Persistence of HIV transmission clusters among people who inject drugs

Rebecca Rose[a], Sissy Cross[a], Susanna L. Lamers[a], Jacquie Astemborski[b], Greg D. Kirk[b], Shruti H. Mehta[b], Matthew Sievers[c], Craig Martens[e], Daniel Bruno[e], Andrew D. Redd[c,d] and Oliver Laeyendecker[c,d]

Objective: We investigated the duration of HIV transmission clusters.

Design: Fifty-four individuals newly infected at enrollment in the ALIVE cohort were included, all of whom had sequences at an intake visit (T1) and from a second (T2) and/or a third (T3) follow-up visit, median 2.9 and 5.4 years later, respectively.

Methods: Sequences were generated using the 454 DNA sequencing platform for portions of HIV *pol* and *env* (HXB2 positions 2717–3230; 7941–8264). Genetic distances were calculated using *tn93* and sequences were clustered over a range of thresholds (1–5%) using HIV-TRACE. Analyses were performed separately for individuals with *pol* sequences for T1 + T2 (*n* = 40, 'Set 1') and T1 + T3 (*n* = 25; 'Set 2'), and *env* sequences for T1 + T2 (*n* = 47, 'Set 1'), and T1 + T3 (*n* = 30; 'Set 2').

Results: For *pol*, with one exception, a single cluster contained more than 75% of samples at all thresholds, and cluster composition was at least 90% concordant between time points/thresholds. For *env*, two major clusters (A and B) were observed at T1 and T2/T3, although cluster composition concordance between time points/thresholds was low (<60%) at lower thresholds for both sets 1 and 2. In addition, several individuals were included in clusters at T2/T3, although not at T1.

Conclusion: Caution should be used in applying a single threshold in population studies where seroconversion dates are unknown. However, the retention of some clusters even after 5 + years is evidence for the robustness of the clustering approach in general.

AIDS 2020, **34**:2037–2044

Keywords: epidemic, genetic linkage, intravenous drug use, next-generation sequencing

Introduction

Identification of transmission clusters using HIV sequence data is a standard tool of molecular epidemiological studies [1–4] due in part to the computational efficiency and ease of use [5]. Cluster analysis is particularly useful when analyzing large datasets, for example, those generated from deep-sequencing technologies or country-wide HIV sequence databases [6]. Cluster analysis can reveal epidemiologically important trends, help identify outbreaks, guide prevention strategies, and uncover significant risk factors for infection [7]. Clustering methods are based on the underlying concept that genetically similar viruses likely shared a recent epidemiological history (e.g. direct transmission, derivation from the same source, or part of the same transmission chain) [5].

Interpretation of cluster dynamics is complicated by the observation that the same distance threshold, used with different methods (e.g. genetic distance [8] vs. tree-based

[a]Bioinfoexperts, LLC, Thibodaux, Lousiana, [b]Bloomberg School of Public Health, [c]Department of Medicine, Johns Hopkins University, [d]Division of Intramural Research, NIAID, NIH, Baltimore, Maryland, and [e]Genomics Unit, Research Technologies Branch, Rocky Mountain Laboratories, Division of Intramural Research NIAID, NIH, Hamilton, Montana, USA.

Correspondence to Rebecca Rose, PhD, 718 Bayou Ln, Thibodaux, LA 70301, USA.

Tel: +1 757 773 7230; e-mail: Rebecca.rose@bioinfox.com

Received: 30 April 2020; revised: 15 July 2020; accepted: 27 July 2020.

DOI:10.1097/QAD.0000000000002662

ISSN 0269-9370

[9,10]), may result in different cluster compositions and capture widely different time spans [11]. Furthermore, samples close to the time of seroconversion are more likely to cluster than samples from well after infection [5,12]. As many cross-sectional and longitudinal population studies include individuals with unknown dates of seroconversion, clustering risk factors may, therefore, erroneously be identified [5] unless explicitly accounted for [1,12,13]. A combination of the two approaches may provide additional epidemiological insight [5].

A threshold for genetic distance can be defined as the median, mean, or maximum genetic distance allowed for any two pairs of sequences within a cluster. Alternatively, 'single linkage' methods can be used, where a sequence is included in a cluster if the distance between it and any other sequence is below the threshold [10]. However, some of these methods (e.g. mean and single linkage) can result in large clusters where sequences are 'chained' together (e.g. sequence A is related to sequence B, B is related to C, C is related to D) resulting in large distances above the threshold for many pairs (e.g. between A and D) [14].

