the plant journal

SEB Society for Experimental Biology

The Plant Journal (2017) **92,** 57–67 doi: 10.1111/tpj.13634

Endogenous sequence patterns predispose the repair modes of CRISPR/Cas9-induced DNA double-stranded breaks in *Arabidopsis thaliana*

Giang T. H. Vu[1,*], Hieu X. Cao[1,†], Friedrich Fauser[2,‡], Bernd Reiss[3], Holger Puchta[2] and Ingo Schubert[1,*]

[1]*Leibniz Institute of Plant Genetics and Crop Plant Research (IPK), D 06466 Gatersleben, Stadt Seeland, Germany,*

[2]*Botanical Institute II, Karlsruhe Institute of Technology, POB 6980, Karlsruhe 76049, Germany, and*

[3]*Max Planck Institute for Plant Breeding Research, 50829, Köln, Germany*

Received 29 May 2017; revised 3 July 2017; accepted 7 July 2017; published online 11 July 2017.

*For correspondence (e-mails vu@ipk-gatersleben.de or schubert@ipk-gatersleben.de).

†Present address: Institute of Biology, Martin-Luther-University Halle-Wittenberg, Weinbergweg 10, D 06120, Halle, Germany.

‡Present address: Department of Molecular Biology, Princeton University, 119 Lewis Thomas Laboratory, Princeton, NJ 08544, USA.

SUMMARY

The possibility to predict the outcome of targeted DNA double-stranded break (DSB) repair would be desirable for genome editing. Furthermore the consequences of mis-repair of potentially cell-lethal DSBs and the underlying pathways are not yet fully understood. Here we study the clustered regularly interspaced short palindromic repeats (CRISPR)/Cas9-induced mutation spectra at three selected endogenous loci in *Arabidopsis thaliana* by deep sequencing of long amplicon libraries. Notably, we found sequence-dependent genomic features that affected the DNA repair outcome. Deletions of 1-bp to <1000-bp size and/or very short insertions, deletions >1 kbp (all due to NHEJ) and deletions combined with insertions between 5-bp to >100 bp [caused by a synthesis-dependent strand annealing (SDSA)-like mechanism] occurred most frequently at all three loci. The appearance of single-stranded annealing events depends on the presence and distance between repeats flanking the DSB. The frequency and size of insertions is increased if a sequence with high similarity to the target site was available in *cis*. Most deletions were linked to pre-existing microhomology. Deletion and/or insertion mutations were blunt-end ligated or via *de novo* generated microhomology. While most mutation types and, to some degree, their predictability are comparable with animal systems, the broad range of deletion mutations seems to be a peculiar feature of the plant *A. thaliana*.

Keywords: amplicon sequencing, DNA double-stranded break repair, genome stability, site-directed mutagenesis, homology-directed repair (HDR), non-homologous end-joining (NHEJ).

INTRODUCTION

The RNA-guided Cas9 nuclease from the bacterial clustered regularly interspaced short palindromic repeats (CRISPR) adaptive immune system provides a powerful and versatile tool for genome engineering of eukaryotic organisms, including plants (Ran, 2013; Bortesi and Fischer, 2015; Cao *et al.*, 2016; Pacher and Puchta, 2016; Puchta, 2017). By simply designing a 20-nt targeting sequence of the guide RNA (Jinek *et al.*, 2012; Cong *et al.*, 2013; Mali *et al.*, 2013), the CRISPR/Cas9 can be targeted to create chromosomal double-stranded breaks (DSBs) at specific sites. To achieve a desired editing outcome, components of the appropriate DNA repair pathway need to be recruited. Recent investigation on DNA repair profiles of CRISPR/Cas9-induced DSB sites in the human genome demonstrates that the resulting DNA repair outcomes are non-random and depend on the target site sequence (van Overbeek *et al.*, 2016). Little information is known about how the DNA sequence pattern of the target locus defines the suitable DNA repair machinery to fix the break or about whether one can precisely predict the DNA repair outcome of any designed CRISPR/Cas9-based genome editing experiment.

In eukaryotic somatic cells, DSBs are typically repaired either by non-homologous end-joining (NHEJ) pathways (Lieber, 2010; Sfeir and Symington, 2015) or by homology-directed repair (HDR) pathways (Puchta, 2005; Heyer *et al.*, 2010; McVey *et al.*, 2016). Non-homologous end-joining pathways require none or short homologies (1–20 bp) and cause deletions and/or insertions of, in many cases, a few base pairs only. However, (micro)homology-mediated

synthesis-dependent (MM-SDSA-like) mechanism(s) can combine deletions with insertions (often copies of nearby sequences) (Gorbunova and Levy, 1997; Salomon and Puchta, 1998; Yu and McVey, 2010); both extending well into the kbp range in plants (Vu *et al.*, 2014, 2017). In contrast, HDR requires extensive DNA end processing and an extended DNA sequence homologous to the target site as a template. While NHEJ is faster, more efficient and active throughout the cell cycle (Mao *et al.*, 2008), HDR is likely to be more prevalent after DNA replication because of the presence of a sister chromatid as a repair template (Hustedt and Durocher, 2017). HDR can yield a sequence conversion via synthesis-dependent strand annealing (SDSA) and/or a reciprocal exchange between identical (or homologous) double helices and typically restores the broken double strand without error. However, when homologous (repetitive) sequences are present in *cis* in the vicinity of the DSB site, a non-conservative HDR mechanism, the so-called single-stranded annealing (SSA) pathway, can result in a large deletion comprising (at least) of one of these homologous repeats and the sequence between them (Figure 1a). As demonstrated using I-*Sce*I-meganuclease-based reporter systems, SSA (Siebert and Puchta, 2002) and SDSA (Puchta, 1998) pathways can be very efficient in somatic plant cells. In terms of application, SSA-based genome manipulation is important for excision of unwanted genomic fragments, marker genes, or transgenes (Aryan *et al.*, 2013), while homologous recombination via SDSA enables gene replacement at chromosomal loci of, at least partial, sequence homology to the transgene (Moehle *et al.*, 2007). The well established CRISPR/Cas9-based technology facilitated gene replacement via homologous recombination in yeast, insect and other animal models and even in higher plants (Steinert *et al.*, 2016), as well as gene editing (site-directed mutagenesis) in all tested plants (Bortesi and Fischer, 2015; Osakabe *et al.*, 2016; Pacher and Puchta, 2016; Zhao *et al.*, 2016; Shen *et al.*, 2017). Nevertheless, more detailed knowledge is needed to understand and predict the modes of CRISPR/Cas9-induced genome editing at distinct endogenous genomic positions in plant and other models.

