J Mol Evol (2005) 60:677–687
DOI: 10.1007/s00239-004-0259-5

JOURNAL OF MOLECULAR EVOLUTION

BoS: A Large and Diverse Family of Short Interspersed Elements (SINEs) in *Brassica oleracea*

Xiaoyu Zhang, Susan R. Wessler

Department of Plant Biology, University of Georgia, Athens, GA 30602, USA

Received: 20 August 2004 / Accepted: 11 October 2004 [*Reviewing Editor:* Dr. Jerzy Jurka]

Abstract. Short interspersed elements (SINEs) are nonautonomous non-LTR retrotransposons that populate eukaryotic genomes. Numerous SINE families have been identified in animals, whereas only a few have been described in plants. Here we describe a new family of SINEs, named *BoS*, that is widespread in Brassicaceae and present at ~2000 copies in *Brassica oleracea*. In addition to sharing a modular structure and target site preference with previously described SINEs, *BoS* elements have several unusual features. First, the head regions of *BoS* RNAs can adopt a distinct hairpin-like secondary structure. Second, with 15 distinct subfamilies, *BoS* represents one of the most diverse SINE families described to date. Third, several of the subfamilies have a mosaic structure that has arisen through the exchange of sequences between existing subfamilies, possibly during retrotransposition. Analysis of *BoS* subfamilies indicate that they were active during various time periods through the evolution of Brassicaceae and that active elements may still reside in some *Brassica* species. As such, *BoS* elements may be a valuable tool as phylogenetic makers for resolving outstanding issues in the evolution of species in the Brassicaceae family.

Key words: *BoS* — SINE — *Brassica* — Mosaic — Hairpin — Widespread — Diverse

Introduction

Class 1 (RNA) transposable elements, or retrotransposons, propagate through a replicative mechanism called retrotransposition, in which an RNA intermediate is reverse-transcribed into a cDNA that can integrate in the genome (for review see Capy et al. 1998). There are two types of retrotransposons: those bordered by long terminal direct repeats (LTRs) are called LTR-retrotransposons while those without LTRs but with poly(A) (or another form of simple nucleotide repeat) at their 3′ end are called non-LTR retrotransposons (Doolittle et al. 1989; Xiong and Eickbush 1990). Non-LTR retrotransposons include long interspersed elements (LINEs) and short interspersed elements (SINEs). LINEs encode a protein(s) that catalyzes their retrotransposition through a mechanism termed target-primed reverse transcription (TPRT) (Eickbush 1992; Luan et al. 1993). SINEs are short (< 500-bp), nonautonomous elements with no coding capacity and their retrotransposition relies on the protein machinery supplied *in trans* by LINEs (Dewannieux et al. 2003; Kajikawa and Okada 2002).

A typical SINE consists of three regions: a 5′ head containing conserved promoter motifs that are recognized by RNA polymerase III (Pol III), a 3′ tail usually consisting of poly(A) and an internal region (Okada 1991a, b). Head regions of most SINEs are ancestrally derived from tRNA genes. Accordingly, transcripts from this region can potentially fold into cloverleaf-like secondary structures reminiscent of their tRNA progenitors. Two exceptions are the primate Alu family and the rodent B1 family, which

Correspondence to: Susan R. Wessler, 4505 Miller Plant Science Building, University of Georgia, Athens, GA 30602; *email:* sue@plantbio.uga.edu

have head regions derived from 7SL RNA genes (Ullu and Tschudi 1984; Weiner 1980). The internal regions of different SINE families are highly variable in size as well as sequence. In several cases, sequence similarity between the internal region of a SINE and the 3′ untranslated region of a LINE has been noted (Ogiwara et al. 1999; Ohshima et al. 1996; Okada and Hamada 1997; Okada et al. 1997; Terai et al. 1998). This feature led to the hypothesis that ancestral SINEs were generated when the 3′ end of a LINE was fortuitously translocated (by retrotransposition or recombination) immediately downstream of a tRNA gene, and the resulting chimeric sequence was mobilized and amplified by LINE-encoded protein(s) (Ohshima et al. 1993; Okada and Hamada 1997).

Several SINE families consist of subfamilies, or groups of elements within a family that share a set of diagnostic nucleotides. Examples include the Alu family in primates (Shen et al. 1991), the AFC family in cichlid fish (Takahashi and Okada 2002), and the S1 family in *Brassica* (Lenoir et al. 1997). In several cases subfamilies appear to have exchanged their head and internal regions, resulting in a new "mosaic" subfamily (Kass et al. 1995; Lenoir et al. 1997; Takahashi and Okada 2002; Zietkiewicz and Labuda 1996).

Integration of SINEs into chromosomes occurs via a TPRT process, where a LINE-encoded protein cleaves the bottom strand of target DNA (called "5′ nicking") and utilizes the exposed 3′ DNA end to prime the reverse transcription of SINE RNA into the first cDNA strand. Another cleavage is then made on the top strand of the target DNA (called "3′ nicking") and the synthesis of second strand cDNA proceeds (Luan et al. 1993). Following the integration of SINEs and the repair of single-stranded gaps, a direct repeat of target site sequence (called the target site duplication; TSD) is generated. Previous analyses of the sequence context near several plant and animal SINE families revealed sequence motifs at the 5′ and/or 3′ nicking sites (Jurka 1997; Lenoir et al. 2001; Tatout et al. 1998).

SINEs are widespread in all three eukaryotic kingdoms but appear to be more diverse and abundant in animals than in plants or fungi. Numerous families of SINEs have been described in animals and some have accumulated to extremely high copy numbers. For example, the human genome harbors ~1.5 million Alu SINEs, accounting for ~10% of the genome (Lander et al. 2001; Schmid 1996). In contrast, only a few families of SINEs have been reported in plants, where their copy number is generally much lower. Plant SINEs include *p-SINE1* in rice (~6500 copies), *TS* in tobacco (~50,000 copies), S1 in *Brassica* (~500 copies in *B. napus*), and *AtSN1*/RAthE3, *AtSN2*/RathE1, and RAthE2 in *Arabidopsis thaliana* (~70, ~130–150, and ~60 copies, respectively) (Deragon et al. 1994; Lenoir et al. 2001; Myouga et al. 2001; Umeda et al. 1991; Yoshioka et al. 1993).

Here we describe a new family of SINEs, named *BoS*, that is widespread in the Brassicaceae family and present at ~2000 copies in *B. oleracea*. Over 400 full-length elements were identified from a partial *B. oleracea* genomic database (generated by The Institute for Genomic Research; TIGR) and grouped into 15 distinct subfamilies. *BoS* elements were characterized with regard to their modular primary structure, secondary structure, and target site preference as well as their distribution in the Brassicaceae family. Overall, the *BoS* family exhibits several interesting features, including frequent formation of mosaic subfamilies through sequence exchange between existing subfamilies and an ability of the head regions of their RNAs to form distinctive hairpin-like secondary structures. *BoS* is the most abundant SINE family identified in the Brassicaceae and is one of the most diverse families from any organism identified to date with regard to the number and sequence divergence of its subfamilies.

