Influence of Random Genetic Drift on Human Immunodeficiency Virus Type 1 *env* Evolution During Chronic Infection

Daniel Shriner,* Raj Shankarappa,*,[1] Mark A. Jensen,* David C. Nickle,* John E. Mittler,* Joseph B. Margolick[†] and James I. Mullins*,[‡,2]

**Department of Microbiology and [‡]Departments of Medicine and Laboratory Medicine, University of Washington School of Medicine, Seattle, Washington 98195-8070 and [†]Department of Molecular Microbiology and Immunology, Bloomberg School of Public Health, Johns Hopkins University, Baltimore, Maryland 21205*

Manuscript received January 14, 2003
Accepted for publication December 12, 2003

ABSTRACT

Human immunodeficiency virus type 1 (HIV-1) has high replication and mutation rates that generate large census populations and high levels of genetic variation. We examined the roles of natural selection, population growth, random genetic drift, and recombination in shaping the variation in 1509 C2–V5 *env* sequences derived from nine men with chronic HIV-1 infection. These sequences were obtained from clinical visits that reflect the first 6–13.7 years of infection. Pairwise comparisons of nonsynonymous and synonymous distances, Tajima's *D* test, Fu and Li's *D** test, and a test of recurrent mutation revealed evidence for episodes of nonneutral evolution in a total of 22 out of 145 blood samples, representing six of the nine individuals. Using three coalescent-based maximum-likelihood estimators, we found viral effective population sizes in all nine individuals to be $\sim 10^3$. We also show that a previous estimate of the effective population size of $\sim 10^5$ based on rare haplotype frequencies decreases to $\sim 10^3$ upon correcting a biased sampling procedure. We conclude that the genetic variation in these data sets can be explained by a predominance of random genetic drift of neutral mutations with brief episodes of natural selection that were frequently masked by recombination.

A hallmark of human immunodeficiency virus type 1 (HIV-1) infection is a clinical stage of chronic, generally asymptomatic infection of highly variable length, during which time viral genetic variation accumulates (reviewed in Viscidi 1999). The roles of population genetic forces in shaping viral genetic variation during chronic infection are unclear. The HIV-1 surface protein Env is thought to be under particularly strong natural selection because genetic variation generates antigenic diversity that facilitates escape from immune responses and allows the use of different coreceptors, thus providing alternate cellular targets (Overbaugh and Bangham 2001). Evidence supporting the importance of natural selection has come from immunological studies in which viral escape at $CD8^+$ cytotoxic T lymphocyte (CTL) epitopes within the *env* locus has been noted (Borrow *et al.* 1997; Evans *et al.* 1999; Soudeyns *et al.* 1999; Wilson *et al.* 1999). Genetic evidence for sites within *env* under positive natural selection has also been reported (Seibert *et al.* 1995; Nielsen and Yang 1998; Suzuki and Gojobori 1999; Yamaguchi-Kabata and Gojobori 2000; Yang *et al.* 2000).

The overall role of natural selection in the evolution of *env* is uncertain. First, the occurrence of viral escape within *env* during chronic infection and its importance to HIV-1 disease progression have not been critically evaluated in the context of population genetic theory. Second, none of the aforementioned studies demonstrated that putative escape variants fixed in the population faster than one would predict on the basis of random genetic drift alone. Third, HIV-1 has a high rate of recombination (Zhuang *et al.* 2002), and we and others have shown that high rates of recombination can create the false appearance of natural selection in a standard phylogenetic test in simulated data sets with neutral mutations (Anisimova *et al.* 2003; Shriner *et al.* 2003). Last, application of Tajima's (1989) *D* test to HIV-1 sequences has not revealed evidence for natural selection in *env* (Leigh Brown 1997).

If natural selection is more important than random genetic drift in shaping *env* genetic variation, then deterministic models of population genetics might be appropriate (Li 1976). Under such models, the population size is assumed to be much greater than the inverse mutation rate (Rouzine *et al.* 2001). Census population sizes of HIV-1 within infected individuals have been estimated to be $\geq 10^8$ (Piatak *et al.* 1993; Ho *et al.* 1995; Wei *et al.* 1995), which is substantially greater than the inverse HIV-1 mutation rate of $\sim 10^5$ (Mansky 1996). These large population sizes have been cited as evidence in support of deterministic models of population genetics (Coffin 1995).

[1]*Present address:* Center for Genomic Sciences, Allegheny-Singer Research Institute, Pittsburgh, PA 15212.

[2]*Corresponding author:* Department of Microbiology, University of Washington School of Medicine, Box 358070, Seattle, WA 98195-8070. E-mail: jmullins@u.washington.edu

However, the relevant parameter when modeling genetic changes within populations is the effective, not the census, population size, because it represents the number of individuals who contribute genetic information to successive generations. The effective population size, known as N_e, represents the size of an ideal population that experiences the same magnitude of random genetic drift as the observed population (Wright 1931; Hartl and Clark 1989). Different methods for estimating the viral effective population size have yielded widely divergent estimates, using *in vivo* measurements of HIV-1 *env* diversity. Leigh Brown (1997) studied one individual from whom 77 231-bp sequences encompassing the V3 loop were derived from five samples over 7 years of infection. After showing that these sequences did not depart from neutral expectations by Tajima's *D* test, Leigh Brown used a coalescent theory-based method that assumed neutrality, no recombination, and no population growth to estimate N_e. The results were striking: N_e was estimated to be $\sim 10^3$, substantially less than the inverse mutation rate. Leigh Brown concluded that the results of the test of neutrality and the low effective population size supported stochastic models of population genetics, wherein random genetic drift may be more important than natural selection in shaping the virus population.

In contrast, analysis of the same sequences by Rouzine and Coffin (1999) led to an estimation of N_e of $\sim 10^5$, a population size that supports deterministic models of population genetics. Their analysis used a linkage disequilibrium test based on haplotypes that should be missing in small populations, but are expected to be present in large populations. Recently, Seo *et al.* (2002) developed a pseudo-maximum-likelihood approach on the basis of coalescent theory to estimate N_e using serial sampled data. They analyzed a subset of Shankarappa *et al.*'s (1999) sequences derived from nine individuals and estimated N_e values on the order of 10^3, close to Leigh Brown's (1997) original estimate. However, they analyzed samples only during the period of time when viral divergence accumulated at a linear rate, did not assess neutrality, and did not account for population growth or recombination.

In this study, we explored the related issues of the viral effective population size and the roles of natural selection, population growth, random genetic drift, and recombination in shaping *env* diversity. We analyzed all 1300 sequences from Shankarappa *et al.* (1999), together with 209 follow-up sequences, using multiple statistical tests of neutrality and multiple estimators of N_e, including ones that explicitly account for population growth and recombination. The large number of serially sampled sequences permitted a rigorous statistical analysis of natural selection through time. These analyses, and reanalysis of Rouzine and Coffin's (1999) work, allowed us to propose a resolution to the discrepancy between low *vs.* high estimates of N_e in favor of low estimates. This work leads to a new understanding of the relative roles of natural selection, population growth, random genetic drift, and recombination in shaping *env* diversity during chronic HIV-1 infection.

