Cross-Genome Screening of Novel Sequence-Specific Non-LTR Retrotransposons: Various Multicopy RNA Genes and Microsatellites Are Selected as Targets

Kenji K. Kojima and Haruhiko Fujiwara

Department of Integrated Biosciences, Graduate School of Frontier Sciences, University of Tokyo

Although most LINEs (long interspersed nuclear elements), which are autonomous non–long-terminal-repeat retrotransposons, are inserted throughout the host genome, three groups of LINEs, the early-branched group, the Tx group, and the R1 clade, are inserted into specific sites within the target sequence. We previously characterized the sequence specificity of the R1 clade elements. In this study, we screened the other two groups of sequence-specific LINEs from public DNA databases, reconstructed elements from fragmented sequences, identified their target sequences, and analyzed them phylogenetically. We characterized 13 elements in the early-branched group and 13 in the Tx group.

In the early-branched group, we identified R2 elements from sea squirts and zebrafish in this study, although R2 has not been characterized outside the arthropod group to date. This is the first evidence of cross-phylum distribution of sequence-specific LINEs. The Dong element also occurs across phyla, among arthropods and mollusks. In the Tx group, we characterized five novel sequence-specific families: Kibi for TC repeats, Koshi for TTC repeats, Keno for the U2 snRNA gene, Dewa for the rRNA tandem arrays, and Mutsu for the 5S rRNA gene. Keno and Mutsu insert into the highly conserved region within small RNA genes and destroy the targets. Several copies of Dewa insert different positions of tRNA tandem array, which indicates a certain “site specifier” other than sequence-specific endonuclease. In all three groups, LINEs specific for the rRNA genes or microsatellites can occur as multiple families in one organism. This indicates that the copy number of a target sequence is the primary factor to restrict the variety of sequence specificity of LINEs.

Introduction

Long interspersed nuclear elements (LINEs) are autonomous non–long-terminal-repeat (non-LTR) retrotransposons, encode a reverse transcriptase, and insert into various genomic locations via RNA intermediates. Recent genome sequencing projects have provided remarkable insights into the evolution of LINEs and their role in shaping eukaryotic genomes (Lander et al. 2001; Aparicio et al. 2002). One of the remarkable features of LINEs is their abundance throughout the genome. In contrast, several LINEs show very restricted distribution in the genome, such as ribosomal RNA gene loci or telomere (Jakubczak, Burke, and Eickbush 1991; Okazaki, Ishikawa, and Fujiwara 1995).

Sequence specificity in retrotransposition is seen in three groups of LINEs: the early-branched group, the Tx group, and the R1 clade group. Phylogenetically, LINEs have been classified into at least 15 clades (fig. 1) (Malik, Burke, and Eickbush 1999; Malik and Eickbush 2000; Lovsin, Gubensek, and Kordi 2001; Burke et al. 2002). The early-branched group consists of five clades, Genie, CRE, R4, R2, and NeSL. These include only one open reading frame (ORF) and this ORF encodes a restriction-enzyme–like endonuclease (RLE) at the C-terminal region. Most elements in these five RLE-encoding clades are sequence specific. Their target sequences are such as the spliced-leader exons (Aksoy et al. 1990; Gabriel et al. 1990; Villanueva et al. 1991; Malik and Eickbush 2000), the 28S rRNA gene (Burke, Müller, and Eickbush 1995), microsatellites (Xiong et al. 1993), or the subtelomeric repeats (Burke et al. 2002). In contrast, most of the recently branched clades, which encode two ORFs and an apurinic/apyrimidinic endonuclease-like endonuclease (APE), do not insert in a sequence-specific manner into the host genome. Among the APE-encoding–type LINEs, only two groups are known to be sequence specific. One is the Tx group, which comprises two elements, Tx1L and Tx2L from Xenopus laevis, and is classified into the L1 clade (Garrett, Knutson and Carroll 1989). Tx1L and Tx2L insert into other transposons Tx1D and Tx2D, respectively. The other group comprises several elements within the R1 clade in arthropods.

