EMBO
reports

scientific report

T-DNA integration into the *Arabidopsis* genome depends on sequences of pre-insertion sites

Véronique Brunaud, Sandrine Balzergue, Bertrand Dubreucq[1], Sébastien Aubourg, Franck Samson, Stéphanie Chauvin, Nicole Bechtold[2,+], Corinne Cruaud[3], Richard DeRose[4], Georges Pelletier[2], Loïc Lepiniec[1], Michel Caboche & Alain Lecharny[‡]

URGV, UMR en Génomique Végétale (INRA/CNRS/Université Evry-Val d'Essonne), [3]Génoscope and CNRS UMR 8030, [4]RhoBio, 2 rue Gaston Crémieux, F-91057 Evry, [1]Laboratoire de Biologie des Semences and [2]Station de Génétique et Amélioration des Plantes, INRA, F-78026, Versailles, France

Received July 2, 2002; revised September 16, 2002; accepted September 26, 2002

A statistical analysis of 9000 flanking sequence tags characterizing transferred DNA (T-DNA) transformants in *Arabidopsis* sheds new light on T-DNA insertion by illegitimate recombination. T-DNA integration is favoured in plant DNA regions with an A-T-rich content. The formation of a short DNA duplex between the host DNA and the left end of the T-DNA sets the frame for the recombination. The sequence immediately downstream of the plant A-T-rich region is the master element for setting up the DNA duplex, and deletions into the left end of the integrated T-DNA depend on the location of a complementary sequence on the T-DNA. Recombination at the right end of the T-DNA with the host DNA involves another DNA duplex, 2–3 base pairs long, that preferentially includes a G close to the right end of the T-DNA.

INTRODUCTION

Transferred DNA (T-DNA) from *Agrobacterium tumefaciens* Ti plasmids is a widely used tool for genetic engineering and plant insertional mutagenesis (Galbiati *et al.*, 2000). T-DNA is transferred into plant cell nuclei as a single-stranded molecule attached at the 5′ end to the protein VirD2 and coated with virulence protein E2. T-DNA integrates into the genome (Gelvin, 2000) by illegitimate recombination (Gheysen *et al.*, 1991; Mayerhofer *et al.*, 1991; Zupan *et al.*, 2000) via a largely unknown mechanism. Characterization of a limited number of T-DNA insertions into genes showed an apparently even repartition along *Arabidopsis thaliana* chromosomes with no preferential integration into a specific gene structure (Azpiroz-Leehan and Feldmann, 1997). We have produced flanking sequence tags (FSTs) for >18 000 *A. thaliana* T-DNA transformants (Balzergue *et al.*, 2001). This has allowed an in-depth analysis of the sequence specificity of T-DNA insertion sites (IS) and brings new insights into the integration process.

RESULTS

Distribution of IS in the host genome

The FST distribution in the genome of *A. thaliana* is even throughout the five chromosomes. As an example, the FST distribution along chromosome 3 is given in Figure 1. FSTs are progressively less frequently observed towards the centromere, as shown in Figure 1 with the predicted genes [The Arabidopsis Genome Initiative (AGI), 2000]. About 40% of the integrations are in genes, i.e. in regions defined by the AGI-predicted genes plus 200 base pairs (bp) on each side of them and covering 54% of the genome. There is no apparent category of genes more or less prone to T-DNA insertion. We observed 121 FSTs per Mb in the 200 bp upstream of the start codon and 77 FSTs per Mb in the intergenic regions and 3′ UTR. In genes, FSTs are more frequently found in introns than in exons, with 43 and 33 FSTs, respectively, per Mb.

