The Plant Journal (2006) **45**, 123–132 doi: 10.1111/j.1365-313X.2005.02610.x

TECHNICAL ADVANCE

Generation of a flanking sequence-tag database for activation-tagging lines in japonica rice

Dong-Hoon Jeong[1,†], Suyoung An[1,†], Sunhee Park[1], Hong-Gyu Kang[1], Gi-Gyeong Park[1], Sung-Ryul Kim[1], Jayeon Sim[1], Young-Ock Kim[1], Min-Kyung Kim[2], Seong-Ryong Kim[2], Joowon Kim[3], Moonsoo Shin[3], Mooyoung Jung[3] and Gynheung An[1,4,*]

[1]*Department of Life Science and National Research Laboratory of Plant Functional Genomics, Pohang University of Science and Technology (POSTECH), Pohang 790-784, Republic of Korea,*

[2]*Department of Life Science, Sogang University, Seoul 121-742, Republic of Korea,*

[3]*Department of Industrial and Management Engineering, POSTECH, and*

[4]*Functional Genomics Center, POSTECH, Pohang 790-784, Republic of Korea*

Received 17 June 2005; revised 1 September 2005; accepted 9 September 2005.

[*]For correspondence (fax +82 54 279 0659; e-mail genean@postech.ac.kr).

[†]These authors contributed equally to this work.

Summary

We have generated 47 932 T-DNA tag lines in japonica rice using activation-tagging vectors that contain tetramerized 35S enhancer sequences. To facilitate use of those lines, we isolated the genomic sequences flanking the inserted T-DNA via inverse polymerase chain reaction. For most of the lines, we performed four sets of amplifications using two different restriction enzymes toward both directions. In analyzing 41 234 lines, we obtained 27 621 flanking sequence tags (FSTs), among which 12 505 were integrated into genic regions and 15 116 into intergenic regions. Mapping of the FSTs on chromosomes revealed that T-DNA integration frequency was generally proportional to chromosome size. However, T-DNA insertions were non-uniformly distributed on each chromosome: higher at the distal ends and lower in regions close to the centromeres. In addition, several regions showed extreme peaks and valleys of insertion frequency, suggesting hot and cold spots for T-DNA integration. The density of insertion events was somewhat correlated with expressed, rather than predicted, gene density along each chromosome. Analyses of expression patterns near the inserted enhancer showed that at least half the test lines displayed greater expression of the tagged genes. Whereas in most of the increased lines expression patterns after activation were similar to those in the wild type, thereby maintaining the endogenous patterns, the remaining lines showed changes in expression in the activation tagged lines. In this case, ectopic expression was most frequently observed in mature leaves. Currently, the database can be searched with the gene locus number or location on the chromosome at http://www.postech.ac.kr/life/pfg/risd. On request, seeds of the T_1 or T_2 plants will be provided to the scientific community.

Keywords: activation tagging, flanking sequence tag, rice, T-DNA.

Introduction

Since the release of the rice genome sequence, the most significant challenge has been the large-scale identification of gene functions (Feng *et al.*, 2002; Goff *et al.*, 2002; Sasaki *et al.*, 2002; Yu *et al.*, 2002). Recently, approximately 370 Mb, or >97% of the genome, have been assembled as reference molecules with the release of the 'build 3.0 pseudomolecules' by the International Rice Genome Sequencing Project (Sasaki *et al.*, 2005). Using these non-overlapping genome sequences as templates for annotation, 57 888 genes now have been predicted by the annotation team of The Institute for Genomic Research (TIGR). In addition, the rice cDNA project has generated sequence data for 175 642 full-length cDNAs clustered into 28 469 non-redundant clones (Kikuchi *et al.*, 2003). These data, available through

the Knowledge-based Oryza Molecular biological Encyclopedia (KOME), facilitate gene prediction to make this information more valuable. As expected, these recent successes are also accelerating the need for functional genomics in this genus.

Various methods have been applied to generate loss-of-function mutations, including the use of ethyl methanesulfonate, fast-neutron treatment, antisense and RNA interference technology, and insertion mutations by a transposable element or T-DNA (Hirochika *et al.*, 2004; Jeon *et al.*, 2000; Kolesnik *et al.*, 2004; Miki and Shimamoto, 2004; Miyao *et al.*, 2003). One limitation to these approaches is that they rarely uncover function when the genes are either redundant or essential for early embryo or gametophyte development. Functional redundancy in most eukaryotic genes provides a significant obstacle to the assignment of gene function (Normandy and Bartel, 1999). Among the numerous approaches that have emerged to overcome these problems, the enhanced expression of genes that provide gain-of-function phenotypes has proved to be a productive identification strategy.

The first direct method for performing gain-of-function genetics in plants utilized the enhancer element from the cauliflower mosaic virus (CaMV) 35S gene (Odell *et al.*, 1985). T-DNA vectors containing four copies of this element were used successfully for generating activation-tagging lines and mediating transcriptional activation of nearby genes in Arabidopsis (Weigel *et al.*, 2000). For example, the mechanism for auxin biosynthesis, consisting of multiple pathways, was elucidated by this activation-tagging approach (Zhao *et al.*, 2001). This method has also been widely used in mutant screening to uncover enhancers or suppressors of given mutations (Li *et al.*, 2001, 2002). Although activation tagging has been applied predominantly to gene mining in Arabidopsis, this technology is now being deployed in diverse plant species such as petunia (Zubko *et al.*, 2002); tomato (Mathews *et al.*, 2003); poplar (Busov *et al.*, 2003); Madagascar periwinkle (van der Fits and Memelink, 2000); and rice (Jeong *et al.*, 2002).

The CaMV 35S enhancers, used for most activation tagging, function both upstream and downstream of a gene, in either orientation, and at a considerable distance from the coding regions. Furthermore, in some activation-tagging lines of Arabidopsis or rice, those enhancers cause greater endogenous expression rather than ectopic expression (Jeong *et al.*, 2002; Neff *et al.*, 1999; Weigel *et al.*, 2000). In these cases, identified phenotypes are more likely to reflect the endogenous function of a given tagged gene. Researchers have also developed an alternative gain-of-function approach, which relies on either a tissue-specific promoter to mis-express a gene in a particular tissue, or an inducible promoter to overexpress a gene at a specific time and under certain conditions (Matsuhara *et al.*, 2000; Zuo *et al.*, 2002).

