Chinese Science Bulletin

© 2009 SCIENCE IN CHINA PRESS
Springer

Identification and characterization of *CACTA* transposable elements capturing gene fragments in maize

LI Qing[1], LI Lin[1], DAI JingRui[1], LI JianSheng[1] & YAN JianBing[1,2†]

[1] National Maize Improvement Center of China, China Agricultural University, Beijing 100193, China;
[2] International Maize and Wheat Improvement Center (CIMMYT), Apdo. Postal 6-641, Mexico 06600, Mexico

Transposable elements (TEs)-mediated gene sequence movement is thought to play an important role in genome expansion and origin of genes with novel functions. In this study, a gene, *HGGT*, involved in vitamin E synthesis was used in a case study to discover and characterize transposons carrying gene fragments in maize. A total of 69 transposons that are distributed across the 10 chromosomes and have an average length of 3689 bp were identified from the maize sequence database by using the BLAST search algorithm. Three of these carry gene fragments from the progenitor *HGGT* gene, while the rest (66) contain gene fragments from other cellular genes. Nine of the 69 transposons contain fragments derived from two locations in the genome. By querying the maize Expressed Sequence Tag (EST) database, we found that at least thirteen out of the 69 TEs had corresponding transcripts. More interestingly, two transposons that carry gene fragments from two different chromosomal loci could be expressed as chimeric transcripts.

CACTA elements, gene fragment, genome distribution, chimeric transcripts, maize

Transposable elements (TEs), which move from place to place in the genome and often make duplicates of themselves, have tremendous impact on genome evolution, origin of novel genes and phenotypic variation[1–3]. Based on the form of their transposition intermediate, TEs can be divided into two classes[4]: Class I elements such as LINE and SINE that are transposed via an RNA intermediate and class II elements such as *Mu* and *Helitron* elements that are transposed directly from DNA to DNA. *En*/*Spm* elements belong to class II and are widely distributed in genomes of plants like maize[5,6], sorghum[7], rice[8], soybean[9] and others. *En*/*Spm* elements are also referred to as *CACTA* elements because they carry the conserved 5 bp sequence 'CACTA' in the outermost terminal inverted repeats (TIRs)[10]. *CACTA* elements are characterized by the generation of a 3 bp target site duplication (TSD) upon insertion and the widespread occurrence of sub-terminal repeats (STRs). The STRs are usually regarded as the binding sites of transposase and are thus related to the transposition of TEs[11].

The enhancer (*En*) and suppressor-mutator (*Spm*) elements in maize that were independently identified by Peterson[5] and McClintock[6] are genetically and molecularly homologous[12–16]. Both are autonomous elements since they encode the transposase that is required for transposition. The corresponding non-autonomous inhibitor (*I*)[5] and defective *Spm* (*dSpm*)[16] are deletion derivates of the two autonomous elements[13–15], and are capable of transposition only in the presence of a functional transposase. In addition to those non-autonomous elements derived from autonomous elements, other types of non-autonomous elements that carry gene segments have also been found to be widely distributed across species. The first elements discovered

Received October 22, 2008; accepted December 14, 2008
doi: 10.1007/s11434-009-0061-2
†Corresponding author (email: yjianbing@cau.edu.cn)
Supported by National Natural Science Foundation of China (Grant No. 30500322) and National Hi-Tech Research and Development Program of China (Grant Nos. 2006AA10Z183, 2006AA10A107)

were the *Mu* elements in maize[17], and, subsequently, *Mutator*-like elements (MULE) carrying gene segments were identified in rice[18–20], Arabidopsis[21], Lotus[22] and melon[23]. *Helitron*, which was recently identified in maize, is also capable of acquiring gene fragments[24–26]. *CACTA* elements that capture host DNA segments have also been identified in Japanese morning glory[27,28], soybean[9] and Antirrhinum[29], but there were no reports of such elements in maize. The objectives of this study were to identify and characterize *CACTA* elements carrying gene fragments in maize and elucidate their role in helping to understand the mechanism of genome expansion and origin of genes with novel functions.

1 Materials and methods

1.1 Determination of TE feature sequences

The repeat sequences of TE were identified using the CENSOR software[30]. Analysis of TSD and TIRs were done manually based on their respective characteristics, that is, TSD are usually 3 bp in length and the left and right TIRs are usually reverse complementary to each other. STRs were analyzed with the DNAMAN software (Version 5.2.2, Lynnon Biosoft, Canada).

1.2 The genome-wide search of TEs

Two methods were used to search the maize B73 genome for TEs carrying gene fragments, namely a BLAST method and a search method based on previously characterized TIRs and TSD. In the BLAST method: the repeat sequences identified by CENSOR in T2, T3 and T9 (see results) were aligned using MUSCLE[31]. A consensus sequence (Figure 1 (a)) obtained with BioEdit[32] software was used to query the HTGS database of maize (GenBank Release164, downloaded in June 2008) to identify sequences starting with CACTA and ending with TAGTG. These sequences generated four position numbers in the BACs where they are incorporated (Figure 1 (b)): F1, F2 (the end position of 5′ alignment), F3 (the start position of 3′ alignment) and F4. Sequences that matched the criteria outlined below were recovered: F1<F2<F3<F4 or F1>F2>F3>F4, no gaps and $|F3-F2|-1<10$ kb.

In the search method based on previously characterized TIRs and TSD, two short sequences $N_1N_2N_3$-CACTA and TAGTG$N_1N_2N_3$ (N represents any of the four nucleotides, where the two $N_1N_2N_3$ are the same) were screened for across the whole genome in order to retrieve sequences with a length of less than 10 kb between the two short sequences. If the first and last 14 bp (excluding $N_1N_2N_3$) of the retrieved sequences were reverse complementary to each other, and if either the first 17 bp were $N_1N_2N_3$CACTACACCAAAAT or the last 17 bp were ATTTTGGTGTAGTG$N_1N_2N_3$, the sequences would be used in the non-redundant analysis.

