Genetics: Published Articles Ahead of Print, published on September 21, 2009 as 10.1534/genetics.109.106021

Phylodynamics of infectious disease epidemics

Erik M. Volz[1,2], Sergei L. Kosakovsky Pond[3], Melissa J. Ward[4],

Andrew J. Leigh Brown[4], Simon D.W. Frost[5]

September 6, 2009

[1] Department of Epidemiology, University of Michigan, Ann Arbor, United States [2] Department of Pathology, University of California San Diego, United States [3] Department of Medicine, University of California, San Diego, United States [4] School of Biological Sciences,University of Edinburgh, United Kingdom [5] Department of Veterinary Medicine, University of Cambridge, United Kingdom

Running Head: Phylodynamics

Keywords: Coalescent, SIR, HIV, Infectious Disease, Mathematical Epidemiology

Corresponding author:

Erik Volz

Department of Epidemiology

University of Michigan– Ann Arbor

M5073, SPH II

1415 Washington Heights

Ann Arbor, MI, 48109-2029

USA

Phone: 1 619 757 4846

Fax: 1 734 936 2084

Email: erik@erikvolz.info

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 SI, SIS, and 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 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; DRUMMOND *et al.*, 2005; ROSENBERG and NORDBORG, 2002) and estimation of epidemic prevalence (YUSIM *et al.*, 2001; WILSON *et al.*, 2005; ROBBINS *et al.*, 2003), yet there have been few attempts to formally integrate coalescent theory with standard epidemiological models (PYBUS *et al.*, 2001; GOODREAU, 2006). Whilst epidemiological models such as SIR consider the dynamics of an entire population going forwards in time, the coalescent theory operates on a small sample of an infected sub-population and models the merging of lineages backwards 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, 1982b,a). Numerous extensions have been made for populations with overlapping generations in continuous time, exponential or logistic growth (GRIFFITHS and TAVARE, 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 super-infection 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 HIV and HCV which have a high mutation rate relative to dispersal. Underlying SIR/SI/SIS 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). 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 formalise a link with standard epidemiological models which 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 (GOODREAU, 2006; LEWIS *et al.*, 2008; DRUMRIGHT and FROST, 2008; YERLY *et al.*, 2001; HUE *et al.*, 2005; PAO *et al.*, 2005; BRENNER *et al.*, 2007). 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.

1 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 (KERMACK and MCKENDRICK, 1927; BAILEY, 1975; ANDERSON and MAY, 1991)

$$\dot{S} = -\beta SI, \dot{I} = \beta SI - \gamma I, \dot{R} = \gamma I. \tag{1}$$

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 will 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 backwards in time towards time $t = 0$. Whilst this differs from the conventional temporal notation for coalescent theory (where the sampling, or present, time is denoted 0, and as we move backwards t denotes the number of years before the present), it allows us to develop a system of equations which 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:

- 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 4 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).

1.1 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 backwards 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 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

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

We will 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 will 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

