Rodent L1 Evolution Has Been Driven by a Single Dominant Lineage That Has Repeatedly Acquired New Transcriptional Regulatory Sequences


Department of Microbiology and Immunology, University of North Carolina at Chapel Hill

All mammalian genomes contain approximately 100,000 copies of the transposable element LINES-1 (L1). Phylogenetic analysis indicates that the L1 progenitor predates the mammalian radiation; since that time, the open reading frames encoded in L1 have evolved under selection. The least conserved regions within L1 are the 5'-terminal transcriptional regulatory sequences. In rodents, four types of L1 elements (A, F, and V from mouse and R from rat) have been defined according to the type of apparently nonhomologous promoter sequence present at the 5' end. In this study, we investigate the relationships between these four types of promoters. DNA sequence was determined from approximately 1.5-kb regions from the 5' ends of seven F- and three V-type L1 elements. These sequences were aligned with 29 previously reported L1 elements. Phylogenetic analysis was then performed on the homologous regions of the alignment. The results indicate that in mouse all of the A-, F-, and V-type elements belong to a single dominant lineage but were inserted into the genome during different time periods; V-type elements are the oldest, while A-type elements are the most recently inserted. V-type elements also appear ancestral to the R-type elements found in rat and therefore were replicatively competent prior to the divergence of rat and mouse. Analysis of sequence identity indicates that the different 5' promoters did not derive from a common ancestor. Therefore, the dominant L1 lineage appears to have acquired novel promoter sequences from non-L1 sources. Transposable elements from a wide range of species show similar structural rearrangements, suggesting that acquisition of new sequences may be a common theme in their evolution.

Introduction

LINES-1 (L1) is an interspersed family of repetitive sequences that make up at least 10% of the mammalian genome (Singer 1982; Burton et al. 1986; for reviews see Rogers 1985; Edgell et al. 1987; Hutchison et al. 1989; Martin 1991a). Full-length L1 elements are 6-7 kb in length, but most are truncated at variable distances from a conserved 3' end. Nearly all L1 elements end with a polyadenylation signal followed by an A-rich region. The entire L1 element is bounded by short (5-15-bp) direct repeats (fig. 1). The L1 consensus sequence and some of the younger elements contain two long open reading frames (ORFs). ORF-2 encodes a protein with reverse-transcriptase homology (Hatori et al. 1986; Loeb et al. 1986; Scott et al. 1987; Mathias et al. 1991); ORF-1 may encode a structural protein (Martin 1991b). These features suggest that L1 is a mammalian transposable element that replicates via an RNA intermediate; recently these predictions have been confirmed (Dombroski et al. 1991; Evans and Palmiter 1991). The favored model of L1 replication involves reverse transcription of a polyadenylated L1 transcript and reinsertion into a new genomic location (Singer 1982; Voliva et al. 1984; Martin et al. 1985; Rogers 1985; Hutchison et al. 1989; Luan et al. 1993). L1 insertions occasionally cause mutations in humans by insertional mutagenesis (Kazazian et al. 1988).

L1 transcripts probably originate from a transcriptional regulatory sequence (promoter) located at the 5' end of the element. Four different types of promoters have been classified in rodent L1s: the A-type (Loeb et al. 1986; Severynse et al. 1992; Schichman et al. 1993), the F-type (Wincker et al. 1987; Padgett et al. 1988; Adey et al. 1994), the putative V-type (Jubier-Maurin et al. 1992), and the rat 5' promoter (Furano et al. 1988; Nur et al. 1988), which we term the "R-type." No ho-
mology exists between these promoters or those from human or rabbit L1 (Scott et al. 1987; Furano et al. 1988; Price et al. 1992). However, all of the L1 promoter sequences are G+C rich CpG dinucleotide islands, features shared with mammalian housekeeping-gene promoters (Loeb et al. 1986; Fanning and Singer 1987; Furano et al. 1988). The R-, A-, and F-type promoters exist as tandemly repeating arrays of monomers. The R-type monomer is about 600 bp long, and the A- and F-type monomers are both about 200 bp long.

