

# Timing the Ancestor of the HIV-1 Pandemic Strains

B. Korber,<sup>1,2,\*†</sup> M. Muldoon,<sup>2,3</sup> J. Theiler,<sup>1</sup> F. Gao,<sup>4</sup> R. Gupta,<sup>1</sup>  
A. Lapides,<sup>1,2</sup> B. H. Hahn,<sup>4</sup> S. Wolinsky,<sup>5</sup> T. Bhattacharya<sup>1†</sup>

HIV-1 sequences were analyzed to estimate the timing of the ancestral sequence of the main group of HIV-1, the strains responsible for the AIDS pandemic. Using parallel supercomputers and assuming a constant rate of evolution, we applied maximum-likelihood phylogenetic methods to unprecedented amounts of data for this calculation. We validated our approach by correctly estimating the timing of two historically documented points. Using a comprehensive full-length envelope sequence alignment, we estimated the date of the last common ancestor of the main group of HIV-1 to be 1931 (1915–41). Analysis of a gag gene alignment, subregions of envelope including additional sequences, and a method that relaxed the assumption of a strict molecular clock also supported these results.

Current evidence indicates that human immunodeficiency viruses (HIV-1 and HIV-2) entered the human population through multiple zoonotic infections from simian immunodeficiency virus (SIV)-infected nonhuman primates (1). The simian reservoirs for the human viruses have been defined (1–3) on the basis of genetic relatedness to known SIVs and the close coincidence between the natural primate host range and the geographic region of the greatest diversity of HIV-1 or HIV-2. HIV-1 is most closely related to SIVcpz isolated from the chimpanzee subspecies *Pan troglodytes troglodytes* (1, 2), the most diverse forms of HIV-1 [the HIV-1 main (M), outlier (O), and N (non-M, non-O) groups] are all found in the geographic region corresponding to the range of *P. t. troglodytes* in west equatorial Africa (4, 5), and HIV-1 groups and SIVcpz sequences are interspersed in phylogenetic trees, suggesting that there are shared viral lineages in human and chimpanzee (1, 2). Analogously, HIV-2 is most closely related to SIVsm from sooty mangabeys, the natural range of sooty mangabeys overlaps the geographic region where the most diverse HIV-2s have been identified, and HIV-2 and SIVsm sequences are interspersed in phylogenetic trees (1–3). The HIV-1 M or “main” group has been further broken down into genetically associated clades, referred to as HIV-1 subtypes A

to K [for a current summary of subtypes, see (6)].

Movement of pathogens between species is a common phenomenon, and many viruses have made the leap between simians and humans, including monkeypox (7), simian T cell leukemia virus/human T cell leukemia virus (8), and simian foamy virus (9). Such transspecies infections can be a dead end, with the virus not readily transmitted in the new host species. This may account for the scarcity of N group HIV-1 (5) and of some forms of SIV/HIV-2 that have only been found in a single human each (3). On the other hand, the HIV-1 M group has created a global crisis, already infecting about 50 million people and leaving 16 million dead in its wake (10). Understanding elements that may have contributed to such extreme differences in outcome may help avert future epidemics.

Previous attempts to extrapolate the age of a common ancestor and the time for dispersion of the HIV-1 M group have been hampered by the very limited HIV-1 sequence data (11, 12) that antedates the discovery of HIV-1 in 1983 (13). Furthermore, application of evolutionary models shown to be most appropriate for circumscribed HIV-1 data sets (14) were too computationally intensive for large data sets. Thus, it has been difficult to derive a reliable time scale for molecular evolution of HIV-1 at the population level, yet such a time scale would be useful for understanding the circumstances surrounding the emergence of the acquired immune deficiency syndrome (AIDS) pandemic and the rate at which HIV-1 diverges.

Here we reconstructed the evolutionary history of HIV-1 to estimate the age of the last common ancestor of the HIV-1 M group. We made several improvements relative to earlier attempts to estimate primate lentiviral ages based on a molecular clock, the assumption of a constant rate of evolution (15, 16).

First, we extended established phylogenetic algorithms (17–21) and adapted maximum-likelihood tree-building code to parallel computers. By doing this, it became feasible to analyze very large HIV-1 sequence sets with sophisticated evolutionary models. Envelope sequences from 159 individuals with known sampling dates were available for analysis when we initiated the study; we also did supporting, independent analysis based on 66 gag genes (22, 23). Full-length gene alignments were selected as our primary data, because longer sequences yield more accurate phylogenetic reconstructions. Second, because of the in vivo biology of HIV-1, we developed a model that incorporated uncertainty in the time of origin of a given viral sequence that allows for periods of evolutionary quiescence within a reservoir of persistently infected cells (24–26). Third, we incorporated a bootstrap method to provide 95% confidence intervals (CIs) on our timing and rate-of-evolution estimates (16, 27). Finally, we tested the validity of our approach and the soundness of our assumptions by correctly estimating the time frame of two historically documented points in our phylogenetic reconstructions. Although it is unrealistic to expect that HIV-1 evolution will always rigidly adhere to a molecular clock (16, 28), it is, however, the average behavior of many sequences that we consider here, and our control estimates of known times were accurate.

The phylogenetic methodology we developed to address these questions for HIV is general and can be applied to other evolutionary studies and other organisms. We introduced methods for establishing CIs on timing estimates, incorporated biologically motivated evolutionary models, and allowed for periods of evolutionary quiescence. We extended phylogenetic code capabilities for parallel computing and further developed and tested preexisting strategies that allow the evolutionary rate itself to evolve. Great strides in sequencing technologies have resulted in vast numbers of new genetic sequences, and it is imperative that the analysis tools keep pace with the DNA sequencing technology. The tools developed here extend the power either to simultaneously analyze large data sets representing many different organisms or to analyze multiple genes from a highly variable pathogen such as HIV.

**Phylogenetic methodology.** Our main analysis relies on the assumption of a molecular clock; this hypothesis postulates that molecular change is a linear function of time and that substitutions accumulate according to a Poisson distribution (20). Use of this strategy to estimate the time to a common ancestor in a phylogenetic tree has inherent limitations beyond the fact that evolutionary rates may, in fact, vary [reviewed in (20, 16)]. Examples

<sup>1</sup>Los Alamos National Laboratory, Los Alamos, NM 87545, USA. <sup>2</sup>Santa Fe Institute, Santa Fe, NM 87501, USA. <sup>3</sup>Department of Mathematics, University of Manchester Institute of Technology, Manchester M60 1QD, UK. <sup>4</sup>Department of Medicine and Microbiology, University of Alabama at Birmingham, Birmingham, AL 35294, USA. <sup>5</sup>Division of Infectious Diseases, Department of Medicine, Northwestern University, Chicago, IL 60611, USA.

\*To whom correspondence should be addressed.

†These authors contributed equally to this work.

of problems are that branch lengths from the tips of trees to internal nodes are not truly independent because of common ancestral branches and that undetected recombination events can muddle the evolutionary relationships in the tree (29). In our analysis, inter-subtype recombinants could be excluded, as they are readily detectable. Despite these limitations, a relatively constant average rate of evolution of viral sequences within individuals who have typical rates of disease progression has been reported (24), and an overall linear rate of evolution was observed at the population level in a Dutch cohort of HIV-1 seroconvertors sampled over a 12-year period (30). Also, estimates of divergence times of HIV-1 sequences with a known transmission history have been accurate under the assumption of a molecular clock, provided that a good evolutionary model and a "pretransmission interval" were incorporated (14); a pretransmission interval accounts for the possibility that HIV-1 sequences sampled from a donor and a recipient may share a last common ancestor that existed in the donor sometime before the transmission event. However, cognizant that different rates of evolution may occur in different lineages (31), we also used an adaptation of a method that relaxed the assumption of a strict molecular clock (21).

