DOI: 10.1534/genetics.109.106021

Phylodynamics of Infectious Disease Epidemics

Erik M. Volz,*,†,[1] Sergei L. Kosakovsky Pond,‡ Melissa J. Ward,§ Andrew J. Leigh Brown§ and Simon D. W. Frost**

**Department of Epidemiology, University of Michigan, Ann Arbor, Michigan 48109, †Department of Pathology and ‡Department of Medicine, University of California, La Jolla, California 92093, §School of Biological Sciences, University of Edinburgh, Edinburgh EH9 3JR, United Kingdom and **Department of Veterinary Medicine, University of Cambridge, Cambridge CB3 0ES, United Kingdom*

Manuscript received June 8, 2009
Accepted for publication September 11, 2009

ABSTRACT

We present a formalism for unifying the inference of population size from genetic sequences and mathematical models of infectious disease in populations. Virus phylogenies have been used in many recent studies to infer properties of epidemics. These approaches rely on coalescent models that may not be appropriate for infectious diseases. We account for phylogenetic patterns of viruses in susceptible–infected (SI), susceptible–infected–susceptible (SIS), and susceptible–infected–recovered (SIR) models of infectious disease, and our approach may be a viable alternative to demographic models used to reconstruct epidemic dynamics. The method allows epidemiological parameters, such as the reproductive number, to be estimated directly from viral sequence data. We also describe patterns of phylogenetic clustering that are often construed as arising from a short chain of transmissions. Our model reproduces the moments of the distribution of phylogenetic cluster sizes and may therefore serve as a null hypothesis for cluster sizes under simple epidemiological models. We examine a small cross-sectional sample of human immunodeficiency (HIV)-1 sequences collected in the United States and compare our results to standard estimates of effective population size. Estimated prevalence is consistent with estimates of effective population size and the known history of the HIV epidemic. While our model accurately estimates prevalence during exponential growth, we find that periods of decline are harder to identify.

COALESCENT theory has found wide applications for inference of viral phylogenies (Nee *et al.* 1996; Rosenberg and Nordborg 2002; Drummond *et al.* 2005) and estimation of epidemic prevalence (Yusim *et al.* 2001; Robbins *et al.* 2003; Wilson *et al.* 2005), yet there have been few attempts to formally integrate coalescent theory with standard epidemiological models (Pybus *et al.* 2001; Goodreau 2006). While epidemiological models such as susceptible–infected–recovered (SIR) consider the dynamics of an entire population going forward in time, the coalescent theory operates on a small sample of an infected subpopulation and models the merging of lineages backward in time until a common ancestor has been reached. The original coalescent theory was based on a population of constant size with discrete generations (Kingman 1982a,b). Numerous extensions have been made for populations with overlapping generations in continuous time, exponential or logistic growth (Griffiths and Tavaré 1994), and stochastically varying size (Kaj and Krone 2003). However, infectious disease epidemics are a special case of a variable size population, often characterized by early explosive growth followed by decline that leads to extinction or an endemic steady state.

If superinfection is rare and if mutation is fast relative to epidemic growth, each lineage in a phylogenetic tree corresponds to a single infected individual with its own unique viral population. An infection event viewed in reverse time is equivalent to the coalescence of two lineages and every transmission of the virus between hosts can generate a new branch in the phylogeny of consensus viral isolates from infected individuals. Recently diverged sequences should represent transmissions in the recent past, and branches close to the root of a tree should represent transmissions from long ago. Consequently, branching patterns provide information about the frequency of transmissions over time (Wilson *et al.* 2005). The correspondence between transmission and phylogenetic branching is easiest to detect for viruses such as human immunodeficiency virus (HIV) and hepatitis C virus that have a high mutation rate relative to dispersal. Underlying SIR dynamics also apply to other pathogens, although in some cases it may be more difficult to reconstruct the transmission history.

We examined the properties of viral phylogenies generated by the most common epidemiological models based on ordinary differential equations (ODEs).

Supporting information is available online at http://www.genetics.org/cgi/content/full/genetics.109.106021/DC1.

[1]*Corresponding author:* Department of Epidemiology, University of Michigan, M5073, SPH II, 1415 Washington Heights, Ann Arbor, MI 48109-2029. E-mail: erik@erikvolz.info

We are able to fit epidemiological models to a reconstructed phylogeny for sampled viral sequence data and make inferences regarding the size of the corresponding infected population. Our solution takes the form of an ODE analogous to those used to track epidemic prevalence and thereby provides a convenient link between commonly used epidemiological models and phylodynamics. Virtually all coalescent theory to date has been expressed in terms of integer-valued stochastic processes. Our motivation for using differential equations to describe the coalescent process is a desire to formalize a link with standard epidemiological models that are also expressed in terms of differential equations.

We use our method to calculate the distribution of coalescent times for samples of viral sequences, fit SIR models to a viral phylogeny, and calculate median time to the most recent common ancestor (MRCA) of the sample. Our method also provides equations that describe the time evolution of the cluster size distribution (CSD)—the distribution of the number of descendants of a lineage over time. Clusters of related virus are often interpreted as epidemiologically linked. For example, clusters of acute HIV infections may represent short transmission chains between high-risk individuals (Yerly *et al.* 2001; Hue *et al.* 2005; Pao *et al.* 2005; Goodreau 2006; Brenner *et al.* 2007; Drumright and Frost 2008; Lewis *et al.* 2008). Because our model reproduces the moments of the cluster size distribution, it can be used to predict the level of clustering as a function of epidemiological conditions. The moments could be directly compared to empirical values or they could be used to reconstruct the entire CSD, whereupon standard statistical tests could be used for comparing distributions.

Although our equations describe the macroscopic properties of the population distribution of cluster sizes, we generalize our method to the case of a small cross-sectional sample of sequences. This allows us to develop a likelihood-based approach to fitting SIR models to observed sequences.

By considering variable degrees of incidence and the size of the infected population, our solution sheds light on the relationship between coalescent rates and epidemic dynamics. Coalescent rates are low near peak prevalence, but higher when there is a large ratio of incidence to prevalence. This can occur early on, when the epidemic is entering its expansion phase, as well as late if the epidemic has multiple periods of growth.

METHODS

Consider a population of size N comprising susceptible ($\mathcal{S}$), infected ($\mathcal{I}$), and recovered ($\mathcal{R}$) individuals. The deterministic limiting behavior of $S = |\mathcal{S}|/N$, $I = |\mathcal{I}|/N$, and $R = |\mathcal{R}|/N$ as $N \to \infty$ and with all variables $\gg 1/N$ is described by a set of coupled ordinary differential equations, with time-dependent rates of change from state X to state Y denoted as $f_{XY}(t)$. For instance, the classical mass-action SIR model

$$\dot{S} = -\beta SI,\ \dot{I} = \beta SI - \gamma I,\ \dot{R} = \gamma I \qquad (1)$$

(Kermack and McKendrick 1927; Bailey 1975; Anderson and May 1991) is obtained by setting $f_{SI}(t) = \beta S(t) I(t)$, $f_{IR}(t) = \gamma I(t)$, and all other rates to 0. We omit the explicit dependence of terms on time when it is unambiguous.

Classical coalescent inference operates on a small subsample of the larger evolving population, taken at a single time point, and infers properties of the population at an earlier time point; *e.g.*, What is the expected number of lineages at a given time *t*? Here, we denote the time of sampling by T and consider the evolution of the population backward in time toward time $t = 0$. While this differs from the conventional temporal notation for coalescent theory (where the sampling, or present, time is denoted 0, and as we move backward t denotes the number of years before the present), it allows us to develop a system of equations that link coalescent inference with standard epidemiological models.

We apply the coalescent model to the population of infecteds ($\mathcal{I}$) and draw upon the dynamical system to provide parameters such as the rate of lineage coalescence. The practical questions that we seek to address include the following:

If n individuals are sampled at time T, how many lineages exist at time $t \leq T$?

How many lineages extant at time t have surviving progeny at time T? We define *progeny* of a viral lineage extant from time $t \leq T$ as those individuals infected at time T whose virus can be traced back to that viral lineage at time t. For instance, in Figure 1, from $t = t_1$ the progeny of lineage 6 has four individuals (5, 6, 8, and 9), but from $t = t_2$ the progeny of lineage 6 consists of only 5 and 6.

Can we describe the distribution of the number of progeny from time t (a time t cluster), $\mathbf{X}(t)$, using its distributional moments? For instance, in Figure 1, at time $t = t_2$ this distribution is given by (2, 2, 2), while for $t = t_1$ the distribution is (4, 2).

Note that a transmission does not always result in an observable coalescent event depending on which lineages expire due to recovery or are not sampled (*e.g.*, the transmission from 7 to 10 in Figure 1), and a transmission to an individual that recovers may still correspond to a coalescent event if that person transmits prior to recovering (*e.g.*, the transmission from 6 to 7 in Figure 1).

Coalescent model for SIR epidemics: In an SIR epidemic, a branch in the tree corresponds to a transmission event, and as a lineage is traced backward in time, it traverses multiple infected hosts. A recovery event before the sample time T does not alter the number of lineages with progeny because no progeny

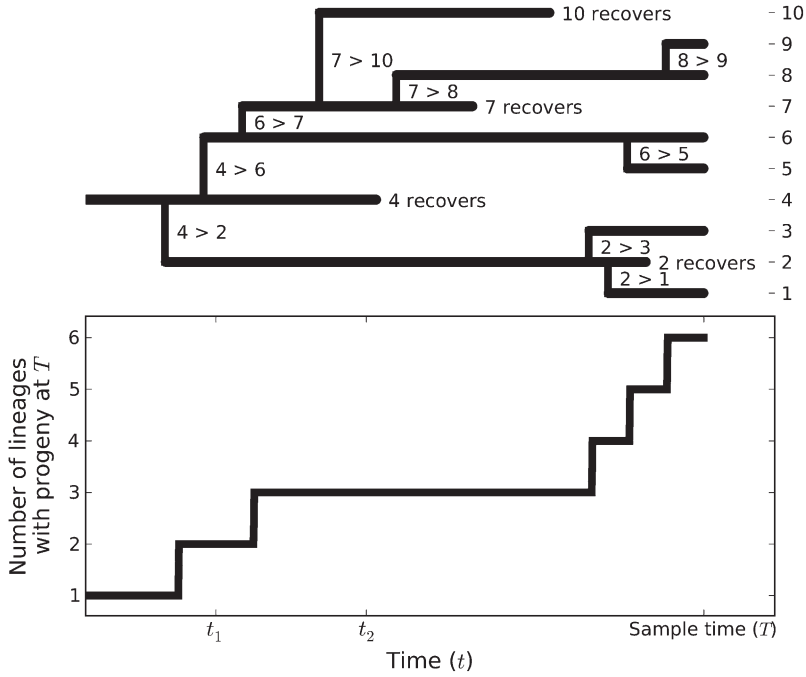

FIGURE 1.—An example of a phylogeny that could be generated by an epidemic process. The number of lineages at time t for a population observed at time T is plotted below. A branch in the tree corresponds to a transmission event, and as a lineage is traced backward in time, it traverses multiple infected hosts.

of this individual can be sampled at a later time. In a standard coalescent model, n lineages merge in reverse time at a rate proportional to $\binom{n}{2}$. Given that a coalescent event occurs among the individuals in $\mathcal{I}$, the probability of observing it among the n observed lineages is

$$\binom{n}{2}\Big/\binom{|\mathcal{I}|}{2} = \frac{n(n-1)}{|\mathcal{I}|\,(|\mathcal{I}|-1)}.$$

We introduce the dimensionless variable $A(t;\ T)$, which is the fraction of the population at t with sampled progeny extant at T. $A(t;\ T)$ is proportional to the number of ancestors of a sample of sequences and is analogous to the integer-valued ancestor function used in standard coalescent theory (GRIFFITHS and TAVARE 1994). We consider how A evolves as t moves into the past, with T fixed.

If a fraction ϕ of the infected population is sampled at time T, then we observe a number $n = \phi|\mathcal{I}(T)|$ lineages. Initially, $t = T$, and $A(T;\ T) = \phi I$ (the ancestor of each sequence is itself). The sample fraction ϕ is not always known, but if $\phi = 1$, our solution will describe the evolution of the fraction of extant lineages for the entire population.

