Dating the common ancestor of SIVcpz and HIV-1 group M and the origin of HIV-1 subtypes by using a new method to uncover clock-like molecular evolution

Marco Salemi,* Korbinian Strimmer, [†] William W. Hall, [‡] Margaret Duffy, [‡] Eric Delaporte, [§] Souleymane Mboup, [‖] Martine Peeters, [§] and Anne-Mieke Vandamme*.

*Rega Institute for Medical Research, Katholieke Universiteit Leuven, B-3000 Leuven, Belgium; [†]MIPS/GSF, Max-Planck-Institut für Biochemie, D-82152 Martinsried, Germany; [‡]Department of Medical Microbiology, Conway Institute of Biomolecular and Biomedical Research, University College Dublin, Dublin 4, Ireland; [§]Laboratoire Retrovirus, IRD, Montpellier, France; and [‖]African Network of HIV variability, Senegal

Corresponding author: Anne-Mieke Vandamme, Rega Institute for Medical Research, Minderbroedersstraat 10, B-3000 Leuven, Belgium. E-mail: annemie.vandamme@uz.kuleuven.ac.be

ABSTRACT

Attempts to estimate the time of origin of human immunodeficiency virus (HIV)-1 by using phylogenetic analysis are seriously flawed because of the unequal evolutionary rates among different viral lineages. Here, we report a new method of molecular clock analysis, called Site Stripping for Clock Detection (SSCD), which allows selection of nucleotide sites evolving at an equal rate in different lineages. The method was validated on a dataset of patients all infected with hepatitis C virus in 1977 by the same donor, and it was able to date exactly the 'known' origin of the infection. Using the same method, we calculated that the origin of HIV-1 group M radiation was in the 1930s. In addition, we show that the coalescence time of the simian ancestor of HIV-1 group M and its closest related cpz strains occurred around the end of the XVII century, a date that could be considered the upper limit to the time of simian-to-human transmission of HIV-1 group M. The results show also that SSCD is an easy-to-use method of general applicability in molecular evolution to calibrate clock-like phylogenetic trees.

Key words: interspecies transmission • molecular clock

Human immunodeficiency virus (HIV)-1 shows a high-sequence variability: The dominant group in the pandemic, group M, exists as different subtypes designated A to K (1–3). The virus was most probably introduced into the human population via zoonotic transmission from simian immunodeficiency virus (SIV)-infected *Pan troglodytes troglodytes* (SIVcpz) (4). However, it remains controversial as to how long HIV-1 has been circulating in humans. The finding of an HIV-1 sequence in an African plasma sample dating from 1959, which clustered near the ancestral node of subtypes B and D in a phylogenetic tree, would certainly imply that the virus was introduced into the African population before this time (5).

In principle, the time of origin of the most recent common ancestor for a clade of contemporary virus strains can be estimated from a phylogenetic tree provided the molecular clock hypothesis holds. In such a tree, the degree of sequence divergence, as measured by the number of substitutions along a branch, is linearly proportional to the time of divergence. With the evolutionary rate known, the clock-like branch lengths can be divided by the rate to calculate divergence times.

Considering the sequence divergence of the different subtypes and an average rate of non-synonymous substitutions—around 10^{-3} substitutions per site per year (6)—the origin of group M has been roughly estimated to be some time in the past 50 years (6, 7), whereas the time of the common ancestor of HIV-1 and SIVcpz remains even more uncertain (7). The limitations of these estimates lie in the non-constant evolutionary rate of the different HIV-1 lineages. The application of molecular clock analysis to HIV-1 phylogeny is unreliable, as different selective pressures can lead to dissimilar rates of evolution of the virus in different individuals (8). A further limitation is the different evolutionary rates of distinct HIV-1 subtypes (8, 9). To overcome these problems, Korber et al. (2000) recently developed a computational-intensive approach to calculate the origin of HIV-1 group M by taking into account evolutionary rate variation among the different viral lineages. They concluded that a common ancestor of group M existed around 1930 (10).

In the present paper we discuss a novel method of molecular clock analysis that allows, in a set of aligned sequences, the removal of nucleotide sites that distort the molecular clock. The method is easy to implement and is based on a different approach than the one developed by Korber and colleagues. Employing this method, we calculated the time of origin of HIV-1 group M radiation and the time of separation between the lineage leading to the common ancestor of HIV-1 group M and the one leading to the currently closest simian relative SIVcpz from *Pan troglodytes troglodytes*.

MATERIALS AND METHODS

Sequence data

Nineteen HCV sequences (360 sites in the E1/E2 region) sampled in 1998 from patients who were infected in 1977 by the same donor through an HCV-genotype-1b-contamined anti-D immunoglobulin preparation were aligned by using the program Clustal W (11). A second alignment was obtained, also including HCV consensus sequences from three other patients infected in 1977 by the same immunoglobulin batch and sampled in 1994 (12).