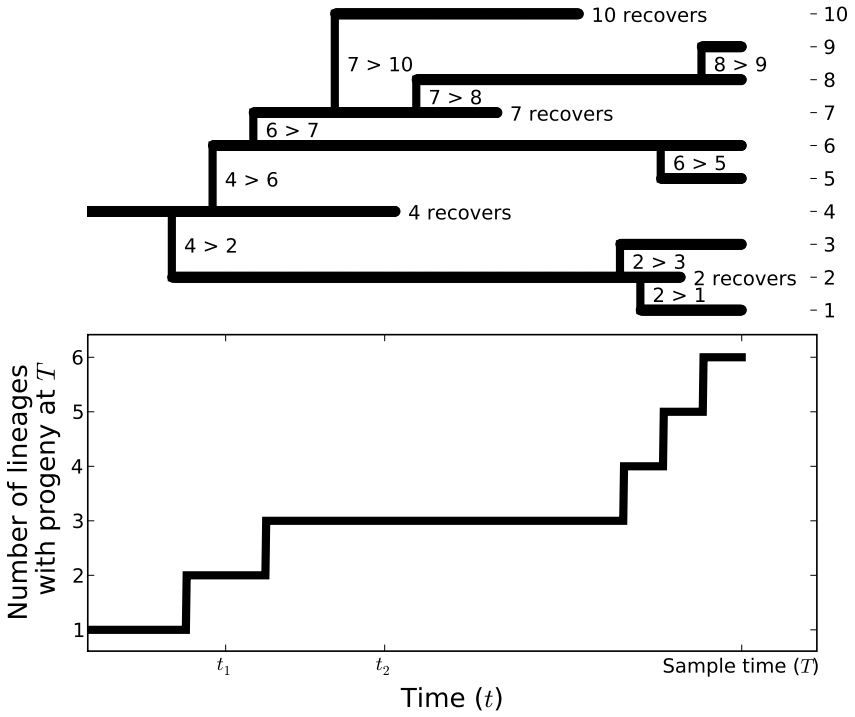

Figure 1: An example of a phylogeny that could be generated by an epidemic process. The number of lineages at time t for a population observed at time T is plotted below. A branch in the tree corresponds to a transmission event, and as a lineage is traced backwards in time, it traverses multiple infected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

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 backwards ordinary differential equation from time T:

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

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.

1.2 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 $x_2 = \mathrm{E}(\mathbf{X}_2)$) will be a random variable which 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

$$-\dot{x}_1(t;T) = f_{SI}(t)I(T)/I(t)^2, \tag{3}$$

$$-\dot{x}_2 = 2\dot{x}_1 \tag{4}$$

backwards 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 $\mathcal{I}(T)$ will be 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 backwards derivative at t yields

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

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

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

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

$$E(\mathbf{X}_1^2) = \sum_i i^2 \text{Pr}\{\mathbf{X}_1 = i\} = \left(\sum_i i \text{Pr}\{\mathbf{X}_1 = i\} \right) \left(\frac{\sum_i i^2 \text{Pr}\{\mathbf{X}_1 = i\}}{\sum_i i \text{Pr}\{\mathbf{X}_1 = i\}} \right) \tag{8}$$

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

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

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

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

Then, the variance in cluster size is

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

Higher moments can also be derived recursively from earlier moments . We now show that the n'th moment of the CSD, M_n, can be found by solving the following differential equation with initial conditions $M_n(T) = 1$:

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

where we define $M_0 := 1$ for convenience. Equations 3 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 n'th moment when two clusters merge. When a cluster of size i merges with a cluster of size j, the n'th 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 n'th moment when two clusters merge, we sum over all possible combinations of clusters of sizes i and j.

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

The product of the coalescent rate $f_{SI}A^2/I^2$ and the factor $1/A$ which accounts for the probability that a focal lineage takes part in a coalescent event, along with the expected size function yields equation 10. In the *Supporting Information*(Figure S1) ,we compare solutions of this equation to the 2nd through 5th moments from simulations. .

1.3 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 which 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

$$-\dot{\overline{A}}/\left(A(T;T)-A(t;T)\right).$$

We will 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 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:

$$\Lambda(t_1,\cdots,t_n-1|\theta)=\sum_{i=1}^{n-1}\log(-\dot{\overline{A}}(t_i)/\left(A(T)-A(t)\right)) \tag{12}$$

$$=-(n-1)\log(A(T;T)-A(t;T))+\sum_{i=1}^{n-1}\log(-\dot{\overline{A}}(t_i;T)), \tag{13}$$

where θ denotes the parameters of the SIR model, such as transmission and recovery rates. In the *Supporting Information*we also present a likelihood function based on the Kolmogorov-Smirnov statistic for comparing distributions.

2 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. 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 most recent common ancestor (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 which 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 the *Supporting Information*(Figure S11).

