Pak. J. Bot., 50(2): 749-756, 2018.

IDENTIFICATION AND INSERTION POLYMORPHISMS OF SHORT INTERSPERSED NUCLEAR ELEMENTS (SINES) IN *BRASSICA* GENOMES

FAISAL NOUROZ[1,2*], SHUMAILA NOREEN[2], MUHAMMAD NAVEED[3], KAFEEL AHMAD[4] AND J.S. HESLOP-HARRISON[2]

[1]*Department of Botany, Hazara University Mansehra, Pakistan*
[2]*Department of Genetics and Genome Biology, University of Leicester, United Kingdom*
[3]*Department of Biotechnology, University of Central Punjab, Lahore, Pakistan*
[4]*Centre of Biotechnology and Microbiology, University of Peshawar, Pakistan*
[*]*Corresponding author's email: faisalnouroz@gmail.com. +92 997414168*

Abstract

The non-LTR retrotransposons (retroposons) are abundant in plant genomes including members of *Brassicaceae*. Of the retroposons, long interspersed nuclear elements (LINEs) are more copious followed by short interspersed nuclear elements (SINEs) in sequenced eukaryotic genomes. The SINEs are short elements and ranged from 100-500 bps flanked by variable sized target site duplications, 5′ tRNA region with polymerase III promoter, internal tRNA unrelated region, 3′ LINEs derived region and a poly adenosine tail. Different computational approaches were used for the identification and characterization of SINEs, while PCR was used to detect the SINEs insertion polymorphisms in various *Brassica* genotypes. Ten previously unidentified families of SINEs were identified and characterized from *Brassica* genomes. The structural features of these SINEs were studied in detail, which showed typical SINE features displaying small sizes, target site duplications, head regions, internal regions (body) of variable sizes and a poly (A) tail at the 3′ terminus. The elements from various families ranged from 206-558 bp, where *BoSINE2* family displayed smallest SINE element (206 bp), while larger members belonged to BoSINE9 family (524-558 bp). The distribution and abundance of SINEs in various *Brassica* species and genotypes (40) at a particular site/locus were investigated by SINEs based PCR markers. Various SINE insertion polymorphisms were detected from different genotypes, where higher PCR bands amplified the SINE insertions, while lower bands amplified the pre-insertion sites (flanking regions). The analysis of *Brassica* SINEs copy numbers from 10 identified families revealed that around 860 and 1712 copies of SINEs were calculated from *B. rapa* and *B. oleracea* Whole-genome shotgun contigs (WGS) respectively. Analysis of insertion sites of *Brassica* SINEs revealed that the members from all 10 SINE families had shown an insertion preference in AT rich regions. The present analysis will be helpful in SINEs annotation in *Brassica* and their identification from related genera. The SINE based molecular markers will also assist to study the diversity among closely related genotypes and cultivars of various species.

Key words: Brassica, SINEs, Retrotransposons, Families, Insertions, Genotypes.

Introduction

Of the transposable elements (TEs) or mobile genetic elements (MGEs), the retrotransposons are more copious in sequenced eukaryotic genomes. In maize 70% of the nuclear DNA is contributed by retrotransposons (SanMiguel & Bennetzen, 1998; Nouroz *et al*., 2015a), a typical result for many species. The genomic and extra-chromosomal copies of retrotransposons proliferate by an RNA intermediate copied into DNA by reverse transcriptase (Feschotte *et al.*, 2002; Kapitonov and Jurka, 2008; Kapitonov *et al.*, 2009). Retroelements have been categorized on the basis of phylogeny of their reverse transcriptase (RT), *gag-pol* domain organization, proliferating devices and structural features into long terminal repeat retrotransposons (LTRs), Non-LTRs retrotransposons which includes long interspersed nuclear elements (LINEs), short interspersed nuclear elements (SINEs), Dictyostelium related sequences (DIRs) and Penelope-like elements (Wicker *et al.*, 2007).

The LINEs includes several families such as CR1, CRE, I, Jockey, L1, NeSL, R2, R4, RandI, Rex1, RTE and Tx1, while SINEs include superfamilies as SINE1, SINE2 and SINE3. The SINEs are small Non-LTR retrotransposons ranging in size from 100-500 bp having internal promoters for RNA polymerase III (Okada *et al.*, 1997; Kapitonov & Jurka, 2003). They have a complex structure, with target site duplications (TSDs) at both ends, a 5′ region similar to tRNA or 7SL RNA genes with polymerase III promoter, internal non-tRNA region of variable sizes, 3′ LINE derived region, and a short segment of A or T at their 3′ terminal end. The SINEs are non-autonomous elements, as they lack their own reverse transcriptase protein necessary for transposition. Despite their non-autonomous nature they are mobile elements and utilize the enzymatic machinery of LINEs for their transposition. Like LINEs, they also generate TSDs upon integration to a new site (Kapitonov & Jurka, 2003; Deragon & Zhang, 2006; Kramerov & Vassetzky, 2011). The tRNA region of the SINEs displays two well conserved sequence motifs called box A and box B, which served as internal promoter for the transcription of SINEs by RNA polymerase III. The SINEs are non-autonomous elements but are mobile and utilize the enzymatic machinery of LINEs for their transposition. *TS* family of SINEs was detected as highly repetitive family among *Solanaceae* crops like *Capsicum annum*, *Lycopersicon esculentum* and *Solanum tuberosum* (Pozueta-Romero *et al.*, 1998).

