Analysis of the Evolutionary Relationships of HIV-1 and SIVcpz Sequences Using Bayesian Inference: Implications for the Origin of HIV-1

D. Paraskevis, P. Lemey,* M. Salemi,* M. Suchard,† Y. Van de Peer,‡ and A.-M. Vandamme**

*Rega Institute for Medical Research, Katholieke Universiteit Leuven, Belgium; †Department of Biomathematics, David Geffen School of Medicine, University of California, Los Angeles; ‡Department of Plant Systems Biology, Flanders Interuniversity Institute for Biotechnology, Gent University, Belgium

The most plausible origin of HIV-1 group M is an SIV lineage currently represented by SIVcpz isolated from the chimpanzee subspecies *Pan troglodytes troglodytes*. The origin of HIV-1 group O is less clear. Putative recombination between any of the HIV-1 and SIVcpz sequences was tested using bootscanning and Bayesian-scanning plots, as well as a new method using a Bayesian multiple change-point (BMCP) model to infer parental sequences and crossing-over points. We found that in the case of highly divergent sequences, such as HIV-1/SIVcpz, Bayesian scanning and BMCP methods are more appropriate than bootscanning analysis to investigate spatial phylogenetic variation, including estimating the boundaries of the regions with discordant evolutionary relationships and the levels of support of the phylogenetic clusters under study. According to the Bayesian scanning plots and BMCP method, there was strong evidence for discordant phylogenetic clustering throughout the genome: (1) HIV-1 group O clustered with SIVcpzANT/TAN in middle *pol*, and partial *vif/env*; (2) SIVcpzGab1 clustered with SIVcpzANT/TAN in 3′*pol/vif*, and middle *env*; (3) HIV-1 group O grouped with SIVcpzCamUS and SIVcpzGab1 in p17/p24; (4) HIV-1 group M was more closely related to SIVcpzCamUS in 3′*gag/pol* and in middle *pol*, whereas in partial gp120 group M clustered with group O. Conditionally independent phylogenetic analysis inferred by maximum likelihood (ML) and Bayesian methods further confirmed these findings. The discordant phylogenetic relationships between the HIV-1/SIVcpz sequences may have been caused by ancient recombination events, but they are also due, at least in part, to altered rates of evolution between parental SIVcpz lineages.

Introduction

Human immunodeficiency viruses types 1 and 2 (HIV-1 and HIV-2) have been characterized as the causative agents of AIDS, and their most plausible origins are ancestral simian viruses, currently known as SIVcpz for HIV-1 and SIVsm for HIV-2, infecting chimpanzees and sooty mangabeys, respectively (Gao et al. 1999; Corbet et al. 2000; Hahn et al. 2000; Sharp et al. 2000, 2001). HIV-1 sequences isolated worldwide have been divided into three main groups: major group M, group O, and group N (Kuiken et al. 2001). Group M viruses have been responsible for most of the AIDS cases worldwide, whereas HIV-1 group O, group N, and HIV-2 are confined mainly to western Africa (Chen et al. 1996). Until now, more than 20 different African primate species have been found to be infected with simian immunodeficiency viruses (SIV) (Hahn et al. 2000; Beer et al. 2001; Kuiken et al. 2001; Santiago et al. 2002; Peeters et al. 2002).

Phylogenetic analysis of all the different HIV-1 groups indicated that they are genetically more closely related to the SIVcpz sequences isolated from the chimpanzee *Pan troglodytes troglodytes* than to any other simian viruses (Gao et al. 1999; Hahn et al. 2000; Sharp et al. 2000; 2001). There are four different subspecies of chimpanzees residing in the geographic area of Africa: *P. t. verus* and *P. t. vellerosus* in western Africa, *P. t. troglodytes* and *P. t. schweinfurthii* in western Africa and Eastern-Equatorial Africa, respectively (Gao et al. 1999; Corbet et al. 2000; Hahn et al. 2000). Until now, there is no evidence for SIV infection of *P. t. verus* (Prince et al. 2002), whereas the only infected chimpanzee identified among *P. t. vellerosus* was a single animal (Cam4) kept together with a *P. t. troglodytes* infected with SIVcpz (Cam3) (Corbet et al. 2000). The SIVcpzCam3 chimpanzee was recovered from Cameroon (Corbet et al. 2000). Until now, five additional SIVcpz strains from *P. t. troglodytes* have been characterized from chimpanzees: one isolated from Cameroon (Cam5), one from an animal kept in the United States (US) (Gao, et al. 1999; Corbet et al. 2000), two from Gabon (Gab1, Gab2) (Peeters et al. 1989), and two more genetically divergent SIV strains (ANT, TAN1) isolated from the subspecies *P.t. schweinfurthii* in the DRC (Congo) and Tanzania, respectively (Peeters et al. 1992; Santiago et al. 2003). The prevalence of SIVcpz antibodies in wild-caught animals, as indicated recently by analysis of 58 chimpanzees from Côte d'Ivoire, Uganda, and Tanzania (Santiago et al. 2002), was found to be as low as 2%, which is far less than in other primate species (Jin et al. 1994; Bibollet-Ruche et al. 1996; Beer et al. 2000; Hahn et al. 2000; Peeters et al. 2002). The extremely low seroprevalence of SIVcpz in the chimps living in the wild could possibly be explained by the limited number of the animals sampled up until now, or by a dramatic loss of the animals in their natural habitat in recent decades (Weiss and Wrangham 1999).

On the one hand, groups M and N of HIV-1 are genetically more closely related to SIVcpz sequences from *P. t. troglodytes*, and it has been suggested that they represent two different cross-species transmissions from this chimpanzee subspecies (Gao et al. 1999; Hahn et al. 2000; Sharp et al. 2001). Moreover, group N consists of genomic regions with discordant phylogenetic relationships with respect to group M and SIVcpz, which was highly suggestive of recombination (Simon et al. 1998; Gao et al 1999). On the other hand, the origin of group O, as well as the evolutionary relationships between the different SIVcpz

Research was done at the Rega Institute.

Key words: HIV, SIVcpz, evolution, Bayesian inference.

E-mail: annemie.vandamme@uz.kuleuven.ac.be.

Mol. Biol. Evol. 20(12):1986–1996. 2003
DOI: 10.1093/molbev/msg207
Molecular Biology and Evolution, Vol. 20, No. 12,

and HIV-1 lineages throughout their complete genome, remains unclear. More specifically, based on phylogenetic analysis of full-length or partial genomic regions, group O was found to be an outgroup to HIV-1 and SIVcpz sequences from *P. t. troglodytes* (Gao et al. 1999; Corbet et al. 2000; Kuiken et al. 2001) and SIVcpzGab1 was found to cluster with the group of the SIVcpz sequences isolated from Cameroon (Corbet et al. 2000).

In the present study, we performed an in-depth analysis of the evolutionary relationships between HIV-1 groups M, O, and SIVcpz sequences isolated from *P. t. troglodytes* and *P. t. schweinfurthii*. Putative recombination between any of the HIV-1 and SIVcpz sequences was tested using bootscanning and Bayesian scanning plots, as well as a new method using a Bayesian multiple change-point (BMCP) model to infer parental sequences and crossing-over points.

