Insertional mutagenesis: Lessons from *Arabidopsis*, gaining experience in rice

EMMANUEL GUIDERDONI1*, PIETRO PIFFANELLI1, CHRISTOPHE SALLAUD1, EMMANUELLE BOURGEOIS1 AND ALEXANDER T. JOHNSON2

1UMR PIA 1096, Cirad-Amis, TA40/03, 34398 Montpellier Cedex 05, France
2Department of Plant Sciences, University of Cambridge, Downing St, Cambridge CB2 3EA, UK

*Corresponding author: guiderdoni@cirad.fr

Abstract

Insertional mutagenesis is a means of analyzing the function of genes through their disruption by foreign DNA followed by observation of the effect on phenotype. Insertion libraries are a key tool in plant functional genomics projects aimed at deciphering major developmental and physiological processes because they permit both the discovery of new genes through forward genetics screens and gene detection systems as well as the validation of genes of known sequence through reverse genetics strategies. Benefiting from experience accumulated over the last decade in the model dicotyledo-
nous species *Arabidopsis*, large insertion line libraries in turn have been generated in rice (*Oryza sativa* L.), the model plant for cereals and grasses, in the last five years. The use of insertion mutagens has varied and encompasses T-DNA, delivered during co-culture with *Agrobacterium*, the genetically engineered maize transposable *Ac/Ds* system and the endogenous *ty1*-copia retroelement *Tos17*, mobilized during tissue cul-
ture. Here we review the current status of rice insertional mutagenesis by concen-
trating on specific features and advantages and drawbacks of each mutagen, and examine the establishment of this new resource from a rice functional genomics perspective.

Introduction

Following the sequencing of the *Arabidopsis* genome (*Arabidopsis* Genome Initiative 2000), the 440 Mbp genome of rice is now nearly fully sequenced by the public International Rice Genome Sequencing Project consortium and pseudomolecules of chromosomes 1, 4 and 10 have already been published (Feng et al. 2002; Sasaki et al. 2002; Rice Chr10 Sequencing Consortium 2003). This breakthrough followed three independent shotgun sequencing projects conducted by two private
companies (Monsanto, see Barry 2001; Syngenta, see Goff et al. 2002) and the Chinese National Genomics Initiative at the Beijing Genomics Institute (Yu et al. 2002) in both *japonica* and *indica* rice (for a comprehensive review see Delseny 2003).

A range of 30,000 to 60,000 genes have been predicted in rice following annotation procedures and sequencing results, compared to the 25,600 genes identified in *Arabidopsis*, which already revealed extensive duplication of two-thirds of the genome and the presence of large multigene families existing as clustered and/or dispersed copies. Eighty per cent of the *Arabidopsis* genes have a homolog in rice whereas nearly half of the predicted rice genes have no homolog in the model dicotyledonous species (Yu et al. 2002). Furthermore, 98% of genes known in cereals are found in the rice genome, confirming the potential of rice as a model monocot for deciphering the function of genes of other cereal crops.

Together with a collection of more than 200,000 ESTs in public databases and the recent release of 28,469 full length cDNAs (The Full Length cDNA Sequencing Consortium 2003) there is now an enormous wealth of rice sequence information available. The next challenge is to clearly assign a biological role to these sequences, the function of which only a few thousand can be defined with great confidence based on sequence similarity with genes of known function.

Recent studies have shown that the determination of gene function in cereals cannot be simply extrapolated from *Arabidopsis* data due to the fact that monocotyledons and dicotyledons separated some 150 million years ago followed by a wide diversification and specialization of gene function within each lineage (Devos and Gale 2000). That limitation can be illustrated by the divergent functions of the *LEAFY* gene which encodes a major transcription factor involved in the formation of floral meristems in *Arabidopsis*, and its rice homolog *RFL* which is involved in panicle branching (Kyozuka et al. 1998).

Inactivation of gene expression followed by observation of the consequence on phenotype is the most direct approach for discovering or validating gene function. The methods that have been developed in higher plants rely on either a targeted inactivation of the expression of a predefined sequence (sequence-specific gene disruption/replacement, through homologous recombination or using self-complementary chimaeric oligonucleotides, or sequence-specific silencing mediated by the expression of sense/antisense/double stranded RNA) or a random disruption created by a physical, chemical or insertion mutagen (transposon, retrotransposon or T-DNA) followed by screening for a particular molecular lesion. First applied to the model dicotyledonous plant *Arabidopsis*, all of these methods have been extended to rice in the last five years. Here we review recent progress in the development of insertion libraries assisting the genome-wide discovery of gene function in this model system.
Methods for inactivating gene function in higher plants

RNA-based inactivation of gene expression

RNA-induced gene silencing, also known as post transcriptional gene silencing, is based on the introduction of doubled stranded RNA (dsRNA) into an organism to induce a sequence-specific RNA degradation mechanism that silence a targeted gene. The method relies on the fact that when naturally occurring viral RNA (that produces dsRNA during its replication), self complementary single-stranded hairpin RNA (HpRNA) or dsRNA is introduced into a plant, it is degraded into dsRNA fragments of 21 nucleotides long, known as small interfering RNAs (siRNA). The siRNA fragments are incorporated into a nuclease containing complex called RISC (RNAi silencing complex) which degrades mRNAs that are complementary to the single-stranded siRNA that is associated with the complex (Waterhouse and Helliwell 2003).

