SUPPLEMENTAL FIGURES

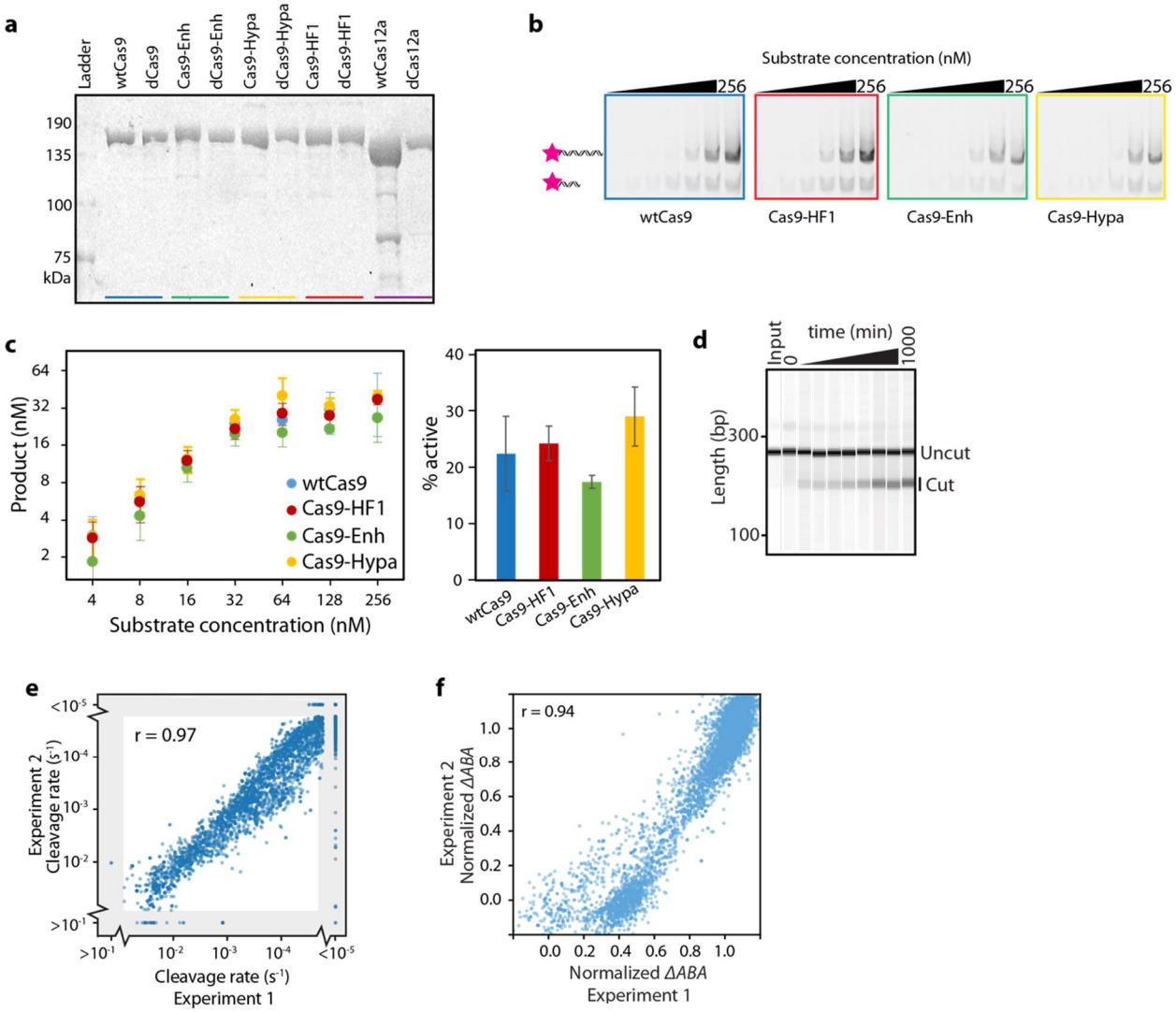

Figure S1. Biochemical characterization of CRISPR-Cas nucleases. (a) Coomassie-stained 10% SDS-PAGE gel of purified nucleases and their catalytically inactive variants. **(b)** Representative active site titration cleavage gels (10% TBE) and **(c)** quantification of three replicates for the indicated Cas9 variants. To determine the activity of each nuclease, 128 nM of the Cas9 RNP was incubated with 2-256nM of ATTO647N-labeled matched DNA (pink star) for 30 minutes. **(c, right)** The active nuclease concentration (mean ± SD; of at least three replicates) was determined from the concentration of product formed at 64-256 nM input DNA concentrations. **(d)** Time course of a wtCas9 nuclease reaction (sgRNA 2), as resolved by capillary electrophoresis. **(e)** Two independent wtCas12a NucleaSeq experiments show the same excellent reproducibility as for wtCas9 (see Figure 1G; Wald χ^2 test statistic, excluding gray regions). Gray regions indicate sequences with cleavage rates that exceed the dynamic range of the experiment. **(f)** Two independent CHAMP experiments show excellent reproducibility (dCas9 with sgRNA 1, Wald χ^2 test statistic).

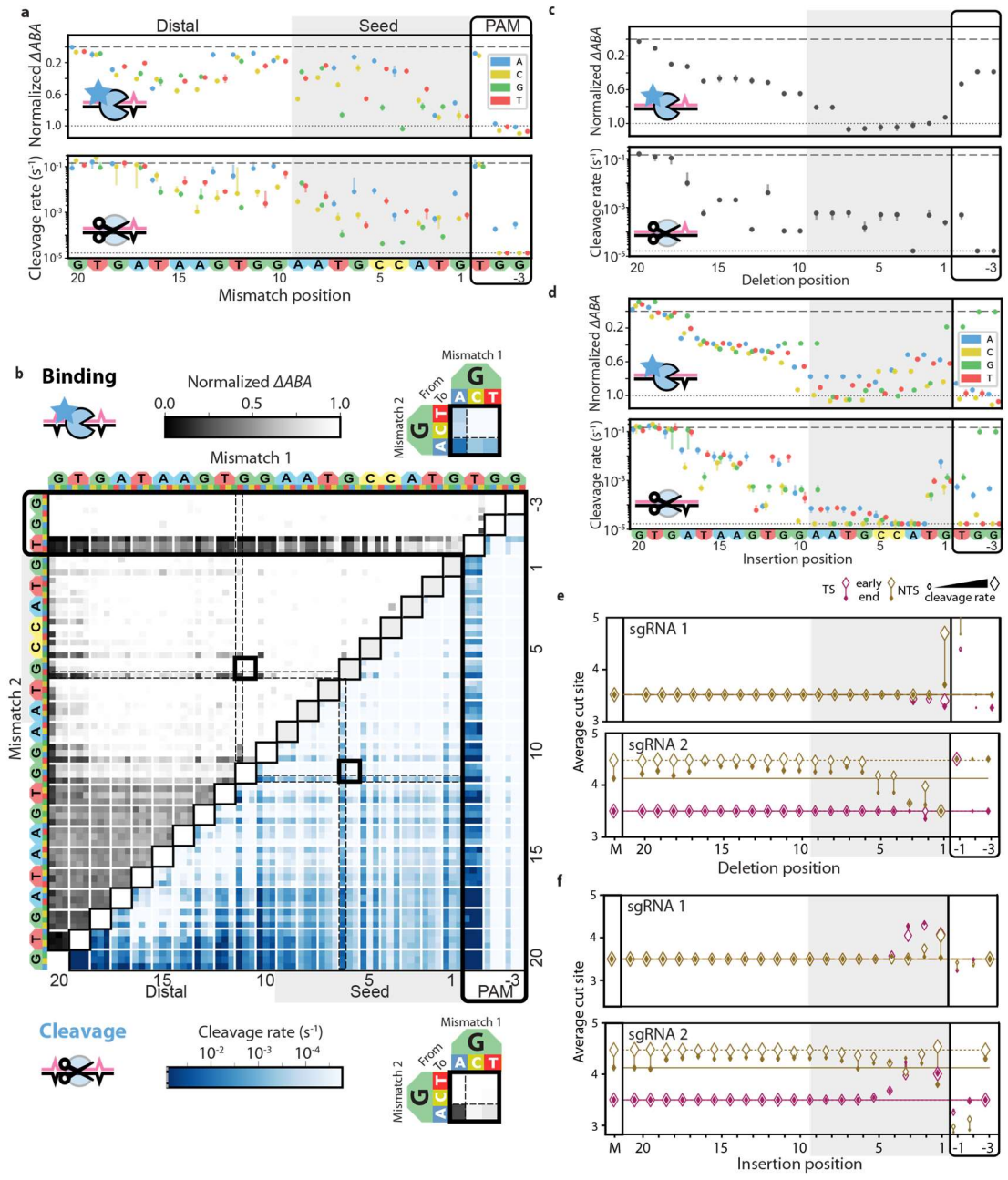
a
Distal
Seed
PAM
Normalized ΔΔABA
Cleavage rate (s⁻¹)
Mismatch position
b
Binding
Normalized ΔΔABA
Mismatch 1
Mismatch 2
Distal
Seed
PAM
Cleavage
Cleavage rate (s⁻¹)
c
Deletion position
d
Insertion position
TS
early end
NTS
cleavage rate
e
sgRNA 1
sgRNA 2
Average cut site
Deletion position
f
Insertion position

