SUPPLEMENTAL DATA

Table A. Characteristics of 67 plant DNA/T-DNA junctions

Junction[1]	Border structure[2]
Co-transformant population	
kg32LB-1	LB0/plant(AC)
kg44LB-2	LB-8/48/plant
kg44LB-3	LB-5/7/plant
kg104LB-4	LB-7/plant(TAT)
kg104LB-5	LB0/8/plant
kg150LB-6	LB-47/23/plant
kg150LB-7	LB0/3/plant
kg150LB-8	LB-6/8/plant
kg158LB-9	LB-17/plant
kg158LB-10	LB-19/plant(CCATT)
kg162LB-11	LB-6/plant(TATA)
kg162LB-12	LB-28/8/plant
kg165LB-13	LB-7/plant(T)
kg165LB-14	LB0/plant(AC)
kg269LB-15	LB-34/plant
kg269LB-16	LB0/33/plant
kg314LB-17	LB-10/plant(CT)
kg348LB-18	LB0/5/plant
kg353LB-19	LB-5/17/plant
kg353LB-20	LB-7/26/plant
kg353LB-21	LB-20/13/plant
kd22LB-22	LB-10/plant(TGA)
kd22LB-23	LB-6/plant(TAT)
kd75LB-24	LB-22/1/plant
kd75LB-25	LB-57/plant
kd75LB-26	LB-18/plant(TTT)
kd315LB-27	LB-57/plant(CA)
kd315LB-28	LB-15/4/plant
kd12LB-29	LB-18/plant(TTT)
kd27LB-30	LB-27/plant(ACATG)
kg7LB-32	LB-38/plant(AGATT)
kg32RB-1	RB0/7/plant
kg44RB-2	RB-1/2/plant
kg135RB-3	RB-1/11/plant
kg135RB-4	RB0/45/plant
kg162RB-5	RB0/plant
kg314RB-6	RB-16/plant(GGTG)
kg314RB-7	RB-1/plant(TG)
kg353RB-8	RB-4/51/plant
kg353RB-9	RB-2/plant(ACT)
kd22RB-10	RB-1/plant(TGACTG)
kd22RB-11	RB0/plant
kd27RB-12	RB-43/plant(CTAATTT)

Single-copy population	
CK2L6LB-1	LB-24/plant(ACTTC)
CK2L7LB-2	LB-18/plant(GGTAAA)
CK2L36LB-3	LB-60/5/plant
CK2L70LB-4	LB-17/3/plant
CK2L72LB-5	LB-48/4/plant
CK2L94LB-6	LB-136/plant(TGC)
CK2L102LB-7	LB-65/plant(TA)
CK2L107LB-8	LB-15/plant(TG)
CK2L111LB-9	LB-43/plant
CK2L148LB-10	LB-39/plant(TA)
CK2L6RB-1	RB-12/plant(AT)
CK2L7RB-2	RB-2/15/plant
CK2L36RB-3	RB-1/plant
CK2L70RB-4	RB0/29/plant
CK2L72RB-5	RB-57/1/plant
CK2L94RB-6	RB-12/plant(AT)
CK2L102RB-7	RB-10/28/plant
CK2L107RB-8	RB0/plant(GG)
CK2L148RB-10	RB0/5/plant
ExtraCK2L-LB	LB-251/plant(CT)
CK2L129LB-10	LB-14/plant(CAA)
CK2L133LB-11	LB-24/plant(C)
CK2L129RB-10	RB-2/plant(CTGACT)
CK2L133RB-11	RB-30/plant(GGG)

[1] Each amplified junction is given a code that is composed of the name of the transgenic line from which the junction is derived, followed by a distinction between left (LB) and right border (RB) junctions and a serial number.

[2] The border structure for each junction is represented as follows: LB (left border) or RB (right border) is followed by the number of bases that have been deleted (indicated with "-" sign), in case the border was cleaved correctly LB or RB is followed by "0". A "+" sign indicates the process of read through. In case microhomology was detected between plant DNA and T-DNA for those junctions without filler DNA, the sequence (such as CTGACT) is represented between brackets. The numbers between the slashes indicate the length of the filler DNA sequences detected.

