Phylogenetically resolving epidemiologic linkage

Ethan O. Romero-Severson, Ingo Bulla, Thomas Leitner

Theoretical Biology and Biophysics Group, Los Alamos National Laboratory, Los Alamos, NM 87545, USA

Submitted to Proceedings of the National Academy of Sciences of the United States of America

While the use of phylogenetic trees in epidemiological investigations has become commonplace, their epidemiological interpretation has not been systematically evaluated. Here, we use a novel 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 6 types of trees, based on monophyletic and paraphyletic relationships and whether the reconstruction is consistent with the true transmission history or not. We show that which of the 6 classes of trees to expect depends on the direction of transmission, and whether unsampled intermediary links or common sources existed. 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. With 20 or more sequences per subject, direction of transmission can often be established when paraphyly exists, intermediary links can be excluded when multiple lineages were transmitted, and when the sampled individuals' HIV populations both are monophyletic a common source was likely the origin. Inconsistent results, where we would infer the wrong transmission direction, were generally rare. 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

INTRODUCTION

Phylogenetic inference of pathogen transmission chains, outbreaks, and epidemics has become 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). Many other similar criminal cases also 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 reconstructions of transmission events: 1) In which direction did the transmission occur? 2) Can intermediary links be excluded? and 3) 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 appeared to only be able to reveal if two persons were "epidemiologically linked" in some way (28). Such a link can be critically tested by asking if the suspected donor and recipient HIV-1 sequence data co-cluster with one another rather than with any other local or database control sequences (1, 7). The insertion of any control sequence, splitting donor and recipient sequences into separate clades, would exclude direct transmission between donor and recipient. 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 (29).

A paraphyletic relationship in a virus phylogeny occurs when a set of sequences from one host is ancestral to a set of sequences from another host suggesting that the direction of transmission is from the ancestor to the descendent. If samples from two hosts in a putative transmission chain are both monophyletic, i.e., sister clades, one cannot infer the direction of transmission as the ancestry is ambiguous. In more recent analyses that used multiple clones of HIV-1 to investigate transmission chains, paraphyly in a set of clones from one individual and monophyly in the set of clones from another individual was successfully argued to indicate the direction of transmission (18, 19). However, 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). Thus, paraphyletic relationships may be more complicated than previously considered.

Until now the state of knowledge of exactly what can and cannot be said about transmission events based on phylogenetic reconstruction has been based largely on logical deduction from implicit models. Hence, the lack of more complete statistical analysis has hampered the interpretation of phylogenetic results in epidemiological investigations. Here we extend a recent model

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.

Reserved for Publication Footnotes

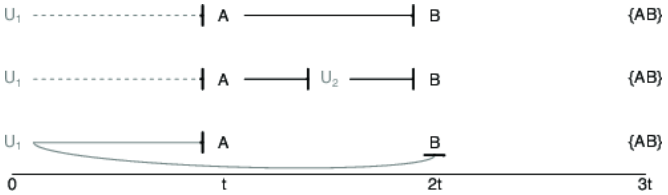

Fig. 1. Epidemiological links between two hosts. Two sampled hosts, A and B, may be linked through transmission in 3 prototypical transmission histories: top row, by having directly infected the other; middle row, by an unsampled intermediary link (U_2); or bottom row, by a common source (U_1). We model these 3 prototypical transmission histories such that samples from A and B are taken at time 3t, A gets infected by an unsampled/unknown donor U_1 at time t, and B gets infected at time 2t. In the indirect transmission case, the unsampled intermediary link (U_2) is infected at time 1.5t.

of within-host dynamics of HIV-1 (34) to investigate different types of transmission histories and probabilistically evaluate the fundamental limitations of paraphyletic inference of direction of transmission when single or multiple lineages were transmitted. We further investigate the probability of the existence of intermediary links and the possibility that epidemiologically linked individuals were infected by an unsampled common source.

