Bayesian phylodynamic inference with complex models

Erik Volz[1*], Igor Siveroni[1],

1 Department of Infectious Disease Epidemiology and MRC Centre for Outbreak Analysis and Modelling, Imperial College London, United Kingdom

* e.volz@imperial.ac.uk

Abstract

Population genetic modeling can enhance Bayesian phylogenetic inference by providing a realistic prior on the distribution of branch lengths and times of common ancestry.The parameters of a population genetic model may also have intrinsic importance, and simultaneous estimation of a phylogeny and model parameters has enabled phylodynamic inference of population growth rates, reproduction numbers, and effective population size through time. Phylodynamic inference based on pathogen genetic sequence data has emerged as useful supplement to epidemic surveillance, however commonly-used mechanistic models that are typically fitted to non-genetic surveillance data are rarely fitted to pathogen genetic data due to a dearth of software tools, and the theory required to conduct such inference has been developed only recently. We present a framework for coalescent-based phylogenetic and phylodynamic inference which enables highly-flexible modeling of demographic and epidemiological processes. This approach builds upon previous structured coalescent approaches and includes enhancements for computational speed, accuracy, and stability. A flexible markup language is described for translating parametric demographic or epidemiological models into a structured coalescent model enabling simultaneous estimation of demographic or epidemiological parameters and time-scaled phylogenies. We demonstrate the utility of these approaches by fitting compartmental epidemiological models to Ebola virus and Influenza A virus sequence data, demonstrating how important features of these epidemics, such as the reproduction number and epidemic curves, can be gleaned from genetic data. These approaches are provided as an open-source package *PhyDyn* for the BEAST phylogenetics platform.

Introduction

Mechanistic models guided by expert knowledge can form an efficient prior on epidemic history when conducting phylodynamic inference with genetic data [1]. Parameters estimated by fitting mechanistic models, such as the reproduction number R_0, are important for epidemic surveillance and forecasting. Compartmental models defined in terms of ordinary or stochastic differential equations are the most common type of mathematical infectious disease model, but in the area of phylodynamic inference, non-parametric approaches based on skyline coalescent models [2] or sampling-birth-death models [3] are more commonly used. Methods to translate compartmental infectious disease models into a population genetic framework have been developed only recently [4–8]. We address the gap in software tools for epidemic modeling and phylogenetic inference by developing a BEAST package, *PhyDyn*, which includes a highly-flexible mark-up language for defining compartmental infectious

disease models in terms of ordinary differential equations. This flexible framework enables phylodynamic inference with the majority of published compartmental models, such as the common susceptible-infected-removed (SIR) model [9] and its variants, which are often fitted to non-genetic surveillance data. The *PhyDyn* model definition framework supports common mathematical functions, conditional logic, vectorized parameters and the definition of complex functions of time and/or state of the system. The *PhyDyn* package can make use of categorical metadata associated with each sampled sequences, such as location of sampling, demographic attributes of an infected patient (age, sex), or clinical biomarkers. Phylogeographic models designed to estimate migration rates between spatial demes [10–12] are special cases within this modeling framework, and more complex phylogeographic models (e.g. time-varying or state-dependent population size or migration rates) can also be easily defined in this framework.

The development of *PhyDyn* was influenced by and builds upon previous efforts to incorporate mechanistic infectious disease models in BEAST. The *bdsir* BEAST package [13] implements a simple SIR model which is fitted using an approximation to the sampling-birth-death process. The *phylodynamics* BEAST package [14] includes simple deterministic and stochastic SIR models which can be fitted using coalescent processes. More recently, the *EpiInf* package has been developed which can fit stochastic SIR models using an exact likelihood with particle filtering [15]. These epidemic modeling packages are, however, limited to unstructured populations (no spatial, risk-group, or demographic population heterogeneity). Other packages have been developed for spatially structured populations with a focus on phylogeographic inference, especially with the aim of estimating pathogen migration rates between discrete spatial locations [16]. The *MultiTypeTree* BEAST package [10] implements the exact structured coalescent model with multiple demes and with constant effective population size in each deme and constant migration rates between demes. Two BEAST packages, *BASTA* [17] and *MASCOT* [11] have been independently developed to use fast approximate structured coalescent models related to approaches developed in [5]. These packages mirror the functionality of *MultiTypeTree*, enabling estimation of time-invariant effective population sizes and migration rates between spatial demes.

The *PhyDyn* BEAST package provides new functionality to the BEAST phylogenetics platform by implementing much more complex structured epidemic models. In a general compartmental model, neither the effective population size nor migration rate between demes need be constant, and in more general frameworks, coalescence is also allowed between lineages occupying different demes. The package includes a flexible mark-up language for compartmental models including common mathematical functions making it simple to develop models which incorporate seasonality or which deviate from the simplistic mass-action premise of basic SIR models. The *PhyDyn* model mark-up language supports vectorised parameters (e.g. an array of transmission rates or population sizes) and simple conditional logic statements, so that epidemic dynamics can change in a discrete fashion, such as from year to year or in response to a public-health intervention. Commonly used phylogeographic models based on the structured coalescent are a special case of the general compartmental models implemented in the *PhyDyn* package, and extensions to the basic phylogeographic model can be implemented, such as by allowing effective population size to vary through time in each deme according to a mechanistic model.

Design and Implementation

In this framework, first described in [5], we define deterministic demographic or epidemiological processes of a general form which includes the majority of

compartmental models used in mathematical epidemiology and ecology. Defining compartmental models within this form facilitates interpretation of the population genetic model developed in the next section. Let there be m demes, and the population size within each deme is given by the vector-valued function of time $Y_{1:m}(t)$. We may also have m' dynamic variables $Y'_{1:m'}(t)$ which are not demes (hence do not correspond to the state of a lineage), but which may influence the dynamics of Y. The dynamics of Y arise from a combination of *births* between and within demes, *migrations* between demes, and *deaths* within demes. We denote these as deterministic matrix-valued functions of time and the state of the system, following the framework in [5]:

- Births: $F_{1:m,1:m}(t, Y, Y')$. This may also correspond to transmission rates between different types of hosts in epidemiological models.
- Migrations: $G_{1:m,1:m}(t, Y, Y')$. These rates may have non-geographic interpretations in some models (e.g. aging, disease progression).
- Deaths: $\mu_{1:m}(t, Y, Y')$. These terms may also correspond to recovery in epidemiological models.

