Using an Epidemiological Model for Phylogenetic Inference Reveals Density Dependence in HIV Transmission

Gabriel E. Leventhal,*[1] Huldrych F. Günthard,[2] Sebastian Bonhoeffer,[1] and Tanja Stadler[1]

[1]Institute of Integrative Biology, ETH Zurich, Zurich, Switzerland

[2]Division of Infectious Diseases and Hospital Epidemiology, University Hospital Zurich, Zurich, Switzerland

***Corresponding author:** E-mail: gabriel.leventhal@env.ethz.ch.

Associate editor: Robin Bush

Abstract

The control, prediction, and understanding of epidemiological processes require insight into how infectious pathogens transmit in a population. The chain of transmission can in principle be reconstructed with phylogenetic methods which analyze the evolutionary history using pathogen sequence data. The quality of the reconstruction, however, crucially depends on the underlying epidemiological model used in phylogenetic inference. Until now, only simple epidemiological models have been used, which make limiting assumptions such as constant rate parameters, infinite total population size, or deterministically changing population size of infected individuals. Here, we present a novel phylogenetic method to infer parameters based on a classical stochastic epidemiological model. Specifically, we use the susceptible-infected-susceptible model, which accounts for density-dependent transmission rates and finite total population size, leading to a stochastically changing infected population size. We first validate our method by estimating epidemic parameters for simulated data and then apply it to transmission clusters from the Swiss HIV epidemic. Our estimates of the basic reproductive number R_0 for the considered Swiss HIV transmission clusters are significantly higher than previous estimates, which were derived assuming infinite population size. This difference in key parameter estimates highlights the importance of careful model choice when doing phylogenetic inference. In summary, this article presents the first fully stochastic implementation of a classical epidemiological model for phylogenetic inference and thereby addresses a key aspect in ongoing efforts to merge phylogenetics and epidemiology.

Key words: **phylodynamics, density dependence, epidemic inference, birth–death, coalescent.**

Introduction

Phylogenetics is becoming increasingly popular thanks to a large availability of genetic sequence information, and consequently phylogenetic methods have successfully been applied to pathogen sequence data in order to obtain estimates of transmission and death rates (Pybus et al. 2001; Rasmussen et al. 2011; Stadler et al. 2012). The basis of these phylogenetic methods is the evolutionary tree reconstructed from the sampled pathogen population (see Pybus and Rambaut [2009] and references therein). If the evolutionary rate of the pathogen is sufficiently high such that the evolutionary and epidemiological timescales are similar, then the evolutionary trees can give insight into the transmission dynamics of the disease (Pybus et al. 2001; Drummond et al. 2003; Grenfell et al. 2004). These phylogenetic methods are of statistical nature and assume an underlying model that describes both the evolutionary and the population dynamics of the genetic sequences (Pybus and Rambaut 2009). The choice of phylogenetic method thus comes with model assumptions that may or may not be appropriate to a specific question. Phylogenetic methods are therefore susceptible to model misspecification that can lead to incorrect parameter estimates.

Models from mathematical epidemiology that describe the spread of a disease in a population are well established (Kermack and McKendrick 1927; Anderson and May 1991). These models are quite different in character from population models traditionally used in phylogenetic inference. There has been great effort to extend previous phylogenetic methods to account for the particularities that accompany the population dynamics of infectious diseases (Volz et al. 2009; Frost and Volz 2010, 2012; Rasmussen et al. 2011). One of the important aspects of many epidemiological models is that they account for saturation effects in the number of infected individuals, meaning that transmission decreases as the pool of susceptible individuals is depleted.

Most of the methods that infer epidemiological parameters from phylogenetic trees assume a population model that is based on the coalescent, which in its original formulation makes strong limiting assumptions (Kingman 1982). To overcome some of these limitations, there has been a range of work that has extended coalescent theory to incorporate more complex epidemiological models (Volz et al. 2009; Frost and Volz 2010; Koelle and Rasmussen 2011; Volz 2012). Yet, these extensions all assume a deterministically changing population size. Stochastic effects, however, have been shown to be of importance when considering epidemiological processes (Rohani et al. 2002).

More recently, Rasmussen et al. (2011) have used coalescent models and sequential Monte Carlo methods together

Article

Fast Track

with stochastic ordinary differential equations (ODEs) to infer parameters of the epidemiological process. Sequential Monte Carlo methods such as particle filters are powerful statistical tools that can approximate the likelihood exactly but come at a large computational cost (Wilkinson 2011).

Alternatively, methods based on the birth–death model can be used for phylogenetic inference (Kendall 1948; Nee 2006; Stadler 2010). The birth–death model is a discrete stochastic description of the process governing the population dynamics. Phylogenetic trees are then produced by adding a sampling process to the birth–death model. Birth–death models naturally account for the stochasticity of the population size in epidemiological models and allow for the estimation of epidemiological parameters when the assumptions of the coalescent are not justified.

The birth–death model has recently been extended to account for saturation effects in the case of species evolution, which has allowed for a much better fit to the number of lineages over time in small species trees (Etienne et al. 2012). This method cannot be directly applied to viral phylogenies as it requires all sequences to be sampled at a single point in time, a condition that is rarely satisfied for disease surveillance data. Furthermore, the method requires the solution of a series of high-dimensional initial value problems, which is computationally challenging and has only been successfully performed for population sizes of the order of tens of species.

Here, we present a new method for phylogenetic inference of epidemiological parameters which is based on the birth–death model. Our method accounts for both sequentially sampled genetic data (Stadler 2010) and saturation effects (Etienne et al. 2012). In particular, we estimate transmission and death rates, as well as the susceptible population size from sequence data under a stochastic susceptible-infected-susceptible (SIS) model. The SIS model is the standard model used to describe the spread of sexually transmitted diseases without immunity (Anderson and May 1991). We derive an expression for the likelihood of a transmission tree, which can then be used to estimate the model parameters in either a maximum likelihood (ML) or Bayesian framework (Drummond et al. 2002, 2005). We use a recently developed method to calculate matrix exponentials (Al-Mohy and Higham 2011) in order to solve the high-dimensional initial value problems required to compute the likelihood. Our method can calculate the likelihood of a single tree for population sizes of the order of 10,000 individuals. Using estimates for the epidemiological rates and the susceptible population size, we calculate the basic reproductive number, R_0 (Anderson and May 1979). We validate our method by reestimating parameters from simulated data using an SIS model. We then estimate epidemiological rates, the susceptible population size, and R_0 for ten transmission clusters of the Swiss HIV epidemic (Kouyos et al. 2010; Schoeni-Affolter et al. 2010; Stadler et al. 2012).

New Approaches

In this section, we will give a brief summary of the new method used and refer to the supplementary information, Supplementary Material online, for a more detailed description. In short, we calculate the likelihood that the observed phylogeny is a realization of a stochastic SIS model (see Materials and Methods).

We assume a susceptible-infected (SI) model with constant total population size N as a model for transmission (see Materials and Methods). An outbreak begins with a single infected individual and the disease is transmitted with rate β/N to susceptible individuals. Infected individuals are removed either through "death" with rate μ or through sampling with rate ψ. Sampling corresponds to the case where individuals are sequenced (e.g., prior to treatment) and become noninfectious thereafter, for example, due to successful treatment or behavior change (Stadler et al. 2013). We assume that a removed individual is immediately replaced by a new susceptible individual, resulting in a constant population size N. Under this assumption, the SI model is equivalent to an SIS model.

