Functional noncoding sequences derived from SINEs in the mammalian genome

Hidenori Nishihara,[1] Arian F.A. Smit,[2] and Norihiro Okada[1,3]

[1]Graduate School of Bioscience and Biotechnology, Tokyo Institute of Technology, Yokohama, Japan; [2]Institute for Systems Biology, Seattle, Washington 98103, USA

Recent comparative analyses of mammalian sequences have revealed that a large number of nonprotein-coding genomic regions are under strong selective constraint. Here, we report that some of these loci have been derived from a newly defined family of ancient SINEs (short interspersed repetitive elements). This is a surprising result, as SINEs and other transposable elements are commonly thought to be genomic parasites. We named the ancient SINE family AmnSINE1, for Amniota SINE1, because we found it to be present in mammals as well as in birds, and some copies predate the mammalian-bird split 310 million years ago (Mya). AmnSINE1 has a chimeric structure of a 5S rRNA and a tRNA-derived SINE, and is related to five tRNA-derived SINE families that we characterized here in the coelacanth, dogfish shark, hagfish, and amphioxus genomes. All of the newly described SINE families have a common central domain that is also shared by zebrafish SINE3, and we collectively name them the DeuSINE (Deuterostomia SINE) superfamily. Notably, of the ~1000 still identifiable copies of AmnSINE1 in the human genome, 105 correspond to loci phylogenetically highly conserved among mammalian orthologs. The conservation is strongest over the central domain. Thus, AmnSINE1 appears to be the best example of a transposable element of which a significant fraction of the copies have acquired genomic functionality.

[Supplemental material is available online at www.genome.org.]

Recent genome sequencing projects have revealed that the protein-coding regions in DNA occupy only ~1.5% of the genome. However, recent cross-species comparative analyses identified a number of sequences phylogenetically conserved among mammalian orthologs (Bejerano et al. 2004a). Regions in the human genome under purifying selection are estimated to comprise about 5% (Lander et al. 2001; Waterston et al. 2002). The remaining 3.5% of conserved DNA may be *cis*-regulatory elements, micro-RNAs, etc. (Dermitzakis et al. 2002), and many other regions may provide essential but unknown roles in the genome. Furthermore, similar comparative analyses have identified many interspersed conserved sequences. For example, Bejerano et al. (2004b) listed ~5000 kinds of sequence categories, each of which is conserved among mammals and is present in more than one copy in the human genome. The detailed functions of most of these regions have not been elucidated.

Eukaryotic genomes generally contain a large number of retroposons that propagate within the host genome via RNA intermediates (Rogers 1985; Weiner et al. 1986; Brosius 1991). For example, 42% of the human genome consists of retroposons that have been analyzed in detail (Lander et al. 2001). SINEs (short interspersed elements) and LINEs (long interspersed elements) are two major classes of retroposons. Typically, SINEs are 75–500 bp long and contain internal promoters for RNA polymerase III (pol III) (Okada 1991a,b). The promoter of almost all known SINE families is derived from tRNA, with two exceptions in which the promoter is derived from 7SL RNA or 5S rRNA. The 7SL RNA-derived SINEs have been identified only in the genomes of primates, rodents, and tree shrew (Ullu and Tschudi 1984; Nishihara et al. 2002), and Kapitonov and Jurka (2003) recently identified one example of a 5S rRNA-derived SINE family, designated zebrafish SINE3, in the zebrafish genome. SINEs do not contain any protein-coding sequence, and each SINE RNA is transcribed by pol III and subsequently reverse transcribed and integrated into the host genome via recognition of the 3′-terminal sequence by a LINE-encoded protein. This model has been verified experimentally (Kajikawa and Okada 2002; Dewannieux et al. 2003) and is illustrated by the fact that many partner SINEs and LINEs have common 3′-terminal sequences (Ohshima et al. 1996; Okada et al. 1997; Ogiwara et al. 2002; Ohshima and Okada 2005).

Originally, SINE families were thought to be limited in their distribution, sometimes being specific to a few species or a single genus (Shedlock and Okada 2000). This concept is exemplified by the SmaI family, which is specific to chum and pink salmon (Kido et al. 1991), and the *Alu* family (Ullu and Tschudi 1984), which is specific to primates. Recently, however, two examples of a group of SINE families have been reported, in which members of each group share the same conserved sequence and are distributed among a wide range of species. The first such example is the superfamily of CORE-SINEs sharing 65 bp of "core" sequence in their central regions (Gilbert and Labuda 1999). Members of CORE-SINEs are distributed in various vertebrates including mammals (e.g., MIR and Mon-1) and several invertebrates. The second example is the superfamily of V-SINEs, members of which are found in various fishes, including lungfish and lamprey, and contain the central sequence specific to V-SINE members (Ogiwara et al. 2002). It is not yet known what the function of the central sequence in these two SINE superfamilies is. Gilbert and Labuda (1999, 2000) proposed that the central region of CORE-SINEs may contribute to retrotranspositional activity or long-term survival of the SINE family, and Ogiwara et al. (2002) proposed a possible function for SINEs in host viability based on high-sequence identity among V-SINEs.

Despite the abundance of SINEs in eukaryote genomes (e.g.,

[3]**Corresponding author.**
E-mail nokada@bio.titech.ac.jp; fax 81-45-924-5835.
Article published online before print. Article and publication date are at http://www.genome.org/cgi/doi/10.1101/gr.5255506. Freely available online through the *Genome Research* Open Access option.

16:864–874 ©2006 by Cold Spring Harbor Laboratory Press; ISSN 1088-9051/06; www.genome.org

constituting 14% and 8% of the human and mouse genomes, respectively), it is unclear whether they are of benefit to the host genomes. Transposable elements are usually regarded as genomic parasites, with their fixed, often inactivated copies considered to be "junk DNA." There are many reports of retroposon copies playing a role in the regulation of transcription or post-transcriptional events (Britten 1997; Brosius 1999b; Peaston et al. 2004); these cases, however, are exceptional, and such regulation is not thought to be an intrinsic function of retroposons. Adoption of a transposable element for a new function by the genome is called "exaptation," a term introduced by Gould and Vrba to describe a feature for which the current use is not derived from its original function through natural selection (Gould and Vrba 1982; Brosius and Gould 1992). Although a retroposon may be exapted in the host genome, the significance of the retroposon's acquired function in the genome is unclear. In other words, to evaluate the significance of an exapted retroposon, it is necessary to demonstrate that the retroposon sequence is under purifying selection, i.e., establish that it is phylogenetically conserved among the orthologs of distant species.

In this report, we describe a third SINE superfamily characterized by a novel shared central domain. It is currently represented by nine separate families in the genomes of mammals, birds, fish, sharks, hagfish, amphioxus, and sea urchin. Most interestingly, the SINE family characterized in mammals and chicken was active in amniotes (mammals, birds, and reptiles) during the Carboniferous period, at least 310 million years ago (Mya), and probably only could be discovered because multiple copies have been highly conserved. The observed conservation strongly indicates that the central domain of these transposable elements have been exapted, i.e., have become a functional component of the mammalian genomes.

Results and Discussion

Characterization of the DeuSINE superfamily

A schematic representation of nine SINE families from 10 phylogenetic groups is shown in Figure 1A. We first characterized a novel SINE family from coelacanth (*Latimeria menadoensis*), designated the LmeSINE1 family. We identified two LmeSINE1 subfamilies, LmeSINE1a and LmeSINE1b, distinguished by differences in the structure of their central domains. We detected other new SINE families using a GenBank FASTA search with the two LmeSINE consensus sequences as queries. All of the SINEs identified by this search had a highly similar sequence in the central region. The new SINE family found in the genomes of rainbow trout (*Oncorhynchus mykiss*), Chinook salmon (*Oncorhynchus tshawytscha*), brown trout (*Salmo trutta*), and Atlantic salmon (*Salmo salar*) was designated OS-SINE1 for its known distribution in two genera, *Oncorhynchus* and *Salmo*. The other four new SINE families were designated as SacSINE1 for dogfish shark (*Squalus acanthias*), EbuSINE1 and EbuSINE2 for hagfish (*Eptatretus burgeri*), and BflSINE1 for amphioxus (*Branchiostoma floridae*). EbuSINE1 and EbuSINE2 families can be distinguished by their 3′-tail sequences (Fig. 1A).

