AIDS RESEARCH AND HUMAN RETROVIRUSES
Volume 17, Number 2, 2001, pp. 185–189
Mary Ann Liebert, Inc.

Sequence Note

Reanalysis of Full-Length HIV Type 1 Group M Subtype K and Sub-Subtype F2 with an MS-DOS Bootscanning Program

GERT VAN DER AUWERA,[1] WOUTER JANSSENS,[1,2] LEO HEYNDRICKX,[1] and GUIDO VAN DER GROEN[1]

ABSTRACT

Five new complete HIV-1 group M genome sequences have been published (Triques *et al.*, AIDS Res Hum Retroviruses 2000;16:139–151). One of these clustered consistently with subtype F sequences, while two others were identified as representatives of a subcluster within the subtype F clade, called F2, and the two remaining sequences were described as a new subtype K. We reanalyzed these sequences by means of bootscanning and phylogeny, using a newly developed MS-DOS bootscanning program. Although our analysis does not contradict the existence of the new subtype K, it also indicates that in some regions the F2 sequences do not cluster with the F1 clade. This suggests that some fragments in the F2 sequences have an uncertain origin, and care should be taken when F2 sequences are used in analyses.

Until now, nine nonrecombinant subtypes have been described within the HIV type 1 group M viruses, that is, A, B, C, D, F, G, H, J, and K. Of each subtype, several complete genome sequences have been determined and analyzed, and a selection was made by the Los Alamos National Laboratory HIV sequence database to serve as a reference framework for investigating the possible recombinant nature of other sequences.[1] Apart from these "pure" subtypes, six epidemiologically significant intersubtype recombinants, called circulating recombinant forms (CRFs), have been described: CRF01_AE (CM240), CRF02_AG (IbNG), CRF03_AB (KAL153), CRF04_cpx (94CY032),[1,2] CRF05_DF (VI1310), and CRF06_cpx (BFP90) (Los Alamos National Laboratory HIV sequence database, http://hiv-web.lanl.gov). Five new full-length HIV-1 sequences have been reported,[3] of which one (96FR-MP411) was shown to belong to subtype F. Two others (95CM-MP255 and 95CM-MP257) were found to be closely related to subtype F, and were grouped in a subcluster called F2, while the original subtype F isolates were renamed as F1 (this new nomenclature will be followed throughout this article). The two remaining sequences (97ZR-EQTB11 and 96CM-MP535) were classified as the new subtype K. The new subtype K and sub-subtype F2 were also included in the HIV-1 nomenclature proposal.[2] We reanalyzed these five new sequences with an in-house developed MS-DOS-based bootscanning program.

Bootscanning[4] is a phylogenetic analysis tool that allows the assessment of the bootstrap support for a given cluster throughout the entire genome sequence. To calculate these bootstrap values, the alignment is divided into overlapping windows, each containing a fixed number of consecutive alignment positions (window size). The number of nucleotides between the start of two successive windows is referred to as the step size. A bootstrap tree is made on the basis of the alignment positions within each window, and plotting the bootstrap support for a predefined cluster in all windows results in the bootscan plot for that cluster. Two software packages are currently available for bootscanning: Bootscan[4] and Simplot.[5] The first package runs on UNIX machines, while Simplot is designed for a Microsoft Windows environment. However, Simplot suffers from the disadvantage that it only allows testing of the relationship of any given sequence relative to the terminal nodes in the tree (i.e., the sequences included), and does not allow examination of internal nodes. This means that for subtyping purposes it only makes sense to include one representative for each subtype in the trees, or to use consensus sequences, as was done by Triques *et al.*[3] We developed a new bootscanning program for MS-DOS

[1]Department of Microbiology, Institute of Tropical Medicine, B2000 Antwerp, Belgium.
[2]Flanders Interuniversity Institute for Biotechnology (VIB), B9052 Ghent, Belgium.

