PLOS COMPUTATIONAL BIOLOGY

RESEARCH ARTICLE

Combining biomarker and virus phylogenetic models improves HIV-1 epidemiological source identification

Erik Lundgren[1], **Ethan Romero-Severson**[1], **Jan Albert**[2,3], **Thomas Leitner**[1]*

1 Theoretical Biology and Biophysics Group, Los Alamos National Laboratory, Los Alamos, New Mexico, United States of America, **2** Department of Microbiology, Tumor and Cell Biology, Karolinska Institutet, Stockholm, Sweden, **3** Department of Clinical Microbiology, Karolinska University Hospital, Stockholm, Sweden

* tkl@lanl.gov

OPEN ACCESS

Citation: Lundgren E, Romero-Severson E, Albert J, Leitner T (2022) Combining biomarker and virus phylogenetic models improves HIV-1 epidemiological source identification. PLoS Comput Biol 18(8): e1009741. https://doi.org/10.1371/journal.pcbi.1009741

Editor: Katrina Abigail Lythgoe, University of Oxford, UNITED KINGDOM

Received: December 9, 2021

Accepted: August 2, 2022

Published: August 26, 2022

Data Availability Statement: Data is available in the open literature, as cited in the paper and in DOI: 10.1371/journal.pcbi.1005316, DOI: 10.1093/molbev/msu179 and DOI: 10.1371/journal.pone.0033484, and in GenBank (Accession numbers JQ698667 – JQ698874). Codes are available at https://github.com/MolEvolEpid/biophybreak. Codes and data for publication simulation and biomarker figures are in supplemental materials.

Funding: This study was supported by NIH/NIAID grants R01AI087520 and R01AI152897 to TL. The

Abstract

To identify and stop active HIV transmission chains new epidemiological techniques are needed. Here, we describe the development of a multi-biomarker augmentation to phylogenetic inference of the underlying transmission history in a local population. HIV biomarkers are measurable biological quantities that have some relationship to the amount of time someone has been infected with HIV. To train our model, we used five biomarkers based on real data from serological assays, HIV sequence data, and target cell counts in longitudinally followed, untreated patients with known infection times. The biomarkers were modeled with a mixed effects framework to allow for patient specific variation and general trends, and fit to patient data using Markov Chain Monte Carlo (MCMC) methods. Subsequently, the density of the unobserved infection time conditional on observed biomarkers were obtained by integrating out the random effects from the model fit. This probabilistic information about infection times was incorporated into the likelihood function for the transmission history and phylogenetic tree reconstruction, informed by the HIV sequence data. To critically test our methodology, we developed a coalescent-based simulation framework that generates phylogenies and biomarkers given a specific or general transmission history. Testing on many epidemiological scenarios showed that biomarker augmented phylogenetics can reach 90% accuracy under idealized situations. Under realistic within-host HIV-1 evolution, involving substantial within-host diversification and frequent transmission of multiple lineages, the average accuracy was at about 50% in transmission clusters involving 5–50 hosts. Realistic biomarker data added on average 16 percentage points over using the phylogeny alone. Using more biomarkers improved the performance. Shorter temporal spacing between transmission events and increased transmission heterogeneity reduced reconstruction accuracy, but larger clusters were not harder to get right. More sequence data per infected host also improved accuracy. We show that the method is robust to incomplete sampling and that adding biomarkers improves reconstructions of real HIV-1 transmission histories. The technology presented here could allow for better prevention programs by providing data for locally informed and tailored strategies.

funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Competing interests: The authors have declared that no competing interests exist.

Author summary

While phylogenetic methods have been successfully used to infer transmission patterns on many scales, substantial uncertainty about the corresponding transmission history has remained. Here, we introduce a formal inference framework for the simultaneous use of HIV biomarkers and HIV sequences derived from persons who may have infected each other. We created a flexible system that can use up to five biomarkers to estimate the time of infection of each person, with the appropriate uncertainty given by an empirical probability distribution. These time-of-infection distributions were jointly modelled with the HIV phylogenetic tree estimation to produce possible transmission histories. We show that adding biomarkers substantially limits the possible transmission histories. This makes it possible to identify transmission risks, to assess confidence in source attributions, and to make efficient resource allocations to prevent further transmissions in local epidemics.

Introduction

To effectively control an infectious disease, limited prevention resources must be allocated to where they are needed most [1]. Thus, identifying hotspots of transmission would allow for efficient resource allocation. The probability of transmission, especially in chronic infections such as HIV, is heterogenous over time, both on the epidemic scale and over a single person's time since infection, leading to episodic transmissions and local outbreaks of epidemiologically closely linked individuals (transmission clusters). Mapping transmission events using traditional epidemiological methods is challenging, expensive, and slow, and may be inaccurate. For example, several studies have reported that interview-based information about sexual contacts where HIV transmission might have taken place was often not in agreement with the phylogenetic history of the transmitted virus [2,3]. Therefore, phylogenetic reconstruction, using existing and growing public health databases, provides an attractive fact-based and less expensive alternative.

Reconstructing the history of an epidemic using phylogenetic methods has become a substantial domain of phylodynamic research involving many different pathogens [4–11]. The primary technical challenge in this domain comes from insufficient sampling, i.e., using samples from a single time point, not adequately representing the within-host diversity, and not providing constraints on when transmissions may have occurred. This applies especially for chronic infections where the pathogen develops substantial within-host diversity, such as in HIV, HBV, HCV, and some bacterial infections [12–16]. This within-host diversity means that when a person infects another, there are many alternative phylogenetic lineages that could have been involved, often more than one, leading to a non-trivial and non-identical correspondence between the transmission history and the pathogen phylogeny [12,17,18]. The extent of this problem was quantified by Hall and Coljin's method that counts the exact number of transmission histories that are logically consistent with a pathogen phylogeny [19]. For example, a phylogeny from 20 infected persons (with 1 sequence/person) could have as many as 102 million transmission histories that are consistent with that phylogeny—the exact number depends on the observed phylogenetic topology. While additional constraints and Bayesian inference can overcome weak non-identifiability, it is desirable to have constraints that are both measurable (i.e., empirical) and based on readily available data. While the basic theoretical underpinnings of how the order of infection events constrains the possible transmission

histories for a given phylogeny have been understood for several years [20], this knowledge currently is not integrated into phylodynamic inference methods. Some software such as SCOTTI [7] are able to include user-defined infection windows, i.e., a fixed period of time that a host was contagious. Unfortunately, for the case of life-long, chronic infections such as HIV, in the absence of treatment such a window would be too long to be useful, and typically the actual start of contagiousness is rarely known.

HIV biomarkers offer an alternative to infection windows, instead estimating when a host was infected [21]. Here, we introduce the use of HIV biomarkers to augment phylogenetic reconstruction and narrow down the possible transmission histories among epidemiologically linked hosts. Some such biomarkers are always available in clinical and public health databases, including HIV *pol* sequences, CD4 cell counts, and viral load measurements, and sometimes quantitative serological assay test results. In addition, there may be information about previous negative HIV test results and other demographic information that may limit possible time of infection. We show that it is possible to enhance transmission history reconstruction by modeling multiple biomarkers in a joint biomarker-phylogeny-transmission history framework.

Materials and methods

Methodological overview

A transmission history is defined as who infected whom and when those transmission(s) occurred. While HIV transmissions can occur in many different contact networks [22], in this work we will consider 5–50 individual hosts that have transmitted HIV in different time intervals. When attempting to reconstruct a local transmission history, one can divide the available information into three levels, i) information from the sampling times, ii) information from the genetic sequences, and iii) information about the infection times (Fig 1). If only the sampling times were known, then the virus phylogeny and transmission history would be almost completely unconstrained. Adding sequence data limits the set of possible transmission histories by revealing temporal evolutionary relationships among sampled pathogens, which constrains the transmission history to some extent, but still leaves a wide variety of possibilities. With the additional information about the host's infection times, the transmission history without considering infection times is still not fully constrained, but there are many fewer possibilities than with sequences alone. In the example in Fig 1 (right panel), the red individual is the first one infected in the cluster, but the order of the other two infections is not certain. In this particular example, there are nine possible configurations among three individuals, out of which all are approximately equally likely given only the sample times (left column). With the additional information from the sequences, all configurations are still plausible, but most of the probability density is concentrated in four configurations (center column). When including the probability density functions for the infection times, however, the number of plausible transmission histories (without infection times) is reduced to three, with half of the total density concentrated in the configuration with the correct infector (right column).

Multiple biomarker model

HIV biomarkers are measurable biological quantities that have some relationship to the amount of time someone has been infected with HIV. Based on the values of a set of biomarkers, we can infer a probability distribution for the amount of time that that person has been infected using a mixed effects Multiple Biomarker Model (MBM). Our model extends the previous model by Giardina et al 2019 [21], adding two additional biomarkers (bringing the total to five) and allowing inference when not all biomarker measurements are available. We also

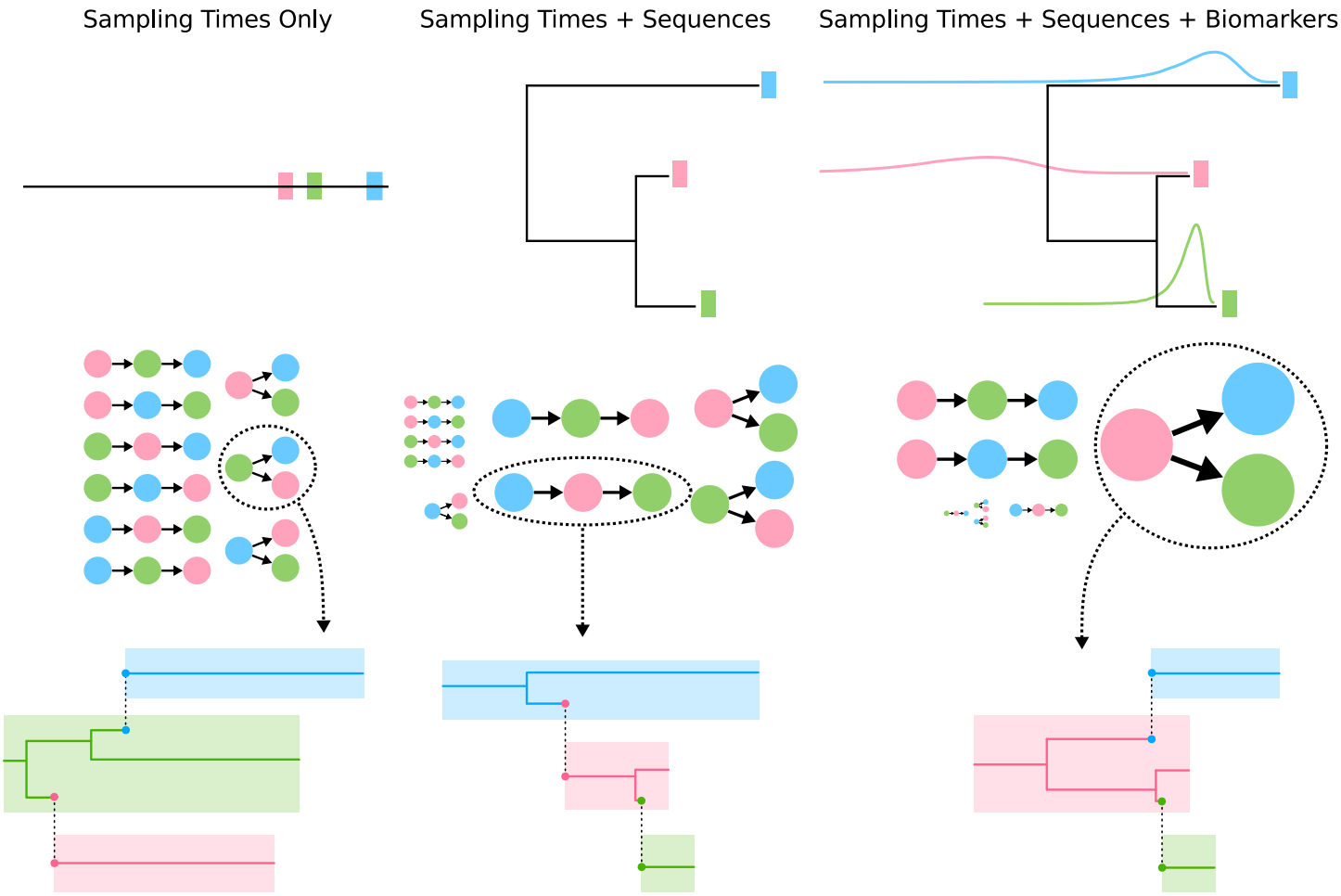

Fig 1. Conceptual Model Motivation. The level of information available for inference (top row) affects the support for the possible transmissions between hosts (filled colored circles in middle row and colored boxes in bottom row). The information from the biomarkers is shown as probability densities of the time of infection for each host. The middle row shows examples of plausible transmission histories (excluding infection times), with the arrows indicating the direction of transmission (pointing from infector to recipient). The relative size of each possibility corresponds to the order of magnitude of the posterior probability for that configuration, with larger configurations having greater support. For each level of information, a transmission history with a phylogeny is shown from a selected configuration, with transmissions indicated by dashed vertical lines (bottom row). The left side of each colored box in the transmission history represents the time of infection, and the right side the time of sampling. Note that the phylogeny is identical in the middle and right panels of the bottom row, but the inferred transmission history is different.

https://doi.org/10.1371/journal.pcbi.1009741.g001

scaled the biomarker values such that the input values to the model are all at the same order of magnitude and used a prior distribution that more closely reflects the expected amount of time between infection and diagnosis.