In addition to these newly characterized SINE families, we found two SINEs in the RepBase database of repetitive elements with similar central domains—the zebrafish SINE3 (Kapitonov and Jurka 2003) and sea urchin SINE2-3_SP (Kapitonov and Jurka 2005). We also found SINEs in catfish (*Ictalurus punctatus*), which we named SINE3_IP, because their consensus is 83% identical to the zebrafish SINE3 consensus.

Using the central domain shared by the SINE families described above to search for homologous SINEs, we identified a novel SINE family in the human and chicken genomes. RepeatMasker analysis, including a consensus sequence derived for these elements, detected over a 1000 copies with an average substitution level from the consensus of over 40%, in both the human and chicken genome. Copies that are more diverged than that tend to escape detection by RepeatMasker, so that it is likely that both the average substitution level and the copy number are underestimates. Given the high divergence and the fact that consensus sequences derived from copies in the human or chicken genome were basically the same, it appeared that the human and chicken repeats may be vestiges of a SINE family that already was propagating in a common ancestor of mammals and sauropsids (birds and reptiles), which lived over 310 Mya (Benton 1997). Indeed, by consulting the UCSC Genome Bioinformatics Database (Karolchik et al. 2003) we found several copies at orthologous sites in the human and chicken genomes; for example, the human copies at (May 2004 hg17 assembly) chr3:70,515,288–70,515,642 and chr15:51,569,670–51,569,822 are orthologous to the chicken copies at (Feb 2004 galGal2 assembly) chr12:15,590,146–15,590,334 and chr10:8,938,046–8,938,218, respectively. As expected, the new SINE family is also present in the genomes of all other mammals, including opossum and platypus, while we found one copy or related sequence in a tortoise (*Gopherus agassizii*) sequence (gi:57334898 pos. 2422–2897). For its Amniota-wide occurrence, we name this novel SINE family AmnSINE1. The discovery of the AmnSINE1 family confirms that many vestiges of ancient repetitive elements are still hidden in the human genome. Tracking such ancient records will improve our understanding of the evolution of the human genome.

All of the SINE families described above, except for the zebrafish SINE3 and sea urchin SINE2-3_SP families, were newly characterized in this study. These nine SINE families have highly similar sequences in their central domains (the average of pairwise identities calculated among the consensus sequence of the SINEs for the central domain is 73.4%), schematically shown as green boxes in Figure 1A, indicating that the domain might have a single evolutionary origin. By considering the phylogenetic distribution of the SINE families in the species, their origin may have been in a common ancestor of Deuterostomia, which includes vertebrates, amphioxus, sea urchins, and their relatives. We therefore name the novel SINE superfamily in Deuterostomia genomes "DeuSINEs" and the central domain of the novel SINE superfamily the "Deu-domain." The sequence of the Deu-domain is completely different from the central regions of CORE-SINEs and V-SINEs. Thus, in addition to CORE-SINEs and V-SINEs, DeuSINEs are the third superfamily in which several distinct SINE families share a common central domain.

Characterization of a chimeric structure of 5S rRNA and tRNA-derived DeuSINEs

In all DeuSINEs, the promoter domain is located in the 5′ region to facilitate pol III transcription. Among them, five new SINE families, LmeSINE1 (a and b), SacSINE1, EbuSINE1, EbuSINE2, and BflSINE1, are tRNA-derived SINEs with promoter regions similar to those of tRNA (Fig. 2A). In addition, the previously reported sea urchin SINE2-3_SP is thought to have been derived from $tRNA^{Lys}$ (Kapitonov and Jurka 2005).

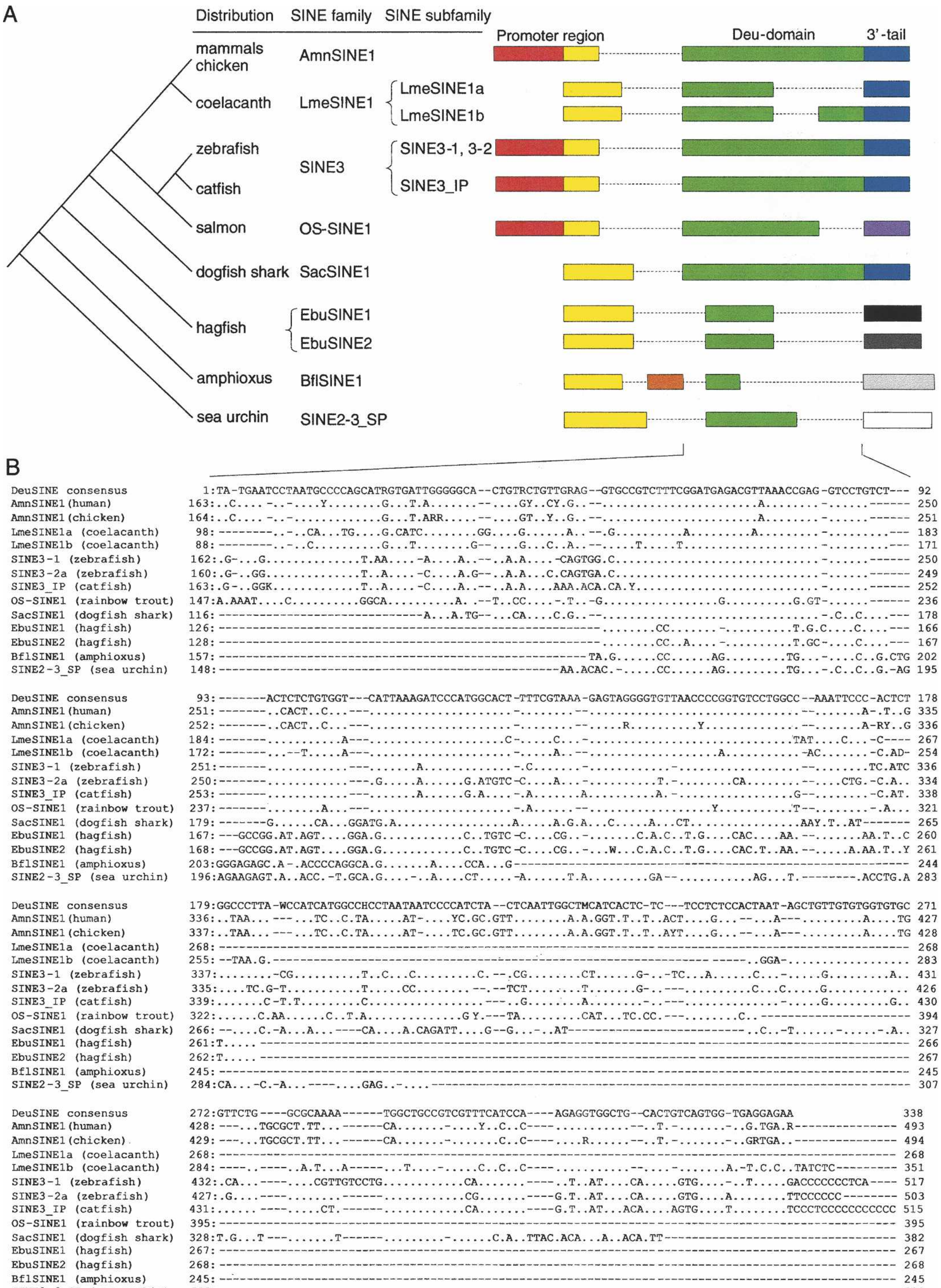


Figure 1. The structure and distribution of DeuSINEs. (*A*) Schematic representation of DeuSINEs and the known phylogenetic relationships between the host species. All of these SINEs were newly characterized in this study except for zebrafish SINE3 and sea urchin SINE2-3_SP. Green boxes represent common sequences among DeuSINEs, and an alignment of the region is shown in *B*. Yellow and red boxes denote promoter regions derived from tRNA (see Fig. 2A) and 5S rRNA (Fig. 2B), respectively. The regions denoted by yellow boxes in 5S rRNA-derived SINEs (AmnSINE1, SINE3, SINE3_IP and OS-SINE1) are similar to tRNA-derived promoter regions of LmeSINE1 (see Fig. 3 for details). Blue boxes represent 3′-tails of SINEs that are similar to that of zebrafish CR1-4_DR LINE (Fig. 4A), whereas the purple box is similar to that of the rainbow trout RSg-1 LINE (Fig. 4B). The 3′-tail sequences of EbuSINE1, EbuSINE2, BflSINE1, and SINE2-3_SP are distinct and of unknown origin. (*B*) Alignment of the common central DeuSINE sequences (Deu-domain; green boxes in *A*) in different host organisms. The SINE3-1 and the SINE3-2a sequences represent two subfamilies of the zebrafish SINE3 family. The *top* sequence is a consensus of the DeuSINE domain. Dots indicate nucleotides identical to those in the consensus sequence, and dashes indicate gaps inserted to improve the alignment.

