Title
The Global Transmission Network of HIV-1

Permalink
https://escholarship.org/uc/item/8jv598zc

Authors
Wertheim, Joel
Leigh Brown, Andrew
Hepler, Lance
et al.

Publication Date
2014-01-01

Peer reviewed
The Global Transmission Network of HIV-1

Joel O. Wertheim,1,2 Andrew J. Leigh Brown,3 N. Lance Hepler,4 Sanjay R. Mehta,2 Douglas D. Richman,1,2,5 Davey M. Smith,2,5 and Sergei L. Kosakovsky Pond2

1Department of Pathology and 2Department of Medicine, University of California, San Diego; 3Institute of Evolutionary Biology, University of Edinburgh, United Kingdom; 4Bioinformatics and Systems Biology Graduate Program, University of California, San Diego; and 5Veterans Affairs San Diego Healthcare System, California

Human immunodeficiency virus type 1 (HIV-1) is pandemic, but its contemporary global transmission network has not been characterized. A better understanding of the properties and dynamics of this network is essential for surveillance, prevention, and eventual eradication of HIV. Here, we apply a simple and computationally efficient network-based approach to all publicly available HIV polymerase sequences in the global database, revealing a contemporary picture of the spread of HIV-1 within and between countries. This approach automatically recovered well-characterized transmission clusters and extended other clusters thought to be contained within a single country across international borders. In addition, previously undescribed transmission clusters were discovered. Together, these clusters represent all known modes of HIV transmission. The extent of international linkage revealed by our comprehensive approach demonstrates the need to consider the global diversity of HIV, even when describing local epidemics. Finally, the speed of this method allows for near-real-time surveillance of the pandemic’s progression.

Keywords. human immunodeficiency virus; transmission network; molecular epidemiology.

The origin and geographic expansion of human immunodeficiency virus type 1 (HIV-1) have been well characterized using phylogenetic approaches [1], but these methods are suboptimal for describing recent HIV transmission. Phylogenies are well suited for differentiating distinct viral lineages but not for identifying transmission partners; it is generally accepted that phylogenetic analysis is most powerful at excluding potential transmission partners, rather than establishing linkage [2–6]. Conversely, transmission networks focus on similarity and directly link genetically similar viruses. Although network methods have previously been used to study HIV [7, 8], partly because recombination in HIV complicates phylogenetic inference, they have not been adopted in the context of global epidemiological patterns.

Previous studies investigating HIV transmission clusters typically begin by inferring a phylogeny and then identifying those clades (ie, subtrees) that have appropriate statistical support. This identification alone, however, is insufficient for epidemiological purposes, because such an analysis lacks the concept of recency; for example, individual HIV-1 subtypes will form well-resolved clades. Hence, the isolates within supported clades are often designated as a transmission cluster if the mean [9], median [10], or maximum [11, 12] genetic diversity (a proxy for time) within these clades falls below a given cutoff. The branching structure within the phylogeny is irrelevant after these clades have been identified. A conceptual problem arises as these clades are not easily resolved into transmission pairs and clusters [11, 13, 14], and, within a transmission cluster, all individuals are treated as equally related/connected; the inferred branching structure within these clusters is again discarded (but see Leigh Brown et al [12] for an alternative approach), wasting the considerable computational effort expended to infer the complete phylogeny. Furthermore, the relationship between statistical support for a clade and
population dynamics may be conceptually ambiguous [15]. High statistical support (eg, bootstrap) for any specific clade indicates that there is no close relative to the clade in question, not that the members of the clade itself are necessarily closely related to each other [16]. A final problem with using phylogenetic methods to infer transmission clusters is that often only a single geographic region is considered, and the data must be subsampled for computational tractability [9, 11, 12, 15, 17–19]. Both of these simplifications can seriously bias the interpretation due to the limited scope of the analysis, as closely related or relevant sequences may be inadvertently excluded during sequence selection.

