Phylogenetically resolving epidemiologic linkage

Ethan O. Romero-Severson[a], Ingo Bulla[a], and Thomas Leitner[a,1]

[a]Theoretical Biology and Biophysics Group, Los Alamos National Laboratory, Los Alamos, NM 87545

Edited by David M. Hillis, The University of Texas at Austin, Austin, TX, and approved January 29, 2016 (received for review November 30, 2015)

Although the use of phylogenetic trees in epidemiological investigations has become commonplace, their epidemiological interpretation has not been systematically evaluated. Here, we use an HIV-1 within-host coalescent model to probabilistically evaluate transmission histories of two epidemiologically linked hosts. Previous critique of phylogenetic reconstruction has claimed that direction of transmission is difficult to infer, and that the existence of unsampled intermediary links or common sources can never be excluded. The phylogenetic relationship between the HIV populations of epidemiologically linked hosts can be classified into six types of trees, based on cladistic relationships and whether the reconstruction is consistent with the true transmission history or not. We show that the direction of transmission and whether unsampled intermediary links or common sources existed make very different predictions about expected phylogenetic relationships: (*i*) Direction of transmission can often be established when paraphyly exists, (*ii*) intermediary links can be excluded when multiple lineages were transmitted, and (*iii*) when the sampled individuals' HIV populations both are monophyletic a common source was likely the origin. Inconsistent results, suggesting the wrong transmission direction, were generally rare. In addition, the expected tree topology also depends on the number of transmitted lineages, the sample size, the time of the sample relative to transmission, and how fast the diversity increases after infection. Typically, 20 or more sequences per subject give robust results. We confirm our theoretical evaluations with analyses of real transmission histories and discuss how our findings should aid in interpreting phylogenetic results.

HIV-1 | transmission | paraphyly | coalescent | phylogeny

Phylogenetic inference of pathogen transmission chains, outbreaks, and epidemics is a popular method to gain insight into otherwise hidden information about the epidemiologic dynamics of transmission. Many viruses, such as HIV-1, evolve faster than transmissions typically occur, making phylogenetic reconstruction an ideal and objective tool for reconstruction of transmission events. For example, an early case where phylogenetic reconstruction was used involved a Florida dentist and several of his patients (1). Because this was the first criminal investigation of HIV-1 transmission it instigated a series of comments and controversy (2–4) and was eventually settled out of court (5). Another criminal investigation involving a Swedish rapist was investigated and became the first case settled in court (6). Subsequently, many other similar criminal cases occurred around the world (7–19). In all of these cases, phylogenetic reconstruction of transmission events was central to the evidence of guilt. However, the interpretation of phylogenetic trees has broader importance beyond criminal investigations. Phylogenetics now plays an increasingly central role in public health investigations and practices (20–24).

Three critical questions have been raised in response to phylogenetic reconstruction of transmission events: (*i*) In which direction did the transmission occur?, (*ii*) Can intermediary links be excluded?, and (*iii*) Can common sources be excluded? In response, it has been claimed that direction of transmission could not be established with most data and the existence of intermediary or common transmission links could never be excluded (7, 25–27). Thus, phylogenetic reconstruction seemed to only be able to reveal whether two persons were "epidemiologically linked" in some way (28). Formally, epidemiologic linkage between two persons (labeled A and B) can occur in one of three ways (Fig. 1): direct transmission (A or B transmits to the other), indirect transmission (transmission from A or B to the other with at least one intervening transmission), or common source (both A and B infected by an unsampled person).

One common method of excluding direct transmission is to look for insertion of local control sequences splitting donor and recipient sequences into separate clades (1, 7). However, because one can never be sure all relevant controls have been sampled, the absence of control sequence(s) inside the A + B monophyletic clade cannot exclude possible intermediary links or common sources. This broad linking of cases, however useful, ignores much of the potential phylogenetic information about the putative transmission history. For example, donor paraphyly was suggested to indicate the source in a transmission chain (18, 19, 29). Several studies have shown that transmission of >1 phylogenetic lineage occurs in 20–40% of transmissions, depending on transmission route and other factors (30–33). This implies that transmission histories may generate more complicated phylogenies than previously considered (i.e., involving combinations of monophyletic, paraphyletic, and polyphyletic relationships). Being able to determine the probability of the phylogenetic topology as a function of epidemiologic relationship, sampling time, and number of samples places the phylogenetic resolution of epidemiologic linkage on a firm theoretical grounding.

Results

Different Transmission Histories Predict Different Expected Phylogenetic Relationships. We used coalescent theory to study the topological signal of a phylogeny of HIV-1 clonal sequences sampled from two putatively linked hosts, labeled A and B, and their true epidemiological relationship (Fig. 2). The six different classes of topologies that are possible are determined by the cladistic relationship between the A and B lineages, which can be dually monophyletic

Significance

Phylogenetic inference of who infected whom has great value in epidemiological investigations because it should provide an objective test of an explicit hypothesis about how transmission(s) occurred. Until now, however, there has not been a systematic evaluation of which phylogeny to expect from different transmission histories, and thus the interpretation of what an observed phylogeny actually means has remained somewhat elusive. Here, we show that certain types of phylogenies associate with different transmission histories, which may make it possible to exclude possible intermediary links or identify cases where a common source was likely but not sampled. Our systematic classification and evaluation of expected topologies should make future interpretation of phylogenetic results in epidemiological investigations more objective and informative.

Author contributions: E.O.R.-S., I.B., and T.L. designed research; E.O.R.-S., I.B., and T.L. performed research; E.O.R.-S., I.B., and T.L. contributed new reagents/analytic tools; E.O.R.-S. and T.L. analyzed data; and E.O.R.-S., I.B., and T.L. wrote the paper.

The authors declare no conflict of interest.

This article is a PNAS Direct Submission.

