*Published online 5 June 2019*

*Nucleic Acids Research, 2019, Vol. 47, No. 15* ***7989–8003***
*doi: 10.1093/nar/gkz487*

# Massively parallel profiling and predictive modeling of the outcomes of CRISPR/Cas9-mediated double-strand break repair

**Wei Chen[1,2], Aaron McKenna[2,†], Jacob Schreiber[3,†], Maximilian Haeussler[4], Yi Yin[2], Vikram Agarwal[2], William Stafford Noble[2,3] and Jay Shendure[2,5,6,7,*]**

[1]Molecular Engineering and Sciences Institute, University of Washington, Seattle, WA 98195, USA, [2]Department of Genome Sciences, University of Washington, Seattle, WA 98195, USA, [3]Paul G. Allen School of Computer Science and Engineering, University of Washington, Seattle, WA 98195, USA, [4]Santa Cruz Genomics Institute, University of California, Santa Cruz, CA 95064, USA, [5]Brotman Baty Institute for Precision Medicine, University of Washington, Seattle, WA 98195, USA, [6]Howard Hughes Medical Institute, Seattle, WA 98195, USA and [7]Allen Discovery Center for Cell Lineage Tracing, Seattle, WA 98195, USA

Received January 28, 2019; Revised April 23, 2019; Editorial Decision May 20, 2019; Accepted May 23, 2019

## ABSTRACT

**Non-homologous end-joining (NHEJ) plays an important role in double-strand break (DSB) repair of DNA. Recent studies have shown that the error patterns of NHEJ are strongly biased by sequence context, but these studies were based on relatively few templates. To investigate this more thoroughly, we systematically profiled ∼1.16 million independent mutational events resulting from CRISPR/Cas9-mediated cleavage and NHEJ-mediated DSB repair of 6872 synthetic target sequences, introduced into a human cell line via lentiviral infection. We find that: (i) insertions are dominated by 1 bp events templated by sequence immediately upstream of the cleavage site, (ii) deletions are predominantly associated with microhomology and (iii) targets exhibit variable but reproducible diversity with respect to the number and relative frequency of the mutational outcomes to which they give rise. From these data, we trained a model that uses local sequence context to predict the distribution of mutational outcomes. Exploiting the bias of NHEJ outcomes towards microhomology mediated events, we demonstrate the programming of deletion patterns by introducing microhomology to specific locations in the vicinity of the DSB site. We anticipate that our results will inform investigations of DSB repair mechanisms as well as the design of CRISPR/Cas9 experiments for diverse applications including genome-wide screens, gene therapy, lineage tracing and molecular recording.**

## INTRODUCTION

Genome engineering conventionally involves using a programmable endonuclease (i.e. a zinc finger nuclease (ZFN), transcription activator-like effector nuclease (TALEN) or RNA guided nuclease Cas9 (clustered regularly interspaced short palindromic repeats-CRISPR associated protein CRISPR/Cas9) to introduce a double-strand break (DSB) at a specific location in the genome. In mammalian cells, such DSBs are primarily repaired by one of two pathways—homology directed repair (HDR) and classical non-homologous end joining (c-NHEJ) (1,2). HDR uses homologous template sequences to repair the DSB, potentially introducing programmed edits via the repair template. In contrast, c-NHEJ directly rejoins the broken ends, often perfectly but occasionally introducing errors, typically in the form of short insertions or deletions (indels) (3). In addition to HDR and cNHEJ, there is evidence for an alternative NHEJ pathway (alt-NHEJ), also termed microhomology mediated end joining (MMEJ), wherein short, homologous sequences in the vicinity of the DSB are used to align the broken ends prior to joining, resulting in deletions or potentially more complex events (4). Below, we use 'NHEJ' to refer to both c-NHEJ and MMEJ/alt-NHEJ, i.e. template-free editing.

In recent years, CRISPR/Cas9 has emerged as a particularly versatile tool for genome editing. For many if not most applications of CRISPR/Cas9-mediated genome engineering, it is used in conjunction with the cell's endogenous NHEJ machinery to introduce short indels in a targeted fashion (5–7), e.g. to disrupt the function of genes or regulatory elements (8–10) or to introduce irreversible changes that record cell lineage or molecular events (11–13). However, despite NHEJ's central role in this transformative

*To whom correspondence should be addressed. Tel: +1 206 685 8543; Fax:+1 206 685 7301; Email: shendure@uw.edu
†The authors wish it to be known that, in their opinion, the second and third authors should be regarded as Joint Second Authors.

tool, our understanding of the processes that determine the rate and patterns of NHEJ-mediated errors remains incomplete.

Recent studies have demonstrated that the error outcomes of NHEJ are strongly dependent on sequence context (14,15). Other studies show that the characteristics of the broken ends (blunt or staggered end; length of any overhang) also affect end-joining patterns both *in vitro* (16) and *in vivo* (16,17). However, a systematic profiling of the sequence determinants of NHEJ repair patterns has yet to be undertaken.

Here, we profiled ∼1.16 million mutational events resulting from *Streptococcus pyogenes* Cas9 (*Sp*Cas9)-mediated cleavage and NHEJ-mediated DSB repair of 6872 synthetic target sequences. From the resulting data, we identify the primary features of sequences adjacent to the sites of DSBs that shape the distribution and relative frequency of NHEJ-mediated mutational outcomes, e.g. nucleotide content and microhomology. We furthermore exploit microhomology to demonstrate the 'programming' of deletion patterns. Finally, we develop a logistic regression model to predict insertions and deletions (**Lindel**) that result from CRISPR/Cas9-mediated cleavage of an arbitrary sequence. A standalone Lindel webtool is freely available (https://lindel.gs.washington.edu), and Lindel predictions have been integrated into the CRISPOR web tool (http://www.crispor.org) (18).

## MATERIALS AND METHODS

### sgRNA and target pair library design:

To generate a library of CRISPR/Cas9 targets that could safely be characterized within human cells, we evaluated ∼1 million random 20mer crRNA sequences, scoring them against the human genome (version hg19) for off-target effects using FlashFry (45). We excluded guides with an exact match or up to two mismatches against any potential target in the human genome, or those with an off-target score <90 (46), resulting in a modest bias towards targets containing CpG dinucleotides (Supplementary Figure S7), and then selected a final library of 70 000 top scoring guides for synthesis. The resulting sgRNA sequence and their corresponding targets were separated by a common 20 bp spacer sequence and ordered as an Agilent SureGuide Unamplified Custom CRISPR Library array (Figure 1A).

To analyze the potential impact of programmed microhomology, we selected a subset of 1000 sgRNA-target pairs from the library above and introduced microhomology with different lengths (2, 4 and 6 bp) matching the last 2, 4 and 6 nucleotides upstream of the cleavage site. Each design was assigned a 4 bp barcode, indicating its programmed microhomology pattern (Figure 5A and B). This library of microhomology sequences was ordered as an oligo pool from Twist Biosciences.

### Library cloning

The lentiGuide-Puro (Addgene #52963) vector was modified with two rounds of PCR to remove the existing tracrRNA and filler sequence (primer P1, P2), and to incorporate two BsmBI restriction site for integration of sgRNA-target pairs (primer P3, P4). The modified vector was digested with BsmBI (NEB, Buffer 3.1) at 55°C for 3 h and gel purified with Monarch DNA Gel Extraction Kit (NEB). This digested and purified vector was used for all downstream cloning.

Oligos with sgRNA-target pairs from Agilent or Twist Bioscience were both resuspended to 10ng/μl. The oligo pool was PCR amplified using KAPA Biosystems HiFi HotStart ReadyMix 2× using primers P5 and P6 and cleaned with the DNA Clean&Concentrator kit (Zymo Research). The purified PCR product was then digested with BsmBI (NEB, buffer 3.1) at 55°C for 1 h to generate compatible sticky ends matching the modified lentiGuide-Puro above, and subsequently cleaned with DNA Clean&Concentrator (Zymo Research). Digested vector and insert were ligated with T4 ligase (NEB) with a molar ratio of 1:3. Ligation products were transformed in to Stable Competent *Escherichia coli* (NEB C3040H). Transformed cells were cultured at 30°C overnight and plasmid DNA was prepared using a ZymoPURE II Plasmid Kit. The subsampled library with 12 917 targets was bottlenecked by seeding transformed cells on plate. Colonies on plates were transferred to liquid medium to expand them. The precision of the number 12 917 follows from the fact that we can simply count the number of unique guide sequences present in deep sequencing of PCR amplicons from cells.

### Cell culture and lentivirus transduction

We generated a mono-clonal 293T cell line expressing Cas9 by transduction of Cas9-blast lentivirus particles (Addgene plasmid #52962). Cells were cultured in DMEM High glucose (GIBCO) supplemented with 10% fetal bovine serum (Rocky Mountain Biologicals) and 1% penicillin–streptomycin (GIBCO) and grown with 5% $CO_2$ at 37°C.

All lentivirus libraries were produced by the Fred Hutchinson Cooperative Center for Excellence in Hematology Vector Production core facility. HEK293T cells were transduced and media was changed to virus free media for 24 h post-transduction. Cells were passed every 48 h with a split ratio of 1:6. Cells were harvested at day 5 after transduction.

### Sequencing library generation

Genomic DNA was extracted with DNeasy Blood & Tissue Kit (Qiagen) following the manufacturer's protocol. 15 bp unique molecular identifiers (UMIs) were added by one initial round of linear PCR using a primer containing a 5′ sequencing adaptor (P7). For each reaction we used 250 ng of genomic DNA, 0.2 μl 100 mM primer and 25 μl HiFi HotStart ReadyMix 2× (KAPA Biosystems). PCR reaction were performed as follows: 95°C 3 min, 98°C 20 s, 5 cycles of 65°C 1 min and 72°C 2 min, 98°C 20 s, 5 cycles of 65°C 1 min and 72°C 2 min. The subsequent PCR product was cleaned with 1.8× AMPure XP beads (Beckman Coulter)

and resuspended in 25 μl of elution buffer. A second round of amplification was performed using primers targeting the 5′ sequencing adaptor (P8) and 50 bp downstream of the cleavage site (P9) for 20 cycles. The resulting PCR product was then size selected using a dual size-selection cleanup of 0.4× and 0.8× AMPure XP beads (Beckman Coulter) to remove genomic DNA and small fragments (<200 bp) respectively. This size-selected product was subsequently re-amplified to add the 3′ sequencing adaptor with primer P8 and P10 for an additional five cycles. The final PCR product was cleaned with 0.75x AMPure XP beads (Beckman Coulter) and was re-amplified to add flow-cell adaptor and sample index for 5 cycles. All PCR reactions used HiFi HotStart ReadyMix 2× (KAPA Biosystems) with the manufacturer's recommended conditions. The library was sequenced on an Illumina NextSeq 500 sequencer using paired-end 150 cycle reads. All primers used are listed in Supplementary Table S2. Sequencing data has been deposited in the Gene Expression Omnibus (GEO) with accession number GSE131421 and the processed data files are deposited in Figshare with a doi link: https://doi.org/10.6084/m9.figshare.7374155.