The elements $F_{kl}(\cdots)$ describe the rate that new individuals in deme l are generated by individuals in deme k. For example, this may represent the rate that infected hosts of type k transmit to susceptible hosts of type l. The elements $G_{kl}(\cdots)$ represent the rate that individuals in deme k change state to type l, but these rates do not describe the generation of new individuals. With the above functions defined, the dynamics of $Y(t)$ can be computed by solving a system of $m + m'$ ordinary differential equations:

$$\dot{Y}_k(t) = -\mu_k(t) + \sum_{l=1}^{m}(F_{lk}(t) + G_{lk}(t) - G_{kl}(t)) \tag{1}$$

The *PhyDyn* package model markup language requires specifying the non-zero elements of $F(t), G(t)$ and $\mu(t)$. There are multiple published examples of simple compartmental models developed in this framework [18–23]. In the following sections, we give examples of simple compartmental models related to infectious diease dynamics and show how these models can be defined within this framework. We provide examples of models fitted to data from seasonal human Influenza virus and Ebola virus as well as a simulation study.

Seasonal human Influenza model

We model a single season of Influenza A virus (IAV) H3N2 and apply this model to 102 HA-1 sequences collected between 2004 and 2005 in New York state [24, 25]. We build on a simple susceptible-infected-recovered (SIR) model which accounts for importations of lineages from the global reservoir of IAV, which we will see is a requirement for good model fit to these data (Figure 1). This model has two demes: The first deme corresponds IAV lineages circulating in New York, and the second deme corresponds to the global IAV reservoir. The global reservoir will be modeled as a constant-size coalescent process. Within New York state, new infections are generated at the rate $\beta I(t)S(t)/N$ where β is the per-capita transmission rate, $I(t)$ is the number of infected and infectious hosts, $S(t)$ is the number of hosts susceptible to infection, and $N = S + I + R$ is the population size. $R(t)$ denotes the number of hosts that have been infected and are now immune to this particular seasonal variant. With the above definitions, we define the matrix-valued function of time:

$$F(t) = \begin{bmatrix} \beta I(t)S(t)/N(t) & 0 \\ 0 & \gamma N_r \end{bmatrix} \tag{2}$$

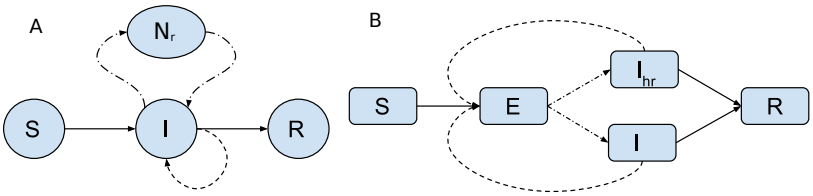


Fig 1. Compartmental diagram representing structure of models for seasonal human Influenza (A) and Ebola virus models (B and C). Solid lines represent flux of hosts between different categories. Dash lines represent migration. Dotted lines represent births (transmission).

Note that births within the reservoir do not vary through time and depend on the effective population size in that deme N_r.

Additionally, we model *deaths* from the pool of infected using

$$\mu(t) = \begin{bmatrix} \gamma I(t) \\ \gamma N_r \end{bmatrix} \tag{3}$$

Births balance deaths in the reservoir population.

Finally, we model a symmetric migration process between the reservoir and New York:

$$G(t) = \begin{bmatrix} 0 & \eta I(t) \\ \eta I(t) & 0 \end{bmatrix} \tag{4}$$

where η is the per-capita migration rate. Note that migration between the reservoir and New York are balanced and do not effect the dynamics of $I(t)$ over time.

These three processes lead to the following differential equation for the dynamics of $I(t)$

$$\dot{I}(t) = \beta I(t) S(t)/N(t) - \gamma I(t)$$

Below, we show a fit of this model where the following parameters are estimated:

- Migration rate η; prior: lognormal (log mean=1.38, log sd = 1)
- Recovery rate γ; prior: lognormal(log mean = 4.8, log sd = 0.25)
- Reproduction number $R_0 = \beta/\gamma$; prior: lognormal(log mean 0, log sd = 1)
- Reservoir size N_r; prior: lognormal(log mean = 9.2, log sd = 1)
- Initial number infected in September 2004; prior: lognormal(log mean = 0, log sd = 1)
- Initial number susceptible in September 2004; lognormal(log mean = 9.2, log sd =1)

Note that the model only had one informative prior, which was for the recovery rate, and was based on the previous study of viral shedding by Cori et al. [26]

Ebola Virus in Western Africa

We develop a susceptible-exposed-infected-recovered (SEIR) model (Figure 1) for the 2014-2015 Ebola Virus (EBOV) epidemic in Western Africa and apply this model to phylogenies previously estimated by Dudas et al. [27]. Phylogenies estimated by Dudas are randomly downsampled to $n = 400$ to alleviate computational requirements.

According to the SEIR model, infected hosts progress from an uninfectious exposed state (E) to an infectious state (I) at rate γ_0 which influences the generation-time distribution of the epidemic. Infectious hosts die or recover at the rate γ_1. The SEIR model has the following form:

$$\frac{d}{dt}E = \beta(t)I(t) - \gamma_0 E(t) \tag{5}$$

$$\frac{d}{dt}I = \gamma_0 E(t) - \gamma_1 I(t) \tag{6}$$

where $\beta(t)$ is the per-capita transmission rate. In a typical mass-action model, we would have $\beta(t) \propto S(t)/(S(t) + E(t) + I(t) + R(t))$, however in order to demonstrate the flexibility of this modeling framework, we will instead use a simple linear function, $\beta(t) = at + b$, and in general a wide variety of parametric and non-parametric functions could be used within the BEAST package to model the force of infection.

There are two demes in this model corresponding to the potential states of an infected hosts. The birth matrix with demes in the order (E, I) is

$$F(t) = \begin{bmatrix} 0 & 0 \\ \beta(t)I(t) & 0 \end{bmatrix} \tag{7}$$

The migration matrix encapsulates all processes which may change the state of a lineage without leading to coalescence of lineages, and this includes progression from E to I:

$$G(t) = \begin{bmatrix} 0 & \gamma_0 E(t) \\ 0 & 0 \end{bmatrix} \tag{8}$$

And finally removals are modeled using

$$\mu(t) = \begin{bmatrix} 0 \\ \gamma_1 I(t) \end{bmatrix} \tag{9}$$

Note that the parametric description of $\beta(t)$ does not require us to model dynamics of $S(t)$ or $R(t)$.

