Inferring epidemiological dynamics with Bayesian coalescent inference: The merits of deterministic and stochastic models

Alex Popinga[*†], Tim Vaughan[*†‡], Tanja Stadler[§], Alexei J Drummond[*†]

December 19, 2014

[*]Department of Computer Science, University of Auckland, Auckland, New Zealand 1010

[†]Allan Wilson Centre for Molecular Ecology and Evolution, New Zealand PN4442

[‡]Massey University, Palmerston North, New Zealand 4442

[§]Department of Biosystems Science and Engineering, ETH Zürich, Basel, Switzerland

Running Head: Bayesian Coalescent Epidemic Inference

Key Words: Bayesian inference, Phylodynamics, Coalescent, Epidemic, Stochastic

Corresponding Author:

Alexei J Drummond

Department of Computer Science

University of Auckland

303 - 379

38 Princes St

Auckland, NZ 1010

(64 9) 373-7599 88298

alexei@cs.auckland.ac.nz

Abstract

Estimation of epidemiological and population parameters from molecular sequence data has become central to the understanding of infectious disease dynamics. Various models have been proposed to infer details of the dynamics that describe epidemic progression. These include inference approaches derived from Kingman's coalescent theory. Here, we use recently described coalescent theory for epidemic dynamics to develop stochastic and deterministic coalescent SIR tree priors. We implement these in a Bayesian phylogenetic inference framework to permit joint estimation of SIR epidemic parameters and the sample genealogy. We assess the performance of the two coalescent models and also juxtapose results obtained with BDSIR, a recently published birth-death-sampling model for epidemic inference. Comparisons are made by analyzing sets of genealogies simulated under precisely known epidemiological parameters. Additionally, we analyze influenza A (H1N1) sequence data sampled in the Canterbury region of New Zealand and HIV-1 sequence data obtained from known UK infection clusters. We show that both coalescent SIR models are effective at estimating epidemiological parameters from data with large fundamental reproductive number R_0 and large population size S_0. Furthermore, we find that the stochastic variant generally outperforms its deterministic counterpart in terms of error, bias, and highest posterior density coverage, particularly for smaller R_0 and S_0. However, each of these inference models are shown to have undesirable properties in certain circumstances, especially for epidemic outbreaks with R_0 close to one or with small effective susceptible populations.

Phylodynamics and the coalescent The epidemiological and evolutionary processes that underpin rapidly evolving species occur on a shared spatiotemporal frame of reference. Unified analyses that include both the dynamics of an epidemic and the reconstruction of the pathogen phylogeny can therefore uncover otherwise inaccessible information to aid in outbreak prevention. Such information includes the rates of pathogen transmission and host recovery, effective population sizes, and the 'time of origin' representing the introduction of the first infected individual into a population of susceptible hosts.

The term *phylodynamics* was popularized by Grenfell *et al.* (2004) to describe the interlaced study of immunodynamics, epidemiology, and evolutionary mechanisms. Several phylodynamic models, both stochastic and deterministic in nature, have since been developed to characterize the phylogenetic history of the pathogen species and compartmentalizations of the host population throughout the epidemic. Such models grant the ability to infer key epidemiological parameters from genetic sequence data and include birth-death branching processes (Stadler *et al.* 2012, 2013; Kühnert *et al.* 2014; Gavryushkina *et al.* 2014), as well as coalescent approaches (Griffiths and Tavaré 1994; Pybus *et al.* 2001a; Koelle and Rasmussen 2012; Rasmussen *et al.* 2011; Dearlove and Wilson 2013; Rasmussen *et al.* 2014) derived from Kingman's coalescent theory (Kingman 1982).

Significant steps toward the unification of epidemiology and statistical phylogenetics were made by Pybus *et al.* (2001b), Dearlove and Wilson (2013), and Volz *et al.* (2009), with the formalization and application of Kingman's n-coalescent to pathogen population dynamics. These methods involved numerical integration of a set of ordinary differential equations (ODEs) to find deterministic approximations to the variation in the number of sampled lineages through time. Volz (2012) extended the tree density calculation from previous work (Volz *et al.* 2009) to allow for serially-sampled and spatially structured genetic sequence data. In this coalescent model, the birth and death rates can vary in time and by the state of the host, so that "the birth rate of a single gene copy is both time- and

state-dependent."

In this paper, we assess the ability of coalescent-based phylodynamic models to infer, in a Bayesian setting, a range of epidemiological parameters from simulated data. While Dearlove and Wilson (2013) paved the way by implementing a coalescent approach for deterministic SI, SIS, and SIR models for Bayesian inference, we implement and rigorously test both deterministic and stochastic coalescent SIR models of epidemic dynamics extended for heterochronously sampled data.

Stochastic and deterministic models Stochasticity and determinism in population sizes each maintain dominant roles in particular stages of an epidemic. Once the infected population has grown considerably large, on the order of 1,000 to 10,000 lineages, the probability densities of stochastically-expressed population size dynamics converge toward the deterministic interpretation (Rouzine *et al.* 2001). However, during the early stages the population size of infected individuals is small, and the dynamics of the epidemic are therefore governed by stochastic processes due to the relative significance of fluctuations in the demographic and rate parameters of the population model (Kühnert *et al.* 2014). Therefore, approximating the prevalence of infection by a deterministic function requires the number of infected hosts within the effective population to be assumed as very large throughout the duration of the described epidemic, i.e., once the exponential growth phase has been reached (Rouzine *et al.* 2001).

Population size is critical to the epidemiological system and, as with any parameter in a Bayesian setting, yields the most accurate estimations when detailed prior information is available and incorporated into the inference (Drummond *et al.* 2006). In our extension and implementation of the coalescent model for epidemics, both stochastic and deterministic population size processes are used for the simulation of trees and/or trajectories for subsequent inference.

Compartmental population models (SIR) Host populations can be compartmentalized simply but effectively in mathematical models that describe epidemic progression. The specific division of the aggregate population depends on the contagion, spanning a range of scenarios where hosts may or may not recover from infection, may or may not be reinfected, etc. Such examples include the SI (Susceptible-Infected), SIS (Susceptible-Infected-Susceptible), and SIR (Susceptible-Infected-Removed) models (Anderson and May 1991; Keeling and Rohani 2008). Each of these compartments can be expressed either (a) by a set of ODEs that describe the deterministic time development of real-valued compartment occupancies, or (b) in terms of integer-valued occupancies governed by continuous-time Markov chains (CTMC) that allow for a degree of uncertainty in the timing and number of events that occur over the course of the epidemic.

In this paper, we concentrate on the SIR model, which describes epidemics that include infected individuals who are at some point in time removed from the effective population by way of immunity, death, behavioral changes, or some other termination of infectiousness. The deterministic variant of this model was introduced by Kermack and McKendrick (1932) and is given by the trio of coupled ODEs,

$$\frac{d}{dt}S(t) = -\beta I(t)S(t), \tag{1}$$

$$\frac{d}{dt}I(t) = \beta I(t)S(t) - \gamma I(t), \tag{2}$$

$$\frac{d}{dt}R(t) = \gamma I(t), \tag{3}$$

where β and γ respectively represent the transition rates from susceptible S to infected I, and infected I to removed R. The model fully defines the population dynamics with initial conditions $S(z_0)$, $I(z_0)$, and $R(z_0)$. It is worth recognizing that, in the closed SIR model used here, there is no demographic change in the host population. Therefore, $\frac{d}{dt}S(t) + \frac{d}{dt}I(t) + \frac{d}{dt}R(t) = 0$ and $S(t) + I(t) + R(t) = N$, where N is the constant total population size. Throughout this paper we refer to the solutions to eq. (1-3) as *deterministic SIR*

trajectories.

The comparable stochastic description is given in terms of the probability of the epidemic state at time t given its initial state and the rate parameters

$$\pi(s,i,r;t) \equiv \Pr(S(t)=s, I(t)=i, R(t)=r | S(0), I(0), R(0), \beta, \gamma), \tag{4}$$

which is governed by the following equation of motion:

$$\begin{aligned}\frac{d}{dt}\pi(s,i,r;t) =& \beta\left[(s+1)(i-1)\pi(s+1,i-1,r;t) - si\pi(s,i,r;t)\right] \\ &+ \gamma\left[(i+1)\pi(s,i+1,r-1;t) - i\pi(s,i,r;t)\right].\end{aligned} \tag{5}$$

An explicit sampling process is incorporated by allowing each removal event to coincide with a sampling event with a fixed probability $\psi/(\psi+\mu)$ where ψ and μ are the overall rates of sampled and unsampled removals, respectively, such that $\gamma = \psi + \mu$. We refer to epidemic histories sampled from this model as *stochastic SIR trajectories*.

Both types of epidemic trajectories can be related to models of sampled transmission tree genealogies. In the deterministic case, this relationship is made via the coalescent distributions described in Volz (2012). We call this the *deterministic coalescent SIR model*. In the stochastic case, genealogies appear naturally from a branching process in which the branching events coincide with the transmission events in the CTMC and only those lineages ancestral to sampled removals are recorded. We call this the *stochastic SIR model*. The *BDSIR model* introduced by Kühnert *et al.* (2014) provides an approximation to the stochastic SIR model.

