The EMBO Journal vol.10 no.3 pp.697-704, 1991

T-DNA integration: a mode of illegitimate recombination in plants

Reinhold Mayerhofer[1], Zsuzsanna Koncz-Kalman[1], Christiane Nawrath[1], Guus Bakkeren[2], Andreas Crameri[2], Karel Angelis[3], George P.Redei[4], Jeff Schell[1], Barbara Hohn[2] and Csaba Koncz[1,5]

[1]Max-Planck-Institut für Züchtungsforschung, D-5000 Köln 30, Carl-von-Linne-Weg 10, FRG, [2]Friedrich-Miescher-Institut, CH-4002 Basel, PO Box 2543, Switzerland, [3]Institute of Experimental Botany, Czechoslovak Academy of Sciences, CS-15 000 Praha 5, CSFR, [4]Department of Agronomy, Curtis Hall, University of Missouri, Columbia, MO 65211, USA and [5]Institute of Plant Physiology, Biological Research Center of Hungarian Academy of Sciences, H-6701 Szeged, PO Box 521, Hungary

Communicated by B.Hohn

Transferred DNA (T-DNA) insertions of *Agrobacterium* gene fusion vectors and corresponding insertional target sites were isolated from transgenic and wild type *Arabidopsis thaliana* plants. Nucleotide sequence comparison of wild type and T-DNA-tagged genomic loci showed that T-DNA integration resulted in target site deletions of 29–73 bp. In those cases where integrated T-DNA segments turned out to be smaller than canonical ones, the break-points of target deletions and T-DNA insertions overlapped and consisted of 5–7 identical nucleotides. Formation of precise junctions at the right T-DNA border, and DNA sequence homology between the left termini of T-DNA segments and break-points of target deletions were observed in those cases where full-length canonical T-DNA inserts were very precisely replacing plant target DNA sequences. Aberrant junctions were observed in those transformants where termini of T-DNA segments showed no homology to break-points of target sequence deletions. Homology between short segments within target sites and T-DNA, as well as conversion and duplication of DNA sequences at junctions, suggests that T-DNA integration results from illegitimate recombination. The data suggest that while the left T-DNA terminus and both target termini participate in partial pairing and DNA repair, the right T-DNA terminus plays an essential role in the recognition of the target and in the formation of a primary synapsis during integration.

Key words: DNA recombination and repair/T-DNA integration model/T-DNA target sites and border junctions/transgenic *Arabidopsis*

Introduction

Virulence (*vir*) genes located on *Agrobacterium* Ti and Ri plasmids encode an inducible DNA processing system that mediates the transfer of any DNA sequence flanked by specific 25 bp direct repeats into plants, where the transferred DNA (T-DNA) is integrated into the nuclear genome (see reviews by Zambryski, 1988; Zambryski *et al.*, 1989). The function of these 25 bp repeats is analogous to that of conjugational transfer origins of bacterial plasmids (Wang *et al.*, 1984; Buchanan-Wollaston *et al.*, 1987). During DNA transfer, induced by plant phenolic compounds, the 25 bp repeats are processed by a complex of VirD1 topiosomerase and VirD2 endonuclease that produces single-strand nicks and double-strand breaks at the T-DNA borders in *Agrobacterium* (Stachel *et al.*, 1985; Stachel and Zambryski, 1986; Winans *et al.*, 1986; Yanofski *et al.*, 1986). Major and minor nick sites are located respectively at the third and first nucleotide positions of the 25 bp borders in the lower strand of the T-DNA (Albright *et al.*, 1987; Wang *et al.*, 1987). The VirD2 subunit of the complex remains covalently attached to the 5′ terminus of the nick site at the right T-DNA terminus (Herrera-Estrella *et al.*, 1988; Young and Nester, 1988; Dürrenberger *et al.*, 1989; Howard *et al.*, 1989). The lower strand of the T-DNA (T-strand) is released by strand-displacement DNA synthesis (Stachel *et al.*, 1986; Albright *et al.*, 1987). The T-strand is probably bound along its entire length and protected against nucleases by a single-stranded DNA-binding (SSDB) protein, VirE2 (Gietl *et al.*, 1987; Citovsky *et al.*, 1989). A second product of processing in induced agrobacteria is a linear double-stranded T-DNA that may give rise to a circular intermediate by recombination of the DNA ends (Koukolikova-Nicola *et al.*, 1985; Jayaswal *et al.*, 1987; Dürrenberger *et al.*, 1989). Since *vir*E2 mutations influence the T-DNA transfer it is believed, but not proven, that the T-strand is probably transferred to plant cells, by a conjugation-like process, through membrane pores formed by VirB proteins (Christie *et al.*, 1988; Ward *et al.*, 1988).

It is not known whether the T-DNA is integrated into the plant genome via a single- or double-stranded DNA intermediate. Agroinfection experiments with T-DNAs carrying genomes of either single- or double-stranded plant virus DNAs in either orientation suggest that the T-strand can be converted to a double-stranded form in bacteria or in plants (Grimsley *et al.*, 1987; Bakkeren *et al.*, 1989). Genetic mapping indicates that T-DNA insertions in plant chromosomes are distributed randomly (Chyi *et al.*, 1986; Wallroth *et al.*, 1986). Notwithstandingly, the high frequency of gene fusions observed after integration of promoterless reporter genes linked to the right T-DNA border indicates that T-DNAs are preferentially integrated into potentially transcribed genomic loci in different plant species (Koncz *et al.*, 1989). Deletions, inversions and duplications of target plant DNA sequences occur during T-DNA transformation (for reviews see Binns and Thomashow, 1988; Zambryski, 1988) and suggest that T-DNA integration might rely on the endogenous recombination system of plants. A possible approach to study the mechanism of T-DNA integration is the comparison of the structure of plant genomic loci before

and after receiving a T-DNA insert (Gheysen *et al.*, 1987).

Here we describe the nucleotide sequence comparison of genomic boundaries and target sites of seven T-DNA inserts in *Arabidopsis thaliana* and show that the T-DNA ends play an important role in the integration events. The data support the notion that T-DNA integration is an example of illegitimate recombination in plants. The integration mechanism probably involves host functions and possibly also the VirD2 protein that mediate the formation of initial synapsis, partial homologous pairing and DNA repair of the junctions between T-DNA inserts and target plant DNA sequences.

