Generation and Analysis of End Sequence Database for T-DNA Tagging Lines in Rice[1]

Suyoung An[2], Sunhee Park[2], Dong-Hoon Jeong, Dong-Yeon Lee, Hong-Gyu Kang, Jung-Hwa Yu, Junghe Hur, Sung-Ryul Kim, Young-Hea Kim, Miok Lee, Soonki Han, Soo-Jin Kim, Jungwon Yang, Eunjoo Kim, Soo Jin Wi, Hoo Sun Chung, Jong-Pil Hong, Vitnary Choe, Hak-Kyung Lee, Jung-Hee Choi, Jongmin Nam, Seong-Ryong Kim, Phun-Bum Park, Ky Young Park, Woo Taek Kim, Sunghwa Choe, Chin-Bum Lee, and Gynheung An*

National Research Laboratory of Plant Functional Genomics, Division of Molecular and Life Sciences, Pohang University of Science and Technology, Pohang 790–784, Korea (S.A., S.P., D.-H.J., D.-Y.L., H.-G.K., J.-H.Y., J.H., S.-R.K., Y.-H.K., M.L., G.A.); Department of Life Science, Sogang University, Seoul 121–742, Korea (S.H., S.-J.K., S.-R.K.); Department of Genetic Engineering, Suwon University, Suwon 445–743, Korea (J.Y., E.K., P.-B.P.); Department of Biology, Sunchon National University, Sunchon 540–742, Korea (S.J.W., K.Y.P.); Department of Biology, Yonsei University, Seoul 120–749, Korea (H.S.C., J-.P.H., W.T.K.); Department of Biology, Seoul National University, Seoul 151–747, Korea (V.C., S.C.); Department of Biology, Dong-eui University, Pusan 614–714, Korea. (H.-K.L., J.-H.C., C.-B.L.); and Institute of Molecular Evolutionary Genetics and Department of Biology, Pennsylvania State University, 208 Mueller Laboratory, University Park, Pennsylvania 16802 (J.N.)

We analyzed 6,749 lines tagged by the gene trap vector pGA2707. This resulted in the isolation of 3,793 genomic sequences flanking the T-DNA. Among the insertions, 1,846 T-DNAs were integrated into genic regions, and 1,864 were located in intergenic regions. Frequencies were also higher at the beginning and end of the coding regions and upstream near the ATG start codon. The overall GC content at the insertion sites was close to that measured from the entire rice (*Oryza sativa*) genome. Functional classification of these 1,846 tagged genes showed a distribution similar to that observed for all the genes in the rice chromosomes. This indicates that T-DNA insertion is not biased toward a particular class of genes. There were 764, 327, and 346 T-DNA insertions in chromosomes 1, 4 and 10, respectively. Insertions were not evenly distributed; frequencies were higher at the ends of the chromosomes and lower near the centromere. At certain sites, the frequency was higher than in the surrounding regions. This sequence database will be valuable in identifying knockout mutants for elucidating gene function in rice. This resource is available to the scientific community at http://www.postech.ac.kr/life/pfg/risd.

Insertional mutagenesis is one of the most useful methods for analyzing gene function. When foreign DNA is inserted into a gene, it not only creates a mutation but also tags the affected gene, facilitating its isolation and characterization (Azpiroz-Leehan and Feldmann, 1997). Transposons and T-DNA have been used most widely as an insertional mutagen (Mathur et al., 1998; Wisman et al., 1998; Krysan et al., 1999; Parinov et al., 1999; Speulman et al., 1999; Tissier et al., 1999). It is believed that T-DNA insertion is a random event and that the inserted sequences are stable through multiple generations (Azpiroz-Leehan and Feldmann, 1997; Parinov and Sundaresan, 2000). Insertional mutant pools have been constructed in Arabidopsis and used for func-

tional analysis of a number of genes (Feldmann, 1991; Koncz et al., 1992; Azpiroz-Leehan and Feldmann, 1997; Bechtold and Pelletier, 1998; Krysan et al., 1999; Galbiati et al., 2000; Parinov and Sundaresan, 2000; Bouché and Bouchez, 2001; Sessions et al., 2002; Szabados et al., 2002). The procedure for T-DNA insertional mutagenesis has also been applied to rice (*Oryza sativa*) using the *Agrobacterium tumefaciens*-mediated transformation method (Hiei et al., 1994). Jeon et al. (2000) have reported the construction of over 20,000 T-DNA-tagged rice lines. A T-DNA insertional mutagen can be modified to trap a gene by inserting a reporter gene, such as *gus* (*β-glucuronidase*), next to the T-DNA border (Sundaresan et al., 1995; Jeon et al., 2000; Springer, 2000). Approximately 5% to 10% of the mutagenized lines are GUS positive, demonstrating the efficiency of this gene-trapping system (Chin et al., 1999; Jeon et al., 2000).

Completion of the genome sequencing for both Arabidopsis and rice has provided new reverse genetic means for assigning biological functions to sequenced genes (Kumar and Hirochika, 2001; Feng et

al., 2002; Goff et al., 2002; Sasaki et al., 2002; Yu et al., 2002). Strategies have been designed to determine a particular gene's function by studying the phenotypes of individuals with mutations in the gene of interest. Two such approaches are now employed to screen for these mutants. The first, a so-called "pooling strategy", combines 100 to 1,000 separate insertion lines. The DNA extracted from these pools is then used to perform PCR screening with gene- and insert-specific primers (Krysan et al., 1999; Meissner et al., 1999 Tissier et al., 1999; Ríos et al., 2002). In rice, PCR screening of 6,000 Ac insertion lines has resulted in the isolation of mutations in two genes (Enoki et al., 1999). Also, 11,809 rice lines carrying 84,975 *Tos17* insertions have been pooled in a three-dimensional matrix (Hirochika, 2001). From that pool, 15 knockout mutants have been identified among the 47 genes screened for an insertion. Although this strategy is useful in identifying such knockout mutants, it is labor intensive and might not be suitable for the analysis of many genes (Kumar and Hirochika, 2001).