The genetic distance threshold chosen to define clusters clearly has a strong impact on the resulting clusters [4,15,16]. Objective methods to select the appropriate distance threshold include using *a priori* knowledge of either the distance between epidemiologically confirmed transmission pairs or the intrahost evolutionary rate of samples in the dataset under study [5]. The optimal threshold is the value, which most accurately separates known linked vs. nonlinked sequences, which will change among datasets depending on the pathogen, gene region, and sample composition [17]. In cases lacking sufficient data for these approaches, conservative genetic distance thresholds for the *pol* gene of 1–2% are commonly used [2,4,5,18], consistent with expected intrahost evolution of this region [4,13,19]. Although fewer studies have used the *env* gene to infer cluster dynamics, a threshold of up to 5.3% was found to correlate with known epidemiological linkage [17,20], consistent with the higher evolutionary rate in this gene [21]. Using a range of threshold values can provide insights that are lost when choosing only one value. For example, lower thresholds will capture the most closely related sequences (often interpreted as the leading edge of an epidemic) [18], although higher thresholds will capture mature epidemics and chronically infected individuals [10]. Furthermore, the mode of transmission [i.e. intravenous drug use (IDU) vs. sexual] can impact transmission network patterns and may require different thresholds [22].

Here, we investigate another approach for determining appropriate thresholds by using sequence data from longitudinally sampled individuals who were all recently infected at enrollment. Individuals were identified as being recently infected by a multiserology assay algorithm validated in an HIV-1 clade B setting that included members of the cohort analyzed in this article [23]. Our objective was to assess the persistence of transmission clusters detected soon after transmission at two follow-up time points (up to 7 years later). High throughput sequencing (HTS) was generated for *pol* and *env*, providing high resolution for determining linkage among individuals and the ability to compare the signal between genes.

Methods

Participants

Individuals were selected from the AIDS Linked to the IntraVenous Experience (ALIVE) cohort, one of the longest running community-based cohorts of people who inject drugs (PWID) prospectively followed from 1988. At the conception of the ALIVE cohort, individuals enrolled were 89% black and 81% men [24]. All individuals were identified as recently infected (<6 months) at enrollment using serological testing as follows: samples from individuals identified as recently infected had a BED-capture EIA value less than 1.0, a BioRad avidity index less than 80%, a viral load above 400 copies/ml and a $CD4^{+}$ cell count greater than 200 cells/μl [25]. A total of 54 individuals were included in the present study, all of whom had sequences at an intake visit (T1) and from a second (T2) and/or a third (T3) follow-up visit, which took place a median of 2.9 and 5.4 years later, respectively. Most of the 54 individuals ($n=45$) were enrolled in the period from February 1988 to April 1989; eight were enrolled between December 1991 and October 1999; and two were enrolled in 2006 (Table S1, http://links.lww.com/QAD/B819). Four individuals were treated with antiretroviral therapy during the course of the study (Table S1, http://links.lww.com/QAD/B819).

Generation of sequence data

Sample extraction and amplification was performed as previously described [26]. Briefly, RNA was extracted from 140 μl plasma. RT-PCR was used to generate portions of *pol* and *env*. PCR products were sequenced using the 454 DNA Sequencing platform (Roche, Branford, Connecticut, USA) as previously described [27]. Sequence data were cleaned and analysis was performed on portions of HIV *pol* (HXB2 position 2717−3230) and *env* (HXB2 position 7941−8264). To determine the HIV subtype, *env* sequences from all individuals were grouped at a 95% similarity threshold using USEARCH [28], and the centroid sequence from each group was subtyped using the REGA HIV-1 Subtyping Tool (version 3.0) available at https://hivdb.stanford.edu/.

Genetic distances

Genetic distances were calculated using the tn93 program (https://github.com/veg/tn93). A receiver operating

Table 1. Number of individuals with sequences in each dataset.

Number of individuals	*pol* Set 1	*pol* Set 2	*env* Set 1	*env* Set 2
20	X	X	x	x
3	X		x	x
4			x	x
2	x	X	x	
10	x		x	
3	x	X		
1	x			x
8			x	
2				x
1	x			
54	**40**	**25**	**47**	**30**

X indicates that the individual has data for that dataset. Totals are given in bold at the bottom.

curve (ROC) analysis was used to determine the optimal threshold to differentiate self vs. nonself using R package *pROC*. Sequences were clustered over a range of thresholds (1–5%) using HIV-TRACE [8], which uses the 'single linkage method' described above. For the purposes of this study, clusters were defined as containing sequences from more than one individual.

Results

Sequences generated

A total of 28 427 (*pol*) and 9548 (*env*) unique consensus sequences were generated representing 995 423 (*pol*) and 339 707 (*env*) sequence reads, which covered the entire region for both genes. Analyses were performed separately for *pol* and *env*. As some individuals did not have sequences for all three visits, sequences for each gene were grouped into sets as follows: individuals with *pol* sequences for T1 and T2 ('*pol* Set 1'), individuals with *env* sequences for T1 and T2 ('*env* Set 1'), individuals with *pol* sequences for T1 and T3 ('*pol* Set 2'), and individuals with *env* sequences for T1 and T3 ('*env* Set 2') (Table 1, Supplemental Table 1, http://links.lww.com/QAD/B819). All sequences were classified as HIV subtype B.