Figure S2. Comprehensive analysis of off-target wtCas9 DNA binding and cleavage with sgRNA 2. (a) dCas9 ΔABAs (upper, 0 indicates matched target) and cleavage rates (lower) for all DNAs with a single mismatch relative to sgRNA 2. Dotted line: normalized matched target ΔABA or cleavage rate. Dashed line: scrambled DNA ΔABA (negative control) or limit of detection for the slowest-cleaving targets. Error bars: SD (upper) or 5% to 95% confidence intervals (lower) as measured by bootstrap analysis. **(b)** ΔABAs (upper, grays) and cleavage rates (lower, blues) for DNAs containing two mismatches relative to sgRNA 2. Black boxes expanded in callouts. **(c)** dCas9 ΔABAs (upper) and Cas9 cleavage rates (lower) for DNAs containing a single nucleotide deletion or **(d)** a single nucleotide insertion compared to sgRNA 2. Error bars: ΔABA SD and cleavage rate 5% to 95% confidence intervals as measured by bootstrap analysis. **(e)** Average cut site positions for each strand (TS, NTS) from DNAs containing one deletion or **(f)** insertion compared to sgRNA 1 (upper) or 2 (lower). Range spans the first timepoint (early, open diamonds) to the final time point (end, filled diamonds). Diamond size indicates the average cleavage rates of the associated DNAs. Dashed and solid horizontal lines indicate average cut site positions for matched DNA at early and late time points.

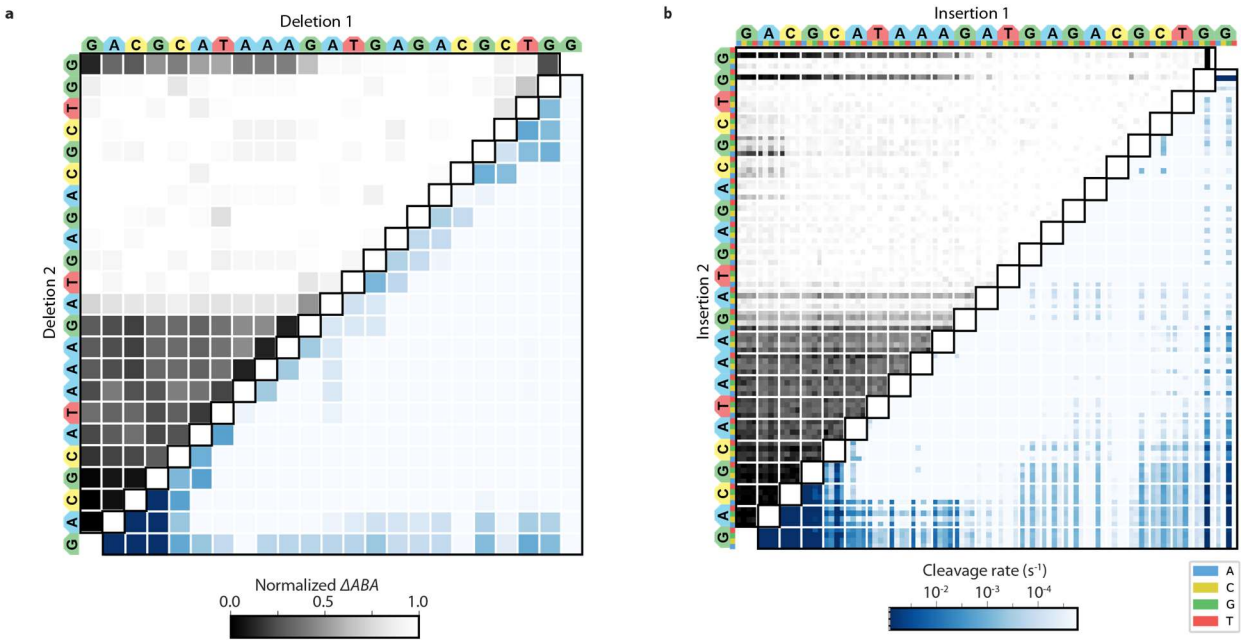

Figure S3 Comprehensive analysis of off-target wtCas9 DNA binding and cleavage of DNAs with insertions or deletions. (a) ΔABAs (upper) and cleavage rates (lower) for DNAs containing two deletions or **(b)** two insertions relative to sgRNA 1.