### Sequence processing pipeline

Across three replicates, we sequenced a total of 148 million paired-end reads on an Illumina NextSeq 500. We first clustered these paired-end reads by their 15 bp UMI sequence and then filtered out reads with <90% identity within their representative UMI clusters. Sequence identity was identified using edlib (47). UMIs with fewer than 10 reads were excluded from downstream analysis. This yielded 4 405 379 UMIs (91 325 700 reads), representing ∼61.8% of our sequencing data (Supplementary Table S3). We then selected the most common forward and reverse read sequence for each UMI for further processing. These forward and reverse reads were merged into a single read using PEAR (48) and aligned in a two step process as follows. First, we sought to identify the 'reference' sequences for each programmed array sequence. We aligned the merged reads to a backbone sequence where the guides and targets were represented by Ns using EMBO's needleall software (21) with the following scoring matrix: match = 5, mismatch = -4, gap-open = –20, gap-extension = –0.5. The mismatch penalty for *N*s was set to 0. The sequence over the guide region was then extracted and matched against the list of programmed array sequences. Guide sequences with more than two mismatches to the designed guides were excluded, with edit distances assessed with UMI-tools (49). Second, merged reads were aligned to their discovered reference, in which *N*s were replaced by the guide/target sequence identified from the first step, using Biopython.pairwise2 (50) with the following scoring matrix: match = 5, mismatch = –4, gap-open = –13, gap-extension = –0.5. All indels were then right aligned (e.g. Supplementary Figure S3A). Aligned reads with indels within –3/+2 bp of the cleavage site were assigned to their indel class. Aligned reads were excluded for downstream analysis if the sgRNA and target sequence didn't match, the result from template switch during lentivirus transduction (19,20), or unexpected mutations introduced during synthesis, cloning, and PCR. A final library of 1.19 million unique reads (UMIs) were identified. Our library of 1000 microhomology sequences were processed by this same pipeline, yielding a final library of 249 039 UMIs from 31 239 645 paired-end sequencing reads. Scripts and other software are available from our GitHub repository: https://github.com/shendurelab/Lindel.

### Data processing and analysis

*kpLogo analysis.* Sequence motif analysis was conducted with kpLogo (24) using default settings with a specified *k*-mer length of 1 or 2. Input sequences were weighted by the frequency of insertion.

*Microhomology identification.* For *n* from 1 to 10 nucleotides, the last *n* nucleotides upstream of each deletion were compared to the last *n* nucleotides of the deleted sequence (as the deletion is right aligned. Supplementary Figure S3A). The length of microhomology was identified as the largest *n* nucleotides match in sequence.

### Machine learning modeling

We phrased our problem of predicting repair outcomes and their frequencies as that of a classification task with 557 classes. Because large mutation events are rare, we limited our classification effort to deletion events <30 bp, and we grouped insertions ≥3 bp into one class. In total, we defined 557 classes of indels. These classes include 536 deletion alleles, 4 possible single nucleotide insertion, and 16 possible dinucleotide insertion and insertions ≥3 bp. There are a total of 550 potential deletion events that are both <30 bp in length and overlap with the –3/+2 window around the cleavage site. We captured 536 deletion alleles in our data; the missing 14 classes are mainly large deletions. As input to our model, we defined 3033 binary features. These are (i) Sequence features: 384 binary features corresponding to the one-hot encoded target sequence (excluding the PAM region), including 80 for single nucleotide content (4 nucleotides × 20 positions) and 304 for dinucleotide content (16 dinucleotides × 19 positions); (ii) Microhomology features: 2649 binary features corresponding to MH tracts; specifically, for each of the possible deletion event class, we defined five binary features (or 2–4, depending on the size of deletion) corresponding to the length of the MH tract, if any (0-4 bp × 519 + 0–3 bp × 7 + 0–2 bp × 6 + 0–1 bp × 4 deletion event classes = total 2649 binary features. Our 4790 programmed sequences were randomly partitioned into a training set of 3,900 sequences, a validation set of 450 sequences, and a test set of 440 sequences.

We trained the logistic regression in a standard manner for machine learning models. However, because each target sequence can generate many possible repair outcomes, we trained our models using soft labels that correspond to the probability that each class is observed, rather than hard labels that force each input to correspond exclusively to one class. Each model was trained using the Adam optimizer (51) with a learning rate of 0.001 and a categorical cross-entropy loss. Training proceeded for a maximum of 100 epochs with a 'patience' of 1, meaning that training was stopped after two epochs with no improvement in

validation set performance. All initializations and the hyperparameters for the Adam optimizer were set to the defaults in Keras v2.1.3 (52) with a backend of Theano v1.0.1 (53). We selected the best model based on performance on the validation set according to the coefficient of determination using grid search over hyperparameters. This search involved separate scans over regularization strengths for L1-regularization and L2-regularization individually with a range of $10^{-10}$ to $10^{-1}$ (Supplementary Figure S5D–F).

### Model comparison

We compared our model to two other models (ForeCasT and inDelphi) in the same setting. All models used 60 bp sequence centered at the cleavage site as an input, while trying to predict the frequency of 557 classes of indels that we defined above. As both Lindel and ForeCasT are predicting all possible unique repair outcomes, it's straightforward to them compare directly. inDelphi only predicts ∼90 classes deletions with locations, and 1–59 bp deletions regardless their locations. We used the classes that inDelphi predicts that overlap with the ForeCasT and Lindel classes, which included all ∼90 classes using microhomology, 1 bp deletion (assuming it's located the cleavage site), and 1 bp insertions. All classes that inDelphi is not predicting were assigned as 0. The performance were measured using MSE on all ∼450 unique classes for each sequence in our test set ($n = 440$) and the ForeCasT test set ($n = 4298$)

## RESULTS

### Development of a massively parallel strategy to profile NHEJ-mediated genome edits

Toward a comprehensive understanding of the sequence determinants of NHEJ-mediated error patterns, we developed a strategy that would allow us to efficiently profile a large number of repair events from each of a large number of sequence contexts (Figure 1). In brief, we designed 70 000 targets balanced in nucleotide content and screened against the human genome for CRISPR/Cas9 single guide RNAs (sgRNAs). We then used array-based oligonucleotide synthesis to encode these targets in *cis* with their corresponding sgRNAs, separated by a 20 bp spacer. We then amplified and cloned these molecules to a lentiviral vector. In our initial experiments, the complexity of the resulting library of synthetic targets and their cognate sgRNAs was such that we obtained relatively few edited templates per target. Therefore, we re-cloned the library under bottlenecking conditions (Materials and Methods), reducing its complexity to 12 917 targets. We then proceeded with viral packaging and transduction, in triplicate, of a monoclonal human embryonic kidney (HEK) 293T cell line that stably expresses *sp*Cas9 (multiplicity of infection of ∼4–8). As such, within any given cell, only one or a few sgRNAs are expressed, and each one directs Cas9-mediated DSBs to a target located immediately adjacent to it. After five days to allow for the introduction of NHEJ-mediated errors at these targets, cells were harvested and genomic DNA isolated. We then PCR amplified the region comprising the targets and corresponding sgRNAs using unique molecular identifiers (UMIs) appended during the first extension cycle to distinguish whether identical edits were derived from the same cell or different cells.

Summing across the three replicates, we sequenced PCR amplicons to a depth of ∼148 million reads, which were reduced to ∼1.19 million reads after collapsing on the basis of identical sequences and UMIs, and filtering of reads with evidence of lentivirus-mediated template switching (19,20) or other unexpected sequences (e.g. synthesis or PCR errors). After further filtering of poorly represented targets (those represented by fewer than 10 UMIs), our dataset consisted of ∼1.16 million UMIs corresponding to 6872 unique targets. On average, each target was represented by 168 UMIs and 24 alleles (where 'allele' refers to a unique post-editing sequence of a given target). Each allele was aligned to its original sequence, known because the corresponding gRNA sequence is part of the same amplicon, using the Needleman-Wunsch algorithm (21). Alleles were categorized as wild-type (*i.e.* unedited), a deletion, or an insertion.

Overall, targets were highly edited, with only 9.8% of UMIs corresponding to the wild-type allele. Of UMIs containing detectable mutations, 63.6% were deletions and 31.5% were insertions (Figure 2A). The remainder (4.9%) contained some combination of substitutions, insertions and deletions, and are excluded from all of our subsequent analyses. Deletions were dominated by small events; only 1.5% were >25 bp, although we note that deletions >150 bp are not captured by our assay (9,22). In contrast, although we believe that our assay should have been able to recover insertions up to ∼500 bp, the overwhelming majority of insertion events were of a single base pair.

### Repair patterns are reproducible but exhibit highly variable 'entropy' between targets

We sought to examine whether repair patterns for any given target were reproducible, as previously shown for a more limited set of templates (14). For each target, we calculated the frequency of each non-wild-type allele. For any given target, the distribution of frequencies for its alleles were highly reproducible in pairwise comparisons of the three replicates (median Pearson's $r = 0.91, 0.93, 0.93$, Figure 2B, left). Meanwhile, if we permute the alleles in one replicate on a target-by-target basis and repeat the pairwise comparison, these correlations are greatly reduced (median Pearson's $r =$ 0.20, Figure 2B, right).

