New Phytologist

Deletion-bias in DNA double-strand break repair differentially contributes to plant genome shrinkage

Giang T. H. Vu[1]*, Hieu X. Cao[1]*, Bernd Reiss[2] and Ingo Schubert[1]

[1]Leibniz Institute of Plant Genetics and Crop Plant Research (IPK), D-06466 Gatersleben, Germany; [2]Max Planck Institute for Plant Breeding Research, 50829 Köln, Germany

Authors for correspondence:
Ingo Schubert
Tel: +49 (0)394825239
Email: schubert@ipk-gatersleben.de

Giang T. H. Vu
Tel: +49 (0)394825291
Email: vu@ipk-gatersleben.de

Received: *23 November 2016*
Accepted: *22 January 2017*

New Phytologist (2017) **214:** 1712–1721
doi: 10.1111/nph.14490

Key words: double-strand break (DSB) repair, genome shrinkage, genome size evolution, nonhomologous end joining (NHEJ), single strand annealing (SSA).

Summary

- In order to prevent genome instability, cells need to be protected by a number of repair mechanisms, including DNA double-strand break (DSB) repair. The extent to which DSB repair, biased towards deletions or insertions, contributes to evolutionary diversification of genome size is still under debate.
- We analyzed mutation spectra in *Arabidopsis thaliana* and in barley (*Hordeum vulgare*) by PacBio sequencing of three DSB-targeted loci each, uncovering repair via gene conversion, single strand annealing (SSA) or nonhomologous end-joining (NHEJ). Furthermore, phylogenomic comparisons between *A. thaliana* and two related species were used to detect naturally occurring deletions during Arabidopsis evolution.
- *Arabidopsis thaliana* revealed significantly more and larger deletions after DSB repair than barley, and barley displayed more and larger insertions. *Arabidopsis* displayed a clear net loss of DNA after DSB repair, mainly via SSA and NHEJ. Barley revealed a very weak net loss of DNA, apparently due to less active break-end resection and easier copying of template sequences into breaks. Comparative phylogenomics revealed several footprints of SSA in the *A. thaliana* genome.
- Quantitative assessment of DNA gain and loss through DSB repair processes suggests deletion-biased DSB repair causing ongoing genome shrinking in *A. thaliana*, whereas genome size in barley remains nearly constant.

Introduction

Alterations of DNA sequence are the source of genomic variability during evolution. Sequence alterations may occur via replication, for example after deamination of 5-methylcytosine into thymidine, but most of the variability at the diploid level might be the outcome of mis-repair of DNA damage, including spreading of mobile elements (Hedges & Deininger, 2007). In particular, genomic DNA double-strand breaks (DSBs) are critical lesions which when unrepaired are usually lethal for dividing cells. To maintain genome integrity, cells evolved DNA damage response and repair mechanisms (for review, see Britt, 1996; Polo & Jackson, 2011; Waterworth *et al.*, 2011). However, these mechanisms do not act absolutely perfectly. Mis-repair of DSBs in genomes can lead either to local changes in DNA sequence such as small deletions, insertions or substitutions, or to large-scale chromosomal rearrangements such as inversions and translocations, large duplications and deletions (if viable) (Aguilera & Gomez-Gonzalez, 2008; Villarreal *et al.*, 2012; Aguilera & Garcia-Muse, 2013).

The two major pathways for DSB repair are nonhomologous end joining (NHEJ) and homologous recombination (HR). DSB repair by HR occurs preferentially after replication, when a closely adjacent sister chromatid offers a homologous template for repair of its damaged counterpart (for review, see San Filippo *et al.*, 2008; Jasin & Rothstein, 2013). Unlike HR, NHEJ acts throughout the cell cycle by joining DNA break ends either directly or by the use of microhomology of 1–25 bp (for review see Lieber, 2010; Symington & Gautier, 2011; Sfeir & Symington, 2015). DSBs repaired by NHEJ can either restore the original sequence or create deletions and/or insertions of different lengths (Vu *et al.*, 2014). Mechanisms and consequences of DSB repair in plants has been reviewed (e.g. Puchta, 2005; Bleuyard *et al.*, 2006; Waterworth *et al.*, 2011). The outcome of inaccurate repair in meristematic cells, if viable, may become fixed and passed to the next generation via germ cells. Even minor modifications of components involved in DSB repair, manifesting as hypo- or hypermorphic variants, might over longer terms profoundly affect genome size and cause progressive shrinkage or expansion of genomes (Orel & Puchta, 2003).

DSB repair has been studied in 40 selected breakpoint junction sequences of *A. thaliana* and tobacco (*Nicotiana tabacum*) calli using the negative selectable marker gene *codA* as transgenic reporter. A tendency for larger deletions was observed in diploid *A. thaliana* which has a 28-fold smaller genome size (157 Mbp) than allotetraploid tobacco (4.5 Gbp) (Kirik *et al.*, 2000).

*These authors contributed equally to this work.

However, it remained unclear whether calli are representative for the *in planta* situation, whether this correlation holds true also for other comparisons of species with small and large genomes, and what proportion individual pathways contribute to a deletion-biased DSB repair.

Based on intron size comparison of homologous genes, a deletion bias during somatic DSB repair is inferred as a reason for genome shrinkage in one of the smallest angiosperm genomes (*Genlisea nigrocaulis,* 86 Mbp) compared with the 18-fold larger *G. hispidula* genome (Vu *et al.*, 2015).

In the present study, we investigated and compared the mutation spectra of the eudicot *A. thaliana* (1C = 157 Mbp, Bennett *et al.*, 2003) and the monocot barley (*Hordeum vulgare*, 1C = 5500 Mbp, Bennett & Leitch, 2005) by sequencing three unique transgenic loci each after targeted DSB induction. We found a similar spectrum of repair events for both species, but noticed quantitative differences. In *A. thaliana*, deletion size and frequency is considerably larger than in barley, whereas in barley insertion size and frequency is larger than in *A. thaliana*. In total, barley revealed a very weak and Arabidopsis a strong net loss of DNA via erroneous DSB repair. Comparison of homeologous sequences between *A. thaliana, A. lyrata* and *Capsella rubella* additionally supported the assumption of evolutionary genome shrinking by deletions during DSB repair via single strand annealing.

Materials and Methods

Plant materials and growth conditions

In order to investigate comparatively the mutation spectra resulting from the repair of I-SceI-induced double-strand breaks (DSBs), three double hemizygous transgenic lines (GU.US × I-SceI; DU.GUS × I-SceI; IU.GUS × I-SceI) of both *Arabidopsis thaliana* and barley *Hordeum vulgare* were used. Each line contains one copy of a modified β glucuronidase gene as the corresponding DSB substrate (either GU.US, DU.GUS or IU.GUS) with an I-SceI recognition site, and the constitutively expressed meganuclease I-SceI gene (Orel *et al.*, 2003; Vu *et al.*, 2014) under the *ubiquitin* promoter. In the GU.US construct, two direct repeats (U) of 614 bp are interrupted by a 39-bp sequence including the 18 bp recognition site for the I-SceI endonuclease. Therefore, the I-SceI-induced DSB can be repaired by single strand annealing (SSA) causing a deletion of 653 bp. In the DU.GUS or IU.GUS constructs, the *cis* homologous sequences (U) in direct (DU) or inverse (IU) orientation may be used as templates for repair via conversion of the interrupted U containing the DSB. DSB repair by gene conversion in DU.GUS and IU.GUS resulted in restoration of correct U and deletion of 30 bp (see Fig. 1a). The transgenic barley cv 'Golden Promise' and *A. thaliana* ecotype Columbia-0 plants were grown in soil at 22°C under long day conditions (16 h light : 8 h dark) after 4 d at 4°C.