MATERIALS AND METHODS

Study samples: The nine HIV-1 subtype B-infected individuals studied were described previously (Rinaldo *et al.* 1998; Shankarappa *et al.* 1999). Briefly, they represented incident cases of HIV infection among individuals followed in the Multicenter AIDS Cohort Study (MACS; Kaslow *et al.* 1987). They had moderate rates of disease progression. Selection for study was based on progression to advanced HIV disease following an asymptomatic period of at least 6 years, including an onset of decline in the total T-cell count, which we have shown precedes AIDS in most cases (Margolick *et al.* 1995; Gange *et al.* 1998; Rinaldo *et al.* 1998; Shankarappa *et al.* 1999). Partial *env* sequences were derived from specimens collected beginning at the first semiannual clinical visit the subject was found to be infected, shortly after seroconversion, and continuing semiannually for 6 to 13.7 years, until AIDS or death in some cases. These sequences averaged 639 bp and spanned the C2–V5 region of *env*. A total of 1509 sequences from 49 samples of plasma viral RNA and 96 samples of peripheral blood mononuclear cell-associated viral DNA were evaluated. The 209 follow-up sequences were generated as described previously (Shankarappa *et al.* 1999).

Sequence analysis: We used four statistical tests of neutrality. In all cases, the null hypothesis was that all mutations were neutral. First, an analysis of nonsynonymous and synonymous distances was performed using MEGA, version 2.1 (Nei and Kumar 2000; Kumar *et al.* 2001). We used the modified Nei-Gojobori method, which corrects for transition/transversion biases. The neutral expectation is that the number of nonsynonymous mutations per potential nonsynonymous site (d_n) should equal the number of synonymous mutations per potential synonymous site (d_s), or $d_n - d_s = 0$ (Nei and Gojobori 1986; Nei and Jin 1989; Nei and Kumar 2000; Kumar *et al.* 2001).

Second, Tajima's (1989) *D* test is a test based on the distribution of mutant frequencies. It compares the number of polymorphic sites (S) to the average pairwise number of nucleotide differences ($\hat{k}$). Assuming infinite sites (which means that each site may be mutated at most once), no population growth, no recombination, and neutral mutations at equilibrium, Watterson (1975) showed that the expected value of $S = a_n\theta$, in which $a_n = \Sigma_{i=1}^{n-1}(1/i)$ is a scalar for n sequences, $\theta = 2N_e\mu L$ is the mutation rate scaled by the effective population size for a haploid organism, and μ is the mutation rate per generation per site with L sites. Under the same assumptions, Tajima (1983) showed that the expected value of $\hat{k} = \theta$. Tajima's (1989) test is therefore based on the expectation that, on average, $\hat{k} - (S/a_n) = 0$. Thus, positive test statistics reflect an excess of intermediate-frequency mutations, suggesting diversifying selection, and negative test statistics reflect an excess of low-frequency mutations, suggesting negative selection (Tajima 1989) or selective sweeps (Braverman *et al.* 1995).

Third, Fu and Li's (1993) *D** test, also a test based on the distribution of mutant frequencies, compares the total number of mutations (η) to the number of singletons (η_s; *i.e.*, mutations that occur on terminal branches on an unrooted phylogeny) among n sequences. η and η_s can be estimated by counting the mutations as they are mapped onto the sample's phylogeny. With the same assumptions as given above for Tajima's test, the test is based on the expectation that, on

average, $(n/(n-1))\eta - a_n\eta_s = 0$. Thus, similar to Tajima's *D* test, positive *D** test statistics reflect an excess of "internal" or "old" mutations, *i.e.*, mutations at intermediate frequency, and negative test statistics reflect an excess of "external" or "new" mutations, *i.e.*, mutations at low frequency.

For Tajima's and Fu and Li's tests, we ascertained *P* values that accounted for recombination from null distributions of 10,000 independent replicates that were created by simulating sequences under a neutral coalescent model. We conditioned upon a population-scaled mutation rate $\theta = 38$ per locus and a population-scaled recombination rate $\rho = 9$ per locus (on the basis of the estimates from the RECOMBINE program, which is described below) with a locus of 639 sites and the appropriate sample size, using DnaSP, version 3.53 (Rozas and Rozas 1999). The sequences simulated for the power analysis of Tajima's test were generated with the same coalescent parameters using TREEVOLVE, version 1.32 (http://evolve.zoo.ox.ac.uk).

These three tests of neutrality each looked at the data in different ways. The test based upon synonymous and nonsynonymous distances uses a codon as its unit, and this test is applicable when the population is not in equilibrium. In contrast, Tajima's *D* test and Fu and Li's *D** test both use an individual nucleotide as their unit, but assume that the population is in equilibrium. Fu and Li's *D** test is generally more powerful in detecting population shrinkage and background selection (which refers to the variation-reducing effect of a negatively selected polymorphism on linked neutral polymorphism), whereas Tajima's *D* test is generally more powerful in detecting population growth and hitchhiking (which refers to the variation-reducing effect of a positively selected polymorphism on linked neutral polymorphism; Fu 1996, 1997).

The fourth test addresses the assumption of infinite sites. We first determined which samples had no sites with more than two cosegregating nucleotide states because the variation at such sites can be most parsimoniously explained by one mutational event. For the remaining samples, we tested whether more sites than expected were unambiguously recurrently mutated (*i.e.*, there were more than two cosegregating nucleotide states). To do this, we calculated how many sites were expected to have been recurrently mutated by assuming a Poisson mutational process with a mean conditioned upon the number of total sites and the number of variable sites. *P* values were generated from 10,000 independent replicates.

Estimating N_e: Effective population sizes were estimated using three coalescent-likelihood programs from the LAMARC package (Kuhner *et al.* 1995a). The first program, COALESCE, estimates θ assuming a single panmictic population of constant size without recombination or natural selection (Kuhner *et al.* 1995b). The second program, FLUCTUATE, estimates θ and *g*, the exponential growth rate (Kuhner *et al.* 1998). The third program, RECOMBINE, estimates θ and *r*, the ratio of the recombination rate to the mutation rate, such that $\rho = r\theta$ (Kuhner *et al.* 2000). Estimates of θ were generated from each of the three coalescent-likelihood programs using a transition/transversion ratio estimated from the data (Swofford 2002) and empirical base frequencies. We then calculated the corresponding values of N_e, assuming $\mu = 2.5 \times 10^{-5}$/site/generation for point substitutions (Mansky 1996). To qualitatively assess the effects of estimating θ for data with recombination under these three sets of assumptions, 10 independent replicates were generated under a neutral coalescent model with recombination, using TREEVOLVE. We conditioned upon a sample size of 10, $\theta = 38$/locus, $\rho = 9$/locus, and a locus of 639 sites, on the basis of the observed data and the results of RECOMBINE.