The mixed effects model is of the form $Y_{ij}^k = f^k(S_{ij} - I_i, \beta_i^k) + \epsilon_{ij}^k$, where Y_{ij}^k is the measured value of the k^{th} biomarker at the j^{th} timepoint for the i^{th} individual, f^k is the function that predicts the value of a biomarker based on time after infection, S_{ij} is the time of the j^{th} sample for the i^{th} individual, I_i is the infection time of the i^{th} individual, β_i^k are the function parameters for the k^{th} biomarker for the i^{th} individual, and ϵ_{ij}^k is the biological and measurement noise. Individual β_i^k values are modeled as draws from a multivariate normal distribution $\boldsymbol{\beta}$ and ϵ_{ij}^k values are independent Gaussian noise terms where the standard deviation depends only on k.

We modeled five biomarkers: BED, the IgG capture BED enzyme immunoassay [23]; LAg, Limiting Antigen Avidity assay [24]; *pol* polymorphism count, the number of multi-state

nucleotide characters in *pol* direct population sequences [25,26]; *pol* NGS diversity, HIV diversity estimated from next-generation sequencing of *pol* on the Illumina platform [27]; and CD4 cell count, the number of CD4 positive T cells in 1 ml plasma. As shown in Fig 2, BED and LAg are modeled as log10 values, normalized with an internal test standard, starting at low concentrations that rise asymptotically over time since infection as in Skar et al [28], with LAg typically rising faster than BED. Both the *pol* polymorphism count and the *pol* NGS diversity (of 3rd codon positions) increase approximately linearly at the timescales that we are interested in. The CD4 cell count is modelled as the square root in order to make the time series trend more linear.

We used *rjags* [29] to draw Markov Chain Monte Carlo (MCMC) samples from the posterior distributions of the infection times. The prior distribution for the infection ages at the times of diagnosis was a Gamma distribution with mean 2 years and standard deviation 1.5 years, chosen to be qualitatively similar to the distributions for time between infection and diagnosis found in Giardina et al. 2019 [21]. We created a flexible system that can use any number or combination of biomarkers (including no biomarkers, which would recover the prior distribution).

Biomarker training data and validation

The training data for our multiple biomarker model came from 30 longitudinally followed Swedish HIV-infected persons with well-defined times of infection. Biomarker data from these patients have been previously used: *pol* polymorphism, BED, and CD4 counts were used in Giardina et al [21]; and *pol* NGS diversity in Puller et al [27]. In this study we added LAg for a total of five biomarkers. The 30 patients were selected to have: 1) a previous negative test and first positive test that were no more than 6 months apart or a known primary HIV infection time, 2) at least three follow up measurements over a long time period (2–5 years), and 3) been treatment naïve for that time period. The biomarker data were measured on stored biobank samples because the inclusion criteria are difficult to fulfill as modern clinical practice is to put patients on antiviral treatment immediately after HIV diagnosis. The full model was trained using all measurements from all 30 patients, using *rjags* with 5×10^4 iterations for an initial sampling phase during which the samplers adapt their behavior to maximize their efficiency, followed by 3×10^5 iterations of burn-in, and 1×10^6 samples, providing estimates for the mean vector and covariance matrix for the multivariate normal $\boldsymbol{\beta}$ distribution and standard deviations of the ϵ^k Gaussian noise terms.

To validate our MBM, we performed a leave one out cross validation using one set of biomarker measurements from each of the 30 patients as testing data while using all measurements from the other 29 patients as training data. The MBM was also used to simulate realistic biomarker values for testing purposes: We first found the maximum likelihood values of the model parameters (trained on all 30 patients), then used those values to simulate new random effects trajectories for the expected values of each of the biomarkers over time, and, finally, simulated new biomarker values by adding Gaussian noise to the expected values.

Joint inference of transmission history and phylogeny

The MBM-derived posterior distributions of the infection times were incorporated into the posterior probability for the transmission history and phylogenetic tree. The general form of the posterior probability remains the same as in Klinkenberg et al 2017 [6]:

$$\Pr(I, M, P, \theta | S, G) \propto \Pr(S, G | I, M, P, \theta) \cdot \Pr(I, M, P, \theta),$$

with unobserved infection times I, infectors M, phylogeny P, parameters θ (the mutation rate,

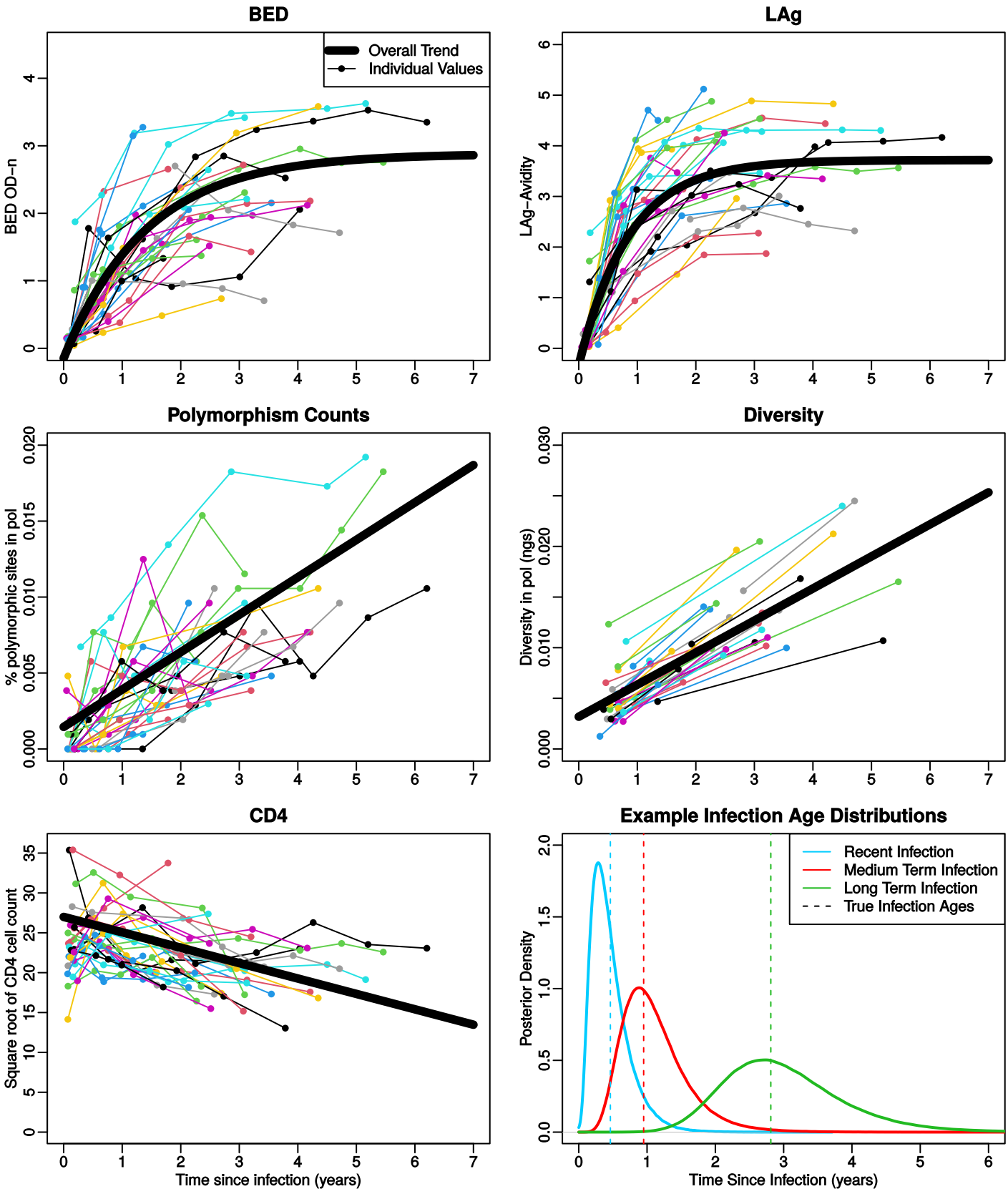
BED
Overall Trend
Individual Values
BED OD−n
LAg
LAg−Avidity
Polymorphism Counts
% polymorphic sites in pol
Diversity
Diversity in pol (ngs)
CD4
Square root of CD4 cell count
Time since infection (years)
Example Infection Age Distributions
Recent Infection
Medium Term Infection
Long Term Infection
True Infection Ages
Posterior Density
Time Since Infection (years)

Fig 2. HIV Biomarker time trends. Time series plots for each of the five biomarkers used in our modeling for a group of 30 patients (colored lines). The best fit line for the fixed part of our mixed effects model is shown for each biomarker (bold black lines). The bottom right panel shows examples of the inferred distributions of infection age from the cross validation using all five biomarker values from a recently infected patient, a patient who has been infected somewhat longer, and a longer-term infected patient. The true infection times for each patient are shown as vertical dashed lines.

https://doi.org/10.1371/journal.pcbi.1009741.g002

within-host parameters α and β, the transmissibility profile function, and the infection age distribution for each individual), observed sampling times S, and genetic sequences G. Again following Klinkenberg, this posterior probability can be split up into four terms for the likelihood of the sequences $\Pr(G|P,\theta)$, the phylogenetic tree $\Pr(P|S,I,M,\theta)$, the difference between the infection and sampling times $\Pr(S|I,\theta)$, and the transmission history considering the transmissibility profile $\Pr(I,M|\theta)$, as well as a term for the prior distributions of the parameters $\Pr(\theta)$:

$$\Pr(I, M, P, \theta|S, G) \propto \Pr(G|P, \theta) \cdot \Pr(P|S, I, M, \theta) \cdot \Pr(S|I, \theta) \cdot \Pr(I, M|\theta) \cdot \Pr(\theta).$$

Given the distributions, we performed the transmission history and phylogenetic inference using a modified version of the R package *phybreak* [6]. We refer to our modified version as *biophybreak* (https://github.com/MolEvolEpid/biophybreak). We modified a version of *phybreak* updated by the original author, with the most notable change since the publication of Klinkenberg et al 2017 [6] being the inclusion of the possibility of a wide transmission bottleneck. This modification is important to model realistic HIV transmission where >1 phylogenetic lineage often is transmitted [30]. The primary modification that we introduced here (in *biophybreak*) is the way in which the likelihood for the interval between infection and sampling is calculated. In the original version of *phybreak*, the length of that interval is a Gamma distribution with a user specified shape and mean that is estimated as a model parameter, with the same parameters for every individual. Our modification allows any non-parametric distribution to be used for this likelihood as well as allow each individual to have their own distribution. Specifically, we used kernel density estimation to obtain posterior distributions of the infection ages using posterior samples from the MBM for each individual. Additionally, we added the Generalized Time Reversible (GTR) substitution model, which has been shown to be the most realistic HIV evolutionary model [31] (instead of the Jukes-Cantor model used in the original *phybreak* package). However, this comes at a computational cost of about four times longer MCMC iteration steps. The third modification we made is similar to the first in that we allow the "transmissibility profile", i.e., the function that specifies how likely an individual is to infect another individual based on how long they have been infected, to be any distribution instead of only a gamma distribution as in the original *phybreak*. This was motivated by the observation that a higher viral load, seen in the acute infection stage, results in a higher risk of transmission [32]. Since it is difficult to know exactly what the shape of the transmissibility profile function should be, but it is known that approximately half of transmissions are from recently (within six months) infected individuals [33], we used a step function that is three times higher in the first six months, resulting in a fairly conservative improper prior distribution, which is fine in this case since only the relative values of the distribution are important. As with the infection time distributions, we allow any function to be used, e.g., modifying transmission probability before and after diagnosis, as described in the next section.