Freely available online through the PNAS open access option.

[1]To whom correspondence should be addressed. Email: tkl@lanl.gov.

This article contains supporting information online at www.pnas.org/lookup/suppl/doi:10.1073/pnas.1522930113/-/DCSupplemental.

www.pnas.org/cgi/doi/10.1073/pnas.1522930113

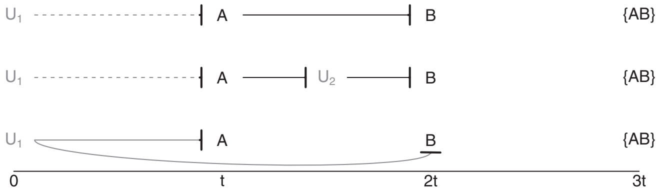

Fig. 1. Epidemiological links between two hosts. Two sampled hosts, A and B, may be linked through transmission in three prototypical transmission histories: (*Top*) by having directly infected the other, (*Middle*) by an unsampled intermediary link (U_2), or (*Bottom*) by a common source (U_1). In our simulations t indicates a single unit of time such that the infection times of A and B and the sampling time are always equidistant from one another. In the indirect transmission case, the unsampled intermediary link (U_2) is infected at time 1.5t.

(MM), paraphyletic–monophyletic (PM), or a combination of paraphyletic and polyphyletic (PP), that in turn assign the root label (A, B, or equivocal). In our framework two forces determine the topological signal: (*i*) the differential and stochastic loss of A and B lineages going backward in time in each host and (*ii*) the coalescence of A and B lineages with one another when they are in a common host (Fig. S1). In all epidemiologic relationships we define the source population as the within-host population from which transmission occurs, generating two derived populations in hosts A and B. The source population may exist in hosts A, B, or some other unsampled host.

We evaluated the topological signal and consistency with the actual transmission events under three possible scenarios (Fig. 1): direct transmission (A transmits to B), indirect transmission (A transmits to an intermediary who transmits to B), and common source (A and B infected by same source). In the case of direct or indirect transmission we call the root label consistent if it agrees with the label of the donor.

The expected phylogenetic relationship of A and B lineages strongly depends on the transmission scenario (Fig. 3):

MM/equivocal, the HIV populations in the hosts' are both monophyletic, that is, no paraphyly exists, and thus there is no indication of the direction of transmission. Typically, common source transmissions result in MM phylogenies. Indirect transmissions, and to a lesser degree direct transmission, may also result in MM trees, especially when β is small ($<2\ d^{-1}$). However, at βs that give normal diversification levels [3–5 d^{-1} (34)], direct and indirect transmissions typically result in PM/consistent trees and common source transmissions typically result in MM trees. Note that MM is consistent with a common source because neither subject infected the other.

PM/consistent, donor's population is paraphyletic and recipient's is monophyletic. PM topologies are only possible in common source transmissions when both a large number of lineages are transmitted (α) and within-host diversification (β) is rapid. In general, however, the PM topology most often results from direct or indirect transmission.

PM/inconsistent, donor's population is monophyletic and recipient's is paraphyletic, which would mislead transmission direction reconstruction. This topology is highly improbable under realistic scenarios.

PP/equivocal, neither donor nor recipient HIV populations are monophyletic and the root cannot be assigned to either. This topology is very rare but may result from direct transmissions where many lineages are transmitted and within-host diversification is high. Interestingly, with this topology direct transmission is highly probable, but we cannot say who the donor was.

PP/consistent, the donor's HIV population is paraphyletic with root label A (and the recipient is polyphyletic), supporting direct transmission from donor to recipient. This topology virtually excludes intermediary links and common sources.

PP/inconsistent, where recipient's HIV population is paraphyletic, and thus transmission appears as recipient to donor. This topology is rare (<1% in common source cases with high β).

Importantly, qualitative aspects of the distribution of the topological signal are robust to times between transmissions [Fig. 3, *Right*, (t)] and sample size (Fig. S2); 20 clones from each population give robust inference of the topology.

Paraphyletic Signal Predicts Direction of Transmission but Decays with Time and Decreasing Sample Size. The inference of the direction of transmission is theoretically possible in PM and PP topologies (Fig. 2). Moving along the reverse time axis in the derived populations, lineages are probabilistically lost to coalescence in each host. The number of lineages with A or B labels that merge in the source population is therefore a random variable determined by the sample size, sampling time, and within-host dynamics. In general, this quantity will be smaller than the HIV-1 within-host population size (35–37), or effective population size (38–40), and consequently sampling plays an important role in the ability of genetic data to resolve an epidemiologic linkage. Furthermore, first principles predict that as the time between transmission and sampling increases, eventually old lineages will be lost in the derived populations, leading to loss of the paraphyletic signal (Fig. S3).

Correct reconstruction of the transmission direction when the donor transmits one lineage is highly probable (>95%), even 3–4 y after transmission and with only 20 sampled sequence clones, when the donor has been infected for several years (Fig. 4). However, this probability decreases substantially when the number of sampled clones becomes small or the donor has only been infected for a short time. Our simulations show that with only five clones there is only a 50% chance to see the correct reconstruction after about 5 y. If the donor had been infected for only 6 mo at time of transmission, the probability of correct transmission direction reconstruction quickly decreases; even with 100 clones from the donor the correct reconstruction drops to 50% chance at about 5 y after transmission. Again, the probability of inconsistent reconstruction, that is, when it would seem as if the recipient infected the donor, was <1% overall.