We previously reported that there are at least eight types of sequence specificity in the R1 clade. However, only three highly repetitive sequences are selected as their target sequences (Kojima and Fujiwara 2003). In this paper, we characterize a variety of LINEs from the other two sequence-specific groups, the early-branched group, and the Tx group and discuss their selection for target sequences.

Materials and Methods

Database Analysis and Element Reconstruction

To intensively screen LINEs belonging to the early-branched group and the Tx group, we performed computer-based nucleotide and protein searches of the GenBank databases, using different Blast search programs (Altschul et al. 1990) in NCBI (www.ncbi.nlm.nih.gov), Nonredundant (NR), dbEST (expressed sequence tags database), dbSTS (sequence tag sites database), dbGSS (genome survey sequences database), HTGS (unfinished high-throughput genomic sequences phase 0, 1, and 2), WGS anopheles (Anopheles gamhiae whole-genome shotgun sequences database). In addition, all eukaryotic genomic sequences in NCBI, including Trace Archive database were used for screening. We also searched
Phylogenetic trees were constructed by the neighbor-joining method, using the MEGA2 program (Kumar et al. 2001). The significance of the various phylogenetic lineages was assessed by bootstrap analysis. All parameters used in both programs were default.

PCR Amplification and Sequence Analyses

Genomic DNAs of the sea squirt, Ciona intestinalis, and the zebrafish, Danio rerio, were kindly provided by Drs. Y. Kimura and S. Kawamura of the University of Tokyo, respectively. To amplify the 3′ junction between 28S rDNA and R2 elements, we used a primer, 28S-R-A (5′-TAGATGACGAGGCAATTTGC-3′), in the highly conserved region of 28S rDNA and several primers within the 3′-UTR regions of respective R2 elements: R2CiA-F3413, 5′-GGCCAGAAAACTCGAGGTAA-3′ for R2Ci-A; R2CiB-F3401, 5′-TGCACTTGGCAACAC-3′ for R2Ci-B; R2CiC-F2004, 5′-TAGCTGAACGAGACAGAAAAG-3′ for R2Ci-C; R2CiD-F3261, 5′-CGATGACGTAGAAAGCTTTG-3′ for R2Ci-D; R2Dr-F4983, and 5′-ACGAGAAACGCAACGC-3′ for R2Dr. PCR was performed for 35 cycles of 96°C for 30 s, 56°C for 30 s, and 72°C for 1 min. PCR products were cloned into the pGEM-T Easy vector (Promega) and sequenced with ABI-310 DNA sequencer (PE Applied Biosystems).

Accession Number

Nucleotide sequences of reconstructed elements in this study are deposited under accession numbers AB097121 to AB097148, AB111947, and AB111948.

Results

Early-Branched LINE

Using the RT domains and RLE domains of various LINEs (see Materials and Methods), we identified several members of the early-branched group. Although these sequences were quite fragmented, we reconstructed the complete or relatively long ORFs of 13 elements (fig. 2), and constructed a phylogenetic tree using their RT domains (fig. 3a). Of these 13 elements newly identified in this study, six elements (R2Ci-A, R2Ci-B, R2Ci-C, R2Ci-D, R2Cs-D, and R2Dr) belong to the R2 clade, three (DongAg, EhRLE2, and EhRLE3) belong to the R4 clade, and the other four (HERODr, HEROTn, HEROFr, and YURECi) belong to the NeSL clade. We could not reconstruct several elements belonging to the early-branched group because of insufficient sequences. Of these, we classified some elements into clades on the basis of sequence similarity (table 1).

