AIDS, Publish Ahead of Print

DOI: 10.1097/QAD.0000000000002105

# Caution is Needed in Interpreting HIV Transmission Chains by Ultra-Deep Sequencing

**Short title:** Analysis of HIV Transmission Chains

Eve TODESCO [1], Marc WIRDEN [1], Ruxandra CALIN [2], Anne SIMON [3], Sophie SAYON [1], Francis BARIN [4], Christine KATLAMA [2], Vincent CALVEZ [1], Anne-Geneviève MARCELIN [1], Stéphane HUÉ[5]

[1] Sorbonne Université, INSERM, Institut Pierre Louis d'Epidémiologie et de Santé Publique (iPLESP), AP-HP, Hôpital Pitié-Salpêtrière, Laboratoire de virologie, F-75013 Paris, France

[2] Sorbonne Université, INSERM, Institut Pierre Louis d'Epidémiologie et de Santé Publique (iPLESP), AP-HP, Hôpital Pitié-Salpêtrière, Department of Infectious Diseases, F-75013 Paris, France.

[3] Department of Internal Medicine, Hôpital Pitié-Salpêtrière, AP-HP, F75013, Paris, France;

[4] Centre National de Référence du VIH, laboratoire de Virologie, CHRU de Tours, Inserm U1259, Université François Rabelais, Tours, France;

[5] Department of Infectious Disease Epidemiology, London School of Hygiene and Tropical Medicine, London, UK.

**Corresponding author:** Eve Todesco, Department of Virology, Bât CERVI, Hôpital Pitié-Salpêtrière, 83 Bd de l'Hôpital, 75013 Paris, France. Email: eve.todesco@aphp.fr. Fax: +33 1 42177411. Phone: +33 1 42177401.

**Conflicts of Interest :** None.

**Source of Funding:** This work was supported by the Agence Nationale de Recherches sur le SIDA et les hépatites virales (ANRS).

**Abstract**

**Objectives**

Molecular epidemiology is applied to various aspects of HIV transmission analyses. With ultra-deep sequencing (UDS), in-depth characterisation of transmission episodes involving minority variants are permitted. We explored HIV-1 epidemiological linkage and evaluated characteristics of transmission dynamics and transmitted drug resistance (TDR) detection through the added value of UDS.

**Design**

HIV *pol* gene fragments were sequenced by UDS and Sanger sequencing (SS) on samples of 70 HIV-1 infected, treatment-naïve recently diagnosed men having sex with men (MSM).

**Methods**

Pairwise genetic distances and maximum likelihood phylogenies were computed. Transmission events were identified as clades with branch support ≥70% and intra-clade genetic difference <4.5%. TDR mutations were recognised from the TDR consensus list. Transmission directionality, directness and inoculum size were inferred from tree topologies.

**Results**

Both datasets concurred in the identification of 7 transmission pairs and 1 cluster of 3 patients. With UDS, direction of transmission was inferred in 4/8 chains. Evidence for multiple founder viruses was found in 2/8 chains. No transmission of minority resistant variants was evidenced. TDR mutations prevalence in protease and reverse transcriptase fragments was 4.3% with SS and 18.6% with UDS.

**Conclusions**

While SS and UDS identified the same transmission chains, UDS provided additional information on founder viruses, direction of transmission and levels of TDR. Nevertheless,

topology of clusters was not always consistent across gene fragments, calling for a cautious interpretation of the data. Moreover, unobserved intermediary links cannot be excluded. Phylogenetic analysis use as a forensic technique for HIV transmission investigations is risky.

**Keywords:** HIV-1; HIV infection; Phylogeny; Cluster analysis; Homosexuality, Male; Drug Resistance, Viral; High-Throughput Nucleotide Sequencing; Minority resistant variants.

## Introduction

Molecular epidemiology is a powerful approach applied to various aspects of HIV transmission analyses, such as the characterisation of patterns of infection within risk or ethnic groups [1–4], the study of transmitted antiretroviral resistance [5–7], the quantification of misreported routes of infection [8,9] or, most recently, the assessment of treatment strategies efficiency in intervention-based trials [10]. This approach implies the phylogenetic identification of viral isolates more genetically similar to each other than expected by chance in an infected population, thus establishing epidemiological linkage between individuals the related viruses are sampled from.

HIV molecular epidemiology predominantly relies on 'convenience' data routinely generated for antiretroviral resistance monitoring, where consensus HIV *pol* sequences are generated by PCR-based Sanger sequencing (SS) prior to treatment initiation [11]. Although this genetic fragment contains sufficient phylogenetic signal to identify transmission clusters [12], it has the shortcomings of a consensus sequence. It does not reflect intra-host genetic diversity of the HIV population, precluding in-depths analyses of the role of minority variants at transmission. It is also notoriously challenging to infer direction of transmission from consensus sequence data, unless sophisticated statistical inference is applied [13,14]. Finally, the incomplete sampling of cases inherent to this data substantially limits attempts to infer directness of transmission within a phylogenetic pair or cluster, as the possibility of unobserved intermediate transmissions is difficult to exclude when individuals are represented by a single viral sequence.

The development of ultra-deep sequencing platforms (UDS) has imposed new data standards. While SS detects viruses accounting for more than 15-20% of the infecting population, UDS technologies permits the detection of minority variants representing <1% of the viral population. The sensitivity and depth of UDS has therefore the potential to unravel transmission episodes involving variants existing at low frequency [15]. Direction of

transmission is also likely to be inferable from the information contained within UDS data, where the viral population of the recipient is expected to be nested within that of the transmitter [16]. Most recently, novel analytical tools were developed to that effect [17,18].

UDS also provides valuable insight into transmitted drug resistance (TDR). Through its increased sensitivity, UDS allows the detection of HIV-1 minority resistant variants (MRV), which may be a cause of virological failure when transmitted, as shown for first line regimens based on first generation NNRTIs [19–21]. Understanding the frequency and dynamics of MRV transmission is therefore of the upmost importance and UDS data provides invaluable information on the matter.

