JOURNAL OF THE ROYAL SOCIETY Interface

rsif.royalsocietypublishing.org

Research

Cite this article: Volz EM, Frost SDW. 2014 Sampling through time and phylodynamic inference with coalescent and birth–death models. *J. R. Soc. Interface* **11**: 20140945. http://dx.doi.org/10.1098/rsif.2014.0945

Received: 22 August 2014
Accepted: 7 October 2014

Subject Areas:
biometrics, computational biology

Keywords:
phylodynamics, coalescent, infectious diseases

Author for correspondence:
Erik M. Volz
e-mail: e.volz@imperial.ac.uk

Electronic supplementary material is available at http://dx.doi.org/10.1098/rsif.2014.0945 or via http://rsif.royalsocietypublishing.org.

Royal Society Publishing

Sampling through time and phylodynamic inference with coalescent and birth–death models

Erik M. Volz[1] and Simon D. W. Frost[2]

[1]Department of Infectious Disease Epidemiology, Imperial College London, London, UK
[2]Department of Veterinary Medicine, University of Cambridge, Cambridge, UK

Many population genetic models have been developed for the purpose of inferring population size and growth rates from random samples of genetic data. We examine two popular approaches to this problem, the coalescent and the birth–death-sampling model (BDM), in the context of estimating population size and birth rates in a population growing exponentially according to the birth–death branching process. For sequences sampled at a single time, we found the coalescent and the BDM gave virtually indistinguishable results in terms of the growth rates and fraction of the population sampled, even when sampling from a small population. For sequences sampled at multiple time points, we find that the birth–death model estimators are subject to large bias if the sampling process is misspecified. Since BDMs incorporate a model of the sampling process, we show how much of the statistical power of BDMs arises from the sequence of sample times and not from the genealogical tree. This motivates the development of a new coalescent estimator, which is augmented with a model of the known sampling process and is potentially more precise than the coalescent that does not use sample time information.

1. Introduction

The genetic diversity of many pathogens is influenced by recent epidemiological history, and a variety of methods exist to estimate features of an epidemic history given random samples of pathogen genetic markers [1]. An issue that is central to how pathogen genetic diversity is understood is how infected individuals are sampled. A great deal of theory has been developed under the assumption of complete sampling, that is, that all infected individuals in the population are sampled and provide at least one pathogen sequence. These methods have found great utility for the study of small outbreaks [2,3], and for certain hospital-acquired infections [4]. A separate body of theory has developed for the study of epidemics where a sample of hosts is obtained for pathogen sequencing, and these methods are derived from classical population genetic models such as the coalescent [5,6] and classical population dynamics models such as the birth–death process [7,8]. This paper considers the scenario of incomplete sampling and the potentially confounding effects of non-random sampling through time on inference using the coalescent and birth–death-sampling formula [9].

The coalescent is a mathematical model of genealogies and describes the structure of genealogies generated by different demographic processes [10]. The coalescent has been the standard tool for demographic inference and is the underlying genealogical model in most phylogenetic software [11,12]. Under the neutral coalescent, the time between consecutive common ancestry events (the internode intervals) is modelled as a point process with a hazard rate $r(t)$ that depends on the effective population size $N_e(t)$ and the number of extant lineages in that interval $A(t)$ at time t in the past. With time in units of the generation interval τ, this becomes

$$r(t) = \frac{\binom{A(t)}{2}}{N_e(t)}.$$

By relating the time of common ancestry to the population size, the coalescent enables estimation of the latter. A variety of non-parametric [13–15] and parametric [13,16,17] models have been developed for N_e as a function of time. The parametric models for $N_e(t)$ tend to be deterministic functions of time, and we will consider such deterministic models in this paper, although there have been several recent attempts to fit stochastic demographic process models using the coalescent [18,19].

Birth–death processes trace their origins to work by Kendall [8], who showed how to calculate the probability that a given number of lineages will survive up to a given point of time in a stochastically growing population. Further results were developed by Thompson [20] and Gernhard [21], who showed how to calculate the probability density of genealogies generated by the birth–death process under complete sampling. These models were subsequently extended to account for incomplete sampling of the population by Stadler [9]. In order to account for incomplete sampling, the birth–death process must be combined with a model of the sampling process. Two sampling processes have thus far been considered in birth–death-sampling models (BDMs): sampling of lineages may take place at a constant rate; or, at a given point in time, a proportion of lineages may be sampled uniformly at random. These sampling processes may be combined, and recently developed methods allow sampling rates to vary through time according to a step function [22].

There are many variations on the coalescent and birth–death models that could be compared. Different coalescent and birth–death models make different assumptions about the demographic and sampling process, and each will be susceptible to different levels of bias by violation of those assumptions. We will focus on two models that have recently received considerable attention and have been used in epidemiological investigations. We use the BDM described in [9], and the coalescent model (CoM) for an unstructured population as described in [17]. Originally, CoMs were based on restrictive assumptions about the proportion of the population sampled and when taxa are sampled. CoMs were also based on strictly deterministic demographic processes, but all of these assumptions have been relaxed since the coalescent was first introduced. BDMs were originally based on census sampling at a single point in time, but that assumption has also been relaxed. Both models have been extended to consider heterogeneous structured populations [17,23].

The probability of observing a genealogy given demographic parameters may be calculated using either the CoM or the BDM, though these two models have very different mathematical foundations. The likelihood functions provided by each approach are difficult to reconcile mathematically, yet they tend to give similar results as we demonstrate below. The BDM has the advantage of accounting for stochasticity of the demographic process in an efficient and natural way. It is also possible to account for stochastically varying effective population size in the coalescent, but this has greater computational requirements [18]. A potential disadvantage of BDMs is that they require a model of the sampling process, whereas the coalescent makes no assumptions about how lineages are sampled through time. If the sampling process deviates from the simplistic processes that form the basis of current BDM theory, it is possible that estimates based on the BDM will be biased.

Both methods have particular advantages and vulnerabilities. Estimates based on CoMs may be biased by noisy demographic processes, and estimates based on the BDM may be biased by misspecification of the sampling process. In this paper, we will evaluate the vulnerability of both methods to these confounders. Because of the additional assumptions about sampling built into BDMs, it is difficult to make a direct comparison of the statistical power of BDMs and CoMs. If the sampling process is correctly specified, the observed sequence of sample times provides a great deal of information about the population size through time, which is not directly accessible with the CoM approach. Indeed, given a sequence of sample times, it is possible to estimate birth and death rates without a genealogy provided that the model of the sampling process is correctly specified (§3.1). We show that much of the statistical power of the BDM approach is derived from information in the sequence of sample times and not in the genealogy. This finding also suggests an enhancement to CoMs: if the sampling process is known, we can augment the CoM likelihood with a separate likelihood for the sequence of sample times. This augmented coalescent method is presented in §3.2.

In sampling at a single time point (homochronously), we show that estimates based on CoMs and BDMs are very similar. In §4.8, we show how the distribution of coalescent times predicted by CoM converges with large sample size to the distribution given by BDM.

2. The demographic and sampling processes

The population size $Y(t)$ is modelled as a continuous-time Markov chain on $[0, \infty)$, which is governed by the following transition probabilities:

$$\left.\begin{aligned} &P(Y(t+\Delta t) = Y(t)+1) = \lambda Y(t)\Delta t + O((\Delta t)^2),\\ &P(Y(t+\Delta t) = Y(t)-1) = \mu Y(t)\Delta t + O((\Delta t)^2)\\ \text{and}\quad &P(Y(t+\Delta t) = Y(t)) = 1-(\lambda+\mu)Y(t)\Delta t + O((\Delta t)^2), \end{aligned}\right\} \tag{2.1}$$

where λ and μ are the *per capita* birth and death rates of the process, respectively. Initially, $Y(0) = 1$.

