Published online 12 July 2022

Nucleic Acids Research, 2022, Vol. 50, No. 13 ***7465–7478***
https://doi.org/10.1093/nar/gkac575

Characterization of sequence contexts that favor alternative end joining at Cas9-induced double-strand breaks

Terrence Hanscom[1,†], Nicholas Woodward[1,†], Rebecca Batorsky[2], Alexander J. Brown[3], Steven A. Roberts[3] and Mitch McVey[1,*]

[1]Department of Biology, Tufts University, 200 Boston Avenue, Suite 4700, Medford, MA 02155, USA, [2]Data Intensive Studies Center, Tufts University, 177 College Ave, Medford, MA 02155, USA and [3]School of Molecular Biosciences, Washington State University, P100 Dairy Road, Pullman, WA 99164, USA

Received May 28, 2021; Revised June 16, 2022; Editorial Decision June 17, 2022; Accepted June 20, 2022

ABSTRACT

Alternative end joining (alt-EJ) mechanisms, such as polymerase theta-mediated end joining, are increasingly recognized as important contributors to inaccurate double-strand break repair. We previously proposed an alt-EJ model whereby short DNA repeats near a double-strand break anneal to form secondary structures that prime limited DNA synthesis. The nascent DNA then pairs with microhomologous sequences on the other break end. This synthesis-dependent microhomology-mediated end joining (SD-MMEJ) explains many of the alt-EJ repair products recovered following I-*Sce*I nuclease cutting in Drosophila. However, sequence-specific factors that influence SD-MMEJ repair remain to be fully characterized. Here, we expand the utility of the SD-MMEJ model through computational analysis of repair products at Cas9-induced double-strand breaks for 1100 different sequence contexts. We find evidence at single nucleotide resolution for sequence characteristics that drive successful SD-MMEJ repair. These include optimal primer repeat length, distance of repeats from the break, flexibility of DNA sequence between primer repeats, and positioning of microhomology templates relative to preferred primer repeats. In addition, we show that DNA polymerase theta is necessary for most SD-MMEJ repair at Cas9 breaks. The analysis described here includes a computational pipeline that can be utilized to characterize preferred mechanisms of alt-EJ repair in any sequence context.

INTRODUCTION

DNA double strand breaks (DSB) are dangerous lesions that must be repaired accurately to maintain genome stability (1). DSB repair generally proceeds through one of two main mechanisms. The first, homologous recombination (HR), typically occurs during the S and G2 phases of the cell cycle and involves resection at the break, ssDNA invasion into a homologous template, synthesis of new DNA, and ligation to complete repair. The second, non-homologous end-joining (NHEJ), takes place throughout the cell cycle and involves direct ligation of the broken ends with minimal processing (2).

While both homologous recombination and NHEJ are largely error-free, NHEJ sometimes generates 1–4 base pair (bp) deletions and insertions.

A more mutagenic form of DSB repair, known as alternative end-joining (alt-EJ), has been shown to occur both in the absence and presence of NHEJ and HR (3). The A-family DNA polymerase theta is responsible for much of this type of repair, which typically involves annealing of latent or newly synthesized 1–8 bp microhomologous sequences present near the ends of the break to facilitate rejoining (4). Because of the central role of pol theta in this process, this type of alt-EJ is often called theta-mediated end joining (TMEJ) (5–8).

Studies in *Caenorhabditis elegans*, Drosophila, and mammalian cells have shown that many alt-EJ repair products are accompanied by insertions templated from nearby flanking sequences (9–11). We have postulated a mechanism by which these insertions can be generated, called synthesis-dependent microhomology-mediated end-joining (SD-MMEJ) (10,12). In this model, short, complementary sequences of 3′ single-stranded DNA (ssDNA) at the DSB, known as primer repeats, form secondary structures such as hairpins and loops, whose 3′ ends can then be extended (Figure 1 and Supplementary Figure S1). The newly synthe-

*To whom correspondence should be addressed. Tel: +1 617 627 4196; Email: mitch.mcvey@tufts.edu
†The authors wish it to be known that, in their opinion, the first two authors should be regarded as Joint First Authors.

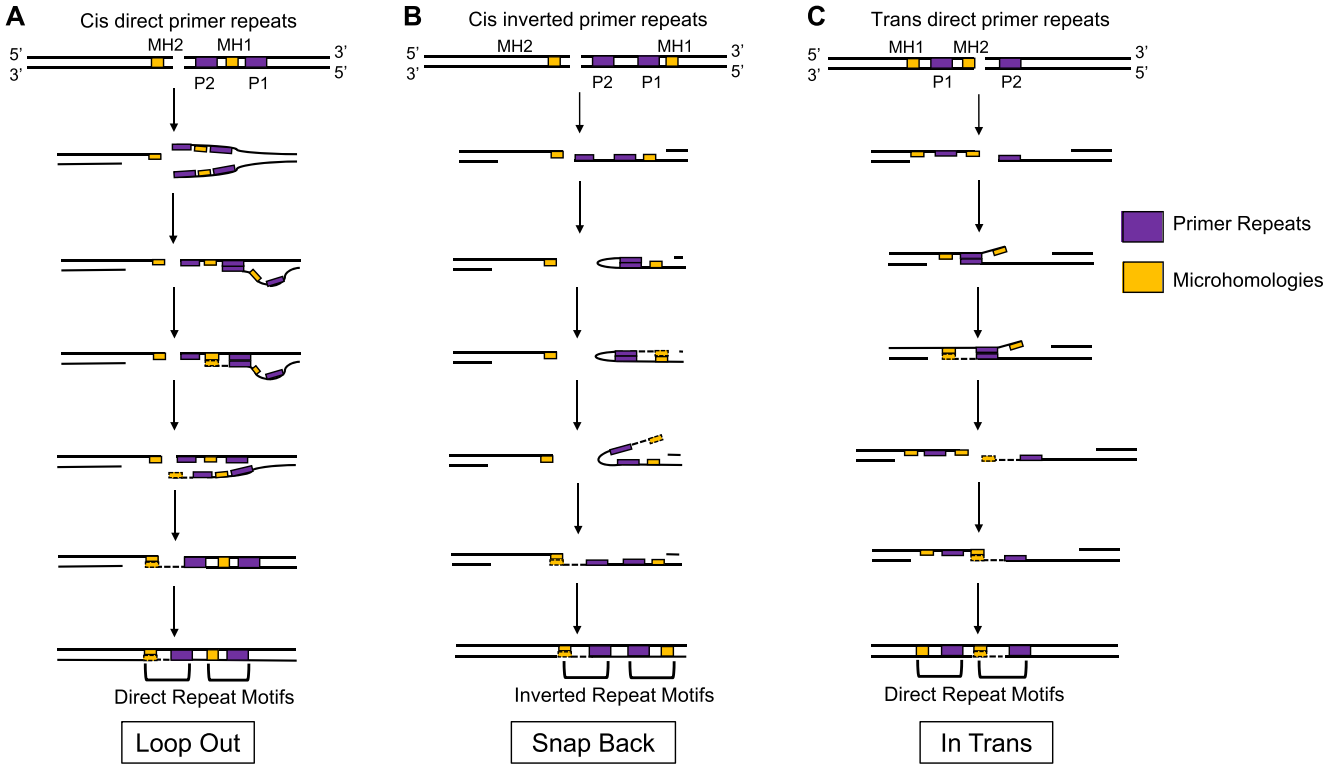


Figure 1. The SD-MMEJ model for alternative end joining repair. (**A**) Loop-out mechanism with DNA unwinding prior to loop formation. (**B**) Snap-back mechanism with DNA resection (or unwinding, not shown) prior to hairpin formation. (**C**) *In trans* mechanism with DNA resection (or unwinding, not shown). All mechanisms utilize annealing of break-proximal primer repeats (P2, purple) to break-distal primer repeats (P1, purple), which primes nascent synthesis that can lead to insertions (black, dashed) and the creation of nascent microhomologous sequences (yellow, dashed). For loop-out and in trans SD-MMEJ, P1 and P2 are direct repeats, while for snap-back they are inverted repeats. Repair concludes with unwinding of secondary structures, annealing of nascent microhomology with MH2 sequences (yellow) on the other side of the break, fill-in synthesis, and ligation. For the repair events shown here, the inserted sequence becomes part of longer direct or inverted repeats. Not shown are the trimming of non-homologous flap intermediates when P2 and MH2 are not directly adjacent to the break site, or simple deletions that are formed when P1 and MH1 are directly adjacent to each other.

sized DNA anneals to microhomologous sequences on the other side of the break, facilitating repair.

Murine pol theta was recently shown to engage in a scanning mechanism to identify microhomologies near double-strand break ends (9). It does this bidirectionally from the 3′ termini for up to 15 nucleotides (nt) and favors break-proximal microhomologies ≥2 bp. Consistent with the SD-MMEJ model, when microhomologies are not present, pol theta can generate insertions of 5 bp or more, templated from sequences within 50 bp of the DSB. These templated insertions are a marker for TMEJ activity and are enriched in BRCA1/2-deficient cells (13). Intriguingly, 5–10% of Cas9-induced DSBs engage in TMEJ in the presence of other, higher fidelity DSB repair mechanisms (5,9).

We have shown that SD-MMEJ is a robust repair process in the context of I-*Sce*I generated DSBs in Drosophila (10,12). We previously demonstrated that specific sequences near an I-*Sce*I generated DSB act as preferred primer repeats during SD-MMEJ. Additionally, we showed that mutating a single base pair in one of these repeats decreases its ability to drive SD-MMEJ repair in a predictable manner (12). To further understand the underlying process of alt-EJ/TMEJ, we developed computational tools that can assess the likelihood that any repair event was created through SD-MMEJ.

SD-MMEJ repair products are identified by sequence signatures called repeat motifs (Figure 1 and Supplementary Figure S1). Repeat motifs are found at the repair junction and in the DNA adjacent to the break; they include one-half of a primer repeat and one-half of a microhomology repeat. Often, a repeat motif also contains an insertion between the primer and microhomology half-repeats. While it is likely that some complex insertions are generated by multiple rounds of annealing, synthesis, and dissociation prior to final pairing (10), for simplicity our computational model requires that all SD-MMEJ consistent repair events must be generated from a single templated synthesis and microhomology annealing event (a single step event). In addition, all SD-MMEJ consistent events require repeat motifs at the break junction of four bp or greater that include one half of the primer repeat, the nascent microhomology, and any inserted nucleotides.

