
Dating the common ancestor of SIVcpz and HIV-1 group M and the origin of HIV-1 subtypes using a new method to uncover clock-like molecular evolution¹

MARCO SALEMI,* KORBINIAN STRIMMER,[†] WILLIAM W. HALL,[‡] MARGARET DUFFY,[‡] ERIC DELAPORTE,[§] SOULEYMANE MBOUP,^{||} MARTINE PEETERS,[§] AND ANNE-MIEKE VANDAMME*²

*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

SPECIFIC AIMS

To investigate the time frame of the common ancestor of HIV-1 group M and its closest simian relative, SIVcpz, we developed a new method of molecular clock analysis, called site stripping for clock detection (SSCD). SSCD allows selection of nucleotide sites evolving at an equal rate in different lineages. We calculated that the origin of HIV-1 group M radiation dates back to the 1920s–1930s and that the coalescence time of HIV-1 group M and its simian counterpart occurred around the end of the XVII century.

PRINCIPAL FINDINGS

1. SSCD is a method of general applicability in molecular evolution to calibrate clock-like phylogenetic trees

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 that 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. Several factors, such as homoplasy, recombination and positive selection can impair this linear relation, but not all the sites in a set of aligned sequences deal with these problems. Our new method, SSCD, helps detecting those sites that are more likely to distort the molecular clock. These sites can be removed from an alignment before calibrating a clock-like phylogenetic tree. The SSCD procedure was validated by using a cohort of HCV-infected women, all infected within a couple of months during 1977. The source of the infection was identified as a contaminated anti-D immunoglobulin batch prepared

from a single blood donation from an HCV-genotype-1b positive donor. We used the SSCD procedure on a set of aligned sequences of HCV strains, sampled from the infected patients in 1998, and we calibrated the evolutionary rate of HCV by using sequences sampled in 1994 from other women infected in 1977 by the same batch. The origin of the common ancestor of the viral strains sampled in 1998 was dated exactly in 1977, with a standard error of 7 years. However, when the molecular clock was calibrated without removing the clock-distorting sites, the resulting estimated date was inaccurate (1955 ± 20).

2. The radiation of the HIV-1 group M subtypes occurred around the 1920 and 1930s

Figure 1 shows the phylogenetic trees of HIV-1 *pol* strains sampled in 1998 and *env* strains sampled in 1992. The trees are not fully resolved, and they should be viewed as a consensus of the evolutionary history of the sequences. The rationale behind using multi-furcating trees is that we are interested mainly in dating the common ancestors of the clades indicated in Fig. 1, for which all the phylogenetic analyses gave robust support. For both trees the molecular clock hypothesis has to be rejected. However, by using the SSCD procedure, we could find which sites in the *pol* and *env* alignment are more likely to distort the molecular clock. After removing these sites, we used the remaining sites to calibrate clock-like branch lengths for the trees in Fig. 1. We estimated the evolutionary rate for the *pol* end *env* clock-like evolving sites by using several HIV-1

¹ To read the full text of this article, go to <http://www.fasebj.org/cgi/doi/10.1096/fj.00-0449fje> To cite this article, use (December 8, 2000) *FASEB J.* 10.1096/fj.00-0449fje

² Correspondence: Rega Institute for Medical Research, Minderbroedersstraat 10, B-3000 Leuven, Belgium. E-mail: annemie.vandamme@uz.kuleuven.ac.be

