Investigating Signs of Recent Evolution in the Pool of Proviral HIV Type 1 DNA during Years of Successful HAART

HELENE MENS,[1] ANDERS G. PEDERSEN,[2] LOUISE B. JØRGENSEN,[3] STEPHANE HUE,[4] YIZI YANG,[2] JAN GERSTOFT,[1] and TERESE L. KATZENSTEIN[1]

ABSTRACT

In order to shed light on the nature of the persistent reservoir of human immunodeficiency virus type 1 (HIV-1), we investigated signs of recent evolution in the pool of proviral DNA in patients on successful HAART. Pro-viral DNA, corresponding to the C2-V3-C3 region of the HIV-1 env gene, was collected from PBMCs isolated from 57 patients. Both "consensus" (57 patients) and clonal (7 patients) sequences were obtained from five time points spanning a 24-month period. The main computational strategy was to use maximum likelihood to fit a set of alternative phylogenetic models to the clonal data, and then determine the support for models that imply evolution between time points. Model fit and model-selection uncertainty was assessed using the Akaike information criterion (AIC) and Akaike weights. The consensus sequence data was also analyzed using a range of phylogenetic techniques to determine whether there were temporal trends indicating ongoing replication and evolution. In summary, it was not possible to detect definitive signs of ongoing evolution in either the bulk-sequenced or the clonal data with the methods employed here, but our results could be consistent with localized expression of archival HIV genomes in some patients. Interestingly, stop-codons were present at the same two positions in several independent clones and across patients. Simulation studies indicated that this phenomenon could be explained as the result of parallel evolution and that some sites were inherently more likely to evolve into stop codons.

INTRODUCTION

SINCE THE INTRODUCTION OF HIGHLY ACTIVE ANTIRETROVIRAL THERAPY (HAART), clinical management of HIV-1 infection has greatly improved.[1,2] However, little is known about the persistence of HIV-1 reservoirs during HAART and the topic remains controversial. Persistence could be a consequence of the long half-life of latently infected, resting CD4[+] T cells, but could also be caused by a low level of ongoing replication. Because of the lack of a biochemical marker for recent HIV-1 infection, most studies of this topic have been based on serially sampled RNA or DNA. If there is ongoing replication, and if the rate of nucleotide substitution is fairly constant across the entire viral population, then the sequences obtained at later time points will be progressively more distant from the root of the

phylogenetic tree. Based on assessments of this type of temporal structure in phylogenetic trees, recent evolution has been found in some,[1,3–5] but not all studies.[6–8] The low abundance of patients developing drug resistance mutations during successful HAART supports the idea of nearly complete inhibition of HIV-1 replication.[9]

In this study proviral DNA was used in lieu of viral RNA, which is difficult to obtain from patients with low viral loads. The coexistence of recent and archival viral variants in the pool of proviral DNA makes it difficult to detect signs of recent evolution. Here, we approach the issue as a model-selection problem.[10–12] In this context it is important to realize that in essence, a mathematical model of a biological system is simply a very stringently phrased scientific hypothesis about how that system works. The parameters of a mathematical model can be esti-

[1]Department of Infectious Diseases, Rigshospitalet, [2]Center for Biological Sequence Analysis, Technical University of Denmark, and [3]Department of Virology at Statens Serum Institut, Copenhagen, Denmark.
[4]Centre for Virology, Department of Immunity and Infection, University College London, London, UK.

TABLE 1.   PATIENT CHARACTERISTICS (CLONAL DATA SET)

Patient	Viremia group	Year of infection[a]	Year of HAART initiation	Date of inclusion	CD4+ count at inclusion (cells/µl)	VL in episodes of low-grade viremia[b]
64	1	1996	1997	28-01-1998	360	—
78	1	1992	1996	16-03-1998	220	—
87	2	1986	1997	12-01-1998	350	77
92	1	1987	1997	06-11-1997	540	—
101	3	1985	1997	13-11-1997	300	1160
106	1	1985	1997	08-06-1998	510	—
109	3	1989	1996	26-11-1997	110	3040, 30

[a]Defined as year of first positive HIV test.
[b]Defined as VL > 20 copies/ml.

mated from the data using maximum likelihood methods, and the likelihood of a fitted model is then a measure of how well the model describes (or "fits") the data. A list of likelihoods can therefore form the basis for stringently selecting one or more models (or hypotheses) that describe the data well. It is this general approach that we take here to decide whether our data sets indicate the presence of ongoing replication. Specifically, we use maximum likelihood to fit a set of probabilistic models of sequence evolution to the clonal data, and then determine the support for those models that imply evolution between time points.[13–15]

MATERIALS AND METHODS

Study population

Fifty-seven subjects were randomly chosen from a well-characterized cohort of HIV-1-infected individuals undergoing treatment with HAART.[16] The cohort was recruited between 1997 and 1998 under the inclusion criterion of a plasma viral load ≤ 200 copies/ml. The patients were followed for 24 months, during which blood samples were drawn every 3 months. Patients were categorized into three viremia groups based on longitudinal plasma HIV-RNA values: group 1 had a viral load persistently ≤20 copies/ml ($n = 18$), group 2 had one or more samples with viral loads >20 copies/ml but ≤ 200 copies/ml ($n = 29$), and group 3 had one or more samples with viral loads >200 copies/ml ($n = 10$). Seven of the 57 patients were randomly selected for clonal analysis; the patients' characteristics are detailed in Table 1.