Materials and Methods

Database Searches and Sequence Analyses

Computer-assisted searches to identify *BoS* elements in *B. oleracea* were performed against the TIGR *B. oleracea* genome database (brassica prelim sequences; http://tigrblast.tigr.org/euk-blast/index.cgi?project = bog1) with no filter but otherwise using default parameters. Preliminary *B. oleracea* sequence data were obtained from the TIGR Web site at www.tigr.org. Blastn searches to identify *BoS* elements in other Brassicaceae species were performed at the National Center for Biotechnology Information Web site (http://www.ncbi.nlm.nih.gov; databases NR and EST; no filter but otherwise using default parameters). *A. thaliana* tRNA sequences were downloaded from the Genomic tRNA Database (http://rna.wustl.edu/GtRDB/At/At-align.html) (Lowe and Eddy 1997). Multiple sequence alignments were performed with the CLUSTALW server available at European Bioinformatics Institute (http://www.ebi.ac.uk/clustalw) using default parameters. Consensus sequences were derived based on simple majority rule. RNA secondary structures were predicted with the Vienna RNA Secondary Structure Prediction server (http://rna.tbi.univie.ac.at/cgi-bin/RNAfold.cgi) (Hofacker 2003). Phylogenetic trees were generated based on the neighbor-joining method, using PAUP* version 4.0b8 with default parameters (Swofford 1999). Bootstrap values were calculated from 1000 replicates.

Determination of BoS *Target Site Preference*

Sequence context of *BoS* insertion sites was analyzed as previously described (Jurka 1997). Briefly, the 5′ flanking sequences of *BoS* elements, including TSDs and 15 additional bases immediately upstream of TSDs, were adjusted so that all TSDs started at the same position, and the 3′ flanking sequences including TSDs and 15 additional bases immediately downstream of TSDs were adjusted so that all TSDs ended at the same position. The nucleotide occurrences at each position were then determined. Chi-square analysis of *BoS* target site preference was performed as described (Tatout et al. 1998) using a significance level of $p < 0.001$ (at 5′ nicking site) or $p < 0.01$ (at 3′ nicking site) for 3 degrees of freedom (df).

Results

Identification of a New Family of SINEs in B. oleracea

In the process of analyzing TEs in the TIGR *B. oleracea* genomic sequence database, a 194-bp segment (clone JBOGL76RB; position 165–358) was identified as an insertion into a LINE element. This segment has several structural features of SINEs, including a 3′ poly(A) track (22 bp) and a flanking direct repeat (12 bp with one mismatch), but does not share significant similarity with previously described SINEs. The sequence of this segment was used as query in a Blastn search to determine if related elements exist in the *B. oleracea* genome. Fifty hits were identified (E value $<10^{-3}$) that displayed ~75–100% identity to the query over their entire length, ended at the 3′ end with a poly(A) tract, and most were flanked by short direct repeats (see below). In addition, 120 hits displayed less sequence identity (60–75% identical) to either the entire query sequence or its 5′ end (~80–90 bp), suggesting the existence of distantly related elements. These hits were resolved into groups based on sequence identity (not shown) and sequences from each group were used as queries to identify additional elements by reiterated Blastn searches. In this way, 377 additional full-length hits were identified. None of the 428 elements (including the original element, the 50 elements identified in the initial Blastn search, and 377 elements identified in subsequent searches) shared significant sequence similarity with previously described SINEs outside of the Pol III promoter motifs (see below). Therefore, these elements represent a new SINE family that was named *BoS* for B. oleracea SINE. In addition to recovering full-length elements, the Blastn searches also detected a large number (~350–400) of elements that were truncated during cloning (located on clone ends) and were not analyzed further. Because each truncated element, statistically, represents half an element, the available database (covering ~30% of the *B. oleracea* genome) contains ~600–650 *BoS* elements. Therefore, the copy number of *BoS* elements was estimated to be ~2000 in the entire genome.

The nucleotide sequences of the 428 full-length *BoS* elements were compared by CLUSTALW multiple alignment and used to generate an unrooted phylogenetic tree (available upon request). In this way, 15 subgroups were resolved that were well supported by bootstrap values. These subgroups were defined as subfamilies and designated *BoS_a* through *BoS_l* (subfamily features summarized in Table 1). Intrasubfamily sequence identity (defined as average nucleotide sequence identity of subfamily members to their respective consensus sequences) ranged from ~80% (*BoS_e*) to ~98% (*BoS_ai2*) and copy numbers ranged from a few (~3, *BoS_l*) to ~270 (*BoS_ab*) (extrapolated to the whole genome). The consensus sequences of *BoS* subfamilies are provided in supplemental data.

Table 1. Summary of *BoS* elements in *B. oleracea*

Subfamily	No. of full-length sequences	Consensus sequence length (bp)	Intrasubfamily sequence identity
BoS_a	16	195	94%
BoS_ab	58	171	95%
BoS_ai2	17	233	98%
BoS_b	38	167	93%
BoS_c	61	218	95%
BoS_d	33	172	93%
BoS_e	51	188	80%
BoS_f	18	229	93%
BoS_g	20	187	82%
BoS_h	17	139	84%
BoS_i1	30	185	89%
BoS_i2	11	235	96%
BoS_j	23	169	85%
BoS_k	34	193	84%
BoS_l	1	214	N/A

Primary Structure of BoS *Elements*

BoS elements exhibit a characteristic SINE structure, including a 5′ head region (~80 bp; containing putative promoter motifs recognized by RNA polymerase III) (Fig. 1; A and B boxes), an internal region (length varies from 50 bp in *BoS_h* to 140 bp in *BoS_f*), and a poly(A) track (average length, ~19 bp). Comparison of the consensus sequences of *BoS* subfamilies revealed significant similarity in their 5′ head regions (on average 66% identity), with the highest level of sequence conservation in blocks designated 1–5 (Fig. 1). Block 1 is located at the 5′ end, where nearly all families begin with the sequence 5′-AACCRRG-3′. Blocks 2 and 4 contain the A and B boxes of Pol III promoter motifs that are highly similar to those found in several previously described plant SINEs as well as *A. thaliana* tRNA genes. Comparison of the consensus sequences of *BoS* head regions to 629 *A. thaliana* tRNA genes showed that *BoS* elements are distantly related to tRNA^{Gln} genes (anticodon CTG and TTG) and tRNA^{Pro} genes (anticodon AGG, CGG and TGG) (~50–60% identity; Fig. 1). While the highest level of sequence similarity was found in blocks 2 and 4, weaker but significant sequence similarity was also detected in other regions. However, neither sequence similarity (Fig. 1) or phylogenetic analysis (not shown) could

