Estimating the basic reproductive number from viral sequence data

Stadler, Tanja; Kouyos, Roger; von Wyl, Viktor; Yerly Ferrillo, Sabine; Böni, Jürg; Bürgisser, Philippe; Klimkait, Thomas; Joos, Beda; Rieder, Philip Alexander; Xie, Dong; Günthard, Huldrych F; Drummond, Alexei J; Bonhoeffer, Sebastian

Estimating the Basic Reproductive Number from Viral Sequence Data

Tanja Stadler,[1,*] Roger Kouyos,[1] Viktor von Wyl,[2] Sabine Yerly,[3] Jürg Böni,[4] Philippe Bürgisser,[5] Thomas Klimkait,[6] Beda Joos,[2] Philip Rieder,[2] Dong Xie,[7] Huldrych F. Günthard,[2] Alexei J. Drummond,[7] Sebastian Bonhoeffer,[1] and the Swiss HIV Cohort Study

[1]Institute of Integrative Biology, Eidgenössische Technische Hochschule (ETH) Zürich, Zürich, Switzerland
[2]Division of Infectious Diseases and Hospital Epidemiology, University Hospital Zürich, Zürich, Switzerland
[3]Laboratory of Virology and AIDS Center, Geneva University Hospital, Geneva, Switzerland
[4]Swiss National Center for Retroviruses, Institute of Medical Virology, University of Zürich, Zürich, Switzerland
[5]Service of Immunology and Allergy, Lausanne University Hospital, Lausanne, Switzerland
[6]Department of Biomedicine, Institute of Medical Microbiology, University of Basel, Basel, Switzerland
[7]Allan Wilson Centre for Molecular Ecology and Evolution, University of Auckland, Auckland, New Zealand

***Corresponding author:** E-mail: tanja.stadler@env.ethz.ch.

Associate editor: Daniel Falush

Abstract

Epidemiological processes leave a fingerprint in the pattern of genetic structure of virus populations. Here, we provide a new method to infer epidemiological parameters directly from viral sequence data. The method is based on phylogenetic analysis using a birth–death model (BDM) rather than the commonly used coalescent as the model for the epidemiological transmission of the pathogen. Using the BDM has the advantage that transmission and death rates are estimated independently and therefore enables for the first time the estimation of the basic reproductive number of the pathogen using only sequence data, without further assumptions like the average duration of infection. We apply the method to genetic data of the HIV-1 epidemic in Switzerland.

Key words: epidemiology, phylogenetics, Bayesian inference.

Introduction

RNA viruses are characterized by short generation time and high mutation rates. Therefore, even over relatively short time spans epidemiological processes (i.e., transmission, recovery, and death) are expected to leave signals in the genetic structure of viral sequences sampled from the host population. Bayesian phylogenetic methods are commonly used for viruses to infer epidemiological processes from genetic data (Pybus et al. 2001; Drummond et al. 2003; Grenfell et al. 2004; Pomeroy et al. 2008). These methods require the specification of a process that generates the phylogenetic trees, which is commonly the coalescent (Kingman 1982; Griffiths and Tavare 1994; Drummond et al. 2002).

The coalescent is an appropriate choice for many applications. However, in the context of virus transmission, it is primarily used for its mathematical convenience rather than the accurate reflection of the underlying transmission process. In particular, there are two main shortcomings of the coalescent as a model of epidemiological transmission. First, the coalescent can detect changes over time in the number of infected people (i.e., population size changes) but not whether such changes are due to a change in transmission rates or a change in death or recovery rates. However, the separate estimation of transmission and death/recovery is key in order to determine central epidemiological quantities such as the basic reproductive number (Anderson and May 1979, 1992). Thus, if the coalescent is used in a phylogenetic study, an independent estimate of the death/recovery rate (or the transmission rate) is required for estimating quantities such as the basic reproductive number. Second, the coalescent is appropriate only if the number of sampled infected hosts is small compared with the total infected host population size. This is a problematic assumption for important diseases such as HIV, where we have particularly dense sampling.

A more appropriate choice for the tree-generating process is the birth–death model (BDM; Kendall 1948; Feller 1968). This model has neither of the two shortcomings of the coalescent. It explicitly assumes a separate transmission (i.e., birth) and death rate. Moreover, it can be applied to situations of sparse or dense sampling because sampling proportion is treated in the model as a separate parameter.

The BDM has been used in phylogenetic analysis for inferring species phylogenies (Patané et al. 2009; Couvreur et al. 2010; Fernández-Mazuecos and Vargas 2010; Weksler et al. 2010), including scenarios of incomplete sequence sampling. However, the sampled sequences were assumed to be from one point in time (Maddison et al. 2007; FitzJohn et al. 2009; Stadler 2009). In viral epidemics, sequences are typically sampled over a long time span. Hence, after an initial application of the BDM to viral sequences from one single time point (Holmes et al. 1995), focus shifted towards assuming the coalescent (Nee et al. 1995), as coalescent-based

methodology became available allowing for sequences being sampled sequentially through time (Rambaut 2000), allowing for changing population sizes through time (Pybus et al. 2000; Drummond et al. 2005), and accounting for phylogenetic uncertainty by using a Bayesian framework (Drummond et al. 2002).

Here, we develop a Bayesian method of phylogenetic analysis for sequentially sampled viral sequence data based on the BDM in order to infer key epidemiological parameters directly from viral sequence data. To this end, we extend previous work (Stadler 2010) to solve the BDM for sequentially sampled viral sequences. We generalize the BDM such that it specifically reflects aspects of viral transmission, determine the accuracy in parameter inference using simulations, and in particular, show that the BDM estimates are more accurate than the coalescent estimates. The basic reproductive number can be estimated very accurately using the BDM, whereas the transmission rate correlates with the death rate and its estimate is thus less accurate. We apply the new BDM method to data from the HIV-1 epidemic in Switzerland and further validate the accuracy of estimates on the basis of other epidemiological estimates.

Materials and Methods

Framework for the Inference of Epidemiological Parameters

In its most general form, a birth–death process is a stochastic description of populations, in which individuals can be born or die at any time point. We use this framework to describe the process of epidemiological transmission. A birth event corresponds to the infection of an individual. A death event corresponds to an individual becoming noninfectious, which can be due to several events such as death, treatment, or behavioral changes of an individual.

We model an epidemic as follows: An infected individual starts a new epidemic at time t_{or} in the past. Each infected individual transmits with a constant rate λ (transmission rate) and becomes noninfectious with a constant rate μ (becoming noninfectious rate). To capture the process of infected individuals being sampled (i.e., included into the data set), we introduce a sampling rate ψ. As sampling in human infectious diseases is typically linked to treatment or behavioral changes, we assume that a sampled individual becomes noninfectious immediately after the sampling. Thus, the term ψ is formally equivalent to a death term.

Our BDM is a forward in time description of the epidemiological process, and the sampling rate ψ allows to specify the sampling intensity; thus, the number and time of sampling points is a random outcome of the process. In contrast, the coalescent is a model backward in time where the number of samples is assumed to be very small compared with the population size (however, the precise sampling fraction cannot be specified), and the analysis is conditioned on the number and time of the sampling points.

Bayesian BDM Method

In Bayesian phylogenetic inference using a Markov chain Monte Carlo (MCMC) approach, the idea is to sample trees and parameters from the posterior distribution $f[\mathcal{T},\eta,\theta|\text{data}]$ where "data" are alignment of sequences sampled through time, θ are the parameters of the sequence evolution model, η are the parameters of the tree-generating model, and $\mathcal{T}$ is the transmission tree describing the epidemiological relationships of the sampled sequences in the alignment. By Bayes' theorem, the posterior distribution is equivalent to

$$f[\mathcal{T},\eta,\theta|\text{data}] = \frac{f[\text{data}|\mathcal{T},\theta]f[\mathcal{T}|\eta]f[\eta]f[\theta]}{f[\text{data}]}.$$

The quantity $f[\text{data}|\mathcal{T},\theta]$ is the probability density of the sequences having evolved on a tree $\mathcal{T}$. This likelihood can be computed efficiently with Felsenstein's pruning algorithm (Felsenstein 2004). $f[\mathcal{T}|\eta]$ is the probability density of the tree given the tree-generating model parameters. This density is known if the coalescent is assumed as the tree-generating model. A prior distribution is assumed for the probability densities of the parameters, $f[\eta]$ and $f[\theta]$. The quantity $f[\text{data}]$ is the normalizing constant and can be disregarded when sampling from the posterior (it is constant for all trees and parameters as the data are fixed).