The parameters estimated and priors for this model are

- $\beta(t)$ slope a, prior: Normal(0, 40)
- $\beta(t)$ intercept b, prior: lognormal(log mean = 4.6, log sd = 1)
- Initial number infected (beginning of 2014), prior: lognormal (log mean=0, log sd = 1)

In order to reconstruct an epidemic trajectory which closely matched the absolute numbers of cases through time, we include additional variables that could influence the relationship between effective population size and the true number of infected hosts. For this purpose we developed a second EBOV model which included higher variance in the offspring distribution, reasoning that a higher variance in the number of transmissions per infected case would lead to higher estimates of the epidemic size [28]. The superspreading model (Figure 1) includes two infectious compartments, I_l and I_h, with per-capita transmission rates $\beta(t)$ and $\tau\beta(t)$ respectively. The factor of $\tau > 1$ represents a transmission risk ratio for the second infectious deme. We specify that a constant fraction p_{hr} progress from E to I_h, with the remainder going to I_l. With demes in the order (E, I_l, I_h), the birth, migration, and death matrices for the superspreading model

are as follows:

$$F(t) = \begin{bmatrix} 0 & 0 & 0 \\ \beta(t)I_l(t) & 0 & 0 \\ \tau\beta(t)I_h(t) & 0 & 0 \end{bmatrix}, \tag{10}$$

$$G(t) = \begin{bmatrix} 0 & (1-p_{hr})\gamma_0 E(t) & p_{hr}\gamma_0 E(t) \\ 0 & 0 & 0 \\ 0 & 0 & 0 \end{bmatrix}, \tag{11}$$

$$\mu(t) = \begin{bmatrix} 0 \\ \gamma_1 I_l(t) \\ \gamma_1 I_h(t) \end{bmatrix} \tag{12}$$

Additional parameters and priors for the superspreading model are

- τ, prior: lognormal(log mean = 1, log sd = 1)
- p_{hr}, fixed at 20%

Simulation model

We developed a simulation model with four demes in order to evaluate the ability of BEAST to identify and estimate birth rates, migration rates, and transmission risk ratios. This model includes two types of hosts, with low and high transmission risk. Additionally, each type of host progresses through two stages of infection, where the first stage is short but has higher transmission rate. The four demes are denoted $Y_{0l}, Y_{1l}, Y_{0h}, Y_{1h}$ where the first subscript denotes stage of infection and the second subscript denotes transmission risk level. The birth matrix is

$$F(t) = \begin{bmatrix} p_l f(t) w_0 Y_{0l}(t)/W(t) & 0 & (1-p_l) f(t) w_0 Y_{0l}(t)/W(t) & 0 \\ p_l f(t) Y_{1l}(t)/W(t) & 0 & (1-p_l) f(t) Y_{1l}(t)/W(t) & 0 \\ p_l f(t) w_0 w_h Y_{0h}(t)/W(t) & 0 & (1-p_l) f(t) w_0 w_h Y_{0l}(t)/W(t) & 0 \\ p_l f(t) w_h Y_{1h}(t)/W(t) & 0 & (1-p_l) f(t) w_h Y_{1h}(t)/W(t) & 0 \end{bmatrix}$$

In this model, a proportion p_l of all transmissions go to the low risk group. Transmissions from stage 1 are proportional to the transmission risk ratio $w_0 > 1$. Transmissions from the high risk group are proportional to the transmission risk ratio $w_h > 1$. The variable $W(t) = w_0 Y_{0l} + Y_{1l} + w_0 w_h Y_{0h} + w_h Y_{1h}$ normalizes the proportion of transmissions attributable to each deme. The variable $f(t)$ gives the total number of transmissions per unit time, and for this we use a SIRS model:

$$f(t) = \beta(Y_{0l} + Y_{1l} + Y_{0h} + Y_{1h})S/N$$

where $S(t)$ is the number susceptible governed by the following equation

$$\dot{S} = -f(t) + \eta S(0) - \eta S(t)$$

and η is the per-capita rate of non-disease related mortality.

The migration matrix captures the disease stage-progression process:

$$G(t) = \begin{bmatrix} 0 & \gamma_0 Y_{0l}(t) & 0 & 0 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & \gamma_0 Y_{0h}(t) \\ 0 & 0 & 0 & 0 \end{bmatrix}$$

The death matrix is

$$\mu(t) = \begin{bmatrix} \eta Y_{0l}(t) \\ (\eta + \gamma_1) Y_{1l}(t) \\ \eta Y_{0h} \\ (\eta + \gamma_1) Y_{1h}(t) \end{bmatrix}$$

To generate simulated data, we simulated epidemics using Gillespie's exact algorithm over a discrete population and an initial susceptible population of two thousand individuals. A random sample of $n = 250$ was collected between times 100 and 250 and the history of transmissions was used to reconstruct a genealogy. BEAST PhyDyn was then used to estimate

- β, prior: lognormal (log mean=-1.6, log sd = 0.5)
- w_0, prior: uniform(0, 50)
- w_h, prior: uniform(0, 50)
- The initial number infected, prior: lognormal (log mean=0, log sd = 1)

Note that BEAST PhyDyn is fitting deterministic models to data generated from a noisy stochastic process and some error should be expected due to this approximation. Supporting figure shows a comparison of a single noisy simulated trajectory and a solution of the deterministic model under the true parameters. All simulation code and BEAST XML files are available at
`https://github.com/emvolz/PhyDyn-simulations`.

Modeling the coalescent process conditioning on a complex demographic history

In this section we review the approximate structured coalescent model described in [5] and describe extensions designed to improve accuracy and reduce computational cost. The new model was first implemented in the *rcolgem* R package (2014) [29]. A special case of this model for phylogeography (constant N_e in demes and constant migration rate matrix) was independently developed by Mueller et al. [11].

