Mobile Element Evolution Playing Jigsaw—SINEs in Gastropod and Bivalve Mollusks

Irina Matetovici[1,2], Szilard Sajgo[1,3], Bianca Ianc[1], Cornelia Ochis[1], Paul Bulzu[1], Octavian Popescu[4], and Annette Damert[1,*]

[1]Institute for Interdisciplinary Research in Bio-Nano-Sciences, Molecular Biology Center, Babes-Bolyai-University, Cluj-Napoca, Romania

[2]Present address: Institute of Tropical Medicine, Unit of Veterinary Protozoology, Antwerpen, Belgium

[3]Present address: Danish Research Institute of Translational Neuroscience, Nordic EMBL Partnership for Molecular Medicine, DANDRITE, Aarhus University, Aarhus, Denmark

[4]Institute of Biology, Romanian Academy, Bucharest, Romania

*Corresponding author: E-mail: annette.damert@gmx.de.

Accepted: December 13, 2015

Abstract

SINEs (Short INterspersed Elements) are widely distributed among eukaryotes. Some SINE families are organized in superfamilies characterized by a shared central domain. These central domains are conserved across species, classes, and even phyla. Here we report the identification of two novel such superfamilies in the genomes of gastropod and bivalve mollusks. The central conserved domain of the first superfamily is present in SINEs in Caenogastropoda and Vetigastropoda as well as in all four subclasses of Bivalvia. We designated the domain MESC (Romanian for MElc—snail and SCoica—mussel) because it appears to be restricted to snails and mussels. The second superfamily is restricted to Caenogastropoda. Its central conserved domain—Snail—is related to the Nin-DC domain. Furthermore, we provide evidence that a 40-bp subdomain of the SINE V-domain is conserved in SINEs in mollusks and arthropods. It is predicted to form a stable stem-loop structure that is preserved in the context of the overall SINE RNA secondary structure in invertebrates. Our analysis also recovered short retrotransposons with a Long INterspersed Element (LINE)-derived 5′ end. These share the body and/or the tail with transfer RNA (tRNA)-derived SINEs within and across species. Finally, we identified CORE SINEs in gastropods and bivalves—extending the distribution range of this superfamily.

Key words: SINE, retrotransposon, Mollusca.

Introduction

Mollusks are a megadiverse phylum. It is subdivided into eight classes, among them are Gastropoda (snails and slugs), Bivalvia (mussels, clams, oysters, etc.), and Cephalopoda (squid, octopus, cuttlefish, nautilus). Mollusks are second in the number of species only to arthropods. Yet, compared with the much less numerous vertebrates, the structure and evolution of mollusk genomes are only poorly understood. Only five complete genomes have been published: Bubble cone (Hu et al. 2011), oyster (Zhang et al. 2012), pearl oyster (Takeuchi et al. 2012), owl limpet (Simakov et al. 2013), and California two-spot octopus (Albertin et al. 2015), and for three more species (pygmy squid, nautilus, and Japanese scallop) partial genome sequences have been reported (Yoshida et al. 2011). The fraction of TE (transposable element)-derived sequences in the mollusk genomes is comparatively smaller than in vertebrate genomes (35–52%) and varies between 2% and 8% (Yoshida et al. 2011; Takeuchi et al. 2012; Zhang et al. 2012; Simakov et al. 2013; Albertin et al. 2015).

SINEs (Short INterspersed Elements) have been characterized in a variety of eukaryotic genomes (for reviews on SINEs, see Okada 1991; Kramerov and Vassetzky 2011). They are nonautonomous non-LTR (long terminal repeat) retrotransposons. A typical SINE is between 80- and 500-bp long. SINEs are derived from small RNAs: tRNA (Ohshima and Okada 1994) and 5S ribosomal RNA (rRNA) (Kapitonov and Jurka 2003). Primate *Alu* and rodent B1 SINEs represent processed 7SL RNA genes (Ullu and Tschudi 1984). Recently, SINEs derived from U1/U2 small nuclear RNA (snRNA) (*SINEU*; Kojima 2015) and from the 3′ end of 28S rRNA

(SINE28; Longo et al. 2015) have been characterized. In its simplest form a SINE consists of a single tRNA-derived module. Such SINEs were first discovered in *Galago crassicaudatus* (Daniels and Deininger 1991). Complex SINEs are either dimers/trimers of tRNA subunits (Piskurek et al. 2003; Schmitz and Zischler 2003; Churakov et al. 2005) or are composed of a tRNA- or 5S rRNA-derived head, a tRNA-unrelated body and a tail. SINE bodies in some cases contain central conserved domains. The origin of additional sequences present in SINE bodies is mostly unknown. SINE 3′ tails are in some cases shared with the LINE autonomous partner that provides the retrotransposition machinery necessary for their mobilization (Okada et al. 1997; Kajikawa and Okada 2002; Ohshima and Okada 2005). SINEs often terminate in either homogeneous or more complex A-rich segments (Borodulina and Kramerov 2001) or carry 3′ terminal tandem repeats (Gilbert and Labuda 1999; Ogiwara et al. 2002). Some authors confine the term "tail" to the 3′ terminal simple repeats and include the LINE-derived region in the SINE body (Vassetzky and Kramerov 2013). As in most of the literature on complex SINEs containing a central conserved domain, the term tail is used for the LINE-derived segment at the 3′ end of SINEs; we will use it in this sense throughout the article. All SINEs described to date are transcribed by RNA polymerase III.

Central conserved domains found in SINE bodies range in length between 45 and 150 bp. To date, eight such domains have been described: CORE (Gilbert and Labuda 1999; Munemasa et al. 2008), V (Ogiwara et al. 2002), Ceph (Akasaki et al. 2010), Deu (Nishihara et al. 2006), Nin-DC (Piskurek and Jackson 2011), Inv/alpha (Luchetti and Mantovani 2013; Vassetzky and Kramerov 2013), beta (Vassetzky and Kramerov 2013), and Pln (Luchetti and Mantovani 2013). In some cases central conserved domains were found to be part of others or to overlap with them: The Nin-DC domain makes up part of the Deu domain (Piskurek and Jackson 2011); Inv/alpha overlaps with the 5′ end of Nin-DC (Luchetti and Mantovani 2013).

Central domains are conserved across species, families, classes, even phyla and are combined with variable head and tail regions as well as additional body sequences in specific SINE families. Two SINE superfamilies with central conserved domains (Ceph and Nin-DC) have been described in mollusks (Akasaki et al. 2010; Piskurek and Jackson 2011).

SINEs present a remarkable variety. As of May 2015, SINE Base (Vassetzky and Kramerov 2013) lists a total of 213 different SINE families. Diversification of SINE families in evolution is frequently the result of the exchange of modules between them and between SINEs and LINEs (Kramerov and Vassetzky 2011). Nonallelic homologous recombination and template switch of the LINE reverse transcriptase (RT) are discussed as mechanisms facilitating the exchange of modules (Nishihara et al. 2006; for review, see Kramerov and Vassetzky 2011).

Taking advantage of the rapidly growing amount of next generation and other sequencing data for nonmodel taxa, we set out to identify and characterize SINE families in gastropod and bivalve mollusks. Here we report two novel superfamilies of SINEs containing a central conserved domain: MESC (Romanian for MElc—snail and SCoica—mussel) SINEs and Snail SINEs. We provide evidence for the presence of Nin-DC SINEs—previously reported in *Aplysia californica* and *Lottia gigantea* (Piskurek and Jackson 2011)—in additional mollusk species. We describe and characterize SINEs related to V-SINEs in bivalves and CORE SINEs in the genomes of *Ap. californica* and *Crassostrea gigas.* Finally, we demonstrate that the head region of SINEs can be derived from LINE elements.

Materials and Methods

SINE Identification and Sequence Retrieval

SINEs in the *Littorina saxatilis* BAC sequences (GenBank accession numbers CR974470, CT476813, CT027673, and CT757510) were identified using PILER (Edgar and Myers 2005). They are listed in supplementary table S1, Supplementary Material online. SINEs in all other species were identified by homology search using BLAST (Altschul et al. 1990) at http://blast.ncbi.nlm.nih.gov/Blast.cgi, last accessed January 6, 2016. Either complete SINEs or individual SINE modules were used as query sequences. The queries used for identification of particular SINE superfamilies or families are specified in the respective sections. In case of matches to SINE modules, between 200 and 500 bp of flanking sequences (depending on the position of the module in the SINE) were retrieved for a smaller number of matches. These were aligned and a provisional consensus sequence was constructed. The provisional consensus was then used to retrieve a larger number of elements. In case of SINEs in species for which the entire genome sequence is available, the provisional consensus sequences were used to repeat-mask the genomic sequence using a locally implemented version of RepeatMasker (Smit et al.). Output files were filtered by "in-house" R scripts in order to keep only elements that lacked extensive truncations.

Multiple alignments were constructed using Bioedit. Bioedit was also used for the generation of consensus and contig sequences and the calculation of identity to the consensus sequence. The databases accessed through BLAST as well as the number and characteristics of individual sequences used for construction of consensus/contig sequences are listed in supplementary table S2, Supplementary Material online.

SINE Annotation and Analysis

Target site duplications (TSDs) were annotated manually. Homology to mollusk tRNAs was established using BLAST (Altschul et al. 1990). Homology to previously described SINE modules and LINEs was identified using RepeatMasker

(Smit et al.), CENSOR (Kohany et al. 2006), and SINE Base (Vassetzky and Kramerov 2013). Subfamilies were identified manually based on shared substitutions when compared with the overall consensus of the SINE family. For the identification of partner LINEs of Vhc_1 SINEs in *Cr. gigas,* 3′ UTR (untranslated region) sequences of autonomous non-LTR is now explained at first occurence in line 49 retrotransposons were obtained from Repbase (Jurka et al. 2005) and subsequently aligned to the tail sequences of Pteriomorpha Vhc_1 SINEs. RNA secondary structure prediction was performed using the Vienna RNA package (Lorenz et al. 2011).

Extraction of Genomic DNA and PCR (polymerase chain reaction) Amplification

Littorina saxatilis were kindly provided by Stefano Mariani, Jennifer Coughlan—University College Dublin/School of Biology and Environmental Science, Dublin, Ireland. Genomic DNA was isolated using the Nucleospin Tissue Kit (Macherey and Nagel) according to the manufacturer's instructions. SINE internal sequences were amplified using the primer A_tRNA 5′-TAGCTCAGTCGGTAGCGCGC-3′ in combination with either 3′REV 5′-GGTTACGTGTGAGGTCGTTTGCC C-3′ (Lsa_6) or Lsa_7_REV 5′-GGAAAGCGAGCTGCCATAC-3′ (Lsa_7). PCR products were subcloned into pTZ57R/T (Fermentas) and Sanger sequenced. Sequences obtained were deposited under the GenBank accession numbers KJ549622– KJ549636.

Results and Discussion

As a starting point, we identified SINE families in *Li. saxatilis* (rough periwinkle) BAC sequences (Wood et al. 2008) by PILER (Edgar and Myers 2005) analysis. Characterization of the repetitive sequences obtained revealed the presence of five different SINE families (Lsa_1 to Lsa_5). While our analysis was under way, two of the families (Lsa_2 [Wood et al. 2008] and Lsa_3 [McInerney et al. 2011]) were reported by other groups. A detailed structural characterization, however, had not been performed. The *Li. saxatilis* SINE sequences were subsequently used in homology searches across bivalve and gastropod sequences available in databases. All SINE families identified, their host species, and the source of the sequences analyzed are listed in table 1.

MESC—A SINE Core Domain for Snails and Mussels

Sixteen copies of Lsa_1 were identified in the *Li. saxatilis* BAC sequences. Six of the copies were found to be 5′ truncated. TSDs could be identified for ten elements, both full length and 5′ truncated (supplementary table S1, Supplementary Material online). The 5′ part of Lsa_1 SINEs can be folded in a tRNA-like structure (not shown). A and B promoter boxes for RNA polymerase III transcription are discernible (fig. 1). To date, *Cr. gigas* is the only mollusk species for which tRNA genes are annotated. Comparison of the Lsa_1 5′ domain with the *Cr. gigas* set provided at Ensembl (http://metazoa.ensembl.org/Crassostrea_gigas/Info/Index, last accessed January 6, 2016) identified tRNA-Arg as its most likely source. A sequence homologous to *Cr. gigas* tRNA-Arg could also be retrieved from a *Li. saxatilis* short-read archive. Among SINE tRNA-derived heads, the Lsa_1 5′ domain is most closely related to that of sea urchin SINE2-1_SP (Kapitonov and Jurka 2005b) and SINE2-2_SP (Kapitonov and Jurka 2005c) (supplementary fig. S1, Supplementary Material online). The tRNA-unrelated region of Lsa_1 does not show any homology to previously described SINE or LINE sequences. Lsa_1 SINEs terminate with CAAA repeats.

