JOURNAL OF VIROLOGY, Sept. 2005, p. 11981–11989 Vol. 79, No. 18
0022-538X/05/$08.00+0 doi:10.1128/JVI.79.18.11981–11989.2005

Molecular Footprint of Drug-Selective Pressure in a Human Immunodeficiency Virus Transmission Chain†

Philippe Lemey,[1,2]* Inge Derdelinckx,[2] Andrew Rambaut,[1] Kristel Van Laethem,[2] Stephanie Dumont,[2] Steve Vermeulen,[2] Eric Van Wijngaerden,[3] and Anne-Mieke Vandamme[2]

Department of Zoology, University of Oxford, South Parks Road, Oxford OX1 3PS, United Kingdom[1]; Rega Institute for Medical Research, KULeuven, Minderbroedersstraat 10, B-3000 Leuven, Belgium[2]; and University Hospitals Leuven, Herestraat 49, B-3000 Leuven, Belgium[3]

Received 17 March 2005/Accepted 23 June 2005

Known human immunodeficiency virus (HIV) transmission histories are invaluable models for investigating the evolutionary and transmission dynamics of the virus and to assess the accuracy of phylogenetic reconstructions. Here we have characterized an HIV-1 transmission chain consisting of nine infected patients, almost all of whom were treated with antiviral drugs at later stages of infection. Partial *pol* and *env* gp41 regions of the HIV genome were directly sequenced from plasma viral RNA for at least one sample from each patient. Phylogenetic analyses in *pol* using likelihood methods inferred an evolutionary history not fully compatible with the known transmission history. This could be attributed to parallel evolution of drug resistance mutations resulting in the incorrect clustering of multidrug-resistant virus. On the other hand, a fully compatible phylogenetic tree was reconstructed from the *env* sequences. We were able to identify and quantify the molecular footprint of drug-selective pressure in *pol* using maximum likelihood inference under different codon substitution models. An increased fixation rate of mutations in the HIV population of the multidrug-resistant patient was demonstrated using molecular clock modeling. We show that molecular evolutionary analyses, guided by a known transmission history, can reveal the presence of confounding factors like natural selection and caution should be taken when accurate descriptions of HIV evolution are required.

The evolution of human immunodeficiency virus (HIV) within and across hosts is known to be remarkably different (11, 31). Within hosts, the viral population is subjected to natural selection as a result of a continuous effort to evade the immune response. This process is frequently reflected in temporal phylogenetic structures showing the continual appearance and extinction of strains through time (11, 34). Across hosts, positive selective pressure does not seem to have an important impact. Instead, HIV genetic diversity is predominately shaped by spatial and temporal factors in the demographic history (11, 31). Understanding how intrahost evolution translates into HIV evolution at the population level is a key factor in determining how drug resistance and immune escape mutations might spread. Treating HIV-1-infected patients with highly active antiretroviral therapy (HAART) has led to a remarkable reduction in HIV-related morbidity and mortality (28). However, no antiretroviral is resistance proof, and when HIV replication is not fully suppressed, drug resistance inevitably appears. In countries where antiretrovirals are widely used, growing concern exists about the transmission of resistant viruses. Many efforts are currently being undertaken to monitor and control its spread. Accurately tracking HIV transmission in the population can provide useful data to investigate this issue. Only recently, an HIV transmission chain provided the proof of principle for transmission of drug resistance (38).

Well-characterized HIV transmission chains are much appreciated in a phylogenetic context. Assuming that the viral phylogeny should be consistent with the "true" transmission history, transmission chains allow assessment of the accuracy and reliability of phylogenetic reconstructions (19). Unlike simulation studies, which cannot fully capture the biological complexity of HIV evolution, transmission chains are informative on the performance of methods for real data. One of the most carefully studied examples involves a Swedish HIV-1 transmission chain consisting of nine individuals from whom 13 samples were obtained (19). Using sequences sampled in the *gag* and *env* genes, several tree reconstruction methods were tested for their ability to reconstruct the true transmission history. Except for one mother-to-child transmission event, the viral tree, reconstructed using maximum likelihood methods and realistic nucleotide substitution models, agreed with the known transmission history (19, 21). In addition, molecular clock analysis revealed that genetic divergence correlated well with the isolation dates of the samples and that significant genetic divergence, referred to as ancestral divergence, existed between the donor and recipient lineage at the time of transmission (17).

Although the characterization of a transmission chain of this scope—in terms of patients involved, separation times between samples, and sequence data obtained—is unique in the epidemic history of HIV, it has been taken as a strong argument for the accuracy of phylogenetic reconstructions (10, 19, 20). It remains, however, uncertain how robust evolutionary analyses

* Corresponding author. Present address: Department of Zoology, University of Oxford, South Parks Road, Oxford OX1 3PS, United Kingdom. Phone: 44 1865 271272. Fax: 44 1865 271249. E-mail: philippe.lemey@zoo.ox.ac.uk.

† Supplemental material for this article may be found at http://jvi.asm.org.

