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

Irina Matetovic1,2, Szilard Sajgo1,3, Bianca Ianc1, Cornelia Ochis1, Paul Bulzu1, Octavian Popescu4, and Annette Damert1,*

1Institute for Interdisciplinary Research in Bio-Nano-Sciences, Molecular Biology Center, Babes-Bolyai-University, Cluj-Napoca, Romania
2Present address: Institute of Tropical Medicine, Unit of Veterinary Protozoology, Antwerpen, Belgium
3Present address: Danish Research Institute of Translational Neuroscience, Nordic EMBL Partnership for Molecular Medicine, DANDRITE, Aarhus University, Aarhus, Denmark
4Institute 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 MEI—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 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). 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 MEIc—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.
Cr. gigas is the only mollusk species for which tRNA genes obtained from Repbase (Jurka et al. 2005) and subsequently explained at first occurrence in line 49 retrotransposons were retrieved region) sequences of autonomous non-LTR is now explained at first occurrence 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’-TAGCTCAGTCCGGTAGCGCGC-3’ in combination with either 3’REV 5’-GGTTACGTGTGAGGTCGTTGCC C-3’ (Lsa_6) or Lsa_7_REV 5’-GGAAAGCCGAGCTGCAATGC-3’ (Lsa_7). PCR products were cloned 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.ensemble.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 sequence (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 Patellolagastropoda, 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 vettigastropod *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 (CACC)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*

---

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).

### Table 1
Summary of the SINE Families Identified in the Study

<table>
<thead>
<tr>
<th>SINE Superfamily</th>
<th>Class_subclass/clade</th>
<th>Species</th>
<th>SINE family</th>
<th>Sequence Source</th>
</tr>
</thead>
<tbody>
<tr>
<td>MESC</td>
<td>Gp_Caenogastropoda</td>
<td>Littorina saxatilis</td>
<td>Lsa_1</td>
<td>Genomic</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Littorina littorea</td>
<td>Li</td>
<td>Transcriptome</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Crepidula fornicata</td>
<td>Cfo</td>
<td>Transcriptome</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Strombus gigas</td>
<td>Sgi</td>
<td>Transcriptome</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Arctica islandica</td>
<td>Ais</td>
<td>Transcriptome</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Bathymodiolus azoricus</td>
<td>Baz</td>
<td>Transcriptome</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Mercenaria mercenaria</td>
<td>Mme</td>
<td>Genomic</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Sinonovacula constricta</td>
<td>Sco</td>
<td>Transcriptome</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Nucula tenuis</td>
<td>Nte</td>
<td>Transcriptome</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Littorina littorea</td>
<td>Lli</td>
<td>Transcriptome</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Littorina saxatilis</td>
<td>Lsa_2</td>
<td>Genomic</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Crepidula fornicata</td>
<td>Cfo</td>
<td>Transcriptome</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Strombus gigas</td>
<td>Sgi</td>
<td>Transcriptome</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Bathymodiolus azoricus</td>
<td>Baz</td>
<td>Transcriptome</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Mercenaria mercenaria</td>
<td>Mme</td>
<td>Genomic</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Sinonovacula constricta</td>
<td>Sco</td>
<td>Transcriptome</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Nucula tenuis</td>
<td>Nte</td>
<td>Transcriptome</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Littorina saxatilis</td>
<td>Lsa_1</td>
<td>Genomic</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Crepidula fornicata</td>
<td>Cfo</td>
<td>Transcriptome</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Strombus gigas</td>
<td>Sgi</td>
<td>Transcriptome</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Bathymodiolus azoricus</td>
<td>Baz</td>
<td>Transcriptome</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Mercenaria mercenaria</td>
<td>Mme</td>
<td>Genomic</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Sinonovacula constricta</td>
<td>Sco</td>
<td>Transcriptome</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Littorina saxatilis</td>
<td>Lsa_2</td>
<td>Genomic</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Crepidula fornicata</td>
<td>Cfo</td>
<td>Transcriptome</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Strombus gigas</td>
<td>Sgi</td>
<td>Transcriptome</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Bathymodiolus azoricus</td>
<td>Baz</td>
<td>Transcriptome</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Mercenaria mercenaria</td>
<td>Mme</td>
<td>Genomic</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Sinonovacula constricta</td>
<td>Sco</td>
<td>Transcriptome</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Littorina saxatilis</td>
<td>Lsa_2</td>
<td>Genomic</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Crepidula fornicata</td>
<td>Cfo</td>
<td>Transcriptome</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Strombus gigas</td>
<td>Sgi</td>
<td>Transcriptome</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Bathymodiolus azoricus</td>
<td>Baz</td>
<td>Transcriptome</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Mercenaria mercenaria</td>
<td>Mme</td>
<td>Genomic</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Sinonovacula constricta</td>
<td>Sco</td>
<td>Transcriptome</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Littorina saxatilis</td>
<td>Lsa_2</td>
<td>Genomic</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Crepidula fornicata</td>
<td>Cfo</td>
<td>Transcriptome</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Strombus gigas</td>
<td>Sgi</td>
<td>Transcriptome</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Bathymodiolus azoricus</td>
<td>Baz</td>
<td>Transcriptome</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Mercenaria mercenaria</td>
<td>Mme</td>
<td>Genomic</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Sinonovacula constricta</td>
<td>Sco</td>
<td>Transcriptome</td>
</tr>
</tbody>
</table>