General simulation experimental design

In order to test our methodology and determine which types of transmission histories may be more or less difficult to correctly infer, we performed a variety of simulations. We varied the number of individuals, the mean time between subsequent infection events anywhere in the transmission history (temporal spacing), the heterogeneity of the number of transmissions per

individual (standard deviation of the network degree, or equivalently offspring number, excluding the most recently infected individual), the level of information about infection time we get from the biomarkers, and the mutation model used in the simulation and inference (mutation parameters according to the HIV-1 *pol* or *env* gene).

For the simulation sets that use randomly generated transmission histories, we first specified the number of individuals, then the amount of time from the first to the last infection, taking into account the number of individuals and temporal spacing. The infection times were then generated using a continuous uniform random variable between the first and last infections, with the option to have a minimum amount of time between any two infection times, which we set to 0.05 years. Next, sampling times were given to each individual. Infectors (direct transmission sources) for each individual were chosen from the pool of previously infected individuals such that the resulting transmission history had a transmission heterogeneity close to the desired value. To facilitate this, weights were assigned to each individual for how likely they were to be a direct transmission source, with the variation in weights depending on the target amount of transmission heterogeneity (higher variance of weights typically allows higher transmission heterogeneity). The weights from the transmissibility profile functions were also taken into account at this time. Note that since the infection times were chosen before the infectors rather than having the new infection times chosen from each infector's transmissibility profile, that function was implicitly changed in a non-trivial way. That is, the differences in infection times between direct transmission sources and recipients would not follow the same distribution as draws from the transmissibility profile. Therefore, we also allowed *biophybreak* to optionally use a penalty for transmission after diagnosis, which may be justified when all patients are successfully and continuously treated after diagnosis [34]. Hence, the weights of the potential infectors can be modified by a factor depending on whether they have been sampled yet. As in the case of the transmissibility profile function, this post-sampling transmission penalty is implicitly changed by the way the infection times are chosen.

Given a transmission history, we created the phylogeny with a coalescence simulator that used a within-host model of linearly increasing the effective population size $N_e(t) = \alpha+\beta t$, where α is the effective population size at the time of infection and β is the growth rate of the effective population size per generation [17,35], with a generation assumed to be 1.5 days [36]. Unless otherwise noted, we used $\alpha = 5$ and $\beta = 5$. Finally, sequences were generated with Seq-Gen [37] using known absolute and relative substitution rate parameters from either the HIV-1 envelope gene (*env*) or polymerase gene (*pol*).

We performed the transmission history and phylogenic inference with *biophybreak* using 2×10^5 MCMC samples after 5×10^4 iterations of burn-in unless otherwise noted. We used an effective sample size (ESS) of 200 for the model parameters ensure proper mixing of the MCMC chains. We used two different measurements of model performance. The first, which we call accuracy, is simply the proportion of individuals in a cluster for which the infector with the highest posterior support was in fact the true infector. For some tests, we also used the mean of the posterior support values for the true infector, which we call the true posterior probability.

Effect of biomarker information

To test the potential of improving transmission history inference using real biomarkers, we used a transmission history with fixed infection times and 15 individuals, mean temporal spacing of 0.5 years, and transmission heterogeneity of about 1.24, generating 100 instances of this transmission history with different sets of sampling times for each instance. For each individual in each history, the time between infection and sampling is determined by independent

samples from a Gamma distribution with mean 2 years and standard deviation 1.5 years. Sequences were generated for each phylogeny using both the *env* and *pol* mutation parameters. We simulated biomarker values for all of the infection ages at the time of sampling, then ran the multiple biomarker model to infer the infection time distributions. In order to assess the effect of the amount of biomarker information, we ran transmission history inference with the infection age distributions using 2, 3, or 5 biomarkers as well as infection age distributions representing no or uninformative biomarkers and near perfect infection age information that effectively provided fixed infection times. In the no biomarker scenario, the infection age distribution was a continuous uniform distribution with minimum 0 and maximum 11 years. In the fixed infection time scenario we used a Gamma distribution with mean equal to the true infection age and standard deviation equal to 0.005 years.

Effects of transmission cluster attributes

To test the effect of various attributes of the transmission cluster itself, one variable at a time, we investigated 1) the number of individuals, ranging from 5 to 50, 2) the temporal spacing, ranging from 0.01 years to 2.5 years, and 3) the transmission heterogeneity, ranging from 0 to 3.74 (the maximum possible for 15 individuals cluster). While each attribute was varied, the other attributes were held constant, with the non-variable values at 15 individuals, temporal spacing of 0.5 years, and transmission heterogeneity around 1 (between 0.8 and 1.2). Both the *env* and *pol* mutation models were used for each history. In all trials, we used simulated infection age distributions using all five biomarkers. We used two subsets of trials, one with more realistic HIV-1 values and one corresponding to an idealized situation. For the realistic values, we again used $\alpha = 5$ and $\beta = 5$ for the within-host model, two years between infection and sampling, and independent biomarkers for each individual. For the idealized situation, we used $\alpha = 0$ and $\beta = 0.1$ (resulting in a short pre-transmission interval [12,18] where the phylogeny closely resembles the transmission history), one year between infection and sampling, and fixed biomarkers for all individuals.

We also tested how different attributes of the transmission clusters may interact with each other, possibly affecting the difficulty of inference. To do this, we simulated histories with all combinations of different values for each attribute. These simulated clusters had 10, 15, 20, or 40 individuals, temporal spacing of 0.1, 0.5, 1, or 2 years, transmission heterogeneity near 0, 0.5, 1, 1.7, 2.3, or 3, with both substitution models (*env* and *pol*) and all five levels of biomarker information used with each cluster.

Effect of multiple sequences per individual

To test whether additional sequence data per patient can help counteract the inference problems inherent with wide transmission bottlenecks, we simulated 200 transmission histories with 3 individuals with temporal spacing of 0.5 years, including both the serial infection scenario and the scenario where the first individual infects the other two. We simulated phylogenies on each transmission history with 4 sampled sequences per individual taken two years after the time of infection, and $\alpha = 5$, $\beta = 5$. Next, we generated subsampled phylogenies, keeping only 1 sequence per individual. Finally, inference was performed on both the full and subsampled datasets using 1×10^6 iterations of MCMC after 4×10^4 iterations of burn-in, with some runs concluding sooner if the target ESS is reached early.

Effect of incomplete sampling of transmission clusters

To investigate how unsampled hosts may impact performance, we ran the inference method on both complete and incomplete simulated transmission histories. We used four different

basic transmission histories with the amount of transmission heterogeneity varying from none to moderately high, with the infectors, infection times, and sampling times fixed within each of the four transmission histories, while the phylogenies and sequences for each replicate were generated independently. The phylogenies were generated using the coalescent simulation with two different values of α, corresponding to wide ($\alpha = 5$) and complete ($\alpha = 0$) transmission bottlenecks, while β remained at 5. In each scenario, transmission history inference was performed on both the complete transmission history as well as that same transmission history with the data from the individual that infected the most other individuals removed (in the no transmission heterogeneity scenario, the sixth individual is removed). In both scenarios, there was no penalty for transmission after diagnosis. Accuracy is defined as before except that in the incomplete sampling scenario, we considered the inference to be "accurate" when the true infector's infector is chosen when the true infector is not sampled. In addition to the overall accuracy, we also looked at the accuracy of the individuals whose true infector is not sampled on their own.

Real HIV transmission cluster data

We demonstrate the application of the inference method on data from 4 real transmission clusters involving 4–14 patients in the larger Swedish HIV epidemic that are believed to be at least close to fully sampled [17,38,39]. These data included sequence data from *pol* drug resistance testing, biomarkers BED, CD4, and *pol* polymorphisms, as well as first positive test dates for all patients, and some patients had a previous negative test date.

We first used the MBM to infer the distributions of the infection times for all individuals. If a previous negative test was available, we assumed that those individuals could not have been infected more than two months before the most recent negative test. Since it is known that individuals who do not undergo regularly scheduled testing are typically infected closer to the first positive test than the last negative test [28], we scaled the prior distribution of the infection ages, using smaller means and standard deviations with more recent negative tests, then truncating the distribution at the earliest plausible time of infection, resulting in a prior distribution that is visually similar to the standard prior distribution, but with density skewed towards the first positive test. Specifically, if the difference between the first positive test date, T_{pos}, and two months before the last negative test date, T_{neg}, is smaller than the 95th percentile of the standard prior distribution, T_{95}, the mean and standard deviation of the prior for that individual are scaled by the ratio of those two times, $(T_{pos}-T_{neg}+2/12)/T_{95}$, where all times are in years. The distribution is then truncated at $(T_{pos}-T_{neg}+2/12)$ years before the first positive test for all individuals with a previous negative test regardless of whether scaling was required. With the numeric distributions for the infection times, we used *biophybreak* with 2×10^6 MCMC iterations (5×10^4 iterations of burn-in) on each transmission cluster. For comparison to when no biomarkers are used, we also ran the inference under the same conditions but with the 0 to 11 years uniform distribution for infection ages instead of the patient-specific distributions from the MBM.

Results

An improved multi-biomarker model for estimation of HIV-1 time of infection

Using the 30-patient training data, we modeled five biomarkers as linear-asymptotic trends for BED and LAg, and linear for *pol* polymorphism count, *pol* NGS diversity, and CD4 cell count (Fig 2). The biomarkers were combined into a mixed effects modeling framework to allow for

Table 1. Biomarker model performance.

Model	Mean Bias	MAE	RMSE
3 Biomarkers, 2019	-0.68	1.01	1.38
5 Biomarkers, 2021	-0.19	0.33	0.47
3 Biomarkers, 2021	-0.23	0.43	0.58
2 Biomarkers, 2021	-0.36	0.49	0.70

Table Footnote: All values are in years relative to actual time of infection. The 2019 model is from Giardina et al [21], shown for comparison; the 2021 models are those developed in this study. 5 biomarkers = BED, LAg, *pol* polymorphisms, *pol* NGS diversity, and CD4 cell count; 3 biomarkers = BED, *pol* polymorphisms, and CD4 cell count; 2 biomarkers = *pol* polymorphisms and CD4 cell count.

https://doi.org/10.1371/journal.pcbi.1009741.t001

patient specific variation and general trends. As expected, a shorter time between infection and sampling typically resulted in a posterior distribution with lower standard deviations, and longer time between infection and sampling resulted in more uncertainty about the time of infection.

Including all five biomarkers, we were able to substantially improve the performance over the previous 3-biomarker model (*pol* polymorphisms + CD4 + BED) used in Giardina et al [21], with 3-fold reductions in mean bias, mean absolute error (MAE), and root mean square error (RMSE) when comparing the medians of the inferred distributions in a cross-validation to the true values (Table 1). We also evaluated the performance of our modified 3-biomarker model (*pol* polymorphisms + CD4 + BED) as well as a 2-biomarker model (*pol* polymorphisms + CD4), which is of practical interest because *pol* polymorphism count and CD4 cell counts almost always exist in HIV-1 clinical databases. Our new 2- and 3-biomarker models also improved over the previous 3-biomarker model.

Biomarker information significantly improves transmission reconstruction

To investigate the expected accuracy of source identification with transmission history inference when using different numbers of biomarkers with a coalescent-based transmission model, we investigated 1,000 simulations with varying times between infection and sampling, and sampling different, possible phylogenies on a fixed transmission history with moderate values for transmission heterogeneity and temporal spacing (Fig 3). In these simulations we used only one sequence per host as that is the standard in clinical and public health databases. Note, however, that our phylogenetic framework models within-host diversity, which can be seen in the reconstructions involving multiple transmitted lineages and super-spreader activity.

