The Plant Cell, Vol. 11, 2263–2270, December 1999, www.plantcell.org © 1999 American Society of Plant Physiologists

GENOMICS ARTICLE

Analysis of Flanking Sequences from *Dissociation* Insertion Lines: A Database for Reverse Genetics in Arabidopsis

Serguei Parinov, Mayalagu Sevugan, De Ye, Wei-Cai Yang, Mande Kumaran, and Venkatesan Sundaresan[1]

Institute of Molecular Agrobiology, National University of Singapore, 1 Research Link, Singapore 117604

We have generated *Dissociation* (*Ds*) element insertions throughout the Arabidopsis genome as a means of random mutagenesis. Here, we present the molecular analysis of genomic sequences that flank the *Ds* insertions of 931 independent transposant lines. Flanking sequences from 511 lines proved to be identical or homologous to DNA or protein sequences in public databases, and disruptions within known or putative genes were indicated for 354 lines. Because a significant portion (45%) of the insertions occurred within sequences defined by GenBank BAC and P1 clones, we were able to assess the distribution of *Ds* insertions throughout the genome. We discovered a significant preference for *Ds* transposition to the regions adjacent to nucleolus organizer regions on chromosomes 2 and 4. Otherwise, the mapped insertions appeared to be evenly dispersed throughout the genome. For any given gene, insertions preferentially occurred at the 5′ end, although disruption was clearly possible at any intragenic position. The insertion sites of >500 lines that could be characterized by reference to public databases are presented in a tabular format at http://www.plantcell.org/cgi/content/full/11/12/2263/DC1. This database should be of value to researchers using reverse genetics approaches to determine gene function.

INTRODUCTION

Rapid progress in global Arabidopsis genome sequencing projects has underscored the need for functional studies of the genome. It is currently estimated that >90% of all Arabidopsis genes remain functionally uncharacterized. In addition, genome sequencing projects have depended on exon prediction algorithms such that the identification of many genes must be regarded as hypothetical. The application of insertional mutagenesis is thus an attractive approach for functional genomics that minimizes the number of steps required to conceptually link a given gene to its function (Spradling et al., 1995). Insertional mutagenesis in Arabidopsis has become particularly attractive due to the development of effective methods for amplification of sequences adjoining the insertion (Liu and Whittier, 1995; Devic et al., 1997; Hui et al., 1998). In this way, disrupted genes can often be readily identified upon reference to published gene sequences. Indeed, with the current availability of Arabidopsis genome sequences from public databases, half of all potential insertions can be matched to sequenced regions of the genome and placed on a physical map. Insertional mutagenesis is especially effective for generating gene knockouts in Arabidopsis because of the high gene density (approximately one gene every 5 kb; Bevan et al., 1998, 1999), which means that on average, one out of two insertions results in gene disruption.

We have been generating insertions in the Arabidopsis genome by means of transposable elements (Sundaresan et al., 1995). The system uses a modified maize *Dissociation* (*Ds*) transposable element carrying a β-glucuronidase (*GUS*) reporter gene acting as either a gene trap or an enhancer trap for detection of genes by their expression pattern. Specifically, various starter lines, each containing a single stable *Ds* insertion, were crossed with lines expressing the *Activator* (*Ac*) element transposase so as to initiate transposition of the *Ds* element. Because the parental *Ac* transposase gene had been linked to the *indole acetic acid hydrolase* (*IAAH*) gene, which confers sensitivity to naphthalene acetamide, progeny could be selected that were free of transposase activity. Each selected *Ds* insertion is thus stable but can be later remobilized by the appropriate cross to a line that expresses the *Ac* transposase. Such remobilization of *Ds* elements is a useful property for confirming the mutational effects of insertions. Another advantage of this system is that most of the insertion lines contain single *Ds* elements that are intact, thereby simplifying the molecular and genetic analyses.

[1] To whom correspondence should be addressed. E-mail director@ima.org.sg; fax 65-872-7012.

On the other hand, *Ds* elements preferentially transpose to sites closely linked to the donor site (Smith et al., 1996; Machida et al., 1997), which is a handicap that must be circumvented to saturate the genome with random insertions. In our system, therefore, the *Ds* element in the T-DNA donor site is also linked to the *IAAH* gene, used in this instance as a counterselectable marker that, in the presence of naphthalene acetamide, eliminates progeny retaining the T-DNA donor site (Sundaresan et al., 1995; see Figure 1). At the same time, plants representing a transposition event can be selected by virtue of the kanamycin resistance gene contained within the *Ds* element. In this way, *Ds* elements that remain closely linked to the T-DNA donor site by virtue of a proximal transposition event can be excluded, and only transposition to more distal sites in the genome results in plants that will survive the selection regime. Here, we evaluate the possibilities of this system for functional genomics by sequence analysis of the insertion sites of 931 independent transposant lines.