The importance of an adequate evolutionary model for accurate phylogenetic reconstructions has been repeatedly demonstrated (14, 32). Evolutionary models incorporate different rates of substitution between bases, different site-specific rates of evolution (14, 19, 20, 32, 33), and different base frequencies (skewed toward high adenosine content in HIV-1 sequences, with 35% A in gag and 37% A in env). Realistic formulation of the evolutionary model is particularly critical in situations where branch lengths are important and evolutionary events are dated (14, 34). The likelihood of the data for the best tree estimated under a particular model can be used to determine if the added complexity introduced by increasing the number of parameters in an evolutionary model is justified for a particular data set (35).

Maximum-likelihood trees based on many taxa are extremely computationally intensive. It would not have been feasible to test alternative evolutionary models with previously available serial implementations of maximum-likelihood tree-building code for the complete envelope sequence alignments (36). We modified G. Olsen's fastDNAmI code (18), a rapid implementation of PHYLIP maximum-likelihood code (17), to incorporate a general-reversible (REV) base-substitution model (32) and to more efficiently use parallel processors (23). A REV model incorporates a reversible substitution rate for each pair of bases. Calculations were performed on up to

512 nodes of the SGI Origin 2000 cluster, called Nirvana, at the Los Alamos Advanced Computing Laboratory ([www.acl.lanl.gov/](http://www.acl.lanl.gov/)).

Following the strategy implemented in the DNARates program (18, 19), we developed a maximum-likelihood method to model rate heterogeneity between sites as an alternative to the more commonly used method of modeling rate heterogeneity among sites based on a gamma distribution (19, 32, 37). The maximum-likelihood model of rate variation among sites is parameter-rich, i.e., as many parameters are estimated from the data as there are positions in the alignment (23). However, a direct comparison with the gamma distribution for the smaller gag sequence analysis with the phylogeny package PAUP (38) indicated that despite the large increase in number of parameters, we were not overfitting the data and our method of assigning rate categories was preferable to using the gamma distribution (23, 39).

The root of the M group is defined as the branching point at which divergence from the common ancestor of the lineage first occurs. The trees generated by our methods are unrooted, and the use of a reversible (REV) model requires additional information to locate the root of the tree. The root is traditionally defined as the branch position of an "outgroup," where an outgroup is specified as a sequence or sequence set that is known to be external to the lineage under consideration (19). The SIVcpz sequences isolated from *P. t. troglodytes* (2) are the closest sequences external to the HIV-1 M group and so are best suited for use as an outgroup, but even these sequences are very distant. Possibly because of this distance, the place where SIVcpz outgroups join the M group is highly unstable (11, 40) and often deep into a well-defined subtype (in this study, A, H, or B, depending on the tree). Such a tree considered in isolation could lead one to the conclusion that the subtype nearest to the root either has a much slower rate of evolution or that the rate of evolution in the internal branches linking the various subgroups was unusually large. We find, however, that the CPZ outgroup position is in a shallow likelihood surface and that such subtype-associated outgroup positions are not statistically significant. Furthermore, the subtype associated with outgroup can differ depending on the phylogenetic methodology, the region of the gene considered, whether the full alignment or third codon positions only (mainly silent substitutions) were included in the alignment, or if different combinations of single and multiple outgroup sequences were used (41); thus, we thought that it was critical to use an alternative strategy.

We therefore constructed a parsimonious candidate for the ancestral sequence, which is the consensus sequence (most common base) of the consensus sequences from each subtype, and used this as an outgroup. This

construction avoids bias resulting from oversampling of some clades relative to others; the resulting sequences did not have preferential association to any clade (Fig. 1C) or have altered base composition relative to natural viral sequences. This outgroup strategy should place the root on a central interior branch, and indeed, our analysis found this (Figs. 1A and 2A). In contrast to the SIVcpz outgroup, the root position defined by a consensus outgroup was stable. For all the sequence sets we tested, the branch length from the consensus to the estimated root was very near zero, as would be expected from a true ancestral sequence (Figs. 1A and 2A). Finally, we considered a third outgroup strategy: We held the branching order obtained with the consensus-outgroup fixed, then replaced the consensus outgroup by the CPZ.US sequence, and then reoptimized the branch lengths. This produced very similar likelihood scores to the trees constructed with CPZ.US as the outgroup from the start, indicating that root positions determined by the consensus sequences were, within statistical fluctuations, compatible with those obtained by the conventional method (23).

The most clocklike behavior, defined as the strongest linear relation when plotting evolutionary distance and time, was observed in trees constructed with the consensus as the outgroup, as expected (see the likelihood of the data assuming a linear relation between branch length and time, presented in Table 1). The more conventional trees constructed by replacing the consensus by the CPZ.US sequence and retaining the same topology while recalculating the branch lengths gave similar, although more dispersed and less "clocklike" results. Arguments can be made in favor of using either method; therefore, we present the full results using both strategies in Table 1.

**Timing estimates assuming a molecular clock.** Given a phylogenetic tree and assuming a uniform rate of evolution, one can plot total branch length (from the tips of the branches to the ancestral node) against the year of sampling and fit a line through the data points. From this inferred linear relation, one can project back to estimate the time associated with zero branch length, i.e., the time of the ancestral sequence. For a standard linear least-squares fit to a line, one 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). But the "error-only-on-branch-length" model for our data fails a goodness-of-fit test: The data are overdispersed compared with a Poisson distribution (20). Furthermore, for HIV-1 sequences, the time a sequence originated is not precisely known for two reasons: The sampling time is generally only recorded to a precision of 1 year,

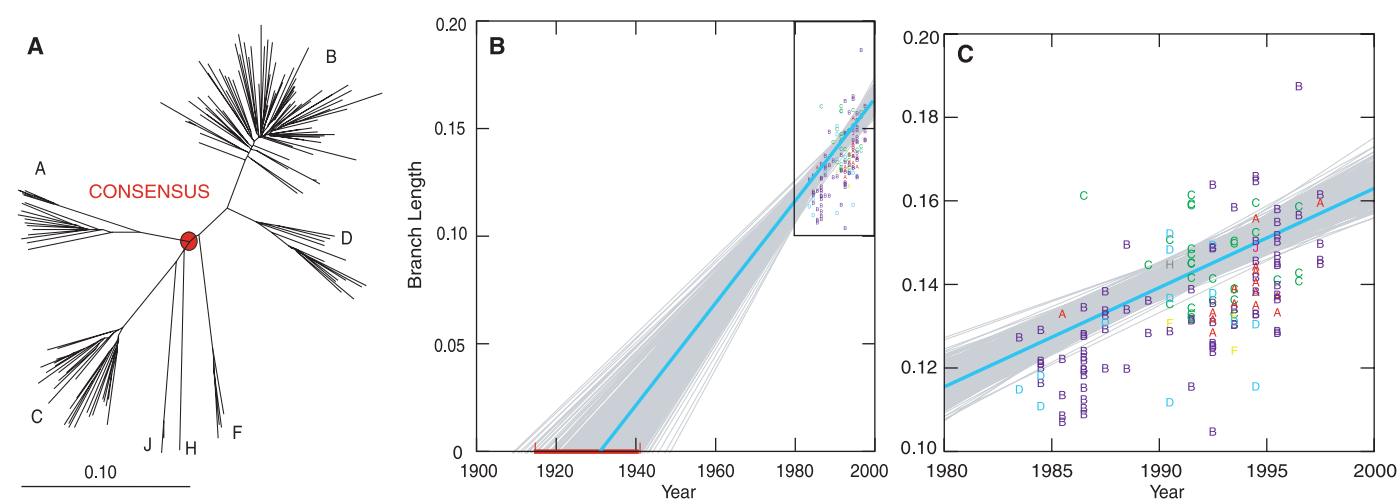
and uncertainty arises from the nature of HIV-1 evolution within a single host. An HIV-1 provirus can be harbored, not evolving, for an extended period of time in persistently infected cells (26), and so viral DNA sampled in a given year may actually have an origin some years earlier (24, 25). Our linear fit was thus based on a combination of the two sources of time uncertainty described above, and a Poisson distribution for branch length error (20) (Fig. 3). Our best fit line maximizes the likelihood of the data given a model with parameters for slope, intercept, and the time scale ( $\tau$ ) of the harbored sequences. The parameter  $\tau$  (Fig. 3), or the exponential decay time for the "age" of sampled sequences, had an average value estimated from the data of 3.4 years with a 95% CI of 1.7 to 7 years, a very reasonable estimate given what is actually observed in intrapatient data sets (24, 25). This method provided a likelihood score for the data given our linear model, which enabled us to compare the relative clocklike behavior of different trees (see the discussion above comparing outgroup strategies and the log-likelihood scores given in Table 1; it is by this measure that the consensus outgroup was more clocklike than the CPZ outgroup). Bootstrap resampling of the data points (480 trials) provided 95% CIs for our estimated timing of ancestral sequences and evolutionary rates (16, 27) (Figs. 1B and 2B and Table 1); Monte Carlo resampling from the estimated model rather than from the data points was also tested and produced similar results (41).

