Plant Molecular Biology (2005) 59:99–110
DOI 10.1007/s11103-005-8532-6

EU-OSTID: a collection of transposon insertional mutants for functional genomics in rice

L. (Ellen) J. G van Enckevort[1,†], Gaëtan Droc[2,†], Pietro Piffanelli[2,†], Raffaella Greco[1,7], Cyril Gagneur[2], Christele Weber[2], Víctor M. González[3], Pere Cabot[3], Fabio Fornara[4], Stefano Berri[5], Berta Miro[6], Ping Lan[6], Marta Rafel[6], Teresa Capell[6], Pere Puigdomènech[7], Pieter B.F. Ouwerkerk[7], Annemarie H. Meijer[7], Enrico Pe'[5], Lucia Colombo[6], Paul Christou[6], Emmanuel Guiderdoni[2] and Andy Pereira[1,*]

*[1]Plant Research International, Wageningen UR, Post Box 16, 6700 AA, Wageningen, The Netherlands (*author for correspondence; e-mail andy.pereira@wur.nl); [2]CIRAD-AMIS UMR PIA 1096, 2477 Avenue d' Agropolis, 34398, Montpellier Cedex 5, France; [3]Plant Molecular Genetics Laboratory CSIC-IRTA, Jordi Girona, 18-26, 08034, Barcelona, Spain; [4]Dipartimento di Biologia Università degli Studi di Milano, via Celoria 26, 20133, Milano, Italy; [5]Dipartimento di Scienze Biomolecolari e Biotecnologie Università degli Studi di Milano, via Celoria 26, 20133, Milano, Italy; [6]Fraunhofer Institute of Molecular Biology and Applied Ecology (FhIME), Auf dem Aberg 1, 57392, Schmallenberg, Germany; [7]Institute of Biology Leiden University, Clusius Laboratory, PO Box 9505, 2300, RA Leiden, The Netherlands; [†]These authors contributed equally to this paper*

Received 2 April 2004; accepted in revised form 8 June 2005

Key words: functional genomics, mutants, reverse genetics, rice, transposons

Abstract

A collection of 1373 unique flanking sequence tags (FSTs), generated from *Ac/Ds* and *Ac* transposon lines for reverse genetics studies, were produced in *japonica* and *indica* rice, respectively. The *Ds* and *Ac* FSTs together with the original T-DNAs were assigned a position in the rice genome sequence represented as assembled pseudomolecules, and found to be distributed evenly over the entire rice genome with a distinct bias for predicted gene-rich regions. The bias of the *Ds* and *Ac* transposon inserts for genes was exemplified by the presence of 59% of the inserts in genes annotated on the rice chromosomes and 41% present in genes transcribed as disclosed by their homology to cDNA clones. In a screen for inserts in a set of 75 well annotated transcription factors, including homeobox-containing genes, we found six *Ac/Ds* inserts. This high frequency of *Ds* and *Ac* inserts in genes suggests that saturated knockout mutagenesis in rice using this strategy will be efficient and possible with a lower number of inserts than expected. These FSTs and the corresponding plant lines are publicly available through OrygenesDB database and from the EU consortium members.

Introduction

The accurate sequencing of the monocot model plant rice genome has been completed by the end of 2004. The International Rice Genome Sequencing Project (IRGSP) has undertaken this project in an effort to complete the sequencing of all the chromosomes of the cultivar Nipponbare (*Oryza sativa* L. ssp. *Japonica*) (http://rgp.dna.affrc.go.jp/IRGSP). The present size of the rice genome is approximately 370 Mb of non-overlapping sequence and estimations of the total number of rice genes vary greatly from 32 000 to 62 000, depending on the quality of the draft sequences and the gene annotation tools (Delseny, 2003). However, the number of predicted rice genes without any homology to genes in the databases seems unusually high, and this makes the release of a fully assembled and precise rice genome sequence only the first step for systematic functional genomic analysis.

High-throughput reverse genetics strategies need to be developed for the functional analysis of genes discovered by genome sequencing. Insertional mutagenesis using T-DNA or transposons is one of the tools for functional analysis that can provide a phenotype as clue to the gene function, but needs collaborative investment to elucidate all gene functions. In several organisms such as *Drosophila*, *Caenorhabiditis* and even *Arabidopsis*, saturation insertional mutagenesis has been used to create a library of insertion mutants first and subsequently select insertions in genes of interest using PCR-based screens (Pereira, 2000). However, with larger genomes saturation mutagenesis is more difficult to achieve, and it is not cost effective to screen for inserts in a gene-for-gene manner. An alternative strategy is to sequence DNA flanking insertion sequences and catalog these in relation to annotated genes in the genome sequence. Several initiatives to generate large rice insertional mutant collections have been initiated (Hirochika *et al*., 2004) and the databases of insertion flanking sequences are now becoming available in the public domain (An *et al*., 2003; Miyao *et al*., 2003; Kolesnik *et al*., 2004; Sallaud *et al*., 2004).

Our multinational European Consortium was formed in 1997 with the aim of developing heterologous transposon mutagenesis strategies for functional genomics in rice (Greco *et al*., 2001a, b). Enhancer trap gene detection constructs using the two-component *Ac*/*Ds* transposon system were introduced into *japonica* rice and a collection of starter lines were identified (Greco *et al*., 2003). In general the *Ac*/*Ds* system revealed high mobility in rice, although inactivation of *Ds* was observed in later generations. The high frequency of T_1 progeny plants bearing independent insertions allowed the effective use of the two-component *Ac*/*Ds* lines for gene tagging, despite the reduced mobility of *Ds* in later generations. As the diversity in the T_2 insertion pool appeared to be related to the initial number of different T_0 regenerants and therefore T_1 families propagated, a large population of early generation plants was produced to generate large numbers of independent stabilized insertions. The T_2 and T_3 lines generated thus provide a collection of stable insertions that can directly be used for reverse genetics screening.

Constructs employing the autonomous *Ac* transposon introduced into *japonica* and *indica* did not lose mobility and showed continuous transposition (Greco *et al*., 2001b; Kohli *et al*., 2001), supporting its further use in the establishment of a tagging system for knockout mutagenesis. In addition, dedifferentiation-mediated increase in transposition frequency showed the effectiveness of *Ac* as a valuable tool for functional genomics in rice by dramatically reducing the number of plants required to achieve saturation insertional mutagenesis in the rice genome (Kohli *et al*., 2004).