RESULTS

We determined the genomic sequences flanking *Ds* insertions from 931 lines, each representing an independent germinal transposition event. Sequences flanking *Ds* insertions were amplified using a thermal asymmetric interlaced polymerase chain reaction protocol (TAIL-PCR; Liu et al., 1995). This method was found to be very effective. For 95% of the lines, at least one flanking fragment longer than 250 bp was obtained (data not shown). Generally, two PCR products (usually the sequences flanking the 3′ and 5′ ends of the *Ds* element) were sequenced and then subjected to BLAST searches (Altschul et al., 1990, 1997) of the NCBI GenBank (Benson et al., 1998) and Arabidopsis GenBank (Flanders et al., 1998) databases.

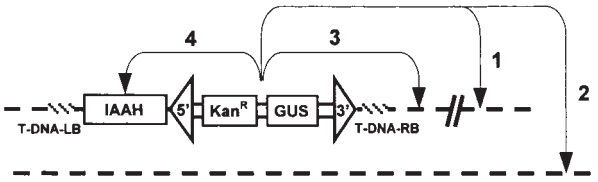

Figure 1. Schematic Diagram of the *Ds* Donor Site and Possible Transposition Events.

Open arrowheads indicate the 5′ and 3′ ends of the transposon. The *Ds* element carries the *NPTII* gene, which confers resistance to kanamycin (Kan^R), and a modified *GUS* reporter gene (Sundaresan et al., 1995). Possible transposition events include the following: (1) unlinked or loosely linked transposition to the same chromosome; (2) transposition to a different chromosome; (3) closely linked transposition; and (4) closely linked transposition disrupting the *IAAH* gene. Hatched boxes represent left and right T-DNA borders (T-DNA-LB and T-DNA-RB, respectively).

Analysis of the sequences that flank insertions of the *Ds* element revealed several classes of insertions (Table 1). Useful sequence information has been obtained in 85% of the transposant lines investigated. Analysis of the remaining 15% yielded sequences that correspond to the donor T-DNA construct. These statistics do not include sequences resulting from seed and PCR cross-contamination, which is estimated to have occurred at a rate of ~70 spurious sequences per 1000 analyses.

More than half (511/931) of all flanking sequences identified are either identical or significantly similar to sequences represented in public databases. Three hundred fifty-four insertions disrupt sequences that correspond to either known or putative structural genes (Table 1) as identified by the analysis of the sequenced genomic BAC and P1 clones (Bevan et al., 1998; Sato et al., 1998). Fifty-three insertions occur within sequences identical to expressed sequence tags that do not match any genomic sequence in the Arabidopsis database. Finally, 46 lines appear to carry insertions into unsequenced genes, as indicated by comparisons of flanking sequences (regarded at the amino acid level) with GenBank protein sequences. A substantial fraction of the flanking sequences (~30%) fails to match, at either the nucleotide or protein level, any sequences in the public databases. However, we can expect the DNA sequences disrupted by these insertions to become available as the Arabidopsis genome sequencing projects progress.

The genes that we found to be disrupted can be classified into all of the major categories described by Bevan et al. (1998). The greatest numbers of these disrupted genes are involved in transcription (11%), signal transduction (8%), and metabolism and energy transduction (15%). We were unable to predict biochemical functions for 33% of the genes. At least 15 groups of insertions represent members of various gene families (e.g., pectin esterases, expansins, peroxidases, and MYB-like proteins). Because many of such genes have potentially redundant functions, the corresponding insertion lines could be used for further studies by generating multiple mutants.

DISCUSSION

Distribution of *Ds* Insertions

Despite the extensive use of *Ac-Ds* transposable elements in various plant hosts for insertional mutagenesis, the lack of systematic large-scale mapping data and the high frequency of short-range transpositions have obscured details of transposition hot spots outside of the donor site. In this study, the physical map positions of 356 *Ds* insertions have been established by comparing their flanking sequences with mapped BAC and P1 clones. This number corresponds

Table 1. Categories of *Ds* Insertions

Analysis of Flanking Sequence[a]	Number of *Ds* Insertions	Percentage of Total Insertions
Positive BLAST results	511	55%
Insertions disrupting genes encoding proteins	354	38%
Insertions into genes with known DNA or mRNA sequence	255	
Expressed sequence tag match only	53	
Partially related to GenBank proteins (including putative and hypothetical proteins)	46	
Insertions into rRNA and tRNA genes	12	1.3%
Identical to genomic clones (negative BLASTX with GenBank protein sequences)	145	15.5%
Negative BLAST results (no similarity)	279	30%
Useful information	790	85%
T-DNA construct	141	15%
Total characterized lines	931	100%

