The University of Manchester

Using human immunodeficiency virus type 1 sequences to infer historical features of the acquired immune deficiency syndrome epidemic and human immunodeficiency virus evolution

Yusim, Karina and Peters, Martine and Pybus, Oliver and Bhattacharya, Tanmoy and Delaporte, Eric and Mulanaga, Claire and Muldoon, Mark and Theiler, James and Korber, Bette

2001

MIMS EPrint: **2006.7**

Manchester Institute for Mathematical Sciences
School of Mathematics

The University of Manchester

Reports available from: http://eprints.maths.manchester.ac.uk/
And by contacting: The MIMS Secretary
School of Mathematics
The University of Manchester
Manchester, M13 9PL, UK

ISSN 1749-9097

THE ROYAL SOCIETY

doi 10.1098/rstb.2001.0859

Using human immunodeficiency virus type 1 sequences to infer historical features of the acquired immune deficiency syndrome epidemic and human immunodeficiency virus evolution

Karina Yusim[1,2], Martine Peeters[3], Oliver G. Pybus[4], Tanmoy Bhattacharya[1], Eric Delaporte[5], Claire Mulanga[3], Mark Muldoon[6], James Theiler[1] and Bette Korber[1,2]

[1]*Los Alamos National Laboratory, Los Alamos, PO Box 1663, NM 87545, USA*
[2]*Santa Fe Institute, 1399 Hyde Park Road, Santa Fe, NM 87501, USA*
[3]*Laboratoire Retrovirus, Institut de Recherche pour le Développement, 911 avenue Agropolis, BP 5045, 34032, Montpellier, France*
[4]*Department of Zoology, University of Oxford, South Parks Road, Oxford OX1 3PS, UK*
[5]*Centre Hospitalier Universitaire (CHU), Gui de Chauliac, 2 avenue Bertin Sans, 34295 Montpellier, France*
[6]*University of Manchester Institute of Science and Technology, PO Box 88, Manchester M60 1QD, UK*

In earlier work, human immunodeficiency virus type 1 (HIV-1) sequences were analysed to estimate the timing of the ancestral sequence of the main group of HIV-1, the virus that is responsible for the acquired immune deficiency syndrome pandemic, yielding a best estimate of 1931 (95% confidence interval of 1915–1941). That work will be briefly reviewed, outlining how phylogenetic tools were extended to incorporate improved evolutionary models, how the molecular clock model was adapted to incorporate variable periods of latency, and how the approach was validated by correctly estimating the timing of two historically documented dates. The advantages, limitations, and assumptions of the approach will be summarized, with particular consideration of the implications of branch length uncertainty and recombination. We have recently undertaken new phylogenetic analysis of an extremely diverse set of human immunodeficiency virus envelope sequences from the Democratic Republic of the Congo (the DRC, formerly Zaire). This analysis both corroborates and extends the conclusions of our original study. Coalescent methods were used to infer the demographic history of the HIV-1 epidemic in the DRC, and the results suggest an increase in the exponential growth rate of the infected population through time.

Keywords: HIV-1; coalescent theory; molecular clock; evolution

1. INTRODUCTION

Acquired immune deficiency syndrome (AIDS) was first detected and defined in 1981 (Gottlieb *et al*. 1981); immediately a search was begun to discover the aetiological agent, culminating in the rapid discovery of human immunodeficiency virus (HIV) (Barre-Sinoussi *et al*. 1983; Gallo *et al*. 1983). Once AIDS had been defined, HIV discovered, and detection methods developed in the early 1980s, it became rapidly apparent that there was an epidemic raging on a very serious scale (Selik *et al*. 1984; Pape *et al*. 1983; Nzilambi *et al*. 1988; Gazzolo *et al*. 1984). Retrospective studies of stored samples allowed some earlier traces of the virus to be identified (Selik *et al*. 1984; Pape *et al*. 1983; Gazzolo *et al*. 1984; Hooper 1999), the oldest being an African sample taken in 1959 from an individual in Léopoldville, Belgian Congo (now Kinshasa, Democratic Republic of the Congo (DRC)), first identified as HIV positive by serological tests (Nahmias *et al*. 1986), and later confirmed by PCR amplification and sequencing (Zhu *et al*. 1998). This virus was a representative of HIV at a relatively early moment in its history, and proved to be very informative (Zhu *et al*. 1998). When subjected to phylogenetic analysis, the viral sequence obtained from this sample yielded a short terminal branch stemming near the centre of the tree, relative to contemporary sequences (Zhu *et al*. 1998). This provided evidence that the sequence itself was valid, because its distinctive behaviour in a phylogenetic tree indicated that it truly was a representative from an older epoch in the history of the epidemic than were modern samples from the 1980s and 1990s. Furthermore, the analysis suggested that the origin of the epidemic was sometime prior to 1959.

An accurate portrait of the epidemic history of HIV can be valuable for understanding how HIV moves through populations, how rapidly it diversifies, and for providing general insight into the emergence of viral epidemics. Yet our understanding is limited by the dearth of clinical and experimental data prior to the definition of the disease and discovery of the virus. What information

do we have to help reconstruct the early events leading to the epidemic? First, a key fact is that simian immunodeficiency virus (SIV) found in chimpanzees (SIVcpz) is closer to HIV-1, in terms of their phylogenetic relationship, than SIV found in any other primate. Therefore the chimpanzee is considered to be the source of HIV-1 in humans (Peeters *et al*. 1989; Gao *et al*. 1999; Hahn *et al*. 2000). Second, as noted above, the single positive sample from 1959 shows that an HIV-1 infected individual was living in central Africa at that time; the position of this sequence in phylogenetic trees suggests that the viral population had already diversified somewhat by the time this virus was sampled (Zhu *et al*. 1998). Third, there are three very distinctive forms of HIV-1, groups M, N and O (Gurtler 1996; Simon *et al*. 1998; Korber *et al*. 1999); group M is the main group globally, and group O and N infections are rare, and may be the result of separate introductions of the virus from chimpanzees. Finally, as HIV-1 was in humans prior to 1959, and AIDS was not detected and defined until 1981, there was clearly a period of time when HIV-1 was moving through the human population undetected. This is not surprising given HIV's long latency and the many different clinical manifestations of the disease (Grmek 1990). HIV-1 may have been present in low levels in populations with limited access to healthcare, and either gradually shifted from low-risk into high-risk populations, or made more sudden shifts into higher-risk populations through specific events (Chitnis *et al*. 2000).

Because useful biological data from earlier decades are sparse, we must fall back on analytical strategies to attempt to deduce the events that could give rise to the variety of viruses circulating in the contemporary epidemic. Through modelling the evolution of the virus, elements of the history leading to the global epidemic can be reconstructed.[1] When interpreting such results, however, it is important to bear in mind the inherent assumptions and limitations of the models used.

There are several strategies that can be used to reconstruct different aspects of the evolutionary and demographic history of a rapidly evolving pathogen, like HIV, from modern data. We will focus on two distinct, but complementary, methods. One method estimates the rate of evolution of the virus. This basic strategy has been applied to the estimation of the time-frames of lentiviral evolution in primates, as well as the most recent common ancestor of the epidemic viral strains (the M group), and for modelling quasispecies evolution within infected individuals (for recent examples, see Shankarappa *et al*. 1999; Leitner & Albert 1999; Sharp *et al*. 2000; Korber *et al*. 2000; Vandamme *et al*. 2000). The other method is based on coalescent theory, and has been used to make inferences about the demographic history of the A and B clades of HIV-1 (Pybus *et al*. 2000; Grassly *et al*. 1999; Holmes *et al*. 1995, 1999). Both methods depend on the accuracy of phylogenetic trees that serve as their foundation.

2. THE TIMING OF THE MOST RECENT COMMON ANCESTOR OF THE M GROUP OF HIV-1

The rate of evolution and the timing of ancestral events based on evolutionary distances to ancestral nodes were estimated for HIV-1 using phylogenetic reconstructions of sequences with known year of sampling (Korber *et al*. 2000). This work was based on the assumption of a molecular clock, essentially a constant rate of evolution (Hillis *et al*. 1996). There are many issues to consider when undertaking and interpreting such an analysis: how well the data meet the assumptions of the models; the quality of the data; what controls are available for testing the performance of the model; the merits of the analytical strategies employed; and generation of appropriate confidence intervals. How these issues were dealt with within the Korber *et al*. HIV-1 timing study is discussed in the following sections.