We conducted a phylogenetic transmission study using HIV *pol* gene sequences sampled from men who have sex with men (MSM). We compared the clustering pattern of sequences generated by SS and UDS, and assessed the ability of the latter to unravel epidemiological features such as direction and directness of infection, multiplicity of founder viruses and frequency of minority resistant variants transmission.

**Methods**

*Study cohort.* The study enrolled 70 treatment-naïve, HIV-positive MSM diagnosed between January 2012 and July 2013. Patients were followed at the Department of Infectious Diseases of the hospital Pitié-Salpêtrière (Paris, France). Clinical information was extracted from the hospital electronic database or medical records and was anonymised prior to analysis. This information included age of the participant, time of HIV diagnosis, time of blood sampling, time of last negative HIV test (when applicable), estimated time of infection (as reported by the patient), viral load and CD4+ T cell count at sampling. All subjects have signed the ethics board-approved informed consent form of the HIV Nadis® electronic patient record database.

*Immunoassay testing.* In order to identify recently HIV-1 infected patients (≤ 6 months), a single, indirect enzyme-linked immunosorbent assay was used to quantify antibodies

directed against two HIV-1 antigens (i.e. consensus peptides of the immunodominant epitope of gp41 and consensus V3 peptides), as described previously [22].

*Viral gene sequencing and sequence data.* HIV *pol* gene fragments were sequenced from the first HIV-positive plasma sample available for each patient, using both (i) SS and (ii) UDS.

(i) SS: full length protease (PR; 297bp; HXB2 position 2253 - 2549) and partial reverse transcriptase (RT; 771bp; position 2550 - 3320) consensus sequences were obtained, according to the routine Agence Nationale de Recherche sur le Sida et les hépatites virales (ANRS) protocol [23]. Sequences found to have a mixture of wild type and mutant residues at a given codon position were considered to have the mutant residue at that position. Subtype determination was performed using the BLAST-based SmartGene HIV Module (SmartGene, Zug, Switzerland).

(ii) UDS: Steps until pyrosequencing on GS Junior (Roche 454® Life Sciences, Branford, CT, United States) were previously described [24]. Primers are available on *hiv.frenchresistance.org*. Pyrosequencing was performed according to the manufacturer recommendations [25]. Due to the length of the gene, reverse transcriptase pyrosequencing was performed in 2 fragments: first fragment RT1 (HXB2 RT amino acid position 17-140), second fragment RT2 (HXB2 RT amino acid position 133-247). GS Amplicon Variant Analyzer (Roche 454® Life Sciences, Branford, CT, United States) was used to analyze the UDS results.

*Drug resistance mutation identification.* TDR mutations were identified from the 2009 restrictive consensus list for surveillance of transmitted HIV-1 drug resistance [26]. Resistance-associated mutations were also identified using the ANRS algorithm V27 (*http://www.hivfrenchresistance.org*), which considers more mutations than the TDR consensus list.

*Transmission cluster identification.* Putative transmission events between the studied patients were identified by phylogenetic means, using UDS data, for each gene fragment (i.e. PR, RT1 and RT2). Intra-host variant sequences were manually aligned using AliView v.1.18 [27] after removal of duplicates. A maximum likelihood (ML) tree was reconstructed under the general time reversible model of nucleotide substitution with gamma-distributed rate heterogeneity (GTR+G), using FastTree v1.2.7 [28]. Branch support was estimated using the Shimodaira-Hasegawa-like likelihood ratio test (SH-aLRT; [29]) implemented in FastTree. Putative transmission clusters were extracted from the PR, RT1 and RT2 phylogenies with ClusterPicker [30]. The criteria used for cluster identification were (i) ≥ 2 monophyletic sequences, (ii) a branch support ≥ 0.70 and (iii) a maximum intra-cluster pairwise genetic distance inferior to a given threshold. Distance thresholds of 1.5%, 2.0%, 2.5%, 3%, 3.5% and 4.5% nucleotide differences were used. Gaps were treated as missing characters. Patients for which viral sequences clustered in all phylogenies (PR, RT1 and RT2) were considered as epidemiologically linked.

Putative transmission clusters were also identified from the SS phylogenies using the same criteria as those applied to the UDS data, for each gene fragment (i.e. PR and RT), as a point of comparison. Publically available control sequences were extracted from GenBank by BLAST search [31], retaining the 10 closest matches to each patient sequence. Additional control sequences from local HIV positive individuals (n = 29) were also added to the dataset. After removal of duplicates, alignments of 643 (PR) and 370 (RT) sequences were manually generated using the sequence editor AliView v.1.18. ML trees were reconstructed using FastTree under the GTR+G model.

Each procedure was repeated after removing codon positions associated with major antiretroviral resistance [26] from the original alignments, in order to check the effect of treatment-induced convergent evolution.

*Pairwise genetic distances calculations.* Intra-cluster pairwise nucleotide differences were calculated for each transmission cluster using the package HyPhy v2.1.2 [32], under the TN93 model of evolution (determined as the best fitting model with the model testing algorithm included in HyPhy), for both SS and UDS sequence data.

*Inference of directionality and directness of transmission.* Directionality (i.e. transmitter/recipient relationship, or 'who infected whom') and directness (i.e. no evidence of an unobserved intermediate transmitter) of transmission were inferred within each UDS cluster as follows.

First, the phylogeny of each individual cluster was rebuilt and rooted against an outgroup formed of sequences from the two patients most closely related to the cluster of interest in the initial ML phylogenies. Direction of transmission between the linked individuals was then inferred from the topological structure of the UDS clusters and the cladistics relationship between the two sequenced populations within them, as formulated by Romero-Severson *et al.* [16]. When both viral populations were monophyletic within a cluster (monophyletic-monophyletic topology, MM; **Figure 1A**), the direction of infection was deemed as equivocal. When one population was paraphyletic and the other monophyletic (paraphyletic-monophyletic topology, PM; **Figure 1B**), the paraphyletic population was assigned to the transmitter. When both populations were paraphyletic (paraphyletic-paraphyletic topology, PP; **Figures 1C & D**), the population whose most recent common ancestor was closest to the ancestral node of the clade was assigned to the transmitter. Directionality was validated and considered unequivocal when the same direction of transmission was observed in all UDS genetic fragments (i.e. PR, RT1 and RT2).