HIV-1 RNA quantification

Plasma viral load was quantified every 3 months using an Amplicor HIV-1 monitor (Roche Diagnostic Systems Inc., Branchburg, NJ). The analyses were performed in real time as described previously.[16]

DNA extraction

EDTA anticoagulated whole blood was collected every 6 months, and peripheral blood mononuclear cells (PBMCs) were isolated with lymphoprep (Nycomed Pharma A/S). The samples were stored at −80°C until use. Cellular DNA was extracted from PBMCs using whole blood specimen solution from Roche (Roche Diagnostic Systems Inc., Branchburg, NJ).

TABLE 2.   CLONAL SEQUENCE DATA CHARACTERISTICS

Patient	Number of sequences	Length of alignment	Nucleotide diversity[a]	Number of sequences with stop in position 1[b]	Number of sequences with stop in position 2[c]	Best fitting model type[d]
64	24	372	0.025	0	NA[e]	DR
78	36	396	0.014	1	21	Clock
87	37	387	0.013	5	0	Clock
92	34	378	0.043	6	0	DR
101	34	393	0.078	1	4	DR
106	36	381	0.047	2	0	DR
109	40	378	0.066	5	NA[e]	DR

[a]Average pairwise sequence difference per site.
[b]Corresponding to amino acid 338 of gp120 in the HXB2 HIV reference strain (HIV/SIV sequence locator, Los Alamos HIV database). HXB2 context of position 1: SRAKWNNTL.
[c]Corresponding to amino acid 379 in gp120 of HXB2. HXB2 context of position 2: SFNCGGEFF.
[d]See Table 4.
[e]Sequence not available.

DNA amplification, cloning, and sequencing

For all patients, the C2V3C3 region of the envelope (*env*) gene was amplified by nested polymerase chain reaction (PCR) using a HOTstart taq amplification kit (Qiagen, Hilden, Germany), as previously described.[17] The correct size of the PCR product (410 bp) was verified by agarose gel electrophoresis (2.5%). All extraction and amplification steps were performed independently and with negative controls in parallel to detect possible contamination. Both strands were then sequenced using the ABI prism BigDye terminator Cycle sequencing Ready Reaction Kit (Perkin Elmer Applied Biosystems, Norwalk, CT) on the ABI Prism 377 Genetic Analyzer (Perkin Elmer). The inner PCR primers were used for the sequencing reaction. In seven patients the PCR products were purified using the QIAprep spin miniprep kit 250 (Qiagen Ltd, UK), and cloned using the Subcloning Efficiency DH5α Chemically Competent *E. coli* cells (Invitrogen Ltd, Paisley, UK) and the pGEM-T vector system I (Promega Corporation, Madison, WI). The gene inserts were then independently amplified, with negative controls in parallel to detect possible contamination, and sequenced as described above. The seven clonal data sets contained from 24 to 40 sequences each, resulting in a total of 241 cloned sequences. Sequence lengths ranged from 372 to 402 with an average of 387 (Table 2).

Phylogenetic reconstruction

For each clonal data set, sequence alignments were constructed using the RevTrans server.[18] Columns containing stop codons or gaps were removed from the alignments. The program MrModeltest[19] was used to find the most appropriate nucleotide substitution model based on the Akaike information criterion.[12] Phylogenetic trees were reconstructed by Bayesian inference using the program MrBayes version 3.0B4.[20] For each patient a consensus tree was constructed using the clonal sequences from all five time points and rooted using a sequence from a different patient as outgroup. Outgroups were subsequently removed, and the trees were used as the basis for all further analysis of the data sets. In all cases Markov chain Monte Carlo (MCMC) sampling was performed for 10,000,000 generations with four chains. Convergence was confirmed by comparing the results of two independent runs. The program Tracer[21] was used to determine burn-in and for further confirmation of proper mixing and adequate run-length of the chains. Phylogenetic trees were also constructed in the manner described above for the three groups of consensus sequence data.

Substitution models

The programs baseml and codeml from the PAML package version 3.14[22] were used to fit a range of nucleotide and codon-based models to the clonal data sets using the trees mentioned above. The nucleotide-based models tested with Baseml were The Jukes and Cantor model (JC),[23] the Kimura two-parameter model (K80),[24] the Felsenstein '81 model (F81),[15] the Felsenstein '84 model (F84),[15] the Hasegawa, Kishino, and Yano '85 model (HKY85),[25] the Tamura-Nei '92 model (T92),[26] the Tamura-Nei '93 model (TN93),[27] and the General time-reversible model (REV/GTR).[28,29] Each of the latter models assumes different patterns of nucleotide frequency and ex-