Despite the usefulness of a phenotype-driven genetic approach, it is somewhat inconvenient for high-throughput screening of rice mutations. First, mutant screening of plants requires more effort because of their larger size and longer life cycles. Second, the phenotypic alterations observed in a T-DNA tagged line are not necessarily due to insertional mutation events but, instead, to the transposition of endogenous mobile elements such as *Tos17* (Miyao *et al.*, 2003). Finally, tissue culture often causes point mutations as well as small deletions or insertions.

These difficulties can be overcome through reverse genetics, in which a database for T-DNA insertion sites is first established and then used for functional analysis of the T-DNA-tagged genes. Although large-scale application of this strategy requires considerable effort (Parinov and Sundaresan, 2000), this database can easily be shared with other scientists, facilitating the distribution of mutant materials and analysis of gene functioning (An *et al.*, 2005). In Arabidopsis, >88 000 independent T-DNA insertion site sequences have been isolated from approximately 127 706 T_1 plants, resulting in the identification of mutations in more than 21 700 of the approximately 29 454 predicted genes (Alonso *et al.*, 2003). Here we report the generation of nearly 50 000 activation-tagging lines and 27 621 insertion-site sequences in rice.

Results

Generation of activation-tagging lines and isolation of tag-end sequences

We previously reported the generation of 13 450 activation-tagging lines of japonica rice using binary vector pGA2715 (Jeong *et al.*, 2002). We have now developed another activation-tagging vector, pGA2772, which is a modified version of pGA2715 containing the pUC18 vector backbone. This new vector is useful for retrieving T-DNA flanking regions if routine PCR methods fail to identify them. Using the *Agrobacterium*-mediated transformation method, we have established an additional 23 009 lines by pGA2715, plus 11 473 lines by pGA2772. Altogether, we have now generated 47 932 activation tag lines in rice.

To facilitate the use of these tagged lines, we isolated genomic sequences flanking the inserted T-DNA via inverse PCR (An *et al.*, 2003). Cutting the genomic DNA with restriction enzymes and using self-circularization yielded a molecule that could be PCR-amplified with primers located in the T-DNA. We designed the primer sets to amplify the genomic sequences flanking either the left or right border of T-DNA. For most lines we performed four sets of amplifications using two different restriction enzymes towards both directions.

By analyzing 31 100 lines of pGA2715 and 10 134 lines of pGA2772, we obtained 22 114 and 5507 flanking sequence

Table 1 Distribution of T-DNA insertions in genic and intergenic regions

Location of T-DNA insertions	pGA2715	pGA2772	Total (%)
Genic	10 017	2488	12 505 (45.3)
5′ UTR (300 bp upstream)	1749	459	2208 (8.0)
Coding exon	3248	820	4068 (14.7)
Intron	3790	884	4674 (16.9)
3′ UTR (300 bp downstream)	1230	325	1555 (5.6)
Intergenic	12 097	3019	15 116 (54.7)
Total	22 114	5507	27 621 (100.0)

Table 2 Distribution of predicted and expressed genes and T-DNA insertions in rice chromosomes

Chromosome	Size [Mb (%)]	Predicted genes [*n* (%)]	Expressed genes [*n* (%)]	T-DNA insertions [*n* (%)]
1	43.2 (11.7)	6905 (11.9)	1853 (12.2)	4126 (14.9)
2	35.9 (9.7)	5422 (9.4)	1774 (11.7)	3177 (11.5)
3	36.3 (9.8)	5986 (10.3)	2067 (13.6)	3642 (13.2)
4	35.0 (9.4)	5534 (9.6)	1370 (9.0)	2505 (9.1)
5	29.7 (8.0)	4636 (8.0)	1319 (8.7)	2065 (7.5)
6	31.2 (8.4)	4837 (8.4)	1313 (8.7)	2137 (7.7)
7	29.7 (8.0)	4635 (8.0)	1147 (7.6)	2015 (7.3)
8	28.3 (7.6)	4326 (7.5)	984 (6.5)	1721 (6.2)
9	22.7 (6.1)	3409 (5.9)	790 (5.2)	1544 (5.6)
10	22.7 (6.1)	3743 (6.5)	816 (5.4)	1565 (5.7)
11	28.4 (7.7)	4286 (7.4)	876 (5.8)	1562 (5.7)
12	27.5 (7.4)	4169 (7.2)	857 (5.7)	1562 (5.7)
Total	370.6 (100)	57 888 (100)	15 166 (100)	27 621 (100)

tags (FSTs), respectively (Supplemental data 1). The isolated FSTs were analyzed by the BLASTN homology search program, using the rice genome sequence database version 3.0 in TIGR. Of the total 27 621 insertions, 12 505 (45.3%) of the T-DNAs were integrated into genic regions and 15 116 (54.7%) were integrated into intergenic regions (Table 1). We considered the 300-bp flanking sequences outside the start ATG and stop codon to be untranslated regions of the genic region (An *et al.*, 2003; Szabados *et al.*, 2002). Our results are similar to those reported previously for non-activation T-DNA tagging lines (An *et al.*, 2003; Chen *et al.*, 2003; Sallaud *et al.*, 2004).

Distribution of T-DNA insertions

Mapping of the 27 621 FSTs revealed that T-DNA integration frequency was generally proportional to chromosome size (Table 2). Insertion was most frequent on the largest, chromosome 1, which also had the greatest number of predicted genes. Chromosomes 9 and 10 were the smallest, and carried the fewest T-DNA insertions.

We found non-uniform distribution of T-DNA insertions when their numbers were plotted per 500-kb interval over the length of each of the 12 chromosomes (Figure 1). Insertion frequency was higher at the distal ends and lower in regions close to the centromeres. In addition, several regions showed extreme peaks and valleys of frequency, suggesting hot and cold spots for T-DNA integration along each chromosome. These results are largely similar to those previously reported with Arabidopsis and rice (Alonso *et al.*, 2003; An *et al.*, 2003; Chen *et al.*, 2003; Sallaud *et al.*, 2004).

