Published online 19 July 2018

Nucleic Acids Research, 2018, Vol. 46, No. 16 ***8417–8434***
doi: 10.1093/nar/gky653

Decoding non-random mutational signatures at Cas9 targeted sites

Amir Taheri-Ghahfarokhi[1,*], Benjamin J.M. Taylor[2], Roberto Nitsch[1,3], Anders Lundin[1], Anna-Lina Cavallo[1], Katja Madeyski-Bengtson[1], Fredrik Karlsson[4], Maryam Clausen[1], Ryan Hicks[1], Lorenz M. Mayr[1,5], Mohammad Bohlooly-Y[1] and Marcello Maresca[1]

[1]Translational Genomics, Discovery Sciences, IMED Biotech Unit, AstraZeneca, Gothenburg, Sweden, [2]Discovery Biology, Discovery Sciences, IMED Biotech Unit, AstraZeneca, Cambridge, UK, [3]Advanced Medicines Safety, Drug Safety and Metabolism, IMED Biotech Unit, AstraZeneca, Gothenburg, Sweden, [4]Quantitative Biology, Discovery Sciences, IMED Biotech Unit, AstraZeneca, Gothenburg, Sweden and [5]GE Healthcare Life Sciences, The Grove Centre, White Lion Road, Amersham HP7 9LL, UK

Received December 28, 2017; Revised July 05, 2018; Editorial Decision July 06, 2018; Accepted July 09, 2018

ABSTRACT

The mutation patterns at Cas9 targeted sites contain unique information regarding the nuclease activity and repair mechanisms in mammalian cells. However, analytical framework for extracting such information are lacking. Here, we present a novel computational platform called Rational InDel Meta-Analysis (RIMA) that enables an in-depth comprehensive analysis of Cas9-induced genetic alterations, especially InDels mutations. RIMA can be used to quantitate the contribution of classical microhomology-mediated end joining (c-MMEJ) pathway in the formation of mutations at Cas9 target sites. We used RIMA to compare mutational signatures at 15 independent Cas9 target sites in human A549 wildtype and A549-POLQ knockout cells to elucidate the role of DNA polymerase θ in c-MMEJ. Moreover, the single nucleotide insertions at the Cas9 target sites represent duplications of preceding nucleotides, suggesting that the flexibility of the Cas9 nuclease domains results in both blunt- and staggered-end cuts. Thymine at the fourth nucleotide before protospacer adjacent motif (PAM) results in a two-fold higher occurrence of single nucleotide InDels compared to guanine at the same position. This study provides a novel approach for the characterization of the Cas9 nucleases with improved accuracy in predicting genome editing outcomes and a potential strategy for homology-independent targeted genomic integration.

INTRODUCTION

The clustered regularly interspaced short palindromic repeats (CRISPR)–Cas9 technology has been widely adopted as a precise genome editing tool in numerous areas of molecular biology that require modification of a specific DNA sequence. Originally harnessed from the bacterial adaptive immune system (1–3), the Cas9 nuclease can be programmed to induce double-strand breaks (DSBs) at a targeted locus in the genome of eukaryotic cells via a single guide RNA (sgRNA) (4–7). In principle, any region of DNA upstream of a 'protospacer adjacent motif' (PAM) can be targeted by the CRISPR–Cas9 system. Cells subjected to genetic DSBs recruit the endogenous DNA repair machinery to fix detrimental DNA damage. Understanding the behaviour of the Cas9 nuclease in living cells and its interplay with the DNA repair machinery could allow scientists to hijack repair pathways and achieve precise and predictable genome editing. Although extensive mechanistic studies have performed crystallography (8–11), *in vitro* (4,12) and *in silico* (13) functional characterization of Cas9, the cleavage activity of Cas9 in living cells remains elusive.

Various factors determine and modulate the types of mutations that occur at genomic loci after targeted cleavage. These factors include cell-dependent preferential repair pathways (14,15), the endonuclease used (16), chemical (17,18) or genetic modulation of DNA repair enzymes (19), cell cycle (20,21), the level and duration of nuclease activity in the target cell, the genomic context of the targeted site (18) and the nucleotides surrounding the DSB (22). However, the mutation signatures following a Cas9 cut at a target site are non-random and consistent under the same experimental conditions (18). DNA repair pathways at the site of DSBs can be categorized into four major pathways (23,24): classical non-homologous end joining (c-NHEJ, NHEJ), alternative end joining (alt-EJ), single-strand an-

*To whom correspondence should be addressed. Tel: +46 725318420; Email: Amir.Taheri-Ghahfarokhi@astrazeneca.com

nealing (SSA) and homologous recombination (HR) (Figure 1A). The HR pathway requires a template with long homology arms (>100 bp), generally provided endogenously by the sister chromatid or exogenously by a linear or circular DNA, and results in a high-fidelity repair events. The SSA pathway typically requires smaller length of homology arms (>20 bp) and results in deletions and translocations. The alt-EJ pathway consists of sub-pathways that all are intrinsically error-prone, frequently resulting in insertions and/or deletions (InDels) events. All alt-EJ sub-pathways require 2–20 bp of microhomology at the break site to promote the repair (23). The NHEJ pathway consists of sub-pathways that recognize broken DNA molecules and promote their direct re-ligation with no requirement for homology. Although the NHEJ pathway is considered to be the predominant repair pathway in mammalian cells (subjected to Cas9-induced DSBs), its precision remains controversial (25–27). NHEJ is distinct from other repair pathways as the associated protein 53BP1, protects the broken ends and inhibit extensive end resection required for alt-EJ, SSA and HR (23). Therefore, it is widely assumed that small genetic alterations are mainly driven by NHEJ. The flexibility and range of NHEJ enzymes (nuclease, polymerase and ligase) allows NHEJ to act on various substrates (e.g., blunt ends, incompatible 5′ or 3′ends, etc.) resulting in different mutational signatures including precise repair as well as small InDels (23). Key protein components of the NHEJ pathway includes 53BP1, Ku70-Ku80, DNA LigIV, DNA-PK, XLF and XRCC4 (23).

In contrast to NHEJ, the category of alt-EJ repair pathway is not yet well understood and has been described with a variety of terms (28). Early studies in yeast and mammalian cells only detected alt-EJ functioning in NHEJ deficient cells (27,29,30), leading to the assumption that alt-EJ has little contribution to a cell's overall repair capacity and operates only when NHEJ fails (31). However, several lines of evidence indicate that alt-EJ has a greater physiological role and coexist with NHEJ in normal and cancer cells (24,32). Biochemical studies suggested distinct kinetics of DSB repair processes: NHEJ is fast ($t_{1/2}$: 5–30 min) while alt-EJ (referred to as backup-NHEJ in the original studies) is slow ($t_{1/2}$: 2–20 h) (30,33). The slower kinetics of alt-EJ increases the chance of concomitant DSBs and therefore a higher chance of translocations (28,34). Multiple proteins have been associated with alt-EJ, including MRE11 (35,36), NBS1, LIG3, XRCC1, FEN1, PARP1 (36,37), and POLQ (polymerase theta; Polθ) (38–45) with several mechanistic models and protein requirements proposed for different types of alt-EJ repair processes (22,24,46,47). The mutational signatures associated with alt-EJ repair suggest the existence of different sub-pathways, however, characterizing alt-EJ pathways by mutational footprints is challenging because the DNA repair outcomes of different pathways can be the same (Figure 1A). In our study, we used the term 'classical Microhomology Mediated End Joining' (c-MMEJ) to specifically refer to deletions that are attributed to microhomology sequences of at least two nucleotides. This definition excludes other alt-EJ sub-pathways like the SD-MMEJ model (22), in which microhomologies are created via limited DNA synthesis at secondary-structure forming sequences.

The mutation patterns at the site of nuclease-mediated DSBs have been subjected to bioinformatic analysis to delineate the underlying repair mechanisms. Several studies have highlighted the importance of considering the target site and its flanking sequence context when choosing DSB sites for genome editing (15,18,22). For example, Bae *et al*. have reported that 52.7% of all deletions induced by Cas9 in human cells are associated with microhomologies flanking the site of DSB (15), and on the basis of their observations, they developed a computer program to predict the microhomology-associated deletion patterns at a given target site. Van Overbeek *et al*. studied the mutation patterns of 96 sgRNAs in HEK293, HCT116, and K562 cells at different time points after transfection and revealed that genome-editing outcome are non-random and change over time (18). Recently, a suite of computational tools (CRISPResso (48)), and a JavaScript-based instant assessment tool (Cas-analyzer) were developed for the deconvolution of mixed NHEJ-HDR (Homology-Directed Repair) outcomes using deep targeted NGS data (49). However, NHEJ in CRISPResso and Cas-analyzer software refers to all non-HDR events including NHEJ and alt-EJ repair events. To the best of our knowledge, there is no bioinformatic tool that discriminates the mutation signatures accountable to c-MMEJ or other non c-MMEJ repair pathways (referred to other-EJ in our study). Here, we present a computational tool called Rational InDel Meta-Analysis (RIMA) that can analyse, categorize, and visualize, deep targeted NGS data from Cas9-induced mutation studies. RIMA collects the variants identified in NGS data from Cas9-targeted loci and generates a detailed graphical report on their mutation pattern. This software can be used to discriminate among DNA repair pathways associated with specific mutations and categorize the insertions and deletions based on their size, type, and location relative to the Cas9 target site. Several datasets from the literature (15,16,18,19) were used to validate RIMA, and here, we show that RIMA can particularly discriminate the c-MMEJ associated deletions from other type of mutations. In addition, we elucidated the role of Polθ in the formation of c-MMEJ associated deletions.

MATERIALS AND METHODS

Plasmid constructs

The plasmids expressing SpCas9-wt, SpCas9-HF1 and FnCas9 were constructed via the synthesis of codon optimized Cas9 sequences fused to nuclear localization signal (NLS) and a self-cleaving enhanced green fluorescent protein (2A-EGFP) cassette under the control of the CMV promoter. The protospacers were cloned into sgRNA expressing plasmids downstream of the human U6 promoter using standard molecular cloning of oligonucleotide duplexes. Previously reported sgRNA scaffold (50) sequences were used to generate the sgRNA expressing plasmids. The selected human gene target sites, primers used to amplify their flanking regions and the expected cut site in the amplicons are listed in Supplementary Table S3. The plasmid expressing TREX2 was purchased from GenScript. The plasmid expressing DNTT was generated by synthesising the DNTT

