Proc. Natl. Acad. Sci. USA
Vol. 93, pp. 10864–10869, October 1996
Evolution

Accurate reconstruction of a known HIV-1 transmission history by phylogenetic tree analysis

THOMAS LEITNER*, DAVID ESCANILLA†, CHRISTER FRANZÉN‡, MATHIAS UHLÉN§, AND JAN ALBERT*¶

*Department of Clinical Virology, Swedish Institute for Infectious Disease Control, Karolinska Institute, S-105 21 Stockholm, Sweden; †Department of Microbiology, Faculty of Medicine, University of Chile, Independencia 1027, Santiago, Chile; ‡Department of Infectious Diseases, Kärnsjukhuset, S-541 85 Skövde, Sweden; and §Department of Biochemistry and Biotechnology, Royal Institute of Technology, S-100 44 Stockholm, Sweden

Communicated by Robert R. Sokal, State University of New York, Stony Brook, NY, July 9, 1996 (received for review February 21, 1996)

ABSTRACT **Phylogenetic analyses are increasingly used in attempts to clarify transmission patterns of human immunodeficiency virus type 1 (HIV-1), but there is a continuing discussion about their validity because convergent evolution and transmission of minor HIV variants may obscure epidemiological patterns. Here we have studied a unique HIV-1 transmission cluster consisting of nine infected individuals, for whom the time and direction of each virus transmission was exactly known. Most of the transmissions occurred between 1981 and 1983, and a total of 13 blood samples were obtained approximately 2–12 years later. The p17 *gag* and *env* V3 regions of the HIV-1 genome were directly sequenced from uncultured lymphocytes. A true phylogenetic tree was constructed based on the knowledge about when the transmissions had occurred and when the samples were obtained. This complex, known HIV-1 transmission history was compared with reconstructed molecular trees, which were calculated from the DNA sequences by several commonly used phylogenetic inference methods [Fitch–Margoliash, neighbor-joining, minimum-evolution, maximum-likelihood, maximum-parsimony, unweighted pair group method using arithmetic averages (UPGMA), and a Fitch-Margoliash method assuming a molecular clock (KITSCH)]. A majority of the reconstructed trees were good estimates of the true phylogeny; 12 of 13 taxa were correctly positioned in the most accurate trees. The choice of gene fragment was found to be more important than the choice of phylogenetic method and substitution model. However, methods that are sensitive to unequal rates of change performed more poorly (such as UPGMA and KITSCH, which assume a constant molecular clock). The rapidly evolving V3 fragment gave better reconstructions than p17, but a combined data set of both p17 and V3 performed best. The accuracy of the phylogenetic methods justifies their use in HIV-1 research and argues against convergent evolution and selective transmission of certain virus variants.**

Phylogenetic analyses are used in essentially all branches of biology; the applications range from studies on the origin of human populations (1) to investigations of important questions about the epidemiology and transmission patterns of human immunodeficiency virus type 1 (HIV-1). Several of the studies on HIV-1 have had direct legal implications, such as if a Florida dentist infected several of his patients (2), if a surgeon infected his patient (3), and if a Swedish rapist infected his victim (4). Other studies have tested hypotheses about the origin of HIV, as well as both global and local transmission patterns (5, 6). The lively debate about the results from some of these studies shows that there is no consensus about accuracy of phylogenetic analyses of HIV sequences (7, 8).

Several different phylogenetic inference methods have been developed (9–11), and their accuracy in reconstructing phylogenetic relationships has been examined by computer simulations (11–16) and studies of experimental phylogenetics (17). However, this type of study may not always accurately reflect the complex evolution of real organisms, because they are based on artificially generated data and oversimplified models of molecular evolution. Studies on real phylogenies have hitherto been lacking because detailed structures of relationships among taxa could not be regarded as known. Such relationships have often been inferred on fossil data, since the slow rate of evolution of higher organisms does not induce enough information to be collected in real time. In contrast, the genomes of RNA viruses, in particular HIV, evolve approximately one million times faster than the nucleic genomes of higher organisms (18), which makes them highly suitable for studying population histories. For the first time we have evaluated and compared the accuracy of several phylogenetic methods on DNA sequences from a known phylogeny of an organism (HIV-1) that has evolved under natural conditions. The known HIV-1 population history was derived from an HIV-1 transmission cluster for which all phylogenetic relationships were exactly known. The elapsed time between the first transmission and the last sampling was 13 years. As indicated above, this corresponds to approximately 13 million years of evolution in higher organisms, i.e., considerably more time than the age of the genus *Homo* (19). The accuracy of seven commonly used phylogenetic inference methods was examined using direct DNA population sequences from parts of the HIV-1 *gag* and *env* genes. Many of the molecular phylogenies were found to be accurate estimates of the true transmission history and, in general, the choice of gene fragment was found to be more important than the choice of tree building method.