In our approach, we assume the BDM instead of the coalescent as a tree-generating model, that is, $\eta = (\lambda,\mu,\psi,t_{or})$. The BDM generates a transmission tree. The "sampled tree" $\mathcal{T}$ is obtained from the transmission tree by suppressing all edges without sampled descendants, see figure 1 and Stadler (2010) for more details. In order to do Bayesian phylogenetic inference using an MCMC approach, the probability density of a sampled tree under the BDM, $f[\mathcal{T}|\lambda,\mu,\psi,t_{or}]$, has to be derived.

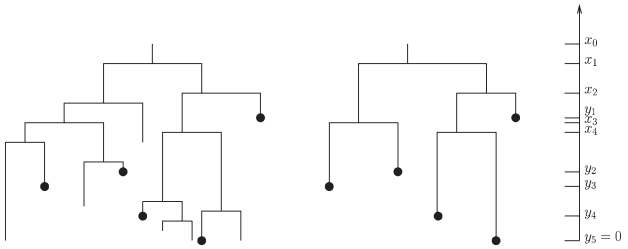

FIG. 1. Example of a transmission tree on the left with the black dots being the sampled individuals ($m = 5$); the corresponding sampled tree is displayed on the right. The time of origin is $t_{or} = x_0$.

Calculation of $f[\mathcal{T}|\lambda,\mu,\psi,t_{or}]$

For calculating the probability density of a sampled tree, we need some notation (see also fig. 1). Let the sampled tree $\mathcal{T}$ have m sampled leaves. Let $x_0 = t_{or}$ be the time of origin of the process. Let $x_1 > \cdots > x_{m-1}$ be the $m-1$ bifurcation times in the sampled tree where time is measured as the distance to the present. Let $y_1 > \cdots > y_m$ be the m sampling times (i.e., times of the leaves).

In order to calculate $f[\mathcal{T}|\lambda,\mu,\psi,t_{or}]$, we define $g_e(t)$ to be the probability density that the infectious individual corresponding to edge e at time t in the past gives rise to the transmission tree between t and the present as observed in $\mathcal{T}$. Then, $g_e(t_{or}) = f[\mathcal{T}|\lambda,\mu,\psi,t_{or}]$ (see fig. 2). In order to

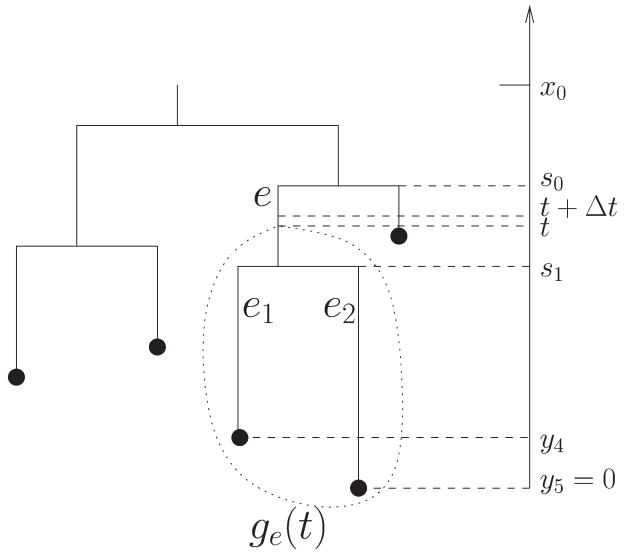

FIG. 2. Illustrating the derivation of the Master equation for $g_e(t)$. e is an edge of the tree, and $g_e(t)$ is the probability density that the infected individual corresponding to edge e at time t gives rise to the observed transmission tree, that is, a tree with two sampled individuals, at time y_4 and y_5, and the transmission at time s_1. The Master equation for $g_e(t)$ is derived via going small time steps Δt back in time.

derive a formula for $g_e(t)$, we need the probability that an infectious individual has no sampled descendants for a time span of length t, $p_0(t)$. The probability $p_0(t)$ is calculated in Stadler (2010):

$$p_0(t) = \frac{\lambda+\mu+\psi+c_1\frac{e^{-c_1 t}(1-c_2)-(1+c_2)}{e^{-c_1 t}(1-c_2)+(1+c_2)}}{2\lambda},$$

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

$$c_2 = -\frac{\lambda-\mu-\psi}{c_1}.$$

We derive the probability density for $g_e(t)$ with a Master equation approach. Let Δt be a very small time step. An event (in our case, transmission, becoming noninfectious, or sampling) happening with a rate α means that in a small time interval Δt, the probability of the event happening once is $\alpha\Delta t$; the event happening several times has probability of order $O(\Delta t^2)$ (here, we do not calculate the exact probability of several events happening during one time interval, as this probability will tend to 0 and thus cancel out, as shown below).

We calculate the probability density $g_e(t+\Delta t)$, that is, the probability density that the individual e at time $t+\Delta t$ gives rise to the the transmission tree between $t+\Delta t$ and the present as observed in $\mathcal{T}$, assuming we know $g_e(t)$ (see fig. 2). Recall that time is measured as a distance to the present, so with an additional time step Δt, we move further into the past, and the tree likelihood is thus calculated going backward in time. The infected individual at time $t+\Delta t$ can either undergo no event or transmit during time Δt (becoming noninfectious would not yield the observed transmission tree). The equation for $g_e(t+\Delta t)$ is therefore

$$g_e(t+\Delta t) = (1-(\lambda+\mu+\psi)\Delta t - O(\Delta t^2))g_e(t) + \lambda\Delta t 2p_0(t)g_e(t) + O(\Delta t^2),$$

where 1) $1-(\lambda+\mu+\psi)\Delta t - O(\Delta t^2)$ is the probability that the individual corresponding to edge e at time $t+\Delta t$ does not undergo an event during time Δt, 2) $\lambda\Delta t$ is the probability that the individual corresponding to edge e infects one individual during time Δt, and $2p_0(t)$ is the probability that one of the two infected individuals has no sampled descendants between time t and time 0, and 3) $O(\Delta t^2)$ summarizes the terms when more than one transmission event happens during time Δt. We transform this equation to

$$\frac{g_e(t+\Delta t)-g_e(t)}{\Delta t} = -(\lambda+\mu+\psi)g_e(t) + 2\lambda p_0(t)g_e(t) + O(\Delta t),$$

which yields in the limit $\Delta t \to 0$ the Master equation for $g_e(t)$,

$$\frac{d}{dt}g_e(t) = -(\lambda+\mu+\psi)g_e(t) + 2\lambda p_0(t)g_e(t).$$

Let the edge e last between time s_0 and s_1, with $s_0 \geqslant t \geqslant s_1$ (see also fig. 2). Note that at time s_1, there are two descending branches (transmission event with rate λ) or no descending branches (sampling event with rate ψ). Thus, we have the initial value at $t = s_1$:

$$g_e(s_1) = \begin{cases} \lambda g_{e_1}(s_1)g_{e_2}(s_1) & \text{if } e \text{ has two descendant edges } e_1, e_2, \\ \psi & \text{if } e \text{ has no descendant edges.} \end{cases}$$

Following Stadler (2010), we can solve the differential equation for $g_e(t)$ and obtain

$$f[\mathcal{T}|\lambda,\mu,\psi,t_{or}=x_0] = g_e(t_{or}) = \lambda^{m-1}\prod_{i=0}^{m-1}\frac{1}{q(x_i)}\prod_{i=1}^{m}\psi q(y_i), \quad (1)$$

where

$$q(t) = 2(1-c_2^2) + e^{-c_1 t}(1-c_2)^2 + e^{c_1 t}(1+c_2)^2.$$

We implemented the Bayesian inference procedure to estimate the parameters of the BDM (using eq. 1) in the software package BEAST (Drummond and Rambaut 2007), replacing the previously used coalescent. We assume that the process is stopped after the last sampled leave, that is, $y_m = 0$, this has the advantage that we reduce the number of parameters by one.

