# The Mid-Depth Method and HIV-1: A Practical Approach for Testing Hypotheses of Viral Epidemic History

*Oliver G. Pybus, Edward C. Holmes, and Paul H. Harvey*

Department of Zoology, University of Oxford

We introduce the mid-depth method, a practical approach for testing hypotheses of demographic history using genealogies reconstructed from sequence data. The relative positions of internal nodes within a genealogy contain information about past population dynamics. We explain how this information can be used to (1) test the null hypothesis of constant population size and (2) estimate the growth rate and current population size of an exponentially growing population. Simulation tests indicate that, as expected, estimates of exponential growth rates are sometimes biased. The mid-depth method is computationally rapid and does not require knowledge of the sample's mutation rate. However, it does assume that the reconstructed genealogy is correct and is therefore best suited to the analysis of variation-rich viral data sets. When applied to HIV-1 sequence data, the mid-depth method provides phylogenetic evidence of different exponential growth rates for subtypes A and B. We posit that this difference in growth rate reflects the different transmission routes and epidemiological histories of the two subtypes.

## Introduction

Homologous DNA sequences sampled from a population carry information about the demographic history of that population. Such information has wide-ranging applications, from virology (Holmes et al. 1995) to anthropology (Harpending et al. 1993). Here, we present a simple and computationally rapid method for the inference of population dynamic history using genealogies reconstructed from sequence data. Our method, called the mid-depth method, is based on coalescent theory and tests hypotheses by comparing reconstructed genealogies against genealogies generated using Monte Carlo simulation.

Previous attempts to infer demographic history from gene sequence data have followed two general approaches. The nonphylogenetic approach involves the calculation of summary statistics directly from sequence data, such as the number of segregating sites in a sample (Tajima 1989) or the pairwise differences between sequences (Slatkin and Hudson 1991; Rogers and Harpending 1992). In contrast, the phylogenetic approach takes into account the genealogical relationships of the sampled sequences. This has been done in two ways: (1) The genealogy is assumed to be unknown. Maximum-likelihood estimates of demographic parameters are calculated by summing or sampling over all possible genealogies, given explicit models of sequence evolution and population change (Griffiths and Tavaré 1994; Kuhner, Yamato, and Felsenstein 1995, 1998). (2) The genealogy is reconstructed using standard phylogenetic methods and assumed to be correct. Demographic information is then inferred from the reconstructed tree (Nee et al. 1995). The method presented here belongs to this latter category.

Felsenstein (1992) demonstrated that estimates of population size based on phylogenetic approaches make more efficient use of the data than do summary statistics. Furthermore, phylogenetic methods for the estimation of population growth rates are expected to outperform methods based on summary statistics when the growth rate is low or negative (Kuhner, Yamato, and Felsenstein 1998). Unfortunately, phylogenetic approaches which integrate over all possible genealogies tend to be computationally intensive and complex. The mid-depth method avoids these problems by separating the task of demographic parameter estimation from that of phylogenetic reconstruction (*sensu* Nee et al. 1995). The drawback of our approach is that uncertainty about the genealogy is not incorporated into the process of parameter estimation (cf. Kuhner, Yamato, and Felsenstein 1998). Instead, the reconstructed genealogy is assumed to be correct, requiring that the sequences in question contain enough phylogenetic signal to accurately infer divergence times. The effect of phylogenetic error on parameter estimation has yet to be quantified; for the time being, we advise caution and recommend that our method should be applied only to rapidly evolving viral genes.

In this paper, we apply the mid-depth method to human immunodeficiency virus type 1 (HIV-1) gene sequence data and find evidence of different growth rates for HIV-1 subtypes A and B.

## Methods

A coalescent tree is a hypothetical phylogeny which describes the genealogical relationships between a small number of individuals sampled from a large population (Kingman 1982). Typically, its tips are contemporaneous, and its internal nodes (coalescence events) are ordered in time (Felsenstein 1992). Here, coalescent trees will be considered solely in terms of the time intervals between their coalescence events. Algorithms are available to generate the time intervals of coalescent trees sampled from constant-sized (Hudson 1990) and exponentially growing (Slatkin and Hudson 1991) populations. Both of these algorithms are derived from probability models with all of the following assumptions: (1) no recombination; (2) phylogenetically-ran-

Key words: coalescent theory, Monte Carlo simulation, HIV-1, maximum likelihood, mid-depth method.

Address for correspondence and reprints: Oliver Pybus, Department of Zoology, University of Oxford, South Parks Road, Oxford, OX1 3PS, United Kingdom. E-mail: oliver.pybus@zoo.ox.ac.uk.

