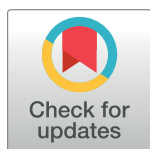RESEARCH ARTICLE

Bayesian phylodynamic inference with complex models

Erik M. Volz [ID] *, **Igor Siveroni** [ID]

Department of Infectious Disease Epidemiology and the MRC Centre for Global Infectious Disease Analysis, Imperial College London, 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 BEAST2 phylogenetics platform.

This is a *PLOS Computational Biology* Software paper.

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 such as 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 BEAST2 package, *PhyDyn*, which includes a highly-flexible markup 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 BEAST2. The *bdsir* BEAST2 package [13] implements a simple SIR model which is fitted using an approximation to the sampling-birth-death process. The *phylodynamics* BEAST2 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* BEAST2 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 BEAST2 packages, *BASTA* [17] and *MASCOT* [11] have been independently developed to use fast approximate structured coalescent models. These packages mirror the functionality of *MultiTypeTree* but include approximations to reduce computational requirements, enabling estimation of time-invariant effective population sizes and migration rates between spatial demes.

The *PhyDyn* BEAST2 package provides new functionality to the BEAST2 phylogenetics platform by implementing a much more complex family of structured coalescent 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 markup language for defining compartmental models within the BEAST2 XML. This includes 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. Models defined with this special syntax can be directly incorporated into BEAST2 XML files for easily reproducing and modifying analyses. The *PhyDyn* model markup 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 and code samples are also provided online. 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 (Fig 1). This model has two demes: The first deme corresponds to 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 per day, $I(t)$ 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:

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

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

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

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.

PhyDyn code for defining these equations can be found at https://github.com/mrc-ide/PhyDyn/wiki/Influenza-Example.

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 (events per year): lognormal (log mean = 1.38, log sd = 1)

- Recovery rate γ; prior (events per year): 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] Previous work [27] on identifiability of parameters in phylodynamic models has shown that it is generally impossible to simultaneously infer transmssion and recovery rates without additional data or strong assumptions about the sampling rate.

Ebola virus in Western Africa

We develop a susceptible-exposed-infected-recovered (SEIR) model (Fig 1) for the 2014-2015 Ebola Virus (EBOV) epidemic in Western Africa and apply this model to phylogenies
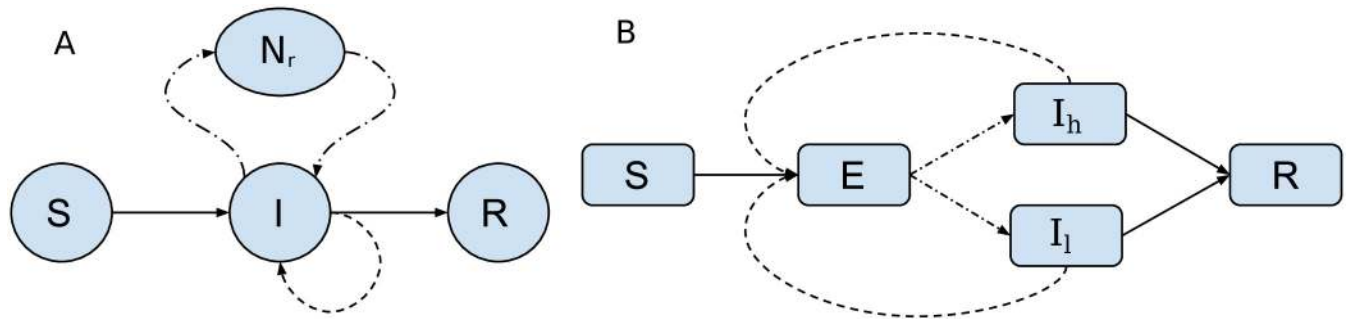
Fig 1. Compartmental diagram representing structure of models for seasonal human Influenza (A) and Ebola virus model (B). Solid lines represent flux of hosts between different categories. Dash lines represent migration. Dotted lines represent births (transmission).

previously estimated by Dudas et al. [28]. 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:

$$\frac{d}{dt}E = \beta(t)I(t) - \gamma_0 E(t),$$
$$\frac{d}{dt}I = \gamma_0 E(t) - \gamma_1 I(t). \tag{5}$$

where $\beta(t)$ is the per-capita transmission rate per year. 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 BEAST2 package to model the force of infection. In addition to demonstrating the flexibility of *PhyDyn*, we chose the affine transmission rate model because the mass action assumption is unrealistic and unnecessary. The number of susceptible individuals was never a limiting factor in this epidemic and incidence declined primarily in response to public health interventions.

