An Integrated Framework for the Inference of Viral Population History From Reconstructed Genealogies

Oliver G. Pybus, Andrew Rambaut and Paul H. Harvey

Department of Zoology, University of Oxford, Oxford OX1 3PS, United Kingdom

Manuscript received May 21, 1999
Accepted for publication March 30, 2000

ABSTRACT

We describe a unified set of methods for the inference of demographic history using genealogies reconstructed from gene sequence data. We introduce the skyline plot, a graphical, nonparametric estimate of demographic history. We discuss both maximum-likelihood parameter estimation and demographic hypothesis testing. Simulations are carried out to investigate the statistical properties of maximum-likelihood estimates of demographic parameters. The simulations reveal that (i) the performance of exponential growth model estimates is determined by a simple function of the true parameter values and (ii) under some conditions, estimates from reconstructed trees perform as well as estimates from perfect trees. We apply our methods to HIV-1 sequence data and find strong evidence that subtypes A and B have different demographic histories. We also provide the first (albeit tentative) genetic evidence for a recent decrease in the growth rate of subtype B.

COALESCENT theory provides a framework for understanding the relationship between a population's demographic history and its genealogy. The rapid accumulation of gene sequence data has prompted the development of many coalescent-based methods with the common aim of inferring the history of population size from samples of gene sequences. These methods fall into three categories: (i) methods that compare observed distributions of pairwise genetic differences with expected distributions derived from coalescent theory (Di Rienzo and Wilson 1991; Slatkin and Hudson 1991; Polanski *et al.* 1998; Grassly *et al.* 1999); (ii) methods that calculate the likelihood of an observed set of sequences given specified models of sequence evolution and demographic change (Griffiths and Tavaré 1994; Kuhner *et al.* 1995, 1998); and (iii) methods that infer past population history from a reconstructed genealogy (Fu 1994; Nee *et al.* 1995; Pybus *et al.* 1999).

Most previous studies have concentrated on estimating the parameters of simple demographic models, such as exponential growth and constant population size. However, estimated demographic parameters are meaningful only if there is a prior reason to believe that the sampled population fits the specified demographic model. How, therefore, can we proceed with parametric estimation if we have sequences sampled from a population with an unknown history? Nee *et al.* (1995) suggested using lineages-through-time (LTT) plots that display the rate of coalescence in a reconstructed genealogy through time. Polanski *et al.* (1998) introduced a method for estimating the history of population size from an observed distribution of pairwise differences. Although both of these methods can be used to infer demographic trends, neither provides a measure of confidence or allows demographic hypotheses to be tested. Furthermore, pairwise difference methods are expected to be less efficient than methods that incorporate genealogical information (Felsenstein 1992). Here, we introduce a new nonparametric estimate of population history and describe a maximum-likelihood (ML) framework for demographic parameter estimation and hypothesis testing. We apply these methods to HIV-1 sequence data and compare our results with those obtained using other techniques.

The rationale of our approach is as follows. Given a set of sequences S, the likelihood of demographic hypothesis H is given by $L[H|S] = \int_G L[H|G] \cdot L[G|S]$, where G is a genealogy with specified branch lengths. $L[H|G]$ is provided by coalescent theory and $L[G|S]$ can be calculated using standard likelihood methods (Felsenstein 1981). However, the integration must be performed over the set of all possible genealogies, which is impractically large. Consequently Griffiths and Tavaré (1994) and Kuhner *et al.* (1995, 1998) have developed general methods for estimating $L[H|S]$ using Monte Carlo integration. Unfortunately these methods are computationally intensive and difficult to implement with complex substitutional and demographic models. We propose a complementary simpler approach in which the likelihood of H is calculated directly from G^*, a genealogy with specified branch lengths reconstructed from S. G^* must be estimated under the assumption of a molecular clock. We also recommend

Corresponding author: Oliver Pybus, Department of Zoology, So. Parks Rd., Oxford OX1 3PS, United Kingdom.
E-mail: oliver.pybus@zoo.ox.ac.uk

that G^* is reconstructed using a maximum-likelihood approach, so that it is an estimate of the most likely tree given the data (the tree with the greatest value of $L[G|S]$). In effect, our approach assumes that (i) very few genealogies are likely to have given rise to S, and (ii) the reconstruction method can reliably locate these genealogies. This method may be practical for sequences that contain much phylogenetic information, particularly those sampled from rapidly evolving RNA virus populations, and we report simulation results suggesting that this is so. Monte Carlo integration methods should also run more quickly on such sequences, in comparison to data sets that contain little variation.

METHOD

Background theory: We are interested in the genealogy of individuals randomly sampled from a large population, the size of which varies deterministically through time. Griffiths and Tavaré (1994) and Donnelly and Tavaré (1995) have shown that such genealogies can be modeled using the variable population size coalescent process. An outline of this process is given below. Consider a large haploid population with no recombination, subdivision, or migration, and let $N_e(x)$ be the effective population size at time x. Time (in units of generations) increases into the past; $N_e(0)$ is the effective population size at the present. The relationship between $N_e(x)$ and the census population size $N(x)$ is determined by the population's reproduction model. If the population reproduces according to the Wright-Fisher model (each offspring selects a parent randomly from the previous generation), then $N_e(x) = N(x)$ (Kingman 1982).

