Multiple Interspecies Transmissions of Human and Simian T-Cell Leukemia/Lymphoma Virus Type I Sequences

Keith A. Crandall

Department of Zoology, University of Texas at Austin

Using two sets of nucleotide sequences of the human and simian T-cell leukemia/lymphoma virus type I (HTLV-I/STLV-I), one consisting of 522 bp of the *env* gene from 70 viral strains and the other a 140-bp segment from the *pol* gene of 52 viral strains, I estimated cladograms based on a statistical parsimony procedure that was developed specifically to estimate within-species gene trees. An extension of a nesting procedure is offered for sequence data that forms nested clades used in hypothesis testing. The nested clades were used to test three hypotheses relating to transmission of HTLV/STLV sequences: (1) Have cross-species transmissions occurred and, if so, how many? (2) In what direction have they occurred? (3) What are the geographic relationships of these transmission events? The analyses support a range of 11–16 cross-species transmissions throughout the history of these sequences. Additionally, outgroup weights were assigned to haplotypes using arguments from coalescence theory to infer directionality of transmission events. Conclusions on geographic origins of transmission events and particular viral strains are inconclusive due to small samples and inadequate sampling design. Finally, this approach is compared directly to results obtained from a traditional maximum parsimony approach and found to be superior at establishing relationships and identifying instances of transmission.

Introduction

A phylogenetically close group of retroviruses, known as primate T-cell lymphoma/leukemia viruses (PTLV), has been the subject of many recent studies in virology and human epidemiology because of their role as etiologic agents in numerous diseases in both humans (Hinuma et al. 1981; Kalyanraman et al. 1982; Zucker-Franklin, Hooper, and Eratt 1992) and other primates (Lee et al. 1985; Sakakibara et al. 1986; Schatzl et al. 1992). This group of retroviruses shares a common ancestor with the bovine leukemia virus (BLV) (Dube et al. 1994) and is more distantly related to the lentivirus group and human immunodeficiency viruses (HIVs) (Yokoyama, Chung, and Gojobori 1988). Within the PTLV group, three viral types have been described; the human T-cell leukemia/lymphoma virus type I (HTLV-I) (Poiesz et al. 1980), the human T-cell leukemia/lymphoma virus type II (HTLV-II) (Kalyanraman et al. 1982), and the simian T-cell leukemia/lymphoma virus type I (STLV-I) (Komuro et al. 1984; Watanabe et al. 1985). The HTLV-II viruses appear to be phylogenetically distinct from the HTLV-I/STLV-I viruses (Dube et al. 1994; Saksena et al. 1994). Both HTLV-I and HTLV-II are found throughout the world as a result of transmission via blood transfusions, intravenous drug abuse, and sexual intercourse (Dube et al. 1994). Likewise, it is suspected that transmission within and possibly among primate species occurs via sexual intercourse as many of these species are able to interbreed (Koralnik et al. 1994). Phylogenetic analyses have revealed no well-supported relationship between the STLV-I and HTLV-I viruses (Saksena et al. 1993, 1994; Dube et al. 1994; Koralnik et al. 1994). An accurate understanding of the phylogenetic relationships among these viral types is important in the development of therapies and vaccine strategies to combat the many diseases associated with these viruses as well as to understand the epidemiology and pathogenesis of these diseases.

The lack of resolution among STLV-I and HTLV-I sequences is due in part to their similarity relative to, for example, HIV sequences (Li, Tanimura, and Sharp 1988; Gessain, Gallo, and Franchini 1992; Kuiken and Korber 1994; e.g., the range of percent sequence divergence relating any pair of haplotypes for the *env* data set in this study is [0.00%, 11.0%] with an average of 0.6%). Traditional phylogeny reconstruction methods (e.g., parsimony, neighbor-joining, and maximum-likelihood) typically have greater statistical power when sequences are more divergent (Huelsenbeck and Hillis 1993). Thus, when divergences among sequences are low, little resolution or support is achieved for phylogenetic relationships. Additionally, various research groups have presented phylogenetic analyses that imply cross-species transmission of the STLV-I and HTLV-I viruses among primates including humans, because virus samples from different host species clustered together in single clades (Saksena et al. 1993; Koralnik et al. 1994). However, the robustness of these clusterings was not tested using statistical procedures. Low resolution

Abbreviations: HTLV-I, human T-cell leukemia/lymphoma virus type I; STLV-I, simian T-cell leukemia/lymphoma virus type I; PTLV, primate T-cell leukemia/lymphoma virus; OTU, operational taxonomic unit; HIV, human immunodeficiency virus; ORF, open reading frame.

Key words: HTLV, STLV, cladogram, parsimony, viral transmission, retrovirus.

Address for correspondence and reprints: Keith A. Crandall, Department of Zoology, University of Texas at Austin, Austin, Texas 78712-1064; e-mail: crandall@phylo.zo.utexas.edu.

Mol. Biol. Evol. 13(1):115–131. 1996

Table 1
70 HTLV and STLV Sequences, Haplotype Designations, and Geographical Origin for *env* Data

Species	Haplotype Designation	Geographical Origin	Accession Number	Reference[a]
Homo sapiens	Hs1	West Africa	U03154	1
H. sapiens	Hs2	West Africa	U03133	1
H. sapiens	Hs3	West Africa	U03134	1
H. sapiens	Hs4	West Africa	U03136	1
H. sapiens	Hs5	West Africa	U03138	1
H. sapiens	Hs6	West Africa	U03135	1
H. sapiens	Hs7	Central Africa	U03142	1
H. sapiens	Hs8	Central Africa	U03141	1
H. sapiens	Hs8	Central Africa	U03140	1
H. sapiens	Hs9	Central Africa	U03139	1
H. sapiens	Hs10	Zaire	M67514	13
H. sapiens	Hs11	Papua New Guinea	L02533	5
H. sapiens	Hs12	Melanesia	M73745	4
H. sapiens	Hs12	Papua New Guinea	M94196	4
H. sapiens	Hs13	Papua New Guinea	M94197	4
H. sapiens	Hs13	Papua New Guinea	U11576	7
H. sapiens	Hs14	Solomon Islands	M94198	4
H. sapiens	Hs15	Solomon Islands	M94199	4
H. sapiens	Hs15	Solomon Islands	M93099	4
H. sapiens	Hs15	Solomon Islands	U11566	7
H. sapiens	Hs15	Solomon Islands	U11580	7
H. sapiens	Hs16	Solomon Islands	M94200	4
H. sapiens	Hs17	Solomon Islands	L02534	5
H. sapiens	Hs18	Solomon Islands	U11568	7
H. sapiens	Hs19	Solomon Islands	U11570	7
H. sapiens	Hs20	Solomon Islands	U11578	7
H. sapiens	Hs21	Melanesia	M94195	4
H. sapiens	Hs22	Melanesia	M93098	4
H. sapiens	Hs23	La Reunion Island	Z28963	10
H. sapiens	Hs23	La Reunion Island	Z28965	10
H. sapiens	Hs24	La Reunion Island	Z28964	10
H. sapiens	Hs24	La Reunion Island	Z28967	10
H. sapiens	Hs25	La Reunion Island	Z28966	10
H. sapiens	Hs26	Australia	M92818	12
H. sapiens	Hs27	Asia	J02029	2
H. sapiens	Hs28	Japan	M86840	9
H. sapiens	Hs29	West Indies	U03155	1
H. sapiens	Hs30	West Indies	D13784	3
H. sapiens	Hs31	Martinique	S92512	6
H. sapiens	Hs32	North America	M67490	13
H. sapiens	Hs33	North America	M69044	13
H. sapiens	Hs34	South America	U03153	1
H. sapiens	Hs35	Brasil	U11559	7
H. sapiens	Hs36	?	L03561	8
H. sapiens	Hs36	?	L03562	8
H. sapiens	Hs36	Japan	M37747	11
Ceropithecus aethiops	Cae1	West Africa	U03132	1
C. aethiops	Cae2	West Africa	U03130	1
C. aethiops	Cae3	West Africa	U03129	1
C. aethiops	Cae4	West Africa	U03131	1
C. aethiops	Cae5	East Africa	U03127	1
C. aethiops	Cae6	East Africa	U03128	1
C. aethiops	Cae7	East Africa	U03126	1
C. aethiops	Cae8	East Africa	U03122	1
C. ascanius	Cas1	East Africa	U03150	1
C. mitis	Cm1	East Africa	U03151	1
C. mitis	Cm2	Central Africa	U03152	1
Macaca mulatta	Mm1	India	U03156	1
M. mulatta	Mm2	India	U03144	1
M. fascicularis	Mf1	Indonesia	U03143	1
Papio cynocephalus	Pc1	Russia	U03145	1
P. cynocephalus	Pc2	East Africa	U03158	1
P. anubis	Pa1	East Africa	U03157	1
P. hamadryas	Ph1	East Africa	U03159	1
P. papio	Pp1	West Africa	U03160	1
Pan troglodytes	Pt1	East Africa	U03124	1
P. troglodytes	Pt2	East Africa	U03149	1
P. troglodytes	Pt3	?	U03148	1
P. troglodytes	Pt4	?	U03147	1
P. troglodytes	Pt5	?	U03146	1

[a] 1 = Koralnik et al. (1994), 2 = Seiki et al. (1983), 3 = Malik, Even, and Karpas (1988), 4 = Gessain et al. (1991), 5 = Gessian et al. (1993), 6 = Gessian, Gallo, and Franchini (1992), 7 = Nerurkar et al. (1994), 8 = Zhao et al. (1993), 9 = Evangelista et al. (1990), 10 = Mahieux et al. (1994), 11 = Gray et al. (1990), 12 = Bastian et al. (1993), 13 = Paine et al. (1991).

and low bootstrap values are characteristic of analyses of HTLV-I and STLV-I sequences (Saksena et al. 1993; Koralnik et al. 1994). Furthermore, previous studies have not explicitly tested a transmission hypothesis versus a null hypothesis of no transmission. In this paper I present an analysis that (1) demonstrates the utility of a statistical parsimony procedure to estimate robust cladograms when few nucleotide differences exist among haplotypes, (2) assigns outgroup weights to haplotypes to allow for hypothesis testing of transmission directionality and geographic origin, (3) utilizes the resulting cladogram as a nested statistical design to identify cases of cross-species transmission and geographic correlation among individuals, and (4) compares these results to a standard analysis using maximum parsimony trees. The analyses are based on nucleotide sequences from two gene regions, *pol* and *env,* from both HTLV-I and STLV-I. Additionally, extensions of outgroup weight calculations (Castelloe and Templeton 1994) and cladogram nesting procedures (Templeton and Sing 1993) are offered for nucleotide sequence data.