BLAST search using the Lsa_1 consensus as query retrieved sequences from 16 gastropod and bivalve species with homology covering the tRNA-related region and the 5′ part of the tRNA-unrelated region. The availability of sufficient sequence information up- and downstream of the elements permitted the identification of TSDs in 10 of the 16 species. For the remaining species, partial sequences retrieved from short-read archives and the expressed sequence tags (EST)/non-redundant (nr) sections of GenBank (*Argopecten*) were used to generate contigs (supplementary table S2, Supplementary Material online). Alignment of the consensus/contig sequences (fig. 1) shows that a 95-bp segment immediately downstream of the head region (as defined by homology to the sea urchin SINEs and marked by an arrowhead in fig. 1) is highly conserved in Caenogastropoda and Vetigastropoda and in SINEs from all four bivalve subclasses (Paleoheterodonta, Heterodonta, Pteriomorpha, and Protobranchia; fig. 2). This central part appears to be restricted to SINEs in Caenogastropoda, Vetigastropoda, and Bivalvia. SINEs containing the conserved domain could not be identified in the available sequences of Heterobranchia and Patellogastropoda, the other two major gastropod clades. Neither could homologous sequences be retrieved in other classes of the phylum Mollusca. As all of the species in which the domain could be identified to date are either snails or mussels, we named it MESC domain.

At the 3′ end the MESC domain is followed by an approximately 50-bp region shared between the SINEs in bivalves and in the vetigastropod *Haliotis* (abalone) (boxed in fig. 1; dark gray boxes in fig. 2). SINE tails are shared by MESC elements in Paleoheterodonta, Heterodonta, and Protobranchia (fig. 1), suggesting that these elements are mobilized by the same as yet unidentified LINE element. The most frequently found terminal tandem repeats are $(CA)_n$ or $(CAAA)_n$. *Haliotis* MESC SINEs terminate with $(CACCT)_n$, and *Bathymodiolus* elements with $(CACT)_n$. Within *Haliotis* MESC SINEs expansion of a $(CA)_n$ microsatellite is observed (fig. 1).

The analysis also recovered potential SINEs from pectinoid bivalves (scallops). Homology between the pectinoid sequences identified extends up- and downstream of the MESC domain (fig. 1). For Farrer's scallop (*Chlamys farreri*/

Table 1

Summary of the SINE Families Identified in the Study[a]

SINE Superfamily	Class_subclass/clade[b]	Species	SINE family	Sequence Source[c]
MESC	Gp_Caenogastropoda	*Littorina saxatilis*	Lsa_1	Genomic
		Littorina littorea	Lli	Transcriptome
		Crepidula fornicata	Cfo	Transcriptome
		Strombus gigas	Sgi	Transcriptome
	Gp_Vetigastropoda	*Haliotis discus/Haliotis* sp.	Hdi	Genomic
	Bv_Pteriomorpha	*Bathymodiolus azoricus*	Baz	Transcriptome
	Bv_Heterodonta	*Mercenaria mercenaria*	Mme	Genomic
		Sinonovacula constricta	Sco	Transcriptome
		Arctica islandica	Ais	Transcriptome
	Bv_Pteriomorpha	*Argopecten irradians*	Air	Transcriptome genomic
		Chlamys farreri	Cfa	Genomic
		Mizuhopecten yessoensis	Mye	Transcriptome
	Bv_Anomalodesmata	*Laternula elliptica*	Lel	Transcriptome
	Bv_Paleoheterodonta	*Elliptio complanata*	Eco	Transcriptome genomic
		Villosa lienosa	Vli	Transcriptome
		Hyriopsis cumingii	Hcu	Transcriptome genomic
	Bv_Protobranchia	*Nucula tenuis*	Nte	Transcriptome
na (Lsa_1 head)	Bv_Pteriomorpha	*Tegillarca granosa*	Tgr_1	Transcriptome
	Bv_Heterodonta	*Meretrix meretrix*	Mme_1	Transcriptome
LP	Gp_Caenogastropoda	*Littorina saxatilis*	Lsa_2	Genomic
	Gp_Vetigastropoda	*Haliotis diversicolor*	Hdi_LP	Genomic
		Haliotis midae	Hmi_LP	Transcriptome
	Bv_Pteriomorpha	*Ostrea* sp.	Oco/Oed_LP	Genomic
LP/Nin		*Saccostrea glomerata*	Sgl_LP	Transcriptome
Nin			Sgl_tRNA_1; Sgl_tRNA_2	Transcriptome
V/Vhc	Bv_Pteriomorpha	*Crassostrea gigas*	Cgi_Vhc_1	Reference genome
		Tegillarca granosa	Tgr_Vhc_1	Transcriptome
		Chlamys farreri	Cfa_Vhc_1; Cfa_Vhc_2	Genomic
		Mizuhopecten yessoensis	Mye_Vhc_1; Mye_Vhc_2	Transcriptome
		Nodipecten subnodosus	Nsu_Vhc_1	Transcriptome
		Argopecten irradians	Air_Vhc_1; Air_Vhc_2	Transcriptome
	Bv_Protobranchia	*Yoldia limatula*	Yli_Vhc	Transcriptome
	Gp_Caenogastropoda	*Strombus gigas*	Sgi_Vhc	Transcriptome
	Insecta	*Thermobia domestica*	Tdo_Vhc	Transcriptome
	Insecta	*Empusa pennata*	Epe_Vhc	Transcriptome
Snail	Gp_Caenogastropoda	*Littorina saxatilis*	Lsa_3; Lsa_4; Lsa_5; Lsa_6	Genomic
		Potamopyrgus antipodarum	Pan_1; Pan_2; Pan_3	Transcriptome
		Nucella lapillus	Nla_1; Nla_2; Nla_3	Transcriptome
		Strombus gigas	Sgi_1	Transcriptome
Nin	Gp_Caenogastropoda	*Littorina saxatilis*	Lsa_Nin_1; Lsa_Nin_2	Genomic transcriptome
		Nucella lapillus	Nla_Nin	Transcriptome
na (Lsa_5/Lsa_6 head)	Gp_Caenogastropoda	*Littorina saxatilis*	Lsa_7; Lsa_8	Genomic
CORE	Gp_Heterobranchia	*Aplysia californica*	Aca_CORE_1; Aca_CORE_2	Reference genome
	Bv_Pteriomorpha	*Crassostrea gigas*	Cgi_CORE_1; Cgi_CORE_2	Reference genome
	Gp_Caenogastropoda	*Cre. fornicata*	Cfo_CORE	Transcriptome

Note.—na, not applicable.

[a]Details for the data sets are provided in supplementary table S2, Supplementary Material online.

[b]Gp, Gastropoda; Bv, Bivalvia.

[c]Genomic sequences were retrieved from the nr and GSS subsets of GenBank. Transcriptome refers to sequences retrieved from short-read archives (SRA), transcriptome assemblies (TSA and other sources), and ESTs.

Azumapecten farreri), a total of 16 sequences spanning the entire length of the potential SINE was retrieved. TSDs could be determined for 11 of them (supplementary fig. S2, Supplementary Material online).

The elements found in Farrer's scallop are 520-bp long. The sequence of the 5′-most 68 bp is repeat-masked by Censor (Kohany et al. 2006). It is similar to the first 68 bp (5′ UTR) of Nimb-17_LMi—an autonomous non-LTR retrotransposon

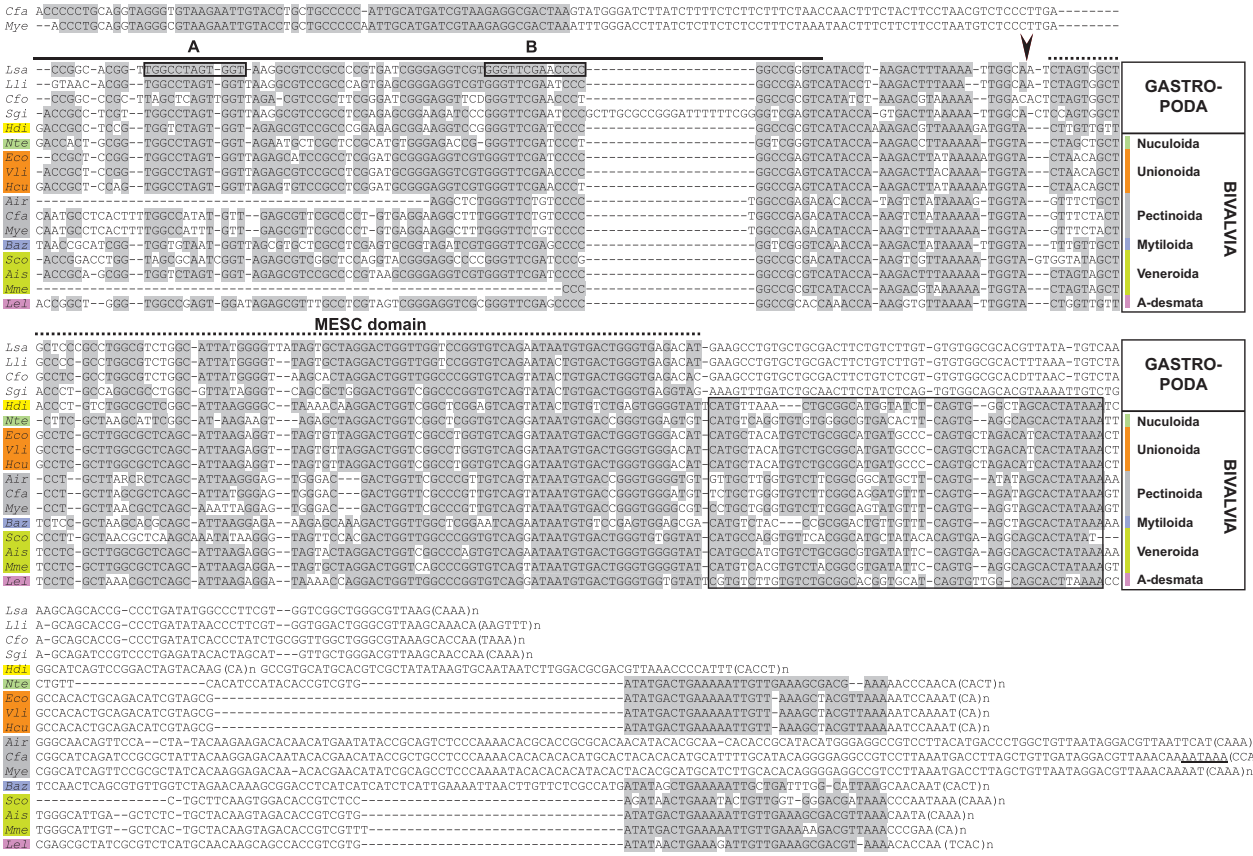


Fig. 1.—MESC SINEs in gastropods and bivalves. A multiple alignment of consensus/contig sequences is shown. The shaded sequence at the 5′ end of *Chlamys farreri* (*Cfa*) and *Mizuhopecten yessoensis* (*Mye*) SINEs are homologous to the 5′ end of Nimb-17_LMi (compare fig. 3*C*, see text for details). The tRNA-derived region is marked with a bold line on top of the alignment block; A and B promoter boxes are boxed in the *Littorina saxatilis* (*Lsa*) sequence. The arrowhead indicates the 3′ end of homology to the heads of *Strongylocentrotus purpuratus* SINE2-1_SP and SINE2-2_SP (for an alignment to these two SINEs, see supplementary fig. S1, Supplementary Material online). The broken line indicates the MESC central conserved domain. The approximately 50-bp sequence shared by MESC SINEs in bivalves and the vetigastropod *Haliotis* is boxed. Conservation of the 3′ end in all bivalve MESC SINEs (except those of Pectinoida) is indicated by identity shading. The putative polyadenylation signal in *Ch. farreri* MESC SINEs is underlined. Color coding of related species corresponds to that in figure 2. *Lli*, *Littorina littorea*; *Cfo*, *Crepidula fornicata*; *Sgi*, *Strombus gigas*; *Hdi*, *Haliotis discus*; *Nte*, *Nucula tenuis*; *Eco*, *Elliptio complanata*; *Vli*, *Villosa lienosa*; *Hcu*, *Hyriopsis cumingii*; *Air*, *Argopecten irradians*; *Baz*, *Bathymodiulus azoricus*; *Mme*, *Mercenaria mercenaria*; *Sco*, *Sinonovacula constricta*; *Ais*, *Arctica islandica*; *Lel*, *Laternula elliptica*.

present in the genome of *Locusta migratoria*. Nimb-17_LMi belongs to the Nimb clade of I-like non-LTR retrotransposons that comprises elements in fish, mollusks, sea squirts, sea urchins, and insects (Kapitonov and Jurka 2009). The LINE-derived 5′ domain of the pectinoid MESC SINEs is followed by an 88-bp CT-rich region of unknown origin (fig. 1). The acquisition of additional sequences upstream of the tRNA-derived head is reminiscent of the situation found in a subgroup of Deu-SINEs in which a 5S rRNA sequence has been fused to the 5′ end of the tRNA-derived head and, following loss of the 3′ part of the tRNA-derived region, has taken over promoter function (Nishihara et al. 2006). As we will demonstrate and discuss in the next section, the LINE-derived sequence as it is found at the 5′ end of MESC SINEs in pectinoid bivalves (fig. 1) cannot only be added to but can also replace tRNA-derived heads in SINEs (fig. 3*B*).