(a) *The evolutionary model*

Incorporating a realistic evolutionary model in maximum-likelihood phylogenies is an important aspect of obtaining correct branching orders and branch length estimates that exhibit reasonably clock-like behaviour in HIV-1 trees (Leitner *et al*. 1996, 1997; Leitner & Albert 1999). Such evolutionary models incorporate base frequencies, relative rates of change between bases, and variation in mutation rates at different positions in a sequence alignment (Hillis *et al*. 1994; Swofford *et al*. 1996; Huelsenbeck 1995; Yang *et al*. 1994; Yang 1996). Maximum-likelihood phylogenies are extremely computationally intensive, and intractable for large numbers of sequences when using a single work-station. An implementation of maximum-likelihood tree building code designed for parallel supercomputers was developed to allow both optimization of the evolutionary model and the use of large comprehensive sets of sequences.

(b) *Incorporation of error on time*

A linear fit through a set of points, plotting branch length from the tips of the branches to the ancestral node of interest against the year of sampling, allows one to project back and estimate the time associated with zero branch length, i.e. the time of the ancestral sequence (Hillis *et al*. 1996). A standard linear least-squares fit to a line implicitly assumes that the data are precisely known on the independent axis (the sampling times) and the best-fit line is chosen to minimize the squared deviation on the dependent axis (the branch lengths). Under the assumption of a molecular clock, one would anticipate Poisson error on branch length (Hillis *et al*. 1996). HIV-1 sequences, however, are likely to have error in both dimensions, branch length and time. Despite known year of sampling, the cumulative time a given sequence has had to evolve is not precisely known for two reasons: (i) the sampling time is generally only recorded to a precision of one year, and (ii) an HIV-1 provirus can be harboured, not evolving, for an extended period of time in persistently infected cells (Furtado *et al*. 1995; Perelson *et al*. 1996; Finzi *et al*. 1999; Gunthard *et al*. 1999; Zhang *et al*. 1999), and so, for example, viral DNA sampled in a given year may actually have last replicated some years earlier; examples of this are often observed in within-patient longitudinal studies, where a sequence sampled in a later year will be nearly identical to a cluster of sequences sampled at an earlier time-point (Ganeshan *et al*. 1997; Wolinsky *et al*. 1996; Shankarappa *et al*. 1999; Rodrigo *et al*. 1999). The linear fit in Korber *et al*. (2000)

(figure 1) was thus based on a two-dimensional probability density for each data point, modelling the combined time uncertainties described above with a Poisson distribution for branch length error. Model parameters for slope, intercept and the time-scale (τ) of the harboured sequences are estimated using maximum likelihood. The log-likelihood for the data given a model is the sum of the contributions from each point given by the integral of this probability density along the model line. The parameter τ, or the exponential decay time for the 'age' of sampled sequences, had an average value estimated from the data of 3.4 years (95% confidence interval (CI) of 1.7–7 years). This decay rate suggests that two-thirds of the samples have an evolutionary delay of less than 3.4 years.

(c) *Use of a Monte Carlo method to provide confidence intervals*

To determine a 95% confidence interval for estimated evolutionary rates and for the timing of ancestral nodes, random-with-replacement bootstrap re-samplings (Efron & Tibshirani 1991) were carried out based on the inferred data points defined by branch length versus time of sampling, recalculating the best-fit line for each of the bootstrap data sets.

(d) *Sequence length*

Longer sequence lengths improve the accuracy of phylogenetic trees (figure 2). Over the last few years many full-length HIV-1 envelope sequences have been generated that could be included in the analysis; 144 HIV envelope gp160 sequences with known time of sampling were available for the Korber *et al*. (2000) study, after excluding all of the obvious interclade recombinant sequences from the set. The parallel maximum-likelihood code enabled the use of the full set. This analysis, being based on the alignment of the longest sequences, gave the best estimate of the timing of the most recent common ancestor of the M group: 1931 (95% CI of 1915–1941) (figure 1; Korber *et al*. 2000).

(e) *Control sequence sets*

There were very little historical data available that could be used as reference points to test the accuracy of our methods; prior to initiating the analysis in Korber *et al*. (2000), two controls were selected as they were considered the best controls available. The first was the 1959 sequence. Treating the time of sampling as an unknown, the branch length between the 1959 sequence tip to the M-group root was used to estimate the time of sampling; this strategy gave an estimate of the time of sampling to be 1957 (95% CI of 1934–1962). The accuracy of this estimate suggests that the method, including the position of the root, was reasonable. A second time-point that was reasonably well established was the internal node in the tree associated with the origin of the E-clade epidemic in Thailand. (Clade E is now called circulating recombinant form 01, or CRF01.) The Thai E-clade ancestral node was the only node that met three important criteria for serving as a control: (i) there was strong epidemiological evidence based on large-scale screening to indicate the timing of the introduction (Wangroongsarb *et al*. 1985; Smith 1990); (ii) the first sequences sampled were all highly conserved (McCutchan *et al*. 1992; Subbarao *et al*. 1998), suggesting the virus was detected within a few years of the initial expansion, thus validating the epidemiological evidence and confirming a single introduction rather than multiple imports of the E-clade virus; and (iii), the pragmatic consideration of there being enough data to serve as a basis for a timing estimate. The epidemiology and sequence data combined indicated that the founder virus of the Thai E-clade virus would have been introduced in the mid-1980s. The date of the ancestral sequence of an internal node was estimated in two ways. The first involved extrapolating back from only E-clade sequences, and this yielded 1986 (95% CI of 1978–1989). The second summed the distance from the Asian E-subtype ancestor to the M-group root, and used the evolutionary rate derived from the full set of M-group sequences to estimate the time of origin of the Asian E subtype; this estimate was 1984 (95% CI of 1980–1986).

The fact that these estimated dates are in excellent accord with dates established by other, external evidence was very encouraging. Both of these controls required using shorter sequences than the primary full-length envelope analysis, so were expected to be less reliable than the gp160 analysis; despite this, they performed well. Only short fragments could be obtained from the 1959 sequence, and the Thai E clade is embedded in a circulating recombinant form that also contains regions of A-clade sequence (CRF01; Robertson 2000); A clade regions were excluded to avoid complications due to recombination. When the timing of the M-group ancestor was re-estimated using the boundaries of these shorter sequences, but with the inclusion of either the 1959 sequence or the Thai E-like region of envelope, the results were in good accord with the full gp160 timing estimate (Korber *et al*. 2000), as was an additional independent analysis based on the HIV-1 gag gene. Furthermore, these results agree with studies that used other approaches to address this same question (Vandamme *et al*. 2000; Sharp *et al*. 2000).

(f) *Recombination*

HIV phylogenetic studies may be confounded by undetected recombination events, or by gene regions subject to different pressures imposed by natural selection. Extensive spatial phylogenetic variation (different branching patterns in different regions of the sequence) has been documented for HIV genes by using the computer program PLATO, which identifies regions where sequences evolve anomalously (Holmes *et al*. 1999). While interclade recombinants can generally be identified and excluded from analysis, intraclade recombinants might still be an issue. Thus a preliminary exploration of the implications of intraclade recombination for timing estimates was undertaken (we intend to extend this study in a systematic and more statistically rigorous way in the future). A maximum-likelihood tree based on 21 gp160 envelope sequences was generated, including five representatives randomly selected from four different subtypes, including the consensus of all subtype consensus sequences as the outgroup. Two sequences were selected at random from each of the four subtypes, and a series of artificial recombinants was generated, including

