Recombination confounds the early evolutionary history of human immunodeficiency virus type 1: Subtype G is a circulating recombinant form

JOURNAL OF VIROLOGY, Aug. 2007, p. 8543–8551 Vol. 81, No. 16
0022-538X/07/$08.00+0 doi:10.1128/JVI.00463-07

Recombination Confounds the Early Evolutionary History of Human Immunodeficiency Virus Type 1: Subtype G Is a Circulating Recombinant Form▽

Ana B. Abecasis,[1,2] Philippe Lemey,[1,3] Nicole Vidal,[4] Túlio de Oliveira,[3]† Martine Peeters,[4] Ricardo Camacho,[2] Beth Shapiro,[3] Andrew Rambaut,[3]‡ and Anne-Mieke Vandamme[1]*

Laboratory for Clinical and Epidemiological Virology, Rega Institute for Medical Research, Katholieke Universiteit Leuven, Minderbroedersstraat 10, B-3000 Leuven, Belgium[1]; Laboratório de Virologia, Serviço de Imunohemoterapia, Hospital de Egas Moniz, Rua da Junqueira, 126, 1349-019 Lisbon, Portugal[2]; Evolutionary Biology Group, Department of Zoology, University of Oxford., South Parks Road, Oxford OX1 3PS, United Kingdom[3]; and Laboratory Retrovirus, IRD, IRD-UMR 145, 911, Av. Agropolis-BP 64501, 34394 Montpellier Cedex 5, France[4]

Received 5 March 2007/Accepted 29 May 2007

Human immunodeficiency virus type 1 (HIV-1) is classified in nine subtypes (A to D, F, G, H, J, and K), a number of subsubtypes, and several circulating recombinant forms (CRFs). Due to the high level of genetic diversity within HIV-1 and to its worldwide distribution, this classification system is widely used in fields as diverse as vaccine development, evolution, epidemiology, viral fitness, and drug resistance. Here, we demonstrate how the high recombination rates of HIV-1 may confound the study of its evolutionary history and classification. Our data show that subtype G, currently classified as a pure subtype, has in fact a recombinant history, having evolved following recombination between subtypes A and J and a putative subtype G parent. In addition, we find no evidence for recombination within one of the lineages currently classified as a CRF, CRF02_AG. Our analysis indicates that CRF02_AG was the parent of the recombinant subtype G, rather than the two having the opposite evolutionary relationship, as is currently proposed. Our results imply that the current classification of HIV-1 subtypes and CRFs is an artifact of sampling history, rather than reflecting the evolutionary history of the virus. We suggest a reanalysis of all pure subtypes and CRFs in order to better understand how high rates of recombination have influenced HIV-1 evolutionary history.

Human immunodeficiency virus type 1 (HIV-1) is a retrovirus characterized by high levels of mutation (ranging between 5×10^{-6} and 9×10^{-5} mutations per nucleotide per cycle of virus replication) (29) and recombination rates (reported as 42.4%/replication cycle, with markers 1 kb apart) (25), making it an ideal model organism for investigating long-term evolutionary processes.

HIV-1 exhibits very high genetic diversity and is classified in three major groups (M, N, and O). Group M, which is responsible for the global HIV-1 pandemic, is further classified into subtypes A, B, C, D, F, G, H, J, and K, each representing distinctive lineages within group M (26). These subtypes have diversified independently following the initial transmission of the HIV-1 group M progenitor to humans. Chance exportation of particular lineages from the initial epidemic region, followed by subsequent local epidemics in previously uninfected regions, likely led to the current global distribution of HIV-1 subtypes (23, 24). Subsubtypes (e.g., A1, A2, A3, and A4 and F1 and F2) are distinctive lineages that are not genetically distant enough to justify designation as a new subtype, and circulating recombinant forms (CRFs) are intersubtype recombinant viruses with a significant epidemic spread (26). According to the Los Alamos National Laboratory database, 34 CRFs are currently characterized, eight of which are mosaic genomes containing gene regions of more than two subtypes (http://www.hiv.lanl.gov/content/index). The database also lists a large number of unique recombinant forms, generated after coinfection or superinfection in a patient with two different subtypes.

The evolutionary history of subtype G. Substantial subtype G prevalence was first noticed and still remains the highest in Central and West African countries (11% and 35% of all infections, respectively) (10), where the highest sequence divergence within subtype G is also reported. A few studies describe the molecular epidemiology of subtype G in Cameroon (5, 6, 31); Nigeria (17, 22); Democratic Republic of Congo (DRC) (35); Senegal, Cameroon, and Gabon (18); and Republic of Congo, where it is more prevalent (20). That the highest genetic diversity of subtype G occurs in DRC (34) is consistent with the currently accepted theory of the origin of HIV-1, which claims that the epidemic emerged in Central West Africa (12, 35).

Since the different subtypes are assumed to have evolved independently, different genome regions are expected to have the same evolutionary history. However, this is apparently not the case for subtype G. The original study describing subtype G reported that some genomic regions within the subtype had greater similarity with subtype A than would be expected for a

* Corresponding author. Mailing address: Katholieke Universiteit Leuven, Laboratory for Clinical and Epidemiological Virology, AIDS Reference Laboratory, Rega Institute and University Hospitals, Minderbroedersstraat 10, B-3000 Leuven, Belgium. Phone: 32-16-332160. Fax: 32-16-332131. E-mail: annemie.vandamme@uz.kuleuven.be.

† Present address: HRC Pathogen Bioinformatics Unit, South African National Bioinformatics Institute, University of the Western Cape, Cape Town, South Africa.

‡ Present address: Institute of Evolutionary Biology, University of Edinburgh, Edinburgh, United Kingdom.

▽ Published ahead of print on 6 June 2007.

pure subtype (2). Several subsequent analyses discussed a putative recombinant origin of subtype G (26); however, no previously reported pure subtype could be assigned to those fragments that did not show genetic similarity to subtype A. Hence, it was decided to keep subtype G as a pure subtype in the classification system (7, 8, 26).

The evolutionary history of the purported CRF CRF02_AG. One of the most prevalent HIV-1 lineages worldwide is CRF02_AG. This proposed CRF makes up 5% of infections, compared to 50% for subtype C (the most prevalent), 12.3% for subtype A, and 10.4% for subtype B. It is most prevalent in West and Central Africa (27.87 and 3.98% of infections, respectively) (10). CRF02_AG was first reported in Nigeria (11) and has currently the highest sequence divergence in several West African countries, but not in DRC and Republic of Congo, where it is almost absent (1, 18, 20, 34, 35). The complete CRF02_AG genome was sequenced for the first time in 1998 (2).

One of the interesting features of the CRF02_AG epidemiology is that it was already considerably prevalent early in the pandemic, which is not expected for a recombinant strain. By 1999, it was already more prevalent than its supposed parental subtype G lineage in its putative region of origin, West Central Africa (1, 2).

