Virology **284,** 250–258 (2001)
doi:10.1006/viro.2000.0887, available online at http://www.idealibrary.com on IDEAL®

Evidence for Positive Selection Driving the Evolution of HIV-1 *env* under Potent Antiviral Therapy

Simon D. W. Frost,*,†,[1] Huldrych F. Günthard,‡ Joseph K. Wong,†,§ Diane Havlir,†
Douglas D. Richman,† and Andrew J. Leigh Brown*,†

*Centre for HIV Research, Institute for Cell, Animal and Population Biology, University of Edinburgh, Scotland; †University of California, San Diego, California 92103; ‡University Hospital of Zurich, Department of Internal Medicine, Division of Infectious Diseases, Zurich, Switzerland; and §San Diego Veterans Affairs Healthscience Center, San Diego, California

Received August 8, 2000; returned to author for revision January 31, 2001; accepted February 25, 2001

In HIV-infected individuals treated with potent antiretroviral therapy, viable virus can be isolated from latently infected cells several years into therapy, due to the long life of these cells, ongoing replication replenishing this population, or both. We have analysed the V3 region of the HIV-1 *env* gene isolated from six patients who have undergone 2 years of potent antiretroviral therapy without frank failure of viral suppression. We show that in two (and possibly three) patients, the sequence changes between baseline virus and virus isolated from infected cells persisting 2 years into infection result from positive selection driving adaptive evolution, occurring either prior to or during therapy. Our analyses suggest low-level replication despite absence of drug resistance due to drug sanctuary sites, or to low-level ongoing replication in the presence of alterations in the selective environment during therapy, perhaps due to a decline in HIV-specific immune responsiveness or changes in target cell pools. In one patient, genetic divergence between baseline plasma and infected cells isolated during therapy may reflect the long half-life of some of these persistent cell populations and the divergence of viral subpopulations that occurred prior to therapy. © 2001 Academic Press

Key Words: HIV-1; *env*; evolution; positive selection; antiviral therapy.

INTRODUCTION

Although the use of highly active anti-retroviral therapy (HAART) to suppress HIV replication can result in dramatic reductions in viral load, viable virus can be isolated from latently infected cells even after several years of HAART, suggesting that therapy may have to be maintained for many years to eradicate HIV from the body (Wong *et al.,* 1997; Finzi *et al.,* 1997; Chun *et al.,* 1997). This persistence may be due to the long life of latently infected cells, ongoing replication that replenishes the infected cell pool, or both (Finzi *et al.,* 1999; Zhang *et al.,* 1999). Even in patients that respond well to therapy, there is substantial variation between patients in their response to HAART, with some individuals showing a slower decline of plasma virus early in therapy and the intermittent reappearance plasma virus late into infection (Günthard *et al.,* 1999). These differences may reflect between-patient variation in levels of replication and/or differences in which cell subpopulations are infected. It has been suggested that activating the immune system may increase turnover of these latently infected cells and hence result in shorter times to eradication and reduction in the duration of HAART required. Understanding the mechanism whereby HIV persists in the body and the extent to which the virus can evolve during potent antiviral therapy is central to understanding how different individuals may respond to such an intervention.

Günthard *et al.* (1999) analysed the divergence of the C2–V3 region of HIV-1 *env* between baseline plasma virus and virus isolated from latently infected cells from six infected individuals. Even though all six patients had responded well to therapy, there was wide variation in the genetic divergence between baseline and year 2-isolated virus. To address whether this divergence had originated prior to or during therapy, the positions of the most common recent ancestors (MRCAs) for the plasma virus and the year 2-isolated virus were analysed. In two patients studied, A and B, baseline and year 2-isolated virus appeared to be very similar. In two patients, K and L, year 2-isolated virus appeared to diverge from baseline virus prior to therapy, suggesting that at least some of the genetic divergence seen between baseline virus and year 2-isolated virus originated prior to therapy. In two other patients, C and M, year 2-isolated virus appeared to diverge directly from baseline virus. As only one sample prior to therapy and one sample during therapy were available, the estimated positions of the MRCAs may be inaccurate, and hence the amount of divergence that occurred prior to therapy relative to during therapy is difficult to establish. However, a positive correlation was found between divergence and the amount of residual replication that occurred during ther-

[1] To whom correspondence and reprint requests should be addressed at Dept. of Pathology, School of Medicine, UCSD Treatment Center, 150 W. Washington St., Suite 100, San Diego, CA 92103. Fax: +1 619 298 0177. E-mail: sdfrost@ucsd.edu.

0042-6822/01 $35.00
Copyright © 2001 by Academic Press
All rights of reproduction in any form reserved.

TABLE 1

Diversity and Divergence Estimates for the Six Patients

Patient	Total diversity d_t, %	Initial diversity d_x, %	Diversity during therapy, d_y, %	Total divergence d_{xy}, %	Net divergence d_a, %
A	3.43 ± 0.59	5.07 ± 0.83	0.00 ± 0.00	3.98 ± 0.74	1.44 ± 0.45
B	0.20 ± 0.07	0.41 ± 0.15	0.05 ± 0.04	0.23 ± 0.08	0.00 ± 0.00
C	2.87 ± 0.52	3.96 ± 0.66	0.12 ± 0.08	3.61 ± 0.72	1.57 ± 0.48
K	2.08 ± 0.51	0.43 ± 0.18	0.30 ± 0.12	3.72 ± 0.98	3.36 ± 0.99
L	1.28 ± 0.18	1.49 ± 0.15	0.25 ± 0.06	1.71 ± 0.32	0.84 ± 0.31
M	5.34 ± 0.81	3.58 ± 0.66	1.07 ± 0.30	7.75 ± 1.30	5.42 ± 1.13

Note. All distances were estimated using the method of Nei and Jin (1989), corrected using a Kimura (1980) two-parameter model, with standard errors estimated from 1000 bootstrap replicates.

apy, consistent with ongoing replication playing a role in the divergence of *env* in chronically infected individuals. The findings of Zhang *et al.* (1999), who described patients starting treatment during early phases of infection, were consistent with these observations.

