

Phylogenetic Reconstruction of a Known HIV-1 CRF04_cpx Transmission Network Using Maximum Likelihood and Bayesian Methods

Dimitrios Paraskevis,^{1,3} Emmanouil Magiorkinis,¹ Gkikas Magiorkinis,¹ Vassilios G. Kiosses,² Philippe Lemey,³ Anne-Mieke Vandamme,³ Andrew Rambaut,⁴ Angelos Hatzakis¹

¹ National Retrovirus Reference Center, Department of Hygiene and Epidemiology, Athens University Medical School, Athens, Greece

² AHEPA University Hospital, Thessalonica, Greece

³ Rega Institute for Medical Research, Katholieke Universiteit Leuven, Leuven, Belgium

⁴ Department of Zoology, University of Oxford, Oxford, UK

Received: 9 February 2004 / Accepted: 1 June 2004 [Reviewing Editor: Dr. Lauren Ancel-Meyers]

Abstract. The CRF04_cpx strains of HIV-1 accounts for approximately 2–10% of the infected population in Greece, across different transmission risk groups. CRF04_cpx was the lineage documented in an HIV-1 transmission network in Thessalonica, northern Greece. Most of the transmissions occurred through unprotected heterosexual contacts between 1989 and 1993. Blood samples were available for six patients, obtained 6–10 years later, except for one patient sampled in 1991. Our objective was to examine whether the transmission history is compatible with the evolutionary tree of the virus, in partial *gag*, partial *env*, and partial *gag + env*.¹ The inferred phylogenetic tree obtained using maximum likelihood and Bayesian methods in partial *gag + env* was much closer to the transmission tree than that using either *env* or *gag* separately. Our findings suggest that the epidemiological relationships among patients who have been infected by a common source correspond almost exactly to the evolutionary trees of the virus, given that enough phylogenetic signal is

present in the alignment. Moreover, we found evidence that recombination is not the most parsimonious explanation for the phylogenetic incongruence between *gag* and *env*. For patients with known infection dates, the estimated dates of the coalescent events obtained using molecular clock calculations based on a newly developed Bayesian method in *gag + env* were in agreement with the actual infection dates.

Key words: HIV-1 — Transmission network — Maximum likelihood — Bayesian — Phylogenetic confidence — Molecular clock

Introduction

Phylogenetic analyses have been used extensively in the molecular epidemiology of HIV-1: for tracing the epidemic spread worldwide (<http://www.unaids.org>, [Holmes et al. 1995]), for dating the origin of the HIV-1 epidemic (Korber et al. 2000; Salemi et al. 2001), or for investigating the HIV-1 transmission in several cases of uncertain history (Ou et al. 1992; Holmes et al. 1993; DeBry et al. 1993; Hillis and Huelsenbeck 1994; Albert et al. 1994; Blanchard et al. 1998; Metzker et al. 2002). Additionally, phylogenetic analyses were used to test the ability of different

¹Isolated sequences from patients belonging to the CRF04_cpx transmission network always correspond to partially characterized *gag*, *env*, and *gag + env* genomic regions.

Table 1. Characteristics of the study population

Isolate name ^a	Infection date ^b	Date of first HIV+ sample	Risk transmission	Source	Isolation date
m2.6.1999	12/89	10/1990	Heterosexual	f1.5	1999
m2.5.1999	6/89	4/1990	Heterosexual	f1.4	1999
f1.4.1999	6/89	12/1991	Heterosexual		1999
f1.2.1999	1/93	8/1993	Heterosexual		1999
f1.5.1991	12/88	7/1991	Heterosexual		1991
f1.5.1999	12/88	7/1991	Heterosexual		1999
f1.3.1999	6/92	12/1993	Heterosexual		1999

^am and f denote males and females, respectively, and the suffix indicates generation spread, identification number, and isolation date. For example, m2.6.1999 corresponds to male case number 6 belonging to the second generation of HIV-1 infection, from whom viral isolates were sampled in 1999.

^bAll the infection dates were according to personal interview of the patients.

methods to reconstruct a known HIV-1 transmission network (Leitner and Albert 1996) and to estimate the HIV-1 nucleotide substitution rate (Leitner et al. 1997; Leitner and Albert 1999).

Phylogenies of known transmission histories provide a framework to improve our knowledge about the evolution of the HIV-1 virus after sequential transmissions and, thus, a way to trace epidemiological relationships in cases with unknown links. Given a number of uncertainties, most notably that the transmission of a population of viruses makes it unclear what the “true tree” really is (Leitner and Fitch 1999), it is important to know the levels of discordance between the evolutionary tree (inferred from DNA sequences) and the transmission tree (based on epidemiological information) using different phylogenetic inference methods, as well as how they are affected by the sequence data in hand. Until now, this issue has been addressed only once in HIV-1 viral sequences with known transmission histories (Leitner and Albert 1996).

In this study we phylogenetically analyzed a known HIV-1 transmission network belonging to CRF04_cpx in the Thessalonica area in northern Greece (Paraskevis et al. 2002; Papa et al. 2002), for which the transmission events and dates were accurately known. CRF04_cpx, initially designated subtype I (Kostrikis et al. 1995), is one of the 16 HIV-1 circulating recombinant forms representing recombinant HIV-1 genomes that have infected three or more persons who are not epidemiologically related (Robertson et al. 2000). CRF04_cpx was documented in Cyprus and Greece and it was found to be comprised of subtypes of at least five distinct groups, A, G, H, K, and unclassified regions (Gao et al. 1998; Nasioulas et al. 1999; Paraskevis et al. 2001).