The evolutionary history of subtype J. Subtype J was discovered in 1995 and originated most probably in DRC. The two published complete genomes for subtype J were both isolated in Sweden from individuals from DRC (13, 14). Few subtype J sequences are available, and these are mostly from DRC, Cameroon, and Senegal (Los Alamos National Laboratory database [http://www.hiv.lanl.gov/content/index]). However, fragments of subtype J are present in many mosaic recombinant forms originating from West Africa (CRF06_cpx) and Central West Africa (CRF11_cpx and CRF13_cpx), suggesting that this subtype, either in a pure or in a recombinant form, is probably prevalent across Central and West Africa.

In this study, we have tested the validity of the current classification of subtype G and CRF02_AG. Given the possible recombinant history of subtype G and the geographical distributions of both subtype G and CRF02_AG, we hypothesize that subtype G is not a pure subtype but is instead a CRF, with the proposed CRF CRF02_AG as a parental lineage. To test this hypothesis, we performed an extensive analysis of the group M phylogeny, using full-genome sequences from all the currently identified pure subtypes and CRFs. Our results have important implications for understanding the geographical epidemiological history of HIV-1 and raise questions about the current classification of HIV-1 subtypes.

MATERIALS AND METHODS

Sequence data. Full-genome alignments of 137 strains of HIV-1 were downloaded from the Los Alamos database and aligned against other subtypes and CRFs using the profile alignment mode of ClustalX (32). The resulting alignment was then manually edited with Se-Al v2.0 (http://evolve.zoo.ox.ac.uk/). The subtype assigned to these sequences in the Los Alamos database was confirmed using the REGA subtyping tool (3). The alignments are available upon request.

Sequencing of new subtype J full-length genome. A previously unpublished subtype J full-length genome (KTB147) was included in the alignment described above. Partial gag and gp160 sequences have been previously reported (35), and full-length sequence was obtained by overlapping of PCR fragments from different genome regions, as previously described (18, 19, 33, 35).

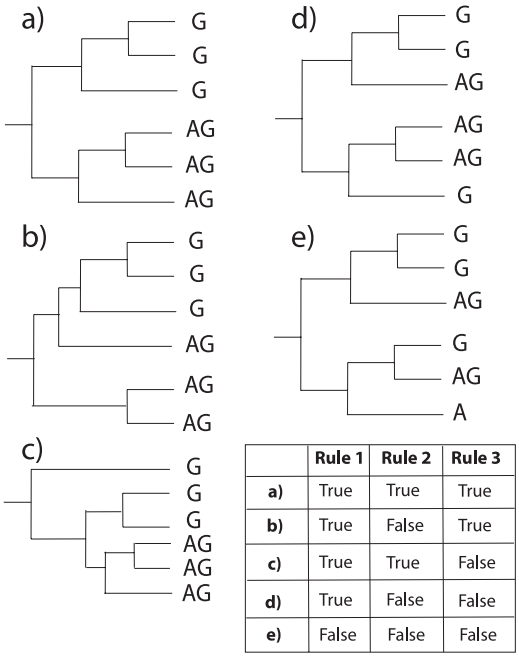
Recombination analysis. To explore putative recombination patterns in the sequences, we used a sliding window approach, which computes a statistic or measure for a successive set of overlapping subregions (windows) of the alignment (28). This allows identification of the putative intersubtype recombination breakpoints of the query sequence by the graphical detection of a change in the phylogenetic signal, which is characterized by a sudden decrease in support for the clustering with a certain subtype and the simultaneous increase of support for the clustering with another subtype. The query sequence is analyzed against the previously described pure subtype reference set. The software Simplot v3.5.1 (15) was initially used to perform similarity and bootscanning analyses of a query sequence against a set of other sequences. The similarity plot measures the similarity/dissimilarity of the query sequence to a set of reference sequences. In the bootscanning plot, the phylogenetic relationship between the query sequence and the reference set is calculated using bootstrap resampling and the bootstrap values are plotted along the genome. In the preliminary analysis, a window size of 500 bp and step size of 100 bp were used, while in the final plots the window size and step size were 350 bp and 50 bp, respectively. This procedure made it possible to maximize the detection of recombination events while maintaining a good phylogenetic signal and, by comparison of the plots, to ensure that the two window sizes generated similar results. In addition, a more rigorous sliding window analysis using a Bayesian phylogenetic approach was performed, to increase our confidence in the inferred recombination breakpoints. Sliding windows of 500 bp, moving in 100-bp steps, were generated using the software SlidingBayes0.94 (21). As this analysis is extremely time-consuming, only one strain of each subtype was included. This involves a phylogenetic analysis using Bayesian inference as implemented in MrBayes v3.1.2 (27). In this analysis, two Monte Carlo Markov chains (MCMCs) are run simultaneously for the number of generations needed for a stationary distribution to be maintained long enough after convergence. Typically, the number of generations was 4×10^6 to 5×10^6, with the initial 10% of these generations discarded as burn-in. To analyze convergence and stability, we used the software Tracer1.3 (A. Rambaut and A. J. Drummond, http://evolve.zoo.ox.ac.uk/), which allowed us to visualize the posterior distribution for each parameter and provided an estimate of the effective sample size, a measure of the number of "effectively independent" samples in each run, as defined by Drummond et al. (4). We also analyzed the convergence using diagnostic measures implemented in MrBayes, in particular the potential scale reduction factor, as defined by Gelman and Rubin (9). We considered a run to have converged when the effective sample size of all parameters was above 100 and when the potential scale reduction factor was approximately 1. We also ensured that the log likelihood reached stability after the burn-in period, which was discarded from the sample. Posterior probabilities for the clustering of the query with the reference strains were plotted along the genome.

The putative recombination breakpoints suggested by the similarity and bootscanning plots and by the posterior probability plots were similar but not identical. Therefore, we used the informative site analysis as implemented in Simplot v3.5.1 to get a more precise breakpoint estimate. Neighbor-joining (NJ) and maximum-likelihood (ML) trees were reconstructed for each of the fragments defined by the recombination breakpoints. The parameters of the evolutionary model were estimated from the data. Finally, 1,000 bootstrap replicates and the zero branch length test were performed to assess the robustness of the clustering. The software PAUP 4.0b10 (30) was used to produce the NJ and ML trees. Phylogenetic analysis was also performed using Bayesian inference, as described above.

Monophyly rules for subtype G and CRF02_AG to discriminate parent from recombinant. A group of sequences is called monophyletic if they form a cluster composed of all descendants from an inferred common ancestor (parent). If a group of sequences do not include all descendants of their inferred most recent common ancestor (MRCA), then those sequences cluster as paraphyletic; they cannot be grouped in a single cluster. In the context of HIV-1 molecular epidemiology, we can expect that the parental subtype will have an MRCA more ancient than that of the CRF originating from it. Therefore, we can expect that the parent pure subtype will be paraphyletic with respect to the CRF, which will cluster monophyletically within the pure subtype cluster.