Consider a set of n gene sequences randomly sampled from the population at the present. The genealogy of the sampled sequences will contain $n - 1$ ordered internode intervals, labeled I_2, I_3, . . . , I_n. The sizes of these intervals are denoted g_2, g_3, . . . , g_n. The subscripts refer to the number of lineages present in the sampled genealogy during each interval. We apply the coalescent approximation throughout, such that the interval sizes g_i are distributed according to the probability density function

$$p(g_i \mid t_i) = \frac{\binom{i}{2}}{N_e\,(g_i + t_i)} \exp\left[- \int_{x=t_i}^{g_i+t_i} \frac{\binom{i}{2}}{N_e(x)} dx \right], \quad (1)$$

where t_i is the time at which interval I_i starts (Griffiths and Tavaré 1994). Equation 1 provides the cornerstone of our framework. However, genealogies reconstructed under the assumption of a molecular clock have internode intervals measured in units of expected substitutions per site, not generations. We therefore rescale the coalescent process into the same units by implementing the change of variable $\gamma_i = \mu g_i$, where γ_i is interval size in substitutions per site and μ is the mutation rate in substitutions per site per generation. After this rescaling, $p(g_i|t_i)$ becomes $p(\gamma_i|\tau_i)$, and many variables become functions of μ (see below).

We call $N_e(x)$ the *demographic model*, as it represents change in population size through time. Here, we consider two tractable and common demographic models, constant population size, $N_e(x) = N_e(0)$, and exponential growth, $N_e(x) = N_e(0)e^{-rx}$, where r is the exponential growth rate. $N_e(0)$ and r are *demographic parameters* that we may wish to estimate. If time is measured in substitutions, then $\theta = N_e(0)\mu$ and $\rho = r/\mu$. A *demographic hypothesis* is a demographic model with specified parameter values. All the methods described in this section are implemented in the program GENIE, available from http://evolve.zoo.ox.ac.uk.

Simulating coalescent trees: If U is a unit uniform random variable, then solving

$$U = \exp\left[- \int_{x=t_i}^{g_i+t_i} \frac{\binom{i}{2}}{N_e(x)} dx \right] \quad (2)$$

for g_i will generate a variate sampled randomly from Equation 1 (Donnelly and Tavaré 1995). Solutions for the constant size and exponential growth models are provided by Hudson (1990) and Slatkin and Hudson (1991). Equation 2 enables coalescent trees to be simulated under almost any demographic hypothesis; the only caveat is that the value of the integral must tend to infinity as x increases, that is, $N_e(x)$ cannot increase indefinitely into the past. Fortunately, all demographic histories that fail this restriction are biologically implausible.

The skyline plot: Here we describe a new nonparametric estimate of demographic history. Since a reconstructed genealogy provides estimates of the random variables g_i and t_i, what can we infer about $N_e(x)$ from these values? Rearranging Equation 2 we obtain the relationship

$$g_i\binom{i}{2} = -\ln(U)H_i, \quad \text{where } H_i = \left(\int_{x=t_i}^{g_i+t_i} \frac{1}{N_e(x)}\, dx/g_i \right)^{-1}. \quad (3)$$

H_i has a meaningful biological interpretation—it is the harmonic mean of effective population size in the range $[t_i, g_i + t_i]$, where $[t_i, g_i + t_i]$ is the time interval delimited by internode interval I_i. If time is measured in substitutions per site, then

$$\gamma_i \binom{i}{2} = -\ln(U)H_i\mu. \quad (4)$$

Since $-\ln(U)$ represents random error, the term $\hat{M}_i = \gamma_i \binom{i}{2}$ is an estimate of $H_i\mu$ that can be calculated from an observed genealogy. Consequently, a plot of $\hat{M}_i$ against time defines a piecewise function that is a nonparametric estimate of demographic history. We name these plots *skyline plots*. $\hat{M}_i$ represents all the information

about $N_e(x)$ that can be inferred from the observed interval I_i. In other words, the most we can infer from I_i is that it defines a time interval $[t_i, g_i + t_i]$, during which the harmonic mean of $N_e(x)\mu$ is estimated to be $\hat{M}_i$. If the substitution rate μ is known, then $g_i(\frac{i}{2})$ can be used to estimate H_i directly.

Figure 1 illustrates the ability of skyline plots to reconstruct population history under a variety of demographic scenarios. $\hat{M}_i$ is equal to $N_e(x)\mu$ only if $N_e(x)\mu$ is constant. Hence, for the constant size model, the arithmetic mean of the $\hat{M}_i$ is equal to the maximum-likelihood estimate of effective population size (Felsenstein 1992). For other demographic models, $\hat{M}_i$ can underestimate $N_e(x)\mu$, because a harmonic mean is always smaller than its corresponding arithmetic mean. As illustrated in Figure 1b, this systematic bias is small when the rate of coalescence is large compared to the rate of population change [that is, when the harmonic and arithmetic means of $N_e(x)\mu$ during an interval are similar].

Parametric estimation: We use a maximum-likelihood approach to parameter estimation. Given an observed interval size γ_i, the likelihood function is

$$L[\psi|\gamma_i] = kp(\gamma_i|\tau_i), \tag{5}$$

where ψ represents the parameter values of the demographic hypothesis and k is an arbitrary constant (Edwards 1972). $l[\psi|G]$, the log likelihood of the γ_2, γ_3, . . . , γ_n from an observed genealogy is simply the sum of the log likelihoods for each interval,

$$l[\psi|G] = \sum_{i=2}^{n} \ln(L[\psi|\gamma_i]). \tag{6}$$

The maximum-likelihood estimate (MLE) of ψ, denoted $\hat{\psi}$, can be found numerically. In addition, the shape of the likelihood surface near $\hat{\psi}$ can be used to obtain approximate confidence limits. We obtained ~95% confidence sets using the likelihood-ratio test (LRT). A point in parameter space, ψ', lies within the confidence set if it satisfies

$$l[\psi'|G] \geq l[\hat{\psi}|G] - \frac{1}{2}\chi^2_d, \tag{7}$$

where d is the number of demographic parameters. Our simulation results (see next section) suggest that this heuristic approach is reasonably accurate. The likelihood framework above can be used to estimate the parameters of almost any demographic model, provided that the probability model is well formed and the MLE can be reliably located.

