

## Estimation of Effective Population Size of HIV-1 Within a Host: A Pseudomaximum-Likelihood Approach

Tae-Kun Seo,<sup>\*,†,1</sup> Jeffrey L. Thorne,<sup>‡</sup> Masami Hasegawa<sup>\*,†</sup> and Hirohisa Kishino<sup>§</sup>

<sup>\*</sup>Department of Biosystems Science, The Graduate University for Advanced Studies, Hayama, Kanagawa, 240-0193, Japan, <sup>†</sup>The Institute of Statistical Mathematics, 4-6-7 Minami-Azabu Minato-ku, Tokyo 106-8569, Japan, <sup>‡</sup>Bioinformatics Research Center, North Carolina State University, Raleigh, North Carolina 27695-7566 and <sup>§</sup>Laboratory of Biometrics, Graduate School of Agriculture and Life Sciences, University of Tokyo, Yayoi 1-1-1 Bunkyo-ku, Tokyo 113-8657, Japan

Manuscript received August 13, 2001  
Accepted for publication January 21, 2002

### ABSTRACT

Using pseudomaximum-likelihood approaches to phylogenetic inference and coalescent theory, we develop a computationally tractable method of estimating effective population size from serially sampled viral data. We show that the variance of the maximum-likelihood estimator of effective population size depends on the serial sampling design only because internal node times on a coalescent genealogy can be better estimated with some designs than with others. Given the internal node times and the number of sequences sampled, the variance of the maximum-likelihood estimator is independent of the serial sampling design. We then estimate the effective size of the HIV-1 population within nine hosts. If we assume that the mutation rate is  $2.5 \times 10^{-5}$  substitutions/generation and is the same in all patients, estimated generation lengths vary from 0.73 to 2.43 days/generation and the mean (1.47) is similar to the generation lengths estimated by other researchers. If we assume that generation length is 1.47 days and is the same in all patients, mutation rate estimates vary from  $1.52 \times 10^{-5}$  to  $5.02 \times 10^{-5}$ . Our results indicate that effective viral population size and evolutionary rate per year are negatively correlated among HIV-1 patients.

ONE of the most striking features of human immunodeficiency virus (HIV)-1 infections is the high variation among patients of the length of the asymptomatic period. During this period, the number of CD4+ T cells decreases slowly, the immune system gradually weakens, and the viral load is roughly constant. The length of the asymptomatic stage can range from a few years to >10 years. At the end of the asymptomatic period, progression to AIDS starts. This progression is characterized by increasing viral loads and a continued decrease in the number of CD4+ T cells (for a review, see VISCIDI 1999).

Mathematical models to account for the variation of the asymptomatic period have been introduced (NOWAK *et al.* 1990; NOWAK and MAY 1992), but its underlying cause is still unknown. Despite the variation in the length of the asymptomatic stage, SHANKARAPPA *et al.* (1999) noted that the pattern of viral evolution during this period can be divided into three stages: (1) an early phase with a linear increase over time in both the amount of extant sequence diversity and in divergence from the HIV-1 sequence that founded the infection; (2) an intermediate phase in which sequence divergence keeps increasing, but diversity stabilizes or decreases; and (3) a late phase in which divergence becomes stable and diversity is stable or decreases.

Because of their high evolutionary rates, RNA viruses such as HIV-1 are potentially more informative regarding evolutionary processes than are more slowly evolving model systems. For example, with the constant rate assumption of a molecular clock, the rate of molecular evolution and the internal node times of a phylogenetic tree can be simultaneously estimated with serially sampled viral data (RAMBAUT 2000). In contrast, rate and time parameters are confounded when all sequences have been isolated at the same date. With slowly evolving organisms, differences of a few years in sequence sampling dates are not sufficient to effectively separate rate and times. Therefore, supplemental information such as fossil data is needed to estimate internal node times for slowly evolving organisms but is fortunately not required for estimating these times from serially sampled viral data. Similarly, with contemporaneously isolated sequences, the molecular clock hypothesis of a constant rate of molecular evolution over time cannot be separated from the more general hypothesis that all evolutionary lineages share a common rate of evolution at a given time but that this common rate changes over time. With serially sampled viral data, these two hypotheses can be distinguished (SEO *et al.* 2002).

As another illustration of the rich information available in serially sampled viral data sets, important population genetic parameters that are confounded for contemporaneously isolated sequence data can be separately estimated for serially sampled data. For example, effective population size and the rate of mutation per genera-

<sup>1</sup>Corresponding author: Bioinformatics Research Center, Box 7566, North Carolina State University, Raleigh, NC 27695-7566.  
E-mail: seo@statgen.ncsu.edu

tion are two of the central parameters in population genetic theory. With contemporaneously isolated sequence data, only their product can be estimated. With serially sampled viral data and with a known generation time, effective population size and mutation rate per generation can be separately estimated. Likewise, with serially sampled viral data and with a known mutation rate, effective population size and generation time can be disentangled (RODRIGO and FELSENSTEIN 1999; RODRIGO *et al.* 1999). Although there is less confounding of parameter estimates from serially sampled data than for contemporaneously isolated data, effective population size, mutation rate per generation, and generation time cannot all three be simultaneously estimated solely from serially sampled sequence data. However, when externally derived values of either generation time or mutation rate are employed, the other two parameters can be separately inferred.

The stages identified by SHANKARAPPA *et al.* (1999) of genetic diversity and divergence during the HIV-1 asymptomatic period are intriguing, but it may also be worthwhile to focus on the differences among patients in HIV-1 evolution. Here, a coalescent-based approach is developed and employed to estimate important population genetic parameters particular to HIV-1 evolution in each of the nine patients who were included in the SHANKARAPPA *et al.* (1999) study. We find a negative correlation between the effective viral population size within a patient and the rate of viral sequence evolution per year and we discuss possible sources of this negative correlation.

In this article, a two-stage estimation procedure is adopted. Times of internal nodes are estimated from sequence data and then these estimated node times serve as the basis for inferring effective population size. Because the main interest is effective population size and not times of internal nodes, internal node times are nuisance parameters in our analysis and the number of these nuisance parameters increases as the number of sequences increases. This situation can be analyzed with a pseudomaximum-likelihood approach (GONG and SAMANIEGO 1981).

THEORY

**Coalescent theory:** Subsequent to the pioneering work of KINGMAN (1982a,b), coalescent theory has attracted widespread interest (for a review see HUDSON 1991). According to coalescent theory, the probability density function of the time duration ( $t_i$ ) of the period with exactly  $i$  different ancestral lineages is

$$p(t_i|N_c) = \frac{i(i-1)}{2N_c} \exp\left\{-\frac{i(i-1)}{2N_c}t_i\right\},$$

where we denote  $N_c$  as the effective population size to emphasize that it is not necessarily the same as the total

population size. Here, the time intervals  $t_i$  are measured in terms of generations. The vector of these intervals is denoted  $\mathbf{t}$ . The mean and variance of  $t_i$  are, respectively,

$$E[t_i] = \frac{2N_c}{i(i-1)}$$

and

$$\text{Var}[t_i] = \frac{4N_c^2}{i^2(i-1)^2}.$$

Initially, we treat the vector of time intervals  $\mathbf{t}$  as observed and we discuss inference of  $N_c$  for this situation. Later, we replace this vector  $\mathbf{t}$  in the equations below by its estimate. When there are  $n$  contemporaneously isolated sequences, the likelihood function  $L_1 = L_1(N_c|\mathbf{t})$  is

$$L_1 = \prod_{i=n}^2 \frac{i(i-1)}{2N_c} \exp\left\{-\frac{i(i-1)}{2N_c}t_i\right\}. \tag{1}$$

Using the log-likelihood equation

$$\frac{\ln L_1}{\partial N_c} = -\frac{(n-1)}{N_c} + \sum_{i=n}^2 \frac{i(i-1)}{2N_c^2}t_i = 0, \tag{2}$$

the maximum-likelihood estimate of the effective population size is

$$\hat{N}_c = \frac{1}{2(n-1)} \sum_{i=n}^2 i(i-1)t_i. \tag{3}$$

This estimate is unbiased and its variance is

$$\text{Var}(\hat{N}_c) = \frac{1}{2^2(n-1)^2} \sum_{i=2}^n i^2(i-1)^2 \text{Var}(t_i) = \frac{N_c^2}{n-1} \tag{4}$$

(*e.g.*, FELSENSTEIN 1992; RODRIGO *et al.* 1999).