Using the definition of A and assuming $A \gg 1/N$, the probability of a transmission event causing a coalescent event to be observed in our sample is

$$p_c(t;\ T) = \lim_{N\to\infty} \frac{\binom{A(t;\ T)N}{2}}{\binom{NI(t)}{2}} = \left(\frac{A(t;\ T)}{I(t)}\right)^2.$$

The rate of coalescence for a sample of sequences is analogous to the rate of change of the ancestor function, A. We can write the coalescence rate for the sample of sequences as the product of the number of transmissions per unit time, $f_{SI}(t)$ and the probability p_c that a transmission results in a coalescence being observed in our sample. The ancestor function $A(t;\ T)$ can be found by integrating the following backward ordinary differential equation from time T:

$$-\frac{dA}{dt} := \bar{A} = -f_{SI}p_c = -f_{SI}\left(\frac{A}{I}\right)^2. \tag{2}$$

This equation works even when $\phi = 1$, in which case A represents the number of ancestors of the entire population of infecteds observed at time T.

Cluster size distribution: Let $\mathbf{X}_1(t;\ T)$ denote the number of progeny at T of a random infected host from time $t \le T$, given that such progeny exist. We denote the expected value of $\mathbf{X}_1$ by $x_1(t;\ T)$ and interpret it as the *mean cluster size* from time t. $\mathbf{X}_2(t;\ T)$ [and $x_2 = E(\mathbf{X}_2)$] is a random variable that describes the size of the cluster if it is selected with probability proportional to the cluster's size. This is the same distribution of cluster sizes as if we select an infected at time T and determine the size of the cluster to which that infected belongs.

Below, we show that x_1 and x_2 can be found by integrating the ordinary differential equations

$$\bar{x_1}(t;\ T) = f_{SI}(t)I(T)/I(t)^2, \tag{3}$$

$$\bar{x_2} = 2\bar{x_1} \tag{4}$$

backward in time from T with initial prevalence $I(T)$ taken from the epidemic model. Also, initially (at $t = T$), all cluster sizes are unity, and $x_1(T;\ T) = x_2(T;\ T) = 1$.

The set of infecteds $\mathcal{I}(T)$ is distributed across a number $A(t;\ T)N$ clusters, and for any $0 \le t \le T$, the

average number of infecteds per time-t cluster is $I(T)/A(t;\,T)$. This implies

$$A(t;\,T) = I(T)/x_1(t;\,T). \tag{5}$$

Evaluating the backward derivative at t yields

$$\bar{A}^{\cdot} = -\bar{x}_1^{\cdot} I(T)/x_1^2. \tag{6}$$

Using Equation 6 in conjunction with Equations 2 and 5 yields Equation 3.

Dynamics of x_2 can be found by directly quantifying the mean field behavior of $\mathbf{X}_2$. Consider the size of a cluster to which a focal individual, a sampled infected at time T, belongs. As before, $p_c \times f_{SI}$ gives the rate of coalescence. Two clusters merge at each coalescent event, so there is a probability proportional to $2/A$ that a focal individual belongs to a cluster that takes part in the event. And given that the individual's cluster coalesces, the average amount by which the cluster increases is x_1. Multiplying these factors and probabilities together yields

$$\bar{x}_2^{\cdot} = p_c f_{SI} \frac{2}{A} x_1 = 2\bar{x}_1^{\cdot}. \tag{7}$$

As with x_1, this can be solved by integrating in reverse time with initial conditions $x_2(T;\,T) = 1$.

The variance of $\mathbf{X}_1$ can be found by noting that

$$\begin{aligned} E(\mathbf{X}_1^2) &= \sum_i i^2 \Pr\{\mathbf{X}_1 = i\} \\ &= \left(\sum_i i \Pr\{\mathbf{X}_1 = i\}\right)\left(\frac{\sum_i i^2 \Pr\{\mathbf{X}_1 = i\}}{\sum_i i \Pr\{\mathbf{X}_1 = i\}}\right). \end{aligned} \tag{8}$$

Recall that $\mathbf{X}_2$ is the size of a cluster selected with probability proportional to size, so

$$\Pr\{\mathbf{X}_2 = i\} = i \Pr\{\mathbf{X}_1 = i\} / \sum_j j \Pr\{\mathbf{X}_1 = j\}.$$

Combining the last two expressions with the definition of $x_1 = \sum_i i \Pr\{\mathbf{X}_1 = i\}$ gives

$$E(\mathbf{X}_1^2) = x_1 x_2.$$

Then, the variance in cluster size is

$$\mathrm{Var}(\mathbf{X}_1) = E(\mathbf{X}_1^2) - (E(\mathbf{X}_1))^2 = x_1 x_2 - x_1^2. \tag{9}$$

Higher moments can also be derived recursively from earlier moments. We now show that the nth moment of the CSD, M_n, can be found by solving the following differential equation with initial conditions $M_n(T) = 1$,

$$\bar{M}_n^{\cdot} = f_{SI} \frac{A}{I^2} \sum_{i=0}^{n-1} \binom{n}{i} M_i M_{n-i}, \tag{10}$$

where we define $M_0 := 1$ for convenience. Equations 3 and 4 could be derived using Equation 10 as a starting point.

Equation 10 is obtained by multiplying the rate at which a cluster merges with other clusters ($f_{SI}A/I^2$) and the expected change in the nth moment when two clusters merge. When a cluster of size i merges with a cluster of size j, the nth moment to be considered will change from that for a cluster of size i to that for a cluster of size $(i + j)$. To find the expected change in the nth moment when two clusters merge, we sum over all possible combinations of clusters of sizes i and j:

$$\begin{aligned} &\sum_i \sum_j \Pr\{\mathbf{X}_1 = i\}\Pr\{\mathbf{X}_1 = j\}(i+j)^n - i^n \\ &\quad = -M_n + \sum_i \Pr\{\mathbf{X}_1 = i\} \sum_j \Pr\{\mathbf{X}_1 = j\} \sum_{m=0}^{n} \binom{n}{m} i^{n-m} j^m \\ &\quad = -M_n + \sum_i \Pr\{\mathbf{X}_1 = i\} \sum_{m=0}^{n} \binom{n}{m} i^{n-m} \sum_j \Pr\{\mathbf{X}_1 = j\} j^m \\ &\quad = -M_n + \sum_i \Pr\{\mathbf{X}_1 = i\} \sum_{m=0}^{n} \binom{n}{m} i^{n-m} M_m \\ &\quad = -M_n + \sum_{m=0}^{n} \binom{n}{m} M_{n-m} M_m \\ &\quad = \sum_{m=0}^{n-1} \binom{n}{m} M_{n-m} M_m. \end{aligned}$$

The product of the coalescent rate $f_{SI}A^2/I^2$ and the factor $1/A$ that accounts for the probability that a focal lineage takes part in a coalescent event, along with the expected size function, yields Equation 10. In supporting information, Figure S1, we compare solutions of this equation to the second through fifth moments from simulations.

Fitting epidemic models to sequence data: If we know the branching times $t_1, t_2, \cdots, t_{n-1}$ for a phylogeny constructed from n sequences, we can use Equation 2 to fit an SIR model. In practice, there is considerable uncertainty about the exact genealogy and branching times given a sample of sequences. The theory developed here is based on a fixed genealogy with no uncertainty about branch lengths, but it should be straightforward to generalize these results to cope with error in the t_i (Drummond *et al.* 2005).

The total number of coalescent events observed between times t and T is proportional to $A(T;\,T) - A(t;\,T)$, and at some time $t < \tau < T$, the fraction of the coalescent events that have occurred is

$$F(\tau) = \frac{A(T;\,T) - A(\tau;\,T)}{A(T;\,T) - A(t;\,T)}. \tag{11}$$

This provides a cumulative distribution function for the distribution of coalescent times. Differentiating with respect to τ yields the density

$$-\bar{A}^{\cdot}/(A(T;\,T) - A(t;\,T)).$$

We make the approximation that when two lineages coalesce, the rates at which other lineages coalesce remain unchanged. Then each coalescent time will be an i.i.d. random variable with the distribution (11). The probability of observing a particular sequence of

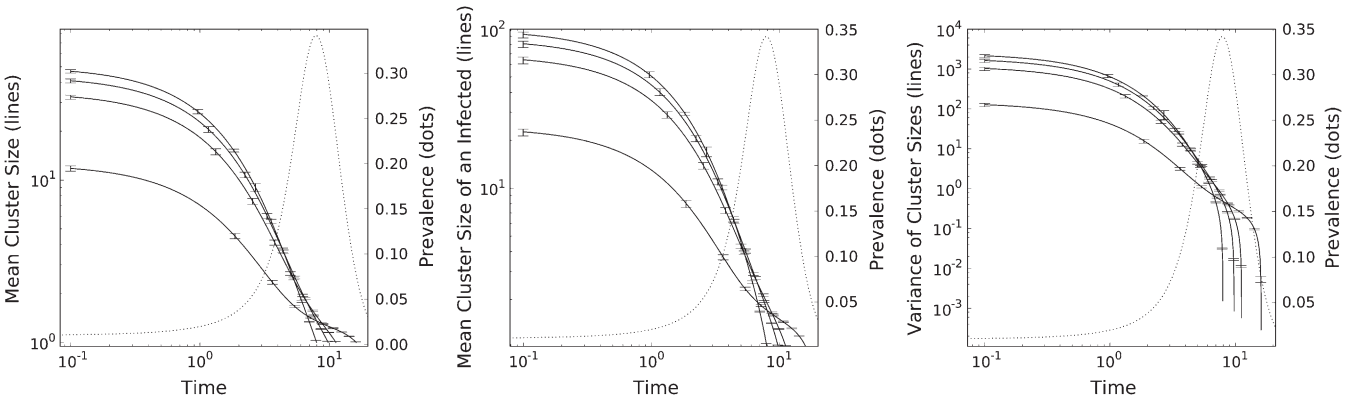

FIGURE 2.—The moments of the cluster size distribution over time as calculated by Equations 3 and 9 (lines, log scale). Four trajectories of the cluster size moments were generated for 4 sample times T. And for each trajectory, simulated moments were calculated for 10 threshold times t. Error bars show the 90% interval for 100 agent-based simulations [$N = 10^5$ and $I(0) = 1\%$]. The SIR model is $\dot{S} = -\beta SI$, $\dot{I} = \beta SI - \gamma I$, $\dot{R} = \gamma I$. Epidemic prevalence (dotted line) is shown on the right axis. Transmission rate $\beta = 1$, and recovery rate $\mu = 0.3$.

branching times will be proportional to the product of the density evaluated at each branching time. Consequently, we can construct the log-likelihood function out of an A-trajectory

$$\begin{aligned}\Lambda(t_1, \cdots, t_{n-1} \mid \theta) &= \sum_{i=1}^{n-1} \log(-\dot{A}(t_i)/(A(T) - A(t))) \\ &= -(n-1)\log(A(T;\, T) - A(t;\, T)) + \sum_{i=1}^{n-1} \log(-\dot{A}(t_i;\, T)),\end{aligned} \tag{12}$$

where θ denotes the parameters of the SIR model, such as transmission and recovery rates. In File S1 we also present a fitting criterion based on the Kolmogorov–Smirnov statistic for comparing distributions.

RESULTS

Equation 3 indicates some simple relationships that govern coalescent rates in epidemics. Coalescent rates are proportional to epidemic incidence (f_{SI}) and inversely proportional to square prevalence (I^{-2}). Rates will be highest when prevalence is low and incidence is high, such as at the beginning of an epidemic, during the expansion phase, or following a trough in prevalence.

Equation 9 implies that variance of the CSD asymptotically approaches the mean squared (Figure S4). This is similar to what is seen in the offspring distribution of forward time branching processes, such as the Galton–Watson process (ATHREYA and NEY 2004).

The point in time where the ancestor function (5) crosses the value $1/N$ is the point at which the phylogeny of the virus has collapsed to a single lineage—the MRCA of the sequences. Therefore, if we collect a sample of size n at time T, and solve Equation 2 to time zero, with $A(T) = n/N$, the time τ that satisfies $A(\tau) = 1/N$ corresponds to the time to the most recent common ancestor of the sample. Although our differential equations should not serve as an adequate description of the discrete valued processes for values close to $1/N$, we find that this approximation works quite well. A demonstration with comparison to simulations is provided in Figure S11.

Simulations: To assess the performance of our model, we carried out stochastic simulations of SIR epidemics. Simulations were individual based and in continuous time. Transmission events and recovery events were queued using exponentially distributed lag times, similar to the Gillespie algorithm. Each transmission event was recorded, which allowed us to simulate viral phylogenies under controlled conditions and to test the accuracy of Equations 3 and 9. The transmission data were then converted into phylogenetic trees with known branching times.

Simulation code was independently written by S. D. Frost and E. M. Volz in Python and C. Results from both models were compared to ensure accuracy.

