Phylogenetic methods inconsistently predict direction of HIV transmission among heterosexual pairs in the HPTN052 cohort

Rebecca Rose[1], Matthew Hall[2], Andrew D. Redd[3], Susanna Lamers[1], Andrew E. Barbier[1], Stephen F. Porcella[4], Sarah E. Hudelson[5], Estelle Piwowar-Manning[5], Marybeth McCauley[6], Theresa Gamble[6], Ethan A. Wilson[7], Johnstone Kumwenda[8], Mina C. Hosseinipour[9], James G. Hakim[10], Nagalingeswaran Kumarasamy[11], Suwat Chariyalertsak[12], Jose H. Pilotto[13], Beatriz Grinsztejn[14], Lisa A. Mills[15], Joseph Makhema[16], Breno R. Santos[17], Ying Q. Chen[7], Thomas C. Quinn[2], Christophe Fraser[2], Myron S. Cohen[9], Susan H. Eshleman[5], Oliver Laeyendecker[2]

1 BioInfoExperts, LLC, Thibodaux, LA, USA

2 Big Data Institute, Oxford University, Oxford, UK

3 Laboratory of Immunoregulation, Division of Intramural Research, National Institute of Allergy and Infectious Diseases (NIAID), National Institutes of Health (NIH), Baltimore, MD, USA

4 Genomics Unit, Research Technologies Section, Rocky Mountain Laboratories, Division of Intramural Research, NIAID, NIH, Hamilton, MT, USA

5 Dept. of Pathology, Johns Hopkins Univ. School of Medicine, Baltimore, MD, USA

6 Science Facilitation Department, FHI360, Durham, NC, USA

7 Vaccine and Infectious Disease Science Division, Fred Hutchinson Cancer Research Institute, Seattle, WA, USA

8 College of Medicine-Johns Hopkins Project, Blantyre, Malawi

9 Dept. of Medicine, Univ. of North Carolina at Chapel Hill, Chapel Hill, NC

10 University of Zimbabwe, Harare, Zimbabwe

11 YRGCARE Medical Centre, Chennai, India

12 Research Institute for Health Sciences, Chiang Mai University, Chiang Mai, Thailand

13 Hospital Geral de Nova Iguaçu and Laboratorio de AIDS e Imunologia Molecular (IOC/Fiocruz), Rio de Janeiro, Brazil

14 Instituto Nacional de Infectologia Evandro Chagas-INI-Fiocruz, Rio de Janeiro, Brazil

15 Centers for Disease Control and Prevention (CDC) Division of HIV/AIDS Prevention/KEMRI–CDC Research and Public Health Collaboration HIV Research Branch, Kisumu, Kenya

16 Botswana Harvard AIDS Institute, Gabarone, Botswana

17 Servico de Infectologia, Hospital Nossa Senhora da Conceicao/GHC, Porto Alegre, Brazil

Corresponding author:

Oliver Laeyendecker MS, MBA, PhD

Staff Scientist, NIAID, NIH

Assistant Professor of Medicine, SOM, JHU

Assistant Professor of Epidemiology, JHSPH

855 North Wolfe St.

Rangos Building, room 538A

Baltimore MD, 21205

Phone: 410-502-3268

Email: olaeyen1@jhmi.edu

summary

We evaluated accuracy of phylogenetic methods to predict direction of HIV transmission in 33 partner-index pairs. Results showed that phylogenetic methods which fail to incorporate phylogenetic uncertainty may provide useful information for population-level analyses but are insufficient in legal contexts.

ABSTRACT

Background

We evaluated use of phylogenetic methods to predict the direction of HIV transmission.

Methods

For 33 index-partner pairs with genetically-linked infection, samples were collected from partners and indexes close to time of partners' seroconversion (SC); 31 indexes also had an earlier sample. Phylogenies were inferred using *env* next-generation sequences (one tree per pair/subtype). Direction of transmission (DoT) predicted from each tree was classified as correct or incorrect based on which sequences (index or partner) were closest to the root. DoT was also assessed using maximum-parsimony to infer ancestral node states for 100 bootstrap trees.