A fourth program in the LAMARC package, MIGRATE, relaxes the assumption of panmixis (Beerli and Felsenstein 1999). However, we generally did not consider population subdivision in our analyses for two reasons. First, we had no data for viruses isolated from anatomical sources other than blood. Second, all of these individuals appeared to be infected from a single source, on the basis of phylogenetic analysis (Shankarappa *et al.* 1999). Thus, we did not have to be concerned with migration from completely isolated populations of virus from other individuals, as would occur in cases of dual infection. Any migration that did occur had to be between anatomical or temporal compartments within an infected individual and had to be of viruses that shared a most recent common ancestor for the entire infection in the individual.

Phylogenetic analysis: Phylogenetic tree reconstruction and model of evolution estimation were performed in PAUP*, version 4.0 (Swofford 2002). Trees were reconstructed by first generating a neighbor-joining tree using maximum-likelihood distances and then swapping under maximum likelihood using the subtree pruning and regrafting (SPR) algorithm (Swofford 2002). Models of evolution were estimated using the general time-reversible model with unequal base frequencies and gamma-distributed rate heterogeneity across sites. Sites with gaps were excluded by pairwise deletion.

Multiple tests: Because we statistically tested a large number of samples, our tests needed to be corrected for multiple comparisons. A strict way to correct for multiple comparisons would be to apply a Bonferroni correction, which is accomplished by dividing the significance level by the number of tests performed (Rice 1989). However, all of the samples derived from a single individual have a shared evolutionary history and hence are correlated to varying degrees, making such a correction too extreme and thereby resulting in overly conservative conclusions. Because infection in each of the nine individuals represents independent iterates of viral evolution, we can justify at minimum nine independent comparisons, resulting in a correction of the significance level from 5% to 0.55%.

Nucleotide sequence accession numbers: The GenBank accession numbers are AF137629–AF138163, AF138166–AF138263, AF138305–AF138703, and AF204402–AF204670 for the 1300 sequences reported previously (Shankarappa *et al.* 1999) and AY348333–AY348528 and AY348532–AY348544 for the 209 follow-up sequences reported herein.

RESULTS

Testing neutrality: Each of the three N_e estimators we employed assumed neutrality, so we first addressed this assumption by applying three two-tailed statistical tests. The first test compared nonsynonymous mutations to synonymous mutations. Of 96 viral DNA samples examined (Figure 1, *), we found 1 sample (from participant 6) for which d_s exceeded d_n and 10 samples (from participants 2, 3, and 9) for which d_n exceeded d_s. Similarly, of 49 plasma viral RNA samples examined, no samples were found for which d_s exceeded d_n and 3 samples (from participants 2 and 7) were found for which d_n exceeded d_s (Figure 1, #). The samples for which neutrality was rejected by this test did not tend to cluster at any particular time during infection (Figure 1). Further, the signal for selection tended to be sporadic, as evident by the single samples in participants 6 and 7 that displayed nonneutral evolution. In summary, by this test, we found evidence for a sporadic excess of synonymous mutations in one of the nine individuals and for spo-

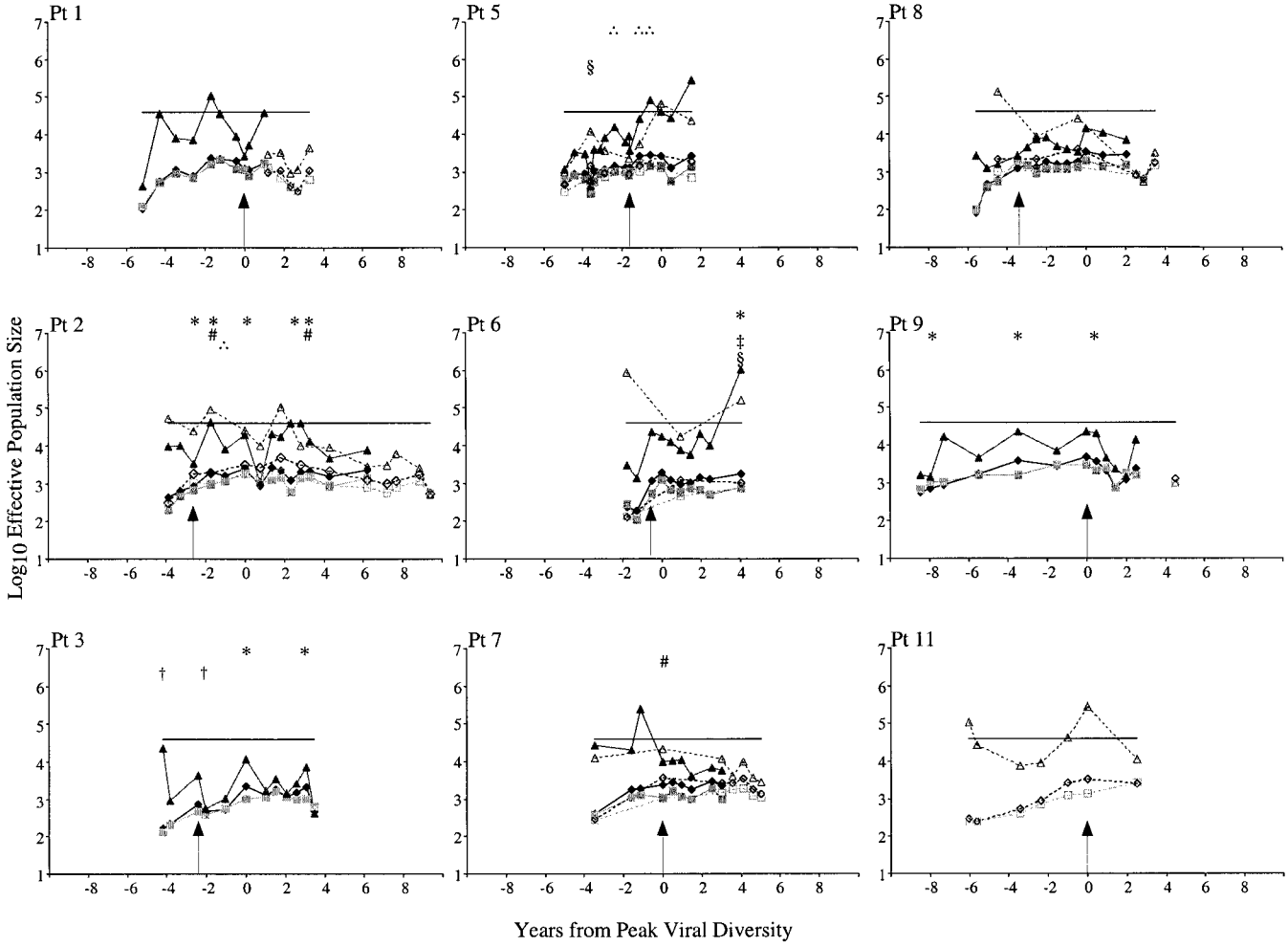

FIGURE 1.—Effective population sizes during chronic infection. Diamonds represent COALESCE estimates, squares represent RECOMBINE estimates, triangles represent FLUCTUATE estimates, and the horizontal line represents the boundary between stochasticity and determinism. Solid lines with solid symbols represent samples derived from peripheral blood cell viral DNA and dotted lines with open symbols represent samples derived from plasma virus RNA. * indicates DNA samples and # indicates RNA samples that tested significant using the modified Nei-Gojobori method. † indicates DNA samples that tested significant by Tajima's *D* test. ‡ indicates DNA samples and § indicates RNA samples that tested significant by Fu and Li's *D** test. ∴ indicates samples that tested significant by the test for recurrent mutation. The arrows point to the time of the first observed fixation event.

radic excesses of nonsynonymous mutations in four of the nine individuals.