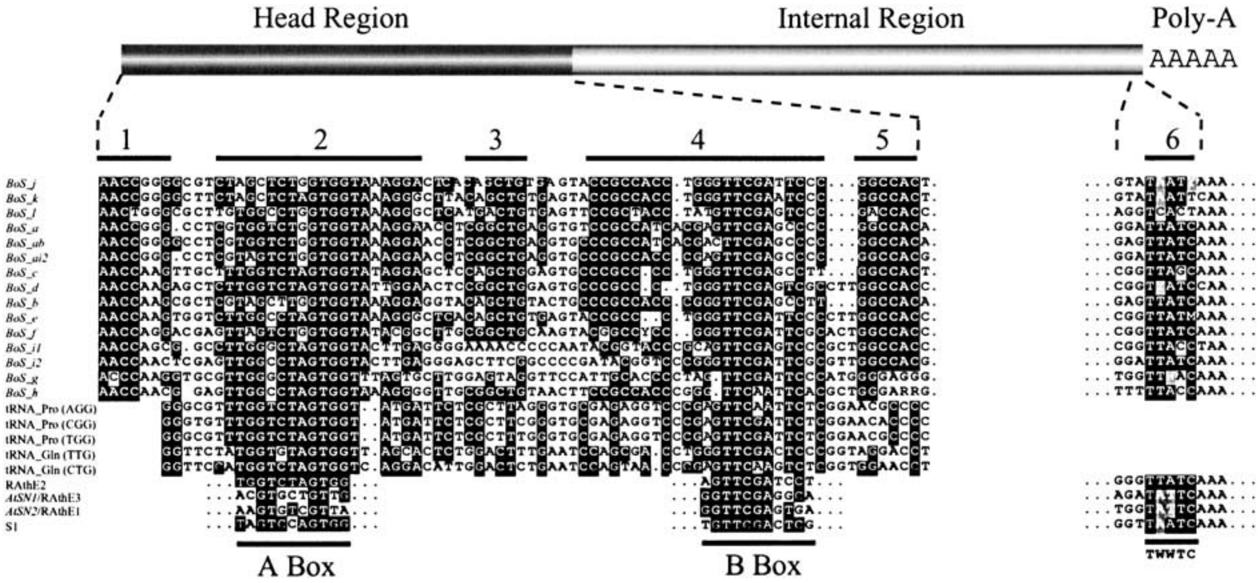

Fig. 1. Primary structure of *BoS* subfamilies. A schematic representation of element structure is shown above the alignment. Alignment of regions conserved in all subfamilies is shown and is discussed in the text. Regions that are highly variable among subfamilies are not shown. Blocks of highly conserved residues (numbered 1–6) are noted by horizontal bars above the alignment, while the A and B boxes of RNA pol III promoter motifs are indicated by horizontal bars below the alignment. *Arabidopsis* tRNA^{Gln} (CTG, TTG) and tRNA^{Pro} (AGG, CGG, TGG) genes are included in the alignment of the head region. The A and B boxes and 3′ termini of SINEs previously described in Brassicaceae (including AtSN1/RAthE3, AtSN2/RAthE1, RAthE2, and S1) are included for comparison.

unambiguously determine which tRNA gene is more closely related to *BoS* elements, and thus the direct ancestor of *BoS* remains to be identified.

The only significant similarity shared by all *BoS* subfamilies in their internal region was a 5-bp motif (TTATC) located at the extreme 3′ end, immediately upstream of the poly(A) track (block 6; Fig. 1). Furthermore, the two nucleotide residues upstream of this motif are usually purines (Fig. 1). A similar motif is also found in all known SINEs from the Brassiceae family, including *AtSN1*/RAtheE3 (GATATTC), *AtSN2*/RAthE1 (GGTATTC), RAthE2 (GGTTATC) from *A. thaliana* and S1 from *Brassica* (GTTAATC) (Deragon et al. 1994; Lenoir et al. 2001; Myouga et al. 2001), but is absent from the tobacco *TS* and the rice *p-SINE1* families (Umeda et al. 1991; Yoshioka et al. 1993). Therefore, it appears that all poly(A) SINEs from Brassicaceae described to date share an 5′-RRTWWTC-3′ motif at the 3′ termini of their internal regions.

Mosaic BoS *Subfamilies*

In several instances the head region of one *BoS* subfamily was found to be highly similar to the head region of a second subfamily, while its internal region is highly similar to a third subfamily. For example, *BoS_a*, *BoS_ab*, and *BoS_ai2* have nearly identical head regions but distinct internal regions (Fig. 2A). Significantly, the internal region of *BoS_ab* is nearly identical to *BoS_b*, whereas *BoS_ai2* shares its internal region with *BoS_i2* (Fig. 2A). Note that the head regions of *BoS_b* and *BoS_i2* are only weakly similar to each other and to the head regions of *BoS_a*/*ab*/*ai2* (42–72% identity). The relationships among these subfamilies determined from their head and internal regions are clearly inconsistent with each other (Fig. 2B). These results suggest that formation of the ancestral *BoS_ab* and *BoS_ai2* subfamilies involved shuffling of the head and internal regions between existing elements (see Discussion). A similar process may also be involved in the formation of *BoS_c* and *BoS_f* subfamilies, as they share 70% identity in the head region but are 96% identical over ~50 bp of their internal region (not shown).

Secondary Structure of BoS *Elements*

Because SINEs transpose via an RNA intermediate, it was of interest to determine the secondary structure of their encoded RNAs. The head regions of all *BoS* subfamily RNAs are predicted to form a relatively stable hairpin structure with free energies ranging from ~−30 to ~−45 kcal/mol (on average, −35.3 kcal/mol) (Fig. 3). This structure is highly conserved despite intersubfamily variation in primary sequence because nucleotide substitutions on one strand of the hairpin are often accompanied by compensatory changes on the other strand. For example, the predicted secondary structures of the head regions of

