Published online 21 April 2020

Nucleic Acids Research, 2020, Vol. 48, No. 9 ***5037–5053***
doi: 10.1093/nar/gkaa231

Systematic *in vitro* profiling of off-target affinity, cleavage and efficiency for CRISPR enzymes

Liyang Zhang[1,2,†], H. Tomas Rube[3,4,5,†], Christopher A. Vakulskas[2], Mark A. Behlke[2], Harmen J. Bussemaker[4,5,†] and Miles A. Pufall[1,*,†]

[1]Department of Biochemistry, Carver College of Medicine, University of Iowa, Coralville, IA 52241, USA, [2]Integrated DNA Technologies, Inc., 1710 Commercial Park, Coralville, IA 52241, USA, [3]Department of Bioengineering, University of California, Merced, New York, NY 10027, USA, [4]Department of Biological Sciences, Columbia University, New York, NY 10027, USA and [5]Department of Systems Biology, Columbia University Irving Medical Center, New York, NY 10032, USA

Received November 27, 2019; Revised March 06, 2020; Editorial Decision March 26, 2020; Accepted March 27, 2020

ABSTRACT

CRISPR RNA-guided endonucleases (RGEs) cut or direct activities to specific genomic loci, yet each has off-target activities that are often unpredictable. We developed a pair of simple *in vitro* assays to systematically measure the DNA-binding specificity (*Spec-seq*), catalytic activity specificity (*SEAM-seq*) and cleavage efficiency of RGEs. By separately quantifying binding and cleavage specificity, *Spec/SEAM-seq* provides detailed mechanistic insight into off-target activity. Feature-based models generated from *Spec/SEAM-seq* data for SpCas9 were consistent with previous reports of its *in vitro* and *in vivo* specificity, validating the approach. Spec/SEAM-seq is also useful for profiling less-well characterized RGEs. Application to an engineered SpCas9, HiFi-SpCas9, indicated that its enhanced target discrimination can be attributed to cleavage rather than binding specificity. The ortholog ScCas9, on the other hand, derives specificity from binding to an extended PAM. The decreased off-target activity of AsCas12a (Cpf1) appears to be primarily driven by DNA-binding specificity. Finally, we performed the first characterization of CasX specificity, revealing an all-or-nothing mechanism where mismatches can be bound, but not cleaved. Together, these applications establish *Spec/SEAM-seq* as an accessible method to rapidly and reliably evaluate the specificity of RGEs, Cas::gRNA pairs, and gain insight into the mechanism and thermodynamics of target discrimination.

INTRODUCTION

Clustered Regularly Interspaced Short Palindromic Repeat (CRISPR)-based RNA-guided endonucleases (RGEs) have emerged as critical tools in genetic research. The CRISPR-associated (Cas) protein 9 isolated from *Streptococcus pyogenes* (SpCas9), is widely used and easily programmed with a CRISPR RNA (crRNA) annealed to a structural tracrRNA or fused as a single guide RNA (sgRNA) to target complementary DNA sequences (1–4). Catalytically deactivated SpCas9 (dSpCas9) has been adapted for manipulation of gene expression and chromatin status as well as visualization of chromosomal loci (3,5–7). However, binding (8–10) and cutting (11–13) by SpCas9 ribonucleoprotein complexes (SpCas9-RNPs) are also prevalent at off-target sites. New CRISPR associated proteins, either engineered from SpCas9 (14–16) or identified *de novo* from microbes are becoming available at an increasingly rapid rate (17,18). More recently identified enzymes include: (i) *Streptococcus canis* Cas9 (ScCas9), which has been reported to have single-base pair specificity within the PAM (5′-NNG-3′) (19), (ii) Cas12a (20) (aka Cpf1), a class II type V CRISPR enzyme that has been reported to have improved targeting properties, and (iii) CasX (a.k.a. Cas12e), another class II CRISPR enzyme (17,18). Rapid and accessible methods are needed to evaluate the specificity and mechanisms of these new enzymes.

The mechanism underlying SpCas9-RNP targeting (1,21) has been extensively studied. The SpCas9-RNP cuts DNA in a two-step process. First, the SpCas9-RNP binds to a protospacer adjacent motif (PAM) composed of a consensus NGG motif and partially melts the DNA duplex. A complementary region of the gRNA then base pairs with the 20-nt protospacer DNA (1) in a sequential fashion, from the PAM toward the distal region of the protospacer (22,23). Mismatches to the PAM-proximal region (positions ∼1–10) disrupt binding by interrupting RNA:DNA 'zipping', with

*To whom correspondence should be addressed. Tel: +1 319 384 1820; Fax: +1 319 335 9570; Email: miles-pufall@uiowa.edu
†These authors contributed equally.

adjacent mismatches having a cooperative effect (24,25). In the second step, base pairing in the PAM-distal region (positions ∼12–20) induces a structural change that moves the HNH and RuvC nuclease domains into position to cut both strands of DNA upstream of position 3 in the protospacer (1). In detailed kinetic studies of a single SpCas9::gRNA complex, most (∼80%) of the on-target sequence is cleaved during a fast phase, with ∼5% being cleaved during a slow phase (21), leaving ∼15% uncut even when enzyme is in vast excess.

The off-target activities of SpCas9 vary greatly by target sequence (reviewed in (26)). Dozens to thousands of off-target cleavage sites have been detected, depending on target-sequence composition (12,27–30). Generally, mismatches have a more pronounced impact on cleavage the closer they are to the PAM (8–10). Some off-target cleavage may be due to off-target binding in the genome, as SpCas9-RNPs have been shown by ChIP-seq to bind hundreds to thousands of sites (8,31). However, occupancy is not a good predictor of cleavage (27), in part because mismatches in the PAM-distal region do not affect binding, but impair cleavage. Moreover, it is not clear why SpCas9-RNPs directed to some target sequences have more off-target activity than others.

Accurate prediction of off-target binding and cleavage for SpCas9-RNPs, or other CRISPR nucleases such as Cas12a, has proven to be challenging. Although thousands of potential binding sites have been identified for some SpCas9:gRNA pairs (8–10), *in vitro* methods to measure DNA-binding specificity, such as HiTS-FLIP (10), and CHAMP (32), have not yielded predictive, quantitative binding models for SpCas9-RNPs. *In vitro* cleavage assays, such as Digenome-seq and CIRCLE-seq, identify hundreds to thousands of off-target sites for SpCas9-RNPs (12,33), but only a fraction of predicted sites can be detected by in cell techniques such as GUIDE-seq (13,29,33,34). This can be partially accounted for by cellular factors such as chromatin context (35–38) and biases in repair outcomes (29,39,40). The Cutting Frequency Determination score (CFD) derived from a large-scale, loss-of-function screening (41) has emerged as a leading prediction algorithm—known as the 'Doench rules'—for *in vivo* off-target cleavage sites.

Assays have been published that provide specificity information for either binding or cleavage by SpCas9, but not both. More importantly, most of the techniques used to measure specificity are technically challenging, expensive, or require special equipment, restricting their wider adoption. As a consequence, characterization of the specificity of other CRISPR nucleases lags far behind that of SpCas9.

To address these issues, we developed parallel assays that together enable systematic quantification of the effect of mutations and mismatches on DNA-binding affinity and catalytic activity for RGEs. The effect of mismatches and mutations on DNA-binding specificity is measured using an adapted version of specificity measured by sequencing (*Spec-seq*) (42,43) while the endonuclease activity is measured concurrently under the same conditions using a new assay we refer to as sequence-specific endonuclease activity measurement by sequencing (*SEAM-seq*). We validated the *Spec/SEAM-seq* approach for the best-characterized RGE, SpCas9, using three SpCas9::gRNA pairs that have been used in previous studies. The models generated from *Spec/SEAM-seq* accurately capture what is known about the specificity of SpCas9-RNPs for different targets and provide new insights into the mechanism of off-target cleavage. We also used *Spec/SEAM-seq* to better understand natural or engineered Cas9 variants and Cas12a from *Acidaminococcus* sp. (AsCas12a), each of which have been reported to have improved specificity. Finally, we profiled the specificity of a previously uncharacterized CRIPSR enzyme, *Deltaproteobacteria* CasX (DpbCasX), which revealed that although the DNA-binding specificity CasX is similar to that of SpCas9, it shows a striking avoidance of off-target cleavage.

MATERIALS AND METHODS

Further information and requests for resources and reagents should be directed to and will be fulfilled by Lead Contact, Miles A. Pufall (miles-pufall@uiowa.edu). Detailed descriptions about mathematical modeling and reagent tables (Supplementary Tables S1 and S2) can be found in Supplementary Materials.

Recombinant protein expression and purification

DNA sequences encoding 6X His-tagged wild type SpCas9, Hifi-SpCas9 (R691A), AsCas12a, ScCas9, DpbCasX were cloned into pET28a vector for protein expression as previously described (44). Point mutations to express catalytically deactivated nucleases (dSpCas9: D10A/H840; dScCas9: D10A/H849A; dAsCas12a: D908A/E993A/D1235A; dDpbCasX: D672A/E769A/D935A) were created by site-directed mutagenesis. To overexpress these proteins, the transformed *E.coli* BL21DE3 cells were first grown in TB medium at 37°C with 250 rpm shaking, and induced by 0.5 mM IPTG when OD_{600} reached 0.6–0.8. The culture was further incubated at 16°C for 16–22 h. All proteins were purified using Ni^{2+} affinity (HisTrap HP) and cation exchange chromatography (HiTrap SPHP or HiTrap Heparin) as previously described (44). The ScCas9 requires further purification by size-exclusion chromatography (HiPrep Sephacryl S-300) to remove nucleic acid contaminated fractions. The purified protein was concentrated by Amicon ultrafiltration device (30-kDa), dialyzed into storage buffer overnight (20 mM Tris–HCl, pH7.4, 300 mM NaCl, 50% Glycerol, and 1 mM DTT), and stored at –20°C. The protein concentration was determined by Nanodrop using extinction coefficients of 120 450 M^{-1} cm^{-1} (SpCas9/ScCas9), 143 940 M^{-1} cm^{-1} (AsCas12a), and 158 140 M^{-1} cm^{-1} (DpbCasX). The DNA sequence for each protein used can be found in the Supplemental file: 'CRISPR_DNA_sequences.docx'.

To purify WT-DpbCasX, it was essential to co-express both protein and sgRNA in *Escherichia coli*. Briefly, the BL21DE3 cells were co-transformed with the pET28 protein expression vector, and a second plasmid constitutive transcribing the sgRNA under the control of J23119 promoter. The assembled CasX:sgRNA in *E.coli* cells was purified by Ni^{2+} affinity and cation exchange chromatogra-

phy (HiTrap Heparin). Apo-WT-CasX with minimal nucleic acid contamination ($A_{260/280} \sim 0.6$) can be recovered by size-exclusion chromatography (HiPrep Sephacryl S-300), which was then concentrated by Amicon ultrafiltration and stored in the buffer as described above, but with 0.5 M NaCl. Without co-expression, WT-CasX eluted from HisTrap HP column was heavily contaminated by nucleic acid, and highly prone to aggregation. Despite numerous efforts to optimize purification conditions, no active WT-CasX free from nucleic acids was able to be isolated.

sgRNA

Single guide RNAs (sgRNAs) for each nuclease were chemically synthesized (Integrated DNA Technologies, IDT), and dissolved in IDTE (10 mM Tris–HCl, 0.1 mM EDTA, pH 8.0) to a final concentration of 100 μM. See Supplementary Table S1 for a list of all gRNAs used.

SELEX-seq library preparation

A DNA library with a 30-bp randomized region (Supplementary Table S1, GTTCAGAGTTCTACAGTCCG ACGATC(N30)TGGAATTCTCGGGTGCCAAGG) was synthesized as single-stranded DNA (IDT). The double-stranded DNA library (dsDNA) was prepared by a Klenow extension reaction using Cy5-labeled TSSR1 primer (Supplementary Table S1) complementary to the 3′-end of the ssDNA library. Briefly, a reaction containing 2.5 μM ssDNA template, 5 μM Cy5-labeled TSSR1, and 150 μM dNTP in NEB buffer 2 (10 mM Tris–HCl, 50 mM NaCl, 10 mM $MgCl_2$, 1 mM DTT) was incubated at 94°C for 3 min, and then cooled to 37°C over 45 min. Klenow enzyme was added to the reaction and incubated at 37°C for 1 h. The enzyme was inactivated at 72°C for 20 min, followed by gradually cooling to 10°C over 45 min. The dsDNA library was purified and concentrated (Qiagen MinElute). The concentration was measured by A_{260} on a Nanodrop, and then diluted to 4 μM in Qiagen buffer EB (10 mM Tris–HCl, pH 8.5).

dSpCas9-RNP SELEX-seq

SELEX-seq of dSpCas9-RNP was performed based on our previous protocol with slight modifications (45). The ribonucleoprotein complex was first assembled using 4 μM dSpCas9 and 4.8 μM sgRNA in the reaction buffer (20 mM Tris–HCl, pH 7.4, 150 mM KCl, 10% glycerol, 5 mM $MgCl_2$ and 1 mM DTT). A 120 μl binding reaction containing 0.1 μM dSpCas9-RNP and 1 μM SELEX library (1:10 ratio) was incubated at 37°C for 1 h (20 mM HEPES–KOH, pH7.5, 150 mM KCl, 10% glycerol, 5 mM $MgCl_2$ and 1 mM TCEP), then resolved on a 4–20% gradient gel (1× TGM: 25 mM Tris-Base, 192 mM glycine, 5 mM $MgCl_2$, pH 8.3) at room temperature. The dSpCas9-RNP:DNA complex was recovered by excision and electroelution (200 V, 4°C in pre-chilled 1× TGM buffer). The bound DNA was purified (Qiagen MinElute) and diluted to 190 μl (EB buffer). Quantitative PCR was performed to determine the optimal number of rounds of amplification before saturation using 1/190 μl recovered DNA (45). The library for the next round was generated by amplifying the remaining recovered DNA (∼180 μl) in 90 PCR reactions of 100 μl each (9 ml total) using the optimal number of rounds. A new library was then generated by purifying the regenerated library (Qiagen MinElute) and diluting to 4 μM for next round of *SELEX*. Five rounds of *SELEX* were performed in total. The initial and final libraries (R0 and R5) were deep-sequenced on an Illumina HiSeq 4000 at a read depth of 25–30 million reads per library.