More specifically, we compared the levels of discordance between the inferred phylogenies and the transmission history in different genomic regions, using maximum likelihood (ML) and Bayesian methods; we examined whether phylogenetic discordance in different genomic regions could be due to

recombination; and, finally, we investigated how accurately we could estimate the dates of the transmission events using a Bayesian method.

Materials and Methods

Study Population. Viral strains were isolated from six HIV-1-infected individuals as shown in Table 1. The history of transmission events for all the HIV-1 patients infected by a common source is shown in Fig. 1.

RNA Extraction and cDNA Synthesis. RNA was extracted from plasma samples using the Total RNA Isolation Kit (Ambion Inc., Austin, TX) and cDNA was synthesized using the GeneAmp RNA PCR Kit (Perkin Elmer, USA), according to the manufacturer’s recommendations.

PCR. Two fragments of approximately 400 and 660 bp from the C2–C4 of gp120 and p17/5’ p24 of *gag*, respectively, were amplified by a single-step PCR using HIV-1 specific primers as described previously (Paraskevis et al. 2001).

Cloning and Nucleotide Sequencing. The PCR products from partial *gag* were cloned into the pCRII plasmid (TA Cloning Kit; Invitrogen Corp., San Diego, CA). An individual clone and the PCR products for partial *gag* and partial *env*, respectively, were sequenced on a VGI automated DNA sequencer using the Cy5/Cy5.5 Dye Primer Kit (Visible Genetics Inc., Toronto, Ontario, Canada), according to the manufacturer’s instructions.

Sequence Analysis. DNA sequence alignment was performed using Clustal W version 1.74 (Thompson et al. 1994). The best-fitting nucleotide substitution model was chosen according to the Akaike information criterion (AIC; Akaike 1973) among 64 different models using the Modeltest (Posada and Crandall 1998) and PAUP*4.0b10 programs (Swofford 1998). Phylogenetic analysis was accomplished using ML with the best-fitting evolutionary model as implemented in PAUP*. Bootstrapping was performed using neighbor joining or ML (1000 replicates) to assess the reliability of the obtained topologies. Phylogenetic trees were also obtained using Bayesian inference with the general time reversible (GTR) model including a Γ distribution rate heterogeneity among sites as implemented in MrBayes v. 3.0 (Huelsenbeck et al. 2001). For Bayesian inference, four Markov chains run for 10^6 generations with burn-in = 2×10^4 were used to

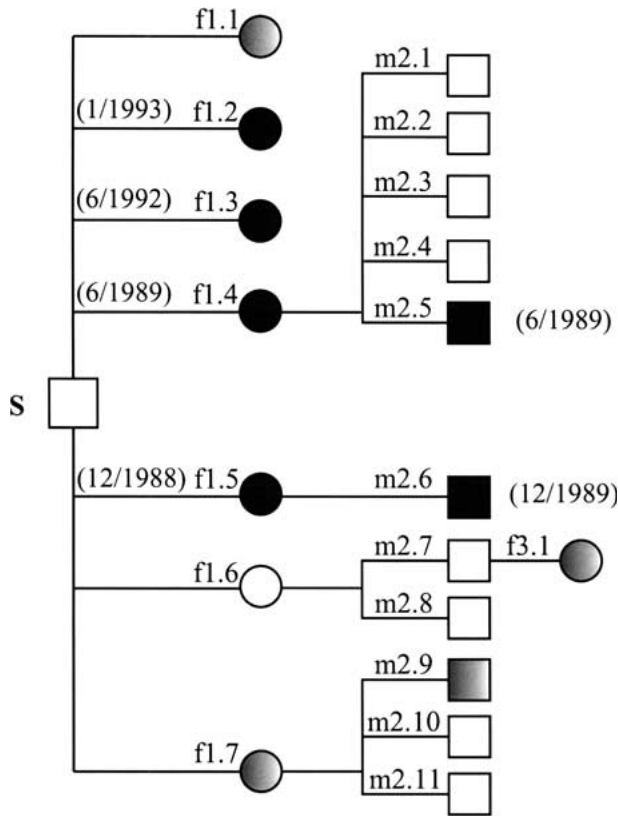


Fig. 1. Schematic representation of the three-generation HIV-1 transmission network from a single source (S). Squares and circles denote males (m) and females (f), respectively. The first and second suffixes denote the generation of the transmission and the sequential number of the patients, respectively. Filled symbols correspond to patients for whom biological samples were available and HIV RNA was detectable in plasma, whereas gray symbols correspond to samples in which HIV RNA was undetectable. Transmission dates are shown in parentheses.

reconstruct the consensus tree. Dissimilarities between the transmission tree and the inferred topologies, the ML and Bayesian inferred trees, as well as between the ML topologies inferred in *gag* and *env* were quantified using the quartets option in COMPONENT (Page 1993). To assess for any significant differences between candidate topologies in the *gag* and *env* regions, we used the approximately unbiased (AU) test (Shimodaira 2000) as implemented in CONSEL (Shimodaira and Hasegawa 2001). For each region, sitewise log-likelihoods were estimated for the candidate trees using PAUP* with the best-fitting evolutionary model.

Simulations. Nucleotide sequences in partial *gag* and *env* were simulated (100 replicates) over the ML estimated branch lengths and evolutionary parameters, as implemented in the Evolver program of PAML (Yang 1997).

Molecular Clock Calculations. Estimation of the substitution rate was performed by inferring simultaneously population and substitution parameters and tree topology using Bayesian inference as implemented in BEAST v. 1.0.3 (Drummond et al. 2002; Drummond and Rambaut 2003) and using a different substitution and evolutionary model (GTR+ Γ) for *gag* and *env* partitions. Five separate MCMC runs were made for 5×10^6 generations, with a burn-in of 5×10^5 .