Interestingly, the more complicated case when 10 lineages were transmitted had roughly the same probabilities (Fig. 4). This is due

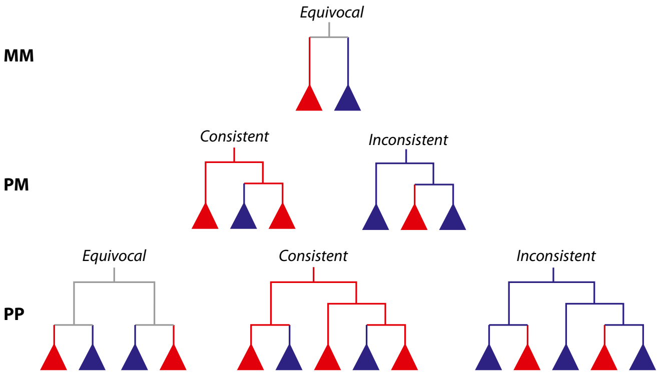

Fig. 2. Classes of topological signal. When one host (red) is epidemiologically linked to another host (blue), the resulting virus populations upon sampling may relate to each other such that both populations are monophyletic (MM), or one is paraphyletic and the other monophyletic (PM), or one is paraphyletic and the other polyphyletic relative to the other (PP). If the red host was infected first, the deduced root label of the phylogeny may be equivocal (the root node could be assigned to either host), consistent (correct root assignment in direct or indirect transmission cases), or inconsistent (incorrect root assignment in direct or indirect transmission cases).

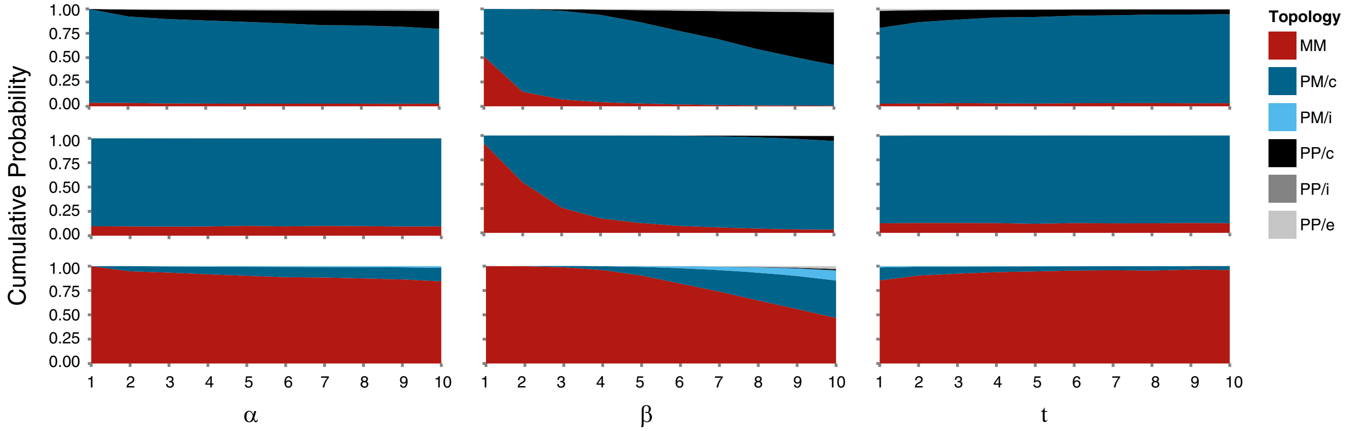

Fig. 3. The distribution of topological signal as a function of within- and between-host dynamics. Color indicates the expected topological signal as a function of number of transmitted lineages (α), population linear growth rate (β), and time between transmissions and samplings (t). In each column the top panel shows the distribution of topological signal in direct transmission, the middle panel in indirect transmission, and the bottom panel in common source transmission. Topological classes are as in Fig. 2; /c, consistent; /e, equivocal; /i, inconsistent. When not indicated the default parameters are α = 5, β = 5 d^{-1}, and t = 1 y (Fig. 1).

to the fact that in the direct transmission case the number of lineages in the source population with the label of the donor will almost always be larger than the number of lineages with the label of the recipient due to the transmission bottleneck. However, in extreme cases such as a very large number of transmitted lineages or a very small sample size in the donor, this may not be true.

PP Trees Indicate Direct Transmission. When a PP tree is observed it is almost certain that no intervening transmission occurred (Fig. 5). Observing a PP topology when in fact an intermediary link existed (A–U_2–B chain in Fig. 1) is highly unlikely because more than one lineage sampled in A must survive not only the transmission bottleneck from U_2 to B but also from A to U_2 (Fig. 1). This may only happen (>1%) when the number of transmitted lineages is very high (α >24). Thus, the only time when a PP relationship is reliably observed is under direct transmission from A to B.

Analysis of Real Cases. We investigated the consistency of our results with three real transmission cases where the transmission history was known (Fig. 6): a common source case where two men had been infected by the same male donor resulted in an MM topology, a case from a gay couple where the recipient was recently infected by the chronically infected partner resulted in a PM topology, and a case where a known HIV-1–positive donor injured a victim in a robbery that resulted in a PP topology. Thus, the topological signal in each case was consistent with the known transmission history (33, 41, 42).

To evaluate whether the inferred trees were consistent with our theoretical analysis, we modeled each case using published epidemiologic data to inform infection and sampling times. Because we could not directly estimate α and β, we tested the range 1–10 (heat maps below each tree in Fig. 6). In the common source case an MM topology was most likely to be observed at low α at any β, or at β = 1 it would be consistent with any α. Note that β = 1 is generally unlikely (34); we show it here only to be fully inclusive. Likewise, the PM/consistent tree was most likely if transmission involved one or two lineages at any β >1. In the robber–victim case we observed a PP/consistent topology, which was to be expected at high α and β. Interestingly, the probability to observe the PP tree was virtually zero at any α <4, suggesting that although only two transmitting lineages were sampled many lineages were likely transmitted in this case. In all three cases inconsistent results, where we would get the transmission direction wrong, were expected to occur <1%.