A thorough search of the HIV database (13) allowed us to select the HIV-1 sequences with the necessary characteristics for carrying out the analysis. We looked for sequences sampled within the same year, which were the longest possible, known as non-recombinant, and representative of the major HIV-1 subtypes. Subtype E is a circulating recombinant form and was therefore excluded. The best dataset resulted in 20 full-length *env* sequences sampled in 1992; they belong to the non-recombinant subtype A, B, C, D, and G of HIV-1 group M. We also retrieved all the available HIV-1 subtype B full-length *env* sequences with known sampling year and used these to estimate the evolutionary rate. In total, 59 HIV-1 entire envelope sequences with known sampling year between 1983 and 1996 were retrieved from the HIV database and SIVcpzGAB,

For estimating the HIV-1 evolutionary rate in *pol* and *env*, several sequences isolated at different time points between 1983 and 1998 are available. Branch lengths connecting subtype B strains to their common ancestor were plotted against their sampling year and a weighted linear regression (the loss function was weighted with the inverse of the standard error of the branch lengths estimated via maximum likelihood) was carried out with the STATISTICA software package (21), in order to find the slope of the curve corresponding to the rate (r).

Divergence times for an ancestral node of a clock-like tree are computed by dividing the branch length h from the tip to the node by the evolutionary rate r. The standard error of the estimated divergence times is obtained by: $(h/r) = \sqrt{1/r^2} \ (se_r^2 \ h^2/r^2 + se_h^2)$, where se_r and se_h represent the standard error of r and h, respectively (21).

RESULTS

Uncovering clock-like evolving sites

Several factors, such as homoplasy and recombination, can be responsible for the distortion of the molecular clock during evolution (22). The effect of homoplasy is usually handled by using models of nucleotide substitutions that include transition transversion bias and rate heterogeneity across sites. In our analysis, we used an evolutionary model by considering these factors (see Materials and Methods) and we also excluded from the analysis strains known to be recombinant. Another factor leading to non-constant rates of evolution among different lineages is that the selective constraints in a genomic region can change (23). However, not all the sites of a given sequence are subjected to the same selective forces; for example, synonymous substitutions, which are less subject to purifying selection, occur much faster than non-synonymous ones, even though the latter can speed up in the presence of positive selection (22). Thus, a good approximation of the clock-like behavior could theoretically be achieved by removing from a set of aligned sequences those sites that contribute more to the distortion of the molecular clock.

To extract clock-like information from a dataset and to determine clock-like branch lengths for the phylogeny underlying that dataset, the following iterative procedure, which has been termed Site Stripping for Clock Detection (SSCD), was carried out. First, we assume that the tree representing the true phylogenetic relationships among the taxa under investigation is known. Relative rates are then assigned to each site of the alignment by a maximum *a posteriori* estimate (24), where 16 categories of rates from the slowest (category 1: invariable sites) to the fastest (category 16: fastest evolving sites) are allowed. Maximum-likelihood branch lengths are then computed separately for each codon position with and without the clock-constraint on the basis of the full alignment. The corresponding likelihood values for the clock and the non-clock tree are compared by using a likelihood ratio test (25). When the clock constraint reduces the likelihood of the tree significantly, the fastest sites are removed progressively (stripped) from the alignment, and the clock-test is repeated until a clock-like behavior is obtained. It is important to point out that this strategy does not imply that the rate of heterogeneity across sites is responsible for the distortion of the molecular clock and that only datasets with no rate heterogeneity across sites would behave in a clock-like manner (which is, in fact, not true). The fastest category is removed because, as we will show below, the sites belonging to the fastest category seem to be

SIVcpzUS, and HIV-O sequences were included as outgroups (2–4). Sequences were aligned with Clustal W, and all the gaps were removed from the final alignment.

To obtain a more recent dataset and to investigate a different genomic region, we used 23 recently sequenced strains, corresponding to nt 2358-3809 of the reference strain HXB2 (13), sampled in 1998 and belonging to the non-recombinant subtype A, B, C, D, G, J, and K of HIV-1 group M (14). All the available subtype B sequences in the same region, with known sampling year between 1983 and 1996, were also retrieved from the HIV database and used to estimate an evolutionary rate. In total an alignment of 45 HIV/SIV sequences was obtained.

All the alignments used in this study are available through the World-Wide-Web at http://kuleuven.ac.be/aidslab/hivclockdata.htm