We investigated three distinct sampling processes for the reconstruction of genealogies from a simulated demographic history:

(1) Continuous sampling through time at constant rate. According to this model, after a lineage dies (with a per-lineage rate μ), it is sampled with independent probability p.
(2) Homochronous sampling. According to this model, every extant lineage at a predetermined time point is sampled with independent probability ρ.
(3) Weighted sampling. According to this model, each unit has a sample weight at the time of death. If $\{t_i\}$ is the set of death times for lineages indexed by i, the sample weights are $w_i = e^{\alpha t_i}$. A sample is taken of n lineages without replacement with selection probabilities proportional to sampling weights.

Note that, with the exception of homochronous sampling, the lineages are only sampled at the time of death. This design is chosen for mathematical convenience, since it eliminates the possibility that a sampled lineage will be directly ancestral

to another sample, which would yield genealogies with zero branch lengths [24], although more complex BDMs and CoMs can be applied in this situation.

In this paper, we use CoMs based on the following deterministic approximation $y(t)$ to the stochastic process $Y(t)$:

$$y(t) = y(0)e^{t(\lambda-\mu)}, \tag{2.2}$$

with real-valued initial conditions $y(0)$ that will be estimated.

Genealogies were generated by simulation of the birth–death process in continuous time using the software MASTER v. 1.7.1 [25]. For simulating genealogies with a time-dependent sampling rate, we developed a custom simulator for the birth–death process in Python (see the electronic supplementary material). We simulated 300 genealogies for each of the three sampling scenarios given above using $\mu = 1$ and $\lambda = 2$ or $\lambda = 1.25$ or $\lambda = 1.1$. In the case of sampling through time, we terminated the simulation when 100 samples were collected and using a sampling probability of $p = 1\%$ or 50%. If sampling homochronously, we sampled 100 taxa after 9.21 or 25 units of time, yielding a sample proportion that varied around 1% or 20%, respectively. Simulations that failed to reach the target sample size were removed.

3. Estimation methods

All models are fitted by maximum likelihood (ML). The choice of ML was motivated by the simplicity of the demographic process, the speed of ML methods and the small number of free parameters. For the exponential growth process, there are four potential parameters that could be estimated: birth rates λ, death rates μ, the initial population size $y(0)$ (needed for CoMs but not BDMs), and a parameter that describes sampling (needed for BDMs but not CoMs). As previously shown in the analysis of BDMs, at most two of these parameters are identifiable from a genealogy alone, and we must therefore choose which parameters to fix according to prior knowledge, and CoMs are subject to the same identifiability constraints. We focus on an epidemiologically plausible scenario, where birth rates and the number of infections are unknown, but independent clinical information provides information on death rates. Consequently, we will assume $\mu = 1$ is known and will focus on the estimation of birth rate λ along with the nuisance parameters describing initial population size (for CoMs) or sampling rates (for BDMs). We will also consider the special case of homochronous sampling, in which we can reparametrize the CoM such that, like the BDM, estimates of the sampling fraction can be obtained.

Throughout the remainder of the paper, we use two symbols to denote time on different axes, and all dynamic variables will be defined on both axes. t will denote time from an arbitrary point in the past, whereas s will denote time before present. It will be useful to define the population genetic models in terms of the retrospective time axis s.

Let $\mathcal{G} = (\mathcal{N}, \mathcal{E}, X)$ represent a genealogy consisting of a set of nodes $\mathcal{N}$, edges $\mathcal{E}$ and a function $X : \mathcal{N} \to \mathbb{R}$ that gives the time s before the present of each node. Every edge corresponds to a 2-tuple (u, v) such that $u, v \in \mathcal{N}$ and the node u is said to be ancestral to v. We will consider only rooted binary genealogies; every internal node has exactly two descendants, and all internal nodes but the root have exactly one ancestor.

For CoMs, we use the likelihood given in [17], and we denote the maximum likelihood estimator (MLE) birth rate $\hat{\lambda}^{C}$.

This likelihood is that of a time-inhomogeneous point process with a hazard rate that depends on the population size and number of extant lineages. Specifically, following the approach in [17], the total population birth rate will be denoted $f(s) = \lambda y(s)$ and the rate of coalescence is

$$r(s) = f(s)\frac{\binom{A(s)}{2}}{\binom{y(s)}{(2)}} = \binom{A(s)}{2}\frac{2\lambda}{y(s)-1}, \tag{3.1}$$

where the first equality can be understood as the hypergeometric probability of selecting two lineages that are ancestral to the sample out of the set of $y(s)$ lineages. Now let x' denote the vector of node times (including sampled tips) in ascending order. The probability of observing the i'th interval is

$$P_i = \begin{cases} e^{-\int_{x_i}^{x_{i+1}} r(s)\mathrm{d}s} & x_{i+1} \text{ is a sample time} \\ r(x_{i+1})e^{-\int_{x_i}^{x_{i+1}} r(s)\mathrm{d}s} & x_{i+1} \text{ is a coalescent time.} \end{cases} \tag{3.2}$$

And the likelihood is

$$\mathcal{L}_{\mathrm{CoM}}(\lambda, \mu, y_0|\mathcal{G}) = \prod_{i=0}^{2n-2} P_i. \tag{3.3}$$

Note that the number of terms in the likelihood is the number of internode intervals $2n - 2$ if all sampling times are distinct. One subtlety arises if more than one lineage is sampled at a single time point, such as with a homochronous sample, in which case we simply deduplicate elements in the vector x' and adjust the number of terms in the likelihood.

For BDMs, we used the ML framework described in [9]. We denote the MLE birth rate $\hat{\lambda}^{BD}$. The R package expoTree [26] was used along with the implementation described here, and all results presented below are based on the best performing of the two implementations of the BDM likelihood. We simplified the likelihood equations in [9] to two situations: sampling occurs at a single time point with sample fraction ρ, or individuals are sampled with probability p at the time of death. Let x denote the vector of times before present for each internal node in $\mathcal{G}$ in descending order. Note that x_0 corresponds to the root of the tree. If the sampling takes place according to the homochronous process, ρ will denote the probability of sampling a lineage at a single point in time. Then,

$$\mathcal{L}_{\mathrm{BDM}}(\lambda,\mu,\rho|\mathcal{G}) = \lambda^{n-1}(4\rho)^n \prod_{i=0}^{n-2} q(x_i, c_2)^{-1}\left(\int_{x_{\mathrm{or}}=x_0}^{\infty} q(x_{\mathrm{or}}, c_2)^{-1}\mathrm{d}x_{\mathrm{or}}\right), \tag{3.4}$$

where $q(\cdot)$ is derived from the birth–death-sampling formula:

$$q(s, c) = 2(1 - c^2) + e^{-c_1 s}(1 - c)^2 + e^{c_1 s}(1 + c)^2, \tag{3.5}$$

and c_1 and c_2 are the following constants:

$$c_1 = |\lambda - \delta| \tag{3.6}$$

and

$$c_2 = -\frac{\lambda - \delta - 2\lambda\rho}{c_1}. \tag{3.7}$$

Note that the integral in the likelihood equation accounts for the unobserved time of origin of the birth–death process.

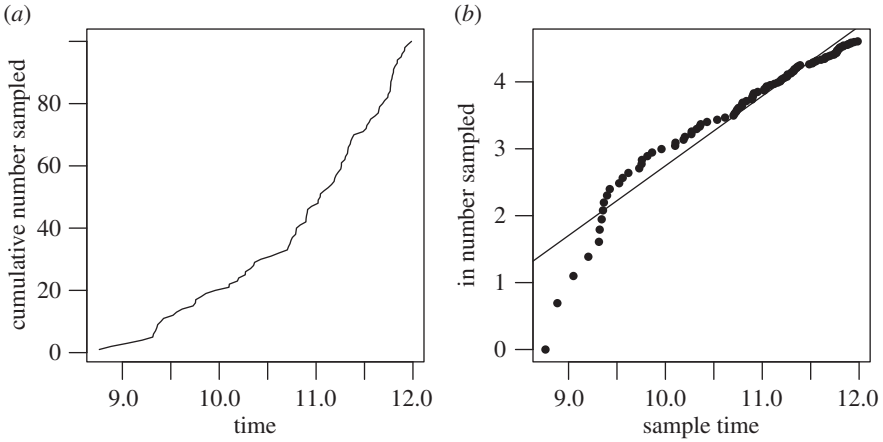

Figure 1. (*a*) Cumulative number of samples through time. (*b*) Log cumulative samples with regression line.