Hypothesis testing: Here we provide two methods for testing demographic hypotheses. First, we describe how to accept or reject specific hypotheses. Given a demographic hypothesis $N_e(x)$, we can use Equation 2 to transform the observed internode intervals γ_2, γ_3, . . . , γ_n into u_2, u_3, . . . , u_n, a series of numbers in the range $[0, 1]$. If $N_e(x)$ represents the true history of the observed genealogy, then the u_i will be distributed according to a unit uniform distribution. The hypothesis $N_e(x)$ can therefore be tested using the one-sample Kolmorgorov-Smirnov (KS) test. The KS test statistic is a measure of the difference between an observed distribution and an expected distribution. Therefore, values of the KS test statistic can also be used to compare the goodness-of-fit of any pair of demographic hypotheses.

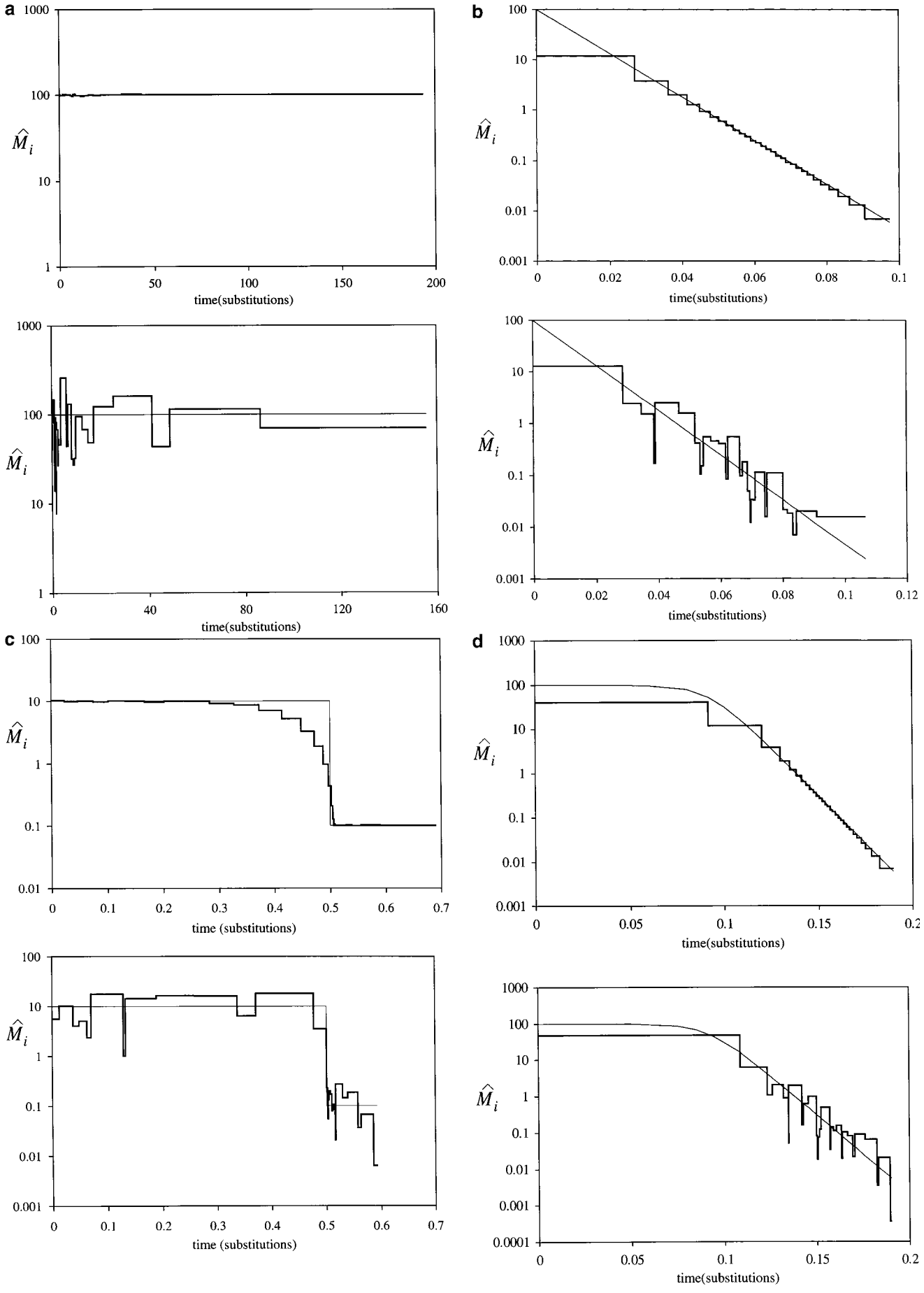
The second method enables us to reject entire classes of hypotheses. More specifically, demographic model A can be rejected in favor of model B, provided A is a special case of B. For example, the constant-size model is a special case of the exponential model (corresponding to $r = 0$). We can reject the constant-size model in favor of the exponential model if the confidence intervals of the MLE of r do not include zero. This test can be similarly applied to any pair of nested demographic models. However, the reliability of this procedure, which is essentially a LRT, depends on the accuracy of the approximate confidence intervals (see above).

