Dating the Origin of the African Human T-Cell Lymphotropic Virus Type-I (HTLV-I) Subtypes

S. Van Dooren, M. Salemi, and A.-M. Vandamme

Rega Institute for Medical Research, Katholiche Universiteit Leuven, Leuven, Belgium

To investigate the origin of the African PTLV-I virus, we phylogenetically analyzed the available HTLV-I and STLV-I strains. We also attempted to date the presumed interspecies transmissions that resulted in the African HTLV-I subtypes. Molecular-clock analysis was performed using the Tamura-Nei substitution model and gamma distributed rate heterogeneity based on the maximum-likelihood topology of the combined long-terminal-repeat and env third-codon-position sequences. Since the molecular clock was not rejected and no evidence for saturation was found, a constant rate of evolution at these positions for all 33 HTLV-I and STLV-I strains was reasonably assumed. The spread of PTLV-I in Africa is estimated to have occurred at least 27,300 ± 8,200 years ago. Using the available strains, the HTLV-II subtype appears to have emerged within the last 3,000 years, and the HTLV-Ia, HTLV-Id, and HTLV-Ie subtypes appear to have diverged between 21,100 and 5,300 years ago. Interspecies transmissions, most probably simian to human, must have occurred around that time and probably continued later. When the synonymous and nonsynonymous substitution ratios were compared, it was clear that purifying selection was the driving force for PTLV-I evolution in the env gene, irrespective of the host species. Due to the small number of strains in some of the investigated groups, these data on selective pressure should be taken with caution.

Introduction

The human and simian T-cell lymphotropic viruses type I (HTLV-I and STLV-I, respectively) share numerous epidemiological, molecular, phylogenetic, and geographical features and are therefore referred to as primate T-cell lymphotropic viruses type I (PTLV-I). Virtually all Old World monkey species from the Cercopithecinae subfamily (Macaca, Cercopithecus, Papio, Mandrillus, Cerczebus and Erythrocebus) and from the Pongidae family (Pongo pygmaeus, Pan troglodytes, and Gorilla gorilla) harbor STLV-I (Lee et al. 1985; Korallnik et al. 1994; Saksema et al. 1994; Fultz et al. 1997; Mahieux et al. 1998; Verschoor et al. 1998). PTLV-I has been associated with both malignant lymphoma and leukemia in humans (Yoshida, Miyoshi, and Hinuma 1982) and nonhuman primates (Miyoshi et al. 1982; Lee et al. 1985; Sakakibara et al. 1986; Tsujimoto et al. 1987; Voevodin et al. 1996). Chronic neurological pathologies (Gessain et al. 1985; Osame et al. 1987) have been described only in humans to date.

According to previously published phylogenetic data, the HTLV-I strains can be classified into six different subtypes: the cosmopolitan HTLV-Ia subtype (Miura et al. 1994, 1997); the Central African subtypes HTLV-Ib (Hahn et al. 1984; Vandamme et al. 1994), HTLV-Id (Mahieux et al. 1997), HTLV-Ie, and HTLV-II (Salemi et al. 1998); and the Australo-Melanesian subtype HTLV-Ic (Gessain et al. 1991; Bastian et al. 1993). Most of these HTLV-I strains are phylogenetically indistinguishable from STLV-I strains. In every genetic region studied, STLV-I and HTLV-I strains from the same geographic origin show a closer phylogenetic relationship than do PTLV-I strains from the same primate species. The only clear separation within the trees inferred from these sequences is found between the African (including the cosmopolitan subtype HTLV-Ia) PTLV-I strains and the Asian-Austronesian strains. The separate and deep-branching pattern of the Australo-Melanesian HTLV-Ic strains and the Asian STLV-I strains more or less according to host genus probably indicates that these strains have undergone a long, independent evolution in their host over a long period. In contrast, there appears to have been a more recent introduction of the PTLV-I virus on the African continent, followed by a spread of this virus, leading to different subtypes. HTLV-I clusters are interspersed with STLV-I strains from different species, suggesting that species and genus barriers have been repeatedly crossed, at least once for each human subtype (Vandamme, Salemi, and Desmyter 1998). Considering the evolutionary inferences, based on these phylogenetic studies, several transmission events between primate genera (Macaca and Papio, Cercopithecus and Papio) and primate “families” (human and simian, Cercopithecus and Pan troglodytes) (Koralnik et al. 1994; Voevodin et al. 1996; Mahieux et al. 1998; Vandamme, Salemi, and Desmyter 1998) must have occurred in the past and are likely still ongoing.

