1 Higher moments and simulations

Figure S1 shows four solutions to
\[
M_n = f_{SI} \frac{A^{n-1}}{T^2} \sum_{i=0}^{n-1} \binom{n}{i} M_i M_{n-i}.
\] (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 equilibrium), 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 $1/(\mu(i))$. Let $T_i$ be the expected time of the $i$'th coalescent event among a sample of $n$ lineages. We have

$$T_m = \sum_{i<m} 1/(\mu\left(\begin{array}{c}i \\ 2\end{array}\right))$$

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

$$p_c = \left(\begin{array}{c}n \\ 2\end{array}\right) / \left(\begin{array}{c}NI^* \\ 2\end{array}\right).$$

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 occurred.

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

\[ D_n = \sup_t \left| \frac{1}{n} \sum_{i=1}^{n} I_{t_i \leq t} - F_A(t) \right| \]

(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 \(\theta\), 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:

<table>
<thead>
<tr>
<th>Parameter</th>
<th>Coverage Probability</th>
</tr>
</thead>
<tbody>
<tr>
<td>Recovery Rate</td>
<td>0.84</td>
</tr>
<tr>
<td>Transmission Rate</td>
<td>0.89</td>
</tr>
<tr>
<td>Final Prevalence</td>
<td>0.92</td>
</tr>
</tbody>
</table>

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)
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.

- $\gamma$: Uniform(1.5, 1.75)
- $T$: Constant = 5
- $N$: Constant = $10^5$

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 $\gamma$ is inaccurate, 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 $\gamma$: Uniform(0, 1.75). The mis-specification of $\gamma$ throws off estimates of both $\beta$ 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 difficulty 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($0.25 \times 10^{-4}$, $100 \times 10^{-4}$)
- 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

\[
\bar{A} = -f_{SI}(A/I)^2
\]

to time zero, with \( A(T) = n/N \), the time \( \tau \) 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 simulations 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 \( \tau \) 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:

\[ \dot{S} = -S^\alpha(\beta_1 I_1 - \beta_2 I_2) + \mu - \mu S \]
\[ \dot{I}_1 = S^\alpha(\beta_1 I_1 + \beta_2 I_2) - \gamma_1 I_1 - \mu I_1 \]
\[ \dot{I}_2 = \gamma_1 I_1 - \gamma_2 I_2 - \mu I_2. \]

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:

\[ \ddot{A}_2 = -\gamma_1 I_1(A_2/I_2) + \beta_2 I_2 S^\alpha(A_1/I_1)((I_2 - A_2)/I_2) \]
\[ \ddot{A}_1 = \gamma_1 I_1(A_2/I_2) - \beta_1 I_1 S^\alpha(A_1/I_1)^2 - \beta_2 I_2 S^\alpha(A_1/I_1). \]

- In forward time, Acute infecteds move to Chronic state at rate \( \gamma_1 \). 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 yields equations S6.

The priors used for fitting ACTG are

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

### 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:

<table>
<thead>
<tr>
<th>(T)</th>
<th>(t)</th>
<th>(t)</th>
<th>(t)</th>
<th>(t)</th>
<th>(t)</th>
<th>(t)</th>
<th>(t)</th>
<th>(t)</th>
<th>(t)</th>
</tr>
</thead>
<tbody>
<tr>
<td>7.96</td>
<td>5.27</td>
<td>0.96</td>
<td>2.68</td>
<td>0.1</td>
<td>6.13</td>
<td>6.99</td>
<td>3.54</td>
<td>7.86</td>
<td>4.41</td>
</tr>
<tr>
<td>9.77</td>
<td>4.35</td>
<td>0.1</td>
<td>9.67</td>
<td>5.42</td>
<td>3.29</td>
<td>1.16</td>
<td>8.61</td>
<td>6.48</td>
<td>2.22</td>
</tr>
<tr>
<td>11.22</td>
<td>4.99</td>
<td>0.1</td>
<td>6.22</td>
<td>1.32</td>
<td>8.67</td>
<td>3.77</td>
<td>9.89</td>
<td>7.44</td>
<td>11.12</td>
</tr>
<tr>
<td>16.06</td>
<td>3.62</td>
<td>5.38</td>
<td>0.1</td>
<td>7.15</td>
<td>1.86</td>
<td>8.91</td>
<td>10.67</td>
<td>12.44</td>
<td>14.20</td>
</tr>
</tbody>
</table>