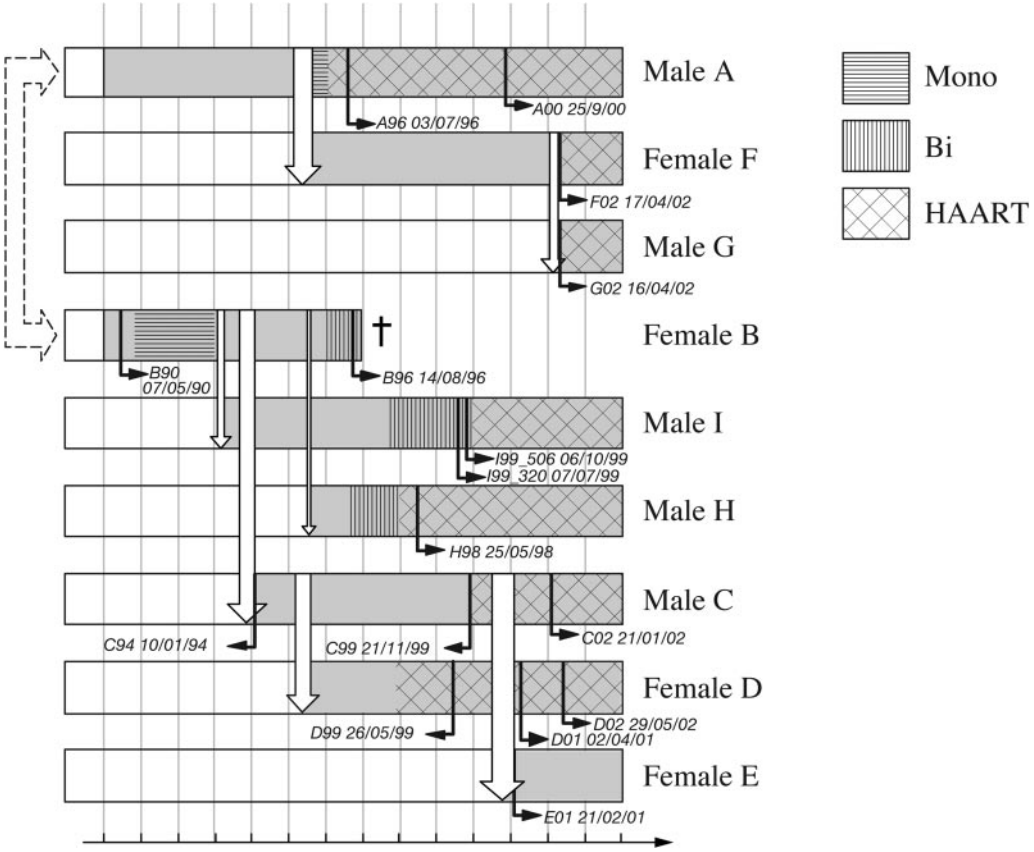

FIG. 1. Real-time HIV-1 transmission history as identified by contact tracing. The open arrows denote the transmission events; the width of the arrows represents the time interval for transmission. The arrow representing the transmission event between patient A and patient B is dashed because the time and direction of transmission could not be determined. The patient's bars filled with gray indicate an HIV-1-infected status with (superimposed) the therapy history. The therapy history is either monotherapy (Mono), bitherapy (Bi), or HAART. †, patient is deceased. The bent black arrows indicate available samples, including date of isolation.

are with respect to biological complications, like recombination and natural selection. Unfortunately, known transmission chains allowing us to address these issues are rare. Here, we present an HIV transmission cluster consisting of nine infected patients for whom the time and direction of each virus transmission were determined by in-depth patient interviews. Reconstructed phylogenies based on *pol* and *env* gp41 gene sequences were evaluated for their compatibility with the known transmission history. A particular clustering in the *pol* tree, topologically incongruent with the transmission history, is attributed to drug-selective pressure in a multidrug-resistant patient. Natural selection in the HIV transmission chain, and in the multidrug-resistant patient in particular, was investigated using codon substitution models; the impact on the evolutionary rate is demonstrated using molecular clock modeling.

MATERIALS AND METHODS

Study population. The epidemiological relationships between nine HIV-1-infected patients attending the University Hospitals Leuven were established through in-depth interviews by physicians experienced in HIV care. A time interval for each transmission event was determined by the following contact tracing criteria (if available): (i) the patient reporting a high-risk contact, (ii) the patient's most recent negative HIV test, and (iii) a history of an acute viral syndrome (which indicates an infection in the range of several days up to 10 weeks in the past). Each patient provided written informed consent, and at least one blood sample was obtained between 1990 and 2002. Epidemiologically unrelated control sequences, having the same subtype as the strains constituting the transmission chain, were obtained from a local database (local controls) and from GenBank using BLAST (2). Control sequences with drug resistance mutations were retrieved from the HIV Drug Resistance Database (http://hivdb.stanford.edu/).

RNA extraction, cDNA synthesis, amplification, and sequencing of the *pol* and *env* region. Plasma was isolated from the blood sample, and HIV RNA was extracted using a QIAamp Viral RNA Mini kit (Westburg, Leusden, The Netherlands). cDNA synthesis and PCR amplification of the *pol* gene region were performed using an in-house protocol (40). *env* gp41 cDNA synthesis and PCR amplification were performed as previously described (41). Direct sequencing of the purified nested PCR products was performed using the ABI PRISM BigDye Terminator v3.1 Ready Reaction Cycle Sequencing kit and analyzed on the ABI3100 genetic analyzer (Applera, Nieuwerkerk a/d Ijssel, The Netherlands). Sequence fragments of 1,069 bp for *pol* and 951 bp for *env* were assembled and analyzed using Sequence Analysis version 3.7 and SeqScape version 2.0 (Applera, Nieuwerkerk a/d Ijssel, The Netherlands).

Phylogenetic inference. Sequences were aligned using CLUSTAL X (39) and manually edited according to their codon-reading frame in Se-Al (http://evolve.zoo.ox.ac.uk). Regions that could not be unambiguously aligned in the *env* gp41 gene were deleted from the alignment. For a representation of the *pol* alignment, see the supplemental material. Hypermutation and recombination were investigated using Hypermut and Simplot v2.5, respectively (22, 33). Appropriate nucleotide substitution models were determined with Modeltest v3.06 (29). Maximum likelihood phylogenetic trees were reconstructed in PAUP* (v4b10) using three different heuristic branch-swapping algorithms (37). Bootstrapping was performed using the stepwise addition algorithm for 1,000 replicates. Maximum a posteriori trees were inferred using MrBayes (v3.0) (13). Synonymous and

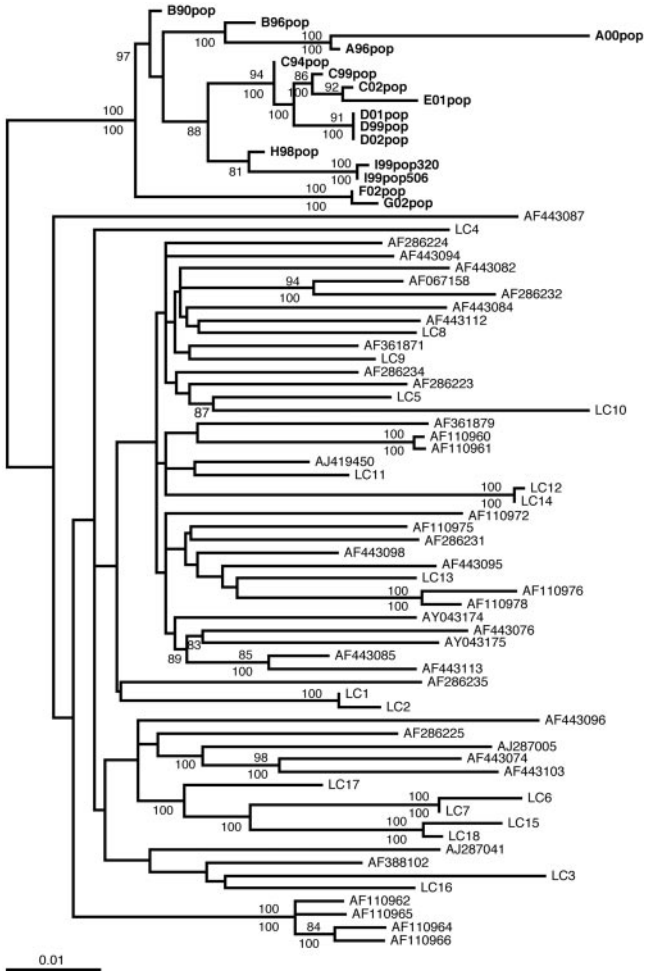

FIG. 2. Maximum likelihood phylogenetic reconstruction for the HIV-1 transmission chain patients and unrelated controls based on the *pol* gene region. The sequences sampled from the transmission chain patients are represented in bold. Subtype C sequences retrieved from a local database are labeled "LC." The tree is rooted at its midpoint. The upper numbers at the nodes indicate the percentage of neighbor-joining bootstrap samples, based on 1,000 replicates, in which the node is supported (only values of >80% are shown). The lower numbers at the nodes represent approximate posterior probabilities obtained from a posterior sample of trees (only values of >80% are shown).

nonsynonymous trees were reconstructed using the neighbor-joining method based on distance matrices estimated in Syn-Scan, which uses a model that includes allelic mixtures (8). Different tree topologies were compared using the Kishino-Hasegawa test and the Shimodaira-Hasegawa test (16, 36).

