VIRUS EVOLUTION

Virus Evolution, 2022, **8(1)**, 1–12

DOI: https://doi.org/10.1093/ve/veac045
Advance access publication date: 2 June 2022
Review Article

Epidemiological inference from pathogen genomes: A review of phylodynamic models and applications

Leo A. Featherstone,[1,*,†] Joshua M. Zhang,[1] Timothy G. Vaughan,[2,3,‡] and Sebastian Duchene[1]

[1]Peter Doherty Institute for Infection and Immunity, University of Melbourne, Melbourne, VIC 3000, Australia, [2]Department of Biosystems Science and Engineering, ETH Zurich, Basel 4058, Switzerland and [3]Swiss Institute of Bioinformatics, Geneva 1015, Switzerland
[†]https://orcid.org/0000-0002-8878-1758
[‡]https://orcid.org/0000-0001-6220-2239
*Corresponding author: E-mail: leo.featherstone@unimelb.edu.au

Abstract

Phylodynamics requires an interdisciplinary understanding of phylogenetics, epidemiology, and statistical inference. It has also experienced more intense application than ever before amid the SARS-CoV-2 pandemic. In light of this, we present a review of phylodynamic models beginning with foundational models and assumptions. Our target audience is public health researchers, epidemiologists, and biologists seeking a working knowledge of the links between epidemiology, evolutionary models, and resulting epidemiological inference. We discuss the assumptions linking evolutionary models of pathogen population size to epidemiological models of the infected population size. We then describe statistical inference for phylodynamic models and list how output parameters can be rearranged for epidemiological interpretation. We go on to cover more sophisticated models and finish by highlighting future directions.

Key words: epidemiological models; phylodynamics; birth-death model; coalescent model.

1. Introduction

Phylodynamics combines evolutionary biology and epidemiology to generate evidence about the spread and source of pathogens. It does this by exploiting the genomic signature left by ongoing evolution during transmission. This allows it to corroborate the answers delivered via nongenomic epidemiological modelling and sometimes offers deeper insight where case numbers collected over time and space fall short. To do this, phylodynamics requires pathogen molecular evolution to occur on the same timescale as transmission, such that accumulated genetic diversity is informative about the timing of transmission. This is known as measurable evolution (Drummond et al. 2003; Grenfell et al. 2004; Biek et al. 2015).

Phylodynamics uniquely contributes to outbreak responses through capturing transmission dynamics in time and space that are otherwise inaccessible with traditional epidemiological analysis. This has notably included applications to the spread of pathogens such as SARS-CoV-2, Ebola, Zika, and HIV (Stadler et al. 2014; Vasylyeva et al. 2019; Giovanetti et al. 2020; Seemann et al. 2020). It is now an established component of coordinated outbreak responses (Rife et al. 2017), and the quantity of genome data made available during the SARS-CoV-2 pandemic demonstrates a new standard of data availability with which to conduct phylodynamics. In light of this, we present a review of phylodynamic models targeted at prospective users of phylodynamics software such as BEAST, a major software platform for using the models discussed here (Drummond et al. 2012; Bouckaert et al. 2019). We begin by outlining some necessary epidemiology for phylodynamics. Later, we consider the input parameters, epidemiological output, and core assumptions necessary for a working knowledge of these analyses. Understanding these assumptions includes recognising the distinction between birth–death and coalescent models, their interface with epidemiological models, skyline models, and models for structured host or pathogen populations. We end by describing some future directions in the field and re-emphasising that amid all the above models and assumptions, an understanding of a pathogen's host population is forever crucial. This allows for sensible assumptions in which to ground phylodynamic analysis where epidemiological knowledge is lacking.

2. Linking epidemiology and phylogenetics

Phylodynamics requires a model of a pathogen's epidemiological dynamics to be linked with a model of how phylogenies of that pathogen evolve over epidemiological timescales. Many publications in phylodynamics already describe epidemiological models, especially the susceptible-infected-recovered (SIR) model (Volz, Koelle, and Bedford 2013; Kühnert et al. 2014; Popinga et al. 2015; Kuhnert et al. 2016). Here, we provide a summary of the SIR model because it is foundational to the phylodynamic methods we discuss.

The SIR model considers a population of size N and partitions it into susceptible, infected, and removed compartments, having sizes $S(t)$, $I(t)$, and $R(t)$, at time t, respectively. We refer to the removed compartment instead of the usual 'recovered' compartment because it is assumed that sampled cases are removed from the infectious population alongside recovered and fatal cases. This is assumed to happen by means such as isolation or treatment coinciding with sampling, with entry to the recovered compartment often termed *becoming uninfectious*. The SIR population is constrained to obey $S(t)+I(t)+R(t)=N$, meaning it is a closed population with no migration in or out. The model begins with any number of infected individuals, although usually one, and describes how case numbers proceed provided all members of the population are equally likely to encounter each other (Kermack 1927). The parameters are the rates of infection/transmission (β, per susceptible–infectious interaction) and removal (δ, per capita). The rate of removal can be further split into rates of sampling, ψ, and recovery, μ with $\delta=\psi+\mu$ (Fig. 1A). The total rate of infection at a given time is then $\beta S(t)I(t)$. This reflects how the number of susceptible individuals is the limiting factor for transmission per infected host ($\beta S(t)$), with fewer susceptibles corresponding to fewer new infections.

Phylodynamics tends to focus on inferring a core set of epidemiological parameters whether assuming an SIR or other related model. The average number of secondary infections stemming from an infected individual in an otherwise susceptible population, R_0, is most commonly inferred (Delamater et al. 2019). For example, R_0 for some early SARS-CoV-2 outbreaks has been estimated at around 2.5, meaning two cases will together on average infect five others over the duration of infection in an otherwise susceptible population (D'Arienzo and Coniglio 2020; Petersen et al. 2020). The assumption of an otherwise susceptible population means that R_0 describes the earlier, exponentially growing, phase of an outbreak when most hosts are still susceptible (Fig. 1A, B). It is critical to note that for any pathogen, R_0 represents the combined effects of the phenotypic propensity for transmission together with all relevant environmental and social determinants of transmission rate (Petersen et al. 2020).

Foundational models in phylodynamics, including the coalescent exponential and constant rate birth–death (described later), model the early phase of an outbreak where the infected population compartment is assumed to be growing exponentially (Fig 1A, 2A, B). R_0 offers a valuable indication of how rapidly an outbreak is spreading ($R_0>1$) or if case numbers are stable or declining ($R_0\leq 1$). It is assumed that $S(t)$ is large compared to the number of infections at any given time (i.e. $S(t)\approx N$) (Frost and Volz 2010; Volz 2012). This is so that the total transmission rate per infected ($\beta S(t)$) is effectively constant over time such that only one transmission parameter needs estimation. We refer to this constant as λ with $\lambda=\beta S(t)$ (Fig. 1B). In other words, infected hosts encounter a stable supply of susceptible others so the number of cases increases. This remains accurate for the early stages of an outbreak, when the susceptible population is relatively far from depletion or immunity in the population is negligible. The assumption of exponential growth is relayed as singular estimates of the basic reproductive number $R_0=\frac{\lambda}{\delta}$ and other epidemiological parameters.

Estimating transmission rate also allows for inference of other quantities such as the doubling time for the number of infections (t_d) and prevalence (the number or proportion of infected individuals). Formulae for these under different phylodynamic models are described below.

Other more sophisticated phylodynamic models, which are discussed later, allow transmission rates to vary dynamically with susceptible population size or directly as a function of time. In these cases, we instead refer to R_e, the time-varying average number of secondary infections, also known as the *effective* reproductive number to distinguish it from R_0 (the basic reproductive number). Changes in R_e reflect depletion of the susceptible population as well as nonpharmaceutical interventions and behavioural changes. However, the exponential phase of an outbreak tended to be the focus of earlier foundational phylodynamic models that present an accessible entry point to the literature. Most subsequent models are an augmentation of these foundations.

3. Foundational models and assumptions

Phylodynamic analysis requires pathogen genome sequences and their sampling times from hosts in the infected and removed compartments of a population. Each sequence is assumed to come from a different host such that epidemiological estimates relate only to between-host dynamics, although some approaches to within-host dynamics have been developed (De Maio, Wu, and Wilson 2016; Didelot et al. 2016).

To adopt these data into a phylogenetic framework, phylodynamics imposes two key assumptions. First, the hypothetical 'true' phylogeny of each individual spreading pathogen mirrors the transmission network, such that transmission events closely correspond to branching events (Fig. 1B, C). This assumption implies that phylogenetic trees inferred from a set of isolates represent subtrees of the underlying transmission network, thus reflecting transmission dynamics (Fig. 1B–D). In reality, branching times may precede transmission times (du Plessis and Stadler 2015).

From a phylodynamic perspective, the phylogenetic tree should be one in which the branch lengths correspond to units of time, known as a time-tree or chronogram. Chronograms are obtained from phylogenetic trees by multiplying branch lengths (units of *substitutions/site*) by an evolutionary clock rate (units of *substitutions/site/time*) to convert branch lengths to units of time. The clock rate, although a phylogenetic rather than phylodynamic parameter, is a key model enabling phylodynamic analysis by relating epidemiological time to evolutionary change. See Bromham et al. (2018) for a review of clock models in the context of Bayesian phylogenetics.