We applied a network approach to analyze global HIV-1 transmission patterns by constructing HIV-1 transmission clusters using only close genetic links to identify potential transmission partners. In contrast with phylogenetic approaches, inclusion in our transmission clusters required only a shared connection with another member of the cluster. We identified potential transmission partners—HIV-positive individuals whose viral pol gene sequences were genetically similar (≤1% nucleotide genetic distance)—and defined transmission clusters as maximal connected groups of potential transmission partners. Focusing on only potential transmission partners (1) maintains the internal structure of the transmission clusters, (2) obviates the reliance on a phylogeny to identify highly supported clades, and (3) eliminates the need to subsample the dataset, which allows for consideration of the global diversity of HIV-1 when inferring transmission clusters.

**METHODS**

**Sequence Dataset**

All HIV-1 group M protease and reverse transcriptase gene sequences (pol, HXB2 coordinates: 2253–3554) at least 500 nucleotides long were downloaded from the Los Alamos National Laboratory (LANL) HIV sequence database on 1 February 2012. Only 1 sequence per individual, based on classification in the LANL database, was retained. These sequences represent all known subtypes, 47 circulating recombinant forms, and many unique recombinant forms. Clonal and “problematic” sequences were removed, leaving 84 757 sequences. Sequences were codon-aligned to HXB2 using an extension of the Smith-Waterman algorithm [20], which uses amino acid homology to directly align nucleotide sequences (more suitable for aligning isolates from different divergent subtypes) and corrects out-of-frame insertions and deletions (likely sequencing errors). Given the evolutionary conservation of this genomic region, codon insertions relative to HXB2 were filtered from downstream analyses.

As a conservative screen for contamination, we inspected clusters that included sequences from the 1980s, as many of them are used as laboratory controls. Eight clusters, containing a total of 230 sequences, were removed as suspected contaminants because they included sequences from both the 1980s and the late 1990s/2000s; a small genetic distance between such heterochronous sequences is implausible, and we attributed it to laboratory contamination or sample mislabeling. The largest of these clusters contained the HIV reference sequence, HXB2, and 188 other sequences sampled between 1983 and 2010 representing 25 countries and regions, with more than half the isolates coming from the United States, China, France, and Italy.

The final curated pol dataset, after removing these potential contaminants, contained 84 527 sequences.

**Network Construction**

Genetic distance (Tamura-Nei 93, or TN93 [21]) was determined for all pairs of pol sequences (approximately 3.5 billion comparisons). TN93 distances, which model a single nucleotide transversion rate and 2 nucleotide transition rates, were used because they are the most complex genetic distances that can be represented by a closed-form solution, allowing for rapid distance calculations. The network was constructed by identifying pairs of sequences (nodes) that overlapped for a minimum of 500 nucleotides and whose divergence was ≤1% (0.01 expected substitutions per site); these sequences were connected (an undirected edge placed between 2 nodes) in our network. Due to the low divergence cutoff used to identify potential transmission partners, network construction is not expected to be noticeably affected by evolutionary model selection [22]; hence, TN93 was chosen as a balance between computational tractability and biological realism.

Network construction was neither a computationally nor memory-intensive process, because sequences were analyzed in a pairwise fashion, and only those pairs whose divergence was below a cutoff were stored in memory (<0.002% of total). Nucleotide ambiguities were averaged during distance calculations (eg, the distance between R:A is the mean of the distances between A:A and G:A) to prevent spurious connections between sequences of poor quality or multiple infections. This analysis ([https://github.com/veg/HIVClustering; https://github.com/veg/TN93; https://github.com/veg/BioExt]), including alignment, distance calculation, and network construction, can be performed in <30 minutes on a small (128 CPUs) computer cluster and in <8 hours on a standard laptop computer (2.53 GHz Intel Core 2 Duo MacBook Pro), making it an extremely efficient tool for rapidly reconstructing the recent history of HIV-1. The addition of a sequence to an existing network has linear complexity in the size of the current network and, in practical terms, can be accomplished nearly instantaneously, making the method very attractive for an online implementation.

**Bootstrap Datasets**

To test the overall stability of our network inference, network replicates were constructed from 1000 bootstrapped codon
alignments. We calculated the frequency of edges from the global network in the bootstrap datasets and explored the variance in the genetic distance these edges represent.