The second strategy involves determination of sequences flanking the insertion sites. These sequences can be obtained from DNA fragments amplified by thermal asymmetric interlaced (TAIL) PCR, inverse PCR (iPCR), or suppression PCR. When numerous flanking sequences are generated, they can then be catalogued in databases (Tissier et al., 1999; Parinov and Sundaresan, 2000; Alonso et al., 2003).

Large-scale application of this alternative strategy requires considerable effort (Parinov and Sundaresan, 2000). However, once established, the database can be easily shared with other scientists, facilitating distribution of the mutant materials and analysis of gene functions. Because sequencing of the entire genomes has been nearly completed in rice and Arabidopsis, the flanking sequence databases will become a powerful tool for systemically analyzing the functions of a large number of genes in those species (Parinov and Sundaresan, 2000; Walbot, 2000; Kumar and Hirochika, 2001; Pan et al., 2003). Databases of the *Ds* transposon insertion site sequences and T-DNA insertion sites already have been established for Arabidopsis (Parinov et al., 1999; Tissier et al., 1999; Ortega et al., 2002; Sessions et al., 2002). In rice, a tagged-sequence database has been derived from the *Tos17* insertional mutant lines (Hirochika, 2001; Yamazaki et al., 2001). In maize (*Zea mays*), DNAs adjacent to the transposed *Ac* elements have also been isolated and sequenced (Cowperthwaite et al., 2002).

Analyses of the insertion sites provide critical information about the characteristics of insertion elements. In Arabidopsis, for example, *Ds* elements tend to transpose near the chromosome ends but rarely near the centromeres (Ito et al., 2002). The distribution of insertions in the genic and intergenic sequences is roughly proportional to the ratio of genic to intergenic sequences throughout the entire genome, with preference being given to the region around the translation start codon (Raina et al., 2002). The retroelement *Tos17* appears to have hot spots for integration, although relatively low target site specificity has been observed at the nucleotide sequence level: No other structural features, e.g. hairpin loops or palindromes, have been seen in the target sequences (Yamazaki et al., 2001). Moreover, analysis of 1,000 T-DNA insertion sites in Arabidopsis has indicated that the majority of T-DNAs land in chromosomal domains of high gene density and that the frequency of insertions is higher in the 5'- and 3'-regulatory regions (Szabados et al., 2002). In the study presented here, we report the sequence analyses of 3,793 insertion ends tagged by T-DNA in rice.

RESULTS

Isolation of Sequences Flanking T-DNA

We previously established T-DNA insertional tagging lines of japonica rice using the binary vector pGA2707 (Jeong et al., 2002). In the current study, genomic DNA was prepared from young seedlings of the tagged lines and was then digested with *Pst*I or *Cla*I, which recognizes T-DNA once and does not cut the vector backbone. iPCR, using restriction enzymes that cut the vector more than once, frequently amplified either T-DNA or the vector backbone sequences (data not shown). After the cut DNA was ligated, the sequences flanking the right or left ends were amplified using the primers located at the T-DNA ends.

In the first attempt, *Pst*I was used, and more than one fragment was frequently amplified by iPCR. These were either different sequences independently tagged by T-DNA or the same sequences with different lengths due to partial digestion of the genomic DNA. Approximately 28% of our amplified DNA sequences were T-DNA or vector backbone sequences. T-DNA is frequently inserted into chromosomes as a tandem or invert repeat (Krizkova and Hrouda, 1998; Jeon et al., 2000; Kumar and Fladung, 2000). The vector backbone is also commonly cotransferred with the T-DNA (Kononov et al., 1997; Wolters et al., 1998; de Buck et al., 2000; Kim et al., 2003). In our study, the nucleotide sequence of the amplified DNA fragments was analyzed by the BLASTN homology search program, using the rice genomic sequence databases in National Center for Biotechnology Information (NCBI) and RiceGD. Our use of *Pst*I resulted in the identification of nonredundant rice DNA fragments flanking the T-DNA with 45% frequency. Lines that failed to generate flanking sequence information in the first experiment were then selected for a second iPCR, using *Cla*I. The frequency for isolating the genomic flanking sequence was 11% on this second attempt.

Analysis of 6,749 lines tagged by pGA2707 resulted in the isolation of 3,793 genomic sequences that flanked the T-DNA (Table I). End sequencing efficiency was higher at the left than at the right ends

Table I. *Distribution of T-DNA insertions in genic and intergenic regions*

Distribution of T-DNA Insertions	No.	Percentage
Genic	1,846	48.7
Exon	618	16.3
Intron	661	17.4
5'-untranslated region (UTR; within 300 bp)	311	8.2
3'-UTR (within 300 bp)	256	6.7
Intergenic	1,864	49.1
NA[a]	83	2.2
Total	3,793	100

[a] T-DNA was inserted into a contig that is less than 4 kb; therefore, annotation was difficult.

(data not shown), probably because inverted repeat T-DNA structures that face the right ends of T-DNA are a frequent event in transgenic rice plants (Kim et al., 2003).

Distribution of T-DNA Insertions

To learn whether T-DNA prefers the genic or the intergenic region, we analyzed its location with respect to putative genes. Among the 3,793 insertions, 1,846 (48.7%) of the T-DNAs were integrated into genic regions, and 1,864 (49.1%) were integrated into intergenic regions (Table I). The remaining 83 loci could not be determined because the contigs were too short to annotate. We considered the 300-bp se-

quences outside the start ATG and stop codon to be part of the intragenic region. If these UTR regions had not been included, insertion frequency in the genic region would have decreased to 1,279 (33.7%). Within the genic regions, T-DNAs were inserted into introns and exons at almost equal frequencies. Feng et al. (2002) have reported that the introns and exons are of approximately equal size in chromosome 4. If this phenomenon was true of the other chromosomes, no T-DNA insertion preference would exist between introns and exons.

