The Plant Journal (2002) **32**, 233–242

TECHNICAL ADVANCE

Distribution of 1000 sequenced T-DNA tags in the *Arabidopsis* genome

László Szabados[1], Izabella Kovács[1], Attila Oberschall[1], Edit Ábrahám[1], Irén Kerekes[1], Laura Zsigmond[1], Réka Nagy[1], Martha Alvarado[1], Inga Krasovskaja[1], Mónika Gál[1], Anikó Berente[1], George P. Rédei[2], Amit Ben Haim[3] and Csaba Koncz[4,*]

[1]*Institute of Plant Biology, Biological Research Center of Hungarian Academy of Sciences, H-6701 Szeged, PO Box 521, Temesvári krt 62, Hungary,*

[2]*3005 Woodbine Ct, Columbia, MO 65203-0906, USA,*

[3]*Vitality Biotechnologies Ltd, Einstein House, Carmel Industry Park, 39101 Haifa, Israel, and*

[4]*Max-Planck Institut für Züchtungsforschung, Carl-von-Linné-Weg 10, D-50829 Köln, Germany*

Received 24 April 2002; revised 16 June 2002; accepted 17 June 2002.
*For correspondence (fax +49 221 5062 213; e-mail koncz@mpiz-koeln.mpg.de).

Summary

Induction of knockout mutations by T-DNA insertion mutagenesis is widely used in studies of plant gene functions. To assess the efficiency of this genetic approach, we have sequenced PCR amplified junctions of 1000 T-DNA insertions and analysed their distribution in the *Arabidopsis* genome. Map positions of 973 tags could be determined unequivocally, indicating that the majority of T-DNA insertions landed in chromosomal domains of high gene density. Only 4.7% of insertions were found in interspersed, centromeric, telomeric and rDNA repeats, whereas 0.6% of sequenced tags identified chromosomally integrated segments of organellar DNAs. 35.4% of T-DNAs were localized in intervals flanked by ATG and stop codons of predicted genes, showing a distribution of 62.2% in exons and 37.8% in introns. The frequency of T-DNA tags in coding and intergenic regions showed a good correlation with the predicted size distribution of these sequences in the genome. However, the frequency of T-DNA insertions in 3′- and 5′-regulatory regions of genes, corresponding to 300 bp intervals 3′ downstream of stop and 5′ upstream of ATG codons, was 1.7–2.3-fold higher than in any similar interval elsewhere in the genome. The additive frequency of insertions in 5′-regulatory regions and coding domains provided an estimate for the mutation rate, suggesting that 47.8% of mapped T-DNA tags induced knockout mutations in *Arabidopsis*.

Keywords: T-DNA, insertion mutagenesis, *Arabidopsis*, sequenced tags, functional genomics.

Introduction

The availability of complete *Arabidopsis* genome sequence offers a unique source of information for characterization of plant gene functions using functional genomics and proteomics approaches (Arabidopsis Genome Initiative, 2000; Chory *et al*., 2000; Somerville, 2000). Despite recent advances in gene targeting (Hanin *et al*., 2001), gene disruption by homologous recombination is not yet applied as routinely in *Arabidopsis* as in other model systems (Coelho *et al*., 2000). To identify gene knockouts, however, several large collections offer insertion mutations induced by either the T-DNA of *Agrobacterium* or heterologous transposons (Bouché and Bouchez, 2001; Martienssen, 1998; Parinov and Sundaresan, 2001). Following an initial assessment of the efficiency of T-DNA mutagenesis, which resulted in the identification of hundreds of mutants with dramatically altered phenotypes (Azpiroz-Leehan and Feldmann, 1997; Feldmann, 1991; Koncz *et al*., 1992), the development of *Agrobacterium*-mediated *in planta* transformation methods provided a technical basis for saturation of the *Arabidopsis* genome with T-DNA insertions (Bechtold and Pelletier, 1998; Bechtold *et al*., 1993; Bent, 2000). In addition to phenotypic screening for loss- and gain-of-function

mutations, the *Arabidopsis* insertion mutant collections have been exploited for the development of efficient reverse genetic approaches using PCR-based identification of tags in known genes (Bouché and Bouchez, 2001; Galbiati *et al.*, 2000; Krysan *et al.*, 1999; Ríos *et al.*, 2002; Sussman *et al.*, 2000; Weigel *et al.*, 2000). The isolation of transposon and T-DNA insert junctions by PCR amplification techniques also provides a simple means for sequencing the tagged genes in order to establish a comprehensive database of sequence-indexed insertion mutations (Balzergue *et al.*, 2001; Parinov *et al.*, 1999; Samson *et al.*, 2002; Tissier *et al.*, 1999). Recently, several databases have become accessible to search for the presence of transposon and T-DNA tags in *Arabidopsis* genes (http://signal.salk. edu/cgi-bin/tdnaexpress; http://www.nadii.com/pages/collaborations/garlic_files/GarlicAnalysis.html; http://flagdb-genoplante-info.infobiogen.fr/projects/fst; http://genetrap. cshl.org; http://www.jic.bbsrc.ac.uk/sainsbury-lab/jonathan-jones/SINS-database/sins.htm).

To estimate the probability of finding an insertion in a gene of interest, it is frequently asked how many transposon or T-DNA tags are necessary to recover at least one mutation in all *Arabidopsis* genes. Although the mechanisms of T-DNA transfer and integration into the plant genome have been studied extensively (for review see Gelvin, 2000; Tinland, 1996; Zupan *et al.*, 2000), it is still unclear whether the T-DNA-induced mutations are indeed distributed randomly throughout the genome. To examine the distribution of T-DNA tags in *Arabidopsis*, we describe the sequence analysis and localization of 1000 PCR amplified T-DNA insert junctions. The data show that T-DNA insertions are rare in interspersed, centromeric, telomeric and rDNA repeats, but enriched in chromosomal domains with high gene density. The distribution of T-DNA tags shows a good overall correlation with the predicted size distribution of coding and intergenic regions in the genome sequence. Nonetheless, the frequency of insertions is significantly higher in 5′- and 3′-regulatory regions of genes, which thus appear to be more accessible for T-DNA integration. The frequency of T-DNA tags in coding regions and their 5′ flanking sequences of 300 bp average size is 47.8%, suggesting that about 160 000 T-DNA tags would probably be sufficient to saturate the *Arabidopsis* genome with insertion mutations at a probability of 95%.