Phylogenetic analysis

Phylogenetic trees for the dataset, including HIV-1 *env* strains isolated in 1992 and the HIV-1 *pol* strains isolated in 1998, were obtained by the distance methods (neighbor-joining, weighted least-squares) implemented in NEIGHBOR and FITCH of the PHYLIP 3.572 software package (15), and by maximum-likelihood methods as implemented in PUZZLE 4.0.2 (16) and DNAML in PHYLIP 3.572. For both datasets SIVcpz strains and HIV-1 group O strains were used as outgroups. Consensus trees from different reconstruction algorithms were also obtained for all the 62 HIV/SIV *env* strains, including sequences with different sampling year, and all the 45 HIV/SIV *pol* strains. In all phylogenetic analyses the distances were calculated with PUZZLE assuming the Tamura and Nei model with Γ-distributed rates across sites (16 rate categories). In each analysis, the parameters of the model, including the α-shape parameter of the Γ-distribution, were estimated by PUZZLE itself with the maximum likelihood method. The clock-like branch lengths for the *pol* phylogenetic tree employing 1st+3rd codon position was obtained with a model of evolution that used a different transition transversion ratio and α for the 1st and the 3rd codon position separately. All the parameters of this model (including the branch lengths) were estimated with the program PAML (17). Bootstrap re-sampling (1,000 bootstrap replicates) was applied to the neighbor-joining and the weighted least-squares trees. Likely ratio clock-tests and calculating assignment of rates to sites was done by using PUZZLE. The SSCD procedure was performed by using PAL version 0.5 (http://members.tripod.de/korbi/pal).

Substitution saturation plots, shown in Figure 4, were obtained by using the program DAMBE (18).

Evolutionary rates and divergence times

When the molecular clock holds, the evolutionary rate of a fast-evolving virus, like HIV or hepatitis C virus (HCV), can be estimated by comparing strains with different sampling years (6, 19, 20). For the HCV dataset, it was possible to use sequences isolated in 1994 and 1998 all from different patients infected by the same donor in 1977. The average number of nucleotide substitutions of the 1998 strains, as well as the 1994 strains to a common outgroup, an HCV subtype 1b strain branching off their clade, was calculated. The evolutionary rate was then obtained by dividing the difference in the number of nucleotide substitutions by the difference in isolation time (6, 20).

the ones distorting the molecular clock more. Even after the removal of one or two of the fastest categories of sites, the datasets that we used in our study maintained a strong rate of heterogeneity, which was taken into account for the estimation of the branch lengths of the phylogenetic trees.

To investigate the molecular clock hypothesis on a tree representing the true phylogenetic relationships among viral strains with SSCD and, subsequently, to infer dates for ancestral nodes of that tree, we need a set of aligned sequences satisfying the following criteria. The sampling year of the viral strains has to be known, and the strains must have been sampled within the same year, possibly no more than a few months apart. Strains that have been shown to be recombinants must be excluded, as the inclusion of a recombinant strain would obscure the phylogenetic relationships. The longest sequence possible should be used to ensure a reliable test of the molecular clock. Finally, we need an estimation of the evolutionary rate (nucleotide substitution per site per year) for the specific sites of the alignment remaining after the SSCD procedure.

A hepatitis C virus dataset with a common ancestor in 1977

To test the reliability of the SSCD procedure, we used a set of 19 HCV strains sampled in 1998 from patients who were infected within a couple of months during 1977. The patients, all women, received an anti-D immunoglobulin preparation from a batch contaminated with HCV-genotype-1b from a single blood donation (11). The HCV sequence from this batch was the common ancestor of the viral quasi-species that was eventually isolated and characterized from them a couple of decades later. The sequence heterogeneity of HCV in the original contaminated batch was investigated, and the virus proved to be completely homogeneous (13). Thus, the topology of the tree shown in Figure 1 represents the true phylogenetic relationships among the 19 HCV strains, and 1977 is the year of their common ancestor. Because the sequences were only 360-nt long (11), we decided not to analyze 1st, 2nd, and 3rd codon positions separately to avoid a further reduction of the sequence length and therefore of the phylogenetic signal.

When clock-like and non-clock-like branch lengths were estimated for the tree in Figure 1, the clock hypothesis had to be rejected (Table 1). However, after removing from the alignment the sites assigned the two-or-more fastest categories, the molecular clock cannot be rejected (Table 1). Thus, the clock-like branch length connecting the HCV strains sampled in 1998 with their common ancestor calculated with the 'stripped' alignment (see Table 1) can be used to date the time of origin of the common ancestor. The evolutionary rate for the 'stripped' alignment was calibrated with the outgroup method as described in Methods. Table 1 shows that, after the SSCD procedure and employing the calculated rate, the correct date of origin of the common ancestor of the HCV infection, 1977, was inferred by using the alignment from which the minimum number of sites was removed to obtain a molecular clock. When removing further categories, the number of variable sites was reduced to such an extent that dates became less reliable with a larger standard error (see Table 1). Without SSCD, molecular clock dating on the non-clock dataset gave wrong estimates. It is worth noting that the virus in the original HCV-contaminated batch is completely homogeneous (13). Thus, the date of the infection found with our method is not a result of an overestimation of the branch lengths.

Because the ancestral sequence of the HCV donor is known (13), we also estimated the HCV evolutionary rates, for stripped and unstripped datasets, by comparing the strains sampled in

1998 and 1994 from different patients with the 1977 sequence from the donor instead of with an outgroup sequence. The same sites had to be stripped to obtain a molecular clock, but the obtained evolutionary rates and the dates were slightly different (1970 ± 11, with 360 sites; 1981 ± 6, with 339 sites; 1978 ± 5, with 325 sites). The fact that the stripped alignment is the one giving the best estimate of the time of origin of the HCV strains is what confirms the validity of our approach. It also shows that larger errors are introduced when the dates are estimated from the unstripped alignment by using an outgroup to calibrate the evolutionary rate. However, this aspect had little effect on our estimation of the date from the stripped alignment, although the confidence interval is somewhat larger when using an outgroup.

