FAVITES: simultaneous simulation of transmission networks, phylogenetic trees, and sequences

Niema Moshiri[1], Manon Ragonnet-Cronin[2], Joel O. Wertheim[2], and Siavash Mirarab[3,*]

[1]Bioinformatics and Systems Biology Graduate Program, UC San Diego, La Jolla, 92093, USA
[2]Department of Medicine, UC San Diego, La Jolla, 92093, USA
[3]Department of Electrical and Computer Engineering, UC San Diego, La Jolla, 92093, USA

[*]To whom correspondence should be addressed.

April 17, 2018

Abstract

Motivation: The ability to simulate epidemics as a function of model parameters allows for the gain of insights unobtainable from real datasets. Further, reconstructing transmission networks for fast-evolving viruses like HIV may have the potential to greatly enhance epidemic intervention, but transmission network reconstruction methods have been inadequately studied, largely because it is difficult to obtain "truth" sets on which to test them and properly measure their performance.

Results: We introduce FAVITES, a robust framework for simulating realistic datasets for epidemics that are caused by fast-evolving pathogens like HIV. FAVITES creates a generative model that produces many items relevant to an epidemic, such as contact networks, transmission networks, phylogenetic trees, and error-prone sequence datasets. FAVITES is designed to be flexible and extensible by dividing the generative model into modules, each of which is expressed as a fixed API that can be implemented using various sub-models. We use FAVITES to simulate HIV datasets resembling the San Diego epidemic and show that the simulated datasets are realistic. We then use the simulated data to make inferences about the epidemic and how it is impacted by increased treatment efforts. We also study two transmission network reconstruction methods and their effectiveness in detecting fast-growing clusters.

Availability and implementation: FAVITES is available at https://github.com/niemasd/FAVITES, and a Docker image can be found on DockerHub (https://hub.docker.com/r/niemasd/favites).

1 Introduction

The spread of many infectious diseases is driven by social and sexual networks (Kelly *et al.*, 1991), and reconstructing their transmission histories from molecular data can greatly enhance intervention. For example, network-based statistics for measuring HIV treatment effects can yield increased statistical power (Wertheim *et al.*, 2011); the analysis of the growth of HIV infection clusters can yield actionable epidemiological information for disease control (Lewis *et al.*, 2008); transmission-aware models can be used to infer HIV evolutionary rates (Vrancken *et al.*, 2014).

A series of events in which an infected individual infects another individual can be shown as a *transmission network*, which itself is a subset of a *contact network*, a graph in which nodes represent individuals and edges represent contacts (e.g. sexual) between pairs of individuals. If the pathogens of the infected individuals are sequenced (e.g. the standard of HIV care in many developed countries), one can attempt to reconstruct the transmission network (or its main features) using molecular data. What gives us hope to reconstruct the network is that some viruses, such as HIV, evolve quickly, and the phylogenetic relationships between viruses are reflective of transmission histories (Leitner *et al.*, 1996), albeit imperfectly (Ypma *et al.*, 2013; Romero-Severson *et al.*, 2014).

Recently, multiple methods have been developed to infer properties of transmission networks from molecular data (e.g. Prosperi *et al.*, 2011; Ragonnet-Cronin *et al.*, 2013; Pond *et al.*, 2018; Mccloskey and Poon, 2017). Efforts have been made to characterize and understand the promise and limitations of these methods. It is suggested that, when combined with clinical and epidemiological data, these methods can provide critical information about drug resistance, associations between sociodemographic characteristics, viral spread within populations, and the time scales over which viral epidemics occur (Grabowski and Redd, 2014). More recently, these methods have become widely used at both local (Campbell *et al.*, 2017) and global scale (Wertheim *et al.*, 2014). Nevertheless, several questions remain to be fully answered regarding the performance of these methods. It is not always clear which method/setting combination performs best for a specific downstream use-case or for specific epidemiological conditions. More broadly, the effectiveness of these methods in helping achieve public health goals is the subject of ongoing clinical and theoretical research.

Transmission networks are difficult to study because controlling parameters of interest such as network shape and transmission rates is not possible. A relatively inexpensive method to investigate questions related to epidemics is via simulation (Villandre *et al.*, 2016). Any simulation of transmission networks needs to combine models of social network, transmission, evolution, and ideally sampling biases and errors. One attempt to build such a simulation tool is `PANGEA.HIV.sim`, an R package developed by the PANGEA-HIV consortium to simulate realistic HIV transmission dynamics, phylogenetic trees, and sequence data (Ratmann *et al.*, 2017). While the `PANGEA.HIV.sim` workflow allows for generality in terms of model parameters, it is restrictive in that the statistical models at each step of the workflow are fixed.

We introduce FAVITES (FrAmework for VIral Transmission and Evolution Simulation), which can simulate numerous models of contact networks, viral transmission, phylogenetic and sequence evolution, data (sub)sampling, and real-world data perturbations, and which was built to be flexible such that users can seamlessly plug in statistical models and model parameters at every step of the simulation process. We show the realism of FAVITES in a series of experiments, study the properties of HIV epidemics as functions of various model and parameter choices, and finally perform simulation experiments to study two transmission network reconstruction methods.

2 Materials and methods

2.1 FAVITES simulation process

FAVITES provides a general workflow for the simulation of viral transmission networks, phylogenetic trees, and sequence data. It breaks the simulation process down into a series of interactions between abstract modules, and users can select the module implementations appropriate to their specific context. In a statistical sense, the end-to-end process creates a complex composite generative model, each module is a template for a sub-model of a larger model, and different implementations of each module correspond to different statistical sub-models. FAVITES is designed to be flexible for developers, a goal achieved by defining APIs for each module and allowing various forms of interaction between modules. These interactions enable sub-models that are described as conditional distributions (via dependence on preceding steps) or as joint distributions (via joint implementation). Module implementations can simply wrap around existing tools, allowing for significant code reuse. To emphasize this, we even wrap around `PANGEA.HIV.sim` (Ratmann *et al.*, 2017).

Simulations start at time zero and continue until a user-specified end criterion is met. Error-free and error-prone transmission networks, phylogenetic trees, and sequences are output at the end. FAVITES has eight steps (Fig. 1), which we describe below with examples of canonical models implemented for each step.

Step 1. The *ContactNetworkGenerator* module generates a contact network; vertices represent individuals, and edges represent contacts between them that can lead to disease transmission (e.g. sexual). Graphs can be created stochastically using existing models (Karoński, 1982), including those that capture properties of real social networks (Watts and Strogatz, 1998; Watts, 1999; Newman *et al.*, 2002). For example, the Erdős-Rényi (ER) model (Erdos and Rényi, 1960) generates graphs with randomly-placed edges, the Random Partition model (Fortunato, 2010) generates communities, the Barabási-Albert model (Barabási and Albert, 1999) generates scale-free networks whose degree distributions follow power-law (suitable for social and sexual contact networks), the Caveman model (Watts, 1999) and its variations (Fortunato, 2010) generate small-world networks, the Watts-Strogatz model (Watts and Strogatz, 1998) generates small-world networks with short average path lengths, and Complete graphs connect all pairs of individuals (suitable for some communicable diseases). We currently have all these models implemented by wrapping around the NetworkX package (Hagberg *et al.*, 2008). In addition, a user-specified network can be used.

Step 2. The transmission network is initialized in two steps. *a*) The *SeedSelection* module chooses the "seed" nodes: individuals who are infected at time zero of the simulation. *b*) For each selected seed node, the *SeedSequence* module generates an initial viral sequence.