In this study, we succeeded in estimating a time frame for the origin of the African HTLV-I subtypes and of the presumed interspecies transmissions that most probably occurred at the origin of these subtypes based on the currently known HTLV-I and STLV-I sequences. Since African STLV-I strains have been found to cluster tightly with African HTLV-I subtypes, except for HTLV-Ia, we were able to investigate the possible selective pressure due to intra- and interspecies transmission using synonymous versus nonsynonymous substitution ratios.

Materials and Methods

Phylogenetic Analysis

Phylogenetic trees were generated from the multiple alignments (made in Geneworks 2.5.1, Oxford Mo-
l Molecular Systems, United Kingdom) of the long-terminal-repeat (LTR) and *env* regions separately, using neighbor-joining (NJ), maximum-parsimony (mpars), and maximum-likelihood (ML) methods implemented in the software package PAUP*, version 4.0b4a (Swofford 1998). The transition/transversion ratios used were scored using Puzzle, version 4.0 (Strimmer and von Haeseler 1997): 4.44 for the LTR alignment and 5.57 for the *env* analysis. To test the robustness of the NJ and mpars tree topologies, 1,000 bootstrap replicates were performed.

The GenBank accession numbers for the LTR phylogenetic analysis were AF012728–AF012730, AF035538–AF035541, AF045929, AF045931–AF045933, AF054627, AF061438, AF061441, AF061837, AF061838, AF061840, AF061847–AF061849, D00294, D23693, D23694, J02029, J02030, L02534, L36905, L47128, L58023, L60024, L60026, L60302, L60303, L60305, L60306, L60307, L67587, L76032, L76033, L76306, L76307, L76309, L76310, L76312, M33063, M33064, M92845, U12806, U12807, U66376, Y13347, Y16475, Y16481, Y17014, Y17016, Y17017, Z32527, and Z46900. The GenBank accession numbers for the *env* phylogenetic analysis were AF035542–AF035545, AF045928, D00294, J02029, L02534, L36905, L47220, L58023, L60026, L75787, L76032, L76033, L76306, L76307, L76309, L76310, L76312, M33063, M33064, M92845, U12806, U12807, X86376, Y13347, Y16475, Y16481, Y17014, Y17016, Y17017, Z32527, and Z46900. The program SplitsTree, version 2.3f (Huson 1998), was used to generate splits graphs for the LTR-*env* (third-codon-position) data set composed for the molecular-clock analysis. The split decomposition method is a transformation-based approach. Evolutionary data are transformed, or, more precisely, “canonically decomposed,” into a sum of “weakly compatible splits” and then represented by a so-called splits graph. For ideal data, this is a tree, whereas less ideal data will give rise to a treelike network that can be interpreted as possible evidence for different and conflicting phylogenies.

**Molecular-Clock Analysis**

The molecular-clock hypothesis, assuming a constant rate of evolution, was tested on the LTR-*env* data set in Puzzle, version 4.0 (Strimmer and von Haeseler 1997), as previously described (Van Dooren et al. 1998). The most appropriate substitution model for HTLV-I, the Tamura-Nei substitution model with a gamma distribution rate heterogeneity (Salemi, Desmyter, and Vandamme 2000), was used.

**Test for Purifying Versus Positive Selection**

ML trees generated in PAUP*, version 4.0b4a (Swofford 1998), were used in PAML, version 3.0 (Yang 1997), to reconstruct the ancestral sequences at the internal nodes of the tree with BaseML. The synonymous (silent) versus nonsynonymous (amino acid change) substitution distances (*K*$_S$ and *K*$_A$, respectively) were estimated in Dambe according to the Li93 method (Xia 2000) between neighboring internal nodes and between each tip of the tree and the node of its most recent common ancestor. In the Li93 method, sequences are pairwise-compared codon by codon and divided into three categories of sites: zero-fold-, two-fold-, and four-fold-degenerate sites. Transitions and transversions are then scored after applying Kimura’s two-parameter method to correct for multiple hits. The presence of purifying selection is assumed when the *K*$_A$/*K*$_S$ ratio is less than 1, whereas *K*$_A$/*K*$_S$ ratios greater than 1 are evidence for positive selection. The statistical significance of the difference between *K*$_A$ and *K*$_S$ is calculated using a paired *t*-test.


