Population Dynamics of HIV-1 Inferred From Gene Sequences

Nicholas C. Grassly, Paul H. Harvey and Edward C. Holmes

Wellcome Trust Centre for the Epidemiology of Infectious Disease, Department of Zoology, University of Oxford, Oxford OX1 3PS, United Kingdom

Manuscript received July 17, 1998
Accepted for publication October 26, 1998

ABSTRACT

A method for the estimation of population dynamic history from sequence data is described and used to investigate the past population dynamics of HIV-1 subtypes A and B. Using both *gag* and *env* gene alignments the effective population size of each subtype is estimated and found to be surprisingly small. This may be a result of the selective sweep of mutations through the population, or may indicate an important role of genetic drift in the fixation of mutations. The implications of these results for the spread of drug-resistant mutations and transmission dynamics, and also the roles of selection and recombination in shaping HIV-1 genetic diversity, are discussed. A larger estimated effective population size for subtype A may be the result of differences in time of origin, transmission dynamics, and/or population structure. To investigate the importance of population structure a model of population subdivision was fitted to each subtype, although the improvement in likelihood was found to be nonsignificant.

HUMAN immunodeficiency virus (HIV) type-1 currently infects >30 million people, and it is estimated that more than 16,000 new infections are generated every day (UNAIDS and WHO 1998). Global viral isolates from this pandemic have revealed extensive genetic diversity, but restriction of this diversity into distinct viral subtypes (Myers *et al.* 1991; Louwagie *et al.* 1993). HIV-1 represents one of a number of lineages of primate lentiviruses (Sharp *et al.* 1994) and can itself be divided into an M and O group (McCutchan *et al.* 1996), and most recently an N group (Simon *et al.* 1998). Most infections are due to the M group, which is currently classified into 10 further subtypes, A through J (McCutchan *et al.* 1996). The reason for the existence of subtypes is unclear. They do not correlate with classical neutralization serotype (Weber *et al.* 1996; Janssens *et al.* 1997), making their maintenance by epistatic selection of epitopes due to the immune system an unlikely explanation (Gupta *et al.* 1996). It seems more probable that they simply reflect the past population dynamic history of the virus.

Consider the two subtypes, A and B, for which large amounts of sequence data are available due to their dominance of the HIV-1 pandemic. Subtype A is common throughout Africa and is spread predominantly by heterosexual sex. Subtype B, in contrast, is found mainly in the United States and parts of Europe (although it has also been found in China, Japan, South America, Australasia, and Southeast Asia), and is usually associated with transmission by needle sharing among intravenous drug users (IDUs) or male homosexual sex. These subtypes may therefore be expected to show different patterns of genetic diversity, as a result of differences in their epidemiological history and mode of transmission.

The inference of past population dynamics from randomly sampled genetic sequence data can be based on either the genealogical structure relating sampled sequences (*e.g.*, Felsenstein 1992; Lundstrom *et al.* 1992; Fu and Li 1993; Fu 1994; Griffiths and Tavare 1994a,b; Kuhner *et al.* 1995, 1998) or summary statistics, such as the number of segregating sites or the distribution of pairwise differences (Watterson 1975; Tajima 1983; Rogers and Harpending 1992; Griffiths and Tavare 1996). The former approach has the advantage that it makes full use of the available data, and consequently estimators of population dynamic parameters tend to have a smaller variance when the number of sequences sampled is finite (Felsenstein 1992; Fu and Li 1993). However, the latter approach is computationally much faster, and can allow easier implementation of complex population dynamic and mutational models.

In this article a statistical framework for the inference of past population dynamics based on summary statistics is described and used to investigate the population dynamics of HIV-1 subtypes A and B. This framework is designed to be flexible and allows not only population dynamic parameter estimation, but also the fit of different population dynamic and mutational models to be tested. Population dynamic models of the epidemic spread of the virus are implemented, with or without further subdivision of the subtypes into partially isolated subpopulations or "demes." A mutational model that allows heterogeneity in the rate of mutation both along sequences and between nucleotides at a single site is

Corresponding author: Nicholas Grassly, Max Planck Institute for Evolutionary Anthropology, c/o Zoology Institute, University of Munich, Luisenstr. 14, Munich D-80333, Germany.
E-mail: grassly@zi.biologie.uni-muenchen.de

used (following their demonstrated importance in HIV-1 evolution; Eigen and Nieselt-Struwe 1990; Leitner *et al.* 1997). The performance of the estimator of effective population size is investigated using Monte Carlo simulation.

Estimates of the rate of spread of the two subtypes and their effective population sizes are obtained from both the *gag* and the *env* genes. The latter parameter is important because it determines the relative roles of genetic drift and selection in determining the evolution of HIV-1. In particular, the spread of drug-resistant mutations, which may be selectively disadvantageous in the absence of the drug (*e.g.*, Goudsmit *et al.* 1996, 1997), is dependent on the relative importance of genetic drift (Leigh Brown 1997; Leigh Brown and Richman 1997). The parameter estimates for subtypes A and B are contrasted, and reasons for their differences discussed.

THEORY

Coalescent process: If we take K sequences, randomly sampled from a large population of size N, it is possible to model their evolutionary or genealogical relationships with a simple process known as the coalescent (Kingman 1982a,b,c). This process is concerned with the rate at which sampled lineages coalesce when following a genealogy backward in time. Under the assumption of neutrality (or more specifically, exchangeability in offspring number; Kingman 1982c), this rate of coalescence of the sampled lineages is dependent solely on the dynamics of the population, and is independent of the mutational process. Conversely, the genealogy of randomly sampled sequences can therefore be used to make inferences about the dynamics of the population from which they were sampled, given a suitable coalescent model, and with a mutational process specified *a priori*.

In this article a model of the coalescent process for an expanding, subdivided haploid population is used, of which a panmictic (unsubdivided) model is a special case. The island model of subdivision due to Wright (1931) is implemented, and it is assumed that selection and recombination are absent (see discussion). If we follow a single subpopulation or deme backward in time there are therefore two types of event:

Coalescence with rate

$$\binom{k_i}{2}\sigma^2 \Big/ N_x,$$

where k_i is the number of lineages under consideration in the ith deme, N_x the population size at generation x, and σ^2 the variance in the number of offspring had by each member of the population [thus for larger σ^2 the chance of having the same parent in the preceding generation increases; under Wright-Fisher reproduction $\sigma^2 = 1.0$ (Fisher 1930; Wright 1931), while for Moran's (1958) overlapping generations model $\sigma^2 = 2N_x^{-1}$].

Migration with rate $m_{i,j}k$, where $m_{i,j}$ is the migration rate per generation out of deme i into deme j. In the implementation here the migration rate is constant for all i and j and assumed to be independent of deme size N_x (thus $m_{i,j} = m$).

If time is measured in units other than generations (*e.g.*, years), and g is the generation time in these units, then assuming that the probability of an event occurring in a unit time is small, we can calculate the probability of nothing occurring in a unit time as

$$P_{\text{nothing}} \approx \left(1 - \frac{m}{g}\right)^{k_i}\left(1 - \binom{k_i}{2}\frac{\sigma^2}{gN_t}\right) \approx 1 - \frac{mk_i}{g} - \binom{k_i}{2}\frac{\sigma^2}{gN_t}, \quad (1)$$

where N_t is the deme size at time t. Because the deme has been growing exponentially, the deme size at time t into the past is given by $N_t = N_0e^{-\lambda t}$, where N_0 is the deme size at time $t = 0$, and λ is the exponential growth rate per unit time. Given that an event has occurred at time a, then the probability that nothing occurs within the next x time units, in other words that the time between events, T, is greater than x, is given by

$$P(T > x) = \prod_{n=a}^{n=a+x}\left[1 - \frac{mk_i}{g} - \binom{k_i}{2}\frac{\sigma^2}{gN_0e^{-\lambda n}}\right] \approx \exp\left[-\sum_{n=a}^{n=a+x}\left(\frac{mk_i}{g} + \binom{k_i}{2}\frac{\sigma^2}{gN_0e^{-\lambda n}}\right)\right]. \quad (2)$$

If we switch to continuous time, this then becomes

$$P(T > x) = \exp\left[-\int_{n=a}^{n=a+x}\left(\frac{mk_i}{g} + \binom{k_i}{2}\frac{\sigma^2}{gN_0e^{-\lambda n}}\right)\right] = \exp\left[-\frac{mk_ix}{g} - \binom{k_i}{2}\frac{\sigma^2}{gN_0\lambda}\left(e^{\lambda(a+x)} - e^{\lambda a}\right)\right]. \quad (3)$$

The cumulative density function (cdf) for an event occurring within the next x time units is simply $1 - P(T > x)$. Because the area under the cdf must sum to one, the distribution of T can be obtained by setting (3) equal to a uniformly distributed random number between 0 and 1 and solving for x. If we ignore population subdivision and migration, then this cdf simplifies to that used by Slatkin and Hudson (1991, p. 559) to simulate coalescent times.