Here we report the current status of the isolation and identification of insertion sites from selected *Ac*/*Ds* and *Ac* lines. Thus far, 6641 *Ac*/*Ds* enhancer trap and 250 *Ac* insertion lines were used in a high-throughput TAIL-PCR analysis for the isolation of *Ds* enhancer trap (*Ds-ET*) and *Ac* flanking sequence tags (FSTs). After filtering the raw sequence data in an automated BLASTN pipeline to remove transposon and vector sequences, the resulting *Ds-ET* and *Ac* FSTs were together with the original T-DNA construct assigned a position on the currently available most accurate rice genome sequence, represented in the TIGR pseudomolecules (Yuan *et al*., 2003).

The *Ac* and *Ds-ET* FST sequences are available in the public database and searchable interface OrygenesDB at http://orygenesdb.cirad.fr, and the corresponding mutant lines are available from the consortium participants. The OrygenesDB interface is a dedicated platform for rice reverse functional genomics searchable by Keyword, Blast and PFAM domains.

Materials and methods

Generation of Ac *insertion lines*

Particle bombardment-mediated transformation of *indica* rice (*Oryza sativa* L. *indica* Pusa

Basmati and Bengal) was carried out as described by Sudhakar *et al.* (1998). Primary and secondary selection of the putative transgenic calli was carried out on medium containing hygromycin as a selective agent. Calli were transferred to fresh medium after 15–18 days and monitored for GFP activity by looking for bright-green sectors while calli were still enclosed in the Petri plates to avoid contamination. The transformation construct and further details pertaining to transgenic plant recovery are as described in Kohli *et al.* (2001).

A few select lines that had exhibited *Ac* copy number amplification in the primary regenerants were re-introduced into culture by inducing callus from seed derived from T_0 plants. Following callus induction and establishment of unique cell lines from each individual seed (Kohli *et al.*, 2004), plants were regenerated using standard methodology. Multiple regenerated plants from the same callus line were shown to have distinct and unique integration patterns for *Ac* in Southern blots confirming that de-differentiation activates further the *Ac* transposon.

Generation of Ac/Ds *enhancer trap insertion lines*

The *Ac/Ds* enhancer trap construct (ET2) was introduced into *Oryza sativa* ssp. *japonica* cv. Nipponbare as described previously by Greco *et al.* (2003). Transposition analyses revealed a strong decrease for *Ac/Ds* activity in the T_2 generation of T_1 active families. Therefore, the T_2 progeny from transpositionally active T_1 families were considered as stabilized lines and used in this study to isolate *Ds* enhancer trap (*Ds-ET*) FSTs. In addition, from a subset of six T_1 families, generated from $4T_0$ primary transformants, and the T_3 generation of the most active T_2 plants was also used to isolate *Ds-ET* FSTs. From 21 families that had shown to be active in T_1 and T_2 generation we propagated all available T_1 seeds to generate additional T_2 insertional mutants for future analyses.

To select for the presence of *Ds*, either young greenhouse grown seedlings were sprayed at day 8 with 2 ml l^{-1}freshly prepared solution of Finale SL 14 (Aventis, 150 g l^{-1} glufosinate ammonium) or leaves were painted with a freshly prepared dilution of Kontour (KB, 60 g l^{-1} ammonium glufosinate).

Generation of T-DNA and transposon flanking sequence tags

Leaf samples collected from young developing greenhouse grown plants were collected in 96 tube-racks and dry-ground using a Mixer Mill MM300 (Retsch, Germany) with 4 mm stainless steel beads. Genomic DNA extraction was performed either with Qiagen DNA extraction kit according to manufacturers' instructions or according to Pereira and Aarts (1998).

ET2 T-DNA flanking sequence tags (FSTs) were amplified using the adapter-anchor PCR method according to Balzergue *et al.* (2001) modified by Sallaud *et al.* (2003). Each DNA sample was digested separately with *Dra*I, *Ssp*I, *Eco*RV, *Nae*I and *Ecl*136 restriction enzymes and ligated to asymmetric adapters. All enzymatic reactions (digestion–ligation, PCR1 and PCR2) were performed with a Qiagen Robot 3000 in a 96-well plate format. Primers in PCR1 were ET2 T-DNA left border primer CWLB3 (5′-CTTGATTTGGG-TGATGGTTCACGTAGTG) or right border primer GFP1b (5′-GCGATCACATGGTCCTGCTG GAGTTC) followed in PCR2 for the left border by CWLB2 (5′-GTTTTTCGCCCTTTGACGTTG GAGTCCA) and for the right border by GFP1c (5′-GGATCACTCTCGGCATGGACGAGCTG TA). PCR2 products were rearranged into 96-well plates and directly sequenced by GenomeExpress (France) using for the left border CWLBSEQ (5′-CACTCAACCCTATCTCGGG CTATTC) and for the right border TnosRBSEQ (5′-ATCCTG TTGCCGGTCTTGCGATGATT) primers.

For the isolation of the *Ds-ET* and *Ac* FSTs we used a three-step TAIL-PCR (Tsugeki *et al.*, 1996) with nested 3′ or 5′ *Ds* transposon primers and six arbitrary degenerate primers as reported in Liu and Whittier (1995) and Tsugeki *et al.* (1996). An automated 96-channel pipettor, the Multimek^{TM} 96 (Beckman Coulter) and the Biomek® 2000 (Beckman Coulter) were used to automate TAIL-PCR and dilution steps. All tertiary TAIL products were picked from 1.8% agarose gels and the re-PCR products were directly sequenced with the *Ds*3-3 primer (Kohli *et al.*, 2004). 5′*Ac* derived FSTs were cloned into pGEMT-easy vector, putative position clones were further identified using the third step of TAIL-PCR and sequenced by the TaKaRa sequence company (Dalian, China).

