medRxiv preprint doi: https://doi.org/10.1101/2023.08.03.23293620; this version posted August 8, 2023. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity.
It is made available under a CC-BY-ND 4.0 International license .

Estimating epidemic dynamics with genomic and time series data

Alexander E. Zarebski[1,2], Antoine Zwaans[3,4], Bernardo Gutierrez[1], Louis du Plessis[3,4], and Oliver G. Pybus[1,5]

[1]Department of Biology, University of Oxford, Oxford, UK
[2]School of Mathematics and Statistics, University of Melbourne, Melbourne, Australia
[3]Department of Biosystems Science and Engineering, ETH Zürich, Basel, Switzerland
[4]Swiss Institute of Bioinformatics, Lausanne, Switzerland
[5]Department of Pathobiology and Population Sciences, Royal Veterinary College London, London, UK

Abstract

Accurately estimating the prevalence and transmissibility of an infectious disease is a critical part of genetic infectious disease epidemiology. However, generating accurate estimates of these quantities, informed by both time series and sequencing data, is challenging. Birth-death processes and coalescent-based models are popular methods for modelling the transmission of infectious diseases, but they struggle with estimating the prevalence of infection.

We extended our approximation of the likelihood for a point process of viral genomes and time series of case counts so it can estimate historical prevalence, and we implemented this in a BEAST2 package called Timtam. In a simulation study the approximation recovered the parameters from simulated data, even when we aggregated the point process data into a time series of daily case counts.

To demonstrate how Timtam can be applied to real datasets, we estimated the reproduction number and the prevalence of infection through time during the SARS-CoV-2 outbreak onboard the Diamond Princess cruise ship using a time series of confirmed cases and sequence data. We found a greater prevalence than previously estimated and comment on how differences in the algorithms used could explain this.

Introduction

In the field of genetic infectious disease epidemiology, there are two perennial questions: "how many people are infected?" (i.e., what is the prevalence?) and "how transmissible is this pathogen?" (i.e., what is the reproduction number?) Prevalence of infection is the number of individuals currently infected and the reproduction number is the expected number of secondary infections per infectious individual.

Birth-death processes are a popular family of methods for modelling the transmission of infectious diseases, because they capture the mechanism of the process and are amenable to analysis. In the birth-death process, *births* represent new infections and *deaths* the end of an infectious period. In Bayesian phylogenetics, the birth-death process enters the analysis as a prior model for the reconstructed phylogeny (the so-called *tree prior*.) Kendall, 1948 demonstrated how to use generating functions to describe birth-death processes when modelling infectious disease. Later, Nee et al., 1994 connected the process to the number of observed species in a phylogeny, and Stadler, 2010; Stadler et al., 2012 demonstrated how this can be applied when analysing pathogen genomes.

Coalescent-based models offer a computationally convenient alternative to birth-death models and are also used to analyse viral genomes (Volz et al., 2013). However, the assumptions required to justify their use may be questionable in this setting, and the lack of an explicit sampling model can complicate matters (although there are exceptions, e.g., Parag et al., 2020). It is challenging to estimate absolute population sizes with either birth-death or coalescent based models (again, there are exceptions, e.g., Kühnert et al., 2014). To address the lack of an explicit finite susceptible population, parameters — particularly the birth rate — are modelled as piece-wise constant functions (Stadler et al., 2013). I.e., to adjust for the depletion of the susceptible pool, the model allows the birth rate to decline over time. While mathematically and computationally convenient, these nonparametric estimates make it difficult to distinguish if a decline in apparent birth rate is due to depletion of the susceptible pool or changes in behaviour, e.g., due to non-pharmaceutical interventions.

Efficiently incorporating unsequenced case data into phylodynamic analyses is a long-standing challenge in the field (Vaughan et al., 2019). Typically, only a small number of cases are sequenced[1] and phylody-

NOTE: This preprint reports new research that has not been certified by peer review and should not be used to guide clinical practice.

[1]No country with a sizeable COVID-19 outbreak sequenced > 20% of reported cases and most sequenced < 5%. Among low and middle income countries this number is often < 1%. Note that these numbers are for *reported* cases and not the true number of infections (Brito et al., 2022).

medRxiv preprint doi: https://doi.org/10.1101/2023.08.03.23293620; this version posted August 8, 2023. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY-ND 4.0 International license.

namic methods rely on this small subset of cases to estimate epidemiological dynamics, taking the stance that genomic sequences are often more informative than case reports alone and facilitate reconstructing the transmission between cases. Nevertheless, the vast amount of unsequenced case data is also informative and can help to refine estimates. The calculations needed to simultaneously analyse both sources of data in an integrated framework are well known (see for example Manceau et al., 2020), however existing ways to compute them are computationally intractable for all but the smallest outbreaks (Andréoletti et al., 2022).

In Zarebski et al., 2022, we described an efficient and accurate way to approximate the likelihood of a point process of viral genomes and a time series of case counts, which we call Timtam. While this resolved a long-standing challenge for the field, i.e., how to efficiently reconcile genetic and classical epidemiological data, there were substantial limitations: a.) while efficient, it is a complicated algorithm lacking a convenient implementation, limiting reuse and making it inaccessible to most potential users; and b.) it only estimated the present-time prevalence, not changes in the prevalence through time. We resolve these outstanding questions in this manuscript. We resolve the first limitation with a BEAST2 package (Bouckaert et al., 2019), called Timtam which is available on CBAN and can be installed via BEAUti. We resolve the second limitation with an extension to the algorithm, which enables conditioning on historical prevalence.

We carried out a simulation study to demonstrate that the methodology leads to well-calibrated estimates, i.e., that approximately 95% of the 95% credible intervals contain the true parameter value from the simulation. We demonstrate the "real-world" use of this package by recreating an analysis by Andréoletti et al., 2022 of SARS-CoV-2 data from an outbreak on the Diamond Princess cruise ship.

Methods

Figure 1A provides an example *transmission* tree, a complete description of who-infected-whom along with the timing of these events, and the surveillance of this process. Sequenced cases appear as filled circles in the figure, unobserved cases end the lineage without a circle, and empty circles indicate scheduled observation of cases without sequencing which happen at three times indicated with dashed lines. Figure 1B depicts the corresponding *reconstructed* tree and the time series of case counts. The reconstructed tree is the subtree of the transmission tree that results from pruning away any leaves not corresponding to a sequenced sample. Unlike the transmission tree, the reconstructed tree is not oriented, consequently it does not specify which of the two descendant lineages of an internal node corresponds to the infector. The leaves of the transmission tree corresponding to unsequenced samples form a separate, but not statistically independent, time series of confirmed cases. Figure 1C depicts the corresponding plots of the prevalence of infection through time (grey line) and the *lineages through time* (LTT) plot for the reconstructed tree (dashed line). The *lineages through time* (LTT) plot describes the number of lineages in the reconstructed tree as a function of time. Typically, the value of the LTT plot will be less than the prevalence of infection[2]. In Figure 1C, a single estimate of the prevalence appears near the present as a star.

We refer to the lineages in the transmission tree that are not in the reconstructed tree as the *hidden* lineages because they are not visible in the raw data. The sum of the number of hidden lineages and the LTT of the reconstructed tree is equal to the total prevalence of infection. We denote by k_t the value of the LTT at time t and by H_t the number of hidden lineages.

In the terminology used by Zarebski et al., 2022 the reconstructed tree consists of sequenced unscheduled data, and the time series of cases is representative of unsequenced scheduled data. We may consider arbitrary combinations of un/sequenced and un/scheduled data, but here we focus on datasets that consist of sequenced unscheduled data and unsequenced scheduled data (i.e. time-stamped sequences and a time-series of cases), since this aligns closest to typical epidemiological datasets.

In an epidemiological setting, we are often interested in the prevalence of infection and $\mathcal{R}_{\mathrm{e}}(t)$, because these quantities are of critical importance when assessing the threat posed by an outbreak of infectious disease. Bayesian phylodynamic methods provide a coherent solution with clear quantification of uncertainty; unfortunately, this usually requires us to evaluate the joint posterior distribution of the model parameters and the reconstructed tree (up to an unknown normalisation constant if we are using MCMC to generate posterior samples), given time-stamped viral genomes and a time series of confirmed cases. To do this, we need to evaluate the log-likelihood function in a computationally efficient way. Alternative simulation based methods exist, but tend to be far more computationally expensive which can reduce the utility of the resulting estimates.

The data of our model consists of $\mathcal{D}_{\mathrm{MSA}}$ and $\mathcal{D}_{\mathrm{cases}}$, where $\mathcal{D}_{\mathrm{MSA}}$ is the multiple sequence alignment (MSA) containing the pathogen genomic data, and $\mathcal{D}_{\mathrm{cases}}$ is the observation of confirmed cases which do not have an associated pathogen genome.

