Parallel Evolution of Drug Resistance in HIV: Failure of Nonsynonymous/ Synonymous Substitution Rate Ratio to Detect Selection

Keith A. Crandall, Chris R. Kelsey,* Hiromi Imamichi,† H. Clifford Lane,‡ and Norman P. Salzman†*

*Department of Zoology, Brigham Young University; †SAIC Frederick, Frederick Cancer Research & Development Center, National Cancer Institute; and ‡Laboratory of Immunoregulation, National Institute of Allergy and Infectious Diseases, National Institutes of Health

Parallel or convergent evolution at the molecular level has been difficult to demonstrate especially when rigorous statistical criteria are applied. We present sequence data from the protease gene from eight patients infected with the human immunodeficiency virus (HIV-1). These patients have been on multiple drug therapies for at least 2 years. We present sequence data from two timepoints: time zero—the initiation of drug therapy—and a subsequent timepoint between 59 and 104 weeks after the initiation of drug therapy. In addition to the sequence data, we present viral load data from both initial and final timepoints. Our phylogenetic analyses indicate significant evolution of virus from initial to final time points, even in three of eight patients who show low viral loads. Of the five patients who escaped drug therapy, identical amino acid replacements were seen in all five patients at two different codon positions, an indication of parallel evolution. We also measured genetic diversity for these patients and found no correlation between genetic diversity and viral load. Finally, we calculated the nonsynonymous and synonymous substitution rates and showed that the ratio of nonsynonymous to synonymous substitution compared to the value of one may be a poor indicator of natural selection.

Introduction

The most successful treatment for acquired immunodeficiency syndrome (AIDS), caused by the human immunodeficiency virus (HIV), has been drug therapies, initially those inhibiting the reverse transcriptase activity of the virus (e.g., AZT). Recent advances in drug therapy have focused on the inhibitors of the protease gene. Indeed, a combination therapy involving both reverse transcriptase and protease inhibitors has proven to be our best ammunition yet in the fight against AIDS (Perelson et al. 1997). In a recent article, Wain-Hobson (1997) has suggested that these combined drug therapies might even be a cure for HIV infection. The data show that multidrug therapies have the ability to reduce viral load greatly in a short period of time. Perelson et al. (1997) showed that the concentration of HIV-1 in plasma could drop by 99% in the first 2 weeks of treatment. Yet, it is the unaccounted for 1% that is troublesome, because HIV is known to hide out in various tissues, creating latent reservoirs of virus (Chun et al. 1997*a*; Finzi et al. 1997). We know HIV can evolve resistance to both reverse transcriptase inhibitors (Kellam and Larder 1995) and protease inhibitors (Borman, Paulous and Clavel 1996; Molla et al. 1996; Zhang et al. 1997). We also know that resistant genotypes exist in HIV populations in reasonable frequencies even in untreated patients (Kozal et al. 1996; Lech et al. 1996). Thus, the major question in the drug therapy of HIV is this: if the virus remains in latent reservoirs, is it still evolving, and will it eventually evolve resistance to drug therapy using a combination of drugs? We addressed this question by sequencing DNA from the protease gene from eight patients who were subjected to combination drug therapy. Multiple sequences were obtained from two timepoints for each patient: day 0, the day drug therapy began, and a timepoint at least 59 weeks after initiation of therapy. Our results indicate that while the combination drug therapy reduces viral load to undetectable levels, the virus is still evolving, and the HIV population is acquiring drug-resistant mutations. Furthermore, these resistant mutations accumulate in a parallel fashion, with different patients obtaining identical amino acid replacements in many positions.

Abbreviations: AIDS, acquired immunodeficiency syndrome; HIV, human immunodeficiency virus; IL, interleukin; nt, nucleotide; PCR, polymerase chain reaction.

Key words: human immunodeficiency virus, phylogeny, drug resistance, drug therapy, convergent evolution, parallel evolution.

Address for correspondence and reprints: Keith A. Crandall, Department of Zoology, 574 Widtsoe Building, Brigham Young University, Provo, Utah 84602-5255. E-mail: keith_crandall@byu.edu.

Mol. Biol. Evol. 16(3):372–382. 1999

Materials and Methods

The Patients and Therapy

Plasma samples were obtained from patients infected with HIV-1 who were treated with Indinavir (Vacca et al. 1994) (600 mg, four times a day) alone (patients 2, 3, 4, 5, and 6) or in combination with intermittent administration of interleukin-2 (IL-2) by continuous infusion for 5 days every 2 months (patients 1, 7, and 8) (table 1). The protocol was approved by the NIAID Institutional Review Board, and all participants provided written informed consent. The patients had previously received nucleoside analogs. At subsequent times, reverse transcriptase inhibitors were also administered in the continued presence of Indinavir, and patients were allowed to add IL-2 to their treatment regime. Particle-associated plasma HIV-1 RNA levels were determined with the version of the standard branched-DNA signal amplification assay (bDNA assay, Chiron) (Dewar et al. 1994), which has a detection limit of 500 copies of HIV-1 RNA per ml.

Sequence Data

HIV-1 RNA was isolated from 130 μl of plasma by using the QIAamp HCV kit (Qiagen Inc). HIV-1

Table 1
Patient Summary for Drug Treatments and Viral Load (Zhang et al. 1997)

Patient	HIV Initial Load (RNA copies/ml)	HIV Final Load (RNA copies/ml)	Initial CD 4+ (cells/μl)	Final CD 4+ (cells/μl)
1.....	195,000	<500	441	1,088
2.....	119,000	<500	136	729
3.....	245,000	<10,000	226	1,201
4.....	346,000	89,560	62	257
5.....	215,000	40,570	87	305
6.....	1,288,000	425,500	19	36
7.....	168,000	500,300	316	322
8.....	559,000	51,230	36	319

RNA was reverse transcribed to cDNA using a primer (minus strand), 5′-TTGTTTTACATCATTAGTGTGGGC-3′ (nucleotides [nt] 3,626–3,650 of HIV-1 NL4.3) and avian myeloblastosis virus reverse transcriptase (cDNA cycle kit, Invitrogen Corporation). HIV-1 DNA corresponding to the *gag* (p7/p1/p6), protease, and part of reverse transcriptase was amplified by polymerase chain reaction (PCR) with a mixture of Klentaq1 (Ab Peptides, Inc.) and *pfu* (Stratagene) by using the following primer pair: forward primer (nt 1,882–1,905) 5′-GAAGCAATGAGCCAAGTAACAAAT-3′ and reverse primer (nt 3,544–3,567) 5′-GATATGTCCATTGGCCTTGCCCCT-3′. Nested PCR was carried out with the following primer pair: forward primer (nt 1,966–1,988) 5′-TTCAATTGTGGCAAAGAAGGGCAC-3′ and reverse primer (nt 3,501–3,524) 5′-TAAGTCTTTTGATGGGTCATAATA-3′. The PCR product was purified by using the QIAquick spin PCR purification kit (Qiagen Inc.). The nucleotide numbering is based on that of HIVNL4.3. The purified PCR products were ligated using the pCRII vector (Invitrogen Corporation). Dye-deoxy-labeled sequencing reactions were performed using the ABI PRISM Dye Terminator Cycle Sequencing Kit with AmpliTaq DNA Polymerase, FS (Perkin Elmer). The reaction products were resolved by electrophoresis on a 4.75% polyacrylamide gel and analyzed with an Applied Biosystems 377 automated sequencing system.