Results

Viral Isolates. In the current study, we characterized six newly identified CRF04_cpx isolates in addition to the four sequences which have been described previously (Kostrikis et al. 1995; Gao et al. 1998; Nasioulas et al. 1999; Paraskevis et al. 1999) (Table 1). The virus isolated from individual f1.5 in 1991 was classified as CRF04_cpx according to phylogenetic analysis of the partial reverse transcriptase (RT) region, which was initially examined for monitoring genotypic resistance to antiretroviral drugs.

Transmission Network. Subject f1.5 attended the AIDS Clinic at AHEPA University Hospital in Thessalonica, and according to a detailed epidemiological contact tracing, she belonged to a transmission network of 7 females that were infected with HIV-1 by a common source and further spread the infection to at least 12 persons. All these 19 subjects acquired the HIV-1 infection heterosexually. The history of the transmission events was reconstructed by interviews with physicians and is depicted in Fig. 1. Overall the transmission events consisted of three distinct “spread generations” among 7 females that were initially infected by a common source (first generation), 11 males (second generation) that acquired the infection by the females of the previous generation, and 1 female (third generation) that was infected from a male from the second generation (Fig. 1). The dates of the first documented HIV-positive sample of the patients included in this study are shown in Table 1. The “source” subject was a male infected by an unknown source, who was hospitalized for the first time in 1993 with progressive multifocal encephalopathy and died 1 month later. The HIV-1-seropositive status was initially documented in 1993 during the hospitalization period. Blood samples for PCR amplification were available for 10 subjects, whereas the rest of them either had died or were no longer attending AHEPA University Hospital. Among the 10 individuals for whom blood samples were available, PCR amplification was attempted for only 6 for which the HIV-RNA was detectable in plasma. Among these six individuals, PCR was unsuccessful for f1.4.1999 in partial *gag* and, thus, excluded from phylogenetic analysis. HIV-1 DNA sequences were obtained directly from the PCR-amplified products in the C2-V3 region of *env* or after DNA cloning for partial *gag* spanning p17 and the 5'-terminus of p24.

Phylogenetic Analyses. Figure 2 depicts the transmission history of the CRF04_cpx sequences as reconstructed from the transmission events including information on isolation dates of viral samples. We investigated the levels of discordance between trans-

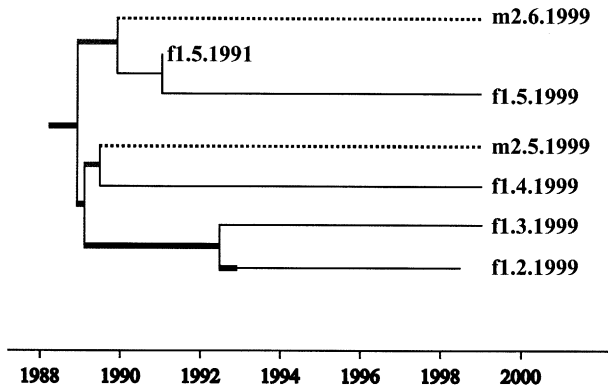


Fig. 2. Tree representing the true transmission history of viral samples collected within the transmission network according to a detailed epidemiological tracing of HIV-1 infections. Internal branches in black represent first-generation transmission events by the source. Dashed lines indicate branches corresponding to viral sequences from patients belonging to the second generation of infections. The first part of the fl.2.1999 branch, until 1993, represents the history of the virus in the source.

mission and evolutionary trees constructed by phylogenetic analyses using partial *gag* (660 nt), *env* (327 nt), and the concatenated alignment of *gag + env*. Phylogenetic trees were inferred by ML using the best-fitting model (Table 2) and Bayesian inference using the general time reversible (GTR) model including Γ -distributed rates heterogeneity among sites.

According to the phylogenetic analysis, in partial *gag* using ML and Bayesian methods, the dissimilarity between the quartets of the transmission and those of the inferred tree was 0.87 (Table 3). More specifically, one misclassified branch between fl.5.1999 and m2.5.1999 was obtained, supported by 72% bootstrap values using ML (Fig. 3A) and a Bayesian posterior probability (PP) of 0.95 (Fig. 3B). In phylogenetic trees inferred by both methods there was an additional misclassified branch between fl.3.1999 and m2.6.1999, albeit with low support (Fig. 3A and B). Phylogenetic analysis in partial *env* showed a lower dissimilarity index between the transmission and the inferred tree than in partial *gag* (0.47 vs 0.87, respectively; Table 3) where the link between the donor (fl.5.1999) and the recipient (m2.6.1999) was accurately reconstructed, whereas a misclassified branch was obtained between fl.3.1999 and m2.5.1999 (Fig. 3C and D). The fl.5.1999/m2.6.1999 cluster inferred by ML received 80 and 64% bootstrap support using NJ and ML methods (BP_{ML} and BP_{NJ}), respectively, similarly to as the misclassified branch between fl.3.1999 and m2.5.1999 (Figure 3C).

Unlike partial *gag* and *env*, the combined *gag + env* region resulted in an improved estimate of the transmission history (Fig. 3E and F) indicated by the diminished dissimilarity score of 0.27 (Table 3). The only discrepancy between the inferred and the

Table 2. Best-fitting evolutionary model for different genomic regions including a different set of sequences

Genomic region	Model selected ^a	α^b	I ^c
p17/5' p24 (<i>gag</i>)	TrN + I	—	0.7
C2–V3 (<i>env</i>)	TVM + G	0.67	—
<i>gag + env</i>	TVM + G	0.3	—

^aGTR is the general time reversible model; TVM and TrN are GTR submodels assuming that the rates of A–G, C–T and of A–C, A–T, G–C, G–T are equal, respectively. G and I indicate models allowing Γ -distributed rates across sites and a proportion of invariable sites, respectively.

