nature COMMUNICATIONS

ARTICLE

https://doi.org/10.1038/s41467-019-09139-4 OPEN

Inferring HIV-1 transmission networks and sources of epidemic spread in Africa with deep-sequence phylogenetic analysis

Oliver Ratmann[1,2], M. Kate Grabowski[3,4], Matthew Hall[5], Tanya Golubchik[5], Chris Wymant[2,5], Lucie Abeler-Dörner[5], David Bonsall[5], Anne Hoppe[6], Andrew Leigh Brown[7], Tulio de Oliveira[8], Astrid Gall[9], Paul Kellam[10], Deenan Pillay[6,11], Joseph Kagaayi[4], Godfrey Kigozi[4], Thomas C. Quinn[3,12], Maria J. Wawer[4,13], Oliver Laeyendecker[3,12], David Serwadda[4,14], Ronald H. Gray[3,4,13] & Christophe Fraser[5], PANGEA Consortium and Rakai Health Sciences Program[#]

To prevent new infections with human immunodeficiency virus type 1 (HIV-1) in sub-Saharan Africa, UNAIDS recommends targeting interventions to populations that are at high risk of acquiring and passing on the virus. Yet it is often unclear who and where these 'source' populations are. Here we demonstrate how viral deep-sequencing can be used to reconstruct HIV-1 transmission networks and to infer the direction of transmission in these networks. We are able to deep-sequence virus from a large population-based sample of infected individuals in Rakai District, Uganda, reconstruct partial transmission networks, and infer the direction of transmission within them at an estimated error rate of 16.3% [8.8–28.3%]. With this error rate, deep-sequence phylogenetics cannot be used against individuals in legal contexts, but is sufficiently low for population-level inferences into the sources of epidemic spread. The technique presents new opportunities for characterizing source populations and for targeting of HIV-1 prevention interventions in Africa.

[1] Department of Mathematics, Imperial College London, London SW72AZ, UK. [2] Department of Infectious Disease, Epidemiology School of Public Health, Imperial College London, London W21PG, UK. [3] Department of Medicine, Johns Hopkins School of Medicine, Baltimore, MD 21205-2196, USA. [4] Rakai Health Sciences Program, Entebbe, P.O.Box 49, Uganda. [5] Oxford Big Data Institute, Li Ka Shing Centre for Health Information and Discovery, Nuffield Department of Medicine, Old Road Campus, University of Oxford, Oxford OX3 7BN, UK. [6] Division of Infection and Immunity, University College London, London WC1E 6BT, UK. [7] School of Biological Sciences, University of Edinburgh, Edinburgh EH9 3FF, UK. [8] College of Health Sciences, University of KwaZulu-Natal, Durban 4041, South Africa. [9] European Molecular Biology Laboratory-European Bioinformatics Institute (EMBL-EBI), Wellcome Genome Campus, Hinxton, Cambridgeshire CB10 1SD, UK. [10] Department of Medicine, Imperial College London, London W12 0HS, UK. [11] Africa Health Research Institute, Private Bag X7, Durban 4013, South Africa. [12] Division of Intramural Research, National Institute of Allergy and Infectious Diseases, NIH, Bethesda, MD 20892-9806, USA. [13] Department of Epidemiology, Johns Hopkins Bloomberg School of Public Health, Baltimore, MD 21205, USA. [14] Makerere University School of Public Health, Kampala 8HQG+3V, Uganda. [#] A full list of consortium members appears at the end of the paper. Correspondence and requests for materials should be addressed to O.R. (email: oliver.ratmann@imperial.ac.uk)

Large generalized epidemics of human immunodeficiency virus type 1 (HIV-1) continue to cause substantial mortality and morbidity across much of sub-Saharan Africa[1]. Rates of new infections have been reduced by adoption of prevention measures, especially antiretroviral therapy and medical male circumcision[1,2]. Despite progress, incidence levels remain well above elimination thresholds[3]. There remains an urgent need to better understand the drivers of transmission such as differential transmission by sex and age groups, especially among young women who account for 74% of new infections among adolescents in sub-Saharan Africa[4]. This may enable better targeting of prevention measures to infected people who most likely act as sources of new infection, and thus reduce transmission amongst groups most likely to sustain the epidemic. HIV-1 evolves faster than transmissions occur, so that viral sequences obtained from an individual tend to be characteristic of that individual within weeks after infection[5,6]. Therefore, viral genetic data have the potential to yield novel insights into the drivers of transmission by identifying who may have been a transmitter, and then by generalizing these findings to identify risk factors that can be directly targeted for prevention[7,8].

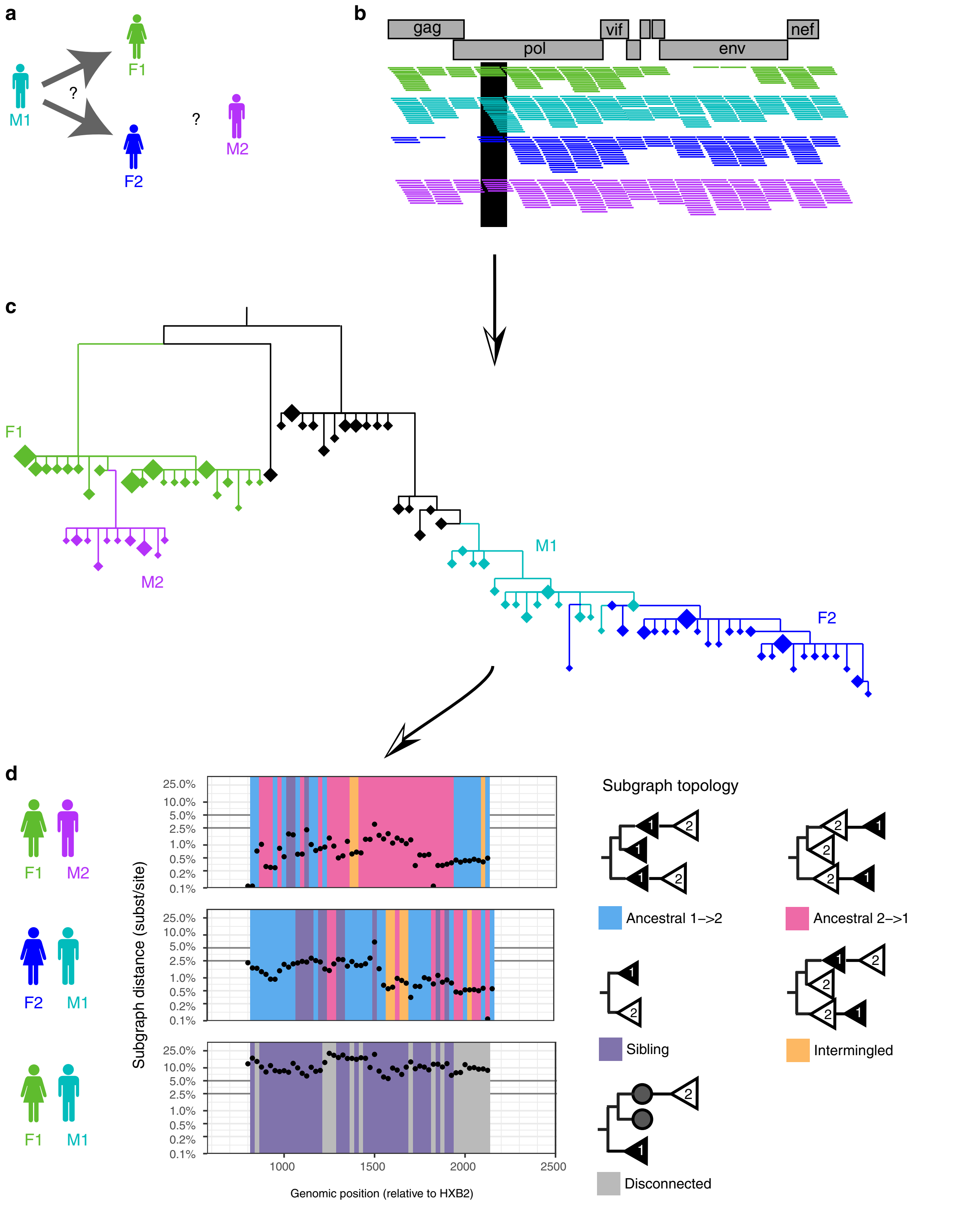
Currently, phylogenetic tools to identify sources of transmission are based on Sanger sequencing, which generates a single HIV-1 consensus sequence per virus sample from an individual[9–13]. Typically one sample per individual is sequenced, and so the entire viral population from one individual is reduced into a single consensus sequence, which is insufficient to determine in which direction infections occurred[14]. For this reason source attribution methods have required data on dates of infection[15–17] or modelling assumptions on the epidemic[9,10,12,18,19]. An advantage of source attribution methods based on additional modelling assumptions is that they may be applied with relatively small sample sizes, although it can be hard to disentangle assumptions from conclusions. For example, in ref. [12], it was assumed that young women are predominantly infected by older men in KwaZulu-Natal, South Africa, and it is unclear to what extent the same conclusion is based on data[20]. There is consequently a need for broadly applicable source attribution methods that are not dependent on external modelling assumptions to provide independent evidence.

Here, we demonstrate that HIV-1 transmission networks and the direction of transmission within them can be reconstructed from deep-sequence data of a large population-based sample of infected individuals with phyloscanner[21], a recently developed software package for viral phylogenetic inference from deep-sequence data. The accuracy in reconstructing the direction of transmission is sufficient to infer source populations, i.e. the most likely drivers of the epidemic, without assumptions on the epidemic. This finding turns into practice the theoretical prediction by Romero-Severson et al.[22] that individuals should be represented by clusters (in short: subgraphs) of viral sequences in phylogenies when many sequence reads per individual are available, and that the phylogenetic ordering of subgraphs should allow inference of the likely direction of transmission between individuals. Figure 1 illustrates this principle. Leitner and Romero-Severson[23] investigated which phylogenetic orderings of subgraphs (in short: subgraph topologies) can be expected among known transmission pairs. The primary aim of this study is the opposite, to establish what epidemiologic inferences can be made from observed patterns in deep-sequence phylogenies. Our population-level analysis is based on deep-sequence data that was cross-sectionally collected from 40 communities in the Rakai region of Southern Uganda. Rakai communities are predominantly small agrarian and semi-urban trading centres as well as fishing communities alongside Lake Victoria. The area was the initial epicentre of the HIV-1 epidemic in Eastern Africa, and today remains among the highest burdened districts in Uganda with an overall adult HIV prevalence that ranges from 9–26% among inland trading and agrarian communities to 38–43% among lakeside fishing communities[24,25].

We report first that it is feasible to obtain population-based samples of HIV-1 deep-sequence data that represent a large proportion of infected individuals with unsuppressed virus in a local setting in Africa. Second, we demonstrate that deep-sequence phylogenetic analysis can be scaled from pairs in whom transmission has been suspected to population-based samples of HIV-1 epidemics. We reconstruct partial transmission networks in the absence of self-reported sexual contact information and identify pairs of individuals in whom transmission and the direction of transmission is phylogenetically inferred with high statistical support, which we call source—recipient pairs. Third, we assess the strength of deep-sequence phylogenetic inferences on direct transmission between two individuals (in short: linkage) in a large population-based sample, and the direction of transmission between two individuals via potentially unsampled intermediates. Our major finding is that the direction of transmission from a source case to a recipient could be frequently estimated with high statistical support, and that accuracy levels are sufficient for inferences into the drivers of epidemic spread at the population-level.

Results

Large deep-sequence data set of an African HIV-1 epidemic. Between August 2011 and January 2015, 25,882 individuals aged 15–49 years were surveyed in 40 communities of the Rakai Community Cohort Study (RCCS) in Uganda (Table 1). The survey included the four largest fishing sites along Lake Victoria because of their high population-level HIV prevalence (~40%)[25] and hypothesized role in epidemic spread. 5142 participants were HIV-positive. Reflecting previous guidelines on initiation of antiretroviral therapy (ART) during the observation period, 3878 (75.4%) infected study participants reported no ART use at time of survey. Self-reported ART use was previously validated as a proxy for actual ART use[26], and 90% of individuals who reported using ART also had suppressed virus titres below 1000 copies per millilitre plasma blood[2]. This prompted us to focus on viral sequencing among individuals who did not report ART use. Deep-sequencing of the virus genomes was performed on 3758/3878 (96.9%) samples using the Gall et al. protocol[27], generating thousands of short viral sequence fragments (reads) per individual. Sequencing success was comparatively modest[28]. We restricted our analysis to samples from 2652 individuals that satisfied minimum criteria on read length and depth for phylogeny reconstruction and subsequent inferences (see Methods and Supplementary Figure 1). Women and individuals of 35 years or more were under-represented in this data set when compared to infected participants, whereas individuals in fishing sites were over-represented. The overall sequence sampling fraction was high, 68.4% (2652/3878) among infected participants who did not report ART use (Fig. 2), and an estimated 65.6% (2652/4043) among infected participants with unsuppressed virus (see Methods). If we assume that individuals who were not present or did not participate at survey visits were infected with unsuppressed virus in proportion to the enrolled population, an additional 1837 individuals likely did not have suppressed viraemia, leading to an estimated sequence sampling fraction of 45.1% (2652/5880) among eligible, infected individuals with unsuppressed virus. Accounting for the previous finding that ~30% of individuals were infected by a person outside the cohort[11], we thus expect that in approximately three of ten cases (0.451 × 0.7), our data contain the transmitter of a sequenced individual.



Scaling deep-sequence phylogenetics to large data sets. We first investigated the types of deep-sequence phylogenetic patterns that arise in known epidemiologic relationships. Our population-based sample comprised 331 concordant HIV-1-positive couples who self-identified as sexual partners. Based on previous partner analyses[16,17], we expected that virus was transmitted in approximately 70% of couples, and that the remaining couples were separately infected by other individuals. Figure 1d illustrates a typical scan of deep-sequence phylogenies across the genome for three male—female pairs. In each phylogeny, subgraphs of reads from two individuals could either be ancestral to each other (pink if virus of the female was ancestral and blue if virus of the male was ancestral), siblings (purple), intermingled (yellow), or disconnected by one or more other individuals (grey, see Methods for full definitions and Supplementary Tables 1–3 for command line specifications of the phyloscanner software). In addition, the

Fig. 1 Inferring the direction of transmission from HIV-1 deep-sequence data. **a** The principles of deep-sequence viral phylogenetic analysis are illustrated on data from male M1 (turquoise) who initially reported partnership with female F1 (green), and later with female F2 (blue). We also included data from another male M2 whose virus was genetically close to that of F1, although a partnership was not reported (see Supplementary Figure 2). **b** Viral genomes from all individuals were deep-sequenced, generating short viral sequence fragments (reads) that cover the genome. Reads were mapped against HIV-1 reference sequences, and are shown as horizontal coloured lines. Genomic windows covering the whole genome were defined; one is highlighted in black. For each window, overlapping reads were extracted, aligned, and a phylogeny was reconstructed using standard methods. **c** Each phylogeny contained many unique reads per individual that tended to cluster in the phylogeny. This enabled us to reconstruct parts of the tree (subgraphs) in which virus was inferred to be in each individual (colours label individuals; diamonds indicate unique read fragments, and the size of diamonds reflects copy number). In the phylogeny shown, virus from M1 (turquoise) was phylogenetically ancestral to that from F2 (blue), suggesting that transmission occurred from M1 to F2. Similarly, virus from F1 (green) was phylogenetically ancestral to that from M2 (purple), suggesting that transmission occurred from F1 to M2. For ease of illustration, only a part of the entire reconstructed deep-sequence phylogeny is shown. HIV-1 reference sequences and virus from another phylogenetically distant individual that is in-between the F1—M2 and M1—F2 pair are shown in black. **d** Viral deep-sequence phylogenies were reconstructed for each 250 bp genomic window to determine the statistical support of inferences on transmission and the direction of transmission. For each pair of individuals, the scan plots show the shortest patristic distance between subgraphs of both individuals (*y*-axis) and the topological relationship between subgraphs of both individuals (colours) across the genome. Deep-sequence data of sufficient quality were available for the HIV-1 *gag* gene, and the genomic position on the *x*-axis indicates the start of each 250 bp read alignment

Table 1 Characteristics of the study population

	Eligible	Participated	HIV-1 positive	Reporting no ART use	Deep-sequenced	Part of phylogenetically inferred transmission chain	Highly supported phylogenetic linkage and direction of transmission
Total	37,645	25,882	5142	3878	2652	1334	554
Women	18,946	13,791	3149	2251	1447	686	279
Age							
15–24	9203 (24%)	5839 (23%)	718 (14%)	610 (16%)	403 (15%)	210 (16%)	91 (16%)
25–34	6158 (16%)	4905 (19%)	1463 (28%)	1104 (28%)	717 (27%)	356 (27%)	141 (25%)
35+	3585 (10%)	3047 (12%)	968 (19%)	537 (14%)	327 (12%)	120 (9%)	47 (8%)
Men	18,699	12,091	1993	1627	1205	648	275
Age							
15–24	7907 (21%)	4845 (19%)	237 (5%)	215 (6%)	163 (6%)	92 (7%)	33 (6%)
25–34	6317 (17%)	4052 (16%)	929 (18%)	817 (21%)	618 (23%)	351 (26%)	145 (26%)
35+	4475 (12%)	3194 (12%)	827 (16%)	595 (15%)	424 (16%)	205 (15%)	97 (18%)

ART, antiretroviral therapy

shortest patristic distance between subgraphs of reads from two individuals (in short: subgraph distance) reflected genetic similarity of their viruses (*y*-axis). Figure 3a summarizes these deep-sequence phylogenetic patterns across known couples. We found, first, that the distribution of subgraph distances separating partners was bimodal (Fig. 3a, showing the median distance per pair across all their phylogenies after standardizing for differences in evolutionary rates across the genome). Most couples were either phylogenetically closely related or distantly related, with intermediate distances being very rare. This suggested that transmission likely occurred among phylogenetically closely related couples, and allowed us to define distance thresholds below which transmission was likely and above which transmission could be ruled out in this population (respectively <0.025 substitutions per site and >0.05 substitutions per site, see Fig. 3a). Additional analysis of whole-genome consensus sequences further supported these findings and thresholds (Supplementary Note 2 and ref. [29]). Second, we found that the large majority (166/178, 93.3%) of phylogenetically close couples also had ancestral subgraphs in most deep-sequence phylogenies, indicating in line with Leitner and Romero-Severson[23] that ancestral subgraph topologies are strongly over-represented among true transmission pairs.

Crucially, molecular epidemiologic analyses aim to infer unknown epidemiologic relationships from observed phylogenetic patterns in a population-based sample. This is a harder analytical problem compared to characterizing phylogenetic patterns among known epidemiologic relationships as in Fig. 3a, because only a tiny proportion of all pairs of individuals in a population-based sample are transmission pairs. We calculated the same phylogenetic patterns among all 3,515,226 possible pairs in our sample of 2562 individuals (see Methods), and summarized them in Fig. 3b as for the couples. With the exception of the 331 couples, sexual contacts were not known among any other of the ~3.5 m possible pairs. We found that ancestral subgraph topologies centred among pairs who were phylogenetically close: of 814 pairs with mostly ancestral subgraphs, 694 (85.3%) had phylogenetically close virus below our threshold for likely direct transmission (0.025 substitutions per site). However, 48 (5.9%) pairs had divergent virus above our threshold for ruling out direct transmission (0.05 substitutions per site). In addition, ancestry missed 118 (14.5%, 118/(694 + 118)) phylogenetically close pairs that had intermingled or sibling subgraphs in most of their deep-sequence phylogenies. Therefore, we used all types of subgraph topologies in combination with subgraph distance for inference of

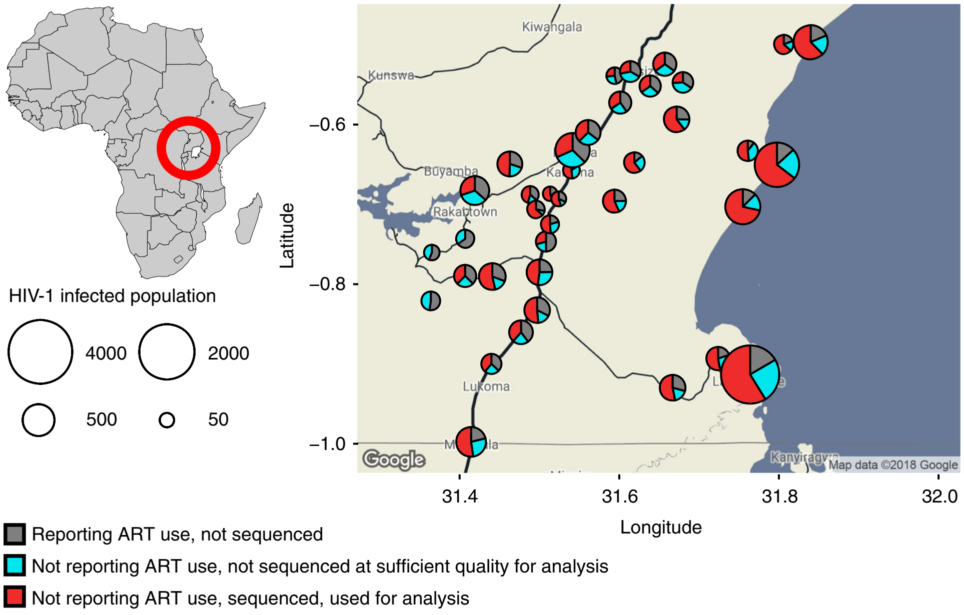

HIV-1 infected population
4000
2000
500
50
Kiwangala
Kunswa
Buyamba
Rakai town
Lukoma
Kanyirago
Google
Map data ©2018 Google
Latitude
−0.6
−0.8
−1.0
31.4
31.6
31.8
32.0
Longitude
Reporting ART use, not sequenced
Not reporting ART use, not sequenced at sufficient quality for analysis
Not reporting ART use, sequenced, used for analysis

Fig. 2 HIV-1 deep-sequencing in the Rakai Community Cohort, Uganda. Individuals aged 15–49 years were surveyed from August 2011 to January 2015 in 40 communities. In all, 5142 men and women were found positive (circles). Of those, 1264 self-reported using antiretrovirals (grey area of circles), and were not considered further as sequencing is challenging when virus is suppressed by treatment. Samples from 3878 individuals were deep-sequenced (see Methods). Of those, samples from 1226 (31.6%) individuals were not of sufficient quality for analysis (blue area of circles). Specifically, for phylogeny reconstruction, only paired-end merged reads of at least 250 base pairs (bp) in length were used, and subsequent deep-sequence inferences were performed on individuals whose reads covered the HIV-1 genome at a depth of at least 30 reads for 750 bp or more. Thus, samples from 2652 individuals (red area of circles) were used for molecular epidemiological analyses, corresponding to an estimated 45.1% of eligible and infected individuals with unsuppressed virus in RCCS communities

transmission networks from deep-sequence data. It is possible to approximate the likelihood of deep-sequence phylogenetic patterns under mathematical models of within-host viral evolution and transmission[30]. However, such models do not fully reproduce empirical observations such as preferential transmission of founder viruses[31], and can be computationally prohibitive at large scales. For these reasons we adopted a statistical approach that is based on counting phylogenetic patterns across the genome, and calculating the proportion of deep-sequence phylogenies in support of no linkage ($\hat{\mu}_{ij}$), linkage ($\hat{\lambda}_{ij}$), and direction of transmission given linkage ($\hat{\delta}_{ij}$); see Fig. 4 and Methods. Starting with subgraph distance, direct transmission could be ruled out for 3,513,800/3,515,226 (99.96%) pairs, leaving only 1426 potential transmission pairs. Next, we also considered information in subgraph topologies. This left 1191 potential transmission pairs that formed 446 transmission networks in the population-based sample of 2562 individuals, i.e. groups of individuals that had predominantly phylogenetically close and topologically adjacent (ancestral, intermingled or sibling) subgraphs.

