Theor Appl Genet (2004) 108:306–314
DOI 10.1007/s00122-003-1423-9

ORIGINAL PAPER

Y. Sha · S. Li · Z. Pei · L. Luo · Y. Tian · C. He

Generation and flanking sequence analysis of a rice T-DNA tagged population

Received: 27 January 2003 / Accepted: 4 July 2003 / Published online: 19 September 2003

Abstract Insertional mutagenesis provides a rapid way to clone a mutated gene. Transfer DNA (T-DNA) of *Agrobacterium tumefaciens* has been proven to be a successful tool for gene discovery in *Arabidopsis* and rice (*Oryza sativa* L. ssp. *japonica*). Here, we report the generation of 5,200 independent T-DNA tagged rice lines. The T-DNA insertion pattern in the rice genome was investigated, and an initial database was constructed based on T-DNA flanking sequences amplified from randomly selected T-DNA tagged rice lines using Thermal Asymmetric Interlaced PCR (TAIL-PCR). Of 361 T-DNA flanking sequences, 92 showed long T-DNA integration (T-DNA together with non-T-DNA). Another 55 sequences showed complex integration of T-DNA into the rice genome. Besides direct integration, filler sequences and microhomology (one to several nucleotides of homology) were observed between the T-DNA right border and other portions of the vector pCAMBIA1301 in transgenic rice. Preferential insertion of T-DNA into protein-coding regions of the rice genome was detected. Insertion sites mapped onto rice chromosomes were scattered in the genome. Some phenotypic mutants were observed in the T_1 generation of the T-DNA tagged plants. Our mutant population will be useful for studying T-DNA integration patterns and for analyzing gene function in rice.

Electronic Supplementary Material Supplementary material is available in the online version of this article at http://dx.doi.org/10.1007/s00122-003-1423-9.

Communicated by D. Mackill

Y. Sha · S. Li · Z. Pei · L. Luo · Y. Tian · C. He (✉)
Laboratory of Plant Biotechnology, Institute of Microbiology,
The Chinese Academy of Sciences,
Zhong Guan Cun, 100080 Beijing, P.R. China
e-mail: hecz@sun.im.ac.cn
Tel.: +86-10-62642536
Fax: +86-10-62548243

Introduction

Rice is one of the most important cereals, the principal food crop for half of the world's population. More than 90% of rice is consumed by humans (Goff 1999). Besides its economic importance, rice has several major advantages over other cereals for genetic and molecular studies. The estimated total genes of the rice genome are 32,000 to 50,000 (Goff et al. 2002) or 46,022 to 55,615 (Yu et al. 2002), yet it has the smallest genome among cereals. The gene density of rice, estimated at one gene in less than 15 kb, is higher than other cereals such as maize and wheat (Goff 1999). Furthermore, cereal genes were predicted to show conservation of gene order and orientation, or synteny (Goff 1999). In addition, a YAC-based map covering 63% of the rice genome and a physical map consisting of 65,287 finger-printed BAC clones are available (Saji et al. 2001; Chen et al. 2002). Cell culture and transformation methods have been developed for rice (Li et al. 1993; Hiei et al. 1994; Yin and Wang 2000). The draft sequence of the rice genome has been published (Goff et al. 2002; Yu et al. 2002). More than 120,000 expressed sequence tag (EST) collections from different rice tissues have been reported to the National Center of Biotechnology Information (NCBI) dbEST (http://www.ncbi.nlm.nih.gov/dbEST/dbEST_summary.html).

Despite such a large amount of rice sequencing data, DNA or protein sequences do not provide much information on biological function for the vast majority of genes. Analysis of the rice genome shows that about 80.6% of the *Arabidopsis* genes had a homolog in rice, whereas only 49.4% of the predicted rice genes had a homolog in *Arabidopsis* (Yu et al. 2002), indicating that genome asymmetry in the monocot and dicot was obvious. Moreover, a large number of genes are also regarded as hypothetical, due to lack of homologous genes and EST sequences in rice and other species. Therefore, defining functions of the numerous genes is extremely important and challenging to plant molecular biologists.

The direct approach to determine a predicted gene is to generate a mutation in the gene and analyze the phenotype (Ramachandran and Sundaresan 2001). Saturation mutagenesis covering the whole genome can be obtained by classic chemical or physical mutagenesis procedures. However, accessing the mutated genes is time-consuming and labor-intensive (Bouchez and Hofte 1998). Insertional mutagenesis, using either transposable elements or transfer DNA (T-DNA) of *Agrobacterium*, provides a rapid way to clone a mutated gene. This method has been proven successful in *Arabidopsis* (Azpiroz-Leehan and Feldmann 1997; Krysan et al. 1999; Parinov et al. 1999; Speulman et al. 1999; Tissier et al. 1999; Weigel et al. 2000) and rice (Chin et al. 1999; Enoki et al. 1999). Besides disrupting the expression of the gene into which it is inserted, the inserted fragment can be used as a marker in subsequent analysis of the mutants. T-DNA insertion has the advantage that additional steps to stabilize the inserts through multiple generations are not needed (Azpiro-Leehan and Feldmann 1997; Ramachandran and Sundaresan 2001). In addition, a low copy number, about 1.5 inserts per diploid genome, was reported in *Arabidopsis* as well as in rice (Feldmann 1991; Jeon et al. 2000).

In this study, we report the generation of 5,200 independent lines of T-DNA tagged rice plants by *Agrobcterium*-mediated transformation. Analysis of the flanking sequences isolated by Thermal Asymmetric Interlaced PCR (TAIL-PCR) revealed complex T-DNA insertion patterns and preferential integration of T-DNA into the protein coding-region of the rice genome. Our results suggest the T-DNA tagged rice population and the database of rice genome-sequences flanking T-DNA inserts will be useful for rice functional genomics.

Materials and methods

Plant transformation

Rice tissue transformation was performed according to Yin and Wang (2000) with slight modification. Vigorously embryogenic calli were induced on scutella from germinated rice seeds of *japonica* rice cultivar Nipponbare and co-cultivated with *Agrobacterium tumefaciens* EHA105 (Hood et al. 1993), carrying the binary vector pCAMBIA1301 (http://www.cambia.org.au/). After 3 days of co-cultivation, the rice tissues were cultured on NB_0 medium (Li et al. 1993) containing 250 mg l^{-1} of cefotaxime, 500 mg l^{-1} of ampicillin, 2 mg l^{-1} of 2, -D and 50 mg l^{-1} of hygromycin at 25°C in the dark for 3–4 weeks. Hygromycin-resistant calli were transferred to the same fresh medium for 2–3 weeks. The vigorously growing hygromycin-resistant calli were subcultured on NB_0 medium containing 1 mg l^{-1} of 6-BA (6-benzylaminopurine), 2 mg l^{-1} of NAA (naphthalaneacetic acid), 5 mg l^{-1} of ABA (abscisic acid) and 50 mg l^{-1} of hygromycin for 3 weeks. Compact, white embryogenic calli were transferred to the NB_0 medium containing 2 mg l^{-1} of 6-BA, 1 mg l^{-1} of NAA, 1 mg l^{-1} of KT (Kinetin), 1 mg l^{-1} of IAA (Indoleacetic acid) and 50 mg l^{-1} of hygromycin, and cultured at 25°C with 14-h light (about 50 μmol m^{-2} s^{-1}) and a 10-h dark period. The regenerated plantlets were subcultured to 1/2 MS medium (Murashige and Skoog 1962) containing 0.1 mg l^{-1} of IBA (indolebutyric acid) to form whole plants. Regenerated plants were transferred to soil and grown in a greenhouse.