We investigated the position of T-DNA to examine whether T-DNA preferred any particular location within the coding regions. Among 3,793 insertion sites obtained, 1,637 sites were located in the sequences annotated in NCBI. Analysis of the insertion sites located within 1,000 bp downstream from the start ATG showed that insertion frequency was higher at the beginning of the coding regions (Fig. 1A). Similarly, analysis of insertion sites located within 1,000 bp upstream from the stop codon indicated that the insertion frequency was higher at the end of the coding regions (Fig. 1B).

We also studied T-DNA insertion frequency in the region upstream of the start ATG codon (Fig. 1A) and in the 3'-downstream region from the stop codon (Fig. 1B). The results showed that regions near the start ATG and stop codon had a higher frequency of insertions than those far from the coding sequences. This result was similar to that observed in Arabidopsis (Szabados et al., 2002).

Figure 1. Distribution of T-DNA insertions within a size interval of 100 bp around the ATG and STOP codons. A, Distribution around the start codon ATG. Data were obtained from 196 sites located within 1 kb downstream from the ATG and from 241 sites located within 1 kb upstream from the start site. B, Distribution around the stop codons. Data were obtained from 244 sites located within 1 kb downstream from the stop codons and from 164 sites located within 1 kb upstream from the stop codons.
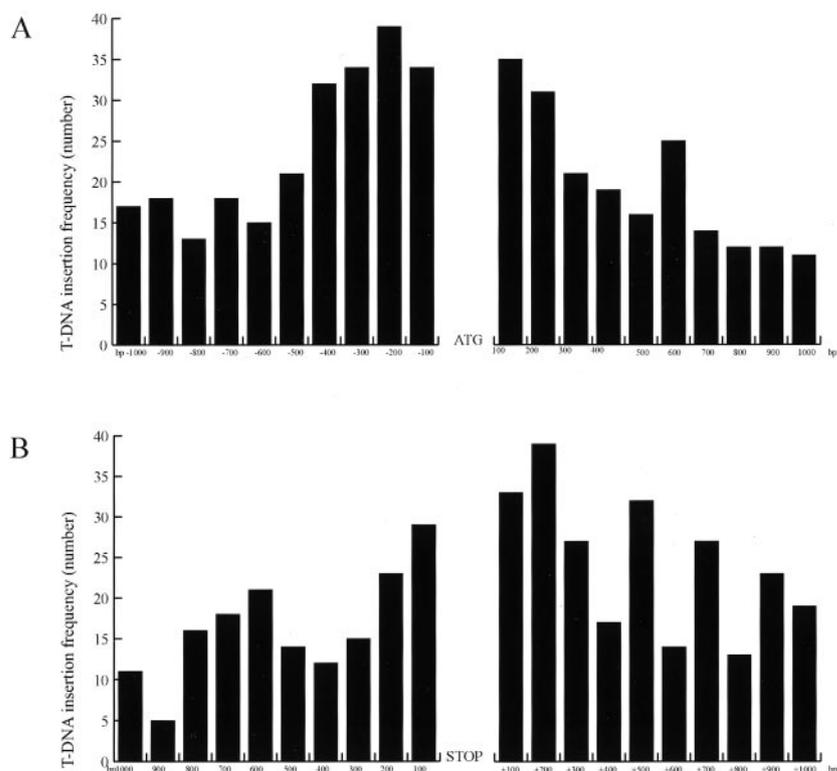
Table II. *GC contents at the T-DNA insertion sites*

Data were calculated from 1,846 insertions in genic regions and 1,864 insertions from intergenic regions.

T-DNA Insertion Sites	Chromosome 1	Chromosome 4	Chromosome 10	Total
		%		
Genic	46.4	44.4	45.0	45.7
Exon	48.3	49.7	48.3	49.3
Intron	44.1	40.7	41.4	42.3
5'-UTR (within 300 bp)	48.1	47.1	47.0	48.4
3'-UTR (within 300 bp)	42.3	41.5	39.5	42.6
Intergenic	43.1	43.8	42.3	43.2
Total	45.4	44.5	43.9	45.2

The nos. of genic insertions were 446, 143, and 134 in chromosomes 1, 4, and 10, respectively. The nos. of intergenic insertions were 318, 184, and 212 in chromosomes 1, 4, and 10, respectively.

Distribution of GC Content in the Exon and Intron Insert Sites

We investigated whether T-DNA preference was influenced by the GC content by examining 3,793 insertion sites. Regions that were within 100 bp upstream or downstream of the insertion position were analyzed. Overall GC content within the 200-bp in-

sertion site region was 49.3% in the exons and 42.3% in the introns (Table II). This level calculated at the insertion sites is close to that observed from the entire rice genome (Yu et al., 2002), where mean GC content is 43.3% and that of the exons and introns is 51.4% and 37.0%, respectively.

A plot of the GC content distribution in the insertion sites showed that the exons displayed a GC-rich tail, whereas the introns did not (Fig. 2). Similar distributions have been observed with exons and introns from the entire genome (Yu et al., 2002), indicating that T-DNA insertion does not favor a particular GC content.