Materials and Methods

Alignment of Viral DNA Sequences

All available full-length sequences of SIVcpz (Cam3, Cam5, US, Gab1, TAN1, and ANT) and several representative sequences of group O (ANT70, MVP5180), group N (YBF30), and group M (subtype A: U455, subtype B: WEAU160, subtype C: C2220 subtype D: 84ZR085) were downloaded from the Los Alamos HIV database (http//hiv-web.lanl.gov). DNA sequence alignment was performed using ClustalW version 1.74 (Thompson, Higgins, and Gibson 1994). The HIV-1/SIV sequence alignment was constructed by excluding all multiple coding regions, the long terminal repeat (LTR), and gaps from the alignment.

Investigation for Discordant Phylogenetic Relationships by Bootscanning Analysis

An exploratory analysis for the presence of any recombination events between all the different HIV-1/SIVcpz sequences was performed by bootscanning plots, using a sliding window of 500 nt, moving in steps of 50 nt and maximum likelihood estimated distances (F84 model with transitions/transversions = 2, default value) (Felsenstein 1984) as implemented in Simplot (3.2 beta version). The F84 model was used because it is the most complex model available in the current version of Simplot. The default value for ti/tv (transitions/transversion rate ratio) provided a good approximation of the value estimated by maximum likelihood (ti/tv = 2.14) using PAUP*4.0b10 (Swofford 1998). Bootscanning plots were then created for each monophyletic cluster separately, against all other HIV-1/SIVcpz sequences.

Bayesian Scanning Analysis

Bayesian scanning was performed including eight representative sequences within the five HIV-1/SIVcpz lineages: (1) SIVcpzGab1, (2) SIVcpzCam3/Cam5, (3) SIVcpzANT/TAN1, (4) ANT70 (HIV-1 group O), and (5) U455/WEAU160 (HIV-1 group M), using the GTR + Γ model and run with a sliding window of 500 nt, moving in steps of 50 nt. The program used for the Bayesian scanning was MrBayes (version 3.0) (Huelsenbeck et al. 2001). For every single window, four Markov chains were run for 10^5 generations with a burn-in of 6×10^3 generations. The Markov chain Monte Carlo (MCMC) chain length for the Bayesian scanning was set to 10^5 generations; we found no differences between MCMC pre-run estimates in several separate regions using 10^5 and 10^6 generations, suggesting that 10^5 steps provided a "realistic" length for a full-length genomic scan using Bayesian inference. Moreover, Bayesian scanning was used as an initial tool to investigate the existence of any topological incongruence between HIV-1/SIVcpz lineages. The fixed number for burn-in was based on an inspection of the number of generations required to reach stationarity in all posterior output files. Burn-in was fixed to 2× the number of generations (3,000th) required to reach equilibrium in all regions of the alignment. The approximate posterior probabilities of the partitions: (1) SIVcpzANT/TAN-HIV-1 group O, (2) SIVcpzANT/TAN-SIVcpzGab1, (3) HIV-1 groupO-SIVcpzCamUS, (4) HIV-1 groupM-SIVcpzCamUS, and (5) SIVcpzGab1-SIVcpzCamUS, for which there was evidence of discordances by the bootscanning analysis, were plotted throughout the complete genome.

Bayesian Multiple Change-Point Model Analysis

The Bayesian multiple change-point (BMCP) model was initially developed to infer spatial phylogenetic variation (SPV) in DNA sequences from organisms that may have undergone recombination (Suchard et al. 2002, 2003). The strength of the BMCP approach is that it allows evolutionary selective pressures, rates, and reconstructed topologies to all vary independently, ensuring that selective pressure and rate changes do not lead to false inference regarding topological changes. The BMCP model employed in this study is a modification of Suchard et al. (2002) that allows for sampling of all 15 possible topologies relating the five distinct HIV-1/SIVcpz lineages. In this way, we did not need to assume a fixed reference tree for a subset of the lineages and test for recombination producing the remaining one; rather, we could test for discordance among all five simultaneously. For BMCP model analysis, we performed several runs using different sequences from SIVcpzCamUS (SIVcpzCam3, SIVcpzUS), group M (U455, WEAU160), and group O (ANT70, MVP5180) lineages together with SIVcpzGab1, and either SIVcpzANT or SIVcpzTAN1. We ran the model for 10^6 generations, assuming a λ value of 10, as was previously described (Suchard et al. 2003). To assess convergence, we ran five independent BMCP chains on a single data set including either SIVcpzANT or SIVcpzTAN1 (10 runs), and then two BMCP chains for each of the data sets including distinct sequences from each HIV-1/SIVcpz lineage (12 runs); thus there were 22 runs in total. Posterior probabilities of the five partitions listed above, and the marginal posterior estimates of the evolutionary parameters were plotted throughout the complete genome. All BMCP analyses were run using the same alignment as was used for the bootscanning and Bayesian scanning, as well for the conditionally independent phylogenetic analyses to be described.

Conditionally Independent Phylogenetic Analysis

In genomic regions with high evidence of discordant evolutionary relationships between the HIV-1/SIVcpz sequences based on the Bayesian approaches, separate phylogenetic analyses were done for the putative non-recombinant fragments. The borders of these fragments were based on the previous Bayesian analyses. The best-fitting nucleotide substitution model was chosen according to the likelihood ratio test (Felsenstein 1981, Goldman 1993) among 64 different models using Modeltest (Posada and Crandall 1998) and PAUP*4.0b10 (Swofford 1998). Conditionally independent phylogenetic analysis was accomplished using a maximum likelihood (ML) approach with the best-fitting evolutionary model as implemented in PAUP* (Table 1). More specifically, all model parameters were initially estimated on a Neighbor-Joining tree, and tree topologies were subsequently evaluated using a heuristic search approach that implemented both tree bisection-reconnection and nearest-neighbor interchange perturbations. Bootstrapping was performed on the Neighbor-Joining trees (1,000 replicates) to assess the reliability of the obtained topologies. Phylogenetic trees were also obtained using Bayesian inference under the general time reversible (GTR) model including a Γ distributed rates heterogeneity among sites as implemented in MrBayes (version 3.0) (Huelsenbeck et al. 2001). For Bayesian inference, four Markov chains run for 10^6 generations with a burn-in of 2×10^4, and four chains run for 10^5 generations with a burn-in of 2×10^3 for the nucleotide and the amino acid based sequences, respectively, were used for the reconstruction of the consensus tree. All MCMC analyses were repeated in five separate runs (using three heating chains and one cold chain) for both nucleotide and amino acid alignments. To assess for any significant differences between candidate topologies in different pieces of the alignment, we used the Approximately Unbiased (AU)-test (Shimodaira 2000) as implemented in CONSEL (Shimodaira and Hasegawa 2001). For each region, site-wise log-likelihoods were estimated for the candidate trees using PAUP* with the best-fitted evolutionary model.

Test for Substitution Saturation

To test for substitution saturation, we plotted the observed number of substitutions (transitionsand transversions) for the 1st plus 2nd codon positions and, for all positions, against the evolutionary distances corrected for multiple substitutions and estimated by maximum-likelihood.

Results

Exploratory Analysis

An exploratory bootscan analysis throughout the complete alignment of the HIV-1/SIVcpz sequences gave indications for discordant phylogenies and, thus, for a putative recombinant origin of the HIV-1/SIVcpz strains. To check whether there were any consistent clusters of different HIV-1 subtypes, groups, or SIVcpz lineages, we examined first which sequences remained monophyletic across the entire alignment. According to the bootscanning plots, all sequences belonging to HIV-1 group M, to SIVcpzCamUS (SIVcpzCam3, SIVcpzCam5 and SIVcpz-US), to group O, as well as to the two SIVcpz sequences (ANT, TAN1) (SIVcpzANT/TAN) isolated from the chimpanzee subspecies *P. t. schweinfurthii*, each remained a monophyletic group throughout the complete alignment. Additionally, none of the strains belonging to these three clusters showed a closer relationship to any of the other sequences than the other strains of the same cluster (data not shown).