Assay of GUS activity

Expression of the *GUS* gene in rice cells was assayed essentially as previously described (Rueb and Hensgens 1998) with 5-bromo-4-chlor-3-indoly glucuromide (X-Gluc) as a substrate. Segments of rice calli, leaf and stems were incubated in phosphate buffer (0.2 M Na_3PO_4, pH 7.0) containing 1 mg ml^{-1} X-Gluc, 0.1% (v/v) Triton X-100 and 10 mM EDTA at 37°C for 4 h. Leaf segments were washed 4–5 times with 75% ethanol before visual examination. Tissues that expressed the *GUS* gene were dark-blue.

PCR primers and reaction

Primers of GUSa (*ACGGCCTGTGGGCATTCAGT*), GUSb (*TGCACCATCAGCACGTTATCG*) and HPTa (*TGCGCCCAAGCTGCATCAT*), HPTb (*TGAACTCACCGCGACGTCTGT*), were designed to amplify a coding region fragment of the *GUS* (β-glucuronidase) gene and the *hpt* (hygromycin phosphotransferase) gene, respectively. The reaction mix (20 μl) for PCR consists of 5–10 ng of rice genomic DNA, 0.2 mM each of dNTPs, 0.2 μM of each PCR primer, 1× PCR buffer (10 mM Tris-HCl, pH 8.3, 50 mM KCl, 1.5 mM $MgCl_2$) and 0.8 units of *Taq* DNA polymerase (TAKARA, Dalian, P. R. China). Thermal cycling was done at 94°C for 1 min followed by 35 cycles of 94°C for 45 s, 60°C for 45 s and 72°C for 90 s. After the final cycle, the reactions were maintained at 72°C for 5 min before completion. Reactions were conducted using a PTC-100-programmable thermal controller (MJ Research, USA).

Isolation of genomic DNA and Southern hybridization

Genomic DNA was extracted from leaf tissues (1–2 g per plant) as described by Komari et al. (1989). Genomic DNA was digested with *Hin*dIII, separated on a 1% agarose gel in 1 × TAE by electrophoresis, blotted onto Hybond-N^+ membranes and hybridized with [^{32}P]dCTP-labeled probe. For the *GUS* gene detection, a coding region of pCAMBIA1301 amplified with primers GUSa and GUSb was used as a probe. For the *hpt* gene detection, a coding region of pCAMBIA1301 amplified with primers HPTa and HPTb was used as a probe. All blot-analysis procedures were carried out as described by Sambrook et al. (1989) and Kang et al. (1998).

TAIL-PCR isolation of the flanking sequences

Genomic sequences flanking the T-DNA insertions were amplified according to Liu et al. (1995), with minor modification of 15 super cycles (consisting of two high-stringency cycles and one reduced-stringency cycle) in the secondary reaction, and 15 super cycles in the tertiary reaction instead of 20 reduced-stringency cycles were performed. Ten to twenty ng of genomic DNA was used as template DNA. We designed three nested and target-specific primers: SP1 (*GGTGACCAGCTCGAATTTCCC*), SP2 (*TGAATCCTGTTGCCGGTCTTG*) and SP3 (*GCGCGCGGTGTCATCTATGT*). Each of these primers was used in combination with three arbitrary degenerate primers: AD4, TG(A/T)GNAG(A/T)ANCA(G/C)AGA; AD8, (G/C)TTGNTA(G/C)TNCTNTGC; and AD9, (A/T)CAGNTG(A/T)TNGTNCTG. TAIL-PCR products were purified by the Gel Extraction Kit (EZNA, USA), cloned into pGEM-T easy vector (Promega, USA) and sequenced.

Database search

With 200–700 bp of T-DNA flanking sequences, Gramene (http://www.gramene.org/gramene/searches/blast), NCBI (http://www.ncbi.nlm.nih.gov/BLAST/), Sygenta (Torry Mesa Research Institute; http://www.tmri.org/index.html) and TIGR (The Institute for Genomic Research; http://www.tigr.org/tdb/e2k1/osa1/

BACmapping/description.shtml) databases were searched for nucleotide similarities by the BLASTN algorithm. It was previously reported that the mean size of rice genomic DNAs between the start and stop codons, including introns, was 2.6 kb, and the mean gene size for rice is about 4.5 kb (Jeon et al. 2000; Yu et al. 2002). To find the disrupted genes located in the 3′ and 5′ sites of inserted T-DNA, 2-kb upstream and downstream of flanking sequences were searched in the same databases using BLASTX algorithm for the similarity of proteins and by BLASTN for the similarity of ESTs. The flanking sequences, which showed no significant similarity to released rice genome sequences through searching the same databases using BLASTN, were submitted to BLASTX for protein similarity and BLASTN for the EST similarity search directly.

In silico mapping of T-DNA insertion sites

To determine the positions of the T-DNA insertion sites, the rice genomic BAC or PAC clones containing the flanking sequences were *in silico* mapped on rice chromosomes based on the map databases of NCGR (National Center for Gene Research, Chinese Academy of Sciences, China; http://www.ncgr.ac.cn/rice/chr4/physicalmap/index.htm), RGP (Rice Genome Program, Japan; http://rgp.dna.affrc.go.jp/cgi-bin/statusdb/status.pl), US genome-sequencing group (http://www.usricegenome.org/) and TIGR (The Institute for Genomic Research; http://www.tigr.org/tdb/e2k1/osa1/BACmapping/description.shtml). The chromosomes were modified from a YAC-based transcript map (http://rgp.dna.affrc.go.jp/publicdata/estmap2001/).