Based on a sampled phylogeny of an epidemic outbreak (see Materials and Methods), we derived an expression for the likelihood of a phylogenetic tree under an SI/SIS model. In the following, time will always be measured going backward from the present t_0 into the past. As with serial sampling in the coalescent framework (Drummond et al. 2002), we can split up a phylogenetic tree into time intervals between sampling times x and branching times y. During these time intervals, the number of lineages in the tree is constant, but it increases by 1 at a sampling event and decreases by 1 at a branching event (see fig. 1). We introduce the probability $p_i(I;\, t)$, which is the probability that within the i-th time interval, exactly I infected individuals at time t gave rise to the phylogeny observed between that time t and the present time t_0. Thus, for every i and t, we can write the probabilities that $0,1,2,\ldots,N$ infected individuals gave rise to the phylogeny as a vector $p_i(t) = (p_i(0;\, t), p_i(1;\, t), \ldots, p_i(N;\, t))$. The time evolution of this vector of probabilities $p_i(t)$ is then governed by a birth–death master equation (see supplementary methods, Supplementary Material online). The master equation translates to a system of linear ODEs for $p_i(t)$ within the i-th time interval,

$$\frac{dp_i(I;\, t)}{dt} = -I(\lambda(I) + \mu + \psi)p_i(I;\, t) + (I + k_i)\lambda(I)p_i(I+1;\, t) + (I - k_i)\mu p_i(I-1;\, t). \quad (1)$$

Here, k_i are the number of tree lineages during the i-th interval, μ is the death rate of individuals, ψ is the sampling rate, and $\lambda(I) = \beta(N - I)/N$ is the rate at which new infections occur in the SIS model. The above equation is linear in the $p_i(I;\, t)$ and its solution can thus be written in matrix form as,

$$p_i(t) = e^{C_i(t - t_{i-1})} p_i(t_{i-1}). \quad (2)$$

The tridiagonal matrix C_i contains all the information of the ODEs within the i-th time interval. This allows us to

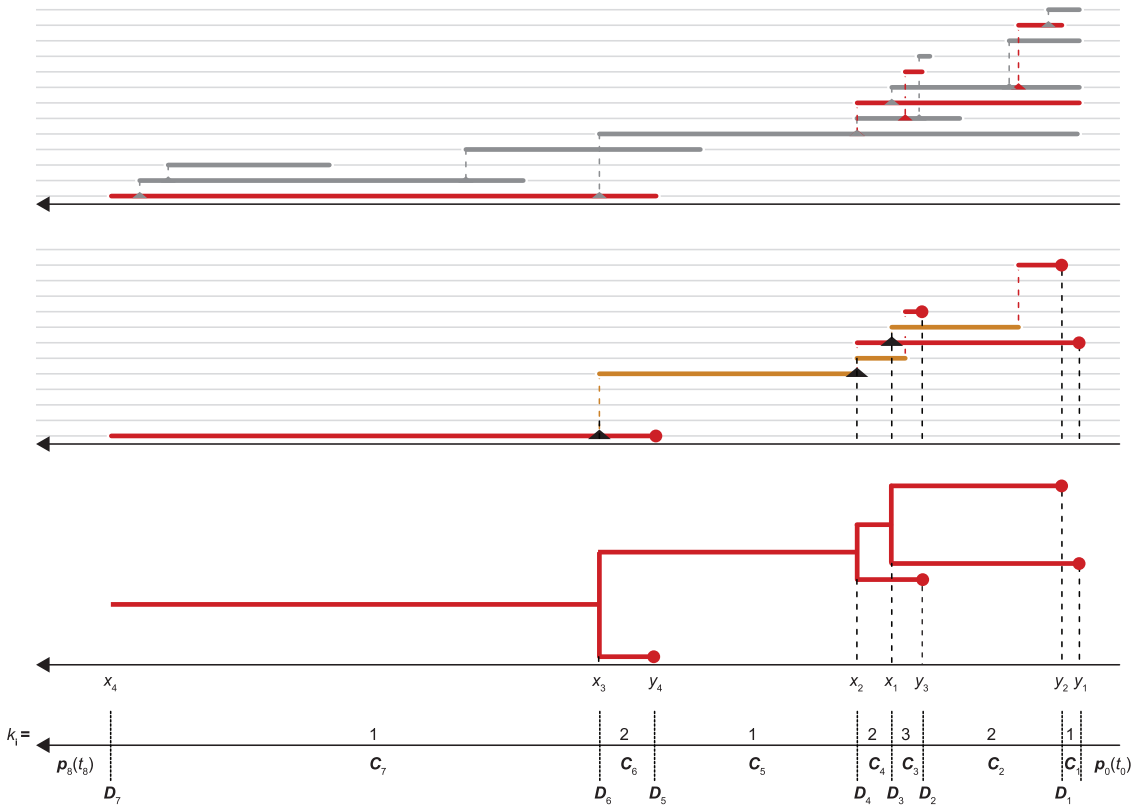

Fig. 1. Example of an epidemic with sampled (red) and unsampled (gray) individuals. The top panel shows the infective periods of all individuals in the epidemic. The middle panel shows the infective periods of only the sampled individuals as well as the recreated transmission tree. The red dots are the sampling times of the individuals and the black triangles the branching times on the sampled phylogeny. Note that while we do not know the exact infectious periods of the sampled individuals, the transmission chain between two events can pass through multiple individuals. The bottom panel shows the corresponding phylogenetic tree with branching times x_i and sampling times y_j. In this example, the joint event time vector is $t = (0,y_1,y_2,y_3,x_1,x_2,y_4,x_3,x_4)$. The axis at the bottom of the figure shows how the matrices in equation (5) are applied to the probability vector $p_i(t)$ from the present time t_0 to the root of the tree t_{2n} (here $n = 4$ and $t_8 = x_4$). The numbers along the axis are the number of extant lineages within that time interval. Note that the C_i matrices are applied as matrix exponentials, i.e., $p_i(t+h) = e^{C_i h} p_i(t)$. The likelihood of the tree given that a single infected individual started the epidemic is then the entry of the vector of probabilities at the root, i.e., $p_7(t_7)$ for which the number of infected individuals $I = 1$.

calculate $p_i(t_i)$ at the end of the i-th time interval given the value of $p_i(t_{i-1})$ at the beginning of the i-th time interval.

At the end of each time interval, the number of tree lineages k_i either decreases by 1 at branching events or increases by 1 at sampling events. At a branching event, the number of tree lineages decreases by 1, so that the vector of probabilities has to be shifted and multiplied with the instantaneous probability that a branching event occurred,

$$p_i(I; t_{i-1}) = 2\lambda(I)\, p_{i-1}(I+1; t_{i-1}). \qquad (3)$$

The factor 2 indicates that either one of the two descendants of the branching event may be the donor-infected individual. At a sampling event, the number of tree lineages increases by 1 and the vector of probabilities is multiplied by the sampling rate ψ,

$$p_i(I; t_{i-1}) = \psi\, p_{i-1}(I-1; t_{i-1}). \qquad (4)$$

These shifts of the vector of probabilities can be summarized in a matrix D_i. We can find the vector of probabilities at the root of the tree by iteratively applying e^{C_i} and D_i until we reach the root at time t_{2n}. This gives us the likelihood of the tree at the root,

$$l(\mathcal{T}) \equiv p_{2n}(t_{2n}) = \left(\prod_{i=1}^{2n} D_i e^{C_i(t_i - t_{i-1})}\right) p_1(t_0). \qquad (5)$$

If we assume that the epidemic was started by a single individual at time t_{2n}, then the likelihood of the tree $\mathcal{L}(\mathcal{T};\theta)$ is the entry of $l(\mathcal{T})$ for which the number of infected individuals $I = 1$ and θ represents the model parameters N,β,μ, and ψ. The vector of initial conditions $p_1(t_0)$ can be chosen to reflect prior knowledge on disease prevalence at the present time. In the absence of any prior information, all prevalence levels are equally likely, $p_1(t_0) = (1,1,\ldots,1)$. Finally, we condition the likelihood $\mathcal{L}(\mathcal{T};\theta)$ on sampling at least one infected individual throughout the epidemic, $\mathcal{L}_0(\theta)$ (see supplementary information, Supplementary Material online). We use the conditioned likelihood $\mathcal{L}(\mathcal{T};\theta)/\mathcal{L}_0(\theta)$ to estimate epidemiological parameters from the sampled phylogenies.