Table 1
Best-Fitting Evolutionary Model for Different Segments of the Alignment

Region of the Alignment	Model Selected[a]	α[b]	I[c]
1–800	GTR+G+I	2.97	0.39
801–1500	GTR+G+I	2.4	0.39
1501–2600	GTR+G+I	2.38	0.42
2601–3300	GTR+G+I	1.17	0.34
3301–3850	GTR+G	0.42	—
3851–4300	GTR+G+I	1.56	0.24
4301–4850	TVM+G+I	1.13	0.11
4851–5937	GTR+G+I	1.76	0.26

[a] GTR is the general time reversible and TVM a GTR sub-model assuming that the rates of A-G and C-T are equal. G and I indicate models allowing Γ-distributed rates across sites and a proportion of invariable sites, respectively.
[b] Shape parameter of the Γ-distribution of rates among sites.
[c] Proportion of invariable sites.

Bootscanning plots of each of these four monophyletic groups and of the SIVcpzGab1 sequence, thus five lineages in total, were performed against the other four lineages. HIV-1 group N was not included, because this group is already known to be recombinant consisting of HIV-1 group M and SIVcpzUS-like regions in the 5′ and 3′ half of its genome, respectively (Gao et al. 1999; Hahn et al. 2000). Such analysis revealed discordant phylogenetic clustering of almost all lineages when comparing different genomic regions (fig. 1A–E).

Although bootscanning plots were highly suggestive about the boundaries of the putative recombinant pieces in several regions such as: 1–800, 800–1500, 2300–3200, and 3200–3600, there was no clear indication about the evolutionary relationships between the HIV-1/SIVcpz sequences in the rest of the genome (positions: 1500–3000, 3600–4900). Phylogenetic analysis in a multiple set of different parts within these genomic regions revealed discordant phylogenies, whereas the borders of each part could not be precisely identified. These difficulties suggest that the evolutionary relationships of highly divergent sequences such as the HIV-1/SIVcpz lineage could not be entirely explored using bootscanning analysis.

Bayesian Scanning and Bayesian Multiple Change-Point (BMCP) Model Analysis

To explore in detail the evolutionary relationships and the boundaries of the genomic regions with discordant phylogenetic relationships between HIV-1/SIVcpz lineages, we performed two different kinds of analyses: (1) a genomic scan of the HIV-1/SIVcpz phylogenetic

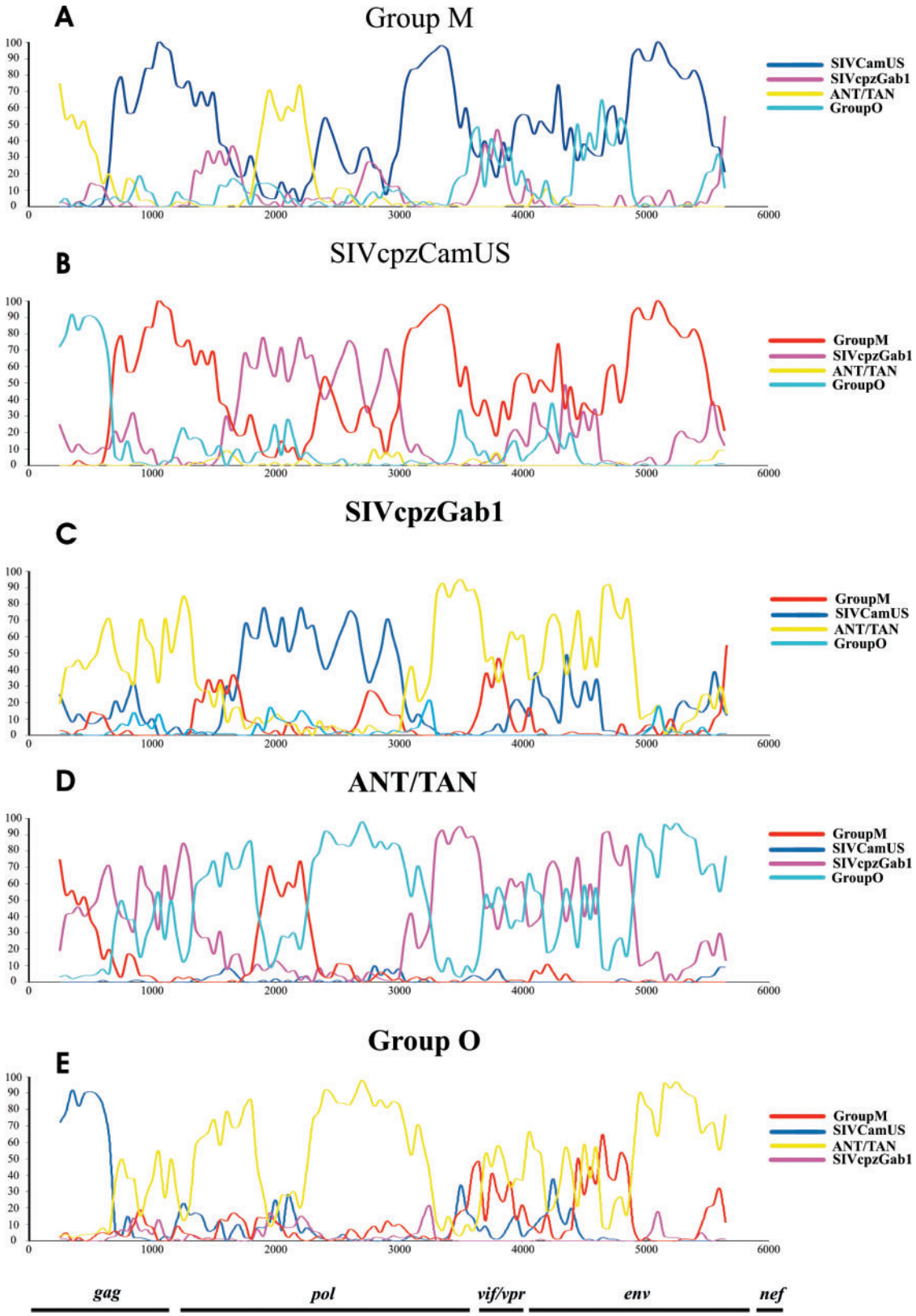

FIG. 1.—Bootscanning plots of the HIV-1/SIVcpz alignment: (*A*) of the HIV-1 group M sequences against SIVcpzGab1, SIVcpzCamUS (Cam3, Cam5 and US), SIVcpzANT/TAN and HIV-1 group O sequences, (*B*) of the SIVcpzCamUs against HIV-1 group M, HIV-1 group O, SIVcpzGab1, and SIVcpzANT/TAN, (*C*) of the SIVcpzGab1 against SIVcpzCamUS, SIVcpzANT/TAN and HIV-1 group M, HIV-1 group O, (*D*) of the SIVcpzANT/TAN against SIVcpzCamUS, SIVcpzGab1, and HIV-1 group M, HIV-1 group O, and (*E*) of the HIV-1 group O against SIVcpzCamUS, SIVcpzANT/ANT, SIVcpzGab1 and HIV-1 group M. All the bootscanning plots were performed with a sliding window of 500 bps moving in steps of 50 bps and were based on the F84 ML estimated distances. The corresponding genomic regions are shown at the bottom of the figure.