**Coalescent likelihood of serially sampled data:** Recently, coalescent theory has been applied to the investigation of serially sampled viral populations (FELSENSTEIN *et al.* 1999; RODRIGO and FELSENSTEIN 1999; RODRIGO *et al.* 1999; DRUMMOND and RODRIGO 2000). For serially sampled data, Equation 1 has to be slightly modified. Suppose that  $n_1$  and  $n_2$  sequences are sampled at times  $k_1$  and  $k_2$ , respectively, where  $k_2$  represents an earlier time than  $k_1$ . We assume that the  $n_1$  sequences sampled at time  $k_1$  have coalesced into  $c$  ancestral lineages when the  $n_2$  sequences are sampled at time  $k_2$  (Figure 1). For the period between times  $k_2$  and  $k_1$ ,  $t_i$  represents the length of time in which the  $n_1$  sequences sampled at time  $k_1$  have exactly  $i$  ancestral lineages. Note that  $t_c^*$  (Figure 1) is a special case in which the end of the time interval is determined not by a coalescent event but by the known time  $k_2$ . For times preceding  $k_2$ , it is convenient to have  $s_i$  denote the amount of time during which there were exactly  $i$  ancestral lineages. The expressions  $e(t_i)$  and  $e(s_i)$ , respectively, specify the points in time at which intervals  $t_i$  and  $s_i$  end (*i.e.*, the interval endpoints that are most recent). The time point  $e(t_{n_1}) = k_1$ . For a vector  $\mathbf{t}$  representing the pertinent time intervals (*e.g.*,  $t_{n_1}, \dots, t_{c+1}, t_c^* + s_{c+n_2}, s_{c+n_2-1}, \dots, s_2$ ), the likelihood  $L_1 = L_1(N_c|\mathbf{t})$

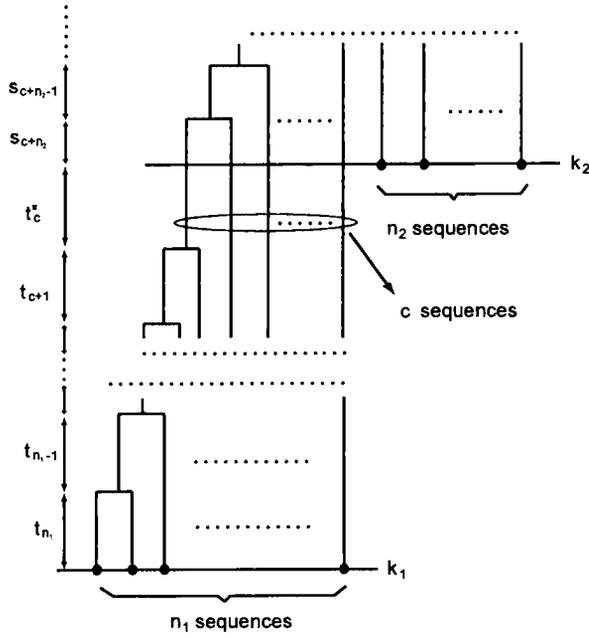


FIGURE 1.—A simple example of serial sampling where  $n_1$  sequences are sampled at time  $k_1$  and  $n_2$  sequences are sampled at an earlier time  $k_2$ . The  $n_1$  sequences coalesce into  $c$  lineages when the  $n_2$  sequences are added at time  $k_2$ .

can be expressed as a product of probability densities of time intervals

$$L_1 = \left[ \prod_{i=n_1}^{c+1} p(t_i | e(t_i), N_c) \right] \times p(t_c^*, s_{c+n_2} | e(t_c^*), k_2, N_c) \times \left[ \prod_{i=c+n_2-1}^2 p(s_i | e(s_i), N_c) \right]. \quad (5)$$

As was shown by RODRIGO and FELSENSTEIN (1999), each of the above factors in the product of  $n_1 + n_2 - 1$  terms can be simply expressed:

$$p(t_i | e(t_i), N_c) = \frac{i(i-1)}{2N_c} \exp\left(-\frac{i(i-1)}{2N_c} t_i\right)$$

$$p(t_c^*, s_{c+n_2} | e(t_c^*), k_2, N_c) = \exp\left(-\frac{c(c-1)}{2N_c} t_c^*\right) \frac{(c+n_2)(c+n_2-1)}{2N_c} \times \exp\left(-\frac{(c+n_2)(c+n_2-1)}{2N_c} s_{c+n_2}\right)$$

$$p(s_i | e(s_i), N_c) = \frac{i(i-1)}{2N_c} \exp\left(-\frac{i(i-1)}{2N_c} s_i\right). \quad (6)$$

Moreover, each factor provides some information regarding the value of  $N_c$ . We define

$$\hat{N}_c^{(i)} = \begin{cases} \operatorname{argmax}_{N_c} p(t_{i-n_2} | e(t_{i-n_2}), N_c) = \frac{(i-n_2)(i-n_2-1)t_{i-n_2}}{2}, \\ i \in \{n_1 + n_2, \dots, c + n_2 + 1\} \\ \operatorname{argmax}_{N_c} p(t_c^*, s_{c+n_2} | e(t_c^*), k_2, N_c) = \frac{c(c-1)t_c^* + (c+n_2)(c+n_2-1)s_{c+n_2}}{2}, \\ i = c + n_2 \\ \operatorname{argmax}_{N_c} p(s_i | e(s_i), N_c) = \frac{i(i-1)s_i}{2}, \\ i \in \{c + n_2 - 1, \dots, 2\}, \end{cases} \quad (7)$$

where  $\operatorname{argmax}_{\theta} f(\theta)$  represents the value of  $\theta$  that maxi-

mizes  $f(\theta)$ . Because the lengths of time intervals between coalescent events are random, the  $\hat{N}_c^{(i)}$  are random variables. By making a transformation of random variables from the time intervals in Equation 6 to the  $\hat{N}_c^{(i)}$ , the distributions of the  $\hat{N}_c^{(i)}$  are seen to be independent of the serial sampling design. Specifically, the probability density  $p(\hat{N}_c^{(i)} | N_c)$  is an exponential distribution with mean  $N_c$ . By expressing Equation 5 in terms of  $\hat{N}_c^{(i)}$  instead of in terms of time intervals, we get

$$L_1 = \frac{1}{N_c^{n_1 n_2 - 1}} \exp\left(-\frac{1}{N_c} \sum_{i=n_1+n_2}^2 \hat{N}_c^{(i)}\right). \quad (8)$$

The maximum-likelihood estimate of  $N_c$  is therefore

$$\hat{N}_c = \frac{1}{n_1 + n_2 - 1} \sum_{i=n_1+n_2}^2 \hat{N}_c^{(i)}. \quad (9)$$

Because the  $\hat{N}_c^{(i)}$  are exponential random variables with mean  $N_c$ ,

$$\operatorname{Var}(\hat{N}_c) = \frac{N_c^2}{n_1 + n_2 - 1}. \quad (10)$$

It is straightforward to show that the approaches used here to derive  $\hat{N}_c$  and  $\operatorname{Var}(\hat{N}_c)$  apply to more general serial sampling designs. Therefore, given the divergence times,  $\operatorname{Var}(\hat{N}_c)$  is affected by the total number of sampled sequences but not by other particulars of the serial sampling design.