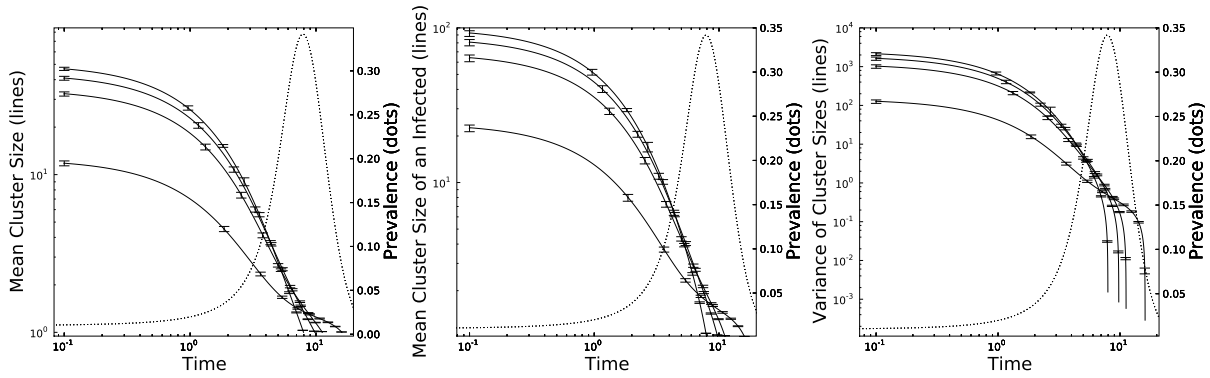

Figure 2: The moments of the cluster size distribution over time as calculated by equations 3 and 9 (lines,log-scale). Four trajectories of the cluster size moments were generated for four sample times T. And for each trajectory, simulated moments were calculated for ten threshold times t. Error bars 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 right axis. Transmission rate $\beta = 1$, and recovery rate $\mu = 0.3$.

2.1 Simulations

In order to assess the peformance 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 SDF and EMV in Python and C. Results from both models were compared to insure 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 2.1 shows equations 3 and 9 (lines) and the 90% confidence intervals from simulations at ten thresholds (t values). The exact values of t and T are reported in the *Supporting Information*. 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 backwards 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 complex function of both t and T, and could therefore not be reduced to a simple forward-looking model.

2.2 Comparison with the generalized skyline

Further simulations were developed to test the suitability of the 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 ten thousand agents.

The simulations were also designed to determine if SIR models that are fit via the 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 the *Supporting Information*.

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 out-performs 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 right, but does not fundamentally change the shape of the distribution. This is easily accounted for by including a parameter which 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 the *Supporting Information*(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 over-estimates 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)).

2.3 The effect of 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, Kingman showed that the probability that n sampled sequences will not have a common ancestor in the preceding generation is