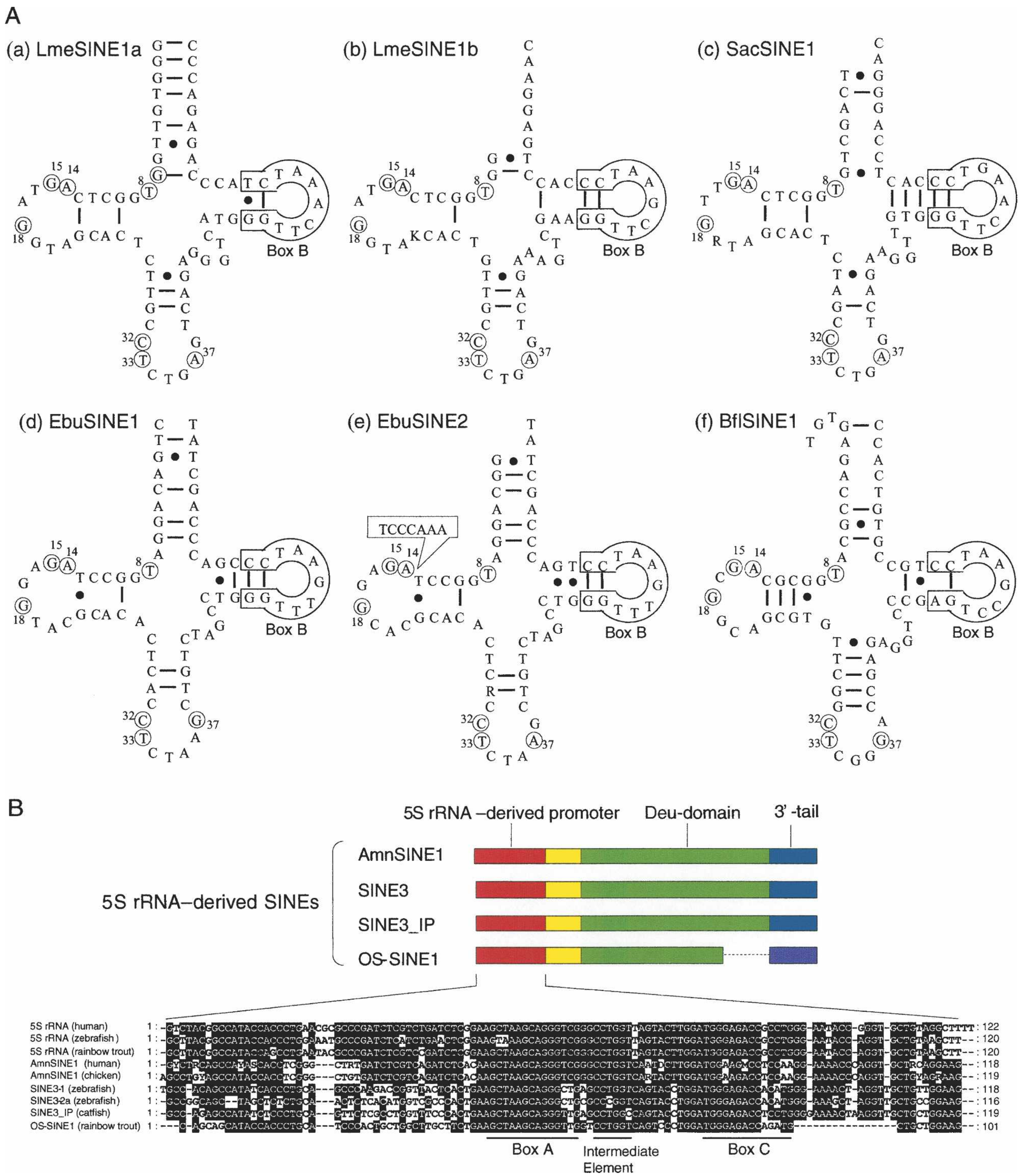


Figure 2. Characterization of promoter regions of DeuSINEs. (*A*) Six tRNA-like structures of the promoter regions of tRNA-derived DeuSINEs: (*a*) LmeSINE1a, (*b*) LmeSINE1b, (*c*) SacSINE1, (*d*) EbuSINE1, (*e*) EbuSINE2, and (*f*) BflSINE1. Standard base pairs and G-T wobble pairs are shown as black dashes and dots, respectively. Nucleotides conserved in functional tRNAs (8T, 14A, 15G, 18G, 32C, 33T, and 37R) are circled, and the Box B promoter sequences are boxed. The numbering system corresponds with that of general tRNA (Gauss et al. 1979). (*B*) An alignment of consensus sequences of the 5S rRNA-related regions (red boxes) of AmnSINE1, SINE3, SINE3_IP, and OS-SINE1. Zebrafish 5S rRNA gene sequences were obtained from Kapitonov and Jurka (2003). The 5S rRNA sequences of human and rainbow trout were obtained from GenBank (accession nos. X51545 and J01861, respectively). Box A, Box C, and Intermediate Element of pol III promoters are denoted with black lines. Nucleotides shaded in black are conserved across sequences.

As mentioned above, SINE3 is the only example of a 5S rRNA-derived SINE to date (Kapitonov and Jurka 2003). In the present study, we add the AmnSINE1, catfish SINE3_IP, and salmon OS-SINE1 to this category (5S rRNA-derived regions are indicated by red boxes in Figure 1). As shown in an alignment of sequences in Figure 2B, the internal promoters of the 5S rRNA gene that are denoted as Box A, Intermediate Element, and Box C, are well conserved. The sequence of human or chicken AmnSINE1 and its composite structure are very similar to that of zebrafish SINE3 (67% identity between the two consensus sequences). Therefore, it is possible that a common ancestor of zebrafish SINE3 and AmnSINE1 dates back to a common ancestor of vertebrates. If so, this SINE family has survived for over 450 Myr (Kumar and Hedges 1998), and there may be fossil sequences of AmnSINE1 in many other vertebrate species.

Surprisingly, a part of the 5S rRNA-derived SINEs is highly similar to a part of the tRNA-derived region of LmeSINEs (see yellow boxes in Fig. 3A). Figure 3A shows the alignment between LmeSINE1a and LmeSINE1b, which contain a tRNA-derived promoter, and AmnSINE1, SINE3, SINE3_IP, and OS-SINE1, which contain a 5S rRNA-derived promoter in addition to a sequence that resembles the tRNA-derived region. A high degree of similarity (e.g., 86.4% identity between this region in LmeSINE1a and AmnSINE1) suggests an evolutionary relationship between this region of the 5S rRNA-derived SINEs and tRNAs. Because the tRNA-derived region of 5S rRNA-derived SINEs lacks a Box B sequence, it probably cannot act as a promoter for pol III transcription. These observations suggest a mechanism for generating the 5S rRNA-derived promoter domain, as proposed in Figure 3B. In this scheme, a primordial tRNA-derived SINE recombined with 5S rRNA to generate the 5S rRNA-derived SINEs containing a tRNA-derived region. Subsequently, the second promoter region of the tRNA-derived region (the 3′ half of the yellow box in Fig. 3B) was lost from the SINE. Thus, strictly speaking, the 5S rRNA-derived SINE originally described in the zebrafish genome (Kapitonov and Jurka 2003) along with those in mammal, chicken, catfish, and salmon genomes described in the present study, represent a chimera of a 5S rRNA and a tRNA-derived SINE.