By using amplicon deep sequencing, we previously explored the outcome of I-*Sce*I-induced DSBs at three β-glucuronidase (GUS) reporter transgene variants (GU.US, IU.GUS, and DU.GUS) (Puchta and Hohn, 2012) regarding the proportions of SSA, SDSA and NHEJ pathways, in barley and in *Arabidopsis thaliana* (Vu *et al.*, 2014, 2017). Here we designed CRISPR/Cas9-mediated cleavage to detect, in addition to NHEJ and SDSA-like pathways, preferentially SSA (with two different spaces between homologous repeats) or SDSA events (involving ectopic homology) at distinct positions on *A. thaliana* chromosomes 3 and 5 (Figure 1). Pacbio-based DNA repair profiling shows a concordant result for endogenous and transgenic reporter sequences, i.e. similar mutation class frequencies. These data and the impact of the sequence context around the break sites, which we found comparing the size and frequency of deletions and inversions between three different target sites, suggest predictable outcomes of genome manipulation approaches. Comparing these results with

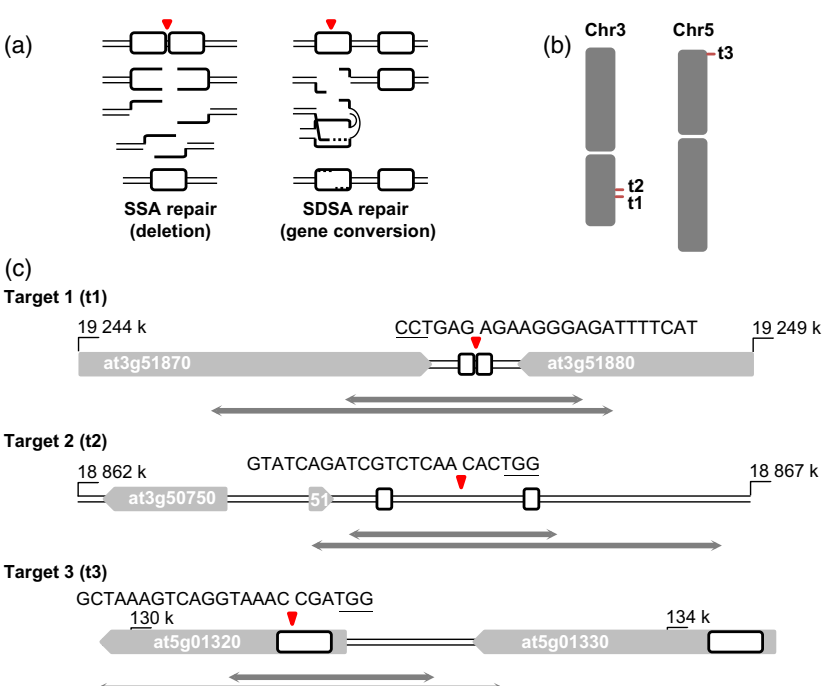

Figure 1. *A. thaliana* loci studied for the repair products of Cas9-mediated DSBs.
(a) Scheme of DSB repair by SSA (targets 1 and 2) and SDSA (target 3) pathways at selected loci.
(b) Chromosomal positions of the three target loci for DSB induction within the *A. thaliana* genome.
(c) The sgRNAs (PAM underlined) designed for the three targets, DSBs = red triangles, grey bars = coding sequences (at3 g51870; at3 g51880; at3 g50750; at3 g50751; at5 g01320; at5 g01330), white boxes = repeats, double arrows = amplicons for sequencing.

the only other large repair profile data available (van Overbeek *et al.*, 2016), suggests, in addition to predictability for both species, clear differences in DSB repair outcome between the plant species *A. thaliana* and human cells.

RESULTS

DNA sequence patterns around the break site define the repair profile of CRISPR/Cas9 induced DBSs

In Arabidopsis, SSA is the prevalent mode of repair for DSBs induced between direct repeats in the transgenic GU.US construct (Orel *et al.*, 2003; Vu *et al.*, 2017). To investigate how the DNA sequence patterns of endogenous loci affect the DNA repair outcome, we designed CRISPR/Cas9-mediated cleavage between homologous repeats at different loci to examine endogenous repair profiles of CRISPR/Cas9 induced blunt-ended DSBs.

The first selected target site for DSB induction and subsequent SSA repair comprises intergenic sequences on *A. thaliana* chromosome 3 that contain two directly adjacent repeats of 111 bp each (Figures 1 and S1; see Experimental Procedures). There are only two single nucleotide polymorphisms (SNPs) that distinguish both copies to ensure annealing between two single-stranded homologous DNA sequences (Figure S1). DSBs induced by CRISPR/Cas9 endonucleases were precisely at the connection between these two tandem repeats (Figure 1). In order to validate the repair profile at this target site, we designed and massively sequenced by single molecule sequencing (PacBio) amplicon libraries of 1770 and 2994 bp, which enabled the detection of deletions or insertions of more than 1000 bp on either side of the break site, respectively (Figure 1c). The results are merged in Table S3.

Of 20 178 informative repair product sequences called 'reads', 18 694 (92.6%) represent the either uncut or precisely restored original sequence. For the remaining 1484 sequences, the mutation spectra of repair events were analysed. We found that indeed SSA repair was favored, representing 1203 (81.1%) of the mutated reads. This result is consistent with our previous observation that SSA is the predominant repair path for the GU.US construct in transgenic *A. thaliana* plants. Furthermore, we found 72 reads (4.85%), belonging to 15 out of 31 mutation classes (each class representing identical reads of a specifically mutated sequence), that showed deletions larger than 1000 bp (ranging from 1104 to 1995 bp) (Table S3; for the remaining mutated reads see below). This observation provokes the question whether breaks between repeats separated by ~1 kbp also yield deletions of >1 kbp and which proportion of these deletions represent SSA events. For studying the potential effect of the distance between repeats flanking the DSB on the repair outcomes, we selected a second target site (target 2) located 297 kbp proximal to the first target on the chromosome 3. This site contains two direct repeats of 97 bp (with only one single mismatch nucleotide) on either side of the gene AT3g50755 (Figures 1 and S2). The DSBs were induced within the AT3g50755 sequence just in the middle between the two repeats. In this situation, SSA repair should cause a deletion of 1004 bp (Figure S2). To assess the outcome of the DSB repair, again, the amplicon sequencing data from two libraries of 1551 bp and 3048 bp, respectively, were generated and analysed (Figure 1c and merged results in Table S4). Interestingly, of 9355 imprecisely repaired sequence reads comprising 33 mutation classes, only 297 reads (3.2%), belonging to five mutation classes, showed deletions of >1000 bp. Remarkably, however, 268 of these reads were repaired by SSA events (deletion 1004 bp), while the other 29 reads are shared by four mutation classes. In general, the number of reads probably overestimates the number of events, because early events could via replication have amplified the corresponding sequence. Conversely, the number of mutation classes underestimates the number of events concerning preferred pathways. The abundance of SSA reads exemplifies this preference, if their occurrence is not in both cases (target 1 and 2) by chance due to an extremely early event. While for target 2, 29 reads of four mutation classes represent deletions >1000 bp, the class of SSA (deleting 1004 bp) comprises 268 reads. Assuming that the four large deletion classes, which are not due to SSA, represent one single event each would mean that, on average ~7 reads (29 reads divided by four mutation classes) correspond to one event causing a deletion of >1000 bp. Under these circumstances, 38 or more SSA events would have occurred. For target 1, 25 mutation classes share 134 reads representing deletions. An average of about five reads per deletion event (134 divided by 25) would mean in target 1 ~53 SSA events might have occurred. Actually, for target 2 the frequency of SSA amounts to 2.9% of all mutated reads but to 90.2% of reads with deletion larger than 1000 bp. These results indicate that the DNA repair outcome of well designed CRISPR/Cas9-based genome editing is predictable. A negative correlation between deletion size and frequency most likely limits the absolute (but not the relative) frequency of SSA repair events, due to an increasing requirement for end resection with increasing distance between direct repeats flanking the broken locus. Taken together, a predisposition for distinct DSB repair modes in dependence of the sequence context at the target locus is given by the presence (and their distance) of repeats around the break (Figure 2), as well as by an apparently species-specific deletion size range. A preference for larger deletions as displayed on average by *A. thaliana* compared with the situation observed for tobacco (Kirik *et al.*, 2000), barley (Vu *et al.*, 2017) or human cells (van Overbeek *et al.*, 2016) enables a higher abundance of SSA between repeats over increasing distances.