Results

Cloning of genomic boundaries and target sites of T-DNA insertions

Transgenic *Arabidopsis* plants were obtained by transformation with *Agrobacterium* gene fusion vectors pPCV621 and pPCV6NFHyg which transfer a promoterless aminoglycoside phosphotransferase [*aph*(3')II] reporter gene linked to the right T-DNA border, a selectable hygromycin resistance gene (*hph*) and a pBR322 plasmid replicon into the plant genome (Koncz *et al.*, 1989). To avoid the bias of nonrandom sample selection and problems caused by possible rearrangements of multiple T-DNA copies after integration, a total of 200 plants was screened for the expression of the *aph*(3')II reporter gene in diverse tissues and the copy number of T-DNA insertions was determined in 48 plants by DNA hybridization. Seven plants carrying single T-DNA inserts and either active or silent *aph*(3')II reporter genes were used in these studies. T-DNA insertions induced transcriptional (plants 621-36, 37 and x34) and translational (plants N6H-c27 and cs4) *aph*(3')II gene fusions that were active either in all vegetative organs (621-36 and x34), or in leaves (N6H-c27), or in stem (N6H-cs4) or in root hairs (621-37) only. No *aph*(3')II gene expression was detected in plants 621-2 and N6H-14. T-DNA insert N6H-14 caused a mutation affecting photosynthesis that was mapped to the *ch*-42 (*chlorata*) locus (Koncz *et al.*, 1989, 1990).

After isolation of T-DNA inserts from transgenic plants by plasmid rescue, the nucleotide sequence of plant DNA fragments flanking the rescued T-DNAs was determined (see Materials and methods). Plant nuclear DNA sequences from untransformed plants corresponding to the sites of T-DNA insertions in isogenic transformants were isolated either by plaque hybridization or by *Taq* polymerase chain reaction (PCR) from a wild type *Arabidopsis* genomic DNA library made in the lambda EMBL 4 vector. Nucleotide sequences of T-DNA insertional target sites and border junctions were deposited at the EMBL and connected DNA sequence databanks.

Nucleotide sequence comparison of wild type and T-DNA-tagged genomic loci

Physical mapping and characterization of genes identified previously by T-DNA inserts 621-37 and N6H-14 (Koncz *et al.*, 1989, 1990) and similar mapping of other insertions (data not shown) indicated that T-DNA integration did not cause large deletions and rearrangements in the *Arabidopsis* genome. Nucleotide sequence comparison of wild type and T-DNA tagged genomic loci, however, demonstrated that all T-DNA insertions induced small deletions at the integration sites without further alteration of flanking plant DNA sequences. The size of deleted plant DNA sequences varied from 29 to 73 bp (Figures 1 and 2). These deleted sequences will further be referred to as target sites.

Comparison of the nucleotide sequence of target sites and those plant DNAs flanking isolated T-DNA inserts, failed to reveal any consensus sequence for T-DNA integration. The occurrence of AT-rich sequences within and around the target sites was characteristic, but not significant, because of the high AT-content (60%) of sequenced *Arabidopsis* genomic DNA fragments.

Five out of seven transgenic plants carried full-length T-DNA inserts that retained segments of both 25 bp border repeats. The fact that in one case 1 and in four cases 3 bases from the right 25 bp border were retained, correlated with the position of major and minor nick sites identified during T-DNA processing in *Agrobacterium* (Wang *et al.*, 1987). This suggested that the right T-DNA end remained intact during transfer and integration. The left border junctions showed more variation. The presence of 22 bp from the 25 bp repeat in the left junction of inserts N6H-14 and cs4 indicated that in these cases the left T-DNA end had remained intact after processing in *Agrobacterium* and integration in plant DNA. No border sequences formed by recombination between the left and right 25 bp T-DNA repeats were detected, in agreement with the conclusion of Bakkeren *et al.* (1989) that circular double-stranded T-DNA intermediates were not involved in the DNA transfer and integration events.

Precise replacements of target sequence

Recombinational exchanges resulting in a defined deletion being replaced by an insert whose break-points are identical with those of the deletion are referred to as 'precise target replacements'. Such events have been observed in plants 621-x34, 36, 37 and N6H-cs4 where T-DNA inserts precisely replaced the target site deletions. Two types of integration events could be distinguished: one caused deletions of the T-DNA ends (in plants 621-x34 and 36), while the other retained segments of the T-DNA 25 bp borders (in plants 621-37 and N6H-cs4; Figure 1).

In cases where deletions removed the T-DNA ends, 5–7 identical nucleotides were observed at the corresponding break-points of target deletions and T-DNA insertions (Figure 1, 621-x34 and 621-36). In plant 621-x34, deletions removed 9 and 33 bp beyond the left and right 25 bp T-DNA borders respectively. ATTGT and TC.GGG terminal sequences of truncated 621-x34 T-DNA were identified at the break-points of the corresponding target site deletion. In plant 621-36 a deletion eliminating 78 bp from the left T-DNA end terminated at an AATTTT internal T-DNA sequence. The same sequence was found at the left break-point of the target deletion. The right end of T-DNA insert 621-36 became identical to the TATGTTT right terminus of the target by deletion of a CAGT internal T-DNA sequence.

When the right T-DNA–plant DNA junctions contained 1 or 3 bp from the 25 bp border (as in plants 621-37 and N6H-cs4, Figure 1), a longer segment of homology was observed only between the left break-points of target deletions and of T-DNA left ends. TATT.AATT and CAG-GAT sequences representing the left ends of T-DNA inserts 621-37 and N6H-cs4 respectively were also detected at the left break-points of the corresponding target deletions.