Directness of transmission was inferred for clusters exhibiting a PP topology only (**Figures 1C & D**), under the assumption that intermixing of transmitter and recipient viral populations is indicative of a recent and direct transmission event. MM and PM clusters were deemed equivocal to that respect (**Figures 1A & B**).

*Multiplicity of the recipient's founder viruses.* For each direct transmission pair, the size of the inoculum was inferred from the topology of the recipient's clade(s) in the tree. In PP clusters, the inoculum was assumed to derive from more than one virus and the minimum number of founding strains was calculated as the number of internal nodes linking viral sequences from both the transmitter and recipient (**Figures 1C & D**). In PM and MM clusters, the size of the inoculum was deemed equivocal since directness of transmission could not be established with certainty (**Figures 1A & B**).

**Results**

*Patients and virus' characteristics.* The median age of the participants was 36 years (range 22-65). Median viral load at sampling was 4.9 $\log_{10}$ HIV RNA copies/mL (IQR=4.4-5.4) and median CD4 cell count was 498/$mm^3$ (IQR=347-585). The median time between the HIV-1 diagnosis and date of the sample used for genotyping was 11 days (IQR: 2.25-29). No multiple HIV infection was detected in the patients. Antibodies quantification was performed on 69/70 serum samples and identified 28 recently HIV-1 infected individuals (infection ≤ 6 months prior to diagnosis). A total of 42/70 patients (60%) were infected by HIV-1 subtype B. Non-B infections included CRF02_AG (15 patients; 21%), subtype C (4 patients; 6%), subtype F2 (3 patients; 4%), subtype F1 (2 patients; 3%), CRF01_AE (2 patients; 3%) and CRF07_BC (2 patients; 3%).

*Ultra-deep sequencing.* An average of 2,633 reads per nucleotide position was amplified. The average error rates in controls (cellular clone 8E5) were 0.0032 and 0.0012 substitutions per base for PR and RT, respectively, allowing an accurate detection of variants down to 1% [33].

*Transmission clusters identification.* The putative transmission clusters identified with UDS data, by genetic fragment, are shown in **Table 1** and **Supplementary Figure 1,** http://links.lww.com/QAD/B408. A total of 38 patients (forming 19 clusters) were linked on the basis of at least one gene fragment. However only 8 clusters, 7 pairs and one triplet, were

identified by phylogenetic clustering of all tested fragments (highlighted in bold in **Table 1**). These 8 clusters were also consistently identified within the SS phylogenies and remained intact after exclusion of drug resistance associated mutations, rejecting the hypothesis of clustering artefacts derived from treatment-induced convergent evolution (data not shown).

These 8 clusters were deemed to represent genuine transmission chains and were included in the subsequent analyses. The subtype of the clustered sequences was B (3/8), C (2/8), CRF_AG (2/8) and CRF07_BC (1/8). Two of the eight transmission clusters contained sequences from the background controls in the SS PR phylogenies (clusters [31,70] and [52, 60] and in the SS RT phylogenies (cluster [31,70]). For cluster [31/70], the origin of the interspersed control sequence matched the country of origin of the patients, adding further credit to the linkage. Sampling interval within the putative transmission clusters ranged from 0 to 355 days (median: 165 days; **Table 1**).

*Intra-cluster pairwise genetic distances.* As expected, UDS intra-cluster pairwise genetic distances in PR and RT were higher than SS pairwise distances (**Supplementary Figure 2,** http://links.lww.com/QAD/B408). However, none of the maximum UDS pairwise distances for a given cluster exceeded 4.5% nucleotide differences.

*Directionality, directness and inoculum size inference.* An example of the different topologies (i.e. MM, PM or PP) observed amongst the UDS transmission clusters is shown in **Figure 2**. Cluster topologies were coherent across all fragments for 4/8 putative transmission clusters, being either PP (clusters [05,41] and [19,62]), PM (clusters [31,62]) or MM (cluster [52,60]) (highlighted in grey **Table 2)**. Three clusters had a PP and PM topology in PR and RT respectively ([08,44], [17,30] and [34,46]). The last cluster included three patients, resulting in more complex clustering patterns (see below). In all PM trees, the branch leading to the monophyletic recipient population was deemed robust, with branch supports ranging from 78% to 100%. No correlation between the clusters' topology and sampling interval between the transmitter/recipient viral populations was observed (data not shown).

Direction of transmission could be inferred in 4/8 clusters (50%; clusters [5,41], [8,44], [17,30] and [31,70]; **Table 2).** In those clusters, the topology was consistent across all fragments. Clusters for which direction was equivocal included a cluster consistently exhibiting a MM topology (cluster [52,60]) or those for which differences in tree topology suggested conflicting scenarios (clusters [19,62] and [34,46]). Direction of transmission could be partially inferred within cluster [25, 33, 45], which comprised three linked individuals. In this cluster, the viral sequences derived from patient 45 were paraphyletic and basal to the cluster in all phylogenies. Sequences from patients 25 and 33 formed distinct sub-clusters, either monophyletic or paraphyletic, branching off the patient 45's viral population (**Supplementary Figure 1,** http://links.lww.com/QAD/B408). These patterns suggest that patient 45 may be at the origin of patients 25 and 33's infections, but whether the latter were independently infected by 45 or infected each other remains unclear. In addition, immunoassay and sample collection dates confirmed that patient 33 was infected later than patient 45, giving credence to phylogeny results.