SD-MMEJ can produce three kinds of repair junctions: insertions/indels, microhomology joins (MHJ), and apparent blunt joins (ABJ). Insertions/indels generate a repeat motif at the repair junction that includes an insertion between the primer and the microhomology. MHJ and ABJ repeat motifs include only the primer and microhomology, with no insertion. While repair events containing insertions can be unambiguously characterized as occurring through

SD-MMEJ (14), MHJ and ABJ events can also be created through classical end-joining repair (10).

SD-MMEJ repair can take place *in cis* or *in trans* (Figure 1 and Supplementary Figure S1). *In cis*, two distinct mechanisms, loop-out and snap-back, generate intra-strand secondary structures that prime new synthesis. Loop-out SD-MMEJ generates direct repeat motifs, while snap-back generates inverted repeat motifs. For *in trans* SD-MMEJ, an interstrand primary synthesis event is mediated by annealing of direct repeats across the break. Of note, loop-out and *in trans* mechanisms can generate the same repair events with swapping of the primer and microhomology repeats. For these events, we are unable to assign mechanism based on the final repair junction (12).

The CRISPR-Cas9 system has become a popular system for generating DSBs in genome editing experiments. Recently, several papers have described predictive algorithms for Cas9-induced break repair products (15–19). These algorithms can satisfactorily predict indels and deletions caused by NHEJ and MMEJ but fall short in their treatment of templated insertions, which also occur during Cas9-genome editing (9). Therefore, we sought to investigate SD-MMEJ repair at single nucleotide resolution in the context of Cas9-induced DSBs in Drosophila, where insertional repair is frequent.

We generated 1100 semi-randomized DNA plasmid constructs, injected them with a targeting sgRNA into Cas9-expressing embryos, recovered the plasmids after a suitable incubation time, and characterized the repair events by deep sequencing. By analyzing SD-MMEJ consistent repair products across all the constructs, we found that SD-MMEJ repair of Cas9 breaks is optimized by the presence of 1–3 nt primer repeats in close proximity to the break. The best primer repeats are separated by 5–6 nt and form stem–loop structures with flexible loops. In addition, the availability of microhomology templates that can pair with ssDNA near the 3′ terminus of the other break end strongly drives SD-MMEJ repair. For researchers wanting to examine whether gene editing sequence contexts are favorable for the recovery of specific alt-EJ repair products, we have made our SD-MMEJ analysis programs freely available.

MATERIALS AND METHODS

Creation of sequence-randomized end-joining substrates

Sequence-randomized constructs were created by designing a 165 nt ssDNA fragment based on R0 with randomized nucleotides adjacent to the I-*Sce*I/Cas9 break site. The PAGE purified semi randomized ssDNA oligo (IDT) was made double stranded via primer extension using Q5 polymerase (NEB). The product was treated with S1 nuclease (ThermoFisher Scientific) to remove remaining ssDNA fragments. 600 μl of phenol:chloroform:isoamyl alcohol (25:24:1) was added to each sample, followed by 15 min on a rocking platform. The samples were centrifuged for 15 min and the aqueous layer was collected. An equal volume of phenol:chloroform:isoamyl alcohol was added, followed by 15 min on a rocking platform and 15 min of centrifugation. The aqueous layer was collected and one-fifth volume of 8 M potassium acetate was added followed by one volume of chloroform. The samples were rocked on a platform for 15 min followed by a 15-min centrifugation. The aqueous layer was collected and 0.7 volumes of isopropyl alcohol and 1 μl glycoblue was added. The samples were incubated at −80°C for 30 min, centrifuged for 15 min and the pellets were washed with 200 μl of 70% ethanol. The samples were centrifuged for 15 min and the pellets were dried and resuspended in ddH_2O.

The dsDNA variant inserts were cloned into pMiniT2.0 using the NEB PCR cloning kit and transformed into NEB 10-beta chemically competent cells. Transformants were grown on LB + ampicillin plates and diluted to yield approximately 150 colonies per plate. Each plate was washed into LB broth with ampicillin, grown overnight, and midi prepped using the Macherey-Nagel NucleoBond Xtra Midi kit. Plasmid DNA libraries were amplified by PCR with Q5 polymerase (NEB) for 19 cycles using an Eppendorf Vapo Protect thermocycler with a pooled set of primers containing one, two, or three random bases at the 5′ end. AMPure bead purification was performed on the PCR products and the purified DNA was subjected to a second PCR to attach indices for amplicon sequencing. A final AMPure purification step was performed to remove all products less than 100 bp.

Samples were pooled with 5% PhiX DNA and sequenced using the Illumina Nano chip for reference library generation and selection of experimental samples. Experimental libraries were chosen based on relative concentration of individual plasmids and combined to generate four distinct injection libraries with a total of 1097 plasmids.

Creation of R0 flex, rigid and T-loop constructs

Site directed mutagenesis was used to create the R0 flex, rigid and T-loop constructs. Primers complementary to R0 were designed to mutate the 5 bp between the GGCC direct and inverted repeat on the right side of the break site (Eton Biosciences). NEB's Site-Directed Mutagenesis Kit was used. PCR products containing the desired mutations were created using Phusion high fidelity DNA polymerase (NEB). The products were subjected to Kinase, Ligase, and DpnI treatment using provided buffer and enzyme mix. Treated products were transformed into NEB 5-alpha competent cells and plated on selective media with 50 μg/ml spectinomycin. Colonies were grown in LB broth with antibiotic selection and plasmids were isolated by alkaline lysis. Individual products were screened by NotI digestion and Sanger sequencing to confirm successful mutagenesis.

Cloning of Cas9 targeting guide RNA expression vector

Cloning of the gRNA expression vector was done using the pU6-BbsI-chiRNA vector (Addgene plasmid #45946) and annealed complementary oligos (Eton Biosciences) for the R0 region of interest. Electroporation of the ligation reaction was performed and transformants were screened via sequencing. A single correct transformant was grown in LB broth with ampicillin and plasmids were purified using the Macherey-Nagel NucleoBond Xtra Midi kit.

Fly stocks

Stocks used in the plasmid injection studies all contained Cas9 and were either wild-type (Bloomington Drosophila

Stock Center stock 54590; y^1, w^*, $M\{w^{+mC}$, *Act5C-Cas9.P}ZH-2A*), *LIG4* deficient (BDSC stock 58492; y^1, w^{1118}, $DNAlig4^{169a}$, $M\{Act5c\text{-}Cas9.P.RFP^-\}ZH\text{-}2A$), or *POLQ* deficient (*Piggybac* insertion 3 nt downstream of the translation start site; w^{1118}, $M\{w^{+mC}$, *Act5C-Cas9.P}ZH-2A; polq*$^{3XFLAG,PBac\{3XP3\text{-}DsRed\}}$). All flies were maintained in bottles containing a cornmeal-agar medium in a 25°C incubator on a 12 h light–dark cycle. Freshly eclosed flies were placed in cages and fed yeast paste on a grape agar substrate to promote embryo laying. Embryos <2 h old were recovered and dechorionated for 2 min in a 50% bleach solution. Embryos ($n = 100$ per experiment) were aligned uniformly on double-stick tape attached to a glass cover slip for injection of plasmids.

Microinjection and DNA recovery

Microinjections were carried out as in (20). Embryos were desiccated for 1 min and coated in halocarbon oil to prevent rupturing of the membrane when injected. Injection libraries were diluted to 500 ng/μl in injection buffer (1 mM sodium phosphate and 50 mM potassium chloride) along with 500 ng/μl gRNA expression vector. The microinjections were done using a Zeiss compound microscope fitted with injection apparatus and a Parker-Hanfin Picospritzer II. Embryos were incubated at 25°C for 24 h to allow for Cas9 cutting and repair of the ensuing double-strand breaks. Halocarbon oil was removed using a 1% sodium dodecyl sulfate solution in 0.7% sodium chloride buffer.

Plasmid DNA was extracted by grinding embryos (two slides using the same injection mix were combined) with a disposable pestle in 200 μl Buffer A (100 mM Tris–HCl, pH 7.5, 100 mM EDTA, 100 mM NaCl, 0.5% SDS). After incubating at 65°C for 30 min, 800 μl of LiCl/KAc solution (1 part 5 M KAc : 2.5 parts 6 M LiCl) was added and tubes were incubated on ice for 10 min. The solution was centrifuged for 10 min and plasmids were precipitated from the supernatant with isopropanol.

Repair junction sequencing

Purified plasmid DNA containing repair junctions were enriched for error-prone repair events by *in vitro* I-*Sce*I treatment for 2 h to remove uncut plasmids or perfect repair events. For high-throughput amplicon sequencing, approximately 165 bp of sequence flanking the Cas9 cut site was amplified from recovered plasmids by PCR with Q5 polymerase (NEB) for 19 cycles using an Eppendorf Vapo Protect thermocycler with a pooled set of primers containing one, two or three random bases at the 5′ end. AMPure bead purification was performed on the PCR products and the purified DNA was subjected to a second PCR to attach indices for amplicon sequencing. A final AMPure purification step was performed to remove all products less than 100 bp. The samples were pooled with 10% PhiX DNA and sequenced on an Illumina HiSeq platform using 2 × 300 paired-end reads.