None of the sequences available for *Argopecten irradians* and *Mizuhopecten yessoensis* covers an entire element. In these cases, contig sequences were generated and used in the alignment. Sequences indicating the presence of MESC-domain SINEs containing a LINE-derived 5′ domain are also found in *Placopecten magellanicus* (sea scallop) and *Pecten maximus* (king scallop) transcriptome shotgun assemblies. However, the frequent occurrence of sequences with identical 5′ flanking sequence/different 3′ flanking sequences or different 5′ flanking sequence/identical 3′ flanking sequences suggests misassembly across highly similar SINE internal sequences.

LP-SINEs: Harnessing a LINE Promoter

Based on the observation that the head regions of all SINEs characterized so far are derived from small cellular RNAs and

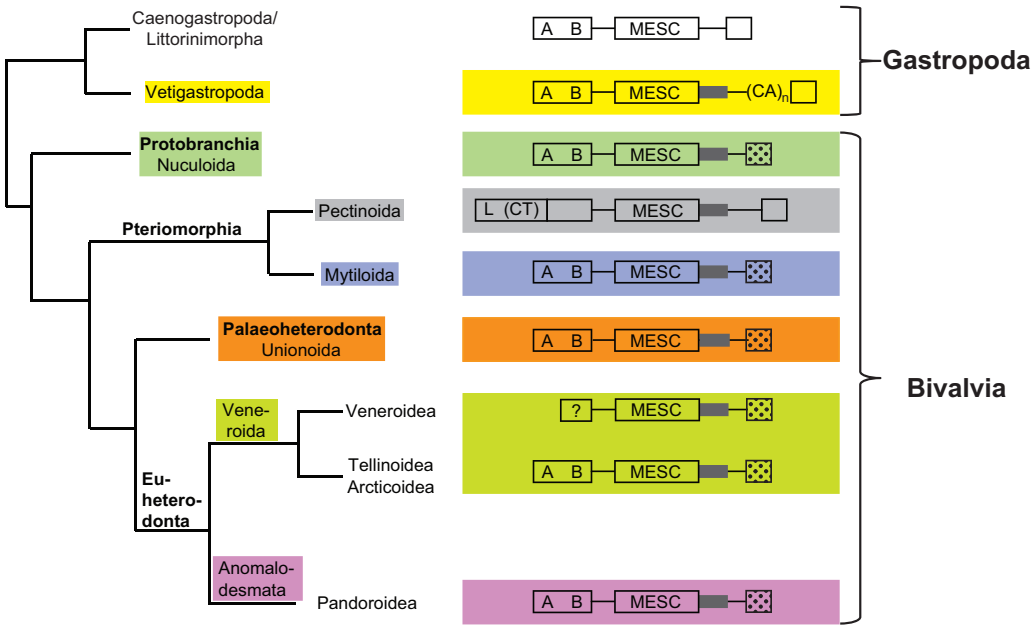


Fig. 2.—MESC SINEs are widely distributed among gastropods and bivalves. A schematic representation of MESC domain SINEs (not drawn to scale) and their distribution in gastropod/bivalve subclasses and families is shown. A and B denote the promoter boxes in the tRNA-derived head regions. Pectinoid MESC domain SINEs are characterized by a LINE element-derived segment (L) and a short C/T-rich sequence at their 5′ ends. Dark gray boxes indicate a sequence stretch conserved in bivalves and the vetigastropod *Haliotis*. $(CA)_n$ denotes a microsatellite present in *Haliotis* MESC SINEs. Dotted boxes represent tail sequences shared between all bivalve MESC SINEs except those in Pectinoida. Empty boxes at the 3′ ends of SINEs in Gastropoda and Pectinoida represent tails of unknown origin that are not shared across clades or subclasses. The topology of the schematic representation for bivalves is based on Gonzalez et al. (2015); the two gastropod clades were added at the appropriate position.

that they are transcribed by RNA polymerase III (for review, see Kramerov and Vassetzky 2011), transcription by this polymerase has been included into the definition of this type of retrotransposon. However, as described above, in the genomes of pectinoid bivalves there are SINEs whose head region/putative promoter is derived from a LINE element. It is, therefore, possible that they are transcribed by RNA polymerase II. The presence of variable size TSDs directly flanking the elements and of 3′ terminal short tandem repeats (typical for SINEs) suggests that they are mobilized by a LINE-dependent mechanism. SINEs with a similar LINE-derived 5′ domain are also found in *Li. saxatilis* (family Lsa_2). Nine copies of Lsa_2 could be identified in the BAC sequences. Six of these display variable size TSDs (supplementary table S1, Supplementary Material online). Homology search using the Lsa_2 head as a query and subsequent analysis of the sequences retrieved revealed that TSD-flanked elements with similar LINE-derived heads are also present in the vetigastropod *Haliotis* (*Haliotis diversicolor* and *Haliotis midae*) and in the ostreoid bivalve *Ostrea* (*Ostrea conchaphila* and *Ostrea edulis*) (fig. 3*A*; for statistics on TSDs, see supplementary table S2, Supplementary Material online). In the Sydney rock oyster (*Saccostrea glomerata*, Ostreoida), a SINE family with the LINE-derived head shares the body and tail with two families of tRNA-derived SINEs (fig. 3*B*). The body of the three *Saccostrea* SINE families contains a partial Nin-DC domain. The *Haliotis* SINEs and the SINEs in *Ostrea* and *Saccostrea* share the tail sequence (fig. 3*A*), which suggests that they are mobilized by the same, as yet unidentified, LINE element.

Its position suggests that the 5′ LINE-derived sequence might function as the promoter of the elements. We will therefore refer to these retrotransposons as "LP" (LINE promoter) SINEs.

Experimental work is necessary to confirm the promoter function of the LINE-derived fragment and to establish which RNA polymerase transcribes this type of retrotransposon. Interestingly, the 3′ terminal tandem repeats in Lsa_2 elements and *Chlamys* MESC SINEs are preceded by a canonical polyA signal (underlined in figs. 1 and 3*A*). This is similar to the setting found in RNA polymerase II transcribed LINEs that terminate with tandem repeats (e.g., UnaL2 [Kajikawa and Okada 2002] and ZfL3 [Ogiwara et al. 2002]). The hexamers found upstream of the *Haliotis* and *Ostrea* LP-SINE tandem repeats (AATATA and ACTAAA; fig. 3*A*) represent polyA signal variants that were shown to be functional in humans (Beaudoing et al. 2000). The sequence found in *Saccostrea* Sgl_LP SINEs (AGTATA) has not been reported as a polyadenylation signal. However, a comprehensive survey of polyA site usage in mollusks is not available.

Outside mollusks a family of SINEs carrying a similar LINE-derived head could be identified in the genome of the Mediterranean fruit fly (*Ceratitis capitata*, data not shown).

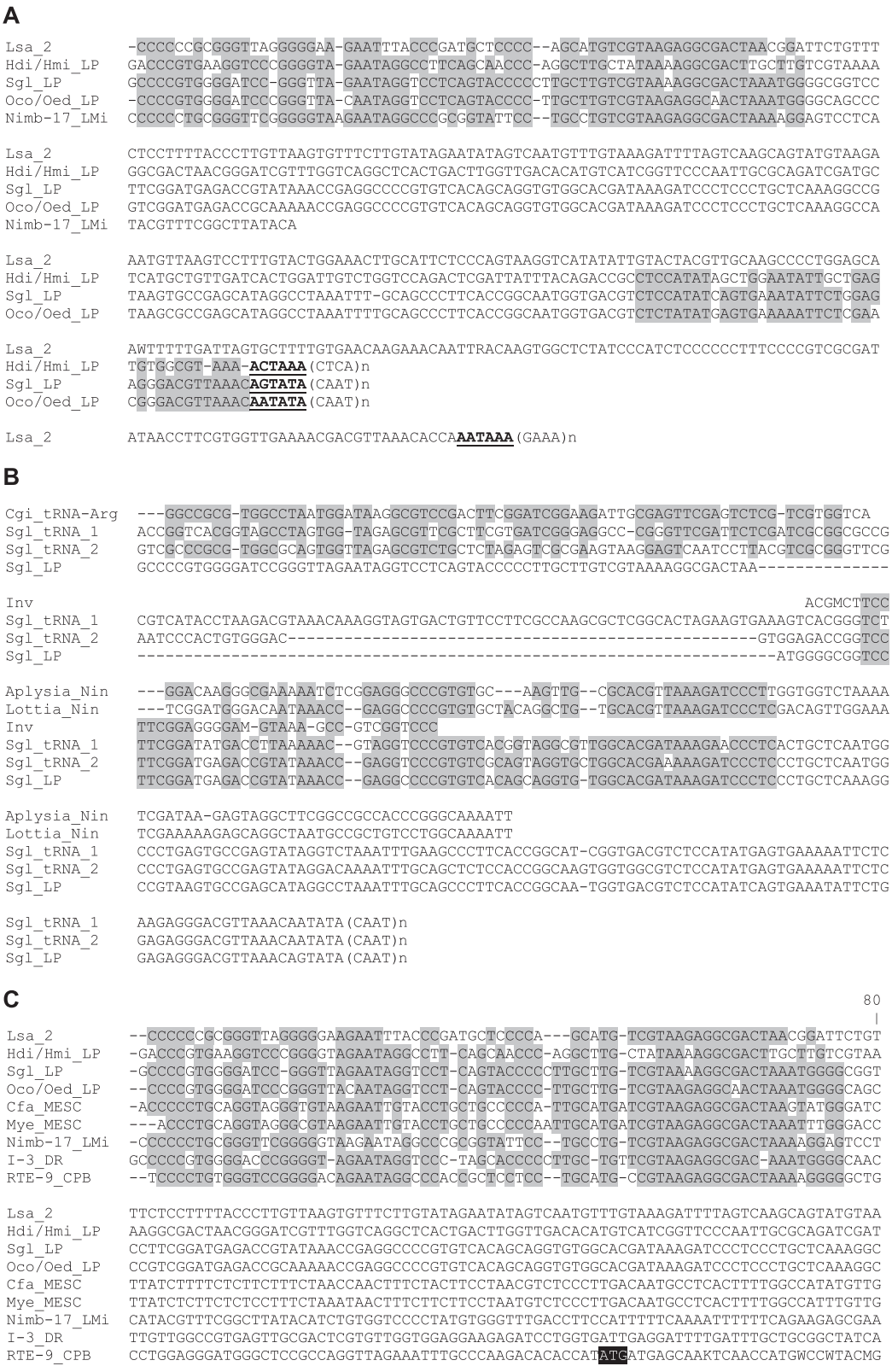


Fig. 3.—SINEs with a LINE-derived head in gastropods and bivalves. (*A*) Multiple alignment of SINE consensus sequences from gastropods (*Littorina saxatilis*, Lsa_2; *Haliotis diversicolor/Haliotis midae*, Hdi/Hmi_LP) and bivalves (*Saccostrea glomerata*, Pteriomorphia, Ostreoida, Sgl_LP; *Ostrea edulis/Ostrea conchaphila*, Pteriomorphia, Ostreoida—Oco/Oed_LP) and the 5′ part of the 5′ UTR of Nimb-17_LMi, a Nimb-clade LINE element from the migratory locust

(continued)

Template switching of the LINE RT has been suggested as the mechanism facilitating the acquisition and exchange of modules in SINE evolution (Akasaki et al. 2010; Kramerov and Vassetzky 2011). This mechanism likely also mediates the recruitment of the LINE 5′ UTR fragment by pectinoid MESC SINEs, the replacement of the tRNA head in the *Saccostrea* SINEs and the generation of the other LP SINE families from as yet unknown ancestors. Interestingly, though, LP SINEs appear to be the only known instance in which LINE sequence constitutes the head of the element. Head regions usually recruited by SINEs are small RNAs—tRNA or 5S rRNA. The preference for switching template to small RNAs is also obvious in the structure of chimeric retrogenes identified in the human genome (Buzdin et al. 2003).