To generate times between events for the entire subdivided population, consisting of L demes, we calculate $\min[x_i]$ for all i $\{i = 1, 2, \ldots L\}$, and assume that an event has occurred at this time in the corresponding deme.

As for any population dynamic model where the population or deme size declines into the past, the coalescent approximations, based on the assumption that not

more than one event can occur per unit time, will begin to break down as N becomes smaller. However, for an exponentially growing population most coalescent events tend to occur at time $t = \ln(2N_0\sigma^{-2}\lambda)/\lambda$ (assuming $g = 1$), at which time the population size is approximately $1/\lambda$ and is independent of N_0 (Slatkin and Hudson 1991). Given that these events occur in a short space of time, the effect of the breakdown of the coalescent approximation on the total genealogical structure should therefore be minimal.

Given that an event (migration, M, or coalescence, C) has occurred in the ith deme, the probability that it is a migration event is

$$P(M|M \text{ or } C) = \frac{mk_i x}{g} \Bigg/ \left[\frac{mk_i x}{g} + \binom{k_i}{2} \frac{\sigma^2}{gN_0\lambda} (e^{\lambda(a+x)} - e^{\lambda a})\right]. \quad (4)$$

Using Equations 3 and 4, it is possible to simulate gene genealogies under a range of population dynamic scenarios. Sequences can subsequently be evolved down these gene genealogies, given a suitable substitution model, to generate sets of aligned sequences. A program to simulate sequence alignments under certain population dynamic scenarios and mutational models, called Treevolve, is available on the World Wide Web at http://evolve.zoo.ox.ac.uk/.

Parameter estimation: Sequence alignments, simulated under certain population dynamic models, can be compared to observed data to not only test the fit of these models, but also to estimate their associated parameters. In the method for estimating population dynamic parameters described here (see also Grassly and Holmes 1998), summary statistics are used to make this comparison. These summary statistics are the mean and variance of the distribution of pairwise differences, and were used because of their well-documented responsiveness to changes in population dynamics (Watterson 1975; Hudson 1987; Slatkin and Hudson 1991; Rogers and Harpending 1992; Simonsen *et al.* 1995; Hey 1997). These can be calculated for the observed data ($\bar{k}_o$ and $S^2_{k,o}$, respectively), and for sets of data simulated under a particular coalescent and mutational model (which we will denote Θ, giving $\bar{k}_{\Theta,i}$ and $S^2_{k,\Theta,i}$; $i = 1, 2, 3, \ldots$). To calculate an approximate likelihood of the model given $\bar{k}_o$ and $S^2_{k,o}$, define the index, $I_i(\Theta)$, as

$$I_i(\Theta) = \begin{cases} 1 & \text{if } (\bar{k}_{\Theta,i} - \bar{k}_o)^2 + (S^2_{k,\Theta,i} - S^2_{k,o})^2 < \partial, \\ 0 & \text{otherwise,} \end{cases} \quad (5)$$

where ∂ is an arbitrary small number (following Weiss and von Haeseler 1998). This gives a measure of the goodness-of-fit of the model Θ, to the observed data ($\bar{k}_o$ and $S^2_{k,o}$). It therefore seems reasonable to assume that the probability of the data given the model can be given by

$$P(\bar{k}_o, S^2_{k,o}|\Theta) \propto \frac{1}{n} \sum_{i=1}^{i=n} I_i(\Theta), \quad (6)$$

where n is a large number corresponding to the number of simulations under Θ. Because the likelihood $L(\Theta|\bar{k}_o, S^2_{k,o}$ is by definition proportional to $P(\bar{k}_o, S^2_{k,o}|\Theta)$ (Fisher 1921; Edwards 1992), maximization of the right-hand side of Equation 6 over Θ gives the maximum-likelihood estimate (MLE), $\hat{\Theta}$. For small n, variance in the simulated $\bar{k}$ and S^2_k can translate into bias in $\hat{\Theta}$. For example, the variance in $\bar{k}$ and S^2_k depends on the total genealogical length, which in turn is linearly proportional to the effective population size, N_0. Thus, for the above method with a finite number of simulations, as the estimator $\hat{N}_0$ ($N_0 \in \Theta$) increases, so too does its variance, leading to an upward bias. However, as $n \to \infty$ and $\partial \to 0$, the variance is expected to tend to zero, and the estimator becomes unbiased.

It is possible to calculate the region of the parameter space about MLEs where simple hypotheses would not be rejected at the 95% significance level using the likelihood ratio. This space determines the set of admissible hypotheses (*sensu.*, Wilks 1938), which are termed here *approximate* confidence limits or intervals (although they do not represent confidence limits in the strictest sense where the type I error rate should equal the significance level, α). At a confidence limit we are effectively restricting the parameter(s) of interest to a particular value(s), θ_R. We can therefore calculate the likelihood ratio as

$$\Lambda = \max[\ln L(\theta_R|\bar{k}_o, S^2_{k,o})] - \max[\ln L(\Theta|\bar{k}_o, S^2_{k,o})]. \quad (7)$$

By definition, θ_R is nested in Θ and hence for large n, -2Λ is approximately χ^2-distributed, with the number of degrees of freedom equal to the number of parameters restricted (Wilks 1938). Thus, for example, the 95% confidence limits about a single MLE are defined by the θ_R, which satisfy $\Lambda = -1.92$.

As the number of estimated parameters increases, the confidence interval becomes wider. Both the population parameters and the two summary statistics used to estimate them are nonindependent to differing extents. Thus it is unclear exactly how many degrees of freedom there are, and whether two unique population genetic parameter estimates can be obtained from the two summary statistics. Fortunately, likelihood surfaces can be calculated about MLEs to reveal the range of plausible parameter estimates (those with a likelihood where Λ is less than the critical, $\frac{1}{2}\chi^2$-distributed, value).

Because of the assumption of neutrality made by the coalescent process, the process of mutation can be decoupled from that generating the gene genealogy (Kingman 1982a,b,c). The parameters governing the

mutational process can therefore be calculated independently and directly from the sequence data, using explicit substitution models (equivalent to mutational models under neutrality). The procedure used here estimates the parameters of a substitution model chosen *a priori*, at the same time as the branch lengths and topology of the phylogenetic tree linking the sampled sequences, by maximizing their likelihood given the observed sequence data (for a description see Swofford *et al.* 1996). The validity of this procedure depends on the accuracy of the chosen substitution model, and of the maximum-likelihood (ML) tree topology. The former is improved by using a wide range of testable substitution models. In the case of the latter, the accuracy of the tree topology tends to have little effect on the inferred substitution model unless the deepest bifurcation of the tree (at the root) partitions the taxa incorrectly (*e.g.*, see Sullivan *et al.* 1996), which is unlikely for any method of phylogenetic reconstruction. Furthermore, even if recombination occurs, invalidating a strictly bifurcating phylogeny, it is unlikely to cause an incorrect phylogenetic partition at such a deep level.

MATERIALS AND METHODS

Sequence data: Complete sequences of the *gag* and *env* genes from subtypes A and B of HIV-1 were extracted from the Los Alamos HIV sequence database (Myers *et al.* 1996), and the published alignments used. Sequences previously demonstrated to be intersubtype recombinants were removed (including *gag* sequences isolated in Thailand, which are subtype A in origin, but constitute the recombinant subtype E; Gao *et al.* 1996; McCutchan *et al.* 1996). In addition, those sequences cloned from a common source (patient or isolate) were removed to promote the random sampling of the sequences. Despite this removal it is likely that sequences were not sampled entirely randomly with respect to geographic origin. The effect of this on estimates of population parameters is considered in the discussion.

In total, 13 *env* and 23 *gag* sequences were obtained for subtype A, and 54 *env* and 26 *gag* sequences for subtype B. The accession numbers, dates, and places of isolation of these sequences are available upon request from the authors. All sequences were isolated between 1983 and 1996 about a mean isolation date of 1991 (SD = 3.65 yr). The coalescent models described here assume sequences are isolated at the same time. The effect of noncontemporaneous-isolation-date estimates of population parameters is analogous to the effects of a substitutional process that is more stochastic than the assumed Poisson process (see discussion).

