

The Structural Features of Thousands of T-DNA Insertion Sites Are Consistent with a Double-Strand Break Repair-Based Insertion Mechanism

Nils Kleinboelting^{1,3}, Gunnar Huep^{1,3}, Ingo Appelhagen¹, Prisca Viehovever¹, Yong Li² and Bernd Weisshaar^{1,*}

¹Center for Biotechnology & Department of Biology, Bielefeld University, Universitaetsstrasse 25, 33615 Bielefeld, Germany

²Department of Medicine IV, University Hospital Freiburg, Berliner Allee 29, 79110 Freiburg, Germany

³These authors contributed equally to this article.

*Correspondence: Bernd Weisshaar (bernd.weisshaar@uni-bielefeld.de)

<http://dx.doi.org/10.1016/j.molp.2015.08.011>

ABSTRACT

Transformation by *Agrobacterium tumefaciens*, an important tool in modern plant research, involves the integration of T-DNA initially present on a plasmid in agrobacteria into the genome of plant cells. The process of attachment of the agrobacteria to plant cells and the transport of T-DNA into the cell and further to the nucleus has been well described. However, the exact mechanism of integration into the host's DNA is still unclear, although several models have been proposed. During confirmation of T-DNA insertion alleles from the GABI-Kat collection of *Arabidopsis thaliana* mutants, we have generated about 34 000 sequences from the junctions between inserted T-DNA and adjacent genome regions. Here, we describe the evaluation of this dataset with regard to existing models for T-DNA integration. The results suggest that integration into the plant genome is mainly mediated by the endogenous plant DNA repair machinery. The observed integration events showed characteristics highly similar to those of repair sites of double-strand breaks with respect to microhomology and deletion sizes. In addition, we describe unexpected integration events, such as large deletions and inversions at the integration site that are relevant for correct interpretation of results from T-DNA insertion mutants in reverse genetics experiments.

Keywords: T-DNA integration, *Agrobacterium tumefaciens*, double-strand break repair, *Arabidopsis thaliana*

Kleinboelting N., Huep G., Appelhagen I., Viehovever P., Li Y., and Weisshaar B. (2015). The Structural Features of Thousands of T-DNA Insertion Sites Are Consistent with a Double-Strand Break Repair-Based Insertion Mechanism. *Mol. Plant.* **8**, 1651–1664.

INTRODUCTION

Transgenic plants, especially from dicotyledonous species, are usually generated using *Agrobacterium tumefaciens*-mediated T-DNA transfer (Gelvin, 1998). T-DNA is originally a part of the *A. tumefaciens* tumor-inducing (Ti)-plasmid (Zambryski et al., 1980; Hoekema et al., 1983). It is flanked by two border sequences of 25 base pairs (bp) in length. These border sequences are similar to each other and were classified as left border (LB) and right border (RB) (Dürrenberger et al., 1989; Tinland et al., 1994). Original T-DNA, which is transferred to plants by naturally occurring agrobacteria, contains (1) sequences with tumor-inducing properties (Van Larebeke et al., 1974), and (2) genes that encode enzymes for the production of opines, which the agrobacteria can use as source of carbon and nitrogen (Schell et al., 1979). It is possible to substitute all the sequences between LB and RB without perturbing the ability of T-DNA to integrate into the plant genome, a feature

that allows the exploitation of T-DNA as a molecular tool (Wang et al., 1984; Peralta and Ream, 1985). The T-DNA-based transformation system has been applied in very diverse experimental setups and also in plant breeding (O'Malley and Ecker, 2010; Que et al., 2014).

The journey of T-DNA into the plant genome starts in *A. tumefaciens* with the creation of a single-stranded copy of T-DNA (T-strand) (Magori and Citovsky, 2011). VirD1 and VirD2, which are encoded by the (Ti)-plasmid, function together as an endonuclease complex. Nicking of T-DNA takes place within LB and RB directly before the recognition sequence 5'-CAGGATATATT-3', shortening LB by 3 bp and RB by 22 bp, respectively (Jasper et al., 1994).

Published by the Molecular Plant Shanghai Editorial Office in association with Cell Press, an imprint of Elsevier Inc., on behalf of CSPB and IPPE, SIBS, CAS.

After binding of VirD2 to the 5' end of the T-strand, the resulting complex is transported to the plant cell by a type IV secretion system (Gelvin, 2010; Pitzschke and Hirt, 2010). VirD2 stays attached to the 5' end of the T-strand during the whole process and is thought to protect the 5' end from exonucleases (Dürrenberger et al., 1989). Within the plant cell and after transport to the nucleus (Lacroix et al., 2005), the T-strand is probably converted into double-stranded T-DNA before integration (Tzfira et al., 2004; Liang and Tzfira, 2013). T-strands can also form T-circles, which lack "free" ends and are probably not integrated (Singer et al., 2012; Rolloos et al., 2014). Most likely, host factors involved in double-strand break repair (DSBR) are exploited for the integration process, which is indicated by altered transformation efficiency in mutants lacking key enzymes for DSBR (Mestiri et al., 2014).

Double-strand breaks (DSBs) can lead to serious problems if left unrepaired (Yoshiyama et al., 2013). In yeast and bacteria, DSBs are mainly repaired via homologous recombination (HR), usually without altering the nucleotide sequence. The main pathway for DSBR in plants is non-homologous end-joining (NHEJ). NHEJ is associated with rearrangements of the target site (see Gorbunova and Levy, 1999 and references therein), and DSBR by NHEJ requires a small microhomology of up to 5 bp between the two broken ends. Annealing at these microhomologies leads to overhanging 3' ends and can result in small deletions, target-site duplications, or, in rare cases, an accurate repair (Waterworth et al., 2011). Incorporation of short DNA stretches unrelated in sequence to the broken region (designated "fillers"; Gorbunova and Levy, 1997) during NHEJ has been explained by the synthesis-dependent strand annealing (SDSA) mechanism whereby open ends of DSBs are elongated by synthesis with a non-homologous template (Gorbunova and Levy, 1999).

Besides HR and NHEJ there is a third, less well characterized pathway for DSBR, which uses substantial microhomology to repair DSBs independent of the key enzymes for NHEJ (Ku70, Ku80, DNA ligase IV, and XRCC4). This so-called microhomology-mediated end-joining (MMEJ) was first described in mammals (see McVey and Lee, 2008 and references therein), but there is also evidence for MMEJ in plants (see Waterworth et al., 2011 and references therein). It can be distinguished from NHEJ by the use of larger microhomologous sequences (5–25 bp) during annealing of broken ends and larger deletions than in NHEJ as a result. MMEJ has also been associated with more complex rearrangements such as translocations or inversions.

A currently popular model for T-DNA integration involves the NHEJ pathway, in which DSBs within the plant genome are targets for the integration (Salomon and Puchta, 1998; Ray and Langer, 2002; Britt and May, 2003; Chilton and Que, 2003; Tzfira et al., 2003). During NHEJ a few nucleotides can become deleted at the DSB site (Weaver, 1995), which has also been observed for some T-DNA insertions (Windels et al., 2003). Moreover, small areas of microhomology between T-DNA and neighboring plant genomic DNA as well as rearrangements such as deletions and/or duplications of target-site sequences have been found during analyses of flanking sequence tags

(FSTs; Brunaud et al., 2002) and several T-DNA insertion sites (Forsbach et al., 2003).

Besides the DSBR-based model for integration, two models for the integration of single-stranded T-DNA have been proposed (Tzfira et al., 2004). In both models, binding of single-stranded T-DNA to one strand of the target DNA takes place in areas of microhomology followed by synthesis of the other strand. A variable size deletion would be the result for both of these models. However, in addition to the integration of the sequences between LB and RB, it is also known that more complex arrangements of T-DNAs including vector backbone sequences from outside of the T-DNA may occur at insertion sites (see Krizkova and Hroudá, 1998 and references therein). Also, parts of the *A. tumefaciens* chromosome sequences are carried over to the plant genome together with T-DNA (Ulker et al., 2008). In contrast to a model where a single-stranded T-DNA is incorporated into the plant DNA, the double-strand model can also explain the integration of such complex T-DNA inserts containing multiple T-DNAs linked together (De Neve et al., 1997; De Buck et al., 1999). It has also been reported that chromosomal translocations were found in 19% of examined T-DNA insertion lines (Clark and Krysan, 2010).

The integration sites within plant genomes are largely randomly distributed (Kim et al., 2007), although it is still debatable whether there are preferences for certain genomic regions. It has been shown that there are preferences for transcription initiation sites and polyadenylation sites and regions outside of centromeric regions, at least after selecting insertion events using a marker gene that needs to be expressed (Szabados et al., 2002; Li et al., 2006). In contrast, it has been reported that the integration sites are distributed within the genome in a completely random manner under non-selective conditions (Kim et al., 2007).

A special application for T-DNA insertions in plant genomes is the generation of large FST-indexed insertion line collections. These collections are a resource for potential knockout alleles, because a T-DNA insertion within or close to a gene usually perturbs this gene's function. Insertion sites were determined by evaluating FSTs (Samson et al., 2002; Alonso et al., 2003; Strizhov et al., 2003; Thole et al., 2009). For *A. thaliana*, several T-DNA insertion line collections exist, of which the SALK and GABI-Kat collections are the largest, comprising about 88 000 and 72 000 indexed lines, respectively. Both collections focused on analyzing junctions to genomic sequences at the LB, because it turned out empirically that the success rate of FST production was higher. At least in part, this can be explained by the fact that T-DNA integrations often appear in an LB-rb:rb-LB configuration that results in LB at both junctions to genomic sequences (Strizhov et al., 2003). The insertion lines can be ordered, either at common stock centers (Scholl et al., 2000) or directly at the Web pages of the collections (Alonso et al., 2003; Kleinboelting et al., 2012). In the case of the GABI-Kat collection, the FST predictions for the insertion sites of T-DNA are confirmed via PCR with border primers and insertion site prediction-specific primers and sequencing before ordered lines are delivered to users. These confirmation sequences are in most cases of better quality than the FSTs, and allow an improved estimation of the exact insertion position.