Results

Efficient transformation of rice mediated by *A. tumefaciens*

Scutella-derived embryonic calli were co-cultivated with *A. tumefaciens* EHA105 carrying the binary vector pCAMBIA1301. Most of the calli that recovered from the first round of selection proliferated on the second selective medium. Approximately 20–30% of the co-cultured calli produced hygromycin-resistant cells. Fifty to sixty percent of hygromycin-resistant calli could be regenerated into whole rice plants. Therefore, an *Agrobacterium*-mediated, stable and efficient rice transformation system was established. With this system, 5,200 independently transformed plants were obtained. Some phenotypic aberrations were observed in the tagged plants and their progeny, such as dwarf, white-striped leaves, lesion-mimic like spotted leaves, long sterile glumes and partial fertility. The study to determine whether these phenotypic aberrations were due to the T-DNA integration is in progress.

Confirmation of the tagged transgenic T_0 plants

GUS histochemical staining, PCR amplification and Southern-blot analysis were carried out to investigate T-DNA integration in randomly selected T_0 plants. Most of the selected plants (22/24, i.e. 92%) were positive in GUS histochemical staining. All of the selected plants showed specific amplified bands of the *hpt* gene using primer pairs HPTa and HPTb, whereas 12 out of 190 randomly selected plants of those regenerated did not show amplified bands of the *GUS* gene using primer pairs of GUSa and GUSb, indicating that deletion of the *GUS* gene has probably occurred during the process of transformation.

The result was also confirmed by Southern analysis of T_0 tagged plants using the *GUS* or *hpt* gene-fragment as a probe. Since the T-DNA of pCAMBIA1301 has a single *Hin*dIII site, and there is no *Hin*dIII site on the probes, genomic DNA of T_0 transgenic plants was digested with *Hin*dIII in Southern analysis. The number of hybridizing bands reflected the copy number of integrated T-DNA in the plants. Among the 39 transgenic T_0 plants examined, most carried only one copy of *GUS* (23, i.e. 59%) or the *hpt* gene (19, i.e. 49%), although multiple copies of *GUS* and *hpt* were observed in some plants (Fig. 1). The average copy number of the *GUS* gene was 1.38, and the average copy number of the *hpt* gene was 1.76. The different copy number of *GUS* and *hpt* in some plants may have been produced by rearrangement of T-DNA during

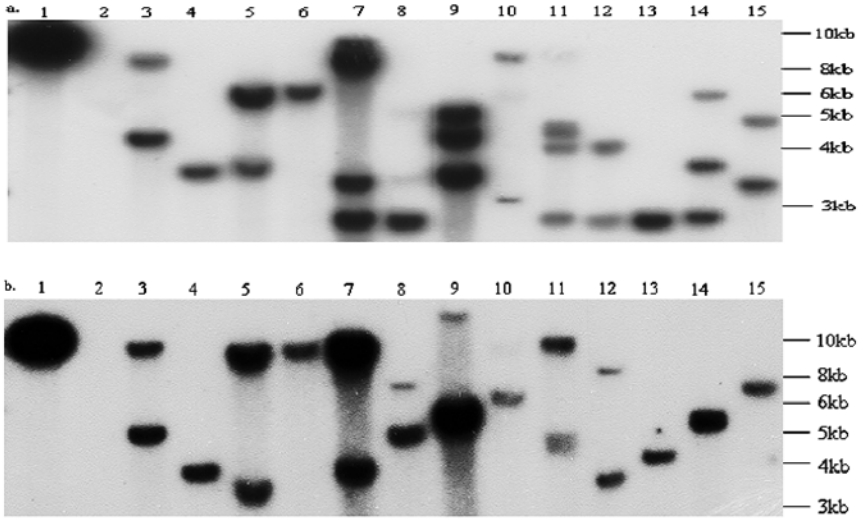


Fig. 1a, b Southern-blot analysis of T-DNA tagged T_0 rice plants. Ti-plasmid pCAMBIA1301, non-transgenic rice genomic DNA and genomic DNAs from different T_0 transgenic rice plants digested with *Hin*dIII were hybridized with *hpt* probe (**a**) or *GUS* probe (**b**). *Lane 1*: Ti plasmid 1301 (positive control); *lane 2*: non-transgenic rice genomic DNA (negative control); *lanes 3–15*: genomic DNA from different transgenic T0 plants

Fig. 2 The nick position of the T-DNA right border in the rice genome. The *capitalized letters* indicate the right border sequences. The *letters in lowercase* indicate T-DNA sequences immediately next to the right border. The *right column* indicates the percent of the T-DNA nick position in 214 rice genome sequences flanking the T-DNA right border. The *asterisk-marked* sequences indicate the T-DNA right border nicked precisely at the same sites found in dicot species

the transformation process (Ohba et al. 1995), or by breakage at some stage during the transfer or integration process (Herman et al. 1990).

Molecular characterization of the T-DNA integration pattern

According to Tinland (1996), a majority of the T-DNA insertions occur within the right-border at specific loci in dicot species. To examine whether it is the same case in our tagged plants, the junction regions between rice genomic DNA and the T-DNA right-border were amplified by TAIL-PCR as previously described (Liu et al. 1995). The success rate for amplification of specific products (>200 base pair) varied according to the arbitrary degenerate (AD) primer used, from about 50% with AD4 to 70% with AD8 and 85% with AD9. With these three AD primers, at least one fragment was amplified in more than 90% of the transformants. Among the examined 361 amplified-products, 214 contained plant sequences flanking T-DNA inserts and 147 showed long T-DNA integration and complex T-DNA integration into the genome. The nick positions at the right border of our T-DNA inserted plants were more complex than those observed in *Arabidopsis* and tobacco transgenic plants. In dicot species, most T-DNAs were nicked after the first or second base of the right border (Yadav et al. 1982; Zambryski et al. 1982). Out of the 214 TAIL-PCR products that contained plant sequences, 59 (27.6%) were nicked precisely at the same site found in dicot species. The other 155 (72.4%) were nicked between 25 bp before the right-border repeat to 4 bp after the right-border repeat (Fig. 2).

Analysis of the conjunction sequences of the T-DNA right-border and rice genome DNA also revealed long T-DNA transfer and complex T-DNA insertion patterns in the rice genome (Fig. 3). Ninety two of 361 sequences showed a long T-DNA transfer. Direct or inverted junction of the T-DNA borders, as well as linkage of

