Molecular Plant • Volume 6 • Number 2 • Pages 470–482 • March 2013 RESEARCH ARTICLE

New Insights into Nested Long Terminal Repeat Retrotransposons in *Brassica* Species

Lijuan Wei[a,2], Meili Xiao[a,2], Zeshan An[a], Bi Ma[a], Annaliese S. Mason[b], Wei Qian[a], Jiana Li[a] and Donghui Fu[c,1]

a Chongqing Engineering Research Center for Rapeseed, College of Agronomy and Biotechnology, Southwest University, Chongqing 400716, China
b School of Agriculture and Food Sciences and Centre for Integrative Legume Research, The University of Queensland, The University of Queensland, Brisbane 4072, Australia
c Key Laboratory of Crop Physiology, Ecology and Genetic Breeding, Ministry of Education, Agronomy College, Jiangxi Agricultural University, Nanchang 330045, China

ABSTRACT Long terminal repeat (LTR) retrotransposons, one of the foremost types of transposons, continually change or modify gene function and reorganize the genome through bursts of dramatic proliferation. Many LTR-TEs preferentially insert within other LTR-TEs, but the cause and evolutionary significance of these nested LTR-TEs are not well understood. In this study, a total of 1.52 Gb of *Brassica* sequence containing 2020 bacterial artificial chromosomes (BACs) was scanned, and six bacterial artificial chromosome (BAC) clones with extremely nested LTR-TEs (LTR-TEs density: 7.24/kb) were selected for further analysis. The majority of the LTR-TEs in four of the six BACs were found to be derived from the rapid proliferation of retrotransposons originating within the BAC regions, with only a few LTR-TEs originating from the proliferation and insertion of retrotransposons from outside the BAC regions approximately 5–23 Mya. LTR-TEs also preferably inserted into TA-rich repeat regions. Gene prediction by Genescan identified 207 genes in the 0.84 Mb of total BAC sequences. Only a few genes (3/207) could be matched to the *Brassica* expressed sequence tag (EST) database, indicating that most genes were inactive after retrotransposon insertion. Five of the six BACs were putatively centromeric. Hence, nested LTR-TEs in centromere regions are rapidly duplicated, repeatedly inserted, and act to suppress activity of genes and to reshuffle the structure of the centromeric sequences. Our results suggest that LTR-TEs burst and proliferate on a local scale to create nested LTR-TE regions, and that these nested LTR-TEs play a role in the formation of centromeres.

Key words: LTR retrotransposons; *Brassica*; centromere; retrotransposon-rich; transposon burst.

INTRODUCTION

Transposons are mobile genetic elements that can be integrated into several different sites over a genome. Transposons were first discovered by Barbara McClintock in the maize genome in the late 1940s (McClintock, 1951), and have since turned out to be a widespread phenomenon. Transposable elements (TEs) are present in high proportions in most genomes, including fruit flies: 10–15% (Vieira et al., 1999), mice: 40% (Smit, 1999), and humans: 44% (Mills et al., 2007). TEs are also important components of the major crop plant genomes, making up 14% of the rice genome (Turcotte et al., 2001), 60% of the maize genome (SanMiguel et al., 1996), and 80% of the wheat genome (Charles et al., 2008). In fact, variation in the size of eukaryotic genomes is primarily dependent on the proportion of repetitive sequences present, including TEs and simple sequences repeats, rather than gene sequences. However, the classes and abundances of retroelements vary with species and variety.

Transposons are divided into two classes based on the presence or absence of RNA as a transposition intermediate. class I elements are called retrotransposons, and are reverse-transcribed using an RNA-intermediate mode of transposition to insert a copy into another location in the genome. Retrotransposons are further subdivided into those with long terminal repeats (LTRs) and those without LTRs. LTR retroelements include the four distinct groups *Ty1-copia*, *BEL*, *DIRS*, and *Ty3-gypsy*, and non-LTR retroelements include long interspersed nuclear elements (*LINE*s) and short interspersed

[1] To whom correspondence should be addressed. E-mail fudhui@163.com, tel. +086-0791–83813142, fax +086-0791–83813185.
[2] These authors contributed equally to this work.

doi:10.1093/mp/sss081, Advance Access publication 28 August 2012
Received 6 March 2012; accepted 10 July 2012

nuclear elements (*SINEs*) (Le et al., 2000). class II TEs move through a 'cut and paste' mode, and include *mariner-Tc1*, *hAT*, *Mu*, *Helitron*, and miniature inverted-repeat transposable elements (*MITEs*) (Wicker et al., 2007). Of the numerous different families of TEs, LTR transposable elements are the most abundant type with the largest genomic effects (Sabot and Schulman, 2006). LTR-TEs generally contain several distinctive structural characteristics, such as long terminal repeat sequences in 5′- and 3′-ends; target site repeats, a polypurine tract, a TG/CA box and the genes encoding reverse transcriptase, integrase, and RNase H enzymes.

TEs greatly contribute to the structure and size of genomes (Kidwell, 2002; Gollotte et al., 2006), driving genetic diversity and evolution through recurrent duplication, transcription, and excision (Flavell et al., 1994; Carareto et al., 1997; Bennetzen, 2000). However, different TEs behave very differently. Low copy number transposons are usually inserted into genes, whereas highly repetitive elements are usually inserted into intergenic regions, such as genic spacer regions, heterochromatin, and even other TEs (SanMiguel et al., 1996). Every transposon has its own target sites, and TEs are not randomly integrated into the host. For example, *Mos1*, a mariner family transposon, is usually inserted into sites such as TATA or TA motifs, or AT-rich regions (Crenes et al., 2011), whereas the *Saccharomyces* Ty5 retrotransposon is integrated into heterochromatin at the telomeres and silent mating loci (Brady et al., 2008). The selection of target sites can also be affected by chromatin structure (Gangadharan et al., 2010) and the structure of the target DNA (Nefedova et al., 2011).

TE bursts, or the rapid expansion and proliferation of TE repeats, can occur under certain conditions. Burst triggers include inactivation of key genes (Naito et al., 2009), environmental stresses (McClintock, 1944), tissue culture (Huang et al., 2009), and interspecific hybridization (Parisod et al., 2009). Furthermore, it seems that certain types of TEs have a greater propensity to burst than other TEs. For example, in the rice genome, the DNA transposon *mPing* increases its copy number by approximately 40 copies per plant per generation (Naito et al., 2006). Interestingly, some retrotransposons may also insert into the region or flanking sequences of other retrotransposons, forming highly TE-rich regions (Kuykendall et al., 2009). These highly clustered and nested genomic regions have been reported in cabbage (Gao et al., 2005), humans (McNaughton et al., 1993), and maize (Tikhonov et al., 1999). However, the evolutionary significance of this phenomenon of clustering TEs and the reasons for its occurrence are unknown.

The *Brassica* genus is a source of important oilseed, condiment, fodder, and vegetable crops. *Brassica napus*, one of the most important oil crops in the *Brassica* genus, is an allotetraploid species derived from interspecific hybridization between diploid vegetable crop species *B. rapa* and *B. oleracea* (U, 1935). TEs constitute 14% of the *B. rapa* genome (Hong et al., 2006) and 20% of the *B. oleracea* genome (Zhang and Wessler, 2004), with retrotransposons comprising 12.3 and 14% of these genomes, respectively (Zhang and Wessler, 2004; Hong et al., 2006). LTR retrotransposons (LTR-TEs) appear to play a significant role in the construction and evolution of *Brassica* genomes. However, the evolutionary history of nested LTR-TEs, the principles of nested LTR-TE bursts and insertions, and the contribution of nested LTR-TEs to genome organization and gene function have been little studied. In the present study, 1.52 Gb of sequences were scanned, and six *Brassica* bacterial artificial chromosome (BAC) fragments containing nested LTR-TEs were selected for detailed analysis. This study reports the characteristics, origins, evolutionary significance, and putative contributions to the genome of nested LTR retrotransposons in *Brassica*.

