Factors influencing HIV-1 phylogenetic clustering

Dennis M. Junqueira[a,b], Zandile Sibisi[a,c], Eduan Wilkinson[a,c], and Tulio de Oliveira[a,c,d]

Purpose of review
A major goal of public health in relation to HIV/AIDS is to prevent new transmissions in communities. Phylogenetic techniques have improved our understanding of the structure and dynamics of HIV transmissions. However, there is still no consensus about phylogenetic methodology, sampling coverage, gene target and/or minimum fragment size.

Recent findings
Several studies use a combined methodology, which includes both a genetic or patristic distance cut-off and a branching support threshold to identify phylogenetic clusters. However, the choice about these thresholds remains an inherently subjective process, which affects the results of these studies. There is still a lack of consensus about the genomic region and the size of fragments that should be used, although there seems to be emerging a consensus that using longer segments, allied with the use of a realistic model of evolution and a codon alignment, increases the likelihood of inferring true transmission clusters. The *pol* gene is still the most used genomic region, but recent studies have suggested that whole genomes and/or sequences from *nef* and *gp41* are also good targets for cluster reconstruction.

Summary
The development and application of standard methodologies for phylogenetic clustering analysis will advance our understanding of factors associated with HIV transmission. This will lead to the design of more precise public health interventions.

Keywords
fragment size, gene target, HIV, phylogenetic, transmission cluster

INTRODUCTION

Phylogenetic analysis has played a major role in understanding the dynamics that have shaped the HIV pandemic since its beginning [1–5]. Based primarily on homologous nucleotide comparison, phylogenetic trees are plain representations of the evolutionary history of a given group of sequences. The tree structure is normally used to provide valuable information on viral diversity, recombination profile, demographic history and ancestral characterization [6–8]. Recently, phylogenetic analysis has been widely used to identify viral linkage between different hosts and to infer putative HIV networks among infected-patients. However, there is still some discussion about what constitutes a phylogenetic cluster, and several different methods are currently being used [9▪].

In a phylogenetic tree, an HIV transmission cluster is commonly described as a set of nonrandom-related sequences that share a common ancestor reflecting one or several events of viral transmission [9▪,10▪▪]. The assumption is that after viral transmission between hosts, HIV strains in the recipient will be similar to the strains found in the transmitting individual and therefore tend to cluster together, despite intrahost evolution [11]. If sampling is complete, a phylogenetic tree could capture the underlying structure of transmission networks within a given population [12–14]. Normally, these analyses involve the generation of viral sequences through one of the several sequencing methods available nowadays, the alignment of sequences

[a]KwaZulu-Natal Research Innovation and Sequencing Platform (KRISP), Nelson R Mandela School of Medicine, University of KwaZulu-Natal, Durban, South Africa, [b]Centro Universitário Ritter dos Reis - UniRitter, Porto Alegre, Rio Grande do Sul, Brazil, [c]College of Health Sciences, School of Laboratory Medicine and Medical Science, University of KwaZulu-Natal, Durban, South Africa and [d]Department of Global Health, University of Washington, Seattle, Washington, USA

Correspondence to Tulio de Oliveira, KwaZulu-Natal Research Innovation and Sequencing Platform (KRISP), Nelson R Mandela School of Medicine, University of KwaZulu-Natal, K-RITH Tower Building, Ground Floor, 719 Umbilo Rd, Congella, Durban, South Africa. Tel: +27 31 260 4898; e-mail: deoliveira@ukzn.ac.za

Curr Opin HIV AIDS 2019, 14:161–172

DOI:10.1097/COH.0000000000000540

KEY POINTS

- Phylogenetic methods can improve our understanding of the structure and dynamics of HIV transmissions by identifying viral linkage between different hosts and inferring putative networks.
- There is confusion surrounding HIV clustering analysis due to differences in sampling, methodological approaches and, most importantly, in the genomic targets (gene fragment and its length) used to reconstruct the phylogenetic trees.
- It is clear that the use of a genetic or patristic distance threshold, allied with a high branching support cut-off, allows the identification of transmission clusters more accurately and improves the quality of the analysis.
- Despite the lack of agreement about the size of fragments used for HIV phylogenetic reconstruction, there seems to be consensus that using longer segments, allied with the use of a realistic model of sequence evolution and an improved alignment, increases the likelihood of inferring true transmission clusters.
- Aside from the *pol* gene, some studies have suggested that longer sequences from *nef* and *gp41* are also good targets for cluster reconstruction.

with one another to obtain position homology, the inference of a phylogenetic tree to reconstruct the evolutionary history of the analysed strains and the identification of clustered sequences that are related by transmission events [15,16]. Phylogenetic clusters are commonly defined by branch support and by genetic or patristic distance threshold (Fig. 1). Finally, after the identification of the phylogenetic clusters, the association between individual traits and cluster membership are evaluated statistically.

Genetic data, coupled with clinical and demographical information, could help define the characteristics of individuals and subpopulations contributing disproportionally to transmission events and, in some cases, this could also help predict or simulate future HIV infections [17,18]. Intervening in risk networks can improve health outcomes and prevent new infections [16,19]. However, there is confusion surrounding HIV clustering analysis due to differences in sampling, methodological approaches and, most importantly, in the genomic targets used to reconstruct the phylogenetic trees. The choice of the size and the subgenomic region to be analysed is crucial for cluster definition and has an important effect on the extent of sequence clustering [20]. The lack of standardization in this field has provoked growing interest in seeking to establish how the choice of gene fragment and its length affect HIV phylogenetic clustering [20,21].

Molecular phylogenies based on single genes can often lead to apparently conflicting results due to phylogenetic incongruence [22,23]. Such differences are related to stochastic error usually attributed to the small gene sequences and/or lack of phylogenetic signal in the data [24]. Using longer gene fragments could theoretically overcome these incongruences but, sometimes, systematic error do not disappear with the addition of data [24,25]. With regards to HIV-1 phylogenetic analysis, the polymerase (*pol*) gene has been predominantly used for phylogenetic reconstruction of transmission events over the past decade [14,26–31]. Although initially the envelope (*env*) gene was considered to present the strongest phylogenetic signal, it was argued that some *env* fragments were too short and too variable for a robust analysis [32] or could be severely affected by convergent and parallel evolution [33,34]. The wide use of *pol* in clustering studies is attributed to the fact that HIV-1 *pol* sequences are generated as part of routine clinical care and thus very large data sets are available for analysis. However, the increasing availability of HIV whole genome sequences in the last few years may stimulate discussions about whether full-length genome trees should be used, and which viral genes provide the most statistically accurate tree for transmission inference.

Here, we outline the latest developments in HIV phylogenetic clustering methods, the effect of gene selection and the parameters that can affect the outcome of the analysis. We also review recent evidence from studies regarding HIV-1 phylogenetic analysis in the developed and developing countries.

DEFINITION MATTERS

Clusters in epidemiology are broadly described as an unusual aggregation of infection [35]. In a phylogenetic approach, clusters contain sequences from different individuals usually separated by low genetic or evolutionary distances but invariably sharing a viral common ancestor [36■■]. These groupings are manifested as subtrees in the phylogeny and are likely to reflect past, recent or ongoing transmission events.