RESULTS

Conceptual model and definitions

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 & B infected by an unsampled person). We define the *joint population* as the within-host population from which transmission occurs generating two *derived populations*. Moving along the reverse time axis from the time at which the A and B were sampled, lineages are lost due to coalescence in the derived populations. Once the derived populations merge into the joint population, the remaining lineages sampled from A and B are free to coalesce with one another as they are in the same population (Fig 2). We define the *phylogenetic signal* of the sample as both the topology of the tree with respect to the tip labels (A or B) and the consistency of the implied temporal order of events. The statistical properties of the phylogenetic signal are determined by 1) the explicit nature of the epidemiologic linkage, 2) the probabilistic loss of lineages through coalescence in the derived populations, and 3) the statistical combination of lineages in the joint population.

Root label determines the consistency of the phylogenetic signal

Figure 3 illustrates the different classes of phylogenetic signal with respect to samples from two hosts labeled A and B. When both populations are monophyletic (MM), the root node is equivocal, i.e., we cannot determine who was infected first. When one population is paraphyletic relative to the other (PM), i.e. the root node is unambiguously labeled A or B, the order of infections is inferred to be from the paraphyletic population to the monophyletic population. In the direct and indirect transmission case, this corresponds to the direction of transmission going from one person to the other. In the common source case, the root label simply implies the temporal sequence of events. If the root label agrees with the actual sequence of events, the phylogenetic signal is considered to be consistent. However, insufficient sampling or the stochastics of the coalescent process may result in the inconsistent inference of transmission direction. Similarly, in the dual paraphyletic case (PP), resulting from transmission of more than one lineage, the inference of transmission direction may be consistent, inconsistent, or equivocal depending on the precise tree topology.

Assuming all tree topologies are equally probable, the root assignment of trees with two host labels (A and B) is determined by the combinatorial space of all possible tree topologies (Fig S1). The probability of root labels, A, B, or equivocal, are determined by the number of A and B labels in the joint population. Thus, when one label is dominant, it will most often determine the root assignment. One example of this situation is when A directly infects B with one lineage, resulting in 1 B label and typically several A labels in the joint population, forming a PM topology. Thus, such a result would be consistent with the true transmission history of A infecting B. However, when the joint population for any reason has a small number of labels the root may be assigned to the less frequent host label. This situation could arise as a function of time or simply a small sample of sequences. In addition, the root assignment more often becomes equivocal when more lineages from both hosts exist in the joint population, which translates to PP topologies with transmission of more than one lineage.

Paraphyletic signal decays with time and decreasing sample size

In the previous section we described what to expect when a fixed number of sampled lineages exist in the joint population (i.e. can coalesce with one another). However, moving along the reverse-time axis, sampled lineages are probabilistically lost to coalescence. The number of lineages with A or B labels that exist in the joint population is a random variable determined by the sample size, sample 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 sizes (38-40), and consequently sampling plays an important role in the ability of genetic data to resolve an epidemiologic linkage. Furthermore, as time passes from the transmission, lineages die out and the paraphyletic signal will eventually be lost (Fig S2).

Figure 4 shows the expected probability of reconstructing the correct transmission direction in the case where a donor directly transmits to a recipient in 4 illustrative examples. Recall that correct inference of the direction of transmission is theoretically possible in PM and PP topologies (Fig 3). In the case were the donor transmits one lineage, the correct reconstruction-of-direction-probability is high (>95%) with 20 or more sampled clones even 3-4 years after transmission if the donor had been infected for 5 years at time of transmission. With only 5 clones, there is only a 50% chance to see the correct reconstruction after about 5 years. If the donor had been infected for only 0.5 years at time of transmission, however, 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 years after transmission. Overall, the probability of inconsistent reconstruction, i.e., 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. This is due to the fact that in the direct transmission case, the number of lineages in the joint population with the label of the actual 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. Curiously, in this case, the probability of correct reconstruction increased in the first year after transmission. This is because the number of lineages that exist at the time of transmission from the recipient's sample are rapidly lost to coalescence due to low diversity in the newly infected recipient. However, the donor has a diverse within-host population and looses lineages to coalescence at a much slower rate. If we hold the number of lineages with the label of the donor in the joint population constant and reduce the number of lineages with the label of the recipient, the probability