Unlike typical phylogenetic clusters[11,12,32,33], these transmission networks contained information on the direction of transmission (Fig. 5). Two hundred and sixty-one networks comprised just two individuals, while 36 had more than five individuals. As expected given the uncertainty in our inferences, larger networks included cycles of possible transmission flows and recipients with more than one probable source case, implying that multiple transmission chains were consistent with our phylogenetic data. We next identified the most likely transmission chains using graph theory (see Methods). This retained 888 phylogenetic linkages in 446 most likely transmission chains, of which 351 linkages had low statistical support ($\hat{\lambda}_{ij} \leq 0.6$, see Fig. 4 and Methods for choice of threshold) and 537 linkages had high statistical support ($\hat{\lambda}_{ij} > 0.6$).

Viral deep-sequence data cannot prove HIV-1 transmission. We hypothesized that many of the 537 highly supported phylogenetic linkages were false discoveries in that transmission did not occur directly between the paired individuals. Our population-based sample did not capture all members of ongoing transmission chains, and so transmission likely occurred via unsampled intermediates in some cases. 80/537 (14.9%) of highly supported phylogenetic linkages were between two women even though HIV-1 is predominantly sexually transmitted in Africa, and extremely rarely transmitted sexually between women[34]. Considering that there were almost twice as many possible male−female combinations than female−female combinations, we calculate in Supplementary Note 3 that up to 35.4% of phylogenetically close male−female pairs of the population-based sample may not represent direct transmission events. Figure 4b illustrates this fundamental problem further: subgraph distances and topologies were not sufficient to clearly separate pairs of individuals from the population sample into two groups of closely related or distantly related pairs.

In prior work, Romero-Severson et al.[22] proposed that direct transmission can be established with near certainty when viral sequences from two individuals are heavily intermingled in deep-sequence phylogenies. This prediction, while based on theoretical evolutionary principles and simulation, implies that deep-sequence phylogenies could be used in criminal cases of HIV-1 transmissions, and thus has important public health and human rights implications.

We revisited this hypothesis in our data, and found 34 phylogenetically close pairs with intermingled subgraphs across

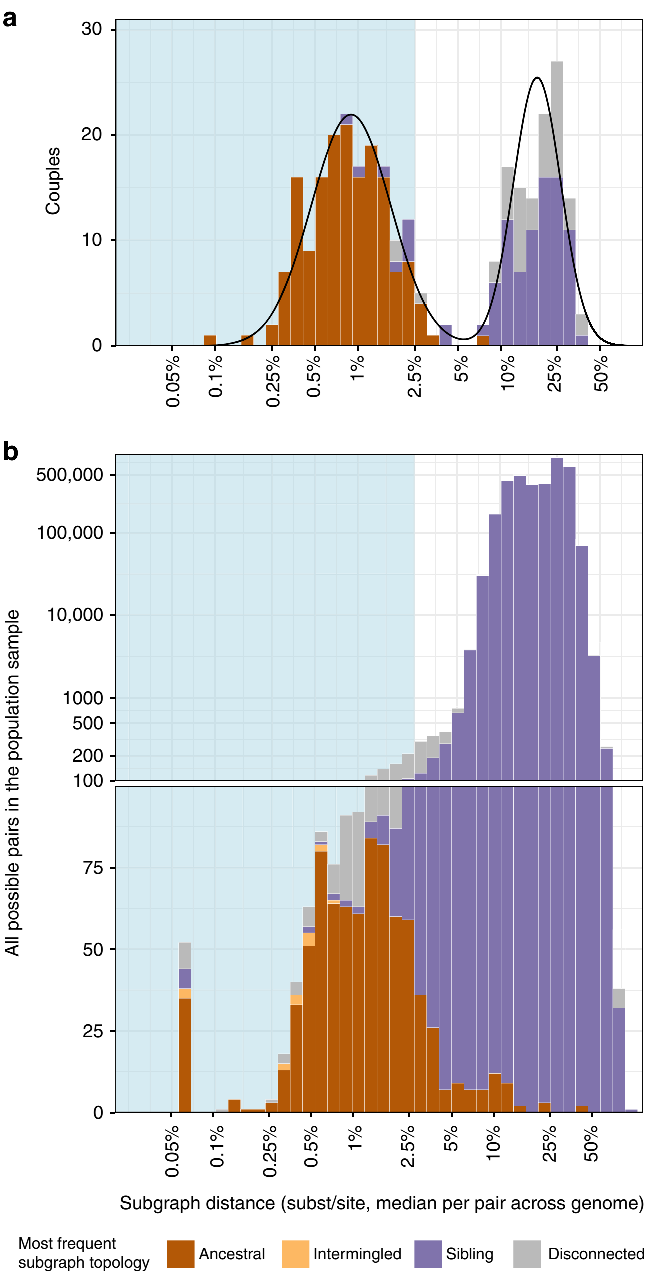


Fig. 3 Deep-sequence phylogenetic data in the population-based sample. To highlight the characteristics of deep-sequence phylogenetic data in a population-based sample, we compared phylogenetic patterns among couples in whom both partners were positive to the patterns in the larger population-based sample. **a** Analysis of 331 couples. For each couple, their subgraph distances and subgraph topologies were calculated in each deep-sequence phylogeny across the genome as shown in Fig. 1d. Subgraph distances were standardized to the average evolutionary rate of the HIV-1 *gag* and *polymerase* genes (see Methods). Information from all deep-sequence phylogenies was summarized by median distance and the most frequent subgraph topology (colours). The distribution of median distances had a clear bimodal shape, separating couples into two groups that were either phylogenetically closely or distantly related. The distribution of median distances was well described by a two-component lognormal mixture model (black lines). 95% of couples in the first component had distances below 0.025 substitutions per site (light blue area) and 99% of couples in the first component had distances below 0.05 substitutions per site. We used these thresholds to classify couples into phylogenetically close and distant. 93.3% of phylogenetically close couples also had mostly ancestral subgraphs. **b** Analysis of 3,515,226 possible pairs in the population-based sample. For visualization purposes, smaller numbers are displayed on natural scale and larger numbers on log scale. The distribution of median distances was not bimodal, and subgraph distances did not clearly separate pairs of individuals into closely or distantly related pairs. 48/814 (5.9%) pairs with mostly ancestral subgraphs were phylogenetically distant as defined by the couples' analysis. One hundred and eighteen phylogenetically close pairs had mostly intermingled or sibling subgraphs and were missed by subgraph ancestry, indicating that all types of subgraph topologies in combination with subgraph distance should be used for inference of population-level transmission networks

the majority of the genome. In two instances, the phylogenetically linked individuals were female (Fig. 6, corresponding deep-sequence phylogenies are reported in Supplementary Data 1), suggesting they were likely infected by a common unobserved male partner. Based on this, the phylogenetic linkages in transmission networks that we inferred from our deep-sequence data may indicate—but cannot prove—direct transmission. The difference between the theoretical expectations of Romero-Severson et al.[22] and our observations may be explained by limited phylogenetic resolution in our reads, or may reflect greater complexity in HIV-1 evolutionary dynamics[35].

These findings put into context that 81 (15.1%) of the 537 highly supported phylogenetic linkages were between two men. Given that the relative proportion of same-sex linkages were equivalent between men and women, our phylogenetic transmission networks provide no evidence of extensive sub-epidemics amongst men who have sex with men in rural Rakai although we cannot rule out the possibility that these may exist due to potential undersampling of widely stigmatized key populations[36].

The direction of transmission can be frequently inferred. We further analysed the remaining 376 highly supported male—female linkages to infer the direction of transmission (i.e. who might have infected whom, potentially via unsampled intermediates). Amongst the population-based sample, we inferred the phylogenetically likely source for 293/376 (77.9%) of linked male—female pairs (Fig. 5, $\hat{\delta}_{ij} > 0.6$, see Methods for choice of thresholds). In comparison, 176/376 (46.8%) of highly supported male—female linkages were between couples, and the phylogenetically likely source could be inferred in 133/176 (75.6%) couples. Inferences of these source—recipient pairs did not depend strongly on our cut-off choices (Supplementary Table 4).

Inferring the direction of transmission has a small error. We cross-validated our findings on the direction of transmission using HIV-1 testing history and clinical data that provided independent evidence that one direction of transmission was much more likely than the other. In 36 pairs (18 couples and 18 pairs between whom sexual contact was not known), one individual tested HIV-1 negative after the other had already tested positive, and the negative individual subsequently seroconverted. The phylogenetically inferred source ($\hat{\lambda}_{ij} > 0.6$ and $\hat{\delta}_{ij} > 0.6$) was consistent with clinical evidence in 27/31 pairs, inconsistent in 4/31 pairs, and could not be inferred reliably in 5/31 pairs (Table 2; corresponding deep-sequence phylogenies are reported in Supplementary Data 2). The false discovery rate for estimating the direction of transmission amongst pairs with epidemiologically known direction of transmission was therefore 12.9% with 95% confidence interval [5.1–28.9%].

In 35 pairs, one individual had a CD4 cell count above 800 cells per mm^3 blood, indicative of being close to time of infection,

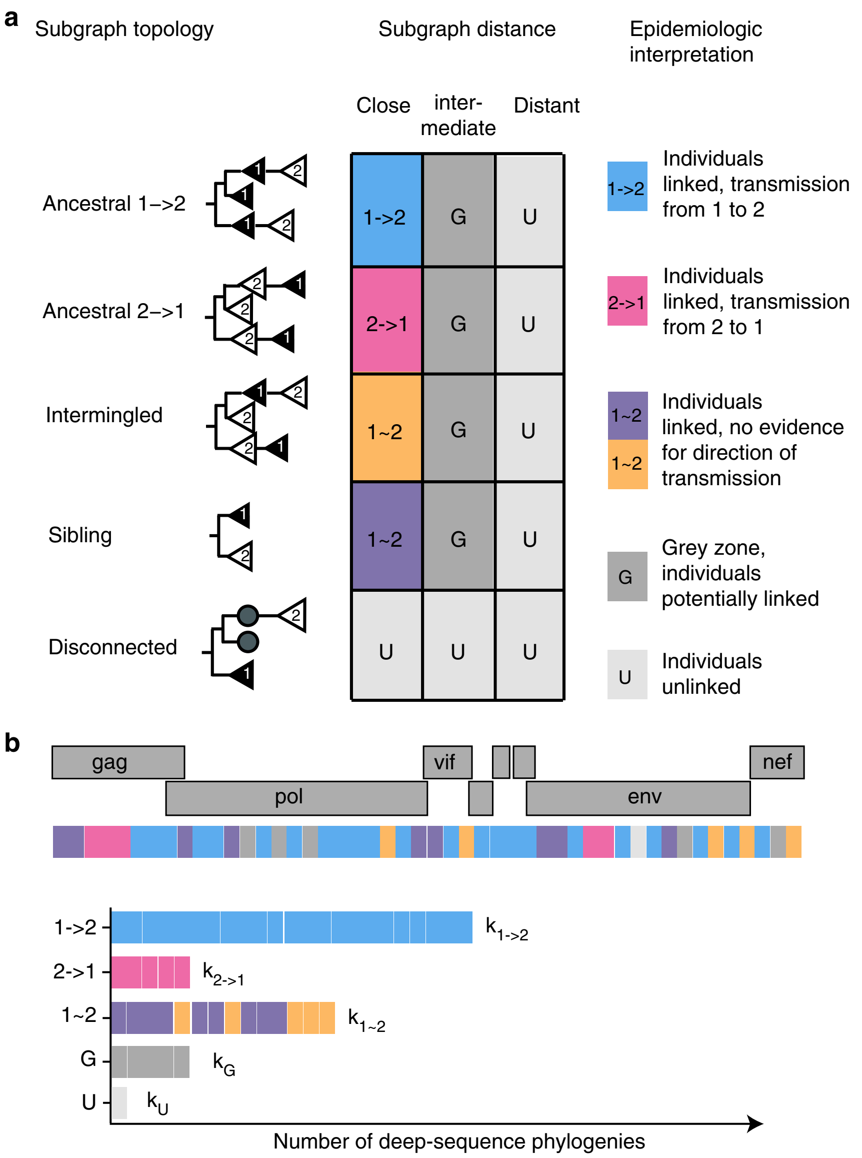

a
Subgraph topology
Subgraph distance
Epidemiologic interpretation
Close
inter-mediate
Distant
Ancestral 1–>2
Ancestral 2–>1
Intermingled
Sibling
Disconnected
1->2
2->1
1~2
G
U
Individuals linked, transmission from 1 to 2
Individuals linked, transmission from 2 to 1
Individuals linked, no evidence for direction of transmission
Grey zone, individuals potentially linked
Individuals unlinked
b
gag
pol
vif
env
nef
k1->2
k2->1
k1~2
kG
kU
Number of deep-sequence phylogenies

Fig. 4 Epidemiological interpretation of deep-sequence phylogenetic data. **a** The 5 × 3 contingency table describes how deep-sequence phylogenetic patterns between two individuals were epidemiologically interpreted. Viral phylogenetic patterns between two individuals were summarized in terms of subgraph distance and subgraph topologies. There are five possible subgraph topologies between two individuals. All subgraphs of person 1 can be disconnected from the subgraphs of person 2 by another individual. If subgraphs of two individuals are adjacent, i.e. not disconnected by another individual, they can be consistently ancestral to each other in the same direction, intermingled in that some subgraphs are ancestral in one direction and others in the opposite direction, or siblings. The subgraph distance between viral subgraphs was stratified into 'close' (<0.025 substitutions per site), 'intermediate' (0.025–0.05 substitutions per site), and 'distant' (>0.05 substitutions per site) based on the couples' analysis shown in Fig. 3a. Epidemiologic interpretations are indicated in colours. When only one sequence per individual is available, subgraphs of individuals correspond to the tips in a phylogeny, are either disconnected or siblings, and thus the direction of transmission is not inferable. **b** To determine the statistical support in inferences on transmission and the direction of transmission, analyses were repeated across the genome and the observed relationship types 1 → 2, 2 → 1, 1 ~ 2, G, U were counted (respectively denoted by $k_{1\to 2}$, $k_{2\to 1}$, $k_{1\sim 2}$, k_G, k_U). To avoid overconfidence, an adjustment was made to account for the fact that overlapping windows are not statistically independent (see Supplementary Note 1). Evidence for no transmission between individuals 1 and 2 was estimated by $\hat{\mu}_{12} = k_U/n$; evidence for transmission between 1 and 2 was estimated by $\hat{\lambda}_{12} = (k_{1\to 2} + k_{1\sim 2} + k_{2\to 1})/n$; and evidence for transmission from 1 to 2 given that transmission occurred between 1 and 2 was estimated by $\hat{\delta}_{12} = k_{1\to 2}/(k_{1\to 2} + k_{2\to 1})$; see Methods for further details

while their partner was already immuno-compromised with a CD4 cell count below 400 cells per mm^3 blood. The phylogenetically inferred source was consistent with clinical evidence in 19/35 pairs, inconsistent in 5/35 pairs, and could not be inferred reliably in 11/35 pairs. In two of the five inconsistent cases, CD4 data were only weakly indicative of the direction of transmission, and it is possible that we overestimated error rates for these pairs with CD4 data to 20.8% [9.2–40.5%] (Supplementary Note 4).

Amongst all pairs, the false discovery rate was 16.3% [8.8–28.3%]. Error rates varied slightly depending on the exact configuration of parameters in the phyloscanner analyses, though not substantially (Supplementary Tables 5–6). Similar error rates were observed in phylogenetic analysis of 454 deep-sequence data over a 320 bp region of the *env* gene among 33 couples with known direction of transmission and confirmed linked infection in the HPTN 052 trial[37]. Our findings are based on

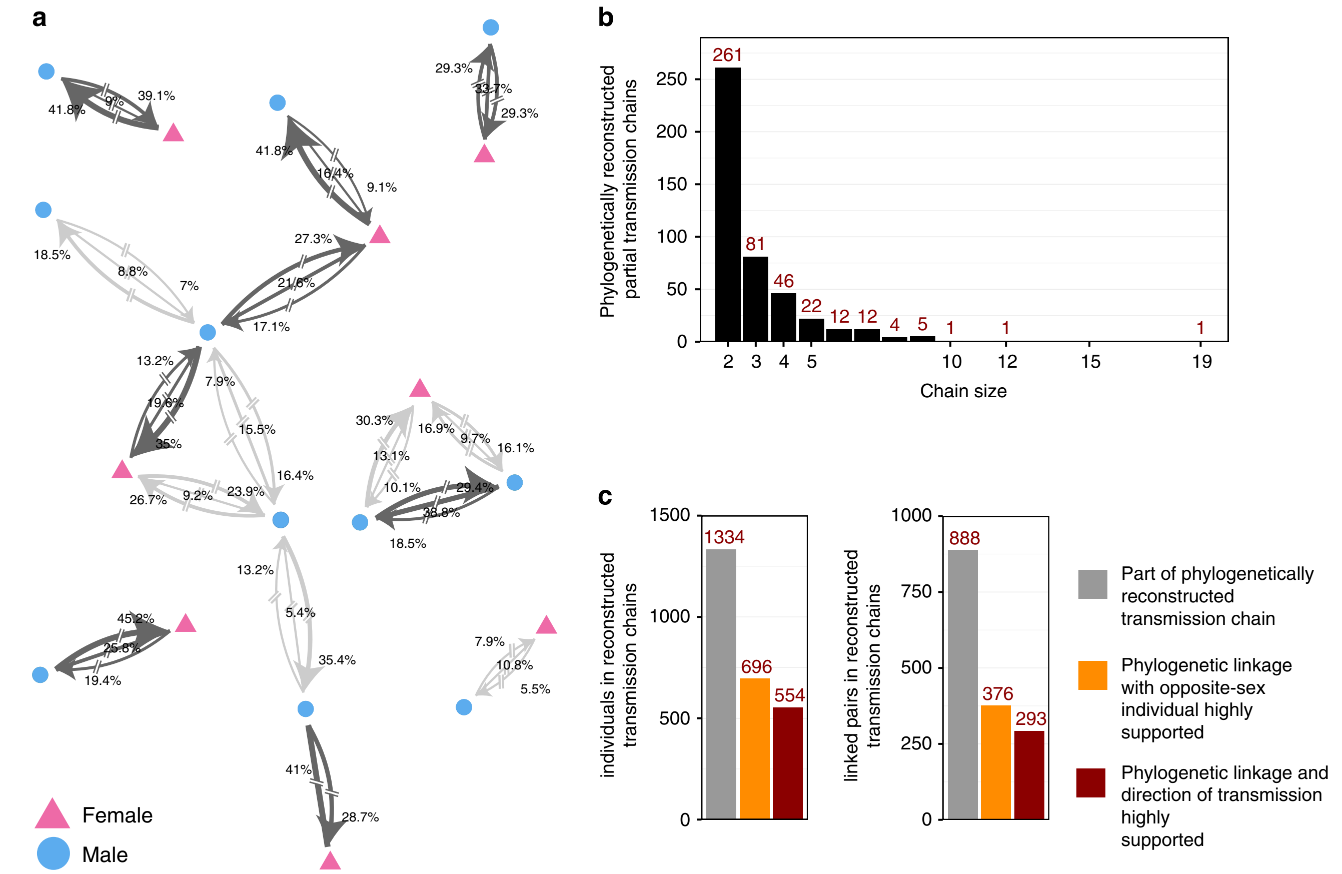


Fig. 5 Phylogenetically reconstructed transmission networks. Four hundred and forty-six transmission networks comprising 1334 individuals and 888 linkages could be reconstructed from the population-based sample. **a** Illustrative set of six transmission networks with nodes indicating gender. In comparison to phylogenetic clustering analyses, deep-sequence phylogenetic analysis provided evidence about the direction of transmission. Edges connecting two individuals were labelled with the statistical support for transmission in the indicated direction (for directed edges), or for transmission with no evidence for direction (for undirected edges), calculated as the proportion of deep-sequence phylogenies supporting each case (see Fig. 4). The sum of the three weights quantified the phylogenetic support for direct transmission on a scale between 0 and 1 ($\hat{\lambda}_{ij}$, see Fig. 4). Pairs of individuals with high support for direct transmission were highlighted in dark grey ($\hat{\lambda}_{ij} > 0.6$). All edges were broken to indicate the possibility of unsampled intermediates. **b** Sizes of reconstructed transmission chains. The majority of transmission chains (261/446, 58.5%) were pairs, though 36 chains had more than five individuals. **c** Numbers of individuals (left) and linked pairs (right) in reconstructed transmission chains. Many linked pairs were weakly supported or between individuals of the same sex, which indicated the presence of unobserved intermediates or common sources. In all, 376 male–female pairs had high support ($\hat{\lambda}_{ij} > 0.6$) (orange bars), and of those, the direction of transmission could be inferred with high support ($\hat{\delta}_{ij} > 0.6$) in 293/376 (77.9%) pairs (burgundy bars)

deep-sequencing of a population-based sample, and thus extend previous results to population-level inferences among individuals between whom sexual contact is not necessarily known a priori.