The second assumption is that the underlying pathogen population infecting hosts, and therefore the isolate tree, evolved according to a model linking epidemiological and phylogenetic dynamics. In the Bayesian framework that predominates in phylodynamics, these models are regarded as a part of the prior and referred to as the 'tree prior'. The tree prior provides an expression for the probability of a tree given a set of parameters governing the epidemiological process generating it. The two foundational tree priors used in phylodynamics are the coalescent and the birth–death.

4. The coalescent

The coalescent originated in the field of population genetics (Kingman 1982; Rosenberg and Nordborg 2002). It models how the ancestry of sampled populations relates to their demographic history. It can be visualised as a genealogy of a set of individuals sampled at various times, with internal nodes corresponding to the times at which they coalesce into their common ancestors (Fig. 2C). It is therefore termed a backwards-in-time process with time starting at the most recent sample and terminating at the most recent common ancestor (MRCA) (Drummond et al. 2002).

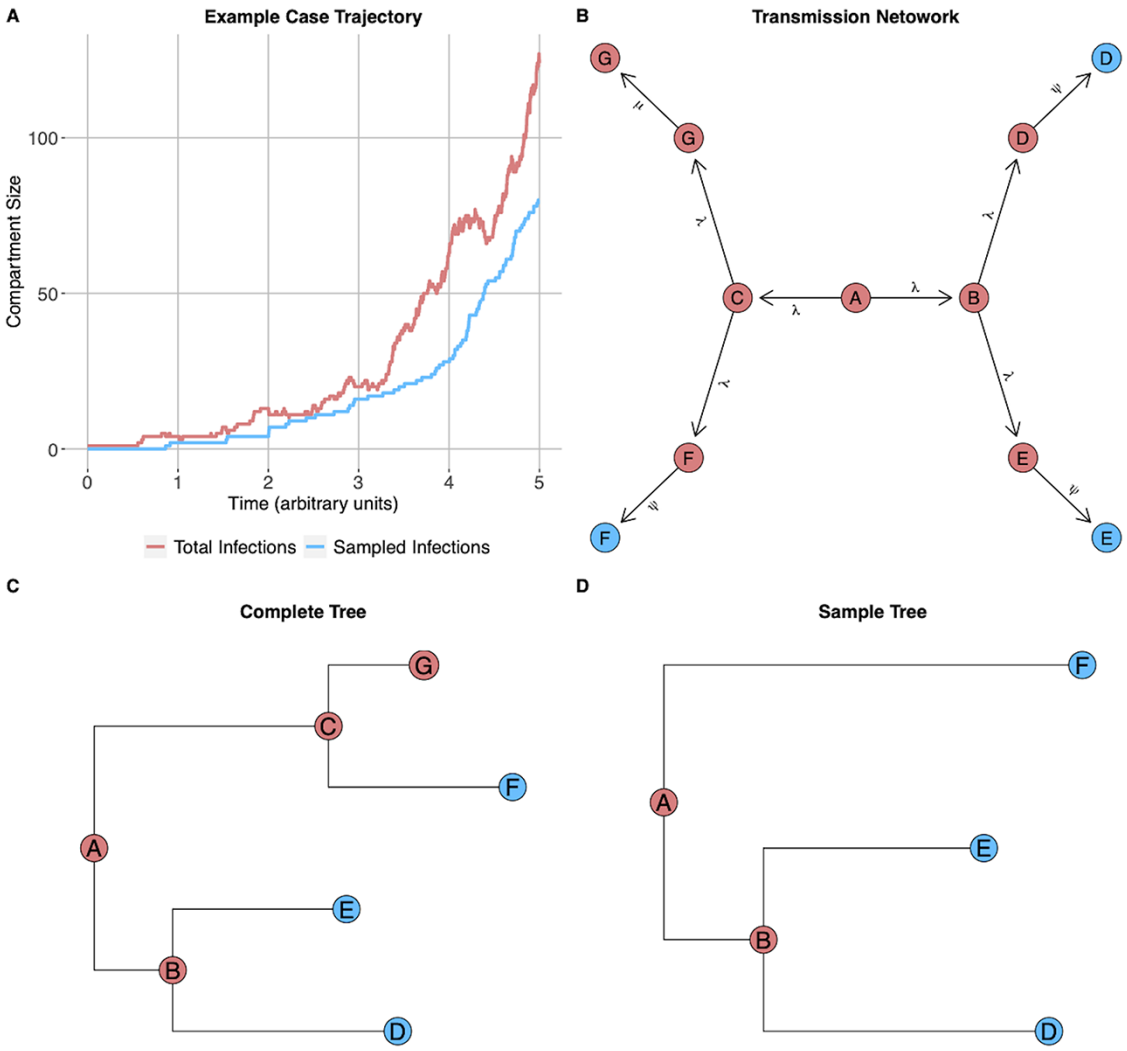

Figure 1. Representations of the key assumptions in standard phylodynamic analysis. (A) It is assumed that the underlying epidemic is growing exponentially and number of samples by extension. This corresponds to a constant rate of transmission (λ) per infected alongside constant rates of sampling infections and terminating infections (ψ and μ, respectively). (B) The transmission network is assumed to grow according to these parameters. Blue nodes correspond to sampled sequences and red nodes correspond to undiscovered or extinct infections. Node A is the first case. (C) It is assumed that the network in (B) corresponds to an underlying phylogeny of the pathogen where transmission co-occurs with branching. (D) An estimated phylogeny from sampled sequences. Phylodynamics uses the sample tree as a means of estimating the parameters driving (A) and (B).

The standalone form of the coalescent, unlinked to any epidemiological model, includes two parameters. These are the effective population size $N_e(t)$ and the generation time g. The coalescent rate for any two co-existing individuals is $\frac{1}{gN_e(t)}$, meaning the coalescence in an inferred tree offers e information about $N_e(t)$. Intuitively, this captures the expectation that the time taken for two randomly chosen individuals to coalesce increases with larger population sizes. It is also often assumed that $gf = 1$ for simplicity.

The coalescent also requires a demographic model, which is usually a deterministic model of how the effective population size changes over time (Pybus, Rambaut, and Harvey 2000). Earlier phylodynamic implementations assumed direct proportionality between $N_e(t)$ and $I(t)$, such that changes in $N_e(t)$ track changes in $I(t)$ (Frost and Volz 2010; Volz 2012). This assumption is still applied in many nonparametric coalescent skyline methods, which are discussed later.

Effective population size was constant in earlier coalescent implementations, but this was later generalised to accommodate any population size trajectory over time (Kingman 1982; Griffiths and Tavare 1994; Drummond et al. 2012). This notably includes exponential growth and logistic growth. Exponential growth is most relevant for modelling populations of rapidly spreading pathogens in susceptible populations, and this introduces an additional demographic parameter, the growth rate r from $N_e = e^{rt}$ (Fig. 1A, Fig. 2A, B). For consistency with surrounding literature, we refer to the coalescent with an exponential growth demographic model as the coalescent exponential from hereon. See Dearlove and Wilson (2013) for a review of the connection between the coalescent and common epidemiological compartmental models as well as how these relate to exponential and logistic demographic models.

Critical work by Frost and Volz (2010), linking the coalescent with epidemiological models such as the SIR, showed that $N_e(t)$ depends on transmission rates (λ) as well as prevalence ($I(t)$), meaning that the earlier assumption of direct proportionality between $N_e(t)$ and $I(t)$ is an exception rather than the rule. As a result, current implementations of the coalescent exponential, such as in BEAST, replace $N_e(t)$ with the *scaled effective population*

