Published in final edited form as:
Nat Microbiol. 2018 September ; 3(9): 983–988. doi:10.1038/s41564-018-0204-9.

Consistent phylogenetic patterns recover HIV epidemiologic relationships and reveal common transmission of multiple variants in known transmission pairs

Thomas Leitner and **Ethan Romero-Severson**
Theoretical Biology & Biophysics group, MS K710, Los Alamos National Laboratory, Los Alamos, NM 87545

Abstract

The growth of HIV sequence databases resulting from drug resistance testing has motivated efforts using phylogenetic methods to assess how HIV spreads[1–4]. Such inference is potentially both powerful and useful for tracking the epidemiology of HIV and allocation of resources to prevention campaigns. We recently used simulation and a small number of illustrative cases to show that certain phylogenetic patterns are associated with different types of epidemiological linkage[5]; our original approach was later generalized for large NGS datasets and implemented as a free computational pipeline[6]. Previous work has claimed that direction and directness of transmission could not be established from phylogeny because one could not be sure there were no intervening or missing links involved[7–9]. Here, we address this issue by investigating phylogenetic patterns from 272 previously identified HIV transmission chains with 955 transmission pairs representing diverse geography, risk groups, subtypes, and genomic regions. These HIV transmissions had known linkage based on epidemiological information such as partner studies, mother-to-child transmission, pairs identified by contact tracing, and criminal cases. We show that the resulting phylogeny inferred from real HIV genetic sequences indeed reveals distinct patterns associated with direct transmission contra transmissions from a common source. Thus, our results establish how to interpret phylogenetic trees based on HIV sequences when tracking who-infected-whom, when, and how genetic information can be used for improved tracking of HIV spread. We also investigate limitations that stem from limited sampling and genetic time-trends in the donor and recipient HIV populations.

Correspondence and requests for materials should be addressed to TL at tkl@lanl.gov.

AUTHOR CONTRIBUTIONS
TL designed the study and compiled the data, TL and ERS analyzed the data and wrote the paper.

COMPETING INTERESTS
The authors declare no competing interests.

Data availability
Supplementary Table 1 lists epidemiological relationship, demographic, and genomic information of all transmission pairs in this study (CSV format). The alignments of each cluster genomic region with subtype outgroup will be made available at the LANL HIV database Special Interest Alignments at https://www.hiv.lanl.gov/content/sequence/HIV/SI_alignments/datasets.html.

Code availability
Computer codes will be made available to researchers upon request.

Phylogenetic analysis of Human Immunodeficiency Virus (HIV) sequences has become a popular method to reveal epidemiological patterns relevant to disease tracking as well as details about transmission. Epidemiological patterns include the fundamental transmission history, which is not possible to directly observe but underlies the observable HIV phylogeny. While it is attractive to assume that these are identical, they may in fact be quite different[10,11]. The main reason for the discrepancy between HIV phylogeny and transmission history is because HIV quickly diversifies in a host. The existence of a highly diverse HIV population also raises the question of how many variants that may be transmitted.

A highly diverse founding population i) makes it harder for the immune system to fight HIV and accelerates the time to AIDS[12–15], ii) increases the probability of transmitting drug-resistant variants and developing future resistance[16], iii) reduces the efficacy of immunological-based prevention technologies[17], and iv) obscures epidemiological relationships[10,18].

Transmission moves some limited number of viral particles from the donor's population to a recipient[19–22]. As HIV within-patient diversity can build up many-years-worth of genetic variation, transmission of even a few particles can represent a highly diverse founding population. Diversity then continues to accumulate[23], and as the adaptive immune system activates, the diversification rate increases as HIV escapes this evolving pressure[24]. Cohort studies of acutely infected persons have used early patterns of diversification to argue that the majority of HIV infections start with a single virus strain, while 20–40% start with more than one HIV strain[22,25,26]. Other studies have investigated individual transmission pairs, or small transmission chains[27,28], showing that a bottleneck at transmission clearly occurs. However, because these studies either only investigated recipient HIV populations or relatively few donor-recipient pairs, they could not study the donor-recipient phylogenetic patterns generally.

Recently, several mathematical modeling studies have investigated population bottlenecks during transmission. To augment sequence data, some methods have inferred transmission histories using other data or made strong assumptions[5,29–31]. Together, these studies showed that transmission leaves characteristic and detectable signals in phylogenetic trees that can elucidate direction, directness, and diversity of the founding population. While such patterns were investigated in a few real transmission cases, the lack of a large-scale analysis of real donor-recipient transmission cases, describing many different epidemiological scenarios, has left researchers skeptical whether general patterns are discernable or not.

In order to evaluate general phylogenetic patterns associated with different modes of HIV transmission, we divided the transmission pairs into known direct or common source transmissions. With HIV DNA sequence samples from hosts A and B, direct transmission corresponds to when A infected B and common source when an unsampled host X infected both A and B. For each such A-B pair, we then reconstructed the joint HIV phylogeny using 30,000 Bayesian posterior phylogenies per pair to take into account phylogenetic reconstruction uncertainty and classified the resulting HIV phylogenies into paraphyletic-polyphyletic or polyphyletic-polyphyletic (PP), paraphyletic-monophyletic (PM), or monophyletic-monophyletic (MM) patterns (Figure 1). To infer the phylogenetic topology, outgroup rooting with specific HIV subtype reference sequences was superior to other rooting methods (see Methods). 71.3% of all datasets had a properly defined outgroup (>95% posterior support for root monophyly). The 28.7% that did not, identified 1) datasets with too little power to reconstruct meaningful phylogenies (27.7% had <10% posterior support), typically with too short genomic sequences, and 2) less frequently (1%), datasets with patients that unlikely had infected each other.