Results

DoT was predicted correctly for both single pair and subtype-specific trees in 22 pairs (67%) using SC samples and 23 pairs (74%) using early index samples. DoT was predicted incorrectly for four pairs (15%) using SC or early index samples. In the bootstrap analysis, DoT was predicted correctly for 18 pairs (55%) using SC samples and 24 pairs (73%) using early index samples. DoT was predicted incorrectly for seven pairs (21%) using SC samples and four pairs (13%) using early index samples.

Conclusions

Phylogenetic methods based solely on tree topology of HIV *env* sequences, particularly without consideration of phylogenetic uncertainty, may be insufficient for determining DoT.

KEY WORDS: networks, epidemiology, viral dynamics

INTRODUCTION

The rapid evolutionary rate of HIV can be used to identify transmission groups based on genetic similarity of HIV [1]. HIV network studies often seek to identify genetically-linked infections, determine when transmission occurred, and infer the likely source of infection. Such studies have provided information about social, community, and global HIV transmission networks [2-6] and informed the design of HIV prevention interventions and interpretation of HIV prevention studies [3, 4, 7]. Phylogenetic analysis of HIV has also been used in court cases to determine the genetic linkage and direction of transmission; however, a great deal of caution is needed when interpreting results of phylogenetic analyses in legal settings [8-12]. Results can be significantly impacted by methodological factors, including the choice of model, the sequencing method, genetic distance threshold, time since infection, and the methods used to address ambiguous nucleotides in sequence alignments [13-16].

Transmission clusters of HIV infections are typically defined using either genetic distance measures alone (e.g. [6, 14]) or in conjunction with branch support values (e.g. [17, 18]). It is possible to confirm genetic linkage if appropriate local controls are included in the analysis and if extensive contact tracing is performed; however, it is often impossible to rule out the possibility that additional linked individuals remain un-sampled [11]. In legal settings, analysis of genetic linkage between two persons should include as many sequences as possible from the local outbreak [11]. However, there are no clear guidelines on the number or relatedness of the reference sequences necessary for accurate determination of the direction of HIV transmission between two individuals.

HIV genetic diversity is often assumed to correlate with time since infection [6, 19, 20]. More sophisticated models that incorporate time-sampled sequences can account for variation in the evolutionary rate and more accurately predict the timing of transmission events [21, 22]. These molecular clock methods are appropriate for small datasets (e.g. consensus sequences from

cross-sectional population surveys or clonal sequences from a few potentially-linked cases [3, 4, 23, 24]). However, inferring the timing of HIV transmission events is complicated by the preferential transmission of ancestral viruses [25] and differences in intra- and inter-host evolutionary rates [26]. Transmission models that take these factors into account may provide greater accuracy [25].

Direction of transmission is difficult to assign using phylogenetic methods since many factors may confound the analysis, including variable viral population size, heterogeneous evolutionary rates, on-going reinfection between long-term partners, unidentified additional partners, drug-resistant mutations creating parallel evolution, transmission of multiple and/or recombinant variants, lack of phylogeny branch support, an inadequate number of sequences and/or time points from the potential donor/recipient, and insufficient sequences from other infected individuals from the local outbreak (i.e. the "background" sequences [11, 27, 28]. However, tree topologies may provide some information [9, 29]. Two informative characteristics of phylogenetic trees are placement of the ancestral node and topological pattern (e.g., monophyly, polyphyly, and paraphyly; see Methods) [9, 30]. The concordance between topological pattern and direction of transmission were substantiated in retrospective analyses of two court cases [9], simulated datasets [31] and, most recently, in documented transmission pairs [30].

Here, we evaluated the accuracy of phylogenetic methods to prediction the direction of transmission in 33 index-partner pairs from the HIV Prevention Trials Network (HPTN) 052 clinical trial [32-34]. The analysis was performed using HIV *env* sequences obtained with next-generation sequencing (NGS). This data set was ideally suited for this study, since the 33 index-partner pairs were previously shown to have genetically-linked infections, and since direction of transmission was known for all pairs (index participants were HIV-infected at study enrollment; partners acquired HIV infection during the study).

HIV sequences were analyzed using different sample sets and phylogenetic methods. All analyses were performed using partner samples collected near the time of the partner's seroconversion (SC). Two different index sample sets were compared: (1) index samples collected close to the time of the partner's seroconversion (SC/SC sample set), and (2) index samples collected at an earlier study visit (early/SC sample set). In the first method, maximum likelihood (ML) trees were inferred using sequences from each index-partner pair as well as using sequences from all index-partner pairs of the same HIV subtype. The direction of transmission was predicted by determining which sequences (index or partner) were closer to the root of the tree based on topological patterns. The second method used maximum parsimony to infer the state of the ancestral node in 100 bootstrap replicates for each index-partner pair.