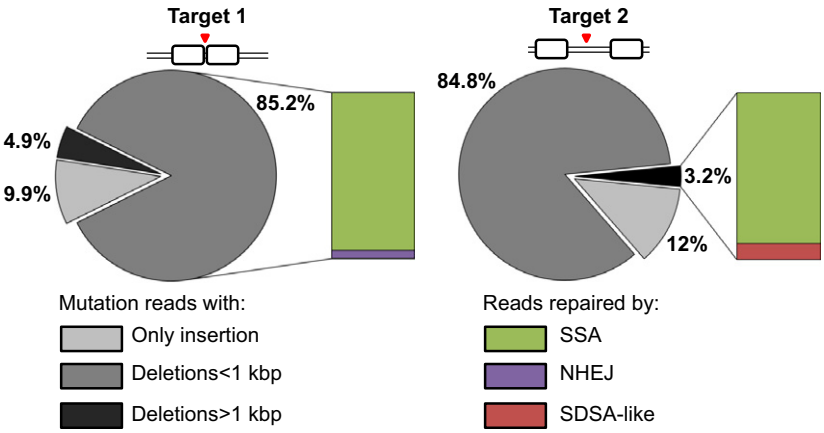

Figure 2. Predictability of DSB repair pathways depending on the sequence context.
For both targets the proportions of reads with only insertions, deletions <1 kbp or >1 kbp are shown in different variants of grey, the relative contribution of different repair types within the deletion fractions harboring SSA events in different color. While the proportions of only small insertions as well as of deletions below and above 1 kbp are rather similar, a negative correlation between deletion size and frequency most likely limits the absolute (but not the relative) frequency of SSA repair events in case of target 1 (DSB between adjacent tandem repeats) compared with target 2 (DSB between two repeats 997 bp apart from each other).

Local homologous sequences initiate longer insertions

Of the 31 mutation classes found for the first target locus that was chosen for detection of SSA, 18 (58.1%) revealed insertions (three classes with 1 bp inserted each) or insertions combined with deletion. For the second locus, with a distance of 907 bp between the repeats, 18 (54.4%) out of the 33 observed mutation classes revealed either insertions (again three classes with one inserted bp) or insertions and deletions together. A detailed analysis of individual mutation classes revealed that the inserted sequences seemed to be templated by sequences near the break ends, thus representing repetitions of these sequences via synthesis primed by microhomology with the respective break end (Table S3). The average insertion sizes calculated based on the number of repair classes are similar for both target sites (28.6 bp and 23.5 bp, respectively). There are only two mutation classes with insertions larger than 80 bp observed for each of these two targets. The maximum insertion size of 225 bp at target 1 (deletion 1920/insertion 225; Table S3) was presumably templated entirely by the homologous sequence of the sister chromatid because the 1920 bp, which were deleted from the mutated site, included the 225 bp of the insertion.

To gain insight into the natural situation in which an unbroken homologous sequence occurs on the same double helix (i.e. the same chromatid, see Figure 1a) and could influence the repair outcome by interaction with the break ends, we used a CRISPR construct that generates a DSB within a 1277 bp sequence (target 3). This sequence has a homologous copy of 1432, 1816 bp apart, differing in size from the 1277 bp by a few indels (Figure S3). In this target locus, the DSB is *within* a repeat sequence and not *between* the homologous repeats as in targets 1 and 2 (Figure 1). We amplified repair products as libraries of 1531 bp and 3038 bp covering the break point. To find out how sequence homology is used as a template for repair synthesis in this situation, we sequenced in-depth the repair products from both libraries of the endogenous target locus 3 (merged results in Table S5) in comparison with targeted loci 1 and 2.

Indeed, although the overall mutation types are similar as for target loci 1 and 2 (Figure 3a), there are more mutations with insertions which resulted from SDSA-like type of repair (Figure 3b). We found 52 (75.7%) out of 66 mutation classes with insertions (52 bp on average), among these, nine classes showed insertions >80 bp (Figure 3c). The largest insert of 554 bp, originated via four copy switches combining five sequences from different positions (114 bp of the insertion overlap partially with the upstream part of the deleted region and are obviously copied from the homologous sequence of the sister chromatid; of the remaining 440 bp, 66 are from the targeted repeat downstream the deletion, while 339, 14 and 21 bp are from the ectopic copy, as recognized by the polymorphisms between both repeats). Compared with the other two loci, the insertions were larger in average and the number of mutation classes with large insertions was increased more than two-fold (from ~6 to 13.6%). This change is probably due to the presence of homology to the break end not only within the sister helix (which is available as repair template only between replication and nuclear division), but also in *cis* (available as repair template all over the cell cycle). The genomic distribution of sequences homologous to the break site in *cis* significantly skews the size of insertions, further indicating that the sequence context around the break position influences the repair outcome (Figure 3).

