An African HIV-1 sequence from 1959 and implications for the origin of the epidemic

Tuofu Zhu*†, **Bette T. Korber**‡§, **Andre J. Nahmias**‖,
Edward Hooper¶, **Paul M. Sharp**# & **David D. Ho***

* Aaron Diamond AIDS Research Center, The Rockefeller University,
455 First Avenue, New York, New York 10016, USA
‡ Theoretical Division, Los Alamos National Laboratory, Los Alamos,
New Mexico 87545, USA
§ The Santa Fe Institute, Santa Fe, New Mexico 87501, USA
‖ Department of Pediatrics, Emory University School of Medicine, Atlanta,
Georgia 30303, USA
¶ PO Box 4087, Worthing BN14 7LQ, UK
# Department of Genetics, University of Nottingham, Queens Medical Centre,
Nottingham NG7 2UH, UK

There is considerable genetic diversity among viruses of different subtypes (designated A to J) in the major group of human immunodeficiency virus type 1 (HIV-1), the form of HIV that is dominant in the global epidemic[1–3]. If available, HIV-1 sequences pre-dating the recognition of AIDS could be crucial in defining the time of origin and the subsequent evolution of these viruses in humans. The oldest known case of HIV-1 infection was reported to be that of a sailor from Manchester who died of an AIDS-like illness in 1959 (refs 4–6); however, the authenticity of this case has not been confirmed[7,8]. Genetic analysis of sequences from clinical materials obtained from 1971 to 1976 from members of a Norwegian family infected earlier than 1971 showed that they carried viruses of the HIV-1 outlier group[9,10], a variant form that is mainly restricted to West Africa[1]. Here we report the amplification and characterization of viral sequences from a 1959 African plasma sample that was previously found to be HIV-1 seropositive[11]. Multiple phylogenetic analyses not only authenticate this case as the oldest known HIV-1 infection, but also place its viral sequence near the ancestral node of subtypes B and D in the major group, indicating that these HIV-1 subtypes, and perhaps all major-group viruses, may have evolved from a single introduction into the African population not long before 1959.

A serological study of 1,213 plasma samples obtained from Africa between 1959 and 1982 has been reported[11]. Twenty-one samples were initially found to be HIV-1 seropositive by immunoassay, but only one was confirmed as reactive with HIV-1 by immunofluorescence, western blotting and radioimmunoprecipitation methods. This positive plasma sample, designated L70, was obtained in early 1959 from an adult Bantu male, with a sickle-cell trait and a glucose-6-phosphate-dehydrogenase deficiency, living in Leopoldville, Belgian Congo (now Kinshasa, Democratic Republic of Congo)[11,12]. The viral sequences contained in this sample might provide insights into the evolution of HIV-1.

Because of the limited amount of plasma available from this sample and uncertainty about its condition, efforts were made to increase the likelihood of recovering HIV-1 sequences by RT-PCR

† Present address: University of Washington, Health Sciences Building, Room T239X, 1959 NE Pacific Avenue, Seattle, Washington DC 98195, USA.

(reverse transcription followed by polymerase chain reaction). Multiple primers were used in a single RT reaction, and all synthesized complementary DNAs were amplified by PCR using primers designed to amplify HIV-1 sequences from all known subtypes (see Methods). However, attempts to amplify HIV-1 fragments of >300 base pairs (bp) were unsuccessful, probably because of a low level of intact viral RNA in this old sample. However, after numerous attempts, four shorter sequences were obtained. As shown in Fig. 1, fragments a, b and c corresponded to env-gene regions containing the V3 loop, the gp120–gp41 junction (including the Rev-responsive element, RRE), and the carboxy-terminal domain of gp41, respectively; fragment d corresponded to a region of the pol gene. Each PCR product was then purified, cloned and sequenced.

