Journal Title Here, 2023, 1–11
doi: DOI HERE
Paper

Predicting Impacts of Contact Tracing on Epidemiological Inference from Phylogenetic Data

Michael D. Kupperman,[1,2] Ruian Ke[1] and Thomas Leitner[1,*]

[1]Theoretical Biology and Biophysics, Los Alamos National Laboratory, New Mexico, United States of America and [2]Department of Applied Mathematics, University of Washington, Washington, United States of America

*Corresponding author. tkl@lanl.gov

FOR PUBLISHER ONLY Received on Date Month Year; revised on Date Month Year; accepted on Date Month Year

Abstract

Robust sampling methods are foundational to many inference problems in the phylodynamic field, yet the impact of using contact tracing, a type of non-uniform sampling used in public health applications, is not well understood. To investigate and quantify how this non-uniform sampling method influences recovered phylogenetic tree structure, we developed a new simulation tool called SEEPS (Sequence Evolution and Epidemiological Process Simulator) that allows for the simulation of contact tracing and the resulting transmission tree, pathogen phylogeny, and corresponding virus genetic sequences. Importantly, SEEPS takes within-host evolution into account when generating pathogen phylogenies and sequences from transmission histories. Using SEEPS, we demonstrate that contact tracing can significantly impact the structure of the resulting tree as described by popular tree statistics. Contact tracing generates phylogenies that are less balanced than the underlying transmission process, less representative of the larger epidemiological process, and affects the internal/external branch length ratios that characterize specific epidemiological scenarios. We also examine a 2007-2008 Swedish HIV-1 outbreak and the broader 1998-2010 European HIV-1 epidemic to highlight the differences in contact tracing and expected phylogenies. Aided by SEEPS, we show that the Swedish outbreak was strongly influenced by contact tracing even after downsampling, while the broader European Union epidemic showed little evidence of universal contact tracing, agreeing with the known epidemiological information about sampling and spread. SEEPS is available at `github.com/MolEvolEpid/SEEPS`.

Key words: Contact tracing, phylodynamics, phylogenetic inference, HIV-1, phylogenetic trees

Introduction

The growth and prevalence of communicable diseases, in which a human individual transmits a pathogen to another individual, without an intermediate vector or reservoir, has led to the development of a variety of detection and surveillance strategies. With the notable exception of zoonotic spillover events, each infection can be attributed to another, older, infection. This basic insight lead to the remarkable development of contact tracing as a core method to efficiently identify closely related infections (Centers for Disease Control, 1986, 1987; Giesecke *et al.*, 1991; Hethcote and Yorke, 1984) resulting in significant contributions to public health (Ramstedt *et al.*, 1990). Indeed, while contact tracing has been successfully used to trace many infectious diseases, including the recent SARS-CoV-2 epidemic (Turcinovic *et al.*, 2022), and been evaluated in mathematical models (Müller and Kretzschmar, 2021; Höhna *et al.*, 2011), there remains little knowledge on how contact tracing may impact and interact with genetic sequence data analyses.

Due to the non-random nature of contact tracing, one might expect that the phylogenetic tree structure of the spreading pathogen could be impacted by such sampling. That raises fundamental questions about the nature of the data that typically is used for phylogenetic and phylodynamic reconstruction of pathogen epidemics: how robust are mathematical assumptions made about the collection of data in practice, and how significant are deviations from these assumptions in real data? While the contact network that pathogens spread across can be informative of the pathogen's phylogeny (Giardina *et al.*, 2017), it remains largely unknown how sampling with contact tracing impacts the observable phylogeny.

A standard form of contact tracing is "iterative contact tracing", in which an initial index case is interviewed to identify contacts which may be infected. Identified contacts are tested, and

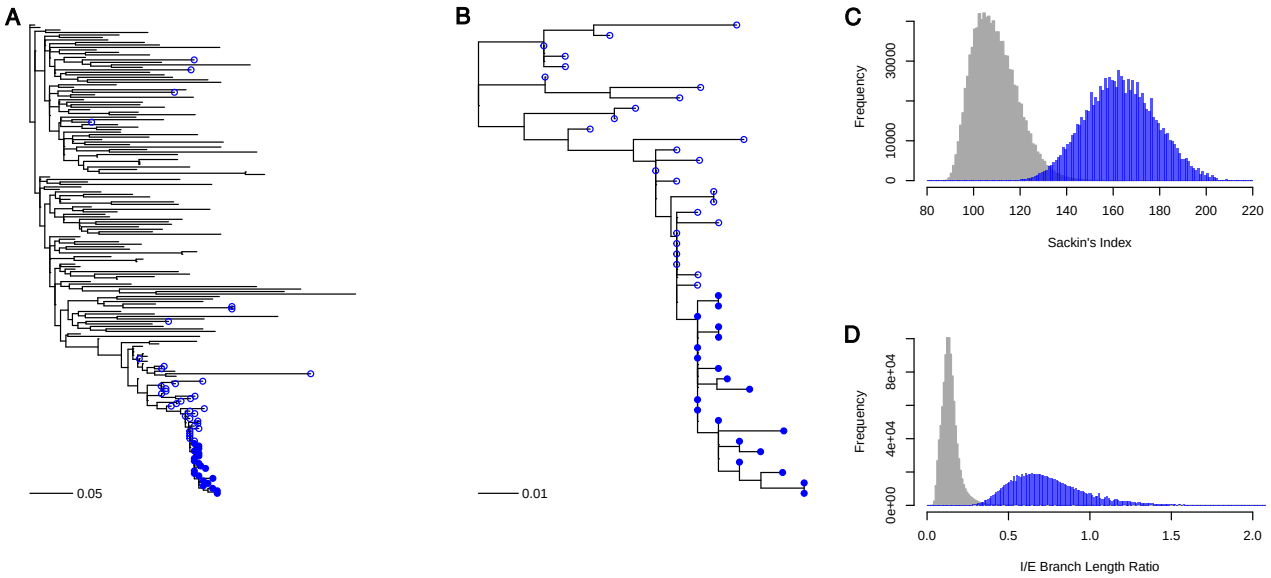

Fig. 1: Contact tracing induces variations in phylogenetic tree structure. Panel A shows the full reconstructed HIV-1 CRF01 phylogeny of sequences collected in Europe, with tips from Sweden in blue. Filled symbols denote 20 closely related samples identified through contact tracing from a known injection drug user outbreak, while unfilled symbols denote additional Swedish samples. Panel B zooms in on the bottom subtree consisting entirely of Swedish sequences. Note that the shape of this subtree is drastically different than the full tree in panel A. To quantify the difference we downsampled 20 tips randomly (without replacement) from the tree without the Swedish subtree taxa in panel A and the Swedish taxa in panel B 1,000,000 times each, and recorded both the Sackin's index and internal to external (I/E) branch length ratios in panels C and D, respectively. The blue distributions are from the Swedish subtree in panel B and the grey distributions from the full European tree without the Swedish subtree taxa. Comparing the distributions with a Kolmogorov-Smirnov test clearly showed very different distributions: D = 0.964, $p < 10^{-16}$ for Sackin's index and D = 0.988, $p < 10^{-16}$ for the I/E branch length ratios. Trees were inferred by maximum-likelihood under a GTR+I+G substitution model (Guindon *et al.*, 2010). Scale bars in panels A and B are in units of substitutions/site.

the interview process is repeated for contacts with positive test status. From a statistical perspective, contact tracing provides a correlation structure to the detection process informed by the transmission network structure. If an individual tests positive for HIV, their contacts have a better than uniformly random probability of being sampled. Importantly, first degree contacts are expected to be evolutionary closer to an index case than could be expected from a random sample of the larger population, hence displaying smaller genetic distances. This could consequently impact both distributions of pairwise distances and phylogenetic tree structures in complex ways.