**Pseudomaximum-likelihood approach for serial samples:** Usually, the time of sampling is measured in chronological units (day, year, etc.) and the times of internal nodes are inferred in the same units. To apply coalescent theory, chronological time should be transformed to time units in terms of generations. We can estimate the evolutionary rate  $r$  (number of substitutions per year) with serially sampled data. To make the problem simple, suppose that  $r$  is almost constant. To easily convert the mutation rate per generation into the substitution rate per year, we also assume that all mutations are selectively neutral. If the generation length  $\tau$  (days/generation) is known, 1 year can be regarded as roughly  $365/\tau$  generations and the mutation rate  $\mu$  (number of mutations/generation) can be calculated as  $r\tau/365$ . If the mutation rate  $\mu$  is known, 1 year can be regarded as roughly  $r/\mu$  generations and the generation length ( $\tau$ ) can be calculated  $365\mu/r$ . Unfortunately, we cannot determine  $\mu$  and  $\tau$  separately with only serially sampled data measured in chronological time units. One of the two must be inferred from other information. Previously estimated values of  $\mu$  include  $4.0 \times 10^{-5}$  (MANSKY 1996) and  $2.5 \times 10^{-5}$  (FU 2001). Previously estimated values of  $\tau$  include 2.6 (PERELSON *et al.* 1996), 1.8 (see reference to personal communication from A. Perelson in RODRIGO *et al.* 1999), 1.78 (FU 2001), and 1.2 (RODRIGO *et al.* 1999).

Equations 3 and 9 apply when the time intervals are

known. In real situations, they are estimated rather than being known with certainty. The uncertainty arises because the sequence data  $X$  are insufficient for exact estimation of the internal node times and topology. Therefore, uncertainty of internal node times needs to be considered when calculating the variance of the maximum-likelihood estimate for the effective population size. Suppose that we know definitely one of  $\mu$  and  $\tau$ . For  $n$  sequences (equivalently,  $n - 1$  coalescent events),

$$\begin{aligned} p(X, \mathbf{t}|N_c, \mu, \tau) &= p(\mathbf{t}|N_c, \mu, \tau)p(X|\mathbf{t}, N_c, \mu, \tau, \mathbf{t}) \\ &= p(\mathbf{t}|N_c, \mu, \tau)p(X|\mathbf{t}, r) \\ &= p(\mathbf{t}|N_c)p(X|\mathbf{t}, r). \end{aligned} \tag{11}$$

The first factor above,  $p(\mathbf{t}|N_c)$ , is itself a product of  $n - 1$  terms that each correspond to a specific coalescent event (see Equations 5 and 6). The second of the factors,  $p(X|\mathbf{t}, r)$ , is the conventional quantity that is calculated for phylogeny reconstruction with Felsenstein's pruning algorithm (FELSENSTEIN 1981). Here, we adopt a pseudo-maximum-likelihood approach (GONG and SAMANIEGO 1981). In this approach,  $\mathbf{t}$  is first estimated and then  $\mathbf{t}$  is later treated as observed in a likelihood function. Letting  $L_1(N_c|\mathbf{t}) = p(\mathbf{t}|N_c)$  and  $L_2(\mathbf{t}) = p(X|\mathbf{t}, r)$ , we get

$$\begin{aligned} l(N_c, \mathbf{t}) &= \log(p(X, \mathbf{t}|N_c, \mu, \tau)) \\ &= \log\{L_1(\mathbf{t}|N_c)L_2(X|\mathbf{t}, r)\} \\ &= l_1(N_c|\mathbf{t}) + l_2(\mathbf{t}). \end{aligned}$$

Also, we note that

$$l_1(N_c|\mathbf{t}) = \sum_{i=n}^2 l_{1,i}(N|t_i),$$

where  $l_{1,i}(N|t_i)$  corresponds to a particular of the  $n - 1$  coalescent events. With pseudomaximum likelihood, all parameters except those of interest are regarded as nuisance parameters. Replacing the nuisance parameters with consistent estimates simplifies the estimation problem. Instead of estimating  $\mathbf{t}$  and  $N_c$  simultaneously, we infer  $\tilde{\mathbf{t}}$  by maximizing  $l_2$  and then estimate  $N_c$  by maximizing  $l_1(N_c|\tilde{\mathbf{t}})$ . We use  $\widehat{N}_c(\tilde{\mathbf{t}})$  to refer to the estimated value of  $N_c$  that is obtained by maximizing  $l_1(N_c|\mathbf{t})$  over  $N_c$  for some fixed value of  $\mathbf{t}$ . Obviously, the estimates  $\tilde{\mathbf{t}}$  and  $\widehat{N}_c(\tilde{\mathbf{t}})$  are not necessarily equal to those obtained by jointly maximizing  $l(N_c, \mathbf{t})$  over  $N_c$  and  $\mathbf{t}$ .

The values of  $\tilde{\mathbf{t}}$  can be inferred with a maximum-likelihood method that incorporates serially sampled data (*e.g.*, the Tipdate software package; RAMBAUT 2000). Once the  $\tilde{\mathbf{t}}$  are obtained,  $\widehat{N}_c(\tilde{\mathbf{t}})$  can be inferred with  $(\partial/\partial N_c)l_1(N_c|\tilde{\mathbf{t}}) = 0$  (Equation 2). This approach is sensible because the maximum-likelihood estimate of  $\mathbf{t}$  is consistent (FELSENSTEIN 1988). As derived in the APPENDIX, the variance of  $\widehat{N}_c(\tilde{\mathbf{t}})$  can be approximated as

$$\begin{aligned} \text{Var}_{\mathbf{t}, X}(\widehat{N}_c(\tilde{\mathbf{t}})) &\approx \frac{N_c^2}{n - 1} \\ &+ \frac{1}{2^2(n - 1)^2} E_{\mathbf{t}} \left\{ \text{Var}_{\mathbf{t}, X} \left( \sum_{i=n}^2 i(i - 1) \tilde{t}_i \right) \right\}. \end{aligned} \tag{12}$$

The right side of Equation 12 is the sum of two terms where the first term,  $N_c^2/(n - 1)$ , is the variance of  $N_c$  when the divergence times are known (see Equation 10). For a given value of  $N_c$ , the divergence times are random and will stochastically vary among coalescent realizations. It is this randomness due to genetic sampling that is captured by  $N_c^2/(n - 1)$ . We can approximate this genetic sampling term with  $\widehat{N}_c(\tilde{\mathbf{t}})^2/(n - 1)$ . The second of the two terms in Equation 12 arises from statistical sampling. With finite sequence lengths, divergence times cannot be perfectly estimated and this second term represents the uncertainty of  $\widehat{N}_c(\tilde{\mathbf{t}})$  that results. The impact of this randomness due to statistical sampling can be handled either via a numerical approximation of the second term of Equation 12 or by simulation (see below).

## METHODS

**Sequence sampling from asymptomatic period:** We analyzed the C2-V5 region of HIV-1 *env* sequences from nine patients. These data were described by SHANKARAPPA *et al.* (1999) who reported on viral sequence data isolated from peripheral blood mononuclear cells (PBMCs) and on a parallel sequence data set isolated from the plasma of each patient. Because there seems to be little difference in patterns of evolution for the PBMC data and the plasma data, we consider only the PBMC data here. The single exception is that we analyzed the plasma-derived sequence data from the patient referred to as p11 by SHANKARAPPA *et al.* (1999). This exception was made because PBMC data from patient p11 were unavailable.