The fastest-evolving sites seem to be the ones more distorting to the molecular clock because they are very likely accelerated by positive selection, which can speed up the evolutionary clock of some lineages with respect to others.

In Figure 2, the distribution of sites belonging to different categories of relative substitution rate of the HCV dataset is reported. More than 80% of the 166 sites assigned to category 1 are 1st or 2nd codon position sites. This finding does not come as a surprise, as category 1 is the slowest-evolving category, representing invariable sites (relative substitution rate = 0; see Fig. 2), which are usually subjected to strong purifying selection (22). Faster-evolving categories show a higher proportion of 3rd codon position sites. For example, within category 13 and 14, more than 60% of sites belong to 3rd codon positions. Again, this finding is in agreement with the observation that replacements at 3rd codon position, mostly synonymous, tend to occur at faster rates than at 1st or 2nd codon position, mostly non-synonymous, when positive selection does not operate (22). However, the situation is rather different for the two fastest-evolving categories (see Fig. 2). More than 75% of sites within category 15 and 16 belong to 1st or 2nd codon position. Because replacements at 1st or 2nd codon position are usually non-synonymous, this finding seems to suggest that positive selection may be, at least in part, operating. To further test this hypothesis, we obtained an alignment including only those codons containing at least one site assigned to category 15 or 16. The sequence of each one of the 19 HCV strains isolated in 1998 was compared with the sequence from the 1977 donor and the ratio of synonymous and non-synonymous substitution (dn/ds) counted with the method of Li (1993), implemented in DAMBE (18). As expected, the average dn/ds was significantly greater than 1 (3.71, p<0.0002), suggesting positive selective pressure. However, when only the codons containing the sites assigned to the remaining categories were analyzed, the average dn/ds was significantly lower than 1 (0.4, p<0.0002), suggesting purifying selection. In conclusion, the data seem to indicate positive selection acting on the sites removed, in agreement with the hypothesis that positively selected sites are more likely to distort the molecular clock.

HIV-1 phylogeny

Phylogenetic trees of HIV-1 *pol* strains sampled in 1998 and *env* strains sampled in 1992 are shown in Figure 3A and 3B, respectively. The relationship of groups of sequences was considered uncertain when different tree topologies were obtained by different reconstruction algorithms. As such, the trees in Figure 3 are not fully resolved; instead, they should be viewed as a consensus of the evolutionary history of the sequences. SIVcpz strains from *Pan troglodytes troglodytes* and very divergent HIV-1 strains belonging to group O served as outgroups. The rationale behind using multifurcating trees is that we are interested mainly in dating the common

ancestors of the clades indicated in Figure 3, for which all the phylogenetic analyses gave a robust support.

Calibrating the HIV-1 molecular clock

For each dataset, sites belonging to the first, second, and third codon positions were analyzed separately, as they are subjected to different selective pressures and thus exhibit different evolutionary patterns (22).

The clock for the second codon positions of *pol* cannot be rejected after removing the three fastest categories, whereas for the 1st and 3rd codon positions no SSCD is necessary to observe a clock-like behavior. In *env*, the two fastest categories for first codon positions, the five fastest for second codon positions and the eight fastest for third codon positions have to be removed. The failure of rejecting the molecular clock at third codon positions of *pol* may be an artifact due to saturation. More precisely, the occurrence of multiple changes at sequence sites tends to level out the distances of the investigated sequences to the root sequence of the assumed tree. In this situation, the data can appear clock-like, even if the contrary is known to be true; consequently, a molecular-clock analysis based on third codon positions is often found to be unreliable (26). Also, first and second codon positions change slower because replacements at these positions usually lead to amino acid changes that are likely to be subjected to purifying selection and, as such, saturation effects are less likely. Because during evolution the number of observed transitions relative to that of transversions gradually decreases with increasing divergence (18), a visual display of substitution saturation can be obtained by plotting the number of transitions and transversions versus divergence for each pair-wise comparison of a given dataset. Figure 4 shows the results of these substitution saturation plots for the first and the third codon position of the *pol* and *env* dataset. The first and third codon positions of *pol* show little evidence of saturation. Thus, first + third codon positions of *pol* can be reliably used (given that the molecular clock holds) to date the divergence time of the nodes indicated in the tree of Figure 3A. On the contrary, both first and third codon position of *env* become completely saturated (transitions outnumber transversions) when the SIVcpz strains are compared with the other HIV-1 group M strains (see Fig. 4). In this situation, it is still possible to obtain reliable estimates of the divergence times for the *env* tree before that saturation had occurred, but the divergence between SIVcpz and HIV-1 would result in an underestimation.