Another way of relating the stochastic SIR model to sampled transmission trees involves drawing a realization of a stochastic SIR epidemic, then using the coalescent distribution in Volz (2012) to produce a tree conditional on the particular piecewise constant infected compartment size corresponding to that realization. We call this approach the *stochastic*

coalescent SIR model. Unlike BDSIR, the stochastic coalescent SIR model does not require the sampling process to be specified explicitly.

Both the transmission rate β and removal rate γ can be estimated using each of the methods considered in this paper from data ascribed to an SIR epidemic.

METHODS

Inference framework All phylodynamic inference discussed in this paper is based on the joint posterior probability density

$$f(\mathcal{T}, \mathcal{V}, \eta, \theta | D) = \frac{\Pr(D|\mathcal{T}, \theta) f(\mathcal{T}|\mathcal{V}, \eta) f(\mathcal{V}|\eta) f(\eta) f(\theta)}{\Pr(D)}, \tag{6}$$

where the sampled transmission tree $\mathcal{T}$, the epidemic trajectory denoted $\mathcal{V} = (\mathcal{S}, \mathcal{I}, \mathcal{R})$, the substitution parameters θ, and the epidemiological parameters $\eta = \{\beta, \gamma, S_0, z_0\}$ are all estimated from the sequence data. The sampled transmission tree $\mathcal{T}$ is assumed to be identical to the pathogen genealogy.

Here, $\mathcal{S}$, $\mathcal{I}$, and $\mathcal{R}$ represent the host compartment sizes from the present time $\tau = 0$ back to the origin z_0, such that: $\mathcal{S}(\tau) = S(z_0 - \tau)$, $\mathcal{I}(\tau) = I(z_0 - \tau)$, and $\mathcal{R}(\tau) = R(z_0 - \tau)$.

The various terms making up the right-hand side of eq. (6) are the tree likelihood $\Pr(D|\mathcal{T}, \theta)$, the tree prior $f(\mathcal{T}|\mathcal{V}, \eta)$, the epidemic trajectory density $f(\mathcal{V}|\eta)$, and the substitution and epidemiological parameter priors $f(\eta)$ and $f(\theta)$. The probability $\Pr(D)$ is merely a normalizing constant and can be ignored. It is the product of the tree prior and trajectory density $f(\mathcal{T}|\mathcal{V}, \eta) f(\mathcal{V}|\eta)$ that distinguishes each of the models considered in this paper.

For both the deterministic and stochastic coalescent SIR models, the tree prior $f(\mathcal{T}|\mathcal{V}, \eta)$ is calculated in the following way. First, consider the time span of a tree divided into segments bracketed by both sampling and coalescent events. By considering intervals ending in sampling events as well as coalescent-ending intervals, we follow previous work that extended coalescent approaches to time-stamped, serially-sampled data (Rodrigo and Felsenstein

1999; Drummond *et al.* 2002). Interval i is spanned by k_i lineages and is the i'th interval when ordered from the most recent tip to the root. The set of intervals A ending in sample events and the set of intervals Y ending in coalescent events together encompass all intervals, $V = A \cup Y$. Let the end time of an interval be τ_i (going back in time), with $\tau_0 = 0$ as the time of the most recent tip and with time increasing into the past. Then the probability density of a genealogy given an epidemic trajectory is

$$f(\mathcal{T}|\mathcal{V}, \eta) = \prod_{i \in Y} \lambda_{k_i}(\tau_i) \prod_{i \in V} \omega(\tau_i, k_i), \tag{7}$$

where $\lambda_{k_i}(\tau)$ is the instantaneous coalescent rate at τ prescribed by Volz (2012)

$$\lambda_{k_i}(\tau) = \binom{k_i}{2} \frac{2\beta \mathcal{S}(\tau)}{\mathcal{I}(\tau)}, \tag{8}$$

and where $\omega(\tau_i, k_i)$ is the survival probability

$$\omega(\tau_i, k_i) = \exp\left(-\int_{\tau_{i-1}}^{\tau_i} \lambda_{k_i}(\tau) d\tau \right). \tag{9}$$

The deterministic coalescent SIR model assumes that the SIR epidemic trajectories are found by integrating the ODEs in eqs. (1)–(3). Therefore, under this model each epidemic trajectory is a deterministic function of its parameters $\mathcal{V}(\eta)$. This means that the trajectory density can be written as

$$f(\mathcal{V}|\eta) = \delta(\mathcal{V} - \mathcal{V}(\eta)), \tag{10}$$

where $\delta(x)$ is the Dirac delta function and represents a point mass concentrated at $x = 0$.

In contrast, the stochastic coalescent SIR model assumes that the epidemic is generated by a jump process corresponding to the master equation given in eq. (5). In this case, the probability $f(\mathcal{V}|\eta)$ is nonsingular and thus contributes to the uncertainty in the final inference result.

In the BDSIR model, $f(\mathcal{V}|\eta)$ is the same as for the stochastic coalescent SIR model, but $f(\mathcal{T}|\mathcal{V},\eta)$ is defined differently. See KÜHNERT *et al.* (2014) for details.

MCMC algorithm We use Markov chain Monte Carlo (MCMC) to sample from the joint posterior density given in eq. (6). Many of the specifics of the algorithm used have been discussed previously; in particular the method for calculating the tree likelihood (FELSENSTEIN 1981, 2004) and mechanism for exploring tree space (DRUMMOND *et al.* 2002). However, the model-specific product $f(\mathcal{T}|\mathcal{V},\eta)f(\mathcal{V}|\eta)$ requires special attention.

As we are primarily interested in parametric inference rather than the epidemic trajectory itself, we can regard $\mathcal{V}$ as a nuisance parameter to be marginalized over. This marginalization can be achieved implicitly by sampling it using MCMC and then ignoring this component of the sampled state, which is the strategy we use when reporting the BDSIR results. It can also be made an explicit part of the likelihood calculation, which is the approach we take with the deterministic and stochastic coalescent SIR models. This marginalization means that the product $f(\mathcal{T}|\mathcal{V},\eta)f(\mathcal{V}|\eta)$ becomes

$$f(\mathcal{T}|\eta) = \int f(\mathcal{T}|\mathcal{V},\eta)f(\mathcal{V}|\eta)\mathrm{d}\mathcal{V}, \tag{11}$$

the probability density of the tree given the epidemiological parameters.

In the case of the deterministic coalescent SIR model, this density reduces to $f(\mathcal{T}|\mathcal{V}(\eta),\eta)$, meaning that the density of the tree given epidemiological parameters η is obtained simply by substituting the numerical solution to eqs. (1)–(3) for those parameters into eq. (7).

The stochastic coalescent SIR model is more complex, as in this case the trajectory density $f(\mathcal{V}|\eta)$ is nonsingular, meaning that computing the integral in eq. (11) is nontrivial. We treat this here using the "pseudo-marginal" approach (BEAUMONT 2003; ANDRIEU and ROBERTS 2009) in which, at each step in the MCMC chain, the marginalized tree density

$f(\mathcal{T}|\eta)$ is replaced by the Monte Carlo estimate

$$\hat{f}(\mathcal{T}|\eta) = \frac{1}{M}\sum_{r=1}^{M} f(\mathcal{T}|\mathcal{V}_r, \eta), \tag{12}$$

where each $\mathcal{V}_r$ is a trajectory sampled independently from $f(\mathcal{V}|\eta)$ using a stochastic simulation algorithm (Sehl *et al.* 2009). Perhaps counterintuitively within an MCMC framework, this stochastic likelihood converges to the true marginal posterior distribution regardless of the number M of realizations used in the estimate. However, the magnitude of M can significantly affect the rate at which the chain produces effectively independent samples from the posterior and must be tuned carefully.

Implementation and validation We have implemented the schemes described above for performing inference under the deterministic and stochastic coalescent SIR models within the BEAST 2 phylodynamics package found at `http://github.com/CompEvol/phylodynamics`. This has a number of advantages over a stand-alone implementation. Foremost, we were able to avoid reimplementing components of the algorithm that are in common with other already-implemented phylogenetic and phylodynamic analyses, such as the MCMC proposal operators used to traverse the parameter space. Furthermore, this greatly increases the usefulness of the implementation, as it can be immediately used in conjunction with a wide variety of nucleotide and amino acid substitution models and parameter priors.

We have taken two steps in order to ensure our implementation is correct. First, we have compared tree probability density $f(\mathcal{T}|\mathcal{V}, \eta)$ values calculated using the main implementation of each of the two models with those calculated using completely independent implementations in R (R Core Team 2014).

Second, we have used the implemented MCMC algorithms to sample transmission trees from the tree density given in eq. (11) for each model. We then compared the distributions of tree height, total edge length, and binary clade count summary statistics from these sampled

ensembles with sample distributions obtained directly via stochastic simulation. As shown in Section 1 (*Sampling from the prior*) in the online supporting information, and in the associated figures, the resulting pairs of distributions agree, providing strong support for our claim that the implementations of the methods described above are correct.

Instructions for downloading and using this package are also available on the project web site located at `http://github.com/CompEvol/phylodynamics`.

Simulation study To evaluate the implementation and extension of the coalescent models, we performed analyses on both sequence data and fixed trees simulated with known parameter values. The median estimated values produced by each model were then used to measure relative error and bias, along with the widths and coverage of 95% highest posterior density (HPD) intervals.