In addition to identical nucleotides at these termini,

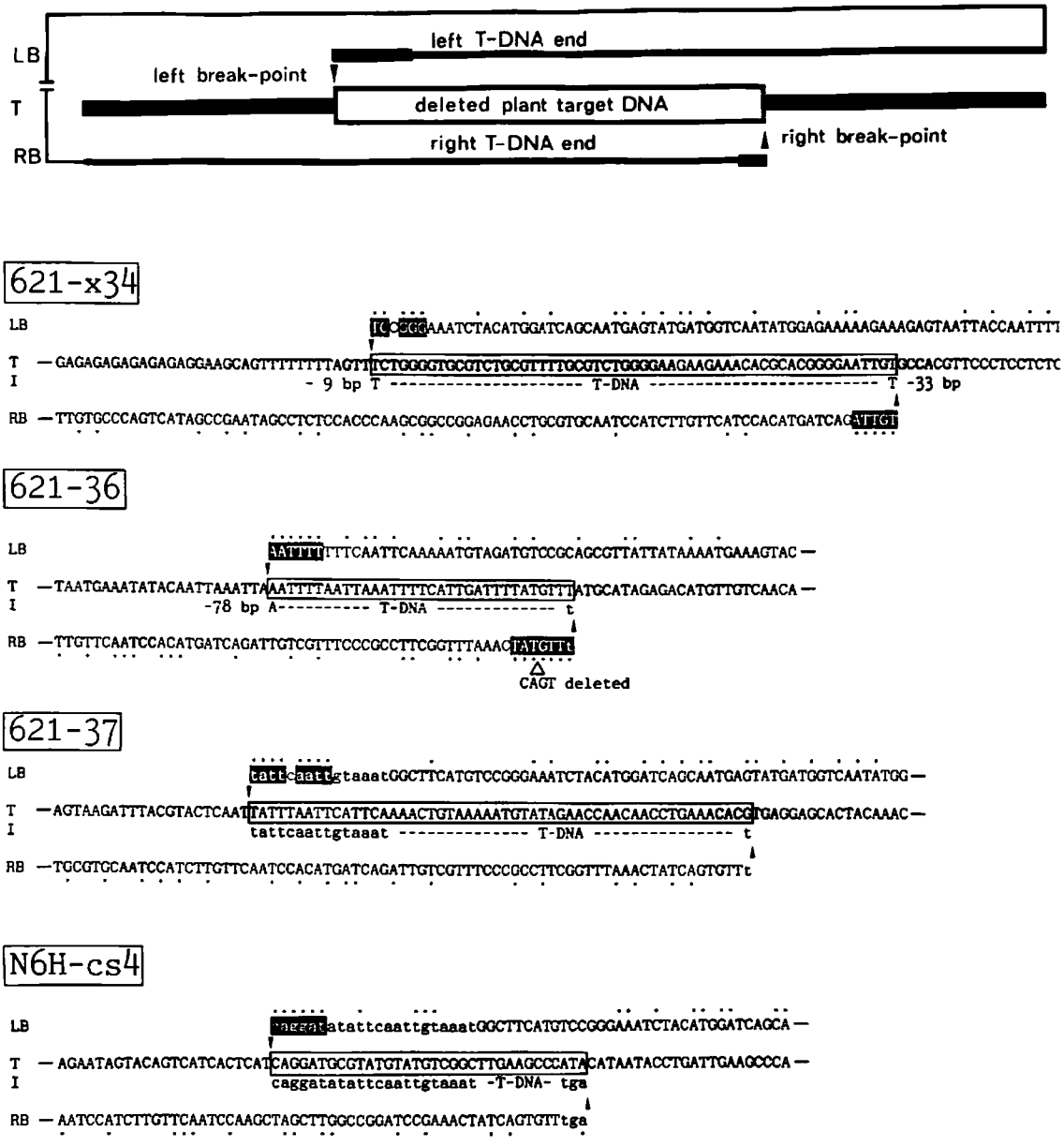

Fig. 1. Precise replacements of target sequence. Nucleotide sequence comparison of target sites and border junctions of T-DNA inserts 621-x34, 36, 37 and N6H-cs4. The scheme at the top illustrates the rationale used to present the DNA sequences below. (This scheme does in no way imply the possible occurrence of a triple-strand intermediate during T-DNA integration and only serves to orient the reader with regard to the presentation of the sequence data.) Termini of the T-DNAs are positioned to correspond with the left and right ends of the deleted plant target DNA. T-DNA sequences starting at the precise break-point of the left T-DNA end observed in a particular transgenic plant (e.g. 621-x34 etc.) and reading into the T-DNA insert are depicted in lines (LB) and were aligned to the last nucleotides remaining from the target DNA sequence and seen in the T-DNA–plant DNA junctions (line T). A similar approach was used to present sequence data of the right ends of T-DNA inserts (depicted in line RB) aligned with corresponding break-points in the target plant DNA. Line T depicts the sequence of plant target DNA before T-DNA integration. Plant DNA sequences actually deleted during T-DNA integration events are boxed. Break-points of target deletions and T-DNA insertions are marked by black vertical arrows. Line I underneath each plant DNA target site describes the actual size of each T-DNA insert and identifies the last nucleotide at the ends of each T-DNA insert. Nucleotides derived from the 25 bp border repeats are printed in lower case. Left 25 bp border repeats of T-DNAs derived from gene fusion vectors pPCV621 and pPCV6NFHyg (Koncz *et al.*, 1989) are identical to the left border of the T_L-DNA of octopine Ti plasmid Ach5, while right border repeats of these T-DNAs are identical to the right 25 bp border repeat of the T-DNA of nopaline Ti plasmid C58 (Koncz and Schell, 1986). Terminal sequences, that in the nucleotide sequence alignment were found to be homologous between corresponding break-points of T-DNA insertions and target deletions, are printed in black background. Sequence homologies throughout alignments of T-DNA ends and plant DNAs are labelled by dots in lines LB and RB. Position of an internal CAGT deletion within the right end of T-DNA insert 621-36 is marked by an arrow. In those cases where the actual T-DNA ends are shorter than those canonical ones that occur within the 25 bp T-DNA border repeats, this is indicated by the number of missing base pairs (left and right in line I).

sequence comparison revealed 20.5 to 42.8% homology between target sites and T-DNA ends.

Imprecise junctions

Imprecise junctions were observed in plants N6H-14, c27 and 621-2 (Figure 2) in that the positions of left or right T-DNA termini did not fit the break-points of target site deletions. These imprecise junctions were characterized by the presence of additional DNA sequences filling the gap between the break-points of target deletions and T-DNA insertions. These additional DNA sequences will further be referred to as 'filler' DNAs.

