THE ROYAL SOCIETY

doi 10.1098/rstb.2001.0864

The origin of acquired immune deficiency syndrome: Darwinian or Lamarckian?

Tom Burr[1*], **J. M. Hyman**[2] **and Gerald Myers**[3]

[1]*Mail Stop E541,* [2]*Mail Stop B284, and* [3]*Mail Stop M888, Los Alamos National Laboratory, Los Alamos, NM 87545, USA*

The subtypes of human immunodeficiency virus type 1 (HIV-1) group M exhibit a remarkable similarity in their between-subtype distances, which we refer to as high synchrony. The shape of the phylogenetic tree of these subtypes is referred to as a sunburst to distinguish it from a simple star phylogeny. Neither a sunburst pattern nor a comparable degree of symmetry is seen in a natural process such as in feline immunodeficiency virus evolution. We therefore have undertaken forward-process simulation studies employing coalescent theory to investigate whether such highly synchronized subtypes could be readily produced by natural Darwinian evolution. The forward model includes both classical (macro) and molecular (micro) epidemiological components. HIV-1 group M subtype synchrony is quantified using the standard deviation of the between-subtype distances and the average of the within-subtype distances. Highly synchronized subtypes and a sunburst phylogeny are not observed in our simulated data, leading to the conclusion that a quasi-Lamarckian, punctuated event occurred. The natural transfer theory for the origin of human acquired immune deficiency syndrome (AIDS) cannot easily be reconciled with these findings and it is as if a recent non-Darwinian process took place coincident with the rise of AIDS in Africa.

Keywords: HIV; AIDS; phylogenetic analysis; coalescent theory; synchronization

1. INTRODUCTION

An understanding of the origin of acquired immune deficiency syndrome (AIDS) must look primarily to Africa and the recorded history of the epidemic(s). Specifically, it must rely on the clinical and epidemiological records, the findings from retrospective serosurveys, and our general knowledge of primate immunodeficiency viruses (PIVs) and their hosts (Hooper 1999). Any molecular account or hypothesis concerning the evolution of the human immunodeficiency viruses (HIVs) must also be in accord with this body of macroscopic data (Myers 1994).

Look-back studies based solely on phylogenetic analysis of molecular sequence data remain one of the mainstream approaches to reconstruction of viral evolution. Such studies tend to assume that PIV was transferred into humans in a natural and gradual way and that HIV then evolved through a continuous Darwinian process. A recent prominent analysis (Korber *et al.* 2000*a*) arrived at a look-back date for the HIV-1 group M not easily reconciled with the macroscopic data. Our purpose in this paper is to set forth a contrasting 'punctuated origin' theory for the recent emergence of the HIV-1 group M viruses, which are largely responsible for the AIDS pandemic. This approach suggests that PIV's entry into humans might be no more ancient than what is to be inferred from the historical rise of AIDS in Africa in the 1970s.

Our paper falls into two parts. In the first part we discuss the recent look-back study (Korber *et al.* 2000*a*). We argue in this context that theoretical look-back studies that do not take population dynamics into account are inadequate for making definitive judgements about the date of entry of PIV into humans. Furthermore, phylogenetic tree analyses of group M viruses reveal an extraordinary phylogenetic pattern, namely a star phylogeny of star phylogenies. We call this pattern a 'sunburst'. The sunburst pattern of the group M viruses signifies a synchronized entry of the ten or so subtype viruses into human populations. Such a phylogenetic pattern is not observed, for example, with feline immunodeficiency viral (FIV) sequences (Bachmann *et al.* 1997), whose phylogenetic pattern we take to be paradigmatic of a natural evolutionary process. The fact that there were ten or so synchronous but distinguishable African epidemics is a definitive feature of AIDS for which the natural transfer theory gives no convincing account. Just as perennial influenza epidemics are thought of as a single epidemic for convenience, African AIDS is often thought of as a single phenomenon merely for convenience.

In the second part of this paper, our analysis turns to the effects that population dynamics has on the pattern of diversification of the virus in an epidemic (or epizootic). Here we use 'forward-process' simulations rather than 'look-back' studies to explore the conditions under which a sunburst phylogenetic pattern involving ten synchronous clades might arise. These forward-process simulations can be viewed as reversing a look-back study to determine likely outcomes under various conditions of mutation rate, population size, transmission rate and so forth.

The general theoretical conditions for achieving star-like phylogenies have been described by Slatkin & Hudson (1991). Following their lead, we employ coalescent theory (Kingman 1982; Fu & Li 1999) first to determine the likely number of star-like clades attainable

[*] Author for correspondence (tburr@lanl.gov).

through a natural Darwinian (and panmictic) process. Simulations of forward processes under representative conditions typically generate two to five star-like clades. This number is significantly smaller than is seen with the HIV-1 group M but consistent with what is seen with the FIV paradigm.

Next we use coalescent theory to quantitatively evaluate the apparent synchronization of the group M subtypes also using coalescent theory. To the degree that the onset of diversification of the numerous group M subtypes (designated A–J) is synchronized and, of necessity, coincident with the various onsets of the macroscopic epidemics, a punctuated event is implied. It is useful in this context to refer to the Luria–Delbruck statistical test for a Darwinian process (Luria & Delbruck 1943). Under the hypothesis of a gradual, natural introduction of PIV over an extended period in the first half of the 20th century, the variance associated with either the inter- or intracladal distances (the so-called summary statistics), should exceed what would be expected from a Poisson process. That is to say, one would not expect synchrony in a natural Darwinian process but would rather expect an asymmetrical pattern of diversification such as is seen in the case of FIV. On the other hand, under the hypothesis of a punctuated event, it would be as if a Lamarckian event took place: the variance of the distances within and between subtypes (clades) would be statistically small in our forward-process simulations. This is precisely what is seen for large samples of HIV-1 group M *env* gp120 and *gag* p17 sequences. The implication of this modelling study is that the origin of HIV-1 group M viruses embodies a statistically extreme, non-Darwinian, synchronized component—in contrast to the original Luria–Delbruck findings for bacterial resistance to virus. We call this exceptional synchronizing event the 'punctuated origin' of the HIV-1 epidemic.

2. LOOK-BACK STUDIES

Phylogenetic tree analyses applied to some of the earliest-known HIV sequences provided minimal look-back estimates for the most recent common ancestor (MRCA) of the human AIDS viruses (Smith *et al.* 1988; Sharp & Li 1988). These studies produced crude estimates that did not take into account variable rates of change across sites, recombination and homoplasy (Eigen & Nieselt-Struve 1990; Swofford *et al.* 1996). The work of Korber *et al.* (2000*a*) brings far greater sophistication to such enquiries, starting with the implementation of parallelized maximum-likelihood analysis. Using the Nirvana supercomputer at Los Alamos National Laboratory, this work produced a look-back estimate to 1930 (plus or minus 20 years) for the M group of HIV-1. That is to say, the study dates the MRCA somewhere between 1910 and 1950.

As the authors themselves acknowledge, the supercomputer-based study cannot tell whether this hypothetical 1930 virus was in humans or animals and so does not show when zoonosis occurred. On the other hand, by assuming that the evolutionary process moving forward from the look-back date is continuous, the authors implicitly assume that the MRCA is an HIV rather than a PIV. Thus, an appropriately qualified summary of the Korber *et al.* (2000*a*) look-back study would be the following: if PIV was in humans in the first half of the 20th century, it may be estimated, given the assumptions of the look-back analysis, that the ancestral HIV-1 group M virus arose at 1930 plus or minus 20 years.

