A Robust Measure of HIV-1 Population Turnover Within Chronically Infected Individuals

G. Achaz, S. Palmer,† M. Kearney,† F. Maldarelli,† J. W. Mellors,‡ J. M. Coffin,† and J. Wakeley**

*Department of Organismic and Evolutionary Biology, Harvard University; †HIV Drug Resistance Program, NCI, NIH, Frederick, Maryland; and ‡Department of Infectious Diseases, University of Pittsburgh

A simple nonparameteric test for population structure was applied to temporally spaced samples of HIV-1 sequences from the *gag-pol* region within two chronically infected individuals. The results show that temporal structure can be detected for samples separated by about 22 months or more. The performance of the method, which was originally proposed to detect geographic structure, was tested for temporally spaced samples using neutral coalescent simulations. Simulations showed that the method is robust to variation in samples sizes and mutation rates, to the presence/absence of recombination, and that the power to detect temporal structure is high. By comparing levels of temporal structure in simulations to the levels observed in real data, we estimate the effective intra-individual population size of HIV-1 to be between 10^3 and 10^4 viruses, which is in agreement with some previous estimates. Using this estimate and a simple measure of sequence diversity, we estimate an effective neutral mutation rate of about 5×10^{-6} per site per generation in the *gag-pol* region. The definition and interpretation of estimates of such "effective" population parameters are discussed.

Introduction

There are at least two levels at which studies of HIV population genetics can be undertaken. The first is at a global level and considers dynamics of the virus in the whole population of infected individuals (Grassly, Harvey, and Holmes 1999). Even more broadly, this might include the whole immunodeficiency virus family (Mindell 1996). The second level, at a smaller scale, focuses on viral populations within an infected individual. The latter represents the intra-host, or intra-individual, population level and is the focus of the present study.

During HIV infection, changes in viral population size are typically characterized by three phases (Coffin 1999). For several weeks after the infection, the first phase is marked by an extensive increase in viral load, associated with a decrease in the CD4+ cells. That phase ends, in chronically infected individuals, when the immune system reacts, leading to a decrease in the viral load of up to two orders of magnitude (Daar et al. 1991). The third and last phase is characterized by a slow increase of the viral load. This phase usually ends when the virus infection overwhelms the individual's immune system, causing immunodeficiency, illness, and death. In some individuals, namely the long-term nonprogressors, this third phase is extended and the viral load can remain relatively low for decades.

In the present study, we analyzed ~1100 base pairs (bp) of the *gag-pol* region of HIV-1 sampled at different time points in two chronically infected patients. We propose a standard measure for analyzing the temporal structure in HIV-1 populations, which is based on a test for geographic population structure that was originally proposed by Hudson, Boos, and Kaplan (1992). The test compares the mean number of pairwise differences between sequences within each population to a theoretical distribution obtained by randomly shuffling the sequence labels. We adapt this straightforward test to the case of temporal structure between two samples of viral sequences taken from the same individual at two different times. We also use the test statistic as a measure of the amount of population turnover.

The evolution of a virus within a host has been shown to be strongly influenced by its environment. Some individuals are overwhelmed by the infection within a few years and others are able to resist disease progression for long periods of time. Two well-known examples of such intra-host environmental constraints are the genotype of the host, e.g., at histocompatibility and coreceptor loci, which can induce selective changes in viral genotype (Moore et al. 2002), and the application of therapeutic drugs, which leads to the emergence of predictable well-characterized drug resistant strains (Shankarappa 1999). Such observations underscore the importance of selection on intra-host evolutionary processes. Because the total population size of HIV in the third phase has been estimated to be very large, on the order of 10^7 to 10^8 infected cells and about 10^{10} individual viruses (Piatak et al. 1993; Haase et al. 1996), it may be appropriate to consider the dynamics of intra-host HIV genetic variation to be deterministic (Coffin 1995), as if the population size were infinite.

It has also been reported that, even if drugs do induce predictable mutations conferring a resistant phenotype, the frequency and timing of the fixation of these mutations is highly variable from one individual to another (Leigh Brown and Richman 1997). This observation has been interpreted as a consequence of random variation of the frequencies of resistant strains preexisting before the start of drug therapy, together with deterministic selection. Other possible explanations for differences among individuals include host-virus genotype interactions that affect either the fitness of resistance mutations or the mutation rate of the virus, and variation in the time for the mutation that confers resistance to appear. In any case, variation in

Key words: intra-host HIV evolution, effective population size, chronically infected individual.

E-mail: gachaz@oeb.harvard.edu.

Mol. Biol. Evol. 21(10):1902–1912. 2004
doi:10.1093/molbev/msh196
Advance Access publication June 23, 2004

the timing of events among individuals would seem to imply an important role for stochasticity, or random genetic drift, during infection.

To reconcile a very large population size with a nonnegligible effect of random genetic drift, it has been proposed that the effective population size (N_e) of viruses inside each individual is several orders of magnitude smaller than the real population size. The effective size is defined as the corresponding size of a hypothetical neutral idealized population (i.e., described by the standard Wright-Fisher model) that gives the same amount of genetic drift observed in the real population. Many factors are known to make the effective size very different from the real size, including selection, population structure, and fluctuating population size. All methods of estimating N_e assume that observed genetic variation is neutral. Indeed, the very concept of an effective population size is based on this notion. In this respect, it is interesting to note that, in a population of infinite size, a locus evolving under directional selection can drive the turnover at a partially linked neutral locus, mimicking genetic drift (Gillespie 2000). The neutral locus changes by a process similar to hitchhiking (Maynard Smith and Haigh 1974; Kaplan, Hudson, and Langley 1989). This "pseudohitchhiking" model (Gillespie 2000) shows that a reduced effective size at a genetic locus can be caused by selection even in an infinite population.

In studies of HIV-1, several estimates of the intra-host N_e have been obtained by analyzing polymorphisms in the *env* region (Leigh Brown 1997; Rodrigo et al. 1999; Shriner et al. 2004). These studies support a relatively small effective population size, on the order of 10^3 to 10^4. An alternative method to estimate N_e, using linkage disequilibrium between polymorphic sites, suggested that N_e is about 10^6 (Rouzine and Coffin 1999). However, a recent reanalysis of the same data suggests that this higher value was due to a bias in the analyzed polymorphisms (Shriner et al. 2004) and that, after correction, this value is on the order of 10^3. All of these estimates of N_e are much smaller than the actual population size of the virus within an infected person, which again may be 10^{10}, and imply an important role for genetic drift in the dynamics of genetic variation. A recent study that modeled resistance to Lamivudine (3TC) argued that even with N_e on the order of 10^6, random drift may still play an important role in *env* region (Frost et al. 2000). Gillespie's (2000) pseudohitchhiking model may help to reconcile these results, since strong selective effects have been observed in the *env* region (Nielsen and Yang 1998; Richman et al. 2003).

We show that temporally spaced samples (often referred to as serial samples) within two chronically infected individuals can be distinguished using the test mentioned above. In addition, we determine the power and size of the test using standard neutral coalescent simulations. A number of previous methods, reviewed in Drummond et al. (2003), have been developed to estimate the mutation rate and the population size from serial samples (Drummond and Rodrigo 2000; Rambaut 2000; Drummond, Forsberg, and Rodrigo 2001; Drummond et al. 2002). All these methods assume that there is no recombination, and they rely on the existence of a single simple coalescent history or genealogy for all sites in the locus. It is not known how such methods will perform when there is recombination. In contrast, the method we propose here does not rely on a common genealogy for all sites, and simulations show that it performs similarly well whether recombination is rampant or completely absent.