[a] This table summarizes results of analysis of flanking sequences from 931 lines. Insertions into genes encoding proteins were ordered into three classes, based on their identification. The first class represents identity to genes whose complete DNA sequence is available in the form of genomic sequence or cDNA sequence. The second class represents flanking sequences identical only to expressed sequence tag sequences without matches to genome sequences. The third class of gene disruptions has been identified solely by homology of the translated flanking sequences with proteins in GenBank. Sequencing data represent the status of the database as of April 2, 1999.

to 45% of all useful insertions sequenced (356/790; see Table 1), which is comparable to the fraction of the sequence of the Arabidopsis genome available from public databases at the time of analysis. Figure 2 shows the distribution of these insertions. For purposes of illustration, we used a public Arabidopsis Sequencing Map (from http://genome-www3.stanford.edu/Arabidopsis) to visually filter out those regions of the genome not yet sequenced.

Two transposition hot spots are apparent at the narrow regions adjacent to nucleolus organizer regions NOR2 and NOR4. A general increase in insertion frequency with increasing proximity to the NOR2 and NOR4 further suggests a positive influence of NORs upon transposition frequency. A closer view of the region immediately adjacent to NOR4, covering ~300 kb, is offered in Figure 3. There is no obvious specificity for insertions within this region, which suggests that it is the proximity of the NOR, rather than the presence of particular sequence cues within the adjacent region, that is responsible for the observed effect on transposition.

Table 2 indicates the fractions of NOR-adjacent insertions arising from various *Ds* starter lines. The relative use of the starter lines in our study was as follows: DsG1, 48%; DsG6, 42%; DsG8, 7%; and DsE, 2%. We have determined precise map positions for the DsG1 and DsG6 donor sites by comparing their flanking sequences with GenBank DNA sequences. The DsG1 donor site is located on chromosome 2 BAC T29F13, whereas the DsG6 donor site is within the coding region of the *FAD7* gene assigned to chromosome 3 (Iba et al., 1993).

The data presented in Table 2 demonstrate that the insertional preference for the NOR4-adjacent region is not due to linkage to any particular donor site. However, a substantial number of the transpositions into NOR2 originated from the DsG1 donor site, which may possibly be due to a higher frequency of intrachromosomal transpositions. Similar observations were made with *Ac* element transposition in maize (Dooner et al., 1994), where many genetically unlinked transpositions in fact proved to have occurred within the same chromosome as that occupied by the donor site. The overall lower number of hits into the NOR2-adjacent region as compared with the NOR4-adjacent region could be due to an unsequenced gap between NOR2 and the closest sequenced BAC clone.

Both NOR2 and NOR4 adjoin the telomeres of the chromosomes on which they reside (Copenhaver and Pikaard, 1996a, 1996b). Each NOR occupies 3.5 to 4.0 Mb and consists of tandemly repeated rRNA gene clusters. The nucleolus is organized around the NORs during interphase and is associated with very active transcription of ribosomal genes by RNA polymerase I. The increasing frequency of insertions into the NOR-adjacent regions could thus be due to higher accessibility of the chromatin in this region to transposase. Alternatively, proximity to the nucleolus could somehow result in a more favorable chromatin structure for *Ds* integration. In contrast to the preference for the NOR-adjacent region, only nine insertions (i.e., 1% of all insertions) have been found to occur within rDNA sequences, which is much less than the 6% contribution of rRNA genes to the Arabidopsis genome. The compartmentalization of rDNA within the nucleolus may be an important factor in restricting rDNA from *Ds* transpositions. We cannot, however, rule out the possibility that for some reason there is a lower efficiency of rDNA amplification and detection by TAIL-PCR so as to result in an apparently lower frequency of rDNA insertions.