Clinical, serological and molecular retrospective studies have all failed to produce any evidence of AIDS or HIV prior to the 1970s, with the exception of a single 1959 sample (Hooper 1999; Nahmias *et al.* 1986; Zhu *et al.* 1998; Myers *et al.* 1993). In the absence of clinical or virological data arising prior to 1959, this study supports only this hypothetical conclusion.

Korber *et al.*'s (2000*a*) study does not make such a qualified statement, and the study has been widely assumed to prove that HIV was in humans in 1930. This conclusion, in turn, has been taken as evidence that the oral polio vaccine (OPV) trials of the late 1950s could not have been the occasion of zoonosis. Indeed, Korber *et al.* (2000*a*) argue at the end of their paper that the OPV hypothesis is an unlikely explanation for the rise of AIDS in Africa. Their arguments for this conclusion are consistent with their study but do not arise directly from it. Given the 1959 sample, it is to be expected (although not logically required) that an ancestral virus would exist prior to the OPV trials of 1957–1959, hence we gain no essentially new information bearing upon the OPV debate directly from this look-back study.

In opposing the OPV hypothesis the authors fall back on the natural theory of zoonosis (the 'cut-hunter' hypothesis), which they take to be the likely alternative. They do not, however, show how such an account adequately explains several defining features of the AIDS epidemic. For example, the 1930 look-back date postulates a 40-year 'pre-epidemic' period during which HIV was present in human populations but was not clinically recognized and for which retrospective serosurveys give no evidence. Korber and colleagues also note that their analysis suggests that 'significant diversification' of HIV-1 group M occurred during this period (Korber *et al.* 2000*a*). The inferred diversification is at odds with the absence of AIDS in that period. In a time-calibrated tree, such as that in the Korber analysis (Korber *et al.* 2000*a*, fig. 4, p. 1793), conspicuous diversification of the group M subtypes coincides, as should be expected (Holmes *et al.* 1999), with the temporal onset of the epidemics in the 1970s. Viral diversification and epidemicity need not go hand-in-hand, but it is difficult to imagine any other human scenario in this case (Holmes *et al.* 1999).

The Korber *et al.* analysis suggests that significant diversification arose during the pre-epidemic period and that ten or so 'pre-epidemics' marched forward in step. Phylogenetic tree analyses of group M sequences invariably show a synchrony in the onset of diversification of the African subtype sequences. Synchrony of this scale has been loosely attributed to urbanization and exponential growth of the epidemic. However, the FIV tree, for which urbanizing forces and rapid epidemiological expansion must also have been in effect (figures 1 and 2), does not show such synchrony.

A major challenge to the natural origin hypothesis of human AIDS—with a 1930 start date—is to account for the extraordinary synchrony in the 1970s of ten or more distinguishable epidemics. How is it possible for so many evolving subtypes to remain in step during a 40-year pre-epidemic period (1930–1970)? Where are the 'jackpots' that signify a Darwinian process (Luria & Delbruck 1943)? We shall attempt to address these questions in terms of coalescent theory (Kingman 1982; Fu & Li 1999). It should be noted here that the intention of the analyses that follow in this paper is not to argue either for or against the OPV hypothesis in particular. It is rather to explore certain features of the AIDS epidemic not easily accounted for by the natural transfer hypothesis and to suggest how a punctuated origin theory might account for those features.

3. A SIMULATED NATURAL PROCESS

(a) *Strategy*

In § 2 we made two observations: (i) there are approximately ten subtypes of HIV-1 group M, implying a minimum of ten distinguishable epidemics; and (ii) the diversification of these subtypes along with their separate epidemics appears to be synchronized. We then raised the question as to whether such multiplicity and synchrony are expected to arise from a natural process. In this section we use forward-process models to investigate how many subtypes tend to emerge through various natural processes and how synchronized the emergence of the subtypes tends to be. These models will include the growth rate and dynamics of the macroscopic epidemics as well as microscopic mutation variables. If the ten or so nearly equidistant HIV-1 subtypes can emerge for certain combinations of the macroscopic and microscopic parameters, then the extraordinary configuration of the group M subtypes will have been shown to be consistent with a natural Darwinian process.

For this and other tests, coalescent theory is employed to model a natural process that mimics HIV evolution by using available and inferred epidemiological data and a molecular substitution model obtained from best current estimates using the *env* gp120 and *gag* p17 regions of the genome (Leitner *et al.* 1997). Coalescent theory, as implemented in Treevolve (Grassley *et al.* 1999), allows us to simulate the evolution of *env* and *gag* sequences (approximately 100 sequences each) by obtaining a representative sample genealogy using the approximating probability distribution for the times that each lineage merges (coalesces). Given the genealogy of the sample, it is straightforward to apply the substitution model to a hypothetical ancestral sequence to generate the 100 sequences.

Coalescent theory involves a minimum of two microscopic and three macroscopic parameters. With respect to the microscopic aspects of the model, we allow the overall rate to vary across sites, invoking a gamma distribution ($\gamma = 0.40$ for *env* and 0.45 for *gag*) to describe the rate variation across sites (Leitner *et al.* 1997). We then specify: (i) an overall substitution rate, μ, which is the probability per unit time of substitution (for the general case, the estimated rate is 0.0035 for *env* gp120 and 0.0026 for *gag* p17 (see Appendix A)); and (ii) a general time-reversible substitution model (GTR), which permits some changes to be more likely than others (Swofford *et al.* 1996; Leitner *et al.* 1997).

The same substitution model for the real and simulated data is used. Therefore, it is not critical to use 'a best possible model' (assuming such exists) because systematic errors from using a suboptimal model to estimate distances in both the real and simulated data will approximately cancel when we compare the real summary statistics with the reference summary statistics (Appendix A). However, because GTR + γ is available in PAUP* and relatively simple to use, there was no reason not to use as realistic a model as possible (Leitner *et al.* 1997; Swofford 1999). Homoplasy is addressed to some degree through this approach but recombination is not: our reasoning is that complexities of recombination would tend more often than not to negate star-like phylogeny, whereas the task herein is to account for the extraordinary star-like configuration (Grassley *et al.* 1999). For further details concerning the microscopic variables, see Appendix A.

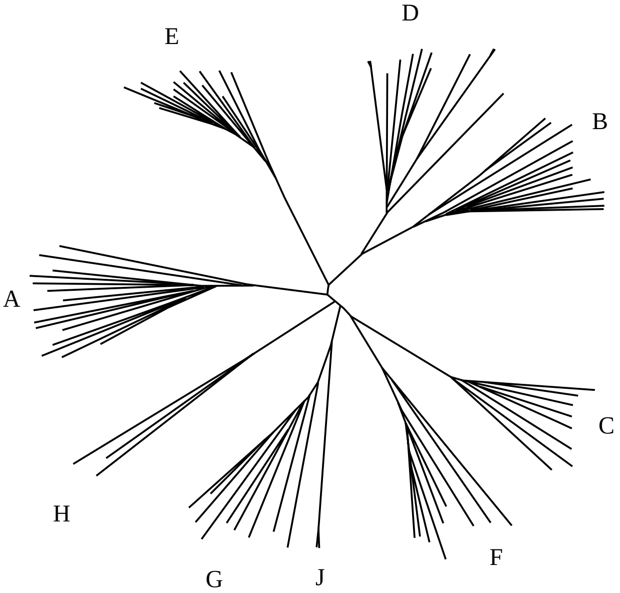

Figure 1. Phylogenetic tree of 100 representative *env* (gp120 region) sequences with nine distinct subtypes, denoted A–J (no I subtypes are included).