Confirming the observations of (14), the diversity of mutations strongly varied from target to target. We calculated the Shannon entropy of mutational outcomes for any given target as $-\sum p_i * \log(p_i)$, where $p_i$ is the frequency of $i$th indel of that target (Figure 2C). Entropy values for any given target were highly reproducible between replicates (Figure 2D) and only modestly correlated with sampling depth (Supplementary Figure S1). Of note, some targets consistently exhibited particularly diverse mutational outcomes consequent to NHEJ—that is, high entropy (e.g. Figure 2E, where the most frequently observed mutation occurs in only 10.1% of mutated templates). Other targets were strongly biased towards a more limited set of mutational outcomes—that

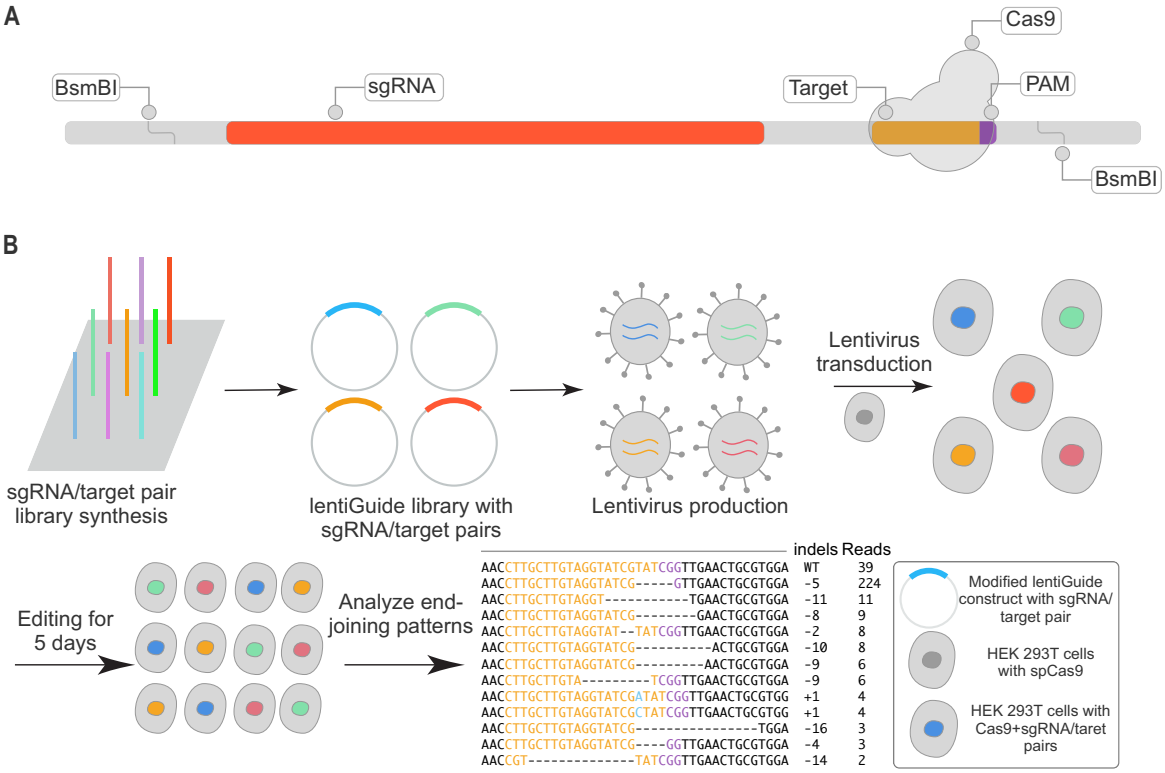


**Figure 1.** An assay for massively parallel profiling of the outcomes of CRISPR/Cas9-mediated double-stranded DNA break repair. (**A**) Schematic of library of 200 bp oligonucleotides encoding sgRNAs (red) targeting a large number of designed 20 bp spacers, with their matched target sequence encoded *in cis* (yellow: target; PAM: purple). In our primary experiment, 70 000 target sequences were designed and cloned. (**B**) After array-based synthesis and PCR amplification of the library, BsmBI restriction sites at either end were used for cloning into a modified lentiviral construct. The library was bottlenecked to 12,286 targets to facilitate greater coverage of independent NHEJ-mediated events corresponding to each target. Monoclonal HEK293T cells expressing Cas9 were transduced with packaged lentivirus. Cells were harvested at 5 days after transduction, and a region including both the spacer and the target was PCR amplified from genomic DNA for high-throughput sequencing. The sequences of mutated targets were aligned to their corresponding unmutated reference, assigned based on the spacer sequence (yellow: target; PAM: purple; green: inserted bases; dashes: deleted bases).

is, low entropy (e.g. Figure 2F, where the most frequently observed mutation occurs in 80.4% of mutated templates).

## Sequence context at the DSB site predicts the frequency of insertions

We next sought to investigate the determinants of insertions at the DSB, which were dominated by 1 bp events (Figure 3A). 84% of 1 bp insertions were predicted (and presumably templated) by the nucleotide immediately upstream of the cleavage site (*i.e.* the 17th nucleotide in target sequence; Figure 3B; Supplementary Figure S2) Although it might have been expected that NHEJ-mediated repair would be symmetric with respect to the site of a DSB, we do not observe templating from the immediately downstream (18th) nucleotide (Figure 3B). Similarly, of 2 bp insertions, a substantially greater than expected proportion (41%) were templated by the sequence immediately upstream of the DSB (i.e. inserted sequence identical to the 16th and 17th nucleotides of the target sequence; Figure 3C). The asymmetric templating of NHEJ-mediated insertions was also described in two other recent studies based on data from yeast and mice (13,23).

Because the ratio of insertions to deletions varied from target to target, we used kpLogo (24) to examine what local sequence features might shape this. We find that the presence of a T or A at the 17th bp of the target was associated with insertion events, while a G or C at this position was associated with deletion events (Figure 3D, left). Additional analyses showed a 'TG' dinucleotide flanking the cleavage site to be the most highly biased toward insertion (57% of events with that context are insertions), while a 'GA' dinucleotide flanking the cleavage site was the most highly biased towards deletion (17% of events with that context are insertions) (Figure 3D, right).

We split 2680 targets associated with both insertion and deletion outcomes into training ($n = 2000$) and test ($n = 680$) sets, and trained a linear regression model to predict the proportion of insertion events based on position-specific content of the hexamer centered on the DSB (single and dinucleotide $k$-mers; 104 binary features; Figure 3E). The model performs reasonably well (Pearson's $r = 0.70$).

Overall, these analyses confirm that local sequence around the DSB site plays an important role in shaping the outcome(s) of NHEJ-mediated errors. In particular, the asymmetry implied by the high rate of identity between 1 and 2 bp insertions and the nucleotides immediately upstream to the DSB, but not the nucleotides immediately downstream to the DSB (Figure 3B-C), suggests that not all CRISPR/Cas9-mediated cleavages are

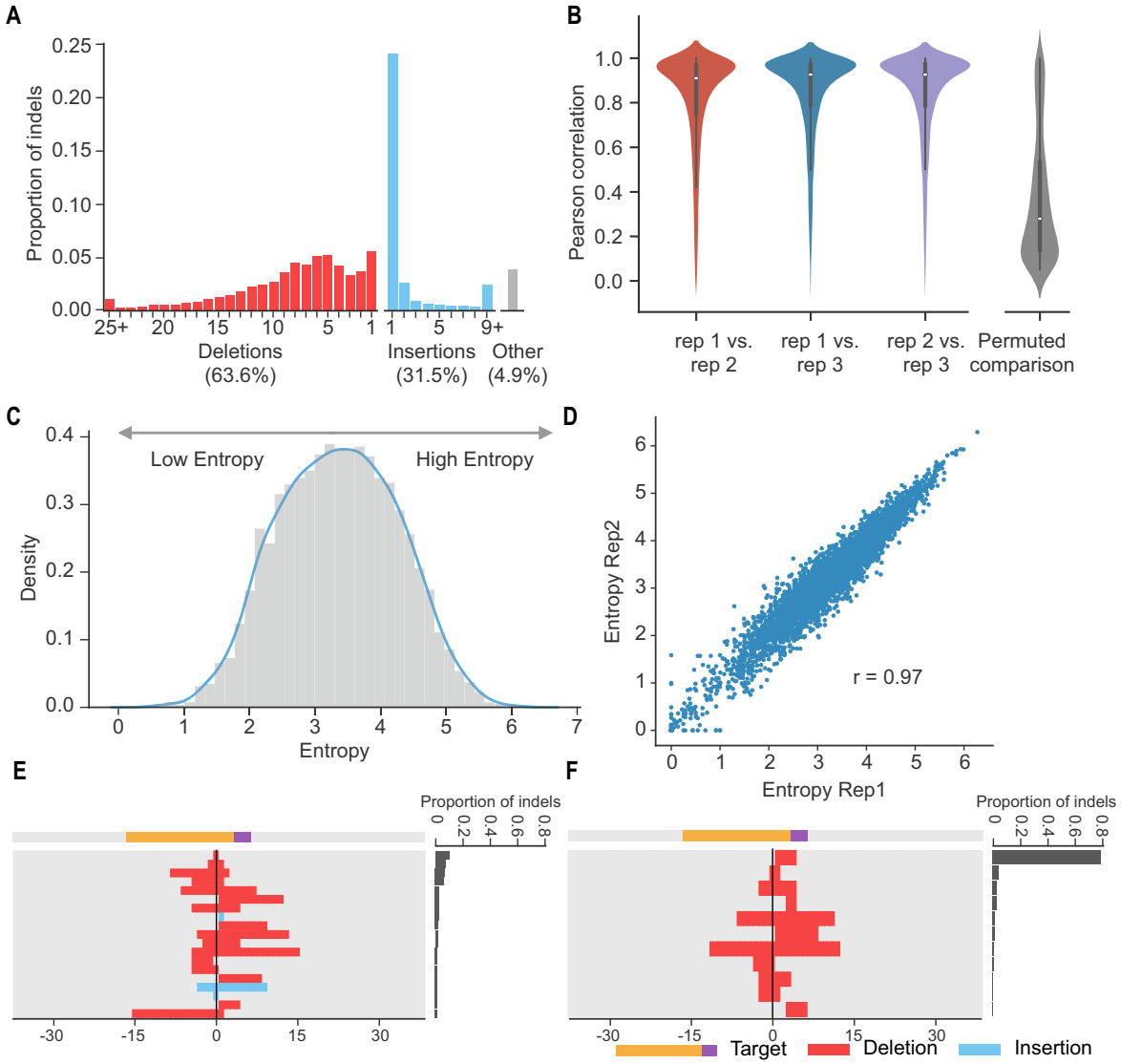


**Figure 2.** Mutation patterns resulting from DSB repair vary greatly between targets, but are highly reproducible for individual targets. (**A**) Overview of indel profiles. The histogram represents the indel rate per target, based on aggregated data from three replicates. The x-axis corresponds to the size of insertion or deletion events. Of all detectably mutated targets, 63.6% were deletions (red) and 31.5% were insertions (blue). The remainder (4.9%) contained some combination of substitutions, insertions and deletions, and are excluded from all subsequent analyses. (**B**) End-joining patterns were highly reproducible for the same target between replicates. Left: violin plot of distribution of correlation coefficients for pairwise comparison of individual targets between replicates. Right: Permuting the allele counts for each target in one replicate and repeating the pairwise comparison greatly reduces the observed correlations. (**C**) Entropy quantifies the diversity of NHEJ outcomes from individual targets. Targets were separated into low, medium and high entropy classes. (**D**) Estimated entropy for individual targets was highly reproducible between replicates (rep1 versus rep2 shown). (**E**, **F**) Example of targets with high and low entropy. High entropy targets had diverse outcomes at appreciable frequencies (E) while low entropy targets were dominated by a single outcome (F).

blunt-ended. Indeed, *in vitro* studies have shown that the non-complementary strand of the target can sometimes be cleaved by Cas9 at multiple sites upstream of the –3 bp position relative to the protospacer adjacent motif (PAM), while the complementary strand is cut only at that site, instances which would result in a 5′ overhang (25,26). The preponderance of 1 bp insertions templated by the 17th rather than 18th base could be explained by fill-in of this overhang followed by blunt-ended ligation (and similarly for the preponderance of 2 bp insertions that are templated by the 16th and 17th bases, rather than the 18th and 19th bases).