Analyzing the 681 pairs of known direct or common source transmission that had a proper phylogenetic root, we observed that most such pairs presented a clear phylogenetic pattern (638 of 681 pairs had >95% support for their phylogenetic class). We found that PP and PM trees were associated with direct transmission, while MM trees typically indicated transmission from a common source ($p=1.8\times10^{-14}$, z-test of logistic regression) (Figure 2). Overall, 52% of direct transmissions resulted in a detected PP tree, and 37% PM and 11% MM, while 76% of common source transmissions resulted in a MM tree. There was no trend in inferred phylogenetic class across the genome. Because we had too few known transmission chains with three serially infected patients, we could not investigate indirect transmission situations (where an intervening link exists between the sampled donor and recipient). Such cases with adequate clonal data are unfortunately extremely rare in the literature. From previous theoretical work, however, we expect PP trees to indicate direct transmission while PM can only indicate direction of transmission[5].

Stratified on transmission risk group, in 167 mother-to-child transmission (MTCT) pairs PP trees dominated (66%), followed by PM (26%) and MM (8%). Thus, this shows that, contrary to previous claims, MTCT most often results in transmission of >1 phylogenetic lineage. A recent study by Kumar et al, using new and independent data, also found that multiple transmitted variants is more common in MTCT than previously thought[32]. For men-who-have-sex-with-men (MSM), 27 direct transmissions resulted in either PP or PM trees at approximately equal frequency, while 83 heterosexual (HET) direct transmissions showed more PM than PP in direct transmissions (61 and 19%, respectively). Since the risk of transmission is higher in MSM than in HET[33], the transmission of more founders in MSM, leading to a PP tree, is in agreement with the sexual transmission mode; and previous results have suggested that MSM often are infected with more variants than HET[34]. We found similar results in male-to-female and female-to-male transmissions, i.e., mostly PM trees (Supplementary Figure 1). MM trees dominated in 121 HET common source transmissions (89%), while the MSM common source situation had too few cases to give a clear picture (3 cases). Other types of transmission risks (34 cases), including nosocomial and unknown risk factors, typically showed PP in both direct transmission and common source. The ‘other’ risk group is shown for completeness, but should be interpreted case by case as the epidemiological situations are typically unusual and different from each other.

While the overall phylogenetic class was strongly associated with transmission mode, there were also cases where the overall pattern did not hold, e.g., 33 out of 292 (11%) direct transmissions resulted in a MM tree (Figure 2A). The reason for observing MM trees in direct transmissions is explained by two mechanisms, 1) loss of phylogenetic lineages over time, and 2) limited sampling of clonal DNA sequences. Figure 3 shows first principle trends of how PP or PM trees decay into MM over time as well as with inadequate sampling. The root host-label should ideally indicate the original HIV population from where the recipient's HIV population was drawn during transmission. Hence, with an adequate sample taken before critical lineage loss has occurred, the donor is identified by the root host-label, as seen in the PP tree in Figure 3. With time, the older lineages die (due to the stochastic birth-death process and amplified by selective mechanisms from, e.g., antiviral drug treatment and immune surveillance). Note also that when lineages are lost or unsampled it is possible in both PP and PM trees that the root label is incongruent with the original population, i.e., it suggests that the recipient's population is older than that of the donor. This type of incongruence is uncommon in PM trees (8% of PM sets had >90% posterior probability of incongruence, 88% had >90% congruence, and about 4% were uncertain), while rather common in PP trees (24% of PP sets had >90% posterior probability of incongruence, 30% had >90% congruence, and 46% were uncertain) (Figure 4AB). The larger uncertainty in root host-label reconstruction among PP trees reflects the theoretical expectation that PP trees may have an equivocal root state, like MM trees always do (Figure 3). Hence, while a PP tree indicates direct transmission, it may not be possible to deduce the donor from a simple root label reconstruction due to loss of lineages over time and inadequate sampling. For accurate donor identification, additional epidemiological data such as exact sampling time, potential transmission times, and individually adjusted population growth parameters can aid the proper donor inference[35].

It is important to point out that this study investigated previously observed transmission pairs where the exact epidemiological relationship is known. The relationship between phylogenetic topology, root label, and the nature of the epidemiologic linkage can be population specific. Using a Bayesian framework we can say

$$\Pr\left(D|G_{\theta}\right) = \frac{\Pr\left(G_{\theta}|D\right)\Pr(D)}{\Pr\left(G_{\theta}|D\right)\Pr(D) + \Pr\left(G_{\theta}|\bar{D}\right)\Pr\left(\bar{D}\right)},$$

where G_{θ} is the phylogenetic topology and root label obtained under the observed conditions, θ (e.g. the sampling times, sequencing technology, and within-host population dynamics), D is direct transmission, and $\bar{D}$ is not direct transmission. In this paper we examined $\Pr\left(G_{\theta}|D\right)$ and $\Pr\left(G_{\theta}|\bar{D}\right)$ under the observed (sampling times) and unobserved (within-host dynamics) aspects of θ for a large population of transmission pairs. However, the probability of direct transmission in a specific case should not be taken as the proportion of direct transmission in the population of PP trees in our study. This is due to the fact that the case-specific aspects of a given case contained in θ may not be well represented in our study. Given that unobservable aspects of each host, such as the within-host evolutionary history, can strongly influence the topology and root label for a fixed sampling scheme, extra care in the form of extensive simulations needs to be taken when attempting to make a principled claim about $\Pr\left(D|G_{\theta}\right)$ in a specific case[35].

Conditional on observing a PP tree in one genomic region, only 62% of the examined datasets displayed a PP tree in another genomic region (and 28% were PM and 10% MM)

(Figure 4C). Furthermore, as time proceeds from the time of infection in the recipient, fewer and fewer polyphyletic clades are observed in the recipient (Figure 4E). Finally, more sequences investigated typically revealed more clades in the recipient (Figure 4D). Together, this shows that the donor and recipient HIV populations often are under-sampled. Thus, our results demonstrate that transmissions with true PP trees, and therefore transmission of multiple founders, are more common than previously thought.