Furthermore, reanalyzing other published transmission pairs with known or assumed direction showed mostly PM/consistent, some PP, and a few MM phylogenies (18, 33, 43, 44). Likewise, in known transmission chains, involving several persons with multiple sampled clones per patient, again most transmissions seemed to be PM/consistent with a few MM topologies among patients that had infected each other, whereas among patients that had not directly infected each other the topology always was MM (19, 45). Hence, overall, many real cases where the true transmission history was known supported our theoretical results.

Discussion

Our simulations demonstrate that different transmission scenarios make very different predictions about expected phylogenetic relationships, validating the use of viral phylogenies to investigate the existence of a transmission event as well as its direction

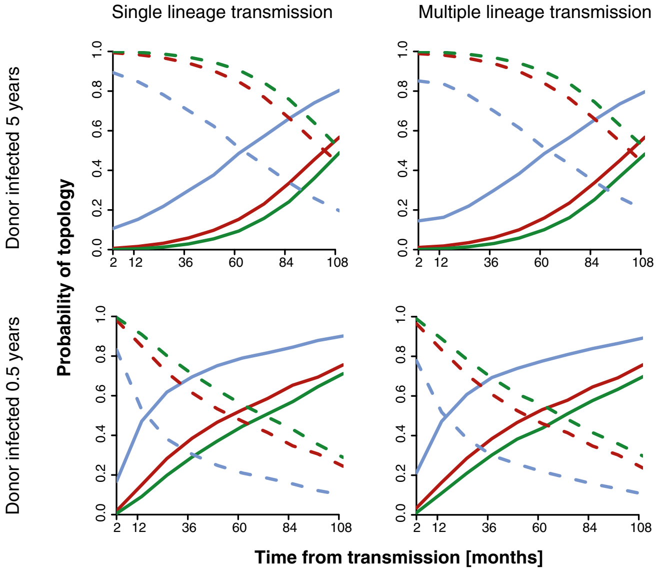

Fig. 4. Paraphyletic reconstruction of direction of transmission. The probability of consistent (dashed lines) and equivocal (solid lines) inference of direction of transmission depends on sample size (green = 100 sequences, red = 20 sequences, blue = 5 sequences) and time from transmission (*x* axis). Rows show examples of direct transmission from a donor who was infected for 5 y or 0.5 y at time of transmission, and columns single or multiple lineage transmission (10 phylogenetic lineages).

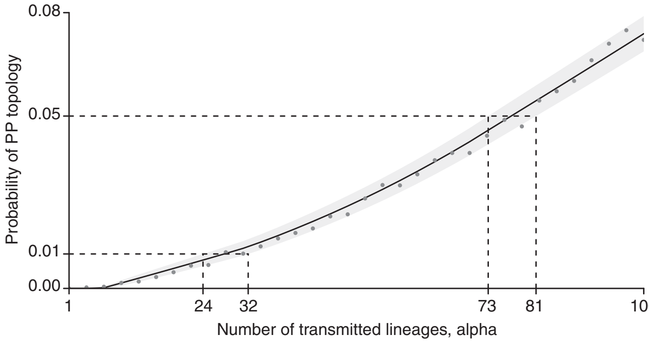

Fig. 5. The probability of observing a PP tree given indirect transmission depends on the number of transmitted lineages α. Each point in the graph shows the mean of 10,000 Monte Carlo simulations at $\beta = 5\ d^{-1}$, where the time interval between transmission events was 1 y and samples were collected 1 y after the last transmission (t = 1 y in Fig. 1). The gray envelope shows the 95% interval of the means, and the line is a LOESS fit to the means.

and directness. Because different transmission histories (direct, indirect, and common source transmissions) impose very different population size dynamics (Fig. 7), mainly determined by the transmission bottleneck(s), the phylogeny in each case has a different expected distribution of topologies (Fig. 3).

The bottleneck(s) has a strong effect on how many lineages can survive through the transmission event back to the source population (Fig. 7). Furthermore, the time that defines the beginning of the source population is different in each transmission history even when the transmission times and sampling times are the same. Together these effects determine the distribution of expected topologies resulting from a transmission scenario. In addition, we show that the resulting inference of the transmission history also depends on the system parameters (i.e., the number of transmitted lineages, the sample size, the time of the sample relative to transmission, and how fast the diversity increases after infection/transmission).

Contrary to claims in the literature asserting that monophyletic reconstruction gives the assurance of proper inference, PP phylogenies provide the most information about who infected whom, because it can virtually exclude intermediary links or common sources. Interestingly, pairs previously judged to be indeterminate show clear transmission direction as PP/consistent trees (see figure 5 in ref. 46 for an example). Note also that with proper rooting many MM phylogenies render PM/consistent, which has information about direction of transmission that MM does not. In fact, the MM phylogeny has the least information about who infected whom because it cannot indicate direction or exclude intermediary links or common sources (7, 25, 26). With proper rooting, the MM phylogeny is typically suggestive of a common source but may also be the result of an intermediary unsampled link, especially when HIV diversification is slow in a host (Fig. 3).

In this study we evaluated the fundamental powers and limitations that can be expected from phylogenetic reconstruction of (potential) transmission pairs. Thus, we evaluated trees as if they were fully resolved, perfectly sampled, and generated by the neutral coalescence processes described by our within-host model. In reality, short branches may not be resolved due to lack of mutations, all relevant lineages may not be sampled, and phylogenetic uncertainty can also occur with large amounts of homoplasy induced by convergent selection or recombination (47–51). In general, fast-evolving genomic regions, such as *env*, have been shown to accumulate enough mutations to robustly recover branches (28). Slower-evolving genomic regions, such as *pol*, have also been used for epidemiological investigations because data are conveniently available from clinical databases due to drug resistance evaluations, and it has been shown to reliably reconstruct known transmission histories (52). When using *pol* sequences, where convergent selection for drug resistance may operate, it is important to strip drug resistance sites before phylogenetic reconstruction (6, 34, 52).