A non-redundant analysis was performed under the condition that TEs were identical across the entire length, or if the identity was less than 100%, the two BACs containing the TEs had to overlap on the maize physical map (http://www.genome.arizona.edu/fpc/WebAGCoL/maize/WebFPC/), the TSD must be the same, and the length of the TEs must be similar (±1).

1.3 Annotation and copy number estimation

Online Blast2GO (http://www.blast2go.de/) was utilized to annotate all the TEs with cutoff *E* value≤−10 (database updated in June 2008). We identified candidate homologous TEs if they were annotated to a similar protein. The identified TEs were used as query sequences

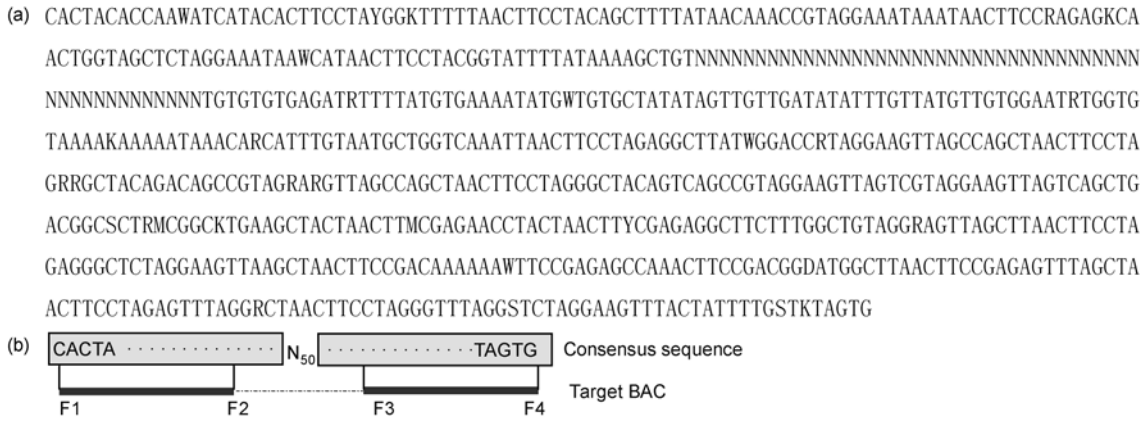

Figure 1 Consensus sequence used in the BLAST method (a) and schematic diagram of the typical blast result (b).

to blast against the maize HTGS database (http://www.ncbi.nlm.nih.gov/). TEs that match each other were defined as copies of the same element.

1.4 PCR amplification of TEs

Primers were designed using Primer3 (http://frodo.wi.mit.edu/). PCR was performed in a 15 μL reaction volume containing 40 ng DNA, a final concentration of 1× Taq buffer (including Mg^{2+}), 200 μmol/L each dNTP, 0.4 μmol/L each primer and 0.5 U Taq polymerase. The PCR program is as follows: 1 cycle at 94℃ for 3 min; 94℃, 30 s, 58℃, 50 s, 72℃, 1 min, 35 cycles; and a final extension cycle at 72℃ for 10 min. Sequencing was carried out at the SunBio Company. Primers used in this study were: 1F: 5′-GATTGTGGCTATGTATTAGGG-3′, 1R: 5′-AGGAAAGAGGCATGGATTA-3′, 2F: 5′-CGCTGCTTCTCGTCTCCAC-3′, 2R: 5′-TCGTATGC-CAATCGCCTGA-3′, 3F: 5′-CCCGCTGCTTCTCATC-TCC-3′ and 3R: 5′-GAGGAATGCCTGCGTTGTC-3′.

1.5 Phylogenetic analysis

The 100 bp sequences from both the 5′ and 3′ ends of the TEs annotated to transposase or cellular genes were extracted and fused to form a 200 bp sequence, which was then aligned using the multiple sequence alignment tool, MUSCLE[31]. The maximum likelihood method in the PHYLIP package[33] was used to construct a phylogenetic tree with a bootstrap value of 100. A consensus tree was constructed using MEGA3.0[34] with a cutoff of value greater than 50.

1.6 Expression analysis

The method used for expression analysis was a modified version of the method described by Jiang et al.[18]. The unmasked sequences with CENSOR were used as queries to search against the maize EST database (updated in June 2008). The TE was considered to have an EST match if, (i) the sequence similarity between the TE and the EST was greater than 99% with a coverage of >95% of the entire EST length, and (ii) the best match of the EST in the maize HTGS database was the original TE and no other better matches existed in the whole genome.

2 Results

2.1 Identification of *CACTA* elements carrying gene fragments in maize

(i) Parts of the *HGGT* gene were captured by *CACTA* elements. Homogentisic acid geranylgeranyl transferase (*HGGT*) catalyzes the committed step of vitamin E biosynthesis[35]. The *HGGT* gene is located on chromosome 9 in maize genome. The BAC clone, AC210150, contains the full-length sequence of *HGGT* gene and another three BACs contain part of *HGGT* gene (Figure 2). For example, AC215796, a BAC clone located on chromosome 9, contains *HGGT* gene sequences from intron 2 to intron 6 and intron 11 to the 3′ UTR (the two parts were directly joined with no other interleaved sequences). The similarities between each of two segments and the progenitor gene are 92% and 96%, respectively (Figure 2).

The upstream and downstream 4 kb sequences flanking the homologous regions in the 3 BACs show high levels of similarity to ENSPM2_ZM, a class of *CACTA* elements, implying that they might also be TEs. The typical 3 bp TSD and 14 bp TIRs were identified in both AC215501 and AC214002. AC215796 also contains an identifiable TSD and TIRs, although some variations were observed (Table 1). The first 5 bp of one of the two TIRs are CACTA, demonstrating they are likely members of *CACTA* elements. The three TEs were named T2, T3 and T9 according to their location on the chromosomes, and are 4202, 3623 and 3349 bp in length, re-