Note.—na, not applicable.

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

Gp, Gastropoda; Bv, Bivalvia.

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.

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.
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 region 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. 3B).

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...
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 identified in the BAC sequences. Six of these display variable size TSDs (supplementary table S1, Supplementary Material online).

In the Sydney rock oyster (*Saccostrea glomerata*, Ostreidae), a SINE family with the LINE-derived head shares the body and tail with two families of tRNA-derived SINES (fig. 3B). 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. 3A), 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 3A). 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. 3A) represent polyA signal variants that were shown to be functional in humans (Beaudouing 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, Haliotis diversicolor/Haliotis midae) and bivalves (Saccostrea glomerata, Ostrea edulis/Ostrea conchaphila) and the 5'0 part of the 5'0 UTR of Nimb-17_LMi, a Nimb-clade LINE element from the migratory locust. 

Mobile Element Evolution Playing Jigsaw


Advance Access publication January 6, 2016

(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 SS 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, Halotis sp., Sa. glomerata, and the pectinoid MESC-domain SINES described above is restricted to the first approximately 70 bp (fig. 3C). 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. 3C). 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. 3C). 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) and 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, Arctica and Meretrix meretrix, Veneridae; 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. 4A). 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. 4C). 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-domain 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. 5A). 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 (Eupusca pennata, Thermobia domestica); consensus sequences for these SINE families are provided in supplementary fig. S4, Supplementary Material online, fig. 5A). We, therefore, suggest to refer to these SINES as “Vhc” (V highly conserved) SINES.

SINES of the first bivalve Vhc superfamily (Vhc_1; fig. 4A) were identified in Pteriomorpha (saltwater clams). They are found in Ostreidae (Cr. gigas), Arcidae (T. granosa), and Pectinidae (Ch. farreiri, M. yessoensis, Nodopererebubnodosus, and A. irritans). The head region of these SINES is
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