Analysis of high-throughput amplicon sequencing

Paired-end reads corresponding to the same end joining event were merged into single consensus reads as FASTQ files using the BBMerge module in Geneious R10 (BioMatters). The 5′ and 3′ ends of the resulting sequences were trimmed to the amplicon primer sequences and an additional 80 nt of chosen DNA sequence was added to each end facilitate proper alignment of the reads. Reads lacking either of chosen 10 bp portions of the amplicon primer sequences were removed as PCR artifacts as they failed to span both sides of the break. Reads were then separated into independent files based upon the presence of unique 5′ and 3′ barcode sequences that correspond to those of the originating DNA substrate that was injected into fly embryos. Reads lacking appropriate barcodes were excluded from further analysis. Reads for each file were subsequently aligned to the corresponding original repair substrate reference sequence using the BWA mem alignment software (21) with default settings. The resulting SAM files were then subjected to Hi-FiBR analysis (22) to classify junction events. Base substitutions within a 15 bp window surrounding the break site were interpreted as repair errors instead of sequencing artifact. Reads were classified as either exact matches to the original repair substrate, deletions (having one or more contiguous bases deleted adjacent to the break site in the reference), insertions (having one or more contiguous inserted bases adjacent to the break site), or complex (having base substitutions, insertions, and/or deletions that are non-adjacent to the break site or contain both deleted and inserted sequences). The number of instances of each specific repair events was counted. Junctions with <10 mapped reads were removed from consideration. Following removal of these junctions, the normalized percentage of reads that contain sequence alterations compared to the reference sequence per junction was calculated by dividing the number of reads per junction by the total number of inaccurate reads that aligned to any junction in the sample. Deletion junctions were characterized as apparent blunt joins (ABJ) or microhomology junctions (MHJ) and were analyzed to determine SD-MMEJ consistency using a novel pure Python suffix tree library, which searched for a break-spanning region that contained both microhomology and primer junctions. Insertion junctions were analyzed using a custom designed R-script to determine SD-MMEJ consistency, as defined in the text.

RESULTS

SD-MMEJ is a prominent repair pathway for Cas9-induced double-strand breaks

Our previous studies of SD-MMEJ used a repair construct, designated R0, that contains an I-*Sce*I nuclease recognition site adjacent to multiple short primer repeats (10,12). To compare SD-MMEJ repair of Cas9 breaks in this sequence context, we designed a sgRNA that targets Cas9 cutting 3 bp downstream from the 3′ end of the I-*Sce*I cut site in R0 (Figure 2A). We injected the R0 plasmid with the sgRNA expressing plasmid into wild-type,

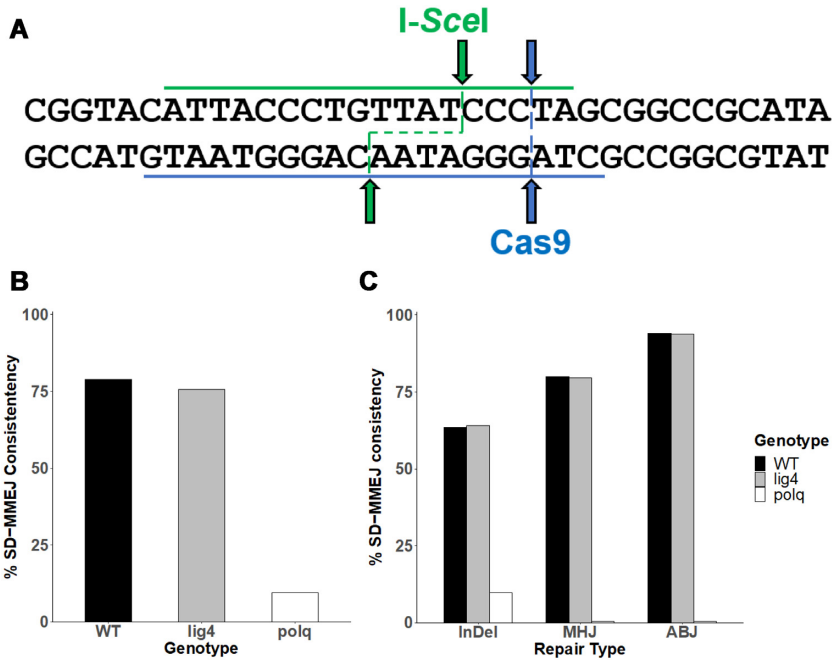


Figure 2. SD-MMEJ repair is heavily reliant on DNA polymerase theta. (**A**) DNA sequence of the R0 construct. Recognition sequences and cut sites for I-*Sce*I (green) and sgRNA-Cas9 (blue) are indicated. (**B**) SD-MMEJ consistency for wild-type (WT), *lig4* and *polq* inaccurate Cas9-induced repair events recovered from the R0 construct. (**C**) SD-MMEJ consistency according to repair junction type for all Cas9-induced inaccurate repair events recovered from the R0 construct. InDel = insertion/deletion junction; MHJ = microhomology join; ABJ = apparent blunt join.

lig4 and *polq* mutant embryos. The absence of DNA ligase 4 abolishes classical NHEJ repair (23,24), while loss of DNA polymerase theta (encoded by PolQ) reduces SD-MMEJ (10). Cut and repaired plasmids were isolated and repair junctions were amplified by PCR, followed by next-generation amplicon sequencing. Error-prone repair events were assigned to each original repair construct and assessed for SD-MMEJ consistency using custom scripts in R and Python.

In both the wild-type and *lig4* mutant backgrounds, more than 75% of the inaccurate repair events that we recovered were consistent with the SD-MMEJ repair model (Figure 2B). Most of these were small insertions created using common primer repeats, suggesting that even when classical end joining is available, SD-MMEJ is frequently used. In contrast, <10% of repair events isolated from *polq* mutant embryos were SD-MMEJ consistent. The number of SD-MMEJ consistent microhomology joins and apparent blunt joins was >75% for wild-type and *lig4* mutants but was reduced to zero in the absence of POLQ. Together, these data indicate that pol theta is vitally important for SD-MMEJ repair of Cas9-induced breaks.

I-*Sce*I produces breaks with four nucleotide 3′ overhangs that are often involved in microhomology annealing during SD-MMEJ (12). Most Cas9-induced breaks have blunt ends (25,26). By comparing the spectrum of inaccurate repair products obtained following cutting by the two nucleases, we could determine the extent to which SD-MMEJ depends on the overhangs generated by I-*Sce*I. Overall, SD-MMEJ consistent repair of a Cas9 break occurred with slightly greater frequency than with an I-*Sce*I break, indicating that 3′ overhangs do not seem to be responsible for driving SD-MMEJ (Supplementary Figure S2).

In the R0 construct, Cas9 cutting creates a blunt-ended break 3 nt downstream of the I-*Sce*I cut site (Figure 2). This repositioning changes the distance from the break site to the frequently used primer repeats, allowing us to query the influence of break proximity on the relative usage of primer repeats. For the analysis presented below, we focused on SD-MMEJ consistent repair products involving single-step insertions.

Overall, we found that the spectrum of SD-MMEJ insertional repair products was similar for both I-*Sce*I and Cas9 breaks (Figure 3). On the right side of the break, a GGCC primer repeat was highly utilized for repair of both types of breaks, although it was more preferentially used for loop-out repair following I-*Sce*I cutting (purple boxes). On the left side of the break, an AT inverted primer repeat and a TTA direct primer repeat were two of the most common primer repeats used for SD-MMEJ repair of both types of breaks (red and blue boxes, respectively).

However, several differences also existed for SD-MMEJ consistent insertional repair of I-*Sce*I and Cas9 breaks. Following I-*Sce*I cutting, the short primer repeats present in the 3′ TTAT/AATA overhangs frequently participate in *trans* SD-MMEJ repair (12). With Cas9, these primer repeats were not used for *in trans* repair, likely because they are not present in pre-resected DNA. In contrast, CC and CCC direct primer repeats at the left 3′ break end were highly utilized for loop-out Cas9 repair (green box), but not for *in trans* repair of I-*Sce*I breaks, where they were further from the break ends.

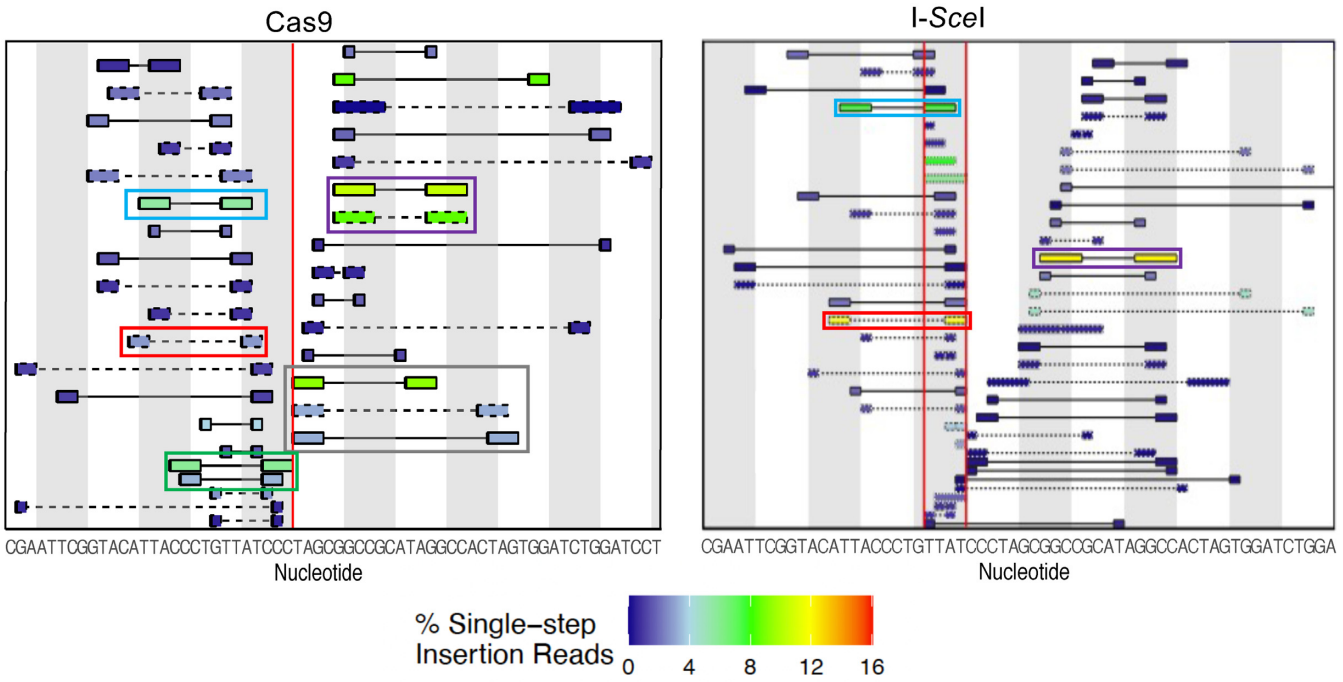


Figure 3. Similar primer repeats are utilized during SD-MMEJ repair of Cas9 and I-*Sce*I induced breaks. All primer pairs used in single step SD-MMEJ insertions are shown for the R0 construct. Different repair junctions may use the same primer pair. Colors correspond to the frequency that each primer pair is utilized with warmer colors indicating greater frequencies. Primer repeats involved in loop-out SD-MMEJ are shown with solid outlines and connecting lines; primer repeats involved in snap-back SD-MMEJ are shown with dashed outlines and connecting lines. Only the top 5′-3′ strand is given. This represents the structure-forming 3′ strand to the left of the break but on the right side the complementary bottom strand is engaging in secondary structure formation. Vertical red lines indicate cutting location with a single cut for Cas9 and TTAT/AATA overhangs produced by I-*Sce*I. I-*Sce*I primer repeat usage data is replicated from (12).

