The Plant Journal (2007) **49**, 947–959 doi: 10.1111/j.1365-313X.2006.03001.x

TECHNICAL ADVANCE

Non-random distribution of T-DNA insertions at various levels of the genome hierarchy as revealed by analyzing 13 804 T-DNA flanking sequences from an enhancer-trap mutant library

Jian Zhang[1,†], Dong Guo[1,†], Yuxiao Chang[1], Changjun You[1], Xingwang Li[1], Xiaoxia Dai[1], Qijun Weng[2], Jianwei Zhang[1], Guoxing Chen[1], Xianghua Li[1], Huifang Liu[1], Bin Han[2], Qifa Zhang[1] and Changyin Wu[1*]

[1]*National Key Laboratory of Crop Genetic Improvement and National Center of Plant Gene Research (Wuhan), Huazhong Agricultural University, Wuhan 430070, China, and*

[2]*National Center for Gene Research, Shanghai Institute for Biological Sciences, Chinese Academy of Sciences, Shanghai 200233, China*

Received 23 July 2006; revised 1 October 2006; accepted 24 October 2006.
[*]For correspondence (fax +86 27 87287092; e-mail cywu@mail.hzau.edu.cn).
[†]These authors contributed equally to this work.

Summary

We isolated 13 804 T-DNA flanking sequence tags (FSTs) from a T-DNA insertion library of rice. A comprehensive analysis of the 13 804 FSTs revealed a number of features demonstrating a highly non-random distribution of the T-DNA insertions in the rice genome: T-DNA insertions were biased towards large chromosomes, not only in the absolute number of insertions but also in the relative density; within chromosomes the insertions occurred more densely in the distal ends, and less densely in the centromeric regions; the distribution of the T-DNA insertions was highly correlated with that of full-length cDNAs, but the correlations were highly heterogeneous among the chromosomes; T-DNA insertions strongly disfavored transposable element (TE)-related sequences, but favored genic sequences with a strong bias toward the 5′ upstream and 3′ downstream regions of the genes; T-DNA insertions preferentially occurred among the various classes of functional genes, such that the numbers of insertions were in excess in certain functional categories but were deficient in other categories. The analysis of DNA sequence compositions around the T-DNA insertion sites also revealed several prominent features, including an elevated bendability from −200 to 200 bp relative to the insertion sites, an inverse relationship between the GC and TA skews, and reversed GC and TA skews in sequences upstream and downstream of the insertion sites, with both GC and TA skews equal to zero at the insertion sites. It was estimated that 365 380 insertions are needed to saturate the genome with $P = 0.95$, and that the 45 441 FSTs that have been isolated so far by various groups tagged 14 287 of the 42 653 non-TE related genes.

Keywords: flanking sequence tag, rice, T-DNA, insertion preference.

Introduction

Rice (*Oryza sativa* L.) is one of the most important crops in the world and has become a model plant in genome research for many reasons: the relatively small genome size; the availability of the whole genome sequence (International Rice Genome Sequencing Project, 2005); the mature T-DNA transformation techniques; the rich genetic and molecular resources; and the genome co-linearity with other monocots. As an effective strategy to study gene function, insertional mutagenesis has been widely adopted for constructing mutant libraries (Hirochika *et al.*, 2004; Jeon *et al.*, 2000; Krysan

et al., 1999; Martienssen, 1998; Osborne and Baker, 1995; Wu *et al.*, 2003). T-DNA is the most frequently used foreign DNA for mutant generation, as it not only disrupts the gene function, which facilitates gene identification, but also provides tags making gene isolation much easier. It has been believed that T-DNA insertion occurs randomly in the plant genome with a low copy number and stable inheritance (Azpiroz-Leehan and Feldmann, 1997). Transformation of T-DNA mediated by Agrobacterium can be highly efficient in many plant species including rice. Large numbers of T-DNA insertion lines of rice have been generated globally in recent years (Hirochika *et al.*, 2004).

An efficient strategy to find the functions of genes in the T-DNA mutant library is to sequence the regions flanking T-DNA insertion sites from large numbers of insertional mutants, and to establish database(s) to make the data of the flanking sequences readily accessible to the research community. So far, several databases with over 10 000 T-DNA flanking sequences have been established in Arabidopsis. For example, the SALK database (http://signal.salk.edu/cgi-bin/tdnaexpress; Alonso *et al.*, 2003) has collected 145 589 T-DNA flanking sequences; GABI-Kat (http://www.gabi-kat.de; Li *et al.*, 2003) and SAIL (Syngenta Arabidopsis Insertion Library; http://www.tmri.org/en/partnership/sail_collection.aspx; Sessions *et al.*, 2002) databases are also composed of 62 524 and 116 000 flanking sequences, respectively. Construction of the T-DNA flanking sequence database in rice has made tremendous progress, and several research groups have established databases containing up to tens of thousands of flanking sequences (An *et al.*, 2003; Chen *et al.*, 2003; Jeong *et al.*, 2006; Sallaud *et al.*, 2004; Zhang *et al.*, 2006). With the growth of available data, and the improvement of the presentation format, the flanking sequence databases will greatly promote functional genomic research, especially using a reverse genetic strategy.

T-DNA flanking sequences have also provided a valuable resource for studying the mechanism of T-DNA transformation. T-DNA integration sites have been extensively investigated in the last decade, especially in Arabidopsis (Pieter *et al.*, 2003; Takano *et al.*, 1997; Tinland, 1996). Alonso *et al.* (2003) reported a small bias of T-DNA insertions toward intergenic regions, but no significant difference was found in insertion frequencies between intron and exon regions, whereas Brunaud *et al.* (2002) reported a biased distribution of T-DNA insertion favoring intronic regions. In rice, Chen *et al.* (2003) analyzed the distribution of 1009 T-DNA flanking sequences, and found that T-DNA preferentially inserted into intron regions and no insertion frequency difference was observed between genic and intergenic regions. However, the analyses of An *et al.* (2003) and Sallaud *et al.* (2004) suggested a small bias of T-DNA insertions towards intergenic regions, but no significant difference in insertion frequencies between intron and exon regions. Such disagreement may be ascribed to the limitation of the sequence information and annotation at the times when the analyses were performed. With the complete sequence of the whole rice genome and progress in the genome annotation, the increasing availability of the flanking sequences may allow a critical assessment of this issue.

We have generated >110 000 independent transgenic lines with the enhancer trap construct using Agrobacterium-mediated T-DNA insertion with three japonica rice varieties (Wu *et al.*, 2003). The system has three built-in strategies for functional analysis of the rice genome. First, T-DNA insertions cause gene mutations, providing an efficient approach for gene identification and isolation. Second, expression of the reporter gene indicates the presence of an enhancer element in the genomic region nearby, which can be used for isolation and characterization of the enhancer. Third, the lines showing either spatial- or temporal-expression patterns of the reporter gene can be used to drive ectopic expression of a transgene, and are thus useful for unveiling latent functions of both unknown and known genes (Liang *et al.*, 2006). Large-scale screening and characterization of the mutant lines are in progress, which have generated comprehensive information of the transformants including the phenotypes, reporter-gene expression patterns, and flanking sequences (http://www.ricefgchina.org/mutant/) (Zhang *et al.*, 2006). The availability of the information provided by the large number of T-DNA flanking sequences has enabled the investigation of the T-DNA distribution and the possible insertion mechanisms.

This paper reports on our analysis of 13 804 rice T-DNA flanking sequences that we isolated in the last few years. The analysis revealed a highly non-random distribution of the T-DNA insertions in the rice genome, and a number of characteristics of sequence composition around the insertion sites. The results may have important implications for understanding the preference and mechanisms of T-DNA insertions in the rice genome.