**Validation by comparisons of predicted dates to documented dates.** Our first control

case was the viral sequence ZR.59, obtained from a blood sample collected in 1959 in the Democratic Republic of the Congo (formerly Zaire) (11). This comparison is important as a control, because we could estimate a known date in the tree that was decades before the rest of the data accrued (Fig. 2). The alignment used for this analysis was restricted to short fragments of the envelope sequence available from the 1959 sample (23). With the ZR.59 date of origin treated as unknown, we calculated an evolutionary rate based on the rest of the data, and then the time of sampling of the 1959 sequence was estimated from its branch length relative to the root. Our estimate for the year of origin of the sequence was 1957 (95% CI 1934–62) (Fig. 2) when the consensus was used as the outgroup; when the CPZ.US replaced the consensus as outgroup, the estimate was 1960 (1943–67). The accuracy of these estimates suggests both that the central location of the root position was reasonable and that our assumption of a molecular clock is consistent with the data, at least back to 1959. However, it could be argued that evolutionary rates might have been variable in more ancient parts of the tree, particularly soon after a cross-species transmission event. In one study that directly addressed that question, the rate of SIVsm evolution was compared in infections of sooty mangabeys, the natural host for SIVsm, and in rhesus macaques, a new host in which SIVsm causes AIDS; the evolutionary rates were found to be slightly slower in the new host, the macaques (42). This suggests that the date of transmission might be farther back in time than our estimate.

The ancestral node of the Thai E subtype is the only internal node in the HIV-1 M group tree that has both adequate information to provide an associated date and enough subsequent sequence data to allow one to make a timing estimate. Two lines of evidence suggest an approximate date for the founder virus of the E subtype in Thailand: epidemiological and sequence data. The first HIV-1-infected people and people with AIDS in Thailand were found in the mid-1980s among a small number of male prostitutes and in two patients with thalassemia (43). Given the risk groups (44) and the fact that subtype E first appeared in the northern part of the country, not in Bangkok, these infections were likely to have been due to subtype B viruses, not subtype E. No HIV-1 was found in the mid-1980s in Thailand in the risk groups where the E subtype ultimately took hold. Hundreds of people at high risk for HIV-1 infection were tested in 1985; furthermore, in 1986, no evidence of HIV-1 was found among thousands of Thais who were tested (43). Yet by 1988, over 5000 Thais were infected with HIV-1, and by 1989–90, E subtype HIV-1 was clearly spreading rapidly in the heterosexual population (43). In addition, DNA sequence analysis showed that the Thai E subtype viruses sampled in 1990 were genetically highly related (45), suggesting that a single founder virus seeded the Thai E subtype epidemic a few years before 1990. Combined, this information suggests a single founder subtype E virus some time near 1986–87.

Subtype E is a mosaic virus, with some regions that cluster closely with subtype A



**Fig. 1.** Estimating the last common ancestor of the HIV-1 M group on the basis of data collected over the last two decades. **(A)** The gp160 phylogenetic tree used for this calculation. **(B)** The branch lengths from each leaf to the root of the tree are plotted against time. The subtype of the sequence is indicated by colored letters. A maximum-likelihood linear fit to the data was generated as described in the text (bold turquoise line). These lines prove to be steeper and offset to the left (into the past) relative to lines that would be drawn with a "standard" linear least squares fit to a line. As it turns out, these two effects, slope and offset,

almost balance each other out for the projections to zero branch length for the M group; for more recent sampling times, the offset in time would be more important, and for times in the more distant past, the slope would dominate. Four hundred eighty bootstrap fits to data points were used to calculate 95% CIs, shown as a red line along the horizontal axis. **(C)** A magnified view of the boxed region in **(B)**, showing that the points derived from different subtypes tend to be reasonably well distributed about the line, a consequence of the approximate equality of the intraclade evolutionary rates.



and other regions that are distinctive and classified as subtype E (6, 46). To avoid inclusion of such mosaic sequences in the analysis, only the distinctive E-like subregion of the envelope gene is included in the alignment. Thus, as for the case of ZR.59, only a portion of the envelope gene was used. Assuming a molecular clock, we estimated the common ancestor of the HIV-1 E subtype virus in Thailand to be 1986 (95% CI 1978–89) on the basis of extrapolating back from E subtype sequences or 1984 (1980–86) on the basis of the distance from the Asian E subtype ancestor to the M group root, using the full set of M group sequences to estimate the rate of evolution (Table 1). These estimates are in excellent accord with the molecular and epidemiological evidence.

An additional, albeit less well-established control point was based on a timing estimate of the B-D ancestral sequence. The ZR.59 branch point indicates that the B and D subtypes split before 1959 (11) (Fig. 2A). Estimates of the B-D common ancestor, even for trees that did not include ZR.59, were all well before 1959, and consistent with this upper bound (Table 1).

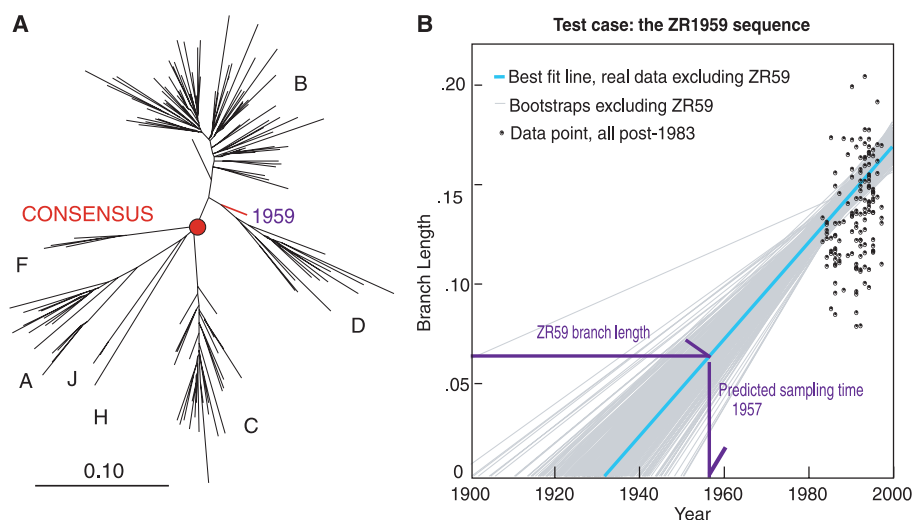
In summary, although there may be some fluctuations in the rate of evolution among HIV-1 lineages, these three observations (the ZR.59 sequence, the Thai E subtype sequence, and the B-D common ancestor sequence) provide convincing evidence that our method was stable enough for estimating the date of ancestral nodes.