Phylogenetic relationships and inference of substitution process: Phylogenetic trees for the four sets of aligned sequences were inferred at the same time as the substitution process, using the maximum likelihood method implemented in a test version of PAUP* made available by the author David Swofford. The general reversible substitution model was used (Lanave *et al.* 1984; Rodriguez *et al.* 1990), together with a discrete gamma model of rate heterogeneity with four rate categories (Yang 1994). In each case the parameters governing the relative rate of substitution between the different nucleotides and the gamma shape parameter, α, were estimated from the data. For each alignment two trees were constructed, one where sequences (tips) were constrained to be contemporaneous with a single rate of substitution, and one where the sequences were not constrained (and thus substitution rates down each branch could vary). Using the likelihood-ratio statistic the goodness-of-fit of the coalescent assumption of contemporary tips and a molecular clock were assessed. The constrained tree is a special case of the unconstrained tree, where the unconstrained tree has $k - 2$ extra parameters, which are free to vary (Felsenstein 1981). Thus the significance of the likelihood ratio was assessed using the χ^2-distribution with $k - 2$ d.f. (Wilks 1938).

The substitution rate of the *gag* and *env* genes of the virus, μ, was set to 5×10^{-3} per site per year for all estimates of population dynamic parameters, consistent with previous estimates for these genes (Li *et al.* 1988). This parameter is completely confounded with the effective population size N_0, and hence estimates of N_0 depend on the accuracy of μ. Although the codon-based structure of the genes has been ignored, the inference of population dynamic history relies on the shape of the genealogy relating the sampled sequences, which is likely to remain the same for all three codon positions due to linkage. Selection does, however, play a role in shaping some of the diversity of HIV-1, and the implications of this for the analysis presented here are considered in the discussion.

Inference of population dynamic history: Initially a panmictic coalescent model was fitted to each of the two HIV-1 subtypes. Such a model is simply derived from Equation 3 by setting $m = 0$ and considering only a single deme. Under this population dynamic model there are two free parameters to be estimated: the exponential growth rate λ, and the compound parameter $g\sigma^{-2}N_0$, which determines the rate of coalescence. This latter parameter can be defined as the current real-time effective population size and is denoted throughout this article as N_τ. The real-time effective population size determines the probability that two randomly sampled gene sequences have a common ancestor within the last year ($\approx N_\tau^{-1}$, for large N_τ) and is analogous to Wright's (1931) effective population number, N_e, which determines the probability that two randomly chosen individuals shared a common ancestor in the preceding generation ($\approx N_e^{-1}$). Because each sequence in the sample of HIV-1 sequences can be considered to represent the viral population from an infected patient, the "individual" of the population genetic model implemented here is an infected patient. Infected individuals transmit their virus to other individuals to generate more "offspring" whose viral sequence resembles that of the "parent." This coalescent model of the transmission process allows variation in offspring number (*i.e.*, $\sigma^2 > 1$), but assumes that there is no correlation across generations in the number of offspring had by a genealogical lineage (no Hill-Robertson effect; Felsenstein 1974). This latter assumption seems valid given that there is so far little evidence for the existence of HIV-1 genetic factors determining transmission rates, or viral virulence. The census population size, N, for this study is therefore the number of individuals infected by HIV-1 subtypes A and B. The effective population size may be smaller or bigger than N, determining the role of genetic drift in the fixation of mutations. It is completely confounded with the substitution rate, and hence when estimating λ any error in μ can be accounted for by allowing N_τ to vary (although of course any error in μ will affect the actual estimates of N_τ).

MLEs of N_τ and λ for the four alignments were obtained and likelihood surfaces about the estimates produced. Using the χ^2-approximation to the likelihood-ratio statistic $\sim$95% confidence limits about these estimates were calculated. Subsequently the exponential growth rate (λ) was fixed at 0.693 yr^{-1}, and estimates of the real-time effective population size were obtained. This growth rate is equivalent to a doubling time of 1 yr, a value consistent with that observed for the incidence

TABLE 1

Maximum-likelihood estimated substitution parameters for constrained (clock) and unconstrained phylogenies inferred from the *env* and *gag* genes isolated from HIV-1 subtypes A and B

Parameters	Subtype A *env* gene	Subtype A *gag* gene	Subtype B *env* gene	Subtype B *gag* gene
Base frequencies π_A, π_C, π_G, π_T	0.346, 0.175, 0.233, 0.247	0.370, 0.188, 0.247, 0.195	0.343, 0.174, 0.235, 0.248	0.372, 0.189, 0.248, 0.191
Relative rate parameters[a] inferred from constrained phylogeny	2.35, 4.55, 0.770, 0.750, 4.36	1.66, 6.98, 0.944, 0.663, 8.73	3.15, 5.70, 0.994, 1.16, 5.19	1.10, 3.58, 0.570, 0.298, 4.98
Gamma parameter, α, from constrained phylogeny	0.339	0.350	0.298	0.265
Relative rate parameters inferred from unconstrained phylogeny	2.31, 4.58, 0.762, 0.807, 4.28	1.63, 6.80, 0.912, 0.653, 8.32	3.26, 5.86, 1.01, 1.27, 5.46	1.09, 3.76, 0.579, 0.333, 4.73
α from unconstrained phylogeny	0.291	0.386	0.323	0.240

[a] The five relative rate parameters of the general reversible model of nucleotide substitution (Lanave *et al.* 1984; Rodriguez *et al.* 1990) are in the order A → C, A → G, A → T, C → G, C → T, with G → T constrained to 1.0 (because parameters are relative to one another).

of HIV-1 infection (where a range from 4.9 to 15.6 mo is seen; May and Anderson 1989).

The subdivided population coalescent model with two demes was also fitted to the observed sequence data. This model has three free parameters: λ, N_τ, and the migration rate, $M = m/g$ (where λ and M have units per year, and N_τ is the real-time effective deme size). Although nonunique due to the use of only two summary statistics ($\bar{k}_o$ and $S^2_{k,o}$), and therefore only 2 d.f., MLEs of these parameters were obtained and the likelihood at these MLEs recorded. The likelihood ratio of the subdivided to the unsubdivided model was calculated for each sequence alignment and its significance assessed using the χ^2-approximation with 1 d.f. (because there is one additional free parameter, M).

Performance of $\hat{N}_\tau$ and the χ^2-approximation to the likelihood ratio: The performance of the estimator of the real-time effective population size and the validity of the χ^2-approximation to the likelihood ratio were assessed using simulations. For each value of N_τ, which was estimated from HIV-1 with the doubling time set at 1 yr (*i.e.*, for A and B subtypes, *gag* and *env* genes), 100 sequence alignments were simulated with the relevant number of sequences and the mutational parameters in Table 1. For each sequence alignment N_τ was estimated and the ML value recorded. The likelihood of the actual N_τ under which the data were simulated was also recorded, and the ratio to the ML value calculated. In this way the performance of both the estimator and the χ^2-approximation to the likelihood ratio could be assessed.

RESULTS

Phylogenetic relationships and inference of substitution process: For each of the four alignments a molecular clock was rejected using the likelihood-ratio statistic [log-likelihood ratio = 14.32 ($P = 0.01$) and 57.45 ($P < 0.005$) for subtype A *env* and *gag* genes, respectively, and 322.01 ($P < 0.005$) and 52.89 ($P < 0.005$) for subtype B *env* and *gag* genes]. This is unsurprising given the noncontemporaneous nature of the viral isolates and the possibility of greater-than-Poisson stochasticity in the substitution process. Furthermore, the likelihood-ratio test of a molecular clock based on a complete phylogeny is very stringent (powerful). A single anomalous branch length can cause rejection of the clock. In contrast, the method for inferring population parameters described here is based on summary statistics, which are less sensitive to such rate heterogeneity. Thus, despite the rejection of a clock by the likelihood ratio, a coalescent analysis is considered appropriate (see discussion for a consideration of the implications of rate heterogeneity for estimates of population parameters).

The substitution processes inferred from both the constrained and unconstrained phylogenies are shown in Table 1. It can be seen that constraint of the tips of the phylogeny has little effect on the estimates of the substitution parameters, despite the lack of fit of a molecular clock. Adenosine (A) is more common than the other nucleotides, a result of the preference for G to A substitution in the HIV-1 genome (Vartanian *et al.* 1994). The gamma parameter, α, governing rate heterogeneity along the sequences is <0.4 in all cases, indicating a fair degree of rate heterogeneity (in agreement with estimates from other *gag* and *env* alignments; Leitner *et al.* 1997). This heterogeneity is most likely a result of functional constraint at certain positions, although there may also be a role of positive selection (see discussion).

