The Plant Journal (2004) **39**, 450–464 doi: 10.1111/j.1365-313X.2004.02145.x

TECHNICAL ADVANCE

High throughput T-DNA insertion mutagenesis in rice: a first step towards *in silico* reverse genetics

Christophe Sallaud[1,†], Céline Gay[2], Pierre Larmande[1], Martine Bès[1], Pietro Piffanelli[1], Benoit Piégu[3], Gaétan Droc[1], Farid Regad[1], Emmanuelle Bourgeois[1], Donaldo Meynard[1], Christophe Périn[1], Xavier Sabau[1], Alain Ghesquière[4], Jean Christophe Glaszmann[1], Michel Delseny[3] and Emmanuel Guiderdoni[1,*]

[1]*Biotrop Program, Cirad-Amis and*

[2]*Inra-Ensam, UMR1096 PIA, Avenue Agropolis, F-34398 Montpellier Cedex 5, France,*

[3]*Laboratoire Génome et Développement des Plantes, UMR5096, CNRS/UP, 52, avenue de Villeneuve, F-66860, Perpignan Cedex, France, and*

[4]*Genetrop, Ird, UMR 5096, BP64501, F-34394 Montpellier Cedex 5, France*

Received 19 March 2004; accepted 10 May 2004.

[*]For correspondence (fax 33 4 67 61 56 05; e-mail guiderdoni@cirad.fr).

[†]Current address: LIBROPHYT, Centre de Cadarache, Bâtiment 185, DEVM 13108 St Paul-Lez-Durance, France.

Summary

A library of 29 482 T-DNA enhancer trap lines has been generated in rice cv. Nipponbare. The regions flanking the T-DNA left border from the first 12 707 primary transformants were systematically isolated by adapter anchor PCR and sequenced. A survey of the 7480 genomic sequences larger than 30 bp (average length 250 bp), representing 56.4% of the total readable sequences and matching the rice bacterial artificial chromosome/phage artificial chromosome (BAC/PAC) sequences assembled in pseudomolecules allowed the assigning of 6645 (88.8%) T-DNA insertion sites to at least one position in the rice genome of cv. Nipponbare. T-DNA insertions appear to be rather randomly distributed over the 12 rice chromosomes, with a slightly higher insertion frequency in chromosomes 1, 2, 3 and 6. The distribution of 723 independent T-DNA insertions along the chromosome 1 pseudomolecule did not differ significantly from that of the predicted coding sequences in exhibiting a lower insertion density around the centromere region and a higher density in the subtelomeric regions where the gene density is higher. Further establishment of density graphs of T-DNA inserts along the recently released 12 rice pseudomolecules confirmed this non-uniform chromosome distribution. T-DNA appeared less prone to hot spots and cold spots of integration when compared with those revealed by a concurrent assignment of the *Tos17* retrotransposon flanking sequences deposited in the National Center for Biotechnology Information (NCBI). T-DNA inserts rarely integrated into repetitive sequences. Based on the predicted gene annotation of chromosome 1, preferential insertion within the first 250 bp from the putative ATG start codon has been observed. Using 4 kb of sequences surrounding the insertion points, 62% of the sequences showed significant similarity to gene encoding known proteins (*E*-value <1.00 e^{-05}). To illustrate the *in silico* reverse genetic approach, identification of 83 T-DNA insertions within genes coding for transcription factors (TF) is presented. Based both on the estimated number of members of several large TF gene families (e.g. Myb, WRKY, HD-ZIP, Zinc-finger) and on the frequency of insertions in chromosome 1 predicted genes, we could extrapolate that 7–10% of the rice gene complement is already tagged by T-DNA insertion in the 6116 independent transformant population. This large resource is of high significance while assisting studies unravelling gene function in rice and cereals, notably through *in silico* reverse genetics.

Keywords: flanking sequence tags, rice, T-DNA mutagenesis, reverse genetics.

Introduction

Rice (*Oryza sativa* L.) has emerged as a model plant for cereal genomics particularly because of the size of its genome (430 Mb), the smallest among graminaceous crops, and the availability of large genetic and molecular resources. These include the complete physical map of the *japonica* cv. *Nipponbare* (Chen *et al.*, 2002), a collection of over 200 000 rice expressed sequence tags (ESTs) in the public database, dbEST (NCBI), a set of 28 469 full-length cDNA (The Rice Full-Length cDNA Consortium, 2003, http://www.cdna01.dna.affrc.go.jp/cDNA/) and a Uniset gene database (OsGI: http://www.tigr.org/tdb/tgi/ogi/). Rice is also highly amenable to *Agrobacterium*-mediated transformation (Delseny *et al.*, 2001 for a review; Hiei *et al.*, 1994; Sallaud *et al.*, 2003). The colinearity of the rice genome with those of other agronomically important monocots was also established (Bennetzen *et al.*, 1998; Devos and Gale, 2000). Recent publication of two drafts of the rice genome (Goff *et al.*, 2002; Yu *et al.*, 2002) and of an elaborated draft of the cv. Nipponbare (http://rgp.dna.affrc.go.jp/IRGSP/index.html) including the near-complete sequences of three chromosomes (Feng *et al.*, 2002; Sasaki *et al.*, 2002; The Rice Chromosome 10 Sequencing Consortium, 2003) has opened a new area for plant biologists (Delseny, 2003). Comparative analysis of the rice and *Arabidopsis* genome sequences has revealed interesting features. In particular, 85% of predicted *Arabidopsis* genes have a homolog in rice whereas only 50% of predicted rice genes have a homolog in *Arabidopsis*. However, nearly 100% of the known genes in wheat, barley and maize have a homolog in rice, strengthening, if needed, its status as a model for unravelling gene function in cereals.

At present, the vast majority of new genes are identified only by their sequence. As a consequence, the function of a large number of genes is unknown or hypothetical. The development of high throughput methods to discover the biological function of these genes has become the next objective of many research groups. In plants, the complete sequencing of the *Arabidopsis* genome in the year 2000 promoted the development of functional genomic approaches and confirmed the importance of using model plants to address fundamental questions in plant biology (Pereira, 2000). Genetic approaches, which rely on the production of a large mutant population by the disruption or replacement of genes, followed by the study of the plant phenotypes, are powerful tools to determine precisely gene function. Whereas the forward genetic screen of such populations has allowed the discovery of the function of a large number of genes in *Arabidopsis*, the large amount of genomic sequence information now available has promoted the development of reverse genetic strategies which consist in identifying a mutant affected in a particular sequence within a mutant population and looking for an associated phenotype.

Depending on the organism studied, different tools are available for implementing such a strategy. Targeted disruption of specific genes by homologous recombination is routinely used in *Escherichia coli*, yeast, mice and *Physcomytrella patens* but is not yet applicable to higher plants despite recent advances (Hanin *et al.*, 2001), notably in rice (Terada *et al.*, 2002). Recent developments in post-transcriptional gene silencing such as RNA interference (RNAi; Fire *et al.*, 1998) and virus-induced gene silencing (VIGS; Baulcombe, 2000) have been effective in studying gene function. Knockouts of more than 4000 genes in *Caenorhabditis elegans* (Fraser *et al.*, 2000; Gonczy *et al.*, 2000) have been produced through RNAi-mediated silencing. Although RNAi is functional in plants, no large-scale silencing programme has been reported yet (Smith *et al.*, 2000; Waterhouse and Helliwell, 2003 for a review; Wesley *et al.*, 2001).

Insertion mutagenesis is based on the generation of a large population of mutants in which foreign DNA is randomly inserted into the genome. Transposable elements (TE) such as transposons or retrotransposons and T-DNA are routinely used for this purpose. Whereas several initiatives have demonstrated the potential of the maize *Ac/Ds* and *En/Spm* transposon systems in *Arabidopsis* (Parinov *et al.*, 1999; Raina *et al.*, 2002; Speulman *et al.*, 1999; Tissier *et al.*, 1999), T-DNA mutagenesis remains the preferred choice for plants for which an efficient genetic transformation procedure is available. This is demonstrated by the increasing number of *Arabidopsis* T-DNA insertion populations produced within the last few years (Bechtold *et al.*, 1993; Feldmann, 1991; Koncz *et al.*, 1992; Krysan *et al.*, 1999; Alonso *et al.*, 2003; Sessions *et al.*, 2002; Szabados *et al.*, 2002). Although an important investment is required initially to obtain the transformed lines, the advantage resides in the stability of the insertion through multiple generations. PCR-based reverse genetics strategies using DNA pools were first developed to identify a mutation in a particular gene (Bouchez and Höfte, 1998 for a review; Krysan *et al.*, 1996; McKinney *et al.*, 1995; Rios *et al.*, 2002; Young *et al.*, 2001). Although large numbers of genes have been isolated by this method, a more systematic approach, which takes advantage of the availability of the *Arabidopsis* genome sequence, is now being undertaken. Each T-DNA flanking sequence tag (FST) is isolated, sequenced and a search for a mutant in a particular gene is performed *in silico* through a WEB database interface (Ortega *et al.*, 2002; Samson *et al.*, 2002; http://www.evry.inra.fr/public/projects/bioinfo/flagdb.html; Sessions *et al.*, 2002; http://www.tmri.org/en/partnership/sail_collection.aspx; Alonso *et al.*, 2003; http://www.signal.salk.edu/cgi-bin/tdnaexpress; Rosso *et al.*, 2003; http://www.mpiz-koeln.mpg.de/GABI-Kat/ see http://www.arabidopsis.org/links/insertion.jsp for other links).