Optimized threshold based on intrahost and interhost comparisons

For *pol*, the ROC-determined optimal distance threshold for Set 1 individuals was 1% at T1 and 1.6% at T2. For *pol* Set 2 individuals, the optimal distance was also 1% at T1, and 2.2% at T3. For *env*, the distance threshold that optimally separated intrahost vs. interhost comparisons for Set 1 individuals was 2.2% for T1 and 3.8% for T2. For *env* Set 2 individuals, the optimal distance for T1 was 1.3%, and 5.1% for T3.

Cluster composition for *pol* Set 1

For the 40 individuals with *pol* sequence data at T1 and T2 (Set 1), for T1, a single cluster contained 31 individuals at 1% (Fig. 1a, Supplemental Figure 1a, http://links.lww.com/QAD/B815). An additional four individuals were added at 2% ($n = 35$) and another three at 3% ($n = 38$). At 4 and 5%, all 40 individuals were in a single cluster. For T2, a single cluster was also observed at

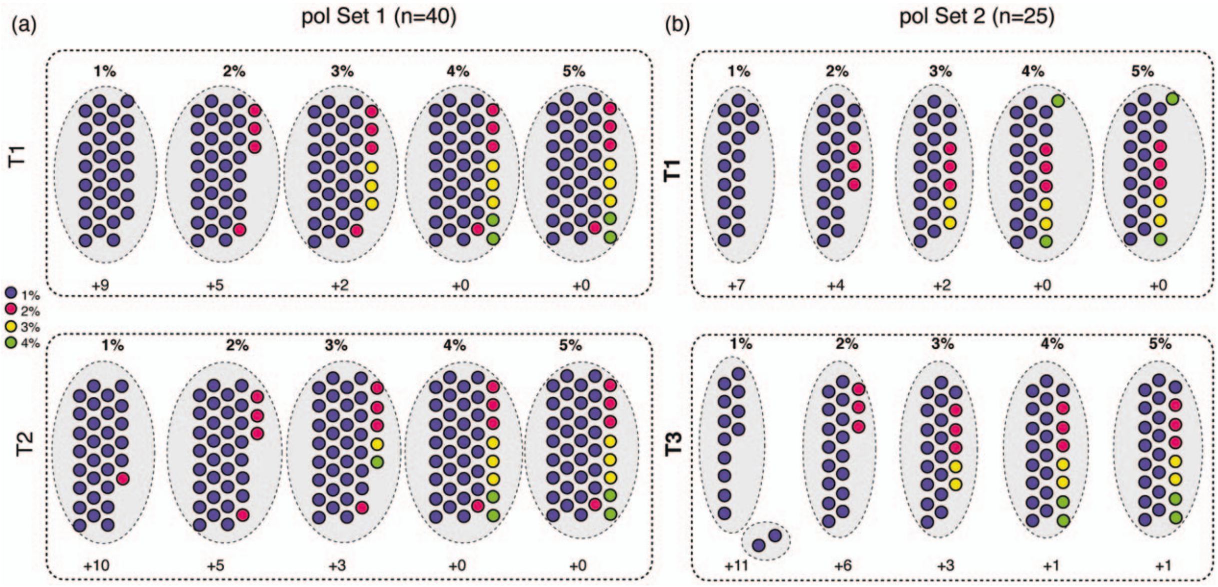

Fig. 1. Cluster composition for composition for *pol* sequences for (a) Set 1 and (b) Set 2. Filled circles represent samples, which are colored according to their cluster designation at T1 for both Sets, as specified in the legend, and retained the same color for the later time point. Larger grey dotted-line circles indicate clusters at the genetic distance thresholds (1–5%) listed across the top for each analysis. Time points for each set are denoted with dotted box: (a) T1, top; T2, bottom; (b) T1, top; T3, bottom. The number of samples that did not cluster is indicated at the bottom of each time point (+*n*).

each threshold, with the composition being similar (although slightly smaller) than those at T1 (Fig. 1a, Supplemental Figure 1b, http://links.lww.com/QAD/B815).

To determine the cluster persistence between time points, we calculated, for each threshold, the number of individuals in the T2 cluster who were also in the cluster at T1, as a proportion of the total number of individuals in the cluster at T1. The results were as follows: 1%: 29/31 (0.94); 2%: 35/35 (1.0); 3%: 36/38 (0.95); 4 and 5%: 40/40 (1.0). Note that at 1 and 3%, the T2 cluster contained one individual who was not part of the T1 cluster at the corresponding threshold.

Cluster composition for *pol* Set 2

For the 25 individuals with *pol* sequence data at T1 and T3 (*pol* Set 2), for T1 a single cluster contained 18 individuals at 1%, which increased by three individuals at 2% ($n = 21$), another two individuals at 3% (n = 23), and the remaining two individuals at 4 and 5% ($n = 25$, Fig. 1b, Supplemental Figure 1c, http://links.lww.com/QAD/B815). All individuals were in the cluster at the highest two thresholds.