Functional Classification of Genes Tagged by T-DNA

The 1,846 genes tagged by T-DNA were functionally classified by their sequence homology to known proteins using the software package INTERPRO (http://www.ebi.ac.uk/interpro). The output was filtered to create sets of the longest domain for each associated protein. Domains were categorized using the Gene Ontology software (http://www.geneontology.org; Goff et al., 2002). Our tagged genes (Table III) could be classified into 10 functional groups, as described by Feng et al. (2002). When it was difficult to assign functions to some predicted coding proteins, they were classified as "other." The most frequently tagged genes were those involved in metabolism. The second group comprised, in almost equal abundance, genes that encode transcription factors and signaling molecules, and the third group contained genes involved in defense and transport. Because these results are quite similar to the gene distribution reported for chromosomes 1, 4, and 10 (Feng et al., 2002; Sasaki et al., 2002; Rice Chromosome 10 Sequencing Consortium, 2003), we conclude that T-DNA insertion is not biased toward a particular class of genes.
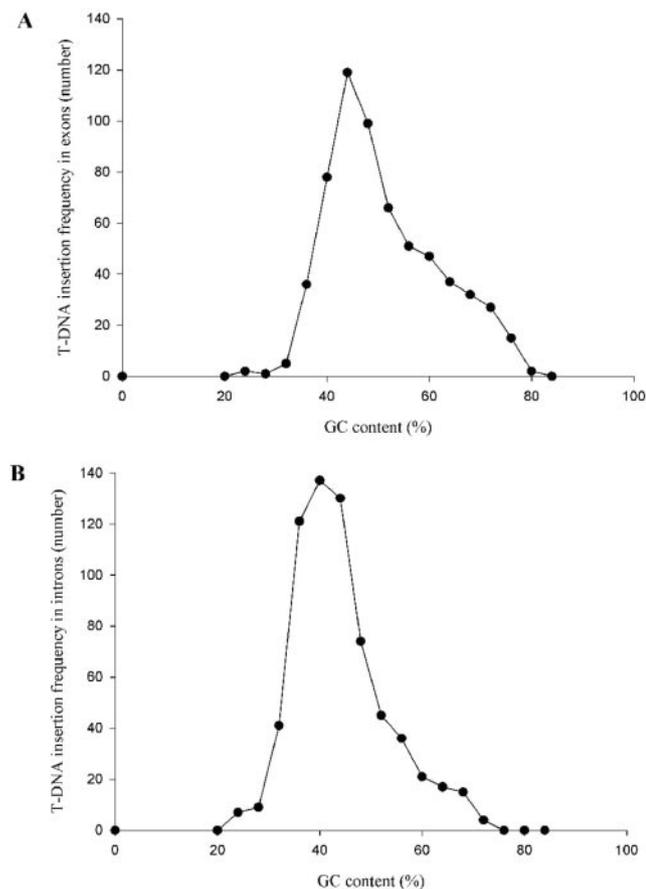
Figure 2. GC content distribution in exons and introns at insertion sites. T-DNA insertions in 618 exons (A) and 661 introns (B) were used for estimating GC contents in the region 100 bp upstream and 100 bp downstream from insertion sites.

Distribution of T-DNA Tags in Rice Chromosomes

Because almost the entire sequences for chromosomes 1, 4, and 10 are available (Feng et al., 2002;

Table III. *Functional classification of the tagged genes*

|	Chromosome 1	Chromosome 4	Chromosome 10	Total
		No. (%)		
Metabolism	46 (10.0)	20 (14.0)	16 (11.9)	210 (11.4)
Transcription	25 (5.5)	18 (12.6)	9 (6.9)	150 (8.1)
Signaling	32 (7.0)	8 (5.6)	12 (9.0)	115 (6.2)
Defense	17 (3.7)	6 (4.2)	5 (3.7)	71 (3.8)
Transporter	11 (2.4)	2 (1.4)	3 (2.2)	53 (2.9)
Protein fate	9 (2.0)	2 (1.4)	5 (3.7)	39 (2.1)
Cell structure	9 (1.9)	1 (0.7)	5 (3.7)	37 (2.0)
Retroelement	6 (1.3)	4 (2.8)	13 (9.7)	49 (2.7)
Unclassified	9 (2.0)	17 (11.9)	11 (8.2)	104 (5.7)
Other	282 (61.4)	65 (45.5)	55 (41.0)	1,018 (55.1)
Total	446	143	134	1,846

Sasaki et al., 2002; Rice Chromosome 10 Sequencing Consortium, 2003), we analyzed the distribution of T-DNA insertions, and found 764, 327, and 346 in chromosomes 1, 4, and 10, respectively. After these insertions were mapped on the bacterial artificial chromosome or phage P1-derived artificial chromosome clones, their frequencies were scored (Fig. 3). The results indicated that T-DNA insertions were not evenly distributed, with frequencies being higher at the ends of the chromosomes and lower in the region near the centromere. T-DNA insertions were also more frequent in certain areas compared with the surrounding region. When gene density was plotted for chromosome 1 (Fig. 3A), we also found that more genes occurred at the ends than near the centromere. However, this difference was not as significant as that for insertion frequency (Fig. 3C).

DISCUSSION

We have established 3,793 rice genomic sequences flanking the T-DNA from 6,749 T-DNA tagging lines. Systematic cataloging of tagging mutants by random sequencing has been achieved via *Ds*-tagged (Pari-

