Systematic *in vitro* specificity profiling reveals nicking defects in natural and engineered CRISPR-Cas9 variants

Karthik Murugan [1, 2, 4], Shravanti K. Suresh [1], Arun S. Seetharam [3], Andrew J. Severin [3] and Dipali G. Sashital [1, 2 *]

Affiliations:

[1] Roy J. Carver Department of Biochemistry, Biophysics & Molecular Biology, Iowa State University, Ames, IA 50011, USA

[2] Molecular, Cellular, and Developmental Biology Interdepartmental Program, Iowa State University, Ames, IA 50011, USA

[3] Genome Informatics Facility, Office of Biotechnology, Iowa State University, Ames, IA 50011, USA

[4] Present address: Integrated DNA Technologies Inc., Coralville, IA 52241, USA

*Correspondence:

* To whom correspondence should be addressed. Tel: +1 (515)-294-5121; Fax: +1 (515)-294-7629 Email: sashital@iastate.edu

ABSTRACT

Cas9 is an RNA-guided endonuclease in the bacterial CRISPR-Cas immune system and a popular tool for genome editing. The commonly used *Streptococcus pyogenes* Cas9 (SpCas9) is relatively non-specific and prone to off-target genome editing. Other Cas9 orthologs and engineered variants of SpCas9 have been reported to be more specific. However, previous studies have focused on specificity of double-strand break (DSB) or indel formation, potentially overlooking alternative cleavage activities of these Cas9 variants. In this study, we employed *in vitro* cleavage assays of target libraries coupled with high-throughput sequencing to systematically compare cleavage activities and specificities of two natural Cas9 variants (SpCas9 and *Staphylococcus aureus* Cas9) and three engineered SpCas9 variants (SpCas9 HF1, HypaCas9, and HiFi Cas9). We observed that all Cas9s tested could cleave target sequences with up to five mismatches. However, the rate of cleavage of both on-target and off-target sequences varied based on target sequence and Cas9 variant. In addition, SaCas9 and engineered SpCas9 variants nick targets with multiple mismatches but have a defect in generating a DSB, while SpCas9 creates DSBs at these targets. Overall, these differences in cleavage rates and DSB formation may contribute to varied specificities observed in genome editing studies.

INTRODUCTION

Cas9 is the well-studied effector protein of type II CRISPR-Cas (clustered regularly interspaced short palindromic repeats-CRISPR associated) bacterial immune systems (1, 2). Cas9 is an endonuclease that uses a dual CRISPR RNA (crRNA) and trans-activating crRNA (tracrRNA) to bind dsDNA targets that are complementary to the guide region of the crRNA and adjacent to a short, conserved protospacer-adjacent motif (PAM) sequence (3, 4). Two nuclease domains in Cas9, HNH and RuvC, cut the target and non-target strand respectively, generating a double-stranded break (DSB) in the dsDNA (4) with little post-cleavage trimming (5, 6). The dual RNAs can be combined into a single guide-RNA (sgRNA) and the targeting region can be varied, making Cas9-sgRNA a readily programmable, two component system for use in various biotechnological applications (4, 7). In particular, DSB formation followed by DNA repair can lead to changes in genomic DNA sequence, enabling genome editing following Cas9 cleavage (8, 9).

Cas9 can tolerate mismatches between the crRNA and the target DNA, which is consistent with its role as a bacterial immune system effector in facilitating defense against rapidly evolving bacteriophages (10–13). Cas9 generally tolerates multiple mismatches in the PAM-distal region while PAM-proximal “seed” mismatches reduce the cleavage activity (14–18). This low fidelity leads to off-target activity when used for genome editing applications, as Cas9 can create DSBs at sites with limited homology to the intended target (16, 17, 19). While the commonly used wildtype (WT) *Streptococcus pyogenes* Cas9 (SpCas9) can tolerate multiple mismatches in the target sequence, other naturally occurring Cas9 orthologs from *Staphylococcus aureus*, *Neisseria meningitidis* and *Campylobacter jejuni* are reported to have higher specificity in genome editing compared to SpCas9 (20–23). Many other strategies have been developed to reduce off-target activity of Cas9 (24). SpCas9 has been engineered to improve the fidelity of target cleavage activity. Some mutations were designed to reduce DNA target interactions, making the requirement for complete complementarity with the crRNA more stringent (25, 26). Mutations rationally introduced in the REC domain of SpCas9 prevent conformational changes required for nuclease domain activation when a target sequence with mismatches is encountered (27, 28). Bacterial screens have also been used to select high-fidelity SpCas9 variants that maintain on-target cleavage but have reduced off-target cleavage activity (29–31).

Several methods have been developed to detect and study off-target activities of Cas9 (24, 32–35). However, methods that measure Cas9 off-target editing in eukaryotic cells are limited because cellular factors like nucleosomes may sequester potential cleavage sites (36, 37). DNA accessibility can also vary depending on cellular processes, which may change the outcome and detection of potential Cas9 off-target editing events. These methods also rely on DSBs in the DNA generated by Cas9 or post-cleavage DNA repair and indel formation, which can vary among cell types and experiments (24, 32–35).

Differences in Cas9/sgRNA delivery methods and cell lines have resulted in discrepancies in the reported specificities of high-fidelity Cas9 variants (25, 27, 31, 35).

To avoid these pitfalls, specificity studies can be performed *in vitro* to detect the native cleavage activities of Cas9 variants (6, 14, 15, 18, 38–41). Here, we used a previously established *in vitro* plasmid library cleavage assay to compare the native cleavage specificity of different Cas9 variants (42). Our method enables the detection of target sequences that may be incompletely cleaved by Cas9, leading to nicking. We tested the cleavage activity of two WT Cas9 orthologs, SpCas9 and *S. aureus* Cas9 (SaCas9), and three engineered SpCas9 variants, SpCas9 HF1 (25), hyper-accurate Cas9 (HypaCas9) (27) and Alt-R ® S.p. HiFi Cas9 (31) against two different target library sequences. Each of these three variants represent a version of high-fidelity Cas9 developed via different strategies discussed above. We show that SpCas9 rapidly cleaves target sequences with up to five mismatches. While the high-fidelity Cas9 variants retained cleavage activity against targets with multiple mismatches, they have reduced rates of cleavage compared to SpCas9. High-fidelity Cas9 variants also nick target sequences with multiple mismatches, resulting in incomplete DSB formation at sites that are fully cleaved by wild-type SpCas9. Overall, our study reveals a target-sequence dependent nicking defect of high-fidelity Cas9 variants that may account for increased specificity observed in genome editing studies that often rely on DSB formation to detect off-target sites.

MATERIAL AND METHODS

Cas9 expression vectors

Expression plasmids for SpCas9 and high-fidelity variants were purchased from Addgene. *Streptococcus pyogenes* Cas9 (SpCas9) (pMJ806) was expressed using expression plasmid pEC-K-MBP, and SpCas9-HF1 (pJSC111) and HypaCas9 (pJSC173) were expressed using expression plasmid pCT10. pMJ806, pJSC111, pJSC173 were gifts from Jennifer Doudna and/or Keith Joung (Addgene plasmid # 39312; http://n2t.net/addgene:39312; RRID:Addgene_39312; Addgene plasmid # 101209; http://n2t.net/addgene:101209; RRID:Addgene_101209; Addgene plasmid # 101218; http://n2t.net/addgene:101218; RRID:Addgene_101218). The gene sequence for *Staphylococcus aureus Cas9* (SaCas9) was synthesized as *Escherichia coli* codon-optimized gBlocks (purchased from Integrated DNA Technologies, IDT). SaCas9 gBlocks were cloned into pSV272 with N-terminal 6X-His sequence, a maltose binding protein (MBP) and a Tobacco Etch Virus (TEV) protease cleavage site via Gibson assembly (New England Biolabs) as per the manufacturer's protocol. All sequences were verified by Sanger sequencing (Eurofins Genomics, Kentucky, USA).

Cas9 expression and purification

All Cas9 proteins were expressed in *Escherichia coli* BL21 (DE3) cells. Overnight cultures of the cells carrying the expression plasmid were used to inoculate 2X TY broth supplemented with corresponding antibiotics in 1:100 ratio. The antibiotics used were kanamycin at 25 µg/mL for SpCas9 (pMJ806), and at 50 µg/mL for SaCas9 (pSV272 construct) and ampicillin at 100 µg/mL for SpCas9-HF1 (pJSC111) and HypaCas9 (pJSC173). Cultures were grown at 37 °C to an optical density (600 nm) of 0.5 − 0.6 and IPTG was added to a final concentration of 0.2 mM to induce protein expression. The incubation was continued at 18 °C overnight (~16 – 18 hours) and harvested the next day for protein purification.

SpCas9 was purified using a previously established protocol (43). Cells were resuspended in Lysis Buffer I (20 mM Tris-HCl pH 8.0, 500 mM NaCl, 10 mM imidazole, and 10% glycerol) supplemented with PMSF. A sonicator was used to lyse the cells and the lysate was centrifuged to remove insoluble material. The clarified lysate was applied to a HisPur™ Ni-NTA Resin (ThermoFisher Scientific) column. After washing the column with Lysis Buffer I, the bound protein was eluted in Elution Buffer I (Lysis Buffer I + 250 mM imidazole final concentration). The Ni-NTA column eluent was concentrated and run on a HiLoad 16/600 Superdex 200 gel filtration column (GE Healthcare) pre-equilibrated with SEC Buffer A (20 mM Tris-HCl, pH 8.0, and 500 mM NaCl). TEV protease was added at 1:100 (w/w) ratio to the pools containing 6X His-MBP tagged Cas9 and incubated on ice, overnight at 4 °C. Samples were reapplied to HisPur™ Ni-NTA Resin (ThermoFisher Scientific) to remove the His-tagged TEV, free 6X His-MBP, and any remaining tagged protein. The flow-through was collected, concentrated and further purified by using a HiLoad 16/600 S200 gel filtration column in SEC Buffer B (20 mM Tris-HCl, pH 8.0, 200 mM KCl, and 1mM EDTA). Peak pools were analyzed on SDS-PAGE gels and the pools with Cas9 were combined, concentrated, flash frozen in liquid nitrogen and stored at −80°C until further use. Cleavage activity of SpCas9 purified using this protocol was similar to commercially available SpCas9 (data not shown).

An alternative previously established purification protocol was used for all other Cas9 variants (44), with the exception of Alt-R ® S.p. HiFi Cas9, which was provided by Integrated DNA Technologies (IDT). Harvested cells were resuspended in Lysis Buffer II (20 mM Tris-HCl pH 8.0, 500 mM NaCl, 5 mM imidazole), supplemented with protease inhibitors (PMSF, cOmplete™ Protease Inhibitor Cocktail Tablet or Halt Protease Inhibitor Cocktail). A sonicator was used to lyse the cells and the lysate was centrifuged to remove insoluble material. The clarified lysate was applied to a HisPur™ Ni-NTA Resin (ThermoFisher Scientific) column. After washing the column with 10 column volumes of Wash Buffer (Lysis Buffer + 15 mM imidazole final concentration), the bound protein was eluted in Elution Buffer I (Lysis Buffer II + 250 mM imidazole final concentration). Fractions containing Cas9 were pooled and TEV protease was added in a 1:100 (w/w) ratio and dialyzed in Dialysis Buffer (10 mM HEPES-KOH pH 7.5, 200 mM KCl, 1 mM DTT) at 4°C overnight. The dialyzed protein was diluted 1:1 with 20 mM HEPES KOH (pH 7.5) and

loaded on a HiTrap Heparin HP (GE Healthcare) column and washed with Buffer A (20 mM HEPES-KOH pH 7.5, 100 mM KCl). The protein was eluted with Buffer B (20 mM HEPES-KOH pH 7.5, 2 M KCl) by applying a gradient from 0% to 50% over a total volume of 60 ml. Eluted peak fractions were analyzed by SDS-PAGE and fractions with Cas9 were combined and concentrated. DTT was added to a final concentration of 1 mM. The protein was fractionated on a HiLoad 16/600 Superdex 200 gel filtration column (GE Healthcare), eluting with SEC buffer (20 mM HEPES-KOH pH 7.5, 500 mM KCl, 1 mM DTT). Peak pools were analyzed on SDS-PAGE gels and the pools with Cas9 were combined, concentrated, flash frozen in liquid nitrogen and stored at −80°C until further use.

Variations in Cas9 purification procedures could lead to differences in activity of the Cas9 variants. However, the level of purity was similar for all variants, and conditions were identical for all reactions (Fig. S1A) (see methods section - *In vitro* cleavage assay and analysis). All Cas9s were frozen as high concentration stocks (~ 61 – 200 µM). Working stock concentrations of the proteins (5 or 10 µM) were made in SEC buffer (20 mM HEPES-KOH pH 7.5, 500 mM KCl, 1 mM DTT).

Library creation

Target libraries were partially randomized to generate a pool of sequences containing mismatches (45). The following probability distribution function was used to determine the randomization/doping frequency,

$$\mathrm{P}\,(\mathrm{n},\mathrm{L},\mathrm{f}) \;=\; \frac{\mathrm{L}!}{\mathrm{n}!(\mathrm{L}-\mathrm{n})!}(\mathrm{f}^{\mathrm{n}})(1\;-\;\mathrm{f})(\mathrm{L}-\mathrm{n})\;\text{[Eq. 1]}$$

where, P is the pool of the population, L is the sequence length, n is the number of mutations/template and f is the probability of mutation/position (doping level or frequency). A randomization/doping frequency (f) of 15% results in a library containing a mixed pool of sequences of 20 nt (L) with a high representation of 2 to 4 mismatches (n). Single-stranded oligonucleotide libraries were ordered from IDT using hand mixed pools (https://www.idtdna.com/pages/products/custom-dna-rna/mixed-bases). For libraries with 15% randomization/doping frequency, if the target sequence has A at a given position, a mix of A:C:G:T would be dispensed in 85:5:5:5 ratio during oligonucleotide synthesis resulting in 85% A at this position and 15% of C, G or T (5% each).

The number of different mutation combinations (MM_c) for a given number of mutations, n, and sequence length, L, regardless of the doping level/frequency is determined by,

$$MM_c = 3^n \frac{\mathrm{L}!}{\mathrm{n}!(\mathrm{L}-\mathrm{n})!}\;\text{[Eq. 2]}$$

The total number of unique target sequences with a single mismatch is 60, with 2 mismatches is 1,710, and with 3 mismatches is 30,780, etc. We used two library sequences that we previously tested for Cas12a (42), a modified protospacer 4 (PS4) sequence from *Streptococcus pyogenes* CRISPR locus (55% GC) and EMX1 gene target sequence (80% GC) (see Supplementary Table 1 for target sequence).

Plasmid and nucleic acid preparation

All DNA oligonucleotides used in this study were synthesized by IDT or Thermo Scientific. RNAs (tracrRNA and crRNA) and single-stranded target or library oligonucleotides were ordered from IDT. Supplementary Table 1 lists the sequences of DNA and RNA oligonucleotides used in this study.

Gibson assembly was used to generate target (pTarget) and library (pLibrary) plasmids (46). The oligonucleotides for the targets or libraries were diluted to 0.2 μM in 1X NEBuffer 2. pUC19 vector was amplified using primers listed in Supplementary table 1 via PCR to insert homology arms. The PCR reaction was subjected to DpnI digestion and PCR clean up (Promega Wizard SV Gel and PCR Clean-Up System), as per the manufacturer's protocol. 30 ng of PCR amplified pUC19, 5 μL of oligonucleotide (0.2 μM) and ddH_2O to bring the volume to 10 μL were mixed with 10 μL 2X NEBuilder HiFi DNA Assembly Master mix (New England Biolabs) and incubated at 50 °C for 1 hour. NEB Stable competent cells were transformed with 2 μL of the assembled product, as per the manufacturer's protocol. Transformants were plated for plasmid preparation (for pTarget plasmids) or to assess transformation efficiencies (for pLibrary plasmids). For pTarget, starter cultures from individual colonies were used to inoculate 50 mL LB media with 100 μg/mL ampicillin. For pLibrary, all of cells in the outgrowth media from the transformation recovery were used to inoculate 50 mL LB with 100 μg/mL ampicillin. Cultures were grown overnight at 37 °C for plasmid propagation and extraction using QIAGEN Plasmid Midi Kit. The following precautions were taken to ensure the plasmid remained supercoiled during plasmid extractions. Cells were cooled on ice before harvesting. All initial steps from lysis to neutralization for plasmid extractions were performed on ice with minimum mechanical stress. Plasmids were stored as aliquots that were used for up to 10 freeze-thaw cycles. Different pLibrary assembly reactions and preparations were used for the replicates of the *in vitro* cleavage assays (Fig. S1B). All pTarget sequences were verified by Sanger sequencing (Eurofins Genomics, Kentucky, USA). For controls, pUC19 was prepared by restriction enzyme digestion using BsaI-HF to linearize the plasmid and Nt.BspQI to nick the plasmid using the manufacturer's protocols (New England Biolabs).