Using the neutral coalescent simulations of serial samples, we describe the rate of change of the test statistic in an evolving population both with and without recombination. This allows the estimation of the effective population size in one of the two patients by a comparison between expected and observed rates of change in the test statistic. Using a parametric bootstrap, i.e., by repeatedly simulating samples and applying the method to them, we can give confidence intervals on these estimates. We show that the intervals for no recombination and for free recombination are closely overlapping. We also calculate an effective mutation rate, which reflects the neutral mutation rate of these sequences.

Material and Methods

Origin and Analysis of the Sequences

Plasma samples were obtained at different times from two untreated individuals with well-established HIV-1 infection. DNA sequences from about 20–50 individual viral genomes were obtained for each sample using single genome RT-PCR sequence (SGS) analysis of approximately 1,098 bp, including the p6 region of gag, protease, and the first 900 nucleotides of RT (see more details on the method in S. Palmer et al. [in preparation]). Summary statistics of the sequences we used are described in table 1.

For each sample, sequences were aligned together by using ClustalW (Thompson, Higgins, and Gibson 1994) with default parameters. All alignments were visually inspected and frameshifts were removed using the sequence editor SEAVIEW (Galtier, Gouy, and Gautier 1996). The gap character was considered a fifth symbol in calculating pairwise differences between the sequences.

Testing for Population Subdivision

We implemented the series of tests for population subdivision described by Hudson, Boos, and Kaplan (1992). The tests were originally proposed to detect associations between genetic structure and geographic structure. However, the design of the tests, in which a matrix of pairwise sequence differences is calculated from the data then randomly permuted to assess the significance of structure, is quite general and nonparametric, so it is easily extended to other situations. Hudson, Boos, and Kaplan (1992) investigated two measures of subdivision, called K_s and K_s^*, defined below, and showed in simulations that the test using K_s^* had more power to detect geographic structure. Let n_1 be the number of sequences in the first sample and n_2 be the number of sequences in the second sample.

K_i is the mean number of differences between pairs of sequences in sample i. K_s is defined as $K_s = w_1K_1 + w_2K_2$, where $w_1 = n_1/(n_1 + n_2)$ and $w_2 = 1 - w_1$.

Table 1
Characteristics of Sequences from Patients A and B

		Date	Days	Number of Sequences	Viral Load (RNA/ml)	K_i (estimated Θ)
Patient A (1,098 sites)	First positive test	1991	—	—	—	—
	Sample 1	11/19/1998	0	16	27,252	8.88
	Sample 2	12/15/1998	26	22	18,604	5.78
	Sample 3	04/20/1999	142	7	24,648	9.52
	Sample 4	08/26/1999	280	17	22,164	8.78
	Sample 5	03/01/2000	468	13	136,760	10.95
	Sample 6	05/26/2000	554	42	10,520	11.31
	Sample 7	06/22/2000	561	15	11,934	9.31
	Sample 8	07/11/2000	600	10	19,285	8.19
	Sample 9	07/12/2000	601	7	15,362	9.95
	Sample 10	07/13/2000	602	7	16,018	12.95
	Sample 11	07/14/2000	603	16	16,446	11.09
	Sample 12	07/15/2000	604	18	16,419	9.88
	Sample 13	07/16/2000	605	16	16,904	9.89
	Sample 14	07/17/2000	606	13	20,392	12.09
	Sample 15	07/18/2000	607	9	18,918	12.00
	Sample 16	07/19/2000	608	14	24,855	11.09
	Sample 17	07/20/2000	609	17	21,600	11.16
	Sample 18	12/27/2001	1,134	21	21,760	11.83
	Sample 19	11/19/2002	1,461	53	30,111	11.00
Patient B (1,313 sites)	First positive test	July 2000		—	—	—
	Sample 1	07/23/2001	0	53	19,783	17.38
	Sample 2	01/07/2002	168	30	3,996	17.26
	Sample 3	01/07/2002	357	5	3,263	16.60

K_i^* is defined as $K_i^* = \sum_{a=1}^{n_i-1} \sum_{b=a+1}^{n_i} \log(1 + D_{ab})/\binom{n_i}{2}$, where D_{ab} is the number of differences between sequence a and sequence b of sample i, and $K_s^* = w_1K_1^* + w_2K_2^*$. Hudson, Boos, and Kaplan (1992) suggest that optimal weights for K_s^* are $w_1 = (n_1 - 2)/(n_1 + n_2 - 4)$ and $w_2 = 1 - w_1$.

This test generates a P value for the probability that the level of structure between two samples of sequences is due simply to chance. To do this, the sequences are randomly relabeled ("population 1" and "population 2") a large number of times, holding n_1 and n_2 constant, and the statistics are computed for each such permutation. Except where specified below, we used 10,000 relabelings/permutations to obtain P values. The P value of the observed statistic is equal to the fraction of times the value for the permuted data is less than or equal to the observed value. This procedure detects patterns of genetic structure in which pairwise differences within samples tend to be smaller than pairwise differences between samples. If the P value is less than the nominal level of significance, which we denote α, the null hypothesis of no structure is rejected.

Coalescent Simulations

We simulated samples of sequences to estimate the size of the test (i.e., the validity of the significance level α) and the power of the test to detect temporal structure as a function of the time between samples. We also used simulations to investigate how the average P value changes with the time between samples. Simulations followed the standard coalescent methods (see, e.g., Hudson 1990), in which a genealogy is constructed and then a Poisson-distributed number of neutral mutations is randomly placed on the genealogy. We assumed that each mutation gave rise to a unique polymorphic site (the infinite-sites mutation model). For each mutation, a branch is chosen randomly in proportion to its length and every descendent of that branch inherits the mutation. We allowed two different possibilities for recombination in this infinite-sites mutation model, either (1) no recombination occurred (Waterson 1975), or (2) recombination occurred freely between all pairs of sites (Kimura 1969). The results below are based on 10,000 simulation replicates for each set of parameters.

To build a genealogy, we first chose a sample size for the first and the second time points, respectively. We then used a standard neutral coalescence process (Kingman 1982*a*, 1982*b*; Tajima 1983), which depends on the neutral mutation parameter $\Theta = 2N_e\mu$ (where μ is the neutral mutation rate per sequence per generation) and on both sample sizes. We simulate the history of the second sample back to the time when the first sample was taken. The number of coalescent events during this part of the history depends on the time t_{2-1} (in number of generations) between the two samples. As usual in the coalescent, this time is rescaled so that the unit of measurement is N generations: $T_{2-1} = (t_{2-1})/N$ (where N is the population size). The expected number of mutations along a single lineage over this time period is equal to $\mu t_{2-1} = T_{2-1} \times \Theta/2$. When the simulation reaches time point 1, the sequences from sample 1 are added to the ancestral lineage(s) remaining from sample 2. The coalescent process continues until the most recent common ancestor of both samples is reached. A realization of such genealogy is shown in figure 1. This approach is identical to the way in which the coalescent process has previously been applied to HIV evolution (Rodrigo and Felsenstein 1999; Rodrigo et al. 1999). In the same manner, it is straightforward to include samples from more than two time points.