Analysis of SELEX-seq data using SelexGLM

Following (45), the *SELEX-seq* data were processed using the R packages *SELEX* (http://bioconductor.org/packages/SELEX, (46)) and *SelexGLM* (http://github.com/BussemakerLab/SelexGLM; Zhang & Martini & Rube & Kribelbauer & Rastogi & FitzPatrick & Houtman & Bussemaker and Pufall (45)). Briefly, a Markov model of order 6 was constructed from the R0 probes using the *selex.mm( )* function from the *SELEX* package, and an affinity table for $k = 18$ was constructed using *selex.affinities( )*. An initial position specific affinity matrix (PSAM; (47)) was constructed from the relative affinity of all single-base mutations of the optimal 18-mer ('NNNNAAGAWKGGAAGNGG'). The PSAM was then expanded to the desired size by adding nine neutral columns on each side to estimate the specificity outside of PAM and protospacer and used as a seed for *SelexGLM*. The final model was plotted as an energy logo using the LogoGenerator tool from the REDUCE Suite (reducesuite.bussemakerlab.org, (47)).

Spec-seq

The *Spec-seq* binding specificity assay was adapted from previously published protocols (42,43). Briefly, individual *Spec-seq* libraries were ordered as ssDNA from IDT, and pooled in equal proportions for Klenow extension as described above. The resulting dsDNA library was purified and concentrated (Qiagen MinElute), and size-selected on a 12% 1× Tris-glycine native gel. The DNA band of correct size was excised, electroeluted, and purified. The concentration was quantified by A_{260} and diluted to 1 μM in buffer EB.

Spec-seq was performed essentially as a single-round *SELEX* experiment. 40 μl binding reactions containing 200 nM dSpCas9-RNP and 50 nM *Spec-seq* DNA library (4:1 ratio) were incubated at 37°C for 1 h and resolved on a 4–20% gradient gel. Both bound and unbound DNA were excised, electroeluted, purified and diluted to 40 μl in buffer EB. Recovered DNA from each fraction was amplified for sequencing. Other than the initial experiments of SpCas9 (Figures 1–3), subsequent experiments on AsCas12a (Figure 4) and SpCas9/ScCas9/Hifi-Cas9 (Figure 5) were performed following the same protocol as SpCas9, but in a reaction buffer supplemented with 20 ng/μl PolydIdC (Thermo, #20148E) or Salmon Sperm DNA (Thermo, #15632011) to reduce non-specific protein:DNA binding. The DpbCasX *Spec-seq* was performed using 800 nM protein and 200 nM DNA library in the buffer with Salmon Sperm DNA.

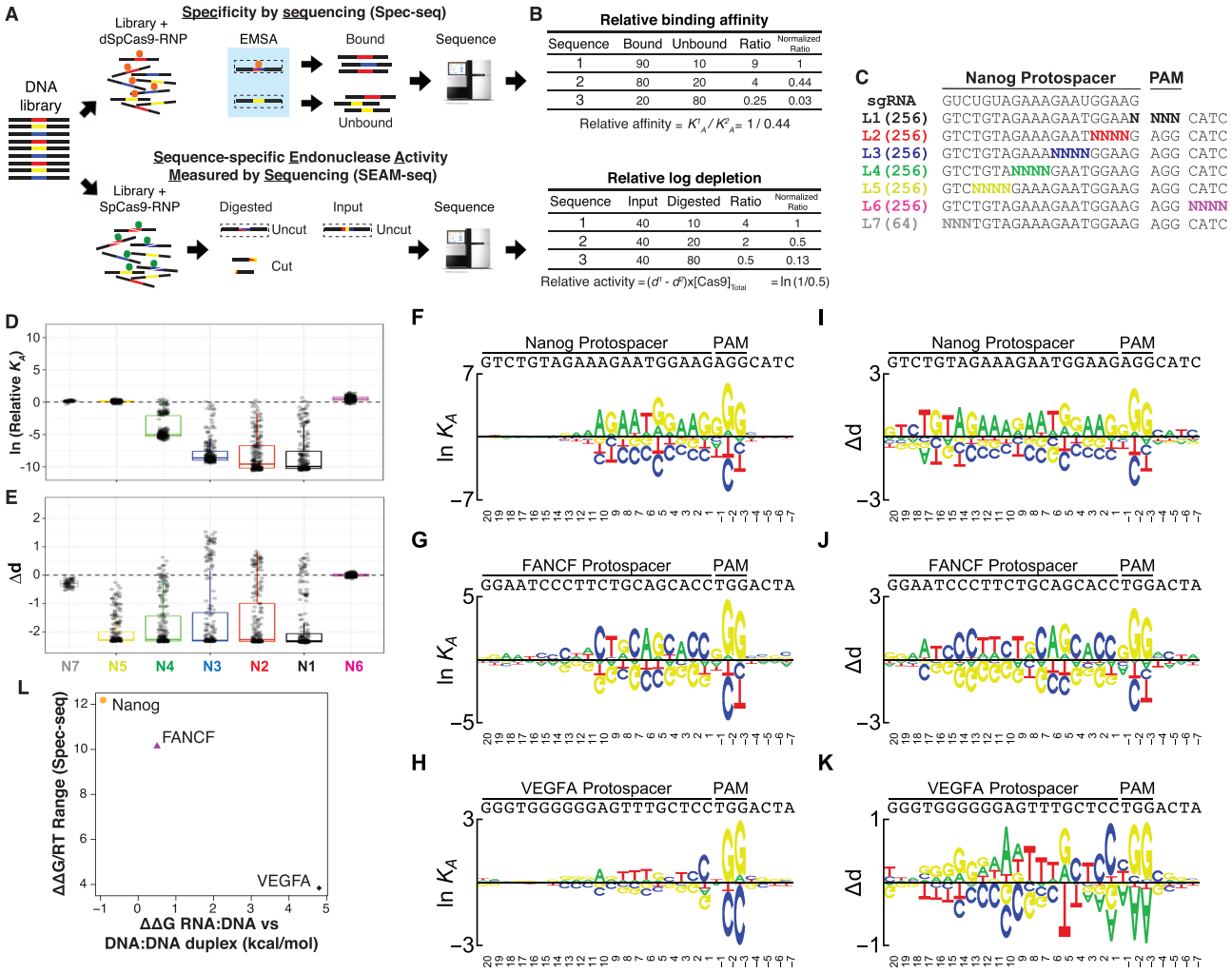

Figure 1. *Spec-seq* and *SEAM-seq* measure the affinity and cleavage activity of RNA-guided endonucleases for on-target and mismatched sequences. (**A**, **B**) Schematic of the *Spec-seq* and *SEAM-seq* protocols. For *Spec-seq*, sequences bound by the catalytic-deactivated dSpCas9-RNP are resolved from unbound on an EMSA gel. Each band is excised and deep-sequenced, with the relative affinity directly calculated (B) from the ratio of bound to unbound (see Materials and Methods). For *SEAM-seq*, the same sequences are cleaved by wild-type SpCas9-RNP. Uncut sequences are amplified by PCR and deep-sequenced. The ratio of uncut sequences in digested to input fraction reflects the relative cleavage of sequences (B), which is normalized to calculate the relative log-depletion (Δd, see Materials and Methods). (**C**) Library design for Nanog. The full library is composed of seven sub libraries; six (L1–6) containing a randomized 4-mer region, and one (L7) a 3 bp randomized region. The sub libraries were mixed equally before use in *Spec-seq* and *SEAM-seq*. (**D**) The natural log of the relative affinity for each sequence is binned by sub library. (**E**) The relative log-depletion for each sequence is binned by sub library. Sequences >0 are cleaved more completely than the on-target sequence after 1 h, those <0 are cleavage less completely. (**F–H**) Sequence logos showing the mononucleotide coefficients in the Mono + NS models of the relative affinity (see Materials and Methods). The height of each letter represents the magnitude of corresponding coefficient (contribution to affinity) and inverted letters below the center line indicate negative values. (**I–K**) Sequence logos representing the mononucleotide coefficients (contribution to cleavage) in the Mono + NS models for the effect of mismatches on the cleavage for each SpCas9-RNP. (**L**) The estimated (Melting 5) difference in the stability for the RNA:DNA duplex minus the stability of the DNA:DNA duplex ($\Delta\Delta G_{exchange}$) is plotted versus the affinity range for the worst mismatched sequences versus best sequences.

SEAM-seq

Sequence-specific Endonuclease Activity Measurement followed by sequencing (*SEAM-seq*) were performed in 50 μl reactions containing 50 nM *Spec-seq* DNA library with either 200 nM Cas-RNP or reaction buffer as mock treated. Reactions were stopped after 1 hour of incubation at 37 °C by adding EDTA (60 mM final) to both digested and input samples. Proteinase K (10 μl, 20 mg/ml, Thermo) was then added to digest proteins for 30 min at room temperature. The remaining DNA was purified (Qiagen MinElute) and eluted in 40 μl. The full-length, uncut DNA library was PCR amplified and deep-sequenced. The reaction buffer for AsCas12a (Figure 4), SpCas9/ScCas9/Hifi-Cas9 (Figure 5) and DpbCasX was also supplemented with 20 ng/μl PolydIdC or Salmon Sperm DNA.

Rate and fraction cleaved for individual sequences

DNA sequences containing the target sites of SpCas9 were ordered as gBlock fragments (IDT) (Supplementary Table S2), PCR amplified, and purified as the substrates for cleavage reactions. The functional gRNAs were assembled us-

ing equal molar ratios of synthetic crRNA and tracrRNA (IDT) in duplex annealing buffer (100 mM potassium acetate, 30 mM HEPES, pH 7.5). The annealing reactions were incubated at 94°C for 3 min, and gradually cooled to 25°C over 30 min. To assemble the SpCas9–RNP complex, the annealed gRNA was incubated with SpCas9 in the reaction buffer (20 mM HEPES, pH 7.5, 150 mM KCl, 10% glycerol, 5 mM $MgCl_2$ and 1 mM DTT) at 37°C for 10 min. The assembled RNP was diluted in reaction buffer, and combined with 10 nM DNA substrate to initiate cleavage. The reaction was quenched at multiple time-points with 50 mM EDTA. Samples were treated with Proteinase K (1 μg/μl, 56°C, 30 min) to digest SpCas9 protein, and then analyzed by capillary electrophoresis (Fragment Analyzer, Agilent). The substrate cleavage (%) was calculated as:

$$\text{Cleavage}\,(\%) = \frac{\frac{C_{F1}+C_{F2}}{2}}{\frac{C_{F1}+C_{F2}}{2} + C_{\text{substrate}}}$$

C_{F1} and C_{F2} are the molar concentration of two cleaved fragments, and $C_{\text{substrate}}$ is the molar concentration of undigested DNA substrate.

Measuring the editing efficiency of Cas9 in HEK293 cells

The ribonucleoprotein complexes were assembled using 12 μM of Cas9 and 20 μM of sgRNA in PBS, and incubated at room temperature for 10 min. The Cas9-RNPs (2 μM) were co-delivered with 3 μM Cas9 electroporation enhancer (IDT) into HEK293 cells (0.24 million cells per reaction) by Lonza SF Cell Line 96-well Nucleofector™ Kit using 96-DS-150 program. The electroporated cells were transferred to a 96-well culture plate, and incubated in 125 μl DMEM (10% FBS) medium for 48 h at 37°C with 5% CO_2. Cells were then washed by PBS, and lysed in 50 μl QuickExtract solution (Lucigen) per well (65°C: 6 min and 98°C: 2 min). The lysate was diluted with 100 μl nuclease-free H_2O (IDT), and stored in −20°C prior to PCR. The HPRT locus from edited cells was amplified using HPRT-F and HPRT-R primers (Supplementary Table S1) using Kapa Hifi HotStart polymerase (Roche, KK2501) (48). The editing efficiency was determined by T7 endonuclease I assay (48).

Quantification, modeling and statistical analysis

Estimation of the binding affinity using Spec-seq data. We followed a previously described framework for interpreting *Spec-seq* data (42,43). DNA sequences matching the Spec-seq library design were counted using SELEX R package (http://bussemakerlab.org/software/SELEX/). The relative DNA-binding affinity was estimated by taking the ratio of sequencing counts between bound and free fractions. Mathematical details for this analysis and subsequent modeling are provided in Supplementary Materials.