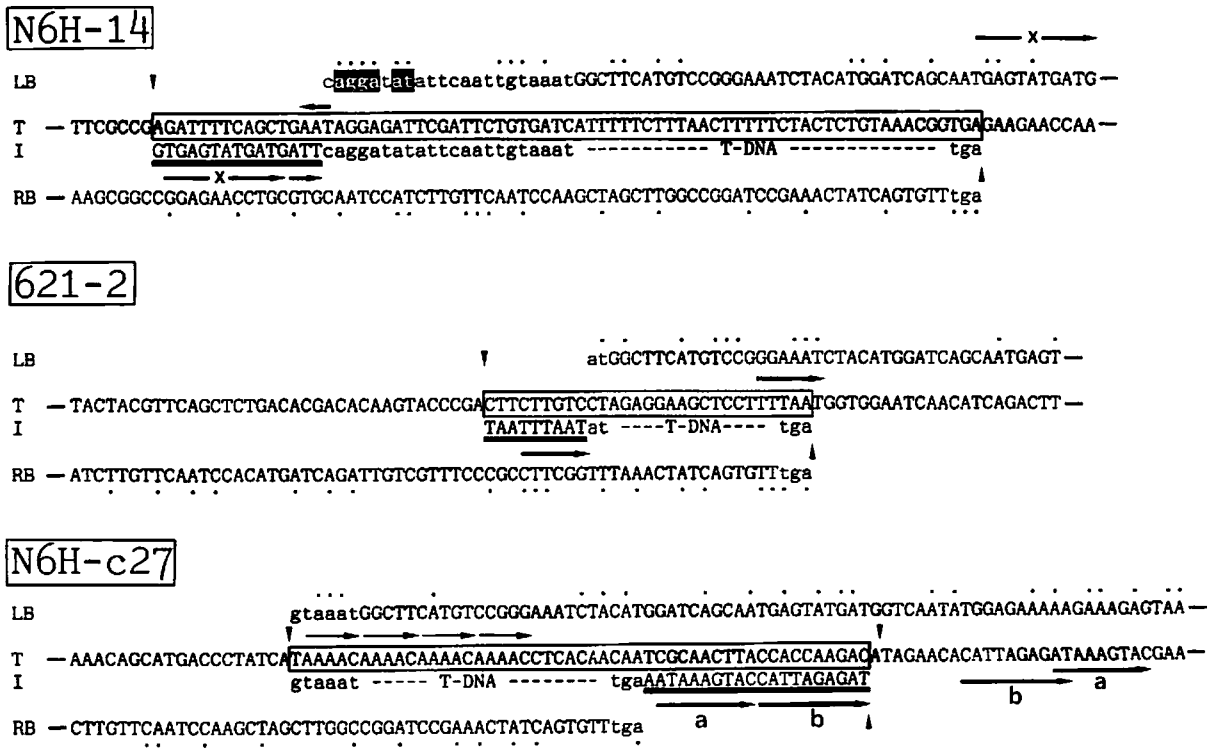

Fig. 2. Imprecise junctions of T-DNA inserts N6H-14, 621-2 and N6H-c27. DNA sequences in lines RB, LB, T and I are depicted as outlined in Figure 1. Sequences thought to be copied from either plant or T-DNA and filling up gaps between break-points of T-DNA inserts and plant target DNA deletions are underlined. 'X' marks a putative duplication of an internal T-DNA segment at the left end of T-DNA insert N6H-14 (lines LB and I). Short arrows in N6H-14 lines I and T indicate the position of a putative template switch at an AAT motif in the target DNA. In 621-2 lines T and I arrows mark a repeated TTTAAT sequence that occurs both in the filler DNA (underlined) and at the right break-point of 621-2 plant target DNA. At the right junction of N6H-c27 target site (line T) arrows 'b' and 'a' mark overlapping plant DNA segments, presumably duplicated in reverse order (arrows 'a' and 'b' in line I) as a filler DNA for the right end T-DNA–plant DNA junction. Above the target N6H-c27 (line T) horizontal arrows label a repeated AAAAC sequence motif. Erroneous annealing of this 5-meric repeated sequence during DNA replication or repair could cause stepwise repetition of 10-meric 'a' and 'b' sequences. [If this was the case, the first nucleotide (A in line I) of N6H-c27 filler DNA (underlined) has probably originated by staggered cuts at the right break-point of the target (marked by two vertical black arrows in line T) and by double-strand break repair.]

In plant N6H-14 a filler DNA of 15 bp was identified between the left break-point of T-DNA insertion and target deletion. The left T-DNA terminus contained an AGGA.AT sequence that also occurred within the target DNA. A TGAGTATGATG motif (Figure 2, lines I and LB, arrow 'x') in the filler DNA, showing no homology to the target, was also detected as an internal T-DNA sequence located 35 bp upstream of the left T-DNA border. In plant 621-2 a TTTAAT sequence (Figure 2, arrows in lines I and T), present at the right break-point of the target deletion, was identical to a sequence detected in a presumed filler DNA sequence (Figure 2, line I, underlined) linked to the left T-DNA end. The filler sequence linking the right end of T-DNA insert N6H-c27 to the right break-point of corresponding target deletion contained two overlapping sequence motifs (Figure 2, line I, arrows a and b), that were also detected in a plant DNA segment starting 8 bp downstream of the right break-point of target deletion (Figure 2, line T, arrows a and b).

Discussion

Genomic boundaries of *Arabidopsis* T-DNA insertions described above represent two classes of recombinant joints: (i) precise junctions between sequences of the T-DNA ends and plant target DNA sequences and (ii) imprecise junctions where the two types of DNA sequences are joined through filler DNAs that were found to repeat sequences located at various distances away from the expected break-points in T-DNA and plant target DNA. Similar imprecise junctions involving at most short homologies have also been observed during extrachromosomal recombination of T-DNAs carrying virus genomes in turnip cells (Bakkeren *et al.*, 1989). Analogous DNA junctions have been described for duplications, inversions and deletions in bacteria (Albertini *et al.*, 1982; Whoriskey *et al.*, 1987), for adenovirus and retrovirus integration (Deuring *et al.*, 1981; Shih *et al.*, 1988; Brown *et al.*, 1989), for intra- and interchromosomal recombination in animal cells (Wilson *et al.*, 1982; Subramani and Berg, 1983; Anderson *et al.*, 1984; Bollag *et al.*, 1989) and for transposon integration and excision in plants (Saedler and Nevers, 1985; Schwarz-Sommer *et al.*, 1987; Coen *et al.*, 1989). In all cases the formation of these types of DNA junctions was shown to involve recombination between DNA sequences that contain only a few identical nucleotides. Such recombination events were therefore referred to as 'illegitimate' (Low, 1988).