Interestingly, there is a dynamic relationship between the proximity of primer repeats to the break and their length in productive SD-MMEJ repair. For example, AT inverted primer repeats were used relatively more often at I-*Sce*I breaks than at Cas9 breaks, where they are located 3 bp further from the break site (red boxes). However, TTA direct repeat primers, which are 2 bp closer to the end of I-*Sce*I cuts, were used only slightly less frequently for SD-MMEJ repair of Cas9 breaks (blue boxes).

On the right side of the break, there was an overall increase in GGCC primer repeat use at Cas9 breaks compared to I-*Sce*I breaks (20% vs. 12%), which correlates with their closer proximity to the Cas9 break (purple boxes). Strikingly, a TAG repeat present at the right 3′ end of the Cas9 break frequently participated in both loop-out and snap-back SD-MMEJ repair (grey box). However, a longer primer repeat containing this TAG (CTAG) was rarely used for repair of I-*Sce*I breaks, likely because it is located farther from the break end.

In summary, although the different end structures at I-*Sce*I and Cas9 breaks precludes us from being able to definitively assign differences in the recovered repair products to one specific property of the sequence, the collective results suggest that both proximity to the break and primer length are important factors that dictate the usage of primer repeats for synthesis events. Short primer repeats are preferentially used when they are closer to the break. When primers are located further than 1–2 nucleotides from the break, increased primer repeat length becomes more important for successful SD-MMEJ repair.

High-throughput characterization of parameters that drive SD-MMEJ repair of Cas9 breaks

To further investigate how sequence context around a Cas9 break influences repair outcomes in a high-throughput manner, we generated a library of ∼1100 plasmids based on the R0 construct, with partially randomized sequences flanking the Cas9 cut site (Figure 4 and Supplementary Table S1, see also materials and methods). By necessity, the sequence recognized by the sgRNA remained static. In addition, we chose to keep the GGCC repeat on the right side of the break constant, as it was the most frequently used primer repeat during SD-MMEJ repair of both I-*Sce*I and Cas9 breaks. By including it in our constructs, we could therefore investigate how changes in the sequence flanking a dominant primer repeat affect its usage.

Pools of these plasmids were injected into *lig4* embryos and inaccurate repair products were recovered and subjected to Illumina amplicon sequencing (Supplementary Figure S3). Each of the repair constructs generated between 100 and 26,000 inaccurate repair events with 4000 to 2,500,000 reads per construct. Inaccurate repair events for each construct consistently comprised 5–10% of all reads; the other reads corresponded to uncut or accurately repaired plasmids. The full data set is available for further analysis at https://www.ncbi.nlm.nih.gov/sra/PRJNA706449. We then analyzed the inaccurate repair events from the entire collection of constructs to gain insight into the most important parameters that drive SD-

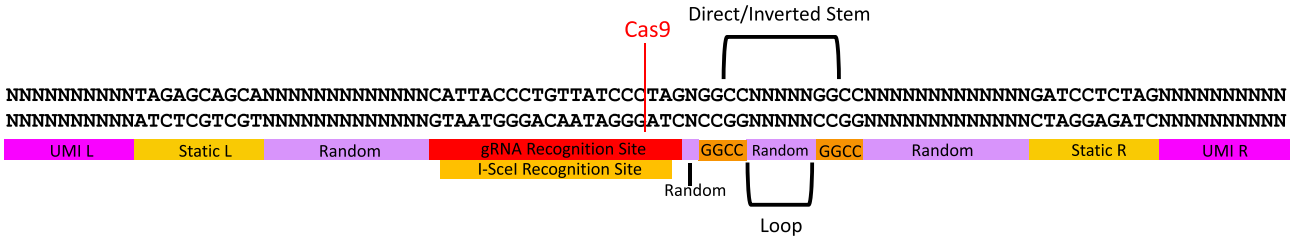


Figure 4. The semi-randomized DNA repair construct used to probe SD-MMEJ sequence preferences. GGCC primer repeats, shown in orange, can form a loop structure with an intervening randomized stem. Other randomized nucleotides are shown in lavender, while unique molecular identifiers (UMIs) for demultiplexing are shown in magenta. The sgRNA recognition sequence and Cas9 cleavage site are shown in red. Static regions, shown in gold, aid in identifying the UMIs.

MMEJ. The computational pipeline and R script for analysis of SD-MMEJ consistency and generation of plots is available at https://gitlab.tufts.edu/mcveylab/sdmmej/.

Primer repeat length and loop length influence the frequency of SD-MMEJ consistent insertions

First, we investigated the relative importance of primer length and loop size during the formation of secondary structures that drive SD-MMEJ synthesis. We previously found that SD-MMEJ repair efficiency of I-*Sce*I breaks decreases as the distance between primer repeats (loop length) increases (12). Analysis of the SD-MMEJ consistent insertion repair events showed that this was also true for Cas9 breaks. Across all constructs, the most utilized primer repeats were separated by fewer than 10 nt (Figure 5A). While we do observe SD-MMEJ consistent repair events created by loop formation of up to 29 nt, these events are rare. Interestingly, 6 nt was the optimal distance between primers, with 40% of all single-step insertions involving the formation of a 6 nt loop. Identical primer repeats are utilized less frequently as the distance between them increases. For example, with the R0 construct, a TAG repeat with an 8 nt loop was used in 9% of insertional junctions, but its use was decreased two-fold when the distance between the repeats increased to 16 nts (Figure 3, grey box).

Intriguingly, short primers were used much more frequently than longer primers. Across all constructs, 90% of all SD-MMEJ consistent insertion events utilized primers that were 3 nt or less (Figure 5B). While shorter primer repeats are more highly represented in our randomized sequences, primer repeats of up to 8 nt can be identified. Because longer primer repeats should form more stable hairpin-loop structures, but they are rarely used during SD-MMEJ repair, we conclude that primer repeats longer than 6 nt are inhibitory to SD-MMEJ.

Similarly, short microhomologies were favored during the final annealing step, with 2 nt microhomologies used in 60% of all single-step insertions (Figure 5C). While these short microhomologies are less thermodynamically stable than longer microhomologies, they appear to be preferred when the complementary sequence is located near the terminus of the other break end (described below).

Examination of SD-MMEJ consistent indel repair events for individual constructs revealed an interplay between primer repeat length and loop length. One example of this can be found in the primer repeat plot for the R983 construct (Supplementary Figure S4A). While 7% of the insertions used a 2 nucleotide TA repeat on the right side of the break during loop-out SD-MMEJ repair when the loop size was 8 nucleotides, the frequency of its use decreased to 2.5% when the distance between these primer repeats increased to 16 nt (Supplementary Figure S4A, blue boxes). In contrast, when the size of the primer repeat was increased by 1 nt to TAG, but the loop length remained at 16 nt, its frequency of usage increased to 7%. Further evidence for can be observed farther from the break on the right side of the R983 construct. A GGCC primer repeat separated by 5 nt was used in 12.5% of all SD-MMEJ consistent indels (Supplementary Figure S4A, red box), The usage of a similarly positioned GGCCT primer repeat was reduced by only 1.5-fold, even though the distance between the repeats was tripled to 15 nt. Previous investigations have found that the rate of closing of stem-loop structures decreases as the loop length increases (27), but our results suggest that this trend can be reversed when additional nucleotides are included in the stem.

Interestingly, there appears to be a mechanism that ensures that secondary structure formation to prime initial SD-MMEJ synthesis occurs near the break. In R983, three GGCC repeats are present, located 4, 13 and 23 nts from the break. The middle and break-distal GGCC primer repeats are never used in successful SD-MMEJ repair events, indicating that, given a choice between identical primer repeats at different distances from the break, the repeats closest to the break are favored.

To further explore positively reinforcing trends between primer repeat length and loop length, we examined insertional repair products from the R958 construct, which has abundant GC content in the randomized region to the right of the break and contains many available overlapping primer repeats which were used in SD-MMEJ repair (Supplementary Figure S4B). We observed that a 6 nt primer repeat was used less often to generate single-step insertions compared to shorter 3 and 4 nt primer repeats contained within the larger repeat (Supplementary Figure S4C). A 5 nt AGGGG primer repeat separated by only 2 nt was used at half the frequency of a 4 nt GGCC primer repeat separated by 4 nt. In addition, repeats of 2 nt or less, or repeats separated by more than 4 nt, were not utilized frequently during the initial step of SD-MMEJ repair for this construct. Thus, during the initial formation of secondary structures during SD-MMEJ, it appears that there are optimal values for both primer repeat length and loop length that provide the highest probability of successful repair. However, because these