If sampling heterochronously at rate $\psi = \mu p$, the likelihood has a different form. Let x denote the vector times before present of each node as above, and let y denote the vector of sample times in any order.

$$\mathcal{L}_{\mathrm{BDM}}(\lambda,\mu,\psi|\mathcal{G}) = \int_{x_{\mathrm{or}}=x_0}^{\infty} \lambda^{n-1}\psi^n \left\{ q(x_{\mathrm{or}},c_2)^{-1} \prod_{i=0}^{n-1} q(y_i,c_2) \prod_{i=0}^{n-2} q(x_i,c_2)^{-1} - q(x_{\mathrm{or}},c_3)^{-1} \prod_{i=0}^{n-1} q(y_i,c_3) \prod_{i=0}^{n-2} q(x_i,c_3)^{-1} \right\} \mathrm{d}x_{\mathrm{or}}, \tag{3.8}$$

and c_1, c_2 and c_3 are the following constants:

$$c_1 = \sqrt{(\lambda-\mu)^2 + 4\lambda\psi}, \tag{3.9}$$

$$c_2 = \frac{\mu-\lambda}{c_1} \tag{3.10}$$

and

$$c_3 = \frac{\lambda+\mu}{c_1}. \tag{3.11}$$

BDMs and CoMs were fitted according to the same numerical algorithm, with maximization of the likelihood accomplished in R using the simplex method. In order to ensure convergence to the global maximum, multiple starting conditions were drawn from a multivariate uniform distribution, and the likelihood optimized for each. The best model fit is reported among the three or five optimizations, although in general they converged to the same value.

3.1. Estimating birth rates using times of sampling

Consider the sequence of sample times in increasing order $\boldsymbol{t} = (t_1, \ldots, t_n)$. If the sampling process is known, the sequence of sample times is informative about population size. We will consider a simplistic sampling process such that individuals are sampled at a constant rate upon death, which is the sampling process underlying current BDMs. If sampling occurs at a constant known rate, it is straightforward to estimate the historical population size from the sample times, since the probability that a sample will be observed at some point in time is proportional to population size at that time. Therefore, it is possible to estimate the population size using sample time information alone, and not using the genealogy. We show here that it is possible to estimate the birth rate, even if the sampling rate is unknown. Two simple estimators are presented. The first is based on a simple regression with the expected cumulative number of samples through time. The second is based on treating the sample times as arising from a point process and using ML.

Let $S(t)$ denote the cumulative number of samples collected up to time t. We show that the cumulative number of samples increases at the same rate as the unknown population size. According to the deterministic model, the expected change in S over time Δt will be

$$\begin{aligned} \Delta S(t) &= (\Delta t) p\mu y(t) + O((\Delta t)^2) \\ &= (\Delta t) p\mu y(0) \mathrm{e}^{(\lambda-\mu)t} + O((\Delta t)^2). \end{aligned} \tag{3.12}$$

Consequently, the logarithm of $S(t) \propto (\lambda - \mu)t$. Regressing the vector $\log(S(\boldsymbol{t}))$ on the vector $\boldsymbol{t}$ yields an estimate of the growth rate $k = \lambda - \mu$, and using knowledge of $\mu = 1$ we have the regression estimator

$$\hat{\lambda}^{\mathrm{R}} = \hat{k} + \mu. \tag{3.13}$$

Figure 1 shows the number of samples through time, for a single simulated genealogy along with the regression line.

The likelihood approach is based on modelling the sequence of sample times as a point process and also makes use of the deterministic approximation to population size. The rate of a sample appearing at time t is

$$f(t) = p\mu y(0) \mathrm{e}^{(\lambda-\mu)t} = a\mathrm{e}^{k_s t},$$

with $a = p\mu y(0)$ and $k_s = \lambda - \mu$. The probability of $\boldsymbol{t}$ is

$$P(\boldsymbol{t}|a,k) = \prod_{i=1} f(t_i) \mathrm{e}^{-\int_{t_{i-1}}^{t_i} f(\tau)\mathrm{d}\tau}. \tag{3.14}$$

As with the regression estimator,

$$\hat{\lambda}^{\mathrm{S}} = \hat{k}_s + \mu. \tag{3.15}$$

3.2. The augmented coalescent model

The genealogy $\mathcal{G}$ and the sample times $\boldsymbol{t}$ are conditionally independent given demographic and sampling parameters $\theta = (\lambda, \mu, p, y(0))$. Therefore, the likelihood of both is the product of the marginal likelihoods given above (equations (3.15) and (3.3)):

$$P(\mathcal{G}, \boldsymbol{t}|\theta) = P(\mathcal{G}|\theta) P(\boldsymbol{t}|\theta). \tag{3.16}$$

We will denote the MLE birth rate as $\hat{\lambda}^{\mathrm{A}}$.

4. Results

The following results demonstrate the level of bias, precision and efficiency of different inference methods when estimating the birth rate from genealogies generated by the birth–death process.

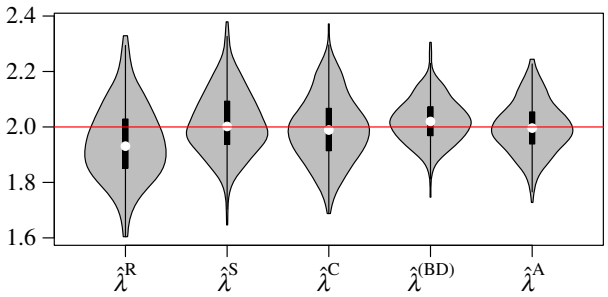

Figure 2. Distribution of MLE birth rates from 300 simulations with constant sampling rate and using five different estimators. The red line shows the true parameter value. $\hat{\lambda}^R$ is a regression estimator, $\hat{\lambda}^S$ is an ML estimator using sample times, $\hat{\lambda}^C$ is the coalescent estimator, $\hat{\lambda}^{(BD)}$ is the BDM estimator and $\hat{\lambda}^A$ is the coalescent estimator that also uses sample times. (Online version in colour.)

4.1. Constant sampling rate

Figure 2 shows the distribution of MLEs for five estimators presented above based on 300 simulated genealogies. Simulations were based on a sampling process with constant sampling probability $p = 1\%$ at the time of death.

Estimators based only on the sequence of sample times $\boldsymbol{t}$ perform well even though they do not use the coalescence times. The ML estimator $\hat{\lambda}^S$ consistently outperforms the simple regression estimator $\hat{\lambda}^R$, presumably reflecting that residuals in the loglinear regression model are not i.i.d. normal.

Comparing a model that only uses sample time information (λ^S) with the CoM that only uses genealogical information (λ^C) we find that that the RMSE of $\hat{\lambda}^S$ is 11.5% compared with 11.8% for $\hat{\lambda}^C$. In this instance, the sample time sequence is actually more informative than the coalescent times for inferring birth rates.

Comparing the BDM and coalescent, we find that the BDM is more precise (RMSE = 0.085) but slightly less accurate; the average bias of the BDM estimator was 0.022 (95% CI: (0.013, 0.031)) compared with 0.009 (95% CI: (−0.022, 0.005)) for the CoM estimator. Comparing the BDM and augmented CoM (a model that uses both coalescence and sample times), we find that the augmented CoM is slightly less precise than the BDM (RMSE of $\hat{\lambda}^A$ is 0.092), which may reflect the use of a misspecifed deterministic population size; however, we did not detect significant bias of $\hat{\lambda}^A$ (95% CI: (−0.012, 0.009)) in contrast to the BDM.