RESULTS

Network Inference

The global transmission network was built from every publicly available HIV-1 group M pol sequence (n = 84,527, 1 sequence per individual), representing 141 countries/regions (Table 1). The network contained 4,342 connected components with ≥2 nodes (clusters) comprising 13,295 nodes (individual sequences) and 51,182 edges (undirected, potential transmission links) (Figures 1 and 2 and Supplementary Figure 1). The average degree (number of edges per node) was 3.84.

Like smaller, local HIV-1 networks [12, 23] and sexual networks in general [24], the inferred global network appears to be scale-free. A classical mechanism for generating scale-free networks is the preferential attachment process, whereby a new member of the network connects to an existing node with probability proportional to the latter’s degree (eg, a connection to an existing node with degree 10 is 5 times more likely than to a node with degree 2). We compared the fit of 4 different degree distributions (negative binomial [not scale free], Pareto, Waring, and Yule [all scale free]) using Bayesian information criterion [12]. The Waring distribution, a preferential-attachment model with a proportion of connections formed randomly [25], provided the best fit to the inferred degree distribution. The characteristic exponent of the Waring distribution (ρ) for the global transmission network is 1.74; a smaller ρ indicates a higher propensity for high-degree nodes.

The network analysis is performed without explicit regard for HIV-1 subtype. Misclassification of subtype or recombinant provenance (a common problem [26]) will not affect the inference of the global network, because no sequences were excluded based on prior classification. Moreover, distance-based approaches are able to handle the inclusion of recombinant sequences, as no single underlying phylogeny is assumed. However, the oversampling of certain subtypes and circulating recombinant forms (eg, B and CRF01_AE) could bias the network toward more high-degree nodes.

International Dimension of HIV Transmission

A global analysis including all published HIV-1 sequences recovered previously characterized transmission clusters, expanded known transmission clusters by incorporating additional isolates and/or smaller clusters, and identified previously undescribed transmission clusters (for representative examples and references, see Figure 3 and Table 2). For example, both UK clusters (shown in Figure 3) were partially characterized in the literature, but these transmission clusters had not been considered in the global context, and hence the international links had been missed. The global network analysis also discovered 2 large international transmission clusters: (1) a 333-node cluster representing 17 countries, primarily comprising heterosexuals and injection drug users in Southeast Asia, and (2) a 674-node cluster representing 18 countries, primarily comprising injection drug users and men who have sex with men in the former Soviet Union (Figure 2). Portions of both of these networks have been previously described in the literature (Supplementary Data); however, the extent of international transmission had been vastly underestimated, because these previous studies narrowly focused on single countries or regions. We argue that a

---

Table 1. Los Alamos National Laboratory pol Sequences in Global Network Analysis and Their Propensity for Clustering