A motivating example to consider is the spread of HIV-1 circulating recombinant form 1 (CRF01) in Europe in the late 1990's and 2000's (Figure 1). HIV-1 CRF01 was originally introduced in southeast Asia from Africa (Gao *et al.*, 1996; McCutchan *et al.*, 1996), and later spread from there to other parts of the world, including several countries in Europe, on many occasions and still ongoing today (Hemelaar *et al.*, 2020). Thus, the available HIV-1 CRF01 sequences from Europe cannot be strongly influenced by contact tracing, as they are not closely related within Europe nor have there been cross-border coordinated sampling efforts. In contrast, a Swedish HIV-1 CRF01 outbreak among injection drug users in 2007-2008 (Skar *et al.*, 2011) elicited a strong public health response resulting in identifying further persons who had been in contact with those infected with this HIV-1 variant, generating many closely related sequences. Hence, part of the resulting European HIV-1 CRF01 phylogeny comes from strong contact tracing while the larger part comes from essentially random sampling. The two parts of the European HIV-1 CRF01 tree highlights the strong impact contact tracing may have on the tree structure, affecting both topological and branch length statistics.

Several simulation techniques have been proposed for generating detailed pathogen phylogenies such as FAVITES (Moshiri *et al.*, 2019), BEAST2 (Bouckaert *et al.*, 2014), or PopART IBM (Pickles *et al.*, 2021). These software tools facilitate simulations of individual, contact network, and population-level models and thus are able to generate phylogenies with a broad variety of features to investigate potential epidemiological assumptions. However, none include contact tracing or similar sampling methods. FAVITES comes close in offering a sampling method weighted by the number of transmission events, but this method results in sampling towards more interconnected individuals, rather than following local structures. As a result, these simulation tools are not suitable for investigating the impact of contact tracing on phylogenies. Since contact tracing is very common practice, and may have strong impact on phylogenetic structures (Figure 1), a new simulation software capable of

emulating contact tracing is needed to formally include its impact on pathogen phylogenies.

Here, we explore the impact of contact tracing on phylogenetic tree structure. To directly address questions about the significance of contact tracing, we developed a new simulation suite in R called SEEPS (Sequence Evolution and Epidemiological Process Simulator) that allows for the simulation of contact tracing, the resulting transmission trees, and pathogen phylogenies. By using an agent-based model, trees can be directly simulated, only specifying simple rules and behaviors. Using SEEPS, we show that both topological and distance based tree measures are sensitive to the presence of contact tracing, demonstrating that contact tracing can significantly impact the structure of the resulting tree. We show that SEEPS can simulate the Swedish HIV-1 CRF01 outbreak and the corresponding broader EU HIV-1 CRF01 epidemics presented in Figure 1, and give a coarse estimate of the performance of contact tracing present in these data. In agreement with known epidemiological data, we find that the Swedish outbreak is strongly influenced by contact tracing, while the broader EU epidemic shows little evidence of contact tracing.

New Approaches

Simple models for contact tracing and transmission dynamics can be directly implemented in a computational environment using an agent based simulation. Building off the agent-based HIV-1 model in (Kupperman *et al.*, 2022), we developed SEEPS (Sequence Evolution and Epidemiological Process Simulator), an end-to-end modern and modular simulator for investigating the connection between evolutionary and epidemiological mechanisms. Written in R (R Core Team, 2022), SEEPS is a flexible and extensible framework for simulating phylodynamic and evolutionary processes at a population level. SEEPS stores the entire transmission history, allowing for models of contact tracing to be run on top of the transmission history and directly compared. Individuals in SEEPS are considered *active* if they are capable of generating secondary infections. SEEPS offers both high level and low level modeling tools, enabling both coarse and fine-grained mechanisms to be modeled.

An experiment in SEEPS begins by simulating a population of infections with user-defined expected offspring generation rates. The population is sampled at user-defined time points, with sampled individuals being removed from the simulation. Once the transmission tree is sampled, SEEPS offers a module for simulating the within-host diversity using a coalescent process from (Lundgren *et al.*, 2022). By modeling each infected individual as a host where further viral evolution can occur and offspring are sampled from, SEEPS can explicitly convert a transmission history into a possible phylogeny by taking the within-host evolutionary diversification into account, often resulting in reordering the host transmission history tree into a pathogen phylogeny.

Both genomic sequences and phylogenetic trees are often used to test analysis pipelines. Sequence simulation is available with a GTR+I+Γ model using Seq-Gen (Rambaut and Grass, 1997) and the PhyClust R package (Chen, 2011). SEEPS can export trees in Newick format for use in other standard phylogenetic analysis software, such as the R package ape (Paradis and Schliep, 2019). Distance matrix representations of the data are also available for export, either using cophenetic distances for trees, or pairwise evolutionary distances (such as TN93) for sequences. A general schematic of simple workflows available in SEEPS is shown in Fig 2.

To study the impact of contact tracing, we implemented a simple algorithm to describe contact tracing in SEEPS. Our model captures the fundamental aspects of iterative contact tracing where each positive contact is discovered with probability $0 \leq p \leq 1$, and is similar to the popular breadth-first-search algorithm (Lee, 1961), but with the variation that the discovery of edges is randomized with a prescribed failure probability $1 - p$.

Using this model, we can generate complex trees reflecting a wide variety of scenarios. In Fig. 3AB, we show two example phylogenies generated by SEEPS. Crucially, SEEPS tracks the entire transmission history of the sampled taxa because it is needed for the proper modeling of the resulting phylogeny, as it informs all intermediate transmission bottlenecks that impact the diversification of the sampled viruses.

In Fig 3CD, we removed the unsampled taxa from the trees, trimmed the resulting internal branches, and collapsed any resulting internal nodes of degree 2. The trees in Fig 3AC were generated with a high contact tracing discovery probability, while the trees in Fig 3BD were generated with a low contact tracing discovery probability. Both scenarios were generated in two stages: First, we simulated an outbreak as exponential growth, and second, a constant population size. Sampled individuals were removed from the active population to reflect that they were either on efficient antiviral treatment or otherwise non-infectious after diagnosis. The trees are visually distinct, with high contact tracing resulting in large clusters being identified which are not closely related to each other. In contrast, low contact tracing identifies small clusters that are loosely related. Thus, varying the sampling method to compare contact tracing against random sampling (which is typically assumed in phylodynamic inferences) may give very different impressions of what appeared to have happened.

Results

Contact tracing makes trees less balanced

In a first set of simulation experiments, we consider the impact of contact tracing on Sackin's index for a collection of taxa taken at a single time point, known as cross-sectional sampling. We simulated 1,000 outbreaks followed by a constant population size for 0 to 10 years (in one year increments), with R_0 uniformly distributed between 1.5 and 5. For each outbreak, we sampled either 15 or 50 taxa, with contact tracing performed at either high ($p = 0.9$) or low ($p = 0.1$) levels. In total, we generated 22,000 transmission trees and 22,000 phylogenies.

We found no clear correlation between Sackin's index and R_0 and no effect of the number of years after the outbreak phase in which a cross-sectional sample was taken. The simulation of the transmission history resulted in an average Sackin's index close to what could be predicted from a Yule model (Kirkpatrick and Slatkin, 1993) when contact tracing performance was low (Fig 4). Conversely, when contact tracing was high ($p = 0.9$), the Sackin's index became elevated above the Yule expectation. Further, adding the within-host diversification process (phylogeny) increased Sackin's index only slightly for both low and high levels of contact tracing. In all configurations, the sampled trees include a Sackin's index close to the minimal possible value for the number of sampled taxa (15 or 50) (Fischer, 2021).