Phylogenetic clusters are often defined by high branch support (approximate likelihood ratio test, bootstrap, transfer bootstrap or posterior probability) and low genetic or patristic distance (Fig. 1) [37]. The use of patristic distances for genetic clustering is essentially an extension of clustering by pairwise genetic distances [36■■,37,39]. However, the choice of threshold cut-offs to define clusters in the literature is *ad hoc* (Table 1). For HIV, bootstraps,

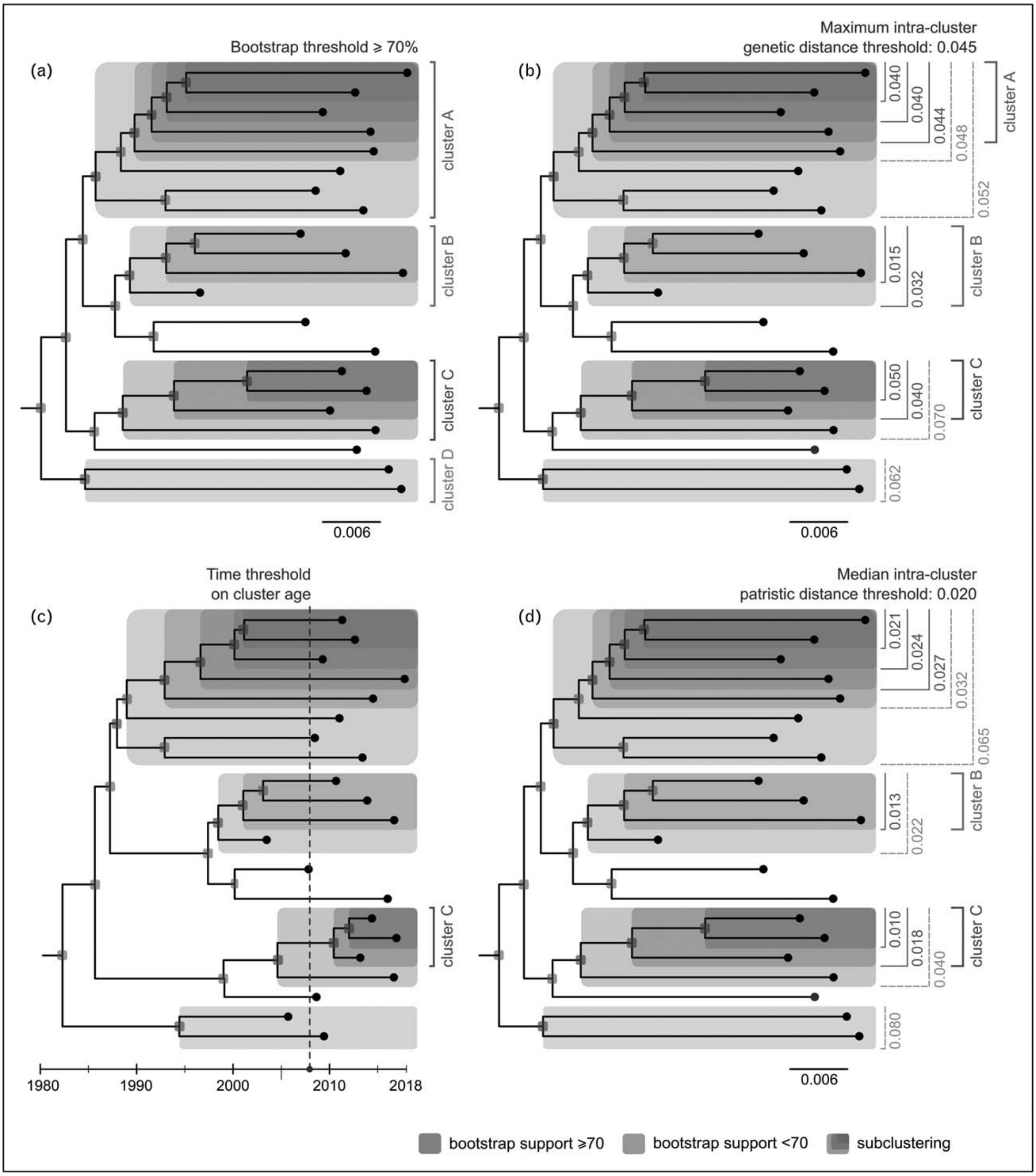

FIGURE 1. Comparison among three criteria used to define HIV phylogenetic transmission clusters. Phylogenetic clusters were identified by brackets. (a) Clusters are defined based solely on a threshold branch support (bootstrap support >70). (b) Putative transmission clusters detected by using the branch support cut-off are evaluated by using a maximum intracluster genetic distance (number of differences per site from between sequences) threshold of 0.045. Only clades with a maximum genetic distance under the threshold will be considered transmission clusters. The values next to the figure represent the maximum pairwise genetic distances of each subcluster. Dashed lines mark those subclusters with a maximum genetic distance above the threshold. (c) After evaluating branch support and the maximum genetic distance within clades, transmission clusters can be defined using a temporal constraint on a time-scaled phylogenetic tree. The dashed red line marks the temporal threshold. (d) Putative transmission clusters detected by using the branch support cut-off are evaluated by using the median intracluster patristic distance threshold (<0.020).

Table 1. Compilation of studies including HIV-1 phylogenetic clustering analysis available on PubMed Central in the last 3 years (August 2015–August 2018). Search was based in the keywords 'HIV', 'cluster' and 'phylogeny' and only included English-written articles