R2 Clade

We characterized four subfamilies (R2Ci-A, R2Ci-B, R2Ci-C, and R2Ci-D) belonging to R2 family in the sea squirt Ciona intestinalis. Of the four R2Ci subfamilies, we could identify only one subfamily (R2Ci-D) in the closely related sea squirt C. savignyi, probably because the sequences of this organism have been analyzed less...
completely than those of *C. intestinalis*. Therefore, it is possible that *C. savignyi* contains the other three subfamilies corresponding to those of *C. intestinalis*. Of the five R2 elements in the two species of sea squirts, R2Ci-D and R2Cs-D are more closely related to R2-Limulus (the R2 element from horseshoe crab) than to other sea squirt R2 subfamilies. The R2 element (R2Dr) from the zebrafish *Danio rerio* is also phylogenetically close to R2-Limulus.

Arthropod R2 elements were proposed to be divided into three groups (Burke et al. 1999). Lineage I includes R2-Limulus and R2-NasoniaB (*Nasonia*; jewel wasp), which contain three zinc-finger motifs and one c-myb DNA binding motif near the N-terminus. Lineage II includes coleopteran five elements represented by R2-PopolliaC, but the N-terminal regions of these elements have not been characterized. The other arthropod R2 elements, such as R2Bm, contain one zinc-finger motif and one c-myb motif and are classified into lineage 3. Our data of RT phylogeny in figure 3a, however, did not clearly support this hypothesis, probably because elements used in the analysis were too divergent. Then, we further investigated the phylogeny among R2 elements using the expanded region from RT to RLE (fig. 3b). The phylogenetic tree supports the two major branches (lineage I and II) in the R2 clade. Lineage I in this paper includes lineage 1 proposed by Burke et al. (1999). Lineage II corresponds to lineage 3. We had no additional data to characterize lineage 2. Among chordate R2 elements, R2Dr, R2Ci-D, and R2Cs-D belong to lineage I, and R2Ci-A, B, and C belong to lineage II. In the view of the N-terminal structure, R2Dr, R2Ci-D, and R2Cs-D have three zinc-finger motifs near the N-terminus (fig. 2), which is characteristic to lineage I. In contrast, R2Ci-A, R2Ci-B, and R2Ci-C have only one zinc-finger motif and one c-myb motif, similar to the more closely related insect R2 elements belonging to lineage II. Their structural features correspond to their phylogeny. These results indicate that the two lineages have been maintained since the early stage of animal evolution.

All identified 3′ end sequences of the chordate R2 elements flank the 28S rRNA gene (table 2), which indicates that these elements insert specifically into the same position of 28S rRNA gene as the arthropod R2 elements. Because these sequences were obtained from the trace archive database, and the possibility of contamination was not completely excluded, we carefully investigated the sources of these sequences. The 200-bp flanking 28S rRNA gene sequences of the five sea squirt R2 elements are completely identical to the reported *C. intestinalis* rRNA gene sequence, except one additional G residue in R2Cs-D (fig. 4a). R2Ci-A and R2Ci-B have variations of the 3′ A-rich tail in length and sequence, and we identified the 5′ truncated copies of R2Ci-D. Multiple copies of these elements suggested that they are not contaminants from other organisms. Furthermore, the existence of two related R2 elements in the sibling species of sea squirts also supported the above idea. The flanking 28S rDNA sequence of R2Dr is consistent with the reported zebrafish sequence and clearly closer to the 28S sequences conserved among vertebrates than those of insects (fig. 4a), indicating that the authentic R2 element exists in the zebrafish genome. Finally, we have confirmed the existence of R2 elements in the *C. intestinalis* and *D. rerio* genome using PCR (fig. 4b). Primers were designed within the 3′-UTR region of each R2 element and within the flanking region of the target site in the 28S rDNA (28S-R-A). As shown by PCR bands in the expected size, we confirmed four R2 elements in the *C. intestinalis* genome and one in the *D. rerio* genome (fig. 4b). Sequence analyses of PCR products revealed that they represent the true band and that R2Ci-A has two 3′-UTR variants (data not shown). From these findings, we concluded that these R2 elements were undoubtedly derived from the chordate genomes. This is the first evidence of cross-phylum occurrence of the same sequence-specific LINE family and the first report of an rDNA-specific LINE in vertebrates. We also searched R2 elements in other vertebrates, including human, mouse, *Fugu rubripes*, and *Tetraodon nigroviridis*, but we could not identify R2 elements.