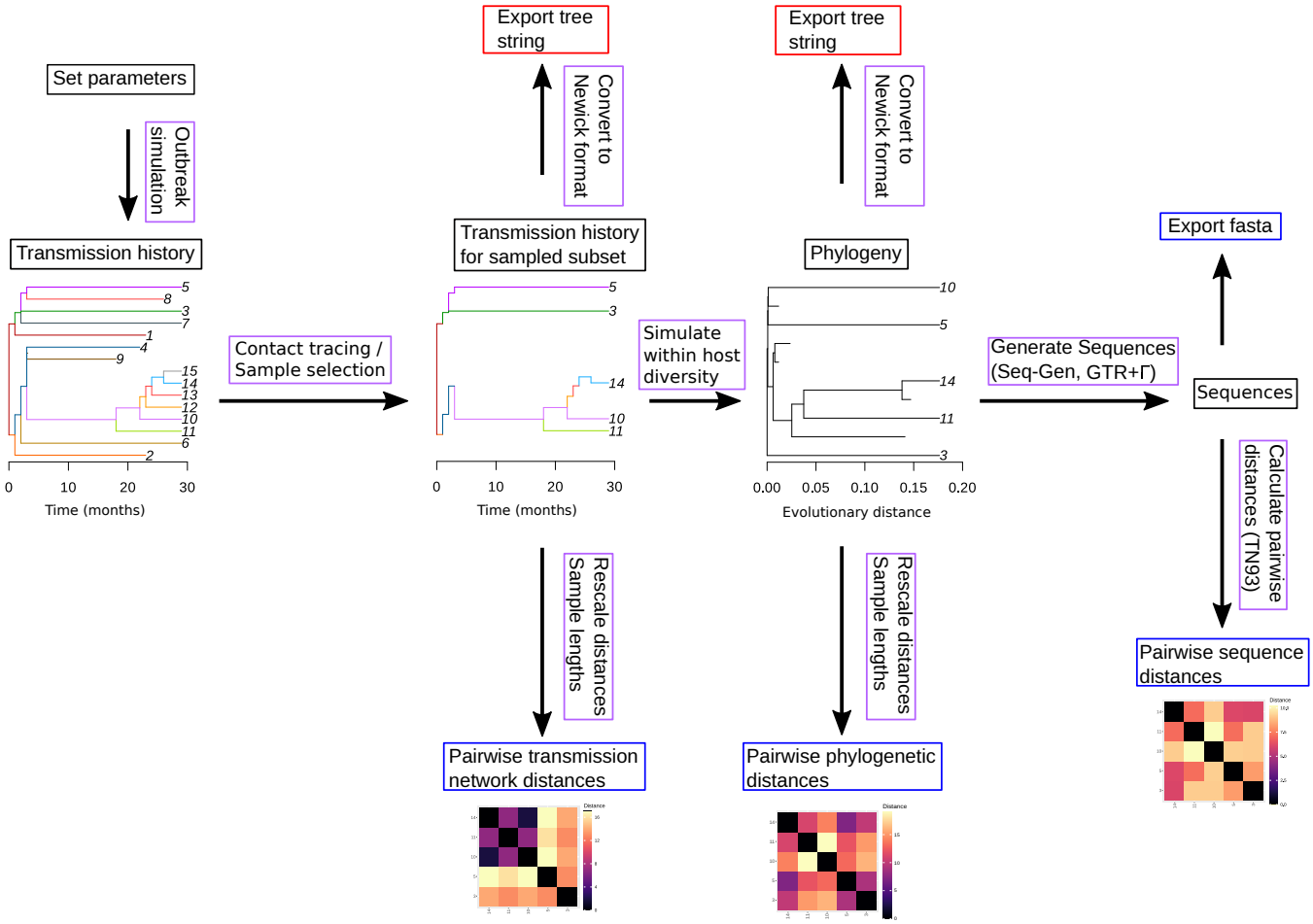

Fig. 2: General workflow for the SEEPS package. Purple boxes denote functionality provided by SEEPS. Blue boxes denote exportable data. Red boxes denote trees that can be read into the R package ape. Black boxes denote internal states or data available for manipulation.

Trees based on contact tracing in a small sample do not represent the larger epidemic

A potential concern with data generated with contact tracing is that it may result in missing, unconnected, or undiagnosed persons, thus making the sample unrepresentative of the larger population (Blum and Tran, 2010). Since we now know that contact tracing biases trees to be more unbalanced, this raises concern about how representative a phylogeny would be of the greater epidemic. Thus, we assessed how representative a second, later in time, sample would be of an earlier sample from the same epidemic.

We simulated outbreaks under varying R_0 in SEEPS to an effective population size of 1,000 infections and sampled 50 active infections as soon as the effective population surpassed 900 active infections. We let the population replace the removed infections while simulating forward for 3, 24, or 120 months. We then drew another 50 taxa, for a total of 100 sampled taxa. The growth rate parameter R_0 took discrete values of $1.1, 1.5, 3, 5$, or 10.

We used a parsimony score to report the number of "transitions" that were required to render the first sample labels into the second sample labels. Thus, this parsimony score represents how much of the original tree structure the second sample recovers; a smaller score would indicate a different tree while a higher score a more similar tree.

We found a strong relationship between the mean parsimony score and both contact tracing and the length of the inter-sampling period (Fig 5 A-E). Increased contact tracing decreased the parsimony score, indicating that the two samples represented different parts of the total epidemic. R_0 only weakly influenced the relationship between parsimony scores and contact tracing performance; R_0 primarily impacted the parsimony score when contact tracing performance was high by increasing the variance. As Fig 5E demonstrates, setting $R_0 = 10$ indicated that the variance increased after approximately $p = 0.5$, while in Fig 5A it occurred close to $p = 1$ when R_0 was close to 1. In Fig 5A-E, the symmetric inner 50% is shown as a shaded region to capture the variance. Fig S2 shows additional bands to provide a more complete picture of this effect.

We next investigated whether the sample size (number of taxa investigated) at 3 months after the first sampling time influenced the parsimony score. In Fig 5F-J, R_0 was varied as before, but