Our BDM is a modification of the framework in Stadler (2010): The previous model in Stadler (2010) allows sampling through time and sampling of present-day individuals (this is relevant when the considered individuals are extinct and extant species, rather than infected hosts). Further, a

sampled individual remained infectious. Modifying the previous model such that there is no sampling of present-day individuals, and assuming that infected individuals become noninfectious when being sampled, yields the model in this paper.

A natural generalization of the model presented here, and the model presented in Stadler (2010) is to assume that a sampled individual becomes noninfectious immediately after the sampling with probability r and remains infectious after sampling with probability $1 - r$. Under this general model, m sampled individuals may have no sampled descendants and k sampled individuals may have sampled descendants. The likelihood of the tree becomes

$$f[\mathcal{T}|\lambda,\mu,\psi,r,t_{or}=x_0] = \lambda^{m-1}(\psi(1-r))^k \times \prod_{i=0}^{m-1}\frac{1}{q(x_i)}\prod_{i=1}^{m}\psi(r+(1-r)p_0(y_i))q(y_i).$$

The current version of Beast cannot use the generalized method, as it cannot account for the k sampled individuals having sampled descendants but requires instead all the sampled individuals being leaves. Thus, we cannot do data analysis yet with this general model.

Inferring Parameters from Simulated Data Sets

We simulated trees under the BDM and then analyzed these trees using the Bayesian BDM method implemented in BEAST (Drummond and Rambaut 2007) in order to validate whether we can accurately re-estimate parameters of the BDM. There are two stages in the simulation process:

Stage 1:

In Stage 1, we simulate trees under the BDM model for fixed $\lambda = 8.23 \times 10^{-4}$, $\mu/\lambda = 0.11$, and $\psi = 2.78 \times 10^{-4}$ (which were the mean estimates obtained from empirical data, see below). We ran an MCMC in Beast without data, which samples trees from the prior (i.e., the BDM model). We used 100 tips without sequence data to simulate the tip times (except the most recent tip whose time is 0), the tree (including all internal nodes and topology), and the time of origin t_{or}. MCMC chain length was set up to 100, 000, 000.

Stage 2:

We picked up ten trees from the sample produced from Stage 1 respectively at the state 100,000,000, 90,000,000, …, 10,000,000 and created ten new MCMC simulations by fixing the tree. In the ten new MCMC simulations, we estimated λ, μ, ψ, and t_{or} and calculated $R_0 = \lambda/(\mu+\psi)$ for the ten chosen trees. MCMC chain length was set up to 100,000,000.

To increase the reliability of our tests, we looped 10 times from Stage 1 to Stage 2 and thus re-estimated $\lambda, \mu, \psi, t_{or}$ based on 100 random trees.

We also simulated sets of ten trees with ten tips to compare the accuracy of the method when having several small trees instead of one big tree.

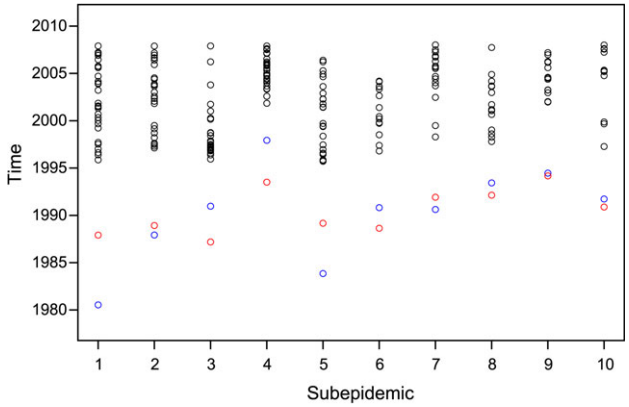

FIG. 3. The time of sampling of the data in the ten subepidemics (black) together with the estimated time of origin t_{or} for each subepidemic using our Bayesian BDM method. Case (i): Red is the time of origin estimated under the assumption that the epidemiological parameters λ, μ, ψ are the same in all subepidemics; Case (ii): blue is the time of origin estimated when the parameters may vary between subepidemics independently.

Inferring Parameters from Swiss HIV-1 Sequence Data

We use sequence data of the HIV polymerase gene from the Swiss HIV Cohort study (SHCS, The Swiss HIV Cohort Study 2010), a cohort where more than 16, 000 Swiss HIV-infected persons are enrolled covering at least 45% of all Swiss HIV-infected individuals. In order to exclude biases due to migration, we determined Swiss transmission clusters in which no migration occurred. We built a maximum likelihood tree using all available Swiss HIV pol sequences plus the same number of randomly selected foreign sequences. We defined a cluster in the tree to be a Swiss cluster if it contains at least 80% Swiss sequences and has bootstrap support of at least 70%. This procedure follows the analysis in Kouyos et al. (2010), with the only two differences that we 1) considered only clusters with a bootstrap support of $>$70%, and that 2) the phylogenetic tree was inferred on the basis of the general time reversible + GAMMA model (instead of the general time reversible with per site rate categories [GTRCAT] model).

The ten largest clusters (subepidemics) contain between 12 and 34 individuals, which have been sampled between 1995 and 2008. The sampling times of the sequences are summarized in figure 3.

From the subepidemics, we estimated transmission and sequence evolution parameters based on the pol sequences, using our Bayesian BDM method. To model the sequence evolution, we partitioned the alignment into two classes of sites: The first class consisted of all first and second codon positions and the second class was made up of third codon positions. Each class was modeled with an independent Hasegawa-Kishino-Yano + Γ model, allowing for a class-specific transition/transversion bias (κ), a class-specific shape parameter (α) describing rate heterogeneity across sites, and a class-specific mean rate of evolution. The prior distribution on the two shape parameters was an exponential distribution with a mean of 1. The prior on the κ parameters was a Gamma distribution with a shape of

Table 1. Accuracy of re-estimating R_0, λ, μ, ψ, $\lambda-\mu-\psi$ under the BDM, and $\lambda-\mu-\psi$ under the coalescent with an exponential growth prior.

True Value		1 Tree (100 Tips)	10 Trees (10 tTips)
R_0 = 2.25	Mean	1.81	3.02
	Relative error	0.20	0.35
	Relative bias	−0.20	0.35
	HPD interval width	1.75	2.75
	95% HPD accuracy	94%	100%
$\lambda = 8.23 \times 10^{-4}$	Mean	15.73×10^{-4}	9.38×10^{-4}
	Relative error	0.91	0.14
	Relative bias	0.91	0.14
	HPD interval width	29.08×10^{-4}	4.69×10^{-4}
	95% HPD accuracy	100%	91%
$\mu = 0.88 \times 10^{-4}$	Mean	8.94×10^{-4}	1.06×10^{-4}
	Relative error	9.15	0.21
	Relative bias	9.15	0.21
	HPD interval width	29.73×10^{-4}	3.28×10^{-4}
	95% HPD accuracy	100%	100%
$\psi = 2.78 \times 10^{-4}$	Mean	2.04×10^{-4}	2.30×10^{-4}
	Relative error	0.27	0.19
	Relative bias	−0.27	−0.17
	HPD interval width	3.03×10^{-4}	1.72×10^{-4}
	95% HPD accuracy	92%	79%
$\lambda-\mu-\psi$=4.57 $\times 10^{-4}$	Mean	4.75×10^{-4}	6.02×10^{-4}
	Relative error	0.14	0.33
	Relative bias	0.04	0.32
	HPD interval width	3.67×10^{-4}	3.93×10^{-4}
	95% HPD accuracy	97%	68%
$\lambda-\mu-\psi = 4.57 \times 10^{-4}$ exponential growth tree prior	Mean	5.27×10^{-4}	6.24×10^{-4}
	Relative error	0.25	0.41
	Relative bias	0.15	0.36
	HPD interval width	2.13×10^{-4}	3.37×10^{-4}
	95% HPD accuracy	55%	52%

0.05 and a scale of 40. In addition, a lognormally distributed uncorrelated relaxed clock model (Drummond et al. 2006) was used to model rate variation across lineages. The S parameter of the lognormal relaxed clock had an exponential prior with a mean of 1/3. We ran the MCMC chain for 200 million generations and neglected the first 10% of output as the burn-in.

