

Available online at www.sciencedirect.com



GENE

Gene 407 (2008) 176-185

www.elsevier.com/locate/gene

Newly discovered young CORE-SINEs in marsupial genomes $\stackrel{\leftrightarrow}{\sim}$

Maruo Munemasa^a, Masato Nikaido^a, Hidenori Nishihara^a, Stephen Donnellan^b, Christopher C. Austin^c, Norihiro Okada^{a,*}

> ^a Graduate School of Bioscience and Biotechnology, Tokyo Institute of Technology, Yokohama, Japan ^b Evolutionary Biology Unit, South Australian Museum, North Terrace, SA 5000, Australia

^c Museum of Natural Science, Louisiana State University, LA 70803, USA

Received 3 July 2007; received in revised form 2 October 2007; accepted 4 October 2007 Available online 12 October 2007 Received by Takashi Gojobori

Abstract

Although recent mammalian genome projects have uncovered a large part of genomic component of various groups, several repetitive sequences still remain to be characterized and classified for particular groups. The short *in*terspersed repetitive *e*lements (SINEs) distributed among marsupial genomes are one example. We have identified and characterized two new SINEs from marsupial genomes that belong to the CORE-SINE family, characterized by a highly conserved "CORE" domain. PCR and genomic dot blot analyses revealed that the distribution of each SINE shows distinct patterns among the marsupial genomes, implying different timing of their retroposition during the evolution of marsupials. The members of Mar3 (*Marsupialia* 3) SINE are distributed throughout the genomes of all marsupials, whereas the Mac1 (*Macropodoidea* 1) SINE is distributed specifically in the genomes of kangaroos. Sequence alignment of the Mar3 SINEs revealed that they can be further divided into four subgroups, each of which has diagnostic nucleotides. The insertion patterns of each SINE at particular genomic loci, together with the distribution patterns of each SINE, suggest that the Mar3 SINEs have intensively amplified after the radiation of diprotodontians, whereas the Mac1 SINE has amplified only slightly after the divergence of hypsiprimnodons from other macropods. By compiling the information of CORE-SINEs characterized to date, we propose a comprehensive picture of how SINE evolution occurred in the genomes of marsupials.

Keywords: CORE-SINE; Marsupials; Phylogeny; Evolution

1. Introduction

Mammalian genomes harbor a large amount of retroposons that propagate their copies in the host genome via an RNA intermediate generated from a "copy and paste" mechanism called retroposition (Rogers, 1985; Weiner et al., 1986; Brosius, 1991; Okada, 1991a,b). Short *interspersed* repetitive *e*lements

* Corresponding author. Department of Biological Sciences, Graduate School of Bioscience and Biotechnology, Tokyo Institute of Technology, 4259-B21,

Nagatsuta-cho, Midori-ku, Yokohama 226-8501, Japan. Tel.: +81 45 924 5742; fax: +81 45 924 5835.

E-mail address: nokada@bio.titech.ac.jp (N. Okada).

0378-1119/\$ - see front matter 0 2007 Elsevier B.V. All rights reserved. doi:10.1016/j.gene.2007.10.008

(SINEs) belong to a class of retroposons that account for more than ten percent of nuclear DNA. The role of SINEs in the host genome still remains to be clarified; however, recent studies, including those from our laboratory, have found that some SINE-derived non-coding sequences are highly conserved (Nishihara et al., 2006a). This implies that these SINEs might have acquired some functionality during the evolution (Nishihara et al., 2006a; Bejerano et al., 2006; Mikkelsen et al., 2007). It may therefore be useful to characterize and categorize the genomic components of various mammals with respect to SINEs. Recent comprehensive genome sequencing projects have allowed us to investigate particular animals on the wholegenome level (e.g. Margulies et al., 2005), providing a very powerful tool for revealing a complete picture of SINE evolution. Indeed, owing to the completion of the human genome project, the contribution of SINEs to the human genome has been clarified in detail — the Alu fraction covers

Abbreviations: SINE, short interspersed repetitive element; LINE, long interspersed repetitive element; PCR, polymerase chain reaction; mya, million years ago.

 $[\]stackrel{\text{\tiny{free}}}{\to}$ The nucleotide sequences reported in this paper have been submitted to GenBank and have been assigned accession numbers AB326393 to AB326416.

more than 13% and *long interspersed repetitive elements* (LINEs) comprise more than 20% of the whole genome (International Human Genome Sequencing Consortium, 2001). Furthermore, recent genome project on short-tailed opossum (*Monodelphis domestica*) have revealed that SINEs cover more than 10% and LINEs comprise more than 29% of its genome (Gentles et al., 2007).

