JOURNAL OF VIROLOGY, Dec. 2007, p. 13050–13056 Vol. 81, No. 23
0022-538X/07/$08.00+0 doi:10.1128/JVI.00889-07

Phylogenetic Surveillance of Viral Genetic Diversity and the Evolving Molecular Epidemiology of Human Immunodeficiency Virus Type 1[▽]

Robert J. Gifford,[1,2]*† Tulio de Oliveira,[3,4]† Andrew Rambaut,[5] Oliver G. Pybus,[3] David Dunn,[6] Anne-Mieke Vandamme,[7] Paul Kellam,[1] and Deenan Pillay,[1,8] on Behalf of the UK Collaborative Group on HIV Drug Resistance

Department of Infection, University College London, London, United Kingdom[1]; Division of Infectious Diseases, Stanford University, Stanford, California[2]; Department of Zoology, University of Oxford, Oxford, United Kingdom[3]; The South African National Bioinformatics Institute, University of the Western Cape, Bellville, South Africa[4]; Institute of Evolutionary Biology, University of Edinburgh, Edinburgh, Scotland[5]; Medical Research Council Clinical Trials Unit, London, United Kingdom[6]; Rega Institute for Medical Research, K.U. Leuven, Belgium[7]; and Centres for Infection, Health Protection Agency, Colindale, United Kingdom[8]

Received 25 April 2007/Accepted 17 September 2007

With ongoing generation of viral genetic diversity and increasing levels of migration, the global human immunodeficiency virus type 1 (HIV-1) epidemic is becoming increasingly heterogeneous. In this study, we investigate the epidemiological characteristics of 5,675 HIV-1 *pol* gene sequences sampled from distinct infections in the United Kingdom. These sequences were phylogenetically analyzed in conjunction with 976 complete-genome and 3,201 *pol* gene reference sequences sampled globally and representing the broad range of HIV-1 genetic diversity, allowing us to estimate the probable geographic origins of the various strains present in the United Kingdom. A statistical analysis of phylogenetic clustering in this data set identified several independent transmission chains within the United Kingdom involving recently introduced strains and indicated that strains more commonly associated with infections acquired heterosexually in East Africa are spreading among men who have sex with men. Coalescent approaches were also used and indicated that the transmission chains that we identify originated in the late 1980s to early 1990s. Similar changes in the epidemiological structuring of HIV epidemics are likely to be taking in place in other industrialized nations with large immigrant populations. The framework implemented here takes advantage of the vast amount of routinely generated HIV-1 sequence data and can provide epidemiological insights not readily obtainable through standard surveillance methods.

As the AIDS pandemic progresses, an increasingly broad range of genetic diversity is being reported within the main (M) group of human immunodeficiency virus type 1 (HIV-1) viruses. Although broad diversity is concentrated in areas of West and Central Africa, where infection is longest established (24, 40–42, 46), it is increasingly evident elsewhere as infection expands globally (9, 13, 14, 28, 30, 34, 44).

Founder effects accompanying the spread of HIV-1 infection have generated an uneven distribution of M group strains among geographic areas and exposure risk populations (27). Consequently, strains often exhibit specific associations with particular geographic regions and/or modes of transmission (8, 18, 20, 32). Tracking these dynamic associations through surveillance of genetic diversity will facilitate epidemiological investigations and inform public health strategies for the prevention of viral spread (5, 16, 18, 43).

The introduction of resistance test sequencing as a standard component of clinical care in many countries provides an abundant source of routinely generated sequence data for the HIV-1 *pol* gene. Such wide-scale generation of HIV sequence data enables us to use phylogenetic approaches to investigate epidemiological hypotheses that cannot be tackled using non-genetic surveillance data alone (16). In particular, it is essential to determine which HIV strains are being introduced into countries, whether these strains are spreading through ongoing transmission, and if so, through what exposure risk.

In a recent analysis of *pol* gene sequence data representing approximately one-fifth of all United Kingdom infections, we revealed the extensive range of HIV-1 genetic diversity present within the United Kingdom (12). Here, we use phylogenetic methods to explore this diversity in depth, showing how epidemiological information can be extracted from sequence data sets of this type. In addition, our analysis provides new, fine-scale information about the geographic structuring of the global HIV-1 epidemic.

MATERIALS AND METHODS

Study population. Sequence data consisted of 5,675 *pol* gene sequences sampled from distinct individuals and linked to patient treatment and demographic data in the United Kingdom Drug Resistance Database (www.hivrdb.org.uk). All sequences were submitted for routine genotypic drug resistance testing between 1996 and 2004. Sequences were generated by population sequencing from plasma samples using a variety of commercial and in-house procedures and were at least 900 nucleotides in length, typically spanning the entire protease (PR) gene region and at least codons 40 to 240 of reverse transcriptase (RT). The median sequence length was 1,497 nucleotides. Demographic data were available for the majority of patients, including nationality (72%), ethnicity (82%), and exposure risk group (93%).

In the United Kingdom, viral sequences may be obtained for resistance testing prior to initiating antiretroviral therapy, as well as in response to treatment

* Corresponding author. Mailing address: Division of Infectious Diseases, Stanford University, Stanford, CA 94305. Phone: (650) 725-2946. Fax: (650) 725-2088. E-mail: rjmg@stanford.edu.
† These authors contributed equally.
▽ Published ahead of print on 26 September 2007.

failure. For all patients in this study, samples taken prior to initiating therapy were preferentially used. Where no pretreatment sample was available, the earliest posttreatment sample was used. In total, 2,821 sequences (50%) were obtained from patients reported as antiretroviral naive at the time of sampling, 2,750 were from patients with some previous treatment history, and 110 were from cases whose treatment history was unknown. The prevalence of drug resistance mutations in the data set was determined using the calculated population resistance (CPR) tool (cpr.stanford.edu/cpr/). In summary, 2,219 sequences (38.2%) had one or more surveillance drug resistance mutations (29). The prevalence of surveillance drug resistance mutations was highly skewed towards mutations at five positions (41, 67, 70, 103, 184, and 215 in RT). Mutations at 12 other positions occurred at a 2 to 9% prevalence (positions 46, 54, 82, 88, and 90 in PR and 70, 74, 101, 181, 190, 210, and 219 in RT). Surveillance drug resistance mutations at other positions occurred at a <2.0% prevalence.