The macroscopic model includes these three major features. (i) Epidemic growth rate (figure 3) as characterized by the number of infected HIV cases (prevalence) over time, from approximately 1970 to 2000. This is expressed as $N = N_0 e^{rt}$ for each of several periods. More specifically, $N(t) = N_i e^{r_i(t-t_i)}$ for t in $[t_i, t_{i+i}]$. Here, N_i is the population size at the beginning of time-period i with rate r_i, and r_i can be negative, zero or positive. Treevolve conveniently provides for piecewise application of exponential epidemic properties. The best statistical evaluation of the early epidemic (taken in its totality as in figure 3) calls for a quadratic process (K. A. Stanecki, personal communication 2000), which is approximated as a piecewise exponential. (ii) Generation time g (imagine each 'crop' of cases dies off and is replaced by a new crop each generation, so that g is *ca.* 1/(infection period)). (iii) Variance, σ^2, in the number of 'offspring' (new cases) that each case produces (some produce zero new cases, others produce many new cases). We use 0.05 as the nominal value of the composite parameter g/σ^2 (Grassley *et al.* 1999).

It follows from coalescent theory that the effective population size is $N_{\text{effective}} = Ng/\sigma^2$. Because time is measured in units of generations, we expect that $N_{\text{effective}}$ would be proportional to N and g. It is not obvious why $N_{\text{effective}}$ is inversely proportional to σ^2; However, for

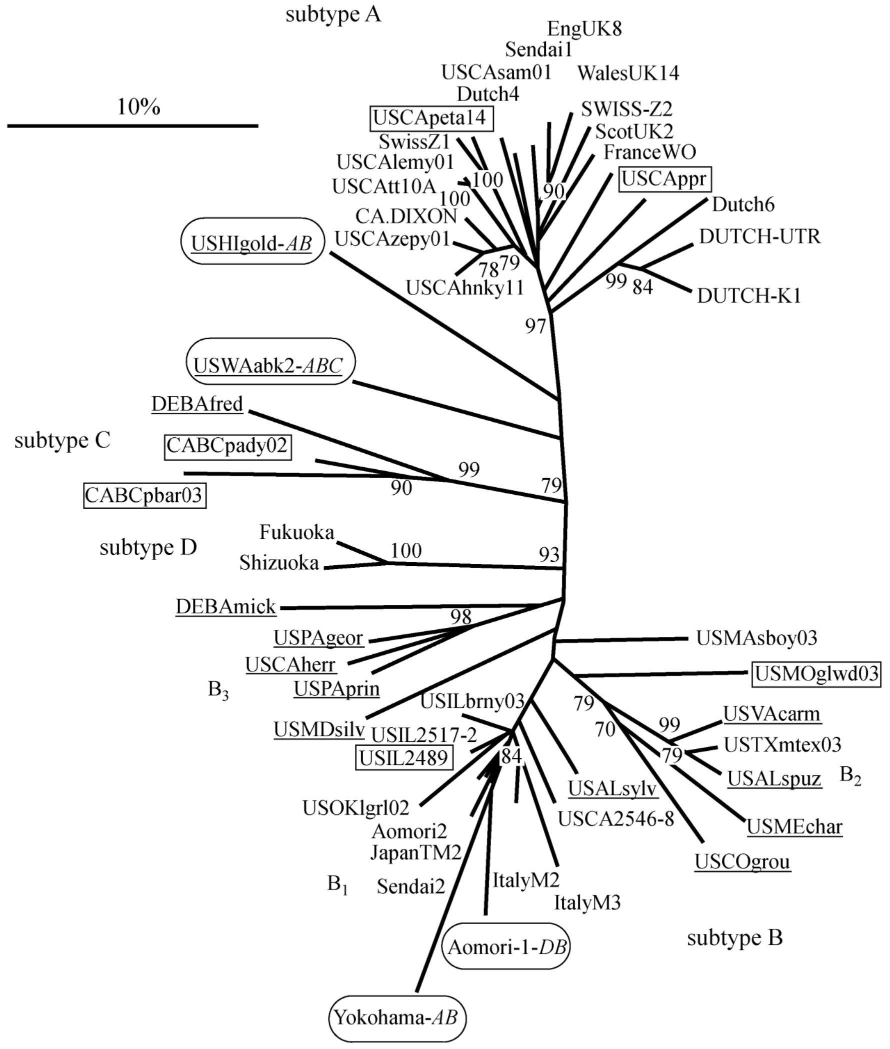

Figure 2. Phylogenetic tree of FIV sequences (Bachmann *et al.* 1997). (Reproduced with permission of Jim Mullins and the *Journal of Virology*.)

larger σ^2, there is a larger probability that two HIV cases arose from the same HIV case in the preceding generation.

To explore parameter space through many simulations, we allowed three macroscopic parameters, N_0, g/σ^2, and the pre-1970 growth rate, to each assume a 'low' and a 'high' value. The same microscopic mutation model is applied in all cases (GTR $+ \gamma$), and we empirically verified that the two microscopic parameters that have the most impact are the average substitution rate μ and rate heterogeneity parameter γ, giving a total of five parameters (three macro and two micro) to vary. Each case is repeated for *gag* and *env*, and several simulations are run per case to provide a reference distribution.

The above-mentioned values for the macroscopic and microscopic parameters are referred to as the nominal values and in the exploration of parameter space they are varied from 0.5 (L = low) to 2.0 (H = high) times their nominal values for N_0, γ, and μ, and from 0.1 (L = low) to 10 (H = high) times their nominal values for g/σ^2 and the pre-1970 growth rate.

When assessing the number of subtypes to be expected in a growing HIV epidemic, considerable attention is given to the specification of the evolutionary model and the associated distance measure. We used a relatively fast and reliable method that could be calibrated to reproduce the number of subtypes in the real data (Burr *et al.* (2000) and Appendix A).

For the purpose of assessing the synchrony of the group M subtypes, an operational definition of synchrony is proposed and extensive comparisons are made between the real sequence data and the simulated data sets. Following the work of Grassley *et al.* (1999) we use summary statistics that describe the distances between all pairs of sequences. Specifically, the summary statistics we use are the estimated number of subtypes (clades) and the averages and standard deviations of the intracladal distances and intercladal distances. We do the same for the simulated data and report where the observed summary statistics lie on the simulated reference distribution.