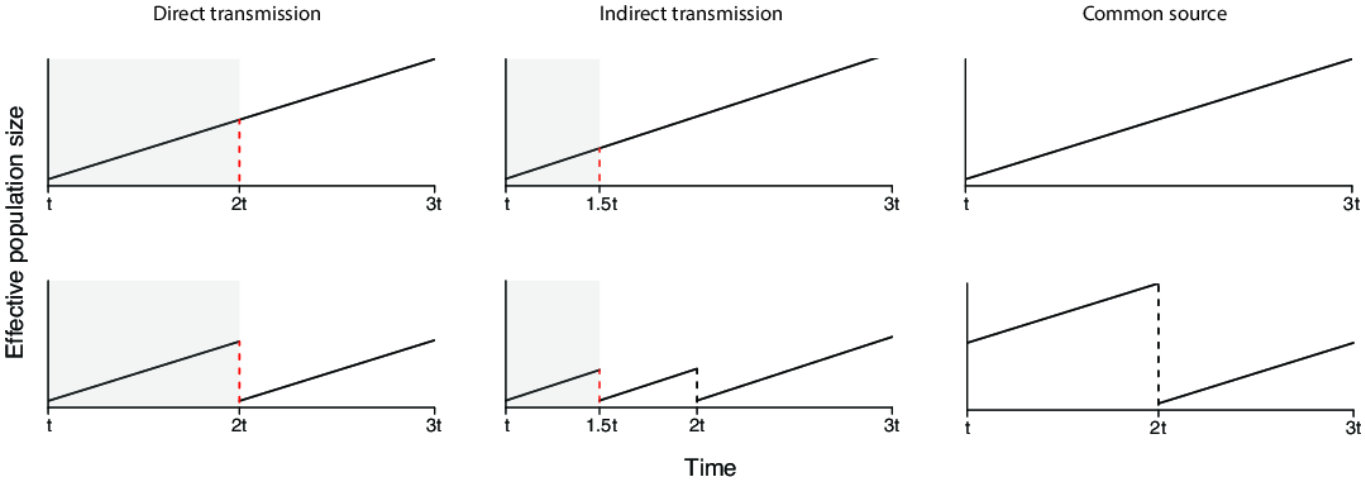

Fig. 2. Population growth profiles in three prototypical transmission histories. The top 3 panels show the population growth in host A, and the bottom 3 panels in host B, respectively, for direct transmission, indirect transmission, and transmissions from a common source. The gray shaded area indicates the times when lineages in A and B can coalesce with one another in the joint population. In the common source transmission the joint population occurs before time t in an unsampled host.

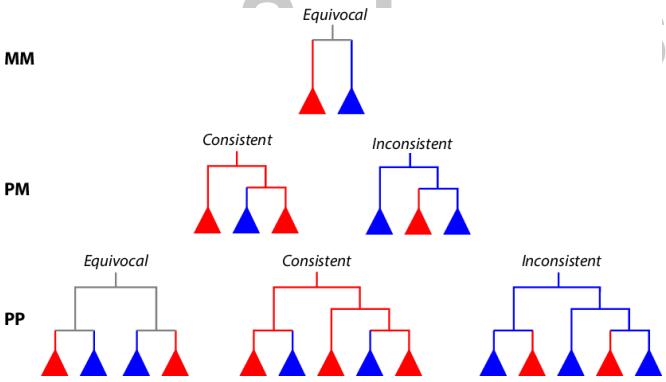

Fig. 3. Classes of phylogenetic signal. When one host (red) is epidemiologically linked to another host (blue), the resulting virus populations upon sampling may relate to each other such that both populations are monophyletic (MM), or one is paraphyletic and the other monophyletic (PM), or both are paraphyletic 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).