***In vitro* cleavage assay and analysis**

The protocol was adapted from previously described methods (47). Cas9:tracrRNA:crRNA complex was formed by incubating Cas9 and tracrRNA:crRNA at a 1:1.5 ratio in reaction buffer (20 mM HEPES, pH 7.4, 100 mM KCl, 5 mM $MgCl_2$, 1 mM DTT, and 5% glycerol) at 37 °C for 10 min. Cas9 RNP complex (final concentration 100 nM Cas9 and 150 nM tracrRNA:crRNA) was mixed with pTarget,

pLibrary or empty plasmid (15 ng/μL, ~9 nM) to initiate cleavage reactions at 37 °C. Phenol-chloroform was used to quench reaction aliquots at 5, 10, 15, 30, 60, 300 and 1800 s for pTarget and at 1, 5, 30, 60 and 180 min for pLibrary. The aqueous layer was extracted and separated on a 1% agarose gel via electrophoresis and stained with SYBR Safe (Invitrogen) or RedSafe (Intron Bio) stain for dsDNA visualization. Excess tracrRNA:crRNA was used in cleavage assays to prevent any RNA-independent cleavage activity (48). All cleavage assays were performed in triplicate.

Bands were visualized and quantified with ImageJ (https://imagej.nih.gov/ij/). Intensities of the band (I) in the uncleaved (supercoiled - SC) and cleaved fractions (nicked - N and linearized - L) were measured. Fractions (FR) cleaved and uncleaved were calculated as follows.

$$Fraction\ cleaved\ (FR_C) = \frac{I_N + I_L}{I_{SC} + I_N + I_L} \text{ [Eq. 3]}$$

$$Fraction\ uncleaved\ or\ supercoiled\ (FR_{SC}) = \frac{I_{SC}}{I_{SC} + I_N + I_L} \text{ [Eq. 4]}$$

$$Fraction\ nicked\ (FR_N) = \frac{I_N}{I_{SC} + I_N + I_L} \text{ [Eq. 5]}$$

$$Fraction\ linearized\ (FR_L) = \frac{I_L}{I_{SC} + I_N + I_L} \text{ [Eq. 6]}$$

The FR_{SC}, FR_N and FR_L were determined for each of the time points 't'. FR for time point 0 (FR_0) was determined for the negative control pLibrary (i.e. pLibrary run on a gel after preparation as represented in Fig. S1B).

The apparent rates of pTarget and pLibrary cleavage were determined by fitting FR_C to a one-phase association equation using GraphPad Prism v 8.4.3 (https://www.graphpad.com/scientific-software/prism/).

$$FR = FR_0 + (FR_{final} - FR_0)(1 - e^{-kt}) \text{ [Eq. 7]}$$

Where t is time, FR is the appropriate FR_c that starts from FR_0 and goes to FR_{final} (FR at the last time point), and k is the apparent rate constant.

Library preparation for HTS

Agarose gel electrophoresis (as described above) was used to separate the plasmid library cleavage products into cleaved (linear and nicked) and uncleaved (supercoiled) products. The bands from the nicked and supercoiled pools from various time points were excised separately and were individually gel purified using QIAquick Gel Extraction Kit (Qiagen). Nextera Adapters (NEA) were designed to amplify across the target region in the pLibrary. Because the PCR primers amplified across the target region, Cas9-mediated linearization of the plasmid due to DSB formation at the target site did not yield any PCR

product while Cas9 mediated nicked plasmid resulted in amplification of the target region via PCR. Standard Nextera unique indices/barcodes were used to multiplex the samples and were added to the first PCR products using another round of PCR (see Supplementary Table 1 for NEA primers). Samples were purified using QIAquick PCR Purification Kit (Qiagen) between the two PCR steps. The size of the PCR products was verified using Agilent 2100 Bioanalyzer. Pooled samples were subjected to NextSeq or MiSeq for paired-end reads of 75 cycles at Admera Health, LLC (New Jersey, USA) or Iowa State DNA Facility (Ames, IA). Samples were pooled and multiplexed to get an average of 100,000 reads per sample (Fig. S2D). To ensure coverage of each sample in a minimal number of NextSeq/MiSeq runs, we included two out of the three replicates performed for the pLibrary cleavage assays. 15% PhiX was spiked in to increase sequence diversity of the sample.

HTS data analysis

Extraction of target sequences, read counts, and number of mismatches per target sequence from HTS data were analyzed using custom bash scripts (see associated GitHub repository: https://github.com/sashital-lab/Cas9_specificity). A simple workflow of the analysis is described in Supplementary figure 3, adapted from our previous study on Cas12a (42). Target sequences were extracted along with the counts of the extracted target sequences and the number of mismatches. The files containing the extracted target sequences and counts are available on Iowa State University Library's DataShare (see Availability for more information). Target sequence information was imported into Microsoft Excel or R for plotting and summarizing, post command-line processing.

In each pool, the fraction of target sequences containing 'n' mismatches (MM) ($F_{n\text{-}MM}$) was calculated as follows.

$$F_{n-MM} = \frac{total\ counts\ of\ sequences\ with\ n\ mismatches}{total\ counts\ of\ all\ sequences\ in\ the\ pool} \quad \text{[Eq. 8]}$$

$F_{n\text{-}MM}$ was normalized to the fraction (FR) of DNA present in the supercoiled or nicked fraction at a given time point 't' to generate an estimated abundance (EA) of a given set of sequences at a given timepoint. FR was calculated for each time point using equations 3 through 6 as described above.

$$EA_{n-MM} = (F_{n-MM}\ of\ S\ at\ t) * (FR\ at\ t) \quad \text{[Eq. 9]}$$

These values were plotted against number of mismatches (n) to generate mismatch distribution curves.

The relative abundance (enrichment and/or depletion) (RA) of a sequence containing 'n' mismatches at each time point 't' compared to the negative control, (i.e. pLibrary run on a gel after preparation as represented in Fig. S1B).

$$RA_S = \frac{EA_{n-MM\ of\ S\ at\ t}}{EA_{n-MM}\ in\ pLibrary}$$ [Eq. 10]

Log-fold change in abundance was calculated as in equation 11 for each time point 't' and plotted as a heatmap to determine overall depletion or accumulation of targets containing a certain number of mismatches.

$$\text{Log} - \text{fold change in adundance} = \log_2(RA_S)\ at\ t$$ [Eq. 11]

The RA for the perfect target sequence (0 MM), RA_{0MM} was calculated using equation 10, where n = 0 at the different time points. The RA for target sequences with 1 to 5 MM at each time point 't', $RA_{1\text{-}5MM\text{-}t}$ was calculated by summing EA for 1 to 5 MM, $EA_{1\text{-}5MM}$ at each time point, and normalizing to the sum of EA of 1 to 5 MM in the negative control (i.e. pLibrary run on a gel after preparation as represented in Fig. S1B) as shown below.

$$RA_{1-5MM-t} = \sum_{n=1}^{n=5} \frac{EA_{n-MM\ of\ S\ at\ t}}{EA_{n-MM}\ in\ pLibrary}$$ [Eq. 12]

The relative cleaved fraction of counts for on-target and off-targets ($RA_{cleaved\ FR}$) was determined by subtracting RA_{0MM} and $RA_{1\text{-}5MM}$ values, respectively from 1 at each time point 't', as shown below and plotted against time.

$$RA_{cleaved\ FR-on-t} = 1 - RA_{0MM-t}$$ [Eq. 13]

$$RA_{cleaved\ FR-off-t} = 1 - RA_{1-5MM-t}$$ [Eq. 14]

The specificity score (SS) for Cas9 cleavage was calculated by dividing the on-target by off-target $RA_{cleaved\ FR}$ at each time point 't'.

$$\text{SS} = \frac{RA_{cleaved\ FR-on-t}}{RA_{cleaved\ FR-off-t}}$$ [Eq. 15]

The specificity scores of SaCas9 and HF Cas9 variants were normalized to WT SpCas9 to determine relative specificity at each time point.

For the heatmaps, the estimated abundance (EA) of sequences containing a particular nucleotide (N = A, G, C, T) at a particular position (P = 1 to 20) for target sequences containing 'n' mismatches at each time point 't' was calculated as above. Relative abundance (RA) was calculated by normalizing EA against the pool of DNA in the original library to eliminate variability in aberrant nicking that may have occurred for individual pLibraries in the negative control.

$$RA_{S-NP} = \frac{EA_{n-MM\ of\ S\ with\ N\ at\ P\ at\ t}}{EA_{n-MM\ of\ S\ with\ N\ at\ P\ in\ pLibrary}} \text{ [Eq. 16]}$$

For the supercoiled pool, we calculated the maximum change in relative abundance (RA) over time as max $\Delta RA_{S\text{-}NP}$ for each sequence containing a particular nucleotide (N = A, G, C, T) at a particular position (P = 1 to 20) for target sequences containing 'n' mismatches over all time points 't' (0 min, 1 min, 5 min, 30 min, 60 min and 180 min). Max $\Delta RA_{S\text{-}NP}$ is indicated as max Δ abundance in the figures for simplicity.

$$Max\ \Delta RA_{S-NP} = \text{Max}\left[\frac{(RA_{S-NP}\ \ at\ t_{T-1}) - (RA_{S-NP}\ \ at\ t_T)}{(\ t_T) - (t_{T-1})}\right] \text{ [Eq. 17]}$$

For the nicked pool, we calculated the average change in relative abundance (RA) over time as $\Delta RA_{S\text{-}NP}$ for each sequence containing a particular nucleotide (N = A, G, C, T) at a particular position (P = 1 to 20) for target sequences containing 'n' mismatches over time points 't' after Cas9 cleavage (1 min, 5 min, 30 min, 60 min and 180 min). $\Delta RA_{S\text{-}NP}$ is indicated as Δ abundance in the figures for simplicity.

$$\Delta RA_{S-NP} = \text{Average}\left[\frac{(RA_{S-NP}\ \ at\ t_T) - (RA_{S-NP}\ \ at\ t_{T-1})}{(\ t_T) - (t_{T-1})}\right] \text{ [Eq. 18]}$$

In the supercoiled pool, we defined the extent of cleavage of a target sequence from the supercoiled pool as $abundance_{min}$ by determining the minimum value of $RA_{S\text{-}NP}$ across all time points for those target sequences. For the nicked pool, we defined the extent of nicking of a target sequence as $abundance_{max}$ by determining the maximum value of $RA_{S\text{-}NP}$ across all time points for those target sequences. $Abundance_{min}$ or $abundance_{max}$ were normalized to the highest value across both pLibraries, Cas9s and mismatches (1 to 5 MM) which allows comparison between Cas9s and mismatches. Using custom scripts in R, the Δ abundance and $abundance_{min}$ and $abundance_{max}$ were used to plot the bubble heatmaps for the supercoiled and nicked pools, respectively. Δ max change and Δ abundance defined the gradient colour and $abundance_{min}$ and $abundance_{max}$ defined the bubble size.

For the analysis of target sequences with two mismatches, the sequences with 2 mismatches were extracted. The distance between the two mismatches and the total counts for sequences separated by that distance were determined. The counts were normalized to the number of possible ways the two mismatches can occur (42), and the max Δ abundance, Δ abundance, $abundance_{min}$ and $abundance_{max}$ were calculated similarly to equations 16, 17 and 18 and plotted versus distance between mismatches.

RESULTS

Cleavage activity of Cas9 against target library

We sought to compare the cleavage activity and specificity of different Cas9 variants in a systematic manner. We performed a previously established *in vitro* plasmid library (pLibrary) cleavage assay with five Cas9 variants (42), WT SpCas9, WT SaCas9 and three high-fidelity variants of SpCas9 – SpCas9 HF1, HypaCas9, and Alt-R ® S.p. HiFi Cas9 (HiFi Cas9) (Fig. 1A, S1A) (25, 27, 31). The three high-fidelity variants of SpCas9 will be collectively referred to as HF Cas9 hereafter. For each Cas9 variant, we used two different crRNA sequences with partner tracrRNA and generated corresponding negatively supercoiled (nSC) plasmids containing the perfect target (pTarget) or target library (pLibrary) (see methods section – Plasmid and nucleic acid preparation) (Fig. S1B). The pLibraries contained a distribution of target sequences with between zero and ten mismatches to the crRNA guide sequence, with a maximum representation of target sequences with two to four mismatches in the libraries (Fig. S1C). The two crRNA and library sequences were designed based on protospacer 4 sequence from *Streptococcus pyogenes* CRISPR locus (55% G/C) and EMX1 gene target sequence (80% G/C), referred to as pLibrary PS4 and pLibrary EMX1 respectively. We employed the native dual crRNA and tracrRNA system for our assay to avoid any differences that may stem from single guide RNA design optimization (49, 50).

We used the differential migration of the nicked (n) and linear (li) cleavage products of negatively supercoiled (nSC) dsDNA plasmid on an agarose gel to analyze Cas9 cleavage activity (51) (Fig. S2A). Linear products represent fully cleaved DNA, in which both strands were cleaved by Cas9. The accumulation of linear DNA over time was used to determine rates of cleavage for pTarget. Cleavage rates of pTarget were significantly variable depending both on target sequence and Cas9 variant (Fig. 1B, C, S2A, B). SpCas9 cleaved pTarget PS4 ~3.6-fold faster than pTarget EMX1. A similar trend was observed for SaCas9, although this ortholog cleaved both pTargets ~7-fold slower than SpCas9 (Fig. 1B, C, S2A, B). Among the HF Cas9s, HiFi Cas9 had cleavage rates that were comparable to WT SpCas9. In contrast, SpCas9 HF1 and HypaCas9 cleaved pTarget PS4 ~36- and ~12-fold slower than SpCas9, respectively (Fig. 1B, C, S2B), similar to previously reported cleavage defects for these two HF Cas9 variants (52). However, cleavage rates for SpCas9 HF1 and HypaCas9 were comparable to SpCas9 for

pTarget EMX1 (Fig. 1C, S2A, B), indicating that cleavage defects for HF Cas9 variants may vary based on target sequence.

For pLibrary cleavage assays, we observed a substantial amount of nicked product, resulting from incomplete cleavage of the target. We therefore determined the apparent rate of overall cleavage (nicked and linearized product) (Fig. 1C, S2A, B, see methods section - *In vitro* cleavage assay and analysis). As expected, rates of pLibrary cleavage were substantially slower than for pTargets, due to the presence of mismatches in the target sequence (Fig. 1B, C, S2A, B). SpCas9 rapidly cleaved more than 50% of both negatively supercoiled pLibraries, with the vast majority of product DNA becoming linearized (Fig. 1B, C). In contrast, for SaCas9 and HF Cas9 variants, we observed greater accumulation of nicked plasmid, especially for pLibrary PS4 (Fig. 1B, C). On average, all other Cas9 variants accumulated significantly more nicked product for pLibrary PS4 than SpCas9 (Fig. 1D). For pLibrary EMX1, SaCas9 and SpCas9 HF1 had significantly more accumulation of nicked product than SpCas9 (Fig. 1D).

We also checked whether cleavage occurred outside of the target region during pLibrary cleavage by testing the cleavage activity of Cas9 against the empty plasmid backbone without and with the different crRNAs (Fig. S2C). The empty plasmid was minimally cleaved by Cas9-tracrRNA:crRNA, except in the case of SpCas9-EMX1 crRNA where a substantial nicked product was observed at the three hour time point. However, we do not observe similar amounts of nicking of the pLibrary EMX1 by Cas9 (Fig. S2C) and further analysis indicated that pLibrary nicking is target-sequence dependent (see below).

To determine which sequences were cleaved by Cas9 variants, we extracted the plasmid DNA from the supercoiled and nicked pools, performed barcoded-PCR amplification and multiplexed, high-throughput sequencing (HTS) to get sufficient coverage of reads for each sample (Fig. 1A, S2D see methods section – Library preparation for HTS). Although we were unable to sequence the linearized pool using PCR amplicon sequencing, for our analysis, we assumed that target sequences absent from both the supercoiled and nicked pools were linearized. We determined the fraction of counts for the target sequences in the HTS data and normalized this fraction with the fraction of DNA present in the pool at a given time point (Fig. 1B, C) to represent an estimated abundance of given target sequences within the pool (see methods section – HTS analysis). Here, target sequences cleaved by Cas9 were depleted from the supercoiled pool while those nicked by Cas9 were enriched in the nicked pool.

We initially evaluated the cumulative effects of mismatches on the cleavage activity of each Cas9 variant by plotting the log-fold change in targets containing different numbers of mismatches with the crRNA guide sequence over time (Fig. 2). We also plotted target abundance as mismatch distribution curves (Fig. S4). Together, the heatmaps and mismatch distribution curves enable overall comparison of cleavage for target sequences containing varying numbers of mismatches with the crRNA across Cas9 variants, across time points for each Cas9 variant (Fig. 2, S4). As expected, the perfect target (zero

mismatch) was rapidly depleted from the supercoiled pool of the pLibrary (Fig. 2A, B, S4A, B). SpCas9 partially cleaved sequences with up to four mismatches in the first time point tested for both pLibraries, as observed in previous *in vitro* and *in vivo* studies on SpCas9 cleavage specificity (17, 18) (Fig. S4A, B). This observation indicates that our *in vitro* pLibrary cleavage assay reproduced a similar specificity profile for SpCas9 as previous studies and can further be used to benchmark against SaCas9 and HF Cas9 variants. Like SpCas9, SaCas9 and HF Cas9 variants cleaved sequences containing up to four mismatches in both pLibraries, although the rate and extent of depletion of these sequences varied (Fig. 2A, B and S4A, B). In general, variations in rates of depletion of mismatched sequences correlated with reduced rates of cleavage of the perfect target (Fig. 2A, B), with SpCas9 HF1 and HypaCas9 showing slowest depletion of all targets in pLibrary PS4 and SaCas9 showing slowest depletion of all targets in pLibrary EMX1.