In rice, significant efforts to develop gene machines through genome-wide coverage with TE inserts have also been accomplished during the last decade. The autonomous maize *Ac* element was introduced into the rice genome in the early 1990s (Shimamoto *et al.*, 1993) where it proved to actively transpose and insert in genes with a three- to fourfold specificity, allowing the establishment of gene machines in cv. Toride 1 and Nipponbare (Enoki *et al.*, 1999; Greco *et al.*, 2001a,b). More sophisticated double-component *AcTpase/Ds* systems have also been recently engineered and have proved to be functional for generating large mutant collections in japonica rice cv. Dong-Jin (Chin *et al.*, 1999) and Nipponbare (Greco *et al.*, 2003; Kolesnik *et al.*, 2004; Upadhyaya *et al.*, 2002). In this latest report, 92% of 1811 unique *Ds* flanking sequences recovered from stabilized lines were mapped on a Nipponbare physical map revealing a low average frequency (6%) of transpositions closely linked to the donor sites, interchromosomal transposition preferences, and a 72% frequency of insertion of the *Ds* element in genic regions (Kolesnik *et al.*, 2004).

The endogenous retrotransposon *Tos 17* is another element, which has been reported for insertion mutagenesis in rice (Hirochika, 2001). Its transposition is induced by *in vitro* tissue culture and the element becomes inactive after plant regeneration, allowing the production of non-transgenic stable mutant plants (Hirochika *et al.*, 1996). Insertion of *Tos17* was found to occur preferentially in low-copy number sequences (Yamazaki *et al.*, 2001) whereas large-scale analyses of sequences at integration sites revealed a 3× insertion preference into genic regions versus intergenic regions as well as hot spots for integration (Miyao *et al.*, 2003). Analysis of the distribution of *Tos17* along chromosome 1 also suggested that integration of the element avoids retrotransposon-rich pericentromeric regions and favours gene-rich subtolemeric regions. A database of over 18 000 *Tos17* flanking sequences is available for similarity search through BLAST queries at http://www.pc7080.abr.affrc.go.jp/~miyao/pub/tos17/index.html.en.

T-DNAs carrying a gene trap or an activation tagging system have also been used for the generation of a large insertion mutant population (Jeong *et al.*, 2002, Jeon *et al.*, 2000; Wu *et al.*, 2003). The characterization of a large number of T-DNA insertion sites was reported only recently (An *et al.*, 2003; Chen *et al.*, 2003). These two studies consistently revealed preferential insertion into gene-rich regions, low frequency insertion in repetitive regions and comparable frequencies of integration in genic versus intergenic regions.

We previously described a high throughput transformation method in japonica rice, which appeared suitable to produce large T-DNA insertion mutant populations (Sallaud *et al.*, 2003). Since the year 2000, our laboratory has been involved in generating a population of 50 000 T-DNA rice lines of cv. Nipponbare in the framework of the French National Plant Genome Initiative 'Génoplante'. The production scheme includes the characterization of T-DNA flanking regions (FST) from DNA extracted from young leaves of primary transformants. In this paper, we describe the characterization of 7480 T-DNA inserts corresponding to 6116 independent lines that have been identified from a first set of 21 295 transformants. A total of 6645 T-DNA insertions (88.8%) on the genome of rice cv. *Nipponbare* were assigned unequivocally with most insertions found to occur at a unique position. The recent release of the 12 assembled chromosome pseudomolecules (http://rgp.dna.affrc.go.jp/IRGSP/index.html) and of more than 28 000 full-length cDNA sequences (The Rice Full-Length cDNA Consortium, 2003) allowed us to establish a genome-wide comparison of distribution of T-DNA insertions, *Tos17* insertions and FL cDNA sequences in rice. To investigate whether T-DNA insertions occur preferentially in genic versus intergenic regions, all insertions on chromosome 1 were annotated using published predicted data (Sasaki *et al.*, 2002) and complementary analyses. Using the reverse *in silico* approach, insertions occurring in the vicinity of genes encoding transcription factors (TF) were also evaluated. The number of lines that will be required to knockout all rice genes is discussed. This resource will be a valuable tool for plant researchers to determine the function of a large number of rice genes.

Results

Agrobacterium-*mediated production of enhancer trap lines*

A large number of T-DNA enhancer trap lines was generated using a high throughput *Agrobacterium*-mediated transformation procedure (Sallaud *et al.*, 2003). Seed embryo-derived embryogenic calli were co-cultured with *Agrobacterium* strains EHA105 or LBA4404 carrying the binary vector pC-4978 in five large-scale transformation experiments (Table S1). The pC-4978 T-DNA contains the *hpt* gene conferring resistance to hygromycin under the control of the Subterranean Clover Mosaic Virus pS4 promoter (Upadhyaya *et al.*, 2002) and of the rice *actin1* first intron region and the *gusA* coding sequence (CDS) fused to the first 46 bp of the 35S CaMV promoter located at the T-DNA right border, thereby creating a typical enhancer trap construct (Figure 1). Overall, 6820 co-cultured calli generated 29 482 primary transformants labelled with a barcode in a 2-year time span. Transformation efficiency averaged 4.3 primary transformants per co-cultured callus. No significant difference in terms of transformation efficiency was observed between the EHA105 and LBA4404 strains. Southern blot analyses of a subset of 400 pC-4978 primary transformants and selected progeny revealed that an average of 2.2 copies of the T-DNA have integrated at 1.4 locus per line, with no difference between bacterial strains used for co-culture (data not shown).

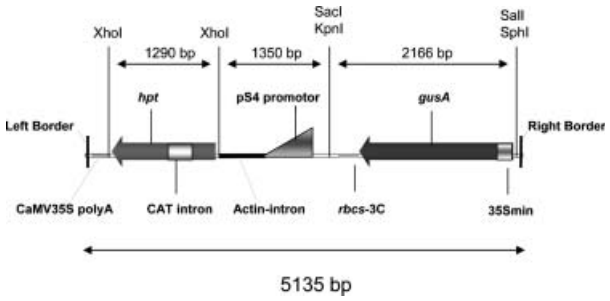

Figure 1. Schematic representation of the T-DNA enhancer trap construct of the binary vector pC-4978.

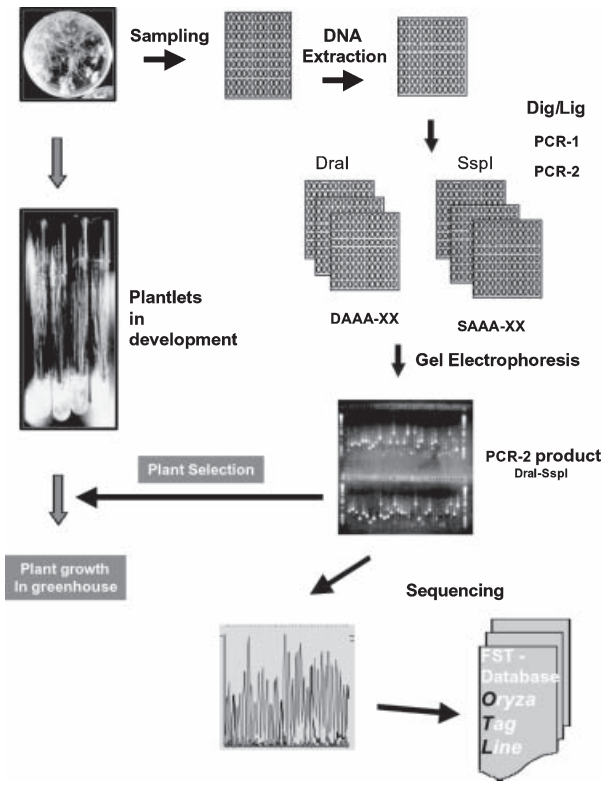

Figure 2. Flowchart of the overall strategy for high throughput isolation and annotation of flanking sequence tags.

Selection of primary transformants based on FST amplification

A prerequisite for developing an *in silico* reverse genetic strategy is the establishment of a high throughput procedure for characterization of T-DNA FST. Isolation of regions flanking the left border of T-DNA inserts was performed using an adapter-anchor PCR method (Siebert *et al.*, 1995), which relies on the use of a blunt-end restriction enzyme, a specific adapter and two steps of nested PCR amplification using primers specific to the adapter and the T-DNA. The method proved to be very efficient for the large-scale isolation of T-DNA flanking sequences in *Arabidopsis* (Balzergue *et al.*, 2001; Ortega *et al.*, 2002). Preliminary studies (Sallaud *et al.*, 2003) confirmed that isolation of FSTs from the right border was less efficient than at the left border because of a higher frequency of inverted T-DNA repeats involving two right borders (De Block and Debrouwer, 1991; De Neve *et al.*, 1997). We first investigated the potential cutting frequency of a range of blunt-end restriction enzymes through *in silico* analyses of the sequences of 10 randomly chosen BAC clones (see Methods). *Dra*I and *Ssp*I were identified as the enzymes potentially generating the highest number of 0–2 kb restriction fragments, a size compatible with further PCR amplification (data not shown). Second, to confirm that this *in silico* analysis reflects experimental results, FST isolation was performed on 192 primary transformants harbouring the pC-4978 T-DNA with a range of blunt-end enzymes (see Methods). The number of plants giving a PCR product was two- to fivefold higher with *Dra*I or *Ssp*I than with any of the other enzymes tested (data not shown).