The widespread use of this method for targeted inactivation of gene expression offers an additional and complementary tool to the traditional knock-outs created by insertional mutagenesis. It is expected to be particularly useful for simultaneously inactivating the expression of all the individual members belonging to a gene family. This indeed has proved to be tedious and time-consuming task to achieve through gene-by-gene based disruption followed by stacking of the mutations in a single line by crossing. It is also possible to envisage the systematic creation of co-suppression lines in a high throughput manner for model species. This would allow a rapid progress in the global understanding of gene function, notably in lineages containing large multigenic families such as the higher plants (see the AGRIKOLA project as one example of such a project in Arabidopsis: www.evry.inra.fr/public/projects/agrikola/agrikola.html).

Transcription of HpRNA appears to be the most efficient method for suppressing the expression of an endogenous gene in plants (Smith et al. 2000) and is replacing less reliable antisense methods that have been used for many years. To construct a gene encoding such a hairpin RNA, a short region specific to the target transcript is first identified to avoid cross-silencing of homologous genes. This 150-600 bp long Gene Specific Tag (GST) is then cloned on both sides of an intron spacer in inverted orientation. Recombination cloning techniques (such as GATEWAY® technology) speeds up the construction of HpGST T-DNA vectors for high-throughput transformation.

Examples of HpRNAi-mediated inactivation of endogenous single genes or gene families are now being accumulated in rice since the first successful report (Wesley et al. 2001). The most detailed study was conducted by Ko Shimamoto’s group at the University of Nara, with the gene-specific and gene family-specific inactivation of expression of the Osrac gene family consisting of seven members (Miki et al. 2003). Other examples include the inactivation of transcription factor expression
such as the AP3 homologue OsMADS16 (Xian et al. 2003) and OsBP73 (Chen et al. 2003a). However, as of now no project for the systematic creation of RNAi lines has been launched in rice. The availability of versatile vectors for monocot transformation (with appropriate promoters and selectable markers), of high throughput rice transformation protocols and a library of 28,000 full length cDNA would however allow embarking on such a project with confidence.

**Homologous recombination**

Following the first report of targeted correction of a transgene in the tobacco genome (Paszkowski et al. 1988), gene targeting (GT) events have been described at several loci but only in dicotyledonous plants and always with a very low frequency ranging from $10^{-3}$ to $10^{-6}$ (e.g. Kempin et al. 1997; Hanin et al. 2001).

Sequence homology-driven disruption of the waxy gene using a T-DNA bearing a powerful diphtheria toxin gene negative selection marker at right and left borders has recently been reported in rice (Terada et al. 2002). The strategy for the targeted disruption was a gene replacement through two crossing-overs at the flanking homologous regions of 6.3-6.8 kb. In this report the gene targeting efficiency attained one homologous recombination (HR) event for 99 escapes of illegitimate recombination, corresponding to a potential of 1,500 integration events when no negative marker is on the T-DNA. This efficiency was similar to that reported in embryogenic stem cells of mice. Extending HR-mediated gene disruption to other loci as well as implementing strategies for enhancing efficiencies through the use of novel introducing vectors are underway and may provide an efficient tool for GT in rice in the next few years (Iida and Terada 2004).

**Chemical and physical mutagenesis**

Conventionally induced mutagenesis relies on the use of chemical (e.g. EMS, diepoxybutane) or physical (e.g. fast neutrons, gamma rays) agents and has been used as a tool in genetics and breeding of crop plants for the last 70 years. Though sequencing of the genome has considerably facilitated the positional cloning of mutant alleles in rice (a recent example was the cloning of a spontaneous mutation in a GRAS family of transcription factor gene Moc1, controlling lateral bud initiation and tillering, Li et al. 2003), the use of traditional mutant populations for reverse genetics was until recently impossible. However, the development of PCR-based technologies for screening single nucleotide point mutations (TILLING; Till et al. 2003) to few kilobases deletions (Delete-a-Gene®; Li et al. 2001) in a particular candidate sequence among pooled DNA samples has placed a new emphasis on traditional mutagenized populations. TILLING is based on the formation of heteroduplexes following denaturation and annealing of PCR products amplified using target gene-specific primer pairs in pooled DNA samples containing mutant and wild type alleles of a target gene. Single nucleotide mismatches in
the heteroduplex are specifically recognized by the endonuclease CEL1, producing shorter strands that are detected differentially on a sequencing gel after appropriate labelling.

Delete-a-Gene® relies on the preferential amplification and subsequent enrichment of smaller fragments resulting from kilobase-scale deletions within the PCR products among pooled DNA samples of fast neutron bombardment-mediated mutagenized populations. Deletion mutagenesis may be the best way to knock out genes tandemly repeated as two or more copies, which are estimated to represent 16 and 22% of Arabidopsis and rice genes, respectively. In rice, a population of 40,000 M$_3$/M$_4$ mutant lines has been established in the lowland cultivar IR64 using these mutagens (Leung et al. 2001). The population is being evaluated for morphological and physiological changes and altered pathogen and drought responses, as well as its amenability to reverse genetics (Manosalva et al. 2003). Such populations are of particular importance for analysis of the indica genome since it is less accessible to insertional mutagenesis than the japonica genome, due to the overall poor tissue-culture response of cultivars in this subspecies.