[2] If every infected individual has their pathogen's genome sequenced, or there is incomplete lineage sorting it is possible for the value of the LTT to exceed the prevalence.

medRxiv preprint doi: https://doi.org/10.1101/2023.08.03.23293620; this version posted August 8, 2023. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity.
It is made available under a CC-BY-ND 4.0 International license .

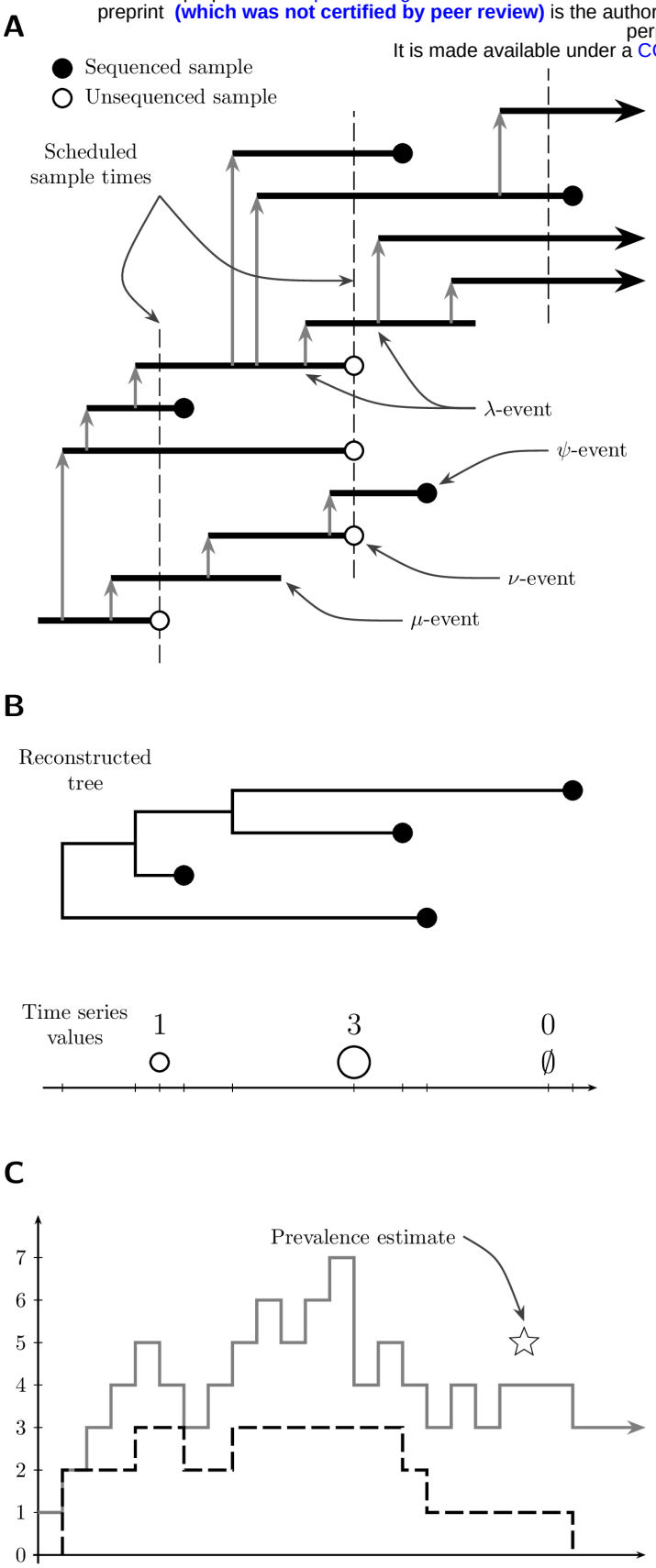

Figure 1: Transmission is viewed as a sequence of events, the observations are processed forward in time to approximate the joint likelihood. Panel **A** demonstrates a *transmission* tree with intervals of time an individual was infected indicated by horizontal lines and the vertical grey arrows indicating transmission. There are three scheduled unsequenced samples. Panel **B** shows the corresponding *reconstructed* tree and time series of confirmed cases at the scheduled observations, in the third sample, no cases where observed. Panel **C** shows the prevalence of infection (grey line) and the LTT (black dashed) along with a single estimate of the prevalence (the star).

The parameters of this process partition into four groups:

- $\mathcal{H}$, the number of hidden lineages at specified points in time (which we use to estimate the prevalence of infection),
- $\mathcal{T}$, the reconstructed tree describing the ancestral relationships between the sequences in $\mathcal{D}_{\text{MSA}}$,
- θ_{evo} the parameters of the evolutionary model, describing how genome sequences change over time (e.g., the clock rate and relative nucleotide substitution rates),
- and θ_{epi}, the parameters of the epidemiological model describing how the outbreak/epidemic spreads over time and how we observe it.

Using the terminology of birth-death processes, θ_{epi} contains the birth rate λ and the death rate μ along with the sequenced sampling rate ψ, the unsequenced sampling rate ω, probability of observation in a scheduled sequenced sample ρ and the probability of observation in a scheduled unsequenced sample ν. Examples of these events are shown in Figure 1A. Throughout this manuscript, we treat these parameters as piecewise constant functions with known change times.

We target the posterior distribution, $f(\mathcal{H}, \mathcal{T}, \theta_{\text{epi}}, \theta_{\text{evo}} \mid \mathcal{D}_{\text{MSA}}, \mathcal{D}_{\text{cases}})$, which we express in terms of simpler components with the factorisation in Equation (1) below.

The likelihood of the sequence data given the reconstructed tree and genomic parameters $f(\mathcal{D}_{\text{MSA}} \mid \mathcal{T}, \theta_{\text{evo}})$, which appears in Equation (1), is sometimes called the *phylogenetic likelihood*. This function is well-known and can be efficiently calculated with Felsenstein's pruning algorithm (Felsenstein, 1981). The likelihood of the time series of cases, reconstructed tree and prevalence, given the epidemiological parameters $f(\mathcal{D}_{\text{cases}}, \mathcal{T}, \mathcal{H} \mid \theta_{\text{epi}})$, which previously has been called the tree prior, might now more accurately be called the *time series likelihood*. Here we make the standard simplifying assumption that the genomic sequences are evolving neutrally, i.e. there is no dependence between the tree structure and the evolutionary process. This means that the phylogenetic likelihood is independent of $\mathcal{D}_{\text{cases}}, \mathcal{H}$ and θ_{epi} and that θ_{epi} and θ_{evo} have independent prior distributions.

We summarise the LTT and the time series of unsequenced observations as a sequence of labelled events, each with an associated time in order to evaluate the time series likelihood $f(\mathcal{D}_{\text{cases}}, \mathcal{T}, \mathcal{H} \mid \theta_{\text{epi}})$. There are four types of events we consider:

1. births, corresponding to the internal nodes of the reconstructed tree;

medRxiv preprint doi: https://doi.org/10.1101/2023.08.03.23293620; this version posted August 8, 2023. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity.
It is made available under a CC-BY-ND 4.0 International license .

$$
\begin{aligned}
f(\mathcal{H},\mathcal{T},\theta_{\text{epi}},\theta_{\text{evo}} \mid \mathcal{D}_{\text{MSA}},\mathcal{D}_{\text{cases}}) &= \frac{f(\mathcal{H},\mathcal{T},\theta_{\text{epi}},\theta_{\text{evo}},\mathcal{D}_{\text{MSA}},\mathcal{D}_{\text{cases}})}{f(\mathcal{D}_{\text{MSA}},\mathcal{D}_{\text{cases}})} \\
&= \frac{f(\mathcal{D}_{\text{MSA}} \mid \mathcal{H},\mathcal{T},\theta_{\text{epi}},\theta_{\text{evo}},\mathcal{D}_{\text{cases}})f(\mathcal{H},\mathcal{T},\theta_{\text{epi}},\theta_{\text{evo}},\mathcal{D}_{\text{cases}})}{f(\mathcal{D}_{\text{MSA}},\mathcal{D}_{\text{cases}})} \\
&= \frac{f(\mathcal{D}_{\text{MSA}} \mid \mathcal{T},\theta_{\text{evo}})f(\mathcal{H},\mathcal{T},\theta_{\text{epi}},\theta_{\text{evo}},\mathcal{D}_{\text{cases}})}{f(\mathcal{D}_{\text{MSA}},\mathcal{D}_{\text{cases}})} \\
&= \frac{f(\mathcal{D}_{\text{MSA}} \mid \mathcal{T},\theta_{\text{evo}})f(\mathcal{D}_{\text{cases}},\mathcal{T},\mathcal{H} \mid \theta_{\text{epi}},\theta_{\text{evo}})f(\theta_{\text{epi}},\theta_{\text{evo}})}{f(\mathcal{D}_{\text{MSA}},\mathcal{D}_{\text{cases}})} \\
&= \frac{f(\mathcal{D}_{\text{MSA}} \mid \mathcal{T},\theta_{\text{evo}})f(\mathcal{D}_{\text{cases}},\mathcal{T},\mathcal{H} \mid \theta_{\text{epi}})f(\theta_{\text{epi}},\theta_{\text{evo}})}{f(\mathcal{D}_{\text{MSA}},\mathcal{D}_{\text{cases}})} \\
&= \frac{\underbrace{f(\mathcal{D}_{\text{MSA}} \mid \mathcal{T},\theta_{\text{evo}})}_{\text{phylogenetic likelihood}}\ \underbrace{f(\mathcal{D}_{\text{cases}},\mathcal{T},\mathcal{H} \mid \theta_{\text{epi}})}_{\text{tree prior/time series likelihood}}\ f(\theta_{\text{epi}})f(\theta_{\text{evo}})}{f(\mathcal{D}_{\text{MSA}},\mathcal{D}_{\text{cases}})}
\end{aligned}
\tag{1}
$$

2. unscheduled sequenced samples, corresponding to the leaves of the reconstructed tree;

3. scheduled unsequenced samples, corresponding to the elements of the time series of cases;

4. and pseudo-observations of the number of hidden lineages at specified times.