METHODS

Study cohort and HIV sequences. HIV sequence data was obtained from samples collected in the HPTN 052 clinical trial [34]. This trial enrolled HIV-serodiscordant pairs and assessed the impact of early antiretroviral treatment (ART) on HIV transmission. A full description of the study protocol and institutional review board oversight is available in the original publication [34]. This report includes analysis of sequence data from HIV-infected participants ("indexes") and participants who acquired HIV infection in the trial ("partners"). HIV sequences from the index-partner pairs included in this study were previously shown to be genetically linked [32, 33]. Genetic linkage of most index-partner pairs was based on phylogenetic and Bayesian analysis of HIV *pol* sequences obtained by bulk Sanger sequencing; in selected cases, linkage was confirmed by neighbor-joining tree analysis of next-generation sequencing (NGS) using 454-Roche Biotechnology [32, 33]. This study only included pairs where both index and partner were infected with a single HIV strain.

The study included index-partner 33 pairs. Samples from newly-infected partners were obtained at the visit when seroconversion was documented or at the next study visit (SC sample, median 91 days after the last HIV negative visit, range 84 to 588 days); partners were not followed in the HPTN 052 trial after HIV infection was confirmed. Index seroconversion (SC) samples were collected at the visit closest to the visit where the partner's infection was documented (>90 prior to partner's SC: N=9; 0-90 days prior to partner's SC: N=14; 1-90 days after partner's SC, N=10; **Figure 1**). For 31 pairs, an additional earlier index sample was also available which was analyzed separately (range: 84-1174 days prior to the index SC sample; median: 362 days prior to the index SC sample).

A total of 450,336 NGS-derived reads from the *env* gene (nucleotides 7941-8264 relative to HXB2) were obtained from plasma samples from the 33 index-partner pairs [34, 35]. From these reads, 9,051 consensus sequences ("sequences") were generated using GS Amplicon Variant Analyzer, version 2.5 (Roche); each consensus sequence represented a cluster of >=10 individual reads. Each sample had an average of 91 sequences representing 4,503 reads. Sequence alignments were manually edited by codon using AliView [36] and frameshift insertions were removed. Sequences were subtyped using REGA (http://dbpartners.stanford.edu:8080/RegaSubtyping/stanford-hiv/typingtool). Reference sequences were obtained from the Los Alamos HIV Database (https://www.hiv.lanl.gov/).

Single-tree method. For each index-partner pair, separate sequence alignments were constructed from two sample sets: 1) partner SC samples with index SC samples (SC/SC sample set), and 2) early index samples with partner SC samples (early/SC sample set). In addition, sequences from all pairs of the same subtype (A1: N=2; AE: N=1; B: N=3, C: N=27) were combined using the SC/SC sample set and the early/SC sample set (pairs with subtypes A1 and AE were combined, giving a total of six analyses, two for each subtype). ML trees were inferred for all alignments using the HKY model of nucleotide substitution with gamma-

distributed among-site variation using PhyML [37] in the Geneious software package (www.geneious.com) and RaxML v8.2.9 [38].

The direction of transmission for each assessment was independently scored by two investigators as described in [31] and discrepancies in scoring were reconciled by a third party. Three topological patterns were assessed: (1) both subjects were monophyletic (all sequences from a given participant shared a common ancestor which excluded sequences from any other subject); (2) paraphyletic/monophyletic (the monophyletic clade of one subject shared a common ancestor with some, but not all, of the other subject's sequences); and (3) paraphyletic/polyphyletic (a mixed clade containing all sequences from one subject shared a common ancestor with some, but not all, of the other subject's sequences; **Supplemental Figure 1**). The direction of transmission was scored as "correct" if index sequences were paraphyletic and partner sequences were monophyletic/polyphyletic, and "incorrect" if partner sequences were paraphyletic and index sequences were monophyletic/polyphyletic. If sequences from both index and partner were monophyletic, the direction of transmission was scored as "equivocal".