**Insertional mutagenesis**

Insertional mutagenesis is another means of disrupting gene function based on the insertion of foreign DNA into the genome which, by integrating into coding sequences, interferes with the expression of any gene. The mutagen further acts as a molecular tag for the identification of the resulting mutation. The insertional mutagen can be T-DNA introduced upon co-culture of plant cells with Agrobacterium or an endogenous or heterologous transposable element (transposon or retrotransposon). Large collections of several hundred thousand insertion lines are now available in Arabidopsis which, in combination, provide the systematic disruption of almost any gene (see the TAIR portal: www.arabidopsis.org/index.jsp). While several initiatives have demonstrated the potential of the maize Ac/Ds and En/Spm transposon systems in Arabidopsis (Parinov et al. 1999; Speulman et al. 1999; Tissier et al. 1999; Raina et al. 2002), T-DNA remains the preferred choice of insertional mutagenesis for plants where an efficient genetic transformation procedure is available (Feldmann 1991; Koncz et al. 1992; Betchold et al. 1993; Krysan et al. 1999; Sessions et al. 2002; Szabados et al. 2002; Alonso et al. 2003). The tagged lines are used for direct genetics screens under standard or specific growth conditions, as well as reverse genetics searches for insertions in particular candidate sequences. Identification of individual plants/lines that carry a particular mutation in a known sequence of interest is based on either PCR screening in DNA pools, using a primer specific to the candidate sequence and another specific to the mutagen (McKinney et al. 1995; Krysan et al. 1999; Young et al. 2001; Rios et al. 2002), or computational searches in Flanking Sequence Tag (FST: a sequence of the genome flanking the mutagen insertion site) databases. Though requiring a large ini-

Several initiatives to develop such gene machines through insertional mutagenesis have been launched in rice over the past five years and FST databases are accessible through the internet: tos.nias.affrc.go.jp; postech.ac.kr/life/risd; www.niab.go.kr; www.pi.csiro.au/fgrttpub/home.htm; www.genoplante-info.infobio-gen.fr/oryzatagline/; www.ricefgchina.org.

Generation of the libraries is based on the use of either the Ac element (Enoki et al. 1999; Greco et al. 2001), a more sophisticated double component Ds/Ac Transposase system (Chin et al. 1999; Upadhyaya et al. 2002; Greco et al. 2003; Van Enckevort et al. 2004), the tissue culture-activated endogenous retrotransponson Tos17 (Hirochika 2001) or T-DNA (Jeon et al. 2000; Jeong et al. 2002; Chen et al. 2003; Wu et al. 2003b; Sallaud et al. 2004). Clearly the most advanced initiative is that of the National Institute of Agrobiological Sciences (NIAS) at Tsukuba, Japan where 50,000 lines comprising 500,000 Tos17 insertions have been generated through tissue culture (Hirochika 2001; Miyao et al. 2003). The initiatives are now merging their efforts in an International Rice Functional Genomics Consortium, coordinated by the International Rice Research Institute (IRRI) at www.iris.irri.org/IRFGC/, which provides useful links to end sequence databases (Hirochika et al. 2004).

**Main features of the insertional mutagens in rice**

**Ac/Ds**

Since the pioneering work of Ko Shimamoto’s group in the early 90s, which demonstrated that the 4.565 kbp Activator (Ac) element originally isolated from maize was active in transformed rice protoplasts, efforts have intensified towards the establishment of gene machines using this system (Enoki et al. 1999; Greco et al. 2001; Kohli et al. 2004). A major problem with such an autonomous system is that transpositions of the Ac element cannot be stabilized, due to the continuous production of transposase (AcTۤase) which catalyzes excision/reinsertion of the element. Moreover, multiple excision/reinsertion events across generations create excision footprint mutations which are no longer tagged by the element. More sophisticated double component systems where the transposase source is dissociated from a non-autonomous Ds element, and may include useful positive-negative selection
markers are therefore preferred (Figure 1). In these systems the AcTpase source and immobilized $Ds$ are introduced linked on the same T-DNA (Chin et al. 1999; Upadhyaya et al. 2002; Greco et al. 2003) or separated on two T-DNAs and combined in the same line by crossing (Nakagawa et al. 2000; Kolesnik et al. 2004). In both cases, the $Ds$ element and the AcTpase source can be segregated away from each other in subsequent generations.