We used three methods for simulating the trees and trajectories, as shown below:

Inference model	Stoch. Coal. SIR	Deter. Coal. SIR	BDSIR
Simulation scheme	Stoch. Coal. SIR	Stoch. Coal. SIR	Stoch. Coal. SIR
	Deter. Coal. SIR	Deter. Coal. SIR	Deter. Coal. SIR
	Stochastic SIR	Stochastic SIR	Stochastic SIR

The stochastic coalescent and deterministic coalescent simulation schemes were used to validate the coalescent SIR inference models. The *stochastic SIR* scheme, contrarily, is emphasized for its realistic properties.

Stochastic SIR trees and trajectories were generated using master equations in the simulation package MASTER (Vaughan and Drummond 2013). Deterministic coalescent trajectories were generated using a Runge-Kutta integrator (Runge 1895; Kutta 1901) with adaptive step sizes to solve a system of first order ODEs. Stochastic coalescent trajectories were generated using Sehl *et al.*'s (2009) SAL tau-leaping algorithm (Sehl *et al.*

2009).

To simulate the stochastic coalescent SIR trees, we used the *stochastic SIR* trajectories, which could be converted to effective population size with the mathematical expression used to obtain Volz's (2012) coalescent rate for the SIR model: $N_e(\tau) = 1/\lambda_2(\tau) = \mathcal{I}(\tau)/(2\beta\mathcal{S}(\tau))$. The sampling times, generated by a sampling rate ψ, for the stochastic coalescent SIR trees were also taken from the MASTER output to allow for direct comparison between the sets of trees. In other words, the underlying epidemic function was the same for both stochastic SIR and stochastic coalescent SIR trees, the latter of which were then simulated under a piecewise constant population function.

Likewise, for the simulation of deterministic coalescent trees we used deterministic SIR trajectories to construct a population function and the relation $N_e = \mathcal{I}/(2\beta\mathcal{S})$ to convert infected and susceptible host population sizes to effective population size. The sampling times were randomly generated from a probability distribution so that the density of samples taken through time were proportional to the number of infected individuals through time, as with the stochastic SIR trees.

We simulated stochastic SIR trees using multiple combinations of parameter values. We were particularly interested in varying the basic reproductive ratio R_0 and the initial susceptible population size S_0, to observe the changes in relative error, bias, and uncertainty in stochastic and deterministic models. To alter the ratio $R_0 = \frac{\beta S_0}{\gamma}$ and still generate sensible trees with a consistent number of tips, one or more of the other parameters (birth rate β, removal rate γ, or S_0) must also change. Table 2, as well as Tables S6, S7, and S9 in the supporting information, show the true values of the parameters for each set of simulations. (The birth rate β is not shown, as our implementation allows either β or R_0 to serve as a parameter in the inference, and R_0 is the parameter of interest. However, β can be calculated via the other three, using $\beta = \frac{R_0\gamma}{S_0}$. For example, when $R_0 = 1.0978$, $S_0 = 499$, and $\gamma = 0.25$, then $\beta = 5.50\text{E-}4$.)

Heterochronous trees We generated 100 trees under each of the three (stochastic SIR, stochastic coalescent SIR, deterministic coalescent SIR) models with parameters S_0, β, and γ. For heterochronously sampled trees, each removal generates a sample with probability $\psi/(\psi + \mu)$, where ψ is the overall rate of sampled removals and μ is the rate of unsampled removals such that $\gamma = \psi + \mu$.

The simulations ended once the number of infected individuals reached zero, i.e., when the last infected individual was removed. This ensured that the simulated trajectories spanned past the exponential growth phase of the epidemic and therefore included samples past the peak of infected individuals. This choice of procedure was motivated by (a) the suggestion of Stadler *et al.* (2014) that the behavior of the coalescent beyond the exponential phase could either inflate or reduce bias, and (b) the observations of Dearlove and Wilson (2013) and Bošková *et al.* (2014) that deterministic coalescent SIR models might be properly fitted only once the epidemic has peaked. Figure 1 shows trajectories of susceptible, infected, and removed individuals underlying the simulation of stochastic SIR trees (Figure 2) generated in MASTER. An example XML for simulating these MASTER trees is provided in the supporting information.

We required that the trees had $n \geq 100$ leaves, filtering out those in which the epidemic died out in the early stages, i.e., when the initial infected individual was removed from the effective population too quickly to infect others. (Note that the inference procedures discussed in this manuscript all implicitly condition on the number of leaves.) The probability that the first event in a given trajectory is the removal (by recovery, death, etc.) of patient zero is given by $\delta/(\beta S_0 + \delta) = 1/(1 + R_0)$. When $R_0 \approx 2.50$, this probability is $\approx 30\%$. In our case, 52/152 ($\approx 34\%$) trees were "empty", or containing only one node. The filtering process left us with a mean of ≈ 160 leaves for the simulated trees.

Homochronous trees A major concern in the comparison between Kühnert *et al.* (2014)'s birth-death-sampling SIR inference model, which includes explicit sampling, and our imple-

mentations of Volz (2012)'s coalescent SIR models, which do not include explicit sampling, is that the former is given extra information via the sampling process. Volz and Frost (2014) have addressed this issue by providing a coalescent SIR model that does incorporate sampling explicitly.

That being said, results from Bošková *et al.* (2014) indicate that the poor performance of the deterministic coalescent SIR model in comparison with birth-death models was due to the lack of handling stochastic population size changes through time rather than the lack of information about the sampling proportion. Their results showed that the coalescent is "very robust to changes in sampling schemes".

Regardless, to ensure a fair comparison of BDSIR and the coalescent SIR models, we simulated an SIR epidemic with homochronous, or contemporaneous, sampling. This type of simulation affords no additional information about the population size for explicit-sampling models, as there is only a single time of sampling.

We selected a simulation time of $t = 20$ for the homochronously sampled trees, with the trajectories being sampled at high prevalence but also past the time of peak prevalence. This is important for distinguishing SIR from SI/SIS outbreaks, as it provides information about the removal parameter γ. In this set of simulations, each lineage was sampled at $t = 20$ with probability 0.7, (the leaf count distribution for varied sampling probabilities is in the supporting information).

Simulated sequences To assess the ability of each SIR model to infer epidemic parameters with the inclusion of phylogenetic uncertainty, we also simulated the evolution of 2000 bp sequences down each simulated tree. We time-stamped the sequences with the tip dates of each corresponding tree and informed the inference with the true Hasegawa-Kishino-Yano (HKY) substitution model (Hasegawa *et al.* 1985), clock rate = 5e-3, and $\kappa = 5$. These choices were made to reflect real data, specifically that of influenza (Vaughan *et al.* 2014).

Along with simulated sequence data, analyses were performed with the simulated trees

fixed (results are in the supporting information), and the parameters R_0, γ, S_0, and the origin of the tree z_0 were estimated with Bayesian prior distributions as listed in Table 4.

Deterministic coalescent SIR on higher R_0 and S_0 Finally, we had particular interest in the effects of varying the population size parameter S_0 on the deterministic coalescent SIR model, as comparisons from initial analyses with lower true R_0 ($\approx$1.5 and $\approx$1.1) and S_0 (=499) showed higher error and bias and lower 95% HPD coverage. Also, it is often assumed that deterministic descriptions will perform well for higher R_0 and larger population sizes. Tables 7 and 9 in the supporting information detail the parameter values we used to explore the behavior of the deterministic coalescent on varied R_0 and S_0 combinations.

Interpretation of results We compared the coalescent SIR, as well as BDSIR, parameter estimations from the simulated data to the true values used to generate the SIR trajectories. Following KÜHNERT *et al.* (2014), the precision and accuracy of these methods were measured by relative error, bias, and highest posterior density (HPD) intervals. We used the posterior median value of the parameter value $\hat{\eta}$ compared with the true parameter $\bar{\eta} \in \{R_0, \gamma, S_0, z_0\}$. Relative error and bias are then gauged by calculating the median value over medians from all 100 trees, such that

$$RE_{\hat{\eta}} = \frac{\sum_{\tau=1}^{100} \frac{|\hat{\eta}-\bar{\eta}|}{\bar{\eta}}}{100}$$

and

$$RB_{\hat{\eta}} = \frac{\sum_{\tau=1}^{100} \frac{\hat{\eta}-\bar{\eta}}{\bar{\eta}}}{100}.$$

Measures of HPD interval widths are given by

$$\frac{95\% \text{ HPD upper bound} - 95\% \text{ HPD lower bound}}{\bar{\eta}}.$$

Tables 1-3 show these results, along with the percentages of posterior estimates that produced 95% HPD intervals containing the true values (i.e., 95% HPD coverage).

H1N1 data analysis To test the efficacy of the coalescent SIR models on real data, epidemic parameters R_0, γ, S_0, and time of origin z_0 were estimated from 42 seasonal influenza A (H1N1) sequences sampled throughout the 2001 flu season in Canterbury, New Zealand.

Influenza infections are well known for their seasonal SIR behavior in non-equatorial populations, as each annual flu season begins with a supply of susceptible hosts and tapers off as the hosts recover with adaptive immunity (Iwasaki and Pillai 2014). Due partly to this seasonal pattern, the influenza virus is both a motivator for the development of specialized models as well as a prime subject for testing phylodynamic models (Koelle *et al.* 2006).

Sampling a particular region bypasses the necessity of specifying geographically-structured populations, and New Zealand is an area of particular interest due to its geographic location and relative isolation from other regions with potentially varying dynamics. It is also assumed to play a key role in the global circulation of influenza strains (Rambaut and Holmes 2009; Bedford *et al.* 2010).