The computation of the above likelihood using traditional matrix multiplication methods suffers from either poor

accuracy or computational intractability. To get around this, we use a novel method that can accurately and efficiently calculate matrix exponentials such as in equation (5) (Al-Mohy and Higham 2011). Our method is available both as an R package expoTree and C function (Leventhal 2013).

Results

Model Validation

We test the validity of our model by simulating 1,000 transmission trees under 11 different parameter combinations using a stochastic SIS model with various population sizes N, recovery rates μ, and sampling rates ψ until n samples are taken. The infection rate $\beta = 1$ in all cases, because it is always possible to rescale time such that $\beta = 1$ without loss of generality. In the density-independent (BD) case, sampling and death rates cannot be independently estimated (Stadler et al. 2013). Because no closed-form solution for the likelihood is available in the density-dependent case, it remains unclear whether all four parameters of the density-dependent model are independent and can be estimated separately (see supplementary methods, section A.7, Supplementary Material online). To account for potential dependence among the parameters in the density-dependent model, we estimate parameters for both an unconstrained (DD) and fixed (DD_{fixed}) value of the sampling probability, $r = \frac{\psi}{\psi + \mu}$.

We then reestimate the parameters for each parameter combination in two ways, illustrating two alternative applications:

(A) We reestimate the parameters for each of the transmission trees using the ML estimator. This corresponds to the case when only a single sampled phylogeny has been reconstructed from the sequence data. The 1,000 parameter estimates can be interpreted as a parametric bootstrap for the confidence interval around the true parameter value.

(B) We estimate the posterior distribution of the parameters for trees #500–590 of the 1,000 simulated transmission trees jointly in a Bayesian framework (see Materials and Methods). This corresponds to the case when multiple independent trees are available with the same underlying stochastic process, for example, independent samples from the posterior distribution of trees estimated using Bayesian Markov chain Monte-Carlo (MCMC) estimation as in BEAST (Drummond and Rambaut 2007).

Method A

The bootstrapped confidence intervals for both the DD_{fixed} and DD model all contain the input parameters (supplementary fig. S1, tables S1 and S2, Supplementary Material online).

The confidence intervals of the DD model fully contain the confidence intervals of the DD_{fixed} model, as expected (except for parameter sets 7 and 9, though the nonoverlap is very small). This is in agreement with previously reported results for the density-independent birth–death model, where μ and ψ only appear as the sum $\mu + \psi$ in the expression for R_0, and the explicit choice of μ given ψ does not change the value of R_0.

Method B

The bias of the posterior mean estimates of the parameters for all 11 parameter sets is listed in table 1. The posterior distribution could not be estimated for parameter sets 2 and 4 in the DD model and for 6 and 7 in both the DD model and DD_{fixed} models, because the Markov chains did not reach a steady state. For these parameter sets, we report ML estimates for the joint likelihood of the subset of trees. In all cases, the density-dependent models provided a better fit to the data on the basis of the deviance information criterion (DIC). For the cases where posterior distributions could not be estimated, we calculated the deviance at the ML estimate. These cannot be directly compared with DIC values, as no correction for complexity has been performed. However, for those cases where DIC values could be computed, the effective number of parameters was generally around 2–4 for the DD_{fixed} model and 4–8 for the DD model. We therefore do not expect the differences in deviance on the order of 10^3 between the BD and DD models for the parameter sets 6 and 7 to be predominantly due to added complexity.

In the DD_{fixed} case, the absolute bias of the estimated parameters is generally less than 5%, with the exception of parameter set 3, where the bias on N was $\xi_N^{(3)} = -0.05$ and $\xi_\beta^{(3)} = 0.06$ on β. For this parameter set, the true values of N,β, and R_0, as well as for R_0 in parameter set 8, fall marginally outside the 95% highest probability density (HPD) interval (see supplementary tables S3 and S4, Supplementary Material online). However, when inferring parameters using different sets of simulated trees, the 95% HPD intervals contain the true parameter (data not shown).

Estimation bias using the posterior mean was generally larger in the DD model compared with the DD_{fixed} model, though the credible intervals were comparatively large and all contained the true parameters, with the exception of ψ and r in parameter set 1. For those parameter sets where $\psi < \mu$ (sets 3, 4, 6, 7, 10, and 11), the bias of the posterior mean estimates was generally small.

In most cases, the density-independent model did not return a posterior HPD interval that contained any of the input parameters, and the posterior means were heavily biased. In the remaining cases, this model was able to correctly estimate μ and ψ but strongly underestimated β. The bias is strongest when N is small (i.e., large saturation effects), because β is estimated as a time-averaged value of the force of infection $\lambda(I) = \beta(1 - I/N)$ (see Materials and Methods), and this time average is smaller than the actual β. This is of particular importance when estimating $R_0 = \beta/(\mu + \psi)$, where strongly underestimating β will result in a strongly underestimated R_0.

We obtain a visual confirmation of the superior fit of the DD and DD_{fixed} models by plotting the lineages-through-time (LTT) of the simulated trees together with the LTT predicted by the estimated parameters (supplementary fig. S2, Supplementary Material online). The density-dependent

Table 1. The Relative Bias of the Estimated Parameters and DIC Values of the Fit.