Transposition behaviour of $Ds$ across generations of rice transformed with a single T-DNA/double component approach has been recently analyzed in detail (Greco et al. 2003). The frequency of transposition is usually high among primary regenerants (62%), with early transpositions occurring during the transformation/regeneration process which are generally incorporated in the germline as well as late transpositions that can also be germinally inherited. There is therefore an advantage to propagating callus tissue or ratooning primary transformants for gene-

![Diagram of double-component Ds-tagging vector](image)

**Figure 1.** An example of double-component $Ds$-tagging vector used in the EU consortium project Cereal Gene Tag (Greco et al. 2003). The *gfp* gene encoding the green fluorescent protein reports excision of the $Ds$ element and presence of the “empty donor” T-DNA. Expression of the *su-1* gene confers sensitivity to the proherbicide R7042 (its product converts R7042 to the herbicide sulfonylurea) and can be used as negative selection marker for eliminating plants still producing the Ac transposase (*AcTpase*). The $Ds$ element is equipped with a β-glucuronidase (*gusA*) enhancer trap and the *bar* selectable gene conferring tolerance to the herbicide ammonium glufosinate which can be used for selecting plants containing the $Ds$ element. Use of appropriate restriction enzymes (e.g. *EcoRI*) and probe (*gusA*) in DNA blot analysis allows confirmation of excision (absence of 3.3 kb signal) and reinsertion (presence of signals larger than 2.3 kb) and number of $Ds$ inserts which can be germinally inherited. LB and RB: left right and borders of the T-DNA, respectively. RJ and LJ: right and left junctions of the $Ds$ element, respectively.
rating more independent germinal insertions. The frequency of actively transposing families in T₁ generation is usually high (83%), whereas retention of activity declines to 32% in T₂ generation, attributed to silencing of the AcTpase gene. However, the autonomous Ac element (Greco et al. 2001) or a Ds element mobilized in trans by an AcTpase source introduced by crossing (Kolesnik et al. 2004) appear to be less prone to this silencing phenomenon. Early silencing of the AcTpase gene is favourable for stabilization of the inserts but not for amplification of copies of the element, which may occur when its transposition takes place during replication, and would eventually allow a reduction of library size. Engineered Ds may be equipped with gene detection and/or activation systems (see below).

An advantage of the transposon system over retrotransposons and T-DNA is that the element can be remobilized to generate a revertant phenotype, thereby avoiding the tedious complementation step, create new mutant alleles in a single gene, or locally saturate a target chromosomal region with insertions. On the other hand, non-transmitted somatic insertions and untagged mutations due to Ds excision footprints are important limitations of this system. Recovery of genomic regions flanking insertion sites is easy due to the clear cut and paste excision/reinsertion mechanism. Though variable, depending on the scheme used for stabilizing the insertions over generations, redundancy of sequenced flanking regions due to germline sharing of inserts among sibling plants may also reduce the efficiency of large scale FST generation (van Enckevort et al. 2005).

**Tos17**

*Tos17* is a Ty1-copia retrotransposon of rice belonging to a 32 member family, of which five members have been found active (*Tos10, 17, 19, 25 and 27*) (Hirochika et al. 1996) (Figure 2). *Tos17* is the most active, exists in low copy number and its transposition is repressed at the transcriptional level during normal culture conditions. New copies resulting from a copy-and-paste mechanism of transposition cumulate as a function of cell culture duration. New *Tos17* copies generally insert at unlinked loci from the endogenous active copy(ies) throughout the genome due to retrotranscription of template mRNA in the cytoplasm and further importation of cDNA to the nucleus for integration in chromosomes. More than ten new inserts have been observed per regenerated line, thereby limiting library size, and are stably transmitted to progeny (Hirochika 2001). T-DNA plants also contain new *Tos17* inserts (3.2 on average in cv. Nipponbare; Bourgeois et al. unpublished) when the cultivar used for transformation contains an active *Tos17* element.

Recovery of genomic regions flanking insertion points is easy due to the mechanism of insertion. However, the presence of resident copies and of multiple new copies may complicate further dissociation of insertions during sequencing of flanking regions and molecular and genetic analyses. The shared origin of cells giving rise to regenerated plants leads to a redundancy of some copies which may make sequencing unproductive (e.g. only 16,784 independent sequences have been obtained from
42,292 sequenced products; Miyao et al. 2003). Utility of Tos17 in reverse and forward genetics has so far been most exemplified in rice through the analysis of several mutants. A high frequency of untagged mutations (95%) resulting from other sources such as somaclonal variation during tissue culture, however, limits its large-scale use in forward genetics. The non transgenic feature of Tos17 lines is a clear comparative advantage over T-DNA and Ds rice lines as far as seed increase, evaluation under field conditions and international exchanges are concerned.

**T-DNA**

A decade of gradual improvements in Agrobacterium-mediated transformation methodologies, from the root explant (Valvekens et al. 1988) and seed transformation (Feldmann and Marks 1987) methods to the floral/whole plant dip techniques (Bechtold et al. 1993), has proved necessary to obtain an output of T-DNA plants compatible with genome-wide insertional mutagenesis in Arabidopsis. Though rice appears to be the cereal species most amenable to transformation (Hiei et al. 1994), high-throughput transformation procedures for rice functional genomics were only recently established. An optimized rice transformation procedure (Sallaud et al. 2003) now attains efficiencies 5-20 times higher than those previously reported for japonica rice with 1-10 transgenic plants produced per co-cultured callus, depending on the cultivar. T-DNA insertions are chemically and physically stable over generations, may carry powerful gene detection and/or activation systems and a Ds element (Figure 3), and are phenotypically tagged through the expression of selectable (and possibly reporter) gene(s). T-DNA is integrated in low-copy number (average of two copies at 1.4 locus per line) thereby facilitating further genetic and molecular analyses, but also necessitating a large library to ensure genome saturation. Some drawbacks of T-DNA insertional mutagenesis are that the mutations

![Figure 2. Structure of the Ty1 copia retrotransposon of rice Tos17 (Hirochika et al. 1996).](image-url)
found in these lines are frequently untagged due to abortive integration of T-DNA and/or somaclonal variation, and the often complex organization of T-DNA inserts, which includes concatenated and/or truncated copies and/or binary vector sequences, may make sequencing of up to 40% of flanking regions unproductive (Sallaud et al. 2004).

