Available online at www.sciencedirect.com
SCIENCE DIRECT
ELSEVIER
Genomics 87 (2006) 645–652
GENOMICS
www.elsevier.com/locate/ygeno

Analysis of T-DNA insertion site distribution patterns in *Arabidopsis thaliana* reveals special features of genes without insertions

Yong Li [a,b], Mario G. Rosso [a,b], Bekir Ülker [a], Bernd Weisshaar [b,*]

[a] *Max Planck Institute for Plant Breeding Research, Carl-von-Linne-Weg 10, D-50829 Cologne, Germany*
[b] *Institute of Genome Research, Center for Biotechnology, Bielefeld University, Universitätsstrasse 25, D-33594 Bielefeld, Germany*

Received 22 September 2005; accepted 20 December 2005
Available online 20 February 2006

Abstract

Large collections of sequence-indexed T-DNA insertion mutants are invaluable resources for plant functional genomics. Flanking sequence tag (FST) data from these collections indicated that T-DNA insertions are not randomly distributed in the *Arabidopsis thaliana* genome and that there are still a fairly high number of annotated genes without T-DNA insertions. We have analyzed FST data from the FLAGdb, GABI-Kat, and SIGnAL mutant populations. The lack of detectable transcriptional activity and the absence of suitable restriction sites were among the reasons genes are not covered by insertions. Additionally, a refined analysis of FSTs to genes with annotated noncoding regions showed that transcription initiation and polyadenylation site regions of genes are favored targets for T-DNA integration. These findings have implications for the use of T-DNA in saturation mutagenesis and for our chances to find a useful knockout allele for every gene.

Keywords: T-DNA; Integration; Insertional mutagenesis; *Arabidopsis thaliana*; Restriction sites; Gene expression

After the completion of the genome sequence of the model plant *Arabidopsis thaliana* [1], research emphasis in the *Arabidopsis* community has shifted toward understanding the function of all the 26,000 genes that are not considered to be pseudogenes. Large-scale insertional mutagenesis approaches, especially by *Agrobacterium*-mediated T-DNA transfer, play important roles in elucidating gene functions in plants [2]. Several T-DNA-mutagenized populations have been generated and indexed in flanking sequence tag (FST)-based databases [3–6]. They are important resources for employing reverse genetics approaches to gene function studies. For example, among 620 *A. thaliana* genes with a known mutant phenotype, 40% were identified by T-DNA tagging, which was the largest fraction in methods used for gene function identification [7].

T-DNA is a segment of the Ti plasmid of *Agrobacterium tumefaciens* flanked by 25-bp imperfect repeats (left and right border). During transformation, the T-DNA is transferred from *Ag. tumefaciens* to the plant cell and imported with the help of several virulence proteins into the nucleus in a single-stranded form. Finally, the T-DNA is integrated into the plant genome [8,9]. PCR-based methods are used to generate DNA fragments spanning from the borders into genomic DNA, which are subsequently sequenced. The availability of large amounts of FST data aided the finding that T-DNA insertion sites are not randomly distributed in the genome and also that insertion distribution bias is present at different levels. T-DNA integration sites were detected preferentially in intergenic regions compared to genic regions [5,10–12], and T-DNA integration events seem to be associated with gene density since higher frequencies of insertions were observed in gene-rich regions and lower frequencies around centromeric regions that contain fewer genes [5,13,14]. It has long been assumed that T-DNA integration prefers transcriptionally active genes [15,16], but this hypothesis could not be confirmed by the SIGnAL FST data combined with expression profiling results [5]. At the gene level, an interesting finding was that T-DNA insertions are enriched in regions before translation start and after translation stop [11,13]. This phenomenon has also been observed in rice [17].

We have retrieved the FST data from three large publicly available T-DNA FST populations: FLAGdb [3], GABI-Kat

* Corresponding author. Fax: +49 521 106 6423.
E-mail address: bernd.weisshaar@uni-bielefeld.de (B. Weisshaar).

0888-7543/$ - see front matter © 2005 Elsevier Inc. All rights reserved.
doi:10.1016/j.ygeno.2005.12.010

[11], and SIGnAL [5]. The high-quality genome annotation [18] in which the majority of protein-coding genes have EST/cDNA support as well as annotated noncoding leader and trailer sequences (UTRs) enabled us to study the insertion site distribution with regard to transcription initiation and polyadenylation (poly(A)) site regions, not only in relation to coding sequences. We present evidence that the integration frequency peaks correspond to transcription initiation and poly(A) site regions rather than translation start and stop. We also analyzed a subset of genes without sequence-indexed T-DNA insertions in any of the three populations and found that these genes, in addition to being small, are short of suitable restriction sites in the surrounding genome sequence and are not likely to be expressed.