Despite the use of the *IAAH* gene to select against linked transpositions, the frequency of insertions near the DsG1 donor site is high. Of all transposants arising from the DsG1 starter line, 9% manifest insertions within 1 Mb of the donor

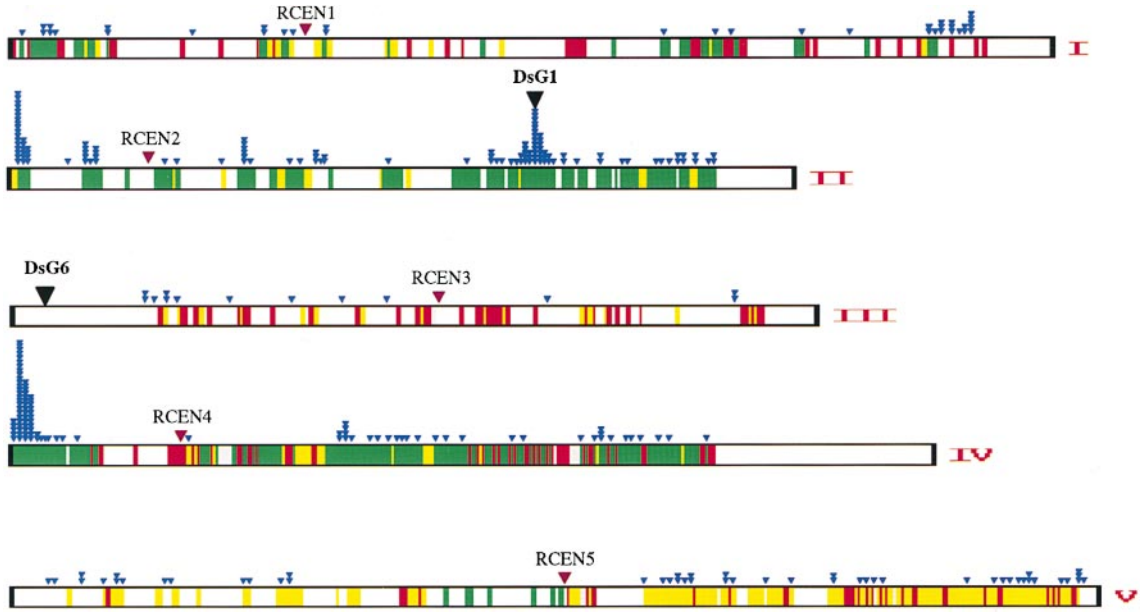

Figure 2. Distribution of *Ds* Insertions on the Arabidopsis Sequencing Map.

The positions of 312 *Ds* insertions are shown. Small arrowheads represent insertion sites. Insertions in the same BAC or P1 clone (average size, ~100 kb) are stacked together to form a single column. Therefore, this insertion map manifests resolution equal to the average size of one BAC clone. The three major hot spots correspond to NOR2- and NOR4-adjacent regions and the region surrounding the DsG1 donor site. The Arabidopsis sequence map was modified from the *Arabidopsis thaliana* Database at Stanford University (http://genome-www.stanford.edu/Arabidopsis). Sequencing data represent the status of Arabidopsis genome sequencing on April 2, 1999. White regions indicate unsequenced regions; green and yellow represent completed sequences available from GenBank; and red indicates sequencing in progress.

site, and one-third of these correspond to the same BAC clone. Nevertheless, a frequency of 9% is much less than the frequency of 25 to 50% obtained in the absence of counterselective measures (Smith et al., 1996; Machida et al., 1997). Several other minor insertional hot spots were noted—for example, BAC T26I20 on chromosome 2 and in the region of chromosome 1 surrounding the *gl-2* locus—that are not due to linkage to any donor site. Further sequencing of genomic DNA that flanks *Ds* insertions is necessary, however, to determine the exact nature of such hot spots.

Insertions into the Donor T-DNA

Another indication of the high frequency of short-range transpositions is the number of insertions that we have observed within the donor T-DNA itself. Of the lines analyzed in this study, 15% carry *Ds* elements inserted somewhere within the donor T-DNA construct (Figure 1); therefore, they provide no useful genomic information (Table 1). Flanking sequences from approximately half of these insertions correspond to the *IAAH* gene. Another fraction of the donor site insertions carry flanking sequences corresponding to the T-DNA border sequence adjacent to the 3′ end of the *Ds* element in the starter line. Various inversions or truncations at or near (within 1 to 30 bp) the border of the donor *Ds* element similarly reflect local transposition events that could in fact lead to partial truncation of the adjacent *IAAH* gene. Because *IAAH* is used as a negative selection marker against linked transpositions, its inactivation by transposition or rearrangement leads to a set of "background" lines, which is an inevitable drawback of this selection scheme (Sundaresan et al., 1995). The introduction of two copies of the *IAAH* gene into the donor T-DNA construct could possibly be used to minimize the accumulation of such "background" insertions.