Within genetic regions where CRF02_AG is currently considered to be of subtype G origin, the parent can be discriminated from the recombinant by investigating their sequence divergence, using the reasoning explained in the previous paragraph. For this purpose, the last 10,000 trees of the posterior distribution of trees generated by each MCMC run, summarizing the phylogenetic uncertainty, were midpoint rooted and the support for all of the following three "monophyly rules," concerning the CRF02_AG/G cluster, was investigated (Fig. 1): (i) monophyly of CRF02_AG plus G, (ii) monophyly of CRF02_AG separately, and (iii) monophyly of subtype G separately.

Rule 1 confirms that the two have a common origin, indicating that the



	Rule 1	Rule 2	Rule 3
a)	True	True	True
b)	True	False	True
c)	True	True	False
d)	True	False	False
e)	False	False	False

FIG. 1. Schematic putative phylogenetic trees of our data set and its classification regarding the monophyly rules defined in Materials and Methods. Rule 1, monophyly of CRF02_AG plus G; rule 2, monophyly of CRF02_AG separately; rule 3, monophyly of subtype G separately. If our hypothesis is confirmed, our output trees should show the pattern of panel b.

analyzed region is appropriate for this investigation. When rules 2 and 3 are both true, the data are concordant with either CRF02_AG or subtype G being parental (Fig. 1a). When one of the two strains is paraphyletic while the other is monophyletic, then the paraphyletic clade is considered to be the parent of the monophyletic clade, since this indicates that the monophyletic clade falls within the diversity of the paraphyletic clade, as explained above (Fig. 1b and c). When both are paraphyletic, specifically when different trees show conflicting paraphyletic relationships, either there is conflicting evidence, there is not enough phylogenetic signal, or multiple recombination events may have occurred. Figure 1 shows an illustration of the possible scenarios.

Nucleotide sequence accession number. The new subtype J sequence (KTB147) was submitted to the GenBank database and assigned accession number EF 614151.

RESULTS

Analysis of the recombination signal in the currently assumed pure subtypes. When using each subtype as query sequence against the remaining subtypes in a Simplot/bootscan/SlidingBayes analysis, we found clear indications for recombination only in subtype G. Apart from a small region in the bp 1000 to 1500 region, where the Bayesian analysis showed a clustering with subtype H, the recombination pattern was consistent for the different approaches, suggesting that subtype G resulted from recombination between subtypes A and J (Fig. 2a). Since the recombination breakpoints with the two methods were similar but not identical, we used the informative site analysis as implemented in Simplot v3.5.1 for the final assignment of the recombination breakpoints. Based on the results of the informative site test, we performed separate phylogenetic analysis of the nonrecombinant fragments. This analysis showed high support for subtype G clustering within subtype A (including A1 and A2 reference strains) in the bp 4316 to 5162 region, while in the bp 5577 to 6083 fragment, G clustered significantly with subtype J (Figure 2b and c, respectively). For the rest of the genome, no significant support was obtained in the phylogenetic tree analysis, although some of the fragments suggested a close relationship to subtype A, subtype H, and subtype J. However, the short size of these fragments makes it difficult to obtain significant support for these regions.

Reanalysis of the recombination signal for CRF02_AG. Since the results of the pure subtype analysis implied that subtype G was a recombinant, we wanted to reanalyze the recombination pattern of CRF02_AG. The exclusion of subtype G from the reference set makes it possible to assess whether CRF02_AG is actually a recombinant strain between subtypes G and A or whether its recombination pattern is an artifact caused by the fact that subtype G is already a recombinant strain that includes subtype A in its genome. As such, CRF02_AG was submitted as a query to similarity, bootscan, and sliding Bayes analysis including all currently considered pure subtypes with (Fig. 3a) and without (Fig. 3b) "subtype G." While the first analysis confirmed the generally accepted recombination pattern of CRF02_AG, removal of the recombinant "subtype G" from the reference sequences resulted in CRF02_AG showing no evidence of recombination (Fig. 3b).

Further phylogenetic analysis of the near-full-length genome, including all subtypes and subsubtypes, revealed that CRF02_AG clustered within subsubtypes A1 and A2 (data not shown). Although the divergence between A1 and A2 and also between CRF02_AG and A1 or A2 is similar to the divergence between some other subtypes (in particular between subtypes B and D), we do not argue for considering CRF02_AG as a separate subtype but rather as a subsubtype of A.

There is, however, a small region (bp 1650 to 2350) where CRF02_AG is not closely related to subtype A. Since ML phylogenetic analysis showed no evidence of CRF02_AG being derived from subtype G, as these two groups formed two separated monophyletic clusters (data not shown), this fragment in CRF02_AG may have been derived from another source.

Investigating whether CRF02_AG or subtype G is the parent of the common fragments. Based on the results above, we performed a scanning analysis in which subtype G was used as a query sequence and the reference sequence set used CRF02_AG as the representative of subtype A (Fig. 4a). The resulting plot suggested a pattern of recombination between CRF02_AG, subtype H, and subtype J, which is confirmed by phylogenetic analysis (Fig. 4b and 2c).

Based on the results above, two plausible hypotheses could explain the origin of CRF02_AG and subtype G: either subtype G is an A/J recombinant and CRF02_AG is a recombinant of this already recombined "subtype G" or subtype G is a CRF02_AG/J recombinant and CRF02_AG is actually the parental "pure" strain. To discriminate between these two hypotheses, we performed phylogenetic analysis for the longest putative subtype G region of CRF02_AG (500 bp belonging to the integrase region), including all currently available full-genome subtype G and CRF02_AG strains, as subtyped by the REGA subtyping tool (3).

If subtype G is an A/J recombinant and CRF02_AG is a recombinant of subtype G and subtype A, then the subtype G fragment is the parent of the CRF02_AG fragment and can