Various TEs were recently identified and characterized from *Brassica* genomes such as LTR retrotransposons (Nouroz *et al*., 2015a), LINEs (Nouroz *et al*., 2017a), DNA transposons such as Mutators, hATs (Nouroz *et al*., 2015b; Nouroz *et al*., 2015c), Harbingers and CACTA (Nouroz *et al*., 2016; Nouroz *et al*., 2017b). SINE elements named *S1* were characterized previously in *Brassica* (Goubely *et al.*, 1999), which were ~170 bp in

size and widely distributed among members of *Brassicaceae*. Another *B. oleracea* specific SINE family (*BoS*) is distributed in *Brassica* having ~4290 estimated copies from different families (Deragon & Zhang, 2006). The *Au* SINEs are very diverse elements detected in *Gramineae* (*Aegilops umbellulata, Triticum aestivam, Zea mays*), *Solanaceae* (*Nicotina tabacum, Solanum esculentum*), *Fabaceae* (*Medicago truncutula, Lotus japonicus, Glycine max*) and others (Fawcett *et al.*, 2006). A survey of SINEs in the rice genome led to the identification of 13487 copies of SINEs, of which *F524* was the most active SINE in rice with highest (119) intact copies. *SINE3_OS* have above 7000 copies but only 10 intact elements were identified, the remaining are all truncated copies (Khan *et al.*, 2011).

The present study aimed to identify the SINE elements in sequenced *Brassica* BACs and characterize their diversity and insertion polymorphisms across *Brassica* germplasms.

Material and Methods

Plant material for *Brassica:* The DNAs from 40 *Brassica* accessions/genotypes from six *Brassica* species were used in the present study listed in Table 1. The seeds were grown in a green house at Department of Genetics and Genome Biology (formerly Department of Biology), University of Leicester, UK. For DNA extraction for PCR analysis, the standard CTAB method (Doyle & Doyle, 1990) was applied.

Computational analysis and data mining: The full length SINEs were identified by dot plot comparison of homeologous BAC sequences in J. Dotter program (Sonnhammer & Durbin, 1995) and were considered as reference elements, which were than run against the *Brassica rapa* and *Brassica oleracea* Whole-genome shotgun contigs (WGS) database in NCBI. The sequences showing > 70% of the query coverage and identity in their entire lengths were retrieved and analysed. The number of the TSDs at terminals and the poly(A) tails at 3′ ends were counted manually.

The SINEs identified by dot plot analysis were characterized on the basis of TSDs, polyA tail and tRNA head regions. For the comparison of tRNA head region of the SINEs, the tRNA sequences were retrieved from *Arabidopsis* Genomic tRNA Database (http://gtrnadb.ucsc.edu/Athal) (Chan & Lowe, 2009). Frequency plots indicating the insertional preference of SINEs families were generated in WebLogo). The Repbase database of eukaryotic TEs (http://www. girinst.org/repbase/ index.html) (Jurka *et al.*, 2005) and Repeat masker of Censor software (http://www. girinst.org/censor/ index.php) implemented in Genetic Information Research Institute (GIRI) were used to characterize the SINEs on homology basis with known elements. Elements that failed to be characterized by the above searches against TE databases were characterized by visual inspection on the basis of their hallmark motifs such as TSDs, poly(A) tail and tRNA head regions. The identified SINEs were classified into respective families by the recommendations of Wicker *et al.*, 2007 on homology basis.

Naming of SINE elements: The names to the novel elements were given on the recommendations of Nouroz *et al.*, (2015a, 2017a). The names were given as *GsXXXXN*, where '*G*' represent genus, small letter '*s*' represent species names, *XXXX* indicate SINEs superfamily and '*N'* indicate the number. Thus *BoSINE1* is indicating the first family of *B. oleracea* SINE.

Table 1. List of *Brassica* species with accessions names used in the present study. ND: Not Determine.

No.	Species	Accession name	No.	Species	Accession name
1.	*B. rapa chinensis*	Pak Choy	21.	*B. juncea*	Tsai Sim
2.	*B. rapa pekinensis*	Chinese Wong Bok	22.	*B. juncea*	W3
3.	*B. rapa chinensis*	San Yue Man	23.	*B. juncea*	Giant Red Mustard
4.	*B. rapa rapa*	Hinona	24.	*B. juncea*	Varuna
5.	*B. rapa rapa*	Vertus	25.	*B. napus*	New
6.	*B. rapa*	Suttons	26.	*B. napus oleifera*	Mar
7.	*B. nigra*	ND	27.	*B. napus biennis*	Last and Best
8.	*B. nigra*	ND	28.	*B. napus napo*	Fortune
9.	*B. nigra*	ND	29.	*B. napus*	Drakker
10.	*B. juncea*	NARC-I	30.	*B. napus*	Tapidor
11.	*B. juncea*	NATCO	31.	*B. carinata*	Addis Aceb
12.	*B. juncea*	NARC-II	32.	*B. carinata*	Patu
13.	*B. oleracea*	De Rosny	33.	*B. carinata*	Tamu Tex-sel Greens
14.	*B. oleracea*	Kai Lan	34.	*B. carinata*	Mbeya Green
15.	*B. oleracea*	Early Snowball	35.	*B. carinata*	Aworks-67
16.	*B. oleracea italica*	Precoce Di Calabria	36.	*B. carinata*	NARC-PK
17.	*B. oleracea capitata*	Cuor Di Bue Grosso	37.	*B. napus* x *B. nigra*	ND
18.	*B. oleracea*	ND	38.	*B. carinata* x *B. rapa*	ND
19.	*B. juncea*	Kai Choy	39.	*B. napus* x *B. nigra*	ND
20.	*B. juncea*	Megarrhiza	40.	*B. napus* x *B. nigra*	ND

Table 2. List of primers, their sequences and sizes of the expected products to amplify the SINEs in *Brassica*.

Sr. No.	Family	TE size	Product size	Primer name	Primer sequence
1.	*BoSINE2*	219	365	BoSINE2F	GAACAAGAAAAATGCAGGG
				BoSINE2R	CGTACCATCACATCTCTTTC
2.	*BoSINE3*	272	585	BoSINE3F	TTCGTTCAAGTTTGATGCCA
				BoSINE3R	AAAGATCCTCACTGGAATCA
3.	*BoSINE9*	524	735	BoSINE9F	AGCTATTACCATGTCGTTCC
				BoSINE9R	ACATAACATTGATACTCCGC
4.	*BrSINE10*	376	615	BrSINE10F	CAAACACTACAAGTGAATAC
				BrSINE10R	GCAAGGTGGAGAAGATAAG

Table 3. Full length SINEs identified by comparative dot plot analysis of *Brassica* BAC sequences.

