A computationally tractable birth-death model that combines phylogenetic and epidemiological data

Alexander E. Zarebski*[1], Louis du Plessis[1], Kris V. Parag[2, †] and Oliver G. Pybus[1, †]

[1]Department of Zoology, University of Oxford
[2]MRC Centre for Global Infectious Disease Analysis, Imperial College London

[†]These authors contributed equally

Abstract

Inferring the dynamics of pathogen transmission during an outbreak is an important problem in both infectious disease epidemiology and phylodynamics. In mathematical epidemiology, estimates are often informed by time-series of infected cases while in phylodynamics genetic sequences sampled through time are the primary data source. Each data type provides different, and potentially complementary, insights into transmission. However inference methods are typically highly specialised and field-specific. Recent studies have recognised the benefits of combining data sources, which include improved estimates of the transmission rate and number of infected individuals. However, the methods they employ are either computationally prohibitive or require intensive simulation, limiting their real-time utility. We present a novel birth-death phylogenetic model, called TimTam which can be informed by both phylogenetic and epidemiological data. Moreover, we derive a tractable analytic approximation of the TimTam likelihood, the computational complexity of which is linear in the size of the dataset. Using the TimTam we show how key parameters of transmission dynamics and the number of unreported infections can be estimated accurately using these heterogeneous data sources. The approximate likelihood facilitates inference on large data sets, an important consideration as such data become increasingly common due to improving sequencing capability.

*For correspondence email: alexander.zarebski@zoo.ox.ac.uk

1 Introduction

Estimating the prevalence of infection and transmission dynamics of an outbreak are central objectives of both infectious disease epidemiology and phylodynamics. In mathematical epidemiology, time series of reported numbers of infections (known as the epidemic curve or time series) are combined with epidemiological models to infer key parameters, such as the basic reproduction number (R_0); a fundamental descriptor of transmission dynamics (Brauer et al., 2008; Grassly and Fraser, 2008). In phylodynamics, as applied to infectious disease epidemiology, phylogenies reconstructed from pathogen genetic sequences sampled over the course of an outbreak are used to estimate either the size or growth rate of the infected population (Pybus and Rambaut, 2009; Stadler et al., 2012).

Combining information from multiple data sources has the potential to improve estimates of transmission rates and prevalence (Rasmussen et al., 2011; Moss et al., 2019), however doing so raises substantial challenges (Angelis et al., 2015). Technical challenges and the diversity of the data types used has meant that phylogenetic and epidemiological inference methods have been developed and examined largely in isolation of each other (Parag and Donnelly, 2020; Ypma et al., 2013).

The two main phylodynamic models used to describe the growth of an infectious disease outbreak are the phylogenetic birth-death (BD) model, which estimates the rate of spread of the pathogen (Nee et al., 1994; Kendall, 1948), and the coalescent process, which estimates the effective size of the infected population (Kingman, 1982; Pybus et al., 2000). Within the coalescent framework, a phylogeny reconstructed from sampled sequences is related to the effective size of the infected population, assuming that the fraction of the population sampled is small (Kingman, 1982). This relationship, when interpreted under a suitable dynamical model, allows the inference of epidemic dynamics (Pybus et al., 2001; Volz et al., 2009). Both differential equation and stochastic epidemic models have been fit to sequence data (Volz et al., 2009; Popinga et al., 2015; Tang et al., 2019), providing estimates of prevalence and R_0. Gill et al., 2016 introduced an additional way to model effective population sizes by considering the association between effective population size and time-varying covariates.

Rasmussen et al., (2011) showed how combining sequence data with an epidemic time series could allow inference of not just the epidemic size but also its growth parameters. However, this approach treated the epidemic time series as independent of the sequence data, an approximation which only holds when the number of sequences is small relative to the outbreak size. Previously, coalescent models have neglected the informativeness of sequence sampling times, although recent work has found estimates of the effective size could be improved by incorporating sampling times (Karcher et al., 2016; Parag et al., 2020). To the best of our knowledge, no coalescent model so far has utilised both epidemic time series and sequence sampling times.

Within the BD framework, births represent transmission events and deaths the cessation of being infectious (e.g. due to death, isolation or recovery). Stadler, (2010)'s birth-death-sampling (BDS) extension of Kendall's BD model (Kendall, 1948) incorporated serially-sampled sequences, which allowed estimation of the underlying epidemic growth and sampling trends. This approach was extended by Kühnert et al., (2014), who linked the BDS process to a stochastic epidemic (SIR) model under strong simplifying assumptions. The resulting model improved estimates of R_0 and provided the first means of inferring the number of unsampled members of the infected population (via estimates of epidemic prevalence). Deterministic SIR models have also been used in both BD (MacPherson et al., 2020) and coalescent frameworks (Volz et al., 2009).

Vaughan et al., (2019) relaxed the assumptions in Kühnert et al., (2014)'s model using a particle-filter approach for inference. The flexibility of the particle-filter enabled the use of both sequence and epidemic time series data. While the particle-filter represents a comprehensive ap-

proach to fusing epidemiological and phylogenetic data, it is computationally intractable, relying on intensive simulation, which can limit its application. Recent work from Manceau et al., (2020) and Gupta et al., (2020) developed a numerical scheme for computing the same likelihood (and so facilitates equivalent estimates). The numerical scheme has a smaller computational overhead, but requires calculations that have a quadratic computational complexity, i.e., that grow as the square of the size of the dataset. Moreover, the approximation used can be numerically unstable under certain conditions.

To the best of out knowledge, there is currently no existing phylogenetic inference method, in either the BD or coalescent frameworks, that can (i) formally combine both epidemiological and sequence data, (ii) estimate the prevalence of infection and (iii) be practically applied to large data sets. As sequencing costs continue to decline and large genome sequence datasets collected over the course of an outbreak become the norm, the need for a tractable solution to these problems grows. Here we present the first steps towards such a solution by approximating, and then generalising, the model of Manceau et al., 2020.

In this manuscript we describe the **T**ime-series **I**ntegration by **M**omen**t** **A**pproxi**m**ation (TimTam), a novel approach for incorporating both epidemiological and sequence data at scale. The novelty of this approach stems from two aspects. First, motivated by a result from Kendall, (1948) we approximate the prevalence distribution (and the number of unobserved lineages) with a negative binomial distribution; this approximation allows us to derive an analytic approximation to the likelihood that has a computational complexity that scales linearly with the size of the dataset. Consequently, our approach can be applied to much larger data sets than was previously possible. Second, the mathematical tractability of TimTam allows us to provide an extension to the sampling models previously considered which more closely represents how epidemiological data is usually recorded in practice. Since epidemiological data is usually only available in the form of binned (e.g. daily or weekly) counts, a time series of such observations align more closely with the data generating process (Wallinga and Teunis, 2004). For example, if a health centre is unable to report new cases over the weekend one can expect scheduled cases at the start of the following week. This is in contrast to sequence data, which is usually reported with the exact sampling date.