Molecular adaptation at individual sites (26, 44), along specific lineages (45), and at individual sites along specific lineages (43) was investigated using codon substitution models as implemented in PAML (v3.13) (42). To test for diversifying selection at individual sites, different models were compared that allow for heterogeneous nonsynonymous/synonymous substitution rate ratios ($\omega = d_N/d_S$) among sites (models M0, M1, M2, M3, M7, and M8; see reference 44). Likelihood ratio testing (LRT) was used to test whether allowing for sites with $\omega > 1$ significantly improves the fit of the model to data (M1-M2, M0-M3, and M7-M8, where M2, M3, M7 can accommodate positively selected sites). To investigate different selective pressure along specific lineages, we tested a model that allows only a single ω for all branches in the tree (M0) against a two-ratio model that allows an additional ω for specific branches in the tree. Positively selected sites along the lineages of interest were identified using branch site models (43). Branch site model A is an extension of the neutral model (M1) because it allows for sites being positively selected along a prespecified lineage while belonging to the class with $\omega_0 = 0$ or $\omega_1 = 1$ in the background. In branch site model B, ω_0 and ω_1 are estimated as free parameters, and thus it is an extension of the discrete model (M3 with two discrete classes of sites).

Molecular clock analysis was performed using maximum likelihood methods implemented in PAML (v3.13b) (42). Evolutionary rates were estimated under the assumption of a constant rate of evolution for sequences that were serially sampled over time (single rate dated tip model [SRDT]) (30). The molecular clock was tested by comparing the SRDT model against an unconstrained different rates (DR) model using LRT (30).

Nucleotide sequence accession number. The sequences described here have been submitted to the GenBank database and assigned accession numbers AF338984, AF338990, AF338992, AF338997, AF339013, AF339017, and AY749169 to AY749196 for the *pol* sequences and AY749197 to AY749208 for the *env* gp41 sequences.

RESULTS

The known transmission history. In this study, we identified a heterosexual HIV-1 transmission chain consisting of nine individuals. The epidemiological information obtained by patient interviews and the clinical data, the time of sampling and the treatment history of the patients are summarized in Fig. 1. The direction of transmission and a relatively narrow time interval were determined for all transmission events, with the exception of patients A and B. Although there was clearly a transmission event between both patients, they reported each other as the original donor and a time interval could not be defined. Since several viral phylogenies might have resulted from this transmission history, we did not attempt to reconstruct a single known phylogenetic tree. Instead, the scheme of the "true" transmission history was considered as a pathway along which the viral population is assumed to have evolved. We have further used the term "compatible" if the viral phylogeny could have been generated under the known transmission history. Similar to the Swedish transmission chain (19), this transmission history spans more than a decade of HIV-1 evolution; however, the transmission chain reported here is situated almost exactly one decade later than the Swedish transmission chain. Reflecting the progress of HIV research and treatment during this decade, almost all patients of our transmission chain have received antiretroviral therapy. We chose to obtain sequence data for the relatively conserved *pol* gene region, which is used to test for resistance against commonly available drugs, and the more variable *env* gp41 gene region, which is anticipated to be used for testing resistance against fusion inhibitors (41).

The viral evolutionary history. In a first analysis, we tested whether the sequences obtained from the transmission chain patients were more closely related to each other than to unrelated control sequences. This represents the general hypothesis test for molecular investigations in forensic settings (5, 18, 25, 27). Phylogenetic analysis of 16 *pol* sequences from the nine transmission chain patients, as well as a set of unrelated controls extracted from a local database and GenBank, confirmed with statistical significance that the transmission chain sequences constituted a monophyletic cluster in the subtype C phylogeny (Fig. 2). A similar analysis of the reverse transcriptase (RT) gene, including additional control sequences with matched drug resistance mutations, still identified the transmission chain as a monophyletic cluster (see the supplemental material). To test if the evolutionary history of the virus was compatible with the known transmission history, molecular

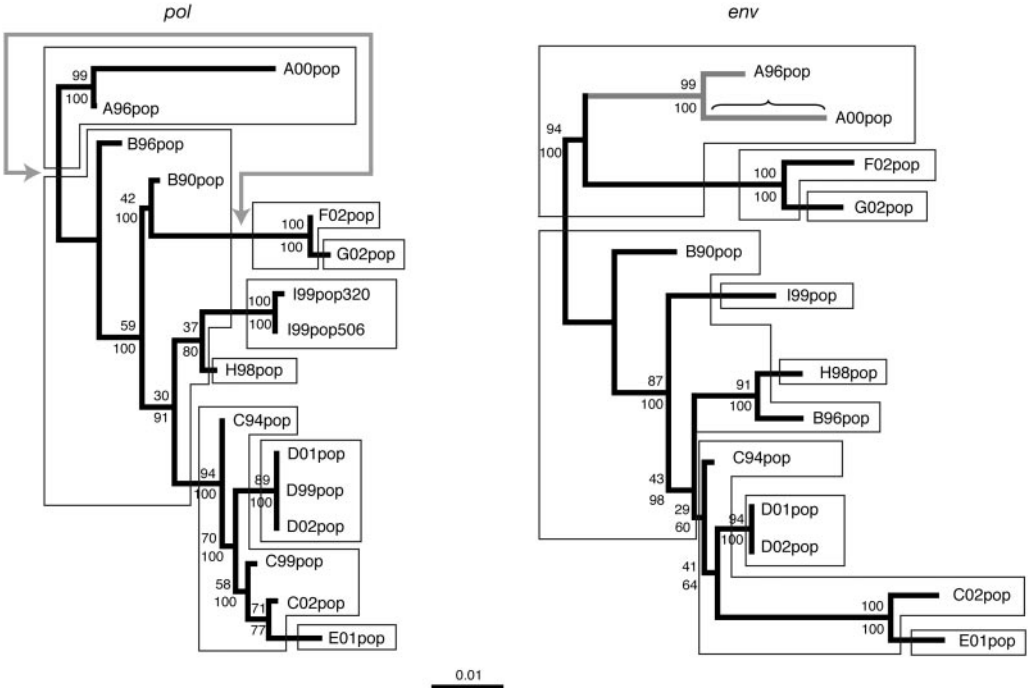

FIG. 3. Phylogenetic trees inferred for the *pol* and *env* gp41 gene regions. Maximum likelihood and Bayesian methods resulted in the same topology for each gene region. Both trees are represented on the same scale and rooted at the position that does not distinguish between patient A and patient B as the original donor for this transmission chain. The most likely host transmission scheme is superimposed onto the viral evolutionary history: hosts are separated arbitrarily along the branch between donor and recipient. For isolates C99pop, D99pop, and I99pop320, not enough sample was left to perform *env* gp41 sequencing. The upper numbers at the nodes indicate the percentage of bootstrap samples, based on 1,000 replicates, in which the node is supported. The lower numbers at the nodes represent approximate posterior probabilities obtained from a posterior sample of trees. The arrow in gray indicates the branch swapping that would make the *pol* phylogeny topologically congruent with the known transmission history. The branch set tested to be under positive selective pressure is indicated in gray in the *env* gp41 tree; the branch tested to have a higher nucleotide substitution rate is indicated with the horizontal bracket.