changeability and was tested with and without the assumption of gamma-distributed rate-variation across sites.[30] Seven codon-based models were tested using codeml: M0, M1a, M2a, M3, M5, M7, and M8.[13,31,32] All seven models were fitted using either the F1x4 (overall nucleotide frequencies) or F3x4 (different nucleotide frequencies for each codon position) approach for estimating codon frequencies. An additional three and nine parameters, respectively, were added to the parameter count given by codeml under F1x4 and F3x4. For each of the above-mentioned 16 (2×8) nucleotide models and 14 (2×7) codon models, we fitted three models with different assumptions about temporal structure in the data (Fig. 1). The three model types were (1) the different rates (DR) model,[15] where each branch has an independent substitution rate; (2) the single rate (SR, or "clock") model, which assumes that all sequences have been isolated at the same point in time and evolves according to a molecular clock (i.e., with a constant rate of substitution)[33]; and (3) the Single Rate with Dated Tips (SRDT, or "tipdate") model,[14] which assumes that sequences have evolved according to a molecular clock, but that individual samples have been obtained at different times and that their distances to the root are therefore proportional to the sampling time. The "tipdate" model implies evolution between sample time points and therefore indicates ongoing viral replication.[34] For each data set we thus fitted 48 nucleotide-based ($3 \times 2 \times 8$) and 42 codon-based ($3 \times 2 \times 7$) models for a total of 90 different models that covered a wide range of assumptions about sequence evolution. Convergence was confirmed by comparing the results of several independent runs started with different parameter vectors.

Model selection

The Akaike Information Criterion (AIC) was used to assess model fits. Briefly, AIC is an estimate of the amount of information that is lost when a given model is used to approximate the full truth (the so-called relative Kullback–Leibler distance). AIC is a function of the maximized log-likelihood (lnL) and the number of estimated parameters (K) for a model. Specifically, $AIC = -2\ln L + 2K$ with lower AIC values being better. From AIC it is also possible to compute Akaike weights, which can be used as the conditional probability of the model given the data and the set of initial models. It is possible to estimate the relative importance of a model feature by summing Akaike weights across a subset of models sharing that feature. Inference can thus be based on a large set of models simultaneously.[11,12] Among other things, this is helpful in avoiding the model selection problems associated with misspecification.[35]

Simulation of stop-codon evolution

Simulation experiments were performed in order to investigate whether the cooccurrence of stop codons at the same two sites in several independent clones and patients could be the result of independent, parallel evolution. First, the maximum likelihood ancestral sequence for each of the seven clonal data sets was reconstructed with the program PAUP* (version 4.0b10)[36] using the tree and substitution parameters that were estimated as part of the Bayesian analysis. Subsequently, we simulated evolution of these ancestral sequences along a star-tree with equal branch lengths, again using the substitution parameters estimated from the original data sets as part of the Bayesian analysis. The program Seq-
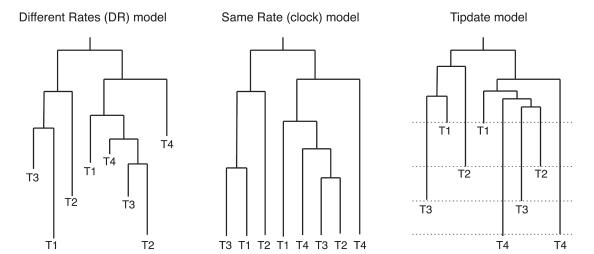
FIG. 1. Schematic presentation of the three different types of time structure used in the phylogenetic models fitted to the clonal data. The different rates model (DR) assumes that individual branches in the tree display different rates of evolution, meaning there is no correlation between sampling time and distance from the root. The single rate, or "clock," model assumes a constant rate of evolution and that all sequences have been sampled at the same time and therefore have the same distance to the root. The "tipdate," or "Single Rate with Dated Tips," model assumes a constant rate of evolution and that individual samples have been obtained at different times (meaning that distance to the root is proportional to the sampling time). If the tipdate model fits the data best, then this indicates that there has been detectable evolution between time points.

Gen[37] was used for simulation. All branches in any given star phylogeny were equally long, and several different lengths were investigated. For each set of conditions 1000 simulations were performed, each resulting in a simulated data set that was subsequently analyzed for the presence of stop codons.

Analysis of viral divergence

Root-to-tip distances were extracted from the Bayesian trees for each patient using software written by the authors. Subsequently root-to-tip distances were plotted versus sampling time and it was tested whether the slope was significantly different from zero.

RESULTS

To investigate whether HIV-1 continued to replicate and evolve in patients on successful HAART, we analyzed serially sampled, proviral DNA from PBMCs isolated from 57 patients with different profiles of low-grade viremia. Samples were collected with 6-month intervals for a period of 24 months (giving five time points). At each time point DNA corresponding to the C2–V3–C3 region of the *env* gene was consensus sequenced. For seven patients we furthermore sequenced 5–10 different clones at each time point (Table 2).