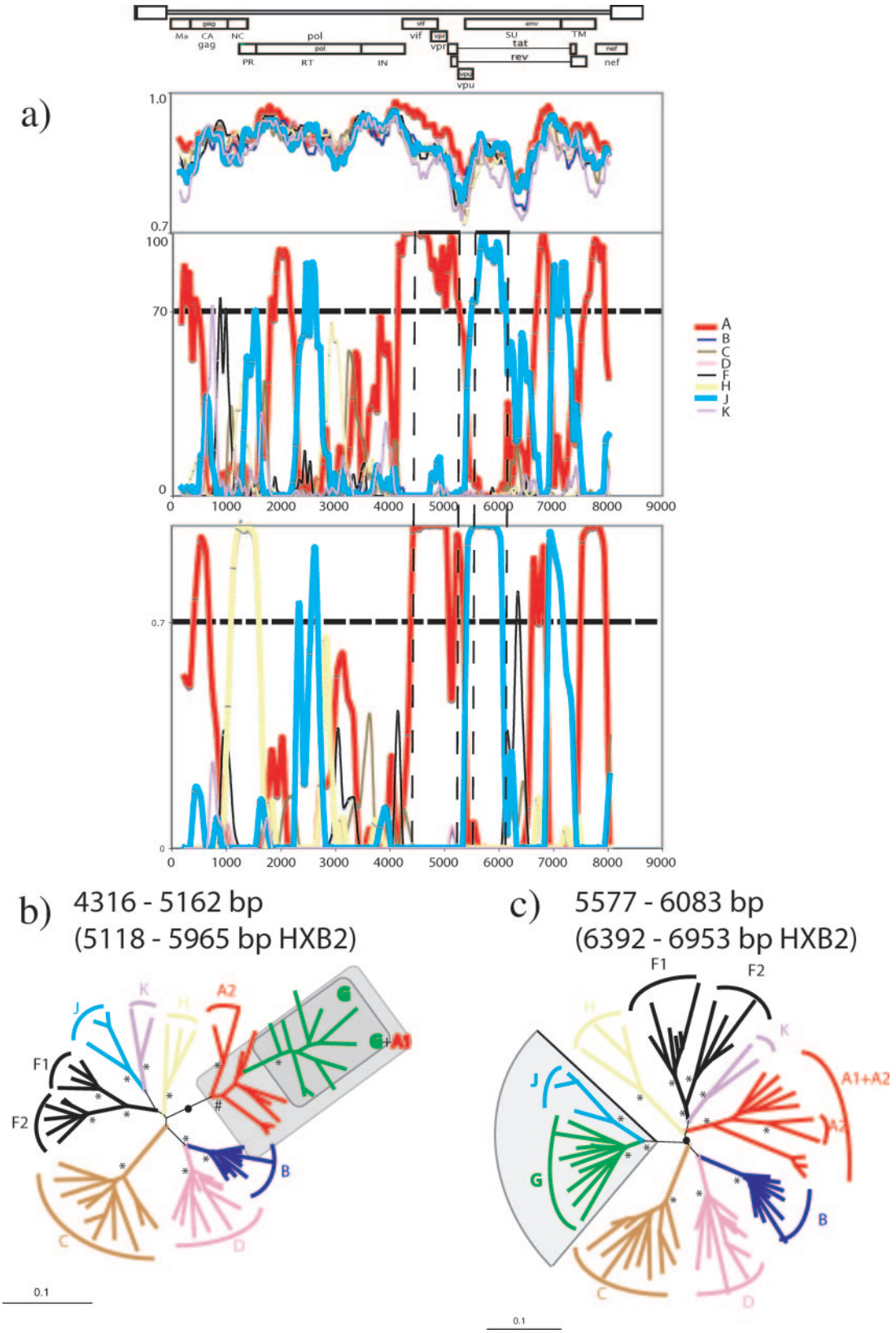


FIG. 2. Recombination analysis of subtype G strains compared to all other pure subtype strains. (a) Similarity (top), bootscanning (middle), and sliding Bayes (bottom) analysis done as described in Materials and Methods, with the gene regions indicated on top and the recombination breakpoints as determined by informative site analysis. (b) ML tree of the genome region between bp 4316 and 5162 as indicated in panel a. (c) ML tree of the genome region between bp 5577 and 6083 as indicated in panel a. The genomic regions illustrated in the tree are indicated in the upper panel. ML trees were generated with PAUP v4b10, as described in Materials and Methods. ●, midpoint root of the tree; ⋆, zero branch length test with $P < 0.001$ and NJ bootstrap support of >70; #, zero branch length test with $P < 0.001$ but NJ bootstrap support of <70.