Clustering identification	Authors	Location	Sample size	HIV-1 genomic region	Maximum genomic size (base pairs)	Thresholds
Use of branching support threshold						
	Konou *et al.* (2016)	Togo	75	*RT*	1200	SH-like >98
	Patino-Galindo *et al.* (2016)	Spain	11 001	*pol*	1080	SH-like >90
	da Guarda-Reis *et al.* (2017)	Brazil	87	*pol*	998	Bootstrap >70
	Hu *et al.* (2017)	Asia	172	NFLG	NFLG	SH-like = 100
	Menza *et al.* (2017)	United States	2784	*pol*	–	SH-like > (unspecified)
	Paraschiv *et al.* (2017)	Romania	117	*pol*	1301	SH-like > 90 + PP = 1
	Patino-Galindo *et al.* (2017)	Spain	1806	*pol*	1302	SH-like > 99
	Sallam *et al.* (2017)	Iceland	106	*pol*	1020	SH-like > 90 + PP = 1
	Sallam *et al.* (2017)	Iceland	230	*pol*	918	SH-like > 90
	Vrancken *et al.* (2017)	Canada	1146	*pol*	1497	PP > 95
	Jagdagsuren *et al.* (2017)	Mongolia	143	*pol* and *env*	1065 and 447	Bootstrap >90
	Yue Yang *et al.* (2018)	China	323	*pol*	1000	SH-like > 90
Use of branching support and genetic distance threshold						
	Chan *et al.* (2015)	United States	1166	*pol*	–	Bootstrap >90%, GD < 3.1%
	Dauwe *et al.* (2015)	Europe	410	*pol* and *env*	857 and 105	SH-like > 90, GD < 4.5%
	Dennis *et al.* (2015)	United States	1673	*pol*	–	SH-like > 90, GD < 1.5%
	Fabeni *et al.* (2015)	Italy	534	*pol*	1302	Bootstrap >90%, GD < 1.5%
	Mbisa *et al.* (2015)	United Kingdom	1140	*pol*	897	Bootstrap >80%, GD < 2.0%
	Robineau *et al.* (2015)	France	547	RT	–	Bootstrap >98, GD < 1.5%
	Tamalet *et al.* (2015)	France	79	*pol*	1200	Bootstrap >98, GD < 4.5%
	Castley *et al.* (2016)	Australia	1021	*pol*	702	Bootstrap >98, GD < 1.5%
	Junqueira *et al.* (2016)	South America	4810	*pol*	999	SH-like > 90, GD < 4.5%
	Marzel *et al.* (2016)	Switzerland	19 604	*pol*	–	Bootstrap >50% and <100%; GD < 1%, <1.5%, <2%, <2.5%
	Ragonnet-Cronin *et al.* (2016)	Europe	83 299	*pol*	1617	Bootstrap >70, >80, >90, >95; GD < 1.5% and <4.5%
	Ragonnet-Cronin *et al.* (2016)	United Kingdom	14 693	*pol*	1206	SH-like > 90, GD < 4.5%
	Yousef *et al.* (2016)	Germany		*pol*	1038	Bootstrap >95, GD breadth-first
	Brenner *et al.* (2017)	Canada	4039	*pol*	918	Bootstrap >95, GD < 1.5%
	Brenner *et al.* (2017)	Canada	3901	*pol*	1496	Bootstrap >98.5, GD < 1.5%
	Castley *et al.* (2017)	Australia	4873	*pol*	957	Bootstrap >98, GD < 1.5%
	Fabeni *et al.* (2017)	Italy	4323	*pol*	1302	Bootstrap >90, GD < 0.015%
	Parczewski *et al.* (2017)	Poland	966	*pol*	1302	SH-like > 90, GD < 3.0%
	Turk *et al.* (2017)	Swiss	101 773	*pol*	1615	–, GD –
	Vanhommerig *et al.* (2017)	Netherlands	6861	*pol*	1140	SH-like > 90, GD < 0.08%
	Wolf *et al.* (2017)	United States	1953	*pol*	1080	SH-like > 95, GD < 1.5%
	Rouet *et al.* (2018)	Cambodia	202	*env*	1305	Bootstrap >99, GD < 0.07%
	Soto-Nava *et al.* (2018)	Mexico	1450	*gag* and *pol*	1634 and 1302	Bootstrap >90, GD < 1.5%
	Yebra *et al.* (2018)	London	365	*pol*	1002	>90, GD < 4.5%

References were only included in this table if all columns could be filled in with the information provided in the publication (articles describing new methods were excluded). *env*, genomic region encoding the viral glycoproteins; *gag*, genomic region encoding the capsid proteins; GD, genetic distance; *pol*, genomic region encoding the viral enzymes; PP, posterior probability; RT, reverse transcriptase; SH-like, Shimodaira-Hasegawa like branch support.

Shimodaira–Hasegawa -like or posterior probability branch support ranging from 50 to 100 [14,40–45] in combination with intracluster genetic distances thresholds between 1 or less and 9.0% or less [31,43,44,46–50] or patristic distance thresholds between 0.020 and 0.080, have been used [19,38,50,51]. A downside of this procedure is that the onus is on the user to determine the appropriate support/distance thresholds; a rationale for the selection of these thresholds is rarely provided. Such decisions can directly affect the study's results and modulate some inferences.

Although genetic or patristic distance is normally combined with branch support threshold, some studies have used only the branching cut-off to detect putative transmission clusters (Table 1). High statistical support (e.g. bootstrap) for any specific clade under a phylogenetic approach indicates that there is no close relative taxon to the clade in question. However, this method does not guarantee that the members of the clade itself are necessarily closely related to each other [52]. The use of a distance threshold within the cluster can help the researcher to exclude such false positive transmission cluster, especially when analysing transmissions over a large geographical range or with a wide sampling period.

To address the effect of the genetic distance and branch support on HIV-1 cluster identification, Marzel *et al.* [45] evaluated the effect of 104 different combinations of intracluster maximum genetic distances ($\leq$1, $\leq$1.5, $\leq$2 and $\leq$2.5%) and bootstrap supports (50–100% in 2% increments). Stricter bootstraps thresholds were associated with the identification of fewer transmission clusters ($P < 0.01$ for all four different genetic distances). In addition, the higher the bootstrap threshold used, the smaller the number of putative transmission links found. As expected, using a high cut-off value for bootstrap (100%) and a very strict threshold for the intraclade variability ($\leq$1.0%) allowed a small number of putative transmission links to be identified. Junqueira *et al.* [31] also evaluated the effect of different maximum genetic distances (between 1.5 and 7.5%, using intervals of 0.5%) and found that the number of clusters detected increased with the genetic distance threshold, reaching a maximum at a genetic distance of 6.5%. However, the proportion of sequences included in these clusters kept increasing with the increment in the genetic distance threshold. This clearly suggests that both genetic distance and branching support thresholds have a great impact on the identification of transmission clusters.

The genetic distance threshold can also impact the age of the clusters being identified. Reducing the cut-off increases the probability of identifying cases related by recent and rapid transmission events [16]. By contrast, if the main objective is to identify all possible transmissions that could be related to a given case, a larger genetic distance threshold should be used. According to the Center for Disease Control, the use of a genetic threshold of 0.5% for HIV-1 phylogenies corresponds to approximately 2–3 years of viral evolution separating those sequences inside any putative transmission cluster [16]. A 1.5% threshold, in this case, would detect networks with a maximum of 7–8 years of viral evolution. However, in a recent study, Marzel *et al.* [45] estimated the fraction of transmissions attributable to recent infections by analysing potential transmission pairs using different combinations of bootstrap and genetic distance. Their results suggest that the fraction of recent transmission increases sharply for higher bootstrap values (>92%) and that this is independent of the genetic distance tested. This indicates that high bootstrap thresholds (rather than genetic distances) may bias the selection towards recently infected transmission pairs.

Some authors have used a temporal and/or geographical constraint to define transmission clusters [53,54–57,58▪]. They did that by using time-resolved trees (which require the application of molecular clock) and selecting only those clusters that a common ancestor is younger than a selected cut-off date [12]. This approach is usually used to detect clusters of recent infections. Other authors used geographical constraints (e.g. city or country) for the analysis of clustering. This method guarantees that only local clusters are analysed. They also used phylogeographical to determine the interrelatedness of local clusters to other geographical regions [31,58▪].

The best method for calculating intracluster genetic distance is another topic of debate. The median and mean of the pairwise genetic distances of clustered sequences, as well as the maximum intracluster genetic distance, have been employed in many studies [12,48,50,59]. Although maximum and median distances are less affected by the number of sequences, the use of the mean tends to normalize the distances by the total number of sequences in the cluster and can potentially affect clustering prediction [59]. A study of methods comparisons have shown that both the use of maximum and median genetic distances have similar performances regarding the number and distribution of the clusters detected [59].