The phylogenetic trees constrained to have contemporaneous tips for the two genes from each subtype are shown in Figure 1. They have a fairly pronounced bush- or star-like topology, which is indicative of exponential growth of the population from which the sequences have been sampled (Slatkin and Hudson 1991). The places of isolation of the sequences are shown on the trees. These can be seen to cluster to a certain extent, indicating a possible role of population subdivision in generating the observed sequence diversity.

Inferred population dynamic history: Under a panmictic model, the population growth rate, λ, and the

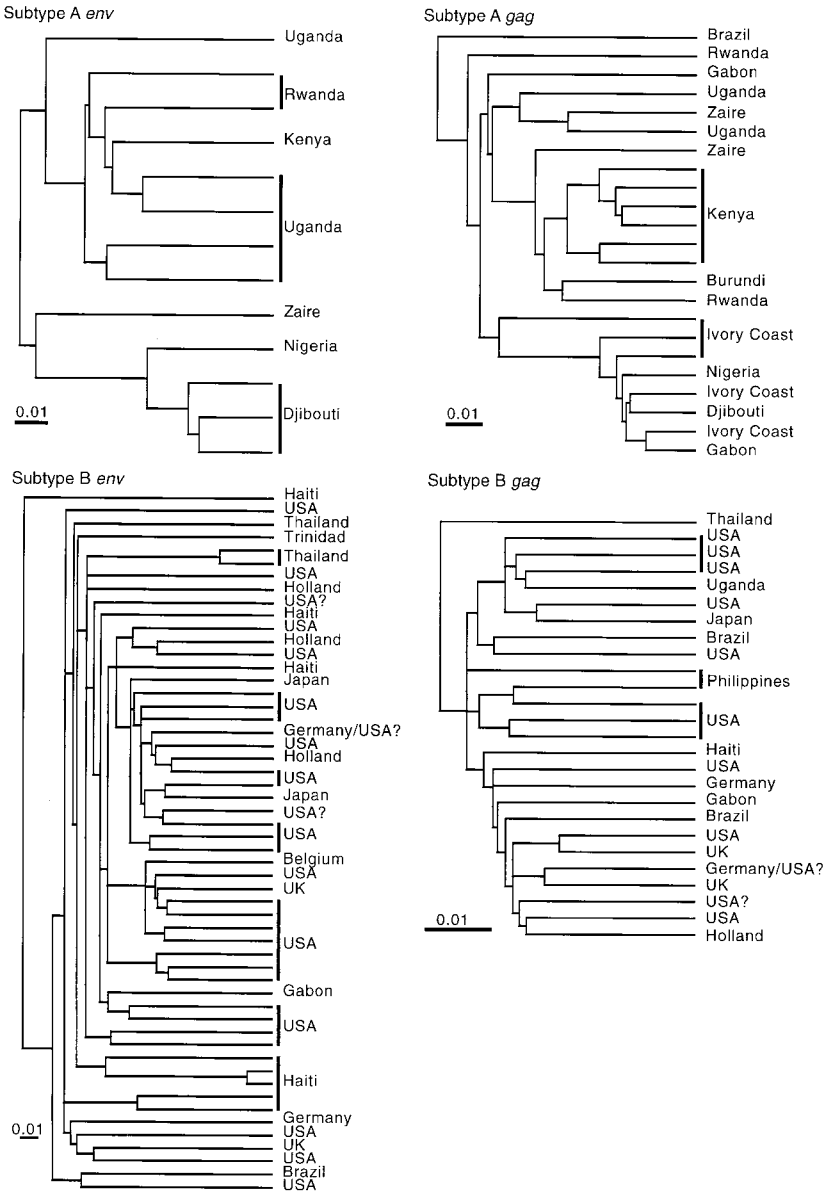

Figure 1.—Maximum-likelihood phylogenetic trees inferred for the *gag* and *env* alignments from HIV-1 subtypes A and B, showing the place of isolation of the sequences. A question mark indicates that the place of isolation was not recorded, and in such cases the location of the authors' academic institute has been substituted.

current real-time effective population size, N_τ, were coestimated to give the MLEs and ~95% confidence limits shown in Figure 2. It can be seen that the likelihood surface is ridged, and indeed the confidence limits are open ended as λ and N_τ increase. This is expected for any coestimate of these two parameters based on the properties of gene genealogies, and is a result of the coupling of the effect of λ and N_τ for large λ. A star-like phylogeny is produced when λ is large, and thus if N_τ further increases, λ can also increase to produce a genealogy of the same shape and length. However, as λ becomes smaller, a more structured phylogeny is produced, and the coupling of λ and N_τ breaks down. This therefore allows a minimum possible λ and N_τ to be estimated. In the case of subtypes A and B of HIV-1 it is meaningful to ask what their minimum rate of spread is. For subtype B this was found to be 0.5 and 0.4 per year for the *env* and *gag* genes, respectively (corresponding to a maximum possible doubling time, $t_{1/2}$, of 17 and 21 mo; Figure 2). This fits with the average $t_{1/2}$ of the HIV-1 pandemic estimated from AIDS case reports to the World Health Organization at ~10 mo (May and Anderson 1989). The maximum possible $t_{1/2}$ of subtype A was 42 and 83 mo for the *env* and *gag* genes, respectively. Although this suggests a less rapid spread of subtype A than subtype B, the confidence limits on the estimate of λ for subtype A will be wider due to fewer available sequences. Hence, it is difficult to draw conclusions about differences in the minimum rate of spread for the two subtypes.

Although conclusions about the exact rate of spread of the two subtypes are difficult, it is clear from Figure 2 that the ratio of $\hat{\lambda}$ to $\hat{N}_\tau$ for subtype B is larger than that for subtype A. This implies that either subtype B is spreading at a faster rate than A (1.2 or 3.2 times faster based on the *env* and *gag* genes, respectively), or

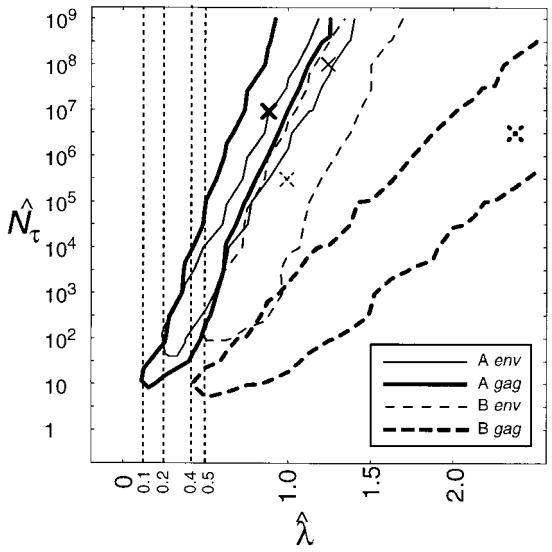

Figure 2.—Approximate 95% confidence limits about the MLEs of λ and N_τ (marked by crosses) for HIV-1 subtypes A and B, based on *gag* and *env* sequence alignments. Because we are interested in the minimum possible value for λ, not N_τ, the confidence limits have been constructed using the χ^2-approximation to the likelihood-ratio statistic with 1 d.f.

that subtype A has a larger current real-time effective population size than B, or a combination of the two. Given our knowledge of the epidemiology of HIV in Africa and in the United States and Europe, it seems unlikely that subtype B is growing at a much faster rate than subtype A. Indeed, the doubling time of HIV incidence in Africa (where subtype A is found) is similar, if somewhat shorter, than in the United States and Europe (where subtype B is found). If $t_{1/2}$ is fixed at 1 yr, consistent with this epidemiological data, it is possible to estimate the corresponding N_τ that maximizes the likelihood for each subtype, as shown in Figure 3. This figure clearly demonstrates that subtype A must have a larger N_τ than subtype B if they have been growing at the same rate, although for the *env* alignments the 95% confidence limits overlap.