Preferential Insertion of the *Ds* Element at the 5′ Ends of Genes

There is considerable evidence that certain transposable elements, such as the Drosophila *P* and yeast *Ty1* elements, preferentially insert into genes at upstream regions (Liebman and Newnam, 1993; Spradling et al., 1995). To determine whether *Ds* elements exhibit any similar preferences, we performed an alignment of insertion sites from our database, which is shown in Figure 4. Specifically, we plotted the

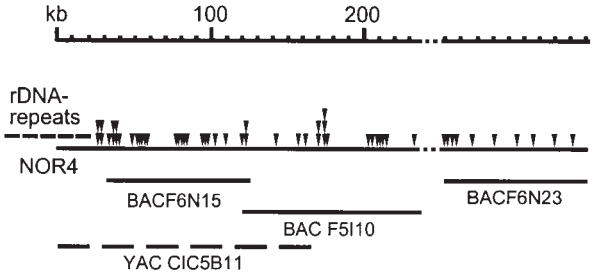

Figure 3. Distribution of *Ds* Insertions in the NOR-Adjacent Region on Chromosome 4.

Insertions are represented by solid triangles. The region shown includes the overlapping BAC F6N15 (25 insertions) and BAC F5I10 (16 insertions). YAC CIC5B11 covers almost the whole region and additionally carries ~26 kb of rDNA repeats of NOR4. Precise distance from BAC F6N23 (10 insertions) to F5I10 is unknown (BAC F19J24 overlapping both clones is not yet sequenced). YAC CIC5B11 (accession number AC004708) is indicated by a broken line because it does not align perfectly with BAC F5I10. Insertions separated by <1 kb are stacked on top of each other.

physical distances of 232 individual insertion sites (not necessarily intragenic) to the closest ATG start codon. Only those insertions that occurred within a range that spanned from 500 bp upstream to 3.5 kb downstream of ATG start codons were plotted; ATG start codons in this case included those based on predicted gene sequences from annotated genomic clones. Insertions into regions with ambiguous shadow exons were not taken into account. The data in Figure 4A suggest a preference for insertions into the 5′ ends of genes, but no other preferences are evident. Because not all existing gene prediction algorithms are perfect, our assessment of insertion distances from 5′ ends is subject to an error rate associated with predictions of open reading frames. However, as seen in Figure 4B, the alignment of insertion sites from a smaller number of well-characterized genes (i.e., genes for which the ATG codon is known with certainty) demonstrates a similar distribution. Indeed, the sole preference appears to be a bias for the 5′ end of the given gene.

Database of *Ds* Insertion Lines

Generating loss-of-function mutations in a specific gene is a cumbersome process that is frequently the rate-limiting step in functional genetic analysis. Current methods of insertional mutagenesis generally require the screening of a large number of lines for insertions within a given gene. To simplify this process, large-scale sequencing of random insertions has been initiated in Drosophila and mice (Spradling et al., 1995; Townley et al., 1997; Zambrowicz et al., 1998). In Arabidopsis, this strategy is particularly useful because genes occupy more than half of the genome, and the entire genome sequence is expected to become available within the next year (Kotani et al., 1997; Bevan et al., 1998). Alternatively, large-scale pooling strategies for PCR screening of transposant libraries have been reported by Tissier et al. (1999).

We have organized the positional information from our set of >500 *Ds* insertion lines into a database, which includes the analysis of disrupted genes, usually along with their map positions (see http://www.plantcell.org/cgi/content/full/11/12/2263/DC1). A sequenced insertion in the database typically corresponds to a stable single insertion line (unpublished results; see also Sundaresan et al., 1995), which can be further inspected for *GUS* staining and mutant phenotype. Such a combination of genetic and functional data makes this reverse genetics database a potentially useful

Table 2. Linkage between Insertions in NOR-Adjacent Regions and Donor Sites

Insertion Site According to Starter Line	Number of Insertions	Percentage of Total (780)
NOR4-adjacent region[a]	54	7%
DsG1 (chromosome 2)	26	
DsG6 (chromosome 3)	18	
DsG8	4	
DsE	6	
NOR2-adjacent region	23	3%
DsG1 (chromosome 2)	17	
DsG6 (chromosome 3)	4	
DsG8	2	
rDNA repeats	9	1%

[a] Numbers of insertions into NOR2- and NOR4-adjacent regions (~400 kb each) and rRNA genes (a total of ~7 Mb in the haploid Arabidopsis genome) are shown. Various starter lines, shown as DsG1, DsG6, DsG8, and DsE, were used to generate the transposants (see text). The two major lines used, DsG1 and DsG6, accounting for 90% of the transposants, have been mapped on chromosomes 2 and 3, respectively; their relative contributions are 48 and 42%. DsE corresponds to three different starter lines—DsE1, DsE2, and DsE3—which together contributed 2%.