Figure 3 sheds some light on why the estimators perform differently by comparing the ML estimated by each method on each simulated genealogy. Electronic supplementary material, figure S1, shows a similar scatter plot of MLE birth rates. The BDM likelihood is highly correlated with that of all other estimators. By contrast, the CoM likelihood is almost independent of the estimators that use sample times only (Pearson correlation = 0.066). The highest correlation is found between the BDM and the augmented CoM (Pearson correlation = 0.95). This illustrates that the CoM is not using sample time information, but the BDM and augmented CoM are using information from both the sample times and genealogy.

4.2. Homochronous sampling

If all samples are collected at a single point in time, and if the sampling proportion is unknown, then the time of sampling and sample size confer no information about population size. The homochronous sampling case with unknown sampling rate therefore provides a fair comparison for BDMs and CoMs. Here, we consider 300 simulations of the birth–death process with a sample of $n = 100$ at $t = 9.2$, so that the sample fraction is around 1%, though it differs between replicates. The birth rate used in the simulations was $\lambda = 2$.

Figure 4 shows the distribution of MLE birth rates. The distributions are very similar and have similar precision (RMSE of $\hat{\lambda}^{BD}$ is 0.106 and RMSE of $\hat{\lambda}^C$ is 0.101). The CoM estimator does not have detectable bias (95% CI of bias: (−0.0183, 0.0048)), but the BD model slightly overestimates birth rates (average bias = 0.036, 95% CI: (0.0242, 0.0470)). Figure 4 also shows that the log likelihoods of the MLEs generated by both methods are highly concordant up to a constant factor. The Pearson correlation of BDM and CoM MLs is 99.6%. The estimated birth rates also have a high correlation coefficient of 86.6%.

Comparing the RMSE of the BDM estimator in both the homochronous and constant sampling rate cases, it appears that having informative sample time information decreases the residual sums-of-squares of the BDM estimator by about 36%, but this gain in precision will certainly depend on parameters of the system and sample size.

We repeated the simulation exercise with a smaller birth rate ($\lambda = 1.25$) in order to assess if the CoM estimator would be less accurate if the population is growing more slowly. The MLEs are depicted in the electronic supplementary material, figure S2. With the smaller birth rate, we do not detect significant bias of the BDM estimator (average bias less than 1×10^{-3}, 95% CI: (−0.0069, 0.0077)), or with the CoM estimator (average bias 0.002, 95% CI of bias: (−0.0057, 0.0096)). The RMSE of the BDM and CoM estimators are similar (0.037 and 0.039, respectively).

4.3. Coverage

To assess the ability of both estimators to estimate accurate confidence intervals, we computed likelihood profiles with the bbmle package in R. We also computed confidence intervals using a parametric bootstrap method for the CoM estimator. 95% CIs were computed for each of 300 simulations with $\lambda = 2$ and 1.25 and homochronous sampling. The BDM estimator provides excellent coverage with profile likelihoods. When $\lambda = 2$, BDM has 95.3% coverage, and when $\lambda = 1.25$, BDM has 95.5% coverage. By contrast, when $\lambda = 2$, the deterministic CoM has 80.5% coverage, and when $\lambda = 1.25$, the deterministic CoM has 75.1% coverage using profile likelihoods.

Because the RMSE of the CoM estimator is similar to that of the BDM estimator, we hypothesized that a bootstrap method would provide more reliable confidence intervals for CoM. For each MLE based on CoM, we simulated 100 coalescent trees using MLE parameters, re-estimated λ for each, and computed confidence intervals based on quantiles of the bootstrap distribution. In order to maximize speed of the bootstrap algorithm, we simulated node heights using the approximate coalescent rates described in §4.8. We find that the CoM estimator has very good coverage with the parametric bootstrap method. When $\lambda = 2$, the deterministic CoM has 95.0% coverage, and when $\lambda = 1.25$, the deterministic CoM has 93.8% coverage.

4.4. Comparison of estimated sample rates

An alternative parametrization of the coalescent is in terms of the population size at the time of sampling in a homochronous scenario. In this case, we can calculate a deterministic

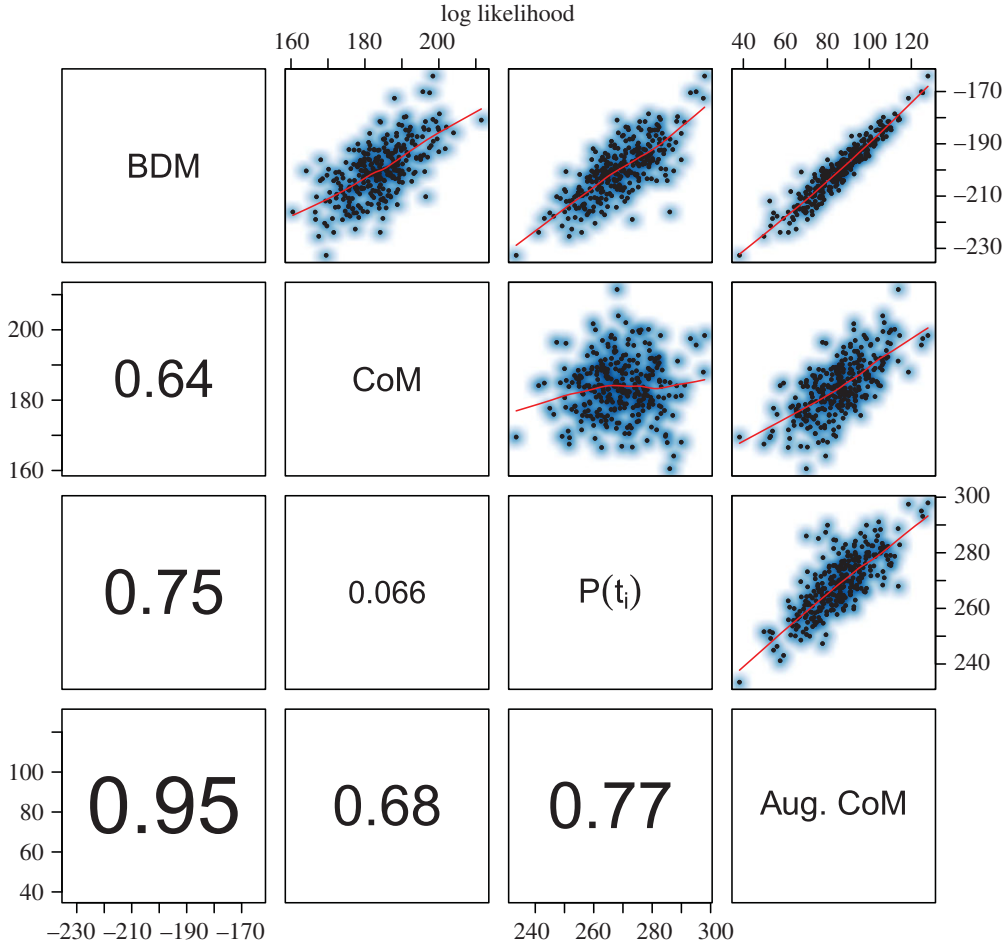

Figure 3. The log likelihood corresponding to MLEs from four estimation methods and based on 300 simulated genealogies. The $P(t_i)$ method refers to the likelihood estimator based on times of sampling only (equation (3.14)). The Pearson correlation coefficient between log likelihoods is shown in the lower panels. The upper panels show a scatter plot with smoothing splines. (Online version in colour.)