The number of HIV clades in the recipient can be interpreted as the minimum number of lineages that were transmitted. We found that with increased number of sequences sampled from the donor and recipient the number of identified transmitted lineages increased (Supplementary Figure 2A). Across all direct transmissions, therefore, both the frequency of PP trees and the number of transmitted founders is likely underestimated. Among detected PP trees, the observed median and mean number of founders was 8.3 and 11.5, respectively, with the distribution significantly skewed towards more founders (Supplementary Figure 2B). These numbers appear high, especially when transmission upon exposure is uncommon[33], mainly due to very few infectious virions in a transmission volume of bodily fluid[36,37]; where one would expect most transmissions resulting in one and rarely two or more founders. While this might be true in many of our HET transmissions, the overall high number of founders in PP trees suggests that many PP trees may be the result from multiple transmission contacts rather than a single transmission of multiple lineages[35]. It is possible that the number of apparent founders could be inflated by within-recipient recombination of a small number of diverse ancestors. However, even in the case of recombination inflating the apparent number of transmitted founders, the true founding population must be highly diverse. Conversely, if recombination occurs outside the examined genomic region, it may hide ancestral lineages that were transmitted by effectively causing lineage death in the partial genomic sequence. For our results presented here, however, recombination cannot falsely generate PP trees from cases where only one lineage was truly transmitted. This means that the phylogenetic patterns determined here are robust to recombination.

The results we present in this study, i.e. that phylogenetic patterns are strongly associated with direct versus common source transmission, support theoretical predictions and justify the foundation of recent bioinformatics applications[6]. On the smaller, pair-wise who-infected-whom level, the strong association between the type of epidemiological linkage and phylogenetic topology opens up possibilities of probabilistic inference of transmission direction using simulations to test alternative scenarios[35].

METHODS

Linked transmission datasets

The LANL HIV database collects and annotates all published HIV sequences[39]. From that database, we retrieved all sequence data from all known HIV “clusters”, i.e. groups of 2 or more patients that have known transmission histories, annotated form the beginning of the recorded HIV research era up until April 2017. The inclusion criteria were: 1) Two or more patients per cluster; 2) 5 or more sequences per genomic region per patient, where the sequences within one genomic region had a start HXB2 coordinate within 80 nucleotides of each other. Besides DNA sequences and a unique patient database code, we collected, when available, the following data: 1) HIV subtype; 2) risk group; 3) sex; 4) time of infection; 5) time of seroconversion; 6) Fiebig stage; and 7) time of sampling. After alignment and initial quality control, this resulted in 272 transmission cluster sequence sets, where 227 (83%) were 2-patient clusters, 19 were 3-patient clusters, 4 4-patient, 4 5-patient, and 9 ≥6-patient clusters. Decomposing these data into epidemiologically linked pairs yielded 955 direct or common source transmission pair sequence sets. 187 (69%) clusters had one genomic region sequenced, while others had 2–13 regions sequenced and some had near full genomes sequenced; the most commonly sequenced genomic region was *env* (Supplementary Figure 3). One cluster was HIV-2, and among HIV-1 clusters subtypes B and C dominated (together 74%), followed by CRF01, D, A1, and G, as well as several recombinants 01/B, 06/A1, CRF07, CRF14, A1/A2, C/D, unclassified (“U”), and group O sequences. 47% of the clusters were mother-to-child transmission (MTCT), 24% heterosexual transmission (HET), 18% men-who-have-sex-with-men transmission (MSM), and the rest had blood transfusion, mixed or unknown transmission risks. In about 35% we had some information on time of infection, and in all cases we had time of sampling (often by year, sometimes month and full date).

Phylogenetic reconstruction

Phylogenetic trees were reconstructed with MrBayes 3.2.6[38] using a GTR+I+Gamma substitution model[40]. The tree topology and branch length priors were both unconstrained (uniform tree prior and non-clock model). We ran 2 chains with 30 million Markov Chain Monte Carlo (MCMC) generations each, sampled every 1000 generations, and discarded the first 50% of the sampled trees as burn-in. Each cluster was thus described by a posterior distribution of 30,000 trees per genomic region.

To assess how rooting affects the phylogenetic reconstruction, different alignments were generated for each genomic region per cluster: 1) alignments with only cluster sequences, 2) alignments with HXB2 included, and 3) alignments with matching subtype reference sequences included[41]. Each such set was aligned using MAFFT v7.305b[42] with the L-INS-i method. We also applied three types of reductions per alignment: 1) none, where all gaps and sequences were included, 2) global gapstripping, where all alignment columns with ≥1 gap were removed, and 3) global gapstripping followed by removal of non-unique sequences per patient. Depending on if and which reference sequences that were included in each genomic region, gapstripping had effects on exactly how many genomic alignments we obtained per cluster. For instance, because the four subtype C reference sequences had gaps in the LTR region, 13 LTR sets were lost due to gapstripping.

Phylogenetic measures

For each phylogenetic tree we measured a set of statistics that we have previously shown both theoretically[5] and empirically[35] to be related to the direction, directness, and frequency of transmission between transmission pairs. First, each tree was classified as paraphyletic-polyphyletic (PP), paraphyletic-monophyletic (PM), or monophyletic-monophyletic (MM)[5]. Here, paraphyly indicates the ancestral population to the joint sample from two epidemiologically linked patients. Computer code for the phylogenetic classification will be made available upon request. Either polyphyly or monophyly of one patient’s sample in combination with paraphyly of the other patient's sample thus indicates that the sequences in the sample are descendants from the paraphyletic population (Figure 1). We argued in previous work that MM trees are most strongly observed when patient pairs were infected by a common source, PM trees are associated with direct or indirect transmission, and PP trees are strongly associated with direct transmission. Here, we classified each transmission pair into PP, PM, or MM categories if greater than 95% of the MrBayes posterior trees fell into one of the three possible categories. Pairs that did not have 95% of the trees in one topological class were not considered in the analysis.