A striking pattern among the nine HIV-1 sequence data sets is that divergence, defined as the evolutionary distance from an early founder sequence, is linearly increasing during the first portion of the infection and then tends to stabilize at a later stage (SHANKARAPPA *et al.* 1999). This linear phase is consistent with an almost constant rate of molecular evolution. The time span of this linearly increasing portion varies among patients. For each patient, we obtained a rough estimate of the time point at which the linear portion of increasing divergence ended (Table 1). Our analyses are based only on viral sequences isolated from a patient during the estimated linear period of divergence.

The sequence set from each patient was aligned via the default options of ClustalW Ver.1.07 (THOMPSON *et al.* 1994). Next, alignments were manually edited. Alignment columns with gaps were removed from the data prior to subsequent analysis. To avoid the elimination of too many columns, cases where only one or two sequences exhibited a deletion at a site relative to other sequences were treated by removing the sequences with the deletion rather than by removing the alignment columns exhibiting the deletion. This procedure resulted in only a relatively small number of sequences being

eliminated from the total number of sequences in the nine data sets (see Table 1).

Phylogenies were then inferred from aligned sequences via maximum-likelihood. Tree reconstruction was performed in three steps. First, a consensus of all sequences with the earliest isolation date from a particular patient was constructed. This consensus sequence was added to the data and was treated as an outgroup. Then, the neighbor-joining method (SAITOU and NEI 1987) was applied to a distance matrix with entries calculated according to the Hasegawa-Kishino-Yano (HKY) model of nucleotide substitution (HASEGAWA *et al.* 1985). With the neighbor-joining tree as an initial guess, local rearrangements as implemented in the Molphy software package (ADACHI and HASEGAWA 1996) were made to search the space of tree topologies with the maximum-likelihood criterion. Last, on the basis of the rooted topology and the HKY model of nucleotide substitution, times of internal nodes and evolutionary rates were simultaneously estimated with the Tiptdate Version 1.1 software (RAMBAUT 2000).

**Estimation of evolutionary rate and effective population size:** With Equation 12, knowledge of  $\text{Var}(\hat{\mathbf{t}})$  is necessary to determine  $\text{Var}(\hat{N}_e(\hat{\mathbf{t}}))$ . According to asymptotic theory, the distribution of maximum-likelihood estimates can be approximated by a multivariate normal distribution. Here, the asymptotic approximation improves as sequence length increases. We approximated the variance of the estimated times of internal nodes with the inverse of a numerically estimated Fisher information matrix. By sampling from the resulting multivariate normal distributions, we simulated new sets of estimated times and then estimated effective population sizes from these simulated times. To estimate  $N_e$  in this fashion, it was necessary to assume a specific value for either  $\tau$  or  $\mu$  and then apply a generalization of Equation 5. For each patient and with each value of  $\tau$  or  $\mu$ , variance estimates were based on 20,000 samples from the appropriate multivariate normal distribution. This computational procedure yields an approximation of the second of the two terms in Equation 12 that are needed to estimate  $\text{Var}(\hat{N}_e(\hat{\mathbf{t}}))$ .

When sampling node times from a multivariate normal distribution, it is possible to observe samples that are inconsistent with the serial sampling design. For example, in Equation 5, if the set of random numbers shows  $> n_1 - 1$  coalescent events after time  $k_2$ , the likelihood function is not defined. In cases where the likelihood function was not defined, the set of sampled node times was discarded.

As an alternative to our strategy of sampling from a multivariate normal distribution to estimate the uncertainty in  $\hat{N}_e(\hat{\mathbf{t}})$  due to uncertainty in node times, a bootstrap approach could be used. The idea would be to sample sites with replacement. For each bootstrap sample, effective population size could be estimated and the

variance of these estimates among bootstrap samples could be determined. As with our multivariate sampling strategy, this bootstrap approach would only estimate the uncertainty in  $\hat{N}_e(\hat{\mathbf{t}})$  that is due to uncertainty in node times. This contribution from node time uncertainty would be added to  $\hat{N}_e(\hat{\mathbf{t}})^2/(n-1)$  to approximate  $\text{Var}(\hat{N}_e(\hat{\mathbf{t}}))$ . Advantages of our multivariate normal sampling strategy over the bootstrap alternative are simplicity and computational feasibility.

## RESULTS

**Correlation between  $N_e$  and  $r$ :** Table 1 shows estimated evolutionary rates and effective population sizes. There appears to be large variation among patients in evolutionary rates. This variation can be explained by any of three scenarios: (1) mutation rates are the same among patients, but generation lengths differ; (2) generation lengths are the same, but mutation rates differ; (3) both mutation rates and generation lengths differ. Because estimates from serially sampled data of the mutation rate and generation length are confounded when the sampling times are measured in chronological units, we considered only the first two of the three potential causes for evolutionary rate variation.

When an identical mutation rate among patients was assumed, we used the value  $2.5 \times 10^{-5}$  (mutations per generation) that was estimated by FU (2001). Contrary to the frequently assumed mutation rate of  $4.0 \times 10^{-5}$  (MANSKY 1996), this lower value was obtained by excluding insertion, deletion, and frameshift mutations. We chose the value  $2.5 \times 10^{-5}$  (mutations per generation) because we too excluded insertions, deletions, and frameshifts.

Estimates of generation time  $\tau$  have been obtained via a wide variety of approaches and these estimates themselves have widely varied. Estimates of  $\tau$  range from 1.2 days (RODRIGO *et al.* 1999) to 2.6 days (PERELSON *et al.* 1996). It seems plausible that there is much variation in viral generation length among patients. Our generation length estimates, on the basis of the simple relation  $\tau = 365\mu/r$ , ranged among patients from 0.76 to 2.64 days with a mean of 1.47 days. This mean value is close to values of 1.78 days (FU 2001) and 1.8 days (RODRIGO *et al.* 1999) that have been previously obtained. These two previous values were obtained via different estimation methods and with different data than analyzed here. Using each estimated generation length, we transformed the chronological units of internal node times to generation time units and then estimated effective population sizes by maximizing the likelihood function (a generalized form of Equation 5).

Figure 2 shows the relation between evolutionary rate and effective population size for two scenarios: constant generation length and constant mutation rate. Assuming a viral generation length of 1.47 days in each patient, we observed a clearly negative correlation between evolutionary rate and effective population size. A negative

**TABLE 1**  
**Estimates of viral evolution parameters from nine different populations**

Patient index	Linear time span (months) <sup>a</sup>	No. of sequences <sup>b</sup>	$\hat{r}$ ( $\times 10^{-3}$ )	$\hat{\tau}$ (days)	$\widehat{N}_e(\hat{\tau})^c$ ( $\mu = 2.5 \times 10^{-5}$ )	$\hat{\mu}$ ( $\times 10^{-5}$ )	$\widehat{N}_e(\hat{\tau})^d$ ( $\tau = 1.47$ )
p1	77	77 (4)	4.73 (0.543)	1.93	3922.8 (692.7)	1.91	5141.7 (907.3)
p2	85	91 (4)	4.83 (0.728)	1.89	4012.1 (724.9)	1.95	5151.8 (930.3)
p3	67	62 (2)	10.9 (1.28)	0.84	1527.7 (258.1)	4.39	869.6 (146.9)
p5	81	151 (0)	4.49 (0.337)	2.03	5480.8 (671.8)	1.81	7553.5 (926.0)
p6	42	56 (1)	12.4 (1.51)	0.73	1789.7 (303.7)	5.02	891.6 (151.3)
p7	74	86 (0)	6.48 (0.620)	1.41	4856.0 (626.4)	2.62	4637.5 (598.3)
p8	81	99 (2)	8.00 (0.732)	1.14	4491.4 (566.7)	3.23	3479.0 (438.8)
p9	122	81 (2)	3.75 (0.374)	2.43	5409.3 (793.1)	1.52	8923.6 (1308.3)
p11	100	39 (0)	10.6 (0.917)	0.86	2473.1 (456.6)	4.29	1441.2 (266.1)
Mean			7.35	1.47			

Entries in parentheses that follow parameter estimates are estimated standard deviations.