We analyzed the subepidemics under two different assumptions: 1) We assume that the epidemiological parameters λ, μ, and ψ are the same in all ten subepidemics but leave the time of origin t_{or} variable and 2) we assume that all subepidemic can have different parameters λ, μ, ψ, and t_{or}. The runs converged in both Case (i) and Case (ii). The effective sampling size (ESS) was usually several thousands and the minimum ESS was 282.

Results

We implemented the Bayesian inference procedure to estimate the parameters of the BDM in the software package BEAST (Drummond and Rambaut 2007). From the estimates obtained with this new method, we can determine the total number of infections caused by an individual over the time during which it is infectious. This number is equivalent to the basic reproductive number R_0 and is given by the ratio of the birth and death rates:

$$R_0 = \frac{\lambda}{\mu + \psi}.$$

Thus, our Bayesian BDM method offers a possibility to estimate a key epidemiological parameter using only sequence data.

Estimates Obtained for Simulated Data

The estimated values for R_0, λ, μ, ψ from the 100 simulated trees are shown in table 1. The relative error for a parameter p was defined as

$$\text{relative\_error} = \frac{\sum_{t=1}^{100} \frac{|\hat{p}-\bar{p}|}{\bar{p}}}{100}, \quad (2)$$

where $\bar{p}$ was the true value of p and $\hat{p}$ was the mean of p. The relative bias was defined as

$$\text{relative\_bias} = \frac{\sum_{t=1}^{100} \frac{\hat{p}-\bar{p}}{\bar{p}}}{100}. \quad (3)$$

The 95% highest probability density (HPD) accuracy is the percentage of trees with 95% HPD intervals containing the true value (where a 95 % HPD interval is defined as the

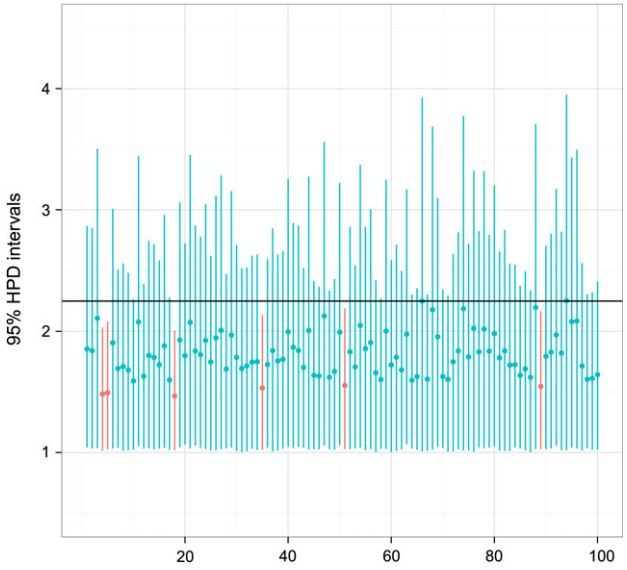

FIG. 4. The estimated mean R_0 together with the 95% HPD for the 100 simulated trees (true $R_0 = 2.29$).

shortest interval, which contains 95% of the posterior probability).

The estimation of R_0 was accurate; the true R_0 was contained in the 95% HPD in 94 of the 100 trees. The HPDs for each tree are plotted in figure 4. The accuracy of all estimates (R_0, λ, μ, ψ, and $\lambda - \mu - \psi$) is given in table 1. Further, the table contains the result of the re-estimation of the net growth $\lambda - \mu - \psi$ based on the same 100 simulated trees but using the coalescent with an exponential growth prior instead of the BDM method for re-estimation.

The parameter estimates for R_0, λ, μ, ψ, and $\lambda - \mu - \psi$ using the Bayesian BDM method were reliable: 95% HPD intervals contained the true parameters in more than 90% of the cases. In particular, the true $\lambda - \mu - \psi$ was contained in the 95% HPD interval in 97% of the cases. However, when using the coalescent and estimating $\lambda - \mu - \psi$, the 95% HPD intervals contained the true parameter in only 55% of the cases.

Using 10 trees with 10 tips instead of 1 tree with 100 tips yielded similar results. The improvement obtained when using ten trees is that the HPD interval width for λ, μ, and ψ become significantly smaller; however, the 95% accuracy for the net growth $\lambda - \mu - \psi$ drops to 68% (from 97%).

We further investigated correlations of parameters and recovered that λ is positively correlated with $\mu + \psi$, see figure 5. Thus, the method can estimate the ratio $\lambda/(\mu+\psi)$ accurately but cannot determine well λ and $\mu + \psi$ separately. This explains the large HPD intervals for λ, μ, ψ, while having a confined HPD interval for $R_0 = \frac{\lambda}{\mu+\psi}$.

Estimates Obtained for HIV-1 in Switzerland

We applied our Bayesian BDM method to HIV-1 sequence data from the SHCS (The Swiss HIV Cohort Study 2010). As explained in the Materials and Methods section, we focused our analysis on ten Swiss HIV-1 subepidemics in order to exclude biases due to migration from outside of Switzerland. In Case (i), we assumed the same epidemiological parameters λ, μ, ψ in the ten subepidemics, and in Case (ii), we allowed for different parameters in the ten subepidemics.

In Case (i), we obtain the following mean estimates: transmission rate $\lambda = 8.23 \times 10^{-4}$/day (or 0.30/year); becoming noninfectious rate $\mu = 9.46 \times 10^{-5}$/day; sampling rate $\psi = 2.78 \times 10^{-4}$/day. The 95% HPD intervals are given in table 2. As the average time until becoming noninfectious is $\frac{1}{\mu+\psi}$, we observe an average time of infectiousness of 7.74 $[4.39, 10.99]$ years. From the estimates of λ, μ, ψ, we calculate the mean posterior $R_0 = 2.29$ with a 95% HPD interval $[1.61, 3.05]$ (see table 2). The times of origin t_{or} of the ten different subepidemics are between 1987 and 1994 (see also fig. 3).

In Case (ii), the estimates lie in the same range as for the analysis of Case (i) (i.e., the 95% HPD intervals largely overlap). The estimated mean rates and R_0 for all analyses together with the 95% HPD intervals are given in table 2. The mean times of origin of the ten subepidemics are shown in figure 3.

The 95% HPD intervals for Case (ii) are much wider than for Case (i), due to less data being available for the estimation of the epidemiological parameters. Further, we note that in Case (ii) compared with Case (i), the mean of λ and μ is always larger. This bias can partially be explained through correlations in parameter estimates. We observe from the simulations and the data that λ correlates linearly with $\mu + \psi$, see figure 5. Determining why larger λ and μ values are estimated for smaller data sizes needs to be investigated in future simulation studies. For the present work, it is important that the biases vanish when considering R_0.