The frequency of successful isolation of T-DNA flanking regions generally ranges from 60 to 80%, irrespective of the procedure used (inverse-PCR, adapter-anchor PCR, Thermal Asymmetric InterLaced-PCR), notably when a large number of samples is processed (Balzergue *et al.*, 2001; Szabados *et al.*, 2002). Considering (i) that the greenhouse space needed to grow the T-DNA primary transformants is limited for rice and (ii) the enormous task to further multiply and maintain and distribute seed stocks of several tens of thousands of lines, it appears preferable to apply a stringent and early selection of the mutant lines on the basis of a well-characterized FST. As a consequence, a general strategy to select T-DNA lines with a putative FST has been developed and is summarized below and illustrated in Figure 2: (i) DNA was extracted from tissues of young shoots of regenerating plantlets before their transfer to test tubes for further development; (ii) isolation of regions flanking the T-DNA left border at the integration site was performed separately from *Dra*I and *Ssp*I digests for each DNA sample; (iii) plants corresponding to tracks exhibiting a unique PCR2 product on agarose gels with either *Dra*I or *Ssp*I were selected and transferred to the greenhouse. This unique PCR2 product was then directly sequenced without any additional purification steps. Overall, 12 707 plants (60%) were selected following the analysis of 21 295 primary transformants in a year time frame. A unique PCR2 product was observed for 7557 (35%) and 8810 (41%) lines following DNA digestion with *Dra*I and *Ssp*I, respectively, whereas 3772 (17%) lines exhibited a unique product for both enzymes. From 14 876 *Dra*I and *Ssp*I PCR2 products sequenced with a T-DNA-specific primer (CAMB6), 13 254 (89%) produced a readable sequence (Table 1). The T-DNA footprint was observed in 12 378 sequences (93.4%) indicating that the amplification was specific. A rather large number of sequences corres-

Table 1 Summary of characterization of T-DNA flanking sequences

Type of sequences	Number of sequences	%*
Genomic sequences >30 bp	7480	56.43
Genomic sequences <30 bp	1956	14.76
T-DNA tandem	1887	14.24
Binary vector	1055	7.96
T-DNA footprint not found	876	6.61
Subtotal (readable sequences)	13 254	100
Bad sequences	1622	10.89
Total	14 876	

T-DNA footprint is detected by homology search with the T-DNA left border sequence using blastn program and deleted. FST size is then calculated to identify genomic sequences with a minimum length of 30 bp. Binary vector and T-DNA tandem repeats are identified by sequence homology.
*The percentage is calculated as the number of sequences over the number of readable sequences (subtotal).

ponded to either T-DNA sequences (14.24%), because of the integration of T-DNA tandem repeats into the host genome, or binary vector sequences (7.96%), which was shown to follow frequently during T-DNA transfer to the plant cell (Kononov *et al.*, 1997). Among the 9436 sequences left (71.2%), 1956 (14.76%) had a genomic sequence shorter than 30 bp. The average size of the remaining 7480 (56.4%) genomic sequences was 250 bp with more than half of them longer than 250 bp. Nevertheless, 972 (13%) of the 7480 FST sequences proved to be redundant because of parallel amplification from the border of the same T-DNA insert in both *Dra*I and *Ssp*I digests. Redundant sequences also resulted from the presence of a few clonal lines deriving from the same transformation event most likely resulting from the rare fragmentation of a hygromycin-resistant cell line and further subculturing of callus pieces. Contamination during DNA extraction or PCR experiments could also be another source of redundancy. The two latter cases represented 391 (6%) redundant sequences. Overall, 6116 independent tagged lines were eventually identified.

Distribution of T-DNA insertions on rice chromosomes

Assignation on the rice genome of T-DNA insertions having an FST sequence longer than 30 bp ($n = 7480$) was determined by similarity searches against Nipponbare BAC/PAC clones representing the almost complete genome sequence (i.e. 412 Mb of an estimated 430 Mb, http://www.rgp.dna.affrc.go.jp/cgi-bin/statusdb/irgsp-status.cgi). Stringent criteria were applied to guarantee a high level of confidence (see Methods). We found that a significant number of FSTs displayed misalignment within the first 20 bp whereas the rest of the sequence perfectly matched the rice genome sequence. This is a result of rearrangement of host DNA that frequently occurs at the T-DNA insertion point (Brunaud *et al.*, 2002; Cluster *et al.*, 1996; Ortega *et al.*, 2002). Therefore, we performed an additional blastn search after deletion of the first 20 bp of the FSTs. The FST size was limited to 250 bp for the analysis as we noticed that sequencing errors appear more frequently beyond that length. Moreover, 250 bp appeared sufficient to assign a T-DNA insertion position on the rice genome with confidence as discussed below. Using this procedure, nearly 15% of additional FSTs were assigned to rice BAC clones. Overall, 6645 of 7480 T-DNA insertions (88.8%) were assigned to at least one position on the rice genome (IRGSP release of April 2003) following exclusion of redundant hits generated at the same loci because of overlapping BAC clone sequences. The percentage of assigned insertions is very consistent with the estimated percentage of the rice genome sequence available (90–95%). This indicates that our procedure of identifying T-DNA flanking sequences is accurate. Distribution of the T-DNA insertions over the 12 rice chromosomes is shown in Table 2. The overall T-DNA insertion density for each chromosome averaged 18.5 insertions per Mb sequenced and ranged from 15.02 (chr.12) to 22.65 (chr. 1) insertions per Mb. Overall, four chromosomes (i.e. 1, 2, 3 and 6) exhibited an apparent higher T-DNA insertion density than the others. Following the same procedure, 90.78% of the 14 643 *Tos17* FSTs deposited at NCBI were successfully assigned to the rice pseudomolecules and the same four chromosomes also exhibited an apparent higher density of *Tos17* inserts. Similar analysis of the distribution of the full-length cDNA sequences (FL cDNA) revealed statistically significant higher and lower densities on chromosomes 1 and 2 and chromosomes 11 and 12, respectively.

Low frequency of T-DNA integration within repetitive sequences

Repetitive sequences such as TE and rDNA were found to represent more than 25% of the total rice genome sequence and are considered to be a major component of the intergenic regions (Goff *et al.*, 2002). We performed an analysis to evaluate the frequency of T-DNA insertions within repetitive sequences based on a blastn similarity search. We considered that if more than four putative locations were assigned for a given FST, the insertion most likely occurred in a repetitive region. Manual examination confirmed this hypothesis. Very few T-DNA insertions (<1%) were found to occur in repetitive sequences, although we consider that in the population, 17% of the sequences was less than 100 bp (Table 3) increasing the probability to match repetitive sequences. As a consequence, unequivocal assignation of the position of most T-DNA insertions on the rice genome was possible. To determine whether a bias for integration within repetitive sequences occurred, we evaluated the frequency of repetitive sequences in four

Table 2 Distribution of T-DNA insertions over the 12 rice chromosomes compared with those of *Tos17* insertions and FL cDNA

Chromosome nb	Chromosome size (Mb)	No. of insertions			Insertion density (per Mb)		
		T-DNA	*Tos17*	FL cDNA	T-DNA	*Tos17*	FL cDNA
1	43	974	1789	4252	22.65	41.6	98.88
2	35.6	732	1420	3367	20.56	39.89	94.58
3	35.1	766	1561	3857	21.82	44.47	109.89
4	34.5	569	1197	2609	16.49	34.7	75.62
5	28.5	524	988	2346	18.39	34.67	82.32
6	30	577	1298	2232	19.23	43.27	74.4
7	29.3	508	994	1985	17.34	33.92	67.75
8	27.7	461	879	1883	16.64	31.73	67.98
9	21.2	329	612	1428	15.52	28.87	67.36
10	22.5	355	719	1530	15.78	31.96	68
11	24.9	446	923	1323	17.91	37.07	53.13
12	26.9	404	914	1487	15.02	33.98	55.28
Total	359.2	6645	13 294	28 299	18.5	37.01	78.78

FSTs of T-DNA (our results) and *Tos17* (Miyao *et al.*, 2003) inserts and FL cDNA (The Rice Full-Length cDNA Consortium, 2003) sequences were assigned by similarity searches (blastn) on rice pseudomolecules (downloaded from ftp://ftp.tigr.org/pub/data/Eukaryotic_Projects/o_sativa/annotation_dbs/pseudomolecules/version_1.0/) according to the criteria described in Methods.
To determine whether the distribution over chromosomes was even, we calculated the theoretical numbers of insertions per Mb for each chromosome, which were tested against a Poisson law for a significant threshold $P < 0.05$ with $\mu =$ average density/Mb for all chromosomes. A significant bias was detected only for FL cDNA and chromosomes 1, 2, 3, 11 and 12.

Table 3 Occurrence of repetitive sequences among six different populations of sequences based on the number of hits on the rice genome

Types of sequences	Hit no.					Total sequences
	1	2	3	4	>5	
FST (all)	3699 (66.34)	1652 (29.63)	153 (2.74)	25 (0.45)	47 (0.84)	5576
100 bp >FST ≤250 bp	3056 (66.22)	1376 (29.82)	131 (2.84)	20 (0.43)	32 (0.69)	4615
Random BAC sequences (250 bp)	2229 (39.97)	2095 (37.57)	292 (5.24)	71 (1.27)	280 (5.02)	4967
Random Chr1 sequences (250 bp)	506 (58.16)	280 (32.18)	31 (3.56)	11 (1.26)	42 (4.83)	870
BAC ends (250 bp)	1284 (62.03)	516 (24.93)	80 (3.86)	27 (1.30)	163 (7.87)	2070
Tos17	2097 (70.89)	767 (25.93)	79 (2.67)	11 (0.37)	4 (0.14)	2958

A hit is identified by sequence similarity search (blastn) on rice BAC/PAC sequences, as previously described for T-DNA assignation (see Methods). A sequence is classified as repetitive when attributed a score higher than four hits. Percentage values are indicated in parentheses.

different populations of rice sequences having an equivalent size and using the same method: (i) 5000 BAC end sequences whose size was limited to 250 bp; (ii) 5000 random sequences of 250 bp extracted from rice BAC clones; (iii) 870 random sequences of 250 bp extracted from chromosome 1 BAC clones; and (iv) *Tos17* flanking sequence population limited to the first set of 4800 sequences deposited in GenBank size (A. Miyao and H. Hirochika, unpublished data), in order to use a comparable population. We observed that the frequency of T-DNA insertions within repetitive sequences is lower than those of rice genomic sequence populations. For the first three populations examined, this frequency was seven- to 11-fold higher than that observed for the FST population of equivalent size (Table 3). The results observed with *Tos17* confirmed a previous report indicating preferential integration of the element within low copy sequences (Miyao *et al.*, 2003; Yamazaki *et al.*, 2001). A manual annotation of all T-DNA insertions within repetitive sequences confirmed that most of them (43 out of 46) correspond to insertions within TE.