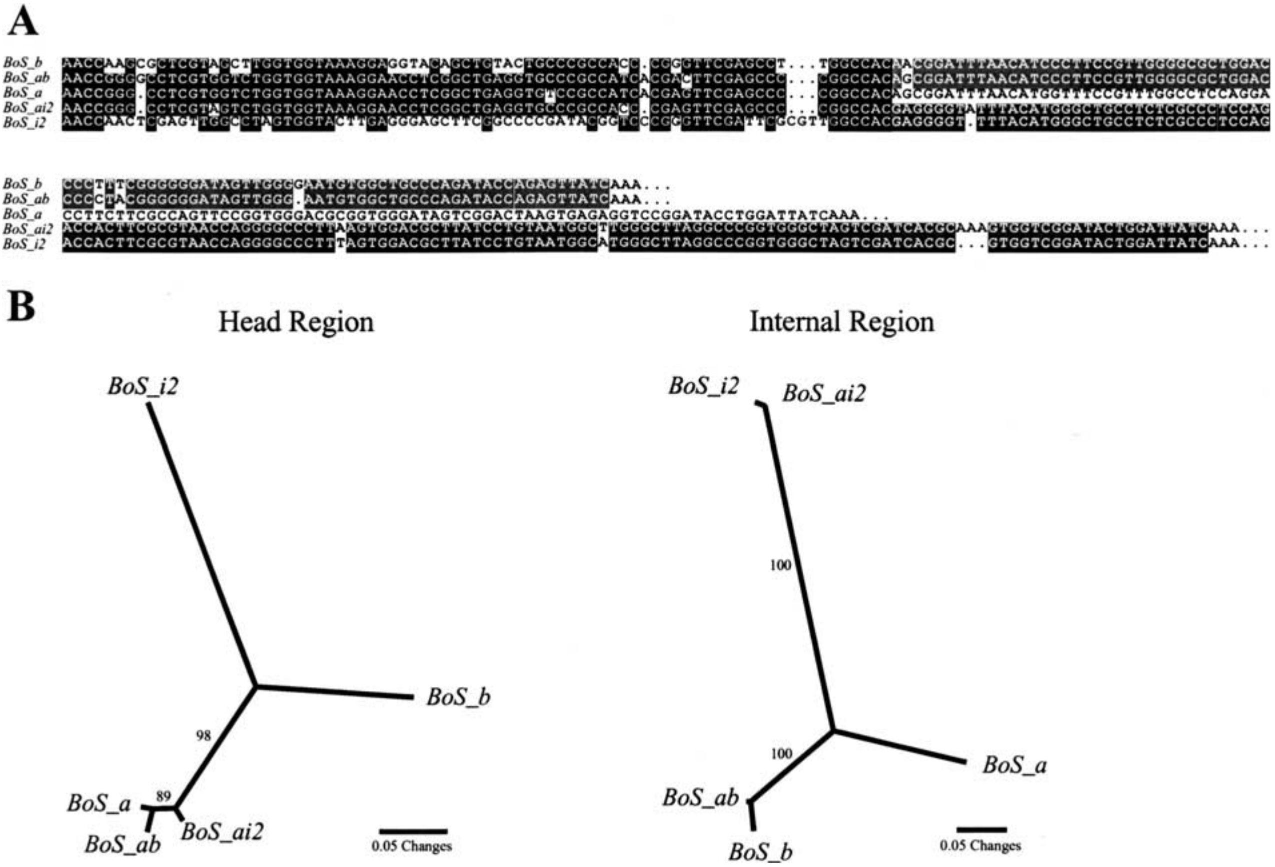
A
BoS_b AACCAAGCGCTCGTAGCTTGGTGGTAAAGGAGGTACAGCTGTACTGCCCGCCACC.CGGGTTCGAGCCT...TGGCCACAACGGATTTAACATCCCTTCCGTTGGGGCGCTGGAC
BoS_ab AACCGGGGCCTCGTGGTCTGGTGGTAAAGGAACCTCGGCTGAGGTGCCCGCCATCACGACTTCGAGCCC...CGGCCACAGCGGATTTAACATCCCTTCCGTTGGGGCGCTGGAC
BoS_a AACCGGG.CCTCGTGGTCTGGTGGTAAAGGAACCTCGGCTGAGGTGTCCGCCATCACGAGTTCGAGCCC...CGGCCACAGCGGATTTAACATGGTTTCCGTTTGGCCTCCAGGA
BoS_ai2 AACCGGG.CCTCGTAGTCTGGTGGTAAAGGAACCTCGGCTGAGGTGCCCGCCACC.CGAGTTCGAGCCC...CGGCCACGAGGGGTATTTACATGGGCTGCCTCTCGCCCTCCAG
BoS_i2 AACCAACTCGAGTTGGCCTAGTGGTACTTGAGGGAGCTTCGGCCCCGATACGGTCCCGGGTTCGATTCGCGTTGGCCACGAGGGGT.TTTACATGGGCTGCCTCTCGCCCTCCAG
BoS_b CCCTTTCGGGGGGATAGTTGGGGAATGTGGCTGCCCAGATACCAGAGTTATCAAA...
BoS_ab CCCCTACGGGGGGATAGTTGGG.AATGTGGCTGCCCAGATACCAGAGTTATCAAA...
BoS_a CCTTCTTCGCCAGTTCCGGTGGGACGCGGTGGGATAGTCGGACTAAGTGAGAGGTCCGGATACCTGGATTATCAAA...
BoS_ai2 ACCACTTCGCGTAACCAGGGGCCCTTAAGTGGACGCTTATCCTGTAATGGCTTGGGCTTAGGCCCGGTGGGCTAGTCGATCACGCAAAGTGGTCGGATACTGGATTATCAAA...
BoS_i2 ACCACTTCGCGTAACCAGGGGCCCTTTAGTGGACGCTTATCCTGTAATGGCATGGGCTTAGGCCCGGTGGGCTAGTCGATCACGC...GTGGTCGGATACTGGATTATCAAA...
B
Head Region
BoS_i2
BoS_b
98
89
BoS_a
BoS_ab
BoS_ai2
0.05 Changes
Internal Region
BoS_i2
BoS_ai2
100
100
BoS_a
BoS_ab
BoS_b
0.05 Changes

Fig. 2. Mosaic *BoS* subfamilies. **A** Comparison of the nucleotide sequences of *BoS_a*, *ab*, *b*, *ai2* and *i2*. Head regions of the subfamilies are aligned (shaded black), while alignment of the internal regions is only shown for comparisons between *BoS_ab* and *BoS_b* (shaded light gray) and between *BoS_ai2* and *BoS_i2* (shaded dark gray). **B** Phylogenetic relationships of *BoS_a*, *ab*, *b*, *ai2*, and *i2* determined based on their head regions (left) and internal regions (right). These unrooted trees were generated based on the neighbor-joining method. Bootstrap values were calculated from 1000 replicates.

BoS_ab and *BoS_b* RNAs are nearly identical despite their sharing only 73% sequence identity. In contrast to the head regions, the internal regions from *BoS* subfamily RNAs adopt a variety of secondary structures (not shown). Folding of the entire *BoS* RNA from different subfamilies, however, reveals that differences in internal region secondary structure do not appear to interfere with the hairpin formed by the head region.

Target Site Specificity of BoS *Elements*

Integration of SINEs into chromosomes involves two cleavage events (referred to as 5′ and 3′ nicking) of the target site DNA (Luan et al. 1993). Several animal and plant SINE families were found to insert preferentially into A/T rich regions of the genome, and for some elements, conserved sequence motifs have been discerned at both the 5′ and the 3′ nicking sites (Jurka 1997; Lenoir et al. 2001; Tatout et al. 1998). To determine whether *BoS* elements also have a target sequence preference, direct repeats flanking *BoS* elements (that could represent target site duplications [TSDs]) were first identified by comparing the sequences flanking each *BoS* element. TSDs were identified for 319 of the 390 elements with sufficient flanking sequence to make this determination. Ten additional elements flanked by TSDs of 1–5 bp were excluded from this analysis as short TSDs might represent coincidental matches. The length of TSDs ranged from 6 to 19 bp (on average, ~13 bp), but most (~70%) were 11–17 bp. Analysis of the sequence context of *BoS* insertion sites indicated that *BoS* elements insert preferentially into A/T-rich regions of the *B. oleracea* genome (Fig. 4). In addition, the 5′ nicking site exhibits a strong preference for 5′-B/AAA-3′ (where B represents C, G, T) ($p < 0.001$). A weaker but significant preference (5′-T-3′; $p < 0.01$) was also detected at the 3′ nicking site.