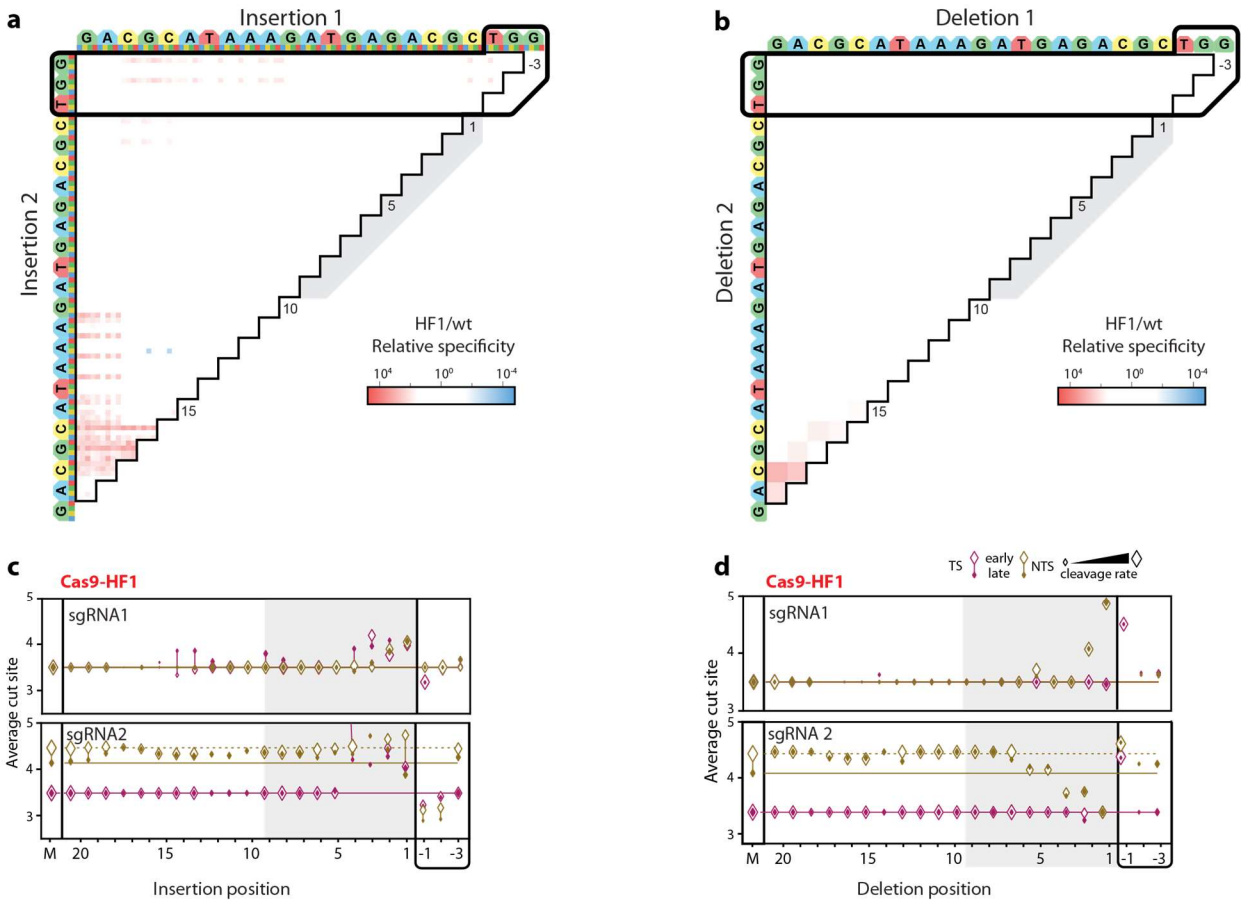

Figure S4. Comparison of Cas9-HF1 and wtCas9 nuclease activities. (a) Ratio of Cas9-HF1 to wtCas9 cleavage rates for all DNAs containing two insertions or **(b)** deletions compared to sgRNA 1. Red: slower cleavage by Cas9-HF1; blue: slower cleavage by wtCas9. **(c)** Average cut site positions generated by Cas9-HF1 for each strand (TS, NTS) for DNAs containing the insertions or **(d)** deletions compared to sgRNA 1 (upper) or sgRNA 2 (lower). Range spans the first timepoint (early, open diamonds) to the final time point (late, filled diamonds). Diamond size indicates the average cleavage rates of the associated DNAs. Dashed and solid horizontal lines indicate average cut site positions for matched DNA at early and late time points.

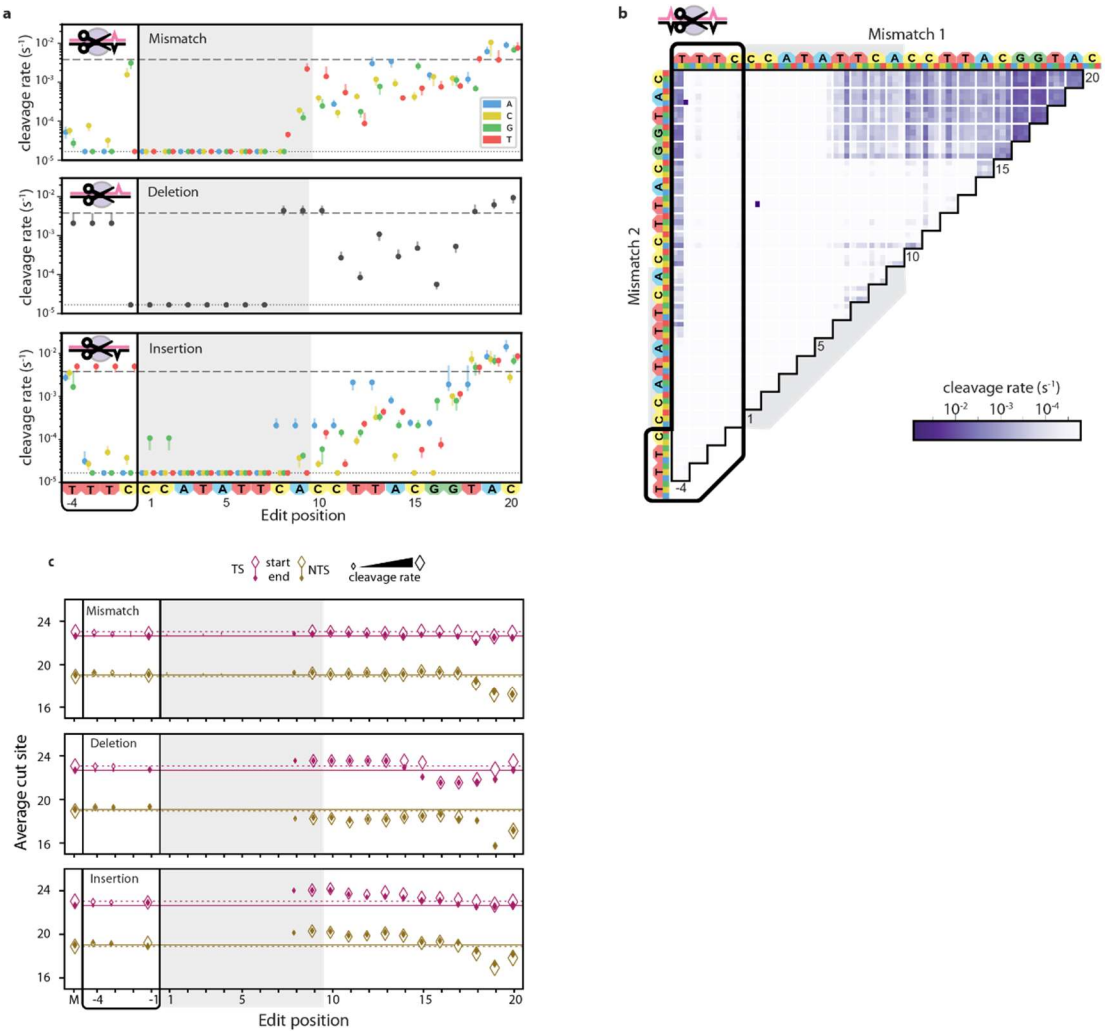

Figure S5. Analysis of off-target wtCas12a cleavage with crRNA 4.

(a) Cleavage rates for all DNAs with a single mismatch (upper), deletion (middle) or insertion (lower). Dotted line: matched target cleavage rate. Dashed line: limit of detection for the slowest-cleaving targets. Error bars: 5% to 95% confidence intervals as measured by bootstrap analysis. **(b)** Cleavage rates for DNAs containing two mismatches. **(c)** Average cut site positions generated by wtCas12a for each strand (TS, NTS) for DNAs containing the mismatches (upper), deletions (middle) or insertions (lower). Range spans the first timepoint (early, open diamonds) to the final time point (late, filled diamonds). Diamond size indicates the average cleavage rates of the associated DNAs, as in (a). Dashed and solid horizontal lines indicate average cut site positions for matched DNA at early and late time points.