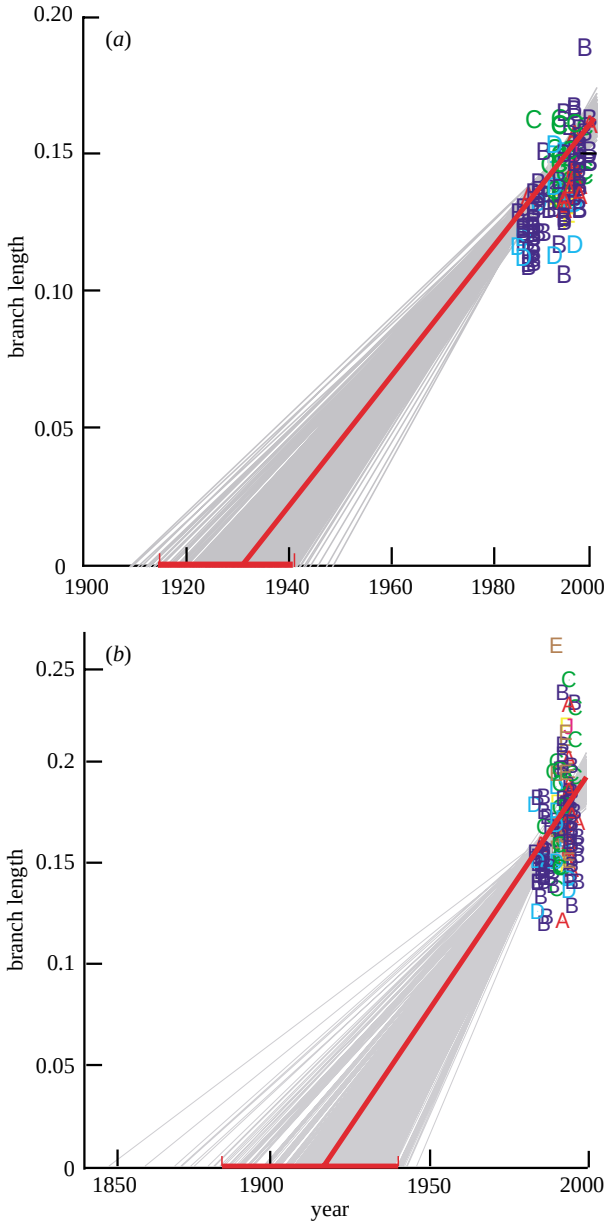

Figure 1. Linear projections to estimate the most recent common ancestor of the M group. Data points are defined by the branch length from branch tips to the ancestral node of the M group for each sequence, plotted as a function of time of sampling. Linear fits include error on both axes using the model discussed in § 2b and are slightly steeper than a traditional least-squares fit, and are offset to the left reflecting the unidirectionality in the error on the time-axis, towards the past. The bold red line indicates the best fit for the real data, the softer grey lines indicate the 480 boostraps. Extrapolating from branch lengths of contemporary sequences, collected over the past two decades, to a point of zero branch length provides an estimate for the most recent common ancestor. (*a*) Estimate for the gp160 tree, included in Korber *et al.* (2000) (144 sequences, 2038 bases), and (*b*) estimate for the V3–V5 envelope tree restricted to the region sequenced for the DRC set, 392 bases. CRF01 sequences were included in the analysis for (*b*), as they are non-recombinant (subtype E) throughout this region of envelope. A feature of note is the greater range in the confidence bounds for (*b*); this is probably a consequence of using only a fragment of envelope compared with the full-length gene, resulting in increased uncertainty for the data points.

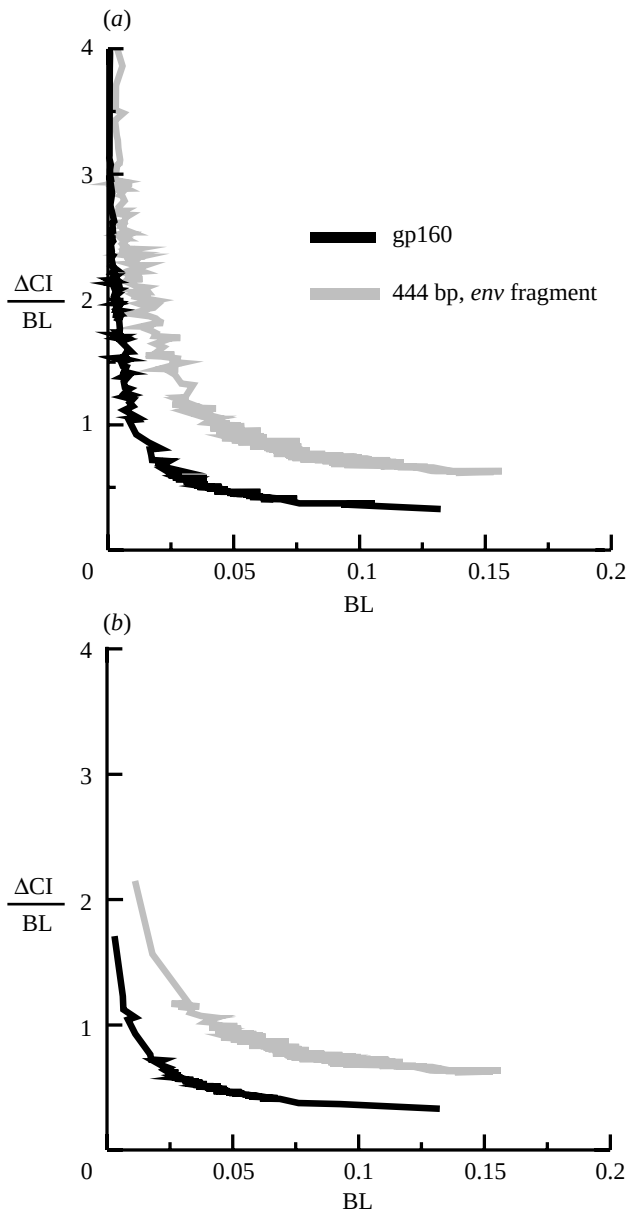

Figure 2. The influence of sequence length on extent of confidence intervals for a given branch length (BL). Two maximum-likelihood trees built by method described in Korber *et al.* (2000) were used as the basis for this analysis. One tree is based on the alignment of 142 gp160 sequences with a length of 2038 bases after gap-stripping, described in Korber *et al.* (2000). The other tree is based on the alignment of 194 DRC sequences with a length of 444 bases (the V3–V5 fragment of envelope). (See more about the alignment in the legend for figure 3). For each tree the relative error was calculated as the length of the confidence interval (CI) divided by the branch length. An example of the estimated branch length is 0.046 (0.025, 0.068) using the random notation BL (CI_1, CI_2), the relative error is $(CI_2 - CI_1)/BL$. The relative error was plotted versus the branch length. The results from the gp160 tree are shown in black, and the results from the V3–V5 fragment in grey. (*a*) Relative errors for all branches in the tree. Branch lengths between all nodes. (*b*) Relative errors for only those branches from the leaves to the nearest internal nodes. Two conclusions can be drawn from these plots. First, shorter sequences yield branch lengths that are known less precisely. Second, the shorter estimated branch length, the larger the relative error.