To assess the accuracy of the equations we have derived, we developed a simulation experiment with 10^3 (1%) initially infected agents out of a population of total size $N = 10^5$ otherwise identical agents. Transmission and recovery rates were such that $R_0 = 10/3$. Figure 2 shows Equations 3 and 9 (lines) and the 90% confidence intervals from simulations at 10 thresholds (t values). The exact values of t and T are reported in File S1. Each trajectory corresponds to a cross-sectional census of the infected population at four time points (T values) corresponding to maximum prevalence, as well as 86, 68, and 22% of maximum prevalence after the peak. As we go backward in time, all moments of the CSD increase, until the entire census of infecteds falls into a single cluster. Many of the trajectories intersect, which demonstrates that the CSD is a complex function of both t and T and could therefore not be reduced to a simple forward-looking model.

Comparison with the generalized skyline: Further simulations were developed to test the suitability of the

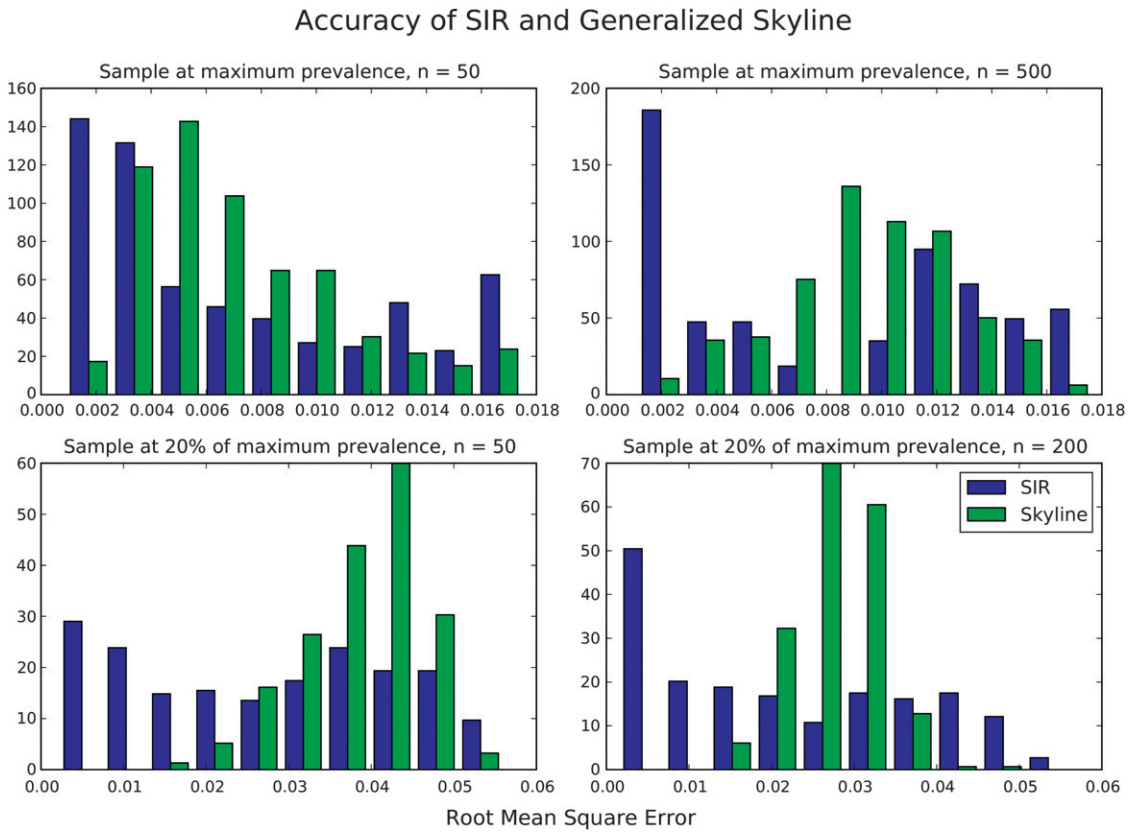

FIGURE 3.—Root mean square error of SIR and generalized skyline estimates of epidemic prevalence. Data are based on 300 simulated epidemics ($R_0 = 2$). RMSE is averaged over 100 time points.

model for estimating epidemiological parameters. When the number of infecteds is small, epidemic dynamics will be subject to large stochastic fluctuations. To determine if Equation 12 can be used to fit SIR models when the population size is small, we conducted a set of simulations with only a single initial infected in a population of 10,000 agents (Figure S5).

The simulations were also designed to determine if SIR models that are fit via likelihood Equation 12 can provide advantages beyond methods that are commonly used to estimate effective population size (N_e). For purposes of comparison, we used the generalized skyline model (OPGEN-RHEIN *et al.* 2005) (ape library in R) and compared the estimated effective population size to the best-fit SIR models and the known epidemic prevalence from simulations. Details of the simulations are provided in File S1.

We found that the accuracy of the best-fit SIR models exceeded that of the generalized skyline by 8–30% as measured by the root mean square error (RMSE) of estimated prevalence. It may seem surprising that the SIR model based on ODEs outperforms the generalized skyline even in the presence of stochasticity at small population sizes. This is due to the fact that population dynamics converge to the deterministic SIR model as the infected population increases in size. Fluctuating incidence due to stochastic effects when the number of infecteds is small has the effect of shifting the distribution of coalescence times to the left or the right, but does not fundamentally change the shape of the distribution. This is easily accounted for by including a parameter that varies the fraction initially infected in the deterministic SIR model.

Figure 3 shows the distribution of RMSE over 300 simulations. The mode of RMSE for the SIR model is zero for all experiments, whereas the skyline is slightly biased. Increasing sample size decreases RMSE for both SIR and skyline. Taking the sample at a later time (corresponding to 20% of peak prevalence) decreases the accuracy of both SIR and skyline, although in general the SIR models cope better with late sample times than does the skyline. In Figure S10, we show several representative SIR and skyline fits. It is usually the case that the generalized skyline fails to detect a decrease in prevalence and overestimates in the latter stages of the epidemic.

The SIR models also provide a quite accurate estimate of R_0 [$R_0 = 2$, $\hat{R}_0 = 1.95$ (95%: 1.71–2.17)].

The effect of a sample fraction: In the Kingman coalescent, the fraction of the population that is sampled is assumed to be small, such that the probability that more than two individuals have the same parent in the preceding generation is negligible. For example,

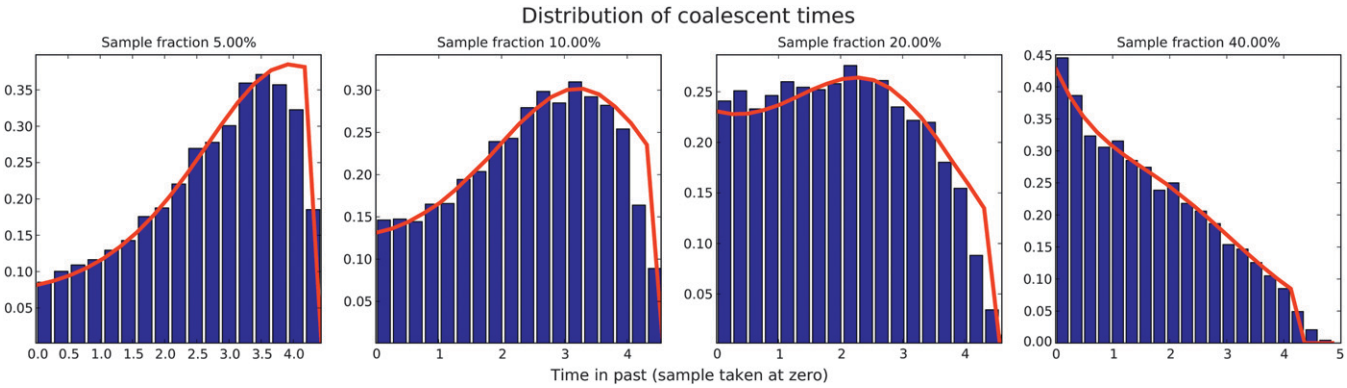

FIGURE 4.—The empirical distribution of coalescence times based on 150 simulated SIR epidemics. Transmission rate = 2, recovery rate = 1. The expected distribution based on Equation 11 is shown in red.

Kingman showed that the probability that n sampled sequences will not have a common ancestor in the preceding generation is

$$\prod_{i<n}(1-i/N) = 1-\sum_{i<n}\frac{i}{N}+O(N^{-2}) = 1-\binom{n}{2}/N+O(N^{-2}).$$

Kingman then made the approximation that the $O(N^{-2})$ terms are zero, which yields a minimum requirement that $n<\sqrt{2N}$.

Analytical work has been carried out to investigate the effect on coalescent processes of violating the assumption of a small sample fraction (see, for example, FU 2006), using discrete mathematics similar to the original Kingman model. Such work has indicated that the Kingman coalescent can be a surprisingly good approximation even when the sample fraction is large.

Nevertheless, our model is not an approximation and takes the sample fraction into account. This gives some insight into how the fraction of the infected population sampled affects the distribution of coalescent times and thus the shape of the reconstructed phylogeny of viral sequences.

Figure 4 shows the empirical distribution of coalescence times for 150 simulations ($R_0 = 2$) with samples taken at peak prevalence. The sample fraction was varied from 5 to 40%. When the sample fraction is small (5%), the distribution is skewed left, meaning the phylogeny is starlike, which is in agreement with conventional notions for an exponentially growing population. However, as the sample fraction is increased to 10, 20, and 40%, the shape of the distribution changes until it is skewed right, which means that most of the branches occur close to the tips. These qualitatively antipodal distributions are generated by the same underlying population dynamics, with only the sample fraction being varied. This observation is of practical as well as theoretical interest, since many serological surveys for HIV may reach >20% of infected individuals within a given locality (LEWIS *et al.* 2008).

Equation 11 gives the analytical distribution of coalescence times and is shown in red in Figure 4. It also provides some simple intuition for why most coalescence events will happen close to the sample time (T) when the sample fraction is large. We use the initial conditions $A(T) = n/N$, so that when n is large, the term $(A(T)/I(T))^2$ is also large, which is the probability that two individuals in a transmission event represent sample lineages. Conversely, if n and $(A(T)/I(T))^2$ are small, fewer coalescent events will occur until I converges to A, which will occur early in the epidemic.

Estimating HIV prevalence: Equation 2 gives the rate of coalescence at any time prior to the sample time (T) and, by extension, the distribution of coalescence times. This allowed us to derive the likelihood function (12), which we used to fit a simple mass-action SIR model to 55 HIV-1 sequences of the *pol* gene collected as part of the ACTG241 clinical trial (D'AQUILA *et al.* 1996; LEIGH BROWN *et al.* 1999). All sequences were collected from men who have sex with men (MSM) over a short period of time (May to July, 1993) within the United States. Because the sequences were collected within a short window of time, it is valid to make the approximation that all sequences were sampled simultaneously. To estimate a phylogeny, we used a general-time-reversible model of nucleotide substitution (TAVARE 1986) with gamma-distributed variation in site-to-site substitution rates. The root giving the most clocklike rates was determined by maximum likelihood and the null hypothesis of a molecular clock could not be rejected at the 5% significance level.

The epidemiology of HIV has several factors that are important to include in a model. Upon infection, individuals progress through an acute phase lasting 1–3 months and then progress to a chronic phase lasting many years. The transmission probability per act is much greater during the acute phase. Furthermore, since we wish to model the epidemic over a period of 25 years, we must consider natural mortality and immigration into the susceptible pool. All of these factors are considered in the following model:

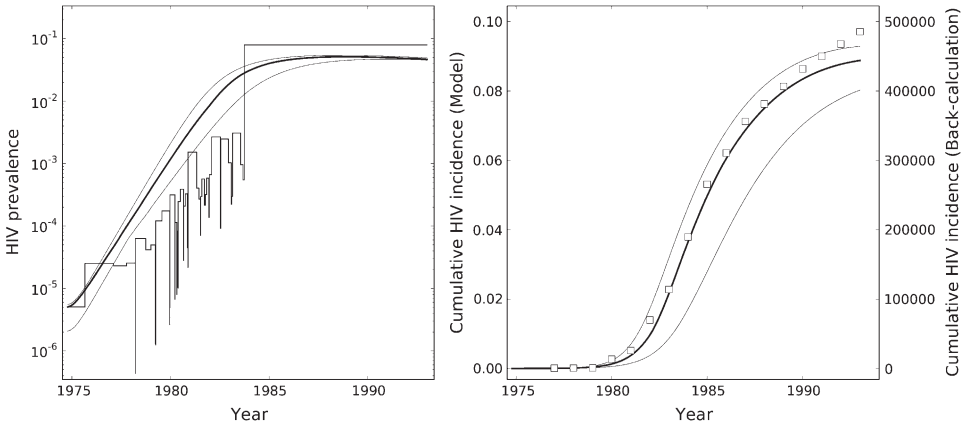

FIGURE 5.—Left: Estimated epidemic prevalence (logarithmic scale) of HIV among MSM in the United States. A solution to Equation 16 is compared to the skyline plot, rescaled such that minimum effective population size equals minimum prevalence. The thin lines show 95% confidence intervals. Right: Estimated cumulative incidence of HIV among MSM *vs.* time (years prior to 1993). A solution to Equation 16 is compared to estimates based on sero-surveillance data (HALL *et al.* 2008).

$$\dot{S} = -S^{\alpha}(\beta_1 I_1 - \beta_2 I_2) + \mu - \mu S \quad (13)$$

$$\dot{I}_1 = S^{\alpha}(\beta_1 I_1 + \beta_2 I_2) - \gamma_1 I_1 - \mu I_1 \quad (14)$$

$$\dot{I}_2 = \gamma_1 I_1 - \gamma_2 I_2 - \mu I_2. \quad (15)$$

I_1 and I_2 respectively represent the fractions of the population that are at the acute and the chronic stages of infection. Parameters we wish to estimate include the following:

β_1: The transmission rate of acute infecteds.
β_2: The transmission rate of chronic infecteds.
μ: The immigration rate into the susceptible population and the natural mortality rate. We consider immigration to balance natural mortality.
α: A parameter that controls how incidence scales with cumulative incidence.
ε: The fraction of the population infected at the TMRCA of the sample.