RESULTS

Characteristics of the Six LTR-TE-Rich BAC Clones and Corresponding LTR Retrotransposons

A total of 1.52 Gb of *Brassica* sequences were downloaded from the NCBI (National Center for Biotechnology Information) website (www.ncbi.nlm.nih.gov/), including 1 894 589 BAC end sequences (BES) and 2020 BACs. Subsequently, 9956 LTR-TEs were detected with the LTR-Finder tool. Finally, six highly LTR-TE-rich BAC clones containing a total of 314 LTR retrotransposons were identified: AC236792 (22), AC183494 (19), CU984542 (136), AC166740 (92), AC189563 (26), and AC189265 (19). These BACs met the criterion that the ratio between the sequence length of the BAC clone and the total number of LTRs in the BAC clone was less than 17.9 ($P < 0.001$, χ^2 test) (Supplemental Figures 1–6). Most LTR-TEs in the six BACs were nested (see Supplemental Figures 1–6). The density of LTR retrotransposons in the BACs ranged from 1.09/kb to 17.9/kb, with an average of 7.24/kb. LTR retrotransposons comprised the major proportion of LTR-TEs. LTR-TEs constituted more than 50% of CU984542 and AC189563, and 20–30% of the other four BAC clones. Using RepeatMasker, we identified other transposons in the six BAC clones: two *LINE*s and one *Helitron* in AC183494 and two *LINE*s in AC236792.

The basic features of the six BACs are listed in Table 1. BAC sequence length ranged from 35 526 to 394 448 bp, and the GC contents were 49, 38.8, 49.4, 36.9, 36.9, and 36.8% for CU984542, AC236792, AC189563, AC189265, AC183494, and AC166740, respectively.

The genes in the six BACs were classified into three classes: class I (genes related to transposition), class II (genes unrelated to transposition), and other genes of unknown function. In total, 207 genes were identified in the six BACs (2–89 per BAC): 31 class I genes (15.0%), 61 class II genes (29.5%), and 115 genes of unknown function (55.5%). The average gene size (including introns) was 4079 bp. Many genes contained transposon insertion sites. Most TEs inserted into introns (29%) and exons (27%), with 8–10% inserting into 5′ UTRs, terminators, 3′ UTRs, and gene spaces, respectively. The activity of the 207 genes was predicted using BlastN analysis

Table 1. The Basic Features of Six LTR-Retrotransposon-Rich *Brassica* BAC Clones.

Genbank accession number	CU984542	AC189563	AC166740	AC236792	AC189265	AC183494	Average
Species origin	*B. rapa*	*B. rapa*	*B. rapa*	*B. napus*	*B. rapa*	*B. oleracea*	–
Sequence length	154 260	35 526	100 288	394 448	130 769	285 752	183 507
GC content (%)	49	39	49	37	37	37	41
Protein-coding DNA regions (%)	74	56	77	52	49	47	59
Non-coding regions (%)	20	1	19	40	36	41	26
Total number of genes	20	2	12	89	27	57	34
No. of transposition-related genes (class I)	7	0	3	11	2	8	5
No. of non-transposition-related genes (class II)	0	0	0	25	15	21	10
No. of genes of unknown function	13	2	9	53	10	28	19
Average gene size (bp)	7206	10 493	8031	4079	4143	4416	6395
Average number of exons per gene	15	2	15	4	4	4	7
Average exon size (bp)	267	6564	302	405	488	324	1392
Average number of introns per gene	13.5	1.5	13.9	2.7	3.2	3.3	6.4
Average intron size (bp)	189	113	179	289	106	297	196
Average spacer size (bp)	529	304	293	355	578	510	428
Total number of LTRs	136	26	92	22	19	19	44
LTR transposons (% of sequence)	54	56	18	23	16	33	36

against the *Brassica* expressed sequence tag (EST) database. Only three genes (all from BAC AC166740) matched the EST library, suggesting the retrotransposon insertions had inactivated the majority of genes present in the BACs.

Alignment of the Six BAC Clones with the Genomes of *Arabidopsis thaliana* and *B. rapa*

The six LTR-TE-rich BAC clones were aligned with the ancestral karyotype of *A. thaliana* genomes by local BlastN with E value set to '$1.0E^{-30}$'. In the results, three of the BACs matched well with the *A. thaliana* genome, but another three BAC clones (AC166740, CU984542, and AC189563) did not match. Supplemental Table 1 shows the alignment results between the BACs and the blocks of *A. thaliana*. The flanking pericentromeric regions in *A. thaliana* were B and C blocks in chromosome 1, G and H blocks in chromosome 2, L and M blocks in chromosome 3, O and P blocks in chromosome 4, and Q and S blocks in chromosome 5. AC189265 showed homology with the H and I blocks in chromosome 2 and with the O and U blocks in chromosome 4. AC183494 had some homology with the G block in chromosome 2, the F, N, and L blocks in chromosome 3, the O and U blocks in chromosome 4, and the Q, S, V, and X blocks in chromosome 5. BAC clone AC236792 showed homology with all chromosomes: with the A and B blocks in chromosome 1, the G block in chromosome 2, the L and M blocks in chromosome 3, the O and U blocks in chromosome 4, and the R, Q, S, and V blocks in chromosome 5. So these three BACs were located in the pericentromeric regions.

To explore the origin of the six BAC clones, a BlastN analysis ($E \leq e^{-30}$) was performed against the entire *B. rapa* genome (Wang et al., 2011) (Supplemental Table 2). The four *B. rapa* BAC clones (CU984542, AC166740, AC189563, and AC189265) were from chromosomes A8 (position 1082959–16539190), A6 (position 441777–25956180), A2 (position 14816623–23048324), and A7 (position 231473–22563732), respectively. The *B. napus* C genome BAC clone (AC236792) was homologous to chromosome A2 of *B. rapa*, from position 69828 to 27826786, and hence is predicted to be on chromosome C2 from known homoeology (Parkin et al., 2005). The *B. oleracea* BAC clone AC183494 was homoeologous to chromosome A6 of *B. rapa*, and from known homoeology (Parkin et al., 2005) may be assumed to be on chromosome C5 or C7.

The Relationship between the Six BAC Clones and Centromeres

From previous research, *Brassica* is known to contain five highly conserved centromere-specific motif sequences (about 170 bp) and other long centromere-specific sequences (Hong et al., 2006; Lim et al., 2007; Pouilly et al., 2008). Three of the BAC clones (AC166740, AC183494, and AC236792) had already been proved to be centromere sequences. The other three BACs (CU984542, AC189563, and AC189265) were aligned with centromere-specific sequences (176-bp repeat unit) (Lim et al., 2005) by Blast2 sequences and the local blast of NCBI ($E \leq e^{-5}$). These results showed that the BAC clone CU984542 shared 63% homology with clone AC166740, clone AC189563 was composed of 176-bp centromere-specific repeat units, and clone AC189265 had less homology with the centromere sequences and hence may not be centromeric. Therefore, five of the six BACs (not AC189265) are either confirmed to be from centromere regions or contain centromeric repeat sequences. In contrast to the other five centromere sequence BACs, AC189265 had the smaller average gene size and larger gene density.

Characteristics of the LTR-TEs in the Six BAC Clones

Basic information for the LTR-TEs is described in Table 2. The 5′ and 3′-regions of the LTR-TEs had an average sequence similarity of 89% (82–93%). The size of both 5′-LTR and 3′-LTR regions averaged 480 bp, but with a 22-fold difference in size between the largest and smallest regions. In four of the six BAC clones, 84.4% of the LTRs on average were located in the antisense strands. In the remaining two BAC clones (AC189563 and AC189265), LTRs tended to be located in the sense strands (77.8% on average).

The 5′ and 3′ flanking sequences (>50 bp) of all LTR retrotransposons were analyzed to determine the basic characteristics of the TE insertion environments. There was no obvious bias in AT content of the LTR flanking sequences. However, the first base of the 5′-end of the LTR retrotransposons was generally a T (75.0%), and the corresponding base at the 3′-end was almost always an A (97.4%) (Figure 1). To determine when LTR-TE accumulation bursts occurred during evolutionary history, we calculated the divergence time of the LTR-TEs according to the formula (T = K/2r) (Figure 2). The LTR-TEs entered the specific BAC regions en masse approximately 5–23 Mya, during the Neogene period of geological time between the Pliocene epoch and the Miocene epoch.

Origin of the LTR-TEs in the Six BAC Clones

The high homology between the 5′-LTR and 3′-LTR regions of the LTR-TEs prompted the selection of 5′-LTRs to analyze the origin of LTR-TE insertions. The six BACs contained a total of 247 LTRs: 22 in AC236792, 18 in AC183494, 122 in CU984542, 59 in AC166740, 9 in AC189563, and 19 in AC189265. The alignments of the LTRs within the six BACs to each other and to external genomic sequences outside the six BACs were used to determine the origins of LTR-TE accumulations. The LTR-TEs in the six BACs could be grouped into three categories: (1) insertions of alien LTR-TEs from outside the BAC, including from other non-tested TE-rich BACs, (2) seed LTR-TEs for retrotransposons inserted from the BAC to external genomic regions, and (3) LTR-TEs duplicated inside the BAC region.