Distribution of T-DNA insertions on chromosome 1

We carried out a more detailed analysis of the T-DNA insertions assigned to rice chromosome 1, for which the sequence and annotation were nearly complete at the time of the study. The objective was to (i) compare the distribution of T-DNA insertions with the distribution of the CDS; (ii) check the possible occurrence of insertion hot spots; and (iii) evaluate the number of genes that have been tagged. In order to eliminate redundant sequences resulting from overlap between BAC clones, we created a pseudomolecule of chromosome 1 (Psmchr1) using the available physical map information (Chen *et al.*, 2002, see Methods). As a

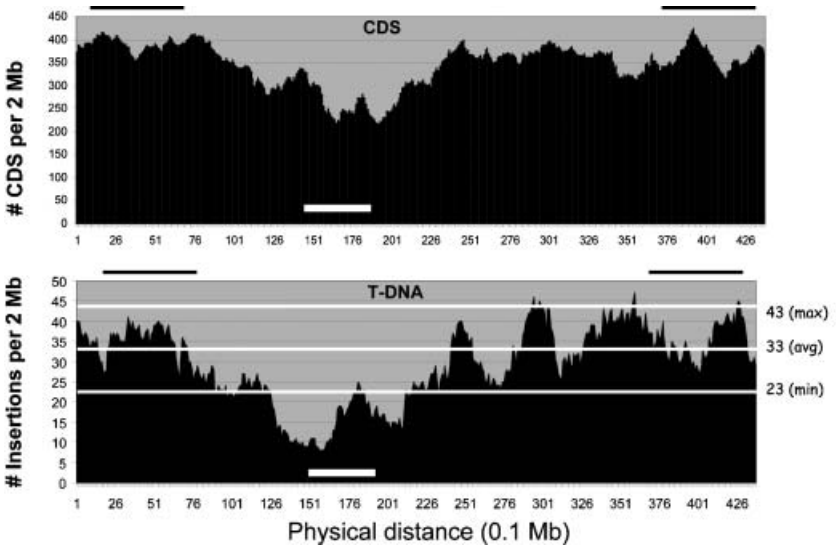

Figure 3. Graphical representation of CDS (upper panel) and T-DNA (lower panel) insertions over the pseudomolecule of chromosome 1. Densities of insertions and CDS were plotted for each 2 Mb with a sliding window of 100 kb. Values for the minimum and maximum T-DNA density probability according to statistical analysis (Poisson law, $P < 0.05$) are indicated on the right hand of the T-DNA density graph. Avg, average of T-DNA insertion density. Black and white bars indicate subtelomeric and centromeric regions, respectively.

result, 11 contigs representing 43.4 Mb were generated and assembled artificially as a pseudomolecule. The theoretical size of chromosome 1 (45.7 Mb) indicates that our analysis was performed with the almost complete sequence. Most gaps were localized in the centromeric region as well as in the 1 Mb of sequence localized in the upper part of the chromosome, which had not yet been sequenced. Overall, 723 T-DNA insertions were assigned on Psmchr1, including 668 (92.8%) to a unique location and 16 (2.2%) in repetitive sequences (i.e. more than four locations).

The distributions of T-DNA insertions and CDS on chromosome 1 were compared by generating a density graph (Figure 3). CDS were retrieved from BAC clones that had been annotated by the software prediction system RiceGAAS (Sakata *et al.*, 2002). T-DNA and CDS densities were analysed for each 2-Mb window with a sliding step of 100 kb. The density of T-DNA insertions showed a fivefold magnitude of variation along the chromosome, ranging from eight to 47 insertions per 2 Mb with an average of 29 insertions per 2 Mb. If one excludes the magnitude of variation and the as yet limited number of T-DNA insertions, the density distribution profile of T-DNA insertions on chromosome 1 was found to be highly consistent with that of CDS. In particular, the centromeric region exhibited a statistically significantly lower T-DNA insertion density, whereas a higher density was observed in the subtelomeric regions of each chromosome arm ($P < 0.05$) (Figure 3).

Comparison of the distribution of T-DNA and Tos17 *insertions and FL cDNA sequences along the 12 rice chromosomes*

We further established the density graphs of the 6645 T-DNA insertions along the 12 pseudomolecules in plotting the FST existing at each 200-kb interval with a sliding window of 10 kb. The distribution of T-DNA inserts was compared with those of the 13 294 *Tos17* insertions (Miyao *et al.*, 2003) and 28 299 FL cDNA sequences (The Rice Full-Length cDNA Consortium, 2003) successfully assigned using the same criteria (Figure 4). Overall, this analysis confirmed the high non-uniform chromosomal distribution of integration events observed in chromosome 1 and the results obtained by other groups (An *et al.*, 2003; Chen *et al.*, 2003). Both T-DNA and *Tos17* graphs followed the same distribution trend with a lower frequency of insertions in the pericentromeric regions and a higher density in subtelomeric regions. The density of T-DNA integration, as well as that of *Tos17*, followed the variation of gene density, reflected by FL cDNA distribution. The influence of the centromere was particularly visible on the short arms of chromosomes 4, 9 and 10, which also exhibited a lower FL cDNA density. These regions are known to be the most heterochromatic in pachytene chromosome observations using DAPI staining (Cheng *et al.*, 2001). Similarly, the three most euchromatic chromosomes identified in the latter report, that is, chromosomes 1, 2 and 3, were also those exhibiting a higher density in FL cDNA sequences and tended to harbour more T-DNA and *Tos17* insertions.

Although the population of *Tos17* FSTs was twice as large as that of T-DNA FSTs, it was obvious from the observation of density graphs (here voluntarily framed at a maximum threshold of 50 inserts per interval) that *Tos17* was more prone to hot and cold spots of integration than T-DNA. In particular, when compared with T-DNA, restriction to *Tos17* integration appears to occur in wider chromosomal segments spanning from the centromere. This may reflect the constitution of these regions, which are rich in TE-related sequences that tend to be avoided by *Tos17* (Miyao *et al.*, 2003).

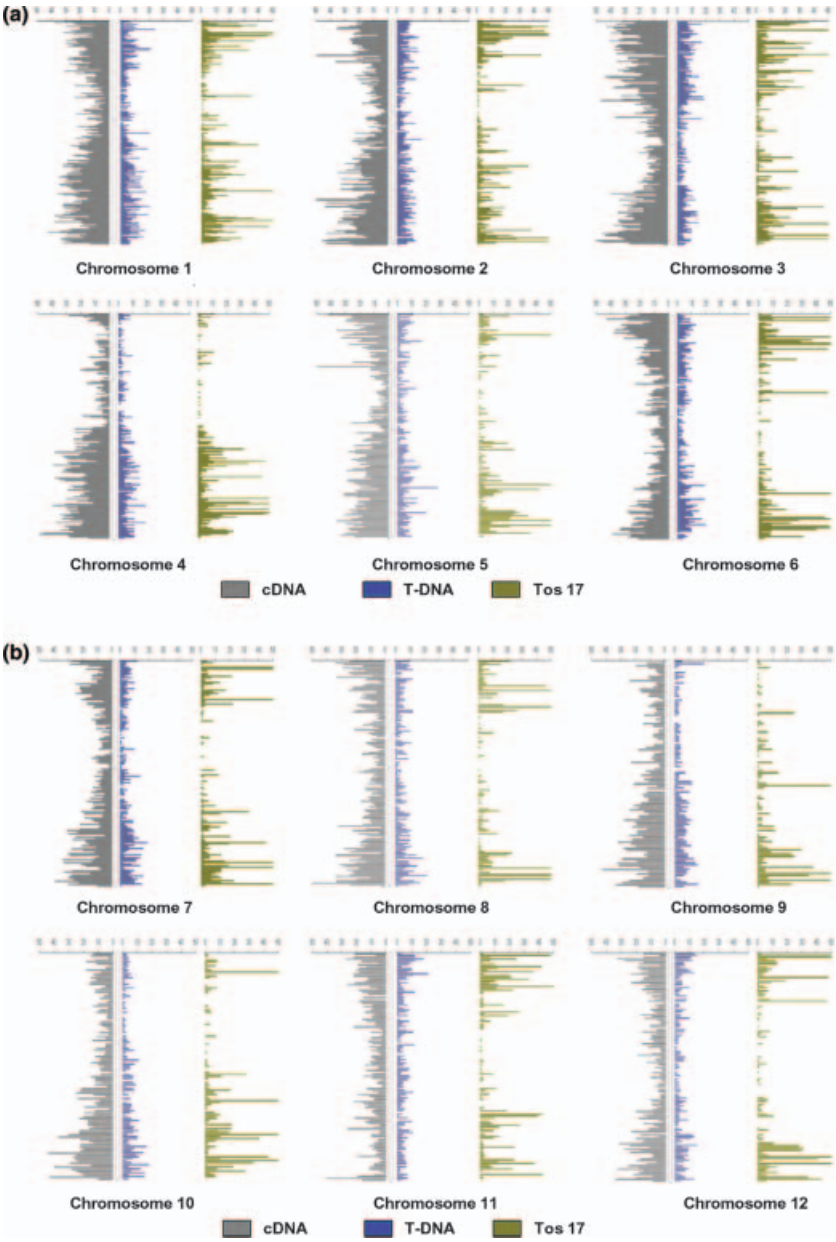

Figure 4. Comparative distributions of T-DNA (this work) and *Tos 17* (Miyao *et al.*, 2003) insertions and of FL cDNAs (The Rice Full-Length cDNA Consortium, 2003) over the 12 pseudomolecules of rice cv. Nipponbare.
Densities of T-DNA and *Tos17* FSTs and FL CDNA were plotted for each 200 kb. The *y*-axes represent the position on the chromosomes while the *x*-axes show the frequency of T-DNA and *Tos17* insertions and genes limited to a threshold of 50 per interval for better comparison between the three populations of sequences. *Tos17* FSTs, FL cDNA sequences and rice pseudomolecules were retrieved from http://www.ncbi.nlm.nih.gov/entrez/query.fcgi?cmd=search&db=nucleotide, http://www.cdna01.dna.affrc.go.jp/cDNA/, and http://www.tigr.org/tdb/e2 k1/osa1/pseudomolecules/info.shtml, respectively.