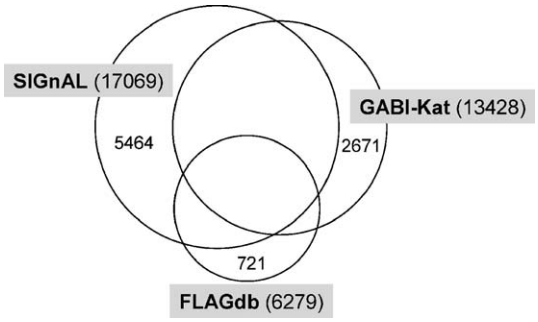

Fig. 1. Venn diagram showing the three sets of genes covered by insertions in each of the three T-DNA populations and their relationships with regard to redundancy of covered genes.

Results

Distribution of insertion sites relative to genes and pseudogenes

After processing the FST data (see Materials and methods), we obtained more than 224,000 insertion sites (Table 1). In total, there are indications for 21,234 of 26,207 protein-coding genes that have T-DNA insertions in the transcribed area. When the T-DNA populations are compared to each other, it becomes clear that each population covers a portion of genes that are uniquely present in only one of the populations (Fig. 1).

We analyzed the distribution of T-DNA integration sites relative to genes with annotated UTRs. The insertion frequency was determined relative to the regions of transcription initiation (start of 5′ UTR) and poly(A) site (end of 3′ UTR) in bins of 100 bp. In parallel, data from simulated random insertions were analyzed in the same way (see Materials and methods). The resulting distribution was quite similar for all three T-DNA populations, displaying two insertion frequency peaks relative to genes (Fig. 2A). The first peak is in the promoter region close to transcription initiation, with the highest insertion frequency just upstream of transcription initiation. The second peak with lower height appears around the poly(A) site region. In addition, we analyzed the T-DNA insertion distribution around the beginning and end of the "ORF" (open reading frame) of pseudogenes (Fig. 2B). The insertion frequency was lower than random in either "genic" or adjacent intergenic regions of pseudogenes. Compared to the results from real genes, the pattern of insertion frequency in and around pseudogenes displays clear differences. Most obvious is the absence of the insertion frequency peaks.

The distribution of the simulated random insertions was found to be clearly distinct from the distributions observed for genes and pseudogenes, but did not appear as a flat baseline. The reason is that if an insertion site is located at some distance from a given gene, it will easily fall into the range of the neighboring gene, so that the insertion frequency obtained for a single gene decreases with distance.

To examine if the insertion frequency peaks are correlated with the "technically relevant restriction sites count" (see Materials and methods) and/or the base composition bias in the genome sequence, we calculated the restriction sites count and the GC content around the regions of transcription initiation and poly(A) sites of genes with annotated UTRs. The GC content is shown in Fig. 2C. The GC content is higher in transcribed regions than in intergenic regions, but is not correlated with the peaks in transcription initiation and poly(A) regions. With regard to the technically relevant restriction sites count, the count value and the frequency distribution around the two insertion peak areas varied greatly between the different populations (Supplementary Fig. F1). This is mostly due to the fact that the recognition sequence of the relevant enzymes have different GC content and that the GC contents in these genomic regions are different (Fig. 2C). From these data it is obvious that the technically relevant restriction sites count has no causal relationship with the two insertion frequency peaks.

Table 1
Insertion sites data summary

Data source	Number of FSTs	Number of insertion sites	Number of distinct genes with insertions
FLAGdb	36,287	30,139	6,279
GABI-Kat	103,033	67,611	13,428
SIGnAL	159,968	126,694	17,069
Total	299,288	224,444	21,234[a]

[a] The total number of distinct genes with insertions is not the sum of genes with insertions in individual populations.

Characteristics of genes without insertions

There was a total of 4973 protein-coding genes (excluding pseudogenes) annotated in TIGR v5 that have no detected insertions in any of the three populations. We focused on genes from this group with a length greater than a variable cut-off value. The length cut-off value was chosen based on the formula by Krysan and colleagues [2],

$$p = 1 - [1 - (X/120{,}000)]^n,$$

where p is the probability of finding an insertion in a given gene, X is the gene length, n is the number of insertions, and 120,000

is the *A. thaliana* genome length in kilobases. It can be generalized to

$$p = 1 - [1 - (X/L)]^n,$$

where L is the genome length considered. Because it is known that T-DNA insertion sites are preferably located in intergenic regions compared to genic regions, we considered only insertions in gene regions. We detected 88,541 insertions in genes (n = 88,541), L = 58,495 kb (the calculated total gene length in the *A. thaliana* genome). If we want to have $p = 0.95$, we get $X \approx 2$ kb. This means that for a gene 2 kb in length, we should have 95% probability that there will be an insertion in this gene when the total number of insertions in genes is 88,541. Therefore, we chose 2 kb as the length cut-off value for selecting a set of genes without insertions. Of course, the calculation above assumes a random distribution of insertion sites, which is not true. Anyway, genes larger than this size should have a high chance of being covered with insertion sites. A total of 830 genes longer than 2 kb were obtained. Among them the largest gene has a length of 8.3 kb. A detailed list of this gene set ("no-insertion genes") is shown in Supplementary Table S1. When the no-insertion genes and a control set of genes (see Materials and methods) were compared, eight Gene Ontology (GO) terms [19] were found to be overrepresented in the no-insertion set, and there were also overrepresented terms in the control set. It seems that some GO terms are overrepresented in genes within a certain length range. We excluded the two of eight GO terms that also appear to be overrepresented in the control set and list in Table 2 the six unique GO terms that are all in the biological process aspect.