It has been suggested that illegitimate recombination results from a cascade of DNA-gyrase catalysed events. Indeed, DNA sequences of illegitimate junctions resemble those of DNA-gyrase cleavage sites (Marvo *et al.*, 1983). Diverse forms of illegitimate recombination, however, cannot be

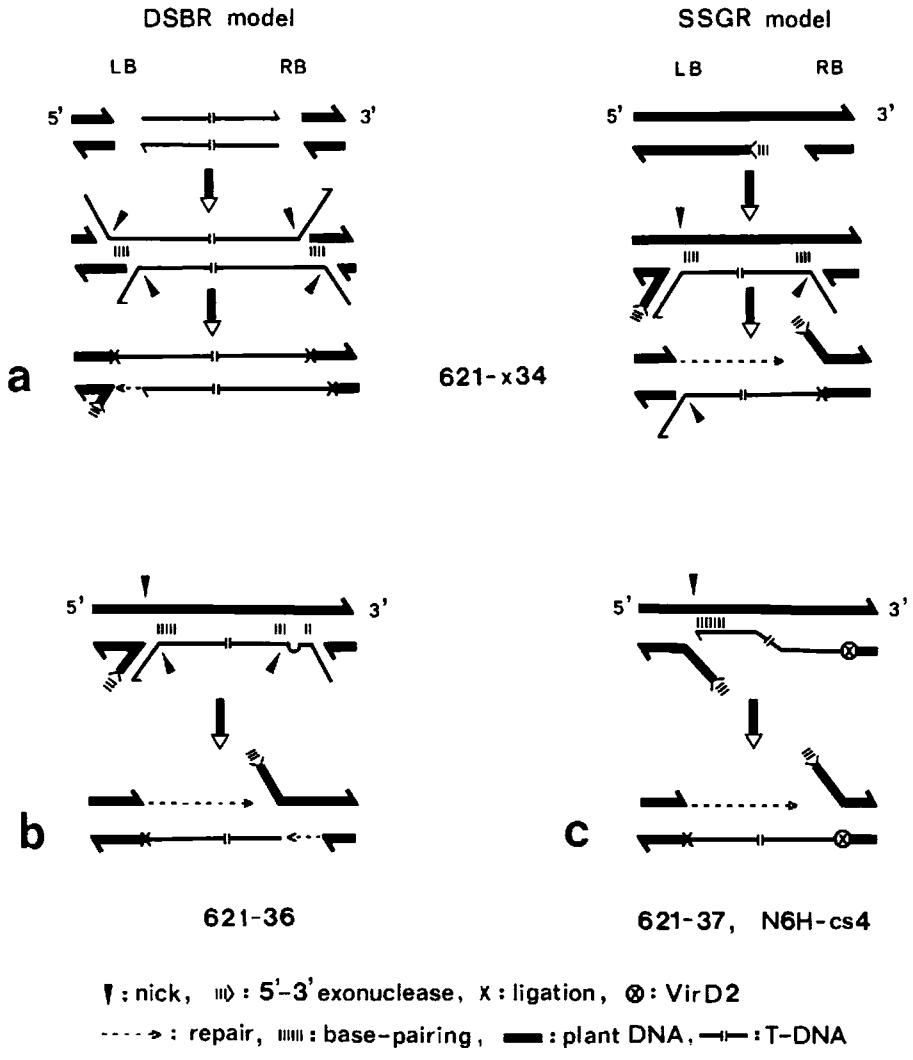

Fig. 3. Illegitimate recombination models for T-DNA integration. (a) Double-strand break-repair (DSBR) and single-strand gap-repair (SSGR) models equally explain the integration events leading to T-DNA insertion 621-x34. As outlined in the Discussion, the DSBR model predicts a double-strand break in the target. Unwound or exonuclease processed ends of the T-DNA and of the target anneal by partial homologous pairing. Single-strand overhangs are removed by endo- or exonuclease digestion; the ends are repaired and ligated. The SSGR model predicts a nick in the target that is converted to a gap by a 5'-3' exonuclease. The T-strand invades the gap and a synapsis is formed by partial pairing between T-DNA ends and target. The overhangs of the T-DNA are removed and the T-strand is ligated to the target. A nick introduced into the second strand of the target initiates repair synthesis of the second strand of the T-DNA. (b) Application of the SSGR model for T-DNA insertion 621-36. As proposed in the model of insertion 621-x34, repair at the ends of invading T-strand may occur before ligation. Absence of an internal CAGT T-DNA sequence from the right junction of insert 621-36 suggests that copying of the target plant DNA led to alteration of the right T-DNA end during integration. Mismatch repair may similarly explain the observed deletion. (c) A modified version of SSGR model explaining possible involvement of VirD2 protein in recombination events that resulted in T-DNA insertions 621-37 and N6H-cs4. T-DNA invasion depicted in the model involves VirD2-mediated recognition of a nick (or gap) that is followed by ligation of the 5' end of the T-strand to the target DNA. The 3' end of the T-strand moves along the target, while the unwound target DNA strand is removed by exonucleolytic digestion. At the position where the 3' end of the T-strand is stabilized by base-pairing, the T-strand is ligated to the target. At the same position a nick is introduced in the second strand of the target that defines the left break-point of target deletion and initiates the replication of the T-strand. Symbols used for presentation of putative recombination events are explained underneath the models.

described by a common model when specific features of each type of recombination are considered. Thus, integration of T-DNA does not seem to be sequence specific (as is the case for retroviruses) and it causes deletions of target sequences in contrast to what happens upon integration of most plant transposons. Notwithstandingly, models proposed to describe illegitimate recombination leading to transposon integration (Schwarz-Sommer *et al.*, 1987) or excision (Saedler and Nevers, 1985) in plants, or intra- and interchromosomal recombination in animal cells (Lin *et al.*, 1984; Wake *et al.*, 1985) were useful for the elaboration of a single-strand gap-repair (SSGR) model for T-DNA integration. Since most recombination models postulate the invasion of double-strand D-loops by single-strand DNA ends (Holliday, 1964; Meselson and Radding, 1975; Rauth *et al.*, 1986), SSGR models are not significantly different from double-strand break-repair (DSBR) models, such as the one proposed by Szostak *et al.* (1983), and T-DNA integration can be explained equally well by both types of models.

A model for T-DNA integration

The results described above indicate that the ends of T-DNA as well as plant target DNA play an important role in T-DNA integration. The main feature being that T-DNA

insertion results in the deletion of target DNA. General recombination models suggest that break-points of deletions result from nicks (or breaks) in DNA that can be converted to recombinogenic single-strand overhangs or gaps by exonuclease action. In addition to enzymes involved in recombination (Cox and Lehman, 1987), errors in DNA replication and repair may also release free DNA ends that can serve as substrates for illegitimate recombinant junctions.