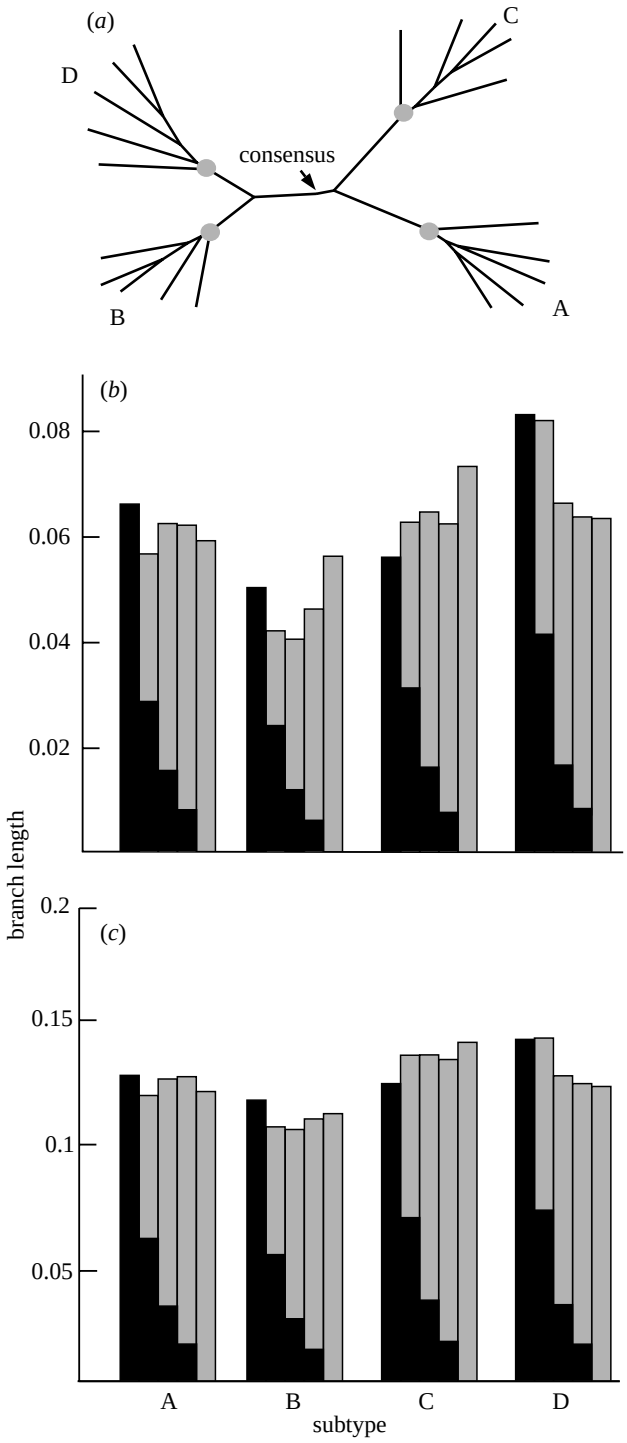

Figure 3. The effects of intraclade recombination on HIV sequence branch lengths. (*a*) A maximum-likelihood tree based on 21 gp160 envelope sequences was generated using a slightly modified version of fastDNAml (Olsen *et al*. 1994).[2] The alignment included five representative sequences of 2373 bases in length after gap-stripping, randomly selected from subtypes A, B, C and D; the consensus of all subtype consensus sequences as the outgroup. Two sequences were selected at random from each subtype, and artificial recombinants were generated based on the two parental strains (one-half, one-quarter, one-eighth). Then the parental sequences were excluded and three additional trees were constructed, with either the one-half, one-quarter, or one-eighth recombinants added. In each case the distances to (*b*) the subtype ancestral nodes, shown by the grey dots on the interior nodes in (*a*), and (*c*), the main group consensus node were calculated for all sequences. In (*b*) and (*c*) the branch lengths of the parental sequences to the node of interest from the first, non-recombinant tree are shown for each subtype. One parental sequence distance is shown in black (at the left for each subtype) and another in grey (on the right of each subtype). The distances of the recombinant sequences to the subtype nodes and main group node taken from the appropriate tree are plotted between the distances of the parental sequences for each subtype. The relative portion of the black and grey shading in the distance bar reflects the contribution of each parental sequence to the recombinant. No recombinant sequences had longer branch lengths to the node of interest than either parental strain, although some had branch lengths shorter than both parental strains.

progressively less of the second sequence (one-half, one-quarter, one-eighth). Three new trees were built that excluded the two parental strains, but included an artificial recombinant. We then compared the branch lengths between the branch tips for the parental strains in the original tree and the artificial recombinants in the subsequent trees, either to the subtype ancestral sequence or to the M-group ancestral sequence. The results are summarized in figure 3. The three non-recombined sequences in each clade had minor random fluctuations in branch length relative to the original tree, and were not significantly altered, and so are not included in the figure. As expected, the artificial recombinants had a branch length either shorter than both parental strains, or intermediate in length between the two parental strains. If such a trend is general, and if recombination was found evenly distributed throughout the time of sample collection, the net result would be that the estimated date of the ancestral node would be artificially close to the present, and timing estimates based on trees with undetected intraclade recombinants would yield dates that are artificially close to the present. If there was a concentration of such recombinants during just one period of sampling, however, then effect on the timing estimate would be unpredictable. Furthermore, recombination can increase the apparent rate heterogeneity among sites in the alignment, which may in turn tend to increase branch length estimates (M. Worobey, personal communication September 2000). Coalescent methods have an additional sensitivity in terms of undetected recombination as compared with timing estimates, in that the branching order as well as the branch length will be modified.

(g) *The M-group root*

Finding the position of the ancestral node in an HIV-1 tree can be problematic. Traditionally, SIVcpz sequences would be used as an outgroup to find the root position. But this strategy gives branch point positions of the outgroup that are unreliable and depend on the precise tree, often much closer to one M-group clade than to all others; usually the likelihood changed only minimally when the SIVcpz outgroup position was shifted to interior nodes of the tree and the branch lengths re-estimated. Therefore, a consensus sequence of the consensus sequences from each subtype was used as the outgroup, which forced a central position of the root. The reasonable timing estimates obtained for the 1959 and Thai E-subtype controls suggest that this was a valid strategy.

In contrast to the M-group root, locating the position of the most recent common ancestor of a given subtype is straightforward, as these nodes are well defined. But for estimates of timing of the origin of a subtype, there are other issues, in particular obtaining enough data to be reliable and representative of the full range of diversity.

(h) *Relaxing the assumption of a molecular clock*

One of the assumptions of the timing methods described above is that the rate of evolution is constant throughout the tree. Thorne *et al*. (1998) developed an interesting alternative strategy that allowed the evolution of the rate of evolution in a maximum-likelihood tree. The branch lengths in such trees reflect time directly, such that each branch has an estimated evolutionary rate and each ancestral node has an estimated time. This strategy was adapted to accommodate HIV-1 sequences that were collected at different time-points. The method has a Bayesian aspect: prior distributions for the position of the root node and the relationship between rates of nearby branches are required as input, and the result is a posterior distribution dependent on both the priors and the data. It was found that to accommodate the relationships of rates between the parental and child branches, the nodes tended to be inappropriately pushed back in time within the constraints of the prior distribution, with the exception of the M-group ancestral node timing estimate, which was robust over a wide range of priors. Because this strategy did not perform as well as the method that assumed a strict molecular clock when compared with our controls, the 1959 sequence and the Thai E clade's most recent common ancestor, we favoured the latter method. The strategy that relaxes the molecular clock does, however, hold promise for future development as evolutionary rates may indeed vary significantly. One interesting aspect of the trees generated using this method, both in Korber *et al*. (2000) and in our more recent study using the DRC sequences (data not shown), is that despite a tendency to compress the internal nodes back in time, a span of several decades is required to cover all of the ancestral nodes of the HIV-1 subtypes, suggesting they did not all arise at one moment in history, but at different times. The timing estimate for the M-group origin using the DRC set with this method was 1932 (95% CI of 1905–1954).

3. ANALYSIS OF DATA FROM THE DRC

(a) *The molecular epidemiology of the DRC set*

In April of 1997, 247 HIV-1 positive blood samples were collected for a molecular epidemiology survey of virus variability from three regions in the former Zaire (Kinshasa, the capital city located in the west, Bwamada in the Equateur Province in the north, and Mbuyi-Maya in the Kasai Province in the south) (Vidal *et al*. 2000). The sequences were obtained from individuals with tuberculosis, pregnant women, sexually transmitted disease patients, blood donors, female sex workers, and asymptomatic adults. V3–V5 sequences were obtained from 197 out of 247 samples. The 50 samples that were not sequenced were classifiable by HMA and subtype-A-specific PCR, which would have introduced a bias towards reduced fraction of subtype A sequences among the set that was sequenced. This set was remarkably diverse; all of the known subtypes were represented (including a single B-subtype sequence). All sequences were examined for inter-subtype recombination breakpoints within the fragment sequenced; the few clear recombinants were excluded from further analysis. Several sequences were unclassifiable in this region, and some of these clustered together in distinctive clades. All sequences that did not have evidence of overt interclade recombination within the fragment sequenced were included in our analysis, including those that were unclassifiable, as they probably represent fragments from distinct and novel lineages (in some cases this was shown by obtaining full-length envelope sequences; Vidal *et al*. 2000). Data from the DRC are very interesting, as they represent a well-established regional epidemic in central Africa. Evidence for this includes the great diversity found in this region; that the oldest known HIV-positive sample comes from the DRC (Zhu *et al*. 1998); and that early samples obtained in 1976 from the DRC were complex recombinants (Choi *et al*. 1997), indicating that the virus was well established with multiple clades co-circulating by this time (Choi *et al*. 1997; Gao *et al*. 1998; Srinivasan *et al*. 1989). The DRC sequence set should be considered to be representative of the geographical region, as there may have been frequent human movement across the boundaries of bordering countries. It is worth emphasizing that it is not known precisely where the epidemic originated, and our work does not address this point.

How do the DRC data differ from the global gp160 data set? First, all samples were collected at the same time, so we cannot plot time versus branch length and fit a line through the points to estimate the rate of evolution; instead there is a spread of branch lengths from one time-point, a purely statistical effect. Second, the data from the DRC are extremely diverse, both in terms of intra-subtype diversity and M-group diversity (Vidal *et al*. 2000; Mokili *et al*. 1999), with more diverse clades represented than had previously been observed. Third, the new alignment has many fewer bases (392 compared with 2038 in the original data set, after gap-stripping), and because of the shorter length of these sequences, both the branching order and distance between leaves on the tree and the node of interest will be known with less certainty in the DRC set than the original gp160 set. As expected, confidence intervals for a given branch length in the tree are broader for the DRC tree than the gp160 tree (figure 2). It was for this reason that the tree based on the gp160 sequence alignment was considered the most reliable test set in the Korber *et al*. (2000) study. On the other hand, a smaller fragment has less opportunity for undetected recombination breakpoints, which may confer an advantage.

(b) *Timing estimates based on the DRC set combined with gp160 sequence*

To enable the estimation of an evolutionary rate for the DRC data, we combined the sequence data from the original gp160 study (Korber *et al*. 2000), which was designed to have the maximum possible spread in time-points of collection (spanning 16 years), with the sequence data from the DRC. We generated a tree and a model for the rate of evolution based on this combined sequence set

using the same maximum-likelihood strategy as we used for gp160. There were 356 taxa in the combined data set, compared with 144 taxa in the gp160 data set; as stated above, the combined alignment includes sequences only 392 bases long. This strategy gave an odd bootstrap result with a bimodal distribution of rates, possibly due to the large number of points from the DRC, all with a sample time of 1997. Thus we excluded the DRC set, but used the same region of 392 bases to recalculate the appropriate evolutionary rate from this region of envelope (figure 1*b*).