Several software applications have been developed to identify transmission clusters but Cluster Picker [59] and PhyloPart [50] are the most commonly used. Cluster Picker is a free software that implements both branch support and maximum cluster genetic distance to identify phylogenetic clusters. PhyloPart is also free and relies on the use of a threshold for the median pairwise patristic distances to detect clusters. A novel software application, called TreeCluster, can identify transmission clusters with different methods, including average, median and maximum genetic distances (https://github.com/niemasd/TreeCluster).

Despite not usually being a criterion chosen by researchers, the sampling density is a limitation worth mentioning because it can affect the choice of genetic distance and branching support. Novitsky *et al.* [60] addressed this question using 1248 *env* gp120 (V1–C5) sequences from a single community in Botswana and 2442 sequences from the LANL HIV Database. Through the simulation of 16 different sampling densities (1–70%) and phylogenetic detection of transmission clusters (bootstrap ≥70, ≥80, ≥90 and 100; genetic distance ≤1–≤15%), they showed that HIV clustering is positively correlated with sampling density. In addition, they found that sampling density below 10% was associated with variable clustering [and broad confidence intervals (CIs)], implying that results from HIV clustering at low sampling density may not reliable. A limitation of this study was that it used a small region of the *env* gene of HIV-1, which is likely to have affected their results.

Analyses of large epidemics (encompassing countries or continent) are usually characterized by poor sampling of HIV-positive individuals and, as a consequence, the transmission network may not be fully understood. The identification of transmission clusters in these cases should adopt more relaxed genetic distances to capture the expected amount of viral genetic variation. However, high branching supports should be used. The same rationale can be applied to data sets sampled over a long period of time: the characterization of transmission patterns might be more problematic in these cases because of intrahost evolution [14]. Sampling concentrated epidemics, nevertheless, theoretically guarantees an oversampling of individuals linked by transmission and stricter genetic distances should be tested. If the objective is to capture recent transmissions, the use of a very small genetic distance and high branching support should be considered. However, whatever data is used, researchers should be cautious about the sampling proportion, since very low sampling seems to result in variable clustering. In either case, genetic transmission clusters tend to group individuals sampled soon after infection, irrespective of whether those infections resulted from higher rates of transmission [36■■]. The evaluations of the parameters that may affect viral clustering outcomes define the quality of the results and are directly reflected in the results of the study.

BIGGER IS BETTER

The choice of genetic marker is important when seeking to identify phylogenetic clustering and perform other phylodynamic analysis. Extensive debate exists concerning the choice of gene(s) in HIV phylogenetic, as there are a number of factors to consider [14,61,62]. Most phylogenetic studies have relied on the HIV-1 *pol* gene, and to a lesser extent on the *env* and *gag* regions (Table 2). The gene choice is easily explained by the extensive efforts made to detect drug resistance mutations in the *pol* gene and not necessarily because this gene is more adequate for phylogenetic clustering.

Initial attempts to detect and infer HIV-1 transmission clusters were based both on *env* and/or *pol* genomic regions [6,63,64]. One of the first studies to use phylogenetic analysis to reconstruct transmission events analysed samples from a criminal investigation [65]. The well known Florida dentist case constituted the first case in the United States in which phylogenetic analysis was used in a criminal court. The investigators used different substitution models and phylogenetic optimality criteria targeting *pol* and *env* genes to link the dentist's sample with the victim's HIV-1 strains. In addition, drug resistance mutations were also used to link the cases and prove that the dentist purposely transmitted the virus to six of his patients. However, there has been considerable debate since then on the adequacy of single gene phylogenies to confidently establish true relationships in transmission cases [61,62].

Phylogenetic analysis can be very informative, but the accuracy of phylogenetic conclusions depends on the method, the gene and the sampling strategy [66,67]. Previous studies indicate that the degree of HIV clustering is commonly associated with the length of viral sequences used for phylogenetic reconstruction (see citations in Table 2). Other studies combined different methods and genes, they found that longer alignments identified transmission clusters more adequately [20,68,69]. However, in the studies of Albert *et al.* [70] and Lemey and Vandamme [68], shorter fragments performed better than lengthened alignments (Table 2). It is important to highlight that in both studies, partial *pol* fragments were used and despite being longer than the other subgenomic region tested, this alignment presented an inferior clustering performance. To establish how fragment length impacts on clustering inference, a specific subgenomic fragment with different lengths should be used to eliminate any noisy phylogenetic signal introduced by different genomic regions.

Table 2. Compilation of studies in which different HIV-1 subgenomics fragments were compared regarding its adequacy in answering clustering questions

HIV-1 genomic region(s)	Authors	Comparison with HIV-1 full genome results available	Fragment(s) and size(s)	Phylogenetic conclusion
All	Lemey *et al.* (1996)	Yes	400, 800 and 1200 bp sliding windows	*vif* (400 bp) and the 3′ part of the *pol* gene up to the *env* (1200 bp) were able to reconstruct the clusters
	Novitsky *et al.* (2015)	Yes	1000 and 2000 bp sliding windows	*pol* performed better than *gag* and *env*, specifically RT region. Extent of HIV clustering was significantly higher for 2000 bp sliding windows
	Yebra *et al.* (2016)	Yes (simulation)	*gag–pol–env* (6987 bp), *gag–pol* (4479 bp), *gag* (1479 bp), *pol* (3000 bp), *env* (2508 bp) and partial *pol* (1302 bp)	Accuracy of the trees was significantly proportional to the length of the sequences used. In addition, the lowest sampling depths (20 and 5%) greatly reduced the accuracy of tree reconstruction, especially when using the shortest gene data sets
	Harris *et al.* (2003)	Yes	*gag* (1470 bp), *pol* (3001), *gag* + *pol* (4275 bp), acessories (1126 bp), *env* (2420 bp), *nef* (382 bp), and C2–V3 (264 bp)	Whole-genome performed better than the subgenomic regions. This study didn't provide any statistical comparison between the adequacies of the subgenomic regions on reproducing data from the whole-genome alignment
gag, *pol* and *env*	Hué *et al.* (2004)	No	*gag* (p17/p24, 690 bp), *pol* (1002 bp) and *env* (V3, 550 bp)	Identical topologies were obtained in trees implemented from *gag*, *pol* and *env* gene alignments
	Rachinger *et al.* (2011)	No	*gag* (1018 bp), *pol* (977 bp) and *env* (C2–C4, 546 bp)	All trees corroborate the same results
gag and *pol*	Albert *et al.* (1994)	No	*gag* (p17, 300 bp) and *pol* (RT, 642 bp)	*gag* performed better than *pol*
pol and *env*	Stürmer *et al.* (2004)	No	*pol* (size not provided) and *env* (C2–V3, size not provided)	PT/RT region cannot be used on its own to prove true relationships between unknown patient isolates. At least two subgenomic regions should be used
	Lemey *et al.* (2005)	No	*pol* (1069 bp) and *env* (gp41, 951 bp)	*env* phylogenetic trees reproduced the previous know transmission history. Controversially, *pol* data set failed to reproduce it
	Amogne *et al.* (2016)	Yes	*pol* (938 bp) and *env* (500 bp)	Whole-genome provides a strong geographical clustering, but only a weak geographical cluster with respect to smaller gene fragments
gag and *env*	Leitner *et al.* (1996)	No	*gag* (p17, 438 bp), *env* (V3, 285 bp) and *gag* + *env* (723 bp)	a combination of p17 and V3 performed best
	Paraskevis *et al.* (2004)	No	*gag* (p17, 400 bp), *env* (C2–C4, 660 bp) and *gag* + *env* (1060 bp)	a combination of *gag* and env performed best

env, genomic region encoding the viral glycoproteins; *gag*, genomic region encoding the capsid proteins; *pol*, genomic region encoding the viral enzymes; PT, protease; RT, reverse transcriptase.