Phylogenetic analysis indicates that L1 has existed for many millions of years and that L1 elements are still actively replicating in the genome (Martin et al. 1985; Hardies et al. 1986). These studies also suggest that only a small subset of L1 elements produce progeny, whereas most elements never replicate. The replicatively competent subset of elements has been termed the "molecular drivers" or "master genes," each of which gives rise to a lineage of elements (Martin et al. 1985; Hardies et al. 1986; Hardies and Rikke 1989; Deininger et al. 1992). In the short term, a number of lineages appear to "compete" to populate the genome. All but one lineage eventually dies out; the successful lineage then gives rise to new competing lineages. Therefore, over the long term, a single lineage of L1 elements exists in a given species (Martin et al. 1985; Hardies et al. 1986; Hardies and Rikke 1989; Deininger et al. 1992; Schichman et al. 1992).

The L1 lineage competition model was derived from phylogenetic studies of L1 elements containing the same type of promoter or from fragments of L1 elements of unknown promoter type. We wanted to further examine this model by extending the phylogenetic analysis to L1 elements with different types of 5' promoters. Specifically, we sought to determine whether A-, F-, V-, and R-type elements represent different lineages that coexist in the genome or whether they represent L1 elements that derived from the same lineage during different time periods. To address this issue, we determined the sequence of the 5' ends of seven F- and three V-elements, aligned their homologous regions with 29 previously reported L1 monomers from the same type of promoter or from fragments of L1 elements of unknown promoter type. The resulting analysis indicates that, in the mouse, A, F, and V belong to the same lineage but were inserted in the genome or whether they represent L1 elements that derived from the same lineage during different time periods. Our phylogenetic analysis also indicates that both F- and R-type elements evolved from V-elements by the acquisition of new promoters. These results greatly strengthen previous suggestions that L1 is capable of acquiring new 5' promoter sequences (Wincker et al. 1987; Furano et al. 1988; Padgett et al. 1988).

**Material and Methods**

Source of L1 Clones Used in This Study

All 30 A- and F-type L1 clones were isolated from mouse. Twenty-one—(L1Md) A1, A3, A4, A5, A7, A8, A10, A11, A12, A14, A15, A19, A21, F1, F2, F3, F11, F13, F14, F16, and F18—were isolated from the same genomic library by selection of plaques that hybridized to at least two probes, one from the first ORF and the other from either the A- or the F-type monomer. The construction and screening of this library are described in detail elsewhere (Schichman et al. 1992). In brief, mouse F9 teratocarcinoma cell DNA was randomly sheared by sonication and end-repaired, and 1.5-kb fragments were isolated by electrophoresis, ligated into M13mp18, and transformed into DH5αF'-cells. (Although these host cells are methylation restrictive, it is unlikely that a strong bias exists against sequences containing CpG dinucleotides, because many CpG-rich clones were readily obtained.) Twenty-nine A- and F-type L1 elements are referenced elsewhere: (L1Md) A2 (Loeb et al. 1986); (L1Md) A13 and 9 (referred to here as "A") (Shehee et al. 1987); (L1Md) A1, A3, A4, A5, A7, A8, A10, A11, A14, A15, A19, and A21 (Schichman et al. 1992); (L1Md) 14RH-1 and 14LH (Begg et al. 1988); (L1Md) A12 and F13 (Aday et al. 1991b); (L1Md) F15 (Aday et al. 1991a); (L1Md) IG-2 and IG-6 (Wincker et al. 1987); and (L1Md) BsNI (Mottez et al. 1986). The accession numbers of the seven F-type L1 elements first described here are (L1Md) F1, M93314; F2, M93315; F3, M93316; F11, M93317; F13, M93318; F16, M93319; and F18, M93320.

Two V-type clones—(L1Md) 1G-5 and (L1Md) 1G-8 (Jubier-Maurin et al. 1992)—were isolated from mouse. The three other V-type L1 clones—(L1Me) V1, (L1Pd)V2, and (L1Mp) V3 (the accession numbers are, respectively, L18786, L18787, and L18788)—were isolated from Mastomys erythroleucus (Me), Praomys daltoni (Pd), and Mus pahari (Mp) genomic libraries (tissue and genomic DNA were a gift from Francois Catzeflis, Institute de Biologie, Montpellier, France). These libraries were constructed as follows: Partial digestions were performed with a mixture of AluI, HaeIII, and RsaI restriction enzymes, and 0.9-1.4-kb fragments were isolated, ligated into M13mp18, and then transformed into DH5αF'-cells. L1 clones were isolated using a probe from L1MdA2 (Loeb et al. 1986) that spanned the last 50 bp of the 5' untranslated region and the first 150 bp of ORF-1. Mastomys, Praomys, Rattus,
and *Mus* are different genera in the subfamily Murinae (for the phylogenetic tree, see Pascale et al. 1990). For additional information concerning the distribution of A-, F-, V-, and R-type L1 elements, see the work of Jubier-Maurin et al. (1987, 1992) and Pascale et al. (1990, 1993). All R-type (our terminology) clones were isolated from *Rattus norvegicus* (rat) and thus are termed "L1Rn: (L1Rn) L3" (D'Ambrosio et al. 1986), "(L1Rn) Lsse" (Soares et al. 1985), and "(L1Rn) B6" and "(L1Rn) B7" (Furano et al. 1988).