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

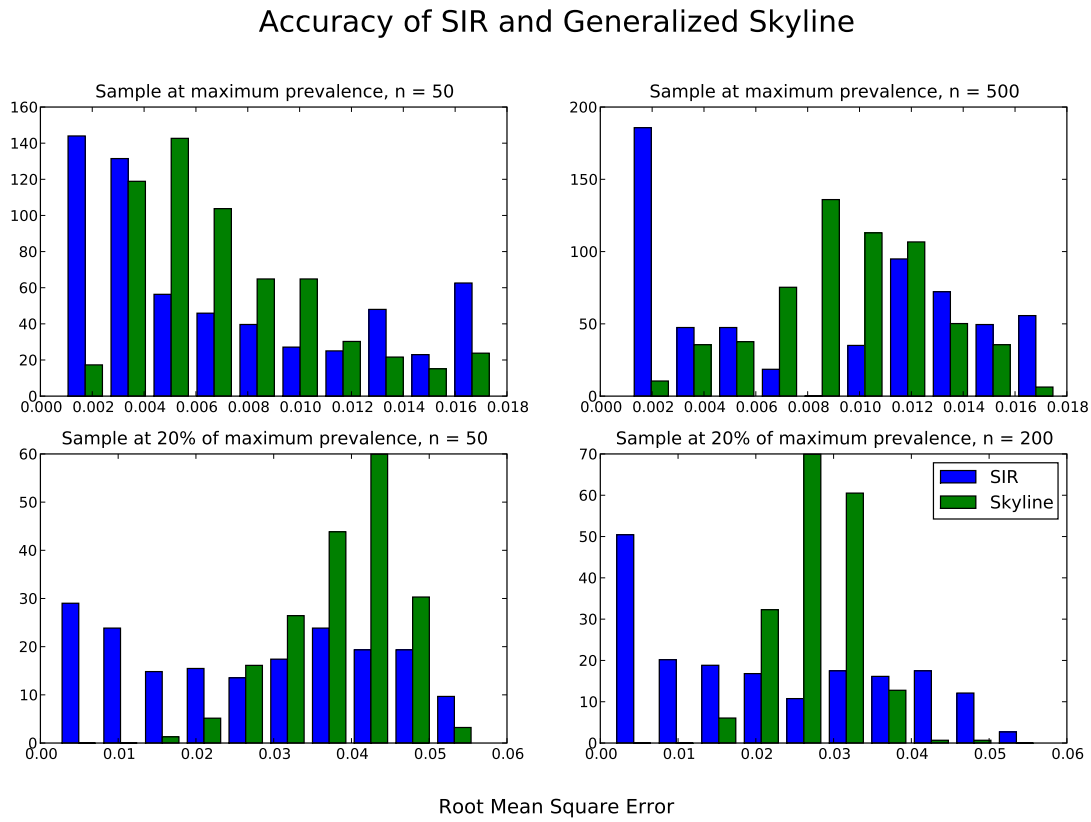

Figure 3: Root mean square error of SIR and generalized skyline estimates of epidemic prevalence. Data are based on three hundred simulated epidemics ($R_0 = 2$). RMSE is averaged over one hundred time points.

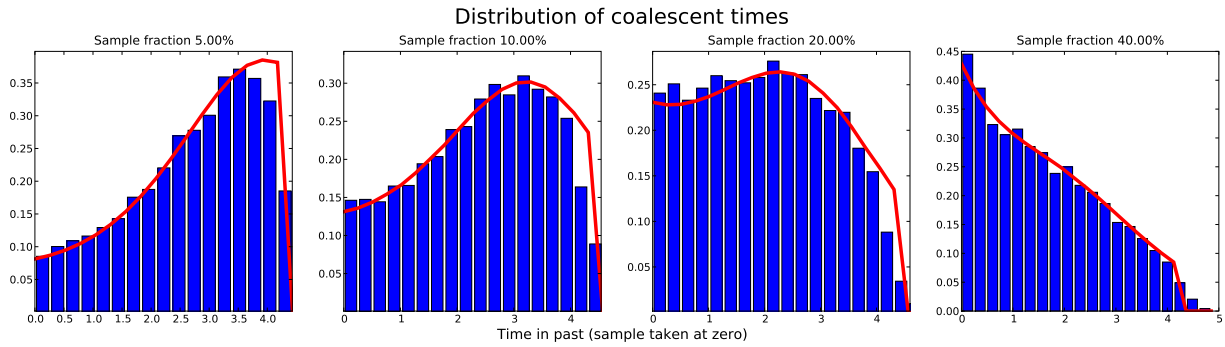

Figure 4: The empirical distribution of coalescence times based on 150 simulated SIR epidemics. Transmission rate = 2, recovery rate = 1. The expected distribution based on equation 11 is shown in red.

Kingman 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 more than 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. 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.