We next addressed the infinite-sites assumption of Tajima's and Fu and Li's tests. Thirty-three samples had no sites with more than two cosegregating nucleotide states (data not shown). For the remaining 112 samples, 1 sample from participant 2 and 3 samples from participant 5 showed significantly more unambiguous recurrent mutation than expected (Figure 1, ∴). This result indicated that the assumption of a Poisson mutational process is largely valid for these samples. Furthermore, because only 4 samples were found to have experienced significant excesses of unambiguous recurrent mutation, diversifying selection was unlikely to have had a predominant role in the evolution of these sequences.

The preceding analysis underestimates recurrent mutation, because it ignores parallel changes and reversions. Recombination can induce spuriously inferred parallel changes and reversions on the sample's phylogeny. Furthermore, recombination is known to make Tajima's test conservative (TAJIMA 1989; WALL 1999; SCHIERUP and HEIN 2000). Figure 2 shows the distributions of Tajima's test statistic for coalescent simulations with and without recombination and demonstrates the reduction in variance induced by recombination. By comparing the 95% confidence intervals we calculated an average gain in power for Tajima's test of 11.8% by accounting for recombination.

Using the null distributions with recombination, significant departures from the neutral expectation of Tajima's test (with excesses of low-frequency mutations) were observed for 2 of the 145 samples, both from participant 3 (Figure 1, †). Only 1 sample that yielded significance occurred after genetic diversity stabilized [taken

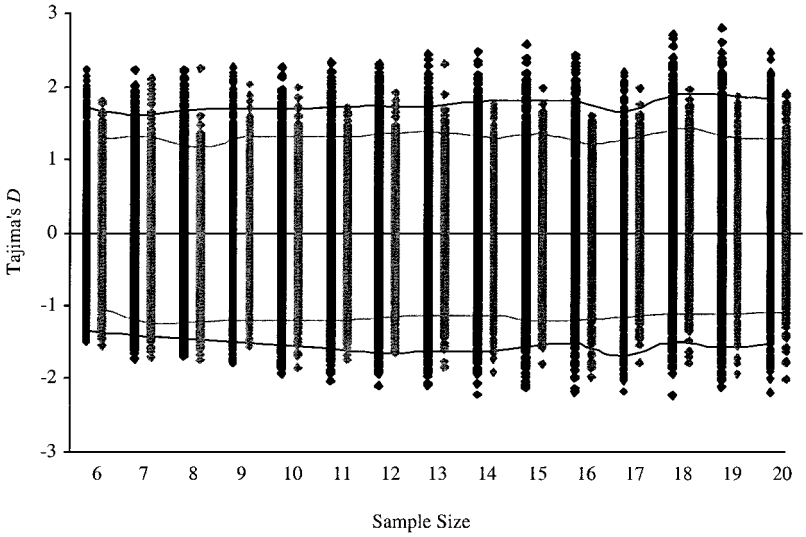

FIGURE 2.—Effect of recombination on null distributions of Tajima's *D* test statistic. Sequences were simulated under the standard neutral coalescent model either without recombination (solid) or with recombination (shaded). The curved lines indicate the 95% confidence intervals.

as the time of peak diversity as reported in SHANKARAPPA *et al.* (1999); Figure 1]. Using Fu and Li's test, the neutral expectation was rejected for 3 of 145 samples (also with excesses of low-frequency mutations; Figure 1, ‡, for the DNA sample and § for the RNA samples). Of these 3, the 2 samples from participant 6 occurred after genetic diversity stabilized. No departures from neutral expectations were detected by either test if recombination remained unacknowledged (data not shown). Accounting for recombination, therefore, revealed stronger evidence for nonneutral evolution in these sequences.

As a further approach for investigating purifying selection, all 1509 sequences were translated and a peptide alignment was constructed. Although in any individual no more than ~21% of nucleotide sites were observed to vary at any given time, only one amino acid residue was observed to be 100% conserved (data not shown). This site was G366, which plays a role in CD4 receptor binding (KWONG *et al.* 1998). Presumably some of the observed *env* alleles do not yield functional proteins; however, the lack of conservation was striking. The observation that every other site tolerated some amount of amino acid variation suggests that *env* evolution is only minimally constrained by functional requirements.

If HIV-1 populations within infected individuals were experiencing recurring selective sweeps, we would expect to detect fixation events occurring with high frequency. We therefore asked when the first fixation event occurred in each individual. The first fixation event was defined as the first time a sample had a derived nucleotide state reach a frequency of 100% when it started at 0% (sites that were initially polymorphic were disregarded because it was unknown when the initial mutation events occurred). It is the neutral expectation that the first fixation event should occur after an average of $2N_e$ generations, with variance on the order of N_e^2 (RODRIGO and FELSENSTEIN 1999), which is also the time when neutral diversity is expected to reach equilibrium. This event happened an average of 3.88 years after seroconversion (Figure 1, ↑), concurrent with the time when genetic diversity stabilized for four of the individuals studied. For the other five individuals, the first fixation event happened before genetic diversity stabilized, as early as 1.5 years after seroconversion. In none of the individuals did the first fixation event happen after genetic diversity stabilized. The observed mean time to the first fixation event of 3.88 years is not significantly less than the mean time to diversity stabilization of 5.06 years. The observation that fixation occurred at an earlier time in five of the nine individuals, but never later, suggests the presence of positive selection.

In summary, the C2–V5 region of HIV-1 *env* appears to evolve in a manner consistent with neutral expectations in 85% of the samples. However, evidence for episodes of nonneutral evolution was found in seven of the nine individuals. Accounting for the reduction of variance introduced by recombination led to stronger evidence for nonneutral evolution.