To further address the effect of potent therapy on the evolution of HIV-1 *env* and what this may tell us about the mechanism of persistence of HIV during therapy, we have analysed the pattern of amino acid divergence during therapy. We show that positive selection had driven adaptive evolution of *env* in two out of six patients studied. We argue that a combination of positive selection prior to (or in the early stages of) therapy and the slow turnover of the latently infected cell pool are very important mechanisms of generating amino acid divergence between plasma virus sampled at the initiation of therapy and virus isolated late into therapy, regardless of whether ongoing replication occurs during therapy. As positive selection had driven the multiple amino acid differences between baseline virus and year 2-isolated virus, they must have had a phenotypic effect on the virus in terms of replication rate, cell tropism, or immune recognition. Hence the virus that may emerge on failure of therapy or during stimulation of the immune system may have different biological properties than that seen on initiation of therapy.

RESULTS

Pairwise analysis of nucleotide diversity and net divergence

The genetic variation in the V3 region of *env* in the sample of viruses isolated 2 years into HAART was lower than that seen in plasma virus at the initiation of therapy (Table 1). For five out of the six patients, genetic diversity in the year 2-isolated virus was very low (0–0.3%), whereas for patient M some diversity (1.07%) appeared to be maintained. Phylogenetic trees illustrating the relationship between sequences sampled at the beginning of therapy and after 2 years of therapy are shown in Fig. 1.

The average pairwise genetic divergence between baseline and year 2-isolated virus was greater than zero and varied greatly between patients. This observation by itself does not prove that the baseline virus and year 2-isolated virus came from two genetically distinct populations, as two samples taken from the same (genetically heterogeneous) population can be divergent from each other. Hence we estimated the "net divergence" between baseline and year 2-isolated virus, which corrects for this effect by taking into account within-sample diversity. If d_{xy} is the average genetic divergence between two populations, x and y, and d_x and d_y are the genetic diversities within populations x and y, respectively, net divergence, d_a, is given by the expression $d_a = d_{xy} - (d_x + d_y)/2$. Although net divergence can be negative when the true divergence between populations is low, we are concerned with high levels of divergence, for which net divergence gives a much better indication than uncorrected distances. Net divergence was even more variable than uncorrected variation, with patients B and L exhibiting low levels of net divergence (<1%), A and C exhibiting medium levels of net divergence (c. 1.5%), and patients K and M exhibiting very high (>3%) levels of net divergence. These results show that the samples of baseline virus and year 2-isolated virus come from genetically distinct populations and are the most distinct in patients K and M.

The low diversity seen in the viral isolates could be due either to the low numbers of infected cells that exist after 2 years of therapy or to an artifact due to sampling a very small number of infected cells during viral isolation, in which case the diversity seen in the isolate may underestimate the true diversity in the latently infected cell pool. In a study of lymph-node RNA in patients with long-term viral suppression during therapy (which included some of the patients analysed here), Gunthard *et al.* (2001) showed that there were essentially no differences between sequences obtained from lymph-node RNA and those obtained from cultured cells. Patient B, who showed very low diversity in the virus isolated from latent cells, also showed very low diversity in sequences

