

The RTE Class of Non-LTR Retrotransposons Is Widely Distributed in Animals and Is the Origin of Many SINEs

Harmit S. Malik and Thomas H. Eickbush

Department of Biology, University of Rochester

RTE-1 is a non-long-terminal-repeat (non-LTR) retrotransposable element first found in the *Caenorhabditis elegans* genome. It encodes a 1,024-amino-acid open reading frame (ORF) containing both apurinic-apyrimidic endonuclease and reverse-transcriptase domains. A possible first ORF of only 43 amino acids overlaps with the larger ORF and may be the site of translation initiation. Database searches and phylogenetic analysis indicate that representatives of the RTE clade of non-LTR retrotransposons are found in the bovine and sheep genomes of mammals and in the silkworm and mosquito genomes of insects. In addition, the previously identified SINEs, Art2 and Pst, from ruminant and viper genomes are shown to be truncated RTE-like retrotransposable elements. RTE-derived SINE elements are also found in mollusc and flatworm genomes. Members of the RTE clade are characterized by unusually short 3' untranslated regions that are predominantly composed of AT-rich trimer, tetramer, and/or pentamer repeats. This study establishes RTE as a very widespread clade of non-LTR retrotransposons. RTE represents the third distinct class of non-LTR retrotransposons in the vertebrate lineage (after Line 1 elements in mammals and CR1 elements in birds and reptiles).

Introduction

Transposable elements have long been considered to be deleterious components of eukaryotic genomes. By virtue of their deleterious effects, it would be expected that selective pressures would tend to eliminate transposable element lineages from host genomes (Charlesworth 1988). Two evolutionary forces would act to maintain transposable elements in a lineage: active transposition or subsequent reintroduction following loss from a host lineage. Both processes have been documented for individual transposable elements (Robertson 1993; Clark, Maddison, and Kidwell 1994; Garcia-Fernandez et al. 1995; Lohe et al. 1995; Springer et al. 1995; Burke et al. 1998).

Among eukaryotes, the most abundant elements are the retrotransposable elements. Retrotransposable elements can be divided into two separate classes based on their structures and modes of integration. The long terminal repeat (LTR) retrotransposable elements, as the name implies, are flanked by direct repeats bearing critical information for their transcription and retrotransposition. LTR retrotransposons rely on a tRNA-mediated mechanism for priming of reverse transcription. The non-LTR class of retrotransposable elements (also called LINE-like for Long Interspersed Nucleotide Elements) are not bounded by element-derived direct repeats and many have a characteristic poly(A) tail at the 3' end. This class of elements is thought to use a nick or break on the host chromosome to prime reverse transcription of its RNA transcript directly onto the target site (Luan et al. 1993).

Abbreviations: AP endonuclease, apurinic-apyrimidic endonuclease; CR1, chicken repeat element; LINEs and SINEs, long and short interspersed nucleotide elements, respectively; LTR, long terminal repeat; ORF, open reading frame; RT, reverse transcriptase.

Key words: AP endonuclease, reverse transcriptase, RTE1, BDDF, Art2, *Caenorhabditis elegans*.

Address for correspondence and reprints: Thomas H. Eickbush, Department of Biology, University of Rochester, Rochester, New York 14627. E-mail: eick@uhura.cc.rochester.edu.

Mol. Biol. Evol. 15(9):1123–1134, 1998

© 1998 by the Society for Molecular Biology and Evolution. ISSN: 0737-4038

It is often difficult to correctly identify non-LTR retrotransposable elements on the basis of only a few copies. Unlike LTR-retrotransposons that generate uniform target site duplications and require the presence of their terminal repeats for integration, non-LTR element copies are often truncated at their 5' ends, and many generate variable target site duplications or even deletions. The characteristic 5' truncations associated with non-LTR integration are thought to be a consequence of the integration of prematurely terminated reverse transcripts initiating at the 3' end of the RNA (Luan et al. 1993). In the absence of a set of uniform structural features, unambiguous indication of the presence of a non-LTR retrotransposable element can sometimes only be accomplished by an examination of its reverse transcriptase (RT) encoding domain. Unfortunately, because the 5' truncations can include the RT domain and because all truncated copies will, with time, accumulate mutations eliminating their ORFs, non-LTR retrotransposons can remain unidentified or be misidentified as SINE insertions. SINEs (for Short Interspersed Nucleotide Elements) are reverse transcripts of short stable RNAs, the best characterized being those derived from 7SL or tRNA (Deininger 1989; Ohshima et al. 1996). SINE insertions also contain poly(A) tails and generate target site duplications that vary in length. SINEs are believed to utilize the machinery of non-LTR retrotransposable elements for their insertion (Luan et al. 1993; Ohshima et al. 1996; Jurka 1997).

In the present study, we examine a lineage of non-LTR elements, the RTE-1 (retrotransposable element) clade, which was first identified in *Caenorhabditis elegans* (Youngman, van Luenen, and Plasterk 1996). We demonstrate that the RTE-1 elements contain a domain with homology to the apurinic-apyrimidic (AP) endonucleases in addition to the previously identified RT domain. We also show that the RTE-like elements are widely distributed in vertebrates and arthropods and that these elements frequently have extensive 5' truncations, giving rise to SINE elements.

Table 1
Survey of RTE1 Elements in the *Caenorhabditis elegans* Genome

Copy Number	Accession Number	Annotations	Target Site Duplication (bp)
1	U00034	2 frameshifts (1–3258)	219
2	AF025462	Uninterrupted ORF (1–3258)	69
3	Z79599	Uninterrupted ORF (1–3258)	1,394
4	Z79755	Uninterrupted ORF (1–3258)	424
5	Z83319	Uninterrupted ORF (1–3258)	439
6	U41008/ U42848 ^b	Uninterrupted ORF (1–3258)	22 ^a
7	Z83109	1 stop codon/2.5 kb insertion (1–3258)	18 ^a
8	U58735/ AF026205 ^b	1 frameshift (1–3258)	190
9	U40801	(790–3258)	ND
10	Z68336	(1005–3258)	72
11	AF025453	(1172–3258)	353 ^a
12	AF039712/ AF043696 ^b	(1311–3258)	133
13	Z73976	(1530–3258)	209
14	U40424	(1918–3258)	415
15	U39850	(2169–3258)	ND
16	Z99277	(2233–3258)	16 ^a
17	U41032	(2253–3258)	ND
18	U41013	(2318–3258)	260
19	U97551	(2804–3258)	11
20	AF016684	(3064–3258)	90
21	U00040	(3116–3258)	20
22	Z68106	(3179–3258)	28 ^a
23	Z82081	(3190–3258)	15 ^a
24	Z81580	(3190–3258)	ND
25	Z46792	(3198–3258)	ND
26	U67955	(3218–3258)	57
27	Z79756	(3218–3258)	9

NOTE.—ND = not detected. RTE-1 homologs containing 3' deletions and, frequently, internal disruptions are listed as follows: Z83114 (1–2896); Z69791 (1–1606); Z81028 (772–1882, 2253–3258); Z92781 (1–782, 796–2392, 2624–3086); AF016663 (1–1937, 2026–2392, 2624–3237); Z69635 (1–251, 408–1175); U42848 (1–519); U58743 (1–125, 89–562); Z81039 (198–633, 1104–1476); AF022967 (2661–3258); Z80219 (33–293); Z69903 (2646–2898); U53335 (37–259); Z81140 (303–536); AF000266 (305–578); Z81514 (116–645, 2931–3040); Z81543 (2986–3237); Z69787 (3012–3237); Z29094 (3042–3238); AL021474 (3000–3238); Z70755 (83–223); AF039037 (800–959); U88172 (123–232); U61950 (749–869).