No.	Reference elements	Family	BAC accession	Species	Size	TSD	Poly (A) Tail	GC%
1.	*BoSINE1-1*	*BoSINE1*	EU642504.1	*B. oleracea*	216	14	CAAAAAAAAAAAAAAAAAA	52.0
2.	*BoSINE2-1*	*BoSINE2*	EU642504.1	*B. oleracea*	219	18	CTTAAAAAAAA	48.4
3.	*BoSINE3-1*	*BoSINE3*	EU642504.1	*B. oleracea*	272	13	CAAAAAAAAA	47.1
4.	*BoSINE4-1*	*BoSINE4*	EU579455.1	*B. oleracea*	443	44	CAAAAAAAA	37.0
5.	*BoSINE5-1*	*BoSINE5*	AC240089.1	*B. oleracea*	225	04	CAAAAAAAA	37.3
6.	*BoSINE6-1*	*BoSINE6*	AC240089.1	*B. oleracea*	335	11	CAAAAAAAAAAAAAAAAAA	37.3
7.	*BoSINE7-1*	*BoSINE7*	AC240089.1	*B. oleracea*	401	08	CAAAAAAAAAAA	41.6
8.	*BoSINE8-1*	*BoSINE8*	AC240089.1	*B. oleracea*	484	13	CAAAAAAAAAAA	37.6
9.	*BoSINE9-1*	*BoSINE9*	EU642504.1	*B. oleracea*	524	11	CAAAAAAAAAA	48.1
10.	*BrSINE10-1*	*BrSINE10*	AC189298.1	*B. rapa*	376	13	CGTTAAAAAAAAA	41.0

Analysis of SINE copy numbers: The copy numbers of SINEs were calculated by blasting the SINE reference element sequences against *Brassica rapa* and *Brassica oleracea* Whole-genome shotgun contigs (WGS). The sequences with >70% query coverage and identity were counted after getting output from BLASTN searches. Only intact elements were counted, while partial elements and remnants were not included.

Polymerase chain reactions (PCRs): The degenerative primer pairs (Table 2) were designed from flanking regions of SINE insertion sites with Primer3 (http://frodo.wi.mit.edu/primer3/). Polymerase chain reaction (PCR) was used for the amplification of SINE fragments. Total volume of reaction mixture was 25 µl. The genomic DNA was used @ 50-75 ng/µl with 10X Kapa Taq buffer A (Kapa Biosystems, UK), additional 1.0 mM $MgCl_2$, 200-250 µM dNTP (2-2.5 mM; YORKBIO), 10 pmoles of each primer (SIGMA-ALDRICH) and 0.5-1 U of 5U/µl Taq polymerase (Kapa Biosystems, UK). The thermal cycling conditions were 3 min denaturation at 94˚C; 35 cycles of 1 min denaturation at 94˚C, 1 min annealing at 52-64˚C (primers dependent) and 1 min extension at 72˚C; than final 5 min extension at 72˚C. PCR products were separated by electrophoresis in 1% agarose gel with TAE buffer according to the standard protocols. Gels were stained with 1-2 µl ethidium bromide for the detection of DNA bands under UV illumination.

Results

Identification of novel SINE families in *Brassica* genome: Ten novel SINE insertions (Table 3) were identified from *Brassica* genome by comparison of homoeologous BAC sequences collected from the GenBank database. By comparing *B. rapa* and *B. oleracea* accessions (AC189298.1 x EU642504.1), a SINE insertion was detected in *B. rapa* and 4 SINEs in *Brassica oleracea*. Similarly, the comparison of *B. rapa* (AC155341.2) x *B. oleracea* (AC240089.1) and *B. rapa* (CU984545.1) x *B. oleracea* (EU579455.1) BAC accessions led to the identification of 4 and 1 SINE insertions respectively. The newly identified SINEs (reference) were used as query in BLASTN searches to identify other relatives residing in *Brassica* species. The sequences were considered as members of the same family, if they generate host TSDs, poly(A) tail at 3′ terminus and >70% coverage in entire lengths. The detail structural analysis resolved them into 10 different families. The families were named as *BoSINE1-BrSINE10* and are represented in figure 1.

Structural features of *Brassica* SINE families: The identified *Brassica* SINEs were small in sizes with typical SINE features displaying TSDs, head regions, internal regions (body) of variable sizes and a poly(A) tail at the 3′ terminus. The smallest SINE was a member of *BoSINE2* family and was 206 bp in size, while larger elements belong to BoSINE9 family. *BoSINE1* displayed 10 members, which ranged from 213-225 bp, flanked by TSDs of 7-14 bp and terminated by a 3′ poly(A) tail of 19-21 bp. The first SINE (*BoSINE1-1*) was identified as an insertion residing in *B. oleracea* (EU642504.1) sequence. The size of the *BoSINE1-1* was 216 bp including 14 by TSDs, and terminated by $C(A)_{18}$ at C-terminal end (Fig. 1; Table 3). *BoSINE2*, presents a low copy number family with members having sizes from 206-219 bp, flanked by TSDs of 13-18 bp and polyadenylation signals of 10-27 bp at their 3′ terminal end (Table 4). The first element (*BoSINE2-1*) from this family was identified in *B. oleracea* (EU642504.1), where a 219 bp insertion was found flanked by 18 bp TSDs and a tail terminating with $CTT(A)_{8.}$ The *BoSINE3* family represents the members ranging in sizes from 256-277 bp including TSDs of 10-17 bp and terminating by a $poly(A)_{9\text{-}11}$ tail. The well characterized member is a 272

bp *BoSINE3-1* having a 13 bp TSDs and 9 bp poly(A) tail. The members of family *BoSINE4* generally ranged in sizes from 361-397 bp with the exception of *BoSINE4-1* (Table 4). The elements were flanked by TSDs of 07-15 bp (except *BoSINE4-1*) and terminated with 8-34 bp poly(A) tail. *BoSINE4-1* is the first element detected (442 bp), which generates largest TSDs (42 bp).