Phylogeny Reconstruction

Phylogenetic relationships among all sequences were estimated using the neighbor-joining method (Saitou and Nei 1987). We determined the appropriate model of evolution by using a maximum likelihood ratio test to test alternative models (Huelsenbeck and Crandall 1997). These tests were implemented using the program Modeltest (Posada and Crandall 1998). We also included sequences from known laboratory strains of HIV to guard against contamination and mislabeling problems (Korber et al. 1995). Relative support for various clades was estimated using 1,000 replications of the bootstrap procedure (Felsenstein 1985), coupled with the neighbor-joining method and optimized model of evolution. The phylogeny estimation, model testing, and bootstrap procedures were performed using PAUP* (Swofford 1998).

Resolution for relationships among sequences within a patient was low, as evaluated by the bootstrap procedure. Therefore, an alternative phylogeny estimation procedure, statistical parsimony, was used to reconstruct relationships among sequences from a single patient (Templeton, Crandall, and Sing 1992). This method has a demonstrated superiority to traditional phylogeny reconstruction techniques when levels of divergence are low among sequences (Crandall 1994, 1996), as they are in HIV sequences from within a single patient. Using this procedure, relationships were established within the 95% confidence limits set by equation (8) of Templeton, Crandall, and Sing (1992). This procedure has the added advantage of checking for recombination and displaying explicitly the changes along the branches that connect different sequences. Therefore, numbers of inferred synonymous and nonsynonymous substitutions were counted directly from these reconstructions and compared to those calculated below.

Genetic Diversity

A characteristic feature of HIV is extensive genetic variation. The importance of genetic variation for the pathogenesis of AIDS is still poorly understood, as is the relationship between genetic variation and disease progression (but see Strunnikova et al. 1995). The importance of HIV genetic variation in the presence of protease-inhibiting drugs is also poorly understood. This study explored this association by measuring genetic diversity at two time points from eight patients.

Genetic diversity can be measured in a number of different ways. We have used two very different methods to estimate relative levels of genetic diversity both within and among patients. The first method, borrowed from conservation biology, estimates the likelihood that two sequences are different in allelic state and incorporates this likelihood into a genetic diversity index (Crozier 1992). The calculations are based on the measured branch lengths of the phylogeny as an uncorrected percent divergence (Crozier and Kusmierski 1994). Branch lengths were estimated using PAUP* (Swofford 1998) and then entered into the program CONSERVE (Agapow 1997) to calculate genetic diversity indices. A 95% confidence interval was calculated for each genetic diversity estimate by analyzing 100 bootstrap trees generated by PAUP*.

Our second approach to estimating genetic diversity was use of a phylogenetic estimate of theta, $\theta = 2N_{ei}\mu$, where N_{ei} is the inbreeding effective population size, and μ is the per nucleotide mutation rate (Kuhner, Yamato, and Felsenstein 1995). The computer program COALESCE provides a maximum likelihood estimate of θ via a Metropolis–Hastings Markov Chain Monte Carlo method. This approach calculates the likelihood of the observed data, given a value of θ by sampling genealogies based on a coalescent distribution (Kuhner, Yamato, and Felsenstein 1995). This phylogenetic approach to the estimation of genetic diversity has been shown to be superior to nonphylogenetic-based methods (Felsenstein 1992). This method does not assume a correct phylogeny but rather samples multiple plausible ge-

Table 2
Likelihood Ratio Tests of Models of Molecular Evolution (Huelsenbeck and Crandall 1997; Posada and Crandall 1998)

Null Hypothesis	Models Compared	$-\ln L_0$	$-\ln L_1$	$-2\ln \lambda$	df	*P*
Equal base frequencies	H_0: JC69 H_1: F81	2,858.78	2,825.19	67.18	3	<0.000001
Equal ti/tv rates	H_0: F81 H_1: HKY85	2,825.19	2,729.03	192.3	1	<0.000001
Equal ti and equal tv rates	H_0: HKY85 H_1: GTR	2,729.03	2,724.15	9.76	3	0.044593
Equal rates among sites	H_0: HKY85 H_1: HKY85+Γ	2,729.03	2,607.36	243.3	1	<0.000001
Proportion of invariable sites	H_0: HKY85+Γ H_1: HKY85+Γ+invar	2,607.36	2,604.74	5.24	1	0.080359

NOTE.—Due to the performance of multiple tests, the significance level of rejection of the null hypothesis should be adjusted via the Bonferroni correction to $\alpha = 0.01$.

nealogies as defined by their likelihood score. The method does assume constant population size, no recombination, no selection, and no migration. A 95% confidence limit was obtained for each estimate by interpolating the θ value corresponding to a log-likelihood difference of two using Mathematica (Wolfram 1991). Theta estimates were obtained for each individual patient, but not for the two timepoints for each patient, due to small sample sizes within timepoints relative to this estimation procedure.

Nucleotide Substitutions

In molecular evolutionary studies, it is valuable to partition nucleotide substitutions into two classes: synonymous substitutions, those causing no change in the amino acid, and nonsynonymous substitutions, those resulting in an amino acid replacement. Strictly random evolution would result in a 3:1 ratio of nonsynonymous to synonymous substitutions. However, the majority of nonsynonymous substitutions are typically eliminated by purifying selection, the result being a predominance of synonymous substitutions (Miyata and Yasunaga 1980). When positive Darwinian selection occurs, then the nonsynonymous rate of substitution accelerates (Hughes and Nei 1988; Messier and Stewart 1997). Thus, the relative rates of synonymous to nonsynonymous substitutions can be good indicators of the amount and types of selection affecting a gene (Sharp 1997). We estimated the number of synonymous substitutions per synonymous site ($\hat{d}_S$) and the number of nonsynonymous substitutions per nonsynonymous site ($\hat{d}_N$), using the method of Nei and Gojobori (1986), with the Jukes–Cantor correction. The estimates were performed using the computer program MEGA (Kumar, Tamura, and Nei 1993).

Results and Discussion

Sequence Data

A minimum of nine early timepoint clones per patient and seven late timepoint clones per patient were sequenced for each patient. For most patients we sequenced 10 or more clones at each timepoint, for a total of 157 sequences of length 297 nt. These sequences have been deposited in GenBank under the following accession numbers: patient 1, AF101320–AF101339; patient 2, AF101340–AF101360; patient 3, AF024721–AF024730 and AF024802–AF024809; patient 4, AF024810–AF024818 and AF024873–AF024880; patient 5, AF024881–AF024890 and AF024937–AF024946; patient 6, AF024947–AF024956 and AF025063–AF025074; patient 7, AF025075–AF025085 and AF025147–AF025155; patient 8, AF025156–AF025164 and AF025276, AF025278–AF025287.