Estimating N_e: Having addressed the assumption of neutrality, we obtained estimates of N_e using three coalescent-likelihood programs. The values of N_e ranged from 311 to 4783 for COALESCE (Figure 1). RECOMBINE, which relaxes the assumption of no recombination, yielded estimates of N_e from 326 to 2886 (Figure 1). As expected, RECOMBINE yielded lower estimates than COALESCE. FLUCTUATE, which relaxes the assumption of a constant effective population size, is known to have an upward bias (KUHNER *et al.* 1998) and yielded the highest estimates (from 439 to 1,063,776; Figure 1). The overestimates from FLUCTUATE were also consistent with sequences analyzed in the presence of unacknowledged recombination (Figure 3). All of the estimates of N_e based on COALESCE and RECOMBINE, and 88% of the estimates of N_e based on FLUCTUATE,

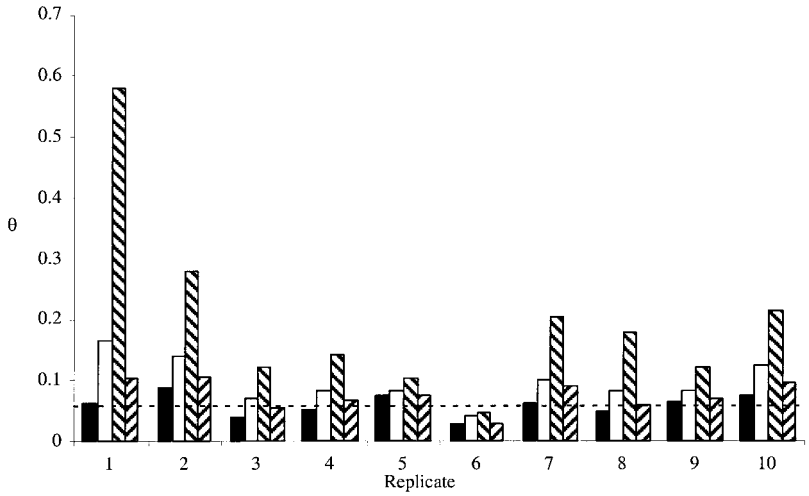

FIGURE 3.—Effects of estimating θ under varying sets of assumptions. Sequences were simulated under a neutral coalescent model with recombination. Solid bars represent Watterson's estimate of θ, open bars represent the estimate of θ from COALESCE, downward-hatched bars represent estimates of θ from FLUCTUATE, and upward-hatched bars represent estimates of θ from RECOMBINE. The horizontal dashed line indicates the true parameter value.

were below the inverse mutation rate, with average estimates of 1933 from COALESCE and 1195 from RECOMBINE, both between one and two orders of magnitude lower than the inverse mutation rate (Figure 1). The median from all estimates was 1810. This, together with the analyses of selection above, suggests that stochastic forces are important influences on these populations and are sufficient to explain much of the observed genetic variation.

Our estimates of N_e are consistent with those of LEIGH BROWN (1997) but not with those of ROUZINE and COFFIN (1999). Rouzine and Coffin compared the observed frequencies of rare haplotypes to frequencies obtained in simulation experiments of an idealized two-site, two-state system. However, their test considers only sites that are highly diverse, *i.e.*, an *a posteriori* selection of sites that are not randomly sampled from all sites. From their data and using their definition of highly diverse sites as those with a mutant frequency of 25–75%, there were 10 sites with high diversity, 15 sites with low diversity, and 209 constant sites ("constant" indicates sites not observed to vary in the sample). If pairs of sites were randomly chosen from all sites, only ~0.2% of pairs would contain 2 highly diverse sites. Thus, their selection of highly diverse sites may have introduced a bias toward higher estimates of diversity and, consequently, higher estimates of N_e.

To test if the effective mutation rate is the same for highly diverse sites as for all sites, we performed a likelihood-ratio test of equal rates across sites *vs.* gamma-distributed rate heterogeneity across sites (YANG 1994; HUELSENBECK and RANNALA 1997). We rejected equal rates across sites, because the test statistic of 44.2 was well outside a χ^2 distribution with 1 d.f. ($P = 3.0 \times 10^{-11}$; SWOFFORD 2002). Thus, the use of a single average per site mutation rate that describes the overall sequences does not adequately describe the few highly diverse sites.

On the basis of the foregoing, we propose that the appropriate mutation rate to use when relating the least frequent haplotype to the product $N_e\mu$ is the effective mutation rate at the two highly diverse sites in question rather than an average across all sites. From the gamma distribution of rate heterogeneity, we estimated the effective mutation rate for the class of highly diverse sites to be ~14, relative to a mean across all sites of 1. Thus, after accounting for Rouzine and Coffin's use of the per site mutation rate $\mu = 10^{-5}$ instead of 2.5×10^{-5}, the mutation rate they used is $14 \times 2.5 \approx 35$ times too low, yielding estimates of N_e that are therefore ~35 times too high. Our proposed corrections estimate N_e to be on the order of 10^3, alleviating the discrepancy between the estimate based on Rouzine and Coffin's method and estimates based on coalescent theory.

DISCUSSION

Although HIV-1 sequence heterogeneity is widely recognized, its biological origins and consequences are not understood. On the basis of a joint consideration of the results of four statistical tests of neutrality and estimates of the effective population size, we conclude that the genetic variation in the C2–V5 region of HIV-1 subtype B *env* observed in the nine individuals studied can be explained by a predominance of random genetic drift of neutral mutations with brief episodes of natural selection, which were frequently masked by recombination. Of 145 samples, 21 showed a departure from neutral expectations by one of four statistical tests of neutrality, and 1 sample showed departures from neutral expectations by two tests. The remaining 85% of the samples showed no departures from neutral expectations by any of the tests. We estimated N_e to be on the order of ~10^3, consistent with estimates of LEIGH BROWN (1997) and SEO *et al.* (2002). Reanalysis of ROUZINE and COFFIN's

(1999) work revealed a systematic bias that, when corrected, also yielded an estimate of N_e on the order of 10^3.

Simmonds *et al.* (1991) suggested that stably integrated proviral DNA may turn over more slowly than plasma viral RNA, thus providing a reason to analyze sequences from these two sources separately. In our previous study (Shankarappa *et al.* 1999), we found no differences in the development of diversity within time points or divergence from the founder strain between cell-associated viral DNA and cell-free viral RNA samples. Similarly, in this study, we found no differences in the patterns of neutrality or the estimates of N_e between viral DNA and RNA samples. This lack of difference may be explained by the fact that our DNA sampling protocol does not differentiate between more labile, intracellular but unintegrated, viral DNA and more stable, integrated, proviral DNA. Furthermore, all three of these genomic forms probably turn over substantially faster than the 6-month interval at which the samples were obtained (Ho *et al.* 1995; Wei *et al.* 1995; Perelson *et al.* 1996; Mittler *et al.* 1999; Ramratnam *et al.* 1999).