Results

Isolation and homology analysis of the flanking sequences

Thermal asymmetric interlaced (TAIL)-PCR (Liu and Whittier, 1995; Liu *et al.*, 1995) was employed to amplify DNA fragments flanking the T-DNA borders from the transformants. The primer pair used in each amplification reaction consisted of a fixed primer, corresponding to either the left or the right border sequence of the T-DNA, and an arbitrary degenerate (AD) primer (see Experimental procedures). After optimizing the conditions for amplification and recovery, PCR products were obtained from approximately 30% of the transformants in the first round of amplification. A second round of PCR using another AD primer was applied to

the remaining transformants. In this way, amplification products could be recovered from approximately 50% of the transformants. So far, a total of 30 578 TAIL-PCR products were obtained and sequenced with the untrimmed length ≥100 bp, including 14 933 sequences in the length range 100–499 bp, 13 933 in the range 500–999 bp and 1712 of length ≥1000 bp, with an overall average length of 523 bp. Of the 30 578 sequences, 29 803 were amplified using primer pairs containing the left border sequence and 775 were amplified using primer pairs containing the right border sequence. Thus, amplifications using primers containing the left border sequence were much more efficient than ones using the right border sequence.

The homology search of the raw sequences, after trimming off the first 20 bp that was usually of low quality, was first performed using BLASTN against the vector. Then the vector sequence was masked with the CROSSMATCH program (Green, 1996), and a homology search of the masked sequences was performed against the TIGR rice pseudomolecules version 4 (ftp://ftp.tigr.org/pub/data/Eukaryotic_Projects/o_sativa/annotation_dbs/pseudomolecules/version_4.0/).

The 30 578 sequences could be divided into six categories according to their similarity with sequences of the rice genome and the vector:

(i) 11 177 (36.6%) sequences were as expected, containing several dozen bases homologous to the T-DNA border, and the remainder having significant (*E*-value < 10^{-5}) similarity with the rice genomic sequence, and thus could be mapped in the rice genome.
(ii) 955 (3.1%) had similarity (mini-score >20) with the vector sequence (either T-DNA with no border sequence or backbone, or both), in addition to significant similarity with the rice genome sequence, but not in the footprint region of the flanking sequence, which is likely to be a consequence of T-DNA rearrangement in the insertion sites.
(iii) 3622 (11.8%) had significant similarity with the rice genome sequence, but did not have significant similarity with the vector sequence.
(iv) 10801 (33.0%) had similarity only with the T-DNA.
(v) 3913 (12.8%) had similarity only with the backbone of the vector;
(vi) 829 (2.7%) had significant similarity with both the T-DNA and the backbone.

Of the 15 754 sequences in the first three categories that had significant similarity with rice genome sequences, 12 683 had distinct sites in the rice genome whereas the remaining 3071 sequences were resolved to 1121 sites, with each site harboring two or more insertions. In the following analysis, only one insertion was included from each of the 1121 sites, resulting in a total of 13 804 sequences with distinct insertion sites, which were henceforth referred to as flanking sequence tags (FSTs).

We also confirmed the FSTs by PCR amplification of the DNA sequences at the insertion sites using primer pairs, each consisting of a primer corresponding to the T-DNA sequence and a primer corresponding to the FST. Expected fragments were amplified for 180 out of the 248 cases tested, giving rise to a confirmed rate of 72.6%, which is similar to the rates reported in previous work on Arabidopsis (Sessions *et al.*, 2002). Seventeen of the 248 FSTs belonged to the category that did not contain the T-DNA footprints, 11 of which were confirmed by PCR amplification resulting in a confirmation rate of 64.7%.

T-DNA insertions favored large chromosomes

Table 1 shows the distribution of the 13 804 FSTs on the 12 rice chromosomes, from which it can be seen that the numbers of insertions on the various chromosomes were obviously related to the chromosome sizes, such that larger chromosomes (e.g. chromosomes 1, 2, 3 and 4) harbored larger numbers of insertions than did the smaller ones.

To investigate how well the numbers of insertions are correlated with chromosome size, we calculated the correlation coefficient between these two attributes, which is 0.92 ($P = 0.00001$), demonstrating that the numbers of insertions were highly dependent on chromosome size. However, a chi-square test for goodness-of-fit between observed numbers of insertions and numbers expected on the basis of chromosome size indicated a highly significant discrepancy ($\chi^2 = 635.3$, $P < 0.000001$). Numbers of observed insertions on chromosomes 1, 2 and 3 are significantly in excess, as indicated by the standardized residues (SRs) between the observed and expected numbers (Table 1), and the number of observed insertions on chromosomes 5–12, excluding chromosome 9, were less than expected whereas the observed and expected numbers on chromosome 4 did not differ significantly. Such distribution indicated that the differences in the numbers of the insertions on various chromosomes were not a result of chromosome size alone.

We thus calculated the densities of the insertions on the chromosomes, which clearly indicated a non-uniform distribution of insertions among the chromosomes (Table 1). The insertion density was highly correlated with chromosome size (correction coefficient $r = 0.74$, $P < 0.01$); insertions occurred more frequently per unit length on the larger chromosomes than on the smaller ones, and were thus biased in favor of larger chromosomes.

Plotting of insertion numbers along the chromosomes in windows of 500 kb (Figure 1) revealed the non-uniformity of the insertions within the chromosomes. In general, insertions occurred more densely in the distal regions of the chromosomes than in other regions, with the lowest density in the centromeric regions, with the possible exceptions of chromosomes 6 and 10.

Table 1 Distribution and density of the T-DNA insertions on the rice chromosomes

Chrom.	No. insertions		SR[b]	Chromosome size (Mb)[c]	Density (insertions/Mb)	r^d
	Observed	Expected[a]				
1	2070	1618	11.2	43.6	47.5	0.62
2	1659	1332	8.9	35.9	46.2	0.64
3	1851	1347	13.7	36.3	51.0	0.57
4	1245	1306	−1.7	35.2	35.4	0.80
5	982	1110	−3.8	29.9	32.8	0.62
6	1049	1158	−3.2	31.2	33.6	0.48
7	1020	1102	−2.5	29.7	34.3	0.75
8	861	1050	−5.8	28.3	30.4	0.82
9	831	853	−0.8	23.0	36.1	0.56
10	715	850	−4.6	22.9	31.2	0.75
11	751	1058	−9.4	28.5	26.4	0.55
12	770	1020	−7.8	27.5	28.0	0.74
Total	13804			372.0	37.1	0.70

[a]Expected number of insertions based on chromosome size. $\chi^2 = 635.3$ ($P < 0.000001$) for the test of goodness-of-fit between the observed and expected numbers of insertions on the 12 chromosomes.
[b]SR: standardized residue [=(observed − expected)/√expected], which follows a normal distribution asymptotically (Dai and Zhang, 1989). Thus an absolute SR value larger than 2.33 indicates statistical significance at $P < 0.01$. A positive value indicates that the observed number is greater than expected, and a negative value indicates that the observed number is smaller than expected.
[c]Data from TIGR (http://www.tigr.org/tdb/e2k1/osa1/pseudomolecules/info.shtml).
[d]*r*, correlation coefficient between the numbers of insertions and full-length cDNAs on various chromosomes in genome intervals of 500 kb.

To investigate the distributional relationship between T-DNA insertions and expressed sequences in the genome, we mapped all the 32 127 rice full-length cDNA sequences from the Knowledge-based Oryza Molecular biological Encyclopedia (KOME) to the TIGR rice pseudomolecules version 4 (Figure 1). We evaluated the correlation between the numbers of cDNAs and insertions in windows of 500 kb (Table 1). The correlation coefficient for the entire dataset of 12 chromosomes was 0.70 ($P < 0.01$), and the correlation coefficients varied from 0.48 to 0.82 among the 12 chromosomes, all of which were significant at the level of $P < 0.01$. A homogeneity test (Steel and Torrie, 1980) of the 12 correlation coefficients resulted in a chi-squared value of 27.1 ($P = 0.004$), indicating significant heterogeneity among the correlations. Correlations between insertions and cDNAs were relatively high for chromosomes 4 and 8, and low for chromosomes 6 and 11, 9 and 3. The higher correlations indicated a larger difference in the distribution of insertions between cDNA-dense regions and cDNA-sparse regions along the chromosomes, whereas the lower correlations indicated a more uniform distribution of the insertions along the chromosomes. Such differences may be related to the degrees of heterochromatin of the chromosomes.