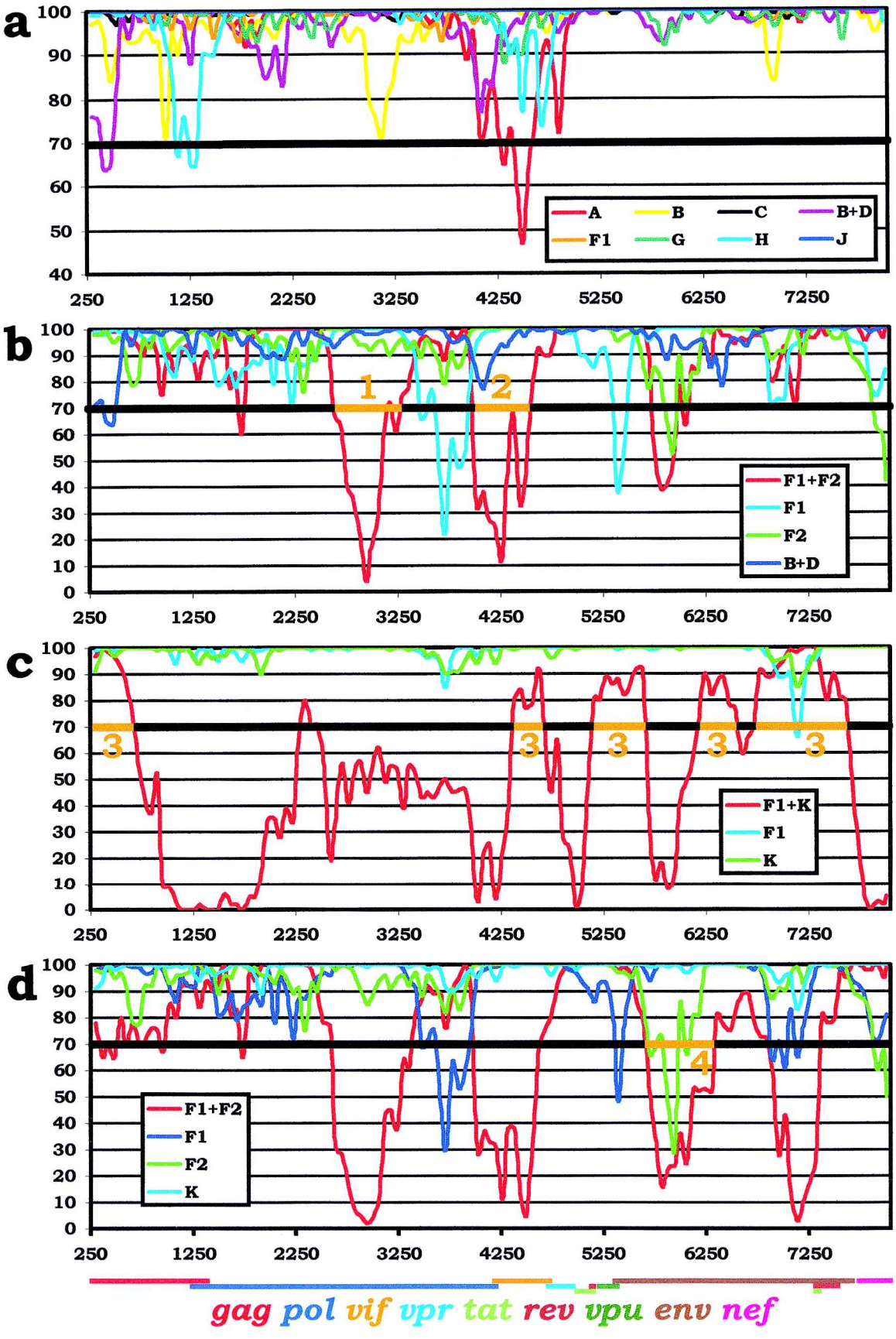
that allows investigation of the bootstrap support for any given cluster (i.e., also internal nodes), as is possible with the Bootscan package for UNIX. The principle of the program is identical to that of UNIX bootscanning, and it also uses the PHYLIP software package[6] (version 3.57c) for the construction of bootstrap trees. The MS-DOS software as used for the analyses in this article can be obtained from the authors on request. It requires installation of the PHYLIP MS-DOS programs prior to use, and uses a sequential (noninterleaved) PHYLIP alignment as input. The program is not as user-friendly as SimPlot, but at least it will allow PC users to perform bootscanning very much in the same way as with the UNIX software.