Bootstrapping method. For each of the 33 pairs, separate sequence alignments were constructed for the two sample sets (SC/SC sample set; early/SC sample set). All alignments also included a reference set consisting of a single random sequence from each of the other index-partner pairs and the HXB2 sequence for rooting. One hundred bootstrap phylogenies of each alignment were generated with RAxML v8.2.9 [38]. For each phylogeny, *Phyloscanner* v1.6.4 was used to infer the ancestral state of each of the internal nodes of the tree using a modified maximum parsimony procedure [39]. Ancestral states were classified as index, partner, or an "unsampled" state representing either a third party or an unclear ancestry.

For each of the 100 trees generated for an index-partner pair, we identified the earliest node(s) in the tree (i.e. the node that had no ancestral nodes with a sampled state). The state of this

node (i.e. index, partner, unsampled) was considered to represent the transmitting subject for that tree.

If there was no such node (i.e. separate clades from each subject with no implied ancestry), then the tree was labelled as 1) “equivocal” if there were no tips from the reference set descended from the most recent common ancestor node of both patients, or 2) “unlinked” if there was at least one tip (**Supplemental Figure 2**).

For each index-partner pair, the direction of transmission was assigned as follows: (1) “correct” if the state of the earliest node was classified as the index in at least twice as many trees as those where it was classified as partner; (2) “incorrect” if the state of the earliest node was classified as partner in twice as many trees as those where it was classified as index; (3) “unlinked” if at least one tip from the reference set was descended from the earliest node in more than half of the trees; and (4) “indeterminate” for all other cases.

RESULTS

Transmission direction predicted using the single-tree method. For each of the 33 index-partner pairs, we first evaluated the predicted direction of transmission using the single-tree method. Two trees were evaluated for each index-partner pair: individual (only sequences from that pair plus subtype reference sequences) and subtype-specific (all sequences of the same subtype combined). The analysis was first performed using the SC/SC sample set (index and partner samples collected near the time of the partner’s seroconversion visit). The predicted direction of transmission was correct in both trees (individual and subtype-specific) for 22 pairs (67%) and incorrect in both trees for four pairs (12%). Trees were discordant for the remaining seven pairs (**Figure 2, Supplemental Table 1**).

The analyses described above were next performed using the early/SC sample set available for 31 pairs (partner samples collected near the time of the partner's seroconversion visit; index samples collected at an earlier visit) (**Figure 2**). The predicted direction of transmission was correct for both trees for 23 pairs (74%), incorrect for both trees for four pairs (13%), and discordant and/or equivocal for four pairs.

We then compared results obtained using the SC/SC and early/SC sample sets to determine whether the timing of sample collection impacted the prediction of the direction of transmission. This analysis was performed for the 31 index-partner pairs who had results from an early index sample. Nineteen pairs (61%) had the correct direction of transmission predicted in both trees (single and subtype-specific) for both sample sets. Four pairs (13%) had the incorrect direction predicted for both trees for both sample sets. Two pairs had the incorrect direction predicted in one tree for both sample sets. The remaining six pairs had discordant predictions for the two sample sets. Incorrect and/or equivocal predictions did not appear to be correlated with the time between collection of the index SC sample and the corresponding partner SC sample.

Transmission direction predicted using the bootstrapping method. We next evaluated the accuracy of predictions of the direction of transmission using replicate bootstrap trees (i.e. bootstrap support for the predicted direction using the modified maximum parsimony approach implemented in *Phyloscanner)*. For the SC/SC sample set, the direction of transmission was predicted correctly for 18 pairs (55%), incorrectly for seven pairs (21%), and indeterminate for eight pairs (24%). (**Figure 3a**, **Supplemental Table 1**). For the 31 pairs with the early/SC sample set, the direction of transmission was predicted correctly for 24 pairs (73%), incorrectly for four pairs (13%), and indeterminate for three pairs (12%) (**Figure 3b**). Sixteen pairs (52%) had the correct direction predicted for both sample sets, and three pairs (10%) had the incorrect direction predicted for both sample sets. The predicted direction of transmission for the remaining 12 pairs was either indeterminate or inconsistent between sample sets.