Among Patient Diversity

Comparing likelihood scores for the various models of evolution, we reject the null hypotheses of equal base frequencies, transition rate equals transversion rate (ti = tv), and rate homogeneity (table 2). Our analysis failed to reject the null hypothesis of equal transition rates and equal transversion rates and the null hypothesis of no invariable sites. We concluded that the HKY85+Γ model incorporating nucleotide frequency differences, ti/tv bias, and rate heterogeneity was the most appropriate for these data (table 2). The nucleotide frequencies averaged across sequences were A = 0.36547, C = 0.16246, G = 0.22699, and T = 0.24508. The ti/tv ratio was estimated by maximum likelihood to be 2.522028 with a kappa = 4.986982. The shape parameter of the gamma distribution for incorporating rate heterogeneity was estimated by maximum likelihood to be 0.421139.

→

FIG. 1.—Midpoint rooted neighbor-joining tree of sequences from the protease gene of HIV-1 collected from eight host individuals. The phylogeny was estimated using the HKY85+Γ model of evolution. Branch lengths are shown proportional to the amount of change along the branches; bootstrap percentages (based on 1,000 replications) are given for nodes defining major groups of sequences. Shaded symbols represent a timepoint greater than 59 weeks after treatment, while white symbols indicate samples at the initiation of drug therapy. Symbols: ● = patient 1; ♦ = patient 2; ⚽ = patient 3; ✚ = patient 4; ▲ = patient 5; ★ = patient 6; † = patient 7; ■ = patient 8.


Con_B
hivd31
hivnl43
hivny5cg
hivhan
hivhxb2
hivlai
hivjrfl
hivjrcsf
hivcam
55
58
85
93
90
hivsf2
61
89
80
78
62
76
79
99
59
hivmn
93
99
68
98
76
hivoyi
hivrf
63
61
72
61
87
63
73
59
93
54
79
81
95

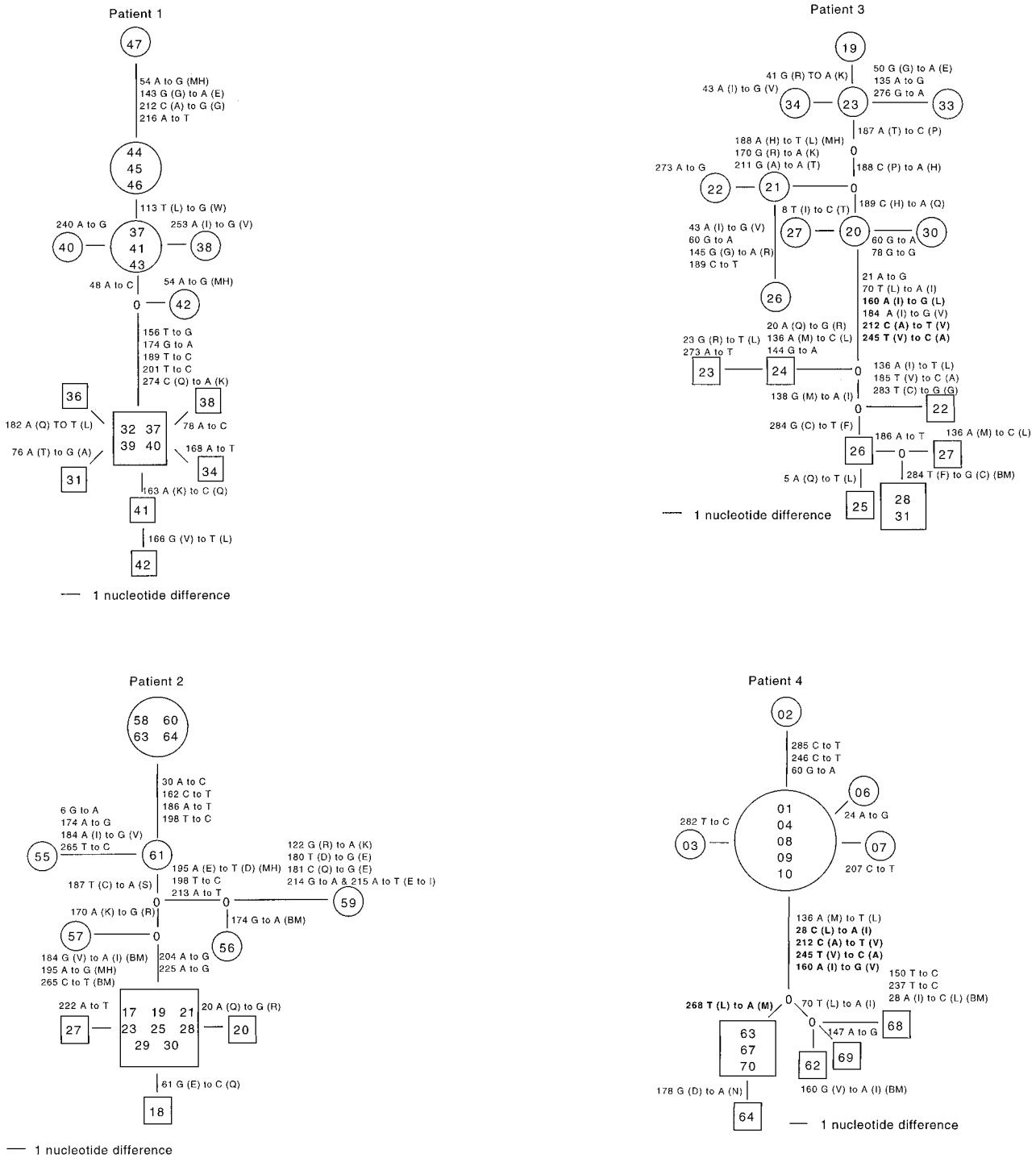
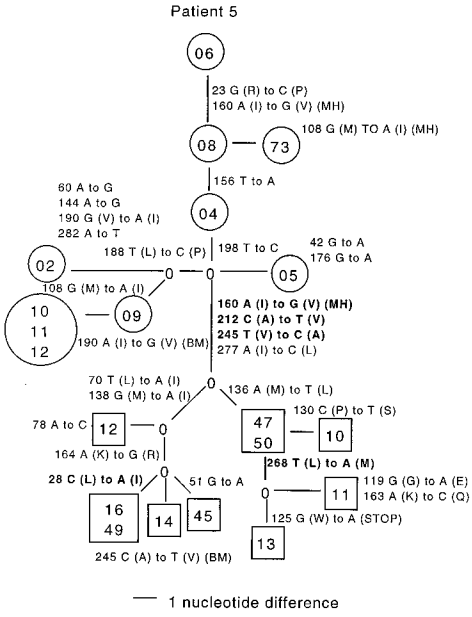
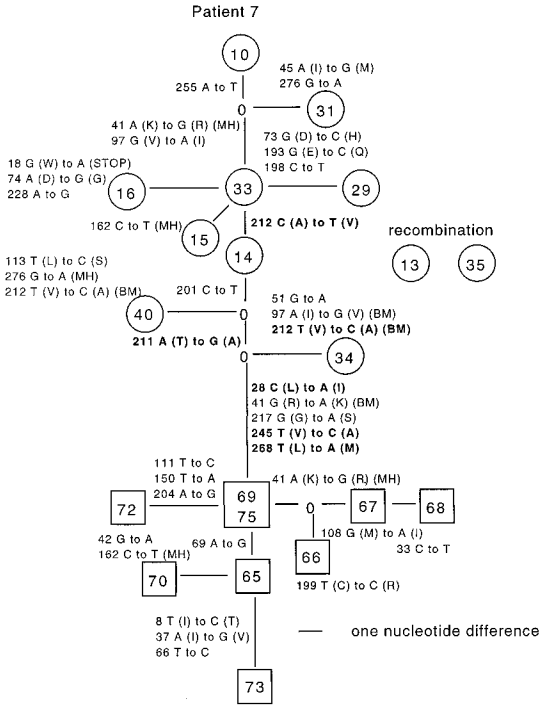
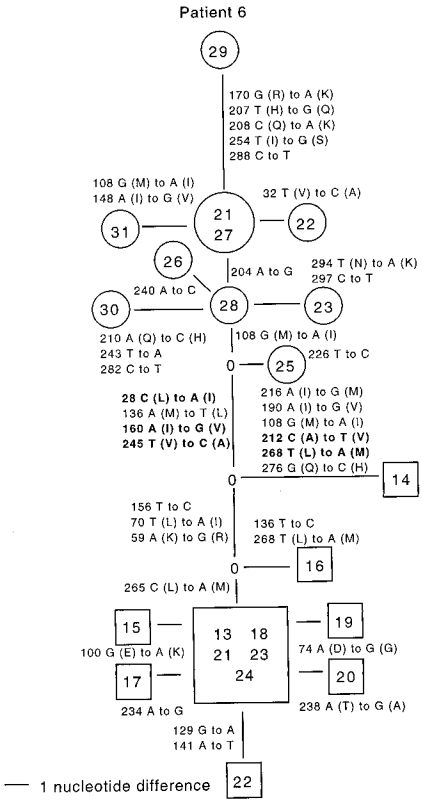


