L1 (LINE-1) Retrotransposon Evolution and Amplification in Recent Human History

Stéphane Boissinot, Pascale Chevret,1 and Anthony V. Furano

Section on Genomic Structure and Function, Laboratory of Molecular and Cellular Biology, National Institute of Diabetes and Digestive and Kidney Diseases, National Institutes of Health, Bethesda, Maryland

L1 (LINE-1) elements constitute a large family of mammalian retrotransposons that have been replicating and evolving in mammals for more than 100 Myr and now compose 20% or more of the DNA of some mammals. Here, we investigated the evolutionary dynamics of the active human Ta L1 family and found that it arose ~4 MYA and subsequently differentiated into two major subfamilies, Ta-0 and Ta-1, each of which contain additional subsets. Ta-1, which has not heretofore been described, is younger than Ta-0 and now accounts for at least 50% of the Ta family. Although Ta-0 contains some active elements, the Ta-1 subfamily has replaced it as the replicatively dominant subfamily in humans; 69% of the loci that contain Ta-1 inserts are polymorphic for the presence or absence of the insert in human populations, as compared with 29% of the loci that contain Ta-0 inserts. This value is 90% for loci that contain Ta-1d inserts, which are the youngest subset of Ta-1 and now account for about two thirds of the Ta-1 subfamily. The successive emergence and amplification of distinct Ta L1 subfamilies shows that L1 evolution has been as active in recent human history as it has been found to be for rodent L1 families. In addition, Ta-1 elements have been accumulating in humans at about the same rate per generation as recently evolved active rodent L1 subfamilies.

Introduction

L1 (LINE-1) elements (fig. 1) are mammalian long interspersed repeated DNA elements that replicate by retrotransposition (Voliva et al. 1984; Hattori et al. 1986; Fanning and Singer 1987a; Mathias et al. 1991; Furano 2000). Most L1 copies are defective and accumulate mutations at the pseudogene (neutral) rate (e.g., Voliva et al. 1983; Hardies et al. 1986; Pascale et al. 1993). They have been replicating and evolving in mammals since before the marsupial/placental divergence (~170 MYA) and account for ~14% of the mass of the human genome (Smit 1999). Thus, L1 activity has had a major and perhaps defining effect on the structure and function of modern genomes.

However, an important unresolved issue is the effect of L1 replication on present-day populations. In addition to causing polymorphisms and genetic defects (reviewed in Kazazian and Moran 1998), L1 insertion could also affect the structural (Usdin and Furano 1989; Schwartz et al. 1998; Moran, DeBerardinis, and Kazazian 1999) and regulatory properties of the genome (Wade et al. 1997; Yang et al. 1998). Although some L1 transposition takes place in modern humans, it has been suggested that L1 amplification occurs at such a low rate that the impact on the genome could be considered minor (DeBerardinis et al. 1998). In contrast, there is clear evidence in murine rodents (Old World rats and mice) for novel, rapidly expanding families of L1 elements (Cabot et al. 1997; DeBerardinis et al. 1998; Naas et al. 1998; Saxton and Martin 1998). These novel L1 families are the latest products of a vigorous evolutionary process that has exemplified rodent L1 evolution for at least the last 10 Myr (e.g., Adey et al. 1994; Casavant and Hardies 1994; Cabot et al. 1997; Saxton and Martin 1998; Verneau, Catzeflis, and Furano 1998).

A subset of human L1 elements called the Ta family (Skowronski, Fanning, and Singer 1988) is the source of 11 of 12 de novo L1 insertions identified so far (quoted in Kimberland et al. 1999), and two active elements belonging to this family have been isolated (Dombroski et al. 1991; Holmes et al. 1994). In addition, several other full-length Ta elements were shown to be active in a cell culture–based retrotransposition assay (Moran et al. 1996; Sassaman et al. 1997). However, the evolutionary dynamics of the Ta family has yet to be examined.

Here, we show that the Ta family emerged ~4 MYA, somewhat after the divergence (6 MYA; Goodman et al. 1998) of humans and chimpanzees. Since then, the Ta family has differentiated into two major subfamilies, Ta-0 and Ta-1, each of which spawned additional subsets. Ta-0 is older than Ta-1, and although Ta-0 retains some active elements, Ta-1 now accounts for about one half of the Ta family and has largely replaced Ta-0 as the replicatively dominant subfamily in humans. The youngest subset of Ta-1, Ta-1d, arose about 1.4 MYA and accounts for about two thirds of the Ta-1 subfamily. The extensive differentiation of the Ta L1 family, typified by the emergence of, and eventual replacement by, novel active subsets, recapitulates the active evolution typical of murine L1 elements (Adey et al. 1994; Casavant and Hardies 1994; Cabot et al. 1997; Saxton and Martin 1998; Verneau, Catzeflis, and Furano 1998). In addition, the Ta family has been expanding at about the same rate per generation as the most active rodent L1 families. Thus, human and murine genomes are being altered to about the same extent by L1 activity.

1 Present address: Institut des Sciences de l’Evolution, Case courrier 064, Université Montpellier II, Montpellier, France.

Key words: L1/LINE-1, human, evolution, polymorphism, retrotransposon.

Address for correspondence and reprints: Anthony V. Furano, National Institutes of Health, Building 8, Room 203, 8 Center Drive MSC 0830, Bethesda, Maryland 20892-0830. E-mail: avf@helix.nih.gov.

