Genetic analysis reveals the complex structure of HIV-1 transmission within defined risk groups

Stéphane Hué1*, Deenan Pillay1†, Jonathan P. Clewley2, and Oliver G. Pybus3

1Centre for Virology, Division of Infection and Immunity, University College London, 46 Clevelend Street, London W1T 4JF, United Kingdom; 2Centre for Infections, Health Protection Agency, 61 Colindale Avenue, London NW9 5HT, United Kingdom; and 3Department of Zoology, University of Oxford, South Parks Road, Oxford OX1 3PS, United Kingdom

Edited by Burton H. Singer, Princeton University, Princeton, NJ, and approved February 7, 2005 (received for review October 11, 2004)

We explored the epidemic history of HIV-1 subtype B in the United Kingdom by using statistical methods that infer the population history of pathogens from sampled gene sequence data. Phylogenetic analysis of HIV-1 pol gene sequences from Britain showed at least six large transmission chains, indicating a genetically variable, but epidemiologically homogeneous, epidemic among men having sex with men. Through coalescent-based analysis, we showed that these chains arose through separate introductions of subtype B strains into the United Kingdom in the early to mid-1980s. After an initial period of exponential growth, the rate of spread generally slowed in the early 1990s, which is more likely to correlate with behavior change than with reduced infectiousness resulting from highly active antiretroviral therapy. Our results provide insights into the complexity of HIV-1 epidemics that must be considered when developing HIV monitoring and prevention initiatives.

Methods

Study Population. HIV-1 subtype B pol gene sequences were generated from plasma samples collected in the U.K. by the Health Protection Agency’s Antiviral Susceptibility Reference Unit. The samples were submitted for routine genotypic drug resistance testing between 1999 and 2003 and included samples from acute infections, chronic but drug-naı¨ve infections, and from patients at the time of therapy failure. The sequences were 952 bp long, including the full protease gene and the first 218 codons of the reverse transcriptase gene. Approximately 85% of these sequences were from men who had sex with men (MSM).

Phylogenetic Reconstruction. To identify HIV-1 lineages derived from single independent introductions of the virus into the U.K. population, a neighbor-joining phylogenetic tree was constructed from 3,429 HIV-1 subtype B pol gene sequences (1,645 U.K. isolates plus 1,784 subtype B reference sequences from throughout the world) (14). The tree was estimated under the Hasegawa–Kishino–Yano model of nucleotide substitution (15). The non-U.K. sequences used for the study were extracted from GenBank (www.ncbi.nlm.nih.gov) and the Los Alamos HIV Sequence Database (www.hiv.lanl.gov). The size of the sequence alignment and the computational power required prevented the use of a more complex evolutionary model.

After identification of U.K. transmission clusters, sequences of non-U.K. origin were removed and the phylogenies of the clusters were reestimated with the program PAUP* by using a maximum likelihood approach (16). The trees were constructed under the General Time Reversible model of nucleotide substitution (17), with proportion of invariable sites and substitution rate heterogeneity, as selected by the program MODELTEST (18). Each U.K. cluster was rooted by using a subtype D pol sequence from our database. The statistical robustness of the maximum likelihood topologies was assessed by bootstrapping with 1,000
results for coalescent analyses, the program TIPDATE (23) was used to model. To fit a constant molecular clock framework, as required (constant population size followed by exponential growth), and piecewise expansion (constant population size followed by exponential growth phase). See ref. 16 for more details of these models. To fit a constant molecular clock framework, as required for coalescent analyses, the program TIPDATE (23) was used to rescale each transmission tree under the Single Rate Dated Tip (SRDT) model.

Estimation of Demographic History and Population Dynamics. The investigation of the epidemic history of the six U.K. clusters involved two steps. First, several models of demographic history, each of which illustrate effective numbers of infections through time, were compared to select the model that best describes the epidemiological history of the U.K. transmission clusters. The demographic models were evaluated by the likelihood ratio test from likelihoods calculated by the program GENIE (22). The five models tested in this study were constant population size, exponential growth, piecewise logistic (exponential growth followed by constant population size), piecewise expansion (constant population size followed by exponential growth), and piecewise con-exp-con (constant growth periods flanking an exponential growth phase). See ref. 16 for more details of these models. To fit a constant molecular clock framework, as required for coalescent analyses, the program TIPDATE (23) was used to rescale each transmission tree under the Single Rate Dated Tip (SRDT) model.

