Assessment of automated genotyping protocols as tools for surveillance of HIV-1 genetic diversity

Robert Gifford[a], Tulio de Oliveira[b], Andrew Rambaut[b], Richard E. Myers[a], Catherine V. Gale[a], David Dunn[d], Robert Shafer[e], Anne-Mieke Vandamme[f], Paul Kellam[a], Deenan Pillay[a,c] on Behalf of the UK Collaborative Group on HIV Drug Resistance*

Background: The routine use of drug resistance testing provides an abundant source of HIV-1 sequence data. However, it is not clear how reliable standard genotyping of these sequences is for describing HIV-1 genetic variation and for detecting novel genetic variants and epidemiological trends.

Objectives: To compare assignment of HIV-1 resistance test sequences to reference strains across commonly used genotyping protocols.

Methods: Subtype assignments were compared across three standard genotyping protocols for 10 537 resistance test sequences, representing approximately one-fifth of all reported infections in the United Kingdom. Sequences that were inconsistently genotyped across methods, or that were unassigned by at least one method, were examined for evidence of recombination using sliding-window-based approaches.

Results: Although agreement across methods was high for subtypes B, C and H, it was generally much lower (< 50%) for other subtypes. Disagreement between methods typically involved closely related, but epidemiologically distinct, groups or involved a significant proportion (~12%) of divergent sequences in which analysis revealed widespread evidence of recombination and a remarkable diversity of unusual recombinant forms.

Conclusions: With frequent long-distance transfer of viral strains and widespread recombination between them, genetic and epidemiological relationships within HIV-1 are becoming increasingly complex. Current methods of subtype assignment vary in their ability to identify novel genetic variants and to distinguish epidemiologically distinct strains. Capturing meaningful epidemiological information from resistance test data will require a critical understanding of the methodologies used in order to appreciate the possible sources of error and misclassification.

AIDS 2006, **20**:1521–1529

Introduction

Phylogenetic analysis of HIV-1 isolates from Africa and other regions of the world reveals three major groups, M, O and N, each of which is proposed to have arisen independently via cross-species transfer from chimpanzees to humans [1–3]. Group M viruses are by far the most widespread and abundant, accounting for more than

From the [a]Department of Infection, University College London, the [b]Department of Zoology, University of Oxford, Oxford, the [c]Centres for Infection, Health Protection Agency, Colindale, the [d]Medical Research Council Clinical Trials Unit, London, UK, the [e]Division of Infectious Diseases, Department of Medicine, Stanford University, California, USA, and the [f]Rega Institute for Medical Research, K.U. Leuven, Belgium.

Correspondence to Dr R. Gifford, Centre for Virology, Department of Infection, Royal Free and University College School of Medicine, Windeyer Institute, 46 Cleveland St, London W1T 4JF, UK.

E-mail: r.gifford@ucl.ac.uk

* See Appendix for study members.

Received: 23 December 2006; accepted: 10 February 2006.

ISSN 0269-9370 © 2006 Lippincott Williams & Wilkins

95% of HIV-1 infections worldwide. The initial dissemination of HIV-1 group M was characterized by founder effects and sampling bias that led to a classification system based on nine subtypes (A, B, C, D, F, G, H, J and K) that are genetically distinct across the entire genome, and a further 16 circulating recombinant forms (CRFs) that have mosaic genomes with consistent inter-subtype breakpoints. Two lineages, originally classified as subtypes E and I, have subsequently been reclassified as CRFs 01 and 04, respectively [1,4,5]. CRF definition is based on sequencing of three complete genomes from epidemiologically unrelated individuals. Mosaic genomes for which these requirements have not been met are designated unique recombinant forms [6].

The subtype structure of the HIV-1 pandemic reflects the genetic founder effect associated with the global dissemination of a small number of group M strains via diverse epidemiological pathways [7]. Throughout the world, prevalence of the various group M subtypes varies dramatically, and often reflects risk group [6]. HIV-1 subtypes thus provide powerful epidemiological markers, and by indicating routes of transmission they may help to define strategies for intervention. Variation between subtypes also has important implications with regard to the design of diagnostic and vaccine strategies in areas where genetic diversity is high [8]. For these reasons, it is important to explore the potential of analytical tools to describe HIV-1 genetic variation accurately using available sequence data.

Although confident assignment of viral subtype requires full-length genome sequences, practical limitations mean that it is often based on subgenomic regions. Traditionally *gag* and *env* sequences have been used, as the genetic variability in these regions is relatively high. In recent years, however, the routine use of drug resistance testing in individuals with virological failure on therapy or before initiating therapy has dramatically increased the volume of *pol* data accrued, particularly for the protease (PR) and reverse transcriptase (RT) genes. One consequence of this has been the increasingly widespread use of resistance test sequences to assign viral subtype, particularly in clinical settings, where it can provide a context in which to improve interpretation of drug resistance mutations. For example, many minor protease inhibitor mutations in subtype B viruses are represented as the wild-type sequence for some non-B viruses [9] and, therefore, would not necessarily suggest prior protease inhibitor treatment. In addition, sequence variation between subtypes may influence the mutational routes to resistance, with implications for cross-resistance patterns [10].