phylogenies were reconstructed based on *pol* and *env* gp41 sequences. Maximum likelihood reconstructions based on both gene regions are depicted in Fig. 3. Both maximum likelihood and Bayesian inference resulted in the exact same topologies. In Fig. 3, the most likely host transmission history is superimposed onto the viral tree, indicating that the evolutionary history in *env* gp41 was perfectly compatible with the known transmission history, while the *pol* tree revealed a particular incompatibility. According to the known transmission history, the sequences for patients A, F, and G were expected to be monophyletic, as correctly inferred in the *env* gp41 tree. However, the *pol* phylogeny did not suggest patient A as the donor for F. As indicated in Fig. 3, this inconsistency could be resolved by a single branch swapping of (A00pop, A96pop) or (F02pop, G02pop) in either direction. It is interesting to note that the ML bootstrap support for the nodes relevant to the incompatible tree was weak compared to the posterior probabilities, suggesting that the more conservative bootstraps might be less misleading in this case. Phylogenetic reconstruction based on the concatenated *pol* and *env* gp41 alignment resulted in the same topology as inferred for the *env* gp41 gene only. For consistency, we have further referred to this topology as the *env* gp41 tree. Using the Kishino-Hasegawa test and the Shimodaira-Hasegawa test, we compared the *env* gp41 tree topology, congruent with the known transmission history, with the *pol* tree (Table 1). Although the *pol* reconstruction yielded a better likelihood for the *pol* alignment, the *env* gp41 reconstruction could not be significantly rejected by any test independent of the evolutionary model used. For the *env* gp41 data, however, the *pol* tree, not fully compatible with the known transmission history, gave a significantly worse fit than the *env* gp41 tree.

TABLE 1. *P* values for the tree incongruence tests

Data[a]	*P* value for tree[b]:			
	pol		*env* gp41	
	KH	SH	KH	SH
pol	Best ML tree	Best ML tree	0.324	0.166
env gp41	<0.001	<0.001	Best ML tree	Best ML tree
polenv	0.033	0.024	Best ML tree	Best ML tree
pol-RT	0.822	0.388	Best ML tree	Best ML tree

[a] *polenv* represents the concatenated *pol* and *env* gp41 alignment, and *pol*-RT represents the *pol* alignment after exclusion of the positions associated with drug resistance in the RT.

[b] The results for the Kishino-Hasegawa (KH) and Shimodaira-Hasegawa (SH) tests are listed for the Hasegawa-Kishino-Yano model of evolution with gamma-distributed rate heterogeneity among sites; the tests for other models of evolution are consistent with these results (data not shown).

To investigate whether the unexpected clustering for the *pol* gene could have resulted from drug-selective pressure, we identified the drug resistance mutations for all samples based on the mutation list available from the International AIDS

TABLE 2. Drug resistance-associated mutations

Patient and sample	Pro mutation(s)[a]	RT mutation(s)
A		
A96pop	M36L	M41L, D67DN, K70KR, L210LW, T215Y
A00pop	L101, L33F, M36L, **G48V** 154A, **V82A**	M41L, E44D, D67N, V118X, M184V, G190AG, L210W, T215HY, K219N, F227FL
B		
B90pop	M36L	
B96	M36L	M41L, E44DE, V118IV, L2190W, T215Y
C		
C94pop	M36L	
C99pop	M36L	
C02pop	M36L	
D		
D99pop	M36L	
D01pop	M36L	
D02pop	M36L	
E		
E01pop	M36L	V75IV
F		
F02pop	K20M, M36L	
G		
G02pop	K20M, M36L	
H		
H98pop	M36L	M184V
I		
I99pop320	M36L	V118X, M184V, T215Y
I99pop506	M36L	M184V, T215Y

[a] Drug resistance-associated mutations in protease (Pro) and RT were identified according to the criteria of the International AIDS Society—USA (14). Protease mutations in bold represent major mutations. The ubiquitous M36L mutation has been shown to confer cross-resistance to atazanavir in combination with other known protease inhibitor resistance mutations, although it has not been selected for by atazanavir either in vitro or in vivo (3). However, no patient in this transmission chain has been treated with this drug. Also the K20M mutation is only important in combination with other protease inhibitor resistance mutations and should here be considered as natural polymorphism.

Society (Table 2) (14). Except for some natural polymorphisms, there was no evidence that resistance mutations were transmitted in this transmission chain. The table reveals that the virus in patient A had a significant number of resistance mutations, and at the second time of sampling, this patient was classified as multidrug resistant under ongoing exposure to therapy. Since such amino acid-altering mutations might have caused an incompatible phylogeny in *pol*, we reconstructed trees for *pol* based on synonymous (silent) and nonsynonymous (amino acid altering) distances separately (Fig. 4). While the synonymous tree was now fully compatible with the known transmission history, the nonsynonymous tree showed again that the sequences from patient A did not cluster with F and G. Instead, the patient A virus clustered with the patient I virus, which had developed resistance mutations that were also all present in the patient A virus (Table 2). Therefore, we could argue that drug-selective pressure had resulted in a pattern of parallel evolution. After exclusion of the codon positions in the RT at which resistance mutations were identified, a *pol* phylogeny fully compatible with the known transmission history could also be inferred (data not shown). For these data, which contained only 10 codon sites less than the original *pol* alignment, the tree incongruence test results were inverted (Table 1). The *env* gp41 topology resulted now in a higher likelihood than the original *pol* topology, but the latter was again not significantly rejected. Comparison of the *pol* and *env* gp41 phylogenies also indicated that different tree topologies could be compatible with the same transmission history. For example, the patient H strain clustered with the patient I strain in *pol* while the former clustered with a patient B isolate in *env* gp41; however, both scenarios were compatible with the known transmission history. In contrast to the *pol* tree, the tree topology reconstructed for *env* gp41 based on nonsynonymous distances was compatible with the known transmission history (data not shown).

Testing for selective pressure. Since we have identified a topological incompatibility in the *pol* reconstruction and attributed it to drug-selective pressure, we could test whether this has left a footprint of positive selection in the viral lineages of interest. The nonsynonymous/synonymous substitution rate ratio ($\omega = d_N/d_S$) provides a qualitative measure of natural selection at the protein level, with $\omega = 1$, $\omega < 1$, and $\omega > 1$, indicating neutral evolution, purifying selection, and positive selection, respectively. Since an averaging approach failed to detect positive selection in a similar case of parallel evolution under drug-selective pressure (4), we used more sensitive codon substitution models to identify positive selection at individual sites (26, 44), along specific lineages (45), and at individual sites along specific lineages (43). As genealogy, we used the tree reconstructed in the *env* gp41 gene (or *pol* plus *env* gp41 concatenated alignment), which was consistent with the transmission history (Fig. 3). The results of the maximum likelihood inference under these models are summarized in Table 3. The two-ratio model that allowed for a different ω for the patient A branches ($\hat{\omega}_1$) compared to all other branches ($\hat{\omega}_0$) fitted significantly better than the one-ratio model (M0) ($P = 0.0001$). Although this provided evidence for a different selective pressure in the patient A lineages, the ω estimate for these lineages ($\hat{\omega}_1 = 1.6230$) was not significantly higher than 1 ($P = 0.3078$).

