ARTICLES
https://doi.org/10.1038/s41477-022-01147-5

nature plants

Distinct mechanisms for genomic attachment of the 5′ and 3′ ends of *Agrobacterium* T-DNA in plants

Lejon E. M. Kralemann[1], Sylvia de Pater[1], Hexi Shen[2], Susan L. Kloet[3], Robin van Schendel[3], Paul J. J. Hooykaas[1] and Marcel Tijsterman[1,3]✉

***Agrobacterium tumefaciens*, a pathogenic bacterium capable of transforming plants through horizontal gene transfer, is nowadays the preferred vector for plant genetic engineering. The vehicle for transfer is the T-strand, a single-stranded DNA molecule bound by the bacterial protein VirD2, which guides the T-DNA into the plant's nucleus where it integrates. How VirD2 is removed from T-DNA, and which mechanism acts to attach the liberated end to the plant genome is currently unknown. Here, using newly developed technology that yields hundreds of T-DNA integrations in somatic tissue of *Arabidopsis thaliana*, we uncover two redundant mechanisms for the genomic capture of the T-DNA 5′ end. Different from capture of the 3′ end of the T-DNA, which is the exclusive action of polymerase theta-mediated end joining (TMEJ), 5′ attachment is accomplished either by TMEJ or by canonical non-homologous end joining (cNHEJ). We further find that TMEJ needs MRE11, whereas cNHEJ requires TDP2 to remove the 5′ end-blocking protein VirD2. As a consequence, T-DNA integration is severely impaired in plants deficient for both MRE11 and TDP2 (or other cNHEJ factors). In support of MRE11 and cNHEJ specifically acting on the 5′ end, we demonstrate rescue of the integration defect of double-deficient plants by using T-DNAs that are capable of forming telomeres upon 3′ capture. Our study provides a mechanistic model for how *Agrobacterium* exploits the plant's own DNA repair machineries to transform it.**

Agrobacterium tumefaciens-mediated transformation (AMT) is the most widely used method for generating transgenic plants. In nature, this soil bacterium transforms dicotyledonous plants by translocating part of its DNA, the transferred (T)-DNA, into plant cells, where it integrates into the plant's genome[1]. Subsequent expression of the *Agrobacterium* genes causes crown gall disease. Within *Agrobacterium* the T-DNA is located on a tumour-inducing (Ti) plasmid flanked by a repeated sequence of 25 bp, the left and right border repeats (LB and RB). These sequences are recognition sites for the virulence proteins VirD1 and VirD2, which generate single-strand DNA (ssDNA) breaks required to liberate T-DNA as a ssDNA molecule, the T-strand[2]. The VirD2 protein remains covalently bound to the 5′ end of the T-strand[3,4] and pilots it into the plant cell through the type 4 secretion system[5] that is created by the *Agrobacterium* virulence programme upon detection of wounded plant cells[6]. The T-DNA is subsequently imported into the nucleus[7] where it integrates at a random position in the genome[8]. The molecular mechanism by which the T-DNA is integrated into the plant genome remained enigmatic until recently when it was found for *Arabidopsis thaliana* that this process depends critically on polymerase theta (Pol θ)[9], a host protein that acts in the repair of DNA double-strand breaks (DSBs) via end joining. Abundant genetic and biochemical research performed over the past few years has established that Pol θ facilitates repair of DSBs in a multitude of species by using (few) complementary bases in 3′ protruding ssDNA break ends to carry out DNA extension on one break end using the other end as a template[10,11]. This biochemical property of the enzyme combined with occasionally occurring primer–template switching provides an explanation for two characteristic features that are observed at sites where Pol θ-mediated end joining (TMEJ) of genomic DSBs takes place, that is microhomology and so-called templated insertions. These features are also prevalent at the junctions of T-DNA integration sites[9,12]—in plant transgenesis, templated insertions have also been described as 'filler' sequences[13,14]. However, although TMEJ presents a logical model for connecting the 3′ end of a T-DNA to a potentially resected genomic break (Fig. 1a), the biochemistry of the capture of its 5′ end has not yet been elucidated. It is currently also unknown how plant cells remove the covalently attached VirD2 from 5′ T-DNA ends to allow integration.

Results

The TRANSGUIDE method. To study the capture of T-DNA by the *Arabidopsis* genome, and in particular attachment of the RB end, we developed a next-generation sequencing-based method that we termed TRANSGUIDE (T-DNA random integration site genome-wide unbiased identification). This method allows us to identify hundreds of T-DNA–genome junctions (both LB and RB) in pools of transformed cells. For this study we chose to collect *Arabidopsis* root-derived callus samples 3 weeks after growth under selection for T-DNA presence (Fig. 1b). We employed custom-made software to filter for high-quality, reliable outcomes and annotate individual T-DNA integration junctions with respect to potentially relevant features, such as genomic position, loss of T-DNA sequences, degree of microhomology and absence or presence of filler DNA. The outcome of this pipeline reliably represents in vivo biology; we used PCR and Sanger sequencing on the input material and could validate 23 of 24 junction sequences (Supplementary Data 1).

[1]Institute of Biology Leiden, Leiden University, Leiden, The Netherlands. [2]School of Municipal and Environmental Engineering, Shandong Jianzhu University, Jinan, Shandong, China. [3]Department of Human Genetics, Leiden University Medical Center, Leiden, The Netherlands. ✉e-mail: M.Tijsterman@lumc.nl

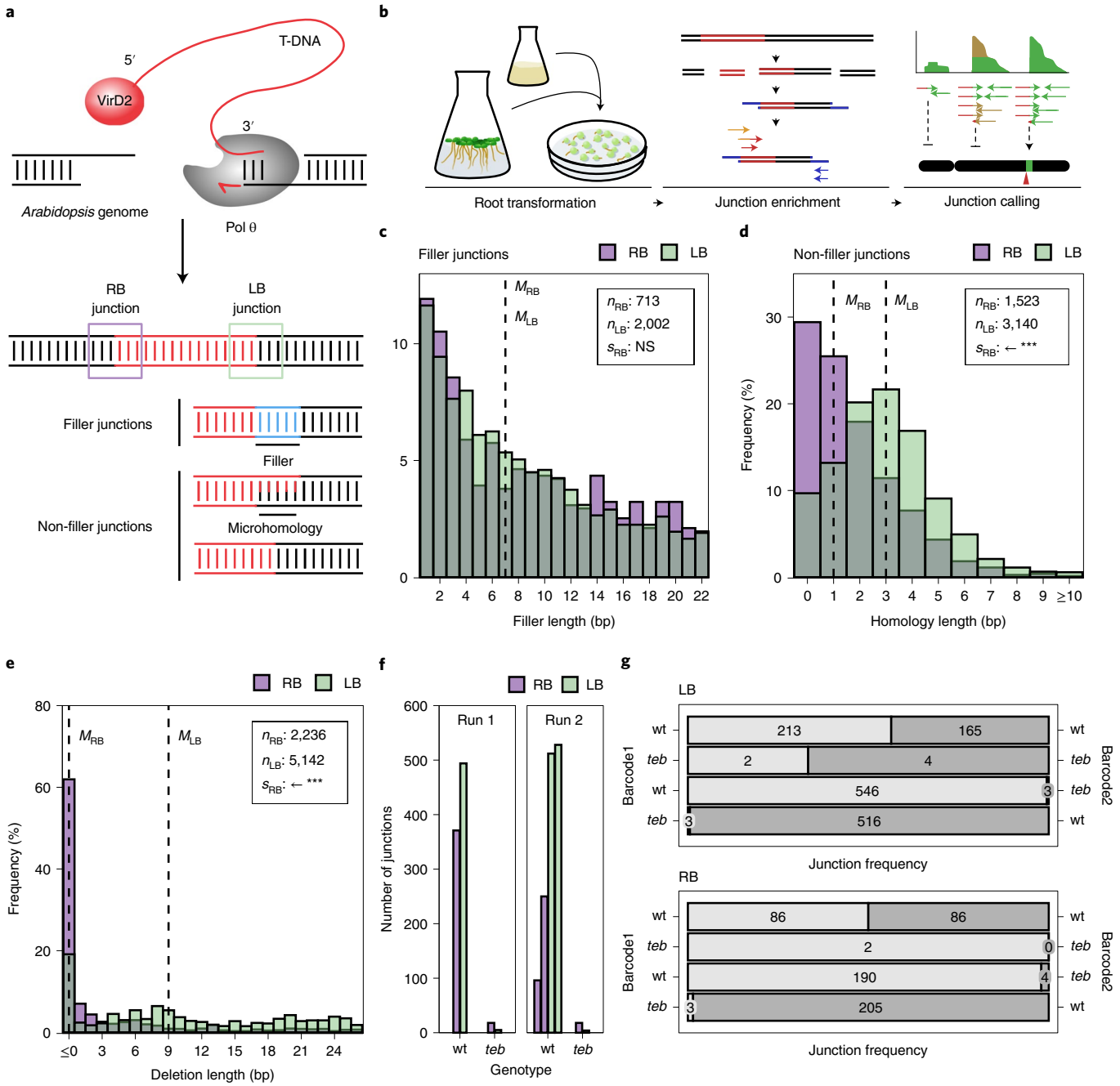


Fig. 1 | TRANSGUIDE reveals different characteristics for LB and RB T-DNA junctions. a, Tentative model for genomic capture of T-DNA via Pol θ action. **b**, Schematic overview of TRANSGUIDE (see Methods for details). **c**–**e**, Overlapping histograms based on combined data from 11 wild-type (wt) samples (transformed with pUBC, pCAS9 or pWY82) showing the filler-length distribution at junctions with fillers (**c**), the extent of microhomology at junctions that do not contain fillers (**d**) and the amount of DNA lost from the T-DNA end at both filler and non-filler junctions (**e**). LB junction data are in light green, RB in purple and the overlap between LB and RB is indicated in olive green. Medians (*M*, dashed lines), number of observations (*n*) and shifts in the RB distribution relative to LB (*s*) are indicated. Wilcoxon rank-sum tests were performed to find the direction and the significance of the shifts ($P_{filler} = 2 \times 10^{-1}$, $P_{homology} = 6 \times 10^{-106}$, $P_{deletion} = 0$). The arrow indicates the direction of the shift. **f**, Comparison of the number of T-DNA–genome junctions in wt ($n = 3$) and *teb* ($n = 2$) samples; each sample contains 20 calli. **g**, Number of junctions after competitive TRANSGUIDE, a variant in which equimolar amounts of genomic DNA of two samples with differently barcoded T-DNA (barcode 1 in light grey, and barcode 2 in dark grey) were combined before junction enrichment. NS, $P \geq 0.05$; *, $P < 0.05$; **, $P < 0.01$; ***, $P < 0.001$.

Unequal Pol θ involvement at RB and LB junctions. Using the aforementioned technology, we obtained a collection consisting of ~2,200 RB–genome junctions and ~5,100 LB–genome junctions upon transformation of the Col-0 ecotype (Supplementary Data 2). Consistent with earlier findings[15], these junctions are scattered across the entire genome, with the exception of the pericentromeric regions (Extended Data Fig. 1). Arguing for a prominent role for Pol θ in integration, we found (very similar) filler DNAs to be abundantly present at both RB–genome and LB–genome junctions (Fig. 1c); however, the percentages were not identical with 31%

fillers at RB–genome junctions versus 39% at LB–genome junctions ($P=6\times10^{-5}$). Also, the degree of junctional microhomology (the median being 1 bp for RB–genome versus 3 bp for LB–genome junctions) and loss of terminal nucleotides (from the T-DNA) were different between RB and LB (Fig. 1d,e and Extended Data Figs. 2 and 3). Because microhomology usage and filler formation are hallmarks of Pol θ activity, these data suggest unequal involvement of this enzyme in the genomic attachment of the two T-DNA ends.