Table 2
52 HTLV and STLV Sequences, Haplotype Designations, and Geographical Origin for *pol* Data

Species	Haplotype Designation	Geographical Origin	Accession Number	Reference[a]
Homo sapiens	Hs1	West Africa	M76751	6
H. sapiens	Hs1	Japan	M86840	10
H. sapiens	Hs2	USA	M99084	4
H. sapiens	Hs3	New Guinea	M76755	6
H. sapiens	Hs4	USA	L20639	1
H. sapiens	Hs4	USA	M99083	4
H. sapiens	Hs4	USA	M99085	4
H. sapiens	Hs4	USA	M99086	4
H. sapiens	Hs4	USA	M99087	4
H. sapiens	Hs4	USA	M99088	4
H. sapiens	Hs4	Japan	J02029	11
H. sapiens	Hs4	Japan	L20641	1
H. sapiens	Hs4	?	L03561	5
H. sapiens	Hs4	?	L03562	5
H. sapiens	Hs4	India	S67866	9
H. sapiens	Hs4	Bellona	M99082	4
H. sapiens	Hs4	Bellona	U12105	14
H. sapiens	Hs4	Brazil	U12108	14
H. sapiens	Hs4	India	U12114	14
H. sapiens	Hs5	Central Africa	S74562	8
H. sapiens	Hs5	Central Africa	L27569	3
H. sapiens	Hs5	Central Africa	L27563	3
H. sapiens	Hs5	Central Africa	L27570	3
H. sapiens	Hs6	Caribbean	D13784	12
H. sapiens	Hs7	Melanesia	L02534	7
H. sapiens	Hs8	USA	L27561	3
H. sapiens	Hs9	Central Africa	L27564	3
Pan troglodytes	Hs9/Pt1	Sierra Leone	U12117	14
H. sapiens	Hs10	Central Africa	L27565	3
H. sapiens	Hs10	Central Africa	L27568	3
H. sapiens	Hs11	Central Africa	L27566	3
H. sapiens	Hs12	Central Africa	L27571	3
H. sapiens	Hs13	India	U12111	14
H. sapiens	Hs14	Australia	U12120	14
Cercopithecus aethiops	Cae1	Central Africa	M92846	13
C. aethiops	Cae2	West Africa	L20355	2
C. aethiops	Cae2	West Africa	L20356	2
C. aethiops	Cae2	West Africa	L20358	2
C. aethiops	Cae2	West Africa	L20361	2
C. aethiops	Cae2	West Africa	L20359	2
C. aethiops	Cae2	West Africa	L20360	2
C. aethiops	Cae3	West Africa	L20357	2
C. aethiops	Cae4	Kenya	U12102	14
Cercopithecus mona	Cm1	Central Africa	L20352	2
Macaca mulatta	Mm1	Asia	L20660	1
M. fuscata	Mf1	Japan	L20646	1
M. fuscata	Mf2	Japan	L20656	1
M. fuscata	Mf2	Japan	L20665	1
Papio cynocephalus	Pc1	Southern Africa	L20651	1
P. doguera	Pd1	Central Africa	L20351	2
Erythrocebus patas	Ep1	West Africa	L20353	2
E. patas	Ep2	Central Africa	L20354	2

[a] 1 = Song et al. (1994), 2 = Saksena et al. (1994), 3 = Dube et al. (1994), 4 = Dube et al. (1993), 5 = Zhao et al. (1993), 6 = Sherman et al. (1992), 7 = Gessian et al. (1993), 8 = Ratner, Philpott, and Trowbridge (1991), 9 = Nerurkar et al. (1993), 10 = Evangelista et al. (1990), 11 = Seiki et al. (1983), 12 = Malik, Even, and Karpas (1988), 13 = Saksena et al. (1993), 14 = Yanagihara et al. (1995).

Materials and Methods

DNA Sequences

The HTLV/STLV genome contains four described open reading frames (ORFs) flanked by long terminal repeats (LTRs) that typically characterize retroviruses (Gessain et al. 1993). ORF I and II encode the three main genes associated with retroviruses; *gag, pol,* and *env*. Seventy nucleotide sequences from a 522-bp segment of the *env* protein from PTLV-Is were obtained from GenBank (table 1). These sequences correspond to nucleotide positions 6,046–6,567 of the HTLV-I prototype sequence (Seiki et al. 1983) and represent 60 distinct haplotypes. In addition, 52 nucleotide sequences representing 26 distinct haplotypes from a 140-bp segment of the *pol* gene were retrieved from GenBank (table 2). This region of the *pol* gene corresponds to nucleotide positions 4,779–4,918 of the HTLV-I prototype sequence. These two gene regions are the only segments available for a large number of individuals in geographically diverse areas. Therefore, they afford the best opportunity to test hypotheses of cross-species transmission and geographic origins. In only six cases were both gene sequences from the same isolate; therefore, these regions were treated as independent data sets and analyzed independently. Sequences were aligned using CLUSTAL V (Higgins, Bleasby, and Fuchs 1992).

Cladogram Estimation

Phylogenetic relationships among HTLV-I and STLV-I sequences were estimated using a statistical parsimony procedure (Templeton, Crandall, and Sing 1992). This method, developed specifically to reconstruct within-species gene trees, sets a statistical criterion for the limits of the parsimony assumption; that is, the probability that a nucleotide difference between a specific pair of sequences is due to a single substitution (the parsimonious state) and not the result of multiple substitutions at a single site (the nonparsimonious state). Thus, I use the term parsimony to refer to the minimum number of differences separating two individual sequences rather than a global minimum tree length based on shared derived characters. The probability that a nucleotide difference between two haplotypes is due to one and only one substitution increases as the number of shared sites between sequences increases. This procedure estimates a set of probable relationships between haplotypes whose cumulative probability is ≥0.95.

The model assumes independence of sites and allows for biases in substitutional patterns of nucleotide changes (Templeton, Crandall, and Sing 1992). The allowance for different mutational biases between each pair of haplotypes being examined results in a lack of a requirement that the mutations be identically distributed, a typical assumption for many reconstruction algorithms. Likewise, the testing for multiple substitutions between a pair of haplotypes assures (at a 95% confidence level) that the infinite alleles model is not violated by the established relationships. This is important when utilizing results from coalescence theory to refine cladogram probabilities and assign outgroup probabilities (Crandall and Templeton 1993; Castelloe and Templeton 1994; Crandall, Templeton, and Sing 1994; see below). Future work will explore further the robustness of this method to violations of these assumptions.

Traditional methods of phylogeny reconstruction were developed to estimate relationships of higher taxonomic groups, e.g., species, genera, families, etc. Consequently, these methods make a number of assumptions that are invalid at the population genetic level (Crandall, Templeton, and Sing 1994). For example, species trees are traditionally regarded as strictly bifurcating. However, in populations, most haplotypes in the gene pool exist as sets of multiple, identical copies because of past DNA replication. In haplotype trees, each gene lineage of the identical copies of a single haplotype are at risk for independent mutation. Consequently, coalescence theory predicts that a single ancestral haplotype will often give rise to multiple, descendant haplotypes, thereby yielding a haplotype tree with multifurcations. Additionally, gene regions examined in populations can undergo recombination. Traditional methods assume recombination does not occur in the region under examination. Furthermore, recombination is an additional reason why the assumption of a strictly bifurcating tree topology is likely to be violated.

Additional differences exist between gene trees at the population level and higher taxa divergences (Pamilo and Nei 1988). Populations typically have lower levels of variation over a given gene region relative to higher taxonomic levels, resulting in fewer characters for phylogenetic analyses. Huelsenbeck and Hillis (1993) have shown that interspecific methods for phylogeny reconstruction perform poorly when few characters are available for analysis. Another difference concerns the treatment of ancestral types. In populations, when one copy of a haplotype in the gene pool mutates to form a new haplotype, it would be extremely unlikely for all the identical copies of the ancestral haplotype to also mutate. Thus, as mutations occur to create new haplotypes, they rarely result in the extinction of the ancestral haplotype. The ancestral haplotypes are thereby expected to persist in the population. Indeed, coalescence theory predicts that the most common haplotypes in a gene pool will tend to be the oldest (Watterson and Guess 1977; Donnelly and Tavaré 1986), and most of these old haplotypes will be interior nodes of the haplotype tree (Crandall and Templeton 1993; Castelloe and Templeton 1994).

The statistical power of the cladogram estimation procedure of Templeton, Crandall, and Sing (1992) is achieved by incorporating the number of shared sites in calculating the probability of multiple mutations at nucleotide positions that differ between a given pair of operational taxonomic units (OTUs). Thus, the fewer differences (more shared sites) between a pair of OTUs, the greater the probability that those few nucleotide substitutions are due only to single mutational event. This estimation procedure has demonstrated statistical power when reconstructing gene trees and greatly outperforms bootstrapping with maximum parsimony when the number of nucleotide substitutions is small and the number of shared positions is large (Crandall 1994), as is the case with the PTLV sequence data.

The statistical parsimony procedure was implemented by first obtaining a distance matrix of absolute differences (in numbers of nucleotides) separating each pair of individual viral sequences using the SHOW DISTANCE MATRIX option in PAUP version 3.1.1 (Swofford 1993). Host individuals with identical viral sequences were collapsed into a single haplotype or OTU. The minimum number of nucleotide differences that separate every pair of OTUs was then obtained from the difference matrix. The probability of a parsimonious connection between OTUs was calculated using equations (6), (7), and (8) of Templeton, Crandall, and Sing (1992). The resulting probabilities of parsimonious connections give a quantitative assessment of the reliability of the particular connection that is analogous to a bootstrap percentile value. When parsimonious connections alone could not be justified using the statistical criterion, nonparsimonious connections between haplotypes were examined using equation (9) of Templeton, Crandall, and Sing (1992).

For a direct comparison with the results from a standard maximum parsimony approach, maximum parsimony trees were estimated using the HEURISTIC SEARCH option in PAUP version 3.1.1 with random sequence addition and multiple replicates (at least 10). For searches resulting in more than a single most parsimonious tree, majority-rule consensus trees were also calculated using PAUP (Swofford 1993). Confidence in the resulting trees was assessed using the bootstrap approach (Felsenstein 1985; Hillis and Bull 1993), with 100 bootstrap replications.

Root Probabilities

The rooting of intraspecific gene trees is particularly difficult using traditional methods such as outgroup rooting. One problem associated with outgroup rooting is that species phylogenies are expected to progress through a transition of polyphyly to paraphyly to monophyly in terms of gene tree relationships relative to a sister group (Neigel and Avise 1986; Takahata and Slatkin 1990). The speed of this progression depends greatly on population size and geographic substructure (Neigel and Avise 1986). While species are within the transitional phases of polyphyly and paraphyly, the effects of ancestral polymorphism and lineage sorting result in the possibility of outgroup OTUs falling within the ingroup. Only when the ingroup has reached the stage of monophyly with respect to a sister species can this outgroup be used to root the cladogram. Unfortunately, this situation leads to another difficulty where the outgroup can be so far removed (genetically) from the ingroup that its connection to the ingroup has no statistical resolution (Templeton 1992, 1993). Similar difficulties are encountered with other methods of rooting such as midpoint rooting (Wills 1992).

Coalescence theory offers great potential for increasing the accuracy of phylogenetic estimations, including directionality of mutational changes (Crandall and Templeton 1993; Excoffier and Smouse 1994). Theoretical studies have established a direct relationship between haplotype frequency and the age of haplotypes (Watterson and Guess 1977; Donnelly and Tavaré 1986). Likewise, coalescence theory predicts that older haplotypes should have more mutational connections and occur preferentially in the interior of the cladogram, whereas recently derived haplotypes should be localized at the tips. Crandall and Templeton (1993) tested these predictions and others using 29 empirical data sets from the literature and showed them to be strongly supported. These predictions are robust over a diversity of biological conditions, because the 29 data sets represented multiple loci over multiple species of *Drosophila*. Furthermore, similar patterns have been indicated in many other species, e.g., humans (Templeton 1993; Excoffier and Smouse 1994), plants (Matos and Schaal 1995), and salamanders (Routman 1993).