DNA Sequencing

The DNA sequence was determined on both strands, and all ambiguities were resolved. Both single- (Bankier et al. 1987) and double-stranded templates (Sequenase, version 2.0; United States Biochemicals) were used. Individual gel readings were assembled using the STADEN programs DBUTIL and DBAUTO. Sequence alignments were done using the GCG programs LINEUP and PRETTY (Devereux et al. 1984). Because of the similarity between rodent L1 sequences, no special alignment protocol was necessary. To align the rabbit and human L1 sequences, the GCG programs COM- PARE, DOTPLOT, and GAP were used.

Phylogenetic Analysis

The phylogenetic tree shown in the left-hand panel of figure 3 was generated by parsimony analysis using the PAUP package, version 3.0s (Swofford 1990). The following conditions were used to run PAUP: CpG positions that showed hypermutability to CpA or TpG were excluded, additional informative positions were added to indicate shared insertion or deletion mutations (except for the ORF-1 length polymorphic region), and character types were weighted so that transversion or length differences are twice the weight of transitions. The tree shown is a strict consensus of the most parsimonious trees determined by branch swapping using the heuristics option. Bootstrap analysis (100 cycles) using heuristics was performed to assess the reliability of the tree. Inclusion of all the highly similar A-type elements (*BstN1*, A2, A3, A5, A11, A13, A15, and A19) greatly increased run times. Because these elements share greater than 99% sequence identity with one another, it is quite certain that they cluster together on the tree. Therefore, only one of the highly similar A-type elements (A2) was used to position the entire subset. A tree containing this subset of elements plus additional elements from the main tree was calculated independently, the branch lengths were adjusted to be proportional to those on the main tree by using clusters of elements present in both trees, and then this subtree was joined to the main tree.

The phylogenetic tree shown in the right-hand panel of figure 3 was generated by neighbor-joining analysis using the programs PSFIND, NJJOIN, and NJBOOT (Thomas Whittam, Pennsylvania State University) with the Jukes-Cantor option and ignoring missing data. Bootstrap analysis (100 cycles) was performed using NJBOOT. Maximum-likelihood analysis was performed using DNAML in the PHYLIP package version 3.2 (Joseph Felsenstein, University of Washington), with transversions weighted twice as much as transitions, the frequency of bases taken from the input sequences, and each input file run five times with a different seed.

Percent Sequence Identity

The percent identity between sequences was calculated using the GCG program DISTANCES. The program was run twice without a denominator (option 5). In the first run, the default symbol-comparison table was used to determine the number of identical bases. In the second run, a symbol-comparison table was used in which matches and mismatches between all bases were given a value of 1.0 in order to determine the number of total positions. The total number of identical bases was then divided by the total number of positions. The percent identity between silent sites was determined using the GCG program DIVERGE. Only category 3 positions were considered. These are defined as positions at which any nucleotide change does not change the encoded amino acid, and, therefore, differences in rates of transitions and transversion should have no effect. Corrections for multiple mutations at the same nucleotide position were done using the formula $b = -\frac{1}{2} \ln (1 - 4/3a)$, where $a$ is the fraction of nonidentical bases and $b$ is the corrected fraction of nonidentical bases (Jukes and Cantor 1969).

In the Results section, the following statement is made: "F-elements are slightly more divergent from A-elements than they are from each other, yet similarity is readily apparent among F-monomers but not between A- and F-monomers." In this argument, the A-type group 3 elements are compared with the F-type elements (average identity 85%, which is 15% divergence) because they were all inserted at approximately the same time and thus have been subject to comparable mutational rates. The F-type groups 1 and 2 are compared with the F-type group 3 and 4 (average divergence 13%) to indicate that homology is still readily apparent between two diverged subsets of F-elements. Divergence between F-type group 1 and 2 monomers and F-type group 3 and 4 monomers is 24% (this is nearly twice the rate observed in the 5' untranslated region). Therefore, if A-monomers evolved from F-monomers, they should still share at least 70% sequence identity. The cause of the accelerated divergence in the monomer region is not known, but it is not solely the result of CpG hypermutability; removal of the mutated CpG dinucleotide...
positions increased the sequence identity an average of just 2%.