^a Not a perfect target site duplication.

^b The complete element is defined in overlapping cosmids.

Materials and Methods

Sequences homologous to the RTE-1 element were identified both within and outside the *C. elegans* genome using the BLASTP (protein query against protein database) and TBLASTN (protein query against all six frames of nucleotide databases) programs (Altschul et al. 1990) against a nonredundant database. We employed RTE-1, RTE-2, and BDDF-bovine (Szemraj et al. 1995) as the query sequences in our search.

All DNA and protein sequence analyses were conducted using the MacVector computer program package (International Biotechnologies). In some cases, as noted in tables 1 and 2, it was necessary to shift frames to maintain an intact ORF. The alignments were carried out

using CLUSTAL W (Thompson, Higgins, and Gibson 1994), and the Neighbor-Joining method (Saitou and Nei 1987) was used to analyze phylogenetic relationships, with bootstrapping done using CLUSTAL W. Maximum-parsimony bootstrapping analyses were carried out using the SEQBOOT, PROTPARS, and CONSENSE programs in the PHYLIP package (Felsenstein 1993).

The new definitions of the RTE-1 ORF structure are summarized in GenBank under accession number AF054983.

Results

Structure of the RTE-1 element of *C. elegans*

The RTE-1 retrotransposon was first identified in *C. elegans* as an insertion into an intron of the *pim related kinase-1 (prk-1)* gene (Youngman, van Luenen, and Plasterk 1996). As defined by the authors, the element was about 3.3 kb in length, was flanked by a direct repeat of 200 bp, and encoded a 600-amino-acid-residue ORF with RT domain. The RT domain had the greatest homology to those of non-LTR retrotransposons. An ORF of 600 amino acids is unusually short for a non-LTR retrotransposon, suggesting that the element might be a 5' truncation. We searched the expanding database generated by the *C. elegans* genome sequencing consortium for a putative full-length RTE-1 element. Fifty additional RTE-1 sequences were identified. The seven longest elements (table 1, copy numbers 2–8) were virtually identical in sequence and length to the original RTE-1 element (copy number 1). These elements extended 3,258 bp from their 5' ends to the termination codon defined by Youngman, van Luenen, and Plasterk (1996). Thus, there was no indication that the originally sequenced RTE-1 was truncated. The remaining 43 RTE-1 sequences identified in the *C. elegans* genome were not full-length. Nineteen of these copies contained 5' truncations typical of the non-LTR mechanism of retrotransposition (copy numbers 9–27). The remaining 24 copies (accession numbers given in the table 1 legend) contained deletions at their 3' ends, and many contained internal disruptions; thus, they are unlikely to represent the structure of the original retrotransposition events.

The ORF of the RTE-1 element had been defined with a methionine as the first amino acid, in keeping with paradigms of eukaryotic translation. However, retrotransposable elements of both the LTR and non-LTR classes have been shown to shift frames or bypass termination codons at a low level during protein translation (Jacks and Varmus 1985). These frameshifts or bypasses allow the low-level expression of a second ORF from an initiation codon located in the first ORF of the element. Consistent with this suggestion, the ORF of the RTE-1 element could be extended nearly 400 codons upstream of the first Met codon (fig. 1B). Conceptual translation of the other seven full-length copies of RTE-1 revealed an identical 400-amino-acid extension upstream of this Met codon. As further evidence that this extension is utilized by RTE-1 elements, the upstream sequence contains an AP endonuclease domain (fig. 2).

Table 2
FTE-1-like Elements in *Caenorhabditis elegans* and Other Genomes

Genome	Accession Number
<i>Aedes aegypti</i> (mosquito)—JAM1	Z86117
<i>Angiostrongylus cantonensis</i> (nematode)	U13191
<i>Bombyx mori</i> (silkworm)	D25321
<i>Bos taurus</i> (cow)—BDDF	M63452
(Art2; bov-2; Pst)	Z25531, M94327, Z25525, X99691, X82671, L13373, Z25530, U39887, AF016539, M95099, Z25526, U19468
(BCNT insertion)	D84513
<i>Caenorhabditis elegans</i> (nematode)—RTE-2	U58775, U0063, g2253129
<i>Helix aspersa</i> (snail)	X55948
<i>Ovis aries</i> (sheep)—Art2	X79703, U65982, AF026566
<i>Ommastrephes sloanei</i> (squid)	M74321
<i>Rattus norvegicus</i> (rat) ^a	M28630
<i>Schistosoma mansoni</i> (bloodfluke)	D87491
<i>Trimeresurus flavoviridis</i> (habu snake)	D31777
<i>Tragulus javanicus</i> (chevrotain)—BCNT	AB005651
<i>Capra hircus</i> (goat)—Art2	M57436
<i>Vipera ammodytes</i> (viper)—Art2	X84017, X76731

NOTE.—Accession numbers in italics indicate the presence of stop codons or frameshifts that need to be invoked to maintain homology in the ORF

^a May represent a bovine DNA contamination (see text).

An AP- endonuclease domain has been identified at the amino terminal end of the second ORF of many non-LTR retrotransposons (Martin et al. 1995; Feng et al. 1996).

Does the expression of the RTE-1 ORF also involve a frameshift or bypass in translation from a first ORF? As shown in figure 1B, conceptual translation of the RTE-1 sequence upstream of its major ORF revealed a 43-amino-acid ORF that overlaps with the beginning of the large ORF. All eight full-length copies of RTE-1 encode this first ORF. This overlap between the first and second ORFs is similar to the arrangement in many other retrotransposable elements. As an additional argument for the use of this short ORF, the Met codon that begins this ORF is the only ATG sequence at the 5' end of the RTE-1 element. Clearly, however, this ORF is much shorter than any previously defined first ORF, and it is unlikely to encode any of the RNA-binding components that have been attributed to the first ORFs of other non-LTR elements (see *Discussion*).

In the original characterization of RTE-1 (Youngman, van Luenen, and Plasterk 1996), it was not clear if the 219-bp direct repeat flanking the element was an integral part of the element or was generated by a target site duplication formed during the integration of the element. Analysis of the sequences of full-length RTE-1 copies in the database clearly indicates that these elements do not have direct repeats. Each full-length RTE-1 element does, however, contain a target site duplication (table 1) which, in some cases, is extremely large (range 18–1,394 bp). Target site duplications (range 11–415 bp) can also be found in 14 of the 19 copies containing 5' truncations. Target site duplications over 100 bp are unusual for non-LTR retrotransposable elements (Eickbush 1992). There was no easily defined sequence similarity in these duplications other than a high AT content, indicating an apparent lack of site-specificity.