Figure 3. Distribution of gene numbers and T-DNA insertion sites along chromosomes 1, 4, and 10. Chromosomes are represented on the *x* axis and are numbered from 1 bp of the telomere on the short arm to the last base pair of the telomere on the long arm by joining up the physical gaps. The *y* axis shows the gene density on rice chromosome 1 (A), average intergenic length (B), and T-DNA insertion frequency on chromosome 1 (C), chromosome 4 (D), and chromosome 10 (E). Gene density, average intergenic length, and T-DNA insertion frequency were obtained by counting the number of events in every 0.5-Mb section. Centromeric regions are indicated by circles.
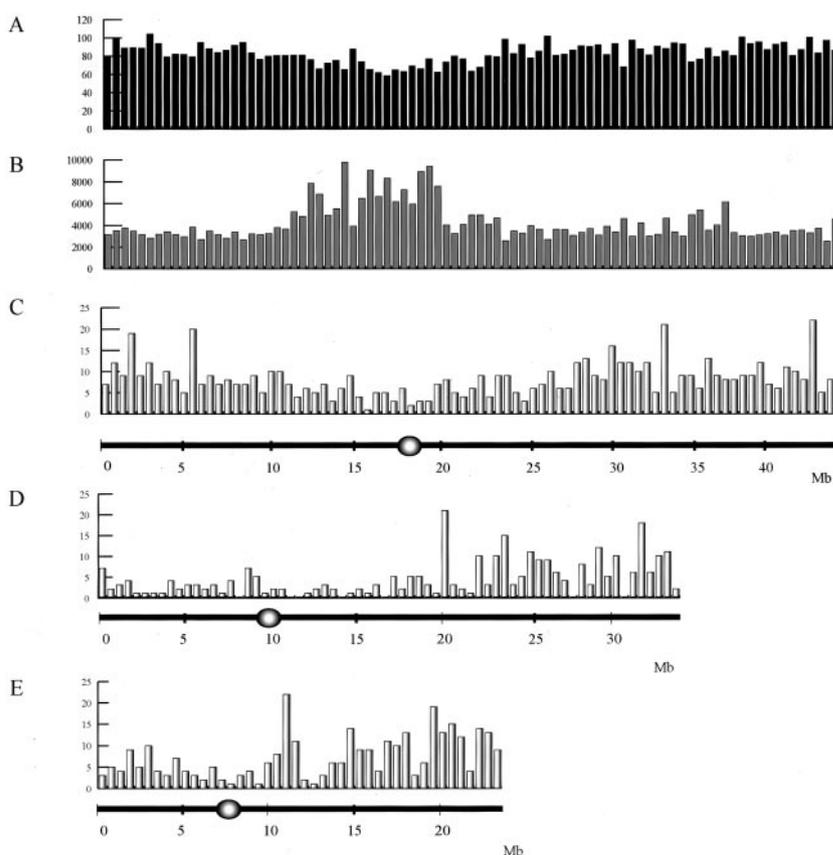
nov et al., 1999; Ito et al., 2002; Raina et al., 2002) and T-DNA-tagged (Ríos et al., 2002; Szabados et al., 2002) Arabidopsis and from mutator-tagged maize (Cowperthwaite et al., 2002). In rice, flanking sequences have been determined for *Tos17*-tagged lines (Hirochika, 2001; Yamazaki et al., 2001). Our results here are the first report, to our knowledge, on a large-scale end sequence database of T-DNA-tagged rice.

Among the end sequences generated in this study, 1,846 (48.7%) insertions occurred within the genic regions. The frequency of insertions into these regions was much higher than the overall frequency of genic regions in the rice genome (i.e. 22%–35%; Feng et al., 2002; Goff et al., 2002: Sasaki et al., 2002). This result indicates that T-DNA prefers genic regions in rice. This insertion frequency for genic regions was similar to the proportion of insertion in genic regions found in the Arabidopsis genome (Szabados et al., 2002).

Plant DNA sequences that flank a known insert have been isolated by several methods, including TAIL-PCR, plasmid rescue, adapter-mediated PCR, and iPCR (Koncz et al., 1990; Liu and Whittier, 1995; Mathur et al., 1998; Yephremov and Saedler, 2000). We employed the iPCR technique in this study. Compared with TAIL-PCR, iPCR uses specific primers and, therefore, achieves a higher success rate. In addition, the iPCR product is longer than that obtained with TAIL-PCR, providing a more unambiguous sequence identification.

In pilot experiments using *Hae*II and *Eco*RI, which cut more than once in the binary vectors, the frequency of isolating sequences of tandem or complex T-DNA insertions was higher than that for rice genomic DNA that flanks T-DNA insertions (data not shown). To enhance the efficiency of isolation, we chose *Pst*I and *Cla*I, which recognize a unique site in T-DNA. Because the recognition site of *Pst*I is located at the center of the T-DNA, DNA sequences flanking both right and left T-DNA ends could be obtained from one enzymatic digestion. In contrast, iPCR efficiency was lower with *Cla*I, perhaps reflecting the

frequency of enzyme sites in rice chromosomes. Alternatively, the efficiency of self-ligation may vary among different enzymes.

We observed that the T-DNA insertion frequency was higher in and near genes, a characteristic that increases the chances of finding an insertion within a given gene. The higher density of T-DNA tags both up- and downstream of the predicted coding regions indicates that these sequences are more accessible for insertion. Because genic regions contain higher GC contents (Feng et al., 2002; Sasaki et al., 2002), we analyzed the sequences at the insertion sites and found that T-DNA preference was not due to any difference in GC contents within the insertion sites. Although our T-DNA preferred genic regions, there was no bias in the functional classes of the tagged genes. Moreover, if actively transcribed regions were targets for T-DNA, genes expressed in callus tissue would be expected to be more frequently found as insertion sites. Therefore, the level of transcriptional activity may not be a determinant of frequency. Another mechanism must exist that recognizes genic regions by T-DNA when it is inserted into the chromosome, e.g. the chromatin structure.

Among our 3,793 flanking sequences, 20.1% (764 insertions) were located in chromosome 1. Chromosome 1 covers 10.1% (43.3 Mb) of the rice genome; in our sampling, T-DNA insertion sites were more frequently located there than in any other chromosome. Because gene density is also greater in chromosome 1, its higher frequency may be a reflection of the tendency for T-DNA insertions to occur in gene-dense regions. In contrast, the frequency we found for chromosome 4 is similar to the proportion of chromosome 4 DNA to the entire genome (Feng et al., 2002).