The same strategy as was used for the gp160 data set was used to find the root of the M group for the V3–V5 region set: a consensus of the subtype consensus sequences was used as the outgroup. The tree with the maximum likelihood resulted in an estimate of the time of origin for the M group of 1916 (95% CI of 1883–1940). The confidence interval was broader for the alignment of shorter sequences, essentially encompassing the gp160 results of 1931 (95% CI of 1915–1941) (figure 1; Korber *et al*. 2000). Finally, as explained above, analysing the DRC sequence data combined with the previous gp160 sequence data set enabled us to estimate a rate of evolution appropriate for the V3–V5 region sequenced in the DRC study: 0.0023 (0.0016–0.0033) substitutions per base pair per year. This rate is roughly comparable with the rate we calculated for the full gp160 sequence even though the V3–V5 region is one of the most variable and rapidly evolving sub-regions in envelope, and might be expected to have had a higher evolutionary rate.

(c) *The DRC set and coalescent theory*

This set of sequences from the DRC has several properties that make it suitable for analysis using coalescent theory, enabling us to make inferences about the demographic history of the HIV-1 epidemic. First, all sequences were sampled at the same time: April 1997. Simultaneous sampling is not a theoretical requirement of coalescent methods (Rodrigo *et al*. 1999), but it simplifies analysis and is assumed in our methods (see below). In contrast, the time of sampling for the gp160 sequences used in Korber *et al*. (2000) spanned almost two decades, a significant fraction of the full history of the M group. Second, the sequences represent a (nearly) random sample of HIV-1 infected individuals in the DRC; for most studies, the sampling is more selective. The DRC set would have been closer to ideal if all 247 samples had been sequenced. (A study of the full DRC set is underway, but not all sequences were available at the time of this writing.) Third, the sample population appears to be relatively well mixed in terms of the diversity of subtypes found in all three sampling locations, approximating a panmictic population. Finally, for the DRC sequence data, we have an estimated rate of evolution, 0.0023 (0.0016–0.0033) substitutions per base pair per year. Hence our estimated demographic history, i.e. the shape of the growth of the epidemic through time, can be transformed into a natural time-scale of years.

4. COALESCENT APPROACH

(a) *Coalescent methods in HIV research*

Coalescent theory aims to describe how population genetic factors (such as recombination or population size) affect the statistical properties of randomly sampled gene sequences (Kingman 1982*a*,*b*). The shared history of such sequences can be visualized as a genealogy, the lineages of which 'coalesce' as we move backwards in time, until the most recent common ancestor of the sample is reached. The shape of a genealogy is determined by factors such as population size; hence genealogies reconstructed from observed nucleotide sequences can be used to infer the size of real populations.

Coalescent theory has been applied to HIV populations at different scales, from single infected individuals to whole countries or continents. For example, the HIV-1 generation time *in vivo* has been estimated using a coalescent approach with sequences drawn serially over time from a single individual (Rodrigo *et al*. 1999). Here we concentrate on modelling HIV at the epidemic level; sampled gene sequences represent infected individuals and coalescent events occur when two branches in a tree merge into a common ancestral point. In essence, a reconstructed molecular phylogeny is used to estimate a population transmission tree (Holmes *et al*. 1995).

Previous coalescent analyses of HIV demographic history (the number of infected individuals through time) have concentrated on differences at the subtype level, especially between the well-characterized and prevalent subtypes A and B. Subtype A contributes substantially to the sub-Saharan pandemic, while subtype B mainly circulates in the developed world. Initial coalescent analysis using lineages-through-time-plots (which display the rate of coalescence in a genealogy through time) indicated that both subtypes have spread at a roughly constant exponential rate (Holmes *et al*. 1999). Grassly *et al*. (1999) compared subtypes A and B using a pairwise difference distribution method that assumed equal growth rates for both subtypes, and concluded that the current effective population size of subtype A was larger than that of subtype B. In contrast, Pybus *et al*. (1999, 2000) inferred that subtype B has a faster growth rate than subtype A. However, their estimates of current population size for subtype B were larger than for subtype A, which disagrees with the known epidemiology. One possible explanation for this discrepancy was suggested by Pybus *et al*. (2000); the demographic model used to fit the data (pure exponential growth) was insufficient to explain the history of subtype B. Other major factors that affect the interpretation of coalescent results include recombination and natural selection of the sequences, non-random sampling due to political and economic reasons, population migration patterns, founder effects and different mutation rates. The analysis of the DRC data set is expected to be more reliable than the A- and B-subtype sets, as it conforms better to the assumptions of coalescent theory.

(b) *Brief outline of the underlying theory*

Coalescent theory provides a probability distribution for the waiting times between coalescent events, when two lineages merge, in a genealogy (coalescent intervals) (Kingman 1982*a*,*b*; Griffiths & Tavaré 1994). This distribution depends on the demographic history of the sampled population. Demographic history is represented mathematically by the function $N_e(t)$, which represents the 'effective' population size at t years before the present.

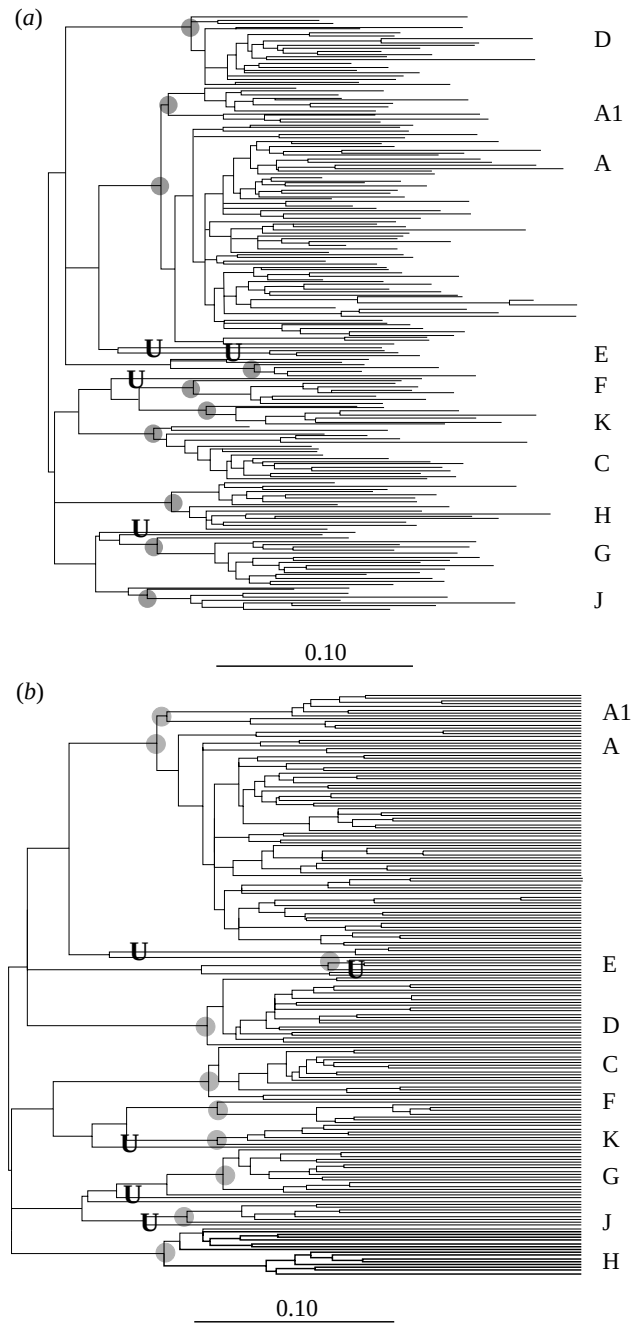

Figure 4. Maximum-likelihood trees based on the DRC sequence set (upper-case letters on the right refer to HIV-1 subtypes). Sequences sampled are provided in Vidal *et al*. (2000). The alignment of these sequences was initially generated using the HMMER method (see http://hmmer.wustl.edu); the sequence alignment was then manually adjusted. While up to 700 bases were obtained from these samples, by the time they were aligned and gap-stripped, only 444 bases were retained in the alignment. The subtype designations originally obtained by Vidal *et al*. were confirmed and clear interclade recombinants were removed from the alignment, so that the final set consisted of 193 sequences. In the trees the grey filled circles indicate subtype ancestral nodes, and 'U' stands for unclassified sequences. Seven sequences classified within the A subtype formed a separate cluster, which we called A1 for convenience of presentation. This should not be confused with a sub-subtype designation, as there is inadequate sequence information available to confirm such a designation; here we consider the A1 sequences to belong to the A subtype. An initial maximum-likelihood tree (A) was generated following the method described in Korber *et al*. (2000), with consensus of all subtype consensus sequences used as the outgroup. The branching order of this initial maximum-likelihood tree was used to generate maximum-likelihood tree (B) with a molecular clock enforced (the method described in Korber *et al*. (2000) was extended by Bhattacharya to include the molecular clock option) (T. Bhattacharya, unpublished data). The log likelihoods of both trees are shown: (*a*) 30830.368 and (*b*) 31130.234.