2 Methods

Birth-death-sampling (BDS) models, as presented in Stadler, (2010) and Stadler et al., (2013), describe sequence data that have either been collected at pre-determined points in time (hereafter scheduled observations), or opportunistically, when cases have presented themselves, (hereafter unscheduled observations). Models such as those considered in Vaughan et al., (2019) and Manceau et al., (2020) incorporate an additional data type in their sampling model, occurrence data, which represents the unscheduled observation (and subsequent removal) of infectious individuals without including them in the reconstructed phylogeny. Such occurrence data may arise e.g. when an individual receives treatment, but the pathogen genome is not sequenced. Unscheduled observations generate a point process of removal events from the infectious population (Stadler et al., 2013). The above suggests a categorisation of observations based on two attributes, (i) whether they were observed at predetermined times (*scheduled* observations) or following a point process (*unscheduled* observations), and (ii) whether the observation is included in the reconstructed phylogeny (a *sequenced* observation), or not (an *unsequenced* observation).

The categorisation above suggests a fourth data type: the scheduled observation of unsequenced (occurrence) data, which corresponds to the removal of multiple individuals from the infectious population at the same time, without incorporating them into the reconstructed phylogeny. There are several benefits to being able to describe such data. First, since epidemiological data are often given as a time series (instead of a point process) this is arguably a more natural way to utilise occurrence data in the transmission process (Wallinga and Teunis, 2004). The same could be said for sequenced samples where there may be multiple samples collected on the same day (Parag et al., 2020). The second benefit is computational. Treating observations as arising from scheduled rather than unscheduled observations simplifies the likelihood, since each scheduled event can account for multiple observations. As sequencing of pathogen genomes becomes more commonplace, the capacity to deal with large data sets becomes increasingly important. As far as we are aware, scheduled occurrence data has not been considered in any phylogenetic inference method. Below we describe this sampling model formally and the TimTam approach to evaluating its likelihood. An implementation of the likelihood is available upon request from the corresponding author.

Phylogenetic Birth-Death Process

The phylogenetic birth-death process starts with a single infectious individual at the origin time, $t = 0$. Infectious individuals "give birth" to new infectious individuals at rate λ, and are removed from the process either through naturally ceasing to be infectious (at rate μ, often called the "death" rate), or through being sampled (an observation). Unscheduled sampling of infectious individuals occurs at different rates depending on whether samples are sequenced (at rate ψ) or not (at rate ω). Individuals can also be removed in scheduled sampling events. Scheduled sampling occurs at predetermined times when each infectious individual is independently removed with a fixed probability; ρ for sequenced samples and ν for unsequenced samples. An illustrative example is shown in Figure 1. For ease of notation we assume that all samples arising from a scheduled sampling event are either sequenced or not. We denote these times r_i for the sequenced sampling events and u_i for the unsequenced ones, and assume these times are known *a priori*, since they are under the control of those observing the system. The parameters of interest in this combined process are $\theta := (\lambda, \mu, \psi, \rho, \omega, \nu)$.

Realisations of the process are binary trees with internal nodes corresponding to infection events and terminal nodes representing one of the removal events as shown in Figure 1. Note that we assume the edges of the tree are labelled with their length to ensure that the nodes appear at

the correct depth. We refer to the resulting tree of all infected individuals as the *transmission tree*. The subtree containing only the terminal nodes corresponding to sequenced samples (both scheduled and unscheduled) is called the *reconstructed tree*. In practice, the reconstructed tree is estimated from pathogen genomes; here we assume the reconstructed tree is known *a priori*.

The reconstructed tree can be summarised by its *lineages through time* (LTT) plot, which depicts the number of lineages in the tree at each point in time. We define the number of *hidden* lineages through time as the count of lineages that appear in the transmission tree but not in the reconstructed tree. The LTT and the unscheduled (point process) and scheduled (time series) of unsequenced samples can all be reduced to a sequence of events, $\mathcal{E}_{1:N}$, starting from the origin and moving forward through time up to the present (i.e., the time of the last observation):

$$\mathcal{E}_{1:N} = \{(\Delta t_i, e_i)\}_{i=1\ldots N} \tag{1}$$

with Δt_i denoting the time since the previous observation and e_i describing the event that was observed at that time: $e_i \in \{\lambda\text{-event}, \psi\text{-event}, \rho\text{-event}, \omega\text{-event}, \nu\text{-event}\}$. The sequence of events identified from Equation (1) will be used to derive the likelihood of the process.

For ease of notation, the time of event number i is denoted t_i, so $t_i := \Delta t_1 + \cdots + \Delta t_i$. The value of K_i is the number of lineages in the reconstructed tree *directly after* the event at time t_i and H_i denotes the number of hidden lineages. Note that while K_i remains constant between observations, $H(t)$ may change. The changes in the LTT and the number of hidden lineages at time t_i is denoted ΔK_i and ΔH_i.

There are some differences between the process described above and that of (Manceau et al., 2020). Manceau et al., 2020 allow for the observation of infectious individuals without removal, i.e., allowing them to appear as an unscheduled sample but potentially able to subsequently give birth to new infections. Accounting for these so-called *sampled ancestors* introduces an additional parameter which is the probability of removal upon sampling (Stadler, 2010; Gavryushkina et al., 2014). As mentioned above, the inclusion of scheduled occurrence data is novel, hence is not part of the process considered by Manceau et al., (2020) or any other work, as far as we know.