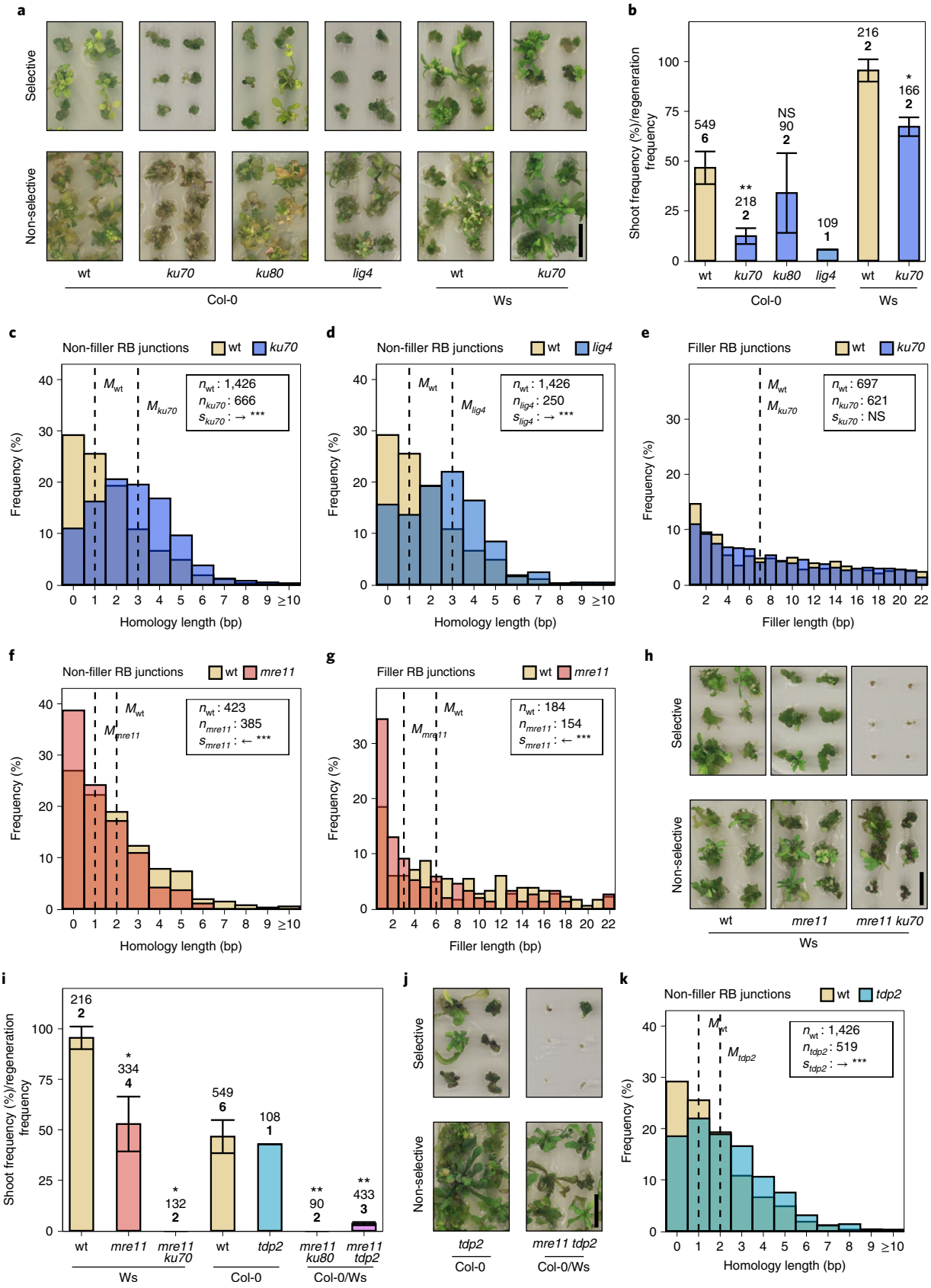
Pol θ deficiency suppresses T-DNA capture. We previously found that *Arabidopsis* plants deficient for Pol θ (*teb* mutants) are completely recalcitrant to AMT, arguing for an essential role for Pol θ in genomic capture of T-DNA. This conclusion is further substantiated here by demonstrating an almost complete absence of T-DNA–genome junctions in DNA isolated from root-transformed Pol θ-deficient plants; instead of finding a few hundred T-DNA integrations, we obtained only a few cases in *teb* calli (Fig. 1f). To exclude potential methodological distortions, for example resulting from PCR steps within TRANSGUIDE, we also performed AMT competition experiments. We mixed equal amounts of DNA from wild-type and *teb* that were transformed with almost identical yet barcoded T-DNA constructs and attributed T-DNA junctions to the appropriate genotype afterwards. These internally controlled experiments corroborate our finding that genomic T-DNA capture is Pol θ dependent (Fig. 1g and Supplementary Data 3). Of note, although the almost complete absence of T-DNA junctions in *teb* material unequivocally demonstrates that TRANSGUIDE outcomes for wild-type plants represent bona fide biology, we cannot conclude that the residual T-DNA–genome junctions found in *teb* samples represent completed T-DNA integration, as opposed to, for example, one-sided capture, in vivo recombination or PCR artefacts. Interestingly, however, and in agreement with a recent report[16], we find the molecules representing genomic capture in *teb* to be almost exclusively RB-to-genome junctions (Fig. 1f). Together with the notion of a reduced signature of Pol θ activity at RB–genome junctions in Pol θ-proficient plants, compared with LB–genome junctions, this result may point to another, redundant, molecular mechanism capable of attaching the 5′ end of the T-DNA to the plant genome.

RB capture can occur via TMEJ and canonical non-homologous end joining. The obvious candidate for end-joining activity other than TMEJ is canonical non-homologous end joining (cNHEJ), another pathway to repair genomic DNA breaks. Previous analysis of AMT in cNHEJ-deficient *Arabidopsis* led to conflicting results: whereas some laboratories reported reduced T-DNA integration[17–20], others found no effects[21–23] or even elevated frequencies[23,24]. We investigated a potential involvement of cNHEJ in T-DNA capture by monitoring shoot development and performing TRANSGUIDE upon root transformation of cNHEJ-deficient *ku70* and *lig4 Arabidopsis* mutants. We observed a reduced number of shoots in cNHEJ-deficient plants (Fig. 2a,b), arguing that cNHEJ affects stable transformation but is not essential. TRANSGUIDE of calli subsequently revealed a profound effect on the composition of T-DNA–genome junctions, specifically at the RB side (Fig. 2c–e); whereas LB–genome junctions found in *ku70* and *lig4* mutant roots are indistinguishable from those found in wild-type, RB–genome junctions isolated from cNHEJ mutant plants were characterized by an increased degree of microhomology (median of 3 bp in *ku70* and *lig4*, versus 1 bp in wild-type). In fact, when plotted for the degree of microhomology, the distribution of RB–genome junctions in cNHEJ mutant conditions is similar to that of the LB–genome junction, in both cNHEJ-deficient and -proficient contexts (Extended Data Fig. 4). This increased usage of microhomology is accompanied by increased loss of T-DNA sequence at the RB end, as well as an increased percentage of junctions containing fillers ($P_{ku70}=1\times10^{-3}$, $P_{lig4}=4\times10^{-3}$, Extended Data Fig. 4), which were of similar length to those observed in wild-type (Fig. 2e). We conclude that capture of the T-DNA 3′ end depends critically on intrinsically mutagenic TMEJ, whereas the 5′ end can be attached to the genome via two redundant activities, TMEJ and cNHEJ.

MRE11 is required for TMEJ-mediated RB capture. The identification of two end-joining pathways capable of attaching the T-DNA 5′ end to the plant genome raises the question of which enzymatic activity removes the bacterial VirD2 protein covalently bound to the outermost 5′ nucleotide of T-DNA? Although the sequence of events leading to completed T-DNA integration is unknown, one can envisage a scenario in which Pol θ-mediated genomic capture of the T-DNA 3′ end leads, simply by DNA synthesis using the T-DNA as a template, to conversion of the single-stranded T-DNA into double-stranded DNA (dsDNA) (Fig. 1a). The resulting structure would have a striking resemblance to DSB ends that are produced during meiotic recombination (by SPO11), or follow from some types of chemotherapy (TOP2 poisons), both of which have proteins covalently attached to their 5′ termini[25,26]. Removal of these end-blocking proteins is a prerequisite for DSB repair and one demonstrated mechanism of their removal involves MRE11-catalysed nicking of the protein-linked strand a short distance from the DSB terminus[27]. Arabidopsis *MRE11* null mutant plants are sterile, hampering their analysis[28]; however, an *mre11* hypomorphic allele (*mre11-2*) exists, which in a homozygous state confers sensitivity towards DNA-damaging agents yet supports plant development[29]. We inspected T-DNA integration in this mutant background and found the RB–genome junction spectrum altered, although inversely to what was observed in cNHEJ mutants. Instead of a more profound TMEJ signature we observed a clear depletion of TMEJ hallmarks in *mre11-2*, namely less microhomology at the junctions and reduced filler size (Fig. 2f,g). We conclude that MRE11 functionality is needed for Pol θ-mediated capture of

Fig. 2 | TMEJ and cNHEJ function redundantly in genomic capture of T-DNA's 5′ end. a, Calli of various genotypes with two different genetic backgrounds (Col-0 and Ws) after transformation with pCAMBIA3301, grown on either selective (+ phosphinothricin) or non-selective medium. **b**, Average percentage of calli with shoots, corrected for potential regeneration defects, after transformation with pCAMBIA3301. Indicated are the total number of scored calli (italic) and the number of experiments (bold) over which the mean (coloured bars) and s.e.m. (error bars) were calculated. **c**–**g**, Frequency of RB junctions with the indicated degree of microhomology (**c**,**d**,**f**) or filler presence (**e**,**g**) for wt (yellow) or mutant (shades of blue for cNHEJ mutants *ku70* and *lig4*, and light red for *mre11*) junctions. Medians (*M*, dashed lines), number of observations (*n*) and shifts in the mutant distribution relative to wt (*s*) are indicated. **h**, Calli after transformation with pCAS9. **i**, Regeneration-corrected shoot formation as in **b**, but for different mutants. **j**, Calli of *tdp2* and *mre11 tdp2* mutants after transformation with pCAMBIA3301. **k**, Frequency of RB junctions with the indicated degree of microhomology for wt (yellow) or *tdp2* mutant (cyan). The overlap is indicated in turquoise. For **b** and **i**, one-sided Student's *t*-tests were performed to test for significant reductions in T-DNA integration efficiency of mutants compared with wt ($P_{ku70c}=5\times10^{-3}$, $P_{ku80}=3\times10^{-1}$, $P_{ku70w}=3\times10^{-2}$, $P_{mre11}=2\times10^{-2}$, $P_{mre11ku70}=2\times10^{-2}$, $P_{mre11ku80}=1\times10^{-3}$, $P_{mre11tdp2}=2\times10^{-3}$). Mutants were compared with the wt of the same genetic background, except for hybrids, which were compared with the Col-0 wt. For **c**–**g** and **k**, Wilcoxon rank-sum tests were performed to find the direction (one-sided tests) and significance level (two-sided tests) of the shifts in homology and filler distributions ($P_{homology\_ku70}=7\times10^{-38}$, $P_{homology\_lig4}=2\times10^{-14}$, $P_{filler\_ku70}=4\times10^{-1}$, $P_{homology\_mre11}=2\times10^{-6}$, $P_{filler\_mre11}=8\times10^{-6}$, $P_{homology\_tdp2}=7\times10^{-11}$). NS, $P\geq0.05$; *, $P<0.05$; **, $P<0.01$, ***, $P<0.001$. Scale bars, 1 cm.

the T-DNA 5′ end—when impaired, only cNHEJ can perform this function. Interestingly, we find a wild-type profile for LB–genome junctions in *mre11-2* mutant plants (Extended Data Fig. 5), which could either mean that MRE11 is not needed to process genomic breaks for capturing T-DNA, or that the hypomorphic *mre11-2* allele encodes a protein still capable of this activity. One prediction



