Supporting material for "Inferring epidemiological dynamics with Bayesian coalescent inference: The merits of deterministic and stochastic models"

File S1

1 Sampling from the prior

In order to assess the correctness of our implementation of the deterministic coalescent SIR and stochastic coalescent SIR models, for each model we used the MCMC algorithm to sample trees from the corresponding distribution $f(\mathcal{T}|\eta)$, and compared these samples with coalescent trees simulated directly under the model.

The chosen η included $\beta = 7.5 \times 10^{-4}$, $\gamma = 0.3$, $S_0 = 999$ and $z_0 = 30$. The comparisons were performed for trees generated from 20 leaves, sampled at integer times 0 through 19, inclusive.

For the deterministic coalescent SIR model, the direct simulation involved numerically solving the Eqs. (1)–(3) in the main text for $t \in [0, 30]$ and using this solution in combination with Eq. (10) in the main text to determine the instantaneous coalescent rate $\lambda(\tau)$. This rate was used to simulate each of the coalescent trees in the usual fashion for heterochronous leaf times. In the case that the MRCA was not reached before the origin time of the epidemic, the tree was discarded and the simulation repeated.

The direct simulation proceeded in a similar way for the stochastic coalescent SIR model, the major difference being that the stochasticity of this model required each coalescent tree to be simulated under a distinct realization of the stochastic trajectory.

Comparisons between the direct simulation and MCMC results are shown in Figures S1 and S2 for three different summary statistics and show very close agreement.

2 Validation through simulated data analysis

As part of the validation of our implementation of the two coalescent SIR models, trees were simulated by their own methods (using stochastically- and deterministically-generated SIR trajectories, as discussed in the Methods section of the main paper), and relevant epidemiological parameters were inferred using the stochastic and deterministic coalescent SIR models. Tables 1 and 2 show the results of these analyses, indicative of correct implementations.

Analyses for varying R_0 (and necessarily, slightly varied other parameters, such as the birth rate β) are provided in Tables S3 and S4. Results from tests of the influence of broader priors (with larger standard deviations in log space) are shown in Table S4. It appears that allowance of broader priors reduces 95% HPD coverage in some cases (e.g., for parameter R_0) when using the deterministic coalescent SIR inference model, as they increase error and bias.

Finally, it was noticed that even for the higher true parameter values of $R_0 = 2.50$ and $S_0 = 999$, under which deterministic coalescent SIR is expected to perform relatively well, there was an inability to accurately estimate the origin parameter z_0. Figure S3 provides some insight into this conundrum by examining the trajectories used for tree simulation and subsequent analysis.

2.1 H1N1 data selection

Initially, the H1N1 dataset contained 45 sequences. The ages of the inferred trees (Figure S4) using the original 45 sequences extended more than 1.5 years into the past for each of the SIR models, which is contrary to what we expect for a single, current strain of seasonal influenza. Three taxa (labelled 32197, 31893, and 31988) were hypothesized to belong to a unique strain, e.g., an additional seeding from outside the Canterbury region or a low-lying previous strain. Removing these three taxa caused the inferred trees to behave as expected, i.e., tree heights and epidemic origin z_0 less than a year old. It also raised the estimated R_0 values for all three SIR models (initially 1.24, 1.10, and 1.55 for stochastic coalescent SIR, deterministic coalescent SIR, and BDSIR, respectively), as well as those for γ (initially 8.74, 12.65, and 11.33 for stochastic coalescent SIR, deterministic coalescent SIR, and BDSIR, respectively).

It will be interesting to further investigate the interplay between influenza strains and its contribution to the overall dynamics. For the closed SIR models discussed in this manuscript, however, this additional complexity leads to increased chance of model misspecification and misleading results. Therefore, we focused our attention on the analyses using 42 sequences.

2.2 HIV-1 data analysis

The original HIV-1 dataset (Hué *et al.* 2005) was agglomerated from both acute and chronic infections sampled in the United Kingdom (UK) and constitutes six phylogenetic clusters, from which the five used here (Clusters 1-4 and 6) were drawn. These particular clusters, with the omission of Cluster 5, were chosen simply for the purpose of direct comparison with Kühnert *et al.* (2014). Our extension to the models allowed us to imprint respective tip dates on the sequence data, sampled from 1999 to 2003, for inclusion in the likelihood computation.

For the selected five clusters, the nucleotide alignments contained 41, 62, 29, 26, and 35 sequences, respectively, each with 952 sites. The substitution scheme chosen for phylogenetic analysis was the symmetric and independent general time reversible model (GTR), with gamma distributed rate variation and explicit proportion of invariable sites (GTR+G+I). Following Hué *et al.* (2005), the substitution rate was set to 2.55E-4 substitutions per site per year. All other parameters were estimated conjointly, and the Bayesian prior distributions are presented in Table 4: Bayesian prior distributions.