To examine whether this bias was due to unequal distribution of genic regions, we downloaded 57 888 predicted genes and 15 166 expressed genes (with expressed sequence tag and/or full-length cDNA evidence from predicted genes) from the TIGR rice genome annotation database version 3.0. Density of T-DNA insertion events was somewhat correlated to the expressed rather than the predicted gene density along each chromosome (Figure 1).

Again, our analysis of 27 621 FSTs revealed that about 45% of the T-DNAs were inserted into the genic region, and about 55% into the intergenic regions (Table 1). Among the 12 505 T-DNA located within the former, some were inserted into the same genes. Therefore removing the multiple alleles within the same gene resulted in the identification of T-DNA knockouts in 9911 (17.1%) predicted genes. When we examined the expressed genes, 4216 (27.8%) had T-DNA inserts (Table 3). Analysis of T-DNA insertions in the intergenic regions showed that 11 309 of the independent predicted genes had the insertion in either the 5′ or 3′ regions. Likewise, 4403 of the expressed genes contained the T-DNA in the flanking regions. T-DNA insertions into intergenic regions were classified by the intergenic regions of predicted or expressed genes adjacent to the left border containing the 35S enhancer.

We also analyzed the *Tos17*-tagged genes, using publicly available information (http://tos.nias.affrc.go.jp). Examination of 14 681 *Tos17* insertion sequences showed that 3380 predicted and 1408 expressed genes were tagged by the element. Among these *Tos17*-tagged genes, 1251 predicted and 552 expressed genes were also tagged by T-DNA. Consequently, 12 040 predicted and 5072 expressed genes were tagged by T-DNA or *Tos17*. This result demonstrates that the chance of finding an insertional mutation in a given gene is higher from the T-DNA insertional database established in this laboratory.

To examine a genome-wide correlation between distribution of T-DNA insertion and genic region, we plotted the numbers of predicted genes in the 500-kb window against the number of tagged genes in the same window (Figure 2a). The correlation coefficiency (r) was 0.34, indicating little relationship between insert distribution and the predicted gene. A similarly low value was obtained between the distribution of intergenic T-DNA insertions and predicted genes (Figure 2b). In contrast, we observed a high correlation between tagged and expressed genes (Figure 2c), as well as between intergenic tags and expressed genes.

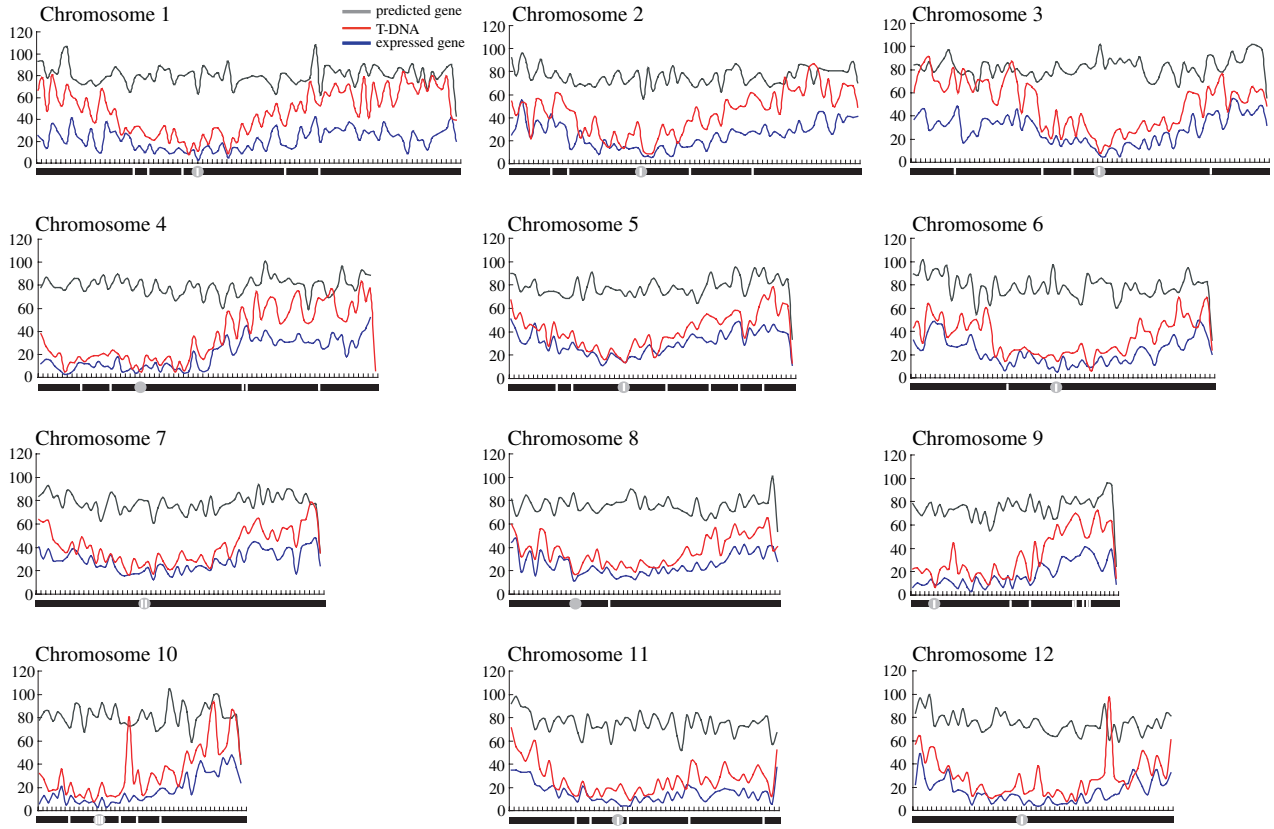

Figure 1. Distribution of T-DNA insertions and expressed and predicted genes along rice chromosomes, divided into 500-kb windows.
Numbers of T-DNA insertions (red), expressed genes (blue) and predicted genes (black) plotted for each window. Centromeric regions, gray circles; positions of physical gaps, white bars.