The large target site duplications bordering RTE-1 elements have enabled a precise definition of the elements' ends. The region upstream of the Met codon of the first ORF, defined as the 5' untranslated region (5' UTR), is only 64 bp in length and is highly conserved in sequence (no variation among the eight full-length copies). The region downstream of the stop codon of the large ORF (3' UTR) is unusually short for a non-LTR retrotransposable element. Unlike the 5' UTR, the 3' UTR exhibits variation in both length and sequence. As shown in figure 3, the 3' UTRs of representative RTE-1 copies vary from 33 to 41 bp in length. They all begin with a 13-bp conserved sequence but end with different numbers of the tetranucleotide repeats, TAAG and TATC. Short AT-rich nucleotide repeats have also been found at the 3' ends of the non-LTR elements I (Fawcett et al. 1986), R1 (Eickbush and Eickbush 1995), L1 (Furano et al. 1994), and CR1 (Silva and Burch 1995). Such heterogeneity found at the 3' junction of non-LTR elements is believed to be introduced by the target-primed reverse transcription mechanism used for integration. The RT encoded by the non-LTR element can either add nontemplated nucleotides or make aborted attempts before the eventual reverse transcription of the transcript near its 3' end (Luan and Eickbush 1995).

In summary, based on the high degree of nucleotide sequence conservation that we see throughout all copies of RTE-1 elements in *C. elegans*, as well as the presence of a 1,024-amino-acid ORF encoding both putative AP- endonuclease and RT activities, we conclude that the original RTE-1 element identified by Youngman, van Luenen, and Plasterk (1996) is a complete non-LTR retrotransposable element. The high level of sequence identity between different copies and the finding that the number and location of RTE-1 elements vary between different geographical strains of *C. elegans* (Youngman,

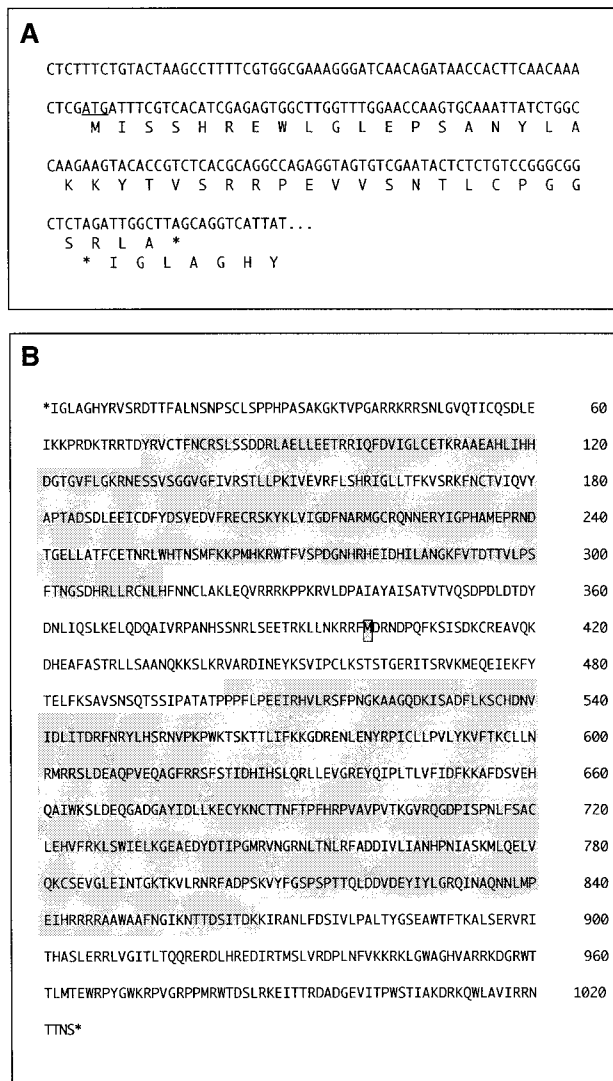


FIG. 1.—The putative ORFs of the RTE-1 element from *C. elegans*. The sequence is derived from the full-length element present in accession number AF025462. **A**, Nucleotide sequence of the 5' end of the RTE-1 element. The first ORF is only 43 amino acid residues long, and overlaps the second ORF by seven nucleotides. The methionine codon beginning this ORF is the only ATG sequence present near the 5' end of the element. **B**, Amino acid sequence of the 1,024-amino-acid second ORF. The AP-endonuclease and RT domains are highlighted by shading. Also indicated is the methionine residue which was originally defined as the beginning of the ORF by Youngman, van Luenen, and Plasterk (1996).

van Luenen, and Plasterk 1996) suggest that the elements have recently been active.

Broad Distribution of RTE-1 Homologous Elements in Animals

We were next interested in examining the distribution of RTE-like elements both within and outside the *C. elegans* genome. The original study (Youngman, van Luenen, and Plasterk 1996) had already identified a highly divergent lineage of RTE-1-like elements in the *C. elegans* genome. These elements are over 50% divergent in nucleotide sequence from the RTE-1 copies; thus we have termed them RTE-2 elements in keeping

with the original nomenclature. Only three copies of the RTE-2 lineage have been sequenced to date and all contain 5' truncations. In figure 4, the ORF structure of the longest RTE-2 element is compared to that of the RTE-1 element. RTE-2 elements contain large ORFs, with RT and endonuclease domains in positions similar to those of RTE-1. The 3' UTRs of the RTE-2 elements are also very short and bear sequence similarity to the RTE-1 elements (fig. 3). In particular, the 3' junctions of RTE-2 elements contain variable numbers of (T/A)AAG and TATC repeats.

Significant matches outside *C. elegans* were found between the RTE-1 RT and endonuclease domains and sequences in mammals and insects. A listing of the sequence matches from GenBank is presented in table 2, and the regions of similarity with RTE-1 are summarized in figure 4. The most extensive homology to RTE-1 was found in the cow, *Bos taurus*. The bovine homolog of RTE-1 was originally named BDDF (for bovine dimer-driven family) and is a 3.1-kb repetitive sequence previously identified as a retroelement that specifically inserts into a bovine Alu-like sequence (Szemraj et al. 1995). This specific BDDF copy contains a number of mutations which generate stop codons or changes in reading frame; thus, the full extent of the coding capacity of this element was not apparent in the original study. Allowing for these mutations (identified by vertical lines in fig. 4), the putative active bovine element would contain an approximately 1,000-amino-acid ORF, similar in organization to the *C. elegans* retrotransposon. Unfortunately, the necessity of introducing frameshifts to maintain the ORF through the endonuclease and RT domains made it impossible to infer whether the BDDF element also contains a small first ORF like that proposed for RTE-1.

More recently a truncated BDDF element encoding just a 280-amino-acid region bearing the AP endonuclease domain was found inserted into the middle of the protein-coding region of a GTPase-activating protein gene, BCNT (Nokubuni et al. 1997). The endonuclease domain is expressed as part of the mature protein in brain, liver, and lung tissues. The BDDF insertion into this gene is also found in *T. javanicus* (chevrotain), suggesting at least a 15-Myr-old association between the BCNT gene and the endonuclease domain in the Ruminantia. The preservation of the AP-endonuclease-encoding ORF in-frame to the BCNT ORF may indicate a selective advantage conferred by this domain. The human homolog to the bovine BCNT gene does not contain this sequence (Nokubuni et al. 1997).