It is probable that limited homologous pairing between plant target sites and T-DNA play a role in T-DNA integration (similar to what happens in retrotransposition or in DNA transformation of animal cells). Schemes depicted in Figure 3a show DSBR and SSGR models that illustrate putative recombination events leading to T-DNA insertion such as observed for 621-x34. The DSBR model predicts a double-strand break in the target that provides recombinogenic single-strand overhangs. According the Szostak *et al.* (1983) such breaks may result from either staggered cuts or from consecutive events involving conversion of a nick to a gap in one DNA strand followed by a nick in the second DNA strand opposite the gap. Application of the DSBR model to DNA sequence data obtained for T-DNA insert 621-x34 suggests that following a double-strand break in plant DNA target site, the unwound (or exonuclease processed) ends of target plant DNA annealed with the ends of double-stranded T-DNA at common TC.GGG and ATTGT sequence motifs. Single-strand overhangs were then removed e.g. by 3′ to 5′ and 5′ to 3′ exonuclease (Lin *et al.*, 1984) or single-strand specific endonuclease activities of the repair system. As depicted in Figure 3a, recombination by double-strand gap repair thus leads to DNA integration and deletion. The size of target deletion and the extent of deletions at the T-DNA ends are precisely defined by the location of short complementary DNA sequences that stabilize the interacting ends of plant DNA and T-DNA during annealing.

The SSGR model suggests that (i) initially a nick was formed in the target that was extended either by partial unwinding or by 5′ to 3′ exonuclease digestion to a gap, (ii) the ends of invading single-stranded T-DNA (T-strand) were located sterically close to each other and formed a heteroduplex with the gap by annealing to complementary TC.GGG and ATTGT sequences, (iii) the unpaired 5′ and 3′ overhangs of the T-strand were removed (as described above) and the T-DNA ends were ligated to the target, and (iv) the target site−T-strand heteroduplex was probably resolved by a nick in the second target DNA strand that provided a free 3′ DNA end as primer for repair synthesis of the second strand of the T-DNA.

Both models are similar to those described by Lin *et al.* (1984) and Rauth *et al.* (1986) for integration of double- and single-stranded DNAs in animal cells and compatible with classical recombination models and enzyme activities predicted therein (Holliday, 1964; Szostak *et al.*, 1983). The DSBR and SSGR models do not define the order of ligation events, however they predict that repair at the annealed DNA ends may lead to gene conversion. Figure 3b outlines how gene conversion or mismatch repair may explain the situation observed for insert 621-36 at the junctions between T-DNA and plant DNA. In animal cells almost exclusively the ends of integrating foreign DNAs are used as templates for DNA repair synthesis (Bollag *et al.*, 1989). Similarly, our data also indicate that T-DNA sequences remained unaltered in most insert junctions. The right junction of insert 621-36 is an exception, since it contains an internal CAGT deletion within the right T-DNA end indicating that in this case the target plant DNA was apparently used as template for repair.

The fact that in five out of seven inserts, sequences derived from the right 25 bp T-DNA border repeat formed a junction with plant DNA suggests that T-DNA integration must involve additional functions. Such functions could be provided by Ti plasmid encoded virulence proteins. The VirD2 protein was found to be covalently attached to T-strands in induced agrobacteria (Herrera-Estrella *et al.*, 1988) as well as to double-stranded linear T-DNAs (Dürrenberger *et al.*, 1989). Homology of VirD2 protein domains to nuclear targeting signals (G.Bakkeren and B.Hohn, unpublished) supports the idea that VirD2 may function as a 'pilot' protein leading the T-DNA into the plant cell nucleus (Zambryski, 1988; A.Herrera-Estrella, personal communication). As a DNA-bound nicking−closing enzyme, VirD2 may also recognize (or even cause) nicks in the plant DNA and mediate joining of the T-DNA and the target. VirE2 protein (Gietl *et al.*, 1987; Citovsky *et al.*, 1989) may also play an auxillary function. This single-stranded DNA-binding (SSDB) protein could protect the T-strand and accelerate DNA replication and recombinational repair by unwinding target DNA strands (Chase and Williams, 1986). VirD2 protein might mediate the formation of precise junctions between the right (5′) T-DNA end and target plant DNA. The left (3′) T-DNA end may be stabilized by partial base-pairing with target plant DNA. This could result in the formation of a nick thus defining the left break-point of target DNA deletion. Repair synthesis using the T-strand as template would complete the integration process. This series of events is illustrated in Figure 3c by a SSGR model that better explains the data described for T-DNA inserts 621-37 and N6H-cs4 than an equivalent DSBR model.

A very different integration pattern was observed by Gheysen *et al.* (1987) in a tobacco tumour line, in which a 27 bp deletion, a 158 bp direct repeat and a 28 bp inverted repeat at the right and left T-DNA junctions appear to have resulted from the integration events. The structure of this tobacco insertion is similar to those observed in *Arabidopsis* plants N6H-14, c27 and 621-2, in which T-DNA insertions generated direct and inverted repeats of recombinational junctions. As outlined in recombination models proposed by Saedler and Nevers (1985), formation of direct and inverted repeats at illegitimate recombinant junctions can be explained by assuming that (i) nicking at the physical break-points of targets can release free 3′ DNA ends that can be used as primers for limited copying of insert T-DNA and target plant DNA sequences located in the vicinity of break-points, (ii) template switch or slippage at the DNA polymerase elongated 3′ ends of T-DNA or target plant DNA may occur (even before interaction between plant DNA target sites with T-DNA inserts) and (iii) these newly synthesized 3′ DNA sequences can serve as templates for a second round of DNA repair synthesis primed by 3′ ends of either T-DNA insert or target plant DNA. It can also be deduced that in such double-strand break structure the priming of the T-DNA replication by both 3′ ends of plant target DNA or a template switch during repair at the T-DNA ends may result in integration of direct or inverted tandem repeats of T-DNA inserts respectively.

The data and models discussed above indicate that DNA

replication and repair are required for T-DNA integration. Preferential integration of the T-DNA into chromosomal loci that are potentially transcribed may therefore reflect a coordinated regulation between DNA replication and transcription in plants, or simply the topological requirements (D-loop formation, nicks, gaps or breaks) for T-DNA integration. If this is the case, it can be predicted that, apart from virulence protein aided processes, essentially the same mechanism will be involved in the integration of any DNA that is by some way introduced in plant cell nuclei.