PacBio amplicon library preparation

DNA was isolated from plants 10 d after germination using a DNeasy Plant Mini Kit (Qiagen). DNA of 10 *A. thaliana* or three barley transgenic plants was pooled. A 1568-bp amplicon library for each GU.US line and 1632-bp amplicon libraries for DU.GUS and IU.GUS lines were prepared. Thus, in total, six PacBio amplicon libraries from the three double-hemizygous lines of *A. thaliana* and of barley were constructed (Fig. 1a). As controls for PCR-generated artifacts and monitoring of endonuclease-independent DSB in studied amplicons, we included four corresponding PacBio amplicon libraries from barley transgenic plants homozygous for GUS, GU.US, DU.GUS and IU.GUS which did not harbor the I-SceI gene. PCR using a BioMix kit (Bioline GmbH, Luckenwalde, Germany) was performed with barcode-tailed PCR primers shown in Supporting Information Table S1. The following program was used for amplification: 3 min denaturation at 94°C, 25 cycles of 30 s at 94°C, 30 s at 59°C, and 1 min 30 s or 3 min at 72°C (depending on the length of amplicons), and final extension of 5 min at 72°C. Amplicons were purified and concentrated using 1.8× volume of pre-washed AMPure XP magnetic beads according to the guidelines for PacBio PCR amplicon preparation. Concentrations of DNA amplicons were measured by Qubit quantification before multiplexing. 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 mutation spectra

Amplicon sequencing was performed on a Pacific Biosciences RSII system (MPIPZ, Cologne, Germany, http://mpgc.mpipz.mpg.de/home/) using P4/C2 chemistry at 180 min movie length. High quality Reads of Inserts (ROI, ≥ 3 full passes and ≥ 90% predicted accuracy) were generated using SMRT Analysis (provided by Pacific Biosciences, Menlo Park, CA, USA). The reads were trimmed to remove low-quality bases (quality limit = 0.05) and demultiplexed on the basis of the 16-bp unique barcodes using the CLC Genomics Workbench (v.5). The resulting reads containing both forward and reverse PCR primer sequences of each construct library were analyzed using Usearch v.8.0.1623 (Edgar, 2010) as described previously (Vu *et al.*, 2014). In short, reads were clustered at ≥ 99% sequence similarity. After filtering out chimeric reads and reads appearing only once (presumably resulting from PCR errors), the remaining reads were aligned against the parental construct sequences using ClustalW to record sequence polymorphisms. Such reads, when containing at least 10 bp of parental construct sequences between both PCR primer sequences, were considered as informative (actually the shortest read represented a 623-bp amplicon indicating a 1019-bp deletion within the IU.GUS construct in barley, Table S2). Comparisons of different size categories were performed by Fisher's exact test using absolute values. Sequences from amplicon libraries of control plants homozygous for the target sites (but lacking the I-SceI gene) showed a sequence accuracy > 97% and no endonuclease-independent indels or deletions caused by PCR artefacts (Table S1).

Phylogenomic analysis for genome-wide SSA repair signatures

Genomic sequence annotations of *A. thaliana* Col-0 (CoGe_id:19865), *A. lyrata* (CoGe_id:19868) and *Capsella*

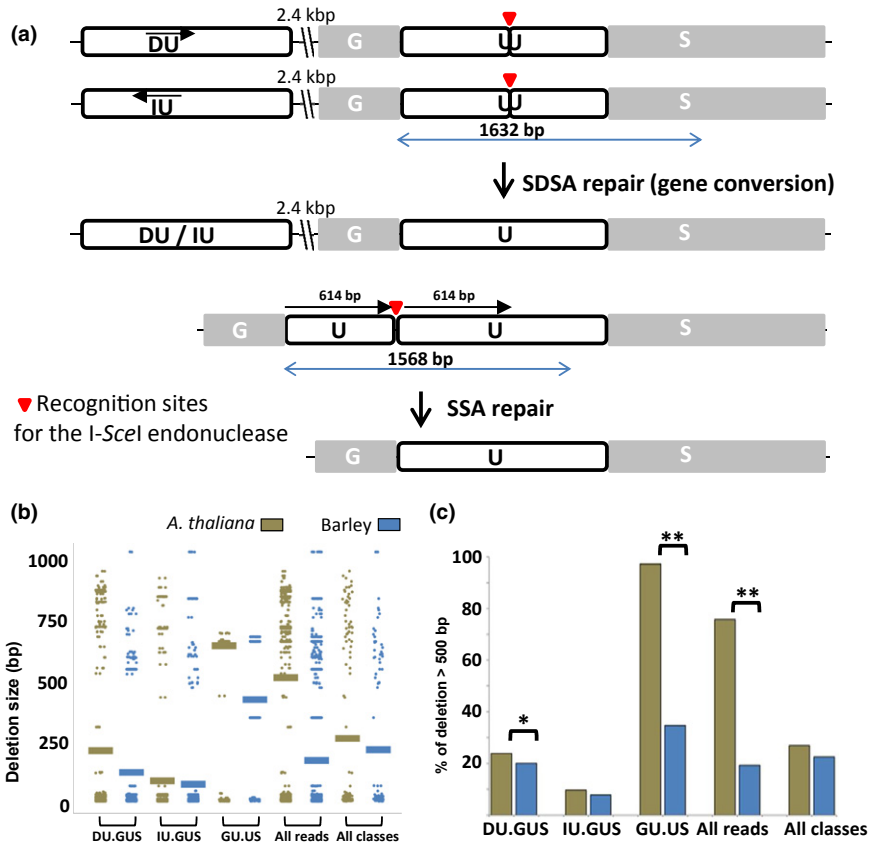

Fig. 1 DNA double-strand break (DSB) repair at DU.GUS, IU.GUS and GU.US loci yielded larger deletions in *Arabidopsis thaliana* than in the barley genome. (a) Three transgenic reporter constructs (DU.GUS, IU.GUS and GU.US) in which the β-glucuronidase (GUS) gene was interrupted by a unique sequence including the 18-bp I-SceI recognition site. In the DU.GUS or IU.GUS constructs, the *cis* homologous sequences (U) in direct (DU) or inverse (IU) orientation may be used as templates for repair via conversion of the interrupted U containing the DSB. DSB repair by gene conversion in DU.GUS and IU.GUS resulted in restoration of correct U and deletion of 30 bp. In the GU.US construct, two direct repeats (U) of 614 bp (black arrowed lines) are interrupted by a 39-bp sequence including a recognition site for the I-SceI endonuclease. Therefore, the I-SceI-induced DSB can be repaired by single strand annealing (SSA) deleting 653 bp. Length and position of amplicon libraries for each construct are indicated by double arrowed blue lines. (b) Deletion sizes (dots) obtained from repair products of each construct calculated based on sequence reads as well as average deletion size of all three constructs per all reads and per all mutation classes in *A. thaliana* (tan) and barley (blue). The average deletion values are indicated by bars. (c) Proportion of sequence reads containing deletions larger than 500 bp calculated for each construct and for all three constructs as well as the proportion of event classes containing deletions larger than 500 bp in all three constructs in *A. thaliana* (tan column) and barley (blue column). *, *P*-value < 0.05; **, *P*-value < 0.01 (Fisher's exact test).