**Estimating unknown ancestral dates assuming a molecular clock.** The majority of our analyses suggest that the last common ancestor of the HIV-1 group M was near 1930, with 95% CIs roughly spanning the first half of the 20th century, depending on the input data and analysis. Table 1 summarizes the complete results for the full-length HIV-1 envelope and gag gene analyses and also the analysis of the ZR.59 and Thai E subtype viral sequences used as controls. Figure 1 shows the full-length envelope sequence analysis, the most informative of the analyses because of the greater nucleotide sequence length and the greater number of taxa. The gag sequence data set had fewer than half the data points and yielded broader CIs; nonetheless, it provided independent support for the timing estimates based on the envelope sequence data. The two validation data sets, ZR.59 and the E subtype, were also used to predict the M-group timing, because ZR.59 provided one additional older time point and the E subtype analysis provided multiple additional viral sequences collected during the 1990s. These results are not independent of the gp160 results, as they are based on fragments of the envelope gene, but are included for comparison. There was a single outlier with a most likely value of 1908

in the eight estimates that assumed a molecular clock (four data sets, two outgroup strategies each, Table 1); this is near the boundary of the 95% CI, and such occasional fluctuations are statistically expected.

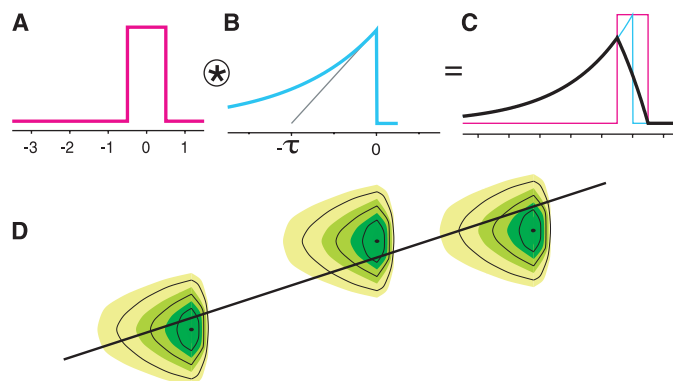
The timing of the origin of the B subtype that dominates the epidemic in the United States and Europe is also of interest. Our best estimate, from the full-length envelope tree analyses assuming a molecular clock, suggests that the founder of the B subtype in the United States originated in 1967 (95% CI 1960–71) (see Table 1 for the complete B subtype results based on all the trees in this study). The seven B subtype sequences from Haiti included in this study branch off earlier than do the other B subtype sequences; this could reflect an older epidemic or possibly be a sampling artifact. These estimates, albeit

earlier than previously thought, are not inconsistent with the epidemiology of the early AIDS cases in the United States and Haiti. AIDS was first identified as a clinical syndrome in 1981 (47). Twelve AIDS cases were retrospectively identified in 1978–79, and by the first quarter of 1983, there were already 1299 cases of clinical AIDS reported in the United States, spread over 35 states (48). In both Haiti and the United States, scattered cases of HIV-1 infection or AIDS were identified in the late 1970s with a handful of possible and probable cases noted in the United States and Haiti between 1972 and 1976 (48, 49). Thus, our timing estimates for an ancestral sequence are plausible, given that HIV-1 infection has an average 10-year asymptomatic period before AIDS develops (50). Our analyses suggest that a commonly



**Fig. 2.** Estimating a known time of sampling: the 1959 sequence. (A) The maximum-likelihood tree. (B) Estimate of the time of origin of the 1959 sequence based on its branch length. A linear fit to these data points, excluding the 1959 sequence, was made with the strategy described in the text (bold turquoise line). The 480 bootstrap replicates were generated by random-with-replacement resampling of the data points to estimate the 95% CI (light gray lines). On the basis of the branch length of the 1959 sequence (horizontal purple arrow), its time of origin was estimated (vertical purple arrow) to be 1957 (95% CI 1934–62).

**Fig. 3.** Error structure for likelihood estimates. (A) There was a 6-month uncertainty in our date of sampling because it was reported only to the nearest year. (B) The delay between the effective age of a sequence and the chronological date of sampling was assumed to be an exponentially distributed random variable with a decay rate of  $\tau$ , where  $\tau$  is a free parameter of the model. (C) The final distribution of uncertainty in time of origin was taken to be a convolution of these two uncertainties. (D) The two-dimensional probability density for a single data point was the product of the time error shown in (C) and a Poisson error on the branch length (vertical axis). The log-likelihood for a model is the sum of the contributions from each point given by the integral of this probability density along the model line.



held view that there was a single founder virus of the HIV-1 B subtype epidemic in the United States in 1976–78 is unlikely. It is more probable that there was a 5- to 15-year “preepidemic” period of subtype B evolution, in which a small number of HIV-1 subtype B infections were present but clinically unrecognized.

The rate of evolution is estimated by the slope of the line relating branch length to time and will depend on the gene region under study

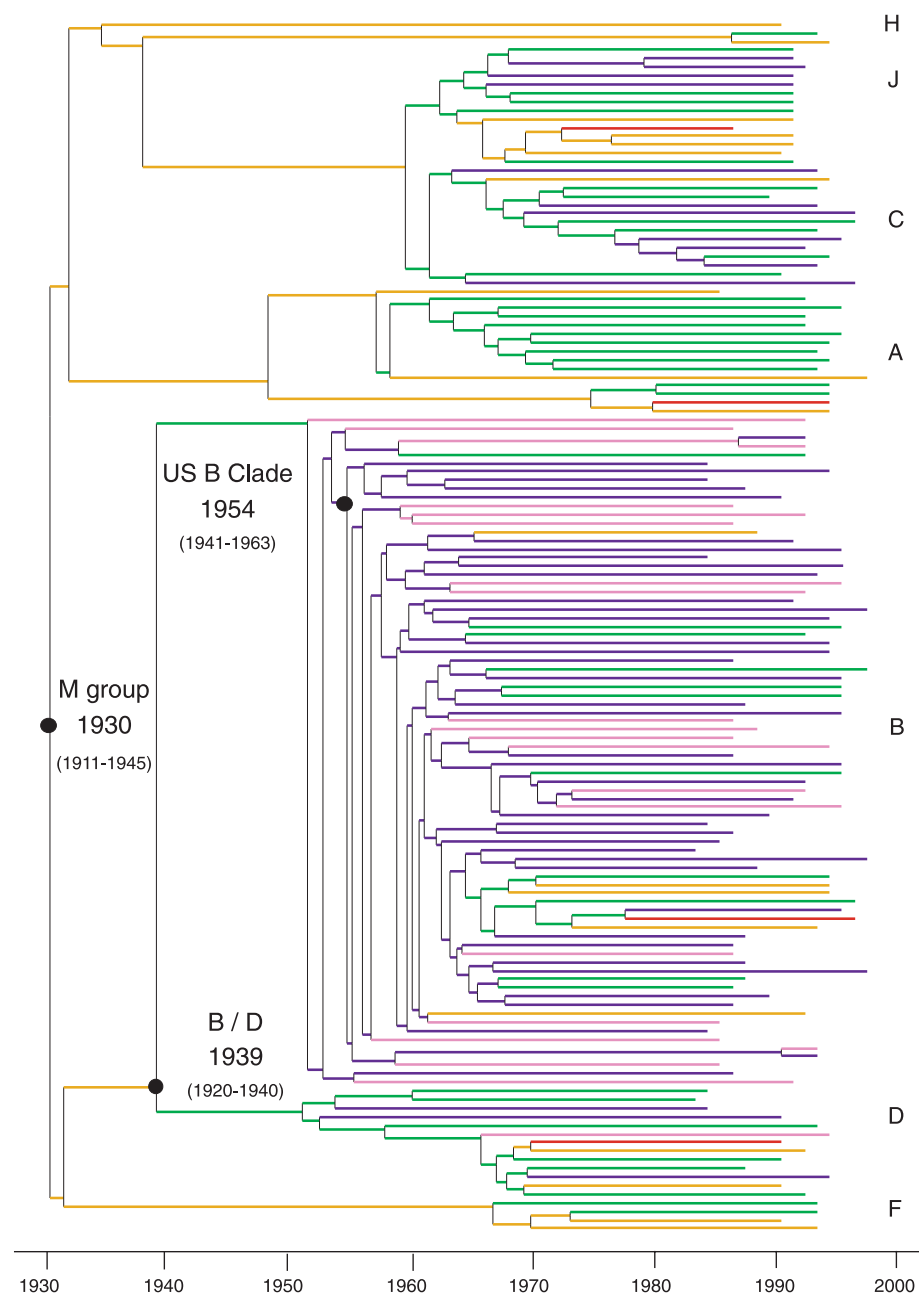
and on the evolutionary model used to calculate the genetic distances. Assuming a molecular clock, we estimated the rate of evolution to be 0.0024 (0.0018 to 0.0028) substitutions per base pair per year for gp160 envelope and 0.0019 (0.0009 to 0.0027) for gag. The envelope gene estimate is similar to our earlier estimate for gp160 of 0.0017 (16), but lower than the estimates of Leitner *et al.* (14), for V3 ( $0.0067 \pm 0.0021$ ) and p17 ( $0.0027 \pm 0.0005$ ) (14). V3 and p17 are relatively variable subre-

