simple SIR model in this comparison. This simulation study shows that our implementations of both LNA-based and ODE-based approaches perform reasonably in this small population setting, but `PhyDyn` does do as well. However, we find that the disagreement between our ODE implementation and `PhyDyn` is artifact of the small population size setting, which leads to the outbreak to be densely sampled. In Appendix Section A-9 we demonstrate that our ODE-based method implementation agrees with R package `PhyDynR` (a predecessor of `BEAST 2 PhyDyn`) under a setup with a large population size, but the two implementations disagree under a small population size setting.

4. Analysis of Ebola outbreak in West Africa. We apply our LNA-based method to the Ebola genealogies reconstructed from molecular data collected in Sierra Leone and Liberia during the 2014–2015 epidemic in West Africa (Dudas et al. (2017)). We use a Sierra Leone genealogy, depicted in the top left plot of Figure 5, which was estimated from 1010 Ebola virus full genomes sampled from 2014-05-25 to 2015-09-12 in 15 cities. The Liberia genealogy, shown in the top left plot of Figure 6, was estimated from a smaller number of samples: 205 Ebola virus full genomes sampled from 2014-06-20 to 2015-02-14. The original sequence data and the reconstructed genealogies are publicly available at https://github.com/ebov/space-time.

When Ebola virus infections were detected in West Africa in mid-Spring of 2014, various intervention measures were proposed and implemented to change behavior of individuals in the populations through which Ebola was spreading. Border closures, encouragement to reduce individual day-to-day mobility, and recommendations on changing burial practices were among the broad spectrum of interventions attempted by multiple countries. It is reasonable to expect that these intervention measures resulted in lowering the contact rates among members of the populations, which, in turn, reduced the infection rate or, equivalently, the basic reproduction number.

When analyzing the Sierra Leone and Liberia genealogies, we rely on conclusions of Dudas et al. (2017) and assume the population in each country to be well mixed. Furthermore, we assume Ebola spread to follow SIR dynamics. For each country, the population size is specified based on its census population size in 2014, with $N = 7{,}000{,}000$ for Sierra Leone and $N = 4{,}400{,}000$ for Liberia. We investigated robustness to population size misspecification in Appendix Section A-8.2 and found that altering population size of Liberia by an order of magnitude in each direction did not appreciably change estimation results. As in our simulation study, we use the lognormal prior for R_0 with $a_1 = 0.7$ and $b_1 = 0.5$ and the lognormal prior for the inverse standard deviation $1/\sigma$ with $a_2 = 3$, $b_2 = 0.2$. Recall that this prior setting ensures that a priori $R_0(t)$ stays within a reasonable range of $[0, 5]$ with probability 0.9. For removal rate γ, we use an informative lognormal prior with mean 3.4 and variance 0.2 based on previous studies (Towers, Patterson-Lomba and Castillo-Chavez (2014)). The parameter $1/\gamma$, interpreted as the length of the infectious period, is expected to be eight to 18 days for each country a priori. The total time span for each genealogy is divided evenly into 40 intervals, which results in grid interval lengths, Δt_is, to be 12.41 days for Sierra Leone and 6.9 days for Liberia. We experimented with two additional grid sizes for the Liberia analysis in Appendix Section A-7 and found that our results are not too sensitive to the choice of grid size.

We run the MCMC algorithm in Section 2.3 for 2,000,000 iterations with nine parallel chains for Sierra Leone data and 750,000 iterations for Liberia data using a single chain. The posterior samples are obtained by discarding the first 100,000 iterations and saving every 30th iteration afterward. The trace plots in Section A-5.4.2 of the Appendix indicate the MCMC algorithm has converged and achieved good mixing in each case.

Figures 5 and 6 show results for Sierra Leone and Liberia, respectively, with intervention events mapped onto the calendar time on the x-axis. Our LNA-based method estimates the