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

$$F(t) = \begin{bmatrix} 0 & 0 \\ \beta(t)I(t) & 0 \end{bmatrix}. \tag{6}$$

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:

$$G(t) = \begin{bmatrix} 0 & \gamma_0 E(t) \\ 0 & 0 \end{bmatrix} \tag{7}$$

And finally removals are modeled using

$$\mu(t) = \begin{bmatrix} 0 \\ \gamma_1 I(t) \end{bmatrix}. \tag{8}$$

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

PhyDyn code for defining these equations can be found at https://github.com/mrc-ide/PhyDyn/wiki/Ebola-Example.

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 [29]. The superspreading model ([Fig 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:

$$F(t) = \begin{bmatrix} 0 & 0 & 0 \\ \beta(t)I_l(t) & 0 & 0 \\ \tau\beta(t)I_h(t) & 0 & 0 \end{bmatrix}, \tag{9}$$

$$G(t) = \begin{bmatrix} 0 & (1-p_{hr})\gamma_0 E(t) & p_{hr}\gamma_0 E(t) \\ 0 & 0 & 0 \\ 0 & 0 & 0 \end{bmatrix}, \tag{10}$$

$$\mu(t) = \begin{bmatrix} 0 \\ \gamma_1 I_l(t) \\ \gamma_1 I_h(t) \end{bmatrix}. \tag{11}$$

Additional parameters and priors for the superspreading model are

- τ, prior: lognormal(log mean = 1, log sd = 1)

- p_{hr}, fixed at 20%

Note that we used an uninformative prior for τ as our previous studies with a related model showed that superspreading parameters are potentially identifiable [21]. This model did not include geographic structure, although the samples were geographically diverse, and some model-misspecification bias is anticipated if migration between spatial demes is sufficiently small.

Simulation model

We developed a simulation model with four demes in order to evaluate the ability of BEAST2 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 model is illustrated as S1 Fig.

The birth matrix is:

$$F(t) = \begin{bmatrix} p_l f(t) w_0 Y_{0l}(t)/W(t) & 0 & (1-p_l)f(t)w_0 Y_{0l}(t)/W(t) & 0 \\ p_l f(t) Y_{1l}(t)/W(t) & 0 & (1-p_l)f(t)Y_{1l}(t)/W(t) & 0 \\ p_l f(t) w_0 w_h Y_{0h}(t)/W(t) & 0 & (1-p_l)f(t)w_0 w_h Y_{0l}(t)/W(t) & 0 \\ p_l f(t) w_h Y_{1h}(t)/W(t) & 0 & (1-p_l)f(t)w_h Y_{1h}(t)/W(t) & 0 \end{bmatrix}. \tag{12}$$

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:

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

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

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

PhyDyn code for implementing this model can be found at https://git.io/ftjg5.

To generate simulated data, we simulated epidemics using Gillespie's exact algorithm over a discrete population and an initial susceptible population of two or five thousand individuals.

A random sample of n = 250 or 500 was collected between times 95 and 250 and the history of transmissions was used to reconstruct a genealogy. *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 *PhyDyn* is fitting deterministic models to data generated from a noisy stochastic process and some error should be expected due to this approximation. S2 Fig shows a comparison of a single noisy simulated trajectory and a solution of the deterministic model under the true parameters. All simulation code and BEAST2 XML files are available at https://github.com/emvolz/PhyDyn-simulations.

Modeling the coalescent process conditioning on a complex demographic history

The coalescent likelihood is based on the conditional density of a genealogy given epidemic and demographic parameters. In BEAST2, the coalescent likelihood is used in tandem with evolutionary models that provide the probability density of a genealogy given a genetic sequence alignment and evolutionary parameters. But the coalescent likelihood can also be used if a time-scaled phylogeny has been estimated independently.