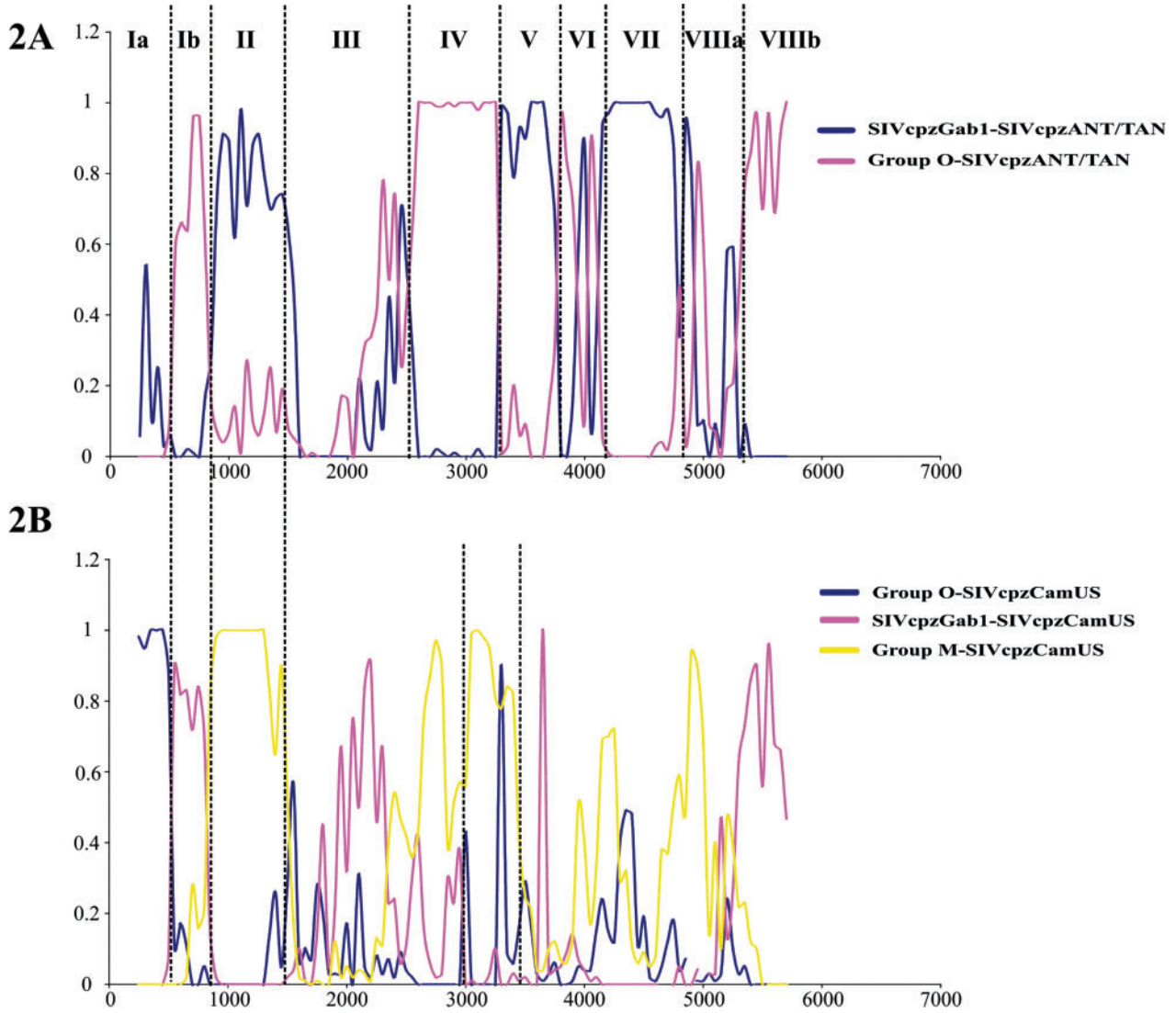

FIG. 2.—Bayesian scanning plots of the HIV-1/SIVcpz sequences: (*A*) approximate posterior probabilities of the SIVcpzGab1-SIVcpzANT/TAN and group O-SIVcpzANT/TAN partitions and (*B*) of the groupM-SIVcpzCamUS, groupO-SIVcpzCamUS and SIVcpzGab1-SIVcpzCamUS plotted throughout the complete alignment. Bayesian analysis was run for four Markov chains and 10^5 generations using the GTR + Γ evolutionary model with a sliding window of 500 bps moving in steps of 50 bps.

relationships by Bayesian inference with the general time reversible (GTR) model including a Γ-distribution of rates among sites, and (2) a Bayesian analysis using a multiple change-point model (BMCP) as described recently (Suchard et al. 2002) and modified for a data set of five taxa.

The first approach resembles the bootscanning analysis; Bayesian inference analysis is performed in a sliding window of a specified length, and then the support of every single clade (approximate posterior probabilities instead of the bootstrap values) is plotted throughout the alignment. The Bayesian multiple change-model was developed to infer spatial phylogenetic variation (SPV) in DNA sequences from organisms that have undergone recombination. The BMCP model to infer recombination breakpoints extends the approaches of Husmeier, Wright and colleagues (McGuire et al. 2000; Husmeier and Wright 2001, 2002; Husmeier and McGuire 2002). Husmeier, Wright, and colleagues detect recombination by modeling multiple site evolution with a hidden Markov model (HMM), where the hidden states at each site along the alignment are the various possible topologies. The BMCP model not only allows for varying evolutionary rates, selective pressures, and topologies along the sequences, it also provides simultaneous estimates of uncertainty of all quantities including the number and locations of the inferred crossover points (Suchard et al. 2002, 2003). The model assumes that the sites along the data separate into an unknown number of contiguous segments, each with possibly different evolutionary relationships between organisms, evolutionary rates, and transition-transversion ratios.

The results of the Bayesian scanning are shown in two plots (fig, 2A, B) where the posterior probabilities of the five different partitions, (1) SIVcpzANT/TAN-group O, (2)