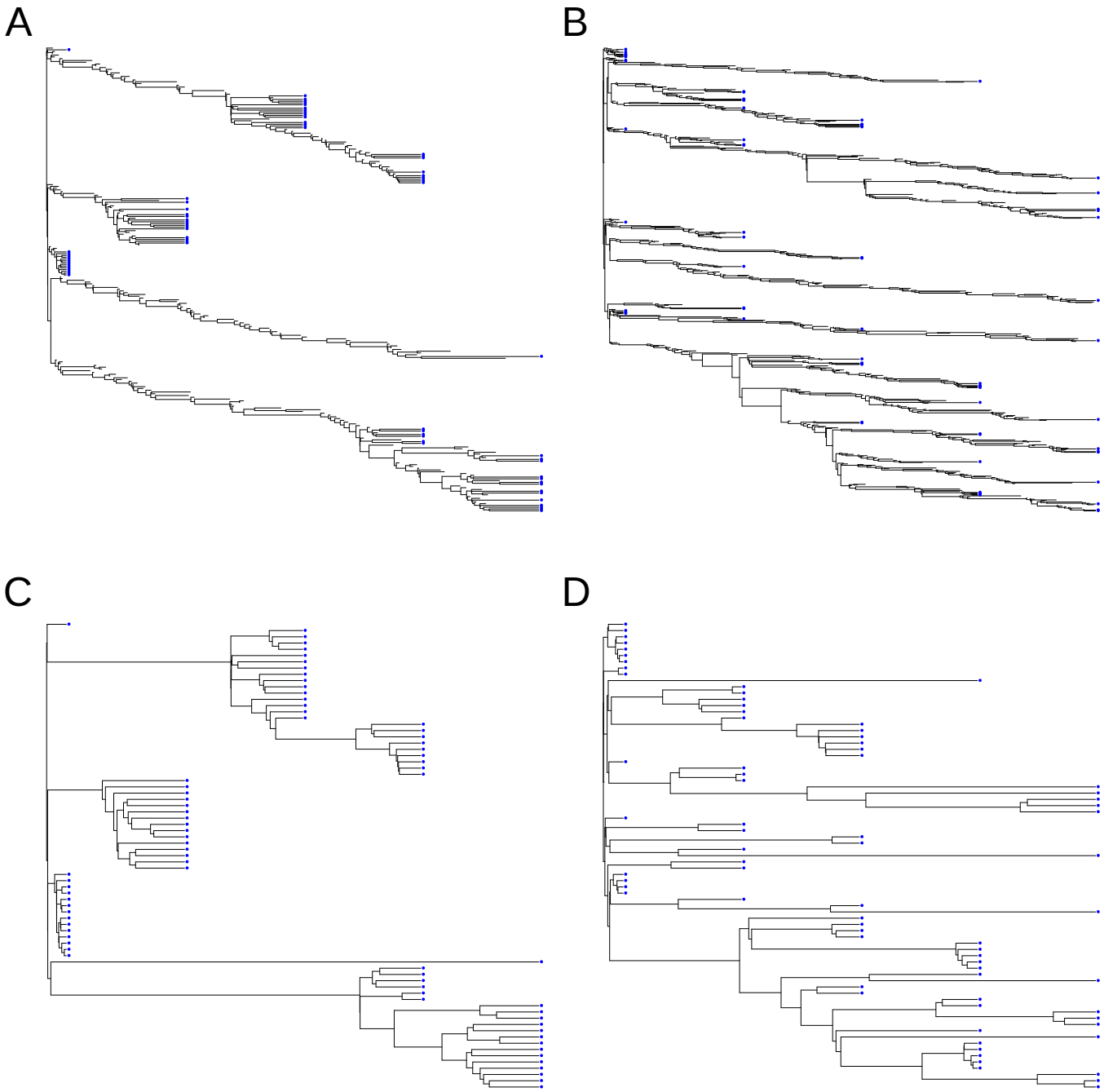

Fig. 3: Examples of how repeated sampling with different contact tracing levels can lead to visually distinct tree structures. Panel A (high contact tracing) and panel B (low contact tracing) show examples of two simulations provided by SEEPS. Note that SEEPS simulates within host diversity for all ancestors, resulting in explicit modeling of the transmission bottlenecks and the realistic population diversity. Panels C and D show only the "observable" history that could be inferred by reconstructing the ancestral relationships between the observed sequences. Sampled taxa are denoted by a blue circle.

we sampled between 5 and 250 taxa at each time point (from a target population size of 1,000). The normalized parsimony score shows little dependence on the sample size, as long as a "small sample" approximation remains reasonable. With as few as five taxa, however, the lowering effect of contact tracing on the parsimony score is diminished because the small tree size limits its range.

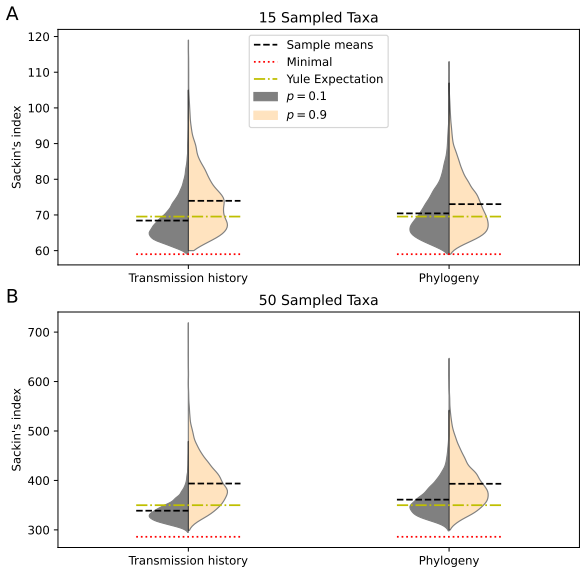

Fig. 4: Violin plots of the distribution of Sackin's index with low performing contact tracing ($p = 0.1$) and high performing contact tracing ($p = 0.9$). Horizontal reference lines are added at the minimal value of Sackin's index (red) for a tree of the given size, the expected value of Sackin's index under a Yule model (yellow), and the sample mean (black). Adding contact tracing increases Sackin's index, regardless of whether the transmission history tree or the sampled phylogenetic tree is considered. In the presence of contact tracing, the Yule model does not well describe the distribution.

Mean internal to external branch length ratio is affected by contact tracing

The mean internal to external (I/E) branch length ratio quantifies the recent evolutionary relationship between sampled taxa. The external branches are the tips of a tree, ending with the sampled taxa, and the internal branches connect all nodes of the reconstructed ancestors of the sampled taxa. Hence, if the ratio is small, then the tips are longer, suggesting that the taxa are not recently related. If the ratio is large, then the samples are recently more related, suggesting the possibility of an epidemiologically significant cluster. Previous work (Giardina *et al.*, 2017) suggests that the branch length ratio can be informative about possible recent outbreaks in a population, but the impact of contact tracing on the I/E ratio has not been evaluated.

Using the same synthetic data set we examined for Sackin's index, we computed the I/E branch length ratio for each phylogenetic tree (Figure 6). The full data used to generate this figure is available in the supplementary materials online Fig S2. While the R_0 growth rate of the outbreak had some impact on the mean I/E branch length ratio, the presence or absence of contact tracing amplified the effect of when sampling occurred relative to the epidemic outbreak on the I/E branch length ratio. R_0 was influential only when it was low; epidemics at $R_0 > 2.5$ taken at the same time point had similar I/E branch length ratios, whereas the I/E branch length ratios increased when R_0 was $R_0 < 2.5$. Interestingly, the I/E branch length ratio was less sensitive to contact tracing immediately after the peak of the outbreak. At low contact tracing ($p = 0.1$), the I/E ratio was never able to rebound past the initial outbreak signal. In contrast, at high contact tracing ($p = 0.9$), the samples taken three years after the end of the exponential growth phase had similar I/E branch length ratios to the samples taken immediately after the peak of the outbreak. Thereafter, the I/E branch length ratio statistic continued to grow with time.

This suggests that some amount of contact tracing early in an epidemic, enough to find a recent nearby infection, is necessary to recover a time signal from the internal branches and indicate the age of the outbreak.

Contact tracing can be observed in real data

Our simulations showed that contact tracing has strong effects on phylogenetic tree reconstructions, and therefore on any epidemiological inference that would be based on such trees. To tests whether we could recover the epidemiological data of our motivating example in the introduction, including the levels of contact tracing in the European and Swedish partitions, we attempted to use SEEPS to simulate the rather complicated European HIV-1 CRF01 epidemic. This epidemic can be divided into three parts: 1) the exponential outgrowth of a new form of HIV-1 in Thailand in the early 1990's, 2) subsequent introductions of several genetically distant lineages into Europe, and 3) an introduction from Europe (Finland) into a previously uninfected injecting drug-user network in Sweden with an explosive outbreak.