Various approximations have been developed for computing the density of a genealogy conditional on a complex demographic history. These differ by the extent to which they account for correlation between co-existing lineages in the genealogy, the extent to which they account for finite size of the population, and the extent to which they account for differences in coalescent rates in different demes. There is a speed/bias tradeoff between these approximations, and consequently *PhyDyn* makes several model variations available. The choice of likelihood approximation depends on time and computational resources available, sample size, and model complexity. Three likelihood approximations are described in S1 Text, and we derive a new approximation which has shown greater accuracy in some situations.

The structured coalescent model in [5] which inspired the development of *PhyDyn* did not account for all correlations between co-existing lineages or all effects stemming from disparate coalescent rates between demes. In [20], a fast likelihood approximation was derived which better accounted for potential bias resulting from highly-disparate coalescent rates in different demes. This model, denoted *QL*, also makes strong approximations regarding lineage independence: In every internode interval, all lineages are updated according to a linear transformation which varies through time but not between lineages. These issues were investigated as a source of bias in the context of phylogeographic models in [30], where yet another likelihood approximation was proposed for models with constant population size and constant migration rates.

In the PhyDyn package, we have developed likelihood approximations based on *QL* which better account for correlation between lineages. These models, denoted *PL1* and *PL2*, work by solving a system of differential equations for each lineage while including terms similar to those in the *QL* model that account for disparate coalescent rates between demes. While these models are demonstrably more accurate in simulation studies, they require more computation. All three likelihood approximations are provided in the PhyDyn package. The new *PL2* model is the suggested default model choice, however the *QL* model may be preferred for some large

datasets or when fitting complex models due to computational advantages. The new models are derived in S1 Text.

Results

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]. Fig 2 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 estimates generated in BEAST2 using a conventional non-parametric Bayesian skyline model which is also shown in Fig 2. 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. We carried out a further comparison with estimates using a GMRF skyride model fitted in BEAST 1.8 [31, 32] (S3 Fig). The skyride model correctly detected a peak in N_e in late 2014 and subsequent decline, however variation $N_e(t)$ was quite small relative to uncertainty in the credible intervals. The peak of N_e was slightly too early, and N_e was also larger prior to the 2014-15 influenza season due to the effects of unmodeled lineage importation from outside New York. Skyline and skyride analysis data and files are available at https://github.com/emvolz/nyflu-skyline. 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 [33] have an interquartile range of 1.18-1.27 for H3N2. Bettancourt et al. [34] 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 *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 [28] and subsequently analyzed in [35]. These phylogenies were
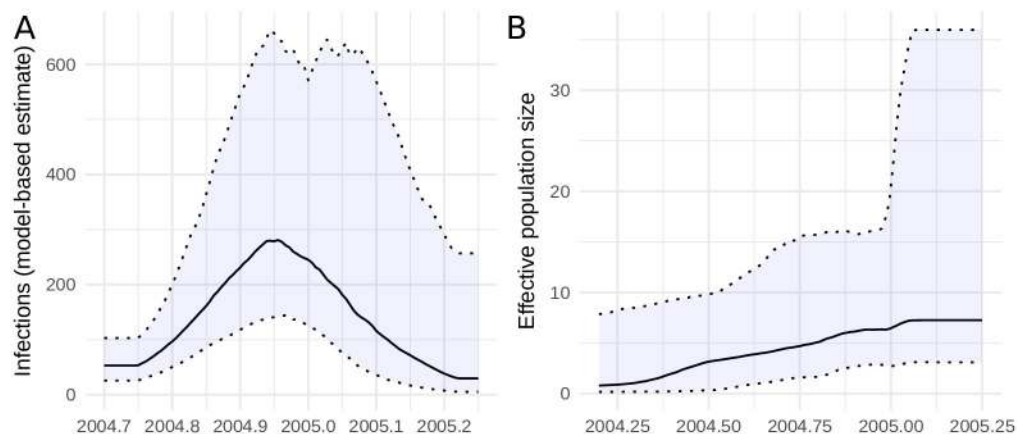