The molecular mechanism for the recombination between 5S rRNA and a tRNA-derived SINE generating a new SINE family is unknown. One possibility is that a tRNA-derived DeuSINE sequence was occasionally integrated near a 5S rRNA gene. Subsequently, the 5S rRNA served as the promoter for the transcription of the SINE, producing a new DeuSINE family. The other possibility is that this recombination may have resulted via template-switching from the tRNA-derived SINE RNA to 5S rRNA during the process of cDNA synthesis in retroposition (Brosius 1999a; Buzdin et al. 2002, 2003). We speculate that the latter possibility is more likely because the whole 5S rRNA (exactly, without any flanking regions) is included in the SINE and because many ex-

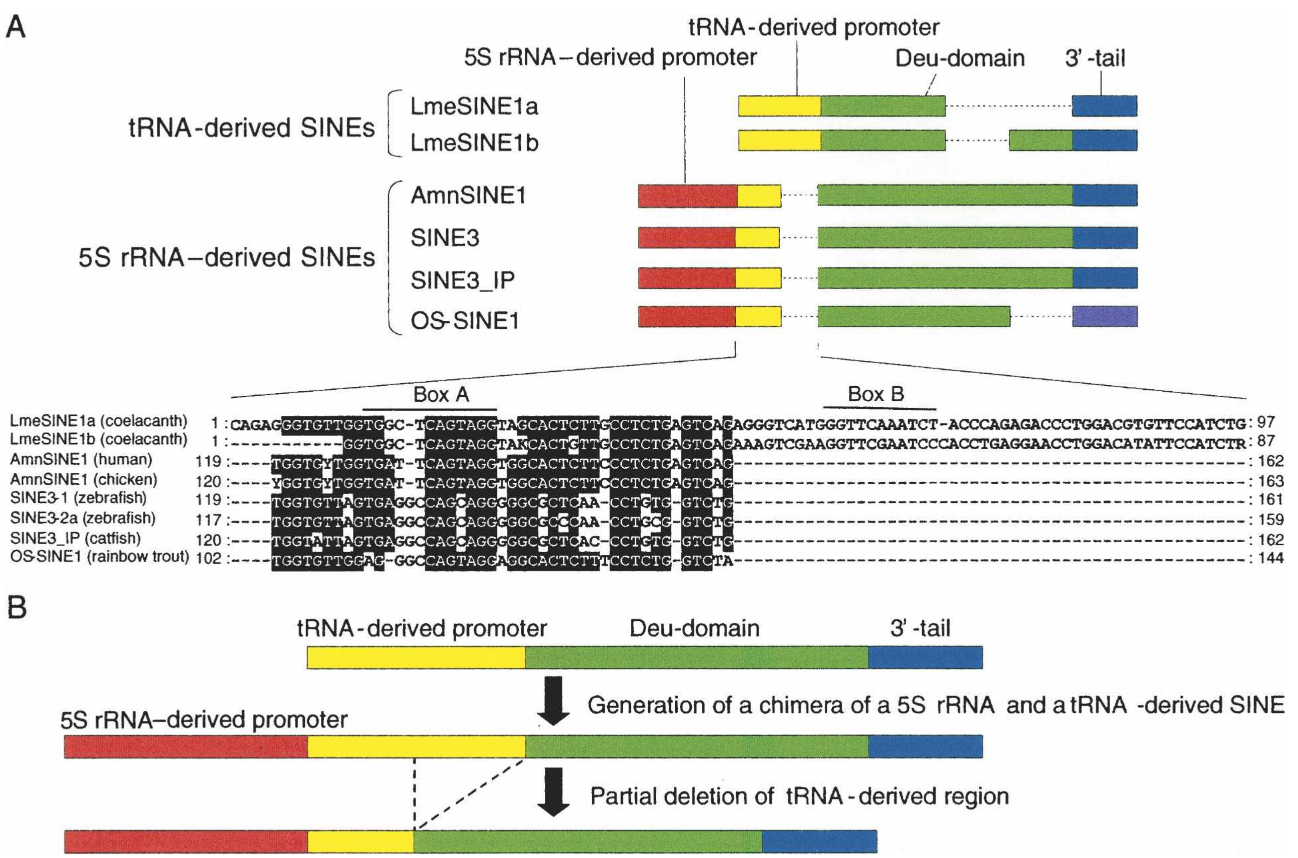


Figure 3. Chimeric structure of 5S rRNA and tRNA-derived DeuSINEs. (*A*) Comparison of the tRNA-derived regions of LmeSINE1a and LmeSINE1b with a part of the 5S rRNA-derived SINEs, AmnSINE1, SINE3, SINE3_IP, and OS-SINE1. Box A and Box B are the pol III promoter sequences of the two LmeSINEs. (*B*) A possible scheme for the structural evolution of the 5S rRNA-derived SINE families. The green boxes and the blue boxes denote the Deu-domain and 3′-tail, respectively. A 5S rRNA sequence (red boxes) became joined with a tRNA-derived SINE, with subsequent partial deletion of the original tRNA-derived promoter region (yellow boxes).

amples of template switching during retrotransposition are known in the human genome (Buzdin et al. 2002).

Variation of the 3′-tail domain in DeuSINEs

The 3′-tail sequence of zebrafish SINE3 is similar to that of CR1-4_DR in zebrafish, which belongs to the L2 LINE clade (Kapitonov and Jurka 2003; Kajikawa et al. 2005). Regarding the other DeuSINE sequences, we found that the 3′-tails of LmeSINE1, SacSINE1, and AmnSINE1 are also quite similar to that of CR1-4_DR (blue boxes in Figs. 1, 4A). We calculated sequence identities between the 3′-tail region of CR1-4_DR (position: 1822–1868) and each of the corresponding sequences of AmnSINE1, LmeSINE1a, SINE3-1, SINE3_IP, and SacSINE1, and their average value was 77%. Therefore, in addition to the zebrafish SINE3, retroposition of these three new SINE families probably depends on LINEs belonging to the L2 clade in each host genome. The eponymous mammalian LINE-2 (L2) family (Smit 1996) is thought to predate the bird-mammalian split 310 Mya (Benton 1997) as well, since ancient copies of elements identical or very similar to it and its associated SINE MIR are also found in the chicken genome (Hillier et al. 2004). However, the 3′ ends of L2 and MIR, conserved in mammals and birds, is unrelated to that of AmnSINE1 and CR1-4_DR. Also, it appears that the L2 element expanded in mammals up to the time of the eutherian radiation, leaving behind hundreds of thousands of recognizable L2 and MIR copies, while AmnSINE1 appears to have become retrotranspositionally inactive much earlier. Thus, AmnSINE1 may not have been retrotransposed by the mammalian L2 but by another, as yet uncharacterized LINE, which is an older CR1-4_DR–like LINE family belonging to the L2 clade. On the other hand, OS-SINE1 shares its 3′-tail sequence (purple boxes in Fig. 4B) with RSg-1 (77% identity), which is a LINE family in rainbow trout (Winkfein et al. 1988; Okada et al. 1997), indicating that RSg-1 is a partner LINE of OS-SINE1. The 3′ tails of the hagfish EbuSINE1 and EbuSINE2, amphioxus BflSINE1, and sea urchin SINE2-3_SP appear to be unrelated, and their origins are unknown (shown as black, dark gray, gray, and white boxes in Fig. 1A). Thus, in addition to having one of two different pol III promoters, these SINEs exhibit variations in the 3′-tail sequences, indicating that DeuSINEs have undergone frequent changes.

Although it is not known why DeuSINEs share a common central region (Deu-domain), this SINE superfamily's long survival (>600 Myr) suggests that the Deu-domain may serve a critical function. There are at least three possible reasons for the maintenance of the Deu-domain. One possibility is that the conserved Deu-domain imparts higher retroposition activity to the SINE, for example, by increasing the stability of either the SINE RNA or a complex of the RNA with a partner LINE protein. Alternatively, this domain may provide a high-frequency recombination opportunity among the SINEs. This would benefit DeuSINEs, because SINEs containing the 3′-tail sequence of more active partner LINEs may have higher retropositional activity. The Deu-domain has survived over all this time by recombining with different pol III promoters and active L2-clade 3′ ends. The third possibility is that the Deu-domain possesses an indepen-