*Mol. Biol. Evol.* 16(7):953–959. 1999

dom sampled sequences; (3) $n \ll N$, where $n$ is the number of different sequences and $N$ is the population size; and (4) Wright-Fisher reproduction (generations are nonoverlapping, with each offspring choosing a parent at random and independent of the previous generation).

The constant-sized (endemic) model has only one parameter, population size ($N$) (Hudson 1990). Most coalescence events in an endemic coalescent tree tend to occur near the tips (Nee et al. 1995). The exponentially growing (epidemic) model has two parameters, current population size ($N_0$) and exponential growth rate ($r$). When both $N_0$ and $r$ are large, the simulated tree resembles a ''star'' phylogeny, with most coalescence events close to the root (Slatkin and Hudson 1991). It is therefore reasonable to predict that as $r$ decreases, coalescence events will tend to move away from the root toward the tips, such that a simulated epidemic tree will come to resemble a simulated endemic tree as $r$ approaches zero.

## The Mid-Depth Method

The above observations suggest that the relative positions of coalescence events within a reconstructed genealogy contain information about population dynamic history. To extract this information, the mid-depth method uses a new tree statistic, called $\sigma$. We define $\sigma$ as the number of coalescence events between the root and the mid-depth point of a genealogy. The mid-depth point is the time exactly halfway between the tree's tips and its root. Importantly, $\sigma$ is independent of any factor which is a constant multiple across all intervals, such as a constant mutation rate or the population size of a constant-sized population.

The internal nodes of the reconstructed genealogy must represent divergence times—a condition which usually requires that the tree reconstruction method used assumes a constant rate of molecular evolution. However, this assumption is not binding, as it is sometimes possible to estimate divergence times in the absence of a constant-rate molecular clock (see, for example, Sanderson 1997). If the molecular clock assumption is incorrectly applied, then the resulting inaccurately reconstructed genealogy may give rise to misleading demographic inferences. The mid-depth method also requires that the genealogy satisfies all of the coalescent model assumptions listed above.

## Testing the Hypothesis of Constant Population Size

If a genealogy represents a sample from a constant-sized population, then, following from the observations above, we expect its $\sigma$ value to be close to zero. Monte Carlo simulation shows this expectation to be correct; more than 95% of trees simulated using Hudson's (1990) endemic coalescent model have $\sigma \leq 3$, provided $2 < n < \approx 200$ (see fig. 1). This result can be used to test the null hypothesis, $H_0$, that a sample of gene sequences has been taken from a constant-sized population. Let $\sigma^*$ be the $\sigma$ value of a genealogy reconstructed from the sample. If $\sigma^* > 3$, then we can reject $H_0$ at the 95% confidence limit.

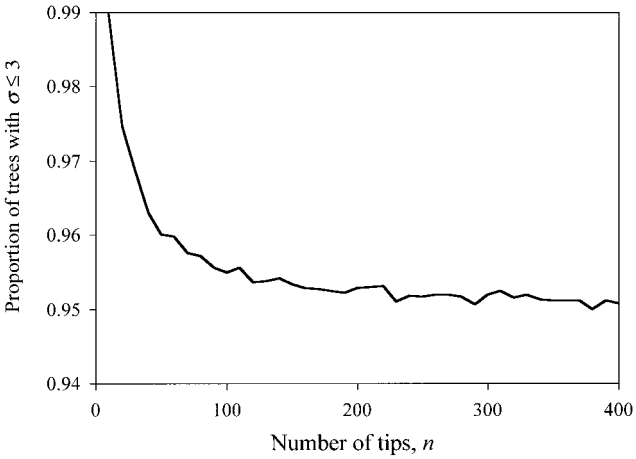


FIG. 1.—Testing the hypothesis of constant population size. For each value of $n$, 10,000 endemic coalescent trees were simulated using the algorithm of Hudson (1990). The ordinate shows the proportion of simulated trees with $n$ tips which have $\sigma \leq 3$. More than 95% of all simulated trees with $n < \approx 200$ tips have $\sigma \leq 3$.

## Estimating Exponential Growth Parameters

If the hypothesis of constant population size is rejected, the next step is to discover the type of population change that has occurred. A useful tool for this purpose is the lineages-through-time (LTT) plot, which can be used to infer the general mode of population change from a genealogy (Nee et al. 1995; Rambaut, Harvey, and Nee 1997). Information from an LTT plot, or perhaps some other external source, may provide a priori evidence that a genealogy has been sampled from a population growing at a constant exponential rate. If (and only if) such evidence exists, then the mid-depth method can be used to estimate $\alpha = N_0 r$, the product of the current population size $N_0$, and exponential growth rate $r$ of that population.