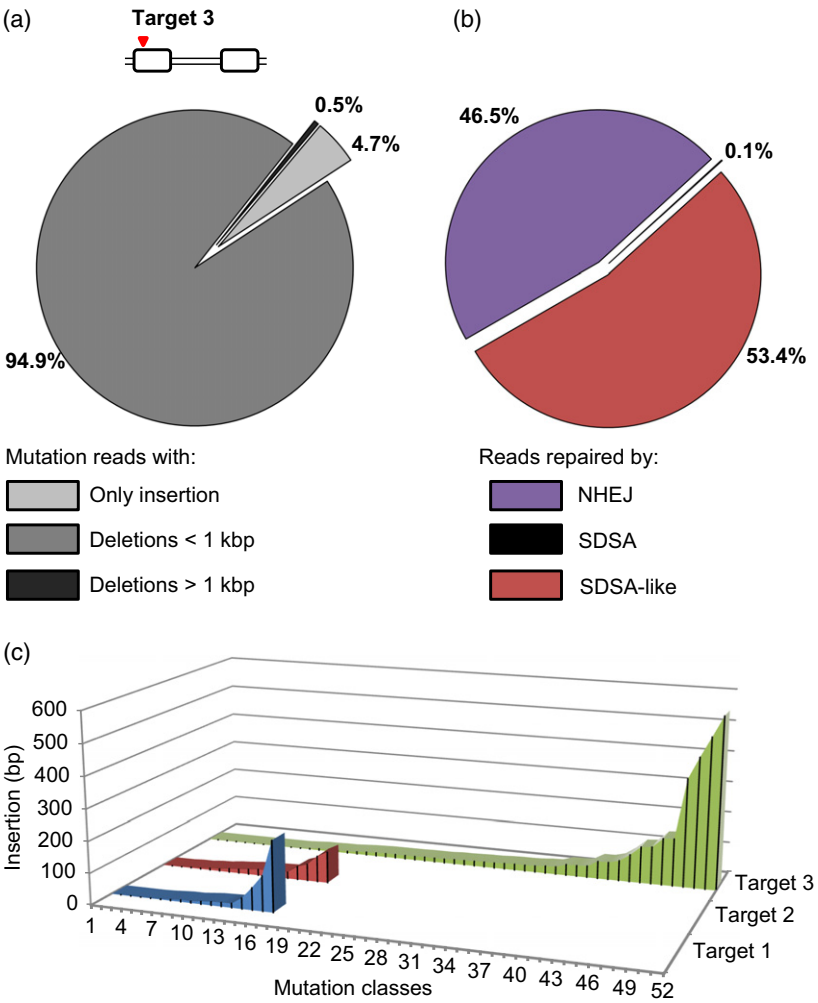

Figure 3. The sequence context at the targets site affects the outcome of DSB repair.
(a, b) (a) The proportions of reads with only insertions, deletions <1 kbp or >1 kbp are shown in different variants of grey. The repair outcome at target 3 is similar to targets 1 and 2 concerning the deletion size and mutations with insertion only, however, (b) the sequences homologous to the break site in *cis* significantly increased the mutation classes with insertions resulting from SDSA and SDSA-like mechanisms (the relative contribution of different repair types are shown in different color).
(c) Frequency and size of mutation classes with insertions of the three targets. *X*-axis: number of mutation classes; *Y*-axis: insertion sizes in bp during imprecise repair; *Z*-axis: the three investigated target sites (blue = target 1, red = target 2, green = target 3). Extended homology in *cis* of the target site (as in target 3) is usable as a template for repair synthesis and initiates more and longer insertions than in targets 1 and 2, where the breakpoint does not reveal homology in *cis*.

Repair of endogenous loci predominantly depends on microhomology for tethering break ends

Error-prone repair processes involve ligation of DNA ends that have lost and/or gained nucleotides prior to ligation. Repair outcomes at all three investigated targets demonstrated that NHEJ as well as HDR repair pathways can be highly mutagenic (Tables S3–S5). We now focus on how the mutagenic break ends were rejoining.

Typically, DNA break ends are resected creating single-stranded overhangs. Short microhomologies on either single-stranded end may subsequently anneal to rejoin the break either by classical NHEJ (using 1–5 bp microhomology) or by microhomology-mediated end-joining (MMEJ using 6–20 bp microhomology; Sfeir and Symington, 2015). The repair using pre-existing microhomologies after resection of blunt ends always creates products containing deletions. Of 11 mutation classes of target 1 containing only deletions (SSA excluded), seven classes revealed 2–4 bp pre-existing microhomology for end-joining repair, and three mutant classes had 1 bp of homology. Similar results were obtained for targets 2 and 3. Seven out of 14 (target 2) and six out of 14 classes (target 3) were joined using 2–4 bp microhomology. Smaller fractions (five and three mutation classes) have been driven by 1 bp microhomology. Only one class at target 2 involved 9-bp microhomology for annealing. The remaining mutation classes with deletions (one for targets 1 and 2, and five for target 3) revealed joined blunt ends or are based on *de novo* synthesis of microhomologies at the deleted end, generating complementarity to the opposite break end (Figure 4a). In contrast to deletions linked with pre-existing microhomologies, the broken ends in mutation classes combining insertions (duplications of nearby flanking sequences) and deletions are either directly joined or via *de novo* synthesized microhomologies (Figure 4b). We analysed the potential use of nascent microhomologies of these remaining mutation classes at the three targets. Intriguingly, nearly all these mutation classes – if not blunt end joined – involved *de novo* synthesized microhomologies for annealing, as exemplified