© 2000 by the Society for Molecular Biology and Evolution. ISSN: 0737-4038
Materials and Methods

Sequence and Phylogenetic Analysis

The BLAST (Altschul et al. 1990) search program was used to find every sequence in the GenEMBL database with a perfect match to a 20mer sequence cognate to different regions of the L1 sequence. For example, to identify all of the Ta elements in the database, we performed the search using the sequence of oligonucleotide 3 depicted in figure 4, which is cognate to the 3′ untranslated region (UTR) A C A motif characteristic of the Ta family. Sequences were initially aligned using the GC G Pileup program (Wisconsin Package, version 10.0, Genetics Computer Group, Madison, Wis.) and refined manually. Phylogenetic analyses were performed using PAUP; version 4.0b2a (Swofford 1998).

Quantification of Ta and Ta-1 Elements

One and two micrograms of human DNA were transferred using a slot-blotter to Bio-Rad Zeta- Probe GT nylon membranes, as well as about 300, 600, 1,200, or 2,400 haploid genomic equivalents of a Ta-1 PCR product mixed with 2 μg of rat DNA. Membranes were hybridized to an oligonucleotide probe cognate to either the Ta (oligonucleotide 3 on fig. 4) or the Ta-1 (oligonucleotide 1 on fig. 4) subfamilies, together with their respective competitor oligonucleotides, as previously described (Verneau, Catzeflis, and Furano 1997). The Ta- and Ta-1-type probes were hybridized overnight at 43°C and 45°C, respectively. Blots were washed three times in 0.3 M NaCl, 0.03 M Nacitrate (pH 7.4). The amount of radioactivity hybridized to each slot was quantified by analyzing the scanned autoradiograms using NIH Image (Wayne Rasband, NIH). In control experiments, we found that oligonucleotide 1 was specific to the Ta-1 subfamily, since it did not hybridize to a cloned Ta-0 element (sequence U09116 in clone JM104-Lre2; Moran et al. 1996).

PCR

PCRs were performed in either an Idaho Technology Air-Thermo Cycler or an MJ Research PTC 100 thermocycler. In both cases, the reactions contained (as suggested by Idaho Technology) 50 mM Tris-Cl (pH 8.3), 2 mM MgCl₂, 0.2 mM of each dNTP, 250 μg/ml bovine serum albumin, 2% sucrose, 0.1 mM Cresol Red (as an electrophoretic dye marker), and 0.5 μM primers. The primers used to determine the phylogenetic distribution of the Ta family are shown on figure 4. The specificity of primers to amplify different L1 sequences was verified with clones of known sequence. Fifty to one hundred nanograms of genomic DNA was amplified in a total volume of 25 μl using the following conditions: denaturation at 94°C for 0 s (30 s in the MJ instrument), primer annealing at 40–50°C (depending on the pair of primers used) for 0 s (30 s), and chain extension at 72°C for 10 s (40 s), for 30 cycles.

We determined the polymorphism of each L1 insert using two PCRs and the primers listed in the appendix: one included the primer pair cognate to the non-L1 flanking sequence, and the second included the primer for the 3′ flank and one specific for a 3′ region of open reading frame (ORF) II (oligonucleotides 2 or 5; fig. 4). All of the human DNAs used in this study were purchased from the Coriell Institute for Medical Research. Nonhuman primate DNAs were gifts from Dr. C. Roos.

Results

The Ta Family Consists of at Least Two Major Subfamilies: Ta-0 and Ta-1

We identified 91 Ta elements in the GenEMBL database (as of August 13, 1999; see appendix). Of these, 18 were obtained in studies designed to select L1 elements (Dombroski et al. 1991; Dombroski, Scott, and Kazazian 1993; Holmes et al. 1994; Sassaman et al. 1997), and 73 were the result of nonselective sequencing. The size distribution of all of the nonselected elements of known size (71 of 73) is shown in figure 1. Interestingly, 34% of the nonselected Ta elements are full-length. This is considerably higher than the previous estimates of 5%–10% full-length elements in primates and mice (Fanning and Singer 1987b), but it is about the proportion of full-length L1 elements in the rat genome (D’Ambrosio et al. 1986). Most (81%) of the partial elements represent simple truncations. However, the remaining 19% contain inversions, some of which contained deletions (see legend to fig. 1). These types of L1 structures have long been noted and are thought to be

FIG. 1.—Ta L1 elements. The top diagram shows a typical full-length human L1 element. The 5′ untranslated region (UTR) has a regulatory function, open reading frame I (ORF I) encodes a 40-kDa nucleic acid–binding protein, ORF II encodes a protein with endonuclease (EN) and reverse transcriptase (RT) activity, and the 3′ UTR contains a conserved G-rich motif (filled rectangle) (reviewed in Furano 2000). Each horizontal line represents one of the 71 Ta elements of known size that were present in the nonredundant human database (GenEMBL) as of August 13, 1999. The filled circle indicates Ta-0 elements (see text). The first (left) vertical line represents the location of two of the nucleotide characters (T, G) that distinguish the Ta-0 and Ta-1 subfamilies, and the second vertical line indicates the location of the ACA motif in the 3′ UTR that is a diagnostic feature of the Ta family (Skowronski, Fanning, and Singer 1988) (see text and fig. 2). The arrows indicate inversions of the corresponding L1 sequence. Some of the inversion-containing inserts also harbor deletions, indicated by the open rectangle. The rectangle represents a member of the youngest Ya5 human Alu (SINE) family (Batzer et al. 1996).
Fig. 2.—Alignment of full-length L1 elements. This alignment includes all of the full-length Ta elements present in GenEMBL (i.e., both the intentionally cloned [selected] L1 elements and those resulting from nonselective [for L1] sequencing). The Ta consensus sequence is shown at the bottom, and only positions at which differences from the consensus are shared between three or more elements are shown. Identity with the consensus is indicated by a dot, and gaps are indicated by dashes. The positions are numbered across the top and the demarcation between the 5' untranslated region (UTR), open reading frame (ORF) I, ORF II, and the 3' UTR is indicated by a vertical gray line. The ORF II
"Ta-1 versus Ta-0"
(T, G) and the 3' UTR ACA characters that correspond to the vertical lines in figure 1 are highlighted in the Ta consensus sequence. Several distinct subsets within either the major Ta-0 or the major Ta-1 subfamilies are outlined in gray boxes (see text). Note, however, that the bottommost pair of Ta-0 sequences (a906677 and u93573) may be the same sequence. These sequences are identical except for the T at position 1154 and share eight differences from all of the other elements shown here. Except for sequence m80340, the distinguishing T at position 1154 is found only in the set of sequences U93562–U93574, reported by Sassaman et al. (1997).
generated during L1 insertion (Rogers 1985; Hutchison et al. 1989; Casavant 1994).