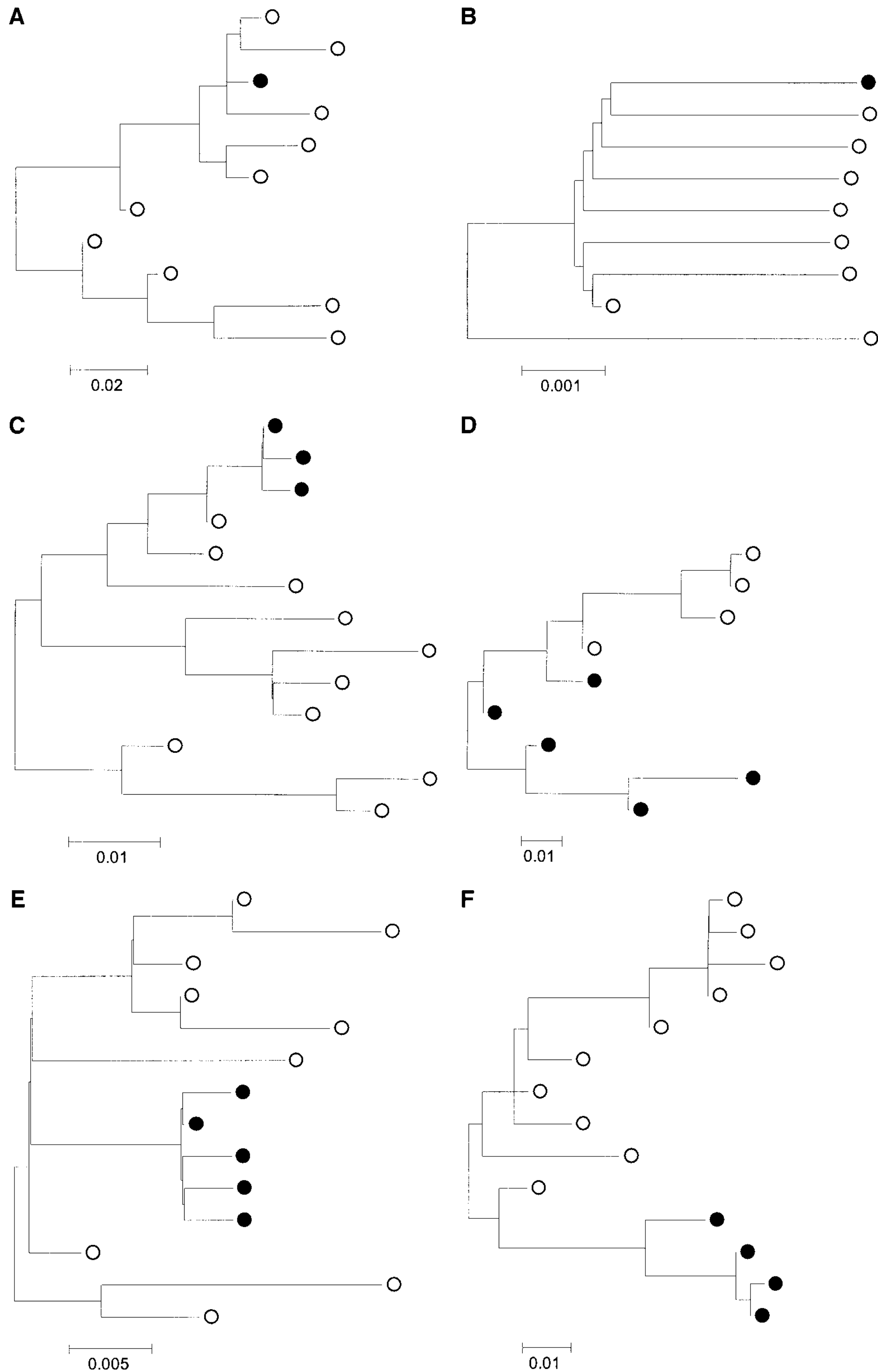

FIG. 1. Phylogenetic trees used for the test for selection for each patient: (A) patient A, (B) patient B, (C) patient C, (D) patient K, (E) patient L, (F) patient M. Open circles denote sequences derived from plasma at the initiation of therapy. Closed circles denote sequences derived from infected cells after 2 years of suppressive therapy. Trees are rooted at their midpoints. Identical sequences have been removed for clarity.

derived from lymph-node RNA. These observations argue that the diversity seen in viral sequences obtained may indeed be representative of that *in vivo*. However, it should be noted that even if therapy is assumed to have no effect on sample diversity, net divergence is still extremely high for patients K (3.29%) and M (4.17%),

TABLE 2

Parameter Estimates and Log-Likelihood Scores Under Model 1 (*dn/ds* ratio equal for all branches in the phylogeny), Model 1* (*dn/ds* = 1 for all branches), and Model 2 (*dn/ds* ratio different for the branch separating the baseline and year 2-isolated virus)

Model	No. of *dn/ds* rates	Patient A	Patient C	Patient K	Patient L	Patient M
1	One	$k = 2.897$	$k = 3.639$	$k = 8.420$	$k = 2.545$	$k = 3.177$
		$dn/ds = 0.838$	$dn/ds = 1.393$	$dn/ds = 1.015$	$dn/ds = 0.323$	$dn/ds = 0.661$
		$\ln L = -859.780$	$\ln L = -774.477$	$\ln L = -555.496$	$\ln L = -646.937$	$\ln L = -835.867$
1*	$dn/ds = 1$	$k = 2.936$	$k = 3.575$	$k = 8.409$	$k = 2.725$	$k = 3.286$
		$dn/ds = 1$	$dn/ds = 1$	$dn/ds = 1$	$dn/ds = 1$	$dn/ds = 1$
		$\ln L = -859.956$	$\ln L = -774.834$	$\ln L = -555.497$	$\ln L = -658.312$	$\ln L = -836.805$
2	Two	$k = 2.898$	$k = 3.6396$	$k = 8.484$	$k = 2.546$	$k = 3.168$
		$dn_1/ds_1 = 0.898$	$dn_1/ds_1 = 1.333$	$dn_1/ds_1 = 0.356$	$dn_1/ds_1 = 0.281$	$dn_1/ds_1 = 0.544$
		$dn_2 = 0.0000$	$dn_2 = 0.0071$	$dn_2 = 0.0425$	$dn_2 = 0.0117$	$dn_2 = 0.0334$
		$ds_2 = 0.0166$	$ds_2 = 0.0000$	$ds_2 = 0.0000$	$ds_2 = 0.0000$	$ds_2 = 0.0000$
		$\ln L = -858.518$	$\ln L = -774.194$	$\ln L = -551.955$	$\ln L = -645.352$	$\ln L = -833.234$
$2\Delta\ln L$ (1* vs. 1)		0.352	0.713	0.001	22.751*	1.876
$2\Delta\ln L$ (2 vs. 1)		2.524	0.567	7.082*	3.169	5.266*
P value (2 vs. 1) (χ^2 approx.)		0.112	0.451	0.008	0.075	0.022
P value (2 vs. 1) (bootstrap)		0.1	0.54	0.02	0.08	0.04

Note. *P* values for the likelihood ratio test were calculated using a χ^2 approximation and using parametric bootstrapping. Asterisks indicate significance using a 2 unit of support criterion.

whereas for the other patients, net divergence is very low or even negative (A, −1.09%; B, −0.18%; C, −0.35%; L, 0.22%). Under the conservative assumption that viral diversity is unaffected by therapy, there is no evidence of mutational differences between the baseline virus and year 2-isolated virus populations in patients A, B, C, and L, but there is strong evidence of divergence in patients K and M.