Characterization of ET2 T-DNA and Ac/Ds *insertion sites*

All raw ET2 T-DNA, *Ds-ET* and *Ac* FST sequence data were screened for binary vector sequences including *Ds* or *Ac* 3′-end sequence by similarity searches (BLASTN) with the ET2 construct sequence as a reference in a pipeline screen using a Perl ad-hoc program (Brunaud *et al.*, 2002). *Ds-ET* and *Ac* FSTs from which the 3′ *Ds* or 5′ *Ac* sequence was removed and with a minimum size of 50 bp were together with the ET2 T-DNA FSTs anchored to the rice genome according to the method described by Sallaud *et al.* (2004). For this assignment we used the TIGR pseudomolecules of the rice genome (ftp://ftp.tigr.org/pub/data/Eukaryotic_Projects/o_sativa/annotation_dbs/pseudomolecules/version_2.0/all_chrs/all.con) and the annotated TIGR rice pseudomolecules (Yuan *et al.*, 2003) (ftp://ftp.tigr.org/pub/data/Eukaryotic_Projects/o_sativa/annotation_dbs/pseudomolecules/version_2.0/all_ch rs/all.con). Furthermore, a similarity search was performed on the KOME Full Length (FL) cDNA database (Kikuchi *et al.*, 2003) (ftp://cdna01.dna.affrc.go.jp/pub/data/current/ine_full_sequence_db.gz) to identify inserts in expressed genes.

Results

Generation of the Ac/Ds *transposon system populations*

To develop an insertional mutagenesis system with maize transposons, we tested a number of *Ac/Ds* and *Ac* transposon constructs by transformation into *japonica* (Greco *et al.*, 2003) and *indica* (Kohli *et al.*, 2004) rice cultivars. The transformants were analyzed for transpositional activity and other parameters (e.g., copy number) useful for generating a large mutant population.

A core collection of 58 *Ac/Ds* enhancer trap (ET2) Nipponbare T_0 rice genotypes (from 26 independent T_0 calli), exhibited active *Ds-ET* transposition, assessed by the diversity of insertional sites revealed by Southern blot analysis. In the T_1 generation 82% of these lines retained activity (51% of the T_1 plants), while in T_2 generation the transposition rate decreased considerably and only 13% of the T_2 progeny plants still displayed new transpositions (Greco *et al.*, 2003). The inactive T_2 plants were considered as a source of stabilized *Ds-ET* insertions to generate a population of T_2 enhancer trap lines for reverse genetics experiments.

On screening 36 T_2 families, comprising 1268 plants originating from 18 T_0 lines, for Basta herbicide resistance, only 10 T_2 families (from 9 T_0 lines) showed the expected segregation of 75–100% resistant plants (data not shown). The other 26 T_2 families showed reduced frequencies of Basta resistant plants ranging from 70 to 0% in a family of T_2 plants. This indicated silencing of the BAR gene in the T_2 generation for 72% of the T_2 families. Therefore, Basta selection was not used further to select for the presence of *Ds* in the T_2 and T_3 generations.

Successive generations that were planted and were available for further analysis are shown on the left in Figure 1. The total population of *Ac/Ds* enhancer trap genotypes advanced comprised of 1421 T_1 plants generating over 200 000 T_2 seed. From these characterized lines, 9958 T_2 plants (from 333 T_1 lines) and 1354 T_3 plants (from 104 T_2 lines), were grown and DNA extracted for further molecular analysis.

Over 3000 *Ac* plants were regenerated from cell lines that were established by culturing T_1 seed-derived from two primary transformants. This culturing and de-differentiation of the cell lines followed by regeneration resulted in the production of independent *Ac* plants showing high levels of new *Ac* insertion sites, as described by Kohli *et al.* (2004).

Isolation of ET2 and Ac/Ds *transposon flanking sequence tags*

The 58 T_0 primary transformants which form the starting generation of our collection of rice *Ds-ET* insertional mutants were derived from 26 independent transformed calli. We adapted a WALK-PCR strategy to amplify FSTs from both Left Border and Right Border sequences of the integrated T-DNA using DNA samples from both T_0 and T_1 plants derived from the 26 original calli. Using this strategy we were able to isolate 18 informative FSTs for 14 ET2 T_0 lines which were anchored to the rice genome. For the 12 remaining lines we did not succeed in the isolation of ET2

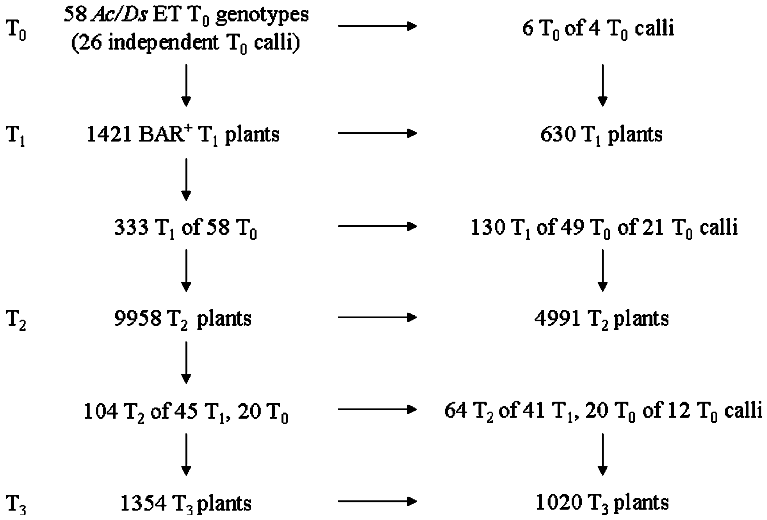

Figure 1. Generation of *Ac/Ds* enhancer trap lines. Flow diagram showing the selection of plants in each generation that comprises the transposon population on the left panel. The right panel describes the number of plants used for TAIL-PCR reactions and the background of these selected plants.

T-DNA FSTs due to the presence of binary vector or T-DNA tandem repeats.

To generate the collection of *Ds-ET* and *Ac* FSTs we first identified the most effective primer combination in TAIL-PCR. For this we tested a series of six AD primers in combination with three nested *Ds* 3′ primers on a representative set of *Ds-ET* samples (data not shown). From these analyses we concluded that the AD2 primer (Liu and Whittier, 1995) and the AD5 primer (Tsugeki *et al.*, 1996) yielded at least one TAIL product for most *Ac/Ds* plants. Therefore, we first used primer AD5 and subsequently primer AD2 to isolate the *Ac* and *Ds-ET* FSTs. In optimizing the TAIL-PCR procedure for the *Ac* lines also *Ds* 5′ nested primers were used in combination with the AD2 and AD5 degenerate primers.