Fig 2. The estimated effective number of H3N2 human influenza infections in 2004-2005 in New York State. A. Estimates obtained using the parametric seasonal influenza model described in the text. B. Effective population size estimated using a conventional Bayesian skyline analysis.

https://doi.org/10.1371/journal.pcbi.1006546.g002

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 [28] and extracted a subtree based on sampling four hundred lineages at random. The *PhyDyn* package was used to fit the models with fixed tree topologies and branch lengths. Co-estimating the phylogeny and epidemic parameters is possible and may lead to more robust credible intervals because the tree prior can influence the topology of the estimated posterior distribution of trees, but this would also require substantialy more computational effort. The trees were fixed in this analyis in order to facilitate comparisons with other software and because of computational tradeoffs. With this fixed tree, *PhyDyn* executes approximately one million MCMC steps per 17 hours using a typical CPU. We also ran the analysis using a fixed tree estimated by maximum likelihood and the *treedater* R package as described in [35], finding similar results.

The transmission rate (per year) $\beta(t)$ was estimated as a linear function with slope -13.22 (95%CI:-14.4587- -12.036) and intercept 85.1(95% CI: 83.93-86.16). We estimated similar reproduction numbers using both models. With the SEIR model, we compute $R_0 = \beta(t)/\gamma_1$. We estimate $R_0 = 1.47(95\%CI: 1.41-1.53)$. With the superspreading-SEIR model, we have a similar estimate of $R_0 = 1.52(95\%CI:1.48-1.54)$. 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. Fig 3 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 [35]. 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 (Fig 4.

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 [29]. 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.
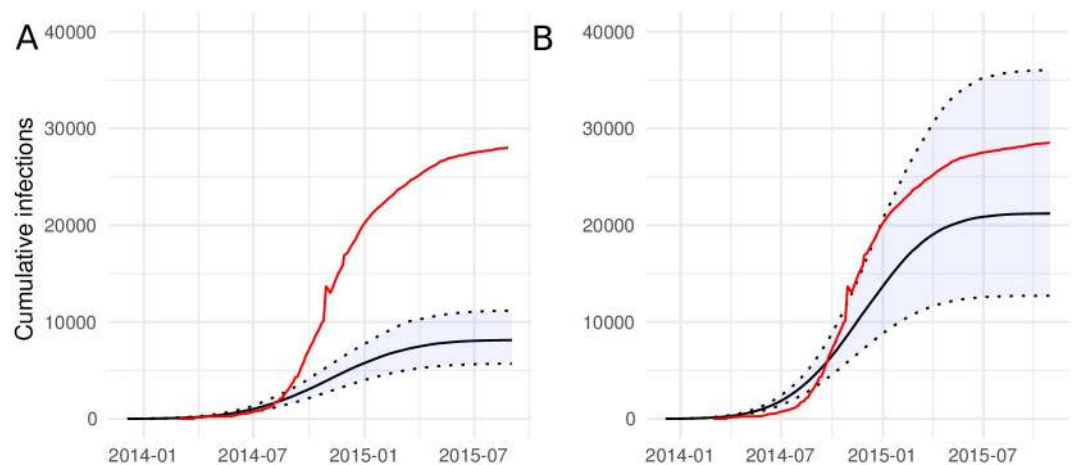
Fig 3. Model-based estimates of cumulative infections through time for the 2014-15 Ebola epidemic in Western Africa. Estimates are shown for the SEIR model (A) and the model which includes super-spreading (B). The red line show the cumulative number of cases reported by WHO [35].