**Results**

**LTR and *env* Phylogenetic Analyses**

Phylogenetic analyses of the LTR region were performed on a 510-nucleotide fragment and separately on a 522-nucleotide fragment of the gp21 of the *env* region using all available known EMBL/GenBank database STLV-I strain sequences (33 strains for LTR and 45 strains for *env*). Since the cosmopolitan HTLV-Ia and the Central African HTLV-Ib are well-established subtypes with good phylogenetic support (Liu et al. 1996; Mahieux et al. 1997; Miura et al. 1997), only a few strains of each subtype, representing the highest divergence within these subtypes, were chosen to illustrate their relationship to the simian strains. All available strains of the other African HTLV-I subtypes (21 HTLV-I strains in total) were also included. The African segment of the tree was rooted using two Asian PTLV-I strains, the HTLV-Ic strain Mel5, and the STLV-I strain TE4.

The topologies of the phylogenetic trees generated by three different methods (NJ, mpars, and ML) were very similar, although the internal branching pattern within some well-defined clusters remained ambiguous. Figures 1 and 2 show the NJ tree of the LTR and *env* regions, respectively, with bootstrap support for the NJ and mpars trees noted on the branches. The separation between the Asian-Austronesian and the African strains was well supported in all analyses (NJ and mpars bootstrap values of >92% and ML *P* < 0.01), implying that the Asian strains are an appropriate outgroup for analyzing the topology of the African part of the tree.

One cluster that was well supported in the LTR region (>94% for NJ and mpars; *P* < 0.01 for ML) but only moderately in the *env* region (>56% for NJ and
Dating the Origin of African HTLV-I Subtypes

**Fig. 1.** Long-terminal-repeat neighbor-joining (NJ) tree of 52 HTLV-I/STLV-I strains. The tree was constructed using PAUP*, 4.0b4a (Swofford 1998), on a 519-nt fragment. Bootstrap support (1,000 replicates) for the NJ (first value) and maximum-parsimony (second value) trees are noted on the branches of the NJ tree. Only the values relevant for the interpretation of the results are given. The strains used are listed in Materials and Methods. All HTLV-I and STLV-I strains are indicated with a symbol according to host species.

HTLV-Ia
- Cosmopolitan

HTLV-Id+
- Gabon/Cameroon

STLV-I

HTLV-Ic
- Indonesia

?
- CAR

HTLV-Ib

?
- Sukhumi

?
- Tanzania

- South Africa

- Gabon

- Kenya

- Gabon

- Africa

- Gabon/Congo

- Africa

- Congo

- Uganda

- Africa

mpars; $P < 0.01$ for ML) contained exclusively human strains: the cosmopolitan HTLV-Ia. All other Central African human strains clustered between and together with African simian strains in four different human subtypes. In addition, several simian clades had no human representative. HTLV-Ib sequences from Central Africa clustered with STLV-I strains from African *P. troglodytes* (with very high bootstrap values for NJ and m pars; $P < 0.01$ for ML). Another Gabonese *Mandrillus sphinx* strain, mnd9, seemed to be almost identical to the described Gabonese strain Lib2, belonging to the newly defined HTLV-Ib subtype.
FIG. 2.—Phylogenetic neighbor-joining (NJ) tree rooted with TE4 and Mel5 of 56 HTLV-I/STLV-I strains constructed in PAUP*, version 4.0b4a (Swofford 1998), using 522 nt of the gp21 env codon region. Bootstrap support (1,000 replicates) for the NJ (first value) and maximum-parsimony (second value) trees are noted on the branches of the NJ tree. Only the values relevant for the interpretation of the results are given. The strains used are listed in Materials and Methods. All HTLV-I and STLV-I strains are classified with a symbol according to host species.

mpars [>99%]; \( P < 0.01 \) for ML). Both strains were closely related to a cluster of STLV-I strains from the Papio genus (NJ: 82%; mpars: 79%; \( P < 0.01 \) in ML for LTR). The African subtype HTLV-Ie, so far containing only one human strain, Efe1, clustered with STLV-I strains from the Cercopithecus genus in LTR and from Papio spp. in env. This clade further clusters within a large group of STLV-I strains from different simian species and genera originating from Congo, Kenya, Tanzania, and South Africa. Although clustering was not well supported by bootstrap analyses, quartet puzzling analysis has shown that clustering with Papio spp. is
Molecular-Clock Analysis