A second RTE-like retrotransposable element sequence in mammals was identified in the rat, *Rattus norvegicus*. This sequence was initially characterized because it contained a SV40 insertion (Bullock, Forrester, and Botchan 1984). The rat sequence is truncated by cloning but contains the entire RT domain (fig. 4). It has been suggested that the sequence flanking the SV40 insertions in this experiment are derived from the calf thymus DNA used as "carrier" (Lenstra 1992). Even if this RTE sequence is actually from the cow, it represents a

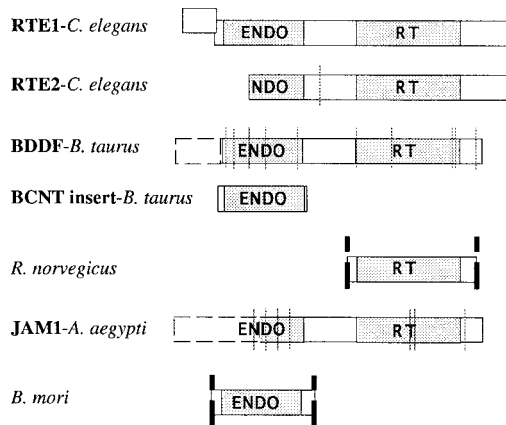
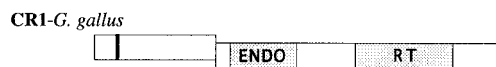
RTE clade**Line1 clade****CR1 clade**

FIG. 4.—Comparison of the ORF structures and enzymatic domains of RTE-like elements. Also shown is one representative of each of the two other known vertebrate lineages of non-LTR retrotransposons: Line 1 elements from *R. norvegicus* (rat) and CR1 from *G. gallus* (chicken). ORFs are indicated by unshaded boxes. The AP endonuclease (ENDO) and RT domains within these ORFs are indicated by darker shading. The thick dashed lines indicate those sequences that are either truncated by cloning or have not been sequenced. The light vertical lines indicate either the bypass of a stop codon or a frameshift that needed to be invoked to retain homology of the ORF. The precise amino terminal ends of the ORFs from the BDDF and JAM1 elements cannot be defined because of the number of such bypasses needed to maintain the ORF through the ENDO domain. While sequenced from an *R. norvegicus* tissue culture cell, the rat sequence is probably of bovine origin, resulting from the use of calf thymus carrier DNA in the cloning experiment (Lenstra 1992).

different lineage from that of the original BDDF element and thus was used in the phylogenetic analysis below.

RTE homologs were also identified in two insect genomes (table 2). A possible complete RTE homolog has been characterized in the mosquito, *Aedes aegypti* (Warren, Hughes, and Crampton 1997). This element, named JAM1, was initially identified as an insertion into the LTR-retrotransposon, *Zebedee*. Multiple frameshifts are necessary, but a large ORF with RT and endonuclease domains can be inferred (fig. 4). While the length of the JAM1 insertion is similar to that of a complete RTE-1 element, the sequence at the 5' end of this element is too mutated to reveal the extent of the ORF or even the full endonuclease domain. The second RTE-1 insect homolog was found in the silkworm, *Bombyx mori*. This RTE-1 homolog is inserted adjacent to the *cecropin B* gene and was not recognized as part of a transposable element. Conveniently, the only part of this element that was sequenced corresponds to the endo-

nuclease domain (fig. 2); thus, it, too, can be used for phylogenetic analysis.

The various non-LTR elements identified in eukaryotes are highly divergent in sequence, which has made it difficult to resolve their phylogenetic relationships (Xiong and Eickbush 1990; Eickbush 1994). Using all available non-LTR sequences, we have recently been able to divide non-LTR elements into several distinct clades (or lineages) (unpublished data). Non-LTR elements from different organisms are suggested to be in the same clade if they have a similar structure and if both distance and parsimony algorithms indicate, with significant bootstrap values, that the elements are more related to each other than to any other element. To demonstrate that the various mammalian and insect sequences described above are truly part of a lineage that includes the *C. elegans* RTE elements and not members of other non-LTR retrotransposable element lineages, the RT domains from the RTE-1 and RTE-2 elements of *C. elegans*, the bovine BDDF sequence, the rat homolog, and the mosquito JAM1 sequence were aligned with representative examples of non-LTR retrotransposable element lineages previously identified in vertebrates, insects, and nematodes. These elements included R4 from *Ascaris lumbricoides*, L1 from *Rattus norvegicus*, CR1 from the chicken *Gallus gallus*, and I, Jockey, R1, and R2 elements from *Drosophila melanogaster*. The RT sequences are aligned in figure 5A with the conserved domains previously identified as common to all reverse transcriptases (Xiong and Eickbush 1990). An additional region which is conserved in all non-LTR retrotransposons is labeled 0. The relationship of the RTE-like elements based on the RT alignment is shown in figure 5B. This analysis clearly groups the two RTE elements from *C. elegans* with the bovine and mosquito elements at significant bootstrap values. We thus conclude that these RTE-like sequences should be considered part of the same lineage.

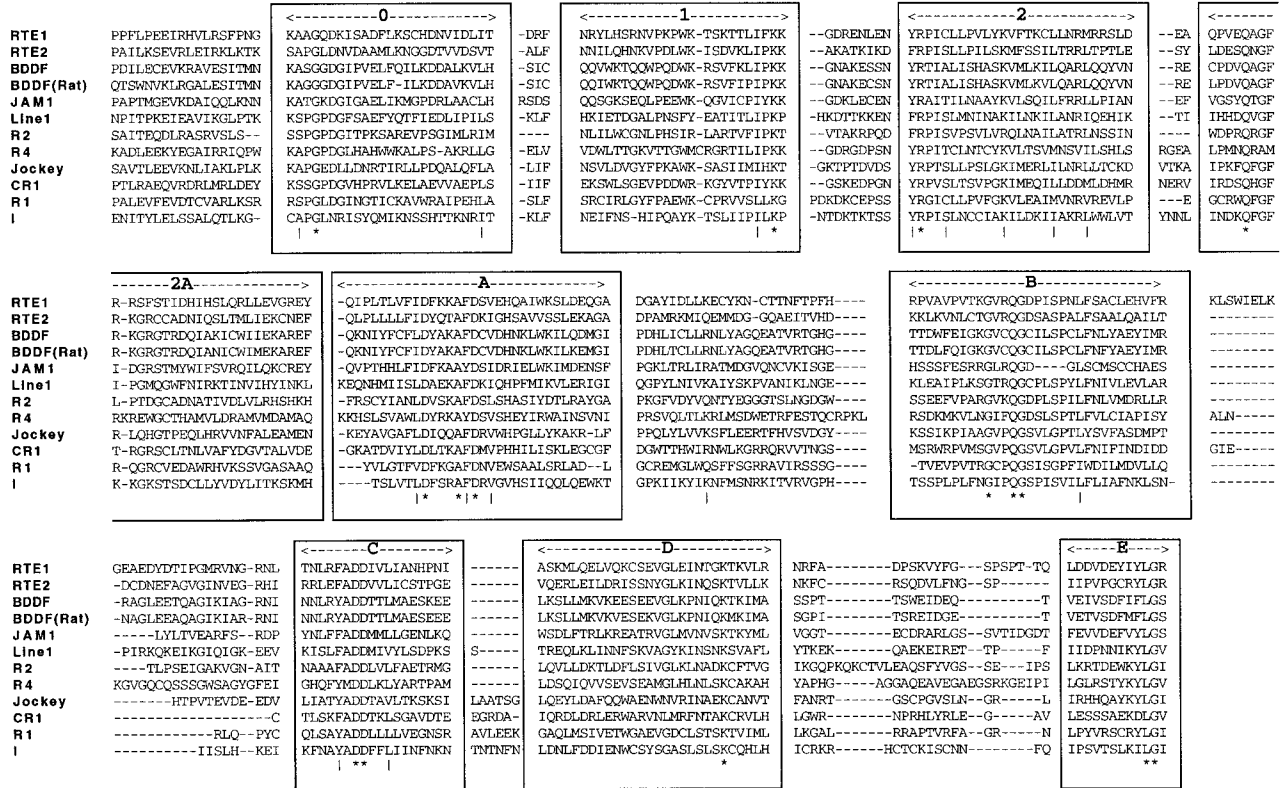
In an attempt to independently confirm the RTE-1 clade, the sequence alignment of the endonuclease domain shown in figure 2 was used in a phylogenetic analysis. The phylogeny based on these sequences was not as reliable as that of the RT domain, but did group the RTE-1 homologs together (data not shown). Thus, phylogenetic analysis for both the RT and endonuclease domains of RTE are consistent with the conclusion that the same lineage of elements is present in nematodes, arthropods, and vertebrates.