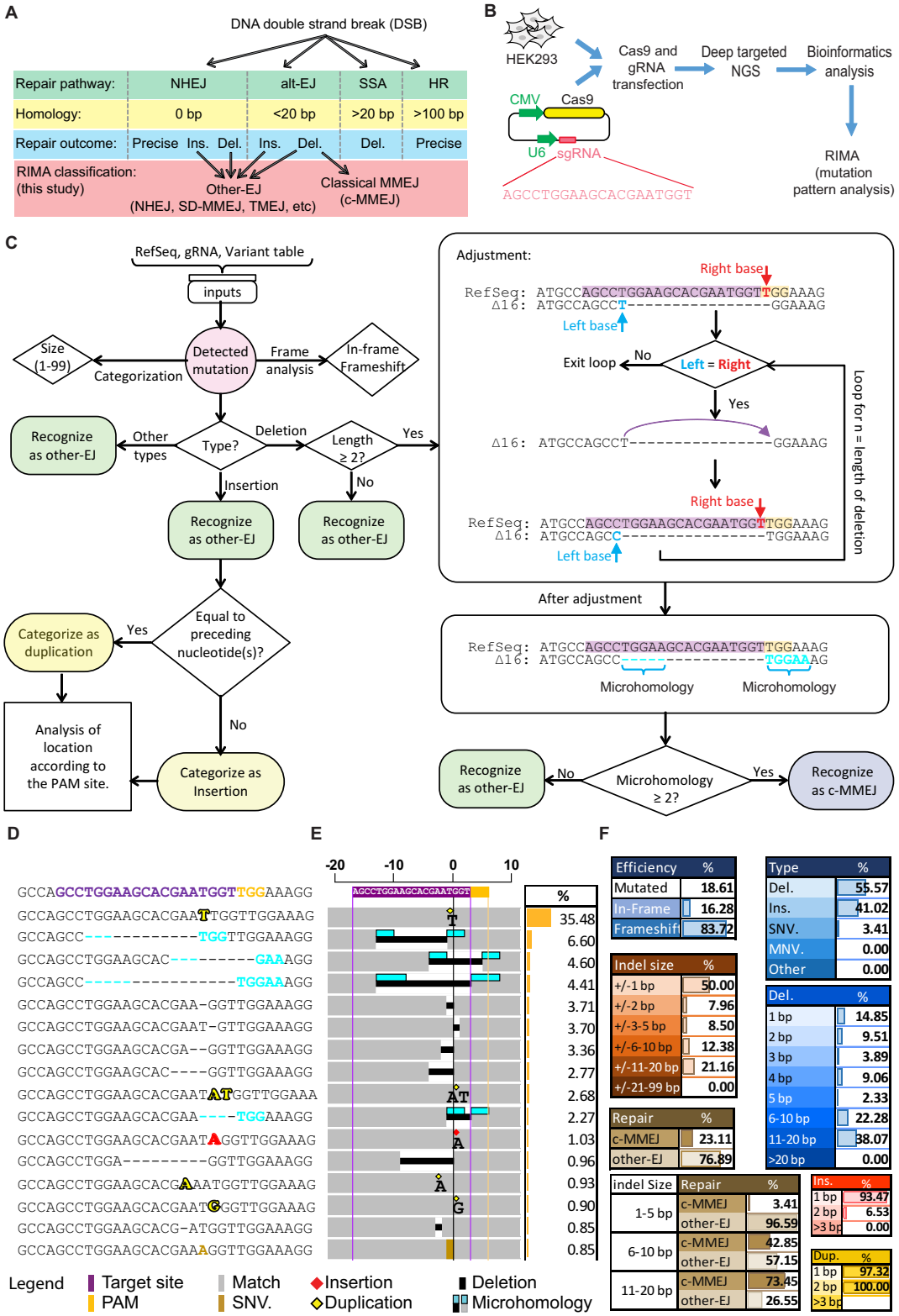

Figure 1. Analysis of DNA repair profiles following Cas9 cleavage of genomic sites in mammalian cells. (**A**) Overview of double strand break (DSBs) repair pathways; Non Homologous End Joining (NHEJ), alternative End Joining (alt-EJ), Single Strand Annealing (SSA), and Homologous Recombination (HR), and their InDel footprints in mammalian cells. (**B**) Schematic of experimental procedures used to detect and analyse the InDel footprints after Cas9 cuts a genomic locus. (**C**) Flowchart and overview of the algorithm used in RIMA to adjust and classify the mutations. sgRNAs and PAM locations are highlighted on the reference sequence (RefSeq) in purple and yellow respectively. Complex indels (e.g. multiple nucleotide variations and replacements) were categorized as 'Other type'. (**D**) RIMA generates a colour-coded alignment of the mutations detected in the NGS data. The wild-type sequence is shown on the top. (**E**) A graphical representation of the alignment shown in (D) generated using RIMA. The orientation of PAM and the sgRNA are

coding sequence (GeneArt, Life Technologies) and cloning into the pcDNA3.1(+) plasmid under the CMV promoter.

Cell culture and transfection

The human embryonic kidney 293 (HEK293), A549 and HCT116 cell lines were maintained in Dulbecco's Modified Eagle's Medium (DMEM) + GlutaMax (Life Technologies) supplemented with 10% FBS and Penicillin-Streptomycin at 37°C and 5% CO_2. All cell culture reagents were obtained from Life Technologies. The cell line identity was validated by STR profiling, and the cells were tested before and after the experiments for mycoplasma contamination. The cells were seeded one day prior to transfection (200K cells per well in 12-well plates or 30K cells per well in 96-well plates). The cells were transfected using FuGENE HD transfection reagents (Promega) at 70–80% confluency following the manufacturer's recommended protocol. Unless otherwise indicated, the cells were seeded in 12-well plates and co-transfected with a total amount of 1100 ng plasmid DNA (for dual sgRNA transfections, 600 ng of Cas9 plasmid and 250 ng of each sgRNA plasmid; for single sgRNA transfections, 800 ng of Cas9 plasmid and 300 ng of sgRNA plasmid; for single sgRNA transfections together with TREX2 or DNTT, 500 ng of Cas9 plasmid, 200 ng of sgRNA plasmid and 400 ng of TREX2 or DNTT plasmid). For the negative control experiments, the Cas9 plasmids were co-transfected with a scramble sgRNA plasmid or an empty vector.

Mirin treatments

The MRN inhibitor Mirin (Sigma, Cat. No. M9948-5MG) was prepared as a 40 mM stock solution dissolved in DMSO and stored at −80°C. Mirin was added to the cells at four concentrations (final concentration: 5, 10, 20 and 40 μM) one hour before the transfection to determine dose-specific responses. The cell lysates were collected 72 h after the transfection.

POLQ experiments

In order to generate POLQ knockout cells, human A549 cells stably expressing Cas9, were transfected with a sgRNAs (5′-CTGACTCCAAAAGCGGTACA-3′) targeting the POLQ gene. Transfected cells were then dissociated and diluted in full media at a single cell suspension level, and plated into 96-well plates. Single cells were expanded and screened for successful gene targeting. The genotype of the POLQ knockout cells was confirmed using Sanger sequencing and TIDE analysis (51). Cells were seeded in 96-well format and were transfected with sgRNAs targeting different endogenous loci. Genomic DNA harvested at three time-points, 12, 24 and 60 h.

Genomic DNA extraction

The cells were lysed 72 h after the transfection to harvest the genomic DNA. For cells cultured in 12-well plates, genomic DNA was extracted using a PureGene Gentra DNA extraction kit (QIAGEN) according to the manufacturer's recommended instructions. For cells cultured in 96-well plates, the genomic lysate was used as the PCR template. Briefly, the media were gently aspirated from the culture plates, 50 μl of EpiBio QuickExtraction DNA extraction solution (Epicenter, QE09050) were added to each well, and the plate was incubated at 37°C for 5 min, following by a 15-min incubation at 65°C and a 10-min incubation at 95°C. The lysates were transferred to PCR plates and stored at −20°C.

Topo cloning and Sanger sequencing

We performed Sanger sequencing to quantify and visualize the mutation patterns of genes targeted by dual sgRNAs. The genomic regions flanking the target site were amplified using the primers listed in Supplementary Table S3. All primers used in this study were designed using Primer-Blast software (52). The amplified fragments were cloned into the pCR2.1-TOPO-TA vector (Thermo Fisher). Single *Escherichia coli* colonies were placed in 200 μl of Terrific broth (TB) growth medium and grown for 20 h. Direct Sanger sequencing of the culture was performed using BigDye Terminator v3.1 sequencing reagents on an Applied Biosystems 3730xl DNA Analyzer. Briefly, 50 μl of the culture were transferred to a 96-well PCR plate and centrifuged for 5 min at 4000 × g to pellet the cells. The supernatant was removed, and the cells were re-suspended in 20 μl of double distilled water. Then, the cells were incubated on dry ice for two minutes, followed by 10 min at 95°C. The cell debris was pelleted by centrifugation for 5 min at the highest speed, and 2 μl of the supernatant were used as the template in a 20 μl sequencing reaction. The sequencing reads were analysed and aligned to the reference sequences using CLC Main Workbench software (version 7.6.2).

Deep targeted amplicon sequencing

To quantify and capture the genetic modifications at genomic loci targeted by a single sgRNA, PCR primers were designed to amplify the genomic DNA surrounding the target site. The PCR primers were linked to the sequences of the Illumina Nextera adapters (listed in Supplementary Table S3). In the first PCR analysis, 50–100 ng of the genomic template were used in a 20 μl reaction with FusionFlash High Fidelity Master Mix (Thermo) and a 500 nM final concentration of each forward and reverse primer. The amplified PCR products were subjected to paired-end sequencing using NextSeq500. The PCR products were amplified using Illumina NextEra XT Index Kit v2 adapters. The libraries were quantified using Qubit HS (Thermo Fisher)

←

shown beneath the scale bar. The length of all deletions is represented by the scale bar on the top. The deletions associated with microhomologies are visualized according to the bars shown in the legend. For the single and double nucleotide insertions or duplications, the corresponding nucleotides are shown under the symbol indicating their position. Only the length of insertions and duplications longer than two nucleotides are indicated. The vertical black line indicates the cut site. (**F**) Classification of the InDels was based on their attributes. The frequency of each class was calculated as a fraction of the mutant reads or a fraction of their parental category. For example, the frequency of the single nucleotide duplications is calculated as the fraction of the single nucleotide insertions equal to the preceding nucleotides.

and Fragment Analyzer (Advanced Analytical Technologies). The indexed libraries were pooled and sequenced with an Illumina NextSeq500 mid-output run using paired-end chemistry with a 150-bp read length.

Collection of data from the literature

In addition to the deep sequencing data generated in this study, we downloaded previously reported NGS data that were relevant to our study from the National Centre for Biotechnology Information (NCBI)-Sequence Read Archive (SRA; https://www.ncbi.nlm.nih.gov/sra/) using the command-line 'prefetch' tool available in the SRA toolkit (version 2.8.1). The previous publications used and SRA accession numbers are listed in Supplementary Table S1, and as appropriate, the corresponding studies from which the NGS data are reanalysed are cited throughout the text and figure legends. In addition, 240 insertion-type mutant sequences at 40 and 12 target sites of ZFNs and TALENs, respectively, were collected from 27 previously reported studies (Supplementary Table S2).

NGS analysis workflow

The SRA data were converted to FASTQ format using the 'fastq-dump' tool available in the SRA toolkit (version 2.8.1). The data from biological replicates were pooled before the analysis using a Windows command prompt program. For example, the command 'copy /b SRR3702339.fastq + SRR3702340.fastq Pool-SRR3702339-40.fastq' was used to pool the SRR3702339 and SRR3702340 samples into the file 'Pool-SRR3702339-40'. CLC Genomics Workbench software (version 9.5, QIAGEN) was used to analyse the FASTQ files (Supplementary Figure S1). Briefly, the data were trimmed based on their quality and mapped to the reference sequence. The mapped reads were locally re-aligned and then subjected to basic variant calling. Variant tables were exported from the CLC Genomics Workbench software into Excel 2010 files. The reference sequences used in this study are provided in Supplementary Table S1. Finally, the variants were analysed using RIMA.

Rational InDel mathematical analysis (RIMA)

The Visual Basic programming for Applications (VBA) in Microsoft Excel 2013 was used to manipulate, analyse, and visualize the mutation patterns. All of the code is freely available in the Supplementary Macro-enabled Excel File (RIMA.xlsm). Initially, the InDels were categorized into five types, including insertions, deletions, single nucleotide variations, multiple nucleotide variations and other (e.g. replacements). Furthermore, deletions with sizes of at least two nucleotides were sub-categorized into other-EJ- and c-MMEJ-associated deletions (Supplementary Note 1). For the identified c-MMEJ-associated deletions, the length of the microhomologies was calculated and used for illustration purposes. The insertions were also examined to determine whether they were duplications. The out-of-frame mutants were recognized based on the length of the InDels. In contrast, an InDel was considered in-frame mutation only if the length of the InDel was evenly divisible by three. All calculations of the InDel types and classes are presented as either a percentage of the total number of mutated reads or a percentage of the parental category. For example, the percentage calculated for single nucleotide duplications represents the proportion of single nucleotide insertions that are indeed duplications. To avoid variations caused by sequencing and amplification errors, we omitted single nucleotide variations (SNVs) from our analysis (except data shown in Figure 1, which SNVs are shown to note the RIMA's graphical output). Only variations within ~30 bp of the predicted cut site and outside the primer regions were considered Cas9-induced mutations to avoid false positive InDels. To enable the automatic analysis of thousands of NGS runs, a batch mode in RIMA was developed. Details and instructions regarding the operation of RIMA are available in the Supplementary text.

Competitive oligo-duplex incorporation assay

The barcoded blunt-ended and staggered-ended oligo-duplexes were prepared by annealing the two oligonucleotides listed in Supplementary Table S4. Each annealing reaction consisted of 5 μl of each 100 μM oligos (purchased from Sigma in the liquid form), 0.2 μl of $MgCl_2$ (0.15 μM) and water up to 50 μl. The following cycling conditions were used to hybridize the oligonucleotides: 95°C, 10 min; 85°C, 1 min; 75°C, 1 min; 65°C, 1 min; 55°C, 1 min; 45°C, 1 min; 35°C, 1 min; and 25°C, 1 min (the temperature ramping between each step was set at −1°C/s), followed by holding at 4°C. The oligo-duplexes were then pooled and transfected into the cells. HEK293 and HCT116 cells were seeded at 30K per well in 96-well plates one day before the transfection. The cells were transfected using FuGENE HD transfection reagents according to the manufacturer's instruction. Briefly, for the transfection of six wells of the 96-well plates, 600 ng of the Cas9 expressing plasmids, 200 ng of the sgRNA expressing plasmid and 300 ng of the oligo-duplexes were mixed in 55 μl of OptiMEM (Life Technologies), followed by the addition of 3.3 μl of FuGene 6 HD transfection reagents (Promega). After pipetting 15 times and a 5-min incubation at room temperature, 8 μl of the transfection reagents were added to each well of the 96-well plates. Three days after the transfection, the cells were lysed using EpiBio QuickExtraction DNA extraction solution, and 2 μl of the cell lysates were used as the template to amplify the junctions of the incorporated oligo-duplexes. The amplified fragments were then subjected to amplicon sequencing. The sequencing reads were automatically demultiplexed using a NextSeq500 Instrument (Illumina), and the paired FASTQ files were analysed using CRISPResso (48). Briefly, the reads with a minimum average quality score of 33 were aligned to the reference sequence. The frequency of the barcodes in each run was calculated based on the alleles detected by CRISPResso, 'Alleles_frequency_table.txt', in the NGS data as follows. First, all detected alleles were imported into Microsoft Excel. Then, the number of reads for alleles with eight identical nucleotide barcodes were consolidated. Finally, the relative frequency of each barcode was calculated. The data were converted to Standard Scores ('Z-scores') using the formula $Z = (x - \mu)/\sigma$, where x is

the value to be standardized (average of relative frequencies of three biological replicates), and μ and σ are the mean and standard deviation, respectively, of all seven target sites tested in this experiment (Supplementary Table S5). Standardized data were used to generate heatmaps in Microsoft Excel.