Materials and methods

T-DNA mutagenesis, insert rescue and isolation of insertional target sites

Construction and characterization of *Agrobacterium* gene fusion vectors pPCV621 and pPCV6NFHyg, *Agrobacterium*-mediated transformation of *Arabidopsis thaliana* (ecotype Columbia) plants, identification and analysis of the expression of aminoglycoside phosphotransferase [*aph*(3')II] gene fusions have been described previously (Koncz *et al.*, 1989). Purification of plant nuclear DNA, mapping and determination of the copy number of T-DNA inserts, construction of *Arabidopsis* genomic DNA library in lambda EMBL 4 vector, DNA hybridization and other DNA manipulations were as described (Koncz and Schell, 1986; Koncz *et al.*, 1989, 1990; Sambrook *et al.*, 1989).

To isolate T-DNA inserts as plasmids, 5–20 μg plant DNA was digested with *Hin*dIII, self-ligated and transformed into *Escherichia coli* strain DH1 (Koncz *et al.*, 1989). Plant DNA fragments flanking the rescued T-DNAs were isolated after *Hin*dIII–*Bcl*I, *Hin*dIII–*Bam*HI and *Hin*dIII–*Cla*I digestions, cut further with various enzymes and subcloned into pUC18 and 19 vectors to determine their nucleotide sequence using T7 DNA polymerase and double-stranded DNA templates (Koncz *et al.*, 1990; Yanisch-Perron *et al.*, 1985; Tabor and Richardson, 1987). Target sites of T-DNA inserts 621-2, 36, 37 and N6H-14 were isolated by screening of *Arabidopsis* genomic DNA library with plant DNA fragments of rescued T-DNA clones as probes. Oligonucleotide primers derived from plant DNA sequences flanking T-DNA inserts 621-x34, N6H-cs4 and c27 were used for amplification of target DNA fragments from the genomic DNA library by *Taq* polymerase chain reaction (PCR) (Innis *et al.*, 1989). An aliquot of genomic library (1×10^6 phages) was heated (70°C, 5 min), then combined with PCR-buffer (25 mM Tris–HCl (pH 9.0), 7.5 mM $(NH_4)_2SO_4$, 3.5 mM $MgCl_2$, 25 mM KCl and 85 μg/ml bovine serum albumin), 0.25 mM dXTPs and 1 μg of each primer in a volume of 100 μl. After denaturation of the DNA (100°C, 10 min), 5 units of *Taq* DNA polymerase was added, the reaction mix was covered with paraffin oil and cycles of denaturation (90°C, 1 min), annealing (55°C, 1 min) and extension (70°C, 3 min) were repeated 30 times. The amplified DNA fragments were phosphorylated by T4 polynucleotide kinase (Sambrook *et al.*, 1989), and cloned into the *Sma*I site of pUC18 to determine their nucleotide sequence.

Nucleotide sequence analysis

Nucleotide sequences of wild type and T-DNA-tagged *Arabidopsis* genomic loci were compared using a WISGEN program package (Deveraux *et al.*, 1984) adapted to VAX/VMS computer version 5.1-1. In recombination models explaining T-DNA integration events conventions introduced by Holliday (1964), Meselson and Radding (1975) and Szostak *et al.* (1983) were followed. The stability of partial DNA duplexes was estimated using the approximation of Tinoco *et al.* (1973). Complete nucleotide sequence of plant DNA fragments flanking T-DNA insertions will appear in the EMBL, Genbank and DDJB Nucleotide Sequence Databases under accession numbers: *A.thaliana* 621-2 DNA, X53920; 621-36 DNA, X53921; 621-37 DNA, X53922; N6H-c27 DNA, X53923; N6H-cs4 DNA, X53924 and 621-x34 DNA, X53925. Nucleotide sequence data of N6H-14 DNA and of the *A.thaliana cs/ch*-42 gene are available under the accession number X51799.

Acknowledgements

We thank Ms A. Radermacher and Ms A.Lossow for skilful technical assistance, Dr S.Schwarz-Sommer for kind help in the construction of *Arabidopsis* genomic library, Drs J.Dangl, B.Kemper, L.Orosz, H.Saedler, P.Starlinger and Z.Koukolikova-Nicola for critical review and Ms G.Kobert for typing of the manuscript. A part of this work was supported by the Deutsche Forschungsgemeinschaft and the Hungarian Academy of Sciences as a joint project between the MPI (Köln) and BRC (Szeged).

References

Albertini,A.M., Hofer,M., Calos,M.P. and Miller,J.H. (1982) *Cell*, **29**, 319–328.

Albright,L.M. Yanofski,M.F., Leroux,B., Ma,D. and Nester,E.W. (1987) *J. Bacteriol.*, **169**, 1046–1055.

Anderson,R.A., Kato,S. and Camerini-Otero,D. (1984) *Proc. Natl. Acad. Sci. USA*, **81**, 206–210.

Bakkeren,G., Koukolikova-Nicola,Z., Grimsley,N. and Hohn,B. (1989) *Cell*, **57**, 847–857.

Binns,A.N. and Thomashow,M.F. (1988) *Annu. Rev. Microbiol.*, **42**, 575–606.

Bollag,R.J., Waldman,A.S. and Liskay,R.M. (1989) *Annu. Rev. Genet.*, **23**, 199–225.

Brown,P.O., Bowerman,B., Varmus,H.E. and Bishop,M.J. (1989) *Proc. Natl. Acad. Sci. USA*, **86**, 2525–2529.

Buchanan-Wollaston,V., Passiatore,J.E. and Cannon,F. (1987) *Nature*, **328**, 172–174.

Chase,J.W. and Williams,K.R. (1986) *Annu. Rev. Biochem.*, **55**, 103–136.

Christie,P.J., Ward,S.J., Winans,S.C. and Nester,E.W. (1988) *J. Bacteriol.*, **170**, 2659–2667.

Chyi,Y.S., Jorgensen,R.A., Goldstein,D., Tanksley,S.D. and Loaiza-Figueroa,F. (1986) *Mol. Gen. Genet.*, **204**, 64–69.

Citovsky,V., Wong,M.L. and Zambryski,P. (1989) *Proc. Natl. Acad. Sci. USA*, **86**, 1193–1197.

Coen,E.S., Robbins,T.P., Almeida,J., Hudson,A. and Capenter,R. (1989) In Berg,E.E. and Howe,M.M. (eds) *Mobile DNA*, ASM, Washington, DC, pp. 413–436.

Cox,M.M. and Lehman,I.R. (1987) *Annu. Rev. Biochem.*, **56**, 229–262.

Deuring,R., Winterhoff,V., Tamanoi,F., Stabel,S. and Doerfler,W. (1981) *Nature*, **293**, 81–84.