Many more parameters could be included in a model for HIV among MSM, but since our purpose is to fit a model to only 55 sequences, we choose to keep the number of free parameters to a minimum. In addition, we assumed an acute phase that lasts 2 months on average ($\gamma_1 = 1/60$) and a chronic phase that lasts 10 years on average [$\gamma_2 = 1/(10 \times 365)$].

Prior distributions are given in File S1.

Given $n = 55$ sequences, we use the initial conditions $A(T) = 55/N$, $I_1(0) = \varepsilon$, and $S(0) = 1 - \varepsilon$. Since we are including equations for two types of infecteds, we must keep track of ancestor functions for both types. A_1 and A_2 are the fractions of the population that are respectively acute and chronic infected and that have sampled progeny at time T. We have

$$\dot{A_2} = -\gamma_1 I_1 (A_2/I_2) + \beta_2 I_2 S^{\alpha}(A_1/I_1)((I_2 - A_2)/I_2) \quad (16)$$

$$\dot{A_1} = \gamma_1 I_1 (A_2/I_2) - \beta_1 I_1 S^{\alpha}(A_1/I_1)^2 - \beta_2 I_2 S^{\alpha}(A_1/I_1). \quad (17)$$

For purposes of fitting the SIR model, we use $A = A_1 + A_2$ and $\dot{A} = \dot{A_1} + \dot{A_2}$. A derivation is provided in File S1.

Fitting the model proceeded in two steps. First, we fit a model using Equation 12 as described above. The second step made use of sero-surveillance data of MSM in the United States (HALL *et al.* 2008). These data provided estimates of HIV incidence based on back calculation for the period 1977–2006. To ameliorate error from uncertainty in the chronological values of phylogenetic branch lengths, we adjusted the timescale of the epidemic and rescaled estimated rates to gain the greatest fit with incidence data by a least-squares criterion.

Figure 5 shows the best-fit SIR model. The median posterior estimates were as follows: acute transmission rate, $\hat{\beta}_1 = 1$ transmission per 47 days; chronic transmission rate, $\hat{\beta}_2 =: 1$ transmission per 1207 days; immigration rate to susceptible state, $\hat{\mu} = 1$ per 19.5 years; and incidence scaling parameter, $\hat{\alpha} = 9.77$. Together, these parameters imply an R_0 value of 2.24 (see File S1). They also imply that 41% of transmissions occur during the acute stage.

For comparison with our SIR model, effective population size (N_e) was calculated using the skyline plot (PYBUS *et al.* 2000). N_e was rescaled so that $\min(N_e) = \min(I)$. Figure 5 shows the rescaled skyline and an SIR trajectory that was integrated with parameters from the median of the posterior distribution. Confidence intervals are also given, which show the upper and lower bounds within which 95% of posterior epidemic prevalence falls. Figure 5 also compares the best-fit SIR model with the estimated cumulative incidence among MSM in the United States based on sero-surveillance data. The SIR model is in broad agreement with the data from public health sources regarding the early rate of growth and saturation in the early 1990s. The skyline also reproduces the growth rate during the expansion phase and the tapering of epidemic growth in the early 1990s. However, the skyline predicts a rise in N_e between 1980 and 1993, which probably overestimates the true prevalence.

We have also compared the CSD mean and variance from our best-fit SIR model to the empirical values from the ACTG241 data (Figure 6). The SIR model successfully reproduces the mean cluster size throughout the

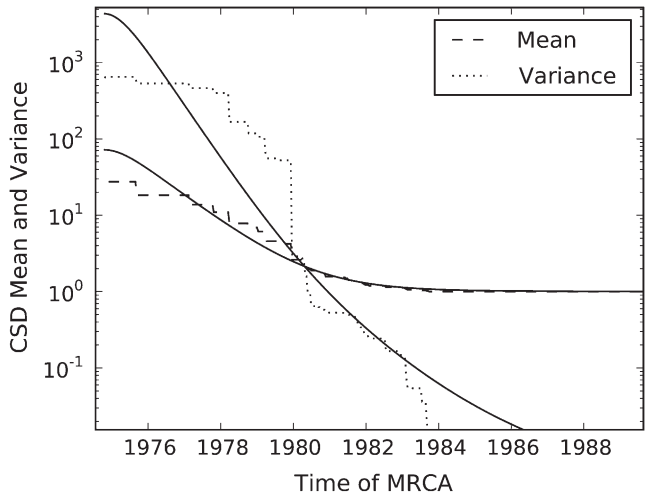

FIGURE 6.—The mean cluster size (dashes) and variance of cluster sizes (dotted line) are calculated from the empirical observations from the ACTG241 sequences (dashed lines) and compared to our best-fit SIR model (solid lines). The horizontal axis gives the clustering threshold as the year of the MRCA of a cluster.

course of the epidemic. However, there is substantial deviation between the actual and the predicted variance of cluster sizes. As the clustering threshold is increased, all sampled infecteds eventually fall within a single cluster, and in a finite population, variance converges to zero (not shown).

DISCUSSION

The distribution of cluster sizes is a function of the time T at which we observe a population, such as by taking a sample of sequences, and $t < T$, which is a clustering threshold (if the MRCA of two sequences occurs after t, then those sequences are clustered). We have derived differential equations that describe how the moments of the CSD change as the threshold t moves into the past. This could be used to calculate the distribution of cluster sizes to arbitrary precision at any time. It is straightforward to use the model to calculate the probability that an infected host will have viral progeny at a later time point and, conversely, the expected number of ancestor lineages of a sample taken at T. The model promises to serve as a null hypothesis for clustering of infecteds under various epidemiological scenarios and could possibly be used to detect effects that may distort the CSD such as selection and population structure.

The CSD is sensitive to details of the underlying population dynamics. Most coalescent approaches take into account only variable population size, such as epidemic prevalence, but not variable birth rates, analogous to epidemic incidence. Such approaches can give misleading results for epidemics. For example, in both susceptible–infected (SI) models (no recovery) and susceptible–infected–susceptible (SIS) models (recovery into the susceptible state), prevalence rapidly approaches an equilibrium. However, a naive coalescent model based on constant population size would erroneously predict identical coalescent patterns in these two cases. In fact, the SIS case is very similar to a standard constant-population size coalescent, but the lineages in an SI epidemic coalesce only during exponential growth, not at equilibrium (Figure S2 and Figure S3).

We observed drastically less precision when estimating recovery rates than when estimating transmission rates. Consequently, decline in prevalence is much harder to detect than growth. This has been observed previously (Lavery *et al.* 1996) in other biological systems due to differences in the timescale of population change and genetic variation. We nevertheless found that our estimation procedure is robust to misspecification of priors that include zero recovery, and it is feasible to distinguish SI from SIR dynamics (Figure S6, Figure S7, Figure S8, and Figure S9).

In conclusion, coalescent-based estimates of effective population size, such as the generalized skyline, have wide applicability and require minimal consideration of underlying population dynamics. However, in the case that the epidemic dynamics are well understood, the potential is raised for a population genetic model that takes into account the precise effects of transmission and recovery, thereby predicting population dynamics with greater accuracy. We have developed a model that provides a step toward the formal integration of phylodynamics and epidemiology and that can be used to estimate epidemiological and demographic parameters directly from viral sequence data.

Fitting population models to data requires biological simplifications to make the model tractable, which presents the danger of making the model useless for real systems (Wilson *et al.* 2005). Pathogens require successful reproduction both within and between hosts, whereas we have focused entirely on transmission of lineages to uninfected and immunologically naive hosts. We have not considered biological nuances such as superinfection and recombination or the possibility that different strains will have different epidemiological characteristics. Consequently, there are many ways that our model could be extended and improved.

We have calculated coalescent rates and CSD moments only for the most simple mass-action SIR models. But modern mathematical epidemiology has progressed in the direction of incorporating variable host susceptibility, pathogen virulence, geographical heterogeneity, and host contact network structure. Reproducing our derivations for such models would be a difficult but worthy enterprise.

While we have focused on variable population size in epidemics, a second pillar of phylodynamics concerns the effects of immune selection on viral phylogenies (Grenfell *et al.* 2004). A major limitation of our

approach is that we adopt the standard assumption of selective neutrality. It is unknown how our method would perform for genes under strong immune selection, such as influenza virus hemagglutinin.

We have made a first attempt at a method for fitting arbitrary SIR models to cross-sectional samples of viral sequences. Many challenges remain for increasing the utility of the method. It may be possible to improve estimation of model parameters when historical prevalence data are available. However, it is not known how to discriminate between competing models when only sequence data are available. The estimation theory developed here is based on a fixed genealogy of virus with no uncertainty about branch lengths; in reality there can be a great deal of uncertainty about the structure of the genealogy, and it should be straightforward to generalize the method to account for this (Drummond *et al.* 2005). Finally, it should also be possible to extend our solutions to heterochronous samples—sequence data collected at multiple time points over the course of an epidemic.

Irene Hall provided estimates of HIV incidence in MSM. The authors acknowledge support from the National Institutes of Health (T32 AI07384, R01 AI47745). S.D.W.F. is supported by a Royal Society Wolfson Research Merit Award. M.J.W. is supported by the Biotechnology and Biological Sciences Research Council.

LITERATURE CITED

Anderson, R. M., and R. M. May, 1991 *Infectious Diseases of Humans: Dynamics and Control.* Oxford University Press, London/New York/Oxford.

Athreya, K. B., and P. E. Ney, 2004 *Branching Processes.* Dover, New York.

Bailey, N. T. J., 1975 *The Mathematical Theory of Infectious Diseases and Its Applications.* Hafner Press, New York.

Brenner, B. G., M. Roger, J. Routy, D. Moisi, M. Ntemgwa *et al.*, 2007 High rates of forward transmission events after acute/early HIV-1 infection. J. Infect. Dis. **195:** 951.

Brown, A. J., H. F. Günthard, J. K. Wong, R. T. D'Aquila, V. A. Johnson *et al.*, 1999 Sequence clusters in human immunodeficiency virus type 1 reverse transcriptase are associated with subsequent virological response to antiretroviral therapy. J. Infect. Dis. **180:** 1043–1049.

D'Aquila, R. T., M. D. Hughes, V. A. Johnson, M. A. Fischl, J. P. Sommadossi *et al.*, 1996 Nevirapine, zidovudine, and didanosine compared with zidovudine and didanosine in patients with HIV-1 infection: a randomized, double-blind, placebo-controlled trial. Ann. Intern. Med. **124:** 1019–1030.

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

Drumright, L. N., and S. D. W. Frost, 2008 Sexual networks and the transmission of drug-resistant HIV. Curr. Opin. Infect. Dis. **21:** 644.

Fu, Y., 2006 Exact coalescent for the Wright–Fisher model. Theor. Popul. Biol. **69:** 385–394.

Goodreau, S. M., 2006 Assessing the effects of human mixing patterns on HIV-1 interhost phylogenetics through social network simulation. Genetics **172:** 2033–2045.

Grenfell, B. T., O. G. Pybus, J. R. Gog, J. L. N. Wood, J. M. Daly *et al.*, 2004 Unifying the epidemiological and evolutionary dynamics of pathogens. Science **303:** 327.

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

Hall, H., R. Song, P. Rhodes, J. Prejean, Q. An *et al.*, 2008 Estimation of HIV incidence in the United States. J. Am. Med. Assoc. **300:** 520.