Finally, we believe it is highly significant that the presumed 3' UTRs of both the cow BDDF and the mosquito JAM1 are extremely short and exclusively composed of tetra- and pentanucleotide repeats reminiscent of those found in RTE-1 and -2 (see fig. 3). Such unusually short 3' UTRs are not found in other non-LTR retrotransposon lineages, clearly supporting the phylogenetic analysis suggesting that these RTE-like elements are from the same lineage.

RTE-1 Elements Have Given Rise to SINEs in Many Lineages

If the 5' truncations associated with non-LTR element insertions are extensive and include both the en-

A



B

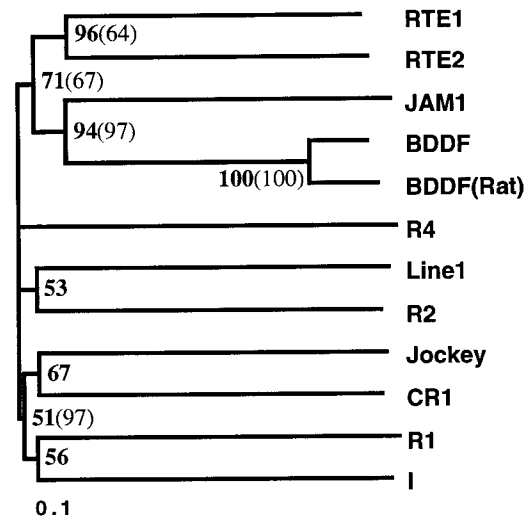


Fig. 5.—Comparison of RT sequences from the RTE-like elements with those from other non-LTR retrotransposons. The BDDF sequence from rat is probably a bovine DNA contamination (see text). A, RT domains are aligned from those RTE-like elements with complete RT domains (see fig. 4). Included in the alignment are those non-LTR retrotransposons lineages also found in vertebrates, insects, and nematodes: R1, R2, I, and Jockey from *D. melanogaster*, Line 1 from *R. norvegicus*, CR1 from *G. gallus*, and R4 from *A. lumbricoides*. The seven conserved RT regions defined in Xiong and Eickbush (1990) are boxed and labeled 1 through E. Two additional domains (labeled 0 and 2A) are also conserved in all non-LTR retrotransposons. “*” and “|” indicate identical and similar amino acids, respectively. B, An unrooted phylogeny of the RT domains constructed using the neighbor-joining method. Numbers next to each node indicate bootstrap values as percentages of 1,000 replicates—a 50% consensus tree is presented here. This topology is also supported by maximum-parsimony methods. Bootstrap values obtained using parsimony analyses are indicated in parentheses wherever the topology is supported by a minimum of 50% of bootstrap replicates. The RT phylogeny places RTE-1, RTE-2, JAM1, BDDF-Bovine, and BDDF-Rat as one clade, supporting the hypothesis that RTE-1 is widely distributed among the three phyla. An amino acid divergence scale is shown at the bottom.



FIG. 6.—Comparison of the carboxyl terminal ends of the ORFs encoded by different RTE homologs with those of putative ORFs encoded by SINEs from different species. Compared are those elements included in figure 3 as well as representatives from another nematode (*A. cantonensis*), a flatworm (*S. mansoni*), a silkworm (*B. mori*), a snail (*H. aspersa*), a squid (*O. sloanei*), two snakes (*V. ammodytes* and *T. flavoviridis*), and two ruminants (*O. aries* and *C. hircus*). The following notations were used in the sequence: *, stop codons that were bypassed; X, ambiguous amino acid due to a shift in frame; ?, loss of homology, possibly due to a subsequent insertion or deletion; and //, truncations by cloning. Numbers indicate amino acids that were omitted from the alignment, and a colon indicates the start of a particular sequence. The Art2 consensus sequence (accession number X82879) is derived from multiple artiodactyl sequences. The underlined residues Y/FLG are the conserved residues of the RT segment labeled E in figure 4A. The numbers in parentheses represent the number of amino acids the RT can be extended upstream for elements not represented in figure 2. At the bottom of the alignment, residues that are nearly invariant are indicated, while similar amino acids in most elements are indicated by “|”.

donuclease and RT domains, it can be difficult to recognize that certain SINE elements are in fact the 3' ends of non-LTR retrotransposons (Silva and Burch 1989; Ohshima et al. 1996). The generation of a very large number of highly truncated insertions by RTE elements appears to have occurred in the bovine genome. In the original characterization of the BDDF element (Szemraj et al. 1995), it was concluded that an Alu-like family of repeat sequences, termed the Pst and Art2 families (Duncan 1987), constituted the target-site for the 3.1-kb BDDF insertion element. More recently, Okada and Hamada (1997) have suggested that it is more likely that these SINE elements represent deletions of the BDDF LINE-like element. As shown in figure 6, the Pst and Art2 elements actually encode the extreme carboxyl terminal end of the BDDF ORF. This clearly confirms that Pst and Art2 elements are indeed extreme 5' truncations of BDDF. These various elements have extremely short 3' UTRs which are composed exclusively of variable numbers of the pentamer repeats found in BDDF (fig. 3). The Pst and Art2 families of SINEs have shown

to be abundant in all genomes in Artiodactyla, suggesting that they are at least 25–40 Myr old (Modi, Gallagher, and Womack 1996).