For the HIV-1 datasets, the reduction of variability at first codon positions of *env* within group M is 17% after SSCD. However, to obtain a clock at the second codon position, 235 sites have to be removed (Table 2) with a reduction of about 65% of the variability, which would suggest an insufficient amount of phylogenetic information for a reliable analysis. The situation for the second codon position of *pol* is even more dramatic, in which only 0.8% of variable sites remain within group M after site stripping.

In conclusion, *pol* first + third codon position and *env* 1st codon positions (after SSCD) were judged to be the most informative and were used to estimate clock-like branch lengths for the *pol* and *env* phylogenetic trees, respectively.

Evolutionary rates of HIV-1 *pol* and *env* were estimated by linear regressi0on analysis as described in Methods. We compared all the available subtype B strains sampled between 1983 and 1998. Subtype B strains were chosen because they represent the largest dataset of strains

with known sampling year. The correlation between sampling years and branch lengths at the first codon positions for *env* ($r^2 = 0.35$, $p<0.03$) and first + third codon position for *pol* ($r^2 = 0.42$, $p<0.03$) was statistically significant. The resulting rates appear in Table 3. Considering that after the SSCD procedure we obtain datasets for which the molecular-clock hypothesis cannot be rejected, an evolutionary rate estimated for a particular clade (subtype B strains in our case) of the stripped dataset is the same for any other lineage included in that dataset and can be used to date ancestral nodes on the corresponding clock-like phylogenetic tree.

It is interesting that most of the sites removed at the first codon position of *env* in order to get a molecular clock belong to the codons recently identified as putative targets of positive selection (27). This result, again, is in agreement with the hypothesis that positively selected sites are more likely to distort the molecular clock.

Origin of HIV-1 subtypes

Evolutionary rates and clock-like branch lengths estimated for the sites for which the clock hypothesis cannot be rejected and inferred divergence times for the main nodes of the HIV-1 phylogeny (see Fig. 3A and 3B) are given in Table 3. The origin of subtype B radiation is estimated to be in the early 1970s (1970 with *pol* and 1972 with *env*) (Table 3). The most recent common ancestor of subtype B and D dates back between the end of the 1940s and the early 1950s (1949 with *pol* and 1954 with *env*). The date of origin of group M radiation is estimated to be between the 1920s and the 1930s; 1920, according to *pol*, and 1937, according to *env* (Table 3).

After removing the fastest-evolving sites responsible for the distortion of the molecular clock, the clock cannot be rejected for the tree, including HIV-1 group M strains as well as SIVcpz strains. Even if considering all sites of the genome HIV-1 and SIV might exhibit different evolutionary rates, the sites remaining in the *pol* and *env* alignment after the SSCD procedure do accumulate mutations at constant rate over time. As we have noticed above, the first codon position of *env* becomes saturated for the deeper branches of the *env* phylogenetic tree and an estimate of the separation between the HIV-1 group M lineage and the SIVcpz strains would be unreliable. However, although first and third codon positions of *pol* show some evidence of saturation, because the curves for transitions and transversions seem to level-off, it is still possible to use them for dating the deep branches because full saturation (crossing of the two clouds) does not occur. We estimated that the most recent common ancestor of HIV-1 group M and SIVcpz (see Fig. 3) dates from around 1675 (99% c.i. 1590–1761).

DISCUSSION

The first goal of this study is to show that it is possible to apply molecular-clock dating methods to a set of aligned nucleotide sequences for which the molecular clock is rejected. This application is achievable by stripping the sites that distort the clock and retaining the clock-like sites that are subsequently used in the calculations. The idea of removing fast-evolving sites to obtain a clock-like tree has been introduced in this paper for the first time. It is a simple but powerful technique that is of general applicability. In principle, different reasons could be responsible for the unequal evolutionary rates observed in HCV, HIV, or other viral or non-viral datasets. Our results here show that as soon as we can uncover the sites for which the molecular

clock fails to be rejected, these sites can be used reliably to construct clock-like phylogenetic trees and to calculate divergence times at ancestral nodes. By identifying the sites that distort the clock, the SSCD procedure also allows us to potentially investigate the selective forces at the basis of differential evolutionary rates. In particular, we have seen that the sites removed to get a clock-like dataset seem to be subjected to positive selection, at least for the HCV dataset (see Results). In addition, when we tried to remove the slowest-evolving sites from the HIV dataset and retain the fastest sites, the molecular clock was always rejected (data not shown). This finding is not surprising, because slow-evolving sites are usually subjected to purifying selection or neutral evolution and they distort the molecular clock to a much lesser extent.

The HCV example demonstrated that if sites belonging to the fastest-evolving categories of a dataset not following the clock are removed, and a clock-like dataset is obtained for which it is possible to calibrate an evolutionary rate, the divergence time estimated is in perfect agreement with empirical facts. On the contrary, assuming a clock-like tree when the molecular-clock hypothesis has, in fact, to be rejected is what leads to erroneous results (see Table 1). Two facts are also worth noting. First, the standard errors are quite large because of the short sequences and low signal content, even if the date of origin of the HCV dataset is inferred correctly when sites are removed. Second, removing variable sites also means loss of phylogenetic signal. For example in the HCV dataset, 54.2% of the 360 sites were variable; whereas in the clock-like dataset after SSCD, only 40.4% of sites were variable. In this case, however, the reduction of some 10% in the variability did not prevent a correct dating of the most recent common ancestor. Removing further phylogenetic signals again reduced the reliability of the data (see Results). As a rule of thumb, we can therefore state that it is important to maintain as much phylogenetic signal as possible, provided the dataset behaves clock-like.