The relationship between $\alpha$ and $\sigma$ was investigated using the simulation algorithm presented by Slatkin and Hudson (1991). Their algorithm implements a renormalized epidemic coalescent model which measures time in units of $1/r$ generations and which has $\alpha$ as its only parameter. The renormalization is equivalent to multiplying all time intervals by the factor $1/r$ and therefore does not affect the value of $\sigma$.

Figure 2 shows the relationship between $\log(\alpha)$ and $\sigma$ for simulated epidemic coalescent trees with 30 and 100 tips. The plot demonstrates that coalescence events move from the tipward half of a tree to the rootward half as $\alpha$ increases. For a wide range of $\alpha$ values, there is an approximately linear relationship between $\sigma$ and $\log(\alpha)$, suggesting that $\sigma$ can be used to estimate $\alpha$. When $\alpha < \approx 1$ or $\alpha > \approx 10^6$, the $\sigma$ statistic is a poor predictor of $\alpha$; all $\alpha > \approx 10^6$ or $\alpha < \approx 1$ generate simulated trees with $\sigma \approx (n - 2)$ or $\sigma \approx 1$, respectively. As trees with $\sigma \leq 3$ are consistent with the hypothesis of constant population size, it is irrelevant that $\sigma$ is a poor estimator of $\alpha$ when $\alpha < 1$. Comparison of the 30-tip and 100-tip curves indicates that using more samples narrows the range of $\alpha$ values which generate uninformative trees with $\sigma \approx (n - 2)$ (see fig. 2).

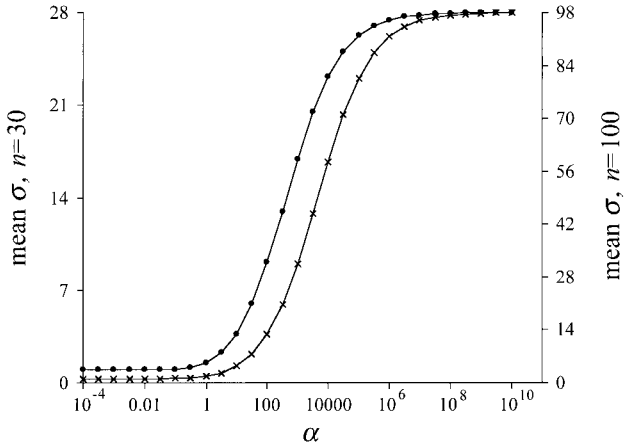


FIG. 2.—The relationship between the tree statistic σ and the demographic parameter α. The ordinates represent the mean σ of 20,000 epidemic coalescent trees simulated using Slatkin and Hudson's (1991) algorithm for each value of α. The curve with circular markers corresponds to the left-hand ordinate (trees with 30 tips). Confidence intervals for each mean σ were calculated using 1.96 × SE(σ) but are not displayed, as they were too small to be noticeable.

Given a reconstructed genealogy, Monte Carlo simulation can be used to calculate a maximum-likelihood estimate (MLE) of α*, the α value of the population from which the genealogy was sampled. Let $\hat{\alpha}$ denote our MLE of α*, and let σ* be the σ value of the reconstructed genealogy. We define $\sigma_x$ as the σ value of an epidemic coalescent tree simulated using Slatkin and Hudson's (1991) algorithm with parameter $\alpha = x$. $P(\sigma_x = \sigma^*)$, the probability that $\sigma_x = \sigma^*$, can be estimated by simulating a large number of epidemic coalescent trees and then calculating the proportion of those trees which satisfy $\sigma_x = \sigma^*$. $P(\sigma_x = \sigma^*)$ is the probability of obtaining result σ* given the hypothesis $\alpha^* = x$ and is therefore proportional to the likelihood of hypothesis $\alpha^* = x$, given the observed σ* (Edwards 1972). Accordingly, $\hat{\alpha}$ is equal to the value of *x* which maximizes $P(\sigma_x = \sigma^*)$. Likelihood curves for $\hat{\alpha}$, which are constructed by calculating $P(\sigma_x = \sigma^*)$ for many values of *x*, become smoother as the number of simulated trees used to calculate each point increases. The likelihood curves are usually unimodal, making location of the MLE straightforward. Conditions under which the likelihood curves become flat are discussed below.