To summarize, we propose a model (Figure 3F) where: (i) some proportion of cleavages of the non-complementary strand by Cas9 occur upstream of the –3 bp PAM cleavage site, while cleavage of the complementary strand always occurs between the 17th and 18th positions, resulting in a 5′ overhang; (ii) 5′ overhangs are preferably repaired by gap-filling and ligation, resulting in the observed bias toward templating by the bases immediately upstream rather than downstream of the DSB; (iii) local sequence context biases the pattern of cleavage on the non-complementary strand, resulting in different frequencies of blunt versus 5′ overhangs for different targets, which in turn biases the ratio

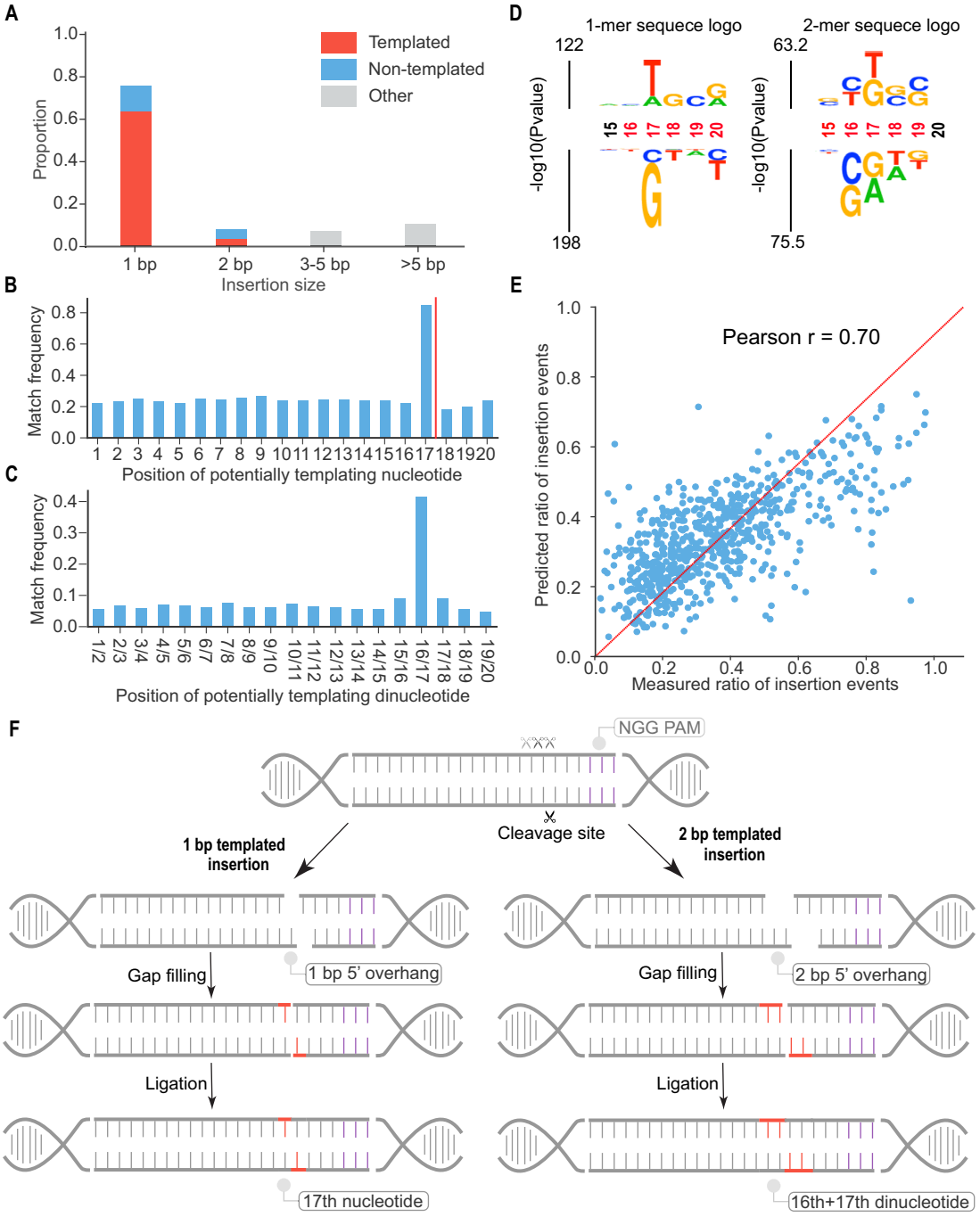


**Figure 3.** A model for asymmetric templating of NHEJ-mediated insertion events at sites of CRISPR/Cas9-mediated DSBs. (**A**) 75.3% of the insertions were 1 bp. Of 1 bp insertions, 85% appear to be templated. (**B**, **C**) Histogram of the number of 1 bp (B) or 2 bp (C) insertion events where the inserted base or dinucleotide is identical to the base at a specific position in the target. The canonical DSB site is between 17th and 18th of the target sequence (red line). The result suggests many 1 bp insertions are templated by the nucleotide at the 17th position but not the 18th position (B) and many 2 bp insertions are templated by dinucleotide at the 16th and 17th positions (C). (**D**) The immediate sequence context surrounding the DSB strongly biases the proportion of NHEJ-mediated outcomes that result in insertions vs. deletions. The 1-mer sequence logo (left) shows that the presence of a 'T' and 'A' at the 17th position increased the ratio of insertions. The 2-mer sequence logo (right) shows that the presence of a 'TG' dinucleotide at the 17th/18th position increased the ratio of insertions, while a 'CG' dinucleotide at the 16th/17th position, or a 'GA' dinucleotide at the 17th/18th position, decreased the ratio of insertions. Significant positions are colored in red. (**E**) A regression model using the nucleotide content of a 6 bp window centered on the DSB site predicted the ratio of insertion-to-deletion events. (**F**) A model for how insertions at CRISPR/Cas9-mediated DSBs are asymmetrically biased by local sequence context. Local sequence context biases the pattern of cleavage of the non-complementary strand to the sgRNA, resulting in different frequencies of blunt vs. 5′ overhangs for different targets. This in turn biases the ratio of insertions vs. deletions, as 5′ overhangs are preferably repaired by gap-filling (red) and ligation, resulting in the observed preponderance of 1 or 2 bp templated insertions (red).

of insertions versus deletions. A similar model was recently proposed by Lemos *et al.* based on asymmetric templating of NHEJ-mediated insertions observed in yeast (23).

### Extensive use of microhomology in NHEJ-mediated deletions

We next examined patterns of deletion. Microhomology (MH) refers to the use of short regions of identical sequence (1–16 bp) that can mediate the alignment of broken ends (Figure 4A) and is relevant to both c-NHEJ and alt-NHEJ/MMEJ (4,27–29). Here, a deletion event is considered to be MH-mediated if the sequence at the 3′ of a rejoined end is identical to the 3′ end of the deleted sequence, and the size of the MH tract refers to the length of that identical sequence. By that definition, we found that over 75% of deletion events in our dataset are MH-mediated. The length of MH tracts ranged from 1 to 10 bp. Nearly all MH-mediated events (94.6%) involved relatively short tracts of microhomology, i.e. 1–4 bp. Longer MH tracts were observed more rarely (Supplementary Figure S3A), probably simply due to the relative paucity of opportunities in our set of target sequences.

The frequencies of tracts of various lengths consistent with MH usage were substantially higher than background expectation for all lengths except 1 bp, with that frequency increasing as a function of tract length (Figure 4B). We further investigated the relevance of 1 bp MH by comparing the proportion of 1 bp deletion events in targets with identical versus non-identical nucleotides immediately spanning the cleavage site. We observe a 3-fold greater proportion of 1 bp deletion events when those nucleotides are identical than when they are not (Supplementary Figure S3B), suggesting that 1 bp MH may play a role in aligning, stabilizing and rejoining the broken ends.

The lengths of MH versus non-MH mediated deletions exhibited distinct distributions (Figure 4C-D). In particular, the distribution of deletion sizes for MH-mediated events peaks at both 1 and 5–6 bp, while an equivalent distribution for non-MH-mediated deletions peaks at both 1–2 and 8 bp. The frequency of longer deletions exhibits an exponential decay for both MH and non-MH mediated events. To investigate this further, we jointly analyzed the frequency of start and end points for deletion events, relative to the position of the canonical cleavage site (Figure 4E and F). Both MH and non-MH mediated deletions exhibited a preference for 'unidirectional' events, i.e. either the start or end point is immediately adjacent to the cleavage site, rather than the deletion spanning the cleavage site.

What explains the excess of deletion events of specific lengths? For MH-mediated events, the excess of 1 bp deletions may simply be attributable to the aforedescribed instances of identical nucleotides spanning the cleavage site (Supplementary Figure S3B). However, the excess of 5–6 bp MH-mediated events is clearly driven by events in the downstream direction (Figure 4E), i.e. deletions between the DSB and the PAM. A potential explanation is that the predilection of PAM-like sequences near the DSB for deletion events (i.e. a G nucleotide at the 17th position or a CG dinucleotide at the 16th/17th position; Figure 3D), coupled with the consistent presence of the CGG PAM sequence at the 21st–23rd position, results in an excess of deletions mediated by CG (5 bp deletion) or G (5-6 bp deletion) microhomology (Figure 4G). Further work would be required to confirm this, as there may be other explanations.