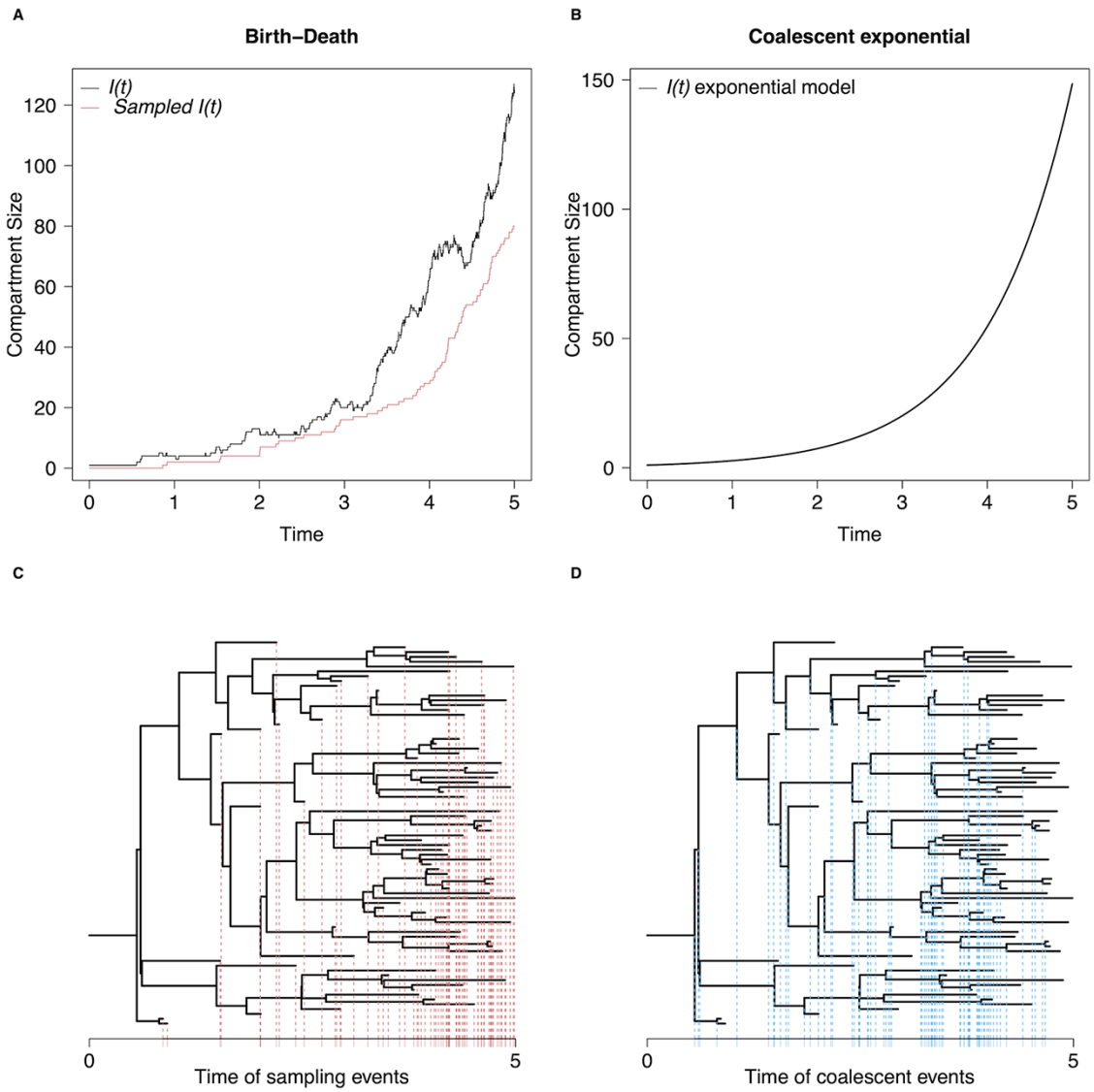

Figure 2. (A) Example stochastic trajectories in the infected compartment and sequenced subset of it. Stochastic exponential growth is accommodated by the constant rate birth–death. (B) An example of the deterministic exponential growth curve in the infected compartment as assumed by the coalescent exponential. (C) Sampling times present a source of information under the birth–death. (D) Coalescent events provide information in the coalescent while conditioning on sampling times.

size $\phi = \frac{I(0)}{2\lambda}$. This relation incorporates the effect of transmission rates and prevalence and relates these back to the coalescence rate via $\delta = \frac{1}{D}$ as for the standalone coalescent $\left(\equiv \frac{1}{gN_e(t)}\right)$. Better linking the coalescent exponential to the SIR with ϕ allows for more accurate estimates of the growth rate r, which is needed to estimate R_0 and the number of cases at the most recent sampling time $I(0)$ (Volz, Koelle, and Bedford 2013) (Table 1).

Capitalising upon the fact that any population size trajectory can be incorporated into the coalescent, further work also linked the SEIR (E for exposed) for improved epidemiological accuracy of the coalescent (Kühnert, Wu, and Drummond 2011; Volz 2012; Popinga et al. 2015). Rasmussen, Volz, and Koelle (2014b) also apply an SIR with compartments for multiple stages of infection to an HIV dataset.

5. The constant rate birth–death

The constant rate birth–death, or birth–death for short, models population growth and has been applied broadly in biology from the levels of speciation to cell division (Novozhilov, Karev, and Koonin 2006). It is a forwards-in-time process beginning with an ancestor in the past which bifurcates into new lineages, generating a tree. The original parameters underlying the birth–death are the birth rate (λ), likened to the transmission rate, and the extinction rate (μ). Phylodynamic applications of the birth–death include a rate of sampling (ψ), such that the total death rate is then the sum of the sampling rate and extinction rate of lineages ($\delta = \mu + \psi$) (Stadler et al. 2012a). This represents the rate at which individuals enter the removed compartment due to death or recovery of the host. Due to the numerous epidemiological events that can result in an individual ceasing to exert infectious pressure, the aggregate rate δ is simply known as the 'becoming uninfectious' rate and is the reciprocal of the average duration of infection (D) (Stadler et al. 2012a). The sampling probability for an infection is then defined as $p = \frac{\psi}{\psi + \mu}$. Phylodynamic birth–death models also include an origin parameter, x_0, which identifies the time of the start of the epidemic.

Table 1. Comparison of the SIR-linked birth–death and coalescent models.

Tree prior	Parameters	Notes	Software
Constant rate birth–death (Stadler et al. 2012b)	R_0, δ, p, and x_0	• $R_0 = \frac{\lambda}{\delta}$ • Is BDSky with one time interval	BEASTv2
BDSIR (Kühnert et al. 2014)	R_e, δ, p, x_0, and aS(0)	• $R^e = \frac{\lambda(t)}{\delta(t)}$ • $R_e = R_0$ if one time interval used (BEASTv2 default) • $S(0) = N$	
Coalescent exponential (Volz et al. 2009)	ϕ and r	• Infected population size at final sample time: $I(0) = e^{rT}$ [b] • $R_0 = rD$[a] $+ 1$	BEASTv1&2
Sampling coalescent SIR (Volz and Frost 2014)	λ, μ, p, and $I(0)$	• $R_0 = \frac{\lambda}{\mu}$ • $r = \lambda - \mu$ • Forwards in time: $I(0) = 1$ by assumption. Referred to as Y (0) in citation	-
Stochastic coalescent SIR (Popinga et al. 2015)	R_0, δ, x_0, and S(0)	• Forwards in time ($S(0) = N$)	BEASTv2
EpiInf (Vaughan et al. 2019)	λ, δ, ψ, μ, p, S(0), and x_0	• $S(0) = N$	

δ referred to as γ in some sources.
[a]The duration of infection in the same time units as r.
[b]T is the posterior tree height.

Given its parameterisation, the birth–death allows inference of λ, p, x_0, and δ. These are then rearranged to calculate the growth rate, basic reproductive number, and doubling time ($R_0 = \frac{\lambda}{\delta}$, $r = \lambda - \delta$, and $t_2 = \frac{ln^2}{r}$, respectively). Note that λ here is identical to the SIR transmission rate assumed to be constant during exponential growth phase, $\lambda = \beta S(0)$. Here, r is also the same growth rate as under the coalescent exponential.

A critical difference between birth–death and coalescent models is that the birth–death naturally encodes stochastic population growth over time, whereas the coalescent typically does not, instead of fitting a deterministic demographic model (Boskova, Bonhoeffer, and Stadler 2014), but see (Volz and Frost 2014; Popinga et al. 2015). Given that the coalescent assumes a low sampling proportion, the birth–death is preferential for modelling small outbreak clusters with dense sampling, where stochastic population growth has a strong impact. Conversely, the coalescent can be more appropriate for larger outbreaks, where sampling (i.e. sequencing) proportion is low, such that deterministic population growth approximates the population trajectory with reasonable accuracy (Stadler et al. 2015). The birth–death also differs by including an explicit sampling rate, whereas the coalescent conditions on sampling times (Fig. 2). The birth–death is consequently more sensitive to biases in sampling, yet also equipped to draw valuable information from sampling times, which often reflect infection prevalence.

6. Generating estimates from data

Phylodynamic models combine multiple population-dynamic parameters such as birth, death, and growth rates to express the likelihood of a phylogeny evolving under specific combinations of these parameters. As a result of their complex interrelation, algebraically solvable likelihood equations with which to infer parameters of best fit are usually unavailable, necessitating more computationally intensive techniques. Bayesian implementations relying on Markov Chain Monte Carlo (MCMC) methods are the most commonly used strategies for inference and are implemented in most popular phylodynamics packages including BEAST 1 (Suchard et al. 2018), BEAST 2 (Bouckaert et al. 2019), phylodyn (Lan et al. 2015; Karcher et al. 2017), and RevBayes (Höhna et al. 2016a). Some maximum likelihood packages are available such as TreeTime (Sagulenko, Puller, and Neher 2018) and TreePar (Stadler 2015).

Bayesian inference requires the combination of independent information, via prior distributions, and the probability of the data given a model, known as the *likelihood*. As such, it accommodates parameter-rich phylodynamic models, and quantification of uncertainty is a natural by-product of inference through the resulting posterior distribution.

Although formally part of the prior, the expression associated with each tree prior is termed the phylodynamic likelihood. Phylodynamic likelihood equations express the probability of a tree for a given set of parameters (Fig. 3). They are therefore central to connecting evolution and epidemiology. Two seminal derivations of phylodynamic likelihood equations are Griffiths and Tavare (1994) for the coalescent and Stadler (2010) for the birth–death. Each likelihood equation contains the parameters of interest under each tree prior, so employing them within an MCMC or related algorithm enables inference of posterior distributions for individual parameters. See Nascimento, Reis, and Yang (2017) for a targeted explanation of MCMC in a phylogenetic context. Also see Fig. 1E from du Plessis and Stadler (2015) for a detailed and accessible breakdown of the posterior probability expression, including the phylodynamic likelihood.