Results and discussion

Isolation and sequencing of T-DNA insert junctions

Plant DNA sequences flanking the T-DNA tags can be isolated by several techniques, including plasmid rescue and amplification by inverse; thermal asymmetric interlaced (TAIL); and adaptor-mediated PCR (Koncz *et al.*, 1990; Liu *et al.*, 1995; Mathur *et al.*, 1998; Yephremov and Saedler, 2000).

To generate a random collection of sequenced T-DNA tags, we have modified the TAIL PCR protocol, which facilitates the amplification of T-DNA insert junctions from undigested genomic DNA. TAIL PCR is performed in three successive amplification steps at alternating high- and low-stringency annealing temperatures with a combination of T-DNA-specific primers of high melting temperature (T_m) and degenerate low-T_m primers that can frequently anneal with plant genomic DNA. In order to enhance the amplification of longer PCR fragments, we used a proofreading ExTaq polymerase (Takara) and extended the time of elongation steps. The improved long-TAIL PCR (LT-PCR) method provided a sufficiently high yield in two amplification steps, eliminating the need for a third PCR reaction (Figure 1). The efficiency of LT-PCR was compared to that of long-range inverse PCR (Li-PCR) performed with ExTaq enzyme, T-DNA-specific primers and template DNAs that were circularized after digestion with *Eco*RI restriction endonuclease, as described by Mathur *et al.* (1998). Whereas the original TAIL PCR protocol provided an amplification of fragments with only 30% of template DNAs (data not shown), over 80% of both LT- and Li-PCR reactions produced a reproducible amplification of T-DNA junction fragments ranging in size between 0.5 and 8 kb (Figure 1a). Using these techniques, PCR reactions were carried out with DNA templates that were purified from M_2 progeny of *Arabidopsis* lines carrying T-DNA insertions of *Agrobacterium* binary vectors pPCV6NFHyg, pPCV621, pPCV730, pPCV5013Hyg, pPCV6NFluxAB, pPCV6NFluxF, pPCVT-GUS and pPCVTac16 (Koncz *et al.*, 1987; Koncz *et al.*, 1989; Koncz *et al.*, 1994; Szabados and Koncz, 2002). About 60% of PCR reactions resulted in the amplification of single DNA fragments, whereas in other cases several products were observed together with some unspecific amplification products (Figure 1a). Following size separation on agarose gels, the amplified DNA fragments were purified by electroelution and used as templates for sequencing with oligonucleotide primers annealing to the T-DNA ends (see Experimental procedures). A comparison of size distribution of PCR fragments recovered by LT-PCR and Li-PCR is shown in Figure 1(b).

Distribution of T-DNA insertions in the Arabidopsis genome

The nucleotide sequence of PCR amplified DNA fragments was analysed in BLASTN homology searches, which revealed that about 35% of all recovered T-DNA insert junctions carried joined ends of tandem T-DNA repeats. On subtraction of these sequences from the database, the junctions between T-DNA and plant DNA were identified in order to remove the tag sequences from the entries, which were then