For non-MH-mediated events, the excess of 8 bp events might be explained by the observation that in the dsDNA-sgRNA-Cas9 complex, the region 1–8 bp downstream of the cleavage site is occupied by Cas9, even after cleavage (26,30). Thus, the enrichment of non-MH deletions 8 bp from the cleavage site could simply correspond to the nearest position lacking Cas9 protection from endonucleases during repair (Figure 4H).

### Generating predictable mutations by programming microhomology tracts

Since MH is widely used in deletion events, we reasoned that we could program a library of targets to generate predictable mutations by introducing MH proximal to the cleavage site. With the same basic experimental scheme (Figure 1), we tested a library of 1000 targets and corresponding guides containing MH tracts of three different lengths (2, 4 or 6 bp) matching the sequence immediately upstream of the expected DSB site, and positioned 6 bp downstream of the cleavage site (Figure 5A-B). The resulting data were processed and analyzed similarly to the previous experiment.

Intentionally programming MH tracts resulted in a high proportion of events corresponding to the expected deletions (8, 10 and 12 bp deletions for 2, 4 and 6 bp MH tracts, respectively; Figure 5B and C, Supplementary Figure S4). We also observe that the ratio of the programmed deletion increases as a function of length of the MH tract (Figure 5C). However, despite the greater predictability of which MH-mediated outcome would occur, the relative proportion of MH-mediated deletions increased only slightly from 76% to 82% (Figure 5D). Furthermore, we did not observe an excess of 'imperfect' MH-mediated events, e.g. an excess of 11 or 13 bp deletions in targets for which a 12 bp deletion was expected (Supplementary Figure S4). Nonetheless, the results show how targets that would result in diverse editing outcomes can be strongly biased towards a specific outcome by the presence of MH tracts (Figure 5E and F).

### A machine learning model to predict editing patterns

The above results above suggest that the NHEJ-mediated repair outcomes for any given target sequence are both reproducible and dependent on sequence context. Accordingly, we next sought to train a machine learning model to predict these outcomes and their relative frequencies. We began by filtering out target sequences that were either poorly reproducible (low correlation between replicates, mainly due to low UMI counts; Supplementary Figure S5A) or poorly edited, resulting in a dataset of ∼1 million UMIs representing 4790 target sequences. On average, each target in this subset of the data used for modeling was represented by 204 UMIs and 28 alleles.

Because larger events are rare in our data, we focused on predicting deletion events <30 bp in length, as well as all possible 1–2 bp insertion events at the DSB. Across all

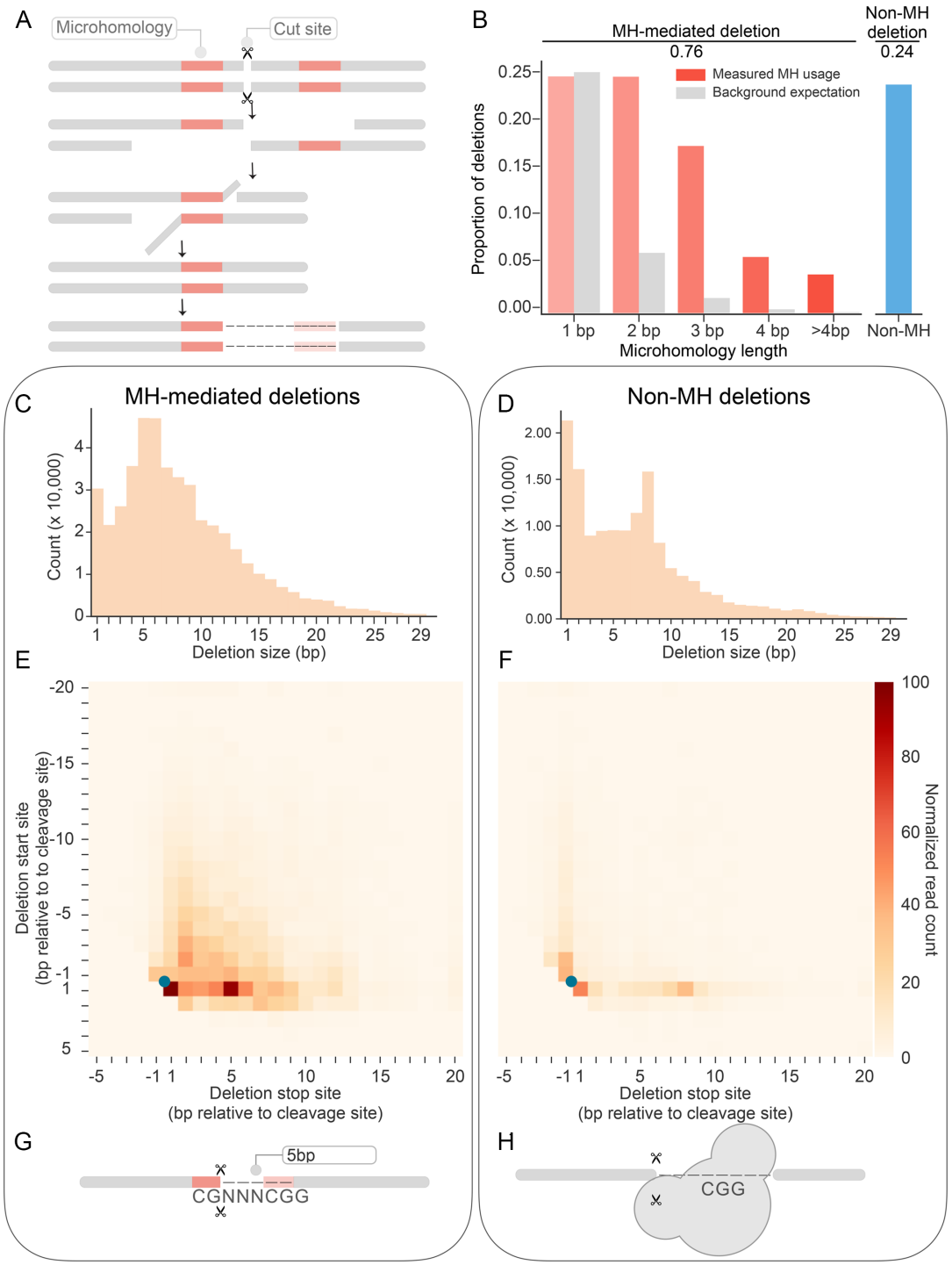


**Figure 4.** Extensive use of microhomology in NHEJ-mediated deletion events. (**A**) Schematic of microhomology (MH) usage in end-joining repair. Tracts of MH (red) in the vicinity of the DSB are used to align the broken ends. The unannealed overhang is cleaved by endonuclease and the gap filled by polymerase. Here, a deletion event is defined as MH-mediated deletion if the sequence at the 3′ of a rejoined end (red, left) is identical to the 3′ end of the deleted sequence (red, right). The size of the MH tract refers to the length of that identical sequence. (**B**) Length distribution of MH tracts in observed MH-mediated events. With the exception of 1 bp deletions, all MH tract lengths occured at substantially greater than expected frequencies. (**C**, **D**) Distribution of deletion sizes of MH-mediated (C) and non-MH (D) events. (**E**, **F**) Heatmap of showing frequency of start/stop sites of MH-mediated (E) and non-MH (F) deletion events. The Y and X axes correspond to the start and stop sites of deletion events, respectively, with positions shown relative to the canonical DSB site (blue dot). Both MH-mediated and non-MH deletions were primarily 'unidirectional' relative to the DSB site, rather than spanning it. (**G**) Schematic of potential explanation for the observed excess of 5–6 bp MH-mediated deletions. PAM-like sequences near the DSB are biased towards deletion events. G: Microhomology between a G at the 17th position or a CG at the 16th/17th position with corresponding sequences in the PAM result in an excess of 5–6 bp deletions. (**H**) Schematic of potential explanation for the observed excess of 8 bp non-MH deletions. In the dsDNA-sgRNA-Cas9 complex, the region 1–8 bp downstream of the cleavage site is occupied by Cas9. The enrichment of non-MH deletions 8 bp from the cleavage site could simply correspond to the nearest position lacking Cas9 protection from endonucleases during repair.