The critical working knowledge surrounding the birth–death likelihood equation is that λ, δ, and p are intertwined such that they are not individually estimable. This is known as the *non-identifiability* problem, and it occurs because λ, δ, and p only occur in the terms $\lambda\delta p$ and $\lambda - \delta$, meaning infinite combinations of each individual parameter can produce identical likelihood values. Nonidentifiability can result in uselessly broad posterior distributions if prior distributions on all nonidentifible parameters are diffuse. Countering this requires the use of additional external information to either fix or set informative prior distributions for some parameters. For example, if one wanted to estimate λ, highly informative priors or fixing of at least one of δ and p would be required for meaningful inference of λ (Boskova, Bonhoeffer, and Stadler 2014; Louca et al. 2021). The reciprocal of the duration of infection is often used to fix the becoming uninfectious rate ($\delta = \frac{1}{D}$) in this situation. For early SARS-CoV-2 variants, independent estimates suggest an infectious period of 10 days (0.0274 years), such that δ could be fixed or given a prior concentrated around $0.1\,\text{days}^{-1} \equiv 36.5\,\text{years}^{-1}$. When applying a coalescent with exponential growth, external information determining D is also used to calculate R_0 from r as in Fig. 3.

7. Further linking the SIR

Coalescent exponential and constant rate birth–death models allow for the inference of R_0, δ, λ I (*at present*), S (*at present*), and the age of the outbreak $T \equiv tMRCA$, the time since the first transmission event (Fig. 3, Table 1). We refer to S and I *at present* here as the coalescent and birth–death view time in opposite directions. I at present is a particularly important parameter as it offers

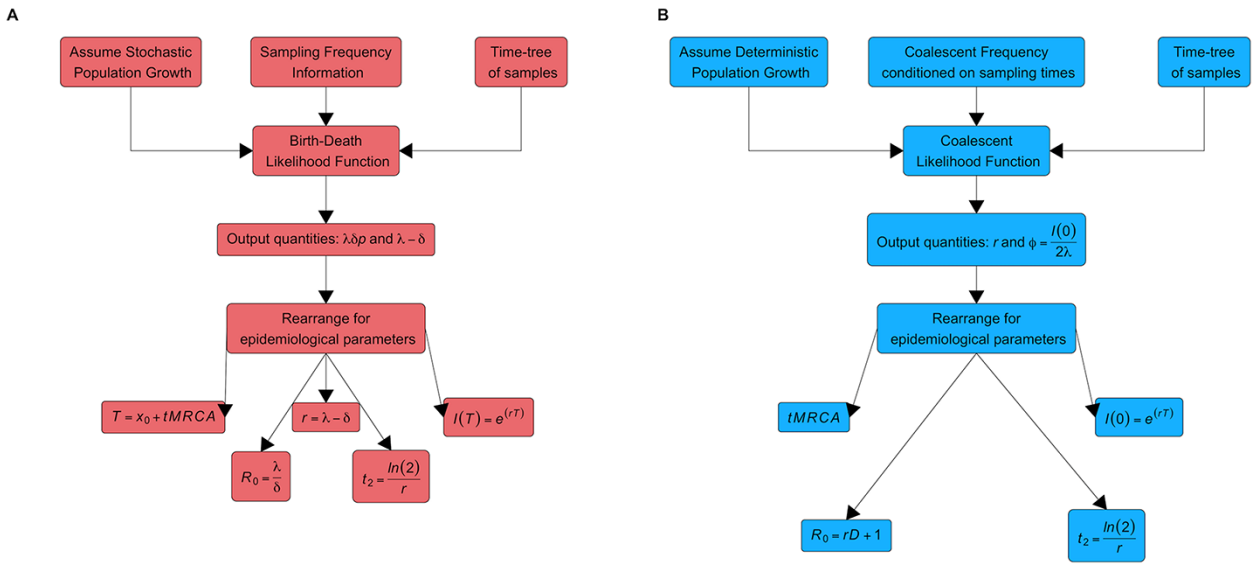

Figure 3. D refers to the duration of infection and T refers to the posterior tree height. (A) A flowchart of assumptions, methods, and inference under the constant rate birth–death. (B) As in (A) but with respect to the coalescent exponential.

insight into the number of infections at the time of the youngest sample and thereby the proportion of un-notified cases. Other tree priors have built upon these foundational models to infer the same parameters by incorporating more sampling information or extending beyond the exponential phase.

The coalescent exponential has been extended to include sampling information and stochastic population growth. Popinga et al. (2015) introduced the stochastic coalescent SIR, which builds upon the coalescent exponential to include stochastic trajectories of the infected compartment. It provides better estimates of epidemiological parameters than the deterministic coalescent exponential, but with the drawback of high computational demand. This approach involves simulating SIR epidemics alongside the inference of a sample tree such that SIR parameters maximising the likelihood of the tree and epidemic trajectory are found.

A coalescent model including a sampling rate is implemented in the 'phylodyn' R package (Karcher et al. 2017). It has been shown to improve estimates of R_0 if the sampling process is correctly specified (Volz and Frost 2014; Karcher et al. 2016).

Birth–death models have also been expanded to cater for post-exponential SIR dynamics. The birth–death SIR (BDSIR) model, introduced by (Kühnert et al. 2014), estimates trajectories in epidemiological compartments rather than singular rates during the exponential phase. The BDSIR model achieves better estimates for each parameter than constant rate birth–death by accommodating this temporal heterogeneity, but at the cost of model simplicity. Its approach is to estimate SIR epidemic trajectories with parameter values drawn from the prior over consecutive time intervals imposed on the time span of the isolate tree. Vaughan et al. (2019) recently introduced a framework to estimate epidemic trajectories using a particle filtering algorithm, which improves the accuracy and generality of this approach. It is available in the EpiInf Beast2 package. It can also incorporate an SIS model. Alternatively, Leventhal et al. (2014) introduced an approach which is able to both accurately and more efficiently infer SIR parameters, but at the expense of reconstructing the epidemic trajectory.

8. Beyond the SIR

There is a wide diversity of phylodynamic models beyond those associated with the SIR. They can include more complex compartmental models or move away from these altogether. A key distinction among them is whether they are parametric or nonparametric. Parametric models explicitly model population parameters, such as size or birth rates, as fixed functions of time. This is similar to the common distinction of parametric statistical tests assuming data follow a fixed distribution. In this framework, the coalescent exponential and birth–death are examples of parametric models (i.e. population growth is assumed to be exponential over time with rate $r = \lambda - \delta$). Nonparametric models instead offer flexibility from fixed models by allowing parameters to vary as piecewise constants over time (Palacios et al. 2014). As such, the significance of nonparametric models is the absence of a fixed expression for parameters over the entire time course of a tree, rather than the absence of parameters altogether as the name suggests. Nonparametric models are also referred to as skyline models due to piecewise constant parameters over time resembling a skyline (Ho and Shapiro 2011).

9. Parametric models

The coalescent can incorporate any model of population size over time, although constant, exponential, and logistic models are commonest (Griffiths and Tavare 1994). For example, Pybus et al. (2001) establish a logistic demographic model for hepatitis C virus (HCV) epidemics in a coalescent context. Pybus et al. (2003) also study the history of HCV in Egypt and apply a model with two constant population sizes separated by a period of exponential growth. Rasmussen, Boni, and Koelle (2014a) consider the dynamics of Dengue virus in Vietnam with parametric models including spatial and seasonal variation in effective population size. Volz and Siveroni (2018) notably introduce PhyDyn, a BEAST v2 package, that allows for any compartmental model to be fit under the coalescent. The authors demonstrate its use with Influenza and Ebola data sets.

Birth–death models can also include time-dependent rates (i.e. not constant/piecewise constant), although they tend to be applied more in macroevolutionary work than genomic epidemiology. Nee, May, and Harvey (1994) provide seminal work on the birth–death likelihood for phylogenetics, including time-dependent birth and death rates. Parag and Pybus (2018); Paradis (2011); Rabosky and Lovette (2008) also introduce key results and methods surrounding inference under the birth–death time-dependent rates. Höhna, May, and Moore (2016b) and Höhna (2013) introduce the capability to simulate under the birth–death with time-dependent rates. The above present a promising path toward replicating the diversity of coalescent parametric methods for parametric birth–death models in epidemiological application.

10. Nonparametric skyline models

Nonparametric skyline models infer parameters as piecewise constants over time to approximate any trajectory. Most infer population size over time ($N_e(t)$), but some infer other parameters. The flexibility to model any demographic history is the greatest advantage of nonparametric skyline methods. They are most appropriate for data sets spanning many generations of infection such that no single mechanistic model of population size is appropriate.