gions of envelope and gag, and our estimates spanned the full genes, which are on average more conserved. The p17 region in our gag alignment was 35% more variable than the entire gag gene. Although the V3 region in our alignment was no more variable than the rest of env, the most variable parts of V3 were removed by “gap stripping,” deleting all positions where a gap is inserted in any sequence to maintain the alignment. Such gaps tend to occur in variable positions; in total, 25% of the V3 region positions were omitted from our analysis through gap stripping. Because the results of Leitner *et al.* (14) were based on a small set of related samples, most V3 variable positions would have been retained in their study. As rate estimates can depend on the gene region, the alignment, the model, and the degree to which variable regions are gap stripped, our results do not contradict previously published rates approaching 1% per year (24, 51) in other studies.

**Results allowing the evolution of evolutionary rates.** Because the rate of evolution may vary in a tree, we felt it was important to reevaluate our estimates using a Bayesian approach recently developed by Thorne *et al.* that can take such variation into account (21). This approach relaxes the molecular-clock assumption and allows the rate of evolution itself to evolve; rates on progeny branches are assumed to be close to those of their parents (23). This method yields a tree with branch lengths that represent separation in time, rather than genetic distances (Fig. 4). Estimated times for each internal node, evolutionary rates for each branch, and standard deviations and 95% CI are calculated (23).

We adapted this strategy for use in our study of HIV evolution by adding three features. First, it was essential to incorporate information about the variation in the ages of the leaf nodes. This is because, in contrast to the ancient tree lineages used as an example by Thorne *et al.* for which sampling time can be considered uniform (the present), the period of 20 years over which our HIV sequences were gathered encompasses substantial molecular evolution. Second, we modified the code to use as input the maximum-likelihood tree topologies based on the evolutionary model we found to be optimum (REV with a maximum-likelihood estimate of the relative rate of variation of different sites); the original code used an F84 model and assumed a constant rate of variation across sites (21). Third, we enabled the code to estimate the timing of a leaf node, so we could use this method to estimate time of origin of the 1959 sequence as a control.

As this method is a Bayesian approach, it requires as input a “prior” probability distribution for the parameters of the model; the result is a “posterior” distribution that depends on the prior and the data. We



**Fig. 4.** Env gp160 tree allowing variable rates of evolution. Estimates of the times of the nodes are averages generated with a Metropolis algorithm; this tree had an M group prior distribution of  $1930 \pm 35$  years and  $\nu$  of  $0.03 \pm 0.015$ . The trees were very robust against changes in the prior distribution on the age of the root node or the rate heterogeneity parameter,  $\nu$  (see Table 2). The B clade node that is marked is ancestral to all sequences from the United States included in this tree. The fastest evolutionary rates were 2.7-fold faster than the slowest in this tree, and the colors represent different categories from fastest to slowest: red, orange, green, blue, and violet.

considered a broad range of priors for both the age of the M group node and the parameter  $\nu$ , which governs the extent of rate variation between branches. Our results are quite robust; the timing estimates for the M group were insensitive to these priors, converging to very similar estimates over the range of input values tested (Table 2).

The M group node from the gp160 tree (Table 2) was estimated at 1930 (95% CI 1911–45) with midrange values of the priors. These results are in excellent agreement with those presented in Table 1 assuming a strict molecular clock. The same consistency was also true of the timing estimates of the United States B subtype node from the gag tree, 1960 (1939–72). The B subtype estimates from the

gp160 envelope-based tree are somewhat further back, however, 1954 (1941–63). Additionally, the estimates of known historical dates in our trees, both the 1959 sequence and the Thai E subtype origin, receded by about a decade (23), inconsistent with the historical record. Also, timing estimates based on trees including Thai E sequences were found to be unduly sensitive to  $\nu$ ; we speculate that this might be a result of the extremely short branch lengths between sequence sampling dates and their nearest ancestral node (some branches approached zero length) found among some of the early E subtype sequences. For further discussion of the E subtype results, see the Web site accompanying this paper (23). Because of this, it is apparent that the systematic errors in this method relaxing the molecular clock are not fully controlled, and so we have greater confidence in the results that assume a strict molecular clock (Table 1). The date most important to this analysis, the origin of the M group node, was highly consistent between methods. It is the node constrained by the largest amount of data for the Bayesian analysis and so is most likely to be most reliable.

**Discussion.** To examine hypotheses about the origin of the HIV-1 M group in the human population, we adapted well-established but computationally intensive maximum-likelihood phylogenetic algorithms for use on a parallel supercomputer. This enabled us to produce maximum-likelihood trees based on unprecedented amounts of HIV-1 sequence information and to use an REV model and allow rates of evolution to vary at different positions. We then estimated dates of internal nodes in the tree, by either assuming a constant rate of evolution or relaxing that assumption. Our estimates of the last common ancestor of the HIV-1 M group point to the first half of the 20th century (Tables 1 and 2). The age of the established viral sequences, used as controls, was better estimated under the assumption of a strict molecular clock than when we relaxed this assumption.

Our analyses do not specifically address the question of when the simian progenitor lentivirus trafficked between species; i.e., they do not provide information as to which species (human or chimpanzee, assuming chimpanzee is the simian source of HIV-1) was infected at the time of the M group expansion. If the most recent common ancestor of group M resided in a human host, then a chimpanzee-to-human transmission event likely occurred earlier than our estimates of the origin of the M group [1931 (95% CI 1915–41) was our best estimate], although it remains unknown whether months, years, or decades passed between the cross-species transmission event and the ancestral virus of the M group pandemic. If, on the other hand,

**Table 1.** Timing estimates based on an assumption of a strict molecular clock. The dates predicted with the consensus as outgroup are noted on the left, followed by their 95% CI and the log-likelihood of the data given the model. The values based on trees with the SIVcpz sequence CPZ.US as the outgroup, after having fixed the topology of the tree with the consensus as an outgroup, are shown on the right. These are not tree log-likelihoods. The log-likelihood values are an indication of how well the data fit the hypothesis of a linear rate of evolution with the model illustrated in Fig. 3. For gag and envelope gp160 data, the M group node, the B subtype node including all sequences from the United States, and the B and D clade ancestral node are given. For the trees based on fragments of envelope, the dates with historical validation and the M group node are shown. In all cases, the estimates were consistent within the boundaries of their CIs, but the most data went into the gp160 sequence analysis, so we consider this the best analysis. Shown are values based on the full gp160 sequence analysis, values based on the full-length gag gene analysis, and values based on the partial envelope alignment including the E subtype sequences. The M group origin was calculated on the basis of the inclusion of data derived from all of the sequences. The Thai E subtype origin was calculated two different ways. See the footnotes. Also shown are values based on the partial env alignment including the 1959 sequence. The 1959 sequence was included in the estimation of the time of origin of the M group.