Upper and lower limits of the MLE are calculated using the likelihood ratio test (Wilks 1938; Edwards 1972). Given data *D*, the difference in the support of hypotheses $H_0$ and $H_1$ is $\lambda = \max[\ln L(H_0|D)] - \max[\ln L(H_1|D)]$. If $H_0$ is a subset of $H_1$, then $-2\lambda$ is approximately $\chi^2$ distributed, with degrees of freedom equal to the difference in the number of free parameters between $H_0$ and $H_1$ (Wilks 1938). Here, $H_0$ is the hypothesis that $\hat{\alpha}$ takes a specific value *v* and is a subset of the hypothesis $H_1$ that $\hat{\alpha}$ is free to vary. If $-2\lambda > 3.841$, then *v* lies outside the approximate 95% confidence limits. The validity of using the likelihood ratio test in our method was confirmed by simulation; simulated values of $-2\lambda$ were found to be approximately $\chi^2$ distributed (results not shown).

**Table 1**
**Expected Estimates of α*, Given 30-Tip or 100-Tip Epidemic Coalescent Trees Simulated with α = α***

| | $E[\hat{\alpha}\,|\,\alpha^*]$[a] | |
|---|---|---|
| Log(α*) | *n* = 100 | *n* = 30 |
| 0 | −1.146 | −1.071 |
| 1 | 0.857 | 0.827 |
| 2 | 1.968 | 1.979 |
| 3 | 2.979 | 3.025 |
| 4 | 3.999 | 4.151 |
| 5 | 5.028 | 6.525 |
| 6 | 6.114 | 10.229 |
| 7 | 8.077 | 12.608 |
| 8 | 11.236 | 13.595 |

[a] Calculated using a C++ program available from the authors on request.

## Estimates of α* Are Sometimes Biased

All estimates of exponential growth rate based on phylogenetic methods are expected to be biased (Kuhner, Yamato, and Felsenstein 1998). We therefore performed extensive simulations to test the accuracy of the mid-depth method. These simulations involved the calculation of $E[\hat{\alpha}\,|\,\alpha^*]$, the expected value of $\hat{\alpha}$ given epidemic coalescent trees simulated with $\alpha = \alpha^*$. If the mid-depth method is biased upward, then $E[\hat{\alpha}\,|\,\alpha^*]$ will be larger than α*.

For trees with *n* tips, $E[\hat{\alpha}\,|\,\alpha^*] = (1/[n-2]) \sum_{i=0}^{n-2} P(\sigma_{\alpha^*} = i)\cdot\hat{\alpha}(i)$. As before, $P(\sigma_{\alpha^*} = i)$ is the probability that the σ value of an epidemic coalescent tree (simulated with $\alpha = \alpha^*$) is equal to *i*. $\hat{\alpha}(i)$ is the mid-depth method MLE of α* given a tree with $\sigma = i$. Table 1 shows the simulation results. As expected, the mid-depth method shows a significant positive bias, but only when log(α*) exceeds a certain value. This value is higher, and the range of unbiased estimation wider, when more samples are used. Comparison of table 1 with figure 2 shows that the range of unbiased estimation approximately coincides with the range of the linear relationship between σ and α. There is also a negative bias when $\log(\alpha^*) \leq 1$. Analysis of the simulation results revealed that the positive bias is almost entirely due to α* being overestimated when $\sigma \approx (n-2)$. If α* is sufficiently large, then all simulated trees have $\sigma \approx (n-2)$, thus causing the likelihood curve for $\hat{\alpha}$ to be flat on one side (see fig. 2). In other words, all sufficiently large values of α* are equally likely explanations of the observed σ; hence, the MLE of α* is biased upward. The negative bias found when α* is small is caused by the same effect; if $\sigma \approx 1$, the part of the likelihood curve representing low α* values becomes flat, biasing the MLE of α* downward. Therefore, $\hat{\alpha}$ is unlikely to be biased unless one of the following circumstances occurs: (1) the population has increased in size very rapidly (hence, $\sigma \approx [n-2]$ is more likely), (2) the population has been decreasing in size ($\sigma \approx 1$ is more likely), (3) *n* is small (there are fewer possible σ values; hence, both $\sigma \approx [n-2]$ and $\sigma \approx 1$ are more likely).

**Table 2**
**Details of the HIV-1 Data Sets**

| Data Set | | Length of Sequences | No. of Sequences | | | Confidence Limits of $\hat{\alpha}$[a] | |
|---|---|---|---|---|---|---|---|
| Subtype | Gene | (bp) | | $\sigma$ | $\hat{\alpha}$[a] | Lower | Upper |
| A | *env* | 600 | 46 | 13 | 178.649 | 39.355 | 734.514 |
| | *gag* | 1,410 | 27 | 10 | 163.305 | 27.989 | 818.465 |
| B | *env* | 2,202 | 54 | 46 | 84,333.5 | 13,740.4 | 636,795 |
| | *gag* | 1,380 | 28 | 24 | 49,888.5 | 4,275.63 | 2,182,730 |
| | *pol* | 1,677 | 25 | 22 | 149,279 | 5,741.16 | 51,880,000 |

[a] Calculated using a C++ program available from the authors on request.