The four tests we used to detect departures from neutrality included pairwise comparisons of nonsynonymous and synonymous distances, Tajima's *D* test, Fu and Li's *D** test, and a test of recurrent mutation. For both Tajima's and Fu and Li's tests, a negative test statistic that achieves significance indicates an excess of low-frequency mutations. Three possible explanations for an excess of low-frequency mutations are directional selection, purifying selection, or population growth (Tajima 1989; Fu and Li 1993; Braverman *et al.* 1995). Directional selection is the most likely of these explanations, on the basis of the following observations from the present study: (1) the estimates of N_e are remarkably constant (rather than varying with census population sizes); (2) the test of nonsynonymous and synonymous distances predominantly revealed excesses of nonsynonymous mutations (the one exception was from participant 6, for which d_s was in excess over d_n, suggesting that the excess of low-frequency mutations was due to purifying selection); and (3) only one amino acid residue was 100% conserved. Similarly, for both Tajima's and Fu and Li's tests, a positive test statistic indicates an excess of intermediate-frequency mutations. Such a result may indicate recurrent mutation, in which case *S* is an underestimate of the true number of mutation events. However, the four samples that displayed an excess of recurrent mutation did not yield significance by either Tajima's or Fu and Li's tests.

Our analyses do not support a hypothesis of recurring selective sweeps affecting *env* during chronic infection. Nucleotide substitutions should happen more rapidly and more frequently when driven by positive selection than when driven by random genetic drift. Thus, we would have expected the first nucleotide substitution to occur much sooner than the average of 3.88 years after seroconversion, assuming that $2N_e$ generations is 5.06 years. Furthermore, the maintenance of a stable amount of diversity, as shown in Figure 1, implies that the strength of natural selection remains equally stable through chronic infection. It seems highly unlikely that the strength of natural selection should remain constant through time and across individuals.

The estimates of N_e derived from COALESCE and RECOMBINE were remarkably consistent within and among individuals, regardless of the presence or absence of a signal for selection. This suggests that the methods were relatively robust to violations of the assumption of no selection in these sequences. The higher estimates from FLUCTUATE were consistent with inappropriately attempting to explain recombination by a model of population growth. The rate heterogeneity we invoked to correct Rouzine and Coffin's estimate to 10^3 can be explained by selection and/or recombination (Yang 1994; Schierup and Hein 2000). We further suggest that if one had adequate data to estimate the least frequent haplotype at less diverse sites, then a more representative estimate would be: $\theta_{average} = f_{con}\theta_{con} + f_{low}\ \theta_{low} + f_{high}\theta_{high}$, in which f_{con}, f_{low}, and f_{high} are the frequencies of constant, low-diversity, and high-diversity sites, respectively, and θ_{con}, θ_{low}, and θ_{high} are the estimates for θ if one applied Rouzine and Coffin's test to haplotype data from these sites. $\theta_{average}$ could then be used in conjunction with the average per site mutation rate to estimate N_e.

Previously, Shankarappa *et al.* (1999) observed that, starting from nearly genetically homogeneous *env* sequences after seroconversion, intrasample diversity initially accumulated at a linear rate and after a certain time stabilized such that there was a zero slope in the curve of diversity as a function of time. There are three possible explanations for this stabilization of genetic diversity:

1. In the deterministic case, N_e is high, and selection is more important than random genetic drift. In this scenario, the stabilization of genetic diversity reflects an equilibrium in which mutant frequencies equal the mutation rate divided by the selection coefficient (Rouzine *et al.* 2001).
2. In the stochastic case, N_e is low, and random genetic drift is more important than selection. In this scenario, the stabilization of genetic diversity reflects the equilibrium between neutral mutations and random genetic drift that is expected to occur after an average of $2N_e$ generations (Rodrigo and Felsenstein 1999).
3. In the alternative stochastic case, N_e is low, but selection is more important than random genetic drift. In this scenario, recurrent selective sweeps (presumably due to immune pressure) reduce diversity and a true equilibrium is never reached. Under this explanation, the reduction in diversity is due to the hitchhik-

ing of neutral mutations linked to the mutation conferring the selective advantage (Maynard Smith and Haigh 1974; Kaplan *et al.* 1989).

Our findings of low N_e values, a predominance of random genetic drift over selection, and a lack of early fixation events favor explanation 2, reflecting a highly stochastic evolutionary process.

Tajima's (1989) *D* test and Fu and Li's (1993) *D** test are applicable only for equilibrium populations. Therefore, although we performed these tests on all of the 145 samples, the presence of significant results among the 70 samples derived from study visits before the time when genetic diversity stabilized may simply reflect preequilibrium conditions. In Figure 1, values of N_e for all samples are shown; however, as with the two tests, samples prior to the peak of viral diversity do not satisfy the assumption of an equilibrium population. Therefore, estimates of N_e derived from these samples are included to depict the expected accumulation of diversity prior to equilibrium starting from near homogeneity (Li 1977).

The basis for the five or more orders of magnitude difference between the census and the effective viral population size is not understood. One partial explanation is a low ratio of infectious units to virus particles (Layne *et al.* 1992; Dimitrov *et al.* 1993; Piatak *et al.* 1993). Other possibilities include high variance in the numbers of progeny virions per infected cell, overlapping (rather than discrete) generations, and periodic or short-lived reductions in population sizes (Kimura 1983). Both positive (directional) and negative selection can remove diversity from the population, thereby reducing N_e. Another hypothesis is that HIV replicates locally in subpopulations that undergo frequent extinction and recolonization (Grossman *et al.* 1998; Frost *et al.* 2001). If recolonization is associated with a small number of founders from other subpopulations, random genetic drift may strongly contribute to shaping the viral population (Slatkin 1977; Maruyama and Kimura 1980; Wakeley and Aliacar 2001; Pannell 2003). These explanations are not mutually exclusive, and all may contribute to reducing N_e to varying degrees.

We note that the course of sampling in the individuals we studied included periods of emergence and dominance in the population of viruses with predicted altered cell tropism, periods of low-impact antiviral therapy, clinical AIDS, and terminal disease (Shankarappa *et al.* 1999). However, samples derived from specimens collected at these potentially clinically or biologically meaningful times showed no consistent departures from neutrality. This, together with the facts that 85% of the samples did not show departures from neutral expectations by any of the four tests and that our data support low estimates of N_e, indicates that the evolution of HIV-1 *env* within these individuals was highly stochastic and that natural selection over the portion of *env* examined played only a minor role in driving diversity (Coffin 1995; Nowak *et al.* 1995, 1996).

We thank Yang Wang for helpful discussions and Allen G. Rodrigo and Gerald H. Learn, Jr. for critically reviewing this manuscript. We thank Dr. Learn for assisting with sequence management. We thank the two anonymous reviewers for their comments. D.S. was a Howard Hughes Medical Institute Predoctoral Fellow. This work was supported by grants from the U.S. Public Health Service and the University of Washington Center for AIDS Research.

LITERATURE CITED

Anisimova, M., R. Nielsen and Z. Yang, 2003 Effect of recombination on the accuracy of the likelihood method for detecting positive selection at amino acid sites. Genetics **164:** 1229–1236.

Beerli, P., and J. Felsenstein, 1999 Maximum-likelihood estimation of migration rates and effective population numbers in two populations using a coalescent approach. Genetics **152:** 763–773.