Table B. Characteristics of 13 junctions between linked T-DNAs

Border[a]	Deleted bases[b]	Filler (bp)	Border	Deleted bases	Microhomology[c]
Tandem junctions					
LB	-46	0	RB	0	3 bp (TGG)
LB	-26	0	RB	-17	1 bp (A)
LB	0	1	RB	0	F
LB	0	21	RB	-36	F
LB	-275	0	RB	0	1 bp (G)
LB	-61	0	RB	-629	NM
LB[d]	-423 (-2007)	38	RB	-12	F
LB[d]	-2309 (-4009)	0	RB	0	NM

Inverted repeat junctions					
LB	-32	0	LB[d]	-4190 (-4969)	3 bp (CGG)
LB	0	0	LB[d]	+710 (+902)	5 bp (TCCTG)
LB	0	21	LB	-2336	F
RB	+1	40	RB[d]	-2679 (-2908)	F
RB	0	4	RB	-3651	F

[a] LB, left border; RB, right border.

[b] 0 means the border has been cleaved correctly; + indicates the process of readthrough.

[c] For the junctions without filler DNA sequences, a screen for microhomology at the transition point between the two T-DNA ends was done. NM, both T-DNA ends do not share homologous bases; F, the junctions harbor filler sequences.

[d] Because the transforming T-DNA plasmids, pAK1202 and pAD1201, used for co-transformation, harbor similar genetic elements that are positioned differently with respect to the T-DNA borders, the T-DNA breakpoint could not be assigned unambiguously in all cases, the alternative possibility is given between parentheses.

Statistical significance test for duplicated identical sequence motifs when a 200-bp plant segment surrounding the T-DNA integration point is used for an identity search

For the plant DNA/T-DNA junctions for which the filler origin was determined and for which sequence motifs identical with a 200-bp plant DNA segment surrounding the T-DNA integration point were observed, we tested whether the reported motifs could have been detected simply by chance. For 14 junctions with filler DNA, we found sequence motifs of 6 bp, 9 bp, 10 bp, 11 bp, 12 bp, 15 bp, 16 bp, 20 bp, and 30 bp that were identical to the plant target (see Table C). To evaluate statistically these identities, we considered the different parameters that influence the statistical relevance of a reported identity. First, the length of the filler is important: the probability of finding a 6-bp sequence motif is higher when a 51-bp filler is analyzed than an 8-bp filler, because out of a 51-bp filler DNA, 92 different 6-bp sequence blocks can be constructed and only six different 6-bp sequence blocks out of an 8-bp filler sequence. Second, the distance between the filler DNA and the position of the template DNA from which the identical sequence motif originated can influence the statistical relevance, namely the probability of finding an identical sequence motif in a 15-bp segment surrounding the T-DNA integration point is lower than that of finding the same sequence motif in a 500-bp segment. Third, the number of observed sequence identities is important. The probability of finding three 6-bp motifs is lower than finding only one. This means that the statistical relevance of each identical sequence motif that is reported is different. Also, because our data show that filler DNA consists of several short duplicated sequence motifs that might arise from chaotic repair synthesis, it is extremely difficult to perform a numeral, statistical analysis of the data.

Table C. Summary of the observed duplicated sequence motifs between filler DNA and a 200-bp plant segment surrounding the T-DNA integration point for 14 plant DNA/T-DNA junctions

T-DNA junction	Length of filler (bp)	Motif (bp)	Distance (bp)
Left			
kg353LB-21	13	9	3
		9	37
		6	TSD[a]
kg353LB-20	26	15	6
		16	0
kg150LB-6	23	10	TSD
kg269LB-16	33	12	10
kg44LB-2	48	6	45
		6	42
		6	35
		6	96
kg44LB-3	7	6	13
kg104LB-5	8	6	27
Right			
kg135RB-4	45	6	66
		6	68
		6	19
		6	39
		6	85
		6	9
CK2L7RB-2	15	12	57
		9	0
CK2L10RB-7	28	6	89
kg32RB-1	7	11	29
kg135RB-3	11	6	84
kg353RB-8	51	20	6
		30	30
CK2L70RB-4	29	16	60
		11	5

[a] TSD, target site deletion; deletion upon integration of the T-DNA, indicates that the observed sequence motif is identical to a sequence block occurring in the TSD.