## Example: The Epidemic Growth of HIV-1

To illustrate the use of the mid-depth method, we applied it to HIV-1 sequences sampled from around the world. Phylogenetic analysis has been used to classify the virus into a series of subtypes, A–J, which have differing geographical distributions (see McCutchan et al. 1996). Two of the most common subtypes (and those for which the most sequences are available) are subtype A, which is most commonly found in sub-Saharan Africa and in migrants from that region, and subtype B, which mainly circulates in the developed world. Two previous analyses of these subtypes using LTT plots indicated that both have spread at a roughly constant exponential rate, with no evidence of a difference in growth rate between them (Holmes et al. 1995; Holmes, Pybus, and Harvey 1999).

To examine whether the mid-depth method provides a different picture of population dynamics, we collected all available *env*, *gag*, and *pol* gene sequences from the Los Alamos HIV database (Myers et al. 1996). Six data sets were constructed, one for each gene from each subtype. Complete (or nearly complete) sequences were used if more than 20 were available; otherwise, the largest region of the gene found in approximately 90% of the sequences was used. Insufficient *pol* sequences from subtype A were available for analysis. The sequence alignments available in the database were used for all data sets.

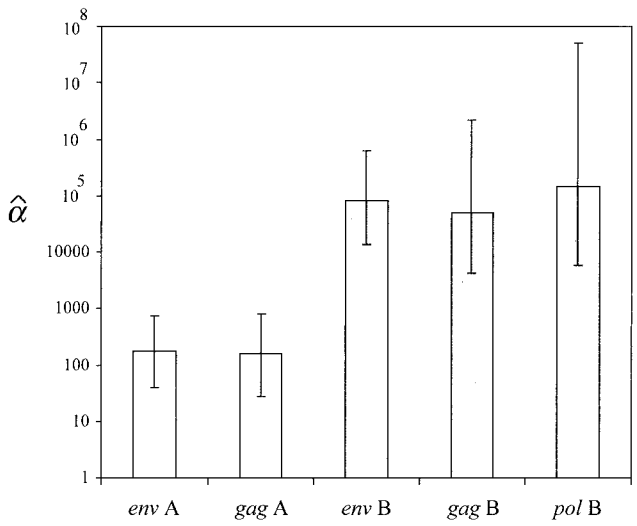


FIG. 3.—Values of $\hat{\alpha}$ for each of the HIV-1 data sets, estimated using the mid-depth method. The error bars represent confidence limits calculated using the likelihood ratio test (see text).

Biases in sample composition affect the reconstruction of population dynamics; an overrepresentation of sequences from very closely related individuals will bias the sample toward recent coalescence events. Therefore, sequences were removed from each data set if they came from the same patient or were related by direct transmission to another sequence in the alignment. This information was obtained from the primary literature. Furthermore, all sequences previously identified as intersubtype recombinants were also removed. Details of the final data sets used are given in table 2.

Genealogies were estimated for each data set using a maximum- likelihood approach. Tree likelihoods were computed using the Hasegawa-Kishino-Yano (HKY-85) substitution model and a codon- position-specific model of rate heterogeneity, under the assumption of a constant-rate molecular clock. Due to the large data sets used, simultaneous coestimation of the tree topology and model parameters was not possible. Therefore, the model parameters were estimated on an initial neighbor-joining tree. Tree topologies were then evaluated using a heuristic search approach which implemented both tree bisection-reconnection and nearest-neighbor interchange perturbations. Finally, the branch lengths of the maximum-likelihood topology were re-estimated using the general reversible (REV) substitution model. The final phylogenies are shown in figure 4. All analyses were performed with a beta test version of PAUP4 kindly provided by David Swofford. Substitution model parameter values and sequence accession numbers are available from the authors on request.

The End-Epi computer package (Rambaut, Harvey, and Nee 1997) was used to analyze the LTT plots of the reconstructed genealogies. In agreement with previous LTT analyses of HIV-1 (Holmes et al. 1995; Holmes, Pybus, and Harvey 1999), all of the plots indicated that the genealogies had been sampled from populations growing at a constant exponential rate (results not shown). All five reconstructed trees had $\sigma > 3$, and therefore the hypothesis of constant population size was rejected for all data sets (see table 2). The mid-depth method was used to calculate $\hat{\alpha}$ and its confidence limits for each data set (see fig. 3 and table 2). It is clear from figure 3 that HIV-1 subtypes A and B have very different values of $\hat{\alpha}$; the $\hat{\alpha}$ values of subtype B are two to three orders of magnitude larger than are those of sub-