SIMULATIONS

The performance of maximum-likelihood estimates: Extensive simulations were carried out to investigate the bias, variability, and type I error rate of ML estimates calculated from simulated coalescent trees. These simulations describe the performance of ML estimates when genealogies are reconstructed without error, and therefore represent the "best case" scenario, against which the performance of other methods should be compared. The simulations were performed as follows:

i. A coalescent tree with 30 tips was simulated with specified parameter values, using Equation 2.
ii. A MLE ($\hat{p}$) and ~95% confidence intervals were obtained for each parameter.
iii. Steps i and ii were repeated 200 times.
iv. The bias of the MLE of parameter value p was calculated as $b(p) = (E[\hat{p}] - p)/p$.
v. The variability of the MLE of parameter value p was calculated as $v(p) = \text{var}[\hat{p}]/p^2$.
vi. The type I error rate of the MLE of parameter value p, denoted $e(p)$, was calculated as the number of simulated trees for which the true parameter value lay outside the 95% confidence intervals of the MLE. $e(p)$ is expected to be binomially distributed with parameters 0.05 and 200.
vii. Steps i–vi were repeated for many different parameter values.

The constant-size model was investigated first and our results agreed with the theoretical values provided by Felsenstein (1992). The MLE of θ is unbiased and has variability $v(\theta) = 1/(n - 1)$, where n is the number of tips in the genealogy. The type I error rate $e(\theta)$ was within its expected range (results not shown).


a
b
c
d
$\hat{M}_i$
time(substitutions)
time (substitutions)

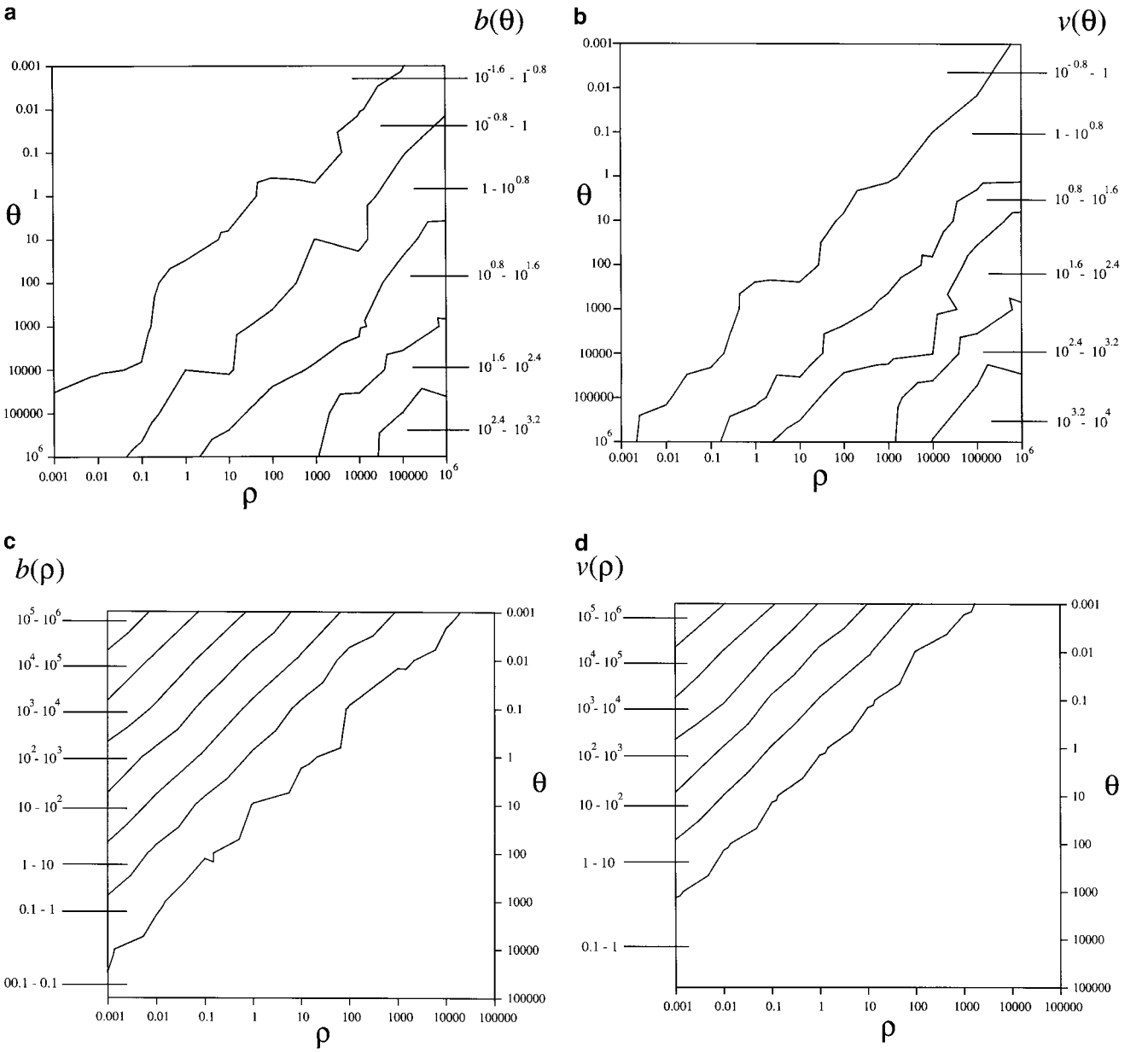


Figure 2.—The bias and variability of MLEs of exponential model parameters. (a) Bias of θ estimates. (b) Variability of θ estimates. (c) Bias of ρ estimates. (d) Variability of ρ estimates.