The Likelihood

Here we describe the likelihood function for the process described above and the distribution of the number of hidden lineages, $H(t)$, conditional upon the observed N events from Equation (1). The joint posterior distribution of these quantities can be factorised as follows:

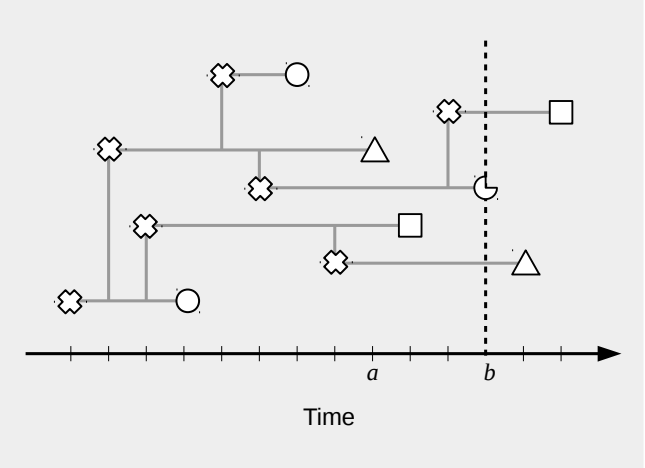
$$f(\theta, H(t) \mid \mathcal{E}_{1:N}) \propto f(H(t) \mid \mathcal{E}_{1:N}, \theta) f(\mathcal{E}_{1:N} \mid \theta) \pi(\theta). \tag{2}$$

with $\pi(\theta)$ as a prior distribution over the parameters of the process, θ. The constant of proportionality is simply that required to normalise the resulting posterior distribution $f(\theta, H(t) \mid \mathcal{E}_{1:N})$.

The remaining two terms compose the process likelihood. The first, which is the distribution of $H(t)$, is calculated incidentally while evaluating $f(\mathcal{E}_{1:N} \mid \theta)$ as explained below. The second term, $f(\mathcal{E}_{1:N} \mid \theta)$, is the likelihood of the observed events given the process parameters. While each observed event depends on all the prior observations, we can factorise this likelihood into sequential terms:

$$f(\mathcal{E}_{1:N} \mid \theta) = \prod_{i=1}^{N} f(\mathcal{E}_i \mid \mathcal{E}_{1:(i-1)}, \theta) = \prod_{i=1}^{N} c_i l_i$$

where c_i is the probability that during the interval (t_{i-1}, t_i) (i.e., between events $\mathcal{E}_{i-1}$ and $\mathcal{E}_i$) there was no observed event, and l_i is the probability of observing event e_i.



ψ (sequenced)
ρ (catastrophe)
ω (observed)
μ (removed)
λ (infection)

Delay	Event
1	λ
2	ψ
1	λ
2	ψ
2	ω
3	ρ^1
1	ω

Figure 1: An illustration of birth-death-sampling process and the resulting data set: the tree depicts the BD process with the shapes indicating which events occurred when; each horizontal line corresponds to the time an individual was infectious. The table shows the corresponding data set with the delay between subsequent observations and the observed event after that delay. At time a there is a ω-event, then after a delay of 3 units of time until time b when there is ρ-event with one individual sampled, denoted ρ^1.

To describe how many hidden linages there are at time t, $H(t)$, we define a generating function $M(t,z)$ for the distribution conditional upon events that have been observed by time t. This is defined as

$$M(t,z) = \sum_h \mathbb{P}(H(t) = h | \mathcal{E}_{1:x} \leq t) z^h,$$

where we have abused our notation by using $\mathcal{E}_{1:x} \leq t$ to denote just the events which occur at times not after t. The c_i and l_i can be computed using properties of the generating function $M(t,z)$. Further, to indicate the generating function over a particular interval of time, let $M_i(t,z)$ denote $M(t,z)$ for $t_i \leq t < t_{i+1}$, i.e., between events $\mathcal{E}_i$ and $\mathcal{E}_{i+1}$. For each of these intervals, we can describe M_i using a PDE derived from the master equations of the process we have described above.

Since the process starts with only a single infected individual (who appears as the root of the tree), by definition $H(0) = 0$. Therefore the initial condition of the generating function is $M_0(0,z) = 1$, and the PDEs describing the M_i are:

$$M_i(t_i, z) = F_i(z) \tag{3}$$

$$\partial_t M_i = (\lambda z^2 - \gamma z + \mu)\partial_z M_i + K_i(2\lambda z - \gamma) M_i. \tag{4}$$

Here $\gamma = \lambda + \mu + \psi + \omega$, the parameters ρ and ν do not appear in the PDE because they only occur in the boundary conditions at the scheduled events. The boundary condition, $F_i(z)$, is the PGF of H after the observation at t_i. The number of lineages in the reconstructed tree, K_i,

only changes when there is a birth, or a sequenced sample and so is a constant in the PDE. The solution during the intervals between observations is[1]

$$M_i(t,z) = F_i\left(p_0(t_{i+1}-t,z)\right)\left(\frac{p_1(t_{i+1}-t,z)}{1-z}\right)^{K_i}. \tag{5}$$

The functions p_0 and p_1 are standard results for birth-death-sampling models (see Stadler, (2010)) and a derivation is given in Supplementary Materials.

Before proceeding, we discuss an important property of the solution given in Equation (5): the coefficients of the generating function M_i do not sum to unity for all t. The normalising constant for $M_i(t_{i+1},z)$ is the probability, c_{i+1}, of there having been no event during the preceding interval of time. Hence, we can calculate the c_{i+1} by evaluating $M_i(t_{i+1},z)$ at $z=1$. Therefore the generating function of the distribution of $H(t_{t+1}^-)$ (i.e., the limiting distribution of the number of hidden distributions prior to an observation) is $\mathcal{M}_i := M_i(t_{i+1}^-)/c_i$, where we use the notation $t^\pm$ to indicate the left and right limits respectively and the inclusion of the denominator c_i ensures that the coefficients sum to 1.

To calculate the PGF of $H(t_{i+1})$, (i.e., the PGF conditioning on the observation at time t_{i+1},) we need l_{i+1}. To do this we consider the changes to the distribution of H that result from observing each possible event. Since λ- and ψ-events do not influence the number of H-lineages, the PGF does not change: $M_{i+1}(t_{i+1}) := \mathcal{M}_i(t_{i+1})$. The likelihood of these events, e_i, is λ and ψ respectively. For a ω-event we need to shift the whole distribution of H and account for the unknown number of hidden lineages that could have been sampled, this is achieved by taking the partial derivative of the generating function[2]. The likelihood of an ω-event is the normalising constant after the differentiation:

$$\begin{aligned} l_{i+1} &= \lim_{z\to 1^-} \omega\partial_z \mathcal{M}_i(t_{i+1},z), \quad \text{and so} \\ M_{i+1}(t_{i+1}) &= \frac{\omega}{l_{i+1}}\partial_z \mathcal{M}_i(t_{i+1},z). \end{aligned} \tag{6}$$