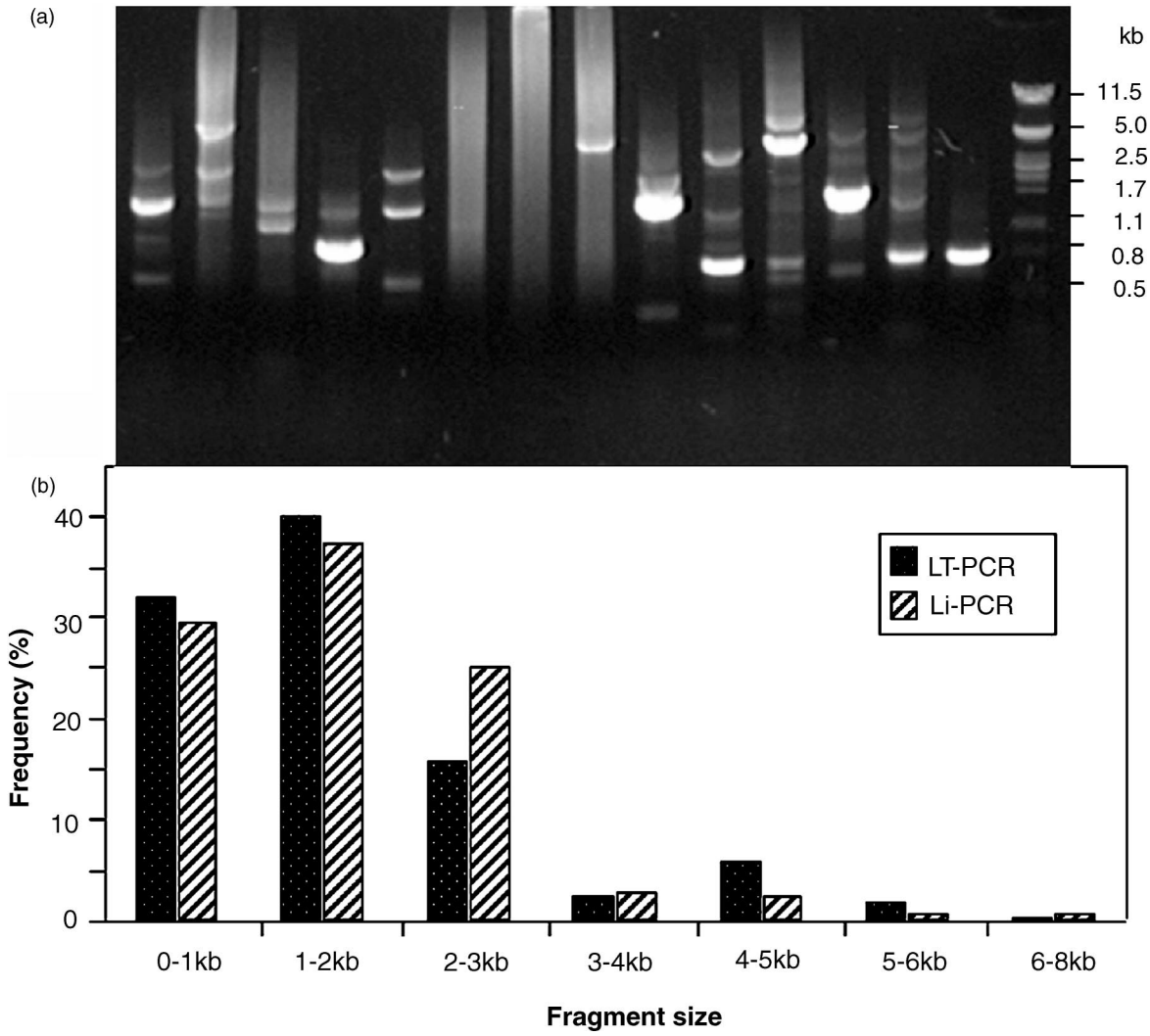

Figure 1. Isolation of T-DNA insert junctions using LT-PCR.
(a) LT-PCR amplification of plant DNA sequences flanking the left border of pTAC16 T-DNA insertions. Bands with higher intensity on ethidium bromide-stained agarose gels indicated specific amplification of junction fragments with over 80% of DNA templates. Weaker bands, corresponding to unspecific amplification products, were also often detected.
(b) A comparison of size distribution of T-DNA junction fragments obtained with 100 DNA templates, which were used in both Li and LT PCR reactions, illustrates that the modified LT-PCR technique is as suitable as the Li-PCR procedure for amplification of larger DNA fragments.

subjected to BLASTN, BLASTX and TBLASTX homology searches by screening the EST and reannotated genome sequence databases at GenBank, MIPS and TAIR (http://mips.gsf.de/proj/thal/db; http://www.arabidopsis.org/Blast). Following exclusion of redundant sequences generated by multiple tags in the same loci, the map positions of 1000 unique sequences were compiled using the MAPVIEWER function of the TAIR database (Huala *et al.*, 2001; Figure 2).

Chromosomal plotting the distribution of T-DNA insertion sites indicated that the T-DNA integration was not random. In contrast to finding one insertion in every 125 kb on average (as expected for even distribution of 1000 insertions in the haploid genome of 125 000 kb), a significantly higher frequency of insertions was observed in chromosomal regions with high gene density. For example, one insert per 55 kb was found in the 1–6 Mb region of chromosome 3, whereas the average size of segment carrying an insertion was 65 kb in the 9–13 Mb region of chromosome 4. In other regions, four to five insertions were localized within chromosomal segments of 60–90 kb, resulting in five to ten times higher insertion density values than the average. Such regions were, for example, represented by the BAC/P1 clones F13K9 in chromosome 1; F14M13 in chromosome 2; F17A9 in chromosome 3; T6G15, F16G20, and F20O9 in chromosome 4; and T30N20 and MRH10 in chromosome 5.

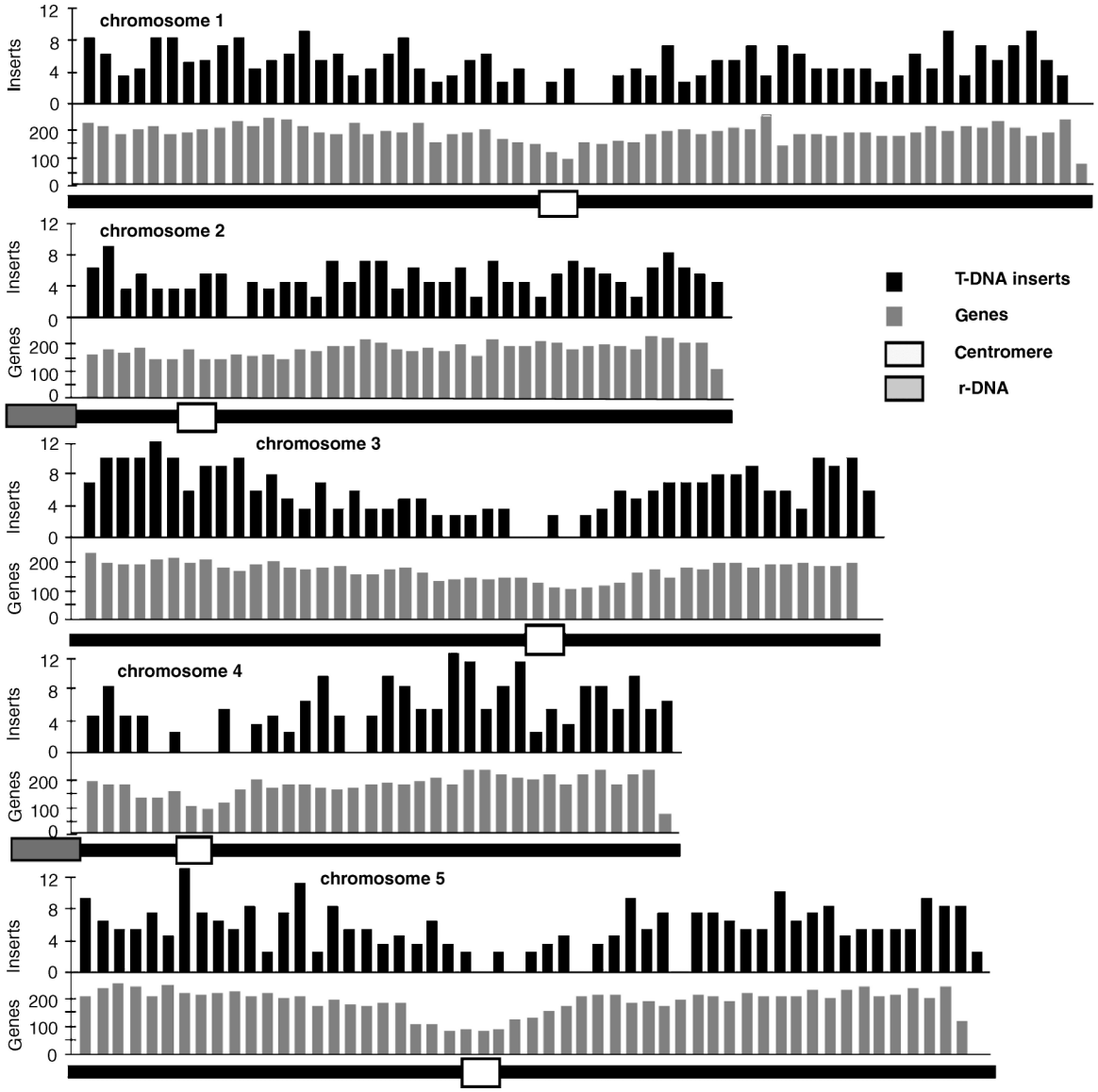

Figure 2. Chromosomal distribution of sequenced T-DNA insert junctions.
The positions of junction sequences in the *Arabidopsis* genome were identified by DNA sequence homology searches and projected to chromosomal intervals using the MAPVIEW facility provided by the TAIR database. The positions of centromeric regions and rDNA repeats are indicated.