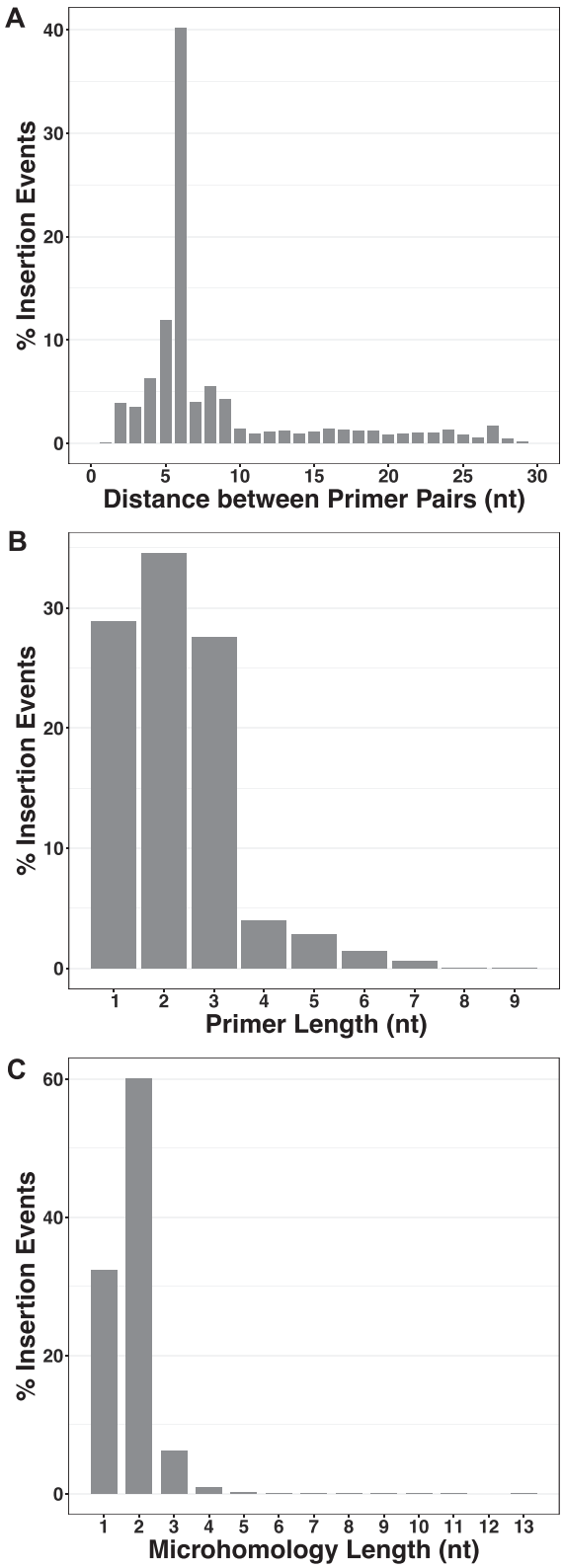


Figure 5. Preferred SD-MMEJ values for single-step insertion events. (**A**) Distance between primer pairs (loop length). (**B**) Primer length. (**C**) Microhomology length (during annealing across the break). Analysis was performed for all 1097 constructs in the library. Each construct was weighted equally in the analysis to prevent overrepresentation of constructs that had more sequencing reads.

can be construct-specific, we continued to probe other variables that might be important.

Flexible loops between primer repeats promote SD-MMEJ while rigid loops inhibit synthesis

We wondered if the DNA sequence between the primer repeats might also impact the formation of stem-loop structures and therefore contribute to SD-MMEJ proficiency. For example, repair could be affected by the flexibility of the loop, or by base pairing within the loop that stabilizes the stem-loop. To investigate these parameters, we examined repair junctions from repair constructs with either homopolymeric A sequences or alternating AT dipolymeric sequences in the single-stranded loop that would form between the GGCC primer repeat. Homopolymeric A sequences are rigid and prevent the formation of stem-loop structures, while AT dipolymeric sequences are flexible and promote stem-loop formation (27,28). We refer to the homopolymer A loops as 'rigid loops,' and alternating AT dipolymer loops as 'flexible loops.'

We identified six constructs with rigid loops between the GGCC repeats (R152, R247, R684, R704, R810 and R1038) and eight constructs with flexible loops (R16, R113, R246, R290, R387, R494, R583 and R983). Interestingly, we recovered many fewer inaccurate sequence reads that mapped to the rigid loop constructs, compared to the flexible loop constructs. Furthermore, the percentage of inaccurate repair events that were consistent with SD-MMEJ repair was overall lower for the rigid constructs, especially for indels and apparent blunt joins (Supplementary Figure S5). The decreased SD-MMEJ consistency for apparent blunt joins suggests that this type of repair often occurs through alt-EJ and not just via classical end joining. Examination of the primer repeat plots for constructs with rigid loops showed that vast majority of the primer repeats that form secondary structures are found on the left side of the break, while flexible loop constructs have a more symmetrical primer repeat distribution (Figure 6A and Supplementary Figure S6A and B). We hypothesize that these trends are due to the inflexibility of the loop between the GGCC repeats, which prevents primer repeat annealing. These results are consistent with the observation that poly A sequences in the loops of stem–loop structures require more energy to close than other sequences (27).

One alternative explanation for these results could be that the reduction in SD-MMEJ events seen in rigid loop constructs is due to other differences in the sequence flanking the GGCC repeats. To explore this possibility, we created constructs identical to R0, but with flexible (TATAT) or rigid (AAAAA) sequences inserted between the GGCC repeat. We named these constructs R0 Flex and R0 Rigid. Overall SD-MMEJ consistency was much higher for the R0 Flex sequence, across all three types of junctions (Figure 6B and C). Primer repeat usage in SD-MMEJ consistent single-step insertions for R0 Flex was similar to the original R0 construct, with many junctions using the GGCC primer repeat (Figures 3 and 6D). In stark contrast, there were no SD-MMEJ consistent indel repair events in the R0 Rigid construct. From these results, we conclude that the repair

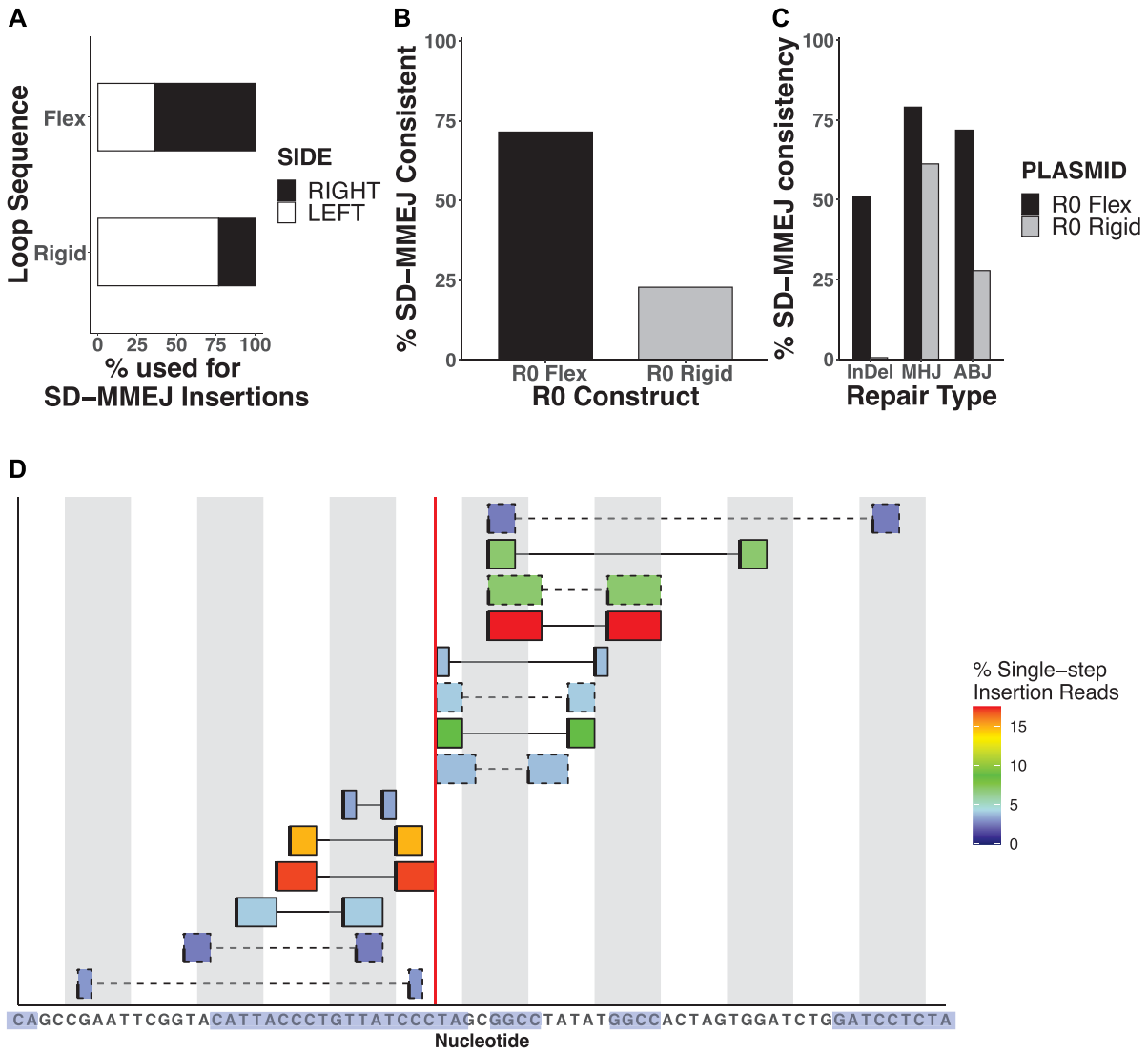


Figure 6. Loop flexibility plays an important role in facilitating SD-MMEJ repair. (**A**) Percentage of primer repeats used for SD-MMEJ consistent single-step insertions on either the right or left sides of the Cas9-induced double-strand break. 5 rigid loop and 8 flexible loop constructs were analyzed, with each construct contributing equal weight to the final values. All constructs in the randomized library that qualify as rigid (TTTTT on top strand between GGCC repeats) and flexible (TATAT or ATATA on top strand between GGCC repeats) were included. (**B**) SD-MMEJ consistency measurements for all inaccurate repair events recovered from R0 Flex and Rigid constructs. (**C**) SD-MMEJ consistency for R0 Flex and R0 Rigid repair products by type. Indel = insertion/deletion junctions; MHJ = microhomology joins; ABJ = apparent blunt joins. (**D**) Primer repeats used for SD-MMEJ consistent indel repair junctions in the R0 Flex construct. Loop-out repair = solid outline and connecting line using direct repeat; snap-back repair = dashed outline and connecting line using inverted repeats. Only the top strand of the construct is shown. Blue shading denotes nucleotides held constant in the semi-randomized design.

differences we observed in the rigid and flexible loop constructs were likely not due to other sequence differences.