The bias and efficiency of the exponential model parameters, θ and ρ, are shown in Figure 2. In agreement with Kuhner *et al.* (1998), both parameters are biased upward, the bias of ρ being more severe than the bias of θ. In addition, our results reveal an important and hitherto undetected pattern—the bias and variability of both parameters depend only on their product, $\alpha = \theta\rho$. α has previously been shown to be important in determining the behavior of the exponential-growth coalescent process (Slatkin and Hudson 1991; Pybus *et al.* 1999). As α increases, the bias and variability of θ estimates increase, but the bias and variability of ρ estimates decrease. Hence, MLEs of ρ are more accurate when $\alpha \ll 1$ and are less accurate when $\alpha \gg 1$. The

Figure 1.—Skyline plots can reconstruct population history under different demographic scenarios. (a) Constant population size, θ = 100. (b) Exponential growth, θ = 100, ρ = 100. (c) A 100-fold instantaneous increase in population size at time 0.5. (d) Logistic growth. The vertical axis shows estimated $N_e\mu$. The horizontal axis represents time in units of substitutions; time is zero at the present. In a–d, the top graph shows the expected skyline plot, obtained by calculating the mean of 5000 plots. The bottom graph shows the skyline plot of a single genealogy simulated under the same conditions. Both the inferred (thick lines) and true (thin lines) demographic histories are shown.

opposite is true for MLEs of θ. It appears that $\alpha = 1$ marks a transition in the behavior of the exponential coalescent process, which behaves similarly to the constant-size process when $\alpha \ll 1$, but generates increasingly star-like trees as α increases. The error rates $e(\theta)$ and $e(\rho)$ were within their expected ranges (results not shown).

The effect of phylogenetic reconstruction: A second set of simulations was performed to investigate the relative performance of MLEs calculated from correct and reconstructed genealogies. The simulations were performed as follows:

i. Coalescent trees with 15 tips were simulated with specified parameter values, using Equation 2.
ii. MLEs of θ and ρ were obtained from each coalescent tree (the true tree).
iii. Sequences, 1000 nucleotides in length, were simulated down each coalescent tree according to the Hasegawa-Kishino-Yano (HKY85) substitution model, with equal base frequencies and a transition:transversion ratio (ti:tv) of 2 (Hasegawa *et al.* 1985). This step was performed with Seq-Gen (Rambaut and Grassly 1997).
iv. Reconstructed trees were obtained from the simulated sequences using two methods. For the first method, a ML distance matrix was estimated using the HKY85 model (ti:tv was estimated), from which a UPGMA tree was constructed. The branch lengths of the UPGMA tree were then reestimated using ML (again under the HKY85 model). The second reconstruction method was a heuristic ML topology search, using the stepwise-addition and nearest-neighbor-interchange algorithms. Tree likelihoods were calculated using HKY85 (ti:tv was estimated) with molecular clock enforced. This step was performed with PAUP4.0b3a (Swofford 1999).
v. MLEs of θ and ρ were calculated from each reconstructed UPGMA and ML tree.
vi. Bias, variability, and error rates were calculated as described in the previous section.

The above procedure was repeated with two sets of demographic parameters ($\theta = 10$, $\rho = 100$) and ($\theta = 1$, $\rho = 1000$), which represent two populations with the same demographic history ($N_e(0) = 10^4$, $r = 0.1$) but different substitution rates ($\mu = 10^{-3}$ and $\mu = 10^{-4}$, respectively). A lower substitution rate will result in less diverse sequences, making accurate tree reconstruction more difficult.

Figure 3 shows the distribution of θ estimates obtained using the $\mu = 10^{-3}$ substitution rate. This distribution is highly skewed with a long upper tail, making $b(\theta)$ and $v(\theta)$ very difficult to estimate with a small number of replicates (and perhaps causing the stochasticity seen in Figure 2, a and b). Table 1, which contains the simulation results, therefore displays percentiles of this distribution. The distribution of ρ estimates was much less skewed so $b(\rho)$ and $v(\rho)$ could be estimated accurately.

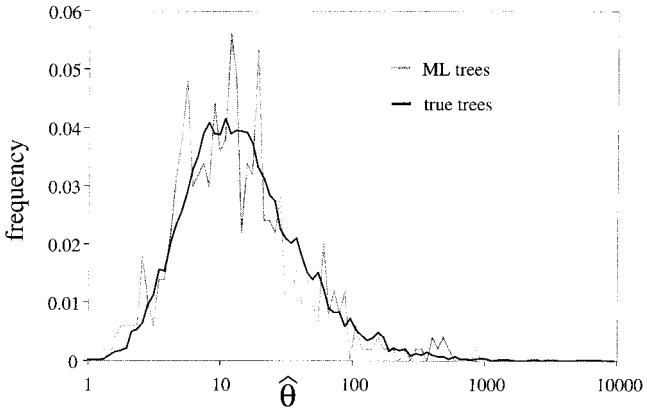


Figure 3.—Distribution of MLEs of θ obtained from true trees (10,000 replicates) and ML-reconstructed trees (500 replicates). The parameters used were $\theta = 10$, $\rho = 100$.

For the faster rate ($\mu = 10^{-3}$), MLEs from reconstructed trees performed as well as MLEs from the correct tree (Table 1; Figure 3). Surprisingly, the UPGMA algorithm did as well as ML tree estimation. For the slower rate ($\mu = 10^{-4}$), the error rates of MLEs from reconstructed trees were higher than those from the correct trees. In this scenario, UPGMA fared worse than ML tree estimation. These results suggest that using a reconstructed genealogy to infer demographic history becomes more reasonable as substitution rate increases (provided that the sequences do not become saturated with substitutions).