Estimation of the binding affinity using Spec-seq data. The binding of a protein P to DNA probe i is governed by a single-site equilibrium binding model:

$$\text{DNA}_i + \text{P} \overset{K_{\text{A},i}}{\leftrightarrow} \text{P} : \text{DNA}_i$$

To estimate the association constant $K_{\text{A},i} = [\text{P} : \text{DNA}_i]/([\text{DNA}_i][\text{P}])$ for each sequence, it is assumed that the (expected) read counts in the bound and unbound *Spec-seq* libraries are proportional to the corresponding concentrations:

$$\frac{n_{\text{bound},i}}{n_{\text{bound},j}} = \frac{[\text{P} : \text{DNA}_i]}{[\text{P} : \text{DNA}_j]}$$

$$\frac{n_{\text{unbound},i}}{n_{\text{unbound},j}} = \frac{[\text{DNA}_i]}{[\text{DNA}_j]}$$

The binding affinity for probe i relative to the on-target probe ($i = 0$) can then be estimated as follows:

$$K^{\text{rel}}_{\text{A},i} \equiv \frac{K_{\text{A},i}}{K_{\text{A},0}} = \left(\frac{[\text{P} : \text{DNA}_i]}{[\text{P} : \text{DNA}_0]}\right) \Big/ \left(\frac{[\text{DNA}_i]}{[\text{DNA}_0]}\right) = \left(\frac{n_{\text{bound},i}}{n_{\text{bound},0}}\right) \Big/ \left(\frac{n_{\text{unbound},i}}{n_{\text{unound},0}}\right)$$

Calculation of relative log-depletion activity using SEAM-seq. To estimate d^i from the data, we assume that the expected read counts in the initial and the digested *SEAM-seq* libraries are proportional to the corresponding concentrations:

$$\frac{n_{\text{input},i}}{n_{\text{input},j}} = \frac{[\text{DNA}_i]_U\,(0)}{[\text{DNA}_j]_U\,(0)}$$

$$\frac{n_{\text{digested},i}}{n_{\text{digested},j}} = \frac{[\text{DNA}_i]_U\,(t_E)}{[\text{DNA}_j]_U\,(t_E)}$$

It follows that it is possible to estimate the different in log-depletion between any two sequences as follows:

$$d_i - d_j = \ln\left[\left(\frac{n_{\text{input},i}}{n_{\text{input},j}}\right) \Big/ \left(\frac{n_{\text{digested},i}}{n_{\text{digested},j}}\right)\right]$$

We define the *relative log-depletion* $\Delta d_i \equiv d_i - d_{\text{on-target}}$ to be the log-depletion of probe i normalized by the mean log-depletion of all the on-target sequences covered by the control library (Cas9: sub-library 6; Cas12a: sub-library 8).

DATA AND SOFTWARE AVAILABILITY

The sequencing data for this project are available from the Sequencing Read Archives (https://www.ncbi.nlm.nih.gov/sra), accession numbers PRJNA547810 and PRJNA608749. Which data sets are available under each accession number can be found in the Key Resources Table of the Supplementary Materials. Scripts and processing software are available upon request.

RESULTS

Unbiased determination of dSpCas9::Nanog binding specificity using SELEX-seq and SelexGLM

The full DNA-binding footprint and specificity of dSpCas9, the catalytically inactive form of SpCas9 in complex with a gRNA targeting the Nanog gene (dSpCas9::Nanog) (3,8,49), was first determined using systematic evolution

of ligands by exponential enrichment measured by deep sequencing (*SELEX-seq*; (46,50)) followed by data analysis using a feature-based generalized linear modeling (*SelexGLM*; (45)) generating a position specific affinity matrix (PSAM) (Materials and Methods and Supplementary Figure S1A–C). After five rounds of enrichment, a motif emerged, dominated by the 3 bp protospacer adjacent motif (PAM; NGG at positions −1 to −3) and an 11bp PAM-proximal 'seed' region within the protospacer (Supplementary Figure S1C). The observed seed is slightly longer than previous estimates in the 8–10 bp range (8,25,31) with little or no sequence preference apparent in the PAM-distal region (+12 through +20). A weak preference for the sequence CGGGGATTT is seen at positions bases −4 to −12 downstream of the PAM, where contacts between the SpCas9 protein and DNA have been observed in Cryo-EM structures (e.g. RCSB 5Y36, (51)). This footprint was used as the basis for designing libraries used in subsequent experiments that encompassed the protospacer and a region downstream of the PAM.

Parallel measurement of DNA binding and cleavage by SpCas9 using Spec-seq and SEAM-seq

SpCas9-RNP endonuclease activity entails the two-step process of DNA binding followed by cleavage (Supplementary Figure S2A). The overall efficiency of the enzyme depends on its binding affinity (K_A) and cleavage rate (k_{cut}). To quantify these two steps, we adapted the *Spec-seq* assay (42,43) to measure dSpCas9 DNA-binding specificity and developed a new assay named *SEAM-seq* to measure cleavage by depletion (Δd) of sequences (Figure 1A, B). Based on the *SelexGLM* footprint we designed libraries that span the protospacer, PAM and 4 (of the 9) bases downstream of the PAM (Figure 1C). We then directly measured the effect of mismatches on relative binding affinity ($-\Delta\Delta G/RT = \ln K_{A,rel}$) and relative nuclease activity (in terms of the log-depletion Δd) under identical conditions (see Materials and Methods and Supplementary Materials). By correcting for occupancy using the K_A for each sequence the effect of mismatches on binding and cleavage are decoupled, allowing calculation of k_{cut}, a measure of the effect of mismatches on cleavage independent of their effect on affinity ('binding independent cleavage'). The relative efficiency of cleavage for off-target sequences is then modeled as the product of affinity and binding-independent cleavage ($K_{eff} = K_A * k_{cut}$).

DNA binding of SpCas9 is mediated by PAM-proximal RNA:DNA base-pairing

The DNA-binding specificity of dSpCas9::Nanog (3,8,49) was measured by *Spec-seq*. The pooled DNA library (Figure 1C) was incubated with the dSpCas9::Nanog at a 1:4 ratio (50 nM DNA:200 nM RNP). The bound and unbound fractions of DNA were separated on a native gel (Supplementary Figure S2B), recovered, and sequenced. After ensuring that sub-libraries were similarly represented in the full library (Supplementary Figure S2D), the relative binding affinity of dSpCas9-RNP for each site was calculated directly as the ratio of the probe frequencies in the bound and unbound fractions (see Materials and Methods). Although qualitatively consistent, the dynamic range of the *Spec-seq* model (Figure 1D) is greater than the SelexGLM model (Supplementary Figure S1C), possibly because the latter does not account for binding saturation or non-specific binding. The affinities for on-target and mismatched sequences spanned four orders of magnitude with high reproducibility (Supplementary Figure S2E). Consistent with the *SELEX-seq*/*SelexGLM* analysis described above, mismatches within the PAM-proximal region of the protospacer and mutations within the PAM site had the strongest effect on the DNA affinity (Figure 1D) with little penalty for mismatches in the PAM-distal region. Contributions from the PAM downstream region were also modest.

To determine the energetic cost of mismatches and mutations, feature-based models for binding specificity were fit to the *Spec-seq* data. A simple weight-matrix-like model based on mononucleotide features, in which each position contributes independently (Model *Mono*), fit the data well ($R^2 = 0.945$, hold-one-out cross-validation) (Supplementary Figure S3A). The fit improved, especially for lower-affinity sites, by incorporating non-specific binding in the model (Model *Mono+NS*, $R^2 = 0.972$, Figure 1F, Supplementary Figure S3B). Adding dinucleotides as predictors did not substantially improve the fit (Model *Di+NS*, $R^2 = 0.975$, Supplementary Figure S3C, D), and nearly quadrupled the fit parameters, and thus all subsequent analyses used the *Mono+NS* model. Comparison between the energy logo representation of the *Spec-seq* model (Figure 1F) and the *SELEX-seq*/*SelexGLM* model (Supplementary Figure S1C) confirmed that although the logos were very similar in footprint and sequence preference, the *Spec-seq* model captured a larger dynamic range.

The tolerance for mismatches is strongly dependent on gRNA sequence

Off-target genomic cleavage by SpCas9-RNPs varies by target sequence and, accordingly, gRNA (12,13). We therefore compared dSpCas9::Nanog to dSpCas9::FANCF, which is similarly specific, and dSpCas9::VEGFA, which has been reported to have substantially higher off-target cleavage activity than SpCas9::FANCF in *in vivo* GUIDE-seq assays (12,13), and *in vitro* SITE-seq assays (52). For all three dSpCas9-RNPs, mismatches in the PAM had the greatest effect on affinity. However, dramatic differences among the three gRNAs were evident in the overall degree of DNA-binding specificity within the seed region (Figure 1F-H). The range of $\Delta\Delta G$ values (compared to the average sequence in the protospacer) encompassed ∼12 RT for dSpCas9::Nanog, ∼8 RT for dSpCas9::FANCF, but only ∼3 RT for dSpCas9::VEGFA (Figure 1F–H). In other words, dSpCas9::Nanog is orders of magnitude more specific than dSpCas9::VEGFA because the effect of a mismatch in dSpCas9::Nanog can be >100× larger than for dSpCas9::VEGFA.

The models also captured the energetic penalty for different types of mismatch. Consistent with previous work, purine:purine (R:R) mismatches tended to be more deleterious than purine:pyrimidine (R:Y) or pyrimidine:pyrimidine (Y:Y) mismatches. For example, for adenines in the gRNA (rA), a mismatch with a dG in the targeted DNA strand

(represented as C in logo, Supplementary Figure S4A) was much less favorable than dA or dC (Supplementary Figure S4B). However, there were examples where this was not the case. For example, dC (represented as G in the logos) was the worst mismatch for rCs in gRNAs. Interestingly, for both rG and rU, the dT mismatch was less detrimental than the other two (Supplementary Figure S4A, B). These trends did not necessarily follow the published stability data of RNA:DNA duplex with single mismatch (such as rA:dG ≈ rA:dC > rA:dA) (53). This may be because the penalty for a mismatch depends on both the stability of the newly gained RNA:DNA duplex and that of the invaded DNA:DNA duplex.

dSpCas9-RNP specificity correlates with exchange from DNA:DNA duplex to RNA:DNA duplex

To explore whether the differences in specificity between dSpCas9-RNPs could be attributed to the energetic difference between the DNA:DNA duplex being invaded and formation of the RNA:DNA duplex in the R-loop, we used MELTING v5 (https://www.ebi.ac.uk/biomodels/tools/melting/) to calculate $\Delta\Delta G_{\text{exchange}}$ (the free energy difference in melting between the corresponding DNA:DNA and RNA:DNA duplexes) (54). Driven by higher GC content, the estimated DNA:DNA annealing free energy for the VEGFA on-target sequence (70% GC, −32.4 kcal/mol) was most favorable, followed by that for FANCF (60% GC, −25.9 kcal/mol), and then Nanog (40% GC, −21.6 kcal/mol). However, because both base pairing and pyrimidine/purine content of RNA affect the stability of RNA:DNA hybrids, the annealing energy of the RNA:DNA hybrids showed a different trend. Although VEGFA formed the most stable RNA:DNA duplex (−27.9 kcal/mol), its DNA:DNA duplex was even more stable, making exchanging a DNA:DNA helix for an RNA:DNA duplex energetically unfavorable ($\Delta\Delta G_{\text{exchange}}$ = 4.5 kcal/mol). In contrast, the RNA:DNA hybrid duplex formed by the Nanog RNA is predicted to be more stable than the DNA:DNA duplex ($\Delta\Delta G_{\text{exchange}}$ = −0.9 kcal/mol). The difference in annealing energy for FANCF falls in between these two cases and is quantitatively consistent with the Spec-seq data (Figure 1L). This correlation suggests that because the exchange from DNA:DNA to RNA:DNA for VEGFA is already energetically unfavorable for the on-target sequence, mismatches have little room to destabilize the complex; the enzyme is therefore less sensitive to them and less specific. On the other hand, because the exchange from DNA:DNA to RNA:DNA for Nanog contributes to the overall binding energy of the dCas9-RNP, mismatches destabilize binding, resulting in greater specificity. Thus the $\Delta\Delta G_{\text{exchange}}$ of RNA:DNA for DNA:DNA may be an important influence on the specificity of SpCas-RNPs for target sequences.

Sequence-specific endonuclease activity measured by sequencing (SEAM-seq)

We next developed the *SEAM-seq* method to measure how the endonuclease activity depends on sequence *in vitro*. This method was *explicitly* designed to be paired with *Spec-seq* (Figure 1A): the DNA library was cleaved for the same amount of time (1 h) and under the same conditions ([DNA] = 50 nM, [SpCas9-RNP] = 200 nM, same buffer) as those used in the *Spec-seq* assay. The remaining uncut DNA ('digested' library) was separated on a gel (Supplementary Figure S2C), amplified by PCR, and sequenced along with a mock-treated uncut control ('input'). The depletion ratio (input/digested) was calculated for each sequence (Figure 1B). The relative depletion, Δd, of the SpCas9-RNP is defined as the logarithm of this ratio, normalized by of the mean depletion of all on-target sequences in sub-library 6 (See Materials and Methods, Supplementary Materials). Sequences with $\Delta d < 0$ are cleaved less completely than the on-target sites, whereas those with $\Delta d > 0$ are cleaved more completely (Figure 1E). As with *Spec-seq*, *SEAM-seq* is reproducible, with a high correlation ($\rho = 0.98$, $R^2 = 0.96$, Supplementary Figure S2F) between replicates. A small fraction of probes (~3%) differed significantly ($\sigma > 3$) between the replicates (a difference that may be due to slightly more protein in Rep 1) with the average Δd from the two replicates was used in downstream analyses.