T-DNA after integration

The sites of VirD2-mediated cleavage of the T-DNA have been determined both *in vivo* (Dürrenberger *et al.*, 1989) and *in vitro* (Pansegrau *et al.*, 1993). FSTs from the T-DNA left border (LB), at the 3′ end of the transferred single-stranded T-DNA, show that in

[+] Present address: Usine des molécules recombinantes, 1020 route de l'église, bureau 600, Sainte Foy, Canada G1V 3V9
[‡] Corresponding author. Tel: +33 1 60 87 45 18; Fax: +33 1 60 87 45 10; E-mail: lecharny@ibp.u-psud.fr

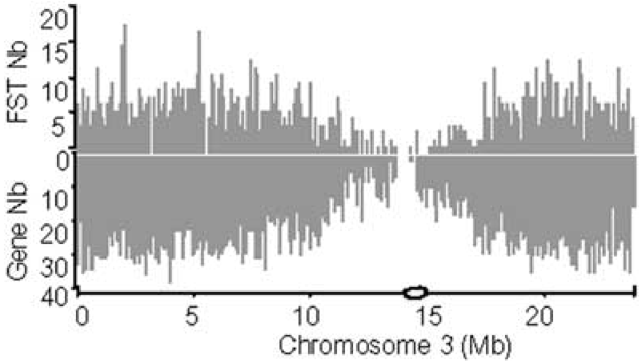

Fig. 1. Gene and FST densities along chromosome 3 of *A. thaliana*. Each peak is a number of FSTs (above) or genes (below) per 100 kb. The centromeric region is represented by a circle on the abscissa.

24% of the cases the T-DNA is integrated with a full-length LB (canonical insertion) (Figure 2A). The number of inserted T-DNAs with an LB truncated by 1–23 bases is relatively even and longer deletions are rare. The right border (RB) of the T-DNA (Figure 2A) is also sometimes fully conserved after integration (19%). The RB is frequently truncated between the second and fifth bases from the canonical IS (36%).

Microsimilarity between the host genome and T-DNA borders

We characterized further the sequences upstream and downstream of 4430 IS. These two regions are defined with respect to the integrated T-DNA: the region upstream of the IS is the region that would be sequenced with a primer identical to a sequence in the LB, and the region downstream of the IS is the region that would be sequenced with a primer designed from the RB. In a region close to the IS, the nucleotide composition is different from the regional composition (Figure 3). We postulated that it might reflect the previously proposed role of microsimilarities (often defined as microhomologies) between host DNA and T-DNA sequences in the integration process (Tinland, 1996). This hypothesis, based on a small number of IS, can be tested with the larger set of IS sequences available: FSTs showing different lengths of deletion in the integrated T-DNA (Figure 2A) might exhibit modified microsimilarity when compared to canonical IS. Consistently, the most frequently observed sequence downstream of the plant IS (Table I, Figure 2B) is related to the sequence at the end of the integrated T-DNA [indicated by arrows (a)–(d) below the T-DNA sequence in the abscissa in Figure 2A]. For instance, when the integrated T-DNA ends at position (c), (Figure 2A), the over-represented nucleotides in the plant IS indicate that the IS consensus sequence is CCCAAC (Table I; Figure 2B downstream), whereas it is AAAAG when the integrated T-DNA ends at position (d). There is, in all cases, an imperfect but striking similarity between the complement of the 3′ sequence of integrated T-DNA and the consensus at the plant IS (Figure 2A and B). Thus, for canonical insertions [case (a) in Figure 2A and B], the consensus sequence 5′-(C/T)(A/C)(G/A)GGA-3′ in the plant genome has some similarity with the complement of the sequence 3′-GTCCT-5′ (i.e. CAGGA) at the end of the T-DNA. The occurrence observed for one nucleotide may be as high as 84.8% at the plant IS (Table I). Nevertheless, the microcomplementarities observed between the T-DNA and plant DNA are not perfect, since, in many cases, two nucleotides may be over-represented at the same position. Indeed, as illustrated for the canonical insertion (Figure 2C), this may be explained by the alignment of two sequences (5′-CAGGAN-3′ and 5′-TCAGGA-3′), complementing the same sequence in the T-DNA LB (3′-GTCCT-5′), but shifted by one nucleotide. This strongly suggests a frequent deletion of one nucleotide downstream of the IS as a consequence of the integration. Inspection of individual sequences downstream of the canonical IS indicates that there is a clear shift due to the presence of one nucleotide before the microsimilarity in 53% of IS, either a T (35%) , an A (11%), a C (6%) or a G (<1%). The relative nucleotidic representation of the first five positions upstream of the IS has been re-computed, taking into consideration, when necessary, the shift of one nucleotide between the apparent and actual IS. Corrected values are 71, 59, 39, 45 and 40% for C, A, G, G and A positions, respectively. Therefore, the consensus sequence of the microsimilarity, at the plant IS, for canonical integrations is clearly 5′-CAGGA-3′. Alignments with no accepted gap of each IS corresponding to canonical T-DNA insertions with the 5′-CAGGA-3′ sequence show that a majority of IS (63%) exhibits an identity of at least 50% with this sequence (Figure 2D). The same alignments with sequences randomly taken from the genome provide only 19% of sequences with this score. Thus, collectively, our data show that, whatever the position of the cut in the T-DNA, there is a microsimilarity between the integrated T-DNA border and the plant IS.