To estimate the level of contact tracing, we simulated epidemics in SEEPS similar to the European and Swedish outbreaks with varying levels of contact tracing and compared the I/E branch length ratios from our simulated trees to that of the real data (Figure 1). The parameter values used to generate the simulated data are shown in table S1. While we used different parameters for each outbreak, we used a similar two-phase simulation for both the European and Swedish partitions: In the first phase, we started the simulation with a single infected individual and allowed the population to grow to a small, fixed size. Once the population reached the fixed size, we let it continue at that size until the end of the phase. In the second phase, we increased the effective simulated population size to our target value, and allowed R_0 to change. As before, the population was allowed to grow until the end of the phase. To mimic the import to Europe of genetically distant lineages from Thailand, we shifted the sampling time of the European sequences forward by 18 years to reflect the lack of available sequences along the long branches that constitute the introductions into Europe. Finally, we sampled 20 individuals with varying levels of contact tracing discovery probability according to the sample years of the EU and Swedish outbreaks, respectively. We then calculated the I/E branch length ratio for each simulated tree and compared the distribution against the real data distribution.

To compare simulated and real I/E ratio distributions, we computed the relative difference of means $(\bar{x} - \mu)/\mu$ and the Kolmogorov-Smirnov test statistic $D = \sup_{x \in [0,\infty]} |F(x) - G(x)|$ where F and G are the cumulative distribution functions for the simulated and the real data, respectively. For these two fitting statistics, we effectively randomized both the population size and R_0 parameters, resulting in 10,000 samples to approximate

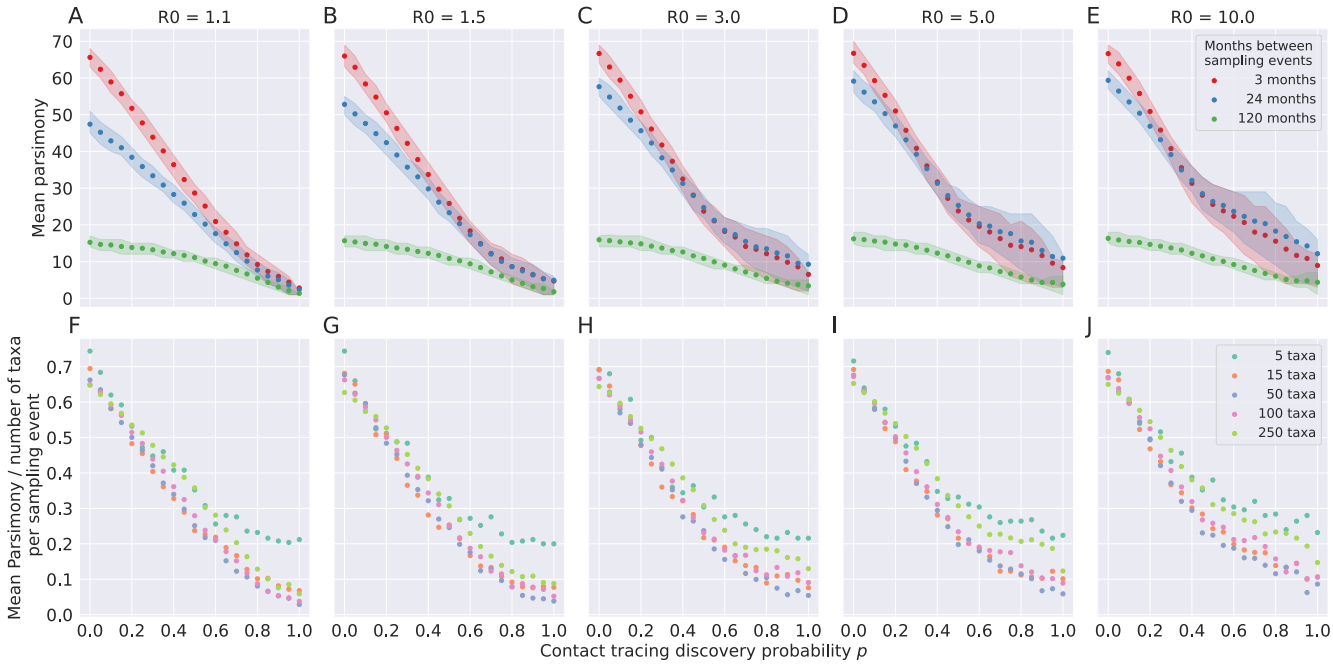

Fig. 5: Parsimony distributions are strongly related to contact tracing performance p and weakly with sample size. The strength of the correlation is primarily dependent on R_0 when contact tracing is good. In A-E, a sample of 100 taxa (pathogen sequences from 100 infected hosts) is obtained at the end of the exponential growth phase, and compared against another sample of 100 taken 3, 24, or 120 months later. The shaded region denotes the symmetric inner 50% of the data. In F-J, the experiment is repeated for the 3-month interval between samples, however the sample size is varied. The parsimony score is normalized against the number of taxa in the sampled tree (two times the reported sample size above). Sampled individuals are removed from the population.

the European epidemic and 320,000 samples to approximate the Swedish outbreak at each level of contact tracing discovery probability p.

Both the KS statistic and the relative difference of means suggest that the European data was generated in a situation with negligible contact tracing, with an upper bound on the contact tracing discovery probability of at most 10% (Figure 7). A visual inspection of the phylogenetic tree (Fig 1A) suggests that the most recent common ancestor between most pairs of sequences in Europe is not in the recent history, or that the European sample is sparsely reported by the available sequences (from the LANL HIV database hiv.lanl.gov). This is consistent with multiple HIV-1 CRF01 introductions into Europe, each with limited local spread.

In contrast, the subsampled Swedish data was consistent with a significant level of contact tracing. Both the KS statistic and relative difference of the means were indicative of a contact tracing discovery probability of approximately $p = 0.6$. We expected that the level of contact tracing would have been very high in this intensely followed outbreak (Skar *et al.*, 2011). However, two effects may have lowered the estimated contact tracing level: 1) Not all infections in the outbreak may have been included in the sequencing, which would affect some I/E ratios, and 2) subsampling the Swedish outbreak phylogeny removed over half of the taxa in each draw, which further may have lowered the estimated contact tracing level. Taking both effects into consideration, our recovery of a contact tracing discovery probability of approximately $p = 0.6$ indicates that a strong effort of contact tracing took place in the discovery of the Swedish outbreak.

Discussion

We developed an epidemiological and evolutionary simulation model that includes contact tracing, available as an R package called SEEPS. Using SEEPS, we showed that there can be a serious impact when pathogen sequences were collected by contact tracing on the resulting phylogenetic tree structure. Overall, contact tracing resulted in a phylogeny that 1) was more unbalanced, 2) was less representative of the larger epidemic, and 3) had its I/E branch length ratio differently impacted depending on when samples were taken relative to an outbreak. We then analyzed a real data set describing a known outbreak of HIV-1 CRF01 in Sweden, derived from Europe, which in turn had derived several lineages from Thailand. This showed that SEEPS was able to simulate a fairly complicated epidemiological scenario and, importantly, also correctly detected contact tracing as it was used in Sweden.