^bShape parameter of the Γ distribution of rates among sites.

^cProportion of invariable sites.

transmission tree was the indirect relationship between the two noncontemporaneous sequences isolated from a single individual (fl.5.1991, fl.5.1999). The donor–recipient relationship between fl.5.1999 and m2.6.1999 was accurately reconstructed, albeit with low support (Fig. 3E and F). The low levels of support between fl.5.1999 and m2.6.1999 (posterior probability of 0.43 and 32% bootstrap by ML) (Fig. 3E and F) could be explained due to the very short time interval (12 months), in which the infection of patients fl.5.1999, m2.6.1999, and m2.5.1999 occurred (Table 1), thus suggesting that the evolutionary distance between the viral sequences at that time was very small. This is in accordance with the very short branch separating the node of the clade fl.5.1999–m2.6.1999 with that of m2.5.1999, as well as the high levels of PP support (0.99) for the common clade connecting fl.5.1999, m2.6.1999, and m2.5.1999 (Fig. 3F).

We should note that in the two regions (*gag* and *env*) examined here, along with the combined data set of both regions, ML and Bayesian methods were highly correlated as indicated by comparisons among all possible quartets using COMPONENT (Table 3).

The confidence of the inferred topology in *gag + env*, reflected by the number of trees within 95% CI, was higher than in phylogenetic trees estimated in *gag* and *env* separately, where a higher number of inferred trees fell within the 95% CI (3 vs 12 and 10, respectively) (Fig. 3B, D, and F). Unlike ML, which infers the single most likely tree, the Bayesian method provides the best-set trees, which means that given a large enough number of generations, the frequency of sampled trees reflects the frequency of those trees in the posterior probability distribution. Thus, the posterior probability of any sampled topology is the probability of this particular topology conditional on the data (Huelsenbeck et al. 2001). Finally, a credible set of the trees (95% CI) can be set by the number of trees with the highest posterior probability that sum up to 0.95. Thus, for a genomic region, the smaller the number of trees with cumulative posterior prob-

Table 3. Quantitative quartet comparisons between the transmission tree and the maximum likelihood (ML) estimated trees (Fig. 2) and between the ML and the Bayesian inferred trees in partial *gag*, *env*, and *gag + env*

Genomic region	Number of all possible quartets	Identical quartets (s)	Resolved and different quartets (d)	Dissimilarity ^a
Comparisons between transmission tree and ML inferred topologies				
p17/5' p24 (<i>gag</i>)	15	2	13	0.87
C2-V3 (<i>env</i>)	15	8	7	0.47
<i>Gag + env</i>	15	11	4	0.27
Comparisons between ML and Bayesian inferred trees				
p17/5' p24 (<i>gag</i>)	15	11	0	0
C2-V3 (<i>env</i>)	15	11	0	0
<i>gag + env</i>	15	15	0	0

^aRatio of dissimilar to total number of quartets.

ability 0.95, the higher the confidence of the inferred topologies for this particular region.

Testing for Phylogenetic Discordance Between gag and env. The significance of dissimilar topologies in partial *gag* and *env* was tested, also, by the approximately unbiased (AU) test (Simodaira et al. 2000), according to which the topologies for *gag* and *env* were significantly rejected for *env* and *gag* alignment regions, respectively ($p < 0.05$) (Table 4). Phylogenetic discordance between *gag* and *env* may have been caused by recombination between different viral lineages in the donor or by other means, such as differential selective pressure between *gag* and *env*, etc. Interestingly, we found similarly that in the Swedish transmission network (Leitner et al. 1996), the topologies of *env* and *gag* were significantly rejected for the *gag* and *env* alignment regions, respectively ($p < 0.05$). However, similarly to the results we present here, the similarity score between the inferred and the transmission trees increased for *gag*, *env*, and *gag + env*, respectively.

To examine further whether the high similarity index in *gag + env* could be observed by chance, by combining the *gag* and *env* regions, we concatenated 100 simulated data sets for partial *gag* and *env* and then phylogenetically analyzed them separately using the ML method. In this case, the consensus tree (based on the inferred topologies of the 100 concatenated data sets) was identical to the topology inferred in *env* but not in *gag + env*: The f1.5.1999/m2.6.1999 cluster was observed at a frequency of 84%, and the m2.5.1999/f1.3.1999 and f1.5.1991/f1.2.1999 clusters at a frequency of 73 and 62%, respectively. Thus, after concatenating two data sets simulated according to the topology and branch lengths inferred in *gag* and *env*, the phylogenetic signal is biased toward the more dominant gene signal (partial *env*).