Figure 2 shows an alignment of all 42 full-length Ta elements (24 of the nonselected plus 18 of the selected elements), along with two "pre-Ta" elements (see below) and five ancestral non-Ta elements (Smit et al. 1995). The 150 (of 6,048) positions (numbers given across the top of the fig. 2) at which a character difference from the consensus is shared by three or more elements are presented. We do not show the additional 108 informative positions at which a difference is shared between just two elements, because these additional data do not change any of the conclusions and make inspection of the alignment unwieldy. We also aligned all of the non–full–length Ta elements and found no subsets other than those revealed here (results not shown). The full alignment of all informative positions of the full-length elements is available in the file Boissinot.Ta-align and can be obtained by anonymous ftp from helix.nih.gov in pub/avf.

Since only five ancestral non-Ta elements were included in the alignment, the consensus sequence is that of the Ta elements. The Ta-defining ACA character in the 3' UTR is boxed. There has been little differentiation of the 3' UTR among the Ta elements. Elements ac00632 and ac007043 (arrows in fig. 2) might be considered "pre-Ta" elements because they contain ACg instead of the ACA character. However, these elements clearly belong in Ta, because they share numerous characters with bona fide Ta elements in other regions of the sequence. In addition, at least some of these elements retain activity, as one generated an insert in the factor VIII gene (Kazazian et al. 1988).

The sequence alignment of the 5' UTR, ORF I, and ORF II reveals obvious subdivisions within the Ta family. For example, the Ta family can be divided into two groups based on the nucleotide pair at positions (5557, 5560) of ORF II. One, which we call Ta-1, has (T, G) at these positions, and the second one, Ta-0, almost always has (G, C). Although some ancestral non-Ta ele-

---

**Fig. 3.**—Divergence between members of L1 subfamilies. The timescale was calibrated using a mutation rate per lineage of 0.25%/Myr (see text). **A,** Frequency distribution of the pairwise differences between the full-length Ta-0 or Ta-1 elements shown in figure 2. **B,** Frequency distribution of the pairwise differences between the Ta-1d or Ta-1nd elements shown in figure 2.
Evolution of L1 Retrotransposons in Humans

The Ta family is human-specific. Primer numbers correspond to the boxed sequences at the bottom. The primate phylogeny (branch lengths not to scale) is adapted from Goodman et al. (1998). In addition to humans, we analyzed four hominoids (the chimpanzee \textit{Pan troglodytes}, the gorilla \textit{Gorilla gorilla}, the orangutan \textit{Pongo pygmaeus}, and the gibbon \textit{Hylobates lar}), one Old World monkey, the macaque \textit{(Macaca arctoides)}, one New World monkey, the squirrel monkey \textit{(Saimiri sciureus)}, and one prosimian, the lemur \textit{(Lemur catta)}. The sequence at the bottom shows a region of open reading frame II and the 3' untranslated region of a non-Ta element. The relationship between this sequence and the corresponding region of Ta-1 and Ta-0 elements is also shown. The unlabeled sequence is derived from a non-Ta element (accession number ac004053) that harbors the Ta-1 \((T, G)\) character at positions 5557 and 5560 (see text and fig. 2). Symbols are as in figure 2.

Fig. 4.—The Ta family is human-specific. Primer numbers correspond to the boxed sequences at the bottom. The primate phylogeny (branch lengths not to scale) is adapted from Goodman et al. (1998). In addition to humans, we analyzed four hominoids (the chimpanzee \textit{Pan troglodytes}, the gorilla \textit{Gorilla gorilla}, the orangutan \textit{Pongo pygmaeus}, and the gibbon \textit{Hylobates lar}), one Old World monkey, the macaque \textit{(Macaca arctoides)}, one New World monkey, the squirrel monkey \textit{(Saimiri sciureus)}, and one prosimian, the lemur \textit{(Lemur catta)}. The sequence at the bottom shows a region of open reading frame II and the 3' untranslated region of a non-Ta element. The relationship between this sequence and the corresponding region of Ta-1 and Ta-0 elements is also shown. The unlabeled sequence is derived from a non-Ta element (accession number ac004053) that harbors the Ta-1 \((T, G)\) character at positions 5557 and 5560 (see text and fig. 2). Symbols are as in figure 2.

ments contain \((T, G)\), \((G, C)\) is quite likely the ancestral character at these positions. First, this \((G, C)\) was invariably found associated with the ancestral LPA2 and LPA3 non-Ta L1 families in an earlier survey of primate L1 sequences (Smit et al. 1995; the Ta family was designated LPA1 in this study). Second, a BLAST search of GenEMBL with a 20mer sequence cognate to this region of ORF II returned 500 L1 entries containing the \((G, C)\) ORF-II sequence (500 is the maximal number returned by BLAST) but only 50 L1 entries containing the \((T, G)\) sequence. Of these, 44 were Ta L1 elements, which we here call Ta-1. Third, as figure 3A shows, the \((G, C)\)-containing Ta-0 subfamily is, on the whole, more divergent than Ta-1. Sequence divergence within an L1 subfamily is positively correlated with its age (Pascale et al. 1993; Adey et al. 1994; Casavant and Hardies 1994; Furano et al. 1994; Verneau, Catzeffis, and Furano 1998) Therefore, the results in figure 3A suggest that most Ta-0 elements have resided longer in the genome than have most Ta-1 elements. This would be consistent with the ancestral nature of the \((G, C)\) character.