MATERIALS AND METHODS

Study Population. A set of HIV-1 infected individuals with known epidemiological relationships (20) was genetically examined in this study. The index case, a Swedish male (p1) who became HIV-1 infected in Haiti in 1980, had eight sexual relations in which six females (p2, p4, p5, p7, p8, and p11) became infected (Fig. 1*A*). Also, two later male sexual partners (p6 and p10) and two children (p3 and p9) of the females became infected. Detailed interviews were done by doctors or nurses with special training in contact tracing. For all subjects it was possible to define a narrow time interval of a few months during which the transmission had occurred, for some subjects this interval could be further narrowed by records of a probable symptomatic primary HIV infection. No other risk factors were identified during the epidemiological investiga-

The publication costs of this article were defrayed in part by page charge payment. This article must therefore be hereby marked "*advertisement*" in accordance with 18 U.S.C. §1734 solely to indicate this fact.

Abbreviations: FM, Fitch-Margoliash; NJ, neighbor joining; ME, minimum evolution; ML, maximum likelihood; MP, maximum parsimony.

Data deposition: The sequences reported in this paper have been deposited in the GenBank data base (accession nos. U68496–U68521).

¶To whom reprint requests should be addressed.

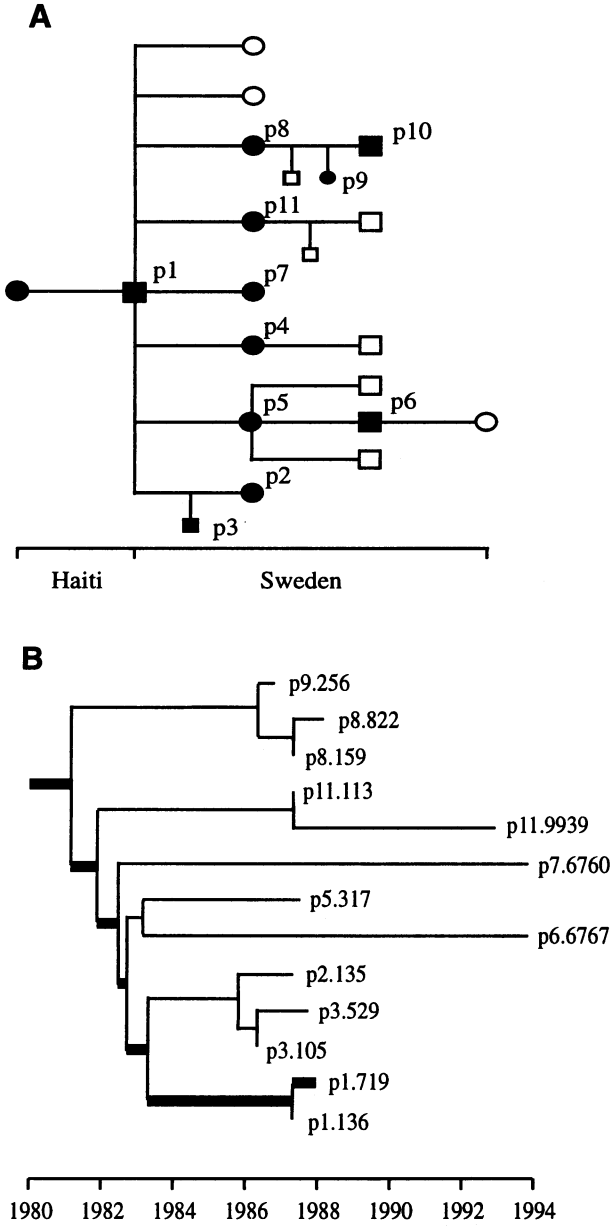

FIG. 1. The real time population history of HIV-1 in a Swedish transmission cluster. (*A*) Pattern of the contact tracing (20). Squares denote males and circles females, smaller symbols denote children. Solid and open symbols denote HIV-1 infected and uninfected persons, respectively. (*B*) The true tree, obtained by combining information about when the virus transmissions occurred and when the samples were obtained. Each ramification denotes a transmission event, where the virus population of the donor and recipient continues as the lower and upper lineages, respectively. The tip of each branch represents a sample, where patient and sample number is given. The thicker line conveys the path of the virus carried by the index case (p1); the path starts in early 1980 when the patient became infected, moves through 1981–1983 when five female sexual partners became infected, and ends with the samplings in 1987. Note that serial samples were obtained from several patients and that no samples were available from patients 4 and 10.

tion. Blood samples were obtained at different time points between 1986 and 1993 and peripheral blood mononuclear cells and plasma were isolated and stored in liquid nitrogen and −70°C, respectively, until analysis in 1994 and 1995. From some individuals, more than one sample was available.

Direct DNA Population Sequencing. The *env* V3 and p17 *gag* regions of the HIV-1 genome were directly sequenced from uncultured peripheral blood mononuclear cells as described (21, 22). Viral DNA corresponding to 100,000 peripheral blood mononuclear cells per reaction (sample p3.105r was derived from plasma RNA), was amplified by nested polymerase chain reaction (PCR) using primers specific for *env* V3 and p17 *gag* (22). To ensure that representative populations were amplified, we determined the viral load in each sample by limiting dilution and pooled the products from several PCRs before performing the sequencing reactions. The pooled PCR products were directly sequenced according to a solid-phase strategy (23). Briefly, the product from the inner PCR, where one primer was biotinylated, was bound to streptavidin-coated magnetic beads (Dynabeads, Dynal, Great Neck, NY). After purification and strand separation by NaOH, both strands were sequenced using fluorescein-labeled primers and T7 polymerase reactions. The Sanger fragments were subsequently detected on automated sequencing machines (A.L.F. DNA sequencer, Pharmacia BioTech). This method directly examines the virus population present in each individual sample, rather than individual clones separately (21). Polymorphic nucleotide positions were given International Union of Pure and Applied Chemistry (IUPAC) codes to indicate that individual members of the virus population differed in sequence at this position.