2.4 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 - 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 clock-like 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 one to three 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:

$$\dot{S} = -S^{\alpha}(\beta_1 I_1 - \beta_2 I_2) + \mu - \mu S \tag{14}$$

$$\dot{I}_1 = S^{\alpha}(\beta_1 I_1 + \beta_2 I_2) - \gamma_1 I_1 - \mu I_1 \tag{15}$$

$$\dot{I}_2 = \gamma_1 I_1 - \gamma_2 I_2 \tag{16}$$

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

- β_1: The transmission rate of acute infecteds.
- β_2: The transmission rate of chronic infecteds.
- μ: The immigration rate into the susceptible population. We consider the total population to have constant size.
- α: A parameter which 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 which lasts two months on average ($\gamma_1 = 1/60$), and a chronic phase that lasts ten years on average ($\gamma_2 = 1/(10 \times 265)$).

Prior distributions are given in the *Supporting Information*.

Given $n = 55$ sequences, we use the initial conditions $A(T) = 55/N$, $I_1(0) = \epsilon$, and $S(0) = 1 - \epsilon$. 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 will be the fractions of the population which are respectively acute and chronic infected and which has sampled progeny at time T. We have:

$$\dot{\bar{A}_2} = -\gamma_1 I_1(A_2/I_2) + \beta_2 I_2 S^\alpha (A_1/I_1)((I_2 - A_2)/I_2) \tag{17}$$

$$\dot{\bar{A}_1} = \gamma_1 I_1(A_2/I_2) - \beta_1 I_1 S^\alpha (A_1/I_1)^2 - \beta_2 I_2 S^\alpha (A_1/I_1) \tag{18}$$

For purposes of fitting the SIR model, we use $A = A_1 + A_2$ and $\dot{\bar{A}} = \dot{\bar{A}_1} + \dot{\bar{A}_2}$. A derivation is provided in the *Supporting Information*.

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

- 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
- Incidence scaling parameter, $\hat{\alpha} = 9.77$

Together, these parameters imply an R_0 value of 2.24 (see *Supporting Information*). 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 re-scaled so that $\min(N_e) = \min(I)$. Figure 5 shows the re-scaled skyline and an SIR trajectory which 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 early nineties. The skyline also reproduces the growth rate during the expansion phase and the tapering of epidemic growth in the early nineties. However, the skyline predicts a rise in N_e between 1980 and 1993, which probably over-estimates the true prevalence.

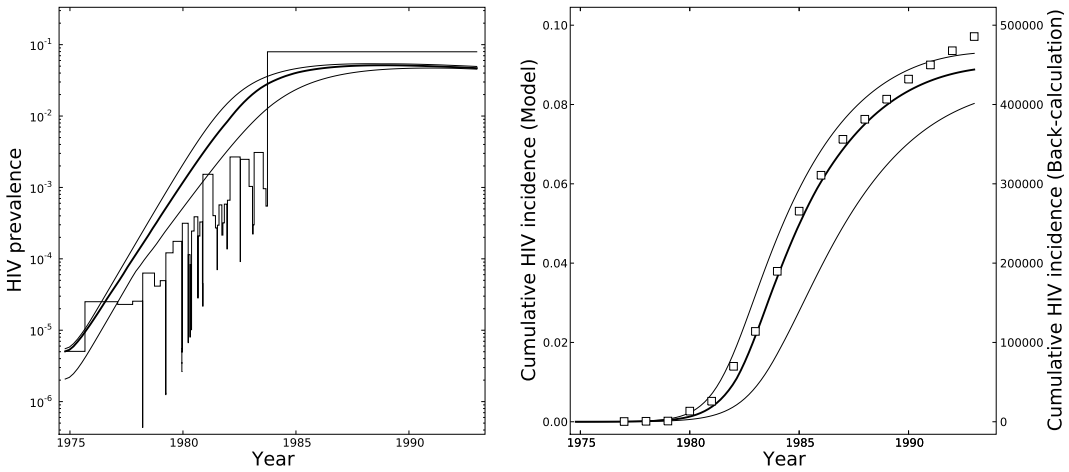

Figure 5: Left: Estimated epidemic prevalence (logarithmic scale) of HIV among MSM in the United States. A solution to equation 17 is compared to the skyline plot, re-scaled such that minimum effective population size equals minimum prevalence. The thin lines show 95% confidence intervals. Right: Estimated cumulative incidence of HIV among MSM versus time (years prior to 1993). A solution to equation 17 is compared to estimates based on sero-surveillance data (HALL *et al.*, 2008).