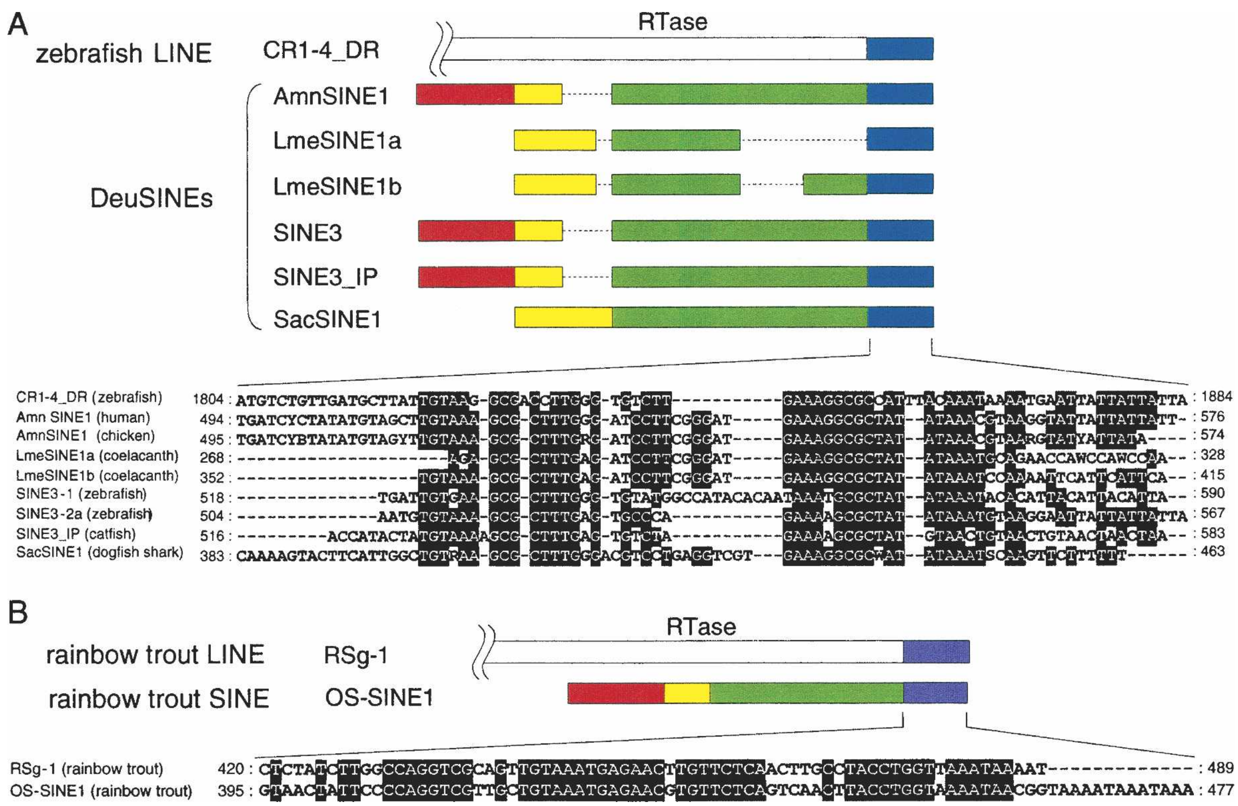


Figure 4. Alignment of the consensus 3′-tail sequences of DeuSINEs with that of the corresponding LINEs. "RTase" denotes the reverse-transcriptase encoded by LINEs. (*A*) The 3′-tail sequences of AmnSINE1, LmeSINE1a, LmeSINE1b, SINE3, SINE3_IP, and SacSINE1 (blue boxes) are similar to that of zebrafish CR1-4_DR LINEs. (*B*) The 3′-tail of OS-SINE1 (purple box) is similar to rainbow trout RSg-1 LINE. Both CR1-4_DR and RSg-1 LINE sequences were obtained from Repbase Update database (Jurka 2000).

dent function in the host genome, and that the presence of multiple copies may be advantageous for host survival. Further biological and genomic analyses will be necessary to confirm or refute these three possibilities.

Conservation of Deu-domain sequences among AmnSINE1 copies in the genomes

We calculated the copy number for each position along the AmnSINE1 consensus sequence using a FASTA search for AmnSINE1 in human and chicken genomes. We found that many AmnSINE1 copies in the genomes of human and chicken lack the promoter region and/or the 3′-tail sequence and only retain vestiges of the Deu-domain. As shown in Figure 5, it is clear that only the Deu-domain sequence, especially in the central 300–400 bp region, is conserved in many copies of both human and chicken AmnSINE1s. Because AmnSINE1 copies were active in a common ancestor of Amniota at least 310 Mya (Benton 1997), our result suggests that such AmnSINE1 sequences could represent full-length inserts, in which the promoter and 3′-tail regions have accumulated substitutions in a more or less neutral fashion to the extent that they have become unrecognizable.

Phylogenetically conserved AmnSINE1 sequences among mammalian species

Although many examples of retroposon exaptation have been reported, it is necessary to demonstrate their conservation among the orthologs of distant species to establish their significance in the genome. In this study, we compared the degree of sequence conservation for each AmnSINE1 locus among mammalian orthologs. For each AmnSINE1 sequence in the human genome, conservation scores per base pair were obtained for the SINE and both its 3′- and 5′-flanking regions (1.5 kbp total) from the UCSC Genome Bioinformatics Database (Karolchik et al. 2003). We found 105 AmnSINE1 copies that are phylogenetically conserved among mammalian orthologs (human, chimpanzee, mouse, rat, and dog). Moreover, we confirmed in the UCSC Genome Bioinformatics Database that almost all loci are also unusually conserved in the opossum genome (e.g., see Fig. 6A). Figure 6B shows conservation graphs for 1.5 kbp sequences around and including 10 representative AmnSINE1s (see also Supplemental Fig. 2 for all 105 conserved loci). Such high conservation provides strong evidence that these AmnSINE1 sequences have been under purifying selection and have a significant function that contributes to host viability. In most cases, flanking regions of those AmnSINE1 loci are also highly conserved, the extent of conserved region being variable from locus to locus, suggesting a variety of involvement of AmnSINE1 in different functions in the host.

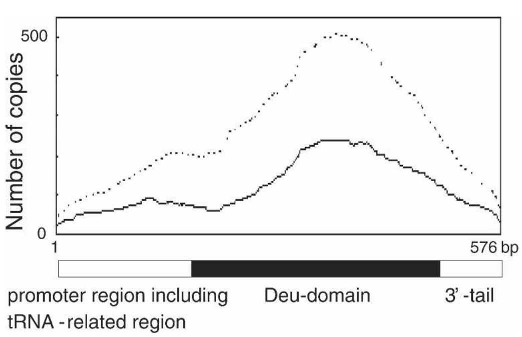
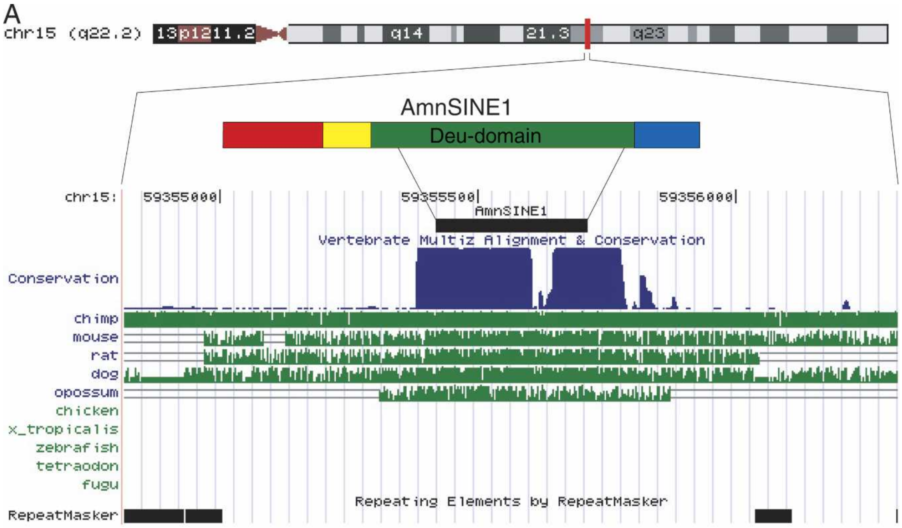


Figure 5. Conservation of Deu-domain sequences (black box) among AmnSINE1 copies in human and chicken genomes. This graph shows the number of copies that include each nucleotide position of the AmnSINE1 in human (the bold line) and chicken (the dotted line). The number of copies of AmnSINE1 analyzed is 380 and 742 for human and chicken, respectively.