The probability that a lineage i in a bifurcating roooted genealogy $\mathcal{G}$ is in deme $k \in 1:m$ at time t will be denoted $p_{ik}(t)$. Usually the state of a lineage will be observed at the time of sampling t_i, so that $p_{ik}(t_i)$ is a point density. We compute the likelihood by solving a system of differential equations for the ancestral states $p_{1:(2n-2),1:m}$ and computing the expected coalescent rate between each pair of lineages.

Let $\mathcal{A}(t)$ denote the set of extant lineages at time t. The expected number of lineages in each deme as a function of time is

$$A_k(t) = \sum_{i \in \mathcal{A}(t)} p_{ik}(t) \tag{13}$$

Given lineages i and $j \in \mathcal{A}(t)$, the rate of coalescence between the pair of lineages is a function of the ancestral state vectors $p_{i,1:m}(t)$ [5, 30]

$$\lambda_{ij}(t) = \sum_{k=1}^{m} \sum_{l=1}^{m} F_{kl}(t) \frac{p_{ik}(t) p_{jl}(t) + p_{jl}(t) p_{ik}(t)}{Y_k(t) Y_l(t)} \tag{14}$$

$$= \rho'_{i,1:m} F(t) \rho_{j,1:m} + \rho'_{j,1:m} F(t) \rho_{i,1:m} \tag{15}$$

where $\rho_{i,1:m}$ is a vector with elements $p_{ik}(t)/Y_k(t)$. Note that $\lambda_{ij} = \lambda_{ji}$. We will also define the rate

$$\tilde{\lambda}_{ij} = \rho'_{i,1:m} F(t) \rho_{j,1:m} \tag{16}$$

which is the rate that a lineage i begets a lineage j conditioning on the states of i and j. And we will refer to the rate

$$\tilde{\lambda}_{i\cdot}(t) = \sum_{j \neq i, j \in \mathcal{A}(t)} \tilde{\lambda}_{ij}(t) \tag{17}$$

which is the rate that i begets any other extant lineage. The total rate of coalescence at time t is

$$\lambda(t) = \sum_{i,j \in \mathcal{A}(t)} \tilde{\lambda}_{ij}(t) \tag{18}$$

using the convention that $\tilde{\lambda}_{ii} = 0$ for all i.

The probability $p(\mathcal{G}|\mathcal{M})$ of a given labeled genealogy given a demographic history $\mathcal{M} = (F, G, Y)$, described in previous publications [5, 30], is that of a point process with intensity $\lambda(t)$ multiplied by the the multinomial density with probabilities $\lambda_{ij}(t)/\lambda(t)$ for all pairs of lineages i and j which coalesce. The form of this likelihood is shared by all structured coalescent models and will not be reviewed further, however approximations to the structured coalescent differ in how ancestral state vectors are derived. In the original publication by Volz [5] in 2012, the following approximation (denoted "com12") was presented which required the solution of the following differential equations:

$$\frac{\mathrm{d}}{\mathrm{d}t} p_i^{com12}(t) = R^{com12}(t)' p_i(t) \tag{19}$$

where R is the $m \times m$ matrix with elements

$$\begin{aligned} R_{kl}^{com12}(t) &= (F_{lk}(t) \frac{Y_l(t) - A_l(t)}{Y_l(t)} + G_{lk}(t))/Y_k(t), \quad k \neq l \\ R_{kk}^{com12}(t) &= -\sum_{l \neq k} R_{kl}^{com12}(t) \end{aligned} \tag{20}$$

Note the inclusion of the term $\frac{Y_l(t) - A_l(t)}{Y_l(t)}$ in equation 20, which was an approximation intended to account for the fact that only lineages not ancestral to the sample could cause a lineage to change state without resulting in coalescence. The form of this equation was found to provide an accurate approximation to the lineages through time in [5], when solving the system of equations

$$\frac{\mathrm{d}}{\mathrm{d}t} A(t) = R^{com12}(t)' A(t) \tag{21}$$

Here we describe an improved continuous time Markov chain (CTMC) model for $p_i(t)$ which employs a different strategy for conditioning on the absence of coalescent events along each lineage. Let $\tilde{p}_i$ represent an augmented state vector where $\tilde{p}_{ik} = p_{ik}$ for $k < m + 1$ and $\tilde{p}_{i,m+1}$ represents the probability that lineage i has coalesced. Note that $m + 1$ is an absorbing state which occurs at the rate $\tilde{\lambda}_{i\cdot}(t)$ (equation 17). The

$m+1 \times m+1$ rate matrix R has elements

$$\begin{aligned} R_{kl}(t) &= (F_{lk}(t) + G_{lk}(t))/Y_k(t), \quad \text{if } k \neq l, k < m+1, l < m+1 \\ R_{k,m+1}(t) &= \sum_{j \neq i, j \in \mathcal{A}(t)} \sum_{l=1}^{m} F_{kl}(t)\rho_{ik}(t)\rho_{jl}(t), \quad \text{if } k < m+1 \\ R_{kl}(t) &= 0, \quad \text{if } k = m+1 \\ R_{kk}(t) &= -\sum_{l \neq k} R_{kl} \end{aligned} \tag{22}$$

With $\tilde{P}(t)$ representing the $(m+1) \times (2n-2)$ matrix of state vectors with columns $(\tilde{p}_1, \cdots, \tilde{p}_{2n-2})$, we can solve for the state vectors with the following system of equations

$$\frac{\mathrm{d}}{\mathrm{d}t}\tilde{P}(t) = R'(t)\tilde{P}(t) \tag{23}$$

using the convention that $p_{ik}(t) = 0$ for all lineages $i \notin \mathcal{A}(t)$.

Note that if the rate of coalescence is non-zero over the history a lineage, $\sum_{k=1}^{m} \tilde{p}_{ik} < 1$. If the ancestor node of a lineage i occurs at a time T_i, we derive $p_i(T_i)$ by renormalizing the distribution computed from equation 23, which provides the state vector conditional on the event that no coalescence was observed:

$$p_{ik}(T_i) = \frac{\tilde{p}_{ik}}{1 - \tilde{p}_{i,m+1}} \tag{24}$$

In [11], a system of equations equivalent to 23 was derived for the special case of a phylogeographic model, which corresponds to non-zero diagonal $F(t)$, time-invariant $Y(t)$ and non-zero off-diagonal $G(t)$, and this system was found to provide a very close approximation to the stochastic structured coalescent.