Art2 SINE families have also been identified in viper genomes (Kordis and Gubensek 1995). This study postulated a horizontal transfer of these SINE families between a mammal and a reptile based on 70% nucleic acid sequence identity of the viper sequence with the bovine sequences. Figure 6 illustrates how these previously defined viper SINE elements (*Vipera ammodytes* 1 and 2) are also homologous to the carboxyl terminal ends of the ORFs of RTE elements. Indeed, sequence identity with the RTE-like elements of *C. elegans* and mosquito can also be identified throughout this short region (conserved residues identified at the bottom of the sequence comparisons). Thus, the preservation of this protein-coding sequence by the parent non-LTR retrotransposon can explain the sequence conservation originally detected between the viper and cow sequences. We have used this short segment of the ORF to reconstruct the phylogenetic relationship of these various

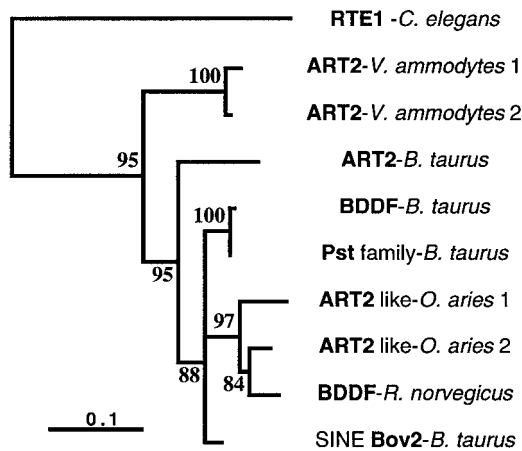


FIG. 7.—Phylogeny of the RTE and SINE sequences from the various vertebrate genomes. The phylogeny presented is based on the nucleotide sequences derived from the amino acid alignment shown in figure 6. The phylogeny is a 50% consensus tree of the vertebrate sequences using the neighbor-joining method and rooted on the RTE-1 sequence of *C. elegans*. Numbers next to each node indicate bootstrap values as percentages of 1,000 replicates. The divergence of some of the lineages within the mammalian genomes might predate the split between rodents and artiodactyls. The viper Art2 copies fall outside the mammalian homologs, as would be expected from a vertical means of inheritance. The nucleotide divergence scale is indicated.

vertebrate RTE-like sequences using RTE-1 of *C. elegans* as the root. As shown in figure 7, the sequence relationships of the SINE/RTE sequences in cow are highly divergent. The elements from sheep (*Ovis aries*) fall within this divergence, as does the rat sequence. The latter finding is consistent with the suggestion of Lenstra (1992) that this sequence may be derived from the cow genome. The viper sequences are clearly outgroups to all of the mammalian sequences. Thus, given the data available to date, there is no need to suggest a horizontal transfer of these elements between a viper and a ruminant genome. Further analysis of RTE sequences from different vertebrates would allow estimates of the rates at which RTE elements evolve. Only then can predictions be made as to whether horizontal transfers are needed to explain the current distribution of these elements.

Using the conserved sequences identified in this analysis at the carboxyl terminal end of the RTE ORF, we were also able to identify truncated RTE-like sequences in a highly diverse set of additional animals. As shown in figure 6, such sequences were found in the genomes of a snail, *Helix aspersa*; a squid, *Ommanstrephes sloanei*; another nematode, *Angiostrongylus cantonensis*; and, finally, a blood fluke, *Schistosoma mansoni*. In the case of the blood fluke, the ORF extends well into the RT domain, and the DNA immediately downstream of the termination codon is composed of pentamer and hexamer repeats similar to those in the RTE lineage (data not shown). Unfortunately, the sequence is truncated by cloning; thus, the precise 5' and 3' ends of the element cannot be identified. Given the conservation of this ORF in this element, it is very likely that full-length RTE elements also exist in this species.

Discussion

The RTE-1 element of *C. elegans* was first identified as a 3.3-kb insertion encoding an RT domain similar to non-LTR retrotransposable elements (Youngman, van Luenen, and Plasterk 1996). Because of the *C. elegans* genomic sequencing project, 8 full-length and 43 partial copies of RTE-1 elements have now been sequenced. The active element appears to encode a 1,024-amino-acid ORF with an AP-endonuclease in addition to the RT domain. The expression of this ORF may require a translational frameshift from a short preceding 43-amino-acid ORF that overlaps with the major ORF. In addition to the sequence relationship of their RT and endonuclease domains, several other features of the RTE-1 elements suggest they are typical non-LTR elements. First, they are not flanked by direct repeats. Second, variable-length target site duplications are generated upon their insertion. Third, over 80% of the RTE-1 copies contain large 5' truncations. Finally, the 3' junctions of RTE-1 elements are characterized by a variable number of short AT-rich nucleotide repeats. These structural features are all consistent with a model in which RTE-1 uses target-primed reverse transcription to insert into the host chromosomes (Luan et al. 1993; Luan and Eickbush 1995). Because of the unusually short first ORF and the less-than-100-bp 5' UTR and 3' UTR, the RTE-1 element of *C. elegans* at 3.3 kb is currently the shortest known non-LTR retrotransposable element. A second highly divergent family of elements is also present in *C. elegans*, RTE-2. This family of elements is not as abundant, and while a full-length element has not yet been sequenced, it is likely to be similar in structure to that of RTE-1.

Using the RTE-1 endonuclease and RT domains, we have identified homologous sequences in mammalian genomes that also appear to be similar in structural organization to the *C. elegans* elements. Thus, this study establishes the RTE-1 clade as the third lineage of non-LTR retrotransposons to be found in vertebrate genomes. The first lineage, which includes the Line1 elements (Tx1), whereas the second lineage, which includes the CR1 elements in chickens, is widespread in birds and reptiles (Haas et al. 1997; Kajikawa, Ohshima, and Okada 1997).

As shown in figure 4, the ORF structure of RTE-1 appears to be more similar to that of the second ORF of CR1 elements than to the second ORF of L1 elements. While all three elements encode both endonuclease and RT domains, CR1 and RTE-1 lack a cysteine-histidine motif found downstream of the RT domain in L1 and many other non-LTR elements. While the function of this cysteine-histidine motif is not known, it is required for L1 retrotransposition (Moran et al. 1996) and may be involved in protein binding to nucleic acids. RTE-1 differs from CR1 and, indeed, from all other non-LTR elements that contain an AP endonuclease domain in that it does not encode a large ORF preceding the endonuclease/RT-encoding ORF. The precise function of this first ORF in the retrotransposition of a non-LTR

element is not known. It is believed to encode an RNA-binding protein based on direct biochemical studies (Hohjoh and Singer 1996; Kolosha and Martin 1997) and by virtue of putative cysteine motifs that are similar to those of retroviral gag proteins (Dawson et al. 1997). The putative 43-amino-acid first ORF of RTE-1 has no recognizable motifs and is not positively charged; thus, it is unlikely to encode for an RNA-binding protein even if fused to the large ORF. Resolution of whether all RTE elements have a first small ORF will come from a more extensive analysis of sequences at the 5' ends of full-length RTE-2 and BDDF sequences.

We have also identified RTE-like elements in insects, molluscs, and flatworms. While the number of sequences currently available is limited, no data exist to suggest that horizontal transfers are needed to explain the very wide distribution of this clade. Thus, it is possible that this lineage of non-LTR retrotransposons might date back to the origin of all metazoa. We are currently conducting a comprehensive evaluation of all available non-LTR retrotransposable elements to determine whether the age of the RTE lineage is consistent with this estimate (unpublished data). However, this hypothesis can be effectively tested only with the identification of RTE elements in many diverse genomes. Given the number of large-scale sequencing projects underway today, such sequences will no doubt become available.