The homology between the LP SINEs identified in *Li. saxatilis*, *Haliotis* sp., *Sa. glomerata,* and the pectinoid MESC-domain SINEs described above is restricted to the first approximately 70 bp (fig. 3*C*). Intriguingly, comparison of the three known LINE families carrying this sequence at their 5′ ends (two Nimb clade families—Nimb-17_LMi and I-3_DR—and an RTE element from the painted turtle—RTE-9_CPB) shows that homology between these is restricted to the same domain as well (fig. 3*C*). In the two Nimb LINEs, the 5′ UTR extends a further ca. 300 bp that do not show any homology. In the turtle RTE LINE, the conserved sequence makes up roughly half of the 5′ UTR (ATG marked in fig. 3*C*). Acquisition of novel 5′ UTRs has been shown to be a frequent event in the evolution of mammalian L1. In these cases, however, the entire UTR differs between subfamilies—with the notable exception of the L1 subfamilies present in the human genome which all share the first approximately 50 bp and the last 20 bp of the UTR (Khan et al. 2006). No hypothesis has been brought forward to explain this phenomenon.

An interesting case in which only the 5′ extremity is conserved between LINE families is represented by the trypanosomatid ingi and L1Tc elements (Bringaud et al. 2002). Promoter (Heras et al. 2007) as well as ribozyme (Sanchez-Luque et al. 2011) activity has been demonstrated for the 77-bp fragment shared between the otherwise very different elements. Recently, Tx1 LINE families from crocodilians that share U1 or U2 snRNA derived 5′ ends with *SINEU* families have been reported (Kojima 2015).

Once additional LINEs and SINEs carrying the particular 5′ end found in the locust, zebrafish, and turtle LINE elements will have been identified, it will be interesting to see how they evolved, and how frequently exchanges between them and SINEs have taken place.

The V Domain in Mollusk SINEs

Whereas the central MESC domain appears to be widely distributed among bivalves, homology search using the Lsa_1 head retrieved SINE sequences from only two bivalve species (*Tegillarca granosa*, Arcoida and *Meretrix meretrix*, Veneroida; supplementary fig. S3, Supplementary Material online). In both cases, it was found to be combined with different bodies and tails. The tail of the *Me. meretrix* elements, however, is shared by yet another superfamily of bivalve SINEs (fig. 4*A*). Surprisingly, these were found to exhibit homology to V-SINEs, which were first described in vertebrates (Ogiwara et al. 2002). Subsequently, using the V-domain as query, we were able to identify a second bivalve SINE superfamily displaying homology to this domain (fig. 4*C*). In both superfamilies homology to the vertebrate V-domain covers its 5′ end and terminates at the position of the 3′ end of the V-domain as it is found in the lamprey Lam1 SINE (Ogiwara et al. 2002). A highly conserved block of 39 bp is found at the 3′ end of the V-homology in the bivalve SINEs. On average, 90% (first superfamily) and 78% (second superfamily) of the nucleotides in this block are identical to the consensus across all fish V-SINEs reported by Ogiwara et al. (2002) (fig. 5*A*). The same subdomain is highly conserved (>80% identity to the "all fish" consensus) also in SINE families identified in a gastropod (*Strombus gigas*), a protobranch bivalve (*Yoldia limatula*) and in arthropods (*Empusa pennata*, *Thermobia domestica*, consensus sequences for these SINE families are provided in supplementary fig. S4, Supplementary Material online, fig. 5*A*). We, therefore, suggest to refer to these SINEs as "Vhc" (V highly conserved) SINEs.

SINEs of the first bivalve Vhc superfamily (Vhc_1; fig. 4*A*) were identified in Pteriomorpha (saltwater clams). They are found in Ostreidae (*Cr. gigas*), Arcidae (*T. granosa*), and Pectinidae (*Ch. farreri*, *Mi. yessoensis*, *Nodopecten subnodosus,* and *Ar. irradians*). The head region of these SINEs is

Fig. 3.—Continued

(*Locusta migratoria*). The Nimb-17_LMi sequence was obtained from Repbase (Jurka et al. 2005). Threshold for identity shading of the head and tail sequences is 50%. Putative polyadenylation signals are boldface and underlined. (*B*) The *Saccostrea* Sgl_LP SINE shares a partial Nin-DC domain and the tail with two tRNA-derived SINEs (Sgl_tRNA_1 and Sgl_tRNA_2). A tRNA-Arg from *Crassostrea gigas* (Cgi; SuperContig scaffold42556: 10,784–10,856), the Nin-DC domains of SINEs from *Aplysia californica* (Aplysia_Nin) and *Lottia gigantea* (Lottia_Nin) (Piskurek and Jackson 2011), and a consensus sequence of the recently described Inv domain (Luchetti and Mantovani 2013) are given for comparison. Note that homology to the Nin-DC domain extends 3′ of the part covered by Inv. Residues identical to *Cr. gigas* tRNA-Arg are highlighted in gray. Threshold for identity shading in the Inv/Nin domain is 50%. (*C*) Homology among LP-SINEs, pectinoid MESC SINEs (*Chlamys farerri*, Cfa_MESC; *Mizuhopecten yessoensis*, Mye_MESC), and LINE elements is restricted to the 5′ part of the LINEs' 5′ UTR. LINE sequences shown are Nimb-17_LMi (*Loc. migratoria*, Nimb clade), I-3_DR (*Danio rerio*, Nimb clade), and RTE-9_CPB (*Chrysemys picta bellii*, RTE clade). The ATG of RTE-9_CPB is indicated, the start codons of the other two LINE elements are located approximately 400 bp downstream of the transcription start site. LINE sequences were retrieved from Repbase (Jurka et al. 2005). Threshold for identity shading is 50%.

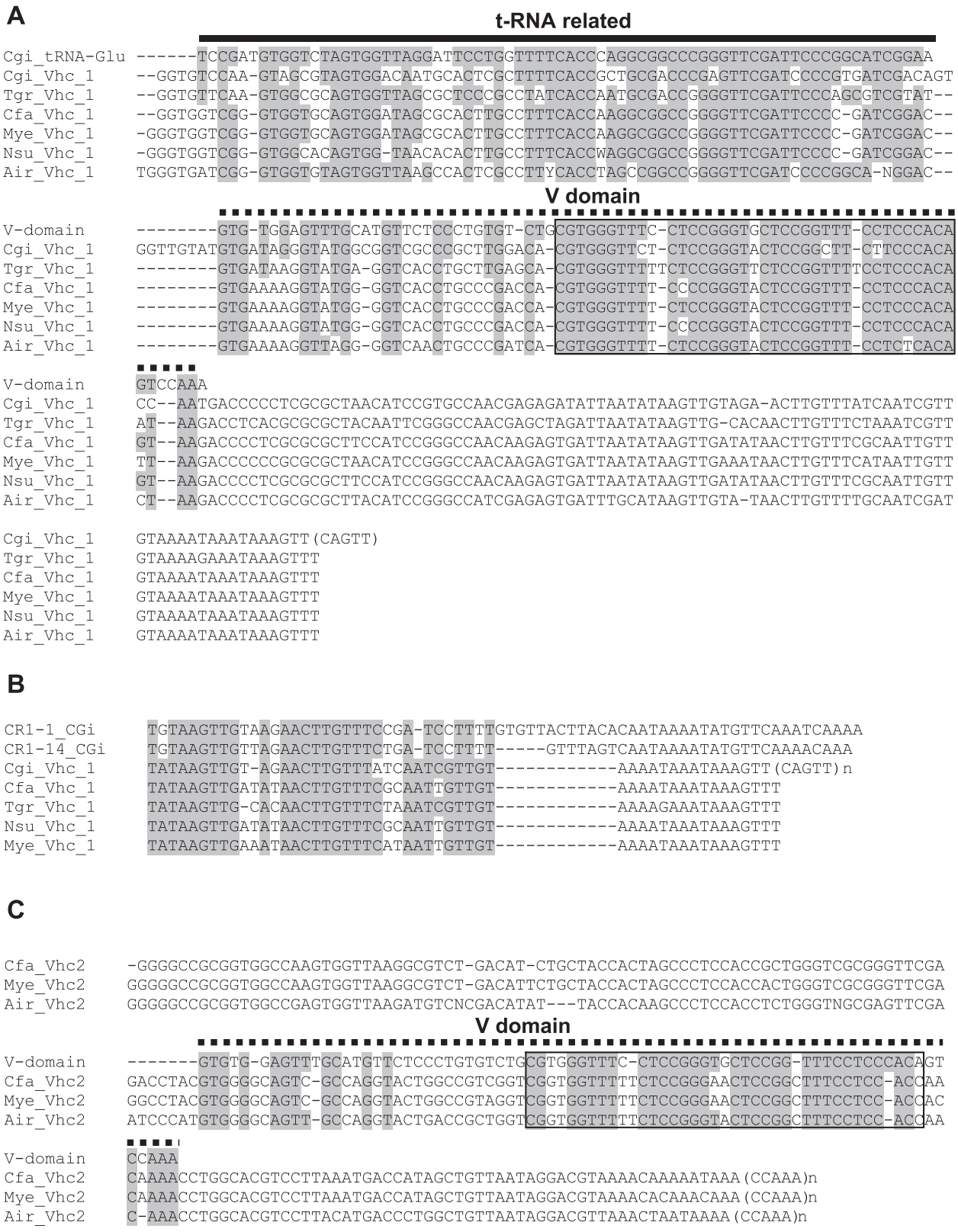


Fig. 4.—Bivalve Vhc SINEs and their partner LINEs. (*A*) Alignment of Vhc_1 SINEs identified in pteriomorph bivalves. V-domain denotes a consensus sequence over the V domains of the V SINEs reported by Ogiwara et al. (2002). Cgi_tRNA-Glu represents a *Crassostrea gigas* tRNA (supercontig:GCA_000297895.1:scaffold474:30309–30980). The tRNA-derived head is marked by a bold line on top of the aligned sequences; the

(continued)

tRNA derived; the best match in the *Cr. gigas* genome being tRNA-Glu. Vhc_1 SINEs lack well-defined 3′ terminal tandem repeats, with exception of *Cr. gigas* (pacific oyster) where some elements terminate in $(CAGTT)_n$. TSDs could be identified for Vhc_1 SINEs in five of the six species in which they are found (supplementary table S2, Supplementary Material online). In *Cr. gigas*, for which the sequence of the entire genome is available, 725 Vhc_1 SINE copies were identified by RepeatMasker (Smit et al.). They cover 0.03% of the genome (supplementary table S3, Supplementary Material online). On average, Cgi_Vhc_1 SINEs show 11% divergence from the consensus. However, there are smaller groups (9–20 elements per group) that are on average up to 99.6% identical to their respective group consensus (supplementary fig. S5, Supplementary Material online). This indicates that Vhc_1 SINEs in *Cr. gigas* might still be active.

Homology between the 3′ tail sequence of Vhc_1 SINEs and those of the *Cr. gigas* autonomous non-LTR retrotransposons CR1-1_CGi and CR1-14_CGi (Jurka 2012; Zhang et al. 2012; Bao and Jurka 2013; fig. 4*B*) suggests that elements of these LINE families could be the autonomous partners of Vhc_1 SINEs. Both LINE elements belong to the CR1 clade.

Elements of the second bivalve Vhc SINE superfamily—Vhc_2 (fig. 4*C*)—are absent from the *Cr. gigas* (Ostreidae) genome and could not be identified in the *T. granosa* (Arcidae) ESTs available in the database. In Pectinidae, Vhc_2 SINEs are present in *Ch. farreri*, *Mi. yessoensis,* and *Ar. irradians.* No closely matching mollusk tRNA could be identified for the Vhc_2 SINE 5′ domain. Potential partner LINEs that share the tail with Vhc_2 SINEs could not be identified in the available database sequences.