Comparing the ratio of replacement to silent changes (*dn/ds*) between time points with a background provides evidence for selection

Having established that the baseline virus and year 2-isolated virus form two distinct populations in at least two patients, K and M, we wished to explore whether the divergence between the two populations had been generated under similar selection pressures to those which generated diversity within each population. Using a maximum-likelihood approach, we tested whether the ratio of nonsynonymous to synonymous substitutions, *dn/ds*, between time points was significantly different from *dn/ds* within time points.

Specifically, we asked whether the branch in the phylogeny representing the net divergence between the two time points was associated with different *dn/ds* relative to the background *dn/ds* ratio estimated using the other branches in the phylogeny. For each patient, twice the difference in the likelihood score of the full model (two *dn/ds* ratios, "model 2") and the simple model (one *dn/ds* ratio, "model 1") was compared to the difference in likelihood for 100 simulated datasets produced assuming one *dn/ds* ratio. This allowed us to determine whether the observed *dn/ds* ratio between the samples is greater than expected by chance alone. To compare the patterns seen in patients K and M with those found in patients with relatively little amino acid divergence between time points, we also analysed patients A, C, and L. The variation seen in patient B was too low for a reliable analysis of this type, but given that viral sequences from lymph-node RNA sampled after 1 and 2 years of therapy also showed very low diversity (Gunthard *et al.,* 2001), it seems unlikely that any replication or evolution occurred in this patient.

Fitting of model 2 to the data revealed that *dn/ds* between baseline plasma virus and year 2-isolated virus was significantly higher than *dn/ds* within populations for patients K and M, with patient L approaching statistical significance (Table 2 and Fig. 2). For all three of these patients, the estimate of the proportion of silent changes, *ds*, for the branch connecting the samples was zero, whereas the estimates of the proportion of replacement changes were high (1.17% in patient L, 3.34% in patient M, and 4.25% in patient K). These results suggest that in patients K and M (and possibly patient L, although the results only border on statistical significance), positive selection was responsible for generating the genetic divergence between plasma virus and year 2-isolated virus, rather than divergence evolving under similar selection pressures to those within populations. Here we define positive selection as that which can generate adaptive amino acid changes, such as mutations which aid immune escape or increase replication.

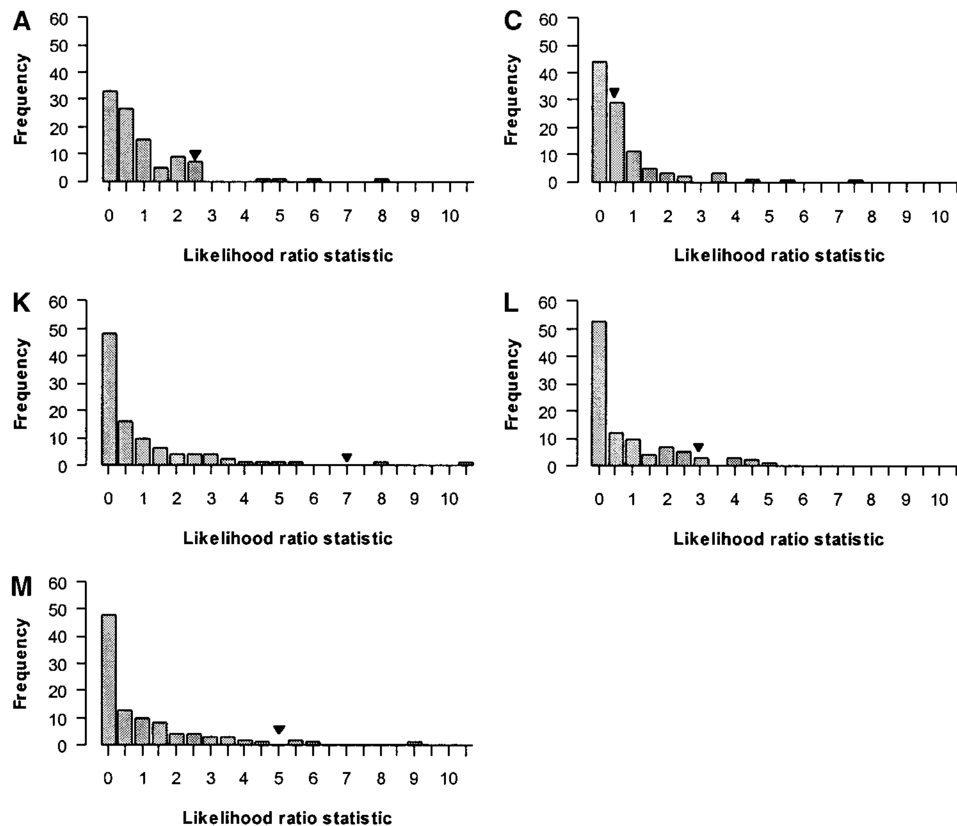

FIG. 2. Null distribution of the likelihood ratio statistic (LRS) for the test for positive selection. The LRS measures the improvement in fit by assuming that the ratio of nonsynonymous to synonymous substitutions between time points is greater than that within time points. For each patient, the null distribution of the LRS is shown, and the LRS is obtained from the data. If positive selection has occurred, then fewer than 5% of the simulations should have a difference in likelihood score greater than that of the data.

Identification of clustered amino acid substitutions between plasma and year 2-isolated virus

To suggest a possible positive selective force, we determined which amino acid sites were divergent between the virus sampled at initiation of therapy and those sampled during therapy by reconstructing the nucleotide sequences at either end of the phylogenetic branch connecting the baseline plasma virus sequences and the year 2-isolated virus sequences (Fig. 3), using a maximum likelihood method. It was found that these sequences could be reconstructed with a high degree of confidence. We then translated the predicted nucleotide sequence to give a predicted amino acid sequence. With one exception (one synonymous change in patient A), all nucleotide differences between the predicted baseline sequence and the predicted year 2 sequence resulted in amino acid changes.