Comparison of predictions from the single-tree and bootstrap methods. In general, results from the two approaches (i.e. single trees vs. bootstrapped trees) were consistent. For the SC/SC sample set, both the bootstrap method and the single-tree method for both individual and subtype-specific trees predicted the correct direction of transmission for 15 pairs (45%) and the incorrect direction in three pairs (9%) (**Figure 3; Supplemental Table 1**). The bootstrap method predicted an indeterminate direction for five pairs (0103, 2912, 0061, 0693, 3108) that were correctly assessed using the single-tree method for both trees, and for one pair (1170) that was incorrectly assessed using the single-tree method for both trees. The bootstrap analysis also predicted the incorrect direction in two pairs (0645, 3283) that were correctly assessed in both trees using the single-tree method. For one of these pairs (0645), approximately 25% of the bootstrap trees predicted the correct direction of transmission; for the other pair (3283), none of the bootstrap trees predicted the correct direction of transmission. Of note, the single trees for this pair (3283) showed that only one index sequence was basal to the whole clade; the remaining index sequences clustered together elsewhere.

For the seven pairs where results from both single trees were inconsistent, the bootstrap method predicted a correct direction in three pairs (2515, 1018, 2318), incorrect in two pairs (3179, 2899), and remained indeterminate for two pairs (2180, 0452 and 3108).

For the early/SC sample set, both the bootstrap method and the single-tree method for both individual and subtype-specific trees predicted the correct direction of transmission for 21 pairs (68%) and the incorrect direction in three pairs (10%) (**Figure 3; Supplemental Table 1**). The bootstrap method predicted an indeterminate direction for two pairs (0103, 2515) that were correctly assessed using the single-tree method for both trees, and for one pair (1170) that was incorrectly assessed using the single-tree method for both trees. For the four pairs with inconsistent or equivocal single trees, the bootstrap analysis predicted a correct direction of transmission for three pairs (0452, 2180, 2187), an incorrect direction for one pair (0061). Taken

together, these results suggest that using a single tree may overestimate the number of cases that are correctly classified for the direction of transmission.

DISCUSSION

We evaluated the accuracy of using tree topology to predict the direction of HIV transmission in 33 genetically-linked index-partner pairs with known direction of transmission. We compared different phylogenetic methods (single tree method; bootstrap method), different sampling strategies (individual index-partner pairs; subtype-specific analysis), and different sample sets (index samples collected near the time of partner seroconversion, or earlier).

The direction of transmission was predicted correctly for both individual and subtype-specific trees in 67% of index-partner pairs when both samples were collected near the time of partner seroconversion (SC/SC sample set). Similarly, direction of transmission was predicted correctly for both trees in 74% of index-partner pairs when the analysis was performed using index samples collected at an earlier date (early/SC sample set). The direction of transmission was predicted correctly for only 61% of the index-partner pairs for both trees and both sample sets. It is concerning that the direction of transmission was predicted incorrectly in 13% of index-partner pairs for both trees and sample sets. In some cases, conflicting results were obtained for the two tree types (individual, subtype-specific); this suggests that the choice and/or number of background sequences may be an important factor in topological reconstruction.

The proportion of cases in this study where the direction of transmission was predicted correctly was lower than that reported in previous studies that used a similar method of basing prediction of direction of transmission using topological patterns [30, 31]. However, low branch support could produce an incorrect result by chance placement of one or a few sequences. To address this, we compared results obtained with the single tree method to results obtained using a

maximum parsimony-based method to infer the state of the ancestral node for 100 bootstrap replicates for each pair. In this analysis, the direction of transmission was correctly predicted for only 18 index-partner pairs (55%) using the SC/SC sample set, and 24 pairs (73%) using the early/SC sample set. Only 16 pairs (52%) had the correct direction predicted using both sample sets. The lower percentage of correct predictions using the bootstrap method demonstrates the potential of stochasticity to skew inferences and suggests that using only a single tree may over-estimate confidence in determining the correct direction of transmission. Additional metrics (e.g., the viral genetic diversity of host vs. recipient) could potentially provide additional information that could enhance phylogenetic methods; however, this avenue has yet to be explored fully.

While both the single tree and the bootstrap methods predicted the correct direction of transmission in more trees using the early/SC sample sets compared to the SC/SC sample sets, generally, there was no clear trend between predicted direction and the timing of index samples relative to the partner's sample. Because partners were not followed in the trial after infection was confirmed, we were not able to evaluate performance of the methods for predicting the direction of transmission when partner samples were collected from individuals with longer-term infections.