Nucleotide sequences from these HIV-1 fragments were designated ZR59a, ZR59b, ZR59c and ZR59d (GenBank accession numbers AF030526–AF030544, AF030637–AF030651, AF030652–AF030671 and AF030672–AF030686 for a, b, c and d, respectively). Sequences were aligned to available sequences in the 1996 Los Alamos Database using multiple aligned sequence editor[13]. Between 6 and 17 HIV-1 sequences were obtained from each region, and each set of ZR59 sequences formed a tight cluster with only 0–3% divergence when analysed phylogenetically using the neighbour-joining method. Each sequence was also run against those in GenBank using basic local alignment search tool[14] to find the closest match. Maximum sequence identity scores of 92% (ZR59a), 96% (ZR59b) and 94% (ZR59c) confirmed that the ZR59 sequences

are indeed unique and unlikely to be the result of PCR contamination.

Consensus sequences for ZR59a, ZR59b and ZR59c were concatenated for in-depth studies. Phylogenetic analyses were performed in the laboratories of B.T.K. and P.M.S. using various approaches and slightly different alternative alignments, but with similar results and conclusions. Three methods were used for phylogenetic analysis of the ZR59a–c sequences, namely, minimum evolution (neighbour joining), maximum likelihood and weighted parsimony (see Methods); the results of all three approaches were similar. The ZR59 sequence was positioned close to the ancestral node of subtypes B, D and F, although with a slightly preferred association with the D lineage (Fig. 2). Although the ZR59d sequence was phylogenetically less informative (because of a higher degree of sequence conservation in pol), analyses of this fragment resulted in a similar conclusion.

Several HIV-1 isolates are hybrids of different subtypes[15], and the phylogenetic position found for ZR59 could be an artefact of it being a recombinant containing parts of modern subtypes D and B or F. Therefore, the ZR59 sequences were scanned using RIP[16], a program that looks for the possibility of recombination in a query sequence relative to a set of control sequences. In this case, ZR59 was compared with the consensus sequence of each subtype within the HIV-1 major group. In a similarity comparison, all four ZR59 fragments were slightly closer to the B-clade consensus, but in the env fragments there were short stretches that were more similar to fragments of the D clade. However, overall, there was no statistically