BoS *Elements Are Widespread in the Brassicaceae Family*

The relatively high level of sequence divergence among different *BoS* subfamilies indicates that *BoS* is

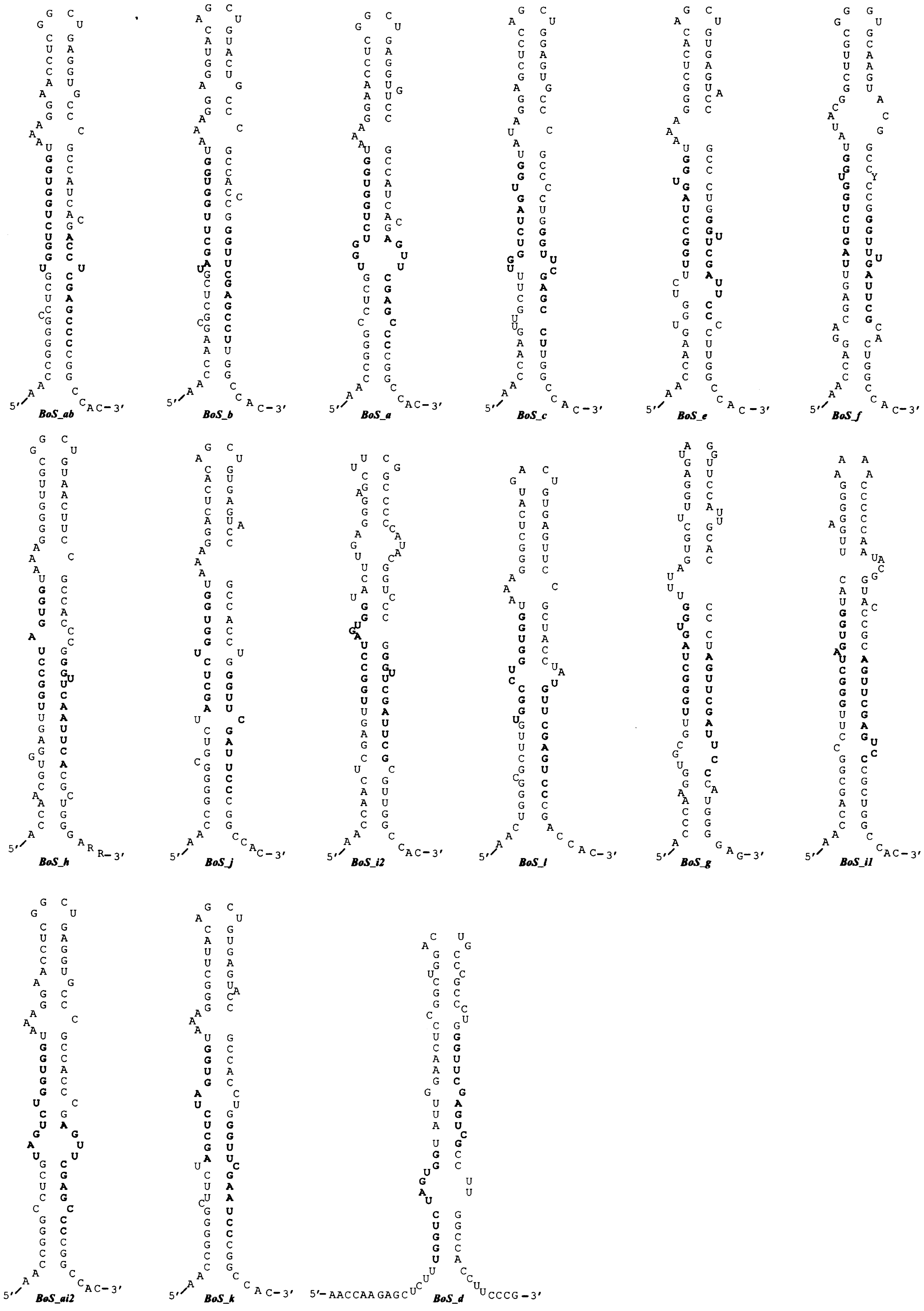

Fig. 3. Predicted secondary structures of the head regions of *BoS* −37.84; *BoS_b*, −38.42; *BoS_c*, −30.59; *BoS_d*, −30.28; *BoS_e*,