For a scheduled sampling event, at time r with removal probability ρ, we need to account for the survival of each of the H-lineages that were not sampled, those that were, and the number of lineages in the reconstructed tree that were not removed during this scheduled sampling. This leads to the following likelihood factor and updated PGF:

$$\begin{aligned} l_{i+1} &= (1-\rho)^{K_{i+1}}\rho^{\Delta K_{i+1}} \lim_{z\to 1^-} \mathcal{M}_i(t_{i+1},(1-\rho)z) \quad \text{and} \\ M_{i+1}(t_{i+1}) &= \frac{(1-\rho)^{K_{i+1}}\rho^{\Delta K_{i+1}}}{l_{i+1}}\mathcal{M}_i(t_{i+1},(1-\rho)z). \end{aligned} \tag{7}$$

Last, we include scheduled unsequenced samples, i.e. the observation and simultaneous removal of multiple lineages without subsequent inclusion in the reconstructed phylogeny. Previously, we noted that a single ω-sampling corresponds to differentiating the PGF of H once. If at time u each lineage is sampled with probability ν and n lineages in total are sampled, then we must take the n-th derivative and accumulate a likelihood factor for the removed and non-removed lineages of $(1-\nu)^K\nu^n$; while scaling z by a factor of $1-\nu$ to account for the H-lineages that were not sampled. Using Equations (6) and (7), the likelihood and updated PGF after a ν-sample are:

[1] This appears as Proposition 4.1 in Manceau et al., 2020.

[2] Differentiation of the PGF achieves this because it shifts the coefficients of the series and weights them by the number of possible ways in which the sample could have been drawn. For example, consider the term of the series $h_j z^j$ which then becomes $jh_j z^{j-1}$ because after the sample the probability of there being $j-1$ hidden lineages is the probability there were previously j and that one of those j lineages was sampled.

$$
\begin{aligned}
l_{i+1} &= (1-\nu)^{K_{i+1}} \nu^{\Delta H_{i+1}} \lim_{z \to 1^-} \partial_z^{\Delta H_{i+1}} \mathcal{M}_i(t_{i+1}, (1-\nu)z) \quad \text{and} \\
M_{i+1}(t_{i+1}) &= \frac{(1-\nu)^{K_{i+1}} \nu^{\Delta H_{i+1}}}{l_{i+1}} \partial_z^{\Delta H_{i+1}} \mathcal{M}_i(t_{i+1}, (1-\nu)z).
\end{aligned} \tag{8}
$$

Above we have derived expressions for c_i and l_i which allow us to compute the likelihood of an observed set of events, $\mathcal{E}_{1:N}$. The outline of the likelihood calculation above is similar to that of Manceau et al., (2020) but with the addition of scheduled unsequenced samples, and greater emphasis on the handling of *repeated* scheduled sequenced sampling. Unfortunately, the expressions above are in terms of limits and derivatives of generating functions that are difficult to manipulate. As noted by Manceau et al., (2020), the evaluation of these expressions becomes increasingly computationally demanding when done numerically and attempts to find a simplified expression using computer algebra systems did not yield suitable results. The strategy followed in Manceau et al., (2020) was to either solve the differential equations numerically or to approximate it with a set of basis functions. The former approach requires truncating an infinite linear system of ordinary differential equations (ODEs) and solving it for each time interval, an operation which is cubic in the size of the truncated system (due to taking the matrix exponential). The latter approach attains quadratic complexity albeit by introducing a further approximation. The accuracy of the numerical solution of Manceau et al., (2020) will increase initially with the size of the truncated system, but at larger values, numerical error from computing the matrix exponential could become significant. The TimTam approach we describe below is our novel approach for avoiding these problems; it has a linear complexity and avoids the need for any numerical integration.

An Analytic Approximation

To apply this approximation, recall that we can evaluate the full PGF point-wise given the boundary condition F_i (see Equation 3). Moreover, as shown in the Supplementary Materials, the generating function of the negative binomial (NB) distribution is closed under partial derivatives (up to a simple multiplicative constant) and partial derivatives of PGFs can be used to calculate the mean and variance of a distribution. Our TimTam model can be described as simply replacing the PGF of H with a NB PGF with equivalent mean and variance whenever necessary. Algorithmically this can be expressed in the following steps:

1. Start at time t_i with M_i and solve for M_i at time t_{i+1}.
2. Define $c_i := M_i(t_{i+1}, 1^-)$.
3. Define $\mathcal{M}_i := M_i / c_i$.
4. Define $\widetilde{\mathcal{M}_i}$ to be the NB approximation to the distribution with PGF $\mathcal{M}_i$.
5. Use $\widetilde{\mathcal{M}_i}$ to compute the likelihood of $\mathcal{E}_{i+1}$ and call it l_i.
6. Define M_{i+1} as the PGF of the distribution obtained by conditioning the NB approximation on $\mathcal{E}_{i+1}$.
7. Increment the log-likelihood by $\log(c_i l_i)$.

The steps involved only require the evaluation of closed form expressions and the amount of computational work is linear in the number of observed events.

Our use of a NB moment-matching approximation is not arbitrary. Kendall observed that the number of lineages descending from a single lineage has a zero-inflated geometric distribution (Kendall, 1948). Moreover, it is well known that the sum of independent and identically distributed geometric random variables follows a NB distribution. Our approach of treating the number of lineages derived from n individuals as a NB random variable is somewhat motivated by combining these two properties. Further support for our approximation is obtained by considering an equivalent BD process, but with the modified total birth rate of $\lambda n + a$ where a is a small offset representing an immigration rate that leads to the removal of the extra (unobservable) zeros. Such processes can be described by NB lineage distributions at all times of their evolution and are stable to the inclusion of additional event types. (Ycart, 1988; Kapodistria et al., 2016).

It is interesting to note that both partial differentiation by z and scaling z by a constant factor in the PGF of a NB random variable produces another PGF for a NB random variable with a multiplicative factor. Put another way, the family of NB PGFs is invariant (up to a multiplicative constant) under the algebraic operations we care about. The significance of this is that if we assume that the distribution of H is NB, then conditioning on an event does not require any further approximation to produce subsequent NB distributions as the initial condition of the next interval.

Additional comments

Conditioning upon observation

The likelihood developed above applies to an arbitrary realisation of the birth-death process. However in practice, the existence of a data set usually means the outbreak has escaped extinction due to stochastic effects. This generates a survivorship bias i.e. we only ever consider the likelihood of realisations which generate at least one observation. In the phylogenetic BD literature, this is readily acknowledged and accounted for by conditioning the process in one of several ways (Nee et al., 1994; Stadler, 2012)[3]. To adjust for this, one should condition upon there being at least $n \geq 1$ observations between the origin and the present. If there were only unscheduled samples in our data set, existing approaches to conditioning the process against extinction could easily be applied to this model. Here we do not consider the problem of conditioning the process against extinction under the repeated scheduled sampling setting.