The sequences from *BoSINE5* were mostly similar in sizes (225-229 bp), flanked by short TSDs (3-4 bp) and a poly(A) tail of 8-11 bp (Table 4). The first element was characterized from *B. oleracea* (AC240089.1) as a 225 bp insertion including 4 bp TSDs and 5′-CAAAAAAAA-3′ C-terminal tail. *BoSINE6* family represents 321-335 bp large members generating TSDs (05-11) and having poly(A) tail. *BoSINE6-1* is the first and well characterized member of the family with a size of 335 bp including 11 bp TSDs at both ends and an 18 bp polyadenylation tail (Table 3). A low copy number family *BoSINE7* is characterized by having representatives ranging in sizes from 392-401 bp, including TSDs (3-8 bp) and a poly(A)$_{11\text{-}13}$ tail (Table 4). The first identified member from the family is *BoSINE7-1* from *B. oleracea* (AC240089.1) residing as an insertion (Fig. 1). The element is 401 bp in size including 8 bp TSDs at both ends and polyadenylation signals of 8 nucleotides. The second largest family of *Brassica* SINEs *BoSINE8* represents diverse members dispersed in *B. rapa* and *B. oleracea* genomic sequences. The elements range in sizes from 480-506 bp including the host TSDs (5-13 bp) and poly(A)$_{11\text{-}28}$ tail adjacent to C-terminal end. Generally the terminal tail have 11-19 bp poly(A) stretch but few elements generate a longer stretch (21-27 bp). *BoSINE8-1* represents the first identified member of the family from *B. oleracea* (AC240089.1) accession with 13 bp TSDs and poly(A)$_{11}$ tail (Fig. 1; Tables 3,4).

A 524 bp insertion (*BoSINE9-1*) flanked by 11 bp TSDs and a poly(A) tail yielded no significant hits but the NCBI EST database yielded two sequences with >85% identity in their entire lengths. The retrieved sequences were *B. napus* cDNA, mRNA sequences and were designated as *BnSINE9-2* and *BnSINE9-3*. The elements were 558 bp in sizes generating 6 bp TSDs and a largest poly(A) tail comprising 50 adenine and a single guanine nucleotide. The copy number estimation in *B. rapa* (26) and *B. oleracea* (74) suggests that this is the lowest copy number family of SINEs studied in present work. The largest family is *BrSINE10* with 505 and 450 members in A and C-genomes respectively. *BrSINE10-1* was a 376 bp large SINE including 13 bp TSDs and a 5′-CGTTAAAAAAAAA-3′ tail. The *BrSINE10* family members ranged from 368-378 bp including TSDs (5-15 bp) and a poly(A) tail of 9-15 bp (Table 4).