For T3, two clusters were detected at 1%, containing 12 and two individuals, respectively (Fig. 1b); however, one individual in the small cluster also had sequences in the larger cluster as well (Supplemental Figure 1d, http://links.lww.com/QAD/B815). At 2%, individuals in both clusters condensed into the single cluster, with seven additional individuals ($n = 19$). At 3%, an additional three individuals were included ($n = 22$), and at 4 and 5%, two additional individuals were included ($n = 24$). One individual did not cluster at all, even at the highest threshold.

For each threshold, the number of individuals in the T3 cluster who were also in the cluster at T1, as a proportion of the total number of individuals in the cluster at T1, were as follows: 1%: 12/18 (0.66); 2%: 19/21 (0.90); 3%: 22/23 (0.96); 4 and 5%: 24/25 (0.96).

Cluster composition for *env* Set 1

For the 47 individuals in Set 1, individuals were broadly grouped into two large clusters (group A and group B) at the lower thresholds (1–3%, Fig. 2a, Supplemental Figure 2a, http://links.lww.com/QAD/B816). For T1 at 1%, two small group A clusters (A1: $n = 5$; A2: $n = 6$) and two small group B clusters (B1: $n = 2$; B2: $n = 3$) were observed. At 2%, clusters A1 and A2 merged with an additional eight individuals, and B1 and B2 merged with an additional five individuals. At 3%, the single group A cluster gained two individuals and the single group B cluster remained the same. At 4%, the group A and group B clusters merged into a single cluster (with an additional three individuals), which then gained another three individuals at 5%. Nine individuals did not cluster at all, even at the highest threshold.

For T2, at the 1% threshold, a single cluster contained a subset ($n = 4$) of the A2 cluster (Fig. 2a, Supplemental Figure 2b, http://links.lww.com/QAD/B816). At 2%, a total of four clusters contained: a subset of A1 ($n = 3$); all of A2 ($n = 6$), with one individual from A1 and two individuals from the larger group A; one individual from group A and one non-A/B individual; and all of B1 ($n = 2$), a subset of B2 ($n = 1$), and one additional individual from group B. At 3%, the single group A cluster now constituted all of A1, all of A2, nine additional group A individuals, one non-A/B individual, plus another individual who was not detected in any cluster at the first time point. The single group B cluster contained all of B1, all of B2, with four additional group B individuals and one non-A/B individual. At 4%, all individuals merged into a single cluster, with three additional individuals. At 5%, two additional individuals were added who were present at in the large single cluster at T1, with another two individuals who were not.

Comparing the cluster composition between time points at the same distance threshold was not as straightforward for *env* as with *pol* as there were multiple clusters at each threshold. We, therefore calculated, for each threshold, the number of individuals in any cluster at T2 who were also in any cluster at T1, as a proportion of the total number of individuals in a cluster at T1. The proportions were as follows: 1%: 4/16 (0.25); 2%: 16/29 (0.55); 3%: 29/31 (0.94); 4%: 35/35 (1.0); 5%: 38/38 (1.0). (Note that some individuals in T2 clusters were not included as they did not appear in T1 clusters at the corresponding threshold.)

For the ROC-defined thresholds (rounded to the nearest whole number), two clusters were found for T1 at 2%, and only one cluster for T2 at 4%.

Cluster composition for *env* Set 2

For the 30 individuals in Set 2, individuals also clustered into two major groups (denoted A* and B*; Fig. 2b, Supplemental Figure 2c, http://links.lww.com/QAD/B816). Individuals in group A* and group B* clusters were largely a subset of those in the group A and B clusters from the Set 1 analysis. For T1 at 1%, two clusters were observed: A1* ($n = 4$) and A2* ($n = 4$). At 2%, two clusters contained: all of A1*, all of A2*, with three other group A* individuals and five B* individuals. At 3%, the group A* cluster remained unchanged, whereas the group B* cluster gained two individuals. At 4%, both clusters merged into a single cluster, with an additional two individuals. At 5%, the single cluster added another five individuals. Five individuals remained unclustered at the highest threshold.

For T3, no clusters were observed at 1 or 2% (Fig. 2b, Supplemental Figure 2d, http://links.lww.com/QAD/B816). At 3%, four clusters were observed: all of A2*; a subset of A1* (n = 2); one individual from A1*, with an