Discussion

A central application of pathogen sequencing is to identify how infectious diseases continue to spread in human populations, and how new infections can be averted most effectively[38–41]. Most molecular epidemiologic studies are based on analysis of Sanger sequences, and typically identify clusters of genetically related infections in an effort to characterize ongoing transmission sources[11,32,33,42]. These approaches fail to distinguish sources from recipients of transmission within such clusters, making epidemiological inferences relevant to public health intervention challenging[7]. In contrast, deep-sequence phylogenetic analyses are based on thousands of reads per individual, and thereby provide more information into the epidemiologic relationship of individuals beyond distance measures, through the topological ordering between subgraphs of viral reads from individuals. Prior work assessed the potential of deep-sequence phylogenetic analyses on simulations and on known transmission pairs for whom at least five viral sequences were available per individual[22,23,43]. Here, we demonstrate that large population-based samples of standard deep-sequence output can be used to infer directed transmission networks of generalized HIV-1 epidemics in sub-Saharan Africa with phyloscanner[21]. Combining the patristic distance between viral subgraphs and their topological ordering in deep-sequence phylogenies, our analysis uncovered 446 partially sampled HIV-1 transmission networks in Rakai comprising 1334 individuals.

We were not able to rule out the possibility that sources were indirectly linked to recipients through unobserved individuals (i.e. intermediate partners) with deep-sequence phylogenetic analysis. One third (161/537) of phylogenetically highly supported linkages were between individuals of the same gender, in line with incomplete sequence coverage. We also found two pairs with phylogenetic patterns previously considered strong enough to virtually exclude the possibility of common sources or recipients, but in whom both individuals were female. These findings have important implications for criminal prosecution of people

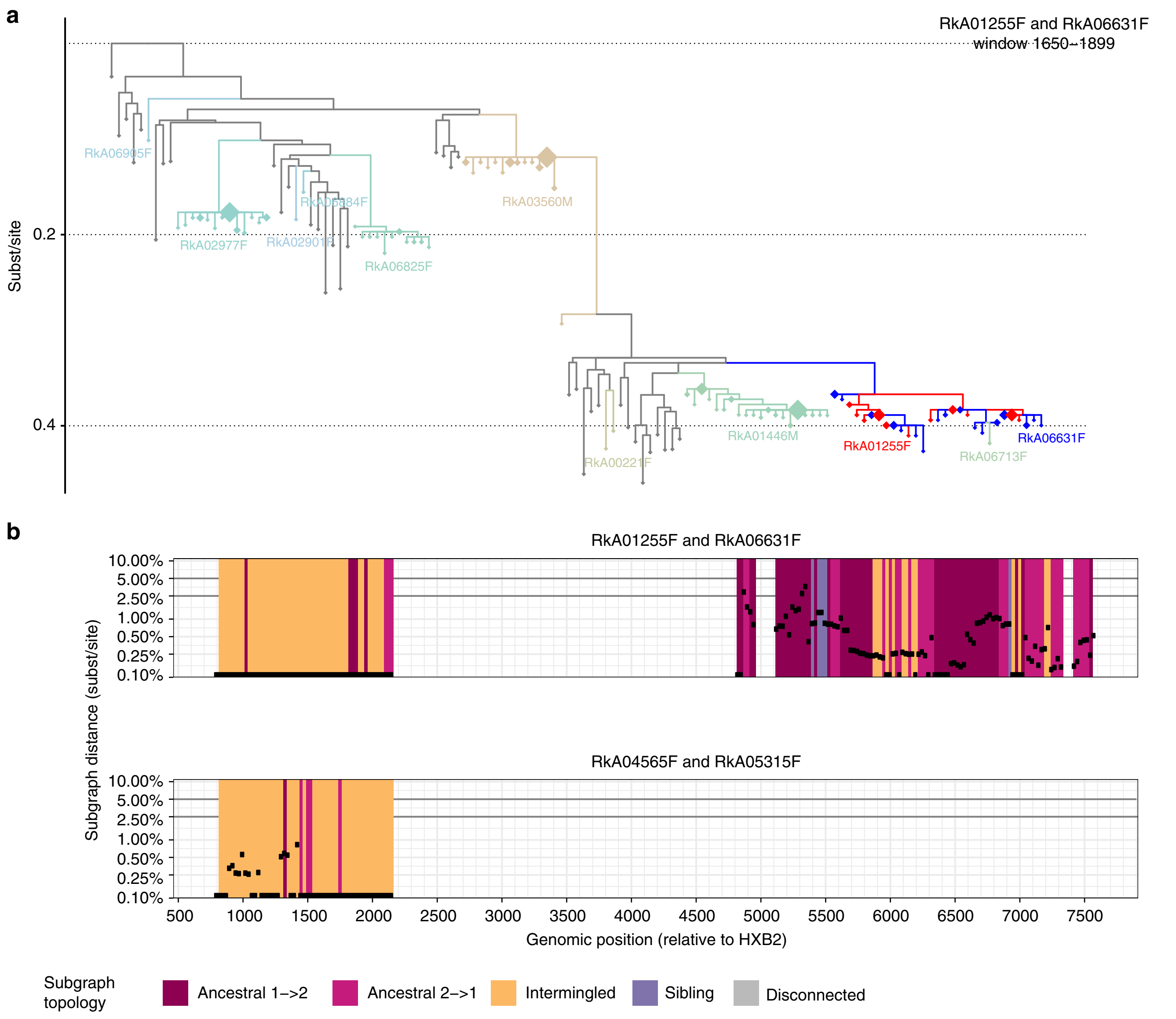


Fig. 6 Direct transmission cannot be established when HIV-1 sequences from two individuals are intermingled in deep-sequence phylogenies. It was previously proposed that certain patterns in deep-sequence phylogenies—intermingled subgraphs of two individuals as shown in panel (**a**) in red and blue—rule out the presence of unobserved common sources and/or intermediates, and could thus prove that direct transmission occurred between two individuals. We revisited this prediction on our data, and found two female—female pairs with mostly intermingled and near identical subgraphs across the genome. These data indicate that such deep-sequence phylogenetic relationships cannot exclude the possibility of unsampled common sources or intermediates. **a** One deep-sequence phylogeny is shown for one female—female pair to illustrate their typical phylogenetic relationships. Reads from the two female—female pairs are shown in red and blue, are intermingled, and often nearly identical. The phylogenetically most closely related individuals that acted as controls are highlighted in colours, and reference sequences are shown in grey. One additional female (RkA06713F) was phylogenetically close to both females, though too poorly sampled to resolve phylogenetic relationship. The other individuals were phylogenetically distant or disconnected from the two females by HIV-1 reference sequences, with no relationship to the two females inferred. Deep-sequence phylogenies of all other windows are shown in Supplementary Data 1. **b** Phyloscan plot of subgraph distances (*y*-axis) and subgraph topologies (colour) across the genome for both female—female pairs. In the majority of deep-sequence phylogenies, both pairs had intermingled subgraphs that were also near identical

living with HIV in at least 72 countries with laws penalizing HIV transmission[14,44]: even with deep-sequencing, transmission of HIV-1 cannot be proven between two individuals. Thus, communicating the limitations of deep-sequencing data is essential to prevent its misuse in criminal prosecutions. For example, we opted to visually interrupt linkages in phylogenetic transmission networks (Fig. 5), in order to highlight the possibility of unsampled cases along inferred source—recipient relationships.

We found that when many reads from different individuals are analysed together, they tend to form subgraphs with consistent ordering in deep-sequence phylogenies from across the genome. This observation enabled us to infer the source of transmission in 77.9% of 376 phylogenetically linked male—female pairs. The accuracy of our viral phylogenetic inferences regarding directionality was validated on 71 male—female pairs with clinical data that suggested transmission in one direction, with an overall false discovery rate of 16.6% [9.1–28.7%], and was thus not substantially different in a population-based sample compared to analysis of couples with known direction of transmission[37]. At this error rate, phyloscanner and similar approaches[21,37,43] allow inferences into population-level transmission networks and the epidemiologic sources of ongoing viral spread from sequence data alone.

Table 2 Error rates in inferring the direction of HIV-1 transmission

Epidemiological evidence for direction of transmission	Phylogenetically linked pairs who reported sexual contact (couples)	Other phylogenetically linked pairs	Total
History of HIV-1 test results[a]			
Total	18	18	36
Direction consistent with clinical evidence	14	13	27
Direction ambiguous	2	3	5
Direction inconsistent with clinical evidence	2	2	4
False discovery rate	12.5% [3.5–36.0%]	13.3% [3.7–37.8%]	12.9% [5.1–28.9%]
Discrepancy in CD4 count[b]			
Total	17s	18	35
Direction consistent with clinical evidence	11	8	19
Direction ambiguous	6	5	11
Direction inconsistent with clinical evidence	0	5	5
False discovery rate	0%	38.5% [17.7–64.5%]	20.8% [9.2–40.5%]
Combined false discovery rate	7.4% [2.1–23.4%]	25% [12.7–43.4%]	16.3% [8.8–28.3%]

[a] Partner 1 tested HIV-negative, while partner 2 tested HIV-positive at or before the same time, and partner 1 was subsequently found HIV-positive
[b] Partner 1 had first CD4 measurement >800 cells per mm^3, while partner 2 had a CD4 measurement <400 cells per mm^3 within 2 years of the first CD4 measurement of partner 1

Our study has several weaknesses. First, sequence sampling of the infected population in RCCS communities remained incomplete. Phylogenetic inferences are expected to improve with higher sampling fraction[45], though in practice, complete sequence sampling is hard to achieve. This study enrolled participants before immediate provision of ART was recommended in national guidelines, so that a relatively large proportion of infected individuals did not report ART use at first study visit, and could be sequenced. To perform similar phylogenetic analyses of ongoing viral spread in sub-Saharan Africa in the future, it is thus important to collect and store samples prior to ART initiation, and to investigate alternative sequencing protocols[46]. Second, relatively modest deep-sequencing quality compromised the length of deep-sequence reads[28]. Analyses were based on relatively short read alignments of 250 bp that primarily covered the *gag* gene, rather than the whole genome (Supplementary Figure 1). It is thus plausible that deep-sequence phylogenetic analyses may be more accurate than reported in this study as deep-sequence output with longer reads and greater coverage is becoming available[47]. Third, we found that inferring the direction of transmission became more challenging as the virus was increasingly closely related within individuals. We thus predict that the direction of transmission may be less frequently inferable in situations when the virus spreads more rapidly between persons, as in high-risk sexual networks among men having sex with men[9,15], or among injecting drug users[48]. For the same reason, sources of infections may be less accurately and/or less frequently inferable for pathogens that generate within-host viral diversity at a slower pace than HIV-1 [39,49,50].

Whole-genome deep-sequencing is now the tool of choice in clinical practice and epidemiologic investigation for a broad range of bacterial infectious disease pathogens, and increasingly used for viral pathogens, and especially HIV-1 [8,38,39,49,50]. Here we establish that HIV-1 phylogenetic analyses can be scaled to large population-based samples of deep-sequence data, and that the direction of transmission can be frequently inferred in reconstructed HIV-1 transmission networks. At present, more than 15,000 individuals have been deep-sequenced and linked to demographic records across sub-Saharan Africa in order to understand who is at the core and driving new infections where the burden of HIV-1 is highest, how the epidemic regenerates from older to younger generations, and how spread can be most effectively interrupted in generalized epidemics[7,8]. The phyloscanner method is applicable to these data, and we hypothesize that this innovation will help identify the key drivers of HIV-1 transmission in regions that are hardest hit by the virus, and in turn facilitate tailoring of interventions to achieve epidemic control.

Methods

Sample selection. Data for this study come from the Rakai Community Cohort Study (RCCS), a population-based study of HIV-1 incidence in Rakai, District Uganda. Procedures for the RCCS have been described in detail elsewhere[2]. Briefly, the RCCS conducts a census in all communities to identify eligible individuals 2 weeks before the survey. Eligible individuals include those able to give consent and between the ages of 15 and 49 years. Eligible individuals who provide written informed consent are administered a survey on their demographs, sexual behaviours and health-care seeking practices. Individuals are also asked to name their cohabitating sexual partners in order to identify couples, and to provide a serum sample for HIV-1 testing and future laboratory studies, including HIV-1 viral sequencing. Data for this particular study were collected between 2011 and 2015 from 40 agrarian, trading and fishing communities.

Ethics. The study was independently reviewed and approved by the Ugandan Virus Research Institute, Scientific Research and Ethics Committee, Protocol GC/127/13/01/16; the Ugandan National Council of Science and Technology; and the Western Institutional Review Board, Protocol 200313317. All study participants provided written informed consent at baseline and follow-up visits using institutional review board-approved forms.

Sampling fraction. To estimate the number of infected participants with unsuppressed virus, we first calculated the expected number of infected participants who did not use antiretrovirals at time of survey, and had thus unsuppressed virus. Participant reported ART use was previously validated as a proxy of actual ART use with a specificity of 99%[26], giving 3878/0.99 individuals. To this, we added the expected number of participants who reported ART use but did not have suppressed virus. Ten per cent of participants reporting ART use had plasma viral loads above 1000 copies/ml plasma blood[2], giving 1264 × 0.9 individuals, and 4043 in total. The sampling fraction was therefore estimated at 2652/4043 (65.6%) among infected participants with unsuppressed virus.

HIV-1 deep-sequencing. Serum samples from HIV-1 seropositive persons who did not self-report ART use over the analysis period were shipped to University College London Hospital, London, United Kingdom for viral RNA extraction. RNA extraction was automated on QIAsymphony SP workstations with the QIAsymphony DSP Virus/Pathogen Kit (Cat. No. 937036, 937055; Qiagen, Hilden, Germany), followed by one-step reverse transcription polymerase chain reaction (RT-PCR)[27]. Deep-sequencing was performed on Illumina MiSeq and HiSeq instruments in the DNA pipelines core facility at the Wellcome Trust Sanger Institute, Hinxton, United Kingdom.

Assembly of HIV-1 reads. Deep-sequencing reads were assembled with the shiver sequence assembly software[51]. Where no contigs could be generated with IVA[52], contigs were generated with SPAdes and metaSPAdes v3.10 [53,54], after excluding reads classified as Homo sapiens by Kraken v0.10.5-beta[55]. Contigs with at least 300 bp matching known HIV-1 diversity were used for shiver analysis.

Read selection. Phyloscanner version 1.1.2 [21] was used to merge paired-end reads, and only merged reads of at least 250 bp in length were retained for phylogeny reconstruction. Subsequent deep-sequence inferences were performed on individuals whose reads covered the HIV-1 genome at a depth of at least 30 reads for 750 bp or more. Individuals who did not have sequencing output meeting these criteria were excluded.

Deep-sequence phylogenetic analysis. It proved computationally intractable to reconstruct viral trees from all deep-sequence reads of all individuals simultaneously. To address this challenge, samples were divided into batches of 50–75 individuals, and phyloscanner was run on all possible pairs of batches to assess deep-sequence phylogenetic relationships in all pairs of individuals in the population-based sample. The phyloscanner command line specification for this first analysis stage is given in Supplementary Tables 1 and 2. Shell scripts were used to handle calculations in parallel, and are available upon request. From stage 1 output, we identified potentially phylogenetically close pairs and, from those, networks of pairs that were connected through at least one common, phylogenetically close individual. Networks were extended to include spouses of partners in networks, couples in no network, and the ten most closely related individuals from stage 1 as controls. For computational considerations, reads of individuals that differed at one nucleotide position were merged. In a second analysis stage, phyloscanner was used to confirm potential transmission pairs by considering also the topological configuration of subgraphs in deep-sequence phylogenies, and to resolve the ordering of transmission events within transmission networks. The phyloscanner command line specification for stage 2 is given in Supplementary Table 3. In this stage, reads of individuals that differed at one nucleotide position were not merged.

Phylogenetic relationships of virus from two individuals. The basis of viral phylogenetic analysis with phyloscanner are subgraphs, sets of tips and internal nodes of a phylogeny that are attributed to one individual with a parsimony-based algorithm[21]. A single individual can have multiple subgraphs in one tree. The following statistics were calculated to characterize the phylogenetic relationship between two individuals i and j in one phylogeny:

- Subgraph distance between i and j (Δ_{ij}): The distance between any two subgraphs u, v is the shortest patristic distance between any nodes or tips of u and v and Δ_{ij} is the minimum patristic distance between subgraphs u from i and v from j. Deep-sequence phylogenies from different parts of the genome had markedly different branch lengths, reflecting evolutionary rate variation across the genome. Prior to calculating subgraph distances, we standardized phylogenies by multiplying branch lengths with the ratio of expected branch lengths in the genomic window from which the tree was reconstructed, divided by the expected branch lengths in the *gag* and *polymerase* genes (Supplementary Table 2).
- Adjacency of i and j (A_{ij}): True if the shortest path between at least one subgraph u from i and v from j is not attributed to any sampled individual other than i and j, and false otherwise.
- Paths from i to (P_{ij}): number of subgraphs from j which have as ancestor a subgraph from i.

Analyses were then based on the following phylogenetic relationship types between two individuals i and j in a viral tree:

- Phylogenetically unlinked (U_{ij}): $A_{ij} = 0$ or $\Delta_{ij} > 0.05$ substitutions per site.
- Phylogenetic linkage grey zone (G_{ij}): $A_{ij} = 1$ and $\Delta_{ij} \in [0.025{-}0.05$ substitutions per site].
- Phylogenetically linked and i source ($i \to j$): $A_{ij} = 1$ and $P_{ij} \geq 1$ and $P_{ji} = 0$ and $\Delta_{ij} < 0.025$ substitutions per site.
- Phylogenetically linked and j source ($j \to i$): $A_{ij} = 1$ and $P_{ji} \geq 1$ and $P_{ij} = 0$ and $\Delta_{ij} < 0.025$ substitutions per site.
- Phylogenetically linked with no evidence for direction of transmission ($i \sim j$): $A_{ij} = 1$ and $P_{ji} \geq 1$ and $P_{ij} \geq 1$ and $\Delta_{ij} < 0.025$ substitutions per site (intermingled), or $A_{ij} = 1$ and $P_{ji} = 0$ and $P_{ij} = 0$ and $\Delta_{ij} < 0.025$ substitutions per site (sibling).

Evidence for transmission and direction of transmission. To capture uncertainty in inferences, relationship types between reads from two individuals were evaluated on a large number of deep-sequence phylogenies that corresponded to sliding and overlapping read alignments (as shown in Fig. 1d). For each pair of individuals, the number of deep-sequence phylogenies in which i and j had one of the above five relationship types were counted (as shown in Fig. 4). The raw counts were adjusted for overlap in read alignments from which the deep-sequence phylogenies were constructed as described in Supplementary Note 1, and are denoted by k_{U} (unlinked), k_{G} (grey zone), $k_{i\,\to\,j}$ (i source), $k_{j\,\to\,i}$ (j source), $k_{i\,\sim\,j}$ (no evidence for direction). After adjusting for overlap, the counts were interpreted as phylogenetic independent observations, leading to Binomial probability models for each count. Evidence for direct transmission (λ_{ij}) was based on the count $k_{\mathrm{L}} = k_{i\,\to\,j} + k_{j\,\to\,i} + k_{i\,\sim\,j} \geq 0$, and binomial model (likelihood)

$$p\left(k_{\mathrm{L}}, n|\lambda_{ij}\right) = \frac{\Gamma(n+1)}{\Gamma(k_{\mathrm{L}}+1)\Gamma(n-k_{\mathrm{L}}+1)}\lambda_{ij}^{k_{\mathrm{L}}}(1-\lambda_{ij})^{n-k_{\mathrm{L}}}, \tag{1}$$

where $n = k_{i\,\to\,j} + k_{j\,\to\,i} + k_{i\,\sim\,j} + k_{\mathrm{G}} + k_{\mathrm{U}} > 0$ and Γ is the Gamma function, with maximum likelihood estimate $\hat{\lambda}_{ij} = k_{\mathrm{L}}/n$. Evidence for ruling out direct transmission (μ_{ij}) was based on k_{U} and total n as above. Evidence for the direction of transmission given linkage (δ_{ij}) was based on $k_{i\,\to\,j}$ and total $k_{i\,\to\,j} + k_{j\,\to\,i}$. Posterior density estimates of λ_{ij}, μ_{ij} and δ_{ij} are available analytically when a Beta prior density on these parameter is chosen. We here chose a flat Beta prior density with scale and shape parameters set to 1, so that e.g. the posterior density for direct transmission is

$$p\left(\lambda_{ij}|k_{\mathrm{L}}, n\right) = \frac{\Gamma(n+1)}{\Gamma(k_{\mathrm{L}}+1)\Gamma(n-k_{\mathrm{L}}+1)}\lambda_{ij}^{k_{\mathrm{L}}}(1-\lambda_{ij})^{n-k_{\mathrm{L}}}. \tag{2}$$

The confidence intervals shown in Supplementary Notes 2 and 4 are 95% highest density intervals of Eq. (2). In principle, the parameters of the Beta prior could be chosen to reflect additional data such as seroconversion histories; however, care should be taken to specify informative priors based on variables such as age differences or age-specific disease prevalence[20], in order to avoid circular inferences on who may have infected whom.