Utilization of other transposon systems such as the En/Spm maize element or the first discovered DNA transposon active in rice, a 430 bp tourist-like MITE lacking ORF called mPing, can be envisaged as possibility for gene tagging in rice. En/Spm has been shown to be transcribed in rice, and although the first trial reported only somatic transpositions (Greco et al. 2004), use of larger TIR regions in the transformation construct has now allowed the system to work efficiently and produce germinally-inherited insertions (www-plb.ucdavis.edu/Labs/sundar/rice/sequence.html). Miniature Ping (mPing) was recently identified in three independent studies through genome sequence analyses (Jiang et al. 2003) analysis of mutability of the slender glume locus (Nakazaki et al. 2003), and observation of its activation by anther culture (Kikuchi et al. 2003). Copy number of mPing in the japonica genome ranges between 60 and 80 by database searches and Southern analyses. This number is unusually low as the copy number of other MITEs is typically in the thousands. mPing appears to be a recent deletion derivative of a 5,353 bp element called Ping, consisting of a mPing sequence separated by two ORFs, the second ORF most likely encoding a transposase. However, Ping appears not to be activat-

![Figure 3](image_url)

**Figure 3.** Example of T-DNA vectors equipped with a gusA enhancer trap (A) and a gusA enhancer trap with a Ds element (B). Southern blot analysis using DNA digested by a restriction enzyme which cuts once in the T-DNA allows the determination of the minimum number of copies (2.2 on average) integrated at one to several loci (1.4 on average).
ed under the same conditions favouring mPing excision (anther culture and gamma ray irradiation) whereas another related element, called Pong, was indeed found to be active under conditions favoring mPing excision and would provide the transposase source in trans (Jiang et al. 2003). A reverse genetics search in DNA from 600 anther-derived calli for insertion in 20 sequences allowed for the identification of an insertion at the waxy locus (Kikuchi et al. 2003), suggesting that the system is effective for gene tagging.

**Forward genetic screens for loss/gain of function**

Forward genetics classically proceeds from the identification of constitutive or conditional mutant phenotypes under standard or altered culture conditions, to the molecular establishment of tagging by the insertional mutagen and, ultimately, isolation of the disrupted gene. This is generally followed by complementation with the wild type gene through crossing or transformation for restoration of wild type phenotype. Systematic forward genetic screens conducted in *Arabidopsis* have shown that only a minority of phenotypes are tagged even when no tissue culture step is involved in the generation of mutant libraries. For instance, only 350 tagged embryo-defective mutants (20.5%) were identified after analysis of 1,700 defective seed phenotypes screened from a collection of more than 120,000 T-DNA insertion lines of the GARLIC library (Mc Elver et al. 2001). Similar frequencies were observed from a comparable screen conducted for the same traits in the Versailles T-DNA library (M. Devic, personal communication).

Large sets of data on tagging frequency are not yet available in rice, but one might speculate that the frequency may even be lower than that for *Arabidopsis* since Tos17 and T-DNA lines are generated through tissue culture procedures which, although limited in terms of duration and subcultures, are known to generate undesirable somaclonal variation. Only 5-10% of the mutations identified in a Tos17 library were found to be caused by insertion of the retroelement (Hirochika 2001). Due to the large size and limited number of seeds produced by the rice plant compared to *Arabidopsis*, and the need to propagate libraries under field conditions for evaluation of agronomic traits (which is a constraint in the case of transgenic lines), systematic screens of whole insertion libraries will likely remain relatively limited in rice. However, several Tos17, Ds and T-DNA libraries are being field propagated in Japan, China, Korea and Colombia, and should yield informative results about the comparative frequency of morphological and physiological alterations effectively tagged by each mutagen. In the case of the Génoplante T-DNA insertion line (Sallaud et al. 2004) which is being propagated at the Centro Internacional de Agricultura Tropical (CIAT) in Colombia, significant numbers of lines exhibited a phenotype for response to fungal pathogens (1.5%; J.B. Morel et al.,
unpublished), seed–related traits (1.5%; P. Perez et al., unpublished), and morphological and physiological traits (20%; M. Lorieux and J. Tohme, unpublished). Despite the rather low expected frequency of tagging, several genes have already been identified in rice following forward genetics screens. For instance, the rather common viviparous mutation (precocious germination), has been observed in 1,400 of 30,000 Tos17 lines. Perfect co-segregation of the mutation was established with one of the transposed Tos17 copies in seven out of 100 mutant lines tested yielding a tagging frequency of 7%. A homozygous mutant exhibited high water loss, no ABA accumulation after drought stress and was found to be deficient in ABA biosynthesis. Isolation and sequencing of the mutated gene revealed that the mutant is defective in the ortholog gene of Arabidopsis, *OsABA1* encoding the zeaxanthin epoxidase, an enzyme that converts zeaxanthin to all-trans-violaxanthin in the ABA biosynthesis pathway (Agrawal et al. 2001). Other mutations created by Tos17 inserts include mutants altered in the *CesA* 4, *CesA* 7 and *CesA* 9 cellulose synthase genes, exhibiting a brittle culm phenotype (fragility of the culm and mature leaves) due to thinner cell walls in cortical fibers (Tanaka et al. 2003). Another Tos17 mutant exhibited high sterility due to excess production of male and female sporophytes leading to disorganized cell walls and appeared to be caused by alteration of *msp1*, a LRR receptor-like kinase determining the number of cells entering sporogenesis and initiating anther cell wall formation (Nonomura et al. 2003). Very recently a late flowering T-DNA mutant, defective in internode elongation, was found to be altered in the *Osmads50* gene, a major flowering activator acting upstream of other AP1 Mads box gene families (Lee et al. 2003). Further overexpression and suppression of *Osmads50* resulted in extremely early and late flowering phenotypes, respectively, and confirmed the important role of the gene. As to the *Ds* element, isolation and characterization of a tagged rice mutant defective in anther dehiscence has also recently been reported (Zhu et al. 2003).