We also observed substantial accumulation of nicked target sequences with two to five mismatches, especially for SaCas9 and HF Cas9 variants cleaving pLibrary PS4 (Fig. 2C, D, S4C, D). These results suggest that SaCas9 and HF Cas9 variants are slower to fully cleave targets containing several mismatches than WT SpCas9, resulting in formation of nicks. Notably, some target sequences with one and two mismatches were initially nicked by SaCas9 or HF Cas9 variants, but subsequently depleted from the nicked pool due to completion of DSB formation (Fig. 2C, D, S4C, D). In addition, we observed differential amounts of accumulation of nicked DNA for targets containing three to five mismatches between the two pLibraries (Fig. 2C, D). Overall, these data suggest that Cas9 variant and crRNA sequence can affect the rate of second-strand cleavage at mismatched targets.

Prolonged exposure reduces specificity of high-fidelity Cas9 variants

Our HTS data allows us to compare the overall cleavage efficiency and specificity of the Cas9 variants. We first determined the efficiency of cleavage of the perfect target and targets with multiple mismatches (one to five MM) in the pLibrary (Fig. 3A – D) (see methods – HTS analysis). Cleavage efficiencies of the perfect target within pLibrary were similar to those observed for pTarget (Fig. 1B, C, S2B). Analysis of mismatched targets indicated differences in cleavage efficiencies in comparison to the perfect target (Fig. 3C, D). For example, while HiFi Cas9 cleaved the PS4 perfect target with similar efficiency to SpCas9, we observed a marked reduction in cleavage of PS4 mismatched targets for HiFi Cas9 (Fig. 3C).

To analyze these differences in cleavage efficiencies, we generated a specificity score that reports the relative efficiency of cleavage of on- and off-target sequences over time for the Cas9 variants relative to SpCas9 (Fig. 3E, F) (See methods – HTS analysis). For the two WT Cas9 orthologs, we did not observe significant differences in specificity scores, suggesting that SpCas9 and SaCas9 have similar specificities for the two target sequences. All three HF Cas9 variants had some significant differences in

specificity scores relative to WT SpCas9 at early time points (Fig. 3E). The relative specificity scores for HF Cas9 variants were substantially larger for pLibrary PS4 than for pLibrary EMX1. Notably, we did not observe significant differences in specificity scores for any Cas9 variants at later time points (≥ 30 min). The lack of specificity differences between WT and HF Cas9 at longer time points indicates that prolonged exposure of HF Cas9 variants can eventually lead to off-target cleavage activity.

Sequence determinants of Cas9 cleavage activity and nicking defects

We next wanted to characterize the effects of mismatch position and type on Cas9 cleavage (Fig. 3E, F). We analyzed the sequences present in both the supercoiled and nicked pools and calculated the relative abundance of target sequences containing one to five mismatches over time (see methods section – HTS analysis). To visualize the effects of mismatches, we used bubble heatmaps that reveal the maximal extent of cleavage (defining bubble size) and the rate of cleavage (defining gradient color) for targets containing a given mismatch type at a given position of the target.

For the supercoiled pool, target sequences that were depleted over time represent sequences that can be cleaved by Cas9. Therefore, the minimum relative abundance value in the time course ($abundance_{min}$) represents the extent of target sequence cleavage by Cas9 (Fig. 4, S5). To estimate the rate of depletion of sequences from the supercoiled pool, we calculated the maximal change in relative abundance between time points (max Δ abundance), colored as depleted (red) or unchanged (white). For SpCas9, the heatmaps reveal cleavage defects in the PAM-proximal "seed" region for target sequences with two to four mismatches, similar to previously reported seed regions comprising eight to ten PAM-proximal nucleotides (6, 17, 18, 53, 54). Seed defects for target sequences with one mismatch were less pronounced. The higher tolerance for a single mismatch in the seed sequence is likely due to the relatively high concentration of Cas9 used for pLibrary cleavage (18). Notably, while seed-dependent defects were evident for other Cas9 variants (Fig. 4, S5), SpCas9 HF1 and HypaCas9 also had substantial cleavage defects for targets containing mismatches located outside of the seed for pLibrary PS4 (Fig. 4). Mismatches located toward the middle of PS4 (positions 11 to 13) were particularly deleterious for HypaCas9 cleaving one to three mismatch targets, while PAM-distal mismatches as far as the second to last position (position 19) from the PAM were highly deleterious for SpCas9 HF1. These results suggest that mismatches are more uniformly deleterious throughout the target for some HF Cas9 variants, although this observation was dependent on target sequence. Despite differences in the rate and extent of cleavage, mismatch specific effects were generally very similar among all Cas9 variants. These effects were more pronounced in the seed, where C-C or U-C mismatches were generally strongly deleterious (Fig. 4, S5). In contrast, G-T mismatches were tolerated well within the seed for all Cas9 variants. These mismatch identity observations for Cas9 are consistent with previous *in vitro* library studies (6, 40).

As noted above, we observed significant accumulation of nicked plasmid for all Cas9 variants in comparison to SpCas9 for pLibrary PS4, and for SaCas9 and Cas9 HF1 for pLibrary EMX1 (Fig. 1D). This accumulation was likely due to a defect in cleavage of the second strand following an initial nicking event. Our HTS data revealed that in the nicked pool, some target sequences initially have a high relative abundance that decreased over time, indicating the eventual formation of a DSB. In contrast, some target sequences were initially uncleaved but accumulated within the nicked pool over time. To visualize these effects, we plotted the maximum abundance ($abundance_{max}$) to define the maximal extent of nicking and colored by the average change in abundance over time (Δ abundance) to define nicked targets that were depleted (red), accumulated (blue), or unchanged (white) following the first time point (Fig. 5, S6). These heatmaps reveal that second strand cleavage defects are highly dependent on mismatch position, and in some cases on mismatch type. While some nicking defects for SaCas9 were caused by seed mismatches, the most notable nicking defects occurred for targets containing mismatches toward the middle of the target sequence (positions 9 to 12). For pLibrary PS4 targets, G-T or U-G mismatches within this region caused a nicking defect that was severely compounded upon addition of further mismatches. For sequences with one or two mismatches, targets containing these mismatches were initially nicked, but rapidly linearized, as visualized by large red circles (Fig. 5). However, when present within three or four mismatch-containing targets, these mismatches caused the target to remain nicked for prolonged periods, as visualized by large blue circles. Similar positional defects in second-strand cleavage were observed for SaCas9 cleaving pLibary EMX1, although the effects were less dependent on mismatch type and less substantial than for PS4 (Fig. 5, S6). Overall, these results suggest that mismatches toward the middle of the target can reduce second-strand cleavage by SaCas9.

For HF Cas9 variants, we observed a similar position-specific defect in second-strand cleavage for pLibrary PS4 (Fig. 5, S6). These defects were not correlated with any particular mismatch type and appeared to be dependent mainly on mismatch location. Mismatches located in the PAM distal region, particularly positions 11 to 16, caused strong nicking defects for all three HF Cas9 variants for pLibrary PS4. A similar position dependence was observed for EMX1 for HF Cas9 variants, although less nicking was observed overall for this target (Fig. 1D, S4D, S5). Notably, although we observed similar patterns of depletion of supercoiled DNA between SpCas9 and HiFi Cas9 (Fig. 4), the mismatch position-dependent nicking defect was substantially greater for HiFiCas9 than for SpCas9, especially for PS4 targets containing three or four mismatches (Fig. 5). This suggests that while HiFi Cas9 can cleave target sequences with similar numbers and types of mismatches as the wild-type protein, accumulation of mismatches in the PAM-distal region results in a defect in cleavage of the second strand for HiFi Cas9 that is not observed for the wild-type protein.

Closely spaced mismatches compound overall and second-strand cleavage defects

For sequences containing multiple mismatches, it has previously been observed that the distance between mismatches can affect the level of cleavage defect by SpCas9 (6, 14, 17, 18, 38, 39, 55). We wished to determine the extent to which mismatch separation affects cleavage by all five Cas9 variants, as well as whether distance between mismatches influenced second-strand cleavage defects. We analyzed sequences containing two mismatches, which were highly represented in our target libraries (Fig. S1C). In a 20-nucleotide sequence, two mismatches can be separated by between 0 (i.e. mismatches located at adjacent positions) and 18 nucleotides (i.e. mismatches located at the beginning and end of the sequence). To determine how this distance affects the rate of cleavage for the Cas9 variants, we analyzed the supercoiled and nicked pool using bubble heatmaps as described above, but now based on the distance between the two mismatches and the location of the two mismatches.

Double mismatches spaced close together (zero to four nucleotides separation) caused substantial decrease in depletion from the supercoiled pool, consistent with previous reports that closely spaced mismatches are deleterious for Cas9-dependent cleavage (17, 18, 38) (Fig. 6A, B). One exception was SaCas9, which did not display a defect for closely spaced double mismatches for pLibrary PS4 (Fig. 6A), although this defect was apparent for pLibrary EMX1 (Fig. 6B). Conversely, SpCas9 HF1 displayed substantial cleavage defects for double mismatches spaced further apart (14 to 17 nucleotides separation) for PS4 (Fig. 6A), a defect that was not observed for EMX1 (Fig. 6B). These results underscore the variability in mismatch effects based on target sequence. To determine whether the effect of mismatch spacing is also influenced by the position within the target, we analyzed the effects of two mismatches separated by between zero and eight nucleotides within the seed or PAM-distal region (Fig. 6C, D). Closely spaced mismatches (five or fewer nucleotides separation) were highly deleterious in the seed. In contrast, mismatch spacing had little impact in the PAM-distal region, where mismatches separated by any distance were similarly tolerated.

For the nicked pool, double mismatches caused similar amounts of nicking defects regardless of spacing across the whole target, as visualized by bubbles of similar sizes (Fig. 6E, F). However, the rate of nicking or linearization of targets was impacted to some degree by mismatch distance. This is especially apparent for SaCas9 and HiFi Cas9 cleaving pLibrary PS4 (Fig. 6E). While most double mismatches led to eventual linearization by SaCas9 and HiFi Cas9, as visualized by bubbles with shades of red or white, mismatches spaced 13 to 16 nucleotides apart were shades of blue, indicating accumulation of these targets in the nicked fraction due to a stronger nicking defect. Mismatches with this spacing necessarily places one mismatch within the PAM-distal region, consistent with the position-dependent nicking defect described above (Fig. 5, S6). Further analysis of mismatch distance in the PAM-distal region revealed marked distance-dependent effects (Fig. 6G, H). In general, for SaCas9 and HF Cas9 variants, mismatches spaced closer together in the PAM-distal region caused second-strand cleavage defects and accumulation in the nicked pool for pLibrary PS4 (Fig. 6G). Double mismatches

separated by four or fewer nucleotides in the PAM-distal region were especially deleterious for SpCas9 HF1 and HypaCas9 (Fig. 6G). A similar defect for closely spaced double mismatches in the PAM-distal region was observed for SpCas9 HF1 for pLibrary EMX1, although the extent of the defect was less substantial (Fig. 6H). For SaCas9 cleaving pLibrary PS4, mismatches spaced further apart (six to eight nucleotides) in either region caused a partial defect in second-strand cleavage resulting in delayed linearization, as visualized by large white or red bubbles (Fig. 6G). Overall, these results indicate that multiple closely spaced mismatches within the PAM-distal region can cause reduced rates of second-strand cleavage for SaCas9 and HF Cas9 variants, albeit in a target-dependent manner.

Validating nicking defects against mismatched targets

Finally, to validate the nicking defect observed in pLibrary cleavage, we verified cleavage of individual target sequences containing two to five mismatches that were present in the nicked pool of pLibrary PS4 at the longest time point (three hours). Targets were subjected to cleavage by each Cas9 variant (Fig. 7A) and the extent of nicking and linearization was quantified at 10 min and 3 h (Fig. 7B). All Cas9 variants linearized targets with two or three mismatches after three hours of incubation. We observed small but significant differences in nicking and linearization between SpCas9 and other Cas9 variants for targets with two or three mismatches. The nicking defect was most notable for SaCas9 cleaving a target containing three mismatches in comparison to SpCas9, which mostly linearized this target. For targets with more than three mismatches, we observed substantially less linearization for all Cas9 variants. A target containing four mismatches within the seed (pTarget 4.1 MM) caused the strongest defect in any type of cleavage, although both SpCas9 and SaCas9 nicked 30 to 40% of the target by 3 h. Cleavage was significantly lower for all three HF Cas9 variants for this target. In contrast, all Cas9 variants cleaved target sequences with four or five mismatches in the PAM-distal region (pTarget 4.2 MM and 5 MM). As expected, based on our HTS analysis, these targets were nicked substantially but not linearized, indicating a second-strand cleavage defect. For pTarget 4.2 MM, SaCas9, HypaCas9 and HiFi Cas9 had significantly more nicked product and significantly less linearized product that SpCas9, indicating a stronger nicking defect for these variants. In contrast, SpCas9 accumulated significantly more nicked product than SaCas9, SpCas9 HF1 and HiFi Cas9 by 3 h for pTarget 5 MM, consistent with the overall lower specificity of SpCas9. Overall, these results validate stronger nicking and overall cleavage defects of SaCas9 and HF Cas9 variants in comparison to SpCas9 for the PS4 target.

DISCUSSION

Cas9 specificity has been the subject of substantial investigation and engineering efforts, due to its importance for genome editing technologies (6, 16–18, 25, 27, 31, 35, 39–41). However, many previous studies investigated individual Cas9 variants separately, focusing on target binding and/or DSB

formation by Cas9. Our *in vitro* library cleavage assay has enabled a comparative study of the cleavage specificity of Cas9 variants, revealing cleavage defects that have previously remained undetected (6, 14, 15, 38, 40, 52). We find that engineered SpCas9 variants display higher specificity than wild-type SpCas9 in a target-dependent manner, although prolonged exposure reduces this specificity. Over time, all Cas9 variants can cleave sequences with up to five mismatches. However, while SpCas9 linearizes most target sequences with multiple mismatches as previously observed, SaCas9 and HF Cas9 variants often only nick these sequences. It is well established that Cas9 binds to sequences with limited similarity to the crRNA, although it has generally been concluded that cleavage may not occur at these sites (6, 40, 56–58). Our results now reveal that partial cleavage can occur at off-target sites, although second-strand cleavage defects prevent DSB formation. Most previous specificity studies tested for DSB and/or indel formation at target and off-target sites (6, 25, 27, 31, 35, 39, 40). Although nicked DNA may be subject to error-prone DNA repair or lead to collapse of replisomes and potential mutagenesis (59–63), nicks may also be repaired by error-free DNA repair pathways. Thus, nicking defects may obscure cleavage that does occur at off-target sites, resulting in higher genome editing specificity for SaCas9 and HF Cas9 variants.

Recent studies have compared the binding and cleavage specificities of SpCas9 and HF variants, including Cas9 HF1 and HypaCas9 (6, 52, 64, 65). Although target binding defects were not observed for these variants, PAM-distal mismatches decreased the rate of cleavage for both variants in comparison to SpCas9. Our results reveal that PAM-distal mismatches not only slow the rate of overall cleavage but can also slow the rate of DSB formation for SaCas9 and HFCas9 variants, leading to nick formation. This second-strand cleavage defect may be due to R-loop collapse and premature target release following nicking of one of the strands, as has been proposed for the overall decreased kinetics of off-target cleavage by HF Cas9 variants (6, 52). Additional defects may be caused by decreased movement of the HNH domain, which is required for cleavage activation of both the HNH and RuvC catalytic domains (4, 27, 28, 66–68). Single-molecule studies of SpCas9 HF1 and HypaCas9 revealed that HNH domain movements were diminished in comparison to wild-type SpCas9, especially in the presence of PAM-distal mismatches (27). Together with our observation of nicking defects caused by PAM-distal mismatches, this suggests that cleavage by the HNH domain is impaired upon binding to targets with PAM-distal mismatches due to loss of domain rearrangements necessary to position the HNH active site for cleavage. However, sufficient HNH domain movement may occur to trigger cleavage of the non-target strand by the RuvC domain, leading to nicking of the non-target strand.

The natural role of Cas effectors is to provide defense against invading genetic elements. Specificity of these effectors has likely been tuned through evolutionary pressures exerted by rapidly evolving phages and other mobile genetic elements. Thus, it is surprising that natural orthologs of Cas effectors, including SaCas9 and various Cas12a orthologs, have been shown to have higher intrinsic

genome editing specificity than SpCas9 (21, 23, 69–71). *In vitro* investigations have been vital for defining the native cleavage specificities of these nucleases to understand their natural role as immune effectors. We and others have observed that Cas9 and Cas12a have similar PAM-distal mismatch tolerance and similar defects for C mismatches and tolerances of T mismatch (6, 42). These findings are consistent with the observation that mismatch position impacts the ability of phages to escape immunity (10, 11, 13), and suggest that the types of mutations that arise may be similarly consequential. We also previously observed that Cas12a, like SaCas9 and HF Cas9 variants, can cleave sequences with several mismatches, but displays a second-strand cleavage defect in the presence of multiple PAM-distal mismatches (42). The ability to nick target sequences with multiple mismatches may allow broader immunity against phages, as nicking within mutated target regions may reduce the rate of phage replication and could still enable target degradation by host nucleases (10, 72). Non-specific nicking activities have also been reported for several Cas effector proteins (42, 72, 73), suggesting that DNA nicking is part of the vast repertoire of nucleic acid cleavage activities employed by CRISPR-Cas systems to neutralize phage infection. Future studies may determine whether single-strand breaks in the invading phage genome are sufficient for CRISPR-mediated immunity.