Borrow, P., H. Lewicki, X. Wei, M. S. Horwitz, N. Peffer *et al.*, 1997 Antiviral pressure exerted by HIV-1-specific cytotoxic T lymphocytes (CTLs) during primary infection demonstrated by rapid selection of CTL escape virus. Nat. Med. **3:** 205–211.

Braverman, J. M., R. R. Hudson, N. L. Kaplan, C. H. Langley and W. Stephan, 1995 The hitchhiking effect on the site frequency spectrum of DNA polymorphisms. Genetics **140:** 783–796.

Coffin, J. M., 1995 HIV population dynamics in vivo: implications for genetic variation, pathogenesis, and therapy. Science **267:** 483–489.

Dimitrov, D. S., R. L. Willey, H. Sato, L.-J. Chang, R. Blumenthal *et al.*, 1993 Quantitation of human immunodeficiency virus type 1 infection kinetics. J. Virol. **67:** 2182–2190.

Evans, D. T., D. H. O'Connor, P. Jing, J. L. Dzuris, J. Sidney *et al.*, 1999 Virus-specific cytotoxic T-lymphocyte responses select for amino-acid variation in simian immunodeficiency virus Env and Nef. Nat. Med. **5:** 1270–1276.

Frost, S. D., M. J. Dumaurier, S. Wain-Hobson and A. J. Brown, 2001 Genetic drift and within-host metapopulation dynamics of HIV-1 infection. Proc. Natl. Acad. Sci. USA **98:** 6975–6980.

Fu, Y.-X., 1996 New statistical tests of neutrality for DNA samples from a population. Genetics **143:** 557–570.

Fu, Y.-X., 1997 Statistical tests of neutrality of mutations against population growth, hitchhiking and background selection. Genetics **147:** 915–925.

Fu, Y.-X., and W.-H. Li, 1993 Statistical tests of neutrality of mutations. Genetics **133:** 693–709.

Gange, S. J., A. Muñoz, J. S. Chmiel, A. D. Donnenberg, L. M. Kirstein *et al.*, 1998 Identification of inflections in T-cell counts among HIV-1-infected individuals and relationship with progression to clinical AIDS. Proc. Natl. Acad. Sci. USA **95:** 10848–10853.

Grossman, Z., M. B. Feinberg and W. E. Paul, 1998 Multiple modes of cellular activation and virus transmission in HIV infection: a role for chronically and latently infected cells in sustaining viral replication. Proc. Natl. Acad. Sci. USA **95:** 6314–6319.

Hartl, D. L., and A. G. Clark, 1989 *Principles of Population Genetics*, Ed. 2. Sinauer Associates, Sunderland, MA.

Ho, D. D., A. U. Neumann, A. S. Perelson, W. Chen, J. M. Leonard *et al.*, 1995 Rapid turnover of plasma virions and CD4 lymphocytes in HIV-1 infection. Nature **373:** 123–126.

Huelsenbeck, J. P., and B. Rannala, 1997 Phylogenetic methods come of age: testing hypotheses in an evolutionary context. Science **276:** 227–232.

Kaplan, N. L., R. R. Hudson and C. H. Langley, 1989 The "hitchhiking effect" revisited. Genetics **123:** 887–899.

Kaslow, R. A., D. G. Ostrow, R. Detels, J. P. Phair, B. F. Polk *et al.*, 1987 The Multicenter AIDS Cohort Study: rationale, organization, and selected characteristics of the participants. Am. J. Epidemiol. **126:** 310–318.

Kimura, M., 1983 *The Neutral Theory of Molecular Evolution.* Cambridge University Press, Cambridge, UK/London/New York.

Kuhner, M. K., J. Yamato, P. Beerli and J. Felsenstein, 1995a LAMARC—likelihood analysis with metropolis algorithm using ran-

dom coalescence (http://evolution.genetics.washington.edu/lamarc.html).

Kuhner, M. K., J. Yamato and J. Felsenstein, 1995b Estimating effective population size and mutation rate from sequence data using Metropolis-Hastings sampling. Genetics **140:** 1421–1430.

Kuhner, M. K., J. Yamato and J. Felsenstein, 1998 Maximum likelihood estimation of population growth rates based on the coalescent. Genetics **149:** 429–434.

Kuhner, M. K., J. Yamato and J. Felsenstein, 2000 Maximum likelihood estimation of recombination rates from population data. Genetics **156:** 1393–1401.

Kumar, S., K. Tamura, I. B. Jakobsen and M. Nei, 2001 MEGA2: molecular evolutionary genetics analysis software. Bioinformatics **17:** 1244–1245.

Kwong, P. D., R. Wyatt, J. Robinson, R. W. Sweet, J. Sodroski *et al.*, 1998 Structure of an HIV gp120 envelope glycoprotein in complex with the CD4 receptor and a neutralizing human antibody. Nature **393:** 648–659.

Layne, S. P., M. J. Merges, M. Dembo, J. L. Spouge, S. R. Conley *et al.*, 1992 Factors underlying spontaneous inactivation and susceptibility to neutralization of human immunodeficiency virus. Virology **189:** 695–714.

Leigh Brown, A. J., 1997 Analysis of HIV-1 env gene sequences reveals evidence for a low effective number in the viral population. Proc. Natl. Acad. Sci. USA **94:** 1862–1865.

Li, C. C., 1976 *First Course in Population Genetics*, pp. 567–592. Boxwood Press, Pacific Grove, CA.

Li, W.-H., 1977 Distribution of nucleotide differences between two randomly chosen cistrons in a finite population. Genetics **85:** 331–337.

Mansky, L. M., 1996 Forward mutation rate of human immunodeficiency virus type 1 in a T lymphoid cell line. AIDS Res. Hum. Retroviruses **12:** 307–314.

Margolick, J. B., A. Muñoz, A. D. Donnenberg, L. P. Park, N. Galai *et al.*, 1995 Failure of T-cell homeostasis preceding AIDS in HIV-1 infection. Nat. Med. **1:** 674–680.

Maruyama, T., and M. Kimura, 1980 Genetic variability and effective population size when local extinction and recolonization of subpopulations are frequent. Proc. Natl. Acad. Sci. USA **77:** 6710–6714.

Maynard Smith, J., and J. Haigh, 1974 The hitch-hiking effect of a favourable gene. Genet. Res. **23:** 23–35.

Mittler, J. E., M. Markowitz, D. D. Ho and A. S. Perelson, 1999 Improved estimates for HIV-1 clearance rate and intracellular delay. AIDS **13:** 1415–1417.

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

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

Nei, M., and S. Kumar, 2000 *Molecular Evolution and Phylogenetics*. Oxford University Press, New York.

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.

Nowak, M. A., R. M. May, R. E. Phillips, S. Rowland-Jones, D. G. Lalloo *et al.*, 1995 Antigenic oscillations and shifting immunodominance in HIV-1 infections. Nature **375:** 606–611.

Nowak, M. A., R. M. Anderson, M. C. Boerlijst, S. Bonhoeffer, R. M. May *et al.*, 1996 HIV-1 evolution and disease progression. Science **274:** 1008–1011.