For dating the origin of HIV-1 with SSCD, we used much longer HIV-1 sequences than the HCV ones and analyzed two different genomic regions: *pol* and *env*. The first and third codon positions of the selected *pol* region behaved already clock-like. For the *env* region, the two fastest categories of sites had to be removed. The estimated dates derived from these *pol* and *env* datasets are in agreement. The date of the most recent common ancestor of HIV-1 group M is not identical in the two datasets, although the confidence intervals overlap: 1920 with *pol* and 1937 with *env*. The discrepancy may be due, in part, to the fact that a longer nucleotide sequence has been used in the *pol* region to estimate the clock-like branch lengths of the phylogenetic tree, and in part to the fact that more sequences were available to estimate the *env* evolutionary rate resulting in a smaller confidence interval. However, the self-consistency of the two gene regions is reassuring: When clock-like datasets are used, the dates estimated employing two different genomic regions and sequences sampled in 1992 (in one case) and in 1998 (in the other), are similar. Also reassuring is that the origin of the B/D node is placed around 1949–1954. This result is fully consistent with the isolation of a HIV-1 strain from a Congolese patient in 1959 that was phylogeneticaly placed near the ancestral B/D node with a very short branch length, suggesting a few years prior to 1959 as the origin of the B/D node itself (5). Finally, it is important to point out that our data agree with the data recently published by Korber and colleague (10). Korber's method of molecular-clock analysis is based on a rather different approach. Comparing the branch lengths of strains sampled at different times and introducing a correction to take into account that different lineages may evolve at different rate, they reported 1930 (95% c.i. 1910–1950) as the date of the most recent common ancestor of group M.

Our new analysis, dating major events in the HIV-1 group M history, does not inform us how HIV-1 was transferred from simians to humans, but it does provide a more accurate view about when this happened. The findings suggest that zoonotic transmission must have occurred before the 1920s–1930s, because by that time the HIV-1 subtypes already started to diverge. Knowing when HIV-1 entered human populations allows us to exclude certain speculations on possible scenarios about how this event may have occurred and instead focus on other speculations that are more appropriate considering the dates reported here. For example, our dating seems to exclude—or at least reduce the likelihood—the theory that vaccination with oral polio vaccine batches contaminated with SIV is at the origin of the HIV epidemics in humans. According to our data, the most recent common ancestor of the HIV-1 subtypes responsible for the AIDS epidemic arose much earlier than 1957–1959, the period when these vaccines were first used on a large scale.

As SIV strains more closely related to HIV-1 other than the SIVcpz from *Pan troglodytes troglodytes* have not been found so far, the separation of the HIV-1/SIVcpz lineages (around the end of the XVII century) is the upper limit of the time of the interspecies transmission that was responsible for the origin of HIV-1 group M in humans. Our dates thus indicate an interval of more than 240 years between the SIVcpz common ancestor and the time when HIV-1 group M started to diverge. However, this does not necessary mean that the time of the interspecies transmission must have occurred more than 200 years ago; the date simply represents the coalescence time of the simian ancestor of HIV-1 group M and its closest related SIVcpz strains. It is important to realize, however, that the lack of reported cases of HIV-1-related disease during these 240 years does not necessarily imply an absence of infection in human populations, particularly when these cases occurred in remote geographical locations. Interspecies transmission of simian retroviruses to humans has occurred repeatedly in the human history. All HTLV-I subtypes resulted from ancient and recent simian-to-human transmissions (28), and foamy viruses were also proven to have crossed simian–human species barriers (29). The factors contributing to a successful simian-to-human transmission and the factors contributing to the expansion of the initial infection into epidemic and then pandemic proportions are not necessarily the same (30). The lack of sterile conditions in vaccination campaigns and the major socio-cultural changes in Africa during the past century might have contributed to the spread of the HIV epidemic, long after the interspecies transmission had occurred (30). Hence, it is not surprising that the dates of HIV-1 origin and the first reported waves of the AIDS epidemic are not very closely linked temporally as is suggested by our analysis.