The results of the site-specific models indicated that the selective pressure on the protein varied greatly among amino acid sites (Table 3). All models allowing for positively selected sites (M2, M3, and M8) provided a significantly better fit to the data than their neutral counterparts (M1, M0, and M7, respectively). Interestingly, the positively selected sites identified by the empirical Bayes' criterion included several drug resistance-associated mutations, mostly observed in more than one patient.

We further tested for sites under positive selection along the lineage of interest using branch site models. Parameter estimates under model A suggested that 65% of sites were highly conserved across all lineages, with $\omega_0 = 0$, and 18% of sites were nearly neutral with $\omega_1 = 1$, whereas the additional 17% of sites were under strong positive selection along the patient A branches, with $\hat{\omega}_2 = 4.74$. In comparison with the simpler neutral model (M1), model A showed a statistically significant improvement ($P = 0.0045$). Parameter estimates under model B suggested that 4.7% of sites were under positive selection in all lineages, with $\hat{\omega}_1 = 4.34$, whereas 18% of sites were only under positive selection in patient A, with $\hat{\omega}_2 = 2.91$. The LRT

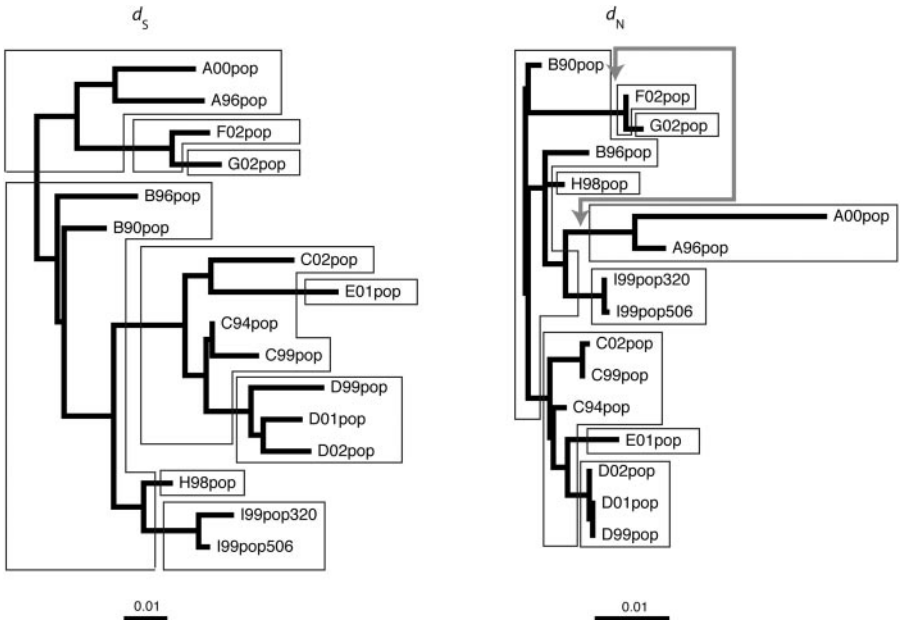

FIG. 4. Phylogenetic trees reconstructed using synonymous (d_S) and nonsynonymous (d_N) distances for *pol*. Both trees are represented as rooted at the position that does not distinguish between patient A and patient B as the original donor for this transmission chain. The most likely host transmission scheme is superimposed onto the viral evolutionary history. The arrow in gray indicates the branch swapping that would make the d_N phylogeny topologically congruent with the known transmission history.

indicated that the branch site model B was significantly better than the site-specific model M3 ($K = 2$) at the 0.05 confidence level ($P = 0.025$), suggesting positive selection in the patient A lineages, in addition to positively selected sites in all lineages. Interestingly, the positively selected sites identified for the background included several positions at which drug resistance mutations were identified in more than one patient (RT, 41, 184, 210, and 215) while the positively selected sites identified for the foreground lineage included additional positions at which drug resistance mutations were only found in patient A (Pro, 10, 33, 48, 54, and 82; RT, 67 and 219). A similar analysis of the *env* gp41 gene region also identified sites under positive selection, but differential selective pressure in the specified lineages was not observed (data not shown).

Molecular clock modeling. Our analyses indicated that several mutations have been fixed under drug-selective pressure in patient A, which could have resulted in a faster evolutionary rate along this lineage. Exploratory linear regression analysis appeared to confirm this effect in the terminal branch of taxon A00pop (data not shown). To test this more formally, we applied molecular clock modeling to the heterochronous sequence data. The results of the maximum likelihood inference are shown in Table 4. The SRDT model, which constrains the tips of the tree to be proportional to the sampling dates, resulted in a nucleotide substitution rate of 0.00121 substitutions/site/year. The single-rate model, which makes no accommodation for the temporal sampling of the isolates (30), was significantly rejected by the LRT in favor of the SRDT model ($P < 0.0001$). This comparison suggested that incorporating isolation dates into a single-rate model significantly improved the likelihood. However, the SRDT model is rejected in favor of the more general different rates (DR) model ($P < 0.0001$), indicating that the assumption of a global molecular clock was violated. A local clock dated tips model, relaxing the molecular clock along the branch leading to the taxon A00pop (Fig. 3), fitted the data significantly better than the SRDT model ($P < 0.0001$). The rate estimated for the branch leading to the multidrug-resistant isolate was significantly higher than the background lineages (0.00616 substitutions/site/year; Table 4). It should be noted that the local clock model was still significantly rejected in favor of the DR model ($P = 0.0014$). Also for the *env* gp41 gene, a global molecular clock was rejected (DR versus SRDT, $P = 0.0007$), but no lineage effect in patient A was observed (local clock dated tips versus SRDT, $P = 0.99$).

DISCUSSION

In this study, we extensively documented a known HIV-1 transmission history almost a decade later than the Swedish transmission chain (19). Although the number of people carrying HIV-1 is estimated to be around 40 million (www.unaids.org), the long time span between the identification of both transmission chains of comparable size reflects the rarity of identifying such data. Sampling genetic data from known transmission histories of fast-evolving pathogens provides the opportunity to investigate the accuracy of phylogenetic reconstructions and the rate and mode at which genetic variation is accumulated (17, 19, 21). Here, we have sequenced the *pol* and *env* gp41 regions of the HIV-1 genome from plasma RNA in at least one sample of nine infected patients. Phylogenetic reconstructions using likelihood methods revealed that, in contrast to the *env* gp41 tree, the *pol* tree was incompatible with the known transmission history. The difference between transmission history and reconstructed topology in *pol* could be attributed to strong drug-selective pressure resulting in a pattern of parallel evolution. Analysis of the *pol* data without the posi-

TABLE 3. Parameter estimates for the codon substitution models applied to the *pol* data