10.1 Skyline models for $N_e(t)$

Pybus, Rambaut, and Harvey (2000) developed the first nonparametric skyline model and provide a lucid introduction to skyline plots. It assumes a known tree and employs maximum likelihood estimation, whereas modern Bayesian techniques also incorporate phylogenetic uncertainty through exploring a posterior distribution of trees. It estimates independent effective population sizes (N_i) between each pair of consecutive coalescent events by exploiting the relationship between rate of coalescence and effective population size (pairwise coalescent rate is $\frac{1}{gN_e(t)}$). This relation underpins all other coalescent *e* skyline methods too.

The earliest model due to Pybus, Rambaut, and Harvey (2000) only considered ultrametric trees (trees relating samples collected at the same time). Subsequent methods were extended to include heterogeneous sampling (nonultrametric trees) and smoothed estimates of population size. These include the Bayesian skyline plot (BSP) (Drummond 2005), the Bayesian multiple change point method (Opgen-Rhein, Fahrmeir, and Strimmer 2005), and the Skyride, which employs Gaussian Markov random fields (GMRF) to estimate a temporally smoothed trajectory of effective population size (Minin, Bloomquist, and Suchard 2008). Extension to continuous change in population size avoids the jumpy output of earlier skyline methods. See Fig. 2 from Ho and Shapiro (2011) for a comparison of these models.

Extensions to heterochronous sampling and continuous population trajectories led to the Skyrgid model (Gill et al. 2013), which also employs GMRF for smoothing. It allows for inference of the effective population size at time points other than coalescent events but can risk overfitting compared to other methods due to this. Gill et al. (2016) then extended the Skygrid to incorporate covariate information to model effective population size against, such as disease prevalence. See Hill and Baele (2019) for a review of skyline methods and an example of how to run a Skygrid model in BEAST 1. See Ho and Shapiro (2011) for a review of all the above-mentioned skyline methods.

Table 2. Coalescent and birth–death-based skyline models.

Tree prior	Parameters	Notes	Software
BSP (Drummond 2005)	N_i	• Prior on N_1	BEASTv1&2
Skyride (Minin, Bloomquist, and Suchard 2008)		• GMRF smoothing prior	
Skygrid (Gill et al. 2013)			
Skygrowth (Volz and Didelot 2018)	$N_e(t)$, r	• Prior only placed on precision parameter, τ	Skygrowth
ESP (Parag, du Plessis, and Pybus 2020)	$N_e(t)$, $\beta(t)$	• β is sampling intensity • Prior on N_1 (first interval for N_e)	BEASTv2
BDSky (Stadler et al. 2012b)	$R^e = \frac{\lambda(t)}{\delta(t)}$ $p(t)$, x_0	• Possible to condition on x_0 • Accommodates simultaneous sampling redefforts with an optional ρ parameter	

10.2 Skyline models beyond $N_e(t)$

Recently, Parag, du Plessis, and Pybus (2020) introduced the epoch sampling plot (ESP), which incorporates coalescent times alongside sampling times for more accurate reconstructions of demographic histories. It infers both sampling intensity and effective population size in predetermined epochs. Parag, du Plessis, and Pybus (2020) apply the ESP to data from human influenza A virus and steppe bison, demonstrating the flexibility of nonparametric skyline methods. The introduction also offers a helpful review of previous work considering sampling information under the coalescent including (Volz and Frost 2014; Karcher et al. 2016).

Volz and Didelot (2018) introduced the skygrowth model, which infers population growth rates over time, and demonstrate its use with data from Rabies virus and *Staphylococcus aureus* epidemics. The skygrowth model is useful for corroborating constant population size in other skyline plots as these may report stable effective population size when a nonzero growth rate suggests otherwise. The model also includes a regression approach to model growth rate against time-varying variables. It is available in the skygrowth R package.

The birth–death skyline model (BDSky) due to Stadler et al. (2012b) offers a skyline model grounded in the birth–death. It infers λ, p, and δ over time, rather than population size or growth rate alone as under the above coalescent models. It also allows for differing numbers of time intervals to be set for each parameter, which can then be further rearranged to infer other epidemiological quantities over time (Table 2). Much like the other birth–death-based models, the BDSky requires strong priors on some parameters to tackle nonidentifiability and infer sharp posteriors on those of interest. The BDSky is similar to the BDSIR model in that it infers parameters as piecewise constant over time but does not simulate epidemic trajectories to achieve this. Both models accommodate sampling efforts where many samples are simultaneously taken with a given probability ρ, leading to ultrametric trees.

11. Structured models

The models discussed so far only consider one pathogen and host population. These are singular populations wherein individuals are assumed to interact randomly, as in the SIR. However, the broader reality is that host and pathogen populations are often structured, meaning they contain distinct subgroups for which transmission may occur at different rates. Structure is often the result of geographical disparities limiting movement across populations but can involve numerous other factors. For example, Stadler and Bonhoeffer (2013) consider a data set of HIV subtype A sequences from Latvia, wherein transmission dynamics differed between heterosexual and intravenous drug using subsets of the population, both comprising the entirety of the dataset. Structure can also arise from differentiation of the pathogen. Additionally, some pathogens exhibit a disease progression requiring additional compartments. For instance, the SEIR model introduces an (E)xposed compartment composed of infected but not yet infectious hosts. In response to the need to consider subpopulations with distinct epidemiological dynamics, structured models have been developed to infer epidemiological dynamics within and between *types* or *demes* in a population. Here, *types* and *demes* interchangeably refer to any differentiating categorical factor such as location, host demographic, or pathogen subtypes.

Structured models can sometimes be challenging to configure in application. This is usually due to the computational demands of a larger set of parameters. See Douglas et al. (2021) for a recent application of some structured models to SARS-CoV-2 transmission.

Lemey et al. (2009) introduced a fundamental approach to modelling migration rates between demes by treating geographical locations as discrete traits between which individuals could shift in a way analogous to nucleotide substitution models. This is often referred to as discrete trait analysis (DTA). DTA allows inference of root state, migration rates, and most probable ancestral demes of a collection of sequences. Lemey et al. (2014) augmented DTA to include explanatory variables such as air-traffic flows to predict migration rates using generalised linear models. Lemey et al. (2020) also recently built upon Lemey et al. (2009) to include the travel history information of samples and applied this to the international spread of SARS-CoV-2.

Modelling locations in the same way as substitution offers valuable simplicity and scalability but also carries unrealistic assumptions. These chiefly include deme-membership not affecting transmission or sampling rates. Sampling biases between demes also impact inference of ancestral states and migration rates (De Maio et al. 2015). In light of this, several methods modelling population structure under structured coalescent and multitype birth–death (MTBD) tree priors have been developed.

The structured coalescent includes rates at which individuals move between demes (Hey 2010). Unlike the coalescent with a single population, inferences focus more on rates of transfer between demes than population size within each. Much of the work on the structured coalescent grappled with its high computational demands by approximating its likelihood. For example, Volz (2012) introduced an approximation to the likelihood of the structured coalescent. An exact method is used in MultiTypeTree due to Vaughan et al. (2014), a BEAST2 package that fits a structured coalescent to sequence data. Vaughan et al. (2014) demonstrated use with a global H3N2 influenza data set.

Noting that the structured coalescent only remained computationally tractable for a few demes and sensitivity to sampling bias within DTA, De Maio et al. (2015) developed an approximation to the coalescent likelihood that allowed for inference with a larger number of demes. It also showed increased accuracy compared to DTA and was released as the BASTA package in BEAST2. De Maio et al. (2015) applied BASTA to data from Ebola and Avian Influenza outbreaks as well as an agricultural virus. BASTA was later employed in SCOTTI, a method that infers transmission dynamics by considering each host as an individual population, thus likening migration rates to transmission rates (De Maio, Wu, and Wilson 2016).

Möller, Rasmussen, and Stadler (2017) developed an exact solution to the structured coalescent likelihood on which they based an improved approximation similar to Volz (2012). It is available in the MASCOT package in BEAST2, which fits structured coalescent models (Möller, Rasmussen, and Stadler 2018). Möller, Dudas, and Stadler (2019) then extended MASCOT to model migration rates against predictor data analogously to Lemey et al. (2014) with generalised linear models.

The MTBD model also accommodates population structure. It allows for estimation of transmission rates, death rates, and sampling proportions across subpopulations. For example, Kühnert et al. (2018) estimate separate transmission rates for drug-sensitive and resistant HIV strains using the MTBD. The mathematical foundation for deriving the likelihood of the MTBD is described in Maddison, Midford, and Otto (2007). Stadler and Bonhoeffer (2013) extended this to heterochronous trees and used maximum likelihood inference. Their method allows estimation of the likelihood of a tree where the 'types' of each sample are known beforehand but can also assess hypotheses about the underlying number of types if tip states are unknown.

Kuhnert et al. (2016) built upon this framework to produce an MTBD model combined with the birth–death skyline that infers parameters across time intervals as in the BDSky. It is available in the BEAST2 bdmm package and allows for both serial and contemporaneous sampling but requires types to be specified a priori, such as knowing the location of each sample. Barido-Sottani, Vaughan, and Stadler (2020) recently developed a model that allows for MTBD inference, including the number of types. It requires a prior on the number of subgroups and can infer transmission rates, death rates, and a total rate of individuals moving between subpopulations alongside a fixed value for the sampling proportion. Scire et al. (2020) proposed a new algorithm that greatly improves the computational efficiency of the MTBD, raising the computational limit in its sample size from roughly 250 to at least 500.