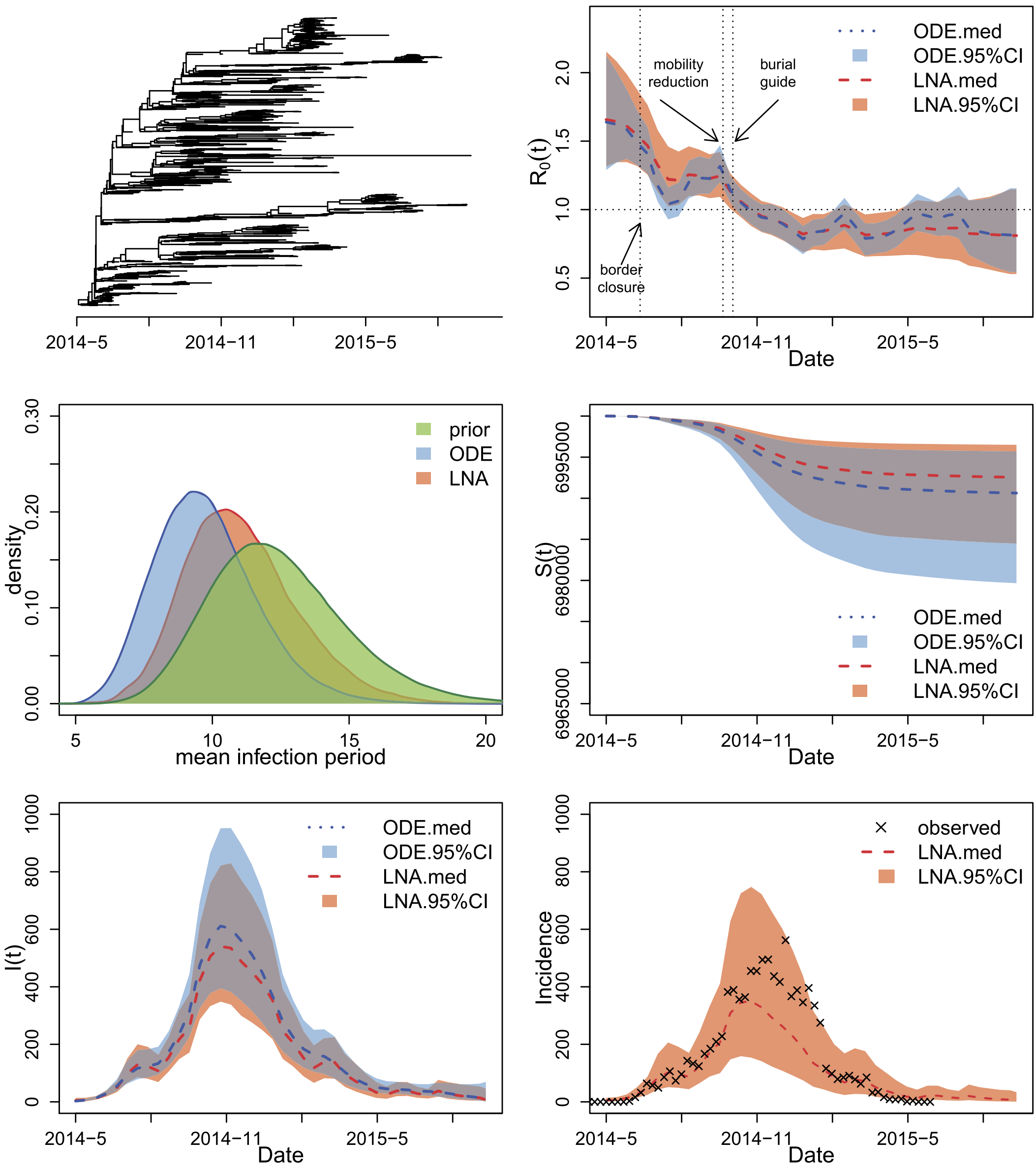

FIG. 5. *Analysis of the genealogy relating Ebola virus sequences collected in Sierra Leone. The top left plot depicts the Ebola genealogy. The top right plot shows the estimated $R_0(t)$, with the red dashed line showing the posterior median and the salmon shaded area showing the 95% BCIs of the LNA-based method. The posterior median, based on the ODE-based method, is plotted as the blue dotted line with blue shading corresponding to the 95% BCIs. The medium left figure shows prior and posterior densities of the mean infection period $1/\gamma$. The prior density is shown in green, while the posterior densities, based on LNA and ODE, are plotted in red and blue, respectively. The medium right and the bottom left figures show the estimated trajectory of $S(t)$ and $I(t)$, using the same legend as in top right plot. The bottom right plot shows the predicted median and 95% BCIs for weekly reported incidence together with the reported incidence from WHO shown as crosses.*

initial R_0 in Sierra Leone during 2014–2015 to be 1.66, with 95% BCI of $(1.31, 2.15)$. Similarly, R_0 in Liberia during 2014–2015 has a point estimate 1.67 and a 95% BCI $(1.29, 2.24)$. using incidence data. Our estimate of initial R_0 in Sierra Leone is consistent with the estimates of Stadler et al. (2014), who fitted multiple birth-death models to 72 sequences at the early stages of the outbreak. Our LNA-based method yields a slightly smaller estimate of the initial R_0 than methods based on susceptible-exposed-infectious-removed (SEIR) mod-