Statistical analysis

No statistical methods were used to predetermine the sample size. All calculations were performed using JMP software (version 13, SAS, Inc.). All results are expressed as the mean ± S.E.M. unless otherwise stated. In the box plot graphs, the central rectangle spans from the first quartile to the third quartile. The segment in the rectangle indicates the median. The outliers are shown outside the whiskers. The percentage data were transformed before performing the statistical test using the formula $ArcSine\left(\sqrt[2]{\frac{(x+0.5)}{100}}\right)$, where x is the value to be transformed. Dunnett's method was performed to compare the means between the control and treatment groups (in the Mirin experiment). Student's t-test was performed to calculate the statistical significance of the results, and a two-tailed $P<0.05$ was considered significant unless otherwise stated. The P-values are shown in figures as asterisks as follows: $^{**}P < 0.01$ and $^{***}P < 0.001$. The calculation of the Z-scores presented in Figure 6 is explained in the previous section and Supplementary Table S5.

Data and software availability

The output of 2738 NGS runs analysed using RIMA are listed in Supplementary Table S1. The macro-enabled version of the Microsoft Excel file (RIMA.xlsm) containing the VBA codes is supplemented and is freely available at github: https://github.com/Ghahfarokhi/RIMA.

RESULTS

Development of the RIMA algorithm

To test the algorithm used for the DNA repair signature detection and visualization, we transfected HEK293 cells with a plasmid expressing *Streptococcus pyogenes* Cas9 (SpCas9) and a sgRNA against the MAP3K1 gene (Figure 1B). We performed deep targeted NGS to detect the SpCas9-induced mutations at the targeted site from genomic DNA harvested 72 h after transfection. We then determined the genetic variants in the NGS data using an analysis workflow in the CLC Genomics Workbench version 9.5 software (Supplementary Figure S1). Subsequently, we designed a flowchart and algorithm for InDel analysis (Figure 1C) using Microsoft Excel (referred to as 'RIMA' in this paper). RIMA uses the reference sequence (RefSeq), sgRNA sequence and variant table as inputs (Supplementary Figure S2) and performs a variety of analyses and generates a graphical report for the visualization of mutations. RIMA categorizes the InDels based on their size (1–99 bp, customizable by the user), reading-frame impact (in-frame and out-of-frame), type (deletion, insertion, single nucleotide variation, multiple nucleotide variations, etc.), location relative to the PAM, and associated repair pathways (c-MMEJ and other-EJ). We used the term 'other-EJ' for collectively referring to the NHEJ, SD-MMEJ, and TMEJ events (Figure 1A). To discriminate genetic signatures associated with the c-MMEJ repair pathway, we developed a code that first arbitrarily adjusts all deletions towards the 5′ end of the reference sequence (Figure 1C, Supplementary Note 1) and then searches for possible microhomologies in the first nucleotides at the beginning and downstream of each re-aligned deletion (Figure 1C). This adjustment allows RIMA to analyse variant tables generated using workflows that do not left align mutations. Microhomology sequences of at least two nucleotides were classified as c-MMEJ-associated mutation and shown in the output mask (Figure 1D–F). This procedure allows for the automated identification of c-MMEJ-mediated DNA rearrangements and enables quantification of the ratio between other-EJ- and c-MMEJ-associated mutations based on the InDel frequencies detected in the deep targeted NGS data. Furthermore, we developed a 'batch mode' to automate the analysis of hundreds of runs. This allowed for analyses of over two thousand Cas9 cleavage sites using NGS data downloaded from the Sequence Read Archive (SRA) (15,16,18,19).

Validation of RIMA

To validate RIMA, we reanalysed previously reported deep targeted NGS datasets (15,18) from CRISPR–Cas9 experiments performed in four human cell lines (Figure 2A). The first dataset consisted of 67 sgRNAs tested in HeLa cells with the percentage of modified reads, frequency of out-of-frame InDels and 'Microhomology scores' available (15). The microhomology score (MHscore), which was developed by Bae *et al.*, is the sum of all theoretically possible mutation patterns that can be formed after repair by the c-MMEJ pathway at the surroundings of the DNA break (15). We used RIMA to measure the percentages of modified reads, the fraction of out-of-frame InDels, and the fraction of c-MMEJ-associated mutations in this dataset. We found a high Pearson's correlation ($r = 0.868$) between the mutagenesis rates reported by Bae *et al*. and those estimated in our study (Figure 2B). Similarly, the Pearson's correlation between the out-of-frame InDels calculated by Bae *et al*. and those calculated in our study was $r = 0.937$ (Figure 2C). However, the MHscores reported by Bae *et al*. and the frequency of c-MMEJ-associated mutations assessed using RIMA were weakly correlated (Pearson's correlation coefficient $r = 0.274$, P-value $= 0.024$, Figure 2D). The MHscore can be used to predict the out-of-frame scores permitting the selection of more potent sgRNAs for knockout experiments (15). However, our results indicated that the MHscore might not represent the frequency of c-MMEJ-associated mutations after Cas9 cuts.

The second dataset reanalysed with RIMA consisted of 96 sgRNAs tested in three human cell lines (HEK293, K562 and HCT116) with NGS amplicon analysis at 4, 8, 16, 24 or 48 h after transfection (18). HEK293 cells 48 h after transfection showed a uniform distribution for the number of detected repair events for each target site, ranging from

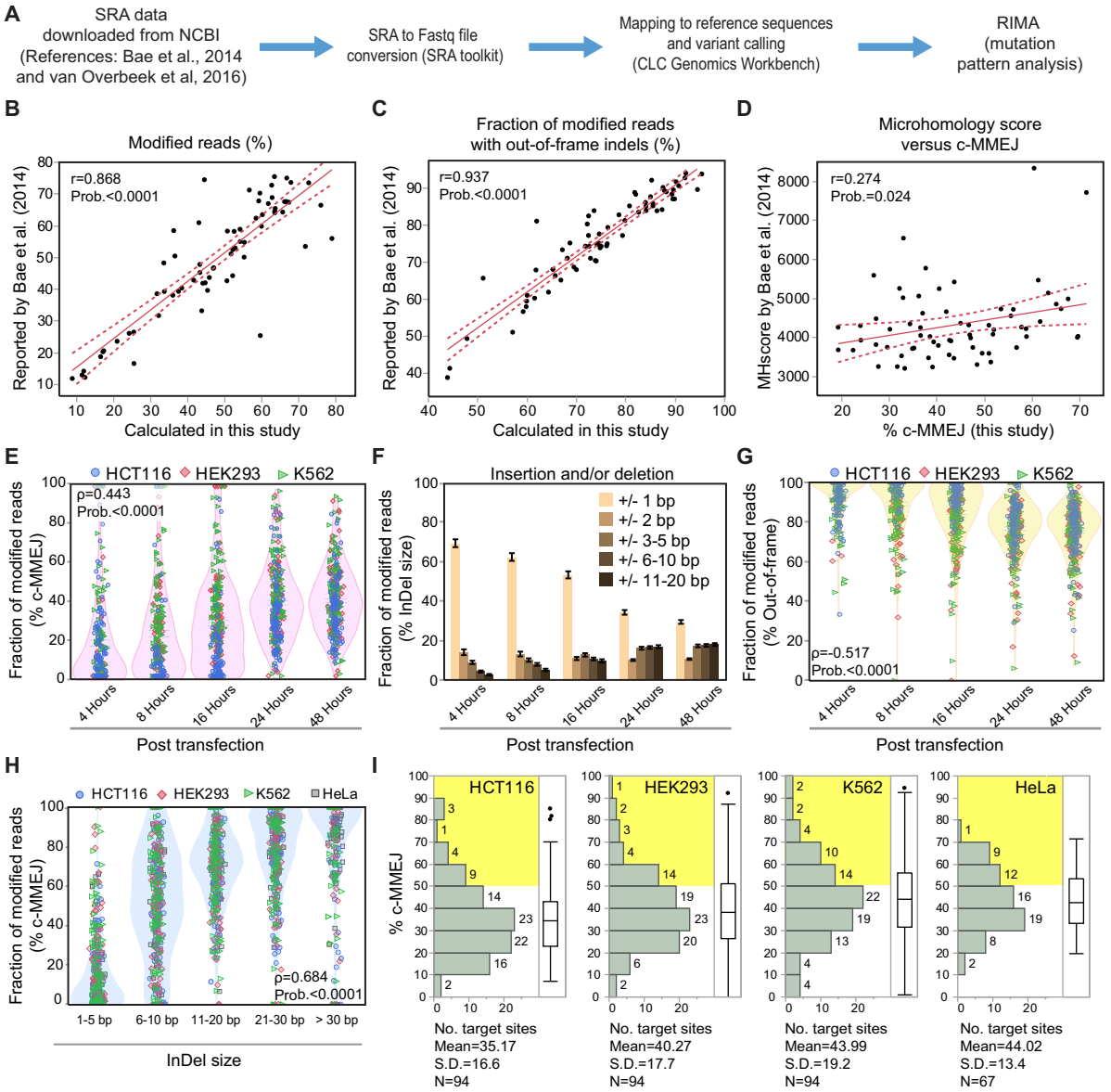

Figure 2. Validation of RIMA using publicly available datasets (15,18). (**A**) Workflow used in this study to download and reanalyse data from previous studies. (**B**) Percentages of modified reads and (**C**) fraction of modified reads with out-of-frame InDels calculated by Bae *et al.* (y-axis) plotted against RIMA calculations (x-axis). (**D**) Microhomology scores (MHscore) reported by Bae *et al.* (y-axis) compared to RIMA generated percentage c-MMEJ (x-axis). The correlation between datasets in B, C and D was calculated by linear regression (solid line) with 95% confidence intervals indicated (dashed line) and Pearson correlation coefficients (*r*) and *P*-values displayed. Time-dependent changes in the mutation patterns after Cas9 cleavage from van Overbeek *et al.* dataset were analysed for the (**E**) c-MMEJ/other-EJ ratio, (**F**) InDel size and frequency of in-frame and (**G**) out-of-frame mutations. (**H**) Comparison of other-EJ/c-MMEJ ratios in InDels with different lengths 48 h after transfection. Corresponding Spearman correlation coefficients (ρ) with *P*-values indicated within the graphs. All error bars indicate the standard errors of the mean (S.E.M.) for $N_{HEK293} = 94$, $N_{HCT116} = 94$, $N_{K562} = 94$. (**I**) Histogram plot of c-MMEJ ratio distribution in different cell lines. The number of target sites with dominant MMEJ repair is highlighted in yellow. Data shown were obtained 48 h after transfection (15,18).

10 to over 200 different mutations (Supplementary Figure S3a). In a total number of 9782 repair events detected in this analysis, deletions and insertions were the most frequently observed mutations (78.4% and 19.2%), while other mutation types, such as multiple nucleotide variants and replacements, were rarely observed (<2.5%, Supplementary Figure S3B). Single nucleotide deletions were the most frequent deletion length (∼15% of the deletions, Supplementary Figure S3C). 60.16% of deletions two nucleotides or larger were c-MMEJ-associated, with microhomology regions of mainly two to three nucleotides (Supplementary Figure S3D and E).

The frequency of c-MMEJ-associated deletions has been suggested to increase over time after Cas9 transfection (18). We analysed the full 1440 sample dataset using RIMA and contextualized the resulting InDel outcomes to either the c-MMEJ or the other-EJ pathways. We detected a significant Spearman's correlation ($\rho = 0.443$) between the inci-

dence of c-MMEJ and the duration of genomic exposure to Cas9 (Figure 2E). Furthermore, it was evident that over time the size of the deletions increased and the frequency of out-of-frame mutations decreased (Figure 2F and G, Supplementary Figure S4). To confirm that the extended DNA deletions were highly predictive of c-MMEJ-mediated repair, we plotted the c-MMEJ/other-EJ ratios against the InDel size and observed a positive Spearman's correlation ($\rho = 0.684$, Figure 2H). The dataset allowed us to determine whether cells preferentially use specific repair pathways following the same type of genomic perturbation. Performing RIMA analysis, we identified that c-MMEJ plays a major role in the repair of Cas9-induced DNA lesions in a considerable number of analysed target sites from all cell types analysed (Figure 2I).