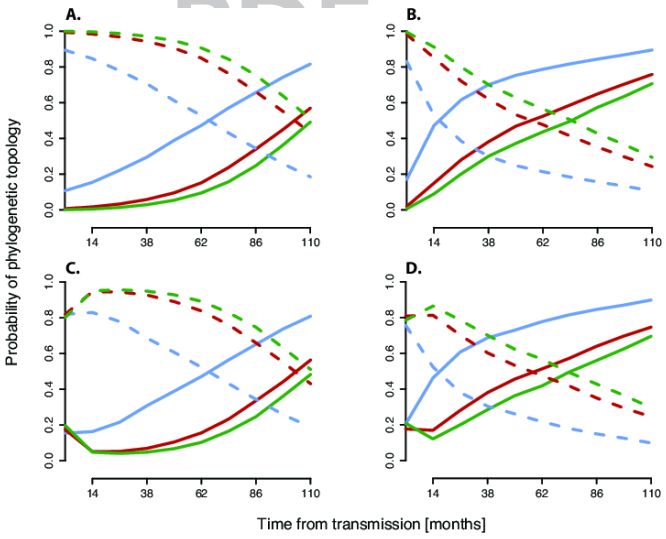

Fig. 4. Paraphyletic reconstruction of direction of transmission. The probability of consistent (dashed lines) and equivocal (solid lines) inference of direction of transmission depends on sample size (green = 100 sequences, red = 20 sequences, blue = 5 sequences) and time from transmission (x-axis).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

of obtaining an equivocal phylogenetic signal decreases. Over longer periods of time, the number of lineages in the donor slowly drops leading to a increased probability of obtaining an equivocal result (Fig S2).

Dual paraphyly indicates direct transmission

We define direct transmission as transmission from donor (A) to recipient (B) without any intermediary (U_2) link (Fig 1). Figure 5 shows the probability of observing a paraphyletic-paraphyletic (PP) A-B relationship when in fact an A-U_2-B chain occurred. When the recipient is infected with a single phylogenetic lineage, a PP relationship is impossible per se. However, if more than one lineage is transmitted, there is some probability of obtaining a PP tree. We found that if a PP tree is observed it is almost certain that no intervening transmission occurred. That is, the only time when a PP relationship is reliably observed is under direct transmission from A to B. This is due to the fact that in the case of indirect transmission, 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 (Figs 1&2). This only happens (>1%) when number of transmitted lineages is implausibly high ($\alpha > 24$).

The probability of the phylogenetic signal as a function of transmission and within-host dynamics

Figure 6 shows the distribution of phylogenetic topologies and their consistency with the actual transmission events under 3 possible scenarios: 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). The distribution of the phylogenetic signal depends on the number of transmitted lineages (α), the growth rate of the effective population (β), times between transmissions and sampling, and number of sampled lineages.

Typically, common source transmissions result in MM phylogenies. From a topological inference perspective, MM is actually consistent with a common source as neither subject infected the other. PM topologies are only possible in common

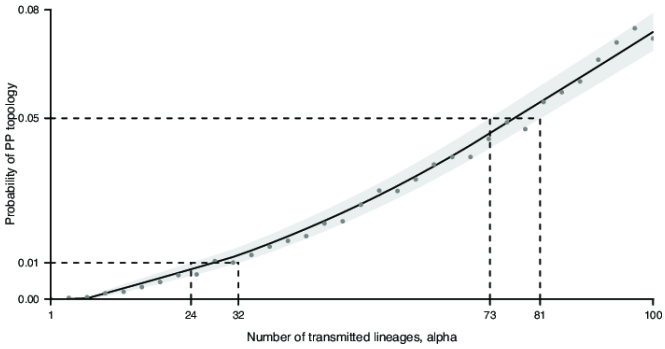

Fig. 5. The topological effect of intermediary links. The probability of observing a PP given indirect transmission depends on the number of transmitted lineages α. Each point in the graph shows the mean of 10,000 Monte Carlo simulations at $\beta = 5$ day^{-1} where the time interval between transmission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