A second alternative explanation for these results could be that DNA resection through the homopolymeric sequence is impaired. To test this, we created the T-loop construct, identical to the R0 Rigid construct but with a homopolymeric T sequence in the single-stranded loop (Supplementary Figure S6C). Primer repeat analysis of single-step insertion products recovered from Cas9 cleavage of the T-loop construct clearly showed a similar profile to the original R0 construct, with frequent usage of the GGCC primer repeat. Based on these results, we conclude that the most likely explanation for our findings is that a homopolymeric A sequence between a frequently used primer repeat strongly inhibits both secondary structure formation and SD-MMEJ repair.

SD-MMEJ efficiency is promoted by the presence of favorable microhomology templates

The final step of SD-MMEJ occurs when a newly synthesized microhomology anneals with single-stranded DNA on the other side of the break, followed by 3′ non-homologous tail clipping (if necessary), fill-in DNA syn-

thesis, and ligation (Figure 1). If microhomology annealing cannot occur, then any prior secondary structure formation and nascent DNA synthesis will not be represented in the recovered repair products. Conversely, if a frequently used primer repeat is directly adjacent to a sequence that can act as a 'microhomology template,' then this sequence context may be even more favorable for SD-MMEJ. Here, the term microhomology template refers to ssDNA adjacent to a primer repeat that can template nascent DNA synthesis, creating a new microhomology complementary to ssDNA across the break (Supplementary Figure S7).

To determine whether successful SD-MMEJ repair utilizes preferred microhomologies, we analyzed the microhomologous sequences used during the annealing step for SD-MMEJ consistent deletions recovered from all constructs. Strikingly, we observed that the most frequently used microhomologies correspond to those directly adjacent to the break site (Figure 7). While 19 of the 20 nucleotides at the break site are fully conserved in all 1100 constructs, 78.5% of all microhomologies used during the final annealing step correspond to CC, C, TA and T, which are immediately to the left and right of the break. Larger microhomologies of 4–5 base pairs, including several G–C rich sequences, are utilized at a much lower frequency (Figure 7). Thus, it appears that while the location of the microhomology template can vary depending on the preferred secondary structure forming sequences, there is a strong preference for these microhomology templates to anneal to complementary sequences located directly proximal to the other break end.

Predicting SD-MMEJ repair events through examination of the relative positions of primer repeats and microhomology templates

In simple MMEJ repair, pre-existing microhomologies frequently used for annealing are usually 2 bp or greater and are associated with high GC content (29). We would expect that newly synthesized microhomologies likely share these parameters. In addition, the amount of templated synthesis in alt-EJ is limited and estimated to be typically between 3–6 bp (9). This constraint predicts that a favorable microhomology template located directly adjacent to a strong primer repeat should promote a high frequency of SD-MMEJ consistent deletions (Supplementary Figure S7). In contrast, the lack of a favorable microhomology template directly adjacent to a strong primer repeat should result in more SD-MMEJ repair products with insertions, as DNA synthesis will need to continue until a more appropriate microhomology is synthesized. In our repair constructs, the 3′ end on the left side of the break is always CCC. Thus, we predict that primer repeats on the right side of the break directly adjacent to microhomology templates that produce G-rich ssDNA should be highly represented in the recovered repair products.

To test this hypothesis, we investigated DNA repair constructs where the break distal GGCC primer repeat is directly adjacent to 0, 2 and 5 bp microhomology templates that can promote the synthesis of ssDNA complementary to the unprocessed 3′ DNA on the left side of the break. Strikingly, we observed that the R0 repair construct with no adjacent microhomology template was associated with a high percentage of indels and few deletions, consistent with our prediction (Figure 8A). Further inspection showed that the R0 construct generated a high proportion of indels via synthesis into the break-proximal GGCC, generating G-rich ssDNA that anneals with the terminal CC on the other side of the break. In contrast, the R790 and R452 constructs, with 2 bp and 5 bp microhomology templates, respectively, generated a greater proportion of deletions/microhomology joins, also consistent with our expectation. SD-MMEJ insertion lengths in these constructs trend shorter as the length of microhomology templates increases (Figure 8B). Interestingly, as the microhomology templates increase in size, so does the proportion of simple MMEJ repair (Figure 8A). This suggests that longer microhomologies, whether present in the original sequence or created *de novo*, promote alt-EJ repair.

DISCUSSION

Alternative end joining was originally viewed as a backup double-strand break repair mechanism that operates in the absence of more dominant repair pathways such as nonhomologous end joining and homologous recombination. In recent years, this perception has changed, particularly with the recognition that alt-EJ repair frequently occurs even when other types of repair are possible (6,29–32). The molecular mechanisms that promote alt-EJ began to come into focus with the identification of DNA polymerase theta, which is critical for TMEJ (5,8,29,33–35). Alt-EJ repair and pol theta expression are concomitantly upregulated in many cancers (36,37) and both become important for cellular survival in the absence of homologous recombination or in situations where cells are subjected to high levels of DNA replication stress (35,38). To gain a more complete understanding of alt-EJ and TMEJ repair mechanisms, the contributions of DNA sequence context must also be considered. This is particularly important in the context of genome editing with double-strand break repair intermediates, as alt-EJ repair can promote both unwanted and desirable editing outcomes.

Using a plasmid-based repair assay in Drosophila embryos, we have now shown that SD-MMEJ, a type of alt-EJ, frequently occurs at Cas9-induced breaks and that pol theta is vital for this type of repair. While SD-MMEJ is a sequence-based model (10,12) and TMEJ is defined by its genetic requirement for pol theta (4,5), our analysis of repair events recovered from *polq* mutant embryos suggests that these two are, in most cases, the same.

Simple MMEJ repair, in which DNA resection at a double-strand break reveals pre-existing microhomologies that pair across the break, is a common form of TMEJ (39). However, we and others have noted that alt-EJ frequently produces insertional repair events that cannot be explained by a simple MMEJ model (10,14). In this study, we focused on mechanisms that promote these types of repair events. We chose to partially randomize the sequence within 30 bp of a single Cas9-induced break and developed a cloning strategy that can quickly generate thousands of unique sequence contexts for the study of SD-MMEJ. We also designed a high-throughput sequencing method and a novel, two-stage demultiplexing strategy using unique molecular

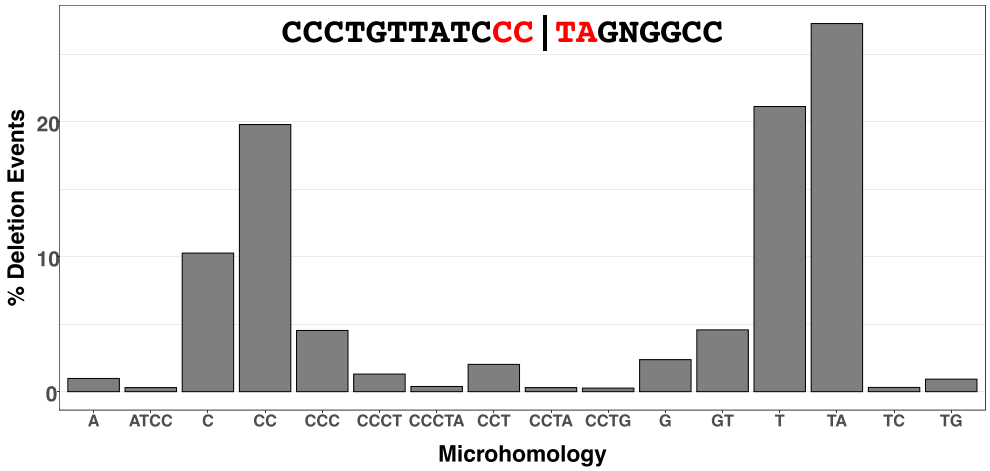


Figure 7. Microhomology usage in SD-MMEJ consistent deletions. Analysis of the microhomologies used during the final annealing step of SD-MMEJ was performed for all 1100 constructs. Each construct was weighted equally to prevent overrepresentation of constructs with more sequencing reads. Although 222 distinct microhomologies were used to create SD-MMEJ consistent deletions across all constructs in the library, only microhomologies used in at least 0.25% of all repair products are listed. Microhomologies correspond to bases in the top strand (shown at the top of the plot) for simplicity. The vertical line represents the Cas9 cut site.

identifiers to characterize large numbers of repair products. To analyze the resulting large-scale amplicon sequencing data, we developed a computational pipeline for streamlined analysis of SD-MMEJ consistency. These resources are freely available for use by the scientific community.

What parameters impact SD-MMEJ repair?

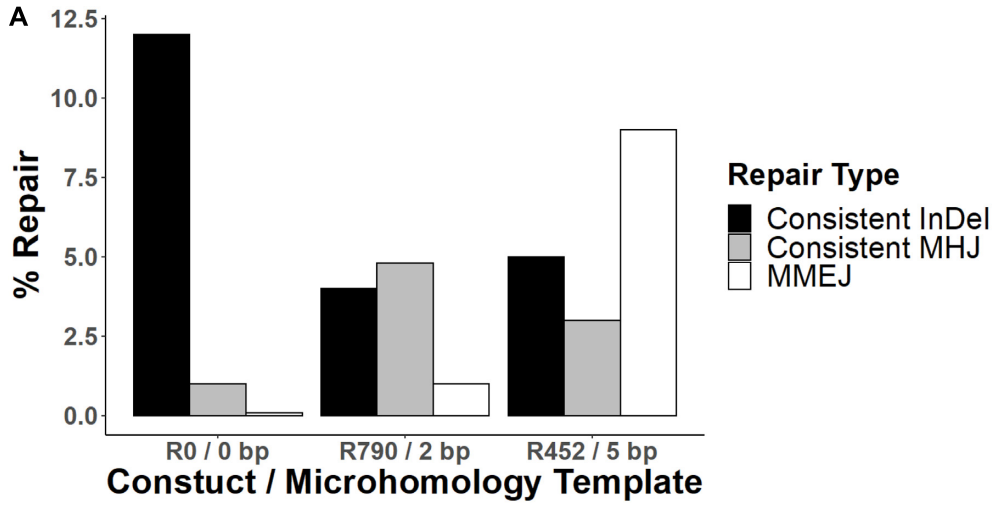
To gain insight into the characteristics of sequence contexts that promote SD-MMEJ, we fully analyzed 1100 semi-randomized constructs that were cut by Cas9 and repaired in Drosophila embryos. Many of the findings from this analysis are consistent with previous studies from our lab and others. For example, secondary structure-forming repeats are more frequently used when they are closer to the break site. Similarly, newly-synthesized microhomologies most frequently anneal to the 3′ terminus of single-stranded DNA on the opposite side of the break. These findings closely match the trends that have been observed with *in vitro* and *in vivo* studies of TMEJ (4,5,11).