Interestingly, the plant genome sequence upstream of the T-DNA IS shows an over-representation of the nucleotide T (Table I). It is particularly striking when FSTs correspond to canonical insertions at the LB. In this case, upstream of the plant IS there is a highly significant occurrence of T at five contiguous positions that cannot be due to sequence matches between the T-DNA and the plant genome. If, as discussed above, the shift of one nucleotide introduced in the apparent IS by a deletion of one nucleotide is taken into consideration, the T representation becomes 38, 40, 45 and 73% at the –5 to –1 position upstream of IS, respectively. Thus, the T-rich region that frequently ends by a T is located immediately upstream of the microsimilarity region. Half of the five nucleotide sequences upstream of the IS contain at least three Ts.

Lastly, we searched for microsimilarities between the RB of the single-stranded T-DNA and the complement of the plant DNA. We observed a significant over-representation of Gs, in the plant DNA, at position –2 from the IS, either characterized by FSTs from the RB of canonical T-DNA insertions [(e) in Figure 2A and B] or FSTs from T-DNAs nicked between the second and the third bases [(f) in Figure 2A and B). These results show that in both cases the nucleotide G is significantly over-represented, in plant sequences, at 2 bp from the IS and it is preceded by an over-representation of A in the case where the T-DNA end is canonical, tttagcacaCT, and T when the T-DNA end is tttagcaCA.

DISCUSSION

We have generated and analysed a large set of integrated T-DNAs and their respective pre-IS. We confirmed and further

