

Mosaic Genomes of the Six Major Primate Lentivirus Lineages Revealed by Phylogenetic Analyses

Marco Salemi,^{1*} Tulio De Oliveira,² Valerie Courgnaud,³ Vincent Moulton,⁴
Barbara Holland,⁵ Sharon Cassol,² William M. Switzer,⁶
and Anne-Mieke Vandamme¹

Rega Institute for Medical Research, KULeuven, Leuven, Belgium¹; Molecular Virology and Bioinformatics Unit, Africa Centre, Nelson Mandela School of Medicine, Durban, South Africa²; Laboratoire Retrovirus, IRD, Montpellier, France³; Linnaeus Center for Bioinformatics, Uppsala University, Uppsala, Sweden⁴; Allan Wilson Centre for Molecular Ecology and Evolution, Massey University, Palmerston North, New Zealand⁵; and Human Immunodeficiency Virus and Retrovirology Branch, Division of AIDS, Sexually Transmitted Diseases, and Tuberculosis, Centers for Disease Control and Prevention, Atlanta, Georgia⁶

Received 23 October 2002/Accepted 3 April 2003

To clarify the origin and evolution of the primate lentiviruses (PLVs), which include human immunodeficiency virus types 1 and 2 as well as their simian relatives, simian immunodeficiency viruses (SIVs), isolated from several host species, we investigated the phylogenetic relationships among the six supposedly nonrecombinant PLV lineages for which the full genome sequences are available. Employing bootscanning as an exploratory tool, we located several regions in the PLV genome that seem to have uncertain or conflicting phylogenetic histories. Phylogeny reconstruction based on distance and maximum-likelihood algorithms followed by a number of statistical tests confirms the existence of at least five putative recombinant fragments in the PLV genome with different clustering patterns. Split decomposition analysis also shows that phylogenetic relationships among PLVs may be better represented by network-based graphs, such as the ones produced by SplitsTree. Our findings not only imply that the six so-called pure PLV lineages have in fact mosaic genomes but also make more unlikely the hypothesis of cospeciation of SIVs and their simian hosts.

The primate lentivirus (PLV) group includes simian immunodeficiency viruses (SIVs) isolated from a number of host species (18) and human immunodeficiency virus types 1 and 2 (HIV-1 and HIV-2). SIVs do not usually cause any disease in their natural hosts (34), but they are designated immunodeficiency viruses after HIV-1, the etiologic agent of AIDS, with which they share genetic and structural similarities (18). These viruses display a high sequence variability, and several circulating HIV and SIV recombinant strains have also been identified (4, 16, 17, 21, 27, 35, 37, 42).

The PLV strains are currently assigned to six approximately equidistant phylogenetic lineages (9, 18): (i) the SIVcpz clade, joining SIVcpz strains from African chimpanzees (*Pan troglodytes*) and HIV-1 groups M, N, and O; (ii) the SIVsmm clade, including HIV-2 subtypes as well as SIVsmm and SIVstm isolated from sooty mangabeys (*Cercocebus atys*) and stump-tailed macaques, respectively; (iii) the SIVagm clade, which clusters together viral strains from three species of African green monkey, namely, vervet monkeys (*Chlorocebus pygerythrus*; SIVagmVer), grivet monkeys (*Chlorocebus aethiops*; SIVagmGri), and tantalus monkeys (*Chlorocebus tantalus*; SIVagmTan); (iv) a group known as the SIVlhoest clade, joining the strains SIVlhoest from L'Hoest monkeys (*Cercopithecus lhoesti*), SIVsun from sun-tailed monkeys (*Cercopithecus solatus*), and SIVmnd from mandrills (*Mandrillus sphinx*); (v) a

divergent strain, SIVsyk, so far isolated from only one Sykes monkey (*Cercopithecus albogularis*); and (vi) the divergent SIVcol strain recently isolated from a guereza colobus monkey.

Phylogenetic analyses also provide evidence for the existence of at least three SIV mosaic genomes: (i) SIVsab, isolated from West African sabaeus monkeys (21); (ii) SIVrcm, isolated from red-capped mangabeys (4, 17); and (iii) SIVmnd2, isolated from mandrills (42).

African monkeys are the natural hosts of PLVs, whereas the SIVs isolated from Asian macaques reflect cross-species transmission from captive sooty mangabeys (5, 19). It has been demonstrated also that HIV-1 and HIV-2 were introduced into the human population through contacts with infected simians (7, 8, 13–15). The six different HIV-2 subtypes seem to have arisen from at least four distinct interspecies transmissions from West African sooty mangabeys (8, 15). At least one zoonotic transmission, and probably three, from SIV-infected *Pan troglodytes troglodytes* (SIVcpz) is responsible for the origin of HIV-1 groups M, O, and N infection in our species (13). Other PLVs appear to be phylogenetically related according to the species rather than the geographic origins of their hosts (18, 47). For example, the SIVagmVer strains, isolated from vervet monkeys living in East and South Africa, are monophyletic and cluster, in turn, with the SIVagmTan strain isolated from a Central African tantalus monkey and with the SIVagmGri strain from an East African grivet monkey (see Fig. 1). This fact has been explained by assuming that the common ancestor of the African green monkey species was infected with the common ancestor of the SIVagm lineage, followed by coevolution of virus and host (12, 18, 28). Another example of

* Corresponding author. Present address: Department of Ecology and Evolutionary Biology, 383 Steinhaus Hall, University of California at Irvine, Irvine, CA 92697. Phone: (949) 824-1056. Fax: (949) 824-2181. E-mail: msalemi68@hotmail.com.

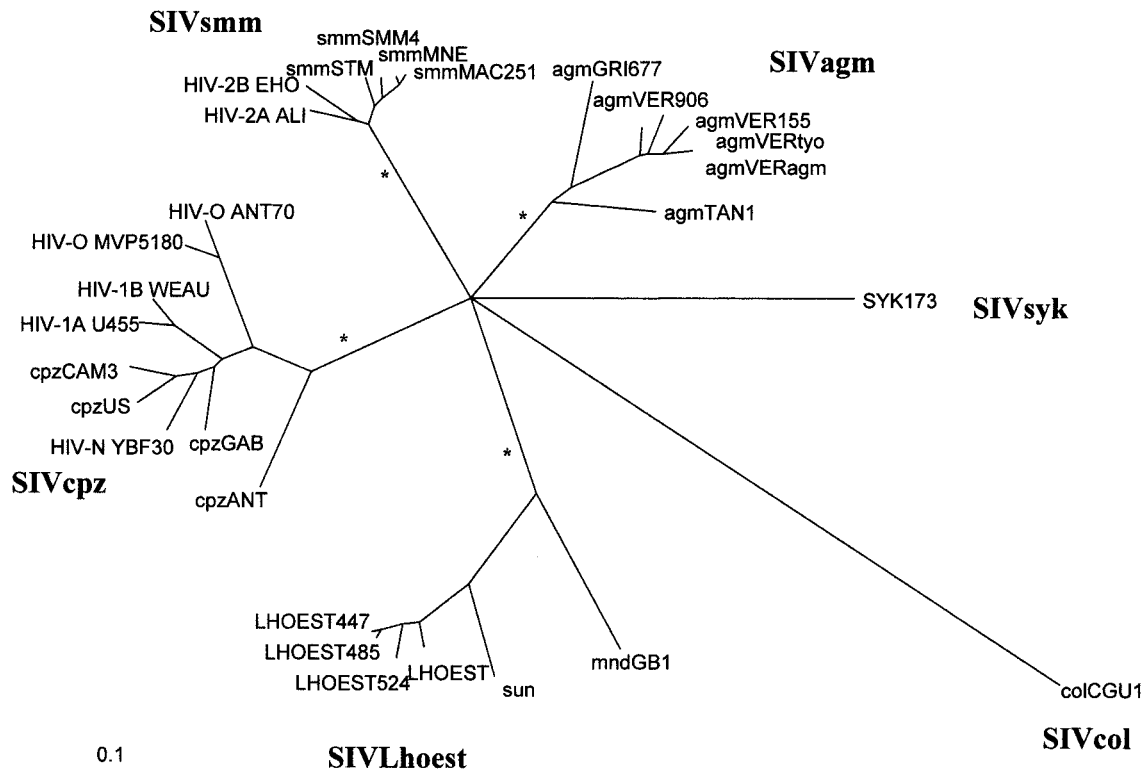


FIG. 1. Consensus phylogenetic tree representing the relationships among the six major lineages of PLVs. Conventional names for the PLV clades are given in bold. Horizontal branch lengths are drawn to scale, with the bar indicating 0.1 nt replacements per site, and were inferred by maximum likelihood with the GTR+ Γ +I model of nucleotide substitution (see Results) by using only first and second codon positions. Asterisks on the branches indicate statistical support ($P < 0.001$) for the monophyletic lineage.

possible cospeciation is the clustering of SIVLhoest, from L'Hoest monkeys, with SIVsun, from sun-tailed monkeys, both members of the genus *Cercopithecus* and belonging to the L'Hoest superspecies (2, 18). On the other hand, the fact that SIVmnd, from mandrills, falls within the same clade may be explained as a cross-species transmission from an unknown source possibly related to L'Hoest and sun-tailed monkeys (18, 20). An SIV transmission from African green monkeys to yellow baboons has also been reported, showing evidence that cross-species transmission has been happening in the wild even in recent times (21).