Because sequence- and phylogeny-based approaches have the potential to reveal otherwise difficult to measure details about how pathogens spread, previous work has evaluated several tree statistics related to the branching structure and the branch lengths, such as Sackin's index and internal to external (I/E) branch length ratios. For example, such statistics have been used to tune MCMC methods that explore the space of phylogenetic trees. Here, we showed that both of these classes of tree

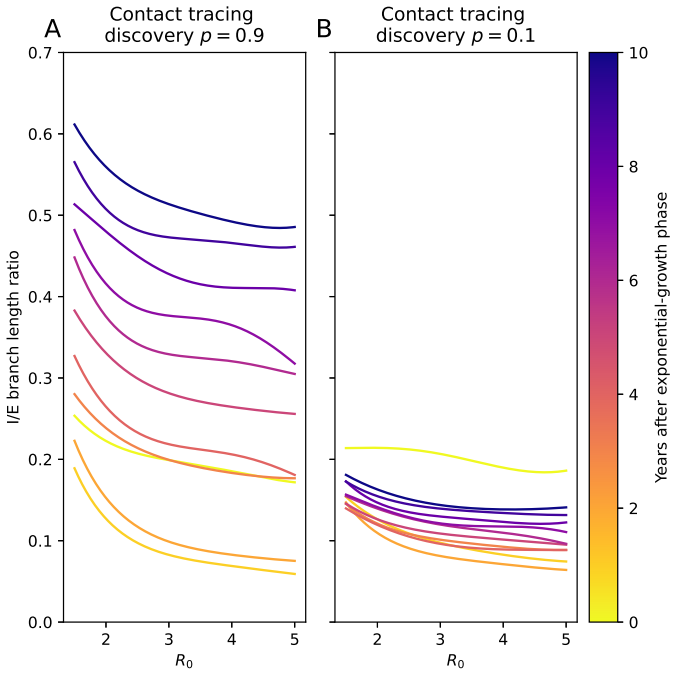

Fig. 6: Estimated mean of the internal to external (I/E) branch length ratios as a function of R_0, stratified by number of years after the peak. The mean trend line was estimated using a Gaussian process model (radial basis function kernel, $\alpha = 3 \times 10^{-5}$, $\ell = 3$) (Pedregosa *et al.*, 2011). The color of the trend line indicates when sampling was performed after the epidemic exponential growth had ended. The I/E ratio interpretation depends on both R_0 and contact tracing level.

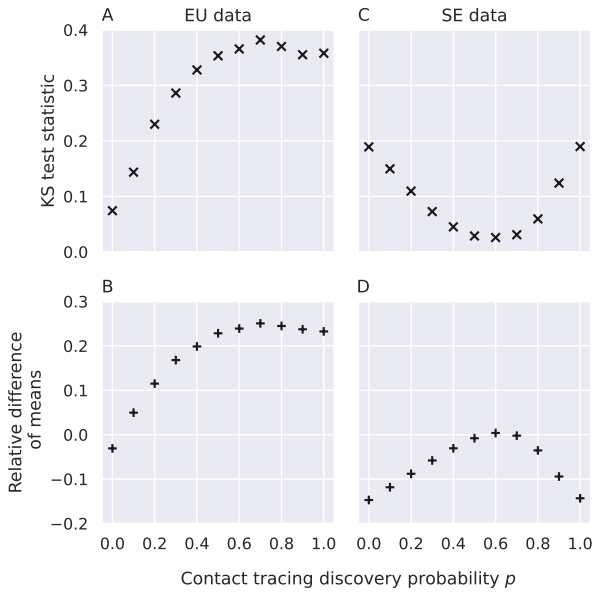

Fig. 7: Comparison of simulated I/E branch length ratio with subsamples taken from the European epidemic and the Swedish outbreak. The contact tracing discovery probability p was varied from 0 to 1. Panels A and B compare simulated data against the European Union subsamples, while panels C and D compare against the Swedish subsamples. The relative difference of means is defined as $(\bar{x} - \mu)/\mu$ where $\bar{x}$ is the observed sample mean and μ is the sample mean from the associated reference distributions.

measurements are affected by contact tracing. Thus, assuming that sequences have been randomly collected, when in fact they were collected as the result from contact tracing may severely mislead analyses and conclusions from sequence- and phylogeny-based epidemiological inferences. One reason for that is that contact tracing results in samples that may be less representative of the larger epidemic.

Contact tracing is a non-random sampling strategy that efficiently finds linked infections (Centers for Disease Control, 1986, 1987; Giesecke *et al.*, 1991; Hethcote and Yorke, 1984). While previously largely ignored in sequence- and phylogeny-based epidemiological inferences, non-uniform sampling strategies such as contact tracing have been investigated in other contexts, such as generalized birth and death models (Höhna *et al.*, 2011) and ordinary differential equation compartment models for outbreaks (Hyman *et al.*, 2003). In (Höhna *et al.*, 2011), it was shown that characteristic rates for the branching process can be incorrectly inferred if an incorrect sampling scheme is assumed, which agrees with our results showing that contact tracing impacts phylogenetic tree structures. In (Hyman *et al.*, 2003) it was shown that the significance of contact tracing as a control mechanism is sensitive to the degree of superspreading. However, those results were based on the assumption of independence and the mechanism of spread through mass action, which breaks down when using sequence- and phylogeny-based epidemiological inferences. While contact tracing may lead to a phylogeny that does not represent the larger epidemic - from a public health perspective, detecting superspreaders is important because they contribute more to overall disease spread. Contact tracing will be much more likely to find superspreaders than random sampling simply because they are more likely to be traced from any one of the people they infected.

SEEPS includes within-host evolution that simulates diversification under a neutral coalescent process. The within-host pathogen diversification is important to account for because it affects the observable phylogeny, which always is different from the non-observable transmission history (Graw *et al.*, 2012; Romero-Severson *et al.*, 2014; Giardina *et al.*, 2017). However, SEEPS does not simulate the selection of escape mutants driven by the host immune system. Because selection also can cause an imbalance in the tree structure, our simulations may be on the conservative side of the impact contact tracing has on the global tree structure of an epidemic. Furthermore, if superspreaders were active, the simulations under our neutral model may show less impact of superspreading than in real epidemics.

Methods that depend on analyzing distances such as HIV-TRACE (Kosakovsky Pond *et al.*, 2018) or machine learning based methods such as convolutional neural network (CNN) models (Kupperman *et al.*, 2022) are inherently sensitive to the distribution of pairwise distances. Contact tracing results in samples that can be significantly closer than random. If clusters are interpreted as outbreaks, then clusters discovered by non-uniform sampling may be correctly labeled as transmission clusters, but erroneously inferred as signs of a larger outbreak. Thus, the performance of HIV-TRACE and the CNN model in detecting outbreaks may be sensitive to how samples were collected. Hence,

popular analytical methods and computational tools used to trace and reconstruct epidemics need to ensure that the impact of contact tracing is not being overlooked or misinterpreted.

Materials and Methods

Simulations in SEEPS

We developed SEEPS to perform generative computational experiments on HIV-1 phylodynamics, generalizing the model and framework from (Kupperman *et al.*, 2022). All experiments in the present study were performed in SEEPS v0.2.0 available at `github.com/molEvolEpid/SEEPS`. Code associated with specific analyses is available at `github.com/molEvolEpid/ContactTracingForPhylogenies`.

SEEPS provides a stochastic forward simulation that tracks the transmission history of the entire simulation (including non-sampled individuals) and maintains a list of active individuals that are capable of generating new offspring. We implemented an agent-based discrete time model of transmission dynamics which randomized the length of each infection between one and three years, with the expected number of lifetime transmissions being equal to R_0. Here, we used a simple biphasic rate for offspring generations where the initial transmission rate was 20-fold higher in the first three months of infection (Graw *et al.*, 2012). The simulation time step was 1 month in duration. When multiple sampling time points were used, the simulation stopped at each, samples were taken and removed from the population, and then the simulation resumed.