Castelloe and Templeton (1994) have used these arguments from neutral coalescence theory in determining root probabilities for intraspecific gene trees. The exact method for determining root probabilities is not feasible for implementation with large data sets such as these. Castelloe and Templeton (1994) suggest a heuristic approach to assigning outgroup weights based on the following observations: (1) interior OTUs are more likely to be closest to the root than tip OTUs, where interior OTUs have more than one connection and tip OTUs have just a single connection, (2) the ratio of OTU frequencies tends to be a conservative estimator for tip haplotype root probabilities, and (3) interior OTUs have high root probability when they are of high frequency or are connected to other OTUs of high frequency. Outgroup weights are then assigned to individual OTUs as follows; tip OTUs are assigned a weight of one-half their frequency in the sample. Interior OTUs are assigned a weight of their frequency plus the frequency of all OTUs one mutational step away. These weights are then summed and normalized to sum to one, resulting in relative weights for each OTU that correlate with the age order of alleles (Castelloe and Templeton 1994). The haplotype (or set of haplotypes) that represents the oldest allele in relative age (has the highest outgroup weight) is then the best candidate to serve as an outgroup for the remainder of the haplotype tree.

In applying this heuristic algorithm to real data sets, at least two complications should be considered. The first complication is that of uncertainty in the estimated haplotype tree topology. Uncertainty can exist in tree topologies due to either homoplasy or recombination (Castelloe and Templeton 1994). Fortunately, the estimation procedure of Templeton, Crandall, and Sing (1992) quantifies the potential for homoplasy by evaluating the probability of multiple substitutions at each linkage between haplotypes. Each linkage within the 95% probable set, by definition, has a probability of 0.95 or greater of not being affected by multiple hits, thereby justifying the use of the infinite sites model assumed by coalescence theory (Crandall and Templeton 1993). When topological ambiguities exist due to homoplasy, outgroup weights can be assigned to a haplotype by incorporating the probabilities of alternative topologies into the outgroup weighting procedure (Crandall and Templeton 1993; Crandall, Templeton, and Sing 1994). Similarly, recombination can complicate the assignment of outgroup weights as one subsection of DNA might have a different "root" than another subsection. Again, the estimation procedure of Templeton, Crandall, and Sing (1992), using the algorithms of Hein (1990, 1993), can subdivide a region resulting from recombination into subregions with little or no recombination. The statistical parsimony method is then applied to each subregion separately.

A second complication comes from the neutrality assumption of coalescence theory (Castelloe and Templeton 1994). There are two classes of selection with effects on estimating outgroup probabilities. The first class is selection that maintains polymorphism, i.e., balancing selection. Takahata (1990), Takahata and Nei (1990), and Takahata, Satta, and Klein (1992) have shown for many models that balancing selection influ-

ences branch lengths among allelic classes but has little impact on the overall topology of the allele genealogy and accentuates the tendency of the common haplotypes to be old and in the interior of the cladogram. Thus, this heuristic method of assigning outgroup weights should be robust to deviations from neutrality involving selectively maintained polymorphism (Castelloe and Templeton 1994).

The second class is directional selection. Golding (1987) has demonstrated that deleterious mutations do not persist long and are found primarily at the tips of cladograms. This result is consistent with the root probability algorithm (Castelloe and Templeton 1994). Directional selection for a favorable mutation, however, can cause erroneous results if the observer is unfortunate enough to make the study during the transient polymorphism phase. During a transient phase of directional selection, a favored mutation will often have high frequency in the sample despite its recent origin (Takahata 1991). However, because of its recent origin, a favorable mutation is still expected to occur topologically on the tips of the cladogram. Because the algorithm of Castelloe and Templeton (1994) incorporates frequency, topological status, and the number of mutational connections, it will still give low outgroup weight to recently derived favorable mutations.

While the resulting outgroup weights will not allow the exact root of the haplotype tree to be determined, they will indicate the most likely places of root location within the haplotype tree and serve as a good indicator of the relative ages of various haplotypes (Castelloe and Templeton 1994). These results are supported by analytical theory (Watterson and Guess 1977; Donnelly and Tavaré 1986), computer simulation (Castelloe and Templeton 1994), and empirical data (Crandall and Templeton 1993).

The method of Castelloe and Templeton (1994) for assigning outgroup weights was developed principally for restriction site data where haplotypes are typically separated by a single restriction site difference. With this sequence data analysis, haplotypes are typically differentiated by more than a single nucleotide difference. Therefore, in calculating relative outgroup weights, I assign to each tip haplotype a weight of one half its frequency and to interior haplotypes, its frequency plus the frequency of all haplotypes to which it is connected, regardless of whether or not the connection is a single mutational difference. These weights are then normalized to sum to one. This allows for the incorporation of the degree of connectedness, which is an important indicator of relative age of alleles (Crandall and Templeton 1993; Castelloe and Templeton 1994). For OTUs of ambiguous topological position, outgroup weights were altered by the probabilities of the alternative resolutions for the ambiguous portion of the cladogram (Crandall, Templeton, and Sing 1994). The resulting outgroup weights are then used to test hypotheses of transmission directionality and geographic origin of viral types.

Nested Statistical Analyses

Templeton, Boerwinkle, and Sing (1987) and Templeton and Sing (1993) have developed statistical procedures for detecting significant associations between phenotype and genotype within a cladogram framework. Their procedures utilize the cladogram structure from the above estimation procedure to define a nested statistical design, thereby allowing the clustering of individuals based on genotype rather that phenotype. The statistical analysis allows ambiguity in the cladogram estimation and is compatible with either quantitative or categorical phenotypes.

The Nesting Procedure

The nesting procedure consists of nesting n-step clades within $(n + 1)$-step clades, where n refers to the number of transitional steps used to define the clade. Thus, n is correlated with, but does not refer to, the number of nucleotide differences separating individual haplotypes. By definition, each haplotype or OTU is a zero-step clade. The $(n + 1)$-step clades are formed by the union of all n-step clades that can be joined together by $(n + 1)$ mutational steps. The nesting procedure begins with tip clades, i.e., those clades with a single mutational connection (e.g., haplotypes H1, H5, and H6 in fig. 1), and proceeds to interior clades. In previous analyses based on restriction site data, missing intermediates were ignored in the nesting procedure as they were inconsequential to these analyses. However, with nucleotide sequence data, there are many more missing intermediates because haplotypes are typically differentiated by more than a single nucleotide difference. These missing intermediates must be considered in the nesting procedure to assure overall consistency. Because these missing intermediates become nested together, the nesting procedure results in a number of empty clades, i.e., two missing intermediates are nested together resulting in a next level clade that represents a missing intermediate as well (see zeros nested together in fig. 1). These next level empty clades are required for the consistency of the nesting procedure to form higher level clades but can be ignored during subsequent statistical analyses since they contain no observations.

Figure 1 offers an example of the nesting procedure performed with all the missing intermediates designated by zeros. The one-step nesting level produces eight clades of which five contain sampled haplotypes. Thus, these five clades are labeled ***1–1*** through ***1–5***, where the first number refers to nesting level and the second is a

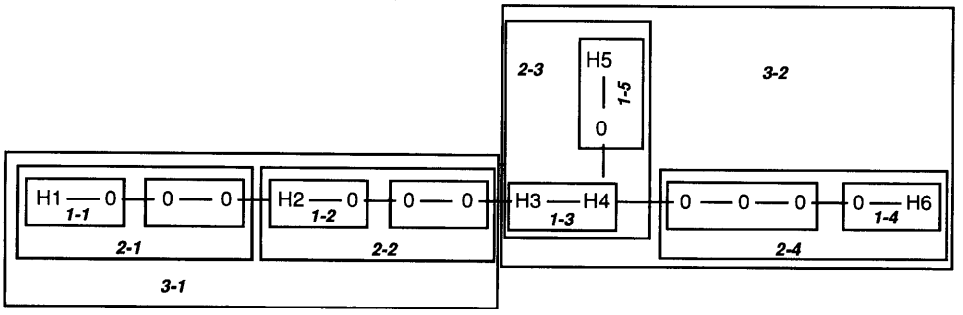

FIG. 1.—A demonstration of the nesting procedure for nucleotide sequence data. H1 through H6 represent the haplotypes under consideration. Lines indicate the mutational pathway interconnecting the six haplotypes with zeros representing missing intermediates. Boxes indicate nesting clades labeled with two numbers in bold and italics. The first number indicates the nesting level and the second is a counter of the clades at that level. Thus, for example, clade ***2–1*** is the first clade formed at the second nesting level. Increasing nesting level corresponds to relatively increasing evolutionary time.

counter for clades containing sampled haplotypes that have been nested at that level. Notice one clade contains three missing intermediates. After nesting in from the H5 and H6 tips, the first missing intermediate to the right of H4 can either nest with the H3–H4 clade or with the clade of missing intermediates to its right. This situation has been termed symmetrically stranded (Templeton and Sing 1993). The placement of the stranded haplotype is initially based on sample size, i.e., it is placed in the clade with the smallest sample size. This results in greater samples within and among clades for hypothesis testing. Therefore, in this example, the missing intermediate is nesting with the other missing intermediates. If both alternatives have the same sample sizes, then one alternative is chosen at random (Templeton and Sing 1993). Now, the two-step nesting begins with the underlying one-step clades as the "haplotypes," resulting in four two-step clades. Nesting continues until the step before all haplotypes are nested into a single clade. Additional rules for nesting with ambiguity are given in Templeton and Sing (1993). The nesting procedure results in hierarchical clades with nesting level directly correlated to evolutionary time, i.e., the lower the nesting level the more recent the evolutionary events relative to higher nesting levels.

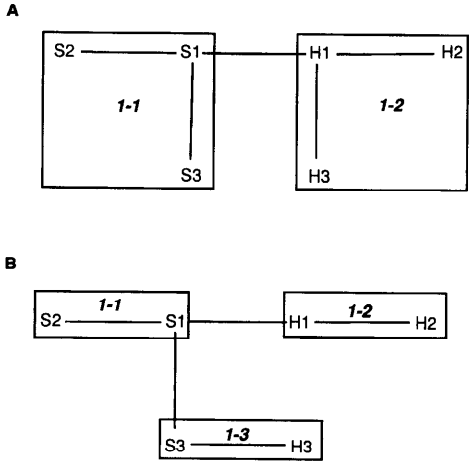

FIG. 2.—(*a*) The null hypothesis of no transmission among species is accepted when all haplotypes from one species (S) are nested together before they are nested with haplotypes from another species (H). (*b*) The null hypothesis of no transmission is rejected when nesting of two species occurs before the nesting of all haplotypes from a single species (H3 and S3).