Phylogenetic Inference. After manual alignment of the DNA population sequences, molecular trees were constructed according to Fitch-Margoliash (FM) (24), neighbor-joining (NJ) (25), minimum-evolution (ME) (26), maximum-likelihood (ML) (27), maximum-parsimony (MP) (28), unweighted pair group method using arithmetic averages (UPGMA) (29), and a FM method assuming a molecular clock (KITSCH) (27). NJ, FM, ML, UPGMA, and KITSCH trees were constructed using the PHYLIP package (27). Substitution models for these programs were according to the Jukes–Cantor one parameter model (30), a modified Kimura two-parameter model (transition/transversion ratio set to 2.0) (K2′) (27, 31), and a more generalized model described by Felsenstein, using empirical nucleotide frequencies and a transition/transversion ratio of 2.0 (F84) (27). ML calculations were done with global rearrangements to find the best tree. Preliminary calculations by ML iterations indicated that the transition/transversion ratio was approximately 2 (T.L. and J.A., unpublished data). ME trees were calculated using the program package METREE (32) under the Jukes–Cantor substitution model where IUPAC ambiguity codes were deleted in each pair of sequences compared. MP trees were constructed using the program PAUP (33), with the branch-and-bound exact algorithm using uniform weighting (also called unweighted) and differential weighting, where the weighting is based on an asymmetric nucleotide substitution matrix calculated for each data set using the program MACCLADE (34). The MSTAXA option in PAUP was set to POLYMORPH during all MP calculations. Amino acid based trees were calculated from the translated nucleotide sequences, using the program PROTDIST under the PAM matrix model and the NJ method in PHYLIP.

Tree Analyses. The dissimilarity between the true tree (Fig. 1*B*) and all reconstructed molecular trees was investigated by subtree comparisons (35). The smallest informative subtree of an unrooted tree is a quartet, in our case with 13 sequences there will be 715 quartets of each tree. Dissimilarity measures can be calculated by comparing the quartets of each molecular tree to the quartets of the true tree. Tree-to-tree distances were also calculated by topological distance (partition metrics) (36). The calculations were performed using the program COMPONENT (37).

Bootstrap analysis (38) was performed to investigate whether low bootstrap values could also support the correct branching order. However, we have only investigated the most accurately reconstructed topology, derived from the combined p17+V3 data set. The NJ method using the K2′ model and 1000 resamplings was used for this purpose.

The phylogenetic signal was estimated by g_1 statistics. Here, 10,000 trees were randomly drawn from all possible trees of each data set. The shape of the frequency distribution of tree lengths can be described by g_1 statistics. A more left-skewed distribution (negative g_1 value) indicates a stronger phylogenetic signal. The critical value ($P = 0.01$) for 15 taxa with 100 and 500 characters is −0.20 and −0.15, respectively (39).

Consistency of information among individual parsimony informative sites in the true tree topology was investigated by average consistency indices (CI), average retention indices (RI), and a rescaled consistency index (RC). CI equals minimum possible tree length/observed tree length, RI equals (maximum possible tree length − actual tree length)/(maximum possible tree length − minimum possible tree length), and RC equals CI * RI. The range of CI, RI, and RC is 0–1, where a higher value indicates a higher degree of agreement between the characters in the data set.

The calculations of the mean pairwise proportion of differences (*p*-distance) included polymorphic positions. Although we have shown that it is possible to quantify the relative base frequencies (21), for this calculation the polymorphisms were assumed to be equally distributed among the bases that the IUPAC code described at the position; for example, R was assumed to be 50% A and 50% G.

RESULTS

A Known HIV-1 Phylogeny. This study involved nine individuals from a unique, well-characterized heterosexual HIV-1 transmission cluster (Fig. 1). The time point and direction for each transmission was exactly known (for details see *Materials and Methods*). The information about when the transmissions had occurred and when the samples were obtained was compiled into a tree, which describes the history of the transmitted virus populations (Fig. 1*B*). This is the first example of a known complex HIV-1 phylogeny. It should be noted that the tree spans more than a decade of HIV-1 evolution since the first transmission occurred in 1981 and the last sample was obtained in late 1993. The true tree was used as reference to investigate the accuracy of trees that were derived from population DNA sequences by several different phylogenetic inference methods.

Construction of Molecular Phylogenetic Trees. To test if the true phylogeny in Fig. 1*B* could be reconstructed by molecular phylogenetic methods we sequenced two regions (p17 *gag* and *env* V3) of the HIV-1 genome from 13 peripheral blood mononuclear cells samples from the nine individuals in the HIV-1 transmission cluster. The sequences represented the viral DNA populations present in uncultured peripheral blood mononuclear cells, because the samples were directly sequenced without cloning. All sequences were found to belong to HIV-1 genetic subtype B.