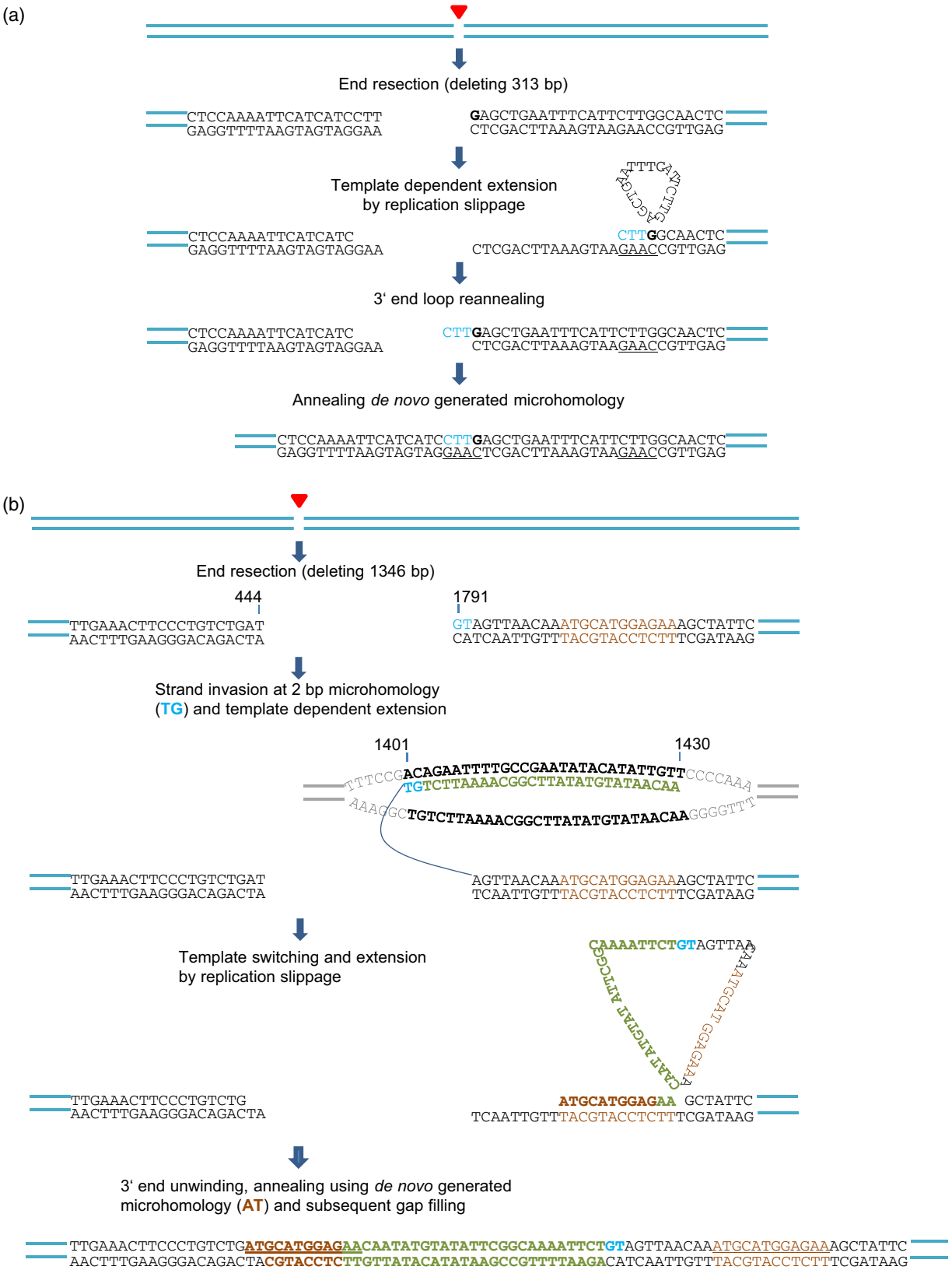

Figure 4. Examples of DSB repair by NHEJ and by SDSA-like repair, respectively.
(a) Mutation class 26 of target 2: NHEJ after 313 bp deletion (308 bp on the left and 5 bp on the right) using de novo-generated microhomology. The 3′ end at the right break site unwinds, forms a loop, extends from G for TTC (=*de novo*-generated microhomology) and re-anneals with the GAA overhang of the resected left break site before ligation. Alternatively, one has to assume blunt end ligation as the second step after deletion.
(b) Mutation class 17 of target 1: SDSA-like repair after 1346 bp deletion (position 444 to 1791; 767 bp on the left and 579 bp on the right). The right 3′-end nucleotides TG invade the sister helix opposite the deleted region and re-synthesize 27 bp (green); the last two nucleotides AA undergo a copy switch and re-synthesize 10 bp (brown) in *cis* (or in *trans*) generating *de novo* 2 bp microhomology before re-annealing with the resected left break end TA, gap filling and subsequent ligation. The SDSA-like mechanism creates an insertion of 35 bp and sequence duplication of 12 bp (underlined). Alternatively, after the copy switch only 8 bp were re-synthesized before blunt end ligation.

in Figure 4. For 124 of 131 investigated mutation classes of the three targets, we detected ligation linked with microhomology. Together, our data suggest that the utilization of (micro)homologies is predominant for end ligation of endogenous breaks by NHEJ. While pre-existing microhomologies lead to only deletions, more complex alterations linking deletions with duplicative insertions of nearby sequences (Gorbunova and Levy, 1997) might be linked via *de novo*-generated microhomologies.

DISCUSSION

Target sequences potentially allow predictions regarding the outcome of DSB repair

From previous studies on transgenic reporter constructs in plants (Kirik *et al.*, 2000; Vu *et al.*, 2014, 2017) it is known that DSBs are frequently mis-repaired, for instance in *A. thaliana*, tobacco and in barley. In addition to specific SSA and SDSA events, the spectrum of deviations from the original sequence comprised base substitutions (very rare), deletions from 1 to >1000 bp and small single insertions (mostly 1 bp), the latter two mutation types due to NHEJ. An interesting class of events combined deletions of varying size with insertions also varying in size, but mostly <100 bp and clearly shorter than the deletions. This type has been described for tobacco (Gorbunova and Levy, 1997; Salomon and Puchta, 1998) and is similar to a phenomenon reported for *Drosophila* (Yu and McVey, 2010). It starts with DNA synthesis at a (resected) break end, primed by microhomology, from a template near the break site. Then, homology ends abruptly (sometimes even template switching may occur) and the junction continues with the other break end after sealing by NHEJ. Such events do not restore the target locus based on homology. They were considered to be 'templated insertion' (Yu and McVey, 2010; Shen *et al.*, 2017) or as a microhomology-mediated SDSA-like type of repair (Gorbunova and Levy, 1997; Salomon and Puchta, 1998; Vu *et al.*, 2014) which occurs frequently in plants and apparently also in mammals (Hartlerode *et al.*, 2016).

Now we have studied the repair products of targeted DSBs at three different endogenous loci within the *A. thaliana* genome, which were selected to detect SSA (targets 1 and 2) and SDSA (target 3), in analogy to previous studies on the transgenic constructs GU.US (for detection of SSA) and IU/DU_GUS (for detection of SDSA).

Analysis of the mutated reads revealed that CRIPSPR/Cas9 induces the same mutation spectra at endogenous sequences as other site-specific endonucleases at transgenic target loci. Our main aim was to investigate the potential impact of sequence context on frequency and type of mutagenic repair pathways. The expected insight is of practical importance for genome editing approaches, because it is of interest to what degree the results of genome editing can be predicted on the basis of the target site sequence. A study with the same aim has recently been published for 223 human genomic positions in human cells (van Overbeek *et al.*, 2016). These authors reported target-specific profiles of indel mutation classes which were highly reproducible for each target between different cell lines. Therefore they concluded that, even in the absence of (external) donor sequences, the repair outcome is not random, but is predictable for the target site by the protospacer sequence. The mutation classes revealed indels of <3 bp resulting from NHEJ to be dominant. Additionally, classes of larger deletions, likely resulting from MMEJ, were found. Deletions rarely exceeded 20 bp and insertions were seldom larger than 4 bp (1-bp insertions were predominant). However, the detection window was only 50 bp on either side of the break, therefore larger indels would have escaped from detections. In most experiments, wild-type reads appeared at similar frequency as mutated ones, but in a few cases wild-type reads were several-fold more abundant. Similar results were described for a smaller number of target sites in human cell lines by Tan *et al.* (2015).

Our comparison of mutation class frequency between the three target sites showed that deletions of varying size and very short insertions, both due to NHEJ, are most abundant for targets 1 and 2, followed by mutation classes resulting from the MM-SDSA-like pathway (Figure 2). SSA events occur when the DSB is targeted between two repeats. This situation was previously found in transgenic constructs (GU.US) with a short distance between the repetitive sequences. Here we could show that the absolute frequency of SSA decreased with the distance between the repeats, i.e. the extension of end resection required for SSA. Nevertheless, the relative frequency of SSA among the reads with deletion of similar size was very high (>80%).