In the study of Novitsky *et al.* [20], which used HIV-1 whole genomes and four levels of bootstrap threshold to identify HIV clusters (≥70, ≥80, ≥90 and 100) in Botswana, the extent of HIV clustering was significantly higher for larger sliding windows (2000 versus 1000 nucleotides) spanning similar regions in the HIV-1 genome. However, this difference gradually decreased with tightening of the bootstrap threshold. Regarding differences in structural gene clustering, this study found that the proportion of *pol* sequences in clusters was larger than *gag* (at any bootstrap threshold used) and *env* (only at bootstrap ≥90). The stability of clusters for *pol* decreased when the bootstrap threshold was raised. In this study, the clusters were not highly supported, mean of 59% (95% CI 52–65%). Similarly, the identification of false clusters, not identified in the whole genome, tree also decreased with the increase in the bootstrap threshold. This study also found that when subgenomic regions were analysed, the reverse transcriptase region of the *pol* gene identified the highest proportion of sequences in clusters. A positive association between the extent of HIV clustering and parameters related to the sequence length, such as the number of variable sites and the number of informative sites, was also found. These results highlight the appropriateness of longer fragments and the apparent adequacy of reverse transcriptase (*pol*) to reproduce data generated on whole genome sequences.

In another study investigating the influence of the fragment size on HIV-1 clustering, Lemey and Vandamme [68] examined three known and distinct transmission cases for which full-genome sequence data were available. The full-genome phylogenetic tree highly supported the three clusters (100% bootstrap), using only a branching support threshold to identify clusters. To evaluate which genomic regions were the most informative for transmission chain reconstruction, they performed a sliding window analysis with 400, 800 and 1200 base pairs (bp) using the same maximum likelihood tree inference method for different window sizes. Extensive variation in gene-specific bootstrap support was observed among the three transmission clusters, suggesting that some genomic regions can perform better than others in reconstructing transmission clusters. For a minimum window size of 400 bp, only the *vif* gene region provided considerable bootstrap support (>90%) for all the transmission clusters. Increasing the window size up to 800 bp resulted, on average, in an increase in transmission cluster support and more distinct patterns of gene-specific support. The 3′ part of the *pol* gene up to the *env* gene appeared to provide good support for all three transmission clusters in all window sizes. However, there was still considerable variability in transmission cluster bootstrap support for *gag* and *env*. A window size of 1200 bp resulted in a further average increase in bootstrap support but still showed that *gag* and *env* regions were not able to detect all three clusters.

To assess which gene target or sequence length is optimal for phylogenetic cluster analysis, it is important to correlate the accuracy of phylogenies inferred from each gene fragment and sequence length. Yebra *et al.* [69] recently used this method to determine which gene(s) provide(s) the best approximation to the real phylogeny by subsampling a simulated data set of 4662 sequences. After subsampling the data set into different combinations of genes (*gag–pol–env*, *gag–pol*, *gag*, *pol*, *env* and partial *pol*), sampling depths (100, 60, 20 and 5%) and replicates ($n = 100$), maximum likelihood trees were constructed and then compared with those of the corresponding true tree. They found a statistically significant positive correlation between the sequence length and the similarity to the true tree, suggesting that the accuracy of the trees was proportional to the length of the sequences used. The *gag–pol–env* data sets showed the best performance considering all the sampling coverage levels together, followed by *gag–pol*, *pol* and *env*. The smaller *gag* (1479 bp) and partial *pol* (1302 bp) showed the worst results. This was also true when analysing the sampling coverage levels individually. In addition, the 60% sampling coverage provided the most similar results to the analyses of the complete data sets. The study concluded that *gag–pol* provides a dependable approximation of the results of nearly whole genome sequences.

The studies referred to above have shown that longer gene sequences are positively associated with tree accuracy. However, the total length of the subgenomic fragments used to identify transmission clusters has been a concern as obtaining longer sequences increases costs (and sometimes it means losing important data sets generated from drug resistance genotyping). Increasing the number of sequences in an alignment, to compensate for the small size of the subgenomic fragments, can also lead to spurious results [71]. The use of a high and strict branching support threshold seems to overcome the use of short sequences (∼1000 bp), irrespective of the targeted genomic region. In addition, the use of a realistic model of sequence evolution and a carefully edited codon alignment can reduce the bias created by data sets with short length sequences [71,72].

THE MORE THE BETTER

The use of longer genetic regions will allow for a more reliable reconstruction of transmission events and better cluster enumeration. Full genome sequences could be considered the top choice for the most informative HIV cluster inference, but such analysis is not always possible. An alternative approach could be based on the use of a selected subgenomic region with an elevated performance for HIV clustering and high tree accuracy.

Hué *et al.* wished to understand the potential of using *pol* gene (protease [PT] + partial reverse transcriptase, 1002 bp) to identify transmission events using phylogenetic trees. Trees based on the corresponding *gag* (p17/p24, 690 bp) and *env* (V3 loop, 550 bp) genes were used to confirm the linkages [14]. Their results show that phylogenetic clustering patterns for these three data sets (*gag*, *pol* and *env*) were identical, with a similar range of statistical significance (bootstraps higher than 95%). Where appropriate information was obtained, three clusters from the twenty three were supported by evidence of epidemiological linkage. In addition, they concluded that resistance mutations induced by antiretroviral therapy are unlikely to bias the reconstruction of relatedness between samples but could potentially help in the study of transmission. Since its publication, this study has been widely accepted by the scientific community (190 citations in August 2018) because it justifies the use of *pol* to identify phylogenetic transmission analysis. However, the use of an uncontrolled data set (lack of transmission link information for clustered sequences) may affect the outcome of the analysis and may limit the extrapolation of these results to other studies.

In contrast, by using a controlled data set with information about direction for three transmission chains, Lemey *et al.* showed an adequacy of the 3′ region of *pol* in inferring epidemiological linkage. Similarly, in the study of Novitsky *et al.*, a data set of the HIV-1 *pol* fragment performed better in reconstructing clustering from whole genome data. Rachinger *et al.* reached similar conclusions. Their study analysed serial samples from three HIV-positive individuals from a triangular relationship during a follow-up of 3 years [7]. Time-measured phylogenies of clonally amplified *gag* (1018 bp), *pol* (PT + partial reverse transcriptase, 977 bp) and *env* (C2–C4, 546 bp) sequences provided compelling evidence for the direction of HIV-1 transmissions [7]. All trees confirmed a single transmission event from patient three to patient one and from patient three to patient two. Despite corroborating the results of other studies, the investigation of Rachinger *et al.* had a sampling peculiarity, since several sequences from the same patient were isolated. This sampling method improved the phylogenetic signal of the data set and gave the results more accuracy but cannot be extrapolated to a 'one sequence per patient' analysis.