We used an HKY nucleotide substitution model, with a substitution rate of 5e-3 as estimated in Vaughan *et al.* (2014), and informed the models with dated sequences. Priors used for the Bayesian inference are shown in Table 4.

HIV-1 data analysis In addition to our analysis of H1N1 sequence data, we selected HIV-1 subtype B nucleotide sequences collected from infected individuals located in the UK. The coalescent SIR results were collated with the results from the BDSIR data analysis performed by Kühnert *et al.* (2014) using the same sequences. More details of this analysis are provided in the supporting information.

RESULTS AND DISCUSSION

Simulation study Results for epidemic parameter inference from nucleotide sequences simulated from stochastic SIR trees are provided in Table 1 for $R_0 \approx 2.50$. Results for inference from fixed trees ($R_0 \approx 2.50$, $R_0 \approx 1.50$, $R_0 \approx 1.10$) are shown in Table 2. Inference results for analyses with true $R_0 = 1.0987$ and varying population size ($S_0 = 499, 999, 1999$) are described in the supporting information, along with results from trees simulated under the stochastic and deterministic coalescent models for validation.

Heterochronous trees For $R_0 \approx 2.50$, all three inference methods performed similarly for parameters R_0 and γ, with high 95% HPD coverage and low error and bias. The most weakly identifiable parameter S_0 yielded the largest HPD intervals for all three inference models. The deterministic coalescent returned higher error (0.52) and bias (0.29) than stochastic coalescent SIR (0.19, -0.03) and BDSIR (0.39, 0.24) and recovered the origin parameter z_0 for only 76 out of 100 simulated trees, while the stochastic coalescent and BDSIR respectively recovered z_0 for 99 and 97 out of 100 simulations.

For $R_0 \approx 1.50$, the relative HPD widths (akin to variance) for three of the four estimated parameters (R_0, γ, and z_0) were smallest for BDSIR. For the parameter S_0, the relative HPD width is largest for BDSIR, although it also had slightly higher 95% HPD coverage than deterministic coalescent SIR and the same as stochastic coalescent SIR. The deterministic coalescent SIR method recovered the truth for 85, 89, 91, and 88 out of 100 trees for parameters R_0, γ, S_0, and z_0, while its stochastic analog recovered the truth for 100, 85, 100, and 99 out of 100 trees for the same parameters. Finally, for stochastic coalescent SIR and BDSIR, error and (absolute) bias were relatively low for R_0, arguably the parameter of most interest to epidemiologists since it represents the number of individuals each infected individual will themselves infect in a naive population. Deterministic coalescent SIR has a higher error (0.24) and bias (0.15) and also has significantly lower coverage for R_0 (85%).

For $R_0 \approx 1.10$, the two stochastic models again outperformed the deterministic coalescent

in error, bias, and 95% HPD coverage. The stochastic coalescent most reliably recovered the truth for R_0 (99 out of 100 simulations), while the deterministic coalescent had more than double the error and bias and still only recovered the truth for 25 of the 100 simulations. BDSIR had the lowest error and bias for R_0 under this scheme, although it only recovered the truth for 75 out of 100 simulations. For removal parameter γ, BDSIR again yielded lower error and bias, in this case returning the truth for 100/100 trees (in contrast to 84 and 86 from the stochastic and deterministic coalescent, respectively).

In the stochastic models, there is a greater tradeoff between parameters due to the impact the relationship between them has on the survival of trajectories at low R_0. A larger estimated removal rate tends to require a larger susceptible population in order for the epidemic to avoid dying out in the early stages. Likewise, a smaller susceptible population implies a smaller estimated γ.

Deterministic coalescent SIR on higher R_0 and S_0 As mentioned in the preceding subsection, the deterministic coalescent model yielded higher error and bias than both the stochastic coalescent and BDSIR for most parameters with $R_0 \approx 1.10$ and $S_0 = 499$.

To investigate the deterministic model's sensitivity to population sizes, we also simulated a range of population sizes ($S_0 = 499$, 999, and 1999) for $R_0 = 1.0987$. Even with $S_0 = 1999$, the deterministic coalescent SIR model's 95% HPD coverage was low. For parameters R_0, γ, S_0, and z_0, this coverage was respectively: 40%, 64%, 66%, and 18%. Table S6 in the supporting information shows these results.

Additionally, we increased both R_0 (to 3.5 and 5) and S_0 (to 4999 and 9999). However, for parameters R_0, γ, and S_0, the deterministic coalescent SIR showed increased error, bias, and HPD widths, and the HPD coverage for z_0 did not improve. These results are shown in Table S9 in the supporting information.

While each of these methods are approximations, the deterministic coalescent particularly suffers from model misspecification since it does not account for the stochasticity that is

always present in the early stages of epidemics, regardless of S_0.

Homochronous trees Results for homochronously sampled trees are given in Table S3 in the supporting information.

All three SIR inference models recover the truth for more than 95/100 trees within their respective 95% HPD widths for epidemic parameters R_0, γ, and S_0. The time of origin z_0 was recovered for 100/100 trees by BDSIR, 95/100 trees by stochastic coalescent SIR, and 73/100 trees by deterministic coalescent SIR. However, relative error and bias also increased consistently across all three models, along with the 95% HPD widths. The deterministic coalescent had the highest error, bias, and HPD width for R_0 and highest error and HPD width for S_0, which is consistent with the heterochronously sampled data.

Further consideration of the effects of sampling rate changes and sampling model misspecification are warranted for BDSIR and coalescent SIR, the latter of which has been facilitated by Volz and Frost (2014).

Simulated sequences Relative error and bias were inflated across all three inference models with the addition of phylogenetic uncertainty, and in certain cases the 95% HPD coverage was lower than with fixed trees. The deterministic coalescent model only recovered the truth within its 95% HPD intervals for 90 or more of the 100 trees in the case of S_0. The true values for the parameters R_0, γ, and z_0 were covered by 95% HPD intervals for 87, 56, and 29 of the 100 trees, respectively. This is contrasted with the performance of the stochastic coalescent (100, 97, 47, and 37 for parameters S_0, R_0, γ, and z_0) and BDSIR (99, 100, 84, and 18 for S_0, R_0, γ, and z_0), as shown in Table 1.

Error, bias, and 95% HPD widths were higher with simulated sequences for all three inference models for parameters γ, S_0, and z_0, than with fixed trees. This indicates the importance of calibrating epidemic parameters of interest. In our case, we emphasize the basic reproductive number R_0, often the parameter of most interest to epidemiologists. For

R_0, stochastic coalescent SIR and BDSIR recovered the truth within their 95% HPD intervals for 97 and 100 of the 100 simulations, respectively. They also showed only slight changes in error and bias compared to inference performed on the fixed trees used to generate the sequences. The deterministic coalescent SIR model recovered R_0 for 87 of the 100 simulations (contrasted with 98/100 for the fixed trees), and with increased error.

Priors and identifiability It is important to understand the impact of selected priors on inference results, as the priors are where the power of Bayesian inference lies. For example, we found relatively weak identifiability in the initial susceptible population parameter S_0, which must either be fixed or estimated alongside the origin parameter z_0.

In addition to allowing each parameter to be either fixed or estimated, we have provided options for parameterization of our models, with either the transmission rate β or R_0 acting as operable parameters in MCMC analysis. For the deterministic coalescent, there is also an option to use the intrinsic growth parameter described by Dearlove and Wilson (2013).

The choice of parameterization necessarily affects the prior that will be used in the inference and should be considered carefully. However, we found that once a parameterization has been selected, our inference models are robust to different prior distributions placed on each parameter. We also used broader prior distributions on the deterministic coalescent to test whether this would increase its lower 95% HPD coverage relative to the stochastic models. We found that doing so increased the error and bias of the results without increasing the accuracy, (shown in Table S4 in the supporting information).

H1N1 data analysis Epidemic parameter estimates from serially-sampled influenza A (H1N1) virus sequence data are shown in Table 3.

The estimated means of the basic reproductive number were R_0 =1.46, 1.35, and 1.61 for the stochastic coalescent, deterministic coalescent, and BDSIR, respectively. Estimates of R_0 from pandemic H1N1 in New Zealand range from about 1.2 to 1.5 (Paine *et al.* 2010;

Roberts and Nishiura 2011; Opatowski *et al.* 2011; Roberts 2013; Biggerstaff *et al.* 2014), and estimates of R_0 for seasonal H1N1 from other countries also range from around 1.2 to 1.5 (Chowell *et al.* 2008). The 95% HPD intervals were very similar across each model, ranging from just over 1.0 to around 2.0.

The population of the Canterbury region in 2001 was reported to be around $481,431$ by the Environment Canterbury Regional Council (ECAN 2001) and $521,832$ by Statistics New Zealand (StatsNZ 2001). The mean estimates of S_0 were considerably lower using the stochastic coalescent ($S_0 = 69,000$), the deterministic coalescent ($S_0 = 120,000$), and BDSIR ($S_0 = 22,200$). However, the *effective* population of susceptibles is assumed to be much smaller, as the total population contains individuals of various susceptibility, e.g., those with partial immunity from vaccination and previous or secondary infections.