The 830 no-insertion genes are spread all over the five chromosomes of the *A. thaliana* genome (Supplementary Fig. F2). However, there are areas in which the no-insertion genes appear in clusters. Interestingly, one of these clusters is found on the long arm of chromosome 4 and is composed of 12 genes encoding receptor-like kinases (AGI codes from At4g20530 to At4g20640, see Supplementary Table S1). This area spans about 40 kb in which the 12 genes form a large tandem duplicated array. The intergenic region between these 12 no-insertion genes contains fewer than average insertions as well.

We also compared the technically relevant restriction site count (see Materials and methods) for the two sets of genes. The mean restriction site counts are 10.7 for the no-insertion set and 15.8 for the control set. The standard deviations are 5.5 and 7.5, respectively. The difference in the restriction site count between the two gene sets is statistically highly significant ($t = -17.80$, $p < 0.0001$).

To determine if the expression levels of genes are relevant for the availability of T-DNA insertions, we examined the expression levels of two sets of genes using expression profiling data. Genes were first ordered as expressed or nonexpressed based on the present/absent calls (see Material and methods) and then compared with regard to the distribution of expressed or nonexpressed genes in the two gene sets. Table 3 shows the data for stage 12 flowers from the AtGenExpress dataset, which obviously can consider only genes that are included on the Affymetrix ATH1 array. Of 729 genes, 208 are nonexpressed in the no-insertion set, while 276 of 2148 genes are nonexpressed in the control set. Nonexpressed genes are significantly enriched in the no-insertion gene set ($\chi^2 = 95.67$, $p < 0.0001$). For all the genes assigned as expressed in the two sets, their average signal values were compared by independent sample t test, but we did not detect any statistically significant difference. It appears that the important determinant for T-DNA insertions to take place is if the gene is expressed at all and not the absolute value of the expression level. Using other flower stages and tissues we obtained similar results (data not shown).

A similar comparison was done by comparing the EST/cDNA support of these two gene sets. In the set of no-insertion genes, 280 have no support and 550 have support, while in the control set 339 genes have no support and 2007 have support. The genes without EST/cDNA support are enriched in the no-insertion gene set ($\chi^2 = 145.31$, $p < 0.0001$).

Discussion

We have examined the T-DNA insertion data from FLAGdb, GABI-Kat, and SIGnAL, the three main publicly available FST populations for patterns in the insertion distribution. Although the three FST populations differ greatly in size, each population has a significant portion of genes with insertions uniquely represented in the given population. Our data indicate that this is at least partly due to the fact that the FSTs from each population were generated using different restriction enzymes in the production process. The T-DNA flanking sequences were recovered by a similar PCR walking method that involves restriction digestion of the genomic DNA as a common step [5,20,21]. As a result, only insertions close to those restriction sites that are experimentally addressed can be recovered. This technical difference (see Materials and methods for details on the different enzymes used for FST production) results in an increased complementarity in the recovered FSTs from the three populations, so that the total number of genes for which insertion alleles are available is significantly increased. This is further supported by the finding that the technically relevant restriction site counts are significantly lower in a set of no-insertion genes than in the control set. We also note that the negative bias imposed by the restriction enzymes used is significantly reduced in the GABI-Kat population compared to the other populations, most probably because of the use of four-cutter restriction enzymes, which cut more often relative to six-cutters in the genome sequence. However, the payoff for this strategy is that insertions in genes that contain a high number of these sites are not detected simply due to generation of fragments that are too small and therefore FSTs that cannot be assigned to certain loci in the genome. This interpretation is supported by a negative correlation between the technically relevant restriction site count and the insertions in the GABI-Kat dataset. To compensate for this effect at least partially, we have used the restriction enzyme *Csp*6I (recognition site 5′-

GTAC-3′) in addition to the initially selected restriction enzyme *Bfa*I (recognition site 5′-CTAG-3′; [21]) in our FST production pipeline.