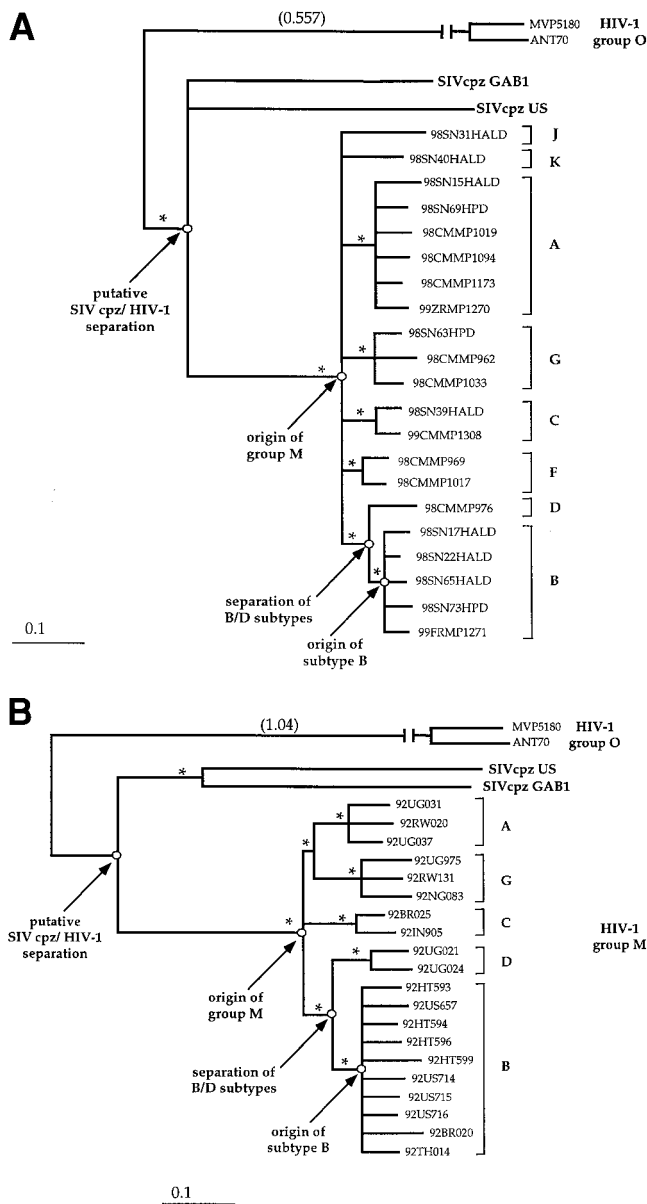


Figure 1. Phylogenetic analysis of HIV-1 group M and SIVcpz from *Pan troglodytes troglodytes* in different gene regions. Trees also include HIV-1 group O as the outgroup. Branch lengths were estimated by assuming unequal evolutionary rates along them. Horizontal branch lengths are drawn to scale with the bar indicating 0.1 nucleotide replacements per site. Nodes with asterisks were supported by greater than 95% bootstrap samples (out of 1,000). The branch length to the outgroup is mentioned on the branch. The strains are described in Materials and Methods. *A)* *pol* region, nt 2358-3809; *B)* full-length envelope.

sequences sampled between 1983 and 1999. The results are shown in **Table 1**. The origin of subtype B radiation is estimated to be in the early 1970s (1970 with *pol*, and 1972 with *env*). The most recent common ancestor of subtypes B and D dates back between the end of the 1940s and the early 1950s (1949 with *pol*, and 1954 with *env*), and the date of origin of group M radiation is estimated to be during the 1920s and the 1930s; 1920 according to *pol*, and 1937 according to *env*.

3. HIV-1 group M and its closest simian relative, SIVcpz, shared a common ancestor in 1675

Once the sites responsible for distorting the molecular clock were removed, the clock could not be rejected for the trees in Fig. 1, including HIV-1 group M strains as well as SIVcpz strains. We estimated that the most recent common ancestor of HIV-1 group M and SIVcpz (see Fig. 1) dates from around 1675 (99% c.i. 1590–1761). The estimation was based on the *pol* gene only, as the *env* dataset shows evidence of nucleotide substitutions saturation when the SIVcpz strains are compared with the other HIV-1 group M strains. Using sites that show saturation would result in an unreliable divergence time.

CONCLUSIONS

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 objective is achievable by stripping from an alignment the sites that distort the clock and by retaining the clock-like sites that are used subsequently in the calculations. Our results show that