We also identified the R2 clade elements in the blood fluke *Schistosoma mansoni* and two plant species, *Zea mays* (maize) and *Oryza sativa* (rice), although we could not characterize their target sequences because the information available was incomplete (table 1). These findings suggest that the R2 elements may spread across kingdoms.
FIG. 3.—(a) Neighbor-joining (NJ) trees of early-branched group constructed using the RT domain. The phylogeny is rooted on the mobile group II intron sequences. The subtree of the L1 clade is shown in figure 6. The number next to each node indicates a value as a percentage of 1,000 replicates. Elements newly identified in this study are shown with circles (○). Sequence-specific elements are shown with asterisks (*). Abbreviations of host organisms of newly identified elements are as follows: Ag = Anopheles gambiae; Eh = Entamoeba histolytica; Dr = Danio rerio; Tn = Tetraodon nigroviridis; Fr = Fugu rubripes; Ci = Ciona intestinalis; Cs = Ciona savignyi. (b) NJ trees of R2 clade elements constructed using the region from RT to RLE.
R4 Clade

DongAg, the Dong element from the African malaria mosquito *Anopheles gambiae*, is similar to the Dong element from the silkworm *B. mori* with regard to phylogenetic position, ORF structure, and sequence specificity. Dong from *B. mori* possesses no additional motifs at the N-terminus and inserts into TAA repeats specifically. DongAg also has no motifs near the N-terminus and inserts sequence specifically into TAA repeats (table 2). We also identified an element belonging to the Dong family from the bloodfluke planorb *Biomphalaria glabrata* (table 1). This element also inserts into TAA repeats and has an RLE domain very similar to those of the other two Dong elements (fig. 5). Thus, the cross-phylum distribution of the Dong family occurs between the phyla Arthropoda and Mollusca. A LINE very closely related to Dong has been reported on teleosts, designated Rex6. However, these elements do not show sequence specificity for TAA repeats but have a preference for microsatellites such as CA or TA repeats (Volff et al. 2001).

EhRLE2 and EhRLE3 are related to EhRLE1 (recently reported as “EhRLE” [Sharma et al. 2001]) and constitute an independent branch in the R4 clade. Previous reports suggested that EhRLE1 has no RLE domain, but we detected a defective RLE domain downstream from the RT domain (figs. 2 and 5). The three EhRLE elements do not have clear sequence specificity. While preparing this paper, Van Dellen et al. reported three LINE families named EhLINE1 to EhLINE3 from *E. histolytica* (Van Dellen et al. 2002). EhRLE1, EhRLE2, and EhRLE3 in this study correspond to EhLINE1, EhLINE3, EhLINE2, respectively.

We also identified other elements belonging to the R4 clade (table 1). The element from the nematode *Strongyloides ratti* is similar to the R4 elements from other nematodes. Elements from the Nile tilapia *Oreochromis niloticus* and the sea squirt *Ciona intestinalis* are similar to Rex6 from the teleosts.

NeSL Clade

Three HERO elements from teleost and YURECi from the sea squirt are the most closely related to NeSL-1 from *C. elegans*, although the monophyly of these elements is not very reliable (fig. 3a). Burke et al. (2002) proposed that NeSL-1 belongs to the R4 clade. However, our results indicate that NeSL-1 and HERO, and possibly YURE, form a distinct and novel branch named NeSL clade. HERO does not have additional motifs at the N-terminus (fig. 2). YURECi has three additional zinc-finger motifs, although NeSL-1 has only two additional zinc-finger motifs near the N-terminus (fig. 2). NeSL-1 has one cysteine protease domain between the N-terminal zinc-finger motifs and the RT domain. However, YURECi does not have a protease domain (fig. 2). The three HERO elements and YURECi do not insert in a sequence-specific manner (table 2), although HEROFr inserts into TA repeats at some extent (8/62). A fragmented sequence similar to HERO has also been identified in the sea urchin *Strongylocentrotus purpuratus* (table 1).