Second, the demographic and evolutionary parameters of the epidemic, together with their confidence intervals, were estimated by Bayesian MCMC inference by using a chain of 10 million states sampled at every 100th generation, as implemented in the program BEAST. The estimated parameters include the date of the most recent common ancestor of the cluster, the effective number of infections at the most recent time of sampling Ne (i.e., the effective number of prevalent infections), and the growth rate during the exponential phase r. The Bayesian MCMC results were used to calculate a marginal posterior distribution of the demographic model for each cluster, a graphical representation of the effective number of infections through time, generated by using the program TRACER (http://evolve.zoo.ox.ac.uk/software.html?iid=tracer).

Results

Introduction of HIV-1 Subtype B into the U.K. The initial neighboring phylogenetic tree constructed from 3,429 U.K. and worldwide subtype B pol sequences is too large to display here (Data Set 1, which is published as supporting information on the PNAS web site). A schematic representation of the clustering patterns seen within the phylogeny is presented in Fig. 1. Three clustering patterns were distinguished, namely sporadic U.K. sequences, non-U.K. transmission clusters, and U.K. transmission clusters. Sporadic U.K. sequences (i.e., those that do not group with other U.K. lineages in the tree) probably represent single, independent introductions of the virus without subsequent spread. Transmission clusters were identified as clades of sequences from a particular location that descend from a common ancestor, indicating spread of the virus in that region. U.K. transmission clusters were differentiated from non-U.K. clusters on the basis of the size of the clade and the proportion of U.K. sequences within it: U.K. transmission clusters were defined as those clades with >25 sequences, 90% or more of which were of U.K. origin. A minimum clade size of 25 was used because smaller sample sizes are unlikely to give reliable coalescent estimates under complex demographic models. A minimum fraction of 90% U.K. sequences was chosen to ensure that the clusters that were identified represent chains of transmission that have overwhelmingly occurred in the U.K. However, we note that this methodology probably underestimates the number of transmission chains identified.

Most of the U.K. sequences represented sporadic lineages (86%) scattered among sequences from other geographical areas, suggesting much geographical mixing and migration of subtype B strains on a worldwide scale. Nonetheless, six U.K. transmission clusters were identified, involving 45, 62, 29, 26, 27, and 34 sequences. These transmission chains were distinct (i.e., reciprocally monophyletic), indicating that at least six independent introductions of subtype B HIV-1 have succeeded in sustaining onward transmission within the U.K. over time and until the present. Each transmission chain contained sequences from a variety of locations within the U.K., and no obvious geographic correlations were observed. The robustness of the clusters within the overall tree could not be statistically evaluated because of the huge size of the data set. Nonetheless, the
branching patterns of the six U.K. lineages showed statistical robustness when compared with subsets of worldwide control sequences by using bootstrap analyses (neighbor-joining method with 1,000 replicates, as implemented in the program PAUP*; data not shown). To further explore the history of these six successful viral lineages, sequences of non-U.K. origin were removed from the six clusters, and the phylogenetic histories of the U.K. sequences were rigorously reestimated by using a maximum likelihood approach. The maximum likelihood trees are available from the authors on request.

Fig. 2. Phylogenetic trees of the six U.K. transmission clusters and their corresponding estimated epidemic histories (all shown on the same time scale). The trees represent the ancestral relationships of sequences belonging to each cluster. (a) Cluster 1. (b) Cluster 2. (c) Cluster 3. (d) Cluster 4. (e) Cluster 5. (f) Cluster 6. The demographic histories were estimated by Bayesian MCMC inference with a model of logistic growth (see text for details) and show change in the effective number of infections through time (time scale in calendar years). The blue line shows the median estimate of the effective number of infections, whereas the black lines show the 95% confidence limits of the estimate.
Estimation of the Rate of Evolution for the HIV-1 Subtype B pol Gene.

The rate of evolution for the subtype B HIV-1 pol gene was calculated by using an independent data set of 106 sequences, sampled between 1983 and 2000 in Amsterdam (20). Using a Bayesian MCMC method, this rate was estimated to be 2.55 × 10⁻³ substitutions per nucleotide site per year (95% confidence intervals: 1.74 × 10⁻³ to 3.51 × 10⁻³). In comparison, previous estimates of HIV-1 evolution rates have typically relied on partial env gene sequences and have ranged from 2.4 × 10⁻³ to 6.7 × 10⁻³ substitutions per site per year (24–26). Our estimate is consistent with the order of magnitude of 10⁻³ expected for an HIV-1 gene. The phylogenetic trees in Fig. 2 are thus shown on a time scale of years.