In order to make our analysis straightforward and realizable, we first determined for which sequence motifs there might be a problem of statistical relevance. According to our calculations, motifs larger than 9 bp were relevant. When we consider the three 9-bp motifs in our analysis, Table C shows that the distance between the sequences from which these motifs originate and the filler DNA is 3 bp, 37 bp, and 0 bp. So, the probability of finding a 9-bp sequence motif by chance in a random 37-bp DNA region (T-DNA integration is assumed random) can be calculated as follows. We know that 4^9 (= 262,144) different 9-bp sequence blocks can be constructed using the four bases A, C, G, and T. If we consider that a 37-bp region is screened for finding a 9-bp sequence motif, we can construct out of this 37-bp region 58 different 9-bp sequence blocks using the formula [[length of the region that is screened - (length of motif – 1)] x 2] because motifs in direct and inverted orientation are taken into account. So, the probability of finding a 9-bp sequence motif in a random 37-bp region is 58/262,144 = 0.000221 (or $p<0.01$). For junction kg353LB-21 we even found two 9-bp sequence blocks in this 37-bp region with a probability of only 5.0^{e-8}. A similar calculation can be done for motifs of 10 bp, 11 bp, 12 bp, 15 bp, 16 bp, 20 bp, and 30 bp. Of course, when the motif size increases, the probability of finding a by chance similarity decreases. Therefore, we conclude that the reported plant DNA-derived sequence identities of 9 bp or longer are statistically significant, certainly when these large sequence duplications are located in the vicinity (see Table C) of the T-DNA integration point.

On the other hand, 6-bp sequence identities might be statistically questionable, even when only a small region of 200 bp of plant DNA is screened. From Table C, we learn that for 6-bp sequence identities the average distance between the filler and the originating sequence block is 47.8 bp. A statistical calculation gives a probability of finding a 6-bp sequence motif in a random 47.8-bp region of $[(47.8-5)x2]/4^6 \rightarrow p=0.0208$ ($p>0.01$). Based on this *p*-value, 6-bp motifs might not be statistically relevant. However, it is difficult to evaluate the statistical relevance of 6-bp motifs based on this *p*-value only, especially because so many different parameters can influence it. Therefore, to test to what extent the reported 6-bp identities are statistically significant, we performed an experimental, statistical analysis. Based on the assumption that the origin of 6-bp sequence motifs in the filler DNA can be attributed to the sequence of the 200-bp plant DNA region surrounding the T-DNA integration point, a statistical difference should be observed between the number of 6-bp sequence identities found with the original plant DNA and with a randomly chosen plant DNA segment. If we assume that a 6-bp motif is not relevant, we should observe a similar number of 6-bp identities when a filler sequence is compared with the plant target surrounding another T-DNA insert. In addition, a similar analysis was done with a 50-bp plant target sequence as well to estimate the impact of the distance between the reported 6-bp sequence blocks and the filler sequence (Table D).

Table D. Sequence identities observed when the actual filler segment is compared with a 200-bp plant DNA surrounding another T-DNA integration site

	kg353LB-21	kg44LB-2	kg44LB-3	kg104LB-5	kg135RB-4	CK2L10RB-7	kg135RB-3
200-bp target							
kg353LB-21		0	6	0	6	0	0
kg44LB-2	0		6, 6, 6, 8	0	0	6	6, 7
kg44LB-3	0	7		0	0	0	0
kg104LB-5	0	0	0		0	0	0
kg135RB-4	0	6, 6	6, 7	6		6, 7, 8	0
CK2L10RB-7	0	0	0	0	0		6
kg135RB-3	0	0	0	0	0	0	
50-bp target							
kg353LB-21		0	6	0	6	0	0
kg44LB-2	0		6, 6	0	0	0	0
kg44LB-3	0	0		0	0	0	0
kg104LB-5	0	0	0		0	0	0
kg135RB-4	0	0	0	0		6, 8	0
CK2L10RB-7	0	0	0	0	0		6
kg135RB-3	0	0	0	0	0	0	

The observed identities are given; 0, no sequence identities were observed; 6, 7, or 8, identical sequence motifs of 6 bp, 7 bp, or 8 bp were observed, respectively.