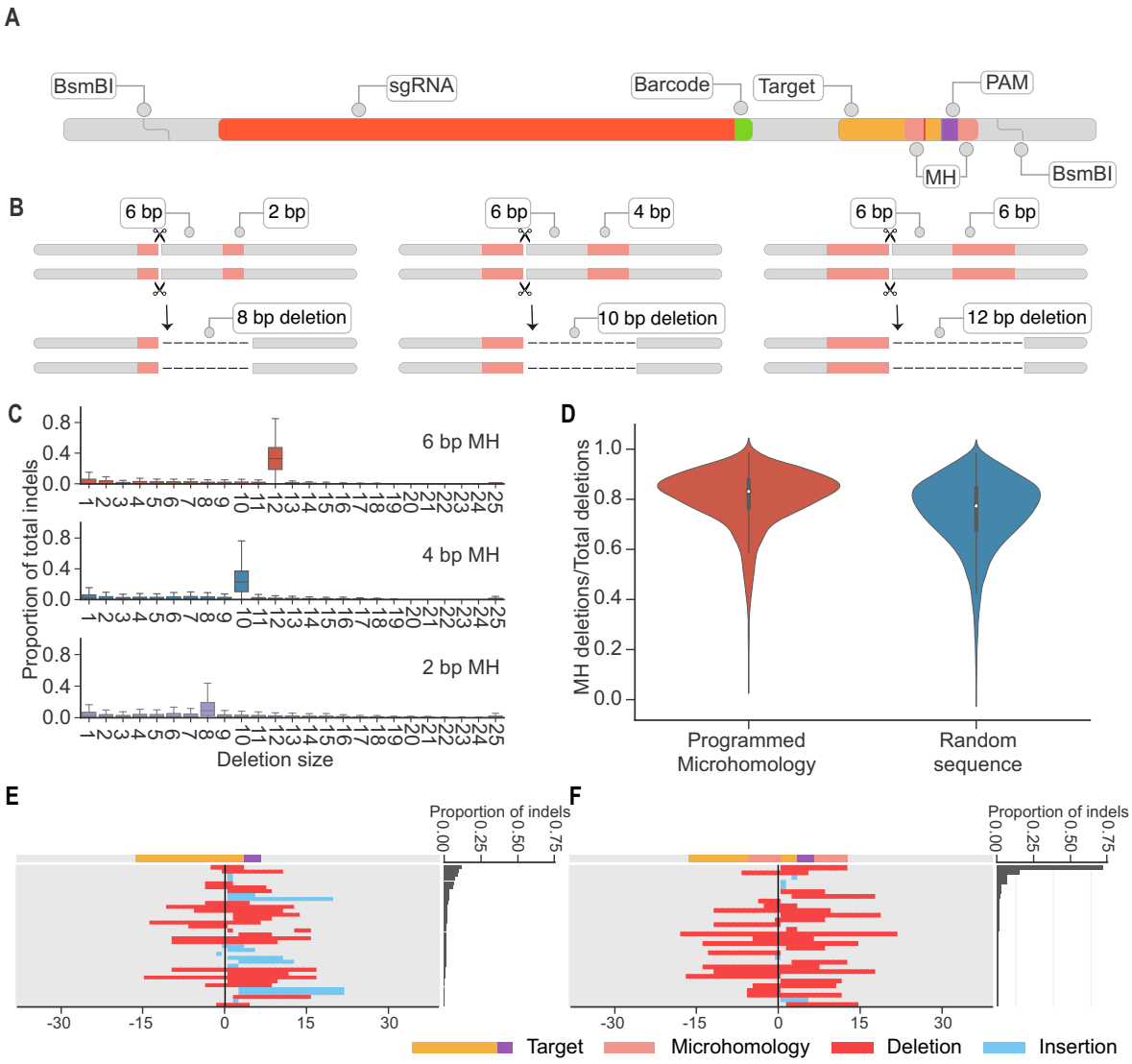


**Figure 5.** Programming microhomology tracts into targets increases predictability of repair outcomes. (**A**, **B**) Schematic of programmed MH tract designs, which starts immediately downstream of PAM sequence (purple), and expected deletion sizes. The distance between the regions of MH (pink) was consistently 6 bp, while the MH tracts were 2, 4 or 6 bp, such that the expected deletion sizes were 8, 10 and 12 bp, respectively. (**C**) Distribution of observed deletion sizes for 1000 targets with programmed MH tracts of various lengths, based on a single replicate/experiment. Each boxplot summarizes the ratio of certain deletion for each target. The box represents 25th percentile, 50th percentile and 75th percentile and whiskers represents 1.5× of the inter-quartile range (IQR). We observe a strong bias towards deletions of the expected lengths, with the proportion increasing for longer MH tracts (median 0.080, 0.209, and 0.318 for 2 bp MH, 4 bp MH and 6 bp MH, respectively). (**D**) MH usage in sequences with (left) or without (right) programmed MH. Despite the strong bias of toward intended deletions when MH occurred, the proportion of MH events only slightly increased from 76% to 82%. (**E**, **F**) Example of a sequence that shows diverse editing outcomes (E). However, when a 6 bp MH tract is introduced onto this sequence backbone, the programmed 12 bp deletion comprises nearly 75% of the editing outcomes (F).

targets, we identified 557 'event classes'. The vast majority of CRISPR/NHEJ-mediated indels arising from any given target sequence should fall into one of these 557 event classes. We therefore framed our machine learning task as one of predicting, for an arbitrary target sequence, the relative frequency of CRISPR/NHEJ-mediated indels falling into each of these 557 event classes. These included 536 deletions (defined solely by their start/end points), all four possible single nucleotide insertions, all 16 possible dinucleotide insertions, and finally, a single event class for insertions greater than 2 bp in length. Of note, the 536 deletion event classes comprise almost all of the 550 possible combinations of start/end positions, with the constraints that deletions must be <30 bp and overlap with the –3/+2 window around the cleavage site. The 14 potential deletions that satisfy these constraints but were not observed in the modeling dataset were mainly large deletions.

We also defined 3033 binary features to characterize the target sequence for which repair outcomes are being predicted. These are (i) Sequence features: 384 binary features corresponding to one-hot encoded sequence, including 80 for single nucleotide content (4 nucleotides × 20 positions)

and 304 for dinucleotide content (16 dinucleotides × 19 positions); (ii) microhomology features: 2649 binary features corresponding to MH tracts; specifically, for each of the possible deletion event class, we defined 2–5 binary features (depending on the size of deletion) corresponding to the length of the MH tract ([0–4 bp × 519] + [0–3 bp × 7] + [0–2 bp × 6] + [0–1 bp × 4] deletion event classes = total of 2649 binary features) (Figure 6A).

We split the 4790 target sequences in our modeling dataset into subsets of 3900 (for training), 450 (for validation) and 440 (for testing).

Our model consists of three components: (i) predicting the ratio of insertions to deletions; (ii) predicting the distribution of 536 classes of deletion events; (3) predicting the distribution of 21 classes of insertion events. The overall distribution of predicted outcomes is then determined by intersecting these three components. Of note, because we define deletion classes using the deletion start site and deletion length, two or more classes can effectively represent identical outcomes due to microhomology, but in a sequence specific manner (Supplementary Figure S5C). To address this, while evaluating performance, we simply collapsed identical outcomes.

We trained predictors for each of the three components independently using logistic regression with a varied number of features and varied strength of L1 or L2 penalties. All of the models were trained on the training set using cross-entropy loss and evaluated on the validation set using the mean squared error (MSE). For the indel ratio predictor, we predicted that microhomology features and sequence context would both be important for prediction. However, including microhomology features did not improve the performance compared to using one-hot encoded sequence alone (MSE = 0.0203 and 0.0201 respectively, Supplementary Figure S5D). For the insertion predictor, it has been shown above that most insertions were templated by the sequence upstream of the cleavage site. We reasoned that sequence context around the cleavage site (±3 bp) should be sufficient to predict insertions. Consistent with this, including the full 20 bp target worsened performance, increasing MSE from 0.00666 to 0.00711 (Supplementary Figure S5E). For the deletion predictor, we compared the performance of models using sequence features only, microhomology features only or all features. The model with all-features performed the best (MSE of 0.000204, as compared with 0.000271 for sequence-only and 0.000208 for microhomology-only) (Supplementary Figure S5F). However, the nearly identical performance of the all-features versus microhomology-only models for predicting deletions is notable. We used the best performing model for each component to build a predictor for the overall distribution of outcomes (Figure 6A).

Applying this model to the test set of 440 target sequences, which had been entirely held out from the training and validation steps, we compared the observed versus predicted frequencies of indels falling into various 'event classes'. Observations and predictions were well matched for most targets, with a MSE of 0.000172 (Figure 6B). As a baseline, we also generated a set of predictions based simply on the aggregate frequencies of event classes in the training and validation datasets; as expected, these predictions performed more poorly (MSE of 0.000359; Figure 6B), confirming the improvement conferred by the model. Poorly predicted targets tended to be those with relatively shallower sampling of editing events, *i.e.* where our observed frequencies are noisier (Supplementary Figure S5B).

### Comparison to other models

While this manuscript was in preparation, several similar studies were published (31–33). Together with this manuscript, all four studies profiled repair outcomes of Cas9-induced DSBs at large numbers of endogenous (33) or synthetic (31,32) targets (Supplementary Table S1). The primary conclusions, *e.g.* that sequence context around the DSB, together with MH, are the major determinants of repair outcomes, are consistent between these studies as well as with earlier studies (14,15). In addition, Shen *et al.* built inDelphi and Allen *et al.* built ForeCasT, as models that predict NHEJ repair outcomes, analogous to the Lindel model described here. The ForeCasT model predicts deletions similarly to Lindel, as well as all possible 1–2 bp insertions (Supplementary Table S1). In contrast, the inDelphi model predicts the frequency of three classes of indels independently, using a neural network model for MH-mediated deletions (90 classes) and non-MH deletions (59 classes corresponding to 1–59 bp deletions, without prediction of location), and *k*-nearest neighbors model for 1 bp insertions (four classes) (Supplementary Table S1).

We compared the three models by measuring the MSE on 440 targets in our test set as well as 4298 targets in ForeCasT test set (of note, the predicted probabilities of event classes not predicted by inDelphi were simply set to 0). Our model performed the best on our test set (MSE = 0.000172, 0.000225, 0.000212 for Lindel, ForeCasT and inDelphi, respectively), while ForeCasT performed the best on its test set (MSE = 0.000173, 0.000152, 0.000182, for Lindel, ForeCasT and inDelphi, respectively). We then combined our training set (3900 sequences) with ForeCast training set (10 725 sequences), resulting in a total of 14 625 sequences. We trained on this aggregated training set using the Lindel modeling approach, which resulted in the best overall performance (MSE = 0.000165 on our test set and 0.000125 on ForeCasT test set; Figure 6C, D). This final Lindel model is more accurate at predicting the ratio of insertions to deletions than ForeCasT (MSE = 0.01 and 0.02 for final Lindel model and ForeCasT, respectively; Supplementary Figure S6A,B). We further investigated the source of the errors for these models. Despite the fact that they implement different modeling approaches, both Lindel and ForeCasT's mispredictions primarily lie with small deletions and 1 bp insertions (Supplementary Figure S6C–F).