The more common conclusion in the literature, so far, is that host-specific virus evolution of PLVs is generally the rule, though clear-cut examples exist of simian-to-simian and simian-to-human cross-species transmissions (18, 40).

Uncertainty remains about the precise phylogenetic relationships among the six major PLV lineages. Several phylogenetic trees have been reported in the literature. Depending on the gene region or the algorithm used to infer the phylogenetic relationships, the relative branching pattern of the major PLV lineages does not appear to be stable. For example, Beer et al. (3) obtained separate neighbor-joining PLV trees for *gag*, *pol*, and *env*: such trees show that the SIVsmm lineage is monophyletic with the SIVagm lineage in *gag* but with the SIVsyk lineage in *pol*. In another example, the maximum-likelihood tree based on full-length PLV Pol protein sequences reported by Hahn et al. (18) clusters SIVagm with SIVLhoest. There is

no doubt that the strains belonging to each of the six equidistant lineages constitute a monophyletic clade, whatever genomic region is used in the analysis; i.e., the full genome sequences within each lineage trace back eventually to a unique common ancestor. The fact that the precise phylogenetic relationships among the major clades remain elusive could be due to the loss of phylogenetic signal near the root of the PLV tree because of substitution saturation or to an inability of the phylogeny inference methods currently used to model properly the evolution of these viruses. Although some regions of the PLV genome appear to have undergone substitution saturation, the first and second codon positions of *gag*, *pol*, and *env* sequences retain enough information to allow a reliable investigation of the PLV phylogeny (10).

In what follows, we show that the supposedly pure PLV lineages have, in fact, mosaic genomes that were probably acquired through several interspecies transmission and recombination events close to the root of the tree. To support this claim, we analyzed full-genome PLV sequences with several tree-building methods, based on distance as well as maximum-likelihood algorithms, and with split decomposition analysis, a network-building approach that can take into account, and visually represent, conflicting phylogenetic relations within the data under investigation. Our findings not only make more unlikely the cospeciation hypothesis but also provide new insights on the origin and evolution of this important group of primate viruses.

TABLE 1. Best-fitting nucleotide substitution model (first and second codon positions only) for the five putative recombinant fragments of the six major PLV lineages

Fragment ^a	Corresponding amino acid position ^b in:			Model	α^c	P _{inv} ^d	Ti/Tv ^e
	Gag	Pol	Env				
PLV_REC1	1–460			HKY+ Γ	0.55		1.12
PLV_REC2	461–529	1–404		TVM+I+ Γ	0.88	0.19	1.62
PLV_REC3		405–591		HKY+I+ Γ	1.37	0.19	1.27
PLV_REC4		592–812		HKY+ Γ	0.61		1.05
PLV_REC5		913–957	1–648	TN+I+ Γ	0.96	0.10	1.08

^a Putative recombinant fragments inferred from the results of the bootscanning analyses shown in Fig. 3 (see the text for more details).

^b According to the absolute amino acid positions of the SIVcpzANT coding sequences used as a reference.

^c Shape parameter of the Γ distribution estimated via maximum likelihood.

^d Proportion of invariable sites estimated via maximum likelihood.

^e Expected transition-transversion ratio estimated via maximum likelihood.

MATERIALS AND METHODS

PLV data set. Full-length genomes representing each of the six major PLV lineages were downloaded from the Los Alamos HIV database (<http://hiv-web.lanl.gov>). We also included HIV-1 M, O, and N strains, which cluster within the SIVcpz clade, and HIV-2 A and B strains, which cluster within the SIVsmm clade. Concatemer sequences of the nonoverlapping *gag*, *pol*, *vif*, *env*, and *nef* nucleotide regions were obtained. Nucleotides were aligned against their predicted amino acid sequences, in order to avoid the introduction of insertions or deletions within a codon, by using the Clustal algorithm implemented in DAMBE (48; X. Xia, DAMBE [data analysis in molecular biology and evolution] version 4.0 software package, Department of Ecology and Biodiversity, University of Hong Kong, Hong Kong).

In a previous study using the same data set, the observed saturation index at each codon position of the PLV concatemer was compared with half of the theoretical index expected in case of full substitution saturation (10). The indexes are computed by using the notion of information entropy (49), and the algorithm is implemented in the program DAMBE (48; Xia, DAMBE software package). The result clearly shows that third codon positions of *gag*, *pol*, *vif*, *env*, and *nef*, as well as *vif* and *nef* first and second codon positions, are significantly saturated (10). Therefore, they were excluded from the analysis. In contrast, first and second codon positions in *gag*, *pol*, and *env* still retain enough information for the reliable inference of phylogenetic relationships (10).

A similar alignment was also obtained for amino acid sequences. In what follows, we analyze only the nucleotide alignment; however, identical results were obtained by using amino acids (data not shown), which is not surprising since the nucleotide alignment includes mostly nonsynonymous positions. The taxa used in the analysis are shown in the consensus tree given in Fig. 1. The complete alignment is available upon request.

PLV consensus tree. The general time-reversible model with Γ -distributed rates across sites (GTR+ Γ model) has been tested and found to be the most appropriate for modeling the nucleotide substitution process in PLVs (22, 24, 40). By applying the hierarchical likelihood ratio test (LRT) strategy implemented in the MODELTEST version 3.06 program (31), we found the GTR+ Γ model with a fraction of sites assumed to be invariable (GTR+ Γ +I model) to be the best fitting for the PLV concatemer data set. In order to perform the test, a star-like tree, like the one shown in Fig. 1, was used. Such a tree is certainly not the best picture of the PLV evolutionary history, but it is not too wrong considering that the monophyletic origin of the strains within each of the major clades is highly supported in each region of the PLV genome. It has been shown that the use of any reasonably good tree for the data, i.e., one that is much better than a randomly chosen tree and includes clades that are well supported under any optimality criterion, will not critically alter the testing of evolutionary models, since parameter estimates do not vary much from tree to tree (43, 46, 50).

A maximum-likelihood tree was obtained by using the selected model. As expected, the tree shows six long branches statistically supported ($P < 0.001$) by the zero-branch-length test (J. Felsenstein, PHYLIP [phylogenetic inference package] version 3.5c software documentation, Department of Genetics, University of Washington, Seattle) and leading to the supposedly pure PLV lineages. However, none of the short internal branches connecting the major clades had results significantly different from zero ($P > 0.05$). Therefore, such branches were collapsed and the resulting star-like tree, shown in Fig. 1, should be thought of as a consensus tree giving a schematic representation of the PLV monophyletic clades, with the central node as a soft polytomy representing our uncertainty about the exact phylogenetic relationships among the major clades.

Bootscanning analysis. Bootscanning plots were obtained by employing the concatenated nucleotide alignment with the SIMPLOT package (S. Ray, SIMPLOT version 2.5 software documentation, 1999). After the sequences were grouped per lineage, a sequence from each major PLV clade was used in turn as a query against those of all the others, with a sliding window of 500 nucleotides (nt) moved in steps of 20 nt and maximum-likelihood-estimated distances. As a control, a bootscanning analysis was also performed on the SIVrcm isolate NG411 (AF349680), which has a known mosaic genome (4), against all the major lineages.

Phylogenetic inference based on tree-building methods. As discussed below in Results, bootscanning plots suggest the existence of at least five recombinant fragments within the PLV genome (see Fig. 3 and Table 1). Separate phylogenetic analyses were carried out for each fragment with distance, as well as maximum-likelihood-based methods, as follows. The best-fitting nucleotide substitution model was evaluated with MODELTEST version 3.06 (31). Minimum evolution and weighted least-square with inverse-square weighting objective functions were used to infer neighbor-joining and Fitch-Margoliash trees, respectively (45). Maximum-likelihood trees were obtained by implementing a heuristic search with Tree Bisection Reconnection branch swapping, since exhaustive or branch and bound searches, which are guaranteed to find the optimal maximum-likelihood tree, are not feasible for more than 10 to 15 taxa due to the exponential increase of the possible topologies. Starting trees for the heuristic search were obtained by both neighbor joining and random sequence addition (10 repeats), leading in each case to the same result. One thousand bootstrapping and jackknifing resamplings were used to evaluate the reliability of the distance-based trees. Statistical support for the maximum-likelihood trees was assessed with the zero-branch-length test (Felsenstein, PHYLIP software documentation). In such a test, each branch of a tree is collapsed in turn. Upon full reoptimization of the tree parameters by maximum likelihood, a probability is calculated, by comparison with the original tree, of obtaining a likelihood ratio as large as, or larger than, the ratio observed under the null hypothesis that the given branch has zero length. Topologies of different trees were compared with the use of the Shimodaira-Hasegawa (S-H) test with resampling-estimated log likelihood (RELL) bootstrapping by using 1,000 bootstrap replicates (41). Phylogenetic analyses and topological tests were performed with PAUP*4.0b10 (D. L. Swofford, PAUP*, phylogeny analysis based on parsimony [*and other methods], version 4, Sinauer Associates, Sunderland, Mass.).