Hue, S., D. Pillay, J. P. Clewley and O. G. Pybus, 2005 Genetic analysis reveals the complex structure of HIV-1 transmission within defined risk groups. Proc. Natl. Acad. Sci. USA **102:** 4425–4429.

Kaj, I., and S. M. Krone, 2003 The coalescent process in a population with stochastically varying size. J. Appl. Probab. **40:** 33–48.

Kermack, W. O., and A. G. McKendrick, 1927 A contribution to the mathematical theory of epidemics. Proc. R. Soc. Lond. Ser. A, *Containing Papers of a Mathematical and Physical Character*, **115:** 700–721.

Kingman, J. F. C., 1982a On the genealogy of large populations. J. Appl. Probab. **19:** 27–43.

Kingman, J. F. C., 1982b The coalescent. Stoch. Proc. Appl. **13:** 235–248.

Lavery, S., C. Moritz and D. R. Fielder, 1996 Genetic patterns suggest exponential population growth in a declining species. Mol. Biol. Evol. **13:** 1106–1113.

Lewis, F., G. J. Hughes, A. Rambaut, A. Pozniak and A. J. Leigh Brown, 2008 Episodic sexual transmission of HIV revealed by molecular phylodynamics. PLoS Med. **5:** e50.

Nee, S., E. C. Holmes, A. Rambaut and P. H. Harvey, 1996 Inferring population history from molecular phylogenies, pp. 66–80 in *New Uses for New Phylogenies*, edited by P. H. Harvey, A. J. Leigh Brown, J. Maynard Smith and S. Nee. Oxford University Press, Oxford.

Opgen-Rhein, R., L. Fahrmeir and K. Strimmer, 2005 Inference of demographic history from genealogical trees using reversible jump Markov chain Monte Carlo. BMC Evol. Biol. **5:** 6.

Pao, D., M. Fisher, S. Hué, G. Dean, G. Murphy *et al.*, 2005 Transmission of HIV-1 during primary infection: relationship to sexual risk and sexually transmitted infections. AIDS **19:** 85.

Pybus, O. G., A. Rambaut and P. H. Harvey, 2000 An integrated framework for the inference of viral population history from reconstructed genealogies. Genetics **155:** 1429–1437.

Pybus, O. G., M. A. Charleston, S. Gupta, A. Rambaut, E. C. Holmes *et al.*, 2001 The epidemic behavior of the hepatitis C virus. Science **292:** 2323–2325.

Robbins, K. E., P. Lemey, O. G. Pybus, H. W. Jaffe, A. S. Youngpairoj *et al.*, 2003 US human immunodeficiency virus type 1 epidemic: date of origin, population history, and characterization of early strains. J. Virol. **77:** 6359–6366.

Rosenberg, N. A., and M. Nordborg, 2002 Genealogical trees, coalescent theory and the analysis of genetic polymorphisms. Nat. Rev. Genet. **3:** 380–390.

Tavare, S., 1986 Some probabilistic and statistical problems in the analysis of DNA sequences, pp. 57–86 in *Lectures on Mathematics in the Life Sciences.* American Mathematical Society, Providence, RI.

Wilson, D. J., D. Falush and G. McVean, 2005 Germs, genomes and genealogies. Trends Ecol. Evol. **20:** 39–45.

Yerly, S., S. Vora, P. Rizzardi, J. P. Chave, P. L. Vernazza *et al.*, 2001 Acute HIV infection: impact on the spread of HIV and transmission of drug resistance. AIDS **15:** 2287.

Yusim, K., M. Peeters, O. G. Pybus, T. Bhattacharya and B. Korber, 2001 Using human immunodeficiency virus type 1 sequences to infer historical features of the acquired immune deficiency syndrome epidemic and human immunodeficiency virus evolution. Philos. Trans. R. Soc. B Biol. Sci. **356:** 855–866.

Communicating editor: M. W. Feldman

GENETICS

Supporting Information
http://www.genetics.org/cgi/content/full/genetics.109.106021/DC1

Phylodynamics of Infectious Disease Epidemics

Erik M. Volz, Sergei L. Kosakovsky Pond, Melissa J. Ward, Andrew J. Leigh Brown and Simon D. W. Frost

DOI: 10.1534/genetics.109.106021

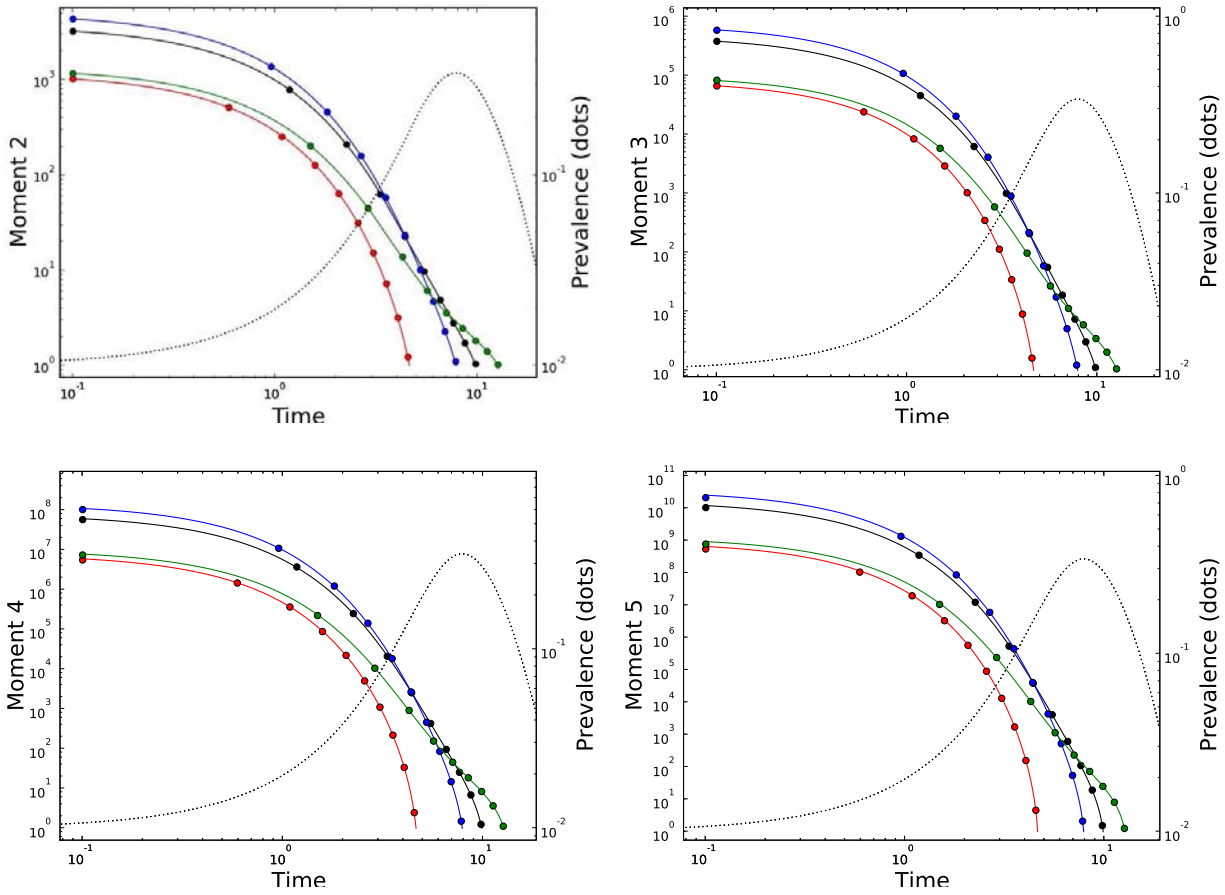

FIGURE S1.—Four trajectories of the 2nd through 5th moments of the cluster size distribution over time. Each trajectory corresponds to a cross-sectional census of the infected population at four time-points (T values). The SIR model is $\dot{S} = -\beta SI, \dot{I} = \beta SI - \gamma I, \dot{R} = \gamma I$. Transmission rate $\beta = 1$, and recovery rate $\gamma = 0.3$. And for each trajectory, simulated moments were calculated for ten threshold times t. The median outcome from one hundred agent-based simulations ($N = 10^5$ and $I(0) = 1\%$) is shown with points. Epidemic prevalence (dotted line) is shown on right axis.

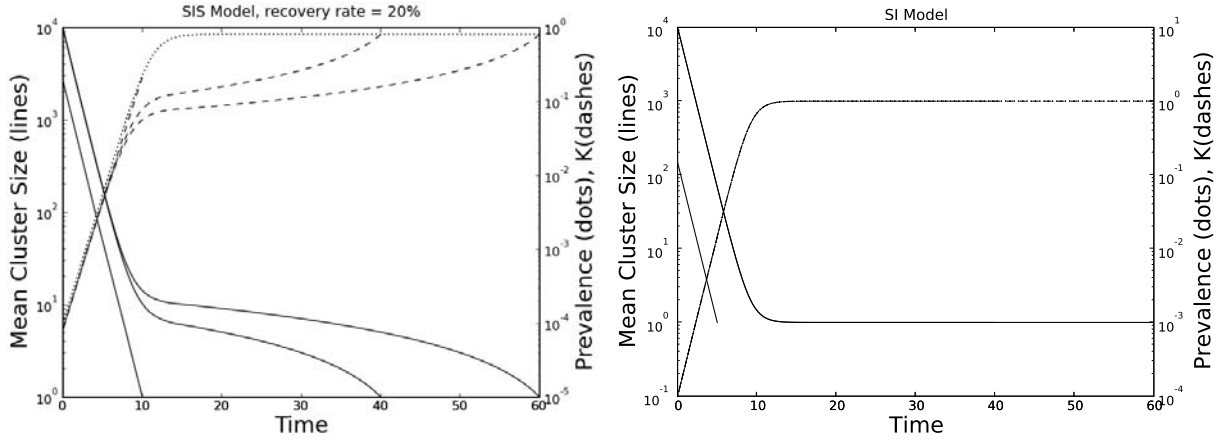

FIGURE S2.—The mean cluster size (x_1) starting from $t = 10, 40$, and 60 (solid line, left axis), is shown with epidemic prevalence (dots, right axis) and A (dashes). The left panel shows an SIS epidemic ($R_0 = 5$) and the right panel shows an SI epidemic.

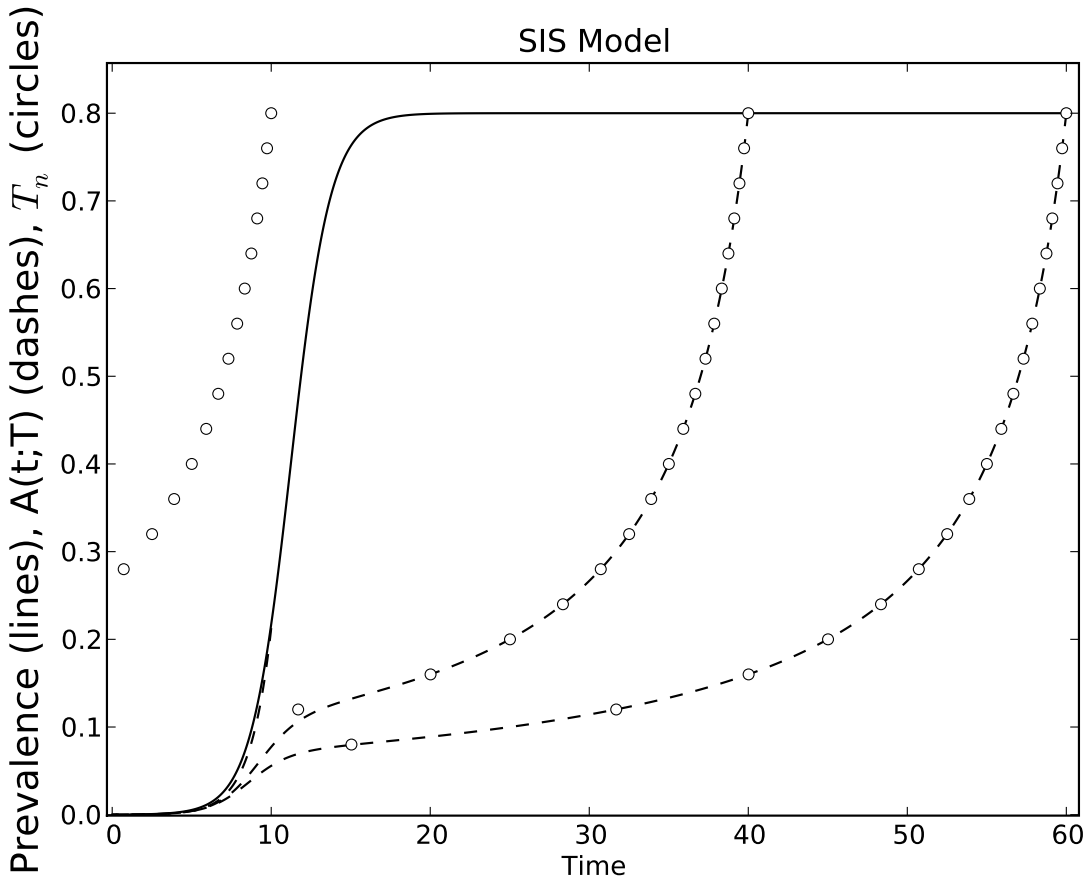

FIGURE S3.—A comparison of standard estimates of the expected time to the m'th coalescent, and the variable A from our model. The model is SIS with $R_0 = 5$.