Molecular Clock Calculations. To estimate the date of the root of the tree and to examine whether the divergence dates of the coalescent events corre-

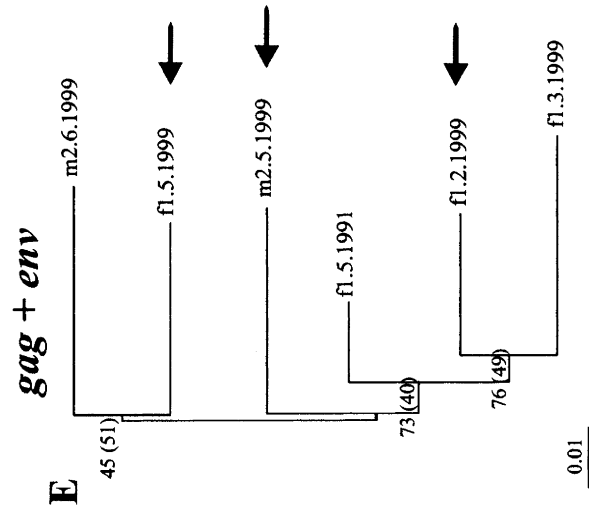
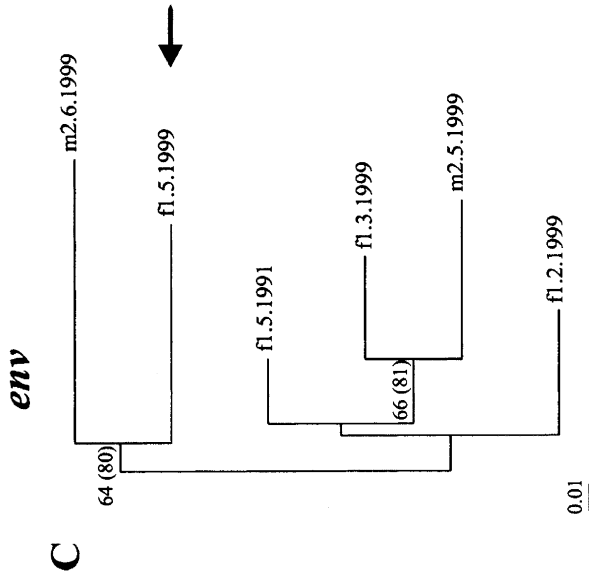
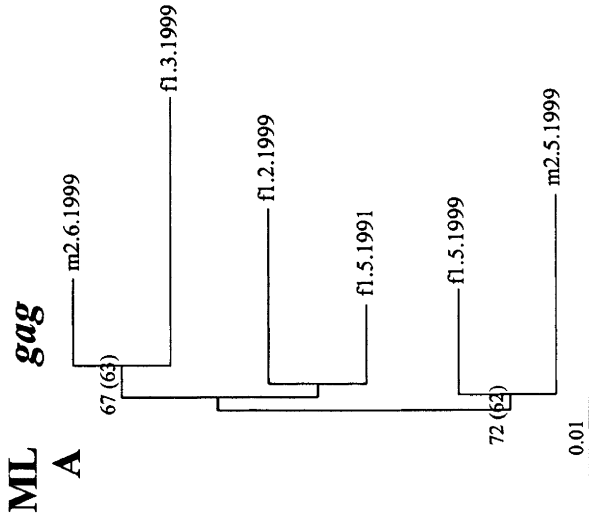
spond to the transmission dates, we estimated the substitution rate by inferring simultaneously population, substitution parameters, and tree topology using Bayesian inference as implemented in BEAST v. 1.0.3 (Drummond et al. 2002; Drummond and Rambaut 2003). More specifically, the divergence dates were estimated in the topology inferred in *gag + env* using a different substitution and evolutionary model (GTR + Γ) for *gag* and *env* partitions. All MCMC independent runs converged to almost-identical values for all parameters (data not shown). The mean substitution rates (combined runs) for *gag* and *env* were estimated as 2.48×10^{-3} (95% highest posterior density interval [HPD]: 6.3×10^{-4} – 4.67×10^{-3}) and 6.55×10^{-3} (95% HPD: 1.88×10^{-3} – 1.21×10^{-2}) substitutions per site per year, respectively, which were almost identical to those estimated previously for the Swedish transmission network (Leitner and Albert 1999). The mean date of the most recent common ancestor (MRCA) of the tree (t_{root}) was estimated as 1985.1 (95% HPD: 1973.5–1991).

Interestingly, for patients with known infection dates the coalescent events were accurately dated: the f1.5.1999/m2.6.1999 node as 1987.5 (95% HPD: 1974.4–1995), which was close to the infection date (12/1989) (Table 1), and the f1.2.1999/f1.3.1999 cluster as 1989.3 (95% HPD: 1981.6–1995.3) (infection date of f1.3: 6/1992). In both cases, the MRCA estimates predated their respective transmissions but these were contained in credible intervals.