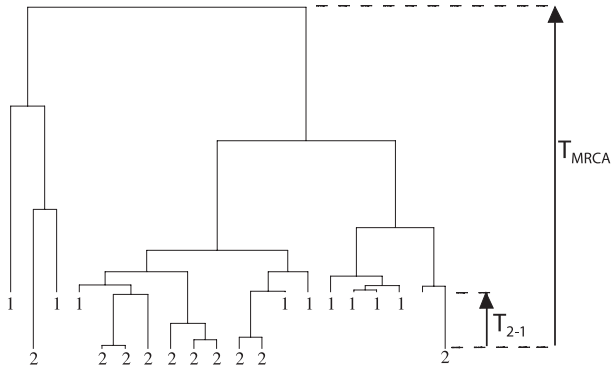

FIG. 1.—Coalescence of two time series samples. An example of a simulated neutral standard coalescent genealogy of two samples from the same population but separated by a defined time interval. In this case, we used $n_1 = 10$ sequences in the first sample (1) and $n_2 = 10$ in the second sample (2). We also used a value of $\Theta(2N\mu)$ of 10 and choose a time interval (T_{2-1}) of 0.4 (in number of N generations); one should note that the expected time to the most recent common ancestor is less than 2 and the expected time for two lineages (like the last two ones) to coalesce is 1.

Estimation of the Effective Population Size with Confidence Intervals

By matching the results of simulations with the results for the data, it is possible to estimate the effective population size N_e of HIV within a patient. There are a variety of ways this might be done. Here, we first estimate of the number of HIV generations between a pair of data samples such that there is a 50% chance of rejecting the null hypothesis at the $\alpha = 0.05$ level. Then, we equate this to the scaled time of separation in simulations that gives 50% power to reject the null hypothesis at the same $\alpha = 0.05$ significance level, and we solve for N_e. The value of 50% power was chosen based on preliminary simulations so we could use linear interpolation between different times of separation (on a log scale) without serious error. In contrast to the simulations in which the time of separation can be controlled, the sampling times for the data samples are fixed (table 1). Therefore, to estimate the 50%-power separation time for the data, we ordered the 171 sample pairs by time of separation then used a sliding window of 20 paired sample points to search (with the aid of interpolation) for the separation time that gave 10/20 rejections of the null hypothesis. We used the mean time of separation among the 20 points in the window as the estimate of the separation time.

We used a parametric bootstrap procedure to obtain confidence intervals for our estimate of N_e. Specifically, we assumed that the true values of N_e and Θ were those we estimated from the data and then simulated 10^4 genealogies of the 19 samples with separation times (table 1) rescaled by N_e, assuming either no recombination or free recombination. For each set of simulated sequences, we performed the test on all pairwise comparisons between samples and estimated N_e exactly as we did for the actual data. The upper and lower 2.5% cutoffs for these simulated distributions of estimates of N_e are taken as the 95% confidence interval.

Results

Because we were interested in finding a useful standard measure to compare population change through time within an individual, we investigated different sample sizes and times of separation of the serial samples in simulations. The coalescent process we used to create pseudosamples of sequences depended on four parameters: the sample sizes n_1 and n_2, the population mutation rate Θ, and the scaled time interval T_{2-1} between two samples. In most of what follows, we discuss results where Θ is equal to 10. This value is very close to the one we estimated for the data from patient A (see table 1) using average pairwise differences. Smaller and larger values for Θ, specifically $\Theta = 1$ and $\Theta = 100$, gave essentially the same results for all analyses and will be discussed later. We applied the test of Hudson, Boos, and Kaplan (1992) to each set of pseudosequences from the simulations to assess whether temporal structure could be detected. We examined the performance of both K_s and K_s^*.

Size of the Test

By setting T_{2-1} to 0, we create two sets of sequences sampled from a single time point. This case represents the null model of no temporal structure and can be used to measure the size of the test (i.e., frequency of false positive outcomes). To do this, we counted the number of times we would reject the null model, at the 5% significance level, for each set of parameter values. This addresses the concern that arises from the fact that an unknown genealogical structure exists and shapes genetic variation in the sample and that small or unbalanced samples ($n_1 < n_2$ or $n_1 > n_2$) might lead spurious rejections of the null hypothesis. Thus, we investigated different values of n_1 and n_2.

Results show that K_s^* is largely insensitive to sample size and to asymmetry of sample sizes from the two time points (fig. 2). The performance of K_s does depend on sample size when n_1 is small and n_2 is large. However, the direction of deviation is conservative (lower than expected chance of rejecting the null hypothesis when it is true). This result shows that these two measures exhibit the expected fraction (or fewer) of false positives even with small samples sizes. The test based on K_s^* was recommended by Hudson, Boos, and Kaplan (1992) because it was more sensitive for detecting geographical structure. In our simulations we found the same tendency and thus present only results from tests using K_s^*. However, using K_s results in only a subtle decrease of the power of the test.

Power to Detect Temporal Structure in a Neutrally Evolving Population

If we set the time interval between the two samples to a nonzero value, we simulate a sampling process at two different time points. To assess the power of the test, we chose a range of times. Again, as in a standard coalescent process, the time of separation T_{2-1} is scaled by the population size $T_{2-1} = (t_{2-1})/N$, where t_{2-1} is the number of generations between the second and the first sample. To assess the effect of recombination we created two series of artificial sequences. In the first series (no recombination; fig. 3*a*), all sites are so tightly linked that they always share the same genealogy. In the second one (free recombination; fig. 3*b*), all sites are segregating independently so that

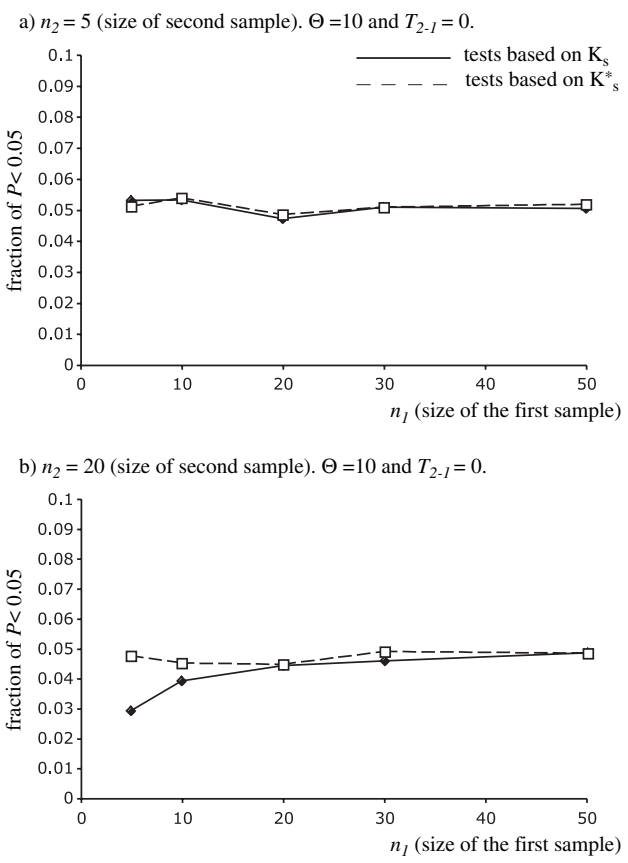

FIG. 2.—Size of the test using four different measures. The frequencies at which the null hypothesis is rejected (at an alpha risk of 5%) are plotted as a function of the first sample size. Since the time between the early and the late sample is $T_{2-1} = 0$, one would expect not to observe more than 5% of false positives. The coalescent parameters used here are $\Theta = 10$. In (*a*), the size of the late sample n_2 is set to 5 and the size of the early sample n_1 varies from 5 to 50. In (*b*), n_2 is set to 20 and n_1 varies from 5 to 50.

each site has its own genealogy. We fixed the sample size from each time point to be equal to 20.