Epidemic History and Parameter Estimation. For each of the six clusters, a model of logistic population growth best fitted the demographic information contained in the tree topologies (likelihoods shown in Table 2, which is published as supporting information on the PNAS web site). Under the logistic model, the effective number of infections Ne grows exponentially at rate r from time t₀ (time of the most recent common ancestor of the cluster) then decreases in growth rate toward the present. A schematic representation of the logistic model is given in Fig. 3. Note that Ne reflects the number of infections contributing to new infections, rather than the total number of prevalent infections within the transmission cluster.

Table 1. Parametric estimates under the logistic growth demographic model for the six lineages

<table>
<thead>
<tr>
<th>Cluster</th>
<th>μ</th>
<th>Ne, no. of infections</th>
<th>r, yr⁻¹</th>
<th>Origin of the tree, yr</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>2.55 × 10⁻³ (0.0017–0.0035)</td>
<td>493 (201–833)</td>
<td>1.08 (0.66–2.56)</td>
<td>1983 (1978–1988)</td>
</tr>
<tr>
<td>2</td>
<td>2.55 × 10⁻³ (0.0017–0.0035)</td>
<td>386 (190–655)</td>
<td>0.47 (0.30–0.95)</td>
<td>1981 (1976–1987)</td>
</tr>
<tr>
<td>3</td>
<td>2.55 × 10⁻³ (0.0017–0.0035)</td>
<td>98 (42–171)</td>
<td>0.50 (0.19–4.62)</td>
<td>1983 (1977–1988)</td>
</tr>
<tr>
<td>4</td>
<td>2.55 × 10⁻³ (0.0017–0.0035)</td>
<td>250 (88–483)</td>
<td>1.38 (0.63–2.50)</td>
<td>1986 (1982–1991)</td>
</tr>
<tr>
<td>5</td>
<td>2.55 × 10⁻³ (0.0017–0.0035)</td>
<td>94 (36–85)</td>
<td>0.68 (0.35–2.10)</td>
<td>1987 (1983–1991)</td>
</tr>
<tr>
<td>6</td>
<td>2.55 × 10⁻³ (0.0017–0.0035)</td>
<td>1,350 (109–5489)</td>
<td>0.67 (0.37–3.85)</td>
<td>1986 (1981–1991)</td>
</tr>
<tr>
<td>U.S.*</td>
<td>6.7 × 10⁻³ (0.0017–0.0035)</td>
<td>4,830 (1,995–26,750)</td>
<td>0.834 (0.72–0.945)</td>
<td>1968 (1966–1970)</td>
</tr>
</tbody>
</table>

Numbers in parentheses are the range with 95% confidence intervals. μ, rate of nucleotide substitution estimated from an independent data set of subtype B pol sequences. r, rate of exponential growth.

Discussion

Our estimates suggest that the HIV-1 subtype B epidemic currently circulating within the U.K. is comprised of at least six established chains of transmission, introduced in the early and mid-1980s. This finding demonstrates the existence of distinct, possibly nonoverlapping sexual networks within the predominant MSM risk group and argues against the hypothesis that one initial entry of HIV-1 was responsible for the spread of the subtype B epidemic. It also emphasizes the role of migration in the HIV-1 epidemic in Britain, as illustrated by the overwhelming prevalence of sporadic lineages (86% of the total U.K. samples) in the genealogy, representing viruses arising from outside the U.K. that have failed to establish a large outbreak.

The transmission clusters we characterized had similar epidemic curves and geographic distributions within the U.K., indicating a concurrent spread under similar demographic pressures, at least during the early stages of the epidemic. The introduction of the earlier viruses in the early 1980s (i.e., clusters 1–3) seems to coincide with the explosion of new infections reported by epidemiological data at the time (www.hpa.org.uk). The coupling of HIV strain “immigration” with epidemiological changes is likely to have favored the emergence and persistence of the transmission chains presently circulating amongst MSM.
However, the first U.K. cases of AIDS were reported in 1982 (www.who.int/emc-hiv/fact_sheets), and these individuals were probably infected within a window of 10 years before that time; hence, the currently circulating strains may not represent the first HIV-1 lineage identified within the U.K. If earlier strains existed, they may have been unsuccessful in sustaining transmission chains until the present and may no longer be of epidemiological significance. However, the absence of older strains could also reflect a sampling bias.