Table 3 Genes tagged by T-DNA or *Tos17*

	Genic regions		Intergenic regions	
	Predicted gene [*n* (%)]	Expressed gene [*n* (%)]	Predicted gene [*n* (%)]	Expressed gene [*n* (%)]
T-DNA	9911 (17.1)	4216 (27.8)	11 309 (19.5)	4403 (29)
Tos17	3380 (5.8)	1408 (9.3)	1783 (3.1)	705 (4.5)
Total	12 040 (20.8)	5072 (33.4)	12 520 (21.6)	4832 (31.9)

A total of 27 621 T-DNA insertions and 14 681 *Tos17* insertions were analyzed. Numbers of predicted and expressed genes in the rice genome were 57 888 and 15 166, respectively. T-DNA insertions into intergenic regions were classified by the intergenic regions of predicted or expressed genes adjacent to the left border containing the 35S enhancer. In the case of *Tos17* insertions into intergenic regions, the genes adjacent to the 3′ LTR of the insertion element were presented.

Therefore this genome-wide comparison strongly suggests that T-DNA insertion prefers the region where expressed genes are clustered.

We also analyzed *Tos17* distribution by the same method (Figure 2a–d), and found a low correlation efficiency between *Tos17* and predicted genes. This value was higher with expressed genes, although not as high as that obtained with T-DNA, suggesting that T-DNA insertion is distributed more randomly in the rice genome. These analyses confirmed the preference of T-DNA and *Tos17* insertions into expressed genes and the results obtained by other groups (Miyao *et al.*, 2003; Sallaud *et al.*, 2004).

Distribution of T-DNA insertions into intergenic regions

Because the tagging vectors pGA2715 and pGA2772 contain multimerized 35S enhancers in the T-DNA, tagging lines transformed with these vectors can be used not only for insertional tagging, but also for activation tagging (Jeong *et al.*, 2002). To investigate how many such lines are candidates for activation of nearby genes, we examined the distributions of intergenic lengths for 57 888 predicted genes. Here distribution displayed a pattern inclined toward shorter lengths (Figure 3a). Among the predicted genes, 32 381 (55.9%) intergenic regions were <3.0 kb long, which was the average length of those regions. We also examined the 15 116 T-DNA insertions located in the intergenic regions (Figure 3b), and found that their distribution displayed a

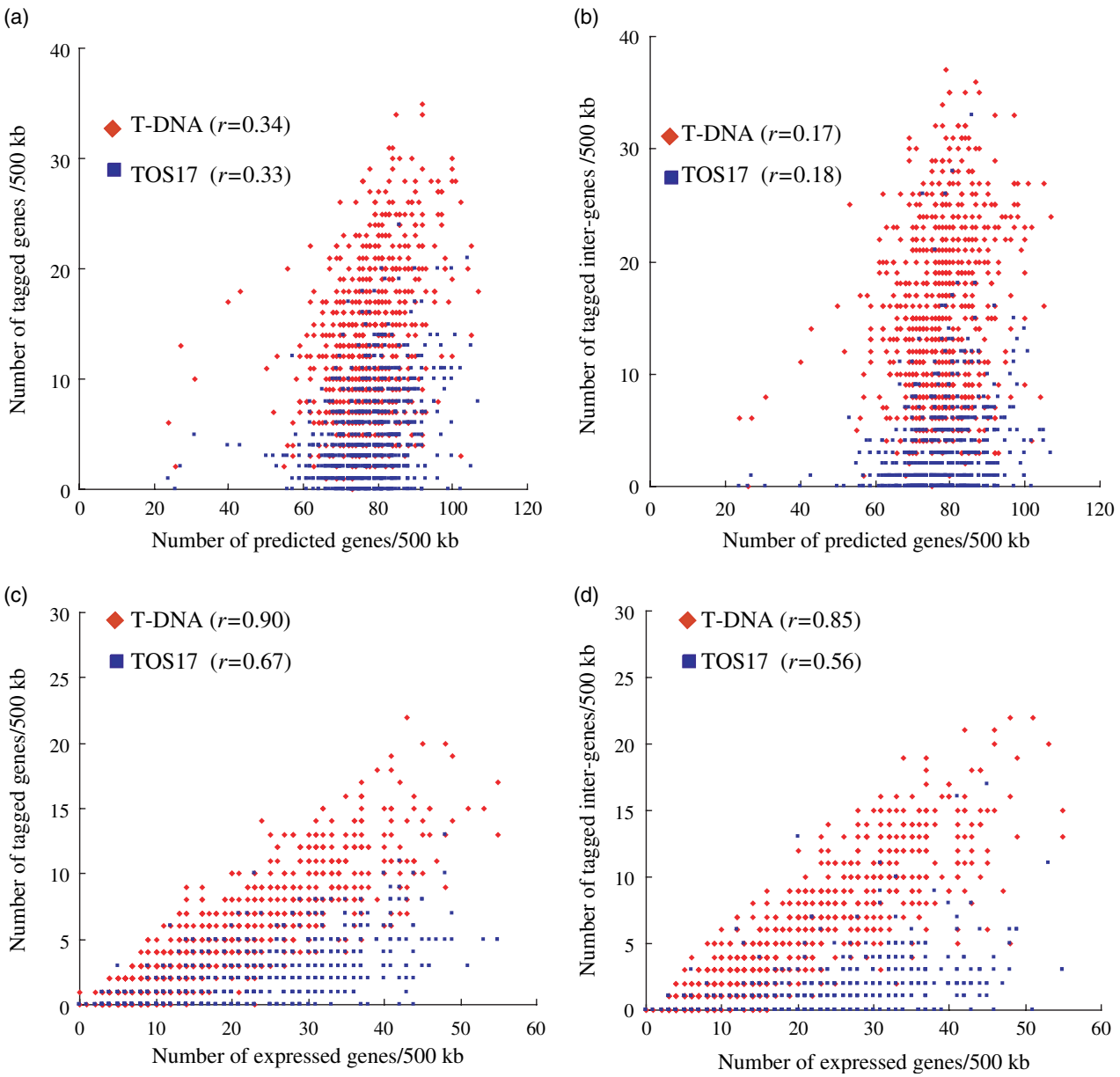

Figure 2. Distribution of genes tagged by T-DNA or *Tos17* compared with distributions of predicted and expressed genes.
Scatter plots comparing distribution of genes tagged by T-DNA or *Tos17* with distributions of predicted genes (a, b) or expressed genes (c, d) within each 500-kb window along rice chromosomes. Taggings by T-DNA or *Tos17* into genic (a, c) or intergenic (b, d) regions are presented separately. Number of genes tagged by T-DNA or *Tos17* per window is plotted by red diamonds or blue rectangles, respectively. Correlation coefficient (*r*) of each comparison is represented.

bell-shaped pattern with an average length of 5.5 kb in those tagged regions. The probability of finding an insertion in the intergenic regions was about 50% when the length was between 3 and 4 kb (Figure 3c). This frequency rose to up to 80% when the length was increased.