The results show that, for $\Theta = 10$, the null hypothesis that the two samples come from the same time point is rejected at greater than 5% frequency if the scaled time of separation between the two samples is larger than ~0.01. They also show that after one scaled time unit (i.e., at $T_{2-1} = 1$ or greater), the samples are essentially always distinguishable from each other. This reflects the fact that many coalescent events will have occurred between the members of the later sample over this amount of time. In fact, from equations 6.1 and 6.2 in Tavaré (1984), the probability that there are more than three ancestral lineages of the later sample remaining at the time ($T_{2-1} = 1$) of the earlier sample is less than 0.05. Comparison of figure 3*a* to figure 3*b* shows that recombination increases the power of the test slightly, although mostly just in the vicinity of $T_{2-1} = 0.1$.

Analyses of other values of Θ show that for smaller values (i.e., $\Theta = 1$) the power of the test decreases and the effect of recombination almost disappears (fig. 3). In contrast, for higher values (i.e., $\Theta = 100$), the power of the test without recombination does not change but the effect of recombination is stronger (it increases the power of the test by an order of magnitude in T_{2-1}).

Application to HIV-1 *gag-pol* Sequences

These results show that the Hudson, Boos, and Kaplan (1992) test of subdivision provides a standard measure of the population structure through time, which can be used to tackle biological questions concerning the timing of population turnover. This can be done either with or without recombination in the sequences. We can now use this test to analyze the extent of intra-host HIV-1 population evolution. To do so, we used sequences sampled from two chronically infected individuals, here called A and B for reference, picked at different time points long after the primary infection (see table 1) and spaced by different time intervals.

As a visual test for structure, we reconstructed trees relating the samples, which are genealogies under the assumption that no recombination had occurred. This was done using the neighbor-joining method (Saitou and Nei 1987) with a Kimura two-parameter distance correction (Kimura 1980). An example tree of samples from two relatively distant time points from individual A is shown in figure 4. Although the tree does appear to show some structure, the significance of this structure is difficult to assess, both because it is just a visual comparison and because the assumption of no recombination is likely to be wrong. Recombination invalidates the usual interpretation that the branches of the tree represent ancestral lineages. Using the test based on K^*_s (or the one based on K_s) leads to rejection of the null hypothesis of no structure for these samples ($P < 0.003$). It is possible that the very shape of the tree implies recombination, because under complete linkage the expected tree should have long internal branches and short external ones (i.e., see fig. 1). As figure 3 shows, the presence of recombination only increases the power of the test, so the null hypothesis is rejected in either case for these samples.

We analyzed samples from 19 time points in individual A and the three time points in individual B. We made all 171 possible pairwise comparisons between samples from different time points in A and between the single pair of time points in B. The results, which are shown in figure 5, indicate that the test systematically rejects the null hypothesis (no population structure over time) after about 666 days, or 22 months. Compared to the very rapid adaptation that can sometimes be observed (e.g., in response to drug therapy; Shankarappa 1999), turnover of the HIV-1 population in these chronically infected individuals appears to be relatively slow, taking more than one year to be detectable at the 5% level using this test.

Estimation of the Effective Population Size

It is difficult to draw conclusions from the few comparisons we have for individual B. However, assuming neutrality of the observed mutations, it is possible to roughly estimate the effective population size of HIV-1 for individual A. The effective size N_e is defined as the