<table>
<thead>
<tr>
<th>Geographic Region</th>
<th>pol Sequences in LANL</th>
<th>Sampling Years</th>
<th>Observed Clustered</th>
<th>Expected Clustered</th>
<th>( \chi^2 ) Test</th>
</tr>
</thead>
<tbody>
<tr>
<td>Central Asia</td>
<td>273</td>
<td>1998–2009</td>
<td>5.43</td>
<td>NS</td>
<td>( **)</td>
</tr>
<tr>
<td>Eastern Asia</td>
<td>4,916</td>
<td>1991–2010</td>
<td>2.38</td>
<td>NS</td>
<td>( **)</td>
</tr>
<tr>
<td>Eastern Europe</td>
<td>2,885</td>
<td>1986–2011</td>
<td>2.02</td>
<td>NS</td>
<td>( **)</td>
</tr>
<tr>
<td>Northern America</td>
<td>17,836</td>
<td>1982–2010</td>
<td>1.82</td>
<td>NS</td>
<td>( **)</td>
</tr>
<tr>
<td>Central America</td>
<td>1,124</td>
<td>2001–2010</td>
<td>1.78</td>
<td>NS</td>
<td>( **)</td>
</tr>
<tr>
<td>Western Asia</td>
<td>333</td>
<td>1994–2009</td>
<td>1.58</td>
<td>NS</td>
<td>( **)</td>
</tr>
<tr>
<td>Northern Europe</td>
<td>4,330</td>
<td>1986–2009</td>
<td>1.55</td>
<td>NS</td>
<td>( **)</td>
</tr>
<tr>
<td>Southern Europe</td>
<td>7,187</td>
<td>1991–2011</td>
<td>1.00</td>
<td>NS</td>
<td>NS</td>
</tr>
<tr>
<td>Northern Africa</td>
<td>306</td>
<td>1998–2010</td>
<td>0.97</td>
<td>NS</td>
<td>NS</td>
</tr>
<tr>
<td>Southern Asia</td>
<td>1,748</td>
<td>1993–2010</td>
<td>0.94</td>
<td>NS</td>
<td>NS</td>
</tr>
<tr>
<td>Southeastern Asia</td>
<td>4,080</td>
<td>1990–2011</td>
<td>0.92</td>
<td>NS</td>
<td>NS</td>
</tr>
<tr>
<td>Caribbean</td>
<td>1,505</td>
<td>1996–2010</td>
<td>0.89</td>
<td>NS</td>
<td>NS</td>
</tr>
<tr>
<td>Western Europe</td>
<td>5,640</td>
<td>1986–2009</td>
<td>0.70</td>
<td>NS</td>
<td>( **)</td>
</tr>
<tr>
<td>Eastern Africa</td>
<td>5,323</td>
<td>1985–2010</td>
<td>0.57</td>
<td>NS</td>
<td>( **)</td>
</tr>
<tr>
<td>Australiad</td>
<td>819</td>
<td>1987–2004</td>
<td>0.49</td>
<td>NS</td>
<td>( **)</td>
</tr>
<tr>
<td>Middle Africa</td>
<td>1,967</td>
<td>1983–2009</td>
<td>0.42</td>
<td>NS</td>
<td>( **)</td>
</tr>
<tr>
<td>South America</td>
<td>8,685</td>
<td>1989–2010</td>
<td>0.40</td>
<td>NS</td>
<td>( **)</td>
</tr>
<tr>
<td>Southern Africa</td>
<td>3,206</td>
<td>1984–2010</td>
<td>0.34</td>
<td>NS</td>
<td>( **)</td>
</tr>
<tr>
<td>Unknown origin</td>
<td>9,861</td>
<td>2001–2003</td>
<td>0.25</td>
<td>NS</td>
<td>( **)</td>
</tr>
<tr>
<td>Western Africa</td>
<td>2,503</td>
<td>1990–2010</td>
<td>0.14</td>
<td>NS</td>
<td>( **)</td>
</tr>
</tbody>
</table>

Abbreviations: LANL, Los Alamos National Laboratory; NS, not significant.

* Propensity for clustering was not uniform across geographic regions \( \chi^2 \) test, \( P < .001, \ df = 19 \). Data are shown from most overrepresented to most underrepresented.

* Based on the United Nations geographic region and composition (http://unstats.un.org/unsd/methods/m49/m49regin.htm).

* Twenty percent of sequences in LANL do not have associated sampling years.

* Includes 1 sequence from Fiji.

* Sequences without associated geographic sampling information in LANL.

** Bonferroni-corrected P value < .001 (df = 1).
global-scale analysis is a prerequisite for capturing the scope of these transmission clusters and should become commonplace in the field.

Our analysis revealed the contemporary pattern of HIV-1 transmission across international borders (Figure 4). Of the 106 countries or regions represented in transmission clusters, individuals from 72 (68%) of these countries/regions had a potential transmission partner from another country/region (see Supplementary Table 1 for a list of the most connected countries). Moreover, 22.6% of potential transmission partners were from different countries/regions, likely due to the high connectedness of larger, international transmission networks (Figure 2), and 3.6% of transmission clusters included potential transmission partners in multiple countries/regions. Of course, substantial sampling and publication biases exist within public sequence databases, particularly regarding the undersampling of isolates from sub-Saharan Africa. However, the international nature of inferred transmission clusters is all the more striking because of the biases toward densely sampling local epidemics.

Transmission Clusters
Inclusion in our transmission clusters required only 1 shared edge with another member of the cluster. Therefore, one would expect to find more and larger transmission clusters in regions with young epidemics and extensive sampling, and indeed the propensity for clustering in our network varied across geographic regions \( (\chi^2 \text{ test}, P < .0001; \text{ Table 1}) \). Isolates from recent epidemics in Central and East Asia clustered more
frequently than isolates from elsewhere in the world, and isolates from older and more diverse African epidemics were less likely to be found in clusters. Furthermore, the 3 largest transmission clusters predominantly from a single country/region were found in younger epidemics from Pakistan, Taiwan, and China (Supplementary Figure 1).