Distribution of T-DNA insertions within genes and intergenic regions

Chromosome 1 annotation data were used to analyse the distribution of T-DNA insertions in genes and intergenic regions (Table 4). A total of 6756 genes and an average 3.4 kb size have been predicted on this chromosome (Sasaki *et al.*, 2002). As a result of uncertainties inherent in current semi-automatic annotation, genes were defined in an interval extending from 1500 bp upstream from the putative ATG codon to 750 bp downstream from the putative STOP codon. Overall, 511 (71.5%) of the insertions were found to fall in to this class, with 272 (37.6%) in the sequence between ATG and STOP codons defined here as coding regions, and 206 (28.5%) within putative non-genic regions defined as intergenic regions. All insertions found in genes were annotated and classified according to the type of predicted proteins (i.e. known, putative, unknown or hypothetical, Table S2). A large number of tagged genes corresponds to hypothetical products (55%), whereas a low number of genes codes for known proteins (5%). However, about 25% of tagged genes are associated with an EST. The high number of identified ESTs on intergenic regions (see Table 4) confirms that gene prediction and annotation on the rice genome is still an imprecise process. Therefore, the distribution of T-DNA insertions in predicted genes and intergenic regions should be treated with caution.

Table 4 Distribution of T-DNA insertions in predicted genes and intergenic regions of rice chromosome 1. Calculations are based on gene information data retrieved from RiceGAAS gene prediction program (Sasaki *et al.*, 2002)

Annotation	Insertion no.	Total (%)	Unit average size (kb)	Total size (kb)	Density of insertion ($kb^{-1} \times 100$)	Average to ratio	EST match	%
5′ upstream, −1500 bp; −750 bp	65	9.0	0.75	5067	1.28	0.81	ND	ND
5′ upstream, −750 bp; −250 bp	53	7.3	0.5	3378	1.57	0.99	ND	ND
5′ upstream, −250 bp	42	5.8	0.25	1689	2.49	1.57	ND	ND
Coding region[a]	272	37.6	3.4	22 970.4	1.18	0.75	ND	ND
Intron	183	25.3	2.3	15 538.8	1.18	0.75	ND	ND
Exon	89	12.3	1.1	7431.6	1.20	0.76	ND	ND
3′ downstream, +250 bp	31	4.3	0.25	1689	1.84	1.16	ND	ND
3′ downstream, +250 bp; +500 bp	27	3.7	0.25	1689	1.60	1.01	ND	ND
3′ downstream, +500 bp; +750 bp	27	3.7	0.25	1689	1.60	1.01	ND	ND
Total intergenic[b]	206	28.5		12 351	1.67	1.06	45	21.8
Total in gene[c]	517	71.5		33 394	1.55	0.98	157	30.4
Total insertion	723	100.0		45745	1.58	1.00	ND	ND

[a]Coding region is defined as the sequence between ATG and STOP codon.
[b]Intergenic sequences are defined as sequences where no predicted gene has been found within 1.5 kb upstream from the ATG codon or 0.75 kb downstream from the STOP codon.
[c]Gene is defined as a sequence 1500 bp upstream from the putative ATG codon to 750 bp downstream from the putative STOP codon.

To identify potential bias of integration within different classes of sequence, the density of insertions per 100 kb was calculated for each class. The average density of insertions in predicted gene sequences was 1.55/100 kb. Insertions in gene regulatory regions (i.e. sequences upstream from the ATG or downstream from the STOP codon) were found to be 1.3–2-fold higher than those within CDS (0.75). A preferential site of integration within the first 250 bp upstream from the putative start codon was detected (2.49). Moreover, we did not find any indication of preferential insertion within exon versus intron sequences (0.75).

The density of insertions within intergenic regions (1.67) was found to be similar to that in gene regions (1.55). To check whether genes could have been missed by the predicted annotation program, identification of cognate ESTs within 1.5 kb of sequences extending upstream and downstream from the T-DNA insertion site was performed using a blastn similarity search (see Methods). Cognate ESTs were identified for 22% of sequences corresponding to intergenic regions on the basis of Chr1 annotations (Table 4). This value is lower than that observed for the sequences corresponding to insertions within predicted gene sequences (30.3%), but is highly significant. This indicates that the number of T-DNA insertions in genes is probably higher than estimated in a first analysis.

T-DNA insertion within TE was evaluated. As an overall strategy, we have extracted the 2-kb sequence extending upstream and downstream from the insertion point to create a 4-kb 'enlarged FST' (eFST) database. Using these sequences, we perform a blastx similarity search against a TE database. The database includes 893 annotated rice TE CDS retrieved from GenBank. A cut-off score of 1.00 e^{-20} on blastx results was applied to validate a result. When compared with a population of chromosome 1 random sequences of similar size (no. 800), a threefold lower percentage of insertions in the vicinity of TE (14% versus 5%) were found irrespective of whether insertions were classified as genic or intergenic (data not shown). This shows a lower frequency of T-DNA integration within TE and confirms our previous results indicating that the frequency of insertions within repetitive sequences is lower than expected from a random integration.

Identification of T-DNA insertions in genes

We further attempted to evaluate the frequency of T-DNA insertions within genes encoding proteins. eFST sequences were used to search for protein similarity (blastx) in several public protein databases (nrProtALL). A high percentage (62%) of eFST sequences showed significant similarity to proteins (*E*-value <1.00 e^{-05}) and for half of these (30%) this similarity was highly significant (*E*-value <1.00 e^{-20}) (Figure S1).

As a first insight into the functional categories of genes tagged by the T-DNA, we analysed the distribution of T-DNA inserts in chromosome 1 genes exhibiting similarities to genes of putative and known function (Figure S2). We observed that all the functional classes were represented, notably those corresponding to defence/stress-related, signal perception and transduction genes.

To gain a broader insight into functional classes of genes interrupted by T-DNA inserts and detect a possible bias of insertion in any of these classes, we conducted a similarity search of the T-DNA FSTs against the FL cDNA sequences which have been successfully classified into biological

process categories (The Rice Full-Length cDNA Consortium, 2003). Overall, 775 (10.3%) FSTs matched the 9734 FL cDNA classified sequences reflecting insertion in exon sequences of these genes (Figure 5). Overall, the distribution in functional categories of the FL cDNA interrupted by T-DNA inserts followed that of the whole FL cDNA population. No significant bias was detected for T-DNA insertions for any of the functional categories except for that of translation. Conversely, distribution of the 3472 (19.3%) *Tos17* insertions matching the 9734 classified FL cDNA sequences exhibited an overall bias, resulting from significant over- or under-representations in four classes, notably an overrepresentation of the communication and defence functional category, confirming the results reported by Miyao and co-workers. The higher frequency of *Tos17* FST matching FL cDNA sequences compared with T-DNA FST also reflects the 3× preference of insertion into genic (intron + exon) versus intergenic regions of *Tos17* (Miyao *et al.*, 2003).

For a further evaluation of the number of genes that could be tagged in our total population, we performed an *in silico* search to find all T-DNA insertion sites in genes encoding TF. TF possess highly conserved domains that are specific to a TF family and are easily identified through similarity searches. The total number of rice TF genes falling into large families (e.g. Zinc-finger, Myb, WRKY, HD-ZIP) was estimated by Goff *et al.* (2002). Using the eFST annotation database, we identified 83 T-DNA insertions in the vicinity of genes encoding TF (Table S3). For each large family, the percentage of tagged TF genes ranged from 7 to 16% with the exception of the Zinc-finger family (27%). T-DNA insertions in TF genes were confirmed by manual annotation for three TF gene families (i.e. MYB, WRKY and HD-ZIP). Overall, these results suggest that approximately 10% of rice genes are tagged in our population of 6116 independent T-DNA lines.

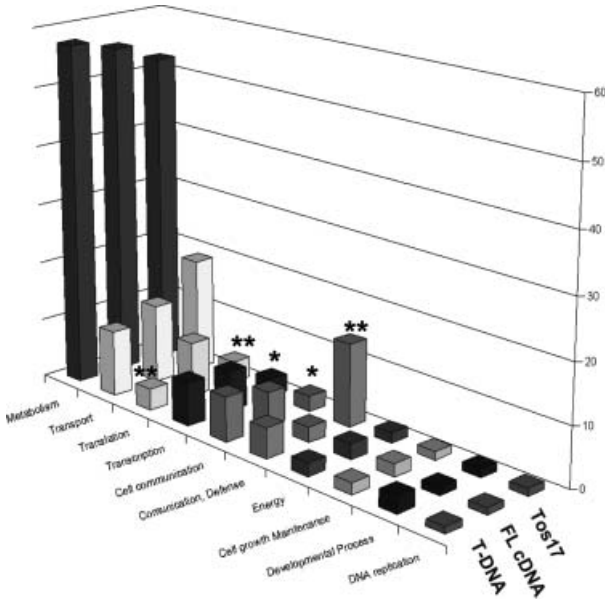

Figure 5. Comparative distribution in Gene Ontology (http://www.geneontology.org) functional categories of the whole population of classified FL cDNA ($n = 9734$), of classified FL cDNA interrupted by a T-DNA insert ($n = 775$) and of classified FL cDNA interrupted by a *Tos17* insert ($n = 3472$). These represented 10.3 and 19.3% of the T-DNA and *Tos17* FSTs used in the analyses. Overall, 29 and 49% of the T-DNA and of the *Tos17* FST populations had hits on the full set of FL cDNAs ($n = 21\,707$) (statistically significant differences in specific classes at: $^*P < 0.05$ and $^{**}P < 0.005$, respectively, in a chi-square test for a fixed distribution hypothesis).