Molecular trees were constructed with the V3 and p17 data sets separately as well with a combined data set of both V3 and p17. In total, seven different tree building methods were applied (FM, NJ, ME, ML, MP, UPGMA, and KITSCH). The four best phylogenetic methods, as well as several different nucleotide substitution models, are presented in Fig. 2. The FM and NJ as well as the ML method gave one best tree estimate. The NJ method is a heuristic search for the ME tree, whereas the real ME tree is the result from an exhaustive search for that topology. Here, the results from both of these searches using the V3 and combined data sets were identical, and the p17 data set generated two alternative trees during the exhaustive search of which one was identical with the NJ topology and the other was less accurate compared with the true phylogeny (data not shown). It has been shown previously that the NJ tree is often identical with the ME tree unless the number of sequences is very large (26, 40). The MP method with uniform weighting gave several equally parsimonious trees (p17, $n = 33$; V3, $n = 4$, and combined p17+V3 fragment, $n = 3$). Therefore, 50% majority rule consensus trees were calculated and used in the subsequent analyses. By applying the differential weighting strategy both the V3 and the combined p17+V3 fragments gave a single most parsimonious tree, and the p17 fragment gave two very similar alternatives. In the latter case comparison to the true tree was conducted on both topologies, and an average of the values was computed.

Accuracy of the Molecular Trees. All of the estimated topologies from the molecular data were compared with the true tree topology by quartet subtree dissimilarity measures and topological distances. A majority of the molecular phylogenies were quite accurate estimates of the true phylogeny (Fig. 2). The analyses suggested that phylogenies obtained with V3 data were more accurate than those obtained with p17 data. However, trees obtained with a combined p17+V3 data set and the NJ, FM, and ML methods gave the most accurate estimates (Fig. 3). These trees contained only a single error, namely, the order of the samples from patients 8 and 9, a mother–child cluster (compare Fig. 1*B* with Fig. 2). Also, the two MP methods performed well; both made only two mistakes on the combined data set. Comparison of Figs. 2 and 3 reveals that the quartet dissimilarity measure sometimes differed between trees with the same number of misclassified branches, this is so because such misclassifications can be more or less severe depending on exactly where in the tree they are located. The two metric measures used to compare trees, quartet subtree dissimilarity and topological distance, showed similar results. The true tree showed large differences in branch lengths, a characteristic that has been reported to influence the relative performance of different methods (14). Therefore, methods that assume a constant molecular clock (such as UPGMA and KITSCH) performed poorly on our data sets (data not shown). In addition, all methods tended to overestimate the length of short branches and underestimate long branches. This error was similar for internal and external branches as well as for intra- and inter-patient relations. The problem appears to be due to the genetic variation within the virus population in each sample and will be thoroughly examined elsewhere (T.L. and J.A., unpublished data). The topologies of the inferred trees did not improve when distances were estimated by more complicated models of nucleotide substitution (Figs. 2 and 3).

Contrary to our findings it has been reported that p17 may be more suitable for phylogenetic analyses than V3 (5). Therefore, we compared the quality and quantity of the sequence data from these two regions in more detail (Table 1). The p17 and V3 fragments showed similar g_1 statistics and average consistency indices, indicating that the variable sites in these two regions were qualitatively similar. In fact, not even the V3 loop itself showed more indications of homoplasy and parallel evolution than the other gene fragments investigated. However, the V3 fragment carried more phylogenetic information than the longer p17 fragment, indicated by greater mean pair-wise distance, higher number of variable and parsimony informative sites. This indicates that the V3 fragment performed better than the p17 fragment because it evolves faster, which gives better opportunities to resolve events close in time. When the two fragments were combined the amount of information increased dramatically, and all investigated inference methods performed very well (Fig. 3). Bootstrap analysis (NJ, 1000 replicates) on the combined data set supported all branches with high values (77–100%), except three short internal branches (56.8%, 56.9% and 57.8%, respectively) (data not shown). Note also that the branches with low bootstrap values were correctly positioned in the trees. These low bootstrap values were due to the fact that the virus populations had not had time to accumulate more than a few mutations on these short branches. Trees generated from translated amino acid sequences were less accurate than those