Among the cell lines tested, HCT116 showed significantly lower overall occurrence of c-MMEJ repair ($P = 0.001$, Figure 2I). This observation could be explained by the presence of a mutant MRE11 allele in HCT116 cells (53); the mutant MRE11 retains the ability to bind DNA but has defective 3′-5′ exonuclease activity (53), which is a critical process for c-MMEJ (54). According to RIMA, whilst c-MMEJ-mediated repair in HCT116 cells are detected at a lower rate, it still represents a considerable fraction of the repair process, suggesting one of the main determinant of c-MMEJ repair is sequence context (Figure 2I, Supplementary Figure S5). The analysed dataset includes 22 multiple target single spacer (MTSS) sgRNAs that are each present in multiple copies throughout the human genome (18). This dataset uniquely allows evaluation of c-MMEJ repair after the cleavage of the same target sequence at different genomic locations and chromatin states. We found that for each sgRNA assessed, the rate of c-MMEJ was similar for each target site across the genome reinforcing the idea that c-MMEJ-mediated editing is primarily correlated to the underlying genomic sequence (Supplementary Figure S6).

Applying RIMA to evaluate the effects of DNA repair inhibitors

Pharmacological inhibitors and transcriptional regulators of key DNA repair enzymes can modulate the balance between DNA repair pathways. To determine the utility of RIMA in delineating the effect of DNA repair inhibitors, HEK293 and HCT116 cells transfected with Cas9 and sgRNA expressing plasmids were treated with different concentrations of the MRN complex inhibitor, Mirin (55,56) (Figure 3A). The MRN complex plays an important role in the initial processing of double-strand DNA breaks prior to repair by homologous recombination or c-MMEJ. The addition of Mirin decreased the overall mutagenesis rate at all three target sites in both cell lines (Figure 3B). However, according to our analysis of the mutation patterns and fraction of c-MMEJ-associated reads, Mirin specifically inhibited c-MMEJ only at concentrations above 20 μM (Figure 3C and D). It is worth noting that MRE11, one of the MRN components, plays specific roles in both NHEJ and alt-EJ (57). In addition, we reanalysed a dataset from cells treated with the DNA-PK inhibitor, NU7441 (18); DNA-PK is required by the NHEJ pathways which is quantified within the other-EJ repair category within RIMA analysis. RIMA identified a strong decrease in other-EJ repair in response to DNA-PK inhibition during CRISPR–Cas9 endonuclease activity, suggesting the presence of enriched c-MMEJ-associated mutations (Supplementary Figure S7). RIMA also identified a strong depletion of single nucleotide insertions in NU7441 treated cells, which might indicate their formation by NHEJ. RIMA correctly identified the expected skew in DNA repair pathways and thus can be used to evaluate the effect of DNA repair inhibitors.

Analysis of the mutation patterns in POLQ knockout cells

Previous work has identified DNA polymerase theta (Polθ; encoded by POLQ gene in human) as playing a key role in alt-EJ repair of DSBs in mammalian cells (44,58). To elucidate any role of Polθ in c-MMEJ, we used RIMA to analyse repaired DSBs in a POLQ knockout setting. A Cas9 encoding transgene was first introduced into human A549 cells into the AAVS1 safe harbour locus using ObLiGaRe methodology (25). Both a pool and single clone of stably integrated cells were derived. We transfected the clonal line with a sgRNA targeting POLQ exon 6 and derived one clonal population in which all alleles harboured frameshifting mutations which should truncate and inactivate Polθ.

The A549-Cas9-Pool, A549-Cas9-Clone and A549-Cas9-Clone POLQ knockout cells were then transfected with 15 sgRNAs targeting independent loci. Genomic DNA was harvested at 12, 24 and 60 h after transfection (Figure 4A) and subjected to amplicon PCR and NGS analysis. The resultant dataset was processed using RIMA to quantify InDels frequencies (Figure 4B) and the fraction of c-MMEJ associated reads among modified reads (Figure 4C). We observed a significant reduction in the relative frequency of c-MMEJ associated mutations in POLQ knockout cells (Figure 4D) indicating a role for Polθ in the c-MMEJ subpathway and repair of Cas9-induced DSBs. We further validated these observations by tracing the c-MMEJ associated mutations at target sites (Figure 4E).

Analysis of single nucleotide InDels at Cas9 targeted sites

In c-MMEJ, the occurrence of deletions depends on the sequences surrounding the site of the DSB (15), while the precision of NHEJ remains controversial (25,26). Single nucleotide insertions and deletions were among the most frequently observed mutations at Cas9 target sites in our analysis (Supplementary Figure S8). The reproducibility of insertions at Cas9 target sites is potentially a consequence of offset nuclease activity of the RuvC and HNH domains to generate a 5′ single nucleotide overhang (59). However, the mechanisms underlying the other-EJ-mediated InDels and the possible role of Cas9 cleavage activity in dictating these mutations remains unclear. To date, various approaches, including *in vitro* biochemical studies (12,60,61), crystallography of Cas9 protein structure (9) and computational molecular dynamic simulations (13), have been performed to gain insight into the mechanism of DNA cleavage by Cas9. The RuvC and HNH nuclease domains within Cas9 mediate the cut in non-targeted and targeted DNA strands, respectively (Figure 5A). However, whether these two nuclease domains always cleave at the same position and whether

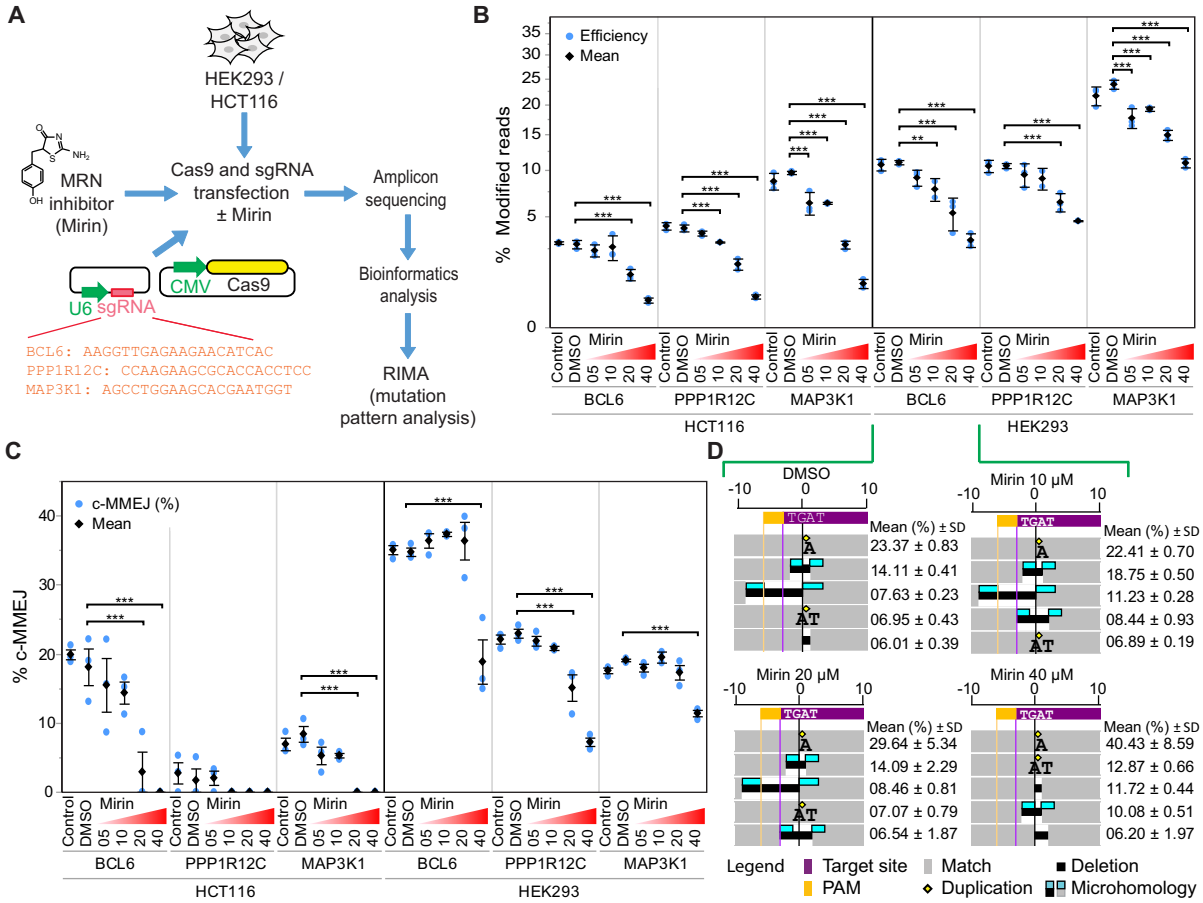

Figure 3. c-MMEJ-associated mutations are insensitive to the inhibition of the MRN complex. (**A**) Schematic of experimental procedures. Mirin was added to the cells 1 h before transfection. Plasmids expressing the Cas9 gene and sgRNA were transiently co-transfected into HEK293 or HCT116 cells. Genomic DNA was harvested 72 h after transfection for NGS and RIMA analysis. The percentages of the (**B**) modified reads and (**C** and **D**) c-MMEJ were determined for different Mirin doses (5, 10, 20 and 40 μM), DMSO vehicle or untreated control samples at three target sites. Data shown were obtained from three independent biological replicates. The TGAT nucleotides shown on the sgRNA are the second to fifth nucleotides before PAM. All error bars indicate the s.e.m. Dunnett's method was performed to statistically analyse and compare to the DMSO group (**$P < 0.01$; ***$P < 0.001$). Relative frequencies of each repair event are presented as the mean ± standard deviation (S.D.).

the nucleotides at the target site affect the cleavage mechanism remains unknown. We speculated that the mutation patterns at the CRISPR–Cas9 targeted sites could provide insight into the Cas9 cleavage activity in mammalian cells.

We therefore applied RIMA to perform a comprehensive in-depth analysis of other-EJ-associated small InDels at Cas9 target sites and to gain insight into the cleavage activity of Cas9 in living cells. We assigned the position of each nucleotide within the protospacer according to its distance from the PAM while considering the orientation of the target site (Figure 5B). First, we investigated whether the ratios of single nucleotide insertions and deletions were equal in different cell types following the same type of endonuclease cut. Interestingly, the cell type appeared to influence the balance between nucleotide insertions and deletions, suggesting that cells possess different capabilities for the end processing of DSBs via NHEJ and thus result in contrasting genome editing outcomes (Figure 5C).

Most single nucleotide insertions observed in our data occur between nucleotide numbers 4 and 3 (4∧3) in all cell types (Figure 5D). These observations suggest that DSBs occur between position 4 and 3, and are consistent with the previously described crystallographic structure (9) and dynamic simulation studies investigating Cas9 (13). Unexpectedly, a considerable percentage of the insertions (28.6% HCT116, 25.3% HEK293, 19.5% HeLa and 13.6% K562) were found between nucleotides 3 and 2 (3∧2), suggesting that the same DNA cleavage events performed by the HNH and RuvC domains occurred at position 3∧2. Remarkably, the single nucleotide insertions at either position (3∧2 and 4∧3) were not random, but these insertions were significantly biased towards duplicating their preceding nucleotides (3∧2 = 3 and 4∧3 = 4, Figure 5E), indicating that DNA break-repair had a precise and predictable underlying mechanism. In contrast to the insertions, the single nucleotide deletions at positions 4 and 3 were comparable among all cell types tested (Figure 5F).