Most people recover from flu symptoms, the time they are likely to be most infectious, within a few days up to two weeks (CDC 2014; WHO 2014). This provides a range of probable true values for the removal parameter γ. The sequence data and molecular clock rate, and therefore the tree, are in units of years. Therefore, our γ range would be 365/14 days to 365/2 days, or $\gamma = 26.1$ to $\gamma = 182.5$. The stochastic coalescent, deterministic coalescent, and BDSIR respectively inferred γ means of: 27.08, 34.50, and 27.72. These estimates are on the low side compared to epidemiological models for influenza that include explicit spatial and household effects (Ferguson *et al.* 2005), but a moderate misfit of the model is not unexpected when fitting a simple closed SIR model with no population substructure.

The root of the tree was very similar across all inference models, respectively: 0.53, 0.54, and 0.49 for stochastic coalescent SIR, deterministic coalescent SIR, and BDSIR. The same was true for the origin z_0, with: 0.69, 0.73, and 0.53 for the stochastic coalescent, deterministic coalescent, and BDSIR. All three inference models returned tree root and origin estimates that are consistent with previous estimates from single flu seasons. That is, the tree age is young and the root coincides with the start of the (winter) influenza season

in the Southern Hemisphere. The time of introduction of influenza into the region, z_0, was one or two months before the root. This supports the notion that the sequences selected represent a single introduction of the strain into the Canterbury population (see supporting information for details of data selection).

The trees estimated by each of the three models are typical for influenza (see Figure 11 for representative trees from each posterior), with branches that are quick to coalesce moving backward in time from the most recently sampled tip.

HIV-1 data analysis Results for inference from HIV-1 sequence data can be found in the supporting information.

Computational efficiency Finally, we supply Table S5 in the supporting information to show comparisons of computation times under each inference model for each type of data analyzed. The deterministic coalescent SIR model is by far the fastest to sample and converge, with stochastic coalescent SIR and BDSIR varying depending on the type of data.

Closing remarks A key reason for the success of coalescent theory in population genetics is its mathematical simplicity and the computational efficiency of calculating the probability density of a sample genealogy. Our results show that a stochastic variant of coalescent theory can be successfully adapted to estimate epidemiological parameters in a true Bayesian inference context. This stochastic coalescent SIR model performs better than the deterministic analog for estimating epidemic parameters in some circumstances. Unfortunately, the stochastic model relies on a computationally demanding Monte Carlo estimate of the coalescent density via simulation of an ensemble of epidemic trajectories, negating one of the main advantages of coalescent theory. In fact, the current implementation is less computationally efficient than the implementation of the BDSIR model. However, an advantage of the stochastic coalescent over the explicit sampling model in BDSIR is its robustness to biased sampling schemes, as has been shown for the case of pure exponential growth dynamics

(Bošková *et al.* 2014).

A more computationally efficient approach to computing the coalescent probability of the sample genealogy in the stochastic setting would be to use particle filtering (Andrieu and Roberts 2009; Andrieu *et al.* 2010; Rasmussen *et al.* 2011, 2014), but there are no theoretical barriers to applying particle MCMC to the exact model (Stadler *et al.* 2014). Therefore, an obvious extension of this work would be to apply particle MCMC algorithms to the exact stochastic SIR model that was used in simulations in this current work. We would anticipate that the exact model would outperform all the methods tested here, especially when R_0 is close to one.

In the meantime, the Bayesian coalescent inference methods developed here make it feasible to estimate epidemic parameters from time-stamped, serially-sampled molecular sequence data, while accurately accounting for uncertainty in the topology and the divergence times of the phylogenetic tree.

[Figure 1 about here.]

[Figure 2 about here.]

[Figure 3 about here.]

[Figure 4 about here.]

[Table 1 about here.]

[Table 2 about here.]

[Table 3 about here.]

ACKNOWLEDGMENTS

AJD was funded by a Rutherford Discovery Fellowship from the Royal Society of New Zealand. AP, TV, TS and AJD were also partially supported by Marsden grant #UOA1324

from the Royal Society of New Zealand. (http://www.royalsociety.org.nz/programmes/funds/marsden/awards/2013-awards/)

We would also like to thank Gabriel Leventhal and Louis Du Plessis (ETH Zürich) for constructive and valuable input and the New Zealand eScience Infrastructure for access to high performance computing facilities. (http://www.nesi.org.nz/)

LITERATURE CITED

Anderson, R. M., and R. M. May, 1991 *Infectious diseases of humans: dynamics and control*. Oxford University Press, Oxford.

Andrieu, C., A. Doucet, and R. Holenstein, 2010 Particle Markov chain Monte Carlo methods. Journal of the Royal Statistical Society: Series B (Statistical Methodology) **72**: 269–342.

Andrieu, C., and G. O. Roberts, 2009 The pseudo-marginal approach for efficient Monte Carlo computations. Ann. Statist. **37**: 697.

Beaumont, M. A., 2003 Estimation of population growth or decline in genetically monitored populations. Genetics **164**: 1139–1160.

Bedford, T., S. Cobey, P. Beerli, and M. Pascual, 2010 Global migration dynamics underlie evolution and persistence of human influenza a (h3n2). PLoS Pathology **6**: e1000918.

Biggerstaff, M., S. Cauchemez, C. Reed, M. Gambhir, and L. Finelli, 2014 Estimates of the reproduction number for seasonal, pandemic, and zoonotic influenza: a systematic review of the literature. BMC Infectious Diseases : 480.

Bošková, V., S. Bonhoeffer, and T. Stadler, 2014 Inference of epidemiological dynamics based on simulated phylogenies using birth-death and coalescent models. PLoS Computational Biology .

CDC, 2014 United states centers for disease control and prevention. retrieved from http://www.cdc.gov/flu/ .

Chowell, G., M. Miller, and C. Viboud, 2008 Seasonal influenza in the united states, france, and australia: transmission and prospects for control. Epidemiology and Infection **6**: 852–864.

Dearlove, B., and D. J. Wilson, 2013 Coalescent inference for infectious disease: meta-analysis of hepatitis C. Philosophical Transactions of the Royal Society B: Biological Sciences **368**: 20120314.

Drummond, A. J., S. Y. W. Ho, M. J. Phillips, and A. Rambaut, 2006 Relaxed phylogenetics and dating with confidence. PLoS Biol **4**: e88.

Drummond, A. J., G. K. Nicholls, A. G. Rodrigo, and W. Solomon, 2002 Estimating mutation parameters, population history and genealogy simultaneously from temporally spaced sequence data. Genetics **161**: 1307–20.

ECAN, 2001 Environment canterbury regional council. http://ecan.govt.nz/about-us/population/how-many/pages/census.aspx .

Felsenstein, J., 1981 Evolutionary trees from DNA sequences: a maximum likelihood approach. J Mol Evol **17**: 368–376.

Felsenstein, J., 2004 *Inferring phylogenies*. Sinauer Associates, Sunderland, Mass.

Ferguson, N., D. Cummings, S. Cauchemez, C. Fraser, S. Riley, *et al.*, 2005 Strategies for containing an emerging influenza pandemic in southeast asia. Nature **437**.

Gavryushkina, A., D. Welch, T. Stadler, and A. Drummond, 2014 Bayesian inference of sampled ancestor trees for epidemiology and fossil calibration. arXiv preprint arXiv:1406.4573 .

Grenfell, B. T., O. G. Pybus, J. R. Gog, J. L. N. Wood, J. M. Daly, *et al.*, 2004 Unifying the epidemiological and evolutionary dynamics of pathogens. Science **303**: 327–332.

Griffiths, R. C., and S. Tavaré, 1994 Ancestral inference in population genetics. Statistical Science **9**: 307–319.

Hasegawa, M., H. Kishino, and T. Yano, 1985 Dating of the human-ape splitting by a molecular clock of mitochondrial dna. Journal of Molecular Evolution **22**: 160–174.

Iwasaki, A., and P. S. Pillai, 2014 Innate immunity to influenza virus infection. Nature Reviews Immunology **14**: 315–328.

Keeling, M. J., and P. Rohani, 2008 *Modeling infectious diseases in humans and animals*. Princeton University Press, Princeton.

Kermack, W., and A. McKendrick, 1932 Contributions to the mathematical theory of epidemics. ii. the problem of endemicity. Proceedings of the Royal Society A **138**.

Kingman, J. F. C., 1982 The coalescent. Stochastic Processes and their Applications **13**: 235–248.

Koelle, K., S. Cobey, B. Grenfell, and M. Pascual, 2006 Epochal evolution shapes the phylodynamics of interpandemic influenza a (h3n2) in humans. Science **314**: 1898–1903.

Koelle, K., and D. A. Rasmussen, 2012 Rates of coalescence for common epidemiological models at equilibrium. J R Soc Interface **9**: 997–1007.

Kühnert, D., T. Stadler, T. G. Vaughan, and A. J. Drummond, 2014 Simultaneous reconstruction of evolutionary history and epidemiological dynamics from viral sequences with the birth-death sir model. J R Soc Interface **11**: 20131106.

Kutta, M. W., 1901 Beitrag zur näherungsweisen integration totaler differentialgleichungen. Zeitschrift für Mathematik und Physik **46**: 435–453.

Opatowski, L., C. Fraser, J. Griffin, E. de Silva, M. Van Kerkhove, *et al.*, 2011 Transmission characteristics of the 2009 h1n1 influenza pandemic: Comparison of 8 southern hemisphere countries. PLoS Pathogens .