As outlined in Figure 2, we analyzed 6641 *Ds-ET* plant DNA samples for TAIL-PCR reactions (their origin is shown on the right in Figure 1). These comprise 1020 T_3 plants (from 64 T_2, 41 T_1, 20 T_0, 12 T_0 calli), 4991 T_2 plants (from 130 $T_{1,}$ 49 T_0, 21 T_0 calli) and 630 T_1 plants (from 6 T_0, 4 T_0 calli). In addition, about 250 *Ac* plants were subjected to TAIL-PCR using both 3′ and 5′ *Ac* nested primers. Up till now we sequenced 5879 putative *Ds-ET* FSTs and 386 putative *Ac*-FSTs. BLASTN homology analyses with all putative FSTs using an automated ad-hoc pipeline data analysis system resulted in the identification of 2738 good-quality *Ds-ET* FSTs and 154 *Ac* FSTs of at least 50 bp in size. A rather large number of 1437 *Ds-ET* FSTs (24.4%) and 232 *Ac* FSTs (60%) were removed from the dataset given their match to the *Ac/Ds* element (derived from re-integration into the *Ds* element) or to poor-quality sequence derived from presence of multiple PCR products (double peaks). Furthermore, 89 *Ds-ET* sequences (1.5%) were removed because they showed complete homology with the T-DNA binary vector sequence and 1166 sequences *Ds-ET* FSTs (19.8%) were shorter than 50 bp and these were not used for further mapping analyses to the rice pseudomolecules.

Genome location of ET2, Ac *and* Ds-ET *insertions*

A BLASTN homology search was conducted with the 18 ET2 informative T-DNA FSTs isolated from 14 T_0 primary transformants enabling their anchoring to one unambiguous genomic location of the rice genome (Table 1). The 18 annotated ET2 T-DNAs are located on 10 of the 12 rice chromosomes.

A BLASTN homology search with the complete set of 2738 good-quality *Ds-ET* FSTs resulted in the anchoring of 2417 of these FSTs to the rice pseudomolecules. By this mapping of all *Ds-ET* FSTs we determined that 1316 *Ds-ET* FSTs (54.4%) were non-redundant unique FSTs. The

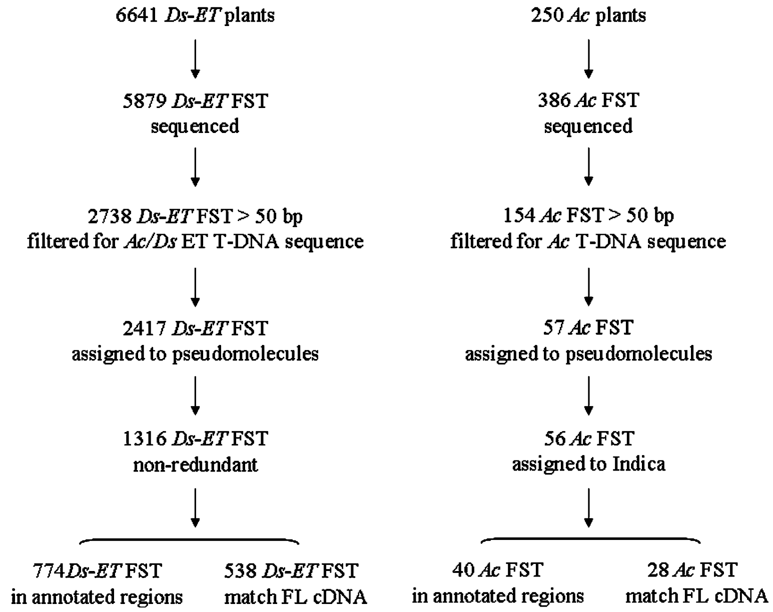
6641 *Ds-ET* plants
5879 *Ds-ET* FST sequenced
2738 *Ds-ET* FST > 50 bp filtered for *Ac/Ds* ET T-DNA sequence
2417 *Ds-ET* FST assigned to pseudomolecules
1316 *Ds-ET* FST non-redundant
774*Ds-ET* FST in annotated regions
538 *Ds-ET* FST match FL cDNA
250 *Ac* plants
386 *Ac* FST sequenced
154 *Ac* FST > 50 bp filtered for *Ac* T-DNA sequence
57 *Ac* FST assigned to pseudomolecules
56 *Ac* FST assigned to Indica
40 *Ac* FST in annotated regions
28 *Ac* FST match FL cDNA

Figure 2. Summary of *Ds-ET* and *Ac* FSTs. Flow diagram showing the number of plants generating the FSTs that were positioned on the rice chromosome pseudomolecules.

redundancy in the dataset derives from FSTs isolated two or more times from a single plant or FSTs from related family members within or between generations. The remaining set of 321 FSTs without homology to the 12 pseudomolecules, were used in a BLASTN search to the yet unanchored *japonica* BAC clones and to BGI (Beijing Genome Institute) *indica* shotgun rice sequences (http://btn.genomics.org.cn:8080/rice/download.php), and 212 of these were found to match rice genomic DNA. It is thus expected that with the completion of the Nipponbare rice sequence, also these FSTs will be positioned on the rice genome.

For the *indica Ac* lines we conducted a pilot project (250 plants) and were able to anchor 154 *Ac*-derived sequences to the *O. sativa japonica* rice genomes which represent a non-redundant set of 57 *Ac* FSTs. Of these, 56 *Ac* FSTs found a match in a BLASTN search to the BGI (Beijing Genome Institute) *indica* shotgun rice sequences.

The distribution of the 1316 *Ds-ET* and 57 *Ac* non-redundant FSTs on the 12 *japonica* rice chromosomes is summarized in Table 1 and Figure 3. Chromosome 1, 2, 3, 4 and 11 contain each 11–12% of the *Ds-ET* and *Ac* FSTs, while 9 and 12 contain each only about 4% of the *Ds-ET* and *Ac* FSTs. The other chromosomes contain numbers varying from 5 to 8% of the total set of 1373 *Ds-ET* and *Ac* FSTs. The overall average density in the rice genome obtained is 3.76 insertions per Mb. Figure 4 shows for three families (callus lines 013, 102 and 288) the distribution of all isolated FSTs for the individual analyzed T_2 and T_3 generation plants of these three families. For all three families a number of insertions are clustered in approximately 1 Mb area around the ET2 T-DNA insertion site showing clearly linked transposition in these areas. In addition, for all three families several T_2 and T_3 generation insertions are distributed over most of the other chromosomes. On chromosome 11 also a clustering of family-related inserts is found, however, for this family from callus 373 the primary ET2 T-DNA insertion was located on chromosome 1.

From the total set of 1373 *Ds*-ET and *Ac* FSTs assigned to the annotated pseudomolecules, 814 (59%) were found in annotated genes. Of these 566 FSTs had also a match in the KOME Full Length cDNA database revealing that 41% of the *Ac*/*Ds* FSTs are integrated in rice expressed genes.