It is possible that some other factor specific to the HPTN 052 trial could have impacted our results. While most of the pairs studied in this report were infected with HIV subtype C (N=27), both correct and incorrect predictions were found for pairs of three different subtypes (A1, B, and C), which suggests that subtype is not a major factor impacting the accuracy of the methods used. Differences in rates of evolution and population growth of the virus may be a factor [27], which could result from ART (although only one of the 64 index samples was collected after the index started ART).

Other factors that may have influenced the accuracy of these methods include the sequence length and genomic location of the *env* sequences analyzed. While diversity of the *env* region likely enhanced the phylogenetic signal, selection bias during sample preparation might have resulted in more frequent variants being preferentially amplified. The HIV *env* gene is also subject to within-host selection pressure, which may have resulted in homoplasies caused by convergent evolution (i.e. identical but independent changes) and/or lost variation; both of these factors could have potentially masked true transmission patterns. Additionally, recombination occurring during amplification/sequencing could have also resulted in homoplasies. We are currently investigating the accuracy of these methods for predicting the direction of transmission using full HIV genome sequences (using methods similar to those described in Wymant *et al* [39]).

The findings here are particularly important because data from phylogenetic analyses have been used as evidence in the criminal and civil justice systems in cases of suspected HIV transmission [11]. Since the repercussions of incorrect conclusions are potentially severe in legal settings, considerable effort has been invested in assessing the appropriateness and accuracy of phylogenetic methods used to assess genetic linkage, timing, and direction of HIV transmission [11]. It is widely acknowledged that current methods are best used for excluding potential persons as the source of infections, and/or for assessing the duration of HIV infections, rather than for determining the direction of transmission (e.g., between a plaintiff and the person suspected of being the source of the plaintiff's infection). Our results strongly indicate that methods to determine the direction of HIV transmission based solely on tree topology of HIV *env* sequences, particularly without consideration of phylogenetic uncertainty, should be considered insufficient for forensic or legal applications, especially in settings where additional epidemiological information is unavailable. However, these methods may provide useful insights

in the context of population level analyses (e.g., to identify factors associated with increased transmission risk).

ACKNOWLEDGMENTS

This project was supported by: (1) The HIV Prevention Trials Network (HPTN) sponsored by the National Institute of Allergy and Infectious Diseases (NIAID), the National Institute on Drug Abuse (NIDA), the National Institute of Mental Health (NIMH), and the Office of AIDS Research, of the National Institutes of Health (NIH), Dept. of Health and Human Services (DHHS) [grant numbers UM1AI068613 (HPTN Network Laboratory – Susan Eshleman, PI), UM1AI068617 (HPTN Statistical and Data Management Center – Deborah Donnell, PI), and UM1AI068619 (HPTN Core and Operations Center – Wafaa El-Sadr, PI)], and (2) the Division of Intramural Research, NIAID, NIH. This project has been funded in whole or in part with federal funds from the National Cancer Institute, National Institutes of Health, under Contract No. HHSN261200800001E.

The content of this publication does not necessarily reflect the views or policies of the Department of Health and Human Services, nor does mention of trade names, commercial products, or organizations imply endorsement by the U.S. Government.

The authors of this manuscript declare no conflict of interest.

This work was accepted as a poster presentation at the 22nd International AIDS Conference, held in the Netherlands, July 2018.

The authors acknowledge the dedication, commitment, and efforts of the entire HPTN 052 team, and acknowledge the invaluable contributions of the participants in the HPTN 052 trial. The authors thank the laboratory staff at Johns Hopkins University, at the Rocky Mountain Laboratories, and at the HPTN 052 study sites for assistance with sample and data management. The authors also thank David Nolan for his assistance with interpretation of data.

1. Hassan AS, Pybus OG, Sanders EJ, Albert J, Esbjörnsson J. Defining HIV-1 transmission clusters based on sequence data. AIDS **2017**; 31:1211-22.

2. Wertheim JO, Kosakovsky Pond SL, Forgione LA, et al. Social and Genetic Networks of HIV-1 Transmission in New York City. PLoS Pathog **2017**; 13:e1006000.

3. Leigh Brown AJ, Lycett SJ, Weinert L, et al. Transmission network parameters estimated from HIV sequences for a nationwide epidemic. J Infect Dis **2011**; 204:1463-9.