source transmissions when both a large number of lineages are transmitted and within-host diversification is rapid. In general, the PM topology most probably results from direct or indirect transmission. MM topologies can also be observed when β is low (<2 day^{-1}) in direct and indirect transmissions. At β's that give normal diversification levels [3—5 day^{-1} (34)], direct and indirect transmissions typically result in PM/consistent trees and common source transmissions typically result in MM trees. When PP trees are observed, they most probably result from direct transmission, making it possible to exclude intermediary links and common sources. Encouragingly, qualitative aspects of the distribution of the phylogenetic signal is robust to times between transmissions and sample size (Fig S3).

Analysis of real cases

We investigated the plausibility of our results with 3 real transmission cases where the transmission history was known (33, 41, 42), and resulted in MM, PM, and PP phylogenies (Fig 7). The MM case came from a common source where two gay men had been infected by the same donor, the PM case came from a gay couple where the recipient was recently infected by the chronically infected partner, and the PP case came from a known HIV-1 positive donor who injured a victim in a robbery. Thus, the phylogenetic signal in each case was consistent with the known transmission histories.

To evaluate if the inferred trees were consistent with our theoretical analysis, we modeled each case where the phylogeny informed α and published epidemiologic data informed infection and sampling times. Since we could not directly estimate β, we tested low, medium, and high (1, 5, 10 day^{-1}) population growth values (bars below each tree in Figure 7). Regardless of β, the MM topology was to be entirely expected in the common source transmission, and likewise the PM/consistent topology was clearly to be expected in the gay couple case. In the robber-victim case we observed a PP/consistent topology, which is to be expected at 29% when β=5—10. The most likely outcome would have been PM/consistent at 54% with β=5—10, and PP/equivocal at 16%. Low β seems unlikely in this case, as β=1 did not show any expectation of a PP topology at all, and low β is unlikely anyway (34). In all 3 cases, inconsistent results, where we would get the transmission direction wrong, were expected to occur <1%.

DISCUSSION

There are two main determinants of phylogenetic signal: first, the population size dynamics are vastly different in the 3 possible transmission histories (Fig 2), which have a strong effect on how many lineages that can survive through the bottleneck(s) of transmission back to the joint population. Second, the time that defines the beginning of the joint population is different in each transmisson history even when the transmission times and sampling times are the same. Together these effects determine the distribution of phylogenetic signal. Consequently, 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. There are 6 possible classes of cladistic relationships between two epidemiologically linked hosts: 1) MM/equivocal, the HIV populations in the hosts' are both monophyletic, i.e., no paraphyly exists, and no indication of the direction of transmission. As we have shown, direct transmission very rarely results in MM trees, which instead typically suggests infection from a common source. 2) PM/consistent, donor's population is paraphyletic and recipient's is monophyletic. This topology is expected in both direct and indirect cases. 3) 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. 4) PP/equivocal, both donor and recipient HIV populations are paraphyletic relative to each other. Interestingly, in this case, it is highly probably that one person infected the other (i.e. direct transmission), but we cannot say who was the donor. 5) PP/consistent, where both HIV populations are paraphyletic and the topology supports direct transmission. This topology virtually excludes intermediary links and common sources. 6) PP/inconsistent, where both are paraphyletic, but transmission appears as recipient to donor. This topology is rare (<1% in common source cases with high β).