Unfortunately, the system of equations 23 can be slow to solve since it requires recursion over extant lineages (twice) and $m+1$ ancestral states (equation 22). We therefore provide an additional approximation which greatly reduces computational cost and is closely related to the approach described in [5]. We define $Q(t,T)$ to be the $m+1 \times m+1$ matrix of transition probabilities such that

$$\tilde{P}^{fast}(T) = Q'(t,T)\tilde{P}^{fast}(t) \tag{25}$$

and $Q(t,t) = I$ is the identity matrix. We can approximate the number of lineages in each deme over the interval using

$$A(\tau) = Q'(t,\tau)A(t) \tag{26}$$

where $t < \tau < T$. Finally, we can modify equation 22 to use the vector $A(\tau)$, avoiding the need to sum over extant lineages:

$$\begin{aligned} R^{fast}_{kl}(t) &= (F_{lk}(t) + G_{lk}(t))/Y_k(t), \quad k \neq l, k < m+1, l < m+1 \\ R^{fast}_{k,m+1}(t) &= \rho_{ik}(t)F'_{k\cdot}(t)\frac{A(t)}{Y(t)}, \quad k < m+1 \\ R^{fast}_{kl}(t) &= 0, \quad k = m+1 \\ R^{fast}_{kk}(t) &= -\sum_{l \neq k} R_{kl} \end{aligned} \tag{27}$$

And $Q(t,T)$ is computed by solving the equations

$$\frac{\mathrm{d}}{\mathrm{d}\tau}Q(t,\tau) = R^{fast}(\tau)'Q(t,\tau) \tag{28}$$

Therefore only $m^2 + m$ differential equations need to be solved over every internode interval. Note that $R^{fast}_{k,m+1}(t) - R_{k,m+1}(t) = F'_{k\cdot}\rho_i$, so that this fast approximation will slightly over-estimate the probability that a lineage coalesces over an interval.

We will denote the model based on equations 22 *comP* and the model based on equations 27 *comQ*. Performance of *comQ* and *comP* is explored in simulations below.

Results

With simulated data, BEAST PhyDyn recovers the correct transmission risk ratios and transmission rates using both the *comP* model (equation 22) and the faster *comQ* model (27). Figure 2 compares estimates across twenty simulations using both variants. The running time of the *comQ* model was approximately five times faster than *comP* in these simulations with trees that have 250 samples and four demes. Good coverage of parameter estimates was observed with the *comP* model. Across 60 parameter estimates (three parameters not counting initial conditions and twenty simulations), estimates did not cover the true value two times. The *comQ* model failed to cover in five of 60 estimates. Greater bias was observed with the *comQ* model, with the greatest bias observed for the w_h parameter (cf equation 13, mean estimate 3.63, truth:5). However the *comQ* model also had superior precision, with a posterior root mean square error of 2.4 versus 4.8 observed with the *comP* model. A similar but less pronounced pattern of bias and precision was observed for other parameters. A complete summary of simulation results is available at `https://github.com/emvolz/PhyDyn-simulations`.

Human Influenza A/H3N2

The seasonal influenza SIR model which accounts for importations from the global reservoir was applied to 102 HA/H3N2 sequences collected from New York state during the 2004-2005 flu season. These data were previously analyzed using non-parametric models by [24]. Figure 3 shows the estimated posterior effective number of infections over the course of the influenza season, and the time of peak prevalence is correctly identified around the end of 2004. We also compared the model-based estimates to non-parametric estimates generated in BEAST using a conventional non-parametric Bayesian skyline model which is also shown in Figure 3. The skyline model does not detect a decrease in prevalence towards the end of the influenza season and does not identify the time of peak prevalence. Use of a well-specified parametric compartmental model imposes a strong prior on the epidemic trajectory which leads to the correct identification of the shape and timing of the epidemic curve.

We estimated the reproduction number $R_0 = 1.16$ (95%CI: 1.07-1.30). This value is similar to many previous estimates based on non-genetic data for seasonal influenza in humans which according to the recent review in [31] have an interquartile range of 1.18-1.27 for H3N2. Bettancourt et al. [32] estimated $R_0 = 1.22$ for the 2004-05 H3N2 seasonal influenza epidemic in the entire USA using weekly case report data. An uninformative prior was used for R_0 in the BEAST/PhyDyn analysis.

Ebola virus in Western Africa

We applied the SEIR and superspreading-SEIR models to Ebola virus phylogenies based on data first described by [27] and subsequently analyzed in [33]. These phylogenies were estimated from whole genome sequences collected 2014-2015 during the West African Ebola epidemic. We derived the maximum clade credibility tree from the analysis by [27] and extracted a subtree based on sampling four hundred lineages at