Most likely transmission chains. Pairs of individuals between whom transmission was not excluded (when $\hat{\mu}_{ij}$>0.6) defined a set of connected graphs, which we call (partially observed) transmission networks. For each network, we defined its adjacency matrix with entries $\hat{\tau}_{ij} = k_{i\to j} + k_{i\sim j}/2$ for $i \neq j$ and $\hat{\tau}_{ij} = 0$. Every spanning tree c of a network defines a possible transmission chain, and was associated with a transmission flow score over its directed edges, $\hat{\tau}_c = \prod_{ij\in c} \hat{\tau}_{ij}$. The most likely transmission chain, defined by $\hat{c}^{\mathrm{ML}} = \mathrm{argmax}_c\, \hat{\tau}_c$, was calculated with Edmonds's algorithm as implemented in the RBGL R package, version 1.55.1 [56].

Classification of linked pairs and sources. Pairs in most likely transmission chains were classified as (epidemiologically) linked when $\hat{\lambda}_{ij} = k_{\mathrm{L}}/n > c$ where n as above and $c = 0.6$, and otherwise as potentially linked. The threshold c was determined as follows. Under model (1), $k_{\mathrm{L}} \sim$ Binomial (n, λ_{ij}), where λ_{ij} indicates the strength of phylogenetic evidence for linkage. The threshold c was motivated by the condition that the posterior probability for $\lambda_{ij} > 50\%$ should be larger than $\alpha = 80\%$ or alternatively $\alpha = 95\%$, i.e.

$$p\left(\lambda_{ij} > 0.5|k_{\mathrm{L}}, n\right) > \alpha. \tag{3}$$

We simplified this criterion by choosing $c \in (0, 1)$ such that Eq. (3) holds for all $k_{\mathrm{L}} > nc$ for a typical whole-genome analysis. For the Rakai analysis, read alignments had a length of 250 bp, resulting in $n = 35$ non-overlapping alignments and deep-sequence phylogenies, and so with Eq. (2), we obtain $c = 0.57$ for $\alpha = 80\%$ and $c = 0.64$ for $\alpha = 95\%$. The thresholds were similar for analyses based on read alignments of length 350 bp, resulting in $n = 25$ deep-sequence phylogenies, and $c = 0.59$ for $\alpha = 80\%$ and $c = 0.67$ for $\alpha = 95\%$. This suggested choosing as default values $c = 0.6$ for $\alpha = 80\%$ and $c = 0.66$ for $\alpha = 95\%$, with the present analysis based on $c = 0.6$ for all linkage and direction classifications.

Reporting Summary. Further information on experimental design is available in the Nature Research Reporting Summary linked to this article.

Data availability

The deep-sequence phylogenies and basic individual-level data analysed during the current study are available in the Dryad repository, https://doi.org/10.5061/dryad.7h46hg2. HIV-1 reads are available on reasonable request through the PANGEA consortium (www.pangea-hiv.org) or the corresponding author. Please contact project manager Lucie Abeler-Dörner (lucie.abeler-dorner@bdi.ox.ac.uk) for further details. Additional individual-level data are available on reasonable request to RHSP or the corresponding author.

Code availability

Code is available from https://github.com/BDI-pathogens/phyloscanner (version 1.1.2) and https://github.com/olli0601/Phyloscanner.R.utilities (version 0.7) under the GNU General Public License v3.0.

Received: 5 October 2018 Accepted: 22 February 2019
Published online: 29 March 2019

References

1. UNAIDS. UNAIDS Data 2017, Document JC2910E. http://www.unaids.org/en/resources/documents/2017/2017_data_book (2017).
2. Grabowski, M. K. et al. HIV prevention efforts and incidence of HIV in Uganda. *N. Engl. J. Med.* **377**, 2154–2166 (2017).
3. UNAIDS. Fast-track: ending the AIDS epidemic by 2030, Document JC2686. http://www.unaids.org/en/resources/documents/2014/JC2686_WAD2014report (2014).
4. UNAIDS. Empower young women and adolescent girls: fast-track the end of the AIDS epidemic in Africa, Document JC2746. http://www.unaids.org/en/resources/documents/2015/JC2746 (2015).
5. Salazar-Gonzalez, J. F. et al. Deciphering human immunodeficiency virus type 1 transmission and early envelope diversification by single-genome amplification and sequencing. *J. Virol.* **82**, 3952–3970 (2008).
6. Maldarelli, F. et al. HIV populations are large and accumulate high genetic diversity in a nonlinear fashion. *J. Virol.* **87**, 10313–10323 (2013).
7. Dennis, A. M. et al. Phylogenetic studies of transmission dynamics in generalized HIV epidemics: an essential tool where the burden is greatest? *J. Acquir. Immune Defic. Syndr.* **67**, 181–195 (2014).
8. Pillay, D. et al. PANGEA-HIV: phylogenetics for generalised epidemics in Africa. *Lancet Infect. Dis.* **15**, 259–261 (2015).
9. Volz, E. et al. HIV-1 transmission during early infection in men who have sex with men: a phylodynamic analysis. *PLoS Med.* **10**, e1001568 (2013).
10. Stadler, T., Kuhnert, D., Bonhoeffer, S. & Drummond, A. J. Birth-death skyline plot reveals temporal changes of epidemic spread in HIV and hepatitis C virus (HCV). *Proc. Natl. Acad. Sci. USA* **110**, 228–233 (2013).
11. Grabowski, M. K. et al. The role of viral introductions in sustaining community-based HIV epidemics in rural Uganda: evidence from spatial clustering, phylogenetics, and egocentric transmission models. *PLoS Med.* **11**, e1001610 (2014).
12. de Oliveira, T. et al. Transmission networks and risk of HIV infection in KwaZulu-Natal, South Africa: a community-wide phylogenetic study. *Lancet HIV* **4**, e41–e50 (2017).
13. Le, Vu,S. et al. Comparison of cluster-based and source-attribution methods for estimating transmission risk using large HIV sequence databases. *Epidemics* **23**, 1–10 (2018).
14. Barre-Sinoussi, F. et al. Expert consensus statement on the science of HIV in the context of criminal law. *J. Int. AIDS Soc.* **21**, e25161 (2018).
15. Ratmann, O. et al. Sources of HIV infection among men having sex with men and implications for prevention. *Sci. Tr. Med* **8**, 320ra2 (2016).
16. Eshleman, S. H. et al. Analysis of genetic linkage of HIV from couples enrolled in the HIV Prevention Trials Network 052 trial. *J. Infect. Dis.* **204**, 1918–1926 (2011).
17. Campbell, M. S. et al. Viral linkage in HIV-1 seroconverters and their partners in an HIV-1 prevention clinical trial. *PLoS ONE* **6**, e16986 (2011).
18. Volz, E. M. et al. Molecular epidemiology of HIV-1 subtype B reveals heterogeneous transmission risk: implications for intervention and control. *J. Infect. Dis.* **217**, 1522–1529 (2018).
19. Didelot, X., Fraser, C., Gardy, J. & Colijn, C. Genomic infectious disease epidemiology in partially sampled and ongoing outbreaks. *Mol. Biol. Evol.* **34**, 997–1007 (2017).
20. Grabowski, M. K. & Lessler, J. Phylogenetic insights into age-disparate partnerships and HIV. *Lancet HIV* **4**, e8–e9 (2017).
21. Wymant, C. et al. PHYLOSCANNER: inferring transmission from within- and between-host pathogen genetic diversity. *Mol. Biol. Evol.* **35**, 719–733 (2017).
22. Romero-Severson, E. O., Bulla, I. & Leitner, T. Phylogenetically resolving epidemiologic linkage. *Proc. Natl. Acad. Sci. USA* **113**, 2690–2695 (2016).
23. Leitner, T. & Romero-Severson, E. Phylogenetic patterns recover known HIV epidemiological relationships and reveal common transmission of multiple variants. *Nat. Microbiol.* **3**, 983–988 (2018).
24. Serwadda, D. et al. Slim disease: a new disease in Uganda and its association with HTLV-III infection. *Lancet* **2**, 849–852 (1985).
25. Chang, L. W. et al. Heterogeneity of the HIV epidemic in agrarian, trading, and fishing communities in Rakai, Uganda: an observational epidemiological study. *Lancet HIV* **3**, e388–e396 (2016).
26. Grabowski, M. K. et al. The validity of self-reported antiretroviral use in persons living with HIV: a population-based study. *AIDS* **32**, 363–369 (2018).
27. Gall, A. et al. Universal amplification, next-generation sequencing, and assembly of HIV-1 genomes. *J. Clin. Microbiol.* **50**, 3838–3844 (2012).
28. Ratmann, O. et al. HIV-1 full-genome phylogenetics of generalized epidemics in sub-Saharan Africa: impact of missing nucleotide characters in next-generation sequences. *AIDS Res. Hum. Retroviruses* **33**, 1083–1098 (2017).
29. Rose, R. et al. Identifying transmission clusters with cluster picker and HIV-TRACE. *AIDS Res. Hum. Retrovir.* **33**, 211–218 (2017).
30. Romero-Severson, E. O. et al. Donor-recipient identification in para- and poly-phyletic trees under alternative HIV-1 transmission hypotheses using approximate Bayesian computation. *Genetics* **207**, 1089–1101 (2017).
31. Carlson, J. M. et al. HIV transmission. Selection bias at the heterosexual HIV-1 transmission bottleneck. *Science* **345**, 1254031 (2014).
32. Hue, S. et al. HIV type 1 in a rural coastal town in Kenya shows multiple introductions with many subtypes and much recombination. *AIDS Res. Hum. Retrovir.* **28**, 220–224 (2012).
33. Novitsky, V. et al. Phylogenetic relatedness of circulating HIV-1C variants in Mochudi, Botswana. *PLoS ONE* **8**, e80589 (2013).
34. Chan, S. K. et al. Likely female-to-female sexual transmission of HIV–Texas, 2012. *Mmwr. Morb. Mortal. Wkly. Rep.* **63**, 209–212 (2014).
35. Fraser, C. et al. Virulence and pathogenesis of HIV-1 infection: an evolutionary perspective. *Science* **343**, 1243727 (2014).
36. Hladik, W. et al. Men who have sex with men in Kampala, Uganda: Results from a bio-behavioral respondent driven sampling survey. *AIDS Behav.* **21**, 1478–1490 (2017).
37. Rose, R. et al. Phylogenetic methods inconsistently predict direction of HIV transmission among heterosexual pairs in the HPTN052 cohort. *J. Infect. Dis.*, https://doi.org/10.1093/infdis/jiy734 (2018).
38. De Silva, D. et al. Whole-genome sequencing to determine transmission of Neisseria gonorrhoeae: an observational study. *Lancet Infect. Dis.* **16**, 1295–1303 (2016).
39. Fifer, H. et al. Sustained transmission of high-level azithromycin-resistant Neisseria gonorrhoeae in England: an observational study. *Lancet Infect. Dis.* **18**, 573–581 (2018).
40. Dellicour, S. et al. Phylodynamic assessment of intervention strategies for the West African Ebola virus outbreak. *Nat. Commun.* **9**, 2222 (2018).
41. Poon, A. F. et al. Near real-time monitoring of HIV transmission hotspots from routine HIV genotyping: an implementation case study. *Lancet HIV* **3**, e231–e238 (2016).
42. Oster, A. M., France, A. M. & Mermin, J. Molecular epidemiology and the transformation of HIV prevention. *JAMA* **319**, 1657–1658 (2018).
43. Skums, P. et al. QUENTIN: reconstruction of disease transmissions from viral quasispecies genomic data. *Bioinformatics* **34**, 163–170 (2018).
44. Bernard, E.J., Cameron, S., HIV Justice Network & GNP+. Advancing HIV Justice 2: Building momentum in global advocacy against HIV criminalisation. http://www.hivjustice.net/wp-content/uploads/2016/05/AHJ2.final2_.10May2016.pdf (2016).
45. Yebra, G. et al. Using nearly full-genome HIV sequence data improves phylogeny reconstruction in a simulated epidemic. *Sci. Rep.* **6**, 39489 (2016).
46. Novitsky, V. et al. Long-range HIV genotyping using viral RNA and proviral DNA for analysis of HIV drug resistance and HIV clustering. *J. Clin. Microbiol.* **53**, 2581–2592 (2015).
47. Bonsall, D. et al. A comprehensive genomics solution for HIV surveillance and clinical monitoring in a global health setting. Preprint at *bioRxiv*, https://www.biorxiv.org/content/early/2018/08/23/397083 (2018).
48. Sypsa, V. et al. Rapid decline in HIV incidence among persons who inject drugs during a fast-track combination prevention program after an HIV outbreak in Athens. *J. Infect. Dis.* **215**, 1496–1505 (2017).
49. Chewapreecha, C. et al. Dense genomic sampling identifies highways of pneumococcal recombination. *Nat. Genet.* **46**, 305–309 (2014).
50. Paterson, G. K. et al. Capturing the cloud of diversity reveals complexity and heterogeneity of MRSA carriage, infection and transmission. *Nat. Commun.* **6**, 6560 (2015).
51. Wymant, C. et al. Easy and accurate reconstruction of whole HIV genomes from short-read sequence data. *Virus Evol.* **4**, vey007 (2018).
52. Hunt, M. et al. IVA: accurate de novo assembly of RNA virus genomes. *Bioinformatics* **31**, 2374–2376 (2015).
53. Bankevich, A. et al. SPAdes: a new genome assembly algorithm and its applications to single-cell sequencing. *J. Comput. Biol.* **19**, 455–477 (2012).
54. Nurk, S., Meleshko, D., Korobeynikov, A. & Pevzner, P. A. metaSPAdes: a new versatile metagenomic assembler. *Genome Res.* **27**, 824–834 (2017).
55. Wood, D. E. & Salzberg, S. L. Kraken: ultrafast metagenomic sequence classification using exact alignments. *Genome Biol.* **15**, R46 (2014).
56. Carey, V., Long, L. & Gentleman, R. RBGL: an interface to the BOOST graph library, version 1.55.1. http://bioconductor.org/packages/release/bioc/html/RBGL.html (2017).

Acknowledgements

We thank the participants of the RHSP RCCS; as well as the PANGEA-HIV steering committee for their input and their comments on a previous version of this article. Computations were performed at the Imperial College Research Computing Service, https://doi.org/10.14469/hpc/2232. This study was supported by the National Institute of Mental Health (K23MH086338, R01MH107275); the National Institute of Allergy and Infectious Diseases (R01AI110324, U01AI100031, R01AI110324, R01AI102939); the National Institute of Child Health and Development (RO1HD070769, R01HD050180); the Division of Intramural Research, National Institute for Allergy and Infectious Diseases, National Institutes of Health; the Bill & Melinda Gates Foundation (22006.02, OPP1084362); the Johns Hopkins University Center for AIDS Research (P30AI094189); and the European Research Council (Advanced Grant PBDR-339251).

Author contributions

O.R., M.K.G., A.L.B., TdO, P.K., D.P., T.C.Q., M.J.W., D.S., R.H.G., C.F. conceived the study; M.K.G., J.K., G.K., O.L., T.C.Q., M.J.W., D.S., R.H.G., A.G., D.B. selected, provided and prepared sequence and patient data; L.A.-D., A.H., T.G. provided managerial and logistical support, including data tracking; M.K.G., C.W., T.G. assembled deep-sequence reads; O.R., M.H. performed computations and statistical analyses; O.R., M.K.G., C.F. evaluated statistical analyses; O.R. wrote the first version of the manuscript; all authors reviewed and approved the statistical analysis and final version of the manuscript.

Additional information

Supplementary Information accompanies this paper at https://doi.org/10.1038/s41467-019-09139-4.

Competing interests: The authors declare no competing interests.

Reprints and permission information is available online at http://npg.nature.com/reprintsandpermissions/

Journal peer review information: *Nature Communications* thanks Denise Kühnert, Thomas Leitner, and the anonymous reviewers for their contribution to the peer review of this work.

Publisher's note: Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

PANGEA Consortium and Rakai Health Sciences Program

Helen Ayles[15], Rory Bowden[16], Vincent Calvez[17], Myron Cohen[18], Ann Dennis[18], Max Essex[19,20], Sarah Fidler[10], Daniel Frampton[6], Richard Hayes[21], Joshua T. Herbeck[22], Pontiano Kaleebu[23], Cissy Kityo[24], Jairam Lingappa[22], Vladimir Novitsky[25], Nick Paton[26], Andrew Rambaut[7], Janet Seeley[21], Deogratius Ssemwanga[23], Frank Tanser[11], Gertrude Nakigozi[4], Robert Ssekubugu[4], Fred Nalugoda[4], Tom Lutalo[4], Ronald Galiwango[4], Fred Makumbi[4], Nelson K. Sewankambo[4], Aaron A. R. Tobian[4], Steven J. Reynolds[3,4], Larry W. Chang[3,4], Dorean Nabukalu[4], Anthony Ndyanabo[4], Joseph Ssekasanvu[4,13], Hadijja Nakawooya[4], Jessica Nakukumba[4], Grace N. Kigozi[4], Betty S. Nantume[4], Nampijja Resty[4], Jedidah Kambasu[4], Margaret Nalugemwa[4], Regina Nakabuye[4], Lawrence Ssebanobe[4], Justine Nankinga[4], Adrian Kayiira[4], Gorreth Nanfuka[4], Ruth Ahimbisibwe[4], Stephen Tomusange[4], Ronald M. Galiwango[4], Sarah Kalibbali[4], Margaret Nakalanzi[4], Joseph Ouma Otobi[4], Denis Ankunda[4], Joseph Lister Ssembatya[4], John Baptist Ssemanda[4], Robert Kairania[4], Emmanuel Kato[4], Alice Kisakye[4], James Batte[4], James Ludigo[4], Abisagi Nampijja[4], Steven Watya[4], Kighoma Nehemia[4], Margaret Anyokot Sr.[4], Joshua Mwinike[4], George Kibumba[4], Paschal Ssebowa[4], George Mondo[4], Francis Wasswa[4], Agnes Nantongo[4], Rebecca Kakembo[4], Josephine Galiwango[4], Geoffrey Ssemango[4], Andrew D. Redd[3,4], John Santelli[4,27], Caitlin E. Kennedy[4] & Jennifer Wagman[4,28]

[15]Zambart Project, Lusaka, P.O. Box 50697, Zambia. [16]Oxford Genomics Centre, The Wellcome Centre for Human Genetics, University of Oxford, Oxford OX3 7BN, UK. [17]Ecole Normale Supérieure de Lyon, Lyon 69007, France. [18]Department of Medicine, University of North Carolina, Chapel Hill, NC 27516, USA. [19]Harvard T.H. Chan School of Public Health AIDS Initiative, Harvard T.H. Chan School of Public Health, Boston, MA 02115, USA. [20]Botswana Harvard AIDS Institute Partnership, Gaborone Private Bag BO 320, Botswana. [21]London School of Hygiene & Tropical Medicine, London WC1E 7HT, UK. [22]Department of Global Health, University of Washington, Seattle, WA 98104, USA. [23]MRC/UVRI, Entebbe, P.O.Box 49, Uganda. [24]Joint Clinical Research Centre, Kampala, P.o.Box 10005, Uganda. [25]Department of Immunology and Infectious Diseases, Harvard T.H. Chan School of Public Health, Boston, MA 02115, USA. [26]Medical Research Council, London WC2B 4AN, UK. [27]Mailman School of Public Health, Columbia University, New York, NY 10032, USA. [28]School of Medicine, University of California San Diego, San Diego, CA 92093, USA

Inferring HIV-1 transmission networks and sources of epidemic spread in Africa with deep-sequence phylogenetic analysis

Ratmann et al.

Supplementary tables and figures

Supplementary Table 1. Specification of deep-sequence phylogenetic analysis at the population-level: inference of deep-sequence phylogenies.