**Activation tagging for gain-of-function mutants**

A limitation learnt from the Arabidopsis experience is that loss-of-function screens rarely identify genes that act redundantly. In addition, genes required during multiple stages of the life cycle and whose loss-of-function results in early embryonic or in gametophytic lethality are difficult to identify (Weigel et al. 2000). Activation tagging has thus been described as an alternative method to isolate genes through the use of inserts carrying strong activating sequences that can quantitatively modify the transcription of genes adjacent to insertion sites, while still retaining their original expression pattern. Activation tagging has been shown functional in Arabidopsis using multimerized transcriptional enhancer sequences from the well characterized CaMV35S promoter (the -343 to -90 fragment) carried by T-DNA (Weigel et al. 2000) or the En/Spm element (Marsch-Martinez et al. 2002), in allowing the isolation of 30 and 31 dominant mutants from 30,000 lines and 2,900 insertions, respectively. In the first
study, overexpressed genes were nearly always found adjacent to the inserted CaMV35S enhancers at distances ranging from 380 bp to 3.6 kbp. This indicates that in small-sized genomes such as that of Arabidopsis, 20,000-30,000 activation tagging insertion lines are sufficient to have a reasonable chance to activate most genes. The possibility of randomly enhancing gene expression through T-DNA-mediated activation tagging has been demonstrated in rice and, to date, 60,000 insertion lines harbouring this system have been generated (Jeong et al. 2002). In this study, four out of ten randomly chosen candidate lines were found to exhibit enhanced expression of nearby genes separated by a distance of 1.5 to 4.3 kb from the enhancer elements, while still maintaining their original expression pattern. These results are comparable to those observed in Arabidopsis and the high efficiency of transformation now routinely attained in rice (see section on “T-DNA”) would thereby allow for the activation of nearly all rice genes in a single T-DNA activation tagging experiment. Indeed, 1,000 embryo-derived embryogenic calli induced from 200 seeds should yield 20,000 hygromycin-resistant cell lines that could be screened in vitro for tolerance at the cell level and/or regenerated to plants.

**Gene detection systems**

**Gene traps and enhancer traps**
To increase chances of detecting genes interrupted by an insertional mutagen or situated in the vicinity of an insertional mutagen, T-DNA or engineered transposable elements can be equipped with a gene or enhancer trap (Figure 4). An enhancer trap (ET) typically consists of a reporter gene fused to a minimal promoter (MP; e.g. -48 bp from the CaMV35S promoter, containing the TATA box and a transcriptional start site) which is not transcriptionally active but whose transcription can be triggered by neighbouring chromosomal enhancer elements (Springer 2000). ET insertions tend to result in a high frequency of gene detection but often may not correspond to disruption of the detected genes. A gene trap (GT) contains a promoterless reporter gene whose expression occurs only when insertion lies within a transcriptional unit and in correct orientation. The presence of one or more splice acceptor sites aligned in all reading frames preceding the reporter gene allows expression of the reporter if insertion occurs in an intron. Frequency of expression is generally lower than that observed with ET, but corresponds to insertion within genes and most likely to knock-outs. When an insertion occurs in an intron or an exon, translational fusion are generated between the reporter gene and upstream exons of the interrupted gene (Springer 2000). This can generate translational fusions which may provide information about protein localization. Sundaresan and coworkers (1995) compared the frequencies of trapping revealed by GUS activity among insertion line populations.
harbouring either a GT or ET $Ds$ element and found relative values of 26 and 48%, respectively. Though such thorough comparison is not yet available in rice, frequencies of gene detection based on GUS activity typically fell in a 1.6-8% and 20-30% range for GT and ET, respectively, in independent studies using either $Ds$ (Chin et al. 1999) or T-DNA (Jeon et al. 2000; Gay et al. unpublished) inserts. Isolation of trapped genes, which is more straightforward in the case of GT and is clearly facilitated by the simple organization of $Ds$ inserts, is underway in rice. Efforts are also being made to use the AT/GT reporter systems to identify genes induced in response to particular stresses. This approach has very recently proven successful to isolate osmotic stress-responsive genes in *Arabidopsis* (Alvarado et al. 2004) and cold stress-responsive genes in rice (Lee et al. 2004).