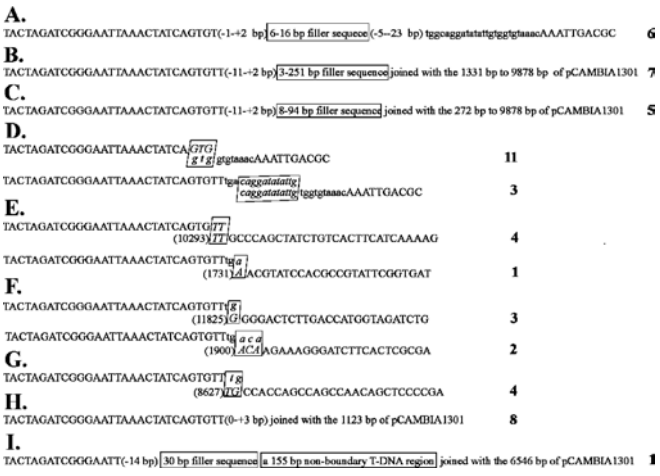
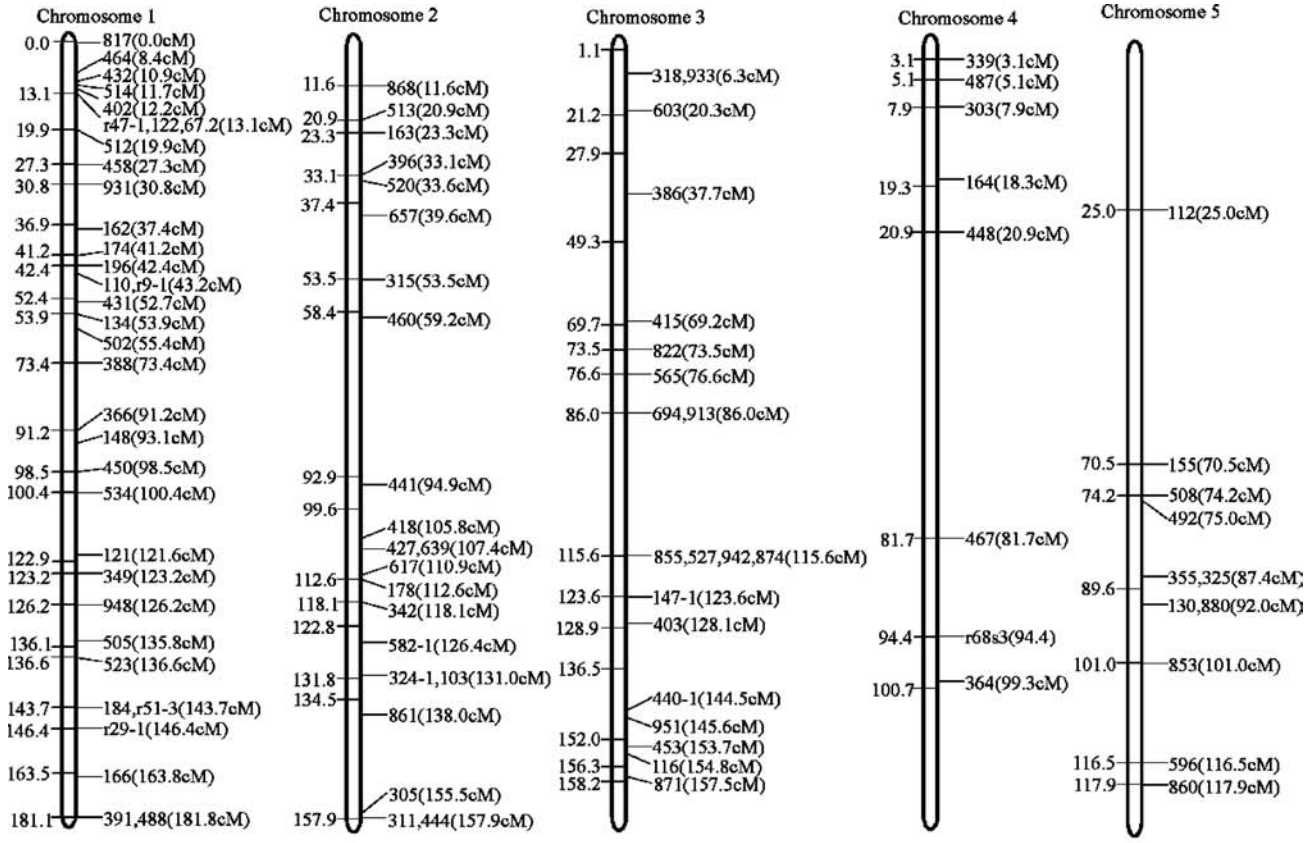


Fig. 3A–I Scheme of complex integration patterns in T-DNA tagged rice plants. **A** Filler sequences at junctions of the right and left border. **B** Filler sequences at junctions of the right border and non-boundary region of T-DNA contiguous to the right border. **C** Filler sequences at junctions of the right border and the non-boundary region of T-DNA contiguous to the left border. **D** Integration of the T-DNA right and left border showing nucleotides of homology. **E** Nucleotides of homology at conjunctions of the right border and non-boundary region of T-DNA contiguous to the right border. **F** Nucleotides of homology at conjunctions of the right border and non-boundary region of T-DNA contiguous to the left border. **G** Nucleotides of homology at conjunctions of the right border and vector backbone of pCAMBIA1301. **H** Direct integration of the right border and non-boundary region of T-DNA. **I** A complex pattern of integration. The *lower case letters* indicate the right border repeat and the left border repeat. The number of clones of each integration pattern is shown *in bold* at the right of each sequence. The *bracket-enclosed numbers* in **A**, **B**, **C**, **H** and **I** indicate nucleotides deleted or elongated, symbolized with – or +, respectively, from the nick position of the T-DNA borders in comparison with dicot species. The *bracket-enclosed numbers* in **E**, **F** and **G** indicate the position at which the other part of pCAMBIA1301 joined with the T-DNA right border. When the junctions are at positions where the right border and the other part of pCAMBIA1301 show nucleotides of homology, *two vertical lines* delineate the region where the right border and other part of pCAMBIA1301 must have joined. The nucleotides of homology are shown in *italics*

the right border to non-boundary regions of T-DNA (T-DNA between right and left borders) and/or non-T-DNA regions of pCAMBIA1301, were found in another 55 sequences. Filler sequences of 3 to 251 bp were found at junctions of the T-DNA borders and between junctions of the right border and the non-boundary region of T-DNA (Fig. 3A, B, C). The junctions between the T-DNA borders, the T-DNA right border and the non-boundary region of T-DNA, and the right border and vector backbone showed 1 to 12 bp of microhomology (Fig. 3D, E, F, G). Direct integration of the right border and the non-boundary region of T-DNA was also observed (Fig. 3H). In these sequences, there are no filler sequences and no microhomologous nucleotide at the conjunctions. The fusion seemed to result from a simple ligation of blunt double-strand ends. A complex pattern of integration that involved a 30-bp filler sequence and a