Searching EU-OSTID for inserts in transcription factors

Although the accurate genome sequencing of Nipponbare is close to completion, the annotation of its gene content remains problematic. Many

Table 1. Distribution of *Ac*, *Ds* and T-DNA *Ds* launching pads on the rice chromosomes.

Chromosome	Size in Mb	T-DNA	*Ds* inserts	*Ac* inserts	*Ds* + *Ac*/Mb
1	42.92	2	149	14	3.80
2	35.44	2	159	7	4.68
3	36.12	3	150	6	4.32
4	34.96		149	7	4.46
5	28.90	3	85	3	3.04
6	30.04	1	112	3	3.83
7	29.61	2	97	4	3.41
8	28.27	2	86	5	3.22
9	21.00	1	47	4	2.43
10	22.70	1	69	3	3.17
11	27.84		154	1	5.57
12	27.11	1	59	0	2.18

genomic sequences are not yet annotated and frequent errors occur in the predicted intron–exon patterns of annotated genes. As a consequence BLAST searches with FSTs can give incorrect hits or insertions in coding sequences can be missed. For a more precise evaluation of the complexity of our FST collection, a search was undertaken with a set of 75 largely known transcription factor genes. This set consisted of 45 homeobox genes, including the complete homeodomain-leucine zipper (HD-Zip) families I, II and III (Meijer *et al.*, 2000; Ouwerkerk *et al.*, unpublished results) and all 12 KNOX (Knotted1-like homeobox) genes (Postma-Haarsma *et al.*, 1999; Sentoku *et al.*, 1999; Ito *et al.*, 2002). In addition, a total of 12 auxin-reponse factors [ARF class (Sato *et al.*, 2001)], eight genes of the TB1 CYC PCF or TCP class (Cubas *et al.*, 1999), five Myb genes (Suzuki *et al.*, 1997) and five bZip genes (Onodera *et al.*, 2001) were also used. *Ds-ET* inserts were found in six of these 75 analyzed genes. *OsARF7b* (ARF class), *HOS66* (KNOX class II) and *Oskn4* (a hitherto un-described member of KNOX class I, closely related to *OSH43*, Ouwerkerk *et al.*, unpublished) contained the *Ds-ET* insert in an exon and *Oshox29* (HD-Zip III, Ouwerkerk *et al.*, unpublished) contained the insert in an intron (Figure 5). Furthermore, *Ds* inserts were found in *Oshox14* (HD-Zip family I) and *OsARF1* (ARF class). Current experiments focus on the identification of homozygous lines which will allow the functional characterization of the corresponding genes.

To compare our *Ds-ET* insertion collection to other insertion types we searched for insertions in the same set of 75 transcription factor genes in three other insertion mutant collections present in OrygenesDB. These are the *Tos17* collection (Miyao *et al.*, 2003) with 18 024 sequences, the French Genoplante T-DNA collection of 7481 FSTs (Sallaud *et al.*, 2004), and another collection of 1072 *Ds* inserts (Kim *et al.*, 2004). In addition to the EU-OSTID *Ds* collection, 7, 15 and 8 insertions were found in the T-DNA, *Tos17* and other *Ds* collections (Table 2). In the cumulative FST database (OrygenesDB), 36 different insertional mutants were identified for 25 of the 75 analyzed genes. Especially, insertions in the KNOX and ARF classes are well represented. From both groups 12 genes were analyzed resulting in eight and 12 alleles, corresponding to five and seven different genes, respectively. Moreover, for eight genes, multiple alleles are available from all four collections which will facilitate validation of possible mutant phenotypes.

Discussion

The genome sequence of rice provides us with a large number of genes whose functions are unknown. Predictions of genes based on homology alone, do not provide information on their exact biological role and function *in planta*. This can be achieved experimentally using high-throughput reverse genetics methods, primarily to provide a phenotypic function to these genes. The method of choice is to generate stable mutations in every gene and systematically identify their knockout function(s). While many mutants would not display a mutant phenotype, probably due to redundancy, the generation of double and multiple mutants of