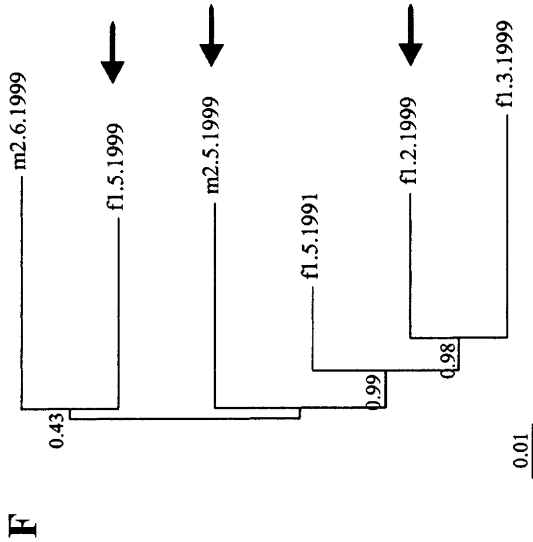
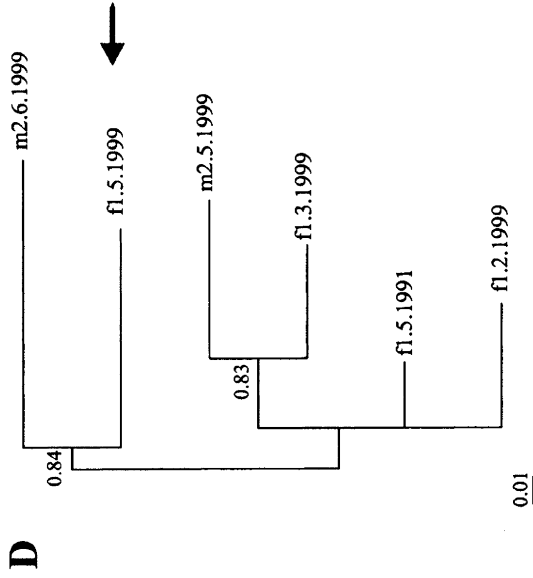
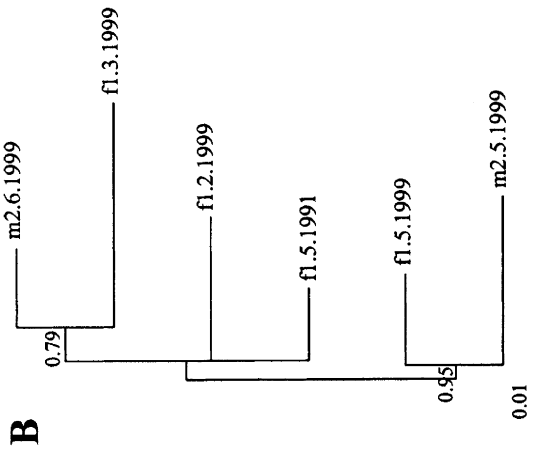
Moreover, the MRCA of the f1.5.1999/m2.6.1999/m2.5.1999 cluster was estimated as 1986.9 (95% HPD: 1973.6–1994.9), and for the f1.2.1999/f1.3.1999/f1.5.1991 branch it was estimated as 1986.6 (95% HPD: 1979–1991).

Discussion

In this study we examined the levels of discordance between the transmission and the evolutionary trees, inferred by ML and Bayesian methods in different genomic regions, for HIV-1 CRF04_cpx sequences isolated from patients belonging to a well-docu-



Bayesian



Number of trees within 95% CI: 12

Number of trees within 95% CI: 10

Number of trees within 95% CI: 3

Table 4. AU test for assessing the significance of different trees in different pieces of the alignment

Piece of alignment	Examined tree ^a	$\Delta \ln L^b$	AU test <i>p</i> value
<i>env</i>	<i>gag</i>	10	0.017
<i>gag</i>	<i>env</i>	7.5	0.038

^aFor each fragment, the maximum likelihood inferred topology using the best-fitting evolutionary model was examined.

^bDifferences in likelihood between the two candidate trees.

mented transmission network. Furthermore, we explored the possibility of recombination between *gag* and *env* and its potential to bias phylogenetic inference in the concatenated alignment.

All the HIV-1 sequences isolated from patients that, according to their personal interview, were infected by a common source were classified as CRF04_cpx, in agreement with epidemiological information for a common source of infection. The transmission network included at least 19 HIV-infected subjects, of whom 6 were characterized by phylogenetic analysis to belong to the CRF04_cpx group. For all these patients, except for the f1.5.1991 sample, viral RNA was isolated approximately 10 years after the infection dates that occurred within the time period between December 1988 and December 1989. The only exceptions were patients f1.2 and f1.3, who were infected in 1993 and 1992, respectively. Thus, most of the infections occurred in a short period of time, and viral sequences were isolated distantly from infection dates; therefore it was challenging to investigate whether the transmission history was in accordance with the evolutionary tree inferred by phylogenetic analysis.

Using two different approaches we found that the similarity index between the transmission tree and the inferred topologies increased for partial *gag*, *env*, and *gag+env*, which was in accordance with a previous study (Leitner and Albert 1996). These findings could be explained by the fact that in *env*, which evolves faster than *gag*, several substitutions were possibly accumulated between the transmission events, thus rendering the *env* region phylogenetically more informative than *gag*.

The inferred topology in *gag+env* was 73% identical to the transmission tree, which was in accordance with the Swedish transmission network (Leitner and Albert 1996). This kind of incongruence

might have been caused by any differences between the viral phylogeny and the transmission events. For instance, branch nodes that reflect the separation of different viral lineages do not necessarily correspond with the time period of transmission events (Leitner and Fitch 1999). In our case, it seems that the viral lineage sampled in 1991 from patient f1.5 (f1.5.1991) was distantly related to the virus sampled in 1999 from the same individual (f1.5.1999) and the virus transmitted to her recipient (m2.6.1999), a finding which is consistent with an earlier estimated coalescent event of the f1.5.1991/f1.2.1999/f1.3.1999 branch node compared to the f1.5.1999/m2.6.1999 node.

Thus our findings based on phylogenetic analysis using ML and Bayesian methods provide evidence that the transmission history is almost entirely compatible with the evolutionary tree of the virus, given that enough phylogenetic signal is present in the alignment. We should stress that in our case, first, most of the transmission events occurred in a very short time interval, and second, viral sequences were isolated distantly from the infection dates, thus probably rendering the reconstruction of the transmission tree rather complex. One of the issues that still remains open is how the ability of phylogenetic methods to identify clusters of epidemiologically linked individuals among unrelated patients is affected by these parameters.