Recently Branched LINE

Using the same approach as we used to characterize the early-branched LINEs shown above, we identified 13 elements (KoshiFr1, KoshiTn1, KibiFr1, KibiTn1, ...
KibiDr1, KibiDr2, MutsuDr1, MutsuDr2, MutsuDr3, DewaDr1, KenoDr1, KenoFr1, and KenoTn1) very similar to Tx1L from three teleost fishes, the zebrafish *Danio rerio*, the green spotted pufferfish *Tetraodon nigroviridis*, and the torafugu *Fugu rubripes* (*Takifugu rubripes*). We also reconstructed these elements, analyzed their phylogenetic relationships (fig. 6), and characterized their target sequences (table 3). Of these elements, two from *F. rubripes* was already described as Tx_Fr1 and Tx_Fr2, although their target sequences had not been analyzed (Aparicio et al. 2002). Tx1L and these new LINEs clearly form a phylogenetic branch distinct from the L1 and swimmer elements, which are the most abundant LINEs in vertebrates. We also identified sea squirt elements (L1Ci-A, L1Ci-B, and L1Ci-C), which map to the L1 branch. This indicates that the L1 and Tx groups coexisted before the evolution of vertebrates.

**Tx Group**

We classified the new elements of the Tx group into Kibi, Koshi, Keno, Dewa, and Mutsu, on the basis of their

---

**Fig. 4.**—Confirmation of existence of R2 elements in chordate genomes. (a) Alignment of 28S rRNA gene sequences downstream of R2 insertion site. The 28S rRNA gene of vertebrates, sea squirts, and insects are shown, in addition to representatives of the raw sequences flanking the R2 elements (shown as “element name_flanking_sequence”). Asterisks (*) indicate the site identical among all species. Bases identical among vertebrates or chordates and not conserved among insects are shaded. Primer 28S-R-A used in PCR analysis is also shown. (b) PCR analysis. Expected size of PCR products are as follows: R2Ci-A = 446 bp; R2Ci-B = 528 bp; R2Ci-C = 698 bp; R2Ci-D = 975 bp; R2Dr = 515 bp. R2Ci-A and R2Ci-B contain variable-length polyA tail. Two bands of R2Ci-A turned out to be corresponding to the 3′-UTR variants by sequencing.

---

KibiDr1, KibiDr2, MutsuDr1, MutsuDr2, MutsuDr3, DewaDr1, KenoDr1, KenoFr1, and KenoTn1) very similar to Tx1L from three teleost fishes, the zebrafish *Danio rerio*, the green spotted pufferfish *Tetraodon nigroviridis*, and the torafugu *Fugu (Takifugu) rubripes*. We also reconstructed these elements, analyzed their phylogenetic relationships (fig. 6), and characterized their target sequences (table 3). Of these elements, two from *F. rubripes* was already described as Tx_Fr1 and Tx_Fr2, although their target sequences had not been analyzed (Aparicio et al. 2002). Tx1L and these new LINEs clearly form a phylogenetic branch distinct from the L1 and swimmer elements, which are the most abundant LINEs in vertebrates. We also identified sea squirt elements (L1Ci-A, L1Ci-B, and L1Ci-C), which map to the L1 branch. This indicates that the L1 and Tx groups coexisted before the evolution of vertebrates.