The present study also presents a simple model for the origin and mobility of an abundant class of SINEs in ruminant genomes. The RTE homolog in cows, called BDDF, was originally thought to be a separate element which exhibited sequence-specific insertion into a series of highly abundant SINE sequences (Art2 and Pst) in the cow genome (Szemraj et al. 1995). Rather than being the target sites, these SINEs are instead deletions of the full-length RTE homolog (Okada and Hamada 1997). The fact that these SINEs encode the carboxyl end of the RTE ORF has led to nucleic acid homology of these elements in different species. This unrecognized ORF and the paucity of RTE-like elements in other mammalian genomes (e.g., human, pig) led to the conclusion that the sequence conservation between Art2 and a SINE element in the viper was a horizontal transfer between bovine and viper genomes (Kordis and Gubensek 1995). While the sequence comparisons are still limited, there is no longer a strong argument to suggest that this is the case.

There are three currently known classes of SINEs, which are all believed to utilize the non-LTR retrotransposition machinery for their insertion. The first class is derived from a pol III RNA transcript and, other than ending with a poly(A) tail, has no apparent similarity to a non-LTR retrotransposable element. The only known example of this class is the primate Alu repeat, derived from 7SL RNA (Deininger 1989; Boeke 1997). The second class of SINEs is also derived from a pol III transcript, but contains the 3' end of a non-LTR retrotransposable element. Okada et al. (1997) have identified a number of SINEs in this class from diverse animals that are all derived from the fusion of a tRNA gene and the

3' end of a non-LTR retrotransposable element. Elements of the third class of SINEs are simply extreme 5' truncations of a non-LTR retrotransposable element and thus should probably not be considered authentic SINEs. The originally defined CR1 element in chickens (Silva and Burch 1989) and the above-mentioned Art2 and Pst repeats of cows (Szemraj et al. 1995) are examples of this group. It will be easier to identify extreme 5' truncated copies or potential new SINE families derived from the RTE lineage of retrotransposable elements than for any other lineage of non-LTR element. Because the 3' UTRs of the RTE family are so short, it is likely that the portion of the element's RNA sequence which is recognized by the RT domain also includes the region encoding part of the ORF. It is much easier to see homology for divergent amino acid sequences than it is for nucleotide sequences. This was certainly the reason we were able to identify distantly related RTE-like elements in divergent mollusc, nematode, and vertebrate genomes. In general, it will not be as easy to recognize SINEs as being derived from other non-LTR retrotransposable element lineages for which, as in the case of the R2 element, the minimum length of sequence required for recognition by the RT lies entirely within the 3' UTR of the element (Luan and Eickbush 1995).

Finally, the present study illustrates the utility of the different genome sequencing projects in identifying transposable elements that would otherwise have been missed. Although the original RTE-1 element was identified due to its insertion into the *prk-1* gene (Youngman, van Luenen, and Plasterk 1996), the critical features of the RTE-1 element or, indeed, the RTE lineage could not have been inferred but for the additional RTE-1 copies obtained by the *C. elegans* genome sequencing effort. In fact, the RTE-1 homolog identified in the genome of the flatworm *S. mansoni* was also available a result of another large sequencing project (M. Tanaka and T. Tanaka, personal communication). As these efforts progress in different genomes, divergent representatives of non-LTR retrotransposon lineages that have already been identified will doubtless be discovered. More enticing will be the discovery of totally novel lineages of retrotransposons, helping to further elucidate the age and evolution of these elements.

Acknowledgments

This work was supported by N.S.F. grant MCB-9601198 to T.H.E. We thank William Burke and Danna Eickbush for their comments. We especially thank Nori Okada for his insightful comments on the manuscript.

LITERATURE CITED

- ALTSCHUL, S. F., W. GISH, W. MILLER, E. W. MYERS, and D. J. LIPMAN. 1990. Basic local alignment search tool. *J. Mol. Biol.* **215**:403–410.
- BOEKE, J. D. 1997. LINES and Alus—the polyA connection. *Nat. Genet.* **16**:6–7.
- BULLOCK, P., W. FORRESTER, and M. R. BOTCHAN. 1984. DNA sequence studies of simian virus 40 chromosomal excision and integration in rat cells. *J. Mol. Biol.* **174**:55–84.