T-DNA insertions strongly disfavored transposable element-related sequences

The genomic sequence could be grossly divided into three classes: genic, transposable element (TE)-related and intergenic sequences (International Rice Genome Sequencing Project, 2005). The numbers of insertions that occurred in each of the three sequence classes are given in Table 2. A chi-square test showed that there is a huge discrepancy ($\chi^2 = 1430$, $P < 0.000001$) between the observed numbers and the numbers based on the quantities of sequences in the three classes. T-DNA insertion density in the TE-related sequences (11.2/Mb) fell far below the average density (37.1/Mb) of the genomic sequence (Table 2), which was highly significant according to the SRs. The number of insertions in the genic sequence is much higher than the expectation, again statistically highly significant. It should be noted that the number of insertions in the intergenic sequence was also significantly greater than the expected number. Taken together, T-DNA insertions strongly disfavored the TE-related sequence among three classes of sequences.

T-DNA insertions preferentially occurred in the 5′ upstream and 3′ downstream regions of the genes

The genic sequences could be tentatively broken down to three disjointed regions: the 1-kb upstream of the translation start codon ATG, the coding region (from ATG to the translation stop codon), and the 500 bp downstream of the translation stop codon. We calculated the observed and expected numbers of T-DNA insertions in each of the regions (Table 2). The 500-bp downstream regions had a much higher density of T-DNA insertions than the genic region on average, followed by the 1-kb upstream region,

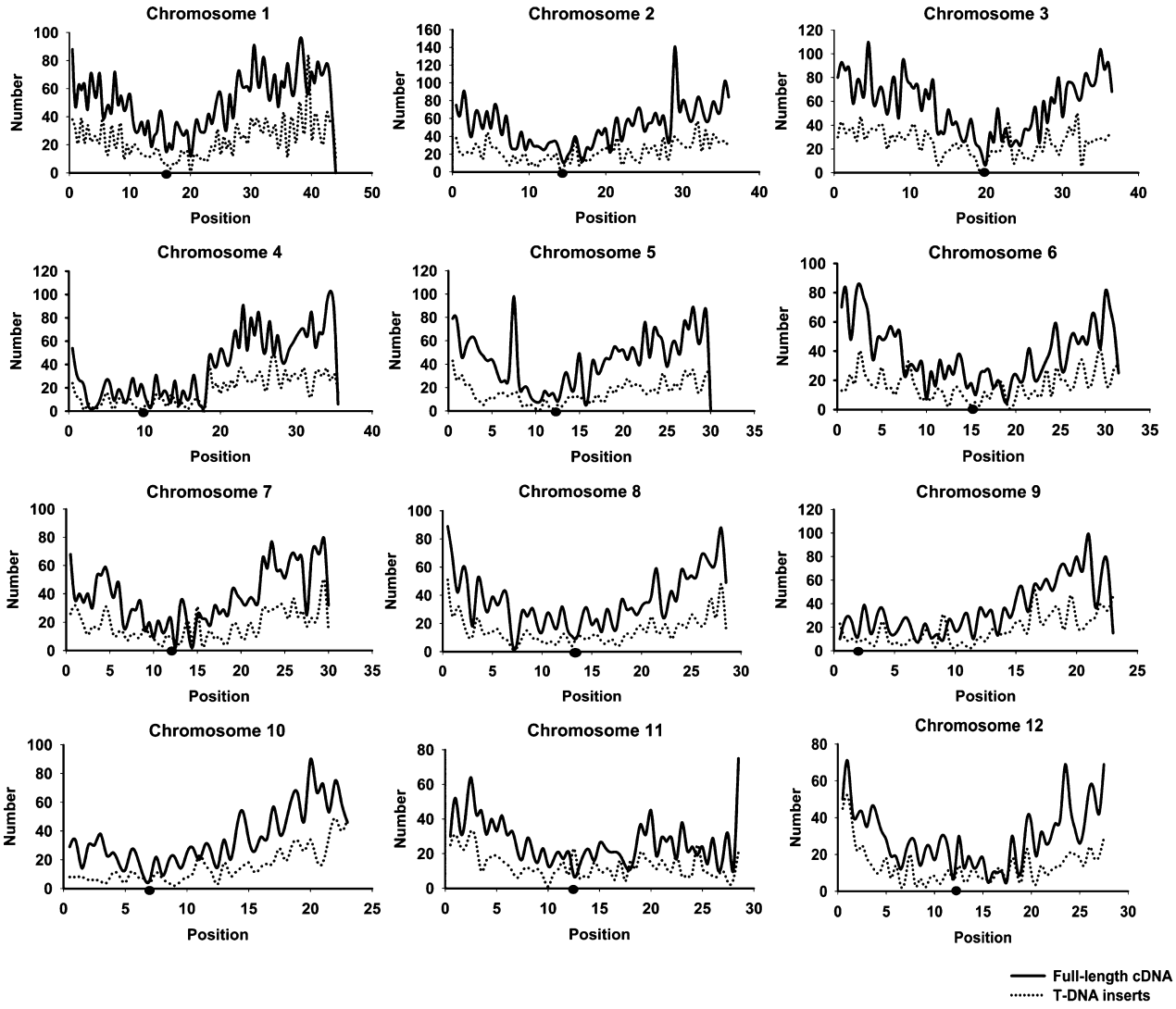

Figure 1. The distributions of the T-DNA insertions and full-length cDNA in the rice chromosome.
Position (*x*-axis) gives the distance (in Mb) from the first base in the rice pseudomolecules with a span of 500 kb. Number (*y*-axis) represents the number of either the insertions or the full-length cDNAs in each 500-kb region.

which was also much higher than the average for the genic region, whereas coding regions had a much lower density of insertions than the average for the genic region. A chi-square test revealed highly significantly uneven distribution of insertions in these three segments ($\chi^2 = 447$, $P < 0.000001$). Thus, T-DNA insertions preferentially occurred in the upstream and downstream regions of the genes.

Each of the coding regions could be further divided into exon(s) and intron(s), and the distribution of the T-DNA insertions in the exonic and intronic regions could again be compared. Such comparison revealed no significant difference of insertion density between introns and exons ($\chi^2 = 2.81$, $P = 0.09$).

Moreover, the data in Table 2 also allowed a statistical assessment of the difference in T-DNA insertion density between the coding and the intergenic sequences. A chi-square test for the null hypothesis of uniform T-DNA insertions in these two classes of sequences revealed that there was a highly significant difference ($\chi^2 = 19.61$, $P = 0.00009$) in the density of T-DNA distributions between intergenic and coding sequences. Thus, T-DNA insertions occurred significantly more frequently in intergenic regions than in coding regions.

T-DNA insertions differentially occurred in certain categories of functional genes

Together 6753 of the 13 804 FSTs corresponded to the TIGR models of functional genes that could be placed into the 11 functional categories using the annotation of 'molecular function' with the Gene Ontology vocabulary

Table 2 Distribution of the 13 804 T-DNA insertions in different regions of the rice genome

Test	Region	No. insertions	Expected[f]	SR[g]	Genome size (Mb)[h]	Density (insertions/Mb)
1	Genic[a]	7152	5952	15.55	160.4	44.6
	TE-related	712	2367	−34.0	63.8	11.2
	Intergenic[b]	5940	5484	6.2	147.8	40.2
2	Coding[c]	3665	4559	−13.2	100.1	36.6
	1 kb upstream[d]	2526	1958	12.8	43.0	58.7
	500 bp downstream[e]	1323	997	10.3	21.9	60.4
3	Intron	2025	1973	1.2	53.9	37.6
	Exon	1640	1691	−1.2	46.2	35.5
4	Intergenic	5940	5727	2.8	147.8	40.2
	Coding	3665	3873	−3.3	100.1	36.6

[a]Genic, including the coding region, and the 1-kb region upstream of the ATG and the 500-bp region downstream of the translation stop codon.
[b]The genomic sequences that are not included in gene sequences.
[c]Coding region, the genomic sequence from ATG to the translational stop codon.
[d]Genomic sequences 1000 bp upstream of the translational start codon (ATG).
[e]Genomic sequences 500 bp downstream of the translational stop codons.
[a,c,d-e]The numbers of insertions in coding, and the 1-kb upstream and the 500-bp downstream do not add up to the number of insertions in the genic sequence because sequences for some of the genes overlap with each other.
[f]Expectations based on null hypotheses for four different tests: (1) at the genome level, a test for homogeneity of insertions in genic, transposable element (TE)-related and intergenic sequences ($\chi^2 = 1430$, $P < 0.0000001$); (2) within the genic sequence, a test for homogeneity of insertions in coding, and the 1-kb upstream and the 500-bp downstream regions ($\chi^2 = 447$, $P < 0.000001$); (3) within the coding region, a test of homogeneity of insertions in introns and exons ($\chi^2 = 2.81$, $P = 0.09$); and (4) a comparison of T-DNA insertion density between coding and intergenic sequences ($\chi^2 = 19.6$, $P = 0.00009$).
[g]See the footnote of Table 1 for a definition of SR.
[h]Data from TIGR (http://www.tigr.org/tdb/e2k1/osa1/pseudomolecules/info.shtml).