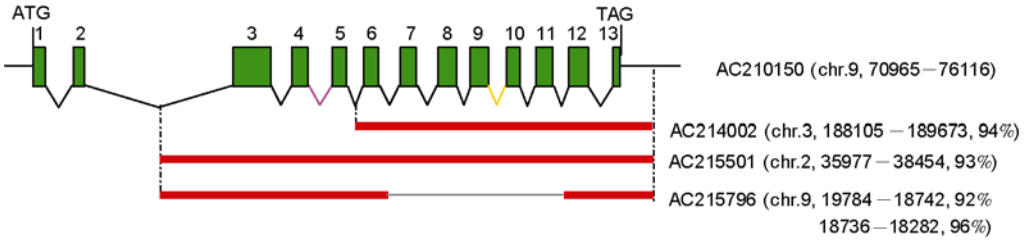

Figure 2 Schematic diagram of matches from the NCBI maize HTGS database that match the maize *HGGT* gene. Filled green boxes represent exons, introns are shown by "V" lines. The red lines indicate the sequences that show similarity to the progenitor *HGGT* gene, and the grey lines represent the sequences deleted from AC215796. The position of matched sequences in the BAC and their percent identity with the progenitor *HGGT* gene are shown on the right. The alternatively spliced introns of AC215501 and AC215796 are highlighted by yellow and pink lines, respectively.

Table 1 Features of T2, T3, T9 and ENSPM2_ZM

Name	BAC	Position	Length (bp)	TSD	TIR[a)]
T2	AC215501	34906–	4202	GAC	CACTACACCAAAAT
		39107		GAC	ATTTTGGTGTAGTG
T3	AC214002	186704–	3623	CAA	CACTACACCAA*T*AT
		190326		CAA	ATTTTG*C*TGTAGTG
T9	AC215796	17640–	3349	TGG	C*CA*TACACCAAAAT
		20988		TGA	ATTTTGGTT*T*AGT*A*
ENSPM2_ZM[b)]	–	–	7628	–	CACTACACCAAAAT
					ATTTTGGTGTAGTG

a) The underlined 5 nucleotides represent the conserved sequence for which *CACTA* elements were named. The italic letters indicate mutated nucleotides compared with the T2 sequence. All the sequences were in 5′ to 3′ orientation. b) ENSPM2_ZM is a consensus sequence deposited in the CENSOR database.

spectively. The TIRs of T2, T3, T9 and ENSPM2_ZM are highly similar (Table 1). The 8 bp motif TAACTTCC, occurring at both the 5′ and 3′ ends as direct or inverted repeats (Figure 5), are identical, indicating that T2, T3 and T9 belong to ENSPM2_ZM family.

(ii) *CACTA* elements that carry gene fragments in B73 genome. A total of 238 TEs were identified in B73 genome using the BLAST method while 233 TEs were tagged in the whole-genome scan using the sequences of previously characterized TIRs and TSD. As most of the TEs identified using the two different methods were overlapped, a total of 219 non-redundant TEs (including T2, T3 and T9) were obtained and these were designated ES2-1 to ES2-219.

As shown in Figure 3, at $E \leqslant -10$, 69 TEs (31.5%) contained sequences corresponding to previously identified proteins involved in 43 different cellular functions, 32 TEs (14.6%) contained sequences similar to transposase or other transposon-related proteins, 20 TEs (9.1%) were homologous to predicted or hypothetical proteins and 98 TEs (44.7%) showed no similarity to either transposon-related, cellular or hypothetical proteins. Therefore, at least 69 ES2 TEs carrying gene fragments were identified in the B73 genome.

2.2 Characteristics of *CACTA* elements carrying gene fragments

(i) Distribution of *CACTA* elements carrying gene fragments in the maize genome. The 69 TEs identified in this study were distributed across the 10 maize chromosomes. Chromosome 3 and chromosome 4 had the highest number of TEs, with each having 11 TEs, while chromosome 8 had the lowest number, having only 2 TEs (Figure 4(a)). The 3 bp TSD was rich in A/T nucleotides. The 20 bp sequences flanking the TSD were also rich in A/T nucleotides with an exception of the innermost nucleotide tending to be G or C (Figure 4(c)). The accumulated length of the 69 TEs is 254.5 kb (sequences between but excluding the TSD), and accounted for 0.01% of the maize genome (254.5 kb/2500 Mb). Of these TEs, 65 had a length between 2 and 5 kb, while four were longer than 5 kb (Figure 4(b)), and the average length was 3689 bp.

The copy number of each TE ranged from one to five (Figure 3). Twenty-six of the 69 TEs were single-copy elements, eight had two copies, six had three copies, one TE had 4 and 5 copies, respectively. TEs with more than five copies were not identified. The average copy num-

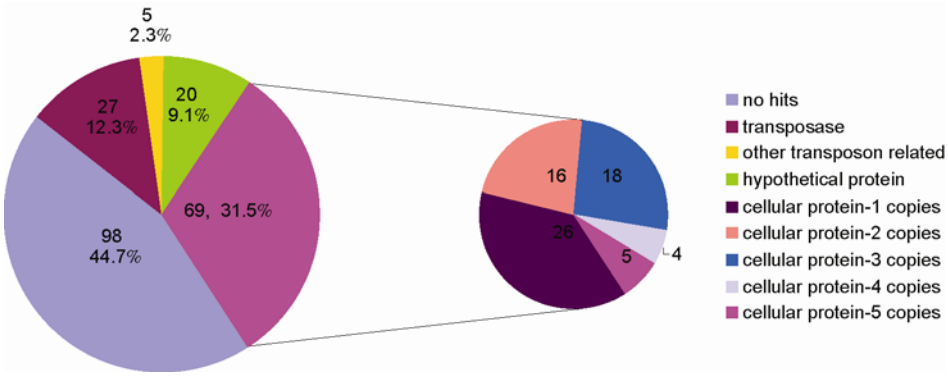

Figure 3 Annotation of ES2 transposable elements.