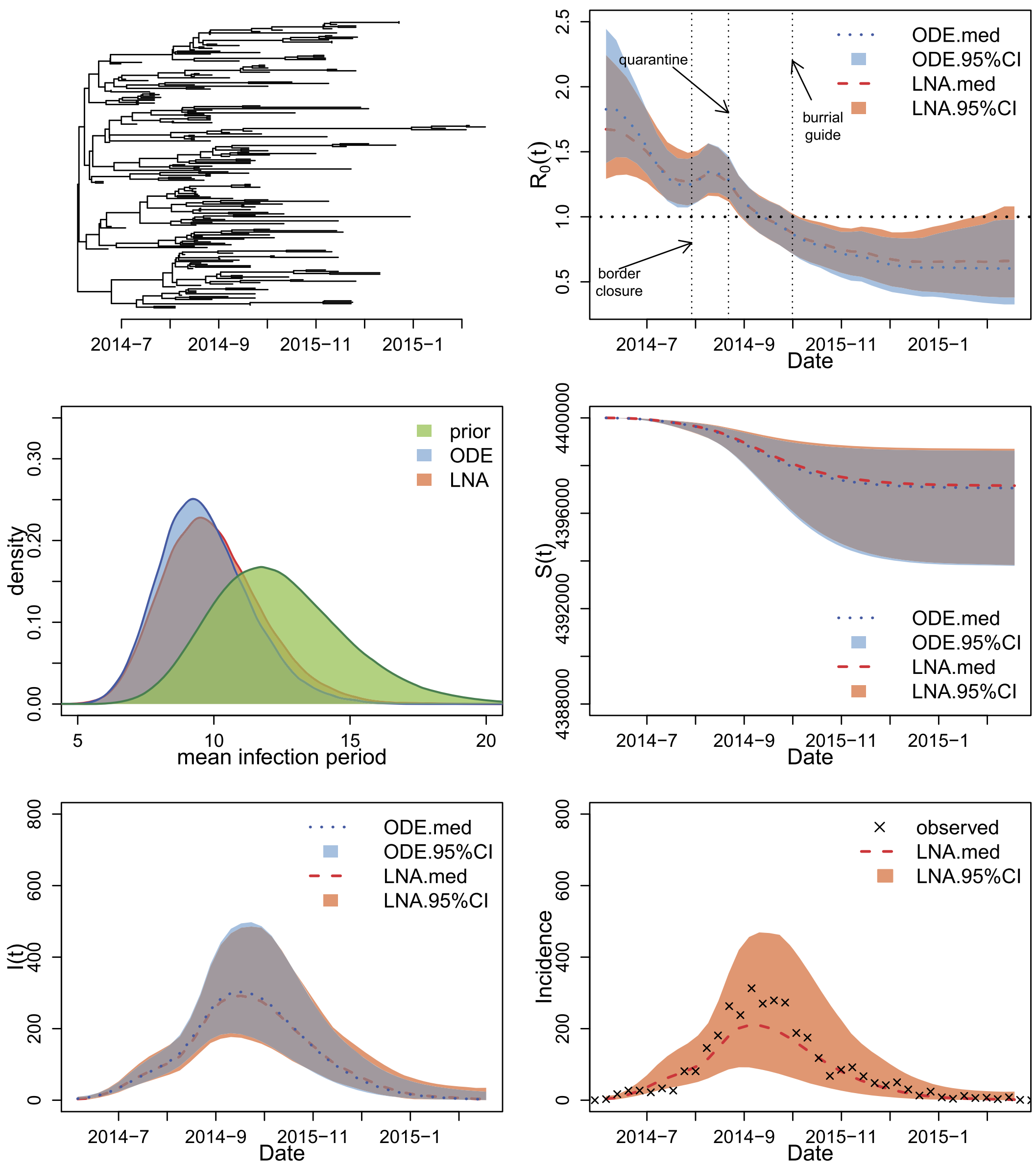

FIG. 6. *Analysis of the genealogy relating Ebola virus sequences collected in Liberia*; *see caption in Figure* 5 *for the explanation of the plots*.

els. For example, Volz and Pond (2014) used a SEIR model with a constant R_0 and estimated it to be 2.40 (CI: $(1.54, 3.87)$). Althaus (2014) assumed an exponentially decaying $R_0(t)$ with an estimated initial R_0 of 2.52 (CI: $(2.41, 2.67)$). The discrepancies between our and SEIR-based estimates are not unexpected, because SEIR models generally yield higher R_0 estimates than SIR models when applied to the same dataset (Keeling and Rohani (2011), Wearing, Rohani and Keeling (2005)). Our estimated R_0 for Liberia is in agreement with results of Althaus (2014), who fitted a SEIR model to incidence data and arrived at an estimated R_0 of 1.59 (CI: (1.57, 1.60)).