Different studies have suggested that the 3′ end of this gene (including reverse transcriptase) has a better performance than the traditional PT + reverse transcriptase fragment (~1300 bp). However, in a well controlled study using different sampling proportions, a combination of genes and different fragment sizes, Yebra *et al.* [69] showed that the *pol* alignment encompassing the first 1302 bp had the worst performance. One possible explanation for these contrasting results is that the subtype may influence the pattern of clustering across different subgenomic regions inside the HIV-1 genome, as different subtypes are being used across distinct investigations (subtype C strains were used in this study as opposed to subtype B in the great majority of the publications in the field). However, to the best of our knowledge, no study has addressed the influence of subtypes in clustering results.

Opposing the assumption that *pol* gene is as a good target for transmission cluster identification, Stürmer *et al.* [73] compared the performance of this gene against *env* sequences sampled from actual transmission cases. A neighbour-joining tree reconstruction using the case strains and several control sequences led the authors to conclude that *pol* sequences alone did not provide enough information to clarify the relationship between transmission. In a reply to this study, Jenwitheesuk and Liu [61] questioned the computational protocols, suggesting that the conclusions would be different if sequences were analysed by other methods of phylogenetic reconstruction (i.e. maximum likelihood or maximum parsimony). In addition, they proposed that the tree reconstruction should be repeated with different seed numbers. In a counter-response, Stürmer *et al.* [62] reanalysed the data using different phylogenetic methods (neighbour-joining, minimum evolution, unweighted pair group, maximum parsimony and maximum likelihood). Despite additional analysis, the results and conclusions were the same. Corroborating these results, Lemey *et al.* tested the efficiency of *pol* compared against gp41 (*env*) to reconstruct well documented transmission events. The phylogenetic analysis for *pol* inferred an evolutionary history not fully compatible with the real transmission history [74]. However, according to this study, the gp41 alignment performed well and was sufficient to detect the transmission cluster.

Aside from the *pol* performance debate, some studies have focused on comparing the adequacy of *gag* and *env* in reconstructing transmission trees. Leitner *et al.* [6] explored the contribution of p17 (*gag*) and V3 (*env*), using a set of HIV-1 sequences from infected individuals with known epidemiological relationships. They showed that combining data on p17 (438 bp) and V3 (285 bp) performed better than data on either p17 or V3 evaluated separately (despite a topology error when compared with the true transmission tree). In addition, they found that the choice of gene fragment was more important than the choice of phylogenetic method and substitution model. In a similar study, Paraskevis *et al.* [75] isolated CRF04_cpx viral strains from six HIV-1-infected individuals with a known history of viral transmissions. Most of the infections occurred in a short period of time, and viral sequences were isolated distantly from infection dates. After amplification and sequencing, maximum likelihood and Bayesian trees were inferred for three different data sets: p17 (*gag*, 400 bp), *env* (C2–C4, 660 bp) and *gag* + *env* (1060 bp). Unlike partial *gag* and *env*, the combined *gag* + *env* alignment resulted in an improved estimate of the transmission history (followed by *env* and *gag*, respectively). In addition, maximum likelihood and Bayesian methods were highly correlated.

Ultimately, despite the extensive use of *pol* as a main target to reconstruct transmission clusters, it seems that the use of a combination of more than one subgenomic region may improve the identification of transmission clusters. The more genes and the longer the alignment, the better the phylogenetic signal and the better the tree resolution. However, the small number of systematic studies comparing full genomes to individual genes and segments against each other precludes a strong conclusion about how many or the best subgenomic regions to combine in the same analysis.

CONCLUSION

The dynamics of HIV-1 transmission networks can be investigated through comprehensive cluster analysis using phylogenetic and network methods. A wide range of factors may influence the results of phylogenetic analysis. In this review, we have outlined the main factors that can bias HIV phylogenetic clustering methods.

It is clear that the use of a genetic or patristic distance threshold (median or maximum), allied with a branching support cut-off, allows the researcher to identify transmission clusters more accurately and improves the quality of the analysis. High branching thresholds seem to perform better in excluding false transmission links and high genetic distances can capture old transmission events. However, the correct combination of these thresholds is uncertain and is still dependent on the sampling proportion, alignment size and targeted genomic region. Simulation analyses have shown that using a sampling proportion under 10% can bias clustering results by creating spurious links. Contrarily, a proportion above 60% could reproduce the results from a 100% sampling. Moreover, it has been shown that the bigger the alignments, the better the clustering results. Longer sequences provide better phylogenetic signal for clustering identification. Similarly, alignments that include more than one subgenomic region can increase the accuracy of the reconstructed tree. If smaller sequences or single subgenomic regions are used, it is important to narrow the branching support to high confidence values (>90%). In such cases, the decision about the genetic distance can set a parameter for finding old or recent transmission clusters.

Several studies have shown the importance of phylogenetic clustering methods to characterize the risk factors associated with HIV transmission rates within and between populations. These methods are now being applied to monitor the HIV epidemic in near real-time and have become a successful and cost-effective resource for public health intervention in localized outbreaks of HIV transmission. Other techniques, such as network [76] and gap partitioning [77], have also been used to track the dissemination of HIV, but, as with phylogenetic clustering, these methods have limitations. Despite these restrictions, the advantages phylogenetic methods and their potential application to public health are undeniable.

Acknowledgements

None.

Financial support and sponsorship

The work was supported by the National Institute of Health Award (5R01DA043409-02), by the South African Medical Research Council (SA MRC) Flagship Grant (MRC-RFA-UFSP-01-2013/UKZN HIVEPI) and by the UKZN Flagship Grant [Title: Afrocentric Precision Approach to Control Health Epidemics (APACHE)]. T.d.O. received support from a UK Royal Society Newton Advanced Fellowship. KRISP is funded by a core award from the Technology Innovation Agency (TIA) of the Department of Science and Technology (DST).

Conflicts of interest

There are no conflicts of interest.

REFERENCES AND RECOMMENDED READING

Papers of particular interest, published within the annual period of review, have been highlighted as:

■ of special interest
■■ of outstanding interest

1. Gao F, Bailes E, Robertson DL, *et al.* Origin of HIV-1 in the chimpanzee *Pan troglodytes* troglodytes. Nature 1999; 397:436–441.
2. Chang SYPY, Bowman BHH, Weiss JBB, *et al.* The origin of HIV-1 isolate HTLV-IIIB. Nature 1993; 363:466–469.

3. Faria NR, Rambaut A, Suchard MA, *et al.* The early spread and epidemic ignition of HIV-1 in human populations. Science 2014; 346:56–61.

4. Hué S, Pillay D, Clewley JP, Pybus OG. Genetic analysis reveals the complex structure of HIV-1 transmission within defined risk groups. Proc Natl Acad Sci U S A 2005; 102:4425–4429.