Strain-level classification of HIV-1 diversity. Nine hundred seventy-nine complete-genome sequences sampled from distinct infections and annotated by country of sampling were used to derive a fine-scale classification of global HIV-1 genetic diversity that reflected geographic associations. Sequences were obtained from the Los Alamos HIV-1 sequence database (www.hiv.lanl.gov) and aligned using a combination of automated protocols and manual adjustment. This data set included representatives of all "pure" HIV-1 M group subtypes (A to D, F to H, J to K) and circulating recombinant forms (CRFs) 01 to 14, 18, and 19. Alignments were used to construct neighbor-joining (NJ) phylogenies with gamma rate distributions and the Hasegawa-Kishino-Yano evolutionary model implemented in PAUP (33). We defined epidemiologically distinct "strains" as monophyletic clusters within established subtype/CRF groupings that comprised two or more sequences sharing the same geographic origin, as defined by country of sampling (countries were grouped into geographic regions according to the classification used by UNAIDS [14]).

The inclusion of recombinant sequences within the data set meant that phylogenetic reconstruction could misrepresent evolutionary relationships, which would be correctly represented by a network rather than a tree. However, the aim here was not to accurately reconstruct the deeper evolutionary history of HIV-1 M group lineages but rather to distinguish more-recent lineages comprising distinct groups of closely related genomes sharing specific geographic associations. To confirm that groups identified in complete-genome phylogenies could be recovered using subgenomic regions, bootstrapped phylogenies were reconstructed using the *gag*, *pol*, and *env* genes, and a concatenated region of the alignment representing a minimum-length resistance test sequence (codons 1 to 99 of PR and 40 to 240 of RT), and 1,000 bootstrap replicates.

Batch genotyping of HIV-1 *pol* gene sequences. HIV-1 *pol* gene sequences were assigned to genotypic groups using a modified version of the Rega automated genotyping tool (http://www.bioafrica.net/subtypingclusters.html). With the Rega tool, each query sequence is analyzed separately with a set of reference sequences representing the various genotypic groups for the sequence under analysis and with a series of phylogenetic procedures incorporating bootstrapped NJ analysis, bootscanning, and likelihood mapping-based analysis of phylogenetic content (4). Assignment of sequences to genotypic groups is based on phylogenetic criteria, including bootstrap support values and the topological relationships of query sequences relative to reference sequences (see reference 4 for details). Sequences that do not fulfill criteria for confident assignment to any group represent poorly characterized diversity and are designated unclassifiable.

Ninety-six *pol* gene reference sequences (including two or more representatives of each global strain) were selected from the original 976 complete-genome sequences used for strain definition. Sequences were classified first according to established subtype/CRF definitions (in this case, assignment to CRFs required identification of recombination breakpoints in the *pol* region analyzed) and second according to the fine-scale, strain-level classification that we had defined by analysis of complete-genome sequences (see above).

Investigation of transmission dynamics. Phylogenetic analysis can indicate whether imported HIV-1 strains are spreading through ongoing local transmission. If strains are not spreading locally, then in a phylogeny representing both (i) sequences from imported strains and (ii) reference sequences representative of HIV-1 genetic diversity in the putative region(s) of original infection, we expect that imported sequences would be randomly distributed among reference sequences, reflecting separate importation events. If, however, we observe larger monophyletic clusters of imported sequences, this suggests either the presence of a local transmission chain or an importation process heavily biased towards closely related strains.

A reference data set comprising 3,201 globally sampled *pol* gene sequences (annotated by country of sampling) was obtained from the Los Alamos HIV sequence database. These sequences were classified by batch genotyping in Rega using the strain-level classification described above. Next, an NJ phylogeny was constructed using a combined United Kingdom and global *pol* sequence data set. For each strain represented in the phylogeny, the statistical significance of clustering among United Kingdom sequences relative to global reference sequences was assessed using a method based on the phylogenetic principle of parsimony (Slatkin-Maddison test [31]). Given a phylogeny in which each tip has been designated on the basis of a given state (in this case, sampling within or outside the United Kingdom), the parsimony algorithm can be used to estimate the minimum number of state changes needed to give rise to the observed distribution of states across a given region of the phylogeny (22). For clusters comprising five or more United Kingdom sequences, states were randomized among all sequences of the same strain, and for each randomization, the minimum number of state changes within that clade was calculated using parsimony. The total number of changes was summed across all 100 randomizations and divided by the number of replicates, giving the expected number of changes under the null hypothesis of random-distribution sequences among countries. The difference between the observed and expected number of changes calculated in this way indicated the significance of clustering.

Investigation of the epidemic history of transmission clusters was carried out using Bayesian evolutionary analysis sampling trees (BEAST) (7). The most appropriate demographic model for each cluster was selected using the likelihood ratio test. A constant molecular clock was applied using mutation rates calculated previously for the same gene region of HIV-1 (18, 21). Demographic and evolutionary parameters of the epidemic, together with their confidence intervals, were estimated by Bayesian Markov chain Monte Carlo inference using a chain of 10 million states sampled every 100th generation. The estimated parameters included the date of the most recent common ancestor of the cluster and the likelihood values of the sampled trees.

RESULTS

A phylogeny of 976 annotated HIV-1 complete-genome sequences was used to derive a fine-grained classification of HIV-1 diversity that reflected geographic associations among strains. In total, we identified 36 global "strains," i.e., robust, monophyletic clusters of closely related viruses exhibiting specific geographic associations (Fig. 1). The strains identified using this approach included all established subtypes and CRFs included in the data set as well as several geographic subdivisions of established subtype groupings, in addition to previously described "sub-subtypes." These included a Ugandan subtype D strain; a Nigerian subtype G strain; subtype A strains associated with East Africa, West Africa, and Central Asia; and African, South American, and Indian strains of subtype C. The geographic associations defined using this approach were in good agreement with previous reports of the global distribution of HIV-1 diversity (14). All 36 strains could be recovered as robustly supported (>70%) monophyletic clades in bootstrapped NJ phylogenies constructed using subgenomic regions, with subgenomic sequences of recombinant viruses forming monophyletic clades within their respective majority subtypes (Fig. 1).