<sup>a</sup> Time period for which linear sequence divergence was estimated from the number provided at <http://ubik.microbiol.washington.edu/HIV/evolution1/>.

<sup>b</sup> The total number of sequences analyzed for each patient. The numbers in parentheses are the numbers of sequences that were excluded from each data set due to gaps that are present at alignment columns in the excluded sequences but that are absent in the analyzed sequences;  $\hat{r}$  is the maximum-likelihood estimate of the evolutionary rate (the number of substitutions/site/year).

<sup>c</sup> Estimated effective population size with the assumption  $\mu = 2.5 \times 10^{-5}$  changes/site/generation;  $\hat{\tau}$  is the estimated generation length with the assumption that  $\mu = 2.5 \times 10^{-5}$  and the simple relation  $\tau = 365\mu/r$ ;  $\hat{\mu}$  is the estimated mutation rate per generation with the assumption that  $\tau = 1.47$  and the simple relation  $\mu = r\tau/365$ .

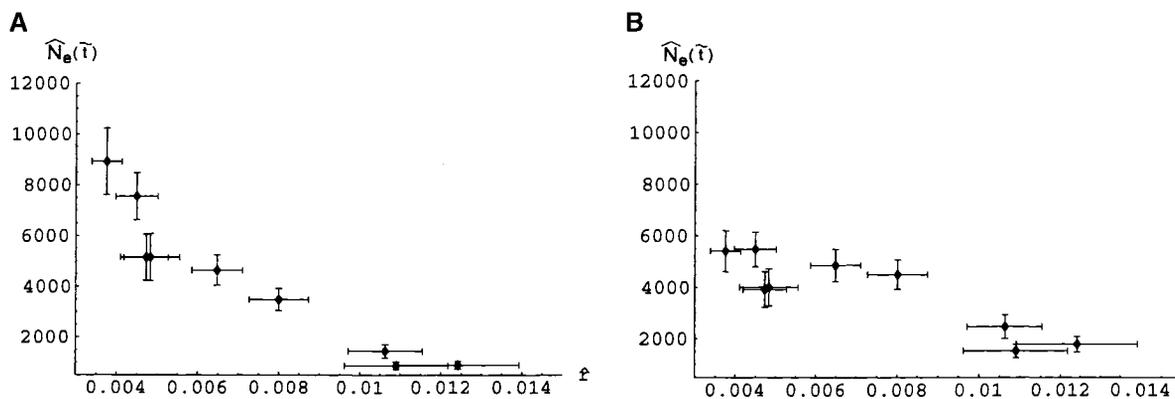
<sup>d</sup> Estimated effective population size with the assumption  $\tau = 1.47$  days per generation.

correlation was also observed when the mutation rate was assumed to be shared among viruses in different patients, but here the correlation was weaker.

**Bias and precision of pseudomaximum-likelihood estimates:** Via simulation, we investigated the effect on effective population size estimates under various circumstances (Table 2). Our simulation scenarios differed according to the number of sequences (5, 10, 15, 20, or 25) that were sampled at each of four sampling times. The interval between sampling times was 1 year, which corresponds to 248.3 generations under the assumption that the generation length ( $\tau$ ) is 1.47 days. The true value of  $N_e$  was 4000, because this is close to the mean estimated effective population size (4232.2) of the nine patients when  $\tau$  was set equal to 1.47 days.

Sequence lengths were simulated to be 600 bases, because the lengths of the aligned sequences from the nine patients ranged from 544 to 600 bases. The evolutionary rate was set to 0.00735 substitutions/year because this is the mean of the estimates from the nine patients. This evolutionary rate corresponds to a mutation rate of  $2.97 \times 10^{-5}$  mutations/generation.

Node times and the tree topology were simulated according to the coalescent process. In the simulations, one additional sequence was added so that the root of the ingroup genealogy could be inferred. This additional sequence had a sampling time that was set to chronologically precede the root of the ingroup by one generation. The time at which the lineage leading to the ingroup joined with the lineage leading to the out-



**FIGURE 2.**—A negative correlation between the evolutionary rate per year  $\hat{r}$  and the effective population size  $\widehat{N}_e(\hat{\tau})$ . (A) Assuming a generation length of 1.47 days. (B) Assuming a mutation rate of  $2.5 \times 10^{-5}$  substitutions/generation.

TABLE 2  
Mean and standard error of  $\widehat{N}_e$

		No. of sequences				
		21	41	61	81	101
$\tau = 1.47$	Case 1 <sup>a</sup>	3991.2 ± 846.7 (4000.0 ± 894.4)	4061.3 ± 663.1 (4000.0 ± 632.5)	3967.9 ± 503.4 (4000.0 ± 516.4)	4017.8 ± 409.9 (4000.0 ± 447.2)	4077.1 ± 443.6 (4000.0 ± 400.0)
	Case 2 <sup>b</sup>	4109.4 ± 1179.4	4241.6 ± 963.6	4059.7 ± 763.6	4101.9 ± 736.3	4086.2 ± 668.7
	Case 3 <sup>c</sup>	4130.1 ± 1178.2	4283.8 ± 937.7	4155.7 ± 892.0	4206.7 ± 802.2	4268.1 ± 878.1
$\mu = 2.97 \times 10^{-5}$	Case 4 <sup>b</sup>	3969.2 ± 897.1	4100.1 ± 685.7	3977.3 ± 550.9	4006.0 ± 440.7	4056.3 ± 510.5
	Case 5 <sup>c</sup>	3979.2 ± 889.7	4109.2 ± 673.8	4031.8 ± 593.3	4035.6 ± 447.9	4130.0 ± 547.8

A total of 21, 41, 61, 81, and 101 sequences are sampled at five sampling times as described in the text. Simulations were performed with 0.00735 substitutions per position per year, a generation time of  $\tau = 1.47$  days, which corresponds to a mutation rate  $\mu = 2.97 \times 10^{-5}$ , and sequences of length 600 bases. Each entry represents the mean and standard error of  $\widehat{N}_e$  from 100 simulated data sets.

<sup>a</sup>In case 1, the tree topology and coalescent times are assumed known. Estimated standard deviations from Equation 10 are shown in parentheses.

<sup>b</sup>In case 2 and case 4, the tree topology is assumed known, but coalescent times are estimated with maximum likelihood and with  $\tau = 1.47$  (case 2) or  $\mu = 2.97 \times 10^{-5}$  (case 4).

<sup>c</sup>In case 3 and case 5, the tree topology is estimated with neighbor-joining plus local rearrangements and coalescent times are then estimated with maximum likelihood. Case 3 sets  $\tau = 1.47$  and case 5 sets  $\mu = 2.97 \times 10^{-5}$ .

group was then randomly determined according to the coalescent process. The Jukes-Cantor model (JUKES and CANTOR 1969) was then used to randomly simulate sequences on the tree relating the ingroup sequences and the outgroup sequence.

After generating the sequences, effective population size was inferred from these data in five different situations. In three cases (referred to as cases 1–3),  $\tau = 1.47$  was assumed. For case 1, the true tree topology and coalescent times were both treated as known. For case 2, the true tree topology was treated as known but the coalescent times were estimated with maximum likelihood. For case 3, the tree topology was estimated with a combination of neighbor-joining and local rearrangements and then the coalescent times were estimated with maximum likelihood. In two situations (referred to as cases 4 and 5),  $\mu = 2.97 \times 10^{-5}$  was assumed. As with case 2, case 4 involved treating the true tree topology as known but estimating coalescent times. As with case 3, case 5 involved reconstructing the tree topology via a combination of neighbor-joining and local rearrangements and then estimating the coalescent times on this topology with maximum likelihood.