Testing the Transmission Hypothesis Within the Nested Framework

Under the null hypothesis of no transmission, the haplotypes within a species are expected to nest together within a single nesting category (clade) before the nesting of additional species at higher nesting levels (fig. 2*A*). In figure 2*A*, we fail to reject the null hypothesis because each clade, ***1–1*** and ***1–2***, has no heterogeneity in the categorical variable of human/simian. All haplotypes of each species are nested together before the joining of additional species. Alternatively, figure 2*B* shows the nesting of haplotypes H3 and S3 (clade ***1–3***) at the one-step level, while other S and H haplotypes are nested in other clades at this level. This clearly rejects the null hypothesis of no transmission. Because nesting level correlates with relative ages of the clades, indicated transmission events can also be ranked in relative evolutionary age. Confidence in the rejection of the null hypothesis is dependent upon the confidence in the topological relationships as inferred by the cladogram estimation procedure.

Results

Cladogram Estimation

The implementation of the cladogram estimation procedure justified parsimonious connections among

Table 3
Probabilities of Multiple Substitutions at Site Differences Between Pairs of Haplotypes Differentiated by the Number of Differences Shown in Column One

MIN.	*env*		*pol*	
DIFF.[a]	ParsProb[b]	ParsPlus1[c]	ParsProb[d]	ParsPlus1
1	1.00	—	1.00	—
2	1.00	—	1.00	—
3	1.00	—	0.98	—
4	1.00	—	0.97	—
5	0.99	—	0.95	—
6	0.98	—	0.93	1.00
8	0.97	—	0.88	1.00
9	0.96	—	NA	NA
10 ...	0.95	—	NA	NA
11 ...	0.94	1.00	(12) 0.75	(12) 0.97
15 ...	0.90	0.99	0.64	0.93
17 ...	0.86	0.99	NA	NA
18 ...	0.85	0.99	NA	NA
19 ...	0.83	0.99	NA	NA

NOTE.—NA, denotes probabilities that are not applicable to the particular data set.

[a] Minimum number of nucleotide differences separating the designated OTU from its nearest neighbor.

[b] One minus the probability of a multiple substitution occurring at the minimum number of differences separating a pair of haplotypes (min. diff.). The probabilities were calculated by equations (6), (7), (8) in Templeton, Crandall, and Sing (1992), with j = min. diff., m = 522 (the total sequence length) − min. diff., u = 1 (the upper bound of the uniform prior on the probability that a nucleotide has experienced a mutation since the two haplotypes first diverged), b = 3 (no transition/transversion bias), and r = 1 (length of the recognition sequence).

[c] Cumulative probability of a parsimonious connection and a parsimonious plus one additional nucleotide substitution calculated by equation (9) in Templeton, Crandall, and Sing (1992) given the parameters above.

[d] As in [b] above except m = 140.

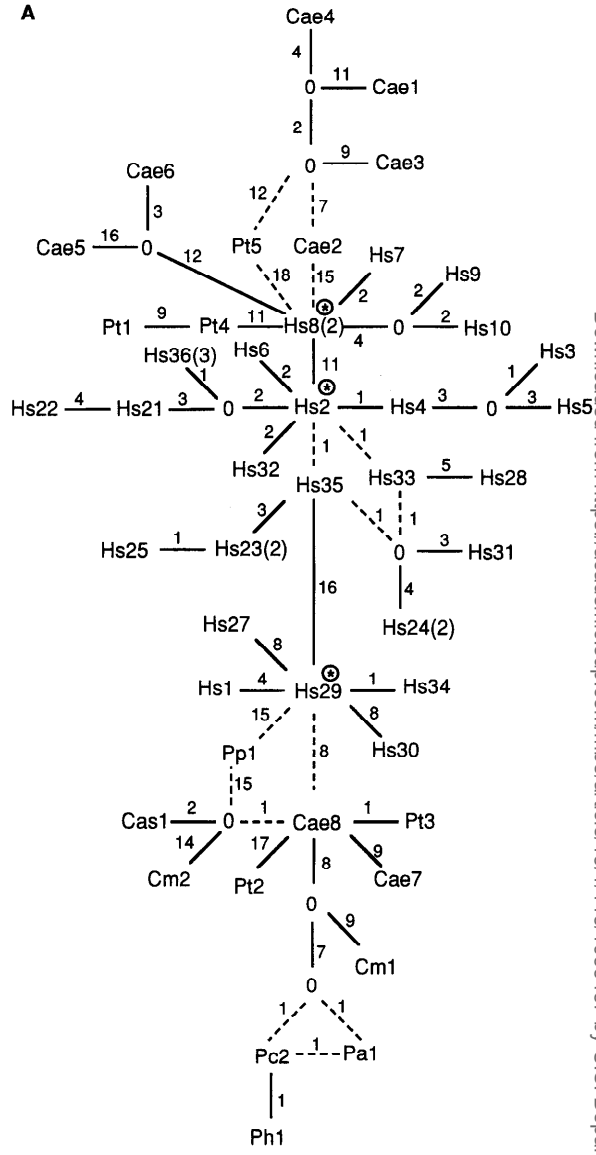

FIG. 3.—(A) Main network of relationships for the *env* sequence data estimated using the method of Templeton, Crandall, and Sing (1992). The lines indicate mutational connections while the number on each line indicates the number of substitutions separating the two connected OTUs. Dashed lines indicate ambiguous connections. The genus and species of the host species is indicated by first and second letters in the haplotype name, except for *Cercopithecus aethiops,* which is indicated by Cae to differentiate it from *C. ascanius.* The number of individual isolates represented by a haplotype is indicated in parentheses when that number is greater than one. Haplotypes with high outgroup weights (>0.05) are indicated by an asterisk within a circle.

OTUs differentiated by 10 or fewer nucleotide substitutions for the *env* data and by five or fewer substitutions for the *pol* data (table 3). Additionally, parsimonious plus one additional mutation connections were also justified for haplotypes separated by 25 or fewer nucleotide substitutions for the *env* data and 11 or fewer for the *pol* data (table 3). These OTUs were connected into parsimony networks for each gene region with the number of nucleotide differences separating each pair shown on the interconnecting linkages (fig. 3 for *env* and fig. 4 for *pol*). OTUs or subtrees separated by more than the number of nucleotide substitutions justified by the procedure formed independent subtrees whose linkage point on the main network could not be statistically justified (fig. 3*B* and fig. 4*B*; table 3). Connections between haplotypes of the human subtree in figure 3*B* and haplotypes from the network in figure 3*A* could not be statistically justified because the fewest number of mutational differences separating any pair of haplotypes from these two was greater than 20 differences. Note that the connection of viruses isolated from members of the genus *Macaca* are made at the highest level (Mm1 12 steps in fig. 4*A*) or form independent subtrees from the other viral isolates (Mf1 and Mf2 in fig. 4*B* and Mm1, Mm2, Mf1 and Pc1 in fig. 3*B*). Because heterogeneity in species name occurs within some of the clades represented in figures 3*A* and 4*A*, these clades will be used to identify transmission events across species. Therefore, only those in-

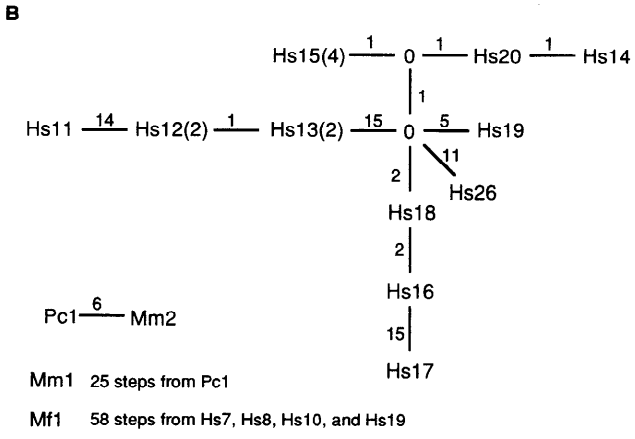

FIG. 3.—(*B*) Other independent trees whose connection to each other or to the main network (*A*) could not be justified by the method of Templeton, Crandall, and Sing (1992). These trees are not used in subsequent analyses of the transmission hypothesis because they contain only single host species.

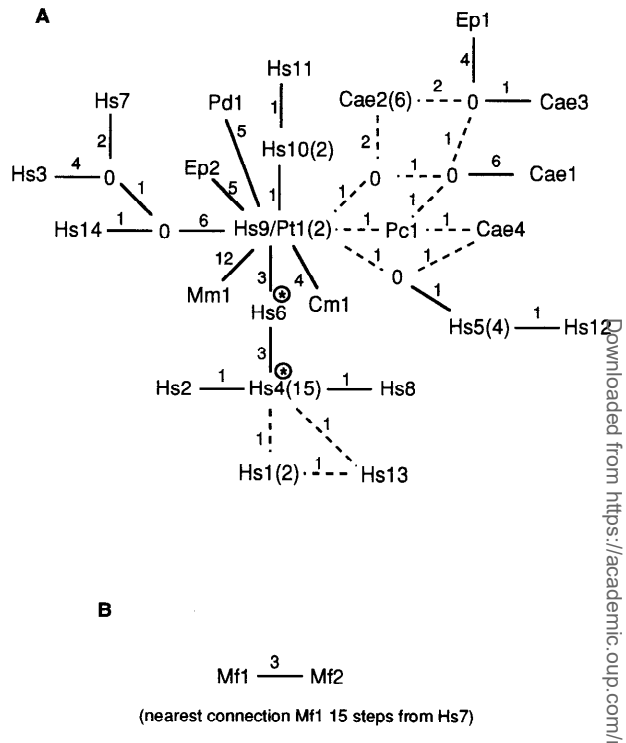

FIG. 4.—Parsimony network estimated using the method of Templeton, Crandall, and Sing (1992) for PTLV-I sequences from the *pol* gene. The haplotype Hs9/Pt1 indicates that the same isolate was found in these two host species. Haplotypes with high outgroup weights (≥ 0.20) are indicated by an asterisk within a circle.

dividuals and haplotypes associated with these clades will be evaluated in the calculations of outgroup weights and in nesting procedures.

Maximum parsimony analysis on the *env* data haplotypes resulted in over 8,000 most parsimonious trees. The computer memory limited the search to 8,000 trees. Thus, the effectiveness of the maximum parsimony search was restricted. The majority-rule consensus tree of this limited search is given in figure 5 with the percentage of times a particular node was represented in the 8,000 parsimony trees shown on that node. Bootstrap analysis was not possible with this data set due to the large number of maximum parsimony trees. The maximum parsimony analysis for the haplotypes of the *pol* data set resulted in 600 most parsimonious trees. The majority-rule consensus tree with bootstrap values (shown in parentheses) is shown in figure 6. Also shown on this tree is the number of changes that occur unambiguously on a branch. For branches involved in polytomies, the number of changes could not be calculated (Maddison and Maddison 1992).

By contrasting the networks in figures 3 and 4 with the unrooted trees presented in figures 5 and 6, the differences in the output of these two procedures is immediately apparent. Representing alternative trees in a network is a more precise way of representing tree ambiguity than the more standard consensus tree representation because it not only identifies which connections are ambiguous but also identifies probable alternatives. Note also that multifurcations are common throughout the networks and the more common haplotypes in the gene pool (those haplotypes represented by multiple individuals) are located in the interior of the main networks. These topological properties, so different from species trees, are exactly what we expect from coalescence theory (Crandall and Templeton 1993).