These events are not the same as the data; events 1 and 4 are parameters. Events 3 and 4 have a value associated with them: for a scheduled unsequenced datum this is the value of the time series, and for the pseudo-observations this is the number of hidden lineages. We denote the event observed at time t_j by $\mathcal{E}_{t_j}$, and the sequence of events that occur up until time t (inclusive) by $\mathcal{E}_{\geq t}$. We use K_j to indicate the value of the LTT of $\mathcal{T}$ at time t_j. In the following we order events using a backward time formulation, with $t_0 = 0$ corresponding to the present (or the most recent sequenced sample in our data) and events further in the past having a larger time: $t_0 > t_1 > \cdots > t_N$. A consequence of specifying time in this way is that events occurring after the time of the last sequenced sample have negative times.

Expressions of the form $f\left(\mathcal{E}_{t_j} \mid \mathcal{E}_{\geq t_{j-1}}\right)$ are simpler, so we will consider the following factorisation:

$$
f\left(\mathcal{E}_{\geq t_N}\right) = \prod_{j=1}^{N} \underbrace{f\left(\mathcal{E}_{t_j} \mid \mathcal{E}_{\leq t_{j-1}}\right)}_{=c_j l_j}. \tag{2}
$$

The factors in the product in Equation (2) are further divided into two parts: c_j, which is the likelihood of the interval of time between t_{j-1} and t_j during which we observed nothing, and l_j the likelihood of the event that we observed at time t_j. We start by considering c_j. For $t_{j-1} > t > t_j$ let $M_t^{(i)} = \Pr\left(H_t = i \mid \mathcal{E}_{\geq t_{j-1}}\right)$, for $i \geq 0$ i.e., the joint distribution of the number of hidden lineages and not having observed any events since t_{j-1}. Evaluated at $t = t_j$ this becomes $M_t^{(i)} = \Pr\left(H_t = i, \mathcal{E}_{t_j} \mid \mathcal{E}_{\geq t_{j-1}}\right)$, i.e., the joint distribution of the number of hidden lineages and the event observed at time t_j.

We are considering the time during which we know there are no observed events, i.e., over an infinitesimal time step, δt, inside the interval $t_{j-1} > t > t_j$. In this case, $M_t^{(i)}$ satisfies the following equation (up to leading order):

$$
\begin{aligned}
M_{t-\delta t}^{(i)} = &\underbrace{(1-\gamma(K_{j-1}+i)\delta t)}_{\text{no event}} M_t^{(i)} + \\
&\underbrace{\lambda(2K_{j-1}+i-1)\delta t}_{\text{unobserved birth}} \mathbb{I}_{i>0} M_t^{(i-1)} + \\
&\underbrace{\mu(i+1)\delta t}_{\text{death}} M_t^{(i+1)},
\end{aligned}
$$

where $\mathbb{I}_x$ is the indicator random variable for event x. The factor of 2 in the term corresponding to births appears because the birth event creates two lineages (moving forward) and there are two ways to select one of them to be a hidden lineage and the other to continue the reconstructed tree.

Re-arranging the terms of these equations and taking the limit as the time step vanishes, we retrieve the master equations for this distribution, i.e., the system of differential equations that describe how it changes across the interval $t_{j-1} > t > t_j$:

$$
\begin{aligned}
\frac{dM_t^{(i)}}{dt} = &-\gamma(K_{j-1}+i)M_t^{(i)} + \\
&\lambda(2K_{j-1}+i-1)\mathbb{I}_{i>0}M_t^{(i-1)} + \\
&\mu(i+1)M_t^{(i+1)}.
\end{aligned}
\tag{3}
$$

Let $M_t(z)$ be the generating function for this system of differential equations: $M_t(z) = \sum_{i=0}^{\infty} M_t^{(i)} z^i$. We can

medRxiv preprint doi: https://doi.org/10.1101/2023.08.03.23293620; this version posted August 8, 2023. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity.
It is made available under a CC-BY-ND 4.0 International license.

write the system in Equation (3) as the following PDE which has the generating function $M_t(z)$ as its solution.

$$\partial_z M_t(z) = (\mu - \gamma z + \lambda z^2)\partial_t M_t(z) + K_{j-1}(2\lambda z - \gamma) M_t(z). \tag{4}$$

Manceau et al., 2020 solved Equation (4) in terms of results from Stadler, 2010. This partial differential equation allows us to update $M_t(z)$ across the intervals where there were no observed events: (t_{j-1}, t_j).

Given the generating function across each interval we can evaluate the c_j used in Equation (2). These come from the observation that $c_j = M_{t_j^+}(1^-)$. Note that we have to take the limit as time decreases to t_j because we are working with backwards time and there is a discontinuity.

The form of $M_{t_j}(z)$ depends on the limit $M_{t_j^+}(z)$ and the event observed at t_j. To simplify the description below, let $M_{t_j}^+(z) := \lim_{x \to t_j^+} M_x(z)$ which is the limiting value of the generating function before the observation. How we transform the generating function depends on $\mathcal{E}_{t_j}$. The expressions for l_j and $M_{t_j}(z)$ are as follows:

- for λ events $l_j := \lambda$ and $M_{t_j}(z) = M_{t_j}^+(z)/M_{t_j}^+(1^-)$,
- for ψ events $l_j := \psi$ and $M_{t_j}(z) = M_{t_j}^+(z)/M_{t_j}^+(1^-)$,
- for ω events, $l_j := \omega \frac{d}{dz}[M_{t_j}^+(z)]\Big|_{z=1}$ and $M_{t_j}(z) = \frac{\omega}{l_j}\frac{d}{dz}[M_{t_j}^+(z)]$,
- for ρ events when ΔK_j individuals were sampled and K_j lineages in the reconstructed tree just after the event, $l_j := \rho^{\Delta K_j}(1-\rho)^{K_j} M_{t_j}^+(1-\rho)$ and $M_{t_j}(z) = \frac{\rho^{\Delta K_j}(1-\rho)^{K_j}}{l_j} M_{t_j}^+((1-\rho)z)$,
- and for ν events when ΔH_j cases were observed, $l_j := (1-\nu)^{K_j}\nu^{\Delta H_j}\frac{d^{\Delta H_j}}{dz^{\Delta H_j}}[M_{t_j}^+(z)]\Big|_{z=1-\nu}$ and $M_{t_j}(z) = \frac{(1-\nu)^{K_j}\nu^{\Delta H_j}}{l_j}\frac{d^{\Delta H_j}}{dz'^{\Delta H_j}}[M_{t_j}^+(z')]\Big|_{z'=(1-\nu)z}$.

When co-estimating the prevalence, there is an additional event corresponding to a pseudo-observation of the number of hidden lineages. When we condition on $H_{t_j} = H_j$, then $l_j := M_{t_j}^{(H_j)+}(1^-)$ (which is the coefficient of z^{H_j} in $M_{t_j}^+(z)$ in the limit as $z \to 1^-$) and $M_{t_j}(z) = z^{H_j}$. Note that while the $\mathcal{H}_j$ are parameters of this model, we still include the corresponding l_j because we are also evaluating their prior distribution under the birth-death process.

Timtam

We can compute the c_j analytically, however, lack a closed form for the l_j. In Zarebski et al., 2022, we describe the time-series integration method through approximation of moments (Timtam). Timtam matches moments to approximates the relevant distribution with a negative binomial distribution. The generating function of the negative binomial distribution allows us to efficiently and accurately approximate the l_j and hence the likelihood of the model.