Except 27 insertions, the location of all T-DNA tagged loci could be determined precisely. One of the insertions with ambiguous map position was found in an unclosed gap of the genome sequence corresponding to the cDNA clone BE524022. Only 48 T-DNA insertions (4.7%) were identified in repeated nuclear DNA sequences, including retro-elements, transposons, and other moderate or middle repetitive interspersed sequences. From these, three insertions were found in 18S ribosomal DNA repeats, two in 5S rDNA repeats, two in telomers, 18 in retroelements, and six in other transposon-like sequences. Remarkably, only four insertions were identified in centromeric DNA repeats, which add up to a total size of 9 Mb. Some insertions could be mapped to 180 bp repeats in chromosomes 3 and 5, and to other moderately or highly repeated DNA sequences in chromosomes 1, 3 and 5. About half of T-DNA insertions in repeats were located in pericentromeric regions. Flanking sequences of six T-DNA tags identified organellar DNAs. Two from four insertions showing homology to the chloroplast genome could be mapped to chromosomes, whereas one of the two insertions with homology to mitochondrial DNA was localized close to the centromeric region of chromosome 2.

From 973 unambiguously mapped T-DNA insertions, 344 were found between predicted ATG and stop codons of genes, 62.2% in putative exons and 37.8% in introns. All but

two (trnG and 18S rRNA) of these T-DNA tagged genes were predicted to be transcribed by RNA polymerase II, and six of them represented putative pseudogenes. Of T-DNA tagged genes, 21% could be assigned to known functions. The analysis of protein sequences derived from the tagged genes in gapped-BLASTP, PSI-BLASTP, SMART, Pfam, InterPro and Blocks searches (see Experimental procedures) revealed conserved protein motifs and functional domains in additional 59% of predicted proteins. The results of sequence analysis were collected in a database that displays the T-DNA insertion sites; predicted functions of T-DNA tagged genes, and features of corresponding hypothetical proteins (http://www.szbk.u-szeged.hu/~arabidop). The results on localization of T-DNA insertions are summarized in Table S1.

Distribution of T-DNA insertions in Arabidopsis genes and intergenic regions

To analyse the distribution of 973 mapped T-DNA tags at a higher resolution, the frequency of insertions in genes and intergenic regions was examined (Table 1). Genes that were defined by sequences between ATG and stop codons with a calculated average size of 2.94 kb carried 35.4% of the insertions. From the remaining 629 T-DNA tags, 121 (12.4%) were found 5′ upstream of predicted ATG codons, whereas 89 (9.1%) were localized 3′ downstream of stop codons, in both cases within a size interval of 300 bp; 419 (43.1%) tags were located outside this 300 bp size limit in intergenic sequences of an average size of about 3.4 kb. The calculated density of T-DNA tags in 5′- and 3′-regulatory regions was 1.7–2.3 higher than in any other interval of similar size in genes and intergenic sequences, suggesting that these sequences may be more accessible for T-DNA integration (Table 1). The number of tags in genes and their flanking 5′- and 3′-regulatory regions was nearly equal when the window size around the ATG and stop codons was increased to 500 bp. Figure 3 illustrates that, at this window, the number of tags observed upstream and downstream of predicted coding domains is two- to threefold higher than in similar intervals within the genes.

T-DNA tags often provide perfect polyadenylation sites when located 3′ downstream of stop codons, resulting in no observable changes in the steady-state levels of mRNAs of tagged genes (see for example Koncz *et al.*, 1989; Németh *et al.*, 1998). By contrast, T-DNA insertions located 5′ upstream of ATG codons, in regions corresponding to non-translated leaders of mRNAs (5′-UTRs) and close surrounding of TATA-box sequences, are expected to result in dramatic changes of gene expression, even if some T-DNA-derived sequences were recognized as functional promoter elements (as, for example, observed by Furini *et al.*, 1996). Therefore, to estimate the frequency of expected knockout mutations, we considered only those tags that were located either between ATG and stop codons, or within a distance of 300 bp 5′ upstream of ATG codons. The additive

Table 1 Distribution of T-DNA insertions in *Arabidopsis* genes and intergenic regions

Distribution of T-DNA tags	Number of inserts	Frequency (%)	Total size of region in the genome (kb)	Density of tags ($kb^{-1} \times 10^6$)	Ratio to average
Coding region	**344**	**35.4**	**51395**	**6.88**	**0.86**
Exon	214	22.0			
Intron	130	13.4			
Intergenic region	629	64.6	125000–51395 = ≈73605	8.78	1.10
5′ upstream, −300 bp	**121**	**12.4**	**25554 × 0.3 = 7666**	**16.2**	**2.03**
5′ upstream, −500 bp	193	19.8	25554 × 0.5 = 12777	15.5	1.94
3′ downstream, +300 bp	89	9.1	25554 × 0.3 = 7666	11.9	1.49
3′ downstream, +500 bp	122	12.5	25554 × 0.5 = 12777	9.81	1.23
Intergenic region/300	419	43.1	73605–7666 = 65939	6.53	0.82
Intergenic region/500	314	32.3	73605–12777 = 60828	5.31	0.66
Coding +300 bp 5′	**465**	**47.8**	**51395 + 7666 = 59061**	**8.09**	**1.01**
Coding +500 bp 5′	537	55.2	51395 + 12777 = 64172	8.60	1.08
All mapped	**973**	**100.0**	**125000**	**8.00**	**1.00**
Repeats	27				
Total	1000				

The second column shows the number of inserts localized in coding regions of genes between predicted ATG and stop codons. The numbers of T-DNA tags in intergenic regions correspond to insert junctions found within and outside of 300 and 500 bp intervals upstream and downstream of genes. The third column shows the frequency (%) of tags in different regions. The fourth column indicates the total size of coding and intergenic regions with or without 5′- and 3′-regulatory sequences of either 300 or 500 bp. The average density of tags in the different regions was calculated by dividing the frequency of tags (shown in the third column) with the total size of examined genomic region and expressed in units of $kb^{-1} \times 10^6$. The sixth column indicates the ratios between the average T-DNA density (8.0) and density of tags in different genomic regions.