We also observed that T-DNA insertions were not evenly distributed on the chromosomes, with integration frequencies being lower near the centromere and higher in the distal regions, where gene density was higher. A similar distribution has been observed in Arabidopsis chromosomes (Szabados et al., 2002). Although T-DNA preference was higher in certain

Table IV. *Primers used for iPCR*

Primer name	Enzyme	Direction	PCR	Sequence
GUS2R	PstI,ClaI	Right	1st, 1st	TTGGGGTTTCTACAGGACGTAAC
HPH4R	PstI	Right	1st	CCATGTAGTGTATTGACCGATTC
GUS1R	PstI,ClaI	Right	2nd,2nd	CAAGTTAGTCATGTAATTAGCCAC
HPH3R	PstI	Right	2nd	TCGTCTGGCTAAGATCGGCCGCA
HPH1F	PstI	Left	1st	GATCGTTATGTTTATCGGCACTT
Ptub3R	PstI	Left	1st	GGTGAATGGCATCGTTTGAA
HPH2F	PstI	Left	2nd	AGTGCTTGACATTGGGGAATTCAG
Ptub2R	PstI,ClaI	Left	1st,2nd	ACAAGCCGTAAGTGCAAGTG
Tnos2F	ClaI	Right	1st	ATGATTAGAGTCCCGCAATT
Tnos1F	ClaI	Right	2nd	ACAAAATATAGCGCGCAAAC
Ttub1F	ClaI	Left	1st	CCTAGTGGCCATTGTGCGTT
Ttub2F	ClaI	Left	2nd	GCAGTTTGCTGCACTTACAAC
Ptub1R	ClaI	Left	2nd	TGTGAAGAAAATTACTTCCTTC

regions compared with the surrounding regions, the bias was not as frequent as the insertion preferences observed from transposons and retrotransposons (Parinov et al., 1999; Yamazaki et al., 2001). This result suggests that T-DNA tagging is more feasible for reverse genetic studies.

Because genome sequencing and annotation of rice will be completed soon, functional analysis of each gene will become an important tool for understanding gene function. Several T-DNA-tagging systems have been constructed for the examination of rice genes (Jeon et al., 2000, 2002). Conventional forward genetic strategies that use these materials are laborious and take considerable time because of the long life cycle and large plant size of rice. Another difficulty inherent to forward genetics is establishing the linkage between visible phenotypes and the insertion. Because the tagging lines are generated via tissue culture procedures after cocultivation of scutellum-derived calli and *A. tumefaciens* that carries the tagging vector, some endogenous transposons, such as *Tos17* and *miniature ping*, become active (Hirochika et al., 1996; Kikuchi et al., 2003). Other unknown mechanisms also generate mutations. Therefore, establishing the end sequence database for insertional mutant sites will be extremely valuable when identifying mutants and sharing precious resources.

MATERIALS AND METHODS

Plant Materials and Growth Conditions

Japonica varieties Dongjin and Hwayoung were used for the generation of transgenic rice (*Oryza sativa*). The vector was pGA2707 (Jeong et al., 2002). For DNA preparation, 20 seeds from the primary transgenic plants were sterilized in 0.025% (w/v) prochloraz (Aventis Crop Science, Yongin, Korea) for 24 h, then imbibed in tap water for 2 d. Seeds were sown in soil and grown for 15 d under greenhouse conditions, with a 14+-h photoperiod and a minimum night temperature of 20°C.

Genomic DNA Preparation

Fully expanded leaves from four seedlings were harvested and placed in 2-mL Safe-Lock microcentrifuge tubes (Eppendorf, Hamburg, Germany). The fresh weight of each leaf was approximately 200 mg. A 3-mm tungsten bit (Qiagen, Hilden, Germany) was put into the tube, and the samples were frozen in liquid nitrogen. Each tube was then vibrated at 1,200 rpm for 1 min with a Grinding Mixer Mill MM300 (Retsch, Haan, Germany). Afterward, the samples were frozen again by submerging the tubes in liquid nitrogen and vibrated at 1,200 rpm for an additional 1 min. DNA was prepared via the modified hexadecylotrimethylammonium (CTAB) method (Chen and Ronald, 1999). In brief, the ground samples were suspended in 750 µL of warm (65°C) CTAB buffer (2% [w/v] CTAB, 1.42 M NaCl, 20 mM EDTA, 100 mM Tris-HCl [pH 8.0], 2% [w/v] polyvinylpyrrolidone-40, and 5.0 mM ascorbic acid). After the addition of 7 µL of RNaseA (20 mg mL^{-1}), the samples were incubated at 65°C for 5 min. They were then extracted with 0.7 volumes of chloroform and centrifuged for 10 min at 15,000 rpm. Each upper aqueous phase was transferred to a new tube, and the DNA was precipitated with 0.7 volumes of isopropanol by immediately spinning the mixture at 12,000 rpm for 10 min. The DNA pellet was washed with 70% (w/v) ethanol, dried, and resuspended in 50 to 100 µL of TE solution (10 mM Tris [pH 8.0] and 1 mM EDTA), resulting in a final DNA concentration of approximately 200 mg L^{-1}.

iPCR

We used the iPCR method (Hui et al., 1998; Spertini et al., 1999) to isolate the flanking sequences of T-DNA, One microgram of genomic DNA was digested with 10 units of *Pst*I or *Cla*I in 50 µL for 10 h. After the enzymes were heat inactivated, the samples were ethanol precipitated and dissolved in 50 µL of ligation buffer, then ligated at 8°C for 12 h, using 1 unit of T4 DNA ligase (Boehringer Mannheim/Roche, Basel). Nested PCR was performed to amplify the flanking sequence. For the first PCR, approximately 1/50 of the ligated DNA and 5 µM of each primer were incubated in 25 µL of a reaction solution containing dNTPs and 0.1 unit of ExTaq polymerase (Takara, Gennevillers, France). PCR was performed with an initial 5-min denaturation at 94°C, followed by 35 cycles (each cycle: 94°C, 1 min; 58°C, 1 min; and 72°C, 4 min), followed by a final 10 min at 72°C. A 0.1-µL aliquot of the first PCR product was then used for the second PCR template, under the same conditions. Primer sequences are shown in Table IV. The PCR products were separated on a 1% (w/v) agarose gel, eluted with a Geneclean III kit (Q-Biogene, Carlsbad, CA), and sequenced using an ABI 3100 Genetic Analyzer (Applied Biosystems, Foster City, CA).