The effective reproduction number

Keeling et al., 2011 describes it as "one of the most critical epidemiological parameters", the reproduction number describes the average number of secondary infections. There are multiple ways to make this definition precise. We choose to define the effective reproduction number $\mathcal{R}_e(t)$ as the expected number of secondary infections generated by a newly infected individual from time t onward.

Without scheduled sampling, the value of $\mathcal{R}_e$ is simple: $\lambda/(\mu+\psi+\omega)$. Including scheduled sampling complicates matters because it combines continuous and discrete sampling. We derive a closed form expression for $\mathcal{R}_e$ in the Supplementary Information. However the result is unwieldy so we will instead make use of a simple approximation.

An approximation to $\mathcal{R}_e$ for scheduled data

Consider the case of unscheduled sequenced and scheduled unsequenced samples at regular intervals of duration Δ_t. From the perspective of an infectious individual, given they are removed during a scheduled sample, the number of intervals until this occurs, which we denote W, has a geometric distribution with probability ν. Given the scheduled samples occur at regular intervals of duration Δ_t, the wait time is approximately $\Delta_t(W+1/2)$. Provided Δ_t and ν are small, this distribution will be similar to an exponential distribution. The rate of an exponential distribution with the same mean is $2\nu/(2\Delta_t - \nu\Delta_t)$. If we consider a process which has unscheduled unsequenced sampling at rate $\tilde{\omega} = 2\nu/(2\Delta_t - \nu\Delta_t)$ and no scheduled unsequenced sampling, $\mathcal{R}_e = \lambda/(\mu+\psi+\tilde{\omega})$.

Figure 2 shows the effective reproduction number calculated using both the recursive method described in the Supplementary Information and the approximation that results from expressing the scheduled sampling as a rate $\tilde{\omega}$. The values of $\mathcal{R}_e$ are greater for longer intervals between scheduled samples, Δ_t, because there is a longer duration during which the individual can infect others.

Model parameterizations

There are multiple ways to parameterize this process. We refer to the parameterization in terms of the rates λ, ψ, ω, and probabilities ρ, and ν as the *canonical parameterization*. We derive the approximate likelihood

medRxiv preprint doi: https://doi.org/10.1101/2023.08.03.23293620; this version posted August 8, 2023. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY-ND 4.0 International license.

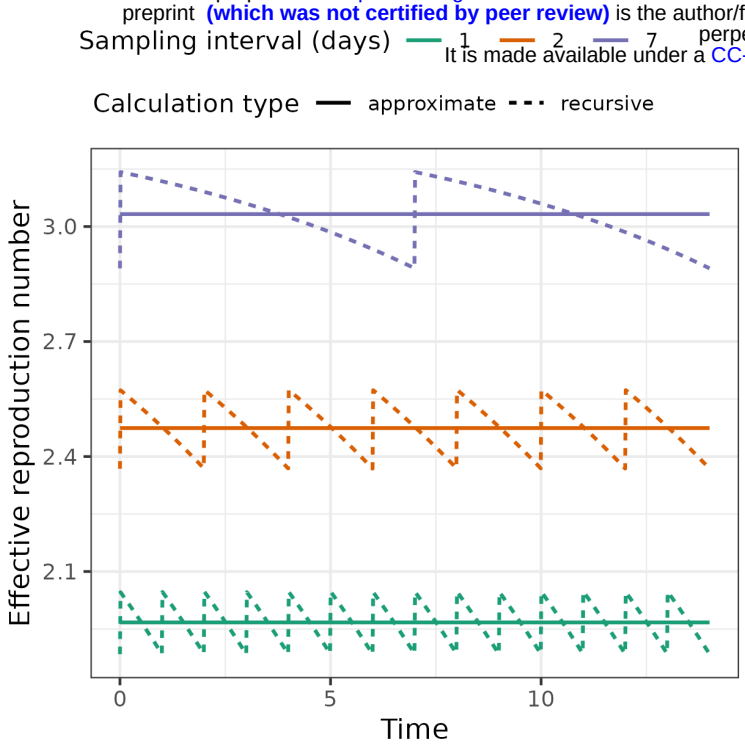

Figure 2: The approximation smooths out the saw-tooth value of the effective reproduction number in the case of scheduled samples. The parameters used for this figure are birth rate of 0.4, death rate of 0.1, sampling rate of 0.02 and a scheduled unsequenced sampling probability of 0.08 (at varying intervals). The solid lines indicate the values obtained with our approximation and the dashed lines indicate the true values accounting for scheduled sampling.

in terms of these parameters. In an epidemiological context, with a time series of confirmed cases and point process sequence data, we prefer to parameterize this in terms of the effective reproduction number, $\mathcal{R}_e$, the net removal rate, $\sigma = \mu + \psi + \tilde{\omega}$ (where $\tilde{\omega}$ is as described above), and the observed proportion of infections captured by the time series, $\tilde{\omega}/\sigma$, and the point process, ψ/σ. Given the focus on the use of time series data, we refer to this parameterization as the *time series parameterization*.

Note that when we use the time series parameterization in the SARS-CoV-2 analysis, we use the approximation of $\tilde{\omega}$ to simplify the model specification. This avoids the issue of having to adjust for future scheduled sampling in $\mathcal{R}_e(t)$.

Sampled ancestors

A natural extension to this model is the inclusion of sampled ancestors (Gavryushkina et al., 2014), which Manceau et al., 2020 and Andréoletti et al., 2022 have already considered. Including sampled ancestors involves a probability r an infected individual is removed upon (unscheduled) observation. Currently we assume all individuals are removed upon observation. We have not yet implemented this extension in Timtam, however we include the relevant expressions and the explanation of some additional details in the Supplementary Information and leave the implementation as an exercise for the reader.

Calibration study

To assess the calibration of Timtam and the validity of our approximation of $\mathcal{R}_e$, we carried out a simulation study. We simulated 100 epidemics for 56 days using `remaster`[3] with the birth rate changing on day 42, i.e., boom-bust dynamics. The prevalence of infection in each of the simulated epidemics is shown in Fig. S1. There is a substantial amount of variability in the prevalence across the simulations, but the boom-bust dynamics can be seen in the average over the simulations.

From each simulation we constructed two datasets: one with unsequenced samples treated as a point process, and a second with these samples aggregated into a time series of daily case counts. The full parameter list is given in Table 1. The parameters used are similar to those used in Zarebski et al., 2022 with an extension for the change in birth rate; we based them on the early dynamics of SARS-CoV-2 in Australia. The code implementing this simulation and the subsequent inference is available at `https://github.com/aezarebski/timtam-calibration-study`.

We estimated the parameters of the model for each simulated data set and compared the estimated values to the true values used in the simulation. Figure 3 show the estimates of prevalence and reproduction numbers across the simulations ordered by the final prevalence in the simulation in the case where the unsequenced data is modelled as a point process. Figure 4 shows the corresponding results when the unsequenced data have been aggregated into a daily time series of counts.

Comparing the simulations with a small final prevalence to those with a large final prevalence we see that, as expected, for simulations with a larger prevalence the estimates of the reproduction number have less uncertainty and are less biased. We attribute this to the strong correlation between the final prevalence and the total number of data points, (as shown in Fig. S2).

Table 1 contains a summary of the rate parameter estimates from the first set of simulations (i.e. the ones with point process data) and Table 2 contains the corresponding summary for the second set of simulations (i.e. with unsequenced samples aggregated into a time series). The credible intervals (CrIs) of both the reproduction number and the prevalence at the time of the last sequenced sample have a coverage that is consis-

[3] `remaster` (`https://github.com/tgvaughan/remaster`) is a re-write of the MASTER simulation package for BEAST2 from Vaughan et al., 2013

medRxiv preprint doi: https://doi.org/10.1101/2023.08.03.23293620; this version posted August 8, 2023. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity.
It is made available under a CC-BY-ND 4.0 International license .

tent with the desired value, suggesting the estimation method is well-calibrated.

The CrI is distinct from the confidence interval (a.k.a. CI). Due to the influence of the prior, we do not necessarily expect 95% of the posterior distributions to contain the true parameters, however, we would like it to be close to this. We performed a hypothesis test of the null hypothesis that 95% of the intervals contain the true parameter. Of course, the truth of this null depends upon the choice of prior distribution, nonetheless, we would like it to be difficult to falsify such a null hypothesis for plausible prior distributions.

In our hypothesis test, we expect 91–99 of the CrIs to contain the true parameter value (out of the total 100 replicates) under the null hypothesis that the CrIs contain the true value 95% of the time. When interpreting the hypothesis test of whether the intervals are well calibrated at 95%, we need to bear in mind that the sampling of the limits of this interval are sparse (by construction) so one would ideally want a far larger effective sample size (ESS) than the one we present. Kruschke, 2014, p. 184 suggests for reliable 95% limits an ESS ≥ 10000 is desirable. For each of our analyses the ESS was ≥ 200 for all variables.