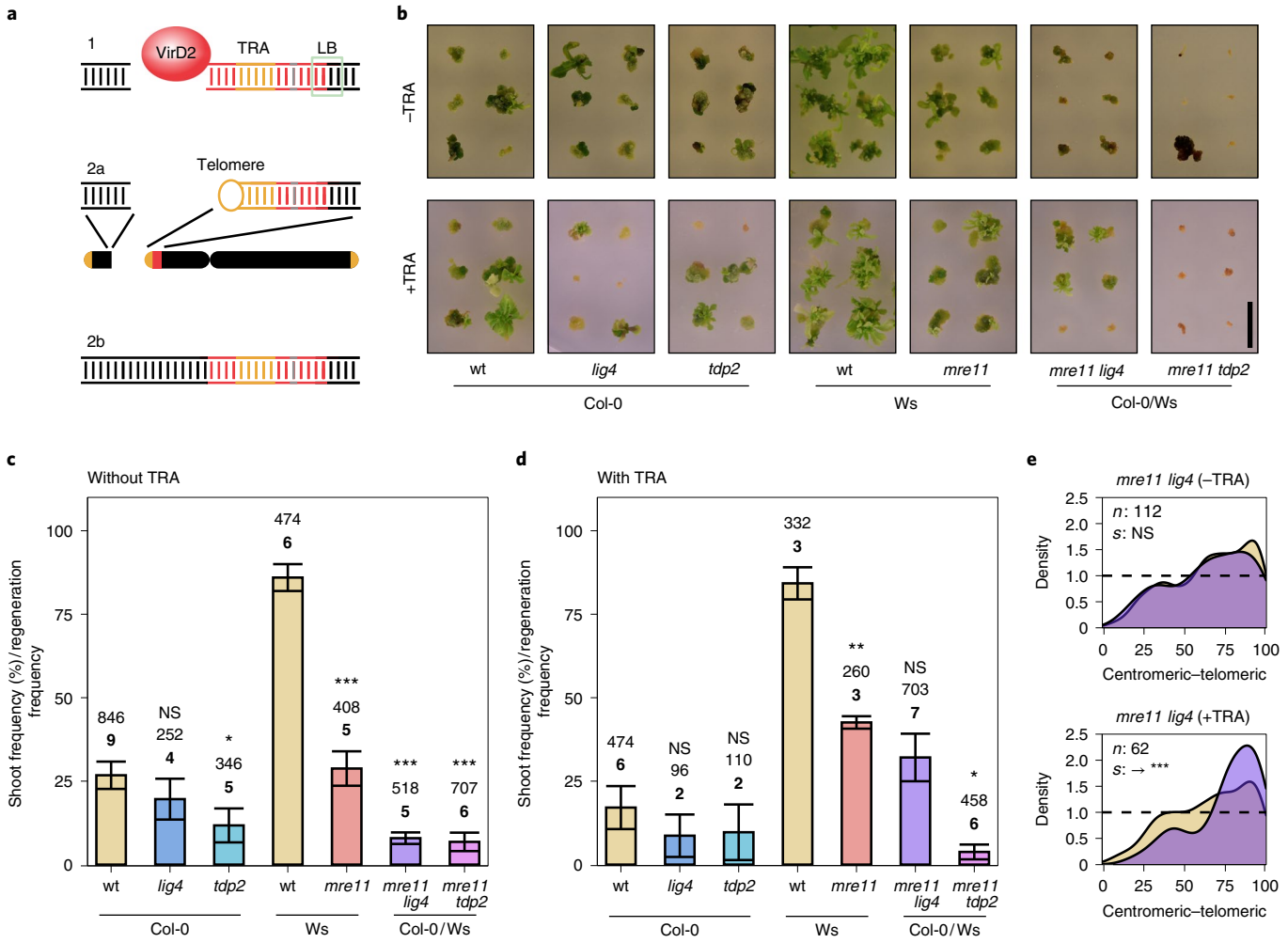


Fig. 3 | MRE11 and TDP2 are required for 5′ attachment of T-DNA. a, Following genomic capture of the 3′ end (1), a 5′ TRA-containing T-DNA can be resolved either via de novo telomere formation (2a) or regular RB capture (2b). In the former case, partial chromosome loss may ensue. Genomic DNA is indicated in black, TRA in orange, selectable marker in grey and the rest of the T-DNA in red. **b**, Calli of various genotypes with different genetic backgrounds (Col-0, Ws or Col-0/Ws hybrid) after transformation without TRA (pUBC, −TRA) or with TRA (pWY82, +TRA) grown on selective medium. Scale bar, 1 cm. **c**,**d**, Average percentage of calli with shoot tissue, corrected for potential regeneration defects, after transformation with pUBC (**c**) or pWY82 (**d**). Indicated are the total number of scored calli (italic) and number of independent experiments (bold) over which the mean (coloured bars) and s.e.m.(error bars) was calculated. One-sided Student's *t*-tests were performed to test for decreased integration compared to wt ($P_{lig4-TRA}=2\times10^{-1}$, $P_{tdp2-TRA}=2\times10^{-2}$, $P_{mre11-TRA}=2\times10^{-6}$, $P_{mre11lig4-TRA}=7\times10^{-4}$, $P_{mre11tdp2-TRA}=3\times10^{-4}$, $P_{lig4+TRA}=2\times10^{-1}$, $P_{tdp2+TRA}=3\times10^{-1}$, $P_{mre11+TRA}=3\times10^{-3}$, $P_{mre11lig4+TRA}=9\times10^{-1}$, $P_{mre11tdp2+TRA}=4\times10^{-2}$). Mutants were compared with the wt of the same genetic background, except for hybrids, which were compared with Col-0 wt. **e**, Relative frequency of LB junctions after transformation with pWY82 (+TRA) or pUBC (−TRA) along all chromosome arms, comparing wt (yellow) and *mre11 lig4* mutant (purple). 0% indicates centromeric position and 100% telomeric; *n* indicates the number of mutant junctions. Wilcoxon rank-sum tests were performed to find the direction (one-sided tests) and significance level (two-sided tests) of the shifts (*s*) in relative position ($P_{-TRA}=7\times10^{-1}$, $P_{+TRA}=9\times10^{-4}$). Only junctions that are represented by more than 20 different DNA molecules (events compatible with multiple cell divisions) were included in this analysis. NS, $P\geq0.05$; *, $P<0.05$; **, $P<0.01$; ***, $P<0.001$.

that follows from our genetic analyses is that although single cNHEJ and *mre11-2* mutant plants are proficient for AMT, albeit at a reduced frequency, double mutants may not be. This is indeed what we observe: whereas 30%–60% of transformed calli with *ku70*, *ku80* or *mre11-2* genotype form shoots on selective medium (which we use as a proxy for stable T-DNA integration), we find none in calli with *mre11-2 ku70* and *mre11-2 ku80* genotypes (Fig. 2h,i, Extended Data Fig. 6 and Supplementary Fig. 1). Corroborating the absence of shoots, we also found a dramatic reduction in the number of junctions in *mre11-2 ku70*, and (to a somewhat lesser extent) in *mre11-2 lig4* calli using TRANSGUIDE competition experiments (Extended Data Fig. 7). Expression of a T-DNA encoded β-glucuronidase marker demonstrates that the absence of T-DNA integration in the double mutants is not caused by impaired T-DNA transfer (Extended Data Fig. 8).

TDP2 is required for cNHEJ-mediated RB capture. The notion of cNHEJ being proficient in attaching the 5′ end of the T-DNA to the genome when MRE11 is impaired argues for another activity able to remove VirD2. The fact that most RB–genome junctions are without loss of the T-DNA's outermost 5′ nucleotides suggests the action of an enzyme able to cleave the phosphotyrosyl bond between VirD2 and the 5′ phosphate of the DNA, generating a ligatable end that can be used by cNHEJ. Previous work in a variety of biological systems has identified that tyrosyl-DNA phosphodiesterase 2 (TDP2) possesses such biochemical activity[30], hence we next

assayed *Arabidopsis* plants deficient for the orthologous protein. Root tissue from such *tdp2* mutant plants was efficiently transformed by *Agrobacterium* as visualized by shoot formation from selected calli, demonstrating that TDP2 is not essential for T-DNA integration (Fig. 2i,j). However, similar to mutations in cNHEJ, TDP2 deficiency also alters the junctional spectrum, specifically of RB–genome junctions, which shifts towards a typical TMEJ profile (Fig. 2k and Extended Data Fig. 9). This outcome is consistent with a model in which TDP2 acts to facilitate cNHEJ and, in line with this interpretation, we find that AMT is severely impaired in *mre11-2 tdp2* double-mutant plants (Fig. 2i,j, Extended Data Fig. 6 and Supplementary Fig. 1).

Bypass of 5′ T-DNA capture by telomere formation. We next reasoned that mutant backgrounds that have impaired T-DNA integration because of an inability to capture the 5′ end would be proficient for AMT in situations in which 3′ attachment of a T-DNA is sufficient to produce cells that stably transmit T-DNA. Such T-DNAs have been created previously, namely T-DNAs that contain so-called telomere repeat arrays (TRAs) at their 5′ side, which are long stretches of sequence consisting exclusively of $(TTTAGGG)_n$ that are able to trigger the formation of new telomeres following genomic capture at their 3′ end[31] (see Fig. 3a for a schematic representation). Two types of outcomes are found upon AMT of TRA-containing T-DNAs: type I with canonical T-DNA integration at a random position in the genome; and type II with telomere formation-dependent integration, which goes together with loss of DNA positioned between the new and former telomere[31]. Probably because they provoke haplo-insufficiency (providing counter-selection for viability) type II integrations are preferentially found near the chromosomal ends (within ~2.5 mb) in full-grown plants[31]. We next performed AMT experiments using TRA-containing T-DNA (in parallel to control T-DNAs) in the aforementioned genetic backgrounds. A *lig4* mutant background was used to assay cNHEJ deficiency because KU70/80 is involved in maintaining telomere homeostasis and also strongly affects de novo telomere formation[31–33]. In agreement with cNHEJ being required for AMT in plants with disturbed MRE11 function, we found profoundly reduced shoot formation in *mre11-2 lig4* mutant plants transformed with control T-DNA (Fig. 3b,c), although not to the same extent as observed earlier for *mre11-2 ku70* and *mre11-2 ku80*, which failed to produce shoots altogether (Fig. 2i, Extended Data Fig. 6 and Supplementary Fig. 1). However, successful AMT with a telomere-forming T-DNA construct did not require functional cNHEJ in the *mre11-2* mutant background (Fig. 3b,d and Supplementary Fig. 2), supporting the conclusion that cNHEJ action is specific to genomic attachment of the 5′ end of T-DNAs. In agreement with the prediction that these integrations are predominantly of type II, we found upon inspection by TRANSGUIDE a profound overrepresentation of LB junctions mapping near the ends of chromosomes (Fig. 3e and Extended Data Fig. 10). The finding that AMT was reduced for *mre11-2 tdp2* mutant roots even with TRA-containing T-DNA, yet not in the respective single mutants (Fig. 3b,d), argues that 5′ covalently bound VirD2 is also a blocking entity to de novo telomere formation.

Discussion

Following our previous elucidation of how, during AMT, the 3′ end of a T-DNA molecule is attached to the *Arabidopsis* genome, we have here identified the mechanisms by which the 5′ end can be attached. In contrast to the T-DNA's 3′ end, which because of its chemical composition (a 3′ hydroxyl at the terminus of a ssDNA molecule) is an ideal substrate for TMEJ, the structure of the 5′ end needs additional processing to create a ligatable end. Our data suggest that MRE11 acts to liberate the 5′ end to facilitate TMEJ, whereas TDP2 acts to allow genomic attachment via cNHEJ.

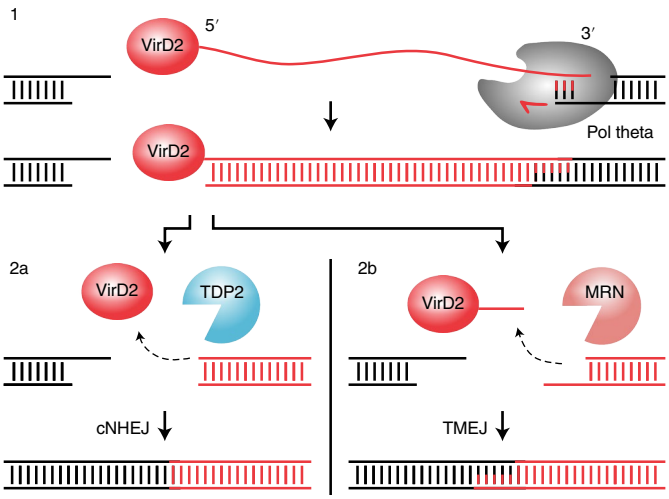


Fig. 4 | Proposed model of T-DNA integration. Capture of the T-DNA's 3′ end essentially requires Pol θ (1). Subsequent attachment of the 5′ end can occur via two redundant pathways: TDP2-mediated removal of VirD2 and attachment of the resulting blunt end to the genome via cNHEJ (2a) or end resection by MRN (of which MRE11 is the enzymatic core), resulting in an 3′ overhang that is suitable for Pol θ action (2b).

Given the biochemical properties of both MRE11 and TDP2 acting on dsDNA, we consider it probable that single-stranded T-DNA molecules are first converted to a double-stranded configuration before 5′ attachment. One potential mechanism for such conversion is genomic capture of the T-DNA 3′ end followed by DNA synthesis using the genomic end as a primer. In this way a new 'extended' DSB end is created (Fig. 4) in which the VirD2 protein blocks 5′ to 3′ resection. Such a structure is conceptually similar to a meiotic SPO11-bound DSB end or a stalled TOP2 cleavage complex, substrates that depend either on MRE11 or on TDP2 for protein removal to facilitate repair. Although multiple biological activities have been described for MRE11—besides removing 5′ bound proteins from DNA, the MRE11-containing MRN complex has also been shown to act as a DNA break sensor and as a DNA resection entity—our data are most parsimoniously explained by a defect in the removal of covalently bound proteins in *mre11-2* mutant plants.

In Fig. 4 we visualize the most simple model for T-DNA integration that is consistent with the data we obtained, but our results do not discriminate between scenarios in which the originally single-stranded T-DNA is converted to dsDNA before, during or after genomic capture. Considering the mechanism of TMEJ, we consider it logical to assume that the 3′ outermost end of the T-DNA is single-stranded before genomic capture. However, it is currently unknown whether this is also the situation for the entire T-DNA sequence: is genomic capture essential for converting T-DNA into a double-strand conformation? Early work aimed at unravelling the mechanism of T-DNA integration provided experimental support for the suggestion that T-DNA molecules can be (at least in part) double-stranded before their integration[34,35]. Furthermore, the observation of relatively proficient 'transient' expression of T-DNA-encoded genes in *Arabidopsis* deficient for Pol θ argues for a dsDNA formation also in the absence of genomic capture. It is conceivable that such free-floating T-DNA molecules can also react with each other via the identified end-joining mechanisms before genomic capture, a process that may underlie two yet unexplained AMT phenomena: extrachromosomal T-circles[36,37], and T-DNA conglomerates that were recently found to make up a large proportion of AMT outcomes[38,39]. Most of the extrachromosomal T-DNAs

may not be able to integrate, but may still be maintained by the cell for a long period and confer growth on selective media.

The observation of cNHEJ-mediated attachment of T-DNA 5′ ends also in Pol θ-proficient cells reveals that a proportion of the integrations used both pathways, that is cNHEJ for 5′ attachment and TMEJ for 3′ attachment, as hypothesized previously[40]. This finding may explain many seemingly contradictory observations in mutant analysis that have confounded AMT research for several decades, namely that usage of cNHEJ over MRE11-stimulated TMEJ to capture the 5′ end may be context dependent with respect to the AMT protocol, the reagents used and the tissue that is targeted. cNHEJ repairs DSBs in G1 and in pre-replicative DNA in S phase[41], whereas recent work in mammalian cells argues for TMEJ in late-S/G2/M phases of the cell cycle[42], and it is thus tempting to speculate that the cell-cycle stage of the host cell when infected may dictate pathway choice and AMT outcome. Indeed, comparing the genome–T-DNA junction signature of AMT events derived from somatic transformation with those from germline transformation reveals that TMEJ is more prominently used to attach the 5′ end of T-DNA in germ cells[9,12] (Extended Data Fig. 2). The notion of Pol θ-independent genomic capture of T-DNA, which we found to occur almost exclusively at the 5′, making Pol θ virtually essential for 3′ capture and thus T-DNA integration, may allow for DNA integration in experimental set-ups, genetic systems or of DNA substrates that are different from those tested here. For instance, in cases where integration may not critically depend on joining ssDNA, or alternatively, in situations where ample homology is already present, making Pol θ obsolete[43].