The $R_0(t)$ dynamics in the two countries share a similar pattern with: (1) a decreasing trend that starts in Spring/Summer of 2014, (2) a stable/constant period until the end of September 2014, and (3) a final decrease below 1.0 (epidemic is contained) around November 2014. Since the number of susceptible individuals did not change significantly over the course of the epidemic, relative to the total population size, the basic and effective reproduction numbers, $R_0(t) = \beta(t)N/\gamma$ and $R_{\text{eff}}(t) = \beta(t)S(t)/\gamma$, are approximately equal. This allows us to compare our $R_0(t)$ estimation results with previously estimated changes in $R_{\text{eff}}(t)$. Our estimation of early $R_0(t)$ dynamics in Sierra Leone agrees with results of Stadler et al. (2013), who concluded that the effective reproduction number did not significantly decrease until mid-June. Our estimated $R_0(t)$ trajectory suggests that later interventions, such as border closures and release of burial guides, may have been helpful in controlling the spread of the disease. The infectious period for Sierra Leone epidemic is estimated to be 10.8 days with a 95% BCI (7.6, 15.6). For Liberia the infection period has a point estimate of 9.8 with a 95% BCI $(6.87, 14.05)$. The posterior median of the total number of infected individuals (final epidemic size) is 7450 and its 95% BCI is (3495, 15518) for Sierra Leone, which is close to 8706 total confirmed number of cases reported by CDC (2019). Liberia had a smaller epidemic than Sierra Leone, with estimated total infected individuals being 2842 and a 95% BCI of $(1296, 6173)$. These results are also in agreement with 3163 total confirmed cases from CDC.

We perform an out-of-sample validation by comparing our results with weekly reported confirmed incidence in Sierra Leone and Liberia from WHO (2016). The posterior predictive weekly incidence at time t, denoted by $\hat{N}(t)$, is approximated by

$$\hat{N}(t) = \hat{\beta}(t)\hat{S}(t)\hat{I}(t) \cdot \Delta t, \tag{19}$$

where $\hat{\beta}(t)$, $\hat{S}(t)$, and $\hat{I}(t)$ are the posterior estimates of the infection rate, number of susceptible, and number of infected individuals at time t, respectively, and $\Delta t := 7/365$ corresponds the time interval of one week. We plot the posterior predictive estimates of weekly incidence together with the corresponding weekly reported confirmed incidence. For both countries, our model-based incidence 95% BCIs cover the reported incidence counts from WHO, suggesting that our time-varying SIR model can estimate incidence well from genetic data alone. Because not all Ebola cases were reported and recorded, we note that our estimated latent incidence should be greater than the reported incidence. However, the discrepancy between latent and reported incidence should not be large, because Ebola reporting rate was high. For example, Scarpino et al. (2014) estimated that 83% of Ebola cases were reported.

We also report results from the ODE-based method and superimpose these results over LNA-based results on Figures 5 and 6. For the relatively small Liberia genealogy, the ODE-based and LNA-based methods yield similar parameter estimates. However, the larger Sierra Leone genealogy produces substantial differences between ODE-based and LNA-based estimates of the $R_0(t)$. The ODE-based method captures the decreasing trend of $R_0(t)$ in Spring and Summer of 2014 but provides narrow BCIs with unrealistic short-term fluctuations in the basic reproduction number trajectory.

5. Discussion. In this paper we propose a Bayesian phylodynamic inference method that can fit a stochastic epidemic model to an observed genealogy estimated from infectious disease genetic sequences sampled during an outbreak. Our statistical model can be viewed as semiparametric with: (1) a parametric SIR model describing the infectious disease dynamics and (2) a nonparametric GMRF-based estimation of the time-varying basic reproduction number. To the best of our knowledge, this is the first method combining a Bayesian nonparametric approach with a deterministic or stochastic SIR model for phylodynamic inference (although, see Xu, Kypraios and O'Neill (2016) for a similar approach applied to more traditional epidemiological data). Our use of LNA allows us to devise an efficient MCMC algorithm to approximate high-dimensional posterior distribution of model parameters and latent variables. Our LNA-based method produces posterior summaries with better frequentist properties than the state-of-the-art ODE-based method, underscoring the importance of working with stochastic models, even in large populations. We showcase our method by applying it to the Ebola genealogies estimated from viral sequences collected in Sierra Leone and Liberia during the 2014–2015 outbreak. Our nonparametric estimates of $R_0(t)$ in Sierra Lione and Liberia suggest that the basic reproduction number decreased in two-stages, where the second stage brought it below 1.0—a sign of epidemic containment. The second stage of $R_0(t)$ decrease closely follows in time implementation of interventions, pointing to their effectiveness.

Our method relies on the assumption that population is well mixed and the population dynamics follow a SIR model. However, it may be desirable to be able to relax these assumptions. For example, in Ebola spread modeling some authors used a SEIR model that assumes a latent period during which an infected individual is not infectious (Althaus (2014), Volz and Siveroni (2018)). Moreover, adding more compartments should allow us to partially relax the unrealistic assumption of homogeneous mixing. For example, stratifying compartments by age group would allow us to account for different contact rates between these groups. One future direction of this work is to generalize the LNA-based method to fit complicated compartmental epidemic models, including models with multistage infections, like SEIR model and models with the population stratified by sex, age, geographic location, or other demographic variables. The structured coalescent likelihoods under these models may not have closed-form expressions. However, Volz (2012), Dearlove and Wilson (2013), and Müller, Rasmussen and Stadler (2017) propose several strategies to approximate structured coalescent likelihoods. Our LNA-based methodology is directly portable to these approximate structured coalescent likelihood approaches, but our current implementation lacks this generality. We hope to remedy this in our future work.