Our framework uses a within-host population growth coalescent model that (*i*) restarts the population growth in each host upon infection (Fig. 7), (*ii*) allows multiple lineages to be transmitted, (*iii*) disallows coalescence of lineages between hosts except at transmission, and (*iv*) is independent of a molecular clock. Although no current Bayesian phylogenetic implementation includes all these features, our approach can easily be used to analyze multiple trees, which can come from Bayesian posterior samples, bootstrap samples, or multiple genomic regions.

As we have shown previously (34), besides α, β, and time between transmissions, which we evaluate here, the maximum population size in our two-phase within-host coalescent model is dominating the impact on the topological outcome. Thus, selective sweeps (due to drug selection or immune escape) may reduce the effective population size and increase the rate of coalescence, possibly altering the expected topological outcome. In the work presented here, we have simplified the two-phase model to a one-phase linear increase because we focus on transmission from a source that has been infected for less than the time it takes to reach the second phase, which may happen 2–8 y after infection, if at all (53).

The inference of donor–recipient relationships we describe here is not restricted to HIV transmissions; it applies to all situations when an original population seeds a new population with a restricted random draw (a bottleneck) of individuals. We use HIV transmission to illustrate the effects because it may aid in contact tracing and untangle outbreak investigations, and the need of statistical guidelines for the interpretation of phylogenetic results in court has been called for (27). Thus, the coalescent model we used is based on HIV diversification (34, 53),

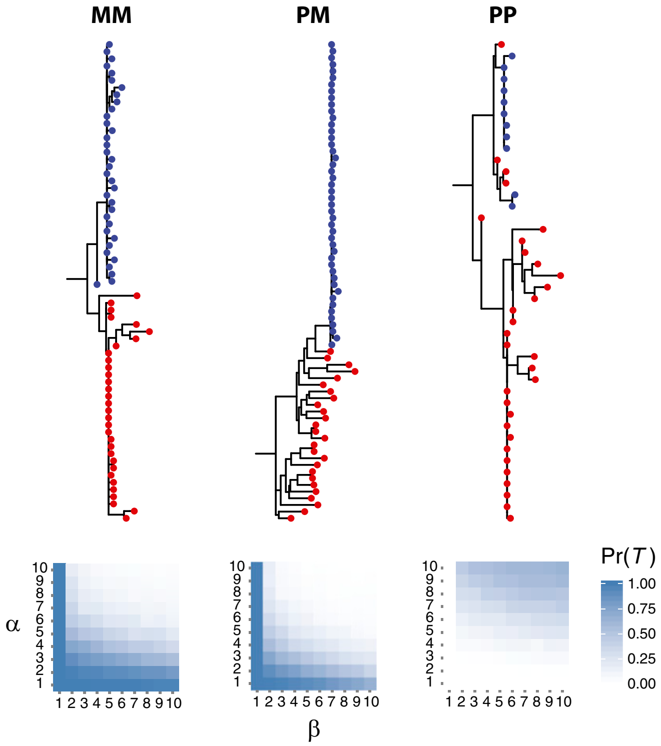

Fig. 6. Examples of real HIV-1 transmission reconstructions. The MM tree came from a common source case, the PM tree came from a recipient that was recently infected by a chronically infected partner, and the PP tree from a case where a robber injured a victim. Each tree was rooted by an outgroup (not shown). Below each tree we show a heat map of the probability of observing the respective topology Pr(*T*).

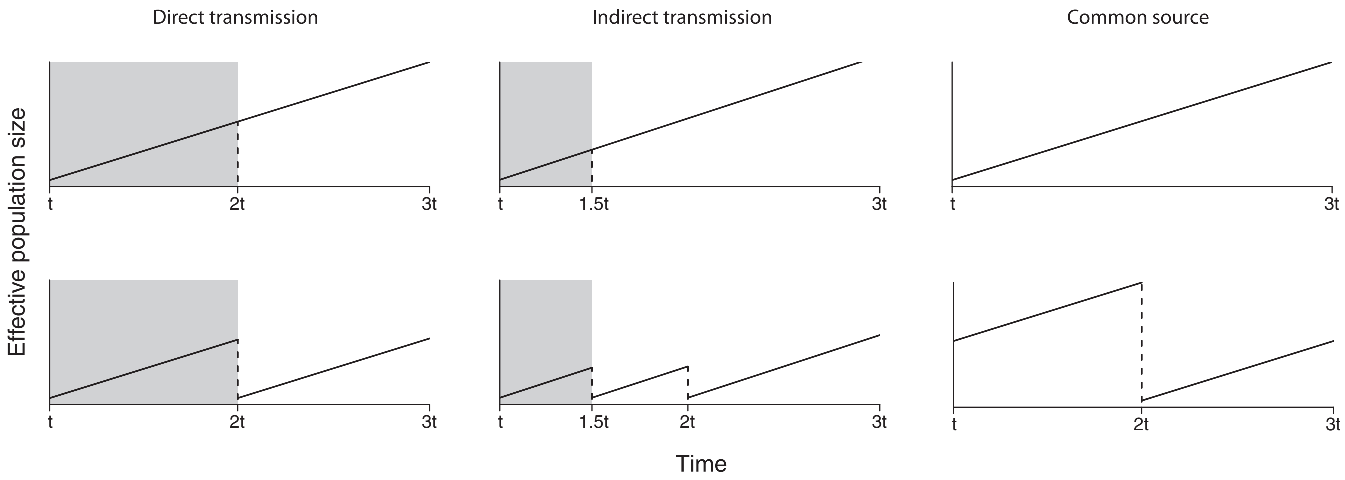

Fig. 7. Population growth profiles in three prototypical transmission histories. The top three panels show the population growth in host A, and the bottom three panels in host B, respectively, for direct transmission, indirect transmission, and transmissions from a common source (Fig. 1). The gray shaded area indicates the times when lineages in A and B can coalesce with one another in the source population. In the common source transmission the source population occurs before time t in an unsampled host.

but with model and parameter adjustments this framework could be used for any diversifying population of organisms.