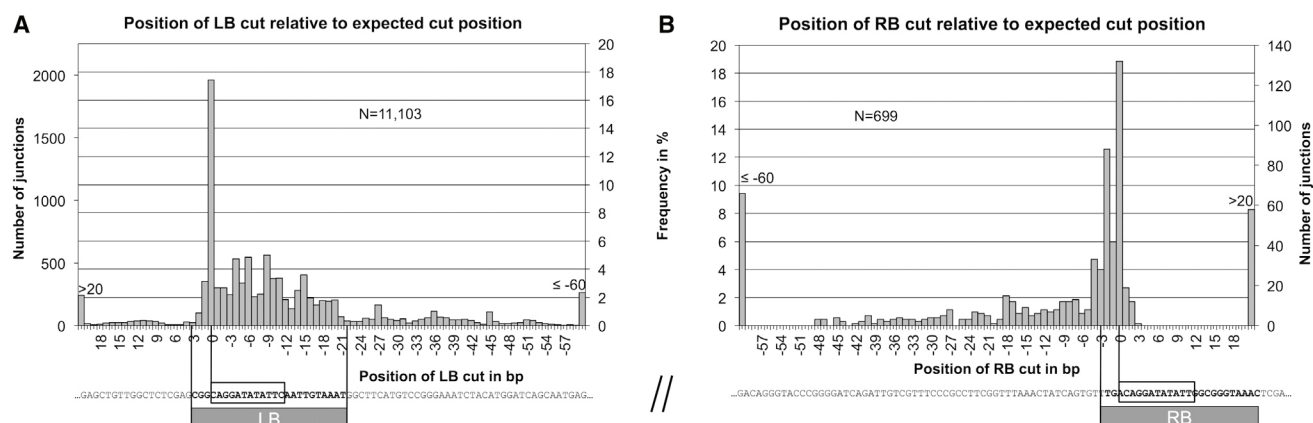


Figure 1. Distribution of Cut Positions in Integrated T-DNA Border Sequences.

Locations of cuts in the border sequence or T-DNA are shown on the x axis. Zero (0) corresponds to the expected cut position within LB (A) or RB (B) sequence; only T-DNA sequences close to LB and RB are shown, and the main part is represented by //. Negative values indicate borders further shortened beyond the expected cut position. Positive values correspond to inclusion of sequences beyond the expected cut position displaying continuous sequence similarity to the T-DNA vector backbone. The highest peak is at the expected location for LB and RB; further shortened border sequences are quite common.

In this study, we present a detailed analysis of confirmation sequences obtained over the years from the GABI-Kat project. For a fairly large number of randomly selected lines, confirmation sequences were created for both junctions of the insertion, which enabled a comprehensive analysis of the size of deletions caused by the integration mechanism. In addition, we investigated the presence of microhomology or fillers at the T-DNA/plant genome fusion, and performed further analysis of filler sequences. We observed some unusual integration events whereby large parts of the genomic DNA were deleted, as well as rare cases whereby a part of the plant DNA was inverted at the insertion site.

RESULTS

Size of RB and LB Sequences Remaining at T-DNA to Genome Junctions

As a first step, we determined the number of bases that were lost from the initially complete borders (LB and RB) of the original T-DNA sequence. The length distribution of T-DNA border sequences detectable after integration is depicted in Figure 1, by assigning position zero (0) to the expected cut position. For LB, 11 103 individual confirmation sequences that each correspond to exactly one junction of an insertion were evaluated. The peak for LB with 17.7% of all sequences was observed at position 0. In 72.3% of all cases, one or more bases were lost, and in 10.0% of all cases the LB sequences were found to contain bases beyond the expected cut position. The largest fraction of these longer border sequences were 1 bp longer (3.2%). In a few cases, sequence from the vector neighboring T-DNA (“outside” of the border) was also integrated (5.6%), with a median length of 13 bp of these additional sequences.

The number of cases available for analysis for RB was lower, because GABI-Kat FSTs were mainly generated from the LB of T-DNA, and therefore fewer confirmation sequences at RB were available for analysis. A total of 699 insertions were analyzed for RB, and also for RB the peak was observed at the expected position (18.9%). In 68.2% of all RB cases one or more bases were

lost. The remaining 12.9% represent cases with a longer RB sequence transferred into the *A. thaliana* genome. In 8.3%, sequences from the vector were integrated, with a median size of 21.5 bp. The median position of the border cut relative to the expected cut position for LB in our dataset was −7 and −2 for RB, indicating that LB gets shortened further than RB (see Discussion). Details on all evaluated confirmation sequences and their cut positions can be found in Supplemental Information File 1.

For both borders, the peak was located immediately upstream of the sequence pattern 5'-CAGGATATATT-3'. We conclude that the main cut position precisely fits the position defined previously, but also that there is considerable heterogeneity with respect to the actual position of the cut, which also happens to be located outside of LB and RB. The nucleotides that fit the original border sequence “outside” of the cut position (numbering up to three) might be contributed by a random fit of sequences contributed by the pre-insertion site. The value for a border one base longer than expected is approximately 25% of the size of the peak for the expected size. This percentage (25%) is also roughly the probability for a random matching base within the plant genomic DNA. This also applies for further values of junctions with two, three, or four additionally fitting bases. The transfer of longer sequences containing vector sequence was much more common than expected as a result from random sequence similarity. The loss of sequences from T-DNA “inside” of the cut position might be a consequence of the integration mechanism.

Microhomology and Fillers at Integration Sites

To examine the insertion sites in more detail, we determined the extent of microhomology of sequences at individual T-DNA ends to the genomic sequence at the insertion site, and the presence or absence of fillers, by sequence comparisons. Figure 2 shows an overview of the size distribution of microhomology and fillers at integration sites. Supplemental Information File 1 contains the

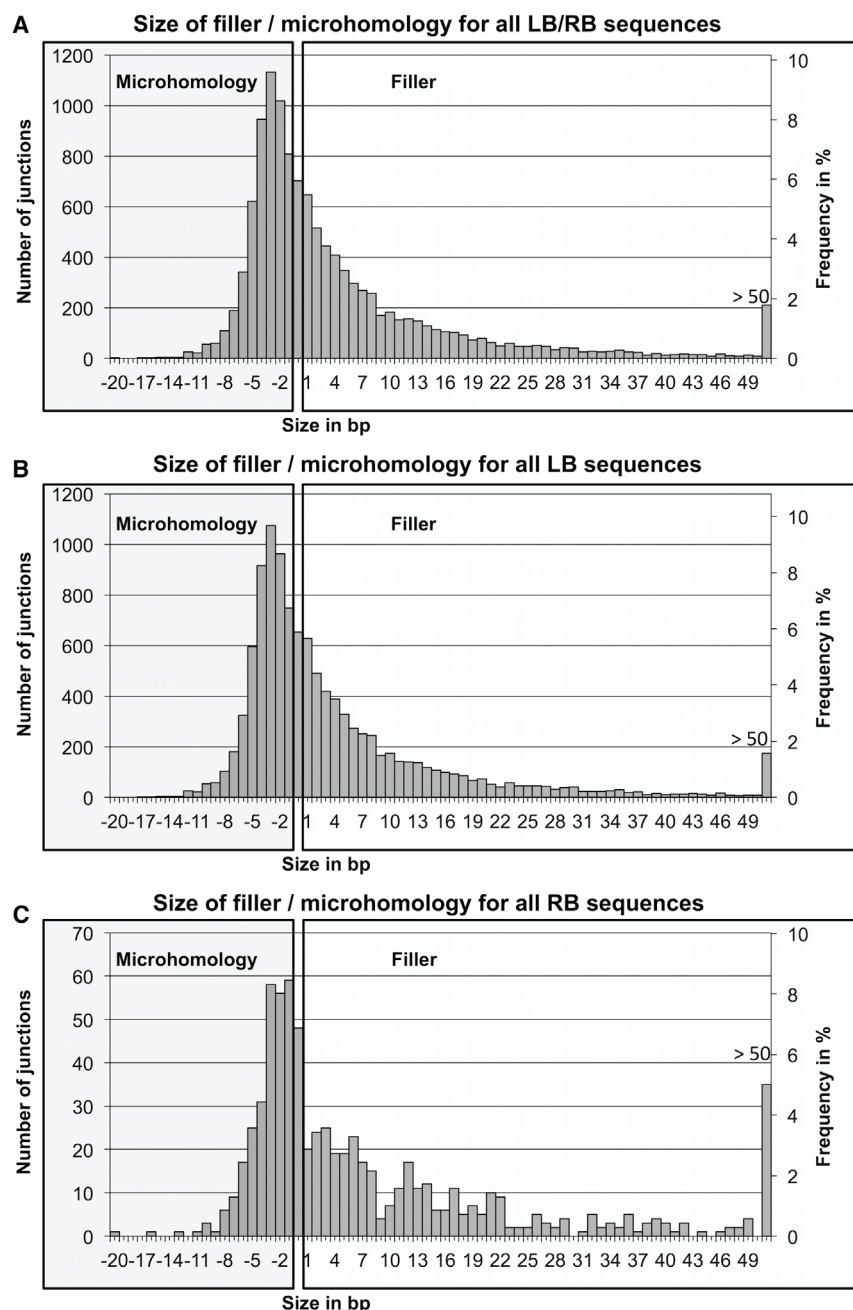


Figure 2. Size Distribution of Microhomologies and Fillers at Integration Sites.

Length of microhomologous sequences and sequences present between the end of T-DNA and plant genomic sequences at the insertion sites (see text for definition). A positive value corresponds to a filler between T-DNA and plant genomic DNA, and a negative value to a microhomology of the size indicated in base pairs. Zero indicates a seamless transition from T-DNA to plant genomic DNA sequence. Cases from RB and LB plotted together show a significant portion of cases with microhomology as well as the common occurrence of fillers (**A**). Microhomology and fillers can be found at RB and LB in similar size distributions (**B** and **C**).