The pathophysiology of HIV is multifarious, and the patterns of its advancement within an infected host change throughout time. In addition to increased complexity potentially caused by recombination events, the transition between HIV's acute and chronic phases alters the host's infectivity (Guss 1994). The SIR compartmental model used for this particular phylodynamic analysis on the UK cluster data does not allow for independent infection rates for the acute and chronic phases (but see Volz *et al.* (2012) and Volz *et al.* (2013)). However, in this study we did not attempt to estimate the infection rate β and thus did not expect such a difference to significantly impact the estimation of the parameters of interest: the basic reproductive number R_0, removal rate γ, size of the initial susceptible population S_0, and origin of the outbreak z_0.

2.2.1 HIV-1 inference results

In regard to parameter inference from the serially-sampled HIV-1 sequence data, the stochastic coalescent SIR, deterministic coalescent SIR, and BDSIR methods were most alike in light of the R_0 results. The medians and HPD intervals for all clusters pertaining to this parameter, (especially Clusters 1, 2, 3, and 6), were very close, and those of Cluster 4 were still congruent across the three analyses (Figure S5).

The coalescent SIR models and BDSIR disagreed with respect to the age of the most recent common ancestor and the origin z_0 (Figure S6). The coalescent SIR models also exhibited much larger 95% HPD intervals for z_0 in each of the clusters; while BDSIR encompassed an average of 16 years, the stochastic coalescent SIR and deterministic coalescent SIR models had averages of 49 and 37 years, respectively. Furthermore, the estimated age of the common ancestor of the tree was older under the coalescent SIR models than the estimates reported by either BDSIR or the original data analysis (Hué *et al.* 2005) for each cluster. This was also true for the time of origin for the epidemic, although for certain clusters the differences between the coalescent estimates of the origin z_0 and the birth-death estimates were much greater than others (e.g., Cluster 3).

The estimates of removal rate γ from Clusters 1 and 6 were very similar across the three methods (Figure S7). However, both coalescent SIR models estimated considerably higher γ values for Clusters 2-4 than BDSIR. This is reflective of the simulation study results, where the two coalescent models did not perform as well as BDSIR for the removal parameter.

Median estimates for the initial susceptible population S_0 were quite similar in all methods for Clusters 1-4, although BDSIR displayed much wider HPD intervals than stochastic coalescent SIR and deterministic coalescent SIR (Figure S8). In Cluster 6, the coalescent SIR models showed the smallest HPD intervals for their individual analyses on each cluster, while the opposite was true for BDSIR. There was also a disparity between the median estimates for the two coalescent approaches and that of BDSIR for Cluster 6. To this effect, it should be noted that the number of infections accrued throughout the duration of the epidemic was reported as $N_e = 1,350$ by Hué *et al.* This casts some suspicion on the low susceptible population estimates obtained by the stochastic coalescent SIR and deterministic coalescent SIR methods (median estimates of $S_0 = 727$ and $S_0 = 693$, respectively), since they appear lower than the estimated number of infected individuals from the original study.

There is disagreement in the literature in regard to the modelling of HIV-1 evolutionary dynamics under stochastic or deterministic processes (Nijhuis *et al.* 1998; Rouzine and Coffin 1999; Achaz *et al.* 2004; Shriner *et al.* 2004). The predicament dwells in the observation that the actual effective population size N_e for HIV-1 is often smaller than the total population size (Kouyos *et al.* 2006). While most of this debate has focused on within-host population dynamics, many of the arguments hold when considering the broader epidemic dynamics of host-to-host transmission. As previously mentioned, the appropriateness of these descriptions is hinged on the magnitude of the infected population, precisely, the effective infected population size. Consequently, even when the total infected population is quite large there may yet be significant stochastic effects in play.

Finally, as mentioned in the main article, the existence of two distinct infectious stages and the possibility of large effects due to recombination are reasons for any discrepancy produced by these SIR inference models.