Phyloscanner input parameter	**Description**	**Value**	**Comments**
phyloscanner_make_trees.py			
Input read file (no prefix)	Input read file	csv file	File specifying bam and reference files for each individual in one phyloscanner run. In total 1896 files were processed in parallel. This corresponded to batches of 50-75 individuals that systematically queried all possible pairwise phylogenetic relationship in the population sample. The aim of the stage 1 analysis (see Methods) was to identify all phylogenetically close pairs in the population sample.
--x-samtools	Samtools options	samtools	Phyloscanner default.
--x-mafft	Alignment options	mafft	Phyloscanner default.
--x-raxml	Phylogeny options	raxmlHPC-AVX -m GTRCAT --HKY85 -p 42	24 models were compared on 27 read alignments with jModelTest2, https://github.com/ddarriba/jmodeltest2[1]: 3 substitution models (all rates equal, unequal transitions/transversions, all rates unequal), 2 base frequencies (equal, unequal), 2 rate variation models (none, Γ4), 2 invariant site models (none, proportion invariant). HKY85+Γ4 had by far the largest sum of all model probabilities across all read alignments and was thus chosen for our analysis.
--alignment-of-other-refs	Background sequences	HIV1_compendium_AD_B_CPX_v2.fasta	Full-genome HIV-1 sequences in the 2012 compendium of the Los Alamos HIV sequence data base[2], that were of subtype A and D, plus HXB2 and CPX AF460972. HXB2 was used for setting default coordinates across the genome, and AF460972 was used for rooting each deep-sequence phylogeny. The alignment is included in the R package Phyloscanner.R.utilities.
--outgroupName	Root	REF_CPX_AF460972	Name of the root sequence in the background sequences file. As sensitivity analysis, a limited number of phyloscanner runs were conducted with group M root sequences. This did not have any measurable impact on tree length and node heights.
--pairwise-align-to	Sequence against which to map genome coordinates	REF_B_K03455	Name of HXB2 in the background sequences file.
--merge-paired-reads	Overlapping mates are merged into one read	Flag set	This value was set since sequencing output consisted of paired-end reads.
--discard-improper-pairs	Paired-end reads that are flagged as improperly paired are discarded	Flag not set	This function was not available at time of analysis, and is now generally recommended.
--quality-trim-ends	Phred quality score to trim ends of reads	23	This value was set to exclude poor quality ends of reads, as determined by the Phred score.
--min-internal-quality	Phred quality score to discard reads with more than one base below threshold after trimming	23	This value was set to excluded reads with poor internal quality, as determined by the Phred score.
--merging-threshold-a	Genetic similarity threshold for merging similar reads	1	Reads that differed by just one base or a one-base indel were merged in stage 1 (see Methods). This enabled us to reconstruct deep-sequence phylogenies from reads of approximately 75 individuals per run, and keeping a computational budget of at most 24 hours per deep-sequence phylogeny reconstruction.
--min-read-count	Minimum count of unique reads so they were included in read alignments	2	Unique reads that occurred, after merging, just once were ignored in stage 1 (see Method). This enabled us to reconstruct deep-sequence phylogenies from reads of approximately 75 individuals per run, and keeping a

			computational budget of at most 24 hours per deep-sequence phylogeny reconstruction.
--check-recombination	Perform triplet recombination check	Flag not set	Computationally too expensive for the read alignments as specified above. No recombination checks were performed.
--dont-check-duplicates	Compare reads between individuals to find duplicates	Flag set	The resulting list of potential duplicates was used to discard potential contaminants at a later stage.
--windows	Start and end coordinates of genomic windows	From 800 to 9400 in 125bp increments of 250bp windows	The window length was chosen so that 75% of subjects were retained in analysis. Windows were incremented by 125bp, which we considered sufficient to identify individuals with phylogenetically close subgraphs.
--num-bootstraps	Number of bootstrap trees reconstructed per read alignment	None	Rather than bootstrapping non-overlapping read alignments, we opted instead to reconstruct deep-sequence phylogenies from tightly overlapping read alignments. This procedure aimed at capturing in addition to phylogenetic uncertainty also uncertainty in deep sequencing and alignment reconstruction.

Supplementary Table 2. Specification of deep-sequence phylogenetic analysis at the population-level: inference of phylogenetically close individuals.

Phyloscanner input parameter	**Description**	**Value**	**Comments**
NormalisationLookupWriter.R			
--norm.file.name	Reference table of tree summary statistics across the genome	hiv.hxb2.norm.constants.rda	To capture changes in evolutionary rates across the HIV-1 genome, Group M sequences in the 2012 compendium alignment of the Los Alamos HIV sequence data base[2] were selected, trimmed to 300bp regions that shifted across the genome by 1bp, phylogenies were reconstructed with RAxML[3] using default options, and several tree summaries were calculated (median pairwise distance, mean pairwise distance, maximum pairwise distance, sum of branch lengths). This file is part of the R package Phyloscanner.R.utilities. Branch lengths of each deep-sequence phylogeny were multiplied with a normalization factor derived from one of these statistics. Specifically, we calculated the average statistic in a reference gene, and then calculated the ratio of that statistic at any base pair divided by the average in the reference gene.
--norm.var	Tree summary statistic used.	median pairwise distance	Phyloscanner default.
--standardize	Normalise summary statistic so that its average on the concatenated *gag+pol* gene equals one.	Flag set	Sets the reference gene to the *gag+pol* gene region. This implied that the evolutionary distances shown in scanplots and reported in the main text can be interpreted as average distances expected in the *pol* gene.
parsimony_based_blacklister.R			
--multifurcation Threshold	Threshold to collapse branches in NGS phylogenies into polytomies.	1e-5	RAxML returns strictly bifurcating trees with minimum-length branches that in fact imply multifurcations. The minimum length can vary, and we set the threshold to the typical minimum branch length value given by RAxML[3].
--sankoffK	K parameter in Sankoff cost matrix	20	This value was chosen by testing different values of *k* on the whole dataset and examining the distribution of multiple infections that they give. From this analysis, we recommend setting the value to the reciprocal of a pairwise genetic diversity (in substitutions per site) that would be unrealistic to see in an infection with a single source. Based on the analysis reported in Figure 3A, that value would be 0.05 substitutions per site.
--rawThreshold	Subgraphs with fewer read counts are flagged	10	Divergent within-host subgraphs containing just one read could be contaminants, and should be excluded from further analysis. We opted for a threshold of 10 after analyzing the

	as potential contaminants and discarded.		frequency of divergent subgraphs with few reads, see supplementary text S2.
--ratioThreshold	Subgraphs, whose tip count divided by that of another subgraph from the same subject is less than this threshold, are flagged as potential contaminants and discarded.	0	Additional and/or alternative threshold for excluding potential contaminants. We only used a threshold on the absolute number of reads in divergent subgraphs.
downsample_reads.R			
--maxReadsPerHost	Downsample reads to at most this number if more reads are present	50	Reads were downsampled to reduce preferential assignment of well-sampled individuals as being ancestral to others. There is currently no strong evidence suggesting that this option is necessary for deep-sequence phylogenetic analysis.
--excludeUnderrepresented	Hosts with less than maxReadsPerHost are discarded	Flag not set	All individuals were kept as controls for pairs of individuals who met minimum read criteria specified at a later point below.
split_hosts_to_subgraphs.R			
--pruneBlacklist	Prune all blacklisted reads from NGS phylogeny before ancestral state reconstruction	Flag not set	All reads were retained to enable investigation of potential contaminants from final output.
--splitsRule	Algorithm for identifying distinct subgraphs among NGS reads of one individual.	Sankoff algorithm	Phyloscanner default.
--kParam	K parameter in Sankoff cost matrix	20	Same as argument --sankoffK above.
--proximityThreshold	Distance parameter that determines when ancestral states return to unsampled individuals	0	This value was set so that ancestral state reconstruction did not depend on phylogenetic branch lengths.
--readCountsMatterOnZeroBranches	Ancestral state reconstruction at parents of zero-branch lengths depends on read counts of children.	Flag set	Generally recommended when there is considerable variation in duplicate read counts.
summary_statistics.R No additional input arguments.			
classify_relationships.R No additional input arguments.			
TransmissionSummary.R			
--minThreshold	Summarize pairwise relationships only when they are not disconnected in at least this many potentially overlapping windows.	1	Summarize all pairwise relationships in csv summary file.
--distanceThreshold	Summarize pairwise relationships only when their subgraph distances are below this threshold.	Inf	Summarize all pairwise relationships in csv summary file.
--allowMultiTrans	If absent, directionality is only inferred between two subjects when both subjects have one subgraph, and the two subgraphs are ancestral.	Flag set	Set so that directionality between two subjects was also inferred when one or both subjects had more than one subgraph, and all subgraphs of one subject were ancestral to all subjects of the other individual. Generally recommended for HIV.

phsc.read.processed.phyloscanner.output.in.directory.Rscript			
--trmw.min.reads	Minimum number of reads for both individuals in one window.	30	A value of 100 is usually recommended. Here we chose a smaller value in order to retain for analysis 75% of individuals for whom deep-sequence data was available. The low value reflects relatively poor deep sequencing quality of our data.
--trmw.min.tips	Minimum number of tips for both individuals in one window.	1	Retain all pairwise relationships; in particular we consider also individuals with no sampled viral diversity.
--trmw.close.brl	Distance parameter to classify subgraphs as phylogenetically close.	0.035 substitutions per site	Based on the couples' analysis reported in Figure 3A, this threshold is 0.025 substitutions per site. To ensure all potentially phylogenetically close pairs were found in stage 1 analysis (see Methods), this value was initially set to 0.035 substitutions per site, and then set to 0.025 substitutions per site in stage 2 analyses.
--trmw.distant.brl	Distance parameter to classify subgraphs as phylogenetically distant.	0.08 substitutions per site	Based on the couples' analysis reported in Figure 3A, this threshold is 0.05 substitutions per site. To ensure all potentially phylogenetically close pairs were found in stage 1 analysis (see Methods), this value was initially set to 0.08 substitutions per site, and then set to 0.05 substitutions per site in stage 2 analyses.
--trmw.min.neff	Minimum number of effectively non-overlapping windows.	3	The phylogenetic relationship between any pair of individuals was not evaluated when data was available from read alignments covering less than 750nt of the HIV-1 genome.
--prior.keff	Hyperparameter on number of effectively non-overlapping windows of one type.	1	Corresponds to flat prior.
--confidence.cut	Confidence threshold for classification.	0.5	We used a cut-off of 60% in stage 2 analyses, see Methods. To ensure all potentially phylogenetically close pairs were found in stage 1 analysis (see Methods), this value was initially set to 50%.
--rel.XXX	Flags to generate output classifications.	Flags set	All output classifications were included for comparative analyses, though this is typically not necessary.

Supplementary Table 3. Specification of deep-sequence phylogenetic analysis at the population-level: inference of transmission networks.

Phyloscanner input parameter	Description	Value	Comments
Input read file (no prefix)	Input read file	csv file	File specifying bam and reference files for each individual in one phyloscanner run. From stage 1 (see Methods), potential networks of phylogenetically close individuals were identified using the criteria in Figure 4 and Methods. To these networks, we added as controls reads from the next 10 phylogenetically closest individuals in stage 1 output. If networks contained only one of two partners who were known to have long-term sexual contact, the second person was added to the network. This resulted in 345 separate phyloscanner runs.
--merging-threshold-a	Genetic similarity threshold for merging similar reads	0	All distinct reads from one individual were kept to retain the entire sampled viral diversity for measuring subgraph relationships. This was a safe option to retain signal and incurred significant computational workload.
--min-read-count	Minimum count of unique reads so they were included in read alignments	1	All distinct reads from one individual were kept to retain the entire sampled viral diversity for measuring subgraph relationships. This was a safe option to retain signal and increased computational workload further.
--windows	Start and end coordinates of genomic windows	From 800 to 9400 in 25bp increments of 250bp windows	The window length was chosen so that 75% of mapped reads were retained in analysis. Windows were incremented by 25bp to capture 99% of mapped reads >250bp in at least one window. In comparison to bootstrap replicates on the

			same read alignment, overlapping windows accounted for uncertainty in read sequencing and the construction of read alignments.
--confidence.cut	Confidence threshold for classification.	0.6	See Methods.

Supplementary Table 4. Inference of phylogenetic transmission networks, sensitivity analyses.

	Phylogenetically inferred transmission chains		**Male-female pairs in inferred transmission chains**			
	Men and women	Links	Phylogenetic linkage highly supported		Phylogenetic linkage and source highly supported	
	(#)	(#)	(#)	(%)**	(#)	(%)***
Subgraphs with fewer read counts are flagged as potential contaminants and discarded (--rawThreshold).						
10 *	1334	888	376	42.3%	293	77.9%
20	1336	889	377	42.4%	290	76.9%
Minimum number of reads for both individuals in one window (--trmw.min.reads).						
10	1366	907	377	41.6%	293	77.7%
20	1362	914	378	41.4%	299	79.1%
30 *	1334	888	376	42.3%	293	77.9%
50	1307	867	374	43.1%	289	77.3%
Threshold to collapse branches in deep-sequence phylogenies into polytomies (--multifurcation Threshold).						
1e-05 *	1334	888	376	42.3%	293	77.9%
1e-03	1336	889	377	42.4%	294	78.0%
Downsample reads to at most this number if more reads are present (--maxReadsPerHost).						
30	1328	881	374	42.5%	298	79.7%
50 *	1334	888	376	42.3%	293	77.9%
100	1339	891	387	43.4%	311	80.4%
1000	1355	910	410	45.1%	326	79.5%
Prune all blacklisted reads from NGS phylogeny before ancestral state reconstruction (--pruneBlacklist).						
No *	1334	888	376	42.3%	293	77.9%
Yes	1329	884	375	42.4%	288	76.8%
K parameter in Sankoff cost matrix (--sankoffK, --kParam)						
10	1350	911	382	41.9%	296	77.5%
20 *	1334	888	376	42.3%	293	77.9%
Proximity parameter in Sankoff cost matrix						
0 substitutions per site *	1334	888	376	42.3%	293	77.9%
0.025 substitutions per site	1268	838	377	45.0%	290	76.9%
Directionality is only inferred between two subjects when both subjects have one subgraph, and the two subgraphs are ancestral (--allowMultiTrans).						
No	1330	885	376	42.5%	293	77.9%

Yes *	1334	888	376	42.3%	293	77.9%
Ancestral state reconstruction at parents of zero-branch lengths depends on read counts of children (----readCounts MatterOnZeroBranches).						
No	1337	891	378	42.4%	287	75.9%
Yes *	1334	888	376	42.3%	293	77.9%
Distance parameter to classify subgraphs as phylogenetically close (--trmw.close.brl).						
0.01 substitutions per site	1284	845	198	23.4%	153	77.3%
0.015 substitutions per site	1313	869	274	31.5%	218	79.6%
0.02 substitutions per site	1326	883	336	38.1%	258	76.8%
0.025 substitutions per site *	1334	888	376	42.3%	293	77.9%
0.03 substitutions per site	1331	887	423	47.7%	334	79.0%
0.035 substitutions per site	1338	891	452	50.7%	351	77.7%
0.04 substitutions per site	1339	892	471	52.8%	369	78.3%
Confidence cut-off on phyloscanner linkage and direction scores						
0.5	1334	888	434	48.9%	417	96.1%
0.55	1334	888	407	45.8%	356	87.5%
0.6 *	1334	888	376	42.3%	293	77.9%
0.65	1334	888	356	40.1%	244	68.5%
0.7	1334	888	328	36.9%	192	58.5%
0.75	1334	888	295	33.2%	130	44.1%
0.8	1334	888	258	29.1%	89	34.5%

* Input specification used in validation and central analysis. ** Proportion of links in inferred transmission chains. *** Proportion of male-female pairs between whom phylogenetic linkage was highly supported.

Supplementary Table 5. Inference of phylogenetically likely transmitters among couples, sensitivity analyses.

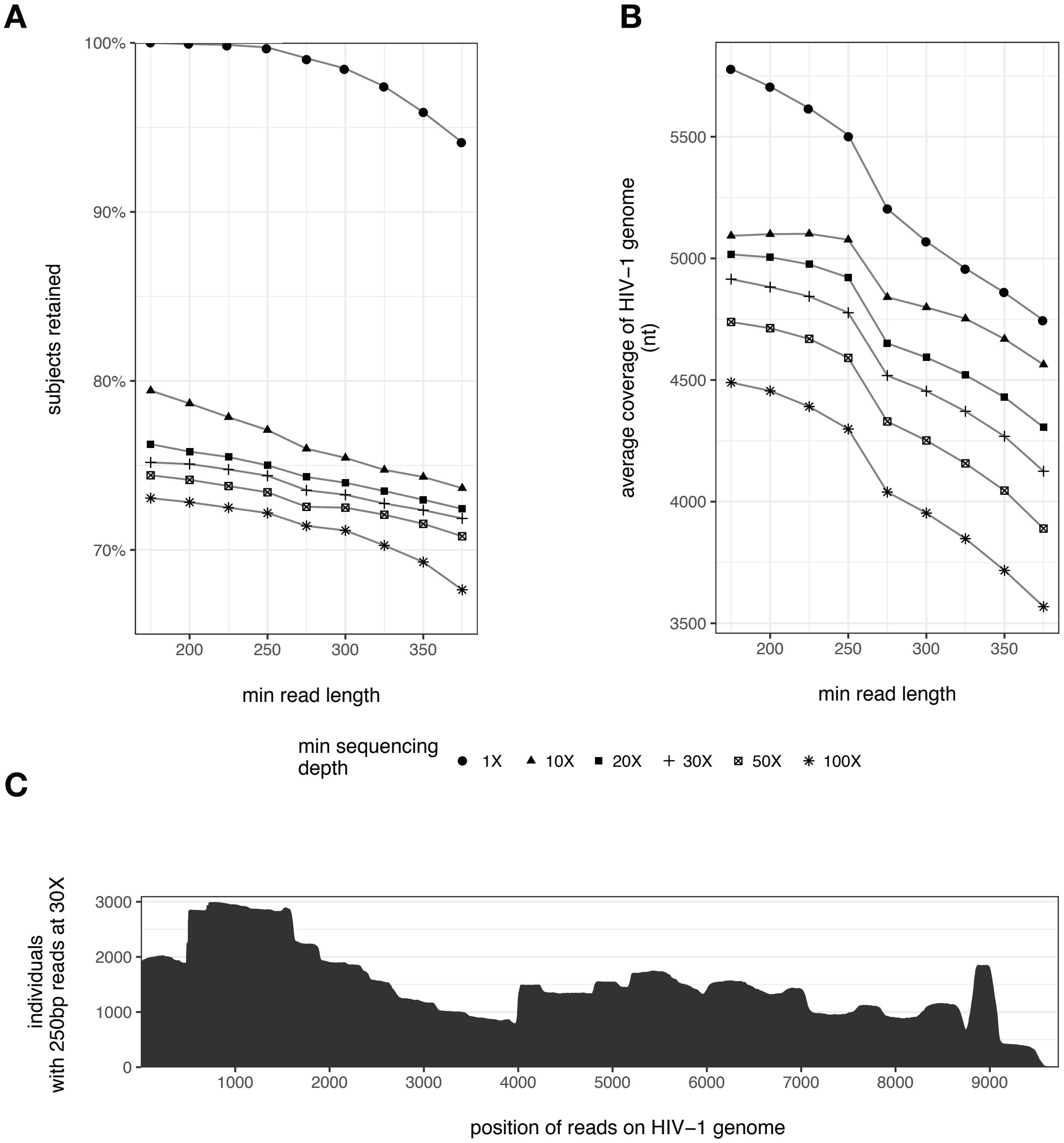
	Phylogenetically linked male-female pairs in population sample with clinical evidence for transmission in one direction, including couples				
	Inferred direction consistent	Direction not inferred	Inferred direction not consistent	False Discovery Rate	
	(#)	(#)	(#)	(point estimate)	(95% confidence interval)
Subgraphs with fewer read counts are flagged as potential contaminants and discarded (--rawThreshold).					
10 *	25	8	2	7.40%	[2.1%-23.4%]
20	18	7	2	10%	[2.8%-30.1%]
Minimum number of reads for both individuals in one window (--trmw.min.reads).					
10	19	8	3	13.60%	[4.7%-33.3%]
20	17	8	4	19%	[7.7%-40%]
30 *	25	8	2	7.40%	[2.1%-23.4%]
50	18	6	5	21.70%	[9.7%-41.9%]
Threshold to collapse branches in deep-sequence phylogenies into polytomies (--multifurcation Threshold).					
1e-05 *	25	8	2	7.40%	[2.1%-23.4%]
1e-03	17	8	2	10.50%	[2.9%-31.4%]
Downsample reads to at most this number if more reads are present (--maxReadsPerHost).					
30	19	5	3	13.60%	[4.7%-33.3%]

50 *	25	8	2	7.40%	[2.1%-23.4%]
100	17	8	2	10.50%	[2.9%-31.4%]
1000	21	7	3	12.50%	[4.3%-31%]
Prune all blacklisted reads from NGS phylogeny before ancestral state reconstruction (--pruneBlacklist).					
No *	25	8	2	7.40%	[2.1%-23.4%]
Yes	25	8	2	7.40%	[2.1%-23.4%]
K parameter in Sankoff cost matrix (--sankoffK, --kParam)					
10	19	6	2	9.50%	[2.7%-28.9%]
20 *	25	8	2	7.40%	[2.1%-23.4%]
Proximity parameter in Sankoff cost matrix					
0 substitutions per site *	25	8	2	7.40%	[2.1%-23.4%]
0.025 substitutions per site	17	7	3	15%	[5.2%-36%]
Directionality is only inferred between two subjects when both subjects have one subgraph, and the two subgraphs are ancestral (--allowMultiTrans).					
No	17	8	2	10.50%	[2.9%-31.4%]
Yes *	25	8	2	7.40%	[2.1%-23.4%]
Ancestral state reconstruction at parents of zero-branch lengths depends on read counts of children (----readCounts MatterOnZeroBranches).					
No	18	7	2	10%	[2.8%-30.1%]
Yes *	25	8	2	7.40%	[2.1%-23.4%]
Distance parameter to classify subgraphs as phylogenetically close (--trmw.close.brl).					
0.01 substitutions per site	10	6	1	9.10%	[0.5%-37.7%]
0.015 substitutions per site	14	5	2	12.50%	[3.5%-36%]
0.02 substitutions per site	16	10	1	5.90%	[0.3%-27%]
0.025 substitutions per site *	25	8	2	7.40%	[2.1%-23.4%]
0.03 substitutions per site	21	7	3	12.50%	[4.3%-31%]
0.035 substitutions per site	21	8	4	16%	[6.4%-34.7%]
0.04 substitutions per site	22	8	4	15.40%	[6.2%-33.5%]
Confidence cut-off on phyloscanner linkage and direction scores					
0.5	31	1	6	16.20%	[7.7%-31.1%]
0.55	28	5	3	9.70%	[3.3%-24.9%]
0.6 *	25	8	2	7.40%	[2.1%-23.4%]
0.65	23	10	2	8%	[2.2%-25%]
0.7	17	17	1	5.60%	[0.3%-25.8%]
0.75	10	20	1	9.10%	[0.5%-37.7%]
0.8	8	19	0	0%	[0%-32.4%]

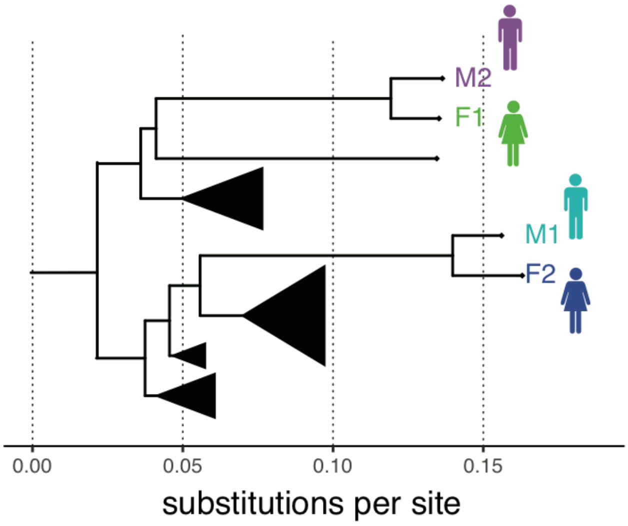
Supplementary Table 6. Inference of phylogenetically likely transmitters in the population-based sample, sensitivity analyses.