AVAILABILITY

HTS data and processed data files from this study have been deposited in the Iowa State University Library's DataShare and can be found at https://doi.org/10.25380/iastate.12245846. HTS data were processed with custom bash scripts which can be found at the GitHub repository https://github.com/sashital-lab/Cas9_specificity.
All other information and data are available from the authors upon request.

SUPPLEMENTARY DATA

Supplementary Data are available at NAR online.

ACKNOWLEDGEMENT

We thank all former and current members of the Sashital Lab for helpful discussions and suggestions on various aspects of the project. We would like to thank Mollie S. Schubert and Aftan Vander Zwaag at Integrated DNA Technologies, Inc. (IDT), Coralville IA for providing initial samples of Alt-R ® S.p. HiFi Cas9. We also thank Heather S. Lewin and Megan N. O'Donnell from the University Library for helping with the data deposition to DataShare.

FUNDING

This work was supported by funds from startup funds to D.G.S from Iowa State University College of Liberal Arts and Sciences and the Roy J. Carver Charitable Trust, the National Science Foundation [grant number 1652661] to D.G.S, and National Institute of Food and Agriculture [grant number IOW05480] to D.G.S.

CONFLICT OF INTEREST

K.M is currently an employee of Integrated DNA Technologies Inc., (IDT). All other authors declare no competing interests.

REFERENCES

1. Jiang,F. and Doudna,J.A. (2017) CRISPR–Cas9 Structures and Mechanisms. Annu. Rev. Biophys., 46, 505–529.

2. Makarova,K.S., Wolf,Y.I., Iranzo,J., Shmakov,S.A., Alkhnbashi,O.S., Brouns,S.J.J., Charpentier,E., Cheng,D., Haft,D.H., Horvath,P., et al. (2020) Evolutionary classification of CRISPR–Cas systems: a burst of class 2 and derived variants. Nat. Rev. Microbiol., 18, 67–83.

3. 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., 109, E2579–E2586.

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. 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.

6. Jones,S.K., Hawkins,J.A., Johnson,N.V., Jung,C., Hu,K., Rybarski,J.R., Chen,J.S., Doudna,J.A., Press,W.H. and Finkelstein,I.J. (2020) Massively parallel kinetic profiling of natural and engineered CRISPR nucleases. Nat. Biotechnol., 10.1038/s41587-020-0646-5.

7. Doudna,J.A. (2020) The promise and challenge of therapeutic genome editing. Nature, 578, 229–236.

8. 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.

9. 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.

10. Tao,P., Wu,X. and Rao,V. (2018) Unexpected evolutionary benefit to phages imparted by bacterial CRISPR-Cas9. Sci. Adv., 4, eaar4134.

11. Paez-Espino,D., Sharon,I., Morovic,W., Stahl,B., Thomas,B.C., Barrangou,R. and Banfield,J.F. (2015) CRISPR Immunity Drives Rapid Phage Genome Evolution in Streptococcus thermophilus. mBio, 6.

12. Deveau,H., Barrangou,R., Garneau,J.E., Labonté,J., Fremaux,C., Boyaval,P., Romero,D.A., Horvath,P. and Moineau,S. (2008) Phage Response to CRISPR-Encoded Resistance in Streptococcus thermophilus. J. Bacteriol., 190, 1390–1400.

13. Barrangou,R., Fremaux,C., Deveau,H., Richards,M., Boyaval,P., Moineau,S., Romero,D.A. and Horvath,P. (2007) CRISPR Provides Acquired Resistance Against Viruses in Prokaryotes. Science, 315, 1709–1712.

14. Fu,B.X.H., St. Onge,R.P., Fire,A.Z. and Smith,J.D. (2016) Distinct patterns of Cas9 mismatch tolerance in vitro and in vivo. Nucleic Acids Res., 44, 5365–5377.

15. Fu,B.X.H., Hansen,L.L., Artiles,K.L., Nonet,M.L. and Fire,A.Z. (2014) Landscape of target:guide homology effects on Cas9-mediated cleavage. Nucleic Acids Res., 42, 13778–13787.

16. Fu,Y., Foden,J.A., Khayter,C., Maeder,M.L., Reyon,D., Joung,J.K. and Sander,J.D. (2013) High-frequency off-target mutagenesis induced by CRISPR-Cas nucleases in human cells. Nat. Biotechnol., 31, 822–826.

17. 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.

18. Pattanayak,V., Lin,S., Guilinger,J.P., Ma,E., Doudna,J.A. and Liu,D.R. (2013) High-throughput profiling of off-target DNA cleavage reveals RNA-programmed Cas9 nuclease specificity. Nat. Biotechnol., 31, 839–843.

19. Cho,S.W., Kim,S., Kim,Y., Kweon,J., Kim,H.S., Bae,S. and Kim,J.-S. (2014) Analysis of off-target effects of CRISPR/Cas-derived RNA-guided endonucleases and nickases. Genome Res., 24, 132–141.

20. Amrani,N., Gao,X.D., Liu,P., Edraki,A., Mir,A., Ibraheim,R., Gupta,A., Sasaki,K.E., Wu,T., Donohoue,P.D., et al. (2018) NmeCas9 is an intrinsically high-fidelity genome-editing platform. Genome Biol., 19, 214.

21. Friedland,A.E., Baral,R., Singhal,P., Loveluck,K., Shen,S., Sanchez,M., Marco,E., Gotta,G.M., Maeder,M.L., Kennedy,E.M., et al. (2015) Characterization of Staphylococcus aureus Cas9: a smaller Cas9 for all-in-one adeno-associated virus delivery and paired nickase applications. Genome Biol., 16, 257.

22. Kim,E., Koo,T., Park,S.W., Kim,D., Kim,K., Cho,H.-Y., Song,D.W., Lee,K.J., Jung,M.H., Kim,S., et al. (2017) In vivo genome editing with a small Cas9 orthologue derived from Campylobacter jejuni. Nat. Commun., 8, 1–12.

23. 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.

24. Vakulskas,C.A. and Behlke,M.A. (2019) Evaluation and Reduction of CRISPR Off-Target Cleavage Events. Nucleic Acid Ther., 29, 167–174.

25. 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.

26. 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.

27. 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.

28. Dagdas,Y.S., Chen,J.S., Sternberg,S.H., Doudna,J.A. and Yildiz,A. (2017) A conformational checkpoint between DNA binding and cleavage by CRISPR-Cas9. Sci. Adv., 3, eaao0027.

29. Hu,J.H., Miller,S.M., Geurts,M.H., Tang,W., Chen,L., Sun,N., Zeina,C.M., Gao,X., Rees,H.A., Lin,Z., et al. (2018) Evolved Cas9 variants with broad PAM compatibility and high DNA specificity. Nature, 556, 57–63.

30. Lee,J.K., Jeong,E., Lee,J., Jung,M., Shin,E., Kim,Y., Lee,K., Jung,I., Kim,D., Kim,S., et al. (2018) Directed evolution of CRISPR-Cas9 to increase its specificity. Nat. Commun., 9, 1–10.

31. Vakulskas,C.A., Dever,D.P., Rettig,G.R., Turk,R., Jacobi,A.M., Collingwood,M.A., Bode,N.M., McNeill,M.S., Yan,S., Camarena,J., et al. (2018) A high-fidelity Cas9 mutant delivered as a ribonucleoprotein complex enables efficient gene editing in human hematopoietic stem and progenitor cells. Nat. Med., 24, 1216–1224.

32. Tsai,S.Q., Nguyen,N.T., Malagon-Lopez,J., Topkar,V.V., Aryee,M.J. and Joung,J.K. (2017) CIRCLE-seq: a highly sensitive in vitro screen for genome-wide CRISPR–Cas9 nuclease off-targets. Nat. Methods, 14, 607–614.

33. 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.

34. Chaudhari,H.G., Penterman,J., Whitton,H.J., Spencer,S.J., Flanagan,N., Lei Zhang,M.C., Huang,E., Khedkar,A.S., Toomey,J.M., Shearer,C.A., et al. (2020) Evaluation of Homology-Independent CRISPR-Cas9 Off-Target Assessment Methods. CRISPR J., 3, 440–453.

35. Schmid-Burgk,J.L., Gao,L., Li,D., Gardner,Z., Strecker,J., Lash,B. and Zhang,F. (2020) Highly Parallel Profiling of Cas9 Variant Specificity. Mol. Cell, 10.1016/j.molcel.2020.02.023.

36. 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.

37. Yarrington,R.M., Verma,S., Schwartz,S., Trautman,J.K. and Carroll,D. (2018) Nucleosomes inhibit target cleavage by CRISPR-Cas9 in vivo. Proc. Natl. Acad. Sci., 115, 9351–9358.

38. Fu,B.X.H., Smith,J.D., Fuchs,R.T., Mabuchi,M., Curcuru,J., Robb,G.B. and Fire,A.Z. (2019) Target-dependent nickase activities of the CRISPR–Cas nucleases Cpf1 and Cas9. Nat. Microbiol., 10.1038/s41564-019-0382-0.

39. Huston,N.C., Tycko,J., Tillotson,E.L., Wilson,C.J., Myer,V.E., Jayaram,H. and Steinberg,B.E. (2019) Identification of Guide-Intrinsic Determinants of Cas9 Specificity. CRISPR J., 2, 172–185.

40. Zhang,L., Rube,H.T., Vakulskas,C.A., Behlke,M.A., Bussemaker,H.J. and Pufall,M.A. (2020) Systematic in vitro profiling of off-target affinity, cleavage and efficiency for CRISPR enzymes. Nucleic Acids Res., 10.1093/nar/gkaa231.

41. Höijer,I., Johansson,J., Gudmundsson,S., Chin,C.-S., Bunikis,I., Häggqvist,S., Emmanouilidou,A., Wilbe,M., den Hoed,M., Bondeson,M.-L., et al. (2020) Amplification-free long-read sequencing reveals unforeseen CRISPR-Cas9 off-target activity. Genome Biol., 21, 290.

42. Murugan,K., Seetharam,A.S., Severin,A.J. and Sashital,D.G. (2020) CRISPR-Cas12a has widespread off-target and dsDNA-nicking effects. J. Biol. Chem., 10.1074/jbc.RA120.012933.

43. Chen,H., Choi,J. and Bailey,S. (2014) Cut Site Selection by the Two Nuclease Domains of the Cas9 RNA-guided Endonuclease. J. Biol. Chem., 289, 13284–13294.

44. Mohanraju,P., Oost,J., Jinek,M. and Swarts,D. (2018) Heterologous Expression and Purification of the CRISPR-Cas12a/Cpf1 Protein. BIO-Protoc., 8.

45. Pollard,J., Bell,S.D. and Ellington,A.D. (2000) Design, Synthesis, and Amplification of DNA Pools for In Vitro Selection. Curr. Protoc. Nucleic Acid Chem., 00, 9.2.1-9.2.23.

46. Gibson,D.G., Young,L., Chuang,R.-Y., Venter,J.C., Hutchison,C.A. and Smith,H.O. (2009) Enzymatic assembly of DNA molecules up to several hundred kilobases. Nat. Methods, 6, 343–345.

47. Anders,C. and Jinek,M. (2014) In Vitro Enzymology of Cas9. In Methods in Enzymology. Elsevier, Vol. 546, pp. 1–20.

48. Sundaresan,R., Parameshwaran,H.P., Yogesha,S.D., Keilbarth,M.W. and Rajan,R. (2017) RNA-Independent DNA Cleavage Activities of Cas9 and Cas12a. Cell Rep., 21, 3728–3739.

49. Dang,Y., Jia,G., Choi,J., Ma,H., Anaya,E., Ye,C., Shankar,P. and Wu,H. (2015) Optimizing sgRNA structure to improve CRISPR-Cas9 knockout efficiency. Genome Biol., 16, 280.

50. Wang,D., Zhang,C., Wang,B., Li,B., Wang,Q., Liu,D., Wang,H., Zhou,Y., Shi,L., Lan,F., et al. (2019) Optimized CRISPR guide RNA design for two high-fidelity Cas9 variants by deep learning. Nat. Commun., 10, 1–14.

51. Oppenheim,A. (1981) Separation of closed circular DNA from linear DNA by electrophoresis in two dimensions in agarose gels. Nucleic Acids Res., 9, 6805–6812.

52. Liu,M.-S., Gong,S., Yu,H.-H., Jung,K., Johnson,K.A. and Taylor,D.W. (2020) Engineered CRISPR/Cas9 enzymes improve discrimination by slowing DNA cleavage to allow release of off-target DNA. Nat. Commun., 11, 3576.

53. Liu,X., Homma,A., Sayadi,J., Yang,S., Ohashi,J. and Takumi,T. (2016) Sequence features associated with the cleavage efficiency of CRISPR/Cas9 system. Sci. Rep., 6, 19675.

54. Sternberg,S.H., Redding,S., Jinek,M., Greene,E.C. and Doudna,J.A. (2014) DNA interrogation by the CRISPR RNA-guided endonuclease Cas9. Nature, 507, 62–67.

55. Anderson,E.M., Haupt,A., Schiel,J.A., Chou,E., Machado,H.B., Strezoska,Ž., Lenger,S., McClelland,S., Birmingham,A., Vermeulen,A., et al. (2015) Systematic analysis of CRISPR–Cas9 mismatch tolerance reveals low levels of off-target activity. J. Biotechnol., 211, 56–65.

56. Wu,X., Scott,D.A., Kriz,A.J., Chiu,A.C., Hsu,P.D., Dadon,D.B., Cheng,A.W., Trevino,A.E., Konermann,S., Chen,S., et al. (2014) Genome-wide binding of the CRISPR endonuclease Cas9 in mammalian cells. Nat. Biotechnol., 32, 670–676.

57. Boyle,E.A., Andreasson,J.O.L., Chircus,L.M., Sternberg,S.H., Wu,M.J., Guegler,C.K., Doudna,J.A. and Greenleaf,W.J. (2017) High-throughput biochemical profiling reveals sequence determinants of dCas9 off-target binding and unbinding. Proc. Natl. Acad. Sci., 114, 5461–5466.

58. Jones,D.L., Leroy,P., Unoson,C., Fange,D., Ćurić,V., Lawson,M.J. and Elf,J. (2017) Kinetics of dCas9 target search in Escherichia coli. Science, 357, 1420–1424.

59. Fukui,K. (2010) DNA Mismatch Repair in Eukaryotes and Bacteria. J. Nucleic Acids, 10.4061/2010/260512.

60. Kuzminov,A. (2001) Single-strand interruptions in replicating chromosomes cause double-strand breaks. Proc. Natl. Acad. Sci., 98, 8241–8246.

61. Vriend,L.E.M. and Krawczyk,P.M. (2017) Nick-initiated homologous recombination: Protecting the genome, one strand at a time. DNA Repair, 50, 1–13.

62. Vrtis,K.B., Dewar,J.M., Chistol,G., Wu,R.A., Graham,T.G.W. and Walter,J.C. (2020) Single-Strand DNA Breaks Cause Replisome Disassembly. bioRxiv, 10.1101/2020.08.17.254235.

63. Maizels,N., Zhang,Y. and Davis,L. (2021) Pathways and signatures of mutagenesis at targeted DNA nicks. bioRxiv, 10.1101/2021.01.08.425852.

64. Singh,D., Wang,Y., Mallon,J., Yang,O., Fei,J., Poddar,A., Ceylan,D., Bailey,S. and Ha,T. (2018) Mechanisms of improved specificity of engineered Cas9s revealed by single-molecule FRET analysis. Nat. Struct. Mol. Biol., 25, 347–354.

65. Okafor,I.C., Singh,D., Wang,Y., Jung,M., Wang,H., Mallon,J., Bailey,S., Lee,J.K. and Ha,T. (2019) Single molecule analysis of effects of non-canonical guide RNAs and specificity-enhancing mutations on Cas9-induced DNA unwinding. Nucleic Acids Res., 47, 11880–11888.

66. Raper,A.T., Stephenson,A.A. and Suo,Z. (2018) Functional Insights Revealed by the Kinetic Mechanism of CRISPR/Cas9. J. Am. Chem. Soc., 140, 2971–2984.

67. 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.

68. Zhu,X., Clarke,R., Puppala,A.K., Chittori,S., Merk,A., Merrill,B.J., Simonović,M. and Subramaniam,S. (2019) Cryo-EM structures reveal coordinated domain motions that govern DNA cleavage by Cas9. Nat. Struct. Mol. Biol., 26, 679–685.

69. Kim,D., Kim,J., Hur,J.K., Been,K.W., Yoon,S. and Kim,J.-S. (2016) Genome-wide analysis reveals specificities of Cpf1 endonucleases in human cells. Nat. Biotechnol., 34, 863–868.

70. Kleinstiver,B.P., Tsai,S.Q., Prew,M.S., Nguyen,N.T., Welch,M.M., Lopez,J.M., McCaw,Z.R., Aryee,M.J. and Joung,J.K. (2016) Genome-wide specificities of CRISPR-Cas Cpf1 nucleases in human cells. Nat. Biotechnol., 34, 869–874.