**Tx Group**

We classified the new elements of the Tx group into Kibi, Koshi, Keno, Dewa, and Mutsu, on the basis of their
target sequences. Kibi inserts into microsatellite TC repeats, and Koshi inserts into TTC repeats (table 3). Interestingly, about half (21/47) of KoshiTn1 insertion sites are 3′ terminal GAA repeats of other LINE elements (data not shown). The target sequences of Keno, Dewa, and Mutsu are the U2 small nuclear RNA gene, the Leu tRNA gene spacer region, and the 5S rRNA gene, respectively (table 3). The insertion site of Keno is 37 nucleotides downstream of the 5′-terminus of U2 snRNA. The 69 nucleotides around the Keno insertion site are identical to that of human U2 snRNA, and 61 nucleotides are identical to that of rice. The Leu tRNA gene repetitive unit of zebrafish is about 1 kb in length, in which there are two copies of tRNA genes. In the spacer region, Dewa inserts into the specific site, which is about 200 bp downstream of one Leu tRNA copy and about 500 bp upstream of the other copy. The Leu tRNA repetitive unit is not found in *Fugu rubripes* and *Tetraodon nigroviridis*. The insertion sites of three Mutsu elements (MutsuDr1-Dr3) are located about 20 nucleotides downstream of the 5′-terminus of the 5S rRNA gene. The insertion site for MutsuDr1 is 2 bp upstream from those for MutsuDr2 and Dr3 (table 3). The insertion of Keno or Mutsu destroys the target genes. We identified Kibi, Koshi, and Keno elements in three teleost species, but we could not identify the Dewa and Mutsu elements in *F. rubripes* and *T. nigroviridis*. Taking the progress of two genome projects into consideration, they could be extinguished in pufferfishes.

Previously, we reported that the close relatives in the R1 clade have the similar target sequences (Kojima and Fujiwara 2003). The target sequences of Kibi and Koshi are similar, but there is no similarity among target sequences of other elements. Although most of DewaDr1 (46/58) insert into a specific position within the tRNA spacer, some (5/58) insert into other positions within the tRNA gene array. In other words, DewaDr1 usually transposes both in a “sequence-specific” and a “site (tRNA)-specific” manner, but sometimes in a “site-specific” but not “sequence-specific” manner. It has been suggested that endonuclease (EN) domains encoded in Tx1L and Tx2L cannot fully distinguish their own target sequences and that other domains, or possibly RNA itself, contribute to their sequence specificities (Christensen, Pong-Kingdon, and Carroll 2000). Recently, human L1 was reported to have a strong preference for the TTAAAA sequence (Szafranski et al. 2002). However, L1 could not transpose to a specific location because of its short target sequence. In contrast, TRE-3 and TRE-5 also insert intensively near tRNA genes, but these elements do not insert sequence specifically (Szafranski et al. 1999). Zepp is another element in the L1 clade that exhibits
"site-specificity" but not "sequence-specificity" (Higashiyama et al. 1997). The insertion sites of DewaDr1, which inserts "site specifically" into the tRNA region, indicate that factors other than the EN domain can recognize target sequences more approximately. It might be possible that sequence specificity in the Tx group is provided with a sequence preference of the EN activity and simultaneously with a site specificity that derives from other unknown domains. That says the "site specifier" could be an additional sequence-specific DNA-binding domain or a domain that can recognize local chromatin or DNA structure.

Phylogenetically, the Tx group consists of two branches: one formed by Kibi, Koshi, and Tx1L, and the other formed by Keno, Dewa, and Mutsu (fig. 6). The target sequences of elements on the two branches are dissimilar. The targets of Kibi/Koshi/Tx1L are microsatellites and transposons, which are dispersed throughout the genome. In contrast, the targets of Keno/Dewa/Mutsu are short RNA genes, which are usually repeated tandemly at one locus. This difference supports the above idea that domains other than the EN domain contribute to target specificity by recognizing the broad structure of target sequences.

**Discussion**

**Cross-Phylum Distribution of Sequence-Specific LINE**

In this study, we identified several target sequence-specific elements. To date, many sequence-specific LINE families have been identified, although they have been reported only in one phylum, generally among very restricted species. For instance, TRAS family is found only in three Lepidopteran insects (Bombyx mori, Samia cynthia, and Saturnia japonica) (Kubo et al. 2001), and the R4 family is found only in three nematode species (Ascaris lumbricoides, Parascaris equorum, and Haemonchus contortus) (Burke, Müller, and Eickbush 1995). R1 and R2 families, which are the most characterized groups in sequence-specific LINEs, have been known only in arthropods (Burke et al. 1998). Here, we identified R2 elements in three chordate species. The cross-phyllum existence of R2 family in chordates and arthropods indicates that sequence specificity for the 28S rRNA gene was generated in the Cambrian era at farthest, before the branch between deuterostomes (including chordates) and protostomes (including arthropods), and conserved for at least 500 Myr.