12. Future directions and final remarks

Future advances in phylodynamics will likely include improved metadata integration, model selection, scalability and performance, accommodation of continuous spatial structure, and adaptive evolution. Many of these areas are continuations of outstanding questions outlined by Frost et al. (2015).

The move to integrate geographic metadata features prominently among advances in data integration (Guindon and De Maio 2021; Hill et al. 2021). Geographic metadata allow for transmission dynamics to be resolved in space. For example, this can allow for the locating of pathogen reservoirs where more traditional case counts may fall short. Although structured models capture geographic structure in a discrete sense, methods modelling spread in continuous space offer finer resolution and are emerging. These use the locations and dates of isolates along with their genome sequences to infer a point source in space that

most likely led to the observed geographic distribution (Lemey et al. 2010; Kalkauskas et al. 2020). This allows for hypothesis testing about geographic spread of a pathogen (Dellicour et al. 2019).

So-called occurrence metadata are also being integrated into phylodynamic analyses. Occurrences are known instances of infection, for example, due to a positive test, without an accompanying sampled genome. These are included alongside sequenced samples so as to offer information about prevalence. These data can improve phylodynamic inference, and the theoretical bases for their inclusion in phylodynamics are growing (Vaughan et al. 2019; Andr´eoletti et al. 2020; Gupta et al. 2020; Featherstone et al. 2021; Zarebski et al. 2022).

Accommodating adaptive evolution offers an opportunity for advance as reliance upon pathogen genomic surveillance increases. Rasmussen and Stadler (2019) recently developed a model based on the MTBD where the evolutionary fitness of lineages is allowed to vary depending on multiple pathogen traits, including the fitness effects of mutations at multiple sites in the genome. This is distinct from many other phylodynamic models which assume selective neutrality between subtypes. Capturing adaptive evolution will likely continue to attract interest, especially following interest in the emergence of variants of concern (Tay et al. 2022).

In the vein of increased computing efficiency, online Bayesian methods which allow for MCMC chains that can be interrupted to include more data are emerging (Gill et al. 2020). This approach is similar to sequential Markov inference (Fourment et al. 2018) and presents a major step towards effectively real-time phylodynamic pathogen surveillance.

Although phylodynamics is expanding in the type and quantity of data it considers, careful consideration must always be given to the assumptions underlying tree prior selection. For instance, the choice of tree prior can affect estimates of the rate of substitution and time of origin (Möoller, du Plessis, and Stadler 2018). Model adequacy programmes such as TreeModelAdequacy in BEAST2 are available to decide whether the model is reasonable (Duchene et al. 2019). However, more work is required to better understand the effects of model misspecification. There is a particular need for the development of methods to assess the suitability of the demographic assumptions that each tree model makes. To this end, Parag, Pybus, and Wu (2022) developed a statistic that measures the relative contribution of genetic information and demographic assumptions in driving skyline trajectories.

13. Extension to bacteria and other pathogens

The ecological and evolutionary complexity of bacteria and other more slowly evolving pathogens offers another frontier for advance in phylodynamics. Much of phylodynamics is focused on viruses, but some bacteria such as *Mycobacterium tuberculosis* also evolve measurably so as to enable phylodynamic inference (Kuhnert et al. 2018). Given that many species of bacteria can survive outside of human hosts, a branching event may represent an environmental acquisition of infection rather between host transmission. This then requires caution regarding the assumption that the isolate tree is a subtree of the transmission tree (Ingle, Howden, and Duchene 2021). Accounting for recombination also presents a consistent challenge in phylodynamics, although some methods to accommodate it have been developed (Didelot et al. 2010; Vaughan et al. 2017; Didelot and Parkhill 2021).

14. Final remarks

The SARS-CoV-2 pandemic has presented a new standard where phylodynamic analyses are conducted closer than ever to real time (Hill et al. 2021). This is in part facilitated by shared sequence repositories such as GISAID, Nextstrain, and forums such as Virological.org (Shu and McCauley 2017; Hadfield et al. 2018). This opportunity has provided valuable insight into practical considerations for real-time phylodynamics. These include the aforementioned need to assess the suitability of tree priors and aspects of sampling. The volume of available data has led to a situation where analysing ever larger data sets is not always more informative relative to tractability, so protocols are required to address when and how to subsample from a large database. Methods are emerging to meet this need, but many questions remain (Duchene et al. 2021; du Plessis et al. 2021). For example, how should one filter sequences originating from household transmission to avoid overinflating transmission rate estimates? Should phylodynamics instead focus on inferring a dispersal distribution of secondary infections to account for the probability of super-spreading events, as has been demonstrated before by Li, Grassly, and Fraser (2017)? Questions regarding sampling are also outstanding with regard to structured models. How should one subsample sequences from individual countries to infer travel-associated transmission rates? Such questions will continue to arise as phylodynamics continues to experience broader application. However, one certainty is that there will always be a need for domain knowledge on the part of local public health researchers to curate suitable datasets with these questions in mind.

In sum, phylodynamics is a burgeoning space. It has and will continue to revolutionise the way we study epidemiology and inform public health responses to future infectious disease spread. It does this by explicitly modelling the timing and or location of transmission events using evolutionary information, which is otherwise inaccessible via traditional epidemiological analyses. It is even poised to enter into the realm of cell biology in measuring how tissues develop (Chodrow et al. 2021; Stadler, Pybus, and Stumpf 2021). In essence, phylodynamics affords us the chance to make more sense of infectious disease epidemiology in the light of pathogen evolution, and as such presents a cutting edge of research to continue into the future.

Supplementary data

Supplementary data is available at *Virus Evolution* online.

Acknowledgements

The authors kindly acknowledge Tanja Stadler's and Ben Phillip's input during the drafting of this paper. They are also grateful for three anonymous reviewers whose comments greatly improved the manuscript.

Funding

L.A.F., J.M.Z., and S.D. were supported by the Australian Research Council (DE190100805) and the Australian Medical Research Council (APP1157586).

Conflict of interest: None declared.

References

Andr´eoletti, J. et al. A Skyline Birth-Death Process for Inferring the Population Size from a Reconstructed Tree with Occurrences. *bioRxiv*. 2020.10.27.356758. 2020.

Barido-Sottani, J., Vaughan, T. G., and Stadler, T. (2020) 'A Multitype Birth–Death Model for Bayesian Inference of Lineage-Specific Birth and Death Rates', *Systematic Biology*, 69: 973–86.

Biek, R. et al. (2015) 'Measurably Evolving Pathogens in the Genomic Era', *Trends in Ecology & Evolution*, 30: 306–13.

Boskova, V., Bonhoeffer, S., and Stadler, T. (2014) 'Inference of Epidemiological Dynamics based on Simulated Phylogenies using Birth-Death and Coalescent Models', *PLOS Computational Biology*, 10: e1003913.

Bouckaert, R. et al. (2019) 'Beast 2.5: An Advanced Software Platform for Bayesian Evolutionary Analysis', *PLOS Computational Biology*, 15: e1006650.

Bromham, L. et al. (2018) 'Bayesian Molecular Dating: Opening up the Black Box', *Biological Reviews*, 93: 1165–91.

Chodrow, P. et al. How Our Cells Become Our Selves: The Cellular Phylodynamic Biology of Growth and Development. Technical report, bioRxiv, Sept. 2021. <https://www.iorxiv.org/content/10.1101/2021.09.22.461268v1> accessed 27 Mar 2022. Section: New Results Type: article.

D'Arienzo, M., and Coniglio, A. (2020) 'Assessment of the SARS-CoV-2 Basic Reproduction Number, R0, Based on the Early Phase of COVID-19 Outbreak in Italy', *Biosafety and Health*, 2: 57–9.

De Maio, N. et al. (2015) 'New Routes to Phylogeography: A Bayesian Structured Coalescent Approximation', *PLOS Genetics*, 11: e1005421.

De Maio, N., Wu, C.-H., and Wilson, D. J. (2016) 'Scotti: Efficient Reconstruction of Transmission within Outbreaks with the Structured Coalescent', *PLOS Computational Biology*, 12: e1005130.

Dearlove, B., and Wilson, D. J. (2013) 'Coalescent Inference for Infectious Disease: Meta-Analysis of Hepatitis C', *Philosophical Transactions of the Royal Society B: Biological Sciences*, 368: 20120314.

Delamater, P. L. et al. (2019) 'Complexity of the Basic Reproduction Number (R0)', *Emerging Infectious Diseases*, 25: 1–4.

Dellicour, S. et al. Phylogeographic and Phylodynamic Approaches to Epidemiological Hypothesis Testing. Technical report. 10 2019.

Didelot, X. et al. (2010) 'Inference of Homologous Recombination in Bacteria using Whole-Genome Sequences', *Genetics*, 186: 1435–49.

Didelot, X., and Parkhill, J. (2021) A Scalable Analytical Approach from Bacterial Genomes to Epidemiology Technical Report, bioRxiv, Nov. <https://www.biorxiv.org/content/10.1101/2021> accessed 27 Mar 2022.

Didelot, X. et al. (2016) 'Within-Host Evolution of Bacterial Pathogens', *Nature Reviews. Microbiology*, 14: 150–62.