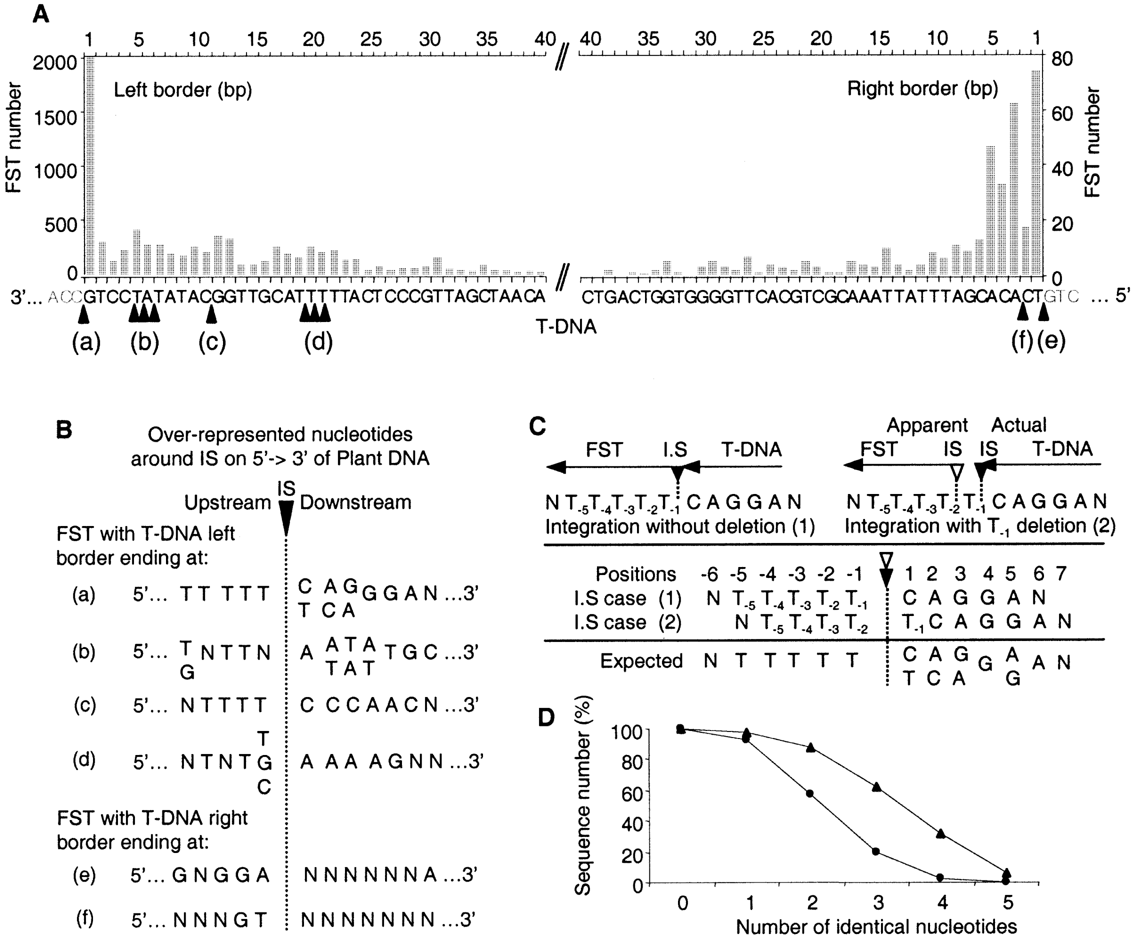

Fig. 2. Deletions in the integrated T-DNA and *A. thaliana* sequences at IS. (**A**) Histogram of the number of FSTs classified on the basis of the ends of the sequence of the integrated T-DNA. 8525 and 394 FSTs were analysed for left and right borders, respectively. The canonical T-DNA sequence runs from (a) to (e). (**B**) The over-representation of nucleotides in the plant sequence before and after IS depends on the end of the integrated T-DNA. (a) to (f) indicate nicking sites as illustrated below the T-DNA sequence in (A). Over-represented nucleotides (shaded boxes in Table I) are shown at five positions downstream and seven positions upstream of IS. (**C**) A deletion of 1 bp upstream of the IS can explain the over-representation of two different nucleotides at different positions in the case of a canonical integration (a). The sequence N(T)$_5$CAGGAN is taken as an example of a plant sequence around an IS. (**D**) Identity scores between the complement of the canonical T-DNA left border, CAGGA and sequences downstream of IS (triangles) or randomly taken from the *A. thaliana* genome (circles).

characterized the involvement of a microsimilarity between the T-DNA LB sequence and host DNA IS previously observed in a limited number of IS (Tinland *et al.*, 1995) and proposed to be the docking force for integration (Tinland, 1996). Interestingly, we found that even in non-canonical insertions a microsimilarity with the host DNA is also present. Most of the microsimilarities involved the first 25 nucleotides of the LB, but some are observed up to the sequence of the oligonucleotide used to prime the FST sequencing. As a consequence of the small length of the microsimilarity and the large region of the T-DNA in which it may be found, the T-DNA may potentially integrate anywhere in the plant genome. However, we showed an over-representation of Ts at five positions immediately upstream of the IS, indicating that T-DNA integration is influenced by the nucleotide composition at the pre-IS independently of any microsimilarity. The preference for a T-rich context for T-DNA integration may explain favoured integrations in the gene region upstream of the start codon, as well as the higher density of FSTs found in introns than in exons. Our data have also increased our knowledge on the recombination reaction at the 5′ end of the T-DNA. Previous results (Tinland *et al.*, 1995) indicated that an identity existed between at least the nucleotide linked to VirD2 and the last nucleotide of the plant IS. We extended these