Adding real biomarker information about time of infection significantly improves the accuracy in reconstructing transmission histories (Fig 4). We compared adding 2, 3, or 5 biomarkers, as well as fixed infection times, to phylogenetic information only. This is a non-trivial problem because node times in the virus phylogenies from epidemiologically linked patients cannot be assumed to be identical to infection times in the transmission histories among the patients (known as the "pre-transmission interval" [18]), nor can the topology of the transmission history be assumed to be identical to the sampled virus phylogeny [12,17,18,40]. Here, we investigated the overall expected probability to infer the correct source in each transmission among 15 patients. While the use of sequence data and phylogenetic reconstruction is much better than a random guess at 1/N, increasing from 6.7% expected accuracy to 30% with no biomarkers, we improved the transmission history inference over the phylogeny alone by on average 12 percentage points using the broadly available 2 biomarkers *pol* polymorphisms and

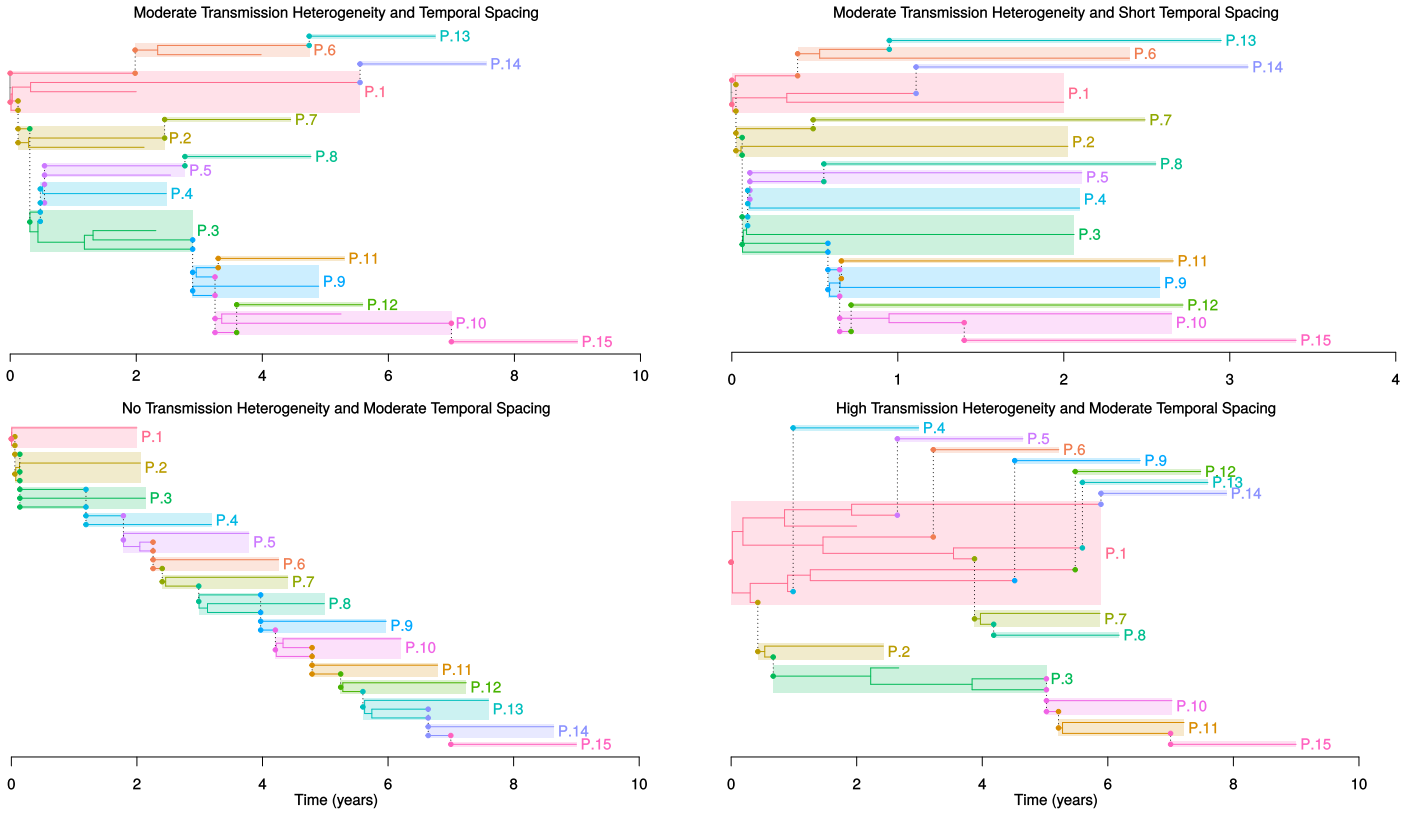

Fig 3. Examples of Simulated Transmission Histories. Four different possible transmission histories and phylogenies. As in Fig 1, each host is indicated by a colored box and transmissions are indicated by dashed vertical lines. Because transmission after diagnosis is not prohibited in these examples, it is possible that the right side of the box for some individuals to extend beyond the sampling time, in which case the right side of the box is the time of the last transmission and the sampling time is the point of termination of a lineage in the interior of the box. Note also that although we only sample one lineage here, the model takes within-host diversity into account, and thus can infer which lineage(s) within the host's diversity that was transmitted. The Top Left Panel shows a transmission history with moderate temporal spacing (0.5 years) and transmission heterogeneity (near 1). The Top Right Panel shows a transmission history with moderate transmission heterogeneity (near 1) and short temporal spacing (0.1 years). The Bottom Left Panel shows a transmission history with no transmission heterogeneity, leading to a straight transmission chain. The Bottom Right Panel shows a transmission history with high transmission heterogeneity involving a super-spreader (P.1).

https://doi.org/10.1371/journal.pcbi.1009741.g003

CD4 cell counts ($p < 1{\times}10^{-9}$, Wilcoxon signed rank test with Bonferroni multiple testing correction). Adding the 3-biomarker model improved the accuracy by >13 percentage points ($p < 4{\times}10^{-10}$) and adding all 5 biomarkers by 16 percentage points ($p < 4{\times}10^{-12}$). We investigated the theoretical limit of using biomarker data to our transmission history inference by adding effectively fixed infection times, which reached on average an accuracy improvement of 29 percentage points over the phylogeny alone. All improvements in accuracy were achieved by an increase in the model posterior prediction score (S1 Fig).

Overall, the *env* gene performed somewhat better than *pol* in the combined biomarker-phylogenetic inference of the transmission history. Part of the explanation likely lies in the fact that *env* evolves faster, thus accumulating more information about genealogical relationships, making the phylogenetic component more robust as previously shown [41,42].

Shorter temporal spacing and increased transmission heterogeneity reduce reconstruction accuracy, but larger clusters are not harder to get right

Heterogeneity in the number of transmitted lineages (phylogenetically separate virus variants), time between infection and onward transmission, time between infection and sampling, and in the number of onward infections a host causes (transmission degree) are all known to occur

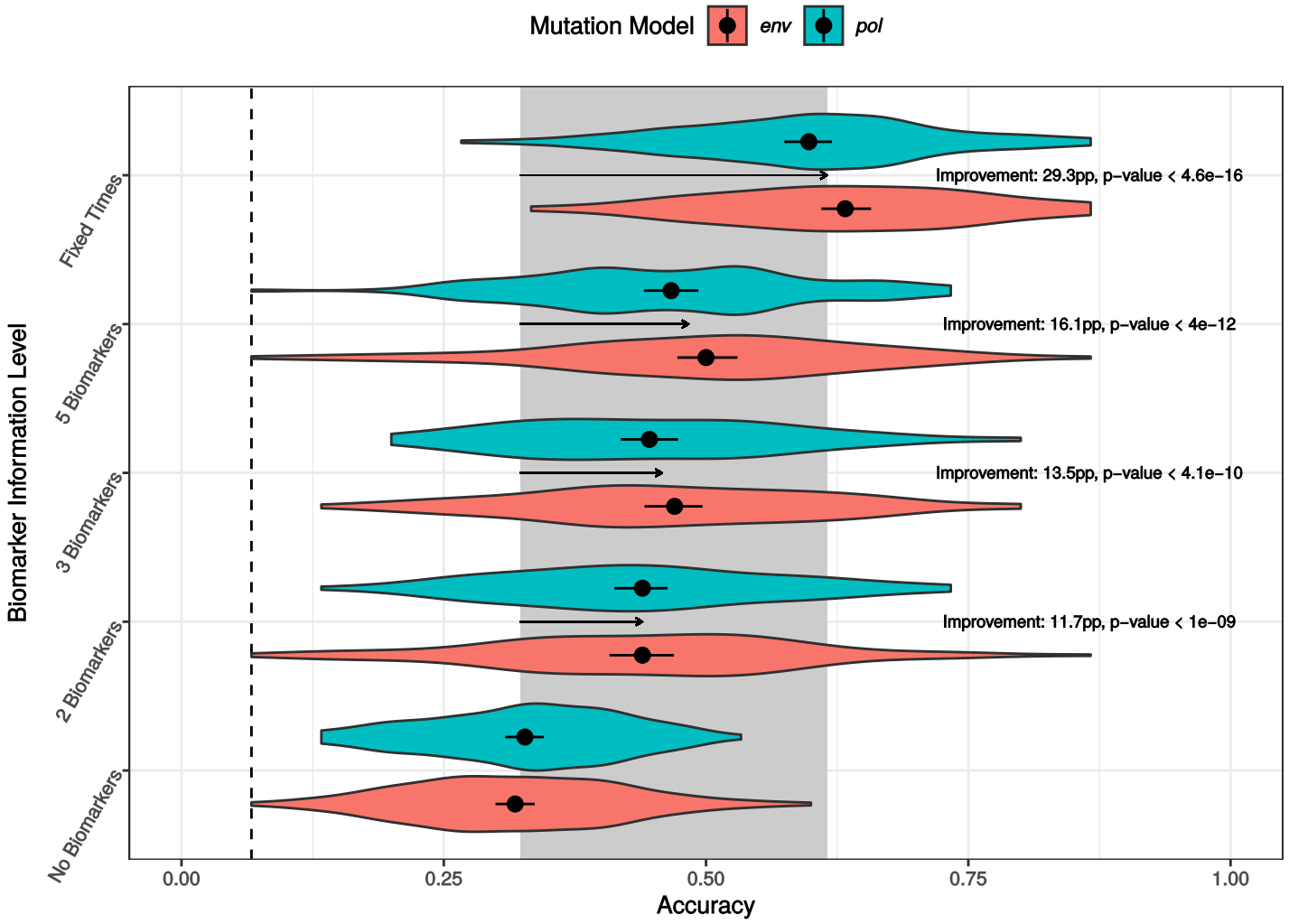

Fig 4. Transmission Inference Improvement with Biomarker Information. Violins show the full distribution of accuracy (proportion of the 15 individuals with the true infector correctly identified) on simulated data for each level of biomarker information and mutation model (genomic region), with the point and line segment in each violin representing the mean and 95% bootstrapped interval for the estimate of the mean. The vertical dashed line represents the random guess accuracy of 1/15. The gray shaded area in the background represents the region between the expected accuracy given effectively no information about infection times and given effectively fixed infection times, with the information level attainable with biomarkers falling between these two extremes. The improvement is shown with an arrow in percentage points (pp) and p value estimated by a Wilcoxon signed rank test with Bonferroni multiple testing correction.

https://doi.org/10.1371/journal.pcbi.1009741.g004

in real transmission histories. Therefore, using 5 biomarkers, we modeled all of these factors, as well as different sizes of transmission clusters, and assessed their effects on the accuracy of transmission history inference. These scenarios cover a wide range of fully sampled, possible, realistic HIV-1 transmission histories (Fig 3).

Under realistic HIV-1 evolutionary within-host parameters ($\alpha = 5, \beta = 5$), where about half of transmissions result in >1 transmitted lineage and within-host diversification is substantial [30], transmission history inference is expected to be quite challenging. For comparison, if only single lineages were transmitted and within-host diversification was very limited ($\alpha = 0, \beta = 0.1$), the overall accuracy reaches about 90% when infections were not close in time and transmission degree was 1 or less (Fig 5A). With realistic HIV-1 parameters, the overall accuracy was about 50% (Fig 5B).