Interestingly, we found evidence for significant phylogenetic discordance between the *gag* and the *env* regions, in a similar way as in the Swedish transmission network, which might have been caused by recombination or other mechanisms. To examine whether the high similarity score between the transmission and the inferred tree in *gag+env* could be observed by chance, which would be the case if recombination had occurred between *gag* and *env*, thus biasing the phylogenetic inference in the concatenated alignment, we performed the following analysis. We reconstructed the phylogeny in *gag+env* for 100 replicates simulated according to the evolutionary trees in *gag* and *env*, separately. In this case, we found evidence that the particular topology in *gag+env* was not obtained by combining two random segments simulated according to the evolutionary process in *gag* and *env*. Thus, although the hypothesis of recombination cannot be entirely excluded, it provides a less parsimonious explanation for the phylogenetic incongruence between *gag* and *env*. On the other hand, differential selective pressure in these genes might have caused this discrepancy, or alternatively the limited number of accumulated mutations in *gag*, within the short time interval in which most of the infections occurred, may seriously diminish the ability to infer the evolutionary process in this region correctly. The latter is in accordance with the relatively low confidence of the phylogenetic tree in partial *gag*.

Fig. 3. Phylogenetic analysis (A, B) in partial *gag*, (C, D) in partial *env*, and (E, F) in *gag+env* using ML with the best-fitting evolutionary model (tree in upper part) or Bayesian inference (tree in lower part). The numbers at the tree nodes (upper part) not in and in parentheses indicate bootstrap values obtained from 1000 replicates using ML and NJ, respectively. Numbers at the tree nodes inferred by the Bayesian method indicate posterior probabilities. Arrows indicate clusters that match the transmission tree.

Given that the real tree is unknown in most of the phylogenetic reconstruction problems, the way that the different measures of phylogenetic confidence reflect accuracy is of crucial importance, since in molecular phylogenetics, apart from the levels of confidence of inferred trees, we also need to know how close we are to the truth. Interestingly, in our case the overall confidence of the inferred trees, reflected by the number of trees within 95% CI, gradually increased in *gag*, *env*, and *gag + env* and directly correlated with the similarity score between inferred topologies and the transmission tree. Assuming that we are closer to the true tree in *gag + env*, this means in turn that the overall confidence of the tree may provide a better estimate of phylogenetic accuracy than PP or BP_{ML} alone, which both may be misleading for assessing phylogenetic accuracy, especially in genomic regions where there is limited phylogenetic signal. However, it is important to stress that in the case of a transmission network the evolutionary tree may not be entirely compatible with the evolutionary tree, thus suggesting that we cannot draw general conclusions about how phylogenetic confidence reflects accuracy in this case.

Dating of the transmission events, using a recently developed Bayesian method, was accurate at least in two cases where direct comparisons could be made between estimated and actual transmission events, whereas the confidence intervals of the estimated dates were wide. In our case we observed also differences between the estimated and the actual transmission events—pretransmission interval—as described previously (Leitner and Albert 1999). This can be explained due to the population diversity in the donor patient, thus suggesting that timing of a divergence between sequences in two different patients can only be a maximum estimate of transmission time (Shankarappa et al. 1999).

Acknowledgments. We wish to acknowledge Jan Albert and Tomas Leitner for their supportive comments and suggestions about this study and Constantinos-Simos Nistiots for statistical advice. D.P. was supported by a Marie Curie fellowship from the European Commission (QLK2-CT2001-51062) and by the Hellenic Center for Infectious Disease Control (KEEL). P.L. was supported by the Flemish Institute for Scientific-Technological Research in Industry (IWT). This work was supported by the Hellenic Scientific Society for the Study of AIDS and Sexually Transmitted Diseases and in part by the Flemish Fonds voor Wetenschappelijk Onderzoek (FWO G.0288.01).

References

Akaike H (1973) Information theory as an extension of the maximum likelihood principle. In: Petrov F, Csaki BN (eds) Second International Symposium on Information Theory. Akademiai Kiado, Budapest, pp 267–281

- Albert J, Wahlberg J, Leitner T, Escanilla D, Uhlén M (1994) Analysis of a rape case by direct sequencing of the human immunodeficiency virus type 1 *pol* and *gag* genes. *J Virol* 68:5918–5924
- Blanchard A, Ferris S, Chamaret S, Guetard D, Montaner L (1998) Molecular evidence for nosocomial transmission of human immunodeficiency virus from a surgeon to one of his patients. *J Virol* 72:4537–4540
- DeBry RW, Abele LG, Weiss SH, Hill MD, Bouzas M, Lorenzo E, Graebnitz F, Resnick L (1993) Dental HIV transmission? *Nature* 361:691
- Douady CJ, Delsuc F, Boucher Y, Doolittle WF, Douzery EJ (2003) Comparison of bayesian and maximum likelihood bootstrap measures of phylogenetic reliability. *Mol Biol Evol* 20:248–254
- Drummond AJ, Rambaut A (2003) BEAST v1.0; available from <http://evolve.zoo.ox.ac.uk/beast/>
- Drummond AJ, Nicholls GK, Rodrigo AG, Solomon W (2002) Estimating mutation parameters, population history and genealogy simultaneously from temporally spaced sequence data. *Genetics* 161:1307–1320
- Gao F, Robertson DL, Carruthers CD, Li Y, Bailes E, Kostrikis LG, Salminen MO, Bibollet-Ruche F, Peeters M, Ho DD, Shaw GM, Sharp PM, Hahn BH (1998) An isolate of human immunodeficiency virus type I represents a complex mosaic comprising three different group M subtypes (A, G, and I). *J Virol* 72:10234–10241
- Hillis DM, Huelsenbeck JP (1994) Support for dental HIV transmission. *Nature* 369:24–25
- Holmes EC, Zhang LQ, Simmonds P, Rogers AS, Leigh Brown AJ (1993) Molecular investigation of human immunodeficiency virus (HIV) infection in a patient of an HIV-infected surgeon. *J Infect Dis* 167:1411–1414
- Holmes EC, Zhang LQ, Robertson P, Cleland A, Harvey E, Simmonds P, Brown AJL (1995) The molecular epidemiology of human immunodeficiency virus type 1 in Edinburgh. *J Infect Dis* 171:45–53
- Huelsenbeck JP, Ronquist F, Nielsen R, Bollback JP (2001) Bayesian inference of phylogeny and its impact on evolutionary biology. *Science* 294:2310–2314
- Korber B, Muldoon M, Theiler J, Gao F, Gupta R, Lapedes A, Hahn BH, Wolinsky S, Bhattacharya T (2000) Timing the ancestor of the HIV-1 pandemic strains. *Science* 288:1789–1796
- Kostrikis LG, Bagdades E, Cao Y, Zhang L, Dimitriou D, Ho DD (1995) Genetic analysis of human immunodeficiency virus type 1 strains from patients in Cyprus: Identification of a new subtype designated subtype I. *J Virol* 69:6122–6130
- Leitner T, Albert J (1996) Accurate reconstruction of a known HIV-1 transmission history by phylogenetic tree analysis. *Proc Natl Acad Sci USA* 93:10864–10869
- Leitner T, Albert J (1999) The molecular clock of HIV-1 unveiled through analysis of a known transmission history. *Proc Natl Acad Sci USA* 96:10752–10757
- Leitner T, Fitch W (1999) The phylogenetics of known transmission histories. In: Crandall KA (ed) *The evolution of HIV*. Johns Hopkins University Press, Baltimore, Maryland, pp 315–345
- Leitner T, Kumar S, Albert J (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
- Metzker ML, Mindell DP, Liu XM, Ptak RG, Gibbs RA, Hillis DM (2002) Molecular evidence of HIV-1 transmission in a criminal case. *Proc Natl Acad Sci USA* 99:14292–14297
- Nasioulas G, Paraskevis D, Magiorkinis E, Theodoridou M, Hatzakis A (1999) Molecular analysis of the full-length genome of the HIV-1 “Subtype I”: evidence of triple recombination. *AIDS Res Hum Retroviruses* 15:745–758