Figure 3 also reveals that both subtypes have a very low current real-time effective population size, N_τ, despite the large numbers of individuals infected with these subtypes. If HIV-1 has been spreading with a doubling time of about 1 yr, then the 95% confidence limits about $\hat{N}_\tau$ for the *env* and *gag* alignments from subtype A range from 2.4×10^4 to 5.8×10^6, while for subtype B they range from 2.6 to 8.8×10^4. At the lower confidence limits of $\hat{N}_\tau$ the coalescent approximations begin to break down. However, at the mean $\hat{N}_\tau$ for subtypes A and B of $\sim 10^5$ and $\sim 10^2$, respectively, the approximations are valid, and at the upper confidence limits for each subtype (5.8×10^6 and 8.8×10^4) the approximations will be accurate. These low values of $\hat{N}_\tau$ are robust to error in the specified mutation rate μ ($= 5 \times 10^{-3}$), which is unlikely to vary by more than an order of magnitude, and therefore may indicate an important role for the stochastic process of genetic drift in the fixation of mutations. However, it is also possible that selection is acting to reduce N_τ, as will be discussed.

The results of fitting a model of population subdivision to the data are shown in Table 2. This table gives the likelihood ratio, Λ, or support, of the subdivided model over the panmictic model, and the MLEs of its associated parameters. The improvement in the fit of the model when subdivision is added can be seen to be small, and absent in one case (subtype B, *env* alignment). Although Λ is on average somewhat higher for subtype A than subtype B, in no case is the improved fit significant at the 5% level when Λ is assessed using the χ^2-approximation ($\frac{1}{2}\chi^2$-value for 1 d.f. = 1.92). Estimates of growth rates are similar to those obtained for the panmictic model.

Performance of $\hat{N}_\tau$ and the χ^2-approximation to the likelihood ratio: The mean estimated N_τ for each simulated dataset and the likelihood of the true value, together with the number of times the true value was rejected at the 95% significance level ($\alpha = 0.05$) using the χ^2-approximation to the likelihood ratio, are shown in Table 3. The percentage of cases where the true $\hat{N}_\tau$ was rejected at the 95% significance level is consistent with this significance level (the 95% confidence interval about the expected α of 0.05 is 0.016–0.113 under the binomial distribution with $n = 100$). Furthermore, the distribution of likelihood ratios was well fitted by a χ^2-distribution (results not shown), as expected given the nesting of the different hypotheses for N_τ (see theory: *Parameter estimation*).

It can be seen that the estimates of N_τ are biased upward, although the true N_τ falls within the 95% confidence limits at the correct frequency. This upward bias is a result of the finite number of simulations carried out when calculating each likelihood value (in this case, $n = 20$ in Equation 6). Simulated values of the summary statistics $\bar{k}$ and S_k^2 under a large N_τ have a high variance and so can result in a reasonable likelihood for a large estimated $\hat{N}_\tau$ even if the sequence data are from a population with a small N_τ. Conversely, simulations under a small N_τ have a low variance, and so the likelihood for a small $\hat{N}_\tau$ is unlikely to be large for data sampled from a big population. As $n \to \infty$ and $\partial \to 0$, this variance will tend to zero and the bias will disappear.

It can be concluded that the small N_τ reported here for HIV-1 subtypes A and B are not a result of bias in the method, which would tend to result in slight over- rather than underestimation of N_τ.

DISCUSSION

The analysis of *gag* and *env* sequences presented here reveals a small current real-time effective population size, N_τ, for subtypes A and B of HIV-1, of $\sim 10^5$ and $\sim 10^2$, respectively. This real-time effective population

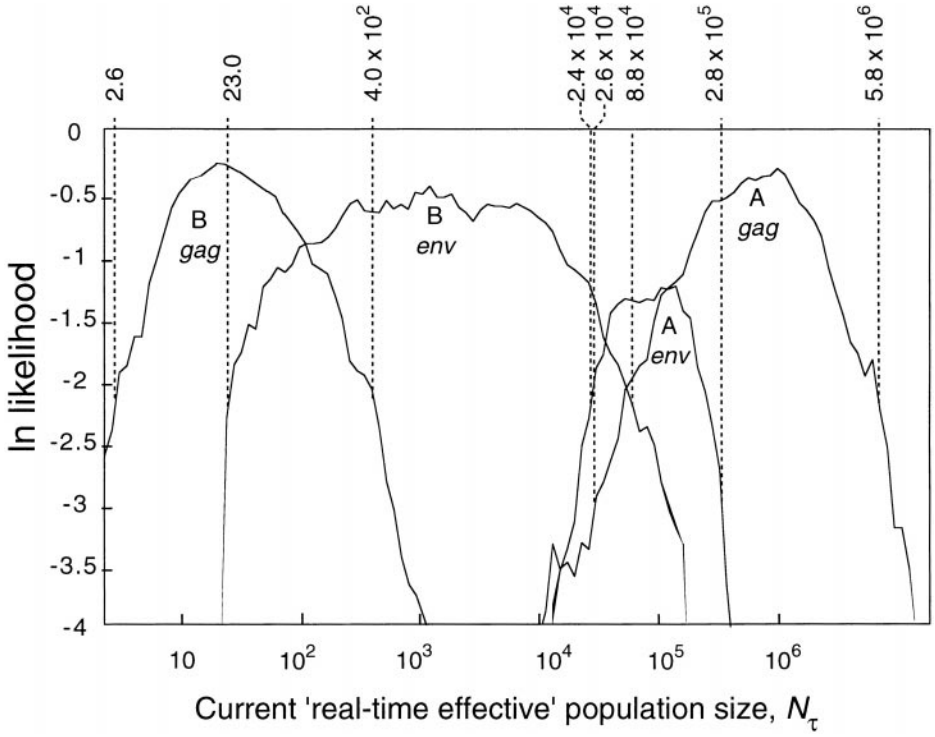

Figure 3.—Likelihood surfaces about the MLE of N_τ on a log scale, for HIV-1 subtypes A and B based on the *env* and *gag* alignments, with λ fixed at 0.693 (corresponding to a doubling time of 1 yr). $\hat{N}_\tau$ for subtype B was 17.8 and 2.1×10^3 for the *gag* and *env* genes, respectively, while for subtype A it was 8.3×10^5 and 1.2×10^5, respectively. Approximate 95% confidence limits about $\hat{N}_\tau$, calculated using the χ^2-approximation (1 d.f.), are shown. These are 2.6 to 400 and 23 to 8.8×10^4 for subtype B *gag* and *env* alignments, respectively, and 2.4×10^4 to 5.8×10^6 and 2.6×10^4 to 2.8×10^5 for subtype A.

size can be converted to Wright's (1931) effective population size, N_e, by dividing by the generation time, *g*. However, it is unclear what the generation time is for HIV-1 infections. Rates of partner change for homosexuals and heterosexuals tend to be on the order of 1 yr (Anderson and May 1988), while it is unclear what the rate of needle sharing is among IDUs (Kaplan 1989; Blower *et al.* 1991). There is a similar ambiguity regarding the probability of transmission in each case. If the generation time is on the order of 1 yr, then $N_\tau = N_e$, while for shorter generation times $N_\tau < N_e$. However, even for a very short generation time of 1 mo, the estimated N_τ imply that the N_e of subtypes A and B is $\sim 10^6$ and $\sim 10^3$, respectively. Both these estimates fall below the census size of HIV-1 infections, particularly for subtype B that predominates in the United States and Europe, where HIV-1 is estimated to infect more than 1.4 million people (UNAIDS and WHO 1998).

Why is N_e small? Two features of HIV-1 may be important in causing a small effective population size: the transmission dynamics and/or natural selection. The transmission dynamics of HIV-1 are such that there is a large variance in the rate at which new infections are generated (*i.e.*, a large σ^2, and hence small N_e). The different modes of transmission of HIV-1, via homosexual or heterosexual sex, needle-sharing among IDUs, or contamination of blood products, are all associated with different rates of transmission of the virus. For example, transmission of the virus through a susceptible network of needle-sharing IDUs is likely to be initially more rapid than through a susceptible heterosexual population (Kaplan 1989; Blower *et al.* 1991). How-

TABLE 2

Fit of model of subdivision to HIV-1 subtypes A and B

Subtype/gene	−ln *L* for panmixis	−ln *L* for subdivision	Likelihood ratio (Λ)	$\hat{N}_\tau$	$\hat{\lambda}$	$\hat{M}$
A *env*	0.931	0.619	0.312	3.2×10^6	0.868	0.76
A *gag*	0.274	0.094	0.180	1.32×10^7	0.749	1.0
B *env*	1.109	1.109	0.000	7.5×10^8	1.61	2.8
B *gag*	0.186	0.117	0.069	2.4×10^6	1.91	1.1

The log likelihoods for the fit of the panmictic and subdivided population dynamic models to the *gag* and *env* alignments from subtypes A and B, together with parameter estimates for the latter model. The likelihood ratio of the subdivided to the panmictic model is given and can be seen to be less than the $\frac{1}{2}\chi^2$-critical value in all cases. Parameters estimated are the current real-time effective deme size, N_τ, the growth rate of the demes, λ, and the migration rate *M*.