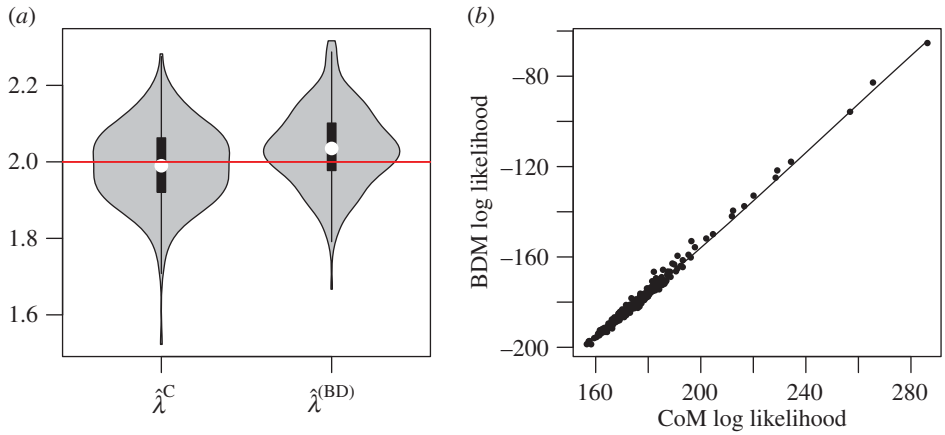

Figure 4. MLEs based on simulations with homochronous sampling. (*a*) Distribution of MLE birth rates from 300 simulations and using BDM and CoM estimators. The red line shows the true parameter value. (*b*) The log likelihoods of the BDM and CoM MLEs with a smoothing spline. (Online version in colour.)

approximation to the population size at time s in the past as

$$y(s) = \frac{n}{\rho} e^{-s(\lambda-\mu)},$$

where n is the sample size and ρ is the sample proportion, and n/ρ is the population size at the time of sampling. According to this parametrization, we replace the nuisance parameter $y(0)$ with ρ, and the coalescent estimates of the sample proportion can be directly compared to estimates with the BDM.

We fit the reparametrized CoM to the same genealogies used in §3.2 with $\lambda = 1.25$ and $\mu = 1$. Figure 5 shows that the estimates are highly concordant with Pearson correlation of 99.7%.

4.5. Small reproduction number and high sample fraction

The CoM based on a deterministic demographic process may be most biased when the population size is small and subject to large stochastic fluctuations. We generated 300 trees from

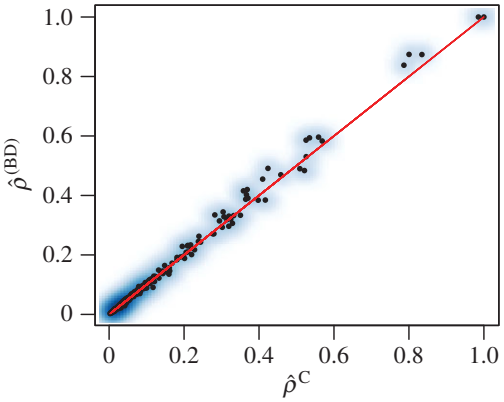

Figure 5. Estimated sample proportions using the coalescent and BDM with homochronous sampling. $\lambda = 1.25$, $\mu = 1$, $n = 100$. (Online version in colour.)

the BD process with $\lambda = 1.1$, $\mu = 1$ and homochronous sampling with $n = 100$ and a variable sample fraction around 50%. The distribution of MLE birth rates is shown in figure 6.

We found small but significant bias in the estimated birth rates using both BDM and CoM methods. The mean bias of the BDM estimator was 0.019 (95% CI: (0.0148, 0.0244)), and the mean bias of the CoM was 0.021 (95% CI of bias: (0.0160, 0.0265)). The BDM had smaller RMSE (0.043 versus 0.47), and the Pearson correlation of estimated birth rates was 95%. A comparison of estimated birth rates is shown in the electronic supplementary material, figure S3.

4.6. Decreasing sample rate and small sample fraction

When the sampling model implicit to the BDM approach is misspecified, the BDM may yield highly biased results. Figure 7 shows the MLE birth rates for both the BDM and CoM estimators when the sampling rate changes through time according to $e^{\alpha t}$ (see §2). One hundred and twenty simulations were carried out, and the sampling rate decreased at a rate of $\alpha = -0.44$. This value was chosen so that the expected sample size would be 100 if taking a weighted sample of all lineages at the time of death. Note that the sampling rate is an exponential function of time, so that the sequence of sample times still appears as though it arises from an exponentially increasing population, and there would be no warning from the sequence of sample times alone that the rate is changing. The BDM estimates are biased downwards by 0.23 (95% CI: (−0.2488, −0.2207)).

In this scenario, the CoM is robust to changing sample rate, since the CoM conditions on observed sample times. The CoM estimates did not have significant bias (95% CI of bias: (−0.0443, 0.0045)).

4.7. Increasing sample rate and large sample fraction

In these experiments, we examine bias in the coalescent due to sampling a large fraction of lineages from a small population growing stochastically. Three hundred genealogies with $n = 100$ were simulated from a birth–death process. Simulations were terminated when the number of deceased lineages reached 200, so that the sample fraction was 50% of deceased lineages and about 25% of all lineages. In the same experiments, we examined bias in BDMs due to a misspecified sampling process. In these experiments, the sampling rate increases from zero at time zero at a rate of $\rho = \mu$.

Figure 8 shows the distribution of MLE birth rates. We do not find detectable bias with the CoM estimator (95% CI: (−0.0271, 0.0260)), despite using a misspecified deterministic approximation to the demographic process, and despite that a large sample of the population was taken and that the population size was only around 400 on average at the time of the last sample.

Because the BDM relies on a misspecified sampling process, the BDM estimator gives highly biased estimates in this scenario. The average bias was 0.46 (95% CI: (0.4460, 0.4920)).

4.8. Asymptotic distribution of coalescent times

Some insight into why CoM and BDM give similar estimates can be gained by comparing the asymptotic distribution of coalescent times predicted by both models in the case of homochronous sampling. The distribution of coalescent times in the limit of large sample size for a deterministic CoM can be easily computed, and we show that this distribution is equivalent to the marginal likelihood of a node given by the birth–death model.

In [27,28], an approximation to the lineages through time for the coalescent process was presented for a population under exponential growth:

$$\frac{\mathrm{d}}{\mathrm{d}s}A = -\binom{A(s)}{2}\frac{2\lambda}{y(s)}. \tag{4.1}$$

If sampling occurs at a single time point, such that $A(0) = n$, this has the unique solution

$$A(s) = \frac{1}{1 - (1/n)(n-1)\mathrm{e}^{-\lambda(\mathrm{e}^{s(\lambda-\mu)}-1)/y_0(\lambda-\mu)}}, \tag{4.2}$$

where y_0 is the population size at the time of sampling. We will call this a doubly deterministic coalescent model (DDCoM) because both the demographic and genealogical processes are modelled with deterministic approximations. The asymptotic distribution of coalescent times for the DDCoM is given by the derivative of $A(s)$ (equation (4.1)) and expanding $y(s)$ and normalizing

$$P_{\mathrm{DDCoM}}(s|\lambda, \mu, \rho, n) = -\frac{\mathrm{d}}{\mathrm{d}s}\frac{A}{n-1} \tag{4.3}$$

$$= \frac{\lambda\rho\mathrm{e}^{\lambda\rho(\mathrm{e}^{s(\lambda-\mu)}-1)/n(\lambda-\mu)+s(\lambda-\mu)}}{\left(n\mathrm{e}^{\lambda\rho(\mathrm{e}^{s(\lambda-\mu)}-1)/n(\lambda-\mu)} - n + 1\right)^2}. \tag{4.4}$$

The factor of $n - 1$ normalizes the distribution since there are $n - 1$ nodes in the tree. In [29], the DDCoM was found to be an excellent approximation to the stochastic coalescent for large populations.