Table 3 The number of T-DNA insertions in various categories of functional genes classified by 'molecular function' using the Gene Ontology vocabulary according to the TIGR model

Category	Number in TIGR model	No. FSTs[a] Observed	Expected[b]	SR[c]
Nutrient reservoir	64	2	18.0	−3.79
Motor	82	22	23.1	−0.23
Enzyme regulator	180	35	50.8	−2.21
Antioxidant	200	102	56.4	6.07
Structural molecule	421	105	118.7	−1.26
Signal transducer	445	137	125.5	1.04
Transporter	1507	457	425.0	1.55
Transcription regulator	2045	526	576.7	−2.11
Ligand binding or carrier	7864	2039	2217.6	−3.79
Catalytic	8893	2692	2507.8	3.68
Molecular function unknown	2246	636	633.4	0.10
Total	23947	6753	6753	

[a]FSTs, flanking sequence tags.
[b]Expectation based on the frequency of the genes in the various categories in the TIGR model.
[c]See footnote of Table 1 for a definition of SR.

(Table 3). To investigate whether the T-DNA insertions preferentially occurred in any of the functional categories, we performed a chi-square test for the null hypothesis of random distribution of the FSTs in the various functional categories (Table 3). It was shown that there was a significant non-random association between T-DNA insertions and the functional categories ($\chi^2 = 93.5$, $P < 0.000001$). To exclude the possibility that the observed association might be the result of a difference in gene length among the various categories, we calculated the correlation between insertion number and gene size in these categories, which was low ($r = 0.05$) indicating that the non-random association was not a result of gene size. Thus, the number of insertions in the category of Antioxidant was much higher than expected (SR = 6.07) according to the frequency of this category in the TIGR model. The number of insertions in the category of Catalytic was also greater than expected (SR = 3.79). In contrast, insertions in the categories of Nutrient reservoir, Enzyme regulator, Transcription regulator or Ligand binding and carrier were in lower than expected (SR values from −2.1 to −3.8).

No obvious tandem repeats, inverse repeats or palindrome structures in the T-DNA insertion sites

It has been suggested that the DNA integration reaction is highly dependent on the DNA configuration of the insertion

sites, which is influenced by the structural features of the DNA molecule (Katz *et al.*, 1998; Withers-Ward *et al.*, 1994). To characterize the DNA structure of the insertion sites, 1-kb sequences upstream and downstream of the insertion points were extracted to create 2-kb insertion-site nearby sequences (ISNS) of the 11 777 typical FSTs, which was analyzed using the EMBOSS (The European Molecular Biology Open Software Suite) program (Rice *et al.*, 2000) to detect possible tandem repeats, inverse repeats and palindrome structures. No obvious tandem repeats, inverse repeats and palindrome structures were found in the ISNS, indicating that these structures are not features of the T-DNA insertion sites.

Elevated bendability of the DNA sequence at the insertion sites

Sequence-dependent DNA bending, like sequence-dependent protein folding, has obvious biological importance in the recognition of particular DNA loci by restriction enzymes, repressors and other regulatory proteins. A pre-formed bend in the DNA would form a site for protein binding. The binding of the TATA-box recognition protein to a double DNA helix is a spectacular example in which major bends in the helix are induced at specific sequence loci (Goodsell and Dickerson, 1994; Satchwell *et al.*, 1986). It was also reported that the bendability is a very important feature in retrotransposon integration (Muller and Varmus, 1994).

Based on the method of Goodsell and Dickerson (1994), the 'banana' function of EMBOSS (Rice *et al.*, 2000) was employed to calculate the DNA bendability of the ISNS based on the 11 777 typical FSTs described above. For comparison, a control was generated by randomly extracting 2000 genome sequences of 2 kb in length from the TIGR rice pesudomolecules version 4. Elevated bendability was observed at positions from −200 to 200 bp around the insertion sites (Figure 2a), as compared with the control of 2000 random sequences (Figure 2c). The profile displayed in Figure 2(a) indicated that the bendability peak is symmetric about the insertion site in the range from −200 to 200 bp around the insertion site, with the highest points at both −10 and 10 bp from the insertion sites. There was an abrupt drop in the bendability value at the insertion sites. Thus, the 200-bp sequences upstream and downstream of the insertion sites are prone to bend to form a configuration that might be recognized by T-DNA insertion.

To confirm whether the bendability profile described above is a general feature of the ISNS in rice, we collected a total of 31 637 FSTs generated by three other groups, Genoplante (France), POSTECH (Korea) and Zhejiang University (China). The bendability of ISNS calculated is displayed in Figure 2(b), from which it can be seen that the

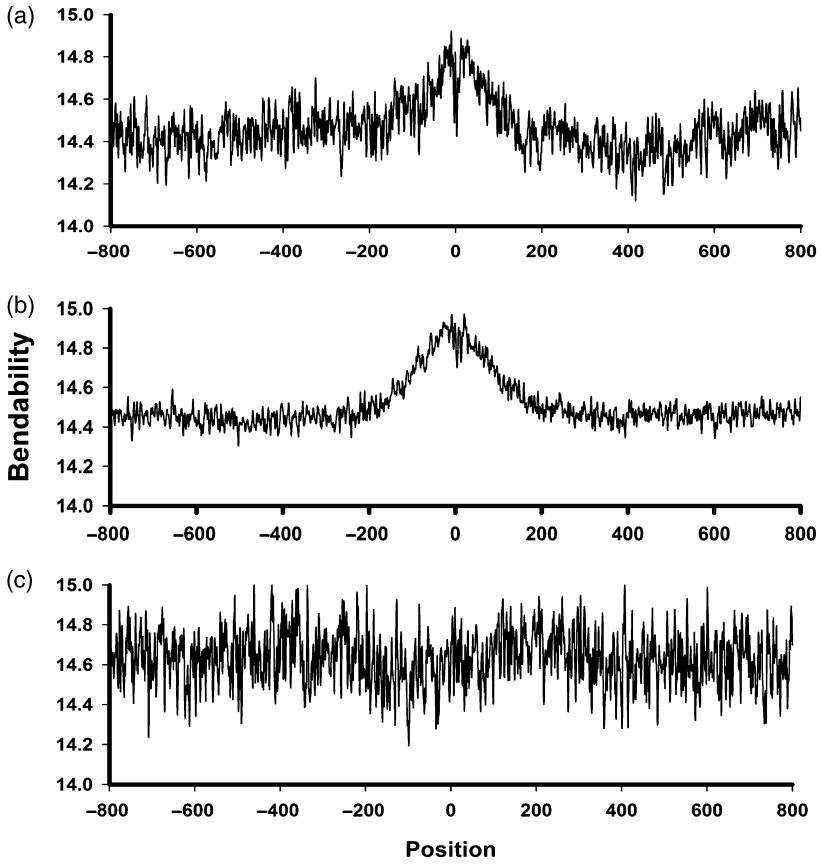

Figure 2. Bendability profiles at the insertion-site nearby sequences (ISNS) of the T-DNA insertion lines against 2000 randomly selected sequences.
(a) Bendability of 11 177 ISNS from our own data.
(b) Bendability of 31 637 ISNS from three other groups (see Experimental procedures).
(c) Bendability of the 2000 randomly selected 2-kb sequences. The *x*-axis represents the distance (in bp) from the insertion point '0'; '−' and '+' indicate upstream and downstream of the insertion point, respectively.
The bendability value at each position is the average of all the ISNS at the position.

feature was even more prominent than the results from our own data, mostly because of the much larger number of sequences that smoothed the oscillations of the curve. Again, there is an elevation in bendability from −200 to 200 bp of the ISNS, with peaks at approximately 10 bp from the insertion sites on both sides, and dropping down substantially at the insertion sites. This suggested that a bendability peak around the insertion sites and a lowered bendability at the insertion point are likely to be a common feature of the T-DNA insertion sites, regardless of the constructs used in the transformation.