Amino acid replacements did not appear to be uniformly distributed along the sequence. To test whether this clustering was statistically significant, we compared the length of sequence over which substitutions were found with 100,000 datasets generated by randomising the positions of the substitutions. The position of variable sites between the reconstructed sequences was also compared to the known 3D structure of the envelope (Kwong *et al.,* 1998).

In patient M, nine substitutions were observed over a stretch of amino acids 34 amino acids long. The median length over which nine amino acid substitutions were seen in the randomised sequence was 96 residues, significantly longer than observed ($P = 0.00023$), showing that amino acid substitutions were concentrated within a relatively short region. Eight out of the nine amino acid replacements were seen within the V3 loop. Although changes in the V3 loop (such as the S to R substitution in residue 11 of the loop seen in patient M) have been associated with changes in coreceptor usage, an *in vitro* MT-2 assay revealed that there was no detectable switch in usage (Günthard *et al.,* 1999). There was also a change in a glycosylation site in the V3 loop which is normally conserved.

In patient K, 10 substitutions were seen over 73 residues, which approached statistical significance (median length = 97, $P = 0.062$). This P value is higher than that seen in patient M as our statistical test is sensitive to outlying mutations such as the two replacements seen toward the C2 end of the sequence. Eight replacements within a 28-residue stretch of sequence were found to-

```
    250       260       270       280       290       300       310       320       330       340       350       360
     |         |         |         |         |     ****|*********|*********|********|*         |         |         |
                               g            g     g     g    p             p   p      g      g                g
AP THGIRPVVSTQLLLNGSLAEEEVIIRSKNITKNSDNIIVQLKDPVEIKCTRPSNNTRKGIHLGPGGAFFTTGAIIGDIRQAHCNISKGNWNNTLKQIVTELRKHF??NKTIVFNQ
AV .......................................................................................E...........................
CP THGIRPVVSTQFLLNGSLAEEEIVIRSENISDNAKTIIVQLNETVKINCTRPNNNTRKSIHIGPGSAFYTTGAIIGDIRKAHCNISREKWEDTLGKIAKKLSEQFE?NKAIVFKN
CV ...............................N.............E.....................................................................
KP THGIRPVVSTQLLLNGSLAEEEVVIRSDNFTNNAKTIIVQLKVPVDINCTRPNNNTRKGIHIGPGSAFYATGDIIGDIRQAHCNISKAKWNDTLKQVVIKLREQFK?NKTIVFNQ
KV .........................................T...E........................................G....N......T.....YE..R..A..H
LP THGIKPVASTQLLLNGSLAEEEVVIRSENFTNNAKTIIVQLNVSVVINCTRPNNNTRKSISLGD?????????IIGDIRQAHCNLSRKAWNNTLKQVVAKLKEQFGNNKTIAFTH
LV ......................................................................................T..........A...............N.
MP THGIRPVVSTQLLLNGSLAEEEVVIRSENFTNNAKTIIVQLNESVTINCTRPNNNTRRSIHLGMGRALYATGRIIGDIRQAHCNLSREAWNRTLKQIVVKLREQYGNNKTIIFNR
MV .............................................I.......Y....R.Q......F.T..S.T...K....................................
```

FIG. 3. Reconstructed sequences for the branch separating plasma sequences (AP-MP) and isolate sequences (AV-MV). Variable sites are shown in different typefaces based upon the predicted secondary structure proposed by Kwong *et al.* (1998); underline = alpha helix; italic = beta sheet; and bold = loop. The V3 loop is marked by asterisks, glycosylation sites by "g," and sites in the loop associated with changes in coreceptor usage by "p."

ward the V4 end of the sequence, and three of these clustered in the exposed E loop were shown by Kwong *et al.* (1998). No substitutions associated with a switch in coreceptor usage were seen.

In patient L, 3 substitutions were observed over 27 residues. Although these substitutions occurred over a shorter region than the randomised datasets (median length = 55), this was not statistically significant ($P = 0.147$). However, the power of the test is low for so few mutations. Two of these three mutations occurred in an alpha helix; three mutations in this region were also seen in patient K.

Our analyses provide support for clustered amino acid changes between plasma and year 2-isolated virus which, as they have arisen due to positive selection, must have had some effect on the biological properties of the virus. However, we could not determine the selective agent involved in driving divergence and what effect these mutations may have on the virus.

DISCUSSION

In two patients out of six studied, K and M (and possibly in a third, L), we have demonstrated that positive selection is responsible for generating extensive genetic divergence in the V3 region of the *env* gene between plasma virus and virus isolated from infected cells which persisted 2 years into antiviral therapy. Our ability to detect selection comes from our approach of using the pattern of genetic variation within time points as an internal control for each patient. A common test for positive selection is to ask whether the ratio of the rate of nonsynonymous to synonymous substitution, *dn/ds*, is greater than 1. This test has identified the presence of positive selection in the C2–V3 region of *env* in a number of studies (Simmonds *et al.,* 1990; Wolfs *et al.,* 1992; Cichutek *et al.,* 1992; Bonhoeffer *et al.,* 1995; Liu *et al.,* 1997; McDonald *et al.,* 1997; Yamaguchi and Gojobori, 1997). However, this test often fails to detect positive selection as the presence of conserved sites within a sequence can hide positive selection on a few sites (Nielsen and Yang, 1998). Crandall *et al.* (1999) demonstrated convergent evolution of the protease gene between individuals on protease therapy, a signature of positive selection; however, *dn/ds* did not differ significantly from 1. This approach also failed to detect selection in our dataset. When we compared the *dn/ds* ratio estimated from all the sequences (Model 1) with a value of 1, there were no significant deviations for patients A, C, K, and M, and the *dn/ds* ratio was significantly less than 1 for patient L (Model 1*, Table 2).