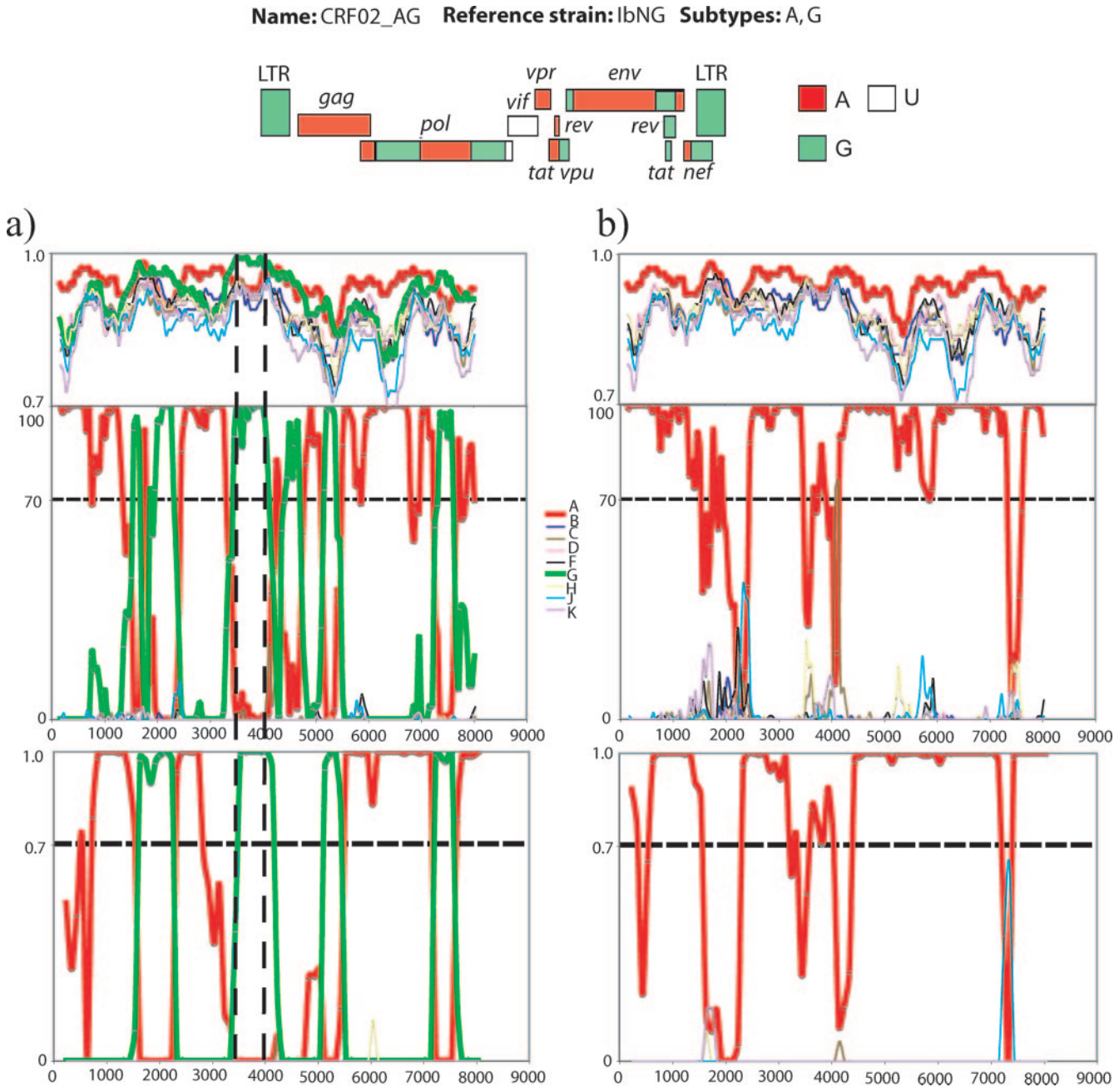
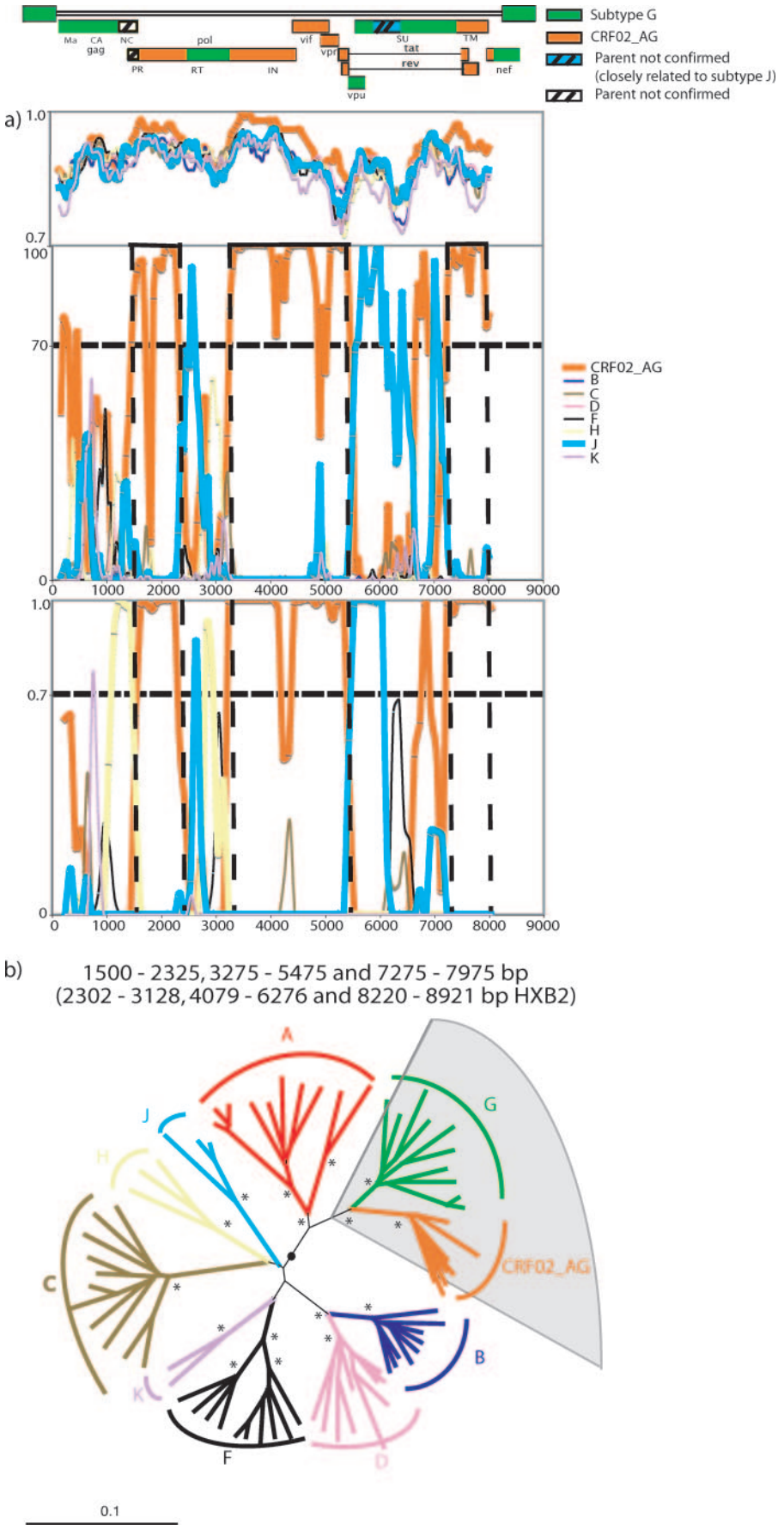


FIG. 3. Recombination analysis of CRF02_AG strains. Similarity (top), bootscanning (middle), and sliding Bayes (bottom) analysis done as described in Materials and Methods, using as subtype reference sequences all pure subtypes including subtype G (a) and all pure subtypes excluding subtype G (b). The recombinant structure as defined in the Los Alamos database is shown on top. The region indicated corresponds to the nonrecombinant region analyzed in the final Bayesian tree (Fig. 4). LTR, long terminal repeat.

therefore be expected to be more diverse, with CRF02_AG clustering within subtype G (CRF02_AG monophyletic and subtype G paraphyletic with respect to CRF02_AG). If the alternative hypothesis is true, the opposite scenario is expected (subtype G monophyletic and CRF02_AG paraphyletic with respect to subtype G) (see definitions of monophyly and paraphyly in Materials and Methods). Bayesian inference with MrBayes (27) showed that the second hypothesis was true: CRF02_AG strains were paraphyletic with respect to the monophyletic subtype G strains, indicating that subtype G arose as a separate lineage from the CRF02_AG diversity and not the other way round (Fig. 5).

We also performed phylogenetic analysis for the other three regions where CRF02_AG is considered to have originated from subtype G. This analysis did not contradict the analysis performed in the integrase region, as subtype G and CRF02_AG formed two separate monophyletic clusters. Furthermore, we also tried to confirm the parental origin of the genomic region


Subtype G
CRF02_AG
Parent not confirmed
(closely related to subtype J)
Parent not confirmed
Ma CA NC
gag
pol
vif
vpr
tat
rev
vpu
SU
TM
nef
PR
RT
IN
a)
1.0
0.7
100
70
0
0 1000 2000 3000 4000 5000 6000 7000 8000 9000
CRF02_AG
B
C
D
F
H
J
K
b)
1500 - 2325, 3275 - 5475 and 7275 - 7975 bp
(2302 - 3128, 4079 - 6276 and 8220 - 8921 bp HXB2)
A
G
J
H
C
CRF02_AG
B
K
F
D
0.1