Directness of transmission and multiplicity of the founder viruses could be inferred in two clusters only, i.e. clusters exhibiting a PP topology in all fragments (pairs [05,41] and [19,62]; **Table 2**). The observed number of founder viruses varied across genetic fragments for a same cluster, ranging from 2 (in RT2) to 8 (in RT1) for cluster [05,41], and from 3 (in RT2) and 10 (in RT1) for cluster [19,62] (**Supplementary Figure 1,** http://links.lww.com/QAD/B408)**.**

*Transmitted drug resistance.* The overall prevalence of TDR mutations in the PR and RT fragments was 4.3% (3/70; 95%CI=0.0%-9.1%) with SS and 18.6% (13/70; 55%CI=9.4%-27.7%) with UDS. Total concordance was found between the two sequencing methods for mutations detected at a frequency >20% with UDS. Although no TDR mutation was detected in the 17 patients involved in transmission chains by SS, UDS detected TDR mutations in 35.3% of them (n=6; 95%CI=12.6%-58.0%). These mutations were present at low frequencies (1.1 to 7.0%) and none of shared within a transmission cluster (**Table 3**). Even

the substitution M46I in PR, detected in patient 45 at a frequency of 7%, was not observed in the plasma of the recently HIV-1 infected patient 33.

When using the ANRS algorithm, polymorphisms to protease inhibitors were found in all transmission chains, on both SS and UDS data. Up to 81.8% (27/33) of the polymorphisms had reached fixation (i.e. were present in all UDS variants in a given individual) and were present in all partners of a given transmission chain (**Table 3**).

**Discussion**

The reconstruction of transmission chains is the bedrock on which many epidemiological interventions are built. For this reason, considerable efforts are invested in the generation of deep sequencing viral data from infected populations, together with the development of phylogenetic frameworks to analyse them [17][34–36]. UDS data provides a snapshot of intra-host viral diversity to unprecedented levels, and we show that such information significantly improves the resolution of HIV transmission studies compared to routinely generated HIV sequencing.

Parallel analyses of SS and UDS data derived from the same primary samples concurred remarkably well in the identification of HIV transmission chains in our study. Good correspondence between maximal pairwise genetic distances was observed with SS and UDS, indicating that distance-based clustering approaches traditionally applied to SS sequence data are applicable to UDS-derived phylogenies while capturing the underlying genetic diversity of minority variants. This also suggests that no minority variant, present at < 20%, the sensitivity limit of SS, was transmitted in the studied cohort. This is in line with the fact that no drug resistant minority variant was transmitted.

Direction of transmission is notoriously difficult to infer in HIV infections. Traditional approaches for inferring a pathogen's direction of transmission rely on known times of infection or time of symptom onset. This is due to the complex within-host evolution dynamics of the viral population, the chronic nature of an HIV infection, and the frequent lack

of epidemiological or clinical information from which a plausible window of transmission can be established. We used the topological structure of phylogenetic clusters to infer directionality in the studied transmission pairs. Our results confirmed the predictions of Romero-Severson's model [16], both in its advantages and limitations. Indeed, directionality remained equivocal in the transmission pairs we identified when the viral populations sampled from the linked individuals both exhibited a monophyletic structure. The same limitations applied to estimating the likelihood of direct transmissions within the clusters. Due to partial sampling of the infected population, the possibility of an intermediate transmitter between sampled linked individual cannot be excluded with certainty. As already raised by Abecasis *and al.,* this point is critical, and the phylogenetic analysis use as a forensic technique for HIV transmission investigations is risky [37].

The weakness of the phylogenetic signal in the selected genetic fragments represented a significant limitation of this study. It resulted in phylogenetic uncertainty, topological discrepancies across fragments and sometimes in conflicting results from one region to another (see cluster [34/46], where the topology of the cluster identified patient 34 as the transmitter in PR and 46 in RT). The HIV *pol* gene exhibits high levels of conservation, due to strong purifying selection. Although routine genotyping fragments were shown to contain enough patient-specific diversity to establish epidemiological linkage [12], the short fragments yielded by the 454 sequencing technology aggravated the phylogenetic signal of the data. Even shorter fragments, such as those obtained with the popular Illumina platforms, may represent a more severe challenge with the phylogenetic reconstruction method used until further technological improvements are reached. A trade-off between length and multiplicity of the sequences must therefore be reached.

The topology and branch uncertainty observed in most phylogenies precluded a precise estimation of the minimum number of transmitted founder variants in most cases. This was emphasized by the fact that topological structure varied across genetic fragments in 4 of the 8 transmission chains identified, and most likely due to differences in phylogenetic signal

and variant sampling bias across genes. However, when multiple, the minimum number of founder variants initiating infection was estimated to range between 2 and 10, depending on the genetic fragment. These estimates are in agreement with studies conducted in individuals with acute or very recent HIV infections [38] and support the notion of a massive reduction of genetic diversity at transmission. Through multiple bottlenecks, both stochastic and selective, the large, genetic diverse population found in a transmitter is reduced to a remarkably small number of established lineages in the recipient. This reduction is believed to be down to a single variant in a majority of sexual transmissions [39], but our framework did not allow us to observe such pattern. Monophyletic populations in the recipient could indeed have explanations other than a single founder virus, such as the presence of an observed intermediate host.

While majority resistant variants were mainly detected among all patients of a cluster, no evidence of transmission of MRV was observed in this work, as recently observed by Chaillon *et al.* [40]. This suggests either preferential transmission of majority variants and/or minority drug-sensitive variants or post transmission reversion to wild type. Carlson *et al.* found that transmitted viruses are the fittest, moderated by factors such as transmitter viral load and recipient genital inflammation [41]. Resistant variants are generally less fit than wild-type variants. On the other hand, rapid change of minority variants levels during early infection had been reported [42], supporting viral lineage modifications in response to adaptive immunity or consistent with the hypothesis that fitness-impacting mutations are efficiently and rapidly removed.

In conclusion, UDS could provide extra information on founder viruses and linkage but the interpretation of the data is still hazardous. The use of whole genome sequencing data or large fragments to improve the phylogenetic analysis by UDS must be studied.

**Acknowledgements**

We thank Christine Katlama, Anne Simon and Ruxandra Calin for following the patients. The virological tests were performed by Marc Wirden and Sophie Sayon (Sanger sequencing), Eve Todesco (UDS), and Francis Barin (quantification of antibodies by immunoassay). Phylogenetic analyses were performed by Stéphane Hué. The paper was written by Stéphane Hué and Eve Todesco. Anne-Geneviève Marcelin, Vincent Calvez, Eve Todesco and Stéphane Hué designed the study.