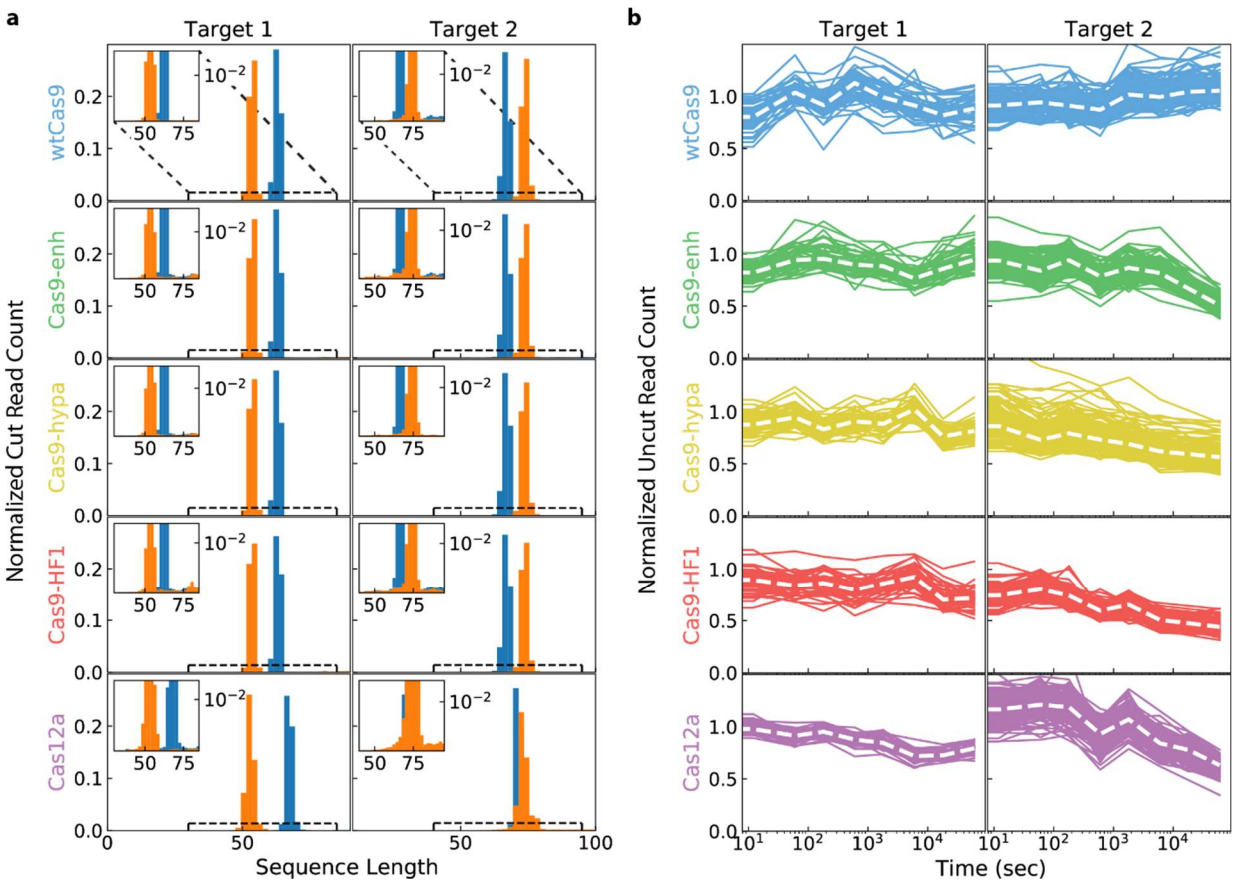

Figure S6. Cas12a exhibits limited *trans* cleavage of NucleaSeq libraries. We looked for two signatures of *trans* cleavage activity. (a) First, the distribution of normalized read counts of DNAs that only have the left (blue) or right (orange) barcodes show limited cut fragments outside the expected enzyme cleavage sites. Inset: a zoomed in view shows that there are few indiscriminately cut products outside of the main peaks, which correspond to the canonical cut site. A comparable amount of short DNA for Cas9 and Cas12 suggests that these fragments arise during library preparation and NGS. Cas12a-catalyzed *trans* cleaved DNA is a relatively minor component of our data (within the noise). Histograms are normalized to have a total area of one. (b) Second, we looked at the time-dependent read counts of ~150 uncut control sequences that are not complementary to the sg/crRNAs (overlapping lines). These sequences are normalized only for total read counts at each time, then with the value at time zero set to one. These values are proportional to the fraction of the library at each time point, not absolute read count. Robust *trans* cleavage by Cas12a would be expected to deplete these values over time at a faster rate than other proteins. However, Cas12a behaves similarly to the engineered Cas9 variants in all cases. Furthermore, we do not observe any individual traces decreasing faster than the group, as would

be observed if trans-cleavage showed sequence bias for some of the control sequences. White dashed line: median.

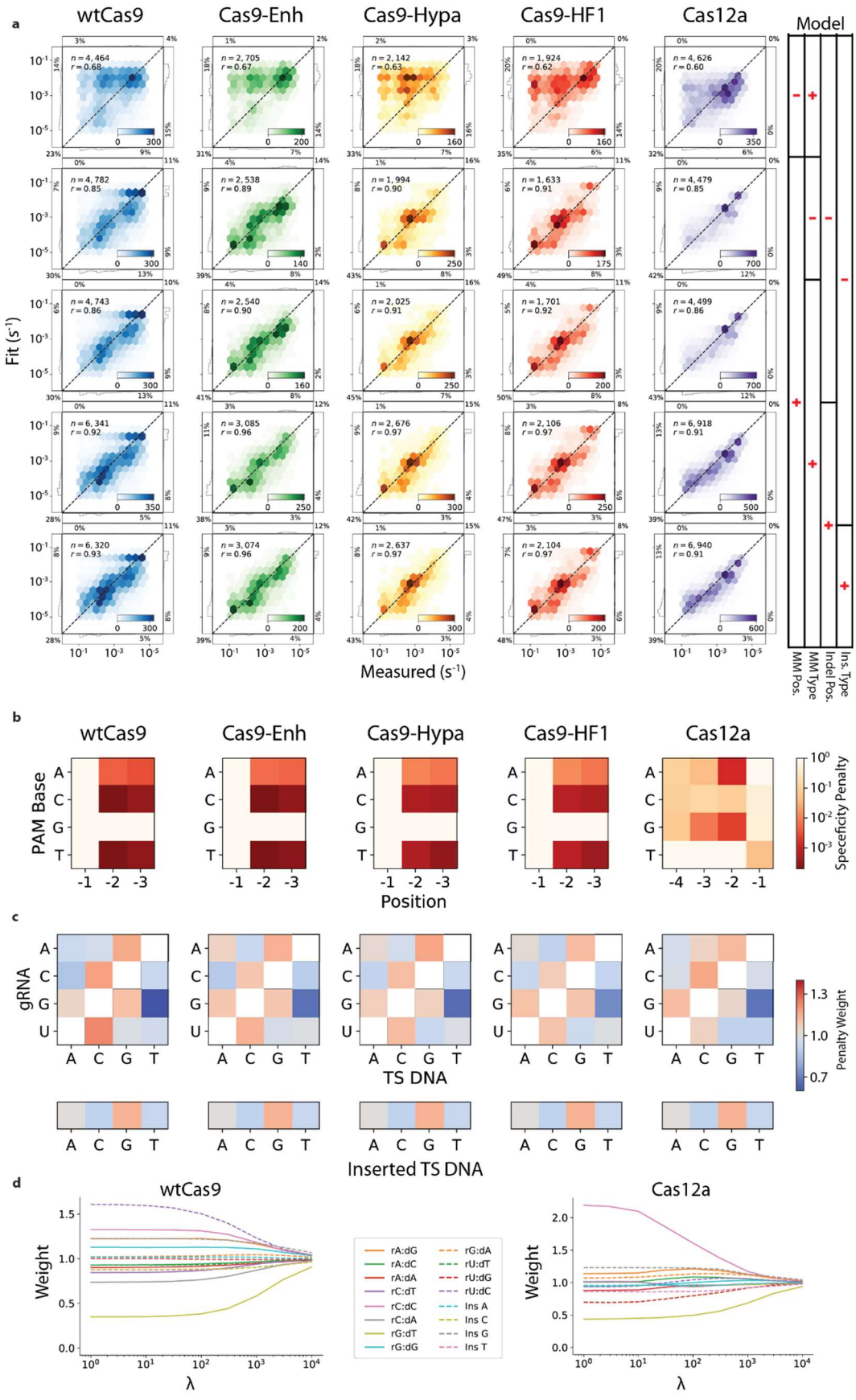
a
wtCas9
Cas9-Enh
Cas9-Hypa
Cas9-HF1
Cas12a
Model
Fit (s⁻¹)
Measured (s⁻¹)
MM Pos.
MM Type
Indel Pos.
Ins. Type
b
PAM Base
Position
Specificity Penalty
c
gRNA
TS DNA
Inserted TS DNA
Penalty Weight
d
Weight
λ
rA:dG
rA:dC
rA:dA
rC:dT
rC:dC
rC:dA
rG:dT
rG:dG
rG:dA
rU:dT
rU:dG
rU:dC
Ins A
Ins C
Ins G
Ins T

Figure S7. Comparison of biophysical models for Cas nuclease specificity. (a) Correlation between measured and fit cleavage rates for each protein for the indicated simplified model. (**b**) PAM position weight matrices computed from the cleavage rates for each protein. (**c**) Penalty weight values of mismatch (top) and insertion types (bottom) for each protein. (**d**) Transition and insertion weights as a function of the regularization parameter, λ (see supplemental computational methods).

MATERIALS AND METHODS

Oligonucleotides, CRISPR RNA, and DNA libraries

Oligonucleotides were purchased from IDT (see **Table S1**). Single guide RNAs (sgRNAs) for Cas9 and CRISPR RNAs (crRNAs) for Cas12a were purchased from Synthego (see **Table S1**). Pooled oligonucleotide libraries were purchased from CustomArray Inc. and Twist Biosciences (see **Table S2**). Libraries were amplified via 12 cycles of PCR with Phusion polymerase (NEB).

DNA Library Design

Each library contains DNAs that are variations of a matched DNA sequence (defined by nuclease PAM preference and RNA guide), termed a 'modified target'. Modified targets include: single and double substitutions, insertions, or deletions, and all sequences with a contiguous subsection changed to the complementary bases. Each modified target is flanked by the following additional sequence elements necessary for NucleaSeq analysis and (5' to 3'): left primer, left barcode, left buffer, modified target, right buffer, variable length buffer, right barcode, right primer (Supplemental File 1). As controls, we included 146 copies of the matched target. Each copy had a unique left and right barcode set. Finally, we included 150 pseudo-random barcoded DNA strands to normalize read depth between time points and biological replicates (see below).