rubella (CoGe_id:16754) were used for genome-wide identification of orthologous regions between these species using the SynMap tool of the CoGe platform (genomevolution.org). The LAST algorithm was used for all-against-all comparison of protein sequences excluding tandem duplicates. A minimum number of five aligned gene pairs and a maximum distance between two matches of 20 genes for a synteny region were applied. Among highly conserved orthologous gene pairs (≥ 90% similarity in amino acid) within the synteny regions, we identified in total 50 *A. thaliana* genes which were at least 1 kbp smaller than their orthologs in *A. lyrata* and *C. rubella* (group A). In addition, 25 orthologous genes which were > 1 kbp shorter in *A. lyrata* than in *A. thaliana* and *C. rubella* (group B) were included. Full-length nucleotide sequences of these genes were pair-wise aligned using NCBI blastn with default parameters without filtering of low complexity regions. In order to detect direct repeats, we looked for hits that covered more than 50 bp. Gene loci containing such direct repeats in *A. lyrata* and *C. rubella* but an equally reduced number of repeat copies and intervening sequences between repeats in *A. thaliana* were considered as indication for SSA repair during genome evolution.

Accession numbers

GenBank accession numbers of three original reporter constructs in Arabidopsis and barley are: JX475903.1–JX475905.1 and KJ817199.1–KJ817201.1, respectively (Table S1).

Results

DSB repair produces larger deletions in the smaller *Arabidopsis* genome

It is indisputable that deletion size and rates can have an impact on genome size of organisms (Lysak *et al.*, 2006; Pettersson *et al.*, 2009; Hu *et al.*, 2011; Vu *et al.*, 2015). However, it is still

unclear to what degree DSB repair affects genome size. To investigate the potential impact of DSB repair on genome size evolution, we compared the mutation spectra resulting from DSB repair. We used three double hemizygous transgenic lines in both *A. thaliana* and barley each with one copy of a corresponding DSB substrate (GU.US, DU.GUS and IU.GUS, respectively; all containing an I-SceI recognition site, see Fig. 1(a) and Orel *et al.*, 2003) to examine repair of DSBs at unique sites targeted by the transgenically expressed meganuclease I-SceI. The three constructs were originally designed to detect by means of restored β-glucuronidase activity DSB repair via either gene conversion (SDSA in case of DU.GUS and IU.GUS) or via deletion (SSA in case of GU.US) (Fig. 1a); we now extend this approach to a much broader spectrum of repair events at the sequence level at different genomic positions. The employed *ubiquitin* promoter mediated strong I-SceI expression ensuring repeated cycles of cutting and repair until inaccurate repair eliminates the recognition site. For inspection of repair events in double-hemizygous *A. thaliana* and barley plants, we designed and massively sequenced by single molecule sequencing (PacBio) similar sized amplicon libraries of the three reporter constructs (1568 bp for GU.US; 1632 bp for DU.GUS and IU.GUS; Fig. 1a; Table S1), in order to validate how the DSBs were repaired. The long single molecule PacBio reads allowed us to sequence and to evaluate each DNA amplicon molecule at least three times within an individual PacBio read to eliminate sequencing errors. Such long-range libraries enabled monitoring of deletions up to *c.* 1400 bp and insertions up to *c.* 3000 bp.

We divided all reads into those representing either an uncut or precisely restored original sequence, and those which revealed a mutated sequence. All reads showing the same mutated sequence were considered as a mutation class. We first investigated the average deletion size in *A. thaliana* (from 6320 imprecisely repaired – i.e. mutated – sequences belonging to 165 mutation classes each formed by identical sequence reads) and barley (from 2163 imprecisely repaired sequences belonging to 128 mutation classes) for all three constructs (Fig. 1; Table 1). Although the number of reads differed for both species, the proportion of mutated reads as well as the number of mutation classes is very similar in both cases. Therefore, differences in size and number of deletions and insertions, respectively, should reflect actual differences in the mode of DSB repair. In case DSB repair has an immediate impact on genome size, we expect that deletions resulting from DSBs mis-repair should be longer in *A. thaliana* than in the 35-fold larger barley genome, if shrinking of the *A. thaliana* genome is still ongoing.

In *A. thaliana*, the average deletion size of all sequence reads for all three constructs at the breakpoint junction is 505 bp, which is three-fold larger than the corresponding value in barley (169 bp; Fig. 1b; Table S2). This larger deletion size is consistent for all three transgenic loci, each at different genomic positions. For GU.US lines, the average deletion size in barley is 417 bp, which is 65.7% of that in *A. thaliana* (636 bp). For the DU.GUS and IU.GUS constructs, characterized by ectopic homologous sequences in direct or inverse orientation which are expected to be used to repair the break via gene conversion, the average deletion sizes in both species were smaller than for the GU.US construct. Nevertheless, the average deletion sizes for both constructs in Arabidopsis (209 bp and 85 bp, respectively) were still larger than in barley (121 bp and 72 bp; Fig. 1b). Moreover, the proportion of mutated reads containing a large deletion (≥ 500 bp) for all three constructs together in *A. thaliana* (75.7%) was significantly higher than in barley (19.3%) (Fig. 1c, $P < 0.01$, Fisher's exact test).

The overall average deletion size calculation based on the number of sequence reads could be an overestimation if early events passed several replication cycles and cell divisions. Therefore, we analyzed the average deletion size from all three constructs for all mutation classes, too. This analysis in turn may underestimate the real average deletion size because identical reads can result from different independent events (e.g. in case of SSA). Notably, the average deletion size per class in *A. thaliana* (250 bp) was again larger in comparison to that of barley (202 bp) (Fig. 1b). Considered together, our results suggest that large deletions caused by DSB repair in *A. thaliana* played a role during the evolution of its small genome.

Microhomology-mediated synthesis-dependent strand annealing (MM-SDSA) generates longer insertions in the larger barley genome

That DSB repair in plants can be linked with insertions was established in the 1990s (Gorbunova & Levy, 1997; Salomon & Puchta, 1998). Sequencing amplicons from targeted transgenic

Table 1 Summary output of PacBio amplicon sequencing libraries

Species	Construct	Informative reads	Original reads*	Imprecise repair reads	Imprecise repair classes	Mutation type (reads/classes)**			
						SSA (HR)	Gene conversion (HR)	MM-SDSA	NHEJ
Arabidopsis	GU.US	4621	6	4615 (99.87%)	19	4477/1		11/5	127/13
	IU.GUS	994	193	801 (80.58%)	58		91/1	50/9	660/48
	DU.GUS	1106	202	904 (81.74%)	88		171/1	147/16	586/71
Barley	GU.US	500	26	474 (94.80%)	15	86/1		73/1	315/13
	IU.GUS	1239	541	698 (56.34%)	55		57/1	24/8	617/46
	DU.GUS	1448	457	991 (68.44%)	58		66/1	193/11	732/46

*Representing uncut or precisely repaired original transgene sequences.
**SSA, single strand annealing; HR, homologous recombination; MM-SDSA, microhomology-mediated synthesis-dependent strand annealing; NHEJ, nonhomologous end-joining.