Model-based analysis of clonal data

The main strategy for detecting signs of recent evolution in the clonal DNA involved using maximum likelihood methods to fit a set of alternative phylogenetic models to the data (Fig. 1), and then determining the support for the type of models that implied evolution between time points. The phylogenetic trees constructed for each patient are presented in Fig. 2. On the basis of these trees, a set of 90 different substitution models, with different assumptions about how the sequences had evolved,

was fitted to each of the seven clonal data sets. We then used an information-theoretic approach (the AIC) to assess relative model fit and cross-model support for recent evolution.[11,12] Here, the model feature we were most interested in was time structure, i.e., how sequences obtained at different time points are placed in the tree. If sequences obtained at later time points show a tendency to be proportionally farther from the root, then the tipdate model (Fig. 1) will fit the data best. This would indicate measurable evolution between the investigated time points, i.e., that there has been ongoing replication despite successful HAART.[14,34] A stronger statistical support for the DR or SR models would be compatible with different scenarios of viral evolution (see discussion).

The results of this analysis are shown in Table 3, which lists the set of best-supported models for each of the seven clonal data sets, and in Table 4, where the overall, cross-model support for important model features is given. In all cases between 1 and 10 models were sufficient to account for more than 95% of the weight (Table 3). For two patients (64 and 101) codon-type models had almost 100% cross-model support, while nucleotide-type models had essentially all support in the remaining five (Table 4). Cross-model support was also computed for the three different types of time structure (tipdate, clock, and DR). In two of the seven patients (78 and 87), the clock model received the highest support, while the DR model more adequately described the remaining five patients (Table 4). We found no data sets where the tipdate model received more support than the alternatives, and we can therefore conclude that we find no definitive evidence of ongoing evolution in the clonal data using the model-based methods.

Analysis of root-to-tip distances in clonal phylogenies

As an additional test of whether the clonal data sets displayed temporal structure, we plotted the root-to-tip distance as a function of sampling time for each phylogenetic tree (data not shown).
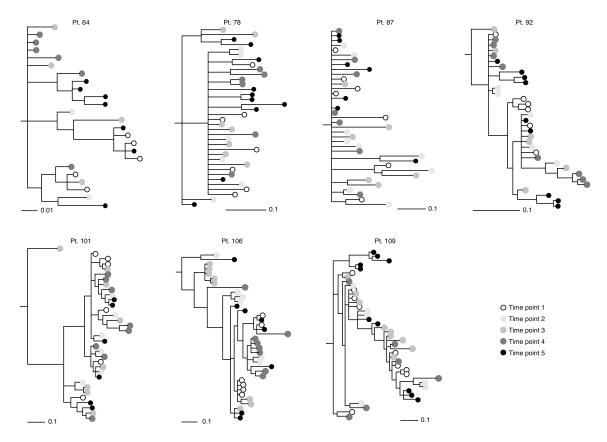
FIG. 2. Bayesian phylogenetic trees constructed from serially sampled clonal sequences spanning the C2–V3–C3 region of the *env* gene. White leafs represent sequences sampled at time point one, light-gray leafs represent sequences sampled at time point two, gray leafs at time point three, dark-gray leafs at time point four, and black leafs at time point five. There is 6 months between all time points.

If sequences from later time points showed a general tendency toward being further removed from the root of the tree, then this was an indication that sequences had evolved between time points. There were no cases in which these plots displayed slopes that were significantly different from zero, again indicating a lack of evidence for recent evolution in the clonal data.

Analysis of colocalized stop codons in clonal data

A number of stop codons were present in the set of cloned sequences. This is to be expected since we are investigating proviral DNA (not viral RNA), and there has consequently been no selection for functionality subsequent to reverse transcription and insertion in the genome. In effect, we are examining the final result of many steps of mutation and selection followed by a single round of unfiltered mutational processes. Interestingly, the 45 stop codons in the data set were all located at only two distinct positions (Table 2). A single sequence from patient 78 contained stop codons at both positions. These observations were puzzling, since it must be assumed that these stop codons will lead to nonfunctional proteins, and they could therefore not have been inherited from common ancestors. If, on the other hand, the stop codons have evolved independently, then it initially seemed surprising that they should all end up at the same two positions in several independent sequences and across six of the seven patients—especially when considering

that of the 61 sense codons 18 can be changed into stop codons by substituting just one single nucleotide.

To investigate whether the stop codons had evolved independently or in parallel we performed simulation experiments using the program Seq-gen.[37] Briefly, we reconstructed ancestral sequences for all seven clonal data sets, and then proceeded to simulate the evolution of these. For each set of conditions 1000 simulations were performed, each resulting in a simulated data set that was subsequently analyzed for the presence of stop codons. If stop codons were to repeatedly evolve at a limited number of sites in the simulated data sets, then this would indicate that the observed colocalization of stop codons could in fact have arisen by chance. The results from this analysis are as follows. First, the site that most frequently changed into a stop codon during simulation was a tryptophan, corresponding to stop codon position 1 in the original data sets. This is understandable since (1) all data sets displayed high G–A substitution rates, and (2) tryptophan is encoded by the singlet codon TGG, which mutates to one of the three stop codons TAG, TGA, or TAA when any or both Gs are substituted with an A. We also note that TGG is in fact the only sense codon that can change into stop as a result of a G to A mutation. Colocalization of stop codons at position 1 is thus explained by the presence of a tryptophan codon at this position and high G–A mutation rates. In the simulation for patient 78, an extra site at position 379 of gp120 (HXB2 coordinates) was found to mu-