More than 30 SINE families have been characterized based on their structure. Usually SINEs are composed of a 5' terminal tRNA-or 7SL RNA-related region containing a pol III promoter and a partner LINE-related 3' tail. Furthermore, several SINE families are grouped into a superfamily based on the presence of a central conserved domain. To date, three superfamilies have been characterized as V-SINEs (Ogiwara et al., 2002), Due-SINEs (Nishihara et al., 2006a) and CORE-SINEs (Gilbert and Labuda, 1999), which exist in vertebrate and invertebrate genomes. Among these three superfamilies, CORE-SINEs are considered to be a rather young group and some intact CORE-SINEs are thought to possess retropositional activity in mammalian (especially non-eutherian) genomes (Gilbert and Labuda, 2000). The CORE element, which is the central conserved domain of CORE-SINEs, was initially reported as mammalian interspersed repeats (MIRs) and is widely distributed among mammalian genomes (Jurka et al., 1995; Smit and Riggs, 1995). Later, this MIR was divided into two families, Ther1 (MIR in RepBase Reports) and Ther2 (MIR3), which are distributed among the genomes of Theria (extant "Theria" consists of all mammals except for platypus and echidnas) (Gilbert and Labuda, 2000). MIRs are the most prevalent repeat in the human genome next to Alu, in that Ther1 shares 2.2% and Ther2 shares 0.3% of the draft human genome sequence (International Human Genome Sequencing Consortium, 2001). The Ther1 and Ther2 are highly divergent, and seem to have lost their retropositional activity before the split of monotremes, marsupials and eutherians, which occurred more than 110 mya (million years ago). Although the CORE-SINEs lack retropositional activity in the genomes of eutherians, they are still active in non-eutherian genomes. Gilbert and Labuda (2000) reported the presence of three additional CORE-SINE families (Mon1, Mar1 and Opo1). The members of these families are