DSB repair constructs revealed in barley a repair mechanism which is based on microhomology-mediated invading and copying of a template sequence by one break end before reannealing with the other break end, and pasting the copied sequence into the break (Vu *et al.*, 2014), similar to the process described for *Drosophila* by Yu & McVey (2010). We call this route of DSB repair MM-SDSA.

Because insertion and deletion mutations (indels) impact genome size by adding or removing sequences (Petrov *et al.*, 2000), we hypothesized that DSB repair in the larger barley genome might be skewed towards shorter deletions and/or larger insertions. To test for occurrence of longer insertions in the larger genome via insertion-biased DSB repair, we compared the insertion size at breakpoint junctions in *A. thaliana* and in barley (Fig. 2a,b). Of 6320 imprecisely repaired sequence reads in Arabidopsis, 338 belonging to 63 mutation classes yielded an insertion at the breakpoint (Table S2). In barley, 496 (53 classes) of 2163 sequence reads showed insertions.

Thus, only 5.3% imprecisely repaired sequences in *A. thaliana* showed insertion, whereas in barley, in addition to the DSB repair biased against large deletions, an accumulation of insertions (22.9% of mutant reads) was observed (Fig. 2b; $P < 0.01$, Fisher's exact test). Furthermore, insertions larger than 100 bp can be observed only from barley DU.GUS and IU.GUS (in total 14 sequences in four different mutation classes, Table S2). The maximal insert length by MM-SDSA repair was 198 bp in barley compared to 65 bp in *A. thaliana* (Fig. 2a). The insertions in almost all cases originated from the construct sequences upstream or downstream flanking the breakpoints (Fig. 2c; Table S2). Finally, the average insertion size in all constructs in barley (18 bp) was 60.4% larger than that in *A. thaliana* (11 bp) (Fig. 2a).

In order to further substantiate our findings, we calculated the average insertion size over all mutation classes. Again, the average insertion size was larger in barley (19 bp) than in *A. thaliana* (14 bp) (Fig. 2a).

The gain of 'filler DNA' can counteract deletion formation during DSB repair by switching from degradation of break ends towards synthesis of insert sequences which eventually compensate for deletion and result in a smaller net loss of DNA at break

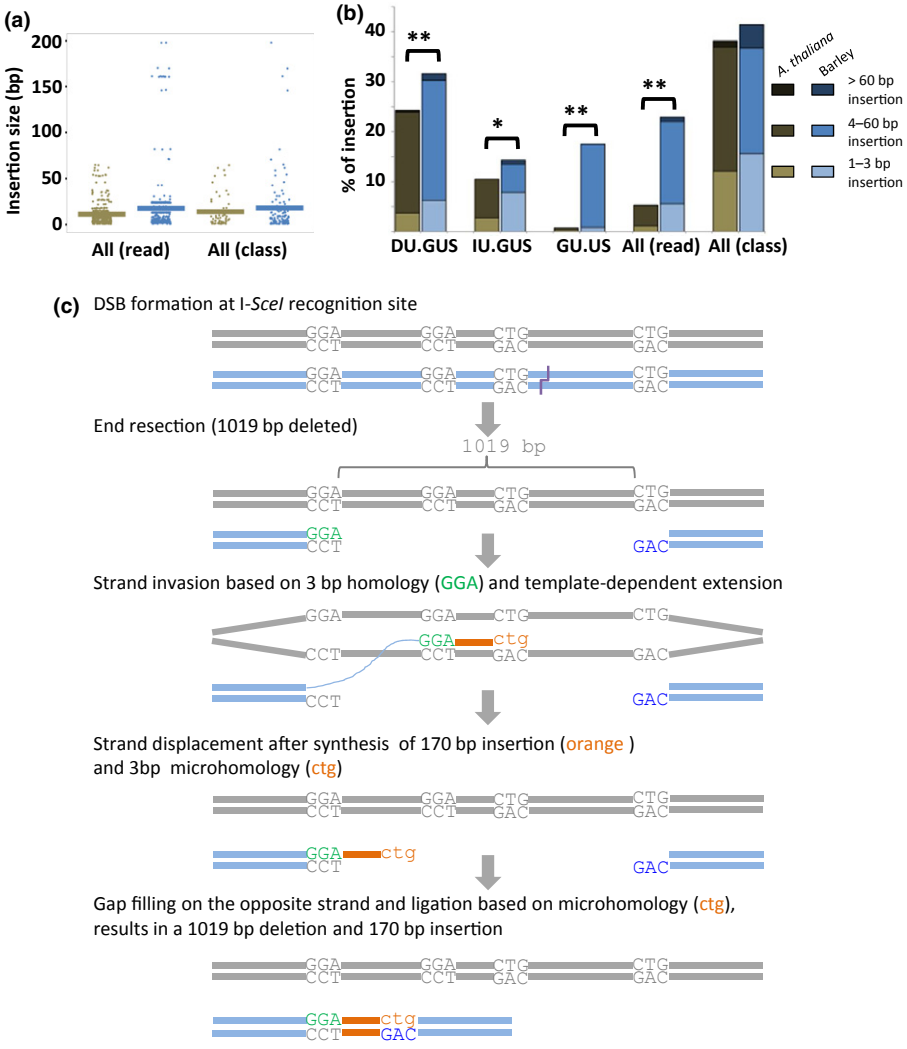

Fig. 2 DNA double-strand break (DSB) repair skewed towards larger insertions in the larger barley genome. (a) Insertion sizes (dots) of all three constructs calculated based on number of sequence reads and per all mutation classes in *Arabidopsis thaliana* (tan) and barley (blue). The average insertion values are indicated by bars. (b) Proportion of sequence reads containing insertions (subdivided into three different size groups) are shown for each construct and for all three constructs together, as well as the proportion of event classes containing insertions in all three constructs in *A. thaliana* (tan columns) and barley (blue columns). *, *P*-value < 0.05; **, *P*-value < 0.01 (Fisher's exact test). (c) Example of microhomology-mediated synthesis-dependent strand annealing (MM-SDSA) as exemplified by read D1019-I170 of the IU.GUS construct in barley (see Supporting Information Table S2). This case represents a combination of deletion (1019 bp) with an insertion (170 bp) during repair.

positions (Salomon & Puchta, 1998). When filler DNA results from MM-SDSA, a process that would be favored in barley, it could contribute to a lower net loss and even to a net gain of DNA by DSB repair in barley. Our sequencing data revealed larger insertions and/or shorter deletions at the breakpoint junctions in the barley genome compared to *A. thaliana*, providing evidence that a net loss of DNA resulting from DSB repair occurred in both species but is inversely correlated with their genome size (Table 2). Possibly, a more efficient protection of free break ends against extensive degradation and at the same time maintenance of the capability of 3′-overhangs for single strand invasion, annealing and copying a template sequence (MM-SDSA), eventually leads to more insertions in barley. Consequently, the DSB repair-mediated net loss of DNA in barley is much smaller than in *A. thaliana* and does not cause a severe evolutionary shrinkage of the barley genome.