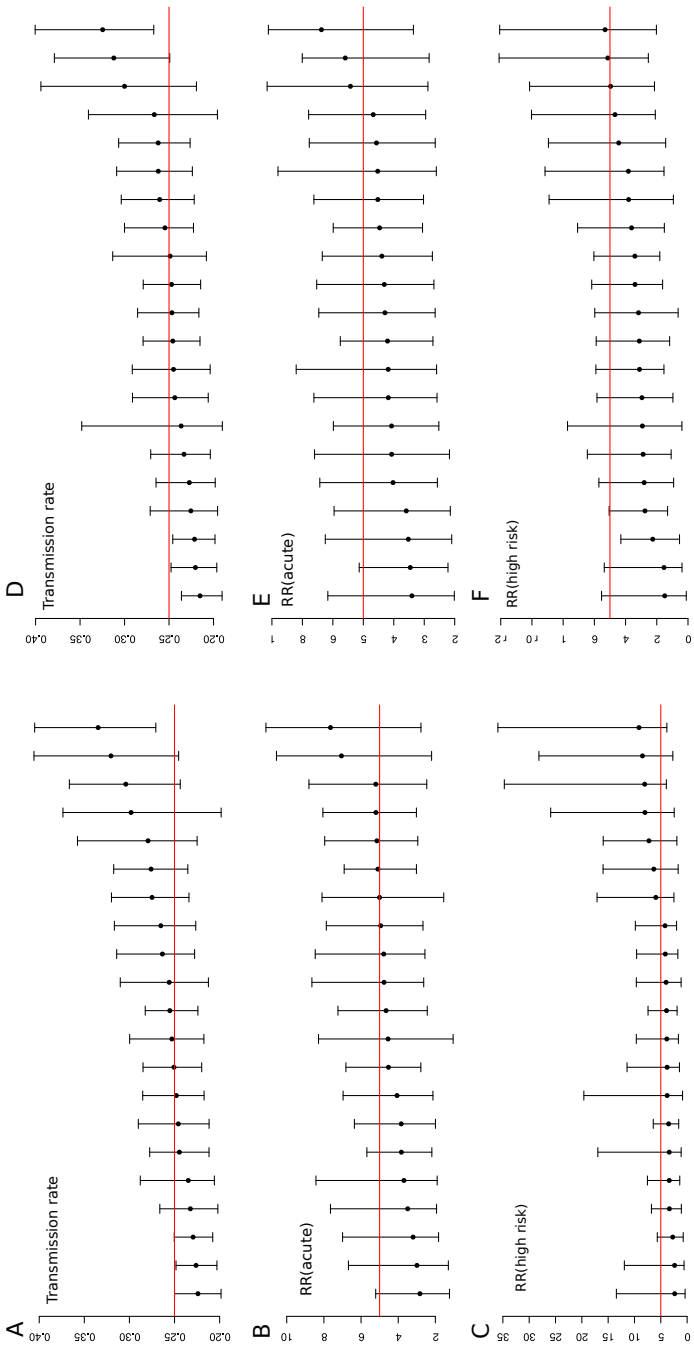


Fig 2. Parameter estimates and credible intervals for 20 simulations. The red line shows the true value. A-C: Results generated using the *comP* model. D-F: Results generated using the *comQ* model.

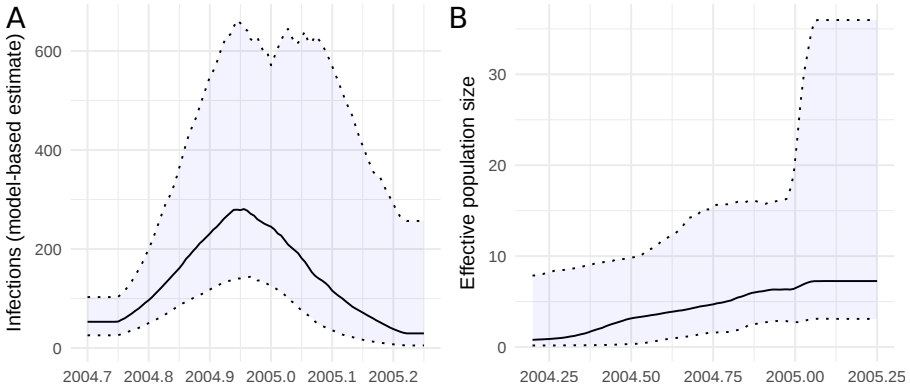


Fig 3. The estimated effective number of H3N2 human influenza infections in 2004-2005 in New York State. A. Estimates obtained using the parametric seasonal influenza model described in the text. B. Effective population size estimated using a conventional Bayesian skyline analysis.

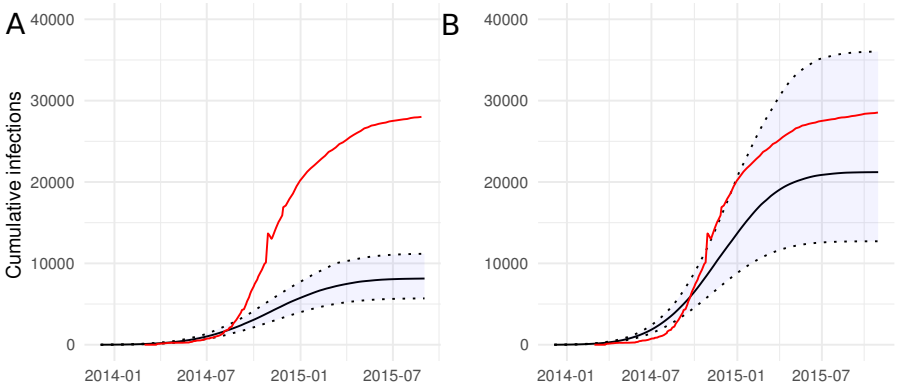


Fig 4. Model-based estimates of cumulative infections through time for the 2014-15 Ebola epidemic in Western Africa. Estimates are shown for the SEIR model (A) and the model which includes super-spreading (B). The red line show the cumulative number of cases reported by WHO [33].

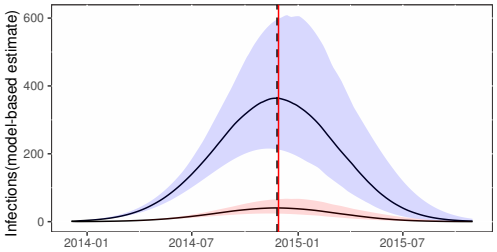


Fig 5. Estimated effective number of infections through time using the superspreading SEIR model for the 2014-15 Ebola epidemic in Western Africa. The red vertical line shows the time of peak prevalence inferred from WHO case reports. The vertical dashed line shows the model estimated time of peak prevalence. The red trajectory shows the proportion of infections in the high-transmission-rate compartment.

random. The BEAST PhyDyn package was used to fit the models with fixed tree topologies and branch lengths. We also ran the analysis using a fixed tree estimated by maximum likelihood and the *treedater* R package as described in [33], finding similar results.

We estimated similar reproduction numbers using both models. With the SEIR model, we estimate $R_0 = 1.47$(95%CI: 1.41-1.53). With the superspreading-SEIR model, we estimate $R_0 = 1.52$(95%CI:1.48-1.54). Note that uninformative priors were used for parameters determining R_0. As anticipated, the model fits provide substantially different estimates of the cumulative number of infections. Figure 4 shows the estimated cumulative infections through time using both models alongside the cumulative number of cases reported by WHO and compiled by the US CDC [33]. Both models provide similar estimates regarding the relative numbers infected through time and the time of epidemic peak. Using the superspreading model, the time of peak incidence is estimated to have occurred on November 25, 2014. According to WHO reports, this occurred only three days later on November 28 (Figure 5.