Outgroup Weights

Outgroup weights were calculated only for those haplotypes connected to the main network for each data set (figs. 3*A* and 4*A*). The calculation of outgroup weights for the main network of the *env* gene indicates three interior OTUs (haplotypes Hs2, Hs8, and Hs29) with outgroup weights >0.09 (table 4*A*). Haplotypes Hs23, Hs35, Cae8, and Pt4 all have outgroup weights between 0.09 and 0.05. The remaining haplotypes have outgroup weights <0.05. For the *pol* data, two haplotypes have outgroup weights of >0.2; Hs4 and Hs6 (table 4*B*). Five additional haplotypes have outgroup weights between 0.10 and 0.05; Hs1, Hs5, Hs9/Pt1, Hs10, Hs13, and Cae2. These resulting outgroup weights can now be incorporated into hypotheses of directionality of inferred transmission events (see below).

Nested Analyses

The one-step nesting on the network (fig. 3*A*) of the *env* gene tree results mainly in the clustering of viral haplotypes from humans (fig. 7). However, two transmission events are indicated. The first is in clade ***1–4***,

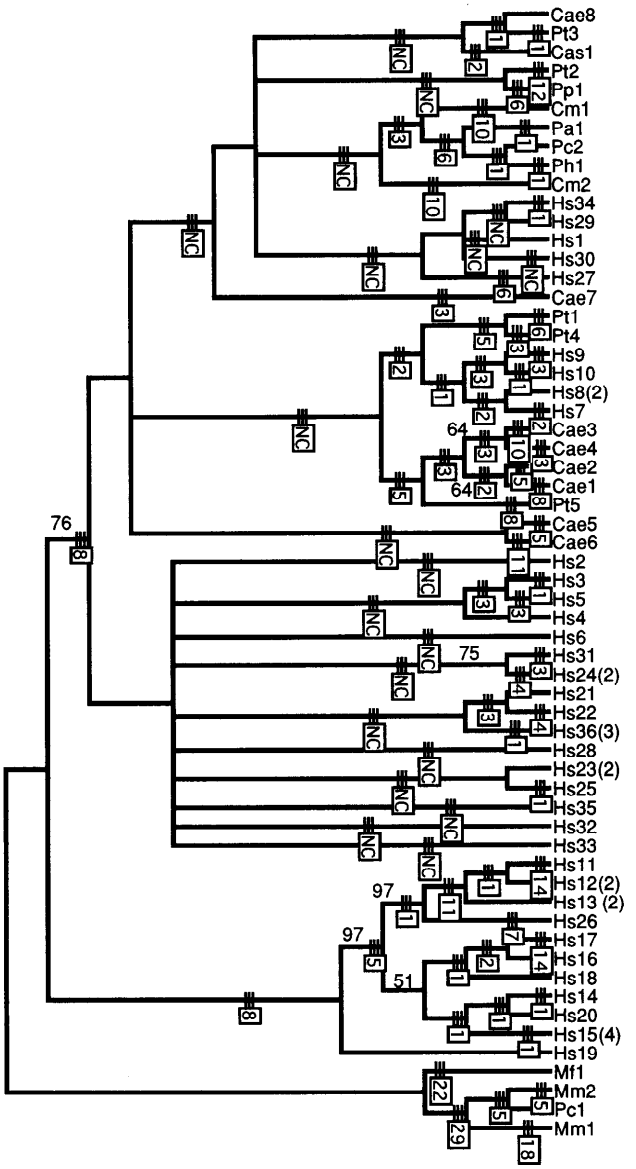

FIG. 5.—Maximum parsimony majority-rule consensus tree of the *env* haplotypes of the 8,000 most parsimonious trees. Numbers on nodes indicate the percent representation among the 8000 most parsimonious trees when less than 100%. The number of unambiguous changes along branches are given in a box on that branch and were reconstructed using MacClade version 3.04 (Maddison and Maddison 1992). Numbers of changes along nonbifurcating branches are not calculated by MacClade and are labeled NC.

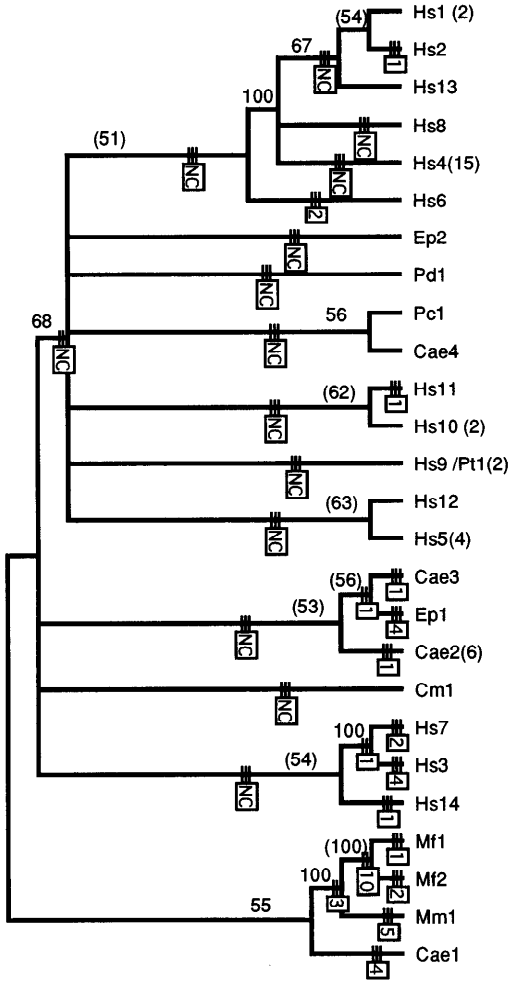

FIG. 6.—Maximum parsimony majority-rule consensus tree of the *pol* haplotypes with bootstrap percentile values given in parentheses on the nodes of those relationships supported by 50% or better based on 100 bootstrap replications. Other percentile values given on nodes indicate nodes that were supported in the majority-rule consensus tree for the 600 most parsimonious trees but failed to be supported at the 50% level in the bootstrap analysis. All nodes with a bootstrap percentile value were supported in the majority-rule consensus tree at 100% except for the Hs1/Hs2 node that had a 67% value in the majority-rule tree.

which nests viral sequences from a *Cercopithecus aethiops* host and a *Pan troglodytes* host. These haplotypes nest before the nesting of all individuals from each host species (e.g., see clade ***4–2*** and ***4–3***), thereby rejecting the null hypothesis of no transmission. The second instance of transmission is indicated in clade ***1–5***, which nests *Papio hamadryas* with *Papio cynocephalus*. Transmission is indicated because the other *P. cynocephalus* host is not even represented in the network (see fig. 3*B*). Thus, due to the nonunion of this species, at least one transmission event is indicated. Two- and three-step nestings show no additional nesting of different host species accept ***2–6***, which nests *Papio anubis* with two other *Papio* species. Because there is only a single sample from *P. anubis* and a transmission event already has been counted for the nonunion of Pc1 and Pc2, no transmission can be inferred. At the four-step nesting level, heterogeneity in host species name exists within the ***4–7*** clade, representing a possible transmission between *C. aethiops* and *Cercopithecus mitis*. Clade ***4–9*** nests the other *C. mitis* haplotype (Cm1) with subclade ***2–6***. So at least one transmission event is indicated in either ***4–***

Table 4A
Outgroup Weights for *env* Network of 45 Haplotypes Representing 50 Individual Isolates

Haplotype[a]	Topological Position	Mutational Connections	Outgroup Weight
Hs1	Tip	1	0.007
Hs2	Interior	6 or 7	0.093
Hs3	Tip	1	0.007
Hs4	Interior	2	0.029
Hs5	Tip	1	0.007
Hs6	Tip	1	0.007
Hs7	Tip	1	0.007
Hs8 (2)	Interior	6 or 7	0.093
Hs9	Tip	1	0.007
Hs10	Tip	1	0.007
Hs21	Interior	2	0.029
Hs22	Tip	1	0.007
Hs23 (2)	Interior	2	0.057
Hs24 (2)	Tip	1	0.014
Hs25	Tip	1	0.007
Hs27	Tip	1	0.007
Hs28	Tip	1	0.007
Hs29	Interior	5 or 6	0.093
Hs30	Tip	1	0.007
Hs31	Tip	1	0.007
Hs32	Tip	1	0.007
Hs33	Interior	2 or 3	0.040
Hs34	Tip	1	0.007
Hs35	Interior	3 or 4	0.068
Hs36 (3)	Tip	1	0.021
Cae1	Tip	1	0.007
Cae2	Ambiguous	1 or 2	0.025
Cae3	Tip	1	0.007
Cae4	Tip	1	0.007
Cae5	Tip	1	0.007
Cae6	Tip	1	0.007
Cae7	Tip	1	0.007
Cae8	Interior	5 or 6	0.068
Cas1	Tip	1	0.007
Cm1	Tip	1	0.007
Cm2	Tip	1	0.007
Pa1	Ambiguous	1 or 2	0.014
Pc2	Interior	2 or 3	0.034
Ph1	Tip	1	0.007
Pp1	Ambiguous	1 or 2	0.018
Pt1	Tip	1	0.007
Pt2	Tip	1	0.007
Pt3	Tip	1	0.007
Pt4	Interior	2	0.057
Pt5	Ambiguous	1 or 2	0.047

[a] Number in parentheses indicates the number of isolates representing a particular haplotype when greater than one.

Table 4B
Outgroup Weights for *pol* Network of 24 Haplotypes Representing 49 Individual Isolates

Haplotype	Topological Position	Mutational Connections	Frequency	Outgroup Weight
Hs1 (2)	Ambiguous	1 or 2	0.041	0.0790
Hs2	Tip	1	0.020	0.0058
Hs3	Tip	1	0.020	0.0058
Hs4 (15)	Interior	4 or 5	0.306	0.235
Hs5 (4)	Interior	2	0.082	0.0590
Hs6	Interior	2	0.020	0.212
Hs7	Tip	1	0.020	0.0058
Hs8	Tip	1	0.020	0.0058
Hs9/Pt1[a] (2) ..	Interior	8, 9, or 10	0.041	0.0873
Hs10 (2)	Interior	2	0.041	0.0590
Hs11	Tip	1	0.020	0.0058
Hs12	Tip	1	0.020	0.0058
Hs13	Ambiguous	1 or 2	0.020	0.0746
Hs14	Tip	1	0.020	0.0058
Cae1	Tip	1	0.020	0.0058
Cae2 (6)	Ambiguous	1 or 2	0.122	0.0706
Cae3	Tip	1	0.020	0.0058
Cae4	Ambiguous	1 or 2	0.020	0.0145
Cm1	Tip	1	0.020	0.0058
Ep1	Tip	1	0.020	0.0058
Ep2	Tip	1	0.020	0.0058
Mm1	Tip	1	0.020	0.0058
Pc1	Ambiguous	1, 2, or 3	0.020	0.0278
Pd1	Tip	1	0.020	0.0058

[a] This isolate was sequenced from both a *Homo sapiens* host and a *Pan troglodytes* host.