Results

Comparison of the 5' Termini from A-, F-, V-, and R-type Ll Elements

The 39 rodent A-, F-, V-, and R-type Ll elements used in this study are shown schematically in figure 1. An Ll element can be divided into the G+C-rich promoter sequences located at the 5' end and the remaining body sequences. As mentioned, A-, F-, V-, and R-type promoters share no significant sequence similarity. However, the body sequences from A- and F-type elements are homologous throughout. Sequence similarity between A- (or F-), V-, and R-elements begins a few bases 5' of the first methionine codon in ORF-1 (fig. 2) and extends to the 3' end of the elements. Although the 5' untranslated sequence from R-elements shares no similarity with A- or F-elements, this sequence is also approximately 200 bp long. The 5' structure of V-elements is less clear. It is difficult to determine where the body begins, since the few elements characterized do not contain a tandemly repeating motif at the 5' terminus. Unfortunately, the 5' sequences from the V-elements reported here are truncated because of cloning procedures, so we cannot address this issue. However, enough sequence was obtained to show that V1, V2, and V3 are indeed V-type elements in that they show sequence homology in the 5' untranslated region with the V-type defining sequences 1G-5 and 1G-8 (fig. 2).

Phylogenetic Analysis of Ll Elements

The relationships among the 39 Ll elements shown in figure 1 are summarized by the phylogenetic trees in figure 3. The lack of sequence homology at the 5' end of the Ll elements, as well as additional missing sequence information due to natural deletions or cloning truncations, necessitated the use of different amounts of sequence data from different Ll elements. The phylogenetic tree shown in the left-hand panel of figure 3 was generated by maximum-parsimony analysis from an alignment containing both the 5' untranslated sequences and the ORF-1 sequences from the A- and F-elements and just the ORF-1 sequences from the V- and R-elements (the 5' untranslated sequence is 205 bp long, and ORF-1 is 1,172 bp long). The root position was determined in a separate analysis using sequences from the 3' half of ORF-1 (see below). Because of the complexity of this data set, the phylogenetic analysis was repeated on a number of different regions of the alignment by using different subsets of input sequences from each region. In addition, three different phylogenetic analysis methods were employed: parsimony, neighbor joining, and maximum likelihood (for additional details, see Material and Methods).

The first data set used for phylogenetic analysis included the 5' untranslated sequence and the entire ORF-1 sequences from the A- and F-elements. The phylogenetic tree generated by parsimony analysis (not shown) was entirely consistent with the phylogenetic tree shown in the left-hand panel of figure 3. A very similar phylogenetic tree was generated by neighbor-joining analysis (fig. 3, right). An interesting difference is that the neighbor-joining tree indicates that the group 3 F-elements are a diverging sublineage, a result weakly supported by maximum-likelihood analysis (not shown). This data set was then divided into three subregions (subregion 1 = 5' untranslated region plus the first 180 bp of ORF-1; subregion 2 = polymorphic region plus 3' 50 bp; and subregion 3 = remaining 3' sequences of ORF-1) and phylogenetic trees derived from each region. The resultant subtrees (not shown) are quite similar to one another and to both trees shown in figure 3, except for a few significant differences involving the placement of A1, A14, and A21. Manual analysis of informative positions (not shown) strongly suggests that these differences are the result of legitimate recombinations between A- and F-elements in which older F-type elements have been fused to younger A-type 5' termini. Interestingly, the recombination break point in each of the three elements occurs near the ORF-1 length-polymorphism region (discussed below). Removal of these three elements from the analysis did not alter the relative arrangement of the remaining elements.

The second data set used for phylogenetic analysis included only the ORF-1 sequences from the V-, R-, F-, and the older A-elements. The results of parsimony analysis are consistent with the tree shown in the left-hand panel of figure 3. Parsimony analysis of different subsets of elements from this data set also produced consistent trees (not shown). Unfortunately, neighbor-joining analysis of this data set produced a tree that was inconsistent with both the neighbor-joining tree described above and the parsimony tree. When neighbor-joining analysis was performed on different subsets of elements from this data set, the resultant trees were inconsistent both with one another and with the tree shown in the right-hand panel of figure 3. Therefore, neighbor-joining analysis of this data set is unreliable and not presented. Maximum-likelihood analysis was unable to resolve the discrepancies.