The spectrum of DSB repair outcome indicates a stronger DSB end resection in Arabidopsis

Previous experiments on extrachromosomal plasmids suggested that different DSB end processing might be responsible for differences in repair-mediated deletion size between *A. thaliana* and tobacco (Orel & Puchta, 2003). To address whether larger deletions in Arabidopsis could be due to a more intense end resection in genomic DNA, we investigated in *A. thaliana* and in barley, the same integrated transgenic reporter construct GU.US designed for the detection of single strand annealing (SSA), a homology-directed repair pathway. The GU.US construct contains a 39-bp sequence including a recognition site for the I-SceI endonuclease between two direct repeats of 614 bp. The classical SSA repair model suggested that the I-SceI-induced DSB at the GU.US construct requires resection of 634 bp at either side of the break to reveal homology for annealing. In this way, the SSA repair restores the functional GUS gene and causes a 653-bp deletion (Fig. 1a). An alternative explanation (shorter resected ends invade by means of RAD51 the second repeat of the sister helix and mimic SSA) is less likely, at least for *A. thaliana*, because loss of RAD51 impairs SDSA but only barely impairs SSA events in *A. thaliana* (Roth *et al.*, 2012; Serra *et al.*, 2013).

Among 4615 imprecisely repaired sequences in a pool of GU.US transgenic Arabidopsis plants, 4477 reads (97%) showed an SSA repair pattern (Tables 1, S2). By contrast, only 18.1% (86 of 474 reads; Tables 1, S2) in a pool of barley GU.US plants were repaired by SSA. Thus, SSA, which requires 1268 bp resection (634 bp from each break end), is a predominant repair path in Arabidopsis, indicating a more extended DSB end resection in Arabidopsis than in barley. Therefore, we speculate that stronger end resection via exonuclease (and/or endonuclease; see Mimitou & Symington, 2009) activities favor SSA between endogenous repeats in *cis* and could be a reason for (ongoing) genome shrinkage by deletion mutations in *A. thaliana*. Blue spots indicating SSA repair after I-SceI-mediated DSB induction in transgenic GU.US plants are very frequent in *A. thaliana* (Orel *et al.*, 2003), but very rare in barley (Vu *et al.*, 2014). For barley we found no SSA at all among the mutation classes in previous experiments (Vu *et al.*, 2014), whereas the other mutation classes appear with similar abundance in both species (Table 1). These results further support our presumption that DSB ends in *A. thaliana* are more strongly resected than in barley.

Table 2 Net loss of product size in bp after targeted double-strand break (DSB) repair in *Arabidopsis thaliana* and barley

	A. thaliana	Barley	Fold difference
Genome size (Mbp)	157	5500	35
Average net loss of product size (per all reads)	504.16	163.23	3.1
Average net loss of product size (per all classes)	218.41	173.07	1.3
Average net loss normalized to genome size (per Mb; over all reads)	3.21	0.03	107*
Average net loss normalized to genome size (per Mb; over all classes)	1.39	0.03	46*

*After normalization to genome size (Mb), *A. thaliana* shows a net loss of DNA between 46- and 107-fold larger than that of barley, for which the net loss is close to equilibrium.

The SSA repair of DSBs is the most likely explanation for many naturally occurring deletions during *Arabidopsis* evolution

DSB repair has been inferred from our experimental systems to mediate genome size evolution, and many deletions in the smaller of two *Arabidopsis* genomes were reported by Hu *et al.* (2011). Whether or not distinct repair activities observed in experiments reflect the repair of natural breaks (which may be of blunt end or staggered end configuration) during evolution is not yet clear. Here, we used a phylogenomic approach to study the repair of naturally occurring chromosomal breaks in evolutionary context. We focused on orthologous sequences to study the frequency of events indicating SSA as a distinct route of deletion-mediating DSB repair. The alignment of *A. lyrata* (~235 Mbp; Hu *et al.*, 2011) and *A. thaliana* genomes, and of the genome of *C rubella* (~255 Mbp; Slotte *et al.*, 2013) as outgroup, allowed studying the evolutionary signature of DSB repair in distinct sequence contexts.

Orthologous genes of *A. lyrata* and *A. thaliana*, containing endogenous direct repeats in *cis* (similar as U.U in the GU.US construct) in *A. lyrata*, but not in *A. thaliana* can be used to deduce SSA events, when (at least one) intergenic sequence between two or more direct repeats and one of the repeat copies were deleted, similar as restoration of GUS from GU.US. An alternative repair involving inter-helix homologous recombination is less likely for Arabidopsis according to the results of Roth *et al.* (2012) that SSA and SDSA are only weakly affected in the absence of SMC6B which promotes sister chromatid recombination after DNA damage (Watanabe *et al.*, 2009). The orthologous genes in *C. rubella* were inspected to make sure that sequences were removed in *A. thaliana* and not amplified in

A. lyrata, suggesting that SSA impacted genome size over evolutionary timescales.

We have identified 12 orthologous genes in the *A. lyrata* and *C. rubella* genomes which are at least 1 kbp larger than orthologous genes in *A. thaliana* (based on the genome alignment; see the Materials and Methods section; Table S3, gene group A) and contain direct repeats in *cis* in *A. lyrata* and *C. rubella*. Four of these 12 orthologous gene pairs showed a similar repeat pattern in *A. lyrata* and *C. rubella* but deletions in *A. thaliana* loci, indicative of SSA processes after separation of both *Arabidopsis* species (Fig. 3a,b; Table S3). In addition, two others (At3g45560 and At5g42320) of the 12 orthologous gene pairs revealed a deviating pattern. The gene At3g45560 revealed SSA before divergence of *A. thaliana* and *A. lyrata*. The gene At5g42320 showed deletions of direct repeats that could be explained by two SSA repair events, either one before and one after divergence of *A. thaliana* and *A. lyrata*, or both after divergence of both species (Fig. 3c).