7 or ***4–9*** depending upon the rooting of the network. At the five-step level clade ***5–2*** and ***5–4*** indicate transmission events. Clade ***5–2*** indicates transmission between *Homo–Pan* and *Homo–Cercopithecus* hosts. Clade ***5–4*** indicates transmission between *Cercopithecus–Pan.* Finally, the six-step level indicates transmissions in both clades ***6–1*** and ***6–2***. Clade ***6–1*** indicates a transmission event between clade ***5–4*** and *Homo* (clade ***5–1***). Because of the diversity of species within clade ***5–4***, it is not possible to discern the exact host species associated with this transmission. Clade ***6–2*** ambiguously indicates either transmissions between *Cercopithecus–Homo* and *Pan–Homo, Cercopithecus–Homo* and *Cercopithecus–Pan,* or *Pan–Homo* and *Pan–Cercopithecus.*

The *n*-step clades for the main network (fig. 4*A*) of the *pol* gene region are shown in figure 8. These data are particularly interesting because two different host species are nested at the zero-step level, i.e., isolates from *P. troglodytes* and *H. sapiens* show identical sequences for this gene region and are therefore placed in the same haplotype (Hs9/Pt1). This nesting indicates a possible transmission event. However, because there is only a single sample from *P. troglodytes,* the inference of a transmission event depends on the root of the network. Notice this haplotype has an outgroup weight greater than 0.05 (0.0873, table 4*B*). Thus, it cannot be ruled out as a possible root haplotype. If it were closest to the true root then no transmission event would be indicated. Only a single clade at the one-step level shows nesting of haplotypes from different host species (***1–1***; fig. 8). This nesting either indicates a transmission event between *Homo* and *P. cynocephalus* or is the high-

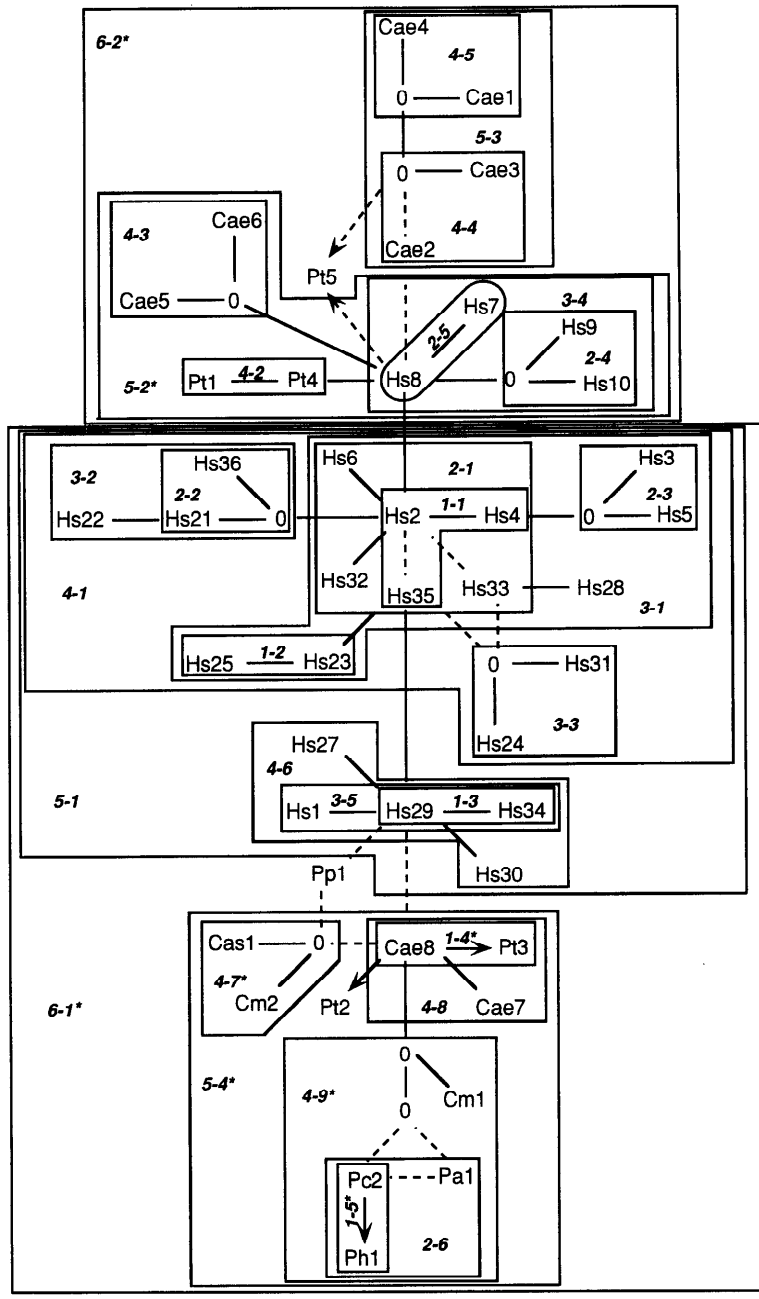

FIG. 7.—Nesting of the *env* network shown in figure 3*A*. The (n + 1) step clades are shown in bold and italics (***n-c***), where ***n*** refers to the nesting level and ***c*** is a counter for the number of clades at that level. The seven clades that contain subclades that reject the null hypothesis of no transmission at 95% confidence are indicated with an asterisk. When directionality of the indicated transmission event could be inferred, it has been labeled with an arrow in that direction.

est nesting level between *Pan* and *P. cynocephalus*. Additional samples from both *P. troglodytes* and *P. cynocephalus* would provide a distinction between these two alternatives. Clade ***2–1*** indicates a transmission event between either *Homo–Pan–Papio* and *C. aethiops* (Cae4). Clade ***3–1*** nests together three different species with the *Homo–Pan–Papio–Cercopithecus* (***2–1***) subclade, yet two of these are the sole representatives of that species (Cm1, Pd1) making inference of transmission questionable. Although, because all human haplotypes have not yet nested together, cross-species transmission is a possibility. There are two samples of *Erythrocebus patas,* one of which nests with *C. aethiops* haplotypes, the other nests with subclade ***2–1***. Thus there is at least one cross-species transmission indicated in either subclade ***3–2*** or ***3–1***. No other transmissions

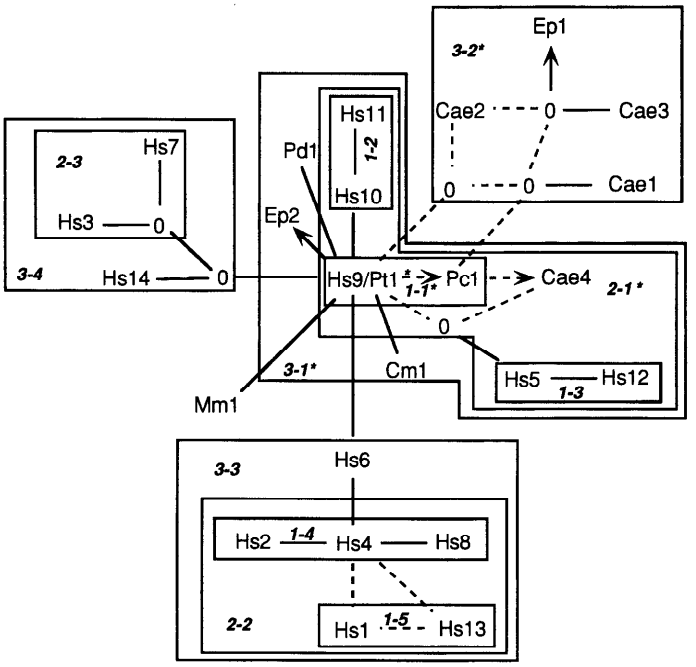

FIG. 8.—Nesting of the *pol* network from figure 4*A*. Nesting designations are as in figure 7.

are indicated as all three-step level subclades are nested together with haplotype Mm1 at the next nesting level.

Discussion

Geographic Origin Hypotheses

The geographical origin of PTLV sequences has been hypothesized to be in Africa (Gallo, Sliski, and Wong-Staal 1983; Watanabe et al. 1985) or in Melanesia (Sherman et al. 1992; Koralnik et al. 1994; Song et al. 1994). Using the data in this analysis, one can examine outgroup weights and determine geographic sampling localities for those haplotypes possessing the highest outgroup weights. Using this strategy on the *env* network, haplotypes Hs2 (West Africa) and Hs8 (Central Africa) have outgroup weights greater than 0.09, but Hs29, from the West Indies, does as well. Additionally, Hs23 (La Reunion Island), Hs35 (Brasil), Cae8 (East Africa), and Pt4 (unknown) have outgroup weights greater than 0.05. These data are somewhat equivocal in their support for an African origin of PTLV sequences. Furthermore, the tree shown in figure 3*B* is made up of entirely Melanesian haplotypes. Thus, the question of origin cannot be answered using this approach. If, however, the tree in figure 3*B* could be shown to be ancestral, this would support the notion of a Melanesian origin of PTLV sequences. The fact that these data form strong geographic clusters is important to consider in the development of vaccines and for epidemiological research. Notice that even within the main network (fig. 3*A*), haplotypes cluster by geographic location (e.g., clade ***5–3*** contains all *C. aethiops* from West Africa). However, counter examples to clustering by geographic location are present (e.g., clade ***1–1*** clusters African hosts with a host from Brazil, then the next level [***2–1***] clusters an additional African host and two North American hosts).

For the *pol* cladogram Hs6 (Caribbean) and Hs4 (USA, India, Japan, Bellona, Brazil) have outgroup weights greater than 0.210. Included in the set of haplotypes with outgroup weight greater than 0.05 are the following: Hs1 (West Africa, Japan), Hs5 (Central Africa), Hs9/Pt1 (West and Central Africa), Hs10 (East Africa), Hs13 (India), and Pc1 (Southern Africa). Thus, there is no good indication of geographic origin from these data either. Here again, clustering by geographic location is noticeable with fewer exceptions. For example, clades ***3–1*** and ***3–2*** contain haplotypes from Africa, whereas clade ***3–4*** contains haplotypes from Melanesia.