Subsequently, we sought to determine whether the nucleotides at the Cas9-target sites are associated with any bias in the occurrence of single nucleotide InDels. Unex-

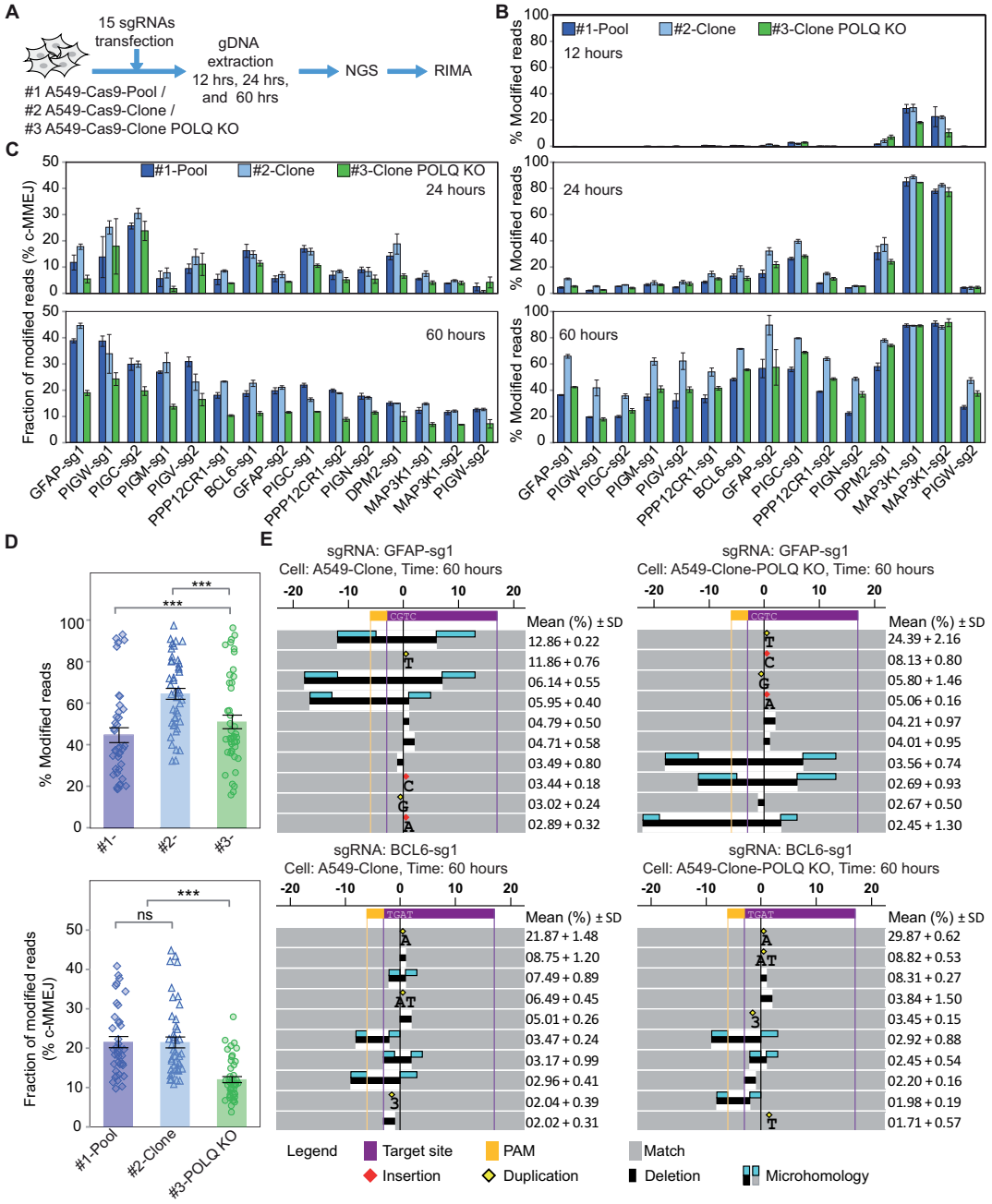

Figure 4. Polθ contributes to the formation of c-MMEJ associated deletions. (**A**) Schematic of experimental procedures. Three cell lines were transfected with one of 15 sgRNA expressing plasmids. Genomic DNA was harvested at 12, 24 and 60 h after transfection and were subjected to NGS and RIMA analysis. (**B**) Overall mutagenesis and (**C**) c-MMEJ rates are shown for all sgRNAs at indicated time point. (**D**) The summary of B and C for data obtained at 60 h after transfection is shown top and bottom, respectively. Significance determined by students t-test: non-significant (ns), ***$P < 0.001$. Error bars on all graphs indicate the S.E.M. for three independent biological replicates. (**E**) Mutation patterns were visualized using RIMA for two sgRNAs (GFAP-sg1 and BCL6-sg1) in A549-Clone and A549-Clone-POLQ-KO cells at 60 h time after transfection. Relative frequencies of each repair event are presented as the mean ± standard deviation (S.D.).

pectedly, the frequency of the single nucleotide InDels was found to be lowest and highest in target sites containing the nucleotides guanine (G) and thymine (T) at position 4, respectively (Figure 5G and H, Supplementary Figure S9A). Interestingly, a 'G' at position 3 was also associated with a two-fold higher occurrence of single nucleotide InDels than a 'G' at position 4 (Figure 5H). Given these observations, we investigated the association among the deletions (Figure 5I), fraction of c-MMEJ-associated deletions, and insertions (Supplementary Figure S9) at target sites containing different nucleotides at position 4. Interestingly, a 'G' at position 4 appeared to promote the formation of deletions rather than insertions at the target sites. This effect appears to be independent of the surrounding nucleotides

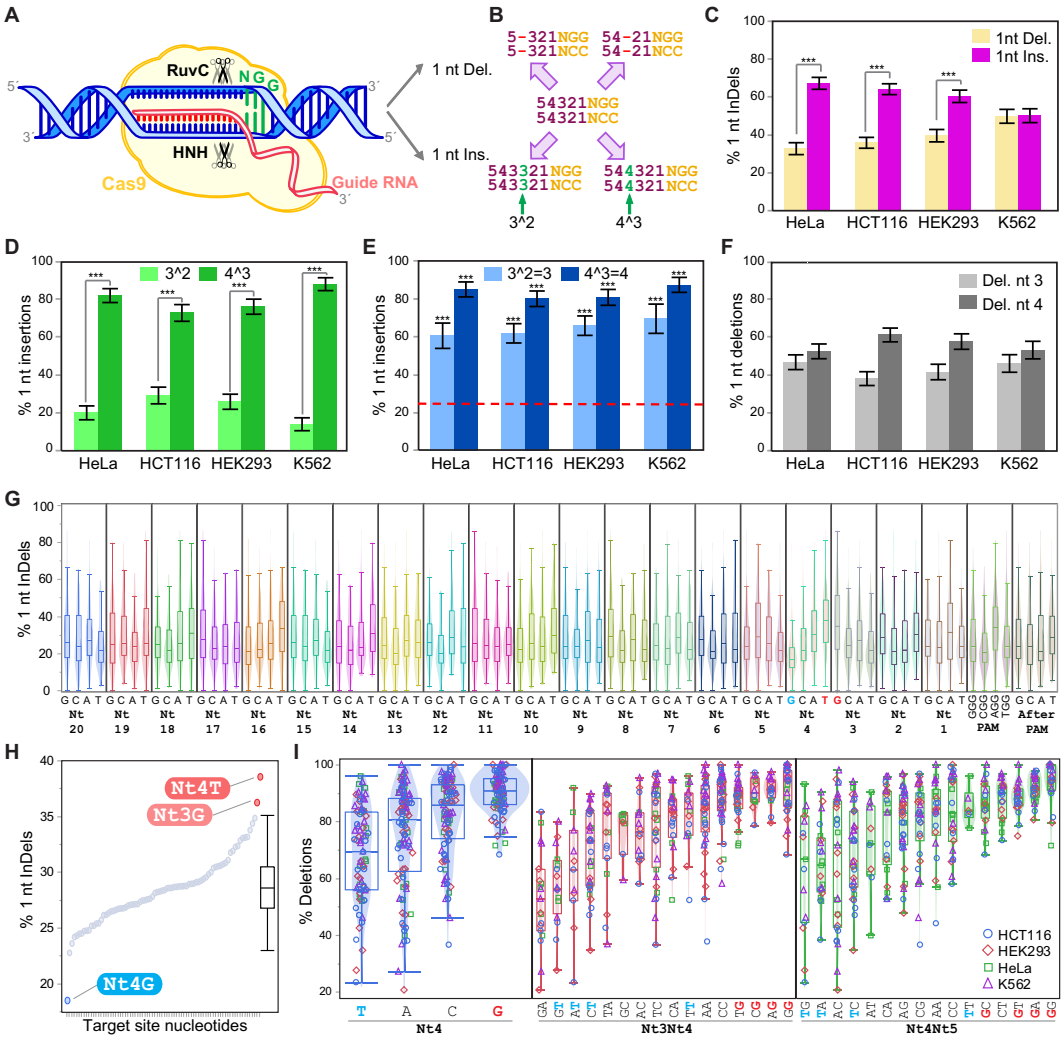

Figure 5. Analysis of single nucleotide insertions/deletions (InDels) at the Cas9 target site in different human cell lines. (**A**) Schematic of RNA-guided Cas9 targeting DNA. Cas9 (yellow) complexed with sgRNA (red) and bound to DNA (blue). RuvC and HNH nuclease domains cut the non-target and target strands, respectively. (**B**) Schematic of frequent single nucleotide insertions and deletions at different positions. For simplicity, the nucleotides were numbered according to their distance from the PAM. (**C**) Comparison of single nucleotide InDel frequencies at Cas9 target sites; nt, nucleotide. $N_{HeLa} = 67$, $N_{HEK293} = 94$, $N_{HCT116} = 94$, and $N_{K562} = 94$, *** $P < 0.001$. (**D**) Frequencies of single nucleotide insertions observed at position three (light green) or position four (dark green) relative to the PAM. To avoid the ambiguity of the mutation locations, only target sites with different nucleotides at positions three and four were analysed. (**E**) Percentage of similarities between the inserted single nucleotide and its 5′ precedent nucleotide compared to baseline. Red dashed line denotes a 25% random chance of a single nucleotide insertion to be similar to the adjacent 5′ nucleotide. ***$P < 0.001$ for comparisons among means and the baseline. (**F**) Observed frequencies of each single nucleotide deletion at the Cas9 target sites. Only target sites with different nucleotides at the cut sites (sgRNAs with different nucleotides at positions 3 and 4) were selected to precisely locate the InDels. All error bars represent the s.e.m.; the numbers of target sites shown in d, e and f were as follows: $N_{HeLa} = 50$, $N_{HEK293} = 58$, $N_{HCT116} = 58$, and $N_{K562} = 58$. Student's *t*-test (one-tailed) was performed for the statistical analysis (***$P < 0.001$). (**G**) The fraction of single nucleotide InDels at target sites with different nucleotides at positions 20 nucleotides before to one nucleotide after PAM. The minimum (blue) and maximum (red) observed InDels rate are highlighted. (**H**) The fraction of single nucleotide InDels is plotted against the target site nucleotides. (**I**) Association between deletions and different nucleotides at position 4. $N_{HeLa} = 67$, $N_{HEK293} = 94$, $N_{HCT116} = 94$ and $N_{K562} = 94$. All results illustrated in this figure were obtained from an analysis of the mutation patterns after 48 h (for HEK293, HCT116 and K562 cells) (18) and 72 h (for HeLa cells) (15).

at positions 3 and 5 (Figure 5I). Further studies are needed to gain better insight into the mechanisms underlying these repair events. However, the nucleotide composition at position 4 could influence the Cas9 nuclease activity, and hence, a distinct repair pattern is induced. Additionally, a 'G' at the broken site could interact with DNA repair enzymes and promote deletions rather than insertions. Regardless of the mechanism, these findings could be used to improve the prediction of genome editing outcomes.

Proposed mechanism of Cas9 nuclease activity in mammalian cells

Based on our findings, we propose a model in which the nuclease domains of Cas9 can catalyse hydrolysis of the phosphate backbone between nucleotides 2 to 5 (Figure 6A). The flexibility of the Cas9 nuclease domains considered in our proposed model could produce both blunt and staggered DNA with a certain degree of predictability. To support the generation of broken blunt ends, we studied the mutation