FIG. 2.—Unrooted cladograms relating nucleotide sequences from within each patient. Circles represent day zero samples; squares represent later timepoint samples as indicated. The clone number or numbers representing a unique sequence are shown inside the circle or square. Zeros represent missing intermediates. Nucleotide and amino acid changes are shown along the branch along which they are inferred to have occurred. The number represents the nucleotide position along the protease gene. BM stands for back mutation; MH indicates inferred multiple hits. The parallel changes are indicated in bold. Patient 7 has two sequences that appear to be recombinants under the criteria of Templeton, Crandall, and Sing (1992) and were therefore not connected to the main network.

Using this model of evolution, the neighbor-joining tree for the entire data set shows that sequences cluster predominantly by host individual (with the exception of patient 3 where the early timepoint sequences do not cluster strongly; fig. 1). Furthermore, no sequences clustered strongly (bootstrap values > 50) with known laboratory strains of HIV subtype B, an indication of no evidence for contamination. Even for those patients with undetectable levels of plasma virus (patients 1–3), it is clear that virus detectable via PCR is still evolving. In







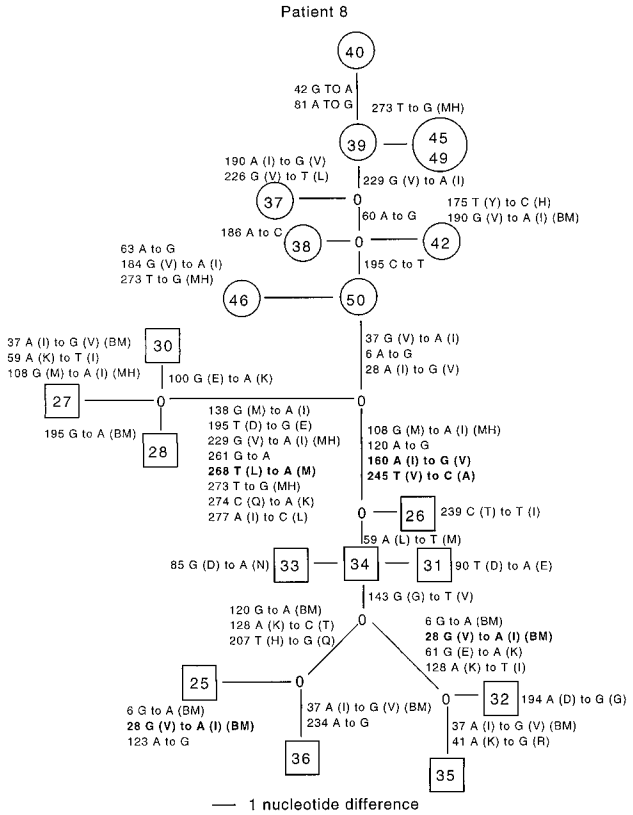


FIG. 2 *(Continued)*

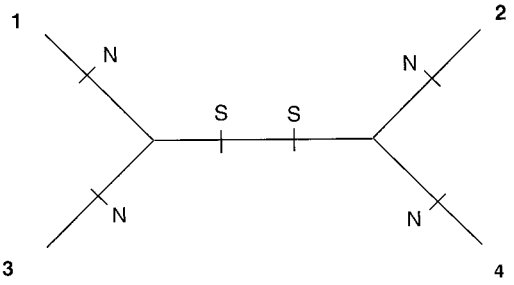


FIG. 3.—Unrooted tree with four terminal taxa (1–4) and synonymous (S) and nonsynonymous (N) changes indicated along the branches. Pairwise estimates of change overestimate changes on the internal branch and therefore can lead to biased estimates.

every case, the later timepoint sample diverged significantly from the initial sample in each patient. Indeed, with the exception of patient 4 (bootstrap = 59), all the later timepoints are supported by bootstrap values in the high 80s or higher.

It is of interest to note that the posttreatment isolates do not always arise from the pretreatment isolates. The expected pattern is shown in patient 2, where a monophyletic cluster of posttreatment timepoints arises from the pretreatment timepoints to form a paraphyletic group of samples from a single host individual. However, in patients 1, 4, 7, and 8 the posttreatment isolates are sister and basal to the pretreatment isolates. There are a number of potential explanations for this phenomenon. The biological explanation is that as virus is being eliminated by drug therapy, populations are being restocked by viral reservoirs. These reservoirs presumably are initially stocked at the stage of initial infection. Thus, one would expect these lineages to fall basal to those that have experienced many generations of evolution.