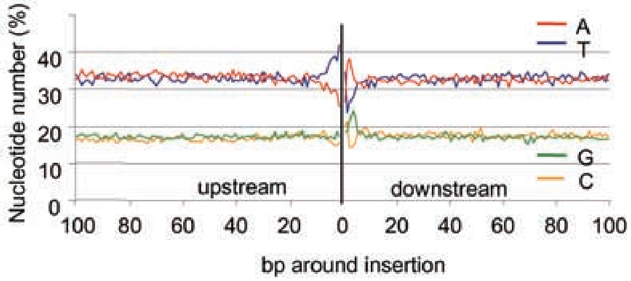

Fig. 3. Relative base composition of the *A. thaliana* DNA at T-DNA IS. IS were mapped in the *A. thaliana* genome by sequence comparison between 4330 FSTs obtained from the T-DNA LB and not flanked by filler DNA.

Table I. Sequences around T-DNA IS in the *A. thaliana* genome

		Integrated T-DNA													
		–6	–5	–4	–3	–2	–1	IS	1	2	3	4	5	6	7
Left border (3′-ACC$_{a}$GTCC$_{b}$T$_{b}$A$_{b}$TATAC$_{c}$GGTTGCAT$_{d}$T$_{d}$T$_{d}$TTA-5′)															
Ending at (a)	T	34.4	**38.4**	**40.2**	**41.5**	**41.5**	**55.2**		**39.4**	***12.1***	***21.9***	***24.0***	***27.1***	31.1	35.2
(762)	C	18.2	17.1	15.5	*14.8*	16.1	*14.7*		**45.2**	**41.3**	***6.6***	***10.8***	***11.3***	***14.0***	18.4
	G	16.3	15.7	18.4	17.4	17.7	***7.6***		***1.6***	***7.3***	**27.9**	**44.4**	**31.9**	17.3	16.0
	A	31.1	28.9	***26***	***26.3***	***24.7***	***22.6***		***13.9***	**39.4**	**43.6**	***20.8***	29.7	**37.6**	30.5
Ending at (b)	T	35.1	39.0	38.0	**39.8**	**40.3**	36.7		36.4	**43.7**	**46.9**	**45.8**	**42.2**	32.5	33.3
(486)	C	19.4	***11.0***	13.9	14.9	*12.8*	19.4		***8.6***	***3.9***	***6.3***	***6.5***	***10.2***	16.2	**23.0**
	G	15.5	**21.7**	18.1	17.5	18.9	19.6		***10.2***	***5.2***	***7.1***	***8.6***	17.0	**22.2**	17.5
	A	30.1	*28.3*	30.1	*27.8*	*28.0*	***24.4***		**44.8**	**47.1**	**39.8**	39.0	30.6	29.1	***26.4***
Ending at (c)	T	37.7	35.9	*39.9*	**43.5**	39.0	**45.7**		30.5	***8.5***	***14.8***	***25.1***	***24.2***	30.0	32.7
(223)	C	21.1	21.5	*13.0*	16.6	13.9	16.6		**50.2**	**84.8**	**44.4**	***8.5***	19.3	**23.8**	17.9
	G	16.1	13.9	16.1	15.7	15.3	***9.9***		***1.0***	***2.2***	***7.2***	13.5	15.7	17.0	19.7
	A	***25.1***	28.7	30.9	***24.2***	31.8	27.8		***18.4***	***4.5***	33.6	**52.9**	**40.8**	29.2	29.6
Ending at (d)	T	33.0	33.3	38.4	36.4	38.9	38.4		***22.7***	***7.0***	***20.2***	32.2	***25.5***	28.6	32.7
(357)	C	18.2	18.2	16.5	14.6	15.1	**22.7**		13.4	***3.6***	***8.1***	***9.5***	16.8	18.2	15.4
	G	17.4	13.4	15.4	17.6	14.3	**20.7**		20.7	***9***	***9.5***	*15.7*	**25.2**	17.9	16.5
	A	31.4	35.0	29.7	31.4	31.6	***18.2***		**43.1**	**80.4**	**62.2**	**42.6**	32.5	35.3	35.3
Right border (3′-TTTAGCACA$_{f}$CT$_{e}$GTC-5′)															
Ending at (e)	T	31.4	35.3	37.3	25.5	***5.9***	*29.4*		29.4	33.3	43.1	35.3	21.6	37.3	23.5
(51)	C	13.7	11.8	9.8	19.6	19.6	*5.9*		9.8	13.7	9.8	17.7	17.7	23.5	17.6
	G	15.7	27.5	23.5	**33.3**	**35.3**	17.7		25.5	15.7	17.7	21.6	15.7	11.8	*7.8*
	A	39.2	25.5	29.4	21.6	39.2	47.1		35.3	37.3	29.4	25.5	45.1	27.5	**51.0**
Ending at (f)	T	34.8	34.8	21.7	43.5	34.8	**56.5**		17.4	26.1	34.8	21.7	17.4	30.4	34.8
(23)	C	17.4	17.4	17.4	17.4	*4.4*	8.7		30.4	26.1	*4.4*	17.4	21.7	8.7	21.7
	G	13.0	17.4	30.4	17.4	**43.5**	8.7		13.4	13.4	21.7	17.4	8.7	8.7	17.4
	A	34.8	30.4	30.4	21.7	17.4	26.1		39.1	34.8	39.1	43.5	52.2	52.2	26.0