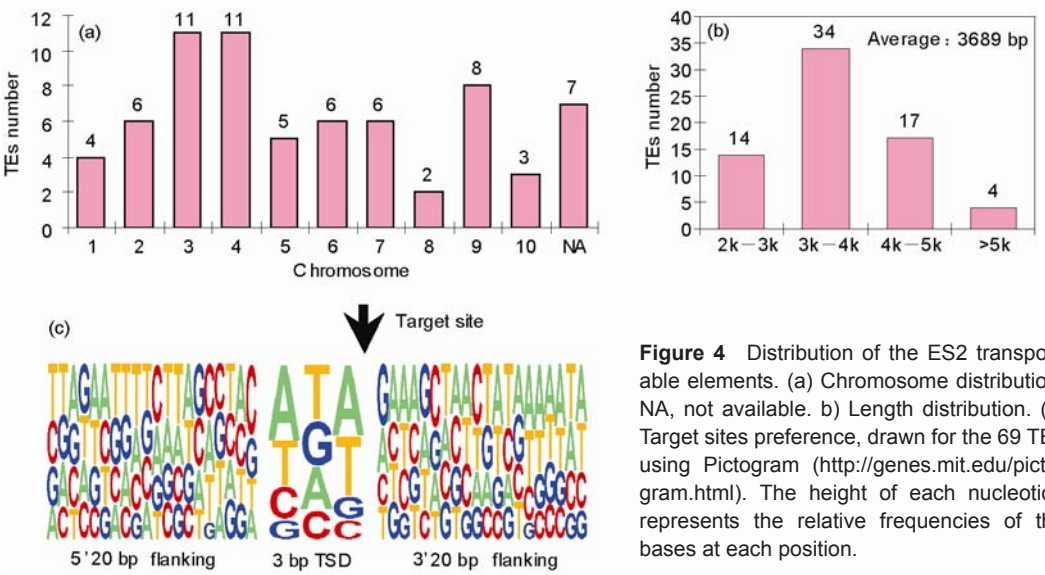

Figure 4 Distribution of the ES2 transposable elements. (a) Chromosome distribution. NA, not available. b) Length distribution. (c) Target sites preference, drawn for the 69 TEs using Pictogram (http://genes.mit.edu/pictogram.html). The height of each nucleotide represents the relative frequencies of the bases at each position.

ber was 1.6 (69/42), implying that the ES2 TEs are low-copy-number elements.

(ii) *CACTA* elements can carry gene fragments from different genomic regions. The internal sequences of T2, T3 and T9 after masking with CENSOR were 3558, 2211 and 2582 bp in length, respectively. Sequences of T2, T3 and T9 inferred from the *HGGT* gene had lengths of 2496, 1526 and 1587 bp, respectively. The length difference between internal sequences of T2, T3, T9 and sequence from the *HGGT* gene indicated that each TE contained a stretch of sequences that did not come from the *HGGT* gene nor did it represent repeats. These sequences were further used to query the HTGS database at NCBI. The results showed that the sequence in T3 was similar to BAC-AC187813 on chromosome 4 with an identity of 94%, while the sequences in T2 and T9 were both similar to BAC-AC206980 on chromosome 2 with an identity of 90% (Figure 5(a)). Two primer pairs, with primer pair 1 specific for T3 and primer pair 2 specific for T2 and T9, were used to amplify the B73 genome. A single band was obtained with primer pair 1 (Figure 5(b), lane2), and its identity to T3 was subsequently confirmed by sequencing. Primer pair 2 also gave a single band (Figure 5(b), lane3), but the sequencing results revealed that both T2 and T9 had been amplified (Figure 5(c)). Collectively, these results showed that T2, T3 and T9 contained fragments from at least 2 different chromosomal loci, and they were not artefacts of sequence assembly errors. Similarly, of the remaining 66 TEs that contained sequences associated with cellular proteins, 6 had homolog to two different proteins (Table 2), implying that they also might have captured gene segments from different genomic regions.

A stretch of 100 bp from each end of TEs carrying gene fragments or TEs with sequences homologous to transposase were extracted for the phylogenetic analysis. The results showed that TEs carrying cellular gene fragments could be clustered together, in a group that was distinct from TEs with sequences homologous to transposase (Figure 6). These results implied that the sequence divergence between the two kinds of TEs might indicate a prerequisite for gene capture.

(iii) Gene fragments contained in the elements have transcriptional activities. The internal sequences of T2, T3 and T9 were used to query the maize EST database at NCBI. The results revealed that both T2 and T9 were represented in the EST database, while T3 had no corresponding EST (Figure 7). Three of the five ESTs having similarity to T2, DV515366, DR785294 and BT034581, corresponded to sequences from the *HGGT* gene and AC206980, indicating that sequences from the two loci had been co-transcribed to form chimeric transcripts (Figure 7(a)). A similar phenomenon was observed for T9, where two of the six ESTs, EB642233 and BT038826, represent chimeric transcripts. The phenomenon of chimeric transcripts was also confirmed by RT-PCR analysis (Figure 7(b)). Alternative splicing was observed in both T2 and T9 in the 9th and 4th introns of the progenitor *HGGT* gene, respectively (Figure 2 and Figure 7(a)). In addition, the mode of splicing for the sequences from AC206980 was different in T2 and T9, as a stretch of sequences spliced out of T2 was retained