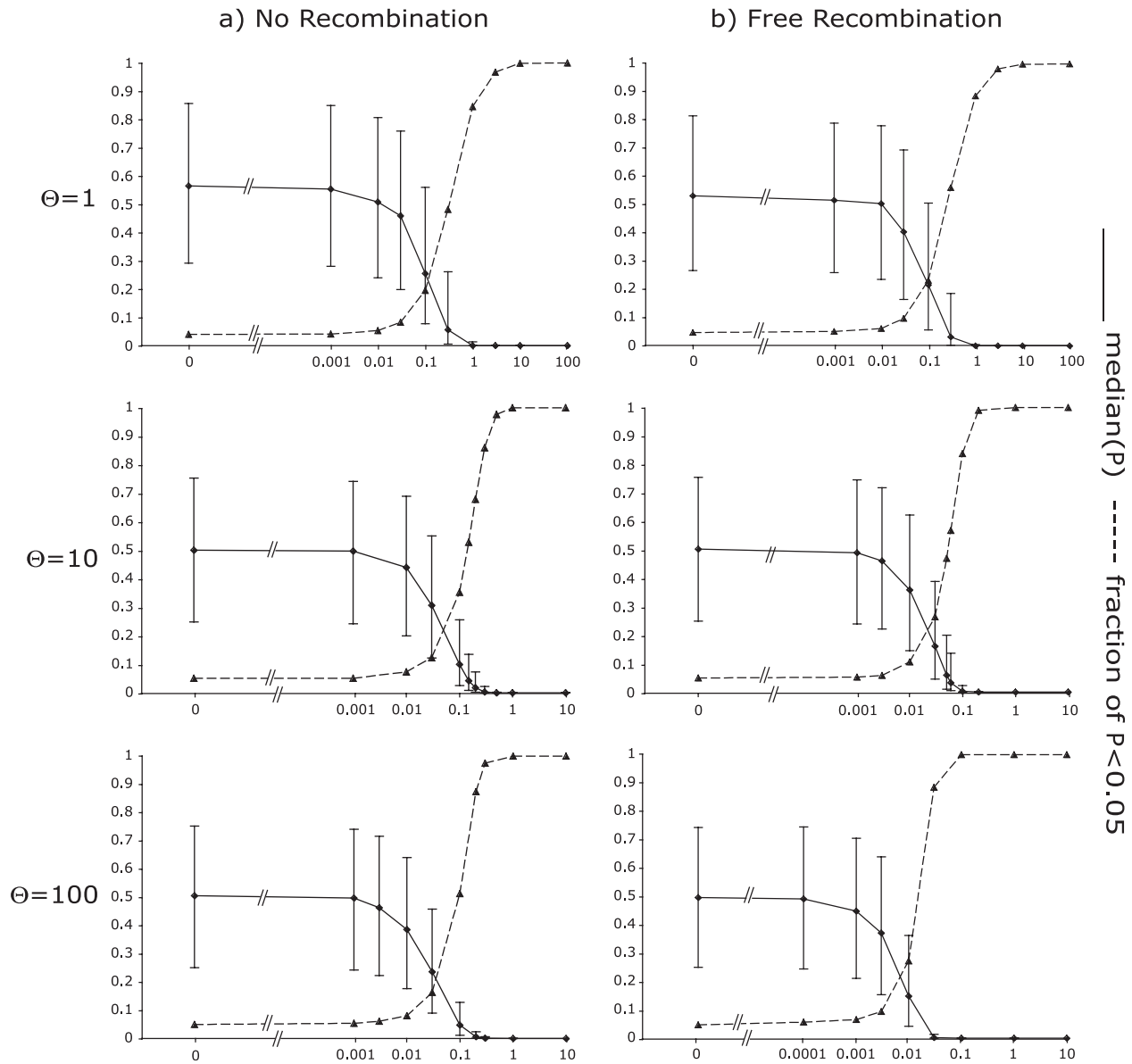

FIG. 3.—Temporal structure in neutrally evolving populations. Diamonds with continuous lines represent the median probability P (estimated by using K_s^*) and its first and third quartiles as a function of the rescaled number of generations (T_{2-1}) that separates the late from the early sample. Triangles and dashed lines represent the frequencies where the null hypothesis is rejected (at an alpha risk of 0.05) by the test as a function of T_{2-1}. We used the following set of parameters: $\Theta = 1$, 10, or 100 and $n_1 = n_2 = 20$. In (*a*) all sites share a single genealogy: there is no recombination and all sites are in complete linkage. In (*b*) each site has its own genealogy: all sites segregate independently.

population size in our simulations that gives the same amount of population change over time (by neutral drift alone) as the one observed in the data. To do so, we compared the expected rate of change for a neutrally evolving population shown in figure 3 to the corresponding data observed in figure 5.

We estimate the scaled time interval for which the test null hypothesis is rejected for half of the paired samples (see *Materials and Methods*). Assuming no recombination (fig. 3*a* with $\Theta = 10$), this time is estimated to be $T_{2-1} = 0.142$. In figure 5*b*, this time corresponds to 223 days. The generation time of HIV-1 in vivo has been estimated to be about 1.5 days (Rodrigo et al. 1999; Fu 2001; Seo et al. 2002; Markowitz et al. 2003). Thus, with the definition of the scaled time $T_{2-1} = (d_{2-1})/(1.5 \times N_e)$, where d_{2-1} is the number of days between the samples, we have $N_e = (d_{2-1})/(1.5 \times T_{2-1})$. This leads to an estimate of N_e equal to 1,047. Finally, by simulating 10^4 genealogies of the 19 samples and by using all pairwise comparisons (see *Materials and Methods*), we estimated the 95% confidence interval to be 445–2,655 under the assumption of no recombination. Assuming free recombination, but otherwise identical methods, we estimated N_e to be equal to 3,026 with a 95% confidence interval of 864–4,955.

We compared these estimates of N_e to that obtained from another method on the data from patient A. We used a recently proposed method that employs Monte Carlo simulations to estimate the likelihood of different

population sizes in a Wright-Fisher model (Anderson, Williamson, and Thompson 2000). The method uses changes in allele frequencies between two time points and tries to fit the real data to the expectation of a simulated time series sampling process of a neutral population. Since the method assumes that no mutations occur between the time points, we used frequency information from sites that are polymorphic in all samples of A. Note that this might lead to an upward bias because we have excluded some of the more extreme changes in allele frequency. We used the 11 samples with 15 sequences or more (see table 1) to increase the number of shared polymorphic sites to seven. This calculation gave an estimated N_e of about 2,800 haploid genomes, which is within the range we estimated using K_s^*.

Estimation of an Effective Mutation Rate

Based on our estimate of N_e, we can estimate an effective mutation rate per generation for the whole sequences corresponding to these data (1,098 sites). We have estimated $\Theta = 2N_e\mu$ to be about 10 using a method (average pairwise differences) that is unbiased under the assumptions of infinite-sites mutation and selective neutrality (Tajima 1983). Thus, μ is interpreted as the effective neutral mutation rate per sequence per generation. Then, $\Theta = 10$ translates into an estimate of $\mu = 10/(2 \times 1{,}047) = 4.8 \times 10^{-3}$, using the estimate of N_e obtained assuming no recombination. This gives a mutation rate per site per generation of 4.35×10^{-6} ($= 4.8 \times 10^{-3}$ /1,098). Using the 95 % confidence interval, we can compute a confidence interval for our estimation of the effective per site mutation rate that ranges from 1.7×10^{-7} to 1.0×10^{-5}. If we assume free recombination, the estimation of the effective mutation rate is 1.5×10^{-7} and the associated confidence interval ranges from 9.2×10^{-7} to 5.2×10^{-6}.

To validate our rough estimation, we used another approach to estimate this effective mutation rate. Fu (2001) developed a framework to estimate the mutation rate per day using multiple samples spaced by time interval. As with ours, this method uses the mean pairwise differences within and between the time points. The estimated effective mutation rate per day found by this method is 0.0058. Assuming a generation time of 1.5 days, we calculate a μ of 0.0024 for the whole sequence (about 1,100 sites) and, thus, a mutation rate per site per generation of 2.18×10^{-6}. This is in good agreement with our estimate based on K_s^*.

Finally, we used an alternate method that estimates both the effective population size and the effective mutation rate. This method reconstructs the likely genealogies (under the assumption of no recombination) using Bayesian statistical inference and Markov Chain Monte Carlo integration (Drummond et al. 2002). A recent version of this strategy was implemented by Drummond and Rambaut (2003) in the program Beast (available at http://evolve.zoo.ox.ac.uk/beast/). Using an HKY model for mutation (with default parameters), we ran Beast on all our sequences. The chain appeared converged after 1.65×10^7 replicates and exhibited ESS (effective sample size) values above 100 (minimum ESS values recommended by the authors). As the estimated values given by Beast were per day, we rescaled them to compare them to our estimations per generation. This gives an effective population size of 9.1×10^3 and a 95% confidence interval ranging from 7.4×10^3 to 1.1×10^4. For the mutation rate, it gives an estimate of 1.9×10^{-5}, with a 95% confidence interval ranging from 1.5×10^{-5} to 2.3×10^{-5}. These estimates are larger than ours, but the estimate of N_e is still much smaller than the actual population size. Our methods and those of Drummond et al. (2002) differ in two significant ways: (1) we assume infinite-sites mutation in estimating Θ whereas Drummond et al. (2002) allow for multiple mutations, and (2) we examine both no recombination and free recombination (and show that our estimates are fairly robust) whereas Drummond et al. (2002) assume no recombination (and account for deviations from this in the data with multiple mutations). Presumably, the differences between the methods explain the differences in parameter estimates, including the fact that Beast gives an inferred value of Θ equal to $2 \times 1{,}098 \times 1.9 \times 10^{-5} \times 9.1 \times 10^3 = 380$, compared to our $\Theta = 10$.