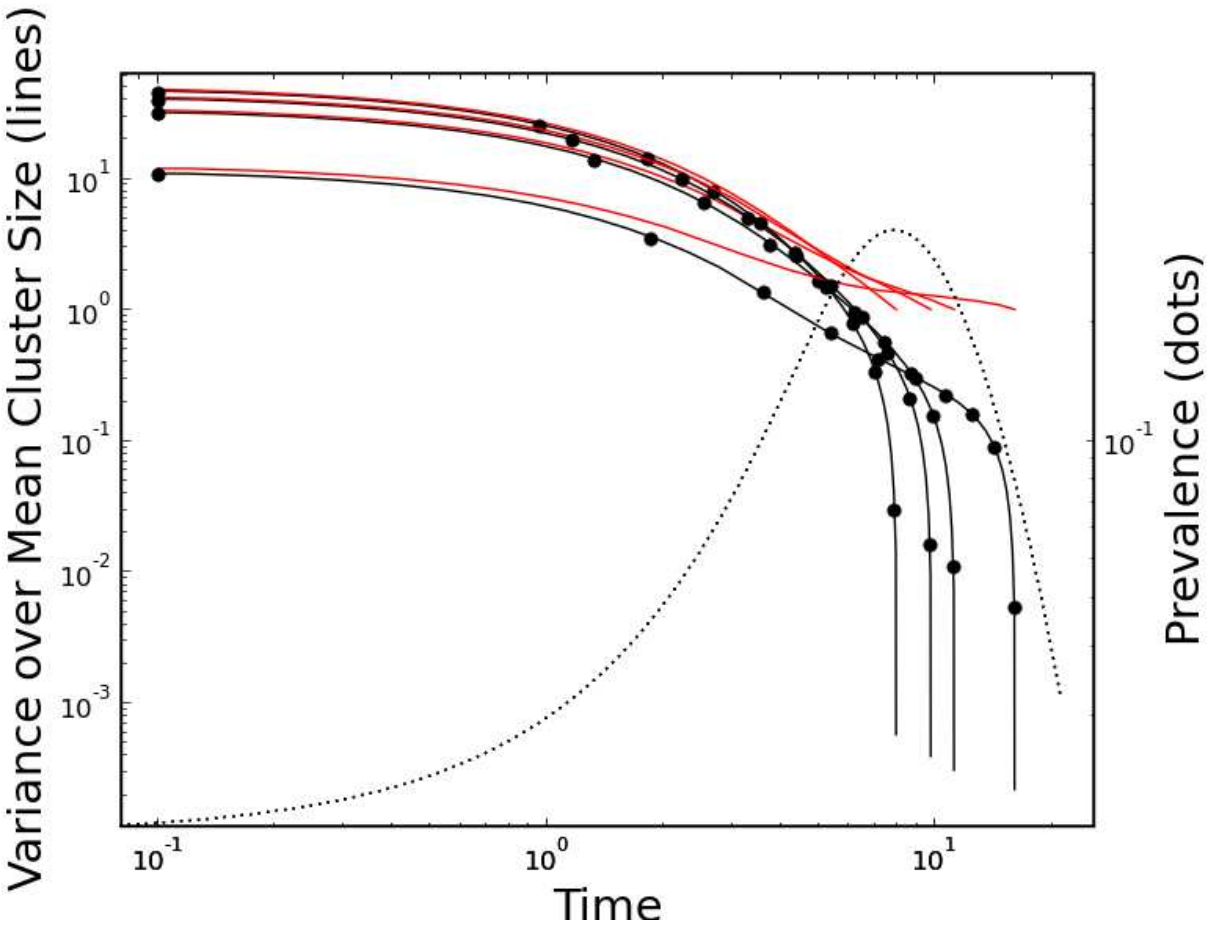

FIGURE S4.—The mean of the CSD (red, left axis) starting from four times over the course of the epidemic, corresponding to prevalence (dots, right axis) of 100%, 86%, 68% and 22% of the maximum. Also shown is the variance of the CSD over mean (solid lines, left axis). Dots (left axis) show the median of simulations.

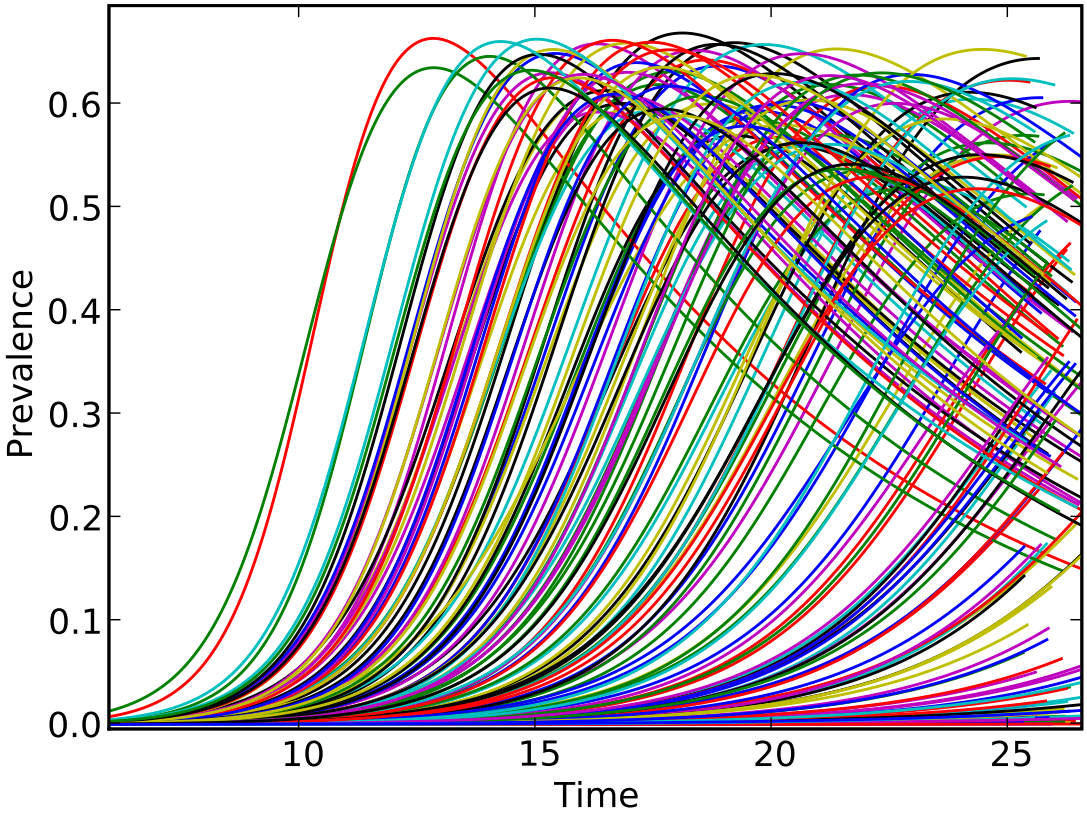

FIGURE S5.—Prevalence over time in 288 solutions to the SIR model. Each replicate is based on a set of parameters drawn independantly from the joint distribution described in Table 1 of the main text.

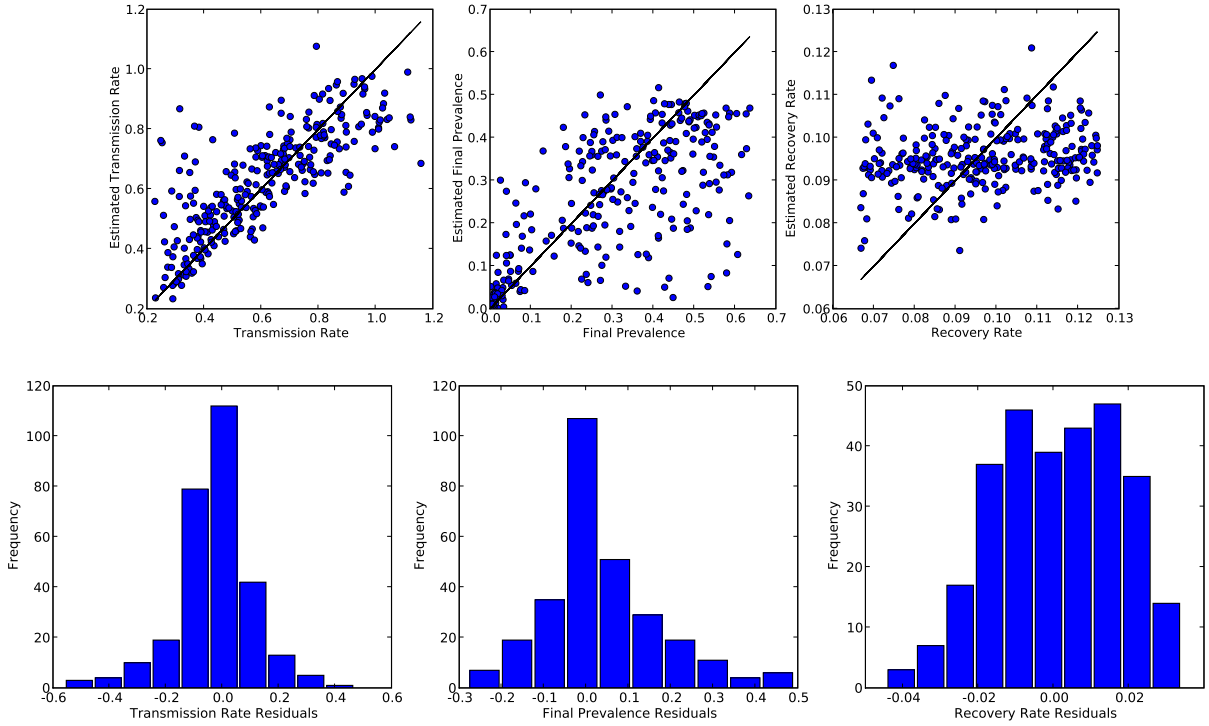

FIGURE S6.—Top: The estimated transmission rate, estimated final prevalence, and estimated recovery rates are shown versus the actual values in 288 replicates. Bottom: Histograms of residuals are shown for transmission rates, final prevalence, and recovery rates.

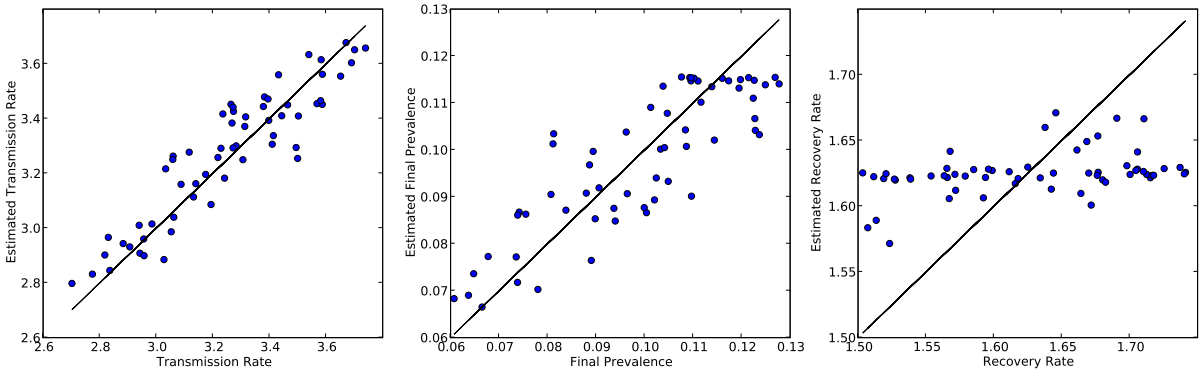

FIGURE S7.—The estimated transmission rate, estimated final prevalence, and estimated recovery rates are shown versus the actual values in 60 replicates. The prior distribution was chosen to have a small R_0 in the range 1-3.