Tree	Method	N	β	μ	ψ	R_0	DIC	p_V
1 ($n = 100$)	SIM	100	1	0.1	0.4	2		
	DD_{fixed}	−0.02	0.01	−0.01	−0.01	0.02	39,106	3.15
	*DD	0.41	0.31	3.23	−0.23[a]	−0.10	38,822	6.67
	BD_{fixed}	—	−0.29[a]	−0.02	−0.02	−0.28[a]	39,291	1.84
2 ($n = 100$)	SIM	500	1	0.1	0.4	2		
	*DD_{fixed}	−0.04	0.01	−0.02	−0.02	0.03	30,475	3.06
	DD[b]	—	—	—	—	—	—	
	BD_{fixed}	—	−0.08[a]	−0.03[a]	−0.03[a]	−0.04[a]	30,519	2.26
3 ($n = 100$)	SIM	100	1	0.4	0.1	2		
	*DD_{fixed}	−0.05[a]	0.06[a]	−0.01	−0.01	0.06[a]	70,059	2.74
	DD	0.11	0.12	0.15	−0.09	0.02	70,064	6.22
	BD_{fixed}	—	−0.44[a]	−0.01	−0.01	−0.44[a]	71,262	2.19
4 ($n = 100$)	SIM	500	1	0.4	0.1	2		
	DD_{fixed}	−0.04	−0.00	−0.00	−0.00	−0.00	53,637	3.18
	*DD[c]	−0.09	−0.03	−0.07	0.03	0.02	53,630[a]	—
	BD_{fixed}	—	−0.23[a]	−0.03[a]	−0.03[a]	−0.20[a]	541,46	2.06
5 ($n = 1000$)	SIM	100	1	0.1	0.4	2		
	*DD_{fixed}	−0.01	0.00	−0.00	−0.00	0.01	512,449	2.89
	DD	0.01	0.02	0.15	−0.02	0.00	512,454	6.04
	BD_{fixed}	—	−0.48[a]	−0.00	−0.00	−0.48[a]	515,752	1.87
6 ($n = 100$)	SIM	1000	1	0.495	0.005	2		
	*DD_{fixed}[c]	−0.00	−0.01	−0.01	−0.01	0.00	110,267[a]	—
	DD[c]	0.03	0.00	0.01	−0.03	−0.01	110,267[a]	—
	BD_{fixed}	—	−0.43[a]	0.06[a]	0.06[a]	−0.46[a]	115,949	1.84
7 ($n = 100$)	SIM	10000	1	0.495	0.005	2		
	DD_{fixed}[c]	0.04	−0.02	−0.03	−0.03	0.02	70,851[a]	—
	*DD[c]	0.04	−0.02	−0.04	−0.03	0.02	70,851[a]	—
	BD_{fixed}	—	−0.15[a]	−0.03	−0.03	−0.13[a]	72,010	2.26
8 ($n = 100$)	SIM	100	1	0.02	0.08	10		
	*DD_{fixed}	0.01	0.02	−0.02	−0.02	0.04[a]	73,229	2.63
	DD	−0.02	0.01	−0.30	0.01	0.07	73,231	4.46
	BD_{fixed}	—	−0.78[a]	−0.10[a]	−0.10[a]	−0.75[a]	80,921	2.09
9 ($n = 100$)	SIM	500	1	0.02	0.08	10		
	*DD_{fixed}	−0.05	0.01	0.04	0.04	−0.03	45,025	2.71
	DD	−0.03	0.02	0.64	0.03	−0.08	45,029	5.7
	BD_{fixed}	—	−0.35[a]	−0.03	−0.03	−0.33[a]	46,647	2.13
10 ($n = 100$)	SIM	100	1	0.08	0.02	10		
	*DD_{fixed}	−0.00	−0.01	−0.00	−0.00	−0.00	112,519	3.37
	DD	0.03	−0.00	0.03	−0.02	−0.02	112,525	7.42
	BD_{fixed}	—	−0.87[a]	0.05[a]	0.05[a]	−0.88[a]	121,208	1.99
11 ($n = 100$)	SIM	500	1	0.08	0.02	10		
	*DD_{fixed}	0.04	−0.02	−0.03	−0.03	0.01	70,854	3.06
	DD	0.06	−0.01	0.00	−0.03	0.01	70,859	6.23
	BD_{fixed}	—	−0.68[a]	0.12[a]	0.12[a]	−0.71[a]	80,572	1.98

NOTE.—The SIM entries are the input parameters to the simulation. DD_{fixed}, density-dependent model with fixed sampling probability; DD, density-dependent model with inferred sampling probability; BD_{fixed}, density-independent model with fixed sampling probability. n is the number of tips in the tree. The entries at N,β,μ,ψ, and R_0 show the relative bias of the estimates. Smaller DIC values indicate a better fit of the model to the data. The model with the smallest DIC value is indicated by an asterisk for each of the parameter sets.

[a]The HPD interval does not contain the true parameter value (shaded cells).

[b]Numerical maximization of the likelihood failed.

[c]The MCMC method did not converge under the Gelman–Rubin diagnostic. We therefore report ML point estimates and the deviance at the ML estimator. This is not equivalent to a DIC, but must be corrected by $2p_V$, where p_V is the effective number of parameters.

models produce LTTs that more accurately reproduce the LTTs of the input trees compared with the density-independent model, especially when N is small. This is most pronounced when the number of sampled individuals is large compared with the total population size ($n > N$), indicating that the epidemic has been in the endemic equilibrium.

Comparison with Logistic Coalescent

In the parametric coalescent, the decrease in effective population size backward in time from an initial value N_0 is described by an arbitrary function (Griffiths and Tavaré 1994). In the case of an SIS model, the solution to the deterministic equations is a logistic function. We therefore compare our method with a coalescent model with population size governed by a logistic function (see supplementary methods, section A.8, Supplementary Material online),

$$M(t) = \frac{\phi M_0}{(\phi - 1)e^{\rho t} + 1}.$$

The parameters of the logistic function are related to the parameters of the SIS model by $\rho = \beta - (\mu + \psi)$ and $\phi = N/M_0$, where N is the carrying capacity of the model (i.e., the total population size in the SIS model). The per lineage coalescent rate in calendar time is $\vartheta^{-1} = (\mu + \psi)/M_0$ (supplementary methods, section A.8, Supplementary Material online). The three parameters of the logistic coalescent are therefore linked to four parameters of the SIS model and M_0. The input parameters of the trees simulated under the SIS model in Model Validation section can therefore only be converted to ϕ and ϑ of the logistic coalescent by an appropriate choice of M_0. The growth rate ρ can be converted independent of the choice of M_0. In order to compare the parameter estimates from the logistic coalescent to our model, we choose M_0 as the total infected population size in a deterministic SIS model that started with a single infected individual after a time $\hat{t}_{2n}$ has passed (supplementary methods, equation A.6.4, Supplementary Material online), where $\hat{t}_{2n}$ is equal to the mean height of the 1,000 simulated trees. For all parameter sets, the ML estimator is either heavily biased or the 95% confidence intervals contain the input parameter but are extremely large (supplementary table S5, Supplementary Material online). Bayesian MCMC was generally not possible, as the Markov chains never reached a steady state.

HIV

We applied both the DD and DD_{fixed} methods to ten transmission clusters of the Swiss HIV Cohort Study (SHCS) (Kouyos et al. 2010; Schoeni-Affolter et al. 2010). For each of the clusters, 90 trees were sampled from the posterior distribution of trees previously determined using BEAST (Drummond and Rambaut 2007; Stadler et al. 2012). The 90 trees were chosen from the MCMC chain, such that they can be considered independent identically distributed (iid) samples from the posterior distribution of trees. We then estimated the posterior distribution of parameters using method B (see Model Validation and Materials and Methods). In our model, "death" corresponds to any process resulting in an individual becoming noninfectious, and a "dead" individual is assumed to be replaced by a susceptible individual. In the case of the SHCS, individuals mainly become noninfectious when they start antiretroviral treatment, upon which their viral load is suppressed and consequently also onward transmission.

Of all the patients who start treatment in Switzerland, around 75% are included in the SHCS (Schoeni-Affolter et al. 2010). This corresponds to the sampling probability, $r = \psi/(\psi + \mu)$. To account for other reasons for becoming noninfectious, we fix the sampling probability in the DD_{fixed} model at three different values of $r = 0.25$, 0.5, and 0.75.

To test for potential bias in the estimates, we simulated transmission trees under the SIS model using the estimated values for each of the clusters and subsequently reinferred the model parameters for the simulated data. We did not find any evidence of estimator bias for the DD_{fixed} model. For the DD model, however, we detected nonnegligible bias of the estimators. We therefore only report the estimated parameters for the DD_{fixed} model with $r = 0.75$ in table 2 and for $r = 0.25$ and $r = 0.5$ in the supplementary materials, Supplementary Material online.

For clusters 3 and 10, the posterior of N was bounded on top by the upper limit of the uniform prior and the MCMC chain did not converge. This indicates that these clusters have a very large total population size and are more appropriately estimated using the density-independent model (supplementary table S8, Supplementary Material online).

In the remaining eight clusters, the population size estimates vary from small, that is, same order of magnitude as the number of sampled individuals (e.g. $N^{(8)} = 14 = n$), to large, that is, $N >> n$ (e.g., $N^{(6)} \in [61.9, 1780]$). This indicates that there are some clusters where the epidemic has saturated and only few new infections are occurring ($N \approx n$), but also other clusters where new infections are still common ($N >> n$). Similarly, estimates of R_0 range from moderate (e.g., $R_0^{(6)} = 2.80$) to very large (e.g., $R_0^{(5)} = 13.7$).