TABLE 3

Performance of $\hat{N}_\tau$ and the χ^2-approximation to the likelihood ratio

Gene	*env*	*gag*	*env*[a]	*gag*
Subtype	A	A	B	B
Actual N_τ	1.2×10^5	8.3×10^5	2.1×10^3	17.8
Mean $\hat{N}_\tau$	1.96×10^5	3.1×10^6	7.8×10^3	72.7
% cases actual N_τ outside confidence limits	5	3	4	1

The mean estimated N_τ for sequence alignments simulated under known N_τ's, together with the number of times the likelihood of the known N_τ was significantly worse than the estimated (maximum-likelihood) value. The four known N_τ's used correspond to the values estimated for HIV-1 subtype A and B, *gag* and *env* genes, when the doubling time was set to 1 yr. In each case the number of simulated sequence alignments used was 100, and the number of sequences in each alignment corresponded to the number of sequences used for the original estimate.

[a] For the subtype B *env* alignment 25 sequences were simulated in each alignment rather than the 54 used originally to estimate N_τ, to save computational time.

ever, even within a sexual mode of transmission, there may be substantial heterogeneity in the rate of spread due to variation in partner exchange rates and transmission probabilities (Anderson *et al.* 1992; Service and Blower 1995). For instance, with regard to the former, for the average rate of homosexual partner exchange (8.7 per year) in England and Wales, the variance is ~600 per year (Anderson and May 1988). With regard to the latter, there is substantial evidence for heterogeneity in the susceptibility of different people to infection. For example, ~10% of the Caucasian population in the United States possess a 32-bp deletion in the chemokine receptor CCR5, and are less susceptible to HIV-1 infection when homozygous (Dean *et al.* 1996; Michael *et al.* 1997).

In an analogous way to heterogeneity in susceptibility, subdivision of the human population according to geographic, social, and behavioral barriers may also play a role in reducing the effective population size below the census size, N. If the migration rate is high, then a subdivided population of total size N_e, consisting of L subpopulations or demes, will begin to approximate a panmictic population of size N_e/L as the migration rate $M \rightarrow \infty$ (for lower migration rates, subdivision can inflate the effective population size, N_e, above N). Subdivision of the HIV-1 subtypes A and B is suggested by a certain degree of clustering of the places of isolation on the phylogenetic trees relating the sequences (see Figure 1), but a simple two-deme model of population subdivision was not found to give a significantly better fit to either subtype (see Table 2). This may reflect the minimal effect subdivision has on patterns of sequence diversity when the population has been growing at a rapid rate. In such cases, a star-like phylogeny where all coalescences occur at approximately the same time will be produced no matter what the level of subdivision, unless single lineages survive into the past in demes of very small size ($N_e < 1$). It is also possible that the particular model of subdivision and the use of summary statistics to make inferences result in a poor improvement in the likelihood when the subdivided model is assessed. The use of a genealogy-based method (*sensu*., Griffiths and Tavare 1994b; Kuhner *et al.* 1998), where place of isolation is an explicit part of the model, would be likely to result in improved power in testing the fit of a model of subdivision.

The analysis of the *gag* and *env* genes also reveals a larger $\hat{N}_\tau$ for subtype A (found in Africa) than subtype B (found in the United States and Europe), although the confidence limits overlap for the *env* gene (see Figure 3). This may simply be the result of a greater age, number of infections, and hence diversity of HIV-1 in Africa than in Europe and the United States. Alternatively, it may be a result of differences in transmission dynamics. Subtype B is associated mainly with the AIDS epidemic in homosexuals and IDUs in developed countries, where transmission rates may be more variable than for the heterosexual transmission associated with subtype A (Anderson and May 1988; Blower *et al.* 1991; Greenhalgh 1996). A larger value of σ^2 for subtype B could therefore explain the smaller estimated N_τ.

Although transmission dynamics shape the observed sequence diversity of HIV-1, natural selection is also likely to play a role. The high levels of rate heterogeneity observed along both the *gag* and *env* genes (Table 1) suggest an important role of functional constraint. This is further evidenced by the restricted number of nonsynonymous as opposed to synonymous changes for these two genes (the d_N:d_S ratio is 0.44 and 0.51 for the *env* subtype A and B alignments, and 0.25 and 0.24 for the *gag* alignments). However, although functional constraint can reduce linked genetic diversity (Charlesworth *et al.* 1993), such a restriction will be reflected by a lower estimated substitution rate, μ, and therefore is unlikely to result in a reduction of $\hat{N}_e$ (which is confounded with μ) unless such constraint is accompanied by occasional selective sweeps of fit mutants through the subtype. Such selective sweeps would reduce linked

variation while maintaining a high substitution rate and hence could therefore decrease $\hat{N}_e$. Selective sweeps may be a feature of HIV-1 evolution, but the high frequency at which HIV-1 recombines (Diaz *et al.* 1995; Robertson *et al.* 1995; Zhu *et al.* 1995; Moutouh *et al.* 1996) will restrict this reduction in diversity to the vicinity of the selected mutation. Thus selective sweeps could be reducing $\hat{N}_\tau$ substantially, but only if they occur fairly often. So far, little evidence concerning whether such sweeps occur at the subtype level or what their frequency is has accumulated.

Role of recombination: Recombination reduces the variance of the distribution of pairwise differences, while little affecting the mean of this distribution. Thus if recombination is ignored, when λ and N_e are coestimated, λ will be overestimated (the phylogeny simply becomes more star-like). However, estimates of N_e for a fixed λ should be little affected because for a star phylogeny $\hat{N}_e$ is mainly determined by the mean of the distribution of pairwise differences. The estimates of N_e for HIV-1 reported here are therefore valid, despite the lack of recombination in the coalescent model. In the future it may be possible to use a coalescent model (*e.g.*, see Hudson 1987; Grassly and Holmes 1998) to estimate the rates of recombination within and between HIV-1 subtypes.

Nonrandom sampling: HIV-1 sequences are unlikely to be sampled randomly with respect to geographic origin for political and economic reasons. Because the estimates of N_e given here are based mainly on mean sequence diversity, such nonrandom sampling will not be too problematic unless subtype diversity is systematically underrepresented. This will occur when countries with prevalent HIV-1 are not included in the sample. In such cases estimates of N_e will be for the regions that are adequately sampled only. For example, because partial V3 sequences from China that group with subtype B (Shao and Wolf 1995) were not included in the alignment, the estimates of N_e presented here reflect subtype B diversity in the United States, Europe, and Japan, but not mainland Asia.

Importance of intrahost dynamics: Both the substitution and genealogical processes modeled focus on interhost evolution. This is because each pair of sequences sampled from subtypes A and B is likely to be separated by a large number of transmission events. A lineage in the coalescent model therefore reflects a population with an effective size of $\sim 10^3$ (the estimated intrahost population size; Leigh Brown 1997; Leigh Brown and Richman 1997) punctuated by repeated bottlenecks occurring at transmission (Holmes *et al.* 1992). For neutral mutations, where substitution rates are independent of population size, it is therefore reasonable to model the interhost substitution process with a standard Poisson model.

If many mutations within the host are nonneutral (*e.g.*, due to immune surveillance), then viral population size within a coalescent lineage becomes important and the rate of interhost substitution may be more variable than the assumed Poisson process (Araki and Tachida 1997). Such rate heterogeneity is supported by the rejection of constrained molecular clock phylogenetic trees by the likelihood-ratio test (see results). The effect of this rate heterogeneity may be further enhanced by the noncontemporaneous nature of the HIV-1 sequence isolates, which can cause sister taxa in the viral genealogy to have unequal branch lengths. The effect of departure from the assumption of a molecular clock and a genealogy with contemporary tips on estimates of N_e is unclear. It is not obvious how it could cause a reduction in estimates of $\hat{N}_e$ based on summary statistics, given that the mean sequence diversity is likely to remain the same. However, for likelihood calculations based on explicit representation of the HIV-1 genealogy, a more accurate model of the substitution process will be required. There is no evidence for systematic differences in the substitution process or rate between subtypes A and B that may be causing the apparent population parameter differences.