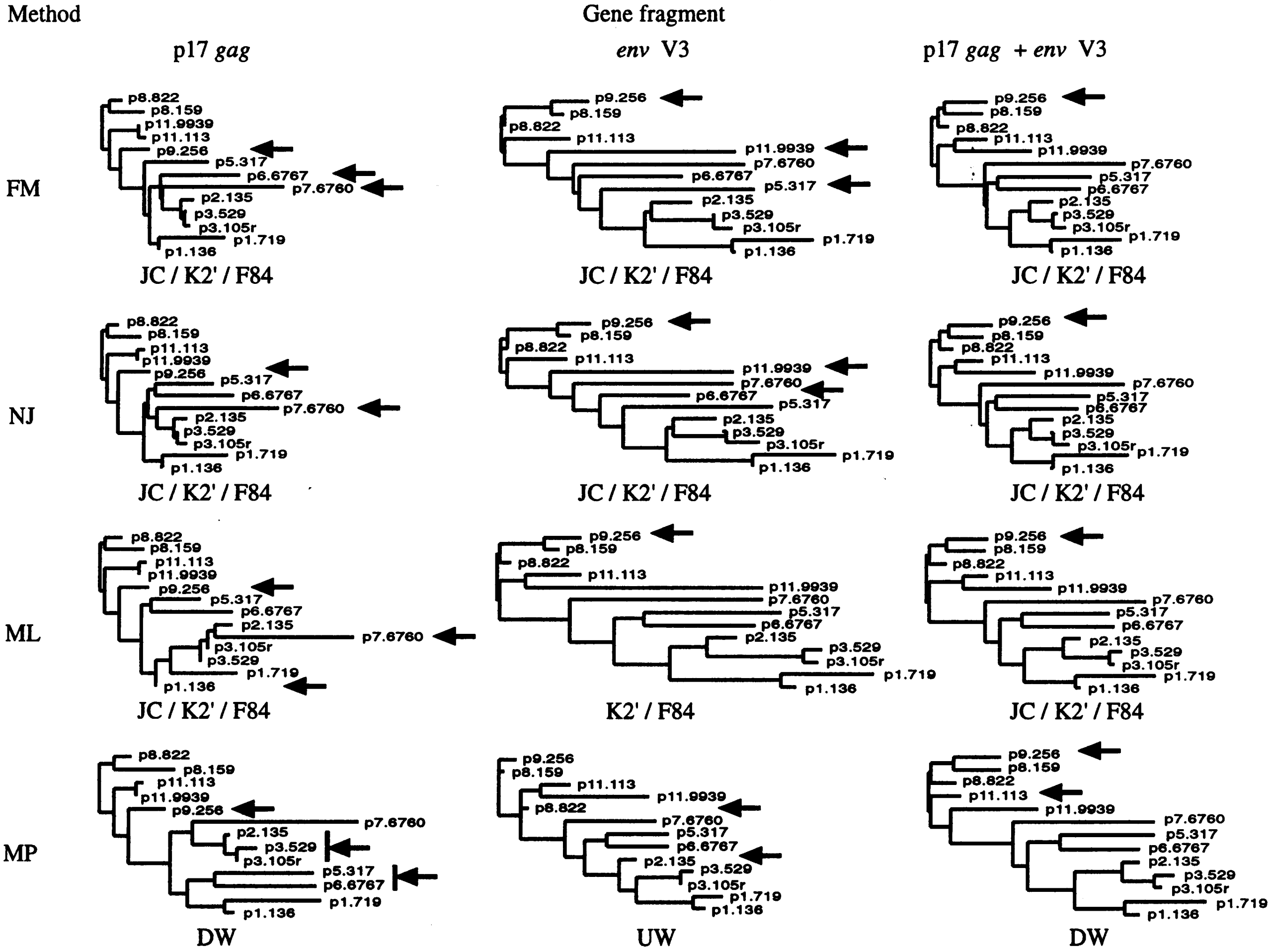

FIG. 2. Molecular trees constructed using the FM, NJ, ML, and MP methods as described. Substitution models for the FM, NJ, and ML methods were according to the Jukes–Cantor one parameter model, a modified Kimura two-parameter model (K2′), and a more generalized model (F84). The presented NJ trees had identical topology with trees generated under an exhaustive search for the ME tree. MP trees were constructed by uniform weighting (UW) and differential weighting (DW). Only the best trees for each method and gene are presented, and the substitution models used in these trees are indicated under each tree (compare with Fig. 3). Arrows identify misclassified tree units.

generated from nucleotide sequences and the order of performance was again p17 + V3 > V3 > p17. Thus, it was important to include synonymous positions to reach maximum resolution on our data set, where some transmissions occurred close in time.

DISCUSSION

In this study we have, to the best of our knowledge, for the first time used a known HIV-1 transmission history to evaluate and compare the accuracy of different phylogenetic methods. Most of these molecular trees were quite accurate estimates of the true phylogeny. The choice of phylogenetic method appeared to be less important than the choice of gene fragment. Thus, the V3 fragment gave better reconstructions than the p17 fragment, but a combined data set of both V3 and p17 performed better than any of the two alone.

The finding that the trees derived from V3 sequences were more accurate than those derived from p17 was surprising. The investigated V3 fragment includes the V3 loop, which is under strong positive selection for amino acid change (5) since it contains the principal neutralization domain and determinants for the biological phenotype. It has been reported that this may lead to convergent and parallel evolution (41, 42), which could make this region less suited for epidemiological inference (5, 8). By contrast, the p17 *gag* encodes the matrix protein, which is less exposed to host immune system. As a result, there is a selection against amino acid change (negative selection) in the p17 region. Based on these differences in selection pressure it has been argued that it would have been more appropriate to use p17 sequences, instead of V3 sequences, for the investigations of the Florida dentist case (8). It is interesting to note that the V3 fragment that was sequenced in the Florida dentist case is almost identical to the fragment that we have analyzed. Thus, our study shows that HIV-1 transmission patterns can be at least as accurately resolved using V3 sequences as with p17 sequences. Detailed analyses of our data set suggested that convergent evolution (41, 43) does not significantly influence the results of phylogenetic analyses of V3 sequences. Even if selection for similar mutations in unrelated samples may occur on certain nucleotide positions in V3, this appears to give very little noise in the phylogenetic signal. Thus, the results from this study justify the use of phylogenetic methods for reconstructing HIV-1 transmission patterns using both p17 and V3 population sequences. However, for maximum resolution, it may be necessary to collect more sequence data; in our study this was achieved by combining the p17 *gag* and V3 data. It should also be stressed that more slowly evolving genes may give more accurate results in studies of distantly related viral species, since the variation in many nucleotide positions would become saturated in fast evolving genes.