Several studies have demonstrated the utility of resistance test data for assignment of viral subtype despite high levels of conservation in *pol* [11,12]. Resistance test sequences can thus provide a valuable source of information with regard to the shifting distribution and diversity of HIV-1, and the epidemiological factors that underlie it. However, with ongoing exchange of viruses between geographic areas, the situation will become increasingly complex. Not only will a broader diversity of HIV-1 strains begin to cocirculate within particular localities, increasing the likelihood of recombination between diverse strains, but also viruses that are genetically quite similar may become epidemiologically distinct through long-distance transfer. It is not clear how reliably we can detect and describe these subtle and complicated dynamics using resistance test sequences and standard genotyping protocols.

In order to assess the robustness of current approaches to classifying HIV-1 resistance test sequences, subtype assignments according to three distinct and widely used batch-genotyping protocols were compared for a large dataset. Data consisted of over 10 000 PR and RT sequences from resistance tests carried out at clinical centres around the United Kingdom since 1996, and representing approximately one-fifth of all reported UK infections. The United Kingdom represents a suitable test case to carry out a study such as this, because the socioeconomic and historical associations of the country with wide-ranging geographic areas contribute to an epidemiologically diverse range of HIV-1 infections, particularly in London [13,14].

Methods

Sequences

Data consisted of 10 537 sequences (10 503 PR and 10 476 RT) derived from resistance tests carried out at clinical centres around the United Kingdom since 1996 and held by the UK HIV Drug Resistance Database, the central repository of resistance test data in the United Kingdom. These sequences are available for use subject to the approval of the database steering committee (http://www.ctu.mrc.ac.uk/hivrdb/index.asp). Sequences were generated in 10 different laboratories using a variety of quality-controlled population sequencing protocols. All sequences were also subject to strict quality-control criteria before use within this study. Sequences that lacked key motifs, contained non-standard indels or contained more than 5% indeterminately sequenced amino acid positions were excluded from the analysis. Both PR and RT genes were contained in 99% of sequences. The majority of PR sequences (> 99%) were close to full length (±5 base pairs), RT sequences typically spanned at least amino acid positions 40–240, with the median length being 1002 base pairs (bp).

Subtype assignment

Subtypes were assigned using three distinct methods, all of which are suitable for processing batches of 100 or more sequences. The three methods used were (1) percentage identity-based genotyping as implemented in Stanford's HIVDB program (http://hivdb.stanford.edu/); (2) position-specific scoring matrix (PSSM)-based genotyping using STAR (http://www.vgb.ucl.ac.uk/star.shtml); and (3) automated neighbour-joining (NJ) phylogeny-based genotyping as implemented in the REGA HIV-1 subtyping tool, version 1.0, which is implemented as a publicly accessible web service on the BioAfrica website (http://www.bioafrica.net/subtypetool/html/), with mirror sites at Stanford (http://dbpartners.stanford.edu/RegaSubtyping/) and the Rega Institute for Medical Research (http://jose.med.kuleuven.be/subtypetool/html/index.html).

Using HIVDB, subtypes are assigned on the basis of *P*-distance as follows. Pairwise alignments are constructed using the PR and RT genes of the query sequence and consensus sequences representing each of the following lineages: group M (subtypes A–D, F–H, J and K, and CRFs 01_AE and 02_AG), group O and group N. Each gene is assigned to one of these subtypes/groups according to which of the consensus sequences it shows the highest percentage identity. If no clear 'best match' is found, an arbitrary decision between the highest matching reference subtypes/CRFs is made. Separate assignments are reported for the PR and RT genes.

In subtype assignment using STAR, an alignment is constructed containing multiple reference sequences from all nine group M subtypes, CRF01_AE and CRF02_AG, and from groups O and N. From this alignment, PSSMs are constructed that represent nucleotide frequencies at each base position for (i) each subset of reference sequences, and (ii) the global alignment including all reference sequences. The query sequence is then compared with each reference subset PSSM and the global PSSM, and a normalized *P*-distance score (*z*-score) is derived, as described in detail elsewhere [15]. An empirically determined *z*-score cut-off is used as the threshold of statistical confidence for assignment of query sequences to reference lineages [15]. Sequences that score below this threshold are left unassigned, indicating that they are potentially divergent and/or recombinant.

In REGA analysis, each query sequence is entered into two global alignments, one including two to four reference sequences from each of the nine group M subtypes (A–D, F–H, J and K), and a second that additionally includes two to three reference sequences for each CRF (01 through to 15). Each alignment is analysed for phylogenetic signal using TREE_PUZZLE software [16], and phylogenetic trees are constructed using the NJ algorithm and the Hasegawa–Kishino–Yano (HKY) evolutionary model [17] with gamma rate heterogeneity as implemented in PAUP [18]. To assess the reliability of phylogenies, 100 bootstrap replicates of the NJ algorithm are performed. Additionally, REGA analysis includes similarity plot and NJ-based 'bootscanning' of query sequences, using a 400 bp sliding window and a step size of 20 bp. A bootscan 'score' is derived from the resulting similarity plot (see de Oliveira *et al.* [19]). A decision tree incorporating the results of both the pure and the CRF analysis, and based on both the NJ bootstrap and the bootscan scores, as well as the position of the query sequence as external or internal to reference clades, is used to assign subtype [19]. Sequences that do not satisfy phylogenetic and decision tree criteria for confident grouping with known subtypes or CRFs are not assigned.