TABLE 3.   BEST FITTING MODELS FOR CLONAL DATA SETS

Patient	Model	K[a]	lnL[b]	AIC[c]	Akaike weight[c]	Cumulated weight[d]
64	DR (M3, flx4)	44	−861.2	1810.3	0.41	0.41
	DR (M3, flx4)	41	−864.5	1811.1	0.28	0.69
	SR (M3, flx4)	21	−885.9	1813.8	0.07	0.78
	DR (M3, flx4)	50	−857.2	1814.4	0.05	0.82
	SR (M5, flx4)	18	−889.3	1814.7	0.05	0.87
	DR (M2, flx4)	43	−864.9	1815.8	0.03	0.90
	DR (M8, flx4)	43	−864.9	1815.8	0.03	0.92
	SRDT (M3, flx4)	22	−885.9	1815.9	0.03	0.95
	SRDT (M5, flx4)	19	−889.4	1816.8	0.02	0.96
78	SR (TN93, γ)	15	−989.7	2009.4	0.37	0.37
	SR (REV, γ)	18	−986.9	2009.8	0.30	0.67
	SRDT (REV, γ)	19	−986.1	2010.3	0.23	0.90
	SRDT (TN93, γ)	16	−990.4	2012.9	0.06	0.97
87	SR (REV, γ)	16	−976.8	1985.6	0.72	0.72
	SRDT (REV, γ)	17	−977.2	1988.3	0.18	0.90
	SR (TN93, γ)	13	−982.0	1990.1	0.08	0.98
92	DR (REV, γ)	59	−1311.0	2740.0	0.96	0.96
101	DR (M3, flx4)	68	−1934.8	4005.7	0.28	0.28
	DR (M2, flx4)	67	−1936.6	4007.3	0.12	0.40
	DR (M5, flx4)	65	−1938.7	4007.3	0.12	0.53
	DR (M8, flx4)	67	−1936.7	4007.4	0.12	0.64
	SRDT (M3, flx4)	36	−1967.9	4007.9	0.09	0.73
	SRDT (M5, flx4)	33	−1971.4	4008.8	0.06	0.79
	SR (M3, flx4)	35	−1969.6	4009.2	0.05	0.84
	SRDT (M2, flx4)	35	−1969.7	4009.4	0.04	0.88
	SRDT (M8, flx4)	35	−1969.8	4009.5	0.04	0.92
	SR (M5, flx4)	32	−1973.0	4010.1	0.03	0.95
106	DR (REV, γ)	67	−1396.6	2927.2	0.57	0.57
	DR (TN93, γ)	64	−1399.9	2927.8	0.42	0.99
109	DR (REV, γ)	74	−1786.6	3721.2	0.56	0.56
	DR (HKY85, γ)	70	−1791.3	3722.7	0.27	0.83
	DR (TN93, γ)	71	−1790.9	3723.7	0.16	0.99

[a]Number of parameters.
[b]Likelihood of model.
[c]Akaike information criterion.
[d]Best fitting models sorted by Akaike weight. For each patient we include models such that the cumulated Akaike weight is at least 95% (the 95% credible set of models).

tate to stop with a much lower frequency. This site was found to correspond to the stop codon at position 2 and to be a CGA (arginine) codon. In the other patients this site was mostly occupied by AGA (arginine) or GGA (glycine) codons, although a few sequences in the data set of patient 101 also contained CGA codons at this position. It thus seems that the observed colocalization of stop codons at position 2, which was observed in patients 78 and 101, can be explained by mutation of CGA to the stop codon TGA.

Analysis of bulk sequenced data

Five sequential consensus sequences were available for the majority of patients. Phylogenetic trees covering all consensus data showed sequences from each patient forming a tight cluster distinct from other patients (data not shown). This indicated the absence of contamination during PCR amplification, and also suggested that these population averages did contain useful information about the underlying sequences. The consensus sequence trees displayed no temporal pattern, with time-dependent, increasing distance from the root of the tree (data not shown). Thus, there were no signs of evolution between time points when examining the phylogenetic trees.

DISCUSSION

We investigated signs of recent evolution in the pool of proviral DNA in patients on successful HAART, with the purpose of shedding light on the mechanisms of viral persistence. Proviral DNA was used in lieu of viral RNA, which is difficult to obtain from patients with low viral loads. It is of course possible that a given piece of proviral DNA encodes a defective virus, and we indeed observed several in-frame stop codons in the investigated sequence data. However, proviral DNA still contains phylogenetic information since it is, by necessity, separated by just one reverse transcription step from a virus that

TABLE 4. CROSS-MODEL SUPPORT FOR MODEL FEATURES

	Type of temporal model			Type of substitution model	
Patient	Tipdate	Clock	DR	Codon[a]	Nucleotide[b]
64	0.05	0.14	0.81	0.99	0.01
78	0.31	0.69	0.00	0.00	1.00
87	0.20	0.80	0.00	0.00	1.00
92	0.00	0.00	1.00	0.00	1.00
101	0.23	0.12	0.64	0.99	0.01
106	0.01	0.00	0.99	0.00	1.00
109	0.00	0.00	1.00	0.00	1.00

Cross-model support for different temporal model types (tip-date, clock, and DR) and for different substitution model types (codon and nucleotide-based). Cross-model support for a given feature was computed by summing Akaike weights for all models including that feature.