In Case (ii), we obtain different R_0 estimates for the different subepidemics. These differences could be due to a variety of factors such as transmission group, the size of the epidemic, the time of origin of the subepidemic, or stochastic fluctuations.

The subepidemics studied here are dominated by different transmission groups (Kouyos et al. 2010). In particular, the subepidemics here are characterized either by predominant transmission between men having sex with men (MSM) or by predominant transmission between mixed groups of heterosexuals (HET) and intravenous drug users (IDU). The composition of the subepidemics according to transmission group are shown in table 3. The subepidemics 3, 6, and 9 are dominated by HET and IDU. The other subepidemics are dominated by MSM. The mean R_0 in the HET/IDU subepidemics shows no trend to be lower or higher than the mean R_0 in the MSM subepidemics: The three HET/IDU subepidemics have 1st, 5th, and 10th largest mean R_0 of the ten mean R_0. This is confirmed by statistical analysis: A nonparametric test (runs statistic, Hogg and Craig 1994) does not reject the null hypothesis, that R_0 is the same in the HET/IDU and MSM transmission groups, with a P value of 0.58.

We test whether the size of the subepidemic correlates with the R_0 by regressing the size of the subepidemic against

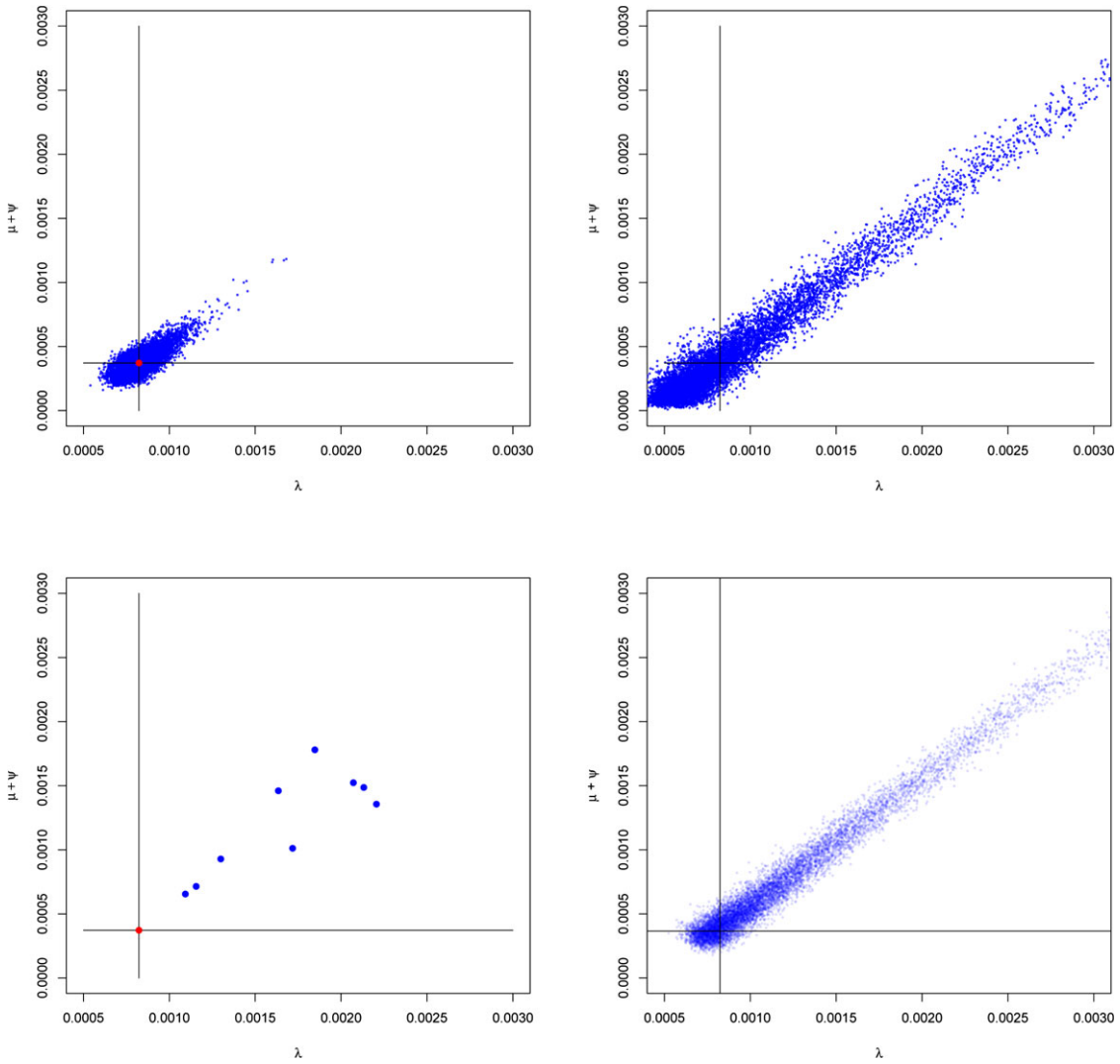

FIG. 5. Correlation between λ and $\mu + \psi$. On the top panels, the posterior distribution of estimates obtained from the empirical analyses is shown (left: Case (i); right: Case (ii) subepidemic 1). On the bottom left panel, the mean estimates from the empirical data analyses are shown (Case (i) analysis in red, Case (ii) analyses in blue). On the bottom right panel, the posterior distribution of estimates obtained from a simulated tree are plotted. The black lines indicate the mean estimates obtained in the empirical Case (i) analysis, which were further used in the simulation study.

R_0. The absence of a significant correlation (Pearson correlation 0.07, P value 0.84) suggests that size is not a major determinant of R_0 (see also fig. 6). Moreover, this also suggests that our estimate of R_0 is not affected by the fact that we only include subepidemics with sample size of 12 and larger in our analysis.

In order to investigate if the R_0 is different in old and new subepidemics, we plot the estimated mean time of origin of each subepidemic against the estimated mean R_0 (see fig. 7). Although the R_0 decreases for younger clusters, the correlation is nonsignificant (Pearson correlation −0.49, P value 0.15).

Clearly, the variation in R_0 in the different subepidemics could also be explained by further factors such as differences in behavior within different subepidemics independent of transmission group or the founder strains of the subepidemics having varying virulence (Alizon et al. 2010). However, in the absence of any such data, we cannot test for the role of these factors. Finally, it should be noted that the differences in R_0 could also be simply due to chance given that the 95% HPD intervals overlap in most cases.

Validation of HIV-1 Estimates

To verify that our estimates are compatible with other data regarding the Swiss HIV epidemic, we compared the estimates obtained from our analysis with estimates obtained through other methods.

First, the mean probability of an infected individual having been sampled before dying, $\frac{\psi}{\psi+\mu}$, is estimated with our method to be 77.6% (95% HPD $[44.7, 100.0]$). This is in very good agreement with the recent Swiss HIV Cohort report (The Swiss HIV Cohort Study 2010), where it is estimated that 75% of all Swiss HIV infected which developed AIDS are enrolled in the Swiss Cohort.

The expected number of secondary infections caused by an infected individual over a year is $365 \times \lambda = 0.30$ with

Table 2. Mean R_0, λ, μ, ψ estimates with 95% HPD intervals for the ten Swiss HIV subepidemics.