As a common use of CRISPR/Cas9 in conjunction with NHEJ is to introduce frameshifting mutations, we also assessed the observed versus predicted ratios of frameshifting indels for each of the 440 targets in our test set and 4298 targets in ForeCasT test set, and found them to be reasonably correlated (Pearson's $r = 0.707$, MSE = 0.0122 on our test set and Pearson's $r = 0.676$, MSE = 0.0098 on ForeCasT test set; Figure 6E; Supplementary Figure S6G). This result compares very favorably with the predictions of a previously published tool that we tested on this same task (Pearson's *r*

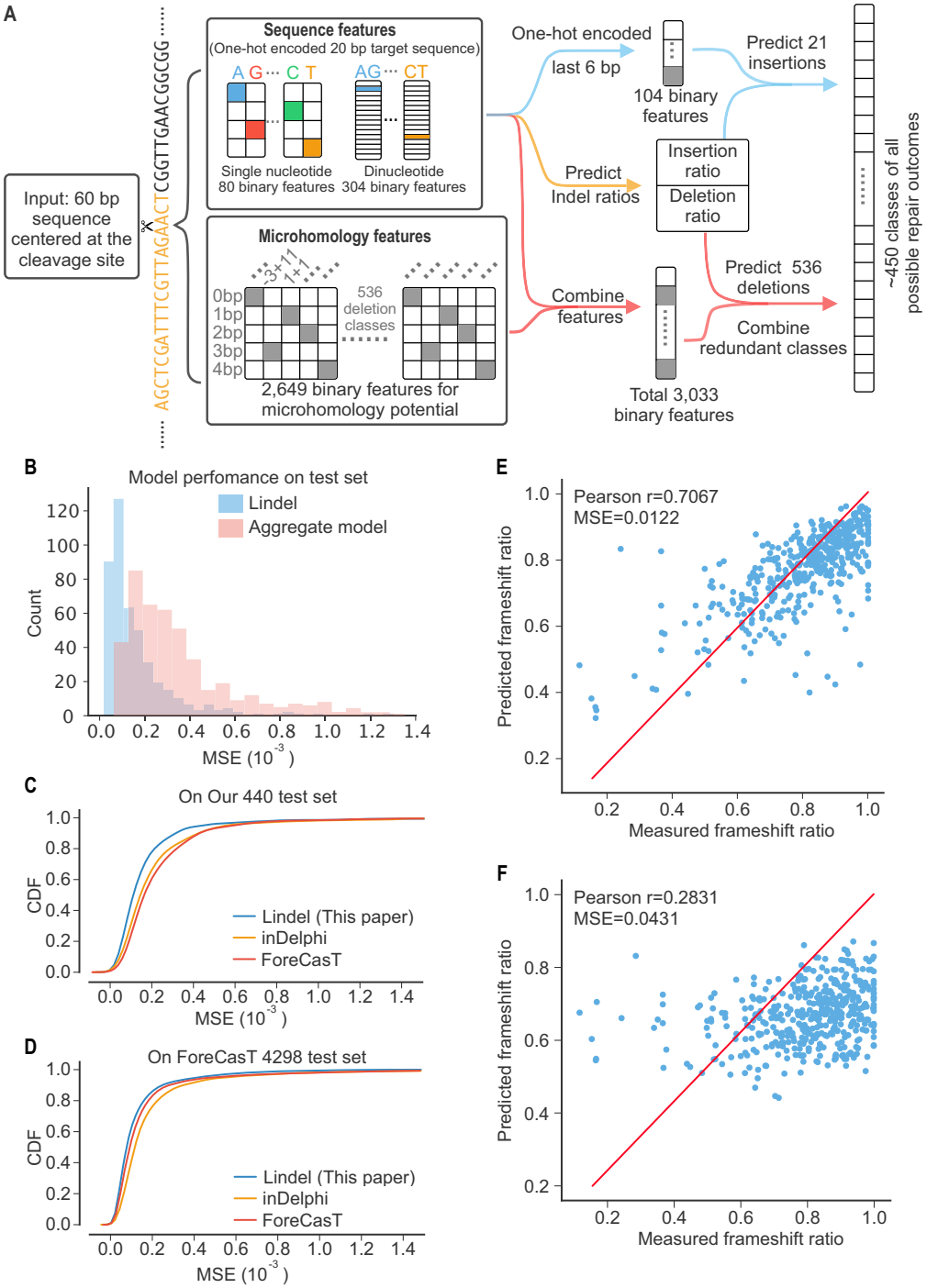


**Figure 6.** End joining patterns are accurately predicted by Lindel. (**A**) Schematic of machine learning framework for Lindel. A 60 bp sequence (±30 bp around the cleavage site) is used as the input to the model. A total of 3033 binary features—2649 corresponding to MH potential and 384 to one-hot encoded mono- and di-nucleotide content of the 20 bp target—are extracted. One-hot encoded sequence features were used to the predict overall ratio of insertion to deletion events. One-hot encoded sequence corresponding to the last 6 bp of the target sequence were used to predict 21 insertion classes. Both sequence features and microhomology features were used to predict deletion classes. Probabilities of redundant deletion classes were combined on a sequence specific manner. (**B**) Performance of Lindel on the test dataset. The distribution of MSE values for the 440 test targets is shown (blue). Poorly predicted targets largely correspond to those that were poorly sampled (see Supplementary Figure S5B). As a baseline to illustrate the improvement conferred by Lindel, we show a similar distribution for the aggregate model, in which the predicted frequencies of 557 indel classes are simply taken from the aggregate frequency at which each is observed in the training and validation datasets (red). (**C**, **D**) The performance of Lindel trained using training data aggregated from this study and the ForeCasT study (3900 sequences from this study and 10 725 from (32)). A Lindel model that was trained on the aggregated training datasets performed best on both the test sets from this study (440 sequences) and the ForeCasT study (4298 sequences). (**E**, **F**) Lindel (E) compared favorably to Microhomology Predictor (34) (F) in predicting the ratio of frameshifting mutations for each of the 440 targets in the test set.

= 0.283, MSE = 0.0431 on our test set and Pearson's $r$ = 0.455, MSE = 0.0315 on the ForeCasT test set; Figure 6F; Supplementary Figure S6H) (34).

## DISCUSSION

In summary, we developed an assay to systematically profile the diversity and relative frequencies of mutational events resulting from CRISPR/Cas9-mediated cleavage and NHEJ-mediated DSB repair of thousands of synthetic sequences. In applying this assay and analyzing the editing outcomes associated with 6,872 target sequences, we confirm that CRISPR/NHEJ-mediated repair outcomes for any given target sequence are reproducible, predictable, and largely shaped by the sequence context around the cleavage site (14,15).

Our results also provide further insights into NHEJ-mediated repair of CRISPR/Cas9-mediated DSBs in human cell lines. First, we observe that insertion events are dominated by 1–2 bp insertions templated by the sequence immediately upstream of the cleavage site. Together with *in vitro* data from the literature (25,26), the data supports a model in which the sequence context around the DSB biases the extent to which cleavages are blunt-ended vs. include a 1–2 bp 5′ overhang. Such 5′ overhangs are repaired by gap-filling and ligation, resulting in asymmetrically templated 1–2 bp insertions. Second, we observe extensive usage of 1–4 bp microhomology in mediating deletion events, and furthermore show that repair outcomes can be strongly biased towards predictable outcomes by intentionally introducing MH tracts at specific distances from the DSB. Notably, however, the introduction of MH tracts did not substantially increase the proportion of MH-mediated events. Third, both MH and non-MH-mediated deletions were overwhelmingly unidirectional (i.e. extending either upstream or downstream from the DSB, rather than spanning it).

Our assay has two main limitations. First, because of the locations of the PCR primer sites, we are only able to recover small deletions, and may be missing the rare, large deletion events that we and others have described (9,22). Greater knowledge of the frequency and determinants of large events is necessary to enable their prediction. Second, the lentiviral-based assay that we used fails to capture the influence of chromatin state on editing efficiency and repair outcomes. Lentivirus integrates to diverse locations across the genome, such that we are effectively observing an average, but integrations are biased towards open chromatin (35,36). As such, the patterns that we observe and model may be biased to this compartment. Furthermore, we are varying the sgRNA spacer sequence within the context of constant neighboring sequence, i.e. the lentiviral backbone. To the extent that this sequence biases nucleosome positioning, and that nucleosome positioning in turn influences Cas9 binding and cleavage (37,38), the patterns that we observe and model may be additionally biased. Additional systematic profiling of repair outcomes, in different compartments (e.g. open versus closed chromatin) and in different sequence contexts, will be necessary to understand the magnitude and nature of each of these potential sources of chromatin-mediated bias (33,39)

In addition to insights into into NHEJ-mediated repair of CRISPR/Cas9-mediated DSBs, our study also provides a new tool for sgRNA design for diverse goals. First, an important application of CRISPR/Cas9 is to achieve gene knockouts, a goal that depends on the efficient introduction of frameshifting indels. Dual cleavage with a variety of different nuclease systems has previously been shown to be an effective strategy for introducing frameshifting mutations (40–44). Our model's accuracy for predicting which sgRNAs/targets are likely to result in a high proportion of frameshifting indels will improve the viability of single cleavage with CRISPR/Cas9 for this same goal. Of note, our approach will not obviate the need for downstream validation to identify clones bearing the intended mutation, although it may reduce the number of clones that need to be screened. Second, for applications focused on mutation correction (e.g. using CRISPR/NHEJ to correct pathogenic mutations), the model may be useful for identifying sgRNAs/targets for which the desired outcome is predicted to occur at a high or sufficient frequency. Third, we and others have recently repurposed CRISPR/Cas9 as a tool for lineage tracing and/or molecular recording (11–13). For some goals (e.g. lineage tracing), the identification of 'high entropy' targets may critically enable the diversity necessary to uniquely label millions or billions of cells. For other goals (e.g. molecular recording), the design of 'low entropy' targets may facilitate predictable sequential editing. More generally, a deeper understanding of CRISPR/NHEJ-mediated mutations will strengthen our ability to precisely orchestrate not only the locations but also the outcomes of genome editing.

## DATA AVAILABILITY

Scripts and other software are available from our GitHub repository: https://github.com/shendurelab/Lindel.

## SUPPLEMENTARY DATA

Supplementary Data are available at NAR Online.

## ACKNOWLEDGEMENTS

We thank members of the Shendure lab and Dezhong Deng for helpful comments and discussion.
*Authors' Contributions*: W.C. designed and performed experiments and led analysis. A.M. contributed to design of the sgRNAs and data analysis. J.M.S. contributed to training and model selection. Y.Y. constructed the 293T Cas9 cell line and contributed to data analysis. V.A. contributed to data analysis. M.H integrated Lindel into CRISPOR. W.C. wrote the first draft of manuscript. J.S. edited the manuscript. All authors contributed to the final manuscript. This work was supervised by W.S.N and J.S.

## FUNDING

Paul G. Allen Frontiers Foundation (Allen Discovery Center grant to J.S.); NIH [R01HG009136 and UM1HG009408 to J.S. and 1K99HG010152-01 to A.M.]. M.H. was funded by National Human Genome Research Institute [4U41HG002371]; St. Baldricks Foundation [427053];

Chan-Zuckerberg Initiative [SVCF 2018-182809]. Funding for open access charge: NIH [R01HG009136].
*Conflict of interest statement.* None declared.