There is a parallel of this approach to the classical Luria–Delbruck fluctuation test (Luria & Delbruck 1943). Luria and Delbruck recognized that the variance-to-mean ratio of the number of mutant bacterial forms provided a way to discriminate between two candidate forward models, a Darwinian model and a Lamarckian model. Here, we hypothesize that the observed between-group distances are 'too similar' (have a small standard deviation, s_{between}) to have arisen from a forward model that does not have some type of synchronization event. To test

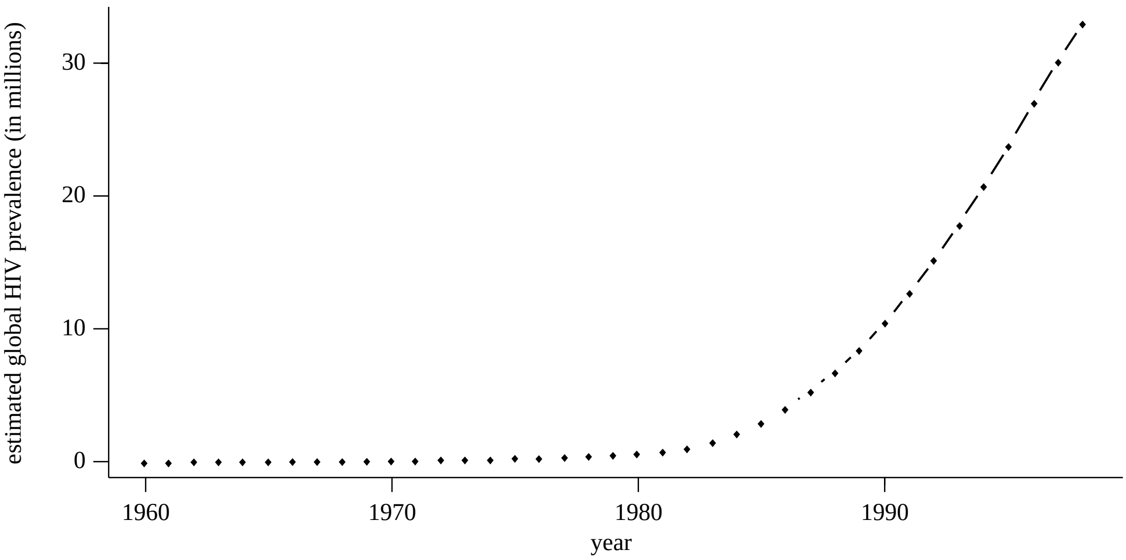

Figure 3. Estimated global HIV prevalence in millions by year.

this hypothesis, one set of the five parameter values (three macro- and two microscopic as already described) will define a case and multiple runs per case will be performed for *gag* and *env.* We then use the summary statistics from the simulated data (for each case separately) as a reference distribution against which to compare the observed summary statistics for the actual *gag* and *env* sequences. If the observed data are in either extreme of the reference distribution (for example, at less than the 0.05 quantile or greater than the 0.95 quantile), then we conclude that the forward model used to simulate the data through a natural process is not a credible explanation for the observations.

Specifically, our approach involves three main steps. (i) Compute summary statistics (mainly the within-subtype average distance d_{within} and the standard deviation s_{between} of the between-subtype distances) for group M *gag* p17 and *env* gp120 sequences. (ii) Simulate sequence data using coalescent theory under forward model A for *gag* and B for *env,* with a quadratically growing epidemic from 1970 to 1990. (Pre-1970 is treated as a stationary period.) The forward model must specify the time dependence of the number of HIV cases and the mutation model; models A and B differ only in the microscopic properties. (iii) Use data simulated for a given case of model A as a reference distribution against which to compare the observed d_{within} and s_{between} for *gag,* and repeat for model B for *env.* If the observed d_{within} and s_{between} are unlikely to have arisen from the forward model used to simulate the reference distribution, then we reject that forward model as an explanation for the observations.

To the extent that there is asymmetry in the onset of diversification of the group M subtypes, questions must be focused upon the timing of the onset of the B epidemic (figure 1). The work of Korber *et al.* (2000*a*) presents two different estimates for the origin of the B clade in terms of *env* sequences, 1967 and 1954, the former estimate presupposing a strict clock, the latter allowing for variable rates among the subtypes. Both estimates suggest a curious retardation of the B epidemic relative to the other subtype epidemics that is inconsistent with the macroscopic epidemiological data pertaining to the North American and African AIDS epidemics (Hooper 1999). There is no evidence of a B epidemic in Africa: isolated B samples have been found in Gabon and Uganda, among other countries, but it is not clear that these are indigenous or old (pre-1990) to those areas. Nor is there evidence of AIDS epidemics in Haiti, the USA or Europe prior to the 1970s. Furthermore, the work of Grassley *et al.* (1999) suggests radically different population dynamics for the African subtype A and North American and European subtype B epidemics. For these reasons, in the analyses that follow we conduct modelling studies both with and without B clade sequences.

The main questions we can now ask are: can a natural process that is evolving according to the model just described produce multiple HIV-1 subtypes—approximately 10; and do the subtypes quantitatively manifest synchrony as implied in figure 1?

(b) *Discriminatory statistics: the number of subtypes*

To formalize the inquiry into the number of group M subtypes, as objectively as possible, a method of choosing the number of subtypes from simulated patterns must be selected. There are several methods for deciding how many subtypes are present and which taxa belong to which subtypes. In a generic sense, this is an unsupervised learning problem, with cluster assignments often depending strongly upon the distance measure used. One common way is to resample the sequences (bootstrap) many times and count the fraction of times that the specified subsets remain clustered using any of several tree-building methods. However, in any method, there is at least an implicit assumption about the evolutionary model; moreover, we have shown that the assumed distance measure (or evolutionary model) can impact the clade assignments regardless of how those assignments are made (Burr *et al.* 2000). Herein we report results from a novel and convenient way to choose subtype assignments: mclust, model-based clustering (Banfield & Raftery 1992) as implemented in available software (S-Plus 1999) that provides a semi-objective way to choose the number of clusters. Qualitatively, mclust is similar to the well-known 'look for diminishing returns' approach that is often used in *k*-means clustering: add clusters until the reduction in within-group sum of squares begins to diminish sharply (see Appendix A).

Figure 4 summarizes results from simulations in which four different epidemiological processes are modelled under the presumption of a naturally unfolding origin of HIV. Figure 4*a* reconstructs an exponential growth process, which Slatkin & Hudson (1991) predicted a decade ago would be a single (monotonic) star phylogeny; our simulation confirms this outcome and illustrates that explanations of the complex HIV-1 star-like phylogeny (figure 1) in terms of exponential growth are erroneous (Hahn *et al.* 2000): exponential growth coupled to a natural process would not generate multiple star-like clades. Stationary epidemics, as illustrated by the one interesting case shown in figure 4*b*, could approach the numerous subtypes observed in figure 1 (we have observed as many as seven subtypes for simulated data from the constant N case), but as Slatkin & Hudson (1991) demonstrated, the outcome is highly random; moreover, stationariness is not the condition of the AIDS epidemics which have prevailed from the 1960s and 1970s into the present. Combining these two processes in tandem—first pre-epidemic stationariness, then exponential growth—as in figure 4*c*, a crudely HIV-like tree is generated, but it has not been possible with these conditions of simulation to produce more than two to five subtypes through a natural Darwinian process. This is, however, what is observed with FIV evolution (figure 2). Finally, a quadratic process, which may best describe the actual early epidemics (figure 3), pushes the star phylogeny closer to what is seen in figure 4*a*, as would be expected. Figure 3 displays estimates for the global HIV prevalence, which we obtained by forming the cumulative HIV incidence data (K. A. Stanecki, personal communication 2000) and subtracting the estimated cumulative death due to HIV where the estimated mean time from infection to death was eight years.