(74.8%) sequences had microhomology of up to 5 bp. The remaining 826 (13.6%) sequences showed microhomologies of 6 bp or more, and the largest microhomology detected was 20 bp in two observed cases.

When separating the data for LB and RB (**Figure 2B** and **2C**), the size distributions were found to be comparable. Fillers were observed in 48.3% (5368 cases) for LB and 54.5% (381 cases) for RB. The fillers were larger than 10 bp in 18% for LB and 29.8% for RB (2001 and 208 cases) and larger than 20 bp in 8.4% for LB and 16.7% for RB (938 and 117 cases). The mean filler was 13 bp for LB and 22.2 bp for RB; the median was 7 bp for LB and 12 bp for RB, indicating slightly larger fillers at RB. When comparing microhomology sizes of LB and RB, 5735 (51.7%) sequences were evaluated for LB and 318 (45.5%) for RB. Among them, 5081 (88.6%) and 270 (85%) sequences showed microhomology of at least 1 bp for LB and RB, respectively; 4296 (74.9%) and 229 (72.0%) of those were between 1 and 5 bp, and 785 (13.7%) and 41 (12.9%) were larger than 5 bp. The mean microhomology was 3.18 bp for the LB and 3.48 bp for the RB, and the

underlying data. In line with what has already been mentioned, we define a filler as a DNA sequence present between the end of T-DNA sequence similarity of a given insertion and the plant genomic DNA sequence at the insertion site that is not a direct extension of either of the two. Overall, 11 802 sequences were analyzed, 11 103 for LB and 699 for RB. Fillers were detected on a regular basis in a total of 5749 cases (**Figure 2A**). Of the 11 802 insertions studied, 48.7% showed fillers of at least 1 bp, 2209 (18.7%) larger than 10 bp, and 1055 (8.9%) even larger than 20 bp. The remaining 6053 junction sequences without fillers were used to examine microhomology between the individual T-DNA sequence ends and the adjacent plant genomic DNA (**Figure 2**). A total of 5351 (88.4%) of these junctions showed microhomology of at least 1 bp, and the largest fraction of 4525

median was 3 bp for both borders. Thus, both borders show a very similar pattern concerning microhomologies.

In conclusion, both borders show substantial microhomology to the *A. thaliana* genome sequence in roughly half of the cases, with similar size distribution. The remaining cases showed fillers, with slightly larger fillers at the RB.

Analysis of Fillers

To gain insights into what might have caused the fillers, we determined the origin of the detected fillers by comparing the sequences bridging the gap between T-DNA and genomic DNA (with a small overlap) to possible origins. **Table 1** gives an

BLAST target	No. of hits
<i>A. thaliana</i> , same chromosome	548/1508 (36.3%)
<i>A. thaliana</i> , other chromosome	883/1508 (58.6%, 14.7% per chromosome)
GABI-Kat T-DNA within 100 bp distance to LB/RB	46/1508 (3.1%)
GABI-Kat T-DNA outside 100 bp distance to LB/RB	2/1508 (0.1%)
GABI-Kat vector backbone within 100 bp distance to LB/RB	1/1508 (0.1%)
GABI-Kat vector backbone outside 100 bp distance to LB/RB	8/1508 (0.5%)
<i>A. tumefaciens</i> (linear, circular, plasmid)	16/1508 (1.1%)
Chloroplast genome	0/1508 (0%)
Mitochondrial genome	2/1508 (0.1%)
Unknown	2/1508 (0.1%)

Table 1. Sources of Fillers.

A total of 1508 fillers were subjected to BLAST analysis, and results classified according to origin of the hit sequence.

overview of the origin of the detected fillers. 94.9% of all best BLAST hits for fillers identify sequences within the *A. thaliana* genome. The second most abundant BLAST hits were found on T-DNA sequence (3.2%). Other hits were rare, and can be explained as random sequence similarity due to the short length of the sequences that had to be studied. Exceptions are sequences that might have been introduced together with T-DNA, namely sequences with hits to the *A. tumefaciens* genome sequence or vector backbone hits. Transfer of genomic DNA of *A. tumefaciens* along with T-DNA has been described previously (Ulker et al., 2008). Here, we detected two fillers larger than 20 bp that are derived from *Agrobacterium* chromosomes (see Supplemental Information File 2).

We examined the distance from the source sequence for the fillers to the location of the insertion site containing the respective filler. Figure 3A shows the distribution of the distances to the corresponding insertion position derived from the confirmation sequence. A large fraction of all hits (36.3% in comparison with 14.7% as mean for the other four chromosomes) was found on the same chromosome. Those fillers displayed a bias toward 1 kbp direct genomic proximity to the insertion site (Figure 3B and 3C). Despite this bias for the “local” source of fillers within 1 kbp of the insertion site, fillers from other genomic regions were randomly distributed within the respective chromosome: without the high peak at close genomic proximity, the distance distribution of positions resembles that of the difference of two equally distributed random variables (one being the insertion site, the other the filler’s source locus). When all fillers were evaluated, the distribution of hits on other chromosomes was equally distributed and scales quite well with the relative size of the source chromosome. When fillers from insertions on an individual chromosome were evaluated for their source location, there was also no convincing bias for a certain source chromosome (Figure 3D).

After identifying the source locus of fillers, we were able to check whether there was substantial microhomology between a given filler and the adjacent sequences at the fusions to either T-DNA or the genome sequence (Figure 3E and 3F). Filler source sequences showed small areas of microhomology to the sequence neighboring the filler at the insertion site in many cases: 22.8% for the fusion to T-DNA and 25.8% for the

fusion to genome sequence. This is approximately half the rate found for the T-DNA/genomic-DNA junction, an observation suggesting that only one side of the filler is repaired using microhomology. The remaining cases contained further “fillers of fillers” forming complex arrangements of sequences from various sources, which we did not analyze further.

In conclusion, the large majority of fillers originate from the same chromosome or at least from the *A. thaliana* genome, and from sequences available within the nucleus at the time of integration (other T-DNA molecules and sequences that “travel” together with T-DNA). There is a bias regarding the source of fillers toward the direct genomic neighborhood and the same chromosome of the integration site. The remaining filler source loci are randomly distributed throughout the genome. Fillers share some microhomology with their neighboring sequences at about half the frequency observed for T-DNA/genome junctions without fillers. All data from the filler analysis are presented in Supplemental Information File 3.

Deletion and Target-Site Duplication of Sequences from the Pre-insertion Site

To determine the effect of the integration on the genomic sequences present at the pre-insertion site, we generated confirmation sequences for both junctions (T-DNA/genome intersections at both ends of the inserted T-DNA) for 1319 individual insertions. These cases were analyzed for deletions and target-site duplications, and the results are shown in Figure 4 with regard to the size of the deletions/duplications detected. Most of the insertions (93.3%, 1252 cases) displayed alterations of the pre-insertion site sequence between a target-site duplication of 19 bp (value of −19 in Figure 4) and a deletion of 99 bp (value of +99 in Figure 4). The remaining 6.7% of cases displayed larger deletions or target-site duplications. A total of 80 insertions (6%) showed deletions of at least 100 bp and larger, and target-site duplications of at least 20 bp and larger were detected in nine cases (0.7%).

Deletions of up to 100 bp were observed in 86% of all examined insertions. The median of all deletion sizes was 19 bp. We observed target-site duplications in 89 cases (6.7%) with a median size of 4 bp. In only 12 of 1319 cases, neither a deletion nor a target-site

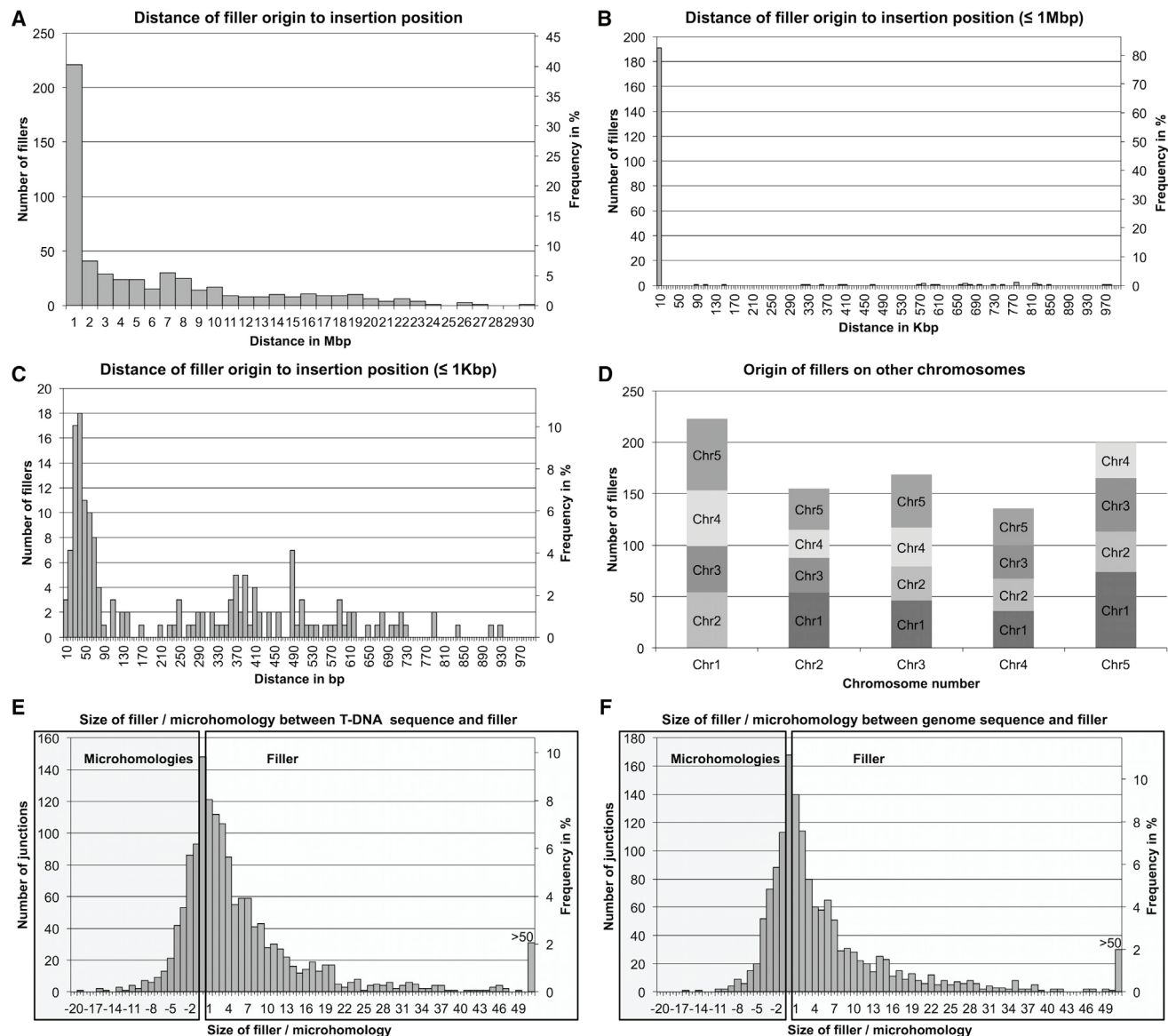


Figure 3. Characteristics of Fillers.