There are, of course, a number of methodological explanations. First, the tree may be wrong due to an incorrect model of evolution or violations of other assumptions of phylogenetic methods (e.g., independent sites). For example, patient 7 appears to have two sequences that may be the result of recombination (see fig. 2g). Standard phylogenetic techniques assume that no recombination has occurred among the sequences and therefore may give spurious relationships when this assumption is violated (Crandall and Templeton 1999). However, eliminating these two sequences from the analysis does not alter the general phenomenon. Second, the tree may be correct but the sampling may be poor; i.e., the initial pretreatment sample did not accurately reflect the entire diversity in the population, and by chance at a latter timepoint, one samples from a different part of the distribution of diversity. This explanation loses credence, for example, in patient 1, where the separation between timepoints is supported by bootstrap values of 99% and 76%. If the distribution of lineages was merely a sampling artifact, one would expect much less support for these nodes. Finally, the posttreatment clade may be driven to a sister and basal position due to a greater abundance of synapomorphic characters associated with drug resistance. However, when one reconstructs the tree based on just 3rd position changes (thereby eliminating parallel changes associated with drug resistance), this phenomenon is still observed (data not shown). One might argue that the placement is simply a rooting artifact, but no matter how one roots the tree, some patients will always show this phenomenon of later timepoints falling basal to earlier timepoints. Thus, the biological argument seems to hold the most weight.

Table 3
Estimates of Genetic Diversity (Crozier 1992) and Nucleotide Diversity ($\hat{\theta}$) (Kuhner et al. 1995)

Patient—Time	Genetic Diversity	95% Confidence Interval	$\hat{\theta}$	95% Confidence Interval
P1	0.0689	0.0676–0.0702	0.0254	0.0119–0.0503
P1—0	0.0409	0.0405–0.0413		
P1—104	0.0309	0.0299–0.0139		
P2	0.0896	0.0891–0.0901	0.0374	0.0206–0.0718
P2—0	0.0246	0.0243–0.0249		
P2—100	0.0712	0.0707–0.0718		
P3	0.1221	0.1211–0.1231	0.0608	0.0256–0.1159
P3—0	0.0737	0.0731–0.0743		
P3—75	0.0612	0.0604–0.0619		
P4	0.0608	0.0603–0.0612	0.0256	0.0194–0.0552
P4—0	0.0413	0.0410–0.0416		
P4—75	0.0203	0.0199–0.0207		
P5	0.1021	0.1011–0.1030	0.0453	0.0232–0.00854
P5—0	0.0638	0.0630–0.0646		
P5–59	0.0477	0.0470–0.0484		
P6	0.1202	0.1198–0.1207	0.0560	0.0237–0.1015
P6—0	0.0728	0.0726–0.0730		
P6—71	0.0557	0.0553–0.0562		
P7	0.1316	0.1305–0.1327	0.0746	0.0269–0.1403
P7—0	0.0688	0.0681–0.0694		
P7—60	0.0754	0.0744–0.0764		
P8	0.1591	0.1583–0.1599	0.0873	0.0277–0.1607
P8—0	0.1211	0.1204–0.1218		
P8—72	0.0490	0.0482–0.0497		

NOTE.—Time is given at the initiation of drug therapy (time 0) and in weeks after drug therapy began.

Within Patient Diversity

Diversity of HIV from within a patient, while less than among patients, was still significant. Relationships among sequences from within each patient showed clear partitioning of the early and late timepoint samples (fig. 2a–h). Early and late timepoint samples were separated by a minimum of 5 nt substitutions (patients 1, 2, 4, 5, and 7) and a maximum of 12 changes (patient 8), even for those patients with viral load levels <500 copies/ml (patients 1 and 2). There was no correlation between initial viral load and initial genetic diversity ($\rho = 0.408$; tables 1 and 3). Likewise, there was no correlation between final viral loads and final genetic diversity, an indication that viral load is a poor indicator of underlying genetic diversity. Because genetic diversity is defined as $\theta = 2N_{ei}\mu$, this lack of correlation also implies, assuming a constant mutation rate, that viral load does not correlate with the inbreeding effective population size. Overall genetic diversity measured by Crozier's (1992) method or Kuhner et al.'s (1995) method does not correlate with initial viral load ($\rho = 0.313$ and $\rho =$

Table 4
The Number of Synonymous Substitutions per Synonymous Site ($\hat{d}_S$), the Number of Nonsynonymous Substitutions per Nonsynonymous Site ($\hat{d}_N$) (Nei and Gojobori 1986), and the Total Numbers of Synonymous (SYN) and Nonsynonymous (NONSYN) Inferred from the Cladograms in Figure 2

Patient—Time	$\hat{d}_S$	SD	$\hat{d}_N$	SD	$\hat{d}_N/\hat{d}_S$	SYN	NON-SYN
P1	0.0481	0.0182	0.0074	0.0031	0.154	10	9
P1—0	0.0139	0.0072	0.0050	0.0028	0.360		
P1—104	0.0056	0.0046	0.0043	0.0022	0.768		
P2	0.0516	0.0167	0.0092	0.0034	0.178	12	10
P2—0	0.0560	0.0187	0.0118	0.0040	0.0211		
P2—100	0.0026	0.0027	0.0016	0.0011	0.0615		
P3	0.0320	0.0118	0.0306	0.0074	0.956	8	23
P3—0	0.0222	0.0098	0.0160	0.0053	0.721		
P3—75	0.0228	0.0125	0.0120	0.0046	0.526		
P4	0.0162	0.0055	0.0154	0.0056	0.951	8	8
P4—0	0.0193	0.0080	0.0000	0.0000	0.000		
P4—75	0.0123	0.0071	0.0089	0.0042	0.724		
P5	0.0199	0.0080	0.0247	0.0066	1.24	9	18
P5—0	0.0291	0.0122	0.0091	0.0042	0.313		
P5—59	0.0102	0.0074	0.0152	0.0054	1.49		
P6	0.0252	0.0100	0.0259	0.0069	1.03	11	24
P6—0	0.0252	0.0107	0.0087	0.0030	0.345		
P6—71	0.0119	0.0053	0.0103	0.0031	0.866		
P7	0.0353	0.0107	0.0227	0.0065	0.643	15	18
P7—0	0.0344	0.0116	0.0116	0.0042	0.337		
P7—60	0.0303	0.0114	0.0069	0.0033	0.228		
P8	0.0421	0.0137	0.0328	0.0068	0.779	11	26
P8—0	0.0339	0.0131	0.0072	0.0035	0.212		
P8—72	0.0331	0.0132	0.0367	0.0078	1.11		

0.243, respectively). These two measures of genetic diversity, however, do correlate strongly with each other ($\rho = 0.985$; table 3), an indication that they are measuring similar properties of the sequence albeit in very different ways.