Douglas, J. et al. (2021) 'Phylodynamics Reveals the Role of Human Travel and Contact Tracing in Controlling Covid-19 in Four Island Nations', *Virus Evolution*, veab052.

Drummond, A. J. (2005) 'Bayesian Coalescent Inference of Past Population Dynamics from Molecular Sequences', *Molecular Biology and Evolution*, 22: 1185–92.

——. et al. (2002) 'Estimating Mutation Parameters, Population History and Genealogy Simultaneously from Temporally Spaced Sequence Data', *Genetics*, 161: 1307–20.

——. et al. (2003) 'Measurably Evolving Populations', *Trends in Ecology & Evolution*, 18: 481–8.

——. et al. (2012) 'Bayesian Phylogenetics with Beauti and the Beast 1.7', *Molecular Biology and Evolution*, 29: 1969–73.

du Plessis, L. et al. (2021) 'Establishment and Lineage Dynamics of the Sars-CoV-2 Epidemic in the UK', *Science*, 371: 708–12.

du Plessis, L., and Stadler, T. (2015) 'Getting to the Root of Epidemic Spread with Phylodynamic Analysis of Genomic Data', *Trends in Microbiology*, 23: 383–6.

Duchene, S. et al. (2019) 'Phylodynamic Model Adequacy using Posterior Predictive Simulations', *Systematic Biology*, 68: 358–64.

——. et al. (2021) 'The Impact of Public Health Interventions in the Nordic Countries during the First Year of SARS-CoV-2 Transmission and Evolution', *Eurosurveillance*, 26: 2001996.

Featherstone, L. A. et al. (2021) 'Infectious Disease Phylodynamics with Occurrence Data', *Methods in Ecology and Evolution*, 12: 1498–507.

Fourment, M. et al. (2018) 'Effective Online Bayesian Phylogenetics via Sequential Monte Carlo with Guided Proposals', *Systematic Biology*, 67: 490–502.

Frost, S. D. W. et al. (2015) 'Eight Challenges in Phylodynamic Inference', *Epidemics*, 10: 88–92.

Frost, S. D. W., and Volz, E. M. (2010) 'Viral Phylodynamics and the Search for an 'Effective Number of Infections'', *Philosophical Transactions of the Royal Society B: Biological Sciences*, 365: 1879–90.

Gill, M. S. et al. (2016) 'Understanding Past Population Dynamics: Bayesian Coalescent-based Modeling with Covariates', *Systematic Biology*, 65: 1041–56.

—— et al. (2013) 'Improving Bayesian Population Dynamics Inference: A Coalescent-based Model for Multiple Loci', *Molecular Biology and Evolution*, 30: 713–24.

—— et al. (2020) 'Online Bayesian Phylodynamic Inference in Beast with Application to Epidemic Reconstruction', *Molecular Biology and Evolution*, 37: 1832–42.

Giovanetti, M. et al. (2020) 'Genomic and Epidemiological Surveillance of Zika Virus in the Amazon Region', *Cell Reports*, 30: 2275–2283.e7.

Grenfell, B. T. et al. (2004) 'Unifying the Epidemiological and Evolutionary Dynamics of Pathogens', *Science*, 303: 327–32.

Griffiths, R. C., and Tavare, S. (1994) 'Sampling Theory for Neutral Alleles in a Varying Environment', *Philosophical Transactions: Biological Sciences*, 344: 403–10.

Guindon, S., and De Maio, N. (2021) 'Accounting for Spatial Sampling Patterns in Bayesian Phylogeography', *Proceedings of the National Academy of Sciences*, 118.

Gupta, A. et al. (2020) 'The Probability Distribution of the Reconstructed Phylogenetic Tree with Occurrence Data', *Journal of Theoretical Biology*, 488: 110115.

Höhna, S. (2013) 'Fast Simulation of Reconstructed Phylogenies under Global Time-Dependent Birth–Death Processes', *Bioinformatics*, 29: 1367–74.

——. et al. (2016a) 'Revbayes: Bayesian Phylogenetic Inference using Graphical Models and an Interactive Model-specification Language', *Systematic Biology*, 65: 726–36.

Höhna, S., May, M. R., and Moore, B. R. (2016b) 'TESS: An R Package for Efficiently Simulating Phylogenetic Trees and Performing Bayesian Inference of Lineage Diversification Rates', *Bioinformatics*, 32: 789–91.

Hadfield, J. et al. (2018) 'Nextstrain: Real-Time Tracking of Pathogen Evolution', *Bioinformatics*, 34: 4121–3.

Hey, J. (2010) 'Isolation with Migration Models for More than Two Populations', *Molecular Biology and Evolution*, 27: 905–20.

Hill, V., and Baele, G. (2019) 'Bayesian Estimation of Past Population Dynamics in BEAST 1.10 using the Skygrid Coalescent Model', *Molecular Biology and Evolution*, 36: 2620–8.

Hill, V. et al. (2021) 'Progress and Challenges in Virus Genomic Epidemiology', *Trends in Parasitology*, 37: 1038–49.

Ho, S. Y. W., and Shapiro, B. (2011) 'Skyline-Plot Methods for Estimating Demographic History from Nucleotide Sequences', *Molecular Ecology Resources*, 11: 423–34.

Ingle, D. J., Howden, B. P., and Duchene, S. (2021) 'Development of Phylodynamic Methods for Bacterial Pathogens', *Trends in Microbiology*, 29: 788–97.

Kalkauskas, A. et al. (2020) Sampling Bias and Model Choice in Continuous Phylogeography: Getting Lost on a Random Walk. Technical report.

Karcher, M. D. et al. (2016) 'Quantifying and Mitigating the Effect of Preferential Sampling on Phylodynamic Inference', *PLoS Computational Biology*, 12: e1004789.

———. et al. (2017) 'Phylodyn: An R Package for Phylodynamic Simulation and Inference', *Molecular Ecology Resources*, 17: 96–100.

Kermack, W. O., and McKendrick, A. G. (1927) 'Contributions to the mathematical theory of epidemics', *Bulletin of Mathematical Biology*, 53: 33–55.

Kingman, J. F. C. (1982) 'The Coalescent', *Stochastic Processes and Their Applications*, 13: 235–48.

Kühnert, D. et al. the Swiss HIV Cohort Study. (2018) 'Quantifying the Fitness Cost of HIV-1 Drug Resistance Mutations through Phylodynamics', *PLOS Pathogens*, 14: e1006895.

——. et al. (2014) 'Simultaneous Reconstruction of Evolutionary History and Epidemiological Dynamics from Viral Sequences with the Birth–Death Sir Model', *Journal of the Royal Society Interface*, 11: 20131106.

Kühnert, D., Wu, C.-H., and Drummond, A. J. (2011) 'Phylogenetic and Epidemic Modeling of Rapidly Evolving Infectious Diseases', *Infection, Genetics and Evolution*, 11: 1825–41.

Kuhnert, D. et al. (2018) 'Tuberculosis Outbreak Investigation using Phylodynamic Analysis', *Epidemics*, 25: 47–53.

——. et al. (2016) 'Phylodynamics with Migration: A Computational Framework to Quantify Population Structure from Genomic Data', *Molecular Biology and Evolution*, 33: 2102–16.

Lan, S. et al. (2015) 'An Efficient Bayesian Inference Framework for Coalescent-based Nonparametric Phylodynamics', *Bioinformatics*, 31: 3282–9.

Lemey, P. et al. (2020) 'Accommodating Individual Travel History and Unsampled Diversity in Bayesian Phylogeographic Inference of Sars-cov-2', *Nature Communications*, 11: 1–14.

—— et al. (2014) 'Unifying Viral Genetics and Human Transportation Data to Predict the Global Transmission Dynamics of Human Influenza H3n2', *PLoS Pathogens*, 10: e1003932.

—— et al. (2009) 'Bayesian Phylogeography Finds Its Roots', *PLoS Computational Biology*, 5: e1000520.

—— et al. (2010) 'Phylogeography Takes a Relaxed Random Walk in Continuous Space and Time', *Molecular Biology and Evolution*, 27: 1877–85.

Leventhal, G. E. et al. (2014) 'Using an Epidemiological Model for Phylogenetic Inference Reveals Density Dependence in HIV Transmission', *Molecular Biology and Evolution*, 31: 6–17.

Li, L. M., Grassly, N. C., and Fraser, C. (2017) 'Quantifying Transmission Heterogeneity using Both Pathogen Phylogenies and Incidence Time Series', *Molecular Biology and Evolution*, 34: 2982–95.

Louca, S. et al. (2021) 'Fundamental Identifiability Limits in Molecular Epidemiology', *Molecular Biology and Evolution*, 38: 4010–24.

Maddison, W. P., Midford, P. E., and Otto, S. P. (2007) 'Estimating a Binary Character's Effect on Speciation and Extinction', *Systematic Biology*, 56: 701–10.

Minin, V. N., Bloomquist, E. W., and Suchard, M. A. (2008) 'Smooth Skyride through a Rough Skyline: Bayesian Coalescent-based Inference of Population Dynamics', *Molecular Biology and Evolution*, 25: 1459–71.