Determination of the root of the phylogenetic tree proved difficult, and therefore less confidence should be placed in the result. Homology between the rodent Ll elements and the human (112) and rabbit (O5) Ll outgroups is weak and exists only in the 3' end of ORF-1,
FIG. 1—a), Schematic of a typical mouse full-length L1Md element (derived from laboratory strains of mice that are hybrids of *Mus domesticus* and *M. musculus*). Indicated are the short direct repeats (SDR) that define the ends of the element, the 5' and 3' untranslated sequences (5'UT and 3'UT), the A-rich region at the 3' end, and the two ORFs (ORF-1 and ORF-2). The element can be divided into the G+C-rich tandemly repeated promoter sequences located at the 5' end (known as “monomers”), and the remaining body sequences. b), Schematic of the 39 L1 clones analyzed in this study. Lines connecting sequence blocks indicate gaps inserted to align homologous sequences. The length of the polymorphic region in ORF-1, defined by Schichman et al. (1992), is indicated; the first repeat unit is redefined as a 66-bp repeat instead of a 42-bp repeat, by inclusion of the upstream 24 bp. Thinner hatching denotes an A-type monomer; thicker hatching denotes an F-type monomer; horizontal lines denote a V-type 5' terminus; cross-hatching denotes an R-type (rat) monomer; vertical lines denote a B1 element (probably inserted into this location in the genome); a rounded end denotes a genomic flanking sequence that exists in the clone; a squared end denotes a clone truncated because of the cloning procedure; X = frameshift relative to L1MdA2; and T = termination codon in ORF-1 after correction for frameshift mutations. Parentheses denote regions that are present in a clone but that were not sequenced. Triangles above sequence blocks indicate non-L1 insertions. Pattern identity indicates sequence homology. The composite L1Md elements are indicated with a superscript "c" (Adey et al. 1991b). The F-monomer sequence located at the 3' end of 14RH-1 may be the result of a cloning artifact; a BamHI site, not present in either sequence, lies at their junction.
Acquisition of Novel Promoters

Because V, F, and A belong to the same lineage, it follows that the new promoters were either acquired from different sequences or evolved from a common ancestor by the accumulation of mutations. Monomers are nearly twice as divergent as unselected sites (5' untranslated region and silent sites in ORF-1) in the bodies from the same elements (table 1), which supports the model that the promoters evolved from a common ancestor. While the cause of this accelerated divergence is not known, we believe that it is not sufficient to account for the lack of homology between A-monomers and F-monomers, for the following reason: the bodies of F-elements are slightly more divergent from the older A-elements than they are from each other, yet similarity is readily apparent among F-monomers but not between A-monomers and F-monomers (see table 2 and Material and Methods). If F-monomers evolved from A-monomers, then the data in table 2 suggest that A and F should still share at least 70% sequence identity. However, homology was not apparent in a variety of DNA sequence-alignment programs (not shown). Furthermore, short branch lengths connecting F-elements to A-elements (fig. 3) suggest that the appearance of A-monomers occurred over a relatively short period of time. Therefore, in the case of A- and F-elements, we strongly favor the model that the A-type sequence was acquired from an unknown source (perhaps viral) by an F-type L1 element. Concurrently or soon after this acquisition event, the F-monomer array was lost. Interestingly, the phylogenetic tree (fig. 3) shows that the composite clade of A/F-elements (A12 and 14LH) appear at approximately the same time. Presumably this narrow window is due to a mechanistic requirement that both A- and F-elements be actively replicating to generate the composite elements (Adey et al. 1991b). This further strengthens the argument that the composite clade
FIG. 3.—Phylogenetic trees of the Ll elements shown in fig. 1. The branches shown in boldface denote lineages of replicatively competent sequences (Hardies et al. 1986). The branch lengths (horizontal lines) are proportional to the number of sequence differences separating elements or groups of elements, except for those indicated as not significant (N.S.). To assess the reliability of the trees, they were each recalculated 100 times in a bootstrap analysis. Listed above the branches is the number of times that a particular node was obtained; <50 = node that was obtained less than half the time. Left, Phylogenetic tree calculated by maximum-parsimony analysis using DNA sequences from an alignment containing the 5′ untranslated sequences and all of the ORF-1 sequences from the A- and F-elements and just the ORF-1 sequences from the V- and R-elements. The placement of the human Ll element H2 (Kazazian et al. 1988) and rabbit Ll element O5 (Demers et al. 1986) outgroups was the result of a separate analysis (see Results). The tree depicts the evolution of Ll in rodents, beginning from the ancestral V-type elements, into the most recently inserted A-type elements in mice and R-type elements in rats. Right, Phylogenetic tree calculated by neighbor-joining analysis using DNA sequences from an alignment containing the 5′ untranslated sequences and all of the ORF-1 sequences from the A- and F-elements only. Inclusion of the V-, R-, human, and rabbit Ll elements gave inconsistent results (see Results).