After prescribing the population dynamics (here, exponential growth or constant size), samples were taken at fixed time points through contact tracing and the transmission history for the subset of sampled individuals was reconstructed. We reconstructed both 1) a reduced transmission history, where unsampled tips were removed and any internal nodes of order two were collapsed, and 2) a complete transmission history for the samples where only pruned branches were removed. The reduced transmission history was converted into a transmission tree. The complete transmission history for the sampled individuals was used to obtain a phylogeny by simulating a (neutral) coalescent process along the transmission history (Lundgren *et al.*, 2022). Within-host diversity is modeled assuming $a = 5$ and $b = 5$ as in (Romero-Severson *et al.*, 2016). This process reconstructs a possible pathogen phylogeny within the transmission tree and forces coalescence events to occur before time 0 (Romero-Severson *et al.*, 2014). Individuals in the transmission history that are not sampled are simulated where needed to ensure that the entire evolutionary history of the samples is correct with respect to transmission bottlenecks and diversification level until the next transmission event. This process can introduce additional tips into the phylogeny that do not correspond to sampled individuals. These were removed in our main analyses, but are available within SEEPS for inspection (as shown in Fig 3).

Both the transmission tree and sampled phylogeny were exported to the R package ape (Paradis and Schliep, 2019).

Contact tracing

To study the impact of contact tracing, we developed a simple algorithm to describe contact tracing as follows: An initial index case is randomly discovered in the population, and all first order contacts (secondary infections and the source itself) are identified by examining the transmission history. Then, each identified contact is *independently* discovered at probability $p \in [0, 1]$. Hence, the parameter p denotes the contact tracing discovery probability of each contact. If $p = 0$, there is no contact tracing, while $p = 1$ corresponds to perfect contact tracing. The discovered individuals are added to a list of discovered individuals, and the identify-and-probabilistically-discover process is repeated for each newly discovered individual. This is repeated until there are no discovered cases to trace, or until a maximum number of *active* individuals have been identified. If the desired number of individuals are not sampled, the process is re-started with a new random index case and repeated until the desired number of individuals have been sampled.

Tree statistics

As phylogenetic trees are complex objects, there are many statistics, indexes, and measures that have been proposed for analyzing trees (Fischer *et al.*, 2021). In this work, we focused on some of the most popular statistics that have been widely used to analyze phylogenetic trees and that we expect to be influenced by contact tracing, including both topological and branch length effects. We considered Sackin's index (Sackin, 1972; Shao and Sokal, 1990) using the R package treebalance (Fischer *et al.*, 2021) to assess topological effects, and the internal/external (I/E) branch length ratio to assess branch length effects.

Sackin's index measures the imbalance of a tree. It is maximized in a caterpillar (ladder-like) topology, while being minimized in a fully balanced tree. In the absence of contact tracing, i.e., with uniform random sampling, we expect the Sackin's index to be low, because then the topology would be informed by random ancestral relationships during the initial exponential growth phase. In the presence of contact tracing, we expect to primarily recover recent information about the transmission history related to epidemiologically significant clusters. While such clusters also are linked together by ancestral relationships, if the contact tracing is good, we expect the majority of the tree to reflect recent transmission events.

We used parsimony to assess how representative a sample of taxa is of a larger epidemic. We did this by sampling twice in an epidemic and compared how representative the second sample was of the phylogeny obtained from the first sample. The taxa of the first sample were labeled "A" and the taxa of the second sample "B". Given the phylogeny of both "A" and "B" labeled taxa, we calculated the number of A $\rightarrow$ B transitions, i.e., the parsimony score of label transitions. We computed the parsimony score using the R package phangorn (Schliep, 2011), and the process was repeated 200 times for each combination of R_0, time point, sample size, and contact tracing discovery probability p. If uniform random sampling was performed, we expect the parsimony score to be high, because the taxa are drawn from the entire epidemic, i.e., the entire phylogeny, independently or nearly independently. If contact tracing was performed, we expect each group of taxa to contain more cluster-like relationships, which would be more informative about local spread but not the entire epidemic. Thus, because more taxa are closely related when contact tracing has occurred, fewer "A" $\rightarrow$ "B" transitions are required, so the resulting tree would have a lower parsimony score.

The I/E ratio is informative of the recent evolutionary relationship between taxa as well as the overall tree structure (Giardina *et al.*, 2017). In the absence of contact tracing, we expect

the ratio to be low, as many external branches will connect the taxa back to an ancestral event in the outbreak phase. In contrast, if there is contact tracing, we expect the ratio to be high as the most recent common ancestor between two taxa in an identified cluster will be much more recent.

HIV-1 CRF01 European sequence data

Data from the European HIV-1 CRF01 epidemic was extracted from the LANL HIV database (hiv.lanl.gov). GenBank accession numbers for all sequences used in this study are available at `github.com/molEvolEpid/ContactTracingForPhylogenies`.

The data consisted of 34 *env* V3 region sequences (approx. 300 nt) from an intravenous drug user (IDU) outbreak in Stockholm, Sweden, in 2006-2007 (Skar *et al.*, 2011) and 155 corresponding European sequences from 2003-2007 (including 23 additional Swedish sequences not involved in the IDU outbreak and 132 sequences from 12 other countries). The entire European HIV-1 CRF01 tree was reconstructed using PhyML v3 under a GTR+I+G model by both NNI and SPR search (Guindon *et al.*, 2010).

Using SEEPS, we simulated 110,000 and 3,520,000 sequences for the EU and Swedish outbreaks respectively with varying levels of contact tracing. Sampling dates were selected by sampling the distribution of sample years that the true sequences were taken from. We then computed the tree statistics for each simulated outbreak. To compare the simulated sample distributions against the bootstrap statistic distributions computed from the real data, we used the two-sample Kolmogorov-Smirnov test and the relative difference of means. As these simulations were large, we refrained from reporting a p-value, but instead reported the test statistic as evidence for how close the simulated distribution were to the real data as the contact tracing parameter was varied. For both test statistics, a value closer to zero implies that the simulated distribution was closer to the real data.

Acknowledgement

This work was supported by the National Institutes of Health (NIH) grant R01AI087520 to TL.

References

Blum, M. G. B. and Tran, V. C. 2010. HIV with contact tracing: a case study in approximate Bayesian computation. *Biostatistics*, 11(4): 644–660.

Bouckaert, R., Heled, J., Kühnert, D., Vaughan, T., Wu, C.-H., Xie, D., Suchard, M. A., Rambaut, A., and Drummond, A. J. 2014. BEAST 2: A Software Platform for Bayesian Evolutionary Analysis. *PLOS Computational Biology*, 10(4): e1003537. Publisher: Public Library of Science.

Centers for Disease Control 1986. Additional Recommendations to Reduce Sexual and Drug Abuse-Related Transmission of Human T-Lymphotropic Virus Type III/Lymphadenopathy-Associated Virus. *Morbidity and mortality weekly report*, 35(10): 152–155.

Centers for Disease Control 1987. Public Health Service Guidelines for Counseling and Antibody Testing to Prevent HIV Infection and AIDS. *Morbidity and Mortality Weekly Report*, 36(31): 509–515. Publisher: Centers for Disease Control & Prevention (CDC).

Chen, W.-C. 2011. *Overlapping Codon Model, Phylogenetic Clustering, and Alternative Partial Expectation Conditional Maximization Algorithm*. Ph.D. thesis, Iowa State University.

Fischer, M. 2021. Extremal Values of the Sackin Tree Balance Index. *Annals of Combinatorics*, 25(2): 515–541.

Fischer, M., Herbst, L., Kersting, S., Kühn, L., and Wicke, K. 2021. Tree balance indices: a comprehensive survey. arXiv:2109.12281 [math, q-bio].

Gao, F., Robertson, D. L., Morrison, S. G., Hui, H., Craig, S., Decker, J., Fultz, P. N., Girard, M., Shaw, G. M., Hahn, B. H., and Sharp, P. M. 1996. The heterosexual human immunodeficiency virus type 1 epidemic in Thailand is caused by an intersubtype (A/E) recombinant of African origin. *Journal of Virology*, 70(10): 7013–7029.