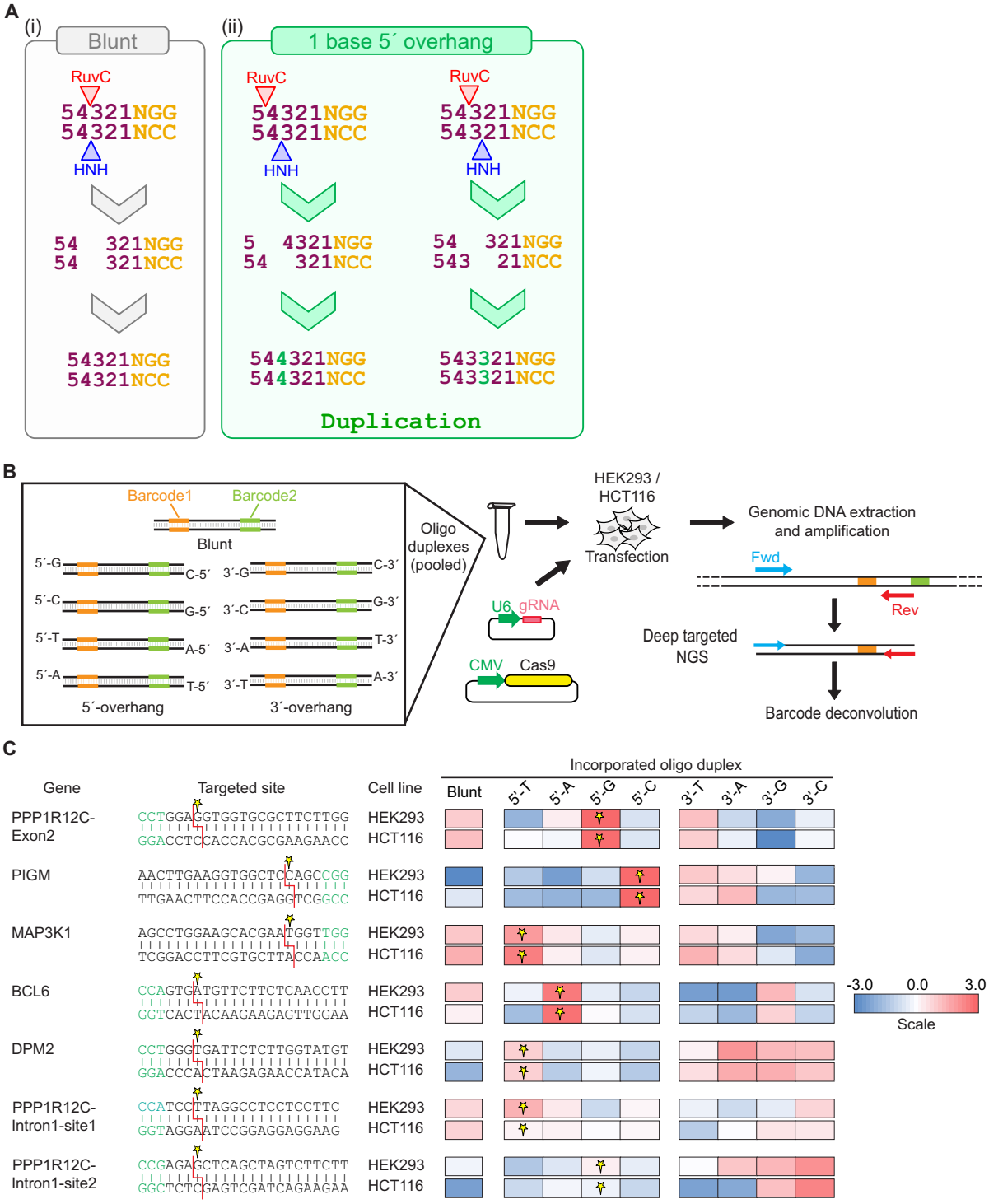

Figure 6. Cas9 endonuclease activity generates frequent 5′ overhangs. (**A**) Model explaining the interplay between DNA repair and the catalytic activity of the Cas9 nuclease domains: (i) shows the cut position of the RuvC and HNH domains to generate blunt ends and subsequent NHEJ-mediated precise repair, (ii) shows the insertion generation by staggered cuts creating a single nucleotide 5′ overhangs followed by DNA polymerases ends filling the overhangs and subsequent NHEJ repair leading to duplication of the preceding 5′ nucleotide. For simplicity, the nucleotides upstream PAM are numbered. (**B**) Schematic of the competitive oligo-duplex incorporation assay. Oligo-duplexes were co-transfected with Cas9 and sgRNA expressing plasmids. Genomic DNA was extracted from cells 72 h after transfection for barcode deconvolution via NGS analysis. (**C**) The pattern of the oligo-duplexes captured at seven target sites. PAM is shown in green. The expected position of the cuts induced by Cas9 resulting in a single nucleotide 5′ overhang is shown by a red line on the protospacer sequence. The staggered nucleotide is indicated by a yellow star. Heatmaps are generated based on normalised values, ranging from high (red) to low (blue) detection frequency. Each cell in the heatmap represents the mean of three independent biological replicates in one experiment.

patterns at genomic regions simultaneously targeted by two adjacent sgRNAs (Supplementary Figure S10). We detected perfect deletions in 87 of 203 mutant sequences in which the DNA sequence between the two Cas9 cut sites (between nucleotides at positions 3 and 4 upstream of the PAM sequence) was precisely deleted. These results imply a considerable proportion of Cas9-mediated blunt cuts and provide evidence of precise repair by NHEJ.

To determine whether Cas9 can generate staggered end, we designed and performed a competitive oligo-duplex incorporation assay in HEK293 and HCT116 cells (Figure 6B). This design is similar to the GUIDE-seq method (62) where cleavage events are identified via incorporation of a doubled stranded DNA template. In this experiment, we co-transfected cells with plasmids expressing Cas9 and sgRNA together with a pool of 48 bp barcoded oligo-duplexes which were blunt or contained all possible single nucleotide overhangs. This pool should report on the type of DNA break that has occurred as blunt oligo-duplexes should be preferentially incorporated into blunt ended breaks whereas overhang containing oligo-duplexes will preferentially be incorporated into breaks with the complementary overhangs. The oligo-duplex pool was co-transfected with seven independent sgRNAs separately and the frequency of oligo-duplexes incorporated at each target site was then determined by deep targeted amplicon sequencing after three days. While the pattern of oligo-duplexes with 3′ overhangs showed no clear pattern amongst the targeted sites, those with 5′ overhangs showed a clear preference (Figure 6C). In all seven target sites and in both cell lines tested, the most frequently integrated oligo-duplex was the one with the complementary overhang predicted by a 3∧4 staggered cut. These results support the notion that Cas9 can generate a variety of cuts with blunt and 3∧4 cleavage events.

The repair of DNA breaks with 5′ overhangs often results in insertions at the repair site with complete or partial duplications of the protruding overhangs. Based on the above observations, we speculated that the Cas9 cuts might generate single-strand 5′ overhangs with some flexibility in length and position relative to the PAM. Although the frequency of insertions larger than one nucleotide was modest in the NGS data, our analysis again showed a high level of duplications among these insertions, supporting the notion that Cas9 can hydrolyse the backbone at different positions.

Comparison of nuclease mutational signatures

To further validate the proposed association between the cleavage mechanism and DNA repair outcome, we analysed the mutation patterns generated by different nucleases with independent cleavage mechanisms. RIMA analysis of ZFN, TALEN and *Fok*I generated breaks revealed the expected insertions/duplications at the targeted loci (Supplementary Figure S11). Four nucleotide duplications were prevalent at *Fok*I targets, which can be caused by duplication of the four nucleotides 5′ overhang generated by *Fok*I catalytic mechanism (Supplementary Figure S11) (63,64). These analyses support the notion that cleavage mechanism can dictate the repair outcome, which can be revealed by RIMA.

Several SpCas9 variants have been reported which reduce off-target activity. To understand whether these mutations would result in differences in catalytic mechanism or repair outcome, we analysed the mutation patterns after cleavage using wild type SpCas9 or the high-fidelity variant, SpCas9-HF1 (65). We compared the overall genome editing efficiencies of SpCas9-wt and SpCas9-HF1 guided by full-length or truncated sgRNAs against a target site in the AAVS1 locus (Supplementary Figure S12A, B). SpCas9-HF1 resulted in low editing rates reducing the possibility of accurately capturing all possible genome editing outcomes. Nevertheless, in cases with a cleavage efficiency above 5%, we observed that while the same base was duplicated in single nucleotide insertions, the frequency of single nucleotide insertions induced by SpCas9-HF1 was two-fold higher than that induced by SpCas9-wt (Supplementary Figure S12C). This data suggests the high fidelity variant results in a subtle difference in alignment of catalytic residues during cleavage favouring the 3∧4 cleavage pattern.

Finally, we analysed the mutation pattern of Cas9 from *Francisella novicida* U112 (FnCas9) which generates four nucleotides 5′ overhangs (50). FnCas9 has the same PAM requirement as SpCas9 allowing the same sgRNA to be used for a direct comparison of these two enzymes (50,66). We used RIMA to compare the mutation patterns at five target sites in the human genome, all of which were targeted using SpCas9 and FnCas9 along with an exonuclease and a polymerase (Figure 7A). In addition, we attempted to test whether coupling Cas9 nucleases with end processing enzymes would change the mutation patterns. TREX2 gene encodes a nuclear protein with 3′ to 5′ exonuclease activity. DNTT is a member of the DNA polymerase type-X family that encodes a template-independent DNA polymerase. Consistent with a previous report (19), the exogenous overexpression of TREX2 exonuclease combined with either SpCas9 or FnCas9 increased the mutagenesis rates (Figure 7B). In all cases, the relative frequency of the insertions was reduced after TREX2 expression (Figure 7C). However, the increased mutagenesis rates after the TREX2 overexpression were not always associated with an increased ratio of c-MMEJ-associated deletions (Figure 7D). Interestingly, for the same target site, we observed that SpCas9 and FnCas9 induce different deletions and insertions; insertions after SpCas9 cleavage were one nucleotide in length, FnCas9-mediated insertions were three to five nucleotides in length. In both cases, the insertions were characterized as duplications of the preceding nucleotides, all of which agrees with the expected catalytic mechanisms. (Figure 7E, Supplementary Figures S13 and S14).

DISCUSSION

In this study, we conducted a comprehensive analysis of Cas9-induced mutational signatures at CRISPR targeted genetic loci in human cells. Our study extends the previous analysis of Cas9-induced non-random mutational signatures (18) and highlights a series of factors that impact on the predictability of the genome-editing outcome:

1) *DNA repair pathways*: alt-EJ sub-pathways seem to be relevant processes used to repair Cas9 induced DSBs and co-exist together with c-NHEJ. However, the kinetics of the two pathways seems to be different. NHEJ-associated

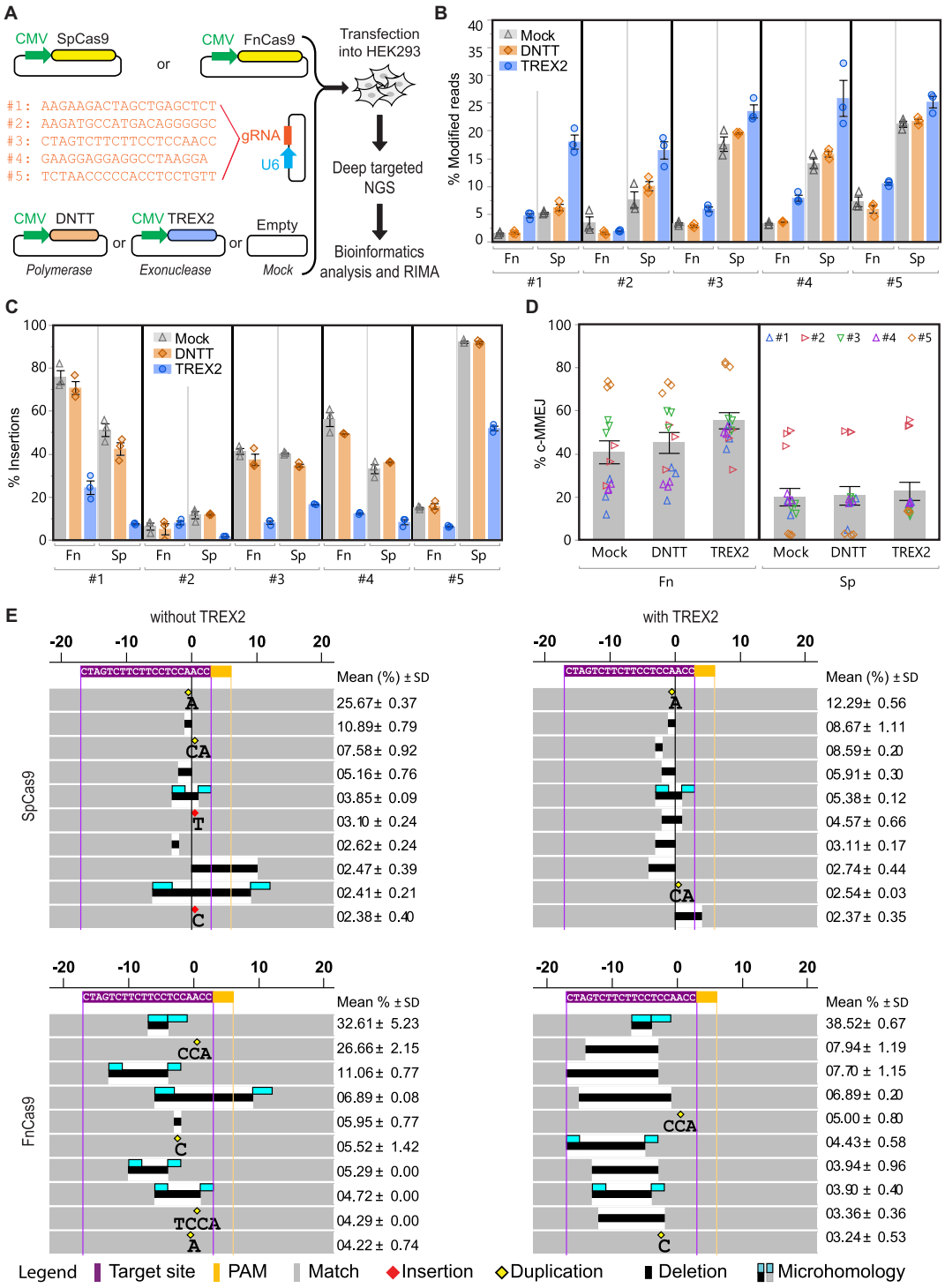

Figure 7. Delineation of mutational mechanism by RIMA: c-MMEJ is unperturbed by exonucleases activity at Cas9 induced breakpoints whilst they increase overall mutations but decrease insertions. (**A**) Schematic of experimental design used to investigate the effect of TREX2 and DNTT on the SpCas9 and FnCas9 mutagenesis rate and mutation patterns. Genetic modifications were identified by performing deep sequencing 72 h after transfection. (**B**) Percentage of modified reads and (**C**) percentage of insertions from cells mock (grey), DNTT (orange) or TREX2 (blue) co-transfected with SpCas9 or FnCas9. (**D**) Quantification of c-MMEJ at all target sites in analysed cell. All error bars represent the S.E.M. of three independent biological replicates in one experiment. (**E**) Visualized mutation patterns at one genomic locus targeted using SpCas9 and FnCas9 with and without the overexpression of TREX2. S.D.: standard deviation.

mutations arise at early time points (first few hours after introduction of the DSB), while c-MMEJ (a sub-pathway of alt-EJ) is much slower and is more clearly detectable after 2–3 days of a Cas9 induced cut. This is in line with early studies that suggested a slower speed for alt-EJ sub-pathways (30,33). Therefore, to experimentally obtain a realistic mutation pattern of a sgRNA in human cells, it is important to use later time points for analysis (around 72 h after transfection). Our findings also highlight that the precision of NHEJ and alt-EJ are not directly comparable; c-NHEJ can be very precise while alt-EJ is intrinsically error-prone. Therefore, measuring the frequency of these two pathways based on the mutation patterns may results in misleading conclusions because the level of NHEJ in this situation is underestimated.