An inverse relation between GC and TA skews at ISNS

We first calculated the average GC content of a 75-bp region extending upstream and downstream of the insertion points (151 bp in total), which averaged as 43.7% for our 11 177 T-DNA lines, identical to the average of 43.7% result from the 2000 random sequences, and almost identical to the GC content (43.6%) of the entire rice genome (International Rice Genome Sequencing Project, 2005). The GC content around the insertion sites of 31 637 FSTs that we collected from other groups was 44.6%, again not very different from the GC content of the rice genome. These results suggested that GC content might not be a feature of the T-DNA insertion sites in the rice genome.

GC skew and TA skew are parameters formerly used to study the substitution pattern of the leading strand and lagging strand in replication error, and are now extensively used to measure the asymmetry of the DNA double strands (Lobry, 1996). In a single strand, a positive GC skew means an overabundance of G against C, and a positive TA skew signifies an excess of T against A. The skew values of 875 bp of DNA sequences both upstream and downstream of the insertion sites from our own 11 177 FSTs are given in Figure 3(a), the skew values of the 31 637 FSTs from other groups are given in Figure 3(b), and the skew values of the 2000 random sequences are given in Figure 3(c). The profiles for the two sets of FSTs (Figure 3a,b) displayed three important features.

(i) The GC and TA skews appeared to be inversely correlated with $r = -0.98$ for data in Figure 3(a), and $r = -0.92$ for data in Figure 3(b), compared with $r = 0.09$ for the 2000 random sequences (Figure 3c).

(ii) The two curves crossed each other at the insertion sites (point 0) at which both GC and TA skews were equal to 0, indicating complete symmetry at this point.

(iii) From the insertion site to 800-bp upstream (−800 to 0 bp), the GC skew was positive and reached a plateau in approximately the region from −300 to −100 bp, indicating more G than C in the strand analyzed, whereas the TA skew was negative and formed a valley in a similar position in approximately the region from −300 to −100 bp, indicating fewer T than A. The reverse was the case from the insertion sites to 800 bp downstream (from 0 to 800 bp), whereas no such features were observed in the 2000 random sequences. These results indicated that the sequence compositions were quite asymmetric about the insertion sites, there were more G than C and less T than A in the upstream regions, and there were less G than C and more T than A in the downstream regions. In particular, the largest GC and TA skews (highly asymmetric in DNA composition) in the region from −200 to −100 bp and from 100 to 200 bp from the insertion sites, together with 0 skews of both GC and TA (completely symmetric), might be a critical configuration for T-DNA targeting.

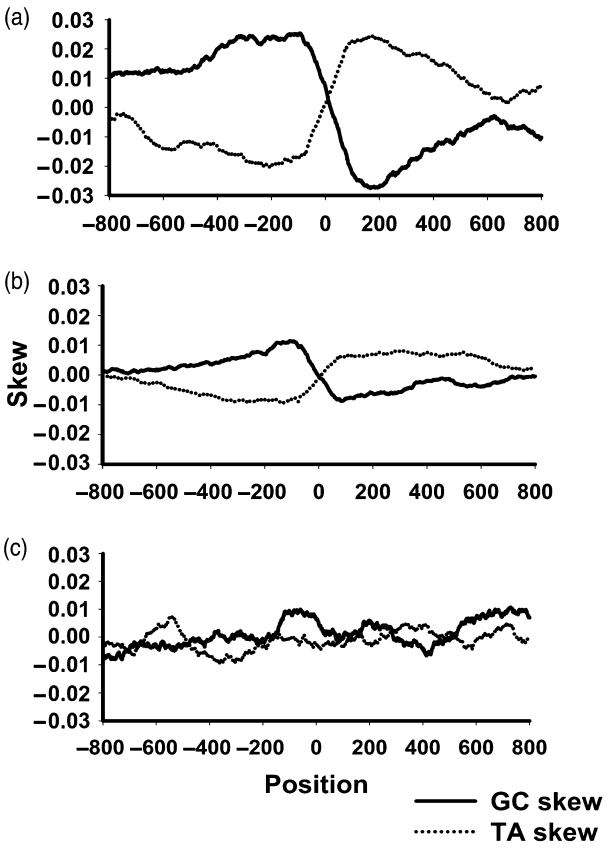

Figure 3. The profiles of GC and TA skews at the insertion-site nearby sequences (ISNS) of T-DNA insertion lines against 2000 randomly selected sequences.
(a) The GC and TA skews of 11 177 ISNS from our own data.
(b) The GC and TA skews of 31 637 ISNS from three other groups (see Experimental procedures).
(c) The GC and TA skews of 2000 randomly selected 2-kb sequences. The *x* axis represents the distance (in bp) from the insertion point '0'; '−' and '+' indicate upstream and downstream of the insertion point, respectively. The skew values were calculated in a range of 75 bp extending upstream and downstream of each position (151 bp in total), and the average value of the corresponding position was taken as the skew value (*y*-axis). The calculation formulas used were: GC skew = (G − C)/(G + C) and TA skew = (T − A)/(T + A), respectively.

Discussion

Distribution of the T-DNA insertions in the rice genome

The comprehensive analysis of the locations of the 13 804 FSTs in the rice genome revealed differential occurrence of T-DNA insertions that could be viewed at several levels.

When viewed at the chromosomal level, it was found that T-DNA insertions were biased in favor of large chromosomes, not only in terms of absolute number of insertions, but also in relative density of the insertions such that larger chromosomes also had a higher density of insertions. Within a chromosome the insertions were more densely populated in the distal ends, and less densely populated in the centromeric regions. Similar findings were also reported previously in Arabidopsis and rice (An *et al.*, 2003; Chen *et al.*, 2003; Jeong *et al.*, 2006; Sallaud *et al.*, 2004; Szabados *et al.*, 2002). This phenomenon may be related to the fact that the centromeric region is highly condensed and thus inaccessible to the T-DNA (Ortega *et al.*, 2002). Alternatively, it is also highly likely that DNA condensation suppresses the expression of the selective marker gene in the T-DNA during the transformation process, reducing the recovery rate of such transgenic events. A possible solution to recover such transformants is to use permissive rather than selective media in transformation in order to allow those transformants harboring silenced selective marker genes to grow normally, as suggested by the results of Francis and Spiker (2005). Such a strategy should receive due consideration in mutant library construction in order to saturate the genome.

It was also found that T-DNA insertions were highly correlated with the full-length cDNAs, similar to the finding in Arabidopsis (Schneeberger *et al.*, 2005). It was speculated that the gene-rich region is usually actively transcribed and more frequently in an 'open' state, and thus more accessible to T-DNA and easier to be integrated (Barakat *et al.*, 2000; Sha *et al.*, 2004). However, our results also showed that the correlations between insertions and full-length cDNAs were highly variable among the chromosomes, which may reflect the different levels of chromatin condensation of the chromosomes. A higher correlation would suggest a higher degree of differentiation of the chromosome regions into condensed (untranscribed) and not condensed (transcribed) regions, and conversely, a low correlation would suggest a lower level of differentiation of the chromosomal regions.

When viewed at the DNA sequence level, there is a highly uneven distribution of T-DNA insertions in different parts of the genic sequences. It was found that T-DNA insertions strongly disfavored TE-related sequences, and favored genic sequences. Within the various portions of the genic sequences, T-DNA insertions were highly biased towards the 5′ upstream and 3′ downstream regions of the genes, which is also similar to the results in Arabidopsis reported by Schneeberger *et al.* (2005). It should be noted that although overall T-DNA insertions occurred more frequently in genic than in intergenic sequences, the insertion density in coding sequences including exons and introns was lower than in intergenic sequences.

There is a possibility that the observed uneven distribution of the FSTs in the various classes of the sequences is the result of biases of the AD primers that may have preferentially amplified certain classes of the sequences. To assess this possibility, we calculated the frequencies of the occurrence of the AD primers in the various classes of genomic sequences (5′ upstream, 3′ downstream, coding, intergenic and TE-related), which were compared with the expectation in each class based on the frequencies of these primer sequences in the whole genome (data not shown). Although there was a highly significant non-random occurrence of these primer sequences in the various classes of DNA sequences, such a non-random occurrence was nonetheless not correlated with the distributions of FSTs, indicating that the biases of the primer sequences are not the major cause for the non-random occurrence of the FSTs.