**Source of Funding:** This work was supported by the Agence Nationale de Recherches sur le SIDA et les hépatites virales (ANRS).

**References**


1 Pao D, Fisher M, Hué S, Dean G, Murphy G, Cane PA, *et al.* **Transmission of HIV-1 during primary infection: relationship to sexual risk and sexually transmitted infections.** *AIDS* 2005; **19**:85–90.

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

3 Hughes GJ, Fearnhill E, Dunn D, Lycett SJ, Rambaut A, Leigh Brown AJ, *et al.* **Molecular phylodynamics of the heterosexual HIV epidemic in the United Kingdom.** *PLoS Pathog* 2009; **5**:e1000590.

4 Oster AM, Pieniazek D, Zhang X, Switzer WM, Ziebell RA, Mena LA, *et al.* **Demographic but not geographic insularity in HIV transmission among young black MSM.** *AIDS* 2011; **25**:2157–2165.

5 Drescher SM, von Wyl V, Yang W-L, Böni J, Yerly S, Shah C, *et al.* **Treatment-naive individuals are the major source of transmitted HIV-1 drug resistance in men who have sex with men in the Swiss HIV Cohort Study.** *Clin Infect Dis* 2014; **58**:285–294.

6 Mbisa JL, Fearnhill E, Dunn DT, Pillay D, Asboe D, Cane PA, *et al.* **Evidence of Self-Sustaining Drug Resistant HIV-1 Lineages Among Untreated Patients in the United Kingdom.** *Clin Infect Dis* 2015; **61**:829–836.

7 Mourad R, Chevennet F, Dunn DT, Fearnhill E, Delpech V, Asboe D, *et al.* **A phylotype-based analysis highlights the role of drug-naive HIV-positive individuals in the transmission of antiretroviral resistance in the UK.** *AIDS* 2015; **29**:1917–1925.

8 Hué S, Brown AE, Ragonnet-Cronin M, Lycett SJ, Dunn DT, Fearnhill E, *et al.* **Phylogenetic analyses reveal HIV-1 infections between men misclassified as heterosexual transmissions.** *AIDS* 2014; **28**:1967–1975.

9 Ragonnet-Cronin M, Hué S, Hodcroft EB, Tostevin A, Dunn D, Fawcett T, *et al.* **Non-disclosed men who have sex with men in UK HIV transmission networks: phylogenetic analysis of surveillance data.** *Lancet HIV* 2018; **5**:e309–e316.

10 Pillay D, Herbeck J, Cohen MS, Oliveira T de, Fraser C, Ratmann O, *et al.* **PANGEA-HIV: phylogenetics for generalised epidemics in Africa.** *The Lancet Infectious Diseases* 2015; **15**:259–261.

11 Shafer RW. **Genotypic testing for human immunodeficiency virus type 1 drug resistance.** *Clin Microbiol Rev* 2002; **15**:247–277.

12 Hué S, Clewley JP, Cane PA, Pillay D. **HIV-1 pol gene variation is sufficient for reconstruction of transmissions in the era of antiretroviral therapy.** *AIDS* 2004; **18**:719–728.

13 Hall M, Woolhouse M, Rambaut A. **Epidemic Reconstruction in a Phylogenetics Framework: Transmission Trees as Partitions of the Node Set.** *PLoS Comput Biol* 2015; **11**:e1004613.

14 Kenah E, Britton T, Halloran ME, Longini IM. **Molecular Infectious Disease Epidemiology: Survival Analysis and Algorithms Linking Phylogenies to Transmission Trees.** *PLoS Comput Biol* 2016; **12**:e1004869.

15 Metzner KJ, Scherrer AU, Preiswerk B, Joos B, von Wyl V, Leemann C, *et al.* **Origin of minority drug-resistant HIV-1 variants in primary HIV-1 infection.** *J Infect Dis* 2013; **208**:1102–1112.

16 Romero-Severson EO, Bulla I, Leitner T. **Phylogenetically resolving epidemiologic linkage.** *Proc Natl Acad Sci USA* 2016; **113**:2690–2695.

17 Wymant C, Hall M, Ratmann O, Bonsall D, Golubchik T, de Cesare M, *et al.* **PHYLOSCANNER: Inferring Transmission from Within- and Between-Host Pathogen Genetic Diversity.** *Mol Biol Evol* Published Online First: 23 November 2017. doi:10.1093/molbev/msx304

18 De Maio N, Worby CJ, Wilson DJ, Stoesser N. **Bayesian reconstruction of transmission within outbreaks using genomic variants.** *PLoS Comput Biol* 2018; **14**:e1006117.

19 Wittkop L, Günthard HF, de Wolf F, Dunn D, Cozzi-Lepri A, de Luca A, *et al.* **Effect of transmitted drug resistance on virological and immunological response to initial combination antiretroviral therapy for HIV (EuroCoord-CHAIN joint project): a European multicohort study.** *Lancet Infect Dis* 2011; **11**:363–371.

20 Li JZ, Paredes R, Ribaudo HJ, Svarovskaia ES, Metzner KJ, Kozal MJ, *et al.* **Low-frequency HIV-1 drug resistance mutations and risk of NNRTI-based antiretroviral treatment failure: a systematic review and pooled analysis.** *JAMA* 2011; **305**:1327–1335.

21 Cozzi-Lepri A, Noguera-Julian M, Di Giallonardo F, Schuurman R, Däumer M, Aitken S, *et al.* **Low-frequency drug-resistant HIV-1 and risk of virological failure to first-line NNRTI-based ART: a multicohort European case-control study using**

**centralized ultrasensitive 454 pyrosequencing.** *J Antimicrob Chemother* 2015; **70**:930–940.

22 Barin F, Meyer L, Lancar R, Deveau C, Gharib M, Laporte A, *et al.* **Development and validation of an immunoassay for identification of recent human immunodeficiency virus type 1 infections and its use on dried serum spots.** *J Clin Microbiol* 2005; **43**:4441–4447.