was generated by transposition and not by ectopic recombination.

The relationship among F-, V-, and R-elements is less clear, because of a smaller number of sequenced elements and a greater degree of divergence. To determine if these Ll elements are also related by acquisition of novel 5′ termini, we examined the sequence similarity among these elements at silent positions in ORF-1 (positions where nucleotide substitutions do not change the encoded protein sequence). Significant sequence similarity does exist between F- and V-elements, R- and V-elements, and F- and R-elements at these silent positions (60%–72%; see table 3) and also between the 3′ untranslated regions of A- or F- and R-elements (Pascale et al. 1990, 1993; authors’ unpublished observations). Therefore, the lack of sequence similarity upstream of ORF-1 must be due to either an accelerated mutation rate or the acquisition of new sequences. An accelerated mutation rate does not exist in the 5′ untranslated regions, because percent similarity among F-elements, among V-elements, or among R-elements is the same in the 5′ untranslated region (81%–91%) as in the silent sites in ORF-1 (80%–91%; see table 1). Thus R-, F-, and V-elements are probably related by acquisition of new 5′ untranslated regions (and presumably the more 5′ promoter regions were acquired in the same event), a result consistent with the hypothesis of Furano et al. (1988).
### Table 1

Percent Sequence Identity among Members of Groups of L1 Elements

<table>
<thead>
<tr>
<th>Type of L1</th>
<th>Monomers 2–4</th>
<th>Monomer 1</th>
<th>5' Untranslated</th>
<th>All Sites</th>
<th>Category 3 Silent Sites</th>
</tr>
</thead>
<tbody>
<tr>
<td>A-type 3</td>
<td>. . . . . . . . .</td>
<td>78, 74</td>
<td>91, 90</td>
<td>93, 93</td>
<td>89, 88</td>
</tr>
<tr>
<td>A-type 2</td>
<td>. ND</td>
<td>89 (84–94), 88</td>
<td>96 (94–98), 96</td>
<td>97 (96–98), 97</td>
<td>98 (96–100), 98</td>
</tr>
<tr>
<td>F-type 3</td>
<td>80 (74–85), 77</td>
<td>78 (72–85), 74</td>
<td>89 (87–92), 88</td>
<td>91 (89–94), 90</td>
<td>86 (82–89), 85</td>
</tr>
<tr>
<td>F-type 1 and 2</td>
<td>81 (78–87), 78</td>
<td>83 (78–88), 81</td>
<td>89 (84–93), 88</td>
<td>89 (88–90), 88</td>
<td>89 (86–92), 88</td>
</tr>
<tr>
<td>lG5 and lG8</td>
<td>. . . . . . . .</td>
<td>ND</td>
<td>86, 85</td>
<td>84, 82</td>
<td>80, 77</td>
</tr>
<tr>
<td>V1, V2, and V3</td>
<td>. . . . . . . .</td>
<td>81 (71–97), 78</td>
<td>87 (83–95), 86</td>
<td>80 (69–96), 82</td>
<td></td>
</tr>
<tr>
<td>R-type</td>
<td>. . . . . . . .</td>
<td>83, 81</td>
<td>91, 90</td>
<td>92, 92</td>
<td>91, 90</td>
</tr>
</tbody>
</table>

* A-type group 3 = (L1Md) A1 and A21; A-type group 2 = A4, A7, A8, A9, A11, A15, and 14RH-1; F-type group 3 = (L1Md) F1, F2, F11, F14, and lG2; F-type groups 1 and 2 = (L1Md) F3, F13, F15, F16, and F18; and R-type elements = (L1Rn) R3 and LSSE. Values for A-type elements were taken from Schichman et al. (1992, 1993).

b Percent identity between two sequences is defined as the number of identical bases divided by the number of positions at which both sequences have a base present. Shown in parenthesis is the range of values from individual comparisons; and shown underlined is the average percent sequence identity corrected to reflect the possibility that multiple substitutions occurred at the same position. ND = not done.

