Genomics 102 (2013) 296–300

Contents lists available at ScienceDirect

Genomics

journal homepage: www.elsevier.com/locate/ygeno

Conserved domains and SINE diversity during animal evolution ☆

Andrea Luchetti *, Barbara Mantovani

Dip. Scienze Biologiche, Geologiche e Ambientali (BiGeA) – Università di Bologna, via Selmi 3, 40126 Bologna, Italy

ARTICLE INFO

Article history:
Received 2 May 2013
Accepted 14 August 2013
Available online 24 August 2013

Keywords:
Short interspersed elements (SINEs)
Conserved domains
Sequence module exchange
Retrotransposons evolution

ABSTRACT

Eukaryotic genomes harbour a number of mobile genetic elements (MGEs); moving from one genomic location to another, they are known to impact on the host genome. Short interspersed elements (SINEs) are well-represented, non-autonomous retroelements and they are likely the most diversified MGEs. In some instances, sequence domains conserved across unrelated SINEs have been identified; remarkably, one of these, called Nin, has been conserved since the Radiata–Bilateria splitting. Here we report on two new domains: Inv, derived from Nin, identified in insects and in deuterostomes, and Pln, restricted to polyneopteran insects. The identification of Inv and Pln sequences allowed us to retrieve new SINEs, two in insects and one in a hemichordate. The diverse structural combination of the different domains in different SINE families, during metazoan evolution, offers a clearer view of SINE diversity and their frequent *de novo* emergence through module exchange, possibly underlying the high evolutionary success of SINEs.

1. Introduction

Eukaryotic genomes host a high number of interspersed repeated elements able to move within and between chromosomes. Mobile genetic elements (MGEs) can be retroelements (class I), moving via a RNA intermediate with a copy-&-paste mechanism, or transposons (class II), moving via a DNA intermediate mainly through a cut-&-paste mode [1]. MGEs are generally well represented in all eukaryotic genomes studied so far and there is increasing evidence of their implication in the evolution of the host genomes [2], especially through exaptation [3,4].

Short interspersed elements (SINEs) are non-autonomous MGEs, i.e. they do not have coding capacity; therefore, they transpose through the enzymatic machinery provided by a "partner" autonomous element (a long interspersed element, LINE) [5]. SINEs are almost ubiquitous in eukaryotes and their structure is composed by a short RNA-related region (the head), an anonymous sequence (the body) and a simple sequence repeat tail (a poly(A) or a microsatellite motif). The head usually derives from a tRNA, but there are examples of 5S- or 7SL-like sequences. Moreover, it can include one, two or three RNA-related sequences, showing therefore a monomeric, a dimeric or even a trimeric structure (reviewed in [6]). The 3′ end of the body and the tail are highly similar to the partner LINE 3′ end and are, thus, involved in the parasitization of its retrotransposition machinery [5,7].

The sequence of the SINEs body is usually defined as anonymous (i.e., not related to any known functional sequence) and element-specific. Yet, in the last two decades, some exceptions have been discovered. In fact, nucleotide blocks conserved among body sequences of different SINE families occur in metazoan genomes: the CORE domain in bilaterian-specific elements, the V domain in vertebrates and the Ceph domain in cephalopods SINEs [8–10]. In deuterostome genomes, a domain named Deu has been found conserved in different SINE families [11]. Recently, a 105–119 bp fragment of the Deu domain, called Nin, has been discovered in SINE families distributed across metazoan, from cnidarian onward [12].

The evolutionary meaning of these conserved domains is still debated: they have been considered advantageous either for the SINE itself, as promoters of recombination or SINE transcript transport [6,13], or for the host genome, being often retrieved in elements inserted within or near functional genes [14,15].

The wide distribution and diversity of SINE sequences suggest they emerged *de novo* many times during evolution. This could be due to recombination either between retropseudogenes and autonomous elements, caused by reverse transcriptase template switch [11,16], or to SINE–SINE recombination [17,18]. Kramerov and Vassetzky [6] suggested a "module exchange" model accounting for the great SINE diversity observed, a mechanisms that has been thought as being at the basis of the evolution of autonomous retroelements [16,19].