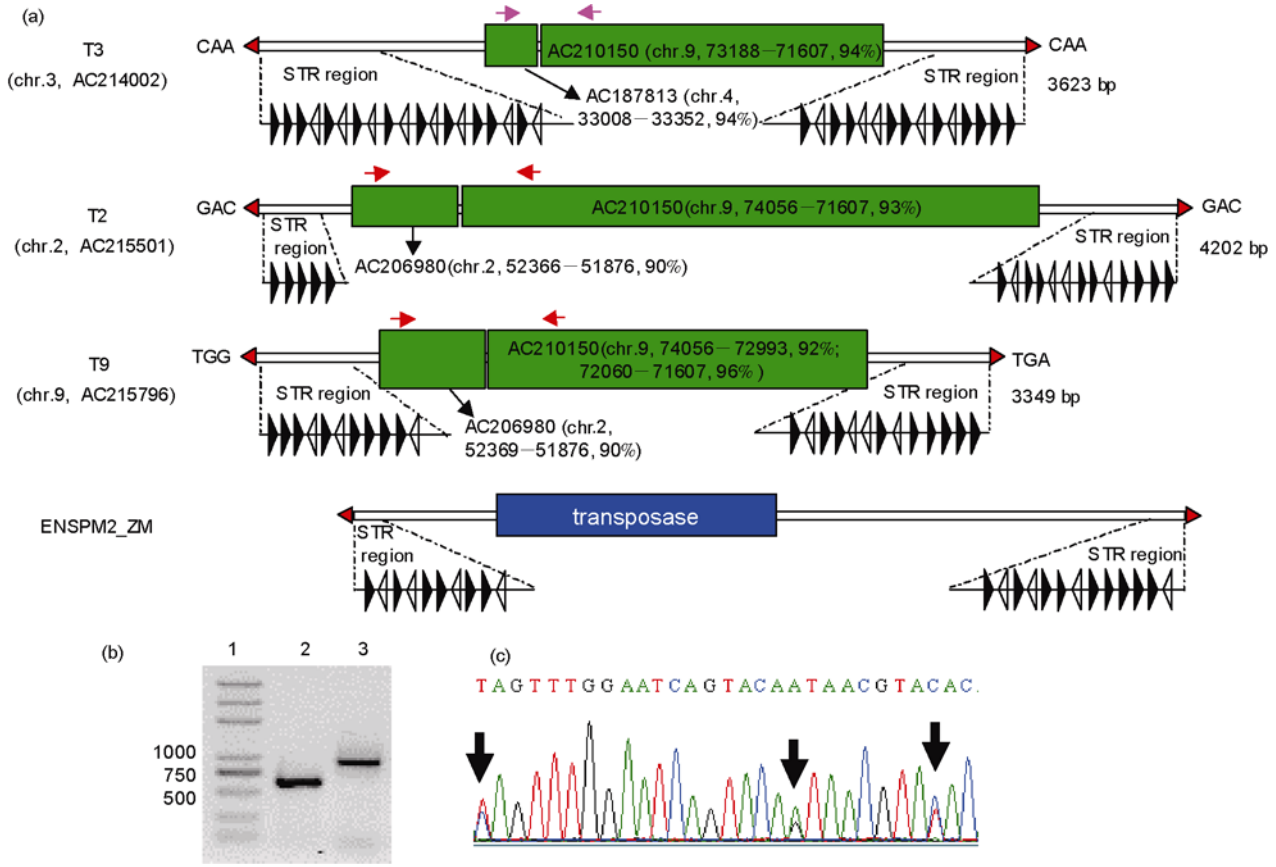

Figure 5 Transposable element structure and PCR amplification. (a) TIRs are indicated by red filled triangles. The filled and open black triangles represent the 8 bp direct and inverted STRs, respectively. Green boxes represent the gene fragments carried by each TE, each fragment's corresponding progenitor sequences and the percent identity between them are also shown. Pink and red arrows indicate primer pair 1 and primer pair 2, respectively. (b) PCR products of primer pair 1 (lane2) and primer pair 2 (lane3) amplified with B73 genome. Lane1 is a DL2000 Plus DNA Marker. (c) Sequencing chromatogram of primer pair 2. Black arrows represent double peaks due to nucleotide acid differences between T2 and T9.

Table 2 Transposable elements with two protein hits

ID	BAC[a)]	Chromosome	Position[b)]	Length (bp)	Copy	*E* value[c)]	Annotation
ES2-91	AC197362	8	101962−105741	3780	3	8.00×10^{-36}	sigma factor protein
						1.00×10^{-16}	CIPK-like protein
ES2-115	AC149034	NA	110312−114443	4132	1	2.00×10^{-19}	DEAD-box ATP-dependent RNA helicase
						1.00×10^{-17}	Os03g0276600
ES2-134	AC214230	NA	81652−86041	4390	2	1.00×10^{-34}	hypothetical protein OsI_007935
						2.00×10^{-31}	U3 snoRNA-associated protein 11
ES2-142	AC186590	6	23277−28334	5058	5	1.00×10^{-50}	putative protein kinase AtSIK
						3.00×10^{-24}	putative 3-ketoacyl-CoA thiolase
ES2-149	AC208430	5	156174−161550	5377	4	7.00×10^{-29}	putative kinase
						2.00×10^{-24}	endoplasmatic reticulum retrieval protein
ES2-154	AC212442	4	171638−178544	6907	2	2.00×10^{-29}	one repeat MYB transcriptional factor
						3.00×10^{-16}	U3 snoRNA-associated protein 11

a) Accession number of the BAC in GenBank; b) start and end position of the TEs in the BAC; c) *E* value of annotation.

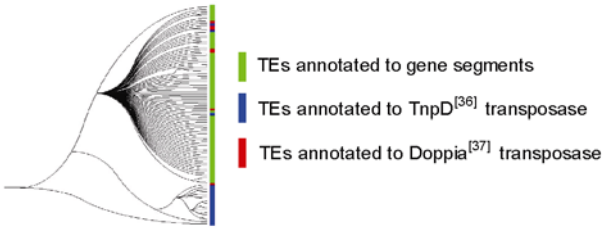

Figure 6 Phylogenetic analysis.

in T9 (Figure 7).

Eleven of the other 66 TEs carrying gene fragments were represented by ESTs (Table 3), of which three had only one EST, and eight had two or more ESTs. ES2-67 was represented by 13 ESTs. The number of corresponding ESTs may reflect the level of transcription. Similar to T2 and T9, alternative splicing phenomenon was observed in 7 of the 8 TEs which were represented by at least 2 ESTs (Table 3). In addition, three TEs, (ES2-115, ES2-142, ES2-154), each associated with the sequences of two genes, are capable of transcription, however, no chimeric transcripts were identified. This lack of chimeric transcripts may be due to the limited number of ESTs in the database.