Chromosome 1
0.0 817(0.0cM)
464(8.4cM)
432(10.9cM)
13.1 514(11.7cM)
402(12.2cM)
19.9 r47-1,122,67.2(13.1cM)
512(19.9cM)
27.3 458(27.3cM)
30.8 931(30.8cM)
36.9 162(37.4cM)
41.2 174(41.2cM)
42.4 196(42.4cM)
110,r9-1(43.2cM)
52.4 431(52.7cM)
53.9 134(53.9cM)
502(55.4cM)
73.4 388(73.4cM)
91.2 366(91.2cM)
148(93.1cM)
98.5 450(98.5cM)
100.4 534(100.4cM)
122.9 121(121.6cM)
123.2 349(123.2cM)
126.2 948(126.2cM)
136.1 505(135.8cM)
136.6 523(136.6cM)
143.7 184,r51-3(143.7cM)
146.4 r29-1(146.4cM)
163.5 166(163.8cM)
181.1 391,488(181.8cM)
Chromosome 2
11.6 868(11.6cM)
20.9 513(20.9cM)
23.3 163(23.3cM)
396(33.1cM)
33.1 520(33.6cM)
37.4 657(39.6cM)
53.5 315(53.5cM)
58.4 460(59.2cM)
92.9 441(94.9cM)
99.6 418(105.8cM)
427,639(107.4cM)
112.6 617(110.9cM)
178(112.6cM)
118.1 342(118.1cM)
122.8 582-1(126.4cM)
131.8 324-1,103(131.0cM)
134.5 861(138.0cM)
305(155.5cM)
157.9 311,444(157.9cM)
Chromosome 3
1.1 318,933(6.3cM)
21.2 603(20.3cM)
27.9 386(37.7cM)
49.3
69.7 415(69.2cM)
73.5 822(73.5cM)
76.6 565(76.6cM)
86.0 694,913(86.0cM)
115.6 855,527,942,874(115.6cM)
123.6 147-1(123.6cM)
128.9 403(128.1cM)
136.5
440-1(144.5cM)
951(145.6cM)
152.0 453(153.7cM)
156.3 116(154.8cM)
158.2 871(157.5cM)
Chromosome 4
3.1 339(3.1cM)
5.1 487(5.1cM)
7.9 303(7.9cM)
19.3 164(18.3cM)
20.9 448(20.9cM)
81.7 467(81.7cM)
94.4 r68s3(94.4)
100.7 364(99.3cM)
Chromosome 5
25.0 112(25.0cM)
70.5 155(70.5cM)
74.2 508(74.2cM)
492(75.0cM)
355,325(87.4cM)
89.6 130,880(92.0cM)
101.0 853(101.0cM)
116.5 596(116.5cM)
117.9 860(117.9cM)
Chromosome 6
0.0
468(5.1cM)
10.4 389(6.9cM)
32.7 482(32.7cM)
61.6 449(61.6-63.2cM)
64.6
65.8 464,307(65.8cM)
70.9 409(72.9cM)
485(73.2cM)
90.5 r1-2(95.4cM)
100.3
109.5 352,354(117.0cM)
119(122.8cM)
122.8 664(123.6cM)
123.9 956(124.4cM)
124.4 953(126.0cM)
Chromosome 7
1.9
3.9 438(2.5cM)
9.6 70s1(9.6cM)
25.4 394(25.4cM)
67.0 826(67.0cM)
69.2 847(70.0cM)
71.6
73.2 r40-2(73.2cM)
76.7 416(78.9cM)
436-1(81.4cM)
89.0 864(91.7cM)
97.4 173(97.4cM)
907(99.3cM)
105.7 105.7(859,636,569(105.7cM)
Chromosome 8
3.6 353(3.6cM)
24.9
31.6 912(29.4cM)
35.7
37.1 839(36.8cM)
54.3 445(54.3cM)
55.4 r18-1(55.4cM)
61.2 101(61.2cM)
89.9 192(90.5cM)
96.6
102.1 320(99.1cM)
495,494(108.2cM)
112.6 408-1,556(111.2cM)
119.6 507(119.3cM)
Chromosome 10
1.1 459(1.2cM)
1.9 374(1.9cM)
3.0 r36-2(3.9cM)
6.8 916(6.8cM)
24.2
32.1 r21-2(30.2cM)
44.3 r10-1(44.0cM)
314(44.8cM)
53.3 602(53.3cM)
53.6 r25-1(53.6cM)
55.3 938(54.7cM)
68.6 305(68.6cM)
72.5 417,r52-2(72.5cM)
151(72.8cM)
426(83.0cM)
83.8 471(83.8cM)
Chromosome 12
5.5 310(6.3cM)
9.4 439-1(9.7cM)
13.3 503(13.3cM)
39.7 407(39.7cM)
66.6 326(67.2cM)
71.7 437(71.7cM)
109.5 940(121.9cM)

155-bp non-boundary region of T-DNA between the right border and the vector backbone is shown in Fig. 3 (I).

Database of T-DNA insertion plants

To assess the distribution of inserts in our population of 5,200 rice tagged-plants, 500 flanking sequences were amplified by TAIL-PCR. After removal of overlapping sequences and vector backbone sequences (i.e. sequences of long T-DNA integration and complex T-DNA integration), a primary database of 214 flanking sequences was generated. Of these, 169 flanking sequences showed greater than 95% homology to sequenced rice genomic BAC or PAC clones, and 45 showed no significant similarity to any released rice genomic BAC or PAC clones. The results of protein similarity and EST similarity indicated that 152 (71%) flanking sequences showed more than 50% identity to predicted protein-sequences, and 70 (32%) flanking-sequences showed more than 90% identity to ESTs (Table 1, Supplemental data). Low identity of some of the flanking sequences with rice ESTs may result from the presence of introns. The T-DNA flanking-sequences database will be useful for mutational analysis since some important genes such as ribosomal protein, the recognition molecule CASPR3, GMP synthase, ammonium transporter and the disease-resistance protein RPM1 were predicted. Phenotypic analysis of these insertion mutants is in progress and the unique function of these genes in rice will be revealed through reverse genetic and biochemical approaches.

Distribution of T-DNA tagged sites

Out of 169 rice genomic BAC or PAC clones containing flanking sequences, 35 (20.7%) were located on chromosome 1, 23 (13.6%) on chromosome 2, 20 (11.8%) on chromosome 3, nine (5.3%) on chromosome 4, 11 (6.5%) on chromosome 5, 15 (8.8%) on chromosome 6, 14 (8.3%) on chromosome 7, 14 (8.3%) on chromosome 8, two (0.1%) on chromosome 9, 16 (9.5%) on chromosome 10, three (0.2%) on chromosome 11 and seven (4.1%) on chromosome 12 (Fig. 4, Table 1 in the supplemental data). Only few flanking sequences were located on chromosome 9 and chromosome 11. This may due to the limited availability of the chromosome 9 and chromosome 11 sequences released. In seven flanking sequences, the corresponding BAC or PAC clones have been sequenced, but precise chromosomal locations have not been determined (Table 1 in the Supplemental data).