To reanalyze the new sequences, they were aligned with the complete genome sequences of the representatives of each subtype and CRF as listed in the Los Alamos National Laboratory HIV sequence database compendium.[1] Regions in which the alignment was unreliable because of excessive length or sequence variation were omitted from the analysis, as were those positions that contained gaps in most sequences. This resulted in an alignment of 8310 nucleotide positions. CRF01_AE is the only intersubtype recombinant of which sequences were included in the analyses, because these sequences contain parts of subtype E, which would otherwise be lacking as no full–length subtype E sequences have been identified yet. For the bootscan plots, distances were calculated with the Kimura two-parameter model, using a transition-to-transversion ratio of 2. The neighbor-joining algorithm was used for tree building, and 100 bootstrap samples were analyzed for each window. These parameters are commonly used for analyzing HIV sequences. The step size for bootscanning was 10% of the window size, resulting in a 90% overlap between two successive windows. After bootscan analysis, some regions were selected for the construction of a bootstrapped phylogenetic tree, using the software package TREECON for Windows.[7] These bootstrap trees were based on sequence dissimilarities instead of estimated distances, because some sequences may have breakpoints in the region used, which makes the current models for converting dissimilarities into distances inaccurate. Again the neighbor-joining algorithm was used, but this time 2000 bootstrap replicates were analyzed.

Before starting the analysis, an appropriate window size had to be chosen for bootscanning. If the windows are too big, small fragments from another subtype may be missed, but if they are too small they may not contain adequate phylogenetic information. Another problem is that the ideal window size can differ along the genome, the more variable or highly conserved regions requiring a bigger window. To determine the window size for our analysis, we constructed bootscan trees including only the reference sequences of the nonrecombinant subtypes, as well as the new F1 sequence 96FR-MP411, using different window sizes (multiples of 100 from 300 to 700). Figure 1a shows the resulting bootscan plots for each of the subtypes A, B, C, B+D, F1, G, H, and J when using a window size of 600 nucleotides. This was found to be the optimal window size because each of the nonrecombinant HIV-1 group M subtypes is supported by bootstrap values of more than 70% throughout the entire genome, except for a small region of subtypes A, B+D, and H. Subtype D was evaluated in combination with subtype B (B+D cluster), as it is well known that subtypes B and D are quite closely related, and it has been suggested that subytpe B can in fact be considered as a sub-subtype within subtype D.[2] The disadvantage of using a window size of 600 is that possible recombinant fragments that are considerably smaller than 600 nucleotides may remain undetected, but on the other hand it assures that relationships can be deducted with a high degree of confidence. With a window size of 700, only the bootstrap support for subtype A dropped below 70% in two small regions, while a window of 500 caused bootstrap values to drop below 70% for subtypes A, B, B+D, and H in some regions (data not shown). From Fig. 1a, it is clear that sequence 96FR-MP411 indeed belongs to subtype F1 over its entire length, as it can be seen that the monophyly of F1 is highly supported (at least 93%) over the entire genome range.

A second bootscanning analysis was performed with the same isolates, but also including CRF01_AE and the two F2 isolates. Figure 1b shows the resulting bootstrap support for the clades F1, F2, subtype F (F1+F2), and the B+D cluster. From these bootscan plots, it is clear that in two substantial regions, positions 2640–3240 (region 1, corresponding to nucleotides 2700–3300 in 96FR-MP411, accession number AJ249238) and 4020–4500 (region 2, nucleotides 4080–4560 in 96FR-MP411), the clustering of F1 with F2 is not supported at the 70% level, as is also the case in a smaller region around position 5850. In general, F1 and F2 both form well-supported monophyletic groups, although in some regions the monophyly of F1 is supported by less than 70% of the bootstrap trees. In these regions, F1+F2 is well supported, which indicates that the resemblance of F1 and F2 is high enough to disturb the F1 group. Although the F1+F2 situation has been compared with the B+D situation,[3] the latter group clearly is monophyletic in each region (Fig. 1a and b), while F1+F2 is not. Regions 1 and 2 from Fig. 1b were used for constructing Fig. 2a and b, respectively, which contain the same isolates as used for bootscanning. Also, from these trees it is apparent that the monophyly of F1+F2 is not supported, although all other subtypes are well supported, indicating that sufficient phylogenetic information is present in this region. Moreover, F2 does not cluster significantly with any known subtype. These results support the possibility that the genome of F2 isolates is in fact recombinant, consisting for the most part of subtype F, interspersed with fragments of unknown origin. On the other hand, it was shown that the distance between F1 and F2 is always smaller than that of F2 relative to any other subtype,[3] although it was found to be larger than all other intrasubtype distances.[2] Hence, these findings too do not contradict the possible recombinant origin of F2. Moreover, according to the new HIV-1 nomenclature proposal,[2] all dis-