It is worthwhile noticing that Vhc_1 and Vhc_2 SINEs display internal stretches of four and five consecutive T residues, respectively. OligodT stretches $\geq$4 bp are also found in the recently described crocodilian SINEU families (Kojima 2015) and in some of the gastropod Snail SINEs described below. Clusters of four or more consecutive T residues were demonstrated to act as RNA polymerase III terminators (Bogenhagen and Brown 1981). A more recent study, however, has shown that the recognition of termination signals by RNA polymerase III is context dependent. In yeast, a C residue immediately downstream of the T stretch weakens its termination potential and favors read-through (Braglia et al. 2005). Interestingly, in the mollusk Vhc and Snail SINEs all potential internal terminators are followed by a C residue; in some cases even by a CT dinucleotide that is even more favorable for read-through (Braglia et al. 2005) (figs. 4 and 7).

As mentioned above, SINE families containing the 37- to 39-bp Vhc subdomain are present in other mollusks (protobranch bivalves and gastropods) and in arthropoda. The corresponding block in the sea anemone (*Nematostella vectensis*, Cnidaria) SINE2-2_NV (Putnam et al. 2007; recently categorized as V-SINE in Vassetzky and Kramerov 2013) is 79% identical to the all fish V-consensus (fig. 5*A*). Thus with representative SINE families in the major eumetazoan phyla Cnidaria, Ecdysozoa (Arthropoda), Lophotrochozoa (Mollusca), and Deuterostomia (as part of the V-domain), the phylogenetic distribution of the Vhc subdomain equals that of the Nin-DC domain (Piskurek and Jackson 2011). The persistence of highly conserved SINE domains across a wide range of species and phyla has been discussed in the context of a possible function of such domains in the host genome, e.g., by providing regulatory sequences (Santangelo et al. 2007; Sasaki et al. 2008). An alternative hypothesis suggests that these domains might be important in maintaining the integrity of the elements and their proliferative capacity (Gilbert and Labuda 2000). The CORE and V conserved domains have been suggested to facilitate the exchange of SINE 3′ ends with active LINE elements (Gilbert and Labuda 2000; Ogiwara et al. 2002). It has also been discussed that the conserved central domain-mediated acquisition of SINE segments might accelerate the formation of complex secondary structures (Sun et al. 2007).

The isolated Vhc subdomain is predicted to form a stable hairpin structure with an internal loop (Minimum Free Energy (MFE) 14.44 kcal/mol; fig. 5*B*). This stem-loop structure is preserved in the context of the entire SINE secondary structure—either as a separate hairpin followed by a stretch of unpaired nucleotides (in Vhc_1 SINEs; fig. 6*A*) or as terminal part of a larger stem loop (in Vhc_2 SINEs; fig. 6*B*). Interestingly, despite the fact that they do not share any other sequence elements with Vhc_1 or Vhc_2 pteriomorph SINEs, the SINEs identified in the protobranch bivalve, the gastropod, and the arthropods fall into either one of the two categories as far as the "embedding" of the Vhc stem loop into the overall secondary structure is concerned: The *Strombus* (Gastropoda) and *Empusa* (Arthropoda) SINEs display the Vhc as separate stem loop comparable with Vhc_1 SINEs (fig. 6*A*), and in the *Yoldia* (Protobranchia), *Thermobia* (Arthropoda), and *Nematostella* (Cnidaria) SINEs the Vhc domain forms the terminal part of a larger stem loop as seen in Vhc_2 SINEs (fig. 6*B*). In the

Fig. 4.—Continued

V domain by a dotted line. The highly conserved Vhc subdomain is boxed. Residues identical to Cgi_tRNA-Glu and to the fish V domain are highlighted in gray. (*B*) Alignment of *Cr. gigas* CR1 element 3′ ends (Jurka 2012; Zhang et al. 2012; Bao and Jurka 2013) to the tails of Vhc_1 SINEs identified in Pteriomorpha. Threshold for identity shading is 50%. (*C*) Multiple alignment of the consensus sequences of Vhc_2 SINEs. The V domain is indicated by a dotted line. The highly conserved Vhc subdomain is boxed. Species names are abbreviated as follows: Cgi, *Crassostrea gigas*; Tgr, *Tegillarca granosa*; Cfa, *Chlamys farreri*; Nsu, *Nodipecten subnodosus*; Mye, *Mizuhopecten yessoensis*; Air, *Argopecten irradians.* Residues identical to the fish V domain are highlighted in gray.

Fig. 5.—Conservation and predicted secondary structure of the Vhc subdomain. (*A*) Multiple alignment of the Vhc subdomains from fish (consensus over all SINEs described in Ogiwara et al. 2002), pteriomorph bivalve Vhc_1 and Vhc_2 SINEs (Cgi, *Crassostrea gigas*; Tgr, *Tegillarca granosa*; Cfa, *Chlamys farreri*; Nsu, *Nodipecten subnodosus*; Mye, *Mizuhopecten yessoensis*; Air, *Argopecten irradians*), and SINEs found in a protobranch bivalve (Yli, *Yoldia limatula*), a gastropod (Sgi, *Strombus gigas*), two arthropods (Epe, *Empusa pennata*; Tdo, *Thermobia domestica*), and the sea anemone (SINE2-2_NV, *Nematostella vectensis*; Jurka et al. 2005). (*B*) Secondary structure of the Vhc subdomain predicted based on the multiple alignment shown in (*A*)—excluding the fish consensus. Complete consensus sequences for all non-pteriomorph SINE families (except SINE2-2_NV) are provided in supplementary fig. S4, Supplementary Material online.

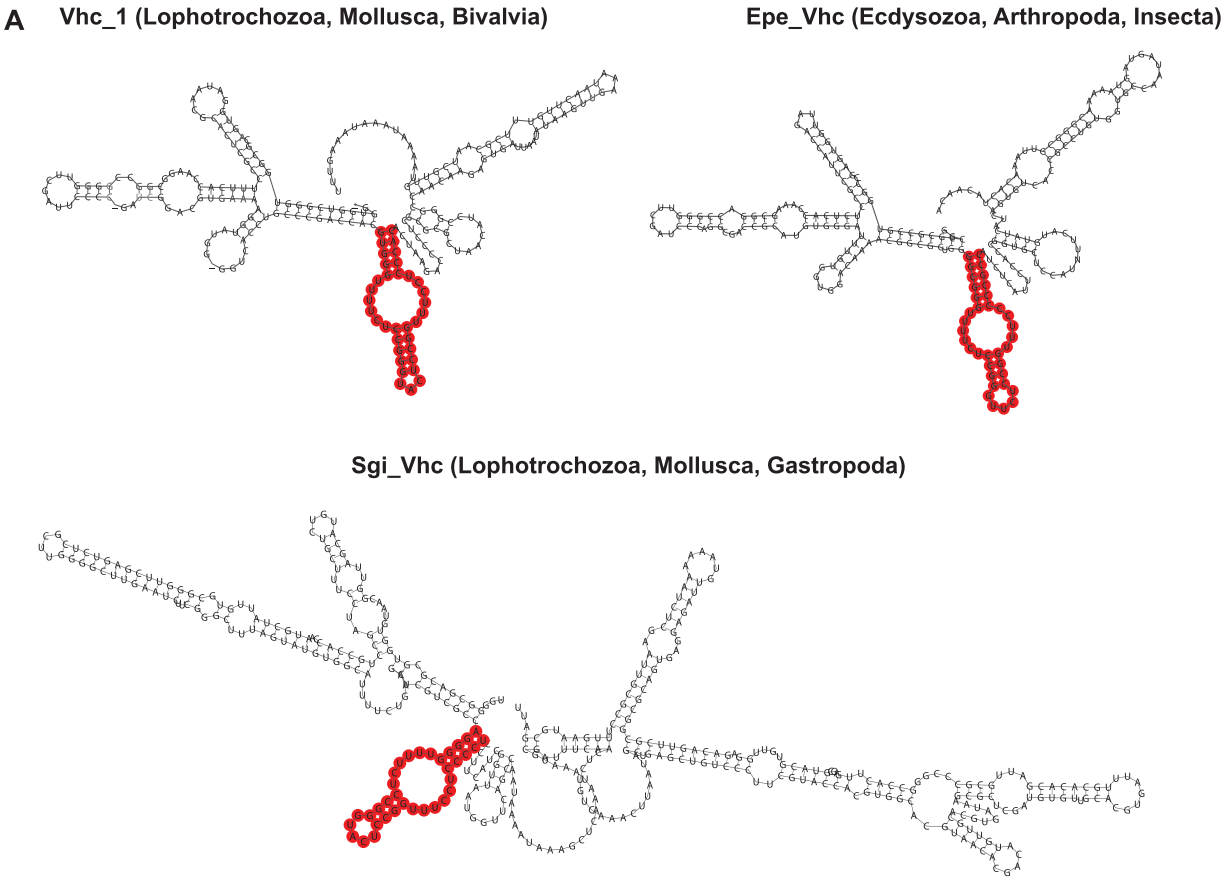


Fig. 6.—The Vhc stem loop is maintained in the context of the overall secondary structure of Vhc SINEs as either a separate stem loop (*A*) or as terminal part of a larger stem loop (*B*). Secondary structure predictions are based on multiple alignments of consensus sequences of Vhc SINEs present in Pteriomorpha (Vhc_1 and Vhc_2) or on single consensus sequences (all other species). The Vhc stem loop is highlighted in red. Complete consensus sequences for all non-pteriomorph SINE families (except SINE2-2_NV) are provided in supplementary figure S4, Supplementary Material online. Epe, *Empusa pennata*; Sgi, *Strombus gigas*; Tdo, *Thermobia domestica*; Yli, *Yoldia limatula*; NV, *Nematostella vectensis*.

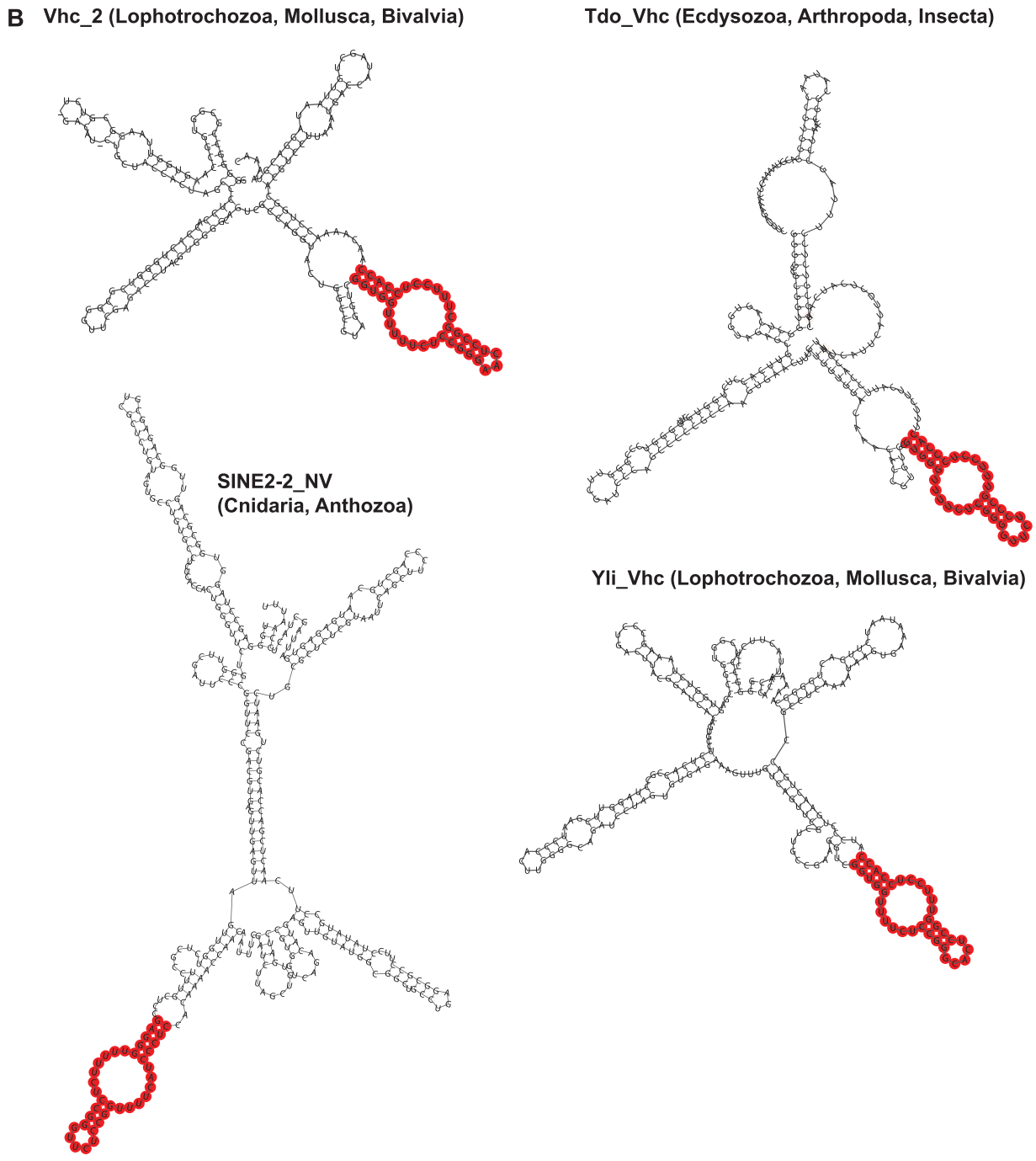


Fig. 6.—Continued.

context of the entire V-domain in fish V-SINEs, in contrast, the Vhc stem loop does not appear to be preserved (data not shown). To establish the significance of the preservation of the Vhc stem loop in invertebrate SINEs, a more detailed structural analysis of a larger number of Vhc SINEs across phyla will be necessary.