Estimates of cumulative infections with the superspreading model are consistent with WHO data, whereas results with the SEIR model are not. The superspreading model accomodates an over-dispersed offspring distribution (the number of transmission per infection), thereby decreasing effective population size per number infected and yielding larger estimates for the number infected [28]. We estimate the transmission risk ratio parameter (ratio of transmission rates between high and low compartments) to be 8.1 (95%CI: 6.68-10.73). This implies that a minority of 10% of infected individuals are responsible for 43%-54% of infections.

Formal model comparison methods such as Bayesian stepping-stone approaches [34] are not yet supported by the *PhyDyn* package, but we note that a much higher mean posterior likelihood was found using the superspreading model (-1006.9) than with the SEIR model (-1068.5).

Availability and Future Directions

The *PhyDyn* package, source code, documentation and examples can be found at `https://github.com/mrc-ide/PhyDyn`. The *PhyDyn* package greatly expands the range of epidemiological, ecological, and phylogeographic models that can be fitted within the BEAST Bayesian phylogenetics framework. Extensions enabled by this package include models with parametric seasonal forcing, non-constant parametric migration or coalescent rates between demes, state-dependent migration or coalescent rates, and discrete changes in migration or coalescent rates in response to perturbation of the system (e.g. a public health intervention). The package also provides a means of utilizing non-geographic categorical metadata which is usually not considered in phylodynamic analyses, such as clinical or demographic attributes of patients in a viral phylodynamics application [19].

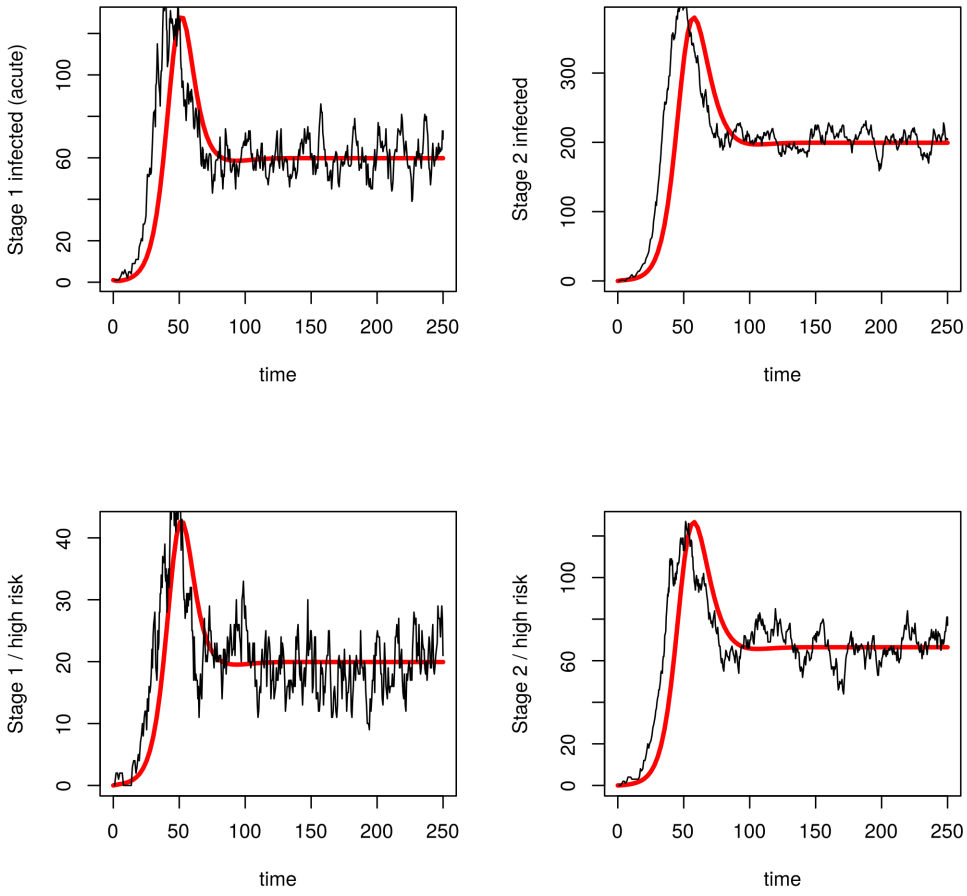
We have demonstrated the utility of this framework using data from Influenza and Ebola virus epidemics in humans, finding epidemic parameters and epidemic trajectories consistent with other surveillance data. In both of these examples, simple structured models were fitted, but notably without using any categorical metadata associated with sampled sequences. This demonstrates potential advantages of structured coalescent modeling even in the absence of informative metadata. In the case of human Influenza A virus, the fitted model included a deme which accounted for evolution in the unsampled global influenza reservoir, which allowed estimation of epidemic parameters within the smaller sub-region which was intensively sampled. The use of a parametric mass-action model allowed *PhyDyn*to correctly detect the time of epidemic peak and epidemic decline, whereas non-parametric skyline methods did not detect epidemic

decline in this case. And in the application to the Ebola virus epidemic in Western Africa, models included un-sampled 'exposed' categories which accounted for realistic progression of disease among patients, as well as a 'super-spreading' compartment which accounted for over-dispersion in the number of transmissions per infected case.

In developing *PhyDyn*, the focus has been on developing a highly flexible framework which is also computationally tractable for moderate sample sizes and model complexity. But flexibility and computational efficiency has come at the cost of some realism, notably in the deterministic nature of the models included in this framework. Future extensions may utilize stochastic epidemic models such as those described by [30]. Other directions for future development include semi-parametric modeling, such as models with a spline-valued force of infection [22] or models utilizing Gaussian processes [35], and approaches for utilizing continuous-valued metadata [36].

Supporting information

S1 Fig. Comparison of stochastic and deterministic trajectories. The stochastic epidemic simulation is shown in black and the deterministic ODE model is shown in red.



Acknowledgments

The authors thank Tim Vaughan for helpful comments and suggestions.

References

1. Volz EM, Koelle K, Bedford T. Viral phylodynamics. PLoS Comput Biol. 2013;9(3):e1002947.

2. Drummond AJ, Rambaut A, Shapiro B, Pybus OG. Bayesian coalescent inference of past population dynamics from molecular sequences. Mol Biol Evol. 2005;22(5):1185–1192.