https://doi.org/10.1371/journal.pcbi.1006546.g003
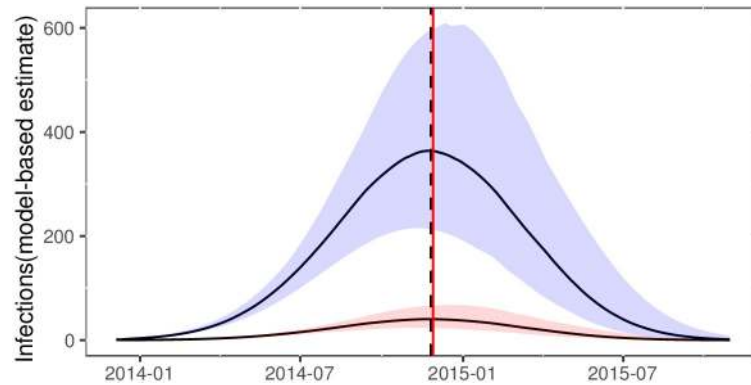
Fig 4. Estimated effective number of infections through time using the superspreading SEIR model for the 2014-15 Ebola epidemic in Western Africa. The red vertical line shows the time of peak prevalence inferred from WHO case reports. The vertical dashed line shows the model estimated time of peak prevalence. The red trajectory shows the proportion of infections in the high-transmission-rate compartment.

https://doi.org/10.1371/journal.pcbi.1006546.g004

Simulations

With simulated tree data, *PhyDyn* recovers the correct transmission risk ratios and transmission rates, although performance depends on which structured coalescent model is used. Fig 5 compares estimates across 25 simulations using *PL2* and *QL* models on epidemics with 5,000 initial susceptible individuals and a sample size of 500 sampled heterochronously shortly after epidemic peak. The transmission risk ratio parameters were varied across simulations between and the per-capita transmission rate was kept constant. S4 Fig shows performance of the *PL1* model which was similar to *PL2* but had slightly higher bias and lower posterior coverage of true parameters. Results for a smaller and noisier epidemic (2000 initial susceptibles) is shown in S5 Fig. The running time of the *QL* model was approximately five times faster than *PL2*
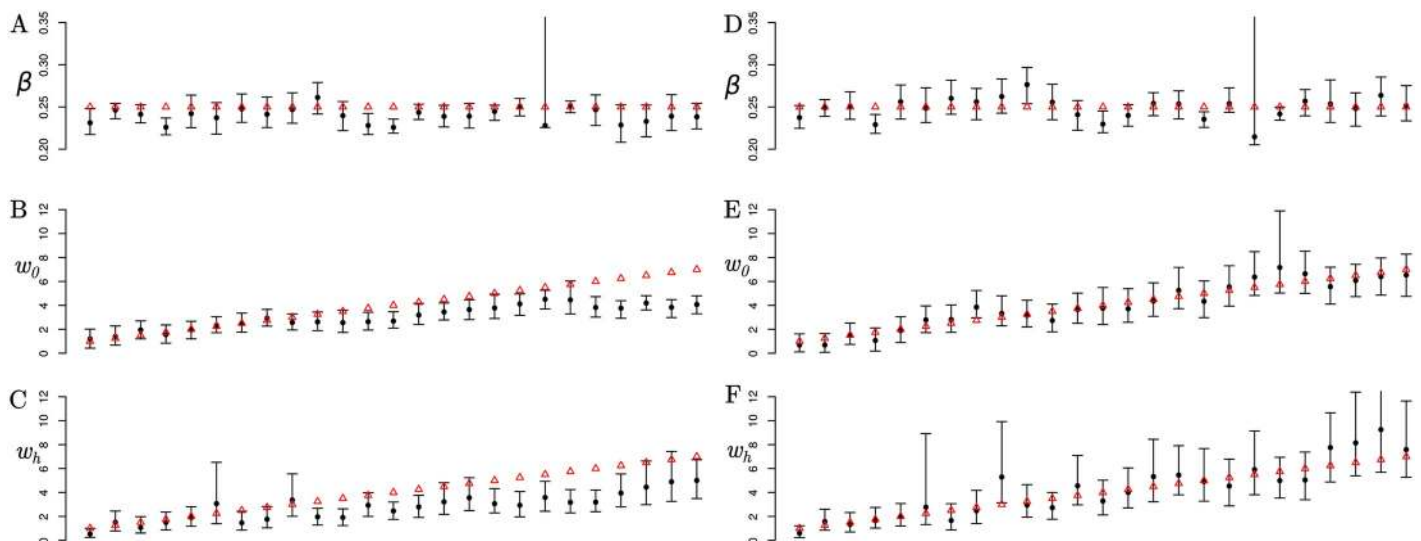
Fig 5. Parameter estimates and credible intervals for 25 simulations with variable transmission risk ratos. The red points show true parameter value. The parameter β is the per-capita transmission rate, and w_0 and w_h are respectively the transmission risk ratios in the first stage of infection and the high risk group (cf. Eq 12). A-C: Results generated using the *QL* model. D-F: Results generated using the *PL2* model. There is one outlier simulation where the transmission rate parameter could not be estimated precisely and upper bound of the CI was > 70% using both methods.