Function of AmnSINE1 in mammalian genomes

Copies of SINEs and other transposable elements are generally thought of as "junk DNA." Just as junk sometimes turns out to be useful, many instances of exapted transposable elements have come to light (Britten 1997; Brosius 1999b; Smit 1999; Peaston et al. 2004). These usually appear to be chance by-products of insertion and subsequent substitution events. In contrast, the consistent localization of the conservation to the AmnSINE1 Deu-domain and the frequency of independent exaptations may suggest that the functionality of the exapted copies was inherent in the transposable element sequence. Like the Deu-domain in AmnSINE1 (Fig. 5), the CORE sequence of the mammalian MIR element, the founding member of the CORE-SINE class, appears to be more conserved than the promoter and tail regions (Smit and Riggs 1995). However, despite a recent report (Silva et al. 2003), it is still unclear whether MIR copies or MIR core sequences are unusually conserved among mammalian species. Such analysis is complicated by the sheer number of recognizable MIR copies in mammalian genomes. In the present study, we describe the first SINE that has been repeatedly exapted by the host, as indicated by the extraordinary conservation of at least 105 copies between mammalian genomes (Fig. 6; Supplemental Fig. 2). Moreover, the conservation appears localized in the Deu-domain region and the recognizable AmnSINE1 copies contain the Deu-domain (Fig. 5) more frequently than the promoter or 3′-tail regions. Thus, it is likely that the functionality of exapted AmnSINE1 sequences in the human genome primarily is provided by the Deu-domain.

Ultimately, it will be necessary to identify the detailed function(s) of the mammalian AmnSINE1s. The genomic function of the exapted AmnSINE1 sequences may differ from case to case. The SINEs may operate as parts of protein-coding sequences, UTRs, promoters, or micro-RNA (Ferrigno et al. 2001; Bejerano et al. 2004b; Smalheiser and Torvik 2005). To examine the possibility that an AmnSINE1 may function as an mRNA, we searched all known exon sequence data for human and mouse using FASTA. We found no sequence similarity except for one example in which a part of the 5′ UTR of mature T cell proliferation 1 (MTCP1; accession no. NM_014221) had 62% identity with the AmnSINE1 consensus sequence. Furthermore, we searched the possible expression of the 105 AmnSINE1s shown in Supplemental Figure 2 against the human spliced EST database and found positive results for three loci (chr8:26,253,307-26,253,397, chr17:19,129,522-19,129,682, and chr12:20,405,782-20,405,989). They code for mRNA of Protein phosphatase 2 regulatory subunit B α isoform, Epsin 2, and cGMP-inhibited 3′,5′-cyclic phosphodiesterase A. Therefore, these AmnSINE1 sequences may function as parts of each mRNA. On the other hand, since some AmnSINE1 are present in introns, it is possible that they are transcribed as pre-mRNA and contribute to the processing process of mRNA such as mRNA splicing. We also examined the possibility that AmnSINE1 functions as a promoter or micro-RNA using the UCSC Genome Bioinformatics and Rfam



B

AmnSINE1 position in human				1.5 kbp including AmnSINE1		
Chr	Start	End	Length	Start	End	Conservation
1	92,167,433	92,167,542	110	92,166,738	92,168,237	
1	102,748,257	102,748,495	239	102,747,626	102,749,125	
1	118,588,292	118,588,566	275	118,587,679	118,589,178	
1	120,372,135	120,372,325	191	120,371,480	120,372,979	
1	142,648,490	142,648,680	191	142,647,835	142,649,334	
1	163,632,135	163,632,326	192	163,631,481	163,632,980	
1	173,788,405	173,788,841	437	173,787,873	173,789,372	
1	206,496,662	206,496,847	186	206,496,005	206,497,504	
2	21,726,650	21,726,771	122	21,725,961	21,727,460	
2	29,495,737	29,495,970	234	29,495,104	29,496,603	

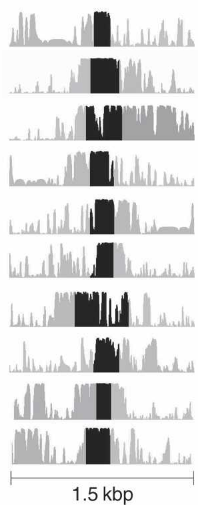


Figure 6. Evidence for purifying selection on AmnSINE1 in mammals. (*A*) An example of conserved AmnSINE1 locus (the window is chr15:59,354,815-59,356,314 in which the AmnSINE1 position is located at chr15:59,355,417-59,355,712) obtained from UCSC Genome Bioinformatics Web site. Note that the AmnSINE1 sequence is conserved in all mammals including opossum. (*B*) The location and conservation of 10 representative AmnSINE1 loci in human. The position information in human was determined from genomic sequence data in UCSC Genome Bioinformatics (ver. hg17). (Chr) Chromosome number. PhastCons conservation scores of the 1.5-kbp region around and including each AmnSINE1 were obtained by comparing human, chimpanzee, mouse, rat, and dog sequences, and the graphs are displayed for each locus. In each graph, the black region denotes the AmnSINE1 sequence and the gray represents the flanking region corresponding to the given position. Detailed information for the 10 loci is available in Supplemental Figure 3.

(Griffiths-Jones et al. 2003; http://www.sanger.ac.uk/Software/Rfam/) databases, respectively, but found no positive correlation. It is also possible that these SINEs may instead serve as enhancers or silencers. Future work will elucidate the contribution of each AmnSINE1 sequence in the human genome.

Although it is likely that the conserved AmnSINE1 copies function in distinct ways, it is also possible that the sequences share a common function, given the high-copy number of AmnSINEs under selection in the mammalian genome. If this is indeed the case, they may function in general cellular processes such as DNA replication or genomic organization (i.e., chromatin structure) (Parada et al. 2004). For example, dispersed yeast tRNA genes cluster at the nucleolus (Thompson et al. 2003) and have been suggested to function in gene silencing (Wang et al. 2005). Thus, additional work is required to resolve the detailed function of the AmnSINE1 sequences not only to understand how they have contributed to evolution of mammalian genomes but also to determine the significance of repetitive elements in the genomes. Furthermore, it is also important to clarify the significance of the Deu-domain shared among the nine DeuSINE families.

The number of retroposons characterized from various eukaryotes continues to increase, and thus retroposons other than DeuSINEs may also serve critical functions at multiple loci in host genomes. Bejerano et al. (2006) recently identified another ancient SINE, which is unrelated to the SINEs in our study, and found that the SINE copies have been exapted at multiple locations in mammalian genomes. Coincidently, they found the closest related modern SINE in the coelacanth as well, thereby reinforcing its status as a living fossil.

Indeed, mammalian genomes contain many dispersed repetitive regions that are highly conserved among distantly related species. Bejerano et al. (2004b) identified ~5000 such regions, and we found that at least one of the categories is actually the AmnSINE1 family (see http://www.soe.ucsc.edu/~jill/dark.html, cluster# 206). Our results provide convincing evidence that conserved noncoding sequences are associated with ancient retroposons. Thus, this study provides the basis for further work directed at assigning a concrete functional role for retroposons in cellular processes. Because all mammals retain highly conserved and exapted AmnSINE1 in many loci of their genomes in common, it is possible to speculate that such extensive exaptation occurred in a common ancestor of mammals and may contribute to innovation of a body plan specific to mammals, such as testicles and mammary glands. Furthermore, since a few AmnSINE1 loci are conserved between mammals and chicken, it is also possible that they contributed to the generation of Amniota-specific morphology. It should be noted that Bejerano et al. (2006) actually demonstrated that one copy of SINEs they characterized (LF-SINE) and conserved among Tetrapoda (mammals, sauropsids, and amphibians) contributes to neural development in mouse. Therefore, it is possible that each of waves of extensive exaptation of SINEs which had occurred in a common ancestor of mammals, amniotes, tetrapods, or vertebrates contributed to the innovation of the new body plan specific to each clade. Experiments on concrete functions for each locus of the exapted AmnSINE1 will shed light on the extent of contribution of this transposable element to morphological innovation of mammals as well as birds and reptiles.