Fig. 4 Chromosomal positions of T-DNA insertions in the rice genome. The chromosomal positions of the rice genomic clones containing T-DNA insertions were determined by searching map databases of RGP, US genome-sequencing group and NCGR. The clone number of T-DNA flanking sequences and the loci of corresponding T-DNA insertion positions are shown on the right side of each chromosome. Chromosomes 9 and 11 are not shown because of few T-DNA insertion sites on these two chromosomes

Discussion

Sequence analysis of the junctions between the T-DNA right-border and rice genome DNA revealed that the T-DNA nick position in the tagged rice plants are more complex than in dicot species. The T-DNA right-boundaries in most of the transgenic rice plants did not correspond to the T-DNA nicking positions found in *A. thaliana,* and tobacco as reported previously (Hiei et al. 1994; Jeon et al. 2000). In our study, the nick position of T-DNA within the right border was complex (Fig. 2). Besides, nicking precisely at the same site found in dicot species, the majority of the sequences at the junction were nicked between 25 bp before the right border repeat to 4 bp after the right border repeat. The result further supports the viewpoint that the nick position of T-DNA in transgenic rice plants is different from that in transgenic dicots (Hiei et al. 1994; Jeon et al. 2000), although the molecular mechanism of the transfer of T-DNA from *Agrobacterium* to dicots and monocots is apparently identical.

Analysis of conjunction sequences also revealed long T-DNA integration and complex T-DNA integration (Fig. 3). Ninety two of the 361 sequences (25.5%) showed long T-DNA transfer. Long T-DNA transfer has been described in previous reports (Hiei et al. 1994; Kononov et al. 1997; Wenck et al. 1997; Jeon et al. 2000), with frequency of 33.2% in rice, 75% in *Nicotiana tabacum* and 62% in *A. thaliana*. Using primer pairs designed from different parts of the Ti-plasmid, Yin and Wang (2000) proved the existence of long T-DNA transfer in rice. Our results provide evidence based on sequences isolated from the T-DNA right-border of transformed rice. According to Wenck et al. (1997), the transfer of long T-DNA may be due to insufficient Vir proteins in the binary system. Two VirD2 proteins are needed per one copy of T-DNA in the normal T-DNA transfer system. One Vir protein associates with the 5′ region of T-DNA and transfers it from *Agrobacterium* to the plant cell. Another VirD2 protein associates with the 5′ end of non-T-DNA and rejoins the 3′ end of the replacement (Lanka and Wilkins 1995). If there are fewer VirD2 proteins in a binary system, T-DNA processing will skip the other border of T-DNA with a long T-DNA of Ti-plasmid transferred into the plant cell (Yin and Wang 2000). Interestingly, the existence of the greater-than-unit-length binary plasmid in rice transformants was also detected in the study of Yin and Wang (2000). In our study, this phenomenon is not found due to the limitation of the size of TAIL-PCR products (about 250 to 750 bp). The presence of long T-DNA transfer would make it difficult to isolate plant genomic sequences flanking T-DNA inserts using PCR approaches. However, the plasmid rescue approach would be applied to isolate the flanking sequence of the long T-DNA insertion (Yin and Wang 2000). Moreover, the presence of the antibiotic gene from the non-T-DNA region in transgenic plants might cause environmental problems in transgenic applications in the field.

Fifty five sequences showed complex T-DNA integration. Illegitimate recombination is observed at junctions between two T-DNA borders and junctions between T-DNA and the plant genome in *A. thaliana* (Gheysen et al. 1991; De Buck et al. 1999). In our study, this phenomenon was observed at junctions between two T-DNA portions in transgenic monocots. Eighteen of the 55 sequences contained a few to 251-bp additional sequences (filler sequence) at the junctions (Fig. 3A, B, C). The filler sequences consisted of additional nucleotides different from the T-DNA sequences adjacent to the junctions. The filler sequences were observed in T-DNA::plant junctions in transgenic *N. tabacum* (Gheysen et al. 1991). According to Gorbunova and Levy (1997) and Salomon and Puchta (1998), the double-strand break repair (DSB) mechanism plays an important role in T-DNA integration into the plant genome. Interaction of the proteins, such as exonucleases, RecA-like proteins, topoisomerases and ligases, results not only in degradation of DNA ends but also the synthesis of filler DNA with multiple template switches. Another 28 sequences showed one to several nucleotides of microhomology between both ends at cross-over points of the right border and other portions of pCAMBIA1301 (Fig. 3D, E, F, G). Roth and Wilson (1986) showed that 1–6 bp of microhomology between single-stranded ends is sufficient for ligation directed by pairing of these homologous nucleotides in the circularization of SV40 molecules in monkey cells. Microhomology was also reported in a previous study of junctions between two T-DNAs and junctions between T-DNA and plant genomic DNA in dicots (Gheysen et al. 1991; De Buck et al. 1999). Interestingly, one sequence showed the most-complex pattern of integration (Fig. 3 I). In this sequence, the T-DNA right border is joined to a 155-bp T-DNA region with a 30-bp filler sequence at the junction. Then the 155-bp T-DNA region was integrated with a vector backbone region of pCAMBIA1301.

Although the exact mechanism of the complex patterns of T-DNA integration is largely unknown, it has been proposed that T-DNA integration involves illegitimate recombination in transgenic dicots (Gheysen et al. 1991; Mayerhofer et al. 1991; Tinland 1996; De Neve et al. 1997). The presence of microhomology at or nearby the junctions is one of the characteristics of illegitimate recombination. Other characteristics of illegitimate recombination include the presence of the filler sequence at the recombinant junctions and the generation of either a microdeletion or the presence of larger target rearrangements (Gheysen et al. 1991). In our study, the existence of microhomology, filler sequences and microdeletion were found at the junctions of T-DNA right borders and other parts of pCAMBIA1301. Thus, we can conclude that illegitimate recombination plays an important role in the formation of complex patterns of T-DNA integration in transgenic rice; and the mechanism of the formation of complex patterns of T-DNA integration in rice is similar to that of dicot species and mammalian cells. It would be interesting to examine whether this pattern is common in other monocot species.

In our initial database of T-DNA flanking sequences, most flanking sequences exhibited different levels of similarity with proteins or predicted genes in public databases. Seventy sequences showed 90% similarity to rice ESTs in public databases of NCBI, TIGR and Gramene. Significant preferential insertion of AC into the protein-coding region was reported previously (Enoki et al. 1999; Greco et al. 2001). Our results suggest preferential insertion of the T-DNA into transcriptionally active regions, which has previously been reported in *Arabidopsis* (Koncz et al. 1989; Kertbundit et al. 1991; Topping et al. 1991) and rice (Barakat et al. 2000). This result may reflect that T-DNA inserts into selective chromatins, which are potentially or actually transcriptionally active, in a relatively "open" status. Only in these types is high-frequency integration of T-DNA possible (Topping et al. 1991). The "open" chromatin structure might facilitate the integration of foreign sequences. A similar result has been suggested for retrovirus insertion in mice (Mooslehner et al. 1990) and the *P*-element in *Drosophila melanaogaster* (Downes 1990). Jeon et al. (2000) predicted that rice genome saturation could be archived with 660,000 insertions. Saturation of the transcriptionally active regions by T-DNA insertion in rice may only require 165,000 insertions due to T-DNA insertion specificity, and the fact that the gene space of rice corresponds to 25% of the genome (Barakat et al. 2000).