		R_0				λ	
		Mean	95% HPD			Mean	95% HPD
Case (i)		2.29	[1.61, 3.05]	Case (i)		8.23	[6.62, 10.17]
	1	2.95	[1.01, 6.82]		1	10.93	[3.58, 24.52]
	2	1.74	[0.92, 3.03]		2	13.00	[4.12, 31.18]
	3	1:06	[0.64, 1.50]		3	18.47	[5.38, 41.70]
	4	1.57	[0.83, 2.67]		4	31.55	[9.72, 72.86]
Case (ii)	5	2.92	[0.93, 7.46]	Case (ii)	5	11.56	[3.69, 27.90]
	6	1.88	[0.90, 3.68]		6	21.32	[5.72, 48.83]
	7	2.85	[1.01, 6.81]		7	17.18	[4.45, 41.34]
	8	1.26	[0.52, 2.21]		8	16.34	[3.57, 37.46]
	9	3.00	[0,77, 7.82]		9	22.06	[5.20, 51.37]
	10	1.70	[0.60, 3.28]		10	20.71	[4.04, 49.71]
		μ				Ψ	
		Mean	95% HPD			Mean	95% HPD
Case (i)		0.95	[0.00, 2.80]	Case (i)		2.78	[1.84, 3.78]
	1	5.88	[0.00, 19.93]		1	0.67	[0.01, 1.71]
	2	7.33	[0.00, 26.19]		2	1.95	[0.17, 3.87]
	3	11.63	[0.00, 37.37]		3	6.16	[1.03, 1128]
	4	18.26	[0.00, 61.33]		4	5.65	[0.72, 11.48]
Case (ii)	5	6.19	[0.00, 22.97]	Case (ii)	5	0.95	[0.00, 2.59]
	6	11.93	[0.00, 39.13]		6	2.93	[0.08, 6.29]
	7	8.96	[0.00, 33.20]		7	1.15	[0.04, 2.80]
	8	10.14	[0.00, 32.79]		8	4.46	[0.22, 9.49]
	9	11.83	[0.00, 40.59]		9	1.73	[0.01, 5.33]
	10	11.99	[0.00, 41.10]		10	3.33	[0.19, 7.19]

NOTE.—Case (i) assumes the same epidemiological parameters λ, μ, ψ in all subepidemics; Case (ii) allows the parameters to vary between the subepidemics $1, \ldots, 10$. The rates λ, μ, ψ are stated in units 10^{-4}/day.

95 % HPD $[0.24, 0.37]$ (table 2). This parameter has been estimated from transmission clusters using the SHCS (The Swiss HIV Cohort Study 2010) together with the Zurich primary HIV infection study (Metzner et al. 2010) to be 1.8 (95% confidence interval $[0.5, 5.8]$) for people infecting in the chronic phase (Rieder et al. 2010). Their confidence interval is not overlapping with our confidence interval, which might be due to real transmission differences in the different data sets (recall that simulations show a 100% accuracy for our confidence interval, table 1).

Thus, we applied our Bayesian BDM method to the largest transmission cluster found in Rieder et al. (2010). For this cluster, we obtained a mean estimate of $365 \times \lambda = 1.14$ with 95 % HPD $[0.31, 2.66]$, which largely overlaps with but is more confined than the previously estimated confidence interval $[0.5, 5.8]$ (Rieder et al. 2010).

Comparison of BDM Analysis to Classical Analysis Using the Coalescent

Under the coalescent with exponential growth, the net growth parameter $\lambda - \mu - \psi$ can be estimated. Recall that based on the simulated trees, the net growth is estimated more accurate under the BDM than under the coalescent (97% vs. 55% HPD accuracy).

Table 3. Number of individuals in the ten considered Swiss HIV subepidemics.

Transmission Group	1	2	3	4	5	6	7	8	9	10
MSM	31	24	0	25	19	0	15	9	0	9
HET	3	4	6	1	3	6	2	5	8	3
IDU	0	0	21	0	2	12	0	0	0	0
Other	0	0	0	0	1	0	0	0	3	0
Unknown	0	1	0	0	0	0	0	0	3	0
Foreign	2	1	0	0	0	1	0	0	0	1

NOTE.—The individuals are sorted with respect to transmission group: men having MSM, HET, IDU, other, unknown, and non-Swiss individual (foreign). Note that the foreign individuals were included to detect Swiss clusters, but they are excluded when estimating the rates in the subepidemics.

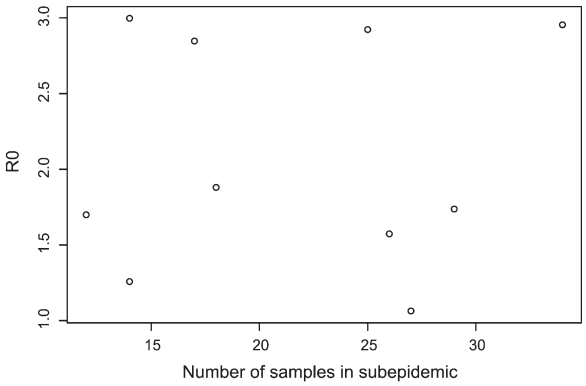

FIG. 6. Sample size of the ten subepidemics against the estimated mean R_0 value.

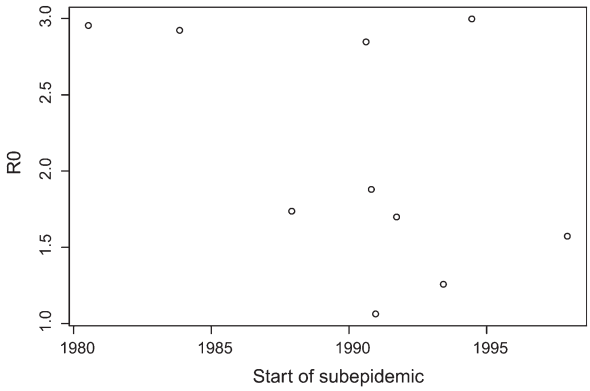

FIG. 7. Estimated mean time of origin of the ten subepidemics against the estimated mean R_0.

To investigate the impact of model choice on empirical data results, we analyzed the ten Swiss HIV subepidemics assuming the coalescent with exponential growth and compared these results with the BDM analysis (Case [i]). Under the coalescent, the mean exponential growth parameter was estimated to be 6.89×10^{-4}/day. The 95% HPD interval is $[3.01 \times 10^{-4}, 11.09 \times 10^{-4}]$. Under the BDM, we estimated for the exponential growth parameter $\lambda - \mu - \psi$ a mean of 4.51×10^{-4}/day with 95% HPD interval $[3.20 \times 10^{-4}, 5.72 \times 10^{-4}]$. The 95% HPD interval of the BDM analysis is fully contained within the HPD interval of the coalescent analysis meaning that the BDM analysis has the power to provide more confined HPD intervals. Further, recall that the BDM method is able to provide the parameters λ, μ, ψ independently such that R_0 can be calculated; the coalescent only provides the net growth $\lambda - \mu - \psi$.

Discussion

Estimation of Key Epidemiological Parameters

Our study presents and applies a method to infer key epidemiological parameters directly from viral sequence data. Various attempts have been made to use genetic sequences in order to estimate epidemiological parameters. However, none of these studies have been able to infer the basic reproductive number R_0 only from sequence data. R_0 of an epidemic (Hepatitis C Virus, HCV) was estimated from viral sequence data for the first time in Pybus et al. (2001). Because these authors used the coalescent as a transmission model, they could not directly infer transmission and death rates but instead required an independent estimate of average duration of infectiousness. In Volz et al. (2009) and Frost and Volz (2010), transmission rates are introduced to the coalescent framework, but an independent estimate is still required for the duration of infectiousness in order to calculate R_0.

Although assuming an estimate of the duration of infectiousness may be appropriate for HCV, the estimation of the time of infectiousness is fraught with difficulties for many infections. In particular in HIV, the duration of infection is highly variable between patients as the time until AIDS can vary between 2 and 20 years. Moreover, the time span over which a patient is (highly) infectious is debated (Yerly et al. 2001, 2007; Brenner et al. 2008; Hollingsworth et al. 2008; Rieder et al. 2010). Given the uncertainties in estimating the duration of infectiousness together with its variability and given the observation that fixing the time of infectiousness to different values yields different R_0 estimates (Pybus et al. 2001; Wallinga and Lipsitch 2007), such coalescent-based methods to infer R_0 for HIV have to be used with great care.