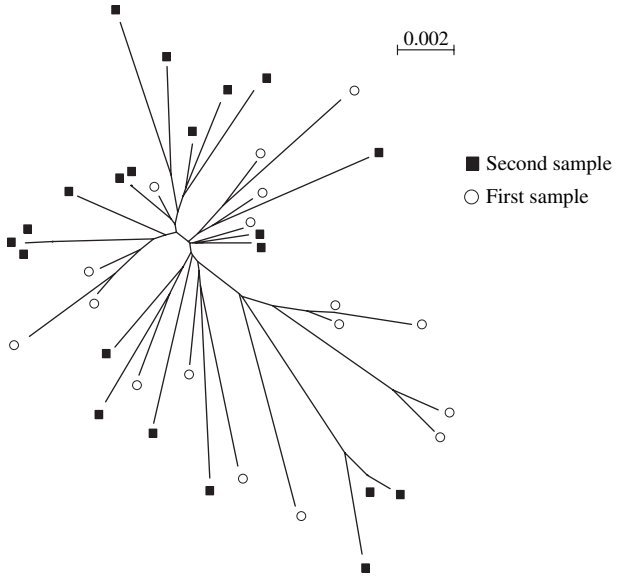

FIG. 4.—Phylogenetic trees of two samples of A. A simple neighbor-joining phylogenetic reconstruction of two samples from individual A. In table 1, the first sample is the "sample 17" and the second one is the "sample 18." These two samples are spaced by 525 days. Using the subdivision test (with 10^5 random labelings for the test), we obtain a probability of $P < 0.003$ that the two samples are picked from the same time point.

Discussion

A Standard Measure for the Rate of Population Change

Although temporal structure in HIV-1 sequences from the region analyzed (comprising ~1,100 bases from the P6 region of *gag* through *pro* and most of the RT region) from a chronically infected individual is not apparent (visually) in genealogical trees reconstructed from the data, we have found that such a structure can be detected if the interval between two samples is about 22 months or more. For this purpose, we used a test that was originally proposed to detect geographic structure

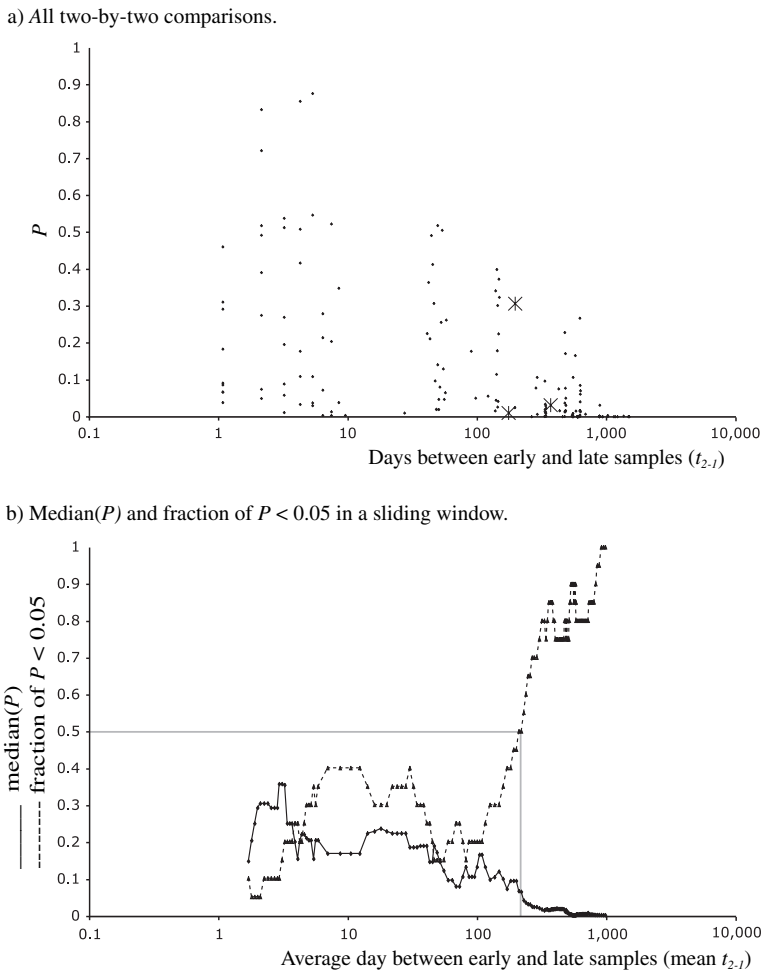

FIG. 5.—Temporal structure in real data. The probability P (estimated by using K_s^*) that two samples could be sampled from the same population is depicted as a function of the number of days between the samples. We used here all 171 possible two-by-two comparisons between the 19 samples of patient A and three comparisons between the three samples of patient B (see table 1). All values of P are plotted as dots for patient A and as stars for patient B. All tests were done using with 10^5 random labelings. In (*a*) all points are shown, whereas in (b) the median P value (diamonds with continuous lines) as well as the frequency where the null hypothesis is rejected, at an alpha risk of 5% (triangles with dashed lines), are given by the average day in a sliding window of 20 data points.

(Hudson, Boos, and Kaplan 1992) as a standard measure of temporal structure. This standard measure could be used to compare the rate of evolution of other populations under various conditions. For example, it might be interesting to measure the rate of evolution of HIV-1 population in a host treated with antiretroviral drugs.

An important assumption of the simulations above is that $\Theta = 2N_e\mu$ is constant over time. Note that this is not an assumption of the nonparametric test we have applied, but it is an important part of using the same statistics to estimate effective size. Changes in the diversity over time would tend to increase the power of the test because: (1) sequences sampled from a population with reduced diversity would tend to cluster together, and (2) bottlenecks between time points would increase the rate of population turnover. Although the mutation rate is contained in Θ, it is doubtful that the mutation rate would change over the times separating these samples. Interestingly, measures of sample sequence diversity and the independent viral load counts (table 1) do not show dramatic changes between time points, at least not in individual A. In individual B, there was a change in viral load but the diversity seems to be almost unaffected by it.

Another phenomenon that would influence the power of the test is recurrent mutations. It has been reported that the G→A and T→C mutations occur at higher frequencies than others (Mansky and Temin 1995). In the sequences of patient A, G/A and T/C pairs represents ~55% and ~30% of all polymorphisms, respectively. It is possible that multiple transition mutations have occurred at these sites, rendering some mutations unobservable. Indeed, we observed more sites that are polymorphic in both of the two patients than expected by chance (data not shown). This complication probably erodes the power of the tests since multiple, unobserved mutations would more likely affect pairwise comparisons between time points (K_{12}) than pairwise comparisons within time points (K_1 and K_2). It could lead to relatively larger K_s or K_s^* than if the infinite-sites mutation model was correct. This would then decrease the power of the test.

A Small Effective Population Size

The definition of effective population size is the size of an idealized population, exactly the population model of our simulations, which would give an equivalent rate of genetic drift as the one observed in the population under study. The "rate of genetic drift" can be defined in a variety of ways (Ewens 1979), leading to slightly different estimates of N_e. We estimated the expected rate of change of a neutrally evolving population by measuring the temporal structure of simulated samples. The comparison of this expected rate of change with the one we observed for real HIV-1 populations within a chronically infected individual leads to an estimate of its effective size of roughly 10^3 to 10^4. Like all previous estimates, this result is several orders of magnitude smaller than the actual count of replicating virus, which may be as high as 10^{10}. Most deviations from the idealized model—in fact all of them except some kinds of population structure—give values of N_e that are smaller than the actual population size. Thus, our observations are consistent with many possible causes.