Model	p[a]	Log L[d]	Estimates of parameters[b]	Positively selected sites[c]
M0 (1 ratio)	1	−2,137.255821	$\hat{\omega} = 0.3918$	None
Branch specific (2 ratios)	2	−2,129.700900	$\hat{\omega}_0 = 0.2495$, $\hat{\omega}_1 =$ **1.6230**	NA
Site-specific				
M1, neutral ($K = 2$)	1	−2,118.947877	$\hat{p}_0 = 0.68985$ ($\hat{p}_1 = 0.31015$)	Not allowed
M2, selection ($K = 3$)	3	−2,107.074525	$\hat{p}_0 = 0.67099$, $\hat{p}_1 = 0.32597$ (**$\hat{p}_2 = 0.00304$**), $\hat{\omega}_2 =$ **52.92262**	RT, 215 ($P > 99$)
M3, discrete ($K = 2$)	3	−2,112.026513	$\hat{p}_0 = 0.93249$, (**$\hat{p}_1 = 0.06751$**) $\omega_0 = 0.17437$, $\hat{\omega}_1 =$ **4.26134**	Pro, 12, 54, 72 (at $0.5 < P < 0.95$), 37 (at $P > 99$) RT, 41, 174, 177, 245 (at $0.5 < P < 0.95$), 36, 135, 184, 210, 211, 215 (at $P > 99$)
M3, discrete (K = 3)	5	−2,104.838467	$\hat{p}_0 = 0.89797$, $\hat{p}_1 =$ **0.09907**, (**$\hat{p}_2 = 0.00297$**) $\hat{\omega}_0 = 0.14630$, $\hat{\omega}_1 =$ **2.75117**, $\hat{\omega}_2 =$ **60.67074**	PRO, 10, 12, 16, 18, 38, 72, 82, (at $0.5 < P < 0.95$), 37, 54, (at $P > 95$) RT, 41, 174, 177, 204 (at $0.5 < P < 0.95$), 210, 211, 215T*, 245, 36, 135, 184 (at $P > 95$)
M7, beta	2	−2,118.943419	$\hat{p} = 0.04111$, $\hat{q} = 0.11138$	Not allowed
M8, beta & ω	4	−2,112.028167	$\hat{p}_0 = 0.93293$, $\hat{p} = 21.08181$, $\hat{q} = 99.00000$ ($\hat{p}_1 =$ **0.06707**), $\hat{\omega} =$ **4.27569**	Same as M3 ($K = 2$)
Branch site				
A	3	−2,113.535572	$\hat{p}_0 = 0.64620$, $\hat{p}_1 = 0.18545$, (**$\hat{p}_2 + \hat{p}_3 = 0.16835$**) $\omega_2 =$ **4.73987**	Sites for foreground: PRO, 10, 13, 33, 37, 38, 41, 48, 62, 82, 89 (at $0.5 < P < 0.95$), 54 (at $P > 95$) RT, 20, 44, 48, 67, 201, 208, 215, 219, 223 (at $0.5 < P < 0.95$)
B	5	−2,108.351559	$\hat{p}_0 = 0.77144$, $\hat{p}_1 =$ **0.04723** (**$\hat{p}_2 + \hat{p}_3 = 0.18132$**)	Sites for background: Pro, 12, 37 (at $0.5 < P < 0.95$) RT, 36, 41, 135, 174, 177, 184, 204, 210, 211, 215, 245 at ($0.5 < P < 0.95$)
			$\hat{\omega}_0 = 0.10933$, $\hat{\omega}_1 =$ **4.34176**, $\hat{\omega}_2 =$ **2.91030**	Sites for foreground: Pro, 10, 13, 33, 38, 41, 48, 54, 62, 72, 82, 89 (at $0.5 < P < 0.95$) RT, 20, 44, 48, 67, 201, 208, 219, 223 (at $0.5 < P < 0.95$)

[a] p is the number of free parameters for the ω ratios.
[b] Parameters indicating positive selection are presented in bold type. Those in parentheses are presented for clarity only but are not free parameters.
[c] An asterisk indicates the site belonging to the class with $\hat{p}_2$ and $\hat{\omega}_2$ in model M3 ($K = 3$). Underlined sites refer to positions associated with drug resistance. NA, not applicable.
[d] Log L, log likelihood.

tions at which known drug resistance mutations were identified resulted in a fully compatible tree.

Although the selective pressure in patient A, due to various drug regimens and eventually leading to multidrug resistance, might be an extreme case of selection in molecular evolution, it had a remarkable impact on our inference. Natural selection, even if confined to a few positions in the sampled gene fragment, can provide sufficient counterweight to the remaining phylogenetic information resulting in incorrect clustering. This observation demands caution when the *pol* gene is used for testing transmission hypotheses in forensic investigations (e.g., references 23 and 25). The presence of resistance mutations does not necessarily invalidate analyses of the *pol* gene for forensic purposes because the hypothesis test of epidemiological relatedness still provided convincing results (Fig. 1). This was also confirmed in the RT gene by using additional drug-resistant control sequences (see the supplemental material). However, it is highly recommended to assess the impact of sites with resistance-associated mutations by removing them from the alignment in a parallel analysis (12). As accessibility to antiretroviral treatment will increase, the footprint of drug-selective pressure might become a common feature in *pol* sequences. Reconstructions based on the *env* gp41 gene region, known to be subject to host immune-selection pressure as

TABLE 4. Molecular clock results

Model	Data	p[a]	Log L[c]	Evolutionary rate (nucleotide substitutions/site/yr)
Different rates	*pol*	30	−2,148.63	NA[b]
	env gp41	28	−2,294.18	
Single rate	*pol*	19	−2,186.81	NA
	env gp41	18	−2,320.77	
Single rate dated tips	*pol*	20	−2,172.89	1.21E-03
	env gp41	19	−2,308.45	2.11E-03
Local clock dated tips	*pol*	21	−2,162.15	Background, 0.973E-03 Foreground, 6.16E-03
	env gp41	20	−2,308.36	Background, 2.06E-03 Foreground, 2.45E-03

[a] p denotes the number of parameters used in the model.
[b] NA, not applicable.
[c] Log L, log likelihood.

reflected in a higher overall d_N/d_S, did not seem to be severely influenced by parallel substitutions. On the contrary, the *env* gp41 data had the power to reject the incompatible *pol* evolutionary history. This showed that obtaining sequence data for multiple HIV genome regions is not a redundant recommendation in forensic investigations (18). It is interesting to note that patient A is now successfully being treated with the fusion inhibitor T20, which targets the HIV-1 gp41 transmembrane glycoprotein (15). Therefore, future sampling of this patient might indeed also reveal an effect of drug-selective pressure in the *env* gp41 gene (32).