Par	True	Median	Error	Bias	Width	Coverage
λ_1	0.185	0.186	0.116	0.004	0.531	94
λ_2	0.092	0.095	0.337	0.032	1.386	93
μ	0.0460	-	-	-	-	-
ψ	0.008	0.010	0.351	0.275	1.811	92
ω	0.046	0.052	0.248	0.140	1.209	96
$\mathcal{R}_e^1$	1.850	1.689	0.180	-0.087	0.677	94
$\mathcal{R}_e^2$	0.925	0.897	0.291	-0.030	1.151	95
H	-	-	0.360	-0.046	-	99

For each parameters (Par), the median over the 100 medians of the estimate, relative error, relative bias and the percentage of credible intervals containing the true value is provided.

Table 1: Posterior parameter estimates and accuracy in the 100 simulations. There are boom-bust dynamics, for the first 42 days of the simulation the birth rate is λ_1 after which it changes to λ_2 for the subsequent 14 days. The death rate is assumed known.

Par	True	Median	Error	Bias	Width	Coverage
λ_1	0.185	0.186	0.121	0.003	0.540	95
λ_2	0.092	0.094	0.337	0.018	1.399	93
μ	0.0460	-	-	-	-	-
ψ	0.008	0.010	0.344	0.267	1.852	90
$\tilde{\omega}$	0.046	0.053	0.265	0.143	1.211	97
$\mathcal{R}_e^1$	1.850	1.670	0.191	-0.097	0.678	89
$\mathcal{R}_e^2$	0.925	0.873	0.287	-0.057	1.158	97
H	0.000	-	0.367	-0.055	-	99

Table 2: Posterior parameter estimates and accuracy in the 100 simulations after we aggregated the unsequenced observations into daily counts and used the resulting time series as data.

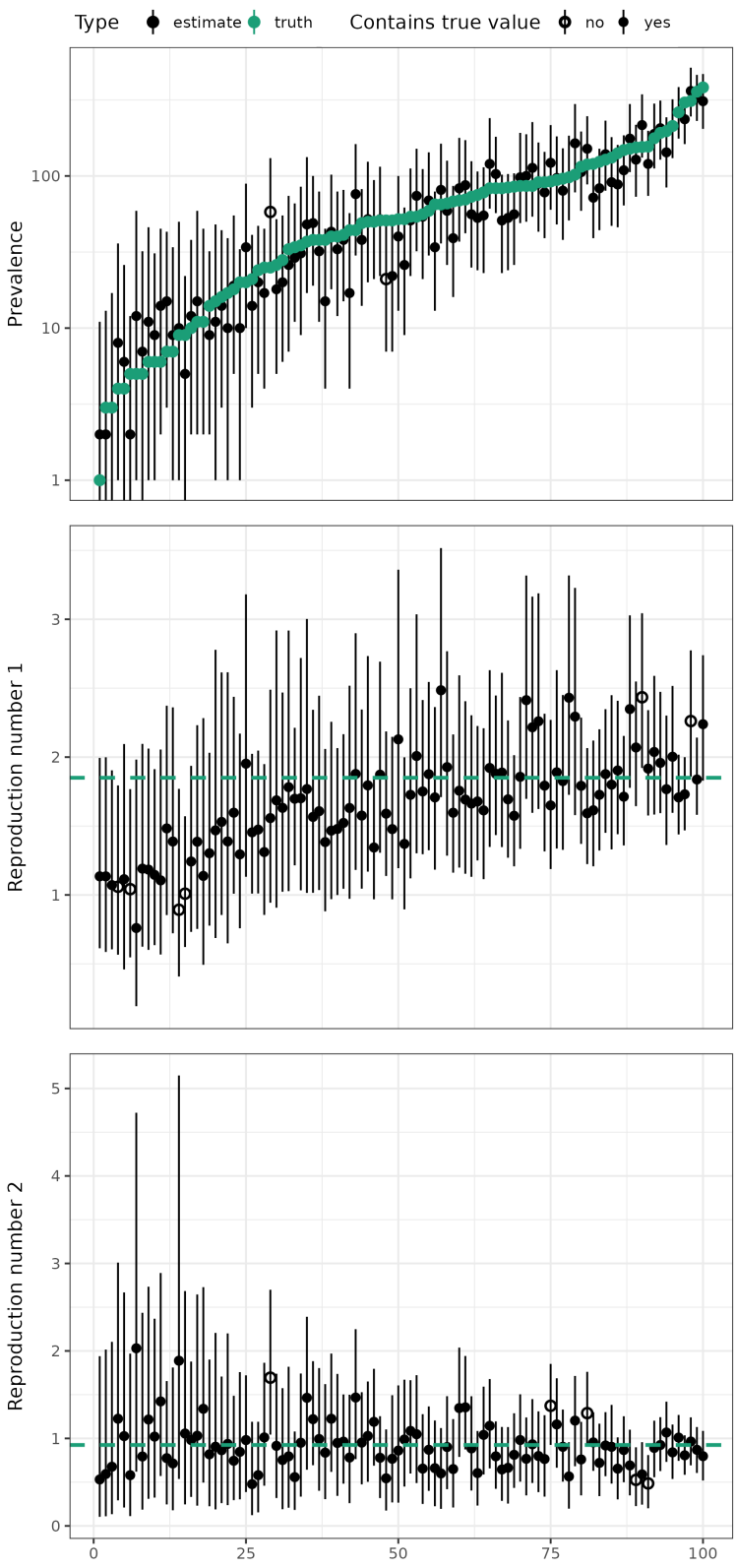

Figure 3: Parameter estimates converge to true values as the data set gets larger. The solid black lines display the credible intervals, and the filled or empty black points indicate the point estimates. The green points and the green dashed lines indicate the true values of the final prevalence and the reproduction number in the boom and bust portions of the simulation. We ordered the replicates by the final prevalence in each simulation.

medRxiv preprint doi: https://doi.org/10.1101/2023.08.03.23293620; this version posted August 8, 2023. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity.
It is made available under a CC-BY-ND 4.0 International license .

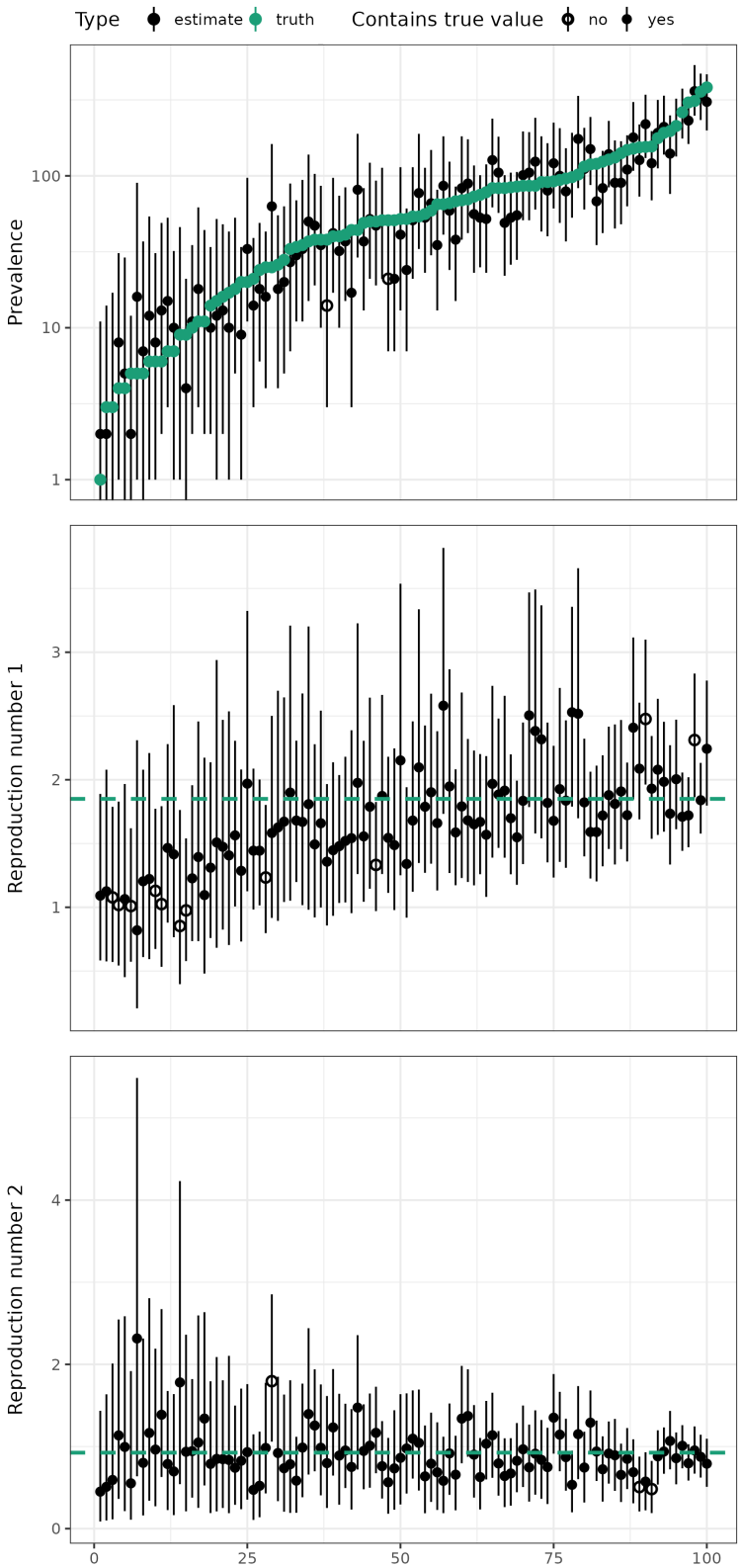

Figure 4: The estimated quantities and their true values from the simulation as shown in Figure 3 when the unsequenced observations were aggregated into a time series of daily case counts.