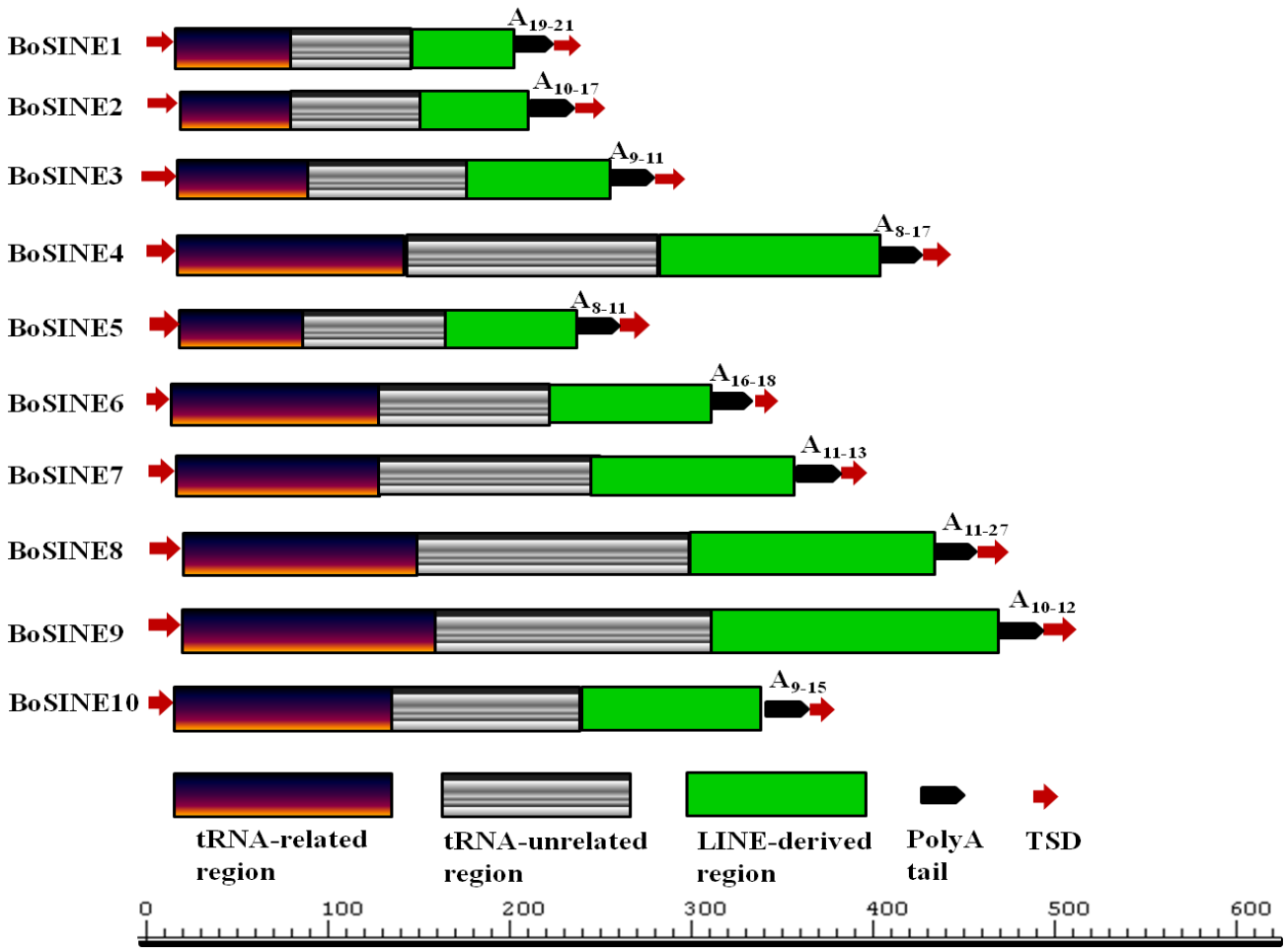

Fig. 1. Schematic representation of *Brassica* SINE families. SINEs are composed of a 5′ tRNA-related region (maroon), a internal tRNA-unrelated region (grey) and a 3′ LINE-related region (green). The variable sized TSDs are represented by arrows at both terminal ends. Scale below is given in base pairs.

Table 4. Average lengths, TSDs, Pre-tail motifs and estimated copy numbers of each SINEs family in *Brassica*. The name of the family is given on the basis of the first element identified in *Brassica*. WGS: Whole-genome shotgun contigs.

No.	Family name	Size of elements	TSDs	Pre-tail motifs	C (A)n	Copy no. in *B. rapa* (AA) WGS	Copy no. in *B. oleracea* (CC) WGS
1.	*BoSINE1*	213-225	07-14	TTATC	$C(A)_{19-21}$	114	190
2.	*BoSINE2*	206-219	13-18	TTATC	$C(A)_{10-17}$	44	130
3.	*BoSINE3*	256-277	10-17	TTTTC	$C(A)_{9-11}$	70	145
4.	*BoSINE4*	361-443	07-15/44	TTAGC	$C(A)_{8-17}$	42	64
5.	*BoSINE5*	225-229	03-04	TTTTC	$C(A)_{08-11}$	106	168
6.	*BoSINE6*	321-335	05-11	TTAGC	$C(A)_{16-18}$	85	114
7.	*BoSINE7*	392-401	03-08	TTACC	$C(A)_{11-13}$	84	108
8.	*BoSINE8*	480-506	05-13	TTGTC	$C(A)_{11-27}$	138	318
9.	*BoSINE9*	524-558	05-11	TGATC	$C(A)_{10-12/51}$	25	138
10.	*BrSINE10*	368-378	05-15	TCAGC	$C(A)_{9-15}$	152	336
					Total	**860**	**1712**