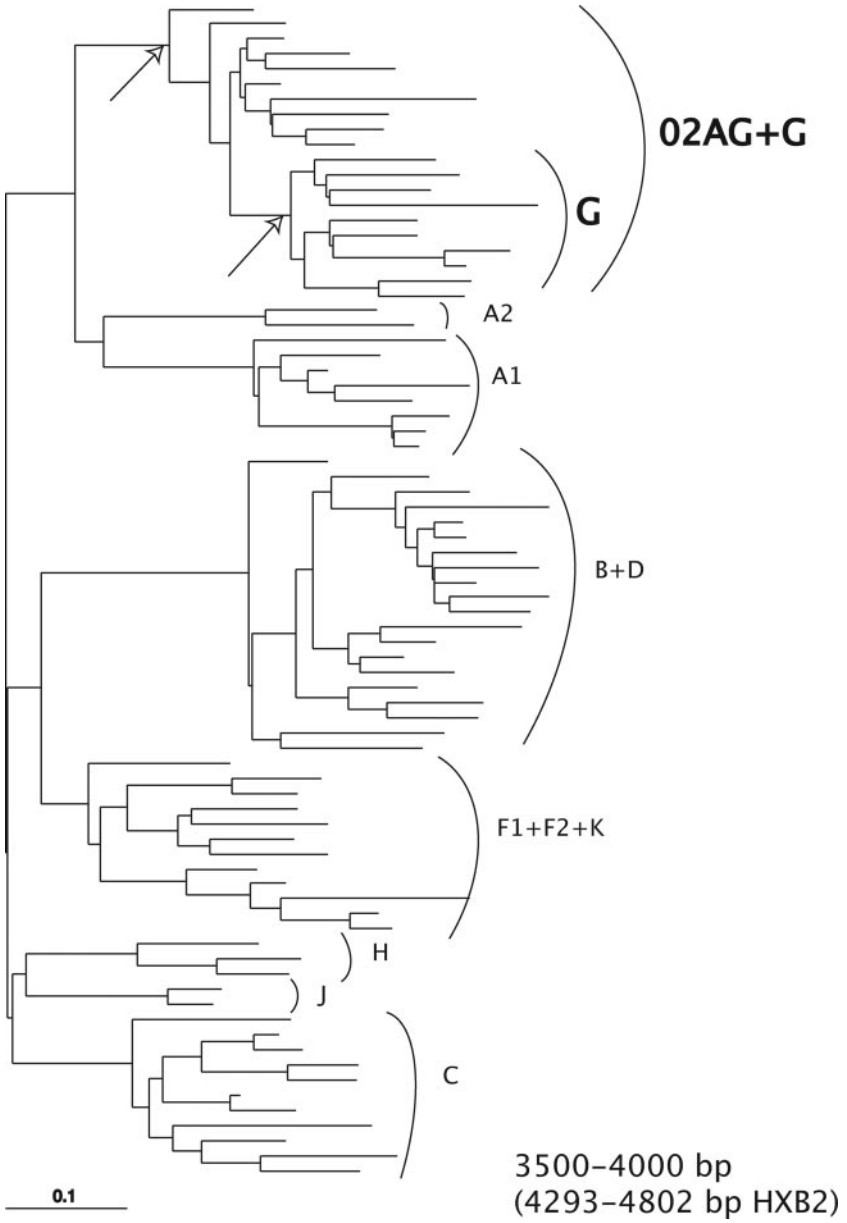


FIG. 5. Phylogenetic analysis to discriminate the parent from the recombinant in the genome region bp 3500 to 4000. The Bayesian tree shown was one of the trees generated by MrBayes in one of two independent MCMC runs. The support of the clustering of CRF02_AG and subtype G was analyzed using the "monophyly rules" described in Materials and Methods. The paraphyletic clade (here CRF02_AG) can be considered the parent, and the monophyletic clade (here subtype G) can be considered the recombinant.

resembling subtype J (bp 5577 to 6083). However, since we obtained two separate monophyletic clusters, we could not make any additional conclusions (data not shown). This could be explained either by the fact that there are very few subtype J strains available or by the fact that subtype J strains are not the real parent strains of subtype G in this region but strains closely related to the parent strains of subtype G. Therefore, we assigned this region as "closely related to subtype J" and draw no conclusions related to which subtype is the parent of this region (Fig. 4a).

FIG. 4. Recombination analysis of subtype G strains compared to all other pure subtype strains and CRF02_AG (considering CRF02_AG as a putative pure subtype representative of subtype A). (a) Similarity (top), bootscanning (middle), and sliding Bayes (bottom) analysis done as described in Materials and Methods and at the top the proposed recombinant structure. (b) ML tree of the merged genome regions bp 1500 to 2325, 3275 to 5475, and 7275 to 7975 as indicated in panel a. ML trees were generated with PAUP v4b10, as described in Materials and Methods. The phylogenetic tree of the J region is shown in Fig. 1c. ●, midpoint root of the tree; ⋆, zero branch length test, $P < 0.001$ and NJ bootstrap support of >70.

To assess the validity of the finding in the integrase region, we used a statistical analysis that records the percentage of posterior trees that had this particular paraphyletic relationship of the CRF02_AG-subtype G cluster through the investigation of three "monophyly rules": monophyletic clustering of CRF02_AG plus subtype G, monophyletic clustering of CRF02_AG alone, and monophyletic clustering of subtype G alone (see Materials and Methods for details). Of trees resulting from both MCMC runs, 99.9% fulfilled the rules concordant with CRF02_AG being the parent of subtype G, with only 10 trees in each run (0.1% per run) showing topologies in which CRF02_AG and "subtype G" formed separated clusters. In these trees, our hypothesis was not confirmed, but it was also not contradicted, since we found separate monophyletic clusters for the two lineages. Therefore, none of the trees resulting from either MCMC run suggested that subtype G was the parent of CRF02_AG.

Recombination pattern of "subtype G." Some regions of the "subtype G" genome could not be assigned to either CRF02_AG or subtype J, and for these regions, we hypothesize a recombinant origin from a putative full-length subtype G (similar to what is assumed for CRF01_AE). "Subtype G" could thus be considered an AGJ recombinant, indicated as in Fig. 4a.

DISCUSSION

The results presented in this work show that the sampling history of subtypes and CRFs has caused a misinterpretation of the evolutionary history of HIV-1 group M. Despite previous analysis suggesting that subtype G is a parental lineage of the proposed recombinant CRF02_AG, our results show that the opposite is most likely. This explains the previous epidemiological findings of an early pandemic of CRF02_AG in Africa and provides an explanation for an early degree of genetic diversity as high as that of the other subtypes (1). The current classification thus reflects the limited sampling at that time for subtype G, CRF02_AG, and subtype A, and the lack of a known second parent for the recombinant "subtype G," subtype J. The fact that even within CRF02_AG some regions are more similar to subtype A than others (Fig. 3b), as has also been observed for CRF01_AE, may be suggestive of unmapped recombination events. This suggests that the evolutionary history of CRF02_AG and other CRFs as it is currently understood may be biased due to incomplete sampling. The classification of a strain as a subtype or a CRF and, in this case, the proposed structure for the recombination structure are therefore highly dependent on the strains that are available at the time of classification.