Most of the fillers originate from the same chromosome with a bias toward close genomic neighborhood to the site of insertion (**A–C**). There was no detectable bias for a specific chromosome as a source of fillers originating from another chromosome, neither when all fillers were evaluated nor when the fillers from insertions of a single chromosome were studied (**D**). Microhomology of a few bases as well as further fillers can be observed between the filler sequence and the neighboring *A. thaliana* genomic DNA (**E**) or T-DNA (**F**).

duplication was found, indicating that deletions and target-site duplications are a very common feature for T-DNA integration, with a strong bias toward deletions. A summary of all examined insertions can be found in [Supplemental Information File 4](#).

In conclusion, integration without some alteration of the pre-insertion site in the form of either a deletion or target-site duplication is very rare. T-DNA insertion usually leads to deletions of 1–50 bp with a smooth transition to target-site duplications, while larger alterations are the exception.

Comparison with Data from SALK Lines

To demonstrate that the observed insertion-site structures are not specific for the T-DNA used at GABI-Kat, we also analyzed a num-

ber of insertions from the SALK collection (see [Supplemental Information Files 5, 6, and 7](#)). With regard to sizes of deletions and target-site duplications, the 69 analyzed cases also showed a bias toward deletions of up to 100 bp (89.9%, 62 cases) and a median deletion size value of 19.5, similar to the result obtained for GK insertions. The dataset for the analysis of the borders of inserted T-DNA, microhomologies, and fillers consisted of 880 SALK cases, 852 for the LB and 28 for the RB. The SALK T-DNA sequences for RB and LB from pROK2 contain the recognition sequence (see above) identical to that found in the GK T-DNA, and we used this consensus to align the position values. The peak of cutting sites in SALK insertions was at the expected position for LB (–3) and RB (–22). However, for SALK LB cuts a smaller peak was located at around –22 to –24 bp, which

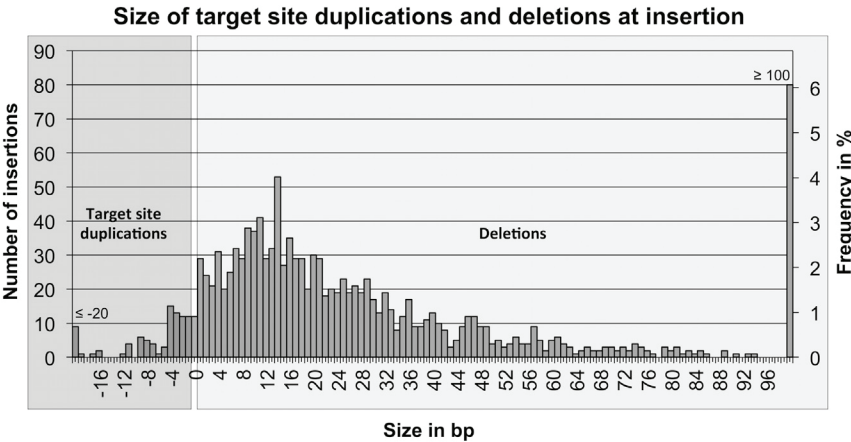


Figure 4. Size Distribution of Deletions and Target-Site Duplications.

Length of deletions and target-site duplications detected at insertion sites. These changes of the pre-insertion site sequences occur on a regular basis with a bias toward smaller deletions. Deletions larger than 100 bp and target-site duplications larger than 20 bp do exist but were observed rarely (see text for details).

resulted in an increased median value for the cut positions of -20 bp relative to the expected site. The median RB cut position was -3.5 bp distance to the expected site, fitting to the observations for GK RB cut positions.

Microhomology was observed in 413 (47%) of the examined SALK cases with a median size of 3 bp (3 for LB and 2 for RB). The median length of fillers was 8 bp (7 for LB and 17 for RB). The origin of fillers was determined for 109 cases, all of them matched to the *A. thaliana* genome sequence, with a small bias to the same chromosome (35 fillers). Frequency and size of microhomologies or fillers and their origin matched to the values obtained from the analysis of GK insertions, at least for the LB, where sufficient cases were analyzed. Taken together, the general picture of insertion-site modifications is similar for GK and SALK T-DNA alleles.

Larger Deletions

We detected several larger deletions (31 insertions with deletions larger than 1000 bp, eight of them larger than 5 kb) at the integration site. Two of these deletions were analyzed in more detail to confirm that the deleted region has really been removed and not translocated.

In line GK_144F03 a deletion of 6.1 kb on Chr3 between position 17 746 865 and 17 753 018 (according to TAIR10, see [Methods](#)) was predicted after evaluation of the confirmation sequences. We sequenced DNA from homozygous T3 plants of this line using an Illumina MiSeq, and mapped the reads (about $32.7\times$

coverage) to the *A. thaliana* (Col) genome sequence. The mapping result clearly confirmed the predicted large deletion because no reads were mapped to the deleted region with respect to the wild-type sequence (see [Figure 5](#) for a coverage plot). Gene *At3g48070*, encoding a RING/U-box superfamily protein, is annotated on Chr3 between position 17 750 711 and 17 752 761. The next-generation re-sequencing result confirms the complete removal of this gene from the genome sequence of GK_144F03, although this is not apparent from the initial FST data. However, a phenotype visible under standard greenhouse conditions was not observed in this mutant.

In line GK_478B05, we predicted a deletion of approximately 6 kb on Chr3 between position 19 022 337 and 19 028 218. In this region, the genes *At3g51230*, *At3g51238*, and *At3g51240* are annotated. *At3g51240* encodes flavanone 3-hydroxylase (*F3H*), an enzyme involved in flavonoid biosynthesis. *F3H* mutants display the *transparent testa* (*tt*) phenotype, which results from a reduced content of brown proanthocyanidins in the seed coat. The locus coding for *At3g51240/F3H* has therefore been named *transparent testa 6* (*tt6*). We phenotypically characterized homozygous offspring of GK_478B05 for the *tt6* phenotype and found all phenotypic features expected for *tt6* ([Figure 6](#)). Therefore, the insertion allele of GK_478B05 does not contain a functional *At3g51240/F3H* gene. We conclude that also in this case, the region predicted to be deleted at the insertion site is absent from the genome of GK_478B05 (if homozygous for the insertion allele). A list of the genes (AGI codes) that were affected by deletions identified in this study but were not detectable from direct FST evaluation is presented in [Supplemental Information File 8](#).

Detection of deletions larger than a few hundred base pairs was only possible if FSTs for both junctions of the insertion had been

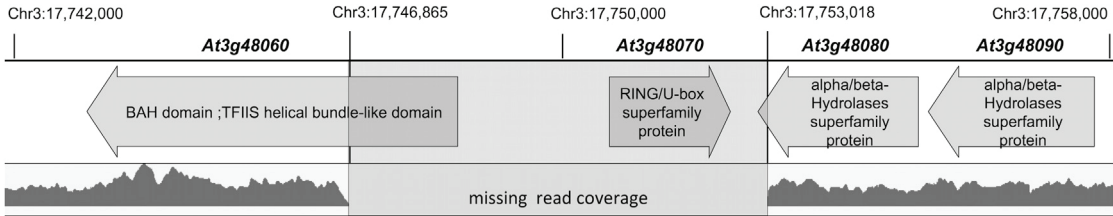


Figure 5. Coverage Plot for the Large Deletion in Line GK_144F03.

Diagram of a region on chromosome 3 that is affected by the deletion caused by the T-DNA insertion. The cumulative coverage for both strands is shown and indicates a deletion affecting three genes, of which *At3g48070* is lost completely. The mapping data were visualized using GenomeView (see [Methods](#)).

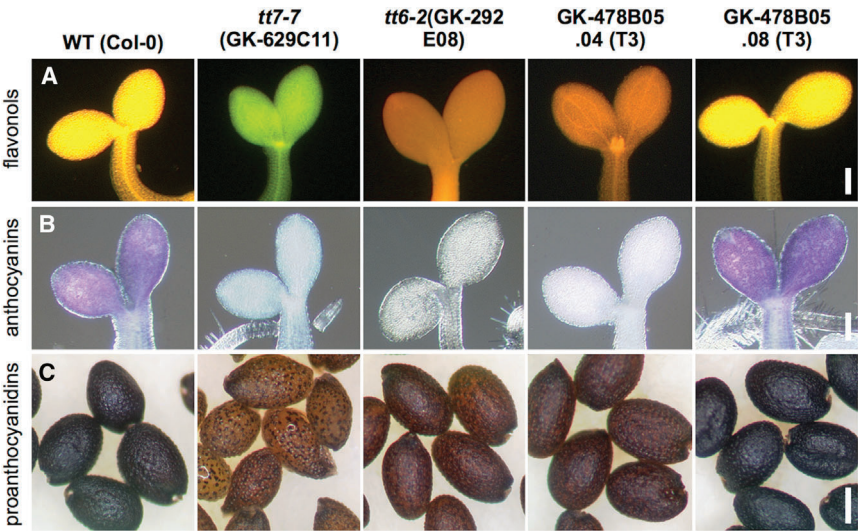


Figure 6. Large Deletion in Line GK_478B05 Affecting *At3g51240* and Causing the Corresponding *tt6* Phenotype.