The experiments in Section 3.1 indicate that one has to pay close attention to parameter identifiability when fitting SIR models to genealogies or to sequence data directly. Identifiability may not be a problem under an assumption of a constant $R_0(t)$. However, the removal rate tends to be only weakly identifiable in the scenarios with a time-varying basic reproduction number in which the estimation can be sensitive to the choice of priors. In Section 3.2 and Appendix Section A-6, we demonstrate that putting a weakly informative prior on the removal rate can cause bias not only in the estimation for removal rate but also can lead to a failure in recovering the reproduction number and latent population dynamics. Therefore, successful inference of SIR model parameters, using genealogical data, should rely on a sound informative prior for the removal rate. This constraint is not a big shortcoming in situations where prior information about the removal rate, or mean length of the infectious period, is available from patient hospitalization data (Team (2014)).

Since parameter identifiability is a recurring problem in infectious disease modeling, integration of multiple sources of information is of great interest. Using particle filter MCMC, Rasmussen, Ratmann and Koelle (2011) demonstrated that jointly analyzing genealogy and incidence case counts considerably reduces the uncertainty in both estimation of latent population trajectory and SIR model parameters, compared with estimation based on a single source of information. We plan to use our LNA-based framework to perform similar integration of genealogical data and incidence time series. Another possible source of information is the distribution of genetic sequence sampling times. Karcher et al. (2016) proposed a preferential sampling approach that explicitly models dependence of the sampling times distribution on the effective population size. The authors demonstrated that accounting for preferential sampling helps decrease bias and results in more precise effective population size estimation. It would be interesting to incorporate preferential sampling into our LNA-based framework by assuming a probabilistic dependency between sampling times and latent prevalence $I(t)$.

Our method assumes a genealogy/phylogenetic tree is given to us. In reality, genealogies are not directly observed and need to be inferred from molecular sequences. Genealogy estimation remains one of the biggest computational bottlenecks in phylodynamics with computational burden of such estimation being typically higher than the burden of phylodynamics methods that use the genealogy as input. Ideally, uncertainty in the genealogy should be handled by building a Bayesian hierarchical model and integrating over the space of genealogies using MCMC. In fact, implementations of such Bayesian hierarchical modeling already exist for nonparametric, birth-death, and ODE-based phylodynamic approaches (Drummond et al. (2005), Minin, Bloomquist and Suchard (2008), Volz and Siveroni (2018), Gill et al. (2013), Stadler et al. (2013)). Therefore, an important future direction will be to extend our LNA framework to fitting stochastic epidemic models to molecular sequences, instead of genealogies. Similarly to the structured coalescent model implementation of Volz and Siveroni (2018), the easiest way to achieve this will be integration of our LNA MCMC algorithm into popular open source phylogenetic/phylodynamic software packages, such as `BEAST`, `BEAST2`, and `RevBayes` (Suchard et al. (2018), Bouckaert et al. (2014), Höhna et al. (2016)).

Acknowledgments. We thank Jon Fintzi for discussing the details of implementing linear noise approximation. We are grateful to Michael Karcher and Julia Palacios for patiently answering questions about nonparametric phylodynamic inference and `phylodyn` package.

Funding. M.T and V.N.M. were supported by the NIH grant R01 AI107034. M.T., T.B., and V.N.M. were supported by the NIH grant U54 GM111274. V.N.M. was supported by the NSF grand DMS-1936833. G.D. was supported by the Mahan postdoctoral fellowship from the Fred Hutchinson Cancer Research Center. This work was supported by NIH grant R35 GM119774-01 from the NIGMS to TB. TB is a Pew Biomedical Scholar.

SUPPLEMENTARY MATERIAL

Appendix (DOI: 10.1214/21-AOAS1583SUPPA; .pdf). Additional tables, figures, and details are included in the Appendix.

Code (DOI: 10.1214/21-AOAS1583SUPPB; .zip). The zip archive contains code implementing our MCMC algorithms (R package) and scripts that allow for reproducing paper results.

REFERENCES

Althaus, C. L. (2014). Estimating the reproduction number of Ebola virus (EBOV) during the 2014 outbreak in West Africa. *PLoS Curr.* **6**. https://doi.org/10.1371/currents.outbreaks.91afb5e0f279e7f29e7056095255b288

Anderson, R. and May, R. (1992). *Infectious Diseases of Humans*: *Dynamics and Control* **28**. Wiley, New York.