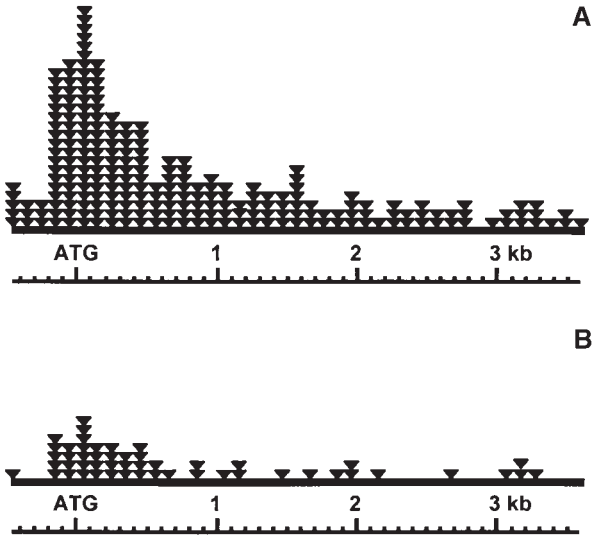

Figure 4. Distribution of *Ds* Insertions within Genes Suggests Preference for 5′ Ends.

(A) Insertion sites were plotted according to their distance to the closest ATG start codon (including hypothetical genes). Insertions <500 bp upstream and 3500 bp downstream of the ATG start are shown. Each triangle indicates the insertion position for a single independent insertion line.

(B) A similar analysis to **(A)** in which only well-characterized genes (i.e., the complete mRNA and genomic sequence has been published) are considered, so that the indicated ATG start codon is more likely to be accurate.

tool for functional genomics. All lines in this database have been sent to the Nottingham Arabidopsis Stock Centre and will be made available for noncommercial research purposes upon request. The database is currently available at http://www.plantcell.org/cgi/content/full/11/12/2263/DC1 and will also be released at http://nasc.nott. ac.uk/ima.html so as to permit wide access.

Use of *Ds* Insertion Lines for the Mutagenesis of Closely Linked Genes

According to Smith et al. (1996), half of all *Ds* transpositional insertions occur within 30 centimorgans of the donor site, with 25% of such transposition events transpiring within a range of 1 Mb and 10% occurring within a range of 200 kb. In a separate system, Machida et al. (1997) reported that 50% of transposition events can occur within 1.7 Mb, and 35% within 200 kb, of the donor site. These high frequencies of short-range transpositions can be effectively utilized for the targeted mutagenesis of closely linked genes by crossing *Ds* insertion lines to plants expressing the *Ac* transposase (Sundaresan, 1996). A PCR screen for knockouts can then be performed using pools of DNA from F_2 seedlings, with primers specific both for the gene of interest and for the ends of the transposon. We have successfully identified *Ds* insertional knockouts of genes closely linked to the original integration site by using as few as 200 F_2 families (M. Kumaran, D. Ye, W.C. Yang, S. Parinov, and V. Sundaresan, unpublished results), but more typically, we estimate that up to 2000 F_2 families may be required to be reasonably certain ($P = 0.95$) of recovering a transposition within a 200-kb distance into a 3-kb gene. Nevertheless, the screening of 2000 F_2 families should be feasible for most laboratories, inasmuch as PCR screening processes that utilize pooling strategies (Zwaal et al., 1993; Das and Martienssen, 1995) are not very labor intensive. Potentially, with 1000 to 2000 sequenced *Ds* insertions distributed at 200-kb intervals along the Arabidopsis genome, it should be possible to utilize this strategy to simplify the task of generating mutations in any specific gene of interest.