As far as possible, the above simulations were designed to mimic the evolution of RNA viruses; $\mu = 10^{-3}$ is slightly less than current estimates of HIV-1 substitution rate (Li *et al.* 1988; Leitner and Albert 1999). However, simplifications have been made for the sake of computational feasibility, so these results should be considered as encouraging but preliminary. A more complete study that incorporates among-site rate heterogeneity, larger trees, different sequence lengths, alternative heuristic search strategies, and more complex substitutional models is necessary.

EXAMPLE: THE HIV-1 EPIDEMIC

Here, we illustrate our methods using four HIV-1 data sets, which contain *env* and *gag* gene sequences from HIV-1 subtypes A and B. These two prevalent subtypes have differing geographical distributions and transmission routes. Subtype A is found mostly in sub-Saharan Africa, where ~90% of transmissions occur through heterosexual intercourse. In contrast, subtype B circulates mainly in the developed world and has been predominately transmitted via intravenous drug use and homosexual intercourse (UNAIDS 1998). The data sets used here, labeled *env*A, *env*B, *gag*A, and *gag*B, were reported in Pybus *et al.* (1999). They were carefully compiled to minimize the effects of nonrandom sam-

TABLE 1

The effect of phylogenetic reconstruction on demographic parameter estimates

Substitution rate	Demographic parameters	Tree reconstruction method	$e(\hat{\theta})$	$e(\hat{\rho})$	Percentiles of the distribution of $\hat{\theta}$			$b(\hat{\rho})$	$v(\hat{\rho})$
					0.025	Median	0.975		
$\mu = 10^{-3}$	$\theta = 10$	True tree[a]	0.062	0.063	2.63	12.45	149.1	0.081	0.045
	$\rho = 100$	UPGMA[b]	0.060	0.076	2.35	11.43	139.1	0.048	0.046
		ML[c]	0.076	0.074	2.21	11.28	172.4	0.057	0.053
$\mu = 10^{-4}$	$\theta = 1$	True tree[a]	0.059	0.060	0.27	1.24	14.4	0.080	0.043
	$\rho = 1000$	UPGMA[b]	0.262	0.476	0.10	0.428	5.04	−0.214	0.053
		ML[c]	0.212	0.312	0.09	0.552	13.6	−0.088	0.089

[a] Values calculated from 10,000 replicates.
[b] Values calculated from 1000 replicates.
[c] Values calculated from 500 replicates.

pling and intersubtype recombination. ML genealogies were estimated for each data set using HKY85 and a codon-position model of rate heterogeneity, under the assumption of a molecular clock. The genealogies are described fully in Pybus *et al.* (1999).

Figure 4 shows the skyline plots obtained from the four HIV-1 genealogies. The subtype A plots indicate a constant-rate exponential increase in population size toward the present, whereas the demographic history of subtype B appears to be logistic. However, the subtype B plots are also consistent with the hypothesis of exponential growth, because genealogies from rapidly expanding populations have large internode intervals near the present (Slatkin and Hudson 1991), resulting in an underestimation of population size (see Figure 1b for example). Previous LTT plot analyses of HIV-1 have only indicated that both subtypes A and B have increased exponentially (Holmes *et al.* 1999).

To further investigate the demographic history of HIV-1, we estimated θ and ρ using the ML framework