Giardina, F., Romero-Severson, E. O., Albert, J., Britton, T., and Leitner, T. 2017. Inference of Transmission Network Structure from HIV Phylogenetic Trees. *PLOS Computational Biology*, 13(1): e1005316. Publisher: Public Library of Science.

Giesecke, J., Granath, F., Ramstedt, K., Ripa, T., Rådö, G., and Westrell, M. 1991. Efficacy of partner notification for HIV infection. *The Lancet*, 338(8775): 1096–1100.

Graw, F., Leitner, T., and Ribeiro, R. M. 2012. Agent-based and phylogenetic analyses reveal how HIV-1 moves between risk groups: Injecting drug users sustain the heterosexual epidemic in Latvia. *Epidemics*, 4(2): 104–116.

Guindon, S., Dufayard, J.-F., Lefort, V., Anisimova, M., Hordijk, W., and Gascuel, O. 2010. New Algorithms and Methods to Estimate Maximum-Likelihood Phylogenies: Assessing the Performance of PhyML 3.0. *Systematic Biology*, 59(3): 307–321.

Hemelaar, J., Elangovan, R., Yun, J., Dickson-Tetteh, L., Kirtley, S., Gouws-Williams, E., Ghys, P. D., and WHO-UNAIDS Network for HIV Isolation and Characterisation 2020. Global and regional epidemiology of HIV-1 recombinants in 1990-2015: a systematic review and global survey. *The lancet. HIV*, 7(11): e772–e781.

Hethcote, H. W. and Yorke, J. A. 1984. *Gonorrhea Transmission and Control*. Number 56 in Lecture Notes in Biomathematics. Springer-Velag, Berlin, Heidelberg.

Höhna, S., Stadler, T., Ronquist, F., and Britton, T. 2011. Inferring Speciation and Extinction Rates under Different Sampling Schemes. *Molecular Biology and Evolution*, 28(9): 2577–2589.

Hyman, J. M., Li, J., and Stanley, E. A. 2003. Modeling the impact of random screening and contact tracing in reducing the spread of HIV. *Mathematical Biosciences*, 181(1): 17–54.

Kirkpatrick, M. and Slatkin, M. 1993. Searching for Evolutionary Patterns in the Shape of a Phylogenetic Tree. *Evolution*, 47(4): 1171–1181. _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1111/j.1558-5646.1993.tb02144.x.

Kosakovsky Pond, S. L., Weaver, S., Leigh Brown, A. J., and Wertheim, J. O. 2018. HIV-TRACE (TRAnsmission Cluster Engine): a Tool for Large Scale Molecular Epidemiology of HIV-1 and Other Rapidly Evolving Pathogens. *Molecular Biology and Evolution*, 35(7): 1812–1819.

Kupperman, M. D., Leitner, T., and Ke, R. 2022. A deep learning approach to real-time HIV outbreak detection using genetic data. *PLOS Computational Biology*, 18(10): e1010598. Publisher: Public Library of Science.

Lee, C. Y. 1961. An Algorithm for Path Connections and Its Applications. *IRE Transactions on Electronic Computers*,

EC-10(3): 346–365. Conference Name: IRE Transactions on Electronic Computers.

Lundgren, E., Romero-Severson, E., Albert, J., and Leitner, T. 2022. Combining biomarker and virus phylogenetic models improves HIV-1 epidemiological source identification. *PLOS Computational Biology*, 18(8): e1009741. Publisher: Public Library of Science.

McCutchan, F. E., Artenstein, A. W., Sanders-Buell, E., Salminen, M. O., Carr, J. K., Mascola, J. R., Yu, X. F., Nelson, K. E., Khamboonruang, C., Schmitt, D., Kieny, M. P., McNeil, J. G., and Burke, D. S. 1996. Diversity of the envelope glycoprotein among human immunodeficiency virus type 1 isolates of clade E from Asia and Africa. *Journal of Virology*, 70(6): 3331–3338.

Moshiri, N., Ragonnet-Cronin, M., Wertheim, J. O., and Mirarab, S. 2019. FAVITES: simultaneous simulation of transmission networks, phylogenetic trees and sequences. *Bioinformatics*, 35(11): 1852–1861.

Müller, J. and Kretzschmar, M. 2021. Contact tracing – Old models and new challenges. *Infectious Disease Modelling*, 6: 222–231.

Paradis, E. and Schliep, K. 2019. ape 5.0: an environment for modern phylogenetics and evolutionary analyses in R. *Bioinformatics*, 35: 526–528.

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., Vanderplas, J., Passos, A., Cournapeau, D., Brucher, M., Perrot, M., and Duchesnay, E. 2011. Scikit-learn: Machine Learning in Python. *Journal of Machine Learning Research*, 12: 2825–2830.

Pickles, M., Cori, A., Probert, W. J. M., Sauter, R., Hinch, R., Fidler, S., Ayles, H., Bock, P., Donnell, D., Wilson, E., Piwowar-Manning, E., Floyd, S., Hayes, R. J., Fraser, C., and Team, H. . P. S. 2021. PopART-IBM, a highly efficient stochastic individual-based simulation model of generalised HIV epidemics developed in the context of the HPTN 071 (PopART) trial. *PLOS Computational Biology*, 17(9): e1009301. Publisher: Public Library of Science.

R Core Team 2022. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria.

Rambaut, A. and Grass, N. C. 1997. Seq-Gen: an application for the Monte Carlo simulation of DNA sequence evolution along phylogenetic trees. *Bioinformatics*, 13(3): 235–238.

Ramstedt, K., Hallhagen, G., Lundin, B. I., Håkansson, C., Johannisson, G., Löwhagen, G. B., Norkrans, G., and Giesecke, J. 1990. Contact tracing for human immunodeficiency virus (HIV) infection. *Sexually Transmitted Diseases*, 17(1): 37–41.

Romero-Severson, E., Skar, H., Bulla, I., Albert, J., and Leitner, T. 2014. Timing and Order of Transmission Events Is Not Directly Reflected in a Pathogen Phylogeny. *Molecular Biology and Evolution*, 31(9): 2472–2482.

Romero-Severson, E. O., Bulla, I., and Leitner, T. 2016. Phylogenetically resolving epidemiologic linkage. *Proceedings of the National Academy of Sciences*, 113(10): 2690–2695. Publisher: Proceedings of the National Academy of Sciences.

Sackin, M. J. 1972. "Good" and "Bad" Phenograms. *Systematic Biology*, 21(2): 225–226.

Schliep, K. P. 2011. phangorn: phylogenetic analysis in R. *Bioinformatics*, 27(4): 592–593.

Shao, K.-T. and Sokal, R. R. 1990. Tree Balance. *Systematic Biology*, 39(3): 266–276.

Skar, H., Axelsson, M., Berggren, I., Thalme, A., Gyllensten, K., Liitsola, K., Brummer-Korvenkontio, H., Kivelä, P., Spångberg, E., Leitner, T., and Albert, J. 2011. Dynamics of Two Separate but Linked HIV-1 CRF01_ae Outbreaks among Injection Drug Users in Stockholm, Sweden, and Helsinki, Finland. *Journal of Virology*, 85(1): 510–518. Publisher: American Society for Microbiology.

Turcinovic, J., Kuhfeldt, K., Sullivan, M., Landaverde, L., Platt, J. T., Doucette-Stamm, L., Hanage, W. P., Hamer, D. H., Klapperich, C., Landsberg, H. E., and Connor, J. H. 2022. Linking contact tracing with genomic surveillance to deconvolute SARS-CoV-2 transmission on a university campus. *iScience*, 25(11): 105337.