Overbaugh, J., and C. R. Bangham, 2001 Selection forces and constraints on retroviral sequence variation. Science **292:** 1106–1109.

Pannell, J. R., 2003 Coalescence in a metapopulation with recurrent local extinction and recolonization. Evolution **57:** 949–961.

Perelson, A. S., A. U. Neumann, M. Markowitz, J. M. Leonard and D. D. Ho, 1996 HIV-1 dynamics in vivo: virion clearance rate, infected cell life-span, and viral generation time. Science **271:** 1582–1586.

Piatak, M., Jr., M. S. Saag, L. C. Yang, S. J. Clark, J. C. Kappes *et al.*, 1993 High levels of HIV-1 in plasma during all stages of infection determined by competitive PCR. Science **259:** 1749–1754.

Ramratnam, B., S. Bonhoeffer, J. Binley, A. Hurley, L. Zhang *et al.*, 1999 Rapid production and clearance of HIV-1 and hepatitis C virus assessed by large volume plasma apheresis. Lancet **354:** 1782–1785.

Rice, W. R., 1989 Analyzing tables of statistical tests. Evolution **43:** 223–225.

Rinaldo, C. R., Jr., P. Gupta, X. Huang, Z. Fan, J. I. Mullins *et al.*, 1998 Anti-HIV-1 memory cytotoxic T lymphocyte responses associated with changes in CD4+ T cell numbers in the progression of HIV-1 infection. AIDS Res. Hum. Retroviruses **14:** 1423–1433.

Rodrigo, A. G., and J. Felsenstein, 1999 Coalescent approaches to HIV population genetics, pp. 233–272 in *The Evolution of HIV*, edited by K. A. Crandall. Johns Hopkins University Press, Baltimore.

Rouzine, I. M., and J. M. Coffin, 1999 Linkage disequilibrium test implies a large effective population number for HIV in vivo. Proc. Natl. Acad. Sci. USA **96:** 10758–10763.

Rouzine, I. M., A. Rodrigo and J. M. Coffin, 2001 Transition between stochastic evolution and deterministic evolution in the presence of selection: general theory and application to virology. Microbiol. Mol. Biol. Rev. **65:** 151–185.

Rozas, J., and R. Rozas, 1999 DnaSP version 3: an integrated program for molecular population genetics and molecular evolution analysis. Bioinformatics **15:** 174–175.

Schierup, M. H., and J. Hein, 2000 Consequences of recombination on traditional phylogenetic analysis. Genetics **156:** 879–891.

Seibert, S. A., C. Y. Howell, M. K. Hughes and A. L. Hughes, 1995 Natural selection on the *gag*, *pol*, and *env* genes of human immunodeficiency virus 1 (HIV-1). Mol. Biol. Evol. **12:** 803–813.

Seo, T.-K., J. L. Thorne, M. Hasegawa and H. Kishino, 2002 Estimation of effective population size of HIV-1 within a host: a pseudomaximum-likelihood approach. Genetics **160:** 1283–1293.

Shankarappa, R., J. B. Margolick, S. J. Gange, A. G. Rodrigo, D. Upchurch *et al.*, 1999 Consistent viral evolutionary changes associated with the progression of human immunodeficiency virus type 1 infection. J. Virol. **73:** 10489–10502.

Shriner, D., D. C. Nickle, M. A. Jensen and J. I. Mullins, 2003 Potential impact of recombination on sitewise approaches for detecting positive natural selection. Genet. Res. **81:** 115–121.

Simmonds, P. P., L. Q. Zhang, F. McOmish, P. Balfe, C. A. Ludlam *et al.*, 1991 Discontinuous sequence change of human immunodeficiency virus (HIV) type 1 env sequences in plasma viral and lymphocyte-associated proviral populations in vivo: implications for models of HIV pathogenesis. J. Virol. **65:** 6266–6276.

Slatkin, M., 1977 Gene flow and genetic drift in a species subject to frequent local extinctions. Theor. Popul. Biol. **12:** 253–262.

Soudeyns, H., S. Paolucci, C. Chappey, M. B. Daucher, C. Graziosi *et al.*, 1999 Selective pressure exerted by immunodominant HIV-1-specific cytotoxic T lymphocyte responses during primary infection drives genetic variation restricted to the cognate epitope. Eur. J. Immunol. **29:** 3629–3635.

Suzuki, Y., and T. Gojobori, 1999 A method for detecting positive selection at single amino acid sites. Mol. Biol. Evol. **16:** 1315–1328.

Swofford, D. L., 2002 *PAUP**: Phylogenetic Analysis Using Parsimony (**and Other Methods)*, Version 4.0. Sinauer Associates, Sunderland, MA.

Tajima, F., 1983 Evolutionary relationship of DNA sequences in finite populations. Genetics **105:** 437–460.

Tajima, F., 1989 Statistical method for testing the neutral mutation hypothesis by DNA polymorphism. Genetics **123:** 585–595.

Viscidi, R. P., 1999 HIV evolution and disease progression via longitudinal studies, pp. 346–389 in *The Evolution of HIV*, edited by K. A. Crandall. Johns Hopkins University Press, Baltimore.

Wakeley, J., and N. Aliacar, 2001 Gene genealogies in a metapopulation. Genetics **159:** 893–905.

Wall, J. D., 1999 Recombination and the power of statistical tests of neutrality. Genet. Res. **74:** 65–79.

Watterson, G. A., 1975 On the number of segregating sites in genetical models without recombination. Theor. Popul. Biol. **7:** 256–276.

Wei, X., S. K. Ghosh, M. E. Taylor, V. A. Johnson, E. A. Emini *et al.*, 1995 Viral dynamics in human immunodeficiency virus type 1 infection. Nature **373:** 117–122.

Wilson, C. C., R. C. Brown, B. T. Korber, B. M. Wilkes, D. J. Ruhl *et al.*, 1999 Frequent detection of escape from cytotoxic T-lymphocyte recognition in perinatal human immunodeficiency

virus (HIV) type 1 transmission: the ARIEL project for the prevention of transmission of HIV from mother to infant. J. Virol. **73:** 3975–3985.

Wright, S., 1931 Evolution in Mendelian populations. Genetics **16:** 97–159.

Yamaguchi-Kabata, Y., and T. Gojobori, 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.

Yang, Z., 1994 Maximum likelihood phylogenetic estimation from DNA sequences with variable rates over sites: approximate methods. J. Mol. Evol. **39:** 306–314.

Yang, Z., R. Nielsen, N. Goldman and A. M. Pedersen, 2000 Codon-substitution models for heterogeneous selection pressure at amino acid sites. Genetics **155:** 431–449.

Zhuang, J., A. E. Jetzt, G. Sun, H. Yu, G. Klarmann *et al.*, 2002 Human immunodeficiency virus type 1 recombination: rate, fidelity, and putative hot spots. J. Virol. **76:** 11273–11282.

Communicating editor: D. Charlesworth