Figure 2 also reveals additional subdivisions within both the Ta-1 and the Ta-0 subfamilies. A deletion at position 74 of the 5' UTR (and several other characters)
the bottommost Ta-0 subset differ at just one position and may actually be the same sequence; see legend to fig. 2.)

Expansion of the Ta Family in Humans

Given the correlation between sequence divergence and age, we converted the percentage of divergence of the Ta-1 and Ta-0 subfamilies into time by using a molecular clock calibrated from the ~7% average divergence between human-specific and orangutan-specific L1 subfamilies (unpublished data). Assuming that humans and orangutans diverged 14 MYA (Goodman et al. 1998), we obtained a nucleotide substitution rate per lineage of ~0.25% per Myr, which is ~15% to ~60% higher than other estimates of the hominid pseudogene rate (e.g., Casane et al. 1997; Easteal and Herbert 1997). Since this difference would not materially change any of our conclusions, we used the L1-derived rate, as it is based on structurally similar sequences.

We estimated that the Ta family emerged as early as ~4 MYA but that most of the amplification occurred in the last 3 MYr (fig. 3A). We confirmed this recent origin of the Ta family by PCR. Ta family–specific primers generated a product only from humans but not from its closest relative, the chimpanzee, which diverged from humans ~6 MYA (Goodman et al. 1998) (fig. 4). In contrast, primers specific for older primate L1 families generated products in both humans and other primates (fig. 4). The data in figure 3A indicate that the Ta-1 subfamily first arose about 2.5 MYA, with most (75%) of the Ta-1 elements having been generated during the last ~1.6 Myr. In contrast, 80% of the Ta-0 elements had apparently already been inserted before ~1.6 MYA.

Slot blot analysis showed that the haploid human genome contains ~700 Ta elements (fig. 5). At the time of our search, ~13% of the human genome had been sequenced and this portion contained 73 Ta elements (after removal of the entries that had been purposely cloned). This would extrapolate to ~560 elements, which agrees reasonably well with the hybridization data (table 1, cf. columns 2 and 5). Of these 73 “nonselected” Ta elements, 55 (75%) were long enough (fig. 1) to be classified as either Ta-1 or Ta-0. Applying this proportion to the ~700 Ta elements detected by hybrid-

![Figure 5](image_url)

**Figure 5.** Hybridization of Ta- and Ta-1-specific oligonucleotides to genomic DNA. The left side of the figure shows a slot blot hybridized with the Ta-specific probe (oligonucleotide 3, fig. 4), and the right side shows a slot blot hybridized with the Ta-1-specific probe (oligonucleotide 1, fig. 4). In both experiments, the genome equivalents of the Ta-1 PCR product correspond to 2 µg of genomic DNA. These samples and the 0-copy control contained 2 µg of rat DNA. The human DNA used here was from a female Biaka pygmy (Coriell Institute repository number NA10471). Hybridization of oligonucleotide 1 to blots of restricted DNA from a Druze female, a Melanesian female, and a Chinese male showed that they contained about the same numbers of Ta-1 elements as the Biaka pygmy (data not shown).

---

**Table 1**

<table>
<thead>
<tr>
<th>Character</th>
<th>Hybridization</th>
<th>GenEMBL</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Copies</td>
<td>Classified</td>
</tr>
<tr>
<td>Ta</td>
<td>700</td>
<td>525&lt;sup&gt;a&lt;/sup&gt;</td>
</tr>
<tr>
<td>Ta-1</td>
<td>300</td>
<td>265&lt;sup&gt;b&lt;/sup&gt;</td>
</tr>
</tbody>
</table>

<sup>a</sup> Based on the fraction (0.75) of nonselected Ta elements in GenEMBL that extend far enough 5' to be classified as Ta-0 or Ta-1 (see fig. 1). All of the calculated copy numbers have been rounded to the nearest 5.

<sup>b</sup> Based on the fraction (0.88) of the L1 (T, G) character (see text) recovered from GenEMBL present in Ta elements.

<sup>c</sup> Based on the number (73) of nonselected Ta elements present in the approximately 13% of the human genome that was represented in the nonredundant GenEMBL database as of August 13, 1999.

<sup>d</sup> Based on the number (43) of the L1 character (T, G) (after removal of the purposely cloned L1 elements) present in ~13% of the human genome.

<sup>e</sup> Based on the number (55) of nonselected classifiable Ta elements present in ~13% of the human genome.

<sup>f</sup> Based on the number (37) of nonselected Ta-1 elements present in ~13% of the human genome.