Paine, S., G. Mercer, P. Kelly, D. Bandaranayake, M. Baker, *et al.*, 2010 Transmissability of 2009 pandemic influenza a(h1n1) in new zealand: effective reproduction number and influence of age, ethnicity, and importations. Eurosurveillance .

Pybus, O. G., M. A. Charleston, S. Gupta, A. Rambaut, E. C. Holmes, *et al.*, 2001a The epidemic behavior of the hepatitis c virus. Science **292**: 2323–2325.

Pybus, O. G., M. A. Charleston, S. Gupta, A. Rambaut, E. C. Holmes, *et al.*, 2001b The epidemic behavior of the hepatitis c virus. Science **292**: 2323–2325.

R Core Team, 2014 *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria.

Rambaut, A., 2007 Figtree. http://tree.bio.ed.ac.uk/software/figtree/.

Rambaut, A., and E. Holmes, 2009 The early molecular epidemiology of the swine-origin a/h1n1 human influenza pandemic. PLoS Currents .

Rasmussen, D. A., O. Ratmann, and K. Koelle, 2011 Inference for nonlinear epidemiological models using genealogies and time series. PLoS Comput Biol **7**: e1002136.

Rasmussen, D. A., E. M. Volz, and K. Koelle, 2014 Phylodynamic inference for structured epidemiological models. PLoS Comput Biol **10**: e1003570.

Roberts, M., 2013 Epidemic models with uncertainty in the reproduction number. Journal of Mathematical Biology : 1463–1474.

Roberts, M., and H. Nishiura, 2011 Early estimation of the reproduction number in the presence of imported cases: Pandemic influenza h1n1-2009 in new zealand. PLoS One .

Rodrigo, A., and J. Felsenstein, 1999 *The evolution of HIV*, chapter Coalescent approaches to HIV population genetics. The Johns Hopkins University Press, 233–272.

Rouzine, I. M., A. Rodrigo, and J. M. Coffin, 2001 Transition between stochastic evolution and deterministic evolution in the presence of selection: general theory and application to virology. Microbiol Mol Biol Rev **65**: 151–85.

Runge, C., 1895 Ueber die numerische auflösung von differentialgleichungen. Mathematische Annalen **46**: 167–178.

Sehl, M., A. V. Alekseyenko, and K. L. Lange, 2009 Accurate stochastic simulation via the step anticipation tau-leaping (sal) algorithm. J Comput Biol **16**: 1195–1208.

Stadler, T., R. Kouyos, V. von Wyl, S. Yerly, J. Böni, *et al.*, 2012 Estimating the basic reproductive number from viral sequence data. Mol Biol Evol **29**: 347–57.

Stadler, T., D. Kühnert, S. Bonhoeffer, and A. J. Drummond, 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.

Stadler, T., T. G. Vaughan, A. Gavruskin, S. Guindon, D. Kühnert, *et al.*, 2014 Population genetics vs. population dynamics: How well can coalescent-based models approximate population dynamic processes? .

StatsNZ, 2001 Statistics new zealand. `http://stats.govt.nz/Census/` .

Vaughan, T., D. Kühnert, A. Popinga, D. Welch, and A. Drummond, 2014 Efficient bayesian inference under the structured coalescent. Bioinformatics **In revision**.

Vaughan, T. G., and A. J. Drummond, 2013 A stochastic simulator of birth-death master equations with application to phylodynamics. Molecular Biology and Evolution .

Volz, E., and S. D. Frost, 2014 Sampling through time and phylodynamic inference with coalescent and birth-death models. Journal of the Royal Society Interface **11**.

Volz, E. M., 2012 Complex population dynamics and the coalescent under neutrality. Genetics **190**: 187–201.

Volz, E. M., S. L. Kosakovsky Pond, M. J. Ward, A. J. Leigh Brown, and S. D. W. Frost, 2009 Phylodynamics of infectious disease epidemics. Genetics **183**: 1421–30.

WHO, 2014 World health organization. retrieved from `http://www.who.int/topics/influenza/en/` .

List of Figures

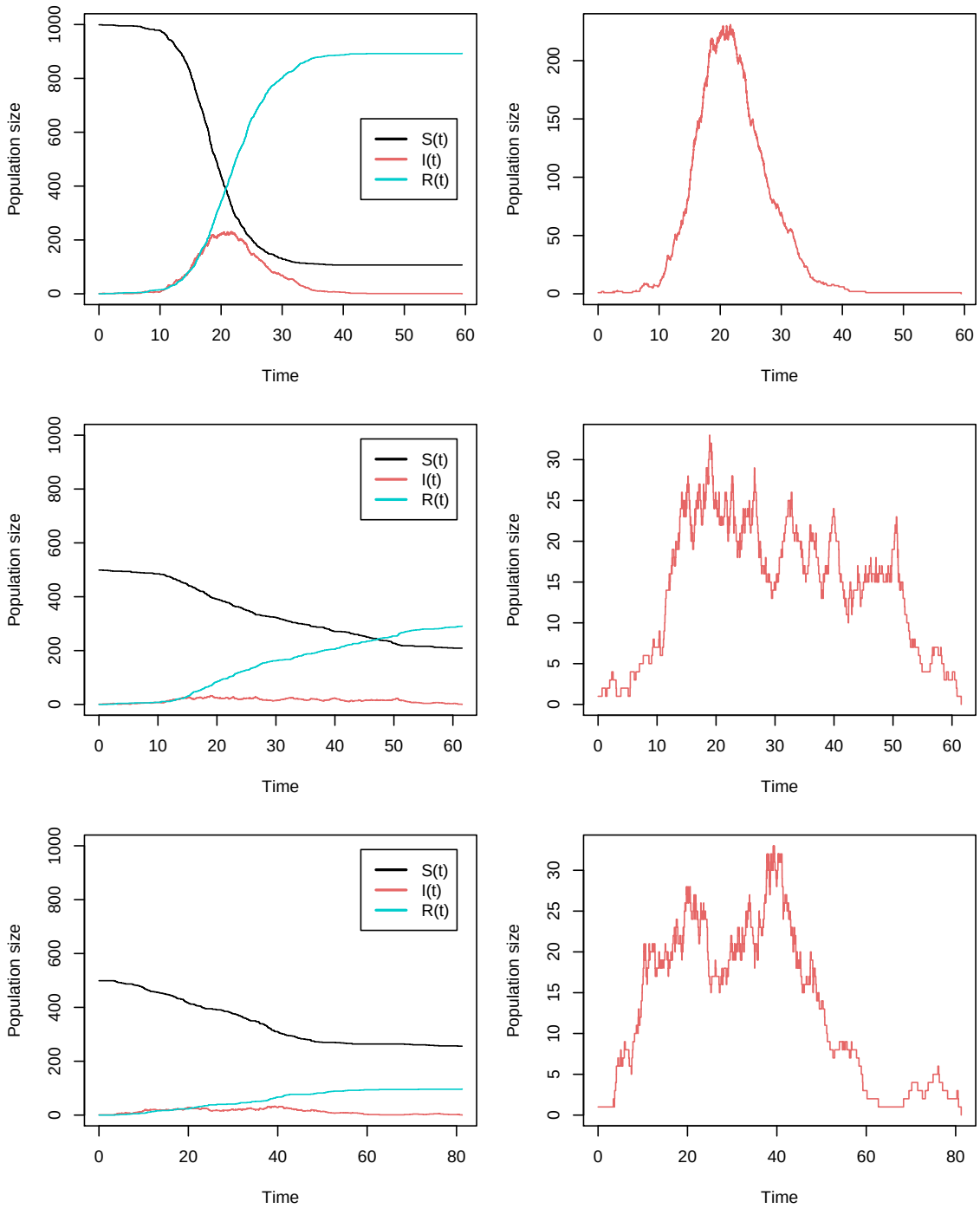

Figure 1: Stochastic SIR trajectories for susceptible S, infected I, and recovered R populations, with (top row) $S_0 = 999$ and $R_0 = 2.4975$, (second row) $S_0 = 499$ and $R_0 = 1.497$, and (bottom row) $S_0 = 499$ and $R_0 = 1.0978$. (The second column shows infected I only.)