23 Descamps D, Delaugerre C, Masquelier B, Ruffault A, Marcelin A-G, Izopet J, *et al.* **Repeated HIV-1 resistance genotyping external quality assessments improve virology laboratory performance.** *J Med Virol* 2006; **78**:153–160.

24 Todesco E, Rodriguez C, Morand-Joubert L, Mercier-Darty M, Desire N, Wirden M, *et al.* **Improved detection of resistance at failure to a tenofovir, emtricitabine and efavirenz regimen by ultradeep sequencing.** *J Antimicrob Chemother* 2015; **70**:1503–1506.

25 Daigle D, Simen BB, Pochart P. **High-throughput sequencing of PCR products tagged with universal primers using 454 life sciences systems.** *Curr Protoc Mol Biol* 2011; **Chapter 7**:Unit7.5.

26 Bennett DE, Camacho RJ, Otelea D, Kuritzkes DR, Fleury H, Kiuchi M, *et al.* **Drug resistance mutations for surveillance of transmitted HIV-1 drug-resistance: 2009 update.** *PLoS ONE* 2009; **4**:e4724.

27 Larsson A. **AliView: a fast and lightweight alignment viewer and editor for large datasets.** *Bioinformatics* 2014; **30**:3276–3278.

28 Price MN, Dehal PS, Arkin AP. **FastTree: Computing Large Minimum Evolution Trees with Profiles instead of a Distance Matrix.** *Mol Biol Evol* 2009; **26**:1641–1650.

29 Guindon S, Dufayard J-F, Lefort V, Anisimova M, Hordijk W, Gascuel O. **New algorithms and methods to estimate maximum-likelihood phylogenies: assessing the performance of PhyML 3.0.** *Syst Biol* 2010; **59**:307–321.

30 Ragonnet-Cronin M, Hodcroft E, Hué S, Fearnhill E, Delpech V, Brown AJL, *et al.* **Automated analysis of phylogenetic clusters.** *BMC Bioinformatics* 2013; **14**:317.

31 Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ. **Basic local alignment search tool.** *J Mol Biol* 1990; **215**:403–410.

32 Pond SLK, Frost SDW, Muse SV. **HyPhy: hypothesis testing using phylogenies.** *Bioinformatics* 2005; **21**:676–679.

33 Wang C, Mitsuya Y, Gharizadeh B, Ronaghi M, Shafer RW. **Characterization of mutation spectra with ultra-deep pyrosequencing: application to HIV-1 drug resistance.** *Genome Res* 2007; **17**:1195–1201.

34 Cornelissen M, Gall A, Kuyl A van der, Wymant C, Blanquart F, Fraser C, *et al.* **Workup of Human Blood Samples for Deep Sequencing of HIV-1 Genomes.** *Viral Metagenomics*.Humana Press, New York, NY; 2018. pp. 55–61.

35 Didelot X, Fraser C, Gardy J, Colijn C. **Genomic Infectious Disease Epidemiology in Partially Sampled and Ongoing Outbreaks.** *Mol Biol Evol* 2017; **34**:997–1007.

36 Campbell F, Strang C, Ferguson N, Cori A, Jombart T. **When are pathogen genome sequences informative of transmission events?** *PLOS Pathogens* 2018; **14**:e1006885.

37 Abecasis AB, Pingarilho M, Vandamme A-M. **Phylogenetic analysis as a forensic tool in HIV transmission investigations.** *AIDS* 2018; **32**:543–554.

38 Keele BF, Giorgi EE, Salazar-Gonzalez JF, Decker JM, Pham KT, Salazar MG, *et al.* **Identification and characterization of transmitted and early founder virus envelopes in primary HIV-1 infection.** *PNAS* 2008; **105**:7552–7557.

39 Tully DC, Ogilvie CB, Batorsky RE, Bean DJ, Power KA, Ghebremichael M, *et al.* **Differences in the Selection Bottleneck between Modes of Sexual Transmission Influence the Genetic Composition of the HIV-1 Founder Virus.** *PLoS Pathog* 2016; **12**:e1005619.

40 Chaillon A, Nakazawa M, Wertheim JO, Little SJ, Smith DM, Mehta SR, *et al.* **No Substantial Evidence for Sexual Transmission of Minority HIV Drug Resistance Mutations in Men Who Have Sex with Men.** *J Virol* 2017; **91**. doi:10.1128/JVI.00769-17

41 Carlson JM, Schaefer M, Monaco DC, Batorsky R, Claiborne DT, Prince J, *et al.* **HIV transmission. Selection bias at the heterosexual HIV-1 transmission bottleneck.** *Science* 2014; **345**:1254031.

42 Kijak GH, Sanders-Buell E, Chenine A-L, Eller MA, Goonetilleke N, Thomas R, *et al.* **Rare HIV-1 transmitted/founder lineages identified by deep viral sequencing contribute to rapid shifts in dominant quasispecies during acute and early infection.** *PLoS Pathog* 2017; **13**:e1006510.

**Figures Legend**

**Figure 1.** Schematic representation of the possible phylogenetic relationships between two viral populations in a transmission cluster and their interpretation (Adapted from Romero-Severson *et al.* 2016 [16]). Ancestral nodes are marked with circles and monophyletic groups of sequences from a given individual collapsed into triangles. For a putative transmission pair involving two individuals X (white) and Y (grey), the phylogenetic cluster formed by the two viral populations can exhibit three possible topologies: (A) When both viral populations were monophyletic (monophyletic-monophyletic topology, MM), the direction of infection was deemed as equivocal; (B) When one population was paraphyletic and the other monophyletic (paraphyletic-monophyletic topology, PM) the paraphyletic population was assigned to the transmitter (here, X); (C & D) When both populations were paraphyletic (paraphyletic-paraphyletic topology, PP), the population inferred to be ancestral to the cluster (marked with circles) was assigned to the transmitter. Directness of transmission was deemed equivocal for MM and PM topologies (A & B). A single founder virus was assumed to be transmitted in PM topologies (B), while PP topologies were considered representative of the transmission of multiple founder viruses (C & D).