2.2.2 Example XML

Below is an example XML for simulating 100 trees and trajectories in MASTER (Vaughan and Drummond 2013). This example is for $R_0 = 2.4975$ and $S_0 = 999$. The simulation ends when the infected I population returns to zero, i.e., when the last infected individual is removed.

```
<beast version=‘2.0’
namespace=‘master.beast:beast.core.parameter:beast.evolution.tree.TreeHeightLogger’>

    <run spec=‘InheritanceEnsemble’
         nTraj=‘100’
         samplePopulationSizes=‘true’
         verbosity=‘1’>

        <model spec=‘InheritanceModel’ id=‘model’>
            <population spec=‘Population’ id=‘S’ populationName=‘S’/>
            <population spec=‘Population’ id=‘I’ populationName=‘I’/>
            <population spec=‘Population’ id=‘R’ populationName=‘R’/>
            <population spec=‘Population’ id=‘Rh’ populationName=‘Rh’/>

            <!-- infection reaction -->
            <reaction spec=‘InheritanceReaction’ reactionName=‘Infection’ rate=‘0.00075’>
                S + I -> 2I
```

```
           </reaction>

           <!-- recovery reaction -->
           <reaction spec=‘InheritanceReaction’ reactionName=‘Recovery’ rate=‘0.25’>
              I -> R
           </reaction>

           <!-- sampling reaction -->
           <reaction spec=‘InheritanceReaction’ reactionName=‘Sampling’ rate=‘0.05’>
              I -> Rh
           </reaction>
       </model>

       <initialState spec=‘InitState’>
           <populationSize spec=‘PopulationSize’ population=‘@S’ size=‘999’/>
           <lineageSeed spec=‘Individual’ population=‘@I’/>
       </initialState>

       <populationEndCondition spec=‘PopulationEndCondition’
                             population=‘@I’
                             threshold=‘0’
                             exceedCondition=‘false’/>

       <inheritancePostProcessor spec=‘LineageFilter’
                               reactionName=‘Sampling’
                               discard=‘false’/>

       <output spec=‘NewickOutput’ fileName=‘SIR.newick’/>
       <output spec=‘NexusOutput’ fileName=‘SIR.nexus’/>
       <output spec=‘JsonOutput’ fileName=‘SIR.json’/>

    </run>
</beast>
```

[Figure 1 about here.]

[Figure 2 about here.]

[Figure 3 about here.]

[Figure 4 about here.]

[Figure 5 about here.]

[Figure 6 about here.]

[Figure 7 about here.]

[Figure 8 about here.]

[Table 1 about here.]

[Table 2 about here.]

[Table 3 about here.]

[Table 4 about here.]

[Table 5 about here.]

[Table 6 about here.]

References

Achaz, G., S. Palmer, M. Kearney, F. Maldarelli, J. W. Mellors, *et al.*, 2004 A robust measure of HIV-1 population turnover within chronically infected individuals. Mol Biol Evol **21**: 1902–12.

Guss, D. A., 1994 The acquired immune deficiency syndrome: An overview for the emergency physician, part 1. The Journal of Emergency Medicine **12**: 375–384.

Hué, 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. PNAS **102**: 4425–4429.

Kouyos, R. D., C. L. Althaus, and S. Bonhoeffer, 2006 Stochastic or deterministic: what is the effective population size of HIV-1? Trends Microbiol **14**: 507–11.

Kühnert, D., T. Stadler, T. G. Vaughan, and A. J. Drummond, 2014 Simultaneous reconstruction of evolutionary history and epidemiological dynamics from viral sequences with the birth-death sir model. J R Soc Interface **11**: 20131106.

Nijhuis, M., C. A. Boucher, P. Schipper, T. Leitner, R. Schuurman, *et al.*, 1998 Stochastic processes strongly influence hiv-1 evolution during suboptimal protease-inhibitor therapy. Proc Natl Acad Sci U S A **95**: 14441–6.

Rouzine, I., and J. Coffin, 1999 Linkage disequilibrium test implies a large effective population number for hiv in vivo. PNAS **96**: 10758–10763.

Shriner, D., R. Shankarappa, M. A. Jensen, D. C. Nickle, J. E. Mittler, *et al.*, 2004 Influence of random genetic drift on human immunodeficiency virus type 1 env evolution during chronic infection. Genetics **166**: 1155–64.

Vaughan, T. G., and A. J. Drummond, 2013 A stochastic simulator of birth-death master equations with application to phylodynamics. Molecular Biology and Evolution .

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

Volz, E. M., E. Ionides, E. O. Romero-Severson, M.-G. Brandt, E. Mokotoff, *et al.*, 2013 Hiv-1 transmission during early infection in men who have sex with men: a phylodynamic analysis. PLoS Med **10**: e1001568; discussion e1001568.

Volz, E. M., J. S. Koopman, M. J. Ward, A. L. Brown, and S. D. W. Frost, 2012 Simple epidemiological dynamics explain phylogenetic clustering of hiv from patients with recent infection. PLoS Comput Biol **8**: e1002552.