For each entry in Table 2, the estimated mean and standard error of  $\widehat{N}_e$  are based upon results from 100 simulated data sets. As expected due to the second term of Equation 12, case 1 exhibits lower standard errors than do the other cases. Another general and expected trend seems to be that the standard error for estimating  $N_e$  decreases as the number of sequences increases. This decrease can be explained by the first term in Equation 12.

As the number of sequences gets larger, the difference grows between the standard errors for the cases where the true tree topology was and was not treated

as known. This indicates that topological uncertainty is more important for data sets with large numbers of sequences. But the impact of topological uncertainty on the standard errors of  $N_e$  seems to be relatively small when compared with the impact of uncertainty of coalescence times.

Results from cases 3 and 5 indicate that the approach introduced here to estimate  $N_e$  has upward bias. The first step of our pseudomaximum-likelihood approach is to estimate  $\mathbf{t}$  by maximizing  $P(\mathbf{X}|\mathbf{t})$ . These estimates of  $\mathbf{t}$  will be asymptotically unbiased and will asymptotically approach a multivariate normal distribution. However, when the sequence length is not long enough, the sampling distribution of the estimates of  $\mathbf{t}$  may greatly differ from a multivariate normal. When there are many sequences, coalescent time intervals are relatively short near the tips and the estimated internal node times are constrained by the times of tips. Thus, the deviation between the sampling distribution of the estimates of  $\mathbf{t}$  and a multivariate normal distribution may become particularly important when there are many sequences. As a result, the sampling distribution of estimates of  $\mathbf{t}$  is truncated and the expected values of  $\hat{\mathbf{t}}$  will be larger than their true values. This could explain the upward bias that we observe for estimating  $N_e$  even when the true tree topology is used. The effect of the truncation of coalescent times was also noted by KUHNER *et al.* (1998).

The estimation properties of the pseudomaximum-likelihood technique are also affected by the fact that one step is to reconstruct a bifurcating topology but this reconstruction step is not sufficiently influenced by the times at which sequences are sampled. As an illustration, consider a data set where sequence A and sequence B

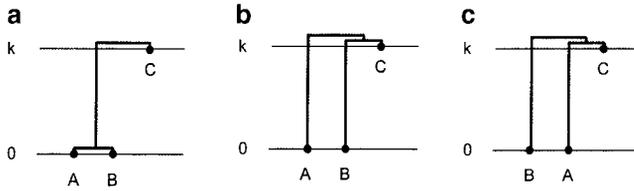


FIGURE 3.—The three possible rooted topologies when sequence C is sampled  $k$  generations earlier than sequence A and sequence B.

were sampled at the same time but sequence C was sampled  $k$  generations earlier. It is possible, especially if the mutation rate is small, that these three sequences are identical because no substitutions occur following their common ancestor. In this case, our topology reconstruction procedure would consider each of the three rooted topologies shown in Figure 3 to be equally likely and one of these three topologies would be arbitrarily selected. If the topology of Figure 3a were selected, the resulting estimate of  $N_c$  for these data would be 0. In contrast, selection of the topologies in Figure 3b or 3c would yield an estimate for  $N_c$  of  $k$ . The necessity to arbitrarily select one bifurcating topology due to branch length estimates being zero becomes more frequent when the number of sequences is large and the impact of these arbitrary selections is likely to be more significant when the number of sequence sampling times is large. The effect of phylogenetic reconstruction on estimating population parameters has also been investigated by Fu (1994), who noted a downward bias for inferring  $\theta = 4N_c\mu$  when the number of sequences is large and when using UPGMA trees without correcting for multiple hits.

The number of generations represented by each branch on the genealogy controls the estimate of effective population size. Our approach is to estimate these numbers of generations from sequence data. Because sequence data can be used to directly estimate the expected number of sequence changes on each branch of the genealogy, the number of generations represented by each branch can be straightforwardly estimated from the numbers of sequence changes on branches and from a known rate  $\mu$  of sequence change per generation. When the chronological time per generation  $\tau$  is assumed known and  $\mu$  is assumed unknown, the rate of sequence change per chronological time unit must also be estimated to permit estimation of the number of generations represented by each branch on the tree. Because rate of sequence change per chronological time unit is subject to estimation uncertainty, the standard errors for estimating  $N_c$  are smaller for cases 4 and 5 than for cases 2 and 3.

## DISCUSSION

Recently, a Bayesian approach that incorporates the uncertainty in the genealogical structure has been devel-

oped (DRUMMOND *et al.* 2002). This Bayesian approach simultaneously estimates the mutation rate, population size, and tree topology. In contrast, our frequentist approach neglects the uncertainty of the estimated tree topology to reduce the required amount of calculation. The accounting for topological uncertainty is a distinct advantage of the Bayesian approach over that presented here.

Although there are multiple Bayesian and frequentist options for model checking, a potential advantage of the approach here over the full Bayesian strategy is ease of model checking. The first step of our procedure begins with reconstructing a genealogy and the structure of the reconstructed genealogy can then be immediately inspected. Inspection of the genealogy permits both formal and informal checks of whether the assumed model of population history is plausible. For example, the structure of the reconstructed genealogy can give an immediate indication as to whether a model of population size increase over time is justified. Because we believe that models are apt to be the weakest point of evolutionary analysis, whereas the method of estimation based upon the model is an important but secondary concern, ease of model checking should not be dismissed when evaluating a procedure.

In our approach, the estimated times of internal nodes are nuisance parameters rather than being of major interest. Our pseudomaximum-likelihood method does not account for uncertainty of these nuisance parameters. To account for this uncertainty, an empirical Bayes approach could be adopted. In empirical Bayes approaches, the marginal-likelihood function is obtained by integrating the conditional distribution over the space of nuisance parameters (O'HAGAN 1996).

In our case, the marginal-likelihood function of  $N_c$  given the sequence data  $X$  could be obtained by integrating over the times  $\mathbf{t}$ ,

$$P(X|N_c) = \int P(X|\mathbf{t})P(\mathbf{t}|N_c) dt.$$

This integration is not analytically simple but a numerical approximation is available (CONGDON 2001). The numerical technique involves sampling times from a multivariate normal distribution that approximates  $P(X|\mathbf{t})$  up to a constant of proportionality. The mean of the multivariate normal distribution is the value of  $\mathbf{t}$  that maximizes  $P(X|\mathbf{t})$ . The covariance structure of the multivariate normal distribution is approximated by the inverse of the Fisher information matrix. The  $i$ th sample from the multivariate normal distribution is denoted  $\mathbf{t}^{(i)}$  and the total number of these samples is  $n$ . Then,  $P(X|N_c)$  should be well approximated by

$$P(X|N_c) \approx \frac{1}{n} \sum_{i=1}^n P(\mathbf{t}^{(i)}|N_c). \quad (13)$$

The marginal likelihood can be calculated with Equation 13 for each  $N_c$  and the maximum-likelihood estimate of  $N_c$  can be found by maximizing  $P(X|N_c)$ .

To facilitate data analysis, we combine the constant rate assumption of the molecular clock hypothesis and Kingman's  $n$ -coalescent (KINGMAN 1982a,b). With a constant rate, divergence time estimation is straightforward. With a simple coalescent process, estimation of effective population size from known divergence times is not difficult. However, neither the clock nor the simple coalescent assumptions are technically correct. Both are violated by realistic forms of natural selection. For example, there is ample evidence that positive selection operates in HIV-1 genes (NIELSEN and YANG 1998). It is unclear whether inaccuracy of the clock and coalescent assumptions would have a large effect on the results of this analysis. Some evolutionary scenarios that invoke selection have been shown to have little effect on the shape of gene genealogies relating contemporaneous data (GOLDING 1997; NEUHAUSER and KRONE 1997; PRZEWORSKI *et al.* 1999).