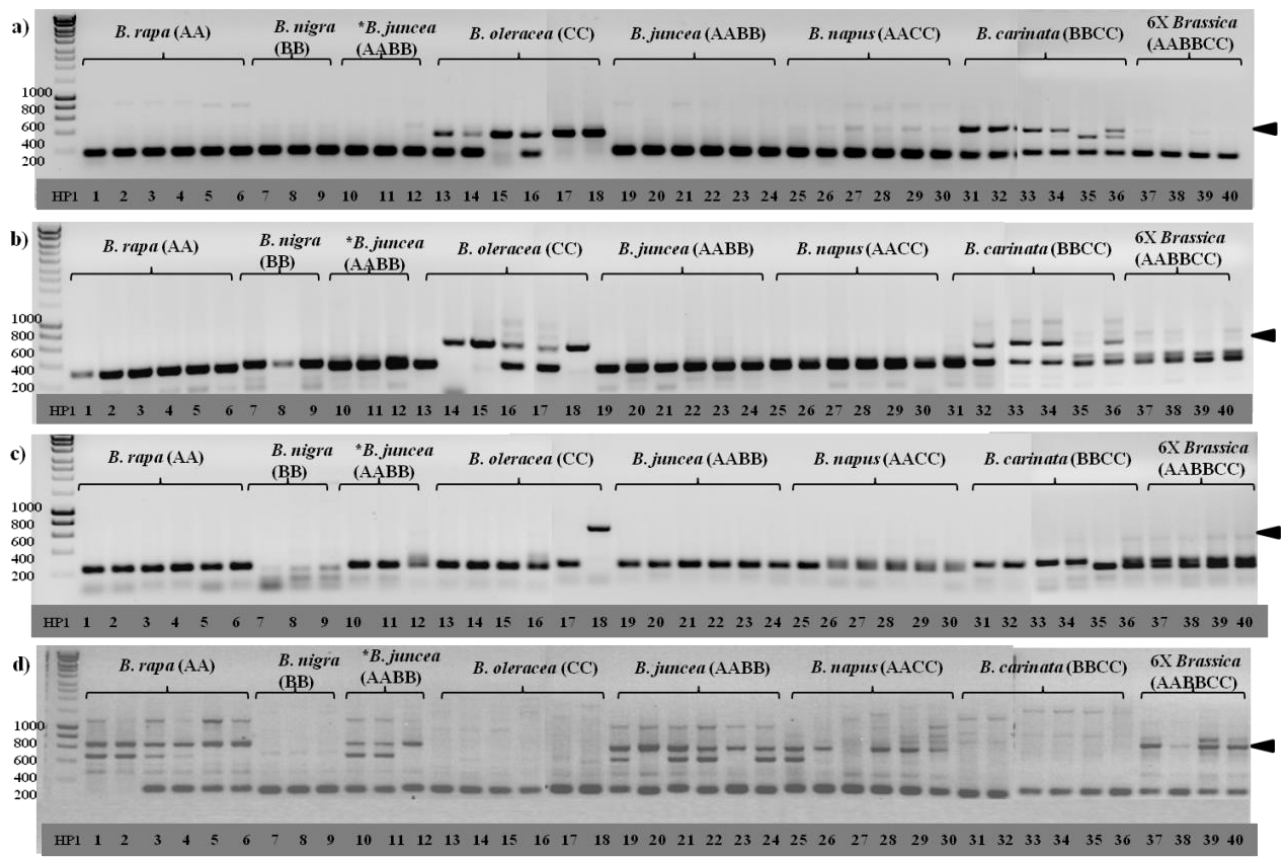

Fig. 2a-d. SINEs insertion polymorphism of various families in *Brassica* species and genotypes: a) *BoSINE2*; b) *BoSINE3*; c) *BoSINE9*; d) *BrSINE10*. Higher bands indicate amplification of SINE insertions, while lower bands represent the pre-insertion sites (no SINE insertions). The numbers below are indicating *Brassica* genotypes/accessions listed in Table 1.

PCR identification of SINE insertion polymorphisms in *Brassica* genotypes: The distribution and insertion polymorphism of SINEs in various *Brassica* species and genotypes was investigated by SINEs based PCR markers (Fig. 2a-d). A total of 40 *Brassica* genotypes were tested for the presence/absence of SINEs at a particular site/locus. Higher bands were amplified with insertions and lower bands amplifying the pre-insertion sites (flanking regions). Four set of primers (Table 2) were used to amplify four SINE families among various *Brassica* species. *BoSINE2-1* (219 bp) amplified by BoSINE2F and BoSINE2R yielded larger (insertion) and smaller (pre-insertion site) bands in various *Brassica* genotypes (Fig. 2a). Six *B. oleracea* and six *B. carinata* accessions amplified the higher bands having *BoSINE2-1* insertions as well as lower bands (without insertion), which indicate their heterozygous genomic nature. All the other *Brassica* SINEs amplified the lower bands with pre-insertional sites (Fig. 2a). A 272 *BoSINE3-1* was tested for its presence in various *Brassica* genotypes. The results showed its amplification from all six *B. oleracea* and six *B. carinata* cultivars. Weak bands were detected in 3 *Brassica* hexaploids (*B. napus* x *B. nigra*) (Fig. 2b). The *B. rapa*, *B. nigra*, *B. juncea* and *B. napus* cultivars amplified pre-insertional sites only.

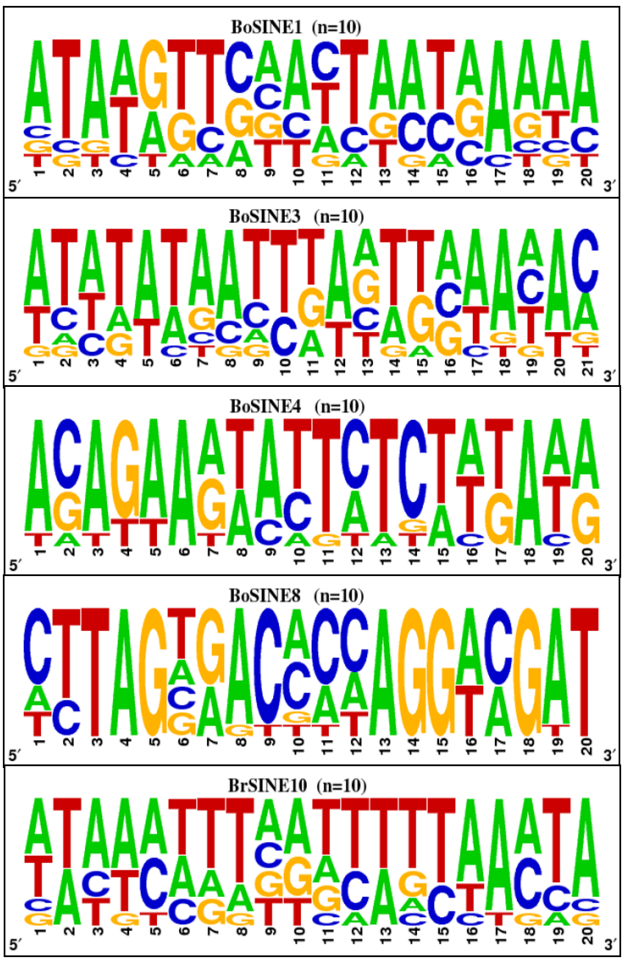

Fig. 3. Frequency plot indicating the insertion preference of five SINE families into AT rich regions. The WebLogo indicates the SINE preference for AT rich regions created by using 20 bp of flanking regions of the SINE insertions for each family.

The primers for *BoSINE9-1* (524 bp) were designed from the flanking regions and were tested against 40 *Brassica* genotypes. Interestingly, only *B. oleracea* accession GK97361 produced the expected product size (735 bp). The amplicon was sequenced and aligned with *BoSINE9-1* achieving >98% identity. All other *Brassica* accessions amplify the pre-insertion sites (~210 bp) (Fig. 2c). The primers designed from flanking regions of *BrSINE10-1* confirmed the abundance of *BrSINE10* family in various *Brassica* genotypes except *B. oleracea* (CC) and its allotetraploid *B. carinata* (BBCC) (Fig. 2d). The amplification of the *BrSINE10-1* was seen in *B. rapa* (AA) and the their allotetraploids (AABB, AACC). Many of the genotypes amplified additional bands of varied sizes concluding that multiple copies of the element are dispersed in *B. rapa* and its allotetraploid genomes.

SINE copy numbers from *Brassica* whole-genome shotgun contigs (WGS): SINEs are diverse retroposons present in the members of *Brassicaceae*. The total numbers of SINEs calculated from Whole-genome shotgun contigs (WGS) of *B. rapa* and *B. oleracea* were 860 and 1712 respectively. The copy number of each SINE family was analysed and low, middle and high copy numbered families were identified (Table 4). *BrSINE10* is the largest and highly diverse family of SINEs with 252 and 336 copies in *B. rapa* and *B. oleracea* genome respectively. *BoSINE8* is considered as second abundant family displaying 138 and 318 copies from A and C-genome *Brassica* respectively. *BoSINE9* family displayed 24 and 138 copies, while *BoSINE4* displayed 42 and 65 copies in A and C-genome *Brassica* respectively (Table 4).