Example implemented models for selecting seeds include Random selection (i.e., uniformly at random) and Edge-Weighted (each node's probability of being selected is directly proportional to its degree). We also designed a Clusters-Bernoulli model: randomly select k seed individuals with equal probability to initiate k "clusters," and for each of the

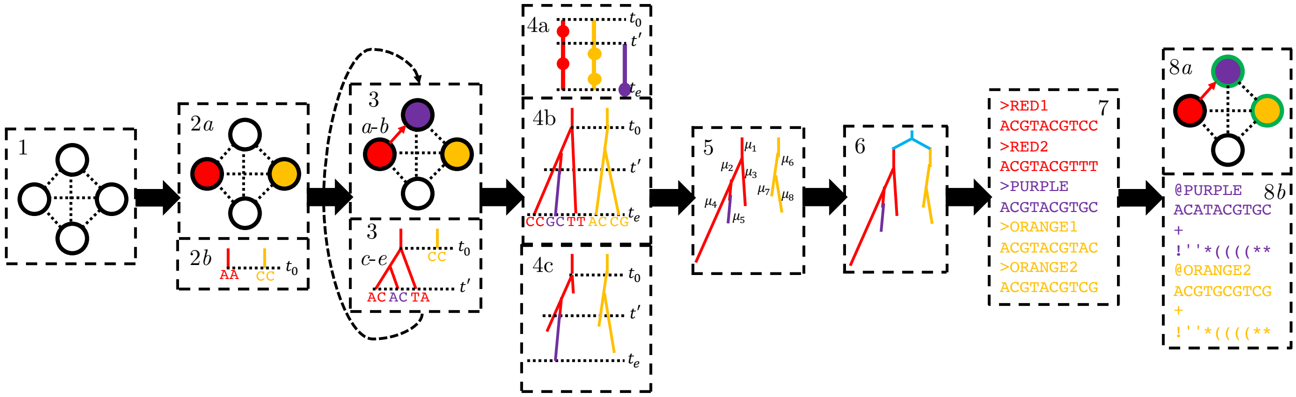


Figure 1: FAVITES workflow. (1) The contact network is generated (nodes: individuals; edges: contacts). (2) *Seed* individuals who are infected at time 0 are selected (2a), and a viral sequence is chosen for each (2b). (3) The epidemic yields a series of transmission events in which the time of the next transmission is chosen (3a), the source and target individuals are chosen (3b), the viral phylogeny in the source node is evolved to the transmission time (3c), viral sequences in the source node are evolved to the transmission time (3d), and a viral lineage is chosen to be transmitted from source to destination (3e). Step (3) repeats until the end criterion is met. Step 3c–3e are optional, as tree and sequence generation can be delayed to later steps. (4) Infected individuals are sampled such that viral sequencing times are chosen for each infected individual (4a), viral phylogenies (one per seed) are evolved to the end time of the simulation (4b), and viral phylogenies (one per seed) are pruned to reflect the viral sequencing times selected (4c). (5) Mutation rates are introduced along the branches of the viral phylogenies and the tree is scaled to the unit of expected mutations. (6) The seed trees are merged using a seed tree (cyan). (7) Viral sequences obtained from each infected individual are finalized. (8) Real-world errors are introduced on the error-free data, such as subsampling of the sequenced individuals (marked as green) (8a) and the introduction of sequencing errors (8b).

k initial individuals, perform a random walk, flipping a coin with probability p at each step to determine if the current individual will be a seed, until m/k seed individuals have been chosen in the current "cluster".

Many models can be used to generate the seed sequence, including random selection. However, seed sequences should ideally emulate the virus of interest. To accomplish this, we implement a model where we use HMMER (Eddy, 1998) to sample each seed sequence from a profile Hidden Markov Model (HMM) specific to the virus of interest. We provide a set of such prebuilt profile HMMs constructed from multiple sequence alignments (MSAs) of viral sequences (currently just for HIV). Our prebuilt HMMs are for the *pol* gene and come from the Los Alamos National Laboratory (LANL) and from a San Diego dataset of HIV-1 subtype B sequences (Little *et al.*, 2014).

When multiple seeds are chosen, we need to model their phylogenetic relationship as well. Thus, we also have a model that samples a *single* sequence from a viral profile HMM using HMMER, simulates a *seed tree* with a single leaf per seed individual (e.g. using Kingman coalescent or birth-death models using DendroPy (Sukumaran and Holder, 2010)), and then evolves the viral sequence down the tree to generate seed sequences using Seq-Gen (Rambaut and Grass, 1997).

Step 3. An iterative series of transmission events occurs under a stochastic transmission model until the end of the simulation, as dictated by the *EndCriteria* module. Each event has five components.

a) The *TransmissionTimeSample* module chooses the time at which the next transmission event will occur and advances the *current* time accordingly, and *b*) the *TransmissionNodeSample* module chooses a source node and target node to be involved in the next transmission event. These two modules are often jointly implemented. Beyond simple models (e.g. selecting times by draws from an exponential and selecting nodes uniformly at random), the main models of transmission treat individuals as Markov processes in which individuals start in some state (e.g. Susceptible) and transition between states of the model (e.g. Infected, Recovered, etc.) over time. These epidemiological models are defined by two sets of transition rates: "nodal" and "edge-based." Nodal transition rates are rates that are independent of interactions with neighbors (e.g. the transition rate from Infected to Recovered in the SIR model), whereas edge-based transition rates are the rate of transitioning from one state to another given that a single neighbor is in a given state (e.g. the transition rate from Susceptible to Infected given that a neighbor is Infected). The rate at which a specific node u transitions from state a to state b is the nodal transition rate from a to b plus the sum of the edge-based transition rate from a to b given neighbor v's state for all neighbors v. We use GEMF (Sahneh *et al.*, 2017) to implement many epidemiological models in this manner, such as Susceptible-Infected (SI), Susceptible-Infected-Susceptible (SIS), Susceptible-Aware-Infected (SAI), Susceptible-Infected-Recovered (SIR), Susceptible-Exposed-Infected-Recovered (SEIR), and Susceptible-Vaccinated-Infected-Treated-Recovered (SVITR). We have also included more sophisticated HIV models, such as the Granich *et al.* (2009) model as well the HPTN 071 (PopART) model (Cori *et al.*, 2014).

c) The *NodeEvolution* module evolves viral phylogenies of the source node to the current time. Many stochastic models of tree evolution have been developed (Hartmann *et al.*, 2010), some simulating forward in time (i.e., starting from the root and generating the tree top-down), and others backward (i.e., starting from leaves and generating back in time, typically based on coalescent theory). For forward models, we wrap around DendroPy for birth-death (Sukumaran and Holder, 2010) and include our own implementation of dual-birth (Moshiri and Mirarab, 2017) and thus Yule. For backward models, we wrap around VirusTreeSimulator (Ratmann *et al.*, 2017) for coalescent models with constant, exponentially-growing, or logistically-growing population size.

d) The *SequenceEvolution* module evolves all viral sequences in the source node to the current time. A commonly-used model for DNA is the General Time-Reversible (GTR) model (Tavaré, 1986), parameterized by nucleotide frequencies and base change rates, with constraints to enforce time-reversibility. Other commonly-used DNA models, e.g. Jukes and Cantor (1969), Kimura (1980), Felsenstein (1981), and Tamura and Nei (1993), are reductions of the GTR model and are thus all also currently available in FAVITES. An extension of the GTR model available in FAVITES is the GTR+Γ model, which incorporates rates-across-sites variation (Yang, 1994). For coding sequences, in which selection occurs at the encoded amino acid level as well, FAVITES currently includes multiple codon-aware extensions of the GTR model, such as mechanistic (Zaheri *et al.*, 2014) and empirical (Kosiol *et al.*, 2007) codon models. Our current implementations internally use Seq-Gen (Rambaut and Grass, 1997) and Pyvolve (Spielman and Wilke, 2015).

e) The *SourceSample* module chooses the viral lineage(s) in the source node to be transmitted.

Substeps $c - e$ are required only if the choice of transmission events after time t depends on the past phylogeny or sequences up to time t. If the choice of future transmission recipients/donors and transmission times are agnostic to past phylogenies and sequences, these can be omitted.

Step 4. The patient sampling (i.e., sequencing) events are determined and phylogenetic trees are updated accordingly. Three sub-steps are involved.

a) For each individual, the *NumTimeSample* module chooses the number of times it will be sequenced, the *TimeSample* module chooses the corresponding sequencing time(s), and the *NumBranchSample* module chooses how many viral lineages will be sampled at each sampling time. A given individual may not be sampled at all, thus simulating incomplete epidemiological sampling efforts. Current implementations include sampling each individual a fixed number of times or a random number of times by draws from a Poisson distribution. Sampling times can be fixed at the end time of the transmission simulation or can be uniformly distributed across the entirety of the individual's span of infection. Individuals can also be sampled the first time they enter a specific state of the transmission model. This model is appropriate when the transmission model includes a state for treatment because the standard of the HIV care in many places is to sequence individuals before the start of antiretroviral therapy. Another model is to draw the sample times from a user-parameterized Truncated Normal distribution over the window(s) of the individual's span of infection to recovery/treatment time(s).

b) The *NodeEvolution* module is called to simulate the phylogenetic trees *given sampling times*. This step can be used *instead of Step 3c* to evolve only lineages that are sampled, thereby reducing the dataset size. In particular, if the tree simulation model is backwards (e.g. coalescent models), *Step 3c* should be ignored, and the full backward simulation process can be performed at once here. *c*) If the tree is simulated in *Step 3c*, it may need to be pruned to only include lineages that are sampled. At this point, we perform the pruning. Sampled lineages at a given time are chosen uniformly at random, but other models can be implemented.