2) *Cas9-nuclease mechanism*: consistent with previous reports (59), our findings suggest that Cas9 nuclease activity followed by NHEJ is the mechanism underlying the duplication of nucleotides at repaired sites (Figure 5A). This finding is also supported by a recent study in budding yeast (67) and a biochemical analysis of Cas9 nuclease activity (68). Interestingly, the latter study also showed that the Cas9 RuvC domain is capable of exonuclease activity on the cleaved target which can lead to its bidirectional degradation (68). This observation could also explain micro-deletions that are not attributed to obvious microhomologies. These collective findings could be used to develop genome-editing tools with more predictable repair outcome.
3) *The sequence of the target site as the substrate for DNA repair pathways*: both NHEJ and alt-EJ are context dependent. The c-MMEJ associated deletions were observed as the most frequent mutation in almost half of the target sites analysed in our work. Therefore, microhomologies near the cut may be used to predict frequent mutations.
4) *Interactions among nuclease, target site, and the repair pathways*: we evaluated the association between the InDel outcome and nucleotide composition of the CRISPR targeted sites and discovered that the identity of fourth and third nucleotides before the PAM play a significant role in the promotion of base deletion or base insertion. However, whether these nucleotides influence Cas9 nuclease activity or directly interact with the DNA repair machinery remain unknown.
5) *Coupling Cas9 with DNA repair modulators*: overexpression or depletion of end processing enzymes, or chemical inhibitors of DNA repair pathways can be further employed to modulate and gain insight into DNA repair pathways or developing new genome-editing methodologies. For example, it has been shown that coupling Cas9 with the exonucleases TREX2 or Artemis, can result in context-dependent increase of mutagenesis rate in human cells, over 20-fold in some cases (19). This indicates how efficiently the cycle of Cas9 cut and precise repair can happen and how additional players are influencing this process.

Our study highlights the power of computational tools to analyse NGS data and help interpret the mutational outcomes from genome editing experiments. RIMA provides a visual user-friendly interface based on an Excel spreadsheet that allows researchers to analyse data without any need for bioinformatics knowledge. RIMA offers the possibility of adjusting parameters, such as mutations detected in primers and variants outside of the Cas9 targeting region, to exclude calls and, therefore, minimizes the false-positive mutations. In this study, we also demonstrated the usefulness of RIMA to classify genetic mutations induced by CRISPR–Cas9 and precisely quantify and discriminate the contribution of c-MMEJ and other-EJ pathways to repair of a DSB. To date, different fluorescent-based assays have been developed to study the DNA repair mechanisms in mammalian cells (69–73). Such assays can be applied to screen small molecule compound libraries with potential inhibitory effects on DNA repair pathways. Nevertheless, developing methods to assess different types of repair outcomes at individual DSBs is challenging. Moreover, fluorescent-based assays require the generation of cellular models (i.e. knock-in of the reporter transgene) and may not be suitable for studying the DNA repair pathways in primary cells. We believe that RIMA can be used to maximally cover the occurrence of genome editing events associated with different DNA repair pathways. Furthermore, RIMA allows for the *in vivo* evaluation of the DNA repair inhibitory effects of small molecule compounds (e.g., in mouse models).

We have analysed newly generated and previously published data (15,16,18,19) to validate RIMA and provide a comprehensive and in-depth analysis of the mutation patterns after Cas9 cleavage in human cells. Our study proposes a new approach to characterize other types of Cas9 endonucleases that are active in mammalian cells based on their induced mutation patterns. Deep understanding of Cas9 catalytic activity is indispensable for certain applications. For example, GUIDE-seq, described as a technology for global detection of Cas9-mediated DSBs, relies on the insertion via NHEJ of blunt oligo-duplexes into the broken DNA ends on both on-target and off-target sites (62). Our results suggest that incorporation of staggered rather than blunt oligo-duplexes into the broken ends may occur more efficiently and therefore improve techniques such as GUIDE-seq to allow a more efficient, sensitive and precise detection of Cas9 off-targets. Similarly, these findings could be useful for designing more efficient NHEJ-based knock-in strategies by exploiting high efficiency ligation between the staggered ends of a Cas9 targeted transgene and target endogenous loci (74,75).

Our findings also suggest that engineered Cas9 variants, such as Cas9-HF1, can generate more homogenous mutational patterns. Therefore, the proper Cas9 choice and an accurate prediction of the targeting outcomes are critical for avoiding unwarranted or potentially adverse genetic alterations in therapeutic genome editing. Recently, several novel CRISPR–Cas9 systems have been developed, broadening the targetable regions in the genome. RIMA could be used to gain insight into the nuclease activity of novel Cas9 orthologue, such as SaCas9 (76), NmCas9 (77), GeoCas9 (78), and Cpf1 (79), and engineered forms of SpCas9, such as espCas9 (80) and HypaCas9 (81), and may provide opportunities for controlling the repair outcome.

CONCLUSIONS

In summary, the data presented here sheds light on the Cas9 catalytic activity inside cells and provides a better understanding of the repair mechanisms underlying Cas9-induced DNA lesions. RIMA offers a user-friendly tool to provide detailed characterisation of mutation patterns from high-throughput NGS data and could be further exploited to study DNA repair pathway selection in cell models. Moreover, our approach provides a unique opportunity to characterize the catalytic activity and processing of not only Cas9 orthologues but also of rationally engineered nucleases. Most importantly, our findings help guide the prediction of Cas9-mediated genome editing outcomes.

SUPPLEMENTARY DATA

Supplementary Data are available at NAR Online.

ACKNOWLEDGEMENTS

We would like to thank the entire AstraZeneca Transgenic team for thoughtful input and discussion. Additional thanks to Damla Etal, Steven Van Der Hoek and Ferdous Ur Rahman for the help with preparing the NGS libraries.

FUNDING

Funding for open access charge: AstraZeneca.
Conflict of interest statement. None declared.

REFERENCES

1. Lillestol,R.K., Redder,P., Garrett,R.A. and Brugger,K. (2006) A putative viral defence mechanism in archaeal cells. *Archaea*, **2**, 59–72.
2. Makarova,K.S., Grishin,N.V., Shabalina,S.A., Wolf,Y.I. and Koonin,E.V. (2006) A putative RNA-interference-based immune system in prokaryotes: computational analysis of the predicted enzymatic machinery, functional analogies with eukaryotic RNAi, and hypothetical mechanisms of action. *Biol. Direct*, **1**, 7.
3. Marraffini,L.A. and Sontheimer,E.J. (2010) CRISPR interference: RNA-directed adaptive immunity in bacteria and archaea. *Nat. Rev. Genet.*, **11**, 181–190.
4. Jinek,M., Chylinski,K., Fonfara,I., Hauer,M., Doudna,J.A. and Charpentier,E. (2012) A programmable dual-RNA-guided DNA endonuclease in adaptive bacterial immunity. *Science*, **337**, 816–821.
5. Mali,P., Yang,L., Esvelt,K.M., Aach,J., Guell,M., DiCarlo,J.E., Norville,J.E. and Church,G.M. (2013) RNA-guided human genome engineering via Cas9. *Science*, **339**, 823–826.
6. Cong,L., Ran,F.A., Cox,D., Lin,S., Barretto,R., Habib,N., Hsu,P.D., Wu,X., Jiang,W., Marraffini,L.A. *et al.* (2013) Multiplex genome engineering using CRISPR/Cas systems. *Science*, **339**, 819–823.
7. Doudna,J.A. and Charpentier,E. (2014) Genome editing. The new frontier of genome engineering with CRISPR–Cas9. *Science*, **346**, 1258096.
8. Nishimasu,H., Ran,F.A., Hsu,P.D., Konermann,S., Shehata,S.I., Dohmae,N., Ishitani,R., Zhang,F. and Nureki,O. (2014) Crystal structure of Cas9 in complex with guide RNA and target DNA. *Cell*, **156**, 935–949.
9. Jinek,M., Jiang,F., Taylor,D.W., Sternberg,S.H., Kaya,E., Ma,E., Anders,C., Hauer,M., Zhou,K., Lin,S. *et al.* (2014) Structures of Cas9 endonucleases reveal RNA-mediated conformational activation. *Science*, **343**, 1247997.
10. Sternberg,S.H., LaFrance,B., Kaplan,M. and Doudna,J.A. (2015) Conformational control of DNA target cleavage by CRISPR–Cas9. *Nature*, **527**, 110–113.
11. Jiang,F., Taylor,D.W., Chen,J.S., Kornfeld,J.E., Zhou,K., Thompson,A.J., Nogales,E. and Doudna,J.A. (2016) Structures of a CRISPR–Cas9 R-loop complex primed for DNA cleavage. *Science*, **351**, 867–871.
12. Geng,Y., Deng,Z. and Sun,Y. (2016) An insight into the protospacer adjacent motif of Streptococcus pyogenes Cas9 with artificially stimulated RNA-guided-Cas9 DNA cleavage flexibility. *RSC Adv.*, **6**, 33514–33522.
13. Zuo,Z. and Liu,J. (2016) Cas9-catalyzed DNA cleavage generates staggered Ends: Evidence from molecular dynamics simulations. *Sci. Rep.*, **5**, 37584.
14. Helleday,T., Eshtad,S. and Nik-Zainal,S. (2014) Mechanisms underlying mutational signatures in human cancers. *Nat. Rev. Genet.*, **15**, 585–598.
15. Bae,S., Kweon,J., Kim,H.S. and Kim,J.S. (2014) Microhomology-based choice of Cas9 nuclease target sites. *Nat. Methods*, **11**, 705–706.
16. Kim,Y., Kweon,J. and Kim,J.S. (2013) TALENs and ZFNs are associated with different mutation signatures. *Nat. Methods*, **10**, 185.
17. Robert,F., Barbeau,M., Ethier,S., Dostie,J. and Pelletier,J. (2015) Pharmacological inhibition of DNA-PK stimulates Cas9-mediated genome editing. *Genome Med*, **7**, 93.
18. van Overbeek,M., Capurso,D., Carter,M.M., Thompson,M.S., Frias,E., Russ,C., Reece-Hoyes,J.S., Nye,C., Gradia,S., Vidal,B. *et al.* (2016) DNA repair profiling reveals nonrandom outcomes at Cas9-Mediated breaks. *Mol. Cell*, **63**, 633–646.
19. Chari,R., Mali,P., Moosburner,M. and Church,G.M. (2015) Unraveling CRISPR–Cas9 genome engineering parameters via a library-on-library approach. *Nat. Methods*, **12**, 823–826.
20. Lin,S., Staahl,B.T., Alla,R.K. and Doudna,J.A. (2014) Enhanced homology-directed human genome engineering by controlled timing of CRISPR/Cas9 delivery. *Elife*, **3**, e04766.
21. Gutschner,T., Haemmerle,M., Genovese,G., Draetta,G.F. and Chin,L. (2016) Post-translational regulation of Cas9 during G1 enhances Homology-Directed repair. *Cell Rep.*, **14**, 1555–1566.
22. 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.
23. 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.
24. Ceccaldi,R., Rondinelli,B. and D'Andrea,A.D. (2016) Repair pathway choices and consequences at the double-strand break. *Trends Cell Biol*, **26**, 52–64.
25. Maresca,M., Lin,V.G., Guo,N. and Yang,Y. (2013) Obligate ligation-gated recombination (ObLiGaRe): custom-designed nuclease-mediated targeted integration through nonhomologous end joining. *Genome Res.*, **23**, 539–546.
26. Betermier,M., Bertrand,P. and Lopez,B.S. (2014) Is non-homologous end-joining really an inherently error-prone process? *PLoS Genet.*, **10**, e1004086.
27. Boulton,S.J. and Jackson,S.P. (1996) Saccharomyces cerevisiae Ku70 potentiates illegitimate DNA double-strand break repair and serves as a barrier to error-prone DNA repair pathways. *EMBO J.*, **15**, 5093–5103.
28. 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.
29. Kabotyanski,E.B., Gomelsky,L., Han,J.O., Stamato,T.D. and Roth,D.B. (1998) Double-strand break repair in Ku86- and XRCC4-deficient cells. *Nucleic Acids Res.*, **26**, 5333–5342.
30. Perrault,R., Wang,H., Wang,M., Rosidi,B. and Iliakis,G. (2004) Backup pathways of NHEJ are suppressed by DNA-PK. *J. Cell. Biochem.*, **92**, 781–794.
31. Corneo,B., Wendland,R.L., Deriano,L., Cui,X., Klein,I.A., Wong,S.Y., Arnal,S., Holub,A.J., Weller,G.R., Pancake,B.A. *et al.* (2007) Rag mutations reveal robust alternative end joining. *Nature*, **449**, 483–486.
32. Sfeir,A. and Symington,L.S. (2015) Microhomology-Mediated end Joining: a back-up survival mechanism or dedicated pathway? *Trends Biochem. Sci.*, **40**, 701–714.
33. Iliakis,G. (2009) Backup pathways of NHEJ in cells of higher eukaryotes: cell cycle dependence. *Radiother. Oncol.*, **92**, 310–315.