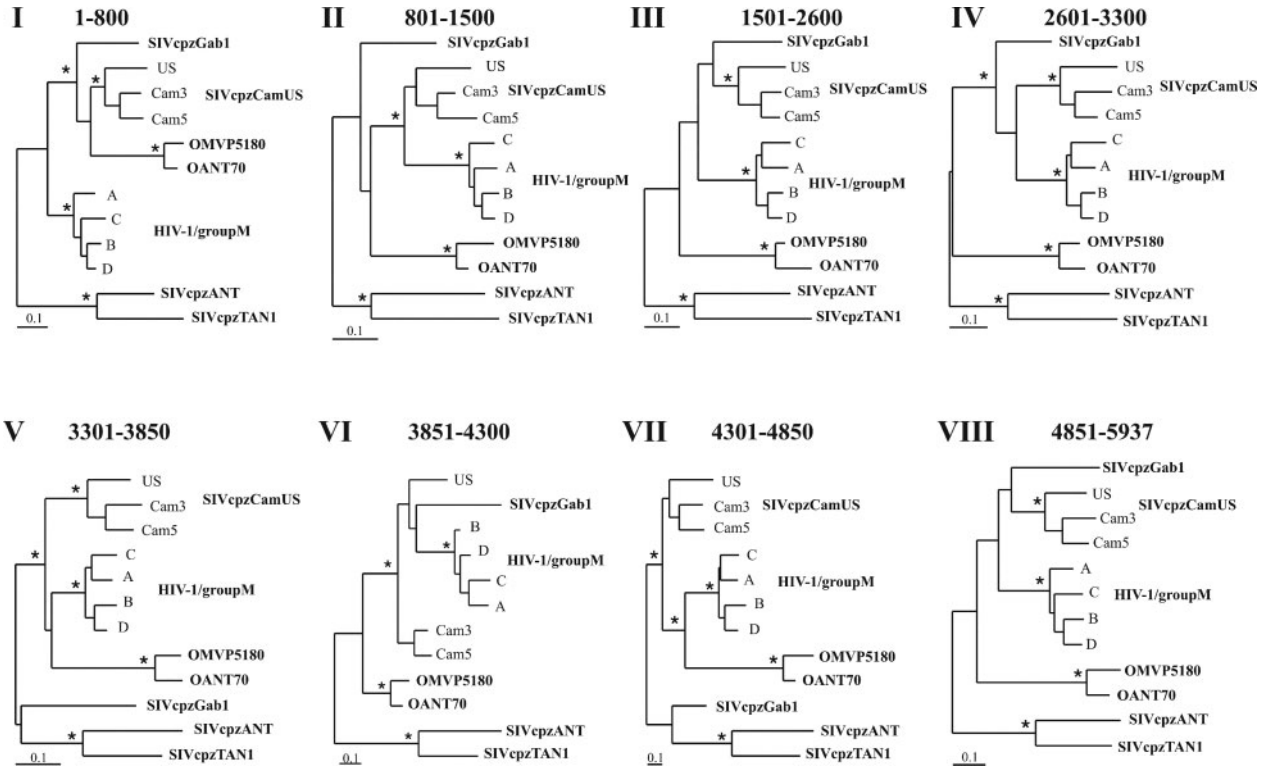

FIG. 3.—Conditionally independent phylogenetic analysis in different parts of the HIV-1/SIVpcz alignment. Asterisks indicate branches supported by (1) bootstrap values > 75% using ML with the best-fitting evolutionary model and (2) posterior probabilities consistently > 0.75 in all independent MCMC runs using both nucleotide and amino acid alignments. Asterisks are not shown for the relationships within group M, group O, and SIVcpzCamUS lineages. Positions in the alignments are shown above each tree. All trees were rooted at their midpoint, as implemented in PAUP.

SIVcpzANT/TAN-SIVcpzGab1, (3) groupO-SIVcpzCamUS, (4) groupM-SIVcpzCamUS, and (5) SIVcpzGab1-SIVcpzCamUS as indicated by the exploratory analysis, were plotted throughout the alignment. According to figure 2A, B, there was strong evidence for discordant phylogenetic variation clustering: (1) SIVcpzANT/TAN with group O (regions IV and VIIIb at positions 2600–3300 and 5400–5937, respectively) or SIVcpzGab1 (regions II, V, and VII at positions 800–1500, 3300–3850, and 4300–4850, respectively), (2) group O with SIVcpzCamUS (region Ia at positions 1–500), and (3) SIVcpzCamUS with group M (region II and between 2900–3400) or SIVcpzGab1 (region VIIIb) (800–1500 and 5400–5937, respectively). At first glance, the results of the Bayesian approaches provided a clearer picture about the boundaries and the levels of confidence of the different partitions than the bootscanning analysis, even in regions where the bootscanning plots were unclear (e.g., position of the breakpoints between regions V, VI and VII, fig. 2A, B).

Conditionally Independent Phylogenetic Analysis

Regarding the evolutionary relationships between the SIVcpzANT/TAN—group O and SIVcpzGab1—we divided the alignment into 10 different pieces as shown in figure 2A, as suggested by the Bayesian scanning results and supported by bootscanning (fig. 1). Phylogenetic analyses using ML or Bayesian inference in both nucleotide and amino acid sequences further confirmed these findings (fig. 3 I–VIII). The analyses showed a similar topology for regions 1–500 (Ia) and 500–800 (Ib), as well as for the last two fragments (VIIIa and VIIIb) (data not shown), which were thus joined into a pooled region, I and VIII, respectively. "Significant" clusters were defined as those supported by (1) bootstrap values >75% using ML with the best-flitting evolutionary model and (2) posterior probabilities consistently > 0.75 in all independent MCMC runs in both nucleotide and amino acid alignments (fig. 3). Thus, discordant phylogenetic relationships between the HIV-1/SIVcpz lineage topologies were inferred in pieces I, II, IV, V, VI, and VII (fig. 3). The significance of dissimilar topologies in different regions was tested also, by the Approximately Unbiased (AU)-test (Shimodaira 2000), according to which the neighboring topologies for regions I, II, V, VI, and VII were significantly rejected ($P < 0.05$) (data not shown). In contrast, in region IV, the topology inferred for region V (suboptimal) did not differ significantly from the ML inferred tree ($P = 0.31$), probably because of the high levels of similarity between the HIV-1/SIVcpz lineages (middle of *pol*) compared to the other parts of the alignment.

The clustering between SIVcpzCamUS and group M was found to be significant in region II (fig. 3II) and 2901–3400, whereas the clustering between SIVcpzCamUS and SIVcpzGab was not significantly supported in region

5401–5937 (region VIIIb) (data not shown). Interestingly, SIVcpzCamUS, SIVcpzGab1, and group O clustered together in the first 800 nt of the alignment (fig. 3I).

To test for substitution saturation, we plotted the transition to transversion ratios for all the sites or for only the 1st plus 2nd codon positions versus the ML estimated evolutionary distances corrected for multiple substitutions for three fragments of the alignment belonging to partial *gag, pol*, and *env* regions, respectively (data not shown). According to the saturation plots, there was no evidence for substantial substitution saturation, especially after the elimination of the 3rd codon position, and thus phylogenetic inference that corresponds mainly to substitutions in the 1st and 2nd codon (analysis of amino acid changes) was found not to be substantially affected by substitution saturation bias.

Multiple Change-Point (BMCP) Model Analysis

To further confirm the findings about the spatial phylogenetic variation, we performed a BMCP model analysis using representative sequences for the five distinct lineages. Because the number of strains that can currently be analyzed with BMCP is restricted, we ran these analyses including only one sequence from each lineage at a time. We performed multiple analysis using different sequences from SIVcpzCamUS, group M and group O lineages, together with SIVcpzGab1 and either SIVcpzANT or SIVcpzTAN1, because there is considerable variability between these last two sequences. The results of all independent runs were almost identical within data sets including SIVcpzANT or within data sets including SIVcpzTAN1 sequences, thus prompting us to average the posterior probabilities for the SIVcpzANT- or SIVcpzTAN1- subset of sequences (fig. 4A). According to the BMCP analysis results, there was evidence of spatial phylogenetic variation throughout the HIV-1/SIVcpz genome (fig. 4A). Comparing the results of the BMCP analysis and the Bayesian scanning, there were some differences in the posterior probabilities of the different regions (fig. 2A, 4A). More specifically, the differences were more pronounced in region II (800–1500), whereas there were some minor differences, also in region III (1500–2600) (fig. 2A, 4A). Interestingly, in region II none of the SIVcpzGab1 or group O sequences clustered significantly with SIVcpzANT/TAN (fig. 3II). We should note here that BMCP and Bayesian scanning were run with five and eight sequences, respectively, and thus the analyzed partitions were not identical. Bayesian scanning based on the exact data set used in BMCP analysis revealed almost identical results with the BMCP analysis (data not shown), thus suggesting that observed differences were due to the distinct datasets used in the analyses.