ACKNOWLEDGMENTS

REFERENCES

1. Sharp, P. M., Robertson, D. L., Gao, F., Hahn B. H. (1994) Origins and diversity of human immunodeficiency viruses. *AIDS* **8**, S27–S42

2. Triques, K., Bourgeois, A., Vidal, N., Mpoudi-Ngole, E., Mulanga-Kabeya, C., Nzilambi, N., Torimiro, N., Saman, E., Delaporte, E., and Peeters, M. (2000) Near-full-length genome sequencing of divergent African HIV type 1 subtype F viruses leads to the identification of a new HIV type 1 subtype designated K. *AIDS Res. Hum. Retrovir.* **16**, 139–151

3. Louwagie, J., Janssens, W., Mascola, J., Heyndrickx, L., Hegerich, P., van der Groen, G., McCutchan, F. E., and Burke, D.S.J. (1995) Genetic diversity of the envelope glycoprotein from human immunodeficiency virus type 1 isolates of African origin. *J. Virol.* **69**, 263–271

4. Gao, F., Bailes, E., Robertson, D. L., Chen, Y., Rodenburg, C. M., Michael, S. F., Cummins, L. B., Arthur, L. O., Peeters, M., Shaw, G. M., Sharp, P. M., and Hahn, B.H.G. (1999) Origin of HIV-1 in the chimpanzee *Pan troglodytes troglodytes*. *Nature* **397**, 436–441

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

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

7. Sharp, P. M., Bailes, E., Robertson, D. L., Gao, F., and Hahn, B. H. (1999) Origins and evolution of AIDS viruses. *Biol. Bull.* **196**, 338–342

8. Korber, B.T.M., Theiler, J., and Wolinsky, S. (1998) Limitations of a molecular clock applied to considerations of the origin of HIV-1. *Science* **280**, 1868–1871

9. Goudsmit, J., and Lukashov, V. (1999) Dating the origin of HIV-1 subtypes [Letter]. *Nature* **400**, 325–326

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

11. Duffy, M., Hegarty, J., Curry, M., Nolan, N., Kelleher, D., McKiernan S., and Hall, W.W. The extent of molecular evolution of hepatitis C virus (HCV) following a single source of infection is not directly related to the severity of chronic liver disease. (submitted to *J. Virol.*).

12. Thompson, J., Higgins, D., and Gibson, T. J. (1994) *Nuc. Acids. Res.* **22**, 4673.

13. McAllister, J., Casino, C., Davidson, F., Power, J., Lawlor, E., Yap, P. L., Simmonds, P., Smith, D. B., and McAllister, J. (1998) Long-term evolution of the hypervariable region of hepatitis C virus in a common-source-infected cohort. *J. Virol.* **72**, 4893–4905

13. Los Alamos, N.M. (1999). Theoretical Biology and Biophysics, Los Alamos National laboratory http://hiv-web.lanl.gov.

14. Vergne, L., Peeters, M., Reynes, J., Mboup S., Liegeois, F. et al. (1999) Molecular diversity of HIV-1 group M protease and RT gene sequences in Africa: No evidence for major drug-resistance mutations in drug naive individuals. In *Seventh European Conference on Clinical Aspects and Treatment of HIV-Infection;* Lisbon, Portugal, October 23–27, Abstract 385.

15. Felsenstein, J. (1993) PHYLIP: Phylogenetic Inference Package, version 3.5c. Seattle: Department of Genetics, University of Washington.

16 Strimmer, K., von Haeseler, A. (1996) *Mol. Biol. Evol.* **13,** 964.

17. Yang, Z. (1997) PAML: a program package for phylogenetic analysis by maximum likelihood. CABIOS **13,** 555–556

18. Xia, X. In *Data Analysis in Molecular Biology and Evolution;* Kluwer Academic Publishers, 2000

19. Hayashida, H., Toh, H., Kikuno, R., and Miyata, T. (1985) Evolution of influenza virus genes. *Mol. Biol. Evol.* **2,** 289–303

20. Salemi, M., Lewis, M., Egan, J. F., Hall, W. W., Desmyter, J., and Vandamme, A. M. (1999) Different population dynamics of human T-cell lymphotropic virus type II in intravenous drug users compared with endemically infected tribes. *Proc. Natl. Acad. Sci.* **96,** 13253–13259

21. STATISTICA for the Macintosh, Statsoft Inc. (Tulsa, Okla.).

22. Li, W. H., Graur, D. *Molecular Evolution;* second edition. Sinauer Associates, Inc. Publishers, 1997

23. Kimura, M. (1987) Molecular evolutionary clock and the neutral theory. *J. Mol. Evol.* **26,** 24–33

24. Felsenstein, J., Churchill, G. A. (1996) A hidden markov model approach to variation among sites in rate of evolution. *Mol. Biol. Evol.* **13,** 93–104

25. Huelsenbeck, J. P., Rannala, B. (1997) Phylogenetic methods come of age: testing hypotheses in an evolutionary context. *Science* **276,** 227–232

26. Philippe, H., Laurent, J. (1998) How good are deep phylogenetic trees? *Curr. Op. Gen. Dev.* **8,** 616–623

27. Yamaguchi-Kabata, Y., Gojobori, T. (2000) Reevaluation of amino acid variability of the human immunodeficiency virus type 1 gp120 envelope glycoprotein and prediction of new discontinuous epitopes. *J. Virol.* **74,** 4335–4350