Apart from providing a mechanistic understanding, we aim to unravel the biology of (T-)DNA integration to allow for improved biotechnological strategies to develop transgenic crops. Recent work demonstrated that homology-directed gene targeting in Pol θ-deficient *Arabidopsis* goes without undesired integration of AMT reagents[44], which otherwise contaminates gene targeting in wild-type conditions. Here, we find that a combinatorial inhibition of *MRE11* and cNHEJ activities, for which inhibitors are available, also precludes random integration. We envisage that an increased understanding on how exogenously provided DNA molecules interact with the genome of a host plant can help in developing precise genome-engineering approaches to benefit crop development.

Methods

Plant lines and growth conditions. Insertional mutants used in this study were *ku70* (refs. [19,45]) (Col-0, SALK_123144), *ku70* (refs. [19,32]) (Ws), *ku80* (ref. [19]) (SALK_016627), *lig4* (ref. [45]) (SALK_044027), *tdp2-1* (SALK _043413), *teb-5* (refs. [9,46]) (SALK_018851), *mre11-2* (refs. [19,29]). Double mutants were created by crossing single mutants. Plants were grown on soil at 20 °C in a 16 h light/8 h dark cycle.

Root transformation for TRANSGUIDE and shoot formation assay. Root transformations were performed as described previously, using disarmed *Agrobacterium tumefaciens* strain AGL1 (ref. [47]) harbouring either pUBC (pUBC-YFP-Dest[48] with *ccdB* cassette removed), pUBC-2 (as pUBC-YFP, but the sequence between secondary TRANSGUIDE primer and LB or RB nick was replaced by a semi-random 56 bp sequence), pWY82 (ref. [49]) or pCAS9 (pDe-CAS9)[50,51] with gRNA against *PPO1;* AT4G01690) or pCAMBIA3301 (Cambia). In brief, seedlings were grown for 11 d, after which roots were removed, and placed on callus-induction medium where they incubated for 4 d. The explants were then co-cultured with *Agrobacterium* on callus-induction medium supplemented with acetosyringone for 2 d. After co-culture root explants were transferred to shoot-induction medium with vancomycin and timentin to kill any remaining bacteria, and phosphinothricin to select for transformed plant cells. After 3 weeks of selection calli were either harvested for TRANSGUIDE analysis (20 per sample), or transferred to fresh selection medium for assaying shoot formation. After a total of 6 weeks of selection, plates were photographed and calli were scored for shoot formation (without prior knowledge of the genotype); any leaf-like protrusions from callus tissue was considered shoot tissue. Calli were also grown on non-selective medium to obtain regeneration frequencies, for which the shoot formation frequencies on selective medium were corrected. Independent experiments were performed to obtain biological replicates for statistical analysis.

Junction enrichment and sequencing. Enrichment of T-DNA–genome junctions was similar to the GUIDEseq procedure[52]. First, DNA extraction from calli was performed with the Wizard genomic DNA isolation kit (Promega). Genomic DNA was sonicated to a suitable size range with a Bioruptor (Diagenode) for six cycles (30 s on, 30 s off) on 'high' intensity. End repair, A-tailing and (home-made) Y-adaptor ligation were performed with the NEBNext ultra II kit (New England Biolabs), and the nested library amplification was performed with Phusion polymerase (ThermoFisher Scientific). The primers used are given in Supplementary Table 1. Sample size selection and clean-up was performed after sonication, adaptor ligation and after both PCR reactions using Ampure XP beads (Beckman Coulter). Sequencing was performed on the Illumina MiSeq (300 bp paired end, v3 chemistry, at the Leiden Genome Technology Center) and on the Illumina NovaSeq 6000 (150 bp paired end, v1.5 chemistry, at GenomeScan). Samples were demultiplexed using bcl2fastq2 conversion software v.2.2 (Illumina).

Junction calling. Reads were clipped to 150 bp and adaptors removed (Trimmomatic, v.0.39)[53]. Reads with identical molecular identifiers (adaptor UMI + 6 bp from forward read + 6 bp from reverse read) were combined into consensus sequences using custom software. Mapping was done with BWA-mem (v.0.7.17), using the default settings. Reads with identical unique molecular identifiers were combined into consensus sequences, and any remaining optical duplicates were excluded from the analysis. Read pairs were grouped into junctions based on their genomic positions. Reads that only aligned to T-DNA were not included in the analysis. Second-in-pair reads were required to start with the (T-DNA part of the) secondary T-DNA primer and end with a genomic sequence. These reads were used to determine the exact genomic position, as well as filler and homology sequences and deletion length. Unique first-in-pair reads (anchors) were counted for each junction, and indicated the number of fragments present in the sample that support the junction. For each junction we generated a consensus sequence and calculated the percentage of reads exactly matching the consensus (consensus match). The junctions were then filtered in R: (1) for duplicate positions between samples (removing all such junctions, in case of barcode-hopping the 'donor' junction of the misassigned reads was kept); (2) for number of anchors (at least three); and (3) for the consensus match (at least 75%). For most analyses (except for junction number comparison) we applied an additional filter for fair comparison, because distances between primer and border were not constant: homology ≤ 57 bp, filler ≤ 22 bp, end deletion ≤ 26 bp. The filtering was performed in R (v.4.1.0).

AMT competition experiment. Roots were transformed with either pUBC (barcode 1) or pUBC-2 (barcode 2). Ten calli were collected per sample, and equal DNA amounts of two samples with different barcodes were combined before junction enrichment. During junction calling the reads were assigned to the sample of origin using the barcode. Junctions with duplicate positions within a sample pair were removed.

Junction validation. Using the same DNA samples as used for TRANSGUIDE, we performed up to two PCR reactions (nested) followed by Sanger sequencing (Macrogen Europe) to determine the correctness of the called junctions. Junctions were selected semi-randomly, making sure different types of junctions (filler/non-filler, intact/non-intact, and so on) were included. The primers used are given in Supplementary Table 1. Junctions were visually inspected using IGV (v.2.8.0).

β-Glucuronidase staining. After co-cultivation, root explants were stained overnight in phosphate buffer (pH 7.3) containing 1 mM $K_3Fe(CN)_6$, 1 mM $K_4Fe(CN)_6$, 10 mM Na_2EDTA, 0.1 % SDS, 0.1 % Triton X-100 and 2 mM X-Gluc, and destained using 70% ethanol.

Statistics. For junction footprint analyses (filler, homology, T-DNA end loss), samples with the same genotype were combined (and sometimes samples with different T-DNAs, see figure legends); each junction of the combined sample constituting a single observation. Data shown for each genotype were thus composed of at least three samples. Wilcoxon rank-sum tests were performed to identify significant shifts in homology, filler, T-DNA end loss and genomic position distributions; one-sided tests were used to identify the direction of the shift, and two-sided tests to determine the *P*-value. For seamless/non-seamless and filler/non-filler ratio analyses, samples were kept separate and each sample constituted an observation. Two-sided Student's *t*-tests were performed to determine whether differences in ratios were significant. For shoot formation analyses, several independent experiments were performed, and for each at least 30 calli (per genotype–T-DNA combination) were examined for shoot formation. For each experiment one frequency value was thus obtained (per genotype–T-DNA combination), which was considered a single observation for statistical purposes. One-sided Student's *t*-tests were performed to determine whether mutants had a significantly lower shoot formation frequency than wild-type. Statistical tests were performed in R.

Data figures were generated in R using the ggplot2 package; composite figures were assembled in Inkscape (v.0.92). Inkscape was also used to generate the model figures.

Reporting Summary. Further information on research design is available in the Nature Research Reporting Summary linked to this article.

Data availability

Mapped sequences are available from NCBI SRA (accession code: PRJNA786733). Junction footprint data is provided in Supplementary Data 2 and 3. The pAC161 data in Extended Data Figs. 2 and 3 is based on previously published data[9,12].

Code availability

The custom java program used for junction calling is available from GitHub (https://github.com/RobinVanSchendel/TRANSGUIDE).

Received: 6 December 2021; Accepted: 30 March 2022;
Published online: 9 May 2022

References

1. Bevan, M. W. & Chilton, M.-D. T-DNA of the *Agrobacterium* Ti and Ri plasmids. *Annu. Rev. Genet.* **16**, 357–384 (1982).
2. Stachel, S. E., Timmerman, B. & Zambryski, P. Generation of single-stranded T-DNA molecules during the initial stages of T-DNA transfer from *Agrobacterium tumefaciens* to plant cells. *Nature* **322**, 706–712 (1986).
3. Ward, E. R. & Barnes, W. M. VirD2 protein of *Agrobacterium tumefaciens* very tightly linked to the 5′ end of T-strand DNA. *Science* **242**, 927 (1988).
4. Scheiffele, P., Pansegrau, W. & Lanka, E. Initiation of *Agrobacterium tumefaciens* T-DNA processing. Purified proteins VirD1 and VirD2 catalyze site- and strand-specific cleavage of superhelical T-border DNA in vitro. *J. Biol. Chem.* **270**, 1269–1276 (1995).
5. van Kregten, M., Lindhout, B. I., Hooykaas, P. J. & van der Zaal, B. J. *Agrobacterium*-mediated T-DNA transfer and integration by minimal VirD2 consisting of the relaxase domain and a type IV secretion system translocation signal. *Mol. Plant Microbe Interact.* **22**, 1356–1365 (2009).
6. Winans, S. C. Two-way chemical signaling in *Agrobacterium*–plant interactions. *Microbiol. Rev.* **56**, 12–31 (1992).
7. Citovsky, V. & Zambryski, P. Transport of nucleic acids through membrane channels: snaking through small holes. *Annu. Rev. Microbiol.* **47**, 167–197 (1993).
8. Kim, S. I., Veena & Gelvin, S. B. Genome-wide analysis of *Agrobacterium* T-DNA integration sites in the *Arabidopsis* genome generated under non-selective conditions. *Plant J.* **51**, 779–791 (2007).
9. van Kregten, M. et al. T-DNA integration in plants results from polymerase-theta-mediated DNA repair. *Nat. Plants* **2**, 16164 (2016).
10. Schimmel, J., van Schendel, R., den Dunnen, J. T. & Tijsterman, M. Templated insertions: a smoking gun for polymerase theta-mediated end joining. *Trends Genet.* **35**, 632–644 (2019).
11. Ramsden, D. A., Carvajal-Garcia, J. & Gupta, G. P. Mechanism, cellular functions and cancer roles of polymerase-theta-mediated DNA end joining. *Nat. Rev. Mol. Cell Biol.* **23**, 125–140 (2022).
12. Kleinboelting, N. et al. 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 (2015).
13. Tinland, B. The integration of T-DNA into plant genomes. *Trends Plant Sci.* **1**, 178–184 (1996).
14. Tzfira, T., Li, J., Lacroix, B. & Citovsky, V. *Agrobacterium* T-DNA integration: molecules and models. *Trends Genet.* **20**, 375–383 (2004).
15. Shilo, S. et al. T-DNA–genome junctions form early after infection and are influenced by the chromatin state of the host genome. *PLoS Genet.* **13**, e1006875 (2017).
16. Nishizawa-Yokoi, A. et al. *Agrobacterium* T-DNA integration in somatic cells does not require the activity of DNA polymerase theta. *N. Phytol.* **229**, 2859–2872 (2021).
17. Friesner, J. & Britt, A. B. Ku80- and DNA ligase IV-deficient plants are sensitive to ionizing radiation and defective in T-DNA integration. *Plant J.* **34**, 427–440 (2003).
18. Li, J. et al. Involvement of KU80 in T-DNA integration in plant cells. *Proc. Natl Acad. Sci. USA* **102**, 19231–19236 (2005).
19. Jia, Q., Bundock, P., Hooykaas, P. J. J. & de Pater, S. *Agrobacterium tumefaciens* T-DNA integration and gene targeting in *Arabidopsis thaliana* non-homologous end-joining mutants. *J. Bot.* **2012**, 989272 (2012).
20. Mestiri, I., Norre, F., Gallego, M. E. & White, C. I. Multiple host–cell recombination pathways act in *Agrobacterium*-mediated transformation of plant cells. *Plant J.* **77**, 511–520 (2014).
21. Gallego, M. E., Bleuyard, J. Y., Daoudal-Cotterell, S., Jallut, N. & White, C. I. Ku80 plays a role in non-homologous recombination but is not required for T-DNA integration in *Arabidopsis*. *Plant J.* **35**, 557–565 (2003).
22. van Attikum, H. et al. The *Arabidopsis AtLIG4* gene is required for the repair of DNA damage, but not for the integration of *Agrobacterium* T-DNA. *Nucleic Acids Res.* **31**, 4247–4255 (2003).
23. Park, S. Y. et al. *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 (2015).
24. Vaghchhipawala, Z. E., Vasudevan, B., Lee, S., Morsy, M. R. & Mysore, K. S. *Agrobacterium* may delay plant nonhomologous end-joining DNA repair via XRCC4 to favor T-DNA integration. *Plant Cell* **24**, 4110–4123 (2012).
25. Hartsuiker, E., Neale, M. J. & Carr, A. M. Distinct requirements for the Rad32Mre11 nuclease and Ctp1CtIP in the removal of covalently bound topoisomerase I and II from DNA. *Mol. Cell* **33**, 117–123 (2009).
26. Hartung, F. et al. The catalytically active tyrosine residues of both SPO11-1 and SPO11-2 are required for meiotic double-strand break induction in *Arabidopsis*. *Plant Cell* **19**, 3090–3099 (2007).
27. Neale, M. J., Pan, J. & Keeney, S. Endonucleolytic processing of covalent protein-linked DNA double-strand breaks. *Nature* **436**, 1053–1057 (2005).
28. Puizina, J., Siroky, J., Mokros, P., Schweizer, D. & Riha, K. Mre11 deficiency in *Arabidopsis* is associated with chromosomal instability in somatic cells and Spo11-dependent genome fragmentation during meiosis. *Plant Cell* **16**, 1968–1978 (2004).
29. Bundock, P. & Hooykaas, P. Severe developmental defects, hypersensitivity to DNA-damaging agents, and lengthened telomeres in *Arabidopsis* MRE11 mutants. *Plant Cell* **14**, 2451–2462 (2002).
30. Zeng, Z., Cortés-Ledesma, F., El Khamisy, S. F. & Caldecott, K. W. TDP2/TTRAP is the major 5′-tyrosyl DNA phosphodiesterase activity in vertebrate cells and is critical for cellular resistance to topoisomerase II-induced DNA damage. *J. Biol. Chem.* **286**, 403–409 (2011).
31. Nelson, A. D., Lamb, J. C., Kobrossly, P. S. & Shippen, D. E. Parameters affecting telomere-mediated chromosomal truncation in *Arabidopsis*. *Plant Cell* **23**, 2263–2272 (2011).
32. Bundock, P., van Attikum, H. & Hooykaas, P. Increased telomere length and hypersensitivity to DNA damaging agents in an *Arabidopsis* KU70 mutant. *Nucleic Acids Res.* **30**, 3395–3400 (2002).
33. Riha, K., Watson, J. M., Parkey, J. & Shippen, D. E. Telomere length deregulation and enhanced sensitivity to genotoxic stress in *Arabidopsis* mutants deficient in Ku70. *EMBO J.* **21**, 2819–2826 (2002).
34. Chilton, M.-D. M. & Que, Q. 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 (2003).
35. Tzfira, T., Frankman, L. R., Vaidya, M. & Citovsky, V. Site-specific integration of *Agrobacterium tumefaciens* T-DNA via double-stranded intermediates. *Plant Physiol.* **133**, 1011–1023 (2003).
36. Bakkeren, G., Koukolikova-Nicola, Z., Grimsley, N. & Hohn, B. Recovery of *Agrobacterium tumefaciens* T-DNA molecules from whole plants early after transfer. *Cell* **57**, 847–857 (1989).
37. Singer, K., Shiboleth, Y. M., Li, J. & Tzfira, T. Formation of complex extrachromosomal T-DNA structures in *Agrobacterium tumefaciens*-infected plants. *Plant Physiol.* **160**, 511–522 (2012).
38. Pucker, B., Kleinbolting, N. & Weisshaar, B. Large scale genomic rearrangements in selected *Arabidopsis thaliana* T-DNA lines are caused by T-DNA insertion mutagenesis. *BMC Genomics* **22**, 599 (2021).
39. Jupe, F. et al. The complex architecture and epigenomic impact of plant T-DNA insertions. *PLoS Genet.* **15**, e1007819 (2019).
40. Levy, A. A. T-DNA integration: Pol θ controls T-DNA integration. *Nat. Plants* **2**, 16170 (2016).
41. Hustedt, N. & Durocher, D. The control of DNA repair by the cell cycle. *Nat. Cell Biol.* **19**, 1–9 (2016).
42. Llorens-Agost, M. et al. POLθ-mediated end joining is restricted by RAD52 and BRCA2 until the onset of mitosis. *Nat. Cell Biol.* **23**, 1095–1104 (2021).
43. Kamp, J. et al. Helicase Q promotes homology-driven DNA double-strand break repair and prevents tandem duplications. *Nat. Commun.* **12**, 7126 (2021).
44. van Tol, N. et al. Gene targeting in polymerase theta-deficient *Arabidopsis thaliana*. *Plant J.* **109**, 112–125 (2021).
45. Du, Y., Hase, Y., Satoh, K. & Shikazono, N. Characterization of gamma irradiation-induced mutations in *Arabidopsis* mutants deficient in non-homologous end joining. *J. Radiat. Res.* **61**, 639–647 (2020).
46. Inagaki, S. et al. *Arabidopsis* TEBICHI, with helicase and DNA polymerase domains, is required for regulated cell division and differentiation in meristems. *Plant Cell* **18**, 879–892 (2006).
47. Lazo, G. R., Stein, P. A. & Ludwig, R. A. A DNA transformation-competent *Arabidopsis* genomic library in *Agrobacterium*. *Biotechnology (N Y)* **9**, 963–967 (1991).
48. Grefen, C. et al. A ubiquitin-10 promoter-based vector set for fluorescent protein tagging facilitates temporal stability and native protein distribution in transient and stable expression studies. *Plant J.* **64**, 355–365 (2010).
49. Yu, W., Lamb, J. C., Han, F. & Birchler, J. A. Telomere-mediated chromosomal truncation in maize. *Proc. Natl Acad. Sci. USA* **103**, 17331–17336 (2006).
50. Fauser, F., Schiml, S. & Puchta, H. Both CRISPR/Cas-based nucleases and nickases can be used efficiently for genome engineering in *Arabidopsis thaliana*. *Plant J.* **79**, 348–359 (2014).
51. Shen, H., Strunks, G. D., Klemann, B. J., Hooykaas, P. J. & de Pater, S. CRISPR/Cas9-induced double-strand break repair in *Arabidopsis* nonhomologous end-joining mutants. *G3 (Bethesda)* **7**, 193–202 (2017).