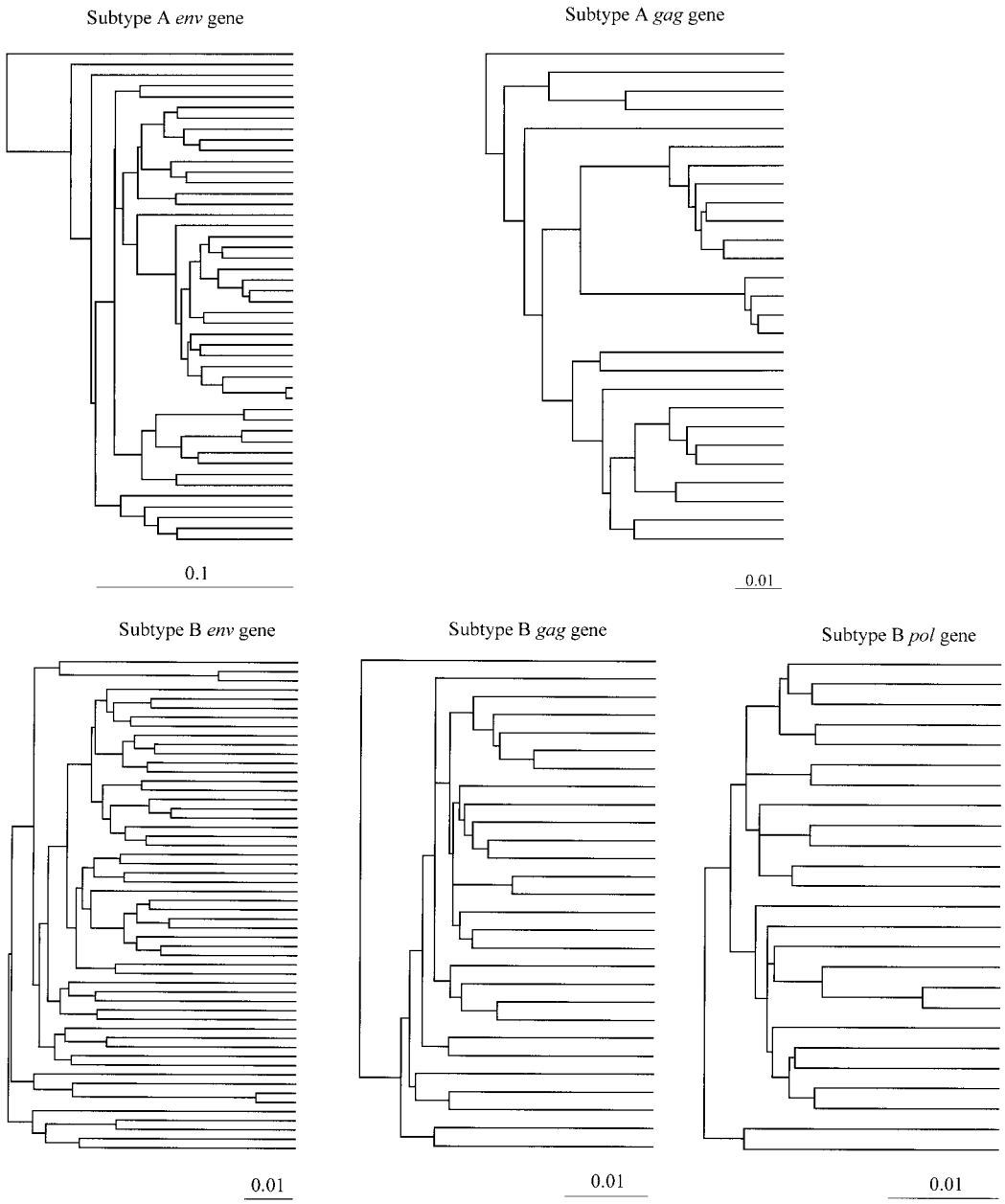


FIG. 4.—Estimated maximum-likelihood phylogenies for each of the five HIV-1 data sets. The scale bars represent expected number of substitutions per nucleotide.

type A. This difference is seen in both the *env* and the *gag* genes and is consistent for all genes within a subtype.

## Interpretation of the HIV-1 Results

The higher $\hat{\alpha}$ values for subtype B suggest that this subtype has either a larger population size ($N_0$) or a higher exponential growth rate ($r$) than subtype A. The former explanation seems unlikely—it is estimated that 7.5% of individuals in sub-Saharan Africa aged 15–49 years are HIV infected, compared with only 0.3% in Western Europe and 0.6% in North America (UNAIDS and WHO 1997). Sub-Saharan Africa is home to two thirds of the world's HIV-infected people (Mann and Tarantola 1998), and it is reasonable to assume that many of these infections are due to subtype A viruses.