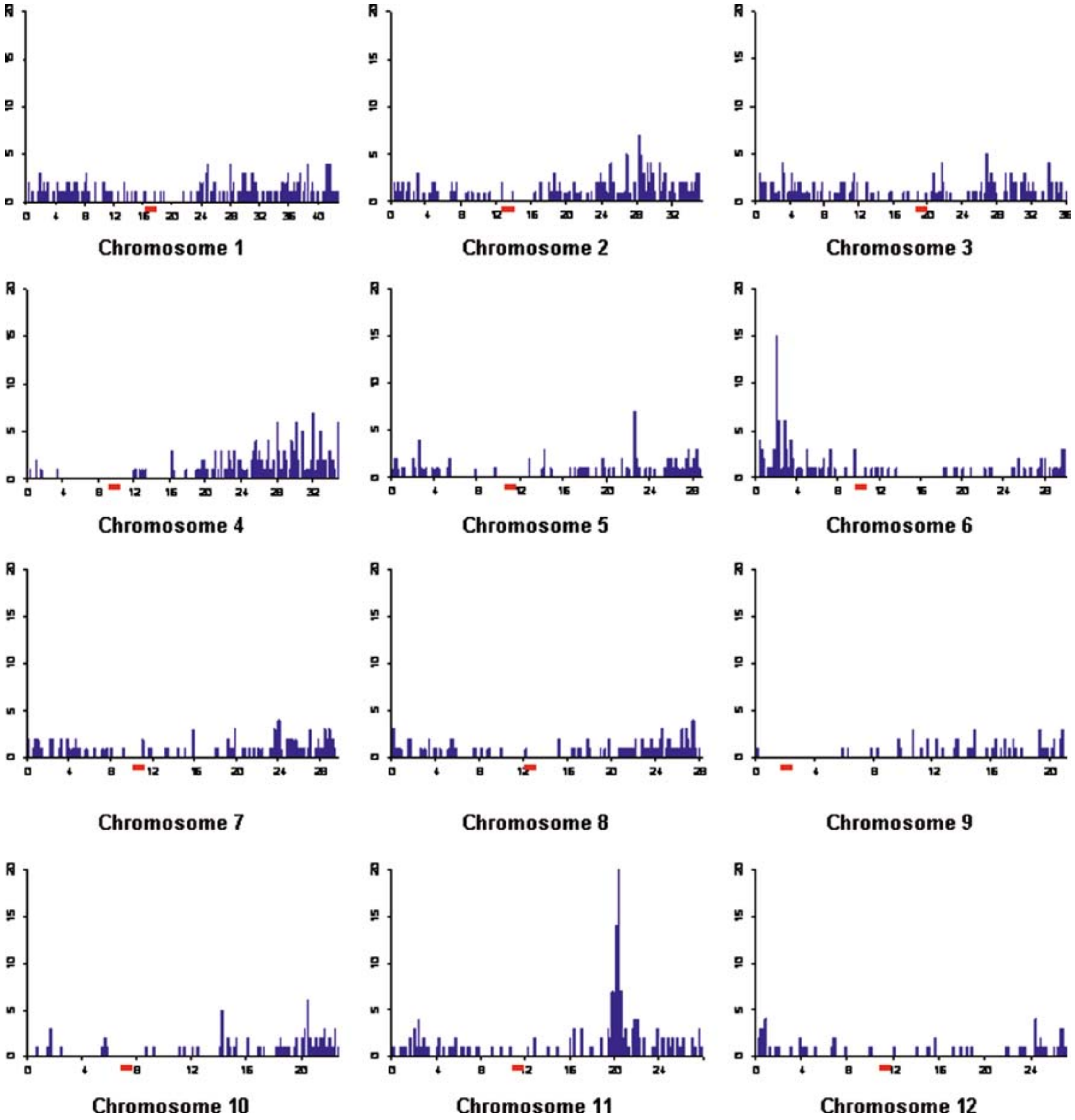

Figure 3. Distribution of 1316 *Ds-ET* and 57 *Ac* unique insertion sites physically mapped on the 12 rice chromosome based upon the reconstructed pseudomolecules derived from Nipponbare genome sequencing. The *x*-axis represents the size of the rice chromosomes in Mb while the *y*-axis shows the number of FSTs (blue bars) in a 200 Kb interval as blue bars. The position of the centromere is indicated as a red square.

the redundant genes would most likely produce a mutant phenotype. The most cost-effective way to catalog insertion mutants is by sequencing the flanking DNA and using the FSTs to position the insertion mutants on the rice genome sequence. These sequence-indexed mutants would provide a permanent resource in the continuous efforts for genome functional analysis.

Our current database of 1373 FSTs from *Ds-ET* and *Ac* insertional mutants, reflects an initial effort on characterization from our transposon collection. This contributes to the growing international rice insertional mutant collections (Hirochika *et al.*, 2004) including endogenous *Tos17* inserts (Miyao *et al.*, 2003), T-DNA lines (An et. al., 2003; Sallaud *et al.*, 2004) and

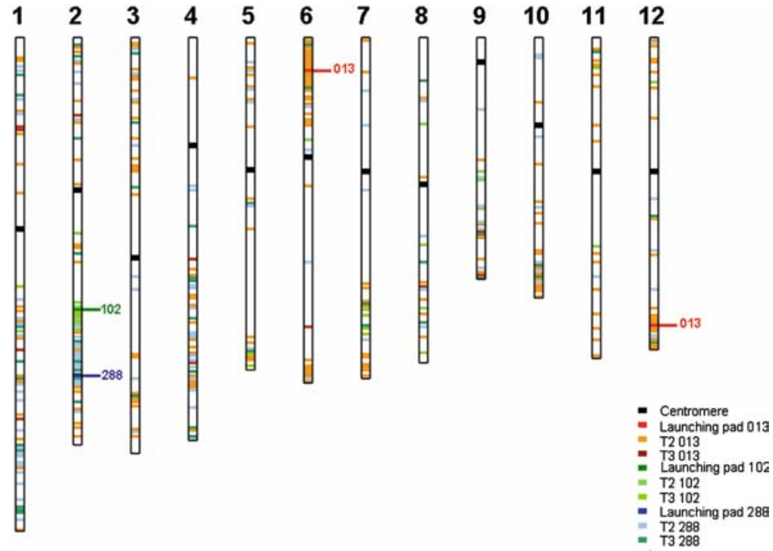

Figure 4. Genomic distribution of *Ds-ET* insertions from T_2 and T_3 generations (see color codes) in the 12 rice pseudomolecules derived from three ET T_0 lines. The position of the original T-DNAs defined as Launching pad is indicated by the corresponding number of the T_0 line, a single copy for lines 102 and 288 and two copies for line 013. The numbers of analyzed T_2 and T_3 generations are for family 013, respectively 218 and 12 plants, for family 102, respectively 16 and 13 plants and for family 288, respectively 105 and 47 plants.

heterologous *Ac*/*Ds* inserts (Kolesnik *et al*., 2004). This diversity of inserts is essential to avoid the bias due to insertional specificity of each insertion element. In contrast to the generation of *Tos17* and T-DNA inserts, both of which have to go through a regeneration phase that induces somaclonal variation, the generation of *Ds* transposon lines in advanced generations can avoid this problem of unwanted background mutations that are segregated out.

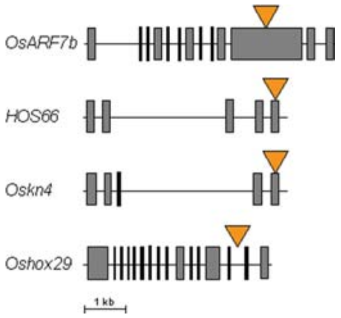

Figure 5. Schematic representation of the genomic organization in exons (boxes) and introns (horizontal lines) of the transcription factor genes *OsARF7b* (Chr. 8), *HOS66* (Chr. 3), *Oskn4* (Chr. 3) and *Oshox29* (Chr. 1) and the presence of the *Ds* inserts (triangles) in these genes (first exon is on the left). The genomic structure for *OsARF7b* is correctly annotated in Genbank Accession AP005509. Nucleotide sequence data and genomic structures reported for *HOS66*, *Oskn4* and *Oshox29* are deposited in the Third Party Annotation Section of the DDBJ/EMBL/GenBank databases under the Accession Numbers TPA: BK005187–BK005189, respectively. *Ds* inserts are between coordinates 52239–52240, 5331–5332, 5150–5151 and 4268–4269 of the Genbank sequences of *OsARF7b*, *HOS66*, *Oskn4* and *Oshox29*, respectively. The scale bar represents 1 kb.