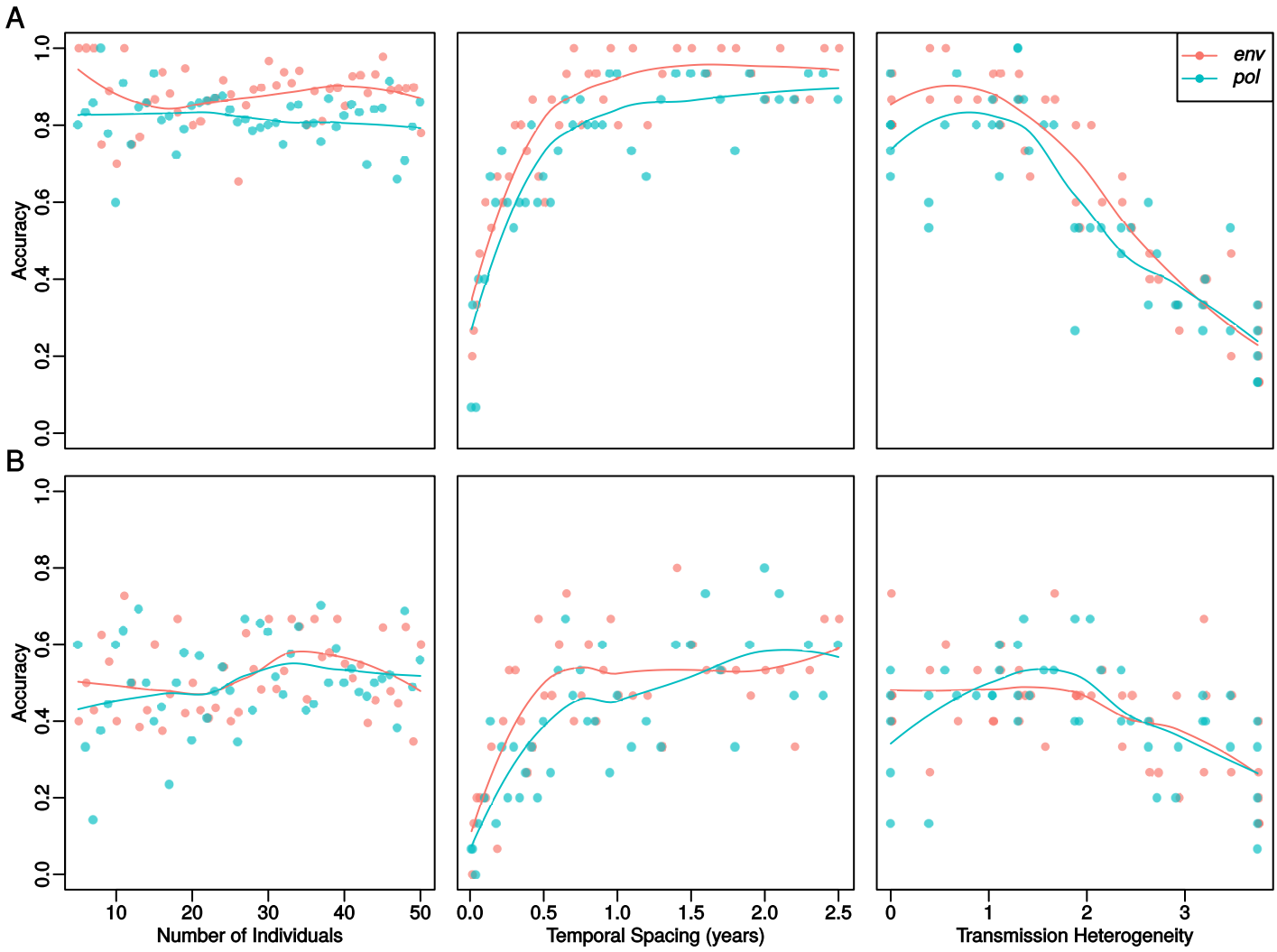

Fig 5. Individual Transmission Cluster Attribute Effects. Each point represents the accuracy of inference on a single transmission history, with the solid lines being the loess lines for each mutation model (*env* or *pol* genomic regions). (A) The top row shows simulation results from idealized situations with complete transmission bottlenecks and low within host diversity ($\alpha = 0$ and $\beta = 0.1$), one year between infection and sampling, and fixed biomarkers. (B) The bottom row shows results from simulations with realistic HIV parameters in terms of transmission bottleneck size variation and within host diversity ($\alpha = 5$ and $\beta = 5$), two years between infection and sampling, and independent biomarkers. In each panel, the non-varied attributes used are 15 individuals, temporal spacing of 0.5 years, and transmission heterogeneity of about 1.0.

https://doi.org/10.1371/journal.pcbi.1009741.g005

The shorter the temporal spacing of transmission events, the harder it was to infer the correct transmission history. Although biomarkers improved the reconstruction accuracy even at very short temporal spacing of only a few days, raising the accuracy over a random guess at 1/N, meaningful accuracy started when the temporal spacing was above a few months. This is because biomarker posterior distributions will greatly overlap when times between transmission are short, making it difficult to order the events in time. With short times between many infections, there was also little time to accumulate mutations that would inform the phylogenetic reconstruction.

Higher degrees of transmission heterogeneity, like in panel 4 in Fig 3, on average also lead to more difficult transmission history inference. At degree levels above 2.5, the average accuracy decreased from about 50% to 25%. Compared to the overall performances and limits in

Fig 4, this reduction was as severe as not having any biomarkers, and clearly constitutes a very difficult to resolve epidemiological situation. The most difficult scenarios were those with short temporal spacing and high transmission heterogeneity. Thus, super-spreader activity can cause particularly difficult to reconstruct epidemiological scenarios where biomarkers may not always help enough to resolve the transmission history.

Most combinations of attributes showed only small amounts of interaction effects. The most notable exception was the temporal spacing and biomarker information level, which had a combined effect on the quality of inference when marginalizing over the other three variables (S2 Fig). With short temporal spacing, the biomarker information offered only small improvements. Having near perfect information about the infection times (the fixed infection time scenario), however, would allow a very large improvement. As the temporal spacing increases, the improvement with better biomarker information increased as well, while also approaching the fixed infection time scenario. This is because as the temporal spacing increases, the infection time distributions become more separated, allowing greater certainty about the infection order. In this way, longer temporal spacing helps both in terms of making the phylogeny more informative and helping the biomarkers to allow more separation.

Promisingly, transmission histories involving more hosts were on average not harder to reconstruct at fixed levels of transmission heterogeneity and temporal spacing (Fig 5). This is encouraging for real-time applications that follow the growth of a public health database because it cannot be known beforehand how many persons that eventually will be part of a transmission cluster. Also encouraging was that *env* and *pol* performed similar in these simulations, as public health databases typically store *pol*, but not *env*, sequences from drug resistance testing.

Additional sequences from hosts improve overall transmission reconstruction

Beyond simply having more data, the conceptual motivation for using >1 sequence/host is that it should increase the chance that the sampled lineages from a recipient will coalesce with at least one of the sampled lineages from the source rather than earlier in the transmission history (S3 Fig). Although the computational burden increased when adding more sequence data per infected host, the accuracy in the transmission history reconstruction did indeed improve. Using 4 sequences instead of 1 sequence per host in 3-person transmission histories showed a 7.4 percentage point (12.2%) improvement in accuracy (on average 0.073 posterior probability (14.9%) improvement; $p < 2.5\times10^{-7}$, Wilcoxon signed rank test) (S4 Fig). While encouraging for future analyses with richer sequence data, further development of computational efficiency will be needed to exploit this enhancement.

The overall accuracy is not significantly affected by incomplete sampling

While the "first" person's source will always be missing, and recipients that have not infected anybody may also be irrelevant to the transmission-history-reconstruction-problem, the problem of missing intermediary links is always a possibility. Thus, in real-life situations it is never known if the sample is complete or not, i.e., one cannot be sure that all *relevant* sources have been sampled. Therefore, we investigated the situation when an intermediary source was missing in transmission histories with 10 hosts, using four levels of transmission heterogeneity (S5 Fig). When missing, we defined accurate source identification as the missing source's source.

The differences in accuracy between the completely and incompletely sampled transmission histories were in general relatively small (Fig 6). In terms of the overall accuracy of the inference, the absolute difference in mean accuracy between the completely and incompletely

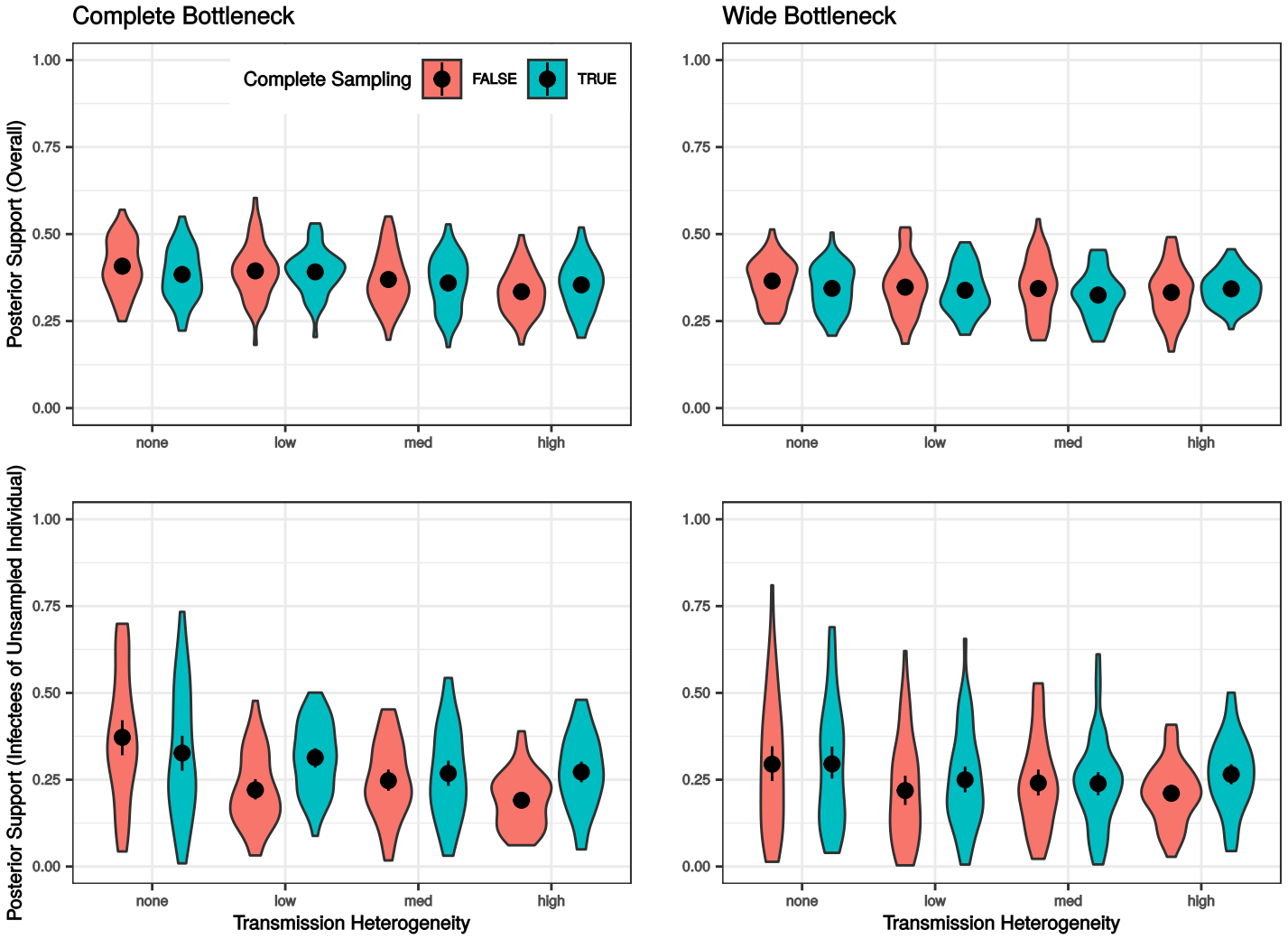

Fig 6. The Effect of Incomplete Sampling. Violins show the distribution of means of the posterior probability for true infectors (sources), or the true infectors' infectors if the true infector is unsampled, with the point and line segment in each violin representing the mean and 95% bootstrapped interval for the estimate of the mean. (Top Row) Overall model performance. (Bottom Row) Model performance for the individuals infected by the unsampled individual.

https://doi.org/10.1371/journal.pcbi.1009741.g006

sampled transmission clusters was less than 2.6 percentage points for any combination of bottleneck size and transmission heterogeneity (all raw p-values > 0.05, Wilcoxon signed-rank tests).

Focusing on the accuracy of assigning the source to the recipients infected by the unsampled source, naturally, was the most challenging. Three out of the eight combinations of bottleneck size and transmission heterogeneity showed moderate drops in accuracy, as well as significant p-values while the other five had only small and insignificant differences. The three situations with moderate differences were the low and high transmission heterogeneity scenarios with complete bottlenecks and the high transmission heterogeneity scenario with the wide bottleneck, with the differences in accuracy at 20.0, 11.2, and 8.4 percentage points, respectively (raw p-values at 0.002, 0.015, and 0.046, Wilcoxon signed-rank test).