Another interesting finding from our analyses is that the T-DNA insertion frequency peaks are actually located at the transcription initiation site and the poly(A) site region and not

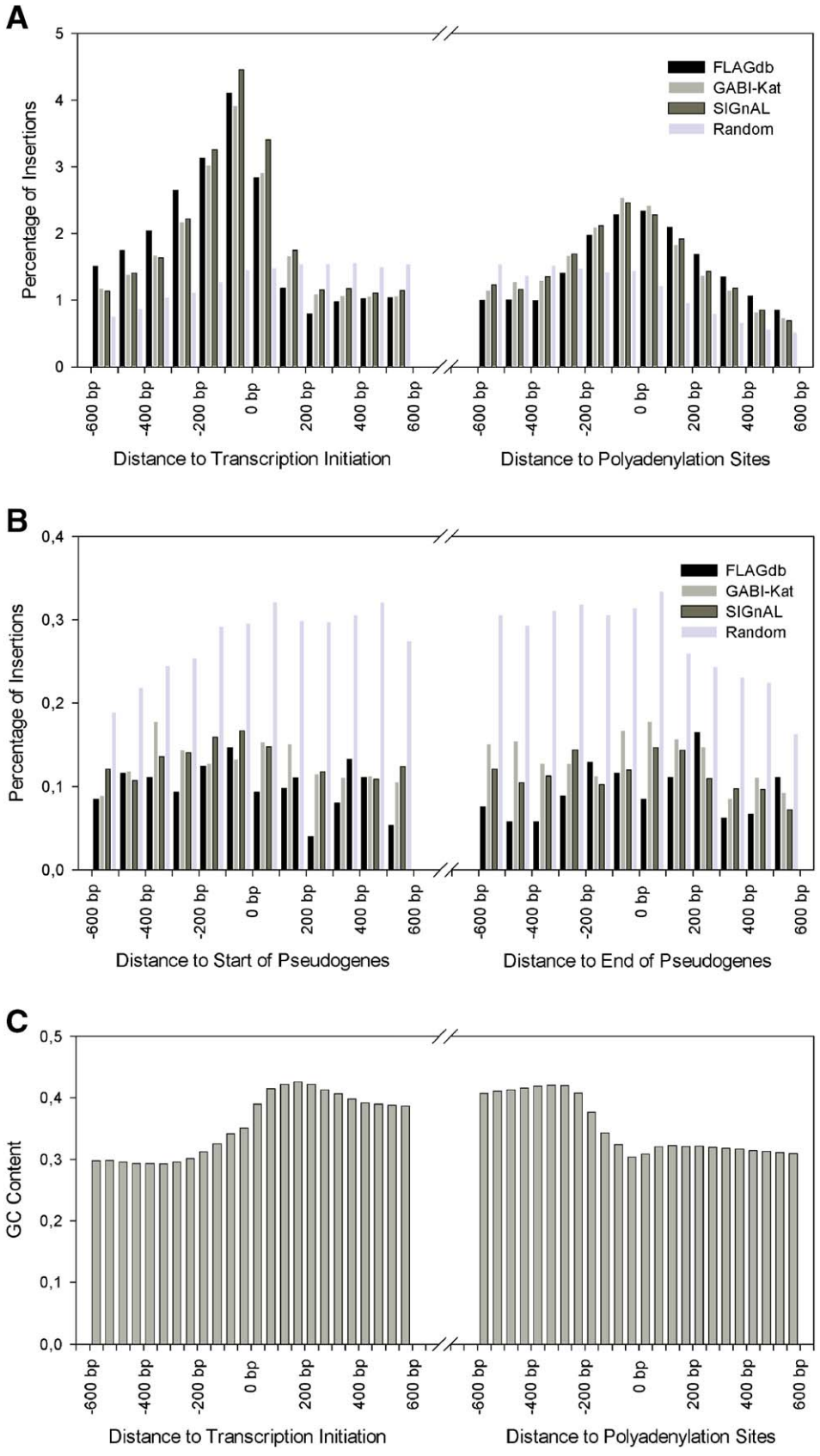

Table 2
Overrepresented Gene Ontology terms in the no-insertion gene set

GO term	GO ID	Aspect	Corrected *p* value	Number of genes annotated
Cell wall organization and biogenesis	GO:0007047	Biological process	0.0056	30
External encapsulating structure organization and biogenesis	GO:0045229	Biological process	0.0056	30
Lipid A biosynthesis	GO:0009245	Biological process	0.0066	6
Lipid A metabolism	GO:0046493	Biological process	0.0066	6
Cell wall biosynthesis (*sensu* Bacteria)	GO:0009273	Biological process	0.0329	10
Peptidoglycan biosynthesis	GO:0009252	Biological process	0.0448	7