71. Kim,H.K., Song,M., Lee,J., Menon,A.V., Jung,S., Kang,Y.-M., Choi,J.W., Woo,E., Koh,H.C., Nam,J.-W., et al. (2017) In vivo high-throughput profiling of CRISPR–Cpf1 activity. Nat. Methods, 14, 153–159.

72. McMahon,S.A., Zhu,W., Graham,S., Rambo,R., White,M.F. and Gloster,T.M. (2020) Structure and mechanism of a Type III CRISPR defence DNA nuclease activated by cyclic oligoadenylate. Nat. Commun., 11, 1–11.

73. Yan,W.X., Hunnewell,P., Alfonse,L.E., Carte,J.M., Keston-Smith,E., Sothiselvam,S., Garrity,A.J., Chong,S., Makarova,K.S., Koonin,E.V., et al. (2019) Functionally diverse type V CRISPR-Cas systems. Science, 363, 88–91.

TABLES, FIGURES AND LEGENDS

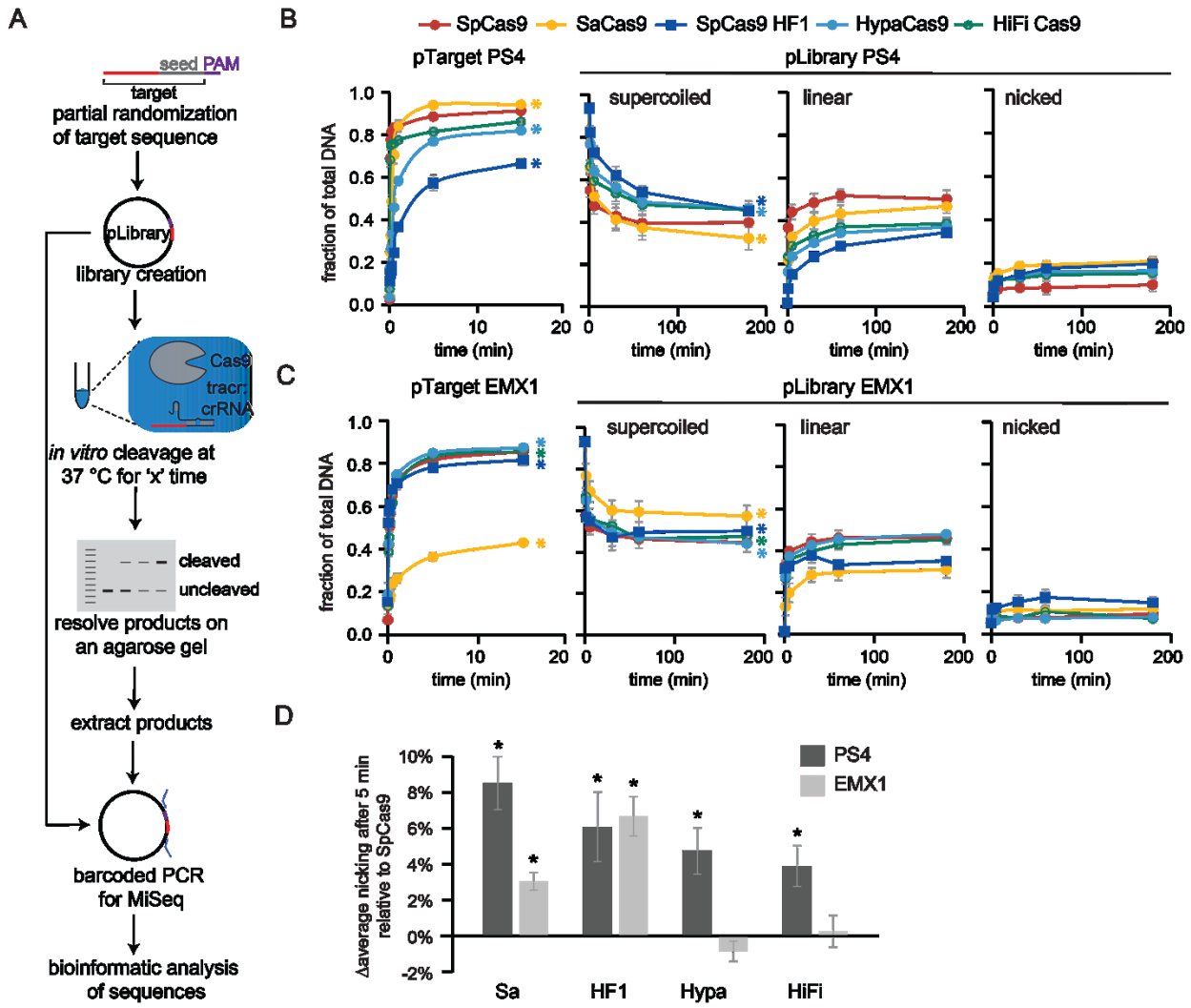


Figure 1. Systematic *in vitro* analysis of Cas9 mismatch tolerance.

(A) Outline and workflow of the *in vitro* pLibrary cleavage assay.

(B, C) Overall cleavage of pTarget and pLibrary (B) PS4 and (C) EMX1 by Cas9 plotted against time. Plot for pTarget shows appearance of linearized pool over time and plots for pLibrary show the decrease in supercoiled (nSC) pool and appearance of nicked (n) and linear (li) pools over time. The 0 time point is the quantification of the negative control pLibrary (i.e. pLibrary run on a gel after preparation as represented in Fig. S1B). The time points for pTarget cleavage are 5 sec, 10 sec, 15 sec, 30 sec, 1 min, 5 min, 15 min and the time points for pLibrary cleavage are 1 min, 5 min, 30 min, 60 min and 180 min. Values plotted represent an average of three replicates for both pTarget and pLibrary. Error bars are SEM. * $P < 0.05$, Student's t-test of for rates of cleavage of pTarget and pLibrary compared to SpCas9.

(D) Average accumulation of pLibrary nicked pool for Cas9 variants compared to SpCas9 over time. Values plotted represent an average of three replicates. Error bars are SEM. * $P < 0.05$, Student's t-test compared to SpCas9.
Sa = SaCas9, HF1 = SpCas9 HF1, Hypa = HypaCas9, HiFi = Alt-R ® S.p. HiFi Cas9.

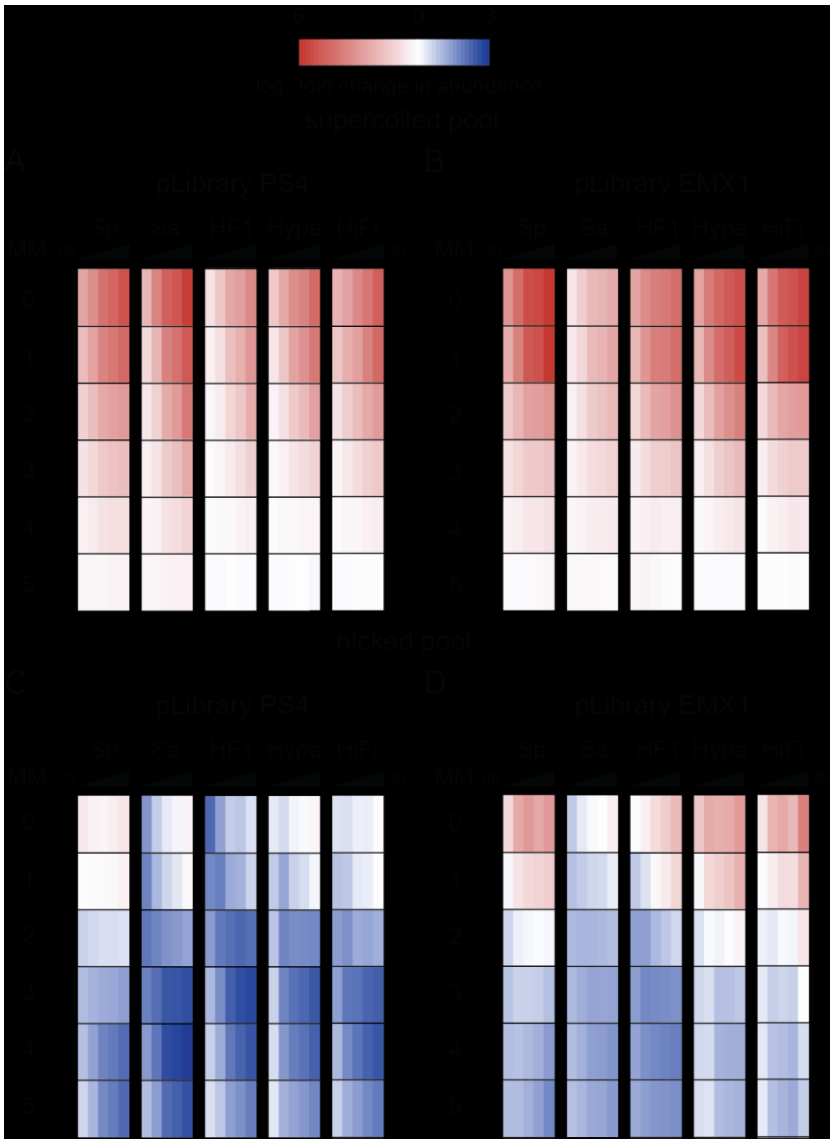


Figure 2. Cas9 cleavage activity against target sequences with different mismatches in pLibrary.

Log_2-fold change in abundance relative to control of target sequences containing different number of mismatches in the (A, B) supercoiled pool and (C, D) nicked pool from pLibrary (A, C) PS4 and (B, D) EMX1 when subjected to cleavage by different Cas9 variants. The time points for pLibrary cleavage are 1, 5, 30, 60 and 180 min. The color gradient represents sequences that were depleted (red), unchanged (white), or enriched (blue) relative to the control. Values plotted represent an average of two replicates. MM = mismatch, Sp = SpCas9, Sa = SaCas9, HF1 = SpCas9 HF1, Hypa = HypaCas9, HiFi = Alt-R ® S.p. HiFi Cas9.

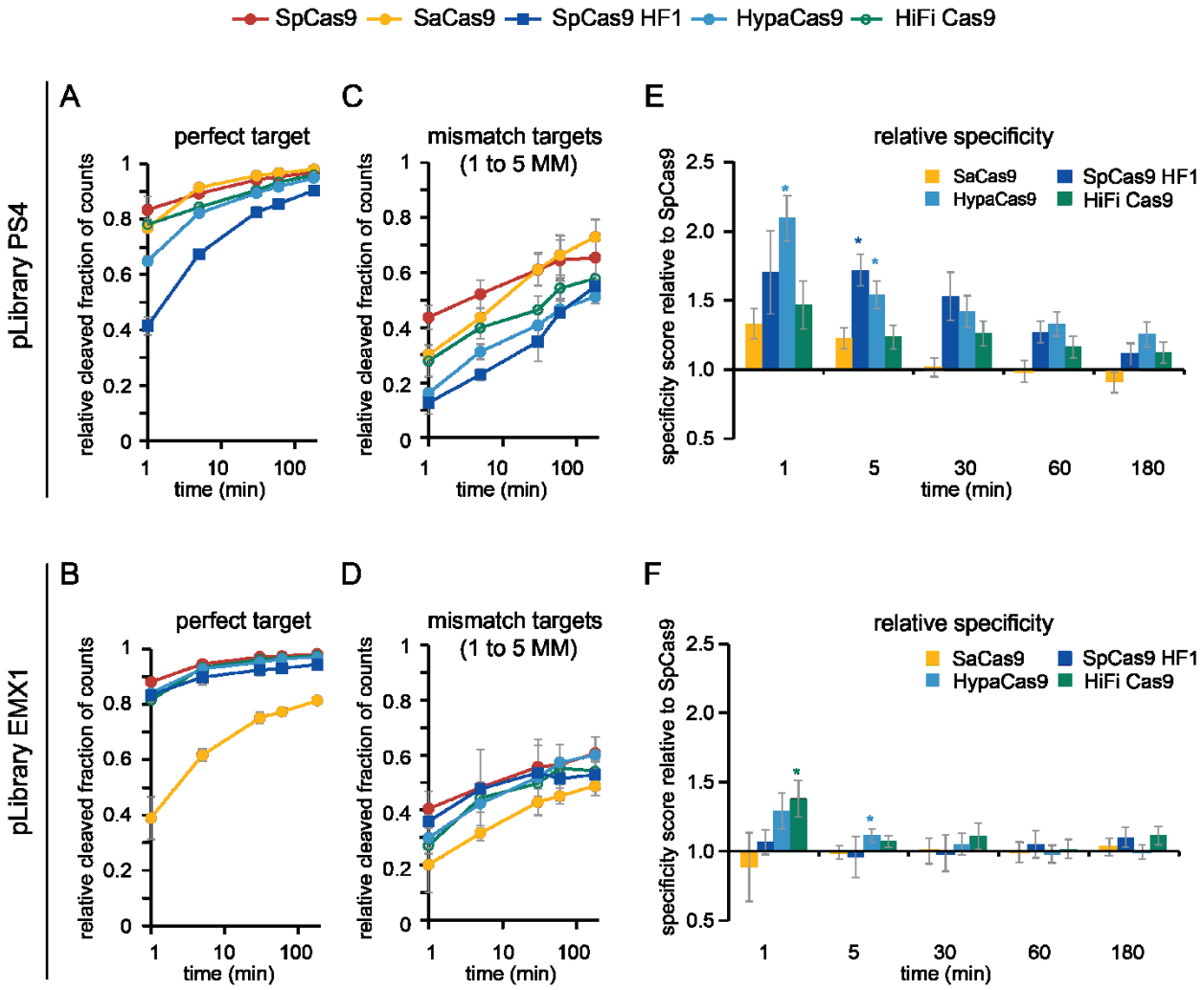


Figure 3. Specificity scores for Cas9 variants.

(A-D) The fractions of the (A, B) perfect target and (C, D) target sequences with 1 to 5 mismatches (MM) that were cleaved by Cas9 variants are plotted versus time for pLibrary (A, C) PS4 and (B, D) EMX1.

(E-F) Specificity scores for each Cas9 variant cleavage of pLibraries (E) PS4 and (F) EMX1 plotted relative to SpCas9 over the time course of the assay (see methods – HTS analysis). Values plotted represent an average of two replicates. Error bars are propagation of SEM. * $P < 0.05$, Student's t-test compared to SpCas9.

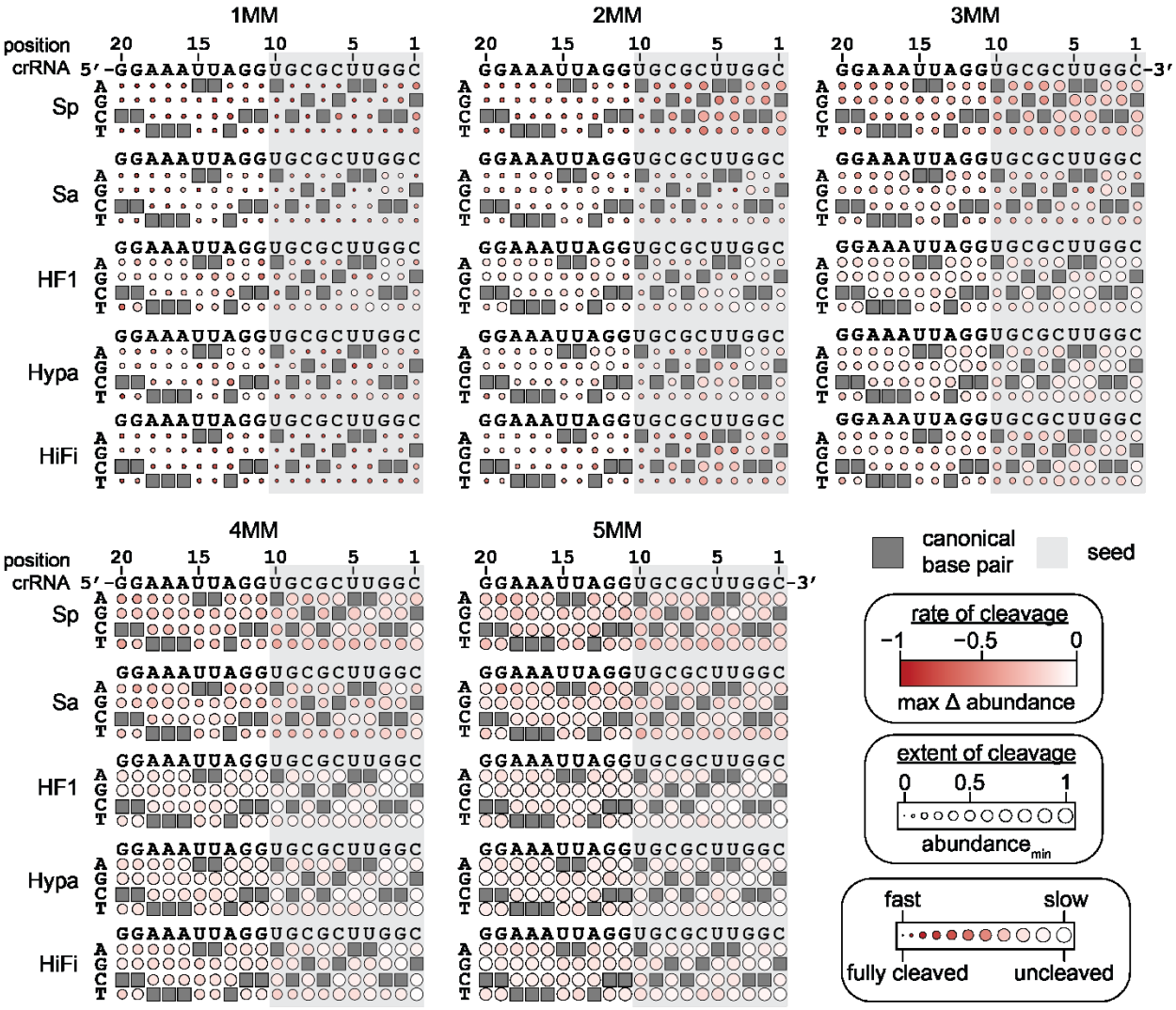


Figure 4. Sequence determinants of Cas9 cleavage activity for pLibrary PS4.