Although there were some scenarios where the accuracy of inference for individuals infected by an unsampled individual was lower, since most differences are small and the

overall accuracy remained at the same level, these results demonstrate that it is appropriate to apply this method even when there might be unsampled infectors in a transmission cluster.

Application to real transmission clusters

To be able to interpret the results effectively when the true history is not known (i.e., using real data), we need a way to link the results of the inference directly to expected accuracy of source identification for individuals. To do this, we use the data from the effect of biomarker information trials, calculating the proportion of times when potential infectors in a certain range of posterior probability values were the true infectors for all potential infectors of each individual in each trial (Fig 7). Generally, the actual probability of being the true infector was only slightly less than the inferred posterior probability. For example, infectors with posterior probability values between 0.45 and 0.55 were the true infector in 45 percent of the actual transmissions. For these analyses, we find the maximum parent credibility (MPC) transmission history, the sampled transmission history that has the highest product of posterior probabilities of the infectors. Note that since the transmission history must have been sampled at least once and must be connected and acyclic, the infectors predicted from the complete MPC transmission history may not be the highest posterior probability infectors for each individual.

We applied our method to four transmission clusters from the general Swedish HIV-1 epidemic [17,38,39]. The data included direct population *pol* gene sequences [25], determined as part of clinical drug resistance testing, BED and CD4 T cell counts, occasionally previous negative tests, and date of sampling. The sequence data was used for phylogenetic inference as well as a biomarker of within-host divergence (*pol* polymorphisms). In all four clusters, the mean of the posterior probability for the highest posterior probability infectors for each individual is higher when using biomarkers than without ($p = 0.008$, Wilcoxon signed rank test) (Table 2). Importantly, about half of the predicted highest posterior probability infectors for each individual changed when adding the biomarker information of infection ages.

Fig 8 shows an inferred chain of four sampled hosts infecting each other serially (Fig 8A) and a more complex transmission history involving 14 sampled hosts that included superspreading (Fig 8B). When longer time from infection to transmission occurred and the biomarker density was narrow, the posterior support for source assignment was high, e.g., the source of P.85.1480 is assigned to P.85.1368 at 0.99 posterior support, who transmitted (at least) two HIV-1 lineages (Fig 8A). Conversely, when there were short time intervals between transmissions and biomarker densities overlap, transmission reconstruction became more difficult, e.g., the source of P.85.1173 was more evenly attributed to 3 out of 4 sampled hosts in the corresponding transmission cluster (Fig 8A), and, similarly, assigning a source to P.24.323 was less certain in the larger cluster (Fig 8B). Because phylogeny, biomarkers, and sampling times interact in non-trivial ways, however, relatively large posterior probabilities may be assigned to one source over many others, e.g., in the transmission to P.24.909, P.24.859 was significantly more likely to be the source than any other sampled source in that cluster (Fig 8B). S6 and S7 Figs show two additional transmission clusters with transmission heterogeneity (degrees 1.7 and 1.2), overlapping biomarkers, and both short overall time (6 transmissions in <1 year) and longer time (8 transmissions over 8 years). These clusters provide further examples of real situations where some source assignments were easier and others harder.

Discussion

In this study we have developed a biomarker-enhanced phylogenetic framework to allow for more accurate inference of pathogen transmission histories. The biomarker component used five real HIV-1 biomarkers from a set of untreated, longitudinally followed HIV-1 infected

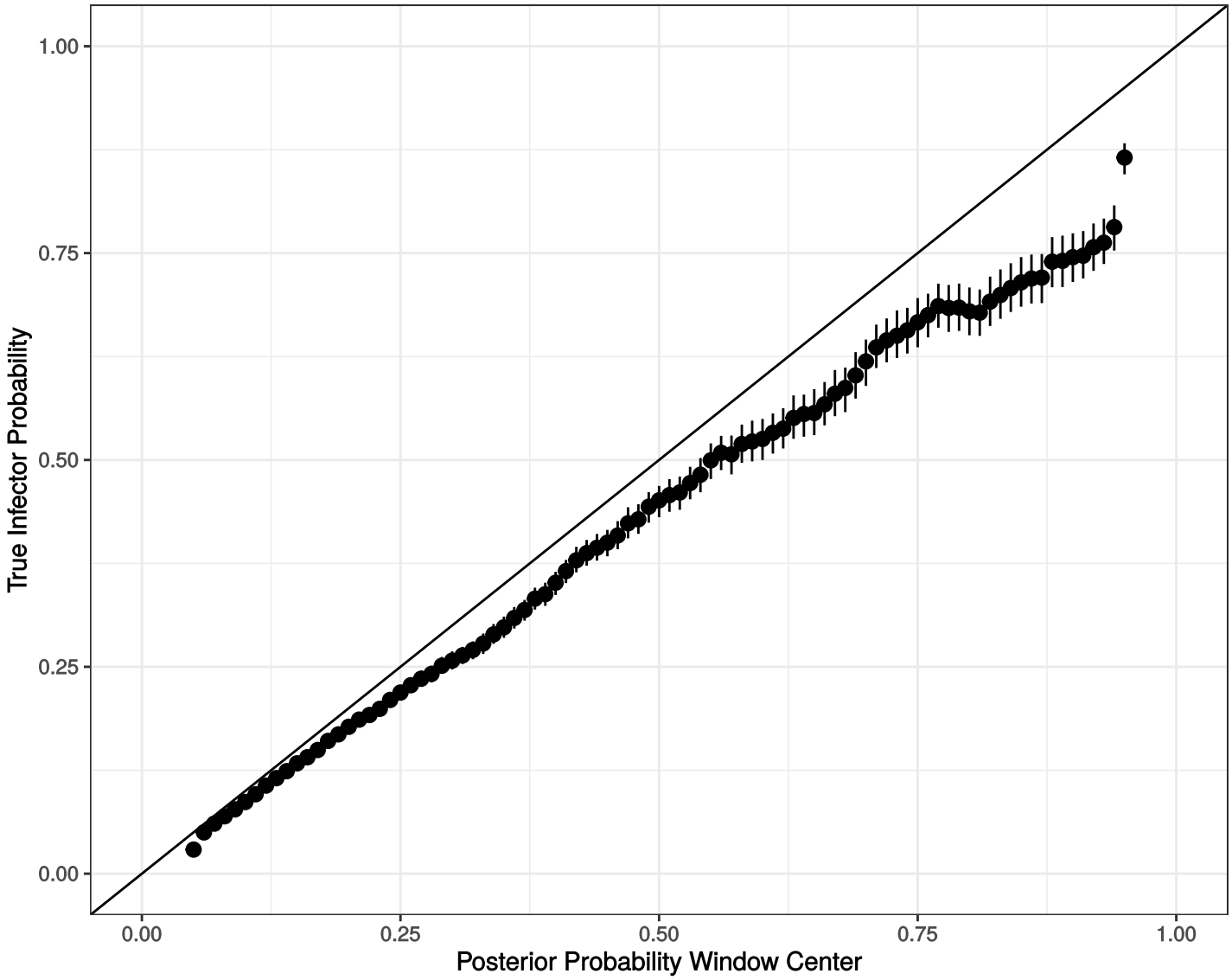

Fig 7. Relationship Between True Infector Probability and Inferred Posterior Probability. Dots with error bars show the probability that a potential infector is the true infector of an individual, given that the inferred posterior probability for that potential infector is within a window of width 0.10 centered at a certain value. The $y = x$ line is shown for comparison. Data for this figure is from the 1000 trials used for testing the effect of biomarker information. Error bars indicate the 95% bootstrapped confidence interval for each window.

https://doi.org/10.1371/journal.pcbi.1009741.g007

Table 2. Performance of inference on real data.

Cluster ID	Mean Posterior Probability (no biomarkers)	Mean Posterior Probability (with biomarkers)	Proportion of Differences in Predicted Infectors
7	0.18	0.39	0.67 (4 of 6)
24	0.42	0.59	0.43 (6 of 14)
79	0.37	0.47	0.50 (4 of 8)
85	0.60	0.70	0.50 (2 of 4)

Table Footnote: Cluster 7 is shown in S6 Fig, cluster 24 in Fig 8B, cluster 79 in S7 Fig, and cluster 85 in Fig 8B.

https://doi.org/10.1371/journal.pcbi.1009741.t002

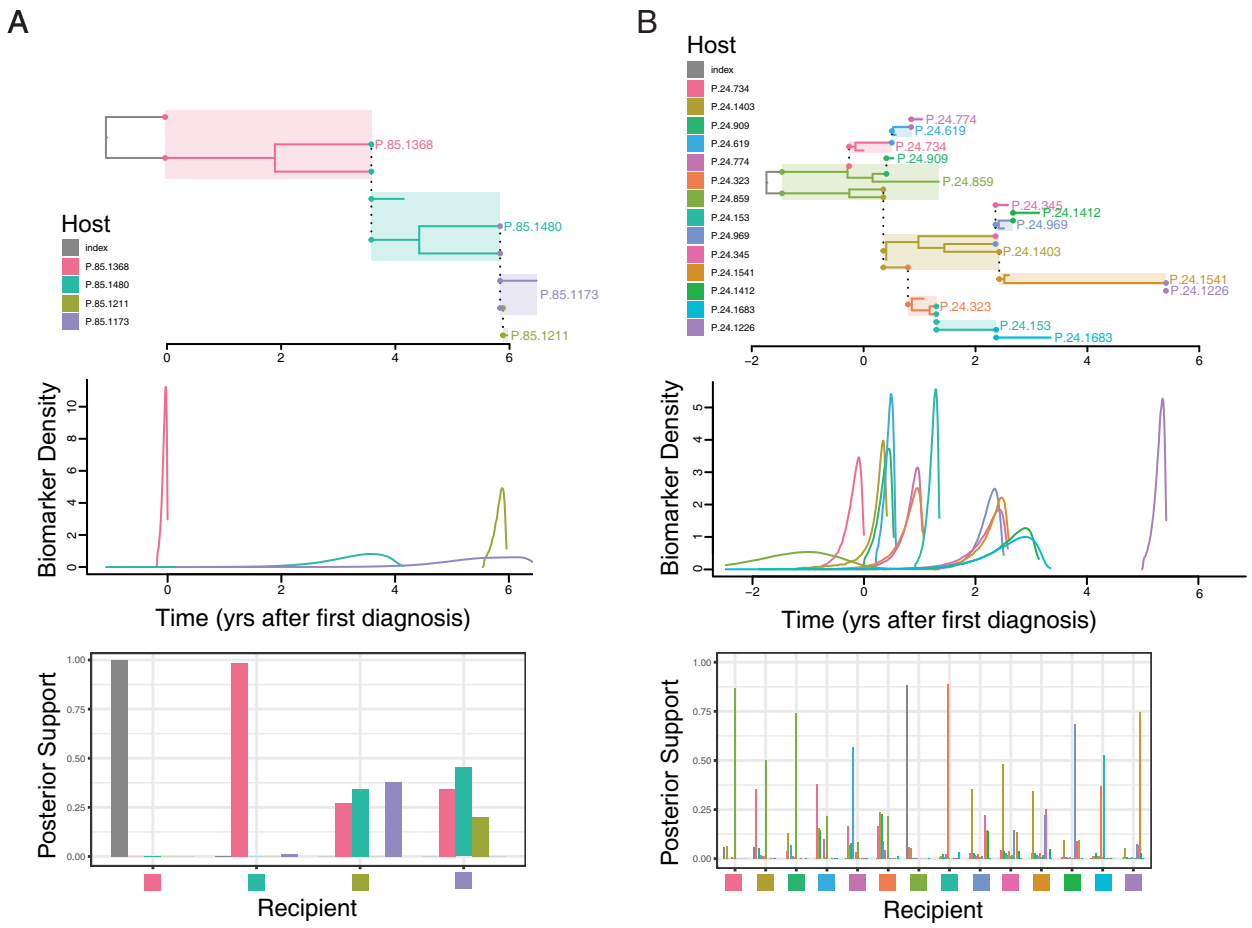

Fig 8. Transmission History Inference on Real HIV Transmission Clusters. Examples of a smaller, simple transmission history (A) and a larger, complex transmission history (B) from the Swedish HIV epidemic. The Top Panel in each transmission cluster shows the inferred maximum parent credibility tree. The Middle Panel shows the distributions of infection times inferred from biomarker values for each individual, using our 3-biomarker model applied to data that existed in a public health database. The Bottom Panel shows posterior support for each individual to be the source for each individual, with each colored square on the x-axis representing one individual and the height of the colored bars represent the posterior support for the corresponding individual to be their infector.

https://doi.org/10.1371/journal.pcbi.1009741.g008

patients. Thus, our results display practical and realistic improvement of the expected accuracy in the inference of HIV-1 transmission histories using data that is typically available in a HIV surveillance system. We investigated a wide variety of transmission scenarios, including heterogeneity in the number of transmitted lineages, in the time between infection and onward transmission, in the time between infection and sampling, in the number of onward infections a host causes, and among different sizes of transmission clusters. Overall, we show that adding biomarkers to the transmission history inference substantially improved accuracy in all the considered scenarios.