3 Discussion

3.1 *CACTA* elements carrying gene segments exist in the maize genome

In this study, we found that parts of the *HGGT* gene had been acquired by 3 *CACTA* elements. Compared with other non-autonomous TEs derived from autonomous elements, the 3 *CACTA* elements lacked any remnants of transposase, but the characteristic TIRs and TSD were retained. The internal sequences were composed of identifiable fragments from at least 2 different genomic regions, the *HGGT* gene and sequences from either chromosome 2 or chromosome 4. The chimeric ESTs in the database, the PCR product with the genomic DNA and the RT-PCR product with the cDNA provide strong evidence that TEs carrying gene segments from different genomic regions exist in the maize genome and are not artefacts of sequence assembly errors. The identification of an additional 66 TEs distributed across the 10 chromosomes of maize by querying the sequences of 16,531 maize BACs suggests that TEs carrying gene fragments widely occur in the maize genome. With an average length of 154 kb per BAC[38], a total of 2546 Mb of sequence (the maize genome is estimated to be 2300—2700 Mb) was used in the whole-genome scan. Considering possible BAC overlaps, the genome coverage may be overestimated, thus, the 69 TEs obtained in this study represent a conserved estimate of the expected number of TEs in the maize genome. As the TEs contain both exons and introns, gene fragments capture is likely to involve the acquisition of genomic DNA rather than cDNA copies. The ability of TEs to capture gene seg-

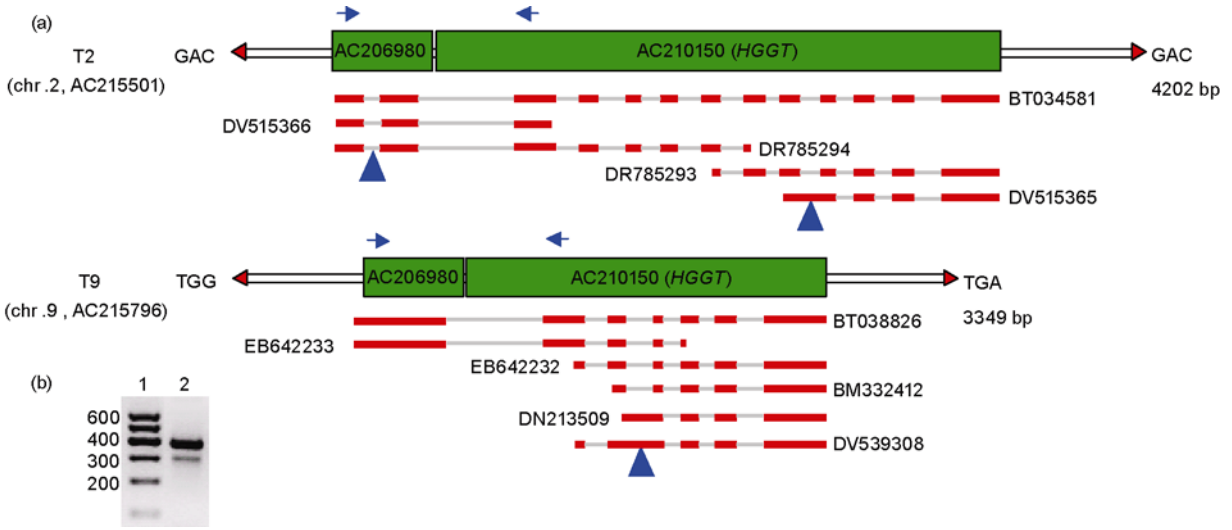

Figure 7 Transcriptional activities of transposable elements. (a) Red triangles represent TIRs, green boxes indicate the gene segments captured by TEs. Exons are shown as red lines, while introns are indicated by grey lines. The GenBank accession number of each EST is also shown. Alternatively spliced introns in different ESTs are indicated by blue triangles. The blue arrows indicate primer pair 3. (b) RT-PCR products of primer pair 3 amplified in the immature ears of inbred zong3 (lane2). Lane1, DNA marker. The 400 and 300 bp bands correspond to T9 and T2 transcripts, respectively.

Table 3 Transposable elements with transcriptional activities

ID	BAC[a)]	Chromosome	Position[b)]	Length (bp)	Copy	EST[c)]	AS[d)]	*E* value[e)]	Annotation
ES2-59	AC190847	NA	96586－99784	3199	1	**EE332027**	n	1.00×10^{-34}	peroxisomal 3-ketoacyl-CoA thiolase 2 precursor
ES2-116	AC209072	5	22161－26344	4184	3	CD941388	n	4.00×10^{-22}	HAD-superfamily hydrolase
ES2-52	AC187920	NA	72752－76143	3392	1	EE047636/DN216834	y	2.00×10^{-21}	cellulose synthase catalytic subunit
ES2-36	AC210796	2	156116－158925	2810	1	DV174513/EC875310/EE028235 /EB159347/DY623477/CF636757 /DN230273/DN229640	y	2.00×10^{-25}	putative auxin efflux carrier
ES2-90	AC194109	8	87232－91228	3997	1	DN560136/DV174286/DT948532 /DY686177/DT948531/CO533420 /**CO533419**/CD996297/DV174287	y	5.00×10^{-10}	26S proteasome regulatory particle non-ATPase subunit12
ES2-67	AC215995	9	81662－85106	3445	3	EE287615/**BM080212**/BE511517 /EC894882/EC894881/DR811361 /CD670441/**DY400559**/BQ667757 /BE639892/CA830085/BE453784 /*DW864676*	y	2.00×10^{-18}	KH domain-containing protein-like
ES2-32	AC213592	3	97729－100343	2615	2	EC894675/DN218861/DR958808 /CD650968/DV173879/DN219854 /DW933796/CD650773/DV173878	y	8.00×10^{-22}	importin alpha 1b
ES2-49	AC193518	2	71305－74396	3092	2	EE177428/DN216421/**BM333663** /**DN226876**/DR827220/DR827219	y	8.00×10^{-22}	importin alpha 1b
ES2-115[f)]	AC149034	NA	110312－114443	4132	1	BM429047/DY398718	n	2.00×10^{-19}	DEAD-box ATP-dependent RNA helicase
								1.00×10^{-17}	Os03g0276600
ES2-142[f)]	AC186590	6	23277－28334	5058	3	DR801744/DV511238 /DR801743/DR831406	y	1.00×10^{-50}	putative protein kinase AtSIK
								3.00×10^{-24}	putative 3-ketoacyl-CoA thiolase
ES2-154[f)]	AC212442	4	171638－178544	6907	2	*CD960507*	n	3.00×10^{-16}	U3 snoRNA-associated protein 11
								2.00×10^{-29}	one repeat MYB transcriptional factor

a) Accession number of the BAC in GenBank. b) The start and end position of the TEs in the BAC. c) Accession number of ESTs in GenBank, ESTs with a polyA tail were bold-faced, ESTs which could be translated into functional protein in one of the six ORFs were italic. d) AS, alternative splicing; y, yes; n, no. e) *E* value of annotation. f) TEs with two protein hits.

ments is remarkable, for example, the complete sequences from intron 2 to the 3′ UTR of the progenitor *HGGT* gene had been captured by T2.