Heatmaps showing the max Δ abundance and abundance$_{min}$ of different mismatched sequences over time for the supercoiled pool in pLibrary PS4 upon cleavage by Cas9 variants. The position of nucleotides in the targeting region of the crRNA and the sequence are indicated on the top. The nucleotides on the left side of the heatmaps indicate the potential base pair or mismatches formed. The crRNA-complementary nucleotides are marked by grey boxes in the heatmap which result in canonical base pairs. The PAM-proximal “seed” sequence is highlighted by the light grey box. The color gradient indicates sequences that were relatively depleted (red) or unchanged (white). Extent of cleavage is represented as the bubble size and varies between 0 to 1. Values plotted represent an average of two replicates. MM = mismatch

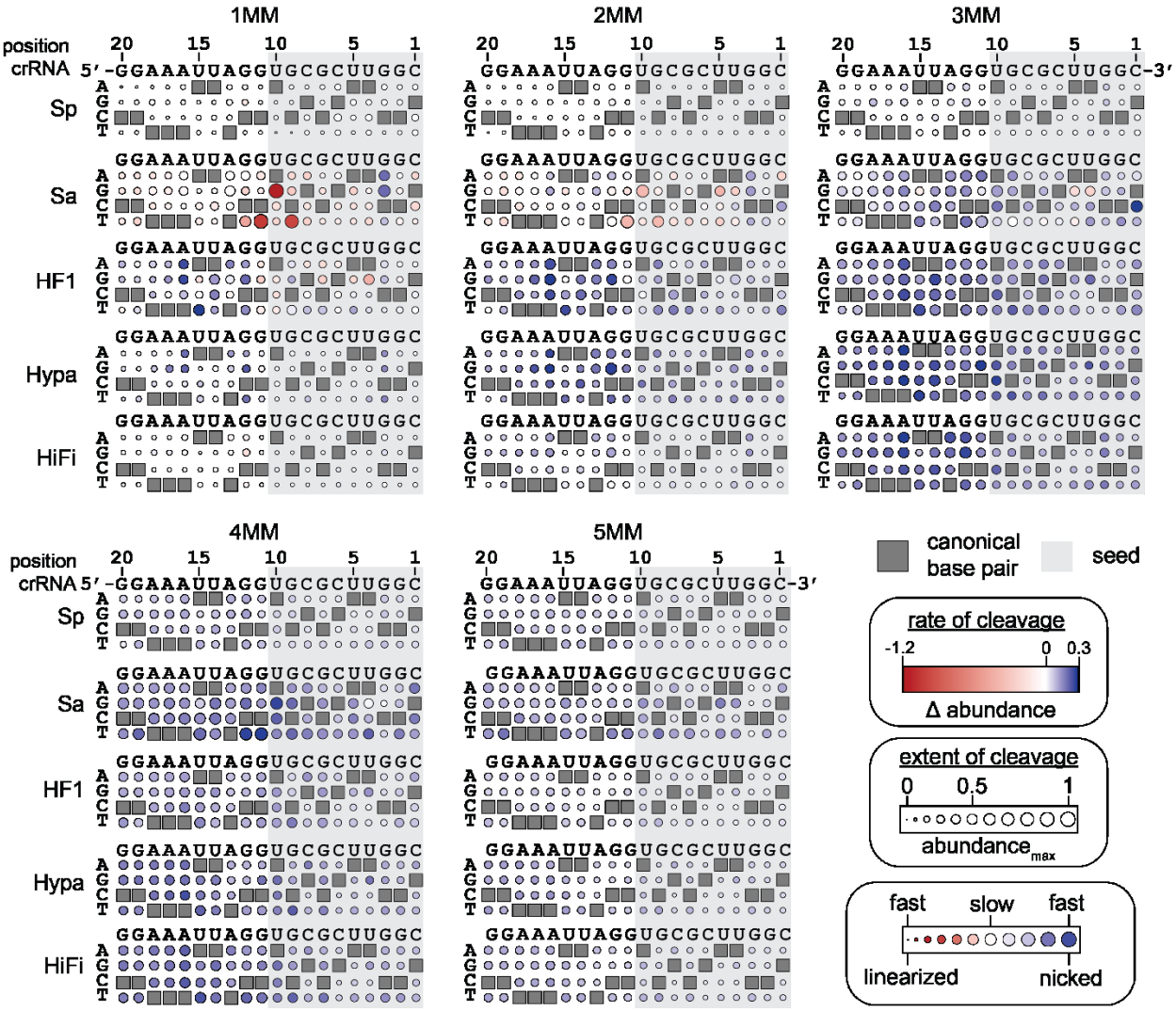


Figure 5. Sequence determinants of Cas9 nicking defect for pLibrary PS4.

Heatmaps showing the Δ abundance and abundance$_{max}$ of different mismatched sequences over time for the nicked pool in pLibrary PS4 upon cleavage by Cas9 variants. The position of nucleotides in the targeting region of the crRNA and the sequence are indicated on the top. The nucleotides on the left side of the heatmaps indicate the potential base pair or mismatches formed. The crRNA-complementary nucleotides are marked by grey boxes in the heatmap which result in canonical base pairs. The PAM-proximal "seed" sequence is highlighted by the light grey box. The color gradient represents sequences that were depleted (red), unchanged (white), or enriched (blue) relative to the control. The extent of nicking is represented as the bubble size and varies between 0 to 1. Values plotted represent an average of two replicates. MM = mismatch

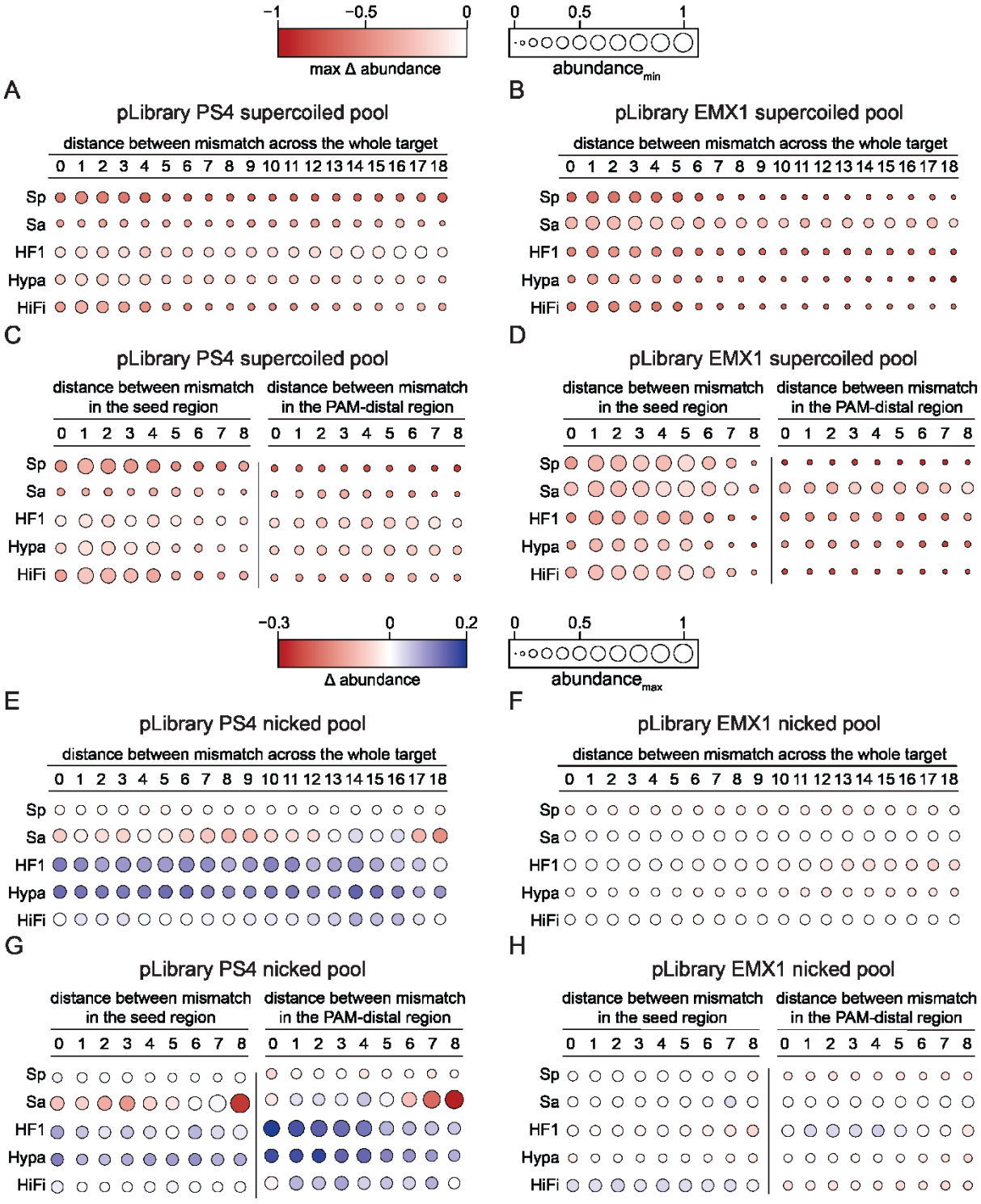


Figure 6. Effect of double mismatches in the target sequence on Cas9 cleavage activity.

Heatmaps plotting the max Δ abundance and abundance$_{min}$ or Δ abundance and abundance$_{max}$ for target sequences with two mismatches in the (A - D) supercoiled pools and (E - H) nicked pools, respectively.

Heatmaps show the effect of two mismatches as a function of distance between the two mismatches (A, B, E, F) across the whole target sequence or (C, D, G, H) in the seed or PAM-distal regions over time upon cleavage by Cas9 variants. The Δ max change or Δ abundance scale for supercoiled and nicked pools are indicated on the top. $Abundance_{min}$ and $abundance_{max}$ are represented as the bubble size and varies between 0 to 1. Values plotted represent an average of two replicates.
MM = mismatch, Sp = SpCas9, Sa = SaCas9, HF1 = SpCas9 HF1, Hypa = HypaCas9, HiFi = Alt-R ® S.p. HiFi Cas9.

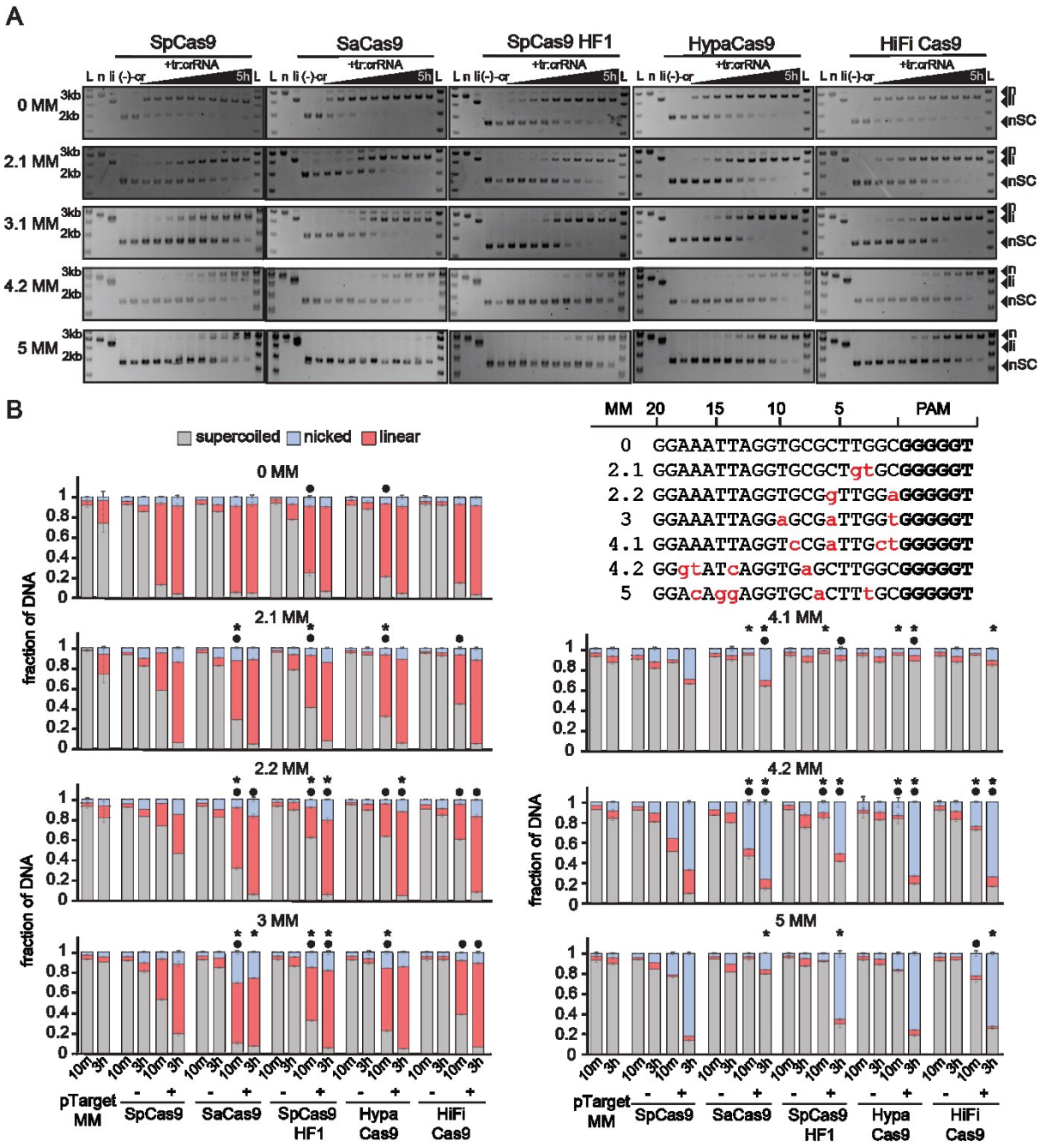


Figure 7. Cas9 variants have different cleavage activities against mismatched targets.

(A) Representative agarose gels showing cleavage of a negatively supercoiled (nSC) plasmid containing the perfect target (0 MM) or mismatched (2 to 5 MM) target over a time course by Cas9 variants, resulting in linear (li) and/or nicked (n) products. Time points at which the samples were collected are 15 sec, 30 sec, 1 min, 2 min, 5 min, 15 min, 30 min, 1 h, 3 h, and 5 h. tr:crRNA = tracrRNA:crRNA.

All controls were performed under the same conditions as the longest time point for the experimental samples. Controls: (-) = pTarget or pLibrary alone incubated at 37 °C for the longest time point in the assay (5 h); (-cr) = pTarget or pLibrary incubated with Cas9 only at 37 °C for the longest time point in the assay (5 h); n = Nt.BspQI nicked pUC19; li = BsaI-HF linearized pUC19

(B) Quantification of supercoiled, linear and nicked pools from cleavage of perfect or fully crRNA-complementary (0 MM) and mismatched (2 to 5 MM) target plasmid by Cas9 after 10 minutes and 3 hours. pTarget MM indicates target plasmid (0, 2 to 5 MM) alone incubated at 37 °C for the time points indicated. Target sequences tested are listed with PAM (bold) and mismatches (lowercase and red) indicated. (-) indicates a cleavage reaction with the target plasmid and Cas9 only, and (+) indicates a cleavage reaction with the target plasmid, Cas9 and cognate tracrRNA:crRNA. Values plotted represent an average of three replicates. Error bars are SEM. * or ● indicate $P < 0.05$ based on Student's t test comparing the fraction of nicked product (*) or fraction of linear product (●) between each Cas9 variant and SpCas9.