## REFERENCES

1. Hsu,P.D., Lander,E.S. and Zhang,F. (2014) Development and applications of CRISPR-Cas9 for genome engineering. *Cell*, **157**, 1262–1278.
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. Bétermier,M., Bertrand,P. and Lopez,B.S. (2014) Is non-homologous end-joining really an inherently error-prone process? *PLoS Genet.*, **10**, e1004086.
4. 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.
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. Jinek,M., East,A., Cheng,A., Lin,S., Ma,E. and Doudna,J. (2013) RNA-programmed genome editing in human cells. *Elife*, **2**, e00471.
8. Wang,T., Birsoy,K., Hughes,N.W., Krupczak,K.M., Post,Y., Wei,J.J., Lander,E.S. and Sabatini,D.M. (2015) Identification and characterization of essential genes in the human genome. *Science*, **350**, 1096–1101.
9. Gasperini,M., Findlay,G.M., McKenna,A., Milbank,J.H., Lee,C., Zhang,M.D., Cusanovich,D.A. and Shendure,J. (2017) CRISPR/Cas9-Mediated scanning for regulatory elements required for HPRT1 expression via thousands of large, programmed genomic deletions. *Am. J. Hum. Genet.*, **101**, 192–205.
10. Klein,J.C., Chen,W., Gasperini,M. and Shendure,J. (2018) Identifying novel enhancer elements with CRISPR-Based screens. *ACS Chem. Biol.*, **13**, 326–332.
11. McKenna,A., Findlay,G.M., Gagnon,J.A., Horwitz,M.S., Schier,A.F. and Shendure,J. (2016) Whole-organism lineage tracing by combinatorial and cumulative genome editing. *Science*, **353**, aaf7907.
12. Frieda,K.L., Linton,J.M., Hormoz,S., Choi,J., Chow,K.-H.K., Singer,Z.S., Budde,M.W., Elowitz,M.B. and Cai,L. (2017) Synthetic recording and in situ readout of lineage information in single cells. *Nature*, **541**, 107–111.
13. Kalhor,R., Kalhor,K., Mejia,L., Leeper,K., Graveline,A., Mali,P. and Church,G.M. (2018) Developmental barcoding of whole mouse via homing CRISPR. *Science*, **361**, eaat9804.
14. 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.
15. Aubrey,B.J., Kelly,G.L., Kueh,A.J., Brennan,M.S., O'Connor,L., Milla,L., Wilcox,S., Tai,L., Strasser,A. and Herold,M.J. (2015) An inducible lentiviral guide RNA platform enables the identification of tumor-essential genes and tumor-promoting mutations in vivo. *Cell Rep.*, **10**, 1422–1432.
16. Chang,H.H.Y., Watanabe,G., Gerodimos,C.A., Ochi,T., Blundell,T.L., Jackson,S.P. and Lieber,M.R. (2016) Different DNA end configurations dictate which NHEJ components are most important for joining efficiency. *J. Biol. Chem.*, **291**, 24377–24389.
17. 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.
18. Haeussler,M., Schönig,K., Eckert,H., Eschstruth,A., Mianné,J., Renaud,J.-B., Schneider-Maunoury,S., Shkumatava,A., Teboul,L., Kent,J. *et al.* (2016) Evaluation of off-target and on-target scoring algorithms and integration into the guide RNA selection tool CRISPOR. *Genome Biol.*, **17**, 148.
19. Sack,L.M., Davoli,T., Xu,Q., Li,M.Z. and Elledge,S.J. (2016) Sources of error in mammalian genetic screens. *G3.*, **6**, 2781–2790.
20. Hill,A.J., McFaline-Figueroa,J.L., Starita,L.M., Gasperini,M.J., Matreyek,K.A., Packer,J., Jackson,D., Shendure,J. and Trapnell,C. (2018) On the design of CRISPR-based single-cell molecular screens. *Nat. Methods*, **15**, 271–274.
21. Needleman,S.B. and Wunsch,C.D. (1970) A general method applicable to the search for similarities in the amino acid sequence of two proteins. *J. Mol. Biol.*, **48**, 443–453.
22. Kosicki,M., Tomberg,K. and Bradley,A. (2018) Repair of double-strand breaks induced by CRISPR–Cas9 leads to large deletions and complex rearrangements. *Nat. Biotechnol.*, **36**, 765–771.
23. 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.
24. Wu,X. and Bartel,D.P. (2017) kpLogo: positional k-mer analysis reveals hidden specificity in biological sequences. *Nucleic Acids Res.*, **45**, W534–W538.
25. 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.
26. 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.
27. 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.
28. Pannunzio,N.R., Li,S., Watanabe,G. and Lieber,M.R. (2014) Non-homologous end joining often uses microhomology: implications for alternative end joining. *DNA Repair*, **17**, 74–80.
29. Conlin,M.P., Reid,D.A., Small,G.W., Chang,H.H., Watanabe,G., Lieber,M.R., Ramsden,D.A. and Rothenberg,E. (2017) DNA Ligase IV guides End-Processing choice during nonhomologous end joining. *Cell Rep.*, **20**, 2810–2819.
30. 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.
31. 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.
32. Allen,F., Crepaldi,L., Alsinet,C., Strong,A.J., Kleshchevnikov,V., De Angeli,P., Páleníková,P., Khodak,A., Kiselev,V., Kosicki,M. *et al.* (2019) Predicting the mutations generated by repair of Cas9-induced double-strand breaks. *Nat. Biotechnol.*, **37**, 64–72.
33. Chakrabarti,A.M., Henser-Brownhill,T., Monserrat,J., Poetsch,A.R., Luscombe,N.M. and Scaffidi,P. (2019) Target-Specific precision of CRISPR-Mediated genome editing. *Mol. Cell*, **73**, 699–713.
34. 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.
35. Yang,S.-H., Cheng,P.-H., Sullivan,R.T., Thomas,J.W. and Chan,A.W.S. (2008) Lentiviral integration preferences in transgenic mice. *Genesis*, **46**, 711–718.
36. Ustek,D., Sirma,S., Gumus,E., Arikan,M., Cakiris,A., Abaci,N., Mathew,J., Emrence,Z., Azakli,H., Cosan,F. *et al.* (2012) A genome-wide analysis of lentivector integration sites using targeted sequence capture and next generation sequencing technology. *Infect. Genet. Evol.*, **12**, 1349–1354.
37. Horlbeck,M.A., Witkowsky,L.B., Guglielmi,B., Replogle,J.M., Gilbert,L.A., Villalta,J.E., Torigoe,S.E., Tjian,R. and Weissman,J.S. (2016) Nucleosomes impede Cas9 access to DNA in vivo and in vitro. *Elife*, **5**, e12677.
38. Chen,X., Rinsma,M., Janssen,J.M., Liu,J., Maggio,I. and Gonçalves,M.A.F.V. (2016) Probing the impact of chromatin conformation on genome editing tools. *Nucleic Acids Res.*, **44**, 6482–6492.
39. Kallimasioti-Pazi,E.M., Chathoth,K.T., Taylor,G.C., Meynert,A., Ballinger,T., Kelder,M.J.E., Lalevée,S., Sanli,I., Feil,R. and Wood,A.J. (2018) Heterochromatin delays CRISPR-Cas9 mutagenesis but does not influence the outcome of mutagenic DNA repair. *PLOS Biol.*, **16**, e2005595.

40. Ran,F.A., Hsu,P.D., Lin,C.-Y., Gootenberg,J.S., Konermann,S., Trevino,A.E., Scott,D.A., Inoue,A., Matoba,S., Zhang,Y. *et al.* (2013) Double nicking by RNA-guided CRISPR Cas9 for enhanced genome editing specificity. *Cell*, **154**, 1380–1389.
41. Tsai,S.Q., Wyvekens,N., Khayter,C., Foden,J.A., Thapar,V., Reyon,D., Goodwin,M.J., Aryee,M.J. and Joung,J.K. (2014) Dimeric CRISPR RNA-guided FokI nucleases for highly specific genome editing. *Nat. Biotechnol.*, **32**, 569–576.
42. Guilinger,J.P., Thompson,D.B. and Liu,D.R. (2014) Fusion of catalytically inactive Cas9 to FokI nuclease improves the specificity of genome modification. *Nat. Biotechnol.*, **32**, 577–582.
43. Wolfs,J.M., Hamilton,T.A., Lant,J.T., Laforet,M., Zhang,J., Salemi,L.M., Gloor,G.B., Schild-Poulter,C. and Edgell,D.R. (2016) Biasing genome-editing events toward precise length deletions with an RNA-guided TevCas9 dual nuclease. *Proc. Natl. Acad. Sci. U.S.A.*, **113**, 14988–14993.
44. Bolukbasi,M.F., Liu,P., Luk,K., Kwok,S.F., Gupta,A., Amrani,N., Sontheimer,E.J., Zhu,L.J. and Wolfe,S.A. (2018) Orthogonal Cas9–Cas9 chimeras provide a versatile platform for genome editing. *Nat. Commun.*, **9**, 4856.
45. McKenna,A. and Shendure,J. (2018) FlashFry: a fast and flexible tool for large-scale CRISPR target design. *BMC Biol.*, **16**, 74.
46. Hsu,P.D., Scott,D.A., Weinstein,J.A., Ran,F.A., Konermann,S., Agarwala,V., Li,Y., Fine,E.J., Wu,X., Shalem,O. *et al.* (2013) DNA targeting specificity of RNA-guided Cas9 nucleases. *Nat. Biotechnol.*, **31**, 827–832.
47. Šošić,M. and Šikić,M. (2017) Edlib: a C/C library for fast, exact sequence alignment using edit distance. *Bioinformatics*, **33**, 1394–1395.
48. Zhang,J., Kobert,K., Flouri,T. and Stamatakis,A. (2014) PEAR: a fast and accurate Illumina Paired-End reAd mergeR. *Bioinformatics*, **30**, 614–620.
49. Smith,T., Heger,A. and Sudbery,I. (2017) UMI-tools: modeling sequencing errors in Unique Molecular Identifiers to improve quantification accuracy. *Genome Res.*, **27**, 491–499.
50. Cock,P.J.A., Antao,T., Chang,J.T., Chapman,B.A., Cox,C.J., Dalke,A., Friedberg,I., Hamelryck,T., Kauff,F., Wilczynski,B. *et al.* (2009) Biopython: freely available Python tools for computational molecular biology and bioinformatics. *Bioinformatics*, **25**, 1422–1423.
51. Kingma,D.P. and Ba,J. (2014) Adam: A Method for Stochastic Optimization. arXiv doi: https://arxiv.org/abs/1412.6980, 22 December 2014, preprint: not peer reviewed.
52. Chollet,F. (2015) Keras. https://github.com/fchollet/keras.
53. Al-Rfou,R., Alain,G., Almahairi,A., Angermueller,C., Bahdanau,D., Ballas,N., Bastien,F., Bayer,J., Belikov,A., The Theano Development Team *et al.* (2016) Theano: a Python framework for fast computation of mathematical expressions. arXiv doi: https://arxiv.org/abs/1605.02688, 09 May 2016, preprint: not peer reviewed.