Discussion

The results here showed the value of comparison of BAC sequences for identification of SINE-like TEs on the basis of their activity and homology. Previous methods have required assumptions about motifs and structures of these elements; but the present results showed that currently adopted method was more efficient in identification of most SINEs, which otherwise were very difficult to identify. This will be valuable in annotation and assisting in assembly of whole genome shotgun (WGS) sequencing data in the future. Current WGS approaches have difficulty in assembling LINE and SINE rich regions of the genomes, even using paired-end strategies, where long and duplicated elements, particularly when heterozygous, prevent conting ends from being overlapped unequivocally. In current analysis, 10 SINE families were identified and characterized from *Brassica* genome. The structural features and distribution of the elements were studied in detail by computational and molecular approaches. The analysis confirmed that SINEs are a diverse group of TEs scattered among *Brassica* genome. The SINEs studied from *Brassica* and *Arabidopsis* genomes display a conserved motif upstream to their poly(A) tail at 3′ terminus. The motif is generally AT rich and is highly conserved within the family members and across various families (Tables 3, 4). The structural analysis of other known *Brassicaceae* SINEs, including the S1, AtSN1/RAtheE3 from *Brassica*, AtSN2/RAtheE1 and RAthE2 from *Arabidopsis thaliana* showed similar conserved motif upstream to poly(A) tail at 3′ terminus (Deragon *et al.*, 1994; Lenoir *et al.*, 2001; Myouga *et al.*, 2001). This suggests that all the SINEs from *Brassicaceae* share more or less similar motifs at their pre-tail ends. The most characterized *BoS* SINE family from *Brassicaceae* exhibits a conserved motif (TTATC) upstream to 3′ terminal end (Zhang & Wessler, 2005).

The SINEs have shown an insertional preference to a heavily populated AT rich regions. To determine the SINEs insertional preference, SINE insertions with extra 20 bp flanking nucleotides on both ends were collected and aligned. Analysis of insertion sites of *Brassica* SINEs revealed that the members from 10 SINE families have shown an insertion preference in AT rich regions (Fig. 3). The *BoS* elements previously identified from *Brassica* also have shown insertional preference in AT rich regions (Zhang & Wessler, 2005). The SINE families investigated in this study are considered as old families, which were present before the separation of *Arabidopsis-Brassica* species. This can be confirmed by the high similarity of *Brassica* SINEs with the *Arabidopsis* genomes (~75%). The BLASTN searches against the GenBank database retrieved many sequences from *A. thaliana* and *A. lyrata*. Some of the hits from *Arabidopsis* showed very high sequence similarity with *Brassica* SINEs suggesting their

common origin from the same ancestor. *Brassica* SINEs were considered as old as the divergence of *Arabidopsis-Brassica*, which is believed to occur 16-19 million years ago (Myo) from a common ancestor (Deragon & Zhang, 2006). *BrSINE10* is considered to be the youngest family due to high homology within its members, as with the passage of time, nucleotide variations are observed. *BoSINE1*, *BoSINE3*, *BoSINE4*, and *BoSINE8* are considered as middle aged families while *BoSINE9* is considered to be recently introduced due to fewer copies and high homology (84-88%) between sequences. The previously identified *Brassica BoS* family is also an old family, whose members are dispersed among various *Brassica* species. It is thought the oldest members have diverged ~20 Mya, whereas the youngest members have originated ~2-3 million years ago (Zhang & Wessler, 2005).

SINEs can be used as molecular markers to investigate the evolutionary relations of species or to trace phylogeny. The first approach is the identification of a specific SINE family in species by PCR analysis or dot hybridization, where species displaying the presence of a specific SINE family are treated as close to each other than to other species, which lack them. The second approach is the site specific insertional polymorphism, where the PCR primers are designed from the common flanking regions around SINEs. The species having SINE insertions generate higher products, while those who lack generate the shorter products (Deragon and Zhang, 2006). Similar methodology was adopted in present work to observe the presence/absence of SINEs at various loci. Several SINE based insertion polymorphisms were observed, which represents their variable proliferating activity in various *Brassica* genotypes (Figs. 2a-d). Species and their genotypes/cultivars sharing the SINE insertions are considered to be close as compared to others, who lack them (Kramerov & Vassetzky, 2011). Thus *B. oleracea* (CC) and *B. carinata* (BBCC) are closer to each other due to site specific amplification of few SINEs as compared to *B. juncea* (AABB), which failed to amplify them. The results revealed that the SINEs can be used as molecular markers in evolutionary studies and to trace the phylogeny among various species and their genotypes.

Conclusions

Non-LTR retrotransposons or retroposons are present in enormous numbers in all eukaryotic genomes. Despite the progress in our understanding of retroposon biology, many aspects remain unclear, especially the identification of SINEs in sequenced genomes due to their short sizes. We set the trends in the field with the identification and characterization of SINEs among *Brassicaceae*, with an emphasis on their distribution in various *Brassica* genotypes. The insertional polymorphisms of SINEs by molecular methodology indicate their absence in some species and genotypes, while presence in others. The analysis will help in annotation and characterization of SINEs like retroposons from *Brassica* and other plant genera. These SINE based loci specific genetic markers will be highly valuable to study the diversity among closely related cultivars and varieties.

Acknowledgements

The work was conducted in highly advanced laboratories at Department of Genetics and Genome Biology (formerly Department of Biology/Genetics), University of Leicester, UK. We are thankful to Dr. Richard Gorner, Dr. Trude Schwarzacher, Dr. Mateus Mondin, Dr. John Bailey, Jeans Liggins, Dr. Niaz Ali, Dr. Farah and all staff for their support during the work. The seeds were provided by Warwick Research Centre, Warwick, UK and DNA of 4 hexaploid species was a gift from Dr. Xian Hong Ge, University of Huazhong Agricultural University, Wuhan, China.