What were the conclusions from this analysis? From Table C, we learn that the 45-bp filler sequence of junction kg135RB-4 harbors six different 6-bp sequence motifs, identical to sequence blocks that are present in the 200-bp plant DNA surrounding the actual T-DNA integration point. Table D shows that this filler sequence harbors a 6-bp, a 7-bp, and an 8-bp identical sequence motif when compared to the plant DNA surrounding T-DNA integration CK2L102RB-7. Also, for other filler sequences similar results were obtained. When the results for the 50-bp region surrounding the T-DNA integration point are considered, we see that the number of identical sequence motifs decreases. Although, still, for four out of seven T-DNA junctions, the origin of the actual filler sequences can be explained to some extent by taking into account the 50-bp plant DNA region of another T-DNA integration event. Therefore, we feel that at least for the large filler insertions (kg44LB-2 and kg135RB-4, 48-bp and 45-bp of filler DNA, respectively) 6, 7, or 8 bp plant DNA-derived sequence motifs might not be statistically significant to explain the origin of the filler insertions discussed in the manuscript. Therefore, a threshold of 9 bp was set for reporting statistically relevant sequence motifs.

Testing the statistical significance of sequence motifs, identical to sequence blocks present in an 100-bp T-DNA border region contiguous with the filler DNA

In order to explain the origin of the filler sequences, we also screened for sequence identities between the T-DNA border immediately adjacent to the filler DNA and the filler sequence itself. A 100-bp region of the T-DNA border was used for this identity search. To evaluate the statistical relevance of T-DNA border-derived sequence motifs, we used a testing method different from that for plant DNA-derived identities, because the T-DNA border sequence is the same for all filler insertions, whereas the plant DNA target is different for each filler sequence.

Therefore, to test the statistical significance of the T-DNA border-derived sequence motifs, the actual filler was shuffled with the SHUFFLESEQ software (EMBOSS software package) that shuffles a given sequence, maintaining the composition of the sequence. If we assume that the reported T-DNA border-derived sequence identities are not statistically significant, then the number of observed T-DNA-derived motifs for the shuffled filler sequences should be statistically similar to that of actual filler insertions. The results of this statistical analysis are shown in Table E.

Table E. Number of identical sequence motifs observed for actual filler DNAs and shuffled filler DNAs when a 100-bp T-DNA border sequence is used for an identity search

T-DNA junction	Actual filler insertions number of motifs (> 6 bp)	Shuffled filler sequences number of motifs (> 6 bp)
kg353LB-21	3	0
kg150LB-8	2	0
kg150LB-6	4	0
kg269LB-16	2	0
kg44LB-2	2	0
kg44LB-3	1	0
kg162LB-12	2	0
kg353LB-19	2	0
CK2L7RB-2	1	1
kg32RB-1	1	0

From Table E, we can conclude that the number of identical sequence motifs found when the actual filler is compared with the 100-bp T-DNA border region differs from that of the shuffled filler sequence. Our results clearly show that short identical sequences are present more often than expected by chance .

We also looked into more detail to the 6-bp sequence motifs, identical to the 100-bp T-DNA border contiguous with the filler DNA(Table F). We see that ten (Table F) and only one (Table E) 6-bp motifs are observed when the actual and shuffled filler sequences are compared with the 100-bp T-DNA border region, respectively. These data clearly show that the reported 6-bp T-DNA-derived sequence identities are statistically relevant for explaining the origin of the observed filler sequences.

Table F. Summary of the observed identities between filler DNA and a 100-bp T-DNA border sequence for 10 plant DNA/T-DNA junctions

T-DNA junction	Length of filler (bp)	Motif (bp)	Distance (bp)
Left			
kg353LB-21	13	6	81
		6	25
		7	22
kg150LB-8	8	6	44
		6	64
kg150LB-6	23	6	62
		6	23
		7	25
		10	24
kg269LB-16	33	7	52
		18	1
kg44LB-2	48	6	81
		7	47
kg44LB-3	7	6	6
kg353LB-19	17	7	13
		12	54
kg162LB-12	8	6	34
		13	92
Right			
CK2L7RB-2	15	9	13
kg32RB-1	7	6	30

Fisher exact test for testing the correlation between T-DNA processing and the outcome of the integration process

We performed a statistical analysis of the data presented in Table I (see manuscript) to verify whether the degree of T-DNA processing might be correlated with the outcome of the T-DNA integration process. Table I is too sparsely populated (more than 20% of the reported values are below 5) for a valid chi-square analysis; therefore, an exact Fisher test for a 2x2 contingency table was performed. Because the intuitive difference we observed between intact and processed T-DNA ends is linked with filler DNA formation and presence of microhomology, we can put these data in a 2x2 contingency table (Table G).