Methods

Application of a computational algorithm to find novel SINEs in coelacanth sequence data

We first obtained 797 kb of sequence data for coelacanth (*Latimeria menadoensis*) from GenBank (AC150283, AC150284, AC150308, AC150309, AC150310, and AC151571) and used a computational algorithm that detects SINE-like sequences from sequence data. This algorithm collects multiple similar sequences, each of which contain a Box B-like sequence. First, the 797-kb sequence of coelacanth DNA was searched for both the top and bottom strands of consensus sequences of the Box B RNA polymerase III promoter (10 nucleotides; GWTYRANNCY) (see Fig. 7). Next, the 5′ (100 bp) and the 3′ (500 bp) flanking sequences of each Box B-like sequence were extracted to obtain 610-bp sequences. A local BLAST search (Altschul et al. 1990) was performed using each extracted 610-bp sequence as the query to search for homology within all of the extracted sequences. We set the E-value standard to 10^{-50}, i.e., we considered the query and the subject sequences to be similar if the E-value was $<10^{-50}$. As shown in Figure 7, an arrow from Sequence1 (Seq1) to Sequence5 (Seq5) indicates that Sequence5 is a hit with an accompanying E-value $<10^{-50}$ by BLAST searching when Sequence1 is used as a query sequence, and vice versa. Next, we divided the sequences that were similar to one another into groups. From the 797 kbp of coelacanth sequences, we obtained 15 groups, each of which included at least two sequences. We aligned the sequences in each group using ClustalX (Thompson et al. 1997) and GENETYX version 6.0 software (GENETYX Co., Ltd.). Among the 15 groups, one group containing seven sequences was ultimately identified as having a novel tRNA-derived SINE sequence. We then classified this group into two subfamilies according to structure and designated them LmeSINE1a and LmeSINE1b (see Fig. 1). Using an additional BLAST homology search of the 797 of kilobase coelacanth sequences, we found a total of 10 and 16 copies of LmeSINE1a and LmeSINE1b, respectively (Supplemental Fig. 1C,D).

Characterization of other members of DeuSINEs

We carried out a FASTA homology search (Pearson and Lipman 1988) through GenBank via DDBJ (http://www.ddbj.nig.ac.jp/) using consensus sequences for LmeSINE1a and LmeSINE1b as queries. We identified similar sequences in various species, namely, zebrafish (*Danio rerio*), catfish (*Ictalurus punctatus*), salmons (*Oncorhynchus mykiss, Oncorhynchus tshawytscha, Salmo trutta*, and *Salmo salar*), dogfish shark (*Squalus acanthias*), hagfish (*Eptatretus burgeri*), amphioxus (*Branchiostoma floridae*), and sea urchin (*Strongylocentrotus purpuratus*). Furthermore, we identified other SINE sequences from human and chicken genomes and designated them AmnSINE1. We aligned the sequences from each organism to construct consensus sequences using GENETYX version 6.0 (Supplemental Fig. 1). The AmnSINE1 copy numbers were investigated using the RepeatMasker program to search human whole-genome data. We used MEGA3 (Kumar et al. 2004) or GENETYX to calculate sequence identities among SINE consensus sequences.

The Deu-domain of AmnSINE1 is more conserved than the promoter and 3′-tail regions

To examine the degree of conservation per site in the AmnSINE1 sequence, we first performed a FASTA search for AmnSINE1 sequences in human and chicken genomes. From the collected sequences, we removed those that were similar only to the 5′ region (1–120 bp) of the consensus sequence to eliminate sequences that represented the 5S rRNA gene or its pseudogenes rather than AmnSINE1. We ultimately obtained 380 and 742 AmnSINE1 copies from the human and chicken genomes, respectively. Among each of the 380 and 742 sequences, we counted the number of copies that included each position in the alignment along the 576 bp and graphed them as shown in Figure 5.

Conservation of AmnSINE1 sequences among mammals

To compare sequence conservation of and around AmnSINE1 among mammals, we used phastCons scores (Siepel et al. 2005) around each AmnSINE1 from the UCSC Genome Bioinformatics Database (Karolchik et al. 2003) (http://hgdownload.cse.ucsc.edu/goldenPath/hg17/phastCons/mzPt1Mm5Rn3Cf1Gg2Fr1Dr1/). The phastCons scores show the level of conservation at each nucleotide among orthologs of human, chimpanzee, mouse, rat, and dog. We obtained the scores for the 1.5 kbp of each AmnSINE1 and its flanking region, which included 750 bp from the middle of the SINE sequence toward both the 3′- and 5′-ends. We then graphed the scores for the 1.5-kbp sequences for each of the 105 loci (see Supplemental Fig. 2); 10 representative graphs are shown in Figure 6B.

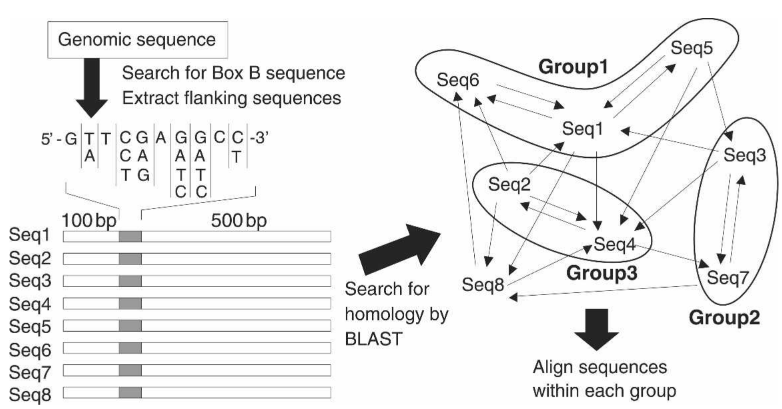


Figure 7. A schematic representation of the method used in this study to find novel SINEs from genomic sequences of the coelacanth. This algorithm consists of the following five steps: (1) detection of the Box B-like sequence; (2) extraction of their flanking sequences (exemplified as Seq1–8); (3) BLAST search for homology among one another to find similar sequences; (4) collection of sequences that are recognized as similar to each other (E-value $<10^{-50}$); (5) alignment of the sequences within each group.

Acknowledgments

We thank three anonymous reviewers for critical and helpful comments on earlier drafts. This work was supported by research grants from the Ministry of Education, Culture, Sports, Science and Technology of Japan (to N.O.).

References

Altschul, S.F., Gish, W., Miller, W., Myers, E.W., and Lipman, D.J. 1990. Basic local alignment search tool. *J. Mol. Biol.* **215:** 403–410.

Bejerano, G., Pheasant, M., Makunin, I., Stephen, S., Kent, W.J., Mattick, J.S., and Haussler, D. 2004a. Ultraconserved elements in the human genome. *Science* **304:** 1321–1325.

Bejerano, G., Haussler, D., and Blanchette, M. 2004b. Into the heart of darkness: Large-scale clustering of human non-coding DNA. *Bioinformatics* (Suppl 1) **20:** I40–I48.

Bejerano, G., Lowe, C., Ahituv, N., King, B., Siepel, A., Salama, S., Rubin, E.M., Kent, W.J., and Haussler, D. 2006. A distal enhancer and ultraconserved exon are derived from a novel retroposon. *Nature* **441:** 87–90.

Benton, M.J. 1997. *Vertebrate paleontology*. Chapman & Hall, New York.

Britten, R.J. 1997. Mobile elements inserted in the distant past have taken on important functions. *Gene* **205:** 177–182.

Brosius, J. 1991. Retroposons–seeds of evolution. *Science* **251:** 753.

———. 1999a. Genomes were forged by massive bombardments with retroelements and retrosequences. *Genetica* **107:** 209–238.

———. 1999b. RNAs from all categories generate retrosequences that may be exapted as novel genes or regulatory elements. *Gene* **238:** 115–134.

Brosius, J. and Gould, S.J. 1992. On "genomenclature": A comprehensive (and respectful) taxonomy for pseudogenes and other "junk DNA". *Proc. Natl. Acad. Sci.* **89:** 10706–10710.

Buzdin, A., Ustyugova, S., Gogvadze, E., Vinogradova, T., Lebedev, Y., and Sverdlov, E. 2002. A new family of chimeric retrotranscripts formed by a full copy of U6 small nuclear RNA fused to the 3′ terminus of l1. *Genomics* **80:** 402–406.

Buzdin, A., Gogvadze, E., Kovalskaya, E., Volchkov, P., Ustyugova, S., Illarionova, A., Fushan, A., Vinogradova, T., and Sverdlov, E. 2003. The human genome contains many types of chimeric retrogenes generated through in vivo RNA recombination. *Nucleic Acids Res.* **31:** 4385–4390.

Dermitzakis, E.T., Reymond, A., Lyle, R., Scamuffa, N., Ucla, C., Deutsch, S., Stevenson, B.J., Flegel, V., Bucher, P., Jongeneel, C.V., et al. 2002. Numerous potentially functional but non-genic conserved sequences on human chromosome 21. *Nature* **420:** 578–582.

Dewannieux, M., Esnault, C., and Heidmann, T. 2003. LINE-mediated retrotransposition of marked *Alu* sequences. *Nat. Genet.* **35:** 41–48.

Ferrigno, O., Virolle, T., Djabari, Z., Ortonne, J.P., White, R.J., and Aberdam, D. 2001. Transposable B2 SINE elements can provide mobile RNA polymerase II promoters. *Nat. Genet.* **28:** 77–81.