A standard classification of 5,675 United Kingdom *pol* gene sequences according to the established subtype/CRF groupings revealed that the majority (74%) belonged to subtype B (the strain most typical of the United Kingdom and Western Europe in general [14]). However, subtypes A and C occurred at frequencies of 6% and 10%, respectively, and a broad range of groups was represented (Table 1).

Further, fine-scale classification using strain-level groupings revealed epidemiological information not apparent when using the standard nomenclature (Table 1). For example, sequences classified as subtype A by the standard approach were seen to comprise viral lineages characteristic of East Africa, West Africa, Central Asia, and Southeast Asia. Most United Kingdom subtype C sequences belonged to the strain prevalent in sub-

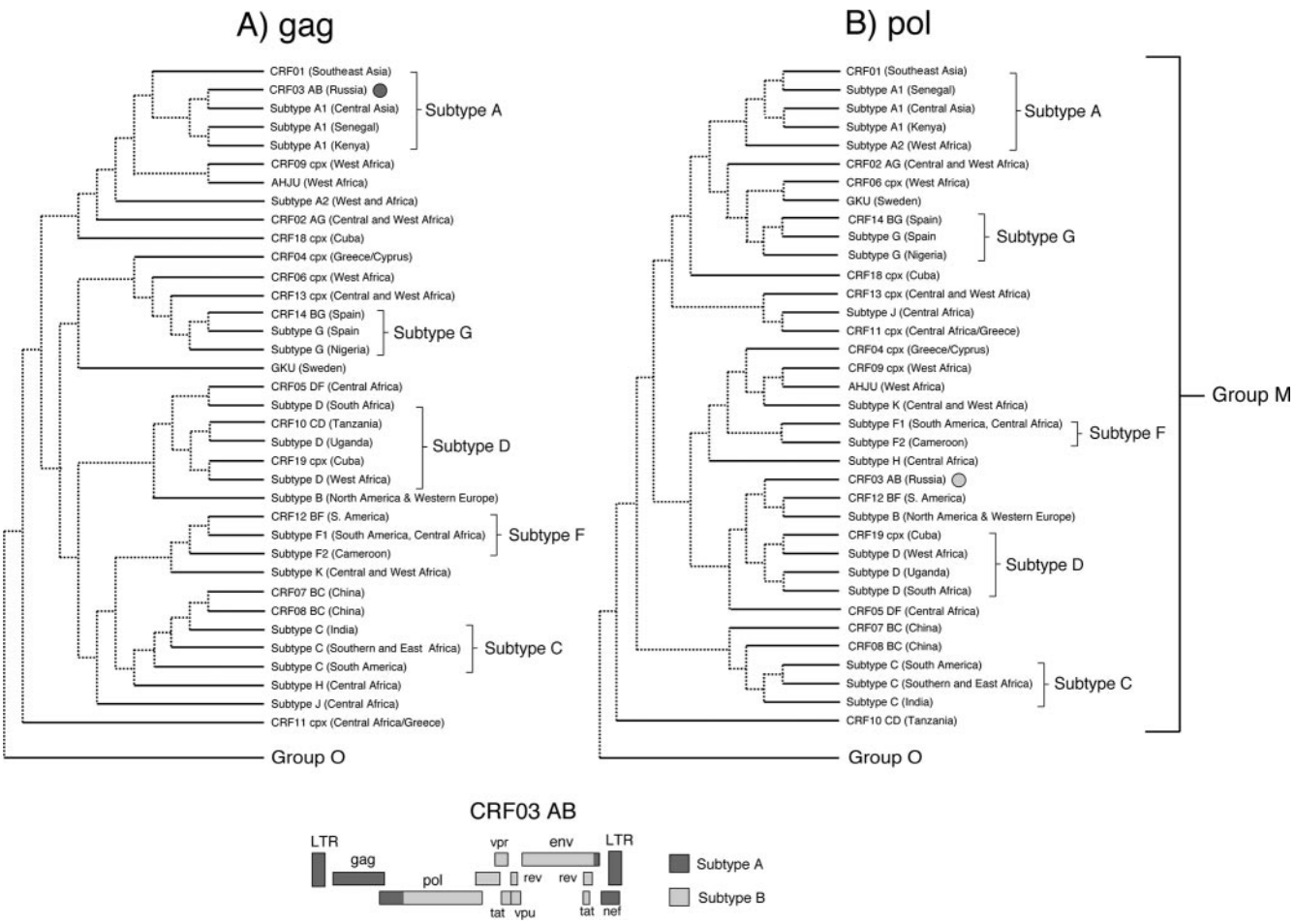

FIG. 1. Strain-level classification of HIV-1 genetic diversity. Thirty-six strain-level groupings were defined by phylogenetic analysis of sequences from the complete HIV-1 genome and were recoverable using subgenomic fragments. In this example, phylograms illustrate that due to the presence of recombinant strains, the internal topologies of phylogenies (shown as dotted lines) differ depending on which subgenomic region is analyzed. Thus, the CRF03 strain illustrated, highlighted in trees by shaded circles, can be seen to group with subtype A in trees constructed using *gag* (A) and with subtype B in trees constructed using *pol* (B). However, the classification that we describe here focuses on closely related sequences sharing specific geographic associations and does not seek to represent deeper evolutionary relationships. As such, it is robust to the effects of recombination; CRF03 and all other CRFs included in the initial data set were recovered as robustly supported monophyletic lineages in trees constructed using subgenomic regions.

Saharan Africa, but the Indian and South American strains were also represented. Approximately 3% of sequences from the United Kingdom represented novel HIV-1 diversity, showing no clear relationship to any previously reported lineage.