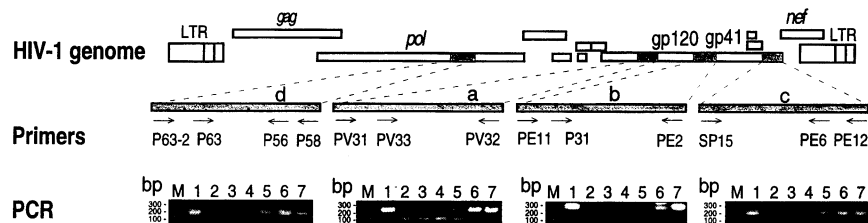
Figure 1 Detection of HIV-1 sequences in the L70 plasma sample by RT-PCR. The HIV-1 genome is shown at the top. ZR59 fragments (a, b, c and d) that were successfully amplified are shown in the middle, as are the primers used. Results of the successful RT-PCR are shown on the bottom. M, molecular marker; lane 1, positive plasma control; lane 2, negative plasma control; lane 3, reagent control (sterile water); lane 4, L70 plasma without reverse transcription before PCR; lane 5, L70 plasma; lane 6, serum from an infected patient identified in 1985; lane 7, serum from an infected patient (H.) identified in 1994. Each gel corresponds to the ZR59 fragment it is shown under. LTR, long terminal repeat.
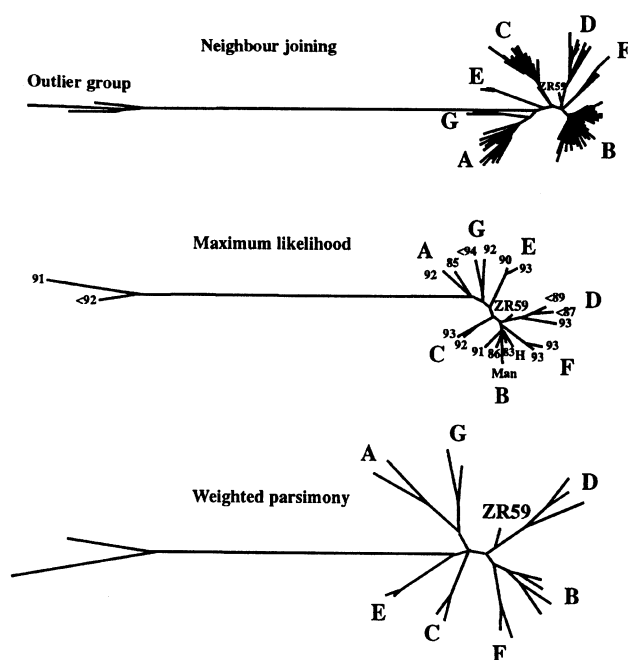
Figure 2 Phylogenetic analyses of the ZR59 sequence. ZR59 branched off the D clade near the B/D/F root in all analyses. The taxa are labelled with the year of sampling of a given sequence, or with '<' and the year of the primary publication if the year of sampling was not specified. The analyses included the following reference sequences and their GenBank accession numbers: A 92, 92RW020 (U08794); A 85, U455 (M62320); B 91, 91HT652 (U08443); B 83, LAI (X01762); B 86, JRFL (U63632); H, the internal control for this study sampled in 1994; Man, the Manchester sailor sequence[7] (U23487); C 92, 92BR025 (U09132); C 93, 93MW965 (U08455); D <89, NDK (M27323); D <87, Z6 (K03458); D 93, 92ZR001 (U27419); E 90, CM240 (U54771); E 93, 93TH966 (U08456); F 93, 93BR020 (U27401); F 93, 93BR029 (U27413); G <94, LBV217 (U09664); G 92, 92UG975 (U22010); outlier group 91, VAU (X80020); and outlier group <92, MVP5180 (L20571). The same set of input taxa was used for the weighted-parsimony and maximum-likelihood trees, except that H was included in the maximum-likelihood tree only. A larger set of taxa was used for the neighbour-joining tree. The scale for branch lengths is comparable for the maximum-likelihood and neighbour-joining trees. For the weighted-parsimony tree, the branch lengths are not directly comparable to the branch lengths of the other two trees, but the branching pattern substantiates the pattern obtained by the other two methods.