3.2 *CACTA* elements and gene amplification

The *CACTA* elements carrying fragments of the *HGGT* gene had three copies in the maize genome. Three possible explanations could be proposed for this phenomenon: firstly, genome duplication; secondly, the insertion of *CACTA* elements into other TEs[39]; thirdly, transposition driven by transposase. The fact that the TSD of the 3 TEs is different excludes the first two possibilities. Under the third conditions, two different processes could be postulated for the amplification of the original progenitor transposon T′ (Figure 8): First, T′ acquired sequence from intron 2 to the 3′ UTR of the progenitor *HGGT* gene (Figure 8, V1, V2 and V3; Figure 2). It was then transposed twice. At one time, T′ acquired sequences from BAC-AC187813 (B) located on chromosome 4 with the loss of V1 (intron 2 to intron 5) to form T3. At another time, T′ acquired sequences from BAC-AC206980 (A) located on chromosome 2 to form T2, which then transposed to a new location losing V2 (intron 6 to intron 11) to form T9. Under the second situation, the formation of T2 and T9 are similar to the process above, the difference is that the formation of T3 is through direct acquisition of sequences from the progenitor *HGGT* gene (V2 and V3) and is not dependent on T'. The fact that the 3′ region was almost split from the same position in the 3 TEs (Figure 2) gives weight to the first situation.

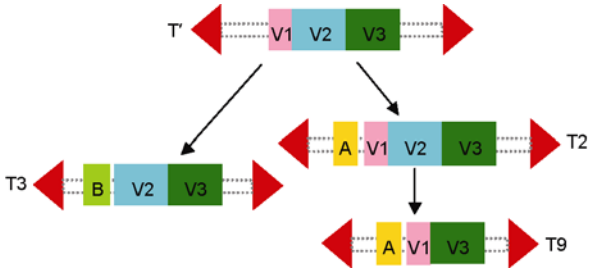

Figure 8 Schematic diagram of amplification of T2, T3 and T9. TIRs are indicated by red filled triangles. V1, V2 and V3 represent stretches of sequences from progenitor *HGGT* gene. Sequences from other regions are indicated by B and A.

In addition to T2, T3 and T9, other TEs with multiple copies were also identified as having a distinct TSD. Therefore, it is quite possible that these TEs had transposition activity in the past. All the TEs identified in this study belong to the family of ENSPM2_ZM elements, which cannot encode transposase, and thus can only be transposed in the presence of corresponding autonomous elements. No such autonomous elements have been reported previously. Some ES2 TEs contained sequences that can be annotated to transposase Doppia[37], implying that Doppia could be the autonomous elements. However, the difference in TIRs and the supposed transposase binding site-STR[11] between ES2 TEs and *Doppia* elements clearly lead us to reject this proposal. Like other cellular gene fragments, the *Doppia* fragments might have been captured by ES2 TEs. The fact that TEs with *Doppia* fragments and TEs with cellular gene fragments could be grouped together into a branch strongly supports this inference (Figure 6). The movement of gene fragments mediated by TEs has a great impact on the host genome. This movement could contribute to the lack of gene collinearity between or within species as found in *Helitron* elements[25,26]. The high similarity between the gene fragments and the progenitor host genes caused great difficulties in gene cloning and functional analysis.

3.3 Gene fragments captured by *CACTA* elements and pseudogenes

Of the 69 TEs that contained sequences annotated to cellular proteins, 13 (18.8%), are represented by ESTs in the queried databases. In addition, the fact that T2 and T9 carry sequences from two different chromosomal loci and that they could be expressed as chimeric transcripts is a circumstance that aids the origin of novel genes. Jiang et al.[18] analyzed over 3000 Pack-MULEs in rice and concluded that more than 10% of them might have been functionally constrained. By contrast, the study by Juretic et al.[20] suggested that virtually all TEs contain pseudogenic features such as frameshifts and premature stop codons. In this study, all the ESTs have features preventing them from encoding functional proteins. However, this does not imply that there was no function. The transcripts could regulate expression of the progenitor gene either at the RNA level by RNAi, or at the protein level by competing with the original functional protein for substrates[20]. Further research on TEs carrying gene fragments should provide more information and help to better understand genome structure, evolution and function in maize.

The authors thank Drs. George Mahuku and Trushar Shah for helping to improve the English writing.