Discussion

The generation of a large number of rice T-DNA lines is no longer a limiting factor because of the establishment of high transformation efficiency procedures (Sallaud *et al.*, 2003). A population of 29 482 T-DNA insertion lines was generated by *Agrobacterium*-mediated transformation over a 2-year period. A high throughput strategy to isolate T-DNA flanking sequences from tissues of young primary transformants has been implemented. The idea was to (i) rapidly generate FSTs during the production process, allowing the creation of a database for *in silico* reverse genetics; and (ii) minimize the handling of uncharacterized lines in the greenhouse by their early selection at the plantlet stage. Using the high throughput amplification protocol, 12 707 plants in which a PCR product corresponding to a putative FST were obtained from 21 295 primary transformants (60%). The efficiency is similar to that described in other reports of large-scale isolation of FSTs from *Arabidopsis* T-DNA collections (Brunaud *et al.*, 2002; Sessions *et al.*, 2002; Szabados *et al.*, 2002). However, only about half of the selected plants produced a genomic sequence of sufficient size (>30 bp) to allow unambiguous identification of the mutation. This major drawback was also reported in the *Arabidopsis* projects mentioned above. It is mainly the result of the frequent occurrence (22%) of sequences corresponding to the binary vector (Kononov *et al.*, 1997) and to T-DNA tandem repeats (Cluster *et al.*, 1996; Szabados *et al.*, 2002). As for other plants, molecular characterization of *Agrobacterium*-mediated transformed rice plants indicated that more than half of the primary transformants harboured repeated tandem T-DNA structures and one-third a partial or full binary vector sequence (Sallaud *et al.*, 2003; Takano *et al.*, 1997; Yin and Wang, 2000). In this context, the system could be improved by sequencing the PCR products during the *in vitro* culture step before transfer of T-DNA plants to the greenhouse. As the rice genome is almost fully sequenced and more accurate annotation will be available in the near future, a more stringent selection could be applied by retaining only the primary transformants in which the T-DNA is inserted into genes.

Distribution of insertions in the rice genome

To determine whether T-DNA integration occurs randomly becomes an important requisite when T-DNA is to be used

as a tool for genome-wide insertion mutagenesis. In particular, this information is necessary to evaluate the population size needed to obtain a knockout in any gene of the genome. In rice, it has been claimed that T-DNA insertion occurs mainly in large gene-rich regions (gene space) (Barakat *et al.*, 2000). However, analysis of the recently completed sequence of chromosome 1 (Sasaki *et al.*, 2002) raises some doubts on the notion of gene space defined by the percentage of GC in large DNA fractions. This prompts the important question of whether or not T-DNA preferentially inserts into particular genome regions. Whereas in *Arabidopsis* the intergenic region is limited to only 8% of the genome, in rice it could represent more than 25%. By different approaches, we evaluated whether a bias of T-DNA integration occurred in rice. Several arguments in favour of such a hypothesis are discussed below.

Using the complete rice physical map and the almost complete sequence of the genome (90–95%), we performed a detailed analysis of the distribution of 6116 independent T-DNA insertions in the rice genome. As a result, 88.8% of these insertions were assigned to rice BAC sequences. At the macromolecular level, our data indicate that T-DNA insertions are rather randomly distributed over the 12 rice chromosomes without significant bias for insertion into a particular chromosome. A similar conclusion was reached following *in situ* hybridization studies in various genomes (Ambros *et al.*, 1986; Chyi *et al.*, 1986; Wallroth *et al.*, 1986) as well as recent *in silico* analyses of the distribution of T-DNA insertions in *Arabidopsis* (Alonso *et al.*, 2003; Brunaud *et al.*, 2002; Sessions *et al.*, 2002; Szabados *et al.*, 2002). However, rice chromosomes 1, 2 and 3, which represent 31.6% of the rice genome size contain a slightly higher percentage of insertions (37.2%). It is interesting to note that cognate genes for 41% of 6591 ESTs are found on the physical map of these three chromosomes, which also suggests that they have a higher euchromatin content and gene density (Wu *et al.*, 2002). By the same token, 40.6% of the FL cDNA was assigned to these three chromosomes. These observations are a first indication that T-DNA insertions may occur preferentially within gene CDS.

A second argument comes from the study performed on T-DNA insertions assigned to chr1. The distribution of the insertions over the chromosome is not random. Moreover, the distribution of the T-DNA insertions and CDS is highly correlated. In particular, both subtelomeric regions present a higher density of insertions than the centromeric region. This profile is also well correlated with the distribution of rice ESTs on this chromosome, for which a density of more than three ESTs per 100 kb have been found in subtelomeric regions compared with less than 1.5 ESTs per 100 kb for the centromeric region (Wu *et al.*, 2002).

A third argument comes from the study of repetitive sequences found in particular in centromeric regions (Sasaki *et al.*, 2002). In *Arabidopsis*, the density of T-DNA insertions within these regions is lower than any other part of the chromosome (Brunaud *et al.*, 2002; Ortega *et al.*, 2002). In chromosomes 1 and 4, this region is known to contain a high density of TE (Feng *et al.*, 2002; Sasaki *et al.*, 2002). If the insertion is localized within repetitive sequences, the small average size of the FSTs (250 bp) could be a drawback to assign an insertion to a unique genome location in rice which contains a high percentage (>25%) of long repeated sequences such as TE (Goff *et al.*, 2002; Yu *et al.*, 2002). We have calculated that the frequency of insertion within repetitive sequences is five- to 10-fold lower than expected if random insertion had occurred. Moreover, T-DNA insertions located on chromosome 1 show a threefold lower insertion frequency within TE elements than expected. As a consequence, more than 90% of the insertions could be assigned unequivocally to a unique genome localization.

Recent analyses of 1009 (Chen *et al.*, 2003) and 3793 (An *et al.*, 2003) T-DNA insertion sites in the rice genome also pointed towards preferential insertion in gene-rich regions and comparable insertion frequency in the intergenic and genic fractions of these regions. These studies also indicated a preferential insertion in 5′ and 3′ regions extending outside the start ATG and stop codons, respectively, which was confirmed in the present report. In *Arabidopsis*, analyses of FST data suggest that T-DNA insertion occurs preferentially in the region upstream from the ATG codon of CDS (Alonso *et al.*, 2003; Brunaud *et al.*, 2002; Sessions *et al.*, 2002; Szabados *et al.*, 2002).

Evaluation of the number of lines needed to knock out all rice genes

According to the bias for integration into gene-rich regions, the size of the population required to knock out any rice gene can be reduced. Krysan *et al.* (1999) established the probability of finding at least one mutation in a gene of x kb in size by the equation $P = 1 - (1 - x/430\,000)^n$ in which 430 000 is the size of the genome in kb and n the number of T-DNA integrations. When applied to our T-DNA population characteristics with 1.4 T-DNA locus per line, and considering that 3.4 kb is the average size of a rice CDS (Sasaki *et al.*, 2002), 271 400 lines would be necessary to knock out any rice genes with 95% probability. With our 6116 tagged lines, the probability of finding an insertion, in a 3.4 kb gene is equal to 4.7%. Using our definition of a gene as 3.4 kb of CDS flanked by 1500 bp upstream from the ATG and 750 bp downstream from the STOP codon, this probability increases to 7.2%. Our analysis indicates that 517 genes of 6756 (7.5%) from chromosome 1 have been tagged, which is very similar to the number expected if random insertion occurred. If the gene definition is restricted to 250 bp upstream and downstream from the ATG and STOP codon,

respectively, then 47% of the insertions are located within gene sequences and thereby generate reliable knockouts. This frequency is remarkably similar to that observed in *Arabidopsis* (Szabados *et al.*, 2002). However, considering that 22% of intergenic regions were found to contain cognate ESTs, the number of tagged genes is probably higher. Moreover, if we consider that the average size of TF genes is similar to the average size of rice genes, the percentage of tagged TF genes (7–10%) also suggests a slight bias towards integration into gene sequences. Overall, we postulate that a range of 5–10% of the rice gene complement is already tagged in our 6116 independent T-DNA lines. Improved knowledge of rice genes identity and genome structure will be needed to confirm that preferential T-DNA insertion occurs within genes versus intergenic region.

In this T-DNA population, the number of tagged genes is not limited to the number of T-DNA insertions. Regenerated primary transformants were found to exhibit an average of 3.2 new copies of the *ty1-copia* endogenous retrotransposon *Tos17*, activated during the tissue culture procedure, which are transmitted to the progeny (E. Bourgeois, unpublished data).

Our current objective is to establish the frequency of tagging of the mutations by the T-DNA, which is an important step to validate the potential of the library for forward genetics screens. Previous experience in *Arabidopsis* has shown that less than 40% of the phenotypes observed are linked to the T-DNA insertion even when no tissue culture step is included in the generation of the transformants (McElver *et al.*, 2001). This frequency might be lower in rice as *Agrobacterium* transformation passes through a tissue culture procedure which, although rather limited in terms of duration and subcultures, is known to generate an undesirable background of somaclonal variations. As a matter of fact, only 5% of the phenotypes observed during field propagation of *Tos17* somaclonal lines were found to be tagged by the retrotransposon (Hirochika, 2001).