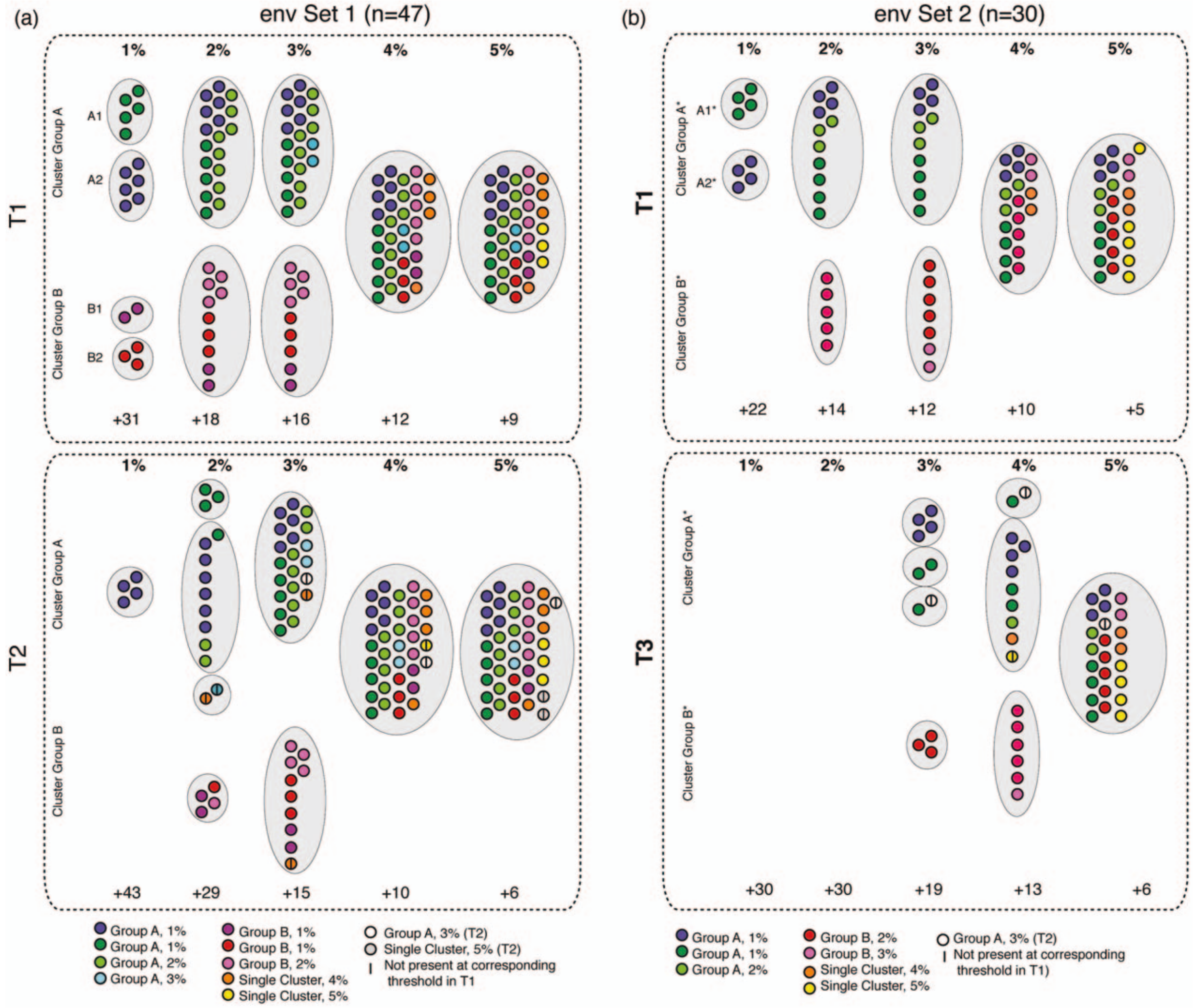

Fig. 2. Cluster composition for composition for *env* sequences for (a) Set 1 and (b) Set 2. Filled circles represent samples; larger grey dotted-line circles indicate clusters at the genetic distance thresholds (1–5%) listed across the top for each analysis. Time points for each set are denoted with dotted box: (a) T1, top; T2, bottom; (b) T1, top; T3, bottom. Samples are colored according to their cluster designation at T1 for both sets, as specified in the legend, and retain the same color for the later time point. Samples only appearing at the later time point are denoted in the legend. The line indicates samples that were not present at the earlier time point for the same threshold. Clusters are labeled (A, B, A*, B*) as described in the text. The number of samples that did not cluster is indicated at the bottom of each time point (+*n*).

additional individual not observed in any cluster at T1; and three cluster group B* individuals. At 4%, three clusters were observed: one individual from A1* with an additional individual not observed in any cluster at T1; all of A2* ($n = 4$), a subset of A1* ($n = 2$), one additional group A* individual, and two non-A*/B* individuals; and six group B* individuals. At 5%, the group A* and B* clusters merged, and included all of A1* ($n = 4$); all of A2* ($n = 4$), two group A* individuals, one individual not observed at T1, seven Cluster B* individuals, and an additional four non-A*/B* individuals. Six individuals remained unclustered at the highest threshold.

For each threshold, the number of individuals in any cluster at T2 who were also in any cluster at T1, as a proportion of the total number of individuals in a cluster at T1, were as follows: 1%: 0/8 (0); 2%: 0/16 (0); 3%: 10/19 (0.53); 4%: 15/20 (0.75); 5%: 23/25 (0.92).

For the ROC-defined thresholds (rounded to the nearest whole number), two clusters were found for T1 at 1%, and one cluster for T3 at 5%.