---

### Age Estimate of L1 Elements

The corrected percent sequence identity of homologous regions can be used to estimate the age of the L1 elements (see tables 1–3). Once inserted, an individual L1 element, like a pseudogene, mutates at a neutral rate (Martin et al. 1985), which has been estimated to be approximately 1%/Myr in rodents (Li et al. 1987). In contrast, an actively replicating L1 lineage has been estimated to diverge at twice this rate (2%), presumably because of repeated cycles of reverse transcription (S. Hardies, personal communication). The active L1 lineages in mouse and rat are represented by L1MdA2 (Schichman et al. 1993) and L1RnR3 (Pascale et al. 1993). The unselected (silent) positions from these elements average 49% sequence identity (corrected for multiple hits), which is 51% divergence from each other or 25.5% divergence from a common ancestor. Using the 2% rate for active lineages, our results suggest that mouse and rat diverged 13 Mya. This result is consistent with the 12 Mya estimated by Pascale et al. (1993) using Lx elements. The age of F-type L1 elements can be estimated in a similar fashion, except that a mutation rate of 0.1%/sequence is used because F-elements are not active lineages. Table 2 indicates that the group 1 and 2 (see fig. 1) F-elements average 88% corrected sequence identity with each other (6% divergence from a common ancestor), which yields an age of 6 Myr. When F-elements are compared with A-elements, a mutation rate of 3% (1% for the inactive F-elements and 2% for the active A lineage) is used. Table 2 indicates that F-elements average 77% corrected sequence identity with the youngest subset of A-elements, so by this method an age of 7.5 Myr is obtained. The different mutation rates of active lineages and individual elements illustrates a potential pitfall if repetitive elements are used to infer phylogenetic relationships between their hosts. Comparison of a young element from one species with a much older element from a second species may lead to erroneous results.

### Analysis of ORF-1

Translation of the ORF-1 homologous sequence from each element yields results that are consistent with the phylogenetic analysis; elements that appear older in the phylogenetic tree also contain a higher number of reading-frame interruptions (see fig. 1). Furthermore, no two elements share the same reading-frame interruption. Because a shared interruption would indicate replication of an L1 element encoding a truncated protein, the lack of such a shared interruption is consistent with the requirement that the encoded product needs to be functional in cis to allow replication. The elements do, however, share length polymorphisms in ORF-1. These polymorphisms consist of a variable number of directly repeated sequence blocks that are the basis of the subclassification into groups (see fig. 1). The blocks can be of different lengths but are always a multiple of 3 nucleotides in length, which maintains the reading frame. This polymorphism has previously been shown to exist in A-elements (Shehee et al. 1987; Schichman et al. 1992), F-elements (Wincker et al. 1987), and R-elements (Furano et al. 1988). We find that it also exists in V-elements. Furthermore, the same three arrangements of blocks found in A-elements (Schichman et al. 1992) also exist in F-elements. These shared arrangements, in some cases, result from sequence recombui-
Table 2
Average Percent Sequence Identity, in 5' Untranslated Region, between Ll Elements

<table>
<thead>
<tr>
<th></th>
<th>F-type 3.4</th>
<th>F-type 1.2</th>
<th>A-type 3</th>
<th>A-type 2</th>
<th>A-type 1</th>
</tr>
</thead>
<tbody>
<tr>
<td>F-type 3.4</td>
<td>89 (87-92), 88</td>
<td>87 (83-92), 86</td>
<td>85 (83-86), 83</td>
<td>83 (80-89), 81</td>
<td>81 (80-82), 78</td>
</tr>
<tr>
<td>F-type 1.2</td>
<td>89 (84-93), 88</td>
<td>85 (82-90), 83</td>
<td>87 (76-90), 81</td>
<td>80 (76-83), 77</td>
<td></td>
</tr>
<tr>
<td>A-type 3</td>
<td>91 (76-90), 90</td>
<td>87 (84-91), 86</td>
<td>85 (83-97), 83</td>
<td></td>
<td></td>
</tr>
<tr>
<td>A-type 2</td>
<td>90 (86-96), 89</td>
<td>93 (86-97), 93</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Note.—Presentation of data is as in table 1. The subsets of L1 elements are as in table 1, except that the F-type group 4 element L1Md1G-6 is included with the F-type group 3 elements and the A-type group 1 elements (which share greater than 99% sequence identity) are represented by L1MdA2.