Gauss, D.H., Gruter, F., and Sprinzl, M. 1979. Compilation of tRNA sequences. *Nucleic Acids Res.* **6:** r1–r19.

Gilbert, N. and Labuda, D. 1999. CORE-SINEs: Eukaryotic short interspersed retroposing elements with common sequence motifs. *Proc. Natl. Acad. Sci.* **96:** 2869–2874.

———. 2000. Evolutionary inventions and continuity of CORE-SINEs in mammals. *J. Mol. Biol.* **298:** 365–377.

Gould, S.J. and Vrba, E.S. 1982. Exaptation; a missing term in the science of form. *Paleobiology* **8:** 4–15.

Griffiths-Jones, S., Bateman, A., Marshall, M., Khanna, A., and Eddy, S.R. 2003. Rfam: An RNA family database. *Nucleic Acids Res.* **31:** 439–441.

Hillier, L.W., Miller, W., Birney, E., Warren, W., Hardison, R.C., Ponting, C.P., Bork, P., Burt, D.W., Groenen, M.A., Delany, M.E., et al. 2004. Sequence and comparative analysis of the chicken genome provide unique perspectives on vertebrate evolution. *Nature* **432:** 695–716.

Jurka, J. 2000. Repbase update: A database and an electronic journal of repetitive elements. *Trends Genet.* **16:** 418–420.

Kajikawa, M. and Okada, N. 2002. LINEs mobilize SINEs in the eel through a shared 3′ sequence. *Cell* **111:** 433–444.

Kajikawa, M., Ichiyanagi, K., Tanaka, N., and Okada, N. 2005. Isolation and characterization of active LINE and SINEs from the eel. *Mol. Biol. Evol.* **22:** 673–682.

Kapitonov, V.V. and Jurka, J. 2003. A novel class of SINE elements derived from 5S rRNA. *Mol. Biol. Evol.* **20:** 694–702.

———. 2005. SINE2-3_SP, a family of SINE2 retrotransposons in the sea urchin genome. *Repbase Reports* **5:** 97.

Karolchik, D., Baertsch, R., Diekhans, M., Furey, T.S., Hinrichs, A., Lu, Y.T., Roskin, K.M., Schwartz, M., Sugnet, C.W., Thomas, D.J., et al. 2003. The UCSC Genome Browser Database. *Nucleic Acids Res.* **31:** 51–54.

Kido, Y., Aono, M., Yamaki, T., Matsumoto, K., Murata, S., Saneyoshi, M., and Okada, N. 1991. Shaping and reshaping of salmonid genomes by amplification of tRNA-derived retroposons during evolution. *Proc. Natl. Acad. Sci.* **88:** 2326–2330.

Kumar, S. and Hedges, S.B. 1998. A molecular timescale for vertebrate evolution. *Nature* **392:** 917–920.

Kumar, S., Tamura, K., and Nei, M. 2004. MEGA3: Integrated software for molecular evolutionary genetics analysis and sequence alignment. *Brief. Bioinform.* **5:** 150–163.

Lander, E.S., Linton, L.M., Birren, B., Nusbaum, C., Zody, M.C., Baldwin, J., Devon, K., Dewar, K., Doyle, M., FitzHugh, W., et al. 2001. Initial sequencing and analysis of the human genome. *Nature* **409:** 860–921.

Nishihara, H., Terai, Y., and Okada, N. 2002. Characterization of novel *Alu*- and tRNA-related SINEs from the tree shrew and evolutionary implications of their origins. *Mol. Biol. Evol.* **19:** 1964–1972.

Ogiwara, I., Miya, M., Ohshima, K., and Okada, N. 2002. V-SINEs: A new superfamily of vertebrate SINEs that are widespread in vertebrate genomes and retain a strongly conserved segment within each repetitive unit. *Genome Res.* **12:** 316–324.

Ohshima, K. and Okada, N. 2005. SINEs and LINEs: Symbionts of eukaryotic genomes with a common tail. *Cytogenet. Genome Res.* **110:** 475–490.

Ohshima, K., Hamada, M., Terai, Y., and Okada, N. 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. 1991a. SINEs. *Curr. Opin. Genet. Dev.* **1:** 498–504.

———. 1991b. SINEs: Short interspersed repeated elements of the eucaryotic genome. *Trends Ecol. Evol.* **6:** 358–361.

Okada, N., Hamada, M., Ogiwara, I., and Ohshima, K. 1997. SINEs and LINEs share common 3′ sequences: A review. *Gene* **205:** 229–243.

Parada, L.A., Sotiriou, S., and Misteli, T. 2004. Spatial genome organization. *Exp. Cell Res.* **296:** 64–70.

Pearson, W.R. and Lipman, D.J. 1988. Improved tools for biological sequence comparison. *Proc. Natl. Acad. Sci.* **85:** 2444–2448.

Peaston, A.E., Evsikov, A.V., Graber, J.H., de Vries, W.N., Holbrook, A.E., Solter, D., and Knowles, B.B. 2004. Retrotransposons regulate host genes in mouse oocytes and preimplantation embryos. *Dev. Cell* **7:** 597–606.

Rogers, J.H. 1985. The origin and evolution of retroposons. *Int. Rev. Cytol.* **93:** 187–279.

Shedlock, A.M. and Okada, N. 2000. SINE insertions: Powerful tools for molecular systematics. *Bioessays* **22:** 148–160.

Siepel, A., Bejerano, G., Pedersen, J.S., Hinrichs, A.S., Hou, M., Rosenbloom, K., Clawson, H., Spieth, J., Hillier, L.W., Richards, S., et al. 2005. Evolutionarily conserved elements in vertebrate, insect, worm, and yeast genomes. *Genome Res.* **15:** 1034–1050.

Silva, J.C., Shabalina, S.A., Harris, D.G., Spouge, J.L., and Kondrashovi, A.S. 2003. Conserved fragments of transposable elements in intergenic regions: Evidence for widespread recruitment of MIR- and L2-derived sequences within the mouse and human genomes. *Genet. Res.* **82:** 1–18.

Smalheiser, N.R. and Torvik, V.I. 2005. Mammalian microRNAs derived from genomic repeats. *Trends Genet.* **21:** 322–326.

Smit, A.F. 1996. The origin of interspersed repeats in the human genome. *Curr. Opin. Genet. Dev.* **6:** 743–748.

———. 1999. Interspersed repeats and other mementos of transposable elements in mammalian genomes. *Curr. Opin. Genet. Dev.* **9:** 657–663.

Smit, A.F. and Riggs, A.D. 1995. MIRs are classic, tRNA-derived SINEs that amplified before the mammalian radiation. *Nucleic Acids Res.* **23:** 98–102.

Thompson, J.D., Gibson, T.J., Plewniak, F., Jeanmougin, F., and Higgins, D.G. 1997. The CLUSTAL_X windows interface: Flexible strategies for multiple sequence alignment aided by quality analysis tools. *Nucleic Acids Res.* **25:** 4876–4882.

Thompson, M., Haeusler, R.A., Good, P.D., and Engelke, D.R. 2003. Nucleolar clustering of dispersed tRNA genes. *Science* **302:** 1399–1401.

Ullu, E. and Tschudi, C. 1984. *Alu* sequences are processed 7SL RNA genes. *Nature* **312:** 171–172.

Wang, L., Haeusler, R.A., Good, P.D., Thompson, M., Nagar, S., and

Engelke, D.R. 2005. Silencing near tRNA genes requires nucleolar localization. *J. Biol. Chem.* **280:** 8637–8639.
Waterston, R.H., Lindblad-Toh, K., Birney, E., Rogers, J., Abril, J.F., Agarwal, P., Agarwala, R., Ainscough, R., Alexandersson, M., An, P., et al. 2002. Initial sequencing and comparative analysis of the mouse genome. *Nature* **420:** 520–562.
Weiner, A.M., Deininger, P.L., and Efstratiadis, A. 1986. Nonviral retroposons: Genes, pseudogenes, and transposable elements generated by the reverse flow of genetic information. *Annu. Rev. Biochem.* **55:** 631–661.
Winkfein, R.J., Moir, R.D., Krawetz, S.A., Blanco, J., States, J.C., and Dixon, G.H. 1988. A new family of repetitive, retroposon-like sequences in the genome of the rainbow trout. *Eur. J. Biochem.* **176:** 255–264.

Received October 11, 2005; accepted in revised form April 18, 2006.