---

divides Ta-1 into two groups, Ta-1<sub>d</sub> (no deletion) and Ta-1<sub>d</sub>, and the latter group itself harbors a subset consisting of the first four elements (in the gray box) in the alignment. As was the case for the Ta-1 and Ta-0 subfamilies, the distinction between the Ta-1<sub>d</sub> and Ta-1<sub>d</sub> subsets based on the alignment is also evident by a difference between their divergence; figure 3B shows that Ta-1<sub>d</sub> is considerably less divergent (younger) than Ta-1<sub>d</sub>. In addition, the Ta-0 subfamily also contains several apparent subsets (boxed). The divergence between the members of the top two subsets are, respectively, ~0.5% and ~0.4% less than the ~1.1% divergence of the Ta-0 subfamily on the whole (results not shown and fig. 3A). As we show below, these younger Ta-1 and Ta-0 subsets were supported by phylogenetic analysis. (The pair of sequences [al096677 and u93573] that compose the bottommost Ta-0 subset differ at just one position and may actually be the same sequence; see legend to fig. 2.)
CGACGCGACGACGGTGTAC CGACGCGACGACGGTGTAC

5' UTR without 5' UTR with
1 bp deletion 1 bp deletion

Fig. 6.—The 5′ untranslated region (UTR) of most Ta-1 elements differs from that of Ta-0 elements. We generated a PCR fragment that contained the 5′ UTR of two different Ta-1 insertions, one deleted at position 74 (on the right; from a Ta-1d element, AC004554) and one with the ancestral state, i.e., the absence of deletion (on the left; from a Ta-1nd element, AC004694). The PCR products were blotted and hybridized to oligonucleotide A (see Sassaman et al. 1997). The efficiency of the hybridization of oligomer A to the deleted version of the 5′ UTR is ~5% of the hybridization to the nondeleted version.

Polymorphic Ta Inserts

Figures 2 and 3 show that the Ta family has a distinct subfamily structure and that these subfamilies are of different but somewhat overlapping divergence. If these differences actually reflect different ages (fig. 3), then we would expect that the loci which contain inserts of the less divergent (younger) subfamilies would be polymorphic (for the presence or absence of the L1 insert) compared with the loci that contain inserts of the older L1 subfamilies. For brevity, we refer to these insert-containing loci as polymorphic and fixed L1 inserts, respectively. We designed oligonucleotide primers cognate to the 5′ (F) and 3′ (R) flanking sequences of each of 14 nonselected Ta-0 and 25 nonselected Ta-1 elements present in GenEMBL (fig. 1). Figure 7 summarizes the results of duplicate PCR reactions with these oligonucleotides (one with the F/R pair, the second with R and an L1 (L) oligonucleotide on a single individual from each of eight populations. Whereas 17 (68%) of 25 Ta-1 inserts were polymorphic, only 3 (22%) of 14 Ta-0 inserts were. These results are consistent with the difference between the divergency (age) of the Ta-1 and Ta-0 subfamilies (fig. 3A). Since we sampled only a single individual from each of the eight populations, it is possible that some “fixed” L1-containing loci may be absent in some individuals. However, this would not change the fact that any arbitrarily sampled Ta-1-containing locus is much rarer in humans than an arbitrarily chosen Ta-0-containing locus.

This correlation was even more dramatic when we mapped polymorphic Ta inserts on a phylogenetic tree of the Ta family. Figure 8 shows a single neighbor-joining tree (Saitou and Nei 1987) of all of the full-length Ta elements. Several nodes (circled) within the Ta family are supported by bootstrap values of ~70% or greater. This would correspond to an ~95% confidence level, since the extents (and presumably the rate) of character change are not very different among the members of our data set (Hillis and Bull 1993). One of these nodes groups the Ta-1d elements as a distinct subset within Ta-1, separate from the Ta-1nd elements. The Ta-1d subset is distinctly less divergent (younger) than the Ta-1nd subset (fig. 3B) and 10 of the 11 (91%) Ta-1d elements tested (here or by others, see legend to fig. 8 and appendix) were polymorphic. In contrast, only one of five tested Ta-1nd elements was polymorphic (fig. 8).

Figure 8 also shows that the generally older Ta-0 subfamily contains at least two well-supported subsets of elements. These subsets are young, as their members are quite similar to each other (connected by short branch lengths). This implies that these elements are re-
Fig. 7.—Polymorphic Ta insertions. The electrophoretic patterns of a pair of PCRs for just one locus are shown. All of the Ta insertions were obtained from the database and include both full-length and partial elements. The PCR shown on the top contained primers cognate to the 5' flank (the F primer) and the 3' flank (R), and the one on the bottom contained a primer (L) cognate to either Ta-1 or Ta-0 (oligonucleotides 2 or 5, fig. 4) and primer R. A plus sign indicates the presence of the insertion, and a minus sign indicates the absence of one. In all reactions, we obtained PCR products of the sizes expected for the locations of the PCR primers. Note that a (−/−) result in all of the individuals tested here nonetheless means that the insert-containing locus is polymorphic, because the tested locus does contain an insert in the individual(s) being
Evolution of L1 Retrotransposons in Humans