Materials and Methods

Real Cases and Phylogenetic Reconstruction. We investigated in detail three real HIV-1 transmission cases that display an MM phylogeny (41), a PM phylogeny (33), and a PP phylogeny (42). The MM case consisted of two male recipients (P1 and P2) that had been infected by a common male donor on the same evening. The samples were taken 63 d after transmission. The donor could not be found. Based on relaxed-clock estimates, the donor had been infected at least 2.82 [95% highest posterior density (HPD) 1.28, 4.54] y before the dual transmission event (41). The PM case consisted of a chronically infected donor who recently had infected a recipient (LACU9000 and HOBR0961). It was unknown how long the donor had been infected, and based on sequence and clinical data analyses it was estimated the recipient was sampled 17 d after transmission (33). The PP case consisted of a robber who injured a victim with a knife and transmitted at least two phylogenetic lineages. Based on previous positive HIV-1 status, the donor (robber) had been infected for at least 1,010 d at time of transmission. The donor and recipient were sampled 225 and 244 d after transmission, respectively (42).

HIV-1 sequences were aligned using MAFFT with the L-INS-i algorithm (54). The MM case had 67 HIV-1 subtype B *gag* sequences (alignment length 788 nt), the PM case 72 subtype B *env* sequences (2,620 nt), and the PP case 42 CRF 07_BC *env* sequences (481 nt). Phylogenetic trees were inferred using PhyML (55) under a GTR+I+G substitution model, four categories of Gamma optimization, with a Bio-NJ starting tree and best of NNI and SPR search.

Within-Host Linear Growth Model. In a single infected host we assume linear growth in the theoretical population size from the time of infection such that $N(t)=\alpha+\beta t$, where α is the number of transmitted lineages and β is the rate of growth. The population size at any point in time depends of the specific epidemiologic relationship. For example, in the case of direct transmission, the population size in the source and derived donor population (the same population in this case) is $N(t)=\alpha_d+\beta_d t$. The population size in the derived recipient population $N(t)=\alpha_r+\beta_r(t-t_{trans})$, where t_{trans} is the time of transmission and d and r subscripts represent parameters of the model in the donor and recipient, respectively. Additional details on the model are given in *Supporting Information*.

Simulation of Topological Signal. Simulation of topological signal as a function of within- and between-host parameters has two components, stochastic simulation of the loss of lineages in the derived populations and deterministic simulation of the topological signal given the number of A and B labeled lineages in the source population.

The root label is determined by propagating the host tip labels to the root. Coalescence between two of the same labels propagates the same label to the parent node, between A and B labels propagates an equivocal label (indicated by *), and between * and A or B propagates an A or B label, respectively. In the neutral case all tree topologies are equally probable in the source population, and thus the probability of root labels A, B, or * is determined by the number of A and B lineages that survive into the source population (Fig. S1).

The number of lineages that survive into the source population can be described by a sequence of random variables giving the time to the next coalescent event as a function of the population size and number of extant lineages. Unfortunately these variables are convoluted and cannot be directly evaluated without evaluating a high-dimensional and unwieldy integral. To simulate the number of lineages that survive into the source population we use the previously derived density of the time to the next coalescent event under the linear growth model that we call Z (34). Integrating and inverting with respect to Z gives the inverse cumulative distribution function

$$F_Z^{-1}(u)=\left(1-(1-u)^{\frac{\beta}{\binom{k}{2}}}\right)(\alpha+\beta t_1)\beta^{-1},$$

where k is the number of extant lineages and t_1 is the index time. If u is a unit uniform random variate, then $F_Z^{-1}(u)$ is a random draw from the distribution of the time to the next coalescent event under the linear growth model. To simulate the number of lineages that survive into the source population we draw a sequence of random variates from Z updating the values of k and t_1 along the sequence (Fig. S4).

Once in the source population we can use a Markov chain and an indicator variable to simulate the topological signal conditional on the number of extant lineages with labels A or B. The initial state of the chain is $[N_A,\ N_B,\ N_*]$, where N_A and N_B are the number of lineages with label A and B, respectively, and N_* is 0; an aggregator variable, I, is also initialized to 0. Steps of the chain represent coalescent events where the number of extant lineages is reduced until a single lineage remains.

There are six possible events (coalescences) that can occur: AA, AB, A*, BB, B*, and **. If the labels are the same, then the probability of coalescence is $\Pr(xx)=C(N_x)$, and if the labels are different, then the probability of coalescence is $\Pr(xy)=(C(N_x+N_y)-C(N_x)-C(N_y))$, where $C(x)=\frac{x(x-1)}{N(N-1)}$ and $N=N_A+N_B+N_*$. If a coalescence occurs between two lineages with the same label, then the number of lineages of that label is decremented by one. If a coalescence occurs with an A and B lineage, the aggregator variable is incremented by one, both N_A and N_B are decremented by one, and N_* is incremented by one. Finally, if a coalescence occurs between a * and either an A or B lineage, N_* is decremented by one. The one exception is that I is not incremented if the final coalescence is between an A and B lineage. The value of I at the final coalescence gives the following topology: If $I=0$ the topology is MM, if $I=1$ the topology is PM, and if $I>1$ the topology is PP. In the PM and PP topology case, I is also the apparent number of transmitted lineages. The label of the single remaining lineage is the root label. Additional details on the simulations are given in *Supporting Information*.