We have also compared the CSD mean and variance from our best-fit SIR model to the empirical values from the ACTG 241 data (figure 6). The SIR model successfully reproduces the mean cluster size throughout the course of the epidemic. However, there is substantial deviation between the actual and 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).

3 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.

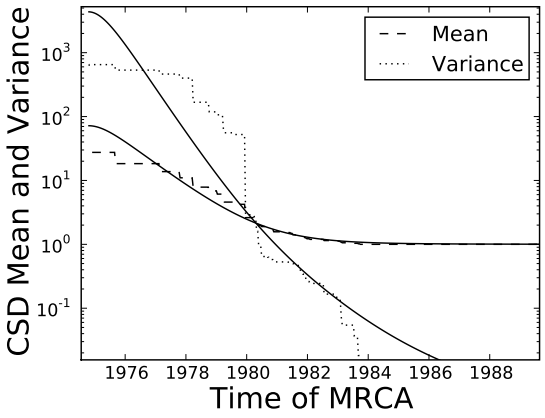

Figure 6: The mean cluster size (dashes) and variance of cluster sizes (dotted line) are calculated from the empirical observations from the ACTG 241 sequences (dashed lines) and compared to our best-fit SIR model(solid lines). The horizontal axis gives the clustering threshold as the year of the MRCA of a cluster.

The CSD is sensitive to details of the underlying population dynamics. Most coalescent approaches only take into account 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 SI models (no recovery) and 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 only coalesce during exponential growth, not at equilibrium (Figures S2 and 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 mis-specification of priors that include zero recovery, and it is feasible to distinguish SI from SIR dynamics(Figures S6-S9).

4 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

which provides a step towards the formal integration of phylodynamics and epidemiology and which 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 both successful reproduction 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 super-infection 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.

Acknowledgements: The authors acknowledge support from the NIH (T32 AI07384,R01 AI47745). SDWF is supported by a Royal Society Wolfson Research Merit Award. MJW is supported by the Biotechnology and Biological Sciences Research Council. Irene Hall provided estimates of HIV incidence in MSM.

References

ANDERSON, R. M. and R. M. MAY, 1991 *Infectious Diseases of Humans: Dynamics and Control*. Oxford University

Press, USA.

ATHREYA, K. B. and P. E. NEY, 2004 *Branching Processes*. Dover Publications.

BAILEY, N. T. J., 1975 *The Mathematical Theory of Infectious Diseases and its Applications*. London.

BRENNER, B. G., M. ROGER, J. ROUTY, D. MOISI, M. NTEMGWA, C. MATTE, J. BARIL, R. THOMAS, D. ROULEAU, J. BRUNEAU, *et al.*, 2007 High rates of forward transmission events after acute/early HIV-1 infection. Journal of Infectious Diseases **195**: 951.

D'AQUILA, R. T., M. D. HUGHES, V. A. JOHNSON, M. A. FISCHL, J. P. SOMMADOSSI, S. LIOU, J. TIMPONE, M. MYERS, N. BASGOZ, and M. NIU, 1996 Nevirapine, zidovudine, and didanosine compared with zidovudine and didanosine in patients with HIV-1 infection: A randomized, double-blind, placebo-controlled trial. Annals of Internal Medicine **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. Molecular Biology and Evolution **22**: 1185–1192.