https://doi.org/10.1371/journal.pcbi.1006546.g005

which required approximately 12 hours to complete 35,000 MCMC iterations, however *QL* has considerable bias at the upper range of transmission risk ratio parameters and corresponding lower posterior coverage.

Good coverage of parameter estimates with estimated 95% credible intervals was observed with the *PL2* model. Across 75 parameter estimates (three parameters not counting initial conditions and 25 simulations), estimates did not cover the true value 4 times. Bias of the mean posterior estimate was quite small; the largest bias was 0.228 for the w_h parameter which varied across simulations between 1 and 9. In contrast, the *QL* model failed to cover much more frequently, however errors were largely confined to larger risk ratios and *QL* had a tendency to underestimate risk ratios. Greater bias was observed with the *QL* model, with the greatest bias observed for the w_h parameter (mean bias:-0.48). However the *QL* model also had good precision with smaller risk ratios as evidenced in the simulation with smaller population size (S5 Fig). In that case, the *PL2* model showed slight bias towards overestimating risk ratios which may be due to the deterministic approximation to the noisy epidemic. 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.

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 BEAST2 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 [36]. 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 [37], and approaches for utilizing continuous-valued metadata [38].

Supporting information

S1 Text. Structured coalescent likelihood and approximations.
(PDF)

S1 Fig. Diagram representing dynamics of simulation model with four demes. This model has two levels of transmission rate (l and h) and two stages of infection with higher transmission in the first stage. Solid lines represents death or stage progression. Dash lines represent transmissions.
(TIF)

S2 Fig. Comparison of stochastic and deterministic trajectories. The stochastic epidemic simulation is shown in black and the deterministic ODE model is shown in red.
(TIF)

S3 Fig. Effective population size of influenza H3N2 in New York 2014-15 estimated using GMRF skyride. The median posterior estimate is shown in the panel on the left, and the panel on the right shows both the median and 95% credible intervals.
(TIF)

S4 Fig. Parameter estimates using the *PL1* coalescent model and credible intervals for 25 simulations with variable transmission risk ratos. The red points show true parameter value. Top: Transmission rate. Middle: Acute stage transmission risk ratio. Bottom: High risk group transmission risk ratio.
(TIF)

S5 Fig. Parameter estimates and credible intervals for 20 simulations. The red line shows the true value. A-C: Results generated using the *PL1* model. D-F: Results generated using the *QL* model. The parameters are in the same order as Fig 5 in the main text.
(TIF)

Acknowledgments

Author Contributions

Conceptualization: Erik M. Volz, Igor Siveroni.

Data curation: Erik M. Volz.

Formal analysis: Erik M. Volz.

Funding acquisition: Erik M. Volz.

Methodology: Erik M. Volz.

Software: Igor Siveroni.

Supervision: Erik M. Volz.

Writing – original draft: Erik M. Volz.

Writing – review & editing: Igor Siveroni.

References

1. Volz EM, Koelle K, Bedford T. Viral phylodynamics. PLoS Comput Biol. 2013; 9(3):e1002947. https://doi.org/10.1371/journal.pcbi.1002947 PMID: 23555203

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. https://doi.org/10.1093/molbev/msi103 PMID: 15703244

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. https://doi.org/10.1073/pnas.1207965110

4. Volz EM, Kosakovsky Pond SL, Ward MJ, Leigh Brown AJ, Frost SDW. Phylodynamics of infectious disease epidemics. Genetics. 2009; 183(4):1421–1430. https://doi.org/10.1534/genetics.109.106021 PMID: 19797047