significant evidence, using RIP[16] or other approaches[15], for a recombination crossover point. Thus, ZR59 is not a mosaic of modern sequences. If ZR59 were a result of a recombination of the B and D subtypes soon after they diverged, the recombination would be of little consequence and would be unlikely to be detectable, especially with the short sequences available.

For most regions of the HIV-1 genome, subtypes B and D are more closely associated with each other than are any other subtypes within the major group[1-3]. However, phylogenetic analysis of the *env* regions showed an unusual association between viruses of subtypes B and F. RIP analysis of the entire *env* gene showed that subtype B is more similar to subtype F than to subtype D, specifically over two short stretches, encompassing the V3 and RRE regions, that coincide with fragments a and b. This finding explains the unusual B/D/F clustering found in our phylogenetic analyses (Fig. 2).

The rate of evolution of HIV-1 sequences has been estimated[17,18], although a precise translation of genetic distances into years is not possible for a single virus because of variation in the rate of evolution among different individuals or different lineages[19,20]. Nevertheless, because of the high evolutionary rate of HIV-1, a virus from 1959 should have evolved to a substantially lesser extent than viruses isolated in the 1980s or later. To test this hypothesis, a comparison was made between the likelihood of the original tree and that of a tree that had an elongated branch between the ZR59 sequence and the D ancestral node, with an imposed branch length comparable to that found in more modern isolates. The branch length of a contemporary subtype D virus sequence to the D ancestral node was found for the sequence NDK, which was then used to fix the branch length for ZR59 to the ancestral node. The tree was subsequently optimized with respect to the new branch length using maximum likelihood, and the log likelihood values of the original tree and the modified tree were compared using the method of Kishino and Hasegawa[21] as implemented by PHYLIP. The difference of 14.85 in log likelihood values for the two trees showed that the modified tree, with a long branch length for ZR59, was significantly less likely ($P < 0.05$). In other words, ZR59 displayed properties of an older sequence. A likelihood-ratio test was also used to test the placement of ZR59 as an early branch-off from the B clade rather than from the D clade, but there was no significant difference between the two alternatives. This conclusion is consistent with results from bootstrap analyses. Although the maximum-likelihood, weighted-parsimony and neighbour-joining trees of Fig. 2 yielded relatively high bootstrap values (83, 75 and 40, respectively), supporting the existence of a branch point that includes the B, D and F subtypes and ZR59, the association between ZR59 and clade D was weaker (bootstrap values of 56, <50 and 38 were obtained from the maximum-likelihood, weighted-parsimony and neighbour-joining trees, respectively). Thus, although a branching-off of ZR59 from the D lineage is the probable result from each of the tree-building methods, this conclusion is not definitive.

We estimated the branch length between ZR59 and the ancestral node of subtypes B, D and F by several methods. Using maximum likelihood, a distance (substitutions per site) of 0.023 was obtained with the fastDNAml and DNArates programs (Fig. 2). A distance of 0.037 was obtained with PAML using a REV model and optimized gamma distribution (see Methods and ref. 22). PHYLIP 3.6 was also used to reconstruct an ancestral sequence based on a maximum-likelihood estimate of the most probable character at each site in the nodal sequence. A distance of only 0.026 separated ZR59 from this reconstructed sequence. Thus, ZR59 is similar to the ancestral sequence of HIV-1 subtypes B, D and F. But are these sequences identical? The maximum-likelihood tree in Fig. 2 was compared with one in which ZR59 has a zero branch length from the B, D and F ancestral sequence. The latter tree was significantly less likely ($P < 0.05$; ref. 21) with a difference in log likelihood of 34.17,

indicating that ZR59 was probably slightly more recent than the B/D/F ancestor.

The short but non-zero distance from the common ancestor of subtypes B and D (and F) to ZR59 indicates that ancestral HIV-1 of subtypes B and D (and F) must have existed before 1959, although probably only a few years before then. This is about a decade earlier than the previous estimate for the divergence of B and D clades[17]. This finding also refutes the suggestion that HIV-1 subtype-B infection was responsible for AIDS-like syndromes beginning in the 1930s in various European populations[23]. Our results also indicate that subtypes B, D and F may have evolved within the human population rather than arising from multiple cross-species transmission events[1,22].

The phylogenetic analyses (Fig. 2) show that ZR59 is not too distant from the internal node for all major-group viruses; however, evolutionary rates may differ in different regions of the phylogenetic tree. Nonetheless, both the major- and the outlier-group viruses were clearly present in humans by 1960 (ref. 10). Our results, the rate of HIV-1 evolution (around 0.005–0.01 nucleotide changes per site per year[17]) and previously described methods of estimation of evolutionary rates[24] indicate that the major-group viruses that dominate the global AIDS pandemic at present shared a common ancestor in the 1940s or the early 1950s. Given their 'starburst' phylogeny, HIV-1 was probably introduced into humans shortly before that time frame, about a decade or two earlier than previously estimated[17,25]. Thus, given the large genetic distance between HIV-1 and HIV-2, the divergence of these viruses could not have occurred in the late 1940s (ref. 25); that branching point must have been considerably earlier[17,26-28]. The diversification of HIV-1 in the past 40–50 years portends even greater viral heterogeneity in the coming decades, and underscores the need for continued surveillance. The factors that propelled the initial spread of HIV-1 in central Africa remain unknown: the role of large-scale vaccination campaigns, perhaps with multiple uses of non-sterilized needles, should be carefully examined, although social changes such as easier access to transportation, increasing population density and more frequent sexual contacts may have been more important. □