34. Yan,C.T., Boboila,C., Souza,E.K., Franco,S., Hickernell,T.R., Murphy,M., Gumaste,S., Geyer,M., Zarrin,A.A., Manis,J.P. *et al.* (2007) IgH class switching and translocations use a robust non-classical end-joining pathway. *Nature*, **449**, 478–482.
35. 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.
36. Sharma,S., Javadekar,S.M., Pandey,M., Srivastava,M., Kumari,R. and Raghavan,S.C. (2015) Homology and enzymatic requirements of microhomology-dependent alternative end joining. *Cell Death Dis.*, **6**, e1697.
37. Wang,M., Wu,W., Wu,W., Rosidi,B., Zhang,L., Wang,H. and Iliakis,G. (2006) PARP-1 and Ku compete for repair of DNA double strand breaks by distinct NHEJ pathways. *Nucleic Acids Res.*, **34**, 6170–6182.
38. 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.
39. Koole,W., van Schendel,R., Karambelas,A.E., van Heteren,J.T., Okihara,K.L. and Tijsterman,M. (2014) A polymerase theta-dependent repair pathway suppresses extensive genomic instability at endogenous G4 DNA sites. *Nat. Commun.*, **5**, 3216.
40. Yousefzadeh,M.J., Wyatt,D.W., Takata,K., Mu,Y., Hensley,S.C., Tomida,J., Bylund,G.O., Doublie,S., Johansson,E., Ramsden,D.A. *et al.* (2014) Mechanism of suppression of chromosomal instability by DNA polymerase POLQ. *PLoS Genet.*, **10**, e1004654.
41. 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.
42. Mateos-Gomez,P.A., Kent,T., Deng,S.K., McDevitt,S., Kashkina,E., Hoang,T.M., Pomerantz,R.T. and Sfeir,A. (2017) The helicase domain of Poltheta counteracts RPA to promote alt-NHEJ. *Nat. Struct. Mol. Biol.*, **24**, 1116–1123.
43. Mateos-Gomez,P.A., Gong,F., Nair,N., Miller,K.M., Lazzerini-Denchi,E. and Sfeir,A. (2015) Mammalian polymerase theta promotes alternative NHEJ and suppresses recombination. *Nature*, **518**, 254–257.
44. 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.
45. Wood,R.D. and Doublie,S. (2016) DNA polymerase theta (POLQ), double-strand break repair, and cancer. *DNA Repair (Amst.)*, **44**, 22–32.
46. Newman,J.A., Cooper,C.D., Aitkenhead,H. and Gileadi,O. (2015) Structure of the helicase domain of DNA polymerase theta reveals a possible role in the microhomology-mediated end-joining pathway. *Structure*, **23**, 2319–2330.
47. Beagan,K. and McVey,M. (2016) Linking DNA polymerase theta structure and function in health and disease. *Cell. Mol. Life Sci.*, **73**, 603–615.
48. Pinello,L., Canver,M.C., Hoban,M.D., Orkin,S.H., Kohn,D.B., Bauer,D.E. and Yuan,G.C. (2016) Analyzing CRISPR genome-editing experiments with CRISPResso. *Nat. Biotechnol.*, **34**, 695–697.
49. Park,J., Lim,K., Kim,J.S. and Bae,S. (2017) Cas-analyzer: an online tool for assessing genome editing results using NGS data. *Bioinformatics*, **33**, 286–288.
50. Chen,F., Ding,X., Feng,Y., Seebeck,T., Jiang,Y. and Davis,G.D. (2017) Targeted activation of diverse CRISPR-Cas systems for mammalian genome editing via proximal CRISPR targeting. *Nat. Commun.*, **8**, 14958.
51. Brinkman,E.K., Chen,T., Amendola,M. and van Steensel,B. (2014) Easy quantitative assessment of genome editing by sequence trace decomposition. *Nucleic Acids Res.*, **42**, e168.
52. Ye,J., Coulouris,G., Zaretskaya,I., Cutcutache,I., Rozen,S. and Madden,T.L. (2012) Primer-BLAST: a tool to design target-specific primers for polymerase chain reaction. *BMC Bioinformatics*, **13**, 134.
53. Wen,Q., Scorah,J., Phear,G., Rodgers,G., Rodgers,S. and Meuth,M. (2008) A mutant allele of MRE11 found in mismatch repair-deficient tumor cells suppresses the cellular response to DNA replication fork stress in a dominant negative manner. *Mol. Biol. Cell*, **19**, 1693–1705.
54. Rass,E., Grabarz,A., Plo,I., Gautier,J., Bertrand,P. and Lopez,B.S. (2009) Role of Mre11 in chromosomal nonhomologous end joining in mammalian cells. *Nat. Struct. Mol. Biol.*, **16**, 819–824.
55. Dupre,A., Boyer-Chatenet,L., Sattler,R.M., Modi,A.P., Lee,J.H., Nicolette,M.L., Kopelovich,L., Jasin,M., Baer,R., Paull,T.T. *et al.* (2008) A forward chemical genetic screen reveals an inhibitor of the Mre11-Rad50-Nbs1 complex. *Nat. Chem. Biol.*, **4**, 119–125.
56. Garner,K.M., Pletnev,A.A. and Eastman,A. (2009) Corrected structure of mirin, a small-molecule inhibitor of the Mre11-Rad50-Nbs1 complex. *Nat. Chem. Biol.*, **5**, 129–130.
57. Xie,A., Kwok,A. and Scully,R. (2009) Role of mammalian Mre11 in classical and alternative nonhomologous end joining. *Nat. Struct. Mol. Biol.*, **16**, 814–818.
58. 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.
59. Li,Y., Park,A.I., Mou,H., Colpan,C., Bizhanova,A., Akama-Garren,E., Joshi,N., Hendrickson,E.A., Feldser,D., Yin,H. *et al.* (2015) A versatile reporter system for CRISPR-mediated chromosomal rearrangements. *Genome Biol.*, **16**, 111.
60. Gasiunas,G., Barrangou,R., Horvath,P. and Siksnys,V. (2012) Cas9-crRNA ribonucleoprotein complex mediates specific DNA cleavage for adaptive immunity in bacteria. *Proc. Natl. Acad. Sci. U.S.A.*, **109**, E2579–E2586.
61. Garneau,J.E., Dupuis,M.E., Villion,M., Romero,D.A., Barrangou,R., Boyaval,P., Fremaux,C., Horvath,P., Magadan,A.H. and Moineau,S. (2010) The CRISPR/Cas bacterial immune system cleaves bacteriophage and plasmid DNA. *Nature*, **468**, 67–71.
62. Tsai,S.Q., Zheng,Z., Nguyen,N.T., Liebers,M., Topkar,V.V., Thapar,V., Wyvekens,N., Khayter,C., Iafrate,A.J., Le,L.P. *et al.* (2015) GUIDE-seq enables genome-wide profiling of off-target cleavage by CRISPR-Cas nucleases. *Nat. Biotechnol.*, **33**, 187–197.
63. Schiml,S., Fauser,F. and Puchta,H. (2014) The CRISPR/Cas system can be used as nuclease for in planta gene targeting and as paired nickases for directed mutagenesis in Arabidopsis resulting in heritable progeny. *Plant J.*, **80**, 1139–1150.
64. Bothmer,A., Phadke,T., Barrera,L.A., Margulies,C.M., Lee,C.S., Buquicchio,F., Moss,S., Abdulkerim,H.S., Selleck,W., Jayaram,H. *et al.* (2017) Characterization of the interplay between DNA repair and CRISPR/Cas9-induced DNA lesions at an endogenous locus. *Nat. Commun.*, **8**, 13905.
65. Kleinstiver,B.P., Pattanayak,V., Prew,M.S., Tsai,S.Q., Nguyen,N.T., Zheng,Z. and Joung,J.K. (2016) High-fidelity CRISPR–Cas9 nucleases with no detectable genome-wide off-target effects. *Nature*, **529**, 490–495.
66. Hirano,H., Gootenberg,J.S., Horii,T., Abudayyeh,O.O., Kimura,M., Hsu,P.D., Nakane,T., Ishitani,R., Hatada,I., Zhang,F. *et al.* (2016) Structure and Engineering of Francisella novicida Cas9. *Cell*, **164**, 950–961.
67. 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.
68. Stephenson,A.A., Raper,A.T. and Suo,Z. (2018) Bidirectional degradation of DNA cleavage products catalyzed by CRISPR/Cas9. *J. Am. Chem. Soc.*, **140**, 3743–3750.
69. Kostyrko,K. and Mermod,N. (2016) Assays for DNA double-strand break repair by microhomology-based end-joining repair mechanisms. *Nucleic Acids Res.*, **44**, e56.
70. Certo,M.T., Ryu,B.Y., Annis,J.E., Garibov,M., Jarjour,J., Rawlings,D.J. and Scharenberg,A.M. (2011) Tracking genome engineering outcome at individual DNA breakpoints. *Nat. Methods*, **8**, 671–676.
71. Bennardo,N., Cheng,A., Huang,N. and Stark,J.M. (2008) Alternative-NHEJ is a mechanistically distinct pathway of mammalian chromosome break repair. *PLoS Genet.*, **4**, e1000110.
72. Guirouilh-Barbat,J., Rass,E., Plo,I., Bertrand,P. and Lopez,B.S. (2007) Defects in XRCC4 and KU80 differentially affect the joining

of distal nonhomologous ends. *Proc. Natl. Acad. Sci. U.S.A.*, **104**, 20902–20907.
73. Chu,V.T., Weber,T., Wefers,B., Wurst,W., Sander,S., Rajewsky,K. and Kuhn,R. (2015) Increasing the efficiency of homology-directed repair for CRISPR–Cas9-induced precise gene editing in mammalian cells. *Nat. Biotechnol.*, **33**, 543–548.
74. 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.
75. Suzuki,K., Tsunekawa,Y., Hernandez-Benitez,R., Wu,J., Zhu,J., Kim,E.J., Hatanaka,F., Yamamoto,M., Araoka,T., Li,Z. *et al.* (2016) In vivo genome editing via CRISPR/Cas9 mediated homology-independent targeted integration. *Nature*, **540**, 144–149.
76. Ran,F.A., Cong,L., Yan,W.X., Scott,D.A., Gootenberg,J.S., Kriz,A.J., Zetsche,B., Shalem,O., Wu,X., Makarova,K.S. *et al.* (2015) In vivo genome editing using Staphylococcus aureus Cas9. *Nature*, **520**, 186–191.
77. Hou,Z., Zhang,Y., Propson,N.E., Howden,S.E., Chu,L.F., Sontheimer,E.J. and Thomson,J.A. (2013) Efficient genome engineering in human pluripotent stem cells using Cas9 from Neisseria meningitidis. *Proc. Natl. Acad. Sci. U.S.A.*, **110**, 15644–15649.
78. Harrington,L.B., Paez-Espino,D., Staahl,B.T., Chen,J.S., Ma,E., Kyrpides,N.C. and Doudna,J.A. (2017) A thermostable Cas9 with increased lifetime in human plasma. *Nat. Commun.*, **8**, 1424.
79. Zetsche,B., Gootenberg,J.S., Abudayyeh,O.O., Slaymaker,I.M., Makarova,K.S., Essletzbichler,P., Volz,S.E., Joung,J., van der Oost,J., Regev,A. *et al.* (2015) Cpf1 is a single RNA-guided endonuclease of a class 2 CRISPR-Cas system. *Cell*, **163**, 759–771.
80. Slaymaker,I.M., Gao,L., Zetsche,B., Scott,D.A., Yan,W.X. and Zhang,F. (2016) Rationally engineered Cas9 nucleases with improved specificity. *Science*, **351**, 84–88.
81. Chen,J.S., Dagdas,Y.S., Kleinstiver,B.P., Welch,M.M., Sousa,A.A., Harrington,L.B., Sternberg,S.H., Joung,J.K., Yildiz,A. and Doudna,J.A. (2017) Enhanced proofreading governs CRISPR–Cas9 targeting accuracy. *Nature*, **550**, 407–410.