$ \begin{array}{c} 13 \\ 0 \\ 0 \\ 0 \\ 0 \\ 0 \\ 0 \\ 0 \\ 0 \\ 0 \\ $	4	A box	B box	+ +	CORI	E-domain
$ \begin{array}{c} 0 \\ 0 \\ 0 \\ 0 \\ 0 \\ 0 \\ 0 \\ 0 \\ 0 \\ 0 $	r3a					
$\begin{array}{c} 1 \\ 0 \\ 1 \\ 1 \\ 1 \\ 1 \\ 1 \\ 1 \\ 1 \\ 1 \\$	043	TCTGCCTTCAAGAAG <u>TTTACAT</u> AAA.GGN.N.TATTATA			TT	A
$ \frac{1}{10} $	04	CAGCCAAACTG <u>GAGAT</u> CACTATCAG	A	.G		TT
10-0 Experimental State A. 1997 A. 199		TAACT <u>GTCAAA</u> AAGGAAATGACCAG.AG.AG.A			A A A	AGT
$ \begin{bmatrix} 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0$			GA	GAC	T	·····A····A····-
$ \frac{1}{3} Transformation of the second s$						1927 (927)
$ \begin{array}{c} 1 \\ 0 \\ 0 \\ 1 \\ 1 \\ 1 \\ 1 \\ 1 \\ 1 \\ 1 \\$						
$ \frac{1}{10} \text{TEAUTITTAACAACATEGETEOC.} A. C. A. A. C. T. A. C. C$				ChCh CTT		m
$ \frac{1}{10} TAATCTTAATAAAAAAAAAAAAAAAAAAAAAAAAAAAA$			GCAT	C AACACTG	а	A
$ \begin{array}{c} 10 \\ 10 \\ 10 \\ 10 \\ 10 \\ 10 \\ 10 \\ 10 $		TARATCTTARATARGAGCACCATG TAR		GACACTT	T	G.T.YG
31 TOUCKYLANDOLLANDALY T. T G						
31 TOUCKYLANDOLLANDALY T. T G	01	TGATTGTTGAAGAT <u>ATAGTGGAC</u> ATCACA			.	
$ \begin{array}{cccccccccccccccccccccccccccccccccccc$	03	TTGGCCATT <u>AAAAGACTAATGATA</u> TA.TGCAAA.			<mark>c</mark>	
$ \begin{array}{cccccccccccccccccccccccccccccccccccc$		ATGTTAAAACAAGGGGTTAGGCAC	т		<mark>c</mark> r	.A
$ \begin{array}{cccccccccccccccccccccccccccccccccccc$		GANATGGCANGTAGCA.TA	AAA.		<mark>c</mark>	
$ \begin{array}{cccccccccccccccccccccccccccccccccccc$		TATTATGCAAAAAAATCAGACCGGGGCAT.A.T.	c		· <mark>C</mark>	·····T·····
$ \begin{array}{cccccccccccccccccccccccccccccccccccc$		ATAGGGAACA <u>AGAGAAACTCAGAG</u> GA.AA.	CA	AACACTT	<mark>c</mark>	
$ \begin{array}{cccccccccccccccccccccccccccccccccccc$		GTATTTTTA <u>AAAAAAAGTTATT</u> GGCATA.T.	CA		····· <mark>c</mark>	
$ \begin{array}{cccccccccccccccccccccccccccccccccccc$		TTTGCCTT <u>AATATAGTCTAGAACT</u> GT	C	MATACTT		G
6 AGGCTGMARTTAL MODELLING. A. T. T. T. T. A. A. A. A. C. A. A. A. C. A. C. M. GALENCETT. C. G. J. AT. A. A. C. C. T. G. A. T. A. A. T. A. T. A. T. A. A. T. A. A. T. T. T. T. A. T. T. T. A. T. A. T. T. T. T. A. T. T. T. T. A. T.						•
95 TCIALAGORIANCELETTETTETT		TTAATACTATTTTAAGACAAATACT	GCA		·····	······································
65 TCANTONYCRAQUADALADATCOCCUPAT A.G., M.G., T. T., C., MARCALLANATCOCCUPATIONAL CONTROL AND		AGGCTGAAGATTAAGAUGCACTAAG. AA. TT. TTC. AG. TT. CA. T. A. T. A. T. T. T. CA. T.		CACACACACTT	<i>a</i>	AT
117 ATATGTTTAAAATGCCCCCCAT				GACACACTT		
18 CATTORCAMACHAGATTACTTAG			G A	C. GACACATTT		A -C
99 AUTOTTOTALANA_CALAGENERTT		CATTTGTCAACATGACTTCATTTAG AG CA T	T. T.A.	GACA CTT		
101 000						
226	Bd					
226	020	TAAGTCAAGAATGGATTGTGTGG				
447	126				222 C2 T	т
347			.			
1 TransfittorAttreatTestercture of the Conservation of the Constructure of the C	036		CAG GG * 70	AACGTG AACGTG AACAC 90	АААС	130 c.
7	n036 n047		GGG. * 70		-AAACT -AAACT -AAACT - 110	130
1a-Marjoi intermidiate 0 .ACCCARCAGECCEGGAAACAATTCCATGGCATTGCGCACTGCCANATGGTTTAACA 13	036 047 3a		• 70	90	AAACT IIII ACT IIII AT	
01	036 047 3a 043 04		GGGG		AAAC	130 · · · · ·
