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

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...
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 <0.2% prevalence.

**Strain-level classification of HIV-1 diversity.** Nineteen 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 219 in RT. Surveillance drug resistance mutations at other positions occurred at a <0.2% prevalence.

**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-
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.
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

### Table 1. Strain-level assignment of 5,675 HIV-1 pol sequences from the United Kingdom and distribution among nationalities

<table>
<thead>
<tr>
<th>Classical subtype or CRF group&lt;sup&gt;a&lt;/sup&gt;</th>
<th>Total no. of sequences</th>
<th>Strain-level grouping and/or geographic association(s)</th>
<th>No. of sequences assigned to group&lt;sup&gt;b&lt;/sup&gt;</th>
<th>No. of sequences for which patient nationality data were available&lt;sup&gt;c&lt;/sup&gt;</th>
<th>Patient nationality&lt;sup&gt;d&lt;/sup&gt;</th>
<th>Other</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td></td>
<td>No. of sequences</td>
<td>%</td>
<td>No. of sequences</td>
<td>%</td>
</tr>
<tr>
<td>Subtype A</td>
<td>362</td>
<td>A1 (Kenya)</td>
<td>122</td>
<td>85</td>
<td>15</td>
<td>17</td>
</tr>
<tr>
<td></td>
<td></td>
<td>A2 (West Africa)</td>
<td>2</td>
<td>1</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td></td>
<td></td>
<td>A1 (Central Asia)</td>
<td>1</td>
<td>1</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td></td>
<td></td>
<td>CRF01 (Southeast Asia)</td>
<td>39</td>
<td>30</td>
<td>14</td>
<td>46</td>
</tr>
<tr>
<td>Subtype B</td>
<td>4,252</td>
<td>North and South America, Western Europe, Caribbean</td>
<td>3,150</td>
<td></td>
<td>1,963&lt;sup&gt;d&lt;/sup&gt;</td>
<td>62&lt;sup&gt;d&lt;/sup&gt;</td>
</tr>
<tr>
<td>Subtype C</td>
<td>568</td>
<td>Southern and East Africa</td>
<td>498</td>
<td>327</td>
<td>48</td>
<td>15</td>
</tr>
<tr>
<td></td>
<td></td>
<td>South America</td>
<td>12</td>
<td>6</td>
<td>3</td>
<td>2</td>
</tr>
<tr>
<td></td>
<td></td>
<td>India</td>
<td>2</td>
<td>2</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>Subtype D</td>
<td>93</td>
<td>Uganda</td>
<td>62</td>
<td>46</td>
<td>3</td>
<td>6</td>
</tr>
<tr>
<td>Subtype F</td>
<td>12</td>
<td>F1 (South America, Central Africa)</td>
<td>10</td>
<td>7</td>
<td>0</td>
<td>5</td>
</tr>
<tr>
<td></td>
<td></td>
<td>F2 (Cameroon)</td>
<td>2</td>
<td>2</td>
<td>0</td>
<td>2</td>
</tr>
<tr>
<td>Subtype G</td>
<td>54</td>
<td>Spain</td>
<td>17</td>
<td>13</td>
<td>4</td>
<td>7</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Nigeria</td>
<td>12</td>
<td>9</td>
<td>1</td>
<td>7</td>
</tr>
<tr>
<td>Subtype H</td>
<td>4</td>
<td>Central Africa</td>
<td>3</td>
<td>0</td>
<td>0</td>
<td>2</td>
</tr>
<tr>
<td>Subtype J</td>
<td>5</td>
<td>Central Africa</td>
<td>3</td>
<td>0</td>
<td>0</td>
<td>2</td>
</tr>
<tr>
<td>CRF02 AG</td>
<td>103</td>
<td>West and Central Africa</td>
<td>63</td>
<td>16</td>
<td>25</td>
<td>40</td>
</tr>
<tr>
<td>CRF06 cpx</td>
<td>15</td>
<td>West and Central Africa</td>
<td>10</td>
<td>5</td>
<td>5</td>
<td>3</td>
</tr>
<tr>
<td>CRF10 CD</td>
<td>2</td>
<td>Tanzania</td>
<td>2</td>
<td>0</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>CRF11 cpx</td>
<td>3</td>
<td>Central Africa, Greece</td>
<td>2</td>
<td>0</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>CRF12 BF</td>
<td>1</td>
<td>South America</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td>CRF13 cpx</td>
<td>2</td>
<td>West and Central Africa</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>CRF18 cpx</td>
<td>7</td>
<td>Cuba</td>
<td>6</td>
<td>3</td>
<td>0</td>
<td>3</td>
</tr>
<tr>
<td>GKU</td>
<td>1</td>
<td>Sweden</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>Unclassified</td>
<td>191</td>
<td>NA</td>
<td>110</td>
<td>23</td>
<td>21</td>
<td>NA</td>
</tr>
<tr>
<td>Total</td>
<td>5,675</td>
<td>NA</td>
<td>4,091</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

<sup>a</sup> Assignments to CRFs using the classical approach required recombinant breakpoints in pol.

<sup>b</sup> 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.

<sup>c</sup> %, 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.

<sup>d</sup> For subtype B, sequences from United Kingdom nationals are included in the proportion of sequences from the strain-associated region.

<sup>e</sup> NA, not applicable.
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 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, Southend Hospital, Bristol; Antony Hale, Leeds Teaching Hospitals NHS Trust; Steve Kaye, St. Marys Hospital, London; Andrew Leigh-Brown, University of Edinburgh, Edinburgh; Chloe Orkin, St. Bartholomews Hospital, London; Anton Pozniak, Chelsea & Westminster Hospital, London; Gerry Robb, Department of Health, London; Erasmus Smit, Health Protection Agency, Birmingham Heartlands Hospitals, 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

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