Relative amounts in the four nucleotides are given for positions around the IS and for integrated T-DNA either canonical [(a) at LB and (e) at RB] or deleted up to positions (b)–(d) at their LB or position (e) at their RB (see Figure 2A). At some positions, we observed nucleotides that are over-represented (grey boxes), under-represented (in italics) or present at the expected number (roman text) as compared with the average nucleotidic composition in the *A. thaliana* genome (T, 32.8; C, 17.1; G, 17.1; A, 33.0). The statistical significance level is indicated as follows: numbers underlined and in bold, $P \leq 0.01$; numbers in bold, $P = 0.05–0.01$; numbers in standard text, $P = 0.10–0.05$. The number of sequences used in each case is indicated in parentheses. The over-represented nucleotides are reported in Figure 2B.

observations and demonstrated that the identity between the 5′ end of the T-DNA and the plant pre-IS often involved the last nucleotide but also a G located immediately downstream. Our data statistically support the model for T-DNA integration previously proposed by Tinland (1996). Taking all our new data into consideration, we propose a model that mainly differs from Tinland's model by the preference for T-DNA integration in the vicinity of a T-rich region (Figure 4). The following five steps are involved. (1) The integration process, often initiated by the 3′ (LB) of the T-DNA invading a poly T-rich site of the host DNA. (2) Upstream of the 3′ end of the single-stranded T-DNA, a more or less perfect duplex with the top strand of the host DNA is formed. Our findings are the first proof of a link between the location of the microsimilarity in the T-DNA and the cut in the LB of the