Step 5. The phylogenetic trees from *Step 4* are in unit of time. To generate sequence data, rates of evolution need to be assumed. The *TreeUnit* module can be used to determine such rates, yielding a tree in unit of per-site expected number of mutations. In current implementations, the mutation rate can be a constant (i.e., all branches are multiplied by a user-given constant), or it can be a constant multiplied by random draws from several commonly-used distributions (e.g. Exponential, Gamma, or Log-Normal).

Step 6. We now have one tree per seed individual. Some implementations of *SeedSequence* use a tree to simulate seed sequences, so the roots of the trees have a phylogenetic relationship. In this case, another call to the *SeedSequence* module in this step handles merging the individual phylogenetic trees into a single global tree by placing each individual tree's root at its corresponding leaf in the seed tree (Fig. 1).

Step 7. A second invocation of the *SequenceEvolution* module is used to finalize the sequences. Two scenarios are possible. If sequences are simulated continuously in *Step 3d*, this step is used to evolve all sequences between the last transmission time and the sampling times. More importantly, to reduce the dataset size and to speed up simulation, when the transmissions and tree evolution are not dependent on the exact sequences, *Step 3d* can be skipped (i.e., a dummy module), and the sequence evolution can be delayed until this point. Here, the *SequenceEvolution* module can perform the full sequence simulation on the final tree(s) at once.

Step 8. Error-free data are now at hand. Noise is introduced onto the complete error-free data in two ways. *a*) The *NodeAvailability* module further subsamples the individuals to simulate lack of accessibility to certain datasets. Note, therefore, that whether a node is sampled is a function of two different modules: *NodeAvailability* and *NumTimeSample* (if *NumTimeSample* returned 0, the individual is not sampled). Conceptually, *NumTimeSample* can be used to model when people are sequenced, while *NodeAvailability* can be used to model patterns of data availability (e.g. sharing of data between clinics). Bernoulli sampling with a fixed probability as well as randomly choosing individuals with sampling probability weighted by the number of transmissions in which the individual was involved are example models currently implemented. *b*) The *Sequencing* module simulates sequencing error on the simulated sequences. In addition to sequencing machine errors, this can incorporate other real-world sequencing issues, e.g. taking the consensus sequence of a sample and introducing of ambiguous characters. The current sequencing machine error models implemented include wrappers around ART (Huang *et al.*, 2012) for simulating Illumina, Roche 454, and SOLiD sequencing error, DWGSIM for simulating Illumina, SOLiD, and Ion Torrent sequencing error, and Grinder (Angly *et al.*, 2012) for simulating Sanger sequencing error, including support for ambiguous characters.

Validation. We provide tools to validate FAVITES outputs (Table S1), which compare real networks, phylogenetic trees, or sequence data to the simulation results. For contact networks, the comparison can be in terms of the average and standard deviation of node degree distributions and the Kolmogorov-Smirnov (KS) test (Smirnov, 1939). For phylogenetic trees, we compare terminal and internal branch length distributions between real and simulated using summary statistics and the KS test. If the user has a Multiple Sequence Alignment (MSA) from real data, a profile Hidden Markov Model (HMM) can be built from the alignment. Then, the simulated sequences can be aligned against the profile HMM using `hmmscan` (Eddy, 1998), and bit-scores can be examined. In addition to post-validation scripts, we have several helper scripts to implement tasks that are likely common to downstream use of FAVITES output (Table S2).

2.2 Experimental setup

We perform a large simulation study of HIV phylodynamics using FAVITES while employing several generative models (datasets available at https://gitlab.com/niemasd/favites-paper). Besides demonstrating its flexibility, we present evidence that the data generated by FAVITES are reasonably similar to real HIV data. We then study properties of the epidemic as a function of the parameters of the underlying generative models. Finally, we compare two transmission cluster inference tools when applied to sequence data generated by FAVITES.

2.2.1 The simulation model

We selected a set of "base" simulation models and parameters and performed experiments in which they were varied. For each parameter set, we ran 10 simulation replicates. The base simulation parameters were chosen to emulate (as much as possible) HIV transmission in San Diego from 2005 to 2014. We start with base parameters.

Contact network. The contact network includes 100,000 individuals to approximate the at-risk community of San Diego. We set the base expected degree ($\mathbb{E}_d$) to 4 edges (i.e., sexual partners over 10 years) per individual. This number is motivated by estimates from the literature (e.g. ≈3 in Wertheim *et al.* (2017b) and 3–4 in Rosenberg *et al.* (2011)), and it is varied in the experiments. We chose the Barabási-Albert (BA) model as the base network model because it can generate power-law degree distributions (Barabási and Albert, 1999), a property commonly assumed of sexual networks (Hamilton *et al.*, 2008).

Seeds. We chose 15,000 total infected seed individuals based on the estimate of total HIV cases in San Diego as of 2004 (Shepard *et al.*, 2005). In the base model, we choose seed individuals uniformly at random.

Epidemiological model. We model HIV transmission as a Markov chain epidemic model (see Section 2.1), with states Susceptible (S), Acute HIV Untreated (AU), Acute HIV Treated with ART (AT), Chronic HIV Untreated (CU), and Chronic HIV Treated with ART (CT). All seed individuals in the AU state and transmissions occur with fixes rates (Fig. 2). Note that this model is a simplification of the model used by Granich *et al.* (2009), which includes eight states of infection (four untreated and four treated) as opposed to our four states (two untreated and two treated).

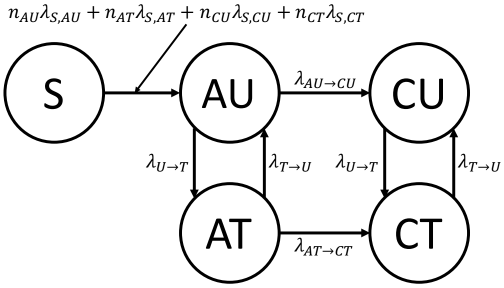


Figure 2: Epidemiological model of HIV transmission with states Susceptible (S), Acute HIV Untreated (AU), Acute HIV Treated with ART (AT), Chronic HIV Untreated (CU), and Chronic HIV Treated with ART (CT). The model is parameterized by the following rates: infectiousness of AU ($\lambda_{S,AU}$), AT ($\lambda_{S,AT}$), CU ($\lambda_{S,CU}$), CT ($\lambda_{S,CT}$) individuals, and by rate to transition from AU to CU ($\lambda_{AU\to CU}$), rate to transition from AT to CT ($\lambda_{AT\to CT}$), rate to start ART ($\lambda_{U\to T}$), and rate to stop ART ($\lambda_{U\to T}$).

Parameter	Parameter Values
Contact Network Model	**Barabási-Albert**, Erdős-Rényi, Watts-Strogatz
Expected Degree($\mathbb{E}_d$)	2, **4**, 8, 16
Seed Selection	**Random**, Edge-Weighted
Mean time to ART ($\mathbb{E}_{ART}$)	⅛, ¼, ½, **1**, 2, 4, 8 (years)

Table 1: Simulation parameters (base parameters in bold)

We set $\lambda_{AU\to CU}$ such that the expected time to transition from AU to CU is 6 weeks (Bellan *et al.*, 2015) and set $\lambda_{AT\to CT}$ such that the expected time to transition from AT to CT is 12 weeks (Cohen *et al.*, 2011). We set $\lambda_{U\to T}$ such that the expected time to start ART is 1 year from initial infection (O'Brien and Markowitz, 2012), and we define $\mathbb{E}_{ART}= 1/\lambda_{U\to T}$. We set $\lambda_{T\to U}$ such that the expected time to stop ART is 25 months from initial treatment (Nosyk *et al.*, 2015). For the rates of infection $\lambda_{S,j}$ for $j \in \{AU, CU, AT, CT\}$, using the infectiousness of CU individuals as a baseline, we set the parameters such that AU individuals are 5 times as infectious (Wawer *et al.*, 2005) and CT individuals are not infectious (i.e., rate of 0). Cohen *et al.* (2011) found a 0.04 hazard ratio when comparing linked HIV transmissions between an early-therapy group and a late-therapy group, so we estimated AT individuals to be 1/20 the infectiousness of CU individuals. We then scaled these relative rates so that the total number of new cases over the span of the 10 years was roughly 6,000 (Macchione *et al.*, 2015), yielding transmission rate from acute untreated set to $\lambda_{S,AU} = 0.1125$ per year.