Bailey, N. T. J. (1975). *The Mathematical Theory of Infectious Diseases and Its Applications*, 2nd ed. Hafner Press, New York. MR0452809

Bouckaert, R., Heled, J., Kühnert, D., Vaughan, T., Wu, C., Xie, D., Suchard, M., Rambaut, A. and Drummond, A. (2014). BEAST 2: A software platform for Bayesian evolutionary analysis. *PLoS Comput*. *Biol*. **10** 1–6.

Buckingham-Jeffery, E., Isham, V. and House, T. (2018). Gaussian process approximations for fast inference from infectious disease data. *Math*. *Biosci*. **301** 111–120. MR3808358 https://doi.org/10.1016/j.mbs.2018.02.003

Dearlove, B. and Wilson, D. (2013). Coalescent inference for infectious disease: Meta-analysis of hepatitis C. *Philos*. *Trans*. *R*. *Soc*. *B* **368** 20120314.

Donnelly, P. and Tavare, S. (1995). Coalescents and genealogical structure under neutrality. *Annu*. *Rev*. *Genet*. **29** 401–421.

DRUMMOND, A., NICHOLLS, G., RODRIGO, A. and SOLOMON, W. (2002). Estimating mutation parameters, population history and genealogy simultaneously from temporally spaced sequence data. *Genetics* **161** 1307–1320.

DRUMMOND, A., RAMBAUT, A., SHAPIRO, B. and PYBUS, O. (2005). Bayesian coalescent inference of past population dynamics from molecular sequences. *Mol. Biol. Evol.* **22** 1185–1192.

DUDAS, G., CARVALHO, L., BEDFORD, T., TATEM, A., BAELE, G., FARIA, N., PARK, D., LADNER, J., ARIAS, A. et al. (2017). Virus genomes reveal factors that spread and sustained the Ebola epidemic. *Nature* **544** 309–315.

FEARNHEAD, P., GIAGOS, V. and SHERLOCK, C. (2014). Inference for reaction networks using the linear noise approximation. *Biometrics* **70** 457–466. MR3258050 https://doi.org/10.1111/biom.12152

FROST, S. D. and VOLZ, E. M. (2010). Viral phylodynamics and the search for an 'effective number of infections'. *Philos. Trans. R. Soc. Lond. B, Biol. Sci.* **365** 1879–1890.

GIAGOS, V. (2010). Inference for Auto-Regulatory Genetic Networks Using Diffusion Process Approximations Ph.D. thesis Lancaster Univ.

GILL, M., LEMEY, P., FARIA, N., RAMBAUT, A., SHAPIRO, B. and SUCHARD, M. (2013). Improving Bayesian population dynamics inference: A coalescent-based model for multiple loci. *Mol. Biol. Evol.* **30** 713–724.

GILLESPIE, D. (1977). Exact stochastic simulation of coupled chemical reactions. *J. Phys. Chem.* **81** 2340–2361.

GRENFELL, B., PYBUS, O., GOG, J., WOOD, J., DALY, J., MUMFORD, J. and HOLMES, E. (2004). Unifying the epidemiological and evolutionary dynamics of pathogens. *Science* **303** 327–332.

GRIFFITHS, R. and TAVARÉ, S. (1994). Sampling theory for neutral alleles in a varying environment. *Philos. Trans. R. Soc. Lond. B, Biol. Sci.* **344** 403–410.

HÖHNA, S., LANDIS, M., HEATH, T., BOUSSAU, B., LARTILLOT, N., MOORE, B., HUELSENBECK, J. and RONQUIST, F. (2016). RevBayes: Bayesian phylogenetic inference using graphical models and an interactive model-specification language. *Syst. Biol.* **65** 726–736.

JOMBART, T., CORI, A., DIDELOT, X., CAUCHEMEZ, S., FRASER, C. and FERGUSON, N. (2014). Bayesian reconstruction of disease outbreaks by combining epidemiologic and genomic data. *PLoS Comput. Biol.* **10** e1003457. https://doi.org/10.1371/journal.pcbi.1003457

KARCHER, M., PALACIOS, J., BEDFORD, T., SUCHARD, M. and MININ, V. (2016). Quantifying and mitigating the effect of preferential sampling on phylodynamic inference. *PLoS Comput. Biol.* **12** e1004789.

KEELING, M. and ROHANI, P. (2011). *Modeling Infectious Diseases in Humans and Animals*. Princeton Univ. Press, Princeton, NJ.