GK_478B05.04, which is homozygous for the insertion, shows the same phenotype as the *tt6* knockout GK_292E08. GK_478B05.08 (a heterozygous sibling) shows the wild-type phenotype. The phenotypic characterization confirms the removal of the gene *At3g51240*. Flavonol accumulation in seedlings was visualized with DPBA and detected using epifluorescence microscopy (A) in homozygous (.04 T3) and heterozygous (.08 T3) plants of GK_478B05, and wild-type (WT) as well as *tt6-2* and *tt7-7* plants as controls. Anthocyanins in seedlings were detected using microscopy after bleaching with norflurazon (B). The proanthocyanidin content of mature seeds was visualized with DMACA staining and detected with a binocular (C). The scale bar always represents 200 μ m.

available. This availability was a result of the FST generation procedure of GABI-Kat, which involved two PCR reactions with primers differing at the 3' terminal base, and which by chance also provided FST for both termini from one insertion in some cases. Thus, the true frequency of larger deletions might be higher than our numbers indicate.

Inversions at the Integration Site

In some cases, we observed a seemingly incorrect orientation of the confirmation sequence relative to T-DNA. Upon detailed analysis of these cases, it turned out that genomic sequences at the insertion site were reversed. These inversions create another breakpoint close to the insertion position. Detection of such inversions was very rare but, similarly to larger deletions, we cannot exclude that they are happening more often and are larger than the ones we observed. Due to our preferred primer distance to the insertion site, and because we (obviously) focus on correctly oriented primers, the confirmation amplicon can only form if a given inversion is smaller than the distance from the primer to the insertion position. Nevertheless, we found 10 inversions of different sizes in our confirmation sequences. Table 2 summarizes the data, including the location and sizes of the inversions.

DISCUSSION

GABI-Kat generated a large, publicly available dataset of sequence-characterized T-DNA insertions with high sequence quality and accurate insertion-site positions in *A. thaliana*. We have analyzed these data to obtain further evidence for or against the available hypotheses on the mechanism of the T-DNA insertion process. The most popular theory assumes that T-DNA integrates as a double-stranded DNA molecule into DSBs, and utilizes plant repair pathways for integration (see above).

Deletions and Target-Site Duplications

We have analyzed the sizes of the deletions and target-site duplications at the integration site by examining insertions where both junctions of the insertion have been characterized by amplicon sequencing. Deletions usually range between 1 and 50 bp, but larger deletions have also been observed, while target-site duplications are shorter and less common. While the simple single-strand annealing repair model for T-DNA integration can explain deletions very well, target-site duplications are only possible when “sticky ends” are created during creation of DSBs. The maximum size of a target-site duplication is thus the distance between the two staggered single-strand breaks. This contradicts a

Line	BAC	Next gene (AGI)	Position of inversion	Size of inversion (bp)
GK_041A02	F24M12	At3g51057	Chr3:18963229–18963303	75
GK_072H03	T28M21	At2g40030	Chr2:16716822–16717129	308
GK_091G10	F22I13	At4g38380	Chr4:17975281–17975444	164
GK_118F06	T2H3	At4g02310	Chr4:1014928–1014979	52
GK_135D02	T32G6	At2g41570	Chr2:17338757–17338972	216
GK_182H05	F13I12	At3g47110	Chr3:17348070–17348138	69
GK_401E04	F6E21	At4g31170	Chr4:15155129–15155593	465
GK_526B04	T17B22	At3g03120	Chr3:717349–717511	163
GK_545H03	MEE6	At5g41050	Chr5:16434876–16434954	79
GK_692B09	F4D11	At4g32640	Chr4:15742746–15742807	62

Table 2. Inversions Detected in GABI-Kat Lines.

model in which the T-DNA integrates as a single-stranded molecule at a position that initially contains only a nick and not a DSB (Tzfira et al., 2004).

Nevertheless, the distribution we observed indicates the creation of sticky ends with randomly distributed length or blunt ends in most cases. Other random factors are the number of bases degraded by exonucleases and the region where microhomology takes place (determining the size of overhanging ends that are removed later). Microhomology-based base pairing taking place at the very ends of the single-stranded DNA of a sticky end during repair leads to target-site duplications, while microhomology-based base pairing distant from the ends leads to removal of overhanging ends and, therefore, to deletions. Both of these factors taken together (degradation of sticky ends and the necessity of microhomology at the ends) favor deletions, in line with the observed insertion structures.

According to Salomon and Puchta (1998), the sizes of deletions observed during DSB repair range from 2 bp to 1.2 kbp, which matches our observations for T-DNA fusions to genomic DNA. Because of our primer design (usually around 450 bp from the predicted insertion site) we rarely found deletions larger than about 300 bp, but the deletion size distribution shows that the main fraction of deletions is up to 30 bp in length, while larger deletions were less frequently observed with increasing size of the deletion. The existence of an undetected high number of large deletions is therefore unlikely. Nevertheless, due to FSTs present on both junctions of the insertion for a few GABI-Kat lines, we were able to verify deletions of several kilobase pairs, as shown in detail for the two cases in GK_144F03 and GK_478B05, and also for a total of 12 cases where a gene not associated with the FST was deleted during integration (see Supplemental Information File 8). The occurrence of events such as large deletions and target-site duplications should be considered during reverse genetics, which relies on T-DNA insertion mutants. It is important to always check both junctions of a T-DNA insertion allele to confirm that phenotypes observed in the mutant are really linked to the single gene apparent from an FST from just one border.

Furthermore, we detected several inversions at the site of T-DNA integration. With respect to the mechanism of integration, inversions have also been observed during MMEJ repair of DSBs (McVey and Lee, 2008). This is another piece of evidence that MMEJ plays a role in T-DNA integration. One explanation for inversions is the occurrence of two DSBs in close chromosomal proximity. After repair, the deleted fragment of genomic DNA is incorporated directly alongside T-DNA. It is unlikely that inversions take place prior to T-DNA integration because all inversions we detected were only found at one end of the T-DNA insertion and not spanning the integration site.

Microhomology and Fillers

The analyses of microhomology and fillers indicates that microhomology seems to be a prerequisite for integration most of the time, further supporting the NHEJ hypothesis in which microhomology between 1 and 4 bp is expected (Gorbunova and Levy, 1999). Nevertheless, we observed a significant number of cases with microhomology of up to 20 bp, indicating that

MMEJ might play a role more often than previously thought. The high abundance of fillers and their origin highlights the importance of the SDSA pathway for T-DNA integration. The fraction of insertion cases showing microhomology is significantly larger than what would be expected as a random observation of sequence similarity at the junctions. We observed microhomology of up to 20 bp on a regular basis. Larger microhomology was not observed, which can be explained by the lack of longer sequence similarity of the area surrounding the T-DNA borders to sequences within the *A. thaliana* genome. Along with the frequent presence of fillers, these results support the hypothesis that T-DNA insertion is largely mediated through the NHEJ pathway assisted by MMEJ. It would be an interesting experiment to compare the sizes of microhomologies in integrations of mutants where Ku70, Ku80, or XRCC4 are knocked out to check whether MMEJ is used more often in such a genetic background.

Regarding the source of the sequences found in fillers, the majority was found to be derived from the same chromosome. In addition, a large fraction of the source hits was found to be in very close genomic proximity to the insertion site. This is in accordance with fillers detected during DSB repair, and can be explained by assuming that SDSA finds its target more likely in close genomic proximity (Gorbunova and Levy, 1997). Fillers share microhomology with their adjacent T-DNA or genomic sequences with half the rate of junctions without fillers. This fits the assumption that a filler is created at one of the open ends (genomic or T-DNA) using SDSA, without the use of microhomology, and that the other part is repaired using plant repair pathways utilizing microhomology. Thus, only one side of the filler is expected to show microhomology, explaining the halved rate. Fillers from templates that were not from the target genome have been observed rarely, and can be explained by the use of sequences as source of the filler that have been imported from *A. tumefaciens* to the site of integration.

Processing of Borders

The analyses of the cut positions at the T-DNA border sequences confirmed the known cleavage site immediately upstream of the recognition sequence 5'-CAGGATATATT-3' (Jasper et al., 1994), but also detected a large fraction of T-DNA border sequences that have been further shortened. This varying degree of shortening might be dependent on the exposure time of T-DNA to exonucleases until integration into a target DSB. The RB of the T-strand is initially protected by VirD2 and is therefore not exposed to exonuclease activity. Consistent with earlier findings (Gheysen et al., 1991; Mayerhofer et al., 1991; Gelvin, 2000), this protection is also reflected in our data indicating fewer base pairs removed at the RB than at the LB. Nevertheless, there is a quite a large number of fusions where some bases were also lost at the RB before or during integration.

During DSB repair by NHEJ or MMEJ, further deletions are possible by removal of overhanging flaps not contributing to microhomology, which explains further shortened LB and RB sequences. Shortened border sequences can be observed in similar frequency for both borders (although the size differs), which leads to the conclusion that the same DSB repair pathway is used for processing of both borders during integration.