Our libraries use unique barcodes appended to either end of each DNA strand (Hawkins et al., 2018). By searching for the barcodes after NGS, any cleaved DNA can be computationally identified from a partial fragment after cleavage. These barcodes are 17 bp, uniquely paired, and are correctly identified despite any combination of up to two substitutions, insertions, or deletions in their sequence. Similarly, primer sequences (common across the library) were selected that help distinguish left barcodes, right barcodes and cleaved ends. They are distinguishable from one another and the cleaved end of any library member cut within 5 bp of a canonical cut site.

Flanking each modified target are left and right 5 bp buffer regions held constant for all sequences to provide a constant local DNA context for nuclease activity. These buffer sequences were randomly generated with nearly equal nucleotide content. Oligos with insertions and deletions also included a variable-length buffer to ensure that these oligos were the same length as the matched target.

Protein cloning and purification

S. pyogenes Cas9 variants were generated via Q5 site-directed mutagenesis (New England Biolabs) of a pET-based plasmid (pMJ806) (Jinek et al., 2012). Nuclease dead Cas9 variants contained the D10A and H840A mutations. Enhanced-, HF1-, and Hypa- Cas9 variants harbored the mutations indicated in **Table S1** (Chen et al., 2017; Kleinstiver et al., 2016a; Slaymaker et al., 2016). An N-terminal 3xFlag epitope was introduced for fluorescent imaging of nuclease dead variants via CHAMP (see below).

Cas9 protein variants were expressed in BL21 star (DE3) cells (Thermo Fisher Scientific) using a previously established protocol with minor modifications (Jinek et al., 2012). A 4L flask containing 1L LB + Kanamycin was inoculated with a single colony and then grown to an optical density (OD) of 0.6 at 30°C with shaking. Protein expression was induced with 1mM IPTG during 18 hours at 18°C with shaking. Cells were collected by centrifugation and lysed by sonication at 4°C in lysis buffer (20 mM Tris-Cl pH 8.0, 250 mM NaCl, 5mM imidazole, 5 µM phenylmethylsulphonyl fluoride, 6 units ml^{-1} DNAse I). The lysate was clarified by ultracentrifugation at 35k RCF, then passed over a nickel affinity column (HisTrap FF 5mL, GE Healthcare) and eluted with elution buffer (20 mM Tris-Cl pH 8.0, 250 mM NaCl, 250 mM imidazole). The His_6-MBP was proteolyzed overnight in dialysis buffer (20 mM HEPES-KOH pH 7.5, 150 mM KCl, 10% glycerol, 1mM DTT, 1mM EDTA) supplemented with TEV protease (0.5 mg per 50 mg protein). The dialyzed protein was resolved on a HiTrap SP FF 5mL column (GE Healthcare) with a linear gradient between buffer A (20 mM HEPES-KOH pH 7.5, 100 mM KCl) and buffer B (20 mM HEPES-KOH pH 7.5, 1 M KCl). Protein-containing fractions were concentrated via dialysis (10 kDa Slide-A-Lyzer, Thermo Fisher Scientific), and then sized on a Superdex 200 Increase 10/300 column (GE Healthcare) pre-equilibrated into storage buffer (20 mM HEPES-KOH pH 7.5, 500 mM KCl). The protein was snap frozen in liquid nitrogen and stored in 10µL aliquots at -80ºC.

Acidaminococcus sp (As) Cas12a was expressed as an N-terminal His_6-TwinStrep-SUMO fusion in a pET19-based plasmid (pIF502) (Strohkendl et al., 2018). The Cas12a fusion protein was expressed in BL21 star (DE3) cells (Thermo Fisher Scientific) using a previously established protocol with minor modifications (Strohkendl et al., 2018). A 20 mL culture of Terrific Broth (TB) + 50 mg mL^{-1} carbenicillin was inoculated with a single colony and grown overnight at 37ºC

with shaking. A 4 L flask containing 1 L of TB was inoculated with 10 mL of the starter culture and then grown to an optical density (OD) of 0.6 at 37°C. Protein expression was induced with 0.5 mM IPTG during 24 hours at 18°C. Cells were collected by centrifugation and lysed by sonication at 4°C in lysis buffer (20 mM Na-HEPES pH 8.0, 1 M NaCl, 1 mM EDTA, 5% glycerol, 0.1% Tween-20, 1 mM PMSF, 2000 U DNase (GoldBio), 1X HALT protease inhibitor (Thermo Fisher)). The lysate was clarified by ultracentrifugation at 35k RCF, applied to a hand-packed StrepTactin Superflow gravity column (IBA Life Sciences), and then eluted (20 mM Na-HEPES, 1 M NaCl, 5 mM desthiobiotin, 5 mM $MgCl_2$ and 5% glycerol). The eluate was concentrated to <1mL using a 30 kDa MWCO spin concentrator (Millipore), SUMO protease was added at 3µM, and then the eluate was incubated overnight on a rotator at 4ºC.The protein was resolved on a HiLoad 16/600 Superdex 200 Column (GE Healthcare) pre-equilibrated with storage buffer (20 mM HEPES-KOH, 150 mM KCl, 5 mM MgCl2, 2 mM DTT buffer). The protein was finally snap frozen in liquid nitrogen and stored in 10µL aliquots at -80ºC.

Cas9 and Cas12a ribonucleoprotein (RNP) complexes were reconstituted by incubating a 2:3 molar ratio of apo protein and RNA (sgRNA and pre-crRNA for Cas9 and Cas12a, respectively) in RNP buffer (20 mM HEPES pH 7.5, 150 mM KCl, 10 mM $MgCl_2$, 2 mM DTT) at room temperature for 30 minutes prior to each experiment. Reconstituted RNPs were diluted in the experimental reaction buffer, used immediately, and discarded after the experiment.

NucleaSeq

DNA libraries were mixed in buffer (20 mM HEPES pH 7.5, 150 mM KCl, 10 mM $MgCl_2$, 2 mM DTT) at room temperature with RNP complex to final concentrations of 10 nM and 100 nM, respectively. Aliquots were transferred to a stop solution (final concentration: 12 mM EDTA and 12 U proteinase K (Thermo Fisher)) at the following time points: 0, 0.2, 0.5, 1, 3, 10, 30, 100, 300 and 1000 minutes. The stopped reactions were incubated at 37°C for 30 minutes to remove Cas9 and Cas12a from their DNA substrates. Each time point was ethanol precipitated and resuspended in TE buffer. Samples were submitted to the University of Texas Genomic Sequencing and Analysis Facility, where sequencing adapters (NEBNext Ultra, NEB) were appended. The samples were sequenced on a MiSeq or NextSeq 500 sequencer (Illumina).

Bioinformatic analysis pipeline

From each paired-end read pair, we inferred the maximum likelihood full-length sequence using the overlapping base pairs as described previously (Jung et al., 2017). Primer and barcode sequences were then used to identify the intended sequence identity and, for cleaved products, the observed side. Observed and intended sequences were aligned using either global alignment (Cock et al., 2009) for uncleaved products or global alignment with cost-free ends (Cieślik et al., 2016) for cleaved products. Throughout this process, sequences were filtered for quality based on length, primer and barcode structure, and number of synthesis and sequencing errors. Sequences with errors were not allowed in the target and buffer regions.

Next, the read counts for each full-length library member in each sample were normalized to account for two sources of variation. First, we normalized the different total numbers of reads across different time points for each sample. Specifically, each member's read count for each sample was normalized by the ratio of total read counts at that time point to the total read count of an input control sample (not treated with nuclease). Second, read counts were normalized to account for changes due to sampling from a library of changing composition. The generation of cleaved products and corresponding depletion of full-length products by nuclease activity changes the number of sampled sequences of all species, including species unaffected by the nuclease. To account for this, we used the 150 non-target control sequences as a reference. For each randomly-generated non-target sequence, there is a small probability it will be susceptible to nuclease cleavage. Hence, we used the median read count value of all the random sequences as a robust measure of changes due only to sampling from a library of changing composition (non-target median). Read counts of each library member at each time point were normalized by the ratio of the non-target median at that time point to the non-target median from the control sample.