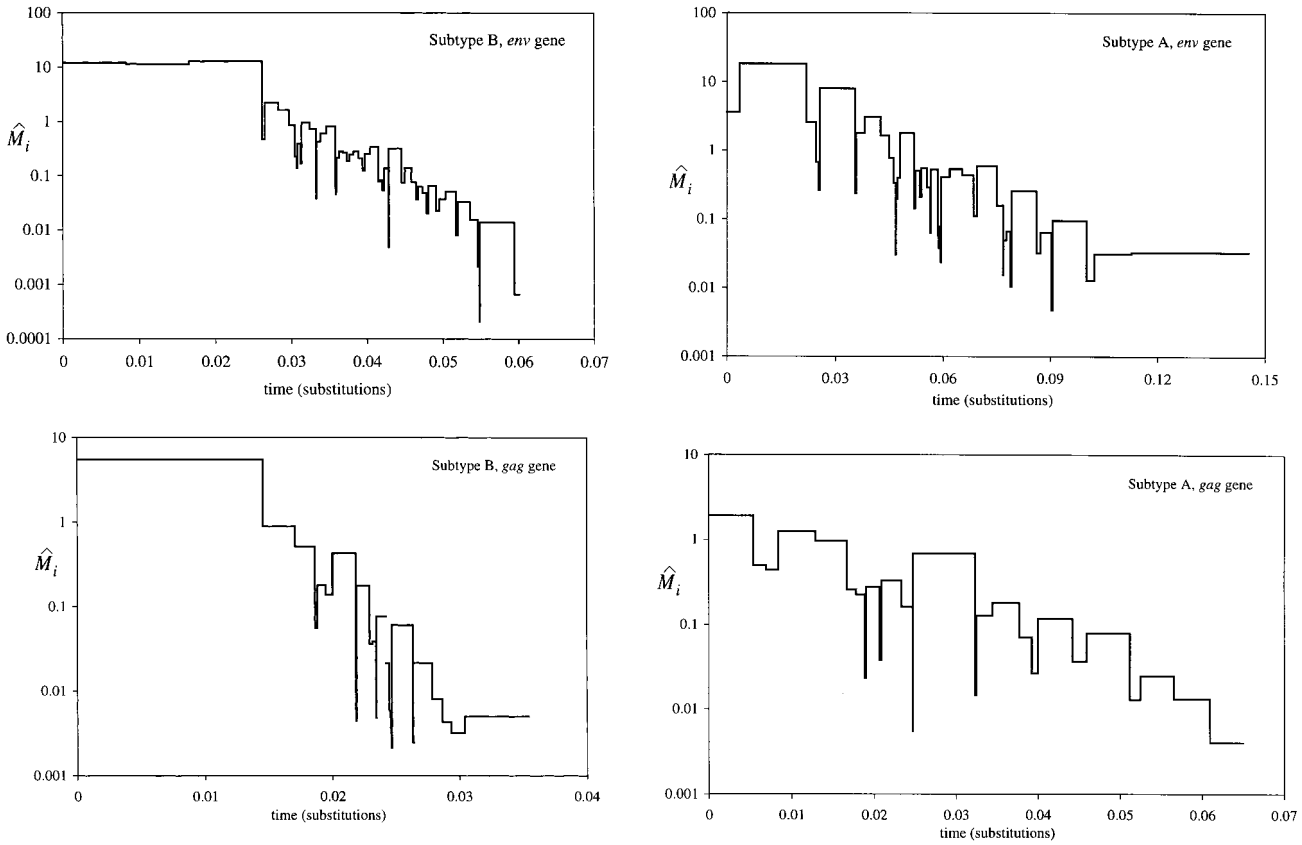


Figure 4.—The skyline plots of the four HIV-1 genealogies (see Figure 1 for details).

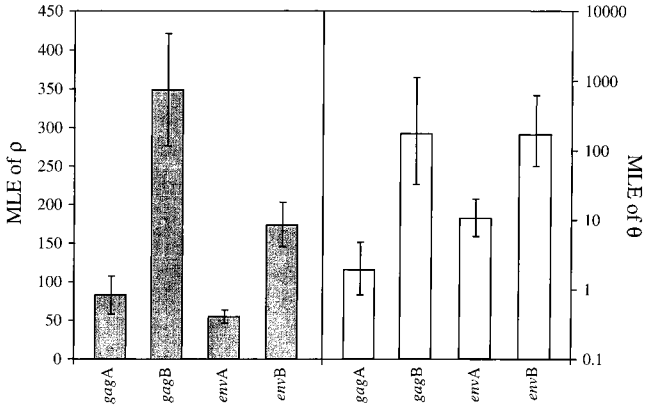


Figure 5.—MLEs of θ (open bars) and ρ (shaded bars) for the four HIV-1 genealogies.

described above (results shown in Figure 5). *gag* and *env* genealogies belonging to the same subtype generate different parameter estimates. This is because θ and ρ are both functions of the substitution rate, which is higher in the *env* gene than in the *gag* gene (Li *et al.* 1988; Leitner and Albert 1999). Hence, for each subtype, *gag* genealogies tend to generate larger ρ estimates and smaller θ estimates than *env* genealogies.

We calculated MLEs of $\alpha = \theta\rho$ using the results in Figure 5. For both subtypes, *gag* and *env* estimates of α are similar, which is expected because α is independent of substitution rate (results not shown). These MLEs were compared with α estimates obtained from the same sequences using the mid-depth method (Pybus *et al.* 1999). The ML and mid-depth estimates were very similar and both suggested that α is greater for subtype B than for subtype A. The mid-depth method confidence intervals were larger, indicating that the method presented here is more powerful.

Returning to Figure 5, both θ and ρ are greater for subtype B than for subtype A. For both subtypes $\alpha \gg 1$, hence ρ estimates are expected to be almost unbiased (Figure 2) and it is safe to infer that subtype B has a faster growth rate than subtype A. We suggest that this result is due to the different modes of transmission that characterized the initial spread of the subtypes. In the developed world, subtype B transmission was aided by the presence of interconnected "standing networks" of intravenous drug users and homosexual men, within which transmission rates were very high (Robertson *et al.* 1986; Jacquez *et al.* 1994). In sub-Saharan Africa, subtype A was spread via heterosexual intercourse, so average waiting times between transmissions were longer (Tarantola and Schwartlander 1997).

Our estimates of current population size, θ, are larger for subtype B than for A. Taken at face value this result is surely wrong: sub-Saharan Africa contains ~70% of the world's HIV-1-infected individuals (UNAIDS 1998), more than half of which appear to be infected with subtype A (Rayfield *et al.* 1998; Robbins *et al.* 1999). However, α values for subtype B are large, so MLEs of θ are expected to be biased upward (Figure 2). In contrast, θ estimates for subtype A are expected to be almost unbiased. We also suggest a second possible explanation for these results; the demographic history of subtype B may not be exponential. The subtype B skyline plots are consistent with both a logistic and an exponential demographic history. If the logistic model is correct, then estimates of θ obtained under the exponential model will be too large. A logistic scenario is partially consistent with epidemiological evidence, as the introduction of behavioral intervention and antiretroviral therapies in western Europe has led to a decrease (although not a cessation) in the number of new infections (UNAIDS 1998).

Grassly *et al.* (1999) studied the population dynamics of HIV-1 using a pairwise difference distribution method. They estimated the current effective population size of subtype A to be larger than that of subtype B. However, they assumed equal growth rates for both subtypes, an assumption that our results suggest is incorrect. Clearly further work is needed to reconcile these differences.

We believe that our results are not an artifact of nonrandom sampling, recombination, selection, or variable substitution rates. As discussed previously in Holmes *et al.* (1999) and Grassly *et al.* (1999), these processes are not expected to be acting differentially at the subtype level. Furthermore, our skyline plots are consistent across different genes, indicating that they are surprisingly robust to different substitution rates and modes of selection (Li *et al.* 1988).