SARS-CoV-2 on the Diamond Princess cruise ship

We replicated the analysis of a SARS-CoV-2 outbreak onboard the Diamond Princess cruise ship reported by Andréoletti et al., 2022. This outbreak is particularly well-suited to analysis because it occurred on an isolated cruise ship (with 3711 people onboard) in a carefully monitored population with detailed accounts of isolation and testing measures. The outbreak appears to have originated from a single introduction of the virus (Sekizuka et al., 2020). Figure 5 displays the cases and sequencing effort across the duration of the quarantine. We obtained a time series of daily confirmed cases from Dong et al., 2020 to use as $\mathcal{D}_{\text{cases}}$ and an alignment of 70 pathogen genomes (Sekizuka et al., 2020) was used as $\mathcal{D}_{\text{MSA}}$. The accession numbers for the sequences are available in the Supplementary Information.

Model

We made minor adjustments to the model to better match standard epidemiological workflows for $\mathcal{R}_{\text{e}}$ estimation, as described in the Supplementary Information. Importantly, we modelled daily case counts of confirmed cases as scheduled samples (i.e., a time series) instead of unscheduled samples (i.e., a point process of occurrences.) Additionally, we put an explicit prior on the reproduction number. Table S1 lists the prior distributions used in the model. The XML files specifying the full analysis and post-processing are available from `https://github.com/azwaans/timtam-diamond-princess`.

Results

Figure 6 shows the estimates of the reproduction number through time along with the 95% credible intervals, calculated as the highest posterior density intervals (HPI). The estimates of the effective reproduction number are consistent with those from previous analyses of these data (Andréoletti et al., 2022, Figure 5). Our estimates differ from those of (Vaughan et al., 2020, Fig. S3). The discrepancy between the estimates from Vaughan et al., 2020 and ours may be due to the different data sets used: our analysis used both the time series of confirmed cases and pathogen genomes, while Vaughan et al., 2020's estimates are based on genomic data alone.

Figure 7 shows the estimates of the prevalence of infection and the 95% HPI credible intervals along with the corresponding values from Andréoletti et al., 2022, Figure 5. Our estimates suggest a larger prevalence of infection than the estimates from Andréoletti et al., 2022. Prevalence estimates from Vaughan et al., 2020

medRxiv preprint doi: https://doi.org/10.1101/2023.08.03.23293620; this version posted August 8, 2023. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity.
It is made available under a CC-BY-ND 4.0 International license .

are not included as they estimated the cumulative number of infections instead of the prevalence.

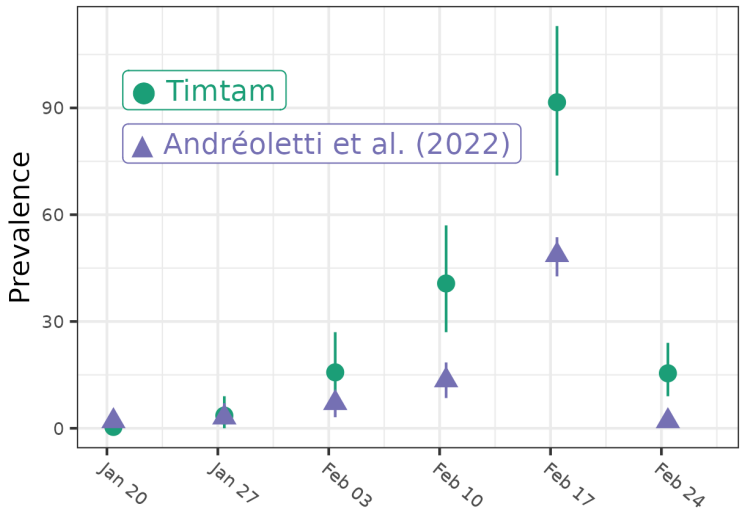

Figure 7: Estimates of the prevalence of infection and the 95% HPI credible intervals onboard the Diamond Princess. In addition to our estimates (shown in green) there are estimates from Andréoletti et al., 2022 (shown in purple).

Discussion

We implemented a tree prior that facilitates the co-estimation of the prevalence and the effective reproduction number, the resulting model can draw on both sequence data and a time series of confirmed cases (where the pathogen genome may not have been sequenced). The algorithm used to compute the (approximate) log-likelihood is fast, requiring a number of steps linear in the number of sequences and length of the time series (Zarebski et al., 2022). The implementation is available as a BEAST2 package and tutorials on the usage of the package are bundled with the source code: `https://github.com/aezarebski/timtam2`.

We performed a simulation study to demonstrate the method is well-calibrated, i.e., that approximately 95% of the 95% credible intervals do contain the true value. The simulation study also demonstrated the performance of the method does not degrade substantially when we aggregated the occurrence data into a time series.

We used the validated method to replicate an analysis carried out by Andréoletti et al., 2022 of an outbreak of SARS-CoV-2 onboard the Diamond Princess. Our estimates of the reproduction number (displayed in Figure 6) are consistent with the values from Andréoletti et al., 2022 and are similar to those from Vaughan et al., 2019. However, Vaughan et al., 2019 only used genomic data which may be the reason the $\mathcal{R}_e$ estimates differ slightly in this case.

Our prevalence estimates are greater than those from Andréoletti et al., 2022 (Figure 7). We attribute this difference to their algorithm having an upper limit of 40 on the number of hidden lineages, which was necessitated by the computational complexity of the numerical integration algorithm used to compute the likelihood. As such, their estimates should be interpreted as lower bounds on the prevalence and not absolute estimates. Timtam overcomes this limitation by approximating the number of hidden lineages with a negative binomial, making it applicable to real-world epidemic scenarios. Instead of estimating the prevalence through time, Vaughan et al., 2019 estimated the *cumulative* number of infections through time, which prevents us from comparing their estimates to ours.

To estimate historical prevalence, we extended the method previously developed in Zarebski et al., 2022. We model the prevalence as a parameter to estimate, this differs from several previous approaches where estimates of the prevalence come from intermediate steps in the likelihood calculation or from post-hoc simulation. Treating the prevalence as a bona fide parameter also means we could incorporate additional data concerning prevalence into the analysis. For example, if survey data on infection in a random sample from the population was available.

For the Diamond Princess data analysis we modelled the sequenced cases as a point process, consistent with previous analyses. Where multiple samples were available for a particular day, we uniformly spaced the sequenced samples across the day the samples were collected. A more nuanced analysis would have modelled these samples as scheduled sequenced samples, however this would make the resulting estimates harder to compare to previous results and complicate the interpretation.

Our implementation does not yet support the use of *sampled ancestors* (Gavryushkina et al., 2014). Extending the approximation to handle this case seems feasible, however there are substantial engineering challenges involved in implementing this on the BEAST2 platform. We include the expressions required for including sampled ancestors in the Supplementary Information.

Acknowledgements

We thank Dr. Timothy Vaughan for patiently answering many questions during the implementation of the Timtam package.

AEZ, BG and OGP are supported by The Oxford Martin Programme on Pandemic Genomics. AZ is supported by the European Research Council (ERC) under the European Union's Horizon 2020 research and innovation programme grant agreement no. 101001077.

medRxiv preprint doi: https://doi.org/10.1101/2023.08.03.23293620; this version posted August 8, 2023. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity.
It is made available under a CC-BY-ND 4.0 International license .