Using the estimated parameters, we plot the LTT as well as the inferred prevalence within the transmission clusters. Two example clusters are shown in figure 2 and all the clusters in supplementary figures S3 and S4, Supplementary Material online. Visual inspection of these LTT plots confirms that the density-dependent model replicates the distribution of phylogenetic trees better than the density-independent model, with the exception of clusters 3 and 10. Interestingly, for cluster 3, neither model is able to adequately produce an acceptable LTT plot, suggesting that an SIS model is not an appropriate representation of this transmission or that the inferred phylogenetic tree is questionable.

Discussion

This study proposes a method that extends previous work on birth–death models for phylogenetic inference (Etienne et al. 2012; Stadler et al. 2012) and combines both density dependence and longitudinal sampling into a single framework. Previous methods based on the coalescent that can infer

Table 2. Epidemiological Parameter Estimates for the 10 Swiss HIV Transmission Clusters Under the DD_{fixed} Model with $r = 0.75$.

Cluster	1[a]	2[a]	3	4[a]	5[a]
n	34	29	27	26	25
N	188 [160,218]	36.1 [32.2,40.6]	—	39.4 [35.5,43.9]	65.7 [58.7,73]
β	0.269 [0.258,0.282]	0.5 [0.458,0.543]	—	1.02 [0.952,1.08]	0.622 [0.594,0.658]
μ	0.00973 [0.009,0.011]	0.0341 [0.032,0.036]	—	0.0808 [0.076,0.086]	0.0113 [0.01,0.012]
ψ	0.0292 [0.027,0.032]	0.102 [0.096,0.108]	—	0.243 [0.227,0.258]	0.034 [0.031,0.037]
R_0	6.93 [6.29,7.62]	3.67 [3.39,3.94]	—	3.15 [2.88,3.4]	13.7 [12.4,15.3]

Cluster	6[a]	7[a]	8[a]	9[a]	10
n	18	17	14	14	12
N	134 [70.5,1780]	175 [128,256]	14 [14,14.4]	63.8 [53,76]	—
β	0.442 [0.39,0.47]	0.396 [0.372,0.42]	0.916 [0.842,0.983]	0.672 [0.623,0.712]	—
μ	0.0394 [0.035,0.042]	0.0154 [0.014,0.017]	0.0427 [0.04,0.045]	0.0196 [0.017,0.023]	—
ψ	0.118 [0.106,0.127]	0.0463 [0.041,0.051]	0.128 [0.119,0.135]	0.0587 [0.051,0.068]	—
R_0	2.8 [2.37,3.06]	6.42 [5.59,7.36]	5.37 [4.76,5.98]	8.59 [7.17,9.93]	—

NOTE.—n is the number of sampled individuals in each subepidemic. $R_0 = \beta/(\mu + \psi)$ is the basic reproductive ratio. The reported values are the posterior mode and the 95% credible intervals from the posterior distribution. For the uniform prior used, the posterior mode corresponds to the ML estimator and only differed negligibly for the estimate of N in cluster 6. In cluster 6, the posterior distribution of N was heavy-tailed and bounded by the uniform prior $N \in [18,2000]$. Therefore, the upper limit of the credible interval is likely an underestimate.

[a]The fit of the density-dependent model is significantly better than the density-independent model. Model comparison is based on DIC values.

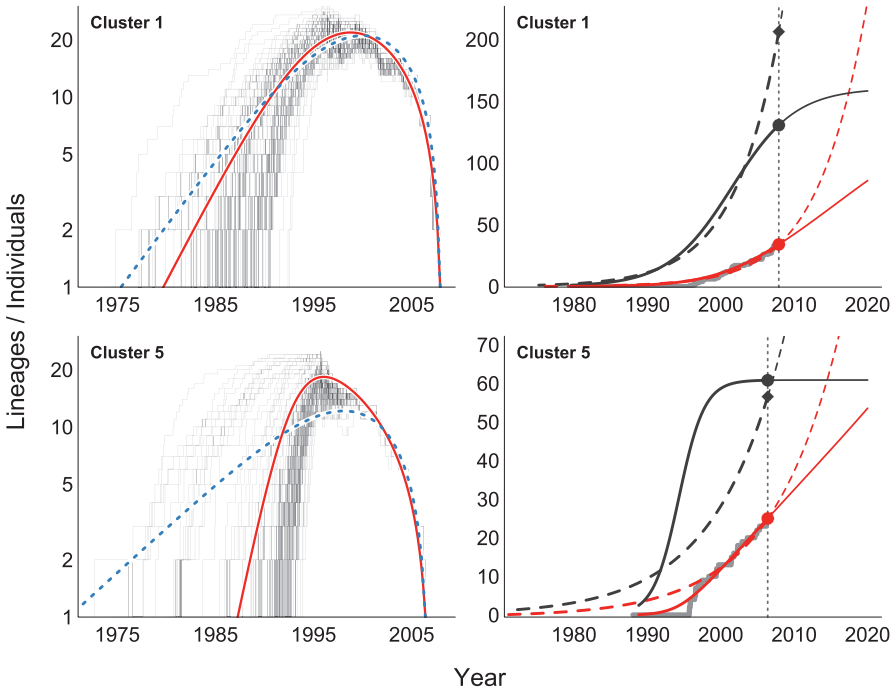

FIG. 2. Lineage through time plots and prevalence curves of two example HIV subepidemics in Switzerland. Left panels: The gray lines are the LTT of the 90 samples from the posterior distribution of trees estimated using BEAST. The solid and dashed lines are the expected number of lineages for the density-dependent and density-independent SIS models, respectively, using the parameter estimates from table 2. Predicted LTT plots are almost identical when using parameter estimates for sampling probabilities $r = 0.25$ and $r = 0.5$. Right panels: Dashed lines correspond to the density-independent (BD) model and solid lines to the density-dependent (DD) model. The vertical dotted line indicates the time of the last sample in the tree. The gray steps are the actual cumulative number of sampled individuals over time and the red curves are the fitted functions. The black lines show the predicted prevalence from the fitted model. The predicted number of infected individuals (black) and cumulative number of sampled individuals (red) for the estimated parameter values. Although both model produce acceptable fits to the cumulative number of samples over time, the BD model predicts both the prevalence and cumulative number of samples to increase exponentially in the future, whereas the DD model can identify subepidemics that are already in the saturated phase.

density-dependent transmission rates rely on a deterministically changing population size and cross-sectionally sampled data (Volz et al. 2009; Frost and Volz 2010; Volz 2012). Other methods based on birth–death models account for longitudinally sampled sequences and stochasticity but neglect density-dependent transmission (Stadler et al. 2012). Alternatively, semiparametric methods such as the various skyline-plot models (Pybus et al. 2001; Ho and Shapiro 2011; Stadler et al. 2013) allow for transmission rates that change over time. These methods, however, require a partitioning of the past into periods of constant population size. Each of these partitions then requires an explicit transmission rate, resulting in a large number of parameters, though simulated trajectories from SIR models can be used to approximate the population sizes within the partitions (Kühnert et al. 2013). Finally, sequential Monte-Carlo methods can be used to estimate parameters (Rasmussen et al. 2011) but are much more computationally intensive than exact likelihood methods.