Interestingly, the BMCP profiles of group O– and SIVcpzGab1– with either SIVcpzTAN1 or SIVcpzANT were quite different in regions 800–1500 and 3800–5400 (data not shown), probably because of the great genetic divergence between the SIVcpzANT and SIVcpzTAN1 sequences. The differences were more pronounced in *env* (3800–5400) because of the high genetic divergence between the HIV-1/SIVcpz sequences in this region, as depicted in figure 4B.

Evolutionary Relationships as Deduced from These Analyses

Thus, phylogenetic analysis using ML with the best-fitting evolutionary model and Bayesian inference in both nucleotide and amino acid sequences revealed discordant evolutionary relationships between the HIV-1/SIVcpz sequences throughout the genome (shown schematically in figure 5). Although unrooted, the trees are drawn with SIVcpzANT/TAN as most external branch, because in most regions the root is stably put on that branch. As depicted in figure 5, our results indicate that, HIV-1 group M clusters significantly with the two SIV lineages from *P. t. troglodytes* (SIVcpsCamUS and SIVcpzGab1) and not with group O in regions IV and VI, and significantly only with SIVcpzCamUS in region II. In regions V and VII, HIV-1 group M clusters significantly with group O and SIVcpzCamUS, and not with SIVcpzGab1. In region I, group O clustered significantly with SIVczGab1 and SIVcpzCamUS and not with group M, whereas in regions III and VIII no significant clustering between the HIV-1/SIVcpz lineages could be found (fig. 5). Interestingly, HIV-1 group M and group O significantly clustered as a separate group in region VII.

Discussion

Previous analyses have described that the origin of HIV-1 group M is an interspecies transmission from *P. t. troglodytes*, because SIVcpz from *P. t. troglodytes* is most closely related to HIV-1 group M (Gao et al. 1999; Hahn et al. 2000; Sharp et al. 2000; 2001). Furthermore, group N was found to be a mosaic consisting of SIVcpz and HIV-1/group M-related sequences (Gao et al. 1999), whereas HIV-1 group O was found to be an outgroup to the HIV-1, SIVcpz sequences isolated from *P. t. troglodytes* (Corbet et al. 2000; Kuiken et al. 2001; Santiago et al. 2003).

To investigate the complex relationships between the HIV-1 and SIVcpz sequences, we performed an in-depth analysis of the five distinct HIV-1/SIVcpz lineages, using bootscanning, a sliding window Bayesian approach, ML and BMCP analysis. We found evidence that in case of highly divergent sequences, such as within HIV-1/SIVcpz, bootscanning analysis based on ML estimated distances (F84) provided an inadequate method for testing for discordant clustering; in several cases it did not provide clear results about the boundaries of the putative recombinant fragments. The inadequacy of the bootscanning method in deep phylogenies is not surprising because of the implementation of an oversimplified and fixed evolutionary model (F84) and a distance based tree-searching algorithm (NJ). In contrast, the Bayesian scanning method provided a clearer picture, not only about the boundaries of the regions with discordant evolutionary relationships but also about the levels of support of the phylogenetic clusters under study in different regions of the alignment.

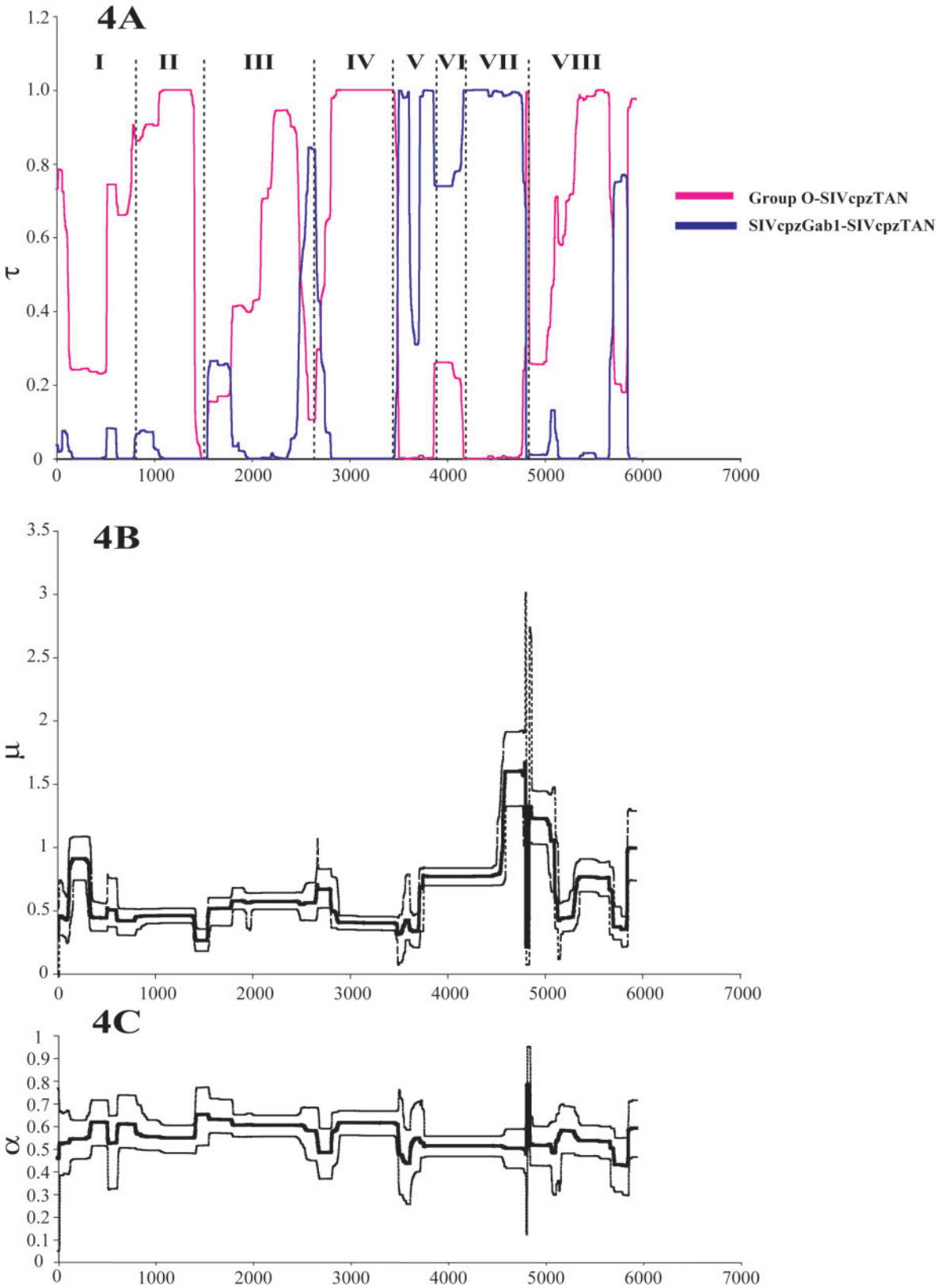

FIG. 4.—Bayesian multiple change-point model results. (*A*) Marginal posterior probabilities of the partitions groupO-SIvcpzTAN1, SIVcpzGab1-SIVcpzTAN throughout the genome averaged over four data sets including different representative sequences for the group O, group M, and the SIVcpzCamUS sequences. The boundaries of the regions with discordant evolutionary relationships between the HIV-1/SIVcpz sequences, identified by Bayesian scanning, are shown with dashed lines. Marginal posterior median (solid line) and 95% Bayesian credible interval (dashed line) of (*B*) the expected average branch length μ and (*C*), the evolutionary parameter α.