Consensus outgroup	Log likelihood	CPZ.US outgroup	Log likelihood
<i>The gp160 analysis</i>			
M group 1931 (1915–41)	–651.6	1912 (1877–34)	–674.4
B/D 1950 (1938–59)		1950 (1939–58)	
USA B 1967 (1960–71)		1967 (1960–72)	
<i>The gag analysis</i>			
M group 1934 (1869–50)	–276.4	1933 (1863–50)	–282.8
B/D 1952 (1899–66)		1953 (1882–67)	
USA B 1972 (1946–78)		1972 (1947–78)	
<i>The analysis based on the env fragment including the E clade</i>			
M group 1940 (1916–51)	–747.3	1944 (1921–54)	–752.2
Thai E* 1986 (1978–89)		1986 (1979–89)	
Thai E† 1984 (1980–86)		1987 (1986–89)	
<i>The analysis based on the env fragment including the 1959 sequence</i>			
M group 1930 (1913–44)	–521.1	1941 (1925–53)	–540.8
1959‡ 1957 (1934–62)		1960 (1943–67)	

\*These results were based on just including E subtype sequences and projecting back to their common ancestor. †The results were based on a projection with the full M group data set, and the branch length between the M group root and the Thai E ancestral node was used to estimate the dates. ‡The estimates of the timing of the 1959 sequence origin were based on treating the 1959 time point as an unknown: The best fit was determined based on all other M group data points, excluding the 1959 sequence, and then the time of origin of the 1959 sequence was estimated based on its branch length relative to the M group root.

**Table 2.** Timing estimates when a variable rate of evolution was allowed. The estimates of the timing of the M group origin were consistent over a wide range of priors. The results are presented as the estimated year and the 95% CI. The timing of the root of the tree requires a prior, and if we changed this from a prior of 1930  $\pm$  35 years, to 1910 or 1950  $\pm$  35 years (shifting our prior by 20 years), our posterior distribution only shifted by a year or two. The prior on  $\nu$  influences the extent of rate variability between parent and child branches; increasing or decreasing  $\nu$  by a factor of 3 also had little effect on the outcome.

Predicted M group	U.S. B clade node	M group prior	$\nu$ prior
<i>Env gp160 tree</i>			
1930 (1911–45)	1954 (1941–63)	1930 $\pm$ 35	0.03
1930 (1911–44)	1955 (1942–64)	1930 $\pm$ 35	0.1
1929 (1907–43)	1954 (1940–63)	1930 $\pm$ 35	0.008
1928 (1906–43)	1954 (1940–63)	1910 $\pm$ 35	0.03
1931 (1911–45)	1955 (1942–64)	1950 $\pm$ 35	0.03
<i>Gag tree</i>			
1927 (1888–51)	1959 (1937–71)	1930 $\pm$ 35	0.03
1929 (1889–52)	1959 (1938–72)	1930 $\pm$ 35	0.1
1922 (1989–52)	1956 (1933–70)	1930 $\pm$ 35	0.008
1922 (1880–48)	1956 (1933–70)	1910 $\pm$ 35	0.03
1929 (1890–52)	1960 (1939–72)	1950 $\pm$ 35	0.03



the last common ancestor of M group strains resided in a chimpanzee, then there must have been multiple cross-species transmissions subsequent to this date, each creating the ancestor of a different clade of HIV. So, strictly speaking, our estimate is neither an upper nor a lower bound on the date of the actual zoonosis. Rather, it is the approximate time of the bottleneck event that was the genesis of the M group and captures the moment of the beginning of the expansion of the M group. If the M group originated in humans, then this would date the founder virus of the pandemic.

If all of the M group subtypes originated in different chimpanzees and entered the human population independently, they all would have had to enter the human population during a short time span of several decades, and all have been capable of spreading in the human host as epidemic strains; this seems to us improbable, but it cannot be ruled out. If, however, the M group subtypes originated in humans, the founder viruses for the subtypes must have seeded distinct epidemics, which may be limited by geography or risk groups. Conditions resulting in splitting off and expansion of genetically related viruses have been recently observed, for example, in Thailand (45) and Kaliningrad (52), demonstrating that such events do occur in human populations. Furthermore, HIV-1 subtype distributions often remain geographically distinct in African populations to this day, demonstrating that epidemics can be bound by geographic as well as risk group considerations (10, 22).

Recently, the possibility that HIV-1 was introduced into the human population iatrogenically through SIVcpz contamination of oral polio vaccines (OPV) used in the vaccination programs that took place in Central Africa between 1957 and 1960 has been revisited (49). The basis for this hypothesis is the belief that polio vaccine stocks may have been grown in kidney cells cultures derived from chimpanzees kept at a research facility close to Stanleyville (present-day Kisangani). Our analyses suggest that the HIV-1 M group ancestral sequence occurred decades before the vaccination programs and that the diverse subtypes were well established by 1957. Thus, for the OPV hypothesis to be consistent with our analyses, at least nine genetically distinct viruses would have had to enter the human population through the vaccine. As SIVcpz infections are rare in captive animals [current data derived from hundreds of animals suggest a prevalence of roughly 1% (2)], and a small number of primate kidneys were needed for OPV cultures (49), this seems implausible. Furthermore, the young age of the captured animals used for research in the facility (53) is inconsistent with the alternative hypothesis that the HIV-1 subtypes re-

sulted from the introduction of diverse quasi-species infecting the kidney of a single animal, because the distance between the subtype progenitors exceeds typical intrahost variation, particularly in infants and juveniles. Given these considerations, we find it unlikely that OPV was the source of HIV-1 transmission to humans.

Two hurdles must be crossed for cross-species transmissions to successfully establish infections capable of spreading widely in a new host: (i) the initial infection and (ii) facile transmission in the new host. If the HIV-1 M group virus did enter the human population and then diversify, then for a 25- to 60-year period before the onset of the pandemic and the first retrospective clinical documentation of AIDS in Africa in the 1970s, the virus must have gone undetected. Given the difficulty in retrospectively diagnosing the myriad of symptoms that make up AIDS, particularly in a rural African setting (54), it is feasible that small numbers of HIV-1 infections could have gone unrecognized. Lack of evidence for early HIV-1 infections is, in this context, uninformative. Historical and epidemiological considerations have been recently summarized by Chitnis *et al.* (55), leading to the suggestion that colonial practices in French Equatorial Africa provided novel opportunity for both zoonosis and subsequent expansion in the human population. Our phylogeny-based dating estimates are in good accord with the dates of the events documented in their paper.