So, do contrasting exponential growth rates explain the different $\hat{\alpha}$ values of subtypes A and B? Two important points must be considered here. First, the growth rate parameter $r$ represents an intrinsic rate of increase—the number of new infections per infected individual per unit time. Hence, the huge number of new infections currently reported in sub-Saharan Africa could be the consequence of a high prevalence rather than a high growth rate. Second, our estimated $\hat{\alpha}$ values reflect the growth rates of subtypes A and B during the start of their spread through human populations; as pointed out by Slatkin and Hudson (1991), random samples of se-

quences are unlikely to contain information about recent decreases in growth rate. Moreover, many of our samples come from individuals who were infected early during the HIV-1 epidemic.

It is clear that the growth rate of the virus when it first emerged in North America and Europe around 20 years ago was much higher than it is today. At that time, before the advent of widespread behavioral intervention, transmission was predominantly localized to high-risk groups, notably homosexual men and injecting drug users. Within these groups, rates of contact or partner exchange were often so high that a fully connected chain of transmission—a ''standing network''—was formed, allowing the virus to spread quickly through a large number of individuals (Jacquez et al. 1994; Watts and May 1992). This process was aided by the fact that many individuals within the network were at their most infectious, having only recently been infected. Such a transmission pattern would generate very high growth rates.

In contrast, most transmissions in Africa occur through unprotected heterosexual intercourse, so that the average waiting times between transmission events are longer than those experienced in standing networks (Tarantola and Schwärtlander 1997). Thus, subtype A's initial growth rate was probably considerably slower than that of subtype B. If we accept that subtype A has a greater population size, then the early growth rate of subtype B must have been at least two orders of magnitude faster than that of subtype A.

### Have the Assumptions of the Method Been Met?

We will now consider the robustness of our results to possible violations of the mid-depth method assumptions. Assuming that the genealogies are correct, we can identify four major factors which will affect the interpretation of our results: (1) the sample is not phylogenetically random, (2) the sequences have recombined, (3) the sequences have been subject to selection, and (4) the sequences have not evolved at a constant rate.

Samples collected from epidemiologically linked patients will be biased toward recently diverged sequences, leading to underestimates of population growth rates. We minimized this problem by removing sequences from our analysis that are clearly from the same source. Unless sequence data are collected in a phylogenetically random manner, this post hoc approach is likely to be the most efficient way to control for sampling bias.

Recombination and natural selection, which are both common occurrences in HIV-1 (Leigh Brown and Holmes 1994; Robertson et al. 1995), significantly complicate the interpretation of our results. However, of singular importance here is whether these two processes occur at different rates in subtypes A and B. Although this is a difficult question to answer, there is no reason to believe that the observed intersubtype differences are due to recombination or natural selection. At present, there is no strong evidence of subtype-level natural selection (see, e.g., Pope et al. 1997) or of high levels of intrasubtype recombination for either subtype, although this surely occurs. Furthermore, all sequences known to be intersubtype recombinants were removed from our analysis, and our results were consistent across all genes, suggesting that they represent populationwide phenomena.

Consideration of the constancy, or inconstancy, of the HIV-1 molecular clock has recently increased following attempts to date the origin of the virus using a strain isolated in 1959 (Korber, Theiler, and Wolinsky 1998; Zhu et al. 1998). Although there is rate variation among different HIV strains (Fauci 1996), estimates of HIV origins based wholly on contemporary sequences are usually close to those made using the ''fossil'' HIV-1 strain (see Kasper et al. 1995). This suggests that short-term rate fluctuations are not necessarily translated into long-term rate heterogeneity between subtypes. More pertinently, there is no evidence that subtypes A and B differ in substitution rate so as to cause the differences observed here.

## Discussion

The mid-depth method provides a straightforward approach to testing hypotheses of population dynamic history. It differs from most other methods in using reconstructed genealogies, which are assumed to be correct, as its raw data. This assumption has benefits and costs. Computational complexity is greatly reduced by separating the task of phylogenetic reconstruction from that of demographic hypothesis testing. This separation also increases the method's flexibility—for each data set, the most apt tree reconstruction procedure can be chosen without modification of the mid-depth method. In addition, our method analyses relative (rather than absolute) interval sizes, allowing demographic parameters to be estimated without knowledge of the sample's mutation rate (*sensu* Nee et al. 1995). However, parameter estimates and their confidence limits do not contain information about the uncertainty of the genealogy. Further work is needed to incorporate this uncertainty into the mid-depth method.