HIV-1 strains may be introduced into a region either by migration of infected individuals from areas where the strain is prevalent or by infections acquired by locals while traveling in those areas (19, 26, 37). In general, the demographic profiles of infected individuals matched the geographic associations of the HIV-1 strain with which they were infected (Table 1). Thus, the majority of subtype B infections were found in British patients, and other, more recently introduced strains were most commonly found in migrants from countries or regions associated with those strains. The most notable exception to this pattern was CRF01, which is prevalent in Southeast Asia (especially Thailand, Cambodia, and Vietnam [14]); almost half of the individuals infected with this strain for whom data were available were white heterosexuals from the United Kingdom. An elevated prevalence (>10%) within British nationals was also observed for the East African strain of subtype A, the southern/East African strain of subtype C, and CRF02 (Table 1). Among divergent, unclassifiable viruses for which patient demographic data were available, 67% were identified in individuals originally from sub-Saharan Africa.

A phylogenetic approach was used to investigate whether recently introduced strains might be spreading through ongoing transmission within the United Kingdom. In phylogenies representing both sequences sampled within the United Kingdom and sequences sampled globally, sequences belonging to strains other than subtype B and sampled in the United Kingdom were generally interspersed among sequences sampled elsewhere (data not shown), suggesting a random process typical of separate importation events occurring either through migration or through infection of British nationals while traveling abroad. However, three clusters were identified for which statistical analysis suggested potential ongoing transmission

TABLE 1. Strain-level assignment of 5,675 HIV-1 *pol* sequences from the United Kingdom and distribution among nationalities[e]

Classical subtype or CRF group[a]	Total no. of sequences	Strain-level grouping and/or geographic association(s)	No. of sequences assigned to group[b]	No. of sequences for which patient nationality data were available[c]	Patient nationality[c]					
					United Kingdom		Strain-associated region		Other	
					No. of sequences	%	No. of sequences	%	No. of sequences	%
Subtype A	362	A1 (Kenya)	122	85	15	17	51	60	19	22
		A2 (West Africa)	2	1	0		1		0	
		A1 (Central Asia)	1	1	0		1		0	
		CRF01 (Southeast Asia)	39	30	14	46	12	40	4	13
Subtype B	4,252	North and South America, Western Europe, Caribbean		3,150	1,963[d]	62[d]	2,941[d]	93[d]	209	7
Subtype C	568	Southern and East Africa	498	327	48	15	251	77	28	8
		South America	12	6	3		2		1	
		India	2	2	1		1		0	
Subtype D	93	Uganda	62	46	3	6	42	91	1	2
Subtype F	12	F1 (South America, Central Africa)	10	7	0		5		2	
		F2 (Cameroon)	2	2	0		2		0	
Subtype G	54	Spain	17	13	4		7		2	
		Nigeria	12	9	1		7		1	
Subtype H	4	Central Africa		3	0		2		1	
Subtype J	5	Central Africa		3	0		2		1	
CRF02 AG	103	West and Central Africa		63	16	25	40	63	7	11
CRF06 cpx	15	West and Central Africa		10	5		3		2	
CRF10 CD	2	Tanzania		2	0		1		1	
CRF11 cpx	3	Central Africa, Greece		2	0		1		1	
CRF12 BF	1	South America		1	1		0		0	
CRF13 cpx	2	West and Central Africa		1	0		1		0	
CRF18 cpx	7	Cuba		6	3		0		3	
GKU	1	Sweden		1	0		0		1	
Unclassified	191	NA		110	23	21	NA	NA	87	79
Total	5,675			4,091						

[a] Assignments to CRFs using the classical approach required recombinant breakpoints in *pol*.
[b] Strain groupings were nested within established subtypes/CRFs. The number of sequences assigned to strains within subtypes may be less than the total number of sequences assigned to that subtype.
[c] %, for strains for which >20 sequences annotated by patient nationality were identified, the percentage of annotated sequences obtained from nationals of the United Kingdom, from nationals of countries within the geographic region associated with the strain, or from nationals of other countries is shown.
[d] For subtype B, sequences from United Kingdom nationals are included in the proportion of sequences from the strain-associated region.
[e] NA, not applicable.

within the United Kingdom (Fig. 2). All three clusters comprised a mixture of resistant and nonresistant sequences, and there was no obvious tendency for sequences to group together according to the resistance mutations that they contained. Within each cluster, the only surveillance drug resistance mutation observed in more than one sequence was K103N (observed in two subtype G sequences), and clusters could be robustly recovered if this position was excluded from phylogenetic analysis.

Two of the three clusters were comprised of sequences belonging to the Kenyan strain of subtype A ($n = 9, 12$). In these clusters, >70% of infected individuals were Caucasian and > 50% were from the United Kingdom, providing further indication that at least some of the infections had been acquired