In Fig. 3 we assumed that the samples were "perfect" (i.e., that every extant variant was sampled). In reality this would correspond to sampling all of the unique genetic variants of the pathogen in an infected host (not necessarily sampling every single pathogen). However, we found that our results were robust as long as more than about 20 clones were sampled (Fig. S2).

ACKNOWLEDGMENTS. We thank the editor and the reviewers for insightful and constructive comments on an earlier version of this paper. Research reported in this publication was supported by National Institute of Allergy and Infectious Diseases/National Institutes of Health under Grant R01AI087520 and Deutsche Forschungsgemeinschaft Fellowship BU 2685/4-1.

1. Ou CY, et al. (1992) Molecular epidemiology of HIV transmission in a dental practice. *Science* 256(5060):1165–1171.
2. Abele LG, DeBry RW (1992) Florida dentist case: Research affiliation and ethics. *Science* 255(5047):903.
3. Smith TF, Waterman MS (1992) The continuing case of the Florida dentist. *Science* 256(5060):1155–1156.
4. Hillis DM, Huelsenbeck JP (1994) Support for dental HIV transmission. *Nature* 369(6475):24–25.
5. Anonymous (1992) No trial to come in Florida dentist case. *Science* 255(5046):787.
6. Albert J, Wahlberg J, Leitner T, Escanilla D, Uhlén M (1994) Analysis of a rape case by direct sequencing of the human immunodeficiency virus type 1 pol and gag genes. *J Virol* 68(9):5918–5924.
7. Leitner T, Albert J (2000) Reconstruction of HIV-1 transmission chains for forensic purposes. *AIDS Rev* 2:241–251.
8. Blanchard A, Ferris S, Chamaret S, Guétard D, Montagnier L (1998) Molecular evidence for nosocomial transmission of human immunodeficiency virus from a surgeon to one of his patients. *J Virol* 72(5):4537–4540.
9. Goujon CP, et al. (2000) Phylogenetic analyses indicate an atypical nurse-to-patient transmission of human immunodeficiency virus type 1. *J Virol* 74(6):2525–2532.
10. Jaffe HW, et al. (1994) Lack of HIV transmission in the practice of a dentist with AIDS. *Ann Intern Med* 121(11):855–859.
11. Holmes EC, Zhang LQ, Simmonds P, Rogers AS, Brown AJ (1993) Molecular investigation of human immunodeficiency virus (HIV) infection in a patient of an HIV-infected surgeon. *J Infect Dis* 167(6):1411–1414.
12. Arnold C, Balfe P, Clewley JP (1995) Sequence distances between env genes of HIV-1 from individuals infected from the same source: Implications for the investigation of possible transmission events. *Virology* 211(1):198–203.
13. Birch CJ, et al. (2000) Molecular analysis of human immunodeficiency virus strains associated with a case of criminal transmission of the virus. *J Infect Dis* 182(3):941–944.
14. Kaye M, Chibo D, Birch C (2009) Comparison of Bayesian and maximum-likelihood phylogenetic approaches in two legal cases involving accusations of transmission of HIV. *AIDS Res Hum Retroviruses* 25(8):741–748.
15. Lemey P, et al. (2005) Molecular testing of multiple HIV-1 transmissions in a criminal case. *AIDS* 19(15):1649–1658.
16. Machuca R, Jørgensen LB, Theilade P, Nielsen C (2001) Molecular investigation of transmission of human immunodeficiency virus type 1 in a criminal case. *Clin Diagn Lab Immunol* 8(5):884–890.
17. Banaschak S, Werwein M, Brinkmann B, Hauber I (2000) Human immunodeficiency virus type 1 infection after sexual abuse: Value of nucleic acid sequence analysis in identifying the offender. *Clin Infect Dis* 31(4):1098–1100.
18. Metzker ML, et al. (2002) Molecular evidence of HIV-1 transmission in a criminal case. *Proc Natl Acad Sci USA* 99(22):14292–14297.
19. Scaduto DI, et al. (2010) Source identification in two criminal cases using phylogenetic analysis of HIV-1 DNA sequences. *Proc Natl Acad Sci USA* 107(50):21242–21247.
20. Volz EM, et al. (2013) HIV-1 transmission during early infection in men who have sex with men: A phylodynamic analysis. *PLoS Med* 10(12):e1001568, discussion e1001568.
21. Lewis F, Hughes GJ, Rambaut A, Pozniak A, Leigh Brown AJ (2008) Episodic sexual transmission of HIV revealed by molecular phylodynamics. *PLoS Med* 5(3):e50.
22. Skar H, et al. (2011) Dynamics of two separate but linked HIV-1 CRF01_AE outbreaks among injection drug users in Stockholm, Sweden, and Helsinki, Finland. *J Virol* 85(1):510–518.
23. Stadler T, Kühnert D, Bonhoeffer S, Drummond AJ (2013) Birth-death skyline plot reveals temporal changes of epidemic spread in HIV and hepatitis C virus (HCV). *Proc Natl Acad Sci USA* 110(1):228–233.
24. Bennett SN, et al. (2010) Epidemic dynamics revealed in dengue evolution. *Mol Biol Evol* 27(4):811–818.
25. Abecasis AB, et al. (2011) Science in court: The myth of HIV fingerprinting. *Lancet Infect Dis* 11(2):78–79.
26. Bernard EJ, Azad Y, Vandamme AM, Weait M, Geretti AM (2007) HIV forensics: Pitfalls and acceptable standards in the use of phylogenetic analysis as evidence in criminal investigations of HIV transmission. *HIV Med* 8(6):382–387.
27. Leitner T (2011) Guidelines for HIV in court cases. *Nature* 473(7347):284.