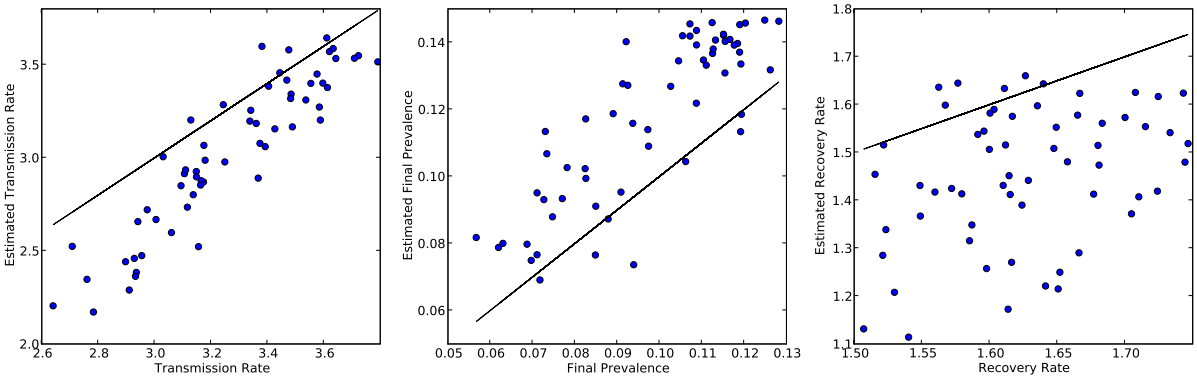

FIGURE S8.—The estimated transmission rate, estimated final prevalence, and estimated recovery rates are shown versus the actual values in 60 replicates. The prior distribution was chosen to have a small R_0 in the range 1-3. When estimating γ, the prior distribution was intentionally mis-specified.

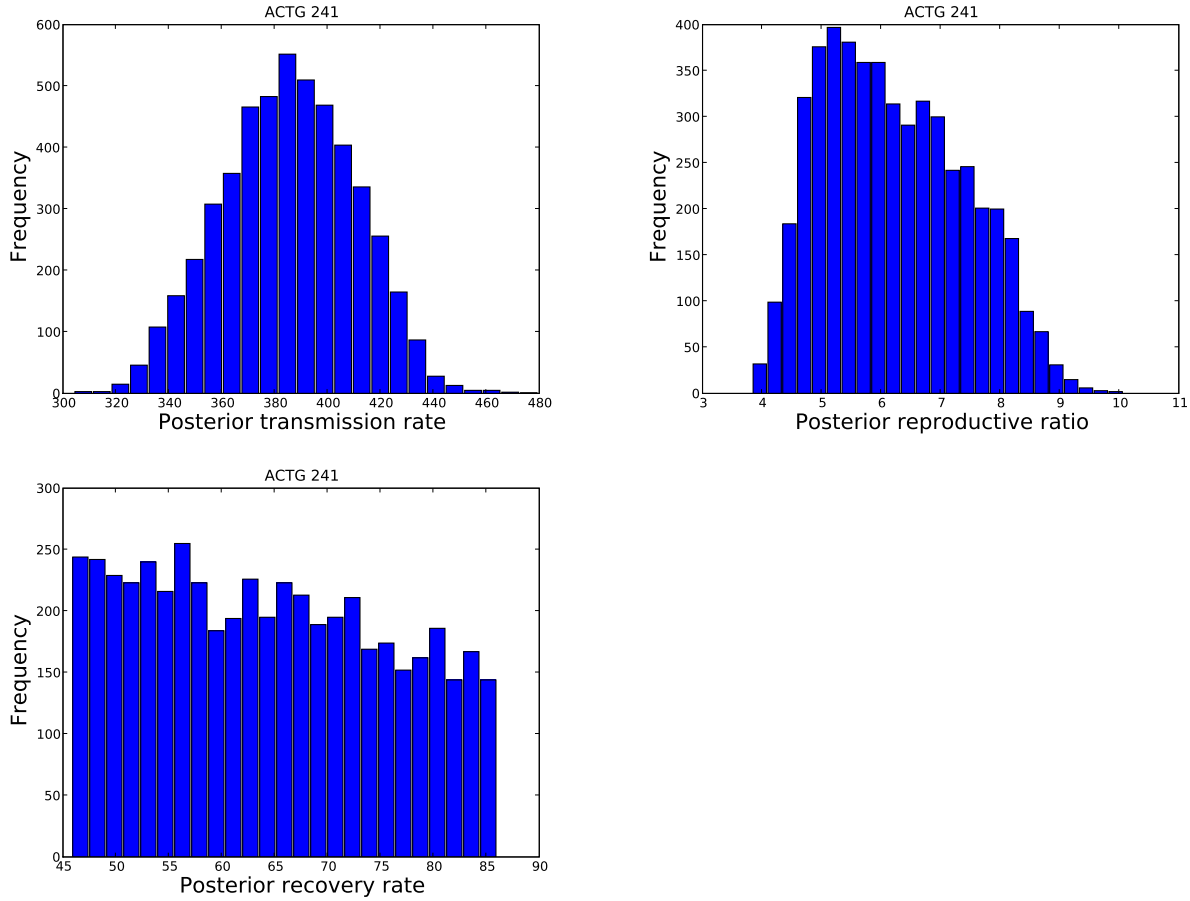

FIGURE S9.—Posterior distributions based on 5000 resamples for the transmission rate, reproductive ratio and recovery rate.

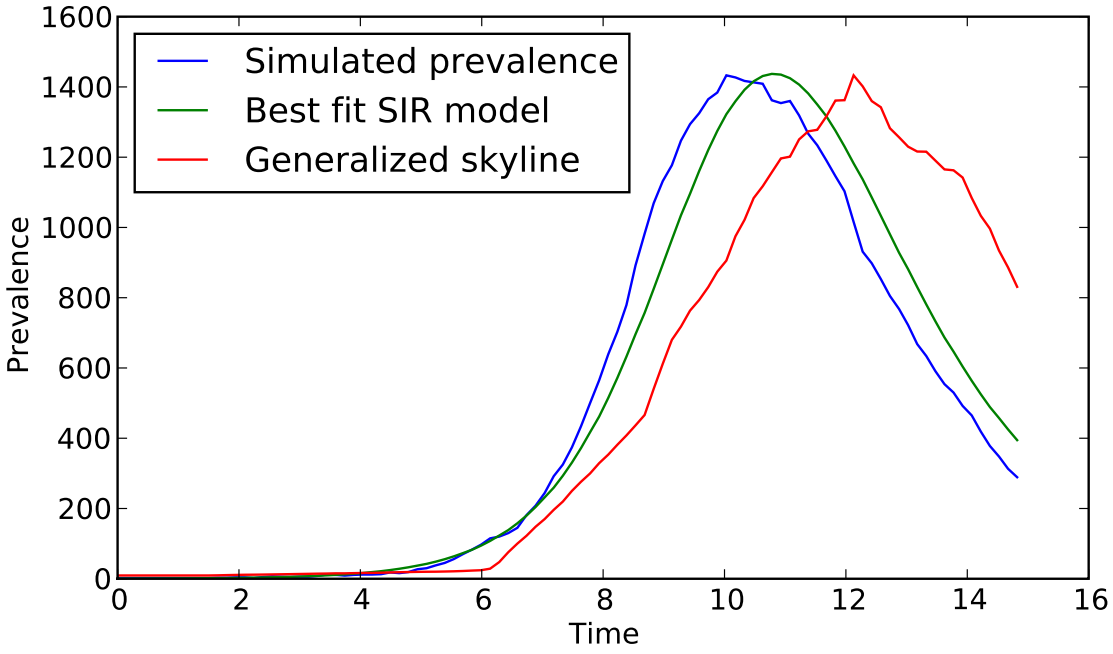

FIGURE S10.—Epidemic prevalence for a single simulation is compared to the best-fit SIR model and the generalized skyline. This instance is typical of many other simulations insofar as the generalized skyline usually fails to detect a drop in prevalence until long after it has occured.

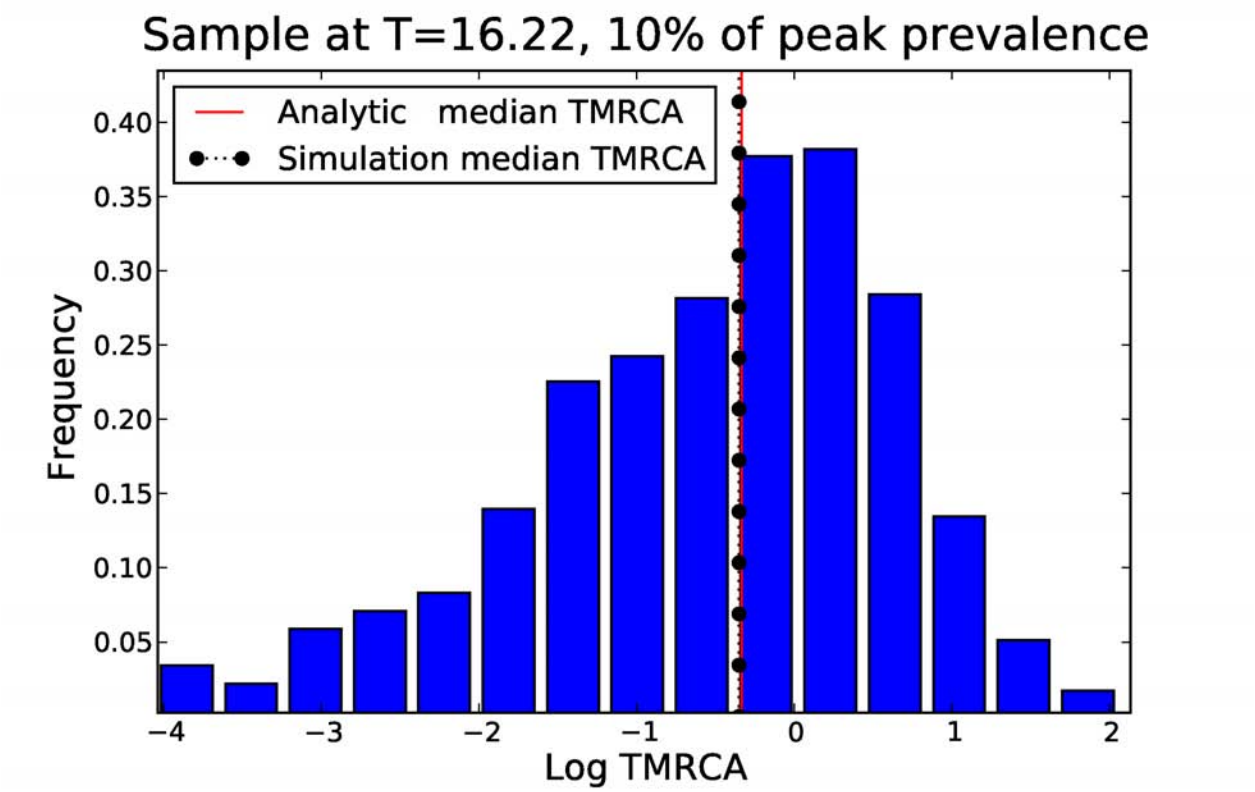

FIGURE S11.—The empirical distribution of the TMRCA of a sample of size 50 in an SIR epidemic. Theoretical and emprical estimates of the median TMRCA are shown as vertical lines.

FILE S1

1 Higher moments and simulations

Figure S1 shows four solutions to

$$\dot{\bar{M}}_n = f_{SI}\frac{A}{I^2}\sum_{i=0}^{n-1}\binom{n}{i}M_i M_{n-i}. \quad (1)$$

The epidemic model is the same as used in the text, with initial T values corresponding to 100% of peak prevalence, and 50% and 85% of peak prevalence before the peak, as well as 50% of peak prevalence after the peak.

2 SIS and SI dynamics

Equations 2, 4 and 5 in the main text correctly predict CSD moments in SI and SIS epidemics as well as the SIR model presented in the main text.

In figure S2 we compare prediction and theory for an SIS model with a recovery rate = 20% and a transmission rate of unity. The model has

$$f_{SI} = \beta SI, \dot{S} = -f_{SI} + \gamma I, \dot{I} = f_{SI} - \gamma I.$$

We also examine an SI model with a recovery rate of 20%, described by

$$f_{SI} = \beta SI, \dot{S} = -f_{SI}, \dot{I} = f_{SI}.$$

The population is observed at three time points: 10, 40 and 60 sec. Prevalence and A (fraction of the population coalesced) are on the right axis. Mean cluster size is on the left axis.

In the SI model, coalescent events do not happen in tail of the epidemic, after all transmissions have occurred. Consequently, the A and MCS curves for $T = 40$ and $T = 60$ coincide.