[a]Codon-based substitution models (M0, M1, M2, M3, M5, M7, and M8).

[b]Nucleotide-based substitution models (JC, K80, K81, F84, HKU85, TN92, TN93, and GTR/REV).

must have been functional. We analyzed both consensus-sequenced data (57 patients) and clonal data (7 patients) from several time points spanning a 2-year period.

Analysis of consensus-sequenced proviral DNA

We found no temporal structure with sampling-time-dependent, increasing distance from the root of the tree in the consensus-sequence phylogenies. This is consistent with a lack of ongoing replication. However, it should be noted that while consensus sequencing is convenient for gaining information on the average properties of a viral population, there are several drawbacks to analyzing it. Thus, consensus sequences essentially give a weighted average of the many different viral sequences that are present at a given time, and it is therefore difficult to recover information on the actual rate of substitution and other aspects of the evolutionary dynamics of the underlying individual sequences using this type of sequence data.

Analysis of root-to-tip distances in clonal phylogenies

As was the case for the consensus-sequenced data, we did not observe temporal structure (i.e., sampling time-dependent, increasing distances between root and tips) when analyzing phylogenies reconstructed from clonal data. This is again consistent with there being no recent evolution, but again does not prove it. Thus, the observation period (24 months) might have been too short to observe increasing distances from the root, although Günthard et al.[3] found increasing root-to-tip distances within 2 years in 3/6 patients on HAART, with suppression of viral load to ≤50 copies/ml, analyzing HIV-1 RNA (before therapy) and proviral DNA (on therapy). Frenkel et al.[5] also found increasing distance to the most recent common ancestor in 2/10 children on HAART, with viral load ≤50 copies/ml, analyzing proviral HIV-1 DNA, but here the median observation period was 5.1 years. Furthermore, we would expect to see

temporal structure in the trees only if individual sequences have been evolving at fairly similar rates.

Model-based analysis of clonal data

When analyzing the clonal sequence data, we also asked whether the proviral DNA sequences displayed signs of recent evolution as a model-selection problem. This allowed us to use rigorous, and highly sensitive, statistical methods for determining the support for a range of alternative hypotheses concerning the evolution of the analyzed sequences. Specifically, 90 alternative phylogenetic models were fitted to each clonal data set using maximum likelihood. Each model can be thought of as a stringently phrased hypothesis about how the investigated sequences have evolved. We then used the AIC to assess cross-model support for tipdate-type models (SRDT), which imply detectable evolution between time points (Fig. 1). In no cases did tipdate receive more support than the alternatives. Instead the standard molecular clock model (SR) received the most support for two data sets, while the DR model was most highly supported for the remaining five. The DR model would be expected to be a good description if there has been no recent evolution and if the proviral DNA corresponds to archival sequences sampled from different times in the past. However, data sets conforming to the DR model could also be the result of ongoing replication and evolution, if the viruses evolve at widely different rates. It is not possible to differentiate between these possibilities based on our sequence data and the models used here (see below for further discussion of this point).

The two cases where the standard molecular clock received the most support (patients 78 and 87) are potentially interesting. In these cases the data do support clock-like evolution, but with no detectable change accumulated during the 2-year observation period. This is consistent with a lack of ongoing replication in the PBMCs of these patients. Supporting this notion is the observation that these two patients are also the ones with the lowest overall diversity, and one of them (patient 78) was furthermore unusual in having a very large number of stop codons (Table 2). Although we cannot rule out the possibility that the observation period has been too short, it should be noted that 2 years was sufficient for detecting signs of evolution in at least one other study.[3]

In summary, it was not possible to detect definitive signs of ongoing evolution in either the bulk-sequenced or the clonal data with the methods employed here.

Modeling the evolution of proviral HIV-1 DNA

It is important to note that the approach used here is quite different from classical hypothesis testing in which a null model is compared to an alternative model, and in which a lack of support for the alternative model does not necessarily imply that the null model is well supported. In the AIC-based model-selection framework employed here, there is no concept of null or alternative models. Instead we investigate a whole range of models in parallel, and determine the support (the Akaike-weight) for each of them. As mentioned, the Akaike-weight can be considered to be the conditional probability that a model is the best one, given the data and the initial set of models. It is of course important that the initial set of models is chosen carefully such that it has a sufficient coverage of relevant hypotheses concerning the investigated system. While we believe that the

set of 90 models used in this study is well chosen in this respect, it is still possible that some model details could be improved. It would be especially interesting to experiment with models that explicitly accounted for the nature of our proviral data set. As mentioned above, the proviral DNA is at most one reverse-transcription step separated from a functional virus, but obviously this last step is quite different from the rest of the evolutionary history of the viral sequence. It is possible that our data set would be more adequately described by a model having two sets of parameter values: one set of values would cover the combined effects of mutation and selection during the part of the viral phylogeny where there is ongoing replication (viruses enter cells, collect mutations, and some functional viruses finally get to perform the next round of infection; model parameters would therefore represent the combined effects of the mutation and selection steps). The other set of parameter values would then cover the final step from entry into the cell through reverse transcription up to insertion of the proviral DNA. Since there is no selection for functionality during this last step, the parameters would mostly represent the mutation process during reverse transcription (and possibly APOBEC3G-mediated editing), and they are therefore likely to be quite different from those involved in the first part of the viral life history. It is unclear whether using such models would have an impact on the relative support for tipdate, molecular clock, and DR-type models. In addition to providing a more accurate description of how the analyzed sequences have evolved, it would also be interesting to use this approach as a way of obtaining estimates of the "raw" mutational rates purged for the effects of selection.