Full SpCas9-RNP endonuclease activity requires base pairing throughout the protospacer

SEAM-seq was performed using the same three gRNAs (Nanog, FANCF, VEGFA), this time with clevage competent SpCas9. As with *Spec-seq*, the data were fit to a *Mono+NS* model (Figure 1I–K). Consistent with previous work (1), the *SEAM-seq* model captured the requirement for base-pairing throughout the 20-nt protospacer for full endonuclease activity, with mismatches at positions +18 through +20 having a less prominent effect. Variations in sequences downstream of the PAM had only a small effect on cleavage, consistent with their effect on affinity. The activity of each SpCas9::gRNA pair was different, with SpCas9::Nanog and SpCas9::FANCF being the most active (~50-fold relative depletion) compared to SpCas9::VEGFA (~4-fold relative depletion (Figure 2, Y-axis)).

SpCas9 affinity does not always correlate with endonuclease activity

Quantification of both binding and cleavage specificity in parallel enabled mechanistic analysis of target discrimination by SpCas9-RNPs. Analysis of affinity versus activity plots (ln K_A versus Δd, Figure 2) show that, in general, reductions in K_A accompany reductions in cleavage. This correlation is, as expected, strongest for alterations of the PAM and mismatches within the PAM-proximal region (positions −3 to +11; black, red, blue, and part of green sub-library). Importantly, the libraries covering the PAM-proximal region exhibit a sigmoidal relationship in the ln K_A versus Δd plot, suggesting that the higher-affinity probes are saturated in the *SEAM-seq* assay (see also below). In contrast, in the PAM-distal region of all three SpCas9-RNPs mismatches affect Δd specifically, consistent with previous observations (1,14), but have little effect on affinity (positions 12–20, yellow, grey and part of green libraries, seen as vertical bands in Figure 2).

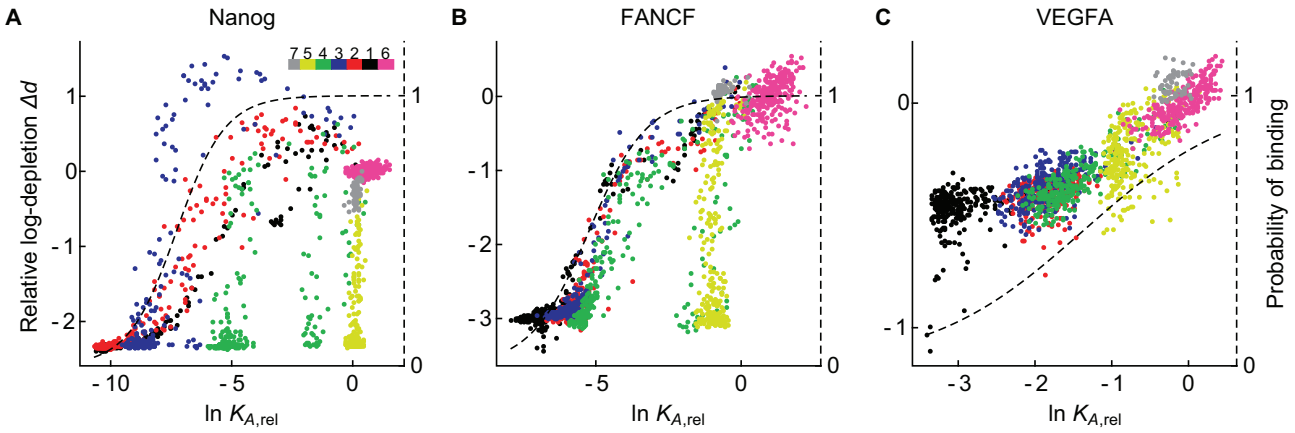

Figure 2. The effect of mismatches on affinity and activity differ by target. Affinity versus Activity plots where each point represents a particular sequence in the library. The horizontal axis corresponds to the relative binding free energy (logarithm of the relative association constant), with low-affinity sequences on the left, and high-affinity sequences on the right. The vertical axis corresponds to the relative depletion of a sequence (Δd), with the most depleted at the top and the least depleted at the bottom. The on-target sequence is by definition located at the origin (0,0) of the plot. (**A**) Scatter plot showing the affinity (x) versus activity (y) for Nanog. The dashed line and corresponding vertical axis on the right denote the probability of binding calculated from the affinity of the dSpCas9-RNP for each sequence under the conditions of the experiment using our model (note that the relative shift and scale of the vertical axes are arbitrary). This probability is used to estimate occupancy and is subsequently used to correct for saturation in the calculation of k_{cut} (see Supplemental Materials). The inset indicates location of the sub library in the target sequence, with black encompassing the PAM. Note alterations downstream of the PAM (pink library, L6) have some effect on affinity, but little effect on cleavage. In contrast, sequences in the PAM distal region (yellow library and part of green) form vertical streaks indicating that mismatches in this region have no effect on binding, but drastically alter cleavage (See Figure 1C). Some single mismatches within the PAM proximal region (library L2, L3, red and blue) have a $\Delta d > 1$, indicating that they are cleaved better under these conditions. (**B**, **C**) Affinity versus activity plots for the VEGFA and FANCF. Note the differences in scale and the absence of sequences with a $\Delta d >$ 1 for both SpCas9-RNPs.

Intriguingly, although most mismatches within the PAM-proximal region impair both binding and endonuclease activity, some resulted in $\Delta d > 0$. We refer to this phenomenon as 'mismatch activation.' It was observed primarily for SpCas9::Nanog (Figure 2A), the most specific SpCas9-RNP tested. Single mismatches throughout the PAM-proximal seed region impair SpCas9::Nanog binding (Figure 2A, Supplementary Figure S5A) while enhancing the cleavage (Figure 2A, Supplementary Figure S5B). This apparent increase in cleavage is most evident between positions +5 and +10 and is largest at positions +7 and +8. Pairs of adjacent mismatches in this region do not exhibit mismatch activation and reduce binding affinity more than expected given the effect of each single mismatch (Supplementary Figure S5C). Similarly, adjacent mismatches within the 'seed' region resulted in less than expected cleavage, but mismatches in the PAM-distal regions were largely additive (Supplementary Figure S5D). The lack of mismatch activation and greater than additive effect of adjacent mismatches on cleavage is thus likely due to their effect on affinity.

Mismatch activation is a result of more complete cleavage of off-target sites by SpCas9::Nanog

To understand the mechanism of mismatch activation, five DNA sequences that exhibited decreased binding affinity but increased cleavage were compared to the on-target sequence (Figure 3A). The cumulative cleavage of each sequence by SpCas9::Nanog was measured by capillary electrophoresis (Fragment Analyzer, AATI) at time points from 20 s to 1 h from reactions incubated with increasing amounts of SpCas9-RNP (Supplementary Figure S6A–E). The eventual fraction cleaved reached a maximum at a SpCas9-RNP:DNA ratio of 10:1, which was then used in downstream analysis.

As described in previous work (21), the cleavage data fit best to a two-phase exponential model composed of a fast and a slow step (Figure 3B). The on-target sequence initially cleaved rapidly ($k_{fast} = 0.161\ s^{-1}$), with a slow second phase ($k_{slow} = 3.35 \times 10^{-3}\ s^{-1}$) but plateaued at 85% after 1 h (Figure 3B–D, red). This incomplete cleavage is observed even at ~100× excess of SpCas9::Nanog (Supplementary Figure S6A), and is consistent with previous observations (21). Each of the mismatch-activated off-target sequences assayed exhibited slower cleavage for both steps ($k_{fast} = 0.025$–$0.122\ s^{-1}$, $k_{slow} = 0.77$–$1.61 \times 10^{-3}\ s^{-1}$, Figure 3C), sometimes taking >20 min to plateau. Despite slower cleavage, each mismatched site was cleaved more completely (89–95% versus 85%, Figure 3B, D) than the perfect match. The fraction eventually cleaved (1 h) correlates well with the results from *SEAM-seq* (Figure 3D), validating this higher-throughput assay. Together, these results indicate that a fraction of on-target bound SpCas9::Nanog does not cleave, and that mismatch-activation results from more complete cleavage, despite slower kinetics.

In contrast, mismatch-activation was not evident in the Affinity versus Activity plots for SpCas9::FANCF or SpCas9::VEGFA (Figure 2B, C). Further, although mismatches have a similar and pronounced effect on Δd for both SpCas9::Nanog and SpCas9::FANCF, the effect of mismatches is weaker for SpCas9::VEGFA. The origin of these differences was tested by measuring the completeness and rate of cleavage for SpCas9::FANCF and SpCas9::VEGFA acting on their respective cognate DNA targets (Figure 3E). The kinetics of cleavage for SpCas9::FANCF is similar to SpCas9::Nanog (k_{fast} =

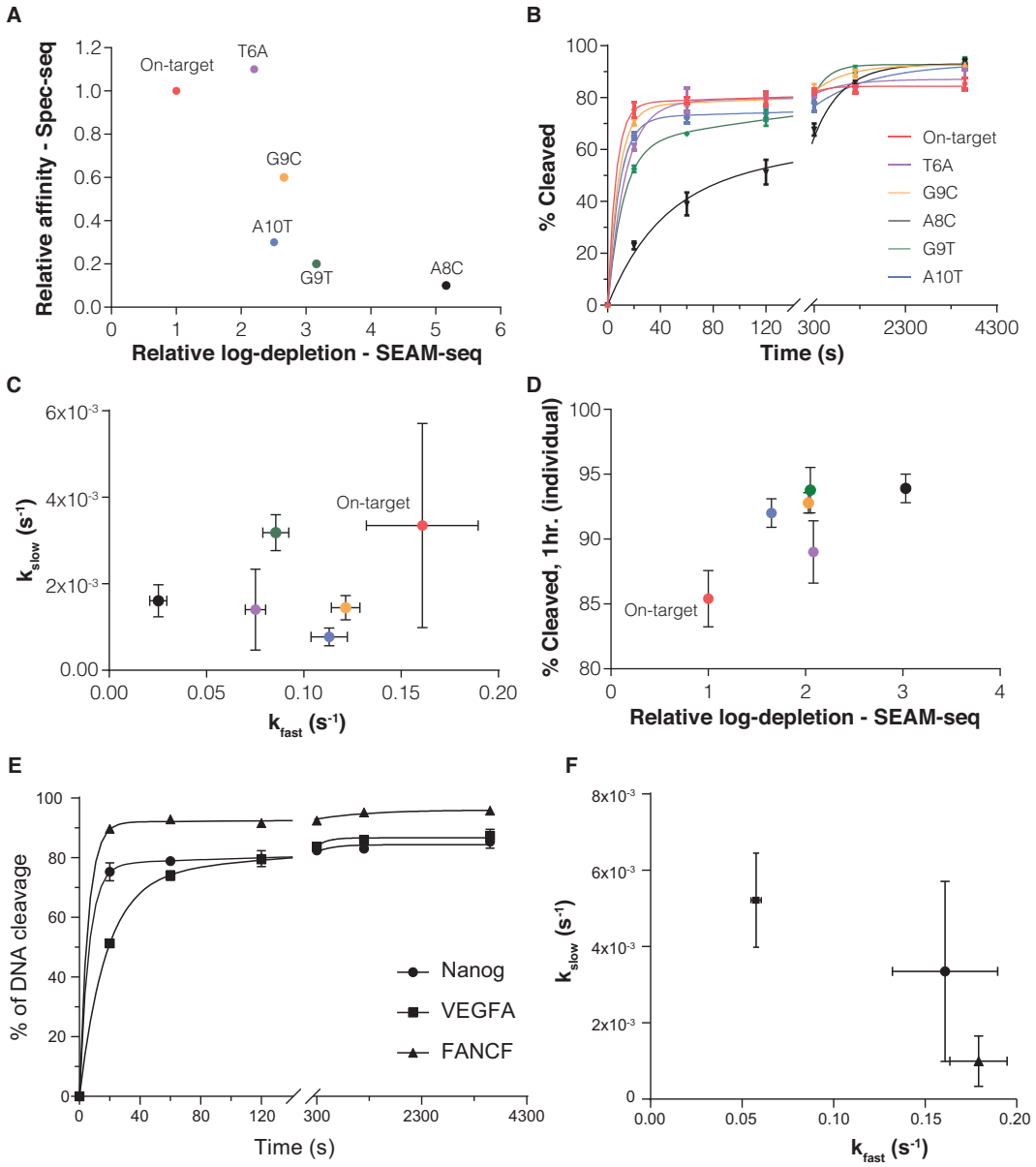

Figure 3. Mismatch activation is the result of more complete cleavage, not an increase in cleavage rate for SpCas9::Nanog. (**A**) Plot of the relative cleavage activity of SpCas9::Nanog for each sequence measured by *SEAM-seq* (x-axis) versus the relative affinity of dSpCas9::Nanog measured by *Spec-seq*. (**B**) Percent of each sequence (10 nM) cleaved over time (in seconds) by SpCas9::Nanog (100 nM). (**C**) The cleavage reactions in (**B**) were fit to a two-phase exponential association equation (GraphPad). Fast (k_{fast}) and slow (k_{slow}) phases are plotted against each other. (**D**) The relative cleavage activity of SpCas9::Nanog for each sequence measured by *SEAM-seq* is plotted against the percent cleaved after one hour for each individual sequence. (**E**) Percent of on-target sequence (10 nM) cleaved over time (in seconds) by each SpCas9-RNP (100 nM). (**F**) The cleavage reactions in (E) were fit to a two-phase exponential association equation (GraphPad). Fast (k_{fast}) and slow (k_{slow}) phases are plotted against each other. Each experiment was performed a minimum of 3 times, with error bars representing the standard deviation.