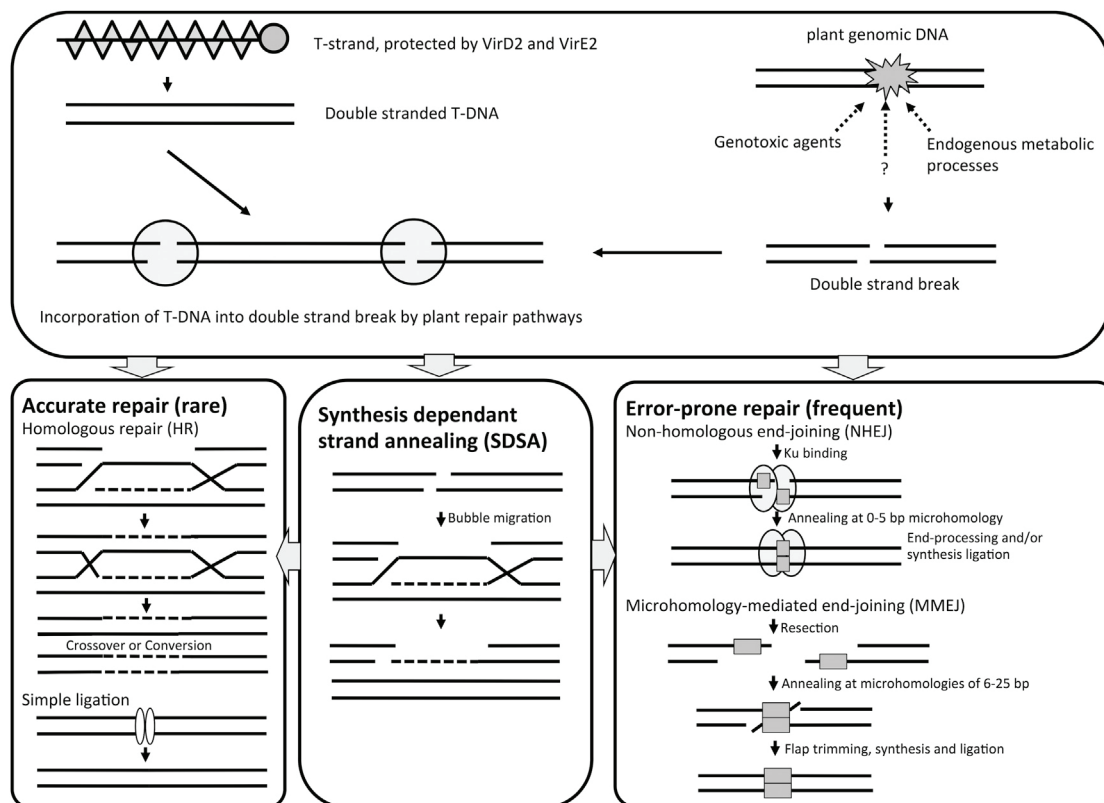


Figure 7. Proposed Repair Mechanisms Used for T-DNA Integration.

Summary of double-strand break repair mechanisms hypothesized to contribute to T-DNA integration. DSBs occurring for various reasons within the host genome are targets for integration of a double-stranded T-DNA molecule. Integration occurs mainly via the NHEJ and MMEJ pathways, both requiring some microhomology. Accurate repair via simple ligation or HR is very rare, and their characteristics were almost absent from our data. Before repair takes place, SDSA may create fillers of additional DNA sequence before one of the repair pathways leads to an intact double-stranded DNA. This may also explain the presence of inversions at the integration site, when two double-strand breaks occur within close genomic proximity and the fragment is reincorporated in the opposite orientation.

On both borders, extended fit of the border sequences by a few bases can be observed. This short extension of range is due to random sequence similarity of the genomic DNA to the border sequence that has a chance of roughly one in four to fit at the next position. In addition to this “extended fit,” some sequence of the binary vector neighboring T-DNA was also found to be integrated, which is consistent with published reports on several transformed plants (Smith, 1998). Transfer of vector sequences has also been observed in yeast, where sequences following the border sequence can be transported into the host cell due to border skipping or incomplete cutting of borders (Rolloos et al., 2014).

Plant Repair Pathway for T-DNA Integration

Our results strongly support the hypothesis that integration utilizes the plants’ repertoire of DSB repair pathways for integration. The main DSB repair pathways in *A. thaliana* are NHEJ and MMEJ supported by SDSA. The microhomologies, fillers, and sizes of deletions or target-site duplications detected at insertion sites match the typical characteristics of DSB repair. In general, both borders show similar characteristics concerning fillers, microhomologies, and positions of the border cuts, favoring a DSB repair model for integration and in contrast to models with a single-stranded

integration that would probably result in differences between the characteristics at both borders (Tzfira et al., 2004). A model for T-DNA integration incorporating these plant repair pathways is summarized in Figure 7.

Recently, two studies have been published (Mestiri et al., 2014; Park et al., 2015) indicating that T-DNA integration still works when key proteins of the NHEJ pathway are knocked out. The integration process may utilize available pathways, including alternative repair pathways such as MMEJ, which may play a larger role in NHEJ-deficient lines. We would expect that in those lines the integration structures show longer stretches of microhomology, a prediction supported by results from DSB repair outcomes in Ku-deficient CHO cells where larger microhomology was observed (Feldmann et al., 2000). We hypothesize that the removal of gene functions required for MMEJ, such as components of the MRX-complex consisting of Mre11, Rad50, and Xrs2(NBS1) (McVey and Lee, 2008), against a background that is already deficient in NHEJ, might lead to plant lines unsuitable for T-DNA integration.

The origin of the initial DSBs used for integration remains an open question. Randomly occurring breaks might be a possible source, but it seems unlikely that they are occurring to an extent

required for the random incorporation of single T-strands present somewhere in the nucleus. One option to solve this problem could be a mechanism that targets the T-DNA to DSBs that are, for example, created by DNA topoisomerases during transcription or replication. It is interesting to note that VirD2 has been shown to interact with yeast histones (Wolterink-van Loo et al., 2015). One may speculate that histone phosphorylation upon DSB detection (Waterworth et al., 2011) might serve as a signal to guide VirD2 and the T-strand to DSBs.

Alternatively, a function is required to mediate DSB formation in the vicinity of the T-strand. However, such functionality has not yet been assigned to one of the Vir proteins associated with T-DNA in plant cells. Given that we detected inversions that require two nearby DSBs, as well as large target-site duplications that require two nicks in opposite DNA strands at quite some distance, one may hypothesize that VirD2, potentially together with plant cell functions, triggers the introduction of DSBs in chromatin.

METHODS

Origin and Processing of Confirmation Sequences

For confirmation of insertion alleles, confirmation amplicons were generated using a primer specific for the T-DNA border from which the FST was derived and an insertion site or genome-specific primer (gsp) as described previously (Huep et al., 2014). After Sanger sequencing with a nested border-specific primer and the gsp, the resulting sequences were processed using PHRED and pregap4 from the Staden package (Staden et al., 2000) for quality trimming. Sequences for submission have been shortened to 1000 bp due to submission guidelines. The shortening was done by leaving the T-DNA end of the sequence intact while removing bases from the other end.

Determination of the Insertion Position from Confirmation Sequences

All analyses in this study were based on the TAIR10 data version for the *A. thaliana* genome. To derive a position for an insertion junction, a MEGABLAST versus the TAIR10 genome sequence was performed. To focus on reliable sequence data, we included the following filters: (1) T-DNA sequence needed to be present at the 5' region (for sequences sequenced with the T-DNA border primer) or at the 3' region (for sequences sequenced with gsp) of the confirmation sequence; (2) MEGABLAST hits in the genome required an e-value of $1e-50$ at most; if there were several sequences fulfilling these criteria for one junction of the insertion, sequences sequenced with the T-DNA specific primer were preferred, and if there were still multiple sequences, the one with the lowest e-value was chosen. The position of the start of the MEGABLAST hit was then determined as the insertion position of this junction of the insertion.

Determination of Deletion Sizes/Target-Site Duplication Sizes

To determine the structure of a given insertion site, we performed the confirmation PCR and amplicon sequencing on both sites of the expected T-DNA, i.e. at both the upstream and downstream junctions of the genome to T-DNA, even if there was only an FST for one of the junctions. Primers were designed independently for each junction using the GABI-Kat primer design tool (Huep et al., 2014). For insertions with unknown border identity at the second junction, which is the usual case for an insertion predicted from a single FST, we used LB- as well as RB-specific T-DNA primers for the junction without FST. The insertion position for the upstream and downstream junction was computed as described above. The size of the deletion was then calculated by the formula: $([\text{downstream_position}] -$

$\text{upstream_position}] - 1)$. If this value was negative, upstream and downstream genomic junction sequences were overlapping, indicating a target-site duplication.

Filtering of Confirmation Sequences for Junction Analysis

Confirmation sequences were filtered to make sure that only one and the highest-quality sequence was evaluated for a given junction. Only the outermost border (BLAST hit) was evaluated, and we required that this outermost border was present in the correct orientation. One confirmation sequence for each junction of an insertion site was chosen according to the following criteria: (1) presence of T-DNA and *A. thaliana* genomic sequence was required; (2) sequences sequenced with the gsp were preferred because they usually contain a larger part of T-DNA sequence, since the read often reaches into the sequence contributed by the border primer; (3) if there were still multiple sequences available, the one with the highest e-value for the hit in the *A. thaliana* genome was preferred. Of all 30 507 confirmation sequences, 11 802 were used for further analysis. Of these, 699 originated from the RB, 11 103 from the LB, and each single read represented one unique T-DNA/genome junction. For SALK lines, 880 sequences selected from 2469 confirmation sequences were analyzed, 852 for LB and 28 for RB.

Sequence Analyses of T-DNA Borders

The limits of T-DNA sequences present at a given insertion site, possibly along with sequences derived from the adjacent plasmid vector, were determined based on BLAST analyses of the sequences that were filtered as described above. First, the BLAST analyses were performed versus the complete binary vector sequence with an e-value cutoff of 10.0 and no filtering for low-complexity regions. The results were analyzed using a dedicated Perl-script along with the LIMS database used at GABI-Kat for storage of sequences and BLAST results (Kleinboelting et al., 2012). The script compared the position of the best BLAST hit with the expected cut position defined by the recognition sequence present in LB and RB. A positive value indicated an integration beyond that position. In some cases a part of the vector was also integrated. Zero indicated an exact cut of T-DNA precisely at the expected cut position (see Figure 1). A negative output value indicated that further parts of the border sequences were missing.