Analysis of colocalized stop codons in clonal data

We also investigated the occurrence of stop codons at only two positions across several independent sequences and across six patients. While initially puzzling, simulation studies indicated that this phenomenon could be explained by the presence of codons that were inherently more likely to change into stop given the relatively high G–A and C–T substitution rates observed in this and other HIV data sets. In particular, the single TGG (tryptophan) codon in our data set was found to be a hot spot for generating stop codons due to G to A mutation. High G–A substitution rates are very likely caused by the action of the APOBEC3G enzyme. This host-encoded enzyme is known to deaminate cytosine residues in the first (minus polarity) strand of reverse-transcribed viral DNA converting these cytosine to uracils and resulting in G to A mutations on the plus strand.[38,39] Since tryptophan is not easily replaced by other amino acids and since it is encoded by just one codon, namely TGG, the virus has difficulty escaping from this hot spot by mutation and must instead rely on the action of the Vif protein.[38] The second observed stop codon (present in two of the seven patients) was found to be explained by a CGA (arginine) codon that was converted to the stop codon TGA due to the relatively high frequency of C to T mutation. High C to T mutation rates have frequently been observed in HIV data sets,[40] although it is unclear what the background for this phenomenon is. While C to T mutation is also the activity displayed by APOBEC3G, this enzyme mostly exerts its effects on the minus polarity strand due to its preference for single-stranded

DNA,[38] arguing against this being the cause for the observed C to T changes on the plus strand. It would be interesting to determine whether an unknown enzyme is in fact responsible for the high C to T rates.

In summary, this study shows that proviral genomes in the pool of PBMC-derived proviral DNA from patients on successful HAART most frequently evolve in a non-clock-like fashion, presumably following activation and proliferation of memory T cells. Models more accurately describing the processes driving the evolution of the persistent reservoir are needed to further investigate intrahost viral dynamics.

SEQUENCE DATA

GenBank accesion numbers DQ468392-642 and DQ463441-681.

ACKNOWLEDGMENTS

REFERENCES

1. Egger M, May M, Chene G, *et al.*: Prognosis of HIV-1-infected patients starting highly active antiretroviral therapy: A collaborative analysis of prospective studies. Lancet 2002;360(9327): 119–129.

2. Mocroft A, Ledergerber B, Katlama C, *et al.*: Decline in the AIDS and death rates in the EuroSIDA study: An observational study. Lancet 2003;362(9377):22–29.

3. Gunthard HF, Frost SD, Leigh-Brown AJ, *et al.*: Evolution of envelope sequences of human immunodeficiency virus type 1 in cellular reservoirs in the setting of potent antiviral therapy. J Virol 1999;73(11):9404–9412.

4. Ramratnam B, Mittler JE, Zhang L, *et al.*: The decay of the latent reservoir of replication-competent HIV-1 is inversely correlated with the extent of residual viral replication during prolonged antiretroviral therapy. Nat Med 2000;6(1):82–85.

5. Frenkel LM, Wang Y, Learn GH, *et al.*: Multiple viral genetic analyses detect low-level human immunodeficiency virus type 1 replication during effective highly active antiretroviral therapy. J Virol 2003;77(10):5721–5730.

6. Persaud D, Pierson T, Ruff C, *et al.*: A stable latent reservoir for HIV-1 in resting CD4(+) T lymphocytes in infected children. J Clin Invest 2000;105(7):995–1003.

7. Ruff CT, Ray SC, Kwon P, *et al.*: Persistence of wild-type virus and lack of temporal structure in the latent reservoir for human immunodeficiency virus type 1 in pediatric patients with extensive antiretroviral exposure. J Virol 2002;76(18):9481–9492.

8. Kieffer TL, Finucane MM, Nettles RE, *et al.*: Genotypic analysis of HIV-1 drug resistance at the limit of detection: Virus production without evolution in treated adults with undetectable HIV loads. J Infect Dis 2004;189(8):1452–1465.

9. Hermankova M, Ray SC, Ruff C, *et al.*: HIV-1 drug resistance profiles in children and adults with viral load of <50 copies/ml receiving combination therapy. JAMA 2001;286(2):196–207.

10. Huelsenbeck JP and Rannala B: Phylogenetic methods come of age: Testing hypotheses in an evolutionary context. Science 1997;276(5310):227–232.

11. Burnham KP and Anderson DR: *Model Selection and Multimodel Inference: A Practical Information-Theretic Approach,* 2nd ed. Springer-Verlag, New York, 2002.