In order to evaluate the clocklike behavior of the African HTLV-I subtypes and their closely related simian strains as indicated in figures 1 and 2, a separate tree was constructed including all STLV-I strains and representatives of the HTLV-I strains for which both the LTR and the env sequences were available. In this way, we obtained a manageable data set with as large a sequence as possible (a 1,031-nt fragment) and as many relevant strains as possible (33 PTLV-I env sequences). The clock hypothesis was rejected when the entire LTR-env data set was used, suggesting a nonconstant rate of evolution among the different HTLV-I subtypes and simian strains (table 1). However, when the LTR data set was combined with only the third codon position (3rdcp) of the gp21 env region (a 683-nt fragment), the molecular-clock could not be rejected, assuming a constant rate of evolution among branches using the LTR-env (3rdcp) data set (table 1). Reestimation of the branch lengths of the LTR-env (3rdcp) ML tree with PAML using a variant of the Tamura-Nei model allowing for different parameters in the LTR and env revealed no statistical difference in the estimates of the branch lengths (data not shown). When the latter data set was tested for substitution saturation using Dambe (Xia 2000), it was clear that no saturation could be observed when transitions and transversions were plotted versus evolutionary distance (fig. 3). The plot shows that transitions and transversions increase linearly with increasing divergence between different PTLV strains. Since transitions occur much more often than transversions, transitions should increase faster than transversions. In the case of substitution saturation, when multiple substitutions have occurred at each site, the phylogenetic signal is essentially lost and its effect is detectable because transversions gradually outnumber transitions. Thus, the graph in figure 3 indicates that no substitution saturation has occurred in the PTLV data set investigated, suggesting the reliability of the dating based on the molecular clock.

To estimate the evolutionary rate and, subsequently, the divergence times on all important nodes of the African part of the LTR-env (3rdcp) tree, one anthropological documented date, supporting a particular node of the tree, is needed. Although no evidence has been found for the presence of simians in Oceania, Indonesian STLV-I strains are both phylogenetically and geographically close to the HTLV-Ic strains. It has been argued that Australo-Melanesian HTLV-Ic, which is found only among non-Austronesian-language-speakers descended from the earliest Melanesian/Australian settlers, most probably arose through interspecies transmission from simians to humans on the migratory pathway of the first humans over Indonesia toward Melanesia (Ibrahim, De Thé, and Gessain 1995; Yamashita et al. 1996; Vandamme, Salemi, and Desmyter 1998). Based on anthropological findings, this migration is dated to around 50,000 ± 10,000 years ago. There is direct archaeological evidence of an early Australian settlement around 40,000 years ago (Cavalli-Sforza, Menozzi, and Piazza 1994) whereas thermoluminescence dating and mitochondrial DNA analysis (Roberts, Jones, and Smith 1990; E. Hageberg, personal communication) suggest migrations as early as 60,000 years ago. Therefore, we assume that HTLV-Ic now found in Australo-Melanesia separated from the rest of the PTLV-I strains around 50,000 years ago, with a 95% confidence interval of 60,000–40,000 years ago.

### Table 1

<table>
<thead>
<tr>
<th>DATA SET</th>
<th>LOG LIKELIHOOD*</th>
<th>2(ln ML – ln MLK)</th>
<th>χ² (0.95) (n – 2 degrees of freedom)</th>
<th>STATISTICAL SIGNIFICANCE LEVELb (P)</th>
<th>REJECTION OF THE MOLECULAR CLOCK</th>
</tr>
</thead>
<tbody>
<tr>
<td>LTR-env (1,031 nt)</td>
<td>−3,932.15</td>
<td>50.78</td>
<td>31</td>
<td>0.0140</td>
<td>Yes</td>
</tr>
<tr>
<td>LTR-env (3rdcp) (683 nt)</td>
<td>−3,062.06</td>
<td>44.83</td>
<td>31</td>
<td>0.0516</td>
<td>No</td>
</tr>
</tbody>
</table>

* ML = maximum-likelihood method; MLK = maximum-likelihood method enforcing a molecular clock using the ML topology.