In our analyses, we included all published full-genome sequences. However, our failure to identify the parental strains of some regions of the subtype G genome suggests that pieces are still missing in the puzzle. Indeed, some of the parental strains may have gone extinct or are as yet undiscovered. We will probably never know the full genetic diversity of HIV at the time of the origin of either CRF02_AG or subtype G. However, our analysis convincingly shows that the current circulating CRF02_AG strains are paraphyletic to the current circulating subtype G strains, so there is no doubt that, for example, for the integrase gene providing the strongest statistical support, the MRCA of the current CRF02_AG strains is ancestral to the MRCA of the current subtype G strains (Fig. 5), indicating that a CRF02_AG-related virus was the parent of the integrase in this recombinant "subtype" G.

Recombination complicates the analysis of the evolutionary history of organisms, as different genomic regions will give discordant results. Here, we show that the high recombination rates observed for HIV can indeed mislead the interpretation of its evolutionary history. Biological interpretations based on the recombinant or nonrecombinant origin of strains should therefore be made with great caution. An example of interpretation based on recombination signal is the current interest in the biological significance of recombination hotspots (16). In such analyses, caution should be taken when assigning the parental strains of the putative recombinants, as the erroneous assignment of parental strains may give rise to misleading results. This is applicable to all viruses known to have high recombination rates and is especially important since most methods for detecting recombination depend on an initial assumption of parental strains.

Finally, our findings urge a reassessment of the HIV-1 evolutionary history. Further detailed analyses will be needed to verify whether the entire notion of "subtype" and "recombinant" applies to HIV-1. As current phylogenetic methods are not capable of accurately reconstructing the evolutionary histories of highly recombinant sequences, it may never be possible to correctly assign for all strains which one is the recombinant and which one is the parent.

ACKNOWLEDGMENTS

A.B.A. was supported by Fundação para a Ciência e Tecnologia (grant no. SFRH/BD/19334/2004). P.L. was supported by an EMBO long-term fellowship. B.S. was funded by the Wellcome Trust. A.R. was funded by the Royal Society. This work was supported by the Flemish Fonds voor Wetenschappelijk Onderzoek (FWO G.0513.06) and by the Early-Stage Host Fellowship GeneTime, from the Marie Curie program of the European Commission (grant no. MEST-CT.2004-007909).

REFERENCES

1. **Carr, J. K., T. Laukkanen, M. O. Salminen, J. Albert, A. Alaeus, B. Kim, E. Sanders-Buell, D. L. Birx, and F. E. McCutchan.** 1999. Characterization of subtype A HIV-1 from Africa by full genome sequencing. AIDS **13:**1819–1826.
2. **Carr, J. K., M. O. Salminen, J. Albert, E. Sanders-Buell, D. Gotte, D. L. Birx, and F. E. McCutchan.** 1998. Full genome sequences of human immunodeficiency virus type 1 subtypes G and A/G intersubtype recombinants. Virology **247:**22–31.
3. **de Oliveira, T., K. Deforche, S. Cassol, M. Salminen, D. Paraskevis, C. Seebregts, J. Snoeck, E. J. van Rensburg, A. M. Wensing, D. A. van de Vijver, C. A. Boucher, R. Camacho, and A. M. Vandamme.** 2005. An automated genotyping system for analysis of HIV-1 and other microbial sequences. Bioinformatics **21:**3797–3800.
4. **Drummond, A. J., G. K. Nicholls, A. G. Rodrigo, and W. Solomon.** 2002. Estimating mutation parameters, population history and genealogy simultaneously from temporally spaced sequence data. Genetics **161:**1307–1320.
5. **Fonjungo, P. N., E. N. Mpoudi, J. N. Torimiro, G. A. Alemnji, L. T. Eno, E. J. Lyonga, J. N. Nkengasong, R. B. Lal, M. Rayfield, M. L. Kalish, T. M. Folks, and D. Pieniazek.** 2002. Human immunodeficiency virus type 1 group M protease in Cameroon: genetic diversity and protease inhibitor mutational features. J. Clin. Microbiol. **40:**837–845.
6. **Fonjungo, P. N., E. N. Mpoudi, J. N. Torimiro, G. A. Alemnji, L. T. Eno, J. N. Nkengasong, F. Gao, M. Rayfield, T. M. Folks, D. Pieniazek, and R. B. Lal.** 2000. Presence of diverse human immunodeficiency virus type 1 viral variants in Cameroon. AIDS Res. Hum. Retrovir. **16:**1319–1324.
7. **Gao, F., S. G. Morrison, D. L. Robertson, C. L. Thornton, S. Craig, G. Karlsson, J. Sodroski, M. Morgado, B. Galvao-Castro, H. von Briesen, S. Beddows, J. Weber, P. M. Sharp, G. M. Shaw, B. H. Hahn, and the WHO and NIAID Networks for HIV Isolation and Characterization.** 1996. Molecular