Second, we calculated the maximum credibility cluster (MCC) set for each transmission pair. For each tree in the posterior sample of trees we counted the frequency of all possible monophyletic clusters. We defined the MCC as the set of clusters that occur the most frequently in the posterior distribution of trees and account for each tip in the phylogenetic tree. The number of clusters in the MCC can be interpreted as the minimum number of transmitted lineages in direct transmission cases.

Quality of HIV phylogenetic data for transmission reconstruction

To classify the reconstructed HIV phylogenies into the topological classes that have theoretically been associated with transmission linkage[5], i.e., PP, PM, or MM trees, we found that correct rooting is essential. Thus, midpoint rooting, i.e., identifying the start of the donor-recipient HIV phylogeny halfway along the longest tip-to-tip path, was inferior to outgroup rooting, where the start of the donor-recipient tree is identified by an unrelated reference (Figure 1). In particular, PM trees that would identify donor-to-recipient transmission direction were often rendered MM using midpoint rooting, with the loss of transmission direction signal. For the two outgroup rootings we tested, using subtype specific reference sequences was superior to universally using HXB2, i.e., rooting with subtype references gave phylogenies that better reflected the known transmission direction. For instance, subtype-specific rooting rendered trees PM that were MM with HXB2 for non-subtype B data. Thus, the reported results are based on using appropriate subtype reference sequences as outgroup.

The use of a rooting outgroup also gave us the ability to ask whether any of the outgroup (subtype reference) sequences phylogenetically mingled with the patient sequences studied. Thus, we tested whether the outgroup reference sequences formed a monophyletic clade (Figure 1 shows 3 examples). Phylogenies that identified donor-recipient pairs where data was either too weak to reconstruct epidemiological linkage or that linkage was unsupported (<95% posterior support) were omitted from further analyses. We also annotated them as "linkage not supported" in the LANL HIV database to avoid future erroneous conclusions about HIV transmission.

No subjective sequence exclusions were done on a case by case level, thus potential outliner sequences would be included in the analyses. Such outliers, if they existed, may have caused non-robust rooting or poor topological signal, and thus such sets would be removed by these quality control procedures.

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

ACKNOWLEDGEMENTS

We thank Nick Hengartner for advice on statistical analyses, Christophe Fraser for suggesting the use of Bayes' rule to illustrate the broader inference problem, and Jennifer Macke and Werner Abfalterer for help with database annotation and searches. This study was supported by an NIH/NIAID grant R01AI087520 and an NIH-DOE interagency agreement AI2013183.

REFERENCES

1. Wertheim JO et al. Social and Genetic Networks of HIV-1 Transmission in New York City. PLoS Pathog 13, e1006000, doi:10.1371/journal.ppat.1006000 (2017). [PubMed: 28068413]
2. Pillay D et al. PANGEA-HIV: phylogenetics for generalised epidemics in Africa. Lancet Infect Dis 15, 259–261, doi:10.1016/S1473-3099(15)70036-8 (2015). [PubMed: 25749217]
3. Lewis F, Hughes GJ, Rambaut A, Pozniak A & Leigh Brown AJ Episodic sexual transmission of HIV revealed by molecular phylodynamics. PLoS Med 5, e50, doi:10.1371/journal.pmed.0050050 (2008). [PubMed: 18351795]
4. Poon AF et al. Near real-time monitoring of HIV transmission hotspots from routine HIV genotyping: an implementation case study. Lancet HIV 3, e231–238, doi:10.1016/S2352-3018(16)00046-1 (2016). [PubMed: 27126490]
5. Romero-Severson EO, Bulla I & Leitner T Phylogenetically resolving epidemiologic linkage. Proc Natl Acad Sci U S A 113, 2690–2695, doi:10.1073/pnas.1522930113 (2016). [PubMed: 26903617]
6. Wymant C et al. PHYLOSCANNER: Analysing Within- and Between-Host Pathogen Genetic Diversity to Identify Transmission, Multiple Infection, Recombination and Contamination. bioRxiv, doi:10.1101/157768 (2017).
7. Leitner T & Albert J Reconstruction of HIV-1 transmission chains for forensic purposes. AIDS Rev 2, 241–251 (2000).
8. Abecasis AB et al. Science in court: the myth of HIV fingerprinting. Lancet Infect Dis 11, 78–79, doi:S1473-3099(10)70283-8 [pii] 10.1016/S1473-3099(10)70283-8 (2011). [PubMed: 21272786]
9. Bernard EJ, Azad Y, Vandamme AM, Weait M & Geretti AM HIV forensics: pitfalls and acceptable standards in the use of phylogenetic analysis as evidence in criminal investigations of HIV transmission. HIV Med 8, 382–387, doi:10.1111/j.1468-1293.2007.00486.x (2007). [PubMed: 17661846]
10. 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 31, 2472–2482, doi: 10.1093/molbev/msu179 (2014). [PubMed: 24874208]
11. Volz EM, Romero-Severson E & Leitner T Phylodynamic inference across epidemic scales. Mol Biol Evol, doi:10.1093/molbev/msx077 (2017).
12. Gottlieb GS et al. Dual HIV-1 infection associated with rapid disease progression. Lancet 363, 619–622, doi:10.1016/S0140-6736(04)15596-7 (2004). [PubMed: 14987889]
13. Grobler J et al. Incidence of HIV-1 dual infection and its association with increased viral load set point in a cohort of HIV-1 subtype C-infected female sex workers. J Infect Dis 190, 1355–1359, doi:10.1086/423940 (2004). [PubMed: 15346349]
14. Yang OO et al. Human immunodeficiency virus type 1 clade B superinfection: evidence for differential immune containment of distinct clade B strains. J Virol 79, 860–868, doi:10.1128/JVI.79.2.860-868.2005 (2005). [PubMed: 15613314]
15. Smith DM et al. Lack of neutralizing antibody response to HIV-1 predisposes to superinfection. Virology 355, 1–5, doi:10.1016/j.virol.2006.08.009 (2006). [PubMed: 16962152]
16. Smith DM et al. Incidence of HIV superinfection following primary infection. JAMA 292, 1177–1178, doi:10.1001/jama.292.10.1177 (2004). [PubMed: 15353529]