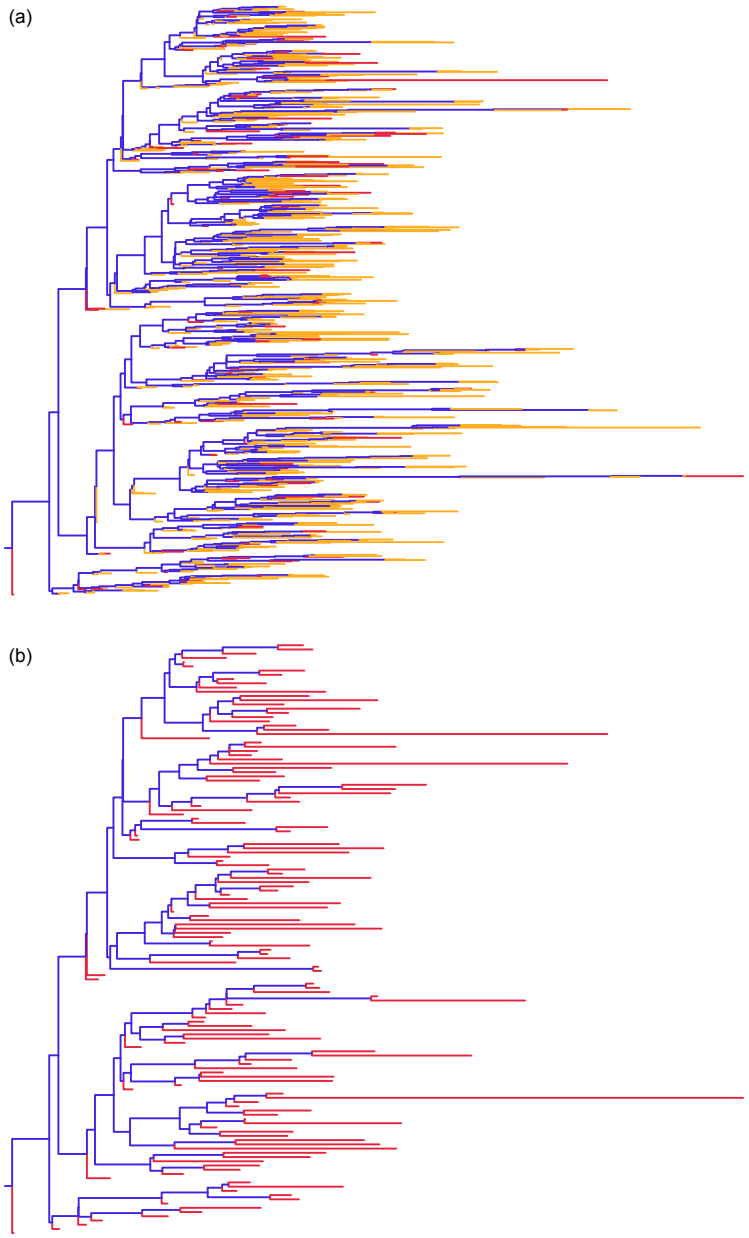

Figure 2: (a) Full stochastic SIR *transmission* tree with both sampled ψ tips, shown in red, and otherwise removed μ tips, shown in yellow. (b) The corresponding 140-tip *sampled* stochastic SIR tree. Figures generated in FigTree (Rambaut 2007).

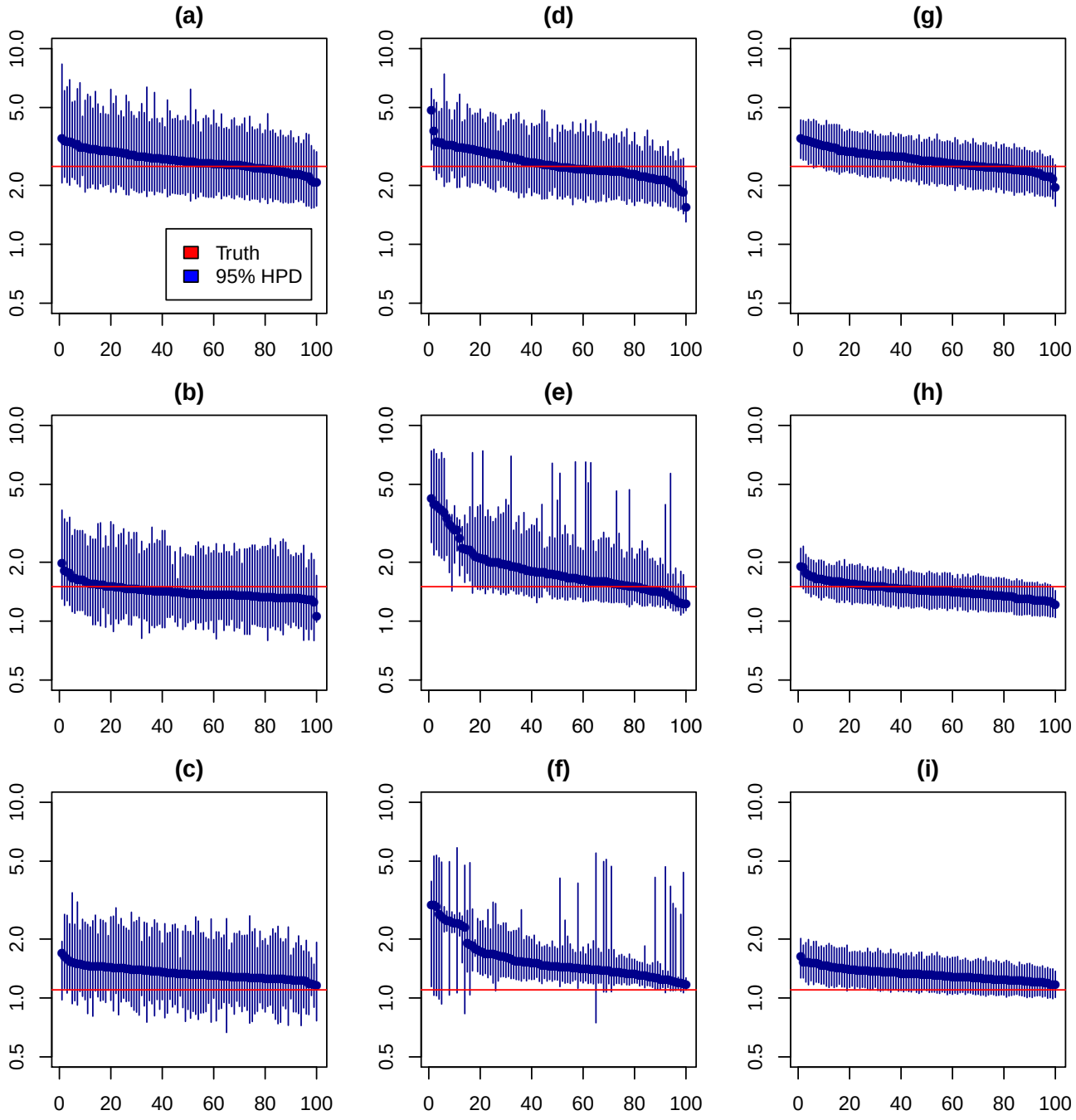

Figure 3: Estimates of $R_{(0)}$ from true stochastic SIR trees using inference methods by column, with stochastic coalescent SIR (a, b, c), deterministic coalescent SIR (d, e, f), and BDSIR (g, h, i). The truth varies by row, with $R_0 = 2.4975$ (a, d, g), $R_0 = 1.4970$ (b, e, h), and $R_0 = 1.0978$ (c, f, i).

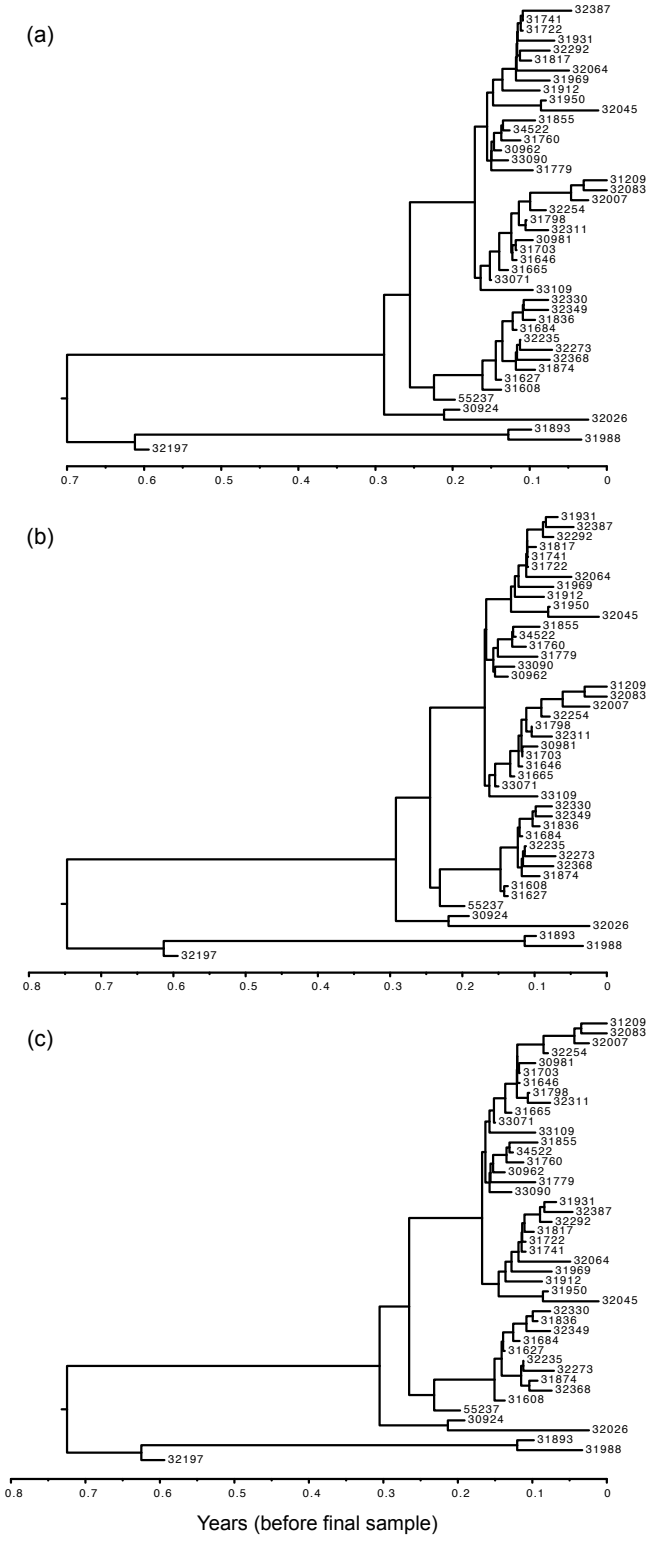

Figure 4: Representative influenza A (H1N1) posterior trees from inference using the (a) BDSIR, (b) stochastic coalescent SIR, and (c) deterministic coalescent SIR inference models.