The Snail Domain Is Related to Nin-DC and Restricted to SINEs in Caenogastropoda

Finally, PILER analysis identified three more SINE families (Lsa_3, Lsa_4 and Lsa_5, for details see supplementary table S1, Supplementary Material online) that are characterized by either identical heads/5′ parts of the body (Lsa_3 and Lsa_4) or

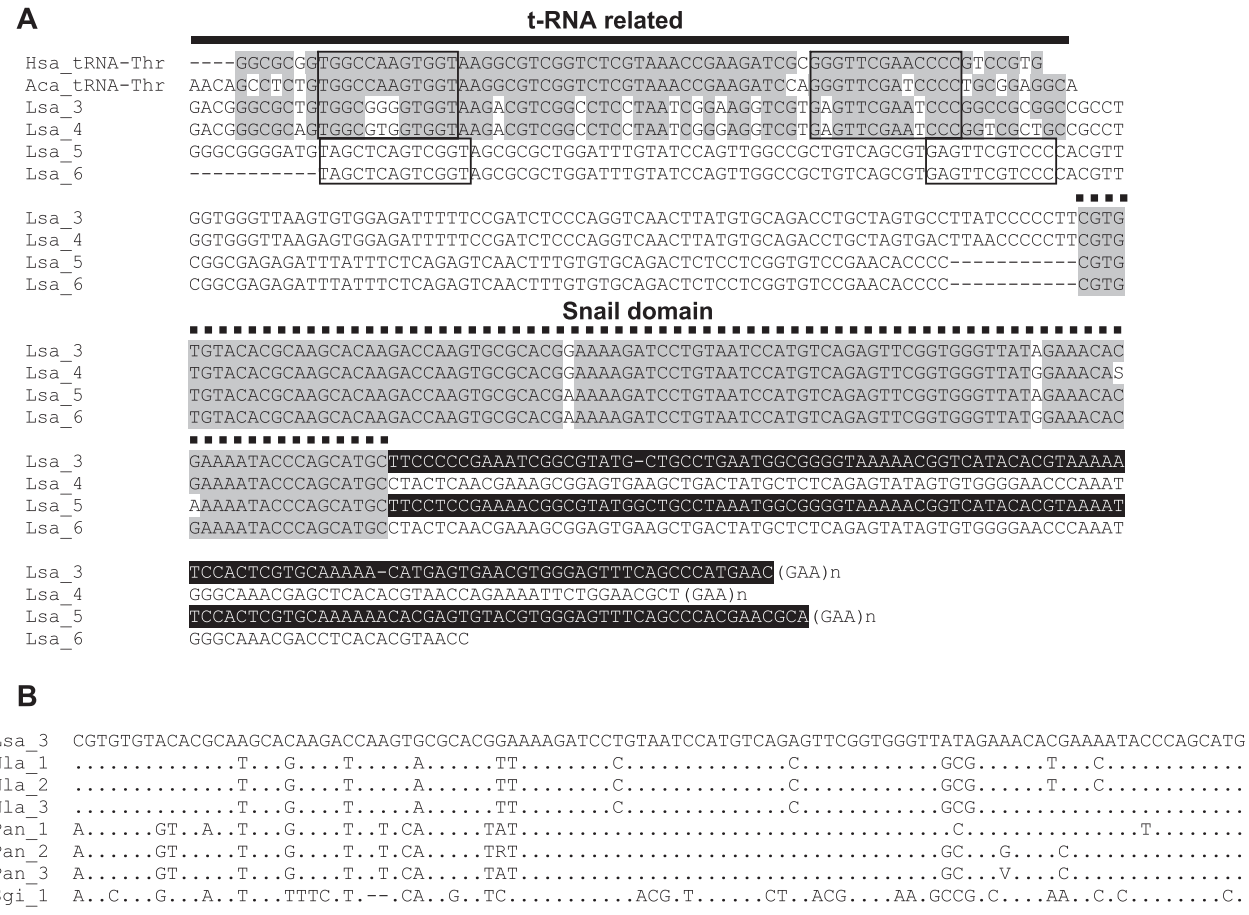
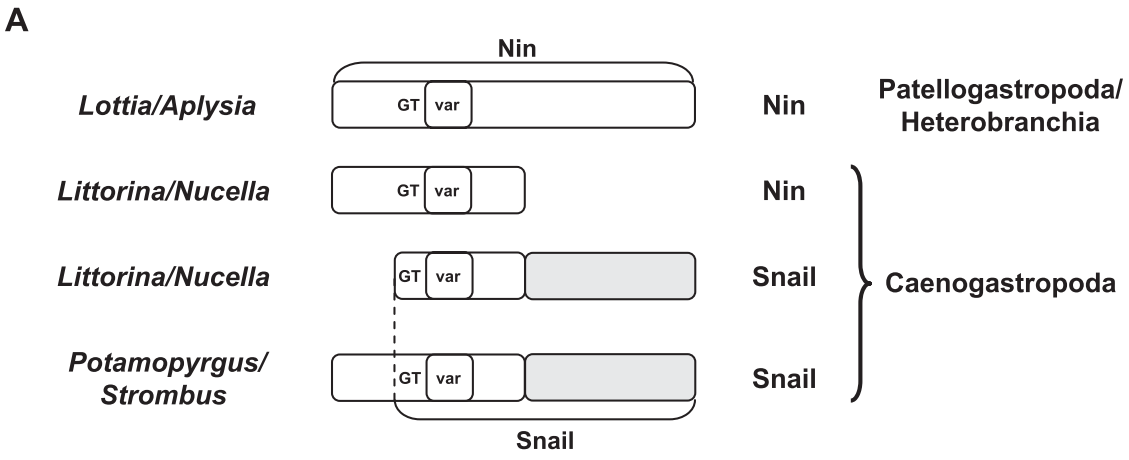


Fig. 7.—The Snail domain is conserved in SINEs in Caenogastropoda. (*A*) A central domain is shared by four SINE families in *Littorina saxatilis* (Lsa_3 to Lsa_6). Note that the heads/5′ part of the body and the 3′ parts of the body/tails are reciprocally combined with the central domain. The heads of Lsa_3 and Lsa_4 are derived from tRNA-Thr. 3′ parts of the body/tails in Lsa_3 and Lsa_5 are highlighted in black to illustrate the reciprocal exchange. A and B RNA polymerase III promoter boxes are boxed. Identity shading in the head regions is relative to the tRNA sequences. Hsa_tRNA-Thr, human tRNA-Thr; Aca_tRNA-Thr, *Aplysia californica* tRNA-Thr (scaffold_1753:118827–118902). Lsa_6 sequences used for consensus generation were obtained by PCR using internal primers. (*B*) The Snail domain is found in SINE families in Littorinimorpha (Lsa, *Littorina saxatilis*; Pan, *Potamopyrgus antipodarum*; Sgi, *Strombus gigas*) and Neogastropoda (Nla, *Nucella lapillus*). Complete consensus sequences for the *Potamopyrgus*, *Strombus*, and *Nucella* SINE families are provided in supplementary figs. S6–S8, Supplementary Material online.

identical 3′ parts of the body/tails (Lsa_3 and Lsa_5) linked by a common central domain (fig. 7*A*). Subsequent PCR amplification revealed that the fourth possible combination of 5′ and 3′ ends (head/5′ part of the body identical with Lsa_5 and 3′ part of the body/tail identical with Lsa_4) is also present in the *Li. saxatilis* genome (Lsa_6; fig. 7*A*).

The SINE head found in Lsa_3 and Lsa_4 is tRNA derived. No corresponding tRNA could be identified in the *Cr. gigas* genome. The sequence shows, however, a good match to human tRNA-Thr. Homology search using this tRNA retrieved five potential tRNAs clustered on *Ap. californica* scaffold 1753 (a representative sequence from the cluster is shown in fig. 7*A*). The head of Lsa_5 and Lsa_6 SINEs can be folded into a tRNA-like structure using tRNA scan (Schattner et al. 2005) (data not shown). However, no corresponding mollusk tRNAs could be identified. Interestingly, the tRNA-derived head regions of Lsa_3 to Lsa_6 SINEs are not directly fused to the central shared domain (see below), but separated from it by an approximately 50-bp "linker" sequence of unknown origin. The linker segments are specific for the respective heads of the elements (fig. 7*A*).

In contrast to the MESC domain described above, the central conserved domain of Lsa_3 to Lsa_6 SINEs appears to be more restricted in its species distribution. It could not be found in the available bivalve sequences and is also absent from the two sequenced (and online available) gastropod genomes of *Ap. californica* (California sea slug, Heterobranchia, sequence available at http://genome.ucsc.edu, last accessed January 6, 2016) and *Lo. gigantea* (owl limpet, Patellogastropoda; Simakov et al. 2013). Analysis of sequences retrieved by homology search suggests that SINEs containing this core domain are present in the genomes of the Atlantic dogwinkle



B

```
Lottia_Nin  AGTCGTTCGGATGGGACAATAAACCGAGGCCCCGT--GTGCTACAGGCT---------------GTGCACGTTAAAGATCCCTCGACAGT
Nla_Nin     AGTCTTTCGGATGAGACGATAAACCGAGGTCCCGT--GTGCAACACGCATTCA-----------GCGCACGTAAAAGAACCCACGGCA-A
Lsa_Nin_1   AGTCTTTCGGATGAGACGAAAAACCGAGGTCCCTTC-GTG-TACACTACATTGGGGT-------GTGCACGTTAAAGATCCCACGATTGA
Lsa_Nin_2   AGTCTTTCGGATGAGACGAAAAACCGAGGTCCCTTC-GTG-TACACTACATTGGGGT-------GTGCACGTTAAAGATCCCACGATTGA
Pan_1       AGTCTTTCGGATGAGACTCAAAACCGAGGTCCAGTGTGTG-TACA--CATGCAGAAGATCATGCACGCACTATAAAGATCCTGTAATCCA
Pan_2       ----CTTCGGATGAGACGTTAAACCGAGGTCTAGTGTGTG-TACG--CATGCAGAAGATCATGCACGCACTRTAAAGATCCTGTAATCCA
Pan_3       TGTCCTTCGGATGAGACGTTAAACCGAGGTCTAGTGTGTG-TACG--CATGCAGAAGATCATGCACGCACTATAAAGATCCTGTAATCCA
Sgi_1       -CTCTTTCGGATGAGAGTATAAACCGAGGTCTAGTCTGTG-CACA--CATGCATTTCATC--GCACGGACTCAAAAGATCCTGACGTTCA
Lsa_3       ------------------------------CGT--GTG-TACACGCAAGCACAAGACCAAGTGCGCACGGAAAAGATCCTGTAATCCA
Lsa_5       ------------------------------CGT--GTG-TACACGCAAGCACAAGACCAAGTGCGCACGAAAAAGATCCTGTAATCCA
Nla_1       ------------------------------CGT--GTG-TACACGCATGCAGAAGATCAAGTACGCACGTTAAAGATCCCGTAATCCA
Nla_2       ------------------------------CGT--GTG-TACACGCATGCAGAAGATCAAGTACGCACGTTAAAGATCCCGTAATCCA
Nla_3       ------------------------------CGT--GTG-TACACGCATGCAGAAGATCAAGTACGCACGTTAAAGATCCCGTAATCCA

Lottia_Nin  TGGAAATCGAAAAAGAGCAGGCTAATGCCGCTGTCCTGGCAAAATT
Nla_Nin     CAAGAGAGTTGTCCCTGGCAAAATTCTGTAGAAAACCCACCCTAAT
Lsa_Nin_1   CAAAAGGGTCTTTCCTGGCAAAATTGTATAGGCATAGATAAAAATG
Lsa_Nin_2   CAAAAGGGTCTTTCCTGGCAAAATTGTATAGGCATAGATAAAAATG
Pan_1       TGTCAGAGTTCGGTGGGTTACAGAAACACGAAAATATCCAGCATGC
Pan_2       TGTCAGAGTTCGGTGGGTTGCAGAGACACCAAAATACCCAGCATGC
Pan_3       TGTCAGAGTTCGGTGGGTTGCAGAWACACCAAAATACCCAGCATGC
Sgi_1       TGTCCTAGACGGGTGAATGCCGGCAACAAAAACACACCCAGCACGC
Lsa_3       TGTCAGAGTTCGGTGGGTTATAGAAACACGAAAATACCCAGCATGC
Lsa_5       TGTCAGAGTTCGGTGGGTTATAGAAACACAAAAATACCCAGCATGC
Nla_1       TGTCAGCGTTCGGTGGGTTGCGGAAACATGAACATACCCAGCATGC
Nla_2       TGTCAGCGTTCGGTGGGTTGCGGAAACATGAAGATACCCAGCATGC
Nla_3       TGTCAGCGTTCGGTGGGTTGCGGAAACACGAAAATACCCAGCATGC
```