Methods

RNA isolation and reverse transcription. The L70 plasma sample (200 μl) was diluted with 10 ml of precooled (4 °C) PBS and centrifuged at 40,000 r.p.m. for 4 h at 4 °C in a swing rotor (Sorvall SW40 TI). The pellet was then re-suspended in 140 μl of precooled (4 °C) PBS. A QIAamp HCV kit (QIAGEN) was used to extract the viral RNA, which was eluted in 50 μl of RNase-free sterile water. The viral RNA was incubated with RNase-free DNase in the presence of 15 mM $MgCl_2$ for 20 min at 37 °C, and then incubated for 5 min at 80 °C to end the reaction. The RNA (25 μl) was mixed with 40 ng of each of the following antisense primers: P58 (residues 3798–3776, according to the HXB2 sequence in the Los Alamos Database; sequence 5'-GAC AAA CTC CCA CTC AGG AAT CCA-3'), PV32 (residues 7258–7235, 5'-TAC CTG TTG TAA AGT GTT ATT CCA-3'), PE2 (residues 7956–7934, 5'-GCC TGG AGC TGT TTA ATG CCC CA-3'), and PE12 (residues 8871–8844, 5'-CTG GCT CAG CTC GTC TCA TTC TTT CCC T-3'). This mixture was heated to 70 °C for 10 min and used to synthesize HIV-1 cDNA at 42 °C for 50 min in a solution containing 50 mM KCl, 10 mM Tris-HCl (pH 9.0), 0.1% Triton X-100, 0.3 mM of each of the deoxynucleoside triphosphates (dNTPs), 2.5 mM $MgCl_2$ and 40 units AMV reverse transcriptase (Promega). The mixture was finally heated at 70 °C for 15 min to inactivate the reverse transcriptase.

PCR and DNA sequencing. All newly synthesized cDNA was used simultaneously to amplify multiple HIV-1 sequences with primers P58 and P63 (residues 3580–3603, 5'-GCC ATT TAA AAA TCT GAA AAC AGG-3') for the *pol* region, PV32 and PV31 (residues 7019–7050, 5'-GCA GAA GAA GAG GTA GTA ATT AGA TCT GAA AA-3') for the V3 region, PE2 and PE11 (residues 7646–7671, 5'-ATG AGG GAC AAT TGG AGA AGT GAA TT-3') for the gp120/gp41-junction region, and PE12 and SP15 (residues 8600-8622, 5'-CTC AAA TAT TGG TGG AAT CTC CT-4') for the C terminus of gp41 (Fig. 1). The reaction buffer was the same as that described for reverse transcription.

except that the reverse transcriptase was replaced by four units of *Thermus aquaticus* DNA polymerase (Promega). Amplification cycles for the first round of the PCR were 96 °C for 4 min, 94 °C for 50 s, 55 °C for 30 s, and 72 °C for 1 min for 32 cycles, followed by a final extension at 72 °C for 10 min. The first-round PCR products (4 μl) were used in a second-round PCR to amplify specific HIV-1 regions with primer pairs P63 and P56 (residues 3757–3734, 5′-TGT CCA CCA TGC TTC CCA TGT TTC-3′) for the *pol* region, PV32 and PV33 (residues 7053–7079, 5′-TCA CAG ACA ATG CTA AAA CCA TAA TAG-3′) for the V3 region, PE2 and P31 (residues 7704–7736, 5′-TAG GAG TAG CAC CCA CCA AGG CAA AGA GAA GAG-3′) for the gp120–gp41-junction region, or SP15 and PE6 (residues 8822–8796, 5′-ACT ACT TTT TGA CCA CTT GCC ACC CAT-3′) for the C terminus of gp41. Amplification conditions were 94 °C for 2 min, 94 °C for 30 s, 55 °C for 25 s, and 72 °C for 1 min for 35 cycles, followed by a final extension at 72 °C for 10 min. All PCR reactions were carried out in the Perkin–Elmer model 9600 thermocycler. PCR products were purified with the QIAEX II gel extraction kit (QIAGEN) and then inserted into the PCR3.1 vector (Invitrogen) before DNA sequencing in an automated sequencer (ABI PRISM 377).