Pryciak and Vamurs (1992) surveyed distribution of retroviral DNA integration sites in DNA and chromatin *in vitro*, and found that the sites of integration into both naked DNA and chromatin did not occur at random, implying that both nucleotide sequence and chromatin status may impose a large effect on the retroviral integration site selection. Whether this observation is relevant to the T-DNA insertion in the transformation remains to be investigated.

Another interesting finding is the preferential occurrence of the T-DNA insertions in certain classes of functional genes, such that the numbers of insertions in the Antioxidant and Catalytic functional catagories were greater than expected, whereas the number of insertions in the Nutrient reservoir, Enzyme regulator, Transcription regulator and Ligand binding or carrier categories were lower than expected. In Arabidopsis Schneeberger *et al.* (2005) compared T-DNA insertions rates between genes with high expression levels and ones with low expression levels, and found that genes with a high average expression level showed significantly higher rates of T-DNA insertion at −400, −100 and 0 bp, with respect to the translation start site. Li *et al.* (2006) also suggested that a lack of detectable transcriptional level is the main reason for the absence of the genes that are not covered by insertions in the T-DNA insertion library. Whether the differential T-DNA insertions observed in this study were related to the status of expression of the various categories of the genes at the time of Agrobacterium inoculation in the transformation process remains to be resolved in future studies.

DNA sequence composition and bendability around T-DNA insertion sites

The primary structure of DNA not only encodes genetic information but also contains sequence-dependent structural information important in the regulation of gene expression, transcription, replication and DNA packaging within the nucleus (Dlakic' and Harrington, 1996). The sequence context effect is significant to DNA physical properties, such as bendability and curvature (Dlakic' and Harrington, 1996).

Our analysis demonstrated several prominent features of DNA sequence compositions around the T-DNA insertions sites. First, there is an elevated bendability at in the region from −200 to 200 bp relative to the insertion sites. Second, there is an inverse relationship between the GC and TA skews at ISNS, and the GC and TA skews were reversed in sequences upstream and downstream of the insertion sites. Thus the sequence compositions are quite asymmetric in both GC and TA comparisons, and the asymmetry was reversed from upstream to downstream. A similar but less prominent feature of GC and TA skews was also observed in Arabidopsis (Schneeberger *et al.*, 2005). Interestingly, both the GC and TA skews were equal to 0 at the insertion sites, indicating that the GC and TA compositions were quite symmetric at this point.

The bendability of the DNA sequence has been reported as a very important feature in retrotransposon integration. For example, Liao *et al.* (2000) analyzed DNA bendability and several other characteristics regarding the physical properties of the DNA sequences and detected a bendability peak around the P element insertion site. A prominent peak in bendability was also detected around the T-DNA insertion site in Arabidopsis (Schneeberger *et al.*, 2005), although the width and height of the peak seemed to be different from the profile based on the rice data of this study. It was suggested that bending of the target DNA would create favorable attaching sites at the outer face of the helix, and thus high bendability is necessary for retroviral integration site selection (Dietrich *et al.*, 2002; Liao *et al.*, 2000; Muller and Varmus, 1994).

Brukner *et al.* (1995) suggested that the sequence asymmetry tends to form a bend DNA configuration that enhances the sensitivity to Dnase I nuclease cleavage, which is crucial in retrovirus and T-DNA integration. Thus the composition asymmetry of the DNA sequence, which affects DNA bendability and configuration, is a likely determinant for the location of T-DNA insertion. From the results described above we suggest that, in the case of rice, the higher bendability in the region of 200 bp upstream and downstream of the insertion sites may be helpful for creating the proper DNA configuration for the protein attachment in the integration reaction.

A genome-wide view of the current global rice T-DNA insertional mutagenesis efforts

With these data in hand, it is relevant to draw a global overview of the current status of T-DNA mutagenesis. We collected a total of 45 441 flanking sequences from four different research groups including our own. Based on the TIGR Transcript Unit (TU) model, 11 945 of the 45 441 T-DNA inserts were in coding regions of non-TE-related genes, 8067 inserts occurred in the 1 kb upstream of the ATG codon and 3482 inserts occurred in the 500 bp downstream of the translational stop codon (data not shown). If all these are considered as insertions in the genic regions, the total insertions in the genic regions would be 23 494, with 14 287 (33.5%) of the genes tagged at least once and the remaining 28 366 remaining untagged (Table 4). Although the majority of the genes were disrupted only once, 29 genes have been disrupted for either 10 times or more. Such results indicated the existence of likely 'hot spots' for insertions in these genes.

For a genome of 372 Mb in size, it can be calculated [using the formula $P = 1 - (1 - x/G)^n$ (Krysan *et al.*, 1999), in which x is the average size of the rice gene, which is 3.05 kb, and G is the genome size, which takes the value of 372 000 kb according to the TIGR model, n is the number of insertions needed to obtain the probability (P) of genome saturation] that a total of 365 380 insertions would be needed to saturate the genome with a probability of 0.95, assuming random integration of the T-DNA. With the view provided by the results of this study that the insertions favored genic regions, compared with intergenic and TE-related regions, the number of insertions needed to saturate the genome could be greatly reduced. According to data in Table 2, the genic region accounted for 43.1% of the genome and 56.9% were intergenic and TE-related sequences. By random integration, 157 479 of the 365 380 insertions would be expected to occur in genic regions. However, data in Table 2 indicated that insertions in genic regions accounted for 51.8% of the total insertions. Thus, to obtain 157 479 insertions in the genic regions, a total of 304 014 insertions rather than 365 380 would be needed. With the ~1.4–2.0

Table 4 Frequencies of insertions in non-transposable element (TE)-related genes in the rice genome based on currently available data from four research groups

Insertion times	No. genes	Frequency (%)
10 or more	29	0.1
6–9	145	0.3
3–5	1909	4.5
2	3205	7.5
1	8999	21.1
0	28 366	66.5
Total	42 653	100

Table 5 The fixed and AD primer sequences used for thermal asymmetric interlaced (TAIL)-PCR amplification

Primer	Usage	Sequence
LSP2	Primary reaction from left border	5′-GAAGTACTCGCCGATAGTGGAAACC-3′
LP2	Primary reaction from left border	5′-CTATCAGAGCTTGGTTGACGGCAATTT-3′
LBT2	Secondary reaction from left border	5′-ATAGGGTTTCGCTCATGTGTTGAGCAT-3′
LBT3	Tertiary reaction from left border	5′-CCAGTACTAAAATCCAGATCCCCCGAAT-3′
PFRB1	Primary reaction from right border	5′-GAGAAAAGGGTCCTAACCAAGAA-3′
PFRB2	Secondary reaction from right border	5′-GGGTCCTAACCAAGAAAATGAAG-3′
PFRB3	Tertiary reaction from right border	5′-CAAGAAAATGAAGGAGAAAAACTAGAA-3′
AD2-1	AD primer	5′-(AGCT)GACGA(GC)(AT)G A(AGCT)A(AT)GA A-3′
AD2a	AD primer	5′-(AGCT)GTCGA(GC)(AT)GA(AGCT)A(AT)GAA-3′
AD8	AD primer	5′-AG(AT)G(AGCT)AG(AT)A(AGCT)CA(AT)AGG-3′
AD10G	AD primer	5′-(AT)GTG(AGTC)AG(AT)A(AGCT)CA(AGCT)AGA-3′
AD11	AD primer	5′-TG(AT)G(AGCT)AG(GC)A(AGCT)CA(GC)AGA-3′
NTLB5	Sequencing from left border	5′-AATCCAGATCCCCCGAATTA-3′
PFRB4	Sequencing from right border	5′-TGCAGGTTCTCTCCAAATGA-3′

insertions per line in the mutant populations, as reported by various groups (Jeon *et al.*, 2000; Wu *et al.*, 2003), 152 007–217 217 transgenic lines in total would be needed to saturate the rice genome with a probability of 0.95. Although the number of transformants generated globally far exceeded the requirement calculated above, a technical difficulty in isolating the flanking sequences seems to exist as the progress has been slow and the total of 45 441 FSTs isolated thus far is far below the calculated value of 304 014 required to saturate the rice genome with 0.95 probability.