	Phylogenetically linked male-female pairs in population sample with clinical evidence for transmission in one direction, including couples				
	Inferred direction consistent	Direction not inferred	Inferred direction not consistent	False Discovery Rate	
	(#)	(#)	(#)	(point estimate)	(95% confidence interval)

Subgraphs with fewer read counts are flagged as potential contaminants and discarded (--rawThreshold).					
10 *	46	16	9	16.4%	[8.9%-28.3%]
20	48	16	8	14.3%	[7.4%-25.7%]
Minimum number of reads for both individuals in one window (--trmw.min.reads).					
10	49	13	8	14.0%	[7.3%-25.3%]
20	45	15	12	21.1%	[12.5%-33.3%]
30 *	46	16	9	16.4%	[8.9%-28.3%]
50	48	14	12	20.0%	[11.8%-31.8%]
Threshold to collapse branches in deep-sequence phylogenies into polytomies (--multifurcation Threshold).					
1e-05 *	46	16	9	16.4%	[8.9%-28.3%]
1e-03	46	16	9	16.4%	[8.9%-28.3%]
Downsample reads to at most this number if more reads are present (--maxReadsPerHost).					
30	49	11	11	18.3%	[10.6%-29.9%]
50 *	46	16	9	16.4%	[8.9%-28.3%]
100	46	15	10	17.9%	[10%-29.8%]
1000	54	14	10	15.6%	[8.7%-26.4%]
Prune all blacklisted reads from NGS phylogeny before ancestral state reconstruction (--pruneBlacklist).					
No *	46	16	9	16.4%	[8.9%-28.3%]
Yes	46	17	8	14.8%	[7.7%-26.6%]
K parameter in Sankoff cost matrix (--sankoffK, --kParam)					
10	48	13	10	17.2%	[9.6%-28.9%]
20 *	46	16	9	16.4%	[8.9%-28.3%]
Proximity parameter in Sankoff cost matrix					
0 substitutions per site *	46	16	9	16.4%	[8.9%-28.3%]
0.025 substitutions per site	45	16	10	18.2%	[10.2%-30.3%]
Directionality is only inferred between two subjects when both subjects have one subgraph, and the two subgraphs are ancestral (--allowMultiTrans).					
No	46	16	9	16.4%	[8.9%-28.3%]
Yes *	46	16	9	16.4%	[8.9%-28.3%]
Ancestral state reconstruction at parents of zero-branch lengths depends on read counts of children (----readCounts MatterOnZeroBranches).					
No	49	14	8	14.0%	[7.3%-25.3%]
Yes *	46	16	9	16.4%	[8.9%-28.3%]
Distance parameter to classify subgraphs as phylogenetically close (--trmw.close.brl).					
0.01 substitutions per site	26	8	5	16.1%	[7.1%-32.6%]
0.015 substitutions per site	35	7	9	20.5%	[11.2%-34.5%]
0.02 substitutions per site	43	16	8	15.7%	[8.2%-28%]
0.025 substitutions per site *	46	16	9	16.4%	[8.9%-28.3%]
0.03 substitutions per site	53	15	12	18.5%	[10.9%-29.6%]
0.035 substitutions per site	56	19	12	17.6%	[10.4%-28.4%]
0.04 substitutions per site	59	20	13	18.1%	[10.9%-28.5%]
Confidence cut-off on phyloscanner linkage and direction scores					
0.5	60	2	19	24.1%	[16%-34.5%]
0.55	52	11	13	20.0%	[12.1%-31.3%]
0.6 *	46	16	9	16.4%	[8.9%-28.3%]
0.65	44	18	8	15.4%	[8.0%-27.5%]
0.7	37	26	6	14.0%	[6.6%-27.3%]
0.75	25	32	3	10.7%	[3.7%-27.2%]
0.8	20	31	1	4.8%	[0.2%-22.7%]



Supplementary Figure 1. Characteristics of deep sequencing output of HIV-1 samples from Rakai District, Uganda. Deep sequencing was performed in high throughput on *Illumina* MiSeq and HiSeq instruments after automated extraction of viral RNA and amplification with a universal HIV-1 primer set[4]. Reads were mapped against de-novo reference sequences with shiver[5]. (**A**) The number of study subjects with deep sequencing output over at least 750nt of the HIV-1 genome decreased relatively steadily as a function of stricter requirements on the minimum sequencing depth at any position (symbols), and as a function of stricter requirements on the minimum length of reads increased (x-axis). 773 individuals were poorly sequenced with a read depth less than 10X. Approximately 3,000 individuals were retained at a minimum read depth of 10X to 30X. Slightly more individuals were lost to further analysis when the minimum read length was increased from 250nt to 275nt, as compared to other 25nt increases in minimum read length. (**B**) Coverage of the HIV-1 genome dropped more markedly between a minimum read length of 250nt and 275nt. This drop corresponded to situations when one of the two reads of a RNA template could be almost fully sequenced (length >250nt), but the second read failed to be sequenced in the opposite direction such that the two mates did not overlap, and did not produce a read of at least 275nt. We therefore set the minimum required read length to 250nt. (**C**) Considering individuals that could be deep sequenced at 30X with reads of at least 250nt over a minimum coverage of 750nt of the HIV-1 genome, most had reads covering the HIV-1 *gag* gene. Overall, in comparison to clinical samples from European HIV-1 subtype B patients, sequencing output on our African samples was of lower quality[6]. The minimum length of reads (250bp) was set lower compared to deep-sequence phylogenetic analyses on European samples (350bp), and chosen as described above by trading off against individuals retained. In general, phylogenetic reconstruction accuracy decays strongly with shorter read lengths[7], suggesting that a stronger phylogenetic signal into HIV-1 transmission networks could likely have been obtained if data had been of similar quality as obtained in Europe.

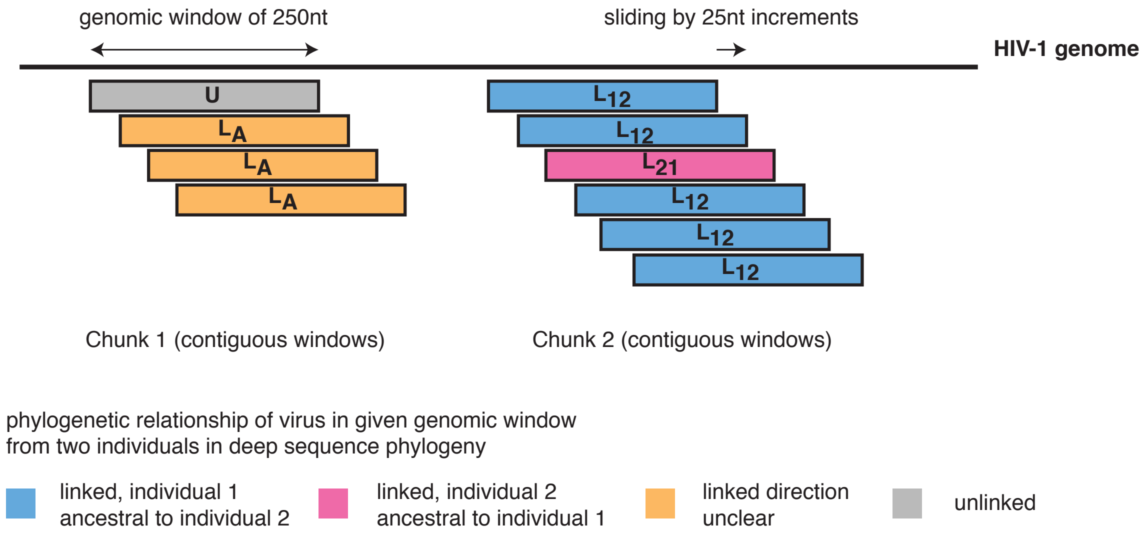


Supplementary Figure 2. Phylogenetic analysis from consensus sequences of the four selected individuals for whom deep-sequence phylogenetic analysis is illustrated in figure 1.

Inferring HIV-1 transmission networks and sources of epidemic spread in Africa with deep-sequence phylogenetic analysis

Supplementary Note 1. Calculation of phyloscanner scores

Consider the following example in which two individuals i and j had reads that overlapped ten genomic windows. Following the specification used on Rakai data, windows were 250nt long and slid by 25nt increments across the HIV-1 genome, with coordinates relative to HXB2 as shown in Supplementary Figure 3.


genomic window of 250nt
sliding by 25nt increments
HIV-1 genome
U
L_A
L_A
L_A
L_{12}
L_{12}
L_{21}
L_{12}
L_{12}
L_{12}
Chunk 1 (contiguous windows)
Chunk 2 (contiguous windows)
phylogenetic relationship of virus in given genomic window from two individuals in deep sequence phylogeny
linked, individual 1 ancestral to individual 2
linked, individual 2 ancestral to individual 1
linked direction unclear
unlinked

Supplementary Figure 3. Overlapping genomic windows. Phylogenetic trees were reconstructed for many genomic windows across the HIV-1 genome, which incremented by 25bp. If reads from individuals did not meet minimum quality criteria in a window, pairwise phylogenetic relationships between that and any other individual were not performed, leading to missing data. A series of contiguous pairwise phylogenetic relationships is referred to as a chunk. Subgraph topologies are indicated in colours.

For each window, phyloscanner constructs read alignments of 250nt in length, uses RAxML to infer corresponding deep-sequence phylogenies, identifies within-host subgraphs in these phylogenies, and characterizes their distance and topological relationship[8]. As illustrated in colours, for each genomic window, pairs are assigned to one of the five categories:

Symbol	Description	Definition (see Methods)
U	Phylogenetically unlinked.	$A_{ij} = 0$ or $\Delta_{ij} > 0.05$ substitutions per site
G	Greyzone phylogenetically linked.	$A_{ij} = 1, \Delta_{ij} \in [0.025 - 0.05$ substitutions per site]

L_{12}	Phylogenetically linked, with subgraphs from 1 ancestral to those of 2	$A_{ij} = 1, \Delta_{ij} < 0.025$ substitutions per site, $P_{ij} \geq 1$, $P_{ji} = 0$
L_{21}	Phylogenetically linked, with subgraphs from 2 ancestral to those of 1	$A_{ij} = 1, \Delta_{ij} < 0.025$ substitutions per site, $P_{ij} = 0$, $P_{ji} \geq 1$
L_A	Phylogenetically linked, with intermingled or sibling subgraphs,	$A_{ij} = 1, \Delta_{ij} < 0.025$ substitutions per site, $P_{ij} \geq 1$. $P_{ji} \geq 1$ or $A_{ij} = 1, \Delta_{ij} < 0.025$ substitutions per site, $P_{ij} = 0$, $P_{ji} = 0$

Observed pairwise relationships are then counted while adjusting for overlap in read alignments with the following algorithm.

Algorithm

Denote the unadjusted counts in order by $\tilde{k}_{\mathrm{U}}, \tilde{k}_{\mathrm{G}}, \tilde{k}_{ij}, \tilde{k}_{ji}, \tilde{k}_{\mathrm{A}}$, and their sum by $\tilde{n}$.

1. Identify genomic chunks c of consecutive genomic windows in which i and j have reads.
2. Calculate the effective number of non-overlapping windows in chunk c,

$$n_c = \frac{\max_{w \in c}(E_w) + 1 - \min_{w \in c}(S_w)}{E_w + 1 - S_w}$$

   where S_w, E_w are the first and last nucleotide positions in window w respectively. The numerator is the length of chunk c, and the denominator is the length of one window.
3. Calculate the effective number of non-overlapping windows in chunk c that are of type t,

$$k_{tc} = \frac{\tilde{k}_{tc}}{\tilde{n}_c / n_c}$$

   where $\tilde{k}_{tc}$ is the number of overlapping windows of type t in chunk c, and $\tilde{n}_c$ is the number of overlapping windows in chunk c.

Sum to obtain $n = \sum_c n_c$, and $k_t = \sum_c k_{tc}$ for all relationship types t.

In the example above, there are two chunks. Chunk 1 consists of 4 read alignments spanning 325nt, and contributes 1.3 effectively independent observations. Similary, chunk 2 consists of 6 read alignments spanning 375nt, and contributes 1.5 effectively independent observations:

Chunk	**Genomic windows**	**Length (nt)**	**Effectively independent observations**
Chunk 1	4	325	1.3 = 325/250
Chunk 2	6	375	1.5 = 375/250
Total	10	700	2.8

The adjusted counts are:

Chunk	**Adjusted counts**				
	$\mathbf{L_{12}}$	$\mathbf{L_{21}}$	$\mathbf{L_{A}}$	**G**	**D**
Chunk 1	0	0	$\frac{3}{4}$ * 1.3	0	$\frac{1}{4}$ * 1.3
Chunk 2	$\frac{5}{6}$ * 1.5	$\frac{1}{6}$ * 1.5	0	0	0
Total	1.25	0.25	0.975	0	0.325

The relative phylogenetic evidence for i and j being epidemiologically unlinked, and infection from i to j, and vice versa were thus:

Strength of phylogenetic evidence (point estimates)		
$\hat{\mu}_{ij}$	$\hat{\lambda}_{ij}$	$\hat{\delta}_{ij}$
0.325/2.8 = 0.12	(1.25+0.975+0.25)/2.8 = 0.88	1.25/(1.25+0.25) = 0.83

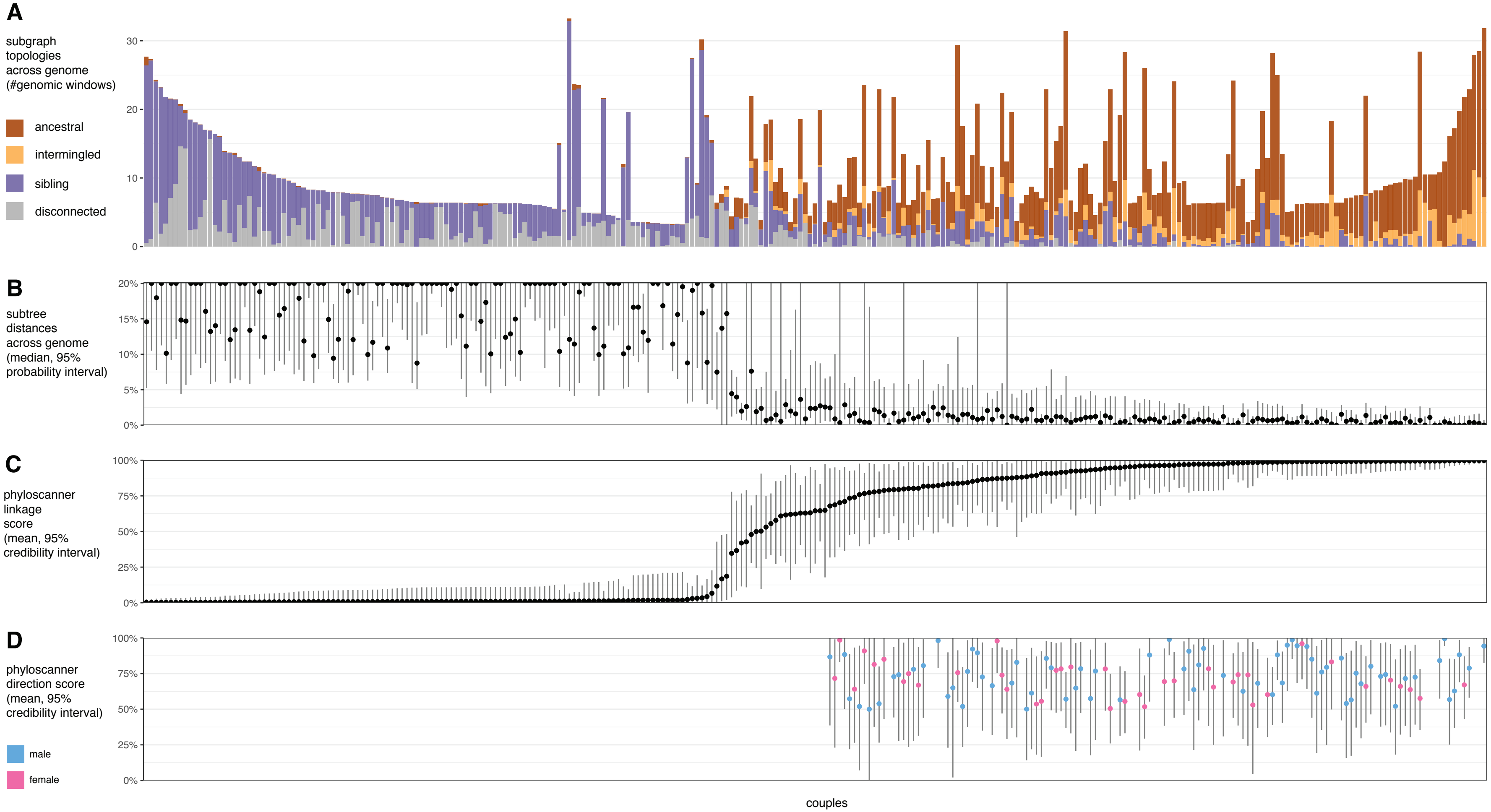
Inferring HIV-1 transmission networks and sources of epidemic spread in Africa with deep-sequence phylogenetic analysis

Supplementary Note 2. Inferring phylogenetic linkage from deep-sequence data compared to consensus sequences

We compared the agreement between phylogenetic linkage analysis from deep-sequence data and consensus sequence data on the couples' data set ($n = 331$ couples). Our primary aim was to assess concordance in estimating phylogenetic linkage on an empirical data set in which linkage is relatively unambiguous to characterize.

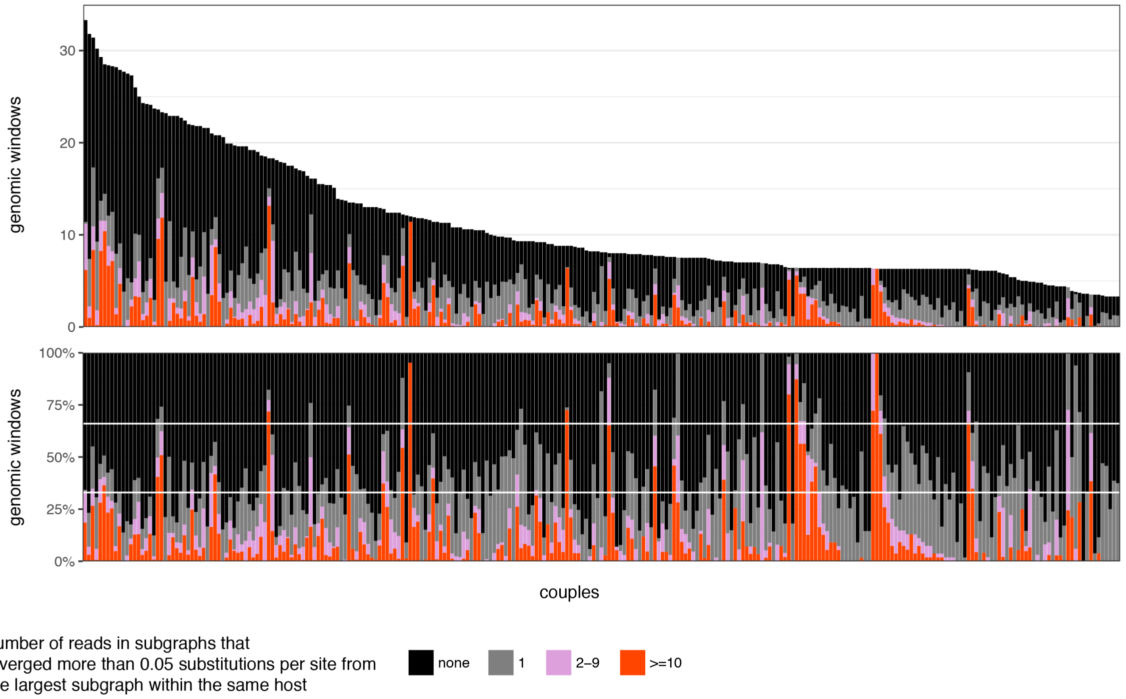
Deep-sequence phylogenetic analysis of couples

Supplementary Figure 4 summarizes deep-sequence viral phylogenetic analysis on the couples. Supplementary Figure 4A shows the number of deep-sequence phylogenies that were evaluated per couple (y-axis), after adjusting for overlap in read alignments. Subgraph topologies between spouses are indicated in colours. Couples did not necessarily both have sequencing output in any one genomic window, and for this reason the number of phylogenetic repeat observations per couple varied considerably (varying heights of bars). Supplementary Figure 4B illustrates median subgraph distances (dots) and empirical 95% confidence interval of subgraph distances per couple, where the median was taken across deep-sequence phylogenies, and after phylogenetic distances were rescaled to reflect typical distances observed in the HIV-1 *pol* gene (see Methods). Very large confidence intervals indicate that in some phylogenies, the subgraphs of couples were very close while in other phylogenies, their subgraphs were highly divergent, which may indicate read contamination, artifacts in tree reconstruction, recombination, or the presence of divergent and cocirculating viral variants in one or both individuals. Supplementary Figure 4C shows the linkage score $\hat{\lambda}_{ij}$ along with Bayesian 95% credibility intervals, which is based on subgraph distances and subgraph topologies as described in Methods. Supplementary Figure 4D shows the direction score $\hat{\delta}_{ij}$ along with corresponding Bayesian 95% credibility intervals.