Phylogeny. A single viral lineage from each individual was assumed to be sampled at the end time of the epidemic simulation (10 years). The viral phylogenetic tree in unit of time (years) was then sampled under a coalescent model in which the viral population in an individual undergoes logistic growth using the same approach as the the PANGEA-HIV methods comparison exercise, setting the initial population to 1, per-year growth rate to 2.851904, and the time back from present at which the population is at half the carrying capacity (`v.T50`) to -2 (Ratmann *et al.*, 2017). Each seed individual is the root of an independent viral phylogenetic tree, and these phylogenetic trees were merged by simulating a seed tree under the pure neutral Kingman coalescent model with one leaf per seed node and an expected height of approximately 40 years to reflect the origin of HIV in the USA in the 1970s (Worobey *et al.*, 2016) using DendroPy (Sukumaran and Holder, 2010). The phylogenetic tree was then scaled from unit of time (years) to unit of expected number of mutations by multiplying each branch length by an evolutionary rate sampled from a log-normal random variable with $\mu = -6.164$ and $\sigma = 0.3$ (Fig. S1) (Ratmann *et al.*, 2017).

Sequence data. We sampled a root sequence from a profile Hidden Markov Model (HMM) generated from 639 HIV-1 subtype B *pol* sequences from San Diego (Little *et al.*, 2014). We evolved down the scaled viral phylogenetic tree under the GTR+Γ model using Seq-Gen (Rambaut and Grass, 1997) with parameters inferred using RAxML (Stamatakis, 2014) from the same San Diego sequence dataset (Table S3).

Varying parameters. We thoroughly explore four parameters (Table 1). For the contact network, in addition to the BA model, we used the Erdős-Rényi (ER) (Erdos and Rényi, 1960) and Watts-Strogatz (WS) (Watts and Strogatz, 1998) models. We also varied the expected degree ($\mathbb{E}_d$) of individuals in the contact network between 2 and 16 (Table 1). For the method of seed selection, we also used "Edge-Weighted," where the probability that an individual is chosen is weighted by the individual's degree. For each selection of contact network model, $\mathbb{E}_d$, and seed selection method, we study multiple rates of starting ART (expressed as $\mathbb{E}_{ART}$). In our discussions, we focus on $\mathbb{E}_{ART}$, a factor that the public health departments can try to impact. Increased effort in testing at-risk populations can decrease the diagnosis time, and the increased diagnosis rate coupled with high standards of care can lead to faster ART initiation. Behavioral intervention could in principle also impact degree distribution, another factor that we vary, but the extent of the effectiveness of behavioral interventions is unclear (Kelly *et al.*, 1991).

2.2.2 Transmission network reconstruction methods

We compare two HIV network inference tools: HIV-TRACE (Pond *et al.*, 2018) and TreeCluster (Moshiri, 2018). HIV-TRACE is a widely-used method (Rose *et al.*, 2017; Wertheim *et al.*, 2017b; Pérez-Losada *et al.*, 2017). Under its default settings, HIV-TRACE clusters individuals such that, for all pairs of individuals u and v, if the Tamura and Nei (1993) (TN93) distance is at most 1.5%, u and v are connected by an edge; each connected component then forms a cluster. We ran HIV-TRACE on the simulation experiment data using its default settings, skipping the alignment step because the simulated sequences did not contain indels. TreeCluster clusters the leaves of a given tree such that the pairwise path length between any two leaves in the same cluster is at most 0.045 (the default threshold), the members of a cluster define a full clade, and the number of clusters is minimized. Trees given to TreeCluster were inferred using FastTree-II (Price *et al.*, 2010) under the GTR+Γ model. TreeCluster is similar in idea to Cluster Picker (Ragonnet-Cronin *et al.*, 2013), which uses sequence distances instead of tree distances. Cluster Picker can infer clusters using bootstrap support in addition to distance, a feature that TreeCluster also supports, but to avoid the time-consuming bootstrapping step, we do not employ it here. We study TreeCluster instead of Cluster Picker because of its improved speed. Our attempts to run PhyloPart (Prosperi *et al.*, 2011) were unsuccessful due to running time.

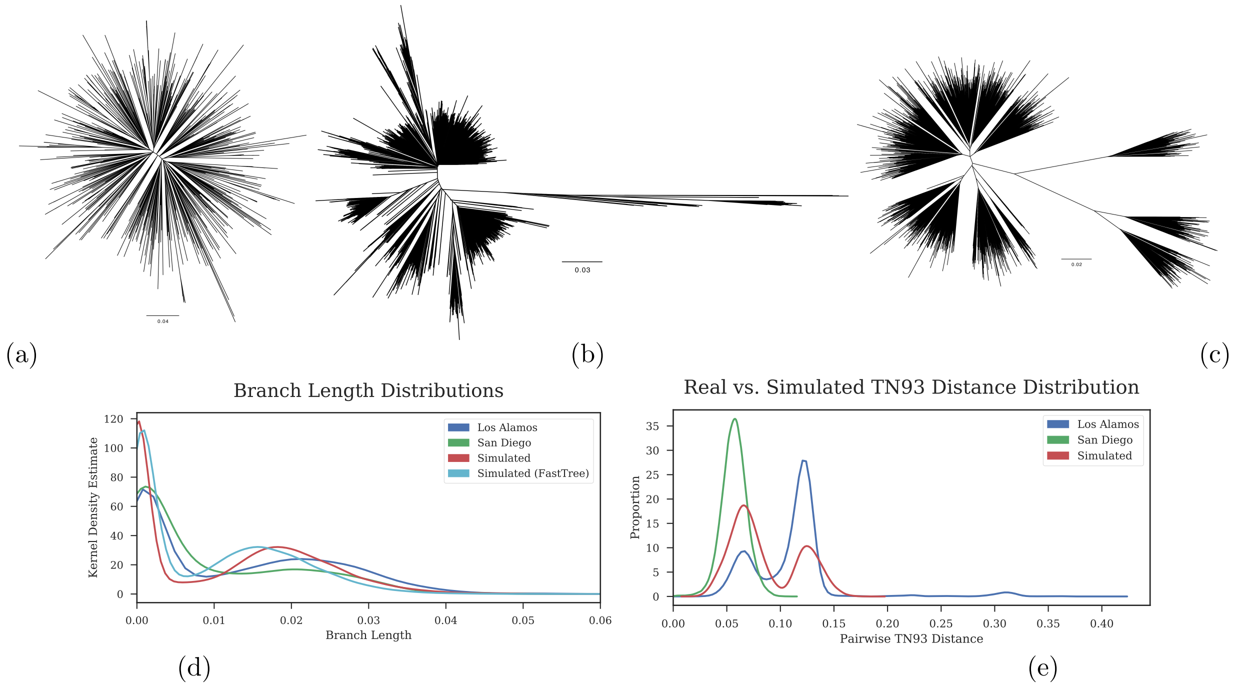
2.2.3 Measuring the predictive power of clustering methods

We now have two sets of clusters at the end of the simulation process (year 10): one produced by HIV-TRACE and one by TreeCluster. Let C^t denote the clustering resulting from removing all individuals infected after year t from a given final clustering C^{10}, let C_i^t denote a single i-th cluster in clustering C^t, and let $g(C_i^t) = \frac{|C_i^t|-|C_i^{t-1}|}{\sqrt{|C_i^t|}}$ denote the growth rate of a given cluster C_i^t; the square root normalization is based on a practice used on real data (Wertheim *et al.*, 2017a). We then compute the average number of individuals who were infected between years 9 and 10 by the "top" 1,000 individuals who were infected at year 9, where we choose top individuals by sorting the clusters in C^9 in descending order of $g(C_i^9)$ (breaking ties randomly) and choosing 1,000 individuals in this sorting, breaking ties in a given cluster randomly if needed (e.g. for the last cluster needed to reach 1,000 individuals). As a baseline, we compute the average number of individuals who were infected between years 9 and 10 by *all* individuals, which is equivalent (in expectation) to a random selection of 1,000 individuals. Our metric, therefore, measures the risk of transmission from the top selected 1,000 individuals (roughly 5% of the total infected population). Our motivation for this metric is to capture whether monitoring cluster growth can help public health intervention efforts with limited resources (hence our limitation on the number of top individuals) in finding individuals with a higher risk of transmitting.