3 G ANACACCEANATCTICUTOTE AGGCCANTATCOCCACTAAACCATTTCCAAACC 4 CCCAAACCACACAAACAACCAAACTAATCTACTCACGAAACCATTCCCAAACC 33 TTCCCCCCTCC G AATTAGTCCTCAGAGAGATTTAAAACCAATTAGCAGAAACCATA 34 TTCCCCCCTCC GC. TTATGGGATCGCAGAGAGATTTAAACCAATGTAGGCAGGAA.CGAGAACCTG.A 35 TTCCCCCCTCC GC. TTATGGGATCGCACAACGAGAGCAGGACGAGGAGGACGTGTCGCAGAACGTTA.A 36 CCC	036 047 3a 043 04 17		• 70 • 70 MAT.T MGG.A SGATC		AAACT Juaaa.aCT 110	
41	036 047 3a 043 04 17 11 3a-Ma		• 70 • 70 MAT.T MGG.A SGATC	сысс-те Висас 90	AAA	130
33 TYCCCCCCTCCG	036 047 3a 043 04 17 11 3a-Ma 10	CCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCC	G		ала, С с 	
11 TTCCCCCCTCC	036 047 8a 043 04 17 11 8a-Ma 10 83	10 10	G		AAA	
bb	036 047 8a 043 04 17 11 8a-Ma 10 33 34	C. G. G. A. C. G. J.	ал		ΔΔΔ Δ	
1	036 047 8a 043 04 17 11 8a-Ma 10 83 84 23	ANALAMAMAMA ANG <u>OTTIAGT</u> TIANTTOGGAGAMAMIAACGAACTCATCATAGAMAMIAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAA	мл.т. мат.т мат.т мат. жа		AAA,	
33	036 047 043 04 04 04 04 04 04 04 04 04 04 04 04 04	ANALAMAMAMA ANG <u>OTTIAGT</u> TIANTTOGGAGAMAMIAACGAACTCATCATAGAMAMIAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAA	мл.т. мат.т мат.т мат. жа		λλλ	.
41	036 047 3a 043 04 17 11 3a-Ma 10 33 34 23 31 3b	ANNANANAN AOGUTTIAGTTIAATTOGGAGAANANTANCATGANCATAGANANTANANTGCAN AOGUTTIAGTTIAATTOGGAGAANANTANCATGANCATAGANANTANANTGCAN AOGUTTIAGTTANTCATTOGCACTOG CANGCACTCOTTOTTOTTCATAGANGTCOT			AAA,	
55	336 047 34 343 34 304 33 34 33 34 33 31 05 01	AAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAA	AAT.T GG.A JGG.A JGATC TAACA LAATC TAACA TAACA TT.A TC.A TC.A TC.A TC.A		λλλ	.
1 TTT	336 047 8a 10 33 34 35 35 35 35 35 35 35 35 35 35 35 35 35	ANALATAGGACTAGACAGGGAAAAA ANTGCAAACTTTCAGCTGAGACCGAACTGGACCGTGGACCATGGAACATGGAACATGGAACATGGAACATGGAACTGGACGGAC			AAA A	
22	336 047 8a 043 04 10 13 34 10 13 13 15 15 10 13 30 4	ALAMAMAMAA AAG <u>OTTAGTE</u> TTAATTGOGAGAAAATAACATGAACATAGAAAATAAAATGCAA J0 J0	AAT.T G.G.A JGG.A JGGAC JGACA LAATC TAACA LAATC TC.A TC.A TC.A TC.A TC.A TAT.A LAAGA		λλλ	
99	36 047 30 43 04 10 13 34 33 43 31 05 01 33 44 33 10 50 10 33 44 30 10 50 10 33 44 30 10 33 44 30 10 10 10 10 10 10 10 10 10 10 10 10 10	ANAMANANANANANANANANANANANANANANANAN			AAA	g .
0	a 43 43 43 43 43 1 1 1 3 4 5 1	CCCCCCCCCC . G			λλλ	
2 Interview	36 047 8443 047 11 8a Me 00 33 12 04 33 12 04 33 12 04 33 12 04 55 12 02 04 7	10 10			AAA	.
6 TICKCCTCC G	36 47 a43 44 7 1 a-Me 34 3 1 b1 3 45 12 9	ALAMAMAMAA AAG <u>UTTAGTT</u> TTAATTGGGAGAAAATAACATGAACATGAACATAGAAAATAAAATGCAAA AAG <u>UTTAGTT</u> TTAATTGGGAGGAAAATAACATGAACATGAACATAGAAAATAACATGAAAATTCAAAATGCTCATTGCTCATTGATTG	AAT.T G.G.A JGG.A JGGA JGATC TAACA LAATC TAACA LAATC TC.A TC.A TC.A TC.A TC.A TC.A TA.A JAAG JAA.A JAA.A JAA.A JAA.G JAA.	срасса - та Бысл - с 90	λ	
9 TOTOCTECTTTTTTTTTAAAAAAAUGTTTTTTTTTTAAAAAAUTTTCATTAAGAACTTTTCATAAAGAATTC 6 TOCOCCETTTCCCCAAGAACAAGATATTAG.TAACTACTTAGAACAGTTACCGCCTAACGAGAACT 17 TOCOCCA	36 47 a 43 47 a 43 47 7 1 a - Ma 0 3 4 3 1 b 1 3 4 5 1 2 9 0 c	ALAMAMAMAA AAG <u>UTTAGTT</u> TTAATTGGGAGAAAATAACATGAACATAGAAAATAAAAT			AAAA , A	