A common convention for detecting the action of positive selection is the nonsynonymous/synonymous substitution rate ratio, a ratio greater than 1 indicating that nonsynonymous mutations offer fitness advantages and have a higher fixation rate than synonymous mutations. Unfortunately, an average ratio usually has little power to detect positive selection (e.g., references 1, 7, and 35). For the evolution of HIV drug resistance in particular, Crandall et al. (4) have shown that this ratio is a poor indicator of natural selection. This is not surprising, since for the conventional drugs, resistance mutations are fixed in the most functionally conserved protein of the HIV genome. Using codon substitution models, we were able to explicitly test differential selective pressure in patient A. Although this approach confirmed a significantly higher d_N/d_S along these lineages compared to the background, we still could not demonstrate any positive selection in the patient A lineages with statistical significance. Only the most complex model for detecting molecular adaptation at individual sites along specific lineages (branch site model B) revealed positively selected sites in patient A, in addition to positively selected sites throughout the complete genealogy. Interestingly, the latter included several positions with resistance mutations identified in more than one patient, while the former included several positions with resistance mutations exclusively seen in patient A. An exception to this was position 44, for which a drug resistance mutation was observed in both patient A and patient B but was identified as a positively selected site only in patient A. However, this mutation is only present as an allelic mixture in the patient B population sequence and therefore was not fully fixed in the population at that time of sampling. It remains to be elucidated whether the remaining sites identified to be under positive selection are resistance associated, selected by the host immune system or simply false positives. For example, it is interesting to note that mutations at positions 20, 203, and 218 in the RT, identified as positively selected in the foreground lineage, were significantly associated with nucleoside RT inhibitor therapy in a recent comprehensive statistical analysis (9).

Finally, we have demonstrated that the increased fixation of resistance mutations in patient A resulted in a significantly higher evolutionary rate along a terminal branch in patient A. Again, this could be explicitly tested by maximum likelihood modeling. However, the local molecular clock model with dated tips was still significantly rejected in favor of the different rates model, which does not assume a molecular clock. A molecular clock in the *env* gp41 gene was also rejected. This was in contrast with the Swedish transmission chain that supported the existence of a molecular clock (17). The difference might have resulted from a different testing approach or from the impact of antiretroviral treatment. On the one hand, it has been shown that effective treatment can result in a slowing down or even cessation of viral evolution (6). On the other hand, suboptimal therapy can lead to drug resistance, which can result in increased mutation rates, in turn increasing the likelihood of further resistance under ongoing therapy (24). The latter might have been the evolutionary pathway in patient A.

In conclusion, we were able to uncover the molecular footprint of drug-selective pressure in a case where the HIV transmission history was known. If the aim would have been to estimate the transmission history, which is much more common in HIV phylogenetics, we would have drawn incorrect conclusions on the basis of the *pol* gene region. Therefore, we suggest that caution should be taken when accurate reconstructions of HIV evolution are required.

ACKNOWLEDGMENTS

This work was supported by the Flemish Fonds voor Wetenschappelijk Onderzoek (FWO G.0288.01). P.L. was supported by the Flemish Institute for the Promotion and Innovation through Science and Technology in Flanders (IWT Vlaanderen). A.R. was supported by the Royal Society. K.V.L. and Y.S. were supported by the Belgian Ministry of Social Affairs through a fund within the Health Insurance System.

We thank Y. Schrooten and B. Maes for expert laboratory assistance.

REFERENCES

1. **Akashi, H.** 1999. Within- and between-species DNA sequence variation and the 'footprint' of natural selection. Gene **238:**39–51.
2. **Altschul, S. F., T. L. Madden, A. A. Schaffer, J. Zhang, Z. Zhang, W. Miller, and D. J. Lipman.** 1997. Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. Nucleic Acids Res. **25:**3389–3402.
3. **Colonno, R. J., A. Thiry, K. Limoli, and N. Parkin.** 2003. Activities of atazanavir (BMS-232632) against a large panel of human immunodeficiency virus type 1 clinical isolates resistant to one or more approved protease inhibitors. Antimicrob. Agents Chemother. **47:**1324–1333.
4. **Crandall, K. A., C. R. Kelsey, H. Imamichi, H. C. Lane, and N. P. Salzman.** 1999. Parallel evolution of drug resistance in HIV: failure of nonsynonymous/synonymous substitution rate ratio to detect selection. Mol. Biol. Evol. **16:**372–382.
5. **DeBry, R. W., L. G. Abele, S. H. Weiss, M. D. Hill, M. Bouzas, E. Lorenzo, F. Graebnitz, and L. Resnick.** 1993. Dental HIV transmission? Nature **361:** 691.
6. **Drummond, A., R. Forsberg, and A. G. Rodrigo.** 2001. The inference of stepwise changes in substitution rates using serial sequence samples. Mol. Biol. Evol. **18:**1365–1371.