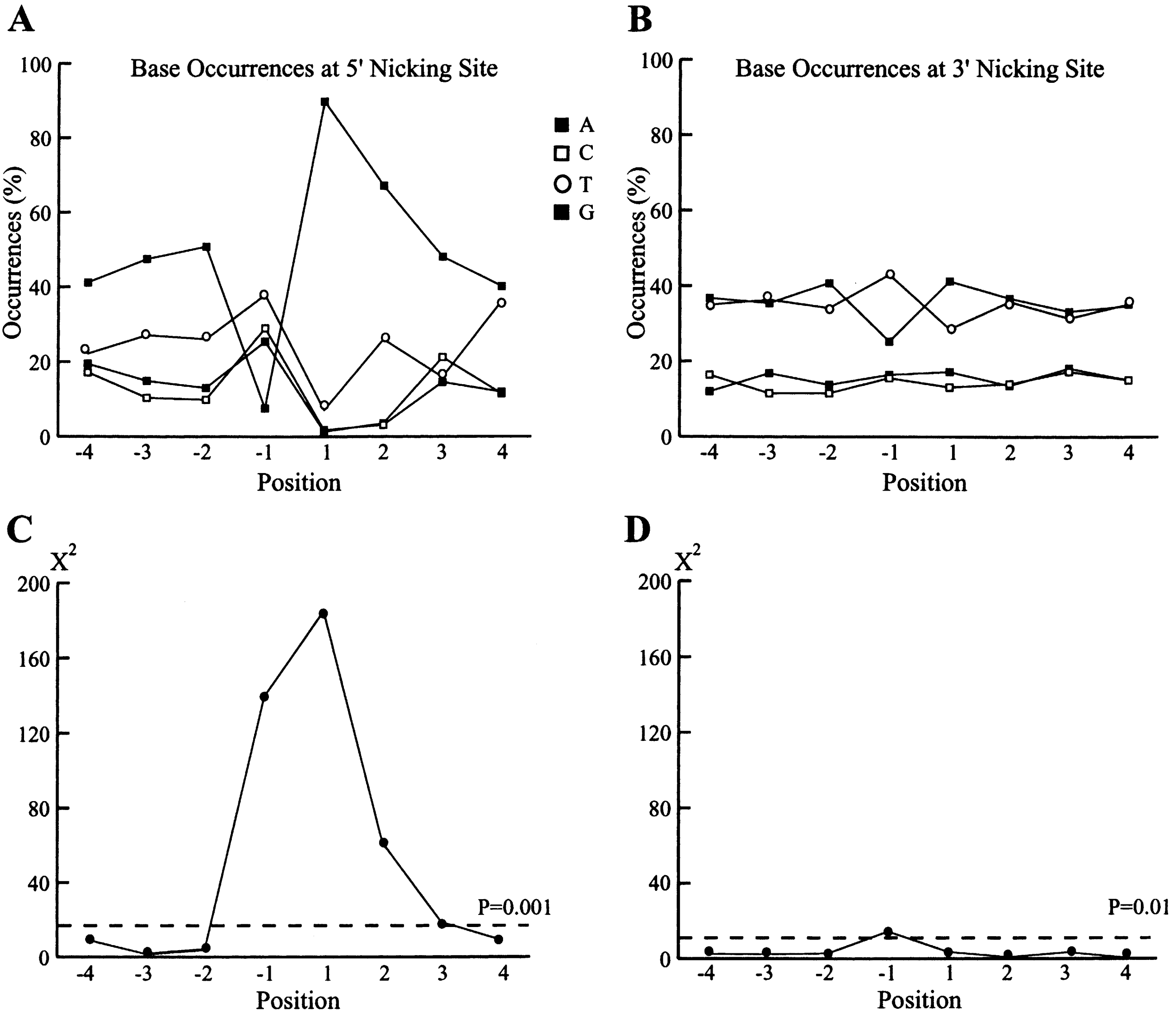

Fig. 4. Insertion site preference of *BoS* elements. **A**, **B** Base occurrences at the 5′ and 3′ nicking sites, respectively. The designations "−1" and "1" are the positions immediately upstream and downstream of the nicking sites, respectively. Adenosine residue data points are represented by black boxes; cytosine, by open boxes; guanosine, by gray boxes; and thymidine, by open and circles. **C**, **D** Chi-square tests of the 5′ and 3′ nicking sites, respectively. Horizontal dashed lines correspond to the significant chi-square values ($p < 0.001$ for 5′ nicking site and $p < 0.01$ for 3′ nicking site).

an old family and, as such, related elements should exist in other plants. In order to identify these elements, the consensus sequence of each *BoS* subfamily was used as query in Blastn searches against the GenBank NR and EST databases. Seven *BoS* elements were identified from the genomic or EST sequences of three other *Brassica* species: *B. napus*, *B. rapa*, and *B. nigra* (Table 2). In each case, the *BoS* element from another *Brassica* species is closely related, over its entire length, to a particular *BoS* subfamily from *B. oleracea*, indicating that each species harbors multiple *BoS* subfamilies and that the divergence of these subfamilies predated the separation of the species. In addition, four *BoS* elements were found in *A. thaliana* (Table 2) that are highly similar to each other over their entire length (~91% identical) but only share sequence similarity with the Brassica *BoS* elements in the head region (~55–60% identity). No *BoS* element was found from species outside the Brassicaceae family.

Discussion

In this study a new family of SINEs, named *BoS*, was identified and found to be widespread in Brassicaceae, where it is present at ~2000 copies in *B. oleracea*. *BoS* elements have typical features of SINEs, including a (i) modular primary structure, (ii) head region distantly related to tRNA genes (Fig. 1), and (iii) strong target site preference similar to previously described plant and animal SINE families (Fig. 4).

Table 2. *BoS* elements in other *Brassica* species and *A. thaliana*

Species	Element name	Similar to *BoS* (Identity %)	Accession No.	Position[a]	Length (bp)	TSD length (bp)
B. napus	*Bn_BoS_a*	*BoS_a* (96%)	CD824681[b,e]	3–186	184	ND[d]
	Bn_BoS_b	*BoS_b* (94%)	AF052241[c]	1,165–1,330	164	14
	Bn_BoS_b	*BoS_b* (92%)	CD836266[b]	446–608	166	11
	Bn_BoS_e	*BoS_e* (91%)	CD836764[b]	378–563	183	13
	Bn_BoS_g	*BoS_g* (92%)	CD812630[b,e]	5–164	160	ND[d]
B. rapa	*Br_BoS_c*	*BoS_c* (96%)	CA991665[b]	141–355	215	ND[d]
B. nigra	*Bg_BoS_c*	*BoS_c* (77%)	AF271220[c]	436–660	225	11
A. thaliana	*At_BoS*	NA[f]	NC_003074	4,635,212–4,635,448	235	13
		NA[f]	AC022492	24,500–24,735	236	6
		NA[f]	AC007843	44.622–44,857	236	8
		NA[f]	AC006577	66,969–67,199	231	NA

[a] Start and end position of *BoS* elements.

[b] EST sequences.

[c] Genomic sequences.

[d] Target site duplication was not determined because one or both ends of *BoS* elements are located at the end of an EST sequence.

[e] These EST sequences consist exclusively of full-length *BoS* sequences and may represent *bona fide BoS* transcripts.

[f] *At_BoS* elements are similar to *Brassica BoS* elements in their head but not their internal regions (see text).

However, with 15 distinct subfamilies, *BoS* is, to our knowledge, one of the most diverse SINE families described to date. Comparison of related but distinct subfamilies has allowed the identification of conserved structural features as well as the origin of new subfamilies from parts of other subfamilies. For example, the head regions of all *BoS* subfamilies can fold into an unusual hairpin-like secondary structure, and existing subfamilies frequently exchanged their head and internal regions, resulting in the formation of new, more active subfamilies. While conserved secondary structure may be necessary for retrotransposition, the origin of new subfamilies from existing subfamilies may have been a major driving force in the evolution of the *BoS* family.

Unusual Hairpin-like Secondary Structure of BoS *Head Regions*

The potential of the head regions of *BoS* elements to form a conserved hairpin-like secondary structure is unusual for SINEs, as previously characterized SINEs derived from cellular RNAs usually adopt the secondary structures of their cognate cellular RNA. For example, the head regions of tRNA-derived SINEs usually fold into a cloverleaf-like secondary structure (Kachroo et al. 1995; Takahashi and Okada 2002; Yoshioka et al. 1993), whereas the head regions of 7SL RNA-derived SINEs fold into so-called "hammer head" (e.g., the B1 family) or "double hammer head" (e.g., the Alu family) structures (Labuda et al. 1991; Labuda and Zietkiewicz 1994; Sinnett et al. 1991). However, the head region of the rodent ID SINE RNA (ancestrally derived from tRNA^{ala}) was shown to form a stable hairpin-like secondary structure that is strikingly similar to that of *BoS* elements (Rozhdestvensky et al. 2001). In addition, although the secondary structure has not been described for several tRNA-derived SINEs from Brassicaceae (*AtSN1*/RAthE3, *AtSN2*/RAthE1, RAthE2, and S1), folding of their consensus sequences indicated that the RNA of their head regions can also form a relatively stable hairpin-like secondary structure (not shown). Taken together, these results suggest that the potential ability to form a hairpin-like structure in the head region may be shared by many tRNA-derived SINEs.

Comparison of distinct *BoS* subfamilies with divergent primary nucleotide sequences provided several lines of evidence suggesting that such a structure has been conserved during the evolution of *BoS* elements and may be functionally important. First, it is the most stable secondary structure formed for all 15 *BoS* subfamilies (~−30 to ~−45 kcal/mol, with an average of ~−35 kcal/mol) (Fig. 3). Second, conservation of this structure does not simply result from conservation of primary nucleotide sequence, as *BoS* subfamilies share on average only ~66% identity in this region. Rather, the base pairing necessary to form the hairpin structure appears to have been maintained by a strong selective pressure despite primary sequence divergence. For example, the head regions of *BoS_ab* and *BoS_b* RNAs differ by a 1-bp indel and 20 substitutions, of which 10 are compensatory changes that occur at five positions (e.g., A-to-G changes on one strand accompanied by T-to-C changes at corresponding positions on the opposite strand). The remaining 10 substitutions and 1-bp indel do not cause any disruption in the hairpin, as they either are in unpaired regions or are A/G transitions when the corresponding nucleotide on the opposite strand is U (Fig. 3). As a result, the secondary structures of the head regions of *BoS_ab* and *BoS_b* RNAs are nearly identical.