# References and Notes

1. B. H. Hahn *et al.*, *Science* **287**, 607 (2000).
2. M. Peeters *et al.*, *AIDS* **10**, 625 (1989); M. Peeters *et al.*, *AIDS* **6**, 447 (1992); F. Gao *et al.*, *Nature* **397**, 436 (1999); S. Corbet *et al.*, *J. Virol.* **74**, 529 (2000). S. Souquiere *et al.*, paper presented at the 7th Conference on Retroviruses and Opportunistic Infections, San Francisco, 2000 ([www.retroconference.org/](http://www.retroconference.org/)).
3. V. Hirsch *et al.*, *Nature* **339**, 389 (1989); F. Gao *et al.*, *Nature* **358**, 495 (1992); F. Gao *et al.*, *J. Virol.* **68**, 7433 (1994); Z. Chen *et al.*, *J. Virol.* **70**, 3617 (1996); A. Chen *et al.*, *J. Virol.* **71**, 3953 (1997).
4. R. De Leys *et al.*, *J. Virol.* **64**, 1207 (1990); P. Charneau *et al.*, *Virology* **205**, 247 (1994); L. Gurtler, *Lancet* **348**, 176 (1996).
5. F. Simon *et al.*, *Nature Med.* **4**, 1032 (1998).
6. D. L. Robertson *et al.*, in *Human Retroviruses and AIDS 1999*, C. Kuiken *et al.*, Eds. (Los Alamos National Laboratory, Los Alamos, NM, in press) (available at [hiv-web.lanl.gov/](http://hiv-web.lanl.gov/)); D. L. Robertson *et al.*, *Science* **288**, 55 (2000).
7. A. Meyer *et al.*, *Med. Trop.* **51**, 53 (1991); D. Heymann, M. Szczeniowski, K. Esteves, *Br. Med. Bull.* **54**, 693 (1998); J. Cohen, *Science* **277**, 312 (1997).
8. A. Voevodin *et al.*, *Virology* **238**, 212 (1997); J. P. Slattery, G. Franchini, A. Gessain, *Genome Res.* **9**, 525 (1999).
9. W. Heneine *et al.*, *Nature Med.* **4**, 403 (1998); M. Callahan *et al.*, *J. Virol.* **73**, 9619 (1999).
10. Report from the Joint United Nations Programme on HIV/AIDS Global HIV/AIDS epidemic update 1999, available at [www.unaids.org/publications/](http://www.unaids.org/publications/).
11. T. Zhu *et al.*, *Nature* **391**, 594 (1998).
12. T. Jonassen *et al.*, *Virology* **231**, 43 (1997).
13. F. Barré-Sinoussi *et al.*, *Science* **220**, 868 (1983); R. C. Gallo *et al.*, *Science* **220**, 865 (1983).
14. T. Leitner, D. Escanilla, C. Franzen, M. Uhlen, J. Albert, *Proc. Natl. Acad. Sci. U.S.A.* **93**, 10864 (1996); T. Leitner, S. Kumar, J. Albert, *J. Virol.* **71**, 4761 (1997); T. Leitner and J. Albert, *Proc. Natl. Acad. Sci. U.S.A.* **96**, 10752 (1999).
15. P. M. Sharp, *Nature* **336**, 315 (1988); T. F. Smith, A. Srinivasan, G. Schochetman, M. Marcus, G. Myers, *Nature* **333**, 573 (1988); P. Kasper *et al.*, *AIDS Res. Hum. Retroviruses* **11**, 1197 (1995).
16. B. Korber, J. Theiler, S. Wolinsky, *Science* **280**, 1868 (1998).
17. J. Felsenstein, *J. Mol. Evol.* **17**, 368 (1981); *Cladistics* **5**, 164 (1989); G. J. Olsen, H. Matsuda, R. Hagstrom, R. Overbeek, *Comput. Appl. Biosci.* **10**, 41 (1994).
18. FastDNAm1 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/](http://rdp.life.uiuc.edu/)).
19. D. L. Swofford, G. J. Olsen, P. J. Waddell, D. M. Hillis, in *Molecular Systematics*, D. M. Hillis, C. Moritz, B. K. Mable, Eds. (Sinauer, Sunderland, MA, 1996), pp. 407–514.
20. D. M. Hillis, B. K. Mable, C. Moritz, in *Molecular Systematics*, 2nd ed., D. M. Hillis, C. Moritz, B. K. Mable, Eds. (Sinauer, Sunderland, MA, 1996), pp. 515–543.
21. J. L. Thorne, H. Kishino, I. S. Painter, *Mol. Biol. Evol.* **15**, 1647 (1998).
22. The alignment was based on those provided in *Human Retroviruses and AIDS*, B. Korber *et al.*, Eds. (Los Alamos National Laboratory, Los Alamos, NM, 1998).
23. A complete description of the alignments, details of the phylogenetic analysis, the sequence alignments, and the links new code written for this study are provided at [www.santafe.edu/btk/science-paper/bette.html](http://www.santafe.edu/btk/science-paper/bette.html).
24. R. Shankarappa *et al.*, *J. Virol.* **73**, 10489 (1999).
25. S. M. Wolinsky *et al.*, *Science* **272**, 537 (1996); S. Ganeshan *et al.*, S. Wolinsky, *J. Virol.* **71**, 663 (1997); R. B. Markham *et al.*, *Proc. Natl. Acad. Sci. U.S.A.* **95**, 12568 (1998); P. Bagnarelli *et al.*, *J. Virol.* **73**, 3764 (1999).
26. M. R. Furtado *et al.*, *J. Virol.* **69**, 2092 (1995); A. S. Perelson *et al.*, *Science* **271**, 1582 (1996); D. Finzi *et al.*, *Science* **278**, 1295 (1997); D. Finzi *et al.*, *Nature Med.* **5**, 512 (1999); H. Gunthard *et al.*, *J. Virol.* **73**, 9404 (1999); L. Zhang *et al.*, *N. Engl. J. Med.* **340**, 1605 (1999); M. R. Furtado, *N. Engl. J. Med.* **340**, 1614 (1999).
27. B. Efron and R. Tibshirani, *Science* **253**, 390 (1991).
28. N. Grassly, P. Harvey, E. Holmes, *Genetics* **151**, 427 (1999).
29. D. L. Robertson, B. H. Hahn, P. M. Sharp, *Mol. Evol.* **40**, 249 (1995); L. Heyndrickx *et al.*, *J. Virol.* **74**, 363 (2000).
30. C. L. Kuiken *et al.*, *AIDS* **10**, 31 (1996).
31. M. Salemi *et al.*, *Proc. Natl. Acad. Sci. U.S.A.* **96**, 13253 (1999); P. Smith and D. Simmons, *J. Virol.* **73**, 5787 (1999); P. Pollyky and E. Holmes, *Mol. Evol.* **49**, 130 (1999).
32. Z. Yang, *J. Mol. Evol.* **39**, 105 (1994); *J. Mol. Evol.* **39**, 306 (1994); *J. Mol. Evol.* **42**, 587 (1996).
33. D. M. Hillis, J. P. Huelsenbeck, C. W. Cunningham, *Science* **264**, 671 (1994); Z. Yang, N. Goldman, A. Friday, *Mol. Biol. Evol.* **11**, 316 (1994); J. Huelsenbeck, *Mol. Biol. Evol.* **12**, 843 (1995); M. Kuhner and J. Felsenstein, *Mol. Biol. Evol.* **11**, 459 (1994).
34. Z. Yang, *Genetics* **139**, 993 (1995).
35. J. P. Huelsenbeck and B. Rannala, *Science* **276**, 227 (1997).
36. Because it is not possible to test all possible tree configurations, tree-building programs use heuristics to estimate the best tree, and the final tree is dependent on the input order of sequences. To optimize the final trees, we randomized the input order of the sequences five to seven times, until the best maximum-likelihood scores were very similar (7). Given the number of taxa we included, and consequently the combinatorially vast potential for different branching orders, we do not expect our trees to be optimal solutions. Limited testing of the final timing estimates based on different input orders of sequences did not significantly affect our calculations of the timing of divergence from a common ancestor. We also compared the likelihood of the data under dif-