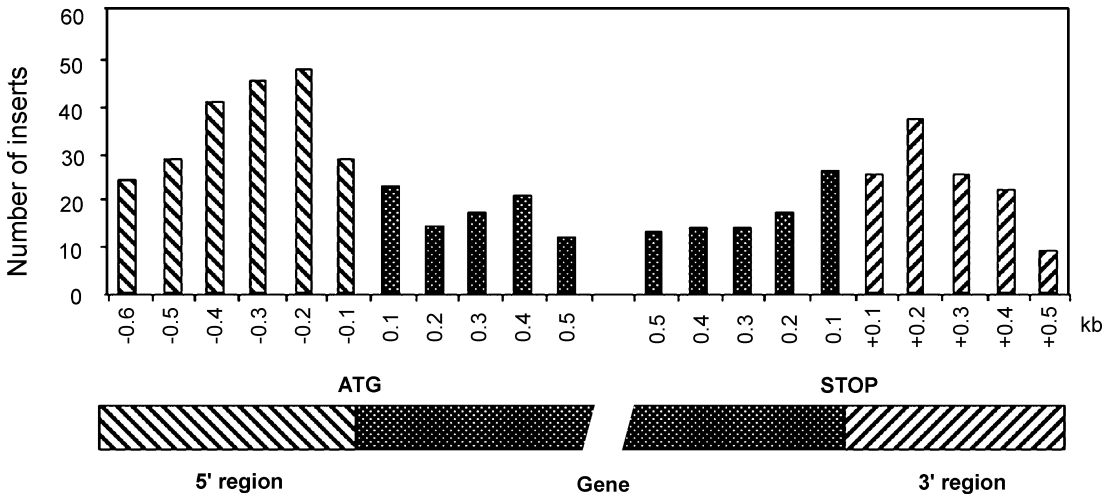

Figure 3. Distribution of T-DNA insertions within a size interval of 500 bp around the ATG and stop codons.
The number of T-DNA tags found within this size interval 3′ downstream of stop and 5′ upstream of ATG codons is two- to threefold higher, respectively, than the number of insertions located at a similar distance within the genes.

frequency of these tags in the population of 973 mapped insertions was $344 + 121/973 = 0.478$, which provided an estimate for the mutation rate.

Estimation of population size required for saturation mutagenesis

To estimate the probability (P) of finding at least one mutation in a gene of size x (kb), previous workers have used the equation

$$P = 1 - (1 - x/125\,000)^n \qquad \text{eqn 1}$$

in which n is the number of T-DNA insertions, and 125 000 (kb) is the size of haploid *Arabidopsis* genome (Krysan *et al.*, 1999). This equation predicts that T-DNA integration occurs with similar probability at any chromosomal interval, and that the frequency of mutations is proportional to the size of target sequences. However, our data indicate that chromosomal distribution of the T-DNA insertions is not random. As T-DNA insertions are rare in repeated sequences, equation 1 is expected to overestimate the n value proportionally with the size contribution of those sequences that are infrequent targets for T-DNA integration.

Nonetheless, the frequency of T-DNA tagged genes sorted to different size classes showed a good correlation with the size distribution of genes (intervals between predicted ATG and stop codons) in the *Arabidopsis* genome (Table 2). In this comparison, the fraction of genome size covered by a given gene class (calculated by dividing the total size of genes in a given size class with the genome

Table 2 Comparison of size distribution of genes in the *Arabidopsis* genome to that of genes identified by sequenced T-DNA insertions

Gene class (kb)	Genes in size class		Size of gene class in the genome (kb)		Expected insertions		Observed insertions		Expected tags in 5′-regions of 300 bp		Observed tags in 5′-regions of 300 bp	
	Total	%	Total	%	Number	%	Number	%	Number	%	Number	%
<1	6028	23.5	3823	3.1	29	7.3	29	8.4	14	23.7	30	24.8
1–2	9627	37.7	14232	11.4	110	27.5	100	29.1	22	37.3	41	33.9
2–3	5489	21.5	13 328	10.7	106	26.5	91	26.4	13	22.0	35	28.9
3–4	2192	8.6	7541	6.0	58	14.5	53	15.5	5	8.5	6	5.0
>4	2218	8.7	12471	10.0	97	24.2	71	20.6	5	8.5	9	7.4
Total	25 554	100	51 395	41.2	400	100	344	100	59	100	121	100

The number of genes in a given size class is shown in the second column, and the total size of genes in the different size classes and their fractions in the genome (according to the reannotated *Arabidopsis* sequence database at the GenBank) are depicted in the third column. The expected number of insertions were calculated as described in the text and compared to numbers of tags observed in 5′-regulatory regions of 300 bp and between ATG and stop codons of predicted genes.

size) was multiplied with the number of tested T-DNA insertions (973). The estimates for the expected number of T-DNA tags yielded by this approximation were similar to the observed number of T-DNA insertions in genes of different size classes. Thus the frequencies of T-DNA tags appeared to be roughly proportional with the size of tagged genes, suggesting that T-DNA integration could probably occur with a similar probability in different genes. To calculate the expected number of inserts in 5′-regulatory regions, the genome size fraction of 300 bp intervals occupied by all genes of a given size class (calculated by multiplying the number of genes in a given size class by 0.3 and dividing the product by the genome size, 125 000 kb) was multiplied by the number of tested T-DNA insertions. In each case, the observed number of T-DNA tags was about twofold the calculated ('expected') values, which reflected our observation that density of T-DNA tags found in a given interval of 5′-regulatory regions was at least twofold higher than anywhere else in the genome. The total number of tags in genes and 5′-regulatory regions expected on the basis of gene size distribution in the *Arabidopsis* genome was 459, closely approaching the observed value of 465.