The 5′-LTR regions were named by BAC code, start position, and end position. For clone AC236792 (*B. napus*), one LTR (AC236792/289691–289866) had weak hits only and the corresponding LTR-TE was assumed to be an 'orphan' TE. Other LTR-TE regions could be classified into two categories according to homology with available *Brassica* LTRs: (1) LTR-TE regions duplicated and inserted into AC236792 (3/21 LTRs) and (2) LTR-TE regions sharing high homology with other TEs outside the six BAC clones ($E \leq e^{-5}$) (18/21 LTRs). To confirm the origin of the 18 LTR-TEs that showed homology with alien TEs, we analyzed the TE divergence times using phylogenetic trees. Twelve LTR-TEs were considered to be derived from the proliferation and insertion of alien TEs from outside AC236792, and six LTR-TEs were seed LTR-TEs, which inserted into external genomic regions from within AC236792 (Figure 3).

Similarly to AC236792, one LTR of AC183494 (AC183494/79929–80234) was an 'orphan'. The remaining 17 LTRs of AC183494 could also be divided into two categories: (1) eight LTRs were homologous to other LTRs outside AC183494, with phylogenetic analysis indicating that four of these eight internally duplicated LTRs were from alien retrotransposons outside AC183494 and that the other four LTRs were ancestral TEs of other TEs and (2) nine LTRs duplicated themselves within AC183494.

In AC166740, 53 LTRs (90%) proliferated and became the major proportion of LTRs in this BAC (Figure 3). Another two LTRs were derived from LTRs outside AC166740. The remaining four LTRs aligned well with LTRs in CU984542, and two of these four LTRs originated from within CU984542. LTRs in CU984542 had similar characteristics to LTRs in AC166740. Up to 93% (113/122) of the TEs were duplicated internally; 5% were derived from alien TEs outside CU984542. The remaining LTRs (2%) were seed TEs, inserting into external genomic regions.

All 19 LTRs originating from AC189265, as well as the retrotransposons inserted into this BAC, rapidly proliferated themselves. Three of the nine LTRs of AC189563 were derived from alien TEs outside AC189563, another three LTRs came from within AC189563, and the remaining three LTRs were seed TEs.

In conclusion, most LTR-TEs in four of the BACs duplicated themselves to insert into the host BAC: 100% of LTR-TEs in AC189265, 53% of LTR-TEs in AC183494, 90% of LTR-TEs in AC166740, and 93% of LTR-TEs in CU984542 (Figure 3).

Table 2. LTR Information for LTR-TEs in Six *Brassica* BAC Sequences.

Genbank accession number	CU984542	AC189563	AC166740	AC236792	AC189265	AC183494	Average
Proportion of sequence similarity between 5′ and 3′-LTR sequences	0.9	0.9	0.9	0.9	0.8	0.9	0.9
Size range of 5′-LTR regions (bp)	63–1562	169–1403	55–2903	52–1366	68–2702	84–901	–
Size range of 3′-LTR regions (bp)	51–1562	169–1424	55–2903	52–1367	59–2696	90–901	–
Average size of 5′-LTRs (bp)	419	1098	294	324	499	249	481
Average size of 3′-LTRs (bp)	415	1103	289	324	507	252	520
Total number of sense strands	9	21	7	6	14	4	10
Total number of antisense strands	127	5	85	16	5	15	42

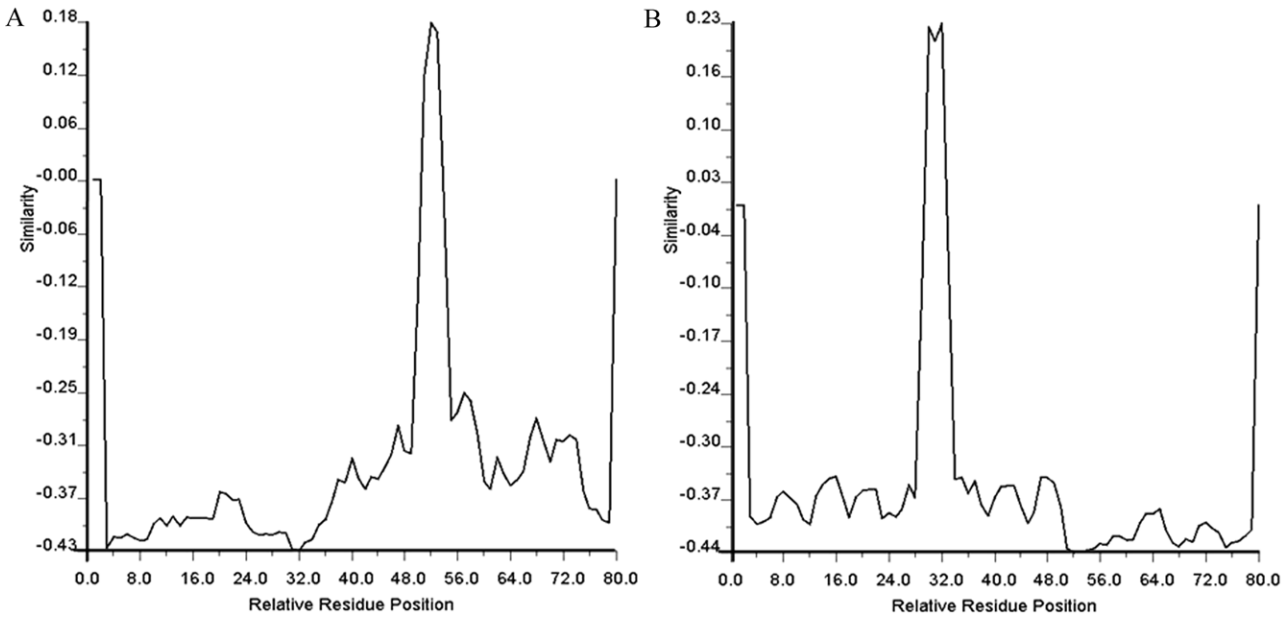

Figure 1. The Conserved Regions of 5′-LTR and 3′-LTR Insertion Sites of LTR Retrotransposons in *Brassica* BACs. **(A)** represents the conserved region of 5′LTR insertion sites, which consisted of 50 bp of 5′-LTR flanking sequences and 30 bp of transposon sequences. **(B)** denotes the conserved region of both 30 bp of transposon sequences and 50 bp of 3′-LTR flanking sequences.

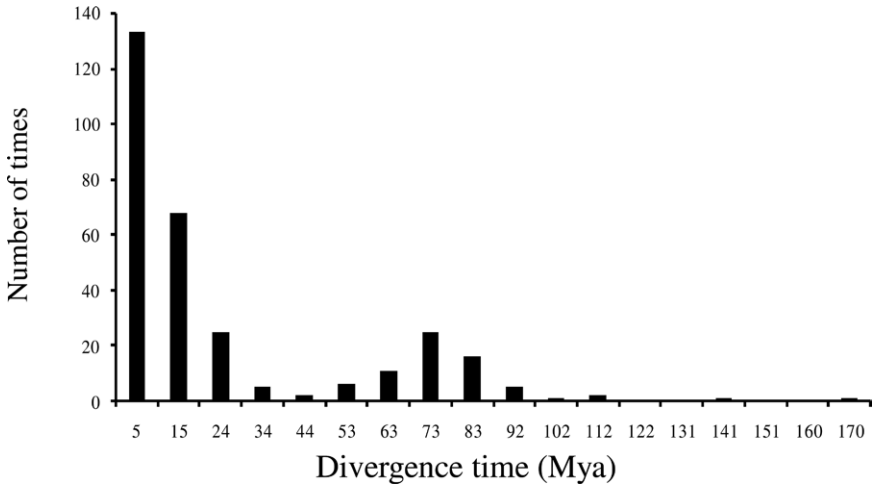

Figure 2. The Distribution of Divergence Time of TEs in Six TE-Rich *Brassica* BACs. The *x*-axis represents time (My) according to the formula (T = K/2r) and the *y*-axis denotes the repeat numbers of the proliferated TEs.

However, in the remaining two BACs, most LTR-TEs originated from alien TEs: 57% of LTR-TEs in AC236792 and 67% of LTR-TEs in AC189563.

Functional Annotations of Coding-Protein Gene Activity

The functional annotations of class I and class II genes are showed in Figure 4. In the annotation of cellular composition, class I genes were mainly operative in the organelles (40%), including the mitochondria and plastids, followed by the nucleus (20%), chromosomes (20%), cytoplasm (10%), and other cell parts (10%). In the annotation of the molecular function of class I genes, binding function occupied the largest proportion (77%), and included DNA binding, RNA binding, nucleic acid binding, and other types of unclassified binding. The remaining gene functions were associated with transferase activity (11%), peptidase activity (5%), catalytic activity (2%), enzyme regulator activity (3%), and hydrolase activity (2%). The biological processes of class I genes were mainly involved with metabolic processes (51%), followed by biosynthetic processes (18%).