12. Posada D and Buckley TR: Model selection and model averaging in phylogenetics: Advantages of Akaike information criterion and Bayesian approaches over likelihood ratio tests. Syst Biol 2004;53(5):793–808.

13. Goldman N and Yang Z: A codon-based model of nucleotide substitution for protein-coding DNA sequences. Mol Biol Evol 1994;11(5):725–736.

14. Rambaut A: Estimating the rate of molecular evolution: Incorporating non-contemporaneous sequences into maximum likelihood phylogenies. Bioinformatics 2000;16(4):395–399.

15. Felsenstein J: Evolutionary trees from DNA sequences: A maximum likelihood approach. J Mol Evol 1981;17(6):368–376.

16. Katzenstein TL, Ullum H, Roge BT, *et al.*: Virological and immunological profiles among patients with undetectable viral load followed prospectively for 24 months. HIV Med 2003;4(1):53–61.

17. Leitner T, Korovina G, Marquina S, Smolskaya T, and Albert J: Molecular epidemiology and MT-2 cell tropism of Russian HIV type 1 variant. AIDS Res Hum Retroviruses 1996;12(17):1595–1603.

18. Wernersson R and Pedersen AG: RevTrans: Multiple alignment of coding DNA from aligned amino acid sequences. Nucleic Acids Res 2003;31(13):3537–3539.

19. Nylander JAA: MrModeltest, 2.2, 2002. Department of Systematic Zoology, Uppsala University, Uppsala, Sweden.

20. Ronquist F and Huelsenbeck JP: MrBayes 3: Bayesian phylogenetic inference under mixed models. Bioinformatics 2003;19(12):1572–1574.

21. Rambaut A and Drummond A: Tracer, 2004. Oxford Evolutionay Biology Group, University of Oxford.

22. Yang Z: PAML: A program package for phylogenetic analysis by maximum likelihood. Comput Appl Biosci 1997;13(5):555–556.

23. Jukes T and Cantor C: Evolution of protein molecules. In: *Mammalian Protein Merabolism*. Academic Press, New York, 1969.

24. Kimura M: A simple method for estimating evolutionary rates of base substitutions through comparative studies of nucleotide sequences. J Mol Evol 1980;16(2):111–120.

25. Hasegawa M, Kishino H, and Yano T: Dating of the human-ape splitting by a molecular clock of mitochondrial DNA. J Mol Evol 1985;22(2):160–174.

26. Tamura K: Estimation of the number of nucleotide substitutions when there are strong transition-transversion and G+C-content biases. Mol Biol Evol 1992;9(4):678–687.

27. Tamura K and Nei M: Estimation of the number of nucleotide substitutions in the control region of mitochondrial DNA in humans and chimpanzees. Mol Biol Evol 1993;10(3):512–526.

28. Yang Z: Estimating the pattern of nucleotide substitution. J Mol Evol 1994;39(1):105–111.

29. Zharkikh A: Estimation of evolutionary distances between nucleotide sequences. J Mol Evol 1994;39(3):315–329.

30. Yang Z: Maximum likelihood phylogenetic estimation from DNA sequences with variable rates over sites: Approximate methods. J Mol Evol 1994;39(3):306–314.

31. Nielsen R and Yang Z: Likelihood models for detecting positively selected amino acid sites and applications to the HIV-1 envelope gene. Genetics 1998;148(3):929–936.

32. Yang Z, Nielsen R, and Hasegawa M: Models of amino acid substitution and applications to mitochondrial protein evolution. Mol Biol Evol 1998;15(12):1600–1611.

33. Goldman N: Statistical tests of models of DNA substitution. J Mol Evol 1993;36(2):182–198.

34. Drummond A, Pybus OG, Rambaut A, Forsberg R, and Rodrigo AG: Measurably evolving populations. Trends Ecol Evol 2003;18:481–488.

35. Zhang J: Performance of likelihood ratio tests of evolutionary hypotheses under inadequate substitution models. Mol Biol Evol 1999;16(6):868–875.

36. Swofford D: PAUP—Phylogenetic Analysis Using Parsimony, 4.0. Sinauer Associates, Sunderland, MA, 2003.

37. Rambaut A and Grassly NC: Seq-Gen: An application for the Monte Carlo simulation of DNA sequence evolution along phylogenetic trees. Comput Appl Biosci 1997;13(3):235–238.

38. Goff SP: Death by deamination: A novel host restriction system for HIV-1. Cell 2003;114(3):281–283.

39. Yu Q, Konig R, Pillai S, *et al.*: Single-strand specificity of APOBEC3G accounts for minus-strand deamination of the HIV genome. Nat Struct Mol Biol 2004;11(5):435–442.

40. Posada D and Crandall KA: Selecting models of nucleotide substitution: An application to human immunodeficiency virus 1 (HIV-1). Mol Biol Evol 2001;18(6):897–906.

Address reprint requests to:
Helene Mens
AIDS-lab
afsn.5702
Rigshospitalet
Henrik Harpestrengsvej 4
DK 2100 Copenhagen, Denmark

E-mail: mens@dadlnet.dk