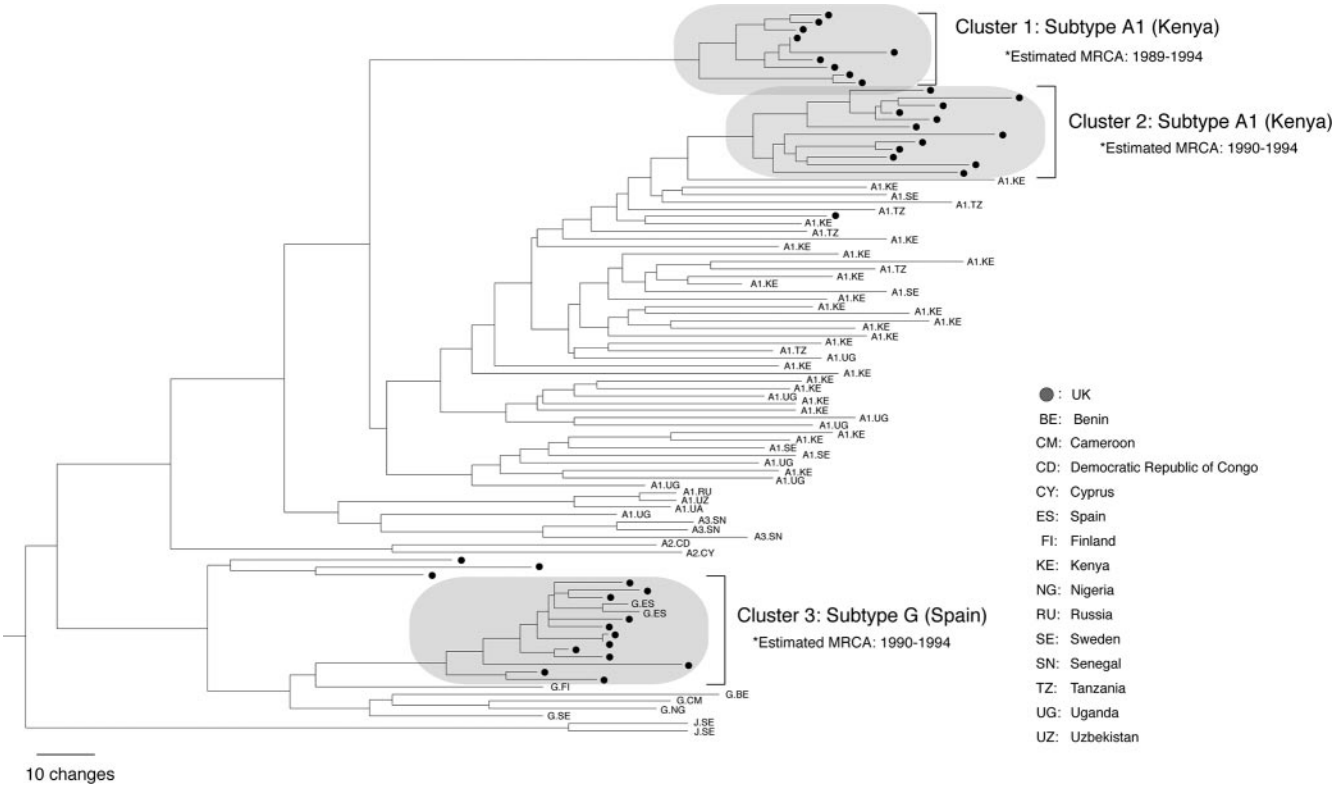

FIG. 2. Clusters of epidemiologically linked HIV-1 infections in the United Kingdom involving recently introduced strains. Epidemiologically linked infections were identified by statistically significant ($P < 0.01$) clustering of sequences sampled in the United Kingdom (UK) relative to globally sampled sequences in phylogenies representative of characterized diversity among the HIV-1 strains involved (East African subtype A [97 sequences] and West African/Iberian subtype G [56 sequences]). These clusters are shown in the phylogeny above with a representative set of globally sampled sequences. Bootstrap support for clusters was >70%. Filled circles indicate sequences from the United Kingdom. Globally sampled reference sequences are labeled according to subtype and country of sampling. The estimated date of the most recent common ancestor (MRCA) is shown for each cluster.

locally. Additionally, the documented of exposure risk in >90% of these cases was sex between men, which also suggests domestic transmission; for the vast majority of subtype A infections acquired in Africa, the probable means of exposure is reported as heterosexual sex (6).

The third cluster was comprised of 12 subtype G sequences obtained from individuals of either Portuguese or Angolan origin, half of whom were intravenous drug users. Although local transmission cannot be ruled out, the demographic data in this case suggested that clustering could reflect biased import, possibly via a network of intravenous drug users in the United Kingdom sharing a connection with Portugal.

Coalescent analysis indicated that all three clusters were established approximately 12 to 20 years ago (Fig. 2). This compares with estimates of ~30 years for the established subtype B lineages in the United Kingdom (18).

DISCUSSION

It is important to monitor the emerging genetic diversity of HIV-1, not only because it has implications for vaccine development, diagnosis, screening of blood products, and the selection of optimal treatment regimens but also because it will facilitate epidemiological investigation of transmission patterns and help define strategies for preventing the spread of infection (3, 5, 15, 16, 18, 25). The distribution of HIV-1 genetic diversity with respect to epidemiological factors such as risk group and geographic location is highly dynamic; novel genetic diversity is continually being generated through mutation and recombination, and travel and migration promote the transfer of diverse viral strains between populations, often across large distances (19, 26, 37). The speed at which genetic diversity is generated by HIV-1 presents a challenge to standard phylogenetic classification systems, as reflected in the growing number of unclassifiable and complex recombinant sequences being reported (13, 34–36, 38, 44), the proliferation of "sub-subtype" nomenclature (11, 23, 39, 45), and accumulating evidence that at least some of the established groupings are artifacts of sampling (1, 2, 10).

In this analysis, we develop a pragmatic approach to the classification of HIV-1 genetic diversity that is tailored to the purposes of epidemiological surveillance. Annotated complete-genome sequences were used to derive groupings focused on representing shared geographic associations among closely related strains, rather than attempting to definitively represent evolutionary relationships. This classification is unique in that it is robust to the effects of recombination. Although it was essential that complete genomes were used for the initial characterization of global diversity (as only complete

or nearly complete sequences can definitively discriminate monophyletic lineages in a population that is continuously being intermixed by recombination), reconstruction of phylogenies using subgenomic regions demonstrated that all groups defined by complete-genome analysis could be recovered using only these regions and the principles that we apply (Fig. 1). Thus, subgenomic fragments could be assigned to recombinant strains. Of course, when assigning subgenomic sequences to strains in this way, we could not rule out the possibility that certain of the sequences were misclassified due to uncharacterized recombination in the parts of the genome that were not analyzed. However, this would not invalidate the epidemiological information that we aim to infer, as robust genetic relatedness between strains, even in a subgenomic fragment, implies an epidemiological link. Furthermore, providing that prevalent recombinant viruses continue to be reported and fully sequenced with regularity, such cases will be relatively rare and unlikely to qualitatively affect our conclusions.

Fine-scale classification of 5,675 *pol* gene sequences using strain-level groupings revealed the underlying complexity of the HIV-1 epidemic in the United Kingdom. For example, for subtypes A, C, and G, the epidemic within the United Kingdom reflects the introduction of infections from two or more geographically distinct populations (Table 1). Reference to patient data confirmed that the geographic associations inferred through strain-level classification were generally concordant with the demographic profiles of infected individuals, which reinforces the epidemiological identities of the strains that we define.