b Statistical significance of rejecting the molecular clock was calculated using the χ² test.
The accuracy of the estimated evolutionary rate depends not only on the accuracy of the time frame but also on the accurate estimation of genetic distances among taxa and the accuracy of the tree topology. To investigate the robustness of the LTR-env trees, two distance-based tree construction methods (NJ and FITCH) and one character-based method (ML) were used with an experimentally determined transition/transversion bias of 7.02. Although the major African clades themselves remained supported as described above, their branching order was dependent on the method used. This could result either from a true multifurcation, from recombination events in the past, or from a lack of data to resolve the multifurcating tree. It is unlikely that recent recombination events took place, since previous analyses showed no evidence of recombination between different subtypes (Salemi, Desmyter, and Vandamme 2000). However, it cannot be excluded that recombination took place very early, before the establishment of the subtypes. Using the splits decomposition method implemented in the software package of SplitsTree (Huson 1998), networks, indicating conflicting topologies, were especially situated close to the central node of the African part of the tree from which all of the African lineages arise (fig. 4). However, overall, SplitsTree tends to support a multifurcation, and this multifurcation at the origin of PTLV-I in Africa as presented in figure 5 was used in order to deal with these nonconsistent tree topologies. For the same reason, considering the low bootstrap support for the topology within the subtypes in figures 1 and 2, we also represented the divergence within the HTLV-I subtypes as multifurcations in the schematic representation of figure 5.

**Dating the Spread of PTLV-I in Africa**

Only a fully resolved tree (without multifurcations) can be used for the calculations of the divergence times. Therefore, the evolutionary rate with a 95% confidence interval was inferred from the ML method enforcing a molecular clock using the ML topology (MLK) LTR-env (3rdcp) tree, which had the highest likelihood in the likelihood ratio test. Using 50,000 ± 10,000 years ago for the separation of HTLV-Ic, a rate of $1.54 \pm 0.43 \times 10^{-6}$ nucleotide substitutions per site per year for PTLV-I LTR-env (3rdcp) was obtained. Preliminary calculations on the HTLV-I evolutionary rate in a case of HTLV-I familial transmission in Zaire (Liu et al. 1994; Van Dooren et al. 2000) do not contradict the molecular-clock estimate. We found that out of seven infected family members in a three-generation family infected vertically, only one mutation in the LTR gene within one isolate of the second generation could be detected compared with the familial HTLV-Ib strain. All seven iso-
lates had identical env HTLV-Ib sequences. Considering that this mutation was not fixed in the next generation, we could calculate that the higher estimate of the evolutionary rate would be roughly one mutation over 683 nt (LTR-env [3rdcp] sequences) accumulated in 189 years (=sum of each birthday-sampling date) by seven individuals, equal to $1.1 \times 10^{-6}$ nucleotide substitutions per site per year for HTLV-I LTR-env. However, given the short observation time and the low number of observed substitutions, only more families and a full Bayesian approach (not yet established) can give a more precise estimation of the rate with confidence intervals. The current rough estimate, however, already indicates that the calculated rate according to the molecular-clock approach is of the same order of magnitude as and not in contradiction to the observed genetic diversity found in vertically transmitted HTLV-I in a three-generation family.

The branch lengths of the MLK tree and their standard errors were then further used to deduce the divergence times (including 95% confidence intervals) at the nodes of the African part of the tree using the estimated evolutionary rate and its standard error. The calculated dates at the origin of the HTLV-I subtypes are based on the currently known sequences and should be seen as lower limits (meaning true dates may be older). This is because current available strains are subject to sampling bias, and the true divergence within the clades may be higher. The African PTLV-I probably originated at least 27,300 ± 8,200 years ago. Considering the strains that we have, all major African HTLV-I subtypes except one (HTLV-If) probably arose within a range of no more than 10,000 years (see schematic multifurcated tree in fig. 5). The Central African HTLV-Ie virus probably arose first, at least 16,200 ± 4,900 years ago. Later, the other subtypes arose: HTLV-Ia 13,600 ± 5,100 years ago, the HTLV-Ia subtype 12,300 ± 4,900 years ago, and the HTLV-Ib subtype 7,800 ± 2,500 years ago. One exception seems to be the HTLV-If subtype, which has a very recent origin compared with the other African subtypes. Since Lib2 and mnd9 had identical sequences in the LTR-env (3rdcp), only a lower estimate for the HTLV-If origin could be calculated based on these two strains, and true dates of origin for this clade may be older, considering the possible sampling bias. Given an evolutionary rate of $1.54 \times 10^{-6}$ nucleotide substitutions per site per year for an LTR-env (3rdcp) fragment of 683 nt, theoretically only one mutation would occur ev-
Table 2
Nonsynonymous (K_A) Versus Synonymous (K_S) Distances and Their Ratios (K_A/K_S) Studied on Three Different African HTLV-1 Subtypes for the env Sequences