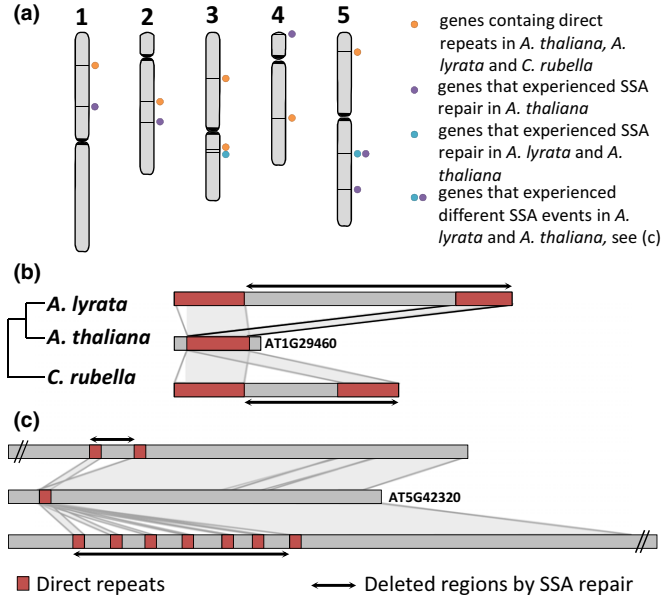

Fig. 3 Evidence of single strand annealing (SSA) repair in *Arabidopsis thaliana* during genome evolution. (a) Location of the 12 studied genes on *A. thaliana* chromosomes (black bars). Deletions in *A. thaliana* genes that are flanked by direct repeats in *A. lyrata* and *Capsella rubella* (mimicking the GU.US arrangement) indicate SSA repair in *A. thaliana* (violet dots). Genes that experienced SSA before the divergence of *A. thaliana* and *A. lyrata* (cyan dots) show deletions in both *Arabidopis* species, whereas flanking direct repeats are present in *C. rubella*. Genes with direct repeats in *A. lyrata* and/or *C. rubella* but without evidence for SSA (due to abundant sequence polymorphism) are shown as orange dots. (b) A cladogram presents the phylogenetic relationship of the three studied species. Sequence alignment of the *A. thaliana* gene AT1G29460 and its orthologs in *A. lyrata* and *C. rubella* illustrates an example of SSA in *A. thaliana*. Homologous sequence regions which were defined by NCBI blastn are indicated as connecting lines with gray background. Red bars represent direct repeats in gene sequences. The deleted parts by SSA repair are shown as black lines. (c) A sequence alignment of *A. thaliana* gene AT5G42320 demonstrating two SSA repair events that occurred either subsequently before and after, or independently after the divergence of *A. thaliana* and *A. lyrata,* and led to shrinkage of this gene in *A. lyrata* and especially in *A. thaliana*. Double slashes indicate discontinued gene sequences.

The other six cases were not considered as SSA repair due to uncertainty in classification, because more than one explanation (SSA and NHEJ or NHEJ) for deletions was possible. Thus, in accordance with our experimental data, it appears that deletions via SSA repair happened relatively frequently (at least half of the 12 testable orthologous genes) between direct repeats in *cis*. We found that deletions, most likely caused by DSB repair, happened more often in the smaller *Arabidopsis* genome (six events) than in the larger one (two events). By contrast, out of 25 orthologous genes which were > 1 kbp shorter in *A. lyrata* than in *A. thaliana* and *C. rubella,* only two contained direct repeats and none showed repeat deletion by SSA in *A. lyrata* (Table S3, gene group B).

These data represent, to our knowledge, the first study on repair of experimentally induced and phylogenomically presumed DSBs in species with different genome sizes and show that deletions with footprints of SSA repair seem to be more frequent in smaller than in larger genomes.

Discussion

Plant genome size varies by more than three orders of magnitude (~2350-fold), ranging from 64 Mbp in *Genlisea* (Greilhuber *et al.*, 2006) to 149 Gbp in *Paris japonica* (Pellicer *et al.*, 2010). The molecular mechanisms responsible for the dynamics of genome size evolution are still under discussion. Polyploidization and transposable element proliferation are two well-known processes which increase genome size (Adams & Wendel, 2005; Feschotte & Pritham, 2007; Bennetzen & Wang, 2014; Soltis *et al.*, 2015). To counteract genome expansion, several mechanisms have been suggested for DNA loss such as the loss of entire chromosomes (in polyploids) as well as transposon-mediated excision, replication slippage and 'illegitimate recombination' (Petrov *et al.*, 1996; Devos *et al.*, 2002; Otto, 2007; Hu *et al.*, 2011; Schubert & Lysak, 2011). All of these mechanisms represent a challenge for the maintenance of genome integrity. The majority of instability-causing events including transposon-mediated excision and 'illegitimate recombination' are considered to be a byproduct of aberrant DSB-repair processes (Hedges & Deininger, 2007; Buchmann *et al.*, 2012; Schubert & Vu, 2016). Importantly, the expansion through retrotransposon proliferation can be reversed by the high rate of DNA removal in the small-sized genomes, as demonstrated in some cotton species (*Gossypium*) by Hawkins *et al.* (2009). These observations suggest that inaccurate DSB repair could be an underestimated but potent force in shaping eukaryotic genomes and a significant source of genome variation. Although direct ligation and repair by HR are pathways that usually maintain sequence accuracy, erroneous DSB repair has the potential to add and/or remove DNA from the genome and thus could play a role in genome size evolution. We analyzed erroneous DSB repair at three different model loci in *Arabidopsis thaliana* and barley. Our data show significant differences in the outcome between *A. thaliana* and barley. These data provide circumstantial evidence that erroneous DSB repair may indeed have a hitherto underestimated role in genome size evolution.

In detail, our results show significantly more and larger deletions after DSB repair for *A. thaliana* than for barley (Fig. 1), but more and larger insertions in barley than *A. thaliana* (Fig. 2). An inverse correlation of the net loss of DNA at break positions (in both species) to genome size has been observed (Table 2), indicating that the global outcome of DSB repair in longer terms has the potential to affect genome size, mediating progressive shrinkage. Alternatively, genome expansion mediated through an insertion bias and 'due to infrequent DNA removal' (e.g. by reduced or lacking DSB repair-mediated deletions according to our model), as claimed by Kelly *et al.* (2015) for *Fritillaria* species (1C ~100 Gbp), is imaginable for the evolutionary origin of very large genomes (>10 Gbp).