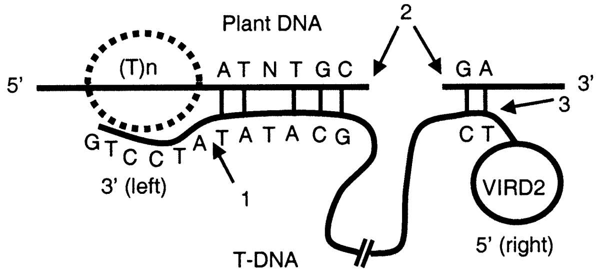

Fig. 4. A model of the T-DNA integration process. The given example is for a putative T-DNA that, as a consequence of its integration is deleted by six nucleotides in the 3′ LB [position (b) in Figure 2A] but complete in its RB. For convenience, only the upper strand of the host DNA is represented. A T-rich region [T(n)] operates as a preferential site of entry of the T-DNA LB. Starting from its 3′ end, the T-DNA scans the plant DNA until it finds a microcomplementarity just downstream of the T-rich region. A nick (2) is generated in the host DNA downstream of the microcomplementarity-based duplex and used as a priming site to synthesize the complementary strand of the T-DNA until the 5′ RB covalently linked to VirD2 is reached. The integration process of the newly synthesized double-stranded DNA frequently brings about a deletion in the host DNA. Recombination between the host bottom strand and the T-DNA operates at sites 1 and 3 after the action of exonucleases on both ends of the T-DNA.

integrated T-DNA. (3) After degradation of the 3′ end portion of the T-DNA downstream of the duplex, the ligation between the digested bottom strand of the host DNA and the 3′ end of the T-DNA is performed by host enzymes. We assume that the 1 bp deletion frequently observed in the plant DNA is a consequence of this digestion-ligation step. (4) A nick in the upper host DNA strand is created downstream of the duplex and used to initiate the synthesis of the complementary strand of the invading T-DNA. The imperfect matches in the duplex are detected and repaired by host enzymes, using the invading T-DNA sequence as a template. (5) The right end of the T-DNA is ligated to the bottom strand of the host DNA. The pairing frequently involves a G and another nucleotide upstream of it. (6) The top strand of the host DNA is degraded between the two microsimilarities and a ligation with the synthesized complementary T-DNA is made. This may result in a deletion of variable length in the host DNA. Out of 180 transformants for which we have FSTs from both sides of the integrated T-DNA, 88 have apparent deletions shorter than 150 bases.

Some of the results presented here differ from previously published data, and one of the possible explanations is that for the first time we used a set of FSTs not biased by an analysis of a particular set of mutants. Secondly, we used a large set of data, enough to be statistically representative of the integration process.

Speculation

The insertion of T-DNA into the genome, may have recruited some of the cellular processes involved in illegitimate recombination (Tzfira and Citovsky, 2002). We demonstrate that the primary docking force of the T-DNA towards the plant IS may be a poly T-rich stretch in the plant DNA. An observed over-representation of T-rich sequences has been observed in other microsimilarity-mediated recombinations (Kohli *et al.*, 1999). An A-T-rich sequence is a region with both a low DNA duplex stability (Breslauer *et al.*, 1986) and a strong bending (Bolshoy *et al.*, 1991). Bending rather than sequence itself has been shown to favour retroviral integration *in vitro* (Müller and Varmus, 1994) and P transposable element integration in the *Drosophila* genome (Liao *et al.*, 2000). Recognition of a bended DNA region might, therefore, be a common feature in the integration of foreign DNA in eucaryote genomes.

METHODS

The collection of T-DNA transformant lines has been generated with the *A. thaliana* ecotype Wassilewskija at the Institut National de la Recherche Agronomique (INRA Versailles), using the *A. tumefaciens* strain C58C1 (pMP90; Bechtold *et al.*, 1993). The protocol used for obtaining FSTs is described in a previous paper (Balzergue *et al.*, 2001), and details on FST sequences are available through the FLAGdb/FST database (Samson *et al.*, 2002; http://genoplante-info.infobiogen.fr). The FST set analysed contained 8919 non-redundant sequences unequivocally mapped at only one locus of the plant genome. This FST set did not contain FSTs indicative of complex IS such as tandem insertions of two T-DNA. Only FSTs not flanked by filler DNA were used. About 20% of FSTs contain stretches of bases downstream of the end of the inserted T-DNA that do not match plant, plasmid or any known sequences. In most of cases, this filler DNA is shorter than 50 bases and its nucleotide composition is the same as the average in the *A. thaliana* genome. The presence of DNA filler at sites of non-homologous recombination is thought to be a consequence of DNA break repairing. BLAST programs (Altschul *et al.*, 1997) were used to align FSTs with the five chromosomes of *A. thaliana*. The pseudo-molecules were downloaded from the TIGR site (http://www.tigr.org), with the associated coding sequence annotations (ID 68170, 51595, 68173, 68164 and 68172). A gene is interrupted by a T-DNA when a FST starts in the genomic region, including the predicted CDS (coding sequence) for this gene and 200 bp on each side.