Compared to previous phylodynamic methods that infer transmission history and direction [4–7,10], our method includes real biomarker-informed time of transmission and allows for wide transmission bottlenecks. Kenah et al previously showed that if the relative order of transmissions is known, transmission history inference should improve [20]. Here, we show that real HIV biomarkers approach this ideal situation, where using more biomarkers is better than fewer, but it is unlikely that we will ever find biomarkers that can resolve all situations. Likewise, factors such as within-host diversity of the virus, the fact that HIV transmission largely is

a random draw of a few variants [17], transmission often involves >1 phylogenetic lineage [30], and that within-host evolution involves lineage death and birth [12,43,44], together put theoretical limits on how accurately we can infer the underlying transmission history from a phylogeny. Note also that *biophybreak* does not consider any kind of super-infection, i.e., it is assumed that a recipient is infected once by one infector.

While our biomarker-enhanced phylogenetic method generally improved transmission history inference, one should not expect such a method to reach 100% accuracy. Here, we show that even if the biomarkers could provide perfect transmission times, having realistic levels of within-host virus diversity induces substantial uncertainty that limits the average possible performance to 40–87% accuracy (95% posterior probability interval) assuming moderate temporal spacing and transmission heterogeneity. Additionally, the infection age distribution inference utilizes a population-appropriate prior distribution for the amount of time between infection and diagnosis based on previously known dynamics of how the population in question behaves [21]. If such dynamics were unknown, a less informative prior would need to be used. There are also certain situations that are particularly difficult to accurately reconstruct. When transmission histories involve large transmission heterogeneity, typically when super-spreader activity has occurred, it becomes difficult to reconstruct all transmission events accurately. This is in part because each time the super-spreader transmits, a random draw of variants is transmitted and thus the phylogenetic ordering of coalescences typically does not follow the transmission order [12,17]. Furthermore, two hosts infected close in time to each other may receive more similar virus than is later sampled in the source, and thus may appear to be linked to each other rather than to the source. This complication is compounded when many transmissions happened over a short time. It is possible that prior identification of super-spreader activity can identify when and where in a phylogeny these problems exist [45].

Another uncertainty is related to the fact that we never have a perfect sample from an ongoing epidemic, meaning that, at any time, we have not yet sampled every actual or soon to be transmitter. This is highlighted by the fact that many populations still are far from the WHO/UNAIDS 90-90-90 goal [46,47], and even in nations where that goal has been reached it takes on average 2 years to detect most infections [48], although some risk groups show shorter or longer times from infection to diagnosis. Here, we show that our biomarker enhanced phylogenetic framework can handle missing links quite well, typically identifying a missing source's source as the origin of the transmission. Also, note that missing individuals that have not infected anyone else leave no trace in a phylogeny. Therefore, a missing link refers to any unsampled person ancestral to the set of sampled persons. While the existence of missing individuals who have not infected any sampled individuals is certainly a concern from a public health perspective, this is not something that can be determined from the available sequence or biomarker data without making substantial assumptions about the expected number of new individuals infected per host.

Phylogenetic reconstruction of transmission histories is a powerful and scientifically sound method because it is objective, can evaluate alternative hypotheses, and, as we show here, can be augmented with additional data. Because HIV infection still causes stigma and legal risks in some jurisdictions, however, both research and public health projects that use such methodology must be ethically justified on the basis of providing public health benefits [49,50]. Here, we show that phylogenetic methods can be made more accurate by adding biomarker data on time of infection. Accuracy is important because it allows public health resources to be directed to where they are needed most, and thus will have the largest reduction in disease spread [51]. Again, we emphasize that it can never reach 100% certainty, typically much less, yet the levels we can reach with the proposed methodology should make public health efforts more efficient.

Improved HIV surveillance, source attribution, and outbreak response depends on advances in HIV prevention, diagnosis, and continuous treatment. Application and further development of the technology presented here could allow for better prevention programs focusing on locally informed and tailored strategies.

Supporting information

S1 Fig. Relationship Between Cluster Accuracy and Posterior Probability of True Infectors. Accuracy and mean true posterior probability values for each of the 1000 trials used while testing the effect of biomarker information shown with the $y = x$ line for comparison. Note that since all trials are on clusters with 15 individuals, the accuracy for each trial can only be one of sixteen possible values, while the means of the true posterior probability are effectively continuous.
(EPS)

S2 Fig. Combined Transmission Cluster Attribute Effects. Violins show the distribution of accuracy for each level of biomarker information for each amount of temporal spacing, with the point and line segment in each violin representing the mean and 95% bootstrapped interval for the estimate of the mean.
(EPS)

S3 Fig. Conceptual Motivation for Additional Sequences per Individual. (Top) A transmission history with 4 sampled sequences per individual. (Bottom) The same history subsampled to 1 sequence per individual.
(EPS)

S4 Fig. Inference with Multiple Sequences. Violins show the distribution of mean true posterior support for 4 and 1 sequence(s) per individual for each mutation model, with the point and line segment in each violin representing the mean and 95% bootstrapped interval for the estimate of the mean.
(EPS)

S5 Fig. Transmission Histories used in Incomplete Sampling Tests. We test the effect of an unsampled individual with four different levels of transmission heterogeneity. The 6^{th} individual is removed in the scenarios of no and low transmission heterogeneity, the 5^{th} individual is removed in the scenario of moderate transmission heterogeneity, and the 3^{rd} individual is removed in the scenario of high transmission heterogeneity.
(EPS)

S6 Fig. Real Transmission Cluster 7. (Top) Inferred maximum parent credibility trees for each transmission cluster. (Middle) Distributions of infection times inferred from biomarker values for each individual. (Bottom) Posterior support for each individual to be the source for each individual, with the height of the colored bars represent the posterior support for the corresponding individual to be their infector.
(PDF)

S7 Fig. Real Transmission Cluster 79. (Top) Inferred maximum parent credibility trees for each transmission cluster. (Middle) Distributions of infection times inferred from biomarker values for each individual. (Bottom) Posterior support for each individual to be the source for each individual, with the height of the colored bars represent the posterior support for the corresponding individual to be their infector.
(PDF)

S1 Data. We provide data and codes for all simulation and biomarker plots in the supplemental file "Figure_data.zip".
(ZIP)

Acknowledgments

We are grateful for helpful discussions with Dr. Don Klinkenberg, and for his unpublished modifications to the R package *phybreak* that allow for wide transmission bottlenecks. We are also grateful to the HIV infected study participants and to Johanna Brodin, Fabio Zanini, Lina Thebo, Eva Eriksson, Christa Lanz, and Göran Bratt for important contributions in the generation of the published data that was analyzed in this study. We thank Helena Skar, Federica Giardina, Tom Britton, Vadim Puller, and Richard Neher for previous developments of the theoretical models of biomarker-based estimation of HIV infection times.

Author Contributions

Conceptualization: Erik Lundgren, Thomas Leitner.

Methodology: Erik Lundgren, Ethan Romero-Severson, Thomas Leitner.

Resources: Jan Albert.

Software: Erik Lundgren.

Writing – original draft: Erik Lundgren, Thomas Leitner.

Writing – review & editing: Erik Lundgren, Ethan Romero-Severson, Jan Albert, Thomas Leitner.

References

1. (CDC) CfDCaP. HIV PREVENTION IN THE UNITED STATES: MOBILIZING TO END THE EPIDEMIC: Centers for Disease Control and Prevention (CDC); 2021 [2021-11-05]. Available from: https://www.cdc.gov/hiv/pdf/policies/cdc-hiv-prevention-bluebook.pdf.
2. Wertheim JO, Kosakovsky Pond SL, Forgione LA, Mehta SR, Murrell B, Shah S, et al. Social and Genetic Networks of HIV-1 Transmission in New York City. PLoS Pathog. 2017; 13(1):e1006000. https://doi.org/10.1371/journal.ppat.1006000 PMID: 28068413; PubMed Central PMCID: PMC5221827.
3. Resik S, Lemey P, Ping LH, Kouri V, Joanes J, Perez J, et al. Limitations to contact tracing and phylogenetic analysis in establishing HIV type 1 transmission networks in Cuba. AIDS Res Hum Retroviruses. 2007; 23(3):347–56. https://doi.org/10.1089/aid.2006.0158 PMID: 17411367.
4. Didelot X, Fraser C, Gardy J, Colijn C. Genomic infectious disease epidemiology in partially sampled and ongoing outbreaks. Mol Biol Evol. 2017. https://doi.org/10.1093/molbev/msw275 PMID: 28100788.
5. Jombart T, Cori A, Didelot X, Cauchemez S, Fraser C, Ferguson N. Bayesian reconstruction of disease outbreaks by combining epidemiologic and genomic data. PLoS Comput Biol. 2014; 10(1):e1003457. https://doi.org/10.1371/journal.pcbi.1003457 PMID: 24465202; PubMed Central PMCID: PMC3900386.
6. Klinkenberg D, Backer JA, Didelot X, Colijn C, Wallinga J. Simultaneous inference of phylogenetic and transmission trees in infectious disease outbreaks. PLoS Comput Biol. 2017; 13(5):e1005495. Epub 2017/05/26. https://doi.org/10.1371/journal.pcbi.1005495 PMID: 28545083; PubMed Central PMCID: PMC5436636.
7. De Maio N, Wu CH, Wilson DJ. SCOTTI: Efficient Reconstruction of Transmission within Outbreaks with the Structured Coalescent. PLoS Comput Biol. 2016; 12(9):e1005130. Epub 2016/09/30. https://doi.org/10.1371/journal.pcbi.1005130 PMID: 27681228; PubMed Central PMCID: PMC5040440.
8. Ypma RJ, van Ballegooijen WM, Wallinga J. Relating phylogenetic trees to transmission trees of infectious disease outbreaks. Genetics. 2013; 195(3):1055–62. https://doi.org/10.1534/genetics.113.154856 PMID: 24037268; PubMed Central PMCID: PMC3813836.