Kingman's  $n$ -coalescent also requires a constant population size. Because clinical measurements such as viral load counts do not change much during the asymptomatic period (reviewed in VISCIDI 1999), the constant population size assumption seems reasonable. If population size is increasing over time, our estimates of it are expected to be intermediate between the early and later population levels. For a high rate of increase, the estimates are expected to be closer to the early population sizes than the later sizes.

Regarding the molecular clock, we focus here on analysis of sequences that were isolated during the approximately linear phase of sequence divergence that is characteristic of the early asymptomatic period in HIV-1 infections. This linear pattern is what would be seen with neutral evolution and with a molecular clock. Because the linear pattern does not extend into later portions of the asymptomatic period, we did not analyze sequences isolated during the later portions. Therefore, instead of using a highly realistic and overly complicated model to analyze the entire data set, we opted here to investigate only the portions of the data that seemed relatively compatible with the simple model of a molecular clock.

Searching for associations between effective population size and measurements that reflect the physiological condition of a patient may also be fruitful. As a potential marker of disease progression, we examined the time points at which CD4+ T cell counts of each patient decreased to 200/ $\mu$ l (see SHANKARAPPA *et al.* 1999). However, we were unable to detect a relationship between these time points and the estimated effective viral population size within a patient. Nonetheless, informative associations between effective population size and other covariates may be detected in more thorough future analyses. These associations have the potential to illuminate the factors that are related to viral adaptation within a patient.

In previous work (SEO *et al.* 2002), we noted that it is desirable to disperse serial sampling times as much

as possible to accurately estimate evolutionary rates and times. The same strategy applies in the estimation of effective population size, because differences in  $\text{Var}(\hat{N}_e(\bar{t}))$  among serial sampling designs that share a common number of sampled sequences will be attributable to differences in  $\text{Var}(\bar{t})$  (see Equation 12). Designs with smaller  $\text{Var}(\bar{t})$  are expected to yield smaller  $\text{Var}(\hat{N}_e(\bar{t}))$ .

In this study, we assumed either that viral mutation rate or viral generation length is constant among patients. In reality, both mutation rate and generation time probably vary to some extent among patients. To investigate this simultaneous variation, more data and especially a more sophisticated model for viral evolution may be warranted.

The negative correlation between evolutionary rate and effective population size (Figure 2) is noteworthy. We cannot formally exclude the possibility that this negative correlation is an artifact of our pseudomaximum-likelihood procedure because the fact that viral generation length and evolutionary rate may both vary among patients makes this exclusion difficult. Because differences in effective population size and evolutionary rate estimates among patients exceed the uncertainty within patients, the possibility that this negative correlation is simply an artifact seems unlikely.

In population genetics, this negative correlation is predicted by both the slightly deleterious mutation model (OHTA 1987) and the nearly neutral mutation model (TACHIDA 1991). In a strictly neutral model, the rate at which mutants become fixed in the population does not depend on the population size but only on the mutation rate. In contrast, in the slightly deleterious or nearly neutral models, the rate of fixation of advantageous mutants depends on both the population size and selective coefficients. In the slightly deleterious model, it is assumed that most mutations are neutral or slightly deleterious and the mean of the distribution of selective coefficient(s) is  $<0$ . A negative correlation between evolutionary rate and effective population size is then predicted. As noted by OVERBAUGH and BANGHAM (2001), if the error rate of reverse transcriptase is high, deleterious mutations would be frequent and negative selection is likely to be a dominant force in viral evolution. In the nearly neutral model, the selection coefficient of mutations has a distribution with variance  $\sigma^2$  and with a mean that can be either negative or positive. It was shown by simulation that the substitution (note that "substitution" here does not refer solely to variants that have been fixed in the population) rate is negatively correlated with  $N_e\sigma$  (TACHIDA 1991). This might result in a negative correlation between evolutionary rate and effective population size for a fixed  $\sigma$ .

It is not clear whether the above classical views of population genetics explain our finding of a negative correlation. There are potentially many population genetic scenarios that could result in this correlation. Also, the correlation could be attributable to the variation of

immune response among patients. If a patient's immune system is strong, viral sequences could experience strong positive selection and this could lead to an increased evolutionary rate and a decreased effective population size. We believe that further characterization of this apparent correlation and its possible sources would facilitate our understanding of the mechanism of viral adaptation within hosts.

We thank two anonymous reviewers for their suggestions. T.-K.S. and H.K. were supported by the Japan Society for the Promotion of Science (JSPS) United States-Japan research collaboration program 12554037; T.-K.S. by the Japanese government scholarship program for foreign students; M.H. and H.K. by grant BSAR-497 from the Ministry of Education, Culture, Sports, Science, and Technology (MECSST); J.L.T. and H.K. by grant 13308013 of MECSST; and J.L.T. by National Science Foundation grants DBI-0077503 and INT-990934.