A phylogenetic exploration of transmission dynamics indicated that the majority of non-B infections in the United Kingdom reflect separate introductions through travel and migration. However, the power of a molecular phylogenetic approach to detect epidemiological shifts of potential significance was illustrated by the identification of two transmission chains involving subtype A strains that are usually associated with heterosexual infections acquired in East Africa (Fig. 2). The observation that >90% of the sequences in these clusters were obtained from patients whose exposure category was defined as sex between men indicates that subtype A, or a novel recombinant epidemiologically linked to it, is spreading within the United Kingdom via the route of men having sex with men. The estimated origin of these two clusters in the late 1980s and early 1990s is concordant with existing epidemiological data, as this period was when the African epidemic was growing at its fastest rate and was prior to the widespread rollout of highly active antiretroviral therapy in the United Kingdom.

HIV-1 *pol* gene sequence data obtained during routine genotypic resistance testing is increasingly abundant in many countries throughout the world. This report illustrates how such opportunistically sampled data can be employed to monitor changes in the molecular epidemiology of national HIV epidemics. Since many sequences are obtained from treated patients (about half of the sequences in our data set), and some level of transmitted drug resistance is likely to be present within the untreated population, these data sets inevitably contain some level of homoplasy (i.e., convergent/parallel evolution) introduced by drug selection. However, previous studies have demonstrated that such data nevertheless contain sufficient phylogenetic signal for subtype assignment and for reconstruction of transmission history (16, 43), and this was reconfirmed here. For the latter purpose, we emphasize the need for careful analysis with respect to shared resistance mutations (17).

As the classification that we developed in this report is dependent on annotated full-genome sequences, the 36 geographic strains that we identify likely represent an under-sampling of HIV-1 diversity. However, given current trends toward more-efficient high-throughput sequencing technologies, we anticipate that more-representative sets of HIV-1 genome sequences will become available in the future. This will allow a richer characterization of the epidemiological associations among strains and enable further detailed characterization of epidemics to be carried out using the framework implemented in this report.

ACKNOWLEDGMENTS

UK Collaborative Group on HIV Drug Resistance Steering Committee members are Sheila Burns, City Hospital, Edinburgh; Sheila Cameron, Gartnavel General Hospital, Glasgow; Patricia Cane, Health Protection Agency, Porton Down; Ian Chrystie, Guy's and St. Thomas' NHS Foundation Trust, London; Duncan Churchill, Brighton and Sussex University Hospitals NHS Trust; Valerie Delpech and Deenan Pillay, Health Protection Agency-Centre for Infections, London; David Dunn, Esther Fearnhill, Hannah Green, and Kholoud Porter, MRC Clinical Trials Unit (Coordinating Centre). London; Philippa Easterbrook and Mark Zuckerman, King's College Hospital, London; Anna Maria Geretti, Royal Free NHS Trust, London; Rob Gifford, Paul Kellam, Deenan Pillay, Andrew Phillips, and Caroline Sabin, Royal Free and University College Medical School, London; David Goldberg, Health Protection Scotland, Glasgow; Mark Gompels, Southmead Hospital, Bristol; Antony Hale, Leeds Teaching Hospitals NHS Trust; Steve Kaye, St. Marys Hospital, London; Andrew Leigh-Brown, University of Edinburgh, Edinbrugh; Chloe Orkin, St. Bartholemews Hospital, London; Anton Pozniak, Chelsea & Westminster Hospital, London; Gerry Robb, Department of Health, London; Erasmus Smit, Health Protection Agency, Birmingham Heartlands Hospital, Birmingham; Peter Tilston, Manchester Royal Infirmary, Manchester; Ian Williams, Mortimer Market Centre, London.

The UK HIV Drug Resistance Database is partially funded by the Department of Health.

The views expressed here are those of the authors and not necessarily those of the Department of Health.

We thank all the clinicians, virologists, data managers, and research nurses in participating centers who assisted with the provision of data.

REFERENCES

1. **Abecasis, A. B., P. Lemey, N. Vidal, T. de Oliveira, M. Peeters, R. Camacho, B. Shapiro, A. Rambaut, and A. M. Vandamme.** 2007. Recombination is confounding the early evolutionary history of human immunodeficiency virus type 1: subtype G is a circulating recombinant form. J. Virol. **81:**8543–8551.
2. **Anderson, J. P., A. G. Rodrigo, G. H. Learn, A. Madan, C. Delahunty, M. Coon, M. Girard, S. Osmanov, L. Hood, and J. I. Mullins.** 2000. Testing the hypothesis of a recombinant origin of human immunodeficiency virus type 1 subtype E. J. Virol. **74:**10752–10765.
3. **Bennett, D.** 2005. HIV [corrected] genetic diversity surveillance in the United States. J. Infect. Dis. **192:**4–9.
4. **de Oliveira, T., K. Deforche, S. Cassol, M. Salminen, D. Paraskevis, C. Seebregts, J. Snoeck, E. J. van Rensburg, A. M. Wensing, D. A. van de Vijver, C. A. Boucher, R. Camacho, and A. M. Vandamme.** 2005. An automated genotyping system for analysis of HIV-1 and other microbial sequences. Bioinformatics **21:**3797–3800.
5. **de Oliveira, T., O. G. Pybus, A. Rambaut, M. Salemi, S. Cassol, M. Ciccozzi, G. Rezza, G. C. Gattinara, R. D'Arrigo, M. Amicosante, L. Perrin, V. Colizzi, and C. F. Perno.** 2006. Molecular epidemiology: HIV-1 and HCV sequences from Libyan outbreak. Nature **444:**836–837.
6. **Dougan, S., V. L. Gilbart, K. Sinka, and B. G. Evans.** 2005. HIV infections acquired through heterosexual intercourse in the United Kingdom: findings from national surveillance. BMJ **330:**1303–1304.
7. **Drummond, A. J., G. K. Nicholls, A. G. Rodrigo, and W. Solomon.** 2002.

Estimating mutation parameters, population history and genealogy simultaneously from temporally spaced sequence data. Genetics **161:**1307–1320.