---

**A**

<table>
<thead>
<tr>
<th>Bivalve Vhc SINEs and their partner LINEs.</th>
<th><strong>t-RNA related</strong></th>
</tr>
</thead>
<tbody>
<tr>
<td>Cgi_tRNA-Glu</td>
<td>TCGAAGCTGCTGTCATAGGTTTCT</td>
</tr>
<tr>
<td>Cgi_Vhc_1</td>
<td>GGTGTATGCTGAGTCCATTCAATAGGTTTCT</td>
</tr>
<tr>
<td>Tgr_Vhc_1</td>
<td>GGTGTATGCTGAGTCCATTCAATAGGTTTCT</td>
</tr>
<tr>
<td>Cfa_Vhc_1</td>
<td>GGTGTATGCTGAGTCCATTCAATAGGTTTCT</td>
</tr>
<tr>
<td>Mye_Vhc_1</td>
<td>GGTGTATGCTGAGTCCATTCAATAGGTTTCT</td>
</tr>
<tr>
<td>Nsu_Vhc_1</td>
<td>GGTGTATGCTGAGTCCATTCAATAGGTTTCT</td>
</tr>
<tr>
<td>Air_Vhc_1</td>
<td>GGTGTATGCTGAGTCCATTCAATAGGTTTCT</td>
</tr>
<tr>
<td><strong>V-domain</strong></td>
<td><strong>t-RNA related</strong></td>
</tr>
<tr>
<td>Cgi_tRNA-Glu</td>
<td>TCGAAGCTGCTGTCATAGGTTTCT</td>
</tr>
<tr>
<td>Cgi_Vhc_1</td>
<td>GGTGTATGCTGAGTCCATTCAATAGGTTTCT</td>
</tr>
<tr>
<td>Tgr_Vhc_1</td>
<td>GGTGTATGCTGAGTCCATTCAATAGGTTTCT</td>
</tr>
<tr>
<td>Cfa_Vhc_1</td>
<td>GGTGTATGCTGAGTCCATTCAATAGGTTTCT</td>
</tr>
<tr>
<td>Mye_Vhc_1</td>
<td>GGTGTATGCTGAGTCCATTCAATAGGTTTCT</td>
</tr>
<tr>
<td>Nsu_Vhc_1</td>
<td>GGTGTATGCTGAGTCCATTCAATAGGTTTCT</td>
</tr>
<tr>
<td>Air_Vhc_1</td>
<td>GGTGTATGCTGAGTCCATTCAATAGGTTTCT</td>
</tr>
</tbody>
</table>

**V-domain**

| Cgi_tRNA-Glu | GT-CTCCAA |
| Cgi_Vhc_1   | GTGAGGTTTCTTACATAGGTTTCT |
| Tgr_Vhc_1   | GGTGTATGCTGAGTCCATTCAATAGGTTTCT |
| Cfa_Vhc_1   | GGTGTATGCTGAGTCCATTCAATAGGTTTCT |
| Mye_Vhc_1   | GGTGTATGCTGAGTCCATTCAATAGGTTTCT |
| Nsu_Vhc_1   | GGTGTATGCTGAGTCCATTCAATAGGTTTCT |
| Air_Vhc_1   | GGTGTATGCTGAGTCCATTCAATAGGTTTCT |

**B**

| CR1-1_CGI | TTTAAATTGTTAAGTACATAGGTTTCT |
| CR1-14_CGI| TTTAAATTGTTAAGTACATAGGTTTCT |
| Cgi_Vhc_1 | TTTAAATTGTTAAGTACATAGGTTTCT |
| Cfa_Vhc_1 | TTTAAATTGTTAAGTACATAGGTTTCT |
| Mye_Vhc_1 | TTTAAATTGTTAAGTACATAGGTTTCT |
| Nsu_Vhc_1 | TTTAAATTGTTAAGTACATAGGTTTCT |
| Air_Vhc_1 | TTTAAATTGTTAAGTACATAGGTTTCT |

**C**

<table>
<thead>
<tr>
<th>Bivalve Vhc SINEs and their partner LINEs.</th>
<th><strong>t-RNA related</strong></th>
</tr>
</thead>
<tbody>
<tr>
<td>Cfa_Vhc_1</td>
<td>GACCTCTGGCGCGCAGT</td>
</tr>
<tr>
<td>Mye_Vhc_1</td>
<td>GACCTCTGGCGCGCAGT</td>
</tr>
<tr>
<td>Air_Vhc_1</td>
<td>GACCTCTGGCGCGCAGT</td>
</tr>
<tr>
<td><strong>V-domain</strong></td>
<td><strong>t-RNA related</strong></td>
</tr>
<tr>
<td>Cfa_Vhc_1</td>
<td>GACCTCTGGCGCGCAGT</td>
</tr>
<tr>
<td>Mye_Vhc_1</td>
<td>GACCTCTGGCGCGCAGT</td>
</tr>
<tr>
<td>Air_Vhc_1</td>
<td>GACCTCTGGCGCGCAGT</td>
</tr>
</tbody>
</table>