**Figure 4.** Enhancer and gene trapping systems equipping the insertion mutagen for gene detection. A: *gusA* enhancer trap B: *gusA* gene trap C: *gal4:Gfp* enhancer trap (see details in the text).
**GAL4:UAS enhancer trap system**

A modification of the classical enhancer trap includes a modified yeast transcriptional activator GAL4 gene (GAL4-VP16) fused to an MP and within the same construct a *gusA* or *gfp* gene fused to tandemly arrayed upstream activating sequence elements (UAS), which are recognized by GAL4 as binding sites. GUS or GFP activity reports GAL4-VP16 expression since GAL4 controls transcription of the *gusA* or *gfp* reporter genes through the binding to the UAS elements (Springer, 2000). After identification of a particular enhancer trap line with interesting *gusA/gfp* expression, secondary constructs can be introduced with genes of interest fused to UAS elements which will be expressed only in tissues and cell types expressing *gusA/gfp*. This strategy has been used successfully in *Drosophila melanogaster* (Brand and Perrimon, 1993; Phelps and Brand, 1998) and was later applied to *Arabidopsis thaliana*, where a *gal4-vp16* fusion gene with modified codon usage was used in a T-DNA-based enhancer trap system (Haseloff 1999; www.plantsci.cam.ac.uk/Haseloff/Home.html). Recently, the system has been successfully incorporated into rice (www.cambia.org/main/r_fg_4.htm#rice%20transgenomics, Wu et al. 2003; Johnson et al. 2005). Enhancer trapping frequency using the GAL4 system in T-DNA rice insertion lines ranged from 29% with GFP (Johnson et al. 2005, www.plantsci.cam.ac.uk/hibberd/aatj.html) to an unexpectedly high 70% with GUS (Wu et al., 2003). As mentioned above, the *gal4*-based system is unique in that it can be used as a tool to trans-activate any gene of interest fused to UAS elements (Bougourd et al. 2000; Kiegle et al. 2000). The possibility to trans-activate a *gusA* reporter gene fused to UAS elements following its introduction in *gal4:gfp* ET lines has recently been demonstrated (Johnson et al. 2005). Other potential uses of the method include ablation of tissues through expression of a lethal gene or gene silencing in specific cell types.

**Reverse genetics**

Reverse genetics comprises a set of methods designed to create or identify lines with inactivated expression of a particular candidate gene sequence in order to assign a function to that gene. Identification of knock-out (KO) mutants in an insertion line library has long relied on the use of PCR screens for the desired insertion in 1-3D pools of DNA samples representing the entire population, using primers specific for both the insertional mutagen and the target gene. Large scale isolation based on PCR methods (iPCR, TAIL PCR or adapter-PCR) and sequencing of chromosomal regions flanking inserts to create insertion databases is becoming more popular because it allows the direct identification of mutant lines through simple worldwide computer searches in public databases. More than 296,000 *Arabidopsis* FST are currently available, likely allowing for the identification of one
or more insertions in any gene of this model species (March 2004 update, www.arabidopsis.org/links/insertion.jsp). In rice, both PCR-based searches in DNA pools and FST databases have enabled the identification of mutants in sequences of interest among Ac (Enoki et al. 1999), Tos17 (Sato et al. 1999; Takano et al. 2001; Yamaguchi et al. 2004) and T-DNA (An, unpublished) mutagenized populations. However, a major lesson gathered from the Arabidopsis experience is that upon analysis under standard culture conditions, only a very small portion of KO mutant lines exhibit an informative phenotype (Bouché and Bouchez 2001). One explanation for this “phenotype gap” is our inability to detect slight physiological alterations and/or to reveal a phenotype under standard growth conditions. As an alternative to the testing of a wide range of environmental conditions, the accumulation of complementary information on the target sequence, ranging from expression profiling data to localization of the gene product, can help define precise conditions for revealing the phenotypic alteration (Bouché and Bouchez 2001). Another explanation for the phenotype gap lies in the redundancy of gene function: gene duplication is indeed frequent in higher plants and most genes belong to gene families with members existing in dispersed and/or clustered copies throughout the genome. This situation is anticipated to be even more frequent in rice than in Arabidopsis since tandemly repeated sequences are more prevalent in the rice gene complement. As mentioned earlier, an alternative to the creation and stacking of mutant alleles in the various members of a gene family through crossing of lines altered in individual genes (which might prove impossible in the case of tightly linked tandem arrays of family members) is to create global KO lines expressing a dsRNA of conserved sequence motifs shared between family members, thus silencing all family members simultaneously. Alternatively, the mobilization of a Ds element from a nearby launching pad to saturate the tandem array with inserts, or the creation of a large deletion in the corresponding chromosomal region, could help to address this problem.