We also analyzed the distribution of T-DNA insertions from the start ATG and stop codons of the next genes. Here 14 548 sites were located within 5 kb upstream from the start ATG, while 12 221 sites were found within 5 kb downstream from the stop codons (Figure 4). 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.

Analysis of activation-tagging patterns

To monitor how activation tagging might perturb gene expression, we randomly selected insertion lines with T-DNA in their intergenic regions. Expression patterns of nearby genes, closest to the tetramerized 35S enhancer

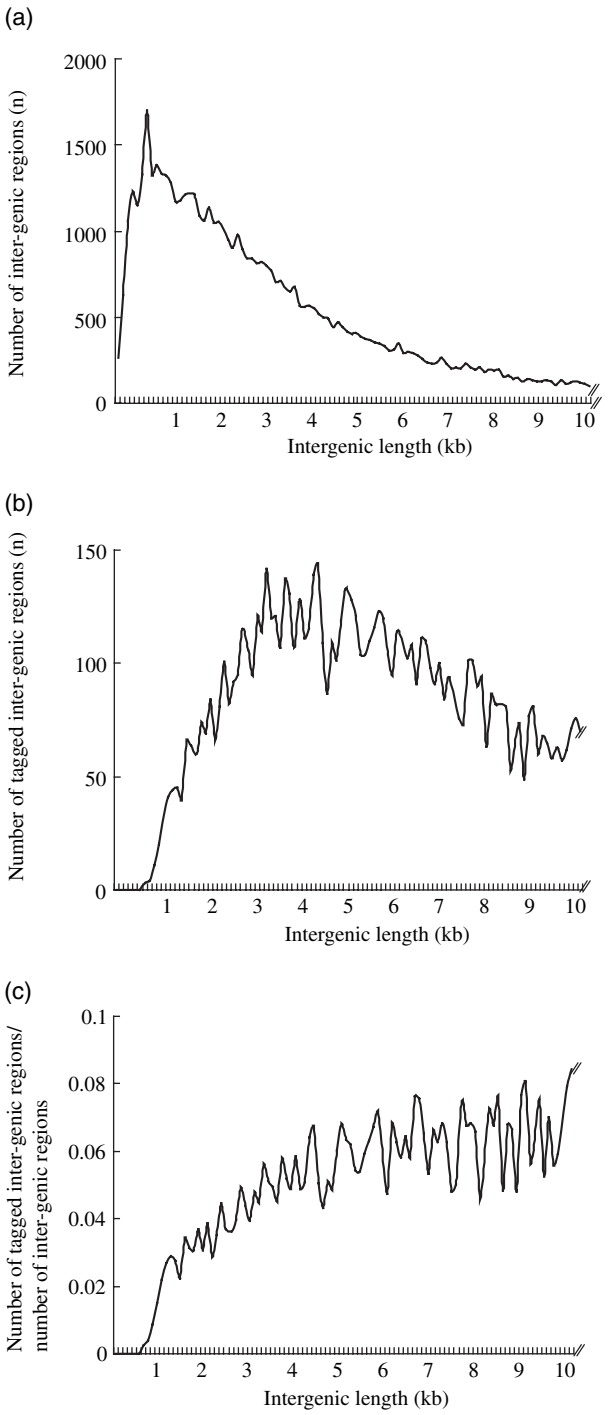

Figure 3. Frequency of T-DNA insertions into intergenic regions.
(a) Numbers of intergenic regions in the entire rice genome plotted against lengths of intergenic regions; (b) distribution of T-DNA-tagged intergenic regions; (c) frequency of T-DNA insertions into intergenic regions.

sequences at the T-DNA left border, were studied via semi-quantitative RT-PCR using gene-specific primers. Levels of expression were measured in the roots and shoots of seedlings, as well as in mature leaves and panicles from

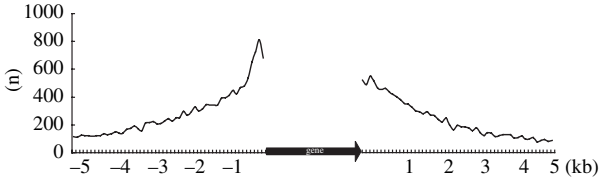

Figure 4. Distribution of T-DNA insertions around start ATG and stop codons.
Data were obtained from 14 548 insertion sites located within 5 kb upstream from ATG, and from 12 221 sites located within 5 kb downstream from stop codons.

tagged lines and wild-type plants. At least half the test lines (52.7%; 59/112) displayed greater expression of the tagged genes (data not shown). In most of the increased lines (69.5%; 41/59), patterns after activation were similar to those in the wild type, maintaining their endogenous expression patterns (Figure 5, lines 2A-00420, 2A-10364, 1B-02632). In the remaining lines (30.5%; 18/59), the patterns changed in the activation-tagged lines, with ectopic expression being most frequently observed in the mature leaves (Figure 5, lines 1B-04331, 2A-10667, 1B-02413).

No good relationship was found between frequency of activation and distance from the 35S enhancers to the gene (data not shown). Similarly, we observed no correlation between degree of activation and distance (Figure 5). For example, strong enhancement was noted in line 1B-02413, where the 35S enhancers were located 10.7 kb upstream from the start codon of the *Os11g40350* gene. Enhancement was observed both upstream and downstream of the tagged genes (Figure 5).