As selection appears to have resulted in multiple amino acid substitutions, at least some of these substitutions must have some adaptive effect on the biological properties of the virus. One potential agent of positive selection is a switch in coreceptor usage from the CCR5 to the CXCR4 molecule. This has been shown to result in an excess of amino acid replacements relative to silent substitutions (Yamaguchi and Gojobori, 1997, 2000). However, no evidence for such a switch in usage was detected *in vitro* for patient M (Günthard *et al.,* 1999), and no amino acid changes associated with CXCR4 usage were detected for patient K.

A commonly cited cause of divergence of *env* is immune selection. However, the reduction in viremia and the resulting decrease in the levels of anti-HIV immunity (Pitcher *et al.,* 1999) may make selection by the immune

system less powerful during therapy and hence less likely to drive divergence. It is possible that immune selection drove divergence prior to therapy, although no data were available to test this hypothesis.

There is no plausible biological mechanism whereby the anti-retroviral drugs used in these patients may directly select for particular variants of *env*. However, drugs may indirectly select for viruses in drug sanctuaries, subpopulations where antiviral drugs have a suboptimal effect and where cycles of replication can continue during therapy. If certain amino acid substitutions were required to infect such a sanctuary, we might expect to see amino acid divergence between plasma virus at the initiation of therapy and isolated virus late into therapy if the drug sanctuary site was seeded before the initiation of therapy.

A shift in *env* genetic variation, similar to those seen in patients K and M, has been observed in two studies of suboptimal protease therapy. In contrast to this study, these shifts in envelope coincided with the emergence of drug resistance. By measuring *env* variation using a heteroduplex mobility assay, Delwart *et al.* (1998) found that in one individual studied, a new variant of *env* arose very early on in therapy, which was later replaced by an *env* variant which was similar to the variant present prior to therapy. Using a sequencing-based approach, Nijhuis *et al.* (1998) demonstrated a similar phenomenon in one patient out of five studied, with a divergent *env* variant emerging early (52 days) into therapy, which was later (84 days into therapy) replaced by variants closely related to the pretreatment strains. In these two cases, the emerging divergent *env* type is so divergent and emerged so quickly, it must have existed prior to therapy, and was linked to resistance mutations, probably in an anatomical compartment which was relatively isolated from the effects of the drug. In contrast to the patterns seen in patient K and M, the divergent *env* type is replaced by the reemergence of the original *env* variant, probably through recombination between resistance mutations and selection for the original *env* variant. The maintenance of divergent virus over several years in patients K and M suggests that recombination may occur at a lower rate during potent therapy than suboptimal therapy. In addition, selection for the baseline *env* variant may not act as strongly, due to the majority of virus being present in latently infected cells and the waning of the immune response during therapy.

The most parsimonious explanation for the divergence seen between plasma virus and isolated virus is that positive selection acted on the envelope prior to or during the early stages of therapy, but that due to the slow rate of turnover of latently infected cells, the positively selected *env* variant seen in the blood had not reached high levels in the latently infected cells at the time of initiation of therapy. Hence even if no ongoing replication had occurred during therapy, a large number of amino acid substitutions may be seen between plasma virus and virus isolated from latently infected cells. The inherent variability between individuals in patterns of selection on *env* will generate different levels of divergence between plasma virus and virus in latently infected cells. Consequently, to determine whether ongoing replication is occurring due to therapy using sequence data, it is necessary to measure divergence in latently infected cells at several time points. In the case of patient A and one other patient not included in this study, we also had time-zero viral isolates available and these more closely resembled one or more plasma sequences than they did the isolates at year 2. This suggests that although samples from infected cells were not obtained for all patients at the beginning of therapy, the effects from time of sampling may be more important than differences between plasma and infected cells at a given time. We also note that the two patients that exhibited multiple amino acid substitutions, K and M, also exhibited intermittent peaks of viremia during therapy.

In summary, we have shown that in two (and possibly three) cases of patients on potent antiviral therapy without frank failure of suppression, the sequence changes between baseline virus and virus isolated from infected cells persisting 2 years into infection result from positive selection occurring either prior to or during therapy. In these patients, the analyses suggest low-level replication despite absence of drug resistance due to sanctuary sites, or to low-level ongoing replication in the presence of alterations in the selective environment during therapy, perhaps due to a decline in HIV-specific immune responses or changes in target cell pools. In one of these patients, K, the findings may alternatively reflect the long half-life of some of these persistent cell populations and the divergence of viral subpopulations that occurred prior to therapy. Understanding the basis of viral persistence in the face of potent anti-retroviral therapy will be extremely important for the development of strategies for phasing out HAART in patients showing optimum response to therapy.

METHODS

Subjects and clinical background

Six subjects were chosen from a cohort of HIV-infected patients who had undergone two years of treatment with indinavir, lamivudine, and zidovudine as described previously (Wong *et al.,* 1997; Gulick *et al.,* 1998; Günthard *et al.,* 1999). By conventional measures, all six patients responded well to therapy, each sustaining a level of plasma RNA below the detection limit of 50 copies/ml, with the exception of patient M, who had 50 copies/ml on the day of sampling for viral isolation only. Pretreatment viral loads for each patient are as follows: A, 62,680 copies RNA/ml plasma; B, 19,200; C, 51,704; K, 39,322; L, 13,644; M, 50,030.