Simulation of recombinant data sets. One hundred recombinant data sets were simulated under the GTR+ Γ +I model by using the parameters estimated for the whole PLV data set with the Seq-Gen program (33). Each simulated set consists of 22 taxa 3,282 nt long with four recombinant fragments of the same lengths as the fragments inferred from the original data (PLV_REC1, PLV_REC2, PLV_REC4, and PLV_REC5; see Table 1), each one following the corresponding tree (see Fig. 4). After a tree for each recombinant fragment in each simulated set was reestimated, the log likelihood of the GTR+ Γ +I model was compared with those of TVM+ Γ +I (which assumes a transition-transversion bias and four different rates for the four different transversions, with Γ -distributed relative nucleotide substitution rates across sites and a fraction of sites being invariable), TN+ Γ +I (which assumes a transition-transversion bias and a purine transition-pyrimidine transition bias, with Γ -distributed relative nucleotide substitution rates across sites and a fraction of sites being invariable), and HKY+ Γ +I (which assumes a transition-transversion bias, with Γ -distributed relative nucleotide substitution rates across sites and a fraction of sites being invariable) by using the LRT.

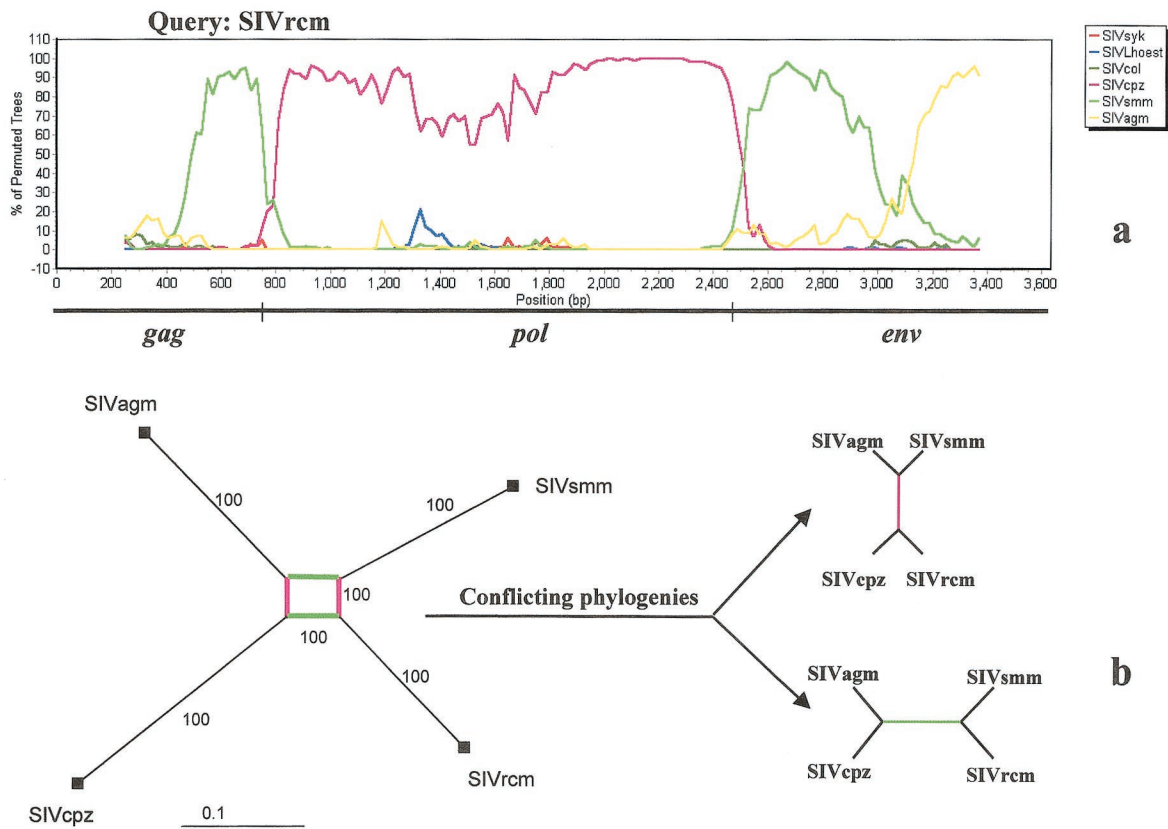


FIG. 2. (a) Results of bootscanning of SIVrcm against the six major PLV lineages. A concatemer with the nonoverlapping regions of *gag*, *pol*, and *env* was obtained. The bootscanning analysis was performed on first and second codon positions with a sliding window of 500 nt (20 nt/step) and 1,000 bootstrap replicates. (b, left panel) Results of split decomposition analysis of the concatemer alignment including only SIVrcm, SIVvagn, SIVsmm, and SIVcpz isolates. Since several isolates of SIVvagn, SIVsmm, and SIVcpz were used in the analysis, such names represent the monophyletic group rather than a particular isolate. Distances were obtained with the GTR+ Γ +I model of nucleotide substitution (see Results), and 1,000 bootstrap replicates were generated to assess the reliability of each edge in the split graph. The percentages of bootstrap replicates are reported on the edge. Edges are drawn to scale, with the bar indicating 0.1 nt replacements per site. (Right panel) The two conflicting phylogenies represented by the split graph are depicted, with internal branches color-coded according to the internal splits in the graph on the left. Each internal split is supported by 100% of bootstrap replicates (see left panel).

Clock-like phylogenetic trees. All the possible rooted trees (for a tree of n taxa, $3n - 2$) were obtained from the corresponding maximum-likelihood tree, and ultrametric branches, where all the tips are equidistant from the root, were estimated for each tree via maximum likelihood by enforcing the molecular clock constraint. The rooted tree with the best likelihood was chosen, and its likelihood was compared with that of the unrooted tree in an LRT (degrees of freedom, $n - 2$) in order to check whether the assumption of rate constancy for all lineages significantly reduced the probability of the data (11).

Split decomposition analysis. Split decomposition analysis allows the canonical decomposition of any distance measure (such as the genetic distances generated from a set of aligned nucleotide or amino acid sequences) into the sum of split metrics plus a split prime residue. More precisely, given a set X of taxa ($X = A, B, C, \dots$, etc.) and a distance $d(A,B)$ in X , the isolation index α_S of any split $S(U,V)$ of X is defined

$$\alpha_S = 1/2 \text{ minimum}[\alpha(AB|CD)]$$

where

$$\alpha(AB|CD) = \text{maximum}[\{d(A,C) + d(B,D)\}, \{d(A,D) + d(B,C)\}] - [d(A,B) + d(C,D)]$$

and the minimum is taken over all A,B in U and C,D in V . The splits with positive isolation index can be represented by a network (1) (see Fig. 2 and 5). The network is a graphic portrayal of the metric in question. As it is an approximation, a fit index giving the percentage of the distances represented by the graph is also computed. A fit index close to 100% indicates true representation,

whereas lower fits indicate reduction in confidence. In general, if the distances supplied are tree-like, the split graph will produce a tree; more complex networks result as the distances deviate from this ideal situation. The advantage of using a split graph rather than a phylogenetic tree to represent the evolutionary relationships between a given set of taxa is that a tree always presumes the underlying evolutionary processes to be either bifurcating or multifurcating but a split graph does not. In the presence of recombination, for example, the data will show conflicting phylogenetic signals because different partitions may support different phylogenies. Split graphs allow more flexibility in representing such data since conflicting signals will not necessarily be forced onto a tree (see below and Fig. 2 for an example).

Split graphs were obtained for the concatenated *gag-pol-env* PLV sequences with the SplitsTree program version 3.0 (available at <http://www.mathematik.uni-bielefeld.de/~huson/phylogenetics/splitstree.html>). Distances were inferred with the GTR+ Γ +I model, the one best fitting the concatemer data set (see above), by using the parameters described in Results. As the measure of support, we calculated the bootstrap value for each edge in the split graph, i.e., the percentage of computed graphs out of 1,000 bootstrap replicates in which the split corresponding to that edge has occurred.

RESULTS

PLV concatemer evolutionary model. The nucleotide substitution model best fitting the PLV concatemer data set is the GTR+ Γ +I model described by the following maximum-like-