Acknowledgements We are grateful to Dr. Guo-liang Wang for critical reading of the manuscript. The work was supported by The Chinese Academy of Sciences (grant No. KXCX2-1-02-01), and The Basic Research Program (The '973' Program, grant No. TG2000016203) and High-Tech Program (The '863' Program, grant No. 2001AA225012) of the Ministry of Science and Technology of China.

References

Azpiroz-Leehan R, Feldmann KA (1997) T-DNA insertion mutagenesis in *Arabidopsis*: going back and forth. Trends Genet 13:152–156

Barakat A, Gallois P, Raynal M, Meatre-Ortega D, Sallaud C, Guiderdoni E, Delseny M, Bernardi G (2000) The distribution of T-DNA in the genomes of transgenic Arabdopsis and rice. FEBS Lett 471:161–164

Bouchez D, H. Höfte (1998) Functional genomics in plants. Plant Physiol 118:725–732

Chen M-S, Presting G, Barbazuk WB, Goicoechea JL, Blackmon B, Fang G-C, Kim H, Frisch D, Yu Y, Sun S-H, Higingbottom S, Phimphilai J, Phimphilai D, Thurmond S, Gaudette B, Li P, Liu J-D, Hatfield J, Main D, Farrar K, Henderson C, Barnett L, Costa R, Williams B, Walser S, Atkins M, Hall C, Budiman MA, Tomkins JP, Luo M-Z, Bancroft I, Salse J, Regad F, Mohapatra T, Singh NK, Tyagi AK, Soderlund C, Dean RA,Wing RA (2002) An integrated physical and genetic map of the rice genome. Plant Cell 14:537–545

Chin HG, Choe MS, Lee SH, Park SH, Koo JC, Kim NY, Lee JJ, Oh BG, Yi GH, Kim SC, Choi HC, Cho MJ, Han CD (1999) Molecular analysis of rice plants harboring an Ac/Ds transpos-

able element-mediated gene trapping system. Plant J 19:615–623

De Buck S, Jacobs A, Van Montagu M, Depicker A (1999) The DNA sequences of T-DNA junctions suggest that complex T-DNA loci are formed by a recombination process resembling T-DNA integration. Plant J 20:295–304

De Neve M, De Buck S, Jacobs A, Van Montagn M, Depicker A (1997) T-DNA integration patterns in co-transformed plant cells suggest that T-DNA repeats originate from ligation of separate T-DNAs. Plant J 11:15–29

Downes M (1990) Preferential insertion of *P* elements into genes expressed in the germ-line of *Drosophila melanogaster*. Mol Gen Genet 222:457–460

Enoki H, Izawa T, Kawahara M, Komatsu M, Koh S, Kyozuka J, Shimamoto K (1999) Ac as a tool for functional genomics of rice. Plant J 19:605–613

Feldmann KA (1991) T-DNA insertion mutagenesis in *Arabidopsis*: mutational spectrum. Plant J 1:71–82

Gheysen G, Villarroel R, Van Montagu M (1991) Illegitimate recombination in plants: a model for T-DNA integration. Genes Dev 5:287–297

Goff SA (1999) Rice as a model for cereal genomics. Curr Opin Plant Biol 2:86–89

Goff SA, Ricke D, Lan TH, Presting G, Wang R, Dunn M, Glazebrook J, Sessions A, Oeller P, Varma H, Hadley D, Hutchison D, Martin C, Katagiri F, Lange BM, Moughamer T, Xia Y, Budworth P, Zhong J, Miguel T, Paszkowski U, Zhang S, Colbert M, Sun WL, Chen L, Cooper B, Park S, Wood TC, Mao L, Quail P, Wing R, Dean R, Yu Y, Zharkikh A, Shen R, Sahasrabudhe S, Thomas A, Cannings R, Gutin A, Pruss D, Reid J, Tavtigian S, Mitchell J, Eldredge G, Scholl T, Miller RM, Bhatnagar S, Adey N, Rubano T, Tusneem N, Robinson R, Feldhaus J, Macalma T, Oliphant A, Briggs S (2002) A draft sequence of the rice genome (*Oryza sativa* L. ssp. *japonica*). Science 296:92–100

Gorbunova V, Levy AA (1997) Non-homologous DNA end-joining in plant cells is associated with deletions and filler DNA insertions. Nucleic Acids Res 25:4650–4657

Greco R, Ouwerkerk PBF, Taal AJC, Favalli C, Beguiristain T, Puigdomenech P, Colombo L, Hoge JHC, Pereira A (2001) Early and multiple Ac transpositions in rice suitable for efficient insertional mutagenesis. Plant Mol Biol 46:215–227

Herman L, Jacobs A, van Montagu M, Depicker A (1990) Plant chromosome/marker gene fusion assay for study of normal and truncated T-DNA integration events. Mol Gen Genet 224:248–256

Hiei Y, Ohta S, Komari T, Kumshiro T (1994) Efficient transformation of rice (*Oryza sativa* L.) mediated by *Agrobacterium* and sequence analysis of the boundaries of the T-DNA. Plant J 6:271–282

Hood EE, Gelvin SB, Melchers LS, Hoekema A (1993) New *Agrobacterium* helper plasmids for gene transfer to plants. Transgenic Res 2:208–218

Jeon JS, Lee S, Jung KH, Jun SH, Jeong DH, Lee J, Kim C, Jang S, Yang K, Nam J, An K, Han MJ, Sung RJ, Choi HS, Yu JH, Choi JH, Cho SY, Cha SS, Kim SI, An G (2000) T-DNA insertional mutagenesis for functional genomics in rice. Plant J 22:561–570

Kang HG, Jeon JS, Lee S, An G (1998) Identification of class B and class C floral-organ identity genes from rice plants. Plant Mol Biol 38:1021–1029

Kertbundit S, De Greve H, Deboeck F, Van Montagu M, Hernalsteens J (1991) In vivo random beta-glucuronidase gene fusions in *Arabidopsis thaliana*. Proc Natl Acad Sci USA 88:5212-5216

Komari T, Saito Y, Nakakido F, Kumashiro T (1989) Efficient selection of somatic hybrids in *Nicotiana tabacum* L. using a combination of drug-resistance markers introduced by transformation. Theor Appl Genet 77:547–552