The actual frequency of SDSA at endogenous target sites is difficult to estimate, because SDSA products are

indistinguishable from the original sequence if the identical repeat of the sister chromatid was the template. However the proportion of SDSA cannot exceed the proportion of reads with the original sequence. In our experimental system only SDSA events involving the ectopic repeat can be detected by the presence of SNPs in the reads, as in target 3. From this target we found indeed one mutation class (52) with nine reads that restored the target sequence, deviating from the original one only by SNPs and indels within the undamaged ectopic template repeat (Figure 3b and Table S5). Remarkably, we uncovered two mutation classes (with two reads each) of incomplete SDSA in which, most likely, the newly synthesized strand was linked with the opposite break end by NHEJ, similar to ectopic or one-sided targeting (Puchta, 1999; Wendeler *et al.*, 2015; Watanabe *et al.*, 2016). Thus, between 0.14 to 26.8% of reads and 4.5 to 6% of mutation classes are based on SDSA, depending on whether only the ectopic sequence served as repair template, or all original sequence reads originated from SDSA too, using the identical sister sequence as template. In addition, 37 (56.1%) out of 66 mutation classes represent SDSA-like events of which the insertion from the sister or the ectopic repeat is flanked on either side by sequence deletion. Furthermore, we discovered that the extended homology to the break site in *cis* (target 3), which (in contrast to the sister chromatid) is available all over the cell cycle for repair synthesis, may serve as an additional potential template not only for the SDSA pathway but in general for more (and longer) insertions compared with the situation in which a DSB is located between two repetitive sequences (templates 1 and 2; Figure 3c). Apparently, except for direct ligation, a DSB can be restored via SDSA or in a more degenerated version by SDSA, which is only at one side homology-directed or, even more degenerating, by an SDSA-like mechanism. Alternatively, DSBs can be repaired yielding only deletions of different size based on none, short (NHEJ, MMEJ) or extended (SSA) homology on either break end. Which one of the different repair pathways is favored is apparently predisposed by the sequence context at the target locus.

Some repair pathways might be intrinsic for all target positions of an organism

In spite of the predictable DSB repair outcome regarding SSA or SDSA, which is linked with the repeat composition of the target locus, the abundance of very short insertions, and of deletions below or above 1 kb is remarkably similar between the target positions (Figures 2 and 3a, b). Although only a few species have been investigated regarding preferential deletion and/or insertion sizes, it is tempting to speculate that specific modes of DSB repair are typical for (groups of) organisms. For instance, a bias of DSB repair either towards deletions in organisms with very small (<500 Mbp) genomes (Ibarra-Laclette *et al.*, 2013; Vu *et al.*, 2015) or towards insertions within expanding genomes has been hypothesized (Kirik *et al.*, 2000; Schubert and Vu, 2016). In fact, *A. thaliana* displayed a clear net loss of DNA during DSB repair, while for the 35-fold larger barley genome loss and gain of DNA were nearly balanced (Vu *et al.*, 2017).

Open questions

Recent studies found DNA polymerase theta (pol θ) was responsible for microhomology-mediated and aborted homology-directed repair in nematodes (Koole *et al.*, 2014) and in mammalian cells (Wyatt *et al.*, 2016). An ortholog of pol θ was found in *A. thaliana* to be involved in T-DNA integration and templated insertion at T-DNA integration sites (van Kregten *et al.*, 2016). The possibility that pol θ acts as a component of MM-SDSA-like error-prone DSB repair, which competes with or substitutes the more correct homology-directed mechanisms, remains to be proven experimentally.

Whether individual mutation classes (except SSA), which display very high read numbers (~1000 or more) in one of the target sites (for target 2: class 3 = D1, class 14 = D1–I20, class 24 = D51, class 31 = D581; for target 3: class 5 = D3–I1, class 35 = D24–I21, class 61 = D28; see Tables S4 and S5), represent random events that occurred very early in development and were amplified by cell divisions, or are frequent and predictable sequence-specific events, as the cases observed in human cells (van Overbeek *et al.*, 2016), remains to be clarified in future experiments.

The amplicon libraries revealed for the three target sites different proportions of reads with the original sequence. While target 2 showed 16.6% and target 3 26.8%, target 1 displayed 92.6% of original sequence reads. Because it is not yet possible to assess directly to what proportion the non-mutated reads represent uncut or precisely repaired sequences, it remains unclear whether target 1 was less efficiently targeted by the endonuclease, or more efficiently repaired in an error-free manner (by either direct ligation, SDSA or sister chromatid exchange) than targets 2 and 3. So far we have found no sequence peculiarity (e.g. varying GC content) that would hint to potential differences between the target loci regarding euchromatic or heterochromatic features, which could be responsible for an altered CRISPR/Cas9 accessibility due to different amounts of methylated C. This is also in line with a recent study in Arabidopsis in which no major differences in the repair patterns of paired nicks induced by the CRISPR/Cas system were found between eu- and heterochromatic sites (Schiml *et al.*, 2016). An occasional over-dominance of wild-type sequence was also found among reads from individual target sequences in human cell lines (van Overbeek *et al.*, 2016). Thus it remains to be solved in future whether there is a tendency for correct restoration of the

original sequences at (a) specific genomic context(s) and, if yes, which pathway (direct ligation, SDSA or sister chromatid exchange) is involved.

CONCLUSION

The presented results led to the following conclusions:

(i) CRISPR/Cas9-mediated blunt end DSBs at endogenous loci cause similar mutations as I-*Sce*I-mediated DSBs with staggered ends at transgenic loci.
(ii) The relative frequency of some mutation classes is similar between different target loci (Figures 2 and 3a).
(iii) The size of DSB-mediated deletions and insertions seems to be species specific and genome-size dependent (Vu *et al.*, 2017).
(iv) The sequence context (presence and arrangement of repeats) around the break has an impact on the repair pathway used, and thus, on the outcome of DSB repair.
(v) DSBs between repetitive sequences facilitate SSA events resulting in larger deletions; their absolute (but not relative) frequency decreases with the distance between repeats.
(vi) Frequency and size of insertions caused by an SDSA-like mechanism increase when breaks occur within a repeat, and a second copy is available in *cis*.
(vii) Most of the mutagenic repair events use pre-existing or (possibly) *de novo*-generated (micro)homology at the resected break ends.
(viii) Whether (i) distinct NHEJ- and SDSA-like-based mutations are predictable; (ii) which enzymatic device is responsible for SDSA-like events in plants; and (iii) to what degree original sequences reflect uncut positions or rather correct repair by direct ligation, SDSA or cross-over-like events remains to be solved in future.

EXPERIMENTAL PROCEDURES

Genomic loci selection for the investigation of DNA repair outcomes

To investigate endogenous repair profiles of CRISPR/Cas9 induced DSBs, we screened the *A. thaliana* genome sequence and selected three distinct genomic loci. Target 1 represents the intergenic region between genes At3g51870 and At3g51880 of chromosome 3 (Figures 1 and S1). This target was selected according to criteria that ensure optimal conditions for SSA repair such as: (i) containing two directly repeats (>80 bp) with a distance less than 100 bp between them; (ii) containing ≤3 SNP between the repeated copies; and (iii) a suitable protospacer can be designed to induce a DSBs between the two repeats. The second target located on the same chromosome between At3g50751 and At3g08155 was selected to test for SSA repair, however, different from target 1, end resection of >1000 bp is required for SSA repair in this case (Figures 1 and S2). Target 3 was selected and designed to test for the HDR repair pathway SDSA. This target is located at a distal position on chromosome 5. The breakpoint is targeted within one of two repeats, 1277 and 1432 bp in length, which are 1816 bp distant from each other and differ by several SNPs and indels responsible for the length difference (Figures 1 and S3).