17. Korber B, Hraber P, Wagh K & Hahn BH Polyvalent vaccine approaches to combat HIV-1 diversity. Immunol Rev 275, 230–244, doi:10.1111/imr.12516 (2017). [PubMed: 28133800]
18. Ypma RJ, van Ballegooijen WM & Wallinga J Relating phylogenetic trees to transmission trees of infectious disease outbreaks. Genetics 195, 1055–1062, doi:10.1534/genetics.113.154856 (2013). [PubMed: 24037268]
19. McNearney T et al. Relationship of human immunodeficiency virus type 1 sequence heterogeneity to stage of disease. Proceedings of the National Academy of Sciences of the United States of America 89, 10247–10251 (1992). [PubMed: 1438212]
20. Wolfs TFW, Zwart G, Bakker M & Goudsmit J HIV-1 genomic RNA diversification following sexual parenteral virus transmission. Virology 189, 103–110 (1992). [PubMed: 1376536]
21. Zhang LQ et al. Selection for specific sequences in the external envelope protein of human immunodeficiency virus type 1 upon primary infection. J. Virol 67, 3345–3356 (1993). [PubMed: 8497055]
22. Salazar-Gonzalez JF et al. Genetic identity, biological phenotype, and evolutionary pathways of transmitted/founder viruses in acute and early HIV-1 infection. The Journal of experimental medicine 206, 1273–1289, doi:10.1084/jem.20090378 (2009). [PubMed: 19487424]
23. Fischer W et al. Transmission of single HIV-1 genomes and dynamics of early immune escape revealed by ultra-deep sequencing. PLoS ONE 5, doi:10.1371/journal.pone.0012303 (2010).
24. Shankarappa R et al. Consistent viral evolutionary changes associated with the progression of human immunodeficiency virus type 1 infection. J. Virol 73, 10489–10502 (1999). [PubMed: 10559367]
25. Keele BF et al. Identification and characterization of transmitted and early founder virus envelopes in primary HIV-1 infection. Proc Natl Acad Sci U S A 105, 7552–7557, doi:0802203105 [pii] 10.1073/pnas.0802203105 (2008). [PubMed: 18490657]
26. Rieder P et al. Characterization of human immunodeficiency virus type 1 (HIV-1) diversity and tropism in 145 patients with primary HIV-1 infection. Clin Infect Dis 53, 1271–1279, doi: 10.1093/cid/cir725 (2011). [PubMed: 21998286]
27. 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 93, 10864–10869 (1996). [PubMed: 8855273]
28. Lemey P et al. Molecular footprint of drug-selective pressure in a human immunodeficiency virus transmission chain. J Virol 79, 11981–11989 (2005). [PubMed: 16140774]
29. Didelot X, Fraser C, Gardy J & Colijn C Genomic infectious disease epidemiology in partially sampled and ongoing outbreaks. Mol Biol Evol, doi:10.1093/molbev/msw275 (2017).
30. Jombart T et al. Bayesian reconstruction of disease outbreaks by combining epidemiologic and genomic data. PLoS Comput Biol 10, e1003457, doi:10.1371/journal.pcbi.1003457 (2014). [PubMed: 24465202]
31. 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 12, e1004869, doi:10.1371/journal.pcbi.1004869 (2016). [PubMed: 27070316]
32. Kumar A et al. Infant Transmitted/Founder HIV-1 Viruses from Peripartum Transmission are Neutralization Resistant to Paired Maternal Plasma. PLoS Pathog, in press (2018).
33. Patel P et al. Estimating per-act HIV transmission risk: a systematic review. AIDS 28, 1509–1519, doi:10.1097/QAD.0000000000000298 (2014). [PubMed: 24809629]
34. Li H et al. High Multiplicity Infection by HIV-1 in Men Who Have Sex with Men. PLoS Pathog 6, e1000890, doi:10.1371/journal.ppat.1000890 (2010). [PubMed: 20485520]
35. Romero-Severson EO et al. Donor-Recipient Identification in Para- and Poly-phyletic Trees Under Alternative HIV-1 Transmission Hypotheses Using Approximate Bayesian Computation. Genetics, doi:10.1534/genetics.117.300284 (2017).
36. Aldovini A & Young RA Mutations of RNA and protein sequences involved in human immunodeficiency virus type 1 packaging result in production of noninfectious virus. J Virol 64, 1920–1926 (1990). [PubMed: 2109098]

37. Rusert P et al. Quantification of infectious HIV-1 plasma viral load using a boosted in vitro infection protocol. Virology 326, 113–129, doi:10.1016/j.virol.2004.05.022 (2004). [PubMed: 15262500]

38. Ronquist F & Huelsenbeck JP MrBayes 3: Bayesian phylogenetic inference under mixed models. Bioinformatics 19, 1572–1574 (2003). [PubMed: 12912839]

39. Foley B et al. HIV Sequence Compendium 2015 (Los Alamos National Laboratory, 2015).

40. 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 71, 4761–4770 (see also correction 1998: 4772; 2565) (1997). [PubMed: 9151870]

41. Leitner T, Korber BT, Daniels M, Calef C & Foley B in HIV Sequence Compendium 2005 (eds Leitner T & et al) 41–48 (Theoretical Biology and Biophysics, Los Alamos National Laboratory, 2005).