- BURKE, W. D., H. S. MALIK, W. C. LATHE III, and T. H. EICKBUSH. 1998. Are retrotransposons longterm hitchhikers? *Nature* **392**:141–142.
- CHARLESWORTH, B. 1988. The maintenance of transposable elements in natural populations. *Basic Life Sci.* **47**:189–212.
- CLARK, J. B., W. P. MADDISON, and M. G. KIDWELL. 1994. Phylogenetic analysis supports horizontal transfer of P transposable elements. *Mol. Biol. Evol.* **11**:40–50.
- DAWSON, A., E. HARTSWOOD, T. PATERSON, and D. J. FINNEGAN. 1997. A LINE-like transposable element in *Drosophila*, the I factor, encodes a protein with properties similar to those of retroviral nucleocapsids. *EMBO J.* **16**:4448–4455.
- DEININGER, P. L. 1989. SINES: short interspersed repeated DNA elements in higher eukaryotes. Pp. 619–636 in D. H. BERG and M. M. HOWE, eds. *Mobile DNA*. American Society of Microbiology, Washington, D.C.
- DUNCAN, C. H. 1987. Novel Alu-type repeats in artiodactyls. *Nucleic Acids Res.* **15**:1340.
- EICKBUSH, D. G., and T. H. EICKBUSH. 1995. Vertical transmission of the retrotransposable elements *R1* and *R2* during the evolution of the *Drosophila melanogaster* species subgroup. *Genetics* **139**:671–684.
- EICKBUSH, T. H. 1992. Transposing without ends: the non-LTR retrotransposable elements. *New Biol.* **4**:430–440.
- . 1994. Origin and evolutionary relationships of retroelements. Pp. 121–157 in S. S. MORSE, ed. *The evolutionary biology of viruses*. Raven Press, New York.
- FAWCETT, D. H., C. K. LISTER, E. KELLETT, and D. J. FINNEGAN. 1986. Transposable elements controlling I-R hybrid dysgenesis in *D. melanogaster* are similar to mammalian LINES. *Cell* **47**:1007–1015.
- FELSENSTEIN, J. 1993. PHYLIP (phylogeny inference package). Version 3.55. Distributed by the author, Department of Genetics, University of Washington, Seattle.
- FENG, Q., J. V. MORAN, H. H. KAZAZIAN JR., and J. D. BOEKE. 1996. Human L1 retrotransposon encodes a conserved endonuclease required for retrotransposition. *Cell* **87**:905–916.
- FURANO, A. V., B. E. HAYWARD, P. CHEVRET, F. CATZEFELIS, and K. USDIN. 1994. Amplification of the ancient murine Lx family of long interspersed repeated DNA occurred during the murine radiation. *J. Mol. Evol.* **38**:18–27.
- GARCIA-FERNANDEZ, J., J. R. BAYASCAS-RAMIREZ, G. MARFANY, A. M. MUNOZ-MARMOL, A. CASALI, J. BAGUNA, and E. SALO. 1995. High copy number of highly similar mariner-like transposons in planarian (Platyhelminth): evidence for a trans-phyta horizontal transfer. *Mol. Biol. Evol.* **12**:412–431.
- HAAS, N. B., J. M. GRABOWSKI, A. B. SIVITZ, and J. B. BURCH. 1997. Chicken repeat 1 (*CR1*) elements, which define an ancient family of vertebrate non-LTR retrotransposons, contain two closely spaced open reading frames. *Gene* **197**:305–309.
- HOHJOH, H., and M. F. SINGER. 1996. Cytoplasmic ribonucleoprotein complexes containing human LINE-1 protein and RNA. *EMBO J.* **15**:630–639.
- JACKS, T., and H. E. VARMUS. 1985. Expression of the Rous sarcoma virus pol gene by ribosomal frameshifting. *Science* **230**:1237–1242.
- JURKA, J. 1997. Sequence patterns indicate an enzymatic involvement in integration of mammalian retrotransposons. *Proc. Natl. Acad. Sci. USA* **94**:1872–1877.
- KAJIKAWA, M., K. OHSHIMA, and N. OKADA. 1997. Determination of the entire sequence of turtle *CR1*: the first open reading frame of the turtle *CR1* element encodes a protein with a novel zinc finger motif. *Mol. Biol. Evol.* **14**:1206–1217.
- KOLOSZA, V. O., and S. L. MARTIN. 1997. In vitro properties of the first ORF protein from mouse LINE-1 support its role in ribonucleoprotein particle formation during retrotransposition. *Proc. Natl. Acad. Sci. USA* **94**:10155–10160.
- KORDIS, D., and F. GUBENSEK. 1995. Horizontal SINE transfer between vertebrate classes. *Nat. Genet.* **10**:131–132.
- LATHE, W. C. III, W. D. BURKE, D. G. EICKBUSH, and T. H. EICKBUSH. 1995. Evolutionary stability of the *R1* retrotransposable element in the genus *Drosophila*. *Mol. Biol. Evol.* **12**:1094–1105.
- LATHE, W. C. III, and T. H. EICKBUSH. 1997. A single lineage of *R2* retrotransposable elements is an active, evolutionarily stable component of the *Drosophila* rDNA locus. *Mol. Biol. Evol.* **14**:1232–1241.
- LENSTRA, J. A. 1992. Bovine sequences in rodent DNA. *Nucleic Acids Res.* **20**:2892.
- LOHE, A. R., E. N. MORIYAMA, D. A. LIDHOLM, and D. L. HARTL. 1995. Horizontal transmission, vertical inactivation and stochastic loss of mariner-like transposable elements. *Mol. Biol. Evol.* **12**:62–72.
- LUAN, D. D., and T. H. EICKBUSH. 1995. RNA template requirements for target DNA-primed reverse transcription by the *R2* retrotransposable element. *Mol. Cell. Biol.* **15**:3882–3891.
- LUAN, D. D., M. H. KORMAN, J. L. JAKUBCZAK, and T. H. EICKBUSH. 1993. Reverse transcription of R2Bm RNA is primed by a nick at the chromosomal target site: a mechanism for non-LTR retrotransposition. *Cell* **72**:595–605.
- MARTIN, F., C. MARANON, M. OLIVARES, C. ALONSO, and M. C. LOPEZ. 1995. Characterization of a non-long terminal repeat retrotransposon cDNA (L1Tc) from *Trypanosoma cruzi*: homology of the first ORF with the Ape family of DNA repair enzymes. *J. Mol. Biol.* **247**:49–59.
- MODI, W. S., D. S. GALLAGHER, and J. E. WOMACK. 1996. Evolutionary histories of highly repeated DNA families among the Artiodactyla (Mammalia). *J. Mol. Evol.* **42**:337–349.
- MORAN, J. V., S. E. HOLMES, T. P. NAAS, R. J. DEBERARDINIS, J. D. BOEKE, and H. H. KAZAZIAN. 1996. High frequency retrotransposition in cultured mammalian cells. *Cell* **87**:917–927.
- NOBUKUNI, T., M. KOBAYASHI, A. OMORI et al. (12 co-authors). 1997. An *Alu*-linked repetitive sequence corresponding to 280 amino acids is expressed in a novel bovine protein, but not in its human homologue. *J. Biol. Chem.* **272**:2801–2807.
- OHSHIMA, K., M. HAMADA, Y. TERAJ, and N. OKADA. 1996. The 3' ends of tRNA-derived short interspersed repetitive elements are derived from the 3' ends of long interspersed repetitive elements. *Mol. Cell. Biol.* **16**:3756–3764.
- OKADA, N., and M. HAMADA. 1997. The 3' ends of tRNA-derived SINES originated from the 3' ends of LINES: a example from the bovine genome. *J. Mol. Evol.* **44**:S52–S56.
- ROBERTSON, H. M. 1993. The mariner transposable element is widespread in insects. *Nature* **362**:241–245.
- OKADA, N., M. HAMADA, I. OGIWARA, and K. OHSHIMA. 1997. SINES and LINES share common 3' sequences: a review. *Gene* **205**:229–243.
- SAITOU, N., and M. NEI. 1987. The neighbor-joining method: a new method for reconstructing phylogenetic trees. *Mol. Biol. Evol.* **4**:406–425.
- SILVA, R., and J. B. BURCH. 1989. Evidence that chicken CR1 elements represent a novel family of retrotransposons. *Mol. Cell. Biol.* **9**:3563–3566.
- SPRINGER, M. S., N. A. TUSNEEM, E. H. DAVIDSON, and R. J. BRITTEN. 1995. Phylogeny, rates of evolution, and patterns

- of codon usage among sea urchin retroviral-like elements, with implications for the recognition of horizontal transfer. *Mol. Biol. Evol.* **12**:219–230.
- SZEMRAJ, J., G. PLUCIENNICZAK, J. JAWORSKI, and A. PLUCIENNICZAK. 1995. Bovine *Alu*-like sequences mediate transposition of a new site-specific retroelement. *Gene* **152**: 261–264.
- THOMPSON, J. D., D. G. HIGGINS, and T. J. GIBSON. 1994. CLUSTAL W: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice. *Nucleic Acids Res.* **22**:4673–4680.
- WARREN, A. M., M. A. HUGHES, and J. M. CRAMPTON. 1997. *Zebedee*: a novel *copia-Ty1* family of transposable elements in the genome of the medically important mosquito *Aedes aegypti*. *Mol. Gen. Genet.* **254**:505–513.
- XIONG, Y., and T. H. EICKBUSH. 1990. Origin and evolution of retroelements based upon their reverse transcriptase sequences. *EMBO J.* **9**:3353–3362.
- YOUNGMAN, S., H. G. A. M. VAN LUENEN, and R. H. A. PLASTERK. 1996. Rte-1, a retrotransposon-like element in *Caenorhabditis elegans*. *FEBS Lett.* **380**:1–7.
- PIERRE CAPY, reviewing editor

Accepted May 18, 1998