In addition to the above two steps, cleaved products were normalized to account for differences in PCR amplification between cleaved products and full-length oligos. We observed that the normalized number of cleaved products should be proportional to the depletion of the corresponding full-length oligos. Stated as an equation, let $|F|_t$ be the number of full-length product reads and $|C|_t^{side}$ be the number of cleaved product reads on a given side at a given time, for a single library member of choice, normalized as above. Then for normalization and proportionality constants Z_t^{side} and k^{side},

$$\frac{|C|_t^{side}}{Z_t^{side}} = k^{side}\left(1 - \frac{|F|_t}{|F|_0}\right)$$

We choose to set the final normalization constant $Z_{t_f}^{side} = 1$ and solve the above for k^{side}. Plugging this back in and rearranging gives normalization constants

$$Z_t^{side} = \frac{|C|_t^{side}}{|C|_{t_f}^{side}}\left(\frac{1 - |F|_{t_f}/|F|_0}{1 - |F|_t/|F|_0}\right)$$

This is intentionally a function only of ratios of read counts, not absolute read counts. This lets us use the median read count ratios from all 146 matched target controls to calculate the normalization constants. These final normalization constants are then used for all library members. Finally, read counts are normalized to range between zero and one. For full-length products, we normalize by the fit value of reads at time zero. For cleaved products, we normalize first by the sum of all cleaved products at all time points, then normalize to set the resulting median sum of all cleaved products at the final time point to the depletion of full-length products, $1 - |F|_{t_f}/|F|_0$.

The normalized read counts were fit to a single exponential decay. We observed that the data was well described by a single exponential, implying a constant reaction rate under the single turnover conditions used in this assay. A small fraction of the starting DNA sequences of each species were never cleaved, possibly indicating some hydrolytically inactive enzymes. We thus fit for exponential decay with a constant offset. For the constant offset, we used the median normalized fraction of uncleaved sequences of the 146 perfect target sequences at the final time point. Error bars give the standard deviation of 50 bootstrap measurements, each of which was calculated by resampling the raw read counts with replacement, renormalizing, and refitting (Efron and Tibshirani, 1993).

Modeling cleavage specificity

Model description

We modeled cleavage specificity (*Model 1*), given as the ratio of the cleavage rate of a given sequence s, k_s, to the cleavage rate of the matched sequence m, k_m, as:

$$\log\frac{k_s}{k_m} = \sum_{i\in\mathcal{P}} \log\Lambda(i, s_i) + \sum_{i\in\mathcal{D}} \log P_D(i) + \sum_{i\in\mathcal{I}} w_I(s_i)\log P_I(i) + \sum_{i\in\mathcal{M}} t_M(r_i, s_i)\log P_M(i)$$

The terms of the model give cleavage rate penalties for the following sequence edits respectively: suboptimal bases in the PAM, target deletions, target insertions, and target mismatches, each with corresponding set of positions with the given sequence edit type: $\mathcal{P}$, $\mathcal{D}, \mathcal{I}$, and $\mathcal{M}$. For suboptimal PAM bases, the cleavage rate penalty is given by the function Λ, a function of both the suboptimal base identity, s_i, and its position i.

For deletions, insertions, and mismatches, the cleavage rate penalty functions P_D, P_I, and P_M are dependent only on the position i, reflecting the fact that position in the target is the primary determinant of the effect of a given sequence edit. This is intuitive for deletions, as they primarily require steric adjustments to realign the matching base pairs. For mismatches, position was determined to be the primary determinant of the cleavage rate penalty via comparison with other models (see Simplified models below). Insertions have a weighting function w_I to allow for different inserted bases to have different penalties. The base identities in the mismatch are modeled via the weighting function $t_M(r_i, s_i)$, a function of the mismatched gRNA base r_i and target strand base s_i.

Within the terms for insertion and mismatch penalties, there is an unconstrained degree of freedom in the relative magnitudes of the weights relative to the log position penalties. To remove this extra degree of freedom, the insertion and mismatch weighting functions w_I and t_M were each constrained to have an average value of 1. This was accomplished through the use of Hadamard matrices, possible because w_I and t_M have 4 and 12 parameters, respectively. Hadamard matrices are maximal-determinant matrices using elements of only 1 and -1. We used Hadamard matrices with negative one in all elements not in the first row or column along diagonals 0, -1, 2, -3, -4, -5, 6, 7, 8, -9, and 10, where 0 is the main diagonal and diagonal indices increase up and to the right. We parameterize a constrained length n weight vector w with a length $(n-1)$ vector x of free parameters as follows. Let H_n be the $n \times n$ Hadamard matrix described above. Due to the inverse identity of Hadamard matrices and the first row and column of H_n being composed entirely of ones, parameterizing with x and using the following conversions enforces an average value of 1 in the weights vector w:

$$\begin{bmatrix} n \\ - \\ x \end{bmatrix} = H_n w, \quad w = \frac{1}{n} H_n^T \begin{bmatrix} n \\ - \\ x \end{bmatrix}$$

Cleavage rates that are shorter than the first time point or longer than the last one cannot be modeled accurately. We therefore constrained the output of our models with the following "bandpass filter" function:

$$B(x) = \begin{cases} x & s \le x \le f \\ s & x < s \\ f & x > f \end{cases}$$

Where s and f are the slowest and fastest detectible cleavage rates, corresponding to half-lives at our first and last time points.

Ridge regularization of the difference of insertion and mismatch weights from one was used to reduce over-fitting of the underlying cleavage data (Hoerl and Kennard, 1970). Figure S7D shows the fit weight values as a function of the regularization parameter λ. The relative parameter values appear to stabilize near $\lambda = 10^3$, which we used to fit the model.

Simplified models

For comparison, we fit our data to four simplified models, each excluding some terms and/or factors in the full model above. The first three simplified models did not include the insertion or deletion terms, modeling the possibility that the recognition channel does not accommodate bulges to realign matching sequences after indels. Under this assumption, for example, a sequence with a single insertion between the first and second bases, but otherwise perfectly matching, would result in about 75% mismatches due to a forced frame shift. These three models were: cleavage rate as a function of only the mismatch base pair identities, only the mismatch position, or both as in the full model above. The fourth simplified model included insertions and deletions but omitted the insertion weights w_I. Each simplified model included the PAM term. We number the models for reference:

$$Model\ 5\text{: } \log\frac{k_s}{k_m} = \sum_{i\in\mathcal{P}} \log\Lambda(i, s_i) + \sum_{i\in\mathcal{M}} \log T_M(r_i, s_i)$$

$$Model\ 4\text{: } \log\frac{k_s}{k_m} = \sum_{i\in\mathcal{P}} \log\Lambda(i, s_i) + \sum_{i\in\mathcal{M}} \log P_M(i)$$

$$Model\ 3\text{: } \log\frac{k_s}{k_m} = \sum_{i\in\mathcal{P}} \log\Lambda(i, s_i) + \sum_{i\in\mathcal{M}} t_M(r_i, s_i) \log P_M(i)$$

$$Model\ 2: \log\frac{k_s}{k_m} = \sum_{i\in\mathcal{P}} \log\Lambda(i, s_i) + \sum_{i\in\mathcal{D}} \log P_D(i) + \sum_{i\in\mathcal{I}} \log P_I(i) + \sum_{i\in\mathcal{M}} t_M(r_i, s_i) \log P_M(i)$$

And Model 1 is the full model above. The mismatching base pairs function in Model 5, $T_M(r_i, s_i)$, is different from the analogous weighting function $t_M(r_i, s_i)$ in the other models as it gives absolute penalty values, not weights constrained to average value of one.

Figure 5 compares these models using the Akaike Information Criterion (AIC) (Akaike, 1974). The significant improvement in AIC between Models 5 and 4 demonstrates that position is in fact the primary determinant of mismatch cleavage rates. Model 3 demonstrates that including the mismatched base pair identities is a useful but relatively small improvement to the position-only model. Similarly, Models 2 and 1 show that adding insertions and deletions to the model provides a significant improvement, while the addition of insertion weights is a relatively small improvement to the model (i.e., insertions are weakly sensitive to the inserted base identity).

Model comparison to previously published datasets

To compare the model's output with prior measures of nuclease specificity, we selected *in vitro* and *in vivo* published datasets for either *Sp*Cas9 or *As*Cas12a that contained at least one measure of specificity per position in the sgRNA (for *Sp*Cas9) or the crRNA (for *As*Cas12a).