In a previous study, species-specific differences of nonhomologous end-joining (NHEJ) have been proposed based on comparison of DSB repair of 40 selected breakpoint junction sequences from transgenic *A. thaliana plants* with 40 sequences from tobacco calli which contained deletions between 200 and 2300 bp (Kirik *et al.*, 2000). No insertions, but larger deleted sequences could be detected in *Arabidopsis*, whereas 16 sequences (40%) showed insertions in tobacco. In a later study, Lloyd *et al.* (2012) observed 'similar patterns of nonhomologous DSB repair in tobacco and *Arabidopsis*'; however, the design of their study prevented detection of deletions larger than 750 bp which were included in Kirik *et al.* (2000) as well as in our study, and might have blurred differences in repair between species in the investigation by Lloyd *et al.* We here show that DSB repair at three different transgenic loci (GU.US, DU.GUS, IU.GUS) resulted in larger deletions in *A. thaliana.* In particular the single strand annealing (SSA) mechanism, leading to sequence deletions between direct repeats, as well as NHEJ, might be candidates responsible for highly efficient genome shrinking (Table S2). This is likely the reason for the very low proportion of mobile elements (23.7%) in the *A. thaliana* genome (vs 29.7% in the *A. lyrata* genome, which is only 80 Mbp larger), and as uncovered by comparative whole-genome sequencing data (Hu *et al.*, 2011). These authors showed an apparent shrinkage by ~80 Mbp of the *A. thaliana* genome compared to the *A. lyrata* genome, mainly as a consequence of hundreds of thousands of deletions (occurring in addition to small insertions) in *A. thaliana,* mostly involving transposons and noncoding DNA. Among the low number of retrotransposons in *A. thaliana*, only ~30% are intact elements. The proportion of truncated ones is a bit larger, and even more are represented as solo large terminal repeats (LTRs; Devos *et al.*, 2002). Although truncated elements may result from NHEJ of DSBs, solo LTRs are most likely the outcome of SSA within or between LTR retro-elements in a similar manner as our phylogenomics data suggest even for repeat-containing genes. Thus, in particular SSA, but also NHEJ, can cause deletions of different sizes and thus can lead to evolutionary genome shrinkage. The truncated retro-elements and solo LTRs suggest that ~70% of retrotransposon sequences in the *A. thaliana* genome are no longer autonomous and thus cannot contribute to genome expansion. Therefore, deletion-biased DSB repair in mobile elements has a two-fold effect: truncation leads to direct genome shrinking and to a low number of potentially active elements. Truncated retroelements can, in addition to epigenetic silencing of complete ones, no longer compensate (or overcompensate) by retro-element proliferation the genome shrinking within the *A. thaliana* genome. The process of DNA loss seems to be still ongoing in *A. thaliana* because segregating deletions are found in the majority of 95 *A. thaliana* individuals and are on average longer than insertions (Nordborg *et al.*, 2005). The risk of loss of essential sequences during continuing decrease of genome size might be overcome by recurrent whole genome duplication, which seems to be frequent in phylogenetic branches with very small genomes (Ibarra-Laclette *et al.*, 2013; Vu *et al.*, 2015; Schubert & Vu, 2016). On the one hand, it seems reasonable to suggest that more extended end resections mediate a significantly deletion-biased DSB repair as a source of genome shrinkage in *A. thaliana.* On the other, a more efficient protection of free break ends against extensive degradation, and at the same time the capability of single strand overhangs to invade, anneal and copy a template sequence, eventually leads to more insertions in barley. Such insertions compensate in terms of length (at least partially) the (shorter) deletions, arising during DSB repair in barley. Therefore, the tendency of genome size evolution in barley seems to be a weak (close to equilibrium) decrease (Table 2). In spite of its 35-times larger size, the barley genome does not display a net insertion bias. Considering that most phyla of higher eukaryotes comprise many species with a genome size *c.* 1 to < 10 Gbp, it seems likely that for many of them balanced DSB repair with no strong bias towards either deletion or insertions (and thus rather stable genome size) is a typical feature (Schubert & Vu, 2016). Data regarding the spectrum of DSB repair products from more species, especially those with very large genomes (> 10 Gbp) (e.g. conifers or lilies), are needed to further elucidate the impact of DSB repair pathways on genome size evolution.

Acknowledgements

We thank: Holger Puchta, KIT Karlsruhe, for seeds of transgenic *Arabidopsis* lines; Armin Meister, IPK, for discussion of statistics; Jörg Fuchs, IPK, Holger Puchta for reading the article and helpful suggestions; and Bruno Huettel, MPIPZ Cologne, for support with PacBio sequencing. We also thank Andrea Kunze for technical assistance. This study is supported by the German Research Foundation (SCHU 951/18-1).

Author contributions

G.T.H.V. and I.S. designed the research; G.T.H.V., H.X.C. and B.R. performed the research; G.T.H.V., H.X.C. and I.S. analyzed data; G.T.H.V. and I.S. wrote the article; and all authors read and approved the article.

References

Adams KL, Wendel JF. 2005. Polyploidy and genome evolution in plants. *Current Opinion in Plant Biology* **8**: 135–141.

Aguilera A, Garcia-Muse T. 2013. Causes of genome instability. *Annual Review of Genetics* **47**: 1–32.

Aguilera A, Gomez-Gonzalez B. 2008. Genome instability: a mechanistic view of its causes and consequences. *Nature Reviews Genetics* **9**: 204–217.

Bennett MD, Leitch IJ. 2005. Nuclear DNA amounts in angiosperms: progress, problems and prospects. *Annals of Botany* **95**: 45–90.

Bennett MD, Leitch IJ, Price HJ, Johnston JS. 2003. Comparisons with *Caenorhabditis* (approximately 100 Mb) and *Drosophila* (approximately 175 Mb) using flow cytometry show genome size in Arabidopsis to be approximately 157 Mb and thus approximately 25% larger than the Arabidopsis genome initiative estimate of approximately 125 Mb. *Annals of Botany* **91**: 547–557.

Bennetzen JL, Wang H. 2014. The contributions of transposable elements to the structure, function, and evolution of plant genomes. *Annual Review of Plant Biology* **65**: 505–530.

Bleuyard JY, Gallego ME, White CI. 2006. Recent advances in understanding of the DNA double-strand break repair machinery of plants. *DNA Repair (Amst)* **5**: 1–12.

Britt AB. 1996. DNA damage and repair in plants. *Annual Review of Plant Physiology and Plant Molecular Biology* **47**: 75–100.

Buchmann JP, Matsumoto T, Stein N, Keller B, Wicker T. 2012. Inter-species sequence comparison of *Brachypodium* reveals how transposon activity corrodes genome colinearity. *Plant Journal* **71**: 550–563.

Devos KM, Brown JKM, Bennetzen JL. 2002. Genome size reduction through illegitimate recombination counteracts genome expansion in Arabidopsis. *Genome Research* **12**: 1075–1079.

Edgar RC. 2010. Search and clustering orders of magnitude faster than BLAST. *Bioinformatics* **26**: 2460–2461.

Feschotte C, Pritham EJ. 2007. DNA transposons and the evolution of eukaryotic genomes. *Annual Review of Genetics* **41**: 331–368.

Gorbunova V, Levy AA. 1997. Non-homologous DNA end joining in plant cells is associated with deletions and filler DNA insertions. *Nucleic Acids Research* **25**: 4650–4657.

Greilhuber J, Borsch T, Müller K, Worberg A, Porembski S, Barthlott W. 2006. Smallest angiosperm genomes found in Lentibulariaceae, with chromosomes of bacterial size. *Plant Biology (Stuttgart)* **8**: 770–777.

Hawkins JS, Proulx SR, Rapp RA, Wendel JF. 2009. Rapid DNA loss as a counterbalance to genome expansion through retrotransposon proliferation in plants. *Proceedings of the National Academy of Sciences, USA* **106**: 17 811–17 816.

Hedges DJ, Deininger PL. 2007. Inviting instability: transposable elements, double-strand breaks, and the maintenance of genome integrity. *Mutation Research* **616**: 46–59.

Hu TT, Pattyn P, Bakker EG, Cao J, Cheng JF, Clark RM, Fahlgren N, Fawcett JA, Grimwood J, Gundlach H *et al.* 2011. The *Arabidopsis lyrata* genome sequence and the basis of rapid genome size change. *Nature Genetics* **43**: 476–481.

Ibarra-Laclette E, Lyons E, Hernandez-Guzman G, Perez-Torres CA, Carretero-Paulet L, Chang TH, Lan T, Welch AJ, Juarez MJ, Simpson J *et al.* 2013. Architecture and evolution of a minute plant genome. *Nature* **498**: 94–98.