## Acknowledgments

Many thanks to Andrew Rambaut, Nick Grassly, Mike Charleston, John Huelsenbeck, and Joe Felsenstein for technical assistance and inspirational discussion. This work was supported by the Wellcome Trust (grant 050275) and The Royal Society.

## LITERATURE CITED

EDWARDS, A. W. F. 1972. Likelihood. Cambridge University Press, Cambridge, England

FAUCI, A. S. 1996. Host factors and the pathogenesis of HIV-induced disease. Nature **384**:529–534.

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

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

HARPENDING, H. C., S. T. SHERRY, A. R. ROGERS, and M. STONEKING. 1993. The genetic structure of ancient human populations. Curr. Anthropol. **34**:483–496.

HOLMES, E. C., S. NEE, A. RAMBAUT, G. P. GARNETT, and P. H. HARVEY. 1995. Revealing the history of infectious disease epidemics through phylogenetic trees. Philos. Trans. R. Soc. Lond. B Biol. Sci. **349**:33–40.

HOLMES, E. C., O. G. PYBUS, and P. H. HARVEY. 1999. The molecular population dynamics of HIV-1. Pp. 177–207 *in* K. A. CRANDALL, ed. The evolution of HIV Johns Hopkins University Press, Baltimore, Md.

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 JR. 1994. Role of primary infection in epidemics of HIV infection in gay cohorts. J. Acqir. Immune Defic. Syndr. **7**: 1169–1184.

KASPER, P., P. SIMMONDS, K. E. SCHNEWEIS, R. KASIER, B. MATZ, J. OLDENBURG, H.-H. BRACKMANN, and E. C. HOLMES. 1995. The genetic diversification of the HIV-1 *gag* p17 gene in patients infected from a common source. AIDS Res. Hum. Retroviruses **11**:1197–1201.

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

KORBER, B., J. THEILER, and S. WOLINSKY. 1998. Limitations of a molecular clock applied to considerations of the origin of HIV-1. Science **280**:1868–1871.

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.

———. 1998. Maximum likelihood estimation of population growth rates based on the coalescent. Genetics **149**:429–434.

LEIGH BROWN, A. J., and E. C. HOLMES. 1994. The evolutionary biology of human immunodeficiency virus. Annu. Rev. Ecol. Syst. **25**:127–165.

MCCUTCHAN, F. E., M. O. SALMINEN, J. K. CARR, and D. S. BURKE. 1996. HIV-1 genetic diversity. AIDS **10**:S13–S20.

MANN, J. M., and D. TARANTOLA. 1998. HIV 1998: the global picture. Sci. Am. **279**:82–83.

MYERS, B., B. KORBER, B. FOLEY, K.-T. JEANG, J. W. MELLORS, and S. WAIN-HOBSON, eds. 1996. Human retroviruses and AIDS 1996. Theoretical Biology and Biophysics Group, Los Alamos National Laboratory, Los Alamos, N.M.

NEE, S., E. C. HOLMES, A. RAMBAUT, and P. H. HARVEY. 1995. Inferring population history from molecular phylogenies. Philos. Trans. R. Soc. Lond. B Biol. Sci. **349**:25–31.

POPE, M., D. D. HO, J. P. MOORE, J. WEBER, M. T. DITTMAR, and R. A. WEISS. 1997. Different subtypes of HIV-1 and cutaneous dendritic cells. Science **278**:786–787.

RAMBAUT, A., P. H. HARVEY, and S. NEE. 1997. End-Epi: an application for inferring phylogenetic and population dynamical processes from molecular sequences. Comput. Appl. Biosci. **13**:303–306.

ROBERTSON, D. L., P. M. SHARP, F. E. MCCUTCHAN, and B. H. HAHN. 1995. Recombination in HIV-1. Nature **374**:124–126.

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

SANDERSON, M. J. 1997. A nonparametric approach to estimating divergence times in the absence of rate constancy. Mol. Biol. Evol. **14**:1218–1231.

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

TAJIMA, F. 1989. The effect of change in population size on DNA polymorphism. Genetics **105**:585–595.

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

UNAIDS and WHO. 1997. HIV/AIDS: the global epidemic. http://www.unaids.org.

WATTS, C. H., and R. M. MAY. 1992. The influence of concurrent partnerships on the dynamics of HIV/AIDS. Math. Biosci. **108**:89–104.

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

ZHU, T. F., B. T. KORBER, A. J. NAHMIAS, E. HOOPER, P. M. SHARP, and D. D. HO. 1998. An African HIV-1 sequence from 1959 and implications for the origin of the epidemic. Nature **391**:594–597.

MICHAEL HENDY, reviewing editor

Accepted March 22, 1999