#### LITERATURE CITED

- ADACHI, J., and M. HASEGAWA, 1996 Programs for molecular phylogenetics based on maximum likelihood, Molphy Version 2.3.
- CONGDON, P., 2001 *Bayesian Statistical Modelling*, pp. 472–474. John Wiley & Sons, New York.
- 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., G. K. HICHOLLS, A. G. RODRIGO and W. SOLOMON, 2002 Estimating mutation rate, population history, substitution model and genealogy simultaneously from temporally spaced sequence data. *Genetics* (in press).
- FELSENSTEIN, J., 1981 Evolutionary trees from DNA sequences: a maximum likelihood approach. *J. Mol. Evol.* **17**: 368–376.
- FELSENSTEIN, J., 1988 Phylogenies from molecular sequences: inference and reliability. *Annu. Rev. Genet.* **22**: 521–565.
- 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.
- FELSENSTEIN, J., M. K. KUHNER, J. YAMATO and P. BEERLI, 1999 Likelihoods on coalescents: a Monte Carlo sampling approach to inferring parameters from population samples of molecular data, pp. 163–185 in *Statistics in Molecular Biology and Genetics* (IMS Lecture Series, Vol. 33), edited by F. SEILLER-MOISEWITSCH. Institute of Mathematical Statistics and American Mathematical Society, Hayward, CA.
- FU, Y.-X., 1994 A phylogenetic estimator of effective population size or mutation rate. *Genetics* **136**: 685–692.
- FU, Y.-X., 2001 Estimating mutation rate and generation time from longitudinal samples of DNA sequences. *Mol. Biol. Evol.* **18**: 620–626.
- GOLDING, G. B., 1997 The effect of purifying selection on genealogies, pp. 271–285 in *Progress in Population Genetics and Human Evolution* (IMA Volumes in Mathematics and Its Applications, Vol. 87), edited by P. DONNELLY and S. TAVARÉ. Springer-Verlag, New York.
- GONG, G., and F. J. SAMANIEGO, 1981 Pseudo maximum likelihood estimation: theory and applications. *Ann. Stat.* **9**: 861–869.
- HASEGAWA, M., H. KISHINO and T. YANO, 1985 Data of the humanape splitting by a molecular clock of mitochondrial DNA. *J. Mol. Evol.* **22**: 160–174.
- HUDSON, R. R., 1991 Gene genealogies and the coalescent process. *Oxf. Surv. Evol. Biol.* **7**: 1–44.
- JUKES, T. H., and C. R. CANTOR, 1969 Evolution of protein molecules, pp. 21–132 in *Mammalian Protein Metabolism*, edited by H. N. MUNRO. Academic Press, New York.
- KINGMAN, J. F. C., 1982a The coalescent. *Stoch. Proc. Appl.* **13**: 235–248.
- KINGMAN, J. F. C., 1982b On the genealogy of large populations. *J. Appl. Probab.* **19A**: 27–43.
- KUHNER, M. K., J. YAMATO and J. FELSENSTEIN, 1998 Maximum likelihood estimation of population growth rates based on the coalescent. *Genetics* **149**: 429–434.
- MANSKY, L. M., 1996 Forward mutation rate of human immunodeficiency virus type 1 in a T lymphoid cell line. *AIDS Res. Hum. Retroviruses* **12**: 307–314.
- NEUHAUSER, C., and S. M. KRONE, 1997 The genealogy of samples in models with selection. *Genetics* **145**: 519–534.
- NIELSEN, R., and Z. YANG, 1998 Likelihood models for detecting positively selected amino acid sites and applications to the HIV-1 envelope gene. *Genetics* **148**: 929–936.
- NOWAK, M. A., and R. M. MAY, 1992 Coexistence and competition in HIV infections. *J. Theor. Biol.* **159**: 329–342.
- NOWAK, M. A., R. M. MAY and R. M. ANDERSON, 1990 The evolutionary dynamics of HIV-1 quasi species and the development of immunodeficiency disease. *AIDS* **4**: 1095–1103.
- O'HAGAN, A., 1996 *Kendall's Advanced Theory of Statistics, Vol. 2B: Bayesian Inference*, pp. 131–132. John Wiley & Sons, New York.
- OHTA, T., 1987 Very slightly deleterious mutations and the molecular clock. *J. Mol. Evol.* **26**: 1–6.
- OVERBAUGH, J., and C. R. M. BANGHAM, 2001 Selective forces and constraints on retroviral sequence variation. *Science* **292**: 1106–1109.
- PERELSON, A. S., A. U. NEUMANN, M. MARKOWITZ, J. M. LEONARD and D. D. HO, 1996 HIV-1 dynamics in vivo: virion clearance rate, infected cell life-span, and viral generation time. *Science* **271**: 1582–1586.
- PRZEWSKI, M., B. CHARLESWORTH and J. D. WALL, 1999 Genealogies and weak purifying selection. *Mol. Biol. Evol.* **16**: 246–252.
- RAMBAUT, A., 2000 Estimating the rate of molecular evolution: incorporating non-contemporaneous sequences into maximum likelihood phylogenies. *Bioinformatics* **16**: 395–399.
- RODRIGO, A. G., and J. FELSENSTEIN, 1999 Coalescent approaches to HIV population genetics, pp. 233–272 in *The Evolution of HIV*, edited by K. A. CRANDALL. Johns Hopkins University Press, Baltimore.
- RODRIGO, A. G., E. G. SHAPER, E. L. DELWART, A. K. IVERSEN, M. V. GALLO *et al.* 1999 Coalescent estimates of HIV-1 generation time in vivo. *Proc. Natl. Acad. Sci. USA* **96**: 2187–2191.
- 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 A viral sampling design for testing the molecular clock and for estimating evolutionary rates and divergence times. *Bioinformatics* **18**: 115–123.
- SHANKARAPPA, R., J. B. MARGOLICK, S. J. GANGE, A. G. RODRIGO, D. UPCHURCH *et al.* 1999 Consistent viral evolutionary changes associated with the progression of human immunodeficiency virus type 1 infection. *J. Virol.* **73**: 10489–10502.
- TACHIDA, H., 1991 A study on a nearly neutral mutation model in finite populations. *Genetics* **128**: 183–192.
- THOMPSON, J. D., D. G. HIGGINS and T. J. GIBSON, 1994 Clustal W: 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.
- VISCIDI, R. P., 1999 HIV evolution and disease progression via longitudinal studies, pp. 346–389 in *The Evolution of HIV*, edited by K. A. CRANDALL. Johns Hopkins University Press, Baltimore.

Communicating editor: G. B. GOLDING

#### APPENDIX

Using the following equation,

$$\frac{\partial}{\partial N_e} l(N_e, \mathbf{t}) = 0,$$

we can estimate  $\hat{N}_e(\hat{\mathbf{t}})$ . With a Taylor expansion of  $(\partial/\partial N_e)l(\hat{N}_e(\hat{\mathbf{t}}), \mathbf{t})$  around  $N_e$ , we get

$$\begin{aligned}
 0 &= \frac{\partial}{\partial N_e} l(\widehat{N}_e(\mathbf{t}), \mathbf{t}) \approx \frac{\partial}{\partial N_e} l(N_e, \mathbf{t}) + \frac{\partial^2}{\partial N_e^2} l(N_e, \mathbf{t}) (\widehat{N}_e(\mathbf{t}) - N_e) \\
 &\approx \frac{\partial}{\partial N_e} l(N_e, \mathbf{t}) - \frac{n-1}{N_e^2} (\widehat{N}_e(\mathbf{t}) - N_e), \tag{A1}
 \end{aligned}$$

because

$$\begin{aligned}
 \frac{\partial^2}{\partial N_e^2} l(N_e, \mathbf{t}) &= \sum_{i=n}^2 \frac{\partial^2}{\partial N_e^2} l_{i,i}(N_e | t_i) \\
 &\approx (n-1) E_i \left[ \frac{\partial^2}{\partial N_e^2} l_{i,i}(N_e | t_i) \right] \\
 &= -\frac{n-1}{N_e^2}.
 \end{aligned}$$

Equation A1 leads to

$$\widehat{N}_e(\mathbf{t}) - N_e \approx \frac{N_e^2}{n-1} \frac{\partial}{\partial N_e} l(N_e, \mathbf{t})$$

and

$$\frac{\partial}{\partial \mathbf{t}^T} \widehat{N}_e(\mathbf{t}) \approx \frac{N_e^2}{n-1} \frac{\partial^2}{\partial N_e \partial \mathbf{t}^T} l(N_e, \mathbf{t}).$$

A Taylor expansion of  $\widehat{N}_e(\tilde{\mathbf{t}})$  around  $\mathbf{t}$  yields

$$\begin{aligned}
 \widehat{N}_e(\tilde{\mathbf{t}}) &\approx \widehat{N}_e(\mathbf{t}) + \frac{\partial}{\partial \mathbf{t}^T} \widehat{N}_e(\mathbf{t})(\tilde{\mathbf{t}} - \mathbf{t}) \\
 &\approx \widehat{N}_e(\mathbf{t}) + \frac{N_e^2}{n-1} \frac{\partial^2}{\partial N_e \partial \mathbf{t}^T} l(N_e, \mathbf{t})(\tilde{\mathbf{t}} - \mathbf{t}) \\
 &= \widehat{N}_e(\mathbf{t}) + \frac{1}{2(n-1)} \sum_{i=n}^2 i(i-1)(\tilde{t}_i - t_i).
 \end{aligned}$$

Therefore,

$$\begin{aligned}
 \text{Var}_{i,\mathbf{x}}(\widehat{N}_e(\tilde{\mathbf{t}})) &= \text{Var}_i\{E_{\mathbf{x}}(\widehat{N}_e(\tilde{\mathbf{t}})|\mathbf{t})\} \\
 &\quad + E_i\{\text{Var}_{\mathbf{x}}(\widehat{N}_e(\tilde{\mathbf{t}})|\mathbf{t})\} \\
 &\approx \text{Var}_i(\widehat{N}_e(\mathbf{t})) \\
 &\quad + \left(\frac{N_e^2}{n-1}\right)^2 \\
 &\quad \times E_i \left\{ \frac{\partial^2}{\partial N_e \partial \mathbf{t}^T} l(N_e, \mathbf{t}) \text{Var}_{\mathbf{x}}(\tilde{\mathbf{t}}|\mathbf{t}) \frac{\partial^2}{\partial N_e \partial \mathbf{t}} l(N_e, \mathbf{t}) \right\} \\
 &= \frac{N_e^2}{n-1} \\
 &\quad + \frac{1}{2^2(n-1)^2} E_i \left\{ \text{Var}_{\mathbf{x}} \left( \sum_{i=n}^2 i(i-1)\tilde{t}_i | \mathbf{t} \right) \right\}. \tag{A2}
 \end{aligned}$$