Dataset 1 (Pattanayak et al., 2013) included representative genes CLTA1 and CLTA2 with sgRNA v2.1 and 100 nM wt Cas9. Published specificity scores were averaged across all single mismatch values at each position. Dataset 2 (Kim et al., 2015) used Digenome-seq and included sgRNAs targeting genes HBB and VEGFA. Dataset 3 (Kleinstiver et al., 2016a) used GUIDE-Seq to profile indels at genes VEGFA-2 and EMX1-1. Values were extracted from the published heatmaps based on RGB values as measured with FIJI (Schindelin et al., 2012). The measured scores were averaged across all single mismatch values at each position. Dataset 4 (Fu et al., 2016) *in vivo* log retention scores for genes UNC-22A and ROL6 were extracted from published graphs with a data digitization tool (Rohatgi, 2019). The measured scores were averaged across all single mismatch values (transitions and transversions) at each position. Dataset 5 (Hsu et al., 2013) used SURVEYOR nuclease to determine mean cleavage results for aggregated EMX1 targets. Values were extracted from the published heatmaps based on position-averaged RGB values as measured with FIJI (Schindelin et al., 2012).

Dataset 6 (Chen et al., 2017) used a T7E1 reporter assay and included representative genes FANCF-1 and FACNF-4. Percent of modification for each gene was extracted from the published heatmaps based on RGB values as measured with FIJI for wtCas9 (Schindelin et al., 2012). Dataset 7 (Yan et al., 2017) used BLISS to generate composite mismatch tolerances for each guide position. Values were extracted from the published graph with a data digitization tool (Rohatgi, 2019). Dataset 8 (Kim et al., 2017) relative indel frequency values at each position were extracted from the published graph with a data digitization tool (Rohatgi, 2019). Dataset 9 (Kleinstiver et al., 2016b) used a T7E1 reporter assay and included representative gene DNMT1, sites 1 and 3. Percent of modification for each gene was extracted from the published graphs with a data digitization tool (Rohatgi, 2019). Since the measure and distribution of data varied from study to study, a nonparametric correlation was used (only requires ordinal data). Each dataset was compared to one another and to our model's average positional mismatch penalty to generate Spearman's rank correlation coefficients (ρ). The average mismatch penalty is denoted as P_M in *Model 1*.

CHAMP (Chip Hybridized Association Mapping Platform)

DNA libraries were sequenced on a MiSeq using 2x75 paired-end chemistry (v3, Illumina) Sequenced MiSeq chips were stored at 4°C in storage buffer (10 mM Tris-Cl, pH 8.0, 1 mM EDTA, 500 mM NaCl) until needed for CHAMP.

Chips were regenerated similarly to our previous strategy (Jung et al., 2017). Each chip was loaded into a custom microscope stage adapter, with temperature controlled by a custom heating element. All solutions were pumped through the chip at 100 µl min^{-1} using a syringe pump (Legato 210, KD Scientific), with reagents added via an electronic injection manifold (Rheodyne MXP9900). Chip DNAs were made single-stranded with 500 µl 60% DMSO, then washed with 500 µl TE buffer. An unlabeled regeneration primer (user DNA specific) and a digoxygenin labeled primer (PhiX DNA specific, for alignment) were annealed over an 85-40°C temperature gradient (30 min) in hybridization buffer (75 mM tri-sodium citrate, pH 7.0, 750 mM NaCl, 0.1% Tween-20), and then excess primers were removed at 40°C with 1 ml wash buffer (4.5 mM Trisodium Citrate, pH 7.0, 45 mM NaCl, 0.1% Tween-20). Annealed primers were extended at 60°C using 0.08 U $µl^{-1}$ Bst 2.0 WarmStart DNA polymerase (New England Biolabs) and 0.8 mM dNTPs in isothermal amplification buffer (20 mM Tris-HCl, pH 8.8, 10 mM $(NH_4)_2SO_4$, 50 mM KCl, 2 mM $MgSO_4$,

0.1% Tween-20), then washed with 500 µl wash buffer. Using 100 µl of rabbit anti-digoxigenin monoclonal antibody (Life Technologies) and 100 µl Alexa488-conjugated, goat anti-rabbit antibody (Thermo Fisher Scientific), PhiX DNA clusters were fluorescently labeled as markers for subsequent image alignment. The MiSeq chips were imaged on a Ti-E microscope (Nikon) in a prism-TIRF configuration (Jung et al., 2017). Images were acquired in OME-TIFF format (uncompressed TIFF plus XML meta data) using the Micro-Manager software (Edelstein et al., 2014).

The dCas9/sgRNA RNP complex was diluted to concentrations of 0.1, 0.3, 1, 3, 10, 30, 100 and 300 nM in CHAMP buffer (20mM Tris-HCl pH 7.5, 100 mM KCl, 5 mM $MgCl_2$, 5% glycerol, 0.2 mg mL^{-1} BSA, 0.1% Tween-20, 1 mM DTT). Starting with the lowest concentration, 100 µl of RNP complex was injected into the regenerated MiSeq chip at room temperature and incubated for 10 minutes. Then, 300 µl of CHAMP buffer containing 4 nM Alexa488-conjugated anti-Flag antibody (Alexa Fluor 488 antibody labeling kit, Thermo Fisher; monoclonal BioM2, Sigma-Aldrich) was injected to wash off unbound RNP and label DNA-bound RNP complex. The chip was then imaged over 420 fields of view (FOVs) with 10 frames of 50 ms each, while illuminated with 10 mW of laser power, as measured at the front face of the prism. Collected images were processed via the CHAMP bioinformatic software for downstream analysis (Jung et al., 2017).

DNA Radiolabeling and PAGE purification

DNA oligonucleotides 308 and 310 were 5'-radiolabeled with [γ-^{32}P] ATP (Perkin-Elmer) using T4-polynucleotide kinase (New England Biolabs). Radiolabeled nucleotides were then purified by electrophoresis in a 12% native polyacrylamide gel, before being eluted in TE buffer (10mM Tris-HCl pH 8, 1 mM EDTA) to a concentration 250nM.

Nuclease active site titration

Atto647N-labeled target DNA was generated with Q5 DNA polymerase (NEB) using oligonucleotides 365, 460 and 371. The DNA was diluted in series from 512 nM to 4nM in reaction buffer (20 mM HEPES pH 7.5, 150 mM KCl, 10 mM $MgCl_2$, 2 mM DTT). RNP complexes were formed by mixing protein and RNA (256nM:384nM) and incubating for 30 minutes at room temperature in the same buffer conditions. Equal volumes of RNP and Atto647N-labeled matched DNA dilutions were combined then incubated for 30 minutes at room temperature. The reaction was stopped by addition of a stop solution (40 mM EDTA and 50 U proteinase K (Thermo Fisher))

and a 30-minute incubation at 37ºC removed RNPs from their DNA substrates. All samples were run in a 10% polyacrylamide native gel and then imaged using a Typhoon FLA9500 gel scanner (GE Healthcare).

Table S1. Nucleic acids used in this work

Name	RNP	Use	Sequence	References
sgRNA 1	Cas9	guide	UAAUACGACUCACUAUAGGACGCAU AAAGAUGAGACGCGUUUUAGAGCU AGAAAUAGCAAGUUAAAAUAAGGC UAGUCCGUUAUCAACUUGAAAAAGU GGCACCGAGUCGGUGCUUUU	This work
sgRNA 2	Cas9	guide	UAAUACGACUCACUAUAGGUGAUAA GUGGAAUGCCAUGGUUUUAGAGCU AGAAAUAGCAAGUUAAAAUAAGGC UAGUCCGUUAUCAACUUGAAAAAGU GGCACCGAGUCGGUGCUUUU	
crRNA 3	Cas12a	guide	GUCAAAAGACCUUUUUAAUUUCUAC UCUUGUAGAUGUGAUAAGUGGAAU GCCAUGUGGA	
crRNA 4	Cas12a	guide	GUCAAAAGACCUUUUUAAUUUCUAC UCUUGUAGAUGACGCAUAAAGAUG AGACGCUGGA	
pr238	dCas9	H840A	TCGTTTAAGTGATTATGATGTCGATG CCATTGTTCCACAAAGTTTCCTTAAA	(Jinek et al., 2012)
pr239			TTTAAGGAAACTTTGTGGAACAATGG CATCGACATCATAATCACTTAAACGA	
pr236		D10A	GAAATACTCAATAGGCTTAGCTATCG GCACAAATAGCGTCG	
pr237			CGACGCTATTTGTGCCGATAGCTAAG CCTATTGAGTATTTC	
pr271	Cas9-enh	K848A	CGATCACATTGTTCCACAAAGTTTCC TTGCAGACGATTCAATAGACAATAA G	(Slaymaker et al., 2016)
pr272			CTTATTGTCTATTGAATCGTCTGCAA GGAAACTTTGTGGAACAATGTGATCG	