Time of samples

We then sought to determine if the time between visits was associated with clustering patterns in *env* (Supplemental Figure 3, http://links.lww.com/QAD/B817). In general, the times between T1 and T2 and T1 and T3 were similar for the individuals in small clusters (A1, A2, B1, and B2) compared with the rest of the individuals.

The exceptions were for *env* Set 1, where the time between T1 and T2 was well below the interquartile range for two A1 and one A2 individuals. These three samples did not have a T3 sample. Finally, we considered whether sampling year was associated with clustering patterns in *env* (Supplemental Figure 4, http://links.lww.com/QAD/B818). Of the 11 individuals who did not cluster in at least one *env* analysis, five were initially sampled in 1994 or later (out of a total of seven), whereas six were initially sampled from 1998 to 1998 (of a total of 41 individuals).

Discussion

Here, we investigated the duration of cluster composition among 54 individuals who were newly infected at study enrollment of the ALIVE cohort using HTS for the *pol* and *env* genes. Our goal was to determine whether transmission clusters present at the first visit (T1) were detectable at the second (T2) and third (T3) visits.

We found a single cluster for *pol* at nearly all time points/thresholds, which likely reflects the shared geographic location, risk group, and early sampling dates of the individuals. However, it is somewhat surprising that 5+ years later, the cluster composition was basically unchanged. This may be explained in part by the single linkage clustering algorithm itself, which adds a sequence to a cluster if the distance between it and at least one other sequence in the cluster is less than the distance threshold (as opposed to the distance between the query sequences and all other sequences being less than the threshold). This can result in clusters containing pairs of sequences separated by distances higher than the threshold. HTS data may exaggerate this behavior as each individual has multiple sequences, rather than just a single consensus. Further investigation of this point may clarify the issue [18]. In this study, the majority of individuals ($n = 47$) provided all samples prior to the introduction of HAART in 1996, and therefore the impact of drug-resistant mutations in *pol* is not expected to impact the results.

For *env*, the ROC-defined thresholds of self vs. nonself also increased with time, although possibly more rapidly than expected given the intrahost *env* evolutionary rate of ~0.005 substitutions/site/year [29,30]. Interestingly, at the ROC-defined thresholds at T2 and T3, only a single large cluster was observed. When comparing cluster composition at the same thresholds, the number of individuals in any cluster at T2 who were also in a cluster at T1, as a proportion of the total number of individuals in a cluster at T1, were highest at the 4–5% thresholds. This behavior is somewhat expected: low thresholds are expected to capture more recent transmission events, and the accumulated divergence among individuals over time will eventually exceed lower diversity thresholds. On the other hand, the clusters at 4–5% also contained the majority of individuals. These results suggest that the use of a single threshold may obscure epidemiologically relevant information.

The composition of the two large clusters (A and B) at T1 in *env* was generally maintained at T2 and T3. However, an entirely new cluster of two individuals was observed at both T2 and T3 that was not present at T1, and individuals who did not cluster at all at T1 were found in clusters at T2. These results are somewhat surprising: while the breakdown of clusters would be expected over time as more lineage-specific mutations accumulate in each individual, the *addition* of individuals into clusters (and the generation of new clusters entirely) at later time points is unexpected as this implies that related sequences accumulate more diversity over time than nonrelated sequences over time. Again, this may have resulted from the single linkage clustering method and/or the use of HTS data. Alternatively, the observed ‘new’ clusters at later time points may actually represent true transmission clusters. Intravenous drug use allows more virus to be transmitted than sexual transmission and the possibility of multiple founder populations [31]. Although HTS is expected to capture greater diversity than older Sanger-sequencing method, it is possible that unsampled lineages at the first time point obscured true transmission events [32].

The difference in clustering patterns between *pol* and *env* are not unexpected, as different regions of the HIV genome are individual to different selective pressures. For example, *env* diversification resulting from host immune pressure and changes in tropism is expected to contribute to a higher substitution rate than *pol*, and recombination can further unlink transmission histories among HIV genes [33]. The use of multiple regions may, therefore, provide enhanced ability to recover epidemiolocal linkages [34].

These results are consistent with those reported by Redd *et al.* [35], which analyzed a subset ($n = 23$) of the individuals in the present study. Redd and co-authors used HIV-TRACE at a single distance threshold (2%) to cluster HTS sequences from the first time point. They found three distinct clusters, which corresponded to the clusters defined here as group A2, group A1/larger group A, and larger group B, respectively. They confirmed these results using single genome amplification of the same samples. The inexact congruence of cluster assignment between the two studies is very likely because of the inclusion of additional individuals in the present analysis. Here, a lower distance threshold (1%) was necessary to obtain similar resolution as found in the earlier study [35], suggesting that that the choice of distance threshold might be further conditioned on the number of individuals included in the analysis.