Fig. 8.—Phylogenetic tree of full-length L1 elements. This tree is a single neighbor-joining tree (using uncorrected p-distances) of all of the full-length elements shown in figure 2 (see Materials and Methods) and is built on the entire sequence exclusive of the 3′ untranslated region (UTR), which showed little or no differentiation (see fig. 2). The number at a particular node indicates its percentage of appearance in 1,000 bootstrap replicates. Only values >50% are indicated. The tree is rooted with a non-Ta L1 element. “F” and “P” represent fixed and polymorphic elements, respectively, as determined either by us (fig. 7) or by others (see appendix). The subsets outlined in gray boxes correspond to those outlined in figure 2. The elements that descend from the node labeled “?S” may be the same (see text and the legend to fig. 2). The “pre-Ta” elements (i.e., those that contain an ACg instead of the ACA in the 3′ UTR; see text and fig. 2) are indicated. The box lists those elements analyzed in earlier studies (Sassaman et al. 1997; Kimberland et al. 1999). These elements are underlined in the tree and listed in the box (by both accession number and previous designation) in the same order as they are on the tree except for the Ta-1nd element, u93572 (L1.25, flanked by “*”). This element clusters with the Ta-0 elements in the neighbor-joining tree even though it lacks several defining features of Ta-0 elements (see fig. 2). Figure 2 shows that u93572 shares a number of nucleotide characters with the Ta-0 element, u93571, with which it groups on the tree. Although some of the shared characters may be due to chance, some may represent regions of common ancestry acquired by recombination (or gene conversion), as has been demonstrated for rodent elements (Hayward, Zavanelli, and Furano 1997; Saxton and Martin 1998). Element af148556 (L1RP) is in intron 1 of the retinitis pigmentosa-2 gene, and af149422 (L1-thal) is in intron 2 of the β-globin gene (see Kimberland et al. 1999 and references therein). The elements in bold type are active in a cell culture retrotransposition assay. The elements flanked by “<” have retrotransposition frequencies of <10% that of the others, and the elements in parentheses contain open reading frames interrupted by a termination codon or a frameshift (or both).

Figure 8 also compares our results with previous studies of Ta elements using the cell culture retrotransposition assay (Sassaman et al. 1997; Kimberland et al. 1999). Although the earlier failure to clone Ta-1d elements (see above) somewhat limits the comparison, there is nonetheless a generally excellent correlation be-

used to generate the human genome database. Samples 10469, 11321, 10975, and 10492 were from males; thus, the X chromosome genotype of these samples is given as /0). The other samples were from females.
between the grouping of an element within a young cluster (i.e., in Ta-1 or the “young” Ta-0 subsets) and whether the element is active in the retrotransposition assay. In particular, those elements that are among the most active in this assay (190088, L1.3; 190902, L1.4; af148856, L1gp) are all Ta-1d elements (Sassaman et al. 1997; Kimberland et al. 1999).

Discussion

Sequence alignment revealed that the human Ta L1 family consists of distinct subsets of elements. We defined two major groups (subfamilies) based on the presence or absence of a number of ancestral nucleotide characters. For example, all members of the Ta-0 subfamily retain the ancestral C at 5560, most retained the ancestral G’s at (4920, 5557) and the ancestral T at 5413, and about one half retained the ancestral A at 2188 and the ancestral G at 2380 (fig. 2). In contrast, none of these characters have been retained by any (or most) of the Ta-1 elements. The presence of the Ta-1 (T, G) character at (5557, 5560) in several non-Ta elements does not invalidate its use as a hallmark of the Ta-1 subfamily. This character was found only about 12% of the time in non-Ta elements, and the ancestral (G, C) character is invariably associated with ancestral primate L1 families (Smit et al. 1995). We are now investigating whether these sites represent mutational hot spots.

Both Ta-0 and Ta-1 harbor additional subsets of elements, and although the various Ta subgroups are clearly distinguished by a number of nucleotide characters, the alignment in figure 2 shows that determining the genealogical relationship between them could be difficult. First, while some characters are exclusively (or largely) confined to a particular subset (e.g., the deletion at position 74 or the T at position 1820), others are shared between several members of two or more subsets: the C at position 155, the G at position 1645, the G at position 2380, the T at position 5131. Furthermore, although the T at position 1820 that distinguishes Ta-1d is an apparent derived character (this T is not present in the five ancestral non-Ta elements), the diagnostic G at position 355 of Ta-1d may be an ancestral character being present in some of the ancestral non-Ta elements. This inconsistent pattern of shared characters and admixture of derived and ancestral characters will confound phylogenetic methods, like maximum parsimony or maximum likelihood, that rely on the pattern of inherited characters. Additionally, all of the Ta sequences are quite similar; most Ta-0 elements are 0.7%–1.1% divergent from most Ta-1 elements. Nonetheless, the neighbor-joining method, which is a distance method and groups sequences based on their overall sequence similarity, did generate some well-supported subsets, including Ta-1d (fig. 8).

The different subsets within the Ta family could also be distinguished by their degree of sequence divergence (fig. 5) and the extent to which their inserts are polymorphic in human populations (figs. 7 and 8). The excellent correlation between the low sequence divergence of a particular Ta subset and the high degree of polymorphism of its members again validates the use of sequence divergence within an L1 subfamily as a measure of its age (Pascale et al. 1993; Adey et al. 1994; Casavant and Hardies 1994; Furano et al. 1994; Verneau, Catzeflis, and Furano 1998). Thus, the differentiation within the Ta family was the result of the successive amplification of different L1 subfamilies over the last ~4 Myr since it first arose.

The distributions of pairwise divergence between members of the Ta-1 and Ta-0 subfamilies (fig. 3A) suggest that the amplification of Ta-1, and particularly Ta-1d (fig. 3B), generally coincided with a decline of transpositional activity of Ta-0 elements, despite the fact that Ta-0 still retains some active subsets (fig. 8). This apparent replacement of a preexisting active subfamily by a more recent one recapitulates the mode of L1 evolution in rats (Cabot et al. 1997; Hayward, Zavanelli, and Furano 1997) and mice (Adey et al. 1994; Casavant and Hardies 1994; Saxton and Martin 1998). The distributions of pairwise divergence shown in figure 3 closely resemble those expected when L1 families are the product of a single lineage of replication-competent elements (the “master model”) (Clough et al. 1996). This model, wherein novel L1 subfamilies belonging to a single dominant lineage are generated successively in time, describes L1 evolution in most of the studied mammalian taxa. Although distinct active L1 subsets may coexist for short periods, a single lineage usually prevails (Rikke, Garvin, and Hardies 1991; Pascale et al. 1993; Adey et al. 1994; Casavant and Hardies 1994; Furano et al. 1994).