The accuracy of the molecular phylogenies have important implications for the understanding of HIV-1 evolution and transmission. First, HIV-1 populations within an infected

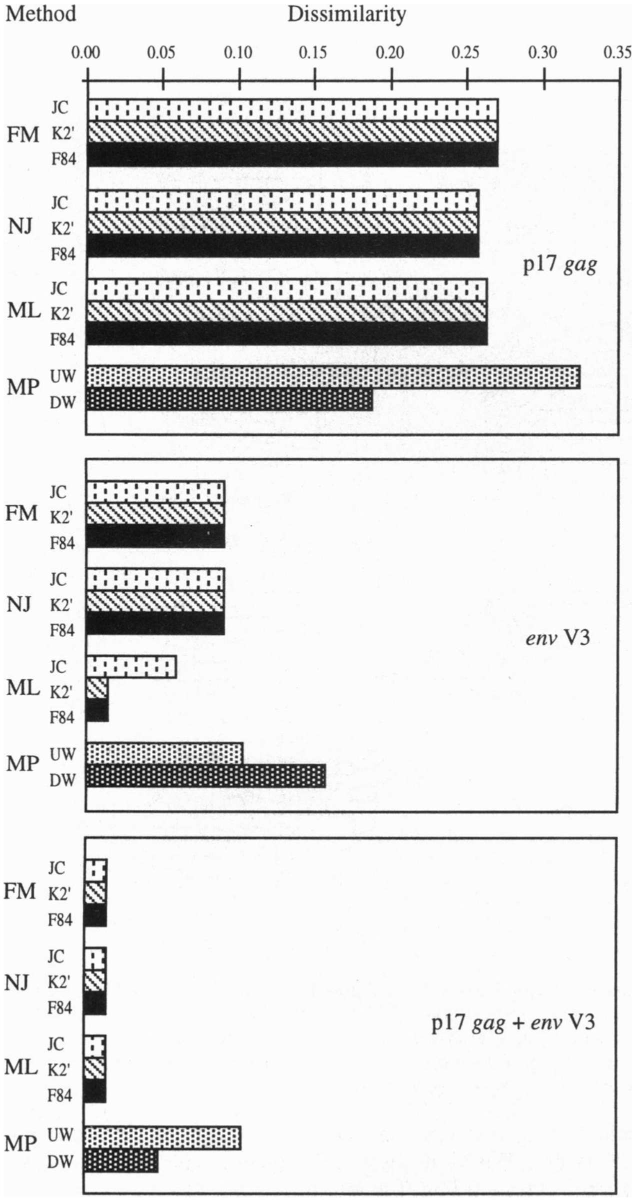

FIG. 3. The dissimilarity between the true tree (Fig. 1*B*) and molecular reconstructed trees as investigated by quartet subtree comparisons, as described. The bars present the proportion of quartets in the reconstructed trees, which explicitly disagree with those of the true tree. Abbreviations as in Fig. 2.

individual appear to be continuously evolving, otherwise it would not have been possible to correctly order the transmissions from the index case to the five female recipients since they all occurred within a 2-year time period. Second, our results argue against a selective transmission of minor variants (42, 44). Our chances to reconstruct the true phylogeny would have been minimal if such a selection had occurred. Instead, our findings suggest that the transmitted virus variant(s) shared many genetic features with the dominant population of the transmitter and that this population had undergone clearly distinguishable changes until the next transmission. Third, our results also argue against the possibility that new infections can only be established by a limited subset of V3 variants as suggested by Zhang *et al.* (42). If this was the case, p1 would have been expected to transmit the same V3 variant to all five female recipients.

In this study we analyzed polymorphic population sequences (21), instead of multiple clones separately. This may positively have influenced the accuracy of the reconstructed phylogenies. First, the possibility that one "odd" clonal sequence would disturb the inferred pattern is removed, because each population is considered as one tree unit. Second, the number of sequences that are analyzed is greatly reduced, which makes it possible to use more sophisticated phylogenetic methods and exhaustive search algorithms. Already with 13 sequences there will be ≈13.7 billions possible tree topologies. If 10 clones from each sample were sequenced this would give 130 sequences and an astronomical number of possible topologies, which no presently available computer can examine. Furthermore, selection during cloning procedures and amplification of *Taq* polymerase errors are avoided. These problems may otherwise bias the estimation of the population structure. Direct population sequencing is also quicker, which gives opportunities to sequence several genes instead of many clones of one fragment in the same amount of time.