Supplementary Materials for

Systematic *in vitro* specificity profiling reveals nicking defects in natural and engineered CRISPR-Cas9 variants

Karthik Murugan [1, 2, 4], Shravanti K. Suresh [1], Arun S. Seetharam [3], Andrew J. Severin [3] and Dipali G. Sashital [1, 2 *]

Affiliations:

[1] Roy J. Carver Department of Biochemistry, Biophysics & Molecular Biology, Iowa State University, Ames, IA 50011, USA

[2] Molecular, Cellular, and Developmental Biology Interdepartmental Program, Iowa State University, Ames, IA 50011, USA

[3] Genome Informatics Facility, Office of Biotechnology, Iowa State University, Ames, IA 50011, USA

[4] Present address: Integrated DNA Technologies Inc., Coralville, IA 52241, USA

*Correspondence:

* To whom correspondence should be addressed. Tel: +1 (515)-294-5121; Fax: +1 (515)-294-7629 Email: sashital@iastate.edu

This file includes:

Figures S1 to S6

Table S1: List of Oligonucleotides

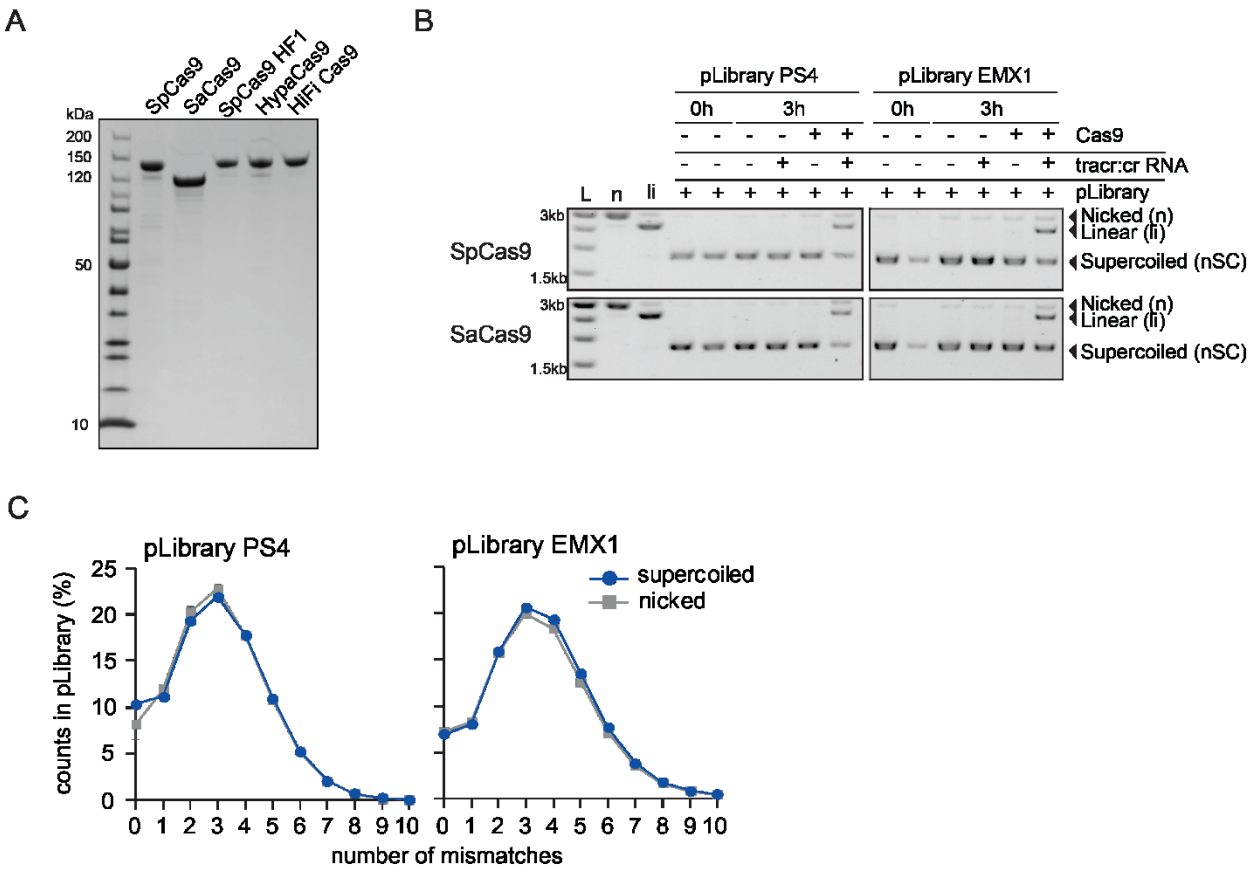


Figure S1: Components of *in vitro* pLibrary cleavage assay.

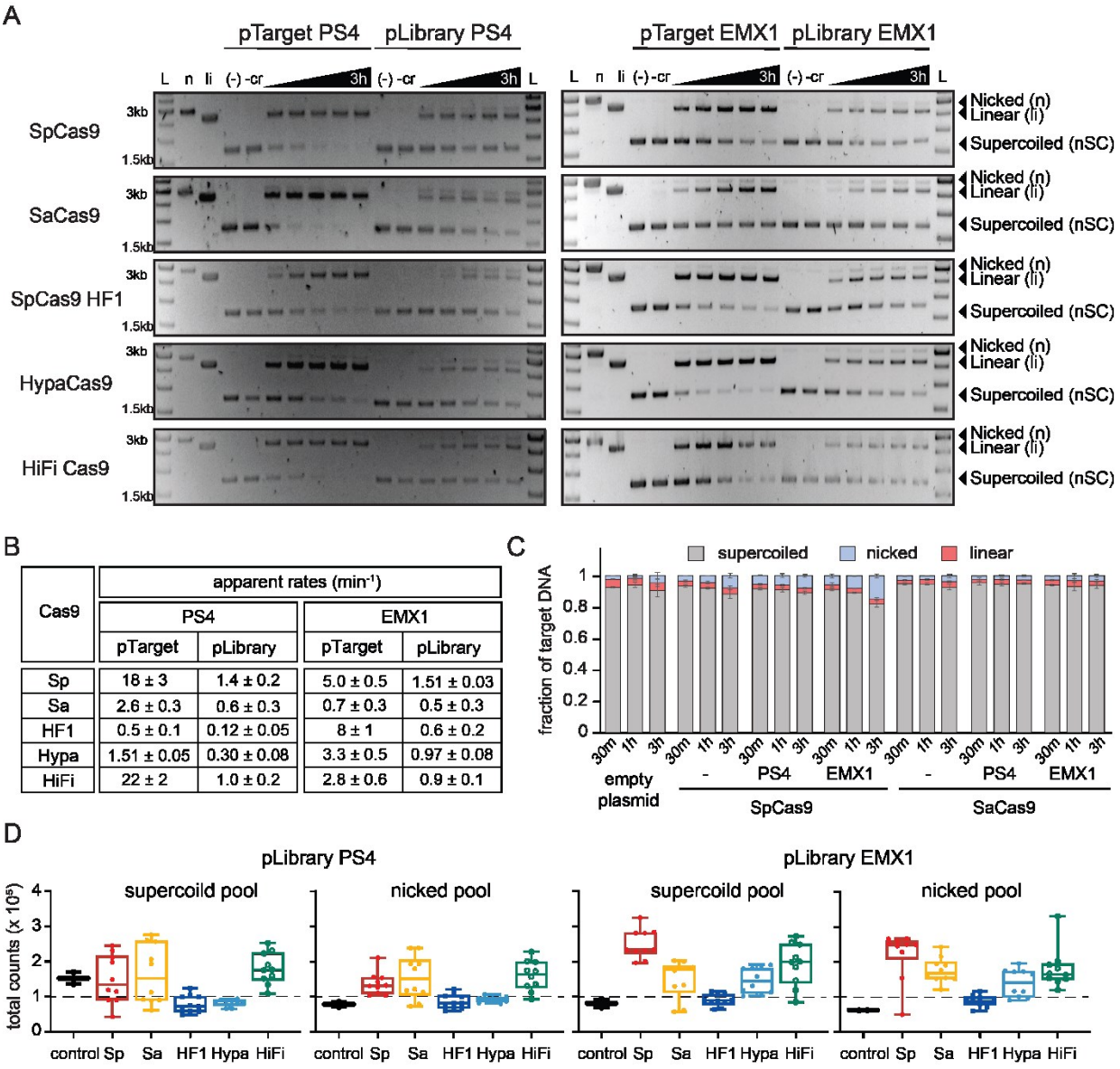
(A) Gradient SDS-PAGE of purified Cas9 variants visualized using Coomassie stain.

(B) Agarose gels showing cleavage activity of SpCas9 and SaCas9 against negatively supercoiled (nSC) plasmid library (pLibrary) only in the presence of cognate tracrRNA:crRNA, resulting in linear (li) and/or nicked (n) products after 3 hours of incubation at 37 °C. Two pLibrary replicates are labelled as 0 h and were used as the negative control pLibrary for analysis shown in panel (C).

All controls were performed under the same conditions as indicated for the longest time point for the experimental samples. n = Nt.BspQI nicked pUC19; li = BsaI-HF linearized pUC19

(C) Mismatch distribution of the supercoiled and nicked fractions from pLibrary 0 h controls (see methods section – Plasmid and nucleic acid preparation). A clear nicked plasmid band was not visible on the gel with SYBR Safe or RedSafe staining. However, a band excised from the gel in the region where the nicked fraction would run produced a similar mismatch distribution to the supercoiled fraction when subjected to HTS. This indicated the presence of trace amounts of nicked pLibrary prior to Cas9 cleavage. The mismatch distribution of this trace amount of nicked plasmid is similar to the supercoiled

fraction which means there is no sequence bias in this nicked pool and any change in the nicked pool would be a result of Cas9-mediated cleavage activity.



Cas9	apparent rates (min⁻¹)			
	PS4		EMX1	
	pTarget	pLibrary	pTarget	pLibrary
Sp	18 ± 3	1.4 ± 0.2	5.0 ± 0.5	1.51 ± 0.03
Sa	2.6 ± 0.3	0.6 ± 0.3	0.7 ± 0.3	0.5 ± 0.3
HF1	0.5 ± 0.1	0.12 ± 0.05	8 ± 1	0.6 ± 0.2
Hypa	1.51 ± 0.05	0.30 ± 0.08	3.3 ± 0.5	0.97 ± 0.08
HiFi	22 ± 2	1.0 ± 0.2	2.8 ± 0.6	0.9 ± 0.1

Figure S2: Cleavage activity of Cas9 on pLibrary PS4 and EMX1.

(A) Representative agarose gel showing time course cleavage of negatively supercoiled (nSC) plasmid containing a fully matched PS4 and EMX1 target (left side on each gel) and the respective plasmid library (right side on each gel) by Cas9 variants, resulting in linear (li) and/or nicked (n) products. Time points at which the samples were collected are 1 min, 5 min, 30 min, 1 h, and 3 h.

All controls were performed under the same conditions as the longest time point for the experimental samples. Controls: (-) = pTarget or pLibrary alone incubated at 37 °C for the longest time point in the assay (3 h); (-cr) = pTarget or pLibrary incubated with Cas9 only at 37 °C for the longest time point in the assay (3 h); n = Nt.BspQI nicked pUC19; li = BsaI-HF linearized pUC19

(B) Apparent rates of cleavage of pTargets and pLibraries PS4 and EMX1 by Cas9. Time points used to calculate rates were 5, 10, 15, 30, 60, 300 and 1800 s for pTarget and 1, 5, 30, 60 and 180 min for pLibrary. Values are an average of three replicates. Error bars are SEM.

(C) Quantification of supercoiled, linear and nicked fractions of empty plasmid upon incubation of Cas9 without and with different tracrRNA:crRNAs at the indicated time points at 37 °C. Values plotted represent an average of three replicates. Error bars are SEM.

(D) Box plots showing total reads from the high-throughput sequencing runs for each Cas9 variant plotted by library and product pool.

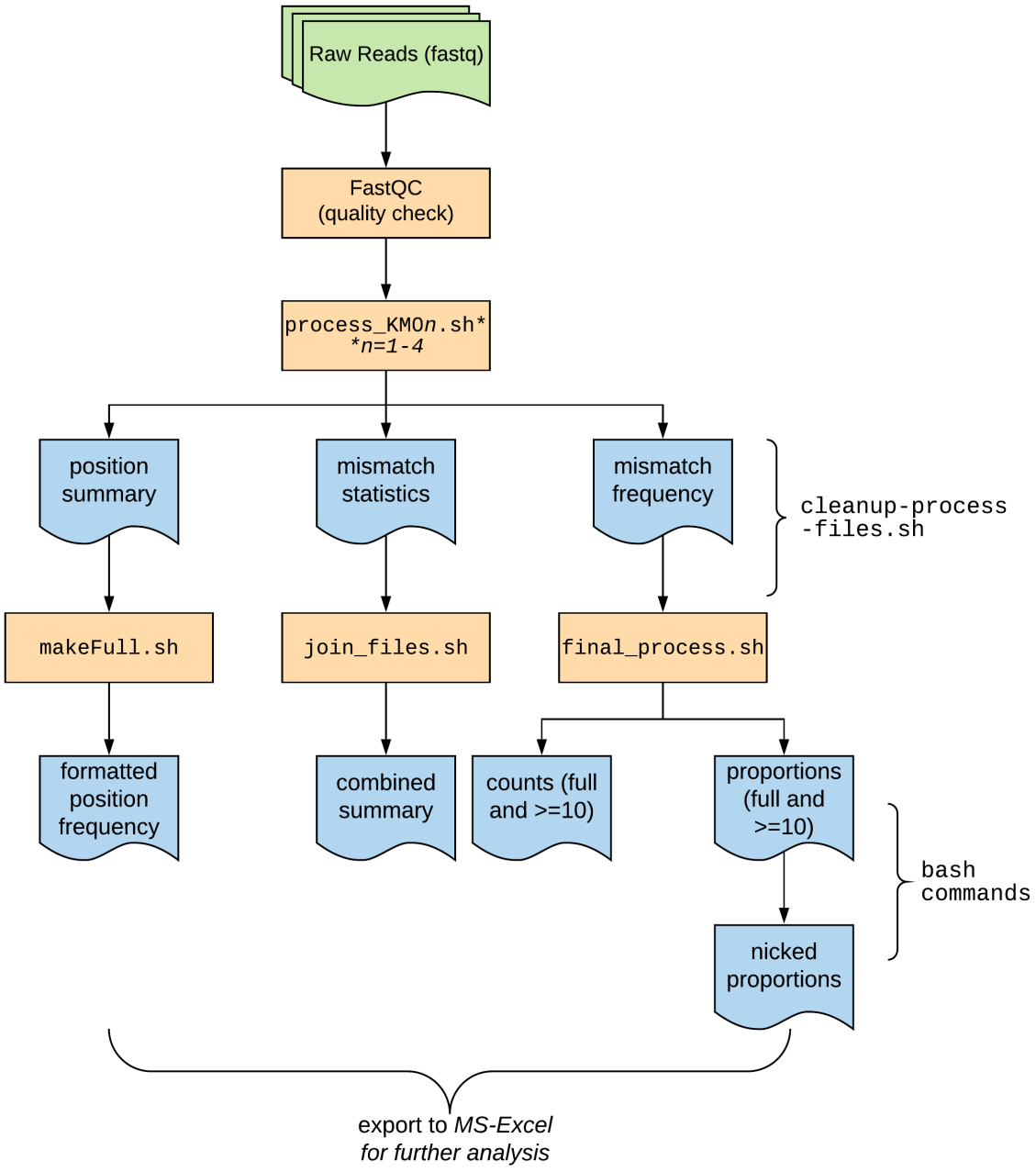


Figure S3: Workflow of the bioinformatic analysis of the HTS data.

Target sequences were extracted from the HTS data using custom scripts, previously used to study Cas12a (Murugan et al., 2020). The mismatch number and position were determined and tables reporting the counts and fractions of each mismatched target sequence were generated for further analysis (see methods – HTS data analysis)

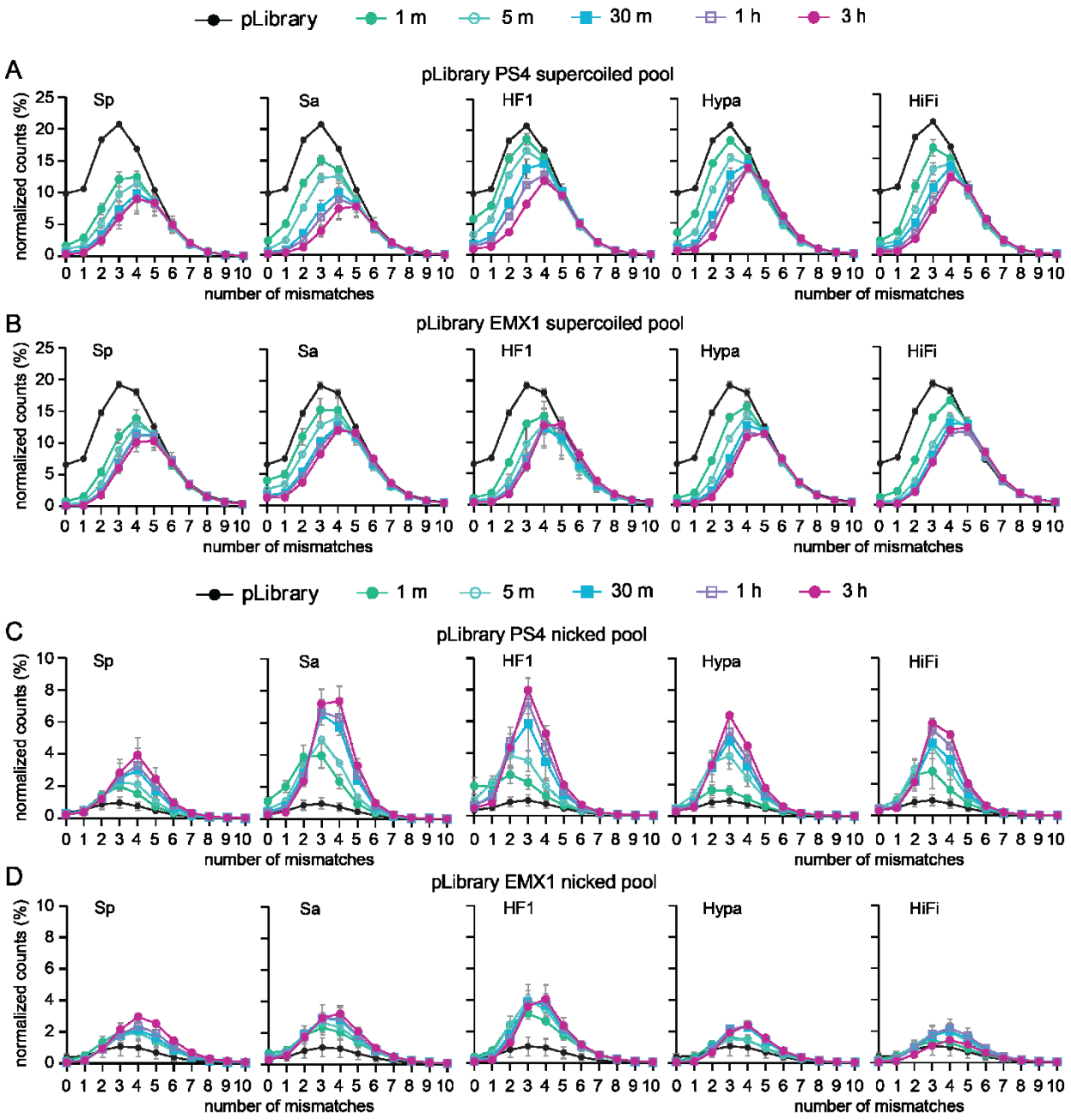


Figure S4. Mismatch distribution curves for Cas9 cleavage activity against pLibrary.