FIG. 1. Bootscanning plots of HIV-1 group M clades. The nonrecombinant subtype references as listed in the Los Alamos National Laboratory HIV sequence database compendium[1] and the F1 sequence 96FR-MP411 were included in the bootstrap trees (**a–d**), supplemented by CRF01_AE (**b–d**), the two sub-subtype F2 sequences (**b,d**), and/or the two subtype K sequences (**c,d**). See text for details concerning the bootscan procedure. *Bottom*: The positions of the various genes on the HIV genome are indicated in different colors (please note that the colors used to indicate the different genes are not linked to the colors used in the bootscan plots).


a
100
90
80
70
60
50
40
250
1250
2250
3250
4250
5250
6250
7250
A
B
C
B+D
F1
G
H
J
b
1
2
F1+F2
F1
F2
B+D
c
3
3
3
3
3
F1+K
F1
K
d
4
F1+F2
F1
F2
K
gag pol vif vpr tat rev vpu env nef

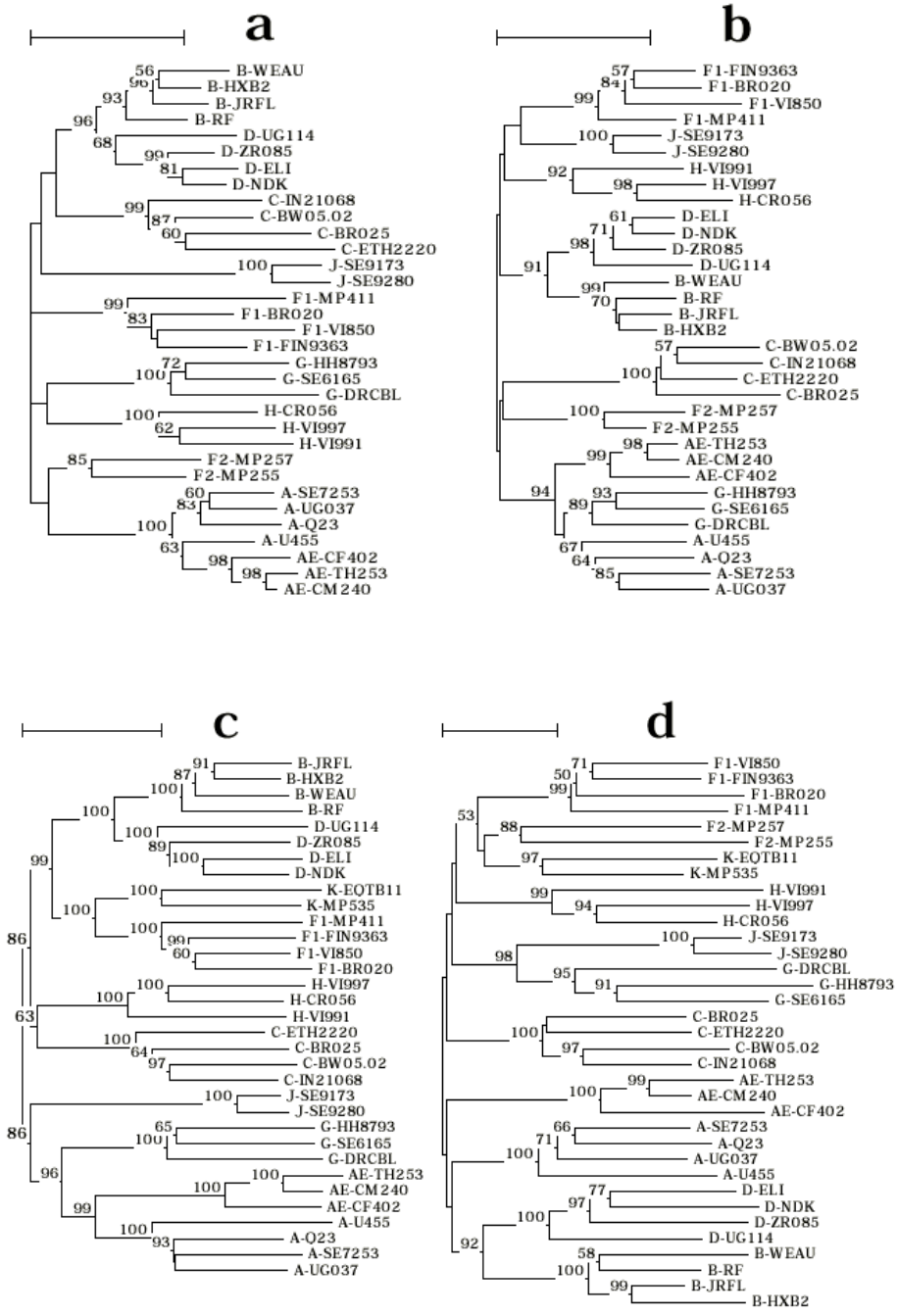


FIG. 2. Neighbor-joining trees based on the regions indicated in Fig. 1. Trees **a**–**d** correspond to regions 1–4, respectively. The trees are based on dissimilarities between the sequences rather than on corrected distances. The distance between two isolates is obtained by summing the lengths of the connecting horizontal branches, using the 5% distance scale at the top of each tree. Bootstrap values from a 2000-replicate analysis are shown in percentages at the internodes if they exceed the 50% level. The subtype or sub-subtype of each isolate precedes the isolate name (AE indicates CRF01_AE).

tance analyses should be backed up by phylogeny, while the bootscan analysis that was published previously[3] failed to check for the clustering of F1+F2, but only checked for the monophyly of F2.

A third bootscanning was performed to evaluate the new subtype K (Fig. 1c). The sequences included for calculating this bootscan plot were the same as used for Fig. 1b, but the two F2 sequences were replaced by the two K sequences. From this bootscan plot it is clear that both subtypes F1 and K are monophyletic in each part of the genome. Also, the bootscan support for the clustering of subtypes K and F1 is shown, as F1 is the