It is possible that selective pressures within the host may increase the effect of intrahost polymorphism on observed differences between sampled sequences. For this reason the large number of immunogenic epitopes along the *env* protein (reflected by a larger d_N:d_S ratio compared with the *gag* polyprotein; Korber *et al.* 1996) may explain why the *env* gene gives a less clear distinction between subtypes A and B. Selection during primary infection for particular configurations of the *env* V3 loop (*e.g.*, see Zhang *et al.* 1993) will not affect the rate of substitution of linked (or unlinked) neutral variants (Birky and Walsh 1988), and so is unlikely to have any effect on rates of interhost evolution, and hence estimates of N_e.

Implications of a small N_e: The small value of $\hat{N}_e$, if not a consequence of past selective sweeps, has important implications for the spread of drug-resistant mutations, which are typically at a selective disadvantage in the absence of the drug (*e.g.*, Goudsmit *et al.* 1996, 1997). In general, for populations of constant size N_e, selective effects that are less than the reciprocal of N_e are negligible compared to the stochasticity of drift. Thus for HIV-1, depending on the future rate of spread of the virus, the small $\hat{N}_e$ may indicate an important role for genetic drift in the fixation of drug-resistant mutations. Although the values of selective coefficients for drug-resistant mutations seem to be variable ($s = -0.004$ to -0.25 within drug-naïve patients; Goudsmit *et al.* 1996), the cases of drug resistance in drug-naïve patients that have begun to be reported (Conlon *et al.* 1994; Najera *et al.* 1995; Goudsmit *et al.* 1996; Kozal *et al.* 1996; Cornelissen *et al.* 1997) suggest that, at least in some cases, s is small enough for genetic drift to occur. Of course, the high mutation rate and rapid turnover of HIV-1 within infected individuals (Ho *et al.* 1995; Wei

et al. 1995) means that single mutations conferring resistance to drugs are likely to arise by chance even in drug-naïve patients (Bonhoeffer and Nowak 1997). However, for drug resistance requiring multiple mutations preexistence is unlikely (Bonhoeffer and Nowak 1997) unless the mutations provide some drug resistance on their own, in which case they may spread and then be brought together in a single viral genome by recombination. The small $\hat{N}_e$ therefore suggests that drug resistance requiring multiple mutations may arise through the stochastic effects of drift. This has important implications for the efficacy of drugs and combination drug therapy, because preexistence of a few drug-resistant viruses within an individual implies that even if the drug therapy is effective in dramatically reducing viral load, resistance will quickly emerge (Bonhoeffer and Nowak 1997; Bonhoeffer *et al.* 1997).

Many thanks to Oliver Pybus for providing early sequence alignments. This work was supported by grants from the Biotechnology and Biological Sciences Research Council (BBSRC), The Royal Society, and The Wellcome Trust.

LITERATURE CITED

Anderson, R. M., and R. M. May, 1988 Epidemiological parameters of HIV transmission. Nature **333:** 514–519.

Anderson, R. M., R. M. May, T. W. Ng and J. T. Rowley, 1992 Age-dependent choice of sexual partners and the transmission dynamics of HIV in sub-Saharan Africa. Philos. Trans. R. Soc. Lond. B Biol. Sci. **336:** 135–155.

Araki, H., and H. Tachida, 1997 Bottleneck effect on evolutionary rate in the nearly neutral mutation model. Genetics **147:** 907–914.

Birky, C. W., and J. B. Walsh, 1988 Effects of linkage on rates of molecular evolution. Proc. Natl. Acad. Sci. USA **85:** 6414–6418.

Blower, S. M., D. Hartel, H. Dowlatabadi, R. M. Anderson and R. M. May, 1991 Drugs, sex and HIV: a mathematical model for New York City. Philos. Trans. R. Soc. Lond. B Biol. Sci. **331:** 171–187.

Bonhoeffer, S., and M. A. Nowak, 1997 Pre-existence and emergence of drug resistance in HIV-1 infection. Proc. R. Soc. Lond. Ser. B Biol. Sci. **264:** 631–637.

Bonhoeffer, S., R. M. May, G. M. Shaw and M. A. Nowak, 1997 Virus dynamics and drug therapy. Proc. Natl. Acad. Sci. USA **94:** 6971–6976.

Charlesworth, B., M. T. Morgan and D. Charlesworth, 1993 The effect of deleterious mutations on neutral molecular variation. Genetics **134:** 1289–1303.

Conlon, C. P., P. Klenerman, A. Edwards, B. A. Larder and R. E. Phillips, 1994 Heterosexual transmission of human immunodeficiency virus type-1 variants associated with zidovudine resistance. J. Infect. Dis. **169:** 411–415.

Cornelissen, M., R. van den Burg, F. Zorgdrager, V. Lukashov and J. Goudsmit, 1997 *pol* gene diversity of five human immunodeficiency virus type 1 subtypes: evidence for naturally occurring mutations that contribute to drug resistance, limited recombination patterns, and common ancestry for subtypes B and D. J. Virol. **71:** 6348–6358.

Dean, M., M. Carrington, C. Winkler, G. A. Huttley, M. W. Smith *et al.*, 1996 Genetic restriction of HIV-1 infection and progression to AIDS by a deletion allele of the CKR5 structural gene. Science **273:** 1856–1862.

Diaz, R. S., E. C. Sabino, A. Mayer, J. W. Mosley and M. P. Busch, 1995 Dual Human Immunodeficiency Virus type-1 infection and recombination in a dually exposed transfusion recipient. J. Virol. **69:** 3273–3281.

Edwards, A. W. F., 1992 *Likelihood.* Extended Edition. The Johns Hopkins University Press, London.

Eigen, M., and K. Nieselt-Struwe, 1990 How old is the immunodeficiency virus? AIDS **4:** S85–S93.

Felsenstein, J., 1974 The evolutionary advantage of recombination. Genetics **78:** 737–756.

Felsenstein, J., 1981 Evolutionary trees from DNA sequences: a maximum likelihood approach. J. Mol. Evol. **17:** 368–376.

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

Fisher, R. A., 1921 On the 'probable error' of a coefficient of correlation deduced from a small sample. Metron **1:** part 4, 3–32.

Fisher, R. A., 1930 *The Genetical Theory of Natural Selection*. Clarendon Press, Oxford.

Fu, Y. X., 1994 A phylogenetic estimator of effective population size or mutation rate. Genetics **136:** 685–692.

Fu, Y. X., and W. H. Li, 1993 Maximum likelihood estimation of population parameters. Genetics **134:** 1261–1270.

Gao, F., D. L. Robinson, S. G. Morrison, H. X. Hui, S. Craig *et al.*, 1996 The heterosexual human immunodeficiency virus type 1 epidemic in Thailand is caused by an intersubtype (A/E) recombinant of African origin. J. Virol. **70:** 7013–7029.

Goudsmit, J., A. de Ronde, D. D. Ho and A. S. Perelson, 1996 Human immunodeficiency virus fitness *in vivo*: calculations based on a single zidovudine resistance mutation at codon 215 of reverse transcriptase. J. Virol. **70:** 5662–5664.

Goudsmit, J., A. de Ronde, E. de Rooij and R. de Boer, 1997 Broad spectrum of *in vivo* fitness of human immunodeficiency virus type 1 subpopulations differing at reverse transcriptase codons 41 and 215. J. Virol. **71:** 4479–4484.

Grassly, N. C., and E. C. Holmes, 1998 The use of Monte Carlo simulation to infer population dynamic history from DNA sequence data, pp. 91–112 in *Proceedings of the Trinational Workshop on Molecular Evolution*, edited by M. K. Uyenoyama, A. von Haeseler and N. Takahata. Duke University Publications Group, Durham, NC.

Greenhalgh, D., 1996 Effects of heterogeneity on the spread of HIV AIDS among intravenous drug users in 'shooting galleries'. Math. Biosci. **136:** 141–186.

Griffiths, R. C., and S. Tavare, 1994a Ancestral inference in population genetics. Stat. Sci. **9:** 307–319.

Griffiths, R. C., and S. Tavare, 1994b Sampling theory for neutral alleles in a varying environment. Philos. Trans. R. Soc. Lond. B Biol. Sci. **344:** 403–410.

Griffiths, R. C., and S. Tavare, 1996 Monte Carlo inference methods in population genetics. Math. Comp. Mod. **23:** 141–158.

Gupta, S., M. C. J. Maiden, I. M. Feavers, S. Nee, R. M. May *et al.*, 1996 The maintenance of strain structure in populations of recombining infectious agents. Nat. Med. **2:** 437–442.