In this paper we describe the evolutionary dynamics of a conserved block that we first found and described in the *Taluc* SINE family isolated in *Reticulitermes lucifugus* (Termitoidea [20]). From previous databases analyses, the 38 bp long domain appeared conserved also in the two *Taluc* elements from *Hodotermopsis sjoestedti* (Termitoidea; *Taluc*_HS)

Abbreviations: LINE, long interspersed elements; LTR, long terminal repeats; MGE, mobile genetic elements; SINE, short interspersed elements; TSD, target site duplications.

☆ Sequence data from this article have been deposited at http://www.mozoolab.net/index.php/downloads.html.

* Corresponding author at: Dip. BiGeA, via Selmi 3, 40126 Bologna, Italy. Fax: +39 0512094286.
E-mail address: andrea.luchetti@unibo.it (A. Luchetti).

0888-7543/$ – see front matter © 2013 Elsevier Inc. All rights reserved.
http://dx.doi.org/10.1016/j.ygeno.2013.08.005

and *Diploptera punctata* (Blattodea; *Taluc*_DP) and in the *Gbim* and *gecko* SINEs from *Gryllus bimaculatus* (Orthoptera [20]) and *Aedes aegypti* (Diptera [21]), respectively. This domain is located downstream the tRNA-related head and contains an inverted repeat: we therefore will refer to it as the "Inv" domain (formerly termed "alpha" domain in [22]). The distribution of the Inv domain along the metazoan phylogeny as well as its relationship with SINE conserved blocks – either already known or newly discovered (the "Pln" domain) – pinpoint to the great evolutionary diversity of SINE sequences, possibly underlying the evolutionary success of these elements.

2. Results

2.1. The Inv domain

In the Repbase database, the Inv domain occurs in the SINE3-1_TC from the flour beetle *T. castaneum* [23], a 5S rRNA-derived element, and in the SINE-3_CQ from the mosquito *C. quinquefasciatus* [24], a tRNA-derived element (Table 1).

In the NCBI EST database, the search for the Inv domain led to the identification of 64 EST sequences from the mosquito *Ochlerotatus triseriatus* carrying a new SINE element (henceforth named *Otri*); the nucleotide identity to the consensus ranges from 77.6% to 100.0% (mean identity 97.4%). *Otri* sequence analysis shows a tRNA-related head with a conserved secondary structure and an RNA pol III promoter sequence (Suppl. Fig. S1; Table 1). The body contains the Inv domain 31 bp downstream the tRNA-like head and the 3′ end presents a poly(A) tail. All retrieved EST sequences are processed SINE RNAs, so that the presence of target site duplications (TSDs) cannot not be verified.

Three SINEs sharing the Inv domain occur, therefore, in Diptera: besides *gecko* from *Aedes aegypti*, in fact, we retrieved SINE-3_CQ from *C. quinquefasciatus* and *Otri* from *O. triseriatus*. Sequence analysis evidenced a dimeric structure for SINE-3_CQ, with two tRNA-related regions separated by an anonymous spacer containing the Inv domain; on the contrary, *Otri* – as *gecko* – is monomeric (Table 1). *Otri* and the first half of SINE-3_CQ (i.e., the tRNA head plus the Inv domain) share an identity of 79.7%; on the other hand, the identity of *gecko* with *Otri* and with the first half of SINE-3_CQ drops below the 70.0%. The second half of SINE-3_CQ (the second tRNA-related region + anonymous body + microsatellite tail) does not show significant identity with any of the other mosquitoes SINEs.

Repeat masking on Repbase database allowed also to identify the occurrence of the Inv domain in SINE2-7_SP, a tRNA-derived element from the sea urchin *Strongylocentrotus purpuratus* [25] (Table 1).

A new SINE (named *Skow*) has been retrieved from the Genome RefSeq, deposited in the NCBI database, of the hemichordate acorn worm *Saccoglossus kowalevskii*. 91 copies of *Skow* SINE have been recovered, with identity to the consensus sequence ranging from 67.8% to 93.3% (mean identity 75.3%). This element is also a tRNA-related SINE with a conserved cloverleaf structure, probably derived from an Alanine tRNA, and RNA pol III boxes (Suppl. Fig. S2; Table 1). The Inv domain is located 70 bp downstream the tRNA-related head and the 3′ end shows a microsatellite-like tail. No evident target site duplications have been found.