Sequences were submitted to all three genotyping protocols and perl programs (available on request) were used to describe discrepancies between the results and to generate agreement matrices. Sequences that were unassigned by at least one method, or that were inconsistently subtyped across methods, plus a randomly selected subset of consistently assigned sequences from each subtype (representing 10% of the total dataset), were manually examined for evidence of recombination using bootscan plots generated by REGA. Reference sequences used in REGA bootscan analysis are listed at (http://www.bioafrica.net/subtypetool/html/subtypealignment.html). Sequences that showed evidence of recombination were additionally examined using similarity plots generated by STAR and the BLAST-based genotyping tool, available on the National Centre for Biotechnology Information (NCBI) website (http://www.ncbi.nih. gov/projects/genotyping/formpage.cgi). NCBI genotyping used all reference sequences for subtypes A1, A2, B, C, D, F1, G, H, J and K and CRFs 01 and 02 (listed at http://www.ncbi.nih.gov/projects/genotyping/formpagex.cgi). With STAR, recombination detection is based on PSSMs for each of the groups listed above, as described by Myers *et al.* [15]. STAR sliding window analysis used a window size of 150 bp and a step size of 1 bp.

Results

Subtype assignments according to *P*-distance (HIVDB and STAR) and phylogenetic methods (REGA) are summarized in Table 1. By all methods, subtype B was most abundant, accounting for an average 71% of the total, followed by subtypes C (12.4%), A (4.7%), D (2.2%) and CRF02_AG (1.8%), with other subtypes represented at < 1%.

Subtype assignment was highly consistent (> 94%) across all three methods for subtypes B, C and H. For most other subtypes, however, there was relatively little agreement across methods. In particular, there were marked differences in assignment to subtype D, with STAR and HIVDB assigning more than six times as many sequences to this subtype than REGA. In many cases where subtype assignment was inconsistent from one method to another, disagreement involved discordant PR

Table 1. Subtype assignment of 10 537 UK resistance test sequences by three methods.

	Assignment [No. (%)]			
Subtype/CRF	HIVDB[a]	STAR	REGA	Consistently assigned[b]
A	389 (3.69)	554 (5.25)	508 (4.82)	285 [672] (42.41)
B	7524 (71.42)	7412 (70.29)	7248 (68.79)	7123 [7545] (94.41)
C	1293 (12.26)	1314 (12.46)	1305 (12.38)	1282 [1321] (97.05)
D	320 (3.03)	301 (2.85)	42 (0.40)	40 [334] (11.98)
F	21 (0.20)	32 (0.30)	25 (0.24)	17 [34] (50.00)
G	72 (0.68)	40 (0.38)	79 (0.75)	31 [98] (31.63)
H	14 (0.13)	13 (0.12)	13 (0.12)	13 [14] (92.86)
J	26 (0.25)	16 (0.15)	7 (0.07)	6 [26] (23.08)
O[c]	1 (0.01)	1 (0.01)	0 (0.00)	N/A
CRF01_AE	151 (1.43)	71 (0.67)	25 (0.24)	23 [151] (15.23)
CRF02_AG	200 (1.98)	196 (1.86)	156 (1.48)	148 [206] (71.84)
CRF03_AB[d]	N/A	N/A	0 (0.00)	N/A
CRF04_cpx[d]	N/A	N/A	0 (0.00)	N/A
CRF06_cpx[d]	N/A	N/A	24 (0.23)	N/A
CRF08_BC[d]	N/A	N/A	0 (0.00)	N/A
CRF10_CD[d]	N/A	N/A	20 (0.19)	N/A
CRF11_cpx[d]	N/A	N/A	6 (0.06)	N/A
CRF12_BF[d]	N/A	N/A	0 (0.00)	N/A
CRF13_cpx[d]	N/A	N/A	2 (0.02)	N/A
CRF14_BG[d]	N/A	N/A	2 (0.02)	N/A
Unassigned	526 (4.91)	587 (5.57)	1075 (10.21)	237 [1447] (16.38)
Assigned	10011 (95.01)	9950 (94.43)	9462 (89.80)	8968

N/A, not applicable.

[a] Sequences that were assigned discordant protease and reverse transcriptase subtypes by HIVDB were considered unassigned.

[b] The percentage consistently assigned sequences for each category was calculated as a proportion of the number of sequences assigned to that category by *any* method (shown in square brackets).

[c] Group only included in STAR and HIVDB.