Our analysis of different sequences contexts flanking a Cas9 break also provided insight into some novel aspects of SD-MMEJ that have not been previously appreciated. First, there seems to be an intricate balance between the length of the primer repeats that form secondary structures and the distance between them. Shorter (1–3 nt) primer repeats are more likely to be used when the inter-repeat distance is small. As the inter-repeat distance increases, a compensatory increase in repeat length is needed to maintain SD-MMEJ repair efficiency. However, primer repeats longer than 3 nt are not utilized for SD-MMEJ repair as often as would be expected based on their thermodynamic stability. Similarly, microhomologies longer than 3 bp are rarely used during the final annealing step. We propose that the geometry of the pol theta active site dictates the preferential formation of certain secondary structures, while the ability of pol theta to unwind transiently formed secondary structures (or their spontaneous dissociation rate) places an upper limit on the length of repeats that are used in SD-MMEJ. Furthermore, because the distributions of preferred primer repeat and microhomology lengths are similar, we speculate that annealing of primer repeats and microhomology repeats that occurs within the pol theta active site may be mechanistically similar.

Second, the flexibility/rigidity of the sequences between primer repeats greatly impacts SD-MMEJ repair. These sequences comprise the loops that are formed during secondary structure formation in both loop-out and snap-back SD-MMEJ. Homopolymeric A sequences, which are known to impair hairpin formation (28,40), drastically reduce the use of flanking primer repeats during SD-MMEJ, while flexible AT loops promote SD-MMEJ. We hypothesize that the thermodynamic forces that promote the formation of SD-MMEJ favorable secondary structures (27) are important factors in the repair process and influence the ability of pol theta to synthesize new microhomologies during SD-MMEJ.

Third, the presence of favorable microhomology templates adjacent to primer repeats strongly promotes SD-MMEJ. During our analysis of insertional repair events, we repeatedly observed that strong primer repeats predicted to be favorable for secondary structure formation were underrepresented in our primer repeat plots. These disfavored primer repeats were not adjacent to microhomology templates that could promote synthesis of ssDNA complementary to the other side of the break. Similarly, predicted weaker repeats of 1–2 bp separated by more than 10 bp were overrepresented in our plots. These repeats tended to be next to good microhomology templates. As a consequence, SD-MMEJ consistent simple deletion junctions are prevalent when microhomology templates are directly adjacent to primer repeats, while insertion junctions are favored when good microhomology templates are separated from primer repeats by one or more bp (Supplementary Figure S7).

Together, our observations strongly suggest that the efficiency of SD-MMEJ repair is largely driven by two factors:



Construct	Sequence (5′ - break - 3′)
R0	**GCCGAATTCGGTA**CATTACCCTGTTATCCC|TAG**C**GGCC**GCATA**GGCC**ACTAGTGGATCTG**
R790	**AAGATTAGGCCTG**CATTACCCTGTTATCCC|TAG**C**GGCC**TCA**CCGGCC**CGCGGAGCGAATC**
R452	**TCTTTTACAACTA**CATTACCCTGTTATCCC|TAG**A**GGCC**ATCCC**GGCC**TTGCCGCCTTATA**

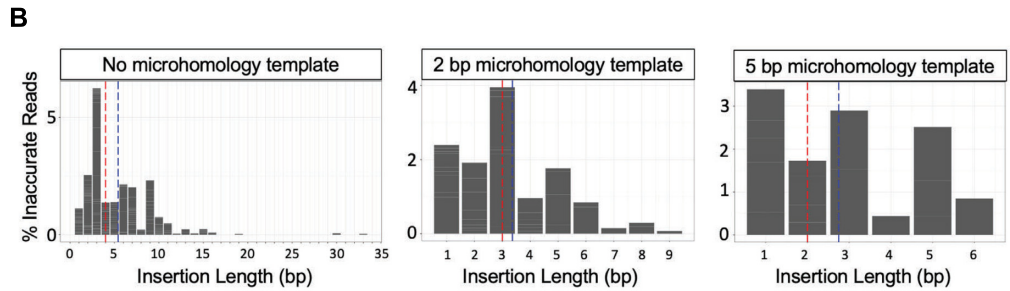


Figure 8. Presence of microhomology templates strongly influences repair outcomes. (**A**) Changes in the length of the microhomology template directly adjacent to the GGCC primer repeat affect its relative use in SD-MMEJ repair. Constructs R0, R790 and R452 possess microhomology templates of 0, 2 and 5 bp that are complementary to ssDNA located at the terminus of the other break end and can be used during loop-out SD-MMEJ repair. Simple MMEJ events not consistent with the SD-MMEJ model were calculated as a percent of all inaccurate reads. (**B**) Insertion size trends longer as the length of microhomology templates decreases. The insertion length distributions are shown for SD-MMEJ consistent repair events from constructs R0, R790 and R452 containing 0, 2 and 5 bp microhomology templates, respectively, for the 3′ terminal microhomology of the other side of the break. Red line = mean, blue line = median.

(i) a balance between secondary structure formation and dissociation following limited repair synthesis and (ii) the presence of microhomology templates directly adjacent to or near these secondary-structure forming repeats. We hypothesize that while many repair options are theoretically possible when a double-strand break is formed, these two principles most strongly influence which repair intermediates ultimately result in a successful repair outcome. Because most SD-MMEJ repair depends on pol theta, these principles may also provide information about the biochemical mechanisms of this enzyme and other proteins involved in alternative end joining.

Limitations and extensions of this study

When designing the semi-randomized constructs, we chose to keep the GGCC primer repeat constant while systematically varying other sequences flanking the break. This lack of total randomization allowed us to make more targeted interpretations of the results but does not represent the full diversity of sequences that could be encountered at any DSB. In addition, because putative loop-out SD-MMEJ events can also be explained via an *in trans* mechanism and some SD-MMEJ consistent repair events are also consistent with simple MMEJ or classical end joining repair, we cannot unambiguously assign mechanism to all the repair junctions.

While our SD-MMEJ modelling software can explain alt-EJ repair events that result in apparent blunt joins, simple deletions, and single-step insertion events, it is currently unable to model multi-step insertion events. We and others have previously recovered large alt-EJ insertion events that are consistent with multiple rounds of annealing, synthesis, and dissociation. Future iterations of the model and computational pipeline should accommodate these types of events.

Our parameters for SD-MMEJ consistency require that the secondary-structure forming primer repeats be located within 30 bp of the break ends, as SD-MMEJ consistent solutions could theoretically be found for all repair events if no search window is imposed (10,11). However, there may be certain contexts where secondary structures form more than 30 bp from the break. For example, depletion of replication protein A, a heterotrimeric ssDNA binding protein that binds 30 nt of DNA in its high affinity mode, might promote long distance SD-MMEJ. Indeed, simple MMEJ occurs more frequently when RPA binding is impaired. (41,42). We are currently testing the effects of RPA binding on SD-MMEJ.

In conclusion, our analysis of a large set of semi-randomized repair constructs has provided new insight into ways that sequences flanking a double-strand break can influence alt-EJ repair. The application of machine learning to our entire data set will yield further insight the mechanisms of alt-EJ, and specifically the sequence contexts that favor insertional repair. Pol theta is likely responsible for the majority of alt-EJ and cells become dependent upon it for double-strand break repair during replication stress (35,38). Therefore, computational modeling of TMEJ repair will be important for genome editing efforts in cells with high pol theta expression or cells subjected to replication stress during chemotherapeutic treatment.

The gold standard for any DSB repair prediction program will be the ability to predict all repair events and their relative frequencies at any DSB, for genome editing purposes. Several predictive programs with impressive accuracy currently exist, but they mostly address repair that occurs through classical end-joining mechanisms (16–19). Including model parameters for SD-MMEJ/TMEJ repair such as those elucidated here will increase the accuracy of predictive programs, particularly in genetic backgrounds where alt-EJ is more prevalent, such as those found in many cancers.

DATA AVAILABILITY

The computational pipeline and R script for analysis of SD-MMEJ consistency and plot generation is available at https://gitlab.tufts.edu/mcveylab/sdmmej/. All sequence reads for the 1100 repair constructs that we analyzed can be found at https://www.ncbi.nlm.nih.gov/sra/PRJNA706449.

SUPPLEMENTARY DATA

Supplementary Data are available at NAR Online.

ACKNOWLEDGEMENTS

We thank the Bloomington Drosophila Stock Center for fly stocks and Manan Krishnamurthy for generating the *polq* mutant stock. We also thank Varandt Khodaverdian for assistance with the bioinformatic analysis, Alice Miller for assistance with the injections, and Sergei Mirkin and members of the McVey lab for helpful discussions throughout the project.

FUNDING

National Science Foundation [MCB-1716039 to M.M.]; National Institutes of Health [R01-CA218112 to S.R.]. Funding for open access charge: Tufts internal account.
Conflict of interest statement. None declared.