References

Chan, P.P. and T.M. Lowe. 2009. GtRNAdb: a database of transfer RNA genes detected in genomic sequence. *Nucleic Acids Res.*, 37: D93-97.

Deragon, J.M. and X. Zhang. 2006. Short interspersed elements (SINEs) in plants: origin, classification, and use as phylogenetic markers. *Syst. Biol.*, 55: 949-956.

Deragon, J.M., B.S. Landry, T. Pelissier, S. Tutois, S. Tourmente and G. Picard. 1994. An analysis of retroposition in plants based on a family of SINEs from *Brassica napus*. *J. Mol. Evol*, 39: 378-386.

Doyle, J.J. and J.L. Doyle. 1990. Isolation of plant DNA from fresh tissue. *Focus*, 12: 13-15.

Fawcett, J.A., T. Kawahara, H. Watanabe and Y. Yasui. 2006. A SINE family widely distributed in the plant kingdom and its evolutionary history. *Plant. Mol. Biol.*, 61: 505-514.

Feschotte, C., N. Jiang and S.R. Wessler. 2002. Plant transposable elements: where genetics meets genomics. *Nat. Rev. Genet.*, 3: 329-341.

Goubely, C., P. Arnaud, C. Tatout, J.S. Heslop-Harrison and J.M. Deragon. 1999. S1 SINE retroposons are methylated at symmetrical and non-symmetrical positions in *Brassica napus*: Identification of a preferred target site for asymmetrical methylation. *Plant. Mol. Biol.*, 39: 243-55.

Jurka, J., V.V. Kapitonov, A. Pavlicek, P. Klonowski, O. Kohany and J. Walichiewicz. 2005. Repbase Update, a database of eukaryotic repetitive elements. *Cytogenet. Genome Res.*, 110: 462-467.

Kapitonov, V.V. and J. Jurka. 2003. A novel class of SINE elements derived from 5S rRNA. *Mol. Biol. Evol.*, 20: 694-702.

Kapitonov, V.V. and J. Jurka. 2008. A universal classification of eukaryotic transposable elements implemented in Repbase. *Nat. Rev. Genet.*, 9: 411-412.

Kapitonov, V.V., S. Tempel and J. Jurka. 2009. Simple and fast classification of non-LTR retrotransposons based on phylogeny of their RT domain protein sequences. *Gene*, 448: 207-213.

Khan, M.F., B.S. Yadav, K. Ahmad and A.K. Jaitly. 2011. Mapping and analysis of the LINE and SINE type of repetitive elements in rice. *Bioinformation*, 7: 276-279.

Kramerov, D.A. and N.S. Vassetzky. 2011. Origin and evolution of SINEs in eukaryotic genomes. *Heredity*, 107: 487-495.

Lenoir, A., L. Lavie and J.L. Prieto. 2001. The evolutionary origin and genomic organization of SINEs in *Arabidopsis thaliana*. *Mol. Biol. Evol.*, 18: 2315-2322.

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

Nouroz, F., S. Noreen and J.S. Heslop-Harrison. 2015a. Identification and characterization of LTR Retrotransposons in *Brassica*. *Turk. J. Biol*., 39: 740-757.

Nouroz, F., S. Noreen and J.S. Heslop-Harrison. 2015b. Molecular characterization and diversity of a novel non-autonomous *mutator-like* transposon family in *Brassica*. *Pak. J. Bot.*, 47(4): 1367-1375.

Nouroz, F., S. Noreen and J.S. Heslop-Harrison. 2015c. Identification, characterization and diversification of nonautonomous hAT transposons and unknown insertions in *Brassica. Genes Genom*., 37: 945-958.

Nouroz, F., S. Noreen and J.S. Heslop-Harrison. 2016. Characterization and diversity of novel *PIF/Harbinger* DNA transposons in *Brassica* genomes. *Pak. J. Bot.*, 48(1): 167-178.

Nouroz, F., S. Noreen and J.S. Heslop-Harrison. 2017b. Identification and evolutionary dynamics of CACTA DNA transposons in *Brassica. Pak. J. Bot.*, 49(2): 789-798.

Nouroz, F., S. Noreen, M.F. Khan, S. Ahmed and J.S. Heslop-Harrison, J.S. 2017a. Identification and characterization of mobile genetic elements LINEs from *Brassica* genome. *Gene*, 627: 94-105.

Okada, N., M. Hamada, I. Ogiwara and K. Ohshima. 1997. SINEs and LINEs share common 3' sequences: a review. *Gene*, 205: 229-243.

Pozueta-Romero, J., G. Houlne and R. Schantz. 1998. Identification of a short interspersed repetitive element in partially spliced transcripts of the bell pepper (*Capsicum annuum*) PAP gene: new evolutionary and regulatory aspects on plant tRNA-related SINEs. *Gene*, 214: 51-58.

SanMiguel, P. and J.L. Bennetzen.1998. Evidence that a recent increase in maize genome size was caused by the massive amplification of intergene retrotransposons. *Ann. Bot.*, 82.

Sonnhammer, E.L. and R. Durbin. 1995. A dot-matrix program with dynamic threshold control suited for genomic DNA and protein sequence analysis. *Gene*, 167: 1-10.

Wicker, T., F. Sabot, A. Hua-Van and J.L. Bennetzen. 2007. A unified classification system for eukaryotic transposable elements. *Nat. Rev. Genet.*, 8(12): 973-982.

Zhang, X. and S.R. Wessler. 2005. BoS: a large and diverse family of short interspersed elements (SINEs) in *Brassica oleracea*. *J. Mol. Evol.*, 60: 677-687.

(Received for publication 18 February 2017)