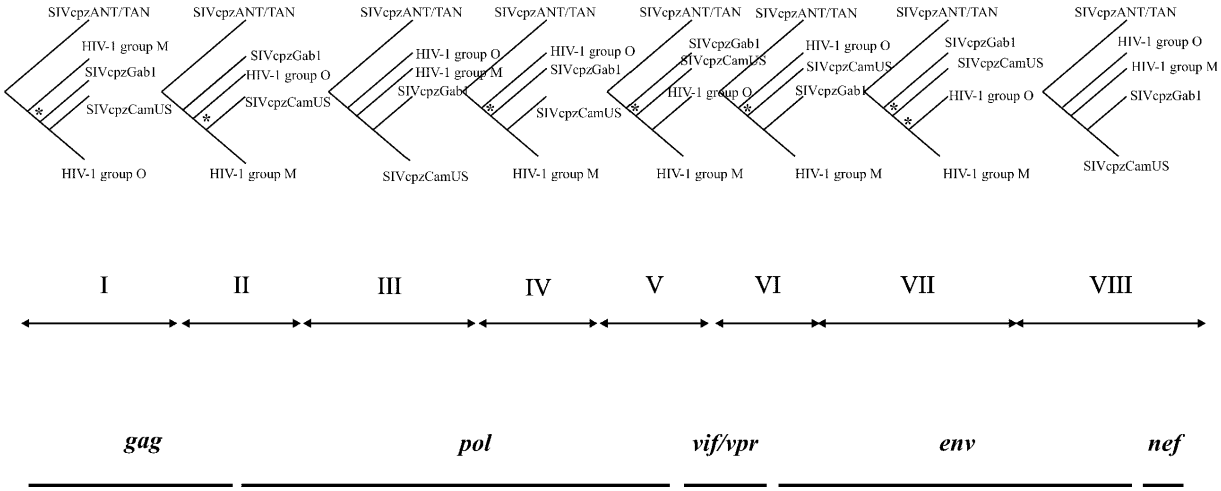

FIG. 5.—Schematic representation of the HIV-1/SIVcpz evolutionary relationships in the eight pieces regarding the SIVcpzGab1—or the HIV-1 group O—SIVcpzANT/TAN partitioning. The corresponding genomic regions for each piece are shown at the bottom of the figure. Asterisks indicate significant clusters inferred by phylogenetic analyses.

Conditionally independent phylogenetic analyses, using different approaches, confirmed further the results obtained by the Bayesian scanning method about the discordant clustering within the HIV-1/SIVcpz lineage. In particular, the BMCP analysis provided additional evidence for spatial phylogenetic variation throughout the HIV-1/SIVcpz genome in accordance with the conditionally independent analyses. The BMCP analysis, which allows for sampling of all 15 possible topologies relating the five distinct HIV-1/SIVcpz lineages, is a new method, used complementarily to Bayesian scanning, to infer the spatial phylogenetic variation of the HIV-1/SIVcpz sequences. Because it is not currently computationally feasible to analyze more than five sequences using a BMCP approach, we chose one representative sequence from each lineage, HIV-1 group M, HIV-1 group O, SIVcpzCamUS, SIVcpzGab1, or SIVcpzANT/TAN. This analysis revealed some incongruence between the data sets when using either SIVcpzANT or SIVcpzTAN1, confirming the high divergence between the latter two sequences.

According to our findings, HIV-1 group M clusters significantly with two SIV lineages from *P. t. troglodytes* (SIVcpsCamUS and SIVcpzGab1) and not with group O in middle *pol*, and in partial *vif*/*env*, and significantly only with SIVcpzCamUS in 3′*gag*/*pol*. We should note here that in middle *pol*, the ML topology did not differ significantly from the phylogenetic tree inferred in neighboring regions, probably because of the high levels of similarity between HIV-1/SIVcpz lineages (middle of *pol*) compared to the other parts of the alignment. In *pol*/*vif* and in middle *env*, including V3, HIV-1 group M clusters significantly with group O and SIVcpzCamUS, and not with SIVcpzGab1, whereas in partial *env* the group M also clustered with group O. The gp120—and especially the V3 region—contain the binding site for the CD4 receptor; thus clustering of groups M and O in this region may be caused by convergent evolution to achieve virus entry into human cells and not to ancient recombination events. Furthermore, in p17/p24, group O clustered significantly with SIVczGab1 and SIVcpzCamUS and not with group M. Taken as a whole, these analyses are more supportive of SIVcpz from *P. t .troglodytes* as the origin of HIV-1 for both HIV-1 group M and group O, than SIVcpz from *P. t. schweinfurthii*.

The discordant phylogeny throughout the HIV-1/SIVcpz sequences suggests that either ancient recombination events between ancestral HIV-1/SIVcpz sequences altered rates of evolution, which in some regions may have been driven by convergent evolution, or a combination of these events has probably shaped the genomic makeup of the different HIV-1/SIVcpz lineages. The hypothesis of altered rates of evolution between the HIV-1/SIVcpz sequences was further supported by the BMCP analysis, where we found differences in the BMCP profiles when comparing analyses with either SIVcpzANT or SIVcpzTAN1, thus suggesting that the inferred evolutionary relationships between the HIV-1/SIVcpz sequences are not stable.

According to the inferred phylogenies, evolutionary distances between the HIV-1 and SIVcpz sequences isolated from *P. t. troglodytes* vary greatly with the genomic region, which is also supportive for a great impact of altered rates of evolution. Such a finding is not surprising because the SIVcpz viruses belonging to SIVcpzCam (Cam3, Cam5) and SIVcpzGab1 lineages were isolated from diverse geographic areas, Cameroon and Gabon, respectively, which means that a number of parameters (selective pressure, number of infected animals, etc.) could have shaped the evolution of the SIVcpz virus infecting chimpanzees, which remained separated in their natural habitat in a different way.

If recombination contributed greatly to the observed phylogenetic discordance, then it should have occurred early in the evolution process of SIVcpz and most probably before the interspecies transmission to human. Such events are not easily identifiable after a prolonged period of sequence divergence of the lineages after their separation.

Interestingly, it has recently been suggested that SIVcpz in chimpanzees originated from a recombination event between ancestral SIVs infecting red-capped mangabeys (SIVrcm) and greater spot-nosed monkeys (SIVgsn) (Courgnaud et al. 2002; Bailes et al. 2003); our findings indicate that additional recombination events or other factors continued to shape the evolutionary history of primate lentiviruses.

In conclusion, we find evidence for discordant phylogenetic relationships between HIV-1 and SIVcpz sequences throughout the genome, as shown schematically in figure 5. This discordance in topologies might have been caused by ancient recombination events, altered rates of evolution, concerted evolution, or a combination of these factors. Our analysis also suggests that SIV from *P. t. troglodytes* is the most probable origin for both HIV-1 group M and group O.