Origin time vs TMRCA

The definition of the likelihood above assumes that the origin of the phylogeny is known a priori or is a parameter to be estimated. This is because we need the initial condition $M_0(0, z) = 1$. In practice this is unlikely to be the case as the phylogeny will likely only be known up to the time of the most recent common ancestor (TMRCA). There are two ways in which this might be remedied. The first, and simplest, is to treat the origin time as an additional parameter to be estimated. The second is to set a boundary condition at the TMRCA and to estimate the distribution of H.

If we were confident that the outbreak stemmed from a single initial case, then the former method would be more suitable, especially if there was prior knowledge that could constrain the estimate of the origin time. On the other hand, if we faced substantial uncertainty about how the outbreak began and sequencing was sparse, i.e., low ψ and ρ, then the TMRCA may be relatively recent and estimating the origin could be particularly challenging. In this case,

[3]There are similar results in the mathematical epidemiology literature, however they are less frequently used, e.g. (Mercer et al., 2011).

the latter approach may be more suitable. This would involve estimating the distribution of $H(t_{\mathrm{TMRCA}})$ and hence its generating function $M_1(t_{\mathrm{TMRCA}}, z)$, presumably from the family of NB distributions.

We have presented the log-likelihood in terms of the assumption of a known origin time, because that is a more mathematically convenient approach, however the most appropriate method will depend on the types of questions to be answered and the potential availability of prior information to inform the estimate of the origin time.

3 Results

Comparison with existing results

In this section we validate and compare our TimTam approach to the method from Manceau et al., (2020), hereafter called the Manceau algorithm. Figure 2 shows the the log-likelihood function evaluated under the TimTam approach and the Manceau algorithm, for 40 simulated data sets. The simulation used the parameters given in Table 1 was repeated to get a range of sample sizes from 5 to 200 observed events (which includes both births and samples). Both methods produce very similar log-likelihood values with the TimTam approach explaining 99% of the variation in the Manceau algorithm values under a linear model.

Since the Manceau algorithm requires a truncation parameter to be specified, we first obtained sensible values on a per simulation basis by increasing this value until the log-likelihood changed by less than 0.1% if the truncation parameter was incremented further. The resulting truncation parameters are shown in Supplementary Figure 1. The full details regarding how the simulated data were generated, how the benchmarks where evaluated and how the truncation parameter for the Manceau algorithm was selected are given in the Supplementary Materials.

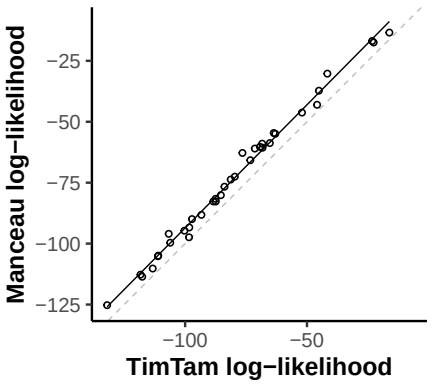


Figure 2: The value of the log-likelihood using the TimTam approximation and the numerical scheme from Manceau et al., (2020). The values are in good agreement with an R^2 of 0.99. The solid black line shows a linear model fit to the data and the grey dashed line follows $y = x$.

To understand the computational efficiency of our new approach, we recorded the time required to evaluate the log-likelihood on each of the data sets considered above. In these simulations we selected the truncation parameter of the Manceau *et al* algorithm before we estimated the evaluation time of the likelihood so that this computation was not included as part of the evaluation time. The average times to evaluate the likelihood are shown in Figure 3. For TimTam the evaluation time grows approximately linearly with the size of the dataset, $\propto n^{1.08}$ where the 95% CI on the exponent is (1.07,1.10). On the other hand, for Manceau et al., (2020)'s numerical scheme the evaluation time grows approximately quadratically, $\propto n^{2.10}$ CI (1.82,2.38). In addition to the improvement in computational complexity, the average evaluation times over the example data sets are orders of magnitude smaller for TimTam, which takes less than a millisecond in comparison to the several seconds required by the implementation presented in

Manceau et al., (2020). We caution against reading too much into the absolute average computation times, since we used Haskell to implement our method, whereas Manceau et al., (2020) used a combination of C and Python, hence it is likely that the faster computation time is a feature of the programming language used and not the algorithm (both implementations are available online). Nonetheless, the computational complexities are features of the respective algorithms and means that the TimTam approach will outperform the Manceau algorithm for large datasets, regardless of the implementation.

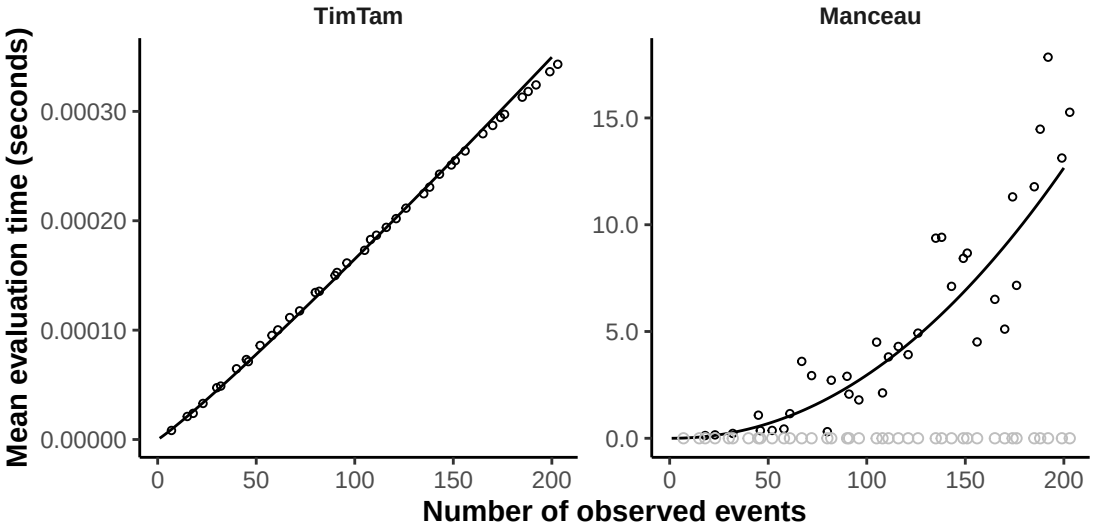


Figure 3: The average time taken to evaluate the likelihood grows approximately linearly with the number of data points: $\propto n^{1.08}$ (1.07,1.10) using TimTam while the algorithm from Manceau *et al* (2020) has approximately quadratic growth $\propto n^{2.10}$ (1.82,2.38). The grey points in the second panel show the TimTam values again on the same scale.