- Ou CY, Ciesielski CA, Myers G et al. (1992) Molecular epidemiology of HIV transmission in a dental practice. *Science* 256:1165–1171
- Page RDM (1993) COMPONENT, version 2.0. Biogeography and Conservation Laboratory, Natural History Museum, London
- Pappa A, Papadimitriou E, Papoutsis A, Kiosses V, Antoniadis A (2002) HIV-1 subtypes and circulating recombinant forms (CRFs) in northern Greece. *Virus Res* 85:85–93
- Paraskevis D, Magiorkinis E, Theodoridou M, Mostrou G, Papaevangelou V, Kiosses VG, Hatzakis A, Matsaniotis N (1999) Molecular epidemiology of vertical HIV-1 transmission in Greece. *J Hum Virol* 2:339–343
- Paraskevis D, Magiorkinis M, Vandamme A-M, Kostrikis LG, Hatzakis A (2001) Re-analysis of human immunodeficiency virus type 1 isolates from Cyprus and Greece, initially designated as ‘subtype I’, reveals a unique complex A/G/H/K/? mosaic pattern. *J Gen Virol* 82:575–580
- Paraskevis D, Magiorkinis E, Magiorkinis G, Kiosses VG, Theodoridou M, Papatizos V, Hatzakis A (2002) Molecular epidemiology of the HIV-1 CRF04_cpx initially designated as “subtype I” in Greece. Poster presentation at the XIV International AIDS Conference
- Posada D, Crandall KA (1998) MODELTEST: Testing the model of DNA substitution. *Bioinformatics* 14:817–818
- Rambaut A (2000) Estimating the rate of molecular evolution: Incorporating noncontemporaneous sequences into maximum likelihood phylogenies. *Bioinformatics* 16:395–399
- Robertson DL, Anderson JP, Bradac JA, Carr JK, Foley B, Funkhouser RK, Gao F, Hahn BH, Kalish ML, Kuiken C, Learn GH, Leitner T, McCutchan F, Osmanov S, Peeters M, Pieniazek D, Salminen M, Sharp PM, Wolinsky S, Korber B (2000) HIV-1 nomenclature proposal. *Science* 288:55–56
- Salemi M, Strimmer K, Hall WW, Duffy M, Delaporte E, Mboup S, Peeters M, Vandamme A-M (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
- Shankarappa R, Margolick JB, Gange SJ, Rodrigo AG, Upchurch D, Farzadegan H, Gupta P, Rinaldo CR, Learn GH, He X, Huang XL, Mullins JI (1999) Consistent viral evolutionary changes associated with the progression of human immunodeficiency virus type 1 infection. *J Virol* 73:10489–10502
- Shimodaira H, Hasegawa M (2001) CONSEL: For assessing the confidence of phylogenetic tree selection. *Bioinformatics* 17:1246–1247
- 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, Palo Alto, CA.
- Suzuki Y, Glazko GV, Nei M (2002) Overcredibility of molecular phylogenies obtained by Bayesian phylogenetics. *Proc Natl Acad Sci USA* 99:16138–16143
- Swofford DL (1998) PAUP*. Phylogenetic Analysis Using Parsimony (*and other methods), Version 4. Sinauer Associates, Sunderland, MA
- Thompson JD, Higgins DG, Gibson TJ (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
- Whittingham LA, Slikas B, Winkler DW, Sheldon FH (2002) Phylogeny of the tree swallow genus, *Tachycineta* (Aves: Hirundinidae), by Bayesian analysis of mitochondrial DNA sequences. *Mol Phylogenet Evol* 22:430–441
- Yang Z (1997) PAML: A program package for phylogenetic analysis by maximum likelihood. *Comput Appl Biosci* 13:555–556