Nonsynonymous and Synonymous Substitutions

Nonsynonymous and synonymous substitution rates were measured at initial timepoints and end timepoints, and averaged across all sequences (table 4). In addition, the raw numbers of inferred nonsynonymous and synonymous substitutions were tabulated from the cladograms in figure 2 (table 4). For all patients except patient 3, the synonymous substitution rate decreased at the later timepoint from the earlier timepoint. The nonsynonymous substitution rate, on the other hand, decreased in the three patients with undetectable levels of virus (patients 1–3) and patient 7, whereas, this rate increased in the other patients. The synonymous substitution rate is always significantly higher ($P < 0.0005$, Student's t) than the nonsynonymous substitution rate at time zero. However, at the terminal timepoint, two patients (5 and 8) show significantly higher nonsynonymous substitution rates ($P < 0.0005$, Student's t), while the remaining patients show less discrepancy between the synonymous and nonsynonymous rates.

A common convention for detecting the action of positive selection is the calculation of the ratio of nonsynonymous to synonymous substitutions, a ratio greater than one indicating strong positive selection (Messier and Stewart 1997). In only two patients examined do we find overall ratios greater than one (patients 5 and 6; table 4). Likewise, the ratio at the later timepoint shows only two patients with a ratio greater than one (patients 5 and 8). However, all patients except 3 and 7 show an increase in this ratio from early timepoint to later timepoint. Sharp (1997) has argued that adaptive changes are very hard to find using this ratio, particularly with the ratio greater than one cutoff. This point is demonstrated here. Clearly these sequences are under positive Darwinian selection as is evident by the parallel changes shown below (table 5). However, the magic ratio of 1:1 is reached in only two of the five patients who have escaped drug therapy. Even in those patients with continued low viral loads (<500 copies/ml), we see a significant shift in the ratio of nonsynonymous to synonymous substitutions in favor of nonsynonymous substitutions. Therefore, Sharp's (1997) prediction that these ''ratios may often be less than one even if some adaptive substitutions have occurred'' is born out in

Table 5
Convergent Amino Acid Replacements in Response to Drug Treatment

	Codon (nucleotides)				
Patient—Time	10 (28–30)	54 (160–162)	71 (211–213)	82 (244–246)	90 (268–270)
P1—0	CTC	ATC	GCT	GTC	TTG
P1—104	No change	No change	No change	No change	No change
P2—0	ATA	ATC	ACA	GTC	TTG
P2—100	No change	No change	No change	No change	No change
P3—0	ATC	ATC	GCT	GTC	TTG
P3—75	No change	GTC	GTT	GCC	No change
P4—0	CTC	ATC	GCT	GTC	TTG
P4—75	ATC	GTC	GTT	GCC	ATG
P5—0	CTT	ATC	GCT	GTC	TTG
P5—59	ATT	GTC	GTT	GCC	ATG
P6—0	CTC	ATC	GCT	TGC	TTG
P6—71	ATC	GTC	GTT	GCC	ATG
P7—0	CTC	ATC	GTT	GTC	TTG
P7—60	ATC	No change	No change	GCC	ATG
P8—0	ATC	ATC	GCT	GTC	TTG
P8—72	GTC	GTC	No change	GCC	ATG

these data. Furthermore, calculating the raw numbers of nonsynonymous and synonymous substitutions directly from the inferred cladogram seemed to give a better indication of the selection pressures experienced by these viruses. For every individual with viral loads greater than 500 copies/ml, the nonsynonymous to synonymous substitution ratio (in raw numbers not in rates) was greater than one. Thus, these raw calculations appear to be better indicators of recent positive selection than the rate estimates, compared to the yardstick of a ratio being greater than or less than one.

Here we have used the Nei and Gojobori method of estimating synonymous and nonsynonymous substitution rates. Maximum likelihood approaches provide enhanced methods for estimating these rates (Goldman and Yang 1994; Muse 1996; Nielsen and Yang 1998; Yang 1998). However, these alternative methods are difficult to implement and therefore are not widely used. We chose to use the Nei and Gojobori method for demonstrative purposes and will explore these other estimation procedures in the future.

There are at least two reasons why the method of Nei and Gojobori could give spurious results. First, as a pairwise method, it counts events on internal branches multiple times. Consider the four taxon statement given in figure 3. The changes on the internal branch are counted multiple times and therefore give an average pairwise estimate of 3:2 nonsynonymous to synonymous. But the actual ratio is 4:2. This underestimate is due to the positioning of the synonymous substitutions on the internal branch and these changes being counted multiple times via pairwise comparisons. Second, while the Nei and Gojobori method corrects for multiple hits, it does not correct for nucleotide frequency differences, transition/transversion bias, or codon frequency bias that have been shown to have a significant effect on the estimation of substitution rates (Muse 1996; Nielsen and Yang 1998). Therefore methods that take into account these biases and do not depend on pairwise comparisons, such as the maximum likelihood methods (Muse and Gaut 1994; Goldman and Yang 1994), are expected to give more reliable results. Perhaps the selection is proportionally on a few sites, such that even if the nonsynonymous rate is estimated correctly, it will still be swamped by synonymous changes at other positions, which is why the magic ratio of 1:1 should not be the ultimate goal for detecting selection.

Parallel Evolution of Drug-Resistant Mutations

Convergent evolution implies adaptive change, by natural selection, resulting in lesser-related entities appearing more related than they are (Doolittle 1994). Parallel changes are those resulting from a common ancestral character state but in independent lineages, whereas, convergent changes are those evolving from different ancestral states. Doolittle (1994) characterized a number of different types of convergent evolution, including sequence convergence. Of sequence convergence, he concluded, ''a convincing case for genuine sequence convergence has yet to be made.'' Indeed, convergent evolution at the nucleotide sequence level seems to be so rare, one cannot find a reference to the phenomenon in the index of Li's (1997) *Molecular Evolution* (the ideas of convergent and parallel evolution are illustrated on p. 69, but without empirical examples). Doolittle presented a purported case of convergent evolution in HIV sequences from the envelope protein. Holmes et al. (1992) argued for convergent evolution based on the crown motif of the V3 region evolving to a GPGSAV motif along two independent lineages from a single patient, one from GPGRAV and the other from GPGSAF. Doolittle (1994) argued that the sample size was too small to conclude convergent evolution.