To find a practical way to describe the observed correlation, we have used the following equation:

$$P = 1 - \mathrm{f}^{n} \qquad \text{eqn 2}$$

(Rédei and Koncz, 1992). As the fraction of T-DNA insertions causing knockout mutations was estimated at $t = 465/973 = 0.478$, the average mutation rate for any of the 25 554 predicted *Arabidopsis* genes (GenBank) was expressed as $t/25\,554$ (1.87×10^{-5}). This average mutation rate thus counts mutations in genes and 5′-regulatory regions without specifically considering the size of genes. It can be deduced from Table 2 that the number of insertions required for finding a mutation in any gene of a given size category is decreased by a factor depending on the ratio of mutations observed in 5′-regulatory regions and coding regions of genes. It can also be shown that the estimates obtained by employing the average mutation rate will be close to those values, which are calculated by counting the additive number of mutations in coding domains and promoter regions of genes in the different size classes. Solving equation 3:

$$P = 1 - (1 - \mathrm{t}/25\,554)^{n} \qquad \text{eqn 3}$$

for n suggests that about 160 000 T-DNA tags would be sufficient to find insertion mutations in all *Arabidopsis* genes at a probability of 95%, whereas about 246 000 insertions would be required to saturate the genome with knockout mutations at a probability of 99%. In comparison, an approximation that accounts both gene size and the ratio of mutations in promoter and coding regions, according to Table 2, would predict that about 230 000 T-DNA insertions would be sufficient for saturation mutagenesis.

Using a PCR-based mutant screening approach with 51 213 T-DNA insertions, Ríos *et al.* (2002) found T-DNA tags in genes at a probability of 47.3%. By solving equation 3 they obtained a t-value of 0.32, predicting that about 240 000 T-DNA tags would be required for saturation T-DNA mutagenesis of *Arabidopsis* genes at a probability of 95%. Based on these two independent estimates, it seems more economic to generate only about 160 000 sequenced indexed tags to approach saturation at an estimated probability of 85–95%, and to fill the remaining gap, if any, using alternative methods such as either PCR screening or sequencing of transposon-tagged mutant collections, or identification of point mutations using the TILLING path technology (Colbert *et al.*, 2001; Parinov and Sundaresan, 2001; Tissier *et al.*, 1999).

Experimental procedures

Plant material and preparation of PCR templates

The optimization of PCR reactions was performed with DNAs prepared from a T-DNA tagged mutant collection, which was generated by tissue culture transformation and has been characterized earlier (Koncz *et al.*, 1992). In addition to lines deposited at the Nottingham *Arabidopsis* Stock Centre (http://nasc.nott.ac.uk), which carried T-DNAs of the kanamycin gene fusion vectors pPCV6NFHyg and pPCV621 (Koncz *et al.*, 1989), this collection included lines transformed with pPCV730, pPCV5013Hyg, pPCV6NFluxAB, pPCV6NFluxF and pPCVT-GUS T-DNAs (Koncz *et al.*, 1987, 1994). All other T-DNA-tagged lines used for sequencing the insert junctions were generated by vacuum infiltration of *Arabidopsis* (Col-0) plants with *Agrobacterium* GV3101 (pMP90RK) carrying the activator tagging vector pTac16 (Bechtold and Pelletier, 1998; Koncz *et al.*, 1994; Szabados and Koncz, 2002).

For DNA isolation, 0.5–1 g frozen leaf material was collected from seedlings representing either M_2 or M_3 progeny of T-DNA-tagged lines, homogenized in 0.9 ml extraction buffer (100 mM Tris–HCl pH 8.0, 50 mM EDTA, 0.5 M NaCl), and incubated at 65°C for 20 min. The crude lysates were supplemented with 0.3 ml 5 M potassium acetate, incubated on ice for 20 min, and cleared by centrifugation in a microcentrifuge at 15 000 rpm (equal with 12 000 ***g***) for 5 min. Following precipitation of the DNA from the cleared lysates with 0.8 ml isopropanol on ice for 1 h, the DNA samples were recovered by centrifugation, washed with ice-cold 70% ethanol, dried, resuspended in 0.4 ml TE buffer (10 mM Tris–HCl pH 8.0, 1 mM EDTA) containing 100 μg ml^{-1} RNase, and incubated for 1 h at 37°C. Subsequently, the samples were subjected to extraction with phenol : chloroform : isoamylalcohol (25 : 24 : 1 v/v), then with chloroform, and precipitated by addition of 50 μl 3 M sodium acetate pH 6.0 and 0.5 ml isopropanol, as described above. Following centrifugation, the purified DNA pellets were dissolved in 100 μl TE buffer and used as PCR templates.

PCR amplification of T-DNA insert junctions

The T-DNA insert junctions were isolated using modified versions of previously described, long-range, inverse PCR (Li-PCR) and TAIL PCR techniques (Liu *et al.*, 1995; Mathur *et al.*, 1998). For Li-PCR amplification, 2 μg plant DNA was digested with *Eco*RI,

Table 3 Oligonucleotide primers for Li- and LT-PCR amplification of T-DNA insert junctions

Primer	Vector, annealing, purpose	Sequence (5′–3′)
LB11	All PCV vectors, left border	CAGACCAATCTGAAGATGAAATGGGTATCTGGG
LB21	All PCV vectors, left border, nested 1	GTGAAGTTTCTCATCTAAGCCCCCATTTGG
LB31	All PCV vectors, left border, nested 2	CTTTCGCCTATAAATACGACGGATCG
LB41	All PCV vectors, left border, nested 3	TTCTCCATATTGACCATCATACTCATTGC
LB51	All PCV vectors, left border, sequencing	TCATACTCATTGCTGATCCATG
PR1	All PCV vectors, left border, Li-PCR	CACATTTCCCCGAAAAGTGCCACCTGACG
PR2	All PCV vectors, left border, Li-PCR	CCTATAAAAATAGGCGTATCACGAGGCCC
RB11	pPCV6NFHyg, right border	GTTTCGCTTGGTGGTCGAATGGGCAGGTAG
RB21	pPCV6NFHyg, right border, nested 1	CCCGGCACTTCGCCCAATAGCAGCCAGTCC
Km21	pPCV6NFHyg, right border, nested 2	CCCAGTCATAGCCGAATAGCCTCTCCACCC
Km3	pPCV6NFHyg, right border, sequencing	CGGAGAACCTGCGTGCAATCCATC
TG1	pTgus, right border	CGATACGCTGGCCTGCCCAACCTTTCGG
TG2	pTgus, right border, nested 1	CATCGGCGAACTGATCGTTAAAACTGCCTGG
TG3	pTgus, right border, sequencing	CGATCCAGACTGAATGCCCACAGG
EH11	pTgus, pPCV6NFHyg, *Eco*RI, Li-PCR	GTCCTGCGGGTAAATAGCTGCGCCGATGG
EH21	pTgus, pPCV6NFHyg, *Eco*RI, Li-PCR	CGTTATGTTTATCGGCACTTTGCATCGGC
AD1	Degenerate primer 1, LT-PCR	NTCA(GC)T(AT)T(AT)T(GC)G(AT)GTT
AD2	Degenerate primer 2, LT-PCR	NGTCGA(GC)(AT)GANA(AT)GAA
AD3	Degenerate primer 3, LT-PCR	(AT)GTGNAG(AT)ANCANAGA