The BDM likelihood takes the form of a product over coalescent times and sample times, including the time of origin. Conditioning on the time of origin, and given a homochronous sample, the likelihood is given by the product of marginal probabilities for each coalescent time. From equation (3.4), expanding c_1,c_2 and simplifying,

$$P_{\mathrm{BDM}}(s|\lambda, \mu, \rho) = \frac{4\lambda\rho}{q(s, c_2)}$$

$$= \frac{4\lambda\rho}{2(1-c_2^2) + \mathrm{e}^{-c_1 s}(1-c_2)^2 + \mathrm{e}^{c_1 s}(1+c_2)^2}, \tag{4.5}$$

where c_1 and c_2 are the following constants:

$$c_1 = |\lambda - \mu| \tag{4.6}$$

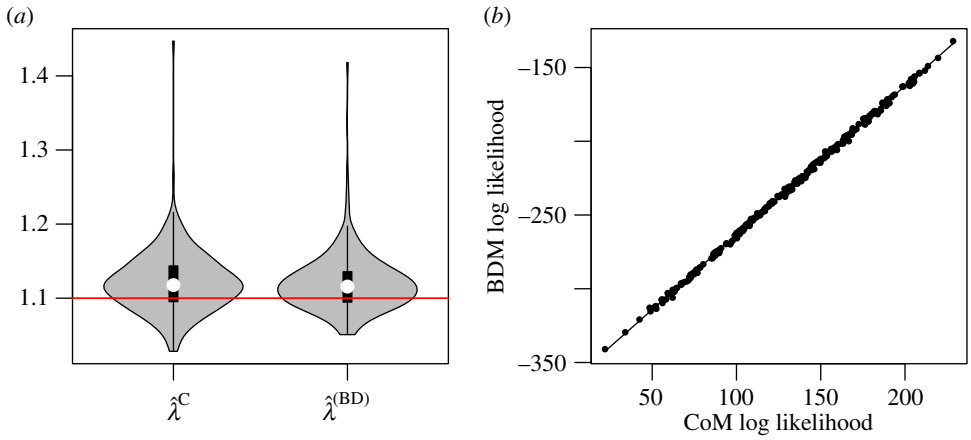

Figure 6. MLEs based on simulations with homochronous sampling, $\lambda = 1.1$, $\mu = 1$ and a variable sample fraction around 50%. (*a*) Distribution of MLE birth rates from 300 simulations and using BDM and CoM estimators. The red line shows the true parameter value. (*b*) The log likelihoods of the BDM and CoM MLEs with a smoothing spline. (Online version in colour.)

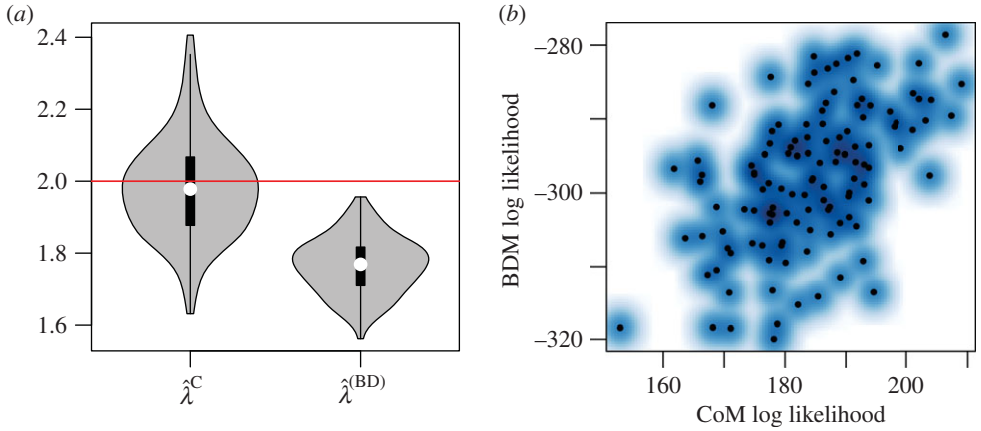

Figure 7. MLE birth rates based on simulations with time-dependent sampling. (*a*) Distribution of MLE birth rates from 120 simulations and using BDM and CoM estimators. (*b*) The log likelihoods of the BDM and CoM MLEs with a smoothing spline. (Online version in colour.)

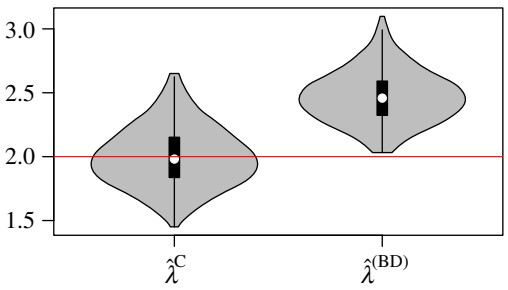

Figure 8. MLE birth rates based on simulations with a 50% sample fraction, $n = 100$, and with a time-dependent sampling rate that increases through time. Left distribution of MLE birth rates from 300 simulations and using BDM and CoM estimators. (Online version in colour.)

and

$$c_2 = -\left(\frac{\lambda - \mu - 2\lambda\rho}{c_1}\right). \tag{4.7}$$

Theorem 4.1. *Given a homochronous sample of a proportion ρ lineages from a population growing exponentially according to the birth–death process with birth rate λ, death rate μ, and $\lambda > \mu$,*

$$\lim_{n\to\infty} P_{\mathrm{DDCoM}}(s|n, \lambda, \mu, \rho) = P_{\mathrm{BDM}}(s|\lambda, \mu, \rho),$$

for all times s.

Proof. By Taylor expansion of the denominator of equation (4.4), we have

$$\left(ne^{\lambda\rho(e^{s(\lambda-\mu)}-1)/n(\lambda-\mu)} - n + 1\right)^2 = \left(1 + \frac{\lambda\rho(e^{s(\lambda-\mu)}-1)}{(\lambda-\mu)} + O\left(\frac{1}{n}\right)\right)^2. \tag{4.8}$$

The limit of the numerator of equation (4.4) is

$$\lim_{n\to\infty} \lambda\rho e^{\lambda\rho(e^{s(\lambda-\mu)}-1)/n(\lambda-\mu)+s(\lambda-\mu)} = \lambda\rho e^{s(\lambda-\mu)}. \tag{4.9}$$

Taking the large n limit of equation (4.8) and computing the ratio of (4.8) and (4.9), and rearranging, we have

$$\lim_{n\to\infty} P_{\mathrm{DDCoM}}(s|\lambda, \mu, \rho) = \frac{\lambda\rho(\lambda-\mu)^2 e^{s(\lambda-\mu)}}{(\lambda - \mu - \lambda\rho + \lambda\rho e^{s(\lambda-\mu)})^2}. \tag{4.10}$$

It may be verified that this is equivalent to P_{BDM} (equation (4.5)). ■

Note that this result applies to the DDCoM and not the CoM used elsewhere in the text. In [29,30], it was shown that the lineages through time given by DDCoMs are generally excellent approximations to lineages through time given by standard CoMs if the sample size is large.

Outside of the large-n limit, we can investigate the similarity of P_{BDM} and P_{DDCoM} numerically. To summarize the

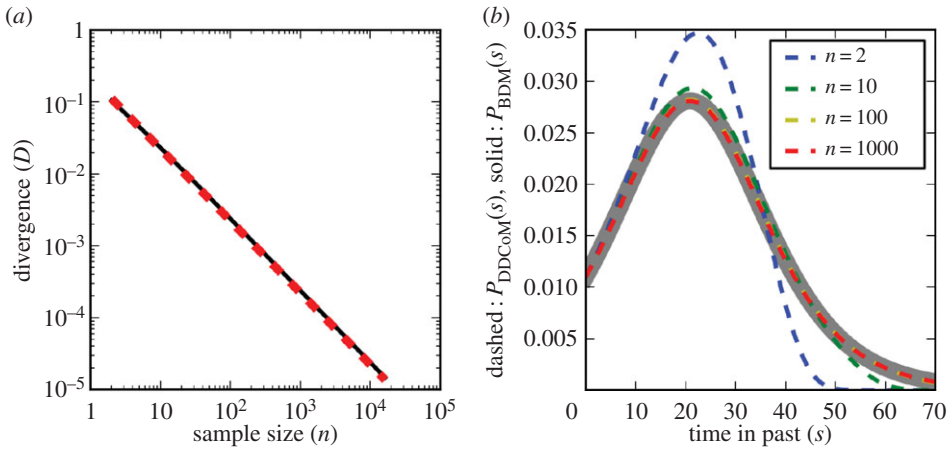

Figure 9. (*a*) The Kullback–Leibler divergence between the coalescent and birth–death distribution (black line) is shown versus the sample size on logarithmic axes. The red line shows a linear approximation ($e^{-3/2}/n$). $\lambda = 1.1$, $\mu = 1$ and $\rho = 0.9$. (*b*) The birth–death marginal density of a node (grey) is compared with the coalescent density based on samples of size $n = 2$, 10, 100 and 1000. $\lambda = 1.1$, $\mu = 1$ and $\rho = 0.01$.