CONCLUSION

Skyline plots are the most appropriate way of graphically displaying the demographic information contained in reconstructed genealogies. They display estimated population size against time, and are therefore more intuitive than Nee *et al.*'s (1995) LTT plots, which must be interpreted using transformations. As skyline plots explicitly incorporate genealogy, they are expected to make more efficient use of the data than Polanski *et al.*'s (1998) more complex pairwise difference distribution method.

Omitting the computationally difficult task of integrating over all possible genealogies greatly simplifies ML parameter estimation. This allows us to use the complex substitution models necessary to accurately represent HIV-1 evolution (Leitner *et al.* 1997). Furthermore, in future work it should be possible to implement more realistic demographic models; our HIV-1 results suggest that the constant size and exponential models alone are not sufficient. However, it is essential to quantify the effects of tree reconstruction on parameter estimation and to determine the conditions under

which our method may be appropriate—the simulations reported here are only preliminary.

Research is also needed to quantify the effects of recombination, selection, subdivision, variable substitution rates, and nonrandom sampling on the accuracy of demographic inference. If these processes, at biologically realistic levels, have a significant effect, then they must be incorporated into the coalescent framework. Significant progress in this area is well underway (see Rodrigo and Felsenstein 1999), although it may be impossible to implement recombination in a framework that considers only a single tree.

Thanks to Mike Charleston, Eddie Holmes, Mike Worobey, and Bob Griffiths for their advice. Thanks to Peter Donnelly for comments on the manuscript and helpful discussions of statistical problems. We also thank the reviewing editor and two anonymous referees for their suggestions. This work was funded by the Wellcome Trust (grant 050275) and The Royal Society.

LITERATURE CITED

Di Rienzo, A., and A. C. Wilson, 1991 Branching pattern in the evolutionary tree for human mitochondrial DNA. Proc. Natl. Acad. Sci. USA **88:** 1597–1601.

Donnelly, P., and S. Tavaré, 1995 Coalescents and genealogical structure under neutrality. Annu. Rev. Genet. **29:** 401–421.

Edwards, A. W. F., 1972 *Likelihood*. Cambridge University Press, Cambridge, United Kingdom.

Felsenstein, J., 1981 Evolutionary trees from gene frequencies and quantitative characters: finding maximum likelihood estimates. Evolution **35:** 1229–1242.

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.

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

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

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

Hasegawa, M., H. Kishino and T. Yano, 1985 Dating the human-ape splitting by a molecular clock of mitochondrial DNA. J. Mol. Evol. **22:** 160–174.

Holmes, E. C., O. G. Pybus and P. H. Harvey, 1999 The molecular population genetics of HIV-1, pp. 177–207 in *The Evolution of HIV*, edited by K. A. Crandall. John Hopkins University Press, Baltimore.

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

Jacquez, J. A., J. S. Koopman, C. P. Simon and I. M. Longini, 1994 Role of primary infection in epidemics of HIV infection in gay cohorts. J. Acquired Immune Defic. Syndr. **7:** 1169–1184.

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

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.

Leitner, T., and J. Albert, 1999 The molecular clock of HIV-1 unveiled through analysis of a known transmission history. Proc. Natl. Acad. Sci. USA **96:** 10752–10757.

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 I populations with a known transmission history. J. Virol. **71:** 4761–4770.

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

Nee, S., E. C. Holmes, A. Rambaut and P. H. Harvey, 1995 Inferring population history from molecular phylogenies. Philos. Trans. R. Soc. Lond. **349:** 25–31.

Polanski, A., M. Kimmel and R. Chakraborty, 1998 Application of a time-dependent coalescence process for inferring the history of population size changes from DNA sequence data. Proc. Natl. Acad. Sci. USA **95:** 5456–5461.

Pybus, O. G., E. C. Holmes and P. H. Harvey, 1999 The mid-depth method and HIV-1: a practical approach for testing hypotheses of viral epidemic history. Mol. Biol. Evol. **16:** 953–959.

Rambaut, A., and N. C. Grassly, 1997 Seq-Gen: an application for the Monte Carlo simulation of DNA sequence evolution along phylogenetic trees. Comput. Appl. Biosci. **13:** 235–238.

Rayfield, M. A., R. G. Downing, J. Baggs, J. Hu, D. Pieniazek *et al.*, 1998 A molecular epidemiologic survey of HIV in Uganda. AIDS **12:** 521–527.

Robbins, K. E., L. G. Kostrikis, T. M. Brown, O. Anzala, S. Shin *et al.*, 1999 Genetic analysis of human immunodeficiency virus type 1 strains in Kenya: a comparison using phylogenetic analysis and a combinatorial melting assay. AIDS Res. Hum. Retro. **15:** 329–335.

Robertson, J. R., A. B. Bucknall, P. D. Welsby, J. J. Roberts, J. M. Inglis *et al.*, 1986 Epidemic of AIDS related virus (HTLV-III/LAV) infection among intravenous drug abusers. Br. Med. J. **292:** 527–529.

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. John Hopkins University Press, Baltimore.

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

Swofford, D. L., 1999 PAUP* version 4.0b3a. Sinauer Associates, Sunderland, MA.

Tarantola, D., and B. Schwartlander, 1997 HIV/AIDS epidemics in sub-Saharan Africa: dynamism, diversity and discrete declines. AIDS **11:** S5–S21.

UNAIDS, 1998 Report on the global HIV/AIDS epidemic. www.who.org/emc-hiv.

Communicating editor: S. Tavaré