5. Brenner BG, Roger M, Moisi DD, *et al.* Transmission networks of drug resistance acquired in primary/early stage HIV infection. AIDS 2008; 22:2509–2515.

6. Leitner T, Escanillat D, Franzent C, *et al.* Accurate reconstruction of a known HIV-1 transmission history by phylogenetic tree analysis. Proc Natl Acad Sci U S A 1996; 93:10864–10869.

7. Rachinger A, Groeneveld PHP, Van Assen S, *et al.* Time-measured phylogenies of gag, pol and env sequence data reveal the direction and time interval of HIV-1 transmission. AIDS 2011; 25:1035–1039.

8. Lemey P, Rambaut A, Pybus OG. HIV evolutionary dynamics within and among hosts. AIDS Rev 2006; 8:125–140.

9. ▪ Hassan AS, Pybus OG, Sanders EJ, *et al.* Defining HIV-1 transmission clusters based on sequence data. AIDS 2017; 31:1211–1222.

The review highlights how HIV-1 transmission clusters can be defined and provide some guidance based on examples from real life datasets.

10. ▪▪ Dearlove BL, Xiang F, Frost SDW. Biased phylodynamic inferences from analysing clusters of viral sequences. Virus Evol 2017; 3:1–10.

The study shows the effect of using a falsely identified transmission cluster of sequences to estimate phylodynamic parameters from trees simulated under several demographical scenarios

11. Centers for Disease Control and Prevention; National Center for HIV/AIDS, Viral Hepatitis, STD and T prevention. Managing HIV and hepatitis C outbreaks among people who inject drugs. Centers for Disease Control and Prevention; 2018.

12. Leigh Brown AJ, Lycett SJ, Weinert L, *et al.* Transmission network parameters estimated from HIV sequences for a nationwide epidemic. J Infect Dis 2011; 204:1463–1469.

13. Brenner BG, Roger M, Routy J, *et al.* High rates of forward transmission events after acute/early HIV-1 infection. J Infect Dis 2007; 195:951–959.

14. 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.

15. Baldauf SL. Phylogeny for the faint of heart: a tutorial. Trends Genet 2003; 19:345–351.

16. Division of HIV/AIDS Prevention & Centers for Disease Control. Detecting, investigating, and responding to HIV transmission clusters. Centers for Disease Control and Prevention; 2018; Available from: http://test.datamonkey.org/hivtrace.

17. Little SJ, Kosakovsky Pond SL, Anderson CM, *et al.* Using HIV networks to inform real time prevention interventions. PLoS One 2014; 9:e98443.

18. Wertheim JO, Kosakovsky Pond SL, Little SJ, De Gruttola V. Using HIV transmission networks to investigate community effects in HIV prevention trials. PLoS One 2011; 6:1–7.

19. Poon AFY, Gustafson R, Daly P, *et al.* Near real-time monitoring of HIV transmission hotspots from routine HIV genotyping: an implementation case study. Lancet HIV 2016; 3:e231–e238.

20. Novitsky V, Moyo S, Lei Q, *et al.* Importance of viral sequence length and number of variable and informative sites in analysis of HIV clustering. AIDS Res Hum Retroviruses 2015; 31:531–542.

21. Harris ME, Maayan S, Kim B, *et al.* A cluster of HIV type 1 subtype C sequences from Ethiopia, observed in full genome analysis, is not sustained in subgenomic regions. AIDS Res Hum Retroviruses 2003; 19:1125–1133.

22. Rokas A, Williams BL, King N, Carroll SB. Genome-scale approaches to resolving incongruence in molecular phylogenies. Nature 2003; 425: 798–804.

23. Rokas A, Chatzimanolis S. From gene-scale to genome-scale phylogenetics: the data flood in, but the challenges remain. Methods Mol Biol 2008; 422:1–12.

24. Jeffroy O, Brinkmann H, Delsuc F, Philippe H. Phylogenomics: the beginning of incongruence? Trends Genet 2006; 22:225–231.

25. Felsenstein J. Cases in which parsimony or compatibility methods will be positively misleading. Syst Zool 1978; 27:401–410.

26. Volz EM, Koelle K, Bedford T. Viral phylodynamics. PLoS Comput Biol 2013; 9:e1002947.

27. Volz EM, Ionides E, Romero-Severson EO, *et al.* HIV-1 transmission during early infection in men who have sex with men: a phylodynamic analysis. PLoS Med 2013; 10:e1001568.

28. Volz EM, Koopman JS, Ward MJ, *et al.* Simple epidemiological dynamics explain phylogenetic clustering of HIV from patients with recent infection. PLoS Comput Biol 2012; 8:2–11.

29. Wertheim JO, Scheffler K, Choi JY, *et al.* Phylogenetic relatedness of HIV-1 donor and recipient populations. J Infect Dis 2013; 207:1181–1182.

30. Brenner BG, Roger M, Routy J-P, *et al.* High rates of forward transmission events after acute/early HIV-1 infection. J Infect Dis 2007; 195:951–959.

31. Junqueira DM, de Medeiros RM, Gräf T, Almeida SEdeM. Short-term dynamic and local epidemiological trends in the South American HIV-1B epidemic. PLoS One 2016; 11:e0156712.

32. DeBry RW, Abele LG, Weiss SH, *et al.* Dental HIV transmission? Nature 1993; 361:691–1691.

33. Strunnikova N, Ray SC, Livingston RA, *et al.* Convergent evolution within the V3 loop domain of human immunodeficiency virus type 1 in association with disease progression. J Virol 1995; 69:7548–7558.

34. Zhang LQ, MacKenzie P, Cleland A, *et al.* Selection for specific sequences in the external envelope protein of human immunodeficiency virus type 1 upon primary infection. J Virol 1993; 67:3345–3356.

35. Watts DJ, Strogatz SH. Collective dynamics of ‘small-world’ networks. Nature 1998; 393:440–442.

36. ▪▪ Poon AFY. Impacts and shortcomings of genetic clustering methods for infectious disease outbreaks. Virus Evol 2016; 2:vew031.

The study applies and compares six genetic methods of clustering detection to trees and sequence alignments simulated under a compartmental epidemic model in a risk-structured population.

37. Lemoine F, Domelevo Entfellner JB, Wilkinson E, *et al.* Renewing Felsenstein’s phylogenetic bootstrap in the era of big data. Nature 2018; 556:452–456.

38. Pommier T, Canbäck B, Lundberg P, *et al.* RAMI: a tool for identification and characterization of phylogenetic clusters in microbial communities. Bioinformatics 2009; 25:736–742.

39. Poon AFY. Phylodynamic inference with kernel ABC and its application to HIV epidemiology. Mol Biol Evol 2015; 32:2483–2495.

40. Bezemer D, van Sighem A, Lukashov VV, *et al.* Transmission networks of HIV-1 among men having sex with men in the Netherlands. AIDS 2010; 24:271–282.

41. Pilon R, Leonard L, Kim J, *et al.* Transmission patterns of HIV and hepatitis C virus among networks of people who inject drugs. PLoS One 2011; 6: 2–8.