[d] CRF lineages only assigned by REGA.

subtypes assigned by HIVDB. TREE-PUZZLE likelihood mapping implemented in REGA indicated that there was insufficient phylogenetic signal within PR for reliable subtype assignment using this gene alone (data not shown). This conclusion was further supported by the higher overall agreement of HIVDB RT results with other methods. For example, although agreement across methods was relatively low for subtype F (50%), it rose to 85% if HIVDB PR subtype was ignored. Amongst the remaining sequences in this category, conflicting assignments generally involved subtypes that are particularly closely related within the sequenced region (B and D), or groups for which the region analysed belonged to the same subtype (e.g. subtype A and CRF01_AE).

By far the majority of disagreements between methods, however, involved a subset of sequences that were unassigned by either REGA or STAR, or by both. A total of 587 sequences were unassigned by STAR (5.4%) and 1075 by REGA (10.21%). Excluding sequences that were < 1000 bp in length, there were 1302 sequences (12%) that were unassigned by at least one of these two methods. Sequences in this subset were examined for evidence of recombination using bootscan plots generated by REGA and sliding-window-based similarity plots generated by STAR and the NCBI genotyping tool. Visualization of bootscan and similarity plots revealed widespread evidence of recombination and a remarkable variety of unique recombinant forms (Fig. 1). As many as 70 distinct subtype compositions were present in this group of sequences, most of which were present as single representatives (Table 2), and only some of which have been identified previously [13]. Numerous diverse mosaic patterns and intersubtype breakpoints were also observed, indicating that mosaic sequences of similar subtype composition had distinct recombinant origins.

Overall, approximately 80% of sequences unassigned in STAR and/or REGA showed some evidence of recombination. Although the potential for polymerase chain reaction-generated recombinants exists, their impact on bulk population sequencing is minimal in the low-ambiguity sequence reads used in this study. However, a precise estimate of the proportion of recombinant sequences was difficult to arrive at, since the output of the various analyses did not always agree and in many cases plots were suggestive of recombination but involved relatively closely related groups. In particular, indications of recombination were frequent between the closely related B and D subtypes, and between subtypes G and A, their various subgroupings (A1, A2) and related CRFs (01_AE, 02_AG). Detection of recombination becomes increasingly subjective the more similar the genotypes involved [20], and while recombination between closely related strains almost certainly occurs, the designation of these sequences as recombinant must