What could be the functional significance of this structure? Mimicry of tRNA or 7SL RNA structure by SINEs is thought to be involved in the localization of SINE RNA in the close proximity of the translation apparatus where LINE-encoded proteins are being synthesized (Labuda and Zietkiewicz 1994). This reflects the necessity for SINE RNA to compete with LINE RNA, as LINE-encoded proteins exhibit a marked *cis* preference for the mRNA from which they are translated. The RNA of ID elements has been shown to interact with several proteins to form an RNA–protein (RNP) complex (West et al. 2002). Interestingly, one component of this RNP complex is poly(A) binding protein (PABP), a regulator of translation initiation, suggesting that the RNP complex may be involved in the targeting of ID RNA to the translation apparatus (Muddashetty et al. 2002; West et al. 2002). Furthermore, incorporation of ID RNA into the RNP complex appears to be dependent on the secondary structure, rather than the primary sequence, of the RNA head region (West et al. 2002). Considering the striking similarity between the secondary structure of *BoS* and ID head regions, it is possible that the conserved hairpin structure in *BoS* RNA (as well as in *AtSN1*/RAthE3, *AtSN2*/RAthE1, RAthE2, and S1) is also involved in their targeting to the translation apparatus by a similar mechanism.

BoS *Elements Emerged and Were Active Over a Long Period of Time*

The *BoS* family must be relatively old because family and subfamily members display a high level of sequence divergence. In addition, the presence of related elements in other species in the Brassicaceae family indicates that *BoS* elements predate species divergence. The age of a family or subfamily of retroelements can be estimated by assuming that nucleotide substitutions accumulate at random after retrotransposition and that 1% sequence divergence corresponds to one million years. Using these criteria, the oldest subfamily, *BoS_e*, may have emerged roughly ~20 Mya, as its members share ~80% identity, whereas members of the youngest subfamily, *BoS_ai2* (97.5% sequence identity), originated about ~2–3 Mya (Table 1). It should be noted that the age of *BoS* subfamilies determined this way remains a rough estimate because, for example, the nucleotide substitution rate in transposable elements is often different from that in nuclear genes, and different genomes or even different regions in the same genome have different substitution rates. Nevertheless, it is apparent that different *BoS* subfamilies emerged and were most active during various evolutionary time periods. Interestingly, of the 15 *BoS* subfamilies, 5 have ~80% to ~85% intrasubfamily sequence identity, whereas members of 8 other subfamilies are ~93% to ~98% identical (Table 1). Thus, the activity of most *BoS* subfamilies appears to be restricted to two periods, one corresponding to the emergence and early diversification of the Brassicaceae family (~15–20 Mya) and the other much more recent (~2–7 Mya).

It is not known whether any *BoS* elements are active in *B. oleracea*. For an element to be active it must be transcribed because the RNA is the transposition intermediate. To date, Blast searches have not turned up any *BoS* ESTs from this species. However, several ESTs that appear to be *bona fide BoS* transcripts were found in *B. napus* (Table 2). Taken together, these results indicate that different *BoS* subfamilies have emerged and were active during various evolutionary time periods and that active *BoS* elements may exist in some Brassica species.

Mosaic BoS *Subfamilies*

Formation of mosaic SINEs through sequence exchange between existing elements appears to be a common phenomenon in plants and animals, and such a process has been proposed as a way to generate new and more active elements (Kass et al. 1995; Lenoir et al. 1997; Takahashi and Okada 2002; Zietkiewicz and Labuda 1996). In this study, sequence comparison among *BoS* subfamilies indicates that formation of mosaic elements has occurred frequently during the evolution of the *BoS* family, with 7 of the 15 subfamilies either participating in or formed by sequence exchange. As described above, the age of a particular *BoS* subfamily can be estimated based on intrasubfamily sequence divergence. In several instances, the progenitor–progeny relationship inferred by intersubfamily sequence comparisons of *BoS* subfamilies is supported by the age of the subfamilies involved. For example, the mosaic subfamily *BoS_ab* is younger than both of its progenitors, *BoS_a* and *BoS_b*. Similarly, the mosaic subfamily *BoS_ai2* is younger than both *BoS_a* and *BoS_i2*. The situation of *BoS_c* and *BoS_f* is less clear, as a third subfamily involved in the sequence exchange was not identified. However, since *BoS_c* is younger than *BoS_f*, it is likely that *BoS_c* acquired its internal region from *BoS_f*. Finally, two of the three mosaic subfamilies (*BoS_ab* and *BoS_c*) have the highest copy numbers among all subfamilies, providing support for the notion that some mosaic elements may be more successful than the elements they derive from.

Two mechanisms, involving either gene conversion or template switching, have been proposed to explain the formation of mosaic SINEs. The gene conversion model posits that mosaic SINEs form through recombination between the genomic copies of two different SINEs (Kass et al. 1995; Lenoir et al. 1997; Zietkiewicz and Labuda 1996). In contrast, the template switching model postulates that during reverse transcription of the RNA of one SINE into cDNA, the reverse transcriptase switches its template to the RNA of a second SINE (Takahashi and Okada 2002). A key difference distinguishing the two models is the requirement for SINE transcription and reverse transcription for template switching to occur but not for gene conversion which occurs between genomic copies. The results from this study indicate that, in the case of *BoS*, the template switching model is more likely, because all donor *BoS* subfamilies (*BoS_a*, *BoS_b*, *BoS_i2*, and *BoS_f*; see above) have been active relatively recently. In contrast, of the six older subfamilies (*BoS_e, g, h, i1, j*, and *k*, with less than 90% sequence identity; see Table 1) where recent activity is unlikely, none were found to be involved in mosaic SINE formation, either with another old subfamily or with one that was recently active. Furthermore, in each case where a mosaic subfamily was formed, the age of the two donor subfamilies are very similar (Table 1), suggesting that they may have been simultaneously active.

In summary, this study identified *BoS* as the most abundant SINE family in Brassicaceae and one of the most diverse families from any organism. In addition to the typical modular primary structure and target site preference of SINEs, *BoS* elements have several unusual features, including the distinct hairpin-like secondary structure formed by the head regions of their RNAs and the frequent formation of new, more successful mosaic subfamilies through sequence exchange between existing subfamilies. Finally, with 15 distinct subfamilies that were active during various time periods and the availability of the sequence information at a large number of insertion sites, *BoS* elements should provide useful phylogenetic makers for future analyses of the evolutionary relationships among species in the Brassicaceae family.