28. Vandamme, A.-M., Salemi, M., and Desmyter, J. (1998) The simian origins of the pathogenic human T cell lymphotropic virus type I. *Trends Microbiol.* **6,** 477–483

29. Heneine, W., Switzer, W. M., Sandstrom, P., Brown, J., Vedapuri, S., Schable, C. A., Khan, A. S., Lerche, N. W., Schweizer, M., Neumann-Haefelin, D., Chapman, L. E., and Folks, T. M. (1999) Identification of a human population infected with simian foamy viruses. *Nature Med.* **4**, 403–407

30. Chitnis, A., Rawls, D., and Moore, J. (1999) Origin of HIV Type 1 in Colonial French Equatorial Africa? *AIDS Res. Hum. Retrovir.* **16**, 5–8

Table 1

Site Stripping for Clock Detection (SSCD) employing 19 HCV individuals infected in 1977 and sampled in 1998.

rate category*	number of nucleotides	number of variable sites retained	p-value Clock/ non Clock tree	clock-like branch length (to the common ancestor)	estimated evolutionary rate	estimated year of origin
1 to 16	360	194	0.03	0.104 ± 0.008	$2.4\ 10^{-3} \pm 1.1\ 10^{-3}$	1955 ± 20
1 to 15	339	173	0.03	0.054 ± 0.005	$2.2\ 10^{-3} \pm 0.8\ 10^{-3}$	1974 ± 9
1 to 14	325	159	**0.16**	0.041 ± 0.004	$1.9\ 10^{-3} \pm 0.6\ 10^{-3}$	1977 ± 7
1 to 13	294	128	**0.88**	0.025 ± 0.003	$6.8\ 10^{-4} \pm 6.6\ 10^{-4}$	1962 ± 35
1 to 12	269	103	**0.15**	0.016 ± 0.002	$2.5\ 10^{-4} \pm 4.8\ 10^{-4}$	1934 ± 123

* each site has been assigned to 16 categories of rates, from 1 (the slowest) to 16 (the fastest). The clock i s tested with the likelihood ratio test employing all the sites and with subsets where sites belonging to the fastest categories are progressively removed. The p-value for which the clock hypothesis cannot be rejected ($p > 0.05$) is indicated in bold. Standard errors of the estimates are given.

Table 2

Site stripping method to test the molecular clock hypothesis for HIV/SIVcpz *pol* and *env* genes

rate category*	*pol*			*env*		
	1st cdp	2nd cdp	3rd cdp	1st cdp	2nd cdp	3rd cdp
1 to 16	484 **(0.14)**	484 (0.002)	484 **(0.21)**	737 (0.01)	737 (0.007)	737 (0.042)
1 to 15	455 **(0.48)**	462 (0.006)	446 **(0.35)**	703 (0. 032)	692 (0.008)	671 (0.0001)
1 to 14		423 (0.047)	423 **(0.35)**	656 **(0.065)**	653 (0.015)	631 $(3\ 10^{-6})$
1 to 13		381 **(0.92)**		606 **(0.19)**	611 (0.0005)	591 (0.0004)
1 to 12				557 **(0.20)**	569 (0.002)	545 (0.003)
1 to 11					512 **(0.11)**	514 (0.002)
1 to 10					494 **(0.17)**	480 (0.039)
1 to 9						443 (0.004)
1 to 8						383 **(0.29)**

* each site has been assigned to 16 categories of rates, from 1 (the slowest) to 16 (the fastest). The clock is tested for each codon positions using all the sites and with subsets where sites belonging to the fastest categories are progressively removed.

The p-values (given in parenthesis) for which the clock hypothesis cannot be rejected ($p > 0.05$) are indicated in bold. The number of nucleotides underlined are those that were chosen to estimate clock-like branch lengths for the *pol* and *env* trees in figure 1 and to calibrate the evolutionary rate.

Table 3

Divergence times of the major nodes of the HIV-1/SIVcpz phylogenetic tree.

node	pol [evolutionary rate: $6.0 \times 10^{-4} \pm 4.9 \times 10^{-5}$]		env [evolutionary rate: $1.0 \times 10^{-3} \pm 0.66 \times 10^{-4}$]	
	branch length*	estimated year of origin #	branch length**	estimated year of origin #
B subtype	0.017 ± 0.001	1970 (1963 - 1978)	0.020 ± 0.002	1972 (1966 - 1978)
B/D node	0.03 ± 0.002	1949 (1935 - 1962)	0.038 ± 0.003	1954 (1944 - 1964)
M group	0.047 ± 0.002	1920 (1902 - 1939)	0.055 ± 0.003	1937 (1925 - 1949)
SIVcpz/HIV-1	0.194 ± 0.012	1675 (1590 - 1761)		

* the length of the branch was estimated assuming a molecular clock and using 1st+3rd codon position of *pol* underlined in Table 2. Standard errors are indicated.

** the height of the branch was estimated assuming a molecular clock and using the stripped 1st codon podition of *env* underlined in Table 2. Standard errors are indicated.

\# 99% confidence intervals for the estimated year of origin are given in parenthesis.