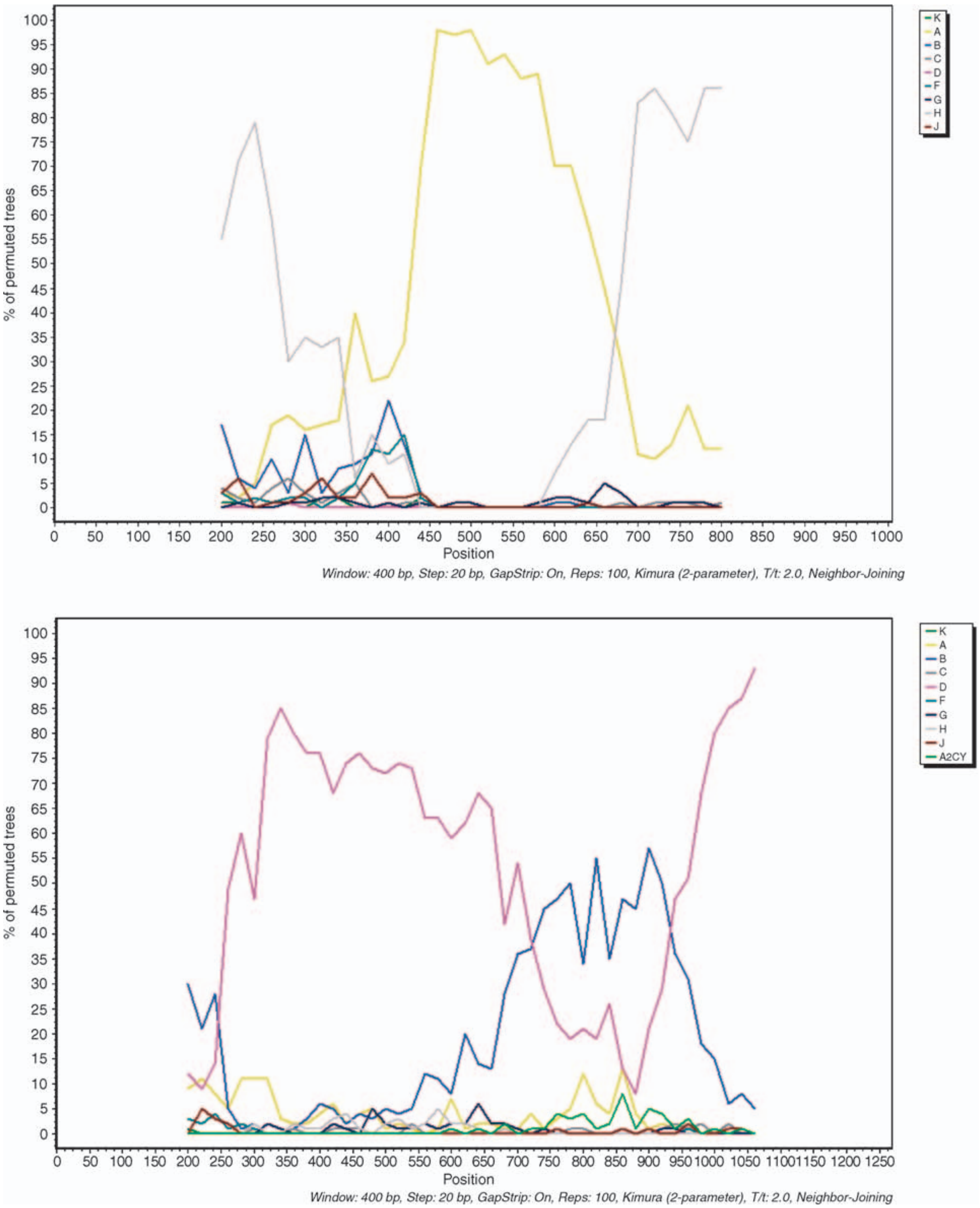


Fig. 1. Examples of sliding-window-based bootscans of reverse transcriptase sequences from unassigned and potentially recombinant viruses. Plots were generated by bootscanning implemented in the SimPlot, available from SCRoftware.

be considered tentative. Nevertheless, these sequences were consistently excluded from all confident grouping with established lineages by phylogenetic and/or statistical criteria. In particular, sequences assigned as subtype D or B by STAR and/or HIVDB frequently clustered paraphyletically to these subtypes, or basal to the entire B–D clade, in trees constructed during REGA analysis, as shown schematically in Fig. 2.

Table 2. Sumary of recombinant sequences by unique subtype composition.

Constituent subtypes	Number[a]	Constituent subtypes	Number[a]
B:D	202	G:H:J	3
A1:A2	105	C:D:H	2
A:G (Not CRF02)	87	A:F:G:J	2
A:D	77	D:cpx	2
B:H	47	D:H	2
A:J	19	D:F:G	2
B:G	19	A:D:J	2
B:F	18	B:D:H	2
B:C	17	F:H	2
D:F	17	A:H:J	2
A:B:D	16	A:B:F	1
A:C	12	C:J	1
A:F	10	B:G:J	1
C:D	10	D:F:H	1
A:B	9	B:D:cpx	1
D:J	7	A:cpx	1
B:D:J	5	A:C:F	1
A:F:G	5	B:C:D:J	1
A:G:J	5	B:D:G	1
A:H	5	A:B:D:G:H	1
A:G:H	5	B:C:D	1
A:B:C:D	5	D:F:J	1
A:K	5	A:D:H	1
A:F:G:H	5	A:B:D:G	1
A:F:H	4	A:H:K	1
B:H:J	4	B:D:K	1
B:F:H	4	A:B:C	1
D:G	4	B:cpx	1
A:D:G	4	B:K	1
G:J	4	B:J	1
B:D:F	4	A:B:G:H	1
A:F:J	3	A:C:H	1
A:G:K	3	A:F:G:H:K	1
C:G	3	B:C:F	1
A:B:H	3	B:F:K	1

[a] Numbers are approximations based on visualization of similarity plots (available on request).