REFERENCES

1. Chang,H.H.Y., Pannunzio,N.R., Adachi,N. and Lieber,M.R. (2017) Non-homologous DNA end joining and alternative pathways to double-strand break repair. *Nat. Rev. Mol. Cell. Biol.*, **18**, 495–506.
2. 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.
3. Deriano,L. and Roth,D.B. (2013) Modernizing the nonhomologous end-joining repertoire: alternative and classical NHEJ share the stage. *Annu. Rev. Genet.*, **47**, 433–455.
4. He,P. and Yang,W. (2018) Template and primer requirements for DNA Pol theta-mediated end joining. *Proc. Natl. Acad. Sci. U.S.A.*, **115**, 7747–7752.
5. 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.
6. Schimmel,J., Kool,H., van Schendel,R. and Tijsterman,M. (2017) Mutational signatures of non-homologous and polymerase theta-mediated end-joining in embryonic stem cells. *EMBO J.*, **36**, 3634–3649.
7. Zelensky,A.N., Schimmel,J., Kool,H., Kanaar,R. and Tijsterman,M. (2017) Inactivation of Pol theta and C-NHEJ eliminates off-target integration of exogenous DNA. *Nat. Commun.*, **8**, 66.
8. Chan,S.H., Yu,A.M. and McVey,M. (2010) Dual roles for DNA polymerase theta in alternative end-joining repair of double-strand breaks in Drosophila. *PLoS Genet.*, **6**, e1001005.
9. Carvajal-Garcia,J., Cho,J.E., Carvajal-Garcia,P., Feng,W., Wood,R.D., Sekelsky,J., Gupta,G.P., Roberts,S.A. and Ramsden,D.A. (2020) Mechanistic basis for microhomology identification and genome scarring by polymerase theta. *Proc. Natl. Acad. Sci. U.S.A.*, **117**, 8476–8485.
10. 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.
11. van Schendel,R., Roerink,S.F., Portegijs,V., van den Heuvel,S. and Tijsterman,M. (2015) Polymerase theta is a key driver of genome evolution and of CRISPR/Cas9-mediated mutagenesis. *Nat. Commun.*, **6**, 7394.
12. Khodaverdian,V.Y., Hanscom,T., Yu,A.M., Yu,T.L., Mak,V., Brown,A.J., Roberts,S.A. and McVey,M. (2017) Secondary structure forming sequences drive SD-MMEJ repair of DNA double-strand breaks. *Nucleic Acids Res.*, **45**, 12848–12861.
13. Zamborszky,J., Szikriszt,B., Gervai,J.Z., Pipek,O., Poti,A., Krzystanek,M., Ribli,D., Szalai-Gindl,J.M., Csabai,I., Szallasi,Z. *et al.* (2017) Loss of BRCA1 or BRCA2 markedly increases the rate of base substitution mutagenesis and has distinct effects on genomic deletions. *Oncogene*, **36**, 5085–5086.
14. Schimmel,J., van Schendel,R., den Dunnen,J.T. and Tijsterman,M. (2019) Templated insertions: a smoking gun for polymerase theta-mediated end joining. *Trends Genet.*, **35**, 632–644.
15. Michlits,G., Jude,J., Hinterndorfer,M., de Almeida,M., Vainorius,G., Hubmann,M., Neumann,T., Schleiffer,A., Burkard,T.R., Fellner,M. *et al.* (2020) Multilayered VBC score predicts sgRNAs that efficiently generate loss-of-function alleles. *Nat. Methods*, **17**, 708–716.
16. Shen,M.W., Arbab,M., Hsu,J.Y., Worstell,D., Culbertson,S.J., Krabbe,O., Cassa,C.A., Liu,D.R., Gifford,D.K. and Sherwood,R.I. (2018) Predictable and precise template-free CRISPR editing of pathogenic variants. *Nature*, **563**, 646–651.
17. Martinez-Galvez,G., Joshi,P., Friedberg,I., Manduca,A. and Ekker,S.C. (2021) Deploying MMEJ using MENdel in precision gene editing applications for gene therapy and functional genomics. *Nucleic Acids Res.*, **49**, 67–78.
18. Allen,F., Crepaldi,L., Alsinet,C., Strong,A.J., Kleshchevnikov,V., De Angeli,P., Palenikova,P., Khodak,A., Kiselev,V., Kosicki,M. *et al.*

(2018) Predicting the mutations generated by repair of Cas9-induced double-strand breaks. *Nat. Biotechnol.*, **37**, 64–72.
19. Leenay,R.T., Aghazadeh,A., Hiatt,J., Tse,D., Roth,T.L., Apathy,R., Shifrut,E., Hultquist,J.F., Krogan,N., Wu,Z. *et al.* (2019) Large dataset enables prediction of repair after CRISPR-Cas9 editing in primary t cells. *Nat. Biotechnol.*, **37**, 1034–1037.
20. Hanscom,T., Khodaverdian,V.Y. and McVey,M. (2018) Recovery of alternative end-joining repair products from drosophila embryos. *Methods Enzymol.*, **601**, 91–110.
21. Li,H. and Durbin,R. (2009) Fast and accurate short read alignment with burrows-wheeler transform. *Bioinformatics*, **25**, 1754–1760.
22. Brown,A.J., Al-Soodani,A.T., Saul,M., Her,S., Garcia,J.C., Ramsden,D.A., Her,C. and Roberts,S.A. (2018) High-Throughput analysis of DNA break-induced chromosome rearrangements by amplicon sequencing. *Methods Enzymol.*, **601**, 111–144.
23. Wilson,T.E., Grawunder,U. and Lieber,M.R. (1997) Yeast DNA ligase IV mediates non-homologous DNA end joining. *Nature*, **388**, 495–498.
24. Grawunder,U., Zimmer,D., Fugmann,S., Schwarz,K. and Lieber,M.R. (1998) DNA ligase IV is essential for V(D)J recombination and DNA double-strand break repair in human precursor lymphocytes. *Mol. Cell.*, **2**, 477–484.
25. Lemos,B.R., Kaplan,A.C., Bae,J.E., Ferrazzoli,A.E., Kuo,J., Anand,R.P., Waterman,D.P. and Haber,J.E. (2018) CRISPR/Cas9 cleavages in budding yeast reveal templated insertions and strand-specific insertion/deletion profiles. *Proc. Natl. Acad. Sci. U.S.A.*, **115**, E2040–E2047.
26. Geisinger,J.M., Turan,S., Hernandez,S., Spector,L.P. and Calos,M.P. (2016) In vivo blunt-end cloning through CRISPR/Cas9-facilitated non-homologous end-joining. *Nucleic Acids Res.*, **44**, e76.
27. Bonnet,G., Krichevsky,O. and Libchaber,A. (1998) Kinetics of conformational fluctuations in DNA hairpin-loops. *Proc. Natl. Acad. Sci. U.S.A.*, **95**, 8602–8606.
28. Nelson,H.C., Finch,J.T., Luisi,B.F. and Klug,A. (1987) The structure of an oligo(dA).oligo(dT) tract and its biological implications. *Nature*, **330**, 221–226.
29. Kent,T., Chandramouly,G., McDevitt,S.M., Ozdemir,A.Y. and Pomerantz,R.T. (2015) Mechanism of microhomology-mediated end-joining promoted by human DNA polymerase theta. *Nat. Struct. Mol. Biol.*, **22**, 230–237.
30. Truong,L.N., Li,Y., Shi,L.Z., Hwang,P.Y., He,J., Wang,H., Razavian,N., Berns,M.W. and Wu,X. (2013) Microhomology-mediated end joining and homologous recombination share the initial end resection step to repair DNA double-strand breaks in mammalian cells. *Proc. Natl. Acad. Sci. U.S.A.*, **110**, 7720–7725.
31. Villarreal,D.D., Lee,K., Deem,A., Shim,E.Y., Malkova,A. and Lee,S.E. (2012) Microhomology directs diverse DNA break repair pathways and chromosomal translocations. *PLoS Genet.*, **8**, e1003026.
32. Howard,S.M., Yanez,D.A. and Stark,J.M. (2015) DNA damage response factors from diverse pathways, including DNA crosslink repair, mediate alternative end joining. *PLoS Genet.*, **11**, e1004943.
33. Beagan,K., Armstrong,R.L., Witsell,A., Roy,U., Renedo,N., Baker,A.E., Scharer,O.D. and McVey,M. (2017) Drosophila DNA polymerase theta utilizes both helicase-like and polymerase domains during microhomology-mediated end joining and interstrand crosslink repair. *PLoS Genet.*, **13**, e1006813.
34. van Schendel,R., van Heteren,J., Welten,R. and Tijsterman,M. (2016) Genomic scars generated by polymerase theta reveal the versatile mechanism of alternative end-joining. *PLoS Genet.*, **12**, e1006368.
35. Ceccaldi,R., Liu,J.C., Amunugama,R., Hajdu,I., Primack,B., Petalcorin,M.I., O'Connor,K.W., Konstantinopoulos,P.A., Elledge,S.J., Boulton,S.J. *et al.* (2015) Homologous-recombination-deficient tumours are dependent on Poltheta-mediated repair. *Nature*, **518**, 258–262.
36. Higgins,G.S., Harris,A.L., Prevo,R., Helleday,T., McKenna,W.G. and Buffa,F.M. (2010) Overexpression of POLQ confers a poor prognosis in early breast cancer patients. *Oncotarget*, **1**, 175–184.
37. Lemee,F., Bergoglio,V., Fernandez-Vidal,A., Machado-Silva,A., Pillaire,M.J., Bieth,A., Gentil,C., Baker,L., Martin,A.L., Leduc,C. *et al.* (2010) DNA polymerase theta up-regulation is associated with poor survival in breast cancer, perturbs DNA replication, and promotes genetic instability. *Proc. Natl. Acad. Sci. U.S.A.*, **107**, 13390–13395.
38. Zhou,J., Gelot,C., Pantelidou,C., Li,A., Yücel,H., Davis,R.E., Farkkila,A., Kochupurakkal,B., Syed,A., Shapiro,G.I. *et al.* (2021) A first-in-class Polymerase Theta Inhibitor selectively targets Homologous-Recombination-Deficient Tumors. *Nat. Cancer*, **2**, 598–610.
39. Black,S.J., Ozdemir,A.Y., Kashkina,E., Kent,T., Rusanov,T., Ristic,D., Shin,Y., Suma,A., Hoang,T., Chandramouly,G. *et al.* (2019) Molecular basis of microhomology-mediated end-joining by purified full-length poltheta. *Nat. Commun.*, **10**, 4423.
40. Varani,G. (1995) Exceptionally stable nucleic acid hairpins. *Annu. Rev. Biophys. Biomol. Struct.*, **24**, 379–404.
41. Kim,C., Paulus,B.F. and Wold,M.S. (1994) Interactions of human replication protein a with oligonucleotides. *Biochemistry*, **33**, 14197–14206.
42. Yates,L.A., Aramayo,R.J., Pokhrel,N., Caldwell,C.C., Kaplan,J.A., Perera,R.L., Spies,M., Antony,E. and Zhang,X. (2018) A structural and dynamic model for the assembly of replication protein a on single-stranded DNA. *Nat. Commun.*, **9**, 5447.