64 TCCCCCCTCTGTCCCMAGACMAGTTATAG_TBACAGCTGAACTAAGTAAGTAAGTTAACCGCTAAACTAGGAGT 17 TCCCCCCTCTTGCCCAACTAACTAAGTAACTAACTAAGTAACCGCTAACTAGGAGT 18 TCCCCT_TTCAGACAAACTACTTCCTAAAGTAACTAACTACTACTCCGAACTCCCAACTGCCAATGTCATATCC 19 TCCCCCT_TTCAGACAAACTCTTCCTTAAATGAACTACTACTAGGACCTGTCGTAATTCC 20	36 047 8 a 043 04 10 10 10 10 10 10 10 10 10 10 10 10 10	10 10		сф.сста бысАс 90	2000 A	g .
17 TCCCCCATEC	36 34 34 34 34 34 34 34 34 34 34	ALAMAMAMAA AAG <u>UTTAGTT</u> TTAGTTTTAATTOGGAGAAAATAACATGAACATAGAAAATAAAAT			Αλλ	
18 TCCCT = TTC	36 36 36 36 36 36 30 47 30 43 30 30 43 30 30 30 30 30 30 30 30 30 3	10 10		сф.сста бысАс 90	AAA , A	g .
23 TTOTOCOGNACT. <u>GCACTACT</u> TAAGGTTCAAGTCAGT.ATTCAATAGTCATATTAAGCACTTGTATGTATGT 24	036 047 3a 043 043 04 10 11 33 44 23 13b 01 33 44 23 13b 01 33 44 23 13b 01 33 44 23 10 30 4 02 5 01 20 4 30 4 30 4 30 4 30 4 30 4 30 4 30 4	ALAMAMAMAA 10 10 30				
14 20	036 047 3a 043 04 17 11 11 10 33 34 10 33 31 10 33 31 004 005 004 005 002 009 10 002 009 10 002 009 10 002 009 10 002 004 7 004 7	10 10			AAA	
20	036 047 8a 043 04 10 11 8a 10 33 34 10 33 34 10 33 34 01 30 34 30 30 34 30 30 30 30 30 30 30 30 30 30 30 30 30	ALAMAMAMAA 10 30				
926 <mark>вс</mark> Алалалалассаласссттто <u>стооо</u> тосоосоласаодастолесстессаа 936 <mark>с.</mark> Алада	036 047 8a 043 04 17 18a-Md 10 33 8b 01 33 223 18 00 33 223 10 20 20 20 20 20 20 20 20 20 20 20 20 20	CCC			AAA	
036 	036 047 3a 043 043 111 33a 10 334 334 231 33b 013 04 501 203 004 502 009 10 302 009 10 302 009 10 302 2009 10 302 2009 10 302 2009 10 302 2009 10 304 304 304 304 304 304 304 304 304 30	ALAMAMAMAA AAG <u>UTTAGTT</u> TAATTGGGAGAAAATAACATGAACATGAACATAGAAAATAAAATGCAAA AAG <u>UTTAGTTAGTT</u> TAATTGGGAGAAAATAACATGAACATGAACATGAAATGCAAAATGCAAATTGCAAATGCACATGATTGAT				
	036 047 3a 043 04 17 11 3a-Me 03 33 4 33 13 03 33 33 33 33 33 33 33 33 33 33 33 33	ADMAMMAMAA ADMAMMAMAAAAAAAAAAAA			AAAA , A	
947	036 047 3a 3043 17 11 17 13a-Me 17 13a-Me 17 13a-Me 17 13a-Me 17 13a-Me 17 13a-Me 17 13a-Me 17 13a-Me 17 13a-Me 17 13a-Me 17 13a-Me 17 13a-Me 17 13a-Me 17 17 13a-Me 17 17 13a-Me 17 17 17 17 17 17 17 17 17 17 17 17 17	ALAMAMAMAM Alag <u>ITTAGTE</u> TTAATTGGGAGAAAATAACATGAACATGAACATAGAAAATAAAATGCAAA AAG <u>ITTAGTTAGTT</u> TAATTGTTAATGGAGAGAAAATAACATGAACATGAACATGAAAATAACATGAACATGAACATGAACATGAAATGCAAAATGCCAAATTGCAATTGATTTACTTTCATTTTATTTTGG G. T. <u>ITTAGTAAATGGCAAGAGAGGAGGAGGAGGAGGAGGAGGAGGAGGAGGAG</u>				g .

Fig. 1. Sequence alignments of the newly identified SINE subfamilies. (A) The Mar3 subfamily is subdivided into Mar3a, b, c, and d. (B) The Mac1 subfamily. The dots indicate nucleotides identical to the consensus sequence at the top. The A box and B box, which are typical for the tRNA region of each SINE are shown by thick bar. The diagnostic nucleotides for each subfamily are shaded in black. The insertions immediately upstream of all Mar3 CORE domains caused by the duplication of the 3' end of the tRNA-related region are indicated by arrows. Underlined nucleotides indicate the target site duplications of each SINE loci.

Download English Version:

https://daneshyari.com/en/article/2819489

Download Persian Version:

https://daneshyari.com/article/2819489

Daneshyari.com