Phylogenetic analyses. The neighbour-joining tree was generated using PHYLIP 3.5 (J. Felsenstein). We used the 111 sequences from the *env* alignment (in the Los Alamos Database) that spanned all three regions of interest, a total of 468 bases in the alignment after gap-stripping. By using all available taxa, we could determine that the phylogenetic behaviour of the ZR59 sequence was unique, and that no sequences from the 1980s or 1990s gave a similar result. The distance matrix for the neighbour-joining tree was created using the F84 option of the program DNADIST with the likelihood estimates of the relative rates of site mutations described below (when a simpler Kimura two-parameter model was used to calculate the distance matrix, the ZR59 sequence preferentially branched off from the B clade in the neighbour-joining tree). We selected a representative subset of the 111 sequences to represent the major-group viruses; we tried to include sequences with known years of sampling. These sequences and the ZR59 sequence were subjected to tree-building methods that are computationally intensive, namely maximum likelihood and weighted parsimony. Maximum-likelihood analyses were done using several programs. FastDNAml[29] was used for the bootstrap analysis[30] and the initial maximum-likelihood tree. PHYLIP 3.5 was used for likelihood-ratio tests to compare the trees. We used PHYLIP 3.6 (J. Felsenstein) to reconstruct the most probable sequence at the ancestral nodes. All likelihood trees incorporated an estimate of rate variation between sites, an important element of accurate tree reconstruction, using the DNArates program (G. Olsen, personal communication) which optimizes the relative rate of substitution at each position in an alignment using maximum likelihood. A maximum-likelihood tree incorporating a REV model with an optimized gamma distribution using PAML (version 1.1, Z. Yang) was also generated[22]; this tree confirmed results obtained with fastDNAml. Statistical comparisons of the maximal-likelihood trees were made using the method of Kishino and Hasegawa[21]. When using weighted parsimony, we estimated substitution rates between the four bases from a parsimony tree generated by PAUP (D. Swofford), using the MacClade program; we then used the inverse of the observed substitution frequency to weight the character changes, by PAUP, in a subsequent parsimony tree. The accuracy of the branching pattern can be improved markedly by this approach, but the branch lengths do not reflect true distances.