By using an SIS model, which is the standard epidemiological model for sexually transmitted diseases without immunity (Anderson and May 1991), we can account for density-dependent transmission with only one additional parameter compared with density-independent (constant rate) birth–death models. This allowed us, for the first time, to estimate not only the transmission and removal rates but also the total susceptible population size of transmission groups from viral sequences. The parametrized SIS model can predict how the number of infected and susceptible individuals will vary over time. Thus, if an SIS model is a good representation of an ongoing epidemic, it is possible to make predictions of how the epidemic will continue to develop.

Furthermore, we have shown that in principle, our method can estimate sampling and death rates independently in some cases, but that this signal is generally weak and estimates obtained without prior knowledge of the sampling probability must be carefully examined for potential biases. Generally, estimator bias for the DD model is small in those simulated data sets where the epidemic reached an endemic equilibrium well before all samples were taken, such that a large part of the samples are from the saturated phase.

Our method is able to estimate parameters with higher accuracy compared with parametric coalescent-based inference with deterministically changing population sizes. This is especially true when the epidemic has already reached the saturated phase. In the deterministic SIS model, the change in the number of infected individuals is described by a logistic function, which can be used to model the population size decline backward in time in a parametric coalescent (Griffiths and Tavaré 1994). The parameters of the parametric coalescent with logistic population decline depend on the effective population size at present, that is, the total number of infected individuals at present. In this context, an important difference between the stochastic SIS model and the deterministic SIS model becomes apparent. In the deterministic model, the infected population size asymptotically approaches the endemic equilibrium. In contrast, in the stochastic model, the process reaches a quasi steady-state, in which the infected population size fluctuates stochastically around an equilibrium value. This means that although the stochastic SIS model can account for processes that have reached a quasi steady-state, the deterministic model interprets small deviations from the asymptotic equilibrium as a signal of how long the epidemic has been in a state close to this equilibrium. This can lead to incorrect parameter estimates.

We have demonstrated that using an epidemiological model together with phylogenetic inference can indeed lead to new insights into ongoing epidemics. Susceptible population sizes in 8 out of 10 transmission clusters of the SHCS were estimated to be small. This is an indication that the subepidemics in these transmission clusters are characterized by an initial rapid spread that subsequently slows down after only a relatively small number of infections. Such a scenario is conceivable when the population is composed of many small susceptible subgroups, and transmission event between subgroups are much rarer than within subgroups. This is compatible with previous findings which showed that HIV epidemics are driven by heterogeneous populations, such as heterogeneity in infection rates or heterogeneity in number of contacts (Liljeros et al. 2001; Kouyos et al. 2010; Leventhal et al. 2012; Stadler and Bonhoeffer 2013).

Differences between transmission clusters are further reflected in estimates of the basic reproductive number, R_0 (Anderson and May 1979). Because the estimate of the susceptible population size is small, the estimated R_0 is larger than previously reported for the density-independent case (Stadler et al. 2012). The reason for the previous underestimation of R_0 is the following: The basic reproductive number is defined as the number of secondary infections caused by an infectious individual in a completely susceptible population. If the pool of susceptible individuals is very large, then R_0 is roughly equal to the number of secondary infections caused by any infectious individual, $R_1(t)$. The density-independent model assumes that $R_1(t)$ is constant throughout the whole epidemic, and it is the time-averaged value of $R_1(t)$ which is estimated by the method. In the early stages of the epidemic, the number of infected individuals grows exponentially, such that $R_1(t)$ is roughly constant and approximately equal to R_0. Thus, when the sampled sequences are confined to the early stages of the epidemic, then the estimate of R_1 using a density-independent model is an acceptable approximation for R_0. As the number of susceptible individuals decreases later on in the epidemic, so does $R_1(t)$ and consequently the time average of $R_1(t)$ is an underestimation of R_0. The new density-dependent model takes the decrease in $R_1(t)$ over time into account, such that more accurate estimates of R_0 can be obtained when sequences are available from all periods of the epidemic.

The underestimation of R_0 is most extreme when the cluster was saturated while most of the individuals were sampled (e.g., SHCS cluster 5: $R_0^{DD} = 13.7, R_0^{\infty} = 2.67$). In this cluster, the initial spread proceeded rapidly, after which

only few infections happened (see fig. 2). LTT plots can be used to get a visual confirmation that the density-dependent model does indeed provide a better fit to the data. Alternatively, when most of the individuals are sampled from the initial phases of the epidemic, both the density-dependent and -independent models give similar estimates of R_0 and the number of lineages in the past (see fig. 2). The estimated total number of infected individuals at the present, however, will then be larger in the density-independent model than in the density-dependent model.

The parameterized SIS model can be used to make predictions on how the epidemic is expected to progress within these clusters. This is in contrast to skyline-plot methods, which can only recreate the history of transmission rate changes in the past unless a time-dependent transmission model is fit to these skyline-plots.

The accuracy of our model predictions will strongly depend on the validity of the assumptions. Here, we assume an SIS model where the rate of becoming noninfectious is equal to the recruitment rate of new susceptibles. This approximation is mainly performed for computational tractability, because the total population size, that is, $I + S$, remains constant over time. Although this assumption is clearly violated for diseases where the expected time to becoming noninfectious is much shorter than the recruitment time of new susceptibles (e.g., influenza), it is plausible when the two processes happen on a similar timescale.

It is possible to extend our method to account for any kind of compartmental epidemiological models, such as a susceptible-infected-recovered (SIR) model. In its present form, however, this extension would come with a significant computational cost, because the number of recovered individuals would need to be tracked in addition to the number of infected individuals. The number of differential equations that would then need to be solved increases from N to N^2. In its present form, this would only be conceivable when the population size is moderate.

The sampling probability, that is, the ratio of sampling rate to total death rate, $r = \psi/(\psi + \mu)$, cannot be estimated using the density-independent model. In fact, it can be shown that for the density-independent model the likelihood function only depends on two out of the three parameters β,μ, and ψ, namely on the net growth rate $\beta - \mu - \psi$ and the product of transmission and sampling rate, $\beta\psi$. For the density-dependent model, we could not show or disprove such a decrease in degrees of freedom of parameter space. However, we observed from our simulation study that it is possible to estimate the sampling probability for certain parameter combinations, though care must be taken to control for potential biases as the signal for r is weak. In particular, the choice of r only has little effect on LTT plots predicted using the estimated parameters, meaning that different r explain the data equally well (recall that our data are essentially a LTT plot). Furthermore, the inferred cumulative number of sampled individuals is not significantly influenced by r, another indication that different r explain the data equally well. Thus, knowledge obtained from other data should be used to supply a prior probability on r. It is important to note that the choice of r does influence the quantitative value of R_0 as well as the estimated current number of infected individuals.

Overall, we have presented a method that is readily available both as an R package and C++ stand-alone programs (Leventhal 2013). The method can be applied to transmission trees inferred from pathogen sequence data in order to obtain better estimates of epidemiological parameters such as R_0, thus providing better insight into the transmission dynamics of SIS-type epidemics. Additionally, when information on the sampling probability is available from other data sources, reliable estimates of the size of the current infected population can be obtained. In summary, by applying our method to pathogen sequence data, we can obtain a better understanding of the intensity of transmission within different transmission clusters, which can help guide and assess public health intervention measures.

Materials and Methods

SIS Model

A common way of modeling the spread of a disease through a population is with the SIS model (Kermack and McKendrick 1927; Anderson and May 1991). Individuals can either be in a susceptible state S or an infected state I. In this article, we use a stochastic SIS model but motivate the choice of the model using a deterministic SIS model.