Given a direct transmission, we expect PM/consistent, PP/equivocal, or PP/consistent phylogenetic signal to de the dominant outcomes (Fig 6). Indeed, a large number of published transmission pairs show PM/consistent and some PP phylogenies, e.g., (18, 19, 33, 43, 44). Across all our simulations in Figure 6, we expect a PP phylogeny in 22% of direct transmissions, 0.9% of indirect transmissions, and 0.7% of common source transmissions, meaning that we can be reasonably sure that no intermediary link or common source existed when we observe a PP tree. Furthermore, among PP trees PP/inconsistent is rare (1.5%, 2.9%, and 29%, respectively). Thus, contrary to claims in the literature that assert that monophyletic reconstruction give 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. (45) for 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 nor 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 6).

Additional data such as sexual partner preference and time of transmission(s) can further constrain the probability of intermediary links in PP trees. For instance, if a putative recipient claims to be infected by suggested donor, and both are strictly heterosexual, that implies at least 2 additional intermediary persons in the chain. Hence, if we observe a PP topology between a putative donor and recipient in that situation, then the probability of several intermediary links, rather than just one, is virtually zero. Also, PP/equivocal situations may be informed about direction if other data indicates who was infected first.

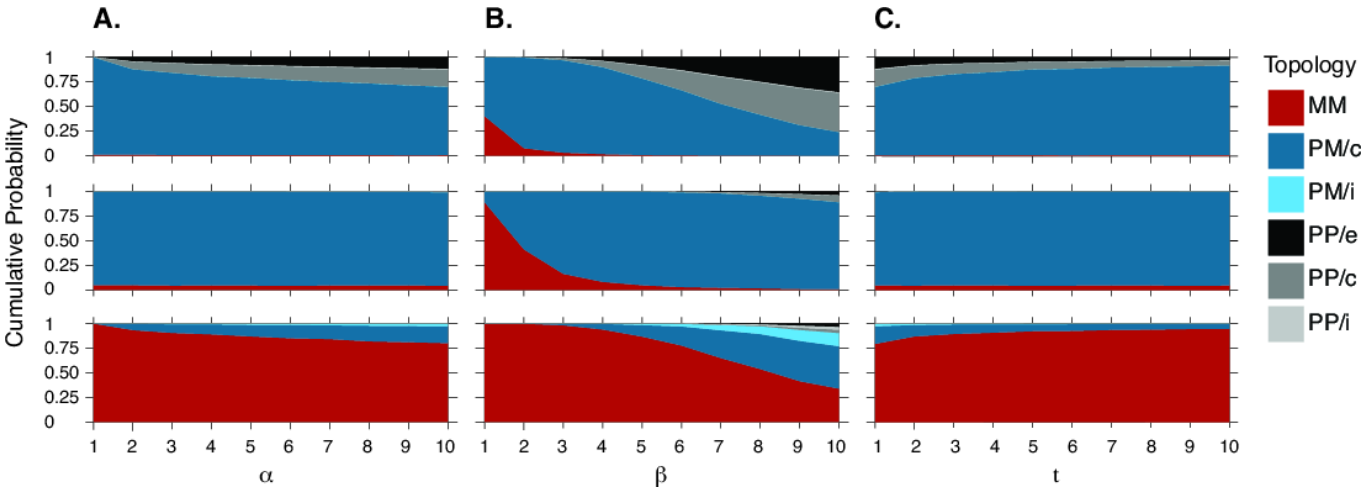

Fig. 6. The distribution of phylogenetic signal. Color indicates the expected phylogenetic signal as a function of number of transmitted lineages (**A**), population linear growth rate (**B**), and time between transmissions and samplings (**C**). In each panel the top subpanel shows the distribution of phylogenetic signal in direct transmission, the middle subpanel in indirect transmission, and the bottom subpanel in common source transmission. When not indicated the default parameters are α=5, β=5 day^{-1}, t=1 year.