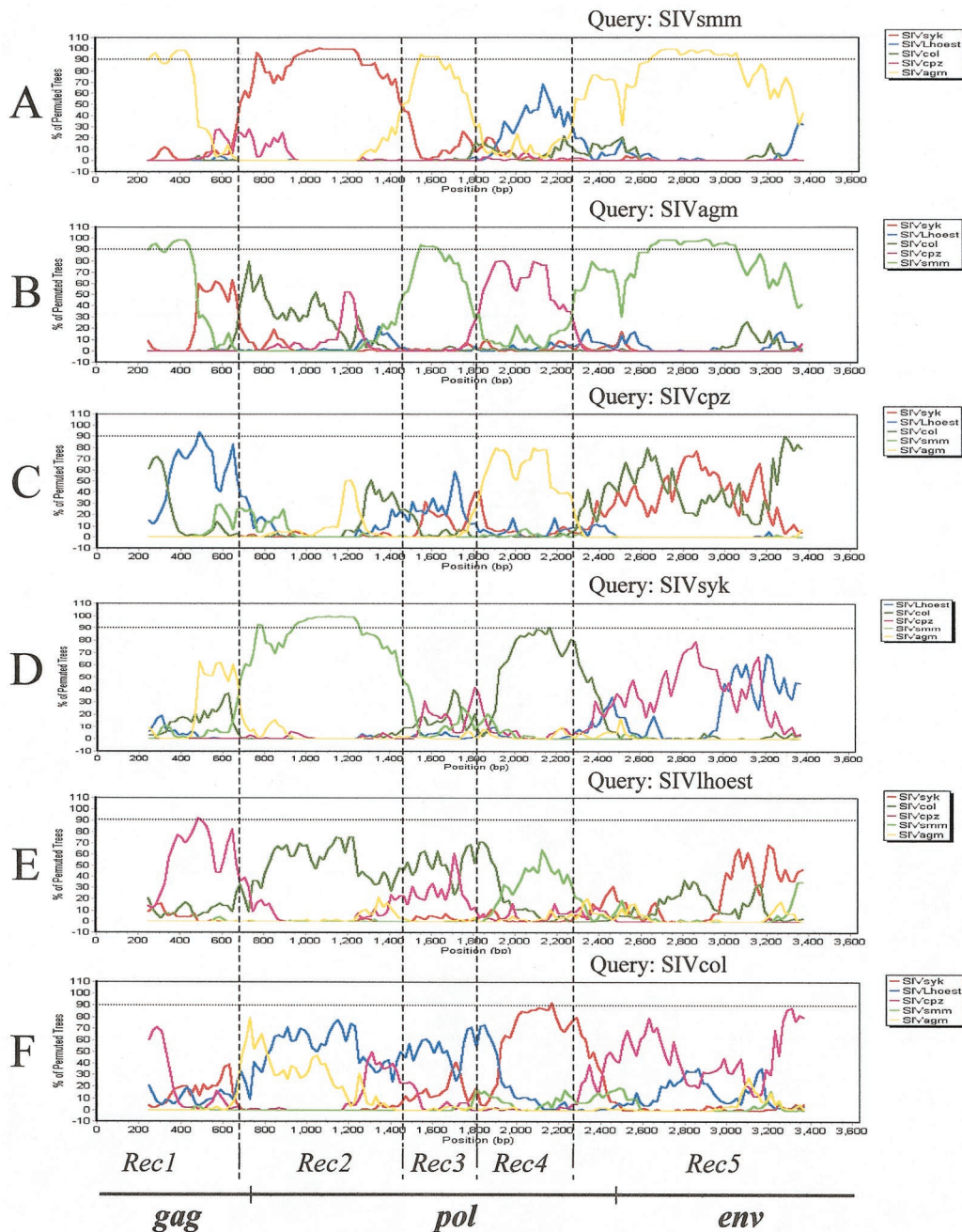


FIG. 3. Results of bootscannings with a sequence from each major PLV lineage as a query sequence against those of all the others. A concatenation with the nonoverlapping regions of *gag*, *pol*, and *env* was obtained. The bootscanning analysis was performed on first and second codon positions with a sliding window of 500 nt (20 nt/step) and 1,000 bootstrap replicates. Significant bootstrap values (>90%) are indicated in each bootscan plot by a horizontal broken line. Vertical broken lines represent attempts to locate the putative recombination break points common to all six PLV lineages. The resulting five fragments, which may have a monophyletic origin, indicated in the figure as Rec1 to Rec5, were chosen to be analyzed in separate phylogenetic analyses (see Tables 1, 2, and 3 and Fig. 4).

likelihood-estimated parameters: $r_{AC} = 4.02$, $r_{AG} = 4.05$, $r_{AT} = 1.79$, $r_{CG} = 3.21$, $r_{CT} = 5.53$, $P_{inv} = 0.19$, and $\alpha = 1.22$, where r_{ij} is the relative rate parameter for the $i \leftrightarrow j$ nucleotide substitution with respect to an r_{GT} of 1 by default, P_{inv} is the proportion of invariable sites, and α is the shape parameter of the discrete Γ distribution (eight rate categories). Simpler models (like the TVM+ Γ +I or the TN+ Γ +I model) compared with

the one above always performed significantly worse ($P < 0.001$).

Phylogenetic patterns in the PLV genome. The maximum-likelihood tree in Fig. 1 shows, as expected, the six major lineages of PLV to be approximately equidistant, with the SIVagm, SIVcpz, SIVsmm, and SIVlhoest clades significantly supported ($P < 0.001$). Results of bootscanning analyses are

TABLE 2. Phylogenetic support for the clustering of the six major PLV lineages in five putative recombinant fragments

Fragment ^a	Monophyletic clade ^b	Percentage of bootstrap/jackknife support for:		<i>P</i> value for ML ^c
		Neighbor-joining tree (1,000 replicates)	Fitch-Margoliash tree (1,000 replicates)	
PLV_REC1	(SIVhoest, SIVcol, SIVcpz)	81.2/80.2	90.8/90.2	<0.001
	(SIVsmm, SIVagm)	68.7/68.7	86.9/90.5	>0.1
PLV_REC2	(SIVsyk, SIVsmm)	89.8/87.9	89.0/88.5	<0.001
	(SIVcol, SIVhoest)	74.8/79.2	76.4/78.1	0.002
PLV_REC3	(SIVagm, [SIVcol, SIVhoest])	52.6/50.5	54.2/52.3	<0.001
	(SIVcol, SIVagm, SIVsmm)	<50/50	<50/50	0.007
PLV_REC4	(SIVcpz, SIVhoest)	<50/50	<50/50	0.029
	(SIVsyk, SIVcol)	88.3/88.1	88.4/88.0	0.01
PLV_REC5	(SIVcpz, SIVagm)	<50/50	<50/50	<0.001
	(SIVcol, SIVcpz)	56.9/56.3	55.9/57.9	0.02
	(SIVagm, SIVsmm)	98.3/98.0	98.5/98.1	<0.001
	(SIVhoest, [SIVagm, SIVsmm])	83.6/84.8	82.2/83.9	<0.001

^a Putative recombinant fragments inferred from the results of the bootscanning analyses (Fig. 3). See the text for more details.

^b Monophyletic clades are indicated between parentheses (Newick notation). Only clades significantly supported by at least one of the phylogenetic methods used are reported.

^c *P* value of the zero-branch-length test for the maximum-likelihood tree (*P* values of >0.05 are not significant).

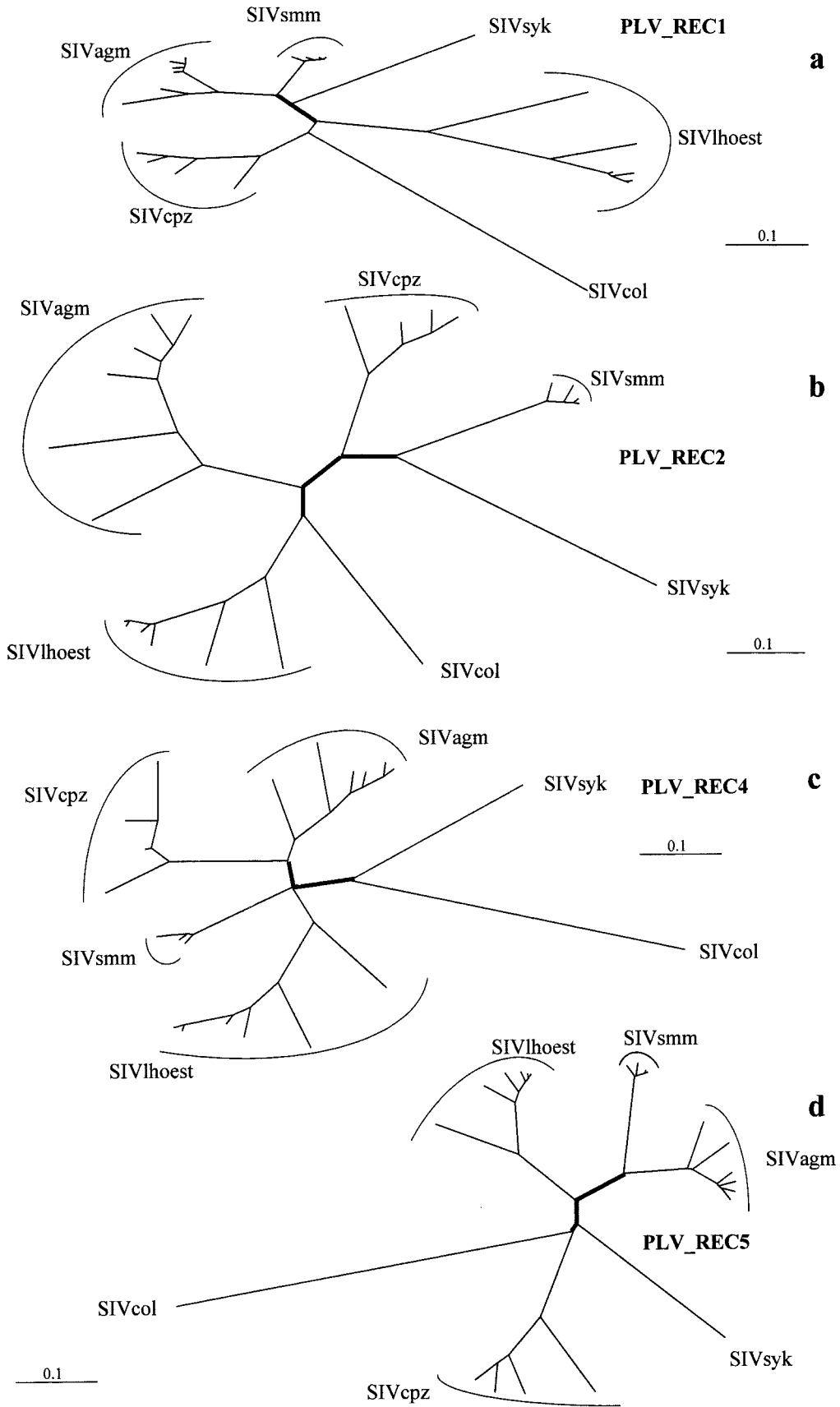
summarized in Fig. 2 and 3. In Fig. 2a, the bootscanning plot of SIVrcm, a previously reported recombinant virus (4), exhibits a typical mosaic pattern, with *gag* and part of *env* clustering with SIVsmm and *pol* clustering with SIVcpz, which is confirmed by the corresponding split graph in Fig. 2b. The bootscanning plots in Fig. 3, where each major lineage is compared with all the others, also show a number of conflicting phylogenies across the PLV genome. For example, from Fig. 3A (and B), it seems clear that SIVsmm and SIVagm are monophyletic in *gag*, in the central and 3'-end part of *pol*, and in *env* but not in the remaining genomic regions. On the other hand, Fig. 3A (and D) shows SIVsmm clustering with SIVsyk in the 5' part of *pol*. In a bootscanning plot, a clustering is considered significant when the percentage of permuted trees is at least 90% or greater (as in the examples discussed above). Thus, it is not possible to infer clear-cut phylogenetic relationships for all the different PLV genomic regions from the data in Fig. 3. However, bootscanning analysis alone is not sufficient evidence to support a hypothesis of extensive recombination across the PLV genome. Bootscanning plots are very sensitive to rate heterogeneity over sites, and the conflicting signals in Fig. 3 could be due to the low resolution of the algorithm in detecting and supporting the correct phylogenetic relationships within each sliding window. What can be deduced from the data in Fig. 3 is an indication of the genomic regions that are problematic in assessing phylogenetic relationships among the PLV major lineages, at least with the simple tools implemented in the SIMPLOT package. When different parts of the aligned viral genomes are analyzed, there seem to exist conflicting phylogenetic histories. Five such regions, flanked by the broken lines in Fig. 3, show relatively consistent patterns among all the major viral lineages. The fragments indicated by the broken lines only approximate the mosaic-like structure depicted by the figure, since recombination break points, as inferred by the bootscanning plots, appear to be at slightly different positions from lineage to lineage (Fig. 3E and F). The goal of this analysis is not to infer accurate recombination break points but to investigate whether or not different PLV genomic regions of the so-called pure lineages actually lead to