It should be noted in this context that backbone sequences beyond the T-DNA borders have been frequently found in Agrobacterium-mediated transgenic plants, with proportions ranging from 15% to 75% of the sequences isolated (De Buck *et al.*, 2000; Kim *et al.*, 2003; Kononov *et al.*, 1997). In rice, Kim *et al.* (2003) reported that 45% of the isolated flanking sequences contained the backbone sequence, which is much higher than the proportion observed in this study (15.5%). Furthermore, similar high frequencies (30% or more) of T-DNA sequences were also reported previously in Agrobacterium-mediated transgenic tobacco and rice populations (Afolabi *et al.*, 2004; Kim *et al.*, 2003; Krizkova and Hrouda, 1998). According to the 'two-phase integration mechanism model' (Kohli *et al.*, 1998), transforming plasmid molecules (either intact or partial) are spliced together in the 'pre-integration' phase, giving rise to rearranged sequences, which upon integration do not contain any interspersed plant genomic sequences. The high rates of T-DNA and backbone sequences in our results are apparently consistent with this model. Thus T-DNA tandem repeats may account for the greater number of the 10 801 (33.0%) sequences that had homology only with T-DNA. The high frequencies of T-DNA and backbone sequences indicate the necessity for developing a more efficient technique for isolating FSTs.

Moreover, with the huge numbers of T-DNA insertion lines generated by various groups globally (Hirochika *et al.*, 2004), efforts in integration and coordination of the data and characterization of the mutants through international collaboration becomes an urgent task for the international rice research community.

Experimental procedures

Plant materials and TAIL-PCR

Three japonica (*O. sativa* ssp. *japonica*) varieties Zhonghua 11, Zhonghua 15 and Nipponbare were used for the generation of a T-DNA mutant library. The vectors were pFX-E24.2-15R (Wu *et al.*, 2003), pSMR-J18R and pEGFP (http://rmd.ncpgr.cn), respectively. The selective gene hygromycin phosphotransferase (*hph*) driven by 35S was adjacent to the left border of the T-DNA, and the GAL4-VP16 element driven by the spliced −48 bp 35S promoter was next to the right border. A reporter gene was placed under the control of 6× UAS. The only difference among the three vectors lies in the reporter genes, such that Botany GUS was used in pFX-E24.2-15R, GFP was used in pSMR-J18R and enhanced GFP in pEGFP. The mutant library and the data generated have been described previously (Wu *et al.*, 2003; Zhang *et al.*, 2006). The TAIL-PCR was performed essentially according to the method described by Liu *et al.* (1995), with minor modifications, such that the number of cycles for the tertiary reaction increased from 20 to 35, and the reaction volume was reduced from 50 to 20 μl. Because the three vectors shared the same left and right borders, the same sets of nested primers AD primers (Table 5; Figure 4) were used for the amplifications.

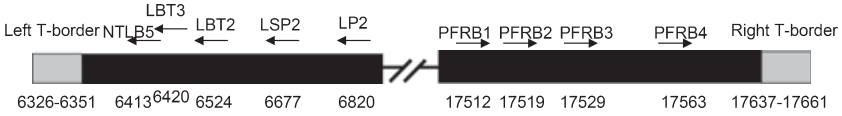

Figure 4. Schematic representation of the position of the thermal asymmetric interlaced (TAIL)-PCR specific primers in T-DNA. See Table 5 for the sequences of the primers. The numbers refer to the positions in the nucleotide sequence of the pFX-E24.2-15R. The arrows indicate the direction from 5′ to 3′ of primer sequence.

TAIL-PCR products purification and sequencing

For each reaction, 7 μl of the TAIL-PCR tertiary product was checked by 1% agarose gel (w/v) electrophoresis. The reactions that produced a specific band longer than 250 bp were subjected to enzyme digestion essentially as described by Sessions *et al.* (2002), with minor modifications. Briefly, the digestion reaction was in a volume of 8 μl containing 5 μl of TAIL-PCR tertiary product, 0.3 μl of 10x PCR buffer, 0.18 μl of 25 mM $MgCl_2$, 5 U Exonuclease I (New England Biolab, MA, USA), 0.25 U of alkaline phosphatase (shrimp) (Takara Company, Dalian, China), with double distilled H_2O to the final volume. The reaction mixture was incubated at 37°C for 60 min, and the reaction was terminated by heating the mixture to 80°C for 10 min. The digested products were directly sequenced with the dideoxy chain termination method using BigDye Terminator Cycle Sequencing V3.1 (Applied Biosystems, CA, USA) processed in an ABI3730xl (Applied Biosystems) sequencing machine.

Sequence analysis

Before conducting a homology search, the first 20 bases of the crude sequences were trimmed because of poor quality for most of the sequences. A homology search was first performed against the vector sequence including T-DNA and the backbone. The CROSSMATCH program (Green, 1996) was used to mask the T-DNA and backbone sequences, and the masked sequences were searched using BLASTN against the 12 rice pseudomolecules in the TIGR database (ftp://ftp.tigr.org/pub/data/Eukaryotic_Projects/o_sativa/annotation_dbs/pseudomolecules/version_4.0/). Insertion sites were determined at $E < 10^{-5}$. Functional classification of the tagged genes and TUs was performed with the GO annotation using terms of 'molecular function' from the TIGR all.xml file. The 'Equicktandem', 'einverted' and 'palindrome' functions of the EMBOSS program were used to detect tandem repeats, inverse repeats and palindrome structure, respectively. The 'banana' function of the EMBOSS program was employed to calculate the bendability. In order to perform batch analysis of bendability, we wrote a short script using PERL. We also wrote short programs for extracting the random sequence control and for calculating skew values using PERL.

Sequence collection

The 7177 flanking sequences of Genoplante (Montpelier, France), are directly downloaded from the URLs: (http://www.ncbi.nlm.nih.gov/entrez/query.fcgi?db=nucleotide&cmd=search&term=Rice%5BORGN%5D+Genoplante). The data of the insertions for the 27 611 flanking sequences isolated by POSTECH, Pohan, Korea, were provided in TIGR pseudomolecules sequences version 3 (http://141.223.132.44/pfg/index.php). The 1118 flanking sequences of Zhejiang University (Hangzhou, Zhejiang Province, China), were kindly provided by Professor Ping Wu.

Acknowledgements

We thank Professor Ping Wu of Zhejiang University for sequence data, and Weibo Xie and Gang Zhou for assistance in sequence analysis. The T-DNA insertion lines were generated by Xiangjun Li, Wenya Yuan, Zhihui Chen, Caishun Li and Xinqiang Gao. This research was supported in part by a grant from the National Special Key Project on Functional Genomics and Biochips of China, and a grant from the National Natural Science Foundation of China. We also acknowledge the support of the Ministry of Education of China.

References

Afolabi, A.S., Worland, B., Snape, J.W. *et al.* (2004) A large-scale study of rice plants transformed with different T-DNAs provides new insights into locus composition and T-DNA linkage configurations. *Theor. Appl. Genet.* **109**, 815–826.

Alonso, J.M., Stepanova, A.N., Leisse, T.J. *et al.* (2003) Genome-wide insertional mutagenesis of *Arabidopsis thaliana*. *Science*, **301**, 653–657.

An, S., Park, S., Jeong, D.H. *et al.* (2003) Generation and analysis of end-sequence database for T-DNA tagging lines in rice. *Plant Physiol.* **133**, 2040–2047.

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

Barakat, A., Gallois, P. and Raynal, M. (2000) The distribution of T-DNA in the genomes of transgenic *Arabidopsis* and rice. *FEBS Lett.* **471**, 161–164.

Brukner, I., Sanchez, R., Suck, D. and Pongor, S. (1995) Sequence-dependent bending propensity of DNA as revealed by DNase I: parameters for trinucleotides. *EMBO J.* **14**, 1812–1818.

Brunaud, V., Balzergue, S., Dubreucq, B. *et al.* (2002) T-DNA integration into the *Arabidopsis* genome depends on sequences of pre-insertion sites. *EMBO Rep.* **12**, 1152–1157.

Chen, S., Jin, W., Wang, M. *et al.* (2003) Distribution and characterization of over 1000 T-DNA tags in rice genome. *Plant J.* **36**, 105–113.