The number of types of cellular composition of class II genes was greater than that of class I genes. The majority of genes

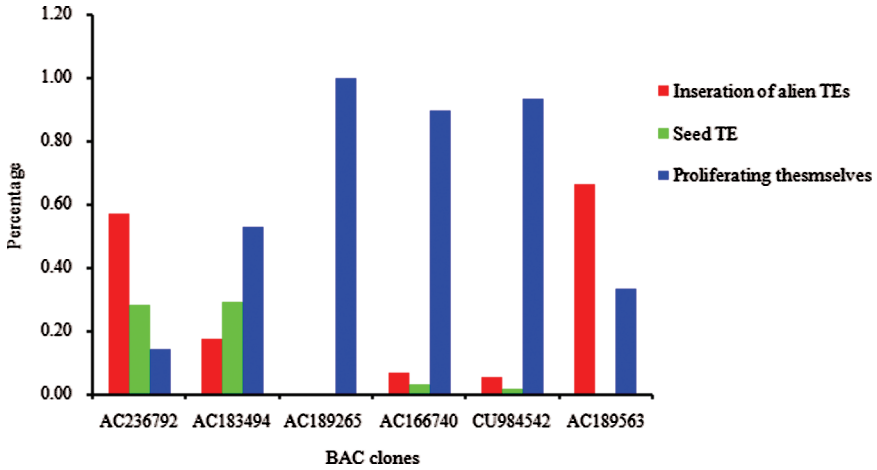

Figure 3. The Origin of Proliferation and Enrichment of TEs in Six TE-rich *Brassica* BACs. All transposons of these six BACs could be grouped into three categories: (1) insertion of alien TEs outside the tested TE-rich BAC, including in other non-tested TE-rich BACs; (2) seed TEs inserting into other places outside the tested TE-rich BAC; and (3) proliferation by self-duplication. The *x*-axis represents six BACs and the *y*-axis denotes the percentage of proliferation types of these transposons in the corresponding BAC.

were expressed in the cell (20%), organelles (30%; 15% in the mitochondria), and plasma membrane (9%). Binding functions occupied the greatest proportion (45%) of the molecular functions of class II genes, similar to class I genes. A larger proportion of class II genes than class I genes were involved with metabolic processes (21%), including carbohydrate metabolic processes, cellular amino acids, and derivative metabolic processes. The secondary types of biological process were responses to stimulus (9%) associated with important environment factors, such as abiotic, biotic, and external stimulus or stress, in which 1-aminocyclopropane-1-carboxylate (ACC) oxidase played a foremost role.

In combining the annotation of class I with that of class II, we found that binding was the key role in molecular function, that metabolic processes comprised the main body of biological processes, and that the genes of the two classes functioned in both organelles and cells. In comparison, genes in the *B. rapa* genome are mainly involved in transcription factor activity (16%), catalytic activity (8.4%), and binding function (6.4%), largely expressed in the chloroplast (23%) and endomembrane system (19%), and are predominantly involved with the response to the environmental adaptability (19.3%) (Wang et al., 2011). The gene functions of genes in nested LTR-TEs in the BACs and genes in *B. rapa* sequences were comparatively consistent, with the exception that the main molecular function of genes in the six BACs was transferase activity, while transcription factor activity genes occupied the higher proportion in the *B. rapa* genome.

SSR (Simple Sequence Repeat) Distribution in the Six Retrotransposon-Rich BACs and LTR Transposons

The six BACs contained 305 SSRs (1–10 bp per repeat unit) and had an SSR density of one SSR per 3.64 kb. SSRs with mononucleotide repeat motifs comprised 74.4% of SSRs (227/305), SSRs with dinucleotide repeats comprised 12.8% (39/305), SSRs with trinucleotide repeats comprised 1.3% (4/305), and SSRs with higher-order motifs accounted for 11.5% (35/305). Table 3 displays the distribution of SSR repeats in the six LTR-rich retrotransposon BACs. There were 212 A/T SSRs, followed by 14 C/G SSR loci. The AT and AG/CT SSRs accounted for 71.8% of the dinucleotide SSRs.

A total of 49 SSRs were detected in the LTR retrotransposons, and the SSR density within the LTRs was one SSR per 6.21 kb. The percentages of mononucleotide, dinucleotide, heptanucleotide, and decanucleotide SSRs among the 49 SSRs were 59.2 (29), 12.2 (6), 26.5 (13), and 2% (1), respectively (Table 4). Similarly to the distribution rules of mononucleotide SSRs in the BACs, almost half of the SSRs in LTRs were A/T-motif SSRs, and the second largest proportion was occupied by C/G-motif SSRs. For the dinucleotide SSRs, the AG-motif SSRs were the most frequent. In summary, the A/T motif comprised the main body of SSRs.

We also analyzed SSR density in the *B. rapa* genome for comparison, and found that the SSR density of the *B. rapa* genome was one SSR per 2.10 kb, higher than in the BACs (one SSR/3.64 kb) and within the LTRs in these BACs (one SSR/6.21 kb). According to Pearson's χ^2 test, the difference in SSR density between the nested LTR-TEs of these six BACs and the *B. rapa* genome was significant ($P<0.01$, $\chi^2=95.09$). This showed that SSR density decreased as LTR retrotransposons proliferated and gathered together.

DISCUSSION

TEs, also called jumping genes, promote gene mutations and affect gene expression. Therefore, TEs can be considered as an engine for gene variation and genome variation. The proliferation and insertion of TEs play a critical role in gene

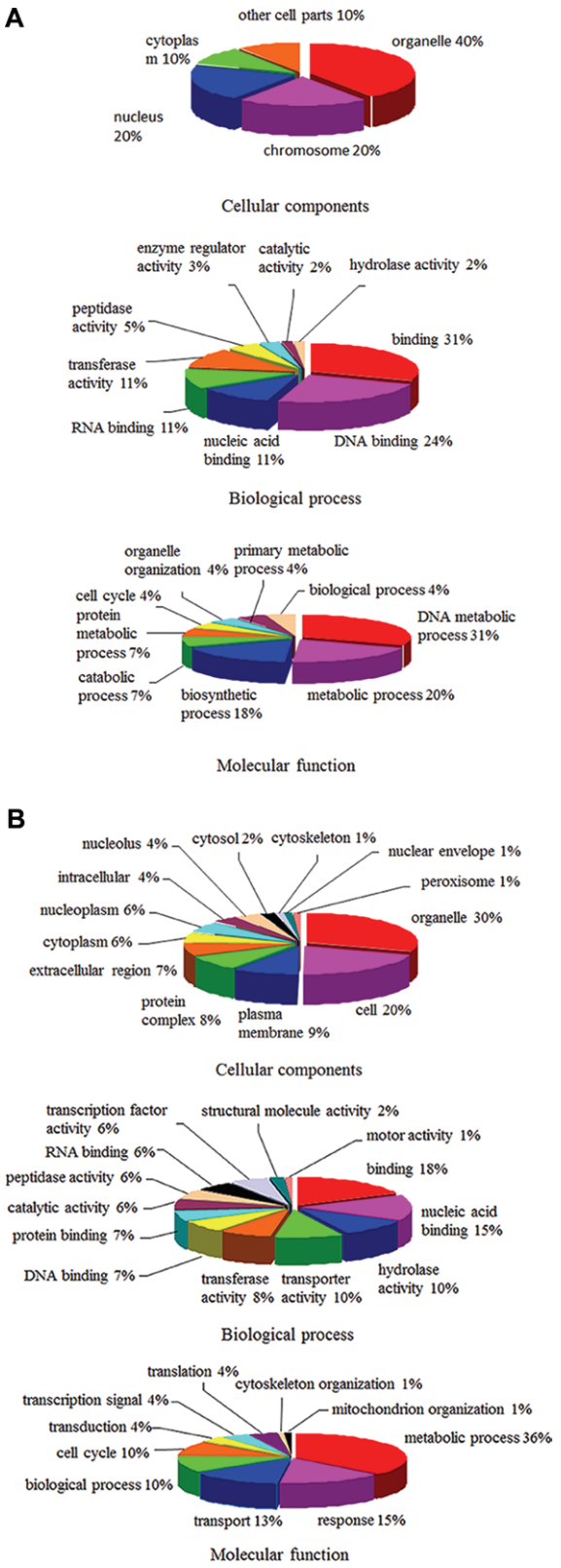

Figure 4. Gene Ontology (GO) Annotation of Genes in Six LTR-TE-Rich *Brassica* BACs.