1 Bennetzen J L. Transposable element contributions to plant gene and genome evolution. Plant Mol Biol, 2000, 42: 251－269
2 Bennetzen J L. Transposable elements, gene creation and genome rearrangement in flowering plants. Curr Opin Genet Dev, 2005, 15: 621－627
3 Wessler S R. Phenotypic diversity mediated by the maize transposable elements *Ac* and *Spm*. Science, 1988, 242(4877): 399－405
4 Feschotte C, Jiang N, Wessler S R. Plant transposable elements: where genetics meets genomics. Nat Rev Genet, 2002, 3: 329－341
5 Peterson P A. A mutable pale green locus in maize. Genetics, 1953, 38: 682－683
6 McClintock B. Mutations in maize and chromosomal aberrations in Neurospora. Carnegie Inst Washington Year Book, 1954, 53: 254－260
7 Chopra S, Brendel V, Zhang J, et al. Molecular characterization of a mutable pigmentation phenotype and isolation of the first active transposable element from *Sorghum bicolor*. Proc Natl Acad Sci USA, 1999, 96(26): 15330－15335
8 Motohashi R, Ohtsubo E, Ohtsubo H. Identification of Tnr3, a suppressor-mutator/enhancer-like transposable element from rice. Mol Gen Genet, 1996, 250(2): 148－152
9 Zabala G, Vodkin L O. The *wp* mutation of *Glycine max* carries a gene-fragment-rich transposon of the CACTA superfamily. Plant Cell, 2005, 17: 2619－2632
10 Tian P F. Progress in plant CACTA elements. Acta Genetica Sinica, 2006, 33(9): 765－774
11 Gierl A, Lütticke S, Saedler H. TnpA product encoded by the transposable element En-1 of *Zea mays* is a DNA binding protein. EMBO J, 1988, 7(13): 4045－4053
12 Peterson P A. A relationship between the Spm and En control systems in maize. Am Nat, 1965, 99(908): 391－398
13 Pereira A, Schwarz-Sommer Z, Gierl A, et al. Genetic and molecular analysis of the Enhancer (En) transposable element system of *Zea mays*. EMBO J, 1985, 4(1): 17－23
14 Gierl A, Schwarz-Sommer Z, Saedler H. Molecular interactions between the components of the En-I transposable element system of *Zea mays*. EMBO J, 1985, 4(3): 579－583
15 Pereira A, Cuypers H, Gierl A, et al. Molecular analysis of the En/Spm transposable element system of *Zea mays*. EMBO J, 1986, 5(5): 835－841
16 Masson P, Surosky R, Kingsbury J A, et al. Genetic and molecular analysis of the *Spm*-dependent *a-m2* alleles of the maize *a* locus. Genetics, 1987, 177: 117－137
17 Talbert L E, Chandler V L. Characterization of a highly conserved sequence related to mutator transposable elements in maize. Mol Biol Evol, 1988, 5(5): 519－529
18 Jiang N, Bao Z, Zhang X, et al. Pack-MULE transposable elements mediate gene evolution in plants. Nature, 2004, 431: 569－573
19 Ohtsu K, Hirano H Y, Tsutsumi N, et al. *Anaconda*, a new class of transposon belonging to the *Mu* superfamily, has diversified by acquiring host genes during rice evolution. Mol Genet Genomics, 2005, 274: 606－615
20 Juretic N, Hoen D R, Huynh M L, et al. The evolutionary fate of MULE-mediated duplications of host gene fragments in rice. Genome Res, 2005, 15: 1292－1297
21 Hoen D R, Park K C, Elrouby N, et al. Transposon-mediated expansion and diversification of a family of *ULP*-like genes. Mol Biol Evol, 2006, 23(6): 1254－1268
22 Holligan D, Zhang X, Jiang N, et al. The transposable element landscape of the model legume *Lotus japonicus*. Genetics, 2006, 174: 2215－2228
23 Leeuwen H V, Monfort A, Puigdomenech P. *Mutator*-like elements identified in melon, Arabidopsis and rice contain ULP1 protease domains. Mol Genet Genomics, 2007, 277: 357－364
24 Gupta S, Gallavotti A, Stryker G A, et al. A novel class of *Helitron*-related transposable elements in maize contain portions of multiple pseudogenes. Plant Mol Biol, 2005, 57: 115－127
25 Lai J, Li Y, Messing J, et al. Gene movement by *Helitron* transposons contributes to the haplotype variability of maize. Proc Natl Acad Sci USA, 2005, 102: 9068－9073
26 Morgante M, Brunner S, Pea G, et al. Gene duplication and exon shuffling by helitron-like transposons generate intraspecies diversity in maize. Nat Genet, 2005, 37(9): 997－1002
27 Takahashi S, Inagaki Y, Satoh H, et al. Capture of a genomic *HMG* domain sequence by the *En/Spm*-related transposable element *Tpn1* in the Japanese morning glory. Mol Gen Genet, 1999, 261: 447－451
28 Kawasaki S, Nitasaka E. Characterization of *Tpn1* family in the Japanese morning glory: *En*/*Spm*-related transposable elements capturing host genes. Plant Cell Physiol, 2004, 45(7): 933－944
29 Roccaro M, Li Y, Sommer H, et al. ROSINA (RSI) is part of a CACTA transposable element, *TamRSI*, and links flower development to transposon activity. Mol Genet Genomics, 2007, 278: 243－254
30 Kohany O, Gentles A J, Hankus L, et al. Annotation, submission and screening of repetitive elements in Repbase: RepbaseSubmitter and Censor. BMC Bioinformatics, 2006, 25(7): 474
31 Edgar R C. MUSCLE: multiple sequence alignment with high accuracy and high throughput. Nucl Acid Res, 2004, 32(5): 1792－1797
32 Hall T A. BioEdit: a user-friendly biological sequence alignment editor and analysis program for Windows 95/98/NT. Nucl Acids Symp, 1999, 41: 95－98
33 Felsenstein J. PHYLIP (Phylongeny Inference Package), version 3.6. (Department of Genetics, University of Washington, Seattle, 2000)
34 Kumar S, Tamura K, Nei M. MEGA3: integrated software for molecular evolutionary genetics analysis and sequence alignment. Brief Bioinform, 2004, 5(2): 150－163
35 Cahoon E B, Hall S E, Ripp K G, et al. Metabolic redesign of vitamin E biosynthesis in plants for tocotrienol production and increased antioxidant content. Nat Biotechnol, 2003, 21: 1082－1087
36 Frey M, Reinecke J, Grant S, et al. Excision of the En/Spm transposable element of *Zea mays* requires two element-encoded proteins. EMBO J, 1990, 9(12): 4037－4044
37 Bercury S D, Panavas T, Irenze K, et al. Molecular analysis of the *Doppia* transposable element of maize. Plant Mol Biol, 2001, 47: 341－351
38 Yim Y S, Davis G L, Duru N A, et al. Characterization of three maize bacterial artificial chromosome libraries toward anchoring of the physical map to the genetic map using high-density bacterial artificial chromosome filter hybridization. Plant Physiol, 2002, 130: 1686－1696
39 Jameson N, Georgelis N, Fouladbash E, et al. *Helitron* mediated amplification of cytochrome P450 monooxygenase gene in maize. Plant Mol Biol, 2008, 67: 295－304