The availability of the rice genome sequences assembled in the form of pseudomolecules enables the exact positioning of the transposon insert FSTs on the 12 rice chromosomes. The insertions are fairly well distributed over all chromosomes (Figure 3) and peaks of insertions represent clustering of inserts isolated from related families that are derived from the same parental callus line(s). For three families (013, 102 and 288) the *Ds-ET* insertion distribution in the T_2 and T_3 generation is visualized in detail in relation to the ET2 T-DNA launching pad (Figure 4). For all three families the existence of high frequency of linked insertion around the original ET2 T-DNA insertion site is clear (approximately 1 Mb around the position of the original T-DNA), a feature observed earlier for *Ac*/*Ds* in rice (Greco *et al*., 2001b). This feature can be used to saturate with *Ac*/*Ds* insertions specific to chromosomal regions of interest using

Table 2. Overview of T-DNA, *Tos17* and *Ds* insertions in a group of 75 genomic sequences corresponding to six different transcription factor (TF) classes from rice.

TF	Gene number	Mutant genes	Mutant alleles	T-DNA	*Tos17*	*Ds* total	EU-OSTID
HD-ZIP	33	9	11	4	3	4	2
KNOX	12	5	8	1	3	4	2
ARF	12	7	12	2	6	4	2
TCP	8	1	1	0	0	1	0
Myb	5	2	3	0	2	1	0
bZip	5	1	1	0	1	0	0
Total	75	25	36	7	15	14	6

specific starting lines. In addition, for all three families of *Ds-ET* lines analyzed in detail *Ds* insertions could be identified in most of the other rice chromosomes, showing that a few founder lines, in this population 26 original transformed calli, are sufficient to reach a good coverage for the whole rice genome.

The bias of *Ac*/*Ds* transposon insertions in coding regions of genes is illustrated by the fact that 59% of the transposon inserts can be found in the coding region of annotated genes and 41% in expressed genes. The insertional bias in genes clearly suggests that *Ac*/*Ds* transposon lines represent an effective tool for generating knockout mutations in genic regions, and a much lower number of plants than what is predicted for random insertion distribution may be necessary to saturate the rice genome with mutations.

Experimental verification of gene insertional specificity was also demonstrated by a screen for insertions in a set of well-annotated transcription factor genes under investigation. In our collection of 1373 FSTs we found inserts in six out of 75 genes this translates to an 8% chance of obtaining an insert in a given gene. In contrast, the *Tos17* collection yielded only 15 inserts (representing 11 different genes) in a collection of 18 024 FSTs, extrapolating to a 20% chance of obtaining an insert in a specific gene of interest. Although the *Tos17* collection is 13 times larger, the chance of finding a gene from our database of 75 transcription factor genes is only 2.5 times more. This difference is likely due to the insertional preference of *Tos17* for stress-related and tissue-culture induced genes and relatively poor insertion in other classes of genes. The cumulative 27 950 FSTs in OrygenesDB revealed insertion mutants for 33% of the 75 transcription factor genes with multiple insertion alleles totaling to 36 mutant alleles.

The main conclusion is that with the insertion mutant collections that are in development at this moment, including EU-OSTID, it is already possible to find mutations in genes-of-interest with a chance of 33%, and with multiple alleles for many of these genes. Together with the observation of *Ac*/*Ds* insertional specificity of about 60% in genes, the total number of insertional mutants required for genome saturation is much less than would be predicted based on random genome coverage.

This database (http://orygenesdb.cirad.fr) and collection of *Ds-ET* and *Ac* transposon lines are a valuable contribution to the growing international efforts (Hirochika *et al*., 2004) to make available functional genomics resources to the scientific community that would be valuable for rice as well as other cereals.

Acknowledgements

We acknowledge the support of the EU 5th framework project QLG2-CT-2001-01453 Cereal-Gene Tags. We would like to thank Dick Geurtsen for greenhouse support, Paul Mooyman and Dr Vanniarajan Chockalingham for help in DNA analysis and Bas te Lintel Hekkert and Dr Christine Bunthof for help in database setup.

References

An, S., Park, S., Jeong, D.-H., Lee, D.-Y., Kang, H.-G. Yu, J.-H., Hur, J., Kim, S.-R., Kim, Y.-H., Lee, M., Han, S., Kim, S.-J., Yang, J., Kim, E., Wi, S.J., Chung, H.S., Hong, J.-P., Choe, V., Lee, H.-K., Choi, J.-H., Nam, J., Kim, S.-R., Park, P.-B., Park, K.Y., Kim, W.T., Choe, S., Lee, C.-B. and An, G. 2003. Generation and analysis of end sequence database for T-DNA tagging lines in rice. Plant Physiol. 133(4): 2040–2047.

Balzergue, S., Dubreucq, B., Chauvin, S., Le Clainche, I., Le Boulaire, F., de Rose, R., Samson, F., Biaudet, V., Lecharny, A., Cruaud, C., Weissenbach, J., Caboche, M. and Lepiniec, L. 2001. Improved PCR-walking for large-scale isolation of plant T-DNA borders. Biotechniques 30: 496–498, 502, 504.

Brunaud, V., Balzergue, S., Dubreucq, B., Aubourg, S., Samson, F., Chauvin, S., Bechtold, N., Cruaud, C., DeRose, R., Pelletier, G., Lepiniec, L., Caboche, M. and Lecharny, A. 2002. T-DNA integration into the *Arabidopsis* genome depends on sequences of pre-insertion sites. EMBO Rep. 3: 1152–1157.

Cubas, P., Lauter, N., Doebley, J. and Coen, E. 1999. The TCP domain: a motif found in proteins regulating plant growth and development. Plant J. 18: 215–222.

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

Greco, R., Ouwerkerk, P.B.F., Sallaud, C., Kohli, A. Colombo, L., Puigdomènech, P., Guiderdoni, E., Christou, P., Hoge, J.H.C. and Pereira, A. 2001a. Transposon insertional mutagenesis in rice. Plant Physiol. 125(3): 1175–1177.

Greco, R., Ouwerkerk, P.B.F., Taal, A.J.C., Favalli, C., Beguiristain, T., Puigdomènech, P., Colombo, L., Hoge, J.H.C. and Pereira, A. 2001b. Early and multiple *Ac* insertions 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, C., 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.

Hirochika, H., Guiderdoni, E., An, G., Hsing Y.-i., Eun, M.Y., Han C.-d., Upadhyaya, N., Ramachandran, S., Zhang, Q., Pereira, A., Sundaresan, V. and Leung, H. 2004. Rice mutant resources for gene discovery. Plant Mol. Biol. 54: 325–334.