Overall, inferred transmission clusters were densely connected, suggesting that most isolates were linked to multiple other isolates in their cluster. The mean genetic distance between nodes was ≤1% for 96% of the clusters comprising ≥3 nodes, and the mean distance for the largest clusters (Figure 2 and Supplementary Figure 1) did not exceed 2%. Thus, many of these clusters could be identified using approaches and thresholds adopted in earlier studies, albeit at a much steeper computational cost and only if the relevant sequences had been considered for inclusion a priori. Notably, our network approach identifies clusters that include all potential transmission partners without excluding peripherally linked partners; these potential partners may have been excluded by previous methods that rely on mean, median, or maximum genetic distance within a cluster. As expected in a scale-free, densely connected network [37], we observed a log-linear relationship between cluster size and the maximum pairwise divergence within a cluster (Figure 5A).

Most viruses from potential transmission partners were sampled within a narrow timeframe of each other (Figure 5B); importantly, the actual number of years of evolution separating these sequences is greater than these times, except when 1 virus is the direct descendant of the other [38]. Viral isolates from 60% of the potential transmission partners were obtained within 1 year of each other. Surprisingly, it appears as if our network approach was able to capture at least some transmission events involving chronic infections: the 95th percentile of years separating potential transmission partners extended to 5 years. Furthermore, we found that nearly 1% of potential transmission partners were separated by at least a decade. This unusually long time span separating viruses with <1% genetic divergence may be explained by (1) transmission of latent virus due to reemergence after antiretroviral therapy [39], (2) transmission of a slowly evolving virus from individuals with low viral load (ie, elite controllers) [40], (3) possible preferential transmission of ancestral variants [41, 42] (although see [43]), and/or (4) incorrectly recorded year of sampling or contamination.

Figure 2. Large international transmission clusters. A, Southeast Asian transmission cluster, CRF01_AE, primarily from Thailand and Vietnam. Additional isolates from Belgium, Cambodia, China, Czech Republic, France, Greece, Indonesia, Japan, Luxembourg, Malaysia, Singapore, Sweden, Taiwan, United Kingdom, and United States. B, Former Soviet Union transmission cluster, subtype A1, primarily from the former Soviet Union. Isolates from Azerbaijan, Belarus, Cyprus, Czech Republic, Denmark, Georgia, Israel, Kazakhstan, Latvia, Moldova, Portugal, Russia, Slovenia, Spain, Sweden, Ukraine, United Kingdom, and Uzbekistan. Edge lengths are optimized for visual presentation and do not represent genetic distances between potential transmission partners. All nodes are of equal size, and colors correspond to geographic regions in Figure 1.

Network Stability
The global transmission network inference appears robust to a number of model parameters and assumptions. More than 99% of edges were supported in a majority-rule bootstrap consensus network (Supplementary Figure 2); this analysis is a network analog of the phylogenetic bootstrap.

To determine the evolution of network properties through time, we reanalyzed the sequences, censoring the most recent...
year in a stepwise fashion. The current network structure appears to have been largely established by the early 2000s, when there were only about 500 inferred clusters (Figure 5C). In our dataset, 17,280 sequences (20%) lacked an associated year of sampling, and excluding these sequences resulted in slightly biased estimates of the network characteristic exponent ($\rho$). This bias is due to the higher frequency with which isolates without known sampling years were found in clusters ($\chi^2$ test, $P < .0001$). To confirm the source of this bias, we reconstructed 100 replicate networks with 17,280 random sequences removed; the network characteristic exponent $\rho$ (95% confidence interval, 1.72–1.75) was indistinguishable from that ($\rho = 1.74$) in the original network.