0.179 versus 0.161 s^{-1}), with SpCas9::VEGFA substantially slower (k_{fast} = 0.058 s^{-1}) (Figure 3F). However, SpCas9::FANCF cleavage of the on-target site was more complete (96%) than SpCas9::Nanog (85%) and VEGFA (86%), even after 1 h (Figure 3E). In other words, both SpCas9::Nanog and SpCas9::FANCF cut quickly, but only SpCas9::FANCF cut to completion, leaving little room for mismatch-activation. On the other hand, the lowest-specificity complex, SpCas9::VEGFA, not only cuts slowly, but also plateaus with a substantial fraction uncut.

Comparison with Doench rules

Doench and coworkers previously constructed a Cutting Frequency Determination (CFD) scoring matrix by analyzing knock-out data for a large panel of on-target and mismatched gRNAs directed against a single gene (41). Under non-saturating conditions on naked DNA, the frequency of cutting off-target sites is determined by the efficiency of the SpCas9-RNP compared to the on-target site. We calculated the efficiency as $K_{eff} = K_{A*}\ k_{cut}$: the product of the affinity (K_A) and a surrogate of cleavage rate, the occupancy-

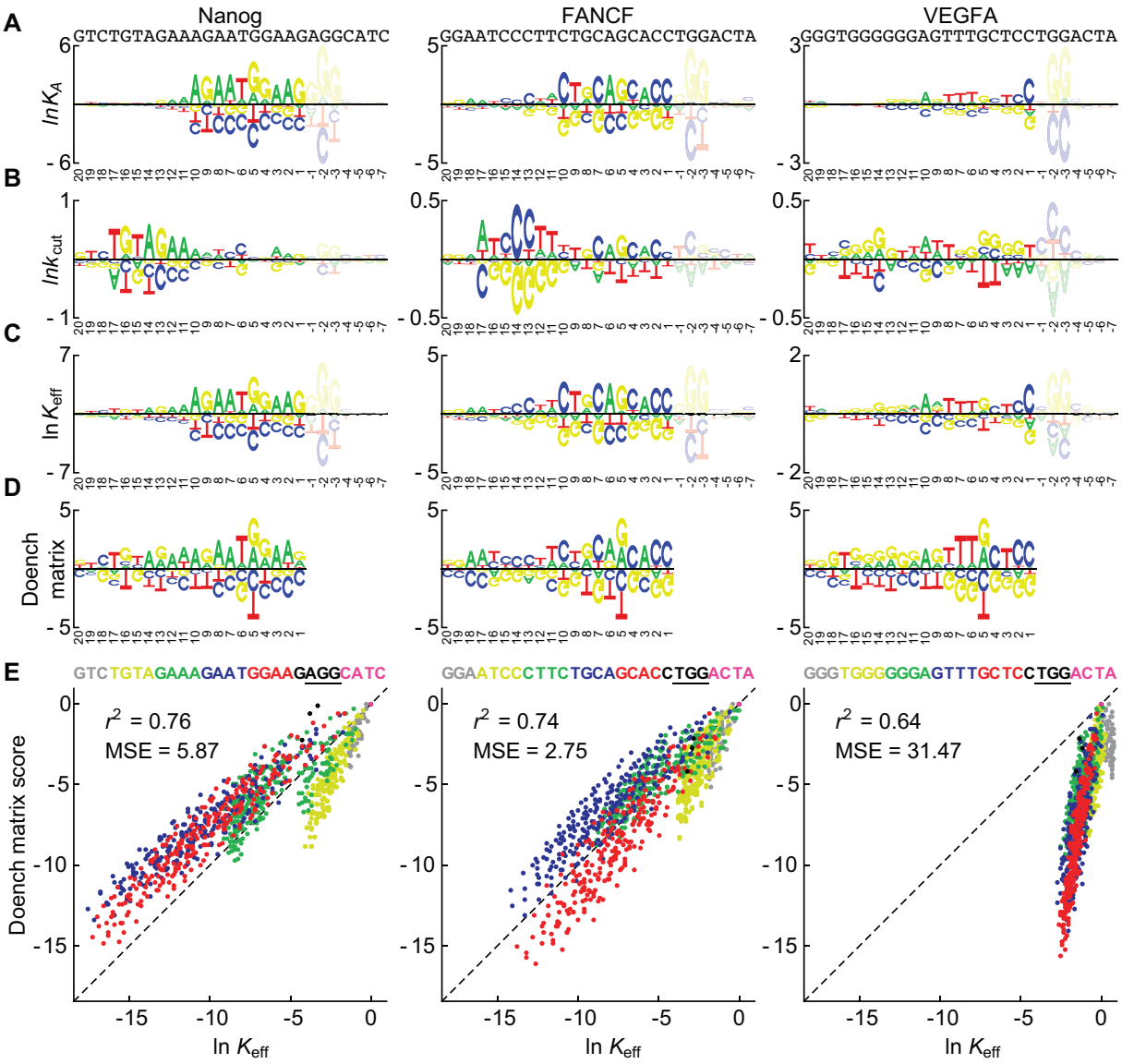

Figure 4. *In vitro* SpCas9-RNP efficiency correlates well with *in vivo* cleavage frequency. (**A**) Sequence logos showing the mononucleotide coefficients in the Mono + NS models of K_A for three different SpCas9-RNPs represent the effect of each base on the binding affinity. PAM region was excluded from the analysis, as these positions have no corresponding values in the Doench matrix. (**B**) Same as (A) but showing coefficients in the models of k_{cut}. (**C**) Logos representing contribution of each base to the efficiency K_{eff} (see Materials and Methods). (**D**) Logo for the predicted effect of mismatches on cleavage *in vivo* from the Doench rules. (**E**) Scatter plot and correlation between the *in vitro* SpCas9-RNP efficiency predicted using *Spec/SEAM-seq* versus the *in vivo* cleavage predicted using the Doench for each SpCas9-RNP. MSE = mean standard error.

corrected cleavage (k_{cut}). To predict the sequence dependence of k_{cut}, the binding affinities derived from the *Spec-seq* data were used to correct for the effects of binding saturation on Δd the *SEAM-seq* assay (assuming that the DNA-binding specificities of catalytically inactivated dSpCas9 and wild-type SpCas9 are identical). Specifically, we assumed that the log-depletion Δd for each sequence is proportional both to (i) the DNA's fractional occupancy by SpCas9-RNP (as computed using the K_A values from the *Spec-seq* data and a free protein concentration parameter inferred from the *SEAM-seq* data; see dashed line in Figure 2) and (ii) the cut rate k_{cut} (see Supplementary Materials). As with K_A and Δd, k_{cut} was modeled as the sum of a term that represents non-specific binding and another term that depends exponentially on a sum of mononucleotide effects.

Two behaviors were evident from the k_{cut} sequence logo models (Figure 4A, B). First, for SpCas9::Nanog and SpCas9::FANCF mismatches in the PAM-distal region have the strongest negative effect on k_{cut}. This agrees with expectation because k_{cut} represents the cleavage rate of bound SpCas9-RNP and mismatches in this region do not affect binding but strongly affect cleavage. SpCas9::VEGFA, which is much less specific overall, does not show the same contrast between PAM-proximal and PAM-distal positions (Figure 4B). Second, only mismatches at positions

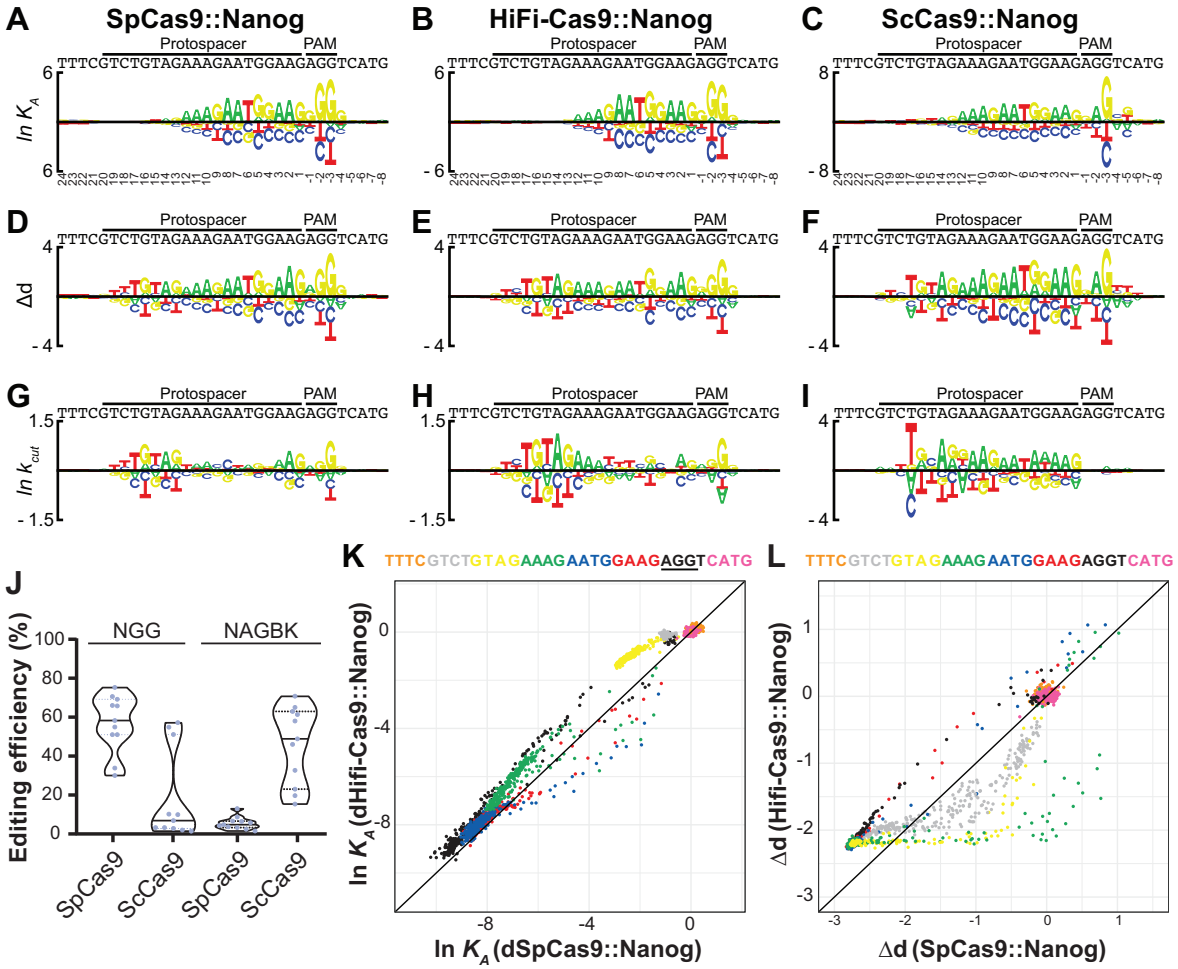

Figure 5. HiFi-Cas9 and ScCas9 derive cleavage specificity from base pairing and an extended PAM. (**A**–**C**) Sequence logo representations of the mononucleotide coefficients for the Mono + NS models of K_A for dSpCas9::Nanog (left), dHiFi-Cas9::Nanog (center) and dScCas9::Nanog (right) represent the effect of each base on the binding affinity. Note that the scale of dScCas9::Nanog is larger. (**D**–**F**) Same as (A–C) but representing the coefficients in the models of relative log-depletion (Δd). (**G-I**) Same as (A-C) but showing coefficients (occupancy independent cleavage) in the k_{cut} models (see Materials and Methods). Note the larger scale for ScCas9::Nanog indicating a greater penalty for mismatches in cleavage activity. (**J**) PAM specificity was tested by efficiency of cleavage for SpCas9 and ScCas9 at 22 sites within the HPRT gene, half of which with an NGG and half with NAGBK (NAG(C/G/T)(G/T)). Efficiency of editing was measured using a T7E1 assay. (**K**) Plot of the relative affinity (ln K_A) for each sequence in the Nanog library for dSpCas9 versus dHiFi-Cas9 as measured by *Spec-seq*. (**L**) Plot of the relative depletion (Δd) for each sequence in the Nanog library for SpCas9 versus HiFi-Cas9 as measured by *SEAM-seq*. Note that the penalty for mismatches in k_{cut} is greater for HiFi-Cas9::Nanog than SpCas9::Nanog. The patterns are consistent with a different target, FANCF (Supplementary Figure S7).