The flanking sequences used to carry out the present work are available on GenBank. Information about the library (including phenotypic data and a graphical representation of the FST annotated sequence environment) has been implemented in an ORACLE database Oryza Tag Line which is open to the public at http://genoplante-info.infobiogen.fr/oryzatagline/. The library described here is being seed increased through T1 generation under field conditions. This is a prerequisite in rice for making seeds available in sufficient number to users. The resource is now becoming available to the scientific community with a 6-month delay respective to seed availability with a first access to T2 seeds in September 2004. A mutant line for a particular gene will then be identified through online blast query at the Oryza Tag Line website and seeds made available through a material transfer agreement. This resource, associated with the recent completion of the rice genome sequence, is an important step towards the systematic elucidation of the function of rice genes. It will also be useful in functional genomic studies for other monocotyledonous crops such as wheat, barley or sorghum in which such a tool is not yet available.

Methods

Production of transgenic lines

Five-week-old secondary, seed embryo-derived callus of cv. Nipponbare (*Oryza sativa* subsp. *japonica*) were co-cultured with *Agrobacterium* strain EHA105 or LBA4404 carrying the pC-4978 binary plasmid following the procedure detailed in Sallaud *et al.* (2003). Upon regeneration, the young shoots were transferred to test tubes, after collection of leaf tissue for DNA isolation (see below) and allowed to grow for 3 weeks. The primary transformants selected on the basis of the amplification of a single PCR product (see below) were transferred to the containment greenhouse first in peat pellets and, 15 days later, to pots. A bar-coded system was used to label the plants and T1 seeds.

Plasmid construction

The binary vector pC-4978 is a derivative of pCAMBIA1300 (Jefferson *et al.*, CAMBIA, Canberra, Australia). The vector was constructed as follows: the intron of the castor bean catalase gene was inserted into the *Pst*1 site of the *hpt* gene of pCAMBIA1300 to build pC-5300 (PBF Ouwerkerk *et al.*, Leiden University, The Netherlands; Accession AF294976 in GenBank). The pS4 promoter fused to the Actin1 intron was inserted between *Bst*XI and *Xho*I sites of pC-5300 resulting in pC-4956. The *gusA* gene fused to the 35S minimum promoter (−46 bp; Pasquali *et al.*, 1994) was then inserted into the *Kpn*I/*Sal*I site of pC-4956 previously deleted in the lacZ region using *Hind*III and *Pme*I enzymes to generate pC-4978 (W. Tucker and A. Betzner, Biocem Pacific, Australia, unpublished data).

FST amplification and sequencing

Fresh leaves (25 mg) of young regenerated plants were collected in a 96-collection tube format, deep frozen in liquid nitrogen and stored at −80°C. Genomic DNA was extracted with the Qiagen DNA extraction kit according to the manufacturer's procedures. FSTs were amplified using the adapter-anchor PCR method according to previous publications (Balzergue *et al.*, 2001 modified by Sallaud *et al.*, 2003) except that each DNA sample was digested separately with both *Dra*I and *Ssp*I restriction enzymes. All enzymatic reactions were performed with a Qiagen robot 3000 in a 96-well plate format. A primary transformant was selected for transfer to the greenhouse if a unique PCR2 product was observed on a 1.2% agarose gel from DNA digested with *Dra*I and/or *Ssp*I. Unique PCR2 products were rearranged into 96-well plates using a Qiagen Robot 3000 and directly sequenced by GenomeExpress or Qiagen companies using a T-DNA left border-specific primer CAM6 (5′-CGCTCATGTGTTGAGCATAT).

Sequence analysis

Identification of the T-DNA footprint and binary vector sequences as well as assignation of T-DNA insertion on rice BAC/PAC sequences were carried out by similarity searches (blastn) in a pipeline using a perl program (Brunaud *et al.*, 2002). BAC/PAC sequences were

downloaded from the NCBI site (http://www.ncbi.nlm.nih.gov/) according to the BAC clone list on the RGP web page (http://www.rgp.dna.affrc.go.jp/cgi-bin/statusdb/irgsp-status.cgi) and pseudomolecules from the TIGR site at http://www.tigr.org/tdb/e2k1/osa1/pseudomolecules/info.shtml. To assign a T-DNA insertion on the rice genome, the following method was applied: similarities of FST sequences to BAC/PAC sequences were found using the blastn program. The ratio between the score obtained for an HSP and the score of the FST aligned with itself was calculated. T-DNA assignation on a rice sequence was validated when this ratio was higher than 0.95. Analyses were only performed with FSTs having a minimum size of 30 bp. For each T-DNA insertion assigned on BAC/PAC sequences, 4 kb around the site of insertion were retrieved to create the eFST database (eFST database). Identification of cognate ESTs for an eFST sequence was performed by batch similarity searches (Blastn) using a perl script. An EST was defined as cognate to the eFST sequence if 95% identity was found over the full length of the EST sequence. eFST sequences were also used for similarity searches using the blastx program to identify similarities with proteins in public databases (non-redundant protein database from EMBL, GB, and DDBJ: nrProtAll) at genoplante-info (http://www.genoplante-info.infobiogen.fr/).

Tos17 FSTs (Miyao *et al.*, 2003) and FL cDNA sequences (The Rice Full-Length cDNA Consortium, 2003) were downloaded from http://www.ncbi.nlm.nih.gov/entrez/query.fcgi?CMD=search&DB nucleotide and http://www.cdna01.dna.affrc.go.jp/cDNA/, respectively, and BlastN search for similarity against the rice genome sequence and FL cDNA sequences followed the same procedure as for T-DNA FSTs. T-DNA FST information was organized in an ORACLE® database named Oryza Tag Line at http://www.genoplante-info.infobiogen.fr/oryzatagline/ (P. Larmande, unpublished data).

Creation of a pseudomolecule of rice chromosome 1

The pseudomolecule of rice chromosome 1 was created as follows. According to the physical map published by Chen *et al.* (2002), all BAC/PAC sequences of each chromosome 1 contig were aligned with Sequencher® software to identify overlap coordinates. A perl script was developed to join each non-overlapping sequence and gaps between contigs were artificially eliminated to create the chromosome 1 pseudomolecule. All annotated information was retrieved according to overlap coordinates to avoid information redundancy.

Acknowledgements

This work was supported by the French Génoplante programme and the Génopole organization of Montpellier Languedoc Roussillon. We greatly acknowledge the skilled technical support of Hélène Vignes, Fabienne Venturoli, Sandra Cobo, Ronan Rivallan, Nadia Chatti and Nadège Lanau – in FST generation – and of Marie Pierre Crouzet, Maryannick Rio, Murielle Portefaix, Laurent Rosso, Thierry Mathieu, Christophe Hémery, Jérôme Veyret, Frederic Salles, Carolle Maisonneuve and Christian Chaine – in mutant production – who have worked on the successive phases of the project. We warmly thank Dr R. Cooke for suggestions and critical reading of the manuscript. We thank Dr M. Yano (RGP, Tsukuba, Japan) for kind supply of Nipponbare seeds. We also acknowledge Dr P. Perez (Biogemma), Dr G. Freyssinet (BayerCropScience), Ms Sandrine Balzergue (INRA), Dr Bertrand Dubreucq (INRA) and Dr Loïc Lepiniec (INRA) for valuable discussions during the course of this project.

Supplementary material

The following material is available from http://www.blackwellpublishing.com/products/journals/suppmat/TPJ/TPJ2145/TPJ2145sm.htm

Figure S1. Data summary on blastx similarity results. The percentage of eFST sequences that show similarity with proteins in public database is indicated. eFSTs are the 2 kb sequences upstream and downstream the T-DNA insertion point. Summary data have been obtained from 4303 eFST sequences.

Figure S2. Distribution in functional classes of chromosome 1 genes tagged by a T-DNA insert and exhibiting similarities with genes of known or putative function ($n = 147$).

Table S1 Summary of *Agrobacterium tumefaciens*-mediated transformation experiments using the pC-4978 T-DNA enhancer trap construct mobilized into EHA105 and LBA4404 strains

Table S2 Product characteristics of genes tagged by T-DNA on chromosome 1 The classification was according to the rule established by the Rice genome sequencing consortium (see web page http://demeter.bio.bnl.gov/Tsukuba02.html). Known: 100% homology to known proteins. Putative: similar to proteins with blastP score >100 or *E*-value $<e^{-20}$. Unknown: homology to unknown ESTs. The homology standard is at least 95% identity at the nucleic acid over ~90% of the length of the entire EST, and should cover two adjacent exons. Hypothetical: Sequences predicted by multiple gene prediction programs with no homology to an EST

Table S3 Summary of T-DNA insertions within a TF gene [a]A TF gene is defined as tagged when the insertion is localized within 1500 bp upstream from the ATG to 750 bp downstream from the STOP codon. [b]Estimation of the total number of TF genes in the rice genome (Goff *et al.*, 2002) except for HD-ZIP. [c]Total number of homeobox gene of HD-ZIP class according to P. Ouwerkerk (personal communication). [d]Estimated value for the entire tagged population (#7480 lines representing #6116 independent lines) according to the result observed for 4300 lines and after eliminating duplicated data (i.e. clonal lines, FST from the same line)

References

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

Ambros, P., Matzke, A. and Matzke, M. (1986) Localization of *Agrobacterium rhizogenes* T-DNA in plant chromosomes by in situ hybridization. *EMBO J.* **5**, 2073–2077.

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.

Balzergue, S., Dubreucq, B. and Chauvin, S. *et al.* (2001) Improved PCR-walking for large-scale isolation of plant T-DNA borders. *Biotechniques*, **30**, 496–498, 502, 504.

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

Baulcombe, D.C. (2000) Molecular biology. Unwinding RNA silencing. *Science*, **290**, 1108–1109.

Bechtold, N., Ellis, J. and Pelletier, G. (1993) In planta *Agrobacterium* mediated gene transfer by infiltration of adult *Arabidopsis Thaliana* plants. *Acad. Sci. Ser. III (Paris)*, **316**, 10–1199.

Bennetzen, J.L., SanMiguel, P., Chen, M., Tikhonov, A., Francki, M. and Avramova, Z. (1998) Grass genomes. *Proc. Natl Acad. Sci. USA*, **95**, 1975–1978.