For all six transmission clusters, the exponential growth phase coincides with a reported augmentation of newly acquired HIV-1 infections within MSM and injecting drug users in the U.K. (www.hpa.org.uk). The average growth rate during the initial exponential phase was estimated to be 0.80 years⁻¹ (ranging from 0.47 to 1.38), approximating a doubling time of 1 year. This value is similar to that estimated for the United States subtype B epidemic (0.83 years⁻¹, 0.72–0.94), suggesting that the two epidemics follow similar trends at the macroepidemiological scale (26). This idea is supported by the effective number of infections estimated for the two epidemics. Despite a wide variation in Ne across the six U.K. transmission clusters, the average effective number of infections among the six U.K. clusters is 445, which is ~2.5% of the infected population. This result is remarkably similar to the values for the United States epidemic, where the effective number of infections and prevalence in 1995 reached 5,000 and 200,000 infections, respectively. Ne represents the number of infections contributing to onward transmission, rather than the larger number of actual infections. Importantly, we observe that the population represented by cluster 6 exhibits a faster doubling time in 2003 than the other five clusters, suggesting a difference in current growth rate among clusters. Current surveillance data for the U.K. reports a very recent increase in infections within MSM (www.hpa.org.uk), and it is reasonable to suppose that the lineage we have identified as cluster 6 has contributed to this recent upturn in infection.

Since 1990, there have been important changes in Britain’s demographic structure, social attitude, and awareness of HIV-1/AIDS (28). Despite a very recent increase in high-risk behavior among MSM (such as the number of sexual partners or concurrent partnerships), a significant increase in consistent condom use has been reported since 1990. Such a change in sexual health, coupled to large-scale educational campaigns over the past decade, could explain the equilibrium reached by the effective number of prevalent infections. The effect of antiretroviral therapy on past epidemic dynamics should also be considered: although such therapy is instituted primarily to reduce progression of disease, it may also impact on transmission through reduction of infectivity. If so, we would expect evidence of a growth rate decrease in the late (rather than early) 1990s, the time that highly active antiretroviral therapy became widely used. In fact, Health Protection Agency data suggests no significant changes in the incidence of HIV-1 within gay men since the late 1980s and an actual increase over the past three years (29). We therefore suggest that antiviral therapy has not had a significant impact on the growth of the epidemic; indeed, some studies suggest that the epidemic is driven by transmissions in primary infection (30–32), before therapy is usually initiated. The current increase in new infections is too recent to be reflected in the growth dynamics of any of the six populations identified by our analysis. Ongoing analyses of the type undertaken here will clarify whether the recent increase in new subtype B infections derive from longstanding viral lineages or newly introduced viruses.

In conclusion, we show that currently circulating HIV-1 subtype B strains entered the U.K. in the mid 1980s and that the rate of spread of these lineages slowed in the early 1990s. It is often assumed that the HIV-1 epidemic within the U.K. is composed of smaller, independent epidemics defined by risk group. We demonstrate here the existence of multiple subepidemics (at least six) within MSM that obey similar demographic constraints during their early stages, yet exhibit differences in their more recent rates of spread. The identification of these multiple lineages within the predominant risk group of the HIV-1 epidemic in the U.K. suggests the existence of subepidemics within groups of MSM, and it is reasonable to assume that this structure exists in comparable risk groups in other countries. Such heterogeneity must therefore be considered when developing HIV monitoring prevention and treatment initiatives.

We thank Dr. Patricia Cane from the Health Protection Agency Antiretroviral Susceptibility Reference Unit and Dr. Vladimir Lukashov from the Academic Medical Center of Amsterdam for kindly providing the sequences used in this study. Dr. Andy Rambaut for support in the phylogenetic analyses, and Prof. Robin Weiss and Dr. Paul Kellam for useful comments on the manuscript. This work was funded by the Health Protection Agency, U.K. O.G.P. was funded by the Wellcome Trust and the Royal Society.