3 Results

3.1 Comparison of simulated and real data

Phylogenetic trees simulated using the base parameters resemble phylogenetic trees inferred from real sequence data under the GTR+Γ model using FastTree 2 (Price *et al.*, 2010), both in terms of topology and branch length (Fig. 3). Reassuringly, our simulated trees, like real trees, include clusters of long terminal branches and short internal branches, with clusters connected via moderately long branches (Fig. 3a–c). Further, the branch length distributions are similar to real data (Fig. 3d), showing a bimodal distribution of short and long branches. The simulated branches are on average slightly shorter than real data, a pattern we do not find surprising given the fact that simulations, unlike real data, enjoy full sampling. Arguably, more important than the branch length distribution is the pairwise sequence distance distribution. Pairwise distances according to the TN93 measure are different between the LANL and San Diego datasets. The simulated distances have distributions that are close to the San Diego dataset, with a second mode that resembles the LANL dataset (Fig. 3e). Overall, both the phylogeny and sequence data seem realistic.



(a) (b) (c)

(d) (e)

Figure 3: Real versus simulated phylogenetic trees and sequence data. Phylogenetic trees inferred from real HIV multiple sequence alignments from San Diego (Little *et al.*, 2014) (a) and the Los Alamos National Laboratory (b) under the GTR+Γ model using FastTree 2 (Price *et al.*, 2010), and a tree simulated by FAVITES using base parameters (c). Kernel density estimates of the branch lengths of the trees (d) and of the pairwise Tamura-Nei 93 (TN93) (Tamura and Nei, 1993) distances between sequences (e). For the simulated pairwise distances in (e), 639 simulated sequences were randomly chosen from the full dataset to represent the subsampling of the San Diego dataset, which also contains 639 sequences.

3.2 Analyzing transmission network properties

As expected, the number of infected individuals increases with time, and the rate of growth is faster for larger $\mathbb{E}_{ART}$ values (Fig. 4a). Interestingly, for all tested values of $\mathbb{E}_{ART}$, the growth of the number of infected individuals is close to linear, indicating that the large at-risk population that we simulated has not saturated in the 10 year simulation period. However, this pattern could not infinitely continue had we simulated beyond 10 years because the number of remaining susceptible people constantly drops. However, in simulations like ours where the network is not dynamically changing, it makes sense to ensure the at-risk population is not exhausted during the simulation period.

The percentage of the infected population on ART at any point in time is also interesting. For example, the 90-90-90 campaign by UNAIDS (2017) aims to have 90% of HIV infected individuals diagnosed, of which 90% should receive treatment, of which 90% (i.e., 72.9% of total) should adhere and be virally suppressed. We observed that the ratio of individuals in the treated and untreated states remains constant (Fig. 4) after an initial time period of roughly a year (initial instability is because we start all individuals in the untreated state; Fig. S2). With $\mathbb{E}_{ART}$= 2 years, we have roughly as many treated people as untreated; decreasing $\mathbb{E}_{ART}$ predictably reduces the portion of untreated people. According to our simulations, reaching the 90-90-90 goals in an epidemic like that of San Diego requires $\mathbb{E}_{ART}$ to be between $^1/_2$ and 1 year (assuming that a lack of viral suppression for treated people is fully attributed to the lack of adherence).

As $\mathbb{E}_{ART}$ decreases, the total number of infected individuals at the end of the simulation time decreases (Fig. 5a). For our parameter set, reducing $\mathbb{E}_{ART}$ from 8 years all the way to $^1/_8$ years keeps reducing the final number of infected people. For example, with degree 4, the average final number of newly infected individuals in the 10 year period is 6686, 4134, and 1273 with $\mathbb{E}_{ART}$ set to 1, $^1/_2$, $^1/_8$ year, respectively.

Beyond the total number of infections, patterns of branch lengths in the true phylogeny also change. The average branch length of the true phylogenetic tree increases when $\mathbb{E}_{ART}$ decreases until it reaches a point of saturation (Fig. 5b). As expected, higher $\mathbb{E}_d$ (which yields faster transmissions) also results in shorter average branch length (Fig. 5b). Note that the same mean branch length could be obtained by various settings of the network degree and the time to ART.

The shortest branches in a phylogeny are generally the most difficult to infer, and thus, we can hope that decreasing $\mathbb{E}_{ART}$ and $\mathbb{E}_d$ coincide with reduced phylogenetic inference error. To test this expectation, we computed the normalized Robinson-Foulds (RF) distance (i.e., the proportion of branches included in one tree but not the other (Robinson and Foulds, 1981)) between the true tree and the estimated tree. For all model conditions, the RF distance is quite high (0.50-0.65). As we hoped, for networks with $\mathbb{E}_d \leq 4$, as $\mathbb{E}_{ART}$ decreases, the RF distance of trees inferred using FastTree (Price *et al.*, 2010) under the GTR+Γ model decreases (Fig. 5c). However, unexpectedly, for networks with $\mathbb{E}_d \geq 8$, as $\mathbb{E}_{ART}$ decreases, the RF distance initially rises and then falls. To further study the cause of this pattern, we analyzed the proportion of extremely short branches versus $\mathbb{E}_{ART}$. For our analysis, we define branches to be "extremely short" if the expected number of mutations along the branch across the entire sequence is less than or equal to 1. As $\mathbb{E}_{ART}$ decreases, the proportion of extremely short branches *increases* (making phylogenetic inference more difficult) despite the fact that the average branch length simultaneously increases (making phylogenetic inference easier). Thus, perhaps, the exact patterns of the fall and rise of RF as a function of decreasing $\mathbb{E}_{ART}$ is the outcome of these two competing factors: increased short branches but also higher average branch lengths.

Surprisingly, the model of contact network and the model of choosing the seed individuals only had marginal effects on epidemiological outcomes. Edge-weighting the seed selections yields a slightly higher total number of infected individuals than the random selection (Figs. S3). The BA model of contact network leads to a slightly higher infection count when

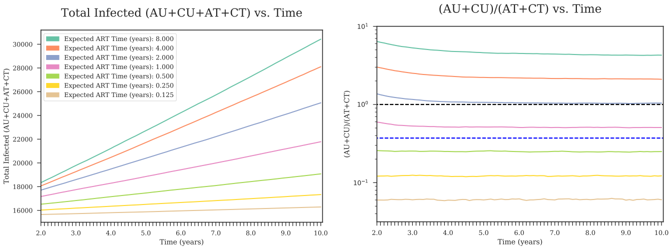


Figure 4: (Left) The total number of infected individuals and (Right) the ratio of the number of untreated vs. the number of treated individuals (log-scale) vs. time for the Barabási-Albert model with various values of expected time to begin ART ($\mathbb{E}_{ART}$, colors) with all other parameters set to base values. In the right, untreated/treated = 1 is shown as a dashed black line, and the value of untreated/treated corresponding to the "90-90-90" goal (UNAIDS, 2017) is shown as a dashed blue line ($^{(1-0.9^3)}/_{0.9^3} \approx 0.37$). We start from year 2 as dynamics are initially unstable because we initialize the transmission network with all infected individuals in the Acute Untreated (AU) state. See Figure S2 for full plots.

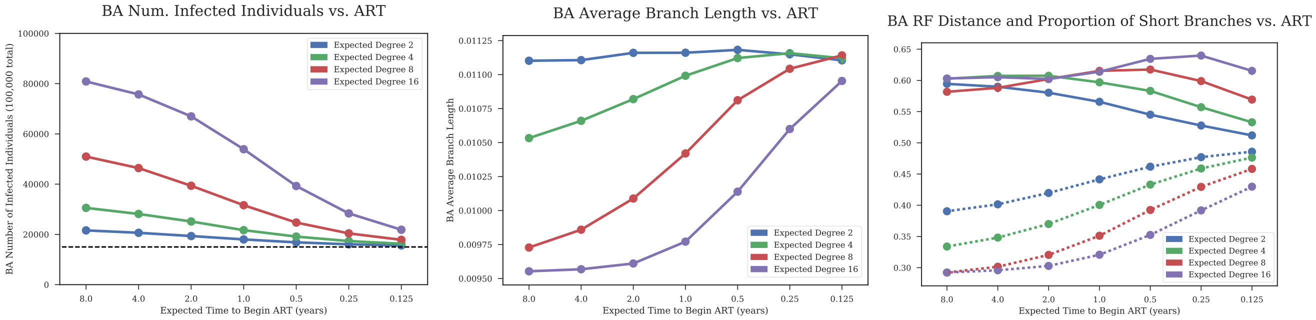


Figure 5: (Left) Total number of infected individuals, (Center) average branch length, and (Right) FastTree RF distance (solid lines) and proportion of "extremely short" branches (dotted lines) vs. expected time to begin ART ($\mathbb{E}_{ART}$) for the Barabási-Albert with various $\mathbb{E}_d$ values (colors) with all other parameters set to base values. In the leftmost plot, the number of seed individuals (15,000) is shown by a black dashed line. In the rightmost plot, we define branches to be "extremely short" if the expected number of mutations along the branch is less than or equal to 1 (i.e., the branch length is less than or equal to the reciprocal of the sequence length).