Sequence Data Analysis

The location of the flanking sequences was determined through BLASTN DNA homology searches using the NCBI GenBank (http://www.ncbi.nlm-.nih.gov/BLAST). When the sequence was not present in that database, the RiceGD database (http://btn.genomics.org.cn/rice) was used for identifying the contig number and insertion position. If a particular sequence had not yet been annotated in the public database, the sequence surrounding the insertion site was annotated using the Softberry program (http://www.softberry.com) and the GeneMark program (http://opal.biology.gatech.edu/GeneMark). The BLASTP program was used to search for homologous genes from the database. Distribution of T-DNA insertion sites in the chromosomes was estimated by counting the number of T-DNA inserts in the bacterial artificial chromosome or phage P1-derived artificial chromosome clones located on the chromosomes and then aligning the clones on the chromosomes. Inserts on overlapping regions between nearby clones were removed to avoid double counting. After collecting the sequences 100 bp upstream and 100 bp downstream from the insert sites with Textpad, the GC content of the region was calculated by domain selection using the MEGA2 program.

Functional classification of the predicted genes was performed with gene ontology programs (http://www.geneontology.org, http://rgp.dna.affrc.go.jp, and http://www.ebi.ac.uk/interpro).

LITERATURE CITED

Alonso JM, Stepanova AN, Leisse TJ, Kim CJ, Chen H, Shinn P, Stevenson DK, Zimmerman J, Barajas P, Cheuk R et al. (2003) Genome-wide insertional mutagenesis of *Arabidopsis thaliana*. Science **301:** 653–657

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

Bechtold N, Pelletier G (1998) In planta *Agrobacterium*-mediated transformation of adult *Arabidopsis thaliana* plants by vacuum infiltration. Methods Mol Biol **82:** 259–266

Bouché N, Bouchez D (2001) Arabidopsis gene knockout: phenotypes wanted. Curr Opin Plant Biol **4:** 111–117

Chen DH, Ronald PC (1999) A rapid DNA minipreparation method suitable for AFLP and other PCR applications. Plant Mol Biol Rep **17:** 53–57

Chin HG, Choe MS, Lee SH, Koo JC, Kim NY, Lee JJ, Oh BG, Yi GH, Kim SC, Choi HC et al. (1999) Molecular analysis of rice plants harboring an *Ac/Ds* transposable element-mediated gene trapping system. Plant J **19:** 615–623

Cowperthwaite M, Park W, Xu Z, Yan X, Maurais SC, Dooner HK (2002) Use of the transposon *Ac* as a gene-searching engine in the maize genome. Plant Cell **14:** 713–726

de Buck S, de Wilde C, van Montagu M, Depicker A (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

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

Feldmann KA (1991) T-DNA insertion mutagenesis in Arabidopsis: mutational spectrum. Plant J **1:** 70–82

Feng Q, Zhang Y, Hao P, Wang S, Fu G, Huang Y, Li Y, Zhu J, Liu Y, Hu X et al. (2002) Sequence and analysis of rice chromosome 4. Nature **420:** 316–320

Galbiati M, Moreno MA, Nadzan G, Zourelidou M, Dellaporta SL (2000) Large-scale T-DNA mutagenesis in Arabidopsis for functional genomic analysis. Funct Integr Genomics **1:** 25–34

Goff SA, Ricke D, Lan TH, Presting G, Wang R, Dunn M, Glazebrook J, Sessions A, Oeller P, Varma H et al. (2002) A draft sequence of the rice genome (*Oryza sativa L.* ssp. japonica). Science **296:** 92–100

Hiei Y, Ohta S, Komari T, Kumashiro 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

Hirochika H (2001) Contribution of *Tos17* retrotransposon to rice functional genomics. Curr Opin Plant Biol **4:** 118–122

Hirochika H, Sugimoto K, Otsuki Y, Tsugawa H, Kanda M (1996) Retrotransposons of rice involved in mutations induced by tissue culture. Proc Natl Acad Sci USA **93:** 7783–7788

Hui EK, Wang PC, Lo SJ (1998) Strategies for cloning unknown cellular flanking DNA sequences from foreign integrants. Cell Mol Life Sci **54:** 1403–1411

Ito T, Motohashi R, Kuromori T, Mizukado S, Sakurai T, Kanahara H, Seki M, Shinozaki K (2002) A new resource of locally transposed *Dissociation* elements for screening gene-knockout lines in silico on the Arabidopsis genome. Plant Physiol **129:** 1695–1699

Jeon J-S, Lee S, Jung K-H, Jun S-H, Jeong D-H, Lee J, Kim C, Jang S, Yang K, Nam J et al. (2000) T-DNA insertional mutagenesis for functional genomics in rice. Plant J **22:** 561–570

Jeong D-H, An S, Kang H-G, Moon S, Han J-J, Park S, Lee HS, An K, An G (2002) T-DNA insertional mutagenesis for activation tagging in rice. Plant Physiol **130:** 1636–1644

Kikuchi K, Terauchi K, Wada M, Hirano HY (2003) The plant MITE mPing is mobilized in anther culture. Nature **421:** 167–170

Kim S-R, Lee J, Jun S–H, Park S, Kang H-G, Kwon S, An G (2003) Transgene structures in T-DNA-inserted rice plants. Plant Mol Biol **52:** 761–773