METHODS

Polymerase Chain Reaction Amplification

We used the Nucleon PhytoPure plant DNA extraction kit (Amersham Life Science) to purify DNA from five to 15 young seedlings of a given line; alternatively, five to 10 young inflorescences were used. Ten to 100 ng of genomic DNA served as template in primary polymerase chain reactions (PCRs) (20 μL). Thermal asymmetric interlaced (TAIL)–PCR was performed according to Liu et al. (1995), with the minor modification that 15 supercycles in the secondary reaction and 30 reduced-stringency cycles in the tertiary reaction were performed. We designed the following nested primers complementary to 3′ and 5′ ends of the *Ds* element: *Ds*3′-1a, GGTTCCCGTCCGATTTCGACT; *Ds*3′-2a, CGATTACCGTATTTATCCCGTTC; *Ds*3′-3a, TCGTTTCCGTCCCGCAAGT; *Ds*5′-1a, ACGGTCGGGAAACTAGCTCTAC; *Ds*5′-2a, TCCGTTCCGTTTTCGTTTTTTAC; and *Ds*5′-3a, CGGTCGGTACGGGATTTTCC. Each of these primers was used in combinations with three arbitrary degenerate primers: AD1, NTCGA(G/C)T(A/T)T(G/C)G(A/T)GTT; AD3, (A/T)GTGNAG(A/T)ANCANAGA; and AD2, (G/C)TTGNTA(G/C)TNCTNTGC (Liu et al., 1995; Tsugeki et al., 1996). Thus, three rounds of nested amplification were made with every DNA sample using six different primer combinations in every round. We found out that even if no specific amplification was detected using this standard protocol, the use of the following alternative set of *Ds*-complementary primers (Grossniklaus et al., 1998) often proved effective: *Ds*3′-1, CGATTACCGTATTTATCCCGTTCG; *Ds*3′-2, CCGGTATATCCCGTTTTCG; *Ds*3′-3, GAAAATGAAAACGGTAGAGGT; *Ds*5′-1, CCGTTTACCGTTTTGTATATCCCG; *Ds*5′-2, CGTTCCGTTTTCGTTTTTTACC; and *Ds*5′-3, CGGTCGGTACGGGATTTTCC.

Purification and Sequencing of PCR Products

PCR products were analyzed on 1.8% agarose gels. Extracted bands were purified using QIAquick Gel Extraction Kit and QIAquick 96 PCR purification Kit (Qiagen, Hilden, Germany). Products were sequenced using ABI Prism dRhodamine Terminator Cycle Sequencing Ready Reaction Kit (PE Applied Biosystems, Foster City,

CA) and an ABI Prism 310 Genetic Analyzer with Data Collection Software (PE Applied Biosystems) supplied by the producer.

Sequence Analysis

Flanking sequences were subjected to BLAST searches of the National Center for Biotechnology Information (NCBI) and the *Arabidopsis thaliana* Database (Stanford University, Stanford, CA). We used gene prediction analysis of BAC clones in GenBank supplied by authors and from Kazusa DNA research institute at http://www.kazusa.or.jp/arabi. We used mapping data from the *Arabidopsis thaliana* Database at http://genome-www.stanford.edu/Arabidopsis and the CSHL genome sequencing center at http://nucleus.cshl.org/protarab.

ACKNOWLEDGMENTS

We thank Sarojam Rajani for contributing some of the lines used in this study, S. Thanumalayan for technical assistance, Megan Griffith for comments on the manuscript, and Ueli Grossniklaus for helpful advice. We are grateful to L.D. Parnell for useful discussions on the NOR4-adjacent region. This research was funded by grants from the National Science and Technology Board of Singapore.

Received July 6, 1999; accepted August 25, 1999.

REFERENCES

Altschul, S.F., Gish, W., Miller, W., Myers, E.W., and Lipman, D.J. (1990). Basic local alignment search tool. J. Mol. Biol. **215,** 403–410.

Altschul, S.F., Madden, T.L., Schaffer, A.A., Zhang, J., Zhang, Z., Miller, W., and Lipman, D.J. (1997). Gapped BLAST and PSI-BLAST: A new generation of protein database search programs. Nucleic Acids Res. **25,** 3389–3402.

Benson, D.A., Boguski, M.S., Lipman, D.J., Ostell, J., and Ouellette, B.F. (1998). GenBank. Nucleic Acids Res. **26,** 1–7.

Bevan, M., et al. (1998). Analysis of 1.9 Mb of contiguous sequence from chromosome 4 of *Arabidopsis thaliana*. Nature **391,** 485–488.

Bevan, M., Bancroft, I., Mewes, H.W., Martienssen, R., and McCombie, R. (1999). Clearing a path through the jungle: Progress in Arabidopsis genomics. Bioessays **21,** 110–120.

Copenhaver, G.P., and Pikaard, C.S. (1996a). RFLP and physical mapping with an rDNA-specific endonuclease reveals that nucleolus organizer regions of *Arabidopsis thaliana* adjoin telomeres on chromosomes 2 and 4. Plant J. **9,** 259–272.

Copenhaver, G.P., and Pikaard, C.S. (1996b). Two-dimensional RFLP analyses reveal megabase-sized clusters of rRNA gene variants in *Arabidopsis thaliana*, suggesting local spreading of variants as the mode for gene homogenization during concerted evolution. Plant J. **9,** 273–282.

Das, L., and Martienssen, R. (1995). Site-selected transposon mutagenesis at the hcf106 locus in maize. Plant Cell **7,** 287–294.