Discussion

Generation of activation-tagging lines to provide wide variety of mutants

Nearly 2000 traits, including both single Mendelian loci/genes and quantitative trait loci, have been identified in rice (Kurata *et al.*, 2005). However, the number of mutants is much smaller than the number of predicted genes found during recent genome sequencing of that species. This might be mainly due to redundancy, because most of its genes are members of one family (Goff *et al.*, 2002; Sasaki *et al.*, 2002; Yu *et al.*, 2002). Therefore classical loss-of-function mutants have limitations when one attempts to elucidate gene functioning. To provide a wide variety of mutants, we have generated binary vectors that contain multimerized 35S enhancer elements immediately next to the left border. Similar vectors have been utilized successfully to produce activation-tagging populations in Arabidopsis and other dicot species (Busov *et al.*, 2003; van der Fits and Memelink, 2000; Li *et al.*, 2001, 2002; Mathews *et al.*,

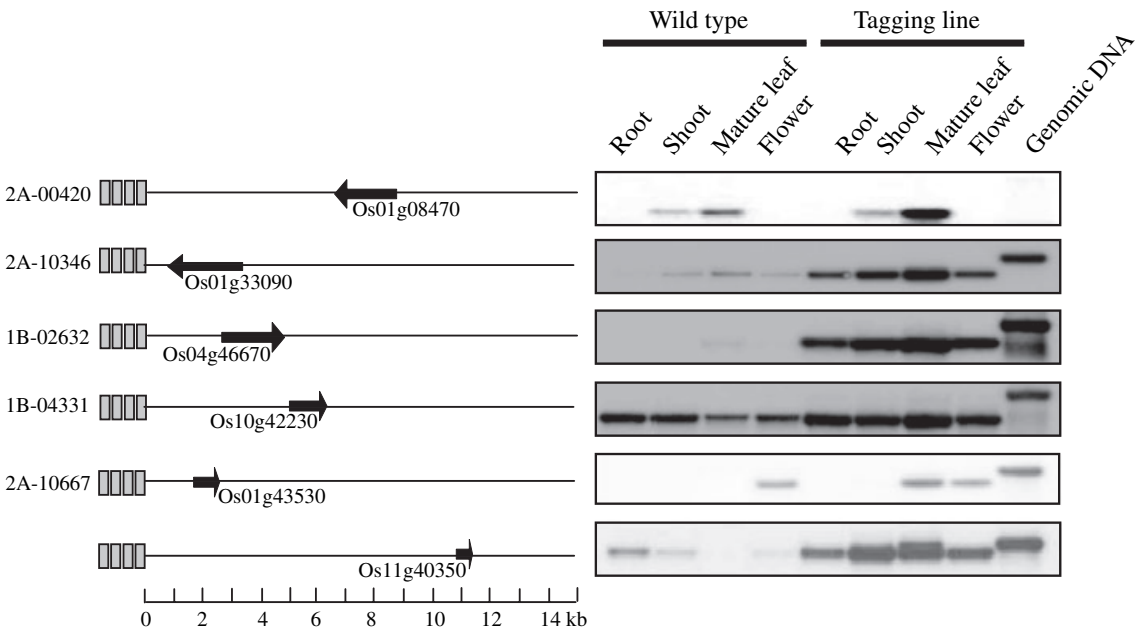

Figure 5. Examples of activation-tagging patterns.
35S enhancer elements in relation to nearby genes are represented schematically on the left. Gray bars, CaMV 35S enhancers. Nearby genes indicated by arrows pointing in direction of transcription, designated with locus ID number represented in TIGR database. Right, expression patterns analyzed by semi-quantitative RT-PCR, using RNA samples prepared from seedling roots and shoots, and mature leaves and flowers of wild-type control and tagging-line plants. RT-PCR products were blotted and hybridized with P^{32}-labeled gene-specific probes. Genomic DNA served as template to verify contamination of genomic DNA during RT-PCR experiments. Amplification products of genomic DNA are distinguished from those of cDNA by different sizes because amplified regions contain an intron. Because the intron in *Os01g8470* is large, genomic DNA was not amplified.

2003; Zubko *et al.*, 2002). In those tagging lines, expression of the gene near the enhancer elements is enhanced, causing dominant gain-of-function phenotypes. Thus activation-tagging mutagenesis presents a phenotypic spectrum that is different from phenotypes generated by loss-of-function mutations. In this study we created nearly 50 000 activation-tagging lines in japonica rice. Because each line contains an average of 1.4 insertion loci (Jeon *et al.*, 2000), approximately 70 000 T-DNA inserts were made. This population should be a valuable resource for researchers in the plant community for functional analysis of rice genes.

Establishing database of T-DNA insertion sites for reverse-genetics approaches

To utilize the mutant population efficiently and facilitate sharing of resources within the scientific community, we determined the genomic sequences flanking the T-DNA insertions, using inverse PCR because that method provides an average of one band after amplification. This is an important factor because this approach does not require gel-separation of PCR bands followed by elution and purification. We directly sequenced the PCR product, thereby analyzing a large number of samples with only limited resources.

From the analysis of 41 234 lines, we obtained 27 621 FSTs. Considering that each tagging line carries an average of 1.4 T-DNA insertion loci, up to 57 700 FSTs might have been retrieved from the analysis. This indicates that our efficiency was approximately 48%. One of the difficulties in isolating FST is a high GC content at the tag sites, which inhibits PCR amplification. Another problem is repetitive sequences that lack the enzyme sites used for inverse PCR analyses. Currently, we are improving the efficiency rate by employing a high-GC buffer. We also plan to determine the flanking sequences of the pGA2772-tagged lines by the plasmid-rescue method.

We have now generated a database with FSTs obtained from the activation-tagging lines. It can be searched with the gene locus number or location on the chromosome at http://www.postech.ac.kr/life/pfg/risd. We are in the process of improving the search engine so the database can be searched with DNA sequences or key words. On request, 15 seeds of the T_1 plants can be made available when >100 seeds are present in the seed stock. If that number is <100, we provide the seeds after their amplification.

More than a quarter of the expressed genes are tagged

Approximately 45% of FSTs are present in the genic region. Our analysis showed that 17.1% of the predicted genes had at least one T-DNA insertion there. In contrast, 27.8% of the expressed genes were tagged by T-DNA. Similarly, higher efficiency in the expressed genes was obtained with the

intergenic insertions. Therefore it seems that T-DNA prefers highly expressed genes compared with those that are poorly expressed. Alternatively, the predicted gene number may be overestimated. Among the 57 888 genes predicted by the TIGR rice genome database, 14 196 genes are transposable elements, which are considered transcriptionally silent. These elements are usually clustered near the centromeres where T-DNA insertion frequency is low. However, even if those transposable elements are not considered, the frequency of FSTs in the predicted genes is still lower than that in the expressed genes. This suggests that the total number of functional genes in rice may be much smaller than the number of annotated genes.