Primers LB11 and RB11 were used in combination with the AD1 degenerate primer 1 in the first steps of LT-PCR reactions. One of the T-DNA nested primers (1, 2 or 3) was combined with the AD2 degenerate primer 2 to perform the second LT-PCR amplification steps. Primers used for sequencing the amplified insert junctions are indicated.

which had one or two recognition site within the different T-DNAs used (Koncz *et al.*, 1994; Szabados and Koncz, 2002). The digested DNAs were circularized by ligation, and aliquots from the ligated samples (0.6–0.8 μg DNA) were used as templates in long-range PCR reactions with Takara ExTaq enzyme (Takara Shuzo Co. represented by BioWhittaker, Taufkirchen, Germany). The PCR reactions were initiated by heating the samples for 2 min at 95°C, followed by 35 cycles performed at 94°C for 15 sec and 68°C for 8 min.

To optimize the amplification of T-DNA insert junctions from undigested DNA templates, the original three-step TAIL PCR protocol (Liu *et al.*, 1995) was modified by using Takara ExTaq instead of Taq polymerase, increasing the annealing temperature of T-DNA specific primers to 68°C, and applying longer elongation periods in order to eliminate the need for a third PCR cycle. The first amplification step was performed with 0.05 μg μl^{-1} DNA template with a T-DNA specific primer and degenerate primer 1 (Table 3). The PCR program included the following steps: 95°C for 2 min; five cycles (94°C for 15 sec and 68°C for 8 min); two cycles (94°C for 15 sec, 45°C for 2 min and 70°C for 8 min), and finally 12 cycles (94°C for 15 sec, 68°C for 8 min, 94°C for 15 sec, 68°C for 8 min, 94°C for 15 sec, 45°C for 30 sec and 70°C for 8 min). After the first amplification step, the PCR reactions were diluted with sterile water to 100 μl and an aliquot of 2 μl from each sample was added to a second PCR reaction of 50 μl which contained a T-DNA-specific nested primer and degenerate primer 2 (Table 3). The second PCR was performed using the program: 94°C for 30 sec, 15 cycles (94°C for 15 sec, 68°C for 8 min, 94°C for 15 sec, 68°C for 8 min, 94°C for 15 sec, 45°C for 30 sec and 70°C for 8 min) and 70°C for 8 min. All PCR reactions were carried out in a thermocycler using 96-well microtitre plates. The PCR products were size-separated in agarose gels, isolated by electroelution, and purified by phenol–chloroform extractions followed by precipitation with isopropanol as described above. The isolated DNA fragments were sequenced with the T-DNA nested primers using a Perkin-Elmer ABI377XL DNA sequencer.

Sequence analysis

Chromosomal positions of T-DNA insertions in the *Arabidopsis* genome sequence were determined by BLASTN DNA homology searches using the NCBI GenBank (Wheeler *et al.*, 2001; http://www.ncbi.nlm.nih.gov/BLAST); TAIR (Huala *et al.*, 2001; http://www. arabidopsis.org/Blast); and MIPS (Mewes *et al.*, 1999; http://mips.gsf.de/proj/thal/db/search/search_frame.html) databases. The map positions of T-DNA tags were projected to chromosomes with the MAPVIEWER function of the TAIR database (http://www.arabidopsis.org/servlets/mapper). Homology searches with protein sequences derived from the predicted genes were performed with gapped-BLASTP and PSI-BLASTP programs (Altschul *et al.*, 1997) by scanning the non-redundant GenBank and EMBL (http://dove.embl-heidelberg.de/Blast2) databases. Searches for conserved protein domains and motifs were carried out using the SMART (Letunic *et al.*, 2002; http://smart.embl-heidelberg.de); Pfam (Bateman *et al.*, 2002; http://www.sanger. ac.uk/Software/Pfam/search. shtml); InterPro (Zdobnov and Apweiler, 2001; ;http://www.ebi.ac. uk/interpro/scan.html); and BLOCKS (Henikoff *et al.*, 2000; http://blocks.fhcrc.org) servers. The sequence analysis data were collected in a computer database using the FILEMAKER PRO 4.1 software.

Acknowledgements

This work was supported by Vitality Biotechnologies Ltd, Haifa, Israel, as well as by grants from the from the Deutsches Zentrum für Luft und Raumfahrt (UNG-027-97), OTKA (No. T-029430 and T-032428), and the European Commission (QLRT-2000-01871).

Supplementary Material

The following material is available from http://www.blackwell-science.com/products/journals/suppmat/TPJ/TPJ1417/TPJ1417sm.pdf

Table S1 Distribution of 1000 sequenced T-DNA tags in the *Arabidopsis* genome.

References

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

Arabidopsis Genome Initiative (2000) Analysis of the genome sequence of the flowering plant *Arabidopsis thaliana*. *Nature*, **408**, 796–815.

Azpiroz-Leehan, R. and Feldmann, K.A. (1997) T-DNA insertion mutagenesis in *Arabidopsis*: going back and forth. *Trends Genet.* **13**, 152–156.

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

Bateman, A., Birney, E., Cerruti, L. *et al.* (2002) The Pfam protein families database. *Nucl. Acids Res.* **30**, 276–280.

Bechtold, N., Ellis, J. and Pelletier, G. (1993) *In planta Agrobacterium* mediated gene transfer by infiltration of adult *Arabidopsis thaliana* plants. *C.R. Acad. Sci. Paris*, **316**, 1194–1199.

Bechtold, N. and Pelletier, G. (1998) *In planta Agrobacterium*-mediated transformation of adult *Arabidopsis thaliana* plants by vacuum infiltration. *Meth. Mol. Biol.* **82**, 259–266.

Bent, A.F. (2000) *Arabidopsis* in plant transformation. Uses, mechanism and prospects for transformation of other species. *Plant Physiol.* **124**, 1540–1547.

Bouché, N. and Bouchez, D. (2001) *Arabidopsis* gene knockout: phenotypes wanted. *Curr. Opin. Plant Biol.* **4**, 111–117.