cloning and analysis of functional envelope genes from human immunodeficiency virus type 1 sequence subtypes A through G. J. Virol. **70:**1651–1667.
8. **Gao, F., D. L. Robertson, C. D. Carruthers, S. G. Morrison, B. Jian, Y. Chen, F. Barre-Sinoussi, M. Girard, A. Srinivasan, A. G. Abimiku, G. M. Shaw, P. M. Sharp, and B. H. Hahn.** 1998. A comprehensive panel of near-full-length clones and reference sequences for non-subtype B isolates of human immunodeficiency virus type 1. J. Virol. **72:**5680–5698.
9. **Gelman, A., and D. B. Rubin.** 1992. Inference from iterative simulation using multiple sequences. Stat. Sci. **7:**457–472.
10. **Hemelaar, J., E. Gouws, P. D. Ghys, and S. Osmanov.** 2006. Global and regional distribution of HIV-1 genetic subtypes and recombinants in 2004. AIDS **20:**W13–W23.
11. **Howard, T. M., D. O. Olaylele, and S. Rasheed.** 1994. Sequence analysis of the glycoprotein 120 coding region of a new HIV type 1 subtype A strain (HIV-1IbNg) from Nigeria. AIDS Res. Hum. Retrovir. **10:**1755–1757.
12. **Keele, B. F., F. Van Heuverswyn, Y. Li, E. Bailes, J. Takehisa, M. L. Santiago, F. Bibollet-Ruche, Y. Chen, L. V. Wain, F. Liegeois, S. Loul, E. Mpoudi Ngole, Y. Bienvenue, E. Delaporte, J. F. Brookfield, P. M. Sharp, G. M. Shaw, M. Peeters, and B. H. Hahn.** 2006. Chimpanzee reservoirs of pandemic and nonpandemic HIV-1. Science **313:**523–526.
13. **Laukkanen, T., J. Albert, K. Liitsola, S. D. Green, J. K. Carr, T. Leitner, F. E. McCutchan, and M. O. Salminen.** 1999. Virtually full-length sequences of HIV type 1 subtype J reference strains. AIDS Res. Hum. Retrovir. **15:**293–297.
14. **Leitner, T., A. Alaeus, S. Marquina, E. Lilja, K. Lidman, and J. Albert.** 1995. Yet another subtype of HIV type 1? AIDS Res. Hum. Retrovir. **11:**995–997.
15. **Lole, K. S., R. C. Bollinger, R. S. Paranjape, D. Gadkari, S. S. Kulkarni, N. G. Novak, R. Ingersoll, H. W. Sheppard, and S. C. Ray.** 1999. Full-length human immunodeficiency virus type 1 genomes from subtype C-infected seroconverters in India, with evidence of intersubtype recombination. J. Virol. **73:**152–160.
16. **Magiorkinis, G., D. Paraskevis, A. M. Vandamme, E. Magiorkinis, V. Sypsa, and A. Hatzakis.** 2003. In vivo characteristics of human immunodeficiency virus type 1 intersubtype recombination: determination of hot spots and correlation with sequence similarity. J. Gen. Virol. **84:**2715–2722.
17. **McCutchan, F. E., J. K. Carr, M. Bajani, E. Sanders-Buell, T. O. Harry, T. C. Stoeckli, K. E. Robbins, W. Gashau, A. Nasidi, W. Janssens, and M. L. Kalish.** 1999. Subtype G and multiple forms of A/G intersubtype recombinant human immunodeficiency virus type 1 in Nigeria. Virology **254:**226–234.
18. **Montavon, C., C. Toure-Kane, F. Liegeois, E. Mpoudi, A. Bourgeois, L. Vergne, J. L. Perret, A. Boumah, E. Saman, S. Mboup, E. Delaporte, and M. Peeters.** 2000. Most env and gag subtype A HIV-1 viruses circulating in West and West Central Africa are similar to the prototype AG recombinant virus IBNG. J. Acquir. Immune Defic. Syndr. **23:**363–374.
19. **Montavon, C., C. Toure-Kane, J. N. Nkengasong, L. Vergne, K. Hertogs, S. Mboup, E. Delaporte, and M. Peeters.** 2002. CRF06-cpx: a new circulating recombinant form of HIV-1 in West Africa involving subtypes A, G, K, and J. J. Acquir. Immune Defic. Syndr. **29:**522–530.
20. **Niama, F. R., C. Toure-Kane, N. Vidal, P. Obengui, B. Bikandou, M. Y. Ndoundou Nkodia, C. Montavon, H. Diop-Ndiaye, J. V. Mombouli, E. Mokondzimobe, A. G. Diallo, E. Delaporte, H. J. Parra, M. Peeters, and S. Mboup.** 2006. HIV-1 subtypes and recombinants in the Republic of Congo. Infect Genet. Evol. **6:**337–343.
21. **Paraskevis, D., K. Deforche, P. Lemey, G. Magiorkinis, A. Hatzakis, and A. M. Vandamme.** 2005. SlidingBayes: exploring recombination using a sliding window approach based on Bayesian phylogenetic inference. Bioinformatics **21:**1274–1275.
22. **Peeters, M., E. Esu-Williams, L. Vergne, C. Montavon, C. Mulanga-Kabeya, T. Harry, A. Ibironke, D. Lesage, D. Patrel, and E. Delaporte.** 2000. Predominance of subtype A and G HIV type 1 in Nigeria, with geographical differences in their distribution. AIDS Res. Hum. Retrovir. **16:**315–325.
23. **Rambaut, A., D. Posada, K. A. Crandall, and E. C. Holmes.** 2004. The causes and consequences of HIV evolution. Nat. Rev. Genet. **5:**52–61.
24. **Rambaut, A., D. L. Robertson, O. G. Pybus, M. Peeters, and E. C. Holmes.** 2001. Human immunodeficiency virus. Phylogeny and the origin of HIV-1. Nature **410:**1047–1048.
25. **Rhodes, T., H. Wargo, and W. S. Hu.** 2003. High rates of human immunodeficiency virus type 1 recombination: near-random segregation of markers one kilobase apart in one round of viral replication. J. Virol. **77:**11193–11200.
26. **Robertson, D. L., J. P. Anderson, J. A. Bradac, J. K. Carr, B. Foley, R. K. Funkhouser, F. Gao, B. H. Hahn, M. L. Kalish, C. Kuiken, G. H. Learn, T. Leitner, F. McCutchan, S. Osmanov, M. Peeters, D. Pieniazek, M. Salminen, P. M. Sharp, S. Wolinsky, and B. Korber.** 2000. HIV-1 nomenclature proposal. Science **288:**55–56.
27. **Ronquist, F., and J. P. Huelsenbeck.** 2003. MrBayes 3: Bayesian phylogenetic inference under mixed models. Bioinformatics **19:**1572–1574.
28. **Salminen, M.** 2003. Detecting recombination in viral sequences, p. 348–377. *In* M. Salemi and A.-M. Vandamme (ed.), The phylogenetic handbook—a practical approach to DNA and protein phylogeny. Cambridge University Press, Cambridge, United Kingdom.
29. **Smith, R. A., L. A. Loeb, and B. D. Preston.** 2005. Lethal mutagenesis of HIV. Virus Res. **107:**215–228.
30. **Swofford, D.** 1998. PAUP* 4.0—Phylogenetic Analysis Using Parsimony (* and Other Methods). Sinauer Associates, Sunderland, MA.
31. **Tebit, D. M., L. Zekeng, L. Kaptue, M. Salminen, H. G. Krausslich, and O. Herchenroder.** 2002. Genotypic and phenotypic analysis of HIV type 1 primary isolates from western Cameroon. AIDS Res. Hum. Retrovir. **18:**39–48.
32. **Thompson, J. D., D. G. Higgins, and T. J. Gibson.** 1994. CLUSTAL W: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice. Nucleic Acids Res. **22:**4673–4680.
33. **Vidal, N., C. Mulanga, S. E. Bazepeo, F. Lepira, E. Delaporte, and M. Peeters.** 2006. Identification and molecular characterization of subsubtype A4 in central Africa. AIDS Res. Hum. Retrovir. **22:**182–187.
34. **Vidal, N., C. Mulanga, S. E. Bazepeo, J. K. Mwamba, J. W. Tshimpaka, M. Kashi, N. Mama, C. Laurent, F. Lepira, E. Delaporte, and M. Peeters.** 2005. Distribution of HIV-1 variants in the Democratic Republic of Congo suggests increase of subtype C in Kinshasa between 1997 and 2002. J. Acquir. Immune Defic. Syndr. **40:**456–462.
35. **Vidal, N., M. Peeters, C. Mulanga-Kabeya, N. Nzilambi, D. Robertson, W. Ilunga, H. Sema, K. Tshimanga, B. Bongo, and E. Delaporte.** 2000. Unprecedented degree of human immunodeficiency virus type 1 (HIV-1) group M genetic diversity in the Democratic Republic of Congo suggests that the HIV-1 pandemic originated in Central Africa. J. Virol. **74:**10498–10507.