42. Katoh K & Standley DM MAFFT multiple sequence alignment software version 7: improvements in performance and usability. Mol Biol Evol 30, 772–780, doi:10.1093/molbev/mst010 (2013). [PubMed: 23329690]

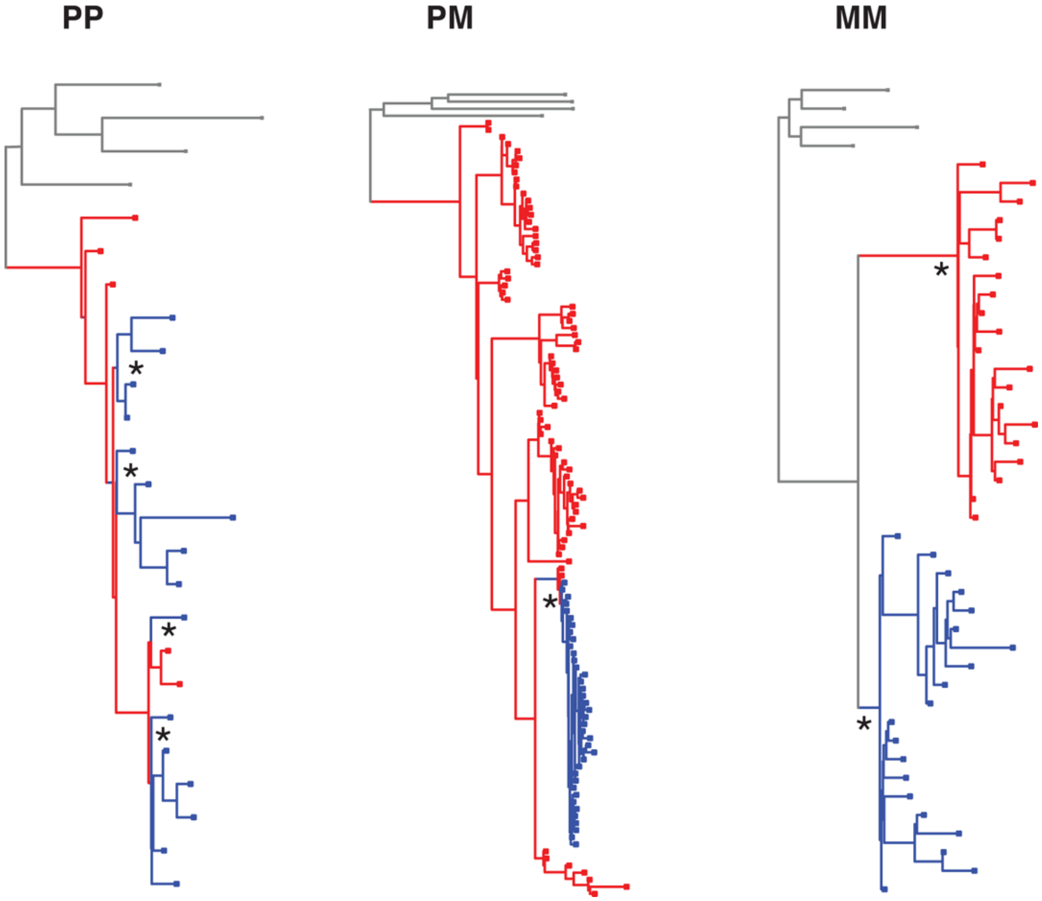

Figure 1. Real examples of PP, PM, and MM trees.

The PP tree comes from a MTCT transmission, the PM tree from a known HET discordant couple transmission, and the MM tree from a HET common source transmission (two recipients from the same known donor source). Each tree shown was randomly selected from 30,000 Bayesian posterior phylogenies per epidemiological pair after burn-in, reconstructed with MrBayes[38], where the topological class had >95% posterior support. The detected recipient lineages are labelled with an asterisk. HIV taxa from two epidemiologically linked hosts are in red or blue, respectively. In PP and PM trees the donor's population is red. Subtype references are in grey. The subtype references correctly root the donor-recipient tree. In PP and PM trees the donor HIV population is paraphyletic, encompassing the recipient's HIV population. In a PP tree, the recipient's HIV population is polyphyletic, in this example with 4 detected clades. In a PM tree, there is only 1 detected clade in the recipient and thus this recipient's population is monophyletic. In a MM tree both patients' HIV populations are monophyletic.