52. Tsai, S. Q. et al. GUIDE-seq enables genome-wide profiling of off-target cleavage by CRISPR–Cas nucleases. *Nat. Biotechnol.* **33**, 187–197 (2015).
53. Bolger, A. M., Lohse, M. & Usadel, B. Trimmomatic: a flexible trimmer for Illumina sequence data. *Bioinformatics* **30**, 2114–2120 (2014).

Acknowledgements

This work was in part funded by ALW OPEN grants (OP.393 and OP.269) from The Netherlands Organization for Scientific Research for Earth and Life Sciences to M.T.

Author contributions

L.E.M.K., S.d.P., P.J.J.H. and M.T. conceived the study. L.E.M.K., R.v.S. and S.L.K. developed the TRANSGUIDE method. L.E.M.K., H.S. and S.d.P. conducted experiments. L.E.M.K. and R.v.S. performed bioinformatic analyses. L.E.M.K. and M.T. wrote the manuscript with input from all authors, who read the manuscript and authorized its publication.

Competing interests

The authors declare no competing interests.

Additional information

Extended data is available for this paper at https://doi.org/10.1038/s41477-022-01147-5.

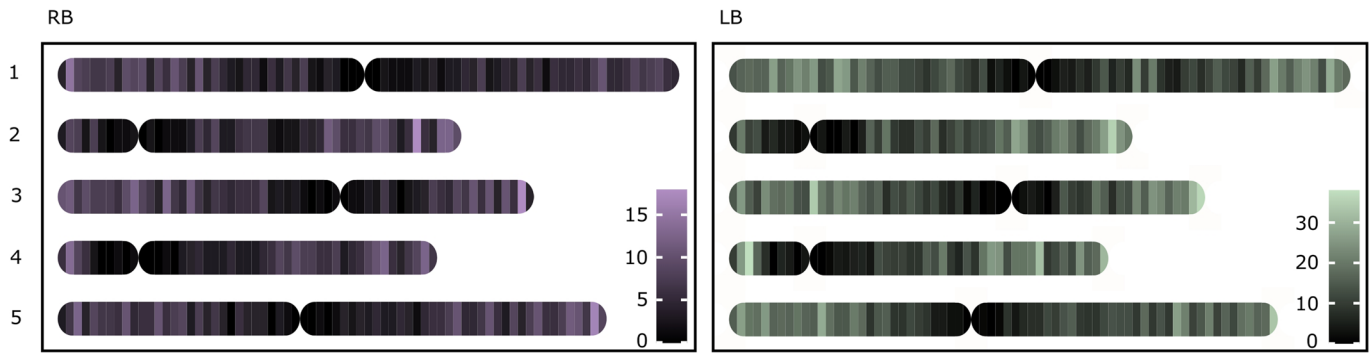
Supplementary information The online version contains supplementary material available at https://doi.org/10.1038/s41477-022-01147-5.

Correspondence and requests for materials should be addressed to Marcel Tijsterman.

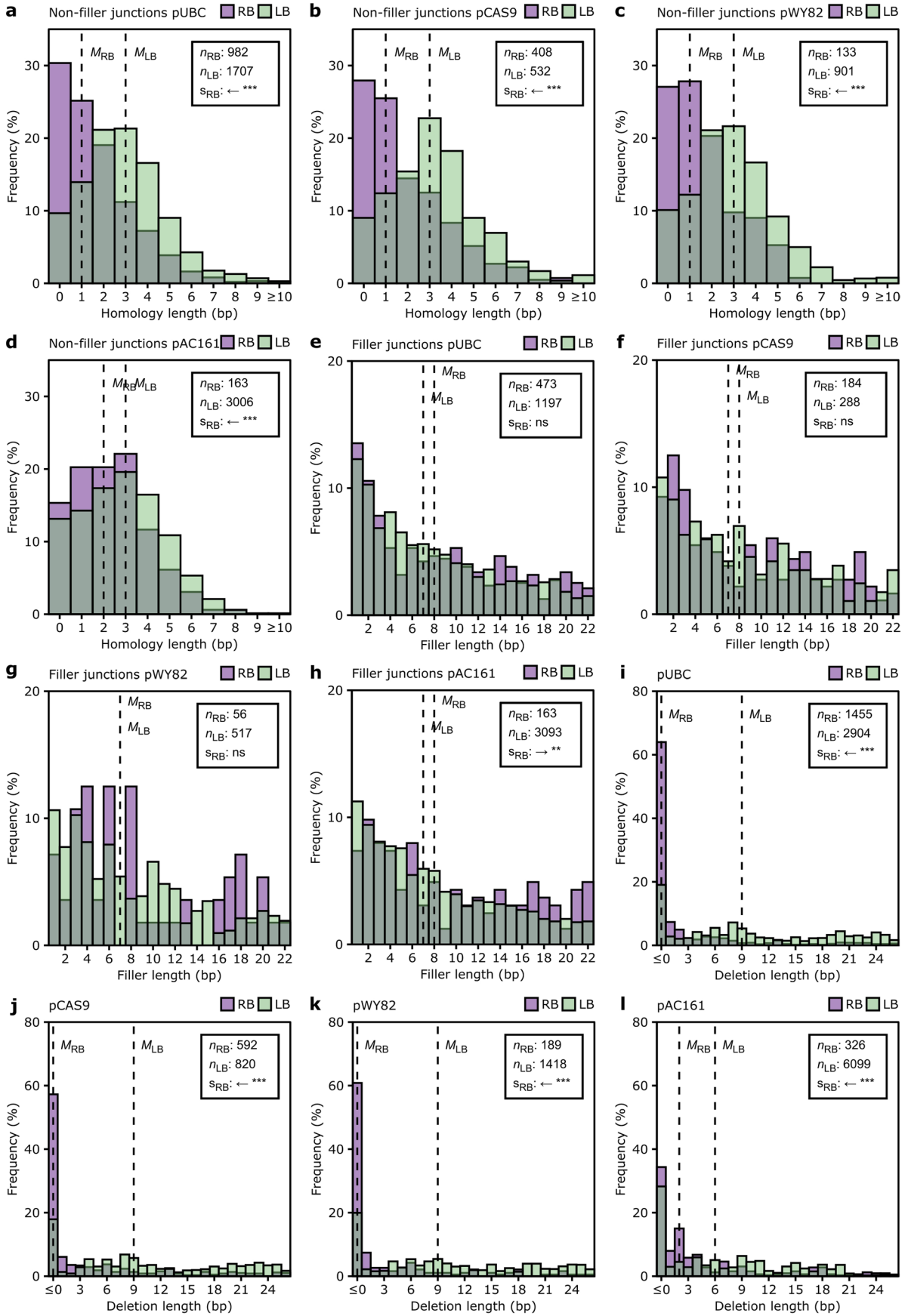
Peer review information *Nature Plants* thanks Anne Britt, Shunping Yan and the other, anonymous, reviewer(s) for their contribution to the peer review of this work.

Reprints and permissions information is available at www.nature.com/reprints.