7. **Endo, T., K. Ikeo, and T. Gojobori.** 1996. Large-scale search for genes on which positive selection may operate. Mol. Biol. Evol. **13:**685–690.
8. **Gonzales, M. J., J. M. Dugan, and R. W. Shafer.** 2002. Synonymous-nonsynonymous mutation rates between sequences containing ambiguous nucleotides (Syn-SCAN). Bioinformatics **18:**886–887.
9. **Gonzales, M. J., T. D. Wu, J. Taylor, I. Belitskaya, R. Kantor, D. Israelski, S. Chou, A. R. Zolopa, W. J. Fessel, and R. W. Shafer.** 2003. Extended spectrum of HIV-1 reverse transcriptase mutations in patients receiving multiple nucleoside analog inhibitors. AIDS **17:**791–799.
10. **Goujon, C. P., V. M. Schneider, J. Grofti, J. Montigny, V. Jeantils, P. Astagneau, W. Rozenbaum, F. Lot, C. Frocrain-Herchkovitch, N. Delphin, F. Le Gal, J.-C. Nicolas, M. C. Milinkovitch, and P. Dény.** 2000. Phylogenetic analyses indicate an atypical nurse-to-patient transmission of human immunodeficiency virus type 1. J. Virol. **74:**2525–2532.
11. **Grenfell, B. T., O. G. Pybus, J. R. Gog, J. L. Wood, J. M. Daly, J. A. Mumford, and E. C. Holmes.** 2004. Unifying the epidemiological and evolutionary dynamics of pathogens. Science **303:**327–332.
12. **Hue, S., J. P. Clewley, P. A. Cane, and D. Pillay.** 2004. HIV-1 pol gene variation is sufficient for reconstruction of transmissions in the era of antiretroviral therapy. AIDS **18:**719–728.
13. **Huelsenbeck, J. P., and F. Ronquist.** 2001. MRBAYES: Bayesian inference of phylogenetic trees. Bioinformatics **17:**754–755.
14. **Johnson, V. A., F. Brun-Vezinet, B. Clotet, B. Conway, R. T. D'Aquila, L. M. Demeter, D. R. Kuritzkes, D. Pillay, J. M. Schapiro, A. Telenti, and D. D. Richman.** 2003. Drug resistance mutations in HIV-1. Top. HIV Med. **11:** 215–221.
15. **Kilby, J. M., S. Hopkins, T. M. Venetta, B. DiMassimo, G. A. Cloud, J. Y. Lee, L. Alldredge, E. Hunter, D. Lambert, D. Bolognesi, T. Matthews, M. R. Johnson, M. A. Nowak, G. M. Shaw, and M. S. Saag.** 1998. Potent suppression of HIV-1 replication in humans by T-20, a peptide inhibitor of gp41-mediated virus entry. Nat. Med. **4:**1302–1307.
16. **Kishino, H., and M. Hasegawa.** 1989. Evaluation of the maximum likelihood estimate of the evolutionary tree topologies from DNA sequence data, and the branching order in hominoidea. J. Mol. Evol. **29:**170–179.
17. **Leitner, T., and J. Albert.** 1999. The molecular clock of HIV-1 unveiled through analysis of a known transmission history. Proc. Natl. Acad. Sci. USA **96:**10752–10757.
18. **Leitner, T., and J. Albert.** 2000. Reconstruction of HIV-1 transmission chains for forensic purposes. AIDS Rev. **2:**241–251.
19. **Leitner, T., D. Escanilla, C. Franzen, M. Uhlen, and J. Albert.** 1996. Accurate reconstruction of a known HIV-1 transmission history by phylogenetic tree analysis. Proc. Natl. Acad. Sci. USA **93:**10864–10869.
20. **Leitner, T., and W. Fitch.** 1999. The phylogenetics of known transmission histories, p. 315–345. *In* K. A. Crandall (ed.), The evolution of HIV. Johns Hopkins University Press, Baltimore, Md.
21. **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.
22. **Lole, K. S., R. C. Bollinger, R. S. Paranjape, D. Gadkari, S. S. Kulkarni, N. G. Novak, R. Ingersoll, H. W. Sheppard, and S. C. Ray.** 1999. Full-length human immunodeficiency virus type 1 genomes from subtype C-infected seroconverters in India, with evidence of intersubtype recombination. J. Virol. **73:**152–160.
23. **Machuca, A., and V. Soriano.** 2000. In vivo fluctuation of HTLV-I and HTLV-II proviral load in patients receiving antiretroviral drugs. J. Acquir. Immune Defic. Syndr. **24:**189–193.
24. **Mansky, L. M.** 2002. HIV mutagenesis and the evolution of antiretroviral drug resistance. Drug Resist. Updates **5:**219–223.
25. **Metzker, M. L., D. P. Mindell, X. M. Liu, R. G. Ptak, R. A. Gibbs, and D. M. Hillis.** 2002. Molecular evidence of HIV-1 transmission in a criminal case. Proc. Natl. Acad. Sci. USA **99:**14292–14297.
26. **Nielsen, R., and Z. Yang.** 1998. Likelihood models for detecting positively selected amino acid sites and applications to the HIV-1 envelope gene. Genetics **148:**929–936.
27. **Ou, C. Y., C. A. Ciesielski, G. Myers, C. I. Bandea, C. C. Luo, B. T. Korber, J. I. Mullins, G. Schochetman, R. L. Berkelman, A. N. Economou et al.** 1992. Molecular epidemiology of HIV transmission in a dental practice. Science **256:**1165–1171.
28. **Palella, F. J., Jr., K. M. Delaney, A. C. Moorman, M. O. Loveless, J. Fuhrer, G. A. Satten, D. J. Aschman, and S. D. Holmberg.** 1998. Declining morbidity and mortality among patients with advanced human immunodeficiency virus infection. N. Engl. J. Med. **338:**853–860.
29. **Posada, D., and K. A. Crandall.** 1998. MODELTEST: testing the model of DNA substitution. Bioinformatics **14:**817–818.
30. **Rambaut, A.** 2000. Estimating the rate of molecular evolution: incorporating non-contemporaneous sequences into maximum likelihood phylogenies. Bioinformatics **16:**395–399.
31. **Rambaut, A., D. Posada, K. A. Crandall, and E. C. Holmes.** 2004. The causes and consequences of HIV evolution. Nat. Rev. Genet. **5:**52–61.
32. **Rimsky, L. T., D. C. Shugars, and T. J. Matthews.** 1998. Determinants of human immunodeficiency virus type 1 resistance to gp41-derived inhibitory peptides. J. Virol. **72:**986–993.
33. **Rose, P. P., and B. T. Korber.** 2000. Detecting hypermutations in viral sequences with an emphasis on G→A hypermutation. Bioinformatics **16:** 400–401.
34. **Shankarappa, R., J. B. Margolick, S. J. Gange, A. G. Rodrigo, D. Upchurch, H. Farzadegan, P. Gupta, C. R. Rinaldo, G. H. Learn, X. He, X.-L. Huang, and J. I. Mullins.** 1999. Consistent viral evolutionary changes associated with the progression of human immunodeficiency virus type 1 infection. J. Virol. **73:**10489–10502.
35. **Sharp, P. M.** 1997. In search of molecular Darwinism. Nature **385:**111–112.
36. **Shimodaira, H., and M. Hasegawa.** 1999. Multiple comparisons of log-likelihoods with applications to phylogenetic inference. Mol. Biol. Evol. **16:**1114–1116.
37. **Swofford, D. L.** 1998. PAUP* 4.0—Phylogenetic Analysis Using Parsimony (*and Other Methods). Sinauer Assoc., Sunderland, Mass.
38. **Taylor, S., P. Cane, S. Hue, L. Xu, T. Wrin, Y. Lie, N. Hellmann, C. Petropoulos, J. Workman, D. Ratcliffe, B. Choudhury, and D. Pillay.** 2003. Identification of a transmission chain of HIV type 1 containing drug resistance-associated mutations. AIDS Res. Hum. Retrovir. **19:**353–361.
39. **Thompson, J. D., T. J. Gibson, F. Plewniak, F. Jeanmougin, and D. G. Higgins.** 1997. The CLUSTAL_X windows interface: flexible strategies for multiple sequence alignment aided by quality analysis tools. Nucleic Acids Res. **25:**4876–4882.
40. **Vandamme, A. M., M. Witvrouw, C. Pannecouque, J. Balzarini, K. Van Laethem, J. C. Schmit, J. Desmyter, and E. De Clercq.** 2000. Evaluating clinical isolates for their phenotypic and genotypic resistance against anti-HIV drugs, p. 223–258. *In* D. Kinchington and R. F. Schinazi (ed.), Antiviral methods and protocols. Humana Press, Inc., Totowa, N.J.
41. **Van Laethem, K., Y. Schrooten, P. Lemey, E. Van Wijngaerden, S. De Wit, M. Van Ranst, and A. M. Vandamme.** 2005. A genotypic resistance assay for the detection of drug resistance in the human immunodeficiency virus type 1 envelope gene. J. Virol. Methods **123:**25–34.
42. **Yang, Z.** 1997. PAML: a program package for phylogenetic analysis by maximum likelihood. Comput. Appl. Biosci. **13:**555–556.
43. **Yang, Z., and R. Nielsen.** 2002. Codon-substitution models for detecting molecular adaptation at individual sites along specific lineages. Mol. Biol. Evol. **19:**908–917.
44. **Yang, Z., R. Nielsen, N. Goldman, and A. M. Pedersen.** 2000. Codon-substitution models for heterogeneous selection pressure at amino acid sites. Genetics **155:**431–449.
45. **Yang, Z., W. J. Swanson, and V. D. Vacquier.** 2000. Maximum-likelihood analysis of molecular adaptation in abalone sperm lysin reveals variable selective pressures among lineages and sites. Mol. Biol. Evol. **17:**1446–1455.