conflicting phylogenies. If any or all of these regions have different monophyletic origins, phylogenetic trees inferred with different algorithms from each region should consistently support, with a high level of confidence, different clustering patterns among the PLV lineages.

We subdivided the PLV genome into five putative recombinant fragments, indicated by the broken lines in Fig. 3, and analyzed them separately by standard phylogeny methods.

Evolutionary patterns and phylogeny of major PLV clades across the genome. Table 1 summarizes the amino acid positioning, relative to the inferred protein sequences of SIVcpzANT, which was used as a reference, of the five PLV genome fragments discussed above and named for convenience PLV_REC1 to PLV_REC5. According to the hierarchical LRT (see Materials and Methods), different fragments follow different nucleotide substitution models (Table 1). The shape parameter α of the Γ distribution ranges from 0.55 in PLV_REC1, which implies strong rate heterogeneity, to 1.37 in PLV_REC3, which indicates a relatively weak nucleotide substitution rate heterogeneity among sites. On the other hand, transition-transversion bias parameters are very similar for each fragment. Note also that the invariant models selected as the best-fitting ones for PLV_REC2, PLV_REC4, and PLV_REC5 show from 10 to 20% of the sites to be invariable, an indication that such sites are severely constrained by strong purifying selection.

Table 2 gives the phylogenetic support, in terms of bootstrap and jackknife values (distance-based trees) or *P* values (maximum-likelihood trees), for the clustering of the six major PLV lineages in the five genomic fragments investigated. Except for PLV_REC3, for which incongruent tree topologies were obtained and which will not be discussed further, all the tree-building algorithms consistently support, for each fragment, different phylogenetic relationships among the major PLV clades. Such relationships are clearly summarized by the neighbor-joining trees shown in Fig. 4. Maximum-likelihood and Fitch-Margoliash trees have the same topologies. The trees were obtained by including only the SIV strains, but similar results were obtained when HIV-1 strains clustering within the



SIVcpz clade and HIV-2 strains clustering within the SIVsmm clade were also included (data not shown).

The SIVlhoest group clusters with the SIVcol and SIVcpz lineages in PLV_REC1, with SIVcol but not with SIVcpz in PLV_REC2, and with SIVsyk in PLV_REC4, and it is monophyletic with SIVsmm and SIVagm in PLV_REC5 (Table 2 and Fig. 4). In turn, SIVcol shares a more recent common ancestor with SIVsyk in PLV_REC4 but with SIVcpz in PLV_REC5. SIVsyk also clusters with SIVsmm (see below) in PLV_REC2. The SIVagm lineage clusters with SIVsmm in PLV_REC1, PLV_REC3, and PLV_REC5 and with SIVcpz in PLV_REC4 (significant support in the maximum-likelihood tree only [Table 2]) and shares in the PLV_REC2 fragment a common ancestor with the SIVcol-SIVlhoest cluster. The SIVsmm group, which includes HIV-2, clusters with SIVagm in PLV_REC1 and PLV_REC5 and with SIVsyk in PLV_REC2. A monophyletic origin of SIVsmm, SIVcol, and SIVagm in the PLV_REC3 fragment ($P < 0.007$) is supported by the maximum-likelihood tree but not by results from distance-based methods (Table 2). The clustering of SIVsmm in the PLV_REC4 fragment appears to be different depending on the tree-building method used, with low support (bootstrap values of $<50\%$ and P values of >0.1 ; data not shown). The SIVcpz lineage shares a significantly supported polytomy with SIVlhoest and SIVcol in PLV_REC1. The phylogenetic position of the SIVcpz clade in the second fragment is different depending on the tree-building method used, with low support (bootstrap values of $<50\%$ and P values of >0.1 ; data not shown). The maximum-likelihood tree (but not distance-based trees) also supports the clustering of SIVcpz with SIVlhoest in PLV_REC3 and with SIVagm in PLV_REC4. Finally, a SIVcpz-SIVcol monophyletic origin of the PLV_REC5 fragment is weakly supported by all methods (Table 2).

Each putative PLV recombinant fragment, except for PLV_REC1 and PLV_REC4, seems to evolve according to a different evolutionary model. All the models reported in Table 1 are simpler variants of the one best fitting the PLV concatenate (see above). Since the nucleotide substitution models considered are nested, it makes sense that when different genomic regions with different evolutionary dynamics are pooled together the most general model, including the simpler ones as a special case, is the one best fitting the data. However, a simpler model may not be rejected with a smaller data set due to the loss of statistical power of the LRT. In other words, with short gene regions, it might not be possible to reject a simple model, not because the simple model is indeed correct but because the data do not provide enough information and the test lacks power. To address this problem, we investigated 100 simulated recombinant data sets, obtained with the GTR+ Γ +I model, by performing a separate LRT for each recombinant fragment in each simulated set (see Materials and Methods). The simulated alignments were generated with four nucleotide parti-

TABLE 3. Results of topological tests for the possible trees of each PLV putative recombinant fragment

Data set ^a	Examined tree ^b	Log likelihood ^c	P value of the S-H test ^d
PLV_REC1	Tree_REC1	6,596.5*	
	Tree_REC2	6,636.7	0.005
	Tree_REC4	6,641.0	0.007
	Tree_REC5	6,626.7	0.029
PLV_REC2	Tree_REC1	8,558.7	0.001
	Tree_REC2	8,510.6*	
	Tree_REC4	8,545.6	0.002
	Tree_REC5	8,567.8	<0.001
PLV_REC4	Tree_REC1	4,352.6	0.02
	Tree_REC2	4,330.7	0.41
	Tree_REC4	4,320.7*	
	Tree_REC5	4,361.5	0.005
PLV_REC5	Tree_REC1	16,143.1	<0.001
	Tree_REC2	16,188.1	<0.001
	Tree_REC4	16,188.8	<0.001
	Tree_REC5	16,073.3*	

^a Putative recombinant fragments inferred from the results of the bootscanning analyses shown in Fig. 3 (see the text for more details).

^b Topologies of Tree_REC1, Tree_REC2, Tree_REC4, and Tree_REC5 are reported in Fig. 4a, b, c, and d, respectively.

^c The nucleotide substitution parameters and the log likelihood of each tree have been reestimated via maximum likelihood by using the best-fitting nucleotide substitution model for each data set according to the data in Table 1. The best tree is statistically compared to the other trees and is indicated by an asterisk.

^d The S-H test was done with RELL bootstraps (1,000 replicates).

tions of the same lengths as PLV_REC1, PLV_REC2, PLV_REC4, and PLV_REC5, each partition following the different phylogenetic tree inferred for the corresponding fragment (Fig. 4). In general, the TVM+ Γ +I model could not be rejected in 53% of the cases, and as expected, failure to reject such a model was greater for the shortest fragment (70%) than for the longest one (29%). On the other hand, simpler models, such as TN+ Γ +I and HKY+ Γ +I, were rejected for 100% of the simulated data sets. Therefore, shorter fragments do reduce the power of the LRT, and the test is very likely to fail in rejecting the TVM+ Γ +I model, which constrains only one of the parameters of the GTR+ Γ +I model (the true one for the simulated data), but not models imposing heavier constraints, like TN+ Γ +I and HKY+ Γ +I. By comparison of this result with the data in Table 1, it seems reasonable to suggest different evolutionary patterns in the four PLV recombinant fragments investigated, although the rejection of the GTR+ Γ +I model for PLV_REC2 could be due to the lack of power of the test.