In our case, in sampling multiple independent individuals, we have identified identical changes occurring in each individual (table 5). We can categorize these changes unambiguously as parallel because we know them to have identical starting sequences. We have summarized the changes at five codon positions known to influence drug resistance (for more details, see Zhang et al. 1997). For example, in codon position 90, five of the five individuals showing increased levels of viral load have evolved a Leu (TTG) to Met (ATG) amino acid replacement. Using a simple binomial distribution with a $p_0 = 0.111$ (i.e., given a change has occurred in the TTG sequence, there is a probability of 1/9 changing to a Met as to any other possible amino acid or one stop codon option [TAG]), five of five individuals changing to the same amino acid has a $P < 0.0001$. Similarly, at position 82, six of six changes are from a Val (GTC) to Ala (GCC) replacement giving a $P < 0.0001$. The other three positions show similar patterns: 71 (four of four show Ala [GCT] to Val [GTT] $P = 0.0001$), 54 (five of six show Ile [ATC] to Val [GTC] $P = 0.0001$), and 10 (three of five show Leu [CTC] to Ile [ATC] $P = 0.0086$). Therefore we show multiple examples of statistically significant parallel evolution at the nucleotide level. The body of evidence demonstrating that convergent/parallel changes play an important evolutionary role is growing from both natural populations (e.g., bacterial drug resistance [Weisblum 1995] and viral drug resistance [Molla et al. 1996]) and experimental systems (Bull et al. 1997; Cunningham et al. 1997).

We chose to examine these five amino acid positions because they were the only five that appeared in five or more of the patients. Eleven positions are known to be associated with Indinavir resistance; 10 (nt 28–30), 20 (nt 58–60), 24 (nt 70–72), 32 (nt 94–96), 46 (nt 136–138), 54 (nt 160–162), 63 (nt 187–189), 71 (nt 211–213), 82 (nt 244–246), 84 (nt 250–252), and 90 (nt 268–270) (Schinazi, Larder, and Mellors 1997). All five of our observed parallel replacements fall within this list. There are many other amino acid replacements along the branches that separate timepoints in these patients. Many of the other known replacements are observed; e.g., in patient 3 (fig. 2c) we see replacements at both amino acid 24 (nt 70) and 46 (nt 136). However, in many individuals, novel amino acid replacements are also observed, e.g., in patient 3 amino acid position 62 (nt 184).

Wong et al. (1997) and Finzi et al. (1997) have argued that while multidrug therapy has failed to stop

replication of virus in latent reservoirs, viral evolution has stopped. Our results clearly show this not to be the case in these patients. In all the patients, viral load was reduced to <500 copies/ml (Zhang et al. 1997), and five of the eight escaped drug therapy via the evolution of drug resistance variants. Wong et al. (1997) defined evolution as the accumulation of drug-resistant amino acid replacements. To an evolutionary audience, this is indeed an unusual definition of evolution, typically defined as a change in gene frequencies at the population genetic level. Their study sampled too few clones per individual (two clones/individual; one early and one late) to get an estimate of frequency changes over time (Wong et al. 1997). Our samples of 10 clones per timepoint per individual are small, but they do show a clear difference in the frequency of drug resistance amino acid replacements at five codon positions. Thus evolution has occurred, implying active replication even when viral load is <500 copies/ml. Other recent studies have similar evidence of continued evolution even when viral loads are <50 copies/ml (Chun et al. 1997*b*; Imamichi et al., unpublished data). This suggests that either extremely small populations (undetectable with current assays) of virus are actively replicating, or latent reservoirs of virus are being activated and then are replicating *and* evolving. If latent reservoirs are being activated, the critical question then becomes: Are these reservoirs being replenished by this active replication or depleted with continual drug therapy? If they are being depleted, then patients could eventually cease drug treatment; if not, drug therapy will be a continual course of treatment. Discussions of discontinuing drug therapy and a cure for HIV (Wain-Hobson 1997) are at this point premature.

Acknowledgments

We thank an anonymous reviewer for helpful comments toward improving the manuscript and Nick Goldman for pointing out the potential difficulty in the pairwise estimates of synonymous/nonsynonymous substitution rates. This work was supported by an award from the Office of Research and Creative Activities at Brigham Young University (C.R.K.), and the Alfred P. Sloan Foundation and a Shannon award from the National Institutes of Health (K.A.C.).

LITERATURE CITED

AGAPOW, P. 1997. Conserve. 3.0.1. LaTrobe University, Melbourne, Australia.

BORMAN, A. M., S. PAULOUS, and F. CLAVEL. 1996. Resistance of human immunodeficiency virus type 1 to protease inhibitors: selection of resistance mutations in the presence and absence of the drug. J. Gen. Virol. **77**:419–426.

BULL, J. J., M. R. BADGETT, H. A. WICHMAN, J. P. HUELSENBECK, D. M. HILLIS, A. GULATI, C. HO, and I. J. MOLINEUX. 1997. Exceptional convergent evolution in a virus. Genetics **147**:1497–1507.

CHUN, T.-W., L. CARRUTH, D. FINZI et al. (15 co-authors). 1997*a*. Quantification of latent tissue reservoirs and total body viral load in HIV-1 infection. Nature **387**:183–188.

CHUN, T.-W., L. STUYVER, S. B. MIZELL, L. A. EHLER, J. M. MICAN, M. BASELER, A. L. LLOYD, M. A. NOWAK, and A. S. FAUCI. 1997*b*. Presence of an inducible HIV-1 latent reservoir during highly active antiretroviral therapy. Proc. Natl. Acad. Sci. USA **94**:13193–13197.

CRANDALL, K. A. 1994. Intraspecific cladogram estimation: accuracy at higher levels of divergence. Syst. Biol. **43**:222–235.

———. 1996. Multiple interspecies transmissions of human and simian T-cell leukemia/lymphoma virus type I sequences. Mol. Biol. Evol. **13**:115–131.

CRANDALL, K. A., and A. R. TEMPLETON. 1999. Statistical approaches to detecting recombination. *In* K. A. CRANDALL, ed. Evolution of HIV. Johns Hopkins University Press, Baltimore, Md. (in press).

CROZIER, R. H. 1992. Genetic diversity and the agony of choice. Biol. Conserv. **61**:11–15.

CROZIER, R. H., and R. M. KUSMIERSKI. 1994. Genetic distances and the setting of conservation priorities. Pp. 227–237 *in* V. LOESCHCKE, J. TOMIUK, and S. K. JAIN, eds. Conservation genetics. Birkhauser Verlag, Basel.

CUNNINGHAM, C. W., K. JENG, J. HUSTI, M. BADGETT, I. J. MOLINEUX, D. M. HILLIS, and J. J. BULL. 1997. Parallel molecular evolution of deletions and nonsense mutations in bacteriophage T7. Mol. Biol. Evol. **14**:113–116.

DEWAR, R. L., H. C. HIGHBARGER, M. D. SARMIENTO, J. A. TODD, M. B. VASUDEVACHARI, R. T. DAVEY, J. A. KOVACS, N. P. SALZMAN, H. C. LANE, and M. S. URDEA. 1994. Application of branched DNA signal amplification to monitor human immunodeficiency virus type 1 burden in human plasma. J. Infect. Dis. **170**:1172–1179.

DOOLITTLE, R. F. 1994. Convergent evolution: the need to be explicit. Trends Biochem. Sci. **19**:15–18.

FELSENSTEIN, J. 1985. Confidence limits on phylogenies: an approach using the bootstrap. Evolution **39**:783–791.

———. 1992. Estimating effective population size from samples of sequences: inefficiency of pairwise and segregating sites as compared to phylogenetic estimates. Genet. Res. Camb. **59**:139–147.