One reason for the long-term hitchhiking of R2 is the conservation of target sequences. R2Bm protein interacts with a region extending from 35 bp upstream to 15 bp downstream of the target site (Xiong and Eickbush 1988). The 50-bp target sequence of 28S rDNA is highly conserved not only between chordates and arthropods but also among all eukaryotes, with one or two substitutions (see fig. 4a). There is one nucleotide substitution near the R2 target sites between D. melanogaster and B. mori. However, R2 of B. mori can insert into the 28S rRNA gene of D. melanogaster (Eickbush, Luan, and Eickbush 2000), suggesting that R2 target recognition allows a few substitutions and that the tolerance for subtle nucleotide changes has supported the long-term hitchhiking of R2 elements.

Targeting to rDNA is not always advantageous to the R2 elements. It has been suggested that the rDNA tandem arrays undergo concerted evolution (Coen, Thoday, and Dover 1982). Concerted evolution homogenizes the target repetitive sequences, which could provide the identical and multiple target sites for sequence-specific LINEs. However, once LINEs insert into the targets, concerted evolution can work disadvantageously on the sequence-specific LINEs. In D. melanogaster, concerted evolution seldom duplicates inserted R1 and R2 elements, rather eliminates insertion sequences from rDNA, and the elimination rate is quite higher than that of other retrotransposons (Perez-Gonzalez and Eickbush 2002). Therefore, R2 must retrotranspose more actively than non-sequence-specific LINEs. Rapid elimination might have taken away the R2 element in several lineages. The copy number of full-length R2 is variable even in one species (Perez-Gonzalez and Eickbush 2002). The extinction of R2 occurred in many lineages in both arthropods and chordates, although many insects still retain R2 elements.

In addition, the cross-phyllum distribution of Dong in arthropods and mollusks also indicates their emergence in the Cambrian era. In contrast to the early-branched LINEs, two sequence-specific groups of recently branched LINEs, Tx and R1 clade groups, have no evidence for the
cross-phylum existence. It remains unclear when these two groups acquired sequence specificity.

Target Selection

Although dozens of sequence-specific LINEs have been discovered, only a few repetitive sequences are selected as targets. We compared the target sequences of three distinct sequence-specific groups (table 4). The target sequences of sequence-specific LINEs are all highly repetitive elements, whether they are necessary or unnecessary, tandem or interspersed. The ribosomal RNA genes are selected as targets by at least seven families of three groups. Of these, the 28S rRNA gene is selected by five families. The 5S rRNA genes are separately organized into their own cluster. Mutsu is the first family identified that inserts into the 5S rRNA gene. The other targets selected by all three groups are microsatellites. The high copy number of target sequences appears to be correlated with the existence of multiple subfamilies inserting into the same target sequences. All families targeting to the 18S-5.8S-28S rRNA gene clusters, except R4, include multiple subfamilies in certain organisms (Burke et al. 1998; Kojima and Fujiwara 2003). Mutsu, which targets the 5S rRNA gene, also includes three subfamilies in zebrafish. The 28S rRNA is generally several hundred. Waldo elements, which inserts into ACAY repeats, Kibi, which inserts into TC repeats, and TRAS, which inserts into telomeric repeats, also include multiple elements (Kubo et al. 2001; Kojima and Fujiwara 2003). The copy numbers of microsatellites and telomeric repeats are virtually limitless. In contrast, Dewa, which is specific for the Leu tRNA gene, and Keno, which is specific for the U2 snRNA gene, consist of single subfamilies in zebrafish. The copy numbers of rRNA gene and U2 snRNA are, at most, several dozens. These facts suggest that one sequence-specific LINE requires several dozens of target repetitive sequences to survive against the selective pressure during evolution.