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

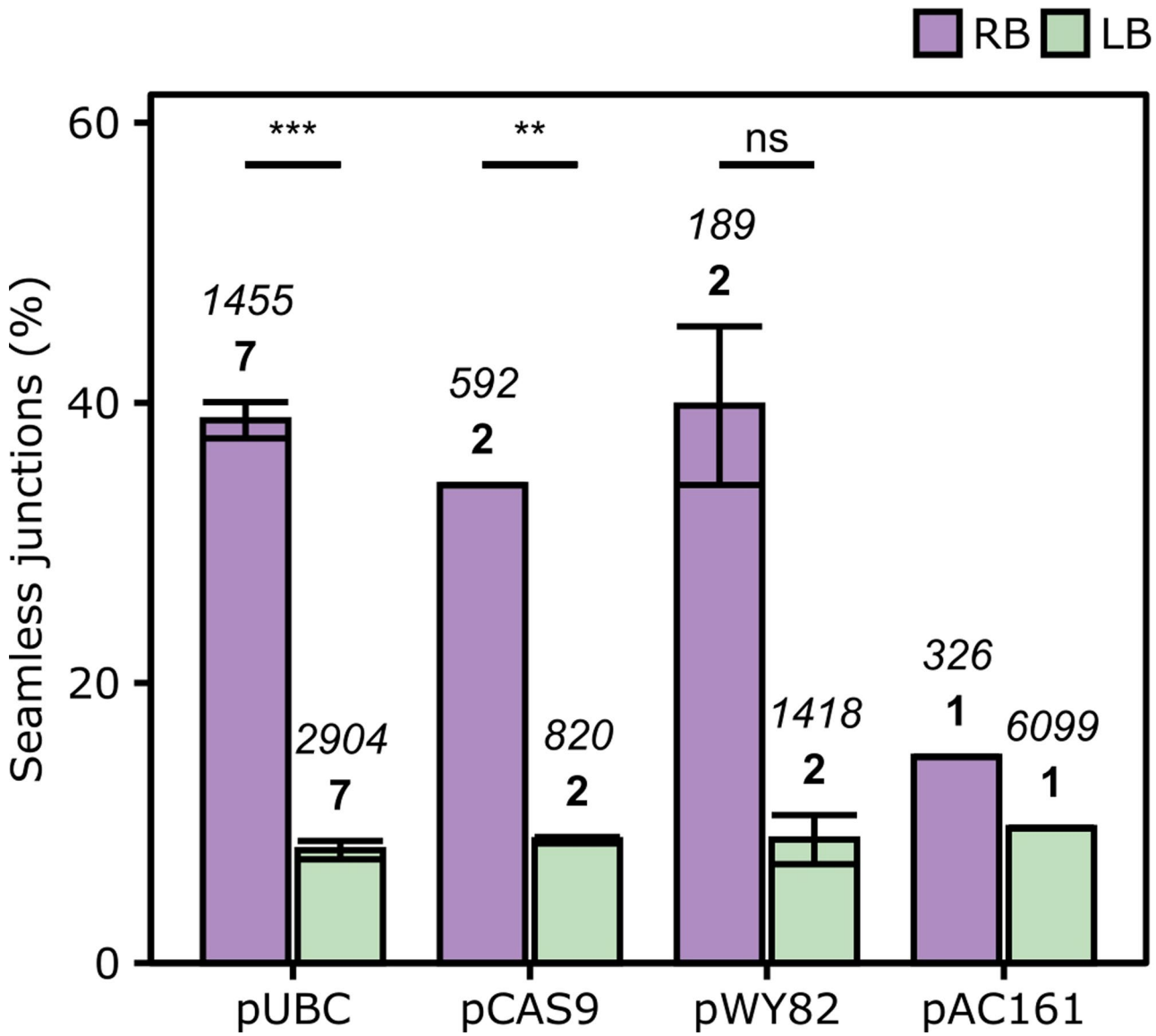


Extended Data Fig. 1 | Genomic position of wild-type junctions obtained with TRANSGUIDE. Arabidopsis chromosomes (1-5) were divided into 0.4 mb bins, in which RB (purple) and LB (green) junctions were counted. The brightness of the colour indicates the number of junctions (note the scales). Centromere positions were rounded to the nearest border between bins. The shown data is from 11 wt samples, transformed with either pUBC, pCAS9, or pWY82.

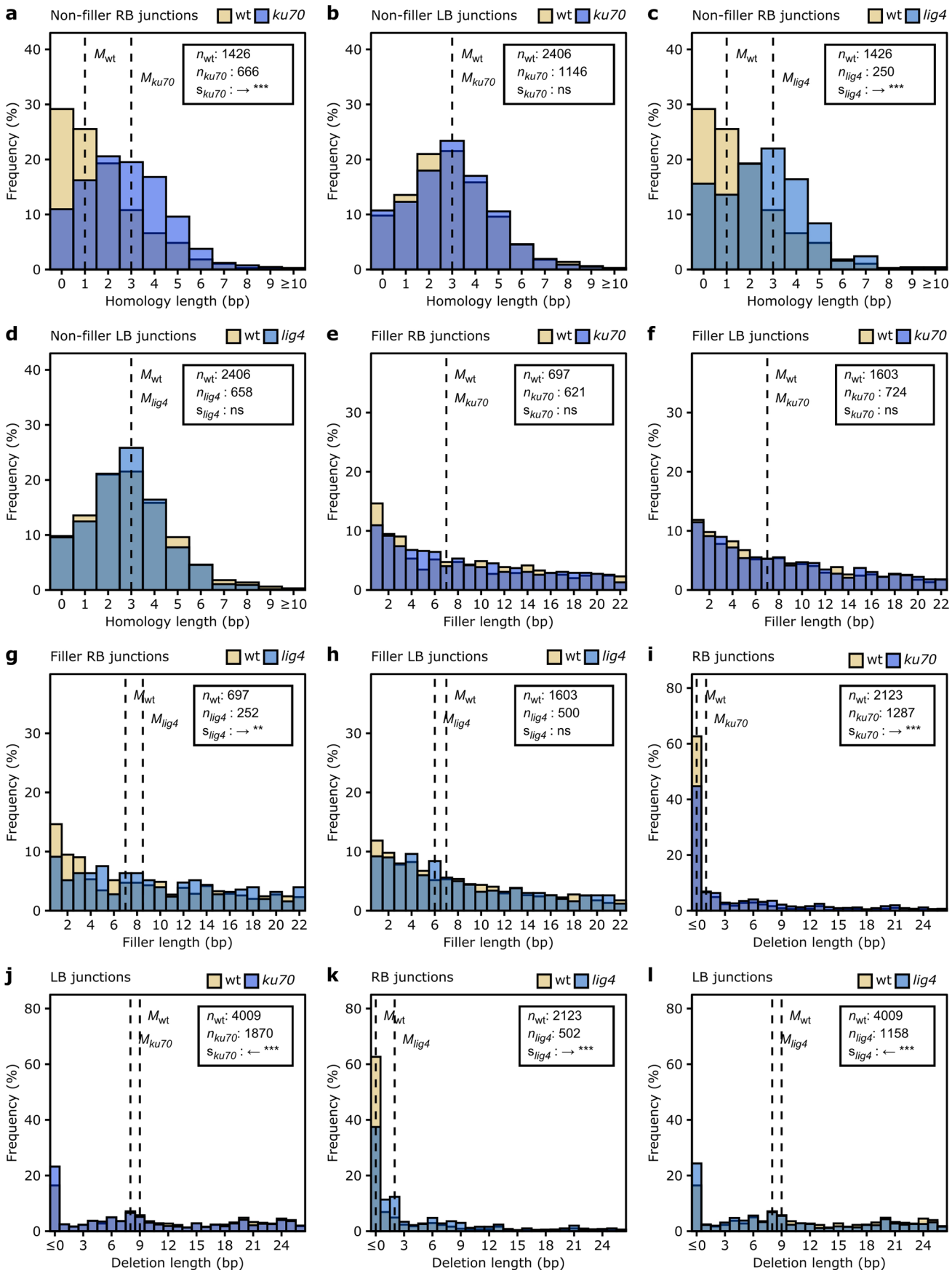


Extended Data Fig. 2 | See next page for caption.

Extended Data Fig. 2 | Homology, filler, and T-DNA loss profiles for 4 different constructs. Frequency of different lengths of microhomology (**a** - **d**), filler (**e** - **h**), or T-DNA loss (**i** - **l**) at RB (purple) and LB (green) junctions, for 3 constructs after somatic transformation (pUBC, pCAS9, and pWY82) and for 1 construct after germ-line transformation (pAC161). The overlap between LB and RB is indicated in olive-green. The medians (dashed lines), the number of observations (*n*), and shifts in the RB distribution relative to LB (s) are indicated. Wilcoxon rank-sum tests were performed to find the direction (one-sided tests) and the significance of the shifts (two-sided tests, $p_{homology\_pUBC} = 7\times10^{-68}$, $p_{homology\_pCAS9} = 7\times10^{-24}$, $p_{homology\_pWY82} = 2\times10^{-14}$, $p_{homology\_pAC161} = 8\times10^{-4}$, $p_{filler\_pUBC} = 2\times10^{-1}$, $p_{filler\_pCAS9} = 9\times10^{-1}$, $p_{filler\_pWY82} = 3\times10^{-1}$, $p_{filler\_pAC161} = 7\times10^{-3}$, $p_{deletion\_pUBC} = 4\times10^{-222}$, $p_{deletion\_pCAS9} = 1\times10^{-63}$, $p_{deletion\_pWY82} = 1\times10^{-32}$, $p_{deletion\_pAC161} = 1\times10^{-11}$). ns: $p \geq 0.05$, *: $p < 0.05$, **: $p < 0.01$, ***: $p < 0.001$.

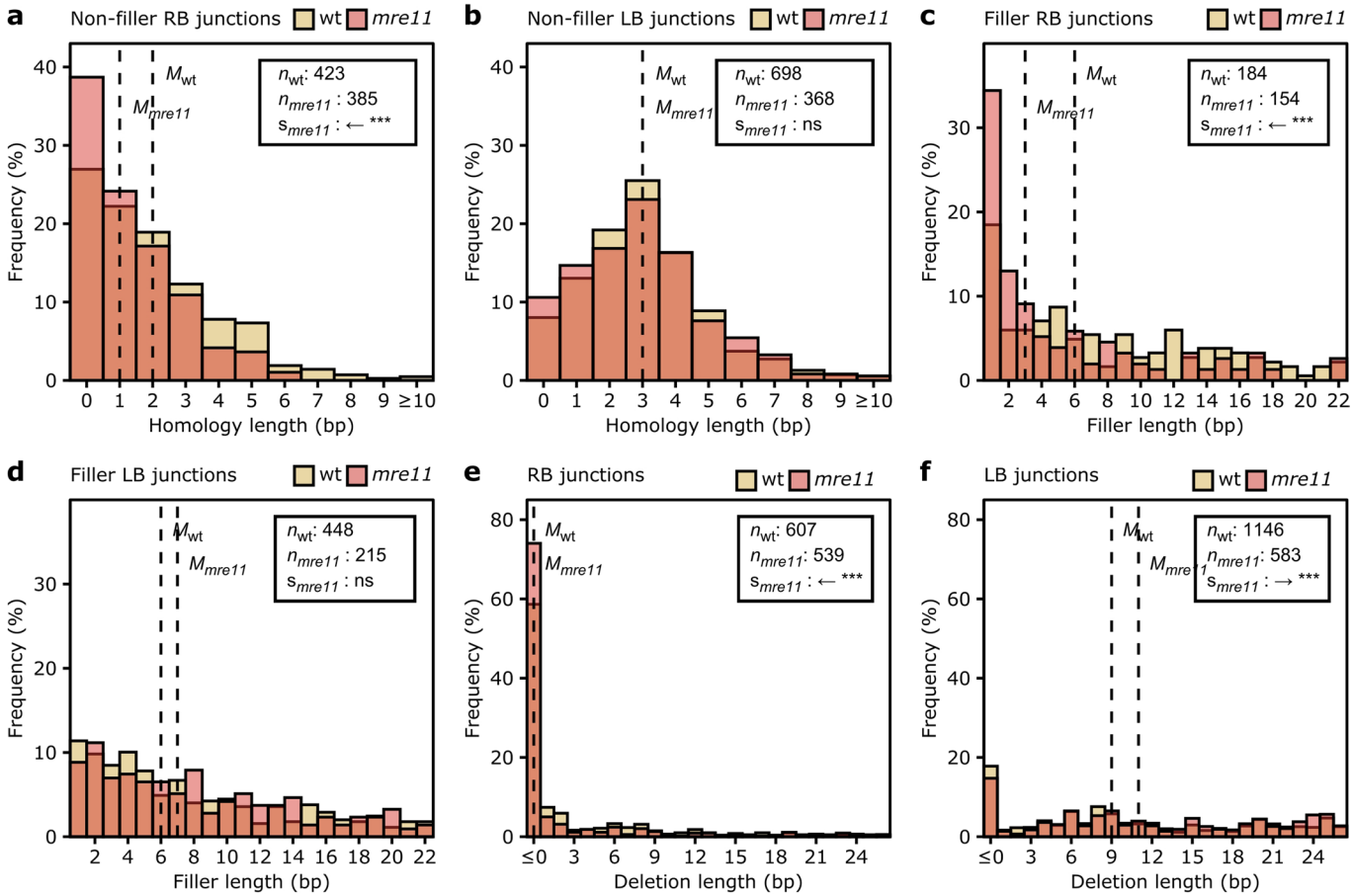


Extended Data Fig. 3 | Seamless junctions. Average percentages of RB and LB junctions without T-DNA loss and without insertions, after somatic transformation (pUBC, pCAS9, pWY82) and germ-line transformation (pAC161). The number in bold indicates the number of samples over which the mean and error bars (standard error of the mean) have been calculated; the number in italic indicates the total number of junctions amongst those samples that were scored for 'seamlessness'. Two-sided Student's t-tests have been performed to test whether the percentage of seamless junctions differed significantly between RB and LB junctions ($p_{pUBC} = 6 \times 10^{-9}$, $p_{pCAS9} = 6 \times 10^{-3}$, $p_{pWY82} = 9 \times 10^{-2}$). ns: $p \geq 0.05$, *: $p < 0.05$, **: $p < 0.01$, ***: $p < 0.001$.