The most obvious deviation from the null model, with regard to HIV-1 intra-host evolution, concerns the neutrality of the observed variation. Clearly HIV-1 is under tremendous selective pressure during an infection. In these data, evidence that selection is operating can be seen in the ratio of the rate nonsynonymous (d_N) to synonymous (d_S) mutations, which is estimated using Nei and Gojobori's method (Nei and Gojobori 1986) with Jukes and Cantor distance (Jukes and Cantor 1969) between 0.05 and 0.01 for the samples listed in table 1. A ratio of d_N/d_S of 1 is expected under neutrality and a ratio smaller than 1 under a purifying selection regime. These results then suggest that the sequences are under a regime of strong purifying selection for protein structure and function. In addition, there may be selection on synonymous sites for translation efficiency, as has been observed in other organisms (Duret and Mouchiroud 1999). There might also be selection for RNA structures in both nonsynonymous and synonymous sites. Selection on synonymous sites would tend to increase the ratio of d_N/d_S. This would reduce d_S and then imply that purifying selection for amino acid replacement is stronger than if synonymous sites where merely neutral. Interestingly, our estimation of the effective size, computed using the *gag-pol* region, is very similar to the estimates computed with the *env* gene. This suggests that even though there is evidence that the selection regime is very likely to be different in those two regions (*gag-pol* being mostly under purifying selection [see above] where *env* is subject to positive selection [Nielsen and Yang 1998; Richman et al. 2003]), other force(s), yet uncharacterized, constrain HIV populations all along their genomes.

A second possible explanation could be that HIV-1 populations do not evolve under panmixia, but rather with some population structure that causes a reduction in effective size. There is evidence that HIV populations are spatially structured because resistant strains can be in different frequencies in different organs (Epstein et al. 1991) or even in different cell types (Potter, Dwyer, and Saksena 2003). It has been suggested recently that HIV populations within patients might exhibit metapopulation structure (Frost et al. 2001), in which local populations of the virus become extinct and are recolonized by propagules from other local populations. It is well known that such patterns of colonization and replacement can reduce N_e dramatically (Slatkin 1977; Whitlock and Barton 1997; Rousset 2003; Wakeley 2004). Note that changes in population size over time are another possible cause of small N_e, but the metapopulation model we apply includes such changes, so we do not consider them as a separate force.

For illustration, we consider both metapopulation dynamics and natural selection as possible causes of the reduced intra-individual effective size of HIV. Prior estimates of intra-host N_e for HIV range from about 10^3 to 10^4 (Leigh Brown 1997; Rodrigo et al. 1999; Rouzine and Coffin 1999; Shriner et al. 2004). The estimates we made here are also in this range, although at the lower end of it. These estimates of N_e cover a broad range, but they are all much smaller than the actual intra-host population size of infected cells, which can be up to 10^{10} (Piatak et al. 1993; Haase et al. 1996). Thus, very roughly, there is between a 10^6-fold and 10^7-fold reduction in HIV effective population size that needs to be explained. Using either the standard metapopulation model (Slatkin 1977) or Gillespie's (2000) "pseudohitchhiking" model, it is possible to make a theoretical prediction for this ratio.

A general model of a metapopulation includes two kinds of dispersal: (1) regular migration and (2) recolonization after extinction (Slatkin 1977). For simplicity, we will assume that there is no regular migration among subpopulations (here cells) and, further, that extinct subpopulations are recolonized by single virus particles. In this case, the ratio of the effective size (N_e) of the population to the total size (N_T) of the population is given by $N_e/N_T = (1 - e_0)/\{N_L[1 - (1 - e_0)^2]\}$, in which e_0 is the proportion of local populations that go extinct and are recolonized every generation and N_L is the size of each local population (see Rousset [2003] or Wakeley [2004]). If we adopt a metapopulation model in which each infected cell is a local population, and we assume N_L to be about 100 viruses (Haase et al. 1996), we infer a high rate of extinction, or turnover, of local populations. In particular, the fraction of cells $(1 - e_0)$ that do contribute to the future intra-host population of HIV is between 10^{-4} to 10^{-5}.

In the pseudohitchhiking model, when no crossover is assumed, the ratio of the effective intra-host population size of HIV to the total intra-host population size is given by $N_e/N_T = 1/(1 + 2N_T\rho)$, where N_T is the total size of the population, which is assumed to be panmictic, and ρ is the per-generation probability of a selective sweep (Gillespie 2000). In this case, using $N_T = 10^{10}$, the rate of sweeps ρ ranges from 5×10^{-4} to 5×10^{-5}. The violation of the no-recombination assumption would lead to higher ρ values, depending on the frequency of crossover events. The high per-site, per-replication mutation rate of HIV, about 3.4×10^{-5} (Mansky and Temin 1995), might appear to violate the Poisson-process assumption of the pseudohitchhiking model. However, if two or more particular mutations are required for selective benefits or if only a small minority of

sites have the potential to drive a selective sweep, then this assumption might be reasonable. In contrast to the case of a metapopulation, under the pseudohitchhiking model even a small per-generation probability of a selective sweep can explain a large reduction in N_e, because N_T is so large and this appears in the denominator of the ratio. Clearly, selection and metapopulation dynamics are just two possibilities to consider, and even these are not mutually exclusive. It seems likely that a combination of factors act together to reduce the intra-host effective population size of HIV-1.

Acknowledgments

We thank all members of the Wakeley lab for their useful advice and their friendly support. We thank Cristian Castillo-Davis, Rob Kulathinal, Richard Watson, Igor Rouzine, and Daniel Shriner as well as the two anonymous reviewers for their constructive comments on the manuscript. We also thank Andrew Rambaut for his generous advice about keeping the Beast running. G.A. was funded by "La Fondation pour le Recherche Médicale." This work was supported by a Presidential Early Career Award for Scientists and Engineers from the NSF (DEB-013760) to J.W. and by a grant from the NIH (R01-CA089441) to J.M.C. J.M.C. was a Research Professor of the American Cancer Society.

Literature Cited

Anderson, E. C., E. G. Williamson, and E. A. Thompson. 2000. Monte Carlo evaluation of the likelihood for $N(e)$ from temporally spaced samples. Genetics **156**:2109–2118.

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

———. 1999. Molecular Biology of HIV. Pp. 3–40 *in* K. A. Crandall, ed. The evolution of HIV. John Hopkins University Press, Baltimore, M.D.

Daar, E. S., T. Moudgil, R. D. Meyer, and D. D. Ho. 1991. Transient high levels of viremia in patients with primary human immunodeficiency virus type 1 infection. New England J. Med. **324**:961–964.

Drummond, A., R. Forsberg, and A. G. Rodrigo. 2001. The inference of stepwise changes in substitution rates using serial sequence samples. Mol Biol. Evol. **18**:1365–1371.

Drummond, A., O. G. Pybus, A. Rambaut, R. Forsberg, and A. G. Rodrigo. 2003. Measurably evolving populations. Trends Ecol. Evol. **18**:481–487.

Drummond, A., and A. G. Rodrigo. 2000. Reconstructing genealogies of serial samples under the assumption of a molecular clock using serial-sample UPGMA. Mol. Biol. Evol. **17**:1807–1815.

Drummond, A. J., and A. Rambaut. 2003. BEAST v.1.0. available from http://evolve.zoo.ox.ac.uk/beast/.