Koncz C, Martini N, Mayerhofer R, Koncz-Kálmán Z, Körber H, Rédei GP Schell J (1989) High-frequency T-DNA-mediated gene tagging in plants. Proc Natl Acad Sci USA 86:8467–8471

Kononov ME, Bassuner B, Gelvin SB (1997) Integration of T-DNA binary vector 'backbone' sequences into the tobacco genome: evidence for multiple complex patterns of integration. Plant J 11:945–957

Krysan P J, Young JC, Sussman MR (1999) T-DNA as an insertional mutagen in *Arabidopsis*. Plant Cell 11:2283–2290

Lanka E, Wilkins BM (1995) DNA processing reactions in bacterial conjunction. Annu Rev Biochem 64:141–169

Li L, Qu R, de Kochko A, Fauquet C, Beachy RN (1993) An improved rice transformation system using the biolistic method. Plant Cell Rep 12:250–255

Liu YG, Mitsukawa N, Oosumi T, Whittier RF (1995) Efficient isolation and mapping of *Arabidopsis thaliana* T-DNA insert junctions by thermal asymmetric interlaced PCR. Plant J 8:457–463

Mayerhofer R, Koncz-Kalman Z, Nawrath C, Bakkeren G, Crameri A, Angelis K, Redei GP, Schell J, Hohn B, Koncz C (1991) T-DNA integration: a mode of illegitimate recombination in plants. EMBO J 10:697–704

Mooslehner K, Karls U, Harbers K (1990) Retroviral integration sites in transgenic Mov mice frequently map in the vicinity of transcribed DNA regions. J Virol 64:3056–3058

Murashige T, Skoog F (1962) A revised medium for rapid growth and bioassays with tobacco tissue cultures. Physiol Plant 15:473-497

Ohba T, Yoshioka Y, Machida C, Machida Y (1995) DNA rearrangement associated with the integration of T-DNA in tobacco: an example for multiple duplications of DNA around the integration target. Plant J 7:157–164

Parinov S, Sevugan M, Ye D, Yang WC, Kumaran M, Sundaresan V (1999) Analysis of flanking sequences from dissociation insertion lines: a database for reverse genetics in *Arabidopsis*. Plant Cell 11:2263–2270

Ramachandran S, Sundaresan V (2001) Transposons as tools for functional genomics. Plant Physiol Biochem 39:243–252

Roth DB, Wilson JH (1986) Non-homologous recombination in mammalian cells: role for short sequence homologies in the joining reaction. Mol Cell Biol 6:4295–4304

Rueb S, Hensgens LAM (1998) Improved histochemical staining for β-D-glucuronidase activity in monocotyledonous plants. Rice Genet Newslett 6:168–169

Saji S, Umehara Y, Antonio BA, Yamane H, Tanoue H, Baba T, Aoki H, Ishige N, Wu JZ, Koike K, Matsumoto T, Sasaki T (2001) A physical map with yeast artificial chromosome (YAC) clones covering 63% of the 12 rice chromosomes. Genome 44:32–37

Salomon S, Puchta H (1998) Capture of genomic and T-DNA sequences during double-strand break repair in somatic plant cells. EMBO J 17:6086–6095

Sambrook J, Fritsch EF, Maniatis T (1989) Molecular cloning: a laboratory manual, 2nd edn. Cold Spring Harbor Laboratory, Cold Spring Harbor, New York

Speulman E, Metz PJ, van Arkel G, Hekkert PL, Stiekema WJ, Pereira A (1999) A two-component enhancer-inhibitor transposon mutagenesis system for functional analysis of the *Arabidopsis* genome. Plant Cell 11:1853–1866

Tinland B (1996) The integration of T-DNA into plant genomes. Trends Plant Sci 1:178–184

Tissier A, Marillonnet S, Klimyuk V, Patel K, Torres MA, Murphy GP, Jones JDG (1999) Multiple independent defective suppressor-mutator transposon insertions in *Arabidopsis*: a tool for functional genomics. Plant Cell 11:1841–1852

Topping JF, Wei WB, Lindsey K (1991) Functional tagging of regulatory elements in the plant genome. Development 112:1009–1019

Weigel D, Ahn JH, Blázquez MA, Borevitz J, Christensen SK, Fankhauser C, Ferrándiz C, Kardailsky I, Malancharuvil EJ, Neff MM, Nguyen JT, Sato S, Wang Z, Xia Y, Dixon RA, Harrison MJ, Lamb CJ, Yanofsky MF, Chory J (2000) Activation tagging in *Arabidopsis*. Plant Physiol 122:1003–1014

Wenck AM, Czako M, Marton L (1997) Frequent collinear long transfer of DNA inclusive of the whole binary vector during *Agrobacterium*-mediated transformation. Plant Mol Biol 34:913–922

Yadav NS, Vanderleyden J, Bennett DR, Barnes WM, Chilton MD (1982) Short direct repeats flank the T-DNA on a nopaline Ti plasmid. Proc Natl Acad Sci USA 79:6322–6326

Yin Z, Wang G-L (2000) Evidence of multiple complex patterns of T-DNA integration into the rice genome. Theor Appl Genet 100:461–470

Yu J, Hu S, Wang J, Wong GK, Li S, Liu B, Deng Y, Dai L, Zhou Y, Zhang X, Cao M, Liu J, Sun J, Tang J, Chen Y, Huang X, Lin W, Ye C, Tong W, Cong L, Geng J, Han Y, Li L, Li W, Hu G, Huang X, Li W, Li J, Liu Z, Li L, Liu J, Qi Q, Liu J, Li L, Li T, Wang X, Lu H, Wu T, Zhu M, Ni P, Han H, Dong W, Ren X, Feng X, Cui P, Li X, Wang H, Xu X, Zhai W, Xu Z, Zhang J, He S, Zhang J, Xu J, Zhang K, Zheng X, Dong J, Zeng W, Tao L, Ye J, Tan J, Ren X, Chen X, He J, Liu D, Tian W, Tian C, Xia H, Bao Q, Li G, Gao H, Cao T, Wang J, Zhao W, Li P, Chen W, Wang X, Zhang Y, Hu J, Wang J, Liu S, Yang J, Zhang G, Xiong Y, Li Z, Mao L, Zhou C, Zhu Z, Chen R, Hao B, Zheng W, Chen S, Guo W, Li G, Liu S, Tao M, Wang J, Zhu L, Yuan L, Yang H (2002) A draft sequence of the rice genome (*Oryza sativa* L. ssp. *indica*). Science 296:79–91

Zambryski P, Depicker A, Kruger K, Goodman HM (1982) Tumor induction by *Agrobacterium tumefaciens* analysis of the boundaries of T-DNA. J Mol Appl Genet 1:361–370