9. Wymant C, Hall M, Ratmann O, Bonsall D, Golubchik T, de Cesare M, et al. PHYLOSCANNER: Inferring Transmission from Within- and Between-Host Pathogen Genetic Diversity. Mol Biol Evol. 2017. Epub 2017/12/01. https://doi.org/10.1093/molbev/msx304 PMID: 29186559; PubMed Central PMCID: PMC5850600.
10. Skums P, Zelikovsky A, Singh R, Gussler W, Dimitrova Z, Knyazev S, et al. QUENTIN: reconstruction of disease transmissions from viral quasispecies genomic data. Bioinformatics. 2018; 34(1):163–70. Epub 2018/01/06. https://doi.org/10.1093/bioinformatics/btx402 PMID: 29304222.
11. De Maio N, Worby CJ, Wilson DJ, Stoesser N. Bayesian reconstruction of transmission within outbreaks using genomic variants. PLoS Comput Biol. 2018; 14(4):e1006117. Epub 2018/04/19. https://doi.org/10.1371/journal.pcbi.1006117 PMID: 29668677; PubMed Central PMCID: PMC5927459.
12. Leitner T. Phylogenetics in HIV transmission: taking within-host diversity into account. Curr Opin HIV AIDS. 2019; 14(3):181–7. Epub 2019/03/29. https://doi.org/10.1097/COH.0000000000000536 PMID: 30920395; PubMed Central PMCID: PMC6449181.
13. Wang HY, Chien MH, Huang HP, Chang HC, Wu CC, Chen PJ, et al. Distinct hepatitis B virus dynamics in the immunotolerant and early immunoclearance phases. J Virol. 2010; 84(7):3454–63. Epub 2010/01/22. https://doi.org/10.1128/JVI.02164-09 PMID: 20089644; PubMed Central PMCID: PMC2838120.
14. Poon AF, Kosakovsky Pond SL, Bennett P, Richman DD, Leigh Brown AJ, Frost SD. Adaptation to human populations is revealed by within-host polymorphisms in HIV-1 and hepatitis C virus. PLoS Pathog. 2007; 3(3):e45. Epub 2007/04/03. https://doi.org/10.1371/journal.ppat.0030045 PMID: 17397261; PubMed Central PMCID: PMC1839164.
15. Didelot X, Walker AS, Peto TE, Crook DW, Wilson DJ. Within-host evolution of bacterial pathogens. Nature reviews Microbiology. 2016; 14(3):150–62. Epub 2016/01/26. https://doi.org/10.1038/nrmicro.2015.13 PMID: 26806595; PubMed Central PMCID: PMC5053366.
16. Worby CJ, Lipsitch M, Hanage WP. Within-host bacterial diversity hinders accurate reconstruction of transmission networks from genomic distance data. PLoS Comput Biol. 2014; 10(3):e1003549. Epub 2014/03/29. https://doi.org/10.1371/journal.pcbi.1003549 PMID: 24675511; PubMed Central PMCID: PMC3967931.
17. Romero-Severson E, Skar H, Bulla I, Albert J, Leitner T. Timing and Order of Transmission Events Is Not Directly Reflected in a Pathogen Phylogeny. Mol Biol Evol. 2014; 31(9):2472–82. https://doi.org/10.1093/molbev/msu179 PMID: 24874208.
18. Leitner T, Albert J. The molecular clock of HIV-1 unveiled through analysis of a known transmission history. Proc Natl Acad Sci USA. 1999; 96:10752–7. https://doi.org/10.1073/pnas.96.19.10752 PMID: 10485898
19. Hall MD, Colijn C. Transmission Trees on a Known Pathogen Phylogeny: Enumeration and Sampling. Mol Biol Evol. 2019; 36(6):1333–43. Epub 2019/03/16. https://doi.org/10.1093/molbev/msz058 PMID: 30873529; PubMed Central PMCID: PMC6526902.
20. Kenah E, Britton T, Halloran ME, Longini IM Jr., Molecular Infectious Disease Epidemiology: Survival Analysis and Algorithms Linking Phylogenies to Transmission Trees. PLoS Comput Biol. 2016; 12(4):e1004869. https://doi.org/10.1371/journal.pcbi.1004869 PMID: 27070316; PubMed Central PMCID: PMC4829193.
21. Giardina F, Romero-Severson EO, Axelsson M, Svedhem V, Leitner T, Britton T, et al. Getting more from heterogeneous HIV-1 surveillance data in a high immigration country: estimation of incidence and undiagnosed population size using multiple biomarkers. Int J Epidemiol. 2019. Epub 2019/05/11. https://doi.org/10.1093/ije/dyz100 PMID: 31074780.
22. Giardina F, Romero-Severson EO, Albert J, Britton T, Leitner T. Inference of Transmission Network Structure from HIV Phylogenetic Trees. PLoS Comput Biol. 2017; 13(1):e1005316. https://doi.org/10.1371/journal.pcbi.1005316 PMID: 28085876.
23. Dobbs T, Kennedy S, Pau CP, McDougal JS, Parekh BS. Performance characteristics of the immunoglobulin G-capture BED-enzyme immunoassay, an assay to detect recent human immunodeficiency virus type 1 seroconversion. J Clin Microbiol. 2004; 42(6):2623–8. Epub 2004/06/09. https://doi.org/10.1128/JCM.42.6.2623-2628.2004 PMID: 15184443; PubMed Central PMCID: PMC427871.
24. Duong YT, Qiu M, De AK, Jackson K, Dobbs T, Kim AA, et al. Detection of recent HIV-1 infection using a new limiting-antigen avidity assay: potential for HIV-1 incidence estimates and avidity maturation studies. PLoS ONE. 2012; 7(3):e33328. Epub 2012/04/06. https://doi.org/10.1371/journal.pone.0033328 PMID: 22479384; PubMed Central PMCID: PMC3314002.
25. Leitner T, Halapi E, Scarlatti G, Rossi P, Albert J, Fenyö EM, et al. Analysis of heterogeneous viral populations by direct DNA sequencing. BioTechniques. 1993; 15:120–6. PMID: 8363827
26. Kouyos RD, von Wyl V, Yerly S, Boni J, Rieder P, Joos B, et al. Ambiguous nucleotide calls from population-based sequencing of HIV-1 are a marker for viral diversity and the age of infection. Clin Infect Dis.

2011; 52(4):532–9. Epub 2011/01/12. https://doi.org/10.1093/cid/ciq164 PMID: 21220770; PubMed Central PMCID: PMC3060900.

27. Puller V, Neher R, Albert J. Estimating time of HIV-1 infection from next-generation sequence diversity. PLoS Comput Biol. 2017; 13(10):e1005775. Epub 2017/10/03. https://doi.org/10.1371/journal.pcbi.1005775 PMID: 28968389; PubMed Central PMCID: PMC5638550.

28. Skar H, Albert J, Leitner T. Towards estimation of HIV-1 date of infection: a time-continuous IgG-model shows that seroconversion does not occur at the midpoint between negative and positive tests. PLoS ONE. 2013; 8(4):e60906. Epub 2013/04/25. https://doi.org/10.1371/journal.pone.0060906 PMID: 23613753; PubMed Central PMCID: PMC3628711.

29. Plummer M. rjags, Bayesian Graphical Models using MCMC. 4–12 ed: CRAN; 2021.

30. Leitner T, Romero-Severson E. Phylogenetic patterns recover known HIV epidemiological relationships and reveal common transmission of multiple variants. Nat Microbiol. 2018; 3(9):983–8. Epub 2018/08/01. https://doi.org/10.1038/s41564-018-0204-9 PMID: 30061758.

31. Leitner T, Kumar S, Albert J. Tempo and mode of nucleotide substitutions in gag and env gene fragments in human immunodeficiency virus type 1 populations with a known transmission history. J Virol. 1997; 71:4761–70 (see also correction 1998: 72; 2565). https://doi.org/10.1128/JVI.71.6.4761-4770.1997 PMID: 9151870

32. Wawer MJ, Gray RH, Sewankambo NK, Serwadda D, Li X, Laeyendecker O, et al. Rates of HIV-1 transmission per coital act, by stage of HIV-1 infection, in Rakai, Uganda. J Infect Dis. 2005; 191(9):1403–9. Epub 2005/04/06. https://doi.org/10.1086/429411 PMID: 15809897.

33. Volz EM, Ionides E, Romero-Severson EO, Brandt MG, Mokotoff E, Koopman JS. HIV-1 Transmission during Early Infection in Men Who Have Sex with Men: A Phylodynamic Analysis. PLoS Med. 2013; 10(12):e1001568. Epub 2013/12/18. https://doi.org/10.1371/journal.pmed.1001568 PMID: 24339751; PubMed Central PMCID: PMC3858227.

34. Cohen MS, Gay CL. Treatment to prevent transmission of HIV-1. Clinical infectious diseases: an official publication of the Infectious Diseases Society of America. 2010; 50 Suppl 3:S85–95. Epub 2010/04/20. https://doi.org/10.1086/651478 PMID: 20397961.

35. Romero-Severson EO, Bulla I, Leitner T. Phylogenetically resolving epidemiologic linkage. Proc Natl Acad Sci U S A. 2016; 113(10):2690–5. https://doi.org/10.1073/pnas.1522930113 PMID: 26903617.

36. Ho DD, Neumann AU, Perelson AS, Chen W, Leonard JM, Markowitz M. Rapid turnover of plasma virions and CD4 lymphocytes in HIV-1 infection. Nature. 1995; 373(123–126). https://doi.org/10.1038/373123a0 PMID: 7816094

37. Rambaut A, Grassly NC. Seq-Gen: an application for the Monte Carlo simulation of DNA sequence evolution along phylogenetic trees. Comput Appl Biosci. 1997; 13(3):235–8. Epub 1997/06/01. https://doi.org/10.1093/bioinformatics/13.3.235 PMID: 9183526.

38. Karlsson A, Bjorkman P, Bratt G, Ekvall H, Gisslen M, Sonnerborg A, et al. Low prevalence of transmitted drug resistance in patients newly diagnosed with HIV-1 infection in Sweden 2003–2010. PLoS ONE. 2012:in press. https://doi.org/10.1371/journal.pone.0033484 PMID: 22448246

39. Lindstrom A, Ohlis A, Huigen M, Nijhuis M, Berglund T, Bratt G, et al. HIV-1 transmission cluster with M41L 'singleton' mutation and decreased transmission of resistance in newly diagnosed Swedish homosexual men. Antivir Ther. 2006; 11(8):1031–9. Epub 2007/02/17. PMID: 17302373.

40. Leitner T, Escanilla D, Franzén C, Uhlén M, Albert J. Accurate reconstruction of a known HIV-1 transmission history by phylogenetic tree analysis. Proc Natl Acad Sci USA. 1996; 93:10864–9. https://doi.org/10.1073/pnas.93.20.10864 PMID: 8855273

41. Leitner T, Escanilla D, Franzen C, Uhlen M, Albert J. Accurate reconstruction of a known HIV-1 transmission history by phylogenetic tree analysis. Proceedings of the National Academy of Sciences of the United States of America. 1996; 93(20):10864–9. Epub 1996/10/01. https://doi.org/10.1073/pnas.93.20.10864 PMID: 8855273; PubMed Central PMCID: PMC38248.

42. Nasir A, Dimitrijevic M, Romero-Severson E, Leitner T. Large Evolutionary Rate Heterogeneity among and within HIV-1 Subtypes and CRFs. Viruses. 2021; 13(9). Epub 2021/09/29. https://doi.org/10.3390/v13091689 PMID: 34578270; PubMed Central PMCID: PMC8473000.

43. Mooers A, Gascuel O, Stadler T, Li H, Steel M. Branch lengths on birth-death trees and the expected loss of phylogenetic diversity. Syst Biol. 2012; 61(2):195–203. Epub 2011/08/26. https://doi.org/10.1093/sysbio/syr090 PMID: 21865336.

44. Romero-Severson EO, Bulla I, Hengartner N, Bartolo I, Abecasis A, Azevedo-Pereira JM, et al. Donor-Recipient Identification in Para- and Poly-phyletic Trees Under Alternative HIV-1 Transmission Hypotheses Using Approximate Bayesian Computation. Genetics. 2017. https://doi.org/10.1534/genetics.117.300284 PMID: 28912340.

45. Zhang Y, Leitner T, Albert J, Britton T. Inferring transmission heterogeneity using virus genealogies: Estimation and targeted prevention. PLoS Comput Biol. 2020; 16(9):e1008122. Epub 2020/09/04. https://doi.org/10.1371/journal.pcbi.1008122 PMID: 32881984; PubMed Central PMCID: PMC7494101.

46. UNAIDS. 90-90-90: An Ambitious Treatment Target to Help End the AIDS Epidemic. Geneva: UNAIDS, 2014.

47. HIV.gov. Ending the HIV epidemic–a plan for America. USA: White House, 2019.

48. Romero-Severson EO, Lee Petrie C, Ionides E, Albert J, Leitner T. Trends of HIV-1 incidence with credible intervals in Sweden 2002–09 reconstructed using a dynamic model of within-patient IgG growth. Int J Epidemiol. 2015; 44(3):998–1006. https://doi.org/10.1093/ije/dyv034 PMID: 26163684; PubMed Central PMCID: PMC4521128.

49. Dawson L, Benbow N, Fletcher FE, Kassaye S, Killelea A, Latham SR, et al. Addressing Ethical Challenges in US-Based HIV Phylogenetic Research. J Infect Dis. 2020; 222(12):1997–2006. Epub 2020/06/12. https://doi.org/10.1093/infdis/jiaa107 PMID: 32525980; PubMed Central PMCID: PMC7661760.

50. Evans D, Dawson A. Ethical considerations for a public health response using molecular HIV surveillance data: a multi-stakeholder approach. Project Inform and Northwestern University: 2018.

51. Romero-Severson E, Nasir A, Leitner T. What Should Health Departments Do with HIV Sequence Data? Viruses. 2020; 12(9). Epub 2020/09/17. https://doi.org/10.3390/v12091018 PMID: 32932642; PubMed Central PMCID: PMC7551807.