4. Hughes GJ, Fearnhill E, Dunn D, et al. Molecular phylodynamics of the heterosexual HIV epidemic in the United Kingdom. PLoS pathogens **2009**; 5:e1000590.

5. Wertheim JO, Oster AM, Hernandez AL, Saduvala N, Bañez Ocfemia MC, Hall HI. The International Dimension of the U.S. HIV Transmission Network and Onward Transmission of HIV Recently Imported into the United States. AIDS Res Hum Retroviruses **2016**.

6. Wertheim JO, Leigh Brown AJ, Hepler NL, et al. The global transmission network of HIV-1. J Infect Dis **2014**; 209:304-13.

7. Wertheim JO, Kosakovsky Pond SL, Little SJ, De Gruttola V. Using HIV transmission networks to investigate community effects in HIV prevention trials. PLoS One **2011**; 6:e27775.

8. Leitner T, Albert J. Reconstruction of HIV-1 transmission chains for forensic purposes AIDS Review **2000**; 2:241-51.

9. Scaduto DI, Brown JM, Haaland WC, Zwickl DJ, Hillis DM, Metzker ML. Source identification in two criminal cases using phylogenetic analysis of HIV-1 DNA sequences. Proc Natl Acad Sci U S A **2010**; 107:21242-7.

10. 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 **2007**; 8:382-7.

11. Abecasis AB, Pingarilho M, Vandamme AM. Phylogenetic analysis as a forensic tool in HIV transmission investigations. AIDS **2018**; 32:543-54.

12. Abecasis AB, Geretti AM, Albert J, Power L, Weait M, Vandamme AM. Science in court: the myth of HIV fingerprinting. Lancet Infect Dis **2011**; 11:78-9.

13. Rose R, Lamers SL, Dollar JJ, et al. Identifying Transmission Clusters with Cluster Picker and HIV-TRACE. AIDS Res Hum Retroviruses **2017**; 33:211-8.

14. Kosakovsky Pond SL, Weaver S, Leigh Brown AJ, Wertheim JO. HIV-TRACE (Transmission Cluster Engine): a tool for large scale molecular epidemiology of HIV-1 and other rapidly evolving pathogens. Mol Biol Evol **2018**.

15. Poon A. Impacts and shortcomings of genetic clustering methods for infectious disease outbreaks. Virus Evol **2016**; 2(2):vew031

16. Le Vu S, Ratmann O, Delpech V, et al. Comparison of cluster-based and source-attribution methods for estimating transmission risk using large HIV sequence databases. Epidemics **2018**; 23:1-10.

17. Prosperi MC, Ciccozzi M, Fanti I, et al. A novel methodology for large-scale phylogeny partition. Nat Commun **2011**; 2:321.

18. Ragonnet-Cronin M, Hodcroft E, Hué S, et al. Automated analysis of phylogenetic clusters. BMC Bioinformatics **2013**; 14:317.

19. Puller V, Neher R, Albert J. Estimating time of HIV-1 infection from next-generation sequence diversity. PLoS Comput Biol **2017**; 13:e1005775.

20. Kouyos RD, von Wyl V, Yerly S, et al. Ambiguous nucleotide calls from population-based sequencing of HIV-1 are a marker for viral diversity and the age of infection. Clin Infect Dis **2011**; 52:532-9.

21. Drummond AJ, Ho SY, Phillips MJ, Rambaut A. Relaxed phylogenetics and dating with confidence. PLoS Biol **2006**; 4:e88.

22. Drummond AJ, Pybus OG, Rambaut A, Forsberg R, Rodrigo AG. Measurably evolving populations. Trends in Ecology and Evolution **2003**; 18:481-8.

23. Lewis F, Hughes GJ, Rambaut A, Pozniak A, Leigh Brown AJ. Episodic sexual transmission of HIV revealed by molecular phylodynamics. PLoS Med **2008**; 5:e50.

24. Jacka B, Applegate T, Poon AF, et al. Transmission of hepatitis C virus infection among younger and older people who inject drugs in Vancouver, Canada. J Hepatol **2016**; 64:1247-55.

25. Vrancken B, Rambaut A, Suchard MA, et al. The genealogical population dynamics of HIV-1 in a large transmission chain: bridging within and among host evolutionary rates. PLoS Comput Biol **2014**; 10:e1003505.