28. Leitner T, Escanilla D, Franzén C, Uhlén M, Albert J (1996) Accurate reconstruction of a known HIV-1 transmission history by phylogenetic tree analysis. *Proc Natl Acad Sci USA* 93(20):10864–10869.
29. Leitner T, Fitch WM (1999) The phylogenetics of known transmission histories. *The Evolution of HIV*, ed Crandall KA (Johns Hopkins Univ Press, Baltimore).
30. Salazar-Gonzalez JF, et al. (2009) Genetic identity, biological phenotype, and evolutionary pathways of transmitted/founder viruses in acute and early HIV-1 infection. *J Exp Med* 206(6):1273–1289.
31. Keele BF, et al. (2008) Identification and characterization of transmitted and early founder virus envelopes in primary HIV-1 infection. *Proc Natl Acad Sci USA* 105(21):7552–7557.
32. Rieder P, et al. (2011) Characterization of human immunodeficiency virus type 1 (HIV-1) diversity and tropism in 145 patients with primary HIV-1 infection. *Clin Infect Dis* 53(12):1271–1279.
33. Li H, et al. (2010) High multiplicity infection by HIV-1 in men who have sex with men. *PLoS Pathog* 6(5):e1000890.
34. Romero-Severson E, Skar H, Bulla I, Albert J, Leitner T (2014) Timing and order of transmission events is not directly reflected in a pathogen phylogeny. *Mol Biol Evol* 31(9):2472–2482.
35. Fraser C, Hollingsworth TD, Chapman R, de Wolf F, Hanage WP (2007) Variation in HIV-1 set-point viral load: Epidemiological analysis and an evolutionary hypothesis. *Proc Natl Acad Sci USA* 104(44):17441–17446.
36. Perelson AS, Neumann AU, Markowitz M, Leonard JM, Ho DD (1996) HIV-1 dynamics in vivo: Virion clearance rate, infected cell life-span, and viral generation time. *Science* 271(5255):1582–1586.
37. Koelsch KK, et al. (2008) Dynamics of total, linear nonintegrated, and integrated HIV-1 DNA in vivo and in vitro. *J Infect Dis* 197(3):411–419.
38. Brown AJ (1997) Analysis of HIV-1 env gene sequences reveals evidence for a low effective number in the viral population. *Proc Natl Acad Sci USA* 94(5):1862–1865.
39. Nijhuis M, et al. (1998) Stochastic processes strongly influence HIV-1 evolution during suboptimal protease-inhibitor therapy. *Proc Natl Acad Sci USA* 95(24):14441–14446.
40. Kouyos RD, Althaus CL, Bonhoeffer S (2006) Stochastic or deterministic: What is the effective population size of HIV-1? *Trends Microbiol* 14(12):507–511.
41. English S, et al.; SPARTAC Trial Investigators (2011) Phylogenetic analysis consistent with a clinical history of sexual transmission of HIV-1 from a single donor reveals transmission of highly distinct variants. *Retrovirology* 8:54.
42. Kao CF, et al. (2011) An uncommon case of HIV-1 transmission due to a knife fight. *AIDS Res Hum Retroviruses* 27(2):115–122.
43. Haaland RE, et al. (2009) Inflammatory genital infections mitigate a severe genetic bottleneck in heterosexual transmission of subtype A and C HIV-1. *PLoS Pathog* 5(1):e1000274.
44. Liu Y, et al. (2008) Env length and N-linked glycosylation following transmission of human immunodeficiency virus Type 1 subtype B viruses. *Virology* 374(2):229–233.
45. Vrancken B, et al. (2014) The genealogical population dynamics of HIV-1 in a large transmission chain: Bridging within and among host evolutionary rates. *PLOS Comput Biol* 10(4):e1003505.
46. Campbell MS, et al.; Partners in Prevention HSV/HIV Transmission Study Team (2011) Viral linkage in HIV-1 seroconverters and their partners in an HIV-1 prevention clinical trial. *PLoS One* 6(3):e16986.
47. Immonen TT, Leitner T (2014) Reduced evolutionary rates in HIV-1 reveal extensive latency periods among replicating lineages. *Retrovirology* 11(1):81.
48. Doyle VP, Andersen JJ, Nelson BJ, Metzker ML, Brown JM (2014) Untangling the influences of unmodeled evolutionary processes on phylogenetic signal in a forensically important HIV-1 transmission cluster. *Mol Phylogenet Evol* 75:126–137.
49. Sato H, et al. (2000) Convergent evolution of reverse transcriptase (RT) genes of human immunodeficiency virus type 1 subtypes E and B following nucleoside analogue RT inhibitor therapies. *J Virol* 74(11):5357–5362.
50. Holmes EC, Zhang LQ, Simmonds P, Ludlam CA, Brown AJ (1992) Convergent and divergent sequence evolution in the surface envelope glycoprotein of human immunodeficiency virus type 1 within a single infected patient. *Proc Natl Acad Sci USA* 89(11):4835–4839.
51. Sanborn KB, Somasundaran M, Luzuriaga K, Leitner T (2015) Recombination elevates the effective evolutionary rate and facilitates the establishment of HIV-1 infection in infants after mother-to-child transmission. *Retrovirology* 12:96.
52. Lemey P, et al. (2005) Molecular footprint of drug-selective pressure in a human immunodeficiency virus transmission chain. *J Virol* 79(18):11981–11989.
53. Shankarappa R, et al. (1999) Consistent viral evolutionary changes associated with the progression of human immunodeficiency virus type 1 infection. *J Virol* 73(12):10489–10502.
54. Katoh K, Toh H (2008) Recent developments in the MAFFT multiple sequence alignment program. *Brief Bioinform* 9(4):286–298.
55. Guindon S, Lethiec F, Duroux P, Gascuel O (2005) PHYML Online–a web server for fast maximum likelihood-based phylogenetic inference. *Nucleic Acids Res* 33(Web Server issue):W557–W559.