Each of the four trees in Fig. 4 is the optimal tree for the corresponding genome region according to both distance- and maximum-likelihood-based criteria. It may be asked whether for each putative recombinant fragment the other trees are significantly worse than the optimal tree. Table 3 shows the

FIG. 4. Neighbor-joining phylogenetic trees of the putative PLV recombinant fragments inferred from the bootscanning plots in Fig. 3. Genetic distances for each fragment were estimated by using the best-fitting substitution model given in Table 1 and the first and second codon positions. Only the names of the major PLV lineages are shown. Thicker internal branches are significantly supported by bootstrap and jackknife values and/or the results of the zero-branch-length test, according to the data in Table 2. Edges are drawn to scale, with the bar indicating 0.1 nt replacements per site. (a) PLV_REC1; (b) PLV_REC2; (c) PLV_REC4; (d) PLV_REC5.

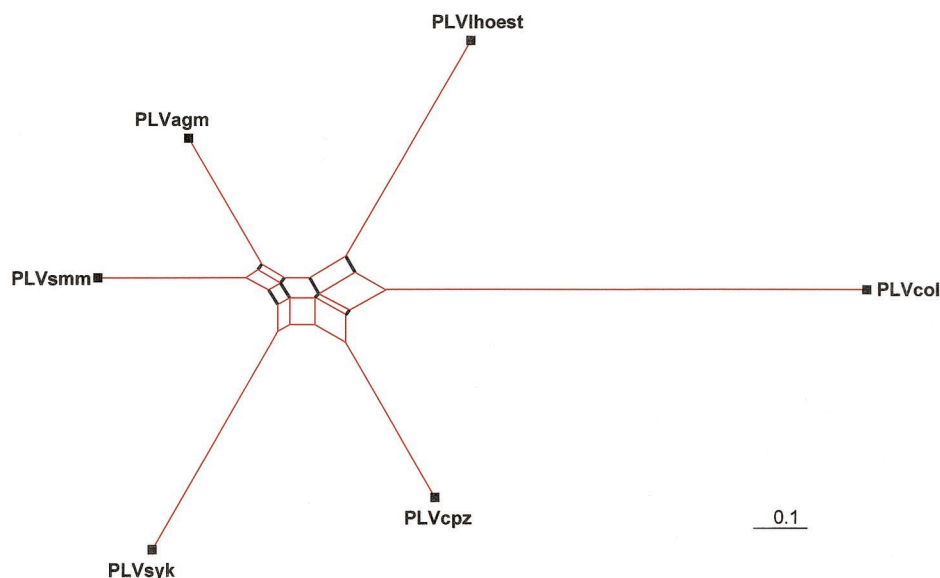


FIG. 5. Results of split-decomposition analysis of the six major PLV lineages. A concatenation with the nonoverlapping regions of *gag*, *pol*, and *env* was obtained. The analysis was performed using only first and second codon positions, with distances inferred by using the GTR+ Γ +I model of nucleotide substitution (see Results). One thousand bootstrap replicates were generated to assess the reliability of each edge in the split graph. Names at the tips of the split graph represent the monophyletic group of each major lineage rather than a particular isolate. Edges are drawn to scale, with the bar indicating 0.1 nt replacements per site. Edges in red have >85% bootstrap support.

results of the S-H RELL test for each fragment. The likelihood of the optimal tree is always significantly better than those of all the others, except for the PLV_REC4 data set for which the likelihood of tree number 2 (the tree with the topology given in Fig. 4b) is not significantly different from the likelihood of the optimal tree as shown in Fig. 4d. Overall, the results in Table 3 confirm that in each of the putative recombinant fragments the inferred optimal tree describes the phylogenetic relationships among the major PLV lineages significantly better than the alternative trees.

The molecular clock hypothesis was tested for the trees in Fig. 4 and could not be rejected for PLV_REC1 (likelihood ratio statistic, 29.6; $P = 0.077$), PLV_REC2 (likelihood ratio statistic, 35.26; $P = 0.019$), and PLV_REC4 (likelihood ratio statistic, 35; $P = 0.020$), but it was rejected for the tree obtained from PLV_REC5 (likelihood ratio statistic, 99.8; $P = 1.4 \times 10^{-12}$). In order to investigate the hypothesis of cospeciation of SIVs and their hosts, we compared, in the clock-like trees, the lengths of the branches connecting the tip to the ancestor (most recent common ancestor) of each clade. We found that the length of the branch connecting the tip with the ancestor of the SIVagm clade is 0.8 to 1.7 times longer than the one leading from the tip to the ancestor of the SIVcpz clade.

Split decomposition analysis. Computer simulations show that the algorithm implemented by SplitsTree becomes unreliable when more than 15 to 20 taxa are analyzed, and with a larger number of strains it tends to give a star-like graph with little or no resolution (data not shown). However, it is possible to group taxa belonging to a monophyletic clade and treat them as a single taxonomic unit. Since the monophyly of each major PLV clade is unequivocal (Fig. 1), the group option in SplitsTree allowed us to analyze six lineages rather than all the

strains in the data set. In practice, the algorithm computes the average distance from each group to any other and, by canonical decomposition of such group distances, infers a split graph representing the relationships among the different groups. The resulting split graph for the PLV *gag-pol-env* concatenation, using only first and second codon positions, is shown in Fig. 5. Note that each branch leading to a major PLV clade is much longer than the internal splits, which are very close to the center of the graph. Even though reticulation in split graphs does not necessarily imply recombination and may be due, for example, to insufficient correction for superimposed mutations (45), Fig. 5 confirms that the data contain conflicting phylogenetic signals and are consistent with early recombination events among the major lineages as shown by the phylogenetic analyses discussed above. The fit index is 98.1, an excellent index meaning that only 1.9% of the distances in the distance matrix are not represented by the graph. Most of the internal splits have a bootstrap support between 85 and 100%. Overall, the graph seems to reliably represent the data, showing the six major PLV clades related by a complex web-like network, as is also evident from the results summarized in Tables 2 and 3 and Fig. 4. By comparison of the split graph with Fig. 4, it is possible to see that the internal splits with high bootstrap support correspond to the topology of each tree in the figure. For example, the horizontal central split partitions the taxa into the two monophyletic subsets PLVhoest-PLVcol-PLVcpz and PLVagm-PLVsmm-PLVsyk, according to the tree in Fig. 4a. On the other hand, the vertical central split and the split on the lower right part of the graph partition the taxa into the two monophyletic subsets PLVsyk-PLVcpz-PLVcol and PLVsmm-PLVagm-PLVhoest, according to the tree in Fig. 4d. The intuitive interpretation of Fig. 5 is that each internal split represents a conflicting phylogeny that would be resolved ar-

bitrarily in a single (wrong) tree by using any of the other tree-building algorithms available.

DISCUSSION

Investigating the origin and understanding the evolution of the PLVs require, first of all, the knowledge of the exact phylogenetic relationships among the PLV clades. The result of this study is that there are no pure lineages in addition to a number of already-established SIV recombinant strains. In fact, PLV isolates from any simian species and any geographic location show complex mosaic genomes: pure lineages as such do not exist. The only difference between the six equidistant major lineages and an acknowledged recombinant, such as SIV_{rcm}, is the time frame of the recombination events. In the case of the major PLV lineages, the events that led to the mosaic genome of each lineage occurred very close to the root of the PLV tree, i.e., very early in time. Subsequently, some of the early recombinants became endemic in a particular geographic location or within a particular species and started to diverge consistently from all the others, leading to the major PLV clades known so far. The results of all the analyses performed in the present study indicate a complex network of phylogenetic relationships among the PLV lineages, summarized by the split graph in Fig. 5. However, it is important to realize that, as stated in Results, the exact recombination break points vary from lineage to lineage and the bootscanning analysis is not powerful enough to infer them unambiguously. Bootscanning has been used in the present work as an exploratory tool to infer possible regions across the PLV genome with conflicting phylogenetic histories. Alternative approaches based on maximum likelihood and Monte Carlo simulations seem to be impracticable for more than a few sequences (data not shown), but they may be promising for the future. In a recent simulation study, Posada and Crandall (32) showed that methods such as bootscanning, based on phylogenetic reconstruction, may fail to detect recombination when recombination events are rare and nucleotide diversity in the data set is low. However, the number of false positives inferred by phylogenetic-based methods in the case of no recombination is around the expectation of 5% (30). The separate phylogenetic analyses carried out on the different genomic regions discussed in Results reveal conflicting phylogenetic relationships from one region to another among the major PLV lineages. Different tree-building algorithms and a number of statistical tests consistently support the same conclusion for at least four regions in the PLV genome. To be conservative, we can say that it is possible that more recombination events along the PLV genome have happened but certainly not fewer than the ones investigated in the present paper. A covarion model, assuming that genetic changes in one region are covarying with mutations seen in a different region, may generate in theory patterns similar to those generated by recombination. However, such a theory does not look very parsimonious since too many of such potential covarying regions and too-extensive regions (too many ad hoc hypotheses) would need to be assumed. In conclusion, it seems well founded to state that the so-called pure PLV lineages are in fact mosaics.