- ferent evolutionary models (1), and over 100 maximum-likelihood trees were run in the course of this study.
37. Z. Yang, *Mol. Biol. Evol.* **10**, 1396 (1993).
  38. D. L. Swofford, PAUP\*. Phylogenetic Analysis Using Parsimony (\*and Other Methods) (Sinauer, Sunderland, MA, 1999).
  39. We also tested other aspects of our evolutionary model. We found that the assignment of base frequencies by means of the phylogenetic trees gave consistently better results than empirical base frequencies. The REV model performed better than an F84 model (2), which only includes rate parameters for transitions and transversions instead of for each pair of bases. Also, for the envelope gene analyses, the improvement in the log-likelihood scores comparing the REV model with a uniform rate of evolution at all sites, to the REV model with rate variation at different sites estimated by the maximum-likelihood method, was many times larger than the number of positions (1), justifying the increase in parameters (3).
  40. B. Korber, P. Sharp, D. Ho, *Nature* **400**, 326 (1999); J. Goudsmit and V. Lukoshov, *Nature* **400**, 325 (1999).
  41. B. Korber et al., data not shown.
  42. V. Courgnaud et al., *Virology* **247**, 41 (1998).
  43. Y. Wangroongsarb et al., *Southeast Asian J. Trop. Med. Public Health* **16**, 517 (1985); D. Smith, *Lancet* **335**, 781 (1990); C. Mason et al., *J. Acquir. Immune Defic. Syndr. Hum. Retrovir.* **19**, 165 (1996); R. Bunnell et al., *AIDS* **13**, 509 (1999).
  44. C. Kuiken et al., *Am. J. Epidemiol.*, in press.
  45. F. McCutchan et al., *J. Virol.* **70**, 3331 (1996); S. Subbarao et al., *AIDS Res. Human Retroviruses* **14**, 319 (1998).
  46. F. Gao et al., *J. Virol.* **70**, 7013 (1996); J. K. Carr et al., *J. Virol.* **70**, 5935 (1996).
  47. M. S. Gottlieb et al., *N. Engl. J. Med.* **305**, 1425 (1981); M. S. Gottlieb et al., *Morb. Mortal. Wkly. Rep.* **30**, 250 (1981).
  48. R. Selik, H. Haverkos, J. Curran, *Am. J. Med.* **76**, 493 (1984); J. Pape, *N. Engl. J. Med.* **309**, 945 (1983); L. Gazzolo, *N. Engl. J. Med.* **311**, 1252 (1984).
  49. E. Hooper, *The River* (Little, Brown, Boston, 1999). See pp. 77–82 and 440–443 for discussion of early cases in the United States and Haiti, and pp. 550, 791, and 1009 for a discussion of the number of primate kidneys required to make OPV.
  50. S. Chevret et al., *J. Epidemiol. Commun. Health* **46**, 582 (1992).
  51. W.-H. Li, M. Tanimura, P. Sharp, *Mol. Biol. Evol.* **5**, 313 (1988); T. Gojobori et al., *Proc. Natl. Acad. Sci. U.S.A.* **340**, 1605, 4108 (1990); J. Kelly, *Genet. Res.* **64**, 1, 1994.
  52. K. Liitsola et al., *AIDS* **12**, 1907 (1998).
  53. A record of the ages of chimpanzees from Camp Lindi
- used for research noted a range from <1 to 10 years, with more than 80% less than 4 years old (S. Plotkin, personal communication; data taken from the laboratory notes of F. Deinhardt).
54. M. Grmek, *History of AIDS Emergence and Origin of a Modern Pandemic* (Princeton Univ. Press, Princeton, NJ, 1990), chaps. 10 and 15.
  55. A. Chitnis, D. Rawls, J. Moore, *AIDS Res. Hum. Retroviruses* **16**, 5 (2000).
  56. We thank D. Pollock, T. Leitner, and B. Bruno for suggestions concerning phylogenetics, maximum likelihood, and estimating the error on time of sampling; G. Shaw for suggesting the 1959 control; S. Wain-Hobson and G. Myers for clarifying discussions on the interpretation and limitations of these results; B. Foley and C. Kuiken for numerous helpful discussions; and K. Rock and J. Shepard for technical support. G. Olsen and J. Thorne generously supplied source code and helped us interpret their work. The research of the Los Alamos authors was supported under internal funds from the Delphi Project, S.W. and B.K. were supported by NIH (RO1-HD37356), B.K. and M.M. were supported through the Pediatric AIDS Foundation, and an anonymous foundation supplied further support for S.W. B.H.H. was supported by grants NO1 AI 85338, RO1 AI 44596, and RO1 AI 40951 from NIH.

16 December 1999; accepted 28 April 2000

# Kinesin Superfamily Motor Protein KIF17 and mLin-10 in NMDA Receptor-Containing Vesicle Transport

Mitsutoshi Setou, Terunaga Nakagawa, Dae-Hyun Seog, Nobutaka Hirokawa\*

Experiments with vesicles containing *N*-methyl-D-aspartate (NMDA) receptor 2B (NR2B subunit) show that they are transported along microtubules by KIF17, a neuron-specific molecular motor in neuronal dendrites. Selective transport is accomplished by direct interaction of the KIF17 tail with a PDZ domain of mLin-10 (Mint1/X11), which is a constituent of a large protein complex including mLin-2 (CASK), mLin-7 (MAL5/Velis), and the NR2B subunit. This interaction, specific for a neurotransmitter receptor critically important for plasticity in the postsynaptic terminal, may be a regulatory point for synaptic plasticity and neuronal morphogenesis.

In mammalian neurons, neurotransmitter receptors such as glutamate receptors, including NMDA receptors, are sorted dynamically and precisely to the dendrites of the cell (1). Although putative anchoring, sorting, and signaling molecules have been colocalized with the receptors (2), it is not yet known how the receptors are sorted. Transport of molecules to specific regions in eukaryotic cells is accomplished by molecular motors (3). In neurons, various microtubule-associated motor proteins have been shown to transport

organelles such as synaptic vesicle precursors and mitochondria to specific regions of the cell; however, the mechanisms by which each motor recognizes its specific cargo are not known (3). Here, we report that KIF17 (4), a neuron-specific microtubule-dependent molecular motor, binds directly and specifically to a PDZ domain (5) of mLin-10 (6) and transports the large protein complex containing the NR2B subunit, which forms the NMDA receptor with the NR1 subunit (7). This complex transports vesicles along microtubules in neurons such as hippocampal pyramidal neurons.

**Identification of KIF17.** Members of the kinesin superfamily (KIFs) support diverse transport systems in cells (3). We cloned KIF17, a neuron-specific motor (Fig. 1A) (8),

to investigate the motors responsible for the sorting of various molecules within neurons. Osm-3 (9), a putative dendritic motor for odorant receptors in *Caenorhabditis elegans*, and KIF17 constitute a family (Fig. 1B). KIF17 is similar to Osm-3 in the head and tail domains (Fig. 1C) and has two putative stalk domains that form an  $\alpha$ -helical coiled coil (Fig. 1D) (10); Osm-3, however, has only one stalk domain. Antibody raised against amino acids 505 to 707 of KIF17 (anti-KIF17) (11) recognized the native KIF17 protein in the brain as a single 170-kD band in SDS-polyacrylamide gel electrophoresis (PAGE) (Fig. 1E).

Although KIFs could be monomeric, homodimeric, heterotrimeric, homotetrameric, or heterotetrameric (3), native KIF17 has a sedimentation coefficient of 3.0 S and a Stokes radius of 170 Å, which suggests that the molecular weight of the native holoenzyme is 215 kD, about twice that calculated from the sequence of the protein (116 kD). The migration of KIF17 in native PAGE was similar to that of the recombinant full-length KIF17 protein (12). Thus, KIF17 probably exists as a homodimer (13).

To measure the direction and velocity of the KIF17 motor activity, we assayed its motility with the use of recombinant KIF17. Recombinant KIF17 could slide an axoneme toward microtubule minus-ends, indicating that KIF17 is a microtubule plus-end-directed motor (Fig. 1F). The microtubule gliding assay showed that the average speed was 0.8 to 1.2  $\mu\text{m/s}$  (14). Thus, KIF17 can act without a coenzyme and can mediate fast intracellular transport.

**KIF17 is a dendrite-specific motor protein.** KIF17 appeared to be brain-specific, present in abundance in the gray matter (es-

Department of Cell Biology and Anatomy, Graduate School of Medicine, University of Tokyo, Bunkyo-ku, Tokyo, Japan.

\*To whom correspondence should be addressed. E-mail: hirokawa@m.u-tokyo.ac.jp