FINZI, D., M. HERMANKOVA, T. PIERSON et al. (12 co-authors). 1997. Identification of a reservoir for HIV-1 in patients on highly active antiretroviral therapy. Science **278**:1295–1300.

GOLDMAN, N., and Z. YANG. 1994. A codon-based model of nucleotide substitution for protein-coding DNA sequences. Mol. Biol. Evol. **11**:725–736.

HOLMES, E. C., L. Q. ZHANG, P. SIMMONDS, C. A. LUDLAM, and A. J. LEIGH BROWN. 1992. Convergent and divergent sequence evolution in the surface envelope glycoprotein of human immunodeficiency virus type 1 within a single infected patient. Proc. Natl. Acad. Sci. USA **89**:4835–4839.

HUELSENBECK, J. P., and K. A. CRANDALL. 1997. Phylogeny estimation and hypothesis testing using maximum likelihood. Annu. Rev. Ecol. Syst. **28**:437–466.

HUGHES, A. L., and M. NEI. 1988. Pattern of nucleotide substitution at major histocompatibility complex class I loci reveals overdominant selection. Nature **335**:167–170.

KELLAM, P., and B. A. LARDER. 1995. Retroviral recombination can lead to linkage of reverse transcriptase mutations that confer increased zidovudine resistance. J. Virol. **69**:669–674.

KORBER, B. T. M., G. LEARN, J. I. MULLINS, B. H. HAHN, and S. WOLINSKY. 1995. Protecting HIV databases. Nature **378**:242–243.

KOZAL, M. J., N. SHAH, N. SHEN et al. (11 co-authors). 1996. Extensive polymorphisms observed in HIV-1 clade B pro-

tease gene using high-density oligonucleotide arrays. Nature Med. **2**:753–759.

KUHNER, M. K., J. YAMATO, and J. FELSENSTEIN. 1995. Estimating effective population size and mutation rate from sequence data using Metropolis–Hastings sampling. Genetics **140**:1421–1430.

KUMAR, S., K. TAMURA, and M. NEI. 1993. MEGA: molecular evolutionary genetics analysis V 1.01. Pennsylvania State University, University Park.

LECH, W. J., G. WANG, Y. L. YANG, Y. CHEE, K. DORMAN, D. MCCRAE, L. C. LAZZERONI, J. W. ERICKSON, J. S. SINSHEIMER, and A. H. KAPLAN. 1996. In vivo sequence diversity of the protease of human immunodeficiency virus type 1: presence of protease inhibitor-resistant variants in untreated subjects. J. Virol. **70**:2038–2043.

LI, W.-H. 1997. Molecular evolution. Sinauer Associates, Inc., Sunderland, Mass.

MESSIER, W., and C.-B. STEWART. 1997. Episodic adaptive evolution of primate lysozymes. Nature **385**:151–154.

MIYATA, T., and T. YASUNAGA. 1980. Molecular evolution of mRNA: a method for estimating evolutionary rates of synonymous and amino acid substitution from homologous nucleotide sequences and its application. J. Mol. Evol. **16**:23–36.

MOLLA, A., M. KORNEYEVA, Q. GAO et al. (17 co-authors). 1996. Ordered accumulation of mutations in HIV protease confer resistance to ritonavir. Nature Med. **2**:760–766.

MUSE, S. V. 1996. Estimating synonymous and nonsynonymous substitution rates. Mol. Biol. Evol. **13**:105–114.

MUSE, S. V., and B. S. GAUT. 1994. A likelihood approach for comparing synonymous and nonsynonymous nucleotide substitution rates, with application to the chloroplast genome. Mol. Biol. Evol. **11**:715–724.

NEI, M., and T. GOJOBORI. 1986. Simple methods for estimating the numbers of synonymous and nonsynonymous nucleotide substitutions. Mol. Biol. Evol. **3**:418–426.

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.

PERELSON, A. S., P. ESSUNGER, Y. CAO, M. VESANEN, A. HURLEY, K. SAKSELA, M. MARKOWITZ, and D. D. HO. 1997. Decay characteristics of HIV-1-infected compartments during combination therapy. Nature **387**:188–191.

POSADA, D., and K. A. CRANDALL. 1998. Modeltest: testing the model of DNA substitution. Bioinformatics (in press).

SAITOU, N., and M. NEI. 1987. The neighbor-joining method: a new method for reconstructing phylogenetic trees. Mol. Biol. Evol. **4**:406–425.

SCHINAZI, R. F., B. A. LARDER, and J. W. MELLORS. 1997. Mutations in retroviral genes associated with drug resistance. Int. Antiviral News **5**:129–142.

SHARP, P. M. 1997. In search of molecular darwinism. Nature **385**:111–112.

STRUNNIKOVA, N., S. C. RAY, R. A. LIVINGSTON, E. RUBALCABA, and R. P. VISCIDI. 1995. Convergent evolution within the V3 loop domain of human immunodeficiency virus type 1 in association with disease progression. J. Virol. **69**:7548–7558.

SWOFFORD, D. L. 1998. PAUP* phylogenetic analysis using parsimony and other methods. 4.0.0beta. Sinauer Associates, Sunderland, Mass.

TEMPLETON, A. R., K. A. CRANDALL, and C. F. SING. 1992. A cladistic analysis of phenotypic associations with haplotypes inferred from restriction endonuclease mapping and DNA sequence data. III. Cladogram estimation. Genetics **132**:619–633.

VACCA, J. P., B. D. DORSEY, W. A. SCHLEIF et al. (23 co-authors). 1994. L-735,524: an orally bioavailable human immunodeficiency virus type 1 protease inhibitor. Proc. Natl. Acad. Sci. USA **91**:4096–4100.

WAIN-HOBSON, S. 1997. Down or out in blood and lymph? Nature **387**:123–124.

WEISBLUM, B. 1995. Erythromycin resistance by ribosome modification. Agents Chemother. **39**:577–585.

WOLFRAM, S. 1991. Mathematica. Addison-Wesley, Redwood City, Calif.

WONG, J. K., M. HEZAREH, H. F. GUNTHARD, D. V. HAVLIR, C. C. IGNACIO, C. A. SPINA, and D. D. RICHMAN. 1997. Recovery of replication-competent HIV despite prolonged suppression of plasma viremia. Science **278**:1291–1295.

YANG, Z. 1998. Likelihood ratio tests for detecting positive selection and application to primate lysozyme evolution. Mol. Biol. Evol. **15**:568–573.

ZHANG, Y.-M., H. IMAMICHI, T. IMAMICHI, H. C. LANE, J. FALLOON, M. B. VASUDEVACHARI, and N. P. SALZMAN. 1997. Drug resistance during indinavir therapy is caused by mutations in the protease gene and its gag substrate cleavage site. J. Virol. **71**:6662–6670.

ANTHONY M. DEAN, reviewing editor

Accepted November 26, 1998