Acknowledgments

We acknowledge Andrew Rambaut, Xuhua Xia, Vincent Moulton, and Ziheng Yang for their supportive comments and suggestions regarding this study. D.P. was supported by a Marie Curie fellowship from the European Commission (QLK2-CT2001–51062). P.L. was supported by the Flemish Institute for Scientific-technological Research in Industry (IWT); M.S. is a postdoctoral fellow with the Fonds voor Wetenschappelijk Onderzoek (FWO). M.A.S. was partially supported by the UCLA Center for AIDS Research (AI28697). This work was supported by the Flemish Fonds voor Wetenschappelijk Onderzoek (FWO G.0288.01).

Literature Cited

Bailes, E., F. Gao, F. Bibollet-Ruche, V. Courgnard, M. Peeters, P. A. Marx, B. H. Hahn, and P. M. Sharp. 2003. Hybrid origin of SIV in chimpanzees. Science **300**:1713.

Beer, B. E., E. Bailes, G. Dapolito, et al. (12 co-authors). 2000. Patterns of genomic sequence 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. S. 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.

Bibollet-Ruche, F., A. Galat-Luong, G. Cuny, P. Sarni-Manchado, G. Galat, J. P. Durand, X. Pourrut, and F. Veas. 1996. Simian immunodeficiency virus infection in a patas monkey (*Erythrocebus patas*): evidence for cross-species transmission from African green monkeys (*Cercopithecus aethiops sabaeus*) in the wild. J. Gen. Virol. **77**:773–781.

Chen, Z., P. Telfier, A. Gettie, P. Red, 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.

Corbet, S., M. C. Müller-Trutwin, P. Versmisse, et al. (12 co-authors). 2000. env sequences of simian immunodeficiency viruses from chimpanzees in Cameroon are strongly related to those of human immunodeficiency virus group N from the same geographic area. J. Virol. **74**:**529–534.**

Courgnaud, V., M. Salemi, X. Pourrut, et al. (11 co-authors). 2002. Characterization of a novel simian immunodeficiency virus with a vpu gene from greater spot-nosed monkeys (*Cercopithecus nictitans*) provides new insights into 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.

———. 1984. Distance methods for inferring phylogenies: a justification. Evolution **38**:16–24.

Gao, F., E. Bailes, D. L. Robertson, et al. (12 co-authors). 1999. Origin of HIV-1 in the chimpanzee Pan troglodytes troglodytes. Nature **397**:436–441.

Goldman, N. 1993. Statistical tests of models of DNA substitution. J. Mol. Evol. **36**:182–198.

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.

Huelsenbeck, J. P., F. Ronquist, R. Nielsen, and J. P. Bollback. 2001. Bayesian inference of phylogeny and its impact on evolutionary biology. Science **294**:2310–2314.

Husmeier D., and F. Wright. 2001. Detection of recombination in DNA multiple alignments with hidden Markov models. J. Computat. Biol. **8**:401–427.

———. 2002. A Bayesian approach to discriminate between alternative DNA sequence segmentations. Bioinformatics **18**:226–234.

Husmeier, D., and G. McGuire. 2002. Detecting recombination with MCMC. Bioinformatics **18**:S345–S353.

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.

Kuiken, C., B. Foley, B. Hahn, P. Marx, F. McCutchan, J. Mellors, S. Wolinsky, and B. Korber. HIV Sequence Compendium 2001. Published by Theoretical Biology and Biophysics Group, Los Alamos National Laboratory, Los Alamos, N. Mex.

McGuire G., F. Wright, and M. J. Prentice. 2000. A Bayesian model for detecting past recombination events in DNA multiple alignments. J. Comp. Biol. **7**:159–170.

Peeters, M., V. Courgnaud, P. Abela et al. (14 co-authors). 2002. Risk to human health from a plethora of simian immunodeficiency viruses in primate bushmeat. Emerg. Infect. Dis. **8**:451–457.

Peeters, M., K. Fransen, E. Delaporte, M. Van der Haesevelde, G. M. Geshry-Damet, L. Kestens, G. Van der Groen, and P. Piot. 1992. Isolation and characterization of a new chimpanzee lentivirus (simian immunodeficiency virus isolate cpz-ant) from a wild-captured chimpanzee. AIDS **6**:447–451.

Peeters, M., C. Honore, T. Huet, L. Bedjabaga, S. Ossari, P. Bussi, R. W. Cooper, and E. Delaporte. 1989. Isolation and partial characterization of an HIV-related virus occurring naturally in chimpanzees in Gabon. AIDS **3**:625–630.

Posada, D., and K. A. Crandall. 1998. MODELTEST: testing the model of DNA substitution. Bioinformatics **14**:817–818.

Prince, A. M., B. Brotman, D. H. Lee, L. Andrus, J. Valinsky, and P. Marx. 2002. Lack of evidence for HIV type 1-related SIVcpz infection in captive and wild chimpanzees (Pan troglodytes verus) in West Africa. AIDS Res. Hum. Retroviruses **18**:657–660.

Santiago, M. L., F. Bibollet-Ruche, E. Bailes et al. (13 co-authors). 2003. Amplification of a complete simian

immunodeficiency virus genome from fecal RNA of a wild chimpanzee. J. Virol. **77**:2233–2242.

Santiago, M. L., C. M. Rodenburg, S. Kamenya et al. (24 coauthors). 2002. SIVcpz in wild chimpanzees. Science **295**:465.

Sharp, P. M., E. Bailes, R. R. Chaudhuri, C. M. Rodenburg, M. O. Santiago, and B. H. Hahn. 2001. The origins of acquired immune deficiency syndrome viruses: where and when? Philos. Trans. R. Soc. Lond. Ser. B. Biol. Sci. **356**:867–876.

Sharp, P. M., E. Bailes, F. Gao, B. E. Beer, V. M. Hirsch, and B. H. Hahn. 2000. Origins and evolution of AIDS viruses: estimating the time scale. Biochem. Soc. Trans. **28**:275–282.

Shimodaira, H. 2000. Another calculation of the *p*-value for the problem of regions using the scaled bootstrap resamplings. Technical Report No. 2000–2035. Stanford University, Stanford, Calif.

Shimodaira, H., and M. Hasegawa. 2001. CONSEL: for assessing the confidence of phylogenetic tree selection. Bioinformatics **17**:1246–1247.

Simon, F., P. Mauclere, P. Roque, I. Loussert-Ajaka, M. C. Müller-Trutwin, S. Saragosti, M. C. Georges-Courbot, F. Barre-Sinoussi, and F. Brun-Vezinet. 1998. Identification of a new human immunodeficiency virus type 1 distinct from group M and group O. Nat. Med. **4**:1032–1037.

Suchard, M. A., R. E. Weiss, K. S. Dorman, and J. S. Sinsheimer. 2002. Oh brother, where art thou? A Bayes factor test for recombination with uncertain heritage. Syst. Biol. **51**:715–728.

———. 2003. Inferring spatial phylogenetic variation along nucleotide sequences: a multiple change-point model. J. Am. Stat. Assoc. (in press).

Swofford, D. L. 1998. PAUP*: phylogenetic analysis using parsimony (*and other methods). Version 4. Sinauer Associates, Sunderland, Mass.

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.

Weiss, R. A., and R. W. Wrangham. 1999. From pan to pandemic. Nature **397**:385–386.

Edward Holmes, Associate Editor

Accepted June 23, 2003