5. Volz EM. Complex population dynamics and the coalescent under neutrality. Genetics. 2012; 190 (1):187–201. https://doi.org/10.1534/genetics.111.134627 PMID: 22042576

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. https://doi.org/10.1098/rstb.2010.0060 PMID: 20478883

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. https://doi.org/10.1098/rstb.2012.0314 PMID: 23382432

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. https://doi.org/10.1093/molbev/msx124 PMID: 28402447

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. https://doi.org/10.1093/bioinformatics/btu201 PMID: 24753484

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. PMID: 10353916

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. https://doi.org/10.1098/rsif.2013.1106 PMID: 24573331

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. https://doi.org/10.1371/journal.pcbi.1000520 PMID: 19779555

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. https://doi.org/10.1371/journal.pgen.1005421 PMID: 26267488

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. https://doi.org/10.1093/molbev/mst203 PMID: 24150038

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. https://doi.org/10.1371/journal.pmed.1001568 PMID: 24339751

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. https://doi.org/10.1093/ve/vex014 PMID: 28775893

21. Volz E, Pond S. Phylodynamic analysis of ebola virus in the 2014 sierra leone epidemic. PLoS Curr. 2014; 6. https://doi.org/10.1371/currents.outbreaks.6f7025f1271821d4c815385b08f5f80e PMID: 25914858

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. https://doi.org/10.1093/molbev/msw217 PMID: 28053012

23. Poon AFY. Phylodynamic Inference with Kernel ABC and Its Application to HIV Epidemiology. Mol Biol Evol. 2015; 32(9):2483–2495. https://doi.org/10.1093/molbev/msv123 PMID: 26006189

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. https://doi.org/10.1371/journal.pcbi.1004789 PMID: 26938243

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. https://doi.org/10.1038/nature06945 PMID: 18418375

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. https://doi.org/10.1016/j.epidem.2012.06.001 PMID: 22939310

27. Volz EM, Frost SD. Sampling through time and phylodynamic inference with coalescent and birth–death models. Journal of The Royal Society Interface. 2014; 11(101):20140945. https://doi.org/10.1098/rsif.2014.0945

28. 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. https://doi.org/10.1038/nature22040 PMID: 28405027

29. Koelle K, Rasmussen DA. Rates of coalescence for common epidemiological models at equilibrium. J R Soc Interface. 2012; 9(70):997–1007. https://doi.org/10.1098/rsif.2011.0495 PMID: 21920961

30. Müller NF, Rasmussen DA, Stadler T. The Structured Coalescent and Its Approximations. Molecular biology and evolution. 2017; 34(11):2970–2981. https://doi.org/10.1093/molbev/msx186 PMID: 28666382

31. Minin VN, Bloomquist EW, Suchard MA. Smooth skyride through a rough skyline: Bayesian coalescent-based inference of population dynamics. Mol Biol Evol. 2008; 25(7):1459–1471. https://doi.org/10.1093/molbev/msn090 PMID: 18408232

32. Drummond AJ, Suchard MA, Xie D, Rambaut A. Bayesian phylogenetics with BEAUti and the BEAST 1.7. Mol Biol Evol. 2012; 29(8):1969–1973. https://doi.org/10.1093/molbev/mss075 PMID: 22367748

33. 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. https://doi.org/10.1186/1471-2334-14-480 PMID: 25186370

34. Bettencourt LMA, Ribeiro RM. Real time bayesian estimation of the epidemic potential of emerging infectious diseases. PLoS One. 2008; 3(5):e2185. https://doi.org/10.1371/journal.pone.0002185 PMID: 18478118

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

36. Rasmussen DA, Volz EM, Koelle K. Phylodynamic inference for structured epidemiological models. PLoS Comput Biol. 2014; 10(4):e1003570. https://doi.org/10.1371/journal.pcbi.1003570 PMID: 24743590

37. Palacios JA, Minin VN. Gaussian Process-Based Bayesian Nonparametric Inference of Population Size Trajectories from Gene Genealogies. Biometrics. 2013;. https://doi.org/10.1111/biom.12003 PMID: 23409705

38. 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. https://doi.org/10.1093/molbev/msq067 PMID: 20203288