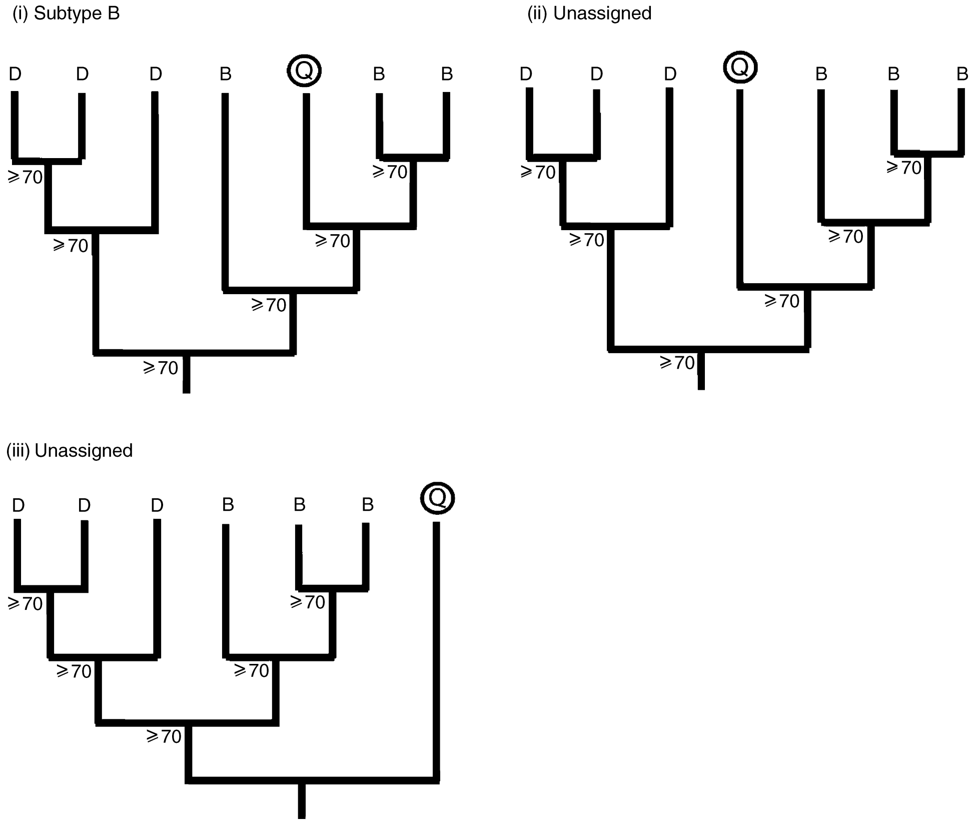
Discussion

In this study, we used three distinct methods to assign subtypes to 10 537 UK resistance test sequences, using batch-processing tools because of the large number of sequences involved. Our results revealed the very broad range of HIV-1 genetic diversity present within the United Kingdom, highlighting the need for robust and reliable protocols for describing HIV-1 genetic variation. Although the UK epidemic is dominated by subtypes B and C, most subtypes and several CRFs are represented. Furthermore, we identified a significant proportion of sequences (12.3%) that could not be confidently assigned to any previously defined subtype or CRF. Bootscan plots and sliding window-based similarity searches suggested a recombinant origin for many of these sequences, reinforcing indications from numerous sources that the role of recombinant forms in the HIV-1 pandemic is increasing [13,20,21]. Furthermore, the range of unique recombinant forms identified was very diverse; even disregarding differences in intersubtype breakpoints, we could distinguish 70 different subtype mosaic combinations (Table 2). The diversity observed amongst mosaic forms suggests that these UK infections originated from individuals infected abroad, most likely areas of sub-Saharan Africa, where coinfection with divergent strains is occurring frequently. Considering that we only examined the PR–RT region, the actual diversity of recombinant genomes circulating in the United Kingdom may be far higher and warrants confirmation by other methods.

Effective surveillance of HIV-1 genetic diversity requires tools capable of reliably distinguishing genetic variants. While the majority of sequences we analysed (85%) could be unambiguously assigned to previously described subtypes and CRFs, this largely reflected the consistency with which the most prevalent groups (subtypes B and C) were assigned by all methods. The consistency of subtype assignment across methods was low for most other subtypes. Overall, disagreements between methods fell into two categories. The first involved subtypes that are particularly closely related within the sequenced region (B and D), or groups for which the region examined belonged to the same subtype (subtype A and CRF01_AE). In these cases, a more fine-grained approach to classification is required (subsubtyping), for which a phylogenetic methodology is the most robust and recommended approach. The application of *P*-distance-based methods relates to earlier studies of HIV-1 group M diversity, which suggested a star-like phylogeny within which all lineages were equally distinct [22].

The second major category of disagreement involved sequences that were revealed by bootscanning to show at least some indication of recombinant origins. The use of phylogenetic criteria and bootscanning implemented in the REGA tool proved most reliable in discriminating these sequences, although the implementation of a statistical threshold for confident subtype assignment in STAR allowed for relatively stringent discrimination of recombinant sequences within a *P*-distance-based protocol.

To some degree, inconsistency across genotyping protocols did not reflect the shortcomings of methods but rather the logical inconsistency of attempting to assign viral strains to reference groups in the face of ongoing divergence within strains and efficient exchange of genetic material between them. This problem is most pertinent when a range of divergent subtypes cocirculate and the risk of coinfection/superinfection is high [8]. It is illustrated by results for subtypes A, G and CRF02_AG, all of which are reported to cocirculate at relatively high prevalence in areas of West and Central Africa [23]. So many distinct recombinants between these strains were identified that the impression was of several lineages merging into a single diverse group.


(i) Subtype B

(ii) Unassigned

(iii) Unassigned

Fig. 2. Schematic showing phylogenetic criteria for discriminating between subtype B and D in REGA. Sequences must cluster *within* a clade of reference sequences with bootstrap support ⩾ 70. In example (i), the query sequence Q clusters within the clade defined by subtype B reference sequences and is, therefore, assigned as subtype B. In examples (ii) and (iii), the basal position of the query sequence relative to the clades defined by subtype B and D reference sequences disqualifies it from confident assignment to either subtype.