only clade that clusters significantly (bootstrap value >70%) with K in some regions, which may compromise the position of the K sequences as a new subtype. An alignment was constructed that includes all the regions designated as 3 in Fig. 1c (positions 300–660, 4380–4680, 5160–5640, 6180–6540, 6720–7620 in the alignment, which correspond to nucleotides 304–684, 4440–4740, 5220–5733, 6374–6797, 6981–7881, in 96FR-MP411). This alignment was used to build the tree in Fig. 2c, which includes the same sequences as used for bootscanning. In this tree, the clustering of F1 with K is supported by a bootstrap value of 100%. However, this in itself is not an argument against subtype K, given that the observed distances between the two groups are sufficiently large to justify the subtype status for the K sequences. It has been suggested that the K clade does not deserve the subtype status because of its relationship to subtype F,[2] which would not be in disagreement with our analysis.

Finally, a bootscan plot was constructed from trees that include again the same sequences as used for Fig. 1b and c, but this time the F1, F2, and K sequences were all used together. Figure 1d shows the resulting bootscan plot for the monophyly of F1, F2, F1+F2, and K. The conclusions drawn from Fig. 1b and c with regard to the monophyly of F1, F2, and K still stand, but the monophyly of F1+F2 is more problematic. Indeed, with the inclusion of the subtype K sequences, there are four instead of two significant regions where the bootstrap support for the clustering of F1 and F2 falls well below the 70% level. This further compromises the designation of the F2 cluster as being part of subtype F, as one of the current criteria for subtype designation is that the sequences must be monophyletic over the entire range of the genome.[2] As an example, a bootstrapped tree based on region 4 (position 5700–6300, corresponding to nucleotides 5793–6506 of 96FR-MP411) is shown in Fig. 2d. Subtype K and sub-subtypes F1 and F2 seem to be somewhat related in this tree, although these relationships are not bootstrap supported.