Hey, J., 1997 Mitochondria and nuclear genes present conflicting portraits of human origins. Mol. Biol. Evol. **14:** 166–172.

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

Holmes, E. C., L. Q. Zhang, P. Simmonds, C. A. Ludlam and A. J. Leigh Brown, 1992 Convergent and divergent sequence evolution in the surface envelope glycoprotein of human immunodeficiency virus type 1 within a single infected patient. Proc. Natl. Acad. Sci. USA **89:** 4835–4839.

Hudson, R. R., 1987 Estimating the recombination parameter of a finite population-model without selection. Genet. Res. **50:** 245–250.

Janssens, W., A. Buve and J. N. Nkengasong, 1997 The puzzle of HIV-1 subtypes in Africa. AIDS **11:** 705–712.

Kaplan, E. H., 1989 Needles that kill: modelling human immunodeficiency virus transmission via shared drug injection equipment in shooting galleries. Rev. Infect. Dis. **11:** 289–298.

Kingman, J. F. C., 1982a The coalescent. Stoch. Process. Appl. **13:** 235–248.

Kingman, J. F. C., 1982b On the genealogy of large populations. J. Appl. Probab. **19A:** 27–43.

Kingman, J. F. C., 1982c Exchangeability and the evolution of large populations, pp. 97–112 in *Exchangeability in Probability and Statistics*, edited by G. Koch and F. Spizzichino. North-Holland, Amsterdam.

Korber, B. T. M., C. Brander, B. D. Walker, R. Koup, J. P. Moore

et al. (Editors), 1996 *HIV-1 Molecular Immunology Database*. Los Alamos National Laboratory, Los Alamos, NM.

Kozal, M. J., N. Shah, N. P. Shen, R. Yang, R. Fucini *et al.*, 1996 Extensive polymorphisms observed in HIV-1 clade B protease gene using high-density oligonucleotide arrays. Nat. Med. **2:** 753–759.

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

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

Lanave, C., G. Preparata, C. Saccone and G. Serio, 1984 A new method for calculating evolutionary substitution rates. J. Mol. Evol. **20:** 86–93.

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.

Leitner, T., S. Kumar and J. Albert, 1997 Tempo and mode of nucleotide substitutions in *gag* and *env* gene fragments in human immunodeficiency virus type 1 populations with a known transmission history. J. Virol. **71:** 4761–4770.

Li, W.-H., M. Tanimura and P. M. Sharp, 1988 Rates and dates of divergence between AIDS virus nucleotide sequences. Mol. Biol. Evol. **5:** 313–330.

Louwagie, J., F. E. McCutchan, M. Peeters, T. P. Brennan, E. Sanders-Buell *et al.*, 1993 Phylogenetic analysis of gag genes from 70 international HIV-1 isolates provides evidence for multiple genotypes. AIDS **7:** 769–780.

Lundstrom, R., S. Tavare and R. H. Ward, 1992 Estimating substitution rates from molecular data using the coalescent. Proc. Natl. Acad. Sci. USA **89:** 5961–5965.

May, R. M., and R. M. Anderson, 1989 The transmission dynamics of human immunodeficiency virus (HIV). Biomath. **18:** 263–311.

McCutchan, F. E., M. O. Salminen, J. K. Carr and D. S. Burke, 1996 HIV-1 genetic diversity. AIDS **10** (Suppl. 3): S13–S20.

Michael, N. L., G. Chang, L. G. Louie, J. R. Mascola, D. Dondero *et al.*, 1997 The role of viral phenotype and CCR-5 gene defects in HIV-1 transmission and disease progression. Nat. Med. **3:** 338–340.

Moran, P. A. P., 1958 Random processes in genetics. Proc. Camb. Philos. Soc. **54:** 60–71.

Moutouh, L., J. Corbeil and D. D. Richman, 1996 Recombination leads to the rapid emergence of HIV-1 dually resistant mutants under selective drug pressure. Proc. Natl. Acad. Sci. USA **93:** 6106–6111.

Myers, G., B. Korber, J. A. Berzofsky, T. F. Smith and G. N. Pavlakis (Editors), 1991 *Human Retroviruses and AIDS 1991*. Theoretical Biology and Biophysics Group, Los Alamos National Laboratory, Los Alamos, NM.

Myers, G., B. Korber, B. Foley, K.-T. Jeang, J. W. Mellors *et al.* (Editors), 1996 *Human Retroviruses and AIDS 1996*. Theoretical Biology and Biophysics Group, Los Alamos National Laboratory, Los Alamos, NM.

Najera, I., A. Holguin, M. E. Quinonesmateu, M. A. Munozfernandez, R. Najera *et al.*, 1995 *pol* gene quasi-species of human immunodeficiency virus: mutations associated with drug resistance in virus from patients undergoing no drug-therapy. J. Virol. **69:** 23–31.

Robertson, D., P. Sharp, F. McCutchan and B. Hahn, 1995 Recombination in HIV-1. Nature **374:** 124–126.

Rodriguez, F., J. L. Oliver, A. Marin and J. R. Medina, 1990 The general stochastic model of nucleotide substitution. J. Theor. Biol. **142:** 485–501.

Rogers, A. R., and H. Harpending, 1992 Population growth makes waves in the distribution of pairwise genetic differences. Mol. Biol. Evol. **9:** 552–569.

Service, S. K., and S. M. Blower, 1995 HIV transmission in sexual networks: an empirical analysis. Proc. R. Soc. Lond. Ser. B Biol. Sci. **260:** 237–244.

Shao, Y., and H. Wolf, 1995 Unpublished. Genbank accession numbers U20001–U20054.

Sharp, P. M., D. L. Robertson, F. Gao and B. H. Hahn, 1994 Origins and diversity of human immunodeficiency viruses. AIDS **8** (Suppl. 1): S27–S42.

Simon, F., P. Mauclère, P. Roques, I. Loussert-Ajaka, M. C. Müller-Trutwin *et al.*, 1998 Identification of a new human immunodeficiency virus type 1 distinct from group M and group O. Nat. Med. **4:** 1032–1037.

Simonsen, K. L., G. A. Churchill and C. F. Aquadro, 1995 Properties of statistical tests of neutrality for DNA polymorphism data. Genetics **141:** 413–429.

Slatkin, M., and R. R. Hudson, 1991 Pairwise comparisons of mitochondrial-DNA sequences in stable and exponentially growing populations. Genetics **129:** 555–562.

Sullivan, J., K. E. Holsinger and C. Simon, 1996 The effect of topology on estimates of among-site rate variation. J. Mol. Evol. **42:** 308–312.

Swofford, D. L., G. J. Olsen, P. J. Waddell and D. M. Hillis, 1996 Phylogenetic inference, pp. 407–514 in *Molecular Systematics*, Ed. 2, edited by D. M. Hillis, C. Moritz and B. Mable. Sinauer Associates, Sunderland, MA.

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

UNAIDS, and WHO, 1998 Report on the global HIV/AIDS pandemic June 1998. http://www.unaids.org.

Vartanian, J. P., A. Meyerhans, M. Sala and H. S. Wain, 1994 G → A hypermutation of the human immunodeficiency virus type 1 genome: evidence for dCTP pool imbalance during reverse transcription. Proc. Natl. Acad. Sci. USA **91:** 3092–3096.

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

Weber, J., E. M. Fenyo, S. Beddows, P. Kaleebu, A. Bjorndal *et al.*, 1996 Neutralization serotypes of human immunodeficiency virus type-1 field isolates are not predicted by genetic subtype. J. Virol. **70:** 7827–7832.

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

Weiss, G., and A. von Haeseler, 1998 Inference of population history using a likelihood approach. Genetics **149:** 1539–1546.

Wilks, S. S., 1938 The large-sample distribution of the likelihood ratio for testing composite hypotheses. Ann. Math. Stat. **9:** 60–62.

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

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

Zhang, L. Q., P. Mackenzie, A. Cleland, E. C. Holmes, A. J. L. Brown *et al.*, 1993 Selection for specific sequences in the external envelope protein of human immunodeficiency virus type-1 upon primary infection. J. Virol. **67:** 3345–3356.

Zhu, T. F., N. Wang, A. Carr, S. Wolinsky and D. D. Ho, 1995 Evidence for coinfection by multiple strains of human immunodeficiency virus type-1 subtype-B in an acute seroconvertor. J. Virol. **69:** 1324–1327.

Communicating editor: M. Slatkin