To summarize these findings regarding the relatively large number of distinct group M subtypes: no set of likely natural conditions (figure 4*a*–*d*) will adequately simulate so many as ten distinguishable subtypes in a complex star-like configuration (figure 1). Only the constant population case (in figure 4*b*, which is not representative of the AIDS epidemic) has generated as many as seven subtypes. The quadratic growth case (figure 4*d*) is most representative of the AIDS epidemic and it typically generates one to four subtypes although on rare occasions we have found five subtypes with quadratic growth.

In our simulations, as described in § 3a–c, rarely are more than five subtypes generated by any set of parameters or time to MRCA. If figure 4*c* best approximates the conditions of the supposed pre-epidemic (1930–1970) and early epidemic (1970–1990) according to the natural theory (Korber *et al.* 2000*a*; Hahn *et al.* 2000), the onus is upon the supporters of the natural theory to account for the unexpectedly large number of HIV-1 subtypes. Exponential growth of the epidemic(s) is not by itself a satisfactory explanation (Hahn *et al.* 2000).

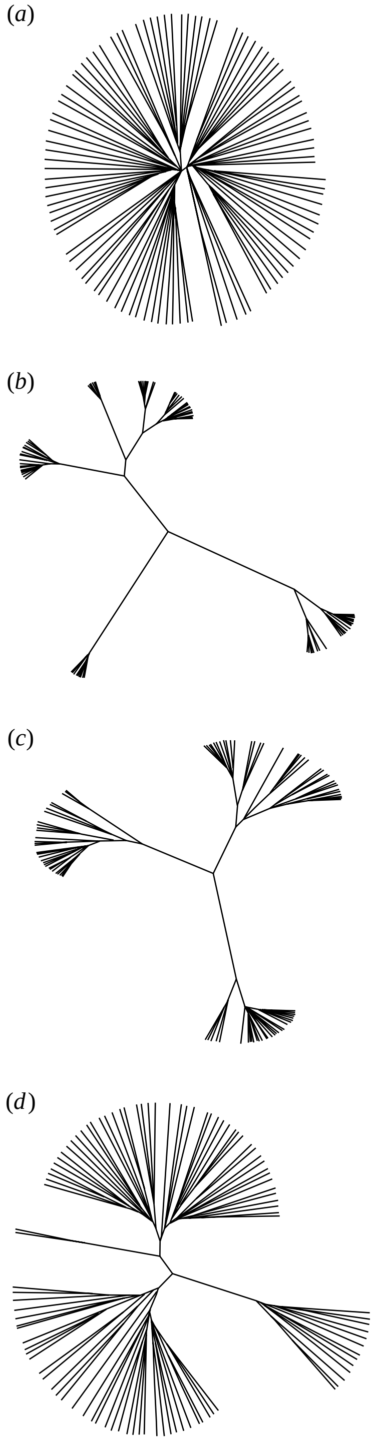

Figure 4. Phylogenetic tree of 100 simulated sequences assuming (*a*) exponential growth, $N = N_0 e^{rt}$, (*b*) $N = N_0$, (*c*) constant N ($N = N_0$) followed by exponential growth ($N = N_0 e^{rt}$), and (*d*) N is quadratic, which is approximated using a piecewise exponential.

(c) *Discriminatory statistics: the extent of synchrony*

We now turn to the question of measuring the extent of synchrony for the diversification of the group M subtypes.

We will define significant synchrony as a small standard deviation s in the n between-subtype distances, where

$$s^2 = (1/(n-1))\sum_{i=1}^{n}(d_i - \bar{d})^2$$

is the sample variance of n between-subtype distances. Note that if the number of subtypes or clades $c = 10$, then the number of distances, n, is 45. Simulations were run at the nominal values of each of the three macroscopic and two microscopic parameters. And, as mentioned, simulations were also run for 32 cases in which each of the five parameters varies in their L to H values in a full-factorial experimental design. Each case was run twice with slightly differing times from 1990 to the MRCA to offer a range of summary statistics for each parameter set. Most often, parameter space was searched by changing parameter values while holding the coalescence time to the MRCA at *ca.* 60 years from 1990, testing the Korber *et al.* (2000*a*) case and at *ca.* 30 years, testing a punctuated origin around 1960, such as could have occurred with the OPV hypothesis. We invariably observed, as should be expected, higher synchrony in cases for which the MRCA was 1960. However, cases having this higher synchrony tended to have much larger intracladal distances (with phylogenies closer to the star phylogeny in figure 4*a*), and the subtype MRCAs occurred much closer to the overall MRCA. Such cases give estimates of one or two clades typically, and for cases with only one or two clades, we cannot define a standard deviation of the intercladal distances. For a simulated case to be considered a viable comparison case, it must produce at least three distinct subtypes as estimated by our application of mclust applied to the principal coordinate data. Accordingly, it becomes highly unlikely that the HIV-1 group M had a natural origin starting close to 1960. The tabulated results of these extensive simulations are too lengthy to be presented here; therefore we present the results for the nominal parameter values (figure 5) and provide the tables with the other parameter values at www.nis7.lanl.gov. We summarize all results below.

The standard deviation of the between-subtype distances ($d_{\text{between,simulated}}$) was evaluated by assuming that there are five clades. In all cases, we found that either or both of the following two summary statistics were very different in the simulated data when compared with the real data for *env* and *gag* (even when we assumed that the simulated data had the same number of clades as the real data). (i) The simulated intercladal distances had higher than the observed standard deviation (expressed below as k_{between}), where $k_{\text{between}} = D_1/s(D_1)$, where

$$D_1 = s(d_{\text{between,real}}) - \text{mean}(s(d_{\text{between,simulated}}))$$

is the difference between the observed $s(d_{\text{between,real}})$ and the average of $s(d_{\text{between,simulated}})$. A large ('statistically significant') value for k_{between} is *ca.* ±2 or more. (ii) The simulated intracladal distances were larger (on average) than the real intracladal distances. This is expressed below as $k_{\text{within}} = D_2/s(D_2)$, where

$$D_2 = (d_{\text{within,real}}) - \text{mean}(d_{\text{within,simulated}})$$

and $s(D_2)$ is the standard deviation of D_2. A large ('significant') value for k_{within} is *ca.* ±2 or more.