Chory, J., Ecker, J.R., Briggs, S. *et al.* (2000) National Science Foundation-Sponsored Workshop Report: 'The 2010 Project': functional genomics and the virtual plant. A blueprint for understanding how plants are built and how to improve them. *Plant Physiol.* **123**, 423–426.

Coelho, P.S.R., Kumar, A. and Snyder, M. (2000) Genome-wide mutant collections: toolboxes for functional genomics. *Curr. Opin. Microbiol.* **3**, 309–315.

Colbert, T., Till, B.J., Tompa, R., Reynolds, S., Steine, M.N., Yeung, A.T., McCallum, C.M., Comai, L. and Henikoff, S. (2001) High-throughput screening for induced point mutations. *Plant Physiol.* **126**, 480–484.

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

Furini, A., Koncz, C., Salamini, F. and Bartels, D. (1996) High level transcription of a member of a repeated gene family confers dehydration tolerance to callus tissue of *Craterostigma plantagineum*. *EMBO J.* **16**, 3599–3608.

Galbiati, M., Moreno, M.A., Nadzan, G., Zourelidou, M. and Dellaporta, S.L. (2000) Large-scale T-DNA mutagenesis in *Arabidopsis* for functional genomic analysis. *Funct. Integr. Genomics*, **1**, 25–34.

Gelvin, S.B. (2000) *Agrobacterium* and plant genes involved in T-DNA transfer and integration. *Annu. Rev. Plant Physiol. Plant Mol. Biol.* **51**, 223–256.

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

Henikoff, J.G., Greene, E.A., Pietrokovski, S. and Henikoff, S. (2000) Increased coverage of protein families with the blocks database servers. *Nucl. Acids Res.* **28**, 228–230.

Huala, E., Dickerman. A.V., Garcia-Hernandez, M. *et al.* (2001) The *Arabidopsis* Information Resource (TAIR): a comprehensive database and web-based information retrieval, analysis, and visualization system for a model plant. *Nucl. Acids Res.* **29**, 102–105.

Koncz, C., Olsson, O., Langridge, W.H.R., Schell, J. and Szalay, A.A. (1987) Expression and assembly of functional bacterial luciferase in plants. *Proc. Natl Acad. Sci. USA*, **84**, 131–135.

Koncz, C., Martini, N., Mayerhofer, R., Koncz-Kálmán, Z., Körber, H., Rédei, G.P. and Schell, J. (1989) High-frequency T-DNA-mediated gene tagging in plants. *Proc. Natl Acad. Sci. U.S.A*, **86**, 8467–8471.

Koncz, C., Mayerhofer, R., Koncz-Kálmán, Z., Nawrath, C., Reiss, B., Rédei, G.P. and Schell, J. (1990) Isolation of a gene encoding a novel chloroplast protein by T-DNA tagging in *Arabidopsis thaliana*. *EMBO J.* **9**, 1337–1346.

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

Koncz, C., Martini, N., Szabados, L., Hrouda, M., Bachmair, A. and Schell, J. (1994) Specialized vectors for gene tagging and expression studies. In *Plant Molecular Biology Manual*, Vol. B2 (Gelvin, S.B., Schilperoort, R.A. and Verma, D.P.S., eds). Dordrecht, the Netherlands: Kluwer Academic Publishers, pp. 1–22.

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

Letunic, I., Goodstadt, L., Dickens, N.J. *et al.* (2002) Recent improvements to the SMART domain-based sequence annotation resource. *Nucl. Acids Res.* **30**, 242–244.

Liu, Y.G., Misukawa, 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.

Martienssen, R.A. (1998) Functional genomics: probing plant gene function and expression with transposons. *Proc. Natl Acad. Sci. USA*, **95**, 2021–2026.

Mathur, J., Szabados, L., Schaefer, S., Grunenberg, B., Lossow, A., Jonas-Straube, E., Schell, J., Koncz, C. and Koncz-Kálmán, Z. (1998) Gene identification with sequenced T-DNA tags generated by transformation of *Arabidopsis* cell suspension. *Plant J.* **13**, 707–716.

Mewes, H.W., Heumann, K., Kaps, A., Mayer, K., Pfeiffer, F., Stocker, S. and Frishman, D. (1999) MIPS: a database for genomes and protein sequences. *Nucl. Acids Res.* **27**, 44–48.

Németh, K., Salchert, K., Putnoky, P. *et al.* (1998) Pleiotropic control of glucose and hormone responses by PRL1, a nuclear WD protein in *Arabidopsis*. *Genes Dev.* **12**, 3959–3073.

Parinov, S. and Sundaresan, V. (2001) Functional genomics in *Arabidopsis*: large-scale insertional mutagenesis complements the sequencing project. *Curr. Opin. Biotechnol.* **11**, 157–161.

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

Rédei, G.P. and Koncz, C. (1992) Classical mutagenesis. In *Methods in Arabidopsis Research* (Koncz, C., Chua, N.-H. and Schell, J., eds). Singapore: World Scientific, pp. 16–82.

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

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

Somerville, C. (2000) The twentieth century trajectory of plant biology. *Cell*, **100**, 13–25.

Sussman, M.R., Amasino, R.M., Young, J.C., Krysan, P.J. and Austin-Phillips, S. (2000) The *Arabidopsis* knockout facility at the University of Wisconsin-Madison. *Plant Physiol.* **124**, 1465–1467.

Szabados, L. and Koncz, C. (2002) Identification of T-DNA insertions in *Arabidopsis* genes. In *Genomics of Plants and Fungi* (Prade, R.A. and Bohnert, H.J., eds). New York: Marcel Dekker, in press.

Tinland, B. (1996) The integration of T-DNA into plant genomes. *Trends Plant Sciz.* **1**, 178–183.

Tissier, A.F., Merillonnet, 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.

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

Wheeler, D.L., Church, D.M., Lash, A.E. *et al.* (2001) Database resources of the National Center for Biotechnology Information. *Nucl. Acids Res.* **29**, 11–16.

Yephremov, A. and Saedler, H. (2000) Display and isolation of transposon flanking sequences starting from genomic DNA or RNA. *Plant J.* **21**, 495–505.

Zdobnov, E.M. and Apweiler, R. (2001) InterProScan – an integration platform for the signature-recognition methods in InterPro. *Bioinformatics*, **17**, 847–848.

Zupan, J., Muth, T.R., Draper, O. and Zambryski, P. (2000) The transfer of DNA from *Agrobacterium tumefaciens* into plants: a feast of fundamental insights. *Plant J.* **23**, 11–28.