Supplementary Figure 4. Viral phylogenetic relationships among 331 couples in Rakai District, Uganda, inferred from deep-sequence data. Please see text for details.

We further investigated wether one or both spouses harboured highly divergent virus, which could indicate dual infection or recombination. To this end, we catalogued for each spouse subgraphs that were highly divergent from the majority subgraph that contained most reads of that spouse in any phylogeny. Within-host subgraphs were considered highly divergent if they were more than 0.05 substitutions per site apart from the majority subgraph, based on the results shown in Figure 3A. Divergent subgraphs were further characterized by read number (1, 2-9, 10+). Supplementary Figure 5 illustrates that spouses frequently had divergent subgraphs of just one read, which could be due to read contamination and/or artifacts in tree reconstruction. 42 of 331 couples (12.7%) had at least one spouse with divergent subgraphs of at least 2 reads in more than 33% of deep-sequence phylogenies (after adjusting counts for overlap in genomic windows as described in Supplementary Note 1). 12 (3.6%) of 331 couples had divergent subgraphs of at least 2 reads in more than 66% of deep-sequence phylogenies.



Supplementary Figure 5. Counts and frequency of divergent virus within spouses. For each of the 331 couples with deep-sequence data (x-axis), deep-sequence phylogenies with divergent subgraphs in one or both spouses were counted, and are shown by the number of reads within them (colour). The number was adjusted for overlap of genomic windows (Supplementary Note 1). Overall, spouses frequently had divergent clades of just one read, indicative of read contamination. For the 6 couples that were classified linked using deep sequencing data but not linked using consensus sequences, at least one spouse had divergent subgraphs in at least 33% of (effective) deep-sequence phylogenies.

Generation of consensus sequences

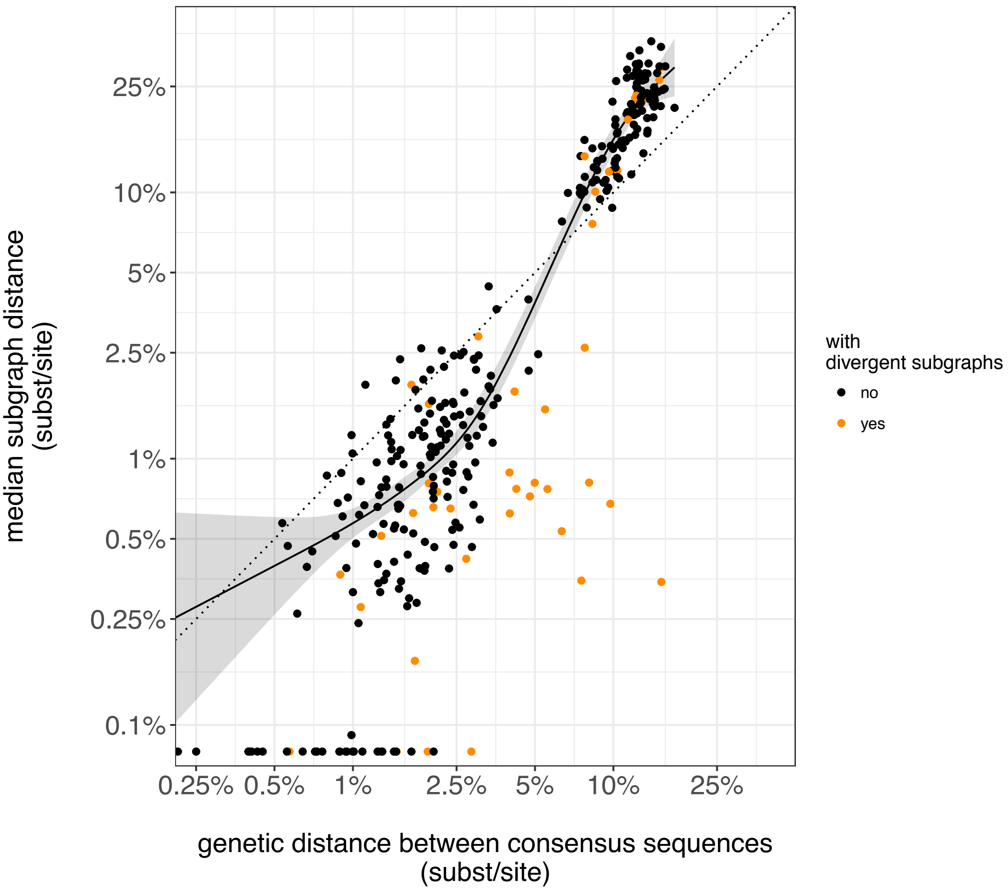
Consensus sequences were generated from mapped read alignments by determining the majority nucleotide call at each base position of the HIV-1 genome, as described in Ref.[6].

Concordance between phylogenetic distances in deep-sequence phylogenies with genetic distances between consensus sequences

For consensus sequences, genetic distances were calculated under three evolutionary models, Tamura-Nei-1993, Tamura-Nei-1993 with Gamma correction, and raw genetic distance using the ape package in R[9,10]. Phylogenetic linkage classification of the spouses from consensus sequences was identical under all three distance matrices, and results are reported for raw genetic distances.

Supplementary Figure 6 illustrates the bivariate relationship between the raw genetic distances obtained from consensus sequences versus median subgraph distances obtained from deep-sequence data. Shown in orange are the 42 couples for whom one or both individuals had divergent subgraphs of at least 2 reads in more than 33% of deep-sequence phylogenies. Overall, the two distance measures were highly correlated (Spearman log rank correlation coefficient $\rho = 0.87$).

To describe the relationship between both distance measures, polynomial splines were fitted to the data after excluding 42 couples with divergent subgraphs and 32 couples with identical subgraphs. A polynomial spline of order 4 provided the best fit and is shown as a line in Supplementary Figure 6.



Supplementary Figure 6. Concordance between median subgraph distances of couples in deep-sequence phylogenies and genetic distances between consensus sequences. Data from 311 couples were available to compare the two distance measures. For each couple, deep-sequence phylogenies were rescaled to account for variation in mutation rates across the genome, and the subgraph distance between couples was determined in all their deep-sequence phylogenies. Genetic distances were determined as described in the text. The plots show the bivariate relationship between median subgraph distances (with median taken over all phylogenies of a couple) and genetic distance between consensus sequences. Couples for whom one or both spouses had divergent subgraphs are shown in orange. For visualization purposes, couples with identical deep-sequence reads in 50% of deep-sequence phylogenies are shown on a horizontal line below 0.1% substitutions per site. The curve shows the best-fitting polynomial transformation between the two distance measures. The two distance measures were highly correlated (Spearman log rank correlation coefficient ρ = 0.87).

Phylogenetic linkage classification

Using deep-sequence data, couples were classified as phylogenetically linked as fully described in the main text by:

- identifying most likely transmission chains in the whole population sample,
- determining if couples were directly linked in a transmission chain,
- classifying a couple as phylogenetically linked with high support when the linkage score exceeded a particular threshold, here 60% ($\hat{\lambda}_{ij} > 0.6$; see Methods).

Using consensus sequences, couples were classified as phylogenetically linked by:

- identifying if the spouse was the genetically closest individual in the whole population sample,
- classifying a couple as phylogenetically linked when their genetic distance did not exceed a particular threshold.

The distance threshold for classifying couples as phylogenetically linked from consensus sequences was based on the transformation function shown in Supplementary Figure 6. Supplementary Table 7 lists corresponding distance thresholds, and further investigation was based on a threshold of 0.025 substitutions per site on subgraph distances (see results in Figure 3A) and the corresponding threshold of 0.041 substitutions per site on genetic distances between consensus sequences.

Supplementary Table 7. Conversion between subgraph distances in scaled deep-sequence phylogenies and genetic distances between consensus sequences

	substitutions per site scaled for HIV-1 *pol* gene								
subgraph distances in scaled deep-sequence phylogenies	0.01	0.015	0.02	0.025	0.03	0.035	0.04	0.045	0.05
genetic distance between consensus sequences	0.022	0.031	0.036	0.041	0.044	0.048	0.051	0.054	0.056

Between the two approaches, phylogenetic linkage classification agreed for 297/331 (89.7%) of couples (Supplementary Table 8). 26 couples were classified linked using consensus sequences but not linked using deep sequencing data. Of those, linkage in 5 couples was excluded because in the overall transmission network, linkage with other individuals was more likely based on our phylogenetic data; linkage in 3 couples was excluded because one of the two individuals had divergent subgraphs; and linkage in 16 couples was excluded because support for phylogenetic linkage was intermediate but not high enough, with $\hat{\lambda}_{ij}$ between 40-60%. This left 2 couples for whom we could not find an immediate explanation why consensus sequences indicated linkage but deep-sequence data did not. For all 8 couples that were classified linked using deep sequencing data but not linked using consensus sequences, at least one spouse had divergent subgraphs in at least 33% of deep-sequence phylogenies. Supplementary Figure 7 shows the couples for whom the two phylogenetic analyses disagreed, confirming that these couples were at the border of the classification thresholds that we used in our analysis. Supplementary Figure 8 illustrates subgraph distances, subgraph topologies and within-host subgraph divergence for 6 of the 8 couples that were classified as linked only when using deep sequencing data. Most couples (except B and F) had highly variable subgraph distances across the genome. These tended to coincide with genomic regions without divergent within-host subgraphs, suggesting that the closely related subgraph still present in their partner was either not sequenced, or lost in the quasi-species. In couples B and F, the closely related subgraphs were sequenced in both spouses, implying small subgraph distances across the sequenced genome but large genetic distance from consensus sequences.

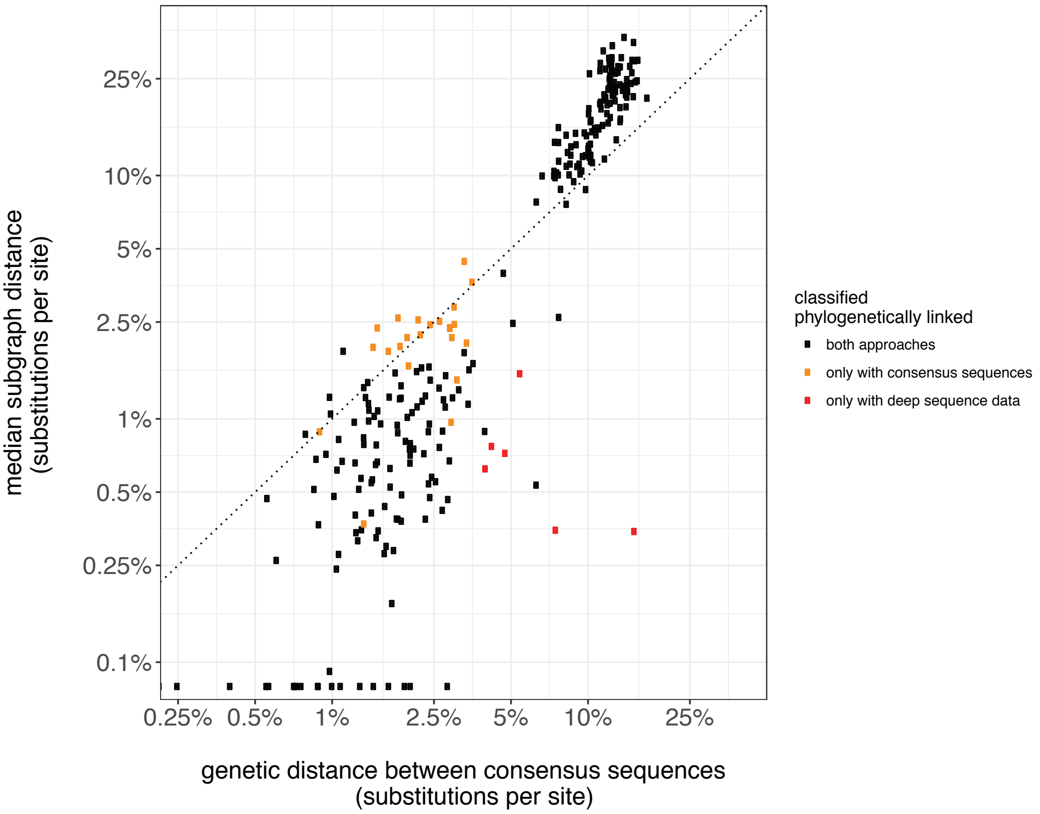
Supplementary Table 8. Comparison of phylogenetic linkage classification based on deep sequencing data and consensus sequences among 331 couples from Rakai District, Uganda.

Phylogenetic linkage classification among long-term sexual partners			**Phyloscanner probability of phylogenetic linkage** [*]			**Proportion of deep-sequence phylogenies with divergent subgraphs in at least one spouse** [**]		
			(mean and 95% empirical confidence interval across couples)			(mean and 95% empirical confidence interval across couples)		
	Deep sequence			Deep sequence			Deep sequence	
Consensus sequence	Not linked	Linked	Consensus sequence	Not linked	Linked	Consensus sequence	Not linked	Linked
Not linked	129	8	Not linked	3% [0%-53%]	77% [67%-90%]	Not linked	12% [0%-61%]	68% [38%-96%]
Linked	26	168	Linked	53% [32%-96%]	90% [67%-100%]	Linked	12% [0%-47%]	13% [0%-64%]

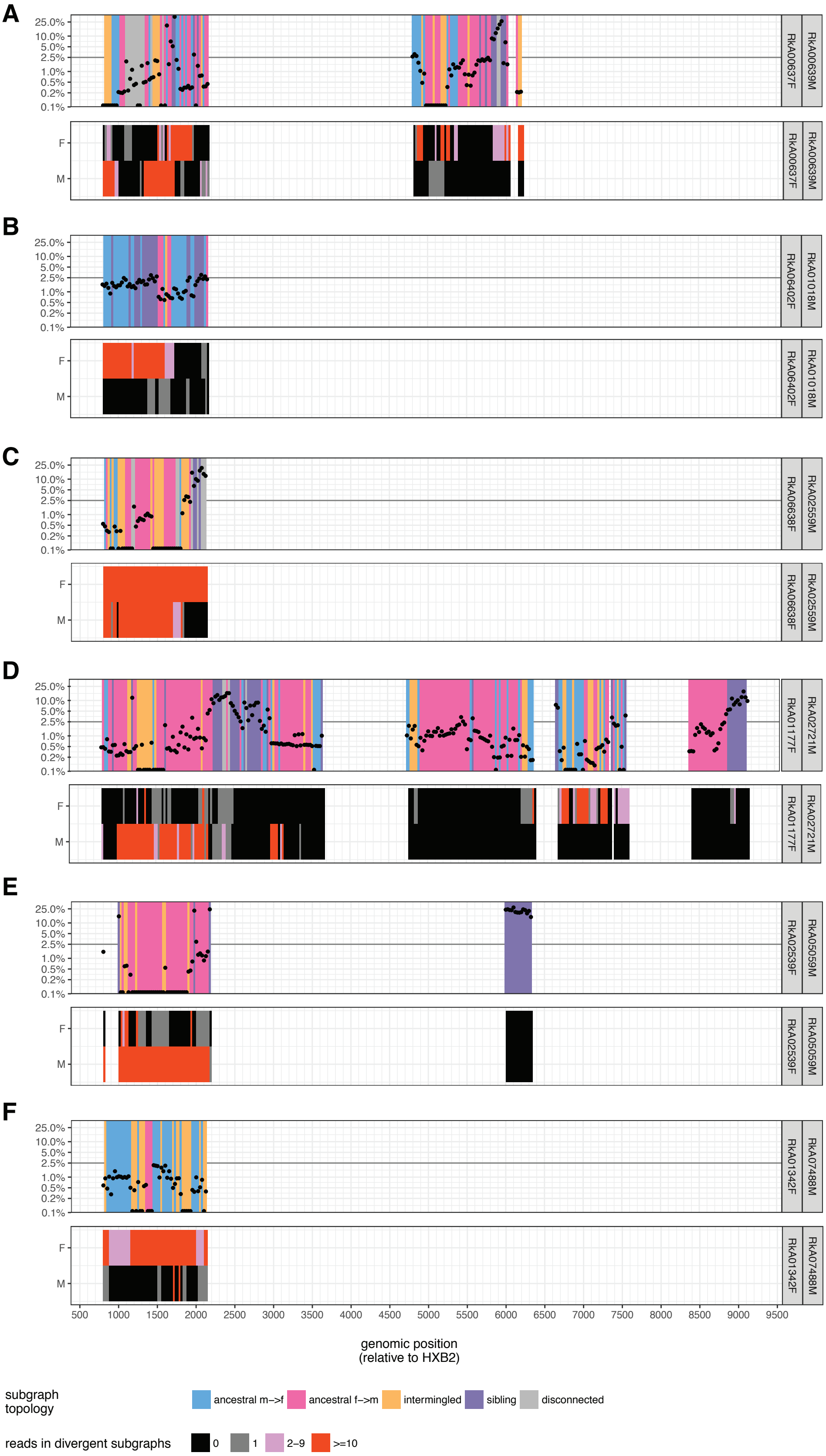
[*] Posterior mode estimate for being phylogenetically linked, see Methods.

[**] Divergent subgraphs in one individual were defined as subgraphs more than 0.05 substitutions per site apart from the individual's main subgraph, which contained at least 2 unique reads.

In summary, we found that phylogenetic linkage estimates from consensus sequences and deep-sequence reads were strongly concordant, in 297/331 (89.7%) of couples. For the majority of the remaining cases, we either found intermediate but not high support for linkage in deep-sequence phylogenies (16/34 (47.1%) of couples), or evidence of highly divergent subgraphs in one or both individuals (11/34 (32.4%) of couples), which typically implied high support for phylogenetic linkage based on deep-sequence reads.



Supplementary Figure 7. Couples for whom linkage classification based on consensus and deep-sequence analysis disagreed. The dotted line shows y=x.



Supplementary Figure 8. Subgraph distance, topology and divergence among couples that were phylogenetically linked using deep sequencing data, but not linked using consensus sequences.

Inferring HIV-1 transmission networks and sources of epidemic spread in Africa with deep-sequence phylogenetic analysis

Supplementary Note 3. Error rates in inferring phylogenetic linkage from deep-sequence data in the population-based sample

HIV-1 is predominantly sexually transmitted, and extremely rarely sexually transmitted between women[11]. This allowed us to characterize error rates in phylogenetic inference of direct transmission between males and females in the population sample.

Denote the number of phylogenetically linked female-female pairs by L_{ff}. For S_f sequenced females and S_m sequenced males, there are S_f*(S_f-1)/2 pairs of sequenced females, and the probability of inferring a phylogenetically linked female-female pair is

$$\frac{L_{ff}}{S_f * (S_f - 1)/2}.$$

If we assume that the probability of incorrectly inferring a phylogenetically linked male-female pair is the same as the above probability of inferring a phylogenetically linked female-female pair, the number of linked male-female pairs between whom transmission did not occur can thus be estimated by

$$\hat{F}_{mf}^{C} = \frac{L_{ff}}{S_f * (S_f - 1)/2} * S_f * S_m,$$

Suppose that L_{mf} male-female pairs were inferred to be phylogenetically linked. An estimate of the false discovery rate is

$$\hat{\rho}_{mf}^{C} = \frac{\hat{F}_{mf}^{C}}{L_{mf}}.$$

This probably overestimates the true false discovery rate because two individuals would have to be missing from the sequence sample to incorrectly infer phylogenetic linkage in a male-female pair, where only one male would have to be missing from the sequence sample to incorrectly infer phylogenetic linkage in a female-female pair. Supplementary Table 9 lists estimates of $\hat{\rho}_{mf}^{C}$ for a range of distance thresholds.

Supplementary Table 9. Estimated error rates in inferring direct transmission from deep sequencing data in Rakai, Uganda.

	Threshold on subgraph distances to define phylogenetically linked individuals in combination with subgraph topology (in substitutions per site)					
	0.01	**0.015**	**0.02**	**0.025**	**0.03**	**0.035**
Phylogenetically linked female-female pairs	25	43	61	80	99	117
Phylogenetically linked male-female pairs between whom transmission did not occur (estimated)	42	72	102	133	165	195
Phylogenetically linked male-female pairs	198	274	336	376	423	452
False discovery rate *	21%	26.10%	30.20%	35.40%	39%	43.10%

* Assuming equal false positive rates among female-female pairs and male-female pairs, see text.