T-DNA constructs and plant transformation

Recombinant CRISPR/Cas9 and sgRNA constructs were designed for unique genomic positions and cloned as described (Fauser *et al.*, 2014). Protospacers (Table S1) were ligated into pEn-Chimera. Respective expression cassettes for sgRNAs were transformed into *Escherichia coli* strain NEB5α and then transferred into pDE-pUbi-Cas9 by Gateway cloning. Finally, three T-DNA constructs were generated. *Arabidopsis thaliana* plants of ecotype Columbia-0 were transformed with T-DNAs (*Agrobacterium tumefaciens* strain GV3101) containing respective nuclease expression cassettes, using the floral dip method (Clough and Bent, 1998).

Plant selection

The Cas9 and sgRNA constructs were stably transformed into *A. thaliana* plants in order to obtain transgenic plant lines to address specifically each of the three selected targets. Transformed T1 plants for each construct were grown and selected on Murashige and Skoog (MS) medium (4.9 g l^{-1} Murashige and Skoog medium, 10 g l^{-1} sucrose, and 8 g l^{-1} agar, pH 5.7) containing 15 mg l^{-1} phosphinothricin and 0.5 g l^{-1} cefotaxime. A batch of 30 primarily transformed 10-day-old seedlings for each construct was pooled for DNA isolation.

Pacbio amplicon library preparation

DNA was isolated from 30 pooled T1 plants for each of three targeted lines using a DNeasy Plant Mini Kit (Qiagen GmbH, Hilden, Germany). Two amplicon libraries with different primer distances from the presumed breakpoint were prepared for each target line (Figure 1). Thus, in total, six PacBio amplicon libraries from the three T1 CRISPR/Cas9 T-DNA lines were constructed (Figure 1). Polymerase chain reaction (PCR) for amplification of the respective sgRNA target sites using a BioMix kit (Bioline GmbH, Luckenwalde, Germany) was performed with barcode-tailed PCR primers shown in Table S2. The following program was used for amplification: 3 min denaturation at 94°C, 25 cycles of 30 sec at 94°C, 30 sec at 60°C, and 1 min 20 sec or 2 min 40 sec at 72°C (depending on the length of amplicons), and final extension of 7 min at 72°C. Pre-washed AMPure XP magnetic beads (1.8× volume) were used to purify DNA amplicons. Library preparation followed the guidelines for SMRT® library preparation using the standard SMRTbell adapters which are subsequently ligated to the barcoded amplicons.

Deep sequencing analysis for DNA repair outcomes

The sequences were analysed as described (Vu *et al.*, 2017). Amplicon sequencing was performed on a Pacific Biosciences RSII system (MPIPZ, Cologne, Germany). Sequence data analysis was performed with the USEARCH version 8.0.1623 (Edgar, 2010) for clustering and with ClustalW for sequence alignment. In brief, reads were clustered at ≥99% sequence similarity. Chimeric reads and reads appearing only once (presumably resulting from PCR errors) were filtered out. The remaining reads were aligned against the original sequence to record polymorphisms. Reads deviating from the original sequence and containing at least 10 bp between both PCR primer sequences were considered as informative. For each respective target sequence, the original sequence (uncut or precisely repaired) and all mutation types (deletions, insertions, substitutions) were recorded and analysed regarding the respective repair path of their origin.

ACKNOWLEDGEMENTS

This study is supported by the German Research Foundation (SCHU 951/18–1) to I.S. and the European Research Council (AdG_20100317 RECBREED to H.P.).

AUTHOR CONTRIBUTIONS

G.T.H.V. and I.S. conceived the study. G.T.H.V., F.F., H.P. and I.S. designed research. G.T.H.V. and F.F. performed research. G.T.H.V., H.X.C., B.R. and I.S. analysed data. G.T.H.V. and I.S. wrote the article. All authors read, edited and approved the article.

CONFLICT OF INTEREST

The authors declare no conflict of interest.

SUPPORTING INFORMATION

Additional Supporting Information may be found in the online version of this article.

Figure S1. Sequence of target 1 selected for DSB induction within the *A. thaliana* genome.

Figure S2. Sequence of target 2 selected for DSB induction within the *A. thaliana* genome.

Figure S3. Sequence of target 3 selected for DSB induction within the *A. thaliana* genome.

Table S1. The CRISPR oligo sequences used for cloning.

Table S2. Design of PacBio amplicon sequencing libraries.

Table S3. Mutation classes detected at target locus 1.

Table S4. Mutation classes detected at target locus 2.

Table S5. Mutation classes detected at target locus 3.

REFERENCES

Aryan, A., Anderson, M.A., Myles, K.M. and Adelman, Z.N. (2013) Germline excision of transgenes in *Aedes aegypti* by homing endonucleases. *Sci. Rep.* **3**, 1603.

Bortesi, L. and Fischer, R. (2015) The CRISPR/Cas9 system for plant genome editing and beyond. *Biotechnol. Adv.* **33**, 41–52.

Cao, H.X., Wang, W., Le, H.T.T. and Vu, G.T.H. (2016) The power of CRISPR-Cas9-induced genome editing to speed up plant breeding. *Int J Genomics*, **2016**, 10.

Clough, S.J. and Bent, A.F. (1998) Floral dip: a simplified method for *Agrobacterium*-mediated transformation of *Arabidopsis thaliana*. *Plant J.* **16**, 735–743.

Cong, L., Ran, F.A., Cox, D. *et al.* (2013) Multiplex genome engineering using CRISPR/Cas systems. *Science*, **339**, 819–823.

Edgar, R.C. (2010) Search and clustering orders of magnitude faster than BLAST. *Bioinformatics*, **26**, 2460–2461.

Fauser, F., Schiml, S. and Puchta, H. (2014) Both CRISPR/Cas-based nucleases and nickases can be used efficiently for genome engineering in *Arabidopsis thaliana*. *Plant J.* **79**, 348–359.

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.

Hartlerode, A.J., Willis, N.A., Rajendran, A., Manis, J.P. and Scully, R. (2016) Complex breakpoints and template switching associated with non-canonical termination of homologous recombination in mammalian cells. *PLoS Genet.* **12**, e1006410.

Heyer, W.D., Ehmsen, K.T. and Liu, J. (2010) Regulation of homologous recombination in eukaryotes. *Annu. Rev. Genet.* **44**, 113–139.

Hustedt, N. and Durocher, D. (2017) The control of DNA repair by the cell cycle. *Nat. Cell Biol.* **19**, 1–9.

Ibarra-Laclette, E., Lyons, E., Hernandez-Guzman, G. *et al.* (2013) Architecture and evolution of a minute plant genome. *Nature*, **498**, 94–98.

Jinek, M., Chylinski, K., Fonfara, I., Hauer, M., Doudna, J.A. and Charpentier, E. (2012) A programmable dual-RNA-guided DNA endonuclease in adaptive bacterial immunity. *Science*, **337**, 816–821.

Kirik, A., Salomon, S. and Puchta, H. (2000) Species-specific double-strand break repair and genome evolution in plants. *EMBO J.* **19**, 5562–5566.