Table G. 2x2 contingency table for Fisher test of statistical significance

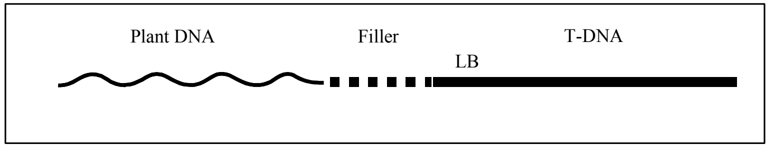
T-DNA ends	Microhomology	Filler formation
Intact	3	8
Processed	30	19
Left border	23	17
Right border	10	10

By performing a Fisher exact test, we compute the probability of obtaining exactly the frequencies observed and any configuration more extreme, given the observed marginal frequencies. This statistical value gives an estimation of the probability of finding by chance a table more extreme than our Table G. When we apply this test to the 2x2 contingency table for the intact and processed T-DNA ends, we see that the probability of obtaining a 2x2 contingency table as or more extreme than that presented in Table G is $p=0.043$ ($p<0.05$).

In order to evaluate whether there is a statistical difference between left and right T-DNA borders with regard to the formation of filler DNA or the presence of microhomology, the 2x2 contingency table for the left and right border was analyzed with chi-square (which is valid because the observed frequencies are high enough). $\chi^2=0.303$, df=1 $\rightarrow$ $p=0.582$. We can conclude that left and right borders do not differ when the formation of filler and the presence of microhomology are considered. In analogy, when the Fisher exact test is applied for the 2x2 contingency table for the left vs. right borders, the probability is $p=0.796$ of finding a table as or more extreme than Table G.

Supplemental data

A) Origin of filler DNA at left border T-DNA/plant DNA junctions



1) kg44LB-2: plant/48/LB-8

* ** *

tagggattcagaaag/20bp/agtgagcacagatcaaa**tgagtcgtgagtgatcgattccccgctttcgttcgagtaaagggatag**ATTCAATTGTA/36bp/AATGAGT/27bp/GAGTAA

31 T-DNA (LB-293 bp) 32 31 32

2) kg44LB-3: plant/7/LB-5

aatacatctgtatctaagaacatgaccaaaggttaa**gatttac**TCAATTGTAAATGGCTTCATGTCCGGGAAATCTACAT

33 33

3) kg162LB-12: plant/8/LB-28

aacgatggatggtggtgatcacgatga**tgatgtag**ATGTCCGGGAAAT/10bp/CAGCAATGAGTATGATGGTC/30bp/ATTTTTTTTCAATTCAAAAATGTAGATGTCCGCAG

34 34 35

35

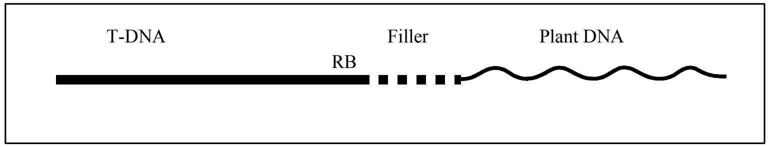
4) kg269LB-16: plant/33/LB 0

36 36

catgggatgaatctacgtgacatttgggggatga**gcaatgtcacgtaggatatattcaattgtatcc**CAGGATATATTCAATTGTAAATGGCTTCATGTCCGGGAAATCTACATGGATCAGCAATGA

37 38 38 37

B) Origin of filler DNA at right border T-DNA/plant DNA junctions



1) CK2L102RB-7: RB-10/28/plant

TCGGGCCTAACTTTTGGTGTGA**cagatagctgggcaatgaaaccatattg**gcaacccaaaagaagaaga

T-DNA (RB-1693 bp)

2) kg32RB-1: RB0/7/plant

39 39

TAATTTCAAACTATTCGGGCCTAACTTTTGGTGTGATGATGCTGACTGG**tttttcg**gggtatatactatactttgtggtatctggcctaaccccgaaaaaagcc

3) CK2L7-RB2: RB-2/15/plant

40

41 41 40

GGGCCTAACTTTTGGTGTGATGATGCTGACT**aattaggtgtttttg**gtgttttggtagtcttaaatatgtgagaagacttatgtttttaacccatacctttcactaattaggtgaactg

42 42

4) kg135RB-4: RB0/45/plant

CCTAACTTTTGGTGTGATGATGCTGACTGG**gcttgcttgaacgcttctacgactaggagataaatggttagcttg**agagta/10bp/tattaaatgta/10bp/tagagagtt/18bp/tgttgcttct

43 T-DNA (RB-638 bp)

43