In the SIS model, the population coalesces even in the tail of the epidemic (at equlibrium), since transmissions are still occurring. The limiting value of x_1 (at $t = 0$) in both cases is 10^4, which is 1 / the fraction initially infected. The equations thereby predict the total population size assuming we infected one person at random.

The limiting value of A (at $t = 0$) is 10^{-4}, which is the fraction initially infected.

Standard coalescent methods based on constant-size populations can be used for SIS dynamics at equilibrium. In a Moran model (overlapping generations in a population of constant size), suppose the expected generation time is $1/\mu$.The expected delay for i lineages to coalesce to $i-1$ is $1/(\mu\binom{i}{2})$. Let T_i be the expected time of the i'th coalescent event among a sample of n lineages. We have

$$T_m = \sum_{i<m} 1/(\mu\binom{i}{2})$$

Suppose S^* and I^* are the fraction susceptible and infected at equilibrium in the SIS model, and N is the size of the entire population. The number of transmissions is proportional to $S^* \times I^* \times N$ and the probability that a transmission corresponds to a coalescent event among n lineages is

$$p_c = \binom{n}{2}/\binom{NI^*}{2}.$$

Then $\mu = (S^* \times I^*/(I^*(I^* \times N - 1)))$.

The quantity A from our model predicts the expected fraction of lineages in the population at any time, and is related to the number of lineages in a coalescent.

$$A \approx (n-m)/N,$$

with sample size n, population size N, and after m coalescent events have occured.

In figure S3, we have compared A with a plot of T_m versus $(n-m)/N$, with $n = 1000$ and $N = 10^5$. These quantities coincide at equilibrium, but not during epidemic growth, when the population size is not constant.

3 Variance and mean of the cluster size distribution

In the main text, we claimed that the variance of the CSD asymptotically approaches the mean squared. Figure S4 demonstrates this by comparing the mean of the CSD to the variance over mean. Parameters are the same as for Figure 2 in the main text.

4 Alternative fitting algorithms

In the paper we proposed a likelihood function for fitting compartmental models to a phylogeny. Alternatively, we can compare the sequence of coalescent times to the predicted distribution of coalescent times.The

Kolmogorov Smirnov test statistic

$$D_n = \sup_t |\frac{1}{n}\sum_{i=1}^{n} I_{t_i \leq t} - F_A(t)| \tag{2}$$

gives the maximum difference between the theoretical distribution F_A and the cumulative empirical distribution of coalescence times. Since F_A is also a function of epidemic parameters θ, this motivates an alternative fitting criterion, which is simply the p-value of the statistic D_n with n degrees of freedom.

5 Simulations: Efficiency of estimation algorithm

The procedure for estimating coverage and bias of our estimation algorithm is as follows:

1. 288 replicates were drawn from the joint-prior distribution for epidemic parameters (table 1 in main text): transmission rate, recovery rate, time population observed, and population size.

2. An SIR model is integrated for each replicate, as well as the variable A in reverse-time. $n = 55$ coalescent times are drawn iid from the distribution of coalescent times with CDF $F_A(t)$.

3. The Bayesian importance sampling algorithm was applied to the sample of coalescent times, which provided posterior estimates (mean of the posterior distribution) of transmission and recovery rates, as well as the final prevalence at t_1. Confidence intervals were also estimated.

4. Estimated values were compared to actual values, and coverage probabilities were calculated by comparing initial replicates from the prior distribution and the estimated confidence intervals.

The simulation prevalence trajectories are shown in figure S5. The results are shown in figure S6.

The estimated coverage probabilities are:

Parameter	Coverage Probability
Recovery Rate	0.84
Transmission Rate	0.89
Final Prevalence	0.92

Our algorithm performs best for the transmission rate, but largely fails to predict the recovery rates.

5.1 Efficiency and recovery rate

A second set of experiments was conducted to see if estimation of recovery rate was more efficient for smaller R_0. 60 epidemics were generated using parameters drawn from the following priors:

- R_0: Uniform(1.75, 2.25)

- γ: Uniform(1.5, 1.75)
- T: Constant = 5
- N: Constant = 10^5

150 lineages were sampled and used to fit an SIR model using the likelihood function based on the KS statistic (equation S2). Coverage values were similar to the 241 case.

- γ: 0.93
- β: 0.98

Figure S7 shows actual versus estimated transmission and recovery rates. As with the 241 data, we see our method accurately estimates transmission rates ($\rho = 0.92$). However performance is poor for estimation of the recovery rate ($\rho = 0.40$).

Although our estimate of γ is innaccurate, it is still robust against mis-specification of the prior distribution. Another set of simulations (figure S8) was conducted with parameters drawn from the same distributions, but our estimation algorithm used a mis-specified prior for γ: Uniform(0, 1.75). The mis-specification of γ throws off estimates of both β and final size, though our method correctly detects the presence of recovery rates greater than zero, and most estimates are near the correct range 1.5-1.75. We have $\rho(\beta, \hat{\beta}) = 0.92$, and $\rho(\gamma, \hat{\gamma}) = 0.34$. These results indicate that our estimation algorithm should at least be able to distinguish between SI and SIR models.

Because of the difficuly of estimating recovery rates, informative priors for these parameters were used for all results presented in the text. Fortunately, information on recovery and the natural history of a disease is usually available for infectious diseases.

6 Comparison with the generalized skyline

The simulations were based on a sample of 50 or 500 sequences at one of two sample times:

1. The time of maximum prevalence
2. The time corresponding to 20% of maximum prevalence after the peak

Transmission and recovery rates were such that $R_0 = 2$. Informative priors were used for the recovery rate and the fraction of the population sampled (see below).RMSE was calculated by averaging the squared deviation of estimated and true prevalence over 100 time points, from 0 to the sample time. When calculating RMSE, we rescaled N_e from the generalized skyline using linear regression which minimizes the squared residuals

with prevalence. The rescaled N_e provides the fairest possible comparison between effective population size and the true prevalence.

The Metropolis-Hastings algorithm was used to fit the SIR model (MCMCpack in R). We began every Markov chain out of equilibrium ($r_0 = 2.5, \mu = 0.5$), so as not to give the deterministic SIR dynamics an unfair advantage over the skyline. To summarize, these experiments were conducted by the following steps:

1. simulate an SIR epidemic, take a standard random sample of agents at time T, and reconstruct the genealogy of transmissions,
2. fit a generalized skyline model to the simulated genealogy,
3. fit an SIR model to the genealogy,
4. determine the goodness of fit of the skyline and SIR models to the actual epidemic prevalence over time.

- transmission rate = 2
- recovery rate = 1
- N = 10^4
- one initial infected.

Fitting the SIR model was conducted using Metropolis-Hastings implemented in MCMCpack in R. The simulations had the following parameters: The Markov chain was started out of equilibrium (transmission rate = 2.5, recovery rate = .5). The Markov chain was iterated for 10000 steps, recording every fifth interval and allowing a 5000 step burn-in. We used the following priors:

- Transmission rate $\sim$ Uniform(0-10)
- Recovery rate $\sim$ Normal(1, .5)
- Fraction initially infected $\sim$ Uniform($.25 \times 10^{-4}$, 100×10^{-4})
- Fraction of the population sampled $\sim$ Normal($n \times 10^{-4}$, $(n/2) \times 10^{-4}$)

The generalized skyline was computed using the *mcmc.popsize* function in the ape package of R. The mcmc.popsize function also uses MCMC, and the Markov chain was iterated for 10000 steps with a 200 step burn-in.

Figure S10 shows the actual estimated prevalence from the skyline and SIR models. These trajectories were picked randomly from the set of 300 simulations with $n = 50$ and the a sample time at 20% of maximum prevalence.

It is usually the case that the generalized skyline fails to detect a decrease in prevalence and over-estimates in the latter stages of the epidemic.

7 Time to most recent common ancestor

The point where $A = 1/N$ represent the point where the genealogy of virus has collapsed to a single lineage– the most recent common ancestor of the sample. Therefore, if we collect a sample of size n at time T, and solve

$$\dot{A} = -f_{SI}(A/I)^2$$

to time zero, with $A(T) = n/N$, the time τ which satisfies $A(\tau) = 1/N$ corresponds to the median time to the most recent common ancestor of the sample.

A demonstration is illustrated in figure S11. The sample time $T = 16.22$ corresponds to 10% of peak prevalence. The simulation parameters are

- $N = 5 \times 10^4$
- $I(0) = 1$
- Transmission rate = 2
- Recovery rate = 1
- Sample size = 50

One thousand simualations were conducted, generating one thousand unique values of TMRCA. The empirical distribution of these values is illustrated in figure S11. The median TMRCA is illustrated with black dots and the time τ is shown as a red line.

The theory was also validated using simulations corresponding to samples at 100% and 50% of peak prevalence.

8 Model for HIV phylodynamics

In the text, we fit the following model to 55 HIV sequences:

$$\dot{S} = -S^{\alpha}(\beta_1 I_1 - \beta_2 I_2) + \mu - \mu S \tag{3}$$

$$\dot{I}_1 = S^{\alpha}(\beta_1 I_1 + \beta_2 I_2) - \gamma_1 I_1 - \mu I_1 \tag{4}$$

$$\dot{I}_2 = \gamma_1 I_1 - \gamma_2 I_2 - \mu I_2. \tag{5}$$

Note that this model implies that the reproduction number, R_0, will be the expected number of transmissions in the acute stage, plus the expected number of transmissions in the chronic stage, provided that the population is susceptible except for a single infected ($S \approx 1$). This is

$$R_0 = \frac{\beta_1}{\gamma_1 + \mu} + \frac{\beta_2}{\gamma_2 + \mu}.$$

This model requires that we compartmentalize the ancestor function by the status (acute or chronic infected) of the ancestor. A_1 denotes the fraction of the population that is acute infected and which has progeny extant at time T. A_2 is the fraction of the population that is chronic infected and which has progeny extant at time T. We now derive the following equations:

$$\bar{\dot{A}}_2 = -\gamma_1 I_1 (A_2/I_2) + \beta_2 I_2 S^{\alpha}(A_1/I_1)((I_2 - A_2)/I_2) \tag{6}$$

$$\bar{\dot{A}}_1 = \gamma_1 I_1 (A_2/I_2) - \beta_1 I_1 S^{\alpha}(A_1/I_1)^2 - \beta_2 I_2 S^{\alpha}(A_1/I_1). \tag{7}$$

- In forward time, Acute infecteds move to Chronic state at rate γ_1. In reverse time, this flow is reversed.
  - A number of chronics proportional to $\gamma_1 A_1$ move to the Acute state.
  - With probability A_2/I_2 the chronic is an ancestral lineage.
  - Consequently, A_1 increases at a partial rate $\gamma_1 I_1 (A_2/I_2)$, and A_2 decreases by the same partial rate.
- A_1 decreases at a partial rate $\beta_1 I_1 S^{\alpha}(A_1/I_1)^2$ which has identical rationale as for A in the standard SIR model.
- Chronic infecteds transmit to susceptibles at the rate $\beta_2 I_2 S^{\alpha}$.
  - With probability A_1/I_1, the new infected is an ancestral lineage.

  - If the transmitting Chronic is an ancestral lineage (with probability (A_2/I_2)), the lineage represented by the Acute is coalesced into the Chronic.
  - If the transmitting Chronic is not an ancestral lineage (with probability $((I_2 - A_2)/I_2)$), the lineage moves to the Chronic state A_2.
- Consequently, A_1 decreases at a partial rate $\beta_2 I_2 S^\alpha (A_1/I_1)$. And, A_2 increases at a partial rate $\beta_2 I_2 S^\alpha (A_1/I_1)((I_2 - A_2)/I_2)$.

Adding the partial rates yeilds equations S6.

The priors used for fitting ACTG are

- $\alpha \sim$ Uniform(1, 30)
- $\beta_1 \sim$ 1/Uniform(35,100)
- $\beta_2 \sim$ 1/Uniform(350,1500)
- $\epsilon \sim$ Uniform(1,20)/N

9 Sample and threshold times for simulations

In Figure 2 of the main text, four trajectories of the cluster size moments were generated for four sample times T. And for each trajectory, simulated moments were calculated for ten threshold times t. The exact values used are as follows:

T	t	t	t	t	t	t	t	t	t	t
7.96	5.27	0.96	2.68	0.1	6.13	6.99	3.54	7.86	4.41	1.82
9.77	4.35	0.1	9.67	5.42	3.29	1.16	8.61	6.48	2.22	7.54
11.22	4.99	0.1	6.22	1.32	8.67	3.77	9.89	7.44	11.12	2.54
16.06	3.62	5.38	0.1	7.15	1.86	8.91	10.67	12.44	14.20	15.96