On the whole, considering also previously identified Inv SINEs (*Taluc*, *Taluc*_HS, *Taluc*_DP, *Gbim* and *gecko*), 10 SINE families have been found sharing the Inv domain: eight in insects and two in invertebrate deuterostomes (the sea urchin and the acorn worm). Sequence analysis confirmed the Inv domain 38 bp long with an overall similarity among the 10 SINE families equal to 75.4% and the 31.6% of identical nucleotide sites (Fig. 1A). The Inv consensus sequence has been compared with the known SINE conserved domains CORE, V, Deu, Ceph and Nin. The Nin domain shows a significant similarity: in particular, Inv and Nin consensus sequences share 29 aligned nucleotide positions, located at the 5′ end of the Nin domain. On the whole, the mean identity between the homologous tracts of the two domains is equal to 75.0% (Fig. 2; Suppl. Fig. S3). The Deu domain, embodying Nin, has the same region homologous to Inv, but the sequence identity falls to 69.2%.

2.2. Pln domain

Eight out of 10 SINEs sharing the Inv domain are from insect genomes. The Inv domain identified in termites, cockroaches and crickets shows an average identity of 75.2%. Interestingly, sequence identity increases to 79.6% in the 33 nucleotidic positions downstream the Inv domain (Fig. 1B). This further conserved region has been found only in polyneopteran taxa, being absent in the other presently analysed organisms, whether insects (Diptera and Coleoptera) or not. It has been, therefore, named Pln (**Pol**y**n**eoptera) domain. Databases have been probed with the Pln domain as query sequence and a significant similarity has been found with five EST clones of the cockroach *Periplaneta americana*. These clones carry copies of a SINE sequence (henceforth named *Pame*; 90%–97% identity from the consensus, mean identity 89.0%; Suppl. Fig. S4) that lacks, however, the Inv domain (Fig. 1B).

3. Discussion

SINEs evolutionary success is likely linked to their ability to reshape themselves, giving raise to new families. In this view, the discovery and the characterization of a number of conserved domains across different SINE families, from different genomes, add to the understanding of such ability [6]. Data presented here show the existence of two new

Table 1
List of identified Inv-SINEs with host species and structural features.

SINE family	Host species[a]	Length[b]	RNA-related head	RNA Pol III promoter sequence[d]	Tail
Taluc	*Reticulitermes lucifugus*	387	tRNA-Gln	TGGACTAGTGG [—] GGTTCGATCCC	[e]
*Taluc*_HS	*Hodotermopsis sjoestedti*	389	tRNA-n.d.	TGGTGTAGTGG [—] GGTTCAAACCT	$(TCA)_n$
*Taluc*_DP	*Diploptera punctata*	401	tRNA-Gln	TGGACTAGTGG [—] GGTTCAAACCC	n.d.
Gbim	*Gryllus bimaculatus*	342	tRNA-Asn	TGATGTAACGG [—] GGTTCAAACCC	$(TTCAA)_n$
gecko	*Aedes aegypti*	188	tRNA-Glu	TGGTGTAGTGG [—] GATCGAATCCC	$(A)_n$
Otri	*Ochlerotatus triseriatus*(1)	252	tRNA-n.d.	TGGTGTAGGGG [—] GGTTCGATCCC	$(A)_n$
SINE-3_CQ	*Culex quinquefasciatus*(2)	402	tRNA-Glu[c]	TGGTGTAGGGG [—] GGTTCGATCCC	$(GATTTTT)_n$
			tRNA-Val[c]	TGGAGTCGCTGG [—] AGTTCGAATCCC	
SINE3-1_TC	*Tribolium castaneum*(2)	250	5S	AGTTAAGCAGCTCT [—] CCCGGT [—] TGGATGGGTGACCGGAT	$(ATATTT)_n$
SINE2-7_SP	*Strongylocentrotus purpuratus*(2)	199	tRNA-Thr	TAGCTCAGTCGG [—] GGTTCGAAACC	$(CCA)_n$
Skow	*Saccoglossus kowalevskii*(3)	352	tRNA-Ala	TTGCTTAGTGG [—] GGTTCAACCCC	$(TAA)_n$

[a] Reference database: (1) NCBI EST database; (2) Repbase; (3) Genome RefSeq.
[b] Length of the consensus sequence excluding the tail.
[c] tRNA-derived dimeric element.
[d] A + B boxes for tRNA-related SINEs; A + IE + C boxes for the 5S rRNA-related SINE.
[e] Multiple microsatellite-like tails [20].