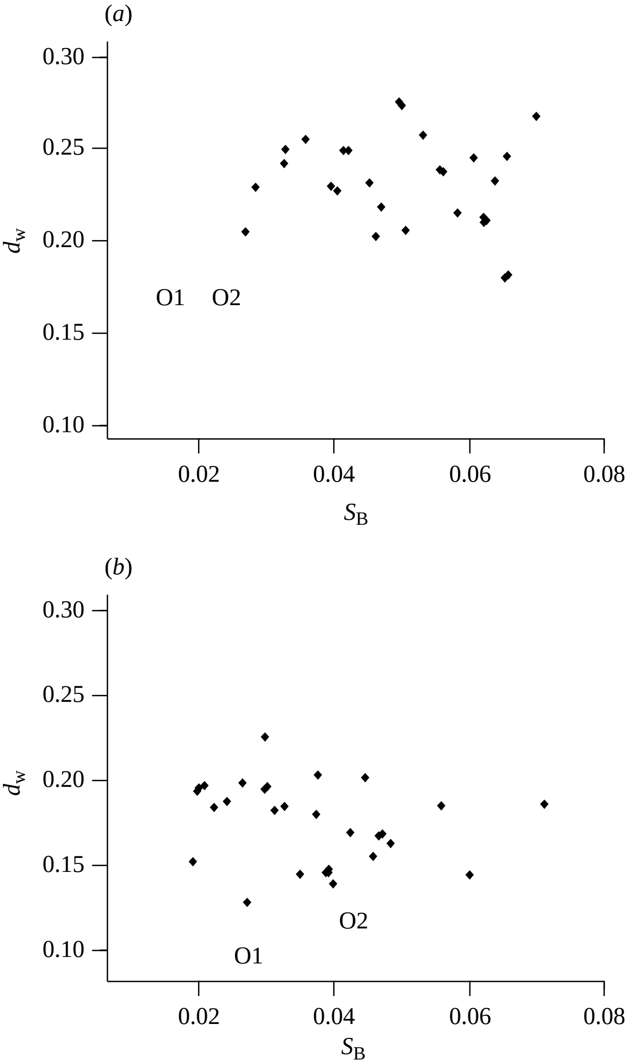

Figure 5. Summary statistics for real data compared with summary statistics from data simulated from the forward model with nominal parameter values. (*a*) *env*, gp120. Note the high surprise level for the observed s_{between} (s_B) and d_{within} (d_W) compared with the simulated reference distribution. (*b*) *gag*, p170. Note that there is a high surprise level for d_W but that s_B is well within the range of the simulated values.

The results for *env* show that at the nominal values for all parameters, the estimated number of clades is three to five in most simulation runs, but (i) k_{between} is *ca.* −2 to −3 (rare event), and (ii) k_{within} is *ca.* −2 to −3 (rare event). At parameter values such as those in cases 3, 4, 9, and 19 in table 1 (www.nis7.lanl.gov), the estimated number of clades is fewer (two or three usually), but the synchrony is sometimes higher than the nominal case. However, even in the most synchronized case, (i) k_{between} is −2 (rare event), and (ii) k_{within} is approximately −3 (rare event).

The results for *gag* show that at the nominal values for all parameters the estimated number of clades is three in most simulation runs, but (i) k_{between} is *ca.* −1 (not a rare event); and (ii) k_{within} is *ca.* 2 to 3 (rare event). At values such as cases 3, 9, and 19 in table 2 (www.nis7.lanl.gov), the estimated number of clades is small (two or three usually), and the synchrony is approximately the same as the nominal case. Therefore, even in the most synchronized case, (i) k_{between} is 0 (not rare event), and (ii) k_{within} is *ca.* -2 to -3 (rare event).

When we combine k_{between} and k_{within} results using the χ_2^2-distribution (Johnson & Wichern 1988), the nominal *env* case gives a *p*-value of *ca.* 0.001 with B and *ca.* 0.0002 without B. When we combine k_{between} and k_{within} results using the χ_2^2-distribution, the nominal *gag* case gives a *p*-value of *ca.* 0.01 with B and *ca.* 0.04 without B.

In figure 5 we present a graphical summary of the results for the nominal case by indicating where the observed summary statistics fall with respect to the simulated reference distribution (using runs that produced at least three clades from the nominal case only) for the summary statistics. Note that qualitatively, the *env* results have a high 'surprise level', while the *gag* results have a high surprise level only for d_{within}. In fact, the d_{between} results for *gag* would not give a 'significant' result if we used a non-parametric test such as giving the rank of the observed d_{between} within the simulated distribution of 30 runs. The rank of the observed d_{between} is 22 out of 31 with the B subtype included, and 8 out of 31 with B omitted. The rank of the observed d_{within} is 1 out of 31 with the B subtype included or omitted. By contrast, the rank of the observed d_{between} or d_{within} is 1 out of 31 for *env* with B or without B.

These results change somewhat depending on whether we include or omit the B sequences and sequences from the subtypes with only a few representative samples. For the *env* sequences, we continue to get (in the nominal case) significant k_{between} and k_{within} values (2–3 in magnitude) no matter whether B is omitted or included or whether the two subtypes with few samples (H and J) are omitted or included. We also get large (two or more) k_{between} and k_{within} values for the cases evaluated in table 1 (www.nis7.lanl.gov). For the *gag* sequences, k_{within} remains significant in all cases (except for case 9 with a coalescent time of *ca.* 200 years) no matter which subtypes are omitted. Overall, the 'surprise level' remains high for *gag* and very high for *env*. Also, for both *gag* and *env*, the 'synchrony level' tends to increase toward the observed only when the intracladal distances also increase beyond the observed as we move toward the figure 4*a* star phylogeny (one or two clades at most would be predicted in such trees).

4. CONCLUDING REMARKS

With the forward model and parameter variation employed herein, it is not possible to simulate a natural Darwinian process that would result in the HIV-1 group M sunburst phylogeny. The model could not reproduce the observed synchrony in the onset of diversification nor the relatively large number of subtypes (signifying at least as many separate epidemics). Also, if we assume a single naturally unfolding epidemic (figure 4), none of the wide range of epidemiological scenarios tested was able to generate more than five distinct subtypes. One could postulate multiple zoonotic transmissions prior to 1970 but these would be nearly impossible to reconcile with the strong synchrony (figure 5) that is responsible for the sunburst phylogeny. The likeliest source of the multiple subtypes and the synchronization of their conspicuous diversification (figure 1) is a punctuated origin. The OPV hypothesis is consistent with a punctuated origin (Hooper 1999) as is urbanization or dirty needles in the African AIDS epidemics (Hooper 1999; Myers 1994). Using a forward model to estimate the timing of the punctuated origin of the human epidemic may be impeded by an associated evolutionary discontinuity in that the initial rates of variation are unlike those characteristics of the viruses that became established in human populations.

A conservative interpretation of our results is that a large mixing population (all of Africa, for example, if we omit the B subtype) cannot produce as many as ten distinct subtypes. In the rare instances when this model can produce up to five subtypes, they are not as highly synchronized as the observed *env* region subtypes. The discrepancy can be accounted for by a synchronizing event (a punctuated origin relative to our model's assumptions) such as urbanization, dirty needles, the OPV (Hooper 1999; Myers 1994), or the sparse sampling from a local outbreak. Alternatively, the associated evolutionary discontinuity may embody initial rates of variation unlike those characteristic of the viruses that became established in human populations. If this happens, then it is not far-fetched to imagine the ten or so clades deriving from a single animal (perhaps immunosuppressed and possessing a swarm of variants) or from a few animals that might have belonged to a single troop or might have been gang-caged together. The number of animals required is secondary to the extent of variation in the source at the time of the zoonotic or iatrogenic event. The OPV hypothesis makes a case for such a punctuated origin (Hooper 1999), although arguments have been put forward for the unlikelihood of the OPV trials as the source of the AIDS epidemics (Korber *et al.* 2000*a*,*b*).

Another possible synchronizing event could be a local outbreak in one African region that later spread distinct, but still closely related, variants of the virus to other parts of Africa and finally to the rest of the world. These distinct variants could then act as seeds (founders) for what later became classified as the different subtypes. The recent discovery of previously uncharacterized M group sequences in Central Africa lends considerable support to the natural origin hypothesis (Vidal *et al.* 2000) with a synchronizing local outbreak. However, a final interpretation of these novel sequences awaits a determination of whether they have arisen from contemporary zoonotic infections, in which case they would throw considerable light upon the origin debate.