Directionality of Transmissions

The nesting procedure applied to the main network of the *env* gene (fig. 3*A*) indicates either eight or nine instances of cross-species transmission with the events partitioned into lower level clades and higher level clades (table 5). No transmission events are indicated at nesting levels 2 or 3. Directionality of transmission events can be inferred using the outgroup weights established above. For example, in clade ***1–4*** the directionality of the transmission event is in the Cae8 to Pt3

Table 5
Clades Resulting in a Rejection of the Null Hypothesis of No Cross-Species Transmission

Clade	Locus	Subclades[a]	Transmission Events	Directionality[b]
1-4	*env*	Cae8–Pt3	1	Cae8 → Pt3
1-5	*env*	Pc2–Ph1	1	Pc2 → Ph1
4-7	*env*	Cas1–Cm2	0 or 1[c]	Ambiguous
4-9	*env*	***2-6***–Cm1	0 or 1	Ambiguous
5-2	*env*	***4-3–4-2–3-4***	2	Ambiguous
5-4	*env*	***4-7***–Pt2–***4-8–4-9***	1	→ Pt2
6-1	*env*	***5-1***–Pp1–***5-4***	1	Ambiguous
6-2	*env*	***5-1***–Pt5–***5-3***	2	→ Pt5
Hs9/Pt1	*pol*	Hs9–Pt1	0 or 1	Ambiguous
1-1	*pol*	Hs9/Pt1–Pc1	0 or 1	Hs9/Pt1 → Pc1
2-1	*pol*	***1-1***–Cae4–***1-3***	1	→ Cae4
3-1	*pol*	Pd1–***2-1***–Ep2–Cm1	0 to 3[d]	→ Ep1 and/or Ep2
3-2	*pol*	Ep1–Cae1–Cae2–Cae3	0 or 1	Ambiguous

[a] Haplotypes are abbreviated as in tables 1 and 2. Subclade are separated by (–).
[b] Directionality of transmission inferred using outgroup weights. In some cases the donating haplotype was questionable, but the directionality to the receiving haplotype could still be inferred.
[c] At least one transmission event within either ***4-7*** or ***4-9***.
[d] At least one transmission event involving either Ep1 or Ep2.

direction because Pt3 has an outgroup weight of 0.007 relative to the rest of the network whose outgroup weight is (1 − 0.007), or 0.993. Thus, Pt3 can be ruled out for rooting purposes at the 95% confidence level. Likewise, the transmission in clade ***1–5*** is inferred to have occurred in the Pc2 to Ph1 direction at the same confidence level. In clade ***5–4***, directionality is indicated from subclade ***4–8*** to Pt2, because Pt2 has an outgroup weight of 0.007. In clade ***6–2***, directionality is indicated from either subclade ***5–2*** or ***5–3*** to Pt5, because Pt5 has an outgroup weight of 0.047.

A range of transmission events is inferred using the *pol* data set, from a minimum of two to a maximum of seven. The *pol* data set indicates directionality of transmissions in clades ***1–1***, ***2–1***, and ***3–1*** (table 5). In all cases, the directionality is away from human species. In no case could a human haplotype be unambiguously assigned as the transmitter of a viral infection to a simian species, but the majority of haplotypes in clades ***1–1*** and ***2–1*** are from human hosts. This could be an artifact of this particular data set, which is not as well sampled for nonhuman species as the *env* data set.

A Comparison with Maximum Parsimony

The first comparison to make between the method of Templeton, Crandall, and Sing (1992) and maximum parsimony is in the tree reconstruction. The network in figure 3*A* gives more resolution than the parsimony alternative shown in figure 5. Furthermore, because of the large number of most parsimonious trees encountered during the search, which was limited by the amount of memory available on the computer, these are not a complete set of most parsimonious trees. Indeed, there may be shorter trees yet to be discovered (Maddison 1991*a*, 1991*b*). But how is this possible when both methods are using the same data set and the same optimality criterion, i.e., parsimony? The advantage of the method of Templeton, Crandall, and Sing (1992) is that it reconstructs pairwise relationships, not global relationships. In doing so, the method limits the amount of homoplasy it is willing to examine using the probability of parsimony calculation. This is obviously a disadvantage if you are interested in the entire tree because even if every linkage were supported unambiguously at the 0.95 level, with the 45 haplotypes listed in figure 3*A* assuming a minimum number of 45 − 1 or 44 linkages, the probability for the overall tree would be $P = 0.95^{44}$ or $P = 0.105$. Similar calculations can be performed on the maximum parsimony tree based on bootstrap values for each node relative to confidence in the entire tree. Except that for this particular example, bootstrap values are not available due to the large number of most parsimonious trees. However, the method of Templeton, Crandall, and Sing (1992) provides accurate linkages between pairs of OTUs (Crandall 1994). In this case, we are more concerned with specific linkages, namely those that link different host species before an entire host species is linked together. For the *pol* data set, only a single island of most parsimonious trees was found, but this island contains 600 most parsimonious trees. The bootstrap majority-rule consensus tree offers little resolution compared to the tree estimated by the method of Templeton, Crandall, and Sing (1992). Thus, the resolving power for pairwise linkages is greater using the method of Templeton, Crandall, and Sing (1992) because it lim-

its the amount of homoplasy considered by linking only those haplotypes that share many sites and differ by few. Furthermore, those sites by which these haplotypes differ have a high probability ($\geq$0.95) of being due to a single substitution event.

The second area of comparison is in the identification of possible instances of cross-species transmission of the PTLVs. For this comparison, I will use only the *pol* data set because the appropriate comparison is between confidence sets, i.e., the network from the method of Templeton, Crandall, and Sing (1992) (fig. 4*A*) and the bootstrapped majority rule consensus tree from maximum parsimony (fig. 6). Besides the obvious case of the Hs9/Pt1 haplotype, one additional transmission event is suggested by the maximum parsimony tree that is supported by a bootstrap value, the Cae3–Ep1 clade. This clade is supported by a bootstrap value of 56%. This example demonstrates the utility of the nested analysis approach and its ability to partition historical events over time. The nested analysis identified additional transmission events, the specific haplotypes (or sets of haplotypes) involved in the transmission events, and the mutational changes associated with each transmission event.

Ambiguity within the statistical parsimony networks did not greatly affect the subsequent nested analyses, since most ambiguity fell within single nesting categories. For example, the ambiguity in the placement of haplotype Hs33 is subsumed in clade ***2–1***. Some ambiguity remained in the nested analysis and in those cases (e.g., the placement of Pt5) each alternative was explored. The traditional approach, however, is severely hampered by ambiguity in the resulting set of most parsimonious trees. I tested hypotheses based on a majority-rule consensus tree. An alternative, and more conservative approach, would be to test the hypotheses against each most parsimonious tree. Clearly, it is desirable to account for the ambiguity in a more efficient manor, as does the nested analysis. There are two distinct advantages to the statistical parsimony approach over traditional analyses: (1) a more accurate estimation of phylogenetic relationships for data with low levels of divergence (Crandall 1994), and (2) a rigorous hypothesis testing framework, which provides a quantitative partitioning of population phenomena across evolutionary time (Templeton 1993; Templeton and Sing 1993). The statistical parsimony approach also allows for uncertainty in the cladogram estimation; therefore, it does not rely on a single estimate of phylogenetic relationships but is robust over a set of plausible alternative phylogenies (Templeton, Crandall, and Sing 1992; Templeton and Sing 1993).

Acknowledgments

I thank Alan Templeton for helpful discussions concerning various aspects of the analyses and Gerald Myers for piquing my interest in the HTLV/STLV data sets and multiple transmission hypotheses. I thank Julian Adams, Jim Bull, Walter Fitch, Alan Templeton, and two anonymous reviewers for critical comments on the manuscript. This work was supported by the National Science Foundation grant DEB-9303258 and a grant from the Alfred P. Sloan Foundation.

LITERATURE CITED

BASTIAN, I., J. F. GARDNER, D. WEBB, and I. GARDNER. 1993. Isolation of a human T-lymphotropic virus type I from Australian aboriginals. J. Virol. **67**:843–851.

CASTELLOE, J., and A. R. TEMPLETON. 1994. Root probabilities for intraspecific gene trees under neutral coalescent theory. Mol. Phylogenet. Evol. **3**:102–113.

CRANDALL, K. A. 1994. Intraspecific cladogram estimation: accuracy at higher levels of divergence. Syst. Biol. **43**:222–235.

CRANDALL, K. A., and A. R. TEMPLETON. 1993. Empirical tests of some predictions from coalescent theory with applications to intraspecific phylogeny reconstruction. Genetics **134**:959–969.

CRANDALL, K. A., A. R. TEMPLETON, and C. F. SING. 1994. Intraspecific phylogenetics: problems and solutions. Pp. 273–297 *in* R. W. SCOTLAND, D. J. SIEBERT, and D. M. WILLIAMS, eds. Models in phylogeny reconstruction. Clarendon Press, Oxford, England.

DONNELLY, P., and S. TAVARÉ. 1986. The ages of alleles and a coalescent. Adv. Appl. Probab. **18**:1–19.

DUBE, D. K., S. DUBE, S. ERENSOY et al. 1994. Serological and nucleic acid analyses for HIV and HTLV infection on archival human plasma samples from Zaire. Virology **202**: 379–389.

DUBE, D. K., M. P. SHERMAN, N. K. SAKSENA et al. 1993. Genetic heterogeneity in human T-cell leukemia/lymphoma virus type II. J. Virol. **67**:1175–1184.

EVANGELISTA, A., S. MAOUSHEK, H. MINNIGAN et al. 1990. Nucleotide sequence analysis of a provirus derived from an individual with tropical spastic paraparesis. Microb. Pathog. **8**:259–278.

EXCOFFIER, L., and P. E. SMOUSE. 1994. Using allele frequencies and geographic subdivision to reconstruct gene trees within a species: molecular variance parsimony. Genetics **136**:343–359.

FELSENSTEIN, J. 1985. Confidence limits on phylogenies: an approach using the bootstrap. Evolution **39**:783–791.

GALLO, R. C., A. SLISKI, and F. WONG-STAAL. 1983. Origin of human T-cell leukaemia–lymphoma virus. Lancet **ii**:962–963.

GESSAIN, A., E. BOERI, R. YANAGIHARA, R. C. GALLO, and G. FRANCHINI. 1993. Complete nucleotide sequence of a highly divergent human T-cell leukemia (lymphotropic) virus type I (HTLV-I) variant from melanesia: genetic and phyloge-

netic relationship to HTLV-I strains from other geographical regions. J. Virol. **67**:1015–1023.

GESSAIN, A., R. C. GALLO, and G. FRANCHINI. 1992. Low degree of human T-cell leukemia/lymphoma virus type genetic drift in vivo as a means of monitoring viral transmission and movement of ancient human populations. J. Virol. **66**: 2288–2295.

GESSAIN, A., R. YANAGIHARA, G. FRANCHINI, R. M. GARRUTO, C. L. JENKINS, A. B. AJDUKIEWICZ, R. C. GALLO, and D. C. GAJDUSEK. 1991. Highly divergent molecular variants of human T-lymphotropic virus type I from isolated populations in Papua New Guinea and the Solomon Islands. Proc. Natl. Acad. Sci. USA **88**:7694–7698.

GOLDING, G. B. 1987. The detection of deleterious selection using ancestors inferred from a phylogenetic history. Genet. Res., Camb. **49**:71–82.

GRAY, G. S., M. WHITE, T. BARTMAN, and D. MANN. 1990. Envelope gene sequence of HTLV-1 isolate MT-2 and its comparison with other HTLV-1 isolates. Virology **177**:391–395.

HEIN, J. 1990. Reconstructing evolution of sequences subject to recombination using parsimony. Math. Biosci. **98**:185–200.

———. 1993. A heuristic method to reconstruct the history of sequences subject to recombination. J. Mol. Evol. **36**:396–405.

HIGGINS, D. G., A. J. BLEASBY, and R. FUCHS. 1992. CLUSTAL V: improved software for multiple sequence alignment. Comput. Appl. Biosci. **8**:189–191.

HILLIS, D. M., and J. J. BULL. 1993. An empirical test of bootstrapping as a method for assessing confidence in phylogenetic analysis. Syst. Biol. **42**:182–192.

HINUMA, Y., K. NAGATA, M. HANAOKA, M. NAKAI, T. MATSUMOTO, K. KINOSHITA, S. SHIRAKAWA, and I. MIYOSHI. 1981. ATL: antigen in an ATL cell line and detection of antibodies to the antigen in human sera. Proc. Natl. Acad. Sci. USA **78**:6476–6480.