DRUMRIGHT, L. N. and S. D. W. FROST, 2008 Sexual networks and the transmission of drug-resistant HIV. Current Opinion in Infectious Diseases **21**: 644.

FU, Y., 2006 Exact coalescent for the Wright–Fisher model. Theoretical Population Biology **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, J. A. MUMFORD, and E. C. HOLMES, 2004 Unifying the epidemiological and evolutionary eynamics of pathogens. Science **303**: 327.

GRIFFITHS, R. C. and S. TAVARE, 1994 Sampling theory for neutral alleles in a varying environment. Philosophical Transactions of the Royal Society B: Biological Sciences **344**: 403–410.

HALL, H., R. SONG, P. RHODES, J. PREJEAN, Q. AN, L. LEE, J. KARON, R. BROOKMEYER, E. KAPLAN, M. MCKENNA, *et al.*, 2008 Estimation of HIV incidence in the United States. JAMA **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. Proceedings of the National Academy of Sciences **102**: 4425–4429.

KAJ, I. and S. M. KRONE, 2003 The coalescent process in a population with stochastically varying size. Journal of Applied Probability **40**: 33–48.

KERMACK, W. O. and A. G. MCKENDRICK, 1927 A contribution to the mathematical theory of epidemics. Proceedings of the Royal Society of London. Series A, Containing Papers of a Mathematical and Physical Character (1905-1934) **115**: 700–721.

KINGMAN, J. F. C., 1982a On the genealogy of large populations. Journal of Applied Probability **19**: 27–43.

KINGMAN, J. F. C., 1982b The coalescent. Stochastic Process. Appl. **13**: 235–248.

LAVERY, S., C. MORITZ, and D. R. FIELDER, 1996 Genetic patterns suggest exponential population growth in a declining species. Molecular Biology and Evolution **13**: 1106–1113.

LEIGH BROWN, A. J., H. F. GCUNTHARD, J. K. WONG, R. T. D'AQUILA, V. A. JOHNSON, D. R. KURITZKES, and D. D. RICHMAN, 1999 Sequence clusters in human immunodeficiency virus type 1 reverse transcriptase are associated with subsequent virological response to antiretroviral therapy. The Journal of Infectious Diseases **180**: 1043–1049.

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 HARVEY, P. H., 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 Evolutionary Biology **5**: 6.

PAO, D., M. FISHER, S. HUÉ, G. DEAN, G. MURPHY, P. A. CANE, C. A. SABIN, and D. PILLAY, 2005 Transmission of HIV-1 during primary infection: relationship to sexual risk and sexually transmitted infections. AIDS **19**: 85.

PYBUS, O. G., M. A. CHARLESTON, S. GUPTA, A. RAMBAUT, E. C. HOLMES, and P. H. HARVEY, 2001 The epidemic behavior of the hepatitis C virus. Science **292**: 2323–2325.

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.

ROBBINS, K. E., P. LEMEY, O. G. PYBUS, H. W. JAFFE, A. S. YOUNGPAIROJ, T. M. BROWN, M. SALEMI, A. M. VANDAMME, and M. L. KALISH, 2003 US human immunodeficiency virus type 1 epidemic: Date of origin, population history, and characterization of early strains. Journal of Virology **77**: 6359–6366.

ROSENBERG, N. A. and M. NORDBORG, 2002 Genealogical trees, coalescent theory and the analysis of genetic polymorphisms. Nature Reviews Genetics **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.

WILSON, D. J., D. FALUSH, and G. MCVEAN, 2005 Germs, genomes and genealogies. Trends in Ecology & Evolution **20**: 39–45.

YERLY, S., S. VORA, P. RIZZARDI, J. P. CHAVE, P. L. VERNAZZA, M. FLEPP, A. TELENTI, M. BATTEGAY, A. L. VEUTHEY, J. P. BRU, *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. Philosophical Transactions of the Royal Society B: Biological Sciences **356**: 855–866.