<table>
<thead>
<tr>
<th>STRAINS^a (#S, #AS)</th>
<th>K_A^b</th>
<th>SE</th>
<th>K_S^b</th>
<th>SE</th>
<th>K_A/K_S</th>
<th>&lt;1 or &gt;1?</th>
<th>(P)</th>
</tr>
</thead>
<tbody>
<tr>
<td>HTLV-Ia (39S, 18AS)</td>
<td>0.015</td>
<td>0.040</td>
<td>0.0075</td>
<td>0.0081</td>
<td>0.1971</td>
<td>&lt;1</td>
<td>0.507</td>
</tr>
<tr>
<td>HTLV-Ib/STLV-I (31S, 19AS)</td>
<td>0.0018</td>
<td>0.0017</td>
<td>0.0113</td>
<td>0.0031</td>
<td>0.1601</td>
<td>&lt;1</td>
<td>0.00721</td>
</tr>
<tr>
<td>HTLV-Id/STLV-I (7S, 4AS)</td>
<td>0.0012</td>
<td>0.0014</td>
<td>0.0243</td>
<td>0.0046</td>
<td>0.0511</td>
<td>&lt;1</td>
<td>1.76 × 10^{-6}</td>
</tr>
</tbody>
</table>

^a Number of strains (#S) and number of ancestral sequences (#AS) are shown in parentheses. The accession numbers of the strains used in this table are listed in Materials and Methods.

^b K_A and K_S were calculated using the Li93 method.

^c The statistical significance of the difference between K_A and K_S was calculated using a paired t-test.

ery 1,000 years in that region (evolutionary rate μ = 1.54 × 10^{-6} × 683 = 1.05 × 10^{-3} substitutions per year). Under a molecular clock following a Poisson distribution, P_n(t) = e^{-μt} × (μt)^n/n!, where P_n(t) is the probability of n substitutions in time t with an evolutionary rate μ, the probability of observing no substitutions in t years between Lib2 and mmd9 in the LTR-env (3rdcp) data set is given by P_n(t) = e^{-μt}. Therefore, we can be 95% confident that the two identical sequences in the LTR-env (3rdcp) data set diverged within the last 3,000 years (t = −ln(0.05)/1.54 × 10^{-6} × 683 nt).

Further dating within these human HTLV-I subtypes to calculate certain interspecies transmission events could not be performed due to the ambiguous internal topology of the HTLV-Ic/STLV-I strains within the clades indicated in figure 5. Several separate STLV-I clades, clearly seen in figures 1 and 2, are not represented in figures 4 and 5, since for none of these strains are both the LTR and env region available.

Test for the Selective Pressure Exerted on the env Region

Synonymous and nonsynonymous substitution distances were investigated for all available strains in the env region of HTLV-Ib, HTLV-Id, and their closely related STLV-I strains and for 39 strains representing the HTLV-Ia cluster.

For HTLV-Id, a branch-and-bound search was performed in order to determine the ML tree with PAUP*, version 4.0b4a (Swofford 1998). For HTLV-Ia and HTLV-Ib/STLV-I, a heuristic search was performed because the number of strains was too large for a branch-and-bound search. ML trees were generated in PAUP*, version 4.0b4a, with the subtree-pruning-regrafting (SPR) branch-swapping algorithm on an NJ starting tree. The ancestral sequence topology, reconstructed on these ML trees in PAML, version 3.0 (Yang 1997), was used to calculate the synonymous versus nonsynonymous substitution distances in Dambe (Xia 2000). An interesting observation is that some of the HTLV-I and STLV-I sequences are identical to those of their reconstructed ancestors. The average K_A/K_S ratios, shown in column 6 of table 2, were always less than 1. Moreover, the difference between K_A and K_S was significant (with P < 0.01) for the HTLV-Ib/STLV-I and HTLV-Id/STLV-I groups using a paired t-test. Within the investigated HTLV-Ia clade, the difference between K_A and K_S was not statistically significant. The divergence between these HTLV-Ia strains is probably too small given the amount of K_A and/or K_S distances of 0 and considering the large standard error for K_A and K_S values differing from 0. Although sporadic K_A values were larger than K_S values between specific nodes, no particular selective pressure was observed along human and simian strains and their ancestral sequences. Thus, there is evidence of strong purifying selection within the investigated HTLV-Ib/STLV-I and HTLV-Id/STLV-I groups. In addition, this purifying selection was equally obvious when human and simian strains were compared. Although the K_A/K_S ratio was less than 1 for HTLV-Ia, t-statistics indicated no evidence for negative selective pressure within this group of human strains. However, these results should still be carefully interpreted, since only the HTLV-Ia and HTLV-Ib subtypes had a reasonable amount of strains present in the clades investigated here.