Möller, N. F., Dudas, G., and Stadler, T. (2019) 'Inferring Time-dependent Migration and Coalescence Patterns from Genetic Sequence and Predictor Data in Structured Populations', *Virus Evolution*, 5: vez030.

Möller, S., du Plessis, L., and Stadler, T. (2018) 'Impact of the tree prior on estimating clock rates during epidemic outbreaks', *Proceedings of the National Academy of Sciences*, 115: 4200–5.

Möller, N. F., Rasmussen, D. A., and Stadler, T. (2017) 'The Structured Coalescent and Its Approximations', *Molecular Biology and Evolution*, 34: 2970–81.

Nascimento, F. F., Reis, D. M., and Yang, Z. (2017) 'A Biologist's Guide to Bayesian Phylogenetic Analysis', *Nature Ecology Evolution*, 1: 1446–54.

Nee, S., May, R. M., and Harvey, P. H. (1994) 'The Reconstructed Evolutionary Process', *Philosophical Transactions of the Royal Society of London. Series B, Biological Sciences*, 344: 305–11.

Novozhilov, A. S., Karev, G. P., and Koonin, E. V. (2006) 'Biological Applications of the Theory of Birth-and-Death Processes', *Briefings in Bioinformatics*, 7: 70–85.

Opgen-Rhein, R., Fahrmeir, L., and Strimmer, K. (2005) 'Inference of Demographic History from Genealogical Trees using Reversible Jump Markov Chain Monte Carlo', *BMC Evolutionary Biology*, 5: 1–13.

Palacios, J. A. et al. (2014) 'Bayesian nonparametric phylodynamics', *Bayesian Phylogenetics: Methods, Algorithms, and Applications*, pp. 229–46.

Paradis, E. (2011) 'Time-Dependent Speciation and Extinction from Phylogenies: A Least Squares Approach', *Evolution*, 65: 661–72.

Parag, K. V., du Plessis, L., and Pybus, O. G. (2020) 'Jointly Inferring the Dynamics of Population Size and Sampling Intensity from Molecular Sequences', *Molecular Biology and Evolution*, 37: 2414–29.

Parag, K. V., and Pybus, O. G. (2018) 'Exact Bayesian Inference for Phylogenetic Birth-Death Models', *Bioinformatics*, 34: 3638–45.

Parag, K. V., Pybus, O. G., and Wu, C.-H. (2022) 'Are Skyline Plot-based Demographic Estimates Overly Dependent on Smoothing Prior Assumptions?' *Systematic Biology*, 71: 121–38.

Petersen, E. et al. (2020) 'Comparing Sars-CoV-2 with Sars-CoV and Influenza Pandemics', *The Lancet Infectious Diseases*, 20: e238–e244.

Popinga, A. et al. (2015) 'Inferring Epidemiological Dynamics with Bayesian Coalescent Inference: The Merits of Deterministic and Stochastic Models', *Genetics*, 199: 595–607.

Pybus, O. G. et al. (2001) 'The Epidemic Behavior of the Hepatitis C Virus', *Science*, 292: 2323–5.

—— et al. (2003) 'The Epidemiology and Iatrogenic Transmission of Hepatitis C Virus in Egypt: A Bayesian Coalescent Approach', *Molecular Biology and Evolution*, 20: 381–7.

Pybus, O. G., Rambaut, A., and Harvey, P. H. (2000) 'An Integrated Framework for the Inference of Viral Population History from Reconstructed Genealogies', *Genetics*, 155: 1429–37.

Rabosky, D. L., and Lovette, I. J. (2008) 'Explosive Evolutionary Radiations: Decreasing Speciation or Increasing Extinction through Time?' *Evolution*, 62: 1866–75.

Rasmussen, D. A., Boni, M. F., and Koelle, K. (2014a) 'Reconciling Phylodynamics with Epidemiology: The Case of Dengue Virus in Southern Vietnam', *Molecular Biology and Evolution*, 31: 258–71.

Rasmussen, D. A., and Stadler, T. (2019) 'Coupling Adaptive Molecular Evolution to Phylodynamics Using Fitness-Dependent Birth-Death Models', *eLife*, 8: e45562.

Rasmussen, D. A., Volz, E. M., and Koelle, K. (2014b) 'Phylodynamic Inference for Structured Epidemiological Models', *PLOS Computational Biology*, 10: e1003570.

Rife, B. D. et al. (2017) 'Phylodynamic Applications in 21st Century Global Infectious Disease Research', *Global Health Research and Policy*, 2: 13.

Rosenberg, N. A., and Nordborg, M. (2002) 'Genealogical Trees, Coalescent Theory and the Analysis of Genetic Polymorphisms', *Nature Reviews. Genetics*, 3: 380–90.

Sagulenko, P., Puller, V., and Neher, R. A. (2018) 'Treetime: Maximum-likelihood Phylodynamic Analysis', *Virus Evolution*, 4: vex042.

Scire, J. et al. (2020) Improved Multi-Type Birth-Death Phylodynamic Inference in Beast 2. Technical report.

Seemann, T. et al. (2020) 'Tracking the COVID-19 Pandemic in Australia Using Genomics', *Nature Communications*, 11: 4376.

Shu, Y., and McCauley, J. (2017) 'GISAID: Global Initiative on Sharing All Influenza Data – from Vision to Reality', *Eurosurveillance*, 22: 30494.

Stadler, T. (2010) 'Sampling-through-Time in Birth–Death Trees', *Journal of Theoretical Biology*, 267: 396–404.

——. *TreePar: Estimating Birth and Death Rates Based on Phylogenies*. (2015), <https://CRAN.R-project.org/package=TreePar> accessed 21 Apr 2020.

Stadler, T., and Bonhoeffer, S. (2013) 'Uncovering Epidemiological Dynamics in Heterogeneous Host Populations using Phylogenetic Methods', *Philosophical Transactions of the Royal Society B: Biological Sciences*, 368: 20120198.

Stadler, T. et al. the Swiss HIV Cohort Study. (2012a) 'Estimating the Basic Reproductive Number from Viral Sequence Data', *Molecular Biology and Evolution*, 29: 347–57.

—— et al. (2014) 'Insights into the Early Epidemic Spread of Ebola in Sierra Leone Provided by Viral Sequence Data', *PLoS Currents*, 6: ecurrents.outbreaks.02bc6d927ecee7bbd33532ec8ba6a25f Oct.

—— et al. (2012b) 'Birth-Death Skyline Plot Reveals Temporal Changes of Epidemic Spread in HIV and Hepatitis C Virus (Hcv)', *Proceedings of the National Academy of Sciences*, 110: 228–33.

Stadler, T., Pybus, O. G., and Stumpf, M. P. H. (2021) 'Phylodynamics for Cell Biologists', *Science*, 371: eaah6266.

Stadler, T. et al. (2015) 'How Well Can the Exponential-Growth Coalescent Approximate Constant-Rate Birth–Death Population Dynamics?' *Proceedings of the Royal Society B: Biological Sciences*, 282: 20150420.

Suchard, M. A. et al. (2018) 'Bayesian Phylogenetic and Phylodynamic Data Integration using BEAST 1.10', *Virus Evolution*, 4: vey016.

Tay, J. H. et al. (2022) 'The Emergence of SARS-CoV-2 Variants of Concern is Driven by Acceleration of the Substitution Rate', *Molecular Biology and Evolution*, 39: msac013.

Vasylyeva, T. I. et al. (2019) 'Tracing the Impact of Public Health Interventions on HIV-1 Transmission in Portugal using Molecular Epidemiology', *The Journal of Infectious Diseases*, 220: 233–43.

Vaughan, T. G. et al. (2014) 'Efficient Bayesian Inference under the Structured Coalescent', *Bioinformatics*, 30: 2272–9.

—— et al. (2019) 'Estimating Epidemic Incidence and Prevalence from Genomic Data', *Molecular Biology and Evolution*, 36: 1804–16.

—— et al. (2017) 'Inferring Ancestral Recombination Graphs from Bacterial Genomic Data', *Genetics*, 205: 857–70.

Volz, E. M. (2012) 'Complex Population Dynamics and the Coalescent under Neutrality', *Genetics*, 190: 187–201.

Volz, E. M., and Didelot, X. (2018) 'Modeling the Growth and Decline of Pathogen Effective Population Size Provides Insight into Epidemic Dynamics and Drivers of Antimicrobial Resistance', *Systematic Biology*, 67: 719–28.

Volz, E. M., and Frost, S. D. W. (2014) 'Sampling through Time and Phylodynamic Inference with Coalescent and Birth–Death Models', *Journal of the Royal Society Interface*, 11: 20140945.

Volz, E. M., Koelle, K., and Bedford, T. (2013) 'Viral Phylodynamics', *PLOS Computational Biology*, 9: e1002947.

Volz, E. M. et al. (2009) 'Phylodynamics of Infectious Disease Epidemics', *Genetics*, 183: 1421.

Volz, E. M., and Siveroni, I. (2018) 'Bayesian Phylodynamic Inference with Complex Models', *PLOS Computational Biology*, 14: e1006546.

Zarebski, A. E. et al. (2022) 'A Computationally Tractable Birth-Death Model that Combines Phylogenetic and Epidemiological Data', *PLOS Computational Biology*, 18: e1009805.