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.

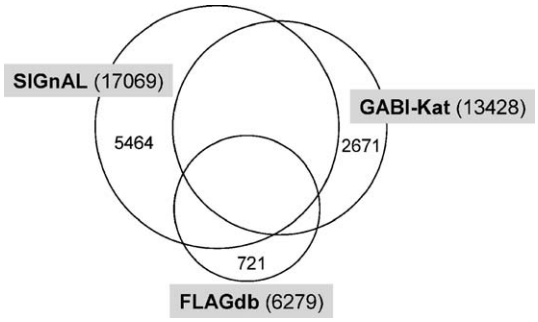

Fig. 1. Venn diagram showing the three sets of genes covered by insertions in each of the three T-DNA populations and their relationships with regard to redundancy of covered genes.

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