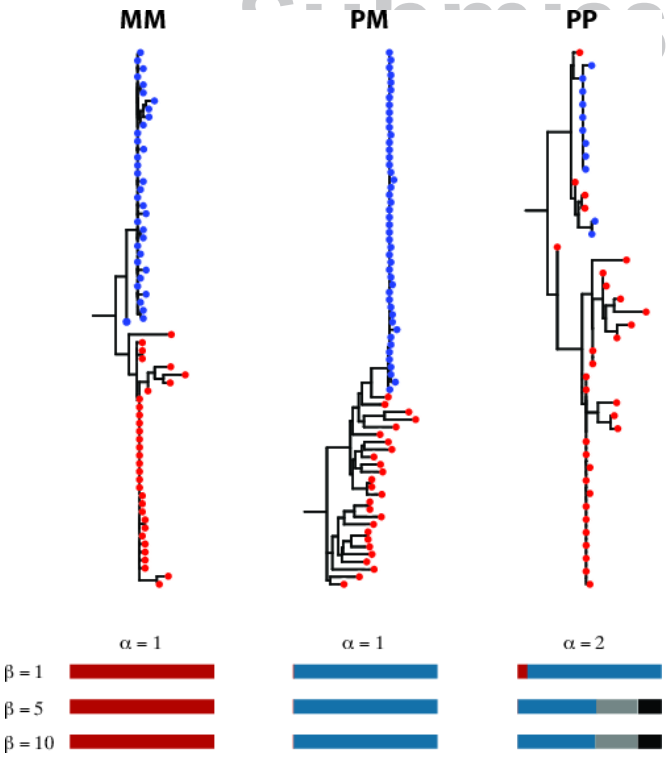

Fig. 7. Examples of real HIV-1 transmission reconstructions. The MM tree came from a common source case, the PM tree came from a recipient that was recently infected by a chronically infected partner, and the PP tree from a case where a robber injured a victim. Each tree was rooted by an outgroup (not shown). Below each tree we show the expected topological distribution (colors as in Fig 6) at the observed (apparent) α, and β = (1, 5, 10).

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, 46), 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 three real HIV-1 transmission cases that display a 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 days after transmission. The donor could not be found. Based on relaxed-clock estimates, the donor had been infected at least 2.82 (95% HPD 1.28, 4.54) years prior to 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 days after transmission (33). The PP case consisted of a robber who injured a victim with a knife and transmitted at least 2 phylogenetic lineages. Based on previous positive HIV-1 status, the donor (robber) had been infected for at least 1010 days at time of transmission. The donor and recipient were sampled 225 and 244 days after transmission, respectively (42).

HIV-1 sequences were aligned using MAFFT with the L-INS-i algorithm (47). The MM case had 67 HIV-1 subtype B *gag* sequences (alignment length 788 nt), the PM case 72 subtype B *env* sequences (2620 nt), and the PP case 42 CRF 07_BC *env* sequences (481 nt). Phylogenetic trees were inferred using PhyML (48) under a GTR+I+G substitution model, 4 categories Gamma optimization, with a Bio-NJ starting tree and best of NNI and SPR search.

Within-host linear growth model

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. Before the time of infection of the index case, the population size is defined and depends on how long the donor has been infected. For example, if the donor is infected at time 0 and transmits at time t_{trans}, then the population size is given by

$$N(t) = \begin{cases} \alpha_r + \beta_r(t - t_{trans}), & t \geq t_{trans} \\ \alpha_d + \beta_d t, & t < t_{trans} \end{cases}$$

where d and r subscripts represent parameters of the model in the donor and recipient respectively.

Simulation of coalescent times

Derivation of the density of coalescent times for the linear growth model is given in (34). Defining Z as the density of times to the next coalescent event from a given index, we can generate random variates from Z with the inverse cumulative distribution function of Z

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

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. To simulate the number of lineages that remain from a sample at some time in the past, we draw a sequence of random variates from Z updating the values of k and t_1 along the sequence.