Inferring HIV-1 transmission networks and sources of epidemic spread in Africa with deep-sequence phylogenetic analysis

Supplementary Note 4: Limitations in inferring the direction of transmission from deep-sequence data

We investigated why the direction of transmission was incorrectly inferred with the phyloscanner method in the nine cases reported in table 2. Given the small number of pairs for whom the direction of transmission was inconsistent with clinical data, this analysis remains largely descriptive. The validation analysis was based on phylogenetically linked pairs of individuals with clinical evidence for the direction of transmission based on seroconversion dates and CD4 cell count measurements, and for whom phylogenetical linkage was inferred with high support. Prior to validation, the selection criteria were specified as follows:

- **Seroconversion data**. Partner 1 tested negative while partner 2 tested positive at or before the same time. Subsequently, partner 1 tested positive. Assuming that transmission occurred between the two individuals, seroconversion data indicates transmission from partner 2 to partner 1.
- **CD4 data**. Partner 1 had first CD4 measurement >800 cells per mm3 within two years of diagnosis, while partner 2 had a CD4 measurement <400 cells per mm3 within two years of diagnosis of partner 1. Assuming that transmission occurred between the two individuals, CD4 data indicates transmission from partner 2 to partner 1.

Detailed epidemiological and phylogenetic characterization of the validation data set.

Detailed timelines on seroconversion dates, CD4 counts, sequencing dates and phyloscanner output for the 55 phylogenetically linked pairs in the validation panel are shown in Supplementary Figures 9–12.

Post-hoc evaluation of the selection criteria by which the validation data set was formed.

We examined potential limitations in these selection criteria. For 36 phylogenetically linked pairs, data on the direction of transmission was available from the seroconversion history, and the direction of transmission could be inferred with phyloscanner in 31 pairs. In 16/31 of pairs, the time between the first positive date of the (epidemiologically inferred) source case and the (epidemiologically inferred) recipient was less than 1 month. Considering limited sensitivity of HIV-1 tests in early infection, it was thus possible (though not very likely) that infection could have occurred the other way round in these pairs. However, the odds ratio for incorrect phylogenetic inference among pairs with very small differences in first positive and last negative dates versus those with larger differences was

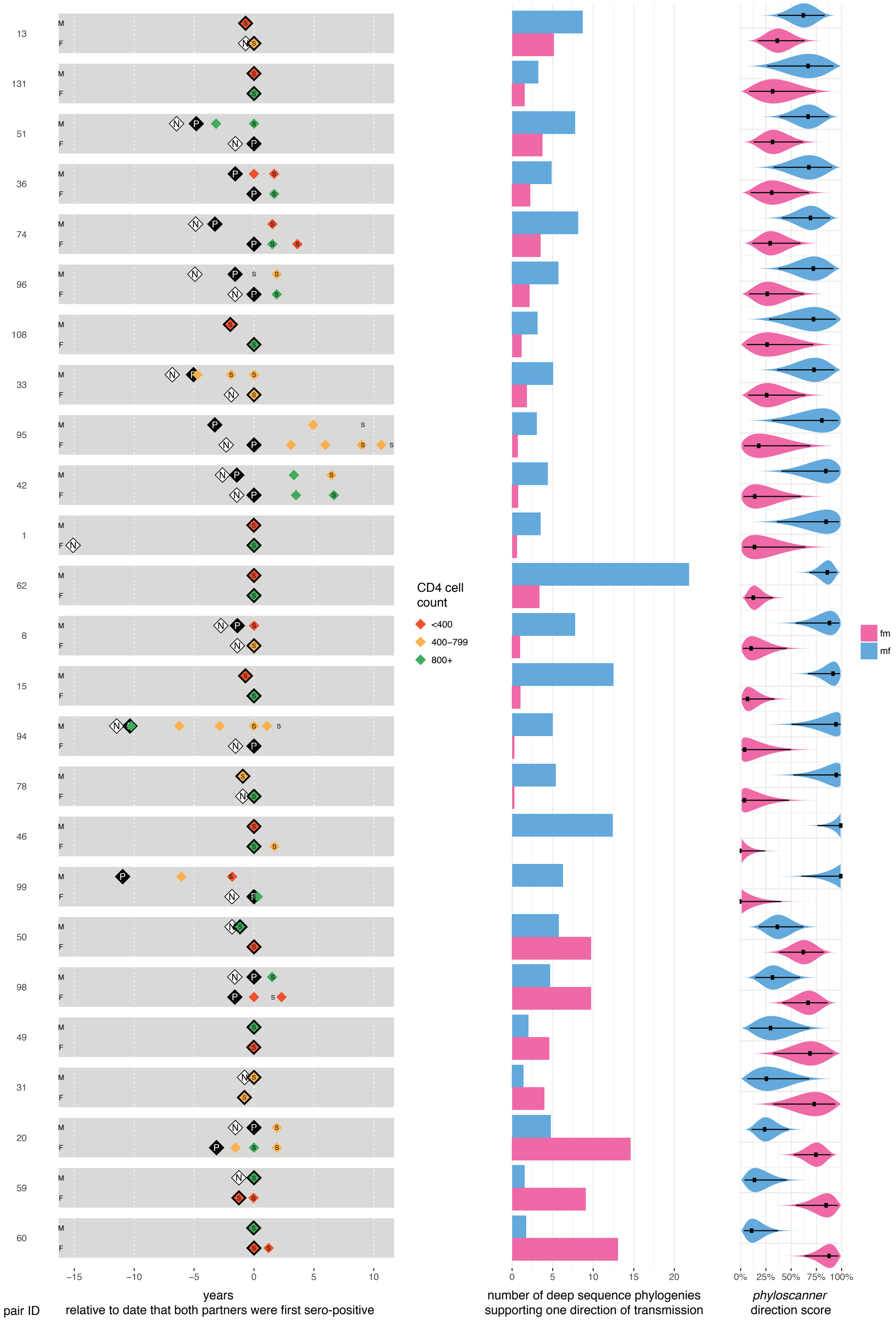
$$(2/14)/(2/13) = 0.93$$

and not significant (Fisher exact test).

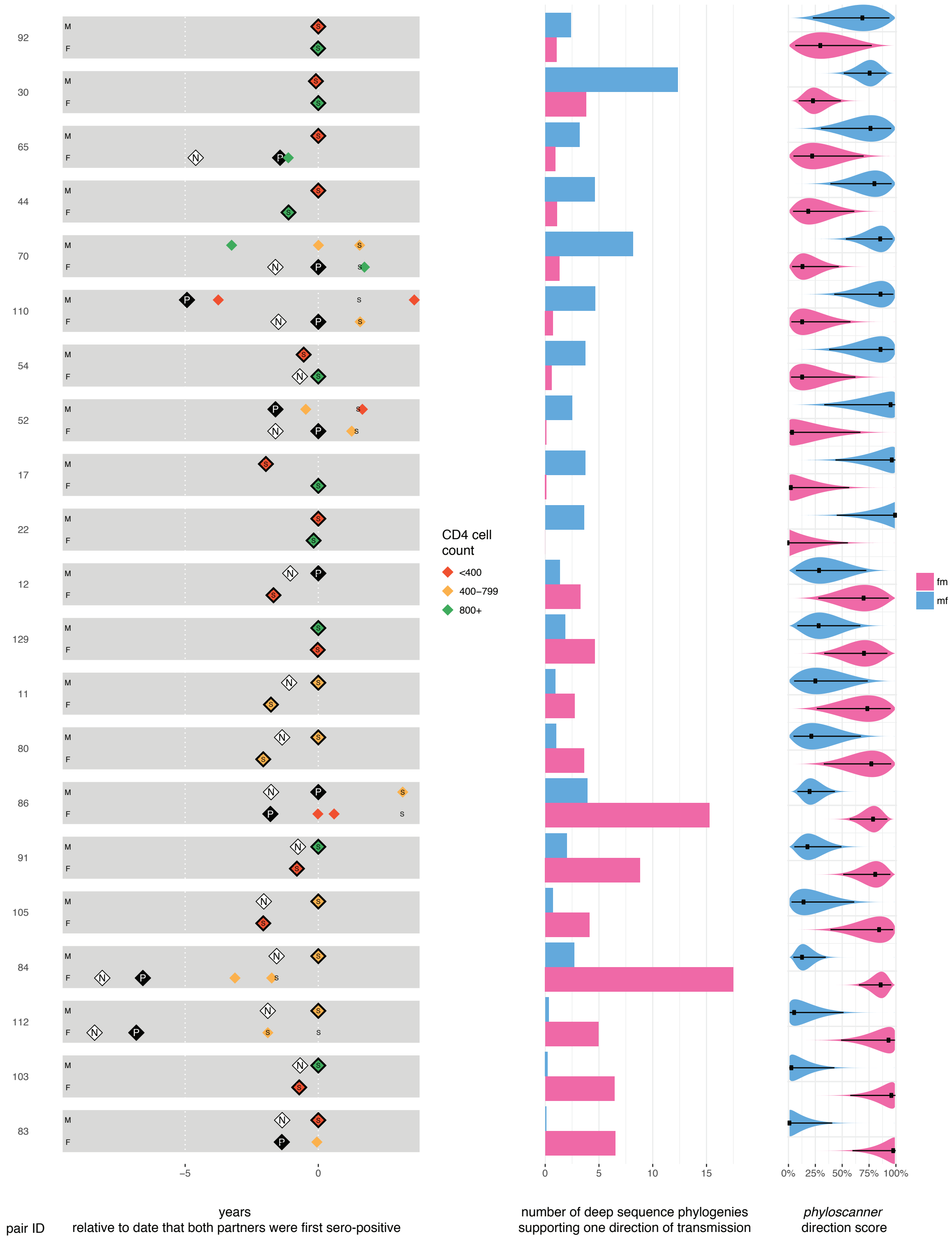
For 35 phylogenetically linked pairs, data on the direction of transmission was available from the CD4 count history, and the direction of transmission could be inferred with phyloscanner in 24 pairs. In 5 pairs, the (epidemiologically inferred) source case had the selected CD4 measurement more than 1 year after the (epidemiologically inferred) recipient. In these pairs, the substantially lower CD4 cell count in the (epidemiologically inferred) source case could have arisen over the difference in measurement times, and it was thus possible that infection could have occurred the other way round. The odds ratio for incorrect phylogenetic inference among pairs with very large negative differences in CD4 measurement dates versus those with larger differences was

$$(2/3)/(3/16) = 3.56,$$

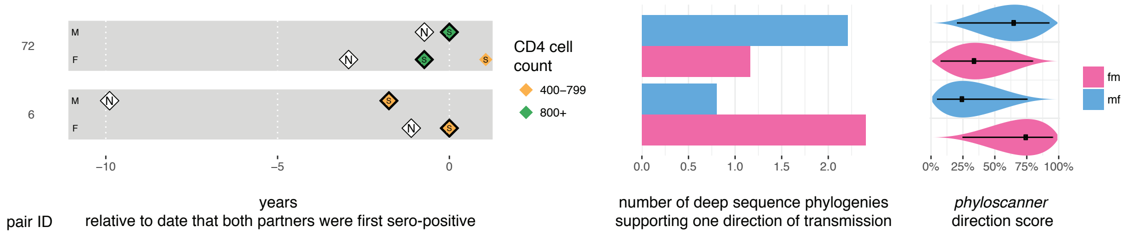
which was again not statistically significant (Fisher exact test, p-value 0.27). However, the magnitude of the odds ratio suggests that it may have been more appropriate to consider pairs with CD4 measurement dates within 1 year of diagnosis as basis for defining the validation data set. Pairs labelled 17, 18, 36, 44, 50, 65, 90, 108 in Supplementary Figures 9–12 did not meet these more stringent selection criteria. The true direction of transmission in pairs 18, 90 could be consistent with phyloscanner inference.



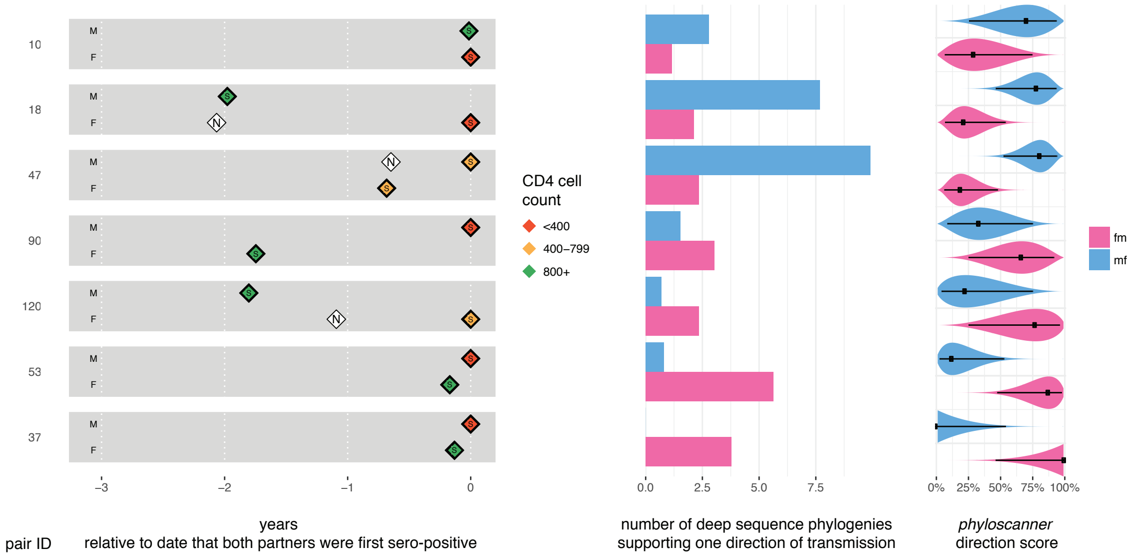
Supplementary Figure 9. Phylogenetically linked couples for whom the phylogenetically inferred direction of transmission was consistent with clinical data. Please see text for details.



Supplementary Figure 10. Phylogenetically linked casual pairs for whom the phylogenetically inferred direction of transmission was consistent with clinical data. Please see text for details.



Supplementary Figure 11. Phylogenetically linked couples for whom the phylogenetically inferred direction of transmission was not consistent with clinical data. Please see text for details.



Supplementary Figure 12. Phylogenetically linked casual pairs for whom the phylogenetically inferred direction of transmission was not consistent with clinical data. Please see text for details.

Potential impact of sequence sampling times on phylogenetic inference into the direction of transmission.

Next, we examined wether particular patterns in sequence sampling times were associated with greater failure to correctly determine the direction of transmission. We hypothesized that true recipients who were sampled earlier might be more likely to appear as source in reconstructed deep-sequence phylogenies. The odds for incorrect phylogenetic inference of the source case were higher when the person, who was the recipient based on epidemiological data, was diagnosed first

$$(5/6)/(4/40) = 0.13,$$

and this was statistically significant (Fisher exact test, p-value 0.011). However, for the large majority individuals in the validation data set, sequencing was performed on the first positive sample (83 of 110). We therefore also considered the difference in times at which the blood sample for sequencing was taken. The odds for incorrect phylogenetic inference of the source case were again higher when the person, who was the recipient based on epidemiological data, was sequenced at an earlier date

$$(5/12)/(4/34) = 0.29,$$

though this was not statistically significant (Fisher exact test, p-value 0.116).

Potential shortcomings of the phyloscanner method on phylogenetic inference into the direction of transmission.

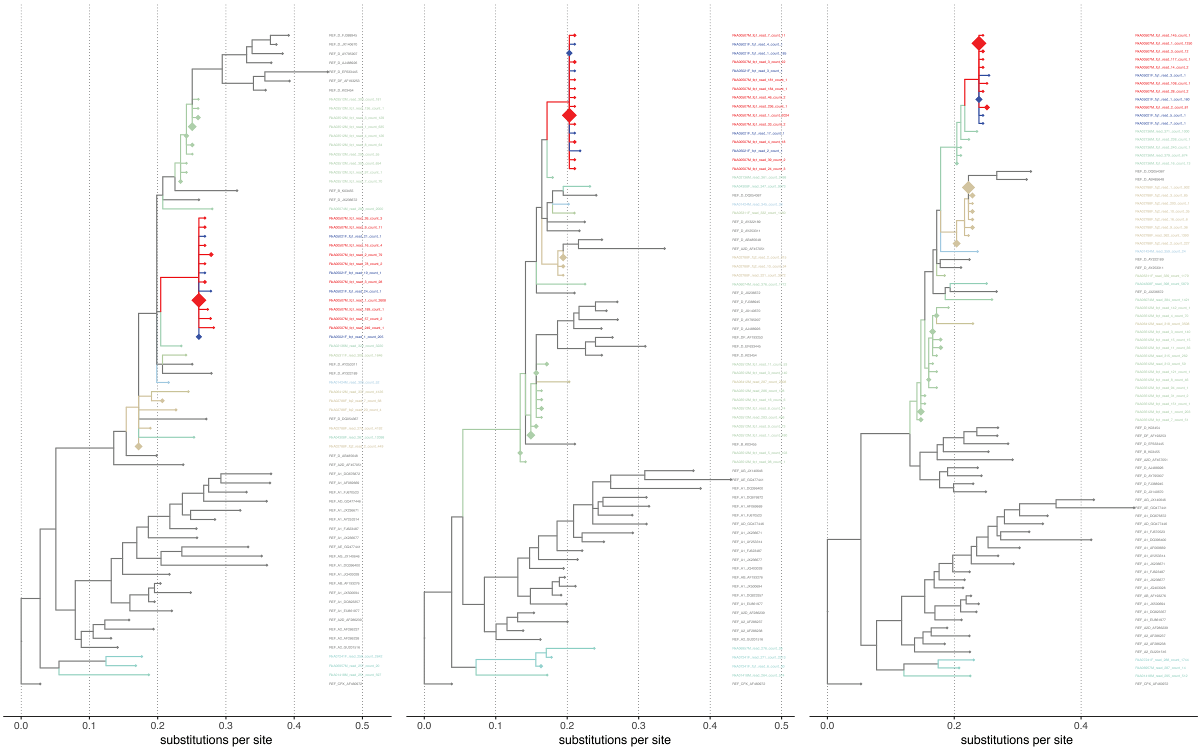
We further examined the deep-sequence phylogenies of the 9 phylogenetically linked pairs for whom the phylogenetically inferred direction of transmission was inconsistent with clinical data. In 10%-20% of those phylogenies, we found that reads from both partners which were essentially identical (subgraph distances below 10^{-6} substitutions per site) and basal in the corresponding subgraphs of both individuals (Supplementary Figure 13). In these cases, inferred ancestry should be in either direction with equal probability. However, due to consistently higher copy number of those reads in one individual, preference was systematically given for ancestral subgraph topologies in one of the two possible directions. This is likely a technical limitation that affected our inferences.

Summary

Supplementary Table 10 summarizes our investigations, indicating that potential reasons for why phylogenetic inference into the directon of transmission was inconsistent with clinical data could be isolated in 8/9 pairs.

Supplementary Table 10. Potential reasons on failure to infer direction of transmission from deep-sequence data.

Pair identifier	Known to have long-term sexual contact	Weak clinical indicator of direction of transmission	Epidemiologically identified recipient sampled before source	Technical limitations in inferring ancestry	Further comments
6	Yes	No	No	Yes	--
10	No	No	yes, a few days	No	No explanation on inconsistent phylogenetic inference
18	No	Yes	yes, 2 years	No	--
37	No	No	yes, two months		Deep sequencing relatively poor compared to most other samples
47	No	No	No	Yes	--
53	No	No	yes, two months	Yes	--
72	Yes	No	No	Yes	--
90	No	Yes	yes, two years	No	--
120	No	No	No	Yes	--



Supplementary Figure 13. Limitations in inferring ancestry between subgraphs with the phyloscanner method. Three consecutive deep-sequence phylogenies are shown, with subgraphs from the male partner (red) and female partner (blue) highlighted. Reads from both partners were basal in the corresponding subgraphs and essentially identical, suggesting that ancestry between the two individuals cannot be established in these phylogenies.

Supplementary References

1. Darriba, D., Taboada, G.L., Doallo, R. & Posada, D. jModelTest 2: more models, new heuristics and parallel computing. *Nat Methods* **9**, 772 (2012).
2. Kuiken, C. *et al*. HIV Sequence Compendium 2012. (ed. Theoretical Biology and Biophysics Group, L.A.N.L.) (NM, LA-UR 12-24653, 2012).
3. Stamatakis, A. RAxML version 8: a tool for phylogenetic analysis and post-analysis of large phylogenies. *Bioinformatics* **30**, 1312-3 (2014).
4. Gall, A. *et al*. Universal amplification, next-generation sequencing, and assembly of HIV-1 genomes. *J Clin Microbiol* **50**, 3838-44 (2012).
5. Wymant, C. *et al*. Easy and Accurate Reconstruction of Whole HIV Genomes from Short-Read Sequence Data. *bioRxiv* (2016).
6. Ratmann, O. *et al*. HIV-1 full-genome phylogenetics of generalized epidemics in sub-Saharan Africa: impact of missing nucleotide characters in next-generation sequences. *AIDS Res Hum Retroviruses* (2017).
7. Yang, Z. & Rannala, B. Molecular phylogenetics: principles and practice. *Nat Rev Genet* **13**, 303-14 (2012).
8. Wymant, C. *et al*. PHYLOSCANNER: Inferring Transmission from Within- and Between-Host Pathogen Genetic Diversity. *Mol Biol Evol* (2017).
9. Tamura, K. & Nei, M. Estimation of the number of nucleotide substitutions in the control region of mitochondrial DNA in humans and chimpanzees. *Mol Biol Evol* **10**, 512-26 (1993).
10. Paradis, E., Claude, J. & Strimmer, K. APE: Analyses of Phylogenetics and Evolution in R language. *Bioinformatics* **20**, 289-90 (2004).
11. Chan, S.K. *et al*. Likely female-to-female sexual transmission of HIV--Texas, 2012. *MMWR Morb Mortal Wkly Rep* **63**, 209-12 (2014).