In conclusion, we find no arguments against the designation of the new subtype K, although our results show that some parts are related to subtype F. However, much care should be taken when including F1 and F2 sequences in phylogenetic analyses. On the one hand, F2 cannot be treated completely separately from F1, as in some regions they are so close that the F2 sequences disturb the monophyly of sub-subtype F1. On the other hand, F1+F2 cannot be treated as a monophyletic group either, because in some regions it clearly is not. Especially when analyzing recombinant sequences that contain regions related to sub-subtypes F1 or F2, these facts must be taken into account. In retrospect, it would probably have been better to classify F2 as recombinant, as may also be the case for subtype K. In view of our increasing knowledge and understanding of HIV diversity, establishing an HIV nomenclature will remain a difficult and challenging task. Because of better analysis tools and an expanding sequence database, it will remain subject to revision. The priority is to keep things as simple as possible, and therefore designations that have already been established and used in the literature should not be changed lightly. For this reason we see no harm in retaining the present classification system,[2] provided that one bears in mind that sub-subtype F2, and possibly also subtype K, have a dubious subtype status.

ACKNOWLEDGMENTS

This work was supported by the Fund of Scientific Research (FWO), Flanders, Belgium (grant number G.0134.97); the Flanders Interuniversity Institute for Biotechnology (VIB), Ghent, Belgium; and the Human Science Foundation, Tokyo, Japan. Gert Van der Auwera is a research fellow of the Fund of Scientific Research (FWO), Flanders, Belgium.

REFERENCES

1. Carr JK, Foley BT, Leitner T, Salminen M, Korber B, and McCutchan F: Reference sequences representing the principal genetic diversity of HIV-1 in the pandemic. In: *Human Retroviruses and AIDS 1998—A Compilation and Analysis of Nucleic Acid and Amino Acid Sequences* (Korber B, Kuiken CL, Foley B, Hahn B, McCutchan F, Mellors JW, and Sodroski J, eds.). Theoretical Biology and Biophysics Group, Los Alamos National Laboratory, Los Alamos, New Mexico, 1998, pp. III10–III19.
2. Robertson DL, Anderson JP, Bradac JA, *et al.*: HIV-1 nomenclature proposal—a reference guide to HIV-1 classification. In: *Human Retroviruses and AIDS 1999—A Compilation and Analysis of Nucleic Acid and Amino Acid Sequences* (Kuiken CL, Foley B, Hahn B, Korber B, McCutchan F, Marx PA, Mellors JW, Mullins JI, Sodroski J, and Wolinsky S, eds.). Theoretical Biology and Biophysics Group, Los Alamos National Laboratory, Los Alamos, New Mexico, 1999, pp. 492–505.
3. Triques K, Bourgeois A, Vidal N, Mpoudi-Ngole E, Mulanga-Kabeya C, Nzilambi N, Torimiro N, Saman E, Delaporte E, and Peeters M: Near-full-length genome sequencing of divergent African HIV type 1 subtype F viruses leads to the identification of a new HIV type 1 subtype designated K. AIDS Res Hum Retroviruses 2000;16:139–151.
4. Salminen MO, Carr JK, Burke DS, and McCutchan FE: Identification of breakpoints in intergenotypic recombinants of HIV type 1 by bootscanning. AIDS Res Hum Retroviruses 1995;11:1423–1425.
5. Lole KS, Bollinger RC, Paranjape RS, Gadkari D, Kulkarni SS, Novak NG, Ingersoll R, Sheppard HW, and Ray SC: Full-length human immunodeficiency virus type 1 genomes from subtype C-infected seroconverters in India, with evidence of intersubtype recombination. J Virol 1999;73:152–160.
6. Felsenstein J: PHYLIP—phylogeny inference package (version 3.2). Cladistics 1989;5:164–166.
7. Van de Peer Y and De Wachter R: TREECON for Windows—a software package for the construction and drawing of evolutionary trees for the Microsoft Windows environment. Comput Appl Biosci 1994;10:569–570.

Address reprint requests to:
Gert Van der Auwera
Institute of Tropical Medicine
Department of Microbiology
Nationalestraat 155
B2000 Antwerp, Belgium

E-mail: gvdauwera@itg.be