By contrast, groupings may be relatively stable within a locality if opportunities for recombination between diverse strains are rare (e.g. in areas where a single subtype predominates). Nevertheless, groups that appear robust within one epidemiological context may be unstable outside it. This phenomenon is well illustrated by the case of the B and D subtypes. The initial introduction of HIV-1 into the United States involved a single subtype B strain that subsequently spread rapidly throughout the United States and Europe, primarily via men who have sex with men. The B subtype within this particular risk group is thus founded on a well-established genetic bottleneck and epidemiological event [24]. The closely related D subtype was defined afterwards, during a period of exploratory sampling in Africa. However, phylogenetic analysis of sequences identified during subsequent sampling suggests that the diversity of 'D-like' viruses has been underrepresented, and the broader range may encompass viruses that are intermediate between the B and D subtypes. In the dataset examined, many sequences were assigned as subtype B by one method but subtype D by another, or showed signs of recombination between these two subtypes. Given the close relationship of the B and D subtypes, these sequences may not actually be recombinant but representative of diversity within a broader category of 'B–D-like' viruses circulating in Africa. This example illustrates the importance of careful sequence analysis in revealing epidemiological information, since the sequences in question may represent infections acquired abroad from areas where diverse B–D-like viruses circulate, rather than typical subtype B infections acquired within the United Kingdom. Previous studies have obtained similar results for CRF01_AE, which may represent a divergent strain of subtype A rather than a recombinant virus [25].

Although the capacity of HIV-1 to generate variation rapidly may challenge any attempt to develop rigid classification systems, the characterization of genetic variation amongst HIV-1 isolates has a demonstrable utility in epidemiological studies, since it represents the evolutionary processes underlying epidemiological trends [26–28]. Wherever possible, assignment of sequences should be based on multiple genome regions. Nonetheless, *pol* data derived from resistance testing provides an opportunity for surveillance of HIV-1 genetic diversity on an unprecedented scale. This is of particular importance in developed countries, where surveillance has identified a diverse range of imported infections. Such countries can act as 'sentinel sites' for monitoring emerging diversity on a global scale. Furthermore, identification of the diverse infections within these countries may help to focus research efforts on strains circulating in the areas most affected by HIV-1 infection. However, meaningful epidemiological information can only be captured by reliably discriminating divergent and/or epidemiologically distinct variants. Criteria for confident assignment to established groupings should, therefore, be conservative, with the emphasis placed firmly on discriminating divergent viruses, rather than forcing them to group with most closely related reference. Sequences that fall outside established groupings can then be subjected to a more detailed analysis, facilitating the identification and epidemiological tracking of novel HIV-1 variants.

Sponsorship: The UK HIV Drug Resistance Database is funded by the Department of Health. Tulio de Oliveira is supported by the Marie Curie Fellowship (010186); Andrew Rambaut is supported by the Royal Society. This work was supported in part by FWO-Viaanderen grant G.0266.04. Note: The views expressed in the publication are those of the authors and not necessarily those of the UK Department of Health.