KINGMAN, J. F. C. (1982). The coalescent. *Stochastic Process. Appl.* **13** 235–248. MR0671034 https://doi.org/10.1016/0304-4149(82)90011-4

KLINKENBERG, D., BACKER, J. A., DIDELOT, X., COLIJN, C. and WALLINGA, J. (2017). Simultaneous inference of phylogenetic and transmission trees in infectious disease outbreaks. *PLoS Comput. Biol.* **13** e1005495. https://doi.org/10.1371/journal.pcbi.1005495

KOMOROWSKI, M., FINKENSTÄDT, B., HARPER, C. V. and RAND, D. A. (2009). Bayesian inference of biochemical kinetic parameters using the linear noise approximation. *BMC Bioinform.* **10** 343. https://doi.org/10.1186/1471-2105-10-343

KUHNER, M., YAMATO, J. and FELSENSTEIN, J. (1998). Maximum likelihood estimation of population growth rates based on the coalescent. *Genetics* **149** 429–434.

KÜHNERT, D., STADLER, T., VAUGHAN, T. G. and DRUMMOND, A. J. (2014). Simultaneous reconstruction of evolutionary history and epidemiological dynamics from viral sequences with the birth-death SIR model. *J. R. Soc. Interface* **11** 20131106. https://doi.org/10.1098/rsif.2013.1106

KURTZ, T. G. (1970). Solutions of ordinary differential equations as limits of pure jump Markov processes. *J. Appl. Probab.* **7** 49–58. MR0254917 https://doi.org/10.2307/3212147

KURTZ, T. G. (1971). Limit theorems for sequences of jump Markov processes approximating ordinary differential processes. *J. Appl. Probab.* **8** 344–356. MR0287609 https://doi.org/10.1017/s002190020003535x

LEVENTHAL, G., GÜNTHARD, H., BONHOEFFER, S. and STADLER, T. (2013). Using an epidemiological model for phylogenetic inference reveals density dependence in HIV transmission. *Mol. Biol. Evol.* **31** 6–17.

MININ, V., BLOOMQUIST, E. and SUCHARD, M. (2008). Smooth skyride through a rough skyline: Bayesian coalescent-based inference of population dynamics. *Mol. Biol. Evol.* **25** 1459–1471.

MÜLLER, N. F., RASMUSSEN, D. A. and STADLER, T. (2017). The structured coalescent and its approximations. *Mol. Biol. Evol.* **34** 2970–2981. https://doi.org/10.1093/molbev/msx186

MURRAY, I., ADAMS, R. and MACKAY, D. (2010). Elliptical slice sampling. In *AISTATS* **13** 541–548.

O'NEILL, P. and ROBERTS, G. (1999). Bayesian inference for partially observed stochastic epidemics. *J. Roy. Statist. Soc. Ser. A* **162** 121–129.

PALACIOS, J. A. and MININ, V. N. (2013). Gaussian process-based Bayesian nonparametric inference of population size trajectories from gene genealogies. *Biometrics* **69** 8–18. MR3058047 https://doi.org/10.1111/biom.12003

PAPASPILIOPOULOS, O., ROBERTS, G. O. and SKÖLD, M. (2007). A general framework for the parametrization of hierarchical models. *Statist*. *Sci*. **22** 59–73. MR2408661 https://doi.org/10.1214/088342307000000014

PYBUS, O., CHARLESTON, M., GUPTA, S., RAMBAUT, A., HOLMES, E. and HARVEY, P. (2001). The epidemic behavior of the hepatitis C virus. *Science* **292** 2323–2325.

RASMUSSEN, D. A., RATMANN, O. and KOELLE, K. (2011). Inference for nonlinear epidemiological models using genealogies and time series. *PLoS Comput*. *Biol*. **7** e1002136. MR2845064 https://doi.org/10.1371/journal.pcbi.1002136

RASMUSSEN, D. A., VOLZ, E. M. and KOELLE, K. (2014). Phylodynamic inference for structured epidemiological models. *PLoS Comput*. *Biol*. **10** e1003570. https://doi.org/10.1371/journal.pcbi.1003570

RUE, H. (2001). Fast sampling of Gaussian Markov random fields. *J*. *R*. *Stat*. *Soc*. *Ser*. *B*. *Stat*. *Methodol*. **63** 325–338. MR1841418 https://doi.org/10.1111/1467-9868.00288

RUE, H. and HELD, L. (2005). *Gaussian Markov Random Fields*: *Theory and Applications*. *Monographs on Statistics and Applied Probability* **104**. CRC Press/CRC, Boca Raton, FL. MR2130347 https://doi.org/10.1201/9780203492024

SCARPINO, S., IAMARINO, A., WELLS, C., YAMIN, D., NDEFFO-MBAH, M., WENZEL, N., FOX, S., NYENSWAH, T., ALTICE, F. et al. (2014). Epidemiological and viral genomic sequence analysis of the 2014 Ebola outbreak reveals clustered transmission. *Clin*. *Infect*. *Dis*. **60** 1079–1082.