Sequences

The C2–V3 region of the *env* gene was sequenced for baseline plasma virus and virus isolated 2 years into therapy for each of the six patients A, B, C, K, L, and M as described previously (Wong *et al.,* 1997; Günthard *et al.,* 1999). For the purposes of our analyses, the ends of sequences were edited to give an alignment of 117 codons to allow meaningful comparisons between individuals. The number of sequences obtained at baseline and after 2 years for each patient are as follows: A (10, 9); B (10, 13); C (10, 10); K (7, 10); L (9, 10); and M (10, 7). The GenBank Accession Nos. for these sequences are AF185823 to AF185937. All sequences were checked for G to A hypermutation using the program HYPERMUT (Rose and Korber, 2000).

Analysis of nucleotide diversity and divergence during therapy

For each patient, we estimated the following: (i) d_t, the average pairwise nucleotide diversity for the two time points taken together; (ii) d_x, the average pairwise diversity within the initial samples of plasma sequences; (iii) d_y, the average pairwise diversity within the virus isolates established after 2 years of potent antiretroviral therapy; (iv) d_{xy}, the average nucleotide divergence between the samples; and (v) d_a, the net nucleotide divergence between the samples, which is the average divergence corrected for within-sample variation using the method of Nei and Jin (1989). Multiple hits were corrected for using a Kimura (1980) model. Standard errors were estimated using 1000 bootstrap replicates. All calculations were performed using SENDBS (N. Takezaki, National Institute of Genetics, Mishima, Japan).

Phylogenetic analysis of amino acid substitution pattern

For each patient, the tree topology was estimated using DNAML v.3.572c (J. Felsenstein, Department of Genetics, University of Washington), assuming a transition/transversion ratio of 2 and rate variation modeled by using eight rate categories taken from a discrete gamma distribution with a shape parameter of 0.4. Codon substitution models were fitted to the data and the estimated topology using the program CODEML (Z. Yang, Department of Biology, University College London). Codon frequencies were calculated using the average nucleotide frequencies at each codon position.

One possible cause of the divergence between plasma virus and year 2 virus is positive selection, where amino acid changes in the envelope confer a selective advantage. To test for positive selection, we estimated the rate of nonsynonymous to synonymous substitutions (*dn*/*ds*) between the two viral populations and compared it to the estimate of *dn*/*ds* within populations. As discussed by Crandall *et al.* (1999), estimates of *dn*/*ds* based on pairwise differences can be biased due to the phylogenetic relationship between sequences. As baseline plasma virus and year 2-isolated virus clustered separately in the maximum likelihood tree obtained with DnaML, we were able to use a phylogenetic test where we estimate the *dn*/*ds* ratio for the branch connecting the most recent common ancestor of the year 2-isolated virus and the plasma virus (which represents the genetic differences between the two populations, Fig. 1). To test for differences in the selection pressure within and between viral samples, we estimated *dn*/*ds* under two models. Model 1 assumed that the *dn*/*ds* ratio is the same for all branches in the phylogeny (Goldman and Yang, 1994). Model 2 assumes that the *dn*/*ds* ratio in the branch connecting the year 2-isolated virus from the plasma virus is different from that in the rest of the phylogeny. The difference in log-likelihood between the two models is a measure of the improvement in fit of model 2 over model 1. The statistical significance of the improvement in fit of model 2 over model 1 was established by comparing the observed improvement in fit to the fit obtained from simulated datasets (a technique known as "parametric bootstrapping"). We used 100 simulated datasets per patient generated using the parameter estimates from model 1 (no difference in *dn*/*ds*) with the program LISTTREE (Z. Yang, Department of Biology, University College, London). Models 1 and 2 were fitted to each of these simulated datasets to obtain a distribution of twice the difference in log-likelihood under the null hypothesis that *dn*/*ds* between time points is the same as within time points, to which the value estimated from the observed dataset was compared.

Estimates of *dn*/*ds* may be unusually high in the presence of G to A hypermutation, which results in a "dead-end" virus. As such substitutions will be found in the tips of the phylogenetic tree, our test for selection, which focuses on substitutions in an internal branch of the phylogeny, and compares it with the rest, is conservative in the presence of G to A hypermutations, which will artificially raise the "background" *dn*/*ds* rate.

Testing for clusters of amino acid substitutions

We determined which amino acids differed between baseline virus and year 2 virus, taking into account the phylogenetic structure. Specifically, the ancestral codon sequence at either end of the phylogenetic branch representing the divergence between baseline plasma virus and year 2-isolated virus was reconstructed for each patient using a marginal reconstruction method.

We wished to determine whether amino acid substitutions between baseline virus and year 2 virus were clumped together. We performed a simple test where the distance (in residues) subtended by the amino acid substitutions was compared to that expected if substitutions

were scattered at random along the sequence. One hundred thousand control datasets per patient were produced by randomising the positions of the substitutions. The proportion of randomised datasets where amino acid substitutions are more clumped than observed (i.e., subtend a shorter length of sequence) gives an estimate of the *P* value. A program written in Mathematica v.3.0 (Wolfram Research) is available on request to perform this test.

ACKNOWLEDGMENTS

We thank Ziheng Yang for help with the parametric bootstrapping and Dan Haydon for comments on the manuscript. This work was funded by a Medical Research Council Research Fellowship G81/298 (S.D.W.F.), the Swiss National Science Foundation (Grant 84AD-046176) (H.F.G.), National Institute of Allergy and Infectious Diseases (Grant AI 01361 (J.K.W.), AI 27670 (D.D.R.), AI 28858 (D.D.R.), AI 36214 (D.D.R.), and AI 29164 (Center for AIDS Research, San Diego)) and an unrestricted gift from Roche Molecular Systems, Alameda, California.

REFERENCES

Bonhoeffer, S., Holmes, E. C., and Nowak, M. A. (1995). Causes of HIV diversity. *Nature* **376,** 125.