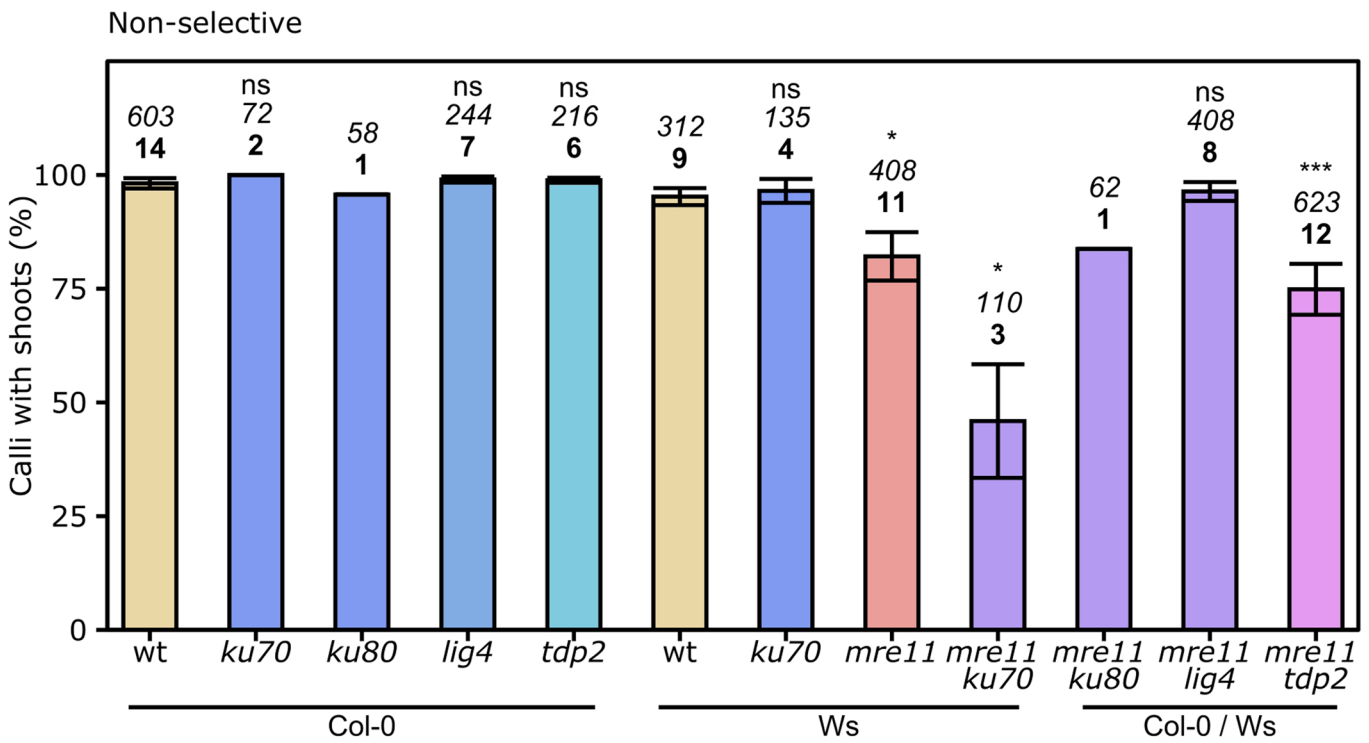


Extended Data Fig. 4 | See next page for caption.

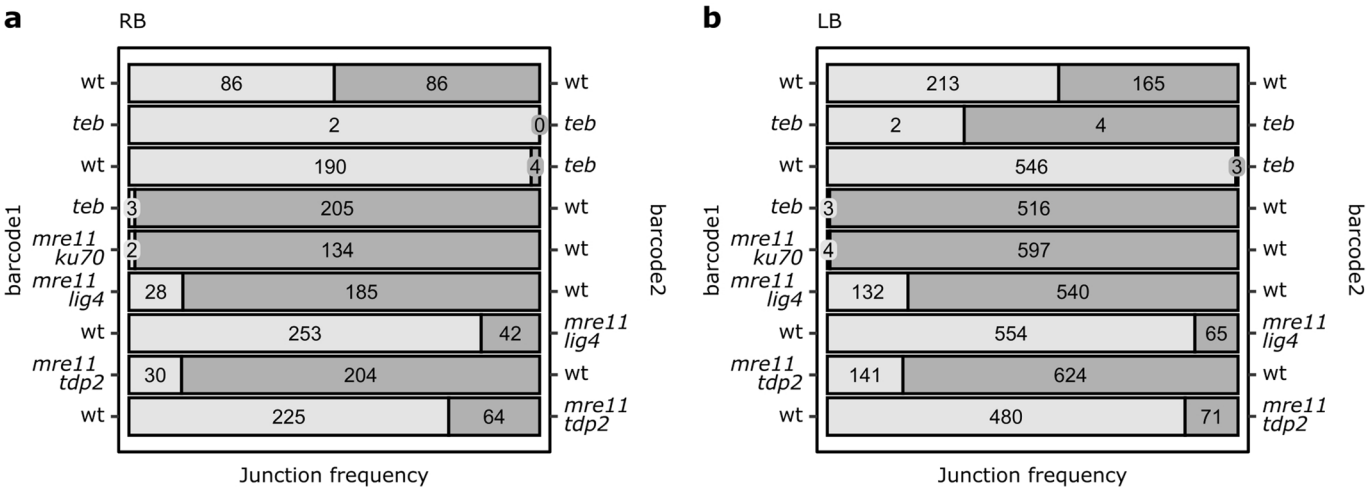
Extended Data Fig. 4 | Homology, filler, and T-DNA loss profiles for cNHEJ mutants. Frequency of different lengths of microhomology (**a** - **d**), filler (**e** - **h**), or T-DNA loss (**i** - **l**) at RB and LB junctions, comparing wt (yellow) with cNHEJ mutants *ku70* and *lig4* (blue). The third colour in each panel indicates the overlapping area. The medians (dashed lines), the number of observations (*n*), and shifts in the mutant distribution relative to wt (s) are indicated. Wilcoxon rank-sum tests were performed to find the direction and the significance of the shifts ($p_{homology\_ku70\_RB} = 7 \times 10^{-38}$, $p_{homology\_ku70\_LB} = 4 \times 10^{-1}$, $p_{homology\_lig4\_RB} = 2 \times 10^{-14}$, $p_{homology\_lig4\_LB} = 5 \times 10^{-1}$, $p_{filler\_ku70\_RB} = 4 \times 10^{-1}$, $p_{filler\_ku70\_LB} = 4 \times 10^{-1}$, $p_{filler\_lig4\_RB} = 4 \times 10^{-3}$, $p_{filler\_lig4\_LB} = 4 \times 10^{-1}$, $p_{deletion\_ku70\_RB} = 3 \times 10^{-28}$, $p_{deletion\_ku70\_LB} = 5 \times 10^{-8}$, $p_{deletion\_lig4\_RB} = 4 \times 10^{-21}$, $p_{deletion\_lig4\_LB} = 1 \times 10^{-12}$). ns: $p \geq 0.05$, *: $p < 0.05$, **: $p < 0.01$, ***: $p < 0.001$.



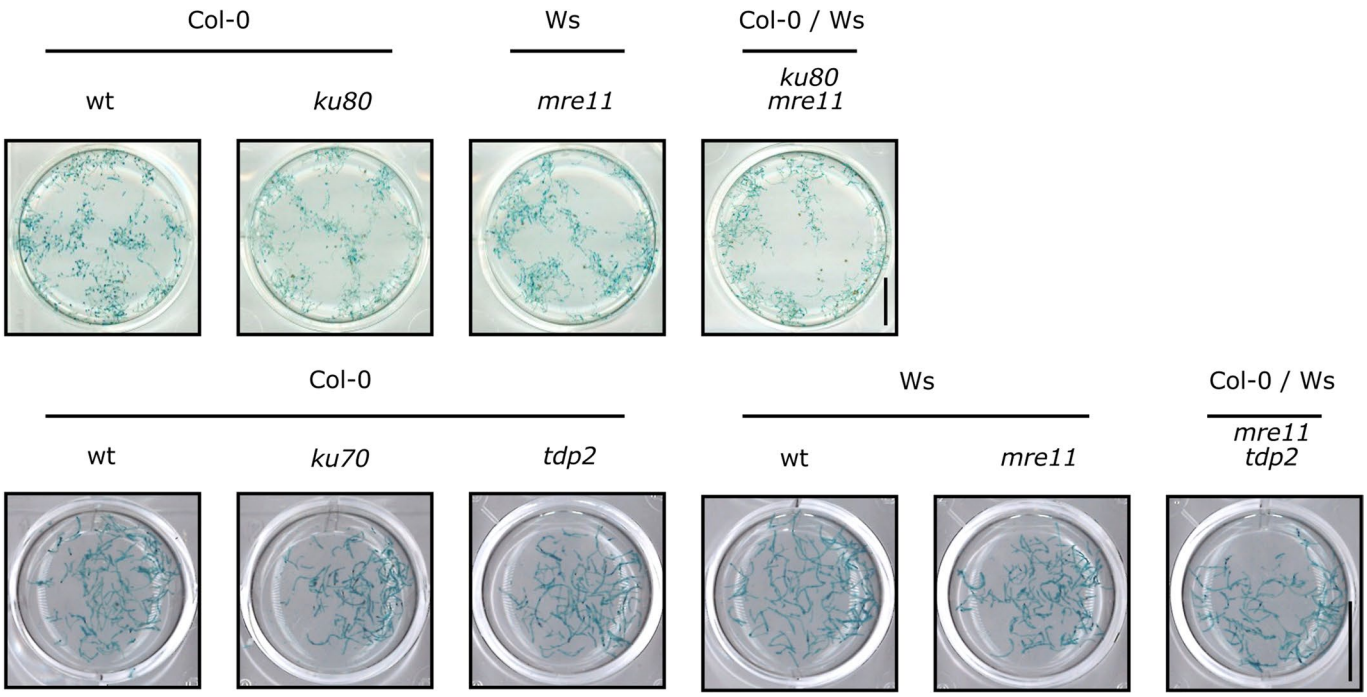
Extended Data Fig. 5 | Homology, filler, and T-DNA loss profiles for *mre11* mutant. Frequency of different lengths of microhomology (**a, b**), filler (**c, d**), or T-DNA loss (**e**, **f**) at RB and LB junctions, comparing wt (yellow) with the *mre11* mutant (light red). The overlapping area is indicated in orange. The medians (dashed lines), the number of observations (*n*), and shifts in the mutant distribution relative to wt (s) are indicated. Wilcoxon rank-sum tests were performed to find the direction (one-sided tests) and the significance of the shifts (two-sided tests, $p_{homology\_RB} = 2 \times 10^{-6}$, $p_{homology\_LB} = 6 \times 10^{-1}$, $p_{filler\_RB} = 8 \times 10^{-6}$, $p_{filler\_LB} = 3 \times 10^{-1}$, $p_{deletion\_RB} = 9 \times 10^{-8}$, $p_{deletion\_LB} = 8 \times 10^{-4}$). ns: $p \geq 0.05$, *: $p < 0.05$, **: $p < 0.01$, ***: $p < 0.001$.



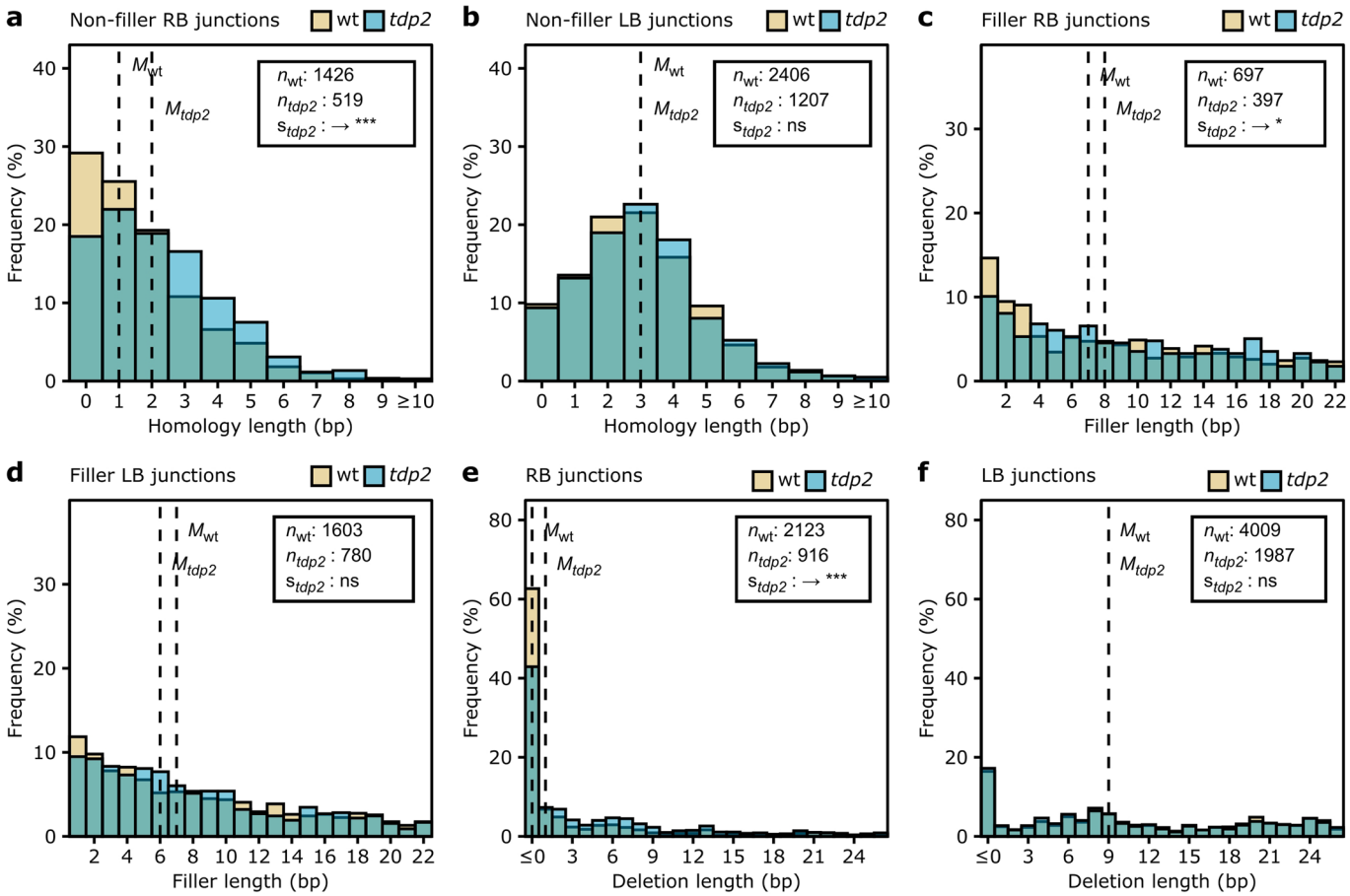
Extended Data Fig. 6 | Regenerative ability. Average percentage of calli with shoot tissue on non-selective plates. The number in italic indicates the total number of calli that were scored for that genotype. The number in bold indicates the number of experiments over which the mean (coloured bars) and standard error of the mean (error bars) were calculated. One-sided Student's t-tests were performed to test for significant reductions in T-DNA integration efficiency of mutant compared to wt ($p_{ku70c} = 9\times10^{-1}$, $p_{lig4} = 7\times10^{-1}$, $p_{tdp2} = 7\times10^{-1}$, $p_{ku70w} = 6\times10^{-1}$, $p_{mre11} = 2\times10^{-2}$, $p_{mre11ku70} = 3\times10^{-2}$, $p_{mre11lig4} = 2\times10^{-1}$, $p_{mre11tdp2} = 8\times10^{-4}$). ns: $p \geq 0.05$, *: $p < 0.05$, **: $p < 0.01$, ***: $p < 0.001$. Mutants were compared to the wt of the same genetic background, except for mutants with a hybrid genetic background, which were compared to the Col-0 wt.