(A) shows the functional annotation of genes related to transposons and **(B)** presents the corresponding annotations of class II.

activity. LTR-TEs are ubiquitous in the eukaryotic genome. Many LTR-TEs recurrently insert into the same region to become either 'nested' or 'clustered' LTRs. The present study systematically characterizes LTR-TE dense sequence structures in *Brassica* species, and speculates that highly TE-rich regions represent regions transitioning from immature centromeres into mature centromeres.

The TE High Accumulation Phenomenon

Most LTR-TEs in four of the six BAC clones were derived from copies of retrotransposons within these BAC regions. Increases in diversity and proliferation of TEs can be realized through three means: (1) horizontal transfer from one species into another species, (2) mutation, which makes TEs diverse and may enhance their ability to transpose, and (3) stochastic TE generation (Lohe et al., 1995). Massive accumulation of LTR-TEs directly leads to increased complexity and density of TEs, and may hence provide a reservoir for TE activation in response to genome variation or environmental stress, or to form the material base of particular structure such as a centromere.

The phenomenon of LTR-TEs gathering in the centromere region has also been found in tomato (Park et al., 2011), *A. thaliana* (Tsukahara et al., 2009), and *Drosophila* (Sun et al., 2003). In the maize genome, LTR-TEs were often inserted into similar regions at high copy numbers to form nested TEs. In addition, in LTR-TE accumulation, MITEs may also give rise to enrichment. For example, in *Brassica*, the MITE transposons accumulate preferentially in the promoter regions and gene-regulating regions but not in protein-coding regions, which indicates that the enrichment of MITEs may play an important role in the regulation of gene expression (Sarilar et al., 2011).

The homology between the TE and the targeted site lays a solid foundation for TE insertion. Although the first bases at the 5′-end and the 3′-end of LTR retrotransposons were found to be T bases (74.99%) and A bases (97.37%), respectively, other specific structures were not observed. However, nested TEs in one BAC often shared the same 5′- and 3′-end, which might facilitate TE insertion into regions containing identical TEs. The LTR-TEs entered the specific BAC regions en masse approximately 5–23 Mya, during the Neogene period of geological time between the Pliocene epoch and the Miocene epoch. This suggests that massive TE proliferation occurred within a relatively narrow time period. In this epoch, the climate became cold, and all living things greatly suffered from climate stress. The burst of LTR-TEs may have been generated during this period as a result of the rugged external environment.

TE Insertion and Gene Function

The predicted gene size in this study ranged from 4.1 to 10.5 kb, with an average of 6.4 kb. This was considerably larger than the 1.7-kb average reported in chromosome A3 of

Table 3. Types of SSR Repeats in Six LTR-Retrotransposon-Rich *Brassica* BAC Sequences.

Motif	'a'/'b'	'a' repeats observed	'b' repeats observed	Total	% of repeat sequences
Monomer	A/T	94	118	212	69.5
	C/G	7	7	14	4.6
Dimer	AT/TA	15	6	21	6.9
	AG/CT	8	5	13	4.3
	TC/GA	4	0	4	1.3
	GT/AC	1	0	1	0.3
Trimer	ATT/AAT	2	0	2	0.7
	CAT/ATG	1	0	1	0.3
	ATC/GAT	1	0	1	0.3
Tetramer	ATAG/CTAT	1	0	1	0.3
Pentamer	CAAAA/TTTTG	1	0	1	0.3
Hexamer	TTTTGT/ACAAAA	2	0	2	0.7
	AAAATT/AATTTT	1	0	1	0.3
Heptamer	CCCGAAA/TTTCGGG	4	0	4	1.3
	AAACCCG/CGGGTTT	4	0	4	1.3
	GAAACCC/GGGTTTC	4	0	4	1.3
	CTAAACC/GGTTTAG	3	1	4	1.3
	GTTTAGG/CCTAAAC	3	0	3	1.0
	TTTAGGG/CCCTAAA	1	0	1	0.3
	CCAAACC/GGTTTGG	1	0	1	0.3
	GTTTGGG/CCCAAAC	1	0	1	0.3
	AGGGTTT/AAACCCT	1	0	1	0.3
	TGGGGTT/AACCCCA	1	0	1	0.3
	TTTGGGG/CCCCAAA	1	0	1	0.3
	AGCGAGG/CCTCGCT	1	0	1	0.3
	AGAGAGA/TCTCTCT	1	0	1	0.3
	AAACCCC/GGGGTTT	1	0	1	0.3
Octamer	GCTCCACC/GGTGGAGC	1	0	1	0.3
Nonamer	TTTAAATTA/TAATTTAAA	1	0	1	0.3
Decamer	TGTCCGTGTG/CACACGGACA	1	0	1	0.3

Table 4. Distribution of SSR Repeats within LTR Retrotransposons.

Motif	'a'/'b'	'a' repeats observed	'b' repeats observed	Total	(%) of SSR '-mer' type	% of repeat sequence
Monomer	A/T	10	14	24	85.7	49.0
	C/G	2	2	4	14.3	8.2
Dimer	AG/CT	4	0	4	66.7	8.2
	GT/AC	1	0	1	16.7	2.0
	AT	1	–	1	16.7	2.0
Heptamer	CCCGAAA/TTTCGGG	4	0	4	30.8	8.2
	GAAACCC/GGGTTTC	4	0	4	30.8	8.2
	AAACCCG/CGGGTTT	3	0	3	23.1	6.1
	AAACCCC/GGGGTTT	1	0	1	7.7	2.0
	CCAAACC/GGTTTGG	1	0	1	7.7	2.0
Decamer	TGTCCGTGTG/CACACGGACA	2	0	2	100	4.1

B. rapa (Mun et al., 2010) and the 2.0-kb average reported in the *B. rapa* genome (Smoot et al., 2011). Gene length may be associated with the evolution of complexity (Xu and Wang, 2007), and the larger gene length may be beneficial for the further evolution of nested LTR-TEs. It is possible more rearrangements occur in the nested LTR-TE genes to extend the

size of genes. However, this finding may also be attributable to the insertion of the LTR-TEs inside gene regions. Only 3/207 genes in the six BACs (all from BAC AC166740) matched the *Brassica* ESTs, which suggests that most genes are inactive and are not expressed after retrotransposons insertion into these regions, as the LTR insertions inactivate the genes. In the human genome, 5.8% of genes are associated with LTR retrotransposons distributed in gene exons (Piriyapongsa et al., 2007). Garber et al. (1999) also showed that the Tpv2 family and LTR retrotransposons are preferentially inserted into the protein-coding regions in *Phaseolus vulgaris*. However, in the maize genome, LTR transposable elements, the main TE type, are often nested inside one another in the intergenic regions. A biased distribution of TEs in the intergenic regions is also found in pepper (Park et al., 2011). Novel LTR retrotransposons of *Aedes aegypti* are also found in introns and intergenic genomic sequences (Minervini et al., 2009). In contrast, the Ty5 family of yeast preferred to localize near telomeres and in silenced mating loci (Zou et al., 1996). In a decreased DNA methylation (*DDM1*) mutant population of *A. thaliana*, LTR retrotransposons were observed to burst and target the centromere region, which is less harmful than insertion into genes (Tsukahara et al., 2009). The universality of TE distribution results in diverse TE effects on the genome. Another reason that the genes in the LTR-TE-rich BACs were not expressed might be the epigenetic regulation of centromere sequences, including methylation, acetylation, and small RNA regulation (Grewal and Rice, 2004; Habu, 2010). In conclusion, the accumulation of LTR-TEs in *Brassica* interferes with genic activity inside or near the TEs.

The Relationship between TE Accumulation and the Centromeres

Centromeres are the structural and functional elements of chromosome segregation at mitosis and meiosis in all eukaryotes. However, centromere sequences lack long conserved regions and vary with species. Generally, highly repetitive satellite DNA and some centromere retrotransposons characterize the centromere regions (Cheng et al., 2002). The repetitive satellite DNAs are arranged head to tail, and are sometimes interspersed with retrotransposons. For instance, the centromeres of *A. thaliana* are composed of the 180-bp repeat family pAL1 interspersed with Athila retrotransposons (Ty3-gypsy family) (Brandes et al., 1997).