Deveraux,J., Haeberli,P. and Smithies,O. (1984) *Nucleic Acids Res.*, **12**, 387–395.

Dürrenberger,F., Crameri,A., Hohn,B. and Koukolikova-Nicola,Z. (1989) *Proc. Natl. Acad. Sci. USA*, **86**, 9154–9158.

Gheysen,G., Van Montagu,M. and Zambryski,P. (1987) *Proc. Natl. Acad. Sci. USA*, **84**, 6169–6173.

Gietl,C., Koukolikova-Nicola,Z. and Hohn,B. (1987) *Proc. Natl. Acad. Sci. USA*, **84**, 9006–9010.

Grimsley,N., Hohn,T., Davies,J.N. and Hohn,B. (1987) *Nature*, **325**, 177–179.

Herrera-Estrella,A., Chen,Z., Van Montagu,M. and Zambryski,P. (1988) *EMBO J.*, **7**, 4055–4062.

Holliday,R. (1964) *Genet. Res.*, **5**, 282–304.

Howard,E.A., Winsor,B.A., DeVos,G. and Zambryski,P. (1989) *Proc. Natl. Acad. Sci. USA*, **86**, 4017–4021.

Innis,M.A., Gelfand,D.H., Sninsky,J.J. and White,T.J. (1989) *PCR Protocols.*, Academic Press, New York.

Jayaswal,R.K., Veluthambi,K., Gelvin,S.B. and Slightom,J.L. (1987) *J. Bacteriol.*, **169**, 5035–5045.

Koncz,C. and Schell,J. (1986) *Mol. Gen. Genet.*, **204**, 383–396.

Koncz,C., Martini,N., Mayerhofer,R., Koncz-Kalman,Zs., Körber,H., Redei,G.P. and Schell,J. (1989) *Proc. Natl. Acad. Sci. USA*, **86**, 8467–8471.

Koncz,C., Mayerhofer,R., Koncz-Kalman,Zs., Nawrath,C., Reiss,B., Redei,G.P. and Schell,J. (1990) *EMBO J.*, **9**, 1337–1346.

Koukolikova-Nicola,Z., Shillito,R.D., Hohn,B., Wang,K., Van Montagu,M. and Zambryski,P. (1985) *Nature*, **313**, 191–196.

Lin,F.L., Sperle,K. and Sternberg,N. (1984) *Mol. Cell. Biol.*, **4**, 1020–1034.

Low,K. (1988) *The Recombination of Genetic Material.* Academic Press, New York.

Marvo,S.L., King,S.R. and Jaskunas,R.S. (1983) *Proc. Natl. Acad. Sci. USA*, **72**, 358–361.

Meselson,M.S. and Radding,C.M. (1975) *Proc. Natl. Acad. Sci. USA*, **72**, 358–361.

Rauth,S., Song,K.Y., Agares,D., Wallace,L., Moore,P.D. and Kucherlapati,R. (1986) *Proc. Natl. Acad. Sci. USA*, **83**, 5587–5591.

Saedler,H. and Nevers,P. (1985) *EMBO J.*, **4**, 585–590.

Sambrook,J., Fritsch,E.F. and Maniatis,T. (1989) *Molecular Cloning: A Laboratory Manual.* Cold Spring Harbor Laboratory Press, Cold Spring Harbor, NY.

Schwarz-Sommer,S., Leclercq,L., Göbel,E. and Saedler,H. (1987) *EMBO J.*, **6**, 3873–3880.
Shih,C.C., Stoye,J.P. and Coffin,J.M. (1988) *Cell*, **53**, 531–537.
Stachel,S. and Zambryski,P. (1986) *Cell*, **46**, 325–333.
Stachel,S., Messens,E., Van Montagu,M. and Zambryski,P. (1985) *Nature*, **318**, 624–628.
Stachel,S., Timmerman,B. and Zambryski,P. (1986) *Nature*, **322**, 706–712.
Subramani,S. and Berg,P. (1983) *Mol. Cell Biol.*, **3**, 1040–1052.
Szostak,J.W., Orr-Weaver,T.L., Rothstein,R.J. and Stahl,F.W. (1983) *Cell*, **33**, 25–35.
Tabor,S. and Richardson,C.C. (1987) *Proc. Natl. Acad. Sci. USA*, **84**, 4767–4771.
Tinoco,I., Borer,P.N., Dengler,B., Levine,M.D., Uhlenbeck,O.C., Crothers,O. and Gralla,J. (1973) *Nature*, **246**, 40–41.
Wake,C.T., Vernaleone,F. and Wilson,J.H. (1985) *Mol. Cell Biol.*, **5**, 2080–2089.
Wallroth,M., Gerats,A.G.M. Roger,S.G., Fraley,R.T. and Horsch,R.B. (1986) *Mol. Gen. Genet.*, **202**, 6–15.
Wang,K., Herrera-Estrella,L., Van Montagu,M. and Zambryski,P. (1984) *Cell*, **38**, 455–462.
Wang,K., Stachel,S.E., Timmerman,B., Van Montagu,M. and Zambryski,P. (1987) *Science*, **235**, 587–591.
Ward,J.E., Akiyoshi,D.E., Regier,D., Datta,A., Gordon,M.P. and Nester,E.W. (1988) *J. Biol. Chem.*, **263**, 5804–5814.
Whoriskey,S.K., Nghiem,V.H., Leong,P.M., Masson,J.H. and Miller,J.H. (1987) *Genes Dev.*, **1**, 227–237.
Wilson,J.M., Berget,P.B. and Pipas,J.M. (1982) *Mol. Cell Biol.*, **2**, 1258–1269.
Winans,S.C., Ebert,P.R., Stachel,S.E., Gordon,M.P. and Nester,E.W. (1986) *Proc. Natl. Acad. Sci. USA*, **83**, 8278–8282.
Yanisch-Perron,C., Vieira,J. and Messing,J. (1985) *Gene*, **33**, 103–119.
Yanofski,M.F., Porter,S.G., Young,C., Albright,C., Gordon,M.P. and Nester,E.W. (1986) *Cell*, **47**, 471–477.
Young,C. and Nester,E.W. (1988) *J. Bacteriol.*, **170**, 3367–3374.
Zambryski,P. (1988) *Annu. Rev. Genet.*, **22**, 1–30.
Zambryski,P., Tempe,J. and Schell,J. (1989) *Cell*, **56**, 193–201.

Received on July 3, 1990; revised on December 20, 1990