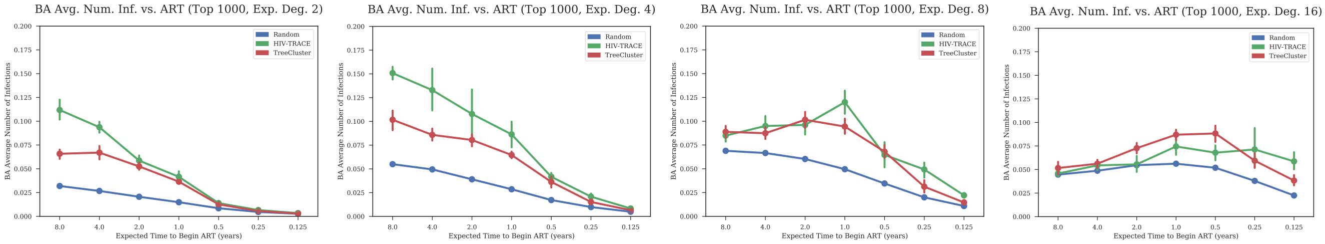


Figure 6: The effectiveness of clustering methods in finding high risk individuals. The average number of new infections between years 9 and 10 of the simulation caused by individuals infected at year 9 in growing clusters. We select 1,000 individuals from clusters, inferred by either HIV-TRACE or TreeCluster, that have the highest growth rate (ties broken randomly). As a baseline control, the average number of infections over all individuals (similar to expectations under a random selection) is shown as well. For a cluster with n_t members at year t, growth rate is defined as $\frac{n_9 - n_8}{\sqrt{n_9}}$. The columns show varying expected degree, and all other parameters are their base values.

compared to the ER and WS models (Figs. S4), but these differences are marginal compared to impacts of $\mathbb{E}_{ART}$ and $\mathbb{E}_d$.

3.3 Evaluating clustering methods

By measuring the number of new infections caused by each person in the clusters with the highest growth rate, we observe that both clustering methods can potentially help target intervention and prevention services (Fig. 6). Over the entire population, the number of new infections caused by each person between years 9 and 10 is 0.029 for our base parameter settings. The top 1,000 people from the fastest growing TreeCluster clusters, in contrast, infect on average 0.065 new people. Thus, public health efforts at treatment and prevention are better spent on the growing clusters according to TreeCluster than just random targeting. HIV-TRACE performs even better than TreeCluster, increasing the per capita new infections among top 1,000 individuals to 0.086 for base parameters. Thus, an individual in the highest-growth HIV-TRACE clusters has about a three times higher chance of transmission compared to the general population. As $\mathbb{E}_{ART}$ decreases, as expected, the total number of per capita new infections reduces; as a result, the positive impact of using clustering methods to find the growing clusters gradually diminishes (Fig. 6). Conversely, reducing $\mathbb{E}_{ART}$ leads to further improvements obtained using TreeCluster versus random selection and using HIV-TRACE versus TreeCluster.

Changing $\mathbb{E}_d$ has a noticable impact on the results (Fig. 6). When $\mathbb{E}_d$ is decreased 4 (the base value) to 2, slowing the epidemic down, both methods remain very effective in finding high-risk individuals, and HIV-TRACE continues to outperform TreeCluster. However, when $\mathbb{E}_d$ is increased, the two methods first tie at $\mathbb{E}_d$= 8, and at $\mathbb{E}_d$= 16, TreeCluster becomes slightly better than HIV-TRACE for most $\mathbb{E}_{ART}$ values (Fig. 6). Interestingly, both methods are only barely better than a random selection of individuals if the epidemic is made very fast growing by setting $\mathbb{E}_{ART}\geq$ 2 and $\mathbb{E}_d$= 16.

However, besides these special cases, both methods seem very effective in most conditions we tested.

4 Discussion

Our results demonstrated that FAVITES can simulate under numerous different models and produce realistic data. We also showed that TreeCluster and HIV-TRACE, when paired with temporal monitoring, can successfully identify individuals most likely to transmit, and HIV-TRACE performs better than TreeCluster under most tested conditions. The ability to find people with increased risk of onward transmission is especially important because it can potentially help public health officials better spend their limited budgets for targeted prevention (e.g. pre-exposure prophylaxis, PrEP) or treatment (e.g. efforts to increase ART adherence).

We studied several models for various steps of our simulations, but we did not exhaustively test all models: FAVITES currently includes 21 modules and a total of 136 implementations across them, and testing all model combinations is infeasible, but we aimed to choose models that best emulate reality. All the sub-models we used can be criticized for imperfect capturing of reality. For example, our contact network remains unchanged with time, whereas real sexual networks are dynamic. Our transmission model does not directly model effective prevention measures such as PrEP. Our sequences include substitutions, but no recombination. Moreover, our models of sequence evolution assume the sites of a sequence are independent and identically-distributed (i.i.d.), ignoring evolutionary constraints across sites. We also ignored infections from outside the network (viral migration), assumed full patient sampling, and we sampled all patients at the end time as opposed to varied-time sampling. While these and other choices may impact results, we note that our goal here was mainly to show the utility of FAVITES. Importantly, beyond the numerous models currently implemented into FAVITES, new models with improved realism can easily be incorporated, and continued model improvement is a reason why we believe frameworks like FAVITES are needed.

We observed relatively high levels of error in inferred phylogenies. This is not surprising given the low rate of evolution and length of the *pol* region (which we emulate). Further, our phylogenetic trees include many super-short branches, perhaps due to our complete sampling. Many transmission cluster inference tools (e.g. PhyloPart, Cluster Picker, and TreeCluster) use phylogenetic trees during the inference process and thus may be sensitive to tree inference error. Other tools, such as HIV-TRACE do no attempt to infer a full phylogeny (only distances). The high levels of tree inference errors may be partially responsible for the relatively lower performance of TreeCluster compared to HIV-TRACE. Nevertheless, TreeCluster had higher per capita new infections in its fastest growing clusters than the population average, indicating that the trees, although imperfect, still include useful signal about the underlying transmission histories.

Nevertheless, we caution that our studies are not meant to be a definitive comparison of TreeCluster and HIV-TRACE, and results should be interpreted with several limitations kept in mind. A major limitation is that both methods we tested use a threshold internally for defining clusters. The specific choice of distance threshold defines a trade-off between cluster sensitivity and specificity, and the trade-off will impact cluster compositions. The best choice of the threshold is likely a function of epidemiological factors, and the default thresholds are perhaps optimal for certain epidemiological conditions, but not others. For example, we observed that, for a minority of our epidemiological settings, TreeCluster is more effective than HIV-TRACE in predicting growing clusters. A thorough exploration of all epidemiological parameters and method thresholds is left for future studies. On a practical note, FAVITES can enable public health officials to simulate conditions similar to their own epidemic and pick the best method/threshold tailored to their situation.

Beyond the methods, the approach we used for evaluating clustering methods, despite its natural appeal, is not the only possible measure. For example, the best way to choose high risk individuals given results of clustering at one time point or a series of time points is not clear. We used a strict ordering based on square-root-normalized cluster growth and arbitrary tiebreakers, but many other metrics and strategies can be imagined. For example, we may want to order individuals within a cluster by some criteria as well and choose certain number of people per cluster inversely proportional to the growth rate of the cluster. We simply choose 1,000 people to simulate a limited budget, but perhaps reducing/increasing this threshold give somewhat different results. We experimented with using 200 or 5,000 individuals as the fixed budget (Fig. S5) and observed that, while general patterns remain consistent, some differences are observed. For example, with increased budgets, as expected, the gap between clustering methods and random clustering is narrower. Less predictably, it appears that, in some conditions (e.g. $\mathbb{E}_d$= 16), lower budgets prefer TreeCluster while higher budgets prefer HIV-TRACE. A thorough exploration of the best method for each budget is beyond the scope of the current work. Similarly, we leave a comprehensive study of the best strategies to allocate budgets based on the results of clustering and better ways of measuring effectiveness to future work.

Funding

This work was supported by NIH subaward 5P30AI027767-28 to SM and NM and an NIH-NIAID K01 Career Development Award (K01AI110181), an NIH-NIAID R01 (AI135992), and a California HIV/AIDS Research Program (CHRP) IDEA Award (ID15-SD-052) to JOW.

References

Angly, F. E., Willner, D., Rohwer, F., Hugenholtz, P., and Tyson, G. W. (2012). Grinder: A versatile amplicon and shotgun sequence simulator. *Nucleic Acids Research*, **40**(12), e94.