Nevertheless, some researchers have found that the satellite DNA of the centromeres shows homology between closely related species (Gindullis et al., 2001; Hass et al., 2003). In *A. thaliana*, repetitive satellite DNA shows more than 58% homology between varieties (Kamm et al., 1995). Hass et al. (2003) found that centromeres were conserved between some related *Oryza* species but not with wild rice. In yeast, special AT-rich DNA sequences are the important organizational elements for centromere function, and determine centromere identity (Baker and Rogers, 2005). In our study, the BAC clones were found to have high homology with five highly conserved *Brasssica* centromere fragments, and two BACs had been confirmed in previous reports to be closely associated with centromeres (Lim et al., 2007; Pouilly et al., 2008).

The processes by which centromeres age and evolve are still unclear. The BAC sequence analysis suggests that LTR-TE accumulation and subsequent gene dysfunction drive centromere formation. Autosome regions are able to evolve into mature centromeres: for instance, human neocentromeres have been observed to form in chromosome arms (Warburton, 2004). Ma and Jackson (2006) revealed that segmental duplication is the underlying mechanism for centromere formation in rice. In rice, Cen8 processing of low gene content in satellite DNA has proven to be an intermediate stage of the centromere body from genic region to mature centromere (Nagaki et al., 2004). We might assume the entire process of TE accumulation has led to the emergence of neocentromeres in the present study. First, some genomic regions may have high homology with LTR-TEs, making these regions preferred sites for LTR-TE insertion. Second, after a 'great shock' to the genome, such as wide hybridization or rugged environmental conditions, LTRs may activate, rapidly proliferate, and insert into the preferred genomic sites. Third, as a result of LTR-TE insertion, genes in this region will lose their function, resulting in phenotypic changes. Fourth, as a result of evolutionary selection for these phenotypic changes, either speciation occurs or lines carrying the 'dead region' become the norm. Lastly, the few remaining genes are gradually abandoned, forming a heterochromatic region that becomes a mature centromere when required (Figure 5). However, as the number of genes in these BACs was still quite large, these BACs might be an intermediate stage in the centromere formation process, namely pre-centromeres.

TE Accumulation and SSRs

SSRs are spread throughout the genomes of many animals and plants, and many SSRs are distributed specifically within TEs. In the *Phyllostachys pubescens* genome, SSRs were more abundant within TEs than in genome survey sequences (GSS) and full-length cDNA sequences (FL-cDNAs) (Zhou et al., 2011).

However, in the present study, SSRs were distributed in all six BACs (one SSR per 3.64 kb) and within LTR retrotransposons in the BACs (one SSR per 6.21 kb). The SSR density of the *B. rapa* genome was higher than that the SSR density in the BACs we analyzed and within the LTRs in these BACs. Similar findings have also been documented in other species. For example, Grover et al. (2007) thought that close associations between TEs and microsatellites are not common in the rice genome. In addition, Schlötterer (2000) stated that a high density of transposons is not consistent with a high density of SSRs in centromeric regions of *A. thaliana*. This phenomenon may result from TEs preferentially inserting into SSR or proto-SSR regions and subsequently 'breaking' SSRs and

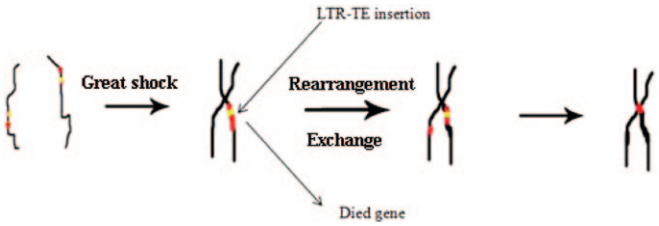

Figure 5. The Maturation Process of the Centromere. A genomic fragment located in an autosomal region had high homology with LTR-TEs. When the genome suffered a 'great shock' (such as wide hybridization or a rugged environment), LTRs were activated and rapidly proliferated and inserted into this specific fragment, resulting in loss of gene function in this region. As a result of evolutionary selection for these phenotypic changes, either speciation occurs or lines carrying the 'dead region' become the norm. Finally, the few remaining genes are gradually abandoned, forming a heterochromatic region which becomes a mature centromere when required.

decreasing SSR density (Ramsay et al., 1999). It is also likely that the structure *per se* and proliferation mechanisms of retrotransposons constrain increases in SSR density (Hong et al., 2007). Sequence conservation of LTR retrotransposons may also not permit SSR variation in LTR-TE regions, because LTR-TEs require higher conservation. Du et al. (2010) found that centromeric retrotransposons were also highly conserved in monocot and eudicot species.

Massive accumulations of LTR-TEs severely truncate genes and affect gene expression, resulting in a process of transition from new centromeres into mature centromeres. These results deepen our understanding of the formation of neocentromeres. However, some questions, such as why organisms select these regions as original sequences for future centromeres, why these LTR-TEs accumulate, how TEs are repeatedly inserted into the same regions, and what is the mechanism of homogenization for these diverse sequences, require further investigation.

To sum up, in this study, we found that extremely nested LTR retrotransposons with few SSRs were enriched in six *Brassica* BACs. The majority of the nested LTR-TEs were derived from the rapid proliferation of retrotransposons within the BAC regions approximately 5–23 Mya, and these LTR-TEs preferentially inserted into TA-rich regions. Most BACs proved to be in centromere regions or contain centromeric repeat sequences. Retrotransposon insertions had also inactivated the majority of genes present in the BACs. Nested LTR-TEs were rapidly duplicated, repeatedly inserted, and accumulated to lead to loss of function of most genes and reshuffled centromere sequence structure. These results suggest that nesting of LTR-TEs plays an evolutionarily significant role in centromere formation.

METHODS

Prediction of Transposon-Rich BAC Clones

Public *Brassica* sequences were downloaded from the NCBI website (www.ncbi.nlm.nih.gov/). LTR retrotransposons were predicted using the LTR-Finder tool (http://tlife.fudan.edu.cn/ltr_finder/) with eukaryotic transposons as reference sequences. The density of LTR transposons (the sequence length of BAC clone/total number of LTR-TE in this BAC) and the transposon-rich BAC clones (TE density of genome versus TE density of this BAC) were tested for significance using Pearson's χ^2 test. Other types of transposons in the LTR-TE-rich BACs were predicted using RepeatMasker with default values (www.repeatmasker.org/) (Smit et al., 2011).

Gene Prediction and Analysis of BAC Origin

The genes of the LTR-TE-rich BAC clones were predicted using GenScan (http://genes.mit.edu/GENSCAN.html; Burge and Karlin, 1997) with *A. thaliana* as the reference species and other default parameters. The GC content of the LTR transposon-rich clones was calculated using BioEdit (Hall, 1999). To understand the origin of these six clones, a local BlastN (www.ncbi.nlm.nih.gov/BLAST/) analysis was performed against the *A. thaliana* (www.arabidopsis.org/) and *B. rapa* genome databases (Smoot et al., 2011) with E-values lower than '$1.0E^{-30}$'.

Classification and Origin Analysis of LTR

LTR sequences, as the specific defining feature of LTR-TEs, were divided into two groups in the present study: (1) LTRs of the transposon-rich BAC clones and (2) LTRs of other *Brassica* BAC clones (non-TE-rich BACs), which are also called 'other LTRs' for short. Given that 5′-LTRs have high similarity with 3′-LTRs for a given LTR-TE, the 5′-LTR of each TE in each BAC clone was selected to conduct the following analysis. First, the 5′-LTR sequences of all *Brassica* LTR-TEs were extracted in batches using the Perl programming language. The overall BlastN analysis was conducted using the local BlastN at E-values lower than '$1.0E^{-10}$'. Second, the LTRs were classified into different classes using our own program and displayed using Cytoscape 2.0, with the layout files set in 'circular' mode (Smoot et al., 2011). Furthermore, the corresponding phylogenetic tree was drawn from each subclass of LTRs of each BAC using MEGA 4.0 (Tamura et al., 2007) with a neighbor-joining mode and 1000 bootstrap tests. The origin and lineage relationship of the BAC LTRs were determined based on divergence time.

Analysis of Flanking Sequences of LTRs

The 50-bp flanking sequences of the 5′-LTR and 3′-LTR of each LTR-TE were extracted and base composition assessed. To plot the conserved blocks of the sequence alignment from the extracted flanking sequences, the Emboss subprogram Plotcon was used with default values (Rice et al., 2000).