Large data set example

Having validated TimTam against the Manceau algorithm, we now showcase our approach as an estimation scheme that merges all the data types considered in this manuscript. We used the parameters listed in Table 2 to generate a larger simulation. To show the effect of the simulation length and the number of observations on statistical power, we truncated our simulation at two timepoints, $t = 12$ and $t = 16$. The numbers of observations of each type in each of the two partial datasets are shown in Table 3. Figure 4 shows cross-sections of the TimTam log-likelihood function generated by fixing the parameters and then varying each element of the parameter vector individually to explore the surface. This was done using the data from $t = 16$ in the simulation centered about two sets of parameters: those used in the simulation and the maximum-likelihood parameter estimates, obtained by numerically optimising the log-likelihood function while fixing the death rate (μ) to its true value. The log-likelihood cross-sections for the datasets truncated at $t = 12$ is shown in Supplementary Figures 2.

We also investigated how well TimTam estimates the prevalence of hidden lineages through time. Figure 5 shows the number of hidden lineages in the simulation at various snapshots, together with the estimated solution to the filtering problem, i.e., estimation of the prevalence

given the incomplete data available at that point in the simulation.

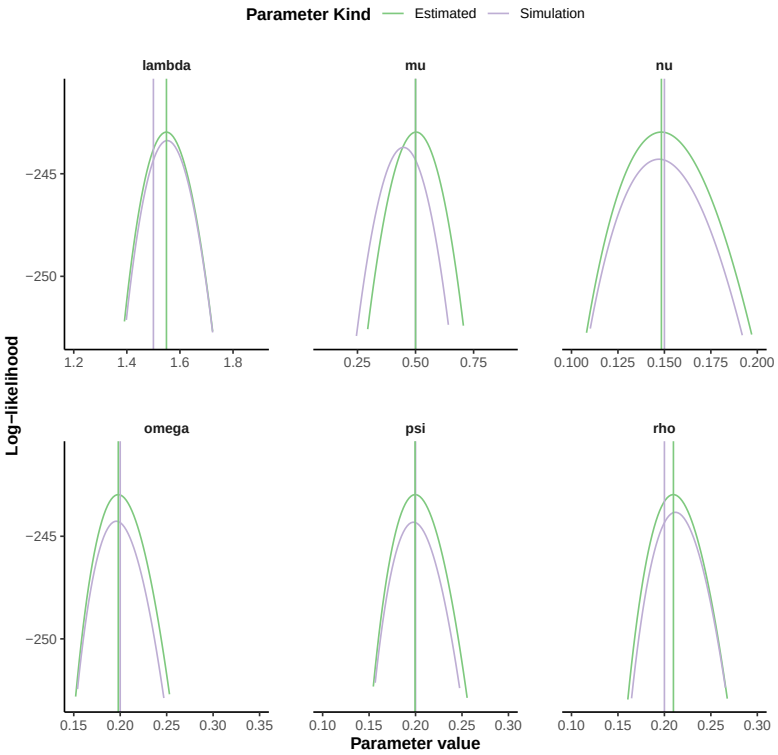


Figure 4: Cross sections of the log-likelihood function taken about the parameter values used in the simulation (lilac) and the estimated values (green), both of which are indicated by vertical lines, given the data that was available at $t = 16$.

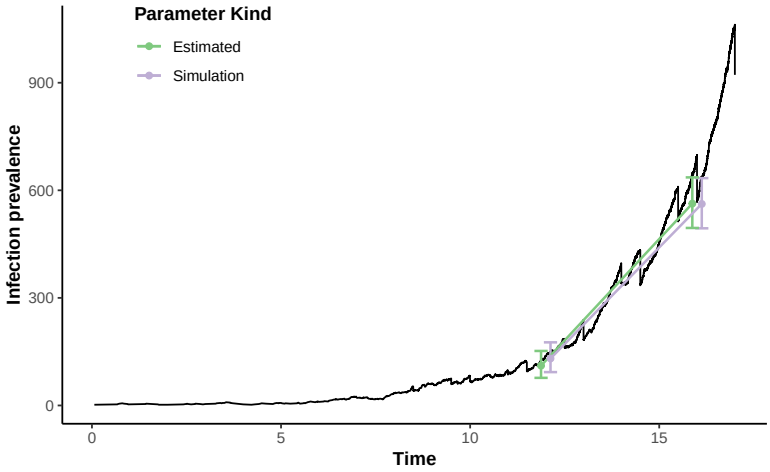


Figure 5: The LTT plot of the simulated data set and the estimates of the prevalence generated at a couple points in the simulation (based on only the data prior to that point in time) using either the "true" parameters responsible for the simulated data or the estimated values conditional upon the data up to that point in the simulation.

4 Discussion

We have described an analytic approximation, called TimTam, for the likelihood of a birth-death-sampling model which can also describe *scheduled data* i.e. cohort sampling at pre-determined times. TimTam can be used with both sequenced and unsequenced samples, i.e., the observations can either be included in the reconstructed tree, or as occurrence data. Our approach generalises previous birth-death estimation frameworks by accommodating and exploiting more data types than have been previously considered and makes it possible to scale existing analyses to larger data sets.

Our work is a step towards more flexible time series-based approaches to phylodynamics, where samples from multiple lineages are considered contemporaneously. This extends the more common point-process based paradigm in which lineages are sampled continuously through time and therefore must be considered individually. In addition, the TimTam likelihood provides an estimate of the distribution of the final prevalence of infection, allowing both the estimation of summary statistics such as R_0 and the total number of cases. Comparison of TimTam to existing algorithms on small to moderate sized data sets suggests this is a faithful representation of the true likelihood function and that the empirical complexity behaves as expected.

The strength of TimTam lies in its use of moment matching. This simplifying assumption allowed us to extend the existing observation models to include the scheduled observation of unsequenced infections. The computational efficiency means that even when we integrate these heterogeneous data sources, our framework remains tractable even for large datasets. Taken together this means phylodynamics can utilise a greater amount and variety of data generated by surveillance and sequencing efforts, both of which are becoming increasingly common in contemporary epidemiology. Moreover, we anticipate this underlying approach will allow us to generate analogous approximations for other phylodynamic models.

At present, we cannot provide rigorous bounds on the error introduced by this approximation (although work is underway on this). However, based on the motivating work from Kendall, (1948), we conjecture that if the probability of extinction becomes large, the zero inflation in the geometric distributions describing the number of descending lineages might become an issue. Since our focus is on large datasets, which will describe established epidemics, we suspect that in practice this situation will rarely arise. Additionally, as the death rate increases, the power of birth-death models as an inference tool is naturally limited by a lack of data (Kubo and Iwasa, 1995; Pyron and Burbink, 2013).