+6 through +9 for SpCas9::Nanog lead to an increase in k_{cut}, consistent with the mismatch-activation described above. Interestingly, the mismatches associated with the greatest activation (C and T) exhibit the most negative effect on affinity.

Having feature-based models for both K_A and k_{cut} allowed estimation of the enzymatic efficiency (K_{eff}, Figure 4C, see Supplemental Materials and Methods). We then compared the *in vivo* off-target scoring matrix generated from CFDs (41) (Figure 4D) to the K_{eff} matrices for each SpCas9-RNP (Figure 4E). The K_{eff} and CFD scores are highly similar for Nanog and FANCF ($r^2 = 0.76$ and 0.74, and mean-square error (MSE) 5.9 and 2.75, for Nanog and FANCF, respectively), suggesting that our *in vitro* models capture the *in vivo* preference of the SpCas9-RNP. Although the two models are highly correlated for VEGFA ($r^2 = 0.64$), the efficiency values span a much narrower range than the CFD scores, and the absolute differences are therefore much larger than for the other gRNAs (MSE = 31.5). Thus, whereas the Doench matrix assigns a shared level of specificity across all sgRNAs, *Spec/SEAM-seq* reveals that there are significant differences in the specificity of binding, cleavage, and efficiency between gRNAs. These differences in specificity are consistent with the finding that SpCas9::FANCF is highly specific, both in cells and *in vitro*, with few off-target cleavage sites observed, while SpCas9::VEGFA is much less specific, cleaving substantially more off-target sites as measured by both in GUIDE-seq and SITE-seq performed under the same conditions (28,29). This raises the question whether the specificity of SpCas9::RNPs can be accurately represented using a single universal scoring matrix with a fixed threshold for predicting off-target sites *in vivo*.

Spec/SEAM-seq captures critical differences in Cas9 variants

Having validated the accuracy of *Spec/SEAM-seq* SpCas9 models, the techniques were then used to profile the specificity of two Cas9 variants with altered specificity. The first, *S. canis* Cas9 (ScCas9), has been reported to have a single-base specificity within the PAM (5′-NNG-3′), but otherwise similar activity to SpCas9 (19). The second, HiFi-Cas9, is an engineered mutant of SpCas9 (R691A) again with similar on-target activity to WT, but substantially lower off-target activity in human cells (55).

Spec/SEAM-seq were performed on unmodified and catalytically dead versions of these enzymes and compared to SpCas9 in complex with the Nanog gRNA using a slightly revised library (Figure 5 and Supplementary Figure S7). The Affinity versus Activity plots shows that SpCas9 and HiFi-Cas9 have similar profiles (Supplementary Figures S7A, B and S7D, E), but a greater penalty for mismatches on cleavage by HiFi-Cas9 (Figure 5K, L). In contrast, mutations within (black library) and downstream of the reported PAM (purple library) had substantial effects on both the affinity and activity of ScCas9 (Supplementary Figure S7C, F). Thus, rather than a minimal PAM, the PAM of ScCas9 is more extended than SpCas9, and mutations have a greater effect on binding affinity.

The DNA-binding specificity (ln K_A) models for SpCas9 (Figure 5A) and HiFi-Cas9 (Figure 5B) are virtually identical (Figure 5K). However, the effect of mismatches on cleavage is larger for HiFi-Cas9 (Figure 5E versus 5D, and 5L). This difference is more pronounced when cleavage rate is corrected for occupancy effects, and represented in terms of k_{cut} (Figure 5H versus 5G). This indicates that reduced off-target cleavage for HiFi-Cas9 is due to an increase in the penalty of PAM-distal mismatches for cleavage, rather than DNA binding. This enhanced 'proofreading' for PAM-distal mismatches has also been observed in other engineered high-fidelity Cas9s (14,16). Interestingly, binding-independent cleavage specificity by ScCas9 is much more strongly affected by PAM-distal mismatches than either SpCas9 and HiFi-Cas9 (Figure 5I and, Supplementary Figure S7O). The mechanism of underlying this increased specificity is not known.

In addition to the difference in cleavage specificity described above, ScCas9 prefers a different PAM sequence. Compared to the PAM of SpCas9, ScCas9: (i) prefers an A at the −2 position, (ii) strongly prefers G at −3 ($\Delta\Delta G/RT = \sim 8$, compared to ∼6 for other Cas9s), (iii) disfavors T at the −4 position, (iv) prefers G/T at position −5. Thus, rather than the optimal 3′-NGG-5′ for SpCas9, or the shorter 3′-NNG-5′ reported for ScCas9, the ScCas9 PAM is the more extended NAGBK (B = A,C,G; K = G,T) (55). To validate this result, we measured the in-cell editing efficiency for SpCas9- and ScCas9-RNPs over 22 randomly chosen targets in the HRPT gene, half of which contained an NGG PAM, and half the extended NAGBK PAM (Figure 5J). All NGG PAM sites were edited efficiently by SpCas9, whereas only 3/11 were edited by ScCas9. Further, no NABGK PAM sites were edited by SpCas9, whereas all were edited by ScCas9. This provides strong evidence that the PAM distinguishes SpCas9 from ScCas9 targets.

In addition to Nanog, SpCas9, HiFi-Cas9 and ScCas9 were tested for binding and cleavage specificity against FANCF (Supplementary Figure S7D-O). Similar properties were evident for the two target sequences: the binding profiles were similar, with mismatches in the PAM-distal region being more deleterious to cleavage by HiFi-Cas9; ScCas9 exhibited a more extended PAM, albeit with an A preferred at the –5 position, and a substantially greater penalty for mismatches on Δd in the protospacer (Supplementary Figure S7H, K, N). Together, these data provide a mechanistic rationale for reduced off-target cleavage by HiFi-Cas9, and a redefined PAM sequence for ScCas9, highlighting the performance of *Spec-seq*/SEAM-in accurately distinguishing between closely related CRISPR enzymes.

AsCas12a uses a mechanism of target discrimination distinct from that of SpCas9

Acidaminococcus sp. Cas 12a (AsCas12a, aka Cpf1) is a Class II type V CRISPR enzyme that has been reported to have less off-target activity than SpCas9 (56), though it's specificity has not been as well characterized. *Spec/SEAM-seq* was performed using three different gRNAs composed of a AsCas12a specific PAM (TTTA) located immediately 5′ of the protospacers identical to the SpCas9 targets (Nanog, FANCF, VEGFA) (20,24). Reactions were run under similar conditions compared to SpCas9, and the data were analyzed identically.

Consistent with previous work, AsCas12a-RNP affinity extends over a substantially longer footprint than SpCas9-RNPs. The first 17 nucleotide positions of the protospacer contribute to K_A of AsCas12a (Figure 6A), compared to the first ∼11 for SpCas9. As with SpCas9-RNPs, the penalty for mismatches is higher closer to the PAM. Also, in accord with previous work, there is a consistent dip in specificity around position +11 (57). However, AsCas12a::RNPs directed against different targets differ little in their specificities, particularly compared to SpCas9-RNPs. Each AsCas12a::RNP is comparable in mismatch penalty to SpCas9::Nanog (Figure 1F), the most specific among the SpCas9-RNPs. Mechanistically, this suggests that AsCas12a::RNPs binding specificity is not as sensitive to differences in stability between DNA:DNA and RNA:DNA duplexes.

Different from SpCas9, the Mono + NS models for AsCas12a Δd revealed essentially the same pattern of mismatch tolerance as those for the binding affinity (Figure 6B). The region of specificity for Δd extends over the same 17 PAM-proximal nucleotide positions of the protospacer with the same dip in specificity at position +11. Moreover, no mismatch activation is evident for AsCas12a. The model for K_{eff} showed similar trends (Figure 6C, D), suggesting that AsCas12a cleavage specificity is dictated by occupancy of DNA, rather than requiring full RNA:DNA duplexing in PAM-distal regions as observed with SpCas9.

CasX has strict cleavage specificity

CasX (Cas12e) is a recently discovered Class II CRISPR enzyme in the same general family as Cas12a, but with little information available about its specificity (17,18). We

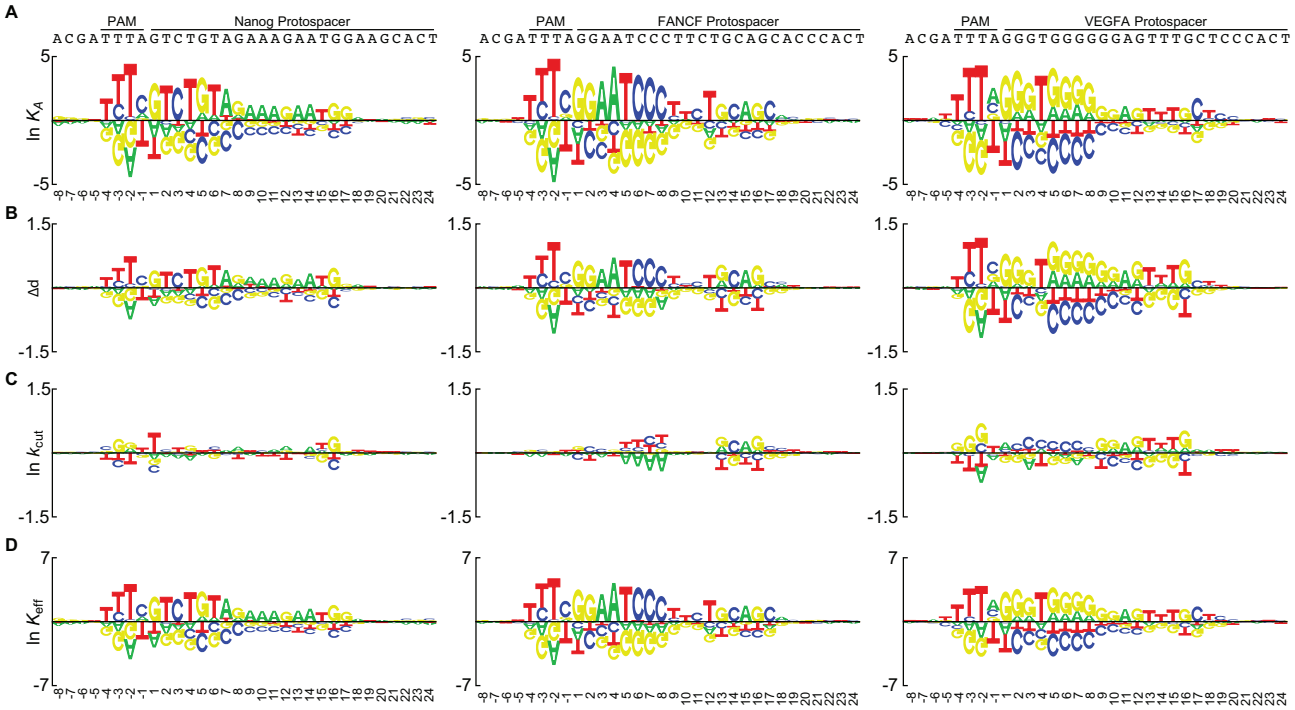

Figure 6. The effect of mismatches on AsCas12a::RNP affinity and cleavage are tightly linked. (**A**) Sequence logo representations of the mononucleotide coefficients in the Mono+NS models of K_A for dAsCas12a::Nanog (left), dAsCas12a::FANCF (center) and dAsCas12a::VEGFA (right) represent the effect of each base on the binding affinity. (**B**) Same as (A) but representing the coefficients in the models of relative log-depletion. (**C**) Same as (A) but showing coefficients (contribution to occupancy independent rate) in the k_{cut} models. (**D**) Logos representing effect of mismatches on the efficiency K_{eff} (see Materials and Methods).

therefore assayed the binding and cleavage specificity of *Deltaproteobacteria* (DpbCasX) for two gRNAs (Target 1 and Target 2, Supplementary Figure S8) used in previous studies (18). Target sites are similar in configuration to AsCas12a, with a TTCA PAM 5′ to the protospacer.

Focusing on Target 1, the binding footprint (10–13 bp) of CasX is similar if slightly longer than SpCas9 (Figure 7), but shorter than AsCas12a (∼17 bp) with little contribution from PAM-distal sequences. Specificity within the PAM is stronger than for SpCas9 and AsCas12a (Figures 1F, 6A, 7A), but is comparable in binding energy to SpCas9:Nanog in the protospacer. To validate this footprint, the DNA-binding affinity of dCasX to four target sites with single mismatches within Target 1 was measured using EMSA. Consistent with the *Spec-seq* model, mutations at position 16 (T16A) and 19 (A19C) retained near full affinity (Figure 7E, F). Mutations in PAM-proximal region (T1G and A10G) has a more pronounced effect on affinity, reflecting what is apparent from the logo: the closer to the PAM, the greater the penalty of a mismatch on affinity. These results demonstrate that the *Spec-seq* accurately delineates the roles of protospacer sub-domains in the binding of CasX, and that CasX is similar in similar in binding specificity to SpCas9 and AsCas12a.