**V-domain**

| Cfa_Vhc_1 | GACCTCTGGCGCGCAGT |
| Mye_Vhc_1 | GACCTCTGGCGCGCAGT |
| Air_Vhc_1 | GACCTCTGGCGCGCAGT |

---

(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 fig. S5, Supplementary Material online). On average, Cgi_Vhc_1 SINEs show 11% divergence from the consensus. However, there are elements that share the tail with Vhc_2 SINEs could not be identified in the available database sequences.

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. 4B) 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. 4C)—are absent from the *Cr. gigas* (Ostreoidea) 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 ≥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. 5A). 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; Ogawa and 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. 5B). 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. 6A) or as terminal part of a larger stem loop (in Vhc_2 SINES; fig. 6B). 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. 6A), 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. 6B). 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, *Nodoplecten subnodosus*; Mye, *Mizuhopecten yessoensis*; Air, *Argospecten irradians*. Residues identical to the fish V domain are highlighted in gray.
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 fig. S4, Supplementary Material online.

A  Vhc_1 (Lophotrochozoa, Mollusca, Bivalvia)  

B  Sgi_Vhc (Lophotrochozoa, Mollusca, Gastropoda)  

Epe_Vhc (Ecdysozoa, Arthropoda, Insecta)
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
identical 3′ parts of the body/tails (Lsa_3 and Lsa_5) linked by a common central domain (fig. 7A). 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. 7A).

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 *Aplysia californica* scaffold 1753 (a representative sequence from the cluster is shown in fig. 7A). 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. 7A).

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 *Aplysia 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...
\textit{Nucella lapillus}, Neogastropoda), of the New Zealand mud snail (\textit{Potamopyrgus antipodarum}, Littorinimorpha), and of the queen conch (\textit{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 (\textit{Li. saxatilis}, \textit{Nu. lapillus}, \textit{Pa. antipodarum}, and \textit{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 \textit{Nu. lapillus}, \textit{Pa. antipodarum}, and \textit{St. gigas} to \textit{Lsa} is provided in supplementary figures S6, S7, and S8, Supplementary Material online. Elements of the three \textit{Nucella} SINE families have the head, "linker," and core domain in common with \textit{Lsa} (supplementary fig. S6, Supplementary Material online). Potential
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) to its 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 30 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 50 part of the domain only. The 30 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 (50 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’ end 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. 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.
Aca_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 Ap.
lysa 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.
10A). 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. 10B).

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 (supple-
mentary 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 (supplemen-
tary fig. S10, Supplementary Material online). Full-length and
nearly full-length (5\(^{\text{0}}\) truncation/C2020 bp) Cgi_Core_2 ele-
ments are on average 85% identical to the consensus.

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 Ap.
lysa 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.
10A). 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. 10B).

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 (supple-
mentary 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 (supplemen-
tary fig. S10, Supplementary Material online). Full-length and
nearly full-length (5\(^{\text{0}}\) truncation \(\leq\) 20bp) 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, Crepudila fornicata (Littorinimorpha; fig. 10A). 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 et al. 2005a), and cephalopods (Aca_CORE_1 and Aca_CORE_2). The Sea urchin genome (supplementary table S3, Supplementary Material online) contains more than three times the number of 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 et al. 2005a), and cephalopods (Aca_CORE_1 and Aca_CORE_2). The Sea urchin genome (supplementary table S3, Supplementary Material online) contains more than three times the number of 3' terminal tandem repeats. TSDs were found for 4 of 34 elements analyzed.

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 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.).

Acknowledgments
The authors thank Joshua Shallom, Eric Hallemman, 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


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


Associate editor: Mar Alba