1. Sharp, P. M., Robertson, D. L., Gao, F. & Hahn, B. H. Origins and diversity of human immuno-deficiency viruses. *AIDS* 8, S27–S43 (1994).
2. Gao, F. *et al.* Molecular cloning and analysis of functional envelope genes from human immuno-deficiency virus type 1 sequence subtypes A through G. *J. Virol.* 70, 1651–1667 (1996).
3. Louwagie, J. *et al.* Phylogenetic analysis of gag genes from 70 international HIV-1 isolates provides evidence for multiple genotypes. *AIDS* 7, 769–780 (1993).
4. Williams, G., Stretton, T. B. & Leonard, J. C. Cytomegalic inclusion disease and pneumocystis carinii infection in an adult. *Lancet* ii, 951–955 (1960).
5. Williams, G., Stretton, T. B. & Leonard, J. C. AIDS in 1959? *Lancet* ii, 1136 (1983).
6. Corbitt, G., Bailey, A. S. & Williams, G. HIV infection in Manchester, 1959. *Lancet* 336, 51 (1990).
7. Zhu, T. & Ho, D. D. Was HIV present in 1959? *Nature* 374, 503–504 (1995).
8. Hooper, E. & Hamilton, W. D. 1959 Manchester case of syndrome resembling AIDS. *Lancet* 348, 1363–1365 (1996).
9. Froland, S. S. *et al.* HIV-1 infection in Norwegian family before 1970. *Lancet* i, 1344–1345 (1988).
10. Jonassen, T. O. *et al.* Sequence analysis of HIV-1 group O from Norwegian patients infected in the 1960s. *Virology* 231, 43–47 (1997).
11. Nahmias, A. J. *et al.* Evidence for human infection with an HTLV-III/LAV-like virus in central Africa, 1959. *Lancet* i, 1279–1280 (1986).
12. Motulsky, A. G. *et al.* Population genetic studies in the Congo. 1. Glucose-6-phosphate dehydrogenase deficiency, hemoglobin S, and malaria. *Am. J. Hum. Genet.* 18, 514–537 (1966).
13. Faulkner, D. V. & Jurka, A. Multiple aligned sequence editor (MASE). *Trends Biochem. Sci.* 13, 321–322 (1988).
14. Altschul, S. F., Gish, W., Miller, W. & Myers, E. W. Basic local alignment search tool (BLAST). *J. Mol. Biol.* 215, 403–410 (1990).
15. Robertson, D. L., Sharp, P. M., McCutchan, F. E. & Hahn, B. H. Recombination in HIV-1. *Nature* 374, 124–126 (1995).
16. Siepel, A. C., Halpern, A. L., Macken, C. & Korber, B. A computer program designed to rapidly screen for HIV-1 intersubtype recombinant sequences. *AIDS Res. Hum. Retroviruses* 11, 1413–1416 (1995).
17. Li, W.-H., Tanimura, M. & Sharp, P. M. Rates and dates of divergence between AIDS-virus nucleotide sequences. *Mol. Biol. Evol.* 5, 313–330 (1988).
18. Kelly, J. K. An application of population genetic theory to synonymous gene sequence evolution in the human immunodeficiency virus (HIV). *Genet. Res.* 64, 1–9 (1994).
19. Lukashov, V. V., Kuiken, C. L. & Goudsmit, J. Intrahost human immunodeficiency virus type 1 evolution is related to length of the immunocompetent period. *J. Virol.* 69, 6911–6916 (1995).
20. Wolinsky, S. *et al.* Adaptive evolution of human immunodeficiency virus type 1 during the natural course of infection. *Science* 277, 537–542 (1996).
21. Kishino, H. & Hasegawa, M. Evaluation of the maximum likelihood estimate of the evolutionary tree topologies from DNA sequence data, and the branching order in Hominoidea. *J. Mol. Evol.* 29, 170–179 (1989).
22. Leitner, T., Escanilla, D., Franzen, C., Uhlen, M. & Albert, J. Accurate reconstruction of a known HIV-1 transmission history by phylogenetic tree analysis. *Proc. Natl Acad. Sci. USA* 93, 10864–10869 (1996).
23. Goudsmit, J. *Viral Sex: the Nature of AIDS.* (Oxford Univ. Press, 1997).
24. Leitner, T. S., Kumar, S. & Albert, J. Tempo and mode of nucleotide substitutions in gag and env gene fragments in human immunodeficiency virus type 1 populations with a known transmission history. *J. Virol.* 71, 4761–4770 (1997).
25. Smith, T. F., Srinivasen, A., Schochetman, G., Marcus, M. & Myers, G. The phylogenetic history of immunodeficiency viruses. *Nature* 333, 573–575 (1988).
26. Sharp, P. M. & Li, W-H. Understanding the origins of AIDS viruses. *Nature* 336, 315 (1988).
27. Eigen, M. & Nieselt-Struwe, K. How old is the immunodeficiency virus? *AIDS* 4 (suppl.) S85–S93 (1990).
28. Gojobori, T. *et al.* Evolutionary origin of human and simian immunodeficiency viruses. *Proc. Natl Acad. Sci. USA* 87, 4108–4111 (1990).
29. Olsen, G. J., Matsuda, H., Hagstrom, R. & Overbeek, R. fastDNAml: a tool for construction of phylogenetic trees of DNA sequences using maximum likelihood. *Comput. Appl. Biosci.* 10, 41–48 (1994).
30. Felsenstein, J. Evolutionary trees from DNA sequences: a maximum likelihood approach. *J. Mol. Evol.* 17, 368–376 (1981).