The evolution of the HIV-1 genome displays a number of features that may cause problems in phylogenetic reconstruction, such as recombination, selection, differences in evolutionary rates among lineages, unbalanced nucleotide frequencies, and differences in nucleotide substitution rates (5, 42, 45, 46). However, our complex population history was accurately inferred, indicating that the methods for phylogenetic reconstruction are robust. For instance, patients 7 and 5, who became infected with only a few months interval and from whom viral sequences were obtained 10 years later, were correctly placed in many of the inferred trees. Computer simulations have earlier shown that the relative efficiencies of various methods depend on the number of nucleotides in the analyzed fragment (11, 12) and on varying rate of substitution on different branches (14). The conclusions from such computer simulations are difficult to translate into real situations, and depending on the experimental setup they have sometimes suggested NJ (15), MP (12), or ML (16) to be superior, whereas under other conditions most methods work well. From these analyses it is clear that the methods have advantages and disadvantages in different situations. We found that for HIV-1, using our data set that describes populations which have undergone natural evolution, the decisive factor in the analysis was the quantity of genetic information. All of the different tree building methods performed better when more information was given.

In conclusion, by using population sequences and combining the V3 and p17 fragments, we obtained excellent estimates of

Table 1. Characteristics for different HIV-1 *gag* and *env* gene fragments

Gene fragment	n_{tot}	n_{pinf}	n_{var}	δ (range)	g_1	CI	RI	RC
p17 *gag*	438	21	64	3.7 (0.3–6.6)	−0.62	0.80	0.68	0.55
env V3	285	45	115	10.2 (3.6–13.9)	−0.65	0.85	0.65	0.55
V3 loop	108	16	46	10.4 (2.7–15.3)	−0.68	0.85	0.67	0.56
V3 flanking	177	29	69	10.0 (4.0–16.4)	−0.68	0.85	0.62	0.53
p17 + V3	723	66	179	6.2 (1.6–9.5)	−0.63	0.84	0.66	0.56

n_{tot}, Total number of nucleotide positions. n_{pinf}, Phylogenetically informative sites for parsimony in the true tree. n_{var}, Number of variable nucleotide positions. δ, Mean pairwise proportion of differences (*p*-distance) in percent. g_1, A parsimony-based estimate of phylogenetic signal. CI, Average consistency indices. RI, Average retention indices. RC, Rescaled consistency index.

the true transmission history with common tree building methods, excluding methods that assume a constant molecular clock. This shows that phylogenetic methods can be used to answer important and complex questions about HIV-1 transmission patterns.

We thank G. Lidin-Janson for invaluable help with clinical information; E. Holmes, S. Kumar, D. Hillis, and J. Bull for valuable suggestions; and M. Axelsson and E. Lilja for expert laboratory assistance. Our work was supported by grants from the Swedish Medical Research Council, the Swedish Agency for Research Cooperation with Developing Countries, and the Swedish National Board for Industrial and Technical Development.