In rice, a precise function has been assigned to several candidate sequences using insertional mutants identified through reverse genetics strategies. For instance, the function of a knotted-1 homeodomain gene involved in internode elongation, Osh15 (Sato et al. 1999) and the YABBY drooping leaf gene regulating carpel specification and midrib development (Yamaguchi et al. 2004) have been elucidated in Tos17 mutants along with traditional allelic mutations. Many other putative KOs in candidate sequences are under examination and should provide a clue to their function in the near future.

Preferential insertion of mutagens

In order to calculate the number of characterized insertion sites needed to reach genome saturation and have likely KOs in any rice gene, the tendency of an insertional mutagen towards random or gene-dense region integration must be
determined. Assuming random insertion occurs in the 430 Mb rice genome, 471,000 T-DNA lines harbouring an average of 1.4 T-DNA inserts have been estimated necessary to have a 99% probability of knocking out every gene (Jeong and An 2001). However, an important finding emerging from the analysis of FST data is that all three mutagens have clear preferred insertion into low copy, gene-rich regions of the genome and rarely integrate into repetitive DNA. This is an important result because, as 55% of the rice genome consists of repetitive DNA, potentially half the number of lines indicated above could be sufficient to knock-out any rice gene. Ds flanking sequences recovered from stabilized lines and mapped on the physical map revealed a low average frequency of transpositions closely linked to the T-DNA donor sites (6%), interchromosomal transposition preferences and a 72% insertion frequency of the Ds element in genic regions defined as an interval extending 600 bp upstream of the ATG start codon to 600 bp downstream of the stop codon of predicted genes (Kolesnik et al. 2004). Analysis of 16,784 Tos17 FSTs revealed that the retroelement exhibits a three-fold preferential insertion into coding sequences (intron + exon) compared to intergenic regions (Miyao et al. 2003), avoiding retrotransposon-rich pericentromeric regions while preferentially integrating into gene-rich subtelomeric regions of the chromosomes. Hot spots and cold spots for integration were also identified within gene-dense regions. An insertion bias toward disease/defense and signal transduction categories of genes was also observed. Characterization of a large number of T-DNA insertion sites (totalling more than 12,000 FSTs) was reported only recently (An et al. 2003; Chen et al. 2003b; Sallaud et al. 2004). Though conducted on different populations, these three studies consistently revealed preferential insertion into gene-rich regions, low frequency insertion in repetitive DNA and equal frequencies of integration in genic vs. intergenic regions. As with Tos17, a non-uniform distribution was observed along chromosomes with a lower insertion density around the centromere region and a higher density in the subtelomeric regions where gene density is higher. The frequencies of insertion in predicted genes were 47% (Sallaud et al. 2004), 49%, (An et al. 2003) and 58% (Chen et al. 2003b) based on intervals extending between 250, 300 and 600 bp upstream the ATG start codon and downstream the stop codon of predicted genes, respectively. T-DNA inserts were preferentially found in 5’ and 3’ regulatory regions of predicted genes but did not exhibit bias for insertion in any specific functional category of genes (An et al. 2003; Sallaud et al. 2004).

Conclusions and prospects

A wide range of molecular tools are now available for the functional characterization of rice genes. In this tool-box, the insertion line libraries currently being generated and characterized in rice will no doubt contribute to an unprecedented boost
in the discovery and validation of cereal gene functions, especially evident because half of the rice genes have no homolog identified in *Arabidopsis*. Given their intrinsic properties and insertional preferences, the combined utilization of all three mutagens appears desirable to achieve genome saturation with insertion sites. The number of insertion lines being generated worldwide in ongoing projects appears sufficient to reach genome saturation with insertion sites within the next five years. Efforts should now be devoted to the sequencing of the insertion sites, which will serve as the basis for efficient identification of particular lines altered in candidate sequences and of international collaborations to rapidly reach a level of genome saturation comparable to that attained in *Arabidopsis*. Creation of allelic series with EMS-induced point mutations, as well as analysis of collections with deletions of tandemly organized gene members with redundant function, is also of particular importance because of their potential to complement insertion libraries. Development of high-throughput systematic RNA interference using shuttle vectors also appears highly desirable. Finally, the construction of a single database displaying graphical representations of the sequence environment of blast queries, including FSTs with all three mutagens and corresponding mutant phenotypes, stress/organ-related expression profiling data, and a link from the physical map to the rice/cereals genetic/QTL maps through sequenced probes would provide an invaluable tool for hastening gene discovery and validation in rice and other cereal crops.

**Acknowledgements**

The financial support of the French National Genomics Initiative Génoplante and of the EU 5th framework project QLG2-CT-2001-01453 Cereal Gene Tags is greatly acknowledged. We also warmly thank colleagues of these projects for making accessible their data and for valuable discussions: Andy Pereira, Mark Tester, Mathias Lorieux, Christophe Périn, Pascual Perez, Isabelle Barnola, Corinne Petit, Jean-Benoit Morel, Alain Ghesquière and Michel Delseny.

**References**


---

**GUIDERDONI ET AL.**

---


Zhu QH, Hoque MS, Dennis ES, Upadhaya NM. (2003) Isolation and characterization of a Ds transposon tagged rice (Oryza sativa L.) mutant defective in anther dehiscence. BMC Plant Biol 3:6