Ito, Y., Hirochika, H. and Kurata, N. 2002. Organ-specific alternative transcripts of KNOX family class 2 homeobox genes of rice. Gene 288: 41–47.

Kikuchi, S., Satoh, K., Nagata, T., Kawagashira, N., Doi, K., Kishimoto, N., Yazaki, J., Ishikawa, M., Yamada, H., Ooka, H., Hotta, I., Kojima, K., Namiki, T., Ohneda, E., Yahagi, W., Suzuki, K., Li, C.J., Ohtsuki, K., Shishiki, T., Otomo, Y., Murakami, K., Iida, Y., Sugano, S., Fujimura, T., Suzuki, Y., Tsunoda, Y., Kurosaki, T., Kodama, T., Masuda, H., Kobayashi, M., Xie, Q., Lu, M., Narikawa, R., Sugiyama, A., Mizuno, K., Yokomizo, S., Niikura, J. Ikeda, R., Ishibiki, J., Kawamata, M., Yoshimura, A., Miura, J., Kusumegi, T., Oka, M., Ryu, R., Ueda, M., Matsubara, K., Kawai, J., Carninci, P., Adachi, J., Aizawa, K., Arakawa, T., Fukuda, S., Hara, A., Hashidume, W., Hayatsu, N., Imotani, K., Ishii, Y., Itoh, M., Kagawa, I., Kondo, S., Konno, H., Miyazaki, A., Osato, N., Ota, Y., Saito, R., Sasaki, D., Sato, K., Shibata, K., Shinagawa, A., Shiraki, T., Yoshino, M. and Hayashizaki, Y. 2003. Collection, mapping, and annotation of over 28 000 cDNA clones from *japonica* rice. Science 301: 376–379.

Kim, C.M., Piao, H.L., Park, S.J., Chon, N.S., Je, B.I., Sun, B., Park, S.H., Park, J.Y., Lee, E.J., Kim, M.J., Chung, W.S., Lee, K.H., Lee, Y.S., Lee, J.J., Won, Y.J., Yi, G.H. Nam, M.H., Cha, Y.S., Yun, D.W., Eun, M.Y. and Han, C.D. 2004. Rapid, large-scale generation of Ds transposant lines and analysis of the Ds insertion sites in rice. Plant J. 39: 252–263.

Kohli, A., Xiong, J., Greco, R., Christou, P. and Pereira, A. 2001. Tagged transcriptome display (TTD) in *indica* rice using *Ac* transposition. Mol. Genet. Genom. 226: 1–11.

Kohli, A., Prynne, M.Q., Miro, B., Pereira, A., Twyman, R.M., Capell, T. and Christou, P. 2004. Dedifferentiation-mediated changes in transposition behavior make the *Activator* transposon an ideal tool for functional genomics in rice. Mol. Breeding 13: 177–191.

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.

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

Meijer, A.H., de Kam, R.J., d'Erfurth, I., Shen, W. and Hoge, J.H.C. 2000. HD-Zip proteins of families I and II from rice: interactions and functional properties. Mol. Gen. Genet. 263: 12–21.

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.

Onodera, Y., Suzuki, A., Wu, C.Y., Washida, H. and Takaiwa, F. 2001. A rice functional transcriptional activator, RISBZ1, responsible for endosperm-specific expression of storage protein genes through GCN4 motif. J. Biol. Chem. 276: 14139–14152.

Pereira, A. and Aarts, M.G.M. 1998. Transposon tagging with the *En–I* system. In: J. Martinez-Zapater and J. Salinas (Eds.), *Arabidopsis* Protocols, Humana Press Inc, Totowa, New York, pp. 329–338.

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

Postma-Haarsma, A.D., Verwoert, I.I.G.S., Stronk, O.P., Koster, J., Lamers, G.E.M., Hoge, J.H.C. and Meijer, A.H. 1999. Characterization of the KNOX class homeobox genes Oskn2 and Oskn3 identified in a collection of cDNA libraries covering the early stages of rice embryogenesis. Plant Mol. Biol. 39: 257–271.

Sallaud, C., Meynard, D., Van Boxtel, J., Gay, C., Bès, M., Brizard, J.P., Larmande, P., Ortega, D., Raynal, M., Portefaix, M., Ouwerkerk, P., Rueb, S., Delseny, M. and Guiderdoni, E 2003. Highly efficient production and characterization of T-DNA plants for rice (*Oryza sativa* L.) functional genomics. Theor. Appl. Genet. 106: 1396 1408.

Sallaud, C., Gay, C., Larmande, P., Bès, M., Piffanelli, P., Piégu, B., Droc, G., Regad, F., Bourgeois, E., Meynard, D., Périn, C., Sabau, X., Ghesquière, A., Glaszmann, J.C., Delseny, M. and Guiderdoni, E. 2004. High throughput T-DNA insertion mutagenesis in rice: a first step towards *in silico* reverse genetics. Plant J. 39: 450–464.

Sato, Y., Nishimura, A., Ito, M., Ashikari, M., Hirano, H.Y. and Matsuoka, M. 2001. Auxin response factor family in rice. Genes Genet. Syst. 76: 373–380.

Sentoku, N., Sato, Y., Kurata, N., Ito, Y., Kitano, H. and Matsuoka, M. 1999. Regional expression of the rice KN1-type homeobox gene family during embryo, shoot, and flower development. Plant Cell 11: 1651–1664.

Sudhakar, D., Duc, L.T., Bong, B.B., Tinjuangjun, P. Maqbool, S.B., Valdez, M., Jefferson, R. and Christou, P. 1998. An efficient rice transformation system utilizing mature seed-derived explants and a portable, inexpensive particle bombardment device. Transgenic Res. 7: 289–294.

Suzuki, A., Suzuki, T., Tanabe, F., Toki, S., Washida, H. Wu, C.Y. and Takaiwa, F. 1997. Cloning and expression of five myb-related genes from rice seed. Gene 198: 393–398.

Tsugeki, R., Kochieva, E.Z. and Fedoroff, N.V. 1996. A transposon insertion in the *Arabidopsis SSR16* gene causes an embryo-defective lethal mutation. Plant J. 10: 479 489.

Yuan, Q., Ouyang, S., Liu, J., Suh, B., Cheung, F., Sultana, R., Lee, D., Quackenbush, J. and Buell, C.R. 2003. The TIGR rice genome annotation resource: annotating the rice genome and creating resources for plant biologists. Nucl. Acids Res. 31: 229–233.