1. Cann, R. L., Stoneking, M. & Wilson, A. C. (1987) *Nature (London)* **325,** 31–36.
2. Ou, C. Y., Ciesielski, C. A., Myers, G., Bandea, C. I., Luo, C. C., Korber, B. T., Mullins, J. I., Schochetman, G., Berkelman, R. L., Economou, A. N., Witte, J. J., Furman, L. J., Satten, G. A., MacInnes, K. A., Curran, J. W., Jaffe, H. W. & Laboratory Investigation Group and Epidemiologic Investigation Group (1992) *Science* **256,** 1155–1161.
3. Holmes, E. C., Zhang, L. Q., Simmonds, P., Rogers, A. S. & Brown, A. J. (1993) *J. Infect. Dis.* **167,** 1411–1414.
4. Albert, J., Wahlberg, J., Leitner, T., Escanilla, D. & Uhlén, M. (1994) *J. Virol.* **68,** 5918–5924.
5. Holmes, E. C., Zhang, L. Q., Robertson, P., Cleland, A., Harvey, E., Simmonds, P. & Leigh-Brown, A. J. (1995) *J. Infect. Dis.* **171,** 45–53.
6. Myers, G., Korber, B., Wain-Hobson, S., Jeang, K.-T., Henderson, L. E. & Pavlakis, G. N. (1994) *Human Retroviruses and AIDS* (Los Alamos Natl. Lab., Los Alamos, NM).
7. DeBry, R. W., Abele, L. G., Weiss, S. H., Hill, M. D., Bouzas, M., Lorenzo, E., Graebnitz, F. & Resnick, L. (1993) *Nature (London)* **361,** 691.
8. Holmes, E. C., Leigh Brown, A. J. & Simmonds, P. (1993) *Nature (London)* **364,** 766.
9. Felsenstein, J. (1988) *Annu. Rev. Genet.* **22,** 521–565.
10. Swofford, D. L. & Olsen, G. J. (1990) in *Molecular Systematics,* eds. Hillis, D. M. & Moritz, C. (Sinauer, Sunderland, MA), pp. 411–501.
11. Nei, M. (1991) in *Phylogenetic Analysis of DNA Sequences,* eds. Miyamoto, M. M. & Cracraft, J. (Oxford Univ. Press, New York), pp. 90–128.
12. Hillis, D. M., Huelsenbeck, J. P. & Cunningham, C. W. (1994) *Science* **264,** 671–677.
13. Nei, M., Takezaki, N. & Sitnikova, T. (1995) *Science* **267,** 253–255.
14. Sourdis, J. & Nei, M. (1988) *Mol. Biol. Evol.* **5,** 298–311.
15. Saitou, N. & Imanishi, T. (1989) *Mol. Biol. Evol.* **6,** 514–525.
16. Saitou, N. (1988) *J. Mol. Evol.* **27,** 261–273.
17. Hillis, D. M., Bull, J. J., White, M. E., Badgett, M. R. & Molineux, I. J. (1992) *Science* **255,** 589–592.
18. Gojobori, T., Moriyama, E. N. & Kimura, M. (1990) *Proc. Natl. Acad. Sci. USA* **87,** 10015–10018.
19. Schreck, F., Bromage, T. G., Betzler, C. G., Ring, U. & Juwayeyi, Y. M. (1993) *Nature (London)* **365,** 833–836.
20. Franzén, C., Albert, J., Biberfeld, G., Lidin-Janson, G. & Löwhagen, G. B. (1988) Fourth International Conference on AIDS (Int. Conf. on AIDS, Stockholm), abstr. 4021.
21. Leitner, T., Halapi, E., Scarlatti, G., Rossi, P., Albert, J., Fenyö, E. M. & Uhlén, M. (1993) *BioTechniques* **15,** 120–126.
22. Leitner, T., Escanilla, D., Marquina, S., Wahlberg, J., Broström, C., Hansson, H. B., Uhlén, M. & Albert, J. (1995) *Virology* **209,** 136–146.
23. Hultman, T., Bergh, S., Moks, T. & Uhlén, M. (1991) *BioTechniques* **10,** 84–93.
24. Fitch, W. M. & Margoliash, E. (1967) *Science* **155,** 279–284.
25. Saitou, N. & Nei, M. (1987) *Mol. Biol. Evol.* **4,** 406–425.
26. Rzhetsky, A. & Nei, M. (1993) *Mol. Biol. Evol.* **10,** 1073–1095.
27. Felsenstein, J. (1993) PHYLIP: Phylogeny Inference Package (Univ. of Washington, Seattle), Version 3.52c.
28. Eck, R. V. & Dayhoff, M. O. (1966) *Atlas of Protein Sequence and Structure* (Natl. Biomed. Res. Found., Silver Spring, MD).
29. Sneath, P. H. A. & Sokal, R. R. (1973) *Numerical Taxonomy* (Freeman, San Francisco).
30. Jukes, T. H. & Cantor, C. R. (1969) in *Mammalian Protein Metabolism,* ed. Munro, H. N. (Academic, New York), pp. 21–132.
31. Kimura, M. (1980) *J. Mol. Evol.* **16,** 111–120.
32. Rzhetsky, A. & Nei, M. (1994) METREE: Program Package for Inferring and Testing Minimum Evolution Trees (Inst. Mol. Evol. Genet., Pennsylvania State Univ., PA), Version 1.4.
33. Swofford, D. L. (1991) PAUP: Phylogenetic Analysis Using Parsimony (Ill. Natl. Hist. Surv., Champaign, IL), Version 3.1.1
34. Maddison, W. P. & Maddison, D. R. (1992) MACCLADE: Analysis of Phylogeny and Character Evolution (Sinauer, Sunderland, MA), Version 3.04.
35. Estabrook, G. F., McMorris, F. R. & Meacham, C. A. (1985) *Syst. Zool.* **34,** 193–200.
36. Robinson, D. F. & Foulds, L. R. (1991) *Math. Biosci.* **53,** 105–147.
37. Page, R. D. M. (1993) COMPONENT (Biogeography and Conservation Lab., Natural History Museum, London), Version 2.0.
38. Felsenstein, J. (1985) *Evolution* **39,** 783–791.
39. Hills, D. M. & Huelsenbeck, J. P. (1992) *J. Hered.* **83,** 189–195.
40. Rzhetsky, A. & Nei, M. (1992) *Mol. Biol. Evol.* **9,** 945–967.
41. Strunnikova, N., Ray, S. C., Livingston, R. A., Rubalcaba, E. & Viscidi, R. P. (1995) *J. Virol.* **69,** 7548–7558.
42. Zhang, L. Q., MacKenzie, P., Cleland, A., Holmes, E. C., Brown, A. J. & Simmonds, P. (1993) *J. Virol.* **67,** 3345–3356.
43. Doolittle, R. F. (1994) *Trends Biochem. Sci.* **19,** 15–18.
44. Wolinsky, S. M., Wike, C. M., Korber, B. T., Hutto, C., Parks, W. P., Rosenblum, L. L., Kunstman, K. J., Furtado, M. R. & Munoz, J. L. (1992) *Science* **255,** 1134–1137.
45. Robertson, D. L., Sharp, P. M., McCutchan, F. E. & Hahn, B. H. (1995) *Nature (London)* **374,** 124–126.
46. Vartanian, J., Meyerhans, A., Sala, M. & Wain-Hobson, S. (1994) *Proc. Natl. Acad. Sci. USA* **91,** 3092–3096.