3. Stadler T, Kühnert D, Bonhoeffer S, Drummond AJ. Birth–death skyline plot reveals temporal changes of epidemic spread in HIV and hepatitis C virus (HCV). Proceedings of the National Academy of Sciences. 2013;110(1):228–233.

4. Volz EM, Kosakovsky Pond SL, Ward MJ, Leigh Brown AJ, Frost SDW. Phylodynamics of infectious disease epidemics. Genetics. 2009;183(4):1421–1430.

5. Volz EM. Complex population dynamics and the coalescent under neutrality. Genetics. 2012;190(1):187–201.

6. Frost SDW, Volz EM. Viral phylodynamics and the search for an 'effective number of infections'. Philos Trans R Soc Lond B Biol Sci. 2010;365(1548):1879–1890.

7. Dearlove B, Wilson DJ. Coalescent inference for infectious disease: meta-analysis of hepatitis C. Philos Trans R Soc Lond B Biol Sci. 2013;368(1614):20120314.

8. Smith RA, Ionides EL, King AA. Infectious Disease Dynamics Inferred from Genetic Data via Sequential Monte Carlo. Mol Biol Evol. 2017;34(8):2065–2084.

9. Anderson RM, May RM, Anderson B. Infectious diseases of humans: dynamics and control. 1992;.

10. Vaughan TG, Kühnert D, Popinga A, Welch D, Drummond AJ. Efficient Bayesian inference under the structured coalescent. Bioinformatics. 2014;30(16):2272–2279.

11. Mueller NF, Rasmussen DA, Stadler T. MASCOT: Parameter and state inference under the marginal structured coalescent approximation; 2017.

12. Beerli P, Felsenstein J. Maximum-likelihood estimation of migration rates and effective population numbers in two populations using a coalescent approach. Genetics. 1999;152(2):763–773.

13. Kühnert D, Stadler T, Vaughan TG, Drummond AJ. Simultaneous reconstruction of evolutionary history and epidemiological dynamics from viral sequences with the birth–death SIR model. J R Soc Interface. 2014;11(94):20131106.

14. Drummond AJ, Bouckaert RR. Bayesian Evolutionary Analysis with BEAST. Cambridge University Press; 2015.

15. Vaughan TG, Leventhal GE, Rasmussen DA, Drummond AJ, Welch D, Stadler T. Directly Estimating Epidemic Curves From Genomic Data; 2017.

16. Lemey P, Rambaut A, Drummond AJ, Suchard MA. Bayesian phylogeography finds its roots. PLoS Comput Biol. 2009;5(9):e1000520.

17. De Maio N, Wu CH, O'Reilly KM, Wilson D. New Routes to Phylogeography: A Bayesian Structured Coalescent Approximation. PLoS Genet. 2015;11(8):e1005421.

18. Rasmussen DA, Boni MF, Koelle K. Reconciling phylodynamics with epidemiology: the case of dengue virus in southern Vietnam. Mol Biol Evol. 2014;31(2):258–271.

19. Volz EM, Ionides E, Romero-Severson EO, Brandt MG, Mokotoff E, Koopman JS. HIV-1 transmission during early infection in men who have sex with men: a phylodynamic analysis. PLoS Med. 2013;10(12):e1001568; discussion e1001568.

20. Volz EM, Ndembi N, Nowak R, Kijak GH, Idoko J, Dakum P, et al. Phylodynamic analysis to inform prevention efforts in mixed HIV epidemics. Virus Evol. 2017;3(2):vex014.

21. Volz E, Pond S. Phylodynamic analysis of ebola virus in the 2014 sierra leone epidemic. PLoS Curr. 2014;6.

22. Ratmann O, Hodcroft EB, Pickles M, Cori A, Hall M, Lycett S, et al. Phylogenetic Tools for Generalized HIV-1 Epidemics: Findings from the PANGEA-HIV Methods Comparison. Mol Biol Evol. 2017;34(1):185–203.

23. Poon AFY. Phylodynamic Inference with Kernel ABC and Its Application to HIV Epidemiology. Mol Biol Evol. 2015;32(9):2483–2495.

24. Karcher MD, Palacios JA, Bedford T, Suchard MA, Minin VN. Quantifying and mitigating the effect of preferential sampling on phylodynamic inference. PLoS Comput Biol. 2016;12(3):e1004789.

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

26. Cori A, Valleron AJ, Carrat F, Scalia Tomba G, Thomas G, Boëlle PY. Estimating influenza latency and infectious period durations using viral excretion data. Epidemics. 2012;4(3):132–138.

27. Dudas G, Carvalho LM, Bedford T, Tatem AJ, Baele G, Faria NR, et al. Virus genomes reveal factors that spread and sustained the Ebola epidemic. Nature. 2017;544(7650):309–315.

28. Koelle K, Rasmussen DA. Rates of coalescence for common epidemiological models at equilibrium. J R Soc Interface. 2012;9(70):997–1007.

29. Volz EM. rcolgem: statistical inference and modeling of genealogies generated by epidemic and ecological processes. R package version 0.0. 5/r154. 2016;.

30. Rasmussen DA, Volz EM, Koelle K. Phylodynamic inference for structured epidemiological models. PLoS Comput Biol. 2014;10(4):e1003570.

31. Biggerstaff M, Cauchemez S, Reed C, Gambhir M, Finelli L. Estimates of the reproduction number for seasonal, pandemic, and zoonotic influenza: a systematic review of the literature. BMC Infect Dis. 2014;14:480.

32. Bettencourt LMA, Ribeiro RM. Real time bayesian estimation of the epidemic potential of emerging infectious diseases. PLoS One. 2008;3(5):e2185.

33. Volz EM, Frost SDW. Scalable relaxed clock phylogenetic dating. Virus Evol. 2017;3(2).

34. Baele G, Li WLS, Drummond AJ, Suchard MA, Lemey P. Accurate model selection of relaxed molecular clocks in Bayesian phylogenetics. Molecular biology and evolution. 2012;30(2):239–243.

35. Palacios JA, Minin VN. Gaussian Process-Based Bayesian Nonparametric Inference of Population Size Trajectories from Gene Genealogies. Biometrics. 2013;.

36. Lemey P, Rambaut A, Welch JJ, Suchard MA. Phylogeography takes a relaxed random walk in continuous space and time. Mol Biol Evol. 2010;27(8):1877–1885.