Table 3
Average Percent Sequence Identity, in ORF-1, between Groups of Ll Elements

<table>
<thead>
<tr>
<th></th>
<th>(V) lG5 and lG8</th>
<th>V1, V2, and V3</th>
<th>R-type</th>
</tr>
</thead>
<tbody>
<tr>
<td>A- and F-type</td>
<td>72 (70-75), 65</td>
<td>81 (77-84), 78</td>
<td>76 (72-80), 71</td>
</tr>
<tr>
<td></td>
<td>60 (48-67), 43</td>
<td>72 (63-80), 65</td>
<td>63 (57-68), 49</td>
</tr>
<tr>
<td>(V) lG5 and lG8</td>
<td>81 (78-87), 78</td>
<td>72 (70-73), 65</td>
<td>62 (52-70), 47</td>
</tr>
<tr>
<td></td>
<td>78 (70-89), 74</td>
<td>79 (76 82), 75</td>
<td>71 (64-77), 63</td>
</tr>
</tbody>
</table>

Note.—Presentation of data is as in table 1; in each cell, the data in the upper line are for all sites, and the data in the lower line are for category 3 silent sites. A- and F-type = (L1Md) A2, F15, F13, and 1G6; and R-type = (L1Rn) R3 and 1SSF.
belong to the same lineage but were inserted during different time periods and that this lineage has repeatedly acquired new 5' promoters. The R-elements in rat probably also arose from V-elements by acquisition of new 5' termini. Thus V may predate the divergence of mouse and rat. Our analysis suggests that V may represent the same class of elements as are termed "Lx" by Pascale et al. (1990, 1993).

The mechanism by which one lineage dominates the replication process is unknown, but the same phenomenon is also observed in phylogenetic studies of short interspersed transposable elements known as "SINES" (reviewed by Deininger et al. 1992). LINEs and SINEs also share the additional feature that the amplification rate of both appears more episodic than continuous (Britten et al. 1988; Deininger and Slagel 1988; Jurka and Smith 1988; Bains and Temple-Smith 1989; Pascale et al. 1990; Krane et al. 1991; Martin 1991a). Our results display an interesting similarity to the episodic model in that the acquisition of the A-type promoter appears to correlate with the demise of a number of F-type lineages. This suggests that a selective advantage for L1 replication may have accompanied the acquisition of new transcriptional regulatory sequences. Transcriptional regulation is therefore one step in the L1 replication-and-dispersal cycle that may play an important role in L1 evolution.

The acquisition of new sequences has also been observed in other types of transposable elements. For example, three subfamilies of SINE elements exist in Galago; one is a composite of the other two (Daniels and Deininger 1991). In humans, two examples exist of an Alu element fused to the 5' end of an I1 element; in both cases, it appears that the composite element inserted as a unit (Miyake et al. 1983). The Txl1 element in Xenopus (Carroll et al. 1989), the F-element in Drosophila (Di Nocera and Casari 1987), and the THF-1 element in human and the INGI element in trypanosomes (both reviewed in Hutchison et al. 1989) contain terminal sequences that can also exist independently in high copy number. In addition, Demers et al. (1989) have suggested that the lack of homology, in the 3' untranslated region, between L1s from different mammals is a result of sequence acquisition. Together, these examples suggest that acquisition of new sequences may play an important role in the evolution of transposable elements.

**Acknowledgments**

The authors thank Stephen Hardies, Anthony Furano, and Sandra Martin for helpful discussions and critical review of the manuscript. François Catzeflis is gratefully acknowledged for providing rodent tissue samples and genomic DNA. Wayne Litaker, Thomas Whittam, and Vincent Illiano are gratefully acknowledged for assistance with the computer and the phylogenetic programs. This work was supported by U.S. Public Health Service grants AI08998 and GM21313 from the National Institutes of Health. S.A.S. was supported by U.S. Public Health Service grant HL07040 from the National Institutes of Health.

**LITERATURE CITED**


is a CpG island can function as a promoter. Nucleic Acids Res. 16:9233–9251.


THOMAS EICKBUSH, reviewing editor

Received February 7, 1994

Accepted April 26, 1994