Determination of Microhomology at Integration Sites

To determine microhomology or fillers at a given integration site, we performed additional BLAST analyses (with the filtered sequences and with respect to the previously performed ones, see paragraph above) versus the *A. thaliana* TAIR10 genome sequence with an e-value cutoff of $5e-3$ and no filtering for low-complexity regions. If the BLAST hit versus the plant genome followed directly before/after the hit versus the vector, integration occurred exactly without use of microhomology and without creating fillers, giving zero as a result. If BLAST hits were overlapping, microhomology was detected and assigned a negative value in the amount of overlapping base pairs. A positive value was assigned if there was a gap between the vector hit and the plant genomic hit, indicating a filler. Although these junctions are classified as junctions with fillers (and without microhomology) in this analysis, it is possible that there is microhomology between the filler and T-DNA or plant genomic DNA. To identify microhomology of fillers, we had to first determine the origin of the fillers.

Examination of Fillers

All fillers of at least 15 bp (see above) were extracted from the respective confirmation sequence with an overlap of 20 bp to the T-DNA part and to the genomic part of the respective confirmation sequence (see Supplemental Information Files 9 for GK and Supplemental Information Files 10 for SALK data). Sequences shorter than 15 bp were not considered because they would not allow generation of meaningful BLAST results. BLAST runs versus sequence datasets from *A. thaliana*, *A. tumefaciens*, chloroplast, mitochondria, T-DNA from LB to RB, and vector backbone (without T-DNA) were performed with an e-value cutoff

of 10.0 to identify the origin of these fillers. To exclude hits mainly based on the overlap of 20 bp, hits covering a larger part of the original filler were preferred. If these criteria did not result in a unique hit, the one with the best score among them was chosen. In about 22% of cases several “best” BLAST hits were obtained, covering the same length of filler-query with identical scores. Initially we excluded these cases. As a result, 1162 sequences were left (389 hits on the same chromosome, 756 on other chromosomes, 15 on the *A. tumefaciens* genome, and 2 on the mitochondrial genome). Unfortunately, many hits within T-DNA or vector were filtered by excluding these ambiguous cases, mostly due to sequence identity of a part of T-DNA with the vector backbone (T-DNA allows plasmid rescue). We decided to assign ambiguous cases to the most probable origin in the order of the ranking observed for almost 80% of all cases. First, hit on the same chromosome, if in doubt the one closest to the insertion position; second, hit on a chromosome other than the one that contained the studied filler; third, hit within T-DNA from LB to RB; fourth, hit on the vector backbone; fifth, hit on the *A. tumefaciens* genome (circular, linear, plasmid); sixth, hit on chloroplast or mitochondrial DNA sequence. This measure did not change the overall trend of the results, but avoided loss of hits to T-DNA or vector backbone. If the hit was on the same chromosome, it was used to compute the distance from the source position of the filler to the insertion position. Microhomology or filler between the filler and the neighboring sequences was then determined by evaluating the range of the BLAST hit: if it reached into the 20 bp overlap, microhomology was present.

Sequencing of Line GK_144F03

Genomic DNA (820 ng) prepared from a pool of 50 T3 offspring seedlings homozygous for the insertion allele affecting the T-DNA insertion locus close to *At3g48060* was used for preparing paired-end libraries according to the Illumina TruSeqv2 Sample Prep Kit. Homozygosity was proved for the parental T2 plant as well as for the pooled T3 plants using primer combinations as described by Bolle et al. (2013). Sequencing was performed on an Illumina MiSeq sequencing instrument with 2 × 250-nt paired-end reads using MiSeqv2 chemistry. Reads were trimmed by Trimmomatic version 0.27 (Bolger et al., 2014) and mapped to the *A. thaliana* TAIR10 genome using BWA (Li and Durbin, 2009) with default parameters. The data were visualized using GenomeView (Abeel et al., 2012).

Histochemical Assays and Microscopy for *tt* Mutant Analyses

Microscopy, binocular analysis, and image acquisition, as well as diphenyl boric acid 2-aminoethylester (DPBA) staining, *p*-dimethylaminocinnamaldehyde (DMACA) staining, and norflurazon treatment of plant material, was performed as described previously (Appelhaugen et al., 2014).

ACCESSION NUMBERS

All confirmation sequences have been submitted to ENA/GenBank, with accession numbers LN484267 to LN515397. Illumina-read data from sequencing of GK_144F03 genomic DNA were deposited at the NCBI Short Read Archive (SRA; <http://www.ncbi.nlm.nih.gov/sra/>) under the accession number SRX879613.

SUPPLEMENTAL INFORMATION

Supplemental Information is available at *Molecular Plant Online*.

FUNDING

The work described in this article was funded by the German Federal Ministry of Education and Research (BMBF) in the context of the German plant genomics program GABI (Förderkennezeichen 0313855).

AUTHOR CONTRIBUTIONS

N.K., G.H., and B.W. conceived the research approach, interpreted the data, and wrote the manuscript. N.K., G.H., Y.L., and B.W. analyzed the data. G.H. conducted and supervised the wet-lab experiments. P.V. designed and performed all Sanger and next-generation sequencing. N.K. carried out database programming and bioinformatics analyses. B.W. de-

signed the research outline. I.A. identified and characterized the GK_478B05/*tt6* mutant and contributed to writing the manuscript. All authors read and approved the manuscript.

ACKNOWLEDGMENTS

The authors thank Mario Rosso, the MPI for Plant Breeding Research, and all former co-workers in the project for their contribution to GABI-Kat, Ute Buerstenbinder, Renate Harder, Eliane Quittschau, Helene Schellenberg, Nina Schmidt, Andrea Voigt, and Marja-Leena Wilke for technical assistance, and Malte Mattheis for database programming to identify inversions. We acknowledge the financial support of the German Research Foundation (DFG) and the Open Access Publication Fund of Bielefeld University for the article processing charge. The authors declare that they have no competing interests.