This study has a several limitations. Epidemiological confirmation of transmission pairs was not available, so the genetic clusters could not be confirmed. As the methods used here did not account for unsampled individuals, and the individuals included in this study were only a small subsample of the PWID epidemic in Baltimore at the time, it is likely that clustered individuals were not involved in direct transmission events. A strength of the study was the use of high-throughput sequencing, which provided much greater depth of diversity than bulk Sanger sequencing. Follow-up studies using phylogenetic methods would provide additional insight into the maintenance of clustering signal over time. Overall, our results suggest caution in equating clusters with transmission events, particularly at higher thresholds. However, the fact that some clustering patterns were retained after 5+ years of divergence, provides confidence for the use of these methods in studies where seroconversion times are unknown.

Acknowledgements

Funding sources: Support was provided by the Division of Intramural Research, NIAID.

Author roles: R.R. and O.L. designed the study. J.A., G.D.K., S.H.M., M.S., C.M., D.B., A.D.D., and O.L. collected samples and generated data. R.R., S.C., S.L.L., and O.L. analyzed data. R.R., A.D.D., and O.L. wrote the manuscript.

Conflicts of interest

There are no conflicts of interest.

References

1. Leigh Brown AJ, Lycett SJ, Weinert L, Hughes GJ, Fearnhill E, Dunn DT, HIV Drug Resistance Collaboration. **Transmission network parameters estimated from HIV sequences for a nationwide epidemic.** *J Infect Dis* 2011; **204**:1463–1469.
2. Wertheim JO, Oster AM, Hernandez AL, Saduvala N, Bañez Ocfemia MC, Hall HI. **The international dimension of the U.S. HIV Transmission Network and onward transmission of HIV recently imported into the United States.** *AIDS Res Hum Retroviruses* 2016; **32**:1046–1053.
3. Wertheim JO, Leigh Brown AJ, Hepler NL, Mehta SR, Richman DD, Smith DM, Kosakovsky Pond SL. **The global transmission network of HIV-1.** *J Infect Dis* 2014; **209**:304–313.
4. Wertheim JO, Kosakovsky Pond SL, Forgione LA, Mehta SR, Murrell B, Shah S, *et al*. **Social and genetic networks of HIV-1 transmission in New York City.** *PLoS Pathog* 2017; **13**:e1006000.
5. Le Vu S, Ratmann O, Delpech V, Brown A, Gill O, Tostevin A, *et al*. **Comparison of cluster-based and source-attribution methods for estimating transmission risk using large HIV sequence databases.** *Epidemics* 2018; **23**:1–10.
6. Dennis AM, Herbeck JT, Brown AL, Kellam P, de Oliveira T, Pillay D, *et al*. **Phylogenetic studies of transmission dynamics in generalized HIV epidemics: an essential tool where the burden is greatest?** *J Acquir Immune Defic Syndr* 2014; **67**:181–195.
7. Grabowski MK, Herbeck JT, Poon AFY. **Genetic cluster analysis for HIV prevention.** *Curr HIV/AIDS Rep* 2018; **15**:182–189.
8. Kosakovsky Pond SL, Weaver S, Leigh Brown AJ, Wertheim JO. **HIV-TRACE (Transmission Cluster Engine): a tool for large scale molecular epidemiology of HIV-1 and other rapidly evolving pathogens.** *Mol Biol Evol* 2018; **35**:1812–1819.
9. Norström MM, Prosperi MC, Gray RR, Karlsson AC, Salemi M. **PhyloTempo: a set of r scripts for assessing and visualizing temporal clustering in genealogies inferred from serially sampled viral sequences.** *Evol Bioinform Online* 2012; **8**:261–269.
10. Ragonnet-Cronin M, Hodcroft E, Hué S, Fearnhill E, Delpech V, Brown AJ, *et al*., UK HIV Drug Resistance Database. **Automated analysis of phylogenetic clusters.** *BMC Bioinformatics* 2013; **14**:317.
11. Gibson KM, Steiner MC, Kassaye S, Maldarelli F, Grossman Z, Pérez-Losada M, Crandall KA. **A 28-Year History of HIV-1 Drug Resistance and Transmission in Washington, DC.** *Front Microbiol* 2019; **10**:369.
12. Poon A. **Impacts and shortcomings of genetic clustering methods for infectious disease outbreaks.** *Virus Evol* 2016; **2**:vew031.
13. Poon AF, Joy JB, Woods CK, Shurgold S, Colley G, Brumme CJ, *et al*. **The impact of clinical, demographic and risk factors on rates of HIV transmission: a population-based phylogenetic analysis in British Columbia, Canada.** *J Infect Dis* 2015; **211**:926–935.
14. Aldous JL, Pond SK, Poon A, Jain S, Qin H, Kahn JS, *et al*. **Characterizing HIV transmission networks across the United States.** *Clin Infect Dis* 2012; **55**:1135–1143.
15. Ragonnet-Cronin M, Ofner-Agostini M, Merks H, Pilon R, Rekart M, Archibald CP, *et al*. **Longitudinal phylogenetic surveillance identifies distinct patterns of cluster dynamics.** *J Acquir Immune Defic Syndr* 2010; **55**:102–108.
16. 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.
17. Rose R, Lamers SL, Dollar JJ, Grabowski MK, Hodcroft EB, Ragonnet-Cronin M, *et al*. **Identifying transmission clusters with cluster picker and HIV-TRACE.** *AIDS Res Hum Retroviruses* 2017; **33**:211–218.
18. Hassan AS, Pybus OG, Sanders EJ, Albert J, Esbjörnsson J. **Defining HIV-1 transmission clusters based on sequence data.** *AIDS* 2017; **31**:1211–1222.
19. Hightower GK, May SJ, Pérez-Santiago J, Pacold ME, Wagner GA, Little SJ, *et al*. **HIV-1 clade B pol evolution following primary infection.** *PLoS One* 2013; **8**:e68188.
20. Eshleman SH, Hudelson SE, Redd AD, Wang L, Debes R, Chen YQ, *et al*. **Analysis of genetic linkage of HIV from couples enrolled in the HIV Prevention Trials Network 052 trial.** *J Infect Dis* 2011; **204**:1918–1926.
21. Alizon S, Fraser C. **Within-host and between-host evolutionary rates across the HIV-1 genome.** *Retrovirology* 2013; **10**:49.
22. Maljkovic Berry I, Ribeiro R, Kothari M, Athreya G, Daniels M, Lee HY, *et al*. **Unequal evolutionary rates in the human immunodeficiency virus type 1 (HIV-1) pandemic: the evolutionary rate of HIV-1 slows down when the epidemic rate increases.** *J Virol* 2007; **81**:10625–10635.
23. Konikoff J, Brookmeyer R, Longosz AF, Cousins MM, Celum C, Buchbinder SP, *et al*. **Performance of a limiting-antigen avidity enzyme immunoassay for cross-sectional estimation of HIV incidence in the United States.** *PLoS One* 2013; **8**:e82772.
24. Vlahov D, Anthony JC, Munoz A, Margolick J, Nelson KE, Celentano DD, *et al*. **The ALIVE study, a longitudinal study of HIV-1 infection in intravenous drug users: description of methods and characteristics of participants.** *NIDA Res Monogr* 1991; **109**:75–100.
25. Laeyendecker O, Brookmeyer R, Cousins MM, Mullis CE, Konikoff J, Donnell D, *et al*. **HIV incidence determination in the United States: a multiassay approach.** *J Infect Dis* 2013; **207**:232–239.
26. Redd AD, Wendel SK, Longosz AF, Fogel JM, Dadabhai S, Kumwenda N, *et al*. **Evaluation of postpartum HIV superinfection and mother-to-child transmission.** *AIDS* 2015; **29**:1567–1573.
27. Redd AD, Collinson-Streng A, Martens C, Ricklefs S, Mullis CE, Manucci J, *et al*., Rakai Health Sciences Program. **Identification of HIV superinfection in seroconcordant couples in Rakai, Uganda, by use of next-generation deep sequencing.** *J Clin Microbiol* 2011; **49**:2859–2867.
28. Edgar RC. **Search and clustering orders of magnitude faster than BLAST.** *Bioinformatics* 2010; **26**:2460–2461.