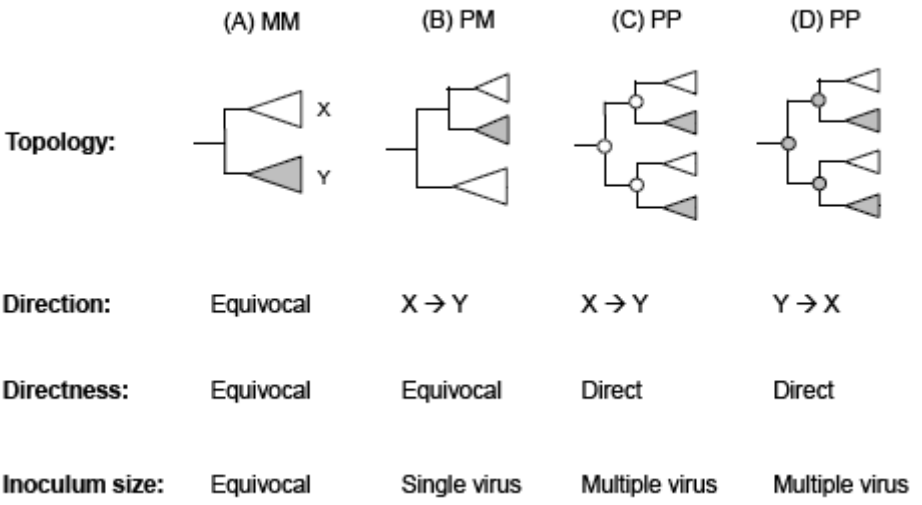

**Figure 2.** Examples of monophyletic-monophyletic (A), paraphyletic-monophyletic (B) and paraphyletic-paraphyletic (C) topologies observed amongst three putative transmission pairs. Sequences from the inferred transmitters (patients 52, 17 and 19) and recipient (patients 60, 30 and 62) are coloured in black and red, respectively.

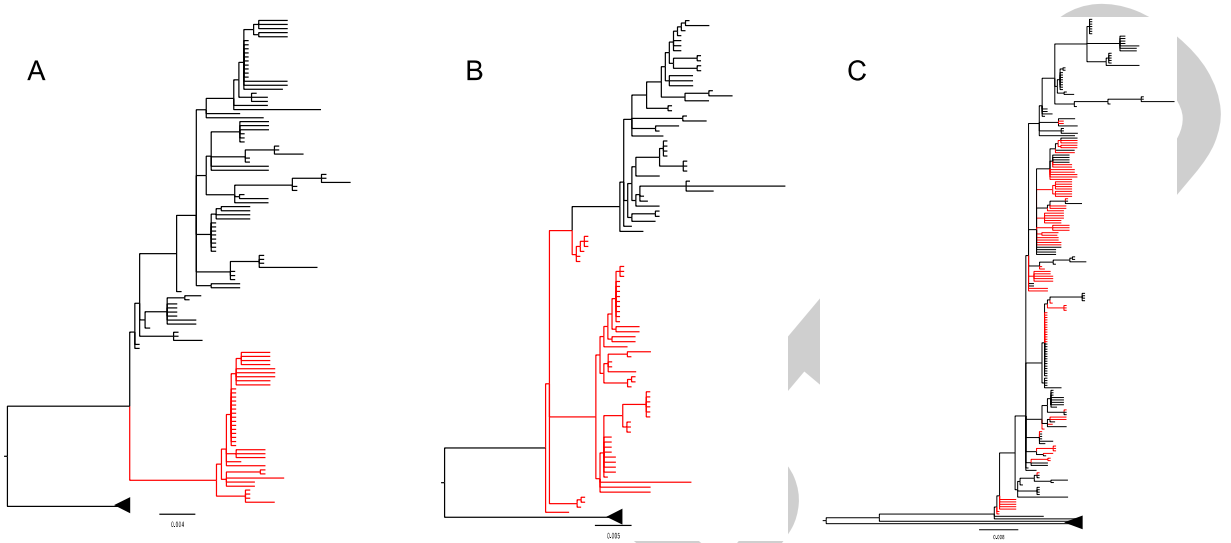

**Table 1.** Putative transmission clusters identified by phylogenetic reconstruction

| Cluster[1] | UDS (GD/BS[2]) | | | Sanger (GD/BS[2]) | | Cluster size | Subtype | Sampling interval (days) | Recently infected [3] |
|---|---|---|---|---|---|---|---|---|---|
| | PR | RT1 | RT2 | PR | RT | | | | |
| [03, 43] | - | - | - | 2.0/1.00 | - | 2 | CRF02_AG | 355 | 43 |
| **[05, 41]** | **2.0/1.00** | **2.5/1.00** | **2.5/0.99** | **1.5/1.00** | **1.5/1.00** | **2** | **B** | **336** | **41** |
| **[08, 44]** | **4.0/1.00** | **4.0/1.00** | **3.0/0.98** | **1.5/0.79** | **1.5/0.97** | **2** | **C** | **340** | - |
| [12, 24] | 0.98 | 4.0/0.91 | - | 2.0/0.92 | 1.5/1.00 | 2 | B | 161 | - |
| [13, 18] | - | - | - | 3.0/0.90 | - | 2 | CRF02_AG | 61 | - |
| [13, 65] | - | 3.0/1.00 | - | - | - | 2 | CRF02_AG | 413 | - |
| [14, 53] | 0.81 | - | - | - | - | 2 | B | 301 | - |
| **[17, 30]** | **2.5/0.93** | **2.5/0.98** | **3.5/0.98** | **1.5/0.93** | **1.5/1.00** | **2** | **B** | **136** | **17, 30** |
| **[19, 62]** | **1.5/0.89** | **3.0/0.97** | **3.0/1.00** | **1.5/0.98** | **1.5/1.00** | **2** | **C** | **313** | - |
| [22, 67] | - | - | 0.76 | - | 3.0/0.91 | 2 | B | 308 | - |
| **[25, 33, 45]** | **2.0/0.93** | **3.5/0.98** | **4.5/0.99** | **1.5/0.86** | **2.0/1.00** | **3** | **CRF02_AG** | **55/116/61** | **33** |
| [26, 55] | - | 4.0/1.00 | 0.99 | - | 3.5/0.92 | 2 | CRF01_AE | 168 | 26,55 |
| [27, 47] | - | - | 0.86 | - | - | 2 | B | 114 | - |
| [28, 66] | - | - | - | - | - | 2 | F2 | 254 | 28 |
| **[31, 70]** | **4.5/0.98** | **4.5/0.98** | **4.5/0.98** | **3.5/0.93** | **3.5/1.00** | **2** | **CRF07_BC** | **43** | - |
| [32, 40] | 0.99 | - | - | - | - | 2 | B | 49 | 32 |
| **[34, 46]** | **1.5/0.94** | **3.0/0.94** | **2.0/0.9** | **1.5/1.00** | **1.5/0.98** | **2** | **CRF02_AG** | **58** | **34, 46** |
| [48, 49] | - | - | 0.94 | - | - | 2 | CRF01_AE | 0 | - |
| **[52, 60]** | **4.5/0.99** | **3.0/1.00** | **3.5/0.98** | **3.5/0.92** | **1.5/1.00** | **2** | **B** | **48** | **60** |