difference between distributions P_{BDM} and P_{DDCoM}, we compute the Kullback–Leibler divergence

$$D(P_{DDCoM}, P_{BDM}|\lambda, \mu, \rho, n) = \int_{s=0}^{\infty} \log\left(\frac{P_{DDCoM}(s|\lambda, \mu, \rho, n)}{P_{BDM}(s|\lambda, \mu, \rho)}\right) P_{DDCoM}(s|\lambda, \mu, \rho, n)\mathrm{d}s.$$

Figure 9 shows the divergence as a function of sample sizes ranging from $n = 2$ to 2^{14} and with $\lambda = 1.1$, $\mu = 1$ and $\rho = 0.9$. We find that divergence is very insensitive to birth rates and sample proportion, so results are only shown for one scenario. When $n = 2$, the divergence is quite high, but it rapidly converges to zero. We observe that, to excellent approximation, the divergence scales in a very simple way as a function of sample size: $D(P_{DDCoM}, P_{BDM}|\lambda\mu, \rho, n) \approx e^{-3/2}/n$, and this is shown by the red line in figure 9.

Figure 9 also shows a comparison of the DDCoM marginal density of coalescent times with the BDM marginal likelihood with several different sample sizes and a smaller sample fraction of $\rho = 0.01$. When $n = 2$, the distributions are quite different, but when $n = 10$ they are very similar and when $n \geq 100$ they are almost indistinguishable.

5. Discussion

Two distinct areas of concern have arisen related to phylodynamic inference using CoMs and BDMs. CoMs based on a deterministic demographic process may be subject to inductive bias if the deterministic process is a bad approximation to the true stochastic demographic process. Similarly, BDMs are subject to bias if the model of the sampling process is misspecified. We have found that the bias due to the deterministic approximation is generally very small for populations growing exponentially, even when sampling 50% of individuals from a small population. Furthermore, errors in CoMs due to a deterministic process can be resolved with additional computational effort, as it is possible to use the coalescent with a stochastic demographic process [19,31]. Such methods may be necessary for populations with very small and noisy population dynamics. Bias is likely to be greatest if the population is small and growing slowly such that population dynamics are relatively noisy. Indeed, we found only one situation where the BDM was noticeably more precise than CoM estimators, which occured with a small R_0 of 1.1 and a

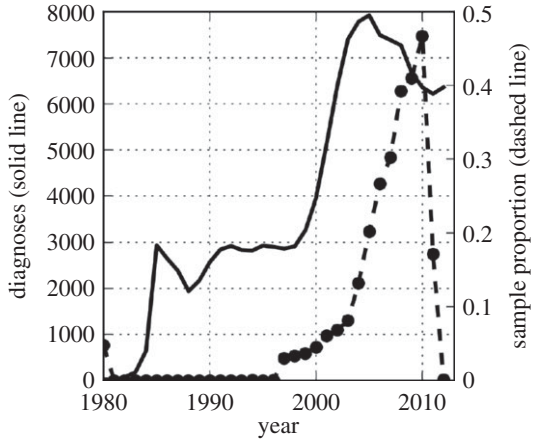

Figure 10. The HIV sequence sample rate through time using data from the UK HIV Drug Resistance Database [33] and the number living diagnosed with HIV through time.

large sample fraction; however, we did not find a situation where the BDM estimator was substantially less biased than the CoM estimator. Confidence intervals based on profile likelihoods have superior coverage if using BDMs rather than deterministic CoMs, which may reflect the explicit incorporation of stochastic population dynamics in the BDM estimator. However, the combined use of the CoM with confidence intervals based on parametric bootstrapping gave estimates with low bias and very good coverage.

We have found that BDMs can yield highly biased estimates if the sampling process is misspecified. It may be hard to detect if the sampling process deviates from the modelled form in many real-world situations, and most real datasets are likely to violate the BDM sampling process assumptions to some degree. An example heterogeneous sampling through time is shown in figure 10 for a dataset which has previously been analysed with BDMs in [32]. Figure 10 shows the sampling proportion through time of HIV sequence samples in the UK HIV Drug Resistance Database [33]. Typical for HIV sequence databases, the sample proportion is essentially zero throughout the 1980s, and there is a rapid increase in sampling effort throughout the late 1990s and early 2000s, followed by a plateau after 2010 due to reporting delays. In [32], a BDM susceptible–infected–recovered (SIR) model was fitted to HIV sequence data from the UK under the assumption that the sampling rate

jumped from zero to a constant rate, but the time span of the estimated phylogenies ranged from 1978 to 2003 over which the true sampling rate varied greatly.

Future work should explore how violation of sampling assumptions may bias estimates of R_0 when fitting BDM SIR models.

The sequence of sample times may be informative about the population size through time if the sampling process can be correctly specified. We have shown how birth rates may be estimated from the sequence of sample times if sampling occurs according to the BDM assumptions, and this is possible even if the sample rate is not known. BDMs implicitly use the sequence of sample times to estimate birth and/or death rates, and this is the case even if the sampling rate is not given, but estimated. Comparisons of CoMs and BDMs should account for the effects of sampling, and a fair comparison can be obtained in the case of homochronous or serial-homochronous sampling with unknown sample rate, so that the sample times contain no information about population size and birth rates.

Previous simulation-based studies on fitting SIR epidemiological models to sequence data [31] have purported to show increased statistical efficiency of BDMs relative to CoMs, but these studies did not control for the informativeness of sample times, and the supposed advantage of BDM in these simulations is likely to be due to the sampling model and not the genealogical model. For example, the simulation studies in [31] did not consider a homochronous sample, a misspecified sampling process, or the possibility of extending the coalescent estimators to use sample time information. The study in [31] used a Bayesian method, in contrast to our ML methods, so some differences may also be due to the choice of priors. Popinga *et al*. [31] hypothesized that the difference in performance of BDMs and CoMs was due to the latter's use of a misspecified deterministic demographic process, but in the context of exponential growth, we found very little bias due to the deterministic approximations of the coalescent, but large biases due to the effects of sampling.

Future research on BDMs may reveal ways to accommodate more realistic sampling processes. For example, in [22], a piecewise constant sampling process was presented; however, this also required the introduction of many more parameters to describe the sampling process. If the sampling process is known, a useful alternative to BDMs is to model the sampling process in tandem with the coalescent. As we have shown, the coalescent likelihood of a genealogy is approximately independent of the likelihood of the sample times, and for complex sampling processes it is much easier to model the genealogical and sampling process separately and combine likelihoods than to derive a joint likelihood. In the case where stochasticity is important but the sampling process is complex, the combined use of the CoM likelihood and parametric bootstrapping offers a means to obtain reliable parameter estimates and associated confidence intervals.

Acknowledgements. The authors thank Alison Brown (PHE UK) for providing HIV diagnosis statistics from the UK. Xavier Didelot and Caroline Colijn (Imperial College London) provided many helpful suggestions. The funders had no role in study design, data collection and analysis, decision to publish or preparation of the paper.