around ATG and STOP of translation as previously suggested [11,13]. By comparison with the data from simulated random insertions, clear changes in T-DNA insertion frequency relative to transcription units become obvious. The insertion frequency is higher than random in promoter and intergenic regions and lower than random in the region covered by exons and introns. While several studies have reported that T-DNA integration prefers intergenic regions over genic regions [5,10–12], the data presented here indicate that the transcription initiation and poly (A) regions are crucial places for integration. This finding can also easily explain why the insertion density is associated with gene density, which was observed in earlier studies [5,13,14]. Alonso and colleagues [5] speculated that the preference in promoter and UTRs could be the result of interactions of virulence proteins with the proteins involved in initiation or termination of transcription. Indeed, the recent finding that the VirD2 protein, which is covalently attached to the T-DNA, interacts with TATA-box-binding protein gives additional evidence of this hypothesis [22]. Transcription initiation regions seem to be preferred integration targets not only for T-DNA, but also for viral DNA transfer. Wu and colleagues [23] reported that the murine leukemia virus prefers integration near the transcription initiation sites in the human genome. One plausible explanation is that the opened chromatin structure and pausing state during initiation of transcription as well as during transcript processing and poly(A) addition increase the accessibility of these target sites for the T-DNA or virus integration machinery.

When considered in the context of evolution, it makes sense that the T-DNA insertion machinery avoids protein coding sequences and transcriptionally silent genome regions. The best chance for successful reprogramming of plant cells toward opine production is to make sure that the T-DNA gets expressed well and also does not destroy essential genes. It seems that the optimal way to reach these two goals is to target the promoter regions of active genes.

Analyses of the base composition of sequences at the T-DNA insertion sites from FLAGdb, GABI-Kat, and SIGnAL all showed a preference for low-GC-content regions [5,12]. Brunaud and colleagues [14] proposed that T-DNA integration prefers a T-rich context. Since the GC content is higher in the transcribed region than in intergenic regions (Fig. 2C), AT richness correlates with the more frequent T-DNA insertion in intergenic regions compared to transcribed regions. However, GC content changes cannot explain the insertion peaks around transcription initiation and poly(A) site regions. Also, pseudogenes are generally not transcribed, but they display major sequence features including GC content of functional genes. The lack of insertion frequency peaks observed around pseudogenes further supports the hypothesis that the T-DNA insertion peaks in transcription initiation and poly(A) regions are related to transcriptional activity, rather than direct sequence features.

The chromosome set of the female gametophyte seems to be the target of T-DNA integration [24,25], and it is generally believed that insertions in essential genes that have a lethal effect on the female gametophyte will not be transmitted to the next generation. Recently, Bechtold and colleagues [26] found that the T-DNA insertion appears to occur at the end of the female haploid phase, and this suggests that only mutations in essential genes required for both the female and the male gametophyte will be lost in the next generation. Pagnussat and colleagues [27] identified 120 genes required for female gametophyte development and function from transposon mutant lines in *A. thaliana*. Comparing their gene list to ours, 4 genes appear in the no-insertion set with length greater than 2 kb (830 genes), while 15 are among the 4973 genes without insertions. These numbers are not sufficient to conclude if a relationship exists between essential genes for the female gametophyte and no-insertion genes.

We also searched overrepresented GO terms in the no-insertion genes set, and we found six significantly overrepresented terms

Fig. 2. T-DNA and random integration distributions in the region of 600 bp (A) before and after transcription initiation and poly(A) sites of genes with annotated UTRs and (B) before and after the beginning and end of the ORF of pseudogenes. (C) GC contents in the same regions for genes with annotated UTRs are shown. The *y* axis in A and B indicates the percentage of insertions in the 100-bp bins of all the insertions in each population. Since the insertion frequency is displayed as the percentage of all insertions per population, and because there are many fewer pseudogenes than genes with annotated UTRs, the average random insertion frequency around pseudogenes (B) is lower than that around genes with annotated UTRs (A) accordingly.

Table 3
Distribution of expressed and nonexpressed genes in the no-insertion gene set and control set based on microarray data of stage 12 flowers[a]

	Nonexpressed	Expressed	Total
No-insertion gene set	208	521	729
Control set	276	1827	2148
Total	484	2393	2877

[a] $\chi^2 = 95.67$, $p < 0.0001$.

that are unique in this set. They refer to processes related to cell wall biosynthesis, lipid A metabolism, and peptidoglycan biosynthesis. It is hard to tell if these processes are essential for the female and the male gametophytes. In addition, the number of genes annotated with these terms is relatively small. Thus from these analyses using the current level of functional annotation of the genome we found no link between no-insertion genes and genes with essential functions.