42. Cuevas M, Fernandez-Garcia A, Sanchez-Garcia A, *et al.* Incidence of non-B subtypes of HIV-1 in Galicia, Spain: high frequency and diversity of HIV-1 among men who have sex with men. Euro Surveill 2009; 14:; Available from: http://www.ncbi.nlm.nih.gov/pubmed/19941808.

43. Chalmet K, Staelens D, Blot S, *et al.* Epidemiological study of phylogenetic transmission clusters in a local HIV-1 epidemic reveals distinct differences between subtype B and non-B infections. BMC Infect Dis 2010; 10:1–9.

44. Kaye M, Chibo D, Birch C. Phylogenetic investigation of transmission pathways of drug-resistant HIV-1 utilizing pol sequences derived from resistance genotyping. J Acquir Immune Defic Syndr 2008; 49:9–16.

45. Marzel A, Shilaih M, Yang WL, *et al.* HIV-1 transmission during recent infection and during treatment interruptions as major drivers of new infections in the Swiss HIV Cohort Study. Clin Infect Dis 2016; 62:115–122.

46. Pilcher CD, Wong JK, Pillai SK. Inferring HIV transmission dynamics from phylogenetic sequence relationships. PLoS Med 2008; 5:e69.

47. Mehta SR, Kosakovsky Pond SL, Young JA, *et al.* Associations between phylogenetic clustering and HLA profile among HIV-infected individuals in San Diego, California. J Infect Dis 2012; 205:1529–1533.

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

49. Yebra G, Frampton D, Cassarino TG, *et al.* A high HIV-1 strain variability in London, UK, revealed by full-genome analysis: results from the ICONIC project. PLoS One 2018; 13:1–18.

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

51. Bezemer D, Cori A, Ratmann O, *et al.* Dispersion of the HIV-1 epidemic in men who have sex with men in the Netherlands: a combined mathematical model and phylogenetic analysis. PLoS Med 2015; 12:e1001898.

52. Felsenstein J. Confidence limits on phylogenies: an approach using the bootstrap. Evolution 1985; 39:783–791.

53. Jacka B, Applegate T, Poon AF, *et al.* Older people who inject drugs in Vancouver. Canada 2017; 64:1247–1255.

54. Kouyos RD, Von Wyl V, Yerly S, *et al.* Molecular epidemiology reveals long-term changes in HIV type 1 subtype B transmission in Switzerland. Epidemiology 2010; 2009:8–11.

55. Vrancken B, Adachi D, Benedet M, *et al.* The multifaceted dynamics of HIV-1 transmission in Northern Alberta: a combined analysis of virus genetic and public health data. Infect Genet Evol 2017; 52:100–105.

56. Sallam M, Esbjörnsson J, Baldvinsdóttir G, *et al.* Molecular epidemiology of HIV-1 in Iceland: early introductions, transmission dynamics and recent outbreaks among injection drug users. Infect Genet Evol 2017; 49:157–163.

57. Wolf E, Herbeck JT, Van Rompaey S, *et al.* Short communication: phylogenetic evidence of HIV-1 transmission between adult and adolescent men who have sex with men. AIDS Res Hum Retroviruses 2017; 33:318–322.

58. ▪ Ragonnet-Cronin ML, Shilaih M, Günthard HF, *et al.* A direct comparison of two densely sampled HIV epidemics: the UK and Switzerland. Sci Rep 2016; 6:1–9.

The study compares the UK HIV Drug Resistance Database and the Swiss HIV Cohort Study using phylogenetic clustering approaches.

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

60. Novitsky V, Moyo S, Lei Q, *et al.* Impact of sampling density on the extent of HIV clustering. AIDS Res Hum Retroviruses 2014; 30:1–10.

61. Jenwitheesuk E, Liu T. Single phylogenetic reconstruction method is insufficient to clarify relationships between patient isolates in HIV-1 transmission case. AIDS 2005; 19:743–744.

62. Stürmer M, Preiser W, Gute P, *et al.* Response to 'Single phylogenetic reconstruction method is insufficient to clarify relationships between patient isolates in HIV-1 transmission case' by Jenwitheesuk and Liu. AIDS 2005; 19:741–743; author reply 743-4.

63. Arnold C, Balfe P, Clewley JP. Sequence distances between env genes of HIV-1 from individuals infected from the same source: implications for the investigation of possible transmission events. Virology 1995; 211:198–203.

64. Hayman A, Moss T, Simmons G, *et al.* Phylogenetic analysis of multiple heterosexual transmission events involving subtype B of HIV type 1. AIDS Res Hum Retroviruses 2001; 17:689–695.

65. Ou C-Y, Ciesielski CA, Myers G, *et al.* Molecular epidemiology of HIV transmission in a dental practice. Science 1992; 256:1165–1171.

66. Lees JA, Kendall M, Parkhill J, *et al.* Evaluation of phylogenetic reconstruction methods using bacterial whole genomes: a simulation based study. Wellcome Open Res 2018; 3:33.

67. Philippe H, Brinkmann H, Lavrov DV, *et al.* Resolving difficult phylogenetic questions: why more sequences are not enough. PLoS Biol 2011; 9:e1000602.

68. Lemey P, Vandamme A-M. Exploring full-genome sequences for phylogenetic support of HIV-1 transmission events. AIDS 2005; 19:1551–1552.

69. Yebra G, Hodcroft EB, Ragonnet-cronin ML, *et al.* Using nearly full-genome HIV sequence data improves phylogeny reconstruction in a simulated epidemic. Sci Rep 2016; 6:39489.

70. Albert J, Wahlberg J, Leitner T, *et al.* Analysis of a rape case by direct sequencing of the human immunodeficiency virus type 1 pol and gag genes. J Virol 1994; 68:5918–5924.

71. Nabhan AR, Sarkar IN. The impact of taxon sampling on phylogenetic inference: a review of two decades of controversy. Brief Bioinform 2012; 13:122–134.

72. Degnan JH, Rosenberg NA. Gene tree discordance, phylogenetic inference and the multispecies coalescent. Trends Ecol Evol 2009; 24:332–340.

73. Stürmer M, Preiser W, Gute P, *et al.* Phylogenetic analysis of HIV-1 transmission: pol gene sequences are insufficient to clarify true relationships between patient isolates. AIDS 2004; 18:2109–2113.

74. Lemey P, Derdelinckx I, Rambaut A, *et al.* Molecular footprint of drug-selective pressure in a human immunodeficiency virus transmission chain molecular footprint of drug-selective pressure in a human immunodeficiency virus transmission chain. J Virol 2005; 79:11981–11989.

75. Paraskevis D, Magiorkinis E, Magiorkinis G, *et al.* Phylogenetic reconstruction of a known HIV-1 CRF04_cpx transmission network using maximum likelihood and Bayesian methods. J Mol Evol 2004; 59:709–717.

76. Wertheim JO, Leigh Brown AJ, Hepler NL, *et al.* The global transmission network of HIV-1. J Infect Dis 2014; 209:304–313.

77. Vrbik I, Stephens DA, Roger M, Brenner BG. The Gap Procedure: For the identification of phylogenetic clusters in HIV-1 sequence data. BMC Bioinformatics 2015; 16:1–9.