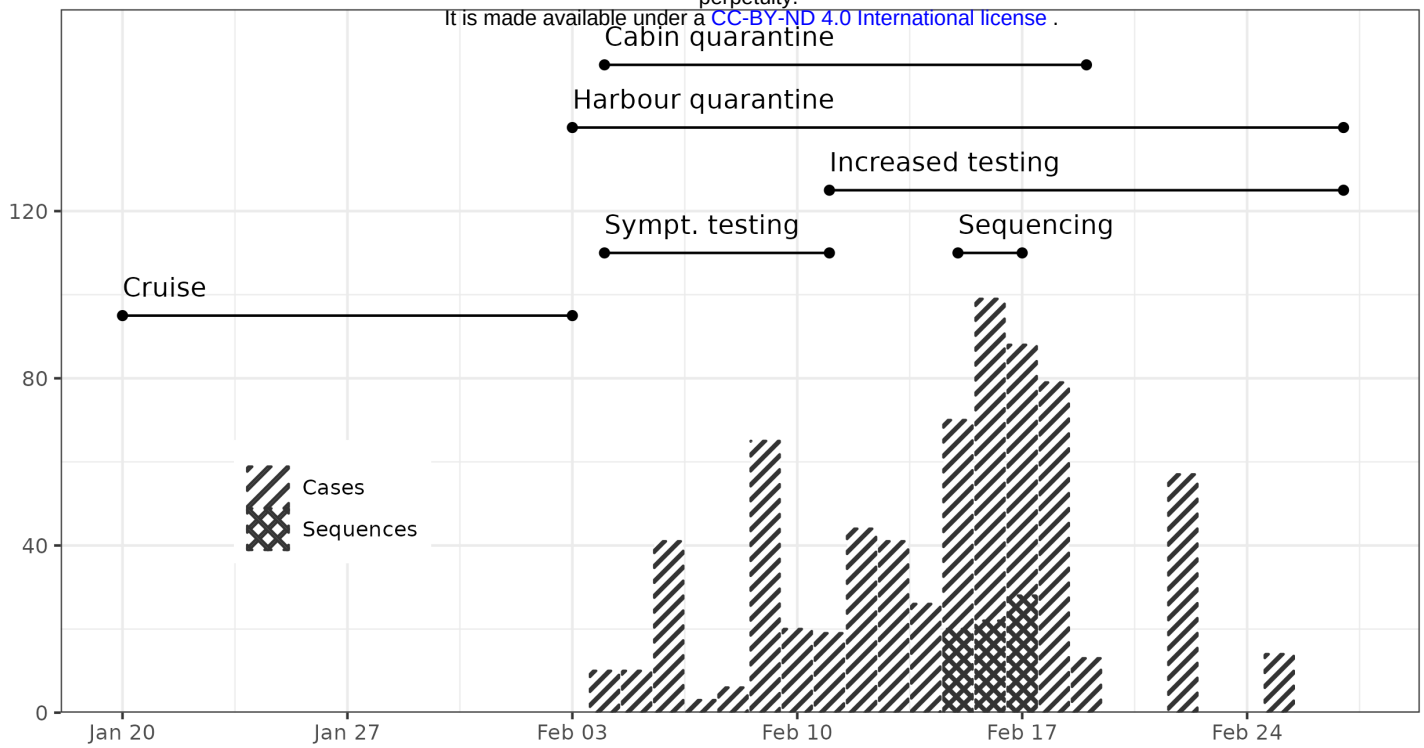

Figure 5: Sequences were collected across three days and testing varied throughout the quarantine. The stacked bar chart shows the daily number of confirmed cases and sequenced samples. We indicate timing of changes to surveillance and quarantine with lines at the top of the figure.

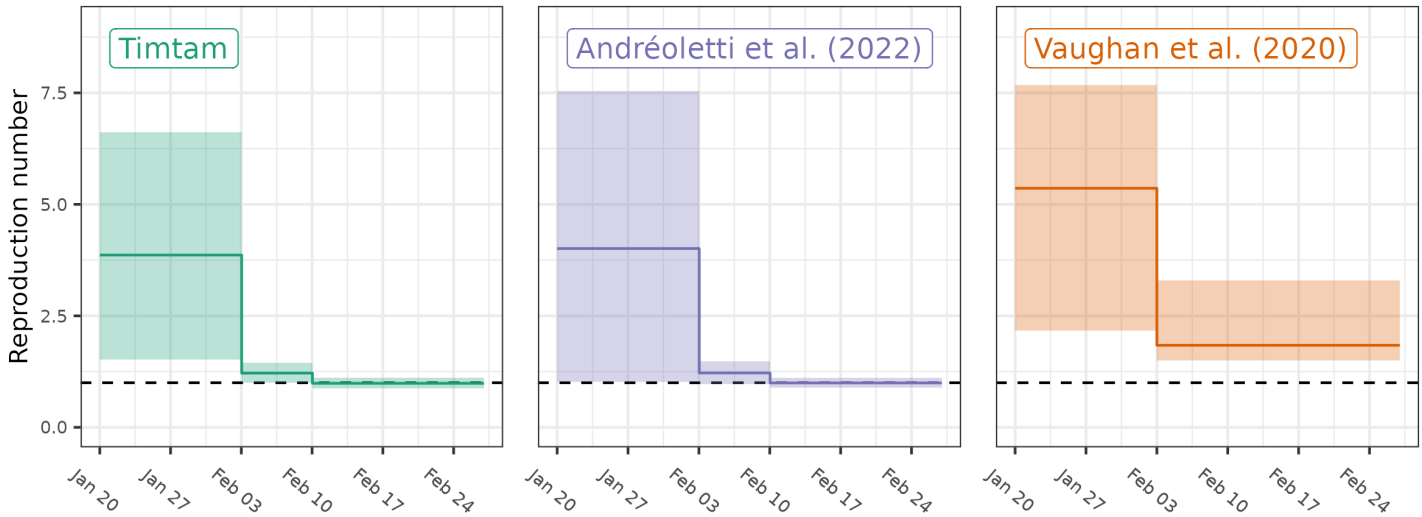

Figure 6: Estimates of the reproduction number and the 95% HPI credible intervals. In addition to our estimates (shown in green) there are estimates from Andréoletti et al., 2022 (shown in purple) and Vaughan et al., 2020 (shown in orange).

References

Andréoletti, Jérémy, Antoine Zwaans, Rachel C M Warnock, Gabriel Aguirre-Fernández, Joëlle Barido-Sottani, Ankit Gupta, Tanja Stadler, and Marc Manceau (May 2022). "The Occurrence Birth-Death Process for combined-evidence analysis in macroevolution and epidemiology". In: *Systematic Biology*. DOI: 10.1093/sysbio/syac037.

Bouckaert, Remco, Timothy G. Vaughan, Joëlle Barido-Sottani, Sebastián Duchêne, Mathieu Fourment, Alexandra Gavryushkina, Joseph Heled, Graham Jones, Denise Kühnert, Nicola De Maio, Michael Matschiner, Fábio K. Mendes, Nicola F. Müller, Huw A. Ogilvie, Louis du Plessis, Alex Popinga, Andrew Rambaut, David Rasmussen, Igor Siveroni, Marc A. Suchard, Chieh-Hsi Wu, Dong Xie, Chi Zhang, Tanja Stadler, and Alexei J. Drummond (Apr. 2019). "BEAST 2.5: An advanced software platform for Bayesian evolutionary analysis". In: *PLOS Computational Biology* 15.4, pp. 1–28. DOI: 10.1371/journal.pcbi.1006650.

medRxiv preprint doi: https://doi.org/10.1101/2023.08.03.23293620; this version posted August 8, 2023. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY-ND 4.0 International license.