A 1% distance cutoff (0.01 substitutions per site) for potential transmission partners was selected because it is a conservative approximation of the level of intrahost diversity early in infection and has been used in previous transmission network analyses in a cohort of recently infected individuals [19, 44]. This strict cutoff favors the detection of recent transmission partners and will likely miss older transmissions between partners with long-term chronic infections. Nevertheless, we found that the overall network structure is stable with respect to the

![Figure 3](http://jid.oxfordjournals.org/)

*Figure 3.* Examples of transmission clusters inferred from the global network shown in Figure 1. Node size is proportional to relative degree (number of edges per node). Edge lengths are optimized for visual presentation and do not represent genetic distances between potential transmission partners. Colors correspond to geographic regions in Figure 1.
distance cutoff used to determine putative transmission partners (Figure 5D). For reasonable levels of intrahost pol diversity (0.002–0.02 substitutions per site), the network characteristic exponent is constrained to a relatively narrow range ($\rho = 1.54$–1.84).

**Drug Resistance**

The structure of the global network supports a hypothesized fitness trade-off between drug resistance-associated mutations (DRAMs) and viral transmission. We observed an inverse relationship between the degree of the node (a proxy for transmissibility) and the number of DRAMs at that node. This analysis was performed on both the original network (ordinal logistic regression, $P < .0001$) and the network constructed after removing the DRAM sites (see below, $P < .01$), adjusting for year and region of sampling. In the absence of antiviral drug treatment, resistance mutations should incur a fitness cost that could result in decreased viral transmission [45]. Nodes with high degrees, an indicator of more frequent transmission events, were less likely to harbor DRAMs. This pattern suggests that in the absence of antiretroviral therapy, resistance mutations should dissipate from the viral population. However, most of the viral isolates analyzed here lacked information about the patient’s drug treatment status. Future studies of transmission networks constructed from higher quality datasets (in which infection stage and drug treatment status are known) will be able to better address the issue of DRAM fitness costs across populations.

To determine whether DRAMs influence network construction, we removed 36 codons associated with major drug resistance mutations (http://hivdb.stanford.edu/pages/download/resistanceMutations_handout.pdf) and repeated the analysis. In the absence of these sites, the network was found to have a

---

**Table 2. Examples of Transmission Clusters in the Global Network**

<table>
<thead>
<tr>
<th>Countries/Regions</th>
<th>No. of Isolates</th>
<th>Sampling Year(s)</th>
<th>Subtype</th>
<th>Prior Description</th>
<th>Risk Factor(s)</th>
<th>References</th>
</tr>
</thead>
<tbody>
<tr>
<td>Greenland</td>
<td>15</td>
<td>1999–2006</td>
<td>B</td>
<td>Yes</td>
<td>Heterosexual</td>
<td>[27, 28]</td>
</tr>
<tr>
<td>Libya</td>
<td>28</td>
<td>1998</td>
<td>CRF02_AG</td>
<td>Yes</td>
<td>Nosocomial</td>
<td>[29]</td>
</tr>
<tr>
<td>Spain/Switzerland/Brazil</td>
<td>41</td>
<td>2005–2011</td>
<td>F1</td>
<td>Yes</td>
<td>MSM/heterosexual</td>
<td>[18, 30–32]</td>
</tr>
<tr>
<td>United Kingdom/France/Austria/Czech Republic</td>
<td>16</td>
<td>2000–2006</td>
<td>B</td>
<td>Partial</td>
<td>MSM</td>
<td>[9, 12, 33, 34]</td>
</tr>
<tr>
<td>United Kingdom/Australia/Singapore</td>
<td>8</td>
<td>2005–2007</td>
<td>B</td>
<td>Partial</td>
<td>MSM</td>
<td>[12, 35, 36]</td>
</tr>
<tr>
<td>France/Martinique</td>
<td>6</td>
<td>2006</td>
<td>B</td>
<td>None</td>
<td>Unknown</td>
<td>NA</td>
</tr>
</tbody>
</table>

Abbreviations: MSM, men who have sex with men; NA, not applicable.

---

**Figure 4.** Global transmission patterns of human immunodeficiency virus type 1. There are 211 putative transmission links between countries/regions. Green countries/regions are sampled in the network; gray indicates absence of sampling. Connections for each country/region originate in the centroid of its map region.
slightly higher characteristic exponent: \( \rho = 1.84 \) (i.e., fewer high-degree nodes). To confirm the source of this shift, we reconstructed 100 replicate networks with 36 random codons removed in which \( \rho \) (mean, 1.75 [95% confidence interval, 1.72–1.78]) was indistinguishable from \( \rho \) in the original network (\( \rho = 1.74 \)).