Devic, M., Albert, S., Delseny, M., and Roscoe, T.J. (1997). Efficient PCR walking on plant genomic DNA. Plant Physiol. Biochem. **35,** 331–339.

Dooner, H.K., Belachew, A., Burgess, D., Harding, S., Ralston, M., and Ralston, E. (1994). Distribution of unlinked receptor sites for transposed Ac elements from the *bz-m2(Ac)* allele in maize. Genetics **136,** 261–279.

Flanders, D., Weng, S., Petel, F.X., and Cherry, J.M. (1998). AtDB, the *Arabidopsis thaliana* database, and graphical-web-display of progress by the Arabidopsis Genome Initiative. Nucleic Acids Res. **26,** 80–84.

Grossniklaus, U., Vielle-Calzada, J.P., Hoeppner, M.A., and Gagliano, W.B. (1998). Maternal control of embryogenesis by MEDEA, a polycomb group gene in Arabidopsis. Science **280,** 446–450.

Hui, E.K., Wang, P.C., and Lo, S.J. (1998). Strategies for cloning unknown cellular flanking DNA sequences from foreign integrants. Cell. Mol. Life Sci. **54,** 1403–1411.

Iba, K., Gibson, S., Nishiuchi, T., Fuse, T., Nishimura, M., Arondel, V., Hugly, S., and Somerville, C. (1993). A gene encoding a chloroplast omega-3 fatty acid desaturase complements alterations in fatty acid desaturation and chloroplast copy number of the fad7 mutant of *Arabidopsis thaliana.* J. Biol. Chem. **268,** 24099–24105.

Kotani, H., Sato, S., Fukami, M., Hosouchi, T., Nakazaki, N., Okumura, S., Wada, T., Liu, Y.G., Shibata, D., and Tabata, S. (1997). A fine physical map of *Arabidopsis thaliana* chromosome 5: Construction of a sequence-ready contig map. DNA Res. **4,** 371–378.

Liebman, S.W., and Newnam, G. (1993). A ubiquitin-conjugating enzyme, RAD6, affects the distribution of Ty1 retrotransposon integration positions. Genetics **133,** 499–508.

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.

Liu, Y.G., Mitsukawa, N., Oosumi, T., and Whittier, R.F. (1995). Efficient isolation and mapping of *Arabidopsis thaliana* T-DNA insert junctions by thermal asymmetric interlaced PCR. Plant J. **8,** 457–463.

Machida, C., Onouchi, H., Koizumi, J., Hamada, S., Semiarti, E., Torikai, S., and Machida, Y. (1997). Characterization of the transposition pattern of the *Ac* element in *Arabidopsis thaliana* using endonuclease I-SceI. Proc. Natl. Acad. Sci. USA **94,** 8675–8680.

Sato, S., Kaneko, T., Kotani, H., Nakamura, Y., Asamizu, E., Miyajima, N., and Tabata, S. (1998). Structural analysis of *Arabidopsis thaliana* chromosome 5. IV. Sequence features of the regions of 1,456,315 bp covered by nineteen physically assigned P1 and TAC clones. DNA Res. **28,** 41–54.

Smith, D., Yanai, Y., Liu, Y.-G., Ishiguro, S., Okada, K., Shibata, D., Whittier, R.F., and Fedoroff, N.V. (1996). Characterization and mapping of *Ds*-GUS-T-DNA lines for targeted insertional mutagenesis. Plant J. **10,** 721–732.

Spradling, A.C., Stern, D.M., Kiss, I., Roote, J., Laverty, T., and Rubin, G.M. (1995). Gene disruptions using P transposable

elements: An integral component of the Drosophila genome project. Proc. Natl. Acad. Sci. USA **92,** 10824–10830.

Sundaresan, V. (1996). Horizontal spread of transposon mutagenesis: New uses for old elements. Trends Plant Sci. **1,** 184–190.

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

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

Townley, D.J., Avery, B.J., Rosen, B., and Skarnes, W.C. (1997). Rapid sequence analysis of gene trap integrations to generate a resource of insertional mutations in mice. Genome Res. **7,** 293–298.

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.

Zambrowicz, B.P., Friedrich, G.A., Buxton, E.C., Lilleberg, S.L., Person, C., and Sands, A.T. (1998). Disruption and sequence identification of 2,000 genes in mouse embryonic stem cells. Nature **392,** 608–611.

Zwaal, R.R., Broeks, A., van Meurs, J., Groenen, J.T., and Plasterk, R.H. (1993). Target-selected gene inactivation in *Caenorhabditis elegans* by using a frozen transposon insertion mutant bank. Proc. Natl. Acad. Sci. USA **15,** 7431–7435.