Our results also suggest that split decomposition methods could be employed to investigate the evolution of PLVs. They

have the advantage of not forcing the data onto the wrong tree but the disadvantage of being less intuitively understood when multiple conflicting signals in the data produce complex networks, like the ones discussed in this paper.

Although clear-cut examples exist of simian-to-simian and simian-to-human cross-species transmissions, it is generally accepted that host-specific virus evolution of PLVs is the rule (18). However, as noticed by Sharp and coworkers (40), the cospeciation hypothesis bears a paradox. Apes and Old World monkeys originated at least 25 million years ago (C. Mulder, Letter, *Nature* **333**:396, 1988), and African green monkeys originated probably around 1,000,000 years ago (25), suggesting an old time scale for the PLV tree. HIV-1 and HIV-2 share the same common ancestor with all other PLVs, but the zoonosis at the origin of their subtypes appears to have occurred much more recently. For example, the ancestor of group M has been dated to the 1920s to 1930s (22, 36), and the interspecies transmissions giving rise to HIV-1 and HIV-2 all occurred somewhere between the end of the 17th and the beginning of the 20th centuries (24a, 36). The finding should not come as a surprise considering the fast evolutionary rate of these viruses, estimated around 10^{-3} nt substitutions per site per year (26, 40, 44). As discussed in Results, the molecular clock hypothesis for the putative PLV recombinant fragments 1, 2, and 4 could not be rejected. In the clock-like trees, the length of the branch connecting the tip with the ancestor of the SIV_{agm} clade is 0.8 to 1.7 times longer than the one leading from the tip to the ancestor of the SIV_{cpz} clade (see Results). This is incompatible with the cospeciation hypothesis, which would require a branch length about 1,000 times longer to fit the dramatically different time scales between the separation of HIV-1 and SIV_{cpz}, dated a few hundred years ago (36), and the origin of the African green monkeys, which occurred around 1,000,000 years ago (26, 40). Moreover, even though the molecular clock hypothesis has been shown in some cases to fit poorly the evolutionary patterns of PLVs (23), rate differences among PLV strains are unlikely to cover such an order of magnitude. In fact, evidence exists of the opposite. For example, the evolutionary rate of SIV *in vivo* has been roughly estimated at around 10^{-2} nt substitutions per site per year in SIV_{agm} (29) and around 6×10^{-3} nt substitutions per site per year in SIV_{smm} (34). As shown above, for at least four of the five putative recombinant fragments across the PLV genome, a molecular clock could not be rejected. Besides molecular clock considerations, jungle graphs have shown that the relatively high degree of cophylogenetic match observed in PLV evolution does not necessarily indicate cospeciation but could in fact have resulted from preferential host interspecies transmission events between more closely related host species (6). The recombinant origin of each PLV lineage discussed in the present paper makes even more unlikely long independent cospeciation of PLV lineages and their hosts. On the contrary, it seems that cross-species transmission events leading to highly complex mosaic genomes have occurred continuously throughout the entire PLV evolutionary history. A few recent studies try to address how this mode of evolution may affect dating strategies (32, 38, 39). Simulations show that the higher the recombination rate in the sequences, the higher the probability that the molecular clock will falsely be rejected (39). Thus, the fact that for three of the four PLV

genome fragments the molecular clock hypothesis cannot be rejected seems to be indirect evidence that within these recombinant fragments further recombination is unlikely.

In conclusion, the findings presented in this paper provide compelling evidence that (i) simian-to-simian transmission of SIV, combined with frequent recombination between diverse lineages of SIV, has played a major role in the evolution of the PLVs; (ii) these recombinogenic events have been ongoing since the beginning of the evolutionary history of PLVs; (iii) as a result of these early events, there are no longer any pure lineages of SIV—instead, most, if not all, contemporary strains of SIV are complex mosaic viruses—and (iv) these mosaic viruses are the ancestral strains of the HIV causing the current HIV infection and AIDS pandemic.

ACKNOWLEDGMENTS

This work was supported by the Flemish Funds voor Wetenschappelijk Onderzoek (FWO grants G.0288.01 and KAN2002 1.5.193.02 and Postdoctoraal Onderzoeker contract 530).

We thank Martine Peeters, Stuart Ray, and Walter Fitch for critical reading of the manuscript and helpful suggestions. We also thank Ziheng Yang and two of the anonymous reviewers for valuable critiques and suggestions.