In this study, we newly identified several target sequences: three small RNA genes, the 5S rRNA, the Leu tRNA, and the U2 snRNA, in addition to the 28S rRNA, the 18S rRNA, and the spliced-leader exons (table 4). Why do these repetitive RNA genes attract the sequence-specific LINEs? Since rRNAs, tRNAs, and snRNAs have universal and essential functions and work as RNA molecules, their functional regions are highly conserved at the nucleotide level against the selective pressure during evolution. On this point, the highly conserved region within the essential RNA genes is an ideal target for sequence-specific LINEs, because its target sequence should not be changed. Actually, the insertion sites of six 28S rDNA–specific LINEs (R1, R2, R4, R6, R7, and RT) reside in the most conserved regions in the 28S rRNA gene (Ben Ali et al. 1999). The insertion sites of Mutsu in the 5S rRNA gene and of Keno in the U2 snRNA gene are also within the most conserved regions of respective RNA genes.

Progress in the various genome-sequencing projects would reveal more new sequence-specific LINEs. Among other RNA genes, is there any candidate to be a target for sequence-specific LINEs? Single-copy genes such as the telomerase RNA and snoRNA genes will be excluded from the candidates. In the human genome, there are more than 10 copies for some RNA genes; U1 and U6 snRNA genes and the Y3 RNA gene, which is a component of Ro RNP (Lander et al. 2001). The Y3 RNA gene, which is the most conserved among Y RNAs in vertebrates (Teunissen et al. 2000), may be one candidate for the target of a sequence-specific LINE.

Protein multigene families, such as histones, are other candidates. The copy numbers of histone families in a species are variable, but some species have quite many copies, up to 1,000 (Kedes 1979). The protein sequence identity among copies is quite high, whereas the protein multigene families evolve according to the model of birth-and-death evolution under strong purifying (negative) selection (Plontkivska, Rooney, and Nei 2002) but not to the concerted evolution. In this case, new copies are created by repeated gene duplication and some are maintained for a long time, but others are deleted or become nonfunctional. The concerted evolution provides nucleotide sequence homogeneity, but purifying selection at the protein level allows synonymous substitutions on the
DNA sequence. In fact, the numbers of synonymous changes among histone H4 member genes on the same genome are nearly at the saturation level (Piontkivska, Rooney, and Nei 2002). Such diversified DNA sequences of multiple histone genes would not be appropriate targets for the sequence-specific LINEs.

Satellite repeats, transposable elements, microsatellites, and telomeric repeats are abundant on the genome (Lander et al. 2001; Aparicio et al. 2002). Satellite repeats usually evolved concertedly (Ugarkovic and Plohl 2002), which gives rise to the sequence homogeneity among copies. Transposable elements are amplified in the genome by active transposition. Repetitive genomic copies of these two elements provide many targets for selfish genes, but they are variable in sequences and may resist long-term inheritance of sequence-specific LINEs. In contrast, similar sequences of microsatellites and telomeric repeats are widely observed among various organisms. In this study, we revealed that Dong microsatellite-specific LINE is distributed across two phyla, suggesting that microsatellite-specific LINE can be inherited for a long time. Taking the copy number, the sequence homogeneity, and the long-term sequence conservation into consideration, RNA genes, microsatellites, and telomeric repeats are best targets for sequence-specific LINEs.

Acknowledgments

This work was supported by grants from the Ministry of Education, Science, and Culture of Japan and by a Grant-in-Aid from the Research for the Future Program of the Japan Society for the Promotion Science.

Literature Cited

Aksoy, S., S. Williams, S. Chang, and F. F. Richards. 1990. SLACS retrotransposon from Typanosoma brucei gambiense is similar to mammalian LINEs. Nucleic Acids Res. 18:785–792.


Thomas Eickbush, Associate Editor

Accepted August 3, 2003