pr273		K1003A	CGTTGGAACTGCTTTGATTAAGAAAT ATCCAGCACTTGAATCGGAGTTTGT	
pr274			ACAAACTCCGATTCAAGTGCTGGATA TTTCTTAATCAAAGCAGTTCCAACG	
pr275		K1060A	ACTTGCAAATGGAGAGATTCGCAAA GCCCCTCTAATCGAA	
pr276			TTCGATTAGAGGGGCTTTGCGAATCT CTCCATTTGCAAGT	
pr263	Cas9-HF1	N497A	CAGCTCAATCATTTATTGAACGCATG ACAGCCTTTGATAAAAATCTTCCAAA TGAAAAAG	(Kleinstiver et al., 2016a)
pr264			CTTTTTCATTTGGAAGATTTTTATCAA AGGCTGTCATGCGTTCAATAAATGAT TGAGCTG	
pr269		Q926A	CCAATTGGTTGAAACTCGCGCAATCA CTAAGCATGTGGCA	
pr270			TGCCACATGCTTAGTGATTGCGCGAG TTTCAACCAATTGG	
pr265		R661A	GCCGTTATACTGGTTGGGGAGCTTTG TCTCGAAAATTGATTA	
pr266			TAATCAATTTTCGAGACAAAGCTCCC CAACCAGTATAACGGC	
pr267	Cas9-HF1	Q695A	GTTTTGCCAATCGCAATTTTATGGCG CTGATCCATGATGATAGTTTG	(Kleinstiver et al., 2016a)
pr268			CAAACTATCATCATGGATCAGCGCCA TAAAATTGCGATTGGCAAAAC	
pr399	Cas9-hypa	H698A, Q695A	GCGCTGATCGCTGATGATAGTTTGAC ATTTAAAGAAGACATTCAAAAAGCA CAA	(Chen et al., 2017)

pr400		M694A, N692A	GCCAAAGCTGCGATTGGCAAAACCA TCTGATTTCAAAAA	
pr309	sgRNA 1, crRNA 4	Matched target 1	ACGCTCTTCCGATCTTTTAGACGCAT AAAGATGAGACGCTGGAGATCGGAA GAGCAC	This work
pr310			GTGCTCTTCCGATCTCCAGCGTCTCA TCTTTATGCGTCTAAAAGATCGGAAG AGCGT	
pr477		Library 1	Oligonucleotide pool. See table S3.	
pr475		Primer for library 1	ATAACTAATTGAGCTGAACGCAC	
pr476			CTGAATAGTCGTGTAGTTGTGCT	
pr307	sgRNA 2, crRNA 3	Matched target 2	ACGCTCTTCCGATCTTTTAGTGATAA GTGGAATGCCATGTGGAGATCGGAA GAGCAC	
pr308			GTGCTCTTCCGATCTCCACATGGCAT TCCACTTATCACTAAAAGATCGGAAG AGCGT	
pr371		Library 2	Oligonucleotide pool. See table S3.	
pr364		Primer for library 2	AACCGCCGAATAACAGAGT	
pr365			AAGAACGCCTCGCACACT	
pr460		Atto647N Primer for library 2	/5atto647n/AACCGCCGAATAACAGAGT	

Table S2. Plasmids used in this work.

Name	Protein	construct	Mutations	References
pIF324	*Sp*Cas9	6xHis-MBP-3xFlag-*Sp*Cas9		(Jinek et al., 2012)
pIF335		6xHis-MBP-3xFlag-d*Sp*Cas9	D10A, H840A	(Jinek et al., 2012)
pIF325		6xHis-MBP-3xFlag-enhanced *Sp*Cas9-1.1	K848A, K1003A, R1060A	(Slaymaker et al., 2016)
pIF326		6xHis-MBP-3xFlag-enhanced d*Sp*Cas9-1.1	D10A, H840A, K848A, K1003A, R1060A	(Slaymaker et al., 2016)
pIF329		6xHis-MBP-3xFlag-*Sp*Cas9-HF1	R661A, Q695A, Q926A	(Kleinstiver et al., 2016a)
pIF330		6xHis-MBP-3xFlag-d*Sp*Cas9-HF1	D10A, R661A, Q695A, H840A, Q926A	(Kleinstiver et al., 2016a)
pIF350		6xHis-MBP-3xFlag-hypa *Sp*Cas9	N692A, M694A, Q695A, H698A	(Chen et al., 2017)
pIF351		6xHis-MBP-3xFlag-hypa d*Sp*Cas9	D10A, N692A, M694A, Q695A, H698A, H840A	(Chen et al., 2017)
pIF502	*As*Cas12a	6xHis-TwinStrep-SUMO-*As*Cas12a-3xFlag		(Strohkendl et al., 2018)

REFERENCES

Akaike, H. (1974). A new look at the statistical model identification. IEEE Trans. Autom. Control *19*, 716–723.

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.

Cieślik, M., Pederson, B., and Arindrarto, W. (2016). Align: polite, proper sequence alignment.

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.

Edelstein, A.D., Tsuchida, M.A., Amodaj, N., Pinkard, H., Vale, R.D., and Stuurman, N. (2014). Advanced methods of microscope control using μManager software. J. Biol. Methods *1*, 10.

Efron, B., and Tibshirani, R.J. (1993). An Introduction to the Bootstrap (New York: Chapman and Hall/CRC).

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.

Hawkins, J.A., Jones, S.K., Finkelstein, I.J., and Press, W.H. (2018). Indel-correcting DNA barcodes for high-throughput sequencing. Proc. Natl. Acad. Sci. *115*, E6217–E6226.

Hoerl, A.E., and Kennard, R.W. (1970). Ridge Regression: Biased Estimation for Nonorthogonal Problems. Technometrics *12*, 55–67.

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.

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.

Jung, C., Hawkins, J.A., Jones, S.K., Xiao, Y., Rybarski, J.R., Dillard, K.E., Hussmann, J., Saifuddin, F.A., Savran, C.A., Ellington, A.D., et al. (2017). Massively Parallel Biophysical Analysis of CRISPR-Cas Complexes on Next Generation Sequencing Chips. Cell *170*, 35-47.e13.

Kim, D., Bae, S., Park, J., Kim, E., Kim, S., Yu, H.R., Hwang, J., Kim, J.-I., and Kim, J.-S. (2015). Digenome-seq: genome-wide profiling of CRISPR-Cas9 off-target effects in human cells. Nat. Methods *12*, 237–243.

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

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.

Kleinstiver, B.P., Pattanayak, V., Prew, M.S., Tsai, S.Q., Nguyen, N.T., Zheng, Z., and Joung, J.K. (2016a). High-fidelity CRISPR-Cas9 nucleases with no detectable genome-wide off-target effects. Nature *529*, 490–495.

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. (2016b). Genome-wide specificities of CRISPR-Cas Cpf1 nucleases in human cells. Nat. Biotechnol. *34*, 869–874.

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.

Rohatgi, A. (2019). WebPlotDigitizer.

Schindelin, J., Arganda-Carreras, I., Frise, E., Kaynig, V., Longair, M., Pietzsch, T., Preibisch, S., Rueden, C., Saalfeld, S., Schmid, B., et al. (2012). Fiji: an open-source platform for biological-image analysis. Nat. Methods *9*, 676–682.

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.

Strohkendl, I., Saifuddin, F.A., Rybarski, J.R., Finkelstein, I.J., and Russell, R. (2018). Kinetic Basis for DNA Target Specificity of CRISPR-Cas12a. Mol. Cell *71*, 816-824.e3.

Yan, W.X., Mirzazadeh, R., Garnerone, S., Scott, D., Schneider, M.W., Kallas, T., Custodio, J., Wernersson, E., Li, Y., Gao, L., et al. (2017). BLISS is a versatile and quantitative method for genome-wide profiling of DNA double-strand breaks. Nat. Commun. *8*, 15058.