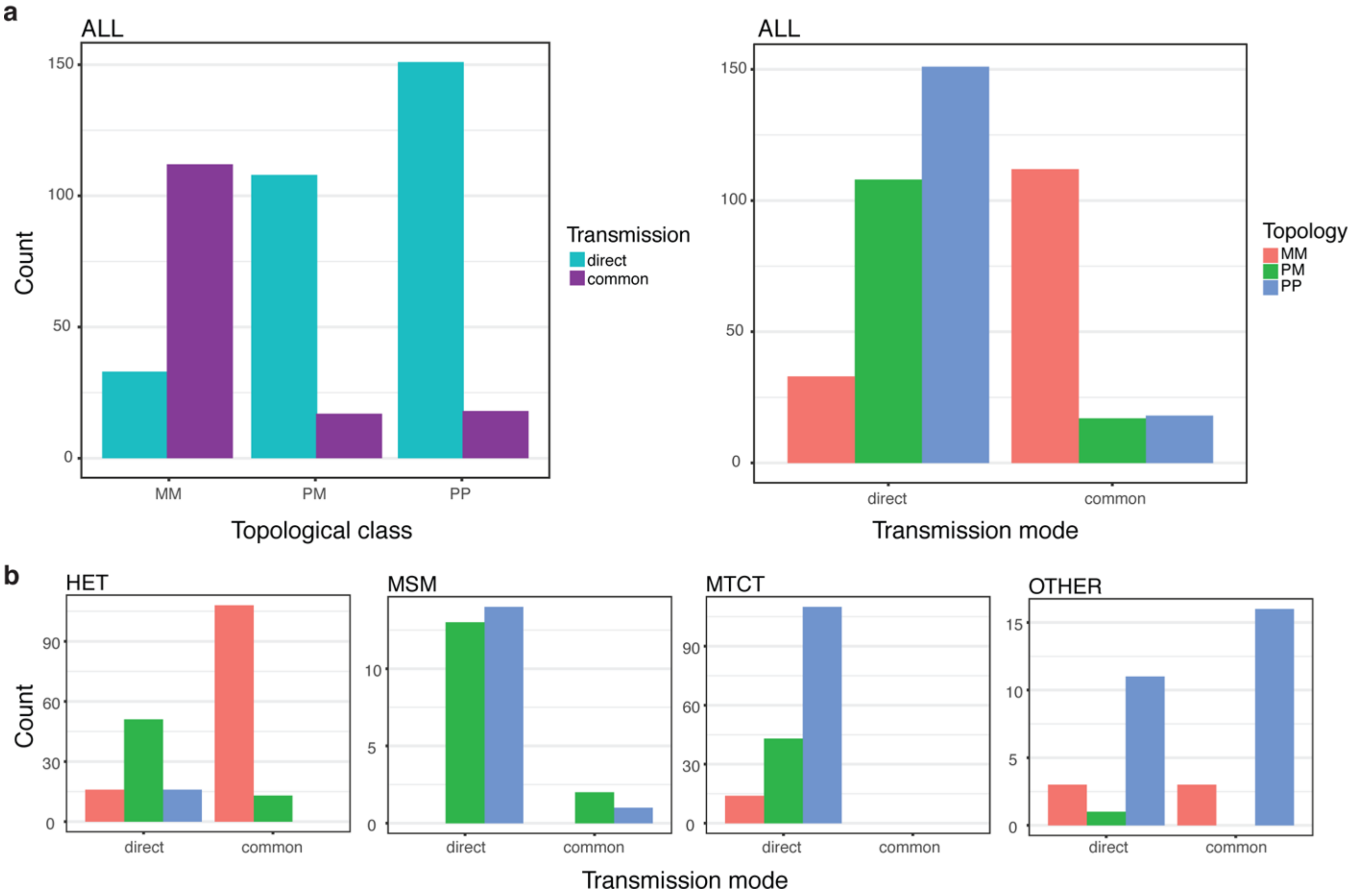

Figure 2. Association of phylogenetic topology and transmission mode.

a, Transmission mode (direct or common source transmission) conditional on topological class (MM, PM, PP) [left panel], and conversely topological class conditional on transmission mode [right panel]. The bars summarize our observations from all transmission risk groups, subtypes, and genomic regions when it was known that transmissions were direct or from a common source, and with good phylogenetic reconstruction (subtype outgroup monophyly at >95% posterior support, and topological support also at >95%; N=438 datasets). **b,** Topological class conditional on transmission mode per risk group. HET, heterosexual transmission; MSM, male homosexual transmission; MTCT, mother to child transmission; OTHER, all other and mixed risk transmissions. Notice different scales.

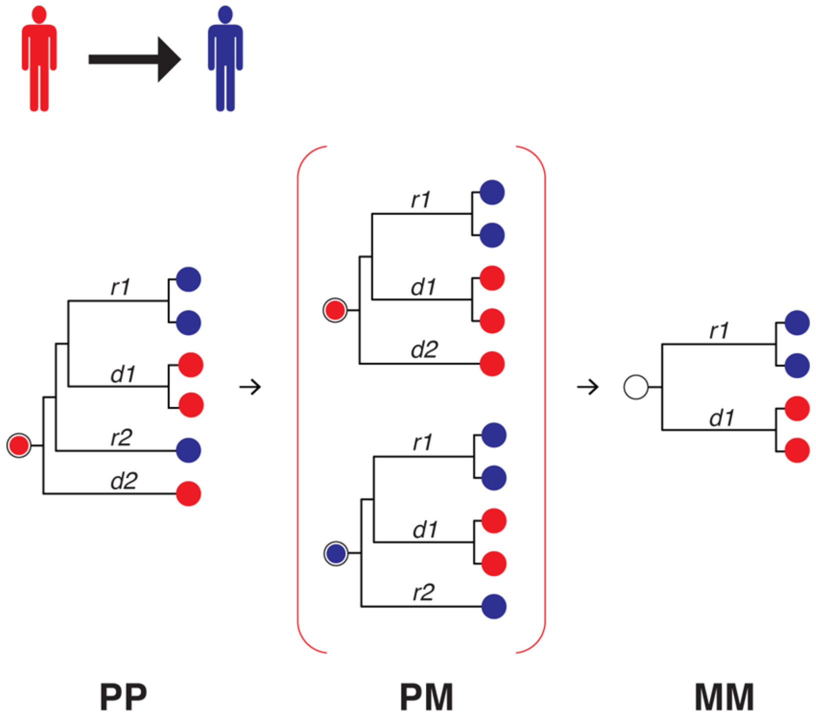

Figure 3. Principal decay of paraphyletic signal.