8. **Esteves, A., R. Parreira, J. Piedade, T. Venenno, M. Franco, J. Germano de Sousa, L. Patricio, P. Brum, A. Costa, and W. F. Canas-Ferreira.** 2003. Spreading of HIV-1 subtype G and envB/gagG recombinant strains among injecting drug users in Lisbon, Portugal. AIDS Res. Hum. Retrovir. **19:**511–517.
9. **Esteves, A., R. Parreira, T. Venenno, M. Franco, J. Piedade, J. Germano De Sousa, and W. F. Canas-Ferreira.** 2002. Molecular epidemiology of HIV type 1 infection in Portugal: high prevalence of non-B subtypes. AIDS Res. Hum. Retrovir. **18:**313–325.
10. **Gao, F., D. L. Robertson, C. D. Carruthers, Y. Li, E. Bailes, L. G. Kostrikis, M. O. Salminen, F. Bibollet-Ruche, M. Peeters, D. D. Ho, G. M. Shaw, P. M. Sharp, and B. H. Hahn.** 1998. An isolate of human immunodeficiency virus type 1 originally classified as subtype I represents a complex mosaic comprising three different group M subtypes (A, G, and I). J. Virol. **72:**10234–10241.
11. **Gao, F., N. Vidal, Y. Li, S. A. Trask, Y. Chen, L. G. Kostrikis, D. D. Ho, J. Kim, M. D. Oh, K. Choe, M. Salminen, D. L. Robertson, G. M. Shaw, B. H. Hahn, and M. Peeters.** 2001. Evidence of two distinct subsubtypes within the HIV-1 subtype A radiation. AIDS Res. Hum. Retrovir. **17:**675–688.
12. **Gifford, R., T. de Oliveira, A. Rambaut, R. E. Myers, C. V. Gale, D. Dunn, R. Shafer, A. M. Vandamme, P. Kellam, and D. Pillay.** 2006. Assessment of automated genotyping protocols as tools for surveillance of HIV-1 genetic diversity. AIDS **20:**1521–1529.
13. **Gomez-Carrillo, M., J. F. Quarleri, A. E. Rubio, M. G. Carobene, D. Dilernia, J. K. Carr, and H. Salomon.** 2004. Drug resistance testing provides evidence of the globalization of HIV type 1: a new circulating recombinant form. AIDS Res. Hum. Retrovir. **20:**885–888.
14. **Hemelaar, J., E. Gouws, P. D. Ghys, and S. Osmanov.** 2006. Global and regional distribution of HIV-1 genetic subtypes and recombinants in 2004. AIDS **20:**W13–W23.
15. **Hu, D. J., T. J. Dondero, M. A. Rayfield, J. R. George, G. Schochetman, H. W. Jaffe, C. C. Luo, M. L. Kalish, B. G. Weniger, C. P. Pau, C. A. Schable, and J. W. Curran.** 1996. The emerging genetic diversity of HIV. The importance of global surveillance for diagnostics, research, and prevention. JAMA **275:**210–216.
16. **Hue, S., J. P. Clewley, P. A. Cane, and D. Pillay.** 2004. HIV-1 pol gene variation is sufficient for reconstruction of transmissions in the era of antiretroviral therapy. AIDS **18:**719–728.
17. **Hue, S., J. P. Clewley, P. A. Cane, and D. Pillay.** 2005. Investigation of HIV-1 transmission events by phylogenetic methods: requirement for scientific rigour. AIDS **19:**449–450.
18. **Hue, S., D. Pillay, J. P. Clewley, and O. G. Pybus.** 2005. Genetic analysis reveals the complex structure of HIV-1 transmission within defined risk groups. Proc. Natl. Acad. Sci. USA **102:**4425–4429.
19. **Lasky, M., J. L. Perret, M. Peeters, F. Bibollet-Ruche, F. Liegeois, D. Patrel, S. Molinier, C. Gras, and E. Delaporte.** 1997. Presence of multiple non-B subtypes and divergent subtype B strains of HIV-1 in individuals infected after overseas deployment. AIDS **11:**43–51.
20. **Liitsola, K., I. Tashkinova, T. Laukkanen, G. Korovina, T. Smolskaja, O. Momot, N. Mashkilleyson, S. Chaplinskas, H. Brummer-Korvenkontio, J. Vanhatalo, P. Leinikki, and M. O. Salminen.** 1998. HIV-1 genetic subtype A/B recombinant strain causing an explosive epidemic in injecting drug users in Kaliningrad. AIDS **12:**1907–1919.
21. **Lukashov, V. V., and J. Goudsmit.** 2002. Recent evolutionary history of human immunodeficiency virus type 1 subtype B: reconstruction of epidemic onset based on sequence distances to the common ancestor. J. Mol. Evol. **54:**680–691.
22. **Maddison, W. P., and D. R. Maddison.** 1992. MacClade: analysis of phylogeny and character evolution. Sinauer, Sunderland, MA.
23. **Meloni, S. T., B. Kim, J. L. Sankale, D. J. Hamel, S. Tovanabutra, S. Mboup, F. E. McCutchan, and P. J. Kanki.** 2004. Distinct human immunodeficiency virus type 1 subtype A virus circulating in West Africa: sub-subtype A3. J. Virol. **78:**12438–12445.
24. **Ndembi, N., J. Takehisa, L. Zekeng, E. Kobayashi, C. Ngansop, E. M. Songok, S. Kageyama, T. Takemura, E. Ido, M. Hayami, L. Kaptue, and H. Ichimura.** 2004. Genetic diversity of HIV type 1 in rural eastern Cameroon. J. Acquir. Immune Defic. Syndr. **37:**1641–1650.
25. **Peeters, M., C. Toure-Kane, and J. N. Nkengasong.** 2003. Genetic diversity of HIV in Africa: impact on diagnosis, treatment, vaccine development and trials. AIDS **17:**2547–2560.
26. **Perrin, L., L. Kaiser, and S. Yerly.** 2003. Travel and the spread of HIV-1 genetic variants. Lancet Infect. Dis. **3:**22–27.
27. **Rambaut, A., D. Posada, K. A. Crandall, and E. C. Holmes.** 2004. The causes and consequences of HIV evolution. Nat. Rev. Genet. **5:**52–61.