We thank Gary Xie (Los Alamos National Laboratory; xie@-lanl.gov) for assistance with phylogenetic tree construction, Oliver Pybus (University of Oxford, Department of Zoology; oliver.pybus@zoo.ox.ac.uk) for helpful conversations regarding the interpretation of the Vidal *et al.* sequences from Central Africa, and Jim Mullins (University of Washington, Department of Microbiology and Medicine; jmullins@u.washington.edu) and the *Journal of Virology* for permission to reproduce figure 2. We are also grateful to Lynda Myers for her editorial help.

APPENDIX A: MATERIAL AND METHODS

(a) *Sequences*

Our HIV-1 group M sequences were taken from the LANL HIV Sequence Database (www.lanl.hiv.gov). These sequences all have isolation times within 0–7 years of 1990. For *gag*, subtypes A, B, C, D, F, G, H and J were included with sample sizes of 20,

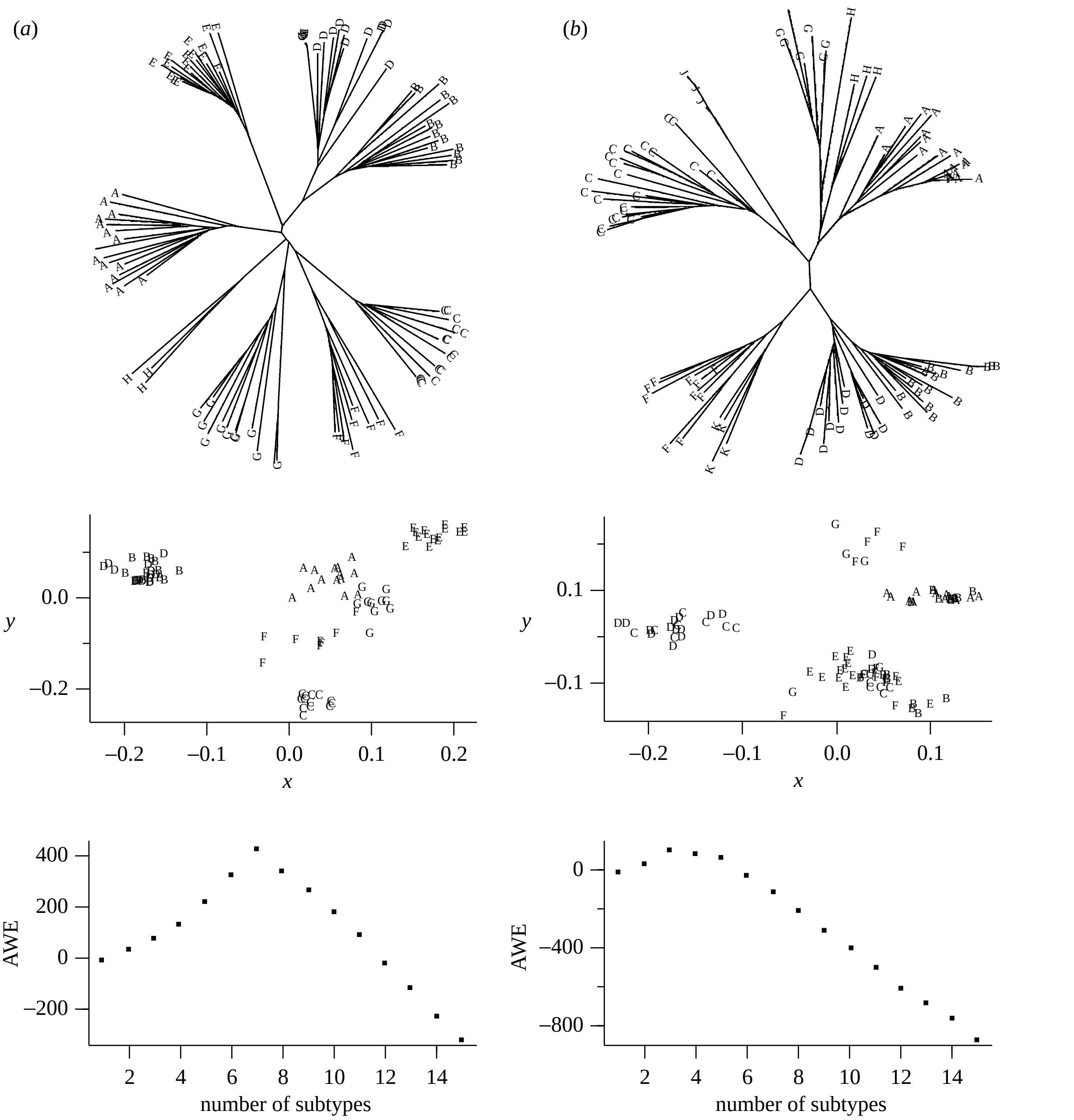

Figure A 1. (*a*) UPGMA tree of distances using the GTR + γ model (top), principal coordinate plot of distances (middle), and AWE plot for candidate numbers of subtypes (bottom) for 96 *env* (gp120) sequences. (*b*) As (*a*), but for 88 *gag* (p17) sequences.

16, 22, 13, 10, 7, 4, 4 and 4, respectively. For *env*, subtypes A, B, C, D, E, F, G, H and J were included with sample sizes of 15, 15, 15, 15, 16, 9, 10, 3 and 2, respectively. Accession numbers are available upon request. We chose the 'all subtypes option' and the associated available alignment option. We do not believe the alignment is optimal, but our results are similar to other published results and our tree methods placed all chosen subtypes into the proper groups so we believe this to be an acceptable approach for our purposes. This resulted in 528 (2531) sites for *gag* (*env*), and 368 (1328) when insertion/deletion alignment characters were removed. We then computed pairwise distances using PAUP* (Swofford 1999) and associated summary statistics using our own software in the statistical programming language S-Plus (S-Plus 1999). Pairwise distances depend on the assumed substitution model. We chose the GTR + γ model, following the recommendation in Leitner *et al.* (1997), and estimated γ to be *ca.* 0.4 for both *gag* (p17) and *env* (gp120) (0.45 for *gag* and 0.40 for *env*), which is in reasonable agreement with Leitner *et al.* (1997).

(b) *Substitution model*

We use the GTR + γ (rate heterogeneity) model. The GTR model has five relative rate parameters for A→C, A→G, A→G, C→G, C→T and G→T changes, plus three relative frequency parameters π_A, π_C and π_T for the A, C and T frequencies. There is also a positive overall substitution rate μ for a total of nine parameters. The parameter μ determines the average rate of change (usually assumed constant over time and for each taxa), and the time-intervals between changes are assumed to be exponentially distributed. Equivalently, the number of changes in a given time is assumed to be Poisson distributed with mean μ. The values of π can be estimated using the observed nucleotide frequencies. The best way to estimate μ requires access to an outgroup taxa, or to have known isolation times as we do for a subset of our real *env* and *gag* data. More recently, it has been demonstrated that allowing μ to vary across sites (rate heterogeneity across sites) is sometimes important, especially so in our context of estimating the number of clades where long branches tend to attract (group together). Typically, μ is assumed to have a gamma distribution with one parameter γ describing its variation across sites (large γ means less variation).