Gene expression in mature leaves is more preferentially enhanced by activation tagging

Transcript levels for genes near the 35S enhancer were increased in about half the activation-tagging lines. In the remaining half, levels and patterns of expression were not significantly changed. In most of the increased lines, expression patterns were conserved but overall enhancement was observed. In these cases the genes were more preferentially expressed in mature leaves. In the remaining lines, expression patterns of the tagged genes were not mature-leaf preferential, but became leaf preferential after activation tagging. Therefore it appears that the 35S enhancer elements increase expression of nearby genes more preferentially in mature leaves, especially when the tagged gene is originally expressed preferentially in other organs. Because only half our tagged genes were enhanced by the tagging vector, there might be a silencing mechanism that inactivates the action of 35S enhancer elements. One possible mechanism is methylation, which is induced by multiple-copy T-DNA integration (Chalfun-Junior *et al.*, 2003).

Experimental procedures

Generation of activation-tagging lines in rice

Scutellum calli derived from *Oryza sativa* var. *japonica* cv. Dongjin or Hwayoung were transformed with *Agrobacterium* that contained the activation-tagging vector pGA2715 or pGA2772 by the procedure reported previously (Jeong *et al.*, 2002; Lee *et al.*, 1999). The pGA2772 vector was constructed by inserting the tetramerized 35S enhancer sequence and the pUC18 vector into the *Xho*I site of pGA2707 (An *et al.*, 2003). In pGA2772, the 35S enhancer elements are located next to the T-DNA left border, and the promoterless GUS reporter gene is located next to the right border. Because pGA2772 contains the origin of replication and the beta lactamase gene, the sequences adjacent to the left border can be retrieved through plasmid rescue. Transgenic plants were grown in the glasshouse at a minimum night temperature of 20°C and with a day length of at least 14 h, supplemented with artificial lights.

Isolation of sequences flanking T-DNA

Preparation of tissue samples and extraction of genomic DNA were performed as described by An *et al.* (2003). To isolate flanking sequences of T-DNA, we used the inverse PCR method described previously by An *et al.* (2003), with the following modifications: 1 μg genomic DNA was digested with 10 U restriction enzymes in 50 μl for 10 h. After the enzymes were heat-inactivated, the cut DNAs were ligated at 8°C for 16 h, using 1 U of T4 DNA ligase (Roche, Mannheim, Germany). Nested PCR was conducted 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, 0.3× Band Doctor solution (Solgent, Daejeon, Korea), and 0.1 U EF *Taq* polymerase (Solgent, Daejeon, Korea). 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), then 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 Supplementary Table S1. Approximately 1/50 of the second PCR product was directly sequenced using Applied Biosystems Big Dye Terminator 3.0 chemistry and then processed on the Applied Biosystems 3730 DNA sequencer.

Analysis of sequences flanking T-DNA

Positions of the T-DNA insertions were deduced from the results of homology alignment of each FST against the TIGR assembly of rice chromosomes 1–12 (sequence accessions AP008207 to AP008218) using the BLASTN program. The vector sequence was masked prior to this homology search. We considered only the alignments with scores higher than 100. The position of the first matching nucleotide between FST and the genomic sequence was used to establish the most likely insertion site. For statistical analyses, identical FSTs that appeared at least twice in the different lines were counted only once. When more than one non-overlapping FST was obtained from a single T-DNA line, each FST was considered an independent insertion. The rice genome annotation data for predicted and expressed genes were downloaded from the TIGR rice genome annotation database (http://www.tigr.org/tdb/e2k1/osa1/index.shtml).

Semi-quantitative RT-PCR analysis

Total RNAs were isolated from each tissue type by an RNA-isolation kit (Tri Reagent, MRC, Inc., Cincinnati, OH, USA). The first-strand cDNA was synthesized, to serve as template, from 2 μg DNAsel-treated total RNA, using M-MLV reverse transcriptase (Promega, Madison, WI, USA). Gene-specific primers were designed for each gene (Supplementary Table S2). After PCR amplification, the products were separated on a 1.5% agarose gel, blotted onto a nylon membrane, and hybridized with ^{32}P-labeled probes. All experiments were performed at least twice to confirm results.

Acknowledgements

We thank Hea-kyung Jung, Hee-Jung Woo, Hyunsook Lee, Jeonghi Lee, Kyoungmi Han, Kyung-Mi Kim, Kyunghwa Jung, Jung-Hwa Yu, Junghe Hur, Ji-sung Han and Sung-Il Ryu for assisting in sequencing the T-DNA insertion sites; In-Soon Park and Kyungsook An for generation of the T-DNA insertional lines; Yoonja Cho, Sangsun An, Soonhee Kim and Soonok Kim for harvest and management of

transgenic seeds; Changduk Jung and Shi-In Kim for growing the transgenic plants, and Priscilla Licht for critical proofreading of the manuscript. This work was funded in part by grants from the Crop Functional Genomic Center, the 21st Century Frontier Program (CG-1111); from the Biogreen 21 Program, Rural Development Administration; and from POSCO.

Supplementary Material

The following supplementary material is available for this article online:

Figure S1. Map of pGA2772.

RB and LB (gray bar), right and left borders of T-DNA, respectively. I, *OsTubA1* intron 2 carrying three putative splicing acceptor and donor sites; GUS, β-glucuronidase; Tn, *nos* terminator; Tt, *OsTubA1* terminator; *hph*, hygromycin phosphotransferase gene; OsTubA1-1, the first intron of *OsTubA1*; pUC18, fragment of pUC18 vector; E, enhancer element of CaMV 35S promoter. The sequence of pGA2772 has been submitted to Genbank (accession number DQ151884).

Table S1 Primers used for iPCR

Table S2 Primers used for RT-PCR

Data S1 Database of 27,621 tag end sequences

This material is available as part of the online article from http://www.blackwell-synergy.com

References

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.

An, G., Lee, S., Kim, S.H. and Kim, S.R. (2005) Molecular genetics using T-DNA in rice. *Plant Cell Physiol.* **46**, 14–22.