28. **Saad, M. D., A. Al-Jaufy, R. R. Grahan, Y. Nadai, K. C. Earhart, J. L. Sanchez, and J. K. Carr.** 2005. HIV type 1 strains common in Europe, Africa, and Asia cocirculate in Yemen. AIDS Res. Hum. Retrovir. **21:**644–648.
29. **Shafer, R. W., S. Y. Rhee, D. Pillay, V. Miller, P. Sandstrom, J. M. Schapiro, D. R. Kuritzkes, and D. Bennett.** 2007. HIV-1 protease and reverse transcriptase mutations for drug resistance surveillance. AIDS **21:**215–223.
30. **Sides, T. L., O. Akinsete, K. Henry, J. T. Wotton, P. W. Carr, and J. Bartkus.** 2005. HIV-1 subtype diversity in Minnesota. J. Infect. Dis. **192:**37–45.
31. **Slatkin, M., and W. P. Maddison.** 1989. A cladistic measure of gene flow measured from the phylogenies of alleles. Genetics **123:**603–613.
32. **Soares, M. A., T. De Oliveira, R. M. Brindeiro, R. S. Diaz, E. C. Sabino, L. Brigido, I. L. Pires, M. G. Morgado, M. C. Dantas, D. Barreira, P. R. Teixeira, S. Cassol, and A. Tanuri.** 2003. A specific subtype C of human immunodeficiency virus type 1 circulates in Brazil. AIDS **17:**11–21.
33. **Swofford, D. L.** 1998. PAUP*. Phylogenetic analysis using parsimony (* and other methods), 4th ed. Sinauer Associates, Sunderland, MA.
34. **Takebe, Y., K. Motomura, M. Tatsumi, H. H. Lwin, M. Zaw, and S. Kusagawa.** 2003. High prevalence of diverse forms of HIV-1 intersubtype recombinants in Central Myanmar: geographical hot spot of extensive recombination. AIDS **17:**2077–2087.
35. **Tee, K. K., X. J. Li, K. Nohtomi, K. P. Ng, A. Kamarulzaman, and Y. Takebe.** 2006. Identification of a novel circulating recombinant form (CRF33_01B) disseminating widely among various risk populations in Kuala Lumpur, Malaysia. J. Acquir. Immune Defic. Syndr. **43:**523–529.
36. **Thomson, M. M., G. Casado, D. Posada, M. Sierra, and R. Najera.** 2005. Identification of a novel HIV-1 complex circulating recombinant form (CRF18_cpx) of Central African origin in Cuba. AIDS **19:**1155–1163.
37. **Thomson, M. M., and R. Najera.** 2005. Molecular epidemiology of HIV-1 variants in the global AIDS pandemic: an update. AIDS Rev **7:**210–24.
38. **Tovanabutra, S., G. H. Kijak, C. Beyrer, C. Gammon-Richardson, S. Sakkhachornphop, T. Vongchak, J. Jittiwutikarn, M. H. Razak, E. Sanders-Buell, M. L. Robb, V. Suriyanon, D. L. Birx, N. L. Michael, D. D. Celentano, and F. E. McCutchan.** 2007. Identification of CRF34_01B, a second circulating recombinant form unrelated to and more complex than CRF15_01B, among injecting drug users in northern Thailand. AIDS Res. Hum. Retrovir. **23:**829–833.
39. **Van der Auwera, G., W. Janssens, L. Heyndrickx, and G. van der Groen.** 2001. Reanalysis of full-length HIV type 1 group M subtype K and subsubtype F2 with an MS-DOS bootscanning program. AIDS Res. Hum. Retrovir. **17:**185–189.
40. **Vidal, N., D. Koyalta, V. Richard, C. Lechiche, T. Ndinaromtan, A. Djimasngar, E. Delaporte, and M. Peeters.** 2003. High genetic diversity of HIV-1 strains in Chad, West Central Africa. J. Acquir. Immune Defic. Syndr. **33:**239–246.
41. **Vidal, N., C. Mulanga, S. E. Bazepeo, J. K. Mwamba, J. W. Tshimpaka, M. Kashi, N. Mama, C. Laurent, F. Lepira, E. Delaporte, and M. Peeters.** 2005. Distribution of HIV-1 variants in the Democratic Republic of Congo suggests increase of subtype C in Kinshasa between 1997 and 2002. J. Acquir. Immune Defic. Syndr. **40:**456–462.
42. **Vidal, N., M. Peeters, C. Mulanga-Kabeya, N. Nzilambi, D. Robertson, W. Ilunga, H. Sema, K. Tshimanga, B. Bongo, and E. Delaporte.** 2000. Unprecedented degree of human immunodeficiency virus type 1 (HIV-1) group M genetic diversity in the Democratic Republic of Congo suggests that the HIV-1 pandemic originated in Central Africa. J. Virol. **74:**10498–10507.
43. **Yahi, N., J. Fantini, C. Tourres, N. Tivoli, N. Koch, and C. Tamalet.** 2001. Use of drug resistance sequence data for the systematic detection of non-B human immunodeficiency virus type 1 (HIV-1) subtypes: how to create a sentinel site for monitoring the genetic diversity of HIV-1 at a country scale. J. Infect. Dis. **183:**1311–1317.
44. **Yang, R., X. Xia, S. Kusagawa, C. Zhang, K. Ben, and Y. Takebe.** 2002. On-going generation of multiple forms of HIV-1 intersubtype recombinants in the Yunnan Province of China. AIDS **16:**1401–1407.
45. **Zhang, M., K. Wilbe, N. D. Wolfe, B. Gaschen, J. K. Carr, and T. Leitner.** 2005. HIV type 1 CRF13_cpx revisited: identification of a new sequence from Cameroon and signal for subsubtype J2. AIDS Res. Hum. Retrovir. **21:**955–960.
46. **Zhong, P., S. Burda, F. Konings, M. Urbanski, L. Ma, L. Zekeng, L. Ewane, L. Agyingi, M. Agwara, Saa, Z. E. Afane, T. Kinge, S. Zolla-Pazner, and P. Nyambi.** 2003. Genetic and biological properties of HIV type 1 isolates prevalent in villagers of the Cameroon equatorial rain forests and grass fields: further evidence of broad HIV type 1 genetic diversity. AIDS Res. Hum. Retrovir. **19:**1167–1178.