References

1. Gao F, Bailes E, Robertson DL, Chen Y, Rodenburg CM, Michael SF, *et al*. **Origin of HIV-1 in the chimpanzee *Pan troglodytes troglodytes*.** *Nature* 1999; **397**:436–441.
2. Santiago ML, Rodenburg CM, Kamenya S, Bibollet-Ruche F, Gao F, Bailes E, *et al*. **SIVcpz in wild chimpanzees.** *Science* 2002; **295**:465.
3. Roques P, Robertson DL, Souquiere S, Apetrei C, Nerrienet E, Barre-Sinoussi F, *et al*. **Phylogenetic characteristics of three new HIV-1 N strains and implications for the origin of group N.** *AIDS* 2004; **18**:1371–1381.
4. Gao F, Robertson DL, Morrison SG, Hui H, Craig S, Decker J, *et al*. **The heterosexual human immunodeficiency virus type 1 epidemic in Thailand is caused by an intersubtype (A/E) recombinant of African origin.** *J Virol* 1996; **70**:7013–7029.
5. Gao F, Robertson DL, Carruthers CD, Li Y, Bailes E, Kostrikis LG, *et al*. **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* 1998; **72**:10234–10241.
6. Perrin L, Kaiser L, Yerly S. **Travel and the spread of HIV-1 genetic variants.** *Lancet Infect Dis* 2003; **3**:22–27.
7. Rambaut A, Robertson DL, Pybus OG, Peeters M, Holmes EC. **Human immunodeficiency virus. Phylogeny and the origin of HIV-1.** *Nature* 2001; **410**:1047–1048.
8. Peeters M, Toure-Kane C, Nkengasong JN. **Genetic diversity of HIV in Africa: impact on diagnosis, treatment, vaccine development and trials.** *AIDS* 2003; **17**:2547–2560.
9. Vergne L, Peeters M, Mpoudi-Ngole E, Toure C, Mboup S, Mulanga-Kabeya C, *et al*. **Genetic diversity of protease and reverse transcriptase sequences in non-subtype-B human immunodeficiency virus type 1 strains: evidence of many minor drug resistance mutations in treatment-naive patients.** *J Clin Microbiol* 2000; **38**:3919–3925.
10. Kantor R, Katzenstein DA, Efron B, Carvalho AP, Wynhoven B, Cane P, *et al*. **Impact of HIV-1 subtype and antiretroviral therapy on protease and reverse transcriptase genotype: results of a global collaboration.** *PLoS Med* 2005; **2**:e112.
11. Pasquier C, Millot N, Njouom R, Sandres K, Cazabat M, Puel J, *et al*. **HIV-1 subtyping using phylogenetic analysis of pol gene sequences.** *J Virol Meth* 2001; **94**:45–54.
12. Yahi N, Fantini J, Tourres C, Tivoli N, Koch N, Tamalet C. **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* 2001; **183**:1311–1317.
13. Barlow KL, Tatt ID, Cane PA, Pillay D, Clewley JP. **Recombinant strains of HIV type 1 in the United Kingdom.** *AIDS Res Hum Retroviruses* 2001; **17**:467–474.
14. Gale CV, Myers R, Tedder RS, Williams IG, Kellam P. **Development of a novel human immunodeficiency virus type 1 subtyping tool, Subtype Analyzer (STAR): analysis of subtype distribution in London.** *AIDS Res Hum Retroviruses* 2004; **20**:457–464.
15. Myers RE, Gale CV, Harrison A, Takeuchi Y, Kellam P. **A statistical model for HIV-1 sequence classification using the subtype analyser (STAR).** *Bioinformatics* 2005; **21**:3535–3540.
16. Schmidt HA, Strimmer K, Vingron M, von Haeseler A. **TREE-PUZZLE: maximum likelihood phylogenetic analysis using quartets and parallel computing.** *Bioinformatics* 2002; **18**: 502–504.
17. Hasegawa M, Kishino H, Yano T. **Dating of the human–ape splitting by a molecular clock of mitochondrial DNA.** *J Mol Evol* 1985; **22**:160–174.
18. Swofford DL. *PAUP**. *Phylogenetic Analysis using Parsimony (*and Other Methods)*, version 4. Sunderland, MA: Sinauer Associates; 1998.
19. de Oliveira T, Deforche K, Cassol S, Salminen M, Paraskevis D, Seebregts C, *et al*. **An automated genotyping system for analysis of HIV-1 and other microbial sequences.** *Bioinformatics* 2005; **21**:3797–3800.
20. Posada D. **Evaluation of methods for detecting recombination from DNA sequences: empirical data.** *Mol Biol Evol* 2002; **19**:708–717.
21. Gomez-Carrillo M, Quarleri JF, Rubio AE, Carobene MG, Dilernia D, Carr JK, *et al*. **Drug resistance testing provides evidence of the globalization of HIV type 1: a new circulating recombinant form.** *AIDS Res Hum Retroviruses* 2004; **20**:885–888.
22. Burr T, Hyman JM, Myers G. **The origin of acquired immune deficiency syndrome: Darwinian or Lamarckian?** *Philos Trans R Soc Lond B Biol Sci* 2001; **356**:877–887.
23. Osmanov S, Pattou C, Walker N, Schwardlander B, Esparza J. **Estimated global distribution and regional spread of HIV-1 genetic subtypes in the year 2000.** *J Acquir Immune Defic Syndr* 2002; **29**:184–190.
24. Robbins KE, Lemey P, Pybus OG, Jaffe HW, Youngpairoj AS, Brown TM, *et al*. **US human immunodeficiency virus type 1 epidemic: date of origin, population history, and characterization of early strains.** *J Virol* 2003; **77**:6359–6366.
25. Anderson JP, Rodrigo AG, Learn GH, Madan A, Delahunty C, Coon M, *et al*. **Testing the hypothesis of a recombinant origin of human immunodeficiency virus type 1 subtype E.** *J Virol* 2000; **74**:10752–10765.
26. Malim MH, Emerman M. **HIV-1 sequence variation: drift, shift, and attenuation.** *Cell* 2001; **104**:469–472.
27. Hue S, Clewley JP, Cane PA, Pillay D. **HIV-1 pol gene variation is sufficient for reconstruction of transmissions in the era of antiretroviral therapy.** *AIDS* 2004; **18**:719–728.
28. Hue S, Pillay D, Clewley JP, Pybus OG. **Genetic analysis reveals the complex structure of HIV-1 transmission within defined risk groups.** *Proc Natl Acad Sci USA* 2005; **102**:4425–4429.

Appendix

UK Collaborative Group on HIV Drug Resistance Steering Committee: Sheila Burns, City Hospital, Edinburgh; Sheila Cameron, Gartnavel General Hospital, Glasgow; Patricia Cane, Health Protection Agency; Ian Chrystie, St Thomas' Hospital, London; Duncan Churchill, Brighton and Sussex

University Hospitals NHS Trust; Valerie Delpech, Health Protection Agency, London; David Dunn, Esther Fearnhill, Kholoud Porter, MRC Clinical Trials Unit, London; Philippa Easterbrook, Mark Zuckerman, King's College Hospital, London; Anna Maria Geretti, Royal Free NHS Trust, London; Robert Gifford, Paul Kellam, Andrew Phillips, Deenan Pillay, Caroline Sabin, Royal Free and University College Medical School, London; David Goldberg, Scottish Centre For Infection and Environmental Health; Mark Gompels, Southmead Hospital, Bristol; Tony Hale, PHLS, Leeds; Steve Kaye, St Marys Hospital, London; Andrew Leigh-Brown, University of Edinburgh; 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; Peter Tilston, Manchester Royal Infirmary; Ian Williams, Mortimer Market Centre.