Chun, T. W., Stuyver, L., Mizell, S. B., *et al.* (1997). Presence of an inducible HIV-1 latent reservoir during highly active antiretroviral therapy. *Proc. Natl. Acad. Sci. USA* **94,** 13193–13197.

Cichutek, K., Merget, H., Norley, S., *et al.* (1992). Development of a quasispecies of human immunodeficiency virus type 1 *in vivo. Proc. Natl. Acad. Sci. USA* **89,** 7365–7369.

Crandall, K. A., Kelsey, C. R., Imamichi, H., Lane, H. C., and Salzman, N. P. (1999). Parallel evolution of drug resistance in HIV: Failure of nonsynonymous/synonymous substitution rate ratio to detect selection. *Mol. Biol. Evol.* **16,** 372–382.

Delwart, E. L., Pan, H., Neumann, A., and Markowitz, M. (1998). Rapid, transient changes at the env locus of plasma human immunodeficiency virus. *J. Virol.* **72,** 2416–2421.

Finzi, D., Kermankova, M., Pierson, T., *et al.* (1997). Identification of a reservoir for HIV-1 in patients on highly active antiretroviral therapy. *Science* **278,** 1295–1300.

Finzi, D., Blankson, J., Siliciano, R., *et al.* (1999). Latent infection of $CD4^+$ T cells provides a mechanism for lifelong persistence of HIV-1, even in patients on effective combination therapy. *Nat. Med.* **5,** 512–517.

Goldman, N., and Yang, Z. (1994). Codon-based model of nucleotide substitution for protein-coding DNA-sequences. *Mol. Biol. Evol.* **11,** 725–736.

Gulick R. M., Mellors, J. W., Havlir, D., *et al.* (1998). Simultaneous vs. sequential initiation of therapy with indinavir, zidovudine, and lamivudine for HIV-1 infection—100-week follow-up. *J. Am. Med. Assoc.* **280,** 35–41.

Günthard, H. F., Frost, S. D. W., Leigh Brown, A. J., *et al.* (1999). Evolution of envelope sequence of HIV-1 in cellular reservoirs in the setting of potent antiviral therapy. *J. Virol.* **73,** 9404–9412.

Günthard, H. F., Havlir, D. V., Fiscus, S., *et al.* (2001). Residual human immunodeficiency virus-1 RNA and DNA in lymph node and HIV RNA in genital secretions and in cerebrospinal fluid after two years of suppression of viremia in the Merck 035 cohort. *J. Inf. Dis.* **183,** 1318–1327.

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

Kwong, P. D., Wyatt, R., Robinson, J., Sweet, R. W., Sodroski, J., and Hendrickson, W. A. (1998). Structure of an HIV gp120 glycoprotein in complex with the CD4 receptor and a neutralising human antibody. *Nature* **393,** 648–659.

Liu, S. L., Schacker, T., Musey, L., *et al.* (1997). Divergent patterns of progression to AIDS after infection from the same source: Human immunodeficiency virus type 1 evolution and antiviral responses. *J. Virol.* **71,** 4284–4295.

McDonald, R. A., Mayers, D. L., Chung, R. C, *et al.* (1997). Evolution of human immunodeficiency virus type 1 env sequence variation in patients with diverse rates of disease progression and T cell function. *J. Virol.* **71,** 1871–1879.

Nei, M., and Jin, L. (1989). Variances of the average numbers of nucleotide substitutions within and between populations. *Mol. Biol. Evol.* **6,** 290–300.

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

Nijhuis, M., Boucher, C. A. B., Schipper, P., Leitner, T., Scuurman, R., and Albert, J. (1998). Stochastic processes strongly influence HIV-1 evolution during suboptimal protease inhibitor therapy. *Proc. Natl. Acad. Sci. USA* **95,** 14441–14446.

Pitcher, C. J., Quittner, C., Peterson, D. M., *et al.* (1999). HIV-1-specific $CD4^+$ T cells are detectable in most individuals with active HIV-1 infection, but decline with prolonged viral suppression. *Nat. Med.* **5,** 518–525.

Rose, P. P., and Korber, B. T. (2000). Detecting hypermutations in viral sequences with an emphasis on G $\rightarrow$ A hypermutation. *Bioinformatics* **16,** 400–401.

Simmonds, P., Balfe, P., Ludlam, C. A., *et al.* (1990). Analysis of sequence diversity in hypervariable regions of the external glycoprotein of human immunodeficiency virus type 1. *J. Virol.* **64,** 5840–5850.

Wolfs, T. F. W, Zwart, G., Bakker, M., and Goudsmit, J. (1992). HIV-1 genomic RNA diversification following sexual and parenteral virus transmission. *Virology* **189,** 103–110.

Wong, J. K., Hezareh, M., Gunthard, H. F., *et al.* (1997). Recovery of replication competent HIV despite prolonged suppression of plasma viremia. *Science* **278,** 1291–1295.

Yamaguchi, Y., and Gojobori, T. (1997). Evolutionary mechanisms and population dynamics of the third variable envelope region of HIV within single hosts. *Proc. Natl. Acad. Sci. USA* **94,** 1264–1269.

Yamaguchi, Y., and Gojobori, T. (2000). Reevaluation of amino acid variability of the human immunodeficiency virus type 1 gp120 envelope glycoprotein and prediction of new discontinuous epitopes. *J. Virol.* **74,** 4335–4350.

Zhang, L., Ramratnam, B., Tenner-Racz, K., *et al.* (1999). Quantifying residual HIV-1 replication in patients receiving combination antiretroviral therapy. *N. Engl. J. Med.* **340,** 1605–1613.