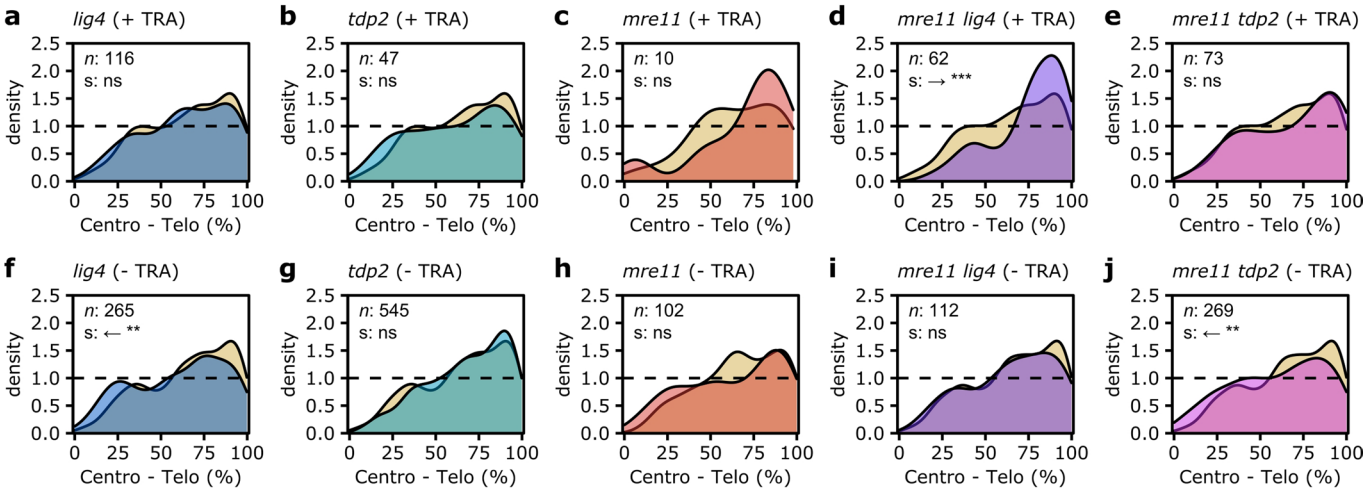
Extended Data Fig. 7 | Comparison of junction numbers by competitive TRANSGUIDE. Number of RB (**a**) and LB (**b**) junctions in competitive TRANSGUIDE, in which equimolar amounts of genomic DNA of two samples with differently barcoded T-DNA (barcode 1 in light grey, and barcode 2 in dark grey) were combined.



Extended Data Fig. 8 | All tested genotypes show (transient) T-DNA expression. Pictures show GUS-stained roots in well plates shortly after co-cultivation with Agrobacterium. The blue colour indicates expression of the T-DNA (pCAMBIA3301). Scale, 1 cm.



Extended Data Fig. 9 | Homology, filler, and T-DNA loss profiles for *tdp2* mutant. Frequency of different lengths of microhomology (**a, b**), filler (**c, d**), or T-DNA loss (**e**, **f**) at RB and LB junctions, comparing wt (yellow) with the *tdp2* mutant (cyan). The overlapping area is indicated in turquoise. The medians (dashed lines), the number of observations (*n*), and shifts in the mutant distribution relative to wt (s) are indicated. Wilcoxon rank-sum tests were performed to find the direction (one-sided tests) and the significance of the shifts (two-sided tests, $p_{homology\_RB} = 7 \times 10^{-11}$, $p_{homology\_LB} = 2 \times 10^{-1}$, $p_{filler\_RB} = 3 \times 10^{-2}$, $p_{filler\_LB} = 7 \times 10^{-1}$, $p_{deletion\_RB} = 2 \times 10^{-24}$, $p_{deletion\_LB} = 6 \times 10^{-2}$). ns: $p \geq 0.05$, *: $p < 0.05$, **: $p < 0.01$, ***: $p < 0.001$.



Extended Data Fig. 10 | Relative genomic position of junctions. Relative frequency of LB junctions after transformation with pWY82 (+ TRA, panels **a**-**e**) or pUBC (- TRA, panels **f**-**j**) along all chromosome arms, comparing wt (yellow) and mutants (other colours). Mutants were compared to wt of the same genetic background, with the exception of the hybrids (*mre11 lig4* and *mre11 tdp2*), which were compared to the Col-0 wt. 0 % indicates centromeric position and 100 % telomeric; *n* indicates the number of mutant junctions. Wilcoxon rank-sum tests were performed to find the direction (one-sided tests) and significance level (two-sided tests) of the shifts (s) in relative position ($p_{lig4+TRA} = 6\times10^{-1}$, $p_{tdp2+TRA} = 4\times10^{-1}$, $p_{mre11+TRA} = 4\times10^{-1}$, $p_{mre11lig4+TRA} = 9\times10^{-4}$, $p_{mre11tdp2+TRA} = 2\times10^{-1}$, $p_{lig4-TRA} = 3\times10^{-3}$, $p_{tdp2-TRA} = 3\times10^{-1}$, $p_{mre11-TRA} = 4\times10^{-1}$, $p_{mre11lig4-TRA} = 7\times10^{-1}$, $p_{mre11tdp2-TRA} = 2\times10^{-3}$). ns: $p \geq 0.05$, *: $p < 0.05$, **: $p < 0.01$, ***: $p < 0.001$. Only junctions that are represented by more than 20 different DNA molecules (thus representing events that are compatible with multiple cell divisions) were included in this analysis.

nature portfolio

Corresponding author(s): Marcel Tijsterman

Last updated by author(s): Mar 22, 2022

Reporting Summary

Nature Portfolio wishes to improve the reproducibility of the work that we publish. This form provides structure for consistency and transparency in reporting. For further information on Nature Portfolio policies, see our Editorial Policies and the Editorial Policy Checklist.

Statistics

For all statistical analyses, confirm that the following items are present in the figure legend, table legend, main text, or Methods section.

n/a	Confirmed
☐	☒ The exact sample size (*n*) for each experimental group/condition, given as a discrete number and unit of measurement
☒	☐ A statement on whether measurements were taken from distinct samples or whether the same sample was measured repeatedly
☐	☒ The statistical test(s) used AND whether they are one- or two-sided *Only common tests should be described solely by name; describe more complex techniques in the Methods section.*
☒	☐ A description of all covariates tested
☒	☐ A description of any assumptions or corrections, such as tests of normality and adjustment for multiple comparisons
☐	☒ A full description of the statistical parameters including central tendency (e.g. means) or other basic estimates (e.g. regression coefficient) AND variation (e.g. standard deviation) or associated estimates of uncertainty (e.g. confidence intervals)
☐	☒ For null hypothesis testing, the test statistic (e.g. *F*, *t*, *r*) with confidence intervals, effect sizes, degrees of freedom and *P* value noted *Give P values as exact values whenever suitable.*
☒	☐ For Bayesian analysis, information on the choice of priors and Markov chain Monte Carlo settings
☒	☐ For hierarchical and complex designs, identification of the appropriate level for tests and full reporting of outcomes
☒	☐ Estimates of effect sizes (e.g. Cohen's *d*, Pearson's *r*), indicating how they were calculated

Our web collection on statistics for biologists contains articles on many of the points above.

Software and code

Policy information about availability of computer code

Data collection	Sequence data was obtained from Illumina MiSeq and NovaSeq machines. Samples were demultiplexed using bcl2fastq2 conversion software v 2.2. Raw paired-end reads with UMIs were combined into consensus sequences, trimmed by TRIMMOMATIC (v 0.39), and mapped to the TAIR10 genome + T-DNA-containing plasmid by BWA-mem (v 0.7.17).
Data analysis	A custom java program was used to call T-DNA-genome junctions (https://github.com/RobinVanSchendel/TRANSGUIDE). Junction filtering and footprint analysis was performed in R (v 4.1.0). Junctions were visually inspected using IGV (v 2.8.0). Figures were assembled in Inkscape (v 0.92).

For manuscripts utilizing custom algorithms or software that are central to the research but not yet described in published literature, software must be made available to editors and reviewers. We strongly encourage code deposition in a community repository (e.g. GitHub). See the Nature Portfolio guidelines for submitting code & software for further information.

Data

Policy information about availability of data

All manuscripts must include a data availability statement. This statement should provide the following information, where applicable:

- Accession codes, unique identifiers, or web links for publicly available datasets
- A description of any restrictions on data availability
- For clinical datasets or third party data, please ensure that the statement adheres to our policy

Mapped sequences are available from NCBI SRA (accession code: PRJNA786733). Junction footprint data is provided in Supplementary Data 2 and 3. The pAC161 data in Extended Data Fig. 2 and 3 is based on previously published data.

Field-specific reporting

Please select the one below that is the best fit for your research. If you are not sure, read the appropriate sections before making your selection.

☒ Life sciences ☐ Behavioural & social sciences ☐ Ecological, evolutionary & environmental sciences

For a reference copy of the document with all sections, see nature.com/documents/nr-reporting-summary-flat.pdf

Life sciences study design

All studies must disclose on these points even when the disclosure is negative.

Sample size	For TRANSGUIDE, per sample 20 large calli were harvested as was the default for root transformation experiments. Our experiments indicated this number was around the saturation point for high-confidence junction calling (so increase is undesirable), and resulted in enough DNA to perform LB and RB TRANSGUIDE on the same sample as well as optional junction validation. Most down-stream analyses included hundreds or even thousands of junctions per comparison group, allowing for the detection of small shifts in distributions. For shoot formation assays the number of calli that were transferred was the maximum that was technically possible. Resulting in at least ~ 100 calli per genotype x construct x root transformation. For GUS staining, hundreds of root explants were processed per sample, and virtually all pieces showed some staining. An increase here would not add any additional information.
Data exclusions	To only analyze high-confidence junctions (and filter out potential artefacts), junctions were filtered for NrAnchorsIncludingPartial >=3, getFractionHighestContributorToMinimumJunction >=0.75. Junctions from different samples occurring on the same position were also filtered out (duplicate positions). In the case of barcode-hopping, the true junction was kept, but the contamination was removed (the true junction had to have at least 10x more anchors than the runner-up). An additional filter was applied to be able to fairly compare data from transformations with different constructs, because distances between primer and border were not constant, and read length limiting: homology needed to be <= 57 bp, filler <= 22 bp, and T-DNA end deletion <= 26 bp. This information is stated in the Method section.
Replication	We followed common practices in the field (2-3 replicates). Findings were additionally reproduced by transformations with different constructs, i.e. highly similar footprints at LB and RB junctions, and similar shoot formation relative to wt .
Randomization	For sample collection calli of the suitable size were collected at random from selective plates. Covariates were controlled by always taking along wild type plants, and when possible always take along also single mutants when studying double mutants. For junction validation, 24 junctions of different types (filler/non-filler, RB/LB, low/high anchor number, etc) were selected randomly.
Blinding	For assaying shoot-formation, all photographs were given code-names and their original labels removed prior to scoring. For junction footprint analysis unbiased programs were used and so blinding was not applicable here. For GUS staining no blinding was done as the result was all-or-nothing.

Reporting for specific materials, systems and methods

We require information from authors about some types of materials, experimental systems and methods used in many studies. Here, indicate whether each material, system or method listed is relevant to your study. If you are not sure if a list item applies to your research, read the appropriate section before selecting a response.

Materials & experimental systems

n/a	Involved in the study
☒	☐ Antibodies
☒	☐ Eukaryotic cell lines
☒	☐ Palaeontology and archaeology
☒	☐ Animals and other organisms
☒	☐ Human research participants
☒	☐ Clinical data
☒	☐ Dual use research of concern

Methods

n/a	Involved in the study
☒	☐ ChIP-seq
☒	☐ Flow cytometry
☒	☐ MRI-based neuroimaging