Our Bayesian BDM method overcomes these problems essentially by independently estimating transmission and death rates together with the transmission chain given sequentially sampled sequence data. Although our method thus represents an advance over coalescent-based methods, it also has certain shortcomings. In particular, being based on a BDM, our approach assumes exponential growth of the epidemic. To fully account for the population dynamics of the epidemic would require a tree-generating model based on more explicit epidemiological models (such as SI, SIR, or SEIR models [S: susceptible, E: exposed, I: infectious, R: recovered]; Keeling and Rohani 2008). Such models would not only account for an initial exponential increase but also a saturation phase as susceptible hosts are becoming limited.

For the application of our method to the HIV data from Switzerland, we believe that saturation is not a major concern. If the considered subepidemics had progressed beyond the exponential phase, late edges in the trees would be expected to be long compared with early edges. Long late edges should result in vanishing estimates for the becoming noninfectious rate (Nee et al. 1994). The fact that we obtain nonzero estimates for the becoming noninfectious rates indicates that the subepidemics did not yet reach the postexponential phase. We emphasize that the considered subepidemics being in the exponential phase does not contradict the Swiss epidemic as a whole being in the postexponential phase.

One study has been published that estimates R_0 in HIV from sequence data (Volz et al. 2009) based on a modified coalescent model. Using a transmission tree inferred from sequence data of HIV-infected individuals sampled at one time point (1993) in the United States, the estimate $R_0 = 2.29$ is obtained, assuming a time of infectiousness of 10 years. A few studies have been published that estimate R_0 in HIV from epidemiological data (Bezemer et al. 2010; Nishiura 2010). These estimates were obtained from the temporal changes of the incidence of the infection and ad hoc estimates for the time of infectiousness. For example, using data for the early HIV epidemic until 1984, the R_0 of HIV in Western Europe was estimated to be between 3.5 and 4.1 (Nishiura 2010) using published estimates for the time span of infectiousness and time-dependent infection intensity (Hollingsworth et al. 2008). In Bezemer et al. (2010), the R_0 for the MSM transmission group in the Netherlands was estimated with a likelihood approach for different time periods showing temporal fluctuations (1980–1983: $R_0 = 2.39$ $[2.17, 2.76]$; 1984–1995: $R_0 = 0.89$ $[0.85, 0.93]$; 1996–1999: $R_0 = 0.76$ $[0.70, 0.86]$; 2000–2003: $R_0 = 1.04$ $[0.98, 1.09]$).

The actual reproductive number, R_a, is defined as the number of secondary infections caused by a single

infected individual for the current frequency of susceptibles (Amundsen et al. 2004), whereas R_0 is defined as the number of secondary infections when the entire population is still susceptible. Provided an epidemic is far from saturation, then the estimates of R_a can be used as a good approximation for R_0.

Assuming an average infectiousness period of 10 years, R_a for HIV in the United Kingdom was estimated to be lower than 1 between 1995 and 2004, R_a remained lower than 1 for HET and above 1 for MSM (White et al. 2006). An independent study estimated for MSM in 1995 an R_a of 0.55 in Denmark, 0.85 in Norway, and 0.58 in Sweden (Amundsen et al. 2004). For IDU, R_a was estimated to be 3.5 in Latvia and 21.7 in Lithuania in 2002 (assuming an average duration of infectiousness of 11 years).

Taken together, the estimates of R_0 and R_a vary considerably. This may be in part due to methodological difference but likely also reflects differences in the epidemics in different countries or different transmission groups. Our estimate of $R_0 = 2.29$ is thus broadly in agreement with these earlier estimates. Note, in particular, that our estimates represent time averages over the epidemic with some of the considered subepidemics ranging back to the 1980.

Comparing our parameter estimates to quantities estimated by independent means from the Swiss HIV Cohort in previous studies (Metzner et al. 2010; Rieder et al. 2010; The Swiss HIV Cohort Study 2010) reinforces our confidence in the method. Specifically, our estimated sampling probability is in good agreement with previous estimates.

Analyzing the transmission in distinct subepidemics in Switzerland, we did not find any significant correlation between R_0 and size of the subepidemic, or age of the subepidemic, or transmission group (HET/IDU versus MSM). The absence of such correlation may be due in part to the limited number of subepidemics that could be studied here. For example, the association between age of the subepidemic and R_0 shows a trend towards decreasing R_0 in younger subepidemics. However, the overall absence of significant associations suggests that neither size, age, nor transmission group of the subepidemic are major explanatory factors of R_0.

A legitimate concern is that our method requires subepidemics that are large enough such that a substantial number of patients are sampled. Therefore, our analysis is biased towards larger subepidemics, which may in turn result from larger R_0. However, as there is no correlation between the size of the subepidemic and R_0 in our study, there is no evidence that our estimate R_0 overestimates the true R_0 in the entire Swiss epidemic. Note, moreover, that a general bias towards larger subepidemics is not a limitation just of our method but more generally applies to all studies mentioned here.

Tree-Generating Models in Bayesian Analyses

The Bayesian BDM method developed and applied here has the advantage over previous methods (which are based on the coalescent) that the underlying tree-generating model reflects more accurately the epidemiological process of disease transmission. This implies that the BDM method provides estimates of key epidemiological parameters, whereas the coalescent only provides an estimate of the net growth of an epidemic. Furthermore, our simulations reveal an improved HPD accuracy for the net growth when using the BDM method instead of the classic coalescent method (97% vs. 55%). The Swiss HIV data analysis reveals three times more confined HPD intervals for the net growth when using the BDM method. Therefore, we consider the Bayesian BDM method to be more accurate and appropriate not only in cases where epidemiological parameters are being inferred but also generally when Bayesian methods are used for phylogenetic analysis of epidemiological sequence data.

Our BDM can only account for the exponential growth phase of an epidemic. Thus, the coalescent is still the only framework under which postexponential epidemic dynamics can be investigated. Having shown the advantages of the BDM over the coalescent, this paper will hopefully stimulate research to also use BDMs in order to describe the postexponential phase of the epidemic.

The method was applied here specifically to HIV but can be used to infer the epidemiology of other viral epidemics. Moreover, it could be adjusted to infer a within-host basic reproductive number R_0 based on sequence samples that are obtained over the course of a viral infection in an individual patient; again, this would circumvent the requirement of an independent estimate of the expected generation time. Hence, our Bayesian BDM method represent a versatile tool for phylogenetic analysis of viral sequence data.

Acknowledgments

We thank the patients for participation in the SHCS and the clinical trials, the physicians and study nurses for excellent patient care, the laboratory technicians of the Swiss resistance laboratories for the quality of the data, SmartGene, Zug, Switzerland for providing excellent technical support, and Brigitte Remy, Martin Rickenbach, and Yannick Vallet from the SHCS data center in Lausanne for the data management. Funding sources: This study has been financed in the framework of the SHCS and supported by the Swiss National Science Foundation (SNF grant number 3345-062041). Further support was provided by the SNF (grants numbers 3247B0-112594 to H.F.G., S.Y., B.L., S.B.; 324730_130865 to H.F.G.; 3100A 0_116408 to S.B.); SHCS project 470, 528 and 569; the SHCS Research Foundation, by a further research grant of the Union Bank of Switzerland in the name of a donor to H.F.G. and by the European Community's Seventh Framework Programme (grant FP7/2007−2013), under the Collaborative HIV and Anti-HIV Drug Resistance Network (grant 223131 to H.F.G.). The funding agencies had no role in conducting the study or in preparing the manuscript.

References

Alizon S, von Wyl V, Stadler T, et al. (19 co-authors). 2010. Phylogenetic approach reveals that virus genotype largely determines HIV set-point viral load. *PLoS Pathog.* 6:e1001123.