If this method is to be used in small outbreaks or, when the reproduction number is low, sensitivity analyses will be necessary to check the fidelity of the NB approximation, as in this situation the zero-inflation lost in our approximation may become substantial. Moreover, we have neither examined the conditions necessary for statistical identifiability of the parameters nor adapted our model likelihood to condition it against extinction (Stadler, 2012; Parag and Pybus, 2018). Calculating extinction probabilities for this model is complicated by the iterated scheduled sampling events and the unsequenced samples.

Our work echoes the frameworks of Vaughan et al., 2019 and Manceau et al., 2020, but trades some generality for simplicity and tractability. Specifically, Vaughan et al., 2019 presented a particle filtering method that can be applied more generally, while Manceau et al., 2020 derived a complete posterior predictive distribution of prevalence over time, which allows the optimal study of historical transmission. While the former is able to describe a greater variety of birth-death processes and the latter can be used to estimate additional properties of the processes considered, there are substantial limitations of scalability in both. While TimTam may not match the current level of generality in Vaughan et al., 2019 or the rigour of Manceau et al., 2020, our method provides a computationally efficient method for handling diverse data types that is

scalable to modern datasets. We are pursuing the aggregation of point-process observations into a time series which provides a closer link to how epidemiological data is usually recorded where it is typically available at a daily or weekly resolution. Moreover, this leads to improvements in performance for large datasets as multiple data points can be handled in a single expression. As the availability of phylogenetic data (derived from sequences or contact tracing) increases and the size of these data grows, such approximation schemes will become increasingly valuable.

Table 1: The parameters used to simulate the data sets for the empirical investigation of the computational complexity.

Parameter	Description	Value
λ	Birth rate	1.5
μ	Death rate	0.3
ψ	Sequenced sampling rate	0.3
ρ	Scheduled sequenced sampling probability	0.5
r	Scheduled sequenced sampling time	6
ω	Unsequenced sampling rate	0.3
	Simulation duration	6

Table 2: The parameters used to simulate the large data set.

Parameter	Description	Value
λ	Birth rate	1.5
μ	Death rate	0.5
ψ	Sequenced sampling rate	0.2
ρ	Scheduled sequenced sampling probability	0.2
r_i	Scheduled sequenced sampling times	$\{2.5, 4, 5.5, 7, 8.5, 10, 11.5, 13, 14.5, 16, 17.5\}$
ω	Unsequenced sampling rate	0.2
ν	Scheduled unsequenced sampling probability	0.15
u_i	Scheduled unsequenced sampling times	$\{2, 3.5, 5, 6.5, 8, 9.5, 11, 12.5, 14, 15.5, 17\}$
	Simulation duration	17
	Inference times	$\{12, 16\}$

Table 3: The number of events that had been observed at each point in the simulation where inference was carried out.

Observation time	Number of observed events
12	315
16	1415

5 Acknowledgements

AEZ, OGP and LdP are supported by The Oxford Martin Programme on Pandemic Genomics. KVP is funded under grant reference MR/R015600/1 by the UK Medical Research Council (MRC) and the UK Department for International Development (DFID)

References

Angelis, Daniela De et al. (2015). "Four key challenges in infectious disease modelling using data from multiple sources". In: *Epidemics* 10. Challenges in Modelling Infectious Disease Dynamics, pp. 83 –87. ISSN: 1755-4365. DOI: https://doi.org/10.1016/j.epidem.2014.09.004. URL: http://www.sciencedirect.com/science/article/pii/S175543651400053X.

Brauer, Fred, Pauline van den Driessche, and Jianhong Wu (2008). *Mathematical Epidemiology*. Springer, Berlin, Heidelberg.

Gavryushkina, Alexandra et al. (2014). "Bayesian Inference of Sampled Ancestor Trees for Epidemiology and Fossil Calibration". In: *PLoS Comput Biol* 10.12, e1003919. DOI: 10.1371/journal.pcbi.1003919. URL: http://dx.doi.org/10.1371/journal.pcbi.1003919 (visited on 08/27/2015).

Gill, Mandev S. et al. (2016). "Understanding Past Population Dynamics: Bayesian Coalescent-Based Modeling with Covariates". In: *Systematic Biology* 65.6, pp. 1041–1056. ISSN: 1063-5157.

Grassly, Nicholas C. and Christophe Fraser (2008). "Mathematical models of infectious disease transmission". en. In: *Nature Reviews Microbiology* 6.6, pp. 477–487. ISSN: 1740-1526. DOI: 10.1038/nrmicro1845. URL: http://www.nature.com/nrmicro/journal/v6/n6/full/nrmicro1845.html (visited on 12/22/2015).

Gupta, Ankit et al. (2020). "The probability distribution of the reconstructed phylogenetic tree with occurrence data". In: *Journal of Theoretical Biology* 488, p. 110115. ISSN: 0022-5193. DOI: https://doi.org/10.1016/j.jtbi.2019.110115. URL: http://www.sciencedirect.com/science/article/pii/S0022519319304849.

Kapodistria, S, T Phung-Duc, and J Resing (2016). "Linear birth/immigration-death process with binomial catastrophes". In: *Prob. Eng. Inform. Sciences* 30, pp. 79–111.

Karcher, M, J Palacios, T Bedford, et al. (2016). "Quantifying and Mitigating the Effect of Preferential Sampling on Phylodynamic Inference". In: *PLoS Comp. Bio* 12.3.

Kendall, David G. (1948). "On the Generalized "Birth-and-Death" Process". In: *The Annals of Mathematical Statistics*. DOI: 10.1214/aoms/1177730285.

Kingman, J (1982). "On the Genealogy of Large Populations". In: *J. Appl. Prob* 19, pp. 27–43.

Kubo, T and Y Iwasa (1995). "Inferring the Rates of Branching and Extinction from Molecular Phylogenies". In: *Evolution* 49.4, pp. 694–704.

Kühnert, D, T Stadler, T Vaughan, et al. (2014). "Simultaneous reconstruction of evolutionary history and epidemiological dynamics from viral sequences with the birth – death SIR model". In: *J. R. Soc. Interface* 11.20131106.

MacPherson, Ailene et al. (2020). "A General Birth-Death-Sampling Model for Epidemiology and Macroevolution". In: *bioRxiv*. DOI: 10.1101/2020.10.10.334383. URL: https://www.biorxiv.org/content/early/2020/10/11/2020.10.10.334383.