26. Lemey P, Rambaut A, Pybus O. HIV evolutionary dynamics within and among hosts. AIDS Rev **2006**; 8:125-40.

27. Romero-Severson EO, Bulla I, Hengartner N, et al. Donor-Recipient Identification in Para- and Poly-phyletic Trees Under Alternative HIV-1 Transmission Hypotheses Using Approximate Bayesian Computation. Genetics **2017**; 207:1089-101.

28. Lemey P, Derdelinckx I, Rambaut A, et al. Molecular footprint of drug-selective pressure in a human immunodeficiency virus transmission chain. J Virol **2005**; 79:11981-9.

29. Siljic M, Salemovic D, Cirkovic V, et al. Forensic application of phylogenetic analyses - Exploration of suspected HIV-1 transmission case. Forensic Sci Int Genet **2017**; 27:100-5.

30. Leitner T, Romero-Severson E. Phylogenetic patterns recover known HIV epidemiological relationships and reveal common transmission of multiple variants. Nat Microbiol **2018**; 3:983-8.

31. Romero-Severson EO, Bulla I, Leitner T. Phylogenetically resolving epidemiologic linkage. Proc Natl Acad Sci U S A **2016**; 113:2690-5.

32. Eshleman SH, Hudelson SE, Redd AD, et al. Treatment as Prevention: Characterization of Partner Infections in the HIV Prevention Trials Network 052 Trial. J Acquir Immune Defic Syndr **2017**; 74:112-6.

33. Eshleman SH, Hudelson SE, Redd AD, et al. Analysis of genetic linkage of HIV from couples enrolled in the HIV Prevention Trials Network 052 trial. J Infect Dis **2011**; 204:1918-26.

34. Cohen MS, Chen YQ, McCauley M, et al. Prevention of HIV-1 infection with early antiretroviral therapy. N Engl J Med **2011**; 365:493-505.

35. Cohen MS, Chen YQ, McCauley M, et al. Antiretroviral Therapy for the Prevention of HIV-1 Transmission. N Engl J Med **2016**; 375:830-9.

36. Larsson A. AliView: a fast and lightweight alignment viewer and editor for large datasets. Bioinformatics **2014**; 30:3276-8.

37. Guindon S, Delsuc F, Dufayard J, Gascuel O. Estimating maximum likelihood phylogenies with PhyML. Methods Mol Biol **2009**; 537:113-37.

38. Stamatakis A. RAxML version 8: a tool for phylogenetic analysis and post-analysis of large phylogenies. Bioinformatics **2014**; 30:1312-3.

39. Wymant C, Hall M, Ratmann O, et al. PHYLOSCANNER: Inferring Transmission from Within- and Between-Host Pathogen Genetic Diversity. Mol Biol Evol **2017**.

Figure 1. Time between collection of the index and partner samples. Blue dots indicate the timing of collection of the early index sample; red dots indicate the timing of the index SC sample. Data are plotted in days between collection of the index sample and the partner SC sample. Positive values indicate that the index sample was collected before the partner sample; negative values indicate that the index sample was collected after the partner sample. The identifier number for each index-partner pair is shown on the x-axis. Pairs are grouped based on the timing of collection of the index SC sample relative to collection of the partner SC sample. Two pairs did not have an early index sample available for analysis. Abbreviation: SC: seroconversion.

Figure 2. Predicted direction of transmission using the single tree method. Each square/circle represents one index-partner pair. Pairs are grouped based on the timing of collection of the index SC sample relative to collection of the partner SC sample. Squares show data obtained for the SC/SC sample set; circles show data obtained for the early/SC sample set (see text). The identifier number for each index-partner pair is shown to the right of the corresponding square. Colors of the squares/circles correspond to the direction of transmission predicted from individual pair trees and subtype-specific trees. Abbreviation: SC: seroconversion.

Figure 3. Predicted direction of transmission using the bootstrap method (inferred ancestral state of 100 bootstrap trees). Each bar shows the percentage of trees with different predicted ancestral states for 100 bootstrap trees, colored according to the legend. The identifier number for each index-partner pair is shown below each bar. Pairs are grouped based on the timing of collection of the index SC sample relative to collection of the partner SC sample. (A) Trees inferred using the SC/SC sample set. (B) Trees inferred using the early/SC sample set.