Effective population size is proportional to actual population size, with a constant of proportionality related to the variance in reproductive success among individuals (see Donnelly & Tavaré (1995) for details). This variance is assumed to be constant through time. Other coalescent model assumptions have been discussed above: no recombination, no subdivision, and randomly sampled sequences obtained at the same time point. We further assume that the reconstructed genealogy used is accurate, and that mutation is clock-like.

Our analysis was conducted using the computer program Genie, which implements a framework for the inference of viral population history from reconstructed phylogenies (Pybus *et al*. 2000). Genie provides parametric and non-parametric estimates of $N_e(t)$, and permits a statistical comparison of model fit. Parametric estimates of $N_e(t)$ are obtained first by specifying a particular demographic model (constant size or exponential growth, for example) and then estimating the parameters of that model using maximum likelihood. Non-parametric estimates of $N_e(t)$, called skyline plots, are obtained by transforming the coalescent intervals of an observed genealogy into a piecewise plot that represents effective population size through time. Skyline plots may show a systematic downward bias if the rate of population change is rapid in comparison with the rate of coalescence (Pybus *et al*. 2000). The parametric and skyline plot estimates of $N_e(t)$ can be superimposed for visual comparison.

(c) *Results of the coalescent approach applied to the DRC data*

Coalescent analysis of the DRC sequences was performed in three steps: (i) the genealogy of the DRC sequences was estimated; (ii) a non-parametric estimate of viral demographic history (the skyline plot) was calculated; and (iii) parametric maximum-likelihood estimates of demographic history were obtained under three models—constant size, exponential growth and expansion growth.

The expansion model represents a population whose growth rate is accelerating through time. It is defined as follows:

$$N_e(t) = N_e(0)(\alpha + (1-\alpha)e^{-rt}),$$
$$\text{where } r \geqslant 0;\ 0 \leqslant \alpha \leqslant 1. \qquad (1)$$

The current size of the population is $N_e(0)$ and it is growing almost exponentially with a rate that will increase to r in the distant future. Moving into the past, the exponential growth rate slows until the population eventually approaches an equilibrium size of $\alpha N_e(0)$. Coalescent theory becomes insensitive to the history of

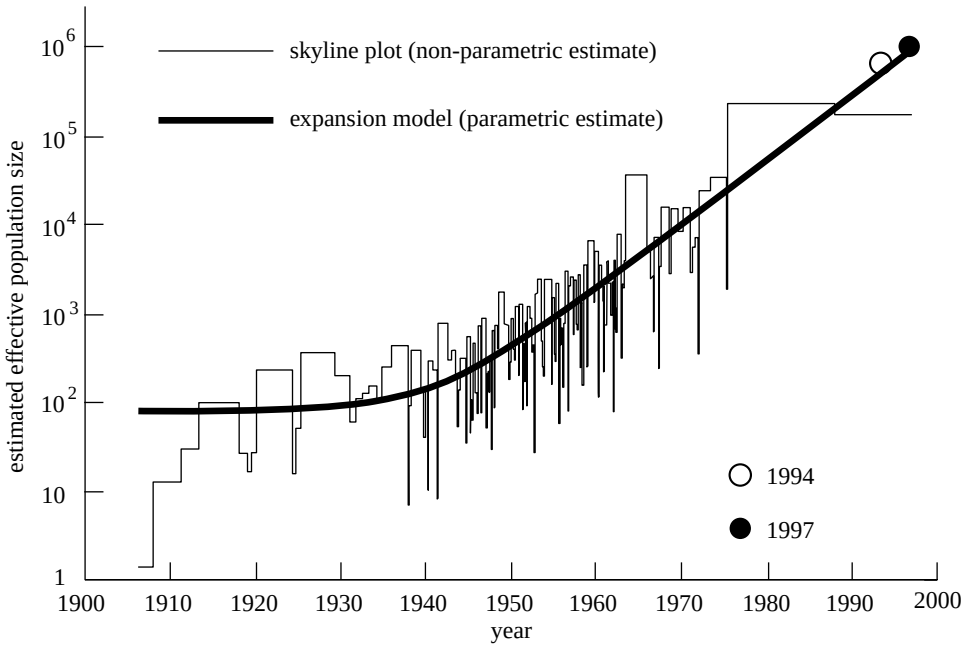

Figure 5. Coalescent results for the DRC data set. The skyline plot (a non-parametric estimate of demographic history) is shown in black. The solid grey curve is the maximum-likelihood estimate of demographic history under the expansion growth model (estimated parameters values were $N_e(0) = 9.24 \times 10^5$, $\alpha = 8.67 \times 10^{-5}$ and $r = 0.1676$ yr^{-1}). $N_e(0)$ represents the estimated effective actual number of infected individuals in 1997 (see § 4b). The two circles represent epidemiological estimates of the actual number of HIV-1 infections in the DRC during 1994 (open circle; Burton & Mertens 1998) and 1997 (filled circle; Schwartlander *et al*. 1999).

the epidemic prior to that time, and would not be able to distinguish between different scenarios. For example, if a large population decreased in size to $\alpha N_e(0)$ and persisted that way for the required number of generations and then followed the expansion model, it would be indistinguishable from a single ancestral virus that came into the population, dispersed slowly or rapidly to $\alpha N_e(0)$, stayed that way for the required number of generations and then followed the expansion model.

The branching order of the genealogy of the 197 DRC sequences was initially reconstructed using the maximum-likelihood method described in Korber *et al*. (2000). Since the coalescent framework we are using requires an absolute time-scale, it was necessary to assume a molecular clock during tree reconstruction. We therefore re-optimized the branch lengths of the maximum-likelihood tree using a general-reversible substitution model with rate heterogenity with a molecular clock enforced (figure 4). By comparing log-likelihood scores, we find that such a strict molecular clock is not justified for this tree; however, it is a necessary baseline assumption for further analysis.

Figure 5 presents the coalescent analysis results for the DRC data. The horizontal axis was converted into the year using the evolutionary rate obtained in § 3(b) (0.0023 substitutions per site per year). In the figure the skyline plot estimate and the maximum-likelihood estimate under the expansion model are superimposed. Both estimates suggest that the number of HIV-1 infections in the DRC has increased exponentially, and that the exponential growth rate has increased through time. Using a likelihood ratio test we could reject the hypotheses of constant size and exponential growth at a fixed rate in favour of the expansion model.

The flattening out of the skyline plot near the present year is possibly an artefact of this technique (Pybus *et al*. 2000). However, Mulanga-Kabeya *et al*. (1998) report epidemiological evidence suggesting that HIV prevalence rates in the DRC remained relatively unchanged between 1991 and 1997. We compared our coalescent estimates of the number of HIV-1 infections in the DRC with existing epidemiological estimates for the years 1994 (680 000 infections; Burton & Mertens 1998), and 1997 (900 000 infections; Schwartlander 1999). Even though we only estimate effective population size (which should be proportional to the actual infected population size), our estimates are similar to these epidemiological results.

5. DISCUSSION

There are two key steps in the generation of a novel epidemic stemming from zoonosis: the initial transfer, and the subsequent adaptation allowing the pathogen to become viable in and transmitted by the new host. Many pathogens move between animal and man, but such movement does not always result in epidemic transmission within the new host species (Meyer 1901; Heymann *et al*. 1998; Voevodin *et al*. 1997; Slattery *et al*. 1999; Callahan *et al*. 1999; Heneine *et al*. 1998; Reid 2000). The timing work discussed here does not prove that initial transfer of virus from primate to man occurred in 1931 (95% CI of 1915–1941; Korber *et al*. 2000; Hillis 2000); instead it indicates that the most recent common ancestor of the HIV-1 M group occurred in that time-frame. The

cross-species transmission might have happened before, concurrently or after that time-point. If the transmission happened before, or if there were multiple cross-species events, perhaps spanning centuries, then the most recent common ancestor indicates a time when a single variant was in the right circumstance with the appropriate genetic potential for its descendants to fuel the pandemic. An alternative scenario is that the most recent common ancestor actually represents an infection in a chimpanzee host, and that multiple diverse strains were subsequently transmitted to man. Epidemic strains are clearly difficult to establish, so this alternative seems less plausible. Additionally, consideration of the phylogenetic relationships among strains within the M group, as well as the pattern and nature of their genetic divergence, all strongly suggest that the M-group ancestor was indeed in the human host (see Sharp *et al.*, this issue).