If one patient (red) infects another (blue), virus in the blue patient should ideally be a subset of the red population, i.e., the red HIV population will be paraphyletic to the blue population. This can be manifested as a PP tree when >1 lineage is transmitted, or a PM tree if only 1 lineage is transmitted (or, theoretically, in a rare instance the oldest lineage in red is transmitted and then dies in red, which would form a MM tree). If a PP tree resulted from the transmission, both lineage death and inadequate sampling could result in a PM tree at time of sampling. Depending which lineage(s) that dies or were not sampled, the observed PM tree could have a host root-label that is incongruent with the true ancestral population (when lineage d2 is not sampled, resulting in blue inferred at root node). Theoretically, under a neutral model, it should be less likely that the sampled PM tree is incongruent, see Figure 4A for an empirical examination and confirmation of this prediction. Eventually, after longer time resulting in more lineage death or a more limited sample (both older lineages, r2 and d2, are unsampled), the tree becomes a MM topology, the absorbing topological state in this phylogenetic system. The MM topology does not allow for an unambiguous root host-label reconstruction, i.e., it cannot infer who the donor was (white node). Starting from a true PM transmission, it should again be more likely that the root host-label in such a tree is congruent with the true ancestral population, also examined and confirmed in Figure 4A.

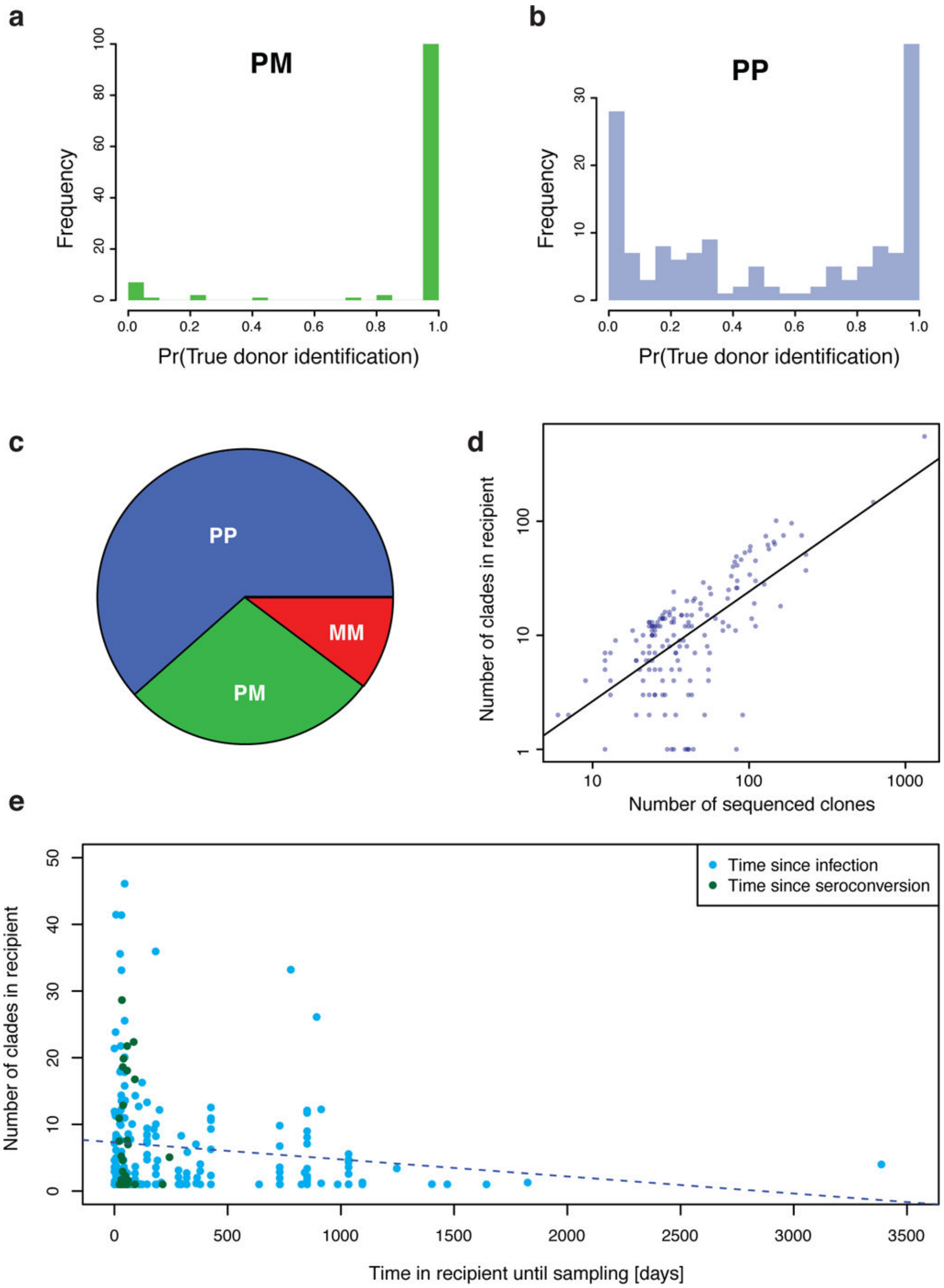

Figure 4. Analyses of the empirical posterior probability of observing the known donor as the root host-label, and PP signal over genomic region, in response to number of sequenced clones, and time since infection.

a, Distribution of the donor host-label posterior support of PM and **b,** PP trees in known direct transmission pairs. Only trees with PM or PP topology posterior support >95%, respectively, were examined (N=262 transmission pairs). Bars represent bins every 5% from 0 (incongruent host-label inferred at root) to 1 (congruent host-label at root). These two distributions are very different ($p<10^{-15}$, two-sided, two-sample Kolmogorov-Smirnov test). **c,** Conditional on observing a PP tree in one genomic region, the pie chart shows the fraction of the detected topological class in another genomic region for transmission pairs that were sequenced in >1 genomic region (N=39 datasets). **d,** Across all transmission pairs with PP trees (PP posterior probability >95%; N=229), the number of observed clades in the recipient grew linearly as more sequences were analyzed from the donor-recipient pairs (R^2=0.44, p<10^{-15}, log-log linear regression with two-sided t test). **e,** As time proceeds from the time of transmission, lineages are lost in the recipient (and donor). Times are based on known time of infection or seroconversion of the recipient (N=231 datasets). The dashed line shows the linear trend (p=0.050, linear regression with two-sided t test).