The differentiation of a single 3’ UTR lineage into distinct subfamilies that have waxed and waned over the past ~4 Myr describes the evolutionary dynamics of the human Ta family. This has also been observed for a recent rat L1 family (Cabot et al. 1997). Figure 2 shows that the Ta 3’ UTR sequence has hardly changed since it first arose in hominids ~4 MYA. In contrast, significant variation has taken place in all other regions of the Ta family. Selective constraint on the 3’ UTR sequence or adaptive changes in the non-3’ UTR sequence (or both) could account for this difference. An essential role for the 3’ UTR sequence in L1 replication, as has been demonstrated for L1-like elements in insects (Luan and Eickbush 1995; Mathews et al. 1997), could account both for the conservation of the Ta 3’ UTR sequence shown here and for the persistence of certain sequence motifs in the 3’ UTR throughout mammalian L1 evolution (Howell and Usdin 1997). On the other hand, adaptive changes in the non-3’ UTR region could have enabled the element to either evade host repression or gain replicative superiority over existing elements. Whatever the case, competition for replicative dominance could explain the successive replacement of existing L1 subfamilies by novel ones (e.g., fig. 3) and the expansion of one L1 subfamily at the expense of another (Casavant and Hardies 1994; Cabot et al. 1997).

In addition to mimicking the evolutionary dynamics of murine L1 evolution, Ta has also been accumulating at about the same rate per generation as its murine counterparts. The average accumulation (haploid) rate of
the Ta family since it began amplifying in earnest ~3.5 MYA (fig. 3A) is ~0.2 elements per 1,000 years (700 ± 3,500; table 1). The most recent active L1 subfamilies in murine rodents have accumulated ~12 elements per 1,000 years for the L1Rnmlvi2-mn subfamily in *Rattus norvegicus* (~5,600 elements per 450,000 years; Cabot et al. 1997) and ~5 elements per 1,000 years for the mouse *Mus musculus* Tf subfamily (~1,825 elements per 325,000 years, the average of three published determinations; DeBerardinis et al. 1998; Naas et al. 1998; Saxton and Martin 1998). Because the impact of L1 amplification would be related to the number of elements accumulating per generation, we normalized the accumulation rates using generation times of ~20 years for humans and ~0.5 years for rodents. The normalized accumulation rates are ~0.006 elements per generation in rats and ~0.0025 elements per generation in mice, as compared with ~0.004 elements per generation in humans. These accumulation rates reflect the rate of both L1 transposition and L1 elimination by either selection (in the case of deleterious or lethal insertions) or genetic drift. Although these factors could be very different between these species, our results suggest that L1 transpositional activity might play equally important roles in the genetic diversity and evolution of both human and rodent genomes.

About 90% of the Ta-1d and a smaller percentage of Ta-1nd and Ta-0 insertions are polymorphic and could be useful for gene mapping or population genetics studies. The number of polymorphic inserts due to Ta-1d alone may well number several hundred in world human populations. L1 insertions have several advantages over commonly used polymorphic markers such as microsatellites and single-nucleotide polymorphisms. Parallelism (i.e., independent retrotransposition into the same chromosomal location) is unlikely, and the ancestral state of the polymorphism is known (i.e., the absence of inserts). Therefore, Ta insertions, like Alu insertions, could be used as robust markers to root population trees (e.g., Batzer et al. 1994; Novick et al. 1998) or allele phylogenies (e.g., Hammer 1994).

**Conclusions**

Evolutionary analysis of the Ta L1 family has shown that the tempo and mode of L1 evolution in recent human history recapitulates the process in murine rodents and that L1 activity may have affected the genomes of these taxa similarly. Furthermore, by identifying the most recently active human L1 subfamily(s), we now might possibly identify those factors responsible for its replicative success. To this end, we recently isolated every member of the Ta-1 subfamily and can compare such parameters as the genetic environment and the regulatory and enzymatic properties of various members of the Ta-1 subsets. Finally, by identifying the entire insertional history of the Ta-1 subfamily in several human populations, we may be better able to estimate the effect of this most current wave of L1 replication on present-day humans.

**Acknowledgments**

We thank C. Roos for the primate DNA samples and Dr. John Moran for clone JM104-Lre2.
## APPENDIX

### Appendix. Ta element list.

<table>
<thead>
<tr>
<th>Sub-family</th>
<th>Genbank</th>
<th>strand</th>
<th>Start</th>
<th>End</th>
<th>Chr</th>
<th>Size</th>
<th>Poly</th>
<th>Forward primer</th>
<th>Reverse primer</th>
</tr>
</thead>
<tbody>
<tr>
<td>Ta-nclass.</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Ta-0</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Ta-1</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

### Data Table:

The table above contains information regarding Ta elements, including their Genbank accession numbers, strand orientation, start and end positions, chromosome (Chr) location, size, poly(A) tail, and the forward and reverse primers used for their detection.