REFERENCES

- Bandelt, H.-J., and A. Dress. 1992. A canonical decomposition theory for metrics on a finite set. *Adv. Mathematics* **9**:47–105.
- Beer, B. E., E. Bailes, R. Goeken, G. Dapolito, C. Coulibaly, S. G. Norley, R. Kurth, J.-P. Gautier, A. Gautier-Hion, D. Vallet, P. M. Sharp, and V. M. Hirsch. 1999. Simian immunodeficiency virus (SIV) from sun-tailed monkeys (*Cercopithecus solatus*): evidence for host-dependent evolution of SIV within the *C. lhoesti* superspecies. *J. Virol.* **73**:7734–7744.
- Beer, B. E., E. Bailes, G. Dapolito, B. J. Campbell, R. M. Goeken, M. K. Axthelm, P. D. Markham, J. Bernard, D. Zagury, G. Franchini, P. M. Sharp, and V. M. Hirsch. 2000. Patterns of genomic diversity among their simian immunodeficiency viruses suggest that L'Hoest monkeys (*Cercopithecus lhoesti*) are a natural lentivirus reservoir. *J. Virol.* **74**:3892–3898.
- Beer, B. E., B. T. Foley, C. L. Kuiken, Z. Tooze, R. M. Goeken, C. R. Brown, J. Hu, M. St. Claire, B. T. Korber, and V. M. Hirsch. 2001. Characterization of novel simian immunodeficiency viruses from red-capped mangabeys from Nigeria (SIVrcmNG409 and -NG411). *J. Virol.* **75**:12014–12027.
- Chakrabarti, L., M. Guyader, M. Alizon, M. D. Daniel, R. C. Desrosiers, P. Tiollais, and P. Sonigo. 1987. Sequence of simian immunodeficiency virus from macaque and its relationship to other human and simian retroviruses. *Nature* **328**:543–547.
- Charleston, M. A., and D. L. Robertson. 2002. Preferential host switching by primate lentiviruses can account for phylogenetic similarity with the primate phylogeny. *Syst. Biol.* **51**:528–535.
- Chen, Z., P. Telfer, A. Gettie, P. Reed, L. Zhang, D. D. Ho, and P. A. Marx. 1996. Genetic characterization of new West African simian immunodeficiency virus SIVsm: geographic clustering of household-derived SIV strains with human immunodeficiency virus type 2 subtypes and genetically diverse viruses from a single feral sooty mangabey troop. *J. Virol.* **70**:3617–3627.
- Chen, Z., A. Luckay, D. L. Sodora, P. Telfer, P. Reed, A. Gettie, J. M. Kanu, R. F. Sadek, J. Yee, D. D. Ho, L. Zhang, and P. A. Marx. 1997. Human immunodeficiency virus type 2 (HIV-2) seroprevalence and characterization of a distinct HIV-2 genetic subtype from the natural range of simian immunodeficiency virus-infected sooty mangabeys. *J. Virol.* **71**:3953–3960.
- Courgnaud, V., X. Pourrut, F. Bibollet-Ruche, E. Mpoudi-Ngole, A. Bourgeois, E. Delaporte, and M. Peeters. 2001. Characterization of a novel simian immunodeficiency virus from guereza colobus monkeys (*Colobus guereza*) in Cameroon: a new lineage in the nonhuman primate lentivirus family. *J. Virol.* **75**:857–866.
- Courgnaud, V., M. Salemi, X. Pourrut, E. Mpoudi-Ngole, B. Abela, P. Auzel, F. Bibollet-Ruche, B. Hahn, A.-M. Vandamme, E. Delaporte, and M. Peeters. 2002. Characterization of a novel simian immunodeficiency virus with a *vpu* gene from greater spot-nosed monkeys (*Cercopithecus nictitans*) provides new insights into the simian/human immunodeficiency virus phylogeny. *J. Virol.* **76**:8298–8309.
- Felsenstein, J. 1981. Evolutionary trees from DNA sequences: a maximum likelihood approach. *J. Mol. Evol.* **17**:368–376.
- Fukasawa, M., T. Miura, A. Hasegawa, S. Morikawa, H. Tsujimoto, K. Miki, T. Kitamura, and M. Hayami. 1988. Sequence of simian immunodeficiency virus from African green monkey, a new member of the HIV/SIV group. *Nature* **333**:457–461.
- Gao, F., E. Bailes, D. L. Robertson, Y. Chen, C. M. Rodenburg, S. F. Michael, L. B. Cummins, L. O. Arthur, M. Peeters, G. M. Shaw, P. M. Sharp, and B. H. Hahn. 1999. Origin of HIV-1 in the chimpanzee *Pan troglodytes troglodytes*. *Nature* **397**:436–441.
- Gao, F., L. Yue, A. T. White, P. G. Pappas, J. Barchue, A. P. Hanson, B. M. Greene, P. M. Sharp, G. M. Shaw, and B. H. Hahn. 1992. Human infection by genetically diverse SIVsm-related HIV-2 in west Africa. *Nature* **358**:495–499.
- Gao, F., L. Yue, D. L. Robertson, S. C. Hill, H. Hui, R. J. Biggar, A. E. Neequaye, T. M. Whelan, D. D. Ho, and G. M. Shaw. 1994. Genetic diversity of human immunodeficiency virus type 2: evidence for distinct sequence subtypes with differences in virus biology. *J. Virol.* **68**:7433–7447.
- Gao, F., L. Yue, D. L. Robertson, C. D. Carruthers, S. G. Morrison, B. Jian, Y. Chen, F. Barre-Sinoussi, M. Girard, A. Srinivasan, A. G. Abimik, 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.
- Georges-Courbot, M. C., C. Y. Lu, M. Makuwa, P. Telfer, R. Onanga, G. Dubreuil, Z. Chen, S. M. Smith, A. Georges, F. Gao, B. H. Hahn, and P. A. Marx. 1998. Natural infection of a household pet red-capped mangabey (*Cercocebus torquatus torquatus*) with a new simian immunodeficiency virus. *J. Virol.* **72**:600–608.
- Hahn, B. H., G. M. Shaw, K. M. De Cock, and P. M. Sharp. 2000. AIDS as a zoonosis: scientific and public health implications. *Science* **287**:607–614.
- Hirsch, V. M., R. A. Olmsted, M. Murphey-Corb, R. H. Purcell, and P. R. Johnson. 1989. An African primate lentivirus (SIVsm) closely related to HIV-2. *Nature* **339**:389–392.
- Hirsch, V. M., B. J. Campbell, E. Bailes, R. Goeken, C. Brown, W. R. Elkins, M. Axthelm, M. Murphey-Corb, and P. M. Sharp. 1999. Characterization of a novel simian immunodeficiency virus (SIV) from L'Hoest monkeys (*Cercopithecus lhoesti*): implications for the origins of SIVmnd and other primate lentiviruses. *J. Virol.* **73**:1036–1045.
- Jin, M. J., H. Hui, D. L. Robertson, M. C. Muller, F. Barre-Sinoussi, V. M. Hirsch, J. S. Allan, G. M. Shaw, P. M. Sharp, and B. H. Hahn. 1994. Mosaic genome structure of simian immunodeficiency virus from west African green monkeys. *EMBO J.* **13**:2935–2947.
- Korber, B., M. Muldoon, J. Theiler, F. Gao, R. Gupta, A. Lapedes, B. H. Hahn, S. Wolinsky, and T. Bhattacharya. 2000. Timing the ancestor of the HIV-1 pandemic strains. *Science* **288**:1789–1796.
- Korber, B. T. M., J. Theiler, and S. Wolinsky. 1998. Limitations of a molecular clock applied to considerations of the origin of HIV-1. *Science* **280**:1868–1871.
- Leitner, T., S. Kumar, and J. Albert. 1997. Tempo and mode of nucleotide substitutions in *gag* and *env* gene fragments in human immunodeficiency virus type 1 populations with a known transmission history. *J. Virol.* **71**:4761–4770.
- Lemey, P., O. G. Pybus, B. Wang, N. K. Saksena, M. Salemi, and A.-M. Vandamme. Tracing the origin and history of HIV-2 epidemics. *Proc. Natl. Acad. Sci. USA*, in press.
- Li, W.-H., M. Tanimura, and P. M. Sharp. 1987. An evaluation of the molecular clock hypothesis using mammalian DNA sequences. *J. Mol. Evol.* **25**:330–342.
- Li, W.-H., M. Tanimura, and P. M. Sharp. 1988. Rates and dates of divergence among AIDS virus nucleotide sequences. *Mol. Biol. Evol.* **5**:313–330.
- McCutchan, F. E., P. A. Hegerich, T. P. Brennan, P. Phanuphak, P. Singharaj, A. Jugsudee, P. W. Berman, A. M. Gray, A. K. Fowler, and D. S. Burke. 1992. Genetic variants of HIV-1 in Thailand. *AIDS Res. Hum. Retrovir.* **8**:1887–1895.
- Muller, M. C., N. K. Saksena, E. Nerrienet, C. Chappey, V. M. Herve, J. P. Durand, P. Legal-Campodonic, M. C. Lang, J. P. Digoutte, and A. J. Georges. 1993. Simian immunodeficiency viruses from central and western Africa: evidence for a new species-specific lentivirus in tantalus monkeys. *J. Virol.* **67**:1227–1235.
- Muller-Trutwin, M. C., S. Corbet, M. Dias Tavares, V. M. A. Hervé, E. Nerrienet, M. C. Georges-Courbot, W. Saurin, P. Sonigo, and F. Barré-Sinoussi. 1996. The evolutionary rate of non-pathogenic simian immunodeficiency virus (SIVagm) is in agreement with a rapid continuous replication in vivo. *Virology* **223**:89–102.
- Posada, D. 2001. Unveiling the molecular clock in the presence of recombination. *Mol. Biol. Evol.* **18**:1976–1978.
- Posada, D., and K. A. Crandall. 1998. MODELTEST: testing the model of DNA substitution. *Bioinformatics* **14**:817–818.
- Posada, D., and K. A. Crandall. 2001. Evaluation of methods for detecting recombination from DNA sequences: computer simulations. *Proc. Natl. Acad. Sci. USA* **98**:13757–13762.
- Rambaut, A., and N. C. Grassly. 1997. Seq-Gen: an application for the Monte Carlo simulation of DNA sequence evolution along phylogenetic trees. *Comput. Appl. Biosci.* **13**:235–238.
- Rey-Cuille, M. A., J. L. Berthier, M. C. Bomsel-Demontoy, Y. Chaduc, L. Montagnier, A. G. Hovanessian, and L. A. Chakrabarti. 1998. Simian immunodeficiency virus replicates to high levels in sooty mangabeys without inducing disease. *J. Virol.* **72**:3872–3886.

35. **Robertson, D. L., P. M. Sharp, F. E. McCutchan, and B. H. Hahn.** 1995. Recombination in HIV-1. *Nature* **374**:124–126.
36. **Salemi, M., K. Strimmer, W. W. Hall, M. Duffy, E. Delaporte, S. Mboup, M. Peeters, and A.-M. Vandamme.** 2001. Dating the common ancestor of SIVcpz and HIV-1 group M and the origin of HIV-1 subtypes using a new method to uncover clock-like molecular evolution. *FASEB J.* **15**:276–278.
37. **Salminen, M. O., J. K. Carr, D. L. Robertson, P. Hegerich, D. Gotte, C. Koch, E. Sanders-Buell, F. Gao, P. M. Sharp, B. H. Hahn, D. S. Burke, and F. E. McCutchan.** 1997. Evolution and probable transmission of intersubtype recombinant human immunodeficiency virus type 1 in a Zambian couple. *J. Virol.* **71**:2647–2655.
38. **Schierup, M. H., and J. Hein.** 2000. Consequences of recombination on traditional phylogenetic analysis. *Genetics* **156**:879–891.
39. **Schierup, M. H., and J. Hein.** 2000. Recombination and the molecular clock. *Mol. Biol. Evol.* **17**:1578–1579.
40. **Sharp, P. M., E. Bailes, D. L. Robertson, F. Gao, and B. H. Hahn.** 2000. Origins and evolution of AIDS viruses. *Biol. Bull.* **196**:338–342.
41. **Shimodaira, H., and M. Hasegawa.** 1999. Multiple comparisons of log-likelihoods with applications to phylogenetic inference. *Mol. Biol. Evol.* **16**:1114–1116.
42. **Souquiere, S., F. Bibollet-Ruche, D. L. Robertson, M. Makuwa, C. Apetrei, R. Onanga, C. Kornfeld, J. C. Plantier, F. Gao, K. Abernethy, et al.** 2001. Variability of human immunodeficiency virus type 2 (HIV-2) infecting patients living in France. *Virology* **280**:19–30.
43. **Sullivan, J., K. E. Holsinger, and C. Simon.** 1996. The effect of topology on estimates of among-site rate variation. *J. Mol. Evol.* **42**:308–312.
44. **Suzuki, Y., Y. Yamaguchi-Kabata, and T. Gojobori.** 2000. Nucleotide substitution rates of HIV-1. *AIDS Rev.* **2**:39–47.
45. **Swofford, D. L., G. J. Olsen, P. J. Waddell, and D. M. Hillis.** 1996. Phylogenetic inference, p. 407–514. *In* D. M. Hillis, C. Moritz, and B. K. Mable (ed.), *Molecular systematics*, 2nd ed. Sinauer Associates Inc., Sunderland, Mass.
46. **Swofford, D. L., and J. Sullivan.** Phylogeny. Inference based on parsimony and other methods using PAUP*. *In* M. Salemi and A.-M. Vandamme (ed.), *The phylogenetic handbook: a practical approach to DNA and protein phylogeny*, in press. Cambridge University Press, New York, N.Y.
47. **van der Kuyl, A. C., C. L. Kuiken, J. T. Dekker, and J. Goudsmit.** 1995. Phylogeny of African monkeys based upon mitochondrial 12S rRNA sequences. *J. Mol. Evol.* **40**:173–180.
48. **Xia, X.** 2000. *Data analysis in molecular biology and evolution*. Kluwer Academic Publishers, Boston, Mass.
49. **Xia, X., Z. Xie, M. Salemi, L. Chen, and Y. Wang.** 2003. An index of substitution saturation and its application. *Mol. Phylogenet. Evol.* **26**:1–7.
50. **Yang, Z.** 1994. Maximum likelihood phylogenetic estimation from DNA sequences with variable rates over sites: approximate methods. *J. Mol. Evol.* **39**:306–314.