The most striking finding was the effect of mismatches on cleavage. An examination of the Affinity versus Activity plot (Figure 7A) reveals a bimodal distribution in Δd dimension: mutations and outside the PAM and protospacer (orange and pink libraries) affect affinity, but not cleavage, whereas mutations in the PAM or mismatches within the protospacer completely prevent cleavage. The only exceptions to this are mutations at the –1 position of the PAM, and at positions 18–20 in the protospacer. Other than at these positions, the models for Δd (Figure 7C) and k_{cut} indicate that CasX exhibits an all-or-nothing behavior with respect to cleavage: no mismatch is tolerated (Figure 7D). Both the binding and cleavage specificity patterns are also evident for Target 2, suggesting that this specificity is not target sequence dependent (Supplementary Figure S8). To validate the Δd model (Figure 7C), we cleaved the on-target sequence as well as the T1G, A10G, T16G, and the A19C mutants and separated the products on a gel. (Figure 7H, I). No cutting was observed for T1G, A10G, and T16A even when a vast excess of WT-CasX (1 μM) was incubated with DNA (10 nM) for 1 h. Base pairing at position 19 appeared to be unimportant, as cleavage A19C was comparable to on-target. A lower concentration of WT-CasX (100 nM) significantly reduced the cleavage of A19C and on-target, providing evidence that CasX binds weakly to its cognate sequence (∼100 nM) under the condition assayed (Figure 7H).

DISCUSSION

Spec/SEAM-seq is a simple, robust and sensitive method for determining the binding and cleavage specificity of RGEs that has several advantages over previous methods. First, it is fast. If an RGE can be expressed and purified, *Spec-seq* and *SEAM-seq* can be performed side-by-side in hours. Libraries can be amplified immediately and sent for sequencing. Second, it is accessible to almost any lab both in terms of expertise and cost. *Spec/SEAM-seq* requires only

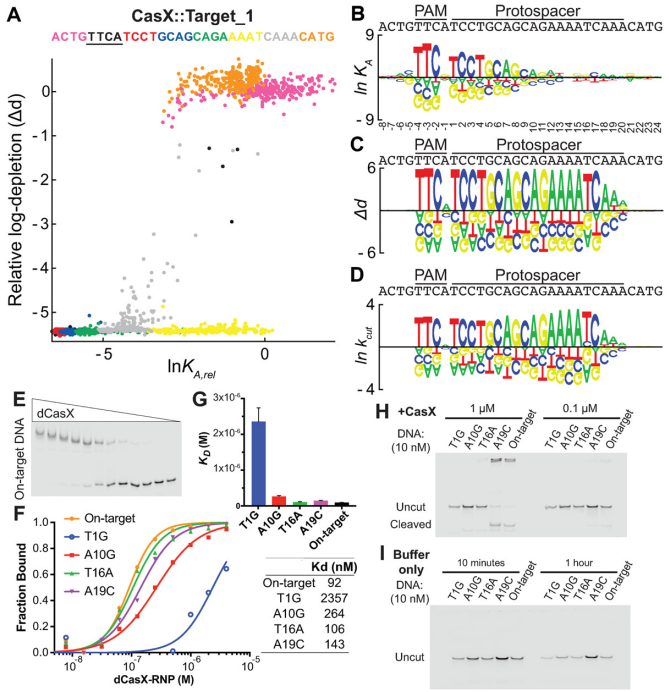

Figure 7. CasX cleavage requires exact matches over most of the PAM and protospacer. (**A**) Affinity versus activity plot (ln K_A versus Δd) for CasX::Target_1. Regions flanking the PAM and protospacer (pink, orange libraries) affect affinity but not cleavage. Almost all other mutations (except PAM @ –1) and mismatches ablate cleavage activity. (**B**) Sequence logo representations of the mononucleotide coefficients for the Mono+NS models of K_A for dCasX::Target_1 represent the effect of each base on the binding affinity. Note that scale is larger than for SpCas9 or AsCas12a (Figures 1 and 6). (**C**) Same as (B) but representing the coefficients in the models of relative log-depletion. (**D**) Same as (B) but showing coefficients (contribution to occupancy independent rate) in the k_{cut} models (see Materials and Methods). Note that the penalty for mismatches in k_{cut} is substantially greater than other RGEs tested. The patterns are consistent with a different target, Target_2 (Supplementary Figure S8). (**E**) Electrophoretic mobility shift assay for measuring the on-target affinity of CasX::Target_1. (**F**, **G**) Fitting of EMSA data for CasX with target mismatches compared to On-target. Numbers correspond to position in the protospacer, error bars represent the standard deviation over three replicates. (**H**) Non-denaturing gel run after cleavage of individual on-target and mismatched sequences with CasX for 1 h, and (**I**) Control gel for the sequences without CasX added.

standard molecular biology equipment (gel boxes, fluorescent imager, PCR machine), and can be performed by researchers with basic training using the protocols and computational scripts provided. *Spec/SEAM-seq* is also cost effective because libraries are composed of 7–8 pools comprising ∼2000 oligos that are subsequently sequenced to a depth of ∼10 million single-end reads. Most importantly, *Spec/SEAM-seq* generates reliable models that dissect the contribution of binding and catalysis to RGE specificity.

Spec/SEAM-seq models are consistent with previous studies

The reliability and utility of *Spec/SEAM-seq* was validated against the most widely used and best characterized system: SpCas9. This enzyme has a complex dependence on its target sequence (26,56,58). *Spec/SEAM-seq* faithfully delineated the disconnect between DNA-binding specificity and mismatch tolerance in cleavage. The DNA-binding specificity of SpCas9 was first measured using an unbiased and comprehensive method: *SELEX-seq/SelexGLM*. Because this method identified specificity downstream of the PAM, *Spec/SEAM-seq* libraries were designed to encompass part of that region. The footprint of the *Spec-seq* binding model was consistent with the *SelexGLM* model, indicating that the PAM, 10–11bp in the PAM proximal region, and sequences downstream of the PAM contribute to binding. This footprint is consistent with some non-equilibrium studies, and is somewhat longer than the previously reported 8–10 bp (8,25). The energetic models are also consistent with the mismatch-specific penalties for binding (10). The models for cleavage (Δd) indicated that base pairing throughout the protospacer is critical for full activity.

Most striking is the correlation between the estimated efficiency of SpCas9, K_{eff}, for off-target cleavage and the Doench rules. Our calculation of K_{eff} was inspired by Michaelis–Menten kinetics. The efficiency of an enzyme is the rate of conversion of substrate to product (k_{cat}) divided by the concentration of substrate at half maximal velocity (K_m), which is approximately equal to K_D ($1/K_A$) if the substrate binds quickly (or k_{cat}*K_A). Using Spec/SEAM-seq the relative affinity (K_A) and the saturation-corrected cleavage (k_{cut}) are calculated, with the product (k_{cut}*K_A) equal to K_{eff}, allowing estimation of the relative efficiency of cleavage for any sequence. In the cell, where potential cleavage sites vastly outnumber the amount of SpCas9-RNP, the probability of cleavage is thus proportional to K_{eff} for each site. The relative K_{eff} for each SpCas9-RNP correlated well with the Doench CFD (Figure 4), defined as the effect of mismatches across the protospacer on cleavage *in vivo* averaged across a number of target sites (41), indicating that the *in vitro* models of efficiency are consistent with *in vivo* off-target activity. This is a key validation and shows that *Spec/SEAM-seq* accurately models the position-specific effect of mismatches on relative efficiency of off-target cleavage.

Lack of DNA-binding specificity and mismatch activation contribute to SpCas9 off-target cleavage

Decoupling the effect of mismatches on affinity and cleavage provided mechanistic information about the source of off-target cleavage. First, the free energy-based models for binding revealed that the penalty for mismatches within the seed region are not consistent among target sequences. The specificity of SpCas9::VEGFA is much less than SpCas9::Nanog and SpCas9::FANCF. This correlates with the energy of exchange ($\Delta\Delta G_{exchange}$) from a DNA:DNA duplex to an RNA:DNA duplex (Figure 1L). $\Delta\Delta G_{exchange}$ had been previously hypothesized to contribute to specificity (59), but binding models were not yet precise enough to support this conclusion. This lack of specificity for VEGFA compared to FANCF is supported by three in vivo studies using different techniques (27–29). This indicates that binding specificity, the free-energy difference between on- and off-target binding, varies substantially between target sequences, and is an important determinant of off-target cleavage.

Paired measurement of affinity and cleavage revealed an unanticipated source of off-target cleavage for SpCas9;

mismatch-activation. Mismatch-activation results from incomplete cleavage of the on-target sequence, but more complete cleavage of an off-target sequence. This was evident in only one of the three targets tested (Nanog), but was consistent not only between replicates, but with a variant of SpCas9, HiFi-Cas9. Incomplete cleavage has also been observed by other groups (21), suggesting that a portion of the bound complex is trapped in an inactive conformation. We speculate that mismatches allow escape from this inactive conformation, resulting in more complete cleavage. In a cellular context where SpCas9 is overexpressed or expressed over a long period of time, mismatch-activation could favor off-target cleavage.

Although $\Delta\Delta G_{\text{exchange}}$ and mismatch activation are potential sources of off-target cleavage, too few target sequences have been tested to reliably be modeled in SpCas9::gRNA design tools (60). As a result, for applications that demand high precision, such as stem cell editing, potential SpCas9::gRNA pairs can be screened using *Spec/SEAM-seq* to avoid promiscuous cutters and select for high activity RNPs.

Spec/SEAM-seq models provide insight into the specificity of other CRISPR enzymes

The reliability of the models generated by *Spec/SEAM-seq* extends to other RGEs, as evidenced by AsCas12a and HiFi-Cas9. AsCas12a has been shown to derive cleavage specificity from a longer footprint (+1 to +18 in the protospacer) (57,61) than SpCas9. *Spec/SEAM-seq* captures this specificity, including a characteristic dip in specificity near the non-target strand cleavage site (+11 to +12). *Spec/SEAM-seq* indicates that, in contrast to SpCas9, cleavage specificity is driven almost exclusively by binding specificity (Figure 6A, B), as the binding-independent cleavage effects of mismatches are minimal (Figure 6C). AsCas12a therefore appears to be more specific overall, and would be a useful RGE for applications that require specific binding in particular, such as CRISPRa/i (62). HiFi-Cas9 has been shown to have fewer off-target cleavage sites *in vivo* (19). The lnK_A models for HiFi-Cas9 show similar binding specificity compared to SpCas9, however cleavage (k_{cut}) is clearly more strongly impaired by mismatches, particularly in the PAM-distal region (Figure 5H).

Spec/SEAM-seq may be most useful in screening of newly isolated RGEs for specificity, and in providing mechanistic insight. To demonstrate this utility, two less-well characterized RGEs were tested: ScCas9 and CasX. Consistent with a previous study, *Spec/SEAM-seq* determined that ScCas9 has a more relaxed specificity at the -2 position within the 3 bp PAM (NNG versus NGG) (55), but revealed a more extended PAM (5′-NAGBK-3′) (Figure 5C, Supplementary Figure S7H). As a result, ScCas9 does not have more flexibility in PAM sequence for binding and cleavage, but rather is predicted to have increased specificity from a greater contribution by –3 G and overall throughout an extended PAM (Figure 5C, Supplementary Figure S7H). Alternatively, the increase in ScCas9 specificity may enable use in precise applications where an extended PAM is present.

DpbCasX is a recently discovered and characterized RGE, for which little was known about specificity (17). Similar to AsCas12a, has a single RuvC nuclease domain that undergoes a conformational change to make staggered cuts in the dsDNA target (18). CasX has two additions domains that contact the non-target strand (NTSB) and target strand (TSL) of the target DNA. The DNA-binding specificity footprint of CasX is nonetheless smaller than AsCas12a (∼10–13 bp versus ∼17 bp), albeit with more pronounced specificity within a different PAM (TTCN versus TTTC) (Figures 6A, 7B and Supplementary Figure S8C, D). However, whereas AsCas12a specificity is driven by binding, CasX catalytic activity has an almost absolute requirement for TTC within the PAM and base pairing from +1 to +17 within the protospacer (Figure 7C, D, and Supplementary Figure S8E–G). This intolerance to mutations and mismatches at all but the -1 position in the PAM render DbpCasX the least prone to off-target cleavage of any of the RGEs tested. This behavior suggests that the NTSB and TSL of CasX aid in 'proofreading' base pairing throughout the protospacer for cleavage rather than aiding in binding. Although many factors affect whether an RGE will be useful in cellular editing (persistence of the gRNA, protein size, and stability), CasX is a promising candidate for applications that demand high precision.

DATA AVAILABILITY

Reads have been deposited to the Sequence Read Archive, accession number: PRJNA547810, PRJNA608749

SUPPLEMENTARY DATA

Supplementary Data are available at NAR Online.

ACKNOWLEDGEMENTS

Author contributions: L.Z. conceived of the study, designed and executed all of the experiment and drafted the manuscript. H.T.R. designed and performed the computational modeling and edited the manuscript. C.A.V. and M.A.B. provided support and guidance for the Cas12a studies and individual sequence testing. H.J.B. and M.A.P. supported and supervised the work, edited the manuscript and performed some data analysis.

FUNDING

National Institutes of Health [R01HG003008 to H.J.B and H.T.R.]; Vagelos Precision Medicine Pilot Program at Columbia University; Columbia University's Shared Research Computing Facility is supported by NIH grant [G20RR030893]; NYSTAR contract [C090171]; University of Iowa was entirely supported by NSF-CAREER [MCB-1552862 to L.Z and M.A.P.]; The Cas12a and individual sequence testing by L.Z., C.A.V. and M.A.B. was supported by Integrated DNA Technologies, Inc. (IDT). Funding for open access charge: NSF.