Mismatch distribution of (A, B) supercoiled pool and (C, D) nicked pool from pLibrary (A, C) PS4 and (B, D) EMX1 when subject to cleavage by different Cas9 variants. Depletion of target sequences from the supercoiled pool indicates cleavage, and enrichment in the nicked pool indicates nicking. The decrease in nicked pool over time indicates linearization of target sequences. Values plotted represent an average of two replicates. Error bars are propagation of SEM.

Sp = SpCas9, Sa = SaCas9, HF1 = SpCas9 HF1, Hypa = HypaCas9, HiFi = Alt-R ® S.p. HiFi Cas9.

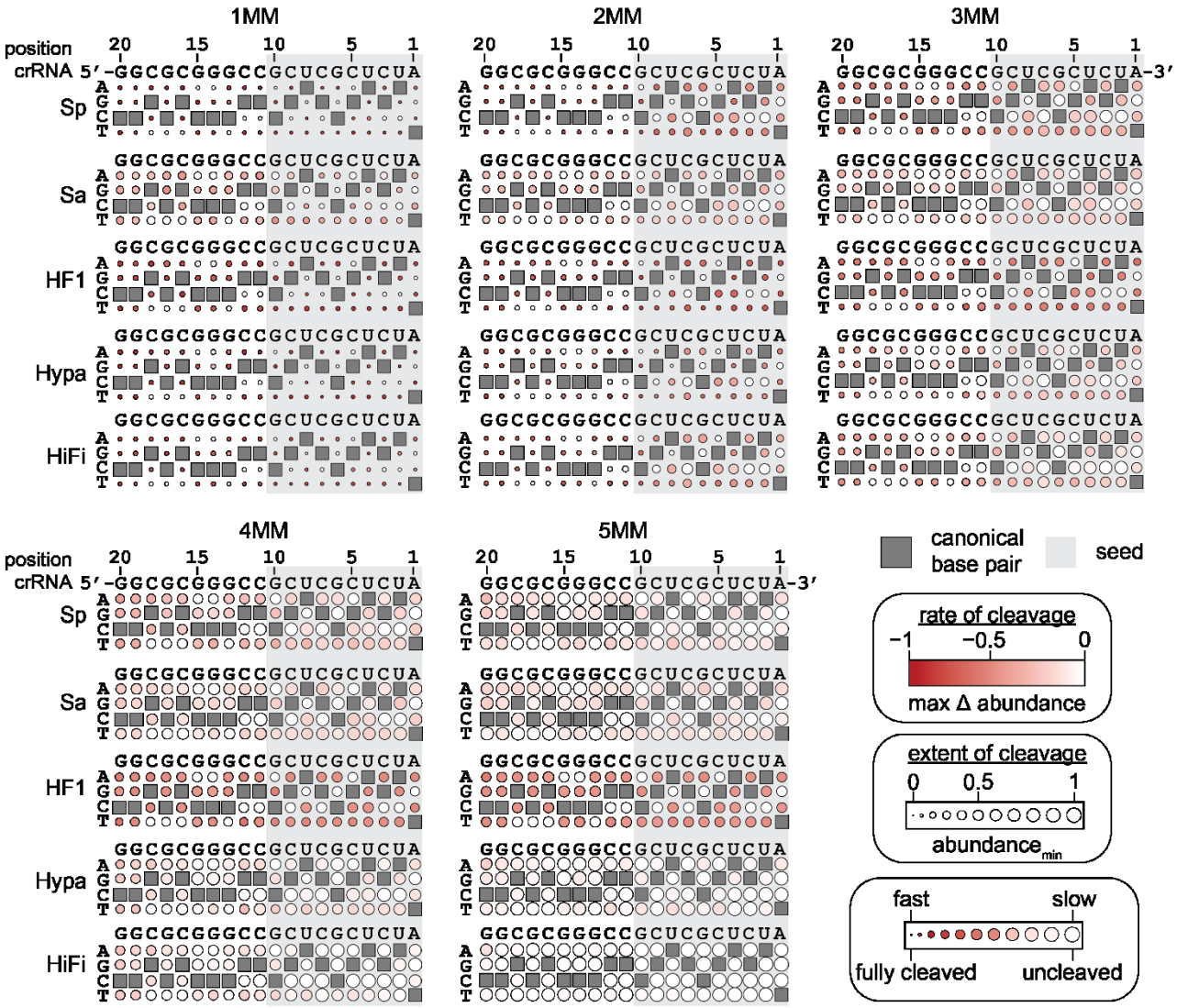


Figure S5. Sequence determinants of Cas9 cleavage activity for pLibrary EMX1.

Heatmaps showing the max Δ abundance and abundance$_{min}$ of different mismatched sequences over time for the supercoiled pool in pLibrary EMX1 upon cleavage by Cas9 variants. The position of nucleotides in the targeting region of the crRNA and the sequence are indicated on the top. The nucleotides on the left side of the heatmaps indicate the potential base pair or mismatches formed. The crRNA-complementary nucleotides are marked by grey boxes in the heatmap which result in canonical base pairs. The PAM-proximal “seed” sequence is highlighted by the light grey box. The color gradient indicates sequences that were relatively depleted (red) or unchanged (white). Extent of cleavage is represented as the bubble size and varies between 0 to 1. Values plotted represent an average of two replicates. MM = mismatch

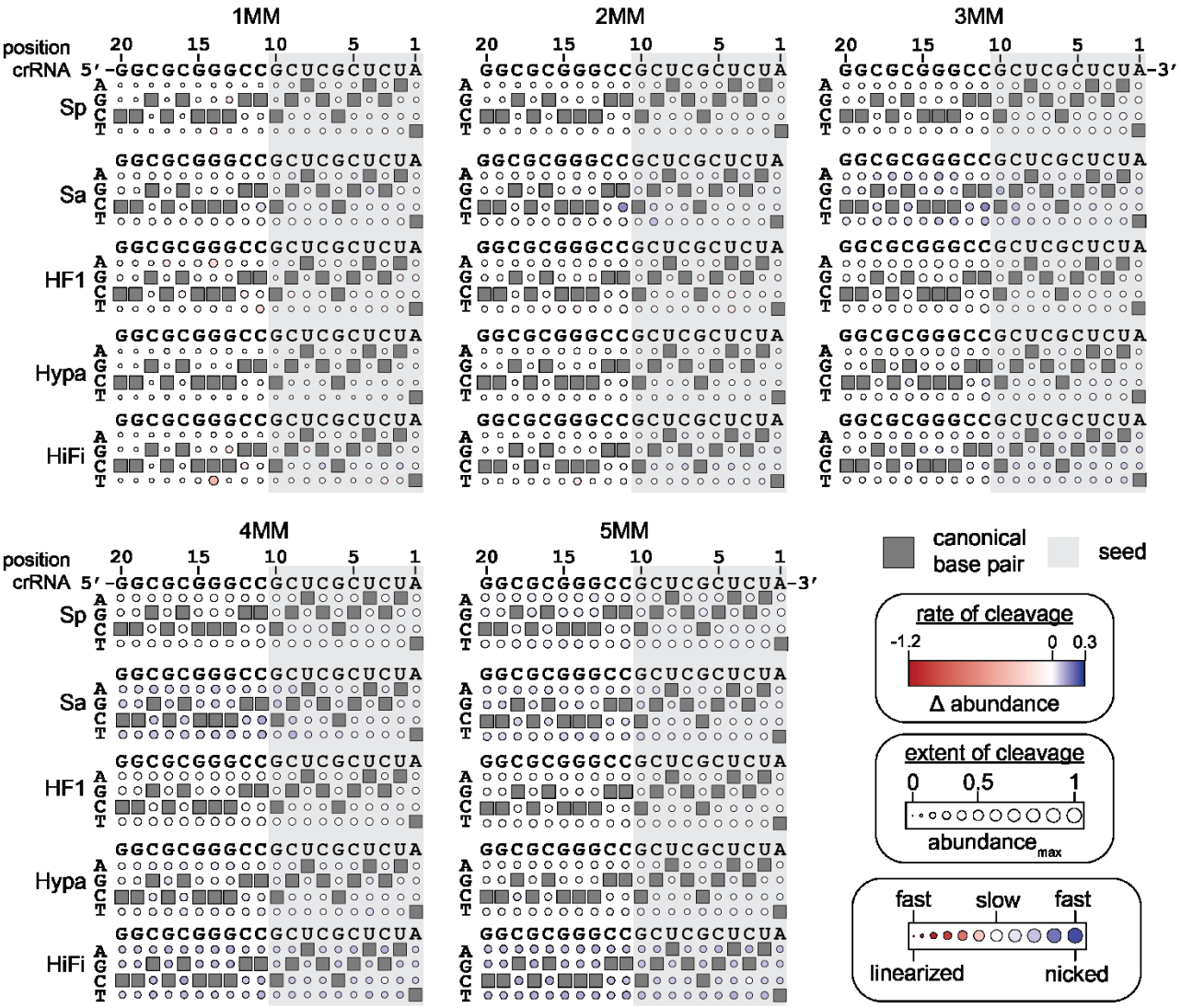


Figure S6. Sequence determinants of Cas9 nicking defect for pLibrary EMX1.

Heatmaps showing the Δ abundance and abundance$_{max}$ of different mismatched sequences over time for the nicked pool in pLibrary EMX1 upon cleavage by Cas9 variants. The position of nucleotides in the targeting region of the crRNA and the sequence are indicated on the top. The nucleotides on the left side of the heatmaps indicate the potential base pair or mismatches formed. The crRNA-complementary nucleotides are marked by grey boxes in the heatmap which result in canonical base pairs. The PAM-proximal "seed" sequence is highlighted by the light grey box. The color gradient represents sequences that were depleted (red), unchanged (white), or enriched (blue) relative to the control. The extent of nicking is represented as the bubble size and varies between 0 to 1. Values plotted represent an average of two replicates. MM = mismatch.

References:

Murugan, K., Seetharam, A.S., Severin, A.J., Sashital, D.G., 2020. CRISPR-Cas12a has widespread off-target and dsDNA-nicking effects. J. Biol. Chem. jbc.RA120.012933. https://doi.org/10.1074/jbc.RA120.012933

Table S1: List of Oligonucleotides

Key:
Bold = target sequence
underlined = PAM
lowercase = mismatch (MM)
RC = reverse complement

Sequence (5' to 3')	Notes
RNA	
GGAACCAUUCAAAACAGCAUAGCAAGUUAAAAUAAGGCUAGUCCGUUAUCAACUUGAAAAAGUGGCACCGAGUCGGUGCUUUUUUU	SpCas9 tracrRNA
AUUGUACUUAUACCUAAAAUUACAGAAUCUACUAAAACAAGGCAAAAUGCCGUGUUUAUCUCGUCAACUUGUUGGCGAGAUUUUU	SaCas9 tracrRNA
GGAAAUUAGGUGCGCUUGGCGUUUUAGAGCUAUGCUGUUUUG	SpCas9 crRNA for modified protospacer 4 (ps4) from Sp CRISPR
GGCGCGGGCCGCUCGCUCUAGUUUUAGAGCUAUGCUGUUUUG	SpCas9 crRNA for EMX1 gene target
GGAAAUUAGGUGCGCUUGGCGUUUUAGUACUCUGUAAUUUUAGGUAUGAGGUAGAC	SaCas9 crRNA for modified protospacer 4 (ps4) from Sp CRISPR
GGCGCGGGCCGCUCGCUCUAGUUUUAGUACUCUGUAAUUUUAGGUAUGAGGUAGAC	SaCas9 crRNA for EMX1 gene target
Target DNA oligonucleotides	
GCATTGCTGTACGAATCGTACAGGGTGCTTCAGGATGGAAATTAGGTGCGCTTGGCGGGGGGTTGGTCAAGCTCGGACATCGTGATTGATAATGCGATGC	Cas9 modified ps4 from Sp CRISPR - 99b target - ssoligo used for Gibson assembly with pUC19 – KMlib002 – pLibrary PS4
GCATTGCTGTACGAATCGTACAGGGTGCTTCAGGTTTAGGCGCGGGCCGCTCGCTCTAGGGGGGTGTCAAGCTCGGACATCGTGATTGATAATGCGATGC	EMX1 gene target - high GC % - 99b target - ssoligo used for Gibson assembly with pUC19 - KMlib003 – pLibrary EMX1
GCATTGCTGTACGAATCGTACAGGGTGCTTCAGGATGGAAATTAGGTGCGCTgtGCGGGGGGTTGGTCAAGCTCGGACATCGTGATTGATAATGCGATGC	Cas9 mismatched target ssoligo - off target for mod protospacer 4, pLibrary PS4 - 2 mismatches 1 (2.1 MM)
GCATTGCTGTACGAATCGTACAGGGTGCTTCAGGATGGAAATTAGGTGCGgTTGGaGGGGGGTTGGTCAAGCTCGGACATCGTGATTGATAATGCGATGC	Cas9 mismatched target ssoligo - off target for mod protospacer 4, pLibrary PS4 - 2 mismatches 2 (2.2 MM)
GCATTGCTGTACGAATCGTACAGGGTGCTTCAGGATGGAAATTAGGaGCGaTTGGtGGGGGGTTGGTCAAGCTCGGACATCGTGATTGATAATGCGATGC	Cas9 mismatched target ssoligo - off target for mod protospacer 4, pLibrary PS4 - 3 mismatches 1 (3.1 MM)
GCATTGCTGTACGAATCGTACAGGGTGCTTCAGGATGGAAATTAGGTcCGaTTGctGGGGGGTTGGTCAAGCTCGGACATCGTGATTGATAATGCGATGC	Cas9 mismatched target ssoligo - off target for mod protospacer 4, pLibrary PS4 - 4 mismatches 1 (4.1 MM)
GCATTGCTGTACGAATCGTACAGGGTGCTTCAGGATGGgtATcAGGTGaGCTTGGCGGGGGGTTGGTCAAGCTCGGACATCGTGATTGATAATGCGATGC	Cas9 mismatched target ssoligo - off target for mod protospacer 4, pLibrary PS4 - 4 mismatches 2 (4.2 MM)

GCATTGCTGTACGAATCGTACAGGGTGCTTCAGGATGGAcAggAGGTGCaCTTtGCGGGGGGTTGGTCAAGCTCGGACATCGTGATTGATAATGCGATGC	Cas9 mismatched target ssoligo - off target for mod protospacer 4, pLibrary PS4 - 5 mismatches
Primers	
GTCAAGCTCGGACATCGTGATTGATAATGCGATGCACTGGCCGTCGTTTTACAACGTC	pUC19 plasmid library assembly - (between) M13 - vector amplification - common primer 1
CCTGAAGCACCCTGTACGATTCGTACAGCAATGCGTCATAGCTGTTTCCTGTGTGAAATTG	pUC19 plasmid library assembly - (between) M13 - vector amplification - common primer 2
TCGTCGGCAGCGTCAGATGTGTATAAGAGACAGGCATCGCATTATCAATCACGATGTC	for Nextera tagmentation - transposase adapter - forward
GTCTCGTGGGCTCGGAGATGTGTATAAGAGACAGGCATTGCTGTACGAATCGTACAGG	for Nextera tagmentation - transposase adapter - reverse
GCATTGCTGTACGAATCGTACAGGGTGCTTCAGGATGTTTACGGTTCGCGTGGtTTAtAGGTGCGTCAAGCTCGGACATCGTGATTGATAATGCGATGC	Cas12a mismatched target ssoligo - TTTA PAM - off target for mod protospacer 4, pLibrary PS4 - 2 mismatches 1 (2.2 MM)