Barabási, A.-L. and Albert, R. (1999). Emergence of Scaling in Random Networks. *Science*, **286**(5439), 509–512.

Bellan, S. E., Dushoff, J., Galvani, A. P., and Meyers, L. A. (2015). Reassessment of HIV-1 Acute Phase Infectivity: Accounting for Heterogeneity and Study Design with Simulated Cohorts. *PLoS Medicine*, **12**(3), e1001801.

Campbell, E. M., Jia, H., Shankar, A., Hanson, D., Luo, W., Masciotra, S., Owen, S. M., Oster, A. M., Galang, R. R., Spiller, M. W., Blosser, S. J., Chapman, E., Roseberry, J. C., Gentry, J., Pontones, P., Duwve, J., Peyrani, P., Kagan, R. M., Whitcomb, J. M., Peters, P. J., Heneine, W., Brooks, J. T., and Switzer, W. M. (2017). Detailed Transmission Network Analysis of a Large Opiate-Driven Outbreak of HIV Infection in the United States. *The Journal of infectious diseases*, **216**(9), 1053–1062.

Cohen, M. S., Chen, Y. Q., McCauley, M., Gamble, T., Hosseinipour, M. C., Kumarasamy, N., Hakim, J. G., Kumwenda, J., Grinsztejn, B., Pilotto, J. H., Godbole, S. V., Mehendale, S., Chariyalertsak, S., Santos, B. R., Mayer, K. H., Hoffman, I. F., Eshleman, S. H., Piwowar-Manning, E., Wang, L., Makhema, J., Mills, L. A., de Bruyn, G., Sanne, I., Eron, J., Gallant, J., Havlir, D., Swindells, S., Ribaudo, H., Elharrar, V., Burns, D., Taha, T. E., Nielsen-Saines, K., Celentano, D., Essex, M., and Fleming, T. R. (2011). Prevention of HIV-1 Infection with Early Antiretroviral Therapy. *New England Journal of Medicine*, **365**(6), 493–505.

Cori, A., Ayles, H., Beyers, N., Schaap, A., Floyd, S., Sabapathy, K., Eaton, J. W., Hauck, K., Smith, P., Griffith, S., Moore, A., Donnell, D., Vermund, S. H., Fidler, S., Hayes, R., and Fraser, C. (2014). HPTN 071 (PopART): A cluster-randomized trial of the population impact of an HIV combination prevention intervention including universal testing and treatment: Mathematical model. *PLoS ONE*, **9**(1), e84511.

Eddy, S. R. (1998). Profile hidden Markov models. *Bioinformatics*, **14**(9), 755–763.

Erdos, P. and Rényi, A. (1960). On the evolution of random graphs. *Magyar Tudományos Akadémia Matematikai Kutató Intézetének Kzleményei*, **5**, 17–61.

Felsenstein, J. (1981). Evolutionary trees from DNA sequences: A maximum likelihood approach. *Journal of Molecular Evolution*, **17**(6), 368–376.

Fortunato, S. (2010). Community detection in graphs. *Physics Reports*, **486**(3-5), 75–174.

Grabowski, M. K. and Redd, A. D. (2014). Molecular tools for studying HIV transmission in sexual networks. *Current Opinion in HIV and AIDS*, **9**(2), 126–133.

Granich, R. M., Gilks, C. F., Dye, C., De Cock, K. M., and Williams, B. G. (2009). Universal voluntary HIV testing with immediate antiretroviral therapy as a strategy for elimination of HIV transmission: a mathematical model. *The Lancet*, **373**(9657), 48–57.

Hagberg, A. A., Schult, D. A., and Swart, P. J. (2008). Exploring Network Structure, Dynamics, and Function using NetworkX. In G. Varoquaux, T. Vaught, and J. Millman, editors, *Proceedings of the 7th Python in Science conference (SciPy 2008)*, pages 11–15, Pasadena.

Hamilton, D. T., Handcock, M. S., and Morris, M. (2008). Degree Distributions in Sexual Networks: A Framework for Evaluating Evidence. *Sexually Transmitted Diseases*, **35**(1), 30–40.

Hartmann, K., Wong, D., and Stadler, T. (2010). Sampling trees from evolutionary models. *Systematic Biology*, **59**(4), 465–476.

Huang, W., Li, L., Myers, J. R., and Marth, G. T. (2012). ART: A next-generation sequencing read simulator. *Bioinformatics*, **28**(4), 593–594.

Jukes, T. H. and Cantor, C. R. (1969). Evolution of protein molecules. *Mammalian Protein Metabolism*, pages 21–132.

Karoński, M. (1982). A review of random graphs. *Journal of Graph Theory*, **6**(4), 349–389.

Kelly, J. A., St. Lawrence, J. S., Diaz, Y. E., Stevenson, L. Y., Hauth, A. C., Brasfield, T. L., Kalichman, S. C., Smith, J. E., and Andrew, M. E. (1991). HIV risk behavior reduction following intervention with key opinion leaders of population: An experimental analysis. *American Journal of Public Health*, **81**(2), 168–171.

Kimura, M. (1980). A simple method for estimating evolutionary rates of base substitutions through comparative studies of nucleotide sequences. *Journal of Molecular Evolution*, **16**(2), 111–120.

Kosiol, C., Holmes, I., and Goldman, N. (2007). An empirical codon model for protein sequence evolution. *Molecular Biology and Evolution*, **24**(7), 1464–1479.

Leitner, T., Escanillat, D., Franzent, C., Uhlen, M., and Albert, J. (1996). Accurate reconstruction of a known HIV-1 transmission history by phylogenetic tree analysis. *Evolution*, **93**(20), 10864–10869.

Lewis, F., Hughes, G. J., Rambaut, A., Pozniak, A., and Brown, A. J. L. (2008). Episodic Sexual Transmission of HIV Revealed by Molecular Phylodynamics. *PLoS Medicine — www.plosmedicine.org March*, **5**(3).

Little, S. J., Pond, S. L. K., Anderson, C. M., Young, J. A., Wertheim, J. O., Mehta, S. R., May, S., and Smith, D. M. (2014). Using HIV networks to inform real time prevention interventions. *PLoS ONE*, **9**(6).

Macchione, N., Wooten, W. J., Waters-Montijo, K., McDonald, E., Bursaw, M., Freitas, L., Tweeten, S., Awa, E., McGann, F., Johnson, M., and Hunter, S. (2015). HIV/AIDS Epidemiology Report. *County of San Diego Health and Human Services Agency Public Health Services*.

Mccloskey, R. M. and Poon, A. F. Y. (2017). A model-based clustering method to detect infectious disease transmission outbreaks from sequence variation. *PLOS Computational Biology*, **13**(11), e1005868.

Moshiri, N. (2018). TreeCluster: Massively scalable transmission clustering using phylogenetic trees. *bioRxiv*.

Moshiri, N. and Mirarab, S. (2017). A Two-State Model of Tree Evolution and Its Applications to Alu Retrotransposition. *Systematic Biology*, **0**(0), 1–15.

Newman, M. E. J., Watts, D. J., and Strogatz, S. H. (2002). Random graph models of social networks. *Proceedings of the National Academy of Sciences*, **99**(Suppl 1), 2566–2572.

Nosyk, B., Lourenço, L., Min, J. E., Shopin, D., Lima, V. D., and Montaner, J. S. (2015). Characterizing retention in HAART as a recurrent event process: insights into cascade churn'. *AIDS*, **29**(13), 1681–1689.

O'Brien, M. and Markowitz, M. (2012). Should We Treat Acute HIV Infection? *Current HIV/AIDS Reports*, **9**(2), 101–110.

Pérez-Losada, M., Castel, A. D., Lewis, B., Kharfen, M., Cartwright, C. P., Huang, B., Maxwell, T., Greenberg, A. E., and Crandall, K. A. (2017). Characterization of HIV diversity, phylodynamics and drug resistance in Washington, DC. *PLoS ONE*, **12**(9), e0185644.

Pond, S. L. K., Weaver, S., Brown, A. J. L., 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*, (msy016).

Price, M. N., Dehal, P. S., and Arkin, A. P. (2010). FastTree 2 - Approximately maximum-likelihood trees for large alignments. *PLoS ONE*, **5**(3).

Prosperi, M. C., Ciccozzi, M., Fanti, I., Saladini, F., Pecorari, M., Borghi, V., Di Giambenedetto, S., Bruzzone, B., Capetti, A., Vivarelli, A., Rusconi, S., Re, M. C., Gismondo, M. R., Sighinolfi, L., Gray, R. R., Salemi, M., Zazzi, M., and De Luca, A. (2011). A novel methodology for large-scale phylogeny partition. *Nature Communications*, **2**(2), 321.