**The gag Transmission Network**

Although the available dataset is much smaller, similar patterns were seen in the global transmission network inferred using gag sequences. The same network inference procedure was repeated for gag (p24) (HXB2 coordinates: 1186–1879), yielding 9426 sequences. There were 592 distinct clusters in the gag network (after excluding 6 clusters, containing 36 sequences, as potentially contaminated), representing 43 putative cross-country transmission events. Using isolates whose sequence spanned both gag and pol regions, 72 clusters were linked, including the 2 large, international pol clusters (Figure 2). Therefore, pol is not unique among HIV genes in its usefulness in reconstructing transmission history, and multiple loci can be used to confirm or supplement network inference.

**DISCUSSION**

Analysis of the global HIV-1 transmission network provides a comprehensive and current picture of the spread of HIV-1 within and between countries; it reveals that “local” epidemics often include international transmission links and can be fully understood only in the context of global HIV diversity. This finding is in agreement with an analysis of transmission partners within a single town in the United Kingdom, which found that only 30% of the transmission events occurred locally [46], as well as with earlier studies in Uganda [47]. The large
international clusters (Figure 2), along with the expanded UK clusters, which were not previously known to extend beyond that country’s borders (Figure 3), explicitly demonstrate the risk incurred by arbitrarily restricting phylogenetic or network analyses to single countries or regions: international transmission links are lost. Furthermore, these and other large transmission clusters, in which each isolate is a potential transmission partner of at least 1 other isolate, could have been missed or broken up by clustering methods that begin with a phylogenetic tree and then identify highly supported clades with low levels of divergence. Even high-resolution phylogenetic analyses of local epidemics would benefit from including relevant isolates identified by a method that is global in scope.

Although edges in the global network are not synonymous with actual transmission events, they likely link individuals in the same epidemiological transmission cluster. The scope and size of many of the clusters presented here expand with a less conservative distance cutoff (Figure 5D), but the network structure remains relatively unchanged. Moreover, the absence of connectedness between countries and clusters should not be interpreted as lack of direct transmission; rather, it reflects incomplete and nonuniform sampling due to biases in publication of HIV-1 sequences. Thus, historical viral migration events in the spread of HIV-1 subtypes around the world (eg, subtype B from Haiti to the United States, subtype F1 from Angola to Brazil and Romania, and subtype C from sub-Saharan Africa to India and China [48–50]) are absent from our network (Figure 4).

Transmission network approaches are well suited for the analysis of HIV surveillance data, which are less susceptible to the effects of sampling and publication bias (eg, oversampling of geographic regions or risk groups) and would produce more robust analysis. Last, our global-network approach will allow researchers to determine, nearly instantaneously, if newly isolated HIV-1 sequences fall within known transmission clusters, enabling near-real-time surveillance of recent and growing local epidemics and patterns of international transmission. As more sequences are deposited in public databases, the accuracy and resolution of this approach will continue to improve.

Supplementary Data

Supplementary materials are available at The Journal of Infectious Diseases online (http://jid.oxfordjournals.org/). Supplementary materials consist of data provided by the author that are published to benefit the reader. The posted materials are not copyedited. The contents of all supplementary data are the sole responsibility of the authors. Questions or messages regarding errors should be addressed to the author.

Notes

Financial support. This work was supported by the National Institutes of Health (grants AI43638, AI47745, AI093163, AI07384, GM093939, MH083552, AI74621, AI100665, DA034978); the University of California, San Diego Center for AIDS Research (grant NIH-AI36214); the University of California Laboratory Fees Research Program (grant 12-LR-236617), and the Department of Veterans Affairs.

Potential conflicts of interest. All authors: No reported conflicts.

All authors have submitted the ICMJE Form for Disclosure of Potential Conflicts of Interest. Conflicts that the editors consider relevant to the content of the manuscript have been disclosed.

References

42. Lythgoe KA, Fraser C. New insights into the evolutionary rate of HIV-1 at the within-host and epidemiological levels. Proc Biol Sci 2012; 279:3367.