Fig. 8.—Relationship between the Nin-DC and Snail domains. (*A*) Schematic representation of the Nin-DC and Snail domains found in gastropod SINEs. The 3′ part of the Snail domain (highlighted in gray) is unrelated to Nin-DC. Snail SINEs in *Littorina* and *Nucella* lack the 5′ part of the Nin-homology. GT denotes a potentially recombinogenic tract of 2–3 GT dinucleotides. Var refers to a variable region interrupting homology (boxed in B). (*B*) Multiple alignment of the Nin-DC and Snail domains of caenogastropod SINE families to the *Lottia gigantea* Nin-DC domain (Lottia_Nin; Piskurek and Jackson 2011). Identity shading in the 5′ part is relative to the *Lo. gigantea* sequence. The variable region interrupting homology is boxed. Identity in the 3′ Nin-DC-unrelated part of the Snail domain is indicated by black shading. Complete consensus sequences for the *Potamopyrgus*, *Strombus*, and *Nucella* SINE families are provided in supplementary figs. S6–S8, Supplementary Material online. Lsa, *Littorina saxatilis*; Pan, *Potamopyrgus antipodarum*; Sgi, *Strombus gigas*; Nla, *Nucella lapillus*.

(*Nucella lapillus*, Neogastropoda), of the New Zealand mud snail (*Potamopyrgus antipodarum*, Littorinimorpha), and of the queen conch (*St. gigas*, Littorinimorpha). However, TSDs could not be identified for the potential SINE elements in these three species, due to insufficient sequence/assembly quality and/or availability of flanking sequences. Figure 7B shows an alignment of the central domain across all four species (*Li. saxatilis*, *Nu. lapillus*, *P. antipodarum,* and *St. gigas*). Based on its limited species distribution, we decided to refer to this domain as Snail domain. Alignments of the consensus sequences of the entire elements found in *Nu. lapillus*, *P. antipodarum,* and *St. gigas* to Lsa_3 is provided in supplementary figures S6, S7, and S8, Supplementary Material online. Elements of the three *Nucella* SINE families have the head, "linker," and core domain in common with Lsa_3 (supplementary fig. S6, Supplementary Material online). Potential

Fig. 9.—*Littorina saxatilis* SINEs sharing the Lsa_5 5′ end. Identity to Lsa_5 is indicated by black and gray shading in the 5′ part and Snail domain, respectively. The Lsa_5 Snail domain is marked with a broken line on top of the sequences.

SINEs in the mud snail and queen conch share the core domain only with Lsa_3 (supplementary figs. S7 and S8, Supplementary Material online).

Interestingly, the sequence directly upstream of the Snail domain (in *P. antipodarum* and *St. gigas*) and its 5′ part (in all Snail SINEs) exhibits homology to the Nin-DC domain (fig. 8). Homology is interrupted by a 15- to 20-bp sequence whose position coincides with that of a variable region in the alignment of Nin-DC domains across distantly related species (Piskurek and Jackson 2011). The 3′ part of the Snail domain is not related to Nin-DC. In an attempt to, possibly, shed more light on the evolution of the Snail domain from the Nin-DC domain, we set out to identify "true" Nin-DC SINEs in the species where Snail SINEs are present. In the available transcriptome sequences of *P. antipodarum* and *St. gigas* (supplementary table S2, Supplementary Material online), no such elements could be identified using the Nin-DC domain as a query. From *Li. saxatilis* and *Nu. lapillus* sequences representing Nin-DC SINEs could be recovered (supplementary fig. S9 and table S2, Supplementary Material online). Homology of their central domains to Nin-DC extends over the 5′ part of the domain only. The 3′ end point of homology coincides with that found in Snail SINEs and in *Ne. vectensis* Nin-DC SINEs (Piskurek and Jackson 2011). However, the sequence interrupting homology to Nin-DC (see above) clearly differs in Nin-DC SINEs and in the (5′ Nin-truncated) Snail SINEs in *Littorina* and *Nucella* (fig. 8). Based on this finding we conclude that the Nin-DC SINEs currently found in the two species are not the direct precursors of their Snail SINEs. The two types of SINEs have most likely evolved in parallel for a longer period of time after a common ancestor acquired the 3′ part of the Snail domain. The replacement of the 5′ part of the Nin-homology in *Littorina* and *Nucella* Snail SINEs occurred at a TG tract. The recombinogenic potential of such TG tracts has been referred to in explaining the acquisition of different tails by V-SINEs (Ogiwara et al. 2002).

Surprisingly, the distribution of Snail SINEs with and without 5′ truncation of the Nin-homology does not match the currently accepted phylogenetic relationships between their host species. All four species in which Snail SINEs are found belong to the clades Caenogastropoda/Hypsogastropoda. Hypsogastropoda are further subdivided into Littorinimorpha and Neogastropoda. The two species displaying Nin-DC 5′ complete Snail SINEs (*P. antipodarum*, Hydrobiidae and *St. gigas*, Strombidae) are Littorinimorpha. Snail SINEs lacking the 5′ part of the Nin-homology are found in a neogastropod (*Nu. lapillus*, Muricidae) and a littorinimorph (*Li. saxatilis*, Littorinidae). In this context, it is worthwhile noticing that at least one study has recovered *Li. saxatilis* within Neogastropoda, albeit with low support (Cunha et al. 2009). In general, a number of molecular studies could not provide support for the monophyly of Neogastropoda, although the latest and—allegedly—most comprehensive one did so (Zou et al. 2011, and references therein). It will be interesting to see how the distribution of the two different types of Snail SINEs fits the phylogeny established using other markers—once sequence information for a larger number of Hypsogastropoda will be available.

The tails of Lsa_3 to Lsa_6 could not be traced in any other genome. Neither are there any potential LINE sequences terminating in homologous sequences present in the available *Littorina* BAC sequences. Homology search using the Lsa_5/

Fig. 10.—CORE SINEs in gastropod and bivalve mollusks. (*A*) Comparison of *Aplysia californica* (Gastropoda, Heterobranchia—Aca), *Crassostrea gigas* (Bivalvia, Pteriomorpha—Cgi) and *Crepidula fornicata* (Gastropoda, Caenogastropoda—Cfo) CORE SINEs with the CORE consensus sequence (Vassetzky and Kramerov 2013). The head regions of the two *Aplysia* SINE families, of Cgi_CORE_1, and of Cfo_CORE are related to tRNA-Thr (compare fig. 7). SINE sequences shown represent consensus sequences. Identity shading is relative to the topmost sequences. (*B*) The tails of Cgi_CORE1, Cgi_CORE_2, and Aca_CORE_1 show homology to the 3′ ends of an L2 clade LINE (L2-52_DR; Jurka et al. 2005) and a SINE mobilized by an L2 element (AFC; Takahashi et al. 1998).

Lsa_6 head as a query identified two more SINE families in *Li. saxatilis*: Lsa_7 and Lsa_8 (fig. 9).

CORE SINEs in Gastropods and Bivalves

The Lsa_3/Lsa_4 head matches a total of four SINE families in the databases. Three of these are found in *Ap. californica* (Gastropoda, Heterobranchia), one in the genome of *Cr. gigas* (Bivalvia, Pteriomorpha). One of the matches in *Aplysia* is the previously identified and characterized *Aca*-Nin-DC-SINE (Piskurek and Jackson 2011). Interestingly, homology between the remaining two *Aplysia* families (Aca_CORE_1 and Aca_CORE_2) and the one identified in *Crassostrea* (Cgi_CORE_1) extends downstream of the tRNA-derived head and covers a 74-bp sequence with 65–70% identity to the CORE domain (Gilbert and Labuda 1999). This domain is also shared by another *Cr. gigas* SINE family (Cgi_CORE_2; fig. 10*A*). The partner LINE of Aca_CORE_1 and the Cgi_CORE SINEs most likely belongs to the L2 clade, as their tails match those of an L2 element from zebrafish and of a SINE mobilized by L2 (fig. 10*B*).

Copy numbers of CORE SINEs differ significantly between *Cr. gigas* and *Ap. californica*. In the *Cr. gigas* genome, Repeatmasker (Smit et al.) analysis identified 47 and 81 copies of Cgi_CORE_1 and Cgi_CORE_2, respectively (supplementary table S3, Supplementary Material online). In case of Cgi_CORE_1, four subfamilies are clearly distinguishable. The members of the subfamilies are on average 97.5% identical to their respective subfamily consensus sequences (supplementary fig. S10, Supplementary Material online). Full-length and nearly full-length (5′ truncation $\leq$ 20 bp) Cgi_CORE_2 elements are on average 85% identical to the consensus.

Both Aca_CORE_1 and Aca_CORE_2 are represented by a higher number of copies in the *Aplysia* genome (supplementary table S3, Supplementary Material online). A total of 9,951 3′ complete Aca_CORE_1 elements harboring at least the CORE domain were identified by Repeatmasker (Smit et al.). In case of Aca_CORE_2, 6,375 elements with these characteristics were found. Divergence from the consensus (by RepeatMasker) is around 10% for both *Aplysia* CORE SINE families. Subfamily structure is discernible in both Aca_CORE_1 and Aca_CORE_2—given the large number of elements a detailed analysis is, however, beyond the scope of this study.

Finally, employing homology search with the Aca_CORE_1 CORE domain as query, a family of CORE SINEs could also be identified in a caenogastropod, *Crepidula fornicata* (Littorinimorpha; fig. 10*A*). Elements of the family lack well-defined 3′ terminal tandem repeats. TSDs were found for 4 of 34 elements analyzed.

CORE SINEs have been described in a wide range of species (for a comprehensive list see SINE Base at http://sines.eimb.ru, last accessed January 6, 2016), mostly vertebrates. Outside vertebrates CORE SINEs have been reported in amphioxus (BfISINE1; Nishihara et al. 2006), sea squirt (Cisc-1; Simmen and Bird 2000), sea urchin (SP-4 [Nisson et al. 1988], SP-5 [Carpenter et al. 1982], and SP-6 [Kapitonov and Jurka 2005a]) and Octopus (Mollusca, Cephalopoda, OR2; Ohshima and Okada 1994).