Deterministic SI/SIS Model

The change in number of susceptible and infected individuals over time can be written as a set of ODEs,

$$\frac{dS}{dt} = -\beta SI/N + \Phi(S,I), \tag{6}$$

$$\frac{dI}{dt} = \beta SI/N - \gamma I. \tag{7}$$

Here, β/N is the infection rate per infectious contact and μ is the death/removal rate of infected individuals. The recruitment of new susceptible individuals is given by the arbitrary function $\Phi(S,I)$. We assume that the total number of individuals in the population, N, is constant. In this case, $\frac{dI}{dt} + \frac{dS}{dt} = 0$, such that $\phi(S,I) = \gamma I$ and the SI model is equivalent to a SIS model (Anderson and May 1991). The force of infection Λ is the rate at which susceptible individuals become infected and is proportional to the number infected individuals in the population,

$$\Lambda(I) \equiv \beta I/N. \tag{8}$$

Using $S = N - I$, we can rewrite equation (7),

$$\frac{dI}{dt} = (\beta(1 - I/N) - \gamma)I. \tag{9}$$

Stochastic SIS Model

We use a continuous-time Markov chain (CTMC) to model the epidemic process and the induced sampled phylogenetic trees. We define $q_I(t)$ as the probability that I individuals

are infected at time t. The transition probabilities for the CTMC process are,

$$\text{Prob}\{I \to I+1\} = (\beta/N)(N-I)Idt, \quad (10a)$$

$$\text{Prob}\{I \to I-1\} = \gamma Idt, \quad (10b)$$

$$\text{Prob}\{I \to I\} = 1-(\beta(N-I)I/N+\gamma I)dt. \quad (10c)$$

We can write down the Kolmogorov forward differential equation for $q_I(t)$,

$$\frac{dq_I}{dt} = q_{I-1}(t)\beta(N-(I-1))(I-1)/N + q_{I+1}(t)\gamma(I+1) - (\beta(N-I)I/N+\gamma)q_I(t). \quad (11)$$

The solution to equation (11) gives us the probability of having I infected individuals at time t. The deterministic SIS model can be used as an upper bound for the expected value of the number of infected individuals at time t (Allen 2008),

$$\frac{d\mathbb{E}[I(t)]}{dt} \leq (\beta(1-\mathbb{E}[I(t)]/N)-\gamma)\mathbb{E}[I(t)]. \quad (12)$$

Sampled Phylogenies of an Epidemic Outbreak

In surveillance data, information about the infectious state is only available for a subset of individuals. We assume that throughout the course of the epidemic, the infected individuals are sampled at a constant rate ψ. Once they are sampled, we assume that these individuals can no longer infect anyone else and are removed. This is an appropriate assumption for many diseases where sampling is usually linked to drug treatment, isolation, or behavior change, after which transmission becomes unlikely (e.g., HIV) or recovery is rapid. The sampled transmission tree (*sampled phylogeny*) results from disregarding all non-sampled individuals from the complete transmission tree (fig. 1). As we assume that sampled individuals are no longer infectious, the removal rate γ in equations (9) and (10b) becomes $\gamma = \mu + \psi$, where μ is the removal rate without sampling and ψ is the removal rate with sampling. We define r as the probability of being sampled upon removal, $r = \psi/(\mu+\psi)$.

Inferring Epidemiological Parameters from Sampled Phylogenies

Our aim is to infer the parameters of the stochastic SIS model, β,μ,ψ,N, defined by equation (11), based on a sampled phylogeny, $\mathcal{T}$. In the supplementary information, Supplementary Material online, we derive the likelihood $\mathcal{L}(\mathcal{T};\theta)$ that an SIS model with parameters $\theta = (\beta,\mu,\psi,N)$ gave rise to the sampled phylogeny.

Parameter inference

We use the likelihood function to obtain parameter estimates in two ways: (A) a ML framework; (B) by estimating the posterior distribution of parameters in a Bayesian framework. In the Bayesian framework, we perform a Metropolis–Hastings (MH) MCMC estimation of the joint likelihood of all the sampled phylogenies to obtain a posterior distribution of the parameters (Metropolis et al. 1953; Hastings 1970). Let $\mathbb{T} = \{\mathcal{T}_1,\mathcal{T}_2,\ldots,\mathcal{T}_m\}$ be a set of iid samples chosen from the distribution $P(\mathcal{T}\mid\theta)$ of sampled phylogenies. The probability density of a tree is the likelihood of the tree from the New Approaches section. Because all the trees are assumed to be iid, the likelihood of $\mathbb{T}$ is the product of the likelihoods of the individual trees,

$$\mathcal{L}(\mathbb{T};\theta) = \prod_{i=1}^{m}\mathcal{L}(\mathcal{T}_i;\theta). \quad (13)$$

The full conditionals are not known, and we need to resort to a MH approach to sample from the posterior distribution of $\mathcal{L}$. Furthermore, the parameters N,β,μ, and ψ are highly correlated, which greatly increases the time to convergence of the MCMC chain when using traditional MH or sequential Gibbs sampling. We thus use Differential Evolution Adaptive Metropolis (DREAM) to estimate the posterior density (Vrugt et al. 2009). Convergence in the DREAM scheme is determined via the Gelman–Rubin convergence diagnostic and is reached when the scale reduction factor $R_c < 1.05$ for all parameters (Brooks and Gelman 1998).

Relative Bias

To determine how well our method estimates the epidemic parameters, we look at the relative bias,

$$\xi(\hat{\theta}_i) = \frac{\hat{\theta}_i - \theta_i}{\theta_i}, \quad (14)$$

where θ_i is the true value of the i-th parameter $\theta = (N,\beta,\mu,\psi)$ and $\hat{\theta}_i$ is the mean of the estimated posterior.

Supplementary Material

Supplementary methods, figures S1–S4, and tables S1–S8 are available at *Molecular Biology and Evolution* online (http://www.mbe.oxfordjournals.org/).

Acknowledgments

The authors thank the patients who participate in the SHCS; the physicians and study nurses for excellent patient care; the resistance laboratories for high-quality genotypic drug resistance testing; SmartGene, Zug, Switzerland, for technical support; Brigitte Remy, Martin Rickenbach, F. Schoeni-Affolter, and Yannick Vallet from the SHCS Data Center in Lausanne for data management; and Daniele Perraudin and Mirjam Minichiello for administrative assistance. The members of the Swiss HIV Cohort Study are Aubert V, Barth J, Battegay M, Bernasconi E, Böni J, Bucher HC, Burton-Jeangros C, Calmy A, Cavassini M, Egger M, Elzi L, Fehr J, Fellay J, Francioli P, Furrer H (Chairman of the Clinical and Laboratory Committee), Fux CA, Gorgievski M, Günthard H (President of the SHCS), Haerry D (deputy of "Positive Council"), Hasse B, Hirsch HH, Hirschel B, Hösli I, Kahlert C, Kaiser L, Keiser O, Kind C, Klimkait T, Kovari H, Ledergerber B,

Martinetti G, Martinez de Tejada B, Metzner K, Müller N, Nadal D, Pantaleo G, Rauch A (Chairman of the Scientific Board), Regenass S, Rickenbach M (Head of Data Center), Rudin C (Chairman of the Mother & Child Substudy), Schmid P, Schultze D, Schöni-Affolter F, Schüpbach J, Speck R, Taffe P, Tarr P, Telenti A, Trkola A, Vernazza P, Weber R, and Yerly S.

G.E.L., T.S., and S.B. thank the Swiss Federal Institute of Technology, Zurich (ETHZ). T.S. thanks the Swiss National Science foundation for funding (SNF grant #PZ00P3 136820). S.B. is supported in part by the European Research Council under the 7th Framework Programme of the European Commission ("PBDR": Grant Agreement Number 268540).