Koncz C, Mayerhofer R, Koncz-Kalman Z, Nawrath C, Reiss B, Redei GP, Schell J (1990) Isolation of a gene encoding a novel chloroplast protein by T-DNA tagging in *Arabidopsis thaliana*. EMBO J **9:** 1337–1346

Koncz C, Németh K, Rédei GP, Schell J (1992) T-DNA insertional mutagenesis in Arabidopsis. Plant Mol Biol **20:** 963–976

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

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

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

Kumar A, Hirochika H (2001) Applications of retrotransposons as genetic tools in plant biology. Trends Plant Sci **6:** 127–134

Kumar S, Fladung M (2000) Transgene repeats in aspen: molecular characterisation suggests simultaneous integration of independent T-DNAs into receptive hotspots in the host genome. Mol Gen Genet **264:** 20–28

Liu YG, Whittier RF (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

Mathur J, Szabados L, Schaefer S, Grunenberg B, Lossow A, Jonas-Straube E, Schell JAZ, Koncz-Kalman Z (1998) Gene identification with sequenced T-DNA tags generated by transformation of Arabidopsis cell suspension. Plant J **13:** 707–716

Meissner RC, Jin H, Cominelli E, Denekamp M, Fuertes A, Greco R, Kranz HD, Penfield S, Petroni K, Urzainqui A (1999) Function search in a large transcription factor gene family in Arabidopsis: assessing the potential of reverse genetics to identify insertional mutations in r2r3 *myb* genes. Plant Cell **11:** 1827–1840

Ortega D, Raynal M, Laudie M, Llauro C, Cooke R, Devic M, Genestier S, Picard G, Abad P, Contard P et al. (2002) Flanking sequence tags in *Arabidopsis thaliana* T-DNA insertion lines: a pilot study. C R Biol **325:** 773–780

Pan X, Liu H, Clarke J, Jones J, Bevan M, Stein L (2003) ATIDB: *Arabidopsis thaliana* insertion database. Nucleic Acids Res **31:** 1245–1251

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

Parinov S, Sundaresan V (2000) Functional genomics in Arabidopsis: large-scale insertional mutagenesis complements the genome sequencing project. Curr Opin Biotechnol **11:** 157–161

Raina S, Mahalingam R, Chen F, Fedoroff N (2002) A collection of sequenced and mapped Ds transposon insertion sites in *Arabidopsis thaliana*. Plant Mol Biol **50:** 93–110

Rice Chromosome 10 Sequencing Consortium (2003) In-depth view of structure, activity, and evolution of rice chromosome 10. Science **300:** 1566–1569

Ríos G, Lossow A, Hertel B (2002) Rapid identification of Arabidopsis insertion mutants by nonradioactive detection of T-DNA tagged genes. Plant J **32:** 243–253

Sasaki T, Matsumoto T, Yamamoto K, Sakata K, Baba T, Katayose Y, Wu J, Niimura Y, Cheng Z, Nagamura Y et al. (2002) The genome sequence and structure of rice chromosome 1. Nature **420:** 312–316

Sessions A, Burke E, Presting G, Aux G, McElver J, Patton D, Dietrich B, Ho P, Bacwaden J, Ko C et al. (2002) A high-throughput Arabidopsis reverse genetics system. Plant Cell **14:** 2985–2994

Spertini D, Beliveau C, Bellemare G (1999) Screening of transgenic plants by amplification of unknown genomic DNA flanking T-DNA. Biotechniques **27:** 308–314

Speulman E, Metz P, van Arkel G, te Lintel Hekkert B, Steikema WJ, Pereira A (1999) A two component *Enhancer-Inhibitor* transposon mutagenesis system for functional analysis of the Arabidopsis genome. Plant Cell **11:** 1853–1866

Springer PS (2000) Gene traps: tool for plant development and genomics. Plant Cell **12:** 1007–1020

Sundaresan V, Springer P, Volpe T, Haward S, Jones JD, Dean C, Ma H, Martienssen R (1995) Patterns of gene action in plant development revealed by enhancer trap transposable element. Genes Dev **9:** 1797–1810

Szabados L, Kovacs I, Oberschall A, Abraham E, Kerekes I, Zsigmond L, Nagy R, Alvarado M, Krasovskaja I, Gal M et al. (2002) Distribution of 1000 sequenced T-DNA tags in the Arabidopsis genome. Plant J **32:** 233–242

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

Walbot V (2000) Saturation mutagenesis using maize transposons. Curr Opin Plant Biol **3:** 103–107

Wisman F, Hartmann U, Sagasser M, Baumann E, Palme K, Hathlbrock K, Saedler H, Weisshaar B (1998) Knock-out mutants from an En-1 mutagenized *Arabidopsis thaliana* population generate phenylpropanoid biosynthesis phenotypes. Proc Natl Acad Sci USA **95:** 12432–12437

Wolters AA, Trindade LM, Jacobsen E, Visser RGF (1998) Fluorescence *in situ* hybridization on extended DNA fibres as a tool to analyse complex T-DNA loci in potato. Plant J **13:** 837–847

Yamazaki M, Tsugawa H, Miyao A, Yano M, Wu J, Yamamoto Y, Matsumoto T, Sasaki T, Hirochika H (2001) The rice retrotransposon *Tos17* prefers low copy sequences as integration targets. Mol Gen Genet **265:** 336–344

Yephremov A, Saedler H (2000) Technical advance: display and isolation of transposon-flanking sequences starting from genomic DNA or RNA. Plant J **5:** 495–505

Yu J, Hu S, Wang J, Wong GK, Li S, Liu B, Deng Y, Dai L, Zhou Y, Zhang X et al. (2002) A draft sequence of the rice genome (*Oryza sativa* L. ssp. indica) Science **296:** 79–92