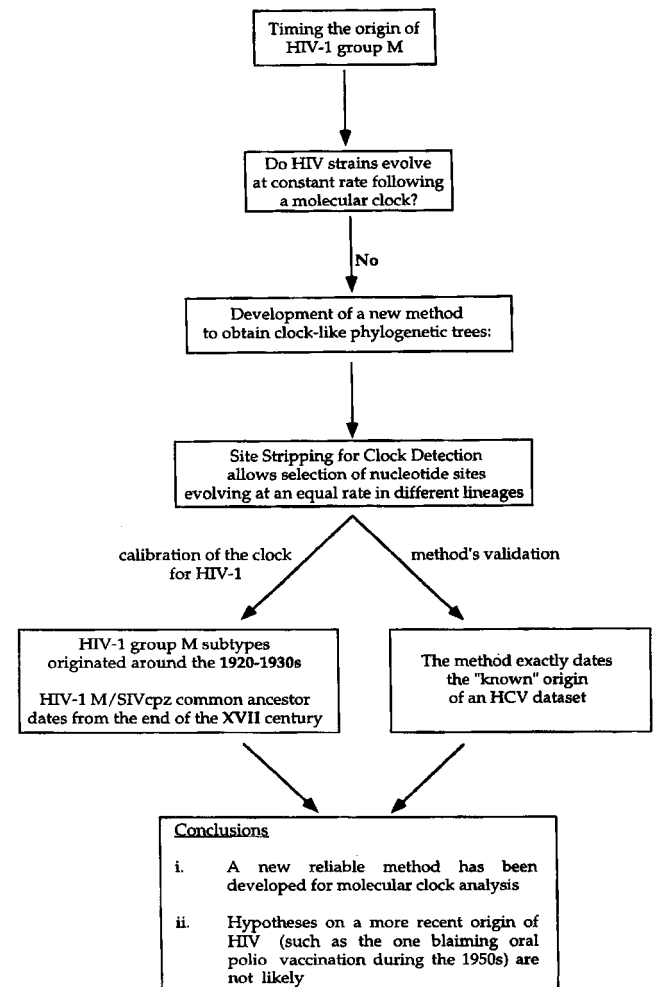


TABLE 1. Divergence times of the major nodes of the HIV-1/SIVcpz phylogenetic tree

Node	<i>pol</i> (Evolutionary rate: $6.0 \cdot 10^{-4} \pm 4.9 \cdot 10^{-5}$)		<i>env</i> (Evolutionary rate: $1.0 \cdot 10^{-3} \pm 0.66 \cdot 10^{-4}$)	
	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 using 1st + 3rd codon position clock-like evolving sites of *pol*. Standard errors are indicated. **The height of the branch was estimated using 656 out of 737 nucleotide clock-like evolving sites at 1st codon position of *env*. Standard errors are indicated. [#]99% confidence intervals for the estimated year of origin are given in parenthesis.

upon identification and removal of the sites for which the molecular clock fails, the remaining sites can be used reliably to construct clock-like phylogenetic trees and to calculate divergence times at ancestral nodes, such as those illustrated by the HCV example.

We applied the new procedure in order to investigate the time of origin of HIV-1 group M by using a dataset of *pol* and *env* viral sequences, respectively. The date of the most recent common ancestor of HIV-1 group M is not identical in the two datasets, but they are in good agreement (see Table 1). 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 phylogenetically placed near the ancestral B/D node with a very short branch length and suggests a few years prior to 1959 as the origin of the B/D node itself.

Our findings suggest that a zoonotic transmission responsible for the introduction of HIV-1 into our species must have occurred before the 1920s–1930s, because by that time the HIV-1 subtypes had already begun to diverge. The dating seems to exclude, or at least to make more unlikely, the theory that vaccination with oral polio vaccine batches contaminated with SIV, between 1957–1959 in the former Belgian Congo, is the origin of the HIV epidemics in humans.

Finally, SIV strains more closely related to HIV-1, other than the SIVcpz from *Pan troglodytes troglodytes*, have not been found thus far. Therefore, the separation of the HIV-1/SIVcpz lineages around the end of the XVII century (see Table 1) is the upper limit of the time of the interspecies transmission responsible for the origin of HIV-1 group M in humans. This finding does not necessarily mean that the time of the interspecies transmission must have occurred more than 300 years ago; the date simply represents the coalescence time of the simian ancestor of HIV-1 group M and its closest related SIVcpz strains. However, it is important to realize that the lack of reported cases of HIV-1-related disease during these years does not necessarily imply an absence of infection in human populations, particularly when this disease has occurred in remote geographical locations. 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. 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. **FJ**