Koole, W., van Schendel, R., Karambelas, A.E., van Heteren, J.T., Okihara, K.L. and Tijsterman, M. (2014) A polymerase theta-dependent repair pathway suppresses extensive genomic instability at endogenous G4 DNA sites. *Nat. Commun.* **5**, 3216.

van Kregten, M., de Pater, S., Romeijn, R., van Schendel, R., Hooykaas, P.J.J. and Tijsterman, M. (2016) T-DNA integration in plants results from polymerase-θ-mediated DNA repair. *Nat. Plants*, **2**, 16164.

Lieber, M.R. (2010) The mechanism of double-strand DNA break repair by the nonhomologous DNA end-joining pathway. *Annu. Rev. Biochem.* **79**, 181–211.

Mali, P., Yang, L., Esvelt, K.M., Aach, J., Guell, M., DiCarlo, J.E., Norville, J.E. and Church, G.M. (2013) RNA-guided human genome engineering via Cas9. *Science*, **339**, 823–826.

Mao, Z., Bozzella, M., Seluanov, A. and Gorbunova, V. (2008) Comparison of nonhomologous end joining and homologous recombination in human cells. *DNA Repair (Amst)*, **7**, 1765–1771.

McVey, M., Khodaverdian, V.Y., Meyer, D., Cerqueira, P.G. and Heyer, W.D. (2016) Eukaryotic DNA polymerases in homologous recombination. *Annu. Rev. Genet.* **50**, 393–421.

Moehle, E.A., Rock, J.M., Lee, Y.L., Jouvenot, Y., DeKelver, R.C., Gregory, P.D., Urnov, F.D. and Holmes, M.C. (2007) Targeted gene addition into a specified location in the human genome using designed zinc finger nucleases. *Proc. Natl Acad. Sci. USA*, **104**, 3055–3060.

Orel, N., Kyryk, A. and Puchta, H. (2003) Different pathways of homologous recombination are used for the repair of double-strand breaks within tandemly arranged sequences in the plant genome. *Plant J.* **35**, 604–612.

Osakabe, Y., Watanabe, T., Sugano, S.S., Ueta, R., Ishihara, R., Shinozaki, K. and Osakabe, K. (2016) Optimization of CRISPR/Cas9 genome editing to modify abiotic stress responses in plants. *Sci. Rep.* **6**, 26685.

van Overbeek, M., Capurso, D., Carter, M.M. *et al.* (2016) DNA repair profiling reveals nonrandom outcomes at Cas9-mediated breaks. *Mol. Cell*, **63**, 633–646.

Pacher, M. and Puchta, H. (2016) From classical mutagenesis to nuclease-based breeding - directing natural DNA repair for a natural end-product. *Plant J.* **90**, 819–833.

Puchta, H. (1998) Repair of genomic double-strand breaks in somatic plant cells by one-sided invasion of homologous sequences. *Plant J.* **13**, 331–339.

Puchta, H. (1999) Use of I-Sce I to induce DNA double-strand breaks in *Nicotiana*. *Methods Mol. Biol.* **113**, 447–451.

Puchta, H. (2005) The repair of double-strand breaks in plants: mechanisms and consequences for genome evolution. *J. Exp. Bot.* **56**, 1–14.

Puchta, H. (2017) Applying CRISPR/Cas for genome engineering in plants: the best is yet to come. *Curr. Opin. Plant Biol.* **36**, 1–8.

Puchta, H. and Hohn, B. (2012) In planta somatic homologous recombination assay revisited: a successful and versatile, but delicate tool. *Plant Cell*, **24**, 4324–4331.

Ran, F.A. (2013) Genome engineering using the CRISPR-Cas9 system. *Nat. Protoc.* **8**, 2281–2308.

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.

Schiml, S., Fauser, F. and Puchta, H. (2016) Repair of adjacent single-strand breaks is often accompanied by the formation of tandem sequence duplications in plant genomes. *Proc. Natl. Acad. Sci. USA*, **113**, 7266–7271.

Schubert, I. and Vu, G.T. (2016) Genome stability and evolution: attempting a holistic view. *Trends Plant Sci.* **21**, 749–757.

Sfeir, A. and Symington, L.S. (2015) Microhomology-mediated end joining: a back-up survival mechanism or dedicated pathway? *Trends Biochem. Sci.* **40**, 701–714.

Shen, H., Strunks, G.D., Klemann, B.J., Hooykaas, P.J. and de Pater, S. (2017) CRISPR/Cas9-induced double-strand break repair in *Arabidopsis* nonhomologous end-joining mutants. *G3*, **7**, 193–202.

Siebert, R. and Puchta, H. (2002) Efficient repair of genomic double-strand breaks by homologous recombination between directly repeated sequences in the plant genome. *Plant Cell*, **14**, 1121–1131.

Steinert, J., Schiml, S. and Puchta, H. (2016) Homology-based double-strand break-induced genome engineering in plants. *Plant Cell Rep.* **35**, 1429–1438.

Tan, E.P., Li, Y., Velasco-Herrera Mdel, C., Yusa, K. and Bradley, A. (2015) Off-target assessment of CRISPR-Cas9 guiding RNAs in human iPS and mouse ES cells. *Genesis*, **53**, 225–236.

Vu, G.T., Cao, H.X., Watanabe, K., Hensel, G., Blattner, F.R., Kumlehn, J. and Schubert, I. (2014) Repair of site-specific DNA double-strand breaks in barley occurs via diverse pathways primarily involving the sister chromatid. *Plant Cell*, **26**, 2156–2167.

Vu, G.T.H., Schmutzer, T., Bull, F. *et al.* (2015) Comparative genome analysis reveals divergent genome size evolution in a carnivorous plant genus. *Plant Genome*, **8**, https://doi.org/10.3835/plantgenome2015.04.0021.

Vu, G.T.H., Cao, H.X., Reiss, B. and Schubert, I. (2017) Deletion-bias in DNA double-strand break repair differentially contributes to plant genome shrinkage. *New Phytol.* **214**, 1712–1721.

Watanabe, K., Breier, U., Hensel, G., Kumlehn, J., Schubert, I. and Reiss, B. (2016) Stable gene replacement in barley by targeted double-strand break induction. *J. Exp. Bot.* **67**, 1433–1445.

Wendeler, E., Zobell, O., Chrost, B. and Reiss, B. (2015) Recombination products suggest the frequent occurrence of aberrant gene replacement in the moss *Physcomitrella patens*. *Plant J.* **81**, 548–558.

Wyatt, D.W., Feng, W., Conlin, M.P., Yousefzadeh, M.J., Roberts, S.A., Mieczkowski, P., Wood, R.D., Gupta, G.P. and Ramsden, D.A. (2016) Essential roles for polymerase theta-mediated end joining in the repair of chromosome breaks. *Mol. Cell*, **63**, 662–673.

Yu, A.M. and McVey, M. (2010) Synthesis-dependent microhomology-mediated end joining accounts for multiple types of repair junctions. *Nucleic Acids Res.* **38**, 5706–5717.

Zhao, Y.P., Zhang, C.S., Liu, W.W. *et al.* (2016) An alternative strategy for targeted gene replacement in plants using a dual-sgRNA/Cas9 design. *Sci. Rep.* **6**, 23890.