The results of our analysis now add two more mollusk classes (Gastropoda—Heterobranchia/Caenogastropoda—and Bivalvia) to the distribution range of CORE SINEs. Interestingly, CORE SINEs are the only superfamily shared between Heterobranchia and Caenogastropoda among the SINE superfamilies identified in this study. Against the background that our analysis recovered a considerable number of SINEs shared between Caenogastropoda and Bivalvia—which are more distantly related to Caenogastropoda than Heterobranchia—the finding that only one of the superfamilies identified is shared between the two gastropod clades is remarkable. Sequence representation in the databases is unlikely to be the reason—the nucleotide section of GenBank contains more than three times the number of sequences from Heterobranchia (excluding *Aplysia*) than from Caenogastropoda (722707 vs. 295888). Nine hundred short-read archives are available for Heterobranchia, 455 for Caenogastropoda. Possibly, in the course of evolution different sets of SINEs have been favored in Heterobranchia when compared with Caenogastropoda (and Bivalvia). It will be interesting to see whether the future analysis of additional genomes and transcriptomes confirms this hypothesis.

Supplementary Material

Supplementary figures S1–S10 and tables S1–S3 are available at *Genome Biology and Evolution* online (http://www.gbe.oxfordjournals.org/).

Acknowledgments

The authors thank Joshua Shallom, Eric Hallerman, Timothy King, and collaborators for making *Elliptio complanata* sequences available before publication. We thank the Broad Institute Genomics Platform and Kerstin Lindblad-Toh for making the sequences for *Aplysia californica* available. We would also like to thank the anonymous reviewers of an earlier version of the manuscript for helpful comments. This work was supported by a grant of the Ministry of National Education, CNCS – UEFISCDI, project number PN-II-ID-PCE-2012-4-0090 (to A.D.) and the strategic grant POSDRU/187/1.5/S/155383 (to P.B.).

Literature Cited

Akasaki T, et al. 2010. Characterization of a novel SINE superfamily from invertebrates: "Ceph-SINEs" from the genomes of squids and cuttlefish. Gene 454:8–19.

Albertin CB, et al. 2015. The octopus genome and the evolution of cephalopod neural and morphological novelties. Nature 524:220–224.

Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ. 1990. Basic local alignment search tool. J Mol Biol. 215:403–410.

Bao W, Jurka J. 2013. Non-LTR retrotransposons from the Pacific oyster genome. Repbase Rep. 13:1473.

Beaudoing E, Freier S, Wyatt JR, Claverie JM, Gautheret D. 2000. Patterns of variant polyadenylation signal usage in human genes. Genome Res. 10:1001–1010.

Bogenhagen DF, Brown DD. 1981. Nucleotide sequences in Xenopus 5S DNA required for transcription termination. Cell 24:261–270.

Borodulina OR, Kramerov DA. 2001. Short interspersed elements (SINEs) from insectivores. Two classes of mammalian SINEs distinguished by A-rich tail structure. Mamm Genome. 12:779–786.

Braglia P, Percudani R, Dieci G. 2005. Sequence context effects on oligo(dT) termination signal recognition by *Saccharomyces cerevisiae* RNA polymerase III. J Biol Chem. 280:19551–19562.

Bringaud F, et al. 2002. Identification of non-autonomous non-LTR retrotransposons in the genome of *Trypanosoma cruzi*. Mol Biochem Parasitol. 124:73–78.

Buzdin A, et al. 2003. The human genome contains many types of chimeric retrogenes generated through in vivo RNA recombination. Nucleic Acids Res. 31:4385–4390.

Carpenter CD, Bruskin AM, Spain LM, Eldon ED, Klein WH. 1982. The 3′ untranslated regions of two related mRNAs contain an element highly repeated in the sea urchin genome. Nucleic Acids Res. 10:7829–7842.

Churakov G, Smit AF, Brosius J, Schmitz J. 2005. A novel abundant family of retroposed elements (DAS-SINEs) in the nine-banded armadillo (*Dasypus novemcinctus*). Mol Biol Evol. 22:886–893.

Cunha RL, Grande C, Zardoya R. 2009. Neogastropod phylogenetic relationships based on entire mitochondrial genomes. BMC Evol Biol. 9:210.

Daniels GR, Deininger PL. 1991. Characterization of a third major SINE family of repetitive sequences in the galago genome. Nucleic Acids Res. 19:1649–1656.

Edgar RC, Myers EW. 2005. PILER: identification and classification of genomic repeats. Bioinformatics 21(Suppl 1):i152–i158.

Gilbert N, Labuda D. 1999. CORE-SINEs: eukaryotic short interspersed retroposing elements with common sequence motifs. Proc Natl Acad Sci U S A. 96:2869–2874.

Gilbert N, Labuda D. 2000. Evolutionary inventions and continuity of CORE-SINEs in mammals. J Mol Biol. 298:365–377.

Gonzalez VL, et al. 2015. A phylogenetic backbone for Bivalvia: an RNA-seq approach. Proc Biol Sci. 282:20142332.

Heras SR, Lopez MC, Olivares M, Thomas MC. 2007. The L1Tc non-LTR retrotransposon of *Trypanosoma cruzi* contains an internal RNA-pol II-dependent promoter that strongly activates gene transcription and generates unspliced transcripts. Nucleic Acids Res. 35:2199–2214.
Hu H, Bandyopadhyay PK, Olivera BM, Yandell M. 2011. Characterization of the *Conus bullatus* genome and its venom-duct transcriptome. BMC Genomics 12:60.
Jurka J. 2012. Non-LTR retrotransposons from the Pacific oyster genome. Repbase Rep. 12:2711.
Jurka J, et al. 2005. Repbase Update, a database of eukaryotic repetitive elements. Cytogenet Genome Res. 110:462–467.
Kajikawa M, Okada N. 2002. LINEs mobilize SINEs in the eel through a shared 3′ sequence. Cell 111:433–444.
Kapitonov VV, Jurka J. 2003. A novel class of SINE elements derived from 5S rRNA. Mol Biol Evol. 20:694–702.
Kapitonov VV, Jurka J. 2005a. SINE2-6_SP, a family of SINE2 retrotransposons in the sea urchin genome. Repbase Rep. 5:164.
Kapitonov VV, Jurka J. 2005b. SINE2-1_SP, a family of SINE2 retrotransposons in the sea urchin genome. Repbase Rep. 5:95.
Kapitonov VV, Jurka J. 2005c. SINE2-2_SP, a family of SINE2 retrotransposons in the sea urchin genome. Repbase Rep. 5:96.
Kapitonov VV, Jurka J. 2009. Nimb—a novel clade of animal non-LTR retrotransposons. Repbase Rep. 9:1535.
Khan H, Smit A, Boissinot S. 2006. Molecular evolution and tempo of amplification of human LINE-1 retrotransposons since the origin of primates. Genome Res. 16:78–87.
Kohany O, Gentles A, Hankus L, Jurka J. 2006. Annotation, submission and screening of repetitive elements in Repbase: RepbaseSubmitter and Censor. BMC Bioinformatics 7:474.
Kojima KK. 2015. A new class of SINEs with snRNA gene-derived heads. Genome Biol Evol. 7:1702–1712.
Kramerov DA, Vassetzky NS. 2011. Origin and evolution of SINEs in eukaryotic genomes. Heredity (Edinb). 107:487–495.
Longo MS, Brown JD, Zhang C, O'Neill MJ, O'Neill RJ. 2015. Identification of a recently active mammalian SINE derived from ribosomal RNA. Genome Biol Evol. 7:775–788.
Lorenz R, et al. 2011. ViennaRNA Package 2.0. Algorithms Mol Biol. 6:26.
Luchetti A, Mantovani B. 2013. Conserved domains and SINE diversity during animal evolution. Genomics 102:296–300.
McInerney CE, Allcock AL, Johnson MP, Bailie DA, Prodohl PA. 2011. Comparative genomic analysis reveals species-dependent complexities that explain difficulties with microsatellite marker development in molluscs. Heredity (Edinb). 106:78–87.
Munemasa M, et al. 2008. Newly discovered young CORE-SINEs in marsupial genomes. Gene 407:176–185.
Nishihara H, Smit AF, Okada N. 2006. Functional noncoding sequences derived from SINEs in the mammalian genome. Genome Res. 16:864–874.
Nisson PE, Hickey RJ, Boshar MF, Crain WR Jr. 1988. Identification of a repeated sequence in the genome of the sea urchin which is transcribed by RNA polymerase III and contains the features of a retroposon. Nucleic Acids Res. 16:1431–1452.
Ogiwara I, Miya M, Ohshima K, 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, Okada N. 1994. Generality of the tRNA origin of short interspersed repetitive elements (SINEs). Characterization of three different tRNA-derived retroposons in the octopus. J Mol Biol. 243:25–37.
Ohshima K, Okada N. 2005. SINEs and LINEs: symbionts of eukaryotic genomes with a common tail. Cytogenet Genome Res. 110:475–490.
Okada N. 1991. SINEs. Curr Opin Genet Dev. 1:498–504.
Okada N, Hamada M, Ogiwara I, Ohshima K. 1997. SINEs and LINEs share common 3′ sequences: a review. Gene 205:229–243.
Piskurek O, Jackson DJ. 2011. Tracking the ancestry of a deeply conserved eumetazoan SINE domain. Mol Biol Evol. 28:2727–2730.
Piskurek O, Nikaido M, Boeadi, Baba M, Okada N. 2003. Unique mammalian tRNA-derived repetitive elements in dermopterans: the t-SINE family and its retrotransposition through multiple sources. Mol Biol Evol. 20:1659–1668.
Putnam NH, et al. 2007. Sea anemone genome reveals ancestral eumetazoan gene repertoire and genomic organization. Science 317:86–94.
Sanchez-Luque FJ, Lopez MC, Macias F, Alonso C, Thomas MC. 2011. Identification of an hepatitis delta virus-like ribozyme at the mRNA 5′-end of the L1Tc retrotransposon from *Trypanosoma cruzi*. Nucleic Acids Res. 39:8065–8077.
Santangelo AM, et al. 2007. Ancient exaptation of a CORE-SINE retroposon into a highly conserved mammalian neuronal enhancer of the proopiomelanocortin gene. PLoS Genet. 3:1813–1826.
Sasaki T, et al. 2008. Possible involvement of SINEs in mammalian-specific brain formation. Proc Natl Acad Sci U S A. 105:4220–4225.
Schattner P, Brooks AN, Lowe TM. 2005. The tRNAscan-SE, snoscan and snoGPS web servers for the detection of tRNAs and snoRNAs. Nucleic Acids Res. 33:W686–W689.
Schmitz J, Zischler H. 2003. A novel family of tRNA-derived SINEs in the colugo and two new retrotransposable markers separating dermopterans from primates. Mol Phylogenet Evol. 28:341–349.
Simakov O, et al. 2013. Insights into bilaterian evolution from three spiralian genomes. Nature 493:526–531.
Simmen MW, Bird A. 2000. Sequence analysis of transposable elements in the sea squirt, *Ciona intestinalis*. Mol Biol Evol. 17:1685–1694.
Smit AF, Hubley R, Green P. RepeatMasker. Available from: http://repeatmasker.org.
Sun FJ, Fleurdepine S, Bousquet-Antonelli C, Caetano-Anolles G, Deragon JM. 2007. Common evolutionary trends for SINE RNA structures. Trends Genet. 23:26–33.
Takahashi K, Terai Y, Nishida M, Okada N. 1998. A novel family of short interspersed repetitive elements (SINEs) from cichlids: the patterns of insertion of SINEs at orthologous loci support the proposed monophyly of four major groups of cichlid fishes in Lake Tanganyika. Mol Biol Evol. 15:391–407.
Takeuchi T, et al. 2012. Draft genome of the pearl oyster *Pinctada fucata*: a platform for understanding bivalve biology. DNA Res. 19:117–130.
Ullu E, Tschudi C. 1984. *Alu* sequences are processed 7SL RNA genes. Nature 312:171–172.
Vassetzky NS, Kramerov DA. 2013. SINEBase: a database and tool for SINE analysis. Nucleic Acids Res. 41:D83–D89.
Wood HM, Grahame JW, Humphray S, Rogers J, Butlin RK. 2008. Sequence differentiation in regions identified by a genome scan for local adaptation. Mol Ecol. 17:3123–3135.
Yoshida MA, et al. 2011. Genome structure analysis of molluscs revealed whole genome duplication and lineage specific repeat variation. Gene 483:63–71.
Zhang G, et al. 2012. The oyster genome reveals stress adaptation and complexity of shell formation. Nature 490:49–54.
Zou S, Li Q, Kong L. 2011. Additional gene data and increased sampling give new insights into the phylogenetic relationships of Neogastropoda, within the caenogastropod phylogenetic framework. Mol Phylogenet Evol. 61:425–435.

Associate editor: Mar Alba