The data used in this study have been financed in the framework of the Swiss HIV Cohort Study, supported by the Swiss National Science Foundation (SNF grant #33CS30-134277) and the SHCS projects #470, 528, 569, the SHCS Research Foundation, the Swiss National Science Foundation (grant #324730-130865 to H.F.G.), the European Community's Seventh Framework Program (grant FP7/2007-2013), under the Collaborative HIV and Anti-HIV Drug Resistance Network (CHAIN; grant 223131 to H.F.G.), by the Yvonne_Jacob Foundation (to H.F.G.) and by a further research grant of the Union Bank of Switzerland, in the name of an anonymous donor to H.F.G., an unrestricted research grant from Gilead, Switzerland, to the SHCS research foundation, and by the University of Zurich's Clinical Research Priority Program (CRPP) "Viral infectious diseases: Zurich Primary HIV Infection Study" (to H.F.G.). The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

References

Allen L. 2008. An introduction to stochastic epidemic models. *Lect Notes Math.* 1945:81–130.

Al-Mohy AH, Higham NJ. 2011. Computing the action of the matrix exponential, with an application to exponential integrators. *SIAM J Sci Stat Comp.* 33:488–511.

Anderson R, May R. 1991. Infectious diseases of humans: dynamics and control. Oxford: Oxford University Press.

Anderson RM, May RM. 1979. Population biology of infectious diseases: part i. *Nature* 280:361–367.

Brooks S, Gelman A. 1998. General methods for monitoring convergence of iterative simulations. *J Comput Graph Stat.* 7: 434–455.

Drummond A, Rambaut A. 2007. BEAST: Bayesian evolutionary analysis by sampling trees. *BMC Evol Biol.* 7:214.

Drummond AJ, Nicholls GK, Rodrigo AG, Solomon W. 2002. Estimating mutation parameters, population history and genealogy simultaneously from temporally spaced sequence data. *Genetics* 161: 1307–1320.

Drummond AJ, Pybus OG, Rambaut A, Forsberg R, Rodrigo AG. 2003. Measurably evolving populations. *Trends Ecol Evol.* 18: 481–488.

Drummond AJ, Rambaut A, Shapiro B, Pybus OG. 2005. Bayesian coalescent inference of past population dynamics from molecular sequences. *Mol Biol Evol.* 22:1185–1192.

Etienne RS, Haegeman B, Stadler T, Aze T, Pearson PN, Purvis A, Phillimore AB. 2012. Diversity-dependence brings molecular phylogenies closer to agreement with the fossil record. *Proc R Soc B.* 279:1300–1309.

Frost SDW, Volz EM. 2010. Viral phylodynamics and the search for an 'effective number of infections'. *Philos Trans R Soc Lond B Biol Sci.* 365:1879–1890.

Grenfell B, Pybus O, Gog J, Wood J, Daly J, Mumford J, Holmes E. 2004. Unifying the epidemiological and evolutionary dynamics of pathogens. *Science* 303:327–332.

Griffiths RC, Tavaré S. 1994. Sampling theory for neutral alleles in a varying environment. *Philos Trans R Soc Lond B Biol Sci.* 344: 403–410.

Hastings WK. 1970. Monte Carlo sampling methods using Markov chains and their applications. *Biometrika* 57:97–109.

Ho SYW, Shapiro B. 2011. Skyline-plot methods for estimating demographic history from nucleotide sequences. *Mol Ecol Resour.* 11:423–434.

Kendall D. 1948. On the generalized birth-and-death process. *Ann Math Stat.* 19:1–15.

Kermack W, McKendrick A. 1927. A contribution to the mathematical theory of epidemics. *Proc R Soc A.* 115:700–721.

Kingman JFC. 1982. The coalescent. *Stoch Proc Appl.* 13:235–248.

Koelle K, Rasmussen DA. 2011. Rates of coalescence for common epidemiological models at equilibrium. *J R Soc Interface.* 9:997–1007.

Kouyos RD, von Wyl V, Yerly S, et al. (19 co-authors). 2010. Molecular epidemiology reveals long-term changes in hiv type 1 subtype b transmission in switzerland. *J Infect Dis.* 201: 1488–1497.

Kühnert D, Stadler T, Vaughan TG, Drummond AJ. 2013.. Simultaneous reconstruction of evolutionary history and epidemiological dynamics from viral sequences with the birth-death SIR model. *arXiv* 1308.5140.

Leventhal G. 2013. expoTree—R package and C++ code. [cited 2013 Oct 11]. Available from: http://www.leventhal.ch/software/.

Leventhal GE, Kouyos R, Stadler T, von Wyl V, Yerly S, Böni J, Cellerai C, Klimkait T, Günthard HF, Bonhoeffer S. 2012. Inferring epidemic contact structure from phylogenetic trees. *PLoS Comput Biol.* 8: e1002413.

Liljeros F, Edling CR, Amaral LAN, Stanley HE, Aberg Y. 2001. The web of human sexual contacts. *Nature* 411:907–908.

Metropolis N, Rosenbluth AW, Rosenbluth MN, Teller AH, Teller E. 1953. Equation of state calculations by fast computing machines. *J Chem Phys.* 21:1087–1092.

Nee S. 2006. Birth-death models in macroevolution. *Annu Rev Ecol Evol. S.* 37:1–17.

Pybus OG, Charleston MA, Gupta S, Rambaut A, Holmes EC, Harvey PH. 2001. The epidemic behavior of the hepatitis C virus. *Science* 292: 2323–2325.

Pybus OG, Rambaut A. 2009. Evolutionary analysis of the dynamics of viral infectious disease. *Nat Rev Genet.* 10:540–550.

Rasmussen DA, Ratmann O, Koelle K. 2011. Inference for nonlinear epidemiological models using genealogies and time series. *PLoS Comput Biol.* 7:e1002136.

Rohani P, Keeling MJ, Grenfell BT. 2002. The interplay between determinism and stochasticity in childhood diseases. *Am Nat.* 159: 469–481.

Schoeni-Affolter F, Ledergerber B, Rickenbach M, Rudin C, Gunthard HF, Telenti A, Furrer H, Yerly S, Francioli P. 2010. Cohort profile: the Swiss HIV Cohort study. *Int J Epidemiol.* 39: 1179–1189.

Stadler T. 2010. Sampling-through-time in birth–death trees. *J Theor Biol.* 267:396–404.

Stadler T, Bonhoeffer S. 2013. Uncovering epidemiological dynamics in heterogeneous host populations using phylogenetic methods. *Philos Trans R Soc Lond B Biol Sci.* 368:20120198.

Stadler T, Kouyos R, von Wyl V, et al. (14 co-authors). 2012. Estimating the basic reproductive number from viral sequence data. *Mol Biol Evol.* 29:347–357.

Stadler T, Kühnert D, Bonhoeffer S, Drummond A. 2013. Birth–death skyline plot reveals temporal changes of epidemic spread

in HIV and hepatitis C virus (HCV). *Proc Natl Acad Sci U S A*. 110: 228–233.

Volz EM. 2012. Complex population dynamics and the coalescent under neutrality. *Genetics* 190:187–201.

Volz EM, Pond SLK, Ward MJ, Brown AJL, Frost SDW. 2009. Phylodynamics of infectious disease epidemics. *Genetics* 183:1421–1430.

Vrugt J, Ter Braak C, Diks C, Robinson B, Hyman J, Higdon D. 2009. Accelerating Markov chain Monte Carlo simulation by differential evolution with self-adaptive randomized subspace sampling. *Int J Nonlin Sci Num*. 10:273–290.

Wilkinson DJ. 2011. Stochastic modelling for systems biology, Vol. 44. Boca Raton (FL): CRC Press.