Drummond, A. J., G. K. Nicholls, A. G. Rodrigo, and W. Solomon. 2002. Estimating mutation parameters, population history and genealogy simultaneously from temporally spaced sequence data. Genetics **161**:1307–1320.

Duret, L., and D. Mouchiroud. 1999. Expression pattern and, surprisingly, gene length shape codon usage in *Caenorhabditis*, *Drosophila*, and *Arabidopsis*. Proc. Natl. Acad. Sci. USA **96**:4482–4487.

Epstein, L. G., C. Kuiken, B. M. Blumberg, S. Hartman, L. R. Sharer, M. Clement, and J. Goudsmit. 1991. HIV-1 V3 domain variation in brain and spleen of children with AIDS: tissue-specific evolution within host-determined quasispecies. Virology **180**:583–590.

Ewens, W. J. 1979. Discrete stochastic models. Effective population size. Pp. 104–112 *in* K. Krickeberg and S. A. Levin, eds. Mathematical population genetics. Springer-Verlag, Berlin.

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

Frost, S. D., M. Nijhuis, R. Schuurman, C. A. Boucher, and A. J. Leigh Brown. 2000. Evolution of lamivudine resistance in human immunodeficiency virus type 1-infected individuals: the relative roles of drift and selection. J. Virol. **74**: 6262–6268.

Fu, Y. X. 2001. Estimating mutation rate and generation time from longitudinal samples of DNA sequences. Mol. Biol. Evol. **18**:620–626.

Galtier, N., M. Gouy, and C. Gautier. 1996. SEAVIEW and PHYLO_WIN: two graphic tools for sequence alignment and molecular phylogeny. Comput. Appl. Biosci. **12**: 543–548.

Gillespie, J. H. 2000. Genetic drift in an infinite population. The pseudohitchhiking model. Genetics **155**:909–919.

Grassly, N. C., P. H. Harvey, and E. C. Holmes. 1999. Population dynamics of HIV-1 inferred from gene sequences. Genetics **151**:427–438.

Haase, A. T., K. Henry, M. Zupancic et al. (14 co-authors). 1996. Quantitative image analysis of HIV-1 infection in lymphoid tissue. Science **274**:985–989.

Hudson, R. R. 1990. Gene genealogy and the coalescent process. Oxford Surv. Evol. Biol. **7**:1–44.

Hudson, R. R., D. D. Boos, and N. L. Kaplan. 1992. A statistical test for detecting geographic subdivision. Mol. Biol. Evol. **9**:138–151.

Jukes, T. H., and C. R. Cantor. 1969. Evolution of protein molecules. Pp. 21–132 *in* H. Munro, ed. Mammalian protein metabolism III. Academic Press, New York.

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

Kimura, M. 1969. The number of heterozygous nucleotide sites maintained in a finite population due to steady flux of mutations. Genetics **61**:893–903.

———. 1980. A simple method for estimating evolutionary rates of base substitutions through comparative studies of nucleotide sequences. J. Mol. Evol. **16**:111–120.

Kingman, J. F. C. 1982*a*. The coalescent. Stochastic processes and their applications. **13**:235–248.

———. 1982*b*. On the genealogy of large population. J. Appl. Prob. **19A**:27–43.

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.

Leigh Brown, A. J., and D. D. Richman. 1997. HIV-1: gambling on the evolution of drug resistance? Nat. Med. **3**:268–271.

Mansky, L. M., and H. M. Temin. 1995. Lower in vivo mutation rate of human immunodeficiency virus type 1 than that predicted from the fidelity of purified reverse transcriptase. J. Virol. **69**:5087–5094.

Markowitz, M., M. Louie, A. Hurley, E. Sun, M. Di Mascio, A. S. Perelson, and D. D. Ho. 2003. A novel antiviral intervention results in more accurate assessment of human immunodeficiency virus type 1 replication dynamics and T-cell decay in vivo. J. Virol. **77**:5037–5038.

Maynard Smith, J., and J. Haigh. 1974. The hitchhiking effect of a favorable gene. Genet. Res. **23**:23–35.

Mindell, D. P. 1996. Positive selection and rates of evolution in immunodeficiency viruses from humans and chimpanzees. Proc. Natl. Acad. Sci. USA **93**:3284–3288.

Moore, C. B., M. John, I. R. James, F. T. Christiansen, C. S. Witt, and S. A. Mallal. 2002. Evidence of HIV-1 adaptation to HLA-restricted immune responses at a population level. Science **296**:1439–1443.

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.

Piatak, M., Jr., M. S. Saag, L. C. Yang, S. J. Clark, J. C. Kappes, K. C. Luk, B. H. Hahn, G. M. Shaw, and J. D. Lifson. 1993. High levels of HIV-1 in plasma during all stages of infection determined by competitive PCR. Science **259**:1749–1754.

Potter, S. J., D. E. Dwyer, and N. K. Saksena. 2003. Differential cellular distribution of HIV-1 drug resistance in vivo: evidence for infection of CD8+ T cells during HAART. Virology **305**:339–352.

Rambaut, A. 2000. Estimating the rate of molecular evolution: incorporating non-contemporaneous sequences into maximum likelihood phylogenies. Bioinformatics **16**:395–399.

Richman, D. D., T. Wrin, S. J. Little, and C. J. Petropoulos. 2003. Rapid evolution of the neutralizing antibody response to HIV type 1 infection. Proc. Natl. Acad. Sci. USA **100**:4144–4149.

Rodrigo, A. G., and J. Felsenstein. 1999. Coalescent approaches to HIV population genetics. Pp. 233–272 *in* K. A. Crandall, ed. The evolution of HIV. Johns Hopkins University Press, Baltimore, M.D.

Rodrigo, A. G., E. G. Shpaer, E. L. Delwart, A. K. Iversen, M. V. Gallo, J. Brojatsch, M. S. Hirsch, B. D. Walker, and J. I. Mullins. 1999. Coalescent estimates of HIV-1 generation time in vivo. Proc. Natl. Acad. Sci. USA **96**:2187–2191.

Rousset, F. 2003. Effective size in simple metapopulation models. Heredity **91**:107–111.

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.

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

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. 1999. Evolution of HIV-1 resistance to antiviral agents. Pp. 3–40 *in* K. A. Crandall, ed. The evolution of HIV. John Hopkins University Press, Baltimore, M.D.

Shriner, D., R. Shankarappa, M. A. Jensen, D. C. Nickle, J. E. Mittler, J. B. Margolick, and J. I. Mullins. 2004. Influence of random genetic drift on human immunodeficiency virus type 1 *env* evolution during chronic infection. Genetics **166**:1155–1164.

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

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

Tavaré, S. 1984. Line-of-descent and genealogical processes, and their applications in population genetics models. Theor. Popul. Biol. **26**:119–164.

Thompson, J. D., D. G. Higgins, and T. J. Gibson. 1994. ClustalW: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice. Nucleic Acids Res. **22**:4673–4680.

Wakeley, J. 2004. Metapopulation models for historical inference. Mol Ecol **13**:865–875.

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

Whitlock, M. C., and N. H. Barton. 1997. The effective size of a subdivided population. Genetics **146**:427–441.

Edward Holmes, Associate Editor

Accepted June 15, 2004