Ragonnet-Cronin, M., Hodcroft, E., Hué, S., Fearnhill, E., Delpech, V., Brown, A. J., and Lycett, S. (2013). Automated analysis of phylogenetic clusters. *BMC Bioinformatics*, **14**(1), 317.

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

Ratmann, O., Hodcroft, E. B., Pickles, M., Cori, A., Hall, M., Lycett, S., Colijn, C., Dearlove, B., Didelot, X., Frost, S., Md Mukarram Hossain, A. S., Joy, J. B., Kendall, M., Kuhnert, D., Leventhal, G. E., Liang, R., Plazzotta, G., Poon, A. F., Rasmussen, D. A., Stadler, T., Volz, E., Weis, C., Brown, A. J., and Fraser, C. (2017). Phylogenetic tools for generalized HIV-1 epidemics: Findings from the PANGEA-HIV methods comparison. *Molecular Biology and Evolution*, **34**(1), 185–203.

Robinson, D. F. and Foulds, L. R. (1981). Comparison of phylogenetic trees. *Mathematical Biosciences*, **53**(1-2), 131–147.

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.

Rose, R., Lamers, S. L., Dollar, J. J., Grabowski, M. K., Hodcroft, E. B., Ragonnet-Cronin, M., Wertheim, J. O., Redd, A. D., German, D., and Laeyendecker, O. (2017). Identifying Transmission Clusters with Cluster Picker and HIV-TRACE. *AIDS Research and Human Retroviruses*, **33**(3), 211–218.

Rosenberg, E. S., Sullivan, P. S., Dinenno, E. A., Salazar, L. F., and Sanchez, T. H. (2011). Number of casual male sexual partners and associated factors among men who have sex with men: Results from the National HIV Behavioral Surveillance system. *BMC Public Health*, **11**(189).

Sahneh, F. D., Vajdi, A., Shakeri, H., Fan, F., and Scoglio, C. (2017). GEMFsim: A stochastic simulator for the generalized epidemic modeling framework. *Journal of Computational Science*, **22**, 36–44.

Shepard, J., Bowen, N., Ginsberg, M., Bursaw, M., Awa, E., Cardoza, L., Freitas, L., Johnson, M., McGann, F., Salgado, S., Tweeten, S., and Van Meter, J. (2005). HIV/AIDS Epidemiology Report. *County of San Diego Health and Human Services Agency Public Health Services*.

Smirnov, N. V. (1939). On the Estimation of the Discrepancy Between Empirical Curves of Distribution for Two Independent Samples. *Bul. Math. de l'Univ. de Moscou*, **2**, 3–14.

Spielman, S. J. and Wilke, C. O. (2015). Pyvolve: A flexible python module for simulating sequences along phylogenies. *PLoS ONE*, **10**(9), e0139047.

Stamatakis, A. (2014). RAxML version 8: A tool for phylogenetic analysis and post-analysis of large phylogenies. *Bioinformatics*, **30**(9), 1312–1313.

Sukumaran, J. and Holder, M. T. (2010). DendroPy: A Python library for phylogenetic computing. *Bioinformatics*, **26**(12), 1569–1571.

Tamura, K. and Nei, M. (1993). Estimation of the Number of Nucleotide Substitutions in the Control Region of Mitochondrial DNA in Humans and Chimpanzees '. *Molecular Biology and Evolution*, **10**(3), 512–526.

Tavaré, S. (1986). Some Probabilistic and Statistical Problems in the Analysis of DNA Sequences. In *Lectures on Mathematics in the Life Sciences*, pages 57–86. American Mathematical Society, 17 edition.

UNAIDS (2017). 90-90-90: An ambitious treatment target to help end the AIDS epidemic. Technical report, UNAIDS, Geneva, Switzerland.

Villandre, L., Stephens, D. A., Labbe, A., Günthard, H. F., Kouyos, R., Stadler, T., Aubert, V., Battegay, M., Bernasconi, E., Böni, J., Bucher, H. C., Burton-Jeangros, C., Calmy, A., Cavassini, M., Dollenmaier, G., Egger, M., Elzi, L., Fehr, J., Fellay, J., Furrer, H., Fux, C. A., Gorgievski, M., Günthard, H., Haerry, D., Hasse, B., Hirsch, H. H., Hoffmann, M., Hösli, I., Kahlert, C., Kaiser, L., Keiser, O., Klimkait, T., Kovari, H., Ledergerber, B., Martinetti, G., Martinez De Tejada, B., Metzner, K., Müller, N., Nadal, D., Nicca, D., Pantaleo, G., Rauch, A., Regenass, S., Rickenbach, M., Rudin, C., Schöni-Affolter, F., Schmid, P., Schüpbach, J., Speck, R., Tarr, P., Telenti, A., Trkola, A., Vernazza, P., Weber, R., and Yerly, S. (2016). Assessment of overlap of phylogenetic transmission clusters and communities in simple sexual contact networks: Applications to HIV-1. *PLoS ONE*, **11**(2), e0148459.

Vrancken, B., Rambaut, A., Suchard, M. A., Drummond, A., Baele, G., Derdelinckx, I., Van Wijngaerden, E., Vandamme, A. M., Van Laethem, K., and Lemey, P. (2014). The Genealogical Population Dynamics of HIV-1 in a Large Transmission Chain: Bridging within and among Host Evolutionary Rates. *PLoS Computational Biology*, **10**(4), e1003505.

Watts, D. J. (1999). Networks, Dynamics, and the Small-World Phenomenon. *American Journal of Sociology*, **105**(2), 493–527.

Watts, D. J. and Strogatz, S. H. (1998). Collective dynamics of 'small-world' networks. *Nature*, **393**(6684), 440–442.

Wawer, M. J., Gray, R. H., Sewankambo, N. K., Serwadda, D., Li, X., Laeyendecker, O., Kiwanuka, N., Kigozi, G., Kiddugavu, M., Lutalo, T., Nalugoda, F., WabwireMangen, F., Meehan, M. P., and Quinn, T. C. (2005). Rates of HIV1 Transmission per Coital Act, by Stage of HIV1 Infection, in Rakai, Uganda. *The Journal of Infectious Diseases*, **191**(9), 1403–1409.

Wertheim, J., Murrell, B. L., Forgione, L., and Torian (2017a). Cluster growth dynamics suggest strategy for targeted intervention in New York City public health HIV-1 surveillance registry. In A. Leigh Brown, A. McLean, E. Hodcroft, J. Albert, M. Kalish, T. Leitner, B. Korber, J. Mullins, S. Kosakovsky Pond, M. Rolland, S. Wolinsky, and M. Worobey, editors, *HIV Dynamics & Evolution*, number 1122, page 38, Sleat, Isle of Skye, Scotland. UC San Diego School of Medicine.

Wertheim, J. O., Kosakovsky Pond, S. L., Little, S. J., and De Gruttola, V. (2011). Using HIV Transmission Networks to Investigate Community Effects in HIV Prevention Trials. *PLoS ONE*, **6**(11), e27775.

Wertheim, J. O., Leigh Brown, A. J., Hepler, N. L., Mehta, S. R., Richman, D. D., Smith, D. M., and Kosakovsky Pond, S. L. (2014). The global transmission network of HIV-1. *Journal of Infectious Diseases*, **209**(2), 304–313.

Wertheim, J. O., Kosakovsky Pond, S. L., Forgione, L. A., Mehta, S. R., Murrell, B., Shah, S., Smith, D. M., Scheffler, K., and Torian, L. V. (2017b). Social and Genetic Networks of HIV-1 Transmission in New York City. *PLoS Pathogens*, **13**(1), e1006000.

Worobey, M., Watts, T. D., McKay, R. A., Suchard, M. A., Granade, T., Teuwen, D. E., Koblin, B. A., Heneine, W., Lemey, P., and Jaffe, H. W. (2016). 1970s and 'Patient 0' HIV-1 genomes illuminate early HIV/AIDS history in North America. *Nature*, **539**(7627), 98–101.

Yang, Z. (1994). Maximum likelihood phylogenetic estimation from DNA sequences with variable rates over sites: Approximate methods. *Journal of Molecular Evolution*, **39**(3), 306–314.

Ypma, R. J., van Ballegooijen, W. M., and Wallinga, J. (2013). Relating phylogenetic trees to transmission trees of infectious disease outbreaks. *Genetics*, **195**(3), 1055–1062.

Zaheri, M., Dib, L., and Salamin, N. (2014). A generalized mechanistic codon model. *Molecular Biology and Evolution*, **31**(9), 2528–2541.