Busov, V.B., Meilan, R., Pearce, D.W., Ma, C., Rood, S.B. and Strauss, S.H. (2003) Activation tagging of a dominant gibberellin catabolism gene (GA2-oxidase) from poplar that regulates tree stature. *Plant Physiol.* **132**, 1283–1291.

Chalfun-Junior, A., Mes, J.J., Mlynarova, L., Aarts, M.G. and Angenent, G.C. (2003) Low frequency of T-DNA based activation tagging in *Arabidopsis* is correlated with methylation of CaMV 35S enhancer sequences. *FEBS Lett.* **555**, 459–463.

Chen, S., Jin, W., Wang, M., Zhang, F., Zhou, J., Jia, Q., Wu, Y., Liu, F. and Wu, P. (2003) Distribution and characterization of over 1000 T-DNA tags in rice genome. *Plant J.* **36**, 105–113.

Feng, Q., Zhang, Y., Hao, P. *et al.* (2002) Sequence and analysis of rice chromosome 4. *Nature*, **420**, 316–320.

van der Fits, L. and Memelink, J. (2000) ORCA3, a jasmonate-responsive transcriptional regulator of plant primary and secondary metabolism. *Science*, **289**, 295–297.

Goff, S.A., Ricke, D., Lan, T.H. *et al.* (2002) A draft sequence of the rice genome (*Oryza sativa* L. ssp. *japonica*). *Science*, **296**, 92–100.

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

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

Kikuchi, S., Satoh, K., Nagata, T. *et al.* (2003) Collection, mapping, and annotation of over 28 000 cDNA clones from japonica rice. *Science*, **301**, 376–379.

Kolesnik, T., Szeverenyi, I., Bachmann, D., Kumar, C.S., Jiang, S., Ramamoorthy, R., Cai, M., Ma, Z.G., Sundaresan, V. and Ramachandran, S. (2004) Establishing an efficient Ac/Ds tagging system in rice: large-scale analysis of Ds flanking sequences. *Plant J.* **37**, 301–314.

Kurata, N., Miyoshi, K., Nonomura, K., Yamazaki, Y. and Ito, Y. (2005) Rice mutants and genes related to organ development, morphogenesis and physiological traits. *Plant Cell Physiol.* **46**, 48–62.

Lee, S.C., Jeon, J.S., Jung, K.H. and An, G. (1999) Binary vector for efficient transformation of rice. *J. Plant Biol.* **42**, 310–316.

Li, J., Lease, K.A., Tax, F.E. and Walker, J.C. (2001) BRS1, a serine carboxypeptidase, regulates BRI1 signaling in *Arabidopsis thaliana*. *Proc. Natl Acad. Sci. USA* **98**, 5916–5921.

Li, J., Wen, J., Lease, K.A., Doke, J.T., Tax, F.E. and Walker, J.C. (2002) BAK1, an *Arabidopsis* LRR receptor-like protein kinase, interacts with BRI1 and modulates brassinosteroid signaling. *Cell*, **110**, 213–222.

Mathews, H., Clendennen, S.K., Caldwell, C.G. *et al.* (2003) Activation tagging in tomato identifies a transcriptional regulator of anthocyanin biosynthesis, modification, and transport. *Plant Cell*, **15**, 1689–1703.

Matsuhara, S., Jingu, F., Takahashi, T. and Komeda, Y. (2000) Heat-shock tagging: a simple method for expression and isolation of plant genome DNA flanked by T-DNA insertions. *Plant J.* **22**, 79–86.

Miki, D. and Shimamoto, K. (2004) Simple RNAi vectors for stable and transient suppression of gene function in rice. *Plant Cell Physiol.* **45**, 490–495.

Miyao, A., Tanaka, K., Murata, K., Sawaki, H., Takeda, S., Abe, K., Shinozuka, Y., Onosato, K. and Hirochika, H. (2003) Target site specificity of the Tos17 retrotransposon shows a preference for insertion within genes and against insertion in retrotransposon-rich regions of the genome. *Plant Cell*, **15**, 1771–1780.

Neff, M.M., Nguyen, S.M., Malancharuvil, E.J. *et al.* (1999) BAS1: a gene regulating brassinosteroid levels and light responsiveness in *Arabidopsis*. *Proc. Natl Acad. Sci. USA* **96**, 15316–15323.

Normandy, J. and Bartel, B. (1999) Redundancy as a way of life – IAA metabolism. *Curr. Opin. Plant Biol.* **2**, 207–213.

Odell, J.T., Nagy, F. and Chua, N.H. (1985) Identification of DNA sequences required for activity of the cauliflower mosaic virus 35S promoter. *Nature*, **313**, 810–812.

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

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.

Sasaki, T., Matsumoto, T., Yamamoto, K. *et al.* (2002) The genome sequence and structure of rice chromosome 1. *Nature*, **420**, 312–316.

Sasaki, T., Matsumoto, T., Antonio, B.A. and Nagamura, Y. (2005) From mapping to sequencing, post-sequencing and beyond. *Plant Cell Physiol.* **46**, 3–13.

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.

Weigel, D., Ahn, J.H., Blazquez, M.A. *et al.* (2000) Activation tagging in *Arabidopsis*. *Plant Physiol.* **122**, 1003–1013.

Yu, J., Hu, S., Wang, J. *et al.* (2002) A draft sequence of the rice genome (*Oryza sativa* L. ssp. *indica*). *Science*, **296**, 79–92.

Zhao, Y., Christensen, S.K., Fankhauser, C., Cashman, J.R., Cohen, J.D., Weigel, D. and Chory, J. (2001) A role for flavin monooxygenase-like enzymes in auxin biosynthesis. *Science*, **291**, 306–309.

Zubko, E., Adams, C.J., Machaekova, I., Malbeck, J., Scollan, C. and Meyer, P. (2002) Activation tagging identifies a gene from *Petunia hybrida* responsible for the production of active cytokinins in plants. *Plant J.* **29**, 797–808.

Zuo, J., Niu, Q.W., Frugis, G. and Chua, N.H. (2002) The *WUSCHEL* gene promotes vegetative-to-embryonic transition in *Arabidopsis*. *Plant J.* **30**, 349–359.