If the initial transfer of HIV-1 did occur in the first half of the 20th century, then one is still left with the puzzle of why AIDS went undetected until 1981. One answer is that AIDS is a complicated disease, leading to a suppressed immune condition that invites other diseases, and therefore it would difficult to recognize; whether or not this explanation is adequate is arguable. But coalescent theory provides an intriguing framework for addressing such questions because it infers the effective population size throughout the history of the epidemic from modern sequence data. In this study, the analysis of the DRC data resulted in the model of epidemic growth illustrated by the skyline plot in figure 5, suggesting that the early years of the M-group epidemic in this region of Africa were characterized by an extended period of slow spread; HIV-1 was keeping a low profile. This analysis suggests that the kind of spread necessary for HIV to have been present but undetected for an extended period is plausible. The ten-year latency from the time of infection to AIDS (Chevret *et al*. 1992) would add a further delay to the recognition of the emerging disease. Unfortunately, coalescent theory imposes assumptions difficult to attain with real data. As good HIV data sets from human populations continue to be hard to obtain, an alternative area of great promise for deeper understanding of HIV-1's origins is more extensive studies of SIV in chimpanzees and other non-human primates. New analysis and new data will be necessary to elucidate more fully the origin of AIDS.

The research of the Los Alamos authors was supported through internal research funds by the Laboratory Directed Research and Development (LDRD) programme; B.K., K.Y. and M.M. were also supported by the Pediatric AIDS Foundation Elisabeth Glaser Scientist Award programme. This paper is US Government work in the public domain in the United States.

ENDNOTES

[1]UNAIDS 2000. These numbers come from the the Joint United Nations Programme on HIV/AIDS Global HIV/AIDS epidemic update 1999, available at http://www.unaids.org/publications.

[2]FastDNAml and DNArates were written by Gary Olsen and colleagues at the Ribosomal Database Project (RDP) at the University of Illinois at Urbana-Champaign, available by anonymous ftp from ftp://rdp.life.uiuc.edu/.

REFERENCES

Barre-Sinoussi, F., Chermann, J. C., Rey, F., Nugeyre, M. T., Chamaret, S., Gruest, J., Dauguet, C., Axler-Blin, C., Vezinet-Brun, F., Rouzioux, C., Rozenbaum, W. & Montagnier, L. 1983 Isolation of a T-lymphotropic retrovirus from a patient at risk for acquired immune deficiency syndrome (AIDS). *Science* **220**, 868–871.

Burton, A. H. & Mertens, T. E. 1998 Provisional country estimates of prevalent infections as of end 1994: a description of the methods. *Int. J. Epidemiol.* **27**, 101–107.

Callahan, M., Switzer, W., Matthews, A., Roberts, B., Heneine, W., Folks, T. & Sandstrom, P. 1999 Persistent zoonotic infection of a human with simian foamy virus in the absence of an intact ORF-2 accessory gene. *J. Virol.* **73**, 9619–9624.

Chevret, S., Costagliola, D., Lefrere, J. J. & Valleron, A. J. 1992 A new approach to estimating AIDS incubation times: results in homosexual infected men. *J. Epidemiol. Comm. Hlth* **46**, 582–586.

Chitnis, A., Rawls, D. & Moore, J. 2000 Origin of HIV type 1 in colonial French Equatorial Africa? *AIDS Res. Hum. Retroviruses* **16**, 5–8.

Choi, D. J., Dube, S., Spicer, T. P., Slade, H. B, Jensen, F. C. & Poiesz, B. J. 1997 HIV type 1 isolate Z321, the strain used to make a therapeutic HIV type 1 immunogen, is intersubtype recombinant. *AIDS Res. Hum. Retroviruses* **13**, 357–361.

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

Efron, B. & Tibshirani, R. 1991 Statistical data analysis in the computer age. *Science* **253**, 390–395.

Finzi, D. (and 16 others) 1999 Latent infection of CD4+ T cells provides a mechanism for lifelong persistence of HIV-1, even in patients on effective combination therapy. *Nature Med.* **5**, 512–517.

Furtado, M. R., Kinsley, L. A. & Wolinsky, S. M. 1995 Changes in the viral mRNA expression pattern correlate with a rapid rate of CD4+ T-cell number decline in human immunodeficiency virus type 1-infected individuals. *J. Virol.* **29**, 2092–2100.

Gallo, R. C. (and 11 others) 1983 Isolation of human T-cell leukemia virus in acquired immune deficiency syndrome (AIDS). *Science* **220**, 865–867.

Ganeshan, S., Dickover, R., Korber, B., Bryson, Y. & Wolinsky, S. 1997 Human immunodeficiency virus type 1 genetic evolution over time in children with different rates of development of disease. *J. Virol.* **71**, 663–667.

Gao, F. (and 12 others) 1998 An isolate of human immunodeficiency virus type 1 originally classified as subtype I represents a complex mosaic comprising three different group M subtypes (A, G, and I). *J. Virol.* **72**, 10 234–10 241.

Gao, F. (and 11 others) 1999 Origin of HIV-1 in the chimpanzee *Pan troglodytes troglodytes*. *Nature* **397**, 436–441.

Gazzolo, L., Gessain, A., Robin, Y., Robert-Guroff, M. & de-The, G. 1984 Antibodies to HTLV-III in Haitian immigrants in French Guiana. *New Engl. J. Med.* **311**, 1252–1253.

Gottlieb, M. S., Schroff, R., Schanker, H. M., Weisman, J. D., Fan, P. T., Wolf, R. A. & Saxon, A. 1981 *Pneumocystis carinii* pneumonia and mucosal candidiasis in previously healthy homosexual men: evidence of a new acquired cellular immunodeficiency. *New Engl. J. Med.* **305**, 1425–1431.

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

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

Grmek, M. 1990 *History of AIDS emergence and origin of a modern pandemic*. Princeton University Press.

Gunthard, H. (and 11 others) 1999 Evolution of envelope sequences of human immunodeficiency virus type 1 in cellular reservoirs in the setting of potent antiviral therapy. *J. Virol.* **73**, 9404–9412.

Gurtler, L. 1996 Difficulties and strategies of HIV diagnosis. *The Lancet* **348**, 176–179.

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

Heneine, W. (and 11 others) 1998 Identification of a human population infected with simian foamy viruses. *Nature Med.* **4**, 403–407.

Heymann, D., Szczeniowski, M. & Esteves, K. 1998 Re-emergence of monkeypox in Africa: a review of the past six years. *Br. Med. Bull.* **54**, 693–702.

Hillis, D. M. 2000 Commentary. *Science* **288**, 1757–1759.

Hillis, D. M., Huelsenbeck, J. P. & Cunningham, C. W. 1994 Application and accuracy of molecular phylogenies. *Science* **264**, 671–677.

Hillis, D. M., Mable, B. K. & Moritz, C. (eds) 1996 Application of molecular systematics: the state of the field and a look to the future. In *Molecular systematics*, 2nd edn, pp. 515–543. Sunderland, MA: Sinauer Associates.

Holmes, E. C., Nee, S., Rambaut, A., Garnett, G. P. & Harvey, P. H. 1995 Revealing the history of infectious disease epidemics through phylogenetic trees. *Phil. Trans. R. Soc. Lond.* B **349**, 33–40.

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

Hooper, E. 1999 *The river: a journey back to the source of HIV and AIDS*. Boston, MA: Little, Brown & Co.

Huelsenbeck, J. 1995 The robustness of two phylogenetic methods: four-taxon simulations reveal a slight superiority of maximum likelihood over neighbor joining. *Mol. Biol. Evol.* **12**, 843–849.

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

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

Korber, B., Sharp, P. & Ho, D. 1999 Reply to 'Dating the origin of HIV-1 subtypes' by J. Goudsmit and V. Lukoshov. *Nature* **400**, 326.

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

Leitner, T. & Albert, J. 1999 The molecular clock of HIV-1 unveiled through analysis of a known transmission history. *Proc. Natl Acad. Sci. USA* **96**, 10 752–10 757.

Leitner, T., Escanilla, D., Franzen, C., Uhlen, M. & Albert, J. 1996 Accurate reconstruction of a known HIV-1 transmission history by phylogenetic tree analysis. *Proc. Natl Acad. Sci. USA* **93**, 10 864–10 869.

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

McCutchan, F. E., Hegerich, P. A., Brennan, T. P., Phanuphak, P., Singharaj, P., Jugsudee, A., Berman, P. W., Gary, A. M., Fowler, A. K. & Burke, D. S. 1992 Genetic variants of HIV-1 in Thailand. *AIDS Res. Hum. Retroviruses* **8**, 1887–1895.