Brito, Anderson F., Elizaveta Semenova, Gytis Dudas, Sarah C. Hasswell, Nikos Avershina, Pål Marstrand, Chunlei Wang, Kamilla Stalteri, Vladimir Roddy, Ana Pyl, Jonathan Wickenhagen, Bo Li, Jenna Tilloy, Kate Nicholay, Marc Taylor, Paul Moyo, Jacob Strong, Mary Fennema, Emma Barroso, Jaqueline Maes, Eduardo Sabino, Luciana Duraes, Vagner Sampaio, Michael Gamal, Valdinete Monteiro, Deborah Guerra, Nuno Fuzeto, Marcos Cunha, Rafael Nascimento, Victor Naveca, Marcilio Odilon, Vanessa Sales, Cecilia Arns, Pedro Souza, Vanessa Tadaiesky, Wagner Fabeni, Jussara Matos, Jaqueline Matos, Paola Gerber, Giovanna Souza, Nelson Siqueira, Mauro Buss, Jeferson Lima, Deivison Pinto, Hygor Sales, Renato Marques, Nuno Cardoso, Renan Penna, Marcio Manuli, Marcelo Cilião, Cintia Cassanjus, Tiagor Rodrigues, Rafaela Rocha, Marcio Miyahara, Sandra Sanabani, Mauricio Nogueira, Nanci Kallas, Edmilson Morais, Ester Sabino, Ronaldo Carvalho, Darren Cromer, Darren Martin, Ana Tanuri, Andrew Rambaut, Nuno Faria, Oliver Pybus, Ester C. Sabino, Luiz Carlos Junior Alcantara, Sunetra Gupta, Ana Bershteyn, Michael Gray, Matthew Hitchings, Lauro Vieira, Joseph Lewis, Marcus Vinicius Silva, Ana Tereza Ribeiro, Karthik Gangavarapu, Alexander Zarebski, Emilia Magnus, Kristian Andersen, Marcos Lemos, Gonzalo Yeo-Teh, Barney Tegally, Jason Hodcroft, Emma Wilkinson, Houriiyah Tegally, Eric Halpin, Karen Lanes, Philip Gerhard, Cassio Peterka, Ricardo Gil, Joice Sousa, Felipe Rabelo, Jayna Gonçalves, Lucianne Sampaio, Yokomizo, Daniel Park, Sarah Abdool, Ingrid Fritsch, Ricardo Fritsch, Marcia Lewandowski, Simone Martines, Cristiano Somasundaram, Rafael Manuli, Alex Zimmermann, Renato Santana, Guilherme Cruz, Alessandro Souza, Vinicius Duarte, Vanessa Barcelos, Karina Tho, Umberto Raboni, Rodrigo Spitz, Bruna Abrahão, Matheus Sabbaga, Barbara Silveira, Luciane Lazari, Paola Cruz, Edivania Almeida, Jean Baiocchi, Ricardo Giglio, Fernando Claro, Luciana Appolinario, Marcelo Luiz, Sandra Bispo, Marilda Pelegrino, Rafael Campagna, Daniel Marques, Victor Boscolo, Nelson Suzuki, Felipe Romero, Erick Gaburo, Carolina Carvalho, Lucimara Rogerio, Matheus Souza, Jaqueline Deccache, Edson Souza, Paola Duarte, Vanessa Morales, Darlan Sales, Luciana Vieira, Fernando Campanini, Alinne Costa, Julia Ferreira, Elaine Campelo, Thiago Moreira, Luciana Oliveira, Kelen Cavalcanti, Poliana Esashika, Maria Maia, Fernanda Morgan, Natalia Ferreira, Matheus Avelino, Josiane Fujiyama, Giuliana Silva, Lilian Sales, Beatriz Segalla, Luiz Lima, Renan Oliveira, Victor Souza, Rafael Evangelista, Gabriela Marqueze, Matheus Prado, Gabriel Abreu, Anderson Almeida, Flavia Giannini, Rafael Moreno, Matheus Bellato, Luiz Nogueira, Fernanda Rodrigues, Jamile Anselmo, Felipe Hoeltgebaum, Ivan Mitchell, Lewi Faria, Maria Mendrone, Rafael Cunha, Ricardo Sobhie, Alexander Morgan, Luiz Gaseta, Tiago Pereira, Anderson Brito, Elizaveta Semenova, Gytis Dudas, et al. (2022). "Global disparities in SARS-CoV-2 genomic surveillance". In: *Nature Communications* 13.1, p. 7003. DOI: 10.1038/s41467-022-33713-y.

Dong, Ensheng, Hongru Du, and Lauren Gardner (2020). "An interactive web-based dashboard to track COVID-19 in real time". In: *The Lancet Infectious Diseases* 20.5, pp. 533–534. DOI: 10.1016/S1473-3099(20)30120-1.

Felsenstein, Joseph (Nov. 1981). "Evolutionary trees from DNA sequences: A maximum likelihood approach". en. In: *Journal of Molecular Evolution* 17.6, pp. 368–376. DOI: 10.1007/BF01734359.

Gavryushkina, Alexandra, David Welch, Tanja Stadler, and Alexei J. Drummond (Dec. 2014). "Bayesian Inference of Sampled Ancestor Trees for Epidemiology and Fossil Calibration". In: *PLOS Computational Biology* 10.12, pp. 1–15. DOI: 10.1371/journal.pcbi.1003919.

Keeling, Matt J and Pejman Rohani (2011). *Modeling Infectious Diseases in Humans and Animals*. Princeton University Press. DOI: 10.2307/j.ctvcm4gk0.

Kendall, David G. (1948). "On the Generalized "Birth-and-Death" Process". In: *The Annals of Mathematical Statistics* 19.1, pp. 1–15. DOI: 10.1214/aoms/1177730285.

Kruschke, John (2014). *Doing Bayesian Data Analysis: A Tutorial with R, JAGS, and Stan*. 2nd ed. Elsevier Science & Technology.

Kühnert, Denise, Tanja Stadler, Timothy G. Vaughan, and Alexei J. Drummond (2014). "Simultaneous reconstruction of evolutionary history and epidemiological dynamics from viral sequences with the birth-death SIR model". In: *Journal of The Royal Society Interface* 11.94, p. 20131106. DOI: 10.1098/rsif.2013.1106.

Manceau, Marc, Ankit Gupta, Timothy Vaughan, and Tanja Stadler (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. DOI: 10.1016/j.jtbi.2020.110400.

Nee, Sean, Robert Mccredie May, and Paul H. Harvey (1994). "The reconstructed evolutionary process". In: *Philosophical Transactions of the Royal Society of London. Series B: Biological Sciences* 344.1309, pp. 305–311. DOI: 10.1098/rstb.1994.0068.

Parag, Kris V, Louis du Plessis, and Oliver G Pybus (Jan. 2020). "Jointly Inferring the Dynamics of Population Size and Sampling Intensity from Molecular Sequences". In: *Molecular Biology and Evolution* 37.8, pp. 2414–2429. DOI: 10.1093/molbev/msaa016.

Shiraishi, Kazuhiro, Kazuki Kizuka, Tsuyoshi Kentaro Itokawa, Tsutomu Kageyama, Shinji Saito, Ikuyo Takayama, Hideki Asanuma, Naganori Nao, Rina Tanaka, Masanori Hashino, Takuri Takahashi, Hajime Kamiya, Takuya Yamagishi, Kensaku Kakimoto, Motoi Suzuki, Hideki Hasegawa, Takaji Wakita, and Makoto Kuroda (2020). "Haplotype networks of SARS-CoV-2 infections in the Diamond Princess cruise ship outbreak". In: *Proceedings of the National Academy of Sciences* 117.33, pp. 20198–20201. DOI: 10.1073/pnas.2006824117.

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

Stadler, Tanja, Roger Kouyos, Viktor von Wyl, Sabine Yerly, Jürg Böni, Philippe Bürgisser, Thomas Klimkait, Beda Joos, Philip Rieder, Dong Xie, Huldrych F. Günthard, Alexei J. Drummond, and Sebastian Bonhoeffer (2012). "Estimating the Basic Reproductive Number from Viral Sequence Data". In: *Molecular Biology and Evolution* 29.1, pp. 347–357. DOI: 10.1093/molbev/msr217.

Stadler, Tanja, Denise Kühnert, Sebastian Bonhoeffer, and Alexei J. Drummond (2013). "Birth-death skyline plot reveals temporal changes of epidemic spread in HIV and hepatitis C virus (HCV)". In: *Proceedings of the National Academy of Sciences* 110.1, pp. 228–233. DOI: 10.1073/pnas.1207965110.

Vaughan, Timothy G. and Alexei J. Drummond (Mar. 2013). "A Stochastic Simulator of Birth-Death Master Equations with Application to Phylodynamics". In: *Molecular Biology and Evolution* 30.6, pp. 1480–1493. DOI: 10.1093/molbev/mst057.

Vaughan, Timothy G, Gabriel E Leventhal, David A Rasmussen, Alexei J Drummond, David Welch, and Tanja Stadler (May 2019). "Estimating Epidemic Incidence and Prevalence from Genomic Data". In: *Molecular Biology and Evolution* 36.8, pp. 1804–1816. DOI: 10.1093/molbev/msz106.

Vaughan, Timothy G., Jérémie Sciré, Sarah A. Nadeau, and Tanja Stadler (2020). "Estimates of outbreak-specific SARS-CoV-2 epidemiological parameters from genomic data". In: *medRxiv*. DOI: 10.1101/2020.09.12.20193284.

Volz, Erik M., Katia Koelle, and Trevor Bedford (Mar. 2013). "Viral Phylodynamics". In: *PLOS Computational Biology* 9.3, pp. 1–12. DOI: 10.1371/journal.pcbi.1002947.

Zarebski, Alexander Eugene, Louis du Plessis, Kris Varun Parag, and Oliver George Pybus (Feb. 2022). "A computationally tractable birth-death model that combines phylogenetic and epidemiological data". In: *PLOS Computational Biology*

medRxiv preprint doi: https://doi.org/10.1101/2023.08.03.23293620; this version posted August 8, 2023. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity.
It is made available under a CC-BY-ND 4.0 International license .

18.2, pp. 1–22. DOI: 10.1371/journal.pcbi.1009805.