### Ta-nclass.

- **Ta-unclass.**
  - **AC004499**
  - **AC004911**
  - **ZT2519**
  - **AC002888**
  - **AC004865**
  - **Ap000084**
  - **AC006465**
  - **LO5367**
  - **Z80899**
  - **AC005530**
  - **AC006222**
  - **AC002069**
  - **AI921101**
  - **AC005105**
  - **AL023284**
  - **Z84814**
  - **AC002122**

### Ta-0.

- **AL030998**
- **AC00491**
- **AC004220**
- **AL020099**
- **AC004784**
- **AC007846**
- **UC1971**
- **AC006986**
- **AC005908**
- **AC006870**
- **AC003808**
- **AL096877**
- **AC004918**
- **U09116**
- **AL096799**
- **U93569**
- **AC002276**
- **U09115**
- **AC004053**
- **AL078632**
- **U93571**
- **U93564**
- **AP14922**
- **U93562**
- **U93563**
- **U93565**
- **U93567**
- **U93570**
- **U93573**

### Ta-1.

- **AC005739**
- **AB20867**
- **Z84814**
- **AL031586**
- **AC006494**
- **AC006027**
- **U93568**
- **MB0343**
- **X501188**
- **AC004142**
- **AC007918**
- **AL049842**
- **AC003100**
- **AC008467**
- **AC005900**
- **AC004883**
- **AC003689**
- **AL02166**
- **AC004000**
- **AC005885**
- **AC004200**
- **AC002135**
- **AC004554**
- **AC005939**
- **AC007556**

The table includes a range of Ta elements from different sub-families, each with its specific Genbank accession number, strand orientation, start and end positions, chromosome location, size, poly(A) tail, and the specific forward and reverse primers used for their detection.
Appendix. Ta element list.

<table>
<thead>
<tr>
<th>Sub-family</th>
<th>Genbank</th>
<th>strand</th>
<th>Start</th>
<th>End</th>
<th>Chr</th>
<th>Size</th>
<th>Poly</th>
<th>Forward primer</th>
<th>Reverse primer</th>
</tr>
</thead>
<tbody>
<tr>
<td>L19088</td>
<td>+</td>
<td>1</td>
<td>6022</td>
<td>14</td>
<td>6022</td>
<td>YES</td>
<td>(Dombroski et al. 1993)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>L19092</td>
<td>-</td>
<td>1</td>
<td>6022</td>
<td>9</td>
<td>6022</td>
<td>YES</td>
<td>(Dombroski et al. 1993)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>ACU07538</td>
<td>-</td>
<td>133547</td>
<td>139570</td>
<td>X</td>
<td>6023</td>
<td>YES</td>
<td>TGGAACCTCTGGCCAATTCT</td>
<td>AAGTTAGCAGCACATTACC</td>
<td></td>
</tr>
<tr>
<td>ACU02990</td>
<td>-</td>
<td>11716</td>
<td>17741</td>
<td>X</td>
<td>6025</td>
<td>YES</td>
<td>GATGAGAGCCTGCTGATTTT</td>
<td>TTTATGACAGACAGTTT</td>
<td></td>
</tr>
<tr>
<td>AP54599</td>
<td>-</td>
<td>74</td>
<td>586</td>
<td></td>
<td>512</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td></td>
</tr>
<tr>
<td>S78214</td>
<td>+</td>
<td>61</td>
<td>580</td>
<td>?</td>
<td>519</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td></td>
</tr>
<tr>
<td>M23332</td>
<td>+</td>
<td>1</td>
<td>681</td>
<td></td>
<td>681</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td></td>
</tr>
<tr>
<td>U08211</td>
<td>-</td>
<td>738</td>
<td>1511</td>
<td>X/Y</td>
<td>773</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td></td>
</tr>
<tr>
<td>ACU08110</td>
<td>+</td>
<td>89047</td>
<td>91737</td>
<td>?</td>
<td>2690</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td></td>
</tr>
<tr>
<td>ACU00100</td>
<td>-</td>
<td>100762</td>
<td>103717</td>
<td>19</td>
<td>2955</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td></td>
</tr>
<tr>
<td>AG02153</td>
<td>+</td>
<td>105741</td>
<td>10924</td>
<td>X</td>
<td>3463</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td></td>
</tr>
<tr>
<td>M23333</td>
<td>+</td>
<td>1</td>
<td>3785</td>
<td></td>
<td>3785</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td></td>
</tr>
<tr>
<td>Z95325</td>
<td>+</td>
<td>9818</td>
<td>14458</td>
<td>X</td>
<td>4640</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td></td>
</tr>
<tr>
<td>Z98754</td>
<td>-</td>
<td>49938</td>
<td>55954</td>
<td>X</td>
<td>6016</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td></td>
</tr>
<tr>
<td>AL078622</td>
<td>-</td>
<td>27243</td>
<td>33261</td>
<td>22</td>
<td>6018</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td></td>
</tr>
<tr>
<td>API145856</td>
<td>+</td>
<td>1</td>
<td>6019</td>
<td>X</td>
<td>6019</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td></td>
</tr>
<tr>
<td>U93566</td>
<td>+</td>
<td>1</td>
<td>6019</td>
<td></td>
<td>6019</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td></td>
</tr>
<tr>
<td>U93572</td>
<td>+</td>
<td>1</td>
<td>6020</td>
<td>?</td>
<td>6020</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td></td>
</tr>
<tr>
<td>M0340</td>
<td>+</td>
<td>1</td>
<td>6021</td>
<td>14</td>
<td>6021</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td></td>
</tr>
</tbody>
</table>

Ta-unclass., elements that did not extend far enough 5' to be classified as Ta-0 or Ta-1. Chr is the chromosomal location of the insert. Size is the length of the insert in bp; ? GenBank entry was truncated in the L1 sequence. Poly indicates whether the insert-containing locus is polymorphic; -, not determined.

LITERATURE CITED


Thomas Eickbush, reviewing editor

Accepted February 22, 2000