Jasin M, Rothstein R. 2013. Repair of strand breaks by homologous recombination. *Cold Spring Harbour Perspectives in Biology* **5**: a012740.

Kelly LJ, Renny-Byfield S, Pellicer J, Macas J, Novak P, Neumann P, Lysak MA, Day PD, Berger M, Fay MF *et al.* 2015. Analysis of the giant genomes of *Fritillaria* (Liliaceae) indicates that a lack of DNA removal characterizes extreme expansions in genome size. *New Phytologist* **208**: 596–607.

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

Lieber MR. 2010. The mechanism of double-strand DNA break repair by the nonhomologous DNA end-joining pathway. *Annual Review of Biochemistry* **79**: 181–211.

Lloyd AH, Wang D, Timmis JN. 2012. Single molecule PCR reveals similar patterns of non-homologous DSB repair in tobacco and Arabidopsis. *PLoS One* **7**: e32255.

Lysak MA, Berr A, Pecinka A, Schmidt R, McBreen K, Schubert I. 2006. Mechanisms of chromosome number reduction in *Arabidopsis thaliana* and related Brassicaceae species. *Proceedings of the National Academy of Sciences, USA* **103**: 5224–5229.

Mimitou EP, Symington LS. 2009. DNA end resection: many nucleases make light work. *DNA Repair* **8**: 983–995.

Nordborg M, Hu TT, Ishino Y, Jhaveri J, Toomajian C, Zheng H, Bakker E, Calabrese P, Gladstone J, Goyal R *et al.* 2005. The pattern of polymorphism in *Arabidopsis thaliana*. *PLoS Biology* **3**: e196.

Orel N, Kyryk A, 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. *The Plant Journal* **35**: 604–612.

Orel N, Puchta H. 2003. Differences in the processing of DNA ends in *Arabidopsis thaliana* and tobacco: possible implications for genome evolution. *Plant Molecular Biology* **51**: 523–531.

Otto SP. 2007. The evolutionary consequences of polyploidy. *Cell* **131**: 452–462.

Pellicer J, Fay MF, Leitch IJ. 2010. The largest eukaryotic genome of them all? *Botanical Journal of the Linnean Society* **164**: 10–15.

Petrov DA, Lozovskaya ER, Hartl DL. 1996. High intrinsic rate of DNA loss in Drosophila. *Nature* **384**: 346–349.

Petrov DA, Sangster TA, Johnston JS, Hartl DL, Shaw KL. 2000. Evidence for DNA loss as a determinant of genome size. *Science* **287**: 1060–1062.

Pettersson ME, Kurland CG, Berg OG. 2009. Deletion rate evolution and its effect on genome size and coding density. *Molecular Biology and Evolution* **26**: 1421–1430.

Polo SE, Jackson SP. 2011. Dynamics of DNA damage response proteins at DNA breaks: a focus on protein modifications. *Genes & Development* **25**: 409–433.

Puchta H. 2005. The repair of double-strand breaks in plants: mechanisms and consequences for genome evolution. *Journal of Experimental Botany* **56**: 1–14.

Roth N, Klimesch J, Dukowic-Schulze S, Pacher M, Mannuss A, Puchta H. 2012. The requirement for recombination factors differs considerably between different pathways of homologous double-strand break repair in somatic plant cells. *Plant Journal* **72**: 781–790.

Salomon S, Puchta H. 1998. Capture of genomic and T-DNA sequences during double-strand break repair in somatic plant cells. *EMBO Journal* **17**: 6086–6095.

San Filippo J, Sung P, Klein H. 2008. Mechanism of eukaryotic homologous recombination. *Annual Review of Biochemistry* **77**: 229–257.

Schubert I, Lysak MA. 2011. Interpretation of karyotype evolution should consider chromosome structural constraints. *Trends in Genetics* **27**: 207–216.

Schubert I, Vu GT. 2016. Genome stability and evolution: attempting a holistic view. *Trends in Plant Science* **21**: 749–757.

Serra H, Da Ines O, Degroote F, Gallego ME, White CI. 2013. Roles of XRCC2, RAD51B and RAD51D in RAD51-independent SSA recombination. *PLoS Genetics* **9**: e1003971.

Sfeir A, Symington LS. 2015. Microhomology-mediated end joining: a back-up survival mechanism or dedicated pathway? *Trends in Biochemical Sciences* **40**: 701–714.

Slotte T, Hazzouri KM, Agren JA, Koenig D, Maumus F, Guo YL, Steige K, Platts AE, Escobar JS, Newman LK *et al.* 2013. The *Capsella rubella* genome and the genomic consequences of rapid mating system evolution. *Nature Genetics* **45**: 831–835.

Soltis PS, Marchant DB, Van de Peer Y, Soltis DE. 2015. Polyploidy and genome evolution in plants. *Current Opinion in Genetics & Development* **35**: 119–125.

Symington LS, Gautier J. 2011. Double-strand break end resection and repair pathway choice. *Annual Review of Genetics* **45**: 247–271.

Villarreal DD, Lee K, Deem A, Shim EY, Malkova A, Lee SE. 2012. Microhomology directs diverse DNA break repair pathways and chromosomal translocations. *PLoS Genetics* **8**: e1003026.

Vu GT, Cao HX, Watanabe K, Hensel G, Blattner FR, Kumlehn J, 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 GTH, Schmutzer T, Bull F, Cao HX, Fuchs J, Tran TD, Jovtchev G, Pistrick K, Stein N, Pecinka A *et al.* 2015. Comparative genome analysis reveals divergent genome size evolution in a carnivorous plant genus. *Plant Genome* **8** doi: 10.3835/plantgenome2015.04.0021.

Watanabe K, Pacher M, Dukowic S, Schubert V, Puchta H, Schubert I. 2009. The STRUCTURAL MAINTENANCE OF CHROMOSOMES 5/6 complex promotes sister chromatid alignment and homologous recombination after DNA damage in *Arabidopsis thaliana*. *The Plant Cell* **21**: 2688–2699.

Waterworth WM, Drury GE, Bray CM, West CE. 2011. Repairing breaks in the plant genome: the importance of keeping it together. *New Phytologist* **192**: 805–822.

Yu AM, McVey M. 2010. Synthesis-dependent microhomology-mediated end joining accounts for multiple types of repair junctions. *Nucleic Acids Research* **38**: 5706–5717.

Supporting Information

Additional Supporting Information may be found online in the Supporting Information tab for this article:

Table S1 Design of PacBio amplicon sequencing libraries

Table S2 Mutation classes detected in GU.US, DU.GUS and IU.GUS construct-derived libraries

Table S3 Orthologous genes in *Arabidopsis thaliana*, *Arabidopsis lyrata* and *Capsella rubella*, which are >1 kbp shorter in *A. thaliana* (group A = 50 genes) or in *A. lyrata* (group B = 25 genes) than in the other two species

Please note: Wiley Blackwell are not responsible for the content or functionality of any Supporting Information supplied by the authors. Any queries (other than missing material) should be directed to the *New Phytologist* Central Office.