List of Tables

Table 1: Results for Simulated Sequences: $R_0 \approx 2.50$, $S_0 = 999$

η	Inference	Truth	Mean	Median	Error	Bias	Relative HPD width	95% HPD accuracy
$\mathcal{R}_0$	Stoch.Coal.SIR	2.50	2.41	2.16	0.13	-0.11	0.97	97.00%
	Deter.Coal.SIR	2.50	2.78	2.03	0.38	0.05	0.79	87.00%
	BDSIR	2.50	3.21	2.84	0.15	0.14	1.86	100.00%
γ	Stoch.Coal.SIR	0.30	0.16	0.13	0.52	-0.52	0.82	47.00%
	Deter.Coal.SIR	0.30	0.25	0.16	0.56	-0.28	0.97	56.00%
	BDSIR	0.30	0.17	0.14	0.52	-0.52	1.13	84.00%
$S_{(0)}$	Stoch.Coal.SIR	999	1805	1148	0.32	0.21	5.12	99.00%
	Deter.Coal.SIR	999	2384	1565	0.66	0.60	6.54	100.00%
	BDSIR	999	4002	2611	1.70	1.70	10.38	99.00%
$z_{(0)}$	Stoch.Coal.SIR	(varies)	51.67	48.89	0.26	0.23	0.61	37.00%
	Deter.Coal.SIR	(varies)	49.13	46.46	0.22	0.20	0.26	29.00%
	BDSIR	(varies)	31.16	29.52	0.51	0.51	0.79	18.00%

Table 2: Simulation Study Results for Fixed Trees: $R_0 \approx 2.50$ and $S_0 = 999$, $R_0 \approx 1.50$ and $S_0 = 499$, $R_0 \approx 1.10$ and $S_0 = 499$

η	Inference	Truth	Mean	Median	Error	Bias	Relative HPD width	95% HPD accuracy
$\mathcal{R}_0$	Stoch.Coal.SIR	2.50	2.84	2.68	0.12	0.09	0.98	100.00%
	Deter.Coal.SIR	2.50	2.68	2.49	0.13	0.04	0.81	98.00%
	BDSIR	2.50	2.73	2.67	0.12	0.08	0.55	94.00%
γ	Stoch.Coal.SIR	0.30	0.27	0.25	0.19	-0.13	1.14	99.00%
	Deter.Coal.SIR	0.30	0.32	0.29	0.16	3.14E-3	1.27	99.00%
	BDSIR	0.30	0.28	0.27	0.13	-0.09	0.62	95.00%
$S_{(0)}$	Stoch.Coal.SIR	999	1390	921	0.19	-0.03	3.85	100.00%
	Deter.Coal.SIR	999	1807	1133	0.52	0.29	4.59	98.00%
	BDSIR	999	1591	1142	0.39	0.24	3.42	99.00%
$z_{(0)}$	Stoch.Coal.SIR	(varies)	41.81	40.35	0.03	0.01	0.20	99.00%
	Deter.Coal.SIR	(varies)	41.17	39.99	0.03	0.01	0.07	76.00%
	BDSIR	(varies)	40.89	39.72	8.65E-4	-5.13E-4	3.43E-3	97.00%
$\mathcal{R}_0$	Stoch.Coal.SIR	1.50	1.48	1.37	0.09	-0.06	0.81	100.00%
	Deter.Coal.SIR	1.50	1.80	1.49	0.24	0.15	0.52	85.00%
	BDSIR	1.50	1.46	1.43	0.08	-0.03	0.47	99.00%
γ	Stoch.Coal.SIR	0.30	0.19	0.17	0.40	-0.40	1.06	85.00%
	Deter.Coal.SIR	0.30	0.26	0.23	0.27	-0.22	1.15	89.00%
	BDSIR	0.30	0.26	0.25	0.18	-0.18	0.72	97.00%
$S_{(0)}$	Stoch.Coal.SIR	499	599	390	0.25	-0.22	3.56	100.00%
	Deter.Coal.SIR	499	562	361	0.44	-0.26	3.36	91.00%
	BDSIR	499	996	714	0.51	0.49	4.63	100.00%
$z_{(0)}$	Stoch.Coal.SIR	(varies)	76.47	68.24	0.55	0.54	0.58	99.00%
	Deter.Coal.SIR	(varies)	91.03	72.51	0.39	0.38	0.42	88.00%
	BDSIR	(varies)	69.11	66.51	0.34	-0.31	0.20	94.00%
$\mathcal{R}_0$	Stoch.Coal.SIR	1.10	1.39	1.32	0.22	0.22	1.09	99.00%
	Deter.Coal.SIR	1.10	1.68	1.44	0.46	0.46	0.59	25.00%
	BDSIR	1.10	1.34	1.32	0.20	0.20	0.51	75.00%
γ	Stoch.Coal.SIR	0.25	0.17	0.15	0.37	-0.36	1.11	84.00%
	Deter.Coal.SIR	0.25	0.22	0.18	0.30	-0.22	1.16	86.00%
	BDSIR	0.25	0.28	0.26	0.12	0.09	0.92	100.00%
$S_{(0)}$	Stoch.Coal.SIR	499	608	398	0.24	-0.18	3.38	100.00%
	Deter.Coal.SIR	499	553	337	0.42	-0.26	3.08	92.00%
	BDSIR	499	1471	1040	1.21	1.21	6.52	99.00%
$z_{(0)}$	Stoch.Coal.SIR	(varies)	91.60	84.55	0.06	0.02	0.60	97.00%
	Deter.Coal.SIR	(varies)	112.79	90.37	0.26	0.26	0.94	85.00%
	BDSIR	(varies)	82.98	80.93	0.02	-0.01	0.08	88.00%

Table 3: Epidemic Parameter Inference from H1N1 Sequences in New Zealand

Inference Model	R_0	γ	S_0	Root of the tree (yr)	Origin z_0 of the epidemic (yr)
Stoch. Coal. SIR	1.46 (1.04 - 2.14)	27.08 (4.20 - 64.03)	6.90E4 (175 - 2.86E5)	0.53 (0.44 - 0.61)	0.69 (0.45 - 1.03)
Deter. Coal. SIR	1.35 (1.05 - 1.84)	34.50 (3.86 - 82.16)	1.20E5 (29 - 4.59E5)	0.54 (0.45 - 0.62)	0.73 (0.47 - 1.04)
BDSIR	1.61 (1.09 - 2.29)	27.72 (6.82 - 55.04)	2.22E4 (259 - 9.38E4)	0.49 (0.41 - 0.56)	0.53 (0.43 - 0.65)

Mean estimates (and 95% HPD intervals) of each epidemic parameter inferred from seasonal influenza A (H1N1) sequence data collected in the Canterbury region of New Zealand throughout the 2001 flu season.

Table 4: Bayesian Prior Distributions

	R_0	γ	$S_{(0)}$	$z_{(0)}$	$\psi/(\psi+\mu)$
$R_0 \approx 2.5$, $S_0 = 999$	LogN(1, 1)	LogN(-1, 1)	LogN(7, 1)	Unif(0, 100)	Beta(1,1)
$R_0 \approx 1.5$, $S_0 = 499$	LogN(0.5, 1)	LogN(-1, 1)	LogN(6, 1)	Unif(0, 500)	Beta(1,1)
$R_0 \approx 1.1$, $S_0 = 499$	LogN(0.1, 1)	LogN(-1.5, 1)	LogN(6, 1)	Unif(0, 500)	Beta(1,1)
* $R_0 \approx 1.1$, $S_0 = 999$	LogN(0.1, 1)	LogN(-1.5, 1)	LogN(7, 1)	Unif(0, 500)	–
* $R_0 \approx 1.1$, $S_0 = 1999$	LogN(0.1, 1)	LogN(-1.5, 1)	LogN(7.5, 1)	Unif(0, 500)	–
* $R_0 \approx 1.2$, $S_0 = 499$	LogN(0.2, 1)	LogN(-1, 1)	LogN(6, 1)	Unif(0, 500)	–
H1N1	Unif(0, 10)	LogN(3, 0.75)	LogN(13, 2)	Unif(0, 10)	Beta(1,1)
HIV-1	LogN(1, 1)	LogN(-1, 1)	LogN(7, 1)	Unif(0, 100)	Beta(1,1)

Prior distributions for the re-estimation of SIR parameters – the reproductive ratio R_0, the rate of removal γ, the number of susceptible individuals at the start of the epidemic $S_{(0)}$, the time of origin $z_{(0)}$, and the sampling proportion $\psi/(\psi+\mu)$ for BDSIR – from the simulated trees, seasonal influenza A (H1N1), and human immunodeficiency virus (HIV-1) data analyses. LogN(M, S) is a log-normal distribution with mean M and standard deviation S in log space. *Only applies to deterministic coalescent SIR, see details in the supporting information.