Clusters identified from all genetic fragments are indicated in bold.

[1] Clusters named after the individuals they include, with the convention [Patient1, Patient2]; some individuals were present in different gene-specific clusters.

[2] GD: intra-cluster maximal genetic distance, in nucleotide substitution per site; BS: Branch support (cluster probability; SH-aLRT)

[3] Patient infected < 6 months prior to sampling

**Table 2.** Inferred direction and directness of transmission within transmission clusters

| Cluster | PR | | | RT1 | | | RT2 | | |
|---|---|---|---|---|---|---|---|---|---|
| | Topology[1] | Direction | Directness | Topology[1] | Direction | Directness | Topology[1] | Direction | Directness |
| [05, 41] | PP | 05 > 41 | Direct | PP | 05 > 41 | Direct | PP | 05 > 41 | Direct |
| [08, 44] | PP | 08 > 44 | Direct | PM | 08 > 44 | Equivocal | PM | 08 > 44 | Equivocal |
| [17, 30] | PP | 30 > 17 | Direct | PM | 30 > 17 | Equivocal | PM | 30 > 17 | Equivocal |
| [19, 62] | PP | 62 > 19 | Direct | PP | 62 > 19 | Direct | PP | 19 > 62 | Direct |
| [25, 33, 45] | MPP | 45 > 25/33 | Equivocal | MMP | 45 > 25/33 | Direct | MPP | 45 > 25/33 | Equivocal |
| [31, 70] | PM | 70 > 31 | Equivocal | PM | 70 > 31 | Equivocal | PM | 70 > 31 | Equivocal |
| [34, 46] | PP | 34 > 46 | Direct | PM | 46 > 34 | Equivocal | PM | 46 > 34 | Equivocal |
| [52, 60] | MM | Equivocal | Equivocal | MM | Equivocal | Equivocal | MM | Equivocal | Equivocal |

[1] MM, monophyletic/monophyletic; PM, paraphyletic/monophyletic; PP, paraphyletic/paraphyletic

Cluster topologies coherent across all fragments are shown in grey.

**Table 3.** Drug resistance mutations in the putative transmission clusters.

| Cluster | Patient | Inferred direction | Recently infected | TDR (from Bennet *et al.* 2009) | | DRMs (ANRS algorythm) | |
|---|---|---|---|---|---|---|---|
| | | | | PR | RT | PR | RT |
| [05, 41] | 5 | Transmitter | No | - | 181C (1.1%) | 60E, 63P | 181C (1.1%) |
| | 41 | Recipient | Yes | - | 188H (1.1%) | 60E, 63P | 188H (1.1%) |
| [08, 44] | 8 | Transmitter | No | 84V (1.7%) | - | 10I, 15V, 69K, 84V (1.7%), 89I | - |
| | 44 | Recipient | No | - | - | 10I, 15V, 69K, 89I | - |
| [17, 30] | 17 | Recipient | Yes | - | - | 62V | - |
| | 30 | Transmitter | Yes | - | - | 62V | - |
| [19, 62] | 19 | N/A | No | - | - | 15V, 36I, 69K, 89M | - |
| | 62 | N/A | No | - | - | 15V, 36I, 62V (2.8%), 69K, 89M | - |
| [25, 33, 45] | 25 | N/A | Yes | - | - | 20I, 36I, 63P, 69K, 89M | - |
| | 33 | N/A | No | - | - | 20I, 36I, 63P, 69K, 89M | - |

| | | | | | | | |
|---|---|---|---|---|---|---|---|
| | 45 | Transmitter | No | 46I (7.0%) | - | 20I, 36I, 46I (7.0%), 63P, 69K, 89M | - |
| [31, 70] | 31 | Recipient | No | - | - | 33V (7.6%), 60E, 63P | - |
| | 70 | Transmitter | No | - | | 15V (12.0%), 60E, 62V, 63P, 77I | 101R |
| [34, 46] | 34 | N/A | Yes | - | - | 16E, 20I, 36I, 69K, 89M | - |
| | 46 | N/A | Yes | - | - | 16E, 20I, 36I, 69K, 89M | - |
| [52, 60] | 52 | N/A | No | - | 219N (1.1%) | 63P, 71T | - |
| | 60 | N/A | Yes | 82A (3.0%) | - | 63P, 71T, 82A (3.0%) | - |

Mutations detected by both UDS and SS are underlined. The frequency of minority resistant variants detected by UDS is indicated in brackets.

DRMs: drug resistance mutations; IN: integrase; PR: protease; RT: reverse transcriptase; SS: Sanger sequencing; TDR: transmitted drug resistance; UDS: ultra-deep sequencing.