Conflict of interest statement. L.Z., C.A.V. and M.A.B. are employees of Integrated DNA Technologies (IDT). The authors declare no competing interests.

REFERENCES

1. 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.
2. 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.
3. Mali,P., Aach,J., Stranges,P.B., Esvelt,K.M., Moosburner,M., Kosuri,S., Yang,L. and Church,G.M. (2013) CAS9 transcriptional activators for target specificity screening and paired nickases for cooperative genome engineering. *Nat. Biotechnol.*, **31**, 833–838.
4. Mali,P., Esvelt,K.M. and Church,G.M. (2013) Cas9 as a versatile tool for engineering biology. *Nat. Methods*, **10**, 957–963.
5. Cheng,A.W., Wang,H., Yang,H., Shi,L., Katz,Y., Theunissen,T.W., Rangarajan,S., Shivalila,C.S., Dadon,D.B. and Jaenisch,R. (2013) Multiplexed activation of endogenous genes by CRISPR-on, an RNA-guided transcriptional activator system. *Nat. Publish. Group*, **23**, 1163–1171.
6. Hilton,I.B., D'Ippolito,A.M., Vockley,C.M., Thakore,P.I., Crawford,G.E., Reddy,T.E. and Gersbach,C.A. (2015) Epigenome editing by a CRISPR-Cas9-based acetyltransferase activates genes from promoters and enhancers. *Nat. Biotechnol.*, **33**, 510–517.
7. Konermann,S., Brigham,M.D., Trevino,A.E., Joung,J., Abudayyeh,O.O., Barcena,C., Hsu,P.D., Habib,N., Gootenberg,J.S., Nishimasu,H. *et al.* (2015) Genome-scale transcriptional activation by an engineered CRISPR-Cas9 complex. *Nature*, **517**, 583–588.
8. 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.
9. Kuscu,C., Arslan,S., Singh,R., Thorpe,J. and Adli,M. (2014) Genome-wide analysis reveals characteristics of off-target sites bound by the Cas9 endonuclease. *Nat. Biotechnol.*, **32**, 677–683.
10. 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. U.S.A.*, **114**, 5461–5466.
11. 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**, 839–837.
12. 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.
13. 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.* (2014) GUIDE-seq enables genome-wide profiling of off-target cleavage by CRISPR-Cas nucleases. *Nat. Biotechnol.*, **33**, 187–197.
14. 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.
15. 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.
16. 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.
17. Burstein,D., Harrington,L.B., Strutt,S.C., Probst,A.J., Anantharaman,K., Thomas,B.C., Doudna,J.A. and Banfield,J.F. (2017) New CRISPR-Cas systems from uncultivated microbes. *Nature*, **542**, 237–241.
18. Liu,J.J., Orlova,N., Oakes,B.L., Ma,E., Spinner,H.B., Baney,K.L.M., Chuck,J., Tan,D., Knott,G.J., Harrington,L.B. *et al.* (2019) CasX enzymes comprise a distinct family of RNA-guided genome editors. *Nature*, **566**, 218–223.
19. 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.
20. Zetsche,B., Gootenberg,J.S., Abudayyeh,O.O., Slaymaker,I.M., Makarova,K.S., Essletzbichler,P., Volz,S.E., Joung,J., van der Oost,J., Regev,A. *et al.* (2015) Cpf1 is a single RNA-Guided endonuclease of a class 2 CRISPR-Cas system. *Cell*, **163**, 759–771.
21. Gong,S., Yu,H.H., Johnson,K.A. and Taylor,D.W. (2018) DNA unwinding is the primary determinant of CRISPR-Cas9 activity. *CellReports*, **22**, 359–371.
22. Szczelkun,M.D., Tikhomirova,M.S., Sinkunas,T., Gasiunas,G., Karvelis,T., Pschera,P., Siksnys,V. and Seidel,R. (2014) Direct observation of R-loop formation by single RNA-guided Cas9 and Cascade effector complexes. *Proc. Natl. Acad. Sci. U.S.A.*, **111**, 9798–9803.
23. 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.
24. Singh,D., Mallon,J., Poddar,A., Wang,Y., Tippana,R., Yang,O., Bailey,S. and Ha,T. (2018) Real-time observation of DNA target interrogation and product release by the RNA-guided endonuclease CRISPR Cpf1 (Cas12a). *Proc. Natl. Acad. Sci. U.S.A.*, **115**, 5444–5449.
25. Singh,D., Sternberg,S.H., Fei,J., Doudna,J.A. and Ha,T. (2016) Real-time observation of DNA recognition and rejection by the RNA-guided endonuclease Cas9. *Nat. Commun.*, **7**, 12778.
26. O'Geen,H., Yu,A.S. and Segal,D.J. (2015) How specific is CRISPR/Cas9 really? *Curr. Opin. Chem. Biol.*, **29**, 72–78.
27. 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.
28. 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.
29. Cameron,P., Fuller,C.K., Donohoue,P.D., Jones,B.N., Thompson,M.S., Carter,M.M., Gradia,S., Vidal,B., Garner,E., Slorach,E.M. *et al.* (2017) Mapping the genomic landscape of CRISPR-Cas9 cleavage. *Nat. Methods*, **14**, 600–606.
30. Wienert,B., Wyman,S.K., Richardson,C.D., Yeh,C.D., Akcakaya,P., Porritt,M.J., Morlock,M., Vu,J.T., Kazane,K.R., Watry,H.L. *et al.* (2019) Unbiased detection of CRISPR off-targets in vivo using DISCOVER-Seq. *Science*, **364**, 286–289.
31. O'Geen,H., Henry,I.M., Bhakta,M.S., Meckler,J.F. and Segal,D.J. (2015) A genome-wide analysis of Cas9 binding specificity using ChIP-seq and targeted sequence capture. *Nucleic Acids Res.*, **43**, 3389–3404.
32. Jung,C., Hawkins,J.A., Jones,S.K. Jr, 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.
33. 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.
34. Haeussler,M., Schönig,K., Eckert,H., Eschstruth,A., Mianné,J., Renaud,J.-B., Schneider-Maunoury,S., Shkumatava,A., Teboul,L., Kent,J. *et al.* (2016) Evaluation of off-target and on-target scoring algorithms and integration into the guide RNA selection tool CRISPOR. *Genome Biol.*, **17**, 148.
35. Abadi,S., Yan,W.X., Amar,D. and Mayrose,I. (2017) A machine learning approach for predicting CRISPR-Cas9 cleavage efficiencies and patterns underlying its mechanism of action. *PLoS Comput. Biol.*, **13**, e1005807.
36. 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 USA*, **115**, 9351–9358.
37. Horlbeck,M.A., Witkowsky,L.B., Guglielmi,B., Replogle,J.M., Gilbert,L.A., Villalta,J.E., Torigoe,S.E., Tjian,R. and Weissman,J.S. (2016) Nucleosomes impede Cas9 access to DNA in vivo and in vitro. *Elife*, **5**, e12677.
38. Kallimasioti-Pazi,E.M., Thelakkad Chathoth,K., Taylor,G.C., Meynert,A., Ballinger,T., Kelder,M.J.E., Lalevee,S., Sanli,I., Feil,R.

and Wood,A.J. (2018) Heterochromatin delays CRISPR-Cas9 mutagenesis but does not influence the outcome of mutagenic DNA repair. *PLoS Biol.*, **16**, e2005595.
39. van Overbeek,M., Capurso,D., Carter,M.M., Thompson,M.S., Frias,E., Russ,C., Reece-Hoyes,J.S., Nye,C., Gradia,S., Vidal,B. *et al.* (2016) DNA repair profiling reveals nonrandom outcomes at Cas9-Mediated breaks. *Mol. Cell*, **63**, 633–646.
40. Allen,F., Crepaldi,L., Alsinet,C., Strong,A.J., Kleshchevnikov,V., De Angeli,P., Páleníková,P., Khodak,A., Kiselev,V., Kosicki,M. *et al.* (2018) Predicting the mutations generated by repair of Cas9-induced double-strand breaks. *Nat. Biotechnol.*, **37**, 64.
41. Doench,J.G., Fusi,N., Sullender,M., Hegde,M., Vaimberg,E.W., Donovan,K.F., Smith,I., Tothova,Z., Wilen,C., Orchard,R. *et al.* (2016) Optimized sgRNA design to maximize activity and minimize off-target effects of CRISPR-Cas9. *Nat. Biotechnol.*, **34**, 184–191.
42. Stormo,G.D., Zuo,Z. and Chang,Y.K. (2015) Spec-seq: determining protein-DNA-binding specificity by sequencing. *Brief. Funct. Genomics*, **14**, 30–38.
43. Zuo,Z. and Stormo,G.D. (2014) High-resolution specificity from DNA sequencing highlights alternative modes of Lac repressor binding. *Genetics*, **198**, 1329–1343.
44. Anders,C. and Jinek,M. (2014) In vitro enzymology of Cas9. *Methods Enzymol.*, **546**, 1–20.
45. Zhang,L., Martini,G.D., Rube,H.T., Kribelbauer,J.F., Rastogi,C., FitzPatrick,V.D., Houtman,J.C., Bussemaker,H.J. and Pufall,M.A. (2018) SelexGLM differentiates androgen and glucocorticoid receptor DNA-binding preference over an extended binding site. *Genome Res.*, **28**, 111–121.
46. Riley,T.R., Lazarovici,A., Mann,R.S. and Bussemaker,H.J. (2015) Building accurate sequence-to-affinity models from high-throughput in vitro protein-DNA binding data using FeatureREDUCE. *eLife*, **4**, 307.
47. Foat,B.C., Morozov,A.V. and Bussemaker,H.J. (2006) Statistical mechanical modeling of genome-wide transcription factor occupancy data by MatrixREDUCE. *Bioinformatics*, **22**, e141-9.
48. Jacobi,A.M., Rettig,G.R., Turk,R., Collingwood,M.A., Zeiner,S.A., Quadros,R.M., Harms,D.W., Bonthuis,P.J., Gregg,C., Ohtsuka,M. *et al.* (2017) Simplified CRISPR tools for efficient genome editing and streamlined protocols for their delivery into mammalian cells and mouse zygotes. *Methods*, **121–122**, 16–28.
49. Perez-Pinera,P., Kocak,D.D., Vockley,C.M., Adler,A.F., Kabadi,A.M., Polstein,L.R., Thakore,P.I., Glass,K.A., Ousterout,D.G., Leong,K.W. *et al.* (2013) RNA-guided gene activation by CRISPR-Cas9–based transcription factors. *Nat. Methods*, **10**, 973–976.
50. Slattery,M., Riley,T., Liu,P., Abe,N., Gomez-Alcala,P., Dror,I., Zhou,T., Rohs,R., Honig,B., Bussemaker,H.J. *et al.* (2011) Cofactor binding evokes latent differences in DNA binding specificity between Hox proteins. *Cell*, **147**, 1270–1282.
51. Huai,C., Li,G., Yao,R., Zhang,Y., Cao,M., Kong,L., Jia,C., Yuan,H., Chen,H., Lu,D. and Huang,Q. (2017) Structural insights into DNA cleavage activation of CRISPR-Cas9 system. *Nat. Commun.*, **8**, 1375.
52. Cameron,P., Fuller,C.K., Donohoue,P.D., Jones,B.N., Thompson,M.S., Carter,M.M., Gradia,S., Vidal,B., Garner,E., Slorach,E.M. *et al.* (2017) Mapping the genomic landscape of CRISPR-Cas9 cleavage. *Nat. Methods*, **14**, 600–606.
53. Lesnik,E.A. and Freier,S.M. (1995) Relative thermodynamic stability of DNA, RNA, and DNA:RNA hybrid duplexes: relationship with base composition and structure. *Biochemistry*, **34**, 10807–10815.
54. Dumousseau,M., Rodriguez,N., Juty,N. and Le Novère,N. (2012) MELTING, a flexible platform to predict the melting temperatures of nucleic acids. *BMC Bioinformatics*, **13**, 101.
55. Chatterjee,P., Jakimo,N. and Jacobson,J.M. (2018) Minimal PAM specificity of a highly similar SpCas9 ortholog. *Sci. Adv.*, **4**, eaau0766.
56. Swarts,D.C. and Jinek,M. (2018) Cas9 versus Cas12a/Cpf1: structure-function comparisons and implications for genome editing. *Wiley Interdiscip. Rev. RNA*, e1481.
57. 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.
58. 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.
59. Farasat,I. and Salis,H.M. (2016) A Biophysical model of CRISPR/Cas9 activity for rational design of genome editing and gene regulation. *PLoS Comput. Biol.*, **12**, e1004724.
60. Bradford,J. and Perrin,D. (2019) A benchmark of computational CRISPR-Cas9 guide design methods. *PLoS Comput. Biol.*, **15**, e1007274.
61. 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.
62. Gilbert,L.A., Horlbeck,M.A., Adamson,B., Villalta,J.E., Chen,Y., Whitehead,E.H., Guimaraes,C., Panning,B., Ploegh,H.L., Bassik,M.C. *et al.* (2014) Genome-scale CRISPR-mediated control of gene repression and activation. *Cell*, **159**, 647–661.