Calculation of Divergence Time of LTRs

One newborn LTR-TE ought to have identical sequence between 5′-LTRs and 3′-LTRs. The base difference between the sequences of the 5′-LTR and the 3′-LTR reflects the evolutionary history and divergence time of the LTR-TE (Kijima and Innan, 2010). Therefore, the divergence time was estimated with the formula $T = K/2r$, where r represents a synonymous substitution rate of 1.5×10^{-8} per site per year (Koch et al., 2000). K is the number of synonymous substitutions per synonymous site (Ks), which was calculated using the MS method using the Ka/Ks calculator (Zhang et al., 2006). To learn about the distribution rules for divergence time of TEs in each BAC, the distributed map of divergence time of the TEs was drawn using DPS 7.55 (Hangzhou Refine Information Tech. Co., Ltd, China).

Functional Annotation

The genes detected in the transposon-rich BAC clones were used to perform functional annotation via Blast2Go (www.hindawi.com/journals/ijpg/2008/619832.html) (Conesa and Gotz, 2008). Genes were grouped into two classes based on function: (1) genes related to transposons (class I), including copia-like, gag-pol, and retroelement polyproteins (subdivided into retrotransposon Ty1-copia and retrotransposon Ty3-gypsy), and (2) non-transposon-related genes (class II), such as genes for 1-aminocyclopropane-1-carboxylate oxidase, cytidine deaminase, sodium dicarboxylate cotransporters, Ras-related GTP-binding, and ethylene-responsive transcription. An analysis of the molecular functions, biological processes, and cellular composition of the two classes was conducted.

Tandem Duplication and SSR Distribution

To explore the particular constructions of LTR-TEs, the type and content of tandem repeats and SSRs were analyzed. Tandem Repeat Finder (http://tandem.bu.edu/trf/trf.html) (Benson, 1999) with default values was used to recognize tandem duplication sequences. In addition, SSR Locator (da Maia et al., 2008) was used to search SSRs in all BACs and inside LTR transposons. The criteria for SSR selection were as follows: mononucleotide motif SSRs had to have ≥10 repeats, di-nucleotide motif SSRs ≥8 repeats, trinucleotide motif SSRs ≥6 repeats, tetra-nucleotide motif SSRs ≥5 repeats, penta- and hexa-nucleotide motif SSRs ≥6 repeats, and hepta- to deca-nucleotide motif SSRs ≥3 repeats.

SUPPLEMENTARY DATA

Supplementary Data are available at *Molecular Plant Online.*

FUNDING

This work was supported by National High Technology Research and Development Program of China (863 Program) (2011AA10A104) and Ministry of Agriculture, Modern Agricultural Industrial Technology System Program (CARS-13).

REFERENCES

Baker, R. E., and Rogers, K. (2005). Genetic and genomic analysis of the AT-rich centromere DNA element II of *Saccharomyces cerevisiae*. Genetics. **171**, 1463–1475.

Bennetzen, J. L. (2000). Transposable element contributions to plant gene and genome evolution. Plant Mol. Biol. **42**, 251–269.

Benson, G. (1999). Tandem repeats finder: a program to analyze DNA sequences. Nucleic Acids Res. **27**, 573–580.

Brady, T. L., Schmidt, C. L., and Voytas, D. F. (2008). Targeting integration of the *Saccharomyces* Ty5 retrotransposon. Methods Mol. Biol. **435**, 153–163.

Brandes, A., Thompson, H., Dean, C., and Heslop–Harrison, J. S. (1997). Multiple repetitive DNA sequences in the paracentromeric regions of *Arabidopsis thaliana* L. Chromosome Res. **5**, 238–246.

Burge, C., and Karlin, S. (1997). Prediction of complete gene structures in human genomic DNA. J. Mol. Biol. **268**, 78–94.

Carareto, C. M.,et al. (1997). Testing transposable elements as genetic drive mechanisms using *Drosophila P* element constructs as a model system. Genetica. **101**, 13–33.

Charles, M., et al (2008). Dynamics and differential proliferation of transposable elements during the evolution of the B and A genomes of wheat. Genetics. **180**, 1071–1086.

Cheng, Z., et al. (2002). Functional rice centromeres are marked by a satellite repeat and a centromere-specific retrotransposon. Plant Cell. **14**, 1691–1704.

Conesa, A., and Gotz, S. (2008). Blast2Go a comprehensive suite for functional analysis in plant genomics. Int. J. Plant Genomics. **2008**, 619832.

Crenes, G., Moundras, C., Demattei, M. V., Bigot, Y., Petit, A., and Renault, S. (2011). Target site selection by the mariner-like element, Mos1. Genetica. **138**, 509–517.

da Maia, L. C., Palmieri, D. A., de Souza, V. Q., Kopp, M. M., de Carvalho, F. I., and Costa de Oliveira, A. (2008). SSR locator: tool for simple sequence repeat discovery integrated with primer design and PCR simulation. Int. J. Plant Genomics. **2008**, 412696.

Du, J., et al. (2010). Evolutionary conservation, diversity and specificity of LTR-retrotransposons in flowering plants: insights from genome-wide analysis and multi-specific comparison. Plant J. **63**, 584–598.

Flavell, A. J., Pearce, S. R., and Kumar, A. (1994). Plant transposable elements and the genome. Curr. Opin. Genet. Dev. **4**, 838–844.

Gangadharan, S., Mularoni, L., Fain–Thornton, J., Wheelan, S. J., and Craig, N. L. (2010). DNA transposon Hermes inserts into DNA in nucleosome-free regions in *vivo*. Proc. Natl Acad. Sci. U S A. **107**, 21966–21972.

Gao, M., Li, G., McCombie, W. R., and Quiros, C. F. (2005). Comparative analysis of a transposon-rich *Brassica oleracea* BAC clone with its corresponding sequence in *A. thaliana*. Theor. Appl. Genet. **111**, 949–955.

Garber, K., et al. (1999). The Tpv2 family of retrotransposons of *Phaseolus vulgaris*: structure, integration characteristics, and use for genotype classification. Plant Mol. Biol. **39**, 797–807.

Gindullis, F., Desel, C., Galasso, I., and Schmidt, T. (2001). The large-scale organization of the centromeric region in *Beta* species. Genome Res. **11**, 253–265.

Gollotte, A., et al. (2006). Repetitive DNA sequences include retrotransposons in genomes of the *Glomeromycota*. Genetica. **128**, 455–469.

Grewal, S. I., and Rice, J. C. (2004). Regulation of heterochromatin by histone methylation and small RNAs. Curr. Opin. Cell Biol. **16**, 230–238.

Grover, A., Aishwarya, V., and Sharma, P. C. (2007). Biased distribution of microsatellite motifs in the rice genome. Mol. Genet. Genomics. **277**, 469–480.

Habu, Y. (2010). Epigenetic silencing of endogenous repetitive sequences by MORPHEUS' MOLECULE1 in *Arabidopsis thaliana*. Epigenetics. **5**, 562–565.

Hall, T. A. (1999). BioEdit: A user-friendly biological sequence alignment editor and analysis program for Windows 95/98/NT. Nucl. Acids Symp. Ser. **41**, 95–98.

Hass, B. L., Pires, J. C., Porter, R., Phillips, R. L., and Jackson, S. A. (2003). Comparative genetics at the gene and chromosome levels between rice (*Oryza sativa*) and wild rice (*Zizania palustris*). Theor. Appl. Genet. **107**, 773–782.

Hong, C. P., et al. (2006). A Survey of the *Brassica rapa* genome by BAC-end sequence analysis and comparison with *Arabidopsis thaliana*. Mol. Cells. **22**, 300–307.

Hong, C. P., et al. (2007). Genomic distribution of simple sequence repeats in *Brassica rapa*. Mol. Cells. **23**, 349–356.

Huang, J., et al. (2009). Identification of a high frequency transposon induced by tissue culture, nDaiZ, a member of the hAT family in rice. Genomics. **93**, 274–281.

Kamm, A., Galasso, I., Schmidt, T., and Heslop–Harrison, J. S. (1995). Analysis of a repetitive DNA family from *Arabidopsis arenosa* and relationships between *Arabidopsis* species. Plant Mol. Biol. **27**, 853–862.

Kidwell, M. G. (2002). Transposable elements and the evolution of genome size in eukaryotes. Genetica. **115**, 49–63.

Kijima, T. E., and Innan, H. (2010). On the estimation of the insertion time of LTR retrotransposable elements. Mol. Biol. Evol. **27**, 896–904.

Koch, M. A., Haubold, B., and Mitchell–Olds, T. (2000). Comparative evolutionary analysis of chalcone synthase and alcohol dehydrogenase loci in *Arabidopsis, Arabis*, and related genera (*Brassicaceae*). Mol. Biol. Evol. **17**, 1483–1498.