Funding statement. E.M.V. is supported in part by US NIH Models of Infectious Disease Agent Study (MIDAS) grant U01GM110749. S.D.W.F. is supported in part by a UK Medical Research Council (MRC) Methodology Research Programme grant (MR/J013862/1). This research is jointly funded by the UK MRC and the UK Department for International Development (DFID) under the MRC/DFID Concordat agreement.

References

1. Volz EM, Koelle K, Bedford T. 2013 Viral phylodynamics. *PLoS Comput. Biol.* **9**, e1002947. (doi:10.1371/journal.pcbi.1002947)
2. Didelot X, Gardy J, Colijn C. 2014 Bayesian inference of infectious disease transmission from whole genome sequence data. *Mol. Biol. Evol*. **31**, 1869–1879. (doi:10.1093/molbev/msu121)
3. Ypma R, Bataille A, Stegeman A, Koch G, Wallinga J, Van Ballegooijen W. 2011 Unravelling transmission trees of infectious diseases by combining genetic and epidemiological data. *Proc. R. Soc. B* **279**, 444–450. (doi:10.1098/rspb.2011.0913)
4. Snitkin ES, Zelazny AM, Thomas PJ, Stock F, Henderson DK, Palmore TN, Segre JA, NISC Comparative Sequencing Program Group. 2012 Tracking a hospital outbreak of carbapenem-resistant Klebsiella pneumoniae with whole-genome sequencing. *Sci. Transl. Med.* **4**, 148ra116. (doi:10.1126/scitranslmed.3004129)
5. Kingman JFC. 1982 The coalescent. *Stoch. Process. Appl.* **13**, 235–248. (doi:10.1016/0304-4149(82)90011-4)
6. Wakeley J. 2009 *Coalescent theory: an introduction*, vol. 1. Greenwood Village, CO: Roberts & Company Publishers.
7. Rannala B, Yang Z. 1996 Probability distribution of molecular evolutionary trees: a new method of phylogenetic inference. *J. Mol. Evol.* **43**, 304–311. (doi:10.1007/BF02338839)
8. Kendall DG. 1948 On the generalized 'birth-and-death' process. *Ann. Math. Stat.* **19**, 1–15. (doi:10.1214/aoms/1177730285)
9. Stadler T. 2010 Sampling-through-time in birth–death trees. *J. Theor. Biol.* **267**, 396–404. (doi:10.1016/j.jtbi.2010.09.010)
10. Rosenberg NA, Nordborg M. 2002 Genealogical trees, coalescent theory and the analysis of genetic polymorphisms. *Nat. Rev. Genet.* **3**, 380–390. (doi:10.1038/nrg795)
11. Guindon S, Dufayard JF, Lefort V, Anisimova M, Hordijk W, Gascuel O. 2010 New algorithms and methods to estimate maximum-likelihood phylogenies: assessing the performance of PhyML 3.0. *Syst. Biol.* **59**, 307–321. (doi:10.1093/sysbio/syq010)
12. Drummond AJ, Suchard MA, Xie D, Rambaut A. 2012 Bayesian phylogenetics with BEAUti and the BEAST 1.7. *Mol. Biol. Evol.* **29**, 1969–1973. (doi:10.1093/molbev/mss075)
13. Pybus OG, Rambaut A, Harvey PH. 2000 An integrated framework for the inference of viral population history from reconstructed genealogies. *Genetics* **155**, 1429–1437.
14. Strimmer K, Pybus OG. 2001 Exploring the demographic history of DNA sequences using the generalized e plot. *Mol. Biol. Evol.* **18**, 2298–2305. (doi:10.1093/oxfordjournals.molbev.a003776)
15. Opgen-Rhein R, Fahrmeir L, Strimmer K. 2005 Inference of demographic history from genealogical trees using reversible jump Markov chain Monte Carlo. *BMC Evol. Biol.* **5**, 6. (doi:10.1186/1471-2148-5-6)
16. Donnelly P, Tavare S. 1995 Coalescents and genealogical structure under neutrality. *Annu. Rev. Genet.* **29**, 401–421. (doi:10.1146/annurev.ge.29.120195.002153)
17. Volz EM. 2012 Complex population dynamics and the coalescent under neutrality. *Genetics* **190**, 187–201. (doi:10.1534/genetics.111.134627)
18. Rasmussen DA, Ratmann O, Koelle K. 2011 Inference for nonlinear epidemiological models using genealogies and time series. *PLoS Comp. Biol.* **7**, e1002136. (doi:10.1371/journal.pcbi.1002136)
19. Rasmussen DA, Volz EM, Koelle K. 2014 Phylodynamic inference for structured epidemiological models. *PLoS Comput. Biol.* **10**, e1003570. (doi:10.1371/journal.pcbi.1003570)

20. Thompson EA. 1975 *Human evolutionary trees*. Cambridge, UK: Cambridge University Press.
21. Gernhard T. 2008 The conditioned reconstructed process. *J. Theor. Biol.* **253**, 769–778. (doi:10.1016/j.jtbi.2008.04.005)
22. Stadler T, Kühnert D, Bonhoeffer S, Drummond AJ. 2013 Birth–death skyline plot reveals temporal changes of epidemic spread in HIV and hepatitis C virus (HCV). *Proc. Natl Acad. Sci. USA* **110**, 228–233. (doi:10.1073/pnas.1207965110)
23. Stadler T, Bonhoeffer S. 2013 Uncovering epidemiological dynamics in heterogeneous host populations using phylogenetic methods. *Phil. Trans. R. Soc. B* **368**, 20120198. (doi:10.1098/rstb.2012.0198)
24. Gavryushkina A, Welch D, Stadler T, Drummond A. 2014 Bayesian inference of sampled ancestor trees for epidemiology and fossil calibration. (http://arxiv.org/abs/14064573)
25. Vaughan TG, Drummond AJ. 2013 A stochastic simulator of birth–death master equations with application to phylodynamics. *Mol. Biol. Evol.* **30**, 1480–1493. (doi:10.1093/molbev/mst057)
26. Leventhal GE, Günthard HF, Bonhoeffer S, Stadler T. 2014 Using an epidemiological model for phylogenetic inference reveals density dependence in HIV transmission. *Mol. Biol. Evol.* **31**, 6–17. (doi:10.1093/molbev/mst172)
27. Frost SD, Volz EM. 2010 Viral phylodynamics and the search for an effective number of infections. *Phil. Trans. R. Soc. B* **365**, 1879–1890. (doi:10.1098/rstb.2010.0060)
28. Maruvka YE, Shnerb NM, Bar-Yam Y, Wakeley J. 2011 Recovering population parameters from a single gene genealogy: an unbiased estimator of the growth rate. *Mol. Biol. Evol.* **28**, 1617–1631. (doi:10.1093/molbev/msq331)
29. Jewett EM, Rosenberg NA. 2014 Theory and applications of a deterministic approximation to the coalescent model. *Theor. Popul. Biol.* **93**, 14–29. (doi:10.1016/j.tpb.2013.12.007)
30. Chen H, Chen K. 2013 Asymptotic distributions of coalescence times and ancestral lineage numbers for populations with temporally varying size. *Genetics* **194**, 721–736. (doi:10.1534/genetics.113.151522)
31. Popinga A, Vaughan T, Stadler T, Drummond A. 2014 Bayesian coalescent epidemic inference: comparison of stochastic and deterministic SIR population dynamics. (http://arxiv.org/abs/14071792)
32. Kühnert D, Stadler T, Vaughan TG, Drummond AJ. 2014 Simultaneous reconstruction of evolutionary history and epidemiological dynamics from viral sequences with the birth–death SIR model. *J. R. Soc. Interface* **11**, 20131106. (doi:10.1098/rsif.2013.1106)
33. Brown AJL, Lycett SJ, Weinert L, Hughes GJ, Fearnhill E, Dunn DT. 2011 Transmission network parameters estimated from HIV sequences for a nationwide epidemic. *J. Infect. Dis.* **204**, 1463–1469. (doi:10.1093/infdis/jir550)