29. Lemey P, Rambaut A, Pybus O. **HIV evolutionary dynamics within and among hosts.** *AIDS Rev* 2006; **8**:125–140.
30. Raghwani J, Redd AD, Longosz AF, Wu CH, Serwadda D, Martens C, *et al*. **Evolution of HIV-1 within untreated individuals and at the population scale in Uganda.** *PLoS Pathog* 2018; **14**:e1007167.
31. Grabowski MK, Redd AD. **Molecular tools for studying HIV transmission in sexual networks.** *Curr Opin HIV AIDS* 2014; **9**:126–133.
32. Ratmann O, Grabowski MK, Hall M, Golubchik T, Wymant C, Abeler-Dörner L, *et al*., PANGEA Consortium and Rakai Health Sciences Program. **Inferring HIV-1 transmission networks and sources of epidemic spread in Africa with deep-sequence phylogenetic analysis.** *Nat Commun* 2019; **10**:1411.
33. Vrancken B, Rambaut A, Suchard MA, Drummond A, Baele G, Derdelinckx I, *et al*. **The genealogical population dynamics of HIV-1 in a large transmission chain: bridging within and among host evolutionary rates.** *PLoS Comput Biol* 2014; **10**:e1003505.
34. Gibson KM, Jair K, Castel AD, Bendall ML, Wilbourn B, Jordan JA, *et al*., DC Cohort Executive Committee. **A cross-sectional study to characterize local HIV-1 dynamics in Washington, DC using next-generation sequencing.** *Sci Rep* 2020; **10**:1989.
35. Redd AD, Doria-Rose NA, Weiner JA, Nason M, Seivers M, Schmidt SD, *et al*. **Longitudinal antibody responses in people who inject drugs infected with similar human immunodeficiency virus strains.** *J Infect Dis* 2020; **221**:756–765.