Distribution of phylogenetic topologies under neutrality

Given two possible labels (A and B), the distribution of topologies with respect to those labels (MM, PM, PP) can be simulated under neutrality with a simple Markov chain. 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. There are 6 possible coalescences with respect to lineage labels. 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 $\Pr(xy) = \left(C(N_x + N_y) - C(N_x) - C(N_y)\right)$, 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 sole exception to the rules 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 topology: if $I = 0$ the topology is MM, if $I = 1$ the topology is PM, if $I > 1$ the topology is PP. In the PM and PP topology case, I is also the apparent number of transmitted lineages.

Consistency of phylogenetic signal

We define the consistency or inconsistency of the phylogenetic topology as when the root label implies a sequence of events in the order in which they actually occurred or not. A phylogenetic signal is said to be equivocal when the sequence of events cannot be discerned from the tree. Hence, MM topologies are always equivocal, as the label at the root cannot be determined; PM topologies can be either consistent or inconsistent; and PP topologies can be any of the three (Fig 3). To determine the consistency of the phylogenetic signal regardless of the topology we use the same basic Markov chain as before, however, disregarding the aggregator variable. The probability of a consistent phylogenetic signal is defined by the distribution of lineage labels when only one lineage remains. Assuming that person A is infected before person B, the phylogeny is consistent with actual events when the root label is A, inconsistent when the root label is B, and equivocal when the root label is *.

ACKNOWLEDGEMENTS.

Research reported in this publication was supported by the NIAID/NIH under award number R01AI087520 and the Deutsche Forschungsgemeinschaft (fellowship BU 2685/4-1) . The content is solely the responsibility of the authors and does not necessarily represent the official views of the NIH.

1. Ou CY, *et al.* (1992) Molecular epidemiology of HIV transmission in a dental practice. *Science* 256:1165-1171.
2. Abele LG & DeBry RW (1992) Florida dentist case: research affiliation and ethics. *Science* 255:903.
3. Smith TF & Waterman MS (1992) The continuing case of the Florida dentist. *Science* 256:1155-1156.
4. Hillis DM & Huelsenbeck JP (1994) Support for dental HIV transmission. *Nature* 369:24-25.
5. Anonymous (1992) No trial to come in Florida dentist case. *Science* 255:787.
6. Albert J, Wahlberg J, Leitner T, Escanilla D, & Uhlén M (1994) Analysis of a rape case by direct sequencing of the HIV-1 *pol* and *gag* genes. *J. Virol.* 68:5918-5924.
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, Guetard D, & Montagnier L (1998) Molecular evidence for nosocomial transmission of human immunodeficency virus from a surgeon to one of his patients. *J. Virol.* 72:4537-4540.
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.* 121:855-859.
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.* 167:1411-1414.
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* 211:198-203.
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.* 182:941-944.
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, Jorgensen LB, Theilade P, & Nielsen C (2001) Molecular investigation of transmission of human immunodeficiency virus type 1 in a criminal case. *Clin diagnost lab imm* 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* 99:14292-14297.
19. Scaduto DI, *et al.* (2010) Source identification in two criminal cases using phylogenetic analysis of HIV-1 DNA sequences. *Proc Natl Acad Sci U S A* 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.
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, Kuhnert 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, Franzen C, Uhlen 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, MD).
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. *The Journal of experimental medicine* 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 U S A* 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 U S A* 104(44):17441-17446.
36. Perelson AS, Neuman 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* 271:1582-1586.
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. Leigh 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* 94:1862-1865.
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.* (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. Campbell MS, *et al.* (2011) Viral linkage in HIV-1 seroconverters and their partners in an HIV-1 prevention clinical trial. *PLoS ONE* 6(3):e16986.
46. Shankarappa R, *et al.* (1999) Consistent viral evolutionary changes associated with the progression of human immunodeficiency virus type 1 infection. *J. Virol.* 73:10489-10502.
47. Katoh K & Toh H (2008) Recent developments in the MAFFT multiple sequence alignment program. *Brief Bioinform* 9(4):286-298.
48. 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:W557-559.