HUELSENBECK, J. P., and D. M. HILLIS. 1993. Success of phylogenetic methods in the four-taxon case. Syst. Biol. **42**: 247–264.

KALYANRAMAN, V. S., M. G. SARANGADHARAN, I. MIYOSHI, D. BLAYNEY, D. GOLDE, and R. C. GALLO. 1982. A new subtype of human T-cell leukemia virus (HTLV-II) associated with a T-cell variant of Hairy cell leukemia. Science **218**:571–573.

KOMURO, A., T. WATANABE, I. MIYOSHI, M. HAYAMI, H. TSUJIMOTO, M. SEIKI, and M. YOSHIDA. 1984. Detection and characterization of simian retroviruses homologous to human T-cell leukemia virus type I. Virology **138**:373–378.

KORALNIK, I. J., E. BEORI, W. C. SAXINGER et al. 1994. Phylogenetic associations of human and simian T-cell leukemia/lymphotropic virus type I strains: evidence for interspecies transmission. J. Virol. **68**:2693–2707.

KUIKEN, C. L., and B. T. M. KORBER. 1994. Epidemiological significance of intra- and inter-person variation of HIV-1. AIDS **8**:S73-S83.

LEE, R. V., A. W. PROWTEN, S. K. SATCHIDANAND, and B. I. S. SRIVASTAVA. 1985. Non-Hodgkin's lymphoma and HTLV-I antibodies in a gorilla. N. Engl. J. Med. **312**:118–119.

LI, W.-H., M. TANIMURA, and P. M. SHARP. 1988. Rates and dates of divergence between AIDS virus nucleotide sequences. Mol. Biol. Evol. **5**:313–330.

MADDISON, D. R. 1991*a*. African origin of human mitochondrial DNA reexamined. Syst. Zool. **40**:355–363.

———. 1991*b*. The discovery and importance of multiple islands of most-parsimonious trees. Syst. Zool. **40**:315–328.

MADDISON, W. P., and D. R. MADDISON. 1992. MacClade: analysis of phylogeny and character evolution. Sinauer Associates, Sunderland, Mass.

MAHIEUX, R., A. GESSAIN, A. TRUFFERT, D. VITRAC, A. HUBERT, J. DANDELOT, C. MONTCHAMP-MOREAU, F. CNUDDE, F. TEKAIA, and G. DE THE. 1994. Seroepidemiology, viral isolation, and molecular characterization of human T cell leukemia/lymphoma virus type I from La Reunion Island, Indian Ocean. AIDS Res. Hum. Retroviruses **10**:745–752.

MALIK, K. T. A., J. EVEN, and A. KARPAS. 1988. Molecular cloning and complete nucleotide sequence of an adult T cell leukaemia virus/human T cell leukaemia virus type I (ATLV/HTLV-I) isolate of Caribbean origin: relationship to other members of the ATLV/HTLV-I subgroup. J. Gen. Virol. **69**:1695–1710.

MATOS, J. A., and B. A. SCHAAL. 1996. Chloroplast evolution in the *Pinus montezumae* complex: II. A coalescent approach to hybridization. Evolution (in press).

NEIGEL, J. E., and J. C. AVISE. 1986. Phylogenetic relationships of mitochondrial DNA under various demographic models of speciation. Pp. 515–534 *in* S. KARLIN and E. NEVO, eds. Evolutionary processes and theory. Academic Press, New York.

NERURKAR, V. R., P. G. BABU, K.-J. SONG, R. R. MELLAND, C. GNANAMUTHU, N. K. SARASWATHI, M. CHANDY, M. S. GODEC, T. J. JOHN, and R. YANAGIHARA. 1993. Sequence analysis of human T cell lymphotropic virus type I strains from southern India: gene amplification and direct sequencing from whole blood blotted onto filter paper. J. Gen. Virol. **74**:2799–2805.

NERURKAR, V. R., K.-J. SONG, I. B. BASTIAN, G. BENOIT, G. FRANCHINI, and R. YANAGIHARA. 1994. Genotyping of human T cell lymphotropic virus type I using Australo-Melanesian topotype-specific olignonucleotide primer-based polymerase chain reaction: insights into viral evolution and dissemination. J. Infect. Dis. **170**:1353–1360.

PAINE, E., J. GARCIA, T. C. PHILPOTT, G. M. SHAW, and L. RATNER. 1991. Limited sequence variation in human T-lymphotropic virus type 1 isolates from North American and African patients. Virology **182**:111–123.

PAMILO, P., and M. NEI. 1988. Relationships between gene trees and species trees. Mol. Biol. Evol. **5**:568–583.

POIESZ, B. J., F. W. RUSCETTI, A. F. GAZDAR, P. A. BUNN, J. D. MINNA, and R. C. GALLO. 1980. Detection and isolation of type C retrovirus particles from fresh and cultured lymphocytes of a patient with cutaneous T-cell lymphoma. Proc. Natl. Acad. Sci. USA **77**:7415–7419.

RATNER, L., T. PHILPOTT, and D. B. TROWBRIDGE. 1991. Nucleotide sequences of North American and African HTLV-I isolates. AIDS Res. Hum. Retroviruses **7**:923–941.

ROUTMAN, E. 1993. Population structure and genetic diversity of metamorphic and paedomorphic populations of the tiger salamander, *Ambystoma tigrinum*. J. Evol. Biol. **6**:329–357.

SAKAKIBARA, I., Y. SUGIMOTO, A. SASAGAWA, S. HONJO, H. TSUJIMOTO, H. NAKAMURA, and M. HAGAMI. 1986. Spontaneous malignant lymphoma in an African green monkey naturally infected with simian lymphotropic virus (STLV). J. Med. Primatol. **15**:311–318.

SAKSENA, N. K., V. HERVE, J. P. DURAND et al. 1994. Seroepidemiologic, molecular, and phylogenetic analyses of simian T-cell leukemia viruses (STLV-I) from various naturally infected monkey species from Central and Western Africa. Virology **198**:297–310.

SAKSENA, N. K., V. HERVE, M. P. SHERMAN et al. 1993. Sequence and phylogenetic analyses of a new STLV-I from a naturally infected Tantalus monkey from Central Africa. Virology **192**:312–320.

SCHATZL, F., L. YAKOVLEVA, B. LAPIN, D. ROSE, L. INZHIA, K. GAEDIGK-NITSCHKO, F. DEINHARDT, and K. VON DER HELM. 1992. Detection and characterization of T-cell leukemia virus-like proviral sequences in PBL and tissues of baboons by PCR. Leukemia **6**:158–160.

SEIKI, M., S. HATTORI, Y. HIRAYAMA, and M. YOSHIDA. 1983. Human adult T-cell leukemia virus: complete nucleotide sequence of the provirus genome integrated in leukemia cell DNA. Proc. Natl. Acad. Sci. USA **80**:3618–3622.

SHERMAN, M. P., N. K. SAKSENA, D. K. DUBE, R. YANAGIHARA, and B. J. POIESZ. 1992. Evolutionary insights on the origin of human T-cell lymphoma/leukemia virus type I (HTLV-I) derived from sequence analyses of a new HTLV-I variant from Papua New Guinea. J. Virol. **66**:2556–2563.

SONG, K.-J., V. R. NERURKAR, N. SAITOU, A. LAZO, J. R. BLAKESLEE, I. MIYOSHI, and R. YANAGIHARA. 1994. Genetic analysis and molecular phylogeny of simian T-cell lymphotropic virus type I: evidence for independent virus evolution in Asia and Africa. Virology **199**:55–66.

SWOFFORD, D. L. 1993. PAUP: phylogenetic analysis using parsimony. 3.1.1. Smithsonian Institution, Washington, D.C.

TAKAHATA, N. 1990. A simple genealogical structure of strongly balanced allelic lines and trans-specific evolution of polymorphism. Proc. Natl. Acad. Sci. USA **87**:2419–2423.

———. 1991. Genealogy of neutral genes and spreading of selected mutations in a geographically structured population. Genetics **129**:585–595.

TAKAHATA, N., and M. NEI. 1990. Allelic genealogy under overdominant and frequency-dependent selection and polymorphism of major histocompatibility complex loci. Genetics **124**:967–978.

TAKAHATA, N., Y. SATTA, and J. KLEIN. 1992. Polymorphism and balancing selection at major histocompatibility complex loci. Genetics **130**:925–938.

TAKAHATA, N., and M. SLATKIN. 1990. Genealogy of neutral genes in two partially isolated populations. Theor. Popul. Biol. **38**:331–350.

TEMPLETON, A. R. 1992. Human origins and analysis of mitochondrial DNA sequences. Science **255**:737.

———. 1993. The "Eve" hypothesis: a genetic critique and reanalysis. Am. Anthropol. **95**:51–72.

TEMPLETON, A. R., E. BOERWINKLE, and C. F. SING. 1987. A cladistic analysis of phenotypic associations with haplotypes inferred from restriction endonuclease mapping and DNA sequence data. I. Basic theory and an analysis of alcohol dehydrogenase activity in *Drosophila*. Genetics **117**:343–351.

TEMPLETON, A. R., K. A. CRANDALL, and C. F. SING. 1992. A cladistic analysis of phenotypic associations with haplotypes inferred from restriction endonuclease mapping and DNA sequence data. III. Cladogram estimation. Genetics **132**:619–633.

TEMPLETON, A. R., and C. F. SING. 1993. A cladistic analysis of phenotypic associations with haplotypes inferred from restriction endonuclease mapping. IV. Nested analyses with cladogram uncertainty and recombination. Genetics **134**:659–669.

WATANABE, T., M. SEIKI, H. TSUJIMOTO, I. MIYOSHI, M. HAYAMI, and M. YOSHIDA. 1985. Sequence homology of the simian retrovirus genome with human T-cell leukemia virus type I. Virology **144**:59–65.

WATTERSON, G. A., and H. A. GUESS. 1977. Is the most frequent allele the oldest? Theor. Popul. Biol. **11**:141–160.

WILLS, C. 1992. Human origins. Nature **356**:389–390.

YANAGIHARA, R., N. SAITOU, V. R. NERURKAR, K. SONG, I. B. BASTIAN, G. FRANCHINI, and D. D. GAJDUSEK. 1995. Molecular phylogeny and dissemination of human T-cell lymphotropic virus type I viewed within the context of primate evolution and human migration. Cell. Mol. Biol. **41S**:S145–S161.

YOKOYAMA, S., L. CHUNG, and T. GOJOBORI. 1988. Molecular evolution of the human immunodeficiency and related viruses. Mol. Biol. Evol. **5**:237–251.

ZHAO, T. M., M. A. ROBINSON, S. SAWASDIKOSOL, R. M. SIMPSON, and T. J. KINDT. 1993. Variation in HTLV-I sequences from rabbit cell lines with diverse in vivo effects. Virology **195**:271–274.

ZUCKER-FRANKLIN, D., W. C. HOOPER, and B. L. ERATT. 1992. Human lymphotropic retroviruses associated with mycosis fungoides: evidence that human T-cell lymphotropic virus type II (HTLV-II) as well as HTLV-I man play a role in the disease. Blood **80**:1537–1545.

JULIAN P. ADAMS, reviewing editor

Accepted July 31, 1995