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

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.* **3**, 1152–1157.

Chen, M.S., Presting, G., Barbazuk, W.B. *et al.* (2002) An integrated physical and genetic map of the rice genome. *Plant Cell*, **14**, 3–545.

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.

Cheng, Z., Buell, C.R., Wing, R.A., Gu, M. and Jiang, J. (2001) Toward a cytological characterization of the rice genome. *Genome Res.* **11**, 2133–2141

Chin, H.G., Choe, M.S., Lee, S.-H., Koo, J.C., Kim, N.Y., Lee, N.J., Oh, B.G., Yi, G.H. and Kim, S.C. (1999) Molecular analysis of rice plants harboring an *Ac/Ds* transposable element-mediated gene trapping system. *Plant J.* **19**, 615–623.

Chyi, Y.S., Jorgensen, R.A., Goldstein, D., Tanksley, S.D. and Loaiza-Figueroa, F. (1986) Locations and stability of *Agrobacterium*-mediated T-DNA insertions in the *Lycopersicon* genome. *Mol. Gen. Genet.* **204**, 64–69.

Cluster, P.D., O'Dell, M., Metzlaff, M. and Flavell, R.B. (1996) Details of T-DNA structural organization from a transgenic *Petunia* population exhibiting co-suppression. *Plant Mol. Biol.* **32**, 1197–1203.

De Block, M. and Debrouwer, D. (1991) Two T-DNA's co-transformed into *Brassica napus* by a double *Agrobacterium tumefaciens* infection are mainly integrated at the same locus. *Theor. Appl. Genet.* **82**, 257–263.

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

Delseny, M. (2003) Towards an accurate sequence of the rice genome. *Curr. Opin. Plant Biol.* **6**, 101–105

Delseny, M., Salses, J., Cooke, R., Sallaud, C., Regad, F., Lagoda, P., Guiderdoni, E., Ventelon, M., Brugidou, C. and Ghesquiere, A. (2001) Rice genomics: present and future. *Plant Physiol. Biochem.* **39**, 324–334.

Devos, K.M. and Gale, M.D. (2000) Genome relationships: the grass model in current research. *Plant Cell*, **12**, 5–646.

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

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

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

Fire, A., Xu, S., Montgomery, M.K., Kostas, S.A., Driver, S.E. and Mello, C.C. (1998) Potent and specific genetic interference by double-stranded RNA in *Caenorhabditis elegans*. *Nature*, **391**, 806–811.

Fraser, A.G., Kamath, R.S., Zipperlen, P., Martinez-Campos, M., Sohrmann, M. and Ahringer, J. (2000) Functional genomic analysis of *C. elegans* chromosome I by systematic RNA interference. *Nature*, **408**, 325–330.

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.

Gonczy, P., Echeverri, C., Oegema, K. *et al.* (2000) Functional genomic analysis of cell division in *C. elegans* using RNAi of genes on chromosome III. *Nature*, **408**, 331–336.

Greco, R., Ouwerkerk, P.B., Sallaud, C., Kohli, A., Colombo, L., Puigdomenech, P., Guiderdoni, E., Christou, P., Hoge, J.H. and Pereira, A. (2001a) Transposon insertional mutagenesis in rice. *Plant Physiol.* **125**, 1175–1177.

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

Greco, R., Ouwerkerk, P.B.F., de Kam, R.J., Sallaud, C., Favalli Colombo, L., Guiderdoni, E., Meijer, A.H., Hoge, J.H.C. and Pereira, A. (2003) Transpositional behaviour of an *Ac/Ds* system for reverse genetics in rice. *Theor. Appl. Genet.* **108**, 10–24.

Hanin, M., Volrath, S., Bogucki, A., Briker, M., Ward, E. and Paszkowski, J. (2001) Gene targeting in *Arabidopsis*. *Plant J.* **28**, 671–677.

Hiei, Y., Ohta, S., Komari, T. and 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 the *Tos17* retrotransposon to rice functional genomics. *Curr. Opin. Plant Biol.* **4**, 118–122.

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

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., 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.

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.

Koncz, C., Nemeth, K., Redei, G.P. and Schell, J. (1992) T-DNA insertional mutagenesis in *Arabidopsis*. *Plant Mol. Biol.* **20**, 963–976.

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.

Krysan, P.J., Young, J.C., Tax, F. and Sussman, M.R. (1996) Identification of transferred DNA insertions within *Arabidopsis* genes involved in signal transduction and ion transport. *Proc. Natl Acad. Sci. USA*, **93**, 8145–8150.

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

McElver, J., Tzafrir, I., Aux, G. *et al.* (2001) Insertional mutagenesis of genes required for seed development in *Arabidopsis Thaliana*. *Genetics*, **159**, 1751–1763.

McKinney, E.C., Ali, N., Traut, A., Feldmann, K.A., Belostotsky, D.A., McDowell, J.M. and Meagher, R.B. (1995) Sequence-based identification of T-DNA insertion mutations in *Arabidopsis*: actin mutants act2–1 and act4–1. *Plant J.* **8**, 613–622.

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

Ortega, D., Raynal, M., Laudié, M. *et al.* (2002) Flanking Sequence Tags in *Arabidopsis Thaliana* T-DNA insertion lines: a pilot study. *C.R. Acad. Sci. Ser. III (Paris)*, **325**, 773–780.

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

Pasquali, G., Ouwerkerk, P.B.F. and Memelink, J. (1994) Versatile transformation vectors to assay the promoter activity of DNA elements in plants. *Gene*. **149**, 373–374.

Pereira, A. (2000) A transgenic perspective on plant functional genomics. *Transgenic Res.* **9**, 245–260.

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

Rios, G., Lossow, A., Hertel, B. *et al.* (2002) Rapid identification of *Arabidopsis* insertion mutants by non- radioactive detection of T-DNA tagged genes. *Plant J.* **32**, 243–253.

Rosso, M.G., Li, Y., Strizhov, N., Reiss, B., Dekker, K. and Weisshaar, B. (2003) An *Arabidopsis thaliana* T-DNA mutagenized population (GABI-Kat) for flanking sequence tag-based reverse genetics. *Plant Mol. Biol.* **53**, 247–259.

Sakata, K., Nagamura, Y., Numa, H. *et al.* (2002) RiceGAAS: an automated annotation system and database for rice genome sequence. *Nucl. Acids Res.* **30**, 98–102.

Sallaud, C., Meynard, D., Van Boxtel, J. *et al.* (2003) Highly efficient production and characterization of T-DNA plants for rice (*Oryza sativa* L.) functional genomics. *Theor. Appl. Genet.* **106**, 1396–1408.

Samson, F., Brunaud, V., Balzergue, S., Dubreucq, B., Lepiniec, L., Pelletier, G., Caboche, M. and Lecharny, A. (2002) FLAGdb/FST: a database of mapped flanking insertion sites (FSTs) of *Arabidopsis Thaliana* T-DNA transformants. *Nucl. Acids Res.* **30**, 94–97.

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

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

Shimamoto, K., Miyazaki, C., Hashimoto, H., Izawa, T., Itoh, K., Terada, R., Inagaki, Y. and Iida, S. (1993) Trans-activation and stable integration of the maize transposable element Ds cotransfected with the *Ac* transposase gene in transgenic rice plants. *Mol. Gen. Genet.* **239**, 354–360.

Siebert, P.D, Chenchick, A., Kellogg, D.E, Lukyanov, K.A and Lukyanov, S.A. (1995) An improved PCR method for walking in uncloned genomic DNA. *Nucl. Acids Res.* **23**, 1087–1088.

Smith, N.A., Singh, S.P., Wang, M.B., Stoutjesdijk, P.A., Green, A.G. and Waterhouse, P.M. (2000) Total silencing by intron-spliced hairpin RNAs. *Nature*, **407**, 319–320.

Speulman, E., Metz, P.L., van Arkel, G., te Lintel, H.B., Stiekema, W.J. and Pereira, A. (1999) A two-component enhancer-inhibitor transposon mutagenesis system for functional analysis of the *Arabidopsis* genome. *Plant Cell*, **11**, 1853–1866.

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. and Wakasa, K. (1997) The structures of integration sites in transgenic rice. *Plant J.* **11**, 353–361.

Terada, R., Urawa, H., Inagaki, Y., Tsugane, K. and Iida, S. (2002) Efficient gene targeting by homologous recombination in rice. *Nat. Biotech.* **20**, 1030–1034.

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

The Rice Full-Length cDNA Consortium (2003) Collection, mapping and annotation of over 28 000 cDNA clones from *japonica* rice. *Science*, **301**, 376–379.

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

Upadhyaya, N.M., Zhou, X-R., Zhu, Q-H. *et al.* (2002) An *iAc/Ds* gene and enhancer trapping system for insertional mutagenesis in rice. *Func. Plant Biol.* **29**, 547–559.

Wallroth, M., Gerats, A.M., Rogers, S.G., Fraley, R.T. and Horsch, R.B. (1986) Chromosomal localization of foreign genes in *Petunia hybrida*. *Mol. Gen. Genet.* **202**, 6–15.

Waterhouse, P.M. and Helliwell, C.A. (2003) Exploring plant genomes by RNA-induced gene silencing. *Nat. Rev. Genet.* **4**, 29–38.

Wesley, S.V., Helliwell, C.A., Smith, N.A. *et al.* (2001) Construct design for efficient, effective and high-throughput gene silencing in plants. *Plant J.* **27**, 581–590.

Wu, J., Maehara, T., Shimokawa, T. *et al.* (2002) A comprehensive rice transcript map containing 6591 expressed sequence tag sites. *Plant Cell*, **14**, 525–535.

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

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

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

Young, J.C., Krysan, P.J. and Sussman, M.R. (2001) Efficient screening of *Arabidopsis* T-DNA insertion lines using degenerate primers. *Plant Physiol.* **125**, 513–518.

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.

Accession number: AF294976