Dai, X. and Zhang, Q. (1989) Genetic diversity of six isozyme loci in cultivated barley of Tibet. *Theor. Appl. Genet.* **78**, 281–286.

De Buck, S., De Wilde, C., Van Montagu, M. *et al.* (2000) T-DNA vector backbone sequences are frequently integrated into the genome of transgenic plants obtained by *Agrobacterium*-mediated transformation. *Mol. Breed.* **6**, 459–468.

Dietrich, C.R., Cui, F., Packila, M.L. *et al.* (2002) Maize Mu transposons are targeted to the 5′ untranslated region of the *gl8* gene and sequences flanking Mu target-site duplications exhibit nonrandom nucleotide composition throughout the genome. *Genetics*, **160**, 697–716.

Dlakic′, M. and Harrington, R.E. (1996) The effects of sequence context on DNA curvature. *Proc. Natl Acad. Sci. USA*, **93**, 3847–3852.

Francis, K.E. and Spiker, S. (2005) Identification of *Arabidopsis thaliana* transformants without selection reveals a high occurrence of silenced T-DNA integrations. *Plant J.* **41**, 464–477.

Goodsell, D.S. and Dickerson, R.E. (1994) Bending and curvature calculations in B-DNA. *Nucleic Acids Res.* **22**, 5497–5503.

Green, P. (1996) Documentation for phrap. *Nature*, **390**, 580–586.

Hirochika, H., Guiderdoni, E., An, G. *et al.* (2004) Rice mutant resources for gene discovery. *Plant Mol Biol.* **54**, 325–334.

International Rice Genome Sequencing Project (2005) The map-based sequence of the rice genome. *Nature*, **436**, 793–800.

Jeon, J.S., Lee, S., Jung, K.H. *et al.* (2000) T-DNA insertional mutagenesis for functional genomics in rice. *Plant J.* **22**, 561–570.

Jeong, D.H., An, S., Park, S. *et al.* (2006) Generation of a flanking sequence-tag database for activation-tagging lines in japonica rice. *Plant J.* **45**, 123–132.

Katz, R.A., Gravuer, K. and Skalka, A.M. (1998) A preferred target DNA structure for retroviral integrase *in vitro*. *J. Biol. Chem.* **273**, 24190–24195.

Kim, S.R., Lee, J., Jun, S.H. *et al.* (2003) Transgene structures in T-DNA-inserted rice plants. *Plant Mol. Biol.* **52**, 761–773.

Kohli, A., Leech, M., Vain, P. *et al.* (1998) Transgene organization in rice engineered through direct DNA transfer supports a two-phase mechanism mediated by the establishment of integration hot sports. *Proc. Natl Acad. Sci. USA*, **95**, 7203–7208.

Kononov, M.E., Bassuner, B. and Gelvin, S.B. (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.

Krizkova, L. and Hrouda, M. (1998) Direct repeats of T-DNA integrated in tobacco chromosome : characterization of junction regions. *Plant J.* **16**, 673–680.

Krysan, P.J., Young, J.C. and Sussman, M.R. (1999) T-DNA as an insertional mutagen in *Arabidopsis*. *Plant Cell*, **11**, 2283–2290.

Li, Y., Rosso, M.G., Strizhov, N. *et al.* (2003) GABI-Kat SimpleSearch: a flanking sequence tag (FST) database for the identification of T-DNA insertion mutants in *Arabidopsis thaliana*. *Bioinformatics*, **19**, 1441–1442.

Li, Y., Rosso, M.G., Ülker, B. and Weisshaar, B. (2006) Analysis of T-DNA insertion site distribution patterns in *Arabidopsis thaliana* reveals special features of genes without insertions. *Genomics*, **87**, 645–652.

Liang, D., Wu, C., Li, C. *et al.* (2006) Establishment of a patterned GAL4-VP16 transactivation system for discovering gene function in rice. *Plant J.* **46**, 1059–1072.

Liao, G.C., Rehm, E.J. and Rubin, G.M. (2000) Insertion site preferences of the P transposable element in *Drosophila melanogaster*. *Proc. Natl Acad. Sci. USA*, **97**, 3347–3351.

Liu, Y.G. and Whittier, R.F. (1995) Thermal asymmetric interlaced PCR: automatable amplification and sequencing of insert end fragments from P1 and YAC clones for chromosome walking. *Genomics*, **25**, 674–681.

Liu, Y.G., Mitsukawa, N., Osumi, T. and Whittier, R.F. (1995) Efficient isolation and mapping of *Arabidopsis thaliana* T-DNA insert junctions by thermal asymmetric interlaced PCR. *Plant J.* **8**, 457–463.

Lobry, J.R. (1996) Asymmetric substitution patterns in the two DNA strands of bacteria. *Mol. Biol. Evol.* **13**, 660–665.

Martienssen, R.A. (1998) Functional genomics: probing plant gene function and expression with transposons. *Proc. Natl Acad. Sci. USA*, **95**, 2021–2026.

Muller, H.P. and Varmus, H.E. (1994) DNA bending creates favored sites for retroviral integration: an explanation for preferred insertion sites in nucleosomes. *EMBO J.* **13**, 4704–4714.

Ortega, D., Raynal, M., Laudie, M. *et al.* (2002) Flanking sequence tags in *Arabidopsis thaliana* T-DNA insertion lines: a pilot study. *C. R. Biol.* **325**, 773–780.

Osborne, B.I. and Baker, B. (1995) Movers and shakers. Maize transposons as tools for analyzing other plant genomes. *Curr. Opin. Cell Biol.* **7**, 406–413.

Pieter, W., Sylvie, D.B., Erik, V.B. *et al.* (2003) T-DNA integration in Arabidopsis chromosomes. Presence and origin of filler DNA sequences. *Plant Physiol.* **133**, 2061–2068.

Pryciak, P.M. and Vamurs, H.E. (1992) Nucleosomes, DNA-binding proteins, and DNA sequence modulate retroviral integration target site selection. *Cell*, **69**, 769–780.

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

Sallaud, C., Gay, C., Larmande, P. *et al.* (2004) High throughput T-DNA insertion mutagenesis in rice: a first step towards *in silico* reverse genetics. *Plant J.* **39**, 450–464.

Satchwell, S.C., Drew, H.R. and Travers, A.A. (1986) Sequence periodicities in chicken nucleosome core DNA. *J. Mol. Biol.* **191**, 659–675.

Schneeberger, R.G., Zhang, K., Tatarinova, T. *et al.* (2005) *Agrobacterium* T-DNA integration in Arabidopsis is correlated with DNA sequence compositions that occur frequently in gene promoter regions. *Funct. Integr. Genomics*, **5**, 240–253.

Sessions, A., Burke, E., Presting, G. *et al.* (2002) A high-throughput *Arabidopsis* reverse genetics system. *Plant Cell*, **14**, 2985–2994.

Sha, Y., Li, S., Pei, Z. *et al.* (2004) Generation and flanking sequence analysis of a rice T-DNA tagged population. *Theor. Appl. Genet.* **108**, 306–314.

Steel, R.G.D. and Torrie, J.E. (1980) *Principles and Procedures of Statistics. A Biometric Approach* 2nd edn. Toronto, Gruario: McGraw-Bill Book Co.

Szabados, L., Kovacs, I., Oberschall, A. *et al.* (2002) Distribution of 1000 sequenced T-DNA tags in the *Arabidopsis* genome. *Plant J.* **32**, 233–242.

Takano, M., Egawa, H., Ikeda, J.E. *et al.* (1997) The structures of integration sites in transgenic rice. *Plant J.* **11**, 353–361.

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

Withers-Ward, E.S., Kitamura, Y., Barnes, J.P. *et al.* (1994) Distribution of targets for avian retrovirus DNA integration *in vivo*. *Genes Dev.* **8**, 1473–1487.

Wu, C.Y., Li, X.J., Yuan, W.Y. *et al.* (2003) Development of enhancer trap lines for functional analysis of the rice genome. *Plant J.* **35**, 418–427.

Zhang, J., Li, C., Wu, C. *et al.* (2006) RMD: a rice mutant database for functional analysis of the rice genome. *Nucleic Acids Res.* **34** (Database issue), D745–748.