Kuykendall, D., Shao, J., and Trimmer, K. (2009). A nest of LTR retrotransposons adjacent the disease resistance-priming gene NPR1 in *Beta* vulgaris L. U.S. Hybrid H20. Int. J. Plant Genomics. **2009**, 576742.

Le, Q. H., Wright, S., Yu, Z., and Bureau, T. (2000). Transposon diversity in *Arabidopsis thaliana*. Proc. Natl Acad. Sci. U S A. **97**, 7376–7381.

Lim, K. B., et al. (2005). Characterization of rDNAs and tandem repeats in the heterochromatin of *Brassica rapa*. Mol. Cells. **19**, 436–444.

Lim, K. B., et al. (2007). Characterization of the centromere and peri-centromere retrotransposons in *Brassica rapa* and their distribution in related Brassica species. Plant J. **49**, 173–183.

Lohe, A.R., Moriyama, E.N., Lidholm, D.A., and Hartl, D.L. (1995). Horizontal transmission, vertical inactivation, and stochastic loss of mariner-like transposable elements. Mol. Biol. Evol. 12, 62–72.

Ma, J., and Jackson, S. A. (2006). Retrotransposon accumulation and satellite amplification mediated by segmental duplication facilitate centromere expansion in rice. Genome Res. **16**, 251–259.

McClintock, B. (1944). The relation of homozygous deficiencies to mutations and allelic series in maize. Genetics. **29**, 478–502.

McClintock, B. (1951). Chromosome organization and genic expression. Cold Spring Harb. Symp. Quant. Biol. **16**, 13–47.

McNaughton, J. C., et al. (1993). A cluster of transposon-like repetitive sequences in intron 7 of the human dystrophin gene. J. Mol. Biol. **232**, 314–321.

Mills, R. E., Bennett, E. A., Iskow, R. C., and Devine, S. E. (2007). Which transposable elements are active in the human genome? Trends Genet. **23**, 183–191.

Minervini, C. F., Viggiano, L., Caizzi, R., and Marsano, R. M. (2009). Identification of novel LTR retrotransposons in the genome of *Aedes aegypti*. Gene. **440**, 42–49.

Mun, J. H., et al. (2010). Sequence and structure of *Brassica rapa* chromosome A3. Genome Biol. **11**, R94.

Nagaki, K., et al. (2004). Sequencing of a rice centromere uncovers active genes. Nat. Genet. **36**, 138–145.

Naito, K., et al. (2006). Dramatic amplification of a rice transposable element during recent domestication. Proc. Natl Acad. Sci. U S A. **103**, 17620–17625.

Naito, K., et al. (2009). Unexpected consequences of a sudden and massive transposon amplification on rice gene expression. Nature. **461**, 1130–1134.

Nefedova, L. N., Mannanova, M. M., and Kim, A. I. (2011). Integration specificity of LTR-retrotransposons and retroviruses in the *Drosophila melanogaster* genome. Virus Genes. **42**, 297–306.

Parisod, C., Salmon, A., Zerjal, T., Tenaillon, M., Grandbastien, M. A., and Ainouche, M. (2009). Rapid structural and epigenetic reorganization near transposable elements in hybrid and allopolyploid genomes in *Spartina*. New Phytol. **184**, 1003–1015.

Park, M., et al. (2011). Comparative analysis of pepper and tomato reveals euchromatin expansion of pepper genome caused by differential accumulation of Ty3/Gypsy-like elements. BMC Genomics. **12**, 85.

Parkin, I. A., et al. (2005). Segmental structure of the *Brassica napus* genome based on comparative analysis with Arabidopsis thaliana. Genetics. **171**, 765–781.

Piriyapongsa, J., Polavarapu, N., Borodovsky, M., and McDonald, J. (2007). Exonization of the LTR transposable elements in human genome. BMC Genomics. **8**, 291.

Pouilly, N., Delourme, R., Alix, K., and Jenczewski, E. (2008). Repetitive sequence-derived markers tag centromeres and telomeres and provide insights into chromosome evolution in *Brassica napus*. Chromosome Res. **16**, 683–700.

Ramsay, L., et al. (1999). Intimate association of microsatellite repeats with retrotransposons and other dispersed repetitive elements in barley. Plant J. **17**, 415–425.

Rice, P., Longden, I., and Bleasby, A. (2000). EMBOSS: the European Molecular Biology Open Software Suite. Trends Genet. **16**, 276–277.

Sabot, F., and Schulman, A. H. (2006). Parasitism and the retrotransposon life cycle in plants: a hitchhiker's guide to the genome. Heredity. **97**, 381–388.

SanMiguel, P., et al. (1996). Nested retrotransposons in the intergenic regions of the maize genome. Science. **274**, 765–768.

Sarilar, V., Marmagne, A., Brabant, P., Joets, J., and Alix, K. (2011). BraSto, a Stowaway MITE from *Brassica*: recently active copies preferentially accumulate in the gene space. Plant Mol. Biol. **77**, 59–75.

Schlötterer, C. (2000). Evolutionary dynamics of microsatellite DNA. Chromosoma. **109**, 365–371.

Smit, A. F. (1999). Interspersed repeats and other mementos of transposable elements in mammalian genomes. Curr. Opin. Genet. Dev. **9**, 657–663.

Smit, A. F.A., Hubley, R., and Green, P. (2011). Current Version: open-3.3.0 (RMLib: 20110419).

Smoot, M. E., Ono, K., Ruscheinski, J., Wang, P. L., and Ideker, T. (2011). Cytoscape 2.8: new features for data integration and network visualization. Bioinformatics. **27**, 431–432.

Sun, X., Le, H. D., Wahlstrom, J. M., and Karpen, G. H. (2003). Sequence analysis of a functional *Drosophila* centromere. Genome Res. **13**, 182–194.

Tamura, K., Dudley, J., Nei, M., and Kumar, S. (2007). MEGA4: Molecular Evolutionary Genetics Analysis (MEGA) software version 4.0. Mol. Biol. Evol. **24**, 1596–1599.

Tikhonov, A. P., SanMiguel, P. J., Nakajima, Y., Gorenstein, N. M., Bennetzen, J. L., and Avramova, Z. (1999). Colinearity and its exceptions in orthologous *adh* regions of maize and sorghum. Proc. Natl Acad. Sci. U S A. **96**, 7409–7414.

Tsukahara, S., Kobayashi, A., Kawabe, A., Mathieu, O., Miura, A., and Kakutani, T. (2009). Bursts of retrotransposition reproduced in *Arabidopsis*. Nature. **461**, 423–426.

Turcotte, K., Srinivasan, S., and Bureau, T. (2001). Survey of transposable elements from rice genomic sequences. Plant J. **25**, 169–179.

U, N. (1935). Genome analysis in *Brassica* with special reference to the experimental formation of *B. napus* and peculiar mode of fertilization. Japanese Journal of Botany. **7**, 389–452.

Vieira, C., Lepetit, D., Dumont, S., and Biémont, C. (1999). Wake up of transposable elements following *Drosophila simulans* worldwide colonization. Mol. Biol. Evol. **16**, 1251–1255.

Wang, X., et al. (2011). The genome of the mesopolyploid crop species *Brassica rapa*. Nat. Genet. **43**, 1035–1039.

Warburton, P. E. (2004). Chromosomal dynamics of human neocentromere formation. Chromosome Res. **12**, 617–626.

Wicker, T., et al. (2007). A unified classification system for eukaryotic transposable elements. Nat. Rev. Genet. **8**, 973–982.

Xu, Z., and Wang, H. (2007). LTR_FINDER: an efficient tool for the prediction of full-length LTR retrotransposons. Nucleic Acids Res. **35**, W265–W268.

Zhang, X., and Wessler, S. R. (2004). Genome-wide comparative analysis of the transposable elements in the related species *Arabidopsis thaliana* and *Brassica oleracea*. Proc. Natl Acad. Sci. U S A. **101**, 5589–5594.

Zhang, Z., Li, J., Zhao, X. Q., Wang, J., Wong, G. K., and Yu, J. (2006). KaKs_Calculator: calculating Ka and Ks through model selection and model averaging. Genomics Proteomics Bioinformatics. **4**, 259–263.

Zhou, M. B., Liu, X. M., and Tang, D. Q. (2011). Transposable elements in *Phyllostachys pubescens* (Poaceae) genome survey sequences and the full-length cDNA sequences, and their association with simple-sequence repeats. Genet Mol. Res. **10**, 3026–3037.

Zou, S., Ke, N., Kim, J. M., and Voytas, D. F. (1996). The *Saccharomyces* retrotransposon Ty5 integrates preferentially into regions of silent chromatin at the telomeres and mating loci. Genes Dev. **10**, 634–645.