Acknowledgments. Preliminary sequence data were obtained from The Institute for Genomic Research Web site at http://www.tigr.org. Sequencing of *Brassica oleracea* was funded by the National Science Foundation. This work was supported by grants from the National Science Foundation and the National Institutes of Health to S.R.W.

References

Capy P, Bazin C, Higuet D, Langin T (1998) Dynamics and evolution of transposable elements. Springer-Verlag, Austin, TX

Deragon JM, Landry BS, Pelissier T, Tutois S, Tourmente S, Picard G (1994) An analysis of retroposition in plants based on a family of SINEs from *Brassica napus*. J Mol Evol 39:378–86

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

Doolittle RF, Feng D-F, Johnson Ms, McClure MA (1989) Origins and evolutionary relationships of retroviruses. Q Rev Biol 64:1–30

Eickbush TH (1992) Transposing without ends: the non-LTR retrotransposable elements. New Biol 4:430–440

Hofacker IL (2003) Vienna RNA secondary structure server. Nucleic Acids Res 31:3429–3431

Jurka J (1997) Sequence patterns indicate an enzymatic involvement in integration of mammalian retroposons. Proc Natl Acad Sci USA 94:1872–1877

Kachroo P, Leong SA, Chattoo BB (1995) Mg-SINE: a short interspersed nuclear element from the rice blast fungus, Magnaporthe grisea. Proc Natl Acad Sci USA 92:11125–11129

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

Kass DH, Batzer MA, Deininger PL (1995) Gene conversion as a secondary mechanism of short interspersed element (SINE) evolution. Mol Cell Biol 15:19–25

Labuda D, Zietkiewicz E (1994) Evolution of secondary structure in the family of 7SL-like RNAs. J Mol Evol 39:506–518

Labuda D, Sinnett D, Richer C, Deragon JM, Striker G (1991) Evolution of mouse B1 repeats: 7SL RNA folding pattern conserved. J Mol Evol 32:405–414

Lander ES, Linton LM, Birren B, Nusbaum C, Zody MC, Baldwin J, et al. (2001) Initial sequencing and analysis of the human genome. Nature 409:860–921

Lenoir A, Cournoyer B, Warwick S, Picard G, Deragon JM (1997) Evolution of SINE S1 retroposons in Cruciferae plant species. Mol Biol Evol 14:934–941

Lenoir A, Lavie L, Prieto JL, Goubely C, Cote JC, Pelissier T, Deragon JM (2001) The evolutionary origin and genomic organization of SINEs in Arabidopsis thaliana. Mol Biol Evol 18:2315–2322

Lowe TM, Eddy SR (1997) tRNAscan-SE: a program for improved detection of transfer RNA genes in genomic sequence. Nucleic Acids Res 25:955–964

Luan DD, Korman MH, Jakubczak JL, Eickbush TH (1993) Reverse transcription of R2Bm RNA is primed by a nick at the chromosomal target site: a mechanism for non-LTR retrotransposition. Cell 72:595–605

Muddashetty R, Khanam T, Kondrashov A, Bundman M, Iacoangeli A, Kremerskothen J, Duning K, Barnekow A, Huttenhofer A, Tiedge H, Brosius J (2002) Poly(A)-binding protein is associated with neuronal BC1 and BC200 ribonucleoprotein particles. J Mol Biol 321:433–445

Myouga F, Tsuchimoto S, Noma K, Ohtsubo H, Ohtsubo E (2001) Identification and structural analysis of SINE elements in the Arabidopsis thaliana genome. Genes Genet Syst 76:169–179

Ogiwara I, Miya M, Ohshima K, Okada N (1999) Retropositional parasitism of SINEs on LINEs: identification of SINEs and LINEs in elasmobranchs. Mol Biol Evol 16:1238–1250

Ohshima K, Koishi R, Matsuo M, Okada N (1993) Several short interspersed repetitive elements (SINEs) in distant species may have originated from a common ancestral retrovirus: characterization of a squid SINE and a possible mechanism for generation of tRNA-derived retroposons. Proc Natl Acad Sci USA 90:6260–6264

Ohshima K, Hamada M, Terai Y, Okada N (1996) The 3′ ends of tRNA-derived short interspersed repetitive elements are derived from the 3′ ends of long interspersed repetitive elements. Mol Cell Biol 16:3756–3764

Okada N (1991a) SINEs. Curr Opin Genet Dev 1:498–504

Okada N (1991b) SINEs: short interspersed repeated elements of the eukaryotic genome. Trends Ecol Evol 6:358–361

Okada N, Hamada M (1997) The 3′ ends of tRNA-derived SINEs originated from the 3′ ends of LINEs: a new example from the bovine genome. J Mol Evol 44:S52–S56

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

Rozhdestvensky TS, Kopylov AM, Brosius J, Huttenhofer A (2001) Neuronal BC1 RNA structure: evolutionary conversion of a tRNA(Ala) domain into an extended stem-loop structure. RNA 7:722–730

Schmid CW (1996) Alu: structure, origin, evolution, significance and function of one-tenth of human DNA. Prog Nucleic Acid Res Mol Biol 53:283–319

Shen MR, Batzer MA, Deininger PL (1991) Evolution of the master Alu gene(s). J Mol Evol 33:311–320

Sinnett D, Richer C, Deragon JM, Labuda D (1991) Alu RNA secondary structure consists of two independent 7 SL RNA-like folding units. J Biol Chem 266:8675–8678

Swofford DL (1999) PAUP*: phylogenetic analysis using parsimony and other methods Sinauer, Sunderland, MA

Takahashi K, Okada N (2002) Mosaic structure and retropositional dynamics during evolution of subfamilies of short interspersed elements in African cichlids. Mol Biol Evol 19:1303–1312

Tatout C, Lavie L, Deragon JM (1998) Similar target site selection occurs in integration of plant and mammalian retroposons. J Mol Evol 47:463–470

Terai Y, Takahashi K, Okada N (1998) SINE cousins: the 3′ end tails of the two oldest and distantly related families of SINEs are descended from the 3′ ends of LINEs with the same genealogical origin. Mol Biol Evol 15:1460–1471

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

Umeda M, Ohtsubo H, Ohtsubo E (1991) Diversification of the rice Waxy gene by insertion of mobile DNA elements into introns. Jpn J Genet 66:569–586

Weiner AM (1980) An abundant cytoplasmic 7S RNA is complementary to the dominant interspersed middle repetitive DNA sequence family in the human genome. Cell 22:209–218

West N, Roy-Engel AM, Imataka H, Sonenberg N, Deininger P (2002) Shared protein components of SINE RNPs. J Mol Biol 321:423–432

Xiong Y, Eickbush TH (1990) Origin and evolution of retroelements based upon their reverse transcriptase sequences. EMBOL J 9:3353–3362

Yoshioka Y, Matsumoto S, Kojima S, Ohshima K, Okada N, Machida Y (1993) Molecular characterization of a short interspersed repetitive element from tobacco that exhibits sequence homology to specific tRNAs. Proc Natl Acad Sci USA 90:6562–6566

Zietkiewicz E, Labuda D (1996) Mosaic evolution of rodent B1 elements. J Mol Evol 42:66–72