Amundsen E., Stigum H., Roettingen J, Aalen O. 2004. Definition and estimation of an actual reproduction number describing past

infectious disease transmission: application to HIV epidemics among homosexual men in Denmark, Norway and Sweden. *Epidemiol Infect*. 132:1139–1149.
Anderson R, May R. 1979. Population biology of infectious diseases: part I. *Nature* 280:361–367.
Anderson R, May R. 1992. Infectious diseases of humans: dynamics and control. New York: Oxford University Press.
Bezemer D, de Wolf F, Boerlijst M, van Sighem A, Hollingsworth T, Fraser C. 2010. 27 years of the HIV epidemic amongst men having sex with men in the Netherlands: an in depth mathematical model-based analysis. *Epidemics* 2(2):66–79.
Brenner B, Roger M, Moisi D, Oliveira M, Hardy I, Turgel R, Charest H, Routy J, Wainberg M, Montreal PHI Cohort and HIV Prevention Study Groups. 2008. Transmission networks of drug resistance acquired in primary/early stage HIV infection. *AIDS* 22:2509.
Couvreur T, Franzke A, Al-Shehbaz I, Bakker F, Koch M, Mummenhoff K. 2010. Molecular phylogenetics, temporal diversification, and principles of evolution in the mustard family (Brassicaceae). *Mol Biol Evol*. 27:55.
Drummond A, Nicholls G, Rodrigo A, Solomon W. 2002. Estimating mutation parameters, population history and genealogy simultaneously from temporally spaced sequence data. *Genetics* 161:1307–1320.
Drummond A, Pybus O, Rambaut A, Forsberg R, Rodrigo A. 2003. Measurably evolving populations. *Trends Ecol Evol*. 18:481–488.
Drummond A, Rambaut A. 2007. BEAST: Bayesian evolutionary analysis by sampling trees. *BMC Evol Biol*. 7:214.
Drummond A, Rambaut A, Shapiro B, Pybus O. 2005. Bayesian coalescent inference of past population dynamics from molecular sequences. *Mol Biol Evol*. 22:1185.
Drummond AJ, Ho SYW, Phillips MJ, Rambaut A. 2006. Relaxed phylogenetics and dating with confidence. *PLoS Biol* 4:e88.
Feller W. 1968. An introduction to probability theory and its applications. Vol. 1. 3rd ed. New York: John Wiley & Sons Inc.
Felsenstein J. 2004. Inferring phylogenies. Vol. 8. Sunderland (MA): Sinauer Associates. p. 8–5.
Fernández-Mazuecos M, Vargas P. 2010. Ecological rather than geographical isolation dominates Quaternary formation of Mediterranean Cistus species. *Mol Ecol*. 19:1381–1395.
FitzJohn R, Maddison W, Otto S. 2009. Estimating trait-dependent speciation and extinction rates from incompletely resolved phylogenies. *Syst Biol*. 58:595.
Frost S, Volz E. 2010. Viral phylodynamics and the search for an 'effective number of infections'. *Philos Trans R Soc Lond B Biol Sci*. 365:1879.
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.
Griffiths R, Tavare S. 1994. Sampling theory for neutral alleles in a varying environment. *Philos Trans R Soc Lond B Biol Sci*. 344: 403–410.
Hogg RV, Craig A. 1994. Introduction to mathematical statistics. 5th ed. Englewood Cliffs (NJ): Prentice Hall.
Hollingsworth T, Anderson R, Fraser C. 2008. HIV-1 transmission, by stage of infection. *J Infect Dis*. 198:687–693.
Holmes E, Nee S, Rambaut A, Garnett G, Harvey P. 1995. Revealing the history of infectious disease epidemics through phylogenetic trees. *Philos Trans R Soc Lond B Biol Sci* 349:33–40.
Keeling M, Rohani P. 2008. Modeling infectious diseases in humans and animals. Princeton (NJ): Princeton University Press.
Kendall DG. 1948. On some modes of population growth leading to R. A. Fisher's logarithmic series distribution. *Biometrika* 35:6–15.
Kingman JFC. 1982. The coalescent. *Stoch Anal Appl*. 13:235–248.
Kouyos RD, von Wyl V, Yerly S, et al. (20 co-authors). 2010. Molecular epidemiology reveals long-term changes in HIV type 1 subtype B transmission in Switzerland. *J Infect Dis*. 201:1488–1497.
Maddison W, Midford P, Otto S. 2007. Estimating a binary character's effect on speciation and extinction. *Syst Biol*. 56:701.
Metzner K, Rauch P, von Wyl V, Leemann C, Grube C, Kuster H, Böni J, Weber R, Günthard H. 2010. Efficient suppression of minority drug-resistant HIV type 1 (HIV-1) variants present at primary HIV-1 infection by ritonavir-boosted protease inhibitor-containing antiretroviral therapy. *J Infect Dis*. 201:1063–1071.
Nee S, Holmes EC, May RM, Harvey PH. 1994. Extinction rates can be estimated from molecular phylogenies. *Philos Trans R Soc Lond B Biol Sci*. 344:77–82.
Nee S, Holmes E, Rambaut A, Harvey P. 1995. Inferring population history from molecular phylogenies. *Philos Trans R Soc Lond B Biol Sci*. 349:25–31.
Nishiura H. 2010. Correcting the actual reproduction number: A simple method to estimate R0 from early epidemic growth data. *Int J Environ Res Public Health*. 7(1):291–302.
Patané J, Weckstein J, Aleixo A, Bates J. 2009. Evolutionary history of Ramphastos toucans: molecular phylogenetics, temporal diversification, and biogeography. *Mol Phylogenet Evol*. 53:923–934.
Pomeroy L, Bjørnstad O, Holmes E. 2008. The evolutionary and epidemiological dynamics of the paramyxoviridae. *J Mol Evol*. 66: 98–106.
Pybus O, Charleston M, Gupta S, Rambaut A, Holmes E, Harvey P. 2001. The epidemic behavior of the hepatitis C virus. *Science* 292: 2323–2325.
Pybus O, Rambaut A, Harvey P. 2000. An integrated framework for the inference of viral population history from reconstructed genealogies. *Genetics* 155:1429.
Rambaut A. 2000. Estimating the rate of molecular evolution: incorporating non-contemporaneous sequences into maximum likelihood phylogenies. *Bioinformatics* 16:395.
Rieder P, Joos B, von Wyl V, et al. (14 co-authors). 2010. HIV-1 transmission after cessation of early antiretroviral therapy among men having sex with men. *AIDS* 24:1177.
Stadler T. 2009. On incomplete sampling under birth-death models and connections to the sampling-based coalescent. *J Theor Biol*. 261:58–66.
Stadler T. 2010. Sampling-through-time in birth-death trees. *J Theor Biol*. 267:396–404.
The Swiss HIV Cohort Study (2010). Cohort profile: the Swiss HIV Cohort Study. *Int J Epidemiol*. 39:1179–1189.
Volz E, Pond K, Sergei L, Ward M, Brown L, Andrew J, Frost S. 2009. Phylodynamics of infectious disease epidemics. *Genetics* 183(4): 1421–1430.
Wallinga J, Lipsitch M. 2007. How generation intervals shape the relationship between growth rates and reproductive numbers. *Proc R Soc Lond B Biol Sci*. 274:599.
Weksler M, Lanier H, Olson L. 2010. Eastern Beringian biogeography: historical and spatial genetic structure of singing voles in Alaska. *J Biogeogr*. 37:1414–1431.
White P, Ward H, Garnett G. 2006. Is HIV out of control in the UK? An example of analysing patterns of HIV spreading using incidence-to-prevalence ratios. *AIDS* 20:1898.
Yerly S, von Wyl V, Ledergerber B, et al. (12 co-authors). 2007. Transmission of HIV-1 drug resistance in Switzerland: a 10-year molecular epidemiology survey. *AIDS* 21:2223.
Yerly S, Vora S, Rizzardi P, et al. (14 co-authors). 2001. Acute HIV infection: impact on the spread of HIV and transmission of drug resistance. *AIDS* 15:2287.