SMITH, R. A., IONIDES, E. L. and KING, A. A. (2017). Infectious disease dynamics inferred from genetic data via sequential Monte Carlo. *Mol*. *Biol*. *Evol*. **34** 2065–2084. https://doi.org/10.1093/molbev/msx124

STADLER, T., KÜHNERT, D., BONHOEFFER, S. and DRUMMOND, A. (2013). Birth–death skyline plot reveals temporal changes of epidemic spread in HIV and hepatitis C virus (HCV). *Proc*. *Natl*. *Acad*. *Sci*. *USA* **110** 228–233.

STADLER, T., KÜHNERT, D., RASMUSSEN, D. and DU PLESSIS, L. (2014). Insights into the early epidemic spread of Ebola in Sierra Leone provided by viral sequence data. *PLoS Curr*. **6**.

SUCHARD, M. A., LEMEY, P., BAELE, G., AYRES, D. L., DRUMMOND, A. J. and RAMBAUT, A. (2018). Bayesian phylogenetic and phylodynamic data integration using BEAST 1.10. *Virus Evol*. **4** vey016. https://doi.org/10.1093/ve/vey016

TANG, M., DUDAS G., BEDFORD, T. and N. MININ, V. (2023). Supplement to "Fitting stochastic epidemic models to gene genealogies using linear noise approximation." https://doi.org/10.1214/21-AOAS1583SUPPA, https://doi.org/10.1214/21-AOAS1583SUPPB

TEAM, W. E. R. (2014). Ebola virus disease in West Africa—the first 9 months of the epidemic and forward projections. *N*. *Engl*. *J*. *Med*. **371** 1481–1495.

TOWERS, S., PATTERSON-LOMBA, O. and CASTILLO-CHAVEZ, C. (2014). Temporal variations in the effective reproduction number of the 2014 West Africa Ebola outbreak. *PLoS Curr*. **6**. https://doi.org/10.1371/currents.outbreaks.9e4c4294ec8ce1adad283172b16bc908

VAN KAMPEN, N. and REINHARDT, W. (1981). *Stochastic Processes in Physics and Chemistry*. North-Holland, Amsterdam. MR0648937

VAUGHAN, T. G., LEVENTHAL, G. E., RASMUSSEN, D. A., DRUMMOND, A. J., WELCH, D. and STADLER, T. (2019). Estimating epidemic incidence and prevalence from genomic data. *Mol*. *Biol*. *Evol*. **36** 1804–1816. https://doi.org/10.1093/molbev/msz106

VOLZ, E. M. (2012). Complex population dynamics and the coalescent under neutrality. *Genetics* **190** 187–201. https://doi.org/10.1534/genetics.111.134627

VOLZ, E. M., KOELLE, K. and BEDFORD, T. (2013b). Viral phylodynamics. *PLoS Comput*. *Biol*. **9** e1002947. MR3048921 https://doi.org/10.1371/journal.pcbi.1002947

VOLZ, E. and POND, S. (2014). Phylodynamic analysis of Ebola virus in the 2014 Sierra Leone epidemic. *PLoS Curr*. **6**. https://doi.org/10.1371/currents.outbreaks.6f7025f1271821d4c815385b08f5f80e

VOLZ, E. and SIVERONI, I. (2018). Bayesian phylodynamic inference with complex models. *PLoS Comput*. *Biol*. **14** e1006546.

VOLZ, E., POND, S., WARD, M., BROWN, A. and FROST, S. (2009). Phylodynamics of infectious disease epidemics. *Genetics* **183** 1421–1430.

WALLACE, E. (2010). A simplified derivation of the linear noise approximation. Arxiv preprint. Available at arXiv:1004.4280.

WEARING, H. J., ROHANI, P. and KEELING, M. J. (2005). Appropriate models for the management of infectious diseases. *PLoS Med*. **2** e174. https://doi.org/10.1371/journal.pmed.0020174

WILKINSON, D. (2011). *Stochastic Modelling for Systems Biology*. CRC press, Boca Raton, FL.

WRIGHT, S. (1931). Evolution in Mendelian populations. *Genetics* **16** 97–159. https://doi.org/10.1093/genetics/16.2.97

XU, X., KYPRAIOS, T. and O'NEILL, P. D. (2016). Bayesian non-parametric inference for stochastic epidemic models using Gaussian processes. *Biostatistics* **17** 619–633. MR3604269 https://doi.org/10.1093/biostatistics/kxw011

YPMA, R. J. F., VAN BALLEGOOIJEN, W. M. and WALLINGA, J. (2013). Relating phylogenetic trees to transmission trees of infectious disease outbreaks. *Genetics* **195** 1055–1062.

CENTERS FOR DISEASE CONTROL AND PREVENTION (2019). 2014–2016 Ebola outbreak in West Africa. https://www.cdc.gov/vhf/ebola/history/2014-2016-outbreak/index.html. Last accessed: Oct, 09, 2022.

WORLD HEALTH ORGANIZATION (2016). Ebola data and statistics. http://apps.who.int/gho/data/node.ebola-sitrep.quick-downloads?lang=en. Last accessed: February 28, 2018.