A possible shortcoming of all models currently in use is their inability to model dependence among sites. Some studies partially address this by considering silent and replacement sites separately, or using a separate model for each codon position. If the real *env* or *gag* sequences are evolving in a way that the site dependence makes a large impact on the true evolutionary distance, then our simulation study will have a systematic error that we have not addressed. As mentioned in the text, systematic errors due to model misspecification of the same type in both the real and simulated data would cancel for our purposes here. However, it is possible that only the real sequence data exhibit departure from the model due to site dependence (which would lead to a bias whose magnitude is currently unknown).

(c) *Distance measure and subtype definition*

All distance measures attempt to compute a distance that is expected (on average) to increase approximately linearly or in a known way with time to the MRCA. We use a distance computed under the GTR $+\gamma$ model (with γ estimated using baseml in PAML (Yang 1997) applied to our real *gag* and *env* sequences) and calculate distances in PAUP and use it to compute distances among all pair of taxa.

(d) *Mclust to choose the number of subtypes*

The top plot in figure A1 is the *env* (*a*) and *gag* (*b*) sequences using unweighted paired group method with arithmetic averages (UPGMA). The middle plot represents the matrix of pairwise distances using principal coordinates, which provides a two-dimensional representation of the data that closely preserves pairwise distances. The bottom plot is the suggested number of clades according to mclust applied to the principal coordinates in the middle plot. We have compared mclust results with the more common 'diminishing returns with *k*-means approach' and have noticed a tendency for mclust to suggest fewer groups than *k*-means suggests. We are currently investigating other methods to choose the number of clades. However, our focus is on comparing the number of subtypes in the *gag* and *env* sequences with the number of subtypes in simulated data, so we report results for only one 'subtype assignments' method (mclust). This is adequate for our purposes because we are choosing a defensible method among many choices and we are using the same method to compare real and simulated data. In figure A1 we present the results of our three steps. Qualitatively, we note that figure A1 (*a*) (*env*) has relatively strong evidence for seven clades, while figure A1 (*b*) (*gag*) has weaker evidence for three to five subtypes. One reason for this difference is that the intersubtype distances are larger for *env* (*ca.* 0.4) than for *gag* (*ca.* 0.3). A quantitative measure of the evidence E for the chosen number of clades is $E=0.35$ for $c=7$ clades in figure A1 (*a*) and $E=0.34$ for $c=3$ clades in figure A1 (*b*). We report the number c that maximizes the estimated approximate weight of evidence (AWE) as defined by mclust. And we define E by normalizing the AWE to NAWE (NAWE is non-negative and sums to one), with $E=\text{NAWE}[c]/(\text{NAWE}[c-1]+\text{NAWE}[c]+\text{NAWE}[c+1])$, which is a measure of how peaked the AWE curve is around the value c where the AWE attains its maximum. Estimation of the number of subtypes and the evidence for that estimate is an important side issue that is currently under investigation. Here we report c (using mclust and AWE) and E as just described.

REFERENCES

Bachmann, M. H., Mathiason-Dubard, C., Learn, G. H., Rodrigo, A. G., Sodora, D. L., Mazzetti, P., Hoover, E. A. & Mullins, J. I. 1997 Genetic diversity of feline immunodeficiency virus: dual infection, recombination, and distinct evolutionary rates among envelope sequence clades. *J. Virol.* **71**, 4241–4253.

Banfield, J. & Raftery, A. 1993 Model-based Gaussian and non-Gaussian clustering. *Biometrics* **49**, 803–821.

Burr, T., Myers, G., Hyman, J. & Skourikhine, A. 2000 Impacts of misspecifying the evolutionary model in phylogenetic tree estimation. In *Proceedings of the International Conference on Mathematics and Engineering Techniques in Medicine and Biological Sciences*, 26–29 June 2000, vol. II (ed. F. Valafar), pp. 481–488. Las Vegas, NV: CSREA Press.

Eigen, M. & Nieselt-Struve, K. 1990 How old is the immunodeficiency virus? *AIDS* **4**(Suppl. 1), S85–S93.

Fu, Y-X. & Li, W-H. 1999 Coalescing into the 21st century: an overview and prospects of coalescent theory. *Theor. Popul. Biol.* **56**, 1–10.

Grassley, N. C., Harvey, P. H. & Holmes, E. C. 1999 Population dynamics of HIV-1 inferred from gene sequences. *Genetics* **151**, 427–438 (www.evolve.ac.uk/software/treevolve/main.html).

Hahn, B. H., Shaw, G., De Cock, K. & Sharp, P. 2000 AIDS as a zoonosis: scientific and public health implications. *Science* **287**, 607–614.

Holmes, E. C., Pybus, O. G. & Harvey, P. H. 1999 The molecular population dynamics of HIV-1. In *The evolution of HIV* (ed. K. Crandell), pp. 177–207. Baltimore, MD: Johns Hopkins University Press.

Hooper E. 1999 *The river: a journey to the source of HIV and AIDS*. New York: Little, Brown and York.

Johnson R. & Wichern, D. 1988 *Applied multivariate statistical analysis*, 2nd edn. Englewood Cliffs, NJ: Prentice-Hall.

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

Korber, B., Muldoon, M., Theiler, J., Gao, R., Gupta, R., Lopedes, A., Hahn, B., Wolinsky, W. & Bhattacharya, T. 2000*a* Timing the ancestor of the HIV-1 pandemic strains. *Science* **288**, 1789–1796.

Korber, B., Bhattacharya, T., Theiler, J., Gupta, R., Lapedes, A., Hahn, B., Gao, F., Muldoon, M. & Wolinsky, S. 2000*b* Letter of response to Search for the Origin of HIV and AIDS. *Science* **289**, 1140–1141.

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

Luria S. E. & Delbruck, M. 1943 Mutations of bacteria from virus sensitivity to virus resistance. *Genetics* **28**, 491–511.

Myers, G. 1994 HIV: between past and future. *AIDS Res. Hum. Retroviruses* **10**, 1317–1324.

Myers, G., MacInnes, K. & Myers L. 1993 Phylogenetic moments in the AIDS epidemic. In *Emerging viruses* (ed. S. A. Morse), pp. 120–137. New York: Oxford University Press.

Nahmias, A. J. (and 11 others) 1986 Evidence for human infection with an HTLV-III/LAV-like virus in Central Africa. *The Lancet* **i**, 1279–1280.

Sharp, P. M. & Li, W.-H. 1988 Understanding the origin of the AIDS viruses. *Nature* **336**, 315.

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

Smith, T. F., Srinivasan, A., Schochetman, G., Marcus, M. & Myers, G. 1988 The phylogenetic history of immunodeficiency viruses. *Nature* **333**, 573–575.

S-Plus 1999 *S-Plus 5.1.* Seattle, WA: MathSoft.

Swofford, D. L. 1999 *PAUP**: *phylogenetic analysis using parsimony*, v. 4. Sunderland, MA: Sinauer Associates.

Swofford, D. L., Olsen, G. J., Waddell, P. J. & Hillis, D. M. 1996 Phylogenetic inference. In *Molecular systematics*, 2nd edn (ed. D. Hillis, C. Moritz & B. Mable), pp. 407–514. Sunderland, MA: Sinauer Associates.

Zhu, T., Korber, B. T., Nahmias, A. J., Hooper, E. & Sharp, P. 1998 An African HIV-1 sequence from 1959 and implications for the origin of the epidemic. *Nature* **391**, 594–597.