Manceau, Marc et al. (2020). "The probability distribution of the ancestral population size conditioned on the reconstructed phylogenetic tree with occurrence data". In: *Journal of Theoretical Biology*, p. 110400. ISSN: 0022-5193. DOI: https://doi.org/10.1016/j.jtbi.2020.110400. URL: http://www.sciencedirect.com/science/article/pii/S0022519320302551.

Mercer, G. N., K. Glass, and N. G. Becker (2011). "Effective reproduction numbers are commonly overestimated early in a disease outbreak". In: *Statistics in Medicine* 30.9, pp. 984–994. DOI: 10.1002/sim.4174.

Moss, Robert et al. (2019). "Accounting for Healthcare-Seeking Behaviours and Testing Practices in Real-Time Influenza Forecasts". In: *Tropical Medicine and Infectious Disease* 4.1. ISSN: 2414-6366. DOI: 10.3390/tropicalmed4010012.

Nee, S, R May, and P Harvey (1994). "The Reconstructed Evolutionary Process". In: *Phil Trans R Soc B* 344, pp. 305–11.

Parag, K and C Donnelly (2020). "Adaptive Estimation for Epidemic Renewal and Phylogenetic Skyline Models". In: *Syst. Biol* syaa035.

Parag, Kris V and Oliver G Pybus (2018). "Exact Bayesian inference for phylogenetic birth-death models". In: *Bioinformatics* 34.21, pp. 3638–3645. ISSN: 1367-4803. DOI: 10.1093/bioinformatics/bty337. eprint: https://academic.oup.com/bioinformatics/article-pdf/34/21/3638/26146996/bty337.pdf. URL: https://doi.org/10.1093/bioinformatics/bty337.

Parag, Kris V, Louis du Plessis, and Oliver G Pybus (2020). "Jointly Inferring the Dynamics of Population Size and Sampling Intensity from Molecular Sequences". In: *Molecular Biology and Evolution* 37.8, pp. 2414–2429. ISSN: 0737-4038. DOI: 10.1093/molbev/msaa016. eprint: https://academic.oup.com/mbe/article-pdf/37/8/2414/33564924/msaa016.pdf. URL: https://doi.org/10.1093/molbev/msaa016.

Popinga, Alex et al. (2015). "Inferring Epidemiological Dynamics with Bayesian Coalescent Inference: The Merits of Deterministic and Stochastic Models". In: *Genetics* 199.2, pp. 595–607. ISSN: 0016-6731. DOI: 10.1534/genetics.114.172791. eprint: https://www.genetics.org/content/199/2/595.full.pdf. URL: https://www.genetics.org/content/199/2/595.

Pybus, Oliver G. and Andrew Rambaut (2009). "Evolutionary analysis of the dynamics of viral infectious disease". en. In: *Nature Reviews Genetics* 10.8, pp. 540–550. ISSN: 1471-0056. DOI: 10.1038/nrg2583. URL: http://www.nature.com/nrg/journal/v10/n8/abs/nrg2583.html (visited on 11/06/2016).

Pybus, Oliver G., Andrew Rambaut, and Paul H. Harvey (2000). "An Integrated Framework for the Inference of Viral Population History From Reconstructed Genealogies". In: *Genetics* 155.3, pp. 1429–1437. ISSN: 0016-6731. eprint: https://www.genetics.org/content/155/3/1429.full.pdf. URL: https://www.genetics.org/content/155/3/1429.

Pybus, Oliver G. et al. (2001). "The Epidemic Behavior of the Hepatitis C Virus". In: *Science* 292.5525, pp. 2323–2325. DOI: 10.1126/science.1058321. URL: https://science.sciencemag.org/content/292/5525/2323.

Pyron, R and F Burbink (2013). "Phylogenetic Estimates of Speciation and Extinction Rates for Testing Ecological and Evolutionary Hypotheses". In: *Trends in Ecology and Evolution* 28.12, pp. 729–36.

Rasmussen, David A., Oliver Ratmann, and Katia Koelle (2011). "Inference for Nonlinear Epidemiological Models Using Genealogies and Time Series". In: *PLOS Computational Biology* 7.8, pp. 1–11. DOI: 10.1371/journal.pcbi.1002136. URL: https://doi.org/10.1371/journal.pcbi.1002136.

Stadler, T, R Kouyos, V von Wyl, et al. (2012). "Estimating the Basic Reproductive Number from Viral Sequence Data". In: *Mol. Biol. Evol* 29.1, pp. 347–57.

Stadler, T et al. (2013). "Birth-death Skyline Plot reveals Temporal Changes of Epidemic Spread in HIV and Hepatitis C Virus (HCV)". In: *PNAS* 110.1, pp. 228–33.

Stadler, Tanja (2010). "Sampling-through-time in birth-death trees". In: *Journal of Theoretical Biology* 267.3, pp. 396–404. ISSN: 0022-5193. DOI: https://doi.org/10.1016/j.jtbi.2010.09.010.

— (2012). "How Can We Improve Accuracy of Macroevolutionary Rate Estimates?" In: *Systematic Biology* 62.2, pp. 321–329. DOI: 10.1093/sysbio/sys073.

Tang, Mingwei et al. (2019). "Fitting stochastic epidemic models to gene genealogies using linear noise approximation". In: *arXiv*. URL: https://arxiv.org/abs/1902.08877.

Vaughan, Timothy G et al. (2019). "Estimating Epidemic Incidence and Prevalence from Genomic Data". In: *Molecular Biology and Evolution* 36.8, pp. 1804–1816. DOI: 10.1093/molbev/msz106.

Volz, Erik M. et al. (2009). "Phylodynamics of Infectious Disease Epidemics". In: *Genetics* 183.4, pp. 1421–1430. DOI: 10.1534/genetics.109.106021.

Wallinga, Jacco and Peter Teunis (2004). "Different Epidemic Curves for Severe Acute Respiratory Syndrome Reveal Similar Impacts of Control Measures". In: *American Journal of Epidemiology* 160.6, pp. 509–516. ISSN: 0002-9262. DOI: 10.1093/aje/kwh255. URL: https://doi.org/10.1093/aje/kwh255.

Ycart, B (1988). "A characterstic property of linear growth birth and death processes". In: *Sankhya A* 50, pp. 184–9.

Ypma, R, W van Ballegooijen, and J Wallinga (2013). "Relating Phylogenetic Trees to Transmission Trees of Infectious Disease Outbreaks". In: *Genetics* 195, pp. 1055–62.