The hypothesis that T-DNA integration prefers transcriptionally active genes has been proposed for more than a decade [15,16]. This seems to be consistent with the recent finding of a high frequency of insertions observed in gene-rich regions and lower frequency around centromeric regions [5,13,14]. However, no significant correlation between the level of gene expression and the frequency of T-DNA integration was observed in the SIGnAL insertion site data [5]. Just recently, Francis and colleagues [28] compared the result of PCR screens (without selection) and kanamycin selection of *A. thaliana* T-DNA transformants and found a higher frequency of transgene silencing in PCR-identified lines. They designated this bias a "selection bias". Our analysis showed that nonexpressed genes are significantly enriched in the no-insertion gene set in comparison to a control set. Although we did not find any difference in expression levels for expressed genes between the two sets of genes, our result does support the hypothesis that T-DNA integration prefers transcriptionally active genes. Because all the FSTs used in our analysis are from transformed plants that passed a selection step, the selection bias could partly explain the observation that genes without insertions are likely to be nonexpressed. In other words, T-DNA might insert into these genes, but due to the position effect (heterochromatin) the selection marker gene in the T-DNA was silenced and therefore transformants never passed the selection, resulting in no detectable insertions in the sequence-indexed mutant collections.

Our findings have important implications for the use of T-DNA in saturation mutagenesis of *A. thaliana*. Different populations complement each other and significantly increase the total number of mutated genes. Despite the higher number of insertions from these populations, we are still far from reaching the goal of having a really useful insertion line for every *A. thaliana* gene. This is mostly due to disfavored integration in the exon-plus-intron region, where insertions are much more likely to cause a knockout allele than in the promoter or downstream of the poly(A) sites. Also, the lack of suitable restriction sites and very low expression levels are reasons many larger genes have no detected insertion alleles. T-DNA insertions detected by the PCR-based screen method are free of the influence of restriction sites that exist in the FST-based populations. Scientists having difficulty in obtaining T-DNA insertion lines in their genes of interest could consider screening other T-DNA populations, for example those of Sussman and colleagues [29] and Rios and colleagues [30], by PCR-based methods. In fact, among 274 genes confirmed only by PCR from the population of Rios and colleagues [30], we found 5 genes that were present among our 830 no-insertion genes (Csaba Koncz, personal communication, Cologne, July 1, 2005). It is worth noting that the SIGnAL and GABI-Kat mutants are of the Col-0 accession, while FLAGdb lines are of the Wassilewskija accession. As a consequence, for combining different alleles into a double or triple mutant, only a fraction of the mutant lines are useful. On the other hand, favored insertions upstream of transcription initiation sites allow a higher chance for activation-tagged lines being generated, as this is the ideal place for activation tagging. As pointed out earlier [12], this is of particular interest in *A. thaliana* research because of functional redundancy generated through genome duplication, and knockout mutants often do not show an obvious phenotype. Therefore gain-of-function mutants by activation tagging are an alternative for the identification of the gene function.

Materials and methods

FST data processing

FSTs of FLAGdb [3] were retrieved from GenBank, GABI-Kat sequences were downloaded from the GABI-Kat Web site (http://www.GABI-Kat.de/ [6]), and SIGnAL sequences were downloaded from the SIGnAL Web site (http://signal.salk.edu/ [5]). The download date was December 20, 2004. TIGR *A. thaliana* chromosome sequences and annotation data v5.0 [18] were downloaded from ftp://ftp.tigr.org/pub/data/a_thaliana/ath1/ and used as the basis for mapping of all FSTs to the genome sequence. There were 26,207 protein-coding genes and 3786 pseudogenes annotated in this dataset. Among the protein-coding genes, 18,094 have annotated UTRs. Each FST was aligned with BLASTN [31] against the five *A. thaliana* pseudomolecules with an *E* value cut-off at 1×10^{-10}. The position on the pseudomolecule corresponding to query start of the best hit was taken as the insertion site. The distance from each predicted insertion site to the beginning or end of the nearest gene (including pseudogenes) was calculated using the annotation data. The beginning and end of the gene are transcription initiation and poly(A) site for genes with annotated UTRs. As a control, 150,000 random insertion sites across the whole genome were generated, and the distance to the beginning or end of the nearest gene was calculated as well. The insertion frequencies (presented as percentage of all insertion sites) in the range before 600 bp and after 600 bp of transcription initiation sites and poly(A) sites of those genes with annotated UTRs were sorted into bins of 100 bp. The same calculation was also performed with regard to pseudogenes. None of the pseudogenes have annotated UTRs; therefore the insertion frequencies were calculated relative to the beginning and end of the ORF itself. The 155 of 3786 pseudogenes with EST/cDNA support were excluded from the analysis.

For each of the three FST populations, a list of protein-coding genes that have at least one insertion in the respective population was generated. We considered insertions only if the insertion site was located between the transcription initiation site and the poly(A) site, for genes with annotated UTR, or between the ATG and the STOP codon for genes without annotated UTR. These regions were also used for calculating gene length and counting restriction

sites. We used a Venn diagram to illustrate the relationships of the three set of genes from the three FST populations [32].

Overrepresented GO terms searching and genome distribution

To build a control set for no-insertion genes, genes having at least one insertion in all of the three populations and a gene length between 2 and 8.3 kb were selected. This set contained a total of 2346 genes. The GO: :TermFinder software [19] was used to scan the two sets of genes to search for overrepresented GO terms that could indicate a functional bias. The GO: :TermFinder was run with default parameters (corrected p value <0.05), and the *Arabidopsis* GO annotation used was taken from TAIR [33].

We used the Chromosome Map Tool at TAIR to draw maps of the distribution of no-insertion genes on *A. thaliana* chromosomes [33].

Restriction site count and GC content calculation

The number of restriction sites was counted in the same way as the T-DNA insertion sites for genes with annotated UTRs, separately for each T-DNA population. The enzymes considered were *Dra*I plus *Eco*RV for FLAGdb [20], *Bfa*I for GABI-Kat [21], and *Hin*dIII plus *Eco*RI for SIGnAL [5]. The GC content was computed in 50-bp sequence windows in the same gene regions. For comparison of the restriction sites count between no-insertion genes and the corresponding control set, the count values for the restriction sites inside the gene for the five enzymes were added up and were compared by t test.

Gene expression level comparison

For comparing the expression levels of the no-insertion gene set and the control set by cross-tabulation, the microarray data of the developmental series of the AtGenExpress project [34] were downloaded from TAIR (ftp://ftp.arabidopsis.org/home/tair/Microarrays/Datasets/ [33]). This dataset contains Affymetrix ATH1 array data from samples of different tissues and developmental stages of *A. thaliana* in triplicate. We focused on flower tissues because flowers are where the T-DNA integration takes place by the floral-dip method [24,25]. The array probe set names were mapped to genes by the Microarray Elements Search tool at TAIR [33]. Genes with fewer than two "present" out of the triplicate set were marked as "nonexpressed", otherwise they were marked as "expressed". Note that nonexpressed must not mean that the genes in question are not expressed at all, it indicates only that the respective expression level is not detectable by the microarray method in the given tissue. The χ^2 test was performed to check if expressed or nonexpressed genes are differently distributed among the no-insertion genes and their controls. For genes marked as expressed, the signal values were averaged and compared between the two sets of genes.

We also extracted the information of whether the gene has EST/cDNA support from the annotation data, since the number of ESTs gives a reasonable approximation of the relative expression level of a given gene [35]. The χ^2 test was used to test if genes with or without EST/cDNA support are differently distributed in the two sets. All the statistics analysis was done using SPSS software (v10.0).

Acknowledgments

We are grateful to FLAGdb/FST, France, and SIGnAL, Salk Institute, USA, for making the T-DNA sequence data public and to Csaba Koncz (MPI for Plant Breeding Research) for sharing unpublished data. We also acknowledge the use of expression profiling data produced by the AtGenExpress Project. Thanks to Martin Werber (MPI for Plant Breeding Research) for helpful discussions and to two anonymous reviewers for valuable suggestions. This work was supported by the BMBF in the context of the German plant genomics program GABI (Förderkennzeichen 0312273).

Appendix A. Supplementary data

Supplementary data associated with this article can be found in the online version at doi:10.1016/j.ygeno.2005.12.010.

References

[1] The Arabidopsis Genome Initiative, Analysis of the genome sequence of the flowering plant *Arabidopsis thaliana*, Nature 408 (2000) 796–815.
[2] P.J. Krysan, J.C. Young, M.R. Sussman, T-DNA as an insertional mutagen in *Arabidopsis*, Plant Cell 11 (1999) 2283–2290.
[3] F. Samson, et al., FLAGdb/FST: a database of mapped flanking insertion sites (FSTs) of *Arabidopsis thaliana* T-DNA transformants, Nucleic Acids Res. 30 (2002) 94–97.
[4] A. Sessions, et al., A high-throughput Arabidopsis reverse genetics system, Plant Cell 14 (2002) 2985–2994.
[5] J.M. Alonso, et al., Genome-wide insertional mutagenesis of *Arabidopsis thaliana*, Science 301 (2003) 653–657.
[6] Y. Li, M.G. Rosso, N. Strizhov, P. Viehoever, B. Weisshaar, GABI-Kat SimpleSearch: a flanking sequence tag (FST) database for the identification of T-DNA insertion mutants in *Arabidopsis thaliana*, Bioinformatics 19 (2003) 1441–1442.
[7] D.W. Meinke, et al., A sequence-based map of *Arabidopsis* genes with mutant phenotypes, Plant Physiol. 131 (2003) 409–418.
[8] B. Tinland, The integration of T-DNA into plant genomes, Trends Plant Sci. 1 (1996) 178–184.
[9] T. Tzfira, J. Li, B. Lacroix, V. Citovsky, Agrobacterium T-DNA integration: molecules and models, Trends Genet. 20 (2004) 375–383.
[10] P.J. Krysan, et al., Characterization of T-DNA insertion sites in *Arabidopsis thaliana* and the implications for saturation mutagenesis, OMICS 6 (2002) 163–174.
[11] M.G. Rosso, et al., An *Arabidopsis thaliana* T-DNA mutagenized population (GABI-Kat) for flanking sequence tag-based reverse genetics, Plant Mol. Biol. 53 (2003) 247–259.
[12] X. Pan, Y. Li, L. Stein, Site preferences of insertional mutagenesis agents in Arabidopsis, Plant Physiol. 137 (2005) 168–175.
[13] L. Szabados, et al., Distribution of 1000 sequenced T-DNA tags in the *Arabidopsis* genome, Plant J. 32 (2002) 233–242.
[14] V. Brunaud, et al., T-DNA integration into the *Arabidopsis* genome depends on sequences of pre-insertion sites, EMBO Rep. 3 (2002) 1152–1157.
[15] C. Koncz, K. Nemeth, G.P. Redei, J. Schell, T-DNA insertional mutagenesis in Arabidopsis, Plant Mol. Biol. 20 (1992) 963–976.
[16] K. Lindsey, et al., Tagging genomic sequences that direct transgene expression by activation of a promoter trap in plants, Transgenic Res. 2 (1993) 33–47.
[17] S. An, et al., Generation and analysis of end sequence database for T-DNA tagging lines in rice, Plant Physiol. 133 (2003) 2040–2047.
[18] J.R. Wortman, et al., Annotation of the *Arabidopsis* genome, Plant Physiol. 132 (2003) 461–468.
[19] E.I. Boyle, et al., GO: :TermFinder—open source software for accessing Gene Ontology information and finding significantly enriched Gene Ontology terms associated with a list of genes, Bioinformatics 20 (2004) 3710–3715.
[20] S. Balzergue, et al., Improved PCR-walking for large-scale isolation of plant T-DNA borders, Biotechniques 30 (2001) 496–504.
[21] N. Strizhov, et al., High-throughput generation of sequence indexes from T-DNA mutagenized *Arabidopsis thaliana* lines, Biotechniques 35 (2003) 1164–1168.
[22] L. Bako, M. Umeda, A.F. Tiburcio, J. Schell, C. Koncz, The VirD2 pilot protein of Agrobacterium-transferred DNA interacts with the TATA box-binding protein and a nuclear protein kinase in plants, Proc. Natl. Acad. Sci. USA 100 (2003) 10108–10113.
[23] X. Wu, Y. Li, B. Crise, S.M. Burgess, Transcription start regions in the human genome are favored targets for MLV integration, Science 300 (2003) 1749–1751.

[24] N. Bechtold, et al., The maternal chromosome set is the target of the T-DNA in the in planta transformation of *Arabidopsis thaliana*, Genetics 155 (2000) 1875–1887.
[25] C. Desfeux, S.J. Clough, A.F. Bent, Female reproduction tissues are the primary target of Agrobacterium-mediated transformation by the Arabidopsis floral-dip method, Plant Physiol. 123 (2000) 895–904.
[26] N. Bechtold, S. Jolivet, R. Voison, G. Pelletier, The endosperm and the embryo of Arabidopsis thaliana are independently transformed through infiltration by *Agrobacterium tumefaciens*, Transgenic Res. 12 (2003) 509–517.
[27] G.C. Pagnussat, et al., Genetic and molecular identification of genes required for female gametophyte development and function in Arabidopsis, Development 132 (2005) 603–614.
[28] K.E. Francis, S. Spiker, Identification of *Arabidopsis thaliana* transformants without selection reveals a high occurrence of silenced T-DNA integrations, Plant J. 41 (2005) 464–477.
[29] M.R. Sussman, R.M. Amasino, J.C. Young, P.J. Krysan, S. Austin-Phillips, The Arabidopsis knockout facility at the University of Wisconsin–Madison, Plant Physiol. 124 (2000) 1465–1467.
[30] G. Rios, et al., Rapid identification of Arabidopsis insertion mutants by non-radioactive detection of T-DNA tagged genes, Plant J. 32 (2002) 243–253.
[31] S.F. Altschul, et al., Gapped BLAST and PSI-BLAST: a new generation of protein database search programs, Nucleic Acids Res. 25 (1997) 3389–3402.
[32] H.A. Kestler, A. Muller, T.M. Gress, M. Buchholz, Generalized Venn diagrams: a new method of visualizing complex genetic set relations, Bioinformatics 21 (2005) 1592–1595.
[33] S.Y. Rhee, et al., The Arabidopsis Information Resource (TAIR): a model organism database providing a centralized, curated gateway to Arabidopsis biology, research materials and community, Nucleic Acids Res. 31 (2003) 224–228.
[34] M. Schmid, et al., A gene expression map of *Arabidopsis thaliana* development, Nat. Genet. 37 (2005) 501–506.
[35] L. Duret, D. Mouchiroud, Expression pattern and, surprisingly, gene length shape codon usage in Caenorhabditis, Drosophila, and Arabidopsis, Proc. Natl. Acad. Sci. USA 96 (1999) 4482–4487.