Meyer, A., Esposito, J., Gras, F., Kolakowski, T., Fatras, M. & Muller, G. 1901 First appearance of monkey pox in human beings in Gabon. *Med. Trop.* **51**, 53–57.

Mokili, J. L., Wade, C. M., Burns, S. M., Cutting, W. A., Bopopi, J. M., Green, S. D., Peutherer, J. F. & Simmonds, P. 1999 Genetic heterogeneity of HIV type 1 subtypes in Kimpese, rural Democratic Republic of Congo. *AIDS Res. Hum. Retroviruses* **15**, 655–664.

Mulanga-Kabeya, C. (and 10 others) 1998 Evidence of stable HIV seroprevalences in selected populations in the Democratic Republic of the Congo. *AIDS* **12**, 905–910.

Nahmias, A. J. (and 10 others) 1986 Evidence for human infection with an HTLV-III/LAV-like virus in central Africa, 1959. *The Lancet* **1**, 1279–1280.

Nzilambi, N., De Cock, K. M, Forthal, D. N., Francis, H., Ryder, R. W., Malebe, I., Getchell, J., Laga, M., Piot, P. & McCormick, J. B. 1988 The prevalence of infection with human immunodeficiency virus over a 10-year period in rural Zaire. *New Engl. J. Med.* **318**, 276–279.

Olsen, G. J., Matsuda, H., Hagstrom, R. & Overbeek, R. 1994 FastDNAml: a tool for construction of phylogenetic trees of DNA sequences using maximum likelihood. *Comput. Appl. Biosci.* **10**, 41–48.

Pape, J. W., Liautaud, B., Thomas, F., Mathurin, J. R., St Amand, M. M., Boncy, M., Pean, V., Pamphile, M., Laroche, A. C. & Johnson Jr, W. D. 1983 Characteristics of the acquired immunodeficiency syndrome (AIDS) in Haiti. *New Engl. J. Med.* **309**, 945–950.

Peeters, M., Honore, C., Huet, T., Bedjabaga, L., Ossari, S., Bussi, P., Cooper, R. W. & Delaporte, E. 1989 Isolation and partial characterization of an HIV-related virus occurring naturally in chimpanzees in Gabon. *AIDS* **10**, 625–30.

Perelson, A. S., Neumann, A., Markowitz, M., Leonard, J. M. & Ho, D. D. 1996 HIV-1 dynamics *in vivo*: virion clearance rate, infected cell life-span, and viral generation time. *Science* **271**, 1582–1586.

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

Pybus, O. G., Rambaut, A. & Harvey, P. H. 2000 An integrated framework for the inference of viral population history from reconstructed genealogies. *Genetics* **155**, 1429–1437.

Reid, A., Fanning, T., Janczewski, T. & Taubenberger, J. 2000 Characterization of the 1918 Spanish influenza virus neuraminidase gene. *Proc. Natl Acad. Sci. USA* **97**, 6785–6790.

Robertson, D. (and 18 others) 2000 HIV-1 nomenclature proposal. *Science* **288**, 55–56.

Rodrigo, A. G., Shpaer, E. G., Delwart, E. L., Iversen, K. N., Gallo, M. V., Brojatsch, J., Hirsch, M. S., Walker, B. D. & Mullins, J. I. 1999 Coalescent estimates of HIV-1 generation time *in vivo*. *Proc. Natl Acad. Sci. USA* **96**, 2187–2191.

Schwartlander, B., Staneski, K. A., Brown, T., Way, P. O., Monasch, R., Chin, J., Tarantola, D. & Walker, N. 1999 Country-specific estimates and models of HIV and AIDS: methods and limitations. *AIDS* **13**, 2445–2458.

Selik, R. M., Haverkos, H. W. & Curran, J. W. 1984 Acquired immune deficiency syndrome (AIDS) trends in the United States, 1978–1982. *Am. J. Med.* **76**, 493–500.

Shankarappa, R. (and 11 others) 1999 Consistent viral evolutionary changes associated with the progression of human immunodeficiency virus type 1 infection. *J. Virol.* **73**, 10489–10502.

Sharp, P. M., Bailes, E., Gao, F., Beer, B. E., Hirsch, V. M. & Hahn, B. H. 2000 Origins and evolution of AIDS viruses: estimating the time-scale. *Biochem. Soc. Trans.* **28**, 275–282.

Simon, F., Mauclere, P., Roques, P., Loussert-Ajaka, I., Muller-Trutwin, M. C., Saragosti, S., Georges-Courbot, M. C., Barre-Sinoussi, F. & Brun-Vezinet, F. 1998 Identification of a new human immunodeficiency virus type 1 distinct from group M and group O. *Nature Med.* **4**, 1032–1037.

Slattery, J. P., Franchini, G. & Gessain, A. 1999 Genomic evolution, patterns of global dissemination, and interspecies transmission of human and simian T-cell leukemia/lymphotropic viruses. *Genome Res*. **9**, 525–540.

Smith, D. 1990 Thailand: AIDS crisis looms. *The Lancet* **335**, 781–782.

Srinivasan, A. (and 10 others) 1989 Molecular characterization of HIV-1 isolated from a serum collected in 1976: nucleotide sequence comparison to recent isolates and generation of hybrid HIV. *AIDS Res. Hum. Retroviruses* **5**, 121–129, 1989.

Subbarao, S. (and 10 others) 1998 HIV type 1 in Thailand, 1994–1995: persistence of two subtypes with low genetic diversity. *AIDS Res. Hum. Retroviruses* **14**, 319–327.

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

Thorne, J. L., Kishino, H. & Painter, I. S. 1998 Estimating the rate of evolution of the rate of molecular evolution. *Mol. Biol. Evol*. **15**, 1647–1657.

Vandamme, A. M., Strimmer, K., Hall, W. W., Delaporte, E., M'Boup, S., Peeters, M. & Salemi, M. 2000 Dating the origin of HIV-1 group M and HIV-1 group M/SIVcpz separation. *Seventh Annual Int. Meeting on HIV Dynamics and Evolution, Seattle, Washington, USA, March 2000*, p. 4. Seattle, WA: University of Washington. (See http://ubik.microbiol.washington.edu/Seattle2000/abstracts/abstract4.html.)

Vidal, N., Peeters, M., Mulanga-Kabeya, C., Nzilambi, N., Robertson, D., Ilunga, W., Sema, H., Tshimanga, K., Bongo, B. & Delaporte, E. 2000 Unprecedented degree of HIV-1 group M genetic diversity in the Democratic Republic of Congo suggests that the HIV-1 pandemic originated in Central Africa. *J. Virol*. **74**, 10 498–10 507.

Voevodin, A., Johnson, B., Samilchuk, E., Stone, G., Druilhet, R., Greer, W. & Gibbs Jr, C. 1997 Phylogenetic analysis of simian T-lymphotropic virus type I (STLV-I) in common chimpanzees (*Pan troglodytes*): evidence for interspecies transmission of the virus between chimpanzees and humans in Central Africa. *Virology* **238**, 212–220.

Wangroongsarb, Y., Weniger, B., Wasi, C., Traisupa, A., Kunasol, P., Rojanapithayakorn, W. & Fucharoen, S. 1985 Prevalence of HTLV-III/LAV antibody in selected populations in Thailand. *Southeast Asian J. Trop. Med. Publ. Hlth* **16**, 517–520.

Wolinsky, S., Korber, B., Neumann, A., Daniels, M., Kuntsman, K., Whetsell, A., Furtado, M., Cao, Y., Ho, D., Safrit, J. & Koup, J. 1996 Adaptive evolution of human immunodeficiency virus type 1 during the natural course of infection. *Science* **277**, 537–542.

Yang, Z. Maximum likelihood models for combined analyses of multiple sequence data. *J. Mol. Evol*. **42**, 587–596.

Yang, Z., Goldman, N. & Friday, A. 1994 Comparison of models for nucleotide substitution used in maximum-likelihood phylogenetic estimation. *J. Mol. Evol*. **11**, 316–324.

Zhang, L., Ramratnam, B., Tenner-Racz, K., He, Y., Vesanen, M., Lewin, S., Talal, A., Racz, P., Perelson, A., Korber, B., Markowitz, M. & Ho, D. 1999 Quantifying residual HIV-1 replication in patients receiving combination antiretroviral therapy. *New Engl. J. Med*. **340**, 1605–1613.

Zhu, T., Korber, B. T., Nahmias, A. J., Hooper, E., Sharp, P. M. & Ho, D. D. 1998 An African HIV-1 sequence from 1959 and implications for the origin of the epidemic. *Nature* **391**, 594–596.