Received: April 10, 2015

Revised: July 28, 2015

Accepted: August 13, 2015

Published: September 4, 2015

REFERENCES

- Abeel, T., Van Parys, T., Saeys, Y., Galagan, J., and Van de Peer, Y. (2012). GenomeView: a next-generation genome browser. *Nucleic Acids Res.* **40**:e12.
- Alonso, J.M., Stepanova, A.N., Leisse, T.J., Kim, C.J., Chen, H., Shinn, P., Stevenson, D.K., Zimmerman, J., Barajas, P., Cheuk, R., et al. (2003). Genome-wide insertional mutagenesis of *Arabidopsis thaliana*. *Science* **301**:653–657.
- Appelhaugen, I., Thiedig, K., Nordholt, N., Schmidt, N., Huep, G., Sagasser, M., and Weisshaar, B. (2014). Update on *transparent testa* mutants from *Arabidopsis thaliana*: characterisation of new alleles from an isogenic collection. *Planta* **240**:955–970.
- Bolger, M.E., Weisshaar, B., Scholz, U., Stein, N., Usadel, B., and Mayer, K.F. (2014). Plant genome sequencing—applications for crop improvement. *Curr. Opin. Biotechnol.* **26**:31–37.
- Bolle, C., Huep, G., Kleinbolting, N., Haberer, G., Mayer, K., Leister, D., and Weisshaar, B. (2013). GABI-DUPLO: a collection of double mutants to overcome genetic redundancy in *Arabidopsis thaliana*. *Plant J.* **75**:157–171.
- Britt, A.B., and May, G.D. (2003). Re-engineering plant gene targeting. *Trends Plant Sci.* **8**:90–95.
- Brunaud, V., Balergue, S., Dubreucq, B., Aubourg, S., Samson, F., Chauvin, S., Bechtold, N., Cruaud, C., DeRose, R., Pelletier, G., et al. (2002). T-DNA integration into the *Arabidopsis* genome depends on sequences of pre-insertion sites. *EMBO Rep.* **3**:1152–1157.
- Chilton, M.D., and Que, Q. (2003). Targeted integration of T-DNA into the tobacco genome at double-stranded breaks: new insights on the mechanism of T-DNA integration. *Plant Physiol.* **133**:956–965.
- Clark, K.A., and Krysan, P.J. (2010). Chromosomal translocations are a common phenomenon in *Arabidopsis thaliana* T-DNA insertion lines. *Plant J.* **64**:990–1001.
- De Buck, S., Jacobs, A., Van Montagu, M., and Depicker, A. (1999). The DNA sequences of T-DNA junctions suggest that complex T-DNA loci are formed by a recombination process resembling T-DNA integration. *Plant J.* **20**:295–304.
- De Neve, M., De Buck, S., Jacobs, A., Van Montagu, M., and Depicker, A. (1997). T-DNA integration patterns in co-transformed plant cells suggest that T-DNA repeats originate from co-integration of separate T-DNAs. *Plant J.* **11**:15–29.
- Dürrenberger, F., Cramer, A., Hohn, B., and Koukolíková-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.
- Feldmann, E., Schmiemann, V., Goedecke, W., Reichenberger, S., and Pfeiffer, P.** (2000). DNA double-strand break repair in cell-free extracts from Ku80-deficient cells: implications for Ku serving as an alignment factor in non-homologous DNA end joining. *Nucleic Acids Res.* **28**:2585–2596.
- Forsbach, A., Schubert, D., Lechtenberg, B., Gils, M., and Schmidt, R.** (2003). A comprehensive characterization of single-copy T-DNA insertions in the *Arabidopsis thaliana* genome. *Plant Mol. Biol.* **52**:161–176.
- Gelvin, S.B.** (1998). The introduction and expression of transgenes in plants. *Curr. Opin. Biotechnol.* **9**:227–232.
- 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.
- Gelvin, S.B.** (2010). Finding a way to the nucleus. *Curr. Opin. Microbiol.* **13**:53–58.
- Gheysen, G., Villarroel, R., and Van Montagu, M.** (1991). Illegitimate recombination in plants: a model for T-DNA integration. *Genes Dev.* **5**:287–297.
- Gorbunova, V., and Levy, A.A.** (1997). Non-homologous DNA end joining in plant cells is associated with deletions and filler DNA insertions. *Nucleic Acids Res.* **25**:4650–4657.
- Gorbunova, V.V., and Levy, A.A.** (1999). How plants make ends meet: DNA double-strand break repair. *Trends Plant Sci.* **4**:263–269.
- Hoekema, A., Hirsch, P.R., Hooykaas, P.J.J., and Schilperoort, R.A.** (1983). A binary plant vector strategy based on separation of vir- and T-region of the *Agrobacterium tumefaciens* Ti-plasmid. *Nature* **303**:179–180.
- Huep, G., Kleinboelting, N., and Weisshaar, B.** (2014). An easy-to-use primer design tool to address paralogous loci and T-DNA insertion sites in the genome of *Arabidopsis thaliana*. *Plant Methods* **10**:28.
- Jasper, F., Koncz, C., Schell, J., and Steinbiss, H.H.** (1994). *Agrobacterium* T-strand production in vitro: sequence-specific cleavage and 5' protection of single-stranded DNA templates by purified VirD2 protein. *Proc. Natl. Acad. Sci. USA* **91**:694–698.
- Kim, S., Veena, and Gelvin, S.** (2007). Genome-wide analysis of *Agrobacterium* T-DNA integration sites in the *Arabidopsis* genome generated under non-selective conditions. *Plant J.* **51**:779–791.
- Kleinboelting, N., Huep, G., Kloetgen, A., Viehoveer, P., and Weisshaar, B.** (2012). GABI-Kat SimpleSearch: new features of the *Arabidopsis thaliana* T-DNA mutant database. *Nucleic Acids Res.* **40**:D1211–D1215.
- Krzkova, L., and Hroudá, M.** (1998). Direct repeats of T-DNA integrated in tobacco chromosome: characterization of junction regions. *Plant J.* **16**:673–680.
- Lacroix, B., Vaidya, M., Tzfira, T., and Citovsky, V.** (2005). The VirE3 protein of *Agrobacterium* mimics a host cell function required for plant genetic transformation. *EMBO J.* **24**:428–437.
- Li, H., and Durbin, R.** (2009). Fast and accurate short read alignment with Burrows-Wheeler transform. *Bioinformatics* **25**:1754–1760.
- Li, Y., Rosso, M.G., Ulker, B., and Weisshaar, B.** (2006). Analysis of T-DNA insertion site distribution patterns in *Arabidopsis thaliana* reveals special features of genes without insertions. *Genomics* **87**:645–652.
- Liang, Z., and Tzfira, T.** (2013). In vivo formation of double-stranded T-DNA molecules by T-strand priming. *Nat. Commun.* **4**:2253.
- Magori, S., and Citovsky, V.** (2011). *Agrobacterium* counteracts host-induced degradation of its effector F-box protein. *Sci. Signaling* **4**:ra69.
- Mayerhofer, R., Koncz-Kalman, Z., Nawrath, C., Bakkeren, G., Cramer, A., Angelis, K., Redei, G.P., Schell, J., Hohn, B., and Koncz, C.** (1991). T-DNA integration: a mode of illegitimate recombination in plants. *EMBO J.* **10**:697–704.
- McVey, M., and Lee, S.E.** (2008). MMEJ repair of double-strand breaks (director's cut): deleted sequences and alternative endings. *Trends Genet.* **24**:529–538.
- Mestiri, I., Norre, F., Gallego, M.E., and White, C.I.** (2014). Multiple host-cell recombination pathways act in *Agrobacterium*-mediated transformation of plant cells. *Plant J.* **77**:511–520.
- O'Malley, R.C., and Ecker, J.R.** (2010). Linking genotype to phenotype using the *Arabidopsis* unimutant collection. *Plant J.* **61**:928–940.
- Park, S.Y., Vaghchhipawala, Z., Vasudevan, B., Lee, L.Y., Shen, Y., Singer, K., Waterworth, W.M., Zhang, Z.J., West, C.E., Mysore, K.S., et al.** (2015). *Agrobacterium* T-DNA integration into the plant genome can occur without the activity of key non-homologous end-joining proteins. *Plant J.* **81**:934–946.
- Peralta, E.G., and Ream, L.W.** (1985). T-DNA border sequences required for crown gall tumorigenesis. *Proc. Natl. Acad. Sci. USA* **82**:5112–5116.
- Pitzschke, A., and Hirt, H.** (2010). New insights into an old story: *Agrobacterium*-induced tumour formation in plants by plant transformation. *EMBO J.* **29**:1021–1032.
- Que, Q., Elumalai, S., Li, X., Zhong, H., Nalapalli, S., Schweiner, M., Fei, X., Nuccio, M., Kelliher, T., Gu, W., et al.** (2014). Maize transformation technology development for commercial event generation. *Front. Plant Sci.* **5**:379.
- Ray, A., and Langer, M.** (2002). Homologous recombination: ends as the means. *Trends Plant Sci.* **7**:435–440.
- Rolloos, M., Dohmen, M.H., Hooykaas, P.J., and van der Zaal, B.J.** (2014). Involvement of Rad52 in T-DNA circle formation during *Agrobacterium tumefaciens*-mediated transformation of *Saccharomyces cerevisiae*. *Mol. Microbiol.* **91**:1240–1251.
- Salomon, S., and Puchta, H.** (1998). Capture of genomic and T-DNA sequences during double-strand break repair in somatic plant cells. *EMBO J.* **17**:6086–6095.
- 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.
- Schell, J., Van Montagu, M., De Beuckeleer, M., De Block, M., Depicker, A., De Wilde, M., Engler, G., Genetello, C., Hernalsteens, J.P., Holsters, M., et al.** (1979). Interactions and DNA transfer between *Agrobacterium tumefaciens*, the Ti-plasmid and the plant host. *Proc. R. Soc. Lond.* **204**:251–266.
- Scholl, R.L., May, S.T., and Ware, D.H.** (2000). Seed and molecular resources for *Arabidopsis*. *Plant Physiol.* **124**:1477–1480.
- Singer, K., Shibolet, Y.M., Li, J., and Tzfira, T.** (2012). Formation of complex extrachromosomal T-DNA structures in *Agrobacterium tumefaciens*-infected plants. *Plant Physiol.* **160**:511–522.
- Smith, N.** (1998). More T-DNA than meets the eye. *Trends Plant Sci.* **3**:85.
- Staden, R., Beal, K.F., and Bonfield, J.K.** (2000). The Staden package, 1998. *Methods Mol. Biol.* **132**:115–130.
- Strizhov, N., Li, Y., Rosso, M.G., Viehoveer, P., Dekker, K.A., and Weisshaar, B.** (2003). High-throughput generation of sequence indexes from T-DNA mutagenized *Arabidopsis thaliana* lines. *BioTechniques* **35**:1164–1168.
- Szabados, L., Kovacs, I., Oberschall, A., Abraham, E., Kerekes, I., Zsigmond, L., Nagy, R., Alvarado, M., Krasovskaja, I., Gal, M., et al.** (2002). Distribution of 1000 sequenced T-DNA tags in the *Arabidopsis* genome. *Plant J.* **32**:233–242.

- Thole, V., Alves, S.C., Worland, B., Bevan, M.W., and Vain, P.** (2009). A protocol for efficiently retrieving and characterizing flanking sequence tags (FSTs) in *Brachypodium distachyon* T-DNA insertional mutants. *Nat. Protoc.* **4**:650–661.
- Tinland, B., Hohn, B., and Puchta, H.** (1994). *Agrobacterium tumefaciens* transfers single-stranded transferred DNA (T-DNA) into the plant cell nucleus. *Proc. Natl. Acad. Sci. USA* **91**:8000–8004.
- Tzfira, T., Frankman, L.R., Vaidya, M., and Citovsky, V.** (2003). Site-specific integration of *Agrobacterium tumefaciens* T-DNA via double-stranded intermediates. *Plant Physiol.* **133**:1011–1023.
- Tzfira, T., Li, J., Lacroix, B., and Citovsky, V.** (2004). *Agrobacterium* T-DNA integration: molecules and models. *Trends Genet.* **20**:375–383.
- Ulker, B., Li, Y., Rosso, M.G., Logemann, E., Somssich, I.E., and Weisshaar, B.** (2008). T-DNA-mediated transfer of *Agrobacterium tumefaciens* chromosomal DNA into plants. *Nat. Biotechnol.* **26**:1015–1017.
- Van Larebeke, N., Engler, G., Holsters, M., Van den Elsacker, S., Zaenen, I., Schilperoort, R.A., and Schell, J.** (1974). Large plasmid in *Agrobacterium tumefaciens* essential for crown gall-inducing ability. *Nature* **252**:169–170.
- Wang, K., Herrera-Estrella, L., Van Montagu, M., and Zambryski, P.** (1984). Right 25 bp terminus sequence of the nopaline T-DNA is essential for and determines direction of DNA transfer from *Agrobacterium* to the plant genome. *Cell* **38**:455–462.
- Waterworth, W.M., Drury, G.E., Bray, C.M., and West, C.E.** (2011). Repairing breaks in the plant genome: the importance of keeping it together. *New Phytol.* **192**:805–822.
- Weaver, D.T.** (1995). What to do at an end: DNA double-strand-break repair. *Trends Genet.* **11**:388–392.
- Windels, P., De Buck, S., Van Bockstaele, E., De Loose, M., and Depicker, A.** (2003). T-DNA integration in *Arabidopsis* chromosomes. Presence and origin of filler DNA sequences. *Plant Physiol.* **133**:2061–2068.
- Wolterink-van Loo, S., Escamilla Ayala, A.A., Hooykaas, P.J., and van Heusden, G.P.** (2015). Interaction of the *Agrobacterium tumefaciens* virulence protein VirD2 with histones. *Microbiology* **161**:401–410.
- Yoshiyama, K.O., Sakaguchi, K., and Kimura, S.** (2013). DNA damage response in plants: conserved and variable response compared to animals. *Biology* **2**:1338–1356.
- Zambryski, P., Holsters, M., Kruger, K., Depicker, A., Schell, J., Van Montagu, M., and Goodman, H.** (1980). Tumor DNA structure in plant cells transformed by *A. tumefaciens*. *Science* **209**:1385–1391.