Discussion

The possible human or nonhuman primate origin and the interspecies transmissions of the PTLV-I virus are still a matter of controversy. The current knowledge of the phylogeny of the HTLV-I/STLV-I viruses clearly indicates a clustering of these strains according to geographic origin rather than host species (Vandamme, Salemi, and Desmyter 1998). This suggests interspecies transmissions between simians and humans, which is correlated with their overlapping natural habitats. These observations reinforce the idea that this virus has been repeatedly crossing species and genus barriers in the past and probably continues to do so (Koralnik et al. 1994).

The Asian STLV-I virus seems to have crossed the simian-human barrier only once to give rise to the Oceanian HTLV-Ic subtype, whereas the multifurcated African part of the phylogenetic tree contains five different HTLV-I subtypes (four of which are Central African) with STLV-I strains between and within these HTLV-I clades (see figs. 1 and 2). STLV-I strains from different simian genera (Cercopithecus and Papio; Cercopithecus and Pan) cluster together between these human subtypes, whereas simian strains, usually from a
single simian genus but from more than one species, seem to be at the origin of the African HTLV-I clades. The simian reservoir for HTLV-Ib seems to be *P. troglodytes*. The Mandrillus genus most likely transmitted the virus to humans, giving rise to the HTLV-Id and HTLV-Iphone subtypes (or vice versa?). The HTLV-Ie subtype Efe1 clusters in the LTR region with low bootstrap support together with *Cercopithecus* spp., whereas in the env region Efe1 clusters with a higher bootstrap support with *Papio* spp. This clustering with STL1-I strains from different genera does not reflect conflicting topologies between the LTR and env tree, but, rather, that for none of these strains are both LTR and env known.

It is now clearly established that PTLV-I originated in Asia (Vandamme, Salem, and Desmyter 1998), implying a migration of the host species from Asia to Africa. One could speculate on whether this host species that introduced this virus in Africa is of human or non-human primate origin. Two factors could point to African PTLV-I having human origins: (1) the Oceanian HTLV-Ic subtype is the closest ancestral clade of the African PTLV-I strains (Ibrahim et al. 1995; Mahieux et al. 1998; Salem, Desmyter, and Vandamme 2000), and (2) most HTLV-I subtypes have a simian counterpart, but the point at which the clade branches is exactly as likely among HTLV-I strains as it is among STL1-I strains, suggesting that human-to-simian transmission cannot be excluded. However, it would be more logical to assume a simian origin of the African PTLV-I virus based on primate behavior and phylogenetic data. First, many different simian species are infected with STLV-I. The virus could have been easily transmitted to different simian species and genera through fighting and to humans through hunting, slaughtering, and consuming raw infected simian meat or keeping simians as house pets. Second, although sampling in humans has been more intensive than sampling in simians, there are several simian STL1-I clades with no human representative and only one human clade (HTLV-Ia) with no simian representative (figs. 1 and 2). From this point of view, the simian origin of the HTLV-Ia subtype remains a mystery. Perhaps the simian strains related to HTLV-Ia have not yet been discovered because of the limited sampling in simians thus far.

We tried to elucidate the time frame for the origin of the HTLV-I subtypes and the presumed interspecies transmissions at the origin of these subtypes. Based on a starting date of 50,000 ± 10,000 years ago, representing the first human migrations from Indonesia toward Melanesia and Australia and most probably coinciding with the separate evolution of HTLV-Ic, a time frame of 27,300 ± 8,200 years ago was inferred for the origin of the African PTLV-I. The migration of the PTLV-I host species from Asia to Africa must therefore have occurred roughly between 60,000 and 20,000 years ago. No evidence for simian migrations from Asia to Africa has yet been found for this time period. It has been suggested that human retrograde flows from West Asia to Africa might have occurred between 60,000 and 40,000 years ago (Cavalli-Sforza, Menozzi, and Piazza 1994). Anthropological findings even suggest that the ancestors of the Khoisan originated in East Africa or even earlier in Arabia and are probably related to West Asians. The southward expansion of the Khoisan must have happened during the last Stone Age, between 20,000 and 8,000 years ago. Another possible introduction of HTLV-I in Africa could be due to the anthropologically documented migration of Indo-European populations from Asia to Europe and North Africa that were at the origin of a late Paleolithic culture called Iberomaurusian, flourishing in Spain and North Africa from 20,000 to 7,500 B.C. (Cavalli-Sforza, Menozzi, and Piazza 1994). If one of these migrations coincided with the origin of the African PTLV-I strains, one would have to assume a human origin for the African PTLV-I strains, possibly via an early human-to-simian transmission that then spread among simian species and possibly later was again transmitted to humans. Phylogenetically, it is more plausible that African PTLV-I is of simian origin; however, this would imply a simian migration from Asia to Africa, possibly coinciding with human migrations. Alternatively, perhaps PTLV-I could have been introduced into Africa by another domestic animal genus migrating along with these humans.

Dating the African primate interspecies transmissions at the origin of the HTLV-I subtypes remains difficult. The calculated dates reflect only the origin of the HTLV-I/STLV-I clades based on the common ancestor of the currently known HTLV-I and STLV-I sequences, subject to sampling bias. True divergences within these clades might possibly be higher than the currently observed ones. From our calculations, all different African human HTLV-I subtypes but one (HTLV-Ie) seem to have arisen at least 16,200 ± 4,900 to 7,800 ± 2,500 years ago. The HTLV-Ie subtype seems to have had its origin more recently, less than 3,000 years ago, based on the Lib2 and mnd9 sequences. However, the simian strains clustering between the human HTLV-I subtypes display ambiguous internal branching order with low bootstrap support, making it impossible to calculate dates of single interspecies transmission events. The only clear conclusion to be drawn from these computations is that simian-to-human (or vice versa?) transmissions at the origin of each HTLV-I subtype happened somewhere between 16,200 ± 4,900 to 7,800 ± 2,500 years ago for HTLV-Ia, HTLV-Ib, HTLV-Id, and HTLV-Ie, with probably still ongoing simian-to-human transmissions within these subtypes. The interspecies transmissions for HTLV-Ie occurred more recently, within the last 3,000 years.

Accurate dating is not only dependent on molecular-clock determination, but also on an accurate estimate of the evolutionary rate and on representative sampling. Preliminary investigations on the HTLV-I sequence variation within families in endemic regions are in complete agreement with the rate used here to calibrate the clock (Van Dooren et al. 2000). This suggests that the estimated rate used is reasonable, although the true di-
vergence dates could still be more ancient than the calculated dates as a result of sampling bias.

Synonymous versus nonsynonymous substitution calculations provided evidence for a strong negative selective pressure, thus supporting purifying selection as a driving force for PTLV-I evolution, irrespective of whether the virus is transmitted within its own host species or whether species barriers are crossed. This implies that the third codon position of the env gene has a different influence on the evolution of PTLV-I than the first and second codon position of env, which have a stronger tendency to stay invariant. It can be said that the evolution of the third codon position is much more relaxed. This is completely compatible with the observations of the molecular-clock analyses, in which all investigated strains were evolving at the same rate when considering sequences of the LTR region combined with only the third codon position of the env region, especially since no significant substitution saturation was observed. Although we have only investigated the env region, we have described that PTLV-I strains in endemic populations evolve at equal rates when only the third codon positions of their total genome (tax and rex excluded) are considered (Salemi, Desmyter, and Vandamme 2000). Positive selection would generally be expected when the virus has to cross species or genus barriers. However, this is not what we observe; on the contrary. The observation of a strong purifying selection could in part explain the stability of the PTLV-I genome in general, independent of the host species. The meaning of this phenomenon deserves further research.

Whether the African PTLV-I virus has a human or a simian origin is far from being resolved. Molecular-clock analyses and selective pressure calculations give us only a rough idea about divergence times and selective pressure exerted on PTLV-I. To find the missing link between Asia and Africa, it would be interesting to analyze the phylogeny, evolutionary rate, and selective pressure aspects of more HTLV-I and STLV-I strains from all over the world, including anthropological specimens, and try to correlate them with the anthropological data.

Acknowledgments

We are grateful to Ria Swinnen for fine editorial help. The Rega Institute is part of the HERN concerted action supported by the Biomed program of the European Commission. This study was supported in part by grant 3009984N of the Belgian Foundation for Scientific Research. M.S. was supported by the Research Council of the Kortholieke Universiteit Leuven.

LITERATURE CITED


Dating the Origin of African HTLV-I Subtypes


Pekka Pamilø, reviewing editor

Accepted December 19, 2000