ACKNOWLEDGEMENTS

The authors are grateful to Mickael Hodges and Ian Small for reading the manuscript. This work was supported by the French Génoplante Program.

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.

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. *et al.* (2001) Improved PCR-Walking for large-scale isolation of plant T-DNA borders. *Biotechniques*, **30**, 496–504.

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

Bolshoy, A., McNamara, P., Harrington, R.E. and Trifonov, E.N. (1991) Curved DNA without A-A: experimental estimation of all 16 DNA wedge angles. *Proc. Natl Acad. Sci. USA*, **88**, 2312–2316.

Breslauer, K.J., Frank, R., Blocker, H. and Marky, L.A. (1986) Predicting DNA duplex stability from the base sequence. *Proc. Natl Acad. Sci. USA*, **83**, 3746–3750.

Dürrenberger, F., Crameri, A., Hohn, B. and Koukolfkova-Nicola, Z. (1989) Covalently bound VirD2 protein of *Agrobacterium tumefaciens* protects the T-DNA from exonucleolytic degradation. *Proc. Natl Acad. Sci. USA*, **86**, 9154–9158.

Galbiati, M., Moreno, M.A., Nadzan, G., Zouyrelidou, M. and Dellaporta, S.L. (2000) Large-scale T-DNA mutagenesis in *Arabidopsis* for functional genomic analysis. *Funct. Integr. Genom.*, **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.

Gheysen, G., Villarroel, R. and Van Montagu, M. (1991) Illegitimate recombination in plants: a model for T-DNA integration. *Genes Dev.*, **5**, 287–297.

Kohli, A., Griffiths, S., Palacios, N., Twyman, R.M., Vain, P., Laurie, D.A. and Christou, P. (1999) Molecular characterization of transforming plasmid rearrangements in transgenic rice reveals a recombination hotspot in the CaMV 35S promoter and confirms the predominance of microhomology mediated recombination. *Plant J.*, **17**, 591–601.

Liao, G.C., Rhem, E.J. and Rubin, G.M. (2000) Insertions site preferences of the P transposable element in *Drosophila melanogaster*. *Proc. Natl Acad. Sci. USA*, **97**, 3347–3351.

Mayerhofer, R. *et al.* (1991) T-DNA integration: a mode of illegitimate recombination in plants. *EMBO J.*, **10**, 697–704.

Müller, H.P. and Varmus, H.E. (1994) DNA bending creates favored sites for retroviral integration: an explanation for preferred insertion sites in nucleosomes. *EMBO J.*, **13**, 4704–4714.

Pansegrau, W., Schoumacher, F., Hohn, B. and Lanka, E. (1993) Site-specific cleavage and joining of single-stranded DNA by VirD2 protein of *Agrobacterium tumefaciens* Ti plasmids: analogy to bacterial conjugation. *Proc. Natl Acad. Sci. USA*, **90**, 11538–11542.

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. *Nucleic Acids Res.*, **30**, 94–97.

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

Tinland, B. (1996) The integration of T-DNA into plant genomes. *Trends Plant Sci.*, **1**, 178–184.

Tinland, B., Schoumacher, F., Gloecker, V., Bravo-Angel, A.M. and Hohn, B. (1995) The *Agrobacterium tumefaciens* virulence D2 protein is responsible for precise integration of T-DNA into the plant genome. *EMBO J.*, **14**, 3585–3595.

Tzfira, T. and Citovsky, V. (